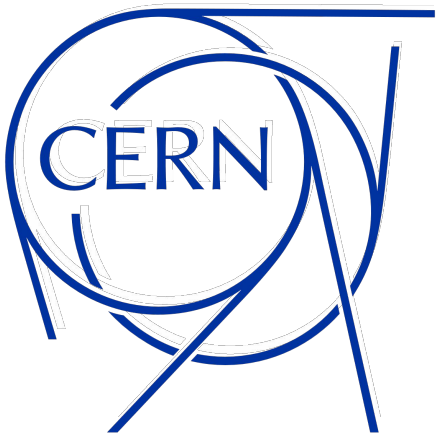


OVN Exchange - CERN/SWITCH

Daniel Failing

16-12-2024

CERN IT

- Support organization services, accelerator complex and experiments
- Currently spanning two datacenters
- 10 000+ servers
  - 5800 physics HTC/HPC
  - 1900 storage servers
    - Over 1 EB dedicated disk storage = 1 Mil TB; 1TB/s read
    - Over 1.5 EB tape storage
- Worldwide LHC Computing Grid (WLCG)
  - 170 sites, 42 countries
  - Network for >2.5Tb/s aggregated storage transfers

Meyrin Datacenter (MDC)

Prevessin Datacenter (PDC)

CERN Cloud

- Private Cloud in production since 2013
- 1800 Hypervisors
- 15000 VMs
- Based on OpenStack
- Offering includes
  - Compute: physical Servers, VMs, Container Orchestrations
  - Storage: Images, Volumes, File and Object Stores
  - Network: Load-Balancers, (Networks)
- Heterogeneous resources
  - X86, ARM, GPU

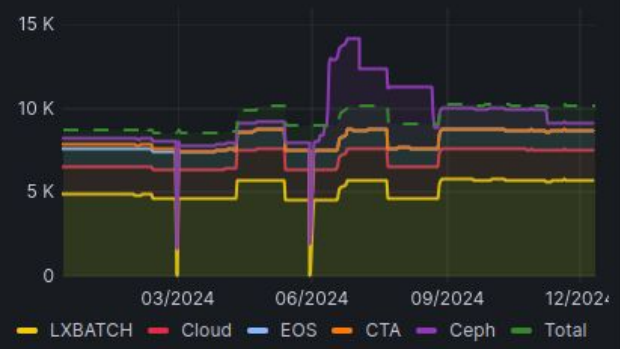
Openstack services statistics

Users	Projects	Loadbalancers	Images	Volumes	Volumes size	File Shares	File Shares siz	Object Store b	Object Store si
3526	4908	431	7159	7568	2.91 PB	5296	2.62 PB	108	7.96 TB

Servers				Cores			RAM			Batch		
Physical	Physical in use	Hypervisors	Virtual	Physical	Hypervisors	Virtual	Physical	Hypervisors	Virtual	Servers	Cores	RAM
10416	10167	1858	14991	346 K	70.2 K	103 K	1.69 PB	496 TB	249 TB	5889	413502	1.62 PB

Time series

Physical servers — Last 1 year



Virtual Machines

Virtual Machines created/deleted — Last 1 week

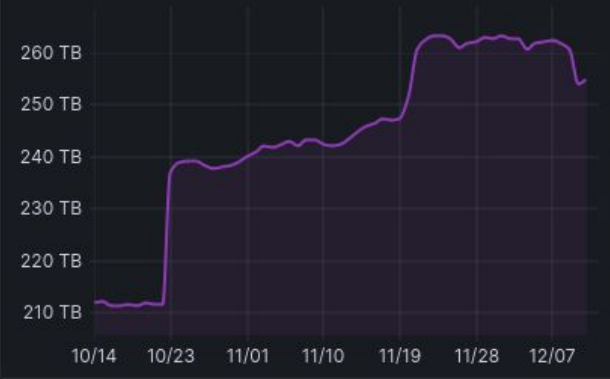
Shared cells availability — Last 1 week

Loadbalancers

Volumes

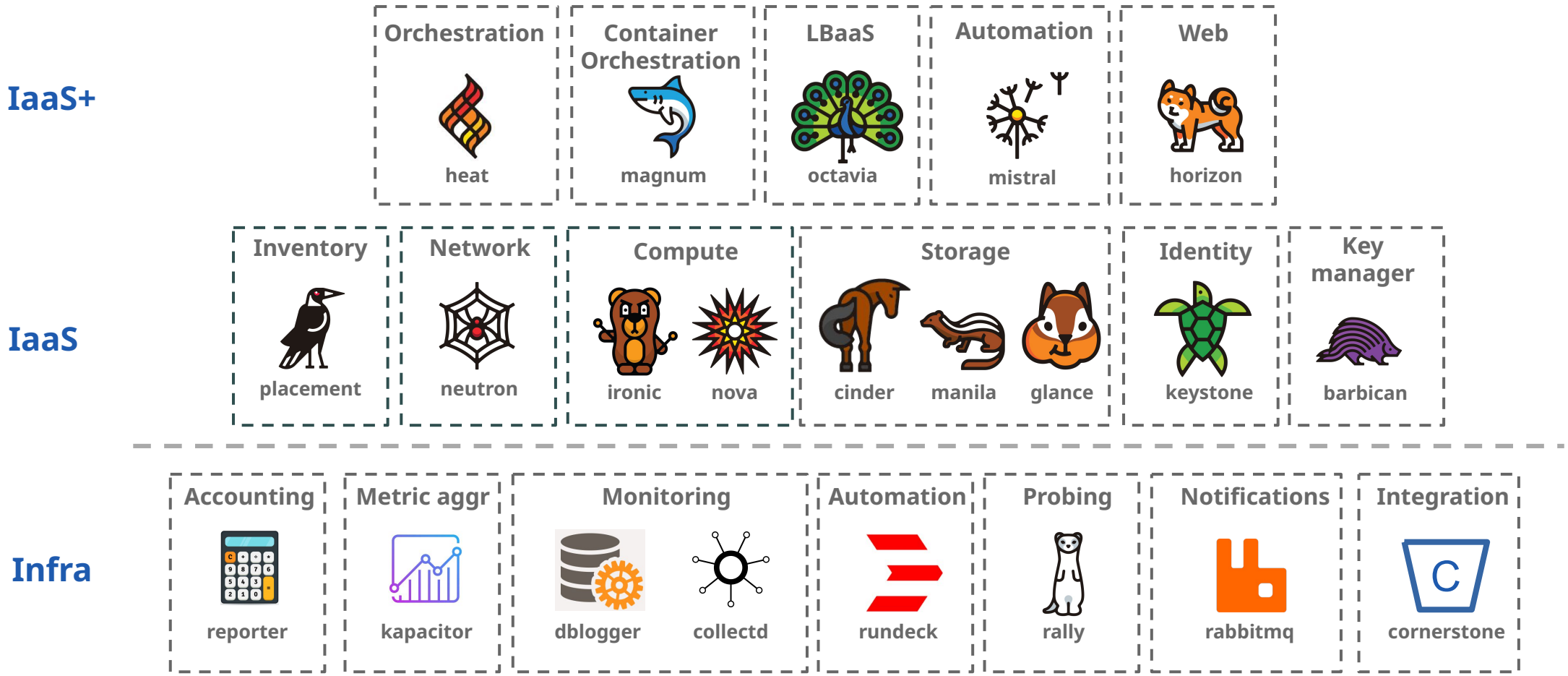
File Shares

Object Store size



https://monit-grafana-open.cern.ch/d/8f4TgzF7z/cern-openstack-overview

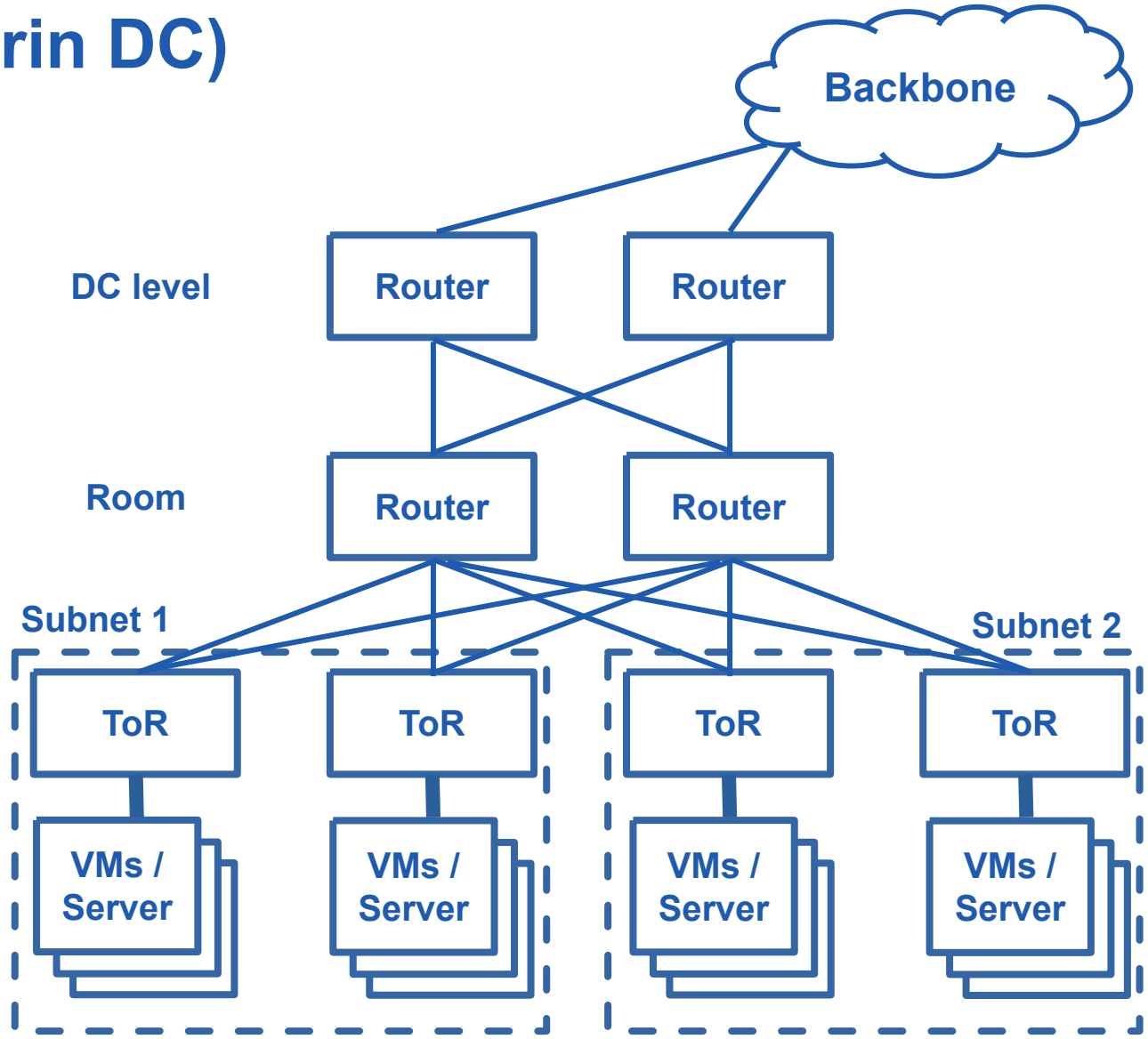
Cloud components



Cloud Networking

Context – Network Layer (Meyrin DC)

- Servers connected to
  - ToR or EoR (per density)
- Multiple routed L2 domains
- Spine/Leaf Routers
- BGP+some remaining OSPF in Spine
- Full Dual-Stack IPv4 / IPv6
- Mix of private and public IPs

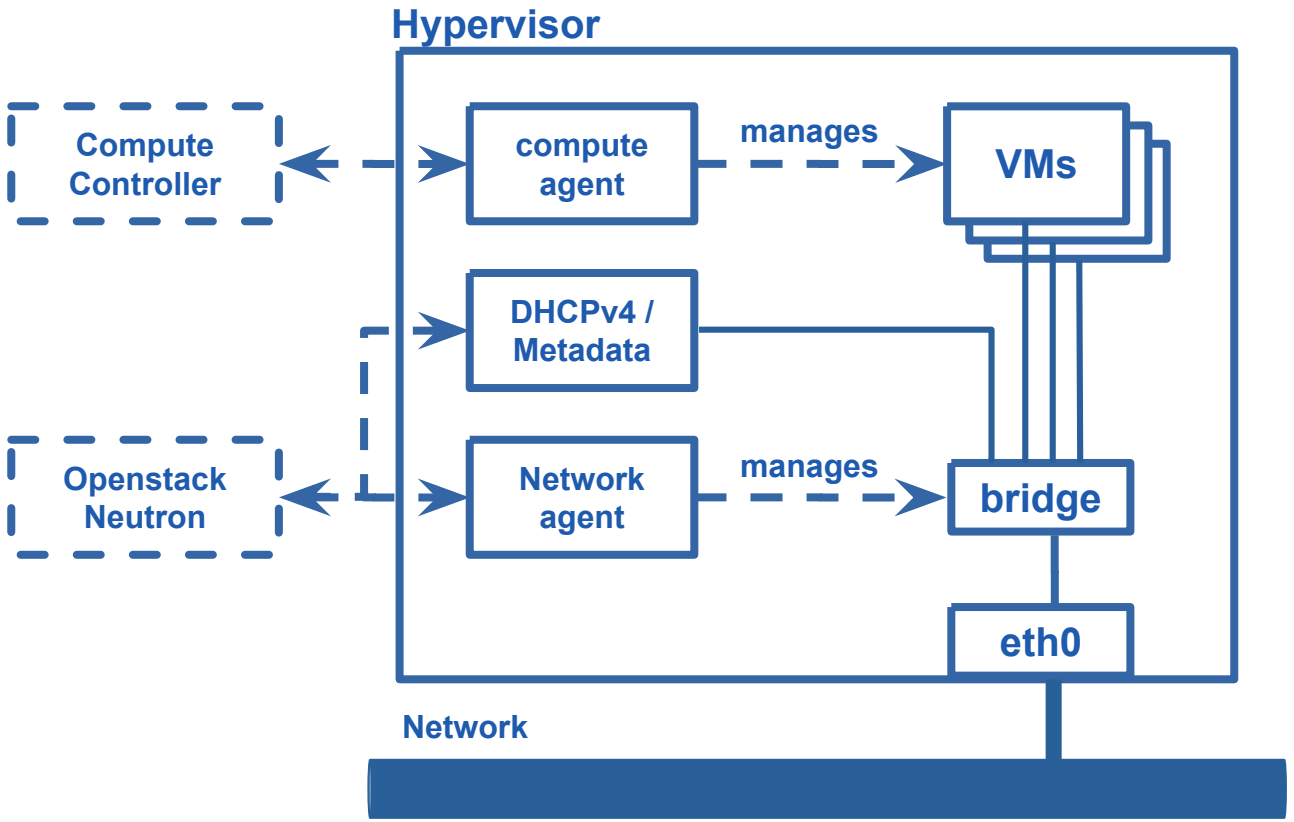


Cloud Network - Physical Servers for users

- One IPMI network port
- One operating system port, one IP
  - Hypervisors have additional subnet on the same port for VMs
- IPs stay (mostly) during the full time
- Network not managed by Cloud team, using site wide DHCP
  - +PXE for setup with OpenStack Ironic

Cloud Network - Current offering for VMs (Meyrin)

- VMs connect over LinuxBridge
- Separated Subnets / Segmented
  - hidden from user
- Mantra: “Everything in same network”
  - => no E/W isolation
- One Interface for everything per hypervisor
- User perspective:
  - 1 VM, 1 port, all in same public network



(New) Cloud Networking Model

- Current state: Meyrin DC uses LinuxBridge (kernel bridges)
- Prevessin DC uses Open Virtual Network (OVN)
- Move from custom implementation to upstream network segments

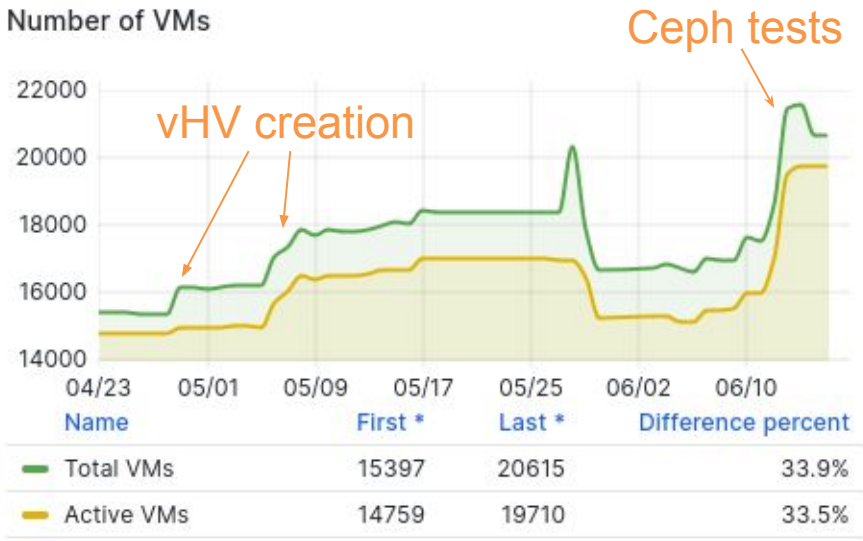
- Multiple reasons to renew network stack
  - Addition of Security Groups
  - Create foundation for more advanced features (e.g. private networks)
  - Deprecation of upstream support for LinuxBridge

- For now the new setup looks similar to users as the “old model” in Meyrin DC
  - One Exception: Security Groups

Network Scalability Tests

- Duplicate the cloud for testing
  - Create a cloud with 2 000 Hypervisors and 14 500 VMs
- Verify the architecture works for the MDC
  - How do we need to scale some of the new components?
  - How many operations can we run on the setup? (vm create/delete)

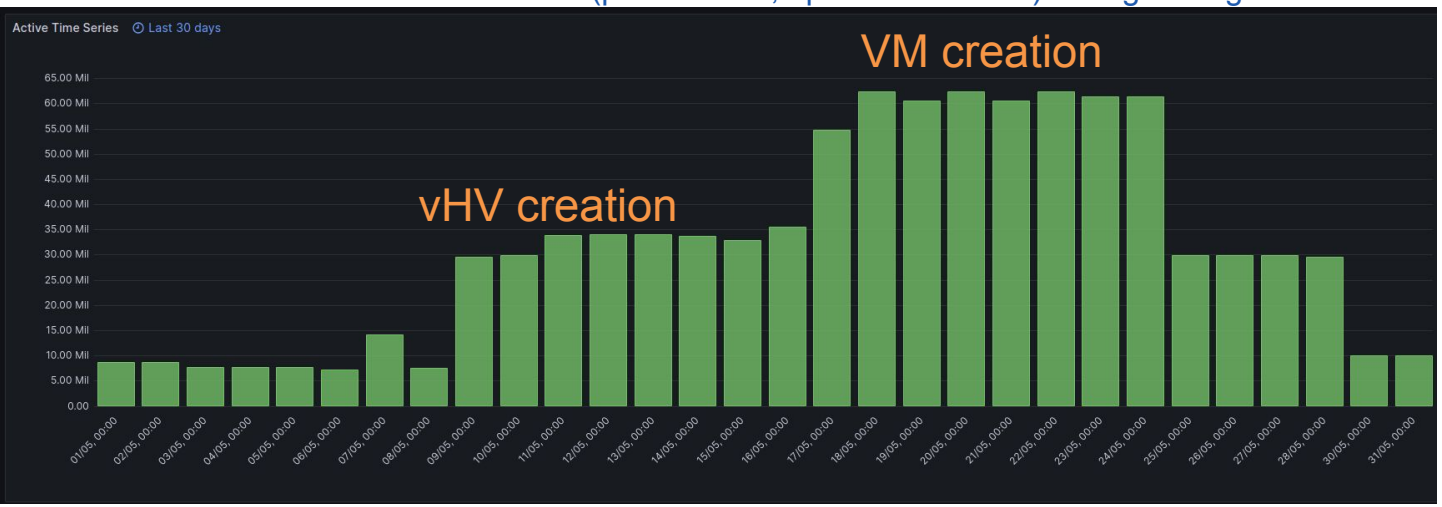
- Solution: Create virtual hypervisors (Hypervisors in VMs)
  - 2 000 VMs in pdc each 30 GB RAM
  - 14 500 VMs in above VMs, each 2 GB RAM
  - Potential to go up to 30 000 VMs in VMs



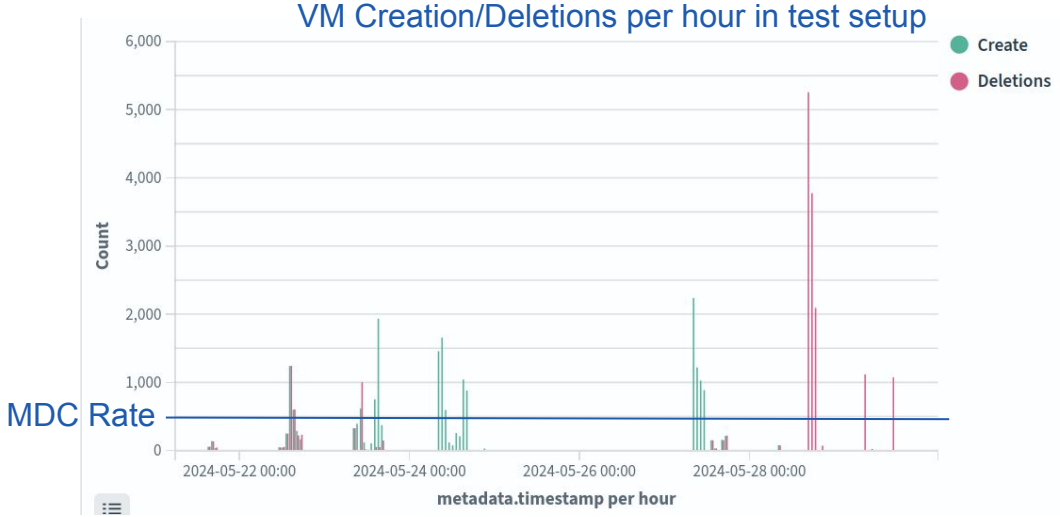
Network Scalability Tests - Results

- Scaled up to 2 000 hypervisors / 14 500 VMs
- Overall stable with smaller API scale then expected
- Minor improvements, mostly unrelated to architecture
- Pushing not only our limits

Active Time Series in monit (prometheus, openstack tenant) during testing



VM Creation/Deletions per hour in test setup



Future Plans

- Migrate the old DC (14 000 VMs, 1800 Hypervisors) to OVN
- Gain Experience
- Promote use of security groups

- Productionize private networks (and support structures, like routers, floating IP)
- BGP additions (towards Active/Active LBaaS)

Recent Fun

- “Transparent” intervention on one power feed resulted in throttling of CPU to 400MHz
- Rebooting triggered tunnels between hypervisors in different L2 domains
  - ➢ Routers polluted with IPv4 duplicates (blocking the IPs)
  - ➢ IPv6 route advertisements forwarded to all VMs and physical machines in L2 domain
    - ■ But we had NDP replies from OVN for those VMs outside of same L2

- Live migration seems different than on LinuxBridge
  - ○ Traffic is tunneled to the destination hypervisor
  - ○ Not everything is smooth (some more glitches in our control plane during migration)

- OVN can intercept and answer DNS queries

Questions?