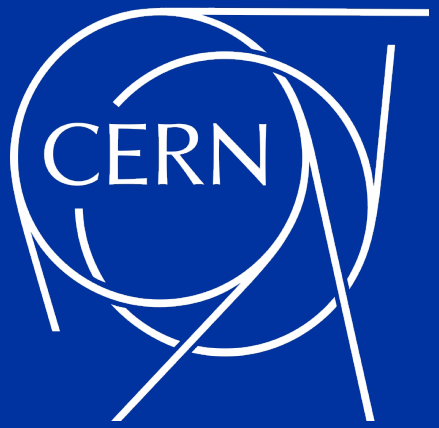


MySQL Community Edition at CERN

Abel Cabezas Alonso

30th January 2025

Abel Cabezas Alonso

➡️ Database Engineer at CERN since 2019
➡️ Transition as DevOps engineer
➡️ Early career as Software Developer

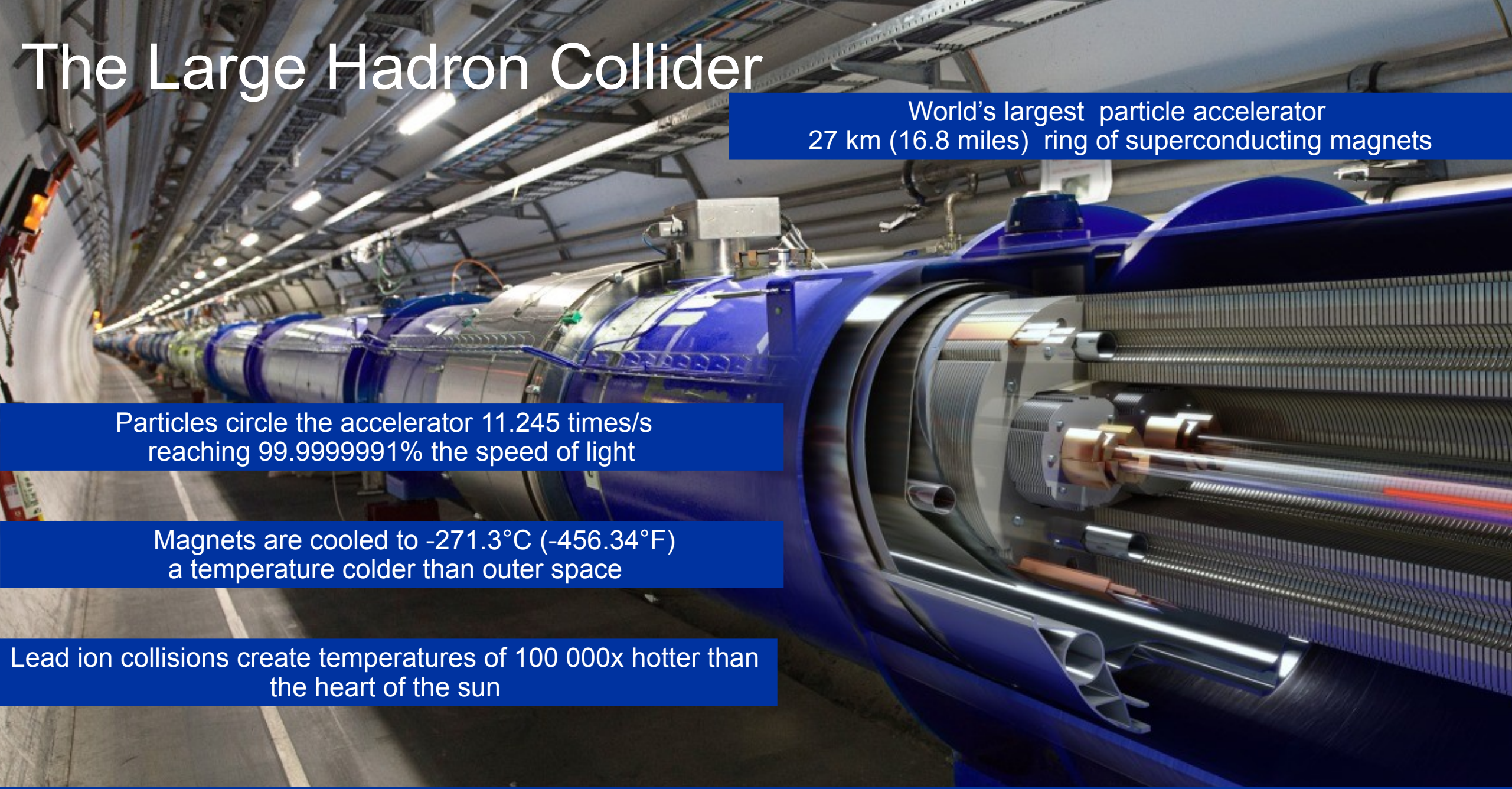
Abel Cabezas Alonso

abel.cabezas.alonso@cern.ch

Our Mission

- Established in 1954
- 24 member states
- Intergovernmental Organisation dedicated to scientific research
- Our goal is to understand the most fundamental laws of the universe
- CERN is the world's biggest laboratory for particle physics
- Unite people from all over the world to push the frontiers of science and technology

The Large Hadron Collider

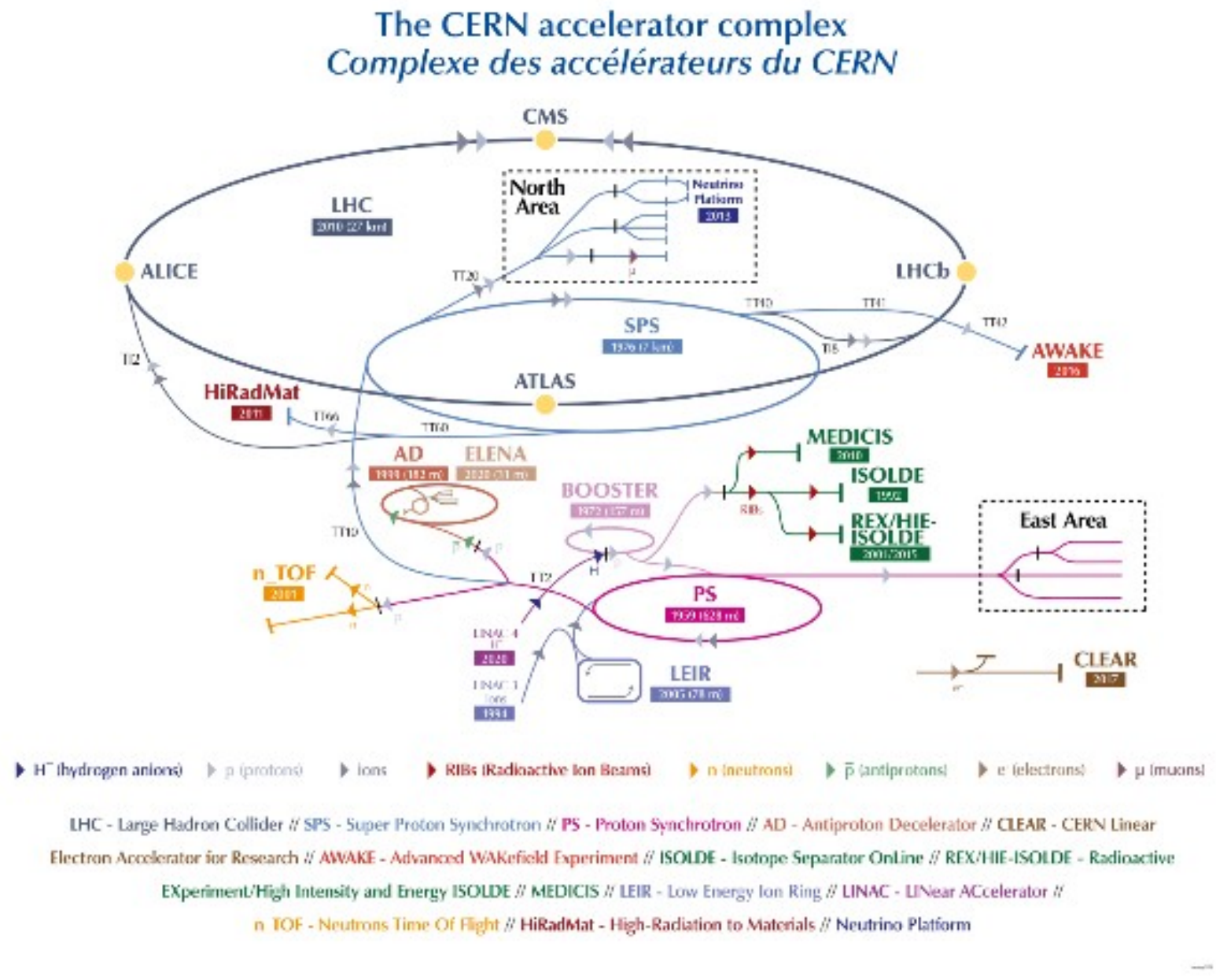


World's largest particle accelerator
27 km (16.8 miles) ring of superconducting magnets

Particles circle the accelerator 11.245 times/s
reaching 99.9999991% the speed of light

Magnets are cooled to -271.3°C (-456.34°F)
a temperature colder than outer space

Lead ion collisions create temperatures of 100 000x hotter than the heart of the sun



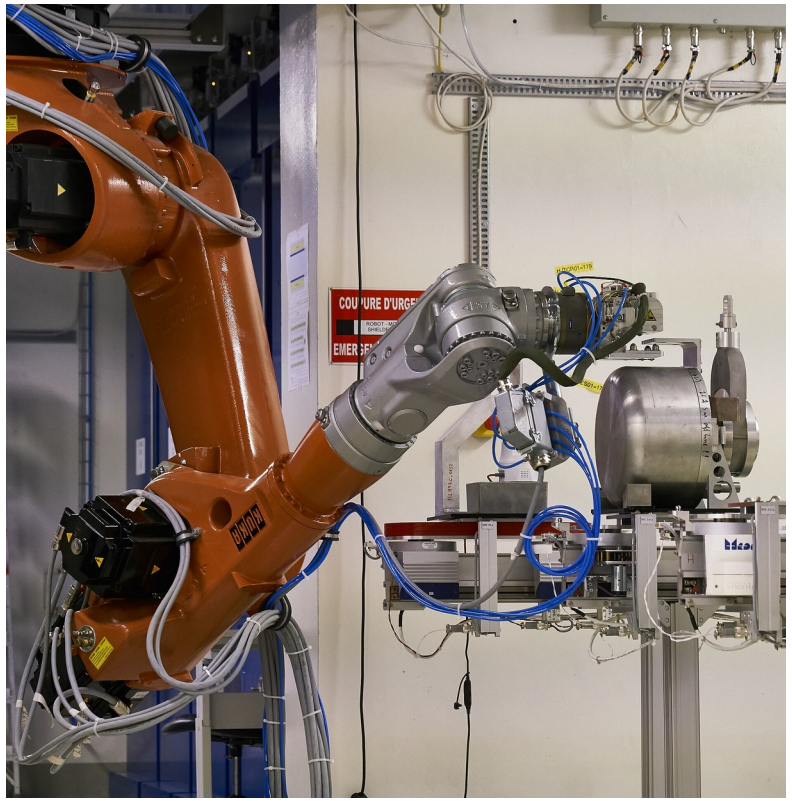
Succession of machines that accelerate particles to increasingly higher energies

Each machine boosts the energy of a beam of particles before injecting it into the next machine in the sequence, being the LHC the last element of this chain

The accelerator complex serves not only the LHC, but also a rich and diverse experimental program.

Beyond the LHC

MEDICIS



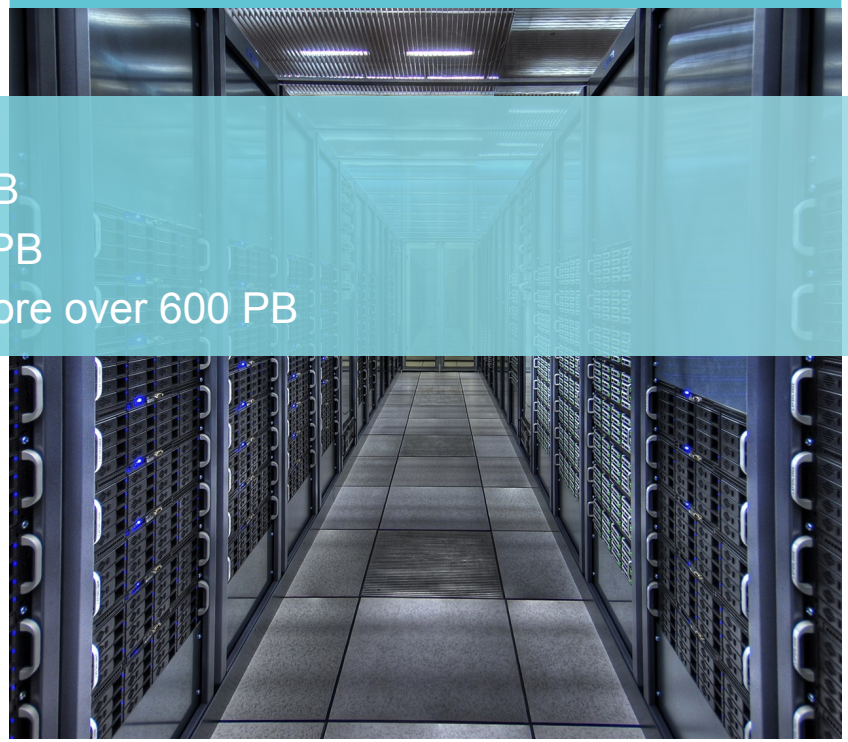
AMS (Alpha Magnetic Spectrometer)

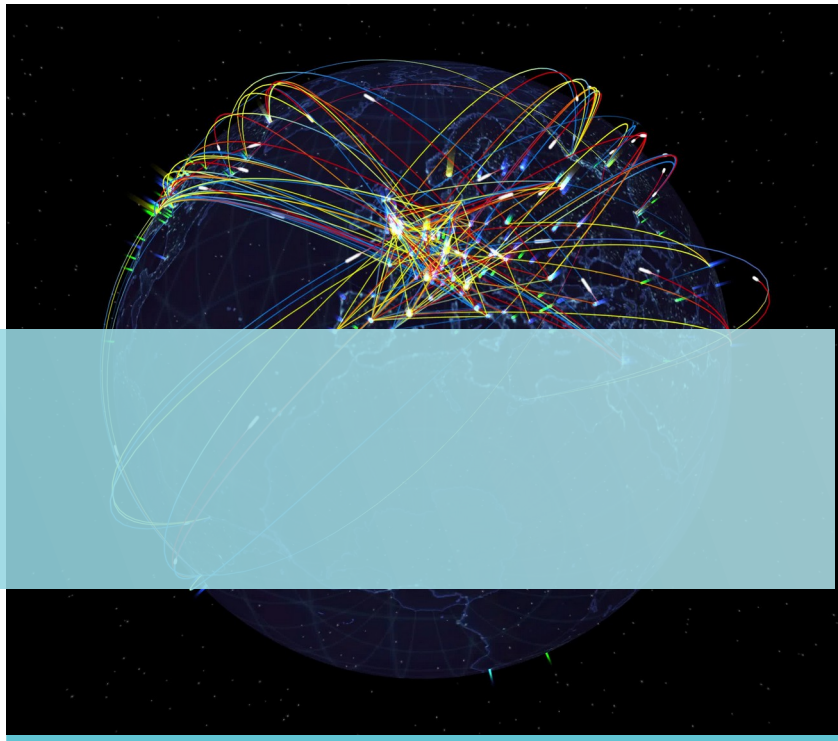
Antimatter factory

The Worldwide LHC Computing Grid (WLCG)

1 PB of data per second
~ 1% of the data is kept (events with specific characteristics)

Tier0:
Data processing and Tape archival happens + data distribution to other tiers
~ 200 PB of data per year





WLCG:
170 collaborating centres
42 countries
Data analysis

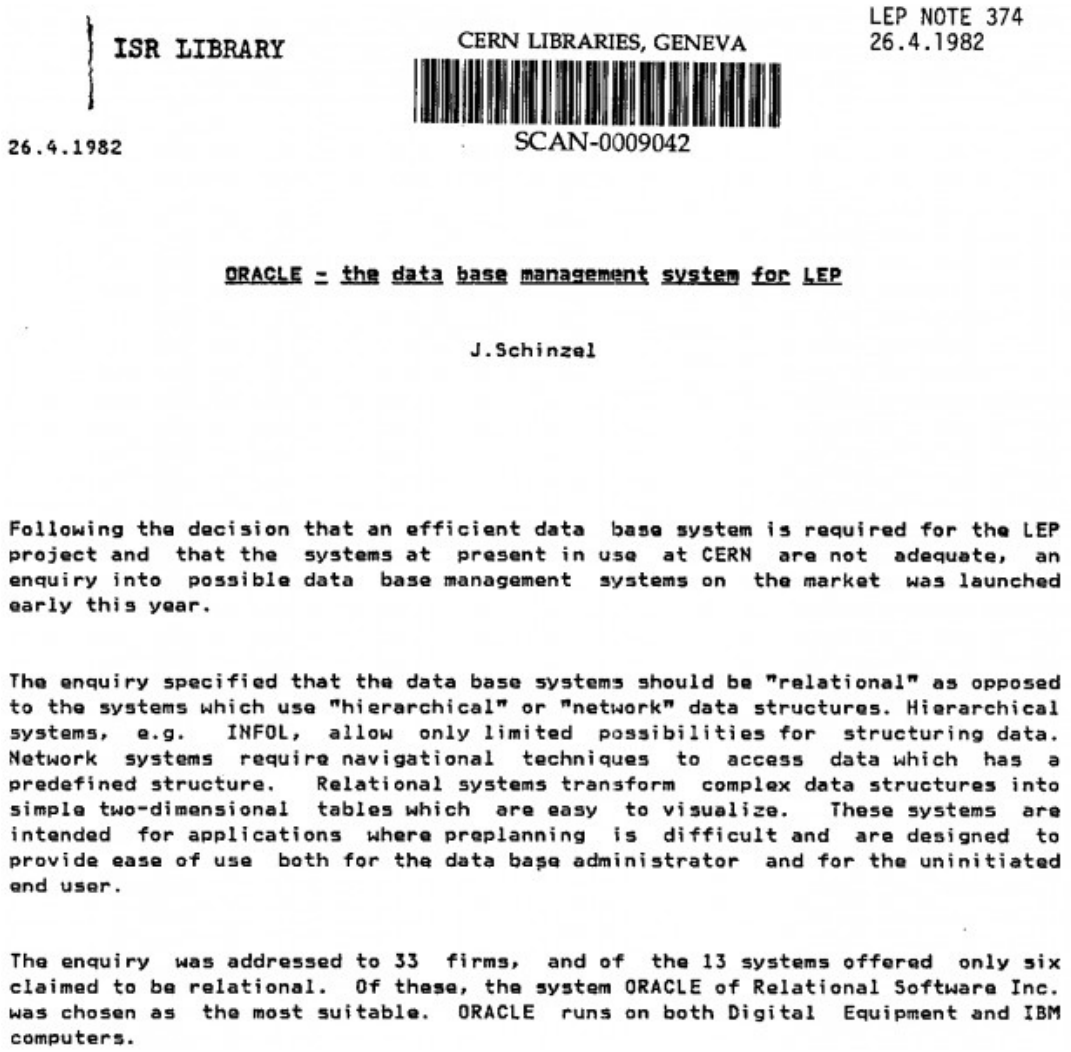
CERN Science is Data Intensive

- Run 1 (2009-2013) we stored 65 PB
- Run 2 (2015-2018) we stored 209 PB
- Run 3 (2022-2026) we expect to store over 600 PB


IT Department

Databases at CERN: Oracle

- **Oracle databases since 1982**
  - 105 Oracle databases,
  - More than 11.800 Oracle
  - RAC, Active DataGuard,OEM, RMAN, Cloud…
  - Complex environment
  - Used by:
    - Administrative Information Services
    - Engineering systems
    - Accelerator and experiments
    - etc.
  - ≈ 5PB of data



ISR LIBRARY

CERN LIBRARIES, GENEVA

LEP NOTE 374
26.4.1982

26.4.1982

SCAN-0009042

ORACLE - the data base management system for LEP

J.Schinzel

Following the decision that an efficient data base system is required for the LEP project and that the systems at present in use at CERN are not adequate, an enquiry into possible data base management systems on the market was launched early this year.

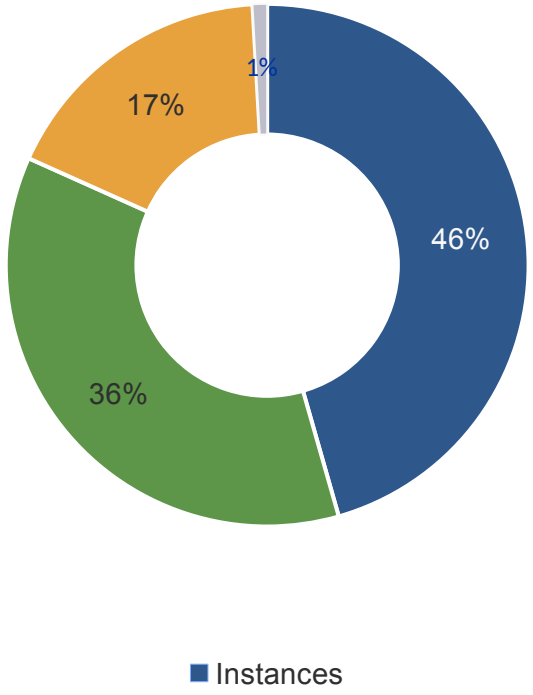
The enquiry specified that the data base systems should be "relational" as opposed to the systems which use "hierarchical" or "network" data structures. Hierarchical systems, e.g. INFOL, allow only limited possibilities for structuring data. Network systems require navigational techniques to access data which has a predefined structure. Relational systems transform complex data structures into simple two-dimensional tables which are easy to visualize. These systems are intended for applications where preplanning is difficult and are designed to provide ease of use both for the data base administrator and for the uninitiated end user.

The enquiry was addressed to 33 firms, and of the 13 systems offered only six claimed to be relational. Of these, the system ORACLE of Relational Software Inc. was chosen as the most suitable. ORACLE runs on both Digital Equipment and IBM computers.

Databases at CERN: DBOD

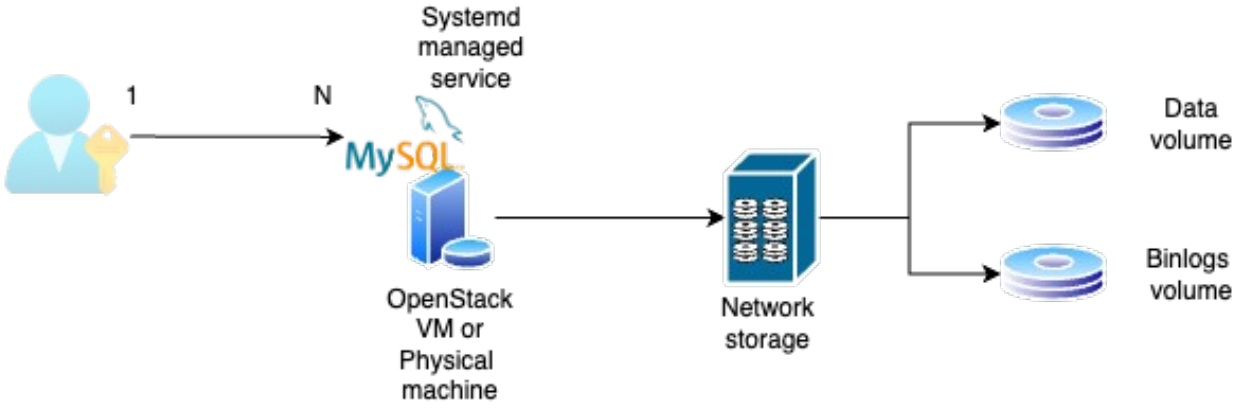
- **Database On Demand (DBoD)**

  - DBaaS conceived in 2011
  - A number of key database applications were running on
  - user-managed MySQL database instances
  - MySQL was the chosen/only supported technology for some applications
  - Empowers users to be their own DBA
  - More than 1200 database server instances
    - ≈600 MySQL, ≈400 PostgreSQL, ≈200 InfluxDB, ≈10 TimescaleDB
  - Flexible architecture allowing to easily integrate other RDBMS
  - Used by:
    - CERN's Single Sign On
    - CERN's private cloud based on Openstack
    - Experiments (ATLAS, LHCb, etc.)
    - WLCG file transfer service
  - ≈150 TB of data



MySQL deployment

- On premise deployment (2 DC)
- Several MySQL binaries per host
- Several database instances per host
- Two different Netapp NFS volumes per DB instance:
  - data directory + binary log directory
- Types of deployment:
  - Single instance
  - Replication for disaster recovery
  - Replication to scale out reads
  - ProxySQL + primary-replica
  - MySQL InnoDB cluster



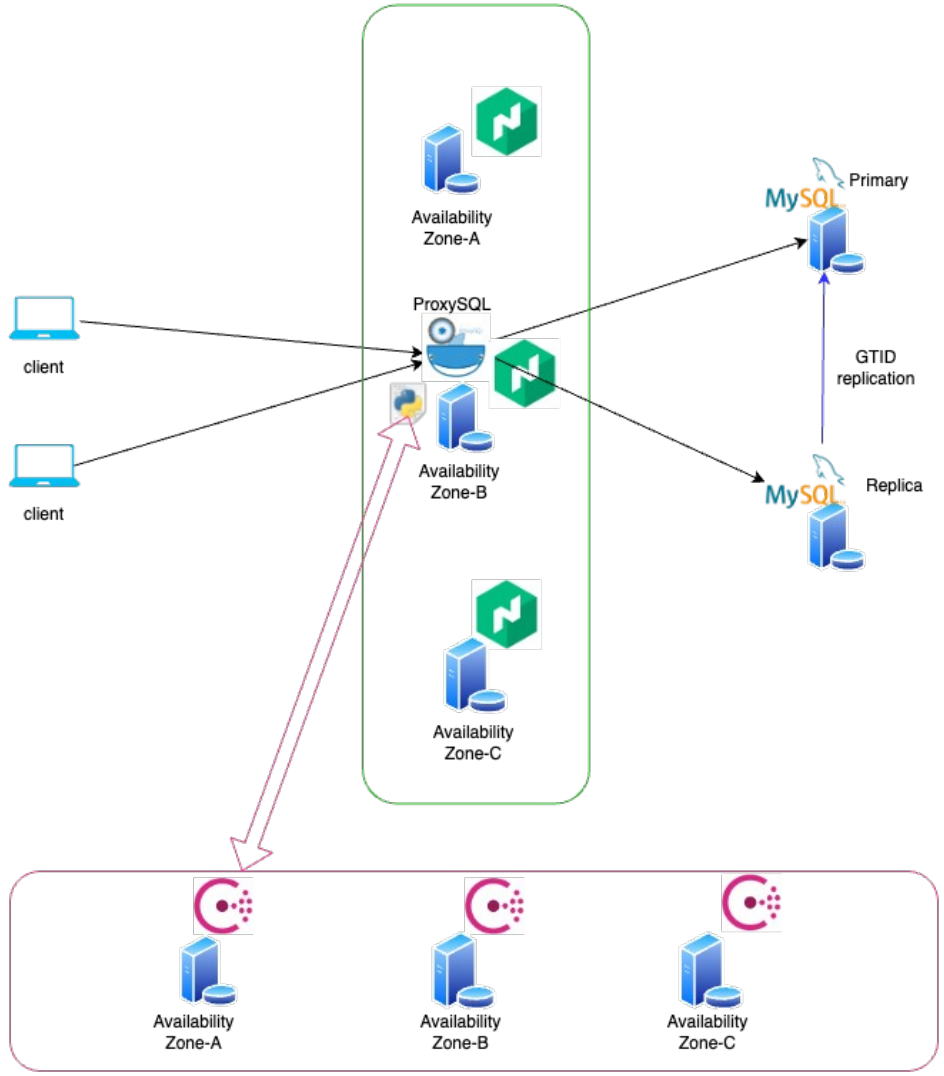
Towards high availability: ProxySQL + semisync GTID replication

Advantages over simple replication with manual failover

- Minimised RTO
- Built-in monitoring module
- Not designed for reconfiguring the topology
- Scheduler module to extend logic:
  - Failover logic
  - Resolution of conflicts (multiple primaries)
  - Monitoring of replication channel

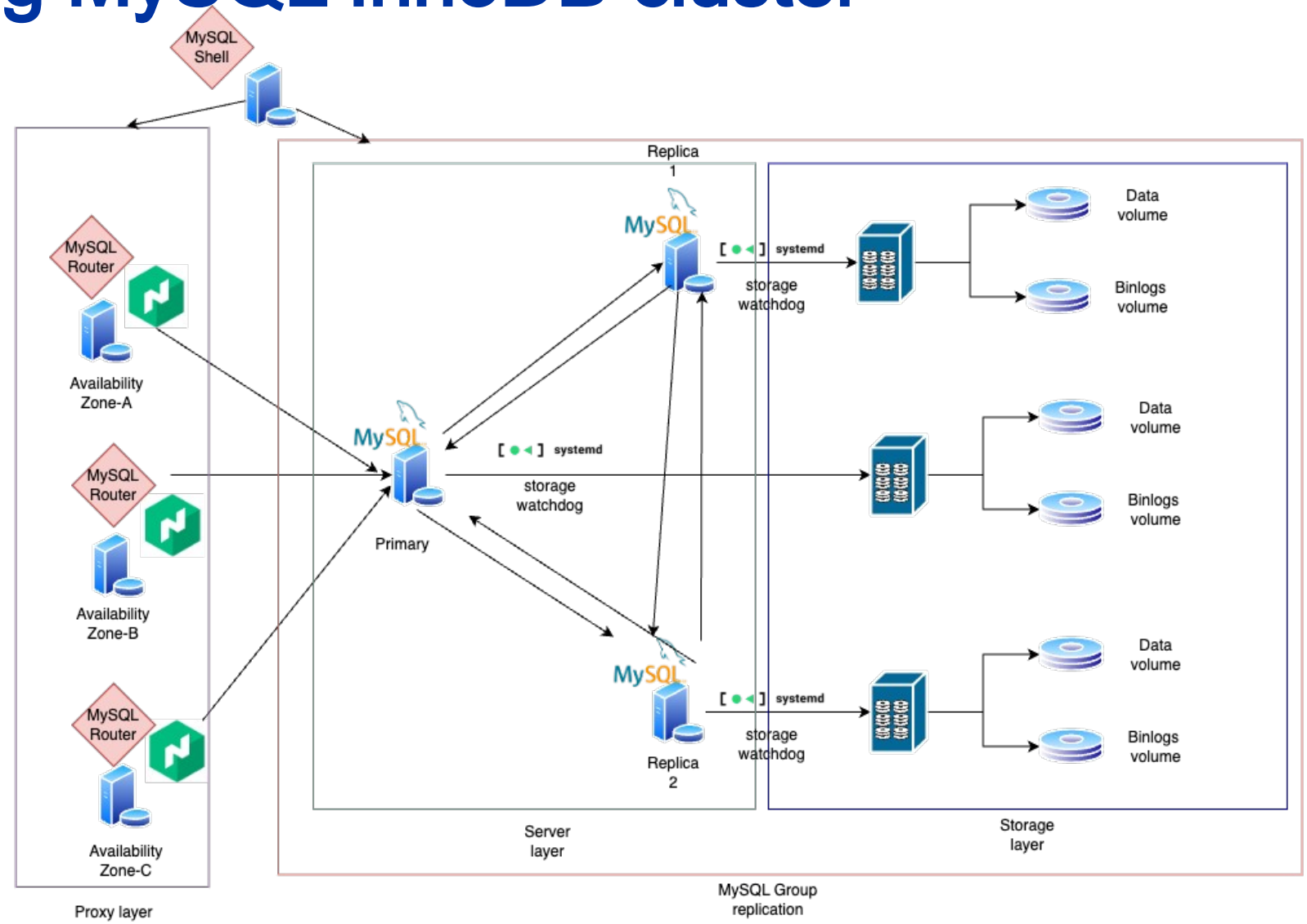
Limitations

- Not a pure HA solution
- SPOF
- Not possible to deploy several proxies for our use case
- No built-in failover/failback
- Big maintenance effort



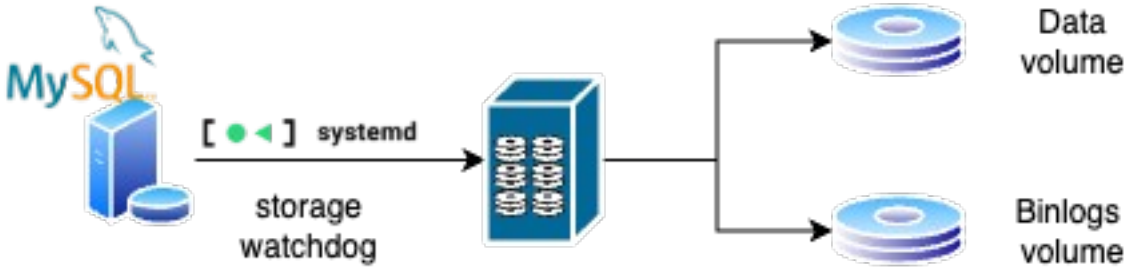
High Availability using MySQL InnoDB cluster

- Integrated solution
  - Easy to deploy, integrate and maintain
- Simplified management/automations with MySQL Shell
  - Configure, force quorum, reboot from complete outage, rescan, dissolve, etc
- No SPOF by deploying several MySQL Routers
- Very good documentation on disaster recovery
- Fully fledged HA solution
- Seamlessly scale out reads through MySQL Router
- Extended functionality
  - Storage watchdog



Storage watchdog for InnoDB cluster instances

- From our experience running DBs with Network attached storage when there are connectivity problems:
  - The MySQL process enters in Ds+ state (uninterruptible sleep)
  - Once the network connectivity is resumed:
    - Crass recovery
    - Process enters in Z state
  - Group replication does not see this as an error, so it won't force a failover
- Our solution:
  - Probe host connectivity with the storage at fixed intervals
  - If the connectivity fails for x consecutive probes -> kill the instance
  - Group replication will take care of promoting the most up to date replica



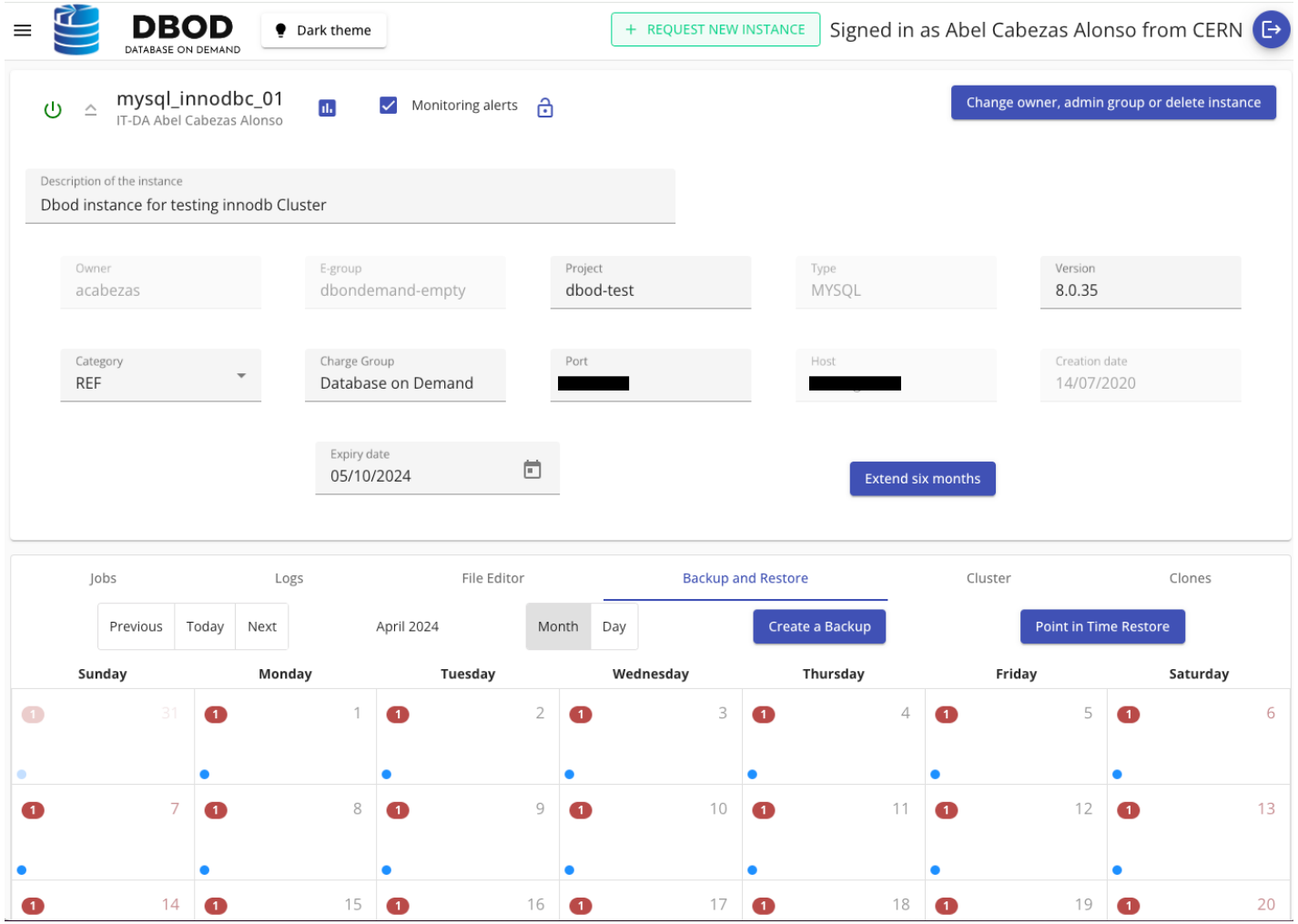
Automation

Web automation

- Automated backup and recovery
- Character set conversion
- MySQL Shell upgrade checker
- Management of configuration files
- Cloning
- Integrated upgrades
  - Primary-replica upgrade logic

Ops automation

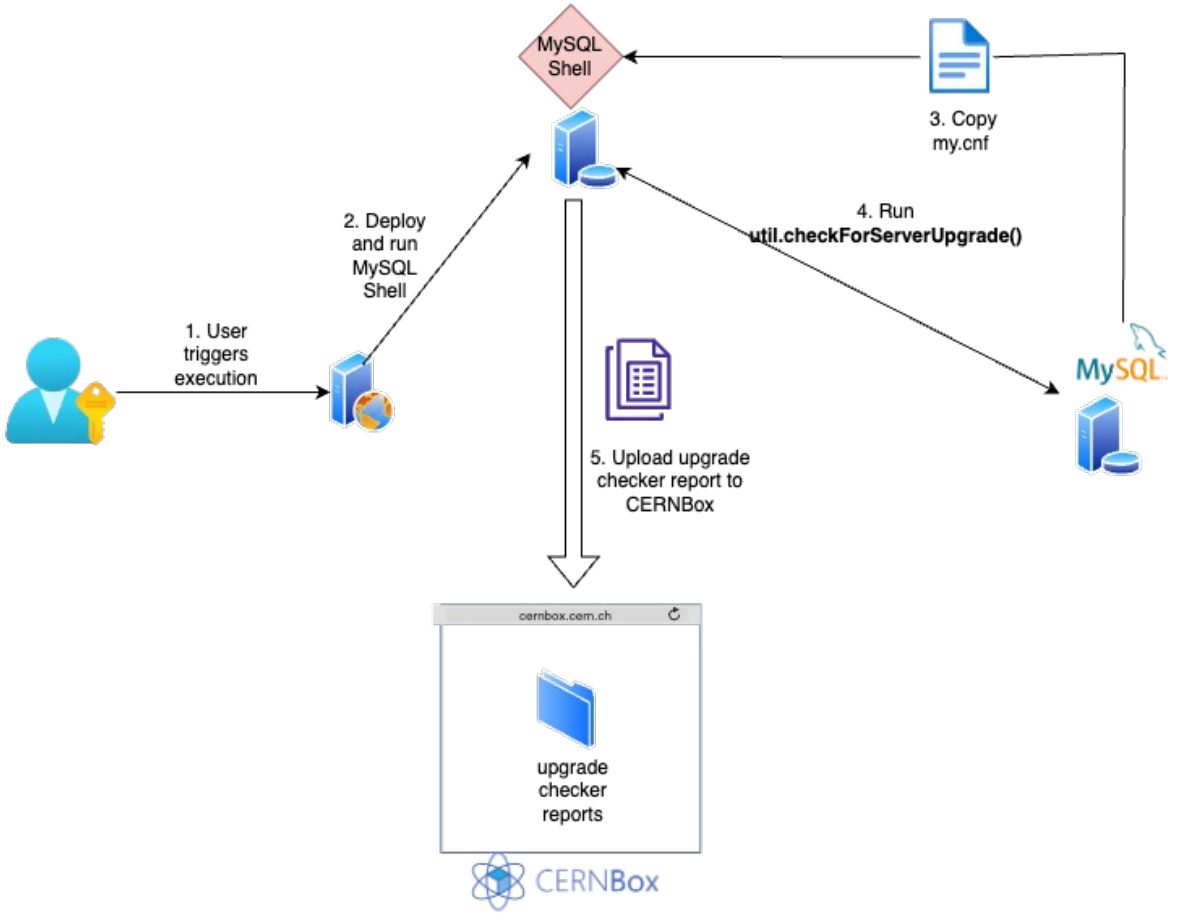
- Continuous validation of backups with PITR
- Instance and storage migration
- Automated replica provisioning
- Automated replication switchover
- Detection of idle instances
- Integrated password cracker



Automating instance upgrades with MySQL Shell

Upgrade checker utility

- MySQL Shell integration with extended logic
- Can be run on demand
- Upgrades disabled by default
- Only enabled once the upgrade checker report is “clean”
- Report shared via cloud storage
- Users can correct errors and warnings before upgrading autonomously
- Extended logic for replication setups
- Exceptionally for upgrades to 8.4 we modify the my.cnf removing any removed variables.



Automating utf8mb3 character set conversion

utf8mb3 is deprecated

- Instances coming from 5.6.x / 5.7.x
- Run once a day an automated check looking for utf8mb3 usage
- Enable the automated charset conversion on the web interface for the affected instances
- Allow dry-run:
  - Generates only DDL to be applied
- Run conversion
  - Generate DDL before and after + conversion log
- Recommended to first test in a cloned instance to avoid surprises like:

  - ERROR 1074
    Column length too big for column 'foo' (max = 16383); use BLOB or TEXT instead

  - A VARCHAR column can only accommodate up to 16383 characters for the utf8mb4 character set

- home.cern