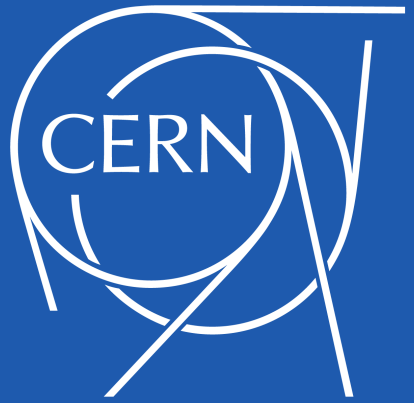

CERN

How to build a Better, Bigger, Stronger Cloud

José Castro León
CERN Cloud Infrastructure

About me

- Technical leader of the CERN Cloud Service
- Joined CERN in 2010 to work on virtualization
- Core Team that built the CERN private Cloud in 2012
- In the OpenInfra community since 2012

Outline

- CERN private Cloud
- Better
  - Cloud Service overhaul
- Bigger
  - Prevessin Data Centre
- Stronger
  - Heterogeneous architectures
  - Software-defined Networking

CERN Cloud Infrastructure

- Infrastructure as a Service
- Production since **July 2013**
- Running on **Redhat Enterprise Linux / AlmaLinux 9**
  - Based on Redhat Distribution of OpenStack (RDO)
- Meyrin and Prevessin Data Centres
- Currently running **Yoga+** release
  - Some services already in Zed release

Openstack services statistics

Users	Projects	Loadbalancers	Images	Volumes	Volumes size	File Shares	File Shares siz	Object Store b	Object Store si
3450	4768	430	6569	7480	4.60 PB	4486	2.27 PB	601	175 TB

Servers

Physical	Physical in use	Hypervisors	Virtual
9145	8946	1860	17863

Cores

Physical	Hypervisors	Virtual
585 K	482 K	116 K

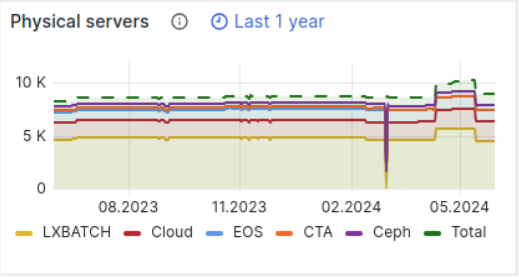
RAM

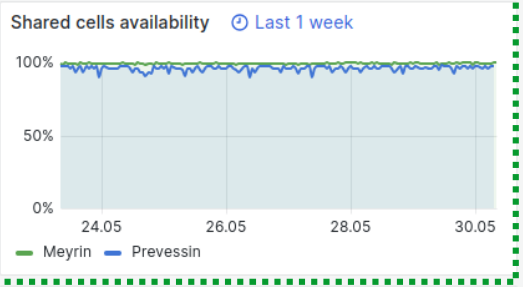
Physical	Hypervisors	Virtual
2.71 PB	492 TB	286 TB

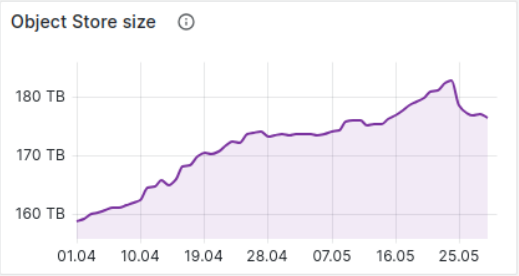
Batch

Servers	Cores	RAM
4684	333259	1.38 PB

Time series

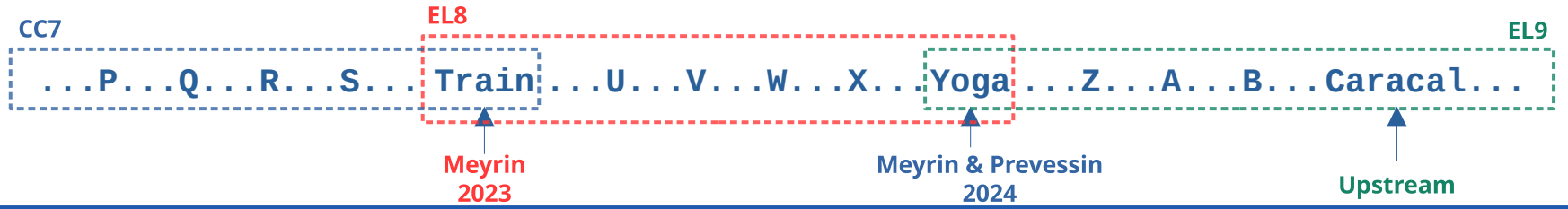






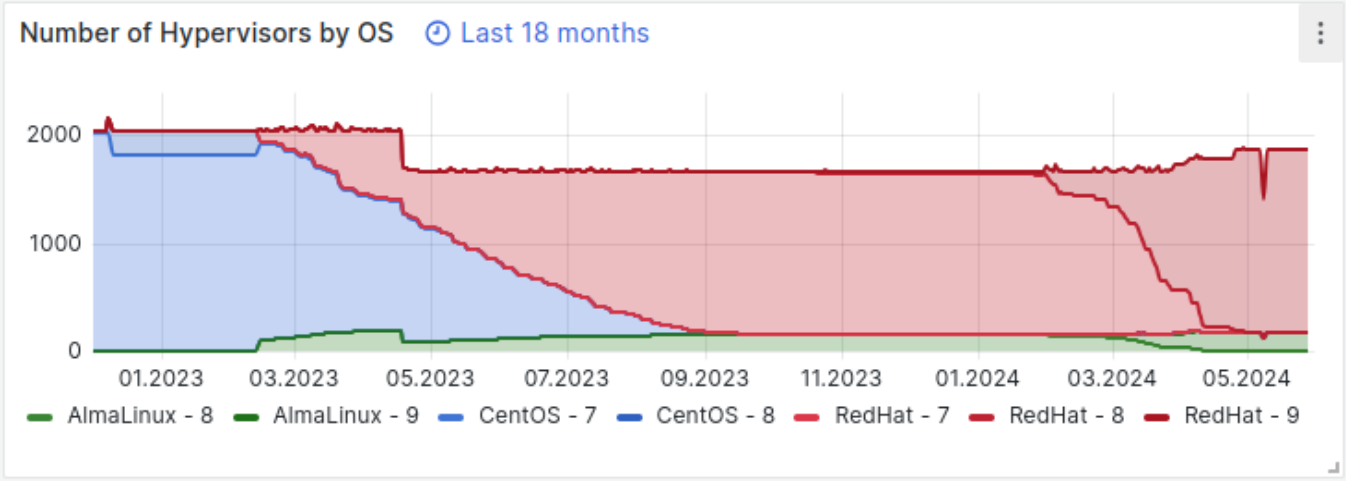
CERN Cloud Service overhaul

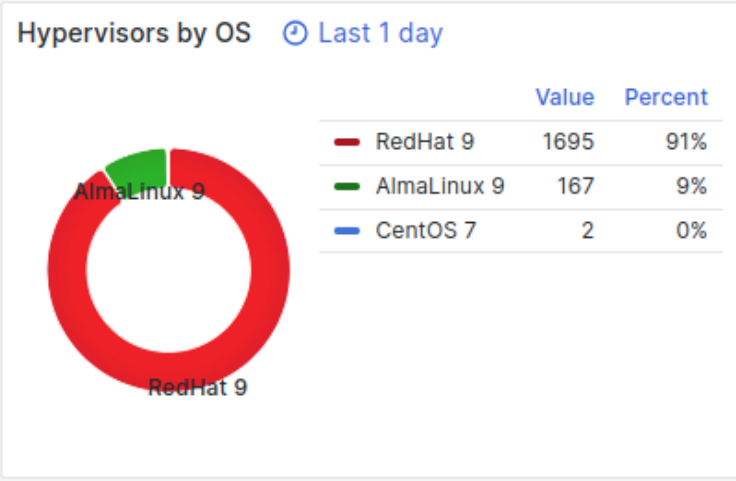
- Last 2 years, huge team effort to **modernize** it:
  - Hypervisor upgrade campaigns (from el7 to el9)
  - LBaaS migration from Tungsten to Octavia
  - Addition of heterogeneous architectures (ARM)
- Currently **working on**:
  - Replace network component in Meyrin Data Centre (MDC)
  - Close the gap with upstream OpenStack releases
  - Remove the boundaries with physical network topology



Hypervisor Upgrade to el8 and then to el9

- Live migrate ~15.000 Virtual machines (twice!)
  - from 6 months (el7 to el8) to 10 weeks (el8 to el9)
- Push the OS EoL support to 2029, and later to 2032 & unblock OpenStack upgrades

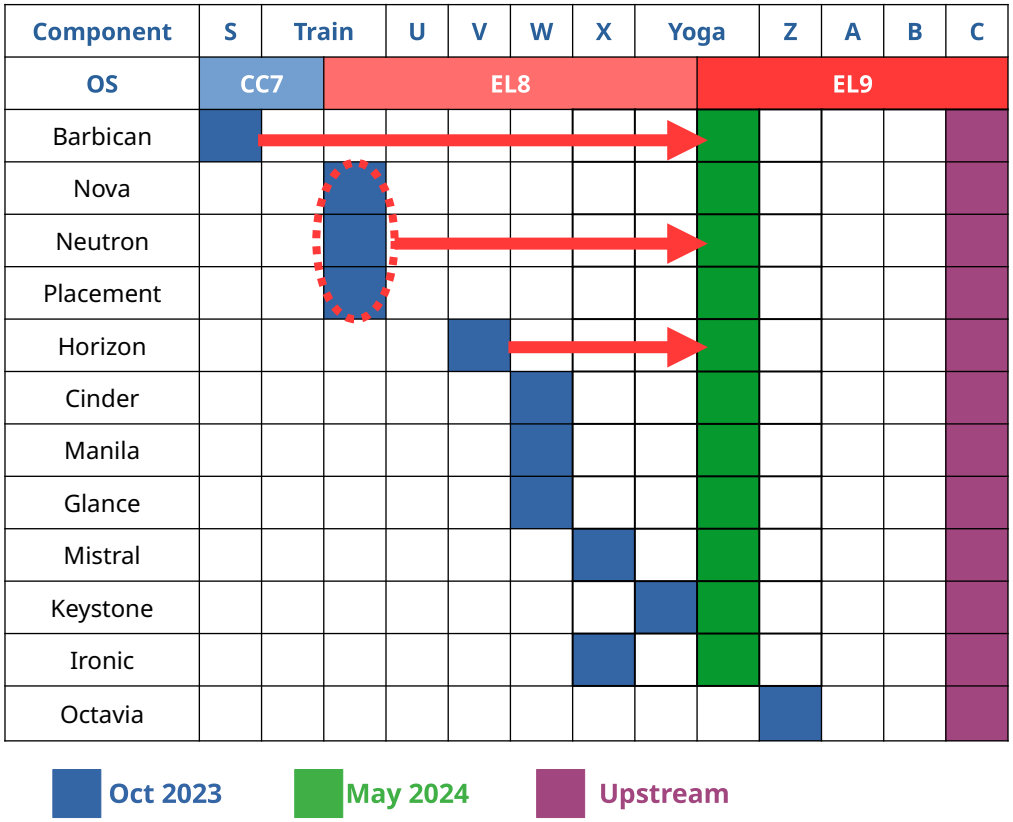




Close the gap with upstream

- Upgrade all components to Yoga in MDC
  - 1st ever great leap forward
    - Nova/Neutron/Placement
      - Long service interruption
  - Small steps for others

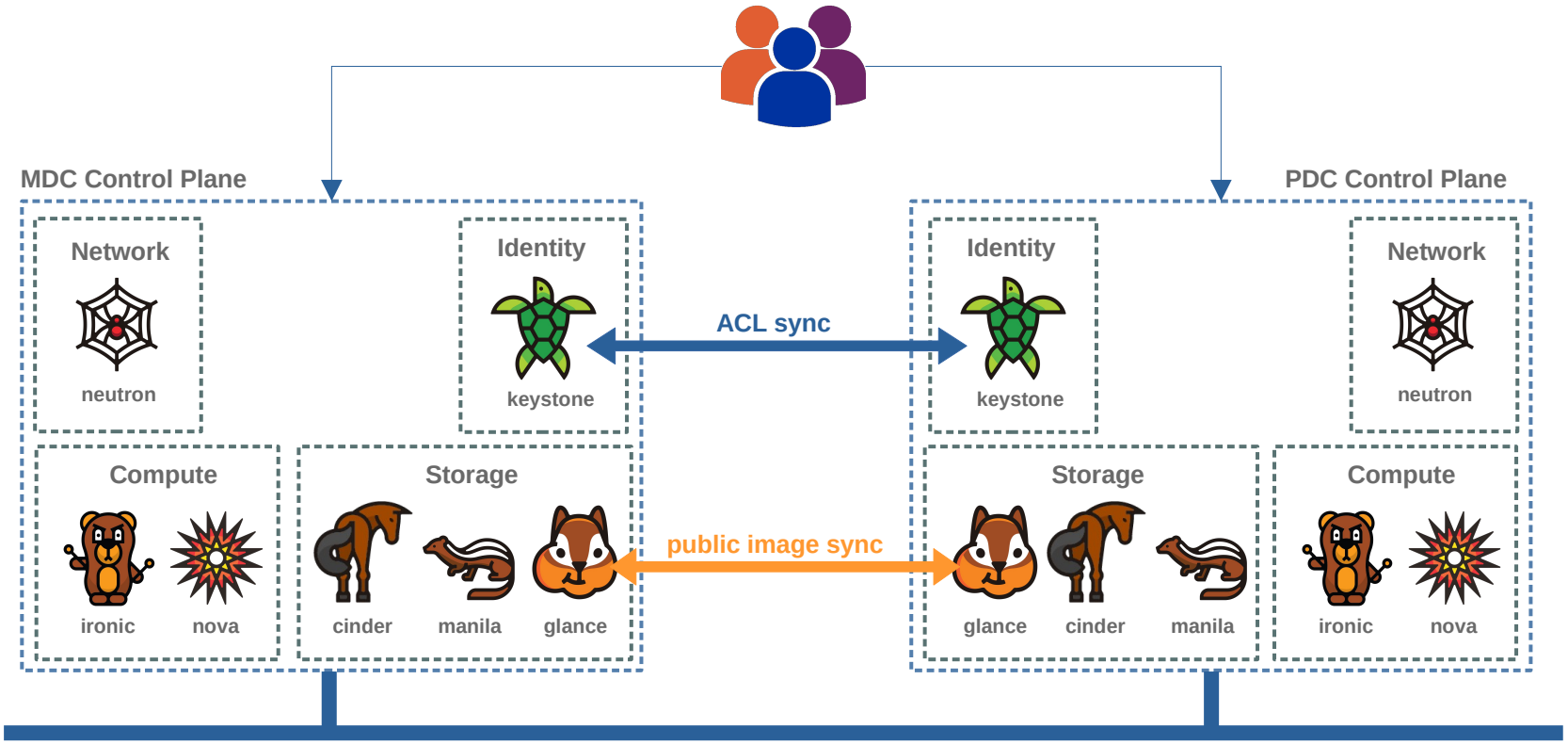
- **Now,** same version deployed in both sites
  - Final push to be closer to upstream



Adding a new datacentre in Prevessin

- Provides **additional capacity** for the upcoming LHC and HL-LHC runs
  - 3 floors with up to 4 MW per floor (**12MW**)
- Green field deployment
  - Review existing shortcomings
  - Opportunity for new features
  - No legacy constraints
- In **production** for Compute intensive workloads
- **Ready** for BC/DR scenarios on IT Services

Minimal interactions between sites



Differences between datacentres

Feature	Meyrin DC	Prevessin DC
OpenStack version	Yoga+	
OS version	RHEL9 / ALMA9	
Availability Zones	3 Compute 3 Storage	1 Compute & Storage
Number of Cells	34	**1**
Cross Zone attachments	YES	**NO**
Anti-/Affinity Filters	Host	**Host, Rack, Room**
Networks	Provider	**Provider & Private**
SDN Features	Load Balancers	**Security Groups** **Load Balancers** **Floating IPs**
Capacity (Memory on HV)	**375TB**	72TB
Capacity (on Diesel)	**12TB**	-
UPS expected lifetime	**15min**	5min

Software Defined Networking (via OVN)

- Addition of **Private Networks** into the portfolio
- **Enforce** Security Groups in the PDC setup
  - Break of current mantra: “Everything can talk with everything”
- Testing at scale to gain **experience** on performance and user feedback

- Future plans to offer **better integration** with routers (e.g. BGP, EVPN)
  - Floating IPs
  - Physical nodes

Manage scarce heterogeneous resources

- Many **distinct** use cases require access to GPUs with different **utilization**
  - deep learning, inference, analysis, simulations, GIS, mechanical, …
- **Multiple** Nvidia models available: T4, V100, V100s, A100, H100 soon …
  - Available as PCI-passthrough, vGPU and MIG
- ARM resources are **available** for users
  - Big initial up-take, additional resources included later on

- Really scarce resources, working on a **lease model**
  - Missing quota handling (unified limits)

Our current focus is on:

- Unified limits
- Cyborg (accelerators)
- Blazar (reservations)
- OVN integration with routers (BGP, EVPN)

Thank you

More info:

https://computing-blog.web.cern.ch/

All our **open source** code is available on:

https://gitlab.cern.ch/cloud-infrastructure

...this won't be possible without the contribution of all cloud team members