Wiring the Future: Open Virtual Networking and Beyond

HEPiX Spring 2024

Daniel Failing, CERN

Outline

- CERN Cloud Overview
- Context
- Requirements
- New features with OVN
- OVN short overview
- Beyond / Plans

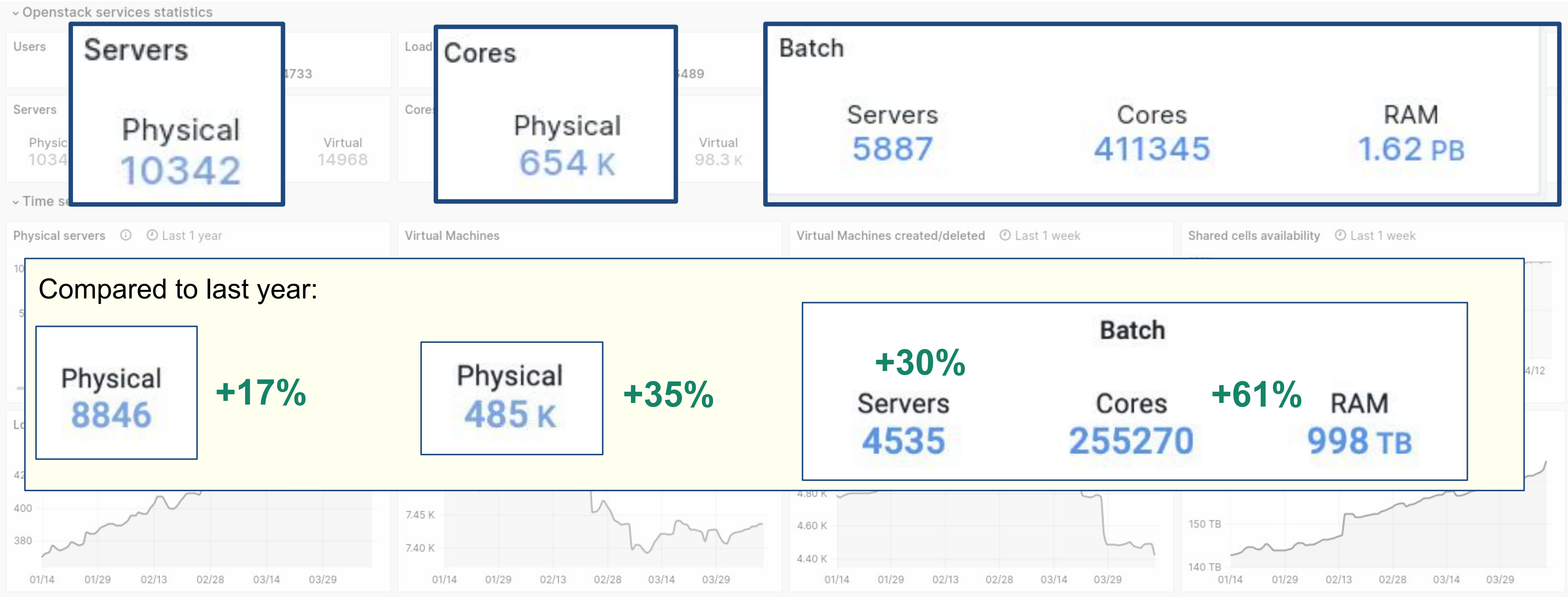
CERN Cloud Infrastructure

- Infrastructure as a Service
  - Production since July 2013
- RHEL/ALMA9
  - Based on RDO
  - x86_64 and Aarch64 architecture
- Meyrin Data Centre (MDC) + new Prevessin DC (PDC)
- Currently running Yoga* release
- Providing VMs/Bare-metal for users in the datacenter
  - Services, Personal VMs, Experiments, Batch

MDC

PDC

CERN Cloud Overview



Openstack services statistics

Servers

Physical
10342

Virtual
14968

Cores

Physical
654 K

Virtual
98.3 K

Batch

Servers
5887

Cores
411345

RAM
1.62 PB

Physical servers · Last 1 year

Virtual Machines

Virtual Machines created/deleted · Last 1 week

Shared cells availability · Last 1 week

Compared to last year:

Physical
8846

+17%

Physical
485 K

+35%

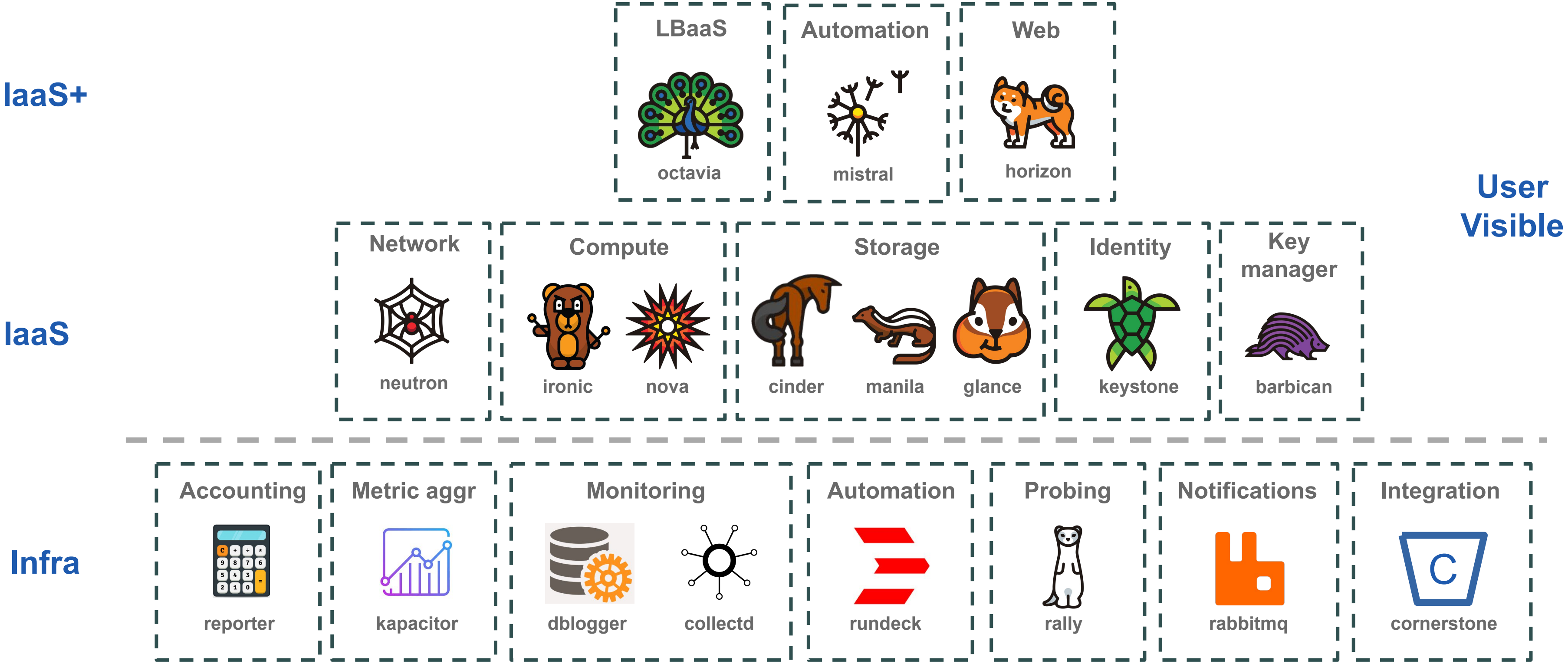
Batch

+30%
Servers
4535

Cores
255270

+61%
RAM
998 TB

Cloud Infrastructure APIs



IaaS+

LBaaS: octavia

Automation: mistral

Web: horizon

User Visible

IaaS

Network: neutron

Compute: ironic, nova

Storage: cinder, manila, glance

Identity: keystone

Key manager: barbican

Infra

Accounting: reporter

Metric aggr: kapacitor

Monitoring: dblogger, collectd

Automation: rundeck

Probing: rally

Notifications: rabbitmq

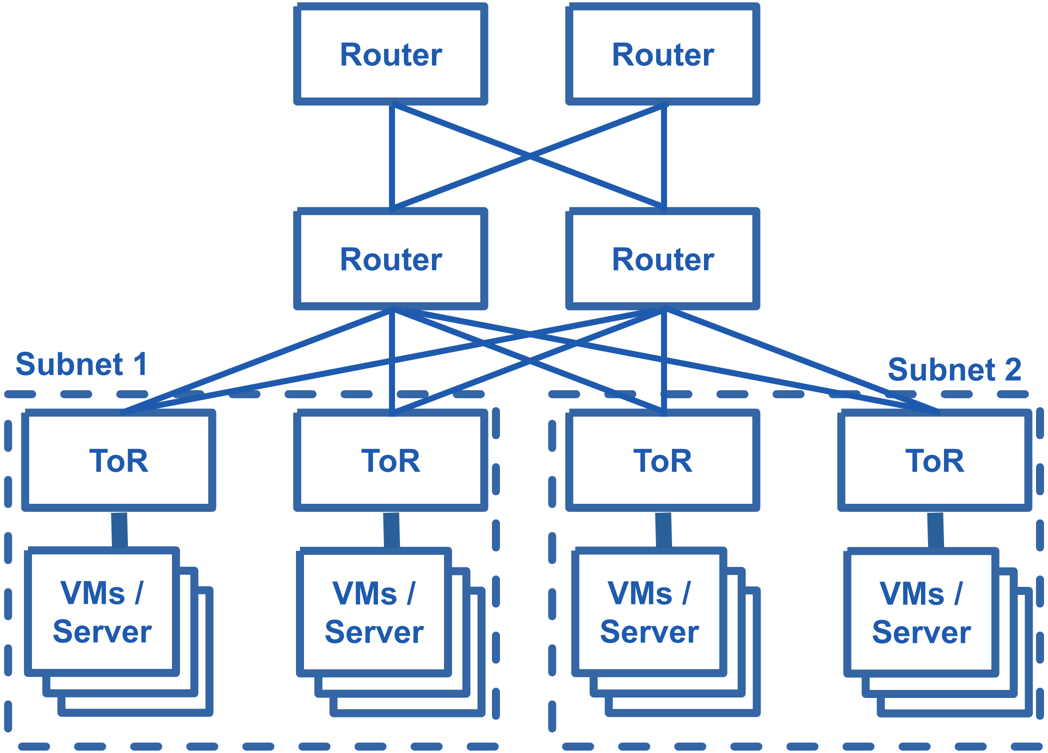
Integration: cornerstone

Context – PDC Deployment Cloud - Goals

- Focus is to **enable** users to deploy their BC/DR scenarios
- Base building blocks should **facilitate** and enable those
    - Fabric, Network, Storage, Compute and Database
- Keep in mind **existing** use-cases and **future** ones
- Use of a **green field deployment**
    - Review existing shortcomings
    - **Opportunity for new features**
    - No legacy constraints

Context – Current setup MDC

- VMs connect via LinuxBridge
- All VMs in the same public network
    - Full Dual-Stack IPv4 / IPv6
- Separated Subnets / Segmented
- Mantra: “Everything in same network”
    - except e.g. tech./control networks



Context – New requirements / Old limitations

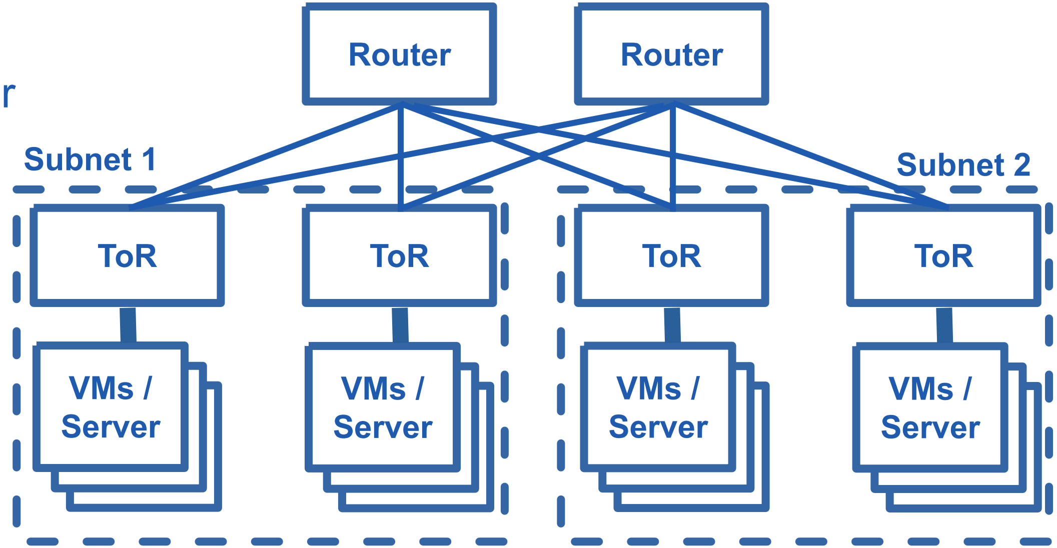
- Add ability to migrate VMs between hypervisors in different subnets
- New hypervisors can easily host over 100 VMs
  - some performance issues seen in current setup with higher VM count
- LinuxBridge implementation upstream marked experimental / unsupported
- Different teams ask for
  - Private Networks
  - Security Groups (Firewall rules on HV level)
  - Floatings IPs

PDC Network / DataCenter architecture

- For better scalability, network is divided into multiple Subnets with up to 1000 IPv4s (for now) + IPv6

- Top-of-Rack switch (ToR) only Layer 2
  - passing traffic outside subnet to Router

- Very similar architecture to other DC



Upstream Network Setups

- Network Service: Openstack Neutron
- Support for multiple vendors:
  - LinuxBridge
    - ➢ currently in use for other DC, marked experimental now
  - OpenVSwitch (OVS)
    - ➢ widely used
  - Open Virtual Network (OVN)
    - ➢ more flexible, widely used and recommended upstream
  - hardware vendor specific drivers
    - ➢ potential vendor lock-in, typically only for hardware switches/routers

Open Virtual Networking (OVN)

Reasons for OVN

- Teams asked for
  - ✓ Private Networks
  - ✓ Security Groups (Firewall rules)
  - ✓ Floatings IPs (under evaluation)
- Add Ability to migrate VMs between hypervisors in different subnets
  - ✓ Yes, across subnets with VMs in private networks for now only
- Aligning software stack with upstream OpenStack
  - ✓ OVN supported upstream
  - ✓ Migration path possible

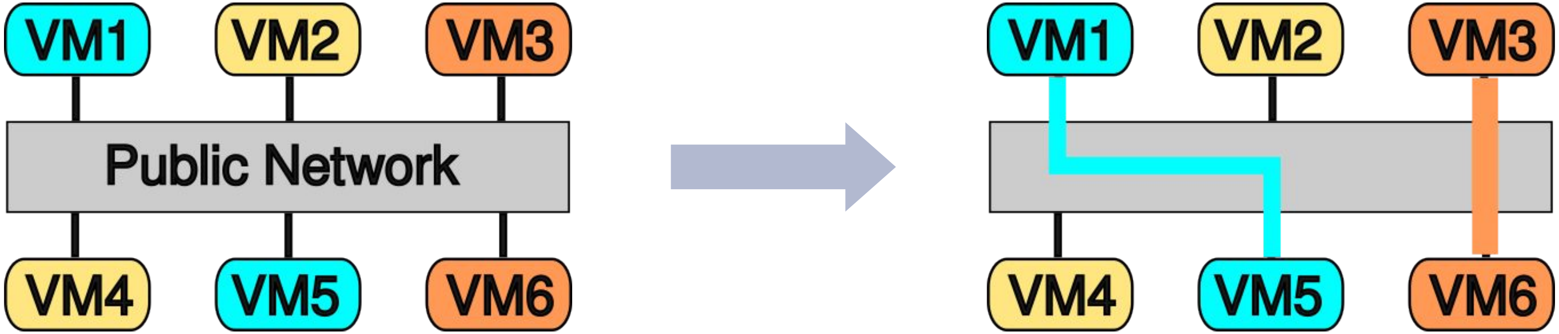
Different network types in PDC

- **Public/Provider** network (from start)
  - Keep existing functionality to simplify on-boarding
  - Reduce in-house patches by leveraging Neutron “Segments”
  - one subnet per 16 servers (approx 1000 IPv4 + IPv6)

- **Private** networks (Q2/2024)
  - Overlay network with OVN
  - Geneve tunnel between hypervisors

Private Networks

- Isolated tunneled network creatable by user
- (Virtual) Routers to connect to other private or public networks
- For now limited to VMs in the setup

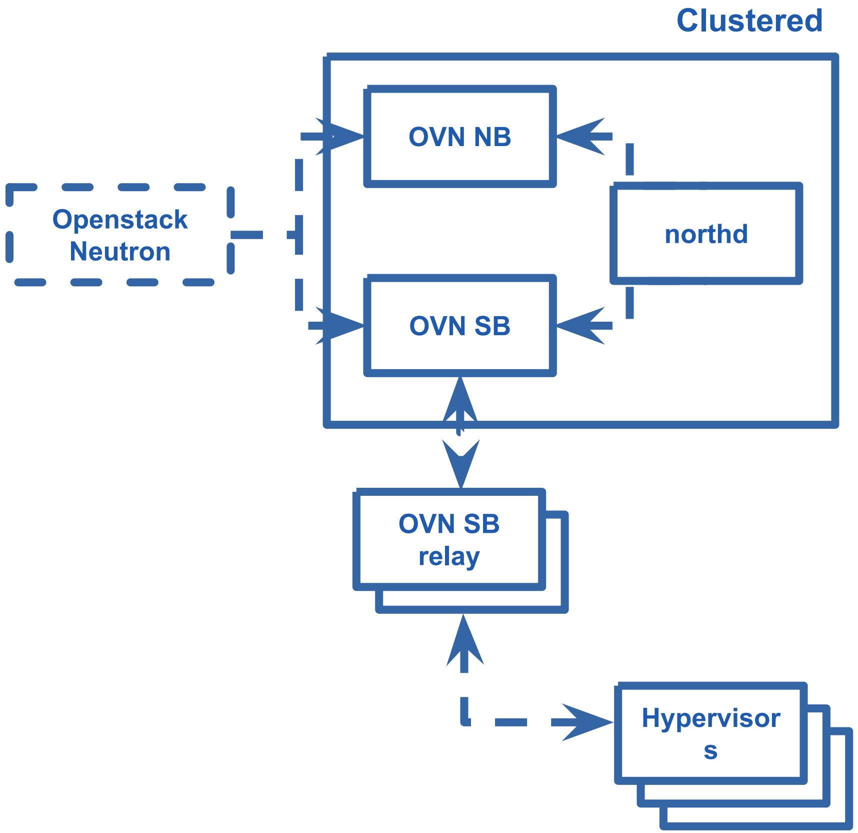


Security Groups

- Firewall for VMs on hypervisor level
- Allow certain groups of servers to talk to each other
- Whitelist approach
- Break of current mantra: “Everything can communicate with everything”
- Experience to be gained for large-scale deployment
  - Performance
  - User feedback

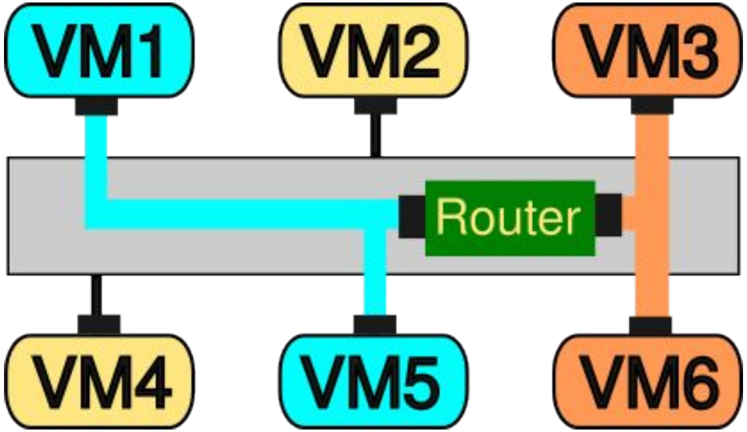
OVN Components

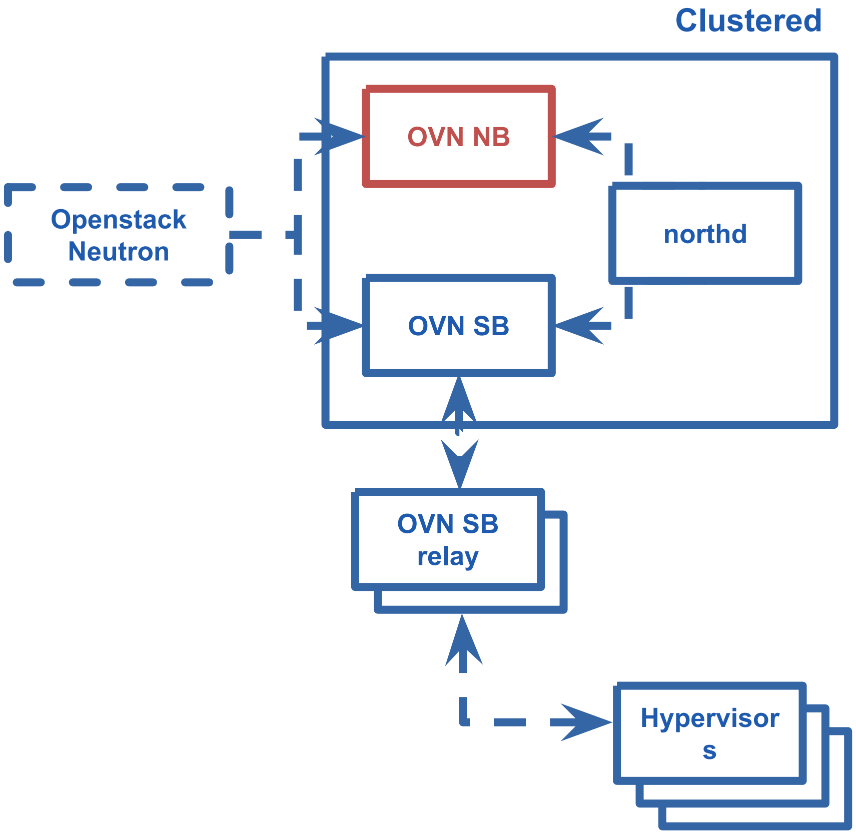
- OVN Northbound (NB) DB
  - ➢ “Port”, “Router”, “Switch”
- OVN Southbound (SB) DB
  - ➢ Hypervisor, Flow rules
- OVN northd
  - ➢ Translation between NB and SB
- OVN SB Relay
  - ➢ Relay/Cache for scaling



OVN Northbound

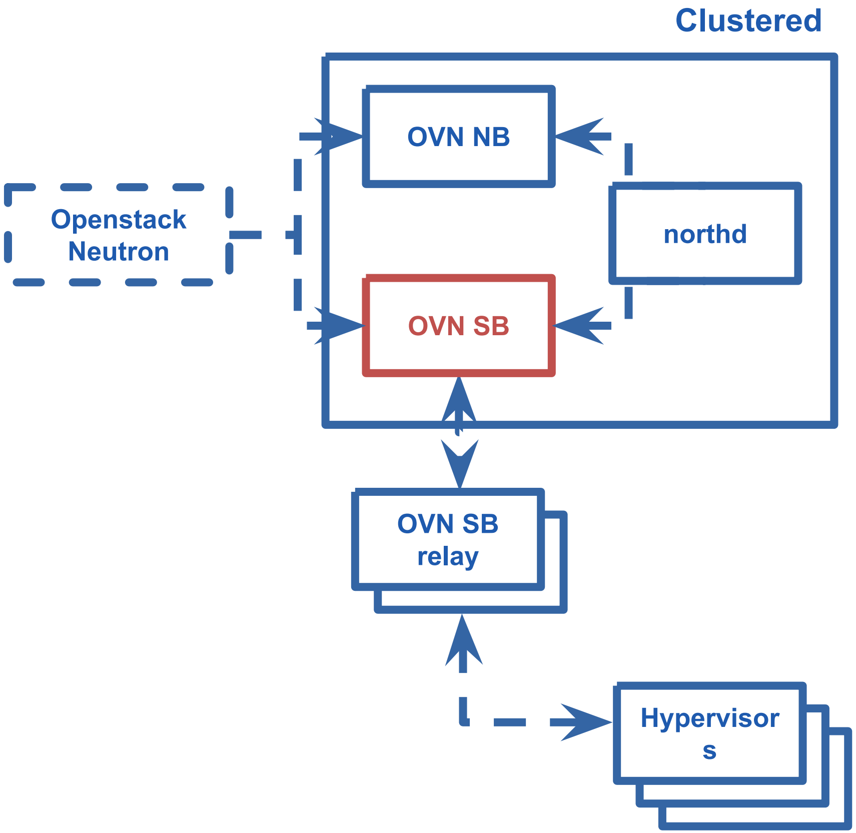
- Stores global abstract network view
- OVN NB: “Port”, “Router”, “Switch”





OVN Southbound

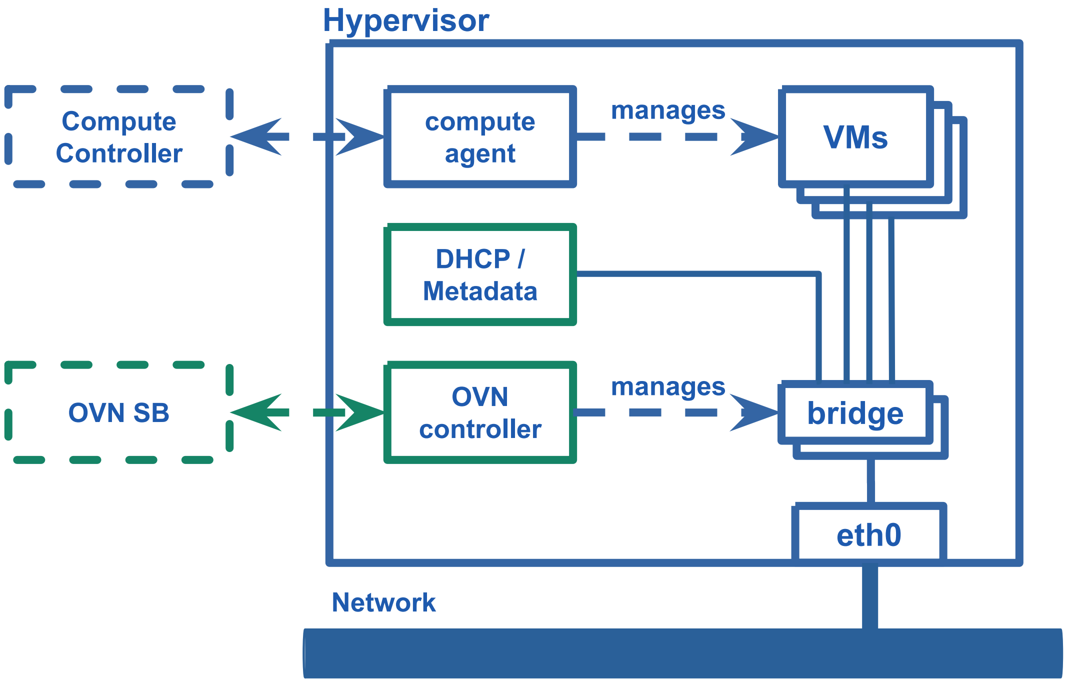
- OVN SB: Hypervisor, Flow rules
- Example FlowRule routed packet:
    - packet from port A
    - verify IP/MAC
    - check TTL
    - modify SRC IP
    - check firewall
    - send out port Z



Hypervisor

- OVN Controller + OpenVSwitch (OVS)
- Central OVN Network DB for configuration
    - Flow rules in local OVS bridge
- DHCP, Metadata for VMs in Hypervisor

- Public networks leaves eth0 directly
- Private networks tunneled to target host



Beyond / Plans

- Short term:
  - ➢ validating functionality with users
  - ➢ scalability test and gain confidence
- Migrate existing setup (old DC) to OVN (~15000 VMs, ~1700 hypervisors)
  - ➢ Migration path not straightforward
- Better integration with routers (e.g. BGP, EVPN)
  - ➢ Floating IPs
  - ➢ Even greater flexibility to move VMs

Thank You!

All our open source code is available on:
https://gitlab.cern.ch/cloud-infrastructure

Multiple contributors over the years helped to achieve this.

For contact:
Daniel Failing - daniel.failing <at> cern.ch