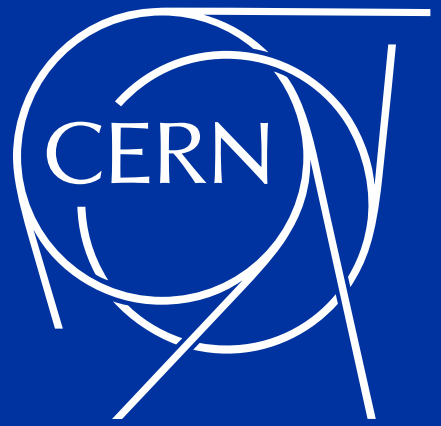


Implementing Disaster Recovery in the Hybrid Cloud: Challenges and Pitfalls

Aimilios Tsouvelekakis
17/11/2021

Agenda

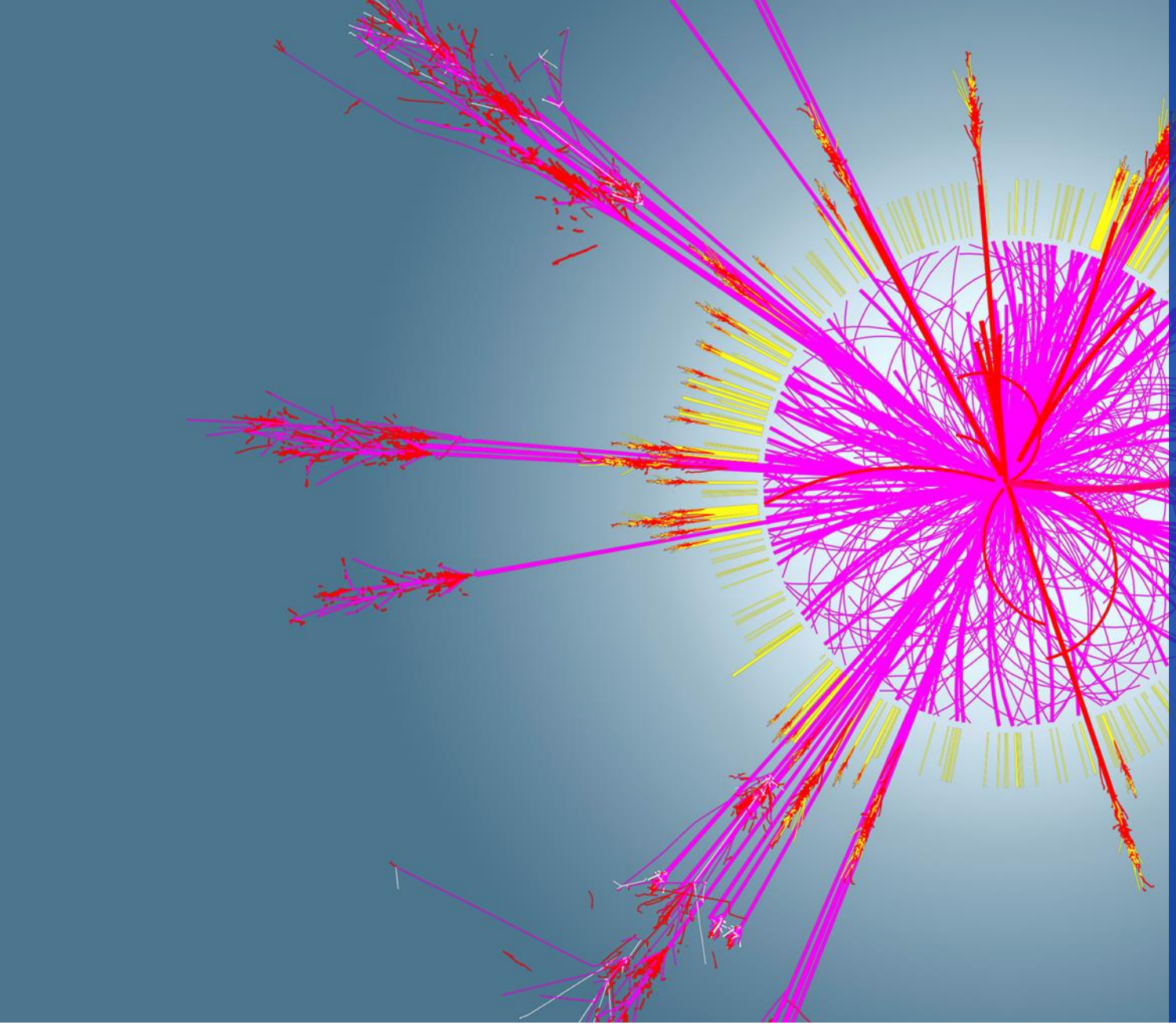
- About CERN
- Network
- Applications
- Databases
- Summary

About CERN

- Established in 1954
- 23 member states
- Primary mission
  - provide a unique range of particle accelerator facilities that enable research at the **forefront of human knowledge**
  - perform world-class research in fundamental physics
  - unite people from all over the world to push the frontiers of science and technology, for the benefit of all

People

- More than 2500 staff
- More than **17500 collaborators from around the world**
- Over 12200 scientists
  - 110 nationalities
  - institutes in more than 70 countries



CERN Pioneer

- **Where the Web was born**
  - https://www.youtube.com/watch?v=pJrAUGpFnPw
  - https://web30.web.cern.ch/
- **Touch screen**
  - https://www.youtube.com/watch?v=tQe5dlzScwU
  - https://cds.cern.ch/record/1248908?ln=en

Project Goals

- Investigate Public Cloud solutions
- Evaluate the Network options
- Application deployment in the Cloud
- Create Standby Databases in the Cloud

Hybrid Cloud
Why?

- Scalability
- Agility
- Combine best of both worlds

Network
High Level Overview

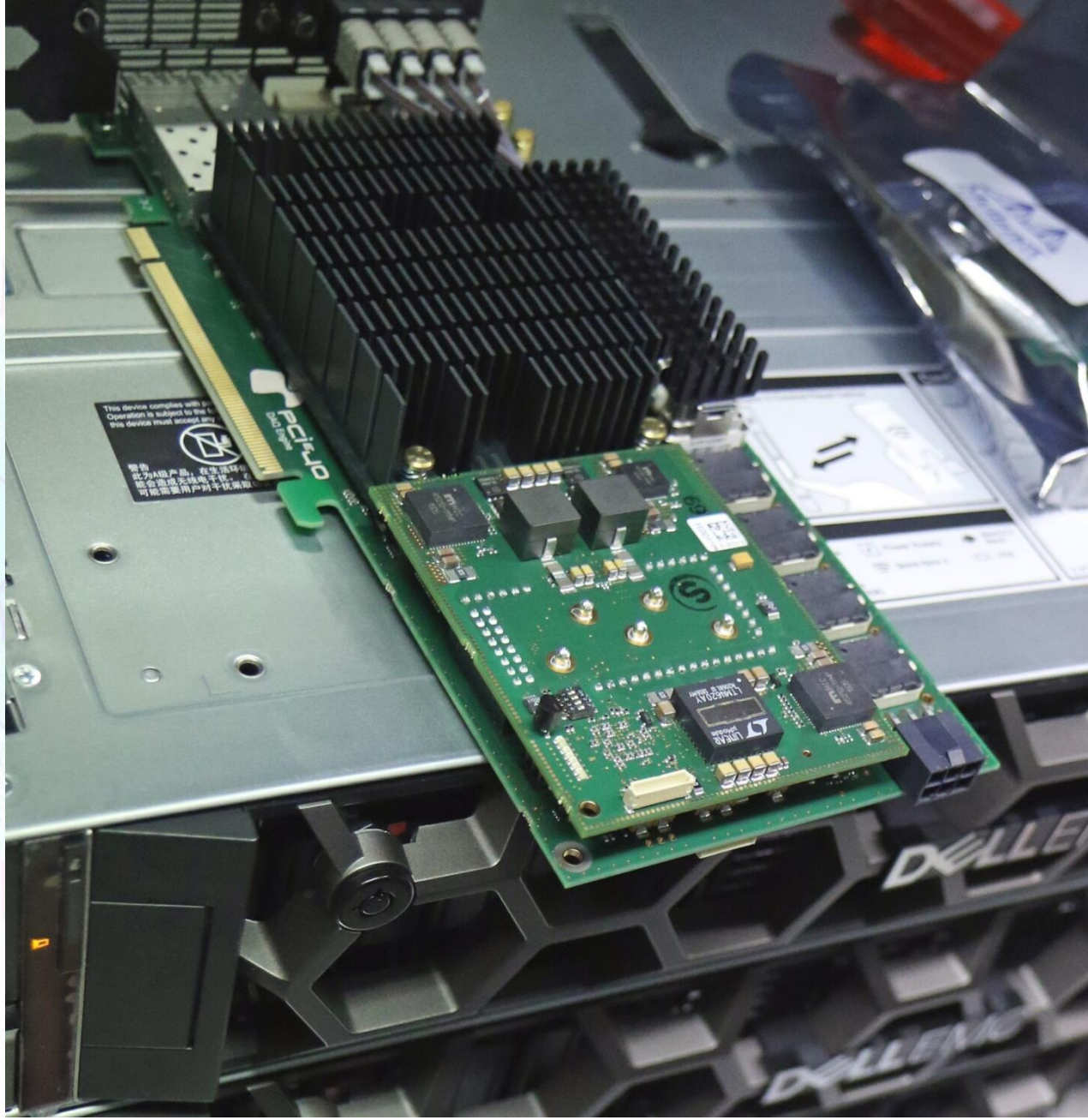
Connecting to OCI

- **VPN**
  - Bandwidth dependent on customer's access
  - IPsec authentication and encryption
  - Goes through the internet
- **FastConnect**
  - High bandwidth
  - Dedicated line bypasses the internet
  - Inbound and outbound traffic is free

CERN to OCI

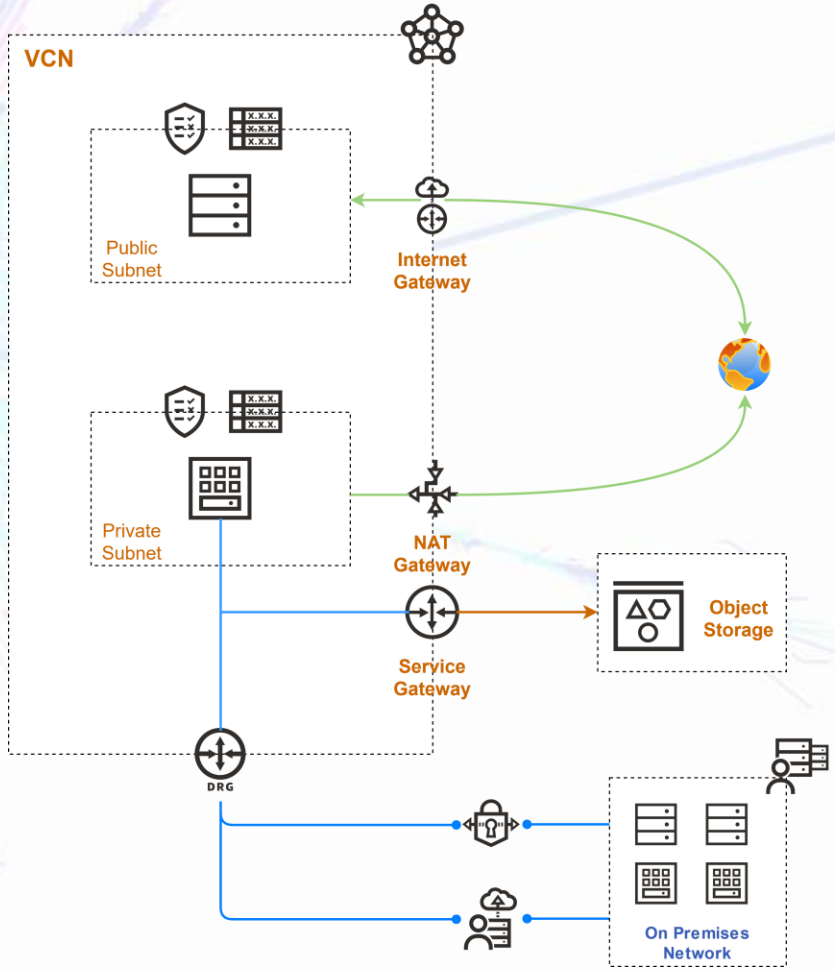
Dedicated connection to OCI Frankfurt

- **We use FastConnect**
  - 2 ports of 10Gbps
  - Private Peering
- **GÉANT is now a partner with Oracle**
- **Achievement: Academic/Research institutions can connect to Oracle Cloud through GÉANT**

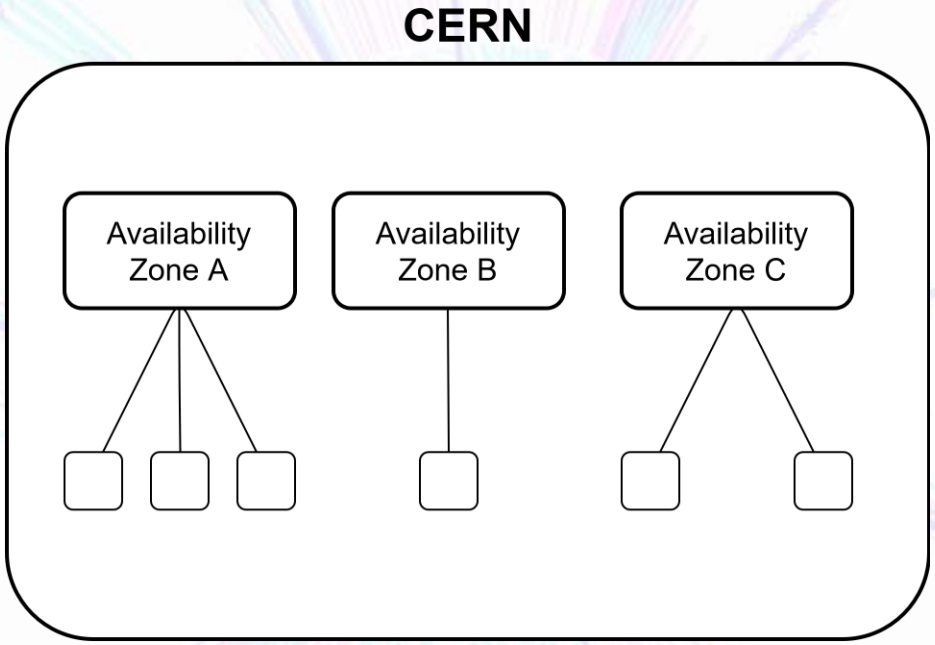


OCI Network Connectivity

- **On Premise to OCI**
  - FastConnect
  - IPSec VPN
- **Public Internet**
  - Internet Gateway
  - NAT Gateway
- **Services Inside OCI**
  - Service Gateway
- **Peering**
  - Remote Peering Gateway (DRG)
  - Local Peering Gateway (LPG)



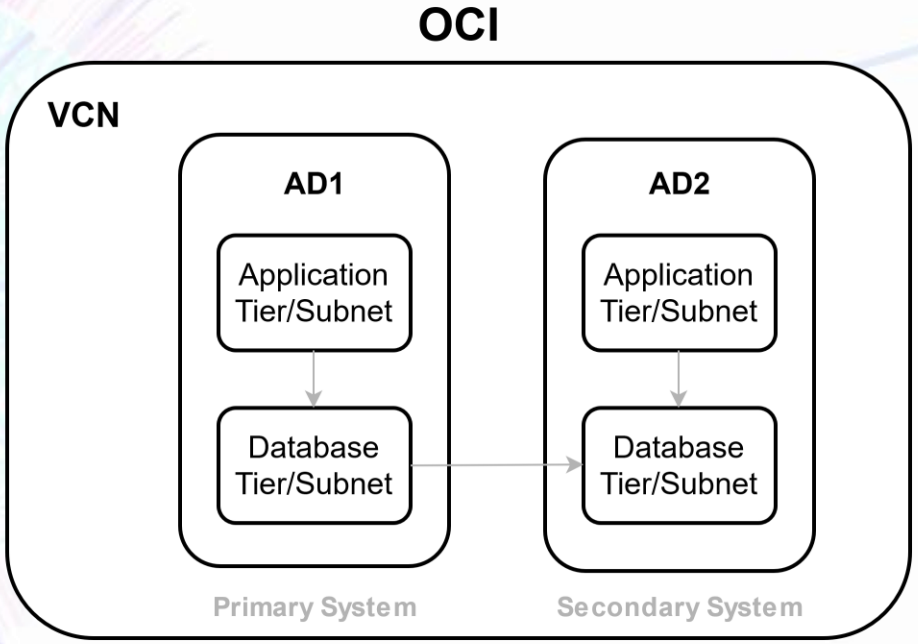
CERN Network



OCI Subnet Grouping

Function Dedicated Subnets

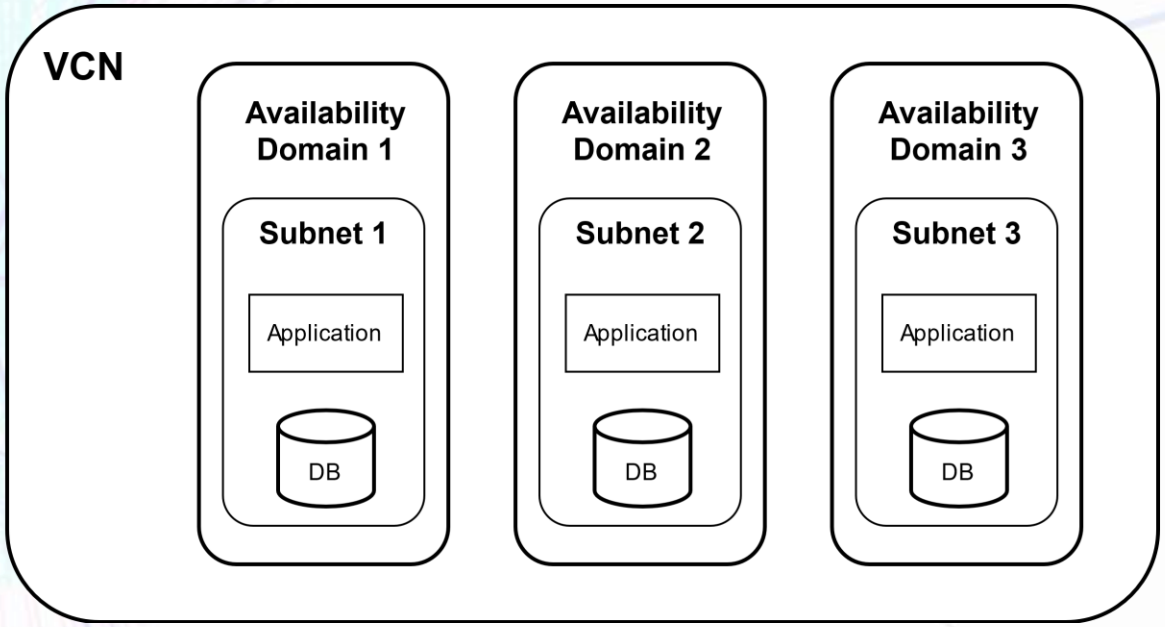
- Subnets can have strict access policies
- Subnets can have different visibility type
- Dedicating subnets for projects improves security
- Requires more time and effort to configure security lists & route tables
- Changing CIDR for existing subnet requires deletion and recreation
- Subnet Structure replication in different ADs



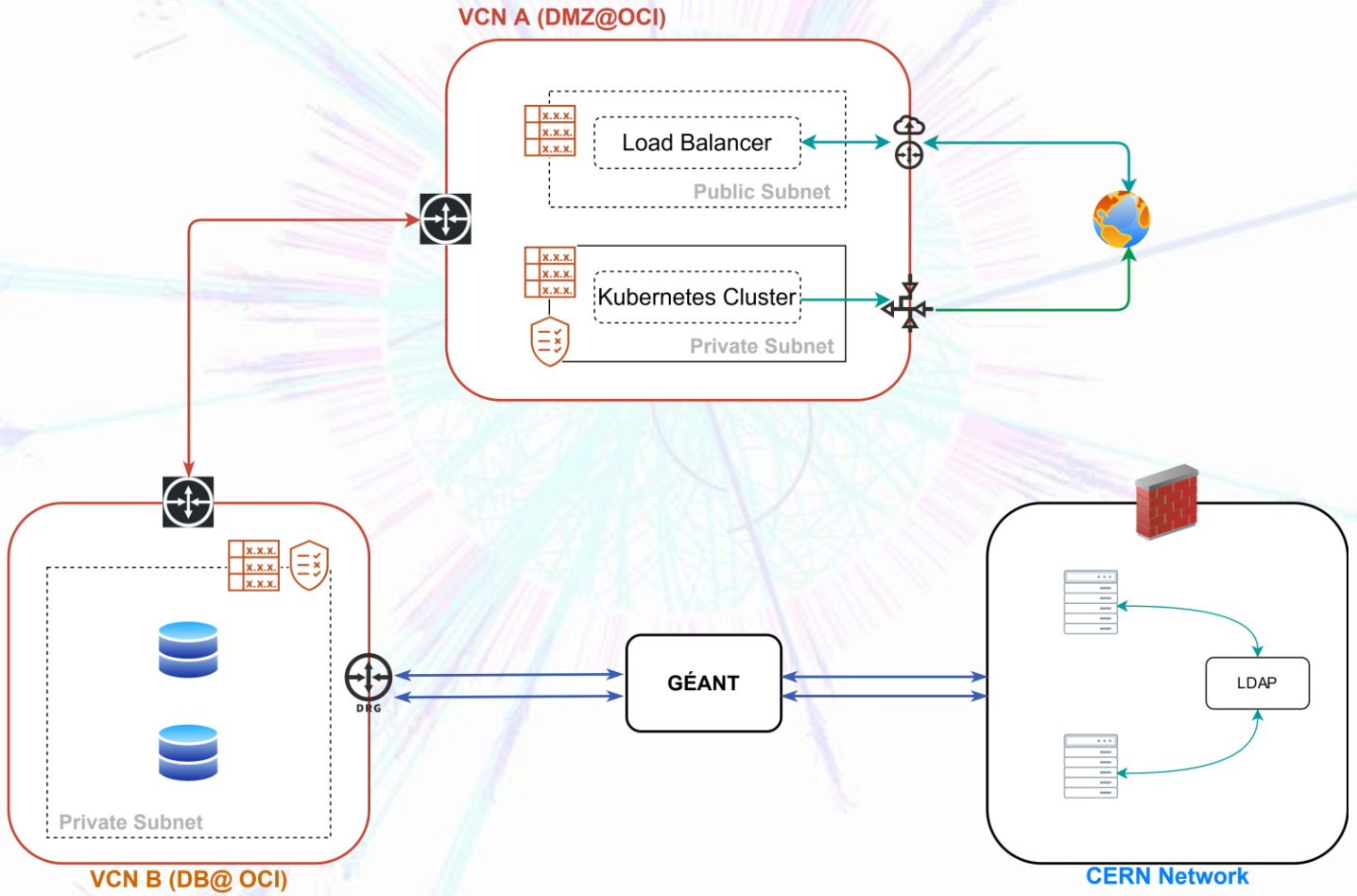
OCI Subnet Grouping

Availability Domain Dedicated Subnets

- Simpler topology
- Less maintenance for CIDR block management
- Possibility to use the same security list and routing table for the subnets
- For HA infrastructure, subnets needs to be replicated in another AD
- Good use case if you need one visibility type in all your ADs



Recommended Design

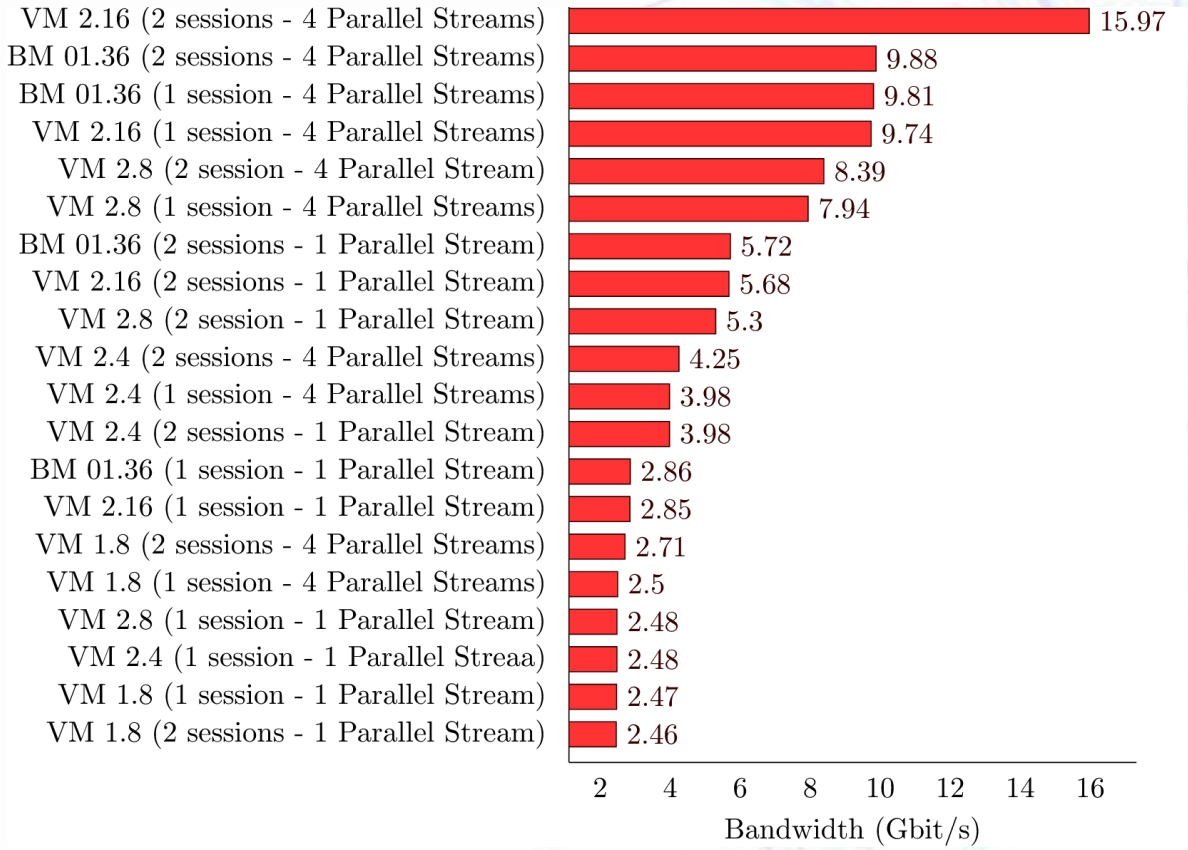


Network Benchmarking
Bandwidth and Latency

- CERN Bare Metal to OCI Virtual Machines and Bare Metal
- Bandwidth benchmarked with iperf
- 1 or Multiple Sessions using Parallel Streams
- Latency tested with mtr & ping
- Investigate performance towards Wigner datacentre

Network Benchmarking

Bandwidth and Latency




```
[opc@aimilios ~]$ mtr --no-dns --report --report-cycles 60 10.x.x.x
Start: Tue Jan  5 16:46:49 2021
HOST: aimilios                    Loss%   Snt   Last   Avg  Best  Wrst StDev
  1.|-- 140.x.x.x                  0.0%    60    0.1   0.1   0.1   0.2   0.0
  2.|-- 192.x.x.x                  0.0%    60    9.6  14.1   9.3  56.9  11.1
  3.|-- 192.x.x.x                  0.0%    60   12.7  11.0   9.3  23.7   2.9
  4.|-- 185.x.x.x                  0.0%    60    9.9  10.1   9.3  23.0   2.0
  5.|-- 10.x.x.x                   0.0%    60    8.7   8.8   8.7   9.5   0.0
```

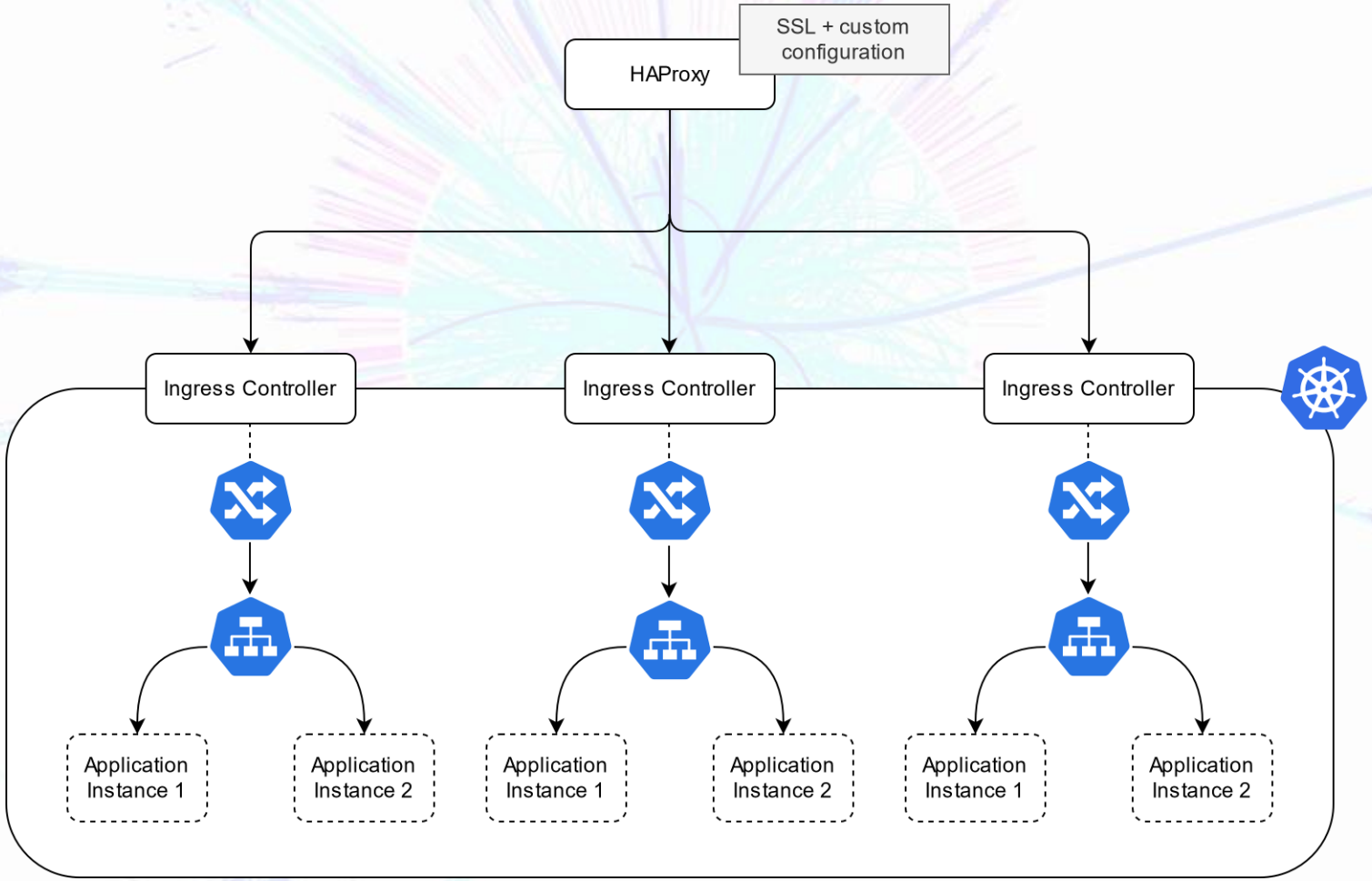
```
[opc@aimilios ~]$ ping 10.x.x.x
PING 10.x.x.x (10.x.x.x) 56(84) bytes of data.
64 bytes from 10.x.x.x: icmp_seq=1 ttl=60 time=8.82 ms
64 bytes from 10.x.x.x: icmp_seq=2 ttl=60 time=8.84 ms
64 bytes from 10.x.x.x: icmp_seq=3 ttl=60 time=8.79 ms
...
64 bytes from 10.x.x.x: icmp_seq=29 ttl=60 time=8.85 ms
64 bytes from 10.x.x.x: icmp_seq=30 ttl=60 time=8.83 ms

--- 10.x.x.x ping statistics ---
30 packets transmitted, 30 received, 0% packet loss, time 29049ms
rtt min/avg/max/mdev = 8.770/8.833/8.923/0.049 ms
```

Applications

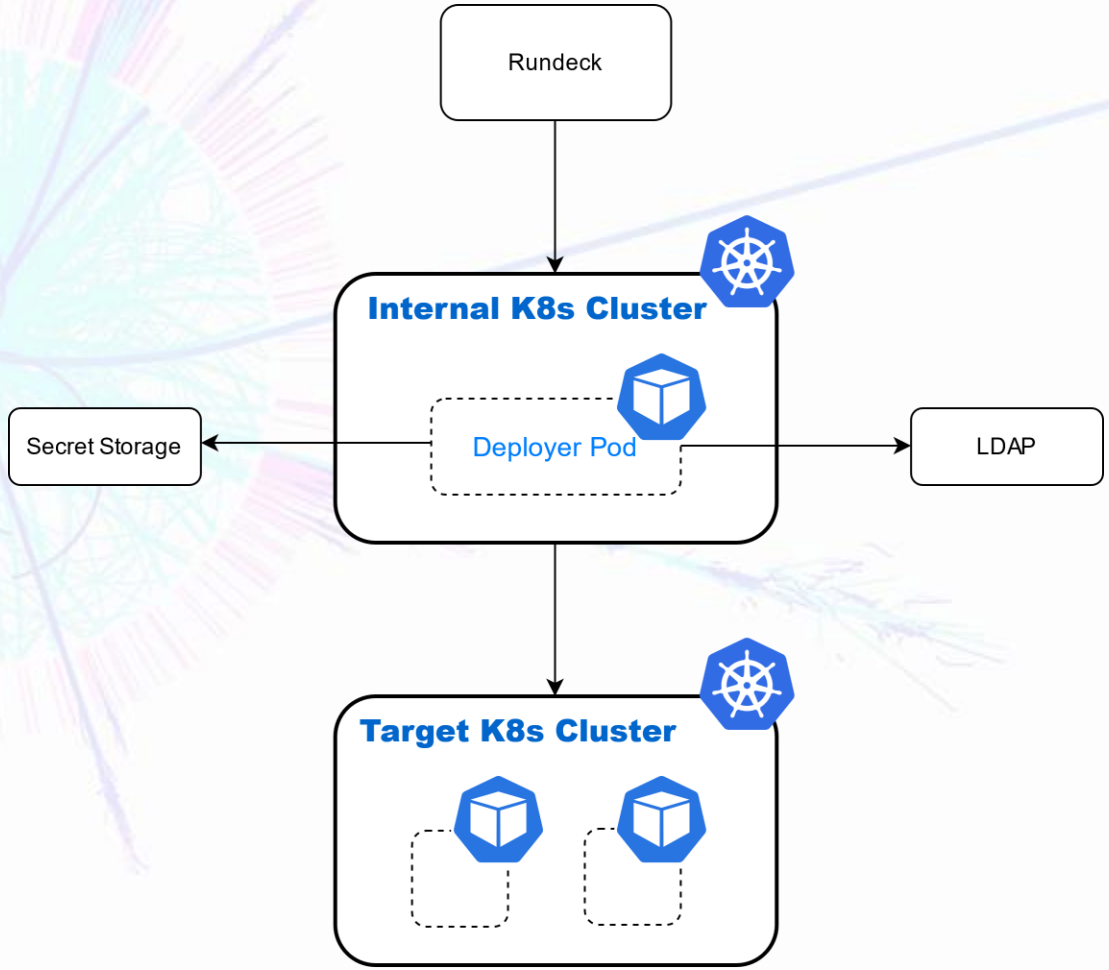
- **80+ Applications**
  - Highly Available
  - Multiple Environments
- **Infrastructure innovations of K8s**
  - Portable solutions
  - Fast provisioning
  - Self Healing

Kubernetes Cluster
@CERN



Application Deployment Process
@CERN

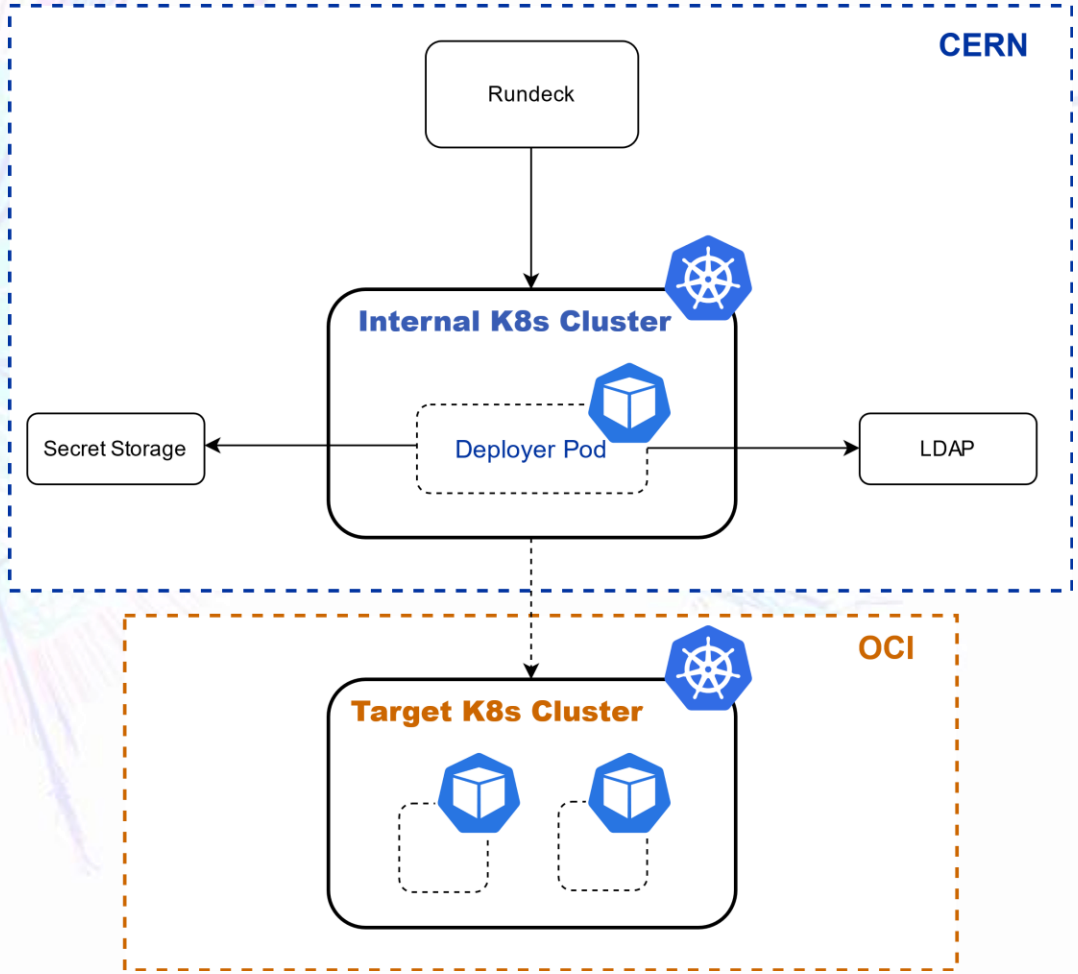
1. Rundeck creates the deployer pod
2. Get secrets from Secret Storage
3. Get configuration from LDAP
4. Generate the helm charts that package our deployments
5. Connect to target cluster and create the K8s resources
6. Destroy the deployer pod



Application Deployment Process
@OCI

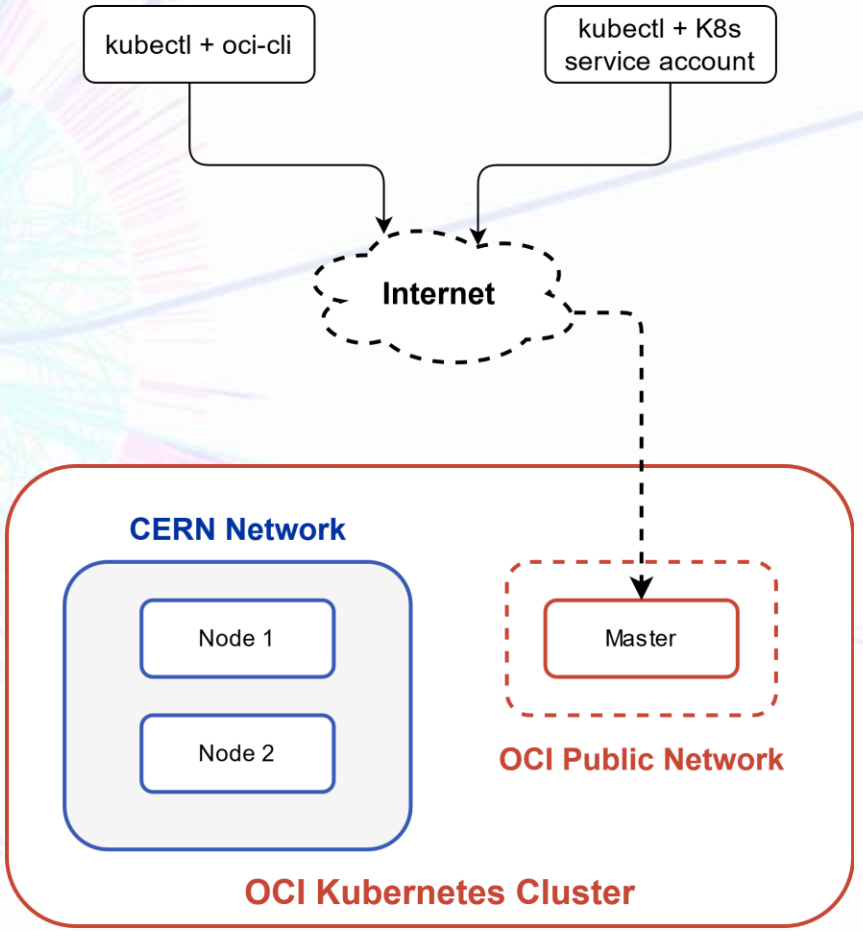
Same procedure with few customizations:

- KUBECONFIG with the k8s service account and token
- Update KUBECONFIG with CI/CD access details
- Store inside the KUBECONFIG cluster access
- Pod creation for active labelled nodes
- Custom configmap for K8s coredns



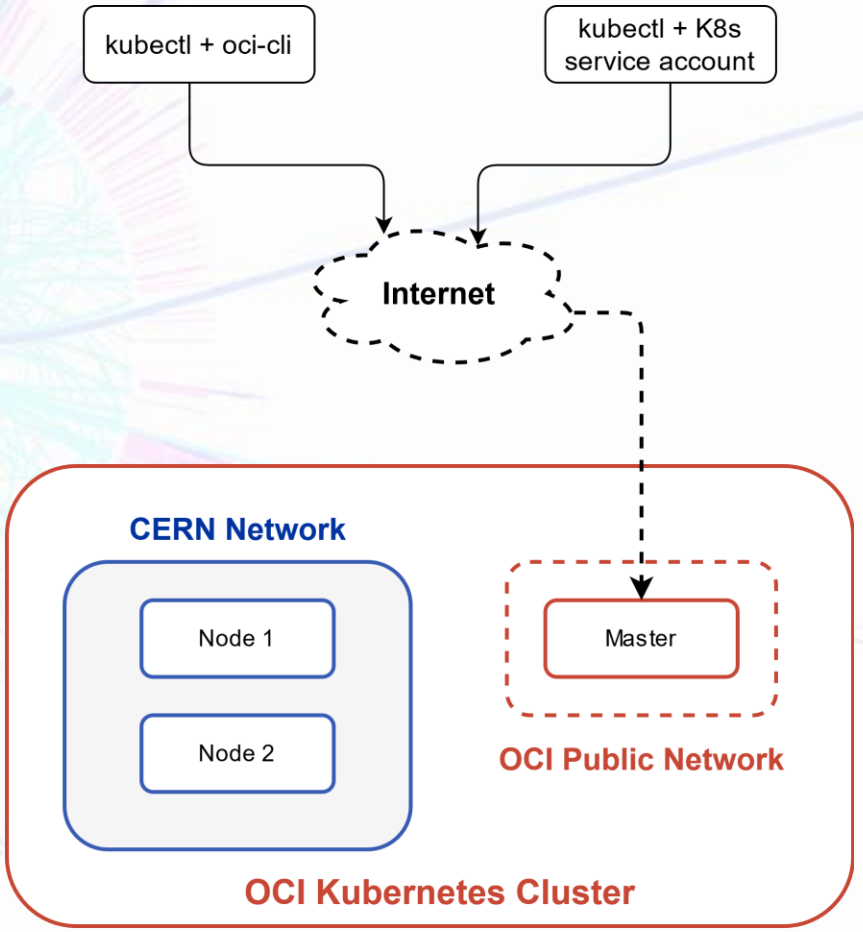
Accessing OCI Kubernetes Cluster

- **CI/CD purposes**
  - Create a Kubernetes local service account with a Kubernetes access token
  - Include the local account with the token in KUBECONFIG
- **Human access**
  - Use oci-cli which creates short lifespan access tokens



OCI Kubernetes Cluster

- Master node lives in the container engine module and is attached to the VCN
- Instantiating node pools requires the definition of the subnet
- For communication between the master and nodes you need to set the appropriate gateway



Application Load Balancing

- **Tested load balancing with 2 ways:**
  - OCI Web Console
  - OKE triggers the load balancer processes
- **What we found out:**
  - For OKE created OCI load balancers there is no possibility to keep the public IP. Consequently, load balancer recreation is essential for DNS Name being up to date
  - OKE creates a fully functional OCI load balancer service but provides very limited customization
    - Individual SSL certificates for different applications
    - Custom headers
    - The aforementioned features are supported from OCI Web Console

Can we automate?

The answer is yes, you should!

- Terraform modules for provisioning cluster creation
- Bash scripts to automate the customization procedure for OCI

Limitations

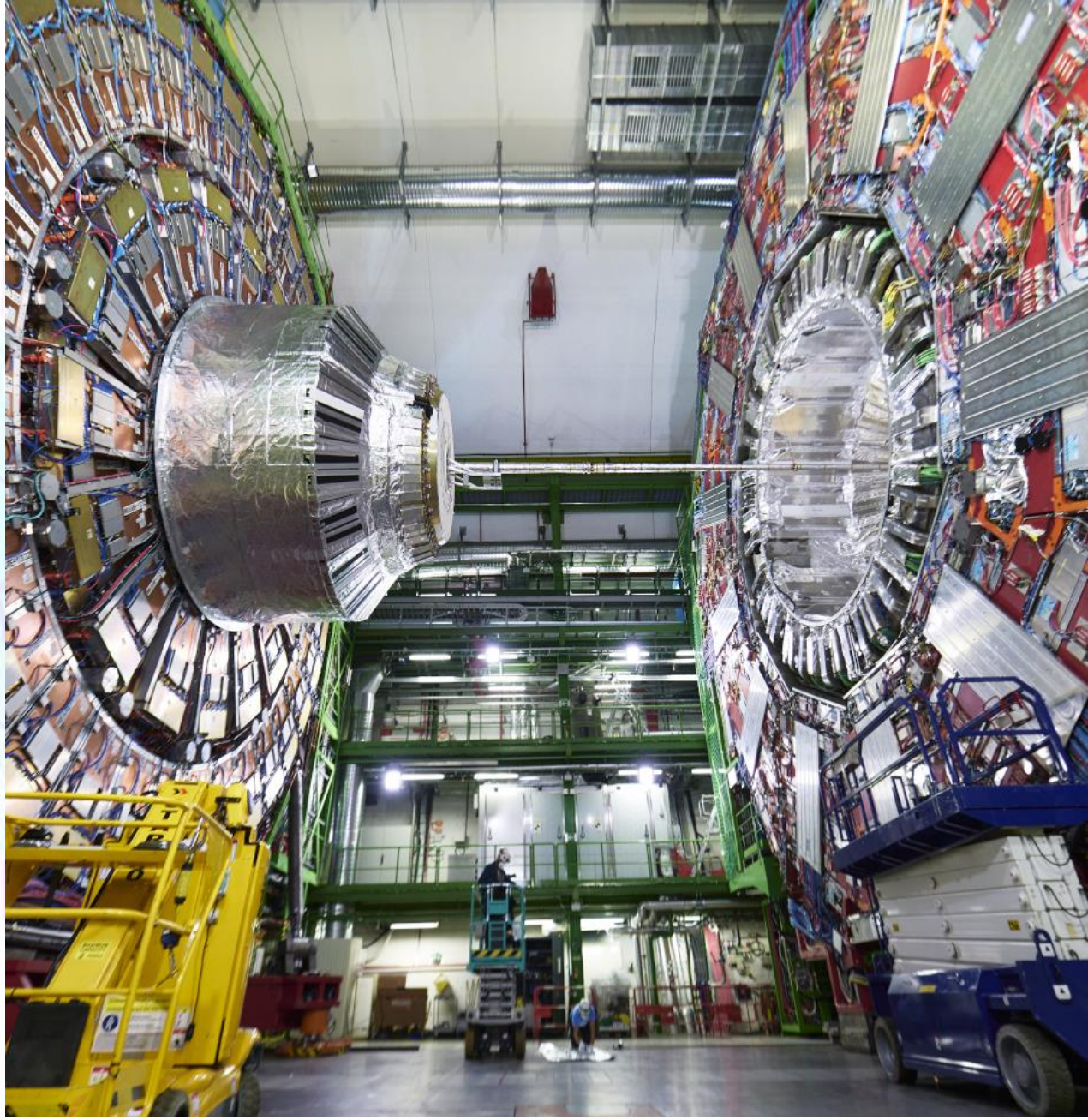
- Applications use components that are still on the CERN side and would need to be migrated to the cloud
- Architectural change on CERN side: some of the customizations on the Load Balancer by OKE could be moved to the ingress controller

Databases at CERN

- **Oracle, MySQL, PostgreSQL, InfluxDB**
- **Total Size: 4.67 PB**
- **Stored on Disk and tapes**
  - Tapes are used for long term storage
- **Storing experiment and user data**

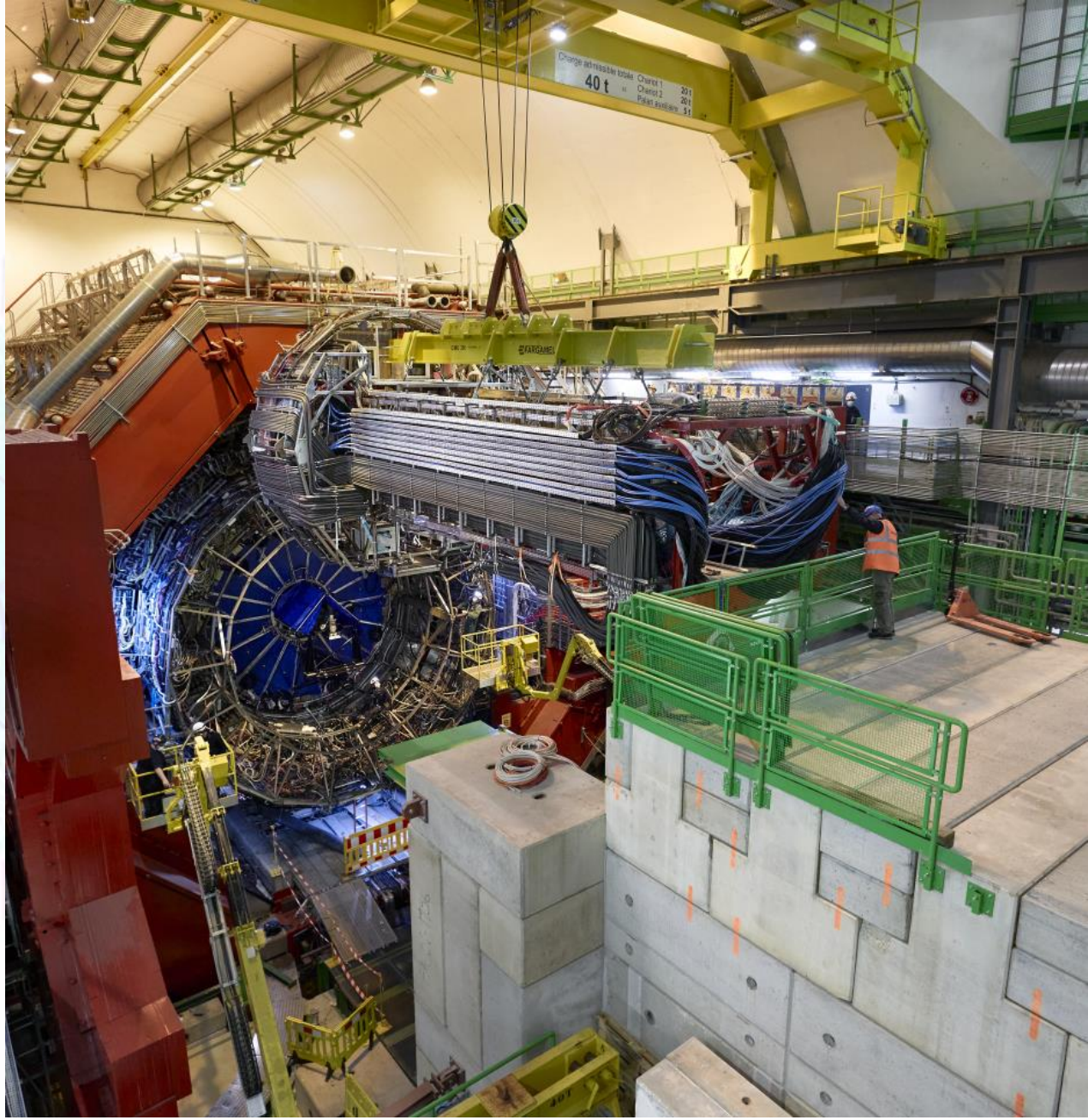
Primary/Standby Setup

- **Database Machine Creation**
  - Enterprise Edition or higher
  - Use the same DB name & version as on premise
  - Create key for opc user
  - DB unique name generated by OCI
- **SSH keys**
  - Create keys for oracle and grid user
- **Prepare the DB**
  - Delete the provisioned database (not with DBCA)
  - Copy password file



Primary/Standby Setup

- **Configure listeners and tnsnames.ora**
- **Configure Wallet (TDE)**
- **Fix RMAN Configuration**
  - Set the channel to disk (if you store in tapes)
  - Configure parallelization
- **Disable container database if you do not have PDBs**
- **Restore from Service**
- **Configure Data Guard Broker**

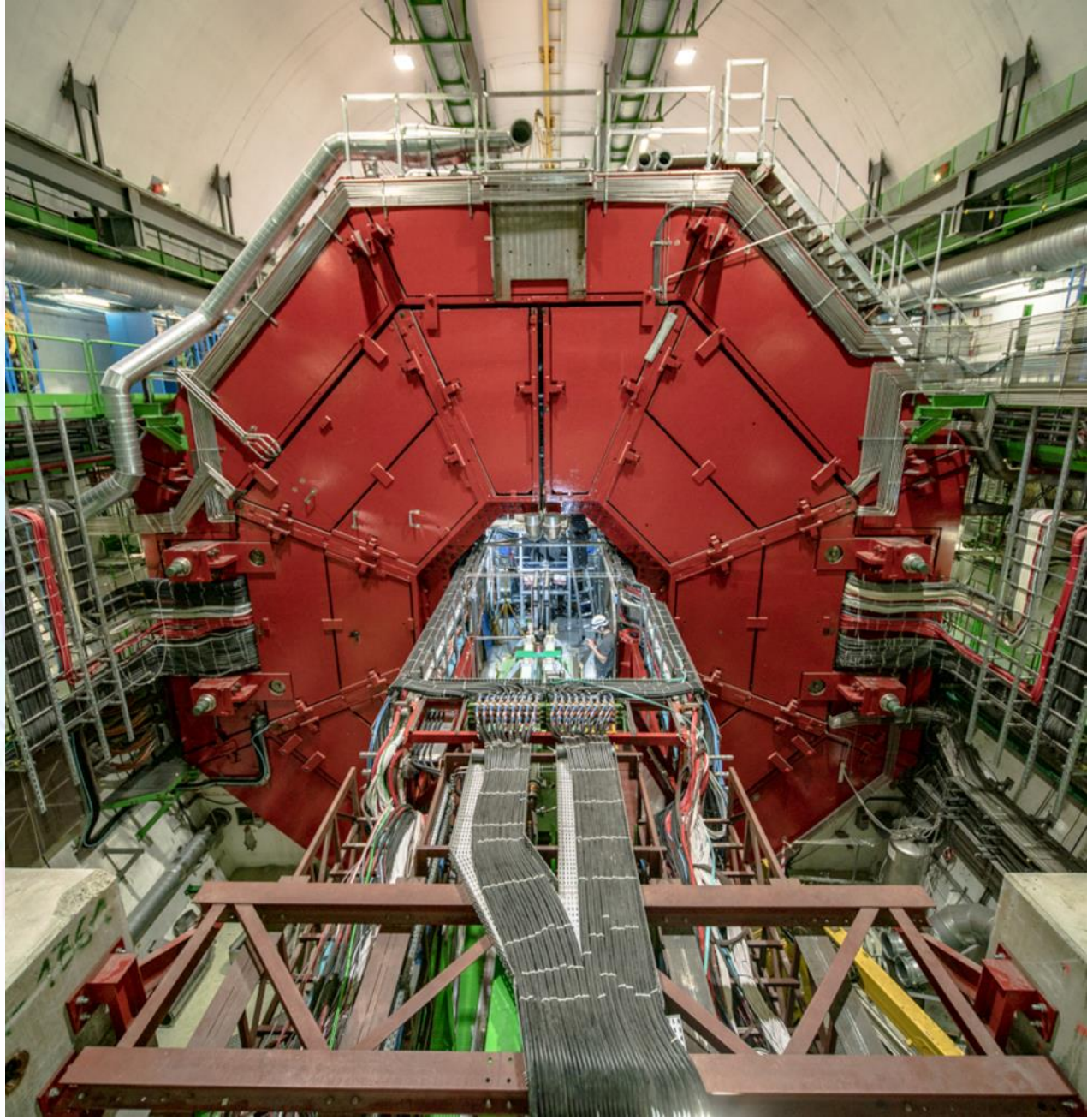


Status & Open Questions

- Provisioned successfully standby databases
- Performed upgrades of CRS and RDBMS
- Developed a bash script to automate the standby database creation
- How do we distribute tnsnames?
- How do we make visible the DBCS OCI machines?

Limitations

- 40TB limit on DBCS Virtual Machines
- Scale down the storage
- Conflicts in OCI custom image creation with our Golden Images (different way of creating them)



Databases in OCI

- 20 Databases
- ~800GB of RAM
- >50 CPU cores
- >100TB of storage

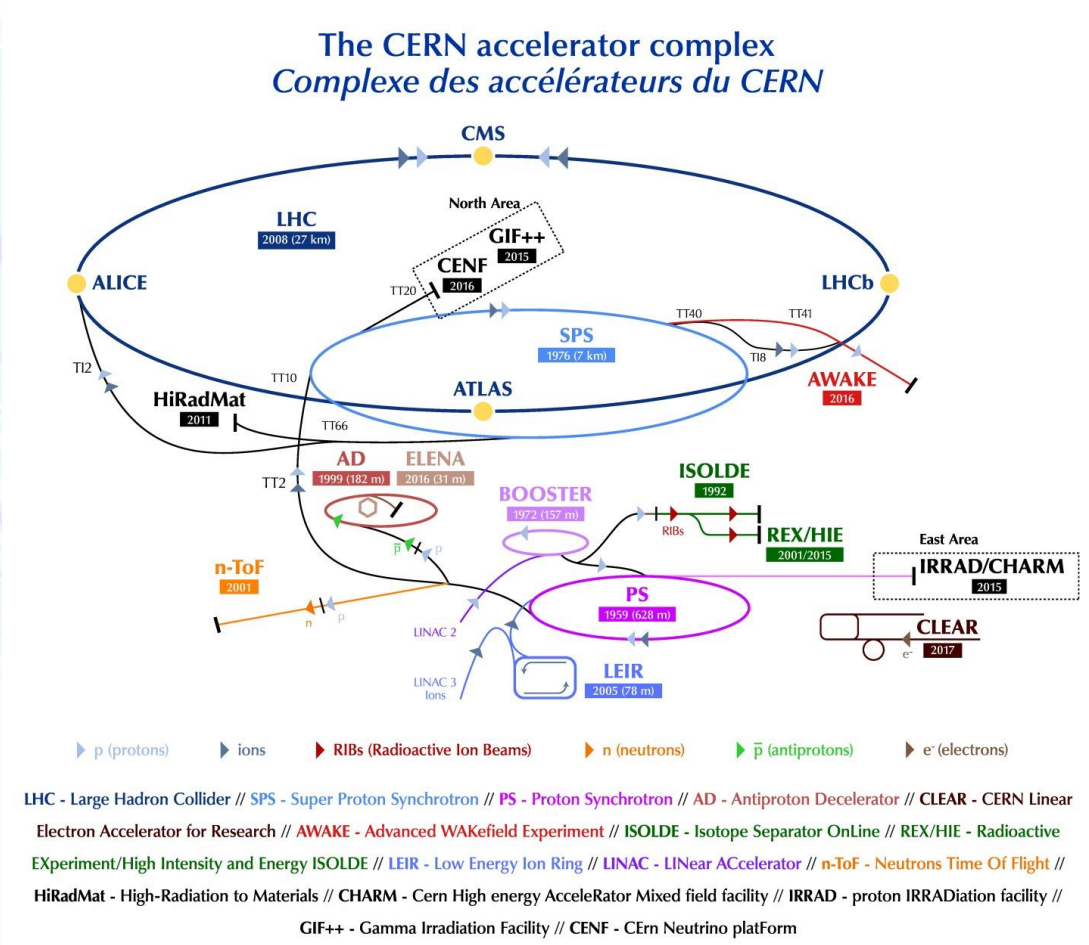
Questions?

Thank you

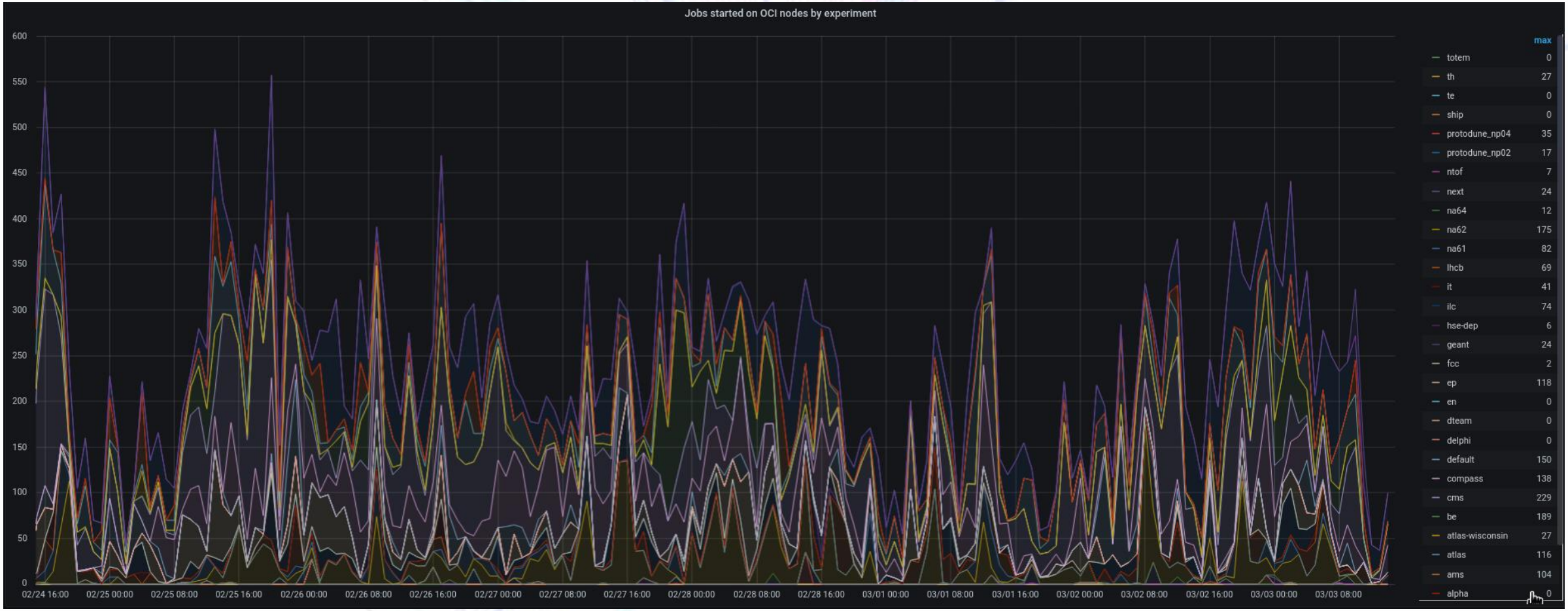
More Hybrid Cloud?

Batch

- **Upload CERN CC7 Image to OCI as custom image**
- **Implemented oci-bs to**
  1. Create a VM in the OCI using the OCI SDK with CERN custom cloudinit userdata
  2. Register Host (MAC/IP address) to CERN DNS/Network database
  3. Execute an internal tool to register the machine to CERN
- **OCI Batch nodes**
  - Set proper Puppet hostgroup
  - 128 nodes (2048cpus, 16GB memory per cpu)



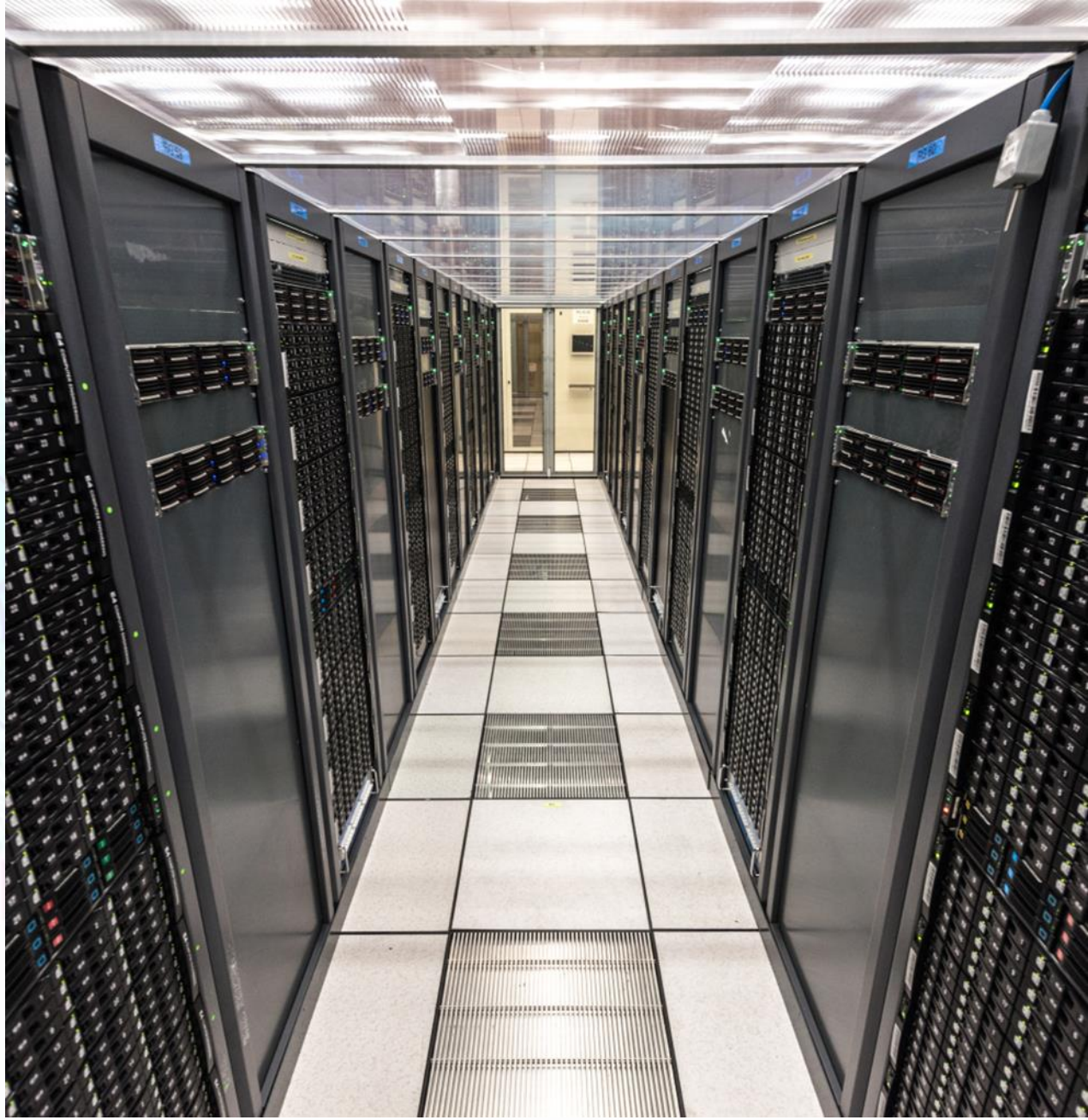
Batch Nodes in OCI



ARM Machines

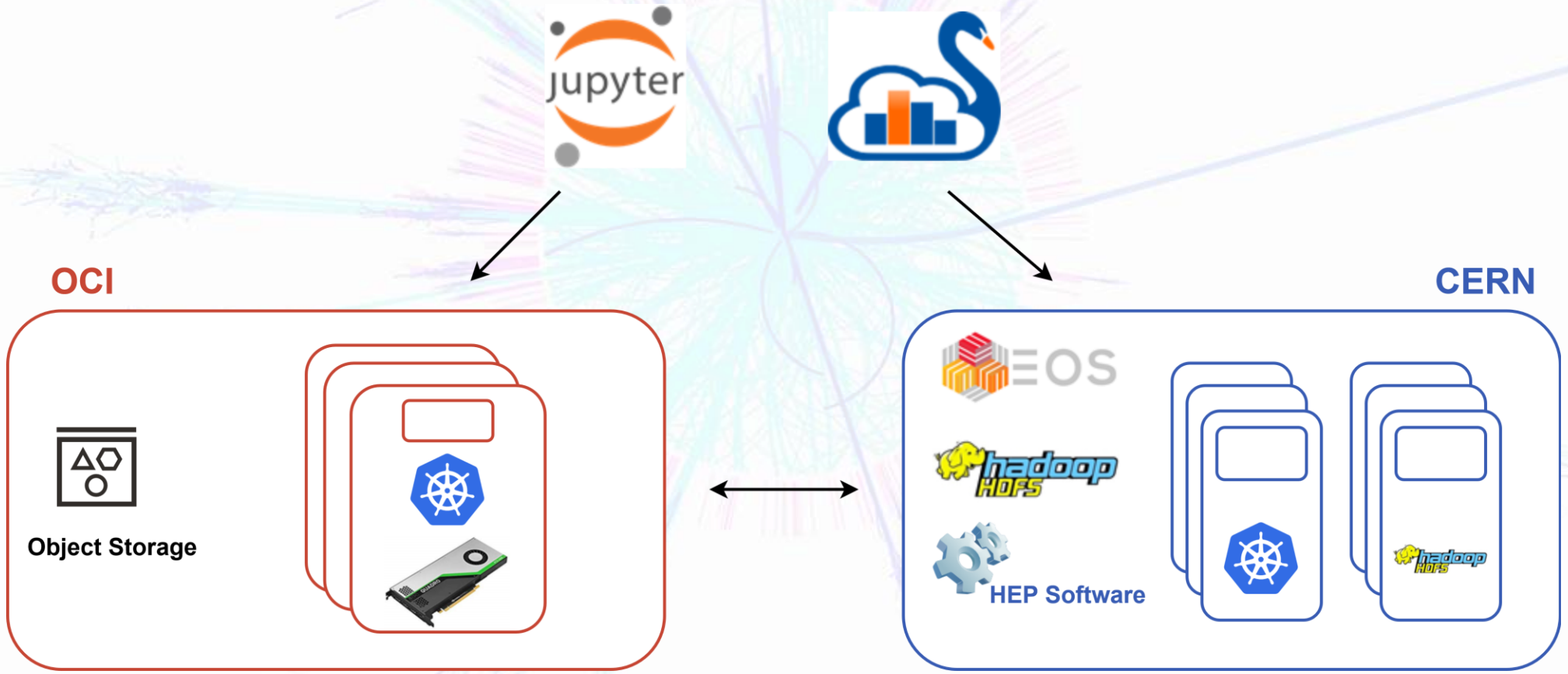
Use Cases

- Linux software building
- Lxplus (interactive shell)
- Gitlab runners



Analytics Platform

Compute across OCI and CERN cloud, storage at CERN



Exadata
Cloud@Customer

- **Shall we treat it like a black box?**
- **How do we handle:**
  - Our procedures
  - Our scripts
  - Our configurations
- **Shall we use for consolidating databases or achieving better performance for heavy databases?**

Challenges

- **Billing**
  - Difficulty in calculating on-premise cost
  - How to compare 2 solutions when some of the parameters for the result are unknown
- **Operations**
  - Backups
  - Upgrades
  - Automating different procedures
  - Components(on-premise and in the cloud) in the hybrid model
  - Monitoring and alerting

Summary

- **Understand your needs**
- **Assess your infrastructure**
  - Network
  - Resources
  - Components
- **Investigate multiple architectural concepts**
- **Evaluate isolation of resources**
  - Compartments
  - Network

Acknowledgments

CERN: Luca Canali, Riccardo Castellotti, Ignacio Coterillo Coz, Eva Dafonte Perez, Lukas Gedvilas, Eric Grancher, Jakub Granieczny , Alina Grigore, Arash Khodabandeh, Viktor Kozlovszky, Manuel Martin Marquez, Sebastien Masson, Antonio Nappi, Nemanja Nedic, Ioannis Panagiotidis, Luis Rodriguez Fernandez , Aimilios Tsouvelekakis, Artur Wiecek

Oracle: Cemil Alper, Giuseppe Calabrese, Michael Connaughton, Dmitrij Dolgušin, David Ebert, Brent Eyler, Maciej Gruszka, Sevgi Guzzella, Gavin Larson, Vincent Leocorbo, Will Lyons, Pauline Mahrer, Marc Meignier, Çetin Özbütün, Oguz Pastirmaci, Cristobal Pedregal-Martin, Arun Ramakrishnan, Alexandre Reigada, Monica Riccelli, Patrice Scattolin, Engin Senel, Garret Swart, Peter Szegedi, Thomas Teske, Reiner Zimmermann

Photo Index

Slide 9: CERN Data Centre

Slide 10: The ALICE detector's First Level Processor (FLP) system

Slide 25: Santa Claus visit in ATLAS control room

Slide 26: LHC cryogenic installations underground near LHCb site

Slide 27: Installation of final sector onto the ATLAS New Small Wheel (NSW)

Slide 28: CMS during the final stages of LS2

Slide 29: ALICE experiment Miniframe installation

Slide 30: The COMPASS experiment

Slide 31: ALICE Magnet

Slide 32: LHCb Removal module in chamber

Slide 35: The accelerator Complex

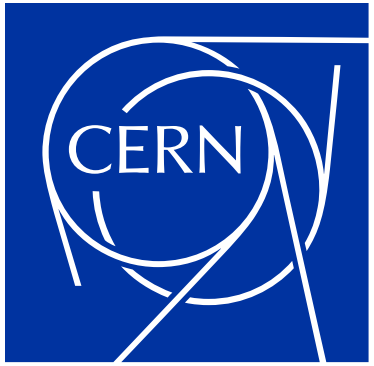
Slide 37: CERN Data Centre

Slide 39: Exadata Machine

Questions?

Thank you

aimilios.tsouvelekakis@cern.ch



home.cern