CERN Datacenter Network Virtualization

Ricardo Rocha - CERN IT-CM

Pre-GDB Workshop Datacenter Network Architectures
https://indico.cern.ch/event/1028690/

General / CERN OpenStack overview

Last 30 days | 15m

Cloud resources

Used	Available	Used	Available	Used	Available
297.1 K cores	257.6 K cores	893.5 TiB RAM	976.3 TiB RAM	7.6 PiB disk	12.5 PiB disk

Openstack services stats

Users	Projects	VMs	Magnum clusters	Hypervisors	Images	Baremetal Nodes
3434	4582	28096	655	5903	3046	7393

Volumes	Volume size	Fileshares	Fileshares size
7018	3.20 PiB	2984	703 TiB

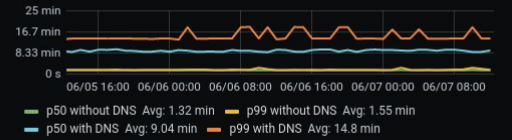
Resource overview by time

VMs created/deleted (last 7days)

Shared cells availability (last 30days)

Total VMs

Average VM boot time



VM changes

Hypervisors

Projects and users

Magnum clusters

Cores

Memory

Disk

Baremetal Nodes

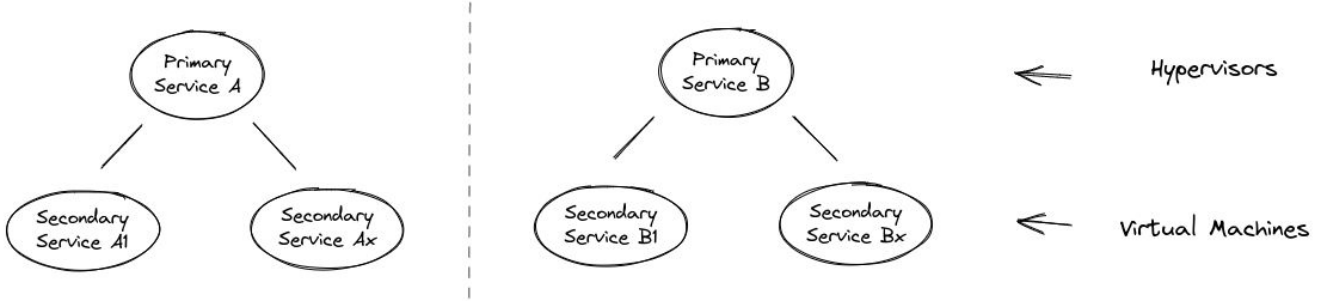
Existing OpenStack Networking

Flat, Provider Network

Network segmentation (L2) into *IP services*, multiple broadcast domains (scalability)

Extensions to OpenStack Neutron LinuxBridge Driver

- Clusters represent Primary Services
- Neutron Query Extension: *“Give me available subnets for host XYZ”*



Software Defined Networking

Project involving people from CERN IT-CM, IT-CS, Security Team

Goal: Improve isolation and flexibility with software defined networks (SDN)

https://sdn.docs.cern.ch (CERN only)

Motivation

IP Mobility

- Overcome existing limitations for IP allocation
  - Can only be done once a hypervisor has been selected
  - VM migration is limited to the same IP service

Floating and Virtual IPs

- Externally assigned and movable IP assignments
- LBaaS, 1-1 NAT, ...

Motivation

Improved Isolation

- Currently limited to Technical Network (TN), ITS (services), LCG
- Networks on demand per project or group of projects

Security Groups

- Firewall as a Service (FWaaS)
- Named network access rules (inbound and outbound)
- Management can be delegated to external entity - e.g. security team

Motivation

Hardware Repurposing

Simplified reassignment of resources accessing different domains

Load Balancing as a Service (LBaaS)

IP based, managed fully by the network layer

Improvement over the existing DNS load balancing

Use Cases

Bastion Hosts

Isolation of personal VMs in ITS

Isolation / network separation for external projects

Firewall as a Service (FWaaS)

- Default firewalls for all / sets of machines
- Access control by security team / operators to sets of machines

Simplified Pacemaker / Corosync style setups

...

Initial Evaluation

	Neutron/OpenVSwitch	OpenDaylight	OVN
DHCP	Neutron	Neutron/Built-in	Built-in
Floating IPs	Yes	Yes	Yes
Distributed Routing	Only with DVR	Yes	Yes
Tunneling Protocols	**vxlan** / **GRE** / geneve	**vxlan** / GRE / geneve	vxlan / **geneve**
Security Groups	IPTables	OpenFlow Native	OpenFlow Native + Logging
Load Balancing	Octavia	Octavia	Octavia / OVN Native
Acceleration	Limited DPDK	DPDK	DPDK
Tracing	tcpdump	tcpdump	ovn-trace
Physical Switch Integr.	L2 / L3	L2 / L3	L2 / L3

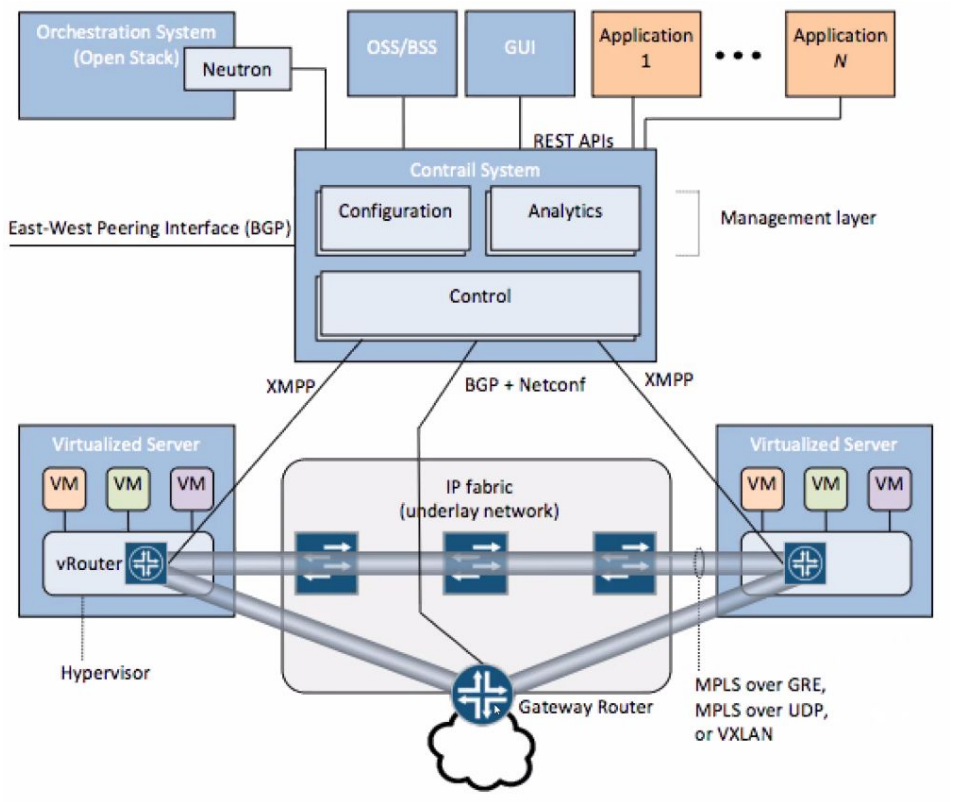
Things Today

https://tungsten.io/

We chose Tungsten Fabric (formerly OpenContrail) as controller

A Linux Foundation project, under LF Networking

Backed by Juniper with contributions from other companies


Orchestration System
(Open Stack)
Neutron
OSS/BSS
GUI
Application
1
Application
N
REST APIs
Contrail System
Configuration
Analytics
Control
Management layer
East-West Peering Interface (BGP)
XMPP
BGP + Netconf
XMPP
Virtualized Server
VM
VM
VM
vRouter
IP fabric
(underlay network)
Virtualized Server
VM
VM
VM
Hypervisor
Gateway Router
MPLS over GRE,
MPLS over UDP,
or VXLAN

What is there

We have deployed two new OpenStack regions with software defined networks

- Separate from our production regions, growing according to demand

sdn1

- Handled in production mode
- Functionality limited to Load Balancing as a Service (LBaaS)

sdn2

- Used for validating changes and testing new functionality
- Functionality includes the full SDN spectrum (private networks, floating IPs, …)

Major Milestones

2021 Jan 05: L7 Support and Security Groups Added to LBaaS

2020 Apr 20: Load Balancing as a Service (LBaaS) Generally Available

2019 Oct 18: Load Balancing as a Service (LBaaS) Pre-Production

Deployment

Control plane deployed on Kubernetes

- OpenStack deployment relying on OpenStack Helm
- Tungsten Fabric deployment using Contrail Helm Deployer

(Some) Hypervisor components also containerized

- Tungsten vRouter Agent and Node Manager
- Nova Compute managed by Puppet

Load Balancing as a Service (LBaaS)

Over 180 LB instances today, in 30+ different projects

https://clouddocs.web.cern.ch/networking/load_balancing.html

Heavy duty handled by Tungsten, with an OpenStack Neutron API on top

Integrated with Kubernetes clusters to offer serviceType: LB (very popular)

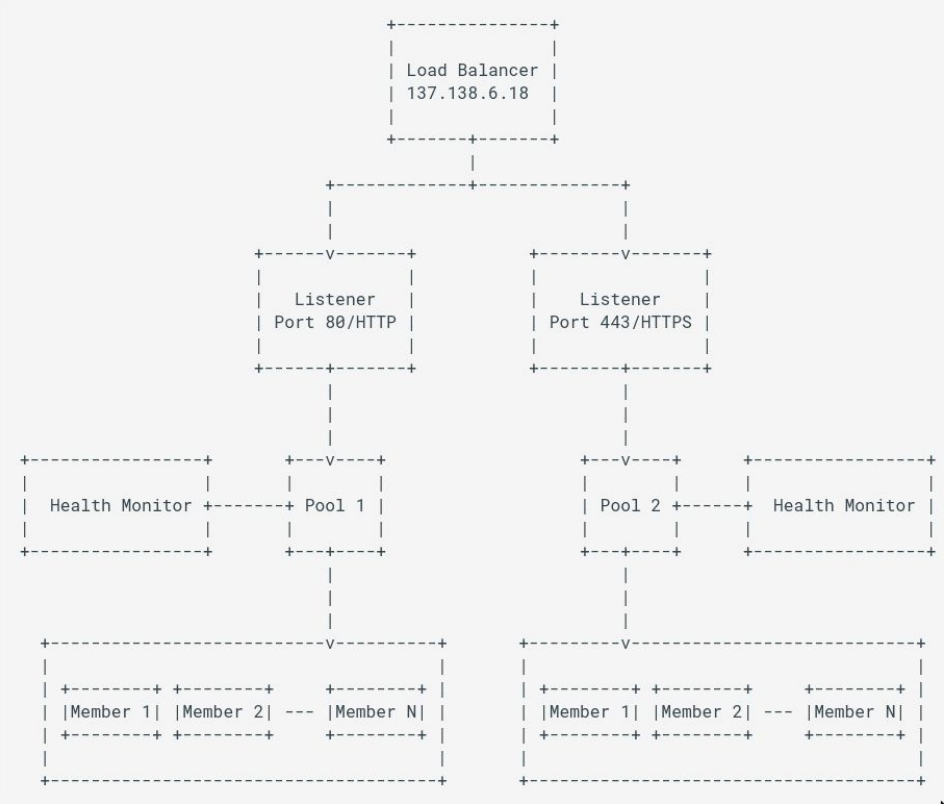
Multiple LB types

- TCP (passthrough), TERMINATED HTTPS, HTTP, HTTPS

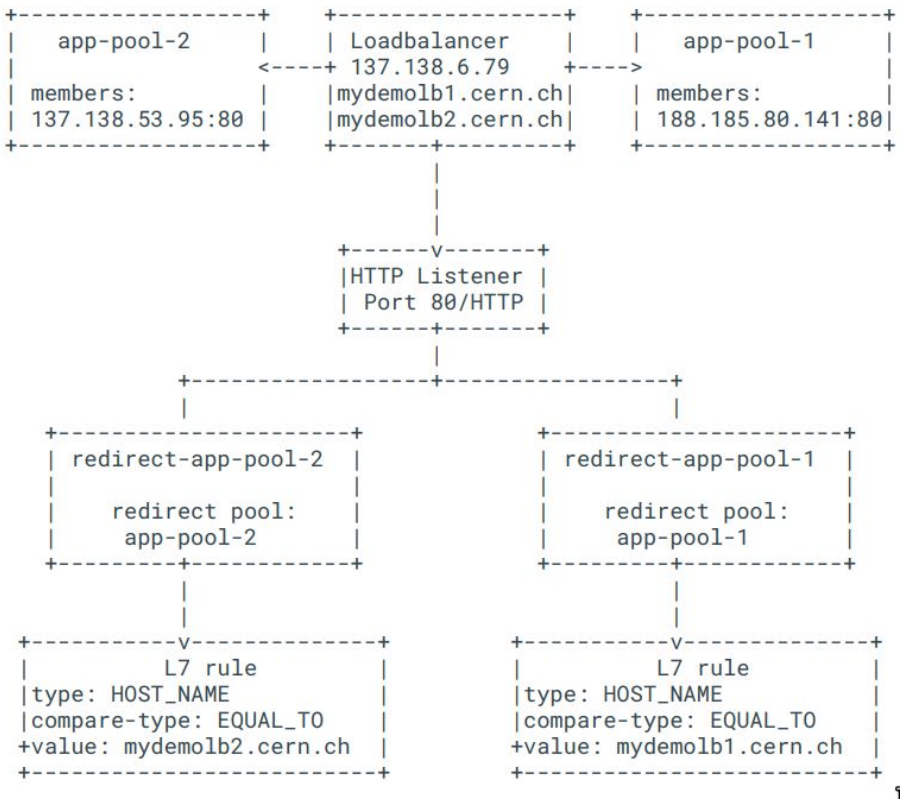
L7 and security groups available since December 2020

- (IT ASDF) https://indico.cern.ch/event/976468/

Load Balancing as a Service (LBaaS)



Load Balancing as a Service (LBaaS)



Software Defined Networks

Validation finished, starting to onboard use cases

Private Networks

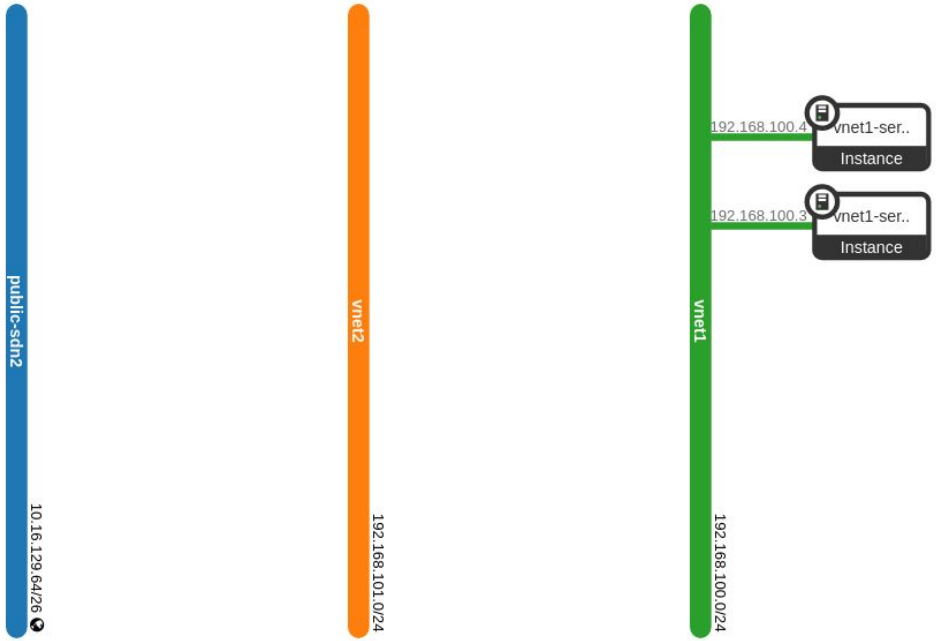
Virtual Routing

Floating and Virtual IPs

Security Groups

public-sdn2
10.16.129.64/26
vnet1
192.168.100.0/24
192.168.100.4
vnet1-ser..
Instance
192.168.100.3
vnet1-ser..
Instance

Create network vnet2

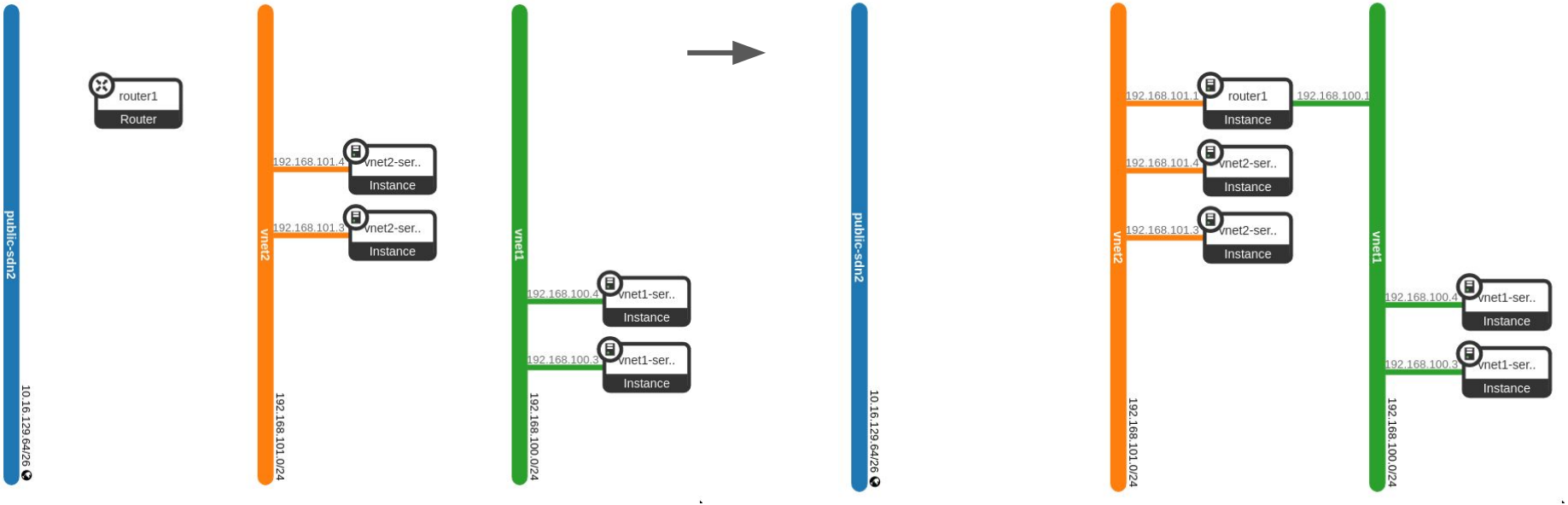



```
$ openstack network create vnet2

$ openstack subnet create --subnet-range 192.168.101.0/24 --network vnet2 subnet2
```

Create vnet2 virtual machines

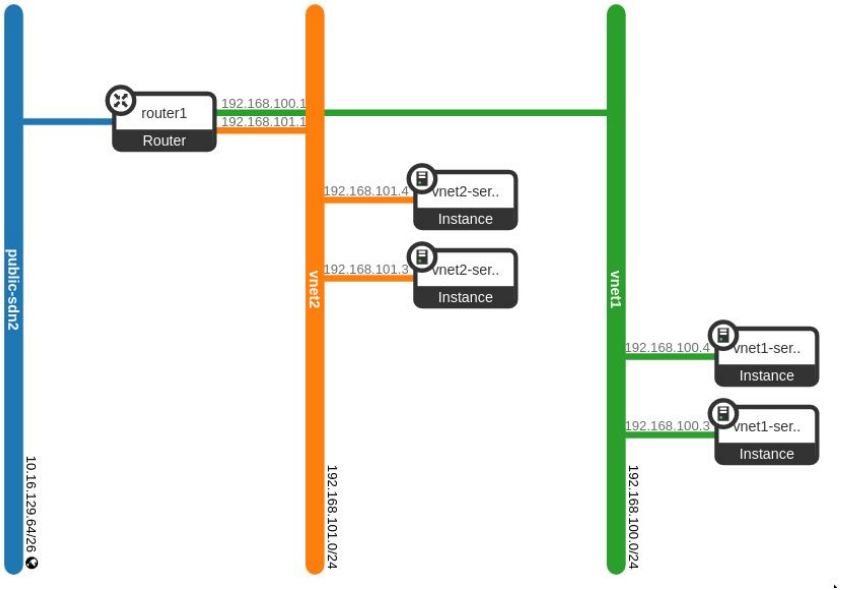

```
$ openstack server create vnet2-server1 --image cirros-image --flavor m2.small
--network vnet2 --property cern-services=false

$ openstack server create vnet2-server2 --image cirros-image --flavor m2.small
--network vnet2 --property cern-services=false
```

Add connectivity between vnet1 and vnet2 (virtual router)



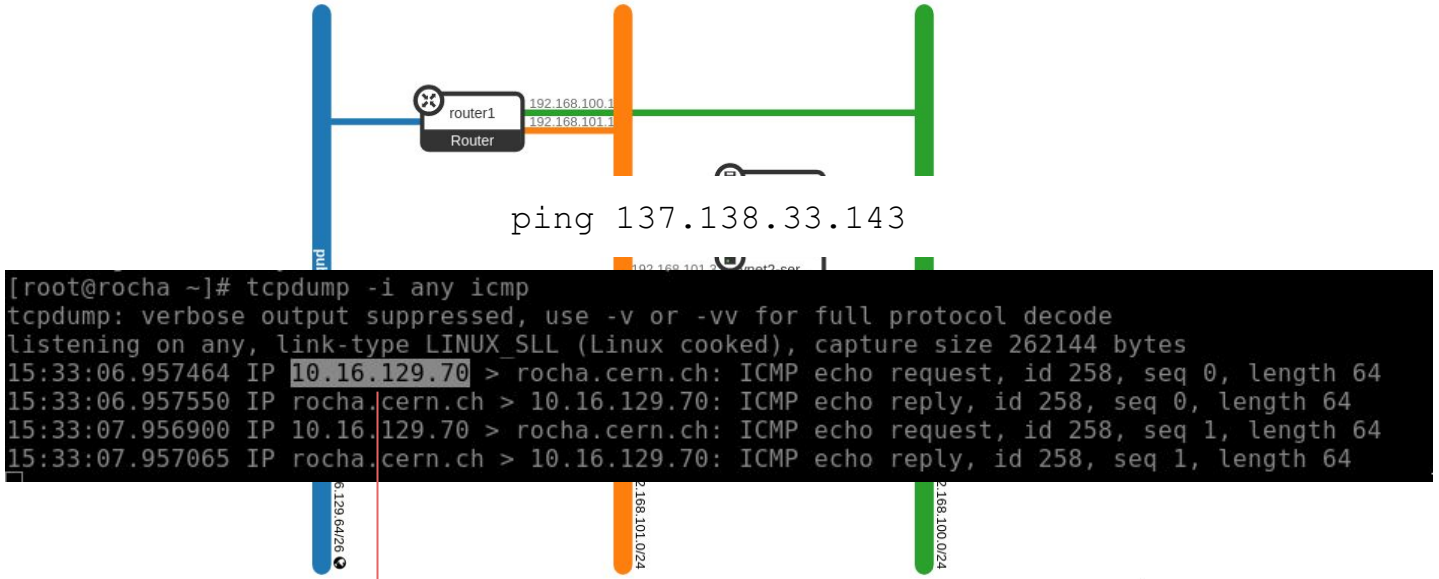

```
$ openstack router create router1

$ openstack router add subnet router1 subnet1

$ openstack router add subnet router1 subnet2
```

External connectivity - SNAT N-to-1 NAT




```
$ openstack router set --external-gateway public-sdn2 router1
```

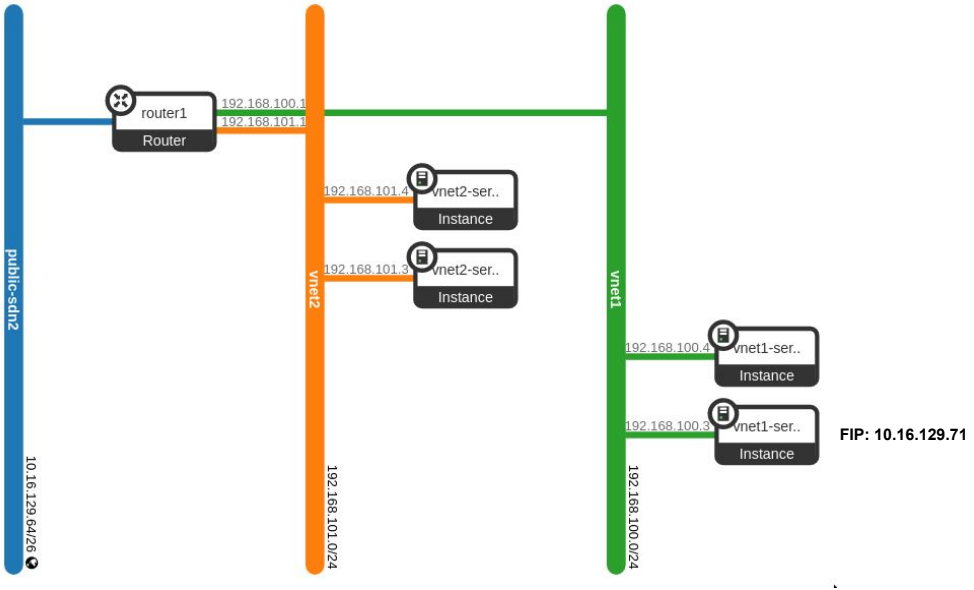
External connectivity - SNAT N-to-1 NAT




```
$ openstack router set --external-gateway public-sdn2 router1
```

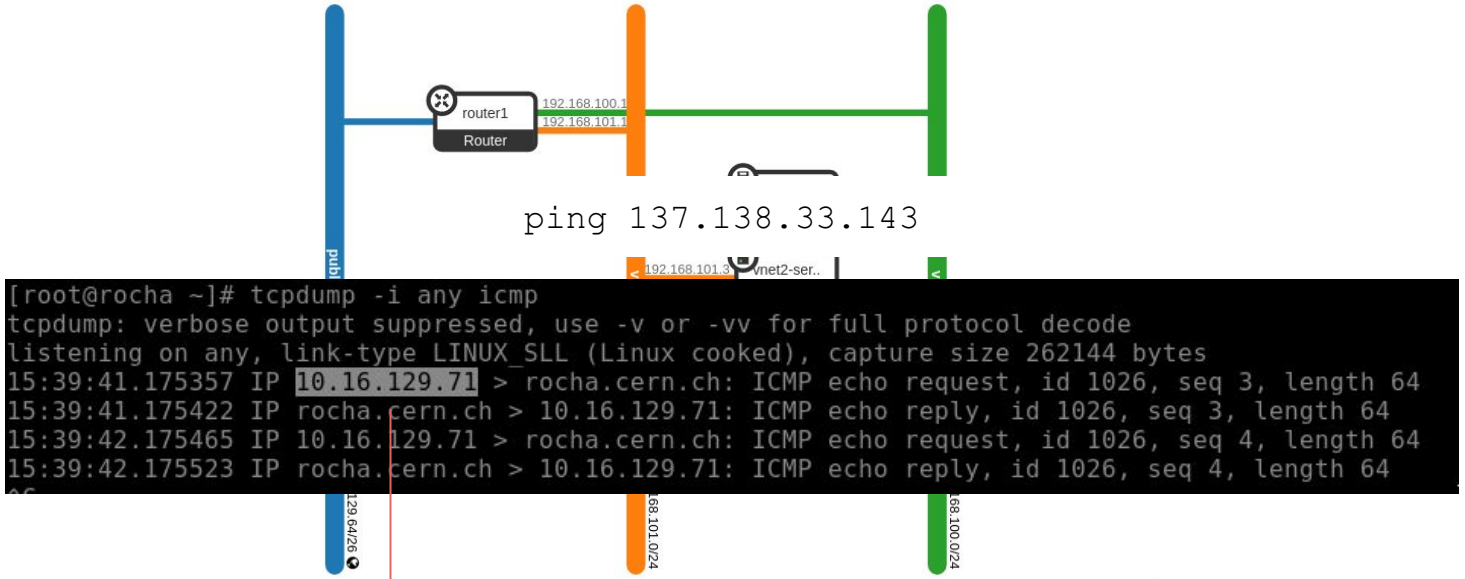
Hypervisor IP

External connectivity - Floating IPs 1-to-1 NAT




```
$ openstack floating ip create public-sdn2

$ openstack server add floating ip vnet1-server1 2b085fe0-338f-42d...1d1cca9
```

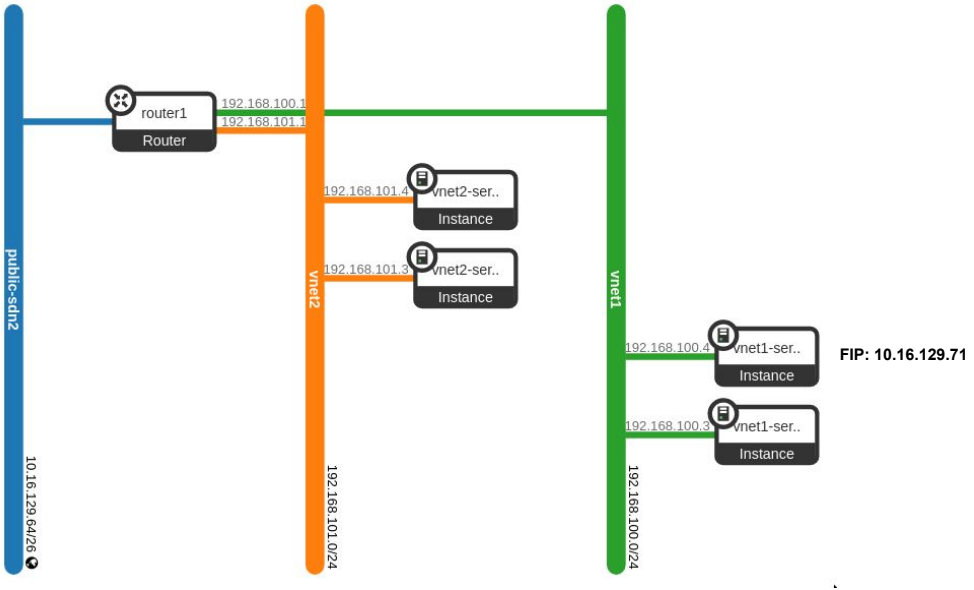
External connectivity - Floating IPs 1-to-1 NAT




```
$ openstack floating ip create public-sdn2

$ openstack server add floating ip vnet1-server1 2b085fe0-338f-42d...1d1cca9
```

Floating IP

Re-assigning Floating IPs




```
$ openstack server remove floating ip vnet1-server1 2b085fe0-338f-42d...1d1cca9

$ openstack server add floating ip vnet1-server2 2b085fe0-338f-42d...1d1cca9
```

Security Groups

Sets of named firewall rules that are applied to networks / ports

Can be updated on the fly, and/or managed by central teams (operations, security, …)

Default security groups are applied to all instances by default

```
$ openstack security group list
+--------------------------------------+-------------------+-------------------------+---------+------+
| ID                                   | Name              | Description             | Project | Tags |
+--------------------------------------+-------------------+-------------------------+---------+------+
| e4b558aa-9463-48ef-83a1-b516da0fad8b | default           | Default security group | None    | []   |
```

```
$ openstack security group rule list default --long
+--------------------------------------+-------------+-----------+-----------+------------+-----------+--------------------------------------+
| ID                                   | IP Protocol | Ethertype | IP Range  | Port Range | Direction | Remote Security Group                |
+--------------------------------------+-------------+-----------+-----------+------------+-----------+--------------------------------------+
| 1057e4cd-de8b-4a80-aeac-6b3a24b7f437 | any         | IPv4      | 0.0.0.0/0 | 0:65535    | ingress   | e4b558aa-9463-48ef-83a1-b516da0fad8b |
| b6daae5d-19e8-4460-a2b4-fd89ec3f1b25 | any         | IPv6      | ::/0      | 0:65535    | ingress   | e4b558aa-9463-48ef-83a1-b516da0fad8b |
| 645b6afe-fbea-4249-bebc-d51fc8d9eacc | any         | IPv4      | 0.0.0.0/0 | 0:65535    | egress    | None                                 |
| 4c769b9e-c967-4543-8fcc-0a404ca85f2b | any         | IPv6      | ::/0      | 0:65535    | egress    | None                                 |
```

Security Groups

Can also be managed by project owners

Defaults usually allow egress only, individual rules can add / remove capabilities

```
$ openstack security group list
+--------------------------------------+-----------------+-------------------------+---------+------+
| ID                                   | Name            | Description             | Project | Tags |
+--------------------------------------+-----------------+-------------------------+---------+------+
| e4b558aa-9463-48ef-83a1-b516da0fad8b | default         | Default security group | None    | []   |
| f090ccb9-b680-45d3-926a-b316c0d8857d | custom-sec-group | custom-sec-group       | None    | []   |
+--------------------------------------+-----------------+-------------------------+---------+------+
```

```
$ openstack security group rule create --ingress --remote-ip 0.0.0.0/0 --protocol icmp custom-sec-group
+--------------------------------------+-------------+-----------+-----------+------------+-----------+-----------------------+
| ID                                   | IP Protocol | Ethertype | IP Range  | Port Range | Direction | Remote Security Group |
+--------------------------------------+-------------+-----------+-----------+------------+-----------+-----------------------+
| 5753d38f-0d0a-4704-b4a4-8dc25de82bbd | any         | IPv4      | 0.0.0.0/0 | 0:65535    | egress    | None                  |
| 08819a02-a430-4bcb-a3ce-5b12999964e8 | any         | IPv6      | ::/0      | 0:65535    | egress    | None                  |
| 41d38dc9-31e0-473d-b3fd-18bc33655dc7 | icmp        | IPv4      | 0.0.0.0/0 |            | ingress   | None                  |
+--------------------------------------+-------------+-----------+-----------+------------+-----------+-----------------------+
```

Main Open Issues and Next Steps

Learning the ropes for debugging, operations… higher complexity, multiple teams

Open Issues

- vRouter builds must follow kernel versions - build on demand procedure
- Limited scalability with IP Fabric Forwarding

Next Steps

- Integration with SDN gateway (advertise to physical routers)
- Propagate audit information to security team
- Integration with Kubernetes clusters (CNI driver)

Questions?