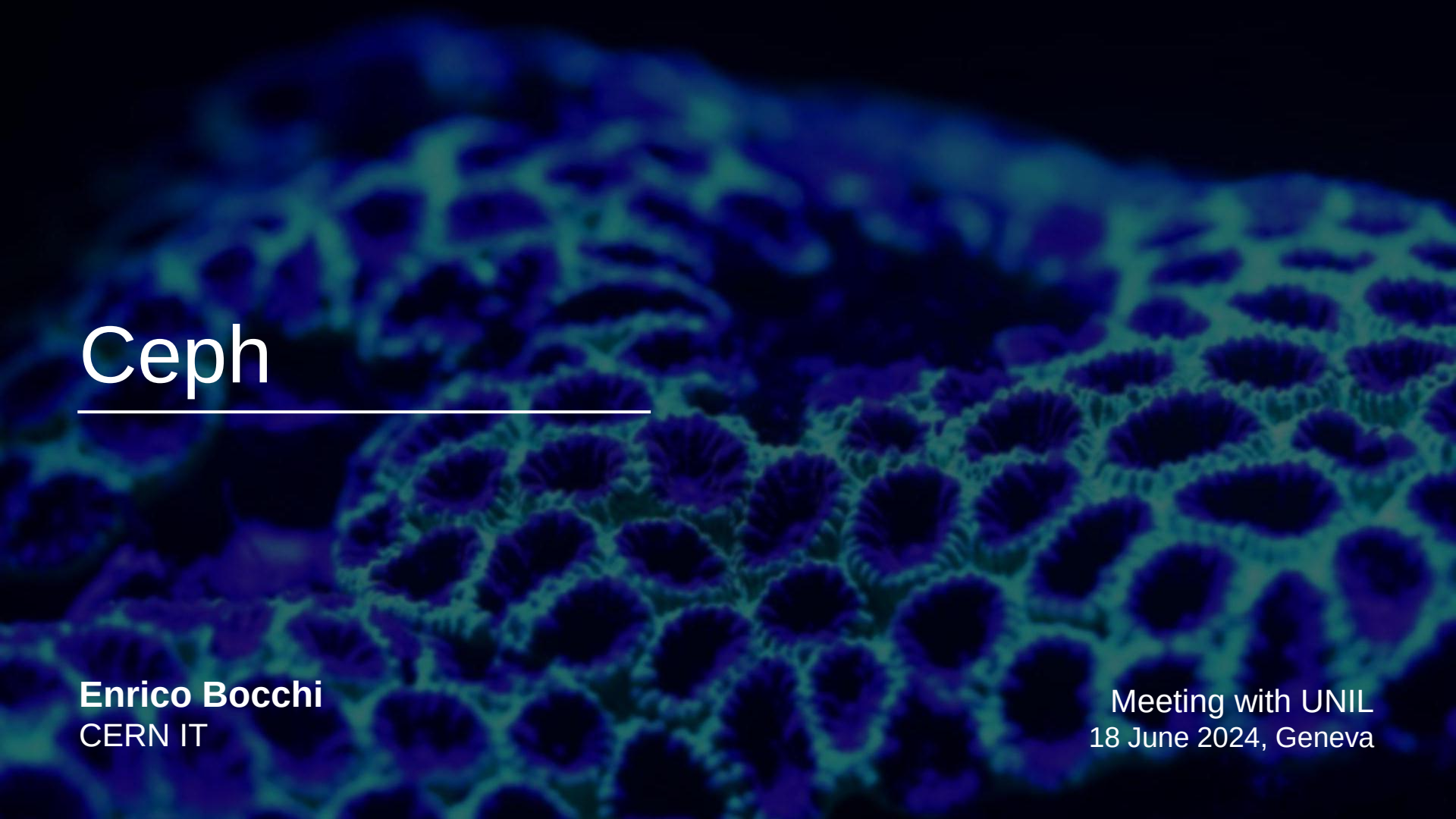


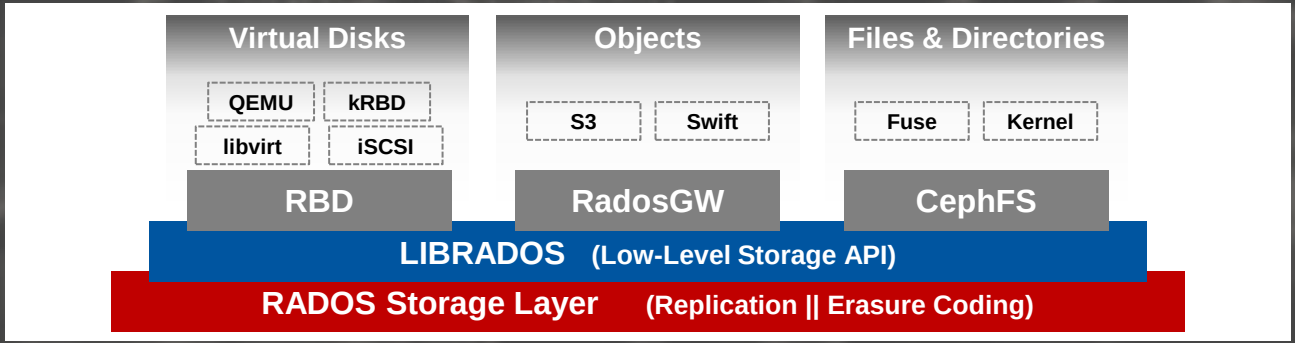
Ceph

Enrico Bocchi
CERN IT

Meeting with UNIL
18 June 2024, Geneva

What is Ceph?

- Distributed Storage System, Open Source
- Reliable storage out of unreliable components:
  - Runs on commodity hardware (IP networks, HDDs/SSDs/NVMes)
  - Favors data consistency and correctness over performance and availability
- Elastic and self-healing:
  - Scale up or out online and under load (or similarly shrink)
  - Automatic recovery from HW failures, res-establishing desired redundancy



Our Cluster Fleet

Application		Size (raw)	Version
RBD (OpenStack Cinder/Glance, `krbd`)	*Production, HDDs*	24.5 PB	Pacific, Quincy
	Production, full-flash (EC 4+2)	643 TB	Pacific
	HyperConverged (HVs with flash storage, EC 2+2)	265 TB	Quincy
CephFS (OpenStack Manila, K8s/OKD PVs, HPC)	*Production, HDDs*	12.6 PB	Pacific, Quincy
	Production, full-flash	1.2 PB	Pacific
	HyperConverged (HVs with flash storage, EC 4+2)	220 TB	Quincy
RGW + RBD Backup (2nd location)	*Production (4+2 EC)*	28.7 PB	Pacific
RGW Multi-Site	*Pre-Production (4+2 EC)*	4.2 PB	Reef
CERN Tape Archive (CTA)	*Tape DB and Disk Buffer*	235 TB	Pacific

A Brief Service History

- 2013: 300TB proof of concept, 3 PB in production for RBD
- 2014-15: Erasure coding, RADOS striper
- 2016-17: 3PB to 6PB with no downtime
- 2018: S3 + CephFS in production
- 2019: Optimizing CephFS for HPC applications
- 2020: Backup cluster in 2nd location (S3)
- 2021: RBD Storage Availability Zones, HW expansion
- 2022: 17 clusters ~65PB, CephFS physical move with 0-downtime
- 2023: kernelRBD in production,
  Explorations in Business Continuity / Disaster Recovery

- **2024: New Data Centre!**

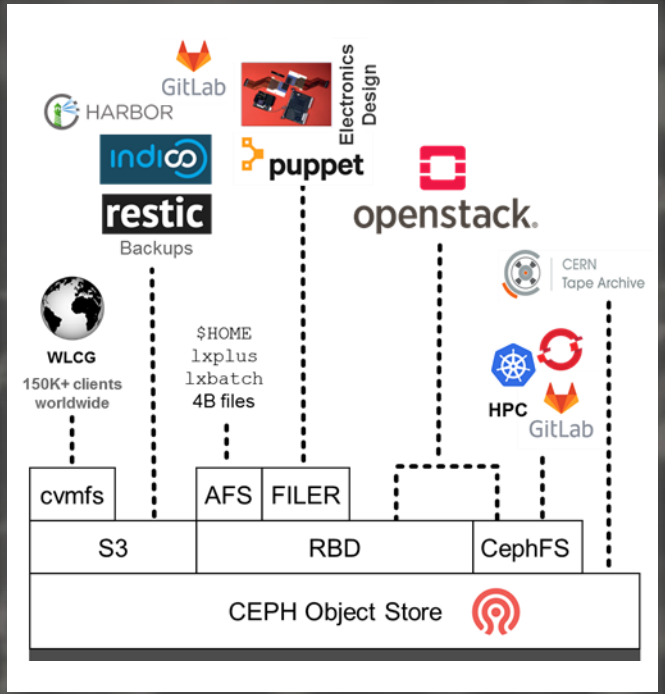
A Brief Service History

12.6 PB HDD
3.6 PB NVMe

- **2024: New Data Centre!**

Applications of Ceph at CERN

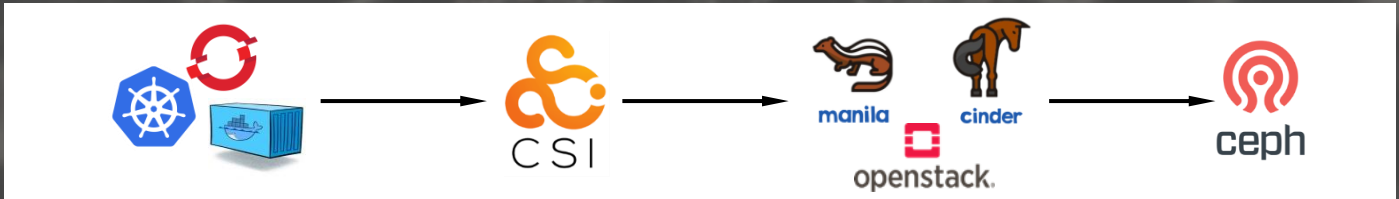
- IT Services:
  - Cloud Infrastructure: OpenStack, K8s, OpenShift
  - Code repositories, Container Registries, GitOps, Agile Infra
  - Monitoring: Open Search, Kafka, Grafana, InfluxDB, Kibana
  - Document Repositories // Web: Indico, Drupal, WordPress
  - Analytics: HTCondor, Slurm, Jupyter Notebooks, Apache Spark
- Other Storage:
  - NFS Filers, AFS, CVMFS, CERN Tape Archive
- Physics Experiments and End-Users:
  - Accelerator Complex Monitoring
  - Microelectronics Design
  - Engineering and Beams



Applications of Ceph at CERN

Provisioning of Ceph Storage to Users

- Self-Service administration of IaaS: Storage, Compute, Network, …
- OpenStack is the entry point for compute and storage resources:
  - Cinder volumes and Glance images on Ceph RBD
  - Keystone as vault for Object Storage keypairs
  - Manila FileShares on CephFS
- Container orchestrators build on top of OpenStack:
  - Container Storage Interface Drivers for RBD and CephFS
  - Declaration of Storage Classes and PVCs propagates to OpenStack + Ceph




2
Block Storage

Block Storage

- Reliable, flexible, virtualized block storage:
  - First Ceph-based storage entering production, oldest cluster is 11yo and rockin'
  - Different QoS (BW + IOps), Media types (HDD/SSD), Availability zones

Volume Type	QoS	Pool Type	Azs
standard	80MB/s, 100 IOps	3x Replicas	3 Zones
io1	120MB/s, 500 IOps		
io2	300MB/s, 1000 IOps	EC 4+2 Full-Flash	-
io3	300MB/s, 5 IO per GB (min 500, max 2000)		
cp1	80MB/s, 100 IOps	3x Replicas	Diesel-backed
cpio1	120MB/s, 500 IOps		

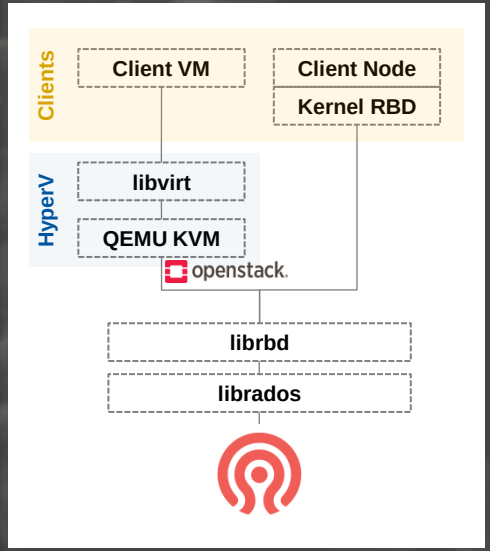
- Block devices for OpenStack VMs:
  - Provisioned through libvirt + QEMU + librdb
  - Each CERN user has a quota of 10 volumes, 250 GB (+20 cores, 20 GB RAM)
  - Tenants for projects can request additional quota + specialized types

Block Storage

- Backend to build other Storage services on top:
  - Virtualization of AFS Disks
    - ✓ Currently biggest single consumer of RBD
  - NFS *“Filers”*
    - ✓ NFS exports of RBD with ZFS on top

- Recent addition of kernel-RBD:
  - Makes Ceph RBD usable by bare-metal nodes
  - Allows for mapping RBD images as devices
  - Client isolation with namespaces and cephx keys

  - Cannot throttle clients (OpenStack Cinder does)
  - Nothing prevents mapping an image on multiple nodes

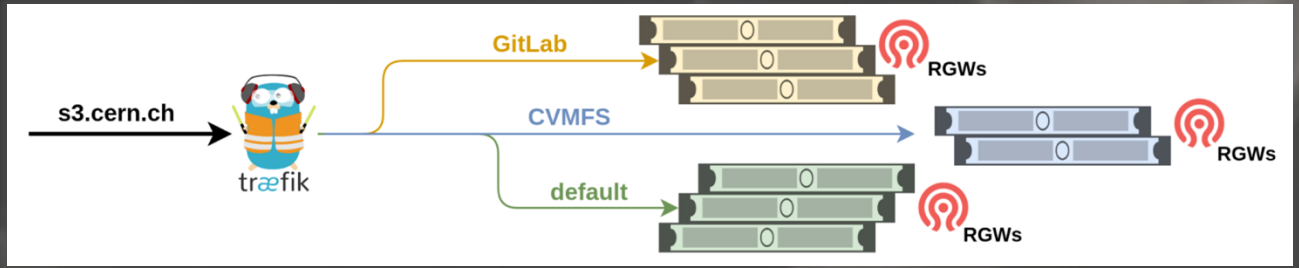


3

Object Storage

Object Storage

- Main production cluster: s3.cern.ch
  - 4+2 EC for data, 3x replicas for Bucket Indices
  - Exposed via 10 load-balanced IPs (round-robin DNS) with Traefik frontend
  - 16 active RadosGWs clustered into groups of users/apps



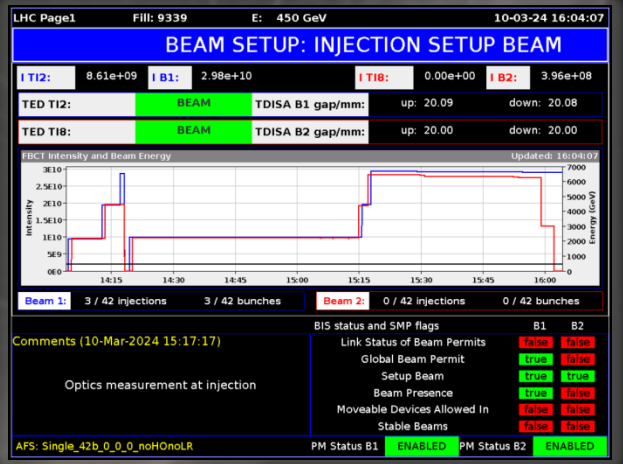
- Second S3 cluster for backups (~5 Km away):
  - ~2000 OSDs, 25 PiB raw (4+2 EC)
  - Backup for File Systems (CephFS, CERNBox, …) via cback, s3-to-s3, and RBDs
  - Fully decoupled from s3.cern.ch – Not a 2nd zone

Object Storage: What for

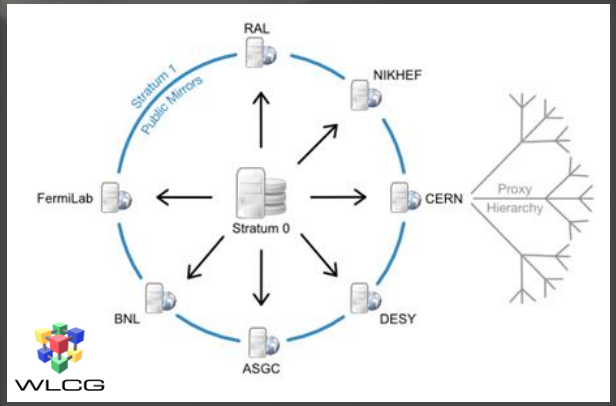
- Cloud native applications:
  - GitLab artifacts, Container Registries, Mattermost, Indico materials, ML workflows, …
  - Prometheus Monitoring

- Accelerator complex monitoring

“LHC Page1”



- Software distribution via CVMFS
  - Repositories of scientific software
  - Distributed over HTTP (and cached), POSIX mounted at **`/cvmfs`** on workers

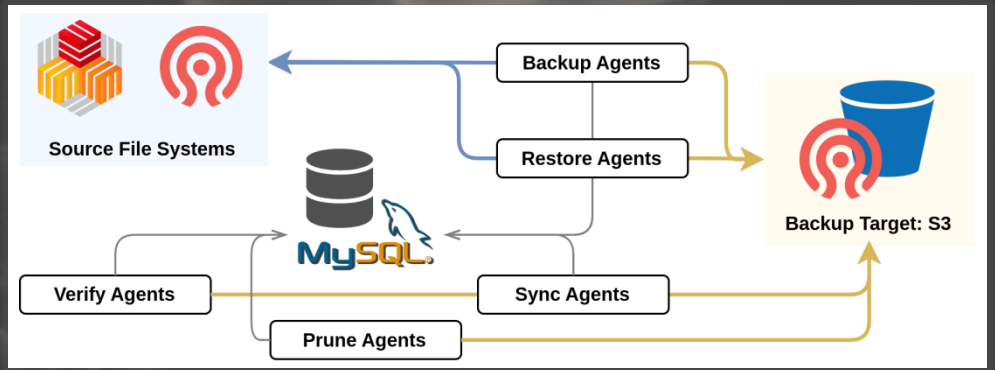


Object Storage: What for

- File system backups with **`cback:`**
  - Backup orchestration tool for File Systems
  - Based on Restic, with the addition of horizontally-scalable agents
  - Centralized queue to keep track of waiting, in-progress, completed jobs
  - Used to backup CERNBox (Sync & Share service) and (some) CephFS

- Source: (virtually) Any mounted file system
- Destination: Ceph S3

- ~40k daily backup jobs
- 1.4+ B files processed per day
- 6.8+ PB backed up to S3




4
Files and Directories

Ceph File System

- First production cluster started operation in 2018:
  - 4.2 PB on HDDs, with metadata on SSDs – 3.5k subvols, 3k+ clients, 350+M files
  - 1 FS, 4 active MDS (+ 4 stand-by), no snapshots
  - Explicit pinning of subdirs to an MDS
    (+ a few selected users on dedicated MDS)

- 2nd flash cluster added in 2020:
  - 0.8 PB on SATA SSDs (data + meta) – 300+ subvols, 500+ clients, 220+M files
  - 1 FS, 1 active MDS (maybe going to multi-active in the future), no snapshots

- Other 4 CephFS clusters for diverse use cases:
  - 2x HPC scratch space and working directories (with standby-replay) for MPI clusters
  - 1x DFS replacement (CephFS kernel mount + SMB export, no **`vfs_ceph`**)
  - 1x general purpose, with snapshots

Ceph File System: What For

- Persistent Storage for K8s + OKD:
  - Web-hosting (including **home.cern**!), Jira, TWiki, OpenSearch, CodiMD, …
  - CSI-enabled clusters create/expand/ mount shares by defining k8s resources

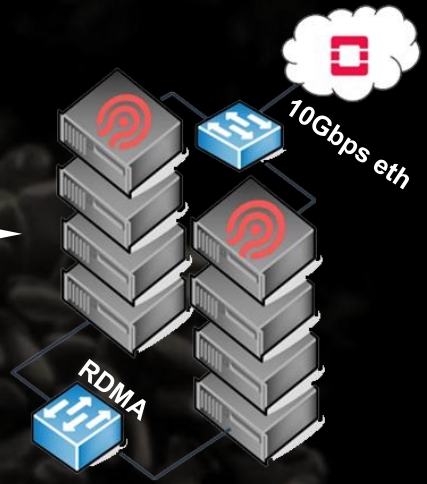
```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: my-manila-vol
spec:
  storageClassName: manila-meyrin-cephfs
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 1Gi
```

- GitLab: On-premise instance for Code Repos, CI/CD, Software Building (rpmci), Pages, Terraform, …

- LinuxSoft / Linux at CERN:
  - Repos to distribute packages to all Linux nodes at CERN – 600k+ RPMs per day
  - Software building through koji (including ceph) – ~500 builds per month

CephFS: A Short HPC Digression

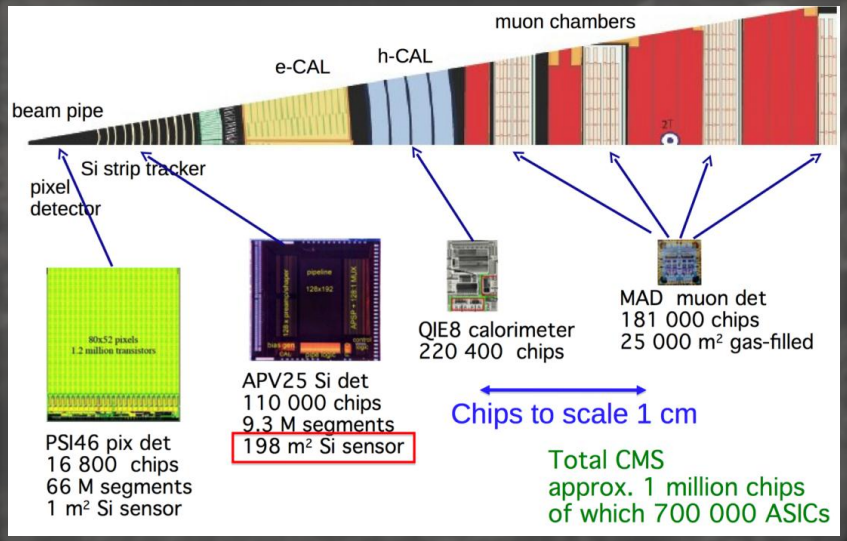
- CERN's "Software Defined HPC":
  - Compute is MPI scheduling with HTCondor + Slurm
  - Storage is CephFS on 2 clusters
    - ✓ General-purpose via OpenStack Manila
    - ✓ Full-flash storage on HPC compute nodes → HyperConverged
    - ✓ Highly parallel, fully-consistent POSIX FS (`LazyIO` is an option)



HyperConverged Setup

- Intel Xeon E5 2630, 128GB
- 4x 960GB Intel S3520 → OSDs
- RDMA + 10 Gbps Ethernet
- CephFS on Quincy 17.2.15
- 1 active MDS (+1 stand-by)
- 3 replicas, rack-aware

Application-Specific
IC Design for CMS Detector




Thank you!