Storage for High Energy Physics

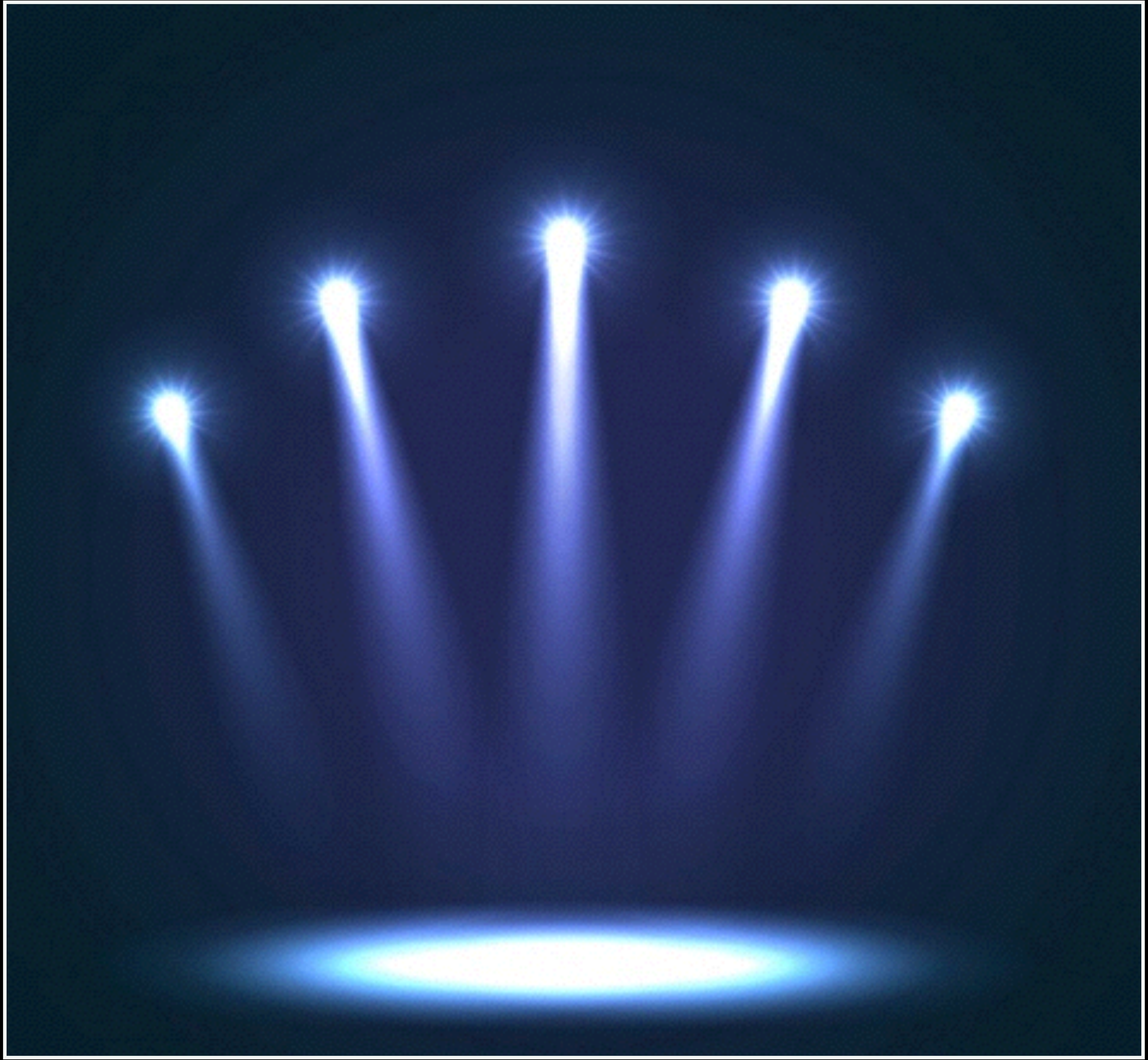
CEPH DAY CERN 2019

Andreas-Joachim Peters

CERN IT Storage Group

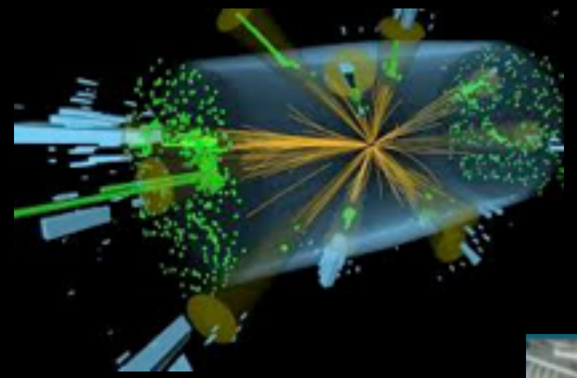
Overview

- Storage in High Energy Physics
  - Past, Present & Future Challenges
  - Data Formats & Access Patterns
  - Storage Software Eco System
  - Remote Access Protocols & Security Mechanism
- Dedicated Storage Systems & Hardware
- Inventory, Summary & Vision



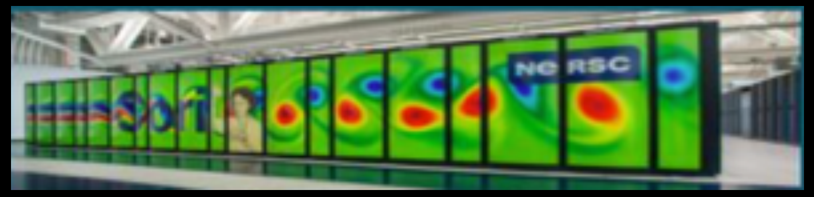
Storage in High Energy Physics

Archival & Backup Storage



Storage for Data Acquisition

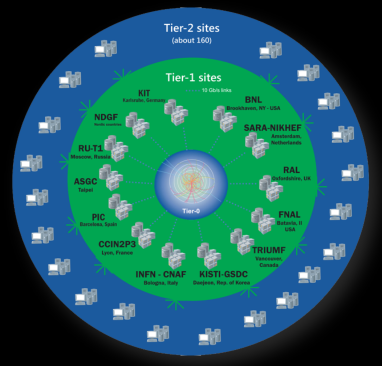
Storage for Home Directories



Storage for HPC

Storage for Applications

Private Cloud Storage



Storage for GRID Computing

Public Cloud Storage

Storage for Software Distribution

Storage for Data Analytics

E=mc²

Storage for Physics Analysis

Storage for Sync&Share

What is HEP Storage?

- **modular stack of layered services** on top of **disk & tape** based persistent storage
- very **heterogeneous resource** composed of s.c. GRID & non-GRID resources
    - various storage systems in computer centres
      filesystems, file stores, object stores, tape storage
    - temporary / time-limited
        - HLT/HPC facilities burst buffers/scratch space
        - public cloud R&D CERN Openlab …

Origins of HEP GRID storage architecture

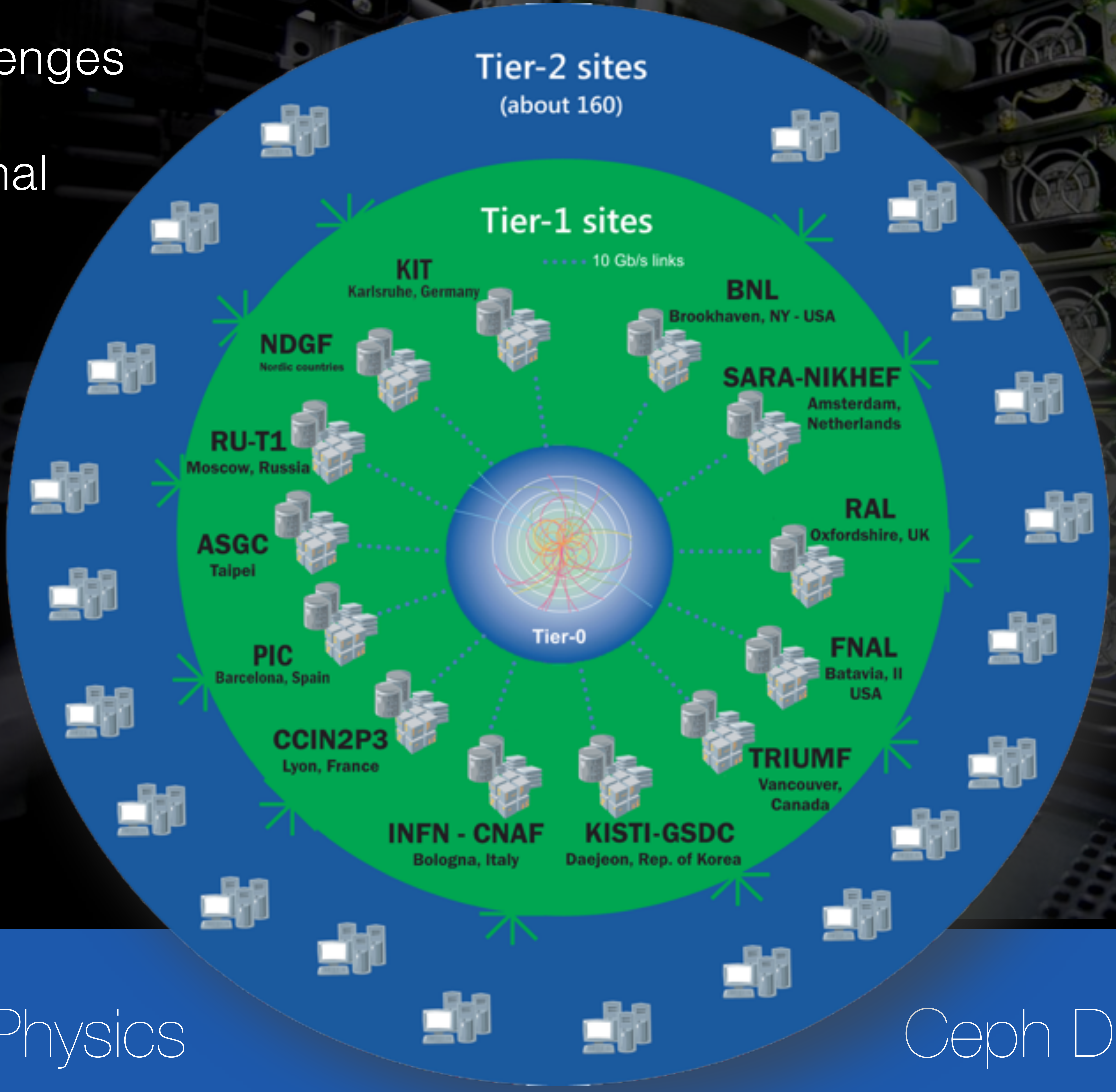
Distributed Computing Model

Since early 2000 HEP embraced
'**the GRID'** model to tackle LHC challenges

- federation of national and international GRID initiatives

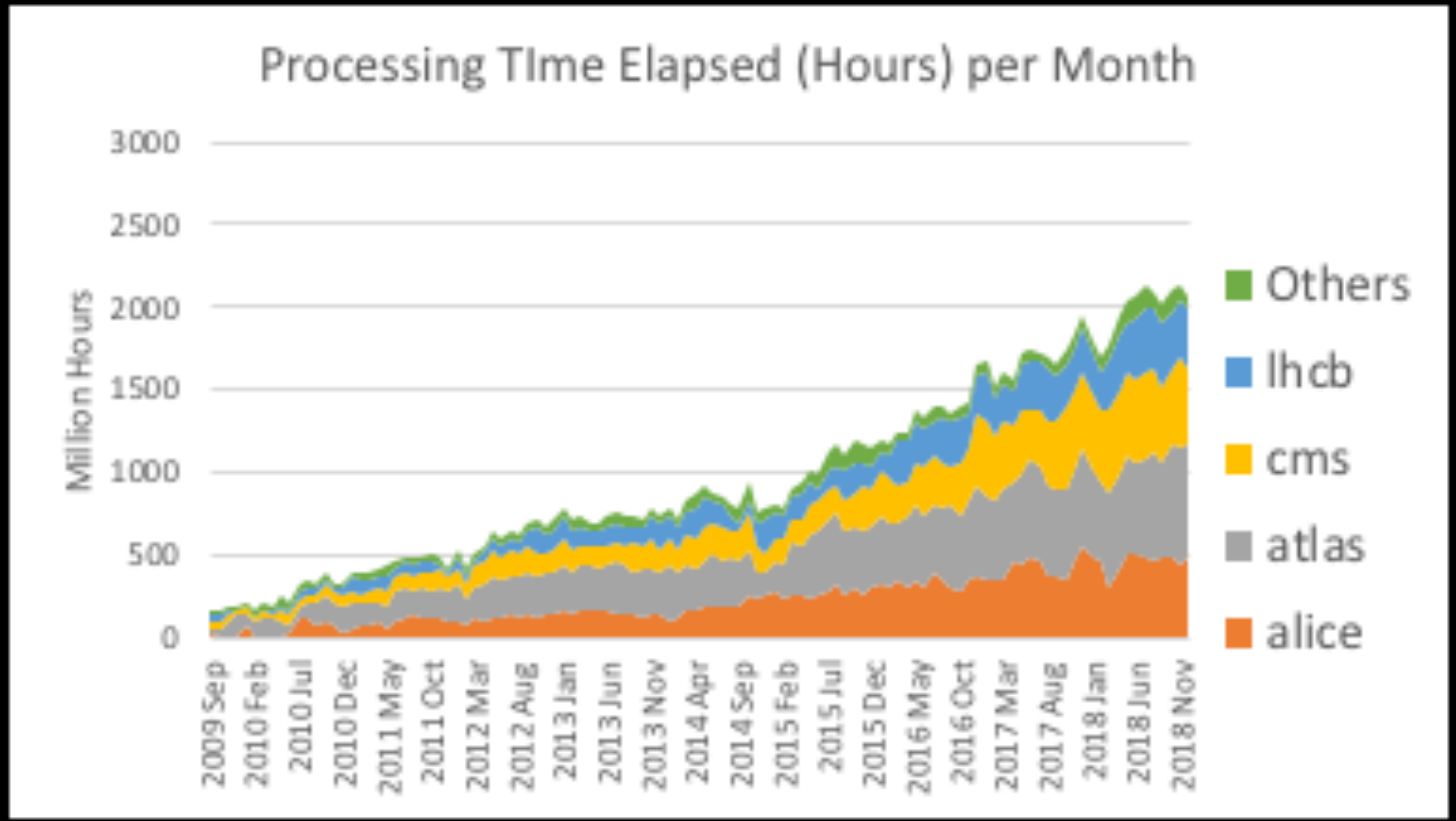
- in line with funding structure

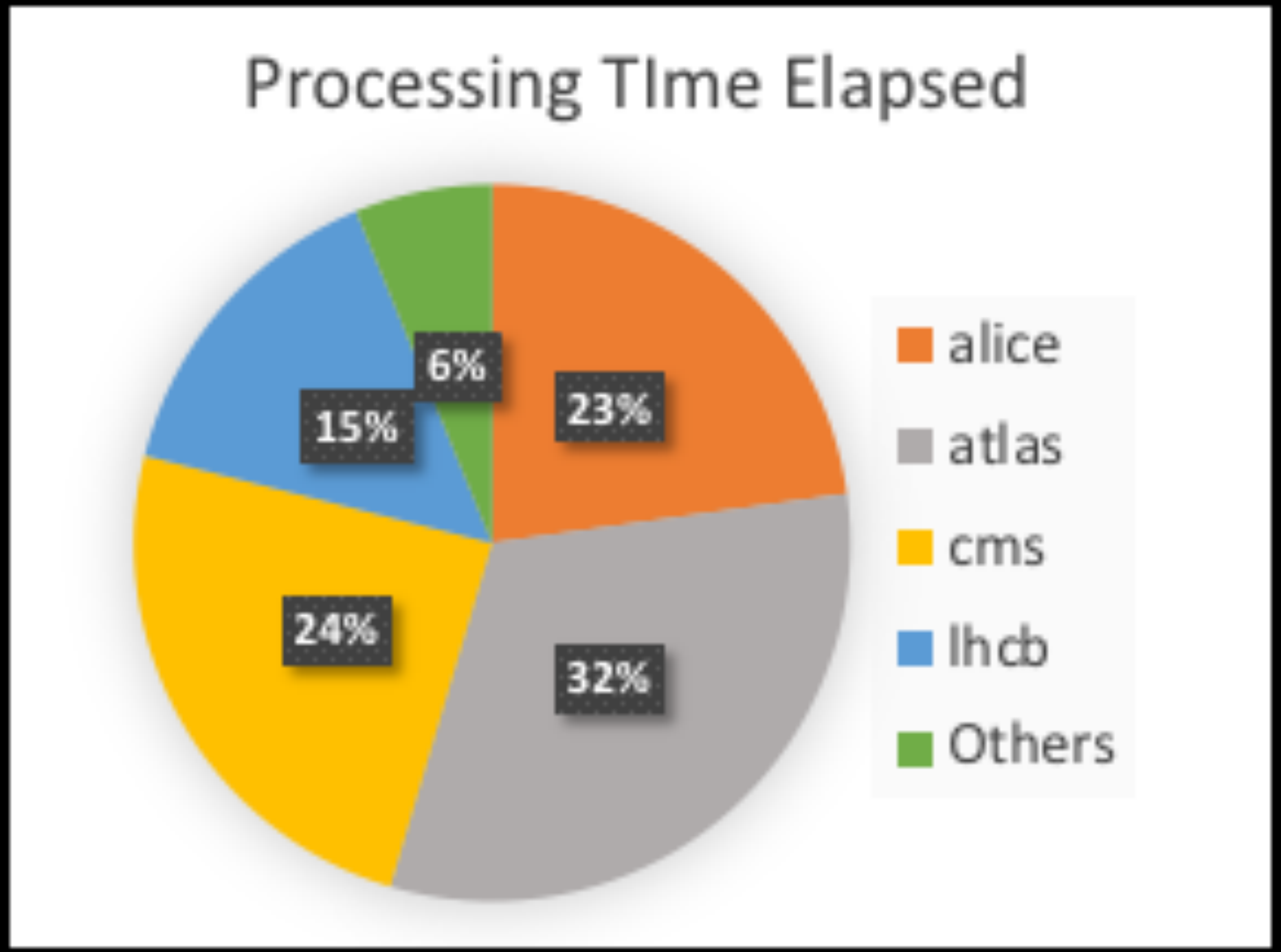
WLCG
Worldwide LHC Computing Grid



HEP Computing Community

- **GRID** resources are shared among many experiments and sciences : Network - CPU - Storage
- **LHC** experiments **consume >> 90%** of the accounted computing capacity
- The remaining part is consumed by 155 other experiments/sciences



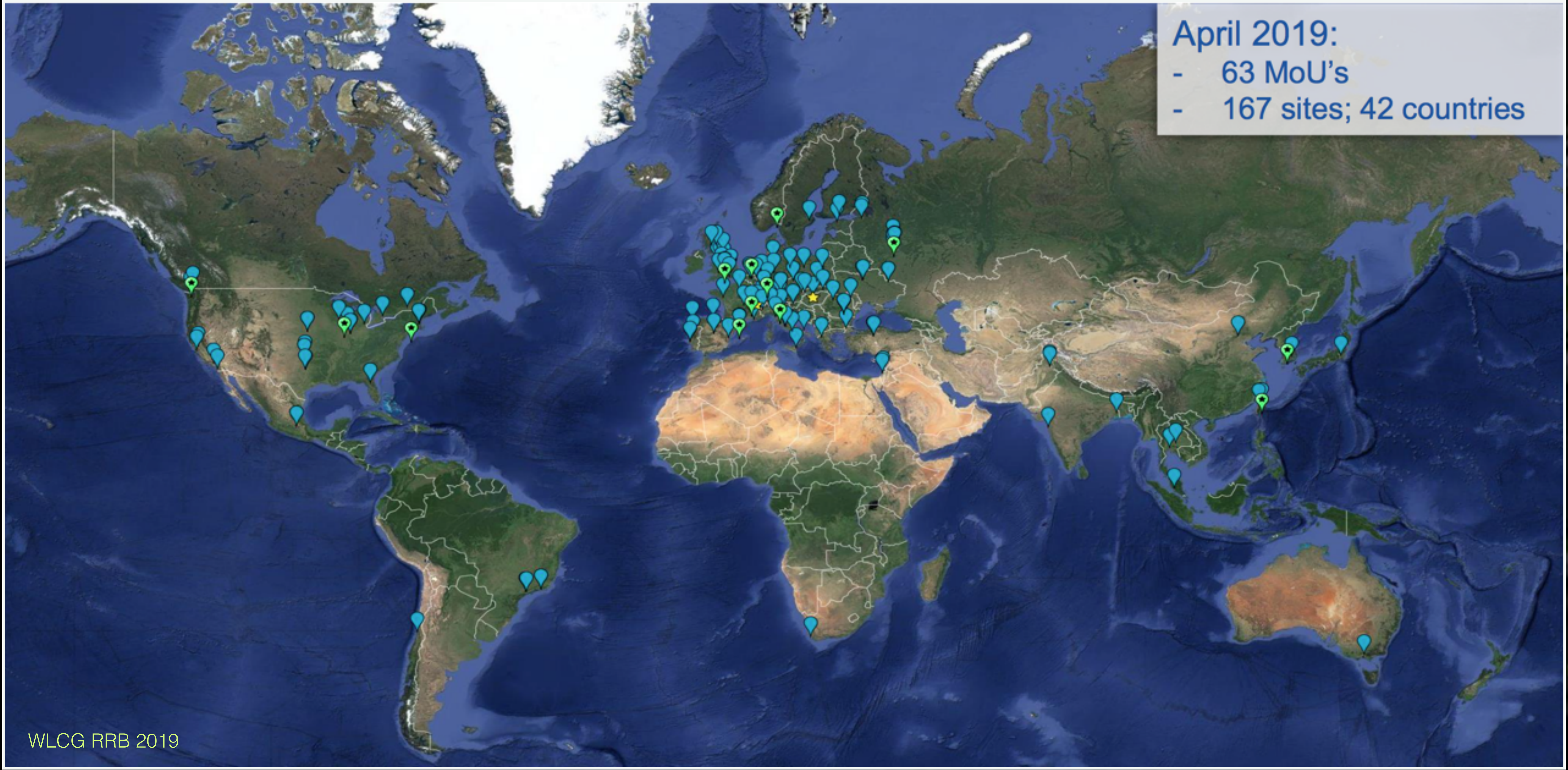


Today **LHC is in a leading position in steering evolution of infrastructure**

S.Campana Granada

Worldwide LHC Computing GRID

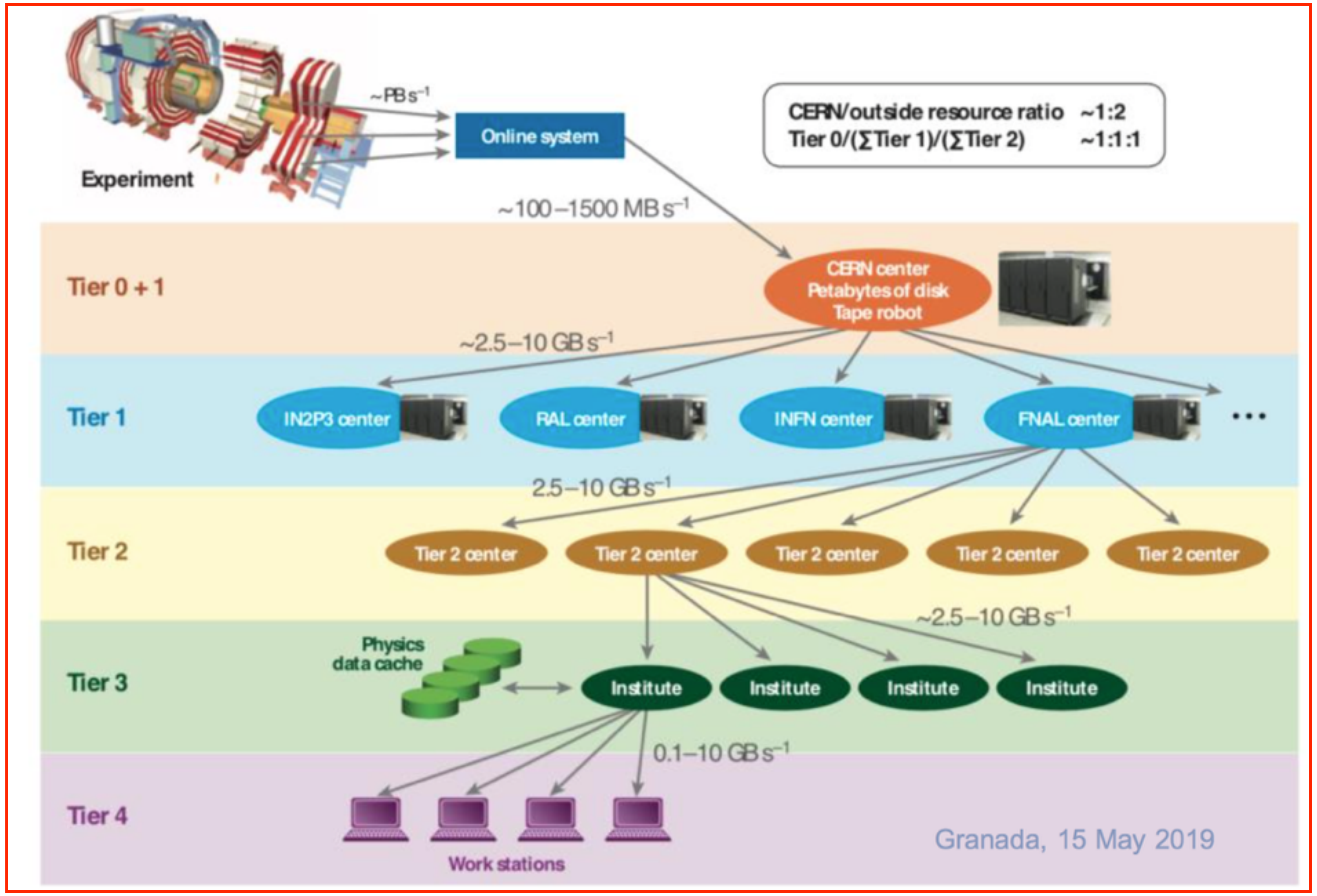
- **WLCG** is a shared resource for ~12.000 physicist
  - 1M cores
  - 1EB storage



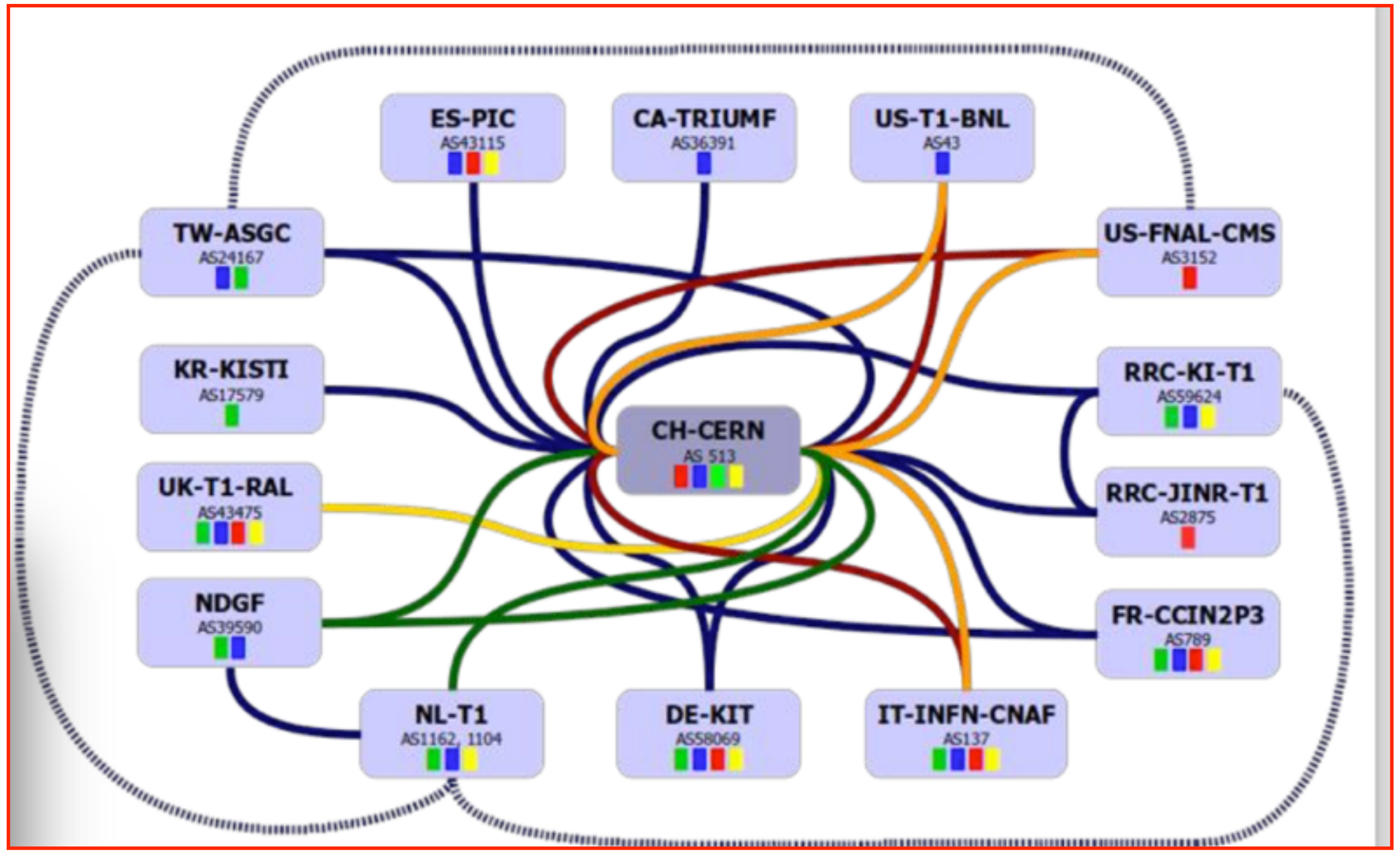
S.Campana Granada

Data Distribution Model

Usage of the **LHC O**ptical **P**rivate **N**etwork



LHCOPN



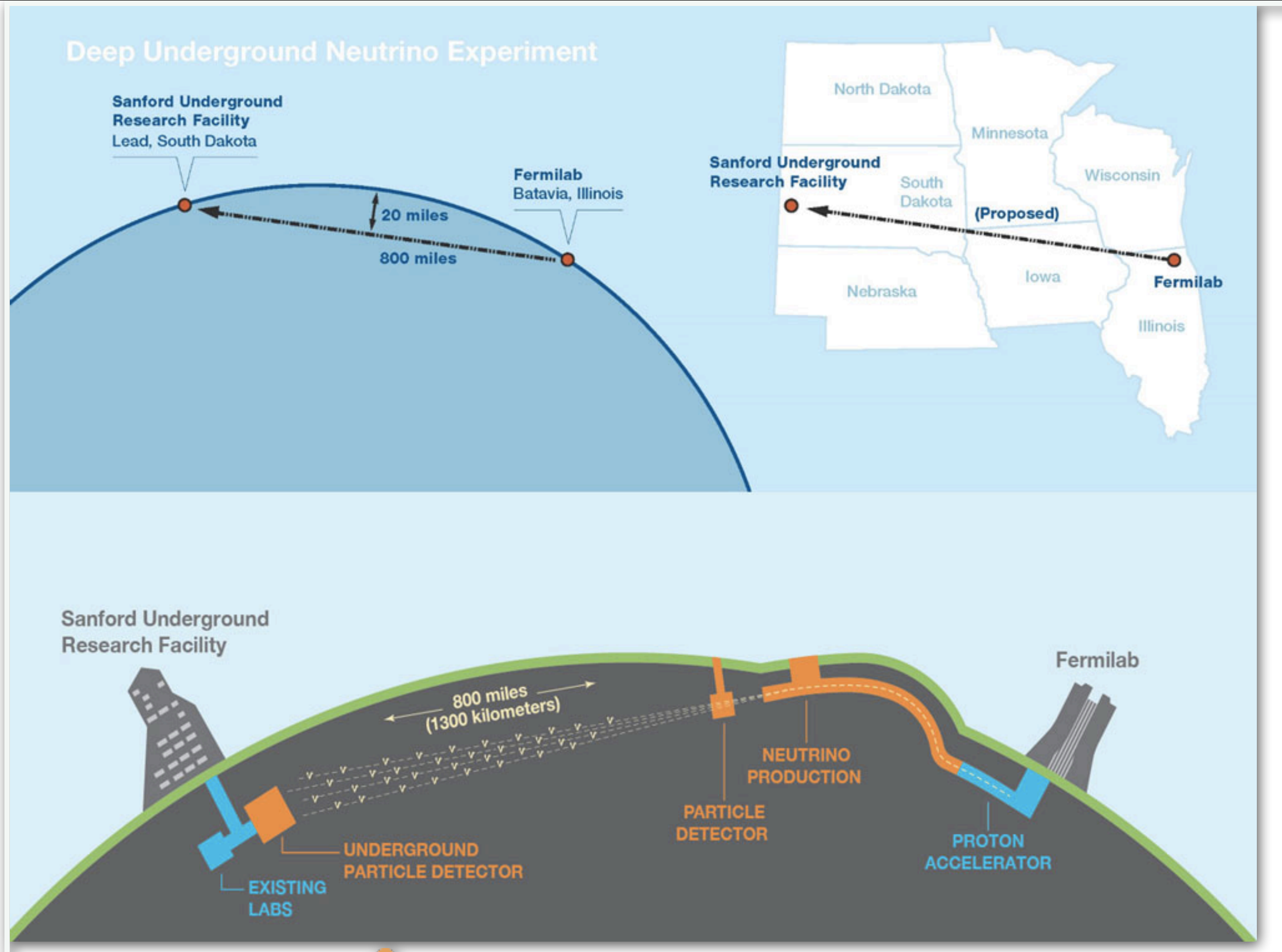
Data Distribution in WLCG

- global transfer rates exceeding regulary **60 GB/s**
- **830 PB** and 1.1B files transferred until end of Run 2
- main **challenge** is to have the **useful data close** to available computing resources
  match storage/compute/network

Running jobs: 365644
Active CPU cores: 807139
Transfer rate: 21.54 GiB/sec

The HEP Data Volume Challenge

LHC future: **ten times more data** … but there is more than just LHC …

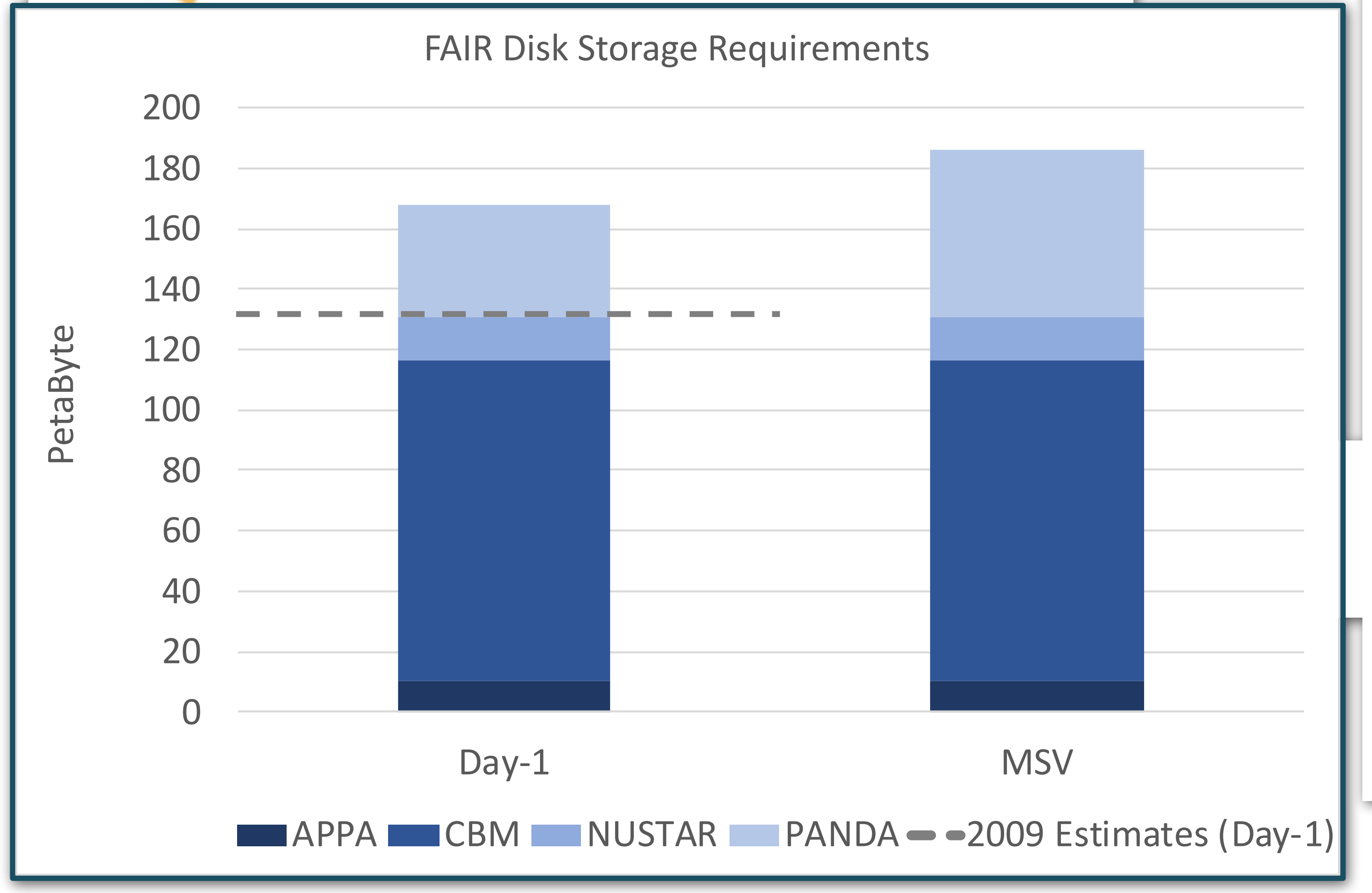


DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

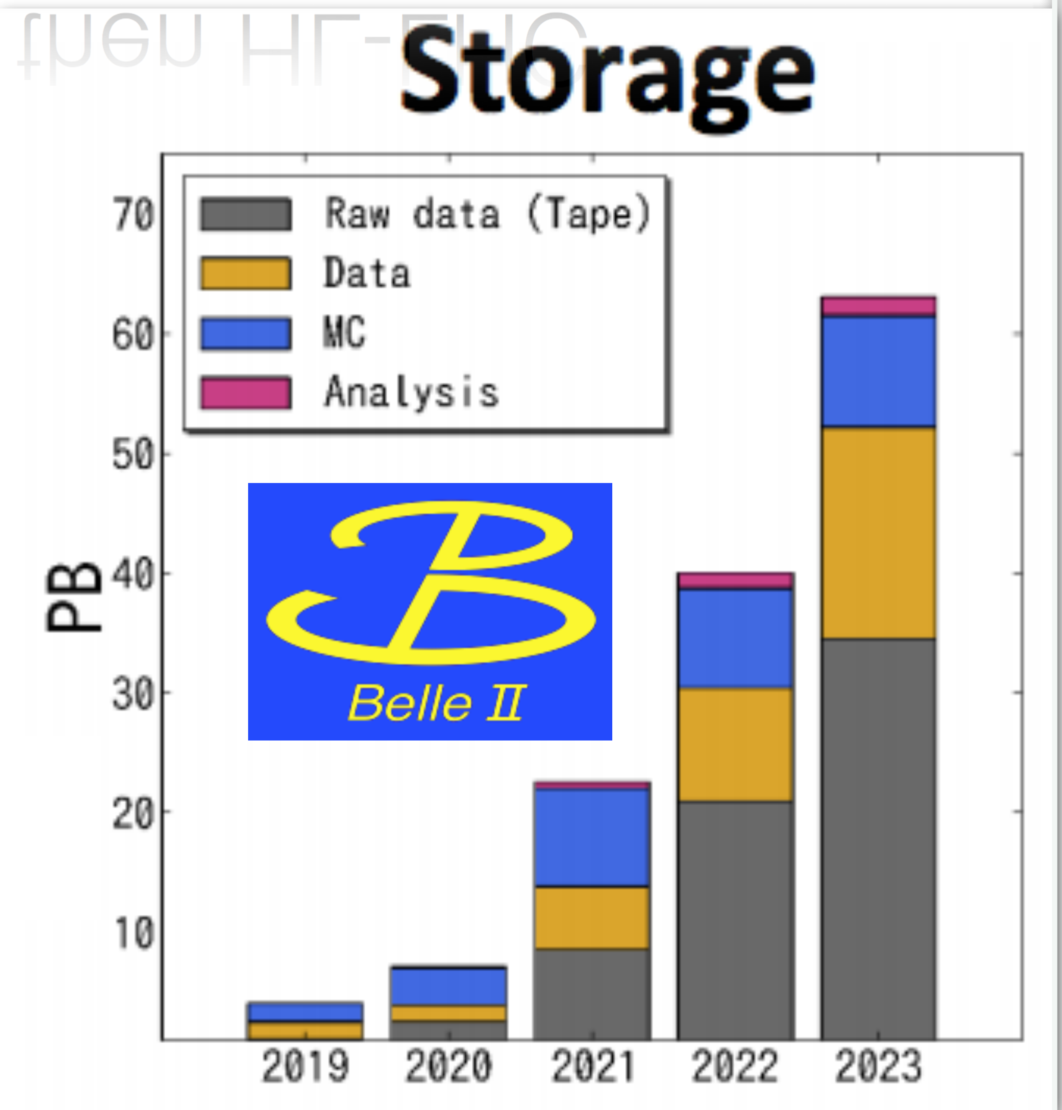
DUNE foresees to produce ~70PB/year in the mid 2020s

Several experiment will require relatively large amount of resources in the future.
Several factors less then HL-LHC

FAIR Comparable data volume to LHC



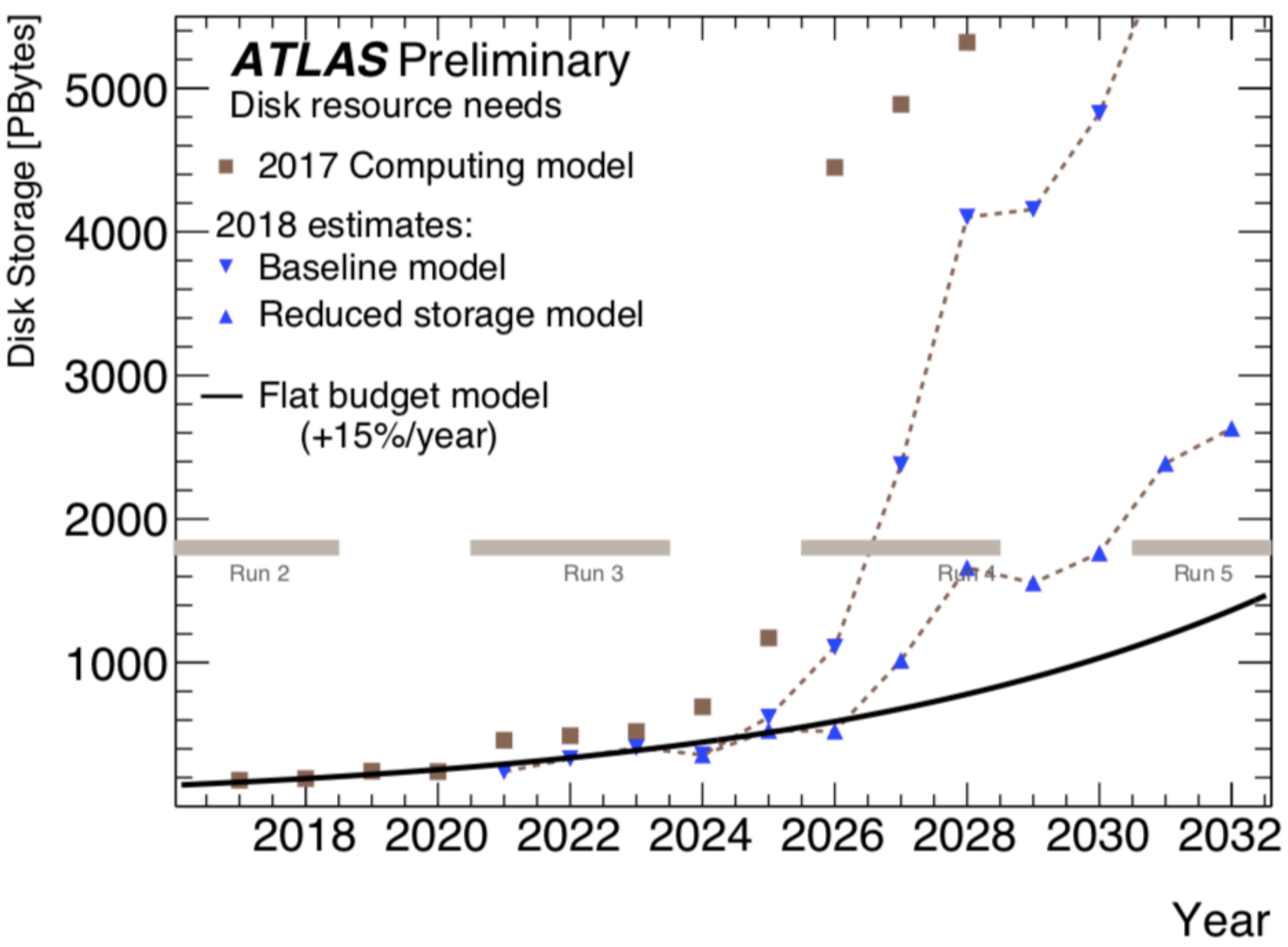
J. Eschke @ ESCAPE kick-off



Y. Kato @ HOW 2019

S.Campana Granada

The LHC Storage Cost Challenge



LHC Challenge of High Luminosity Run 3/4

Assuming a flat budget, the storage requirements of the future cannot be satisfied anymore by technology evolution

Possible improvements

- **changes** to **computing models**
  - **media shift** towards cheapest media combined with experiment driven storage tiering (data carousel)
- **capacity** increase using erasure encoding and optimised replication strategies
  (but often EC ~ RAID volume)
- **Data lake/cloud** model
  - composed of fewer HA storages and satellite caches
  - more efficient storage management and optimised redundancy using QOS interfaces

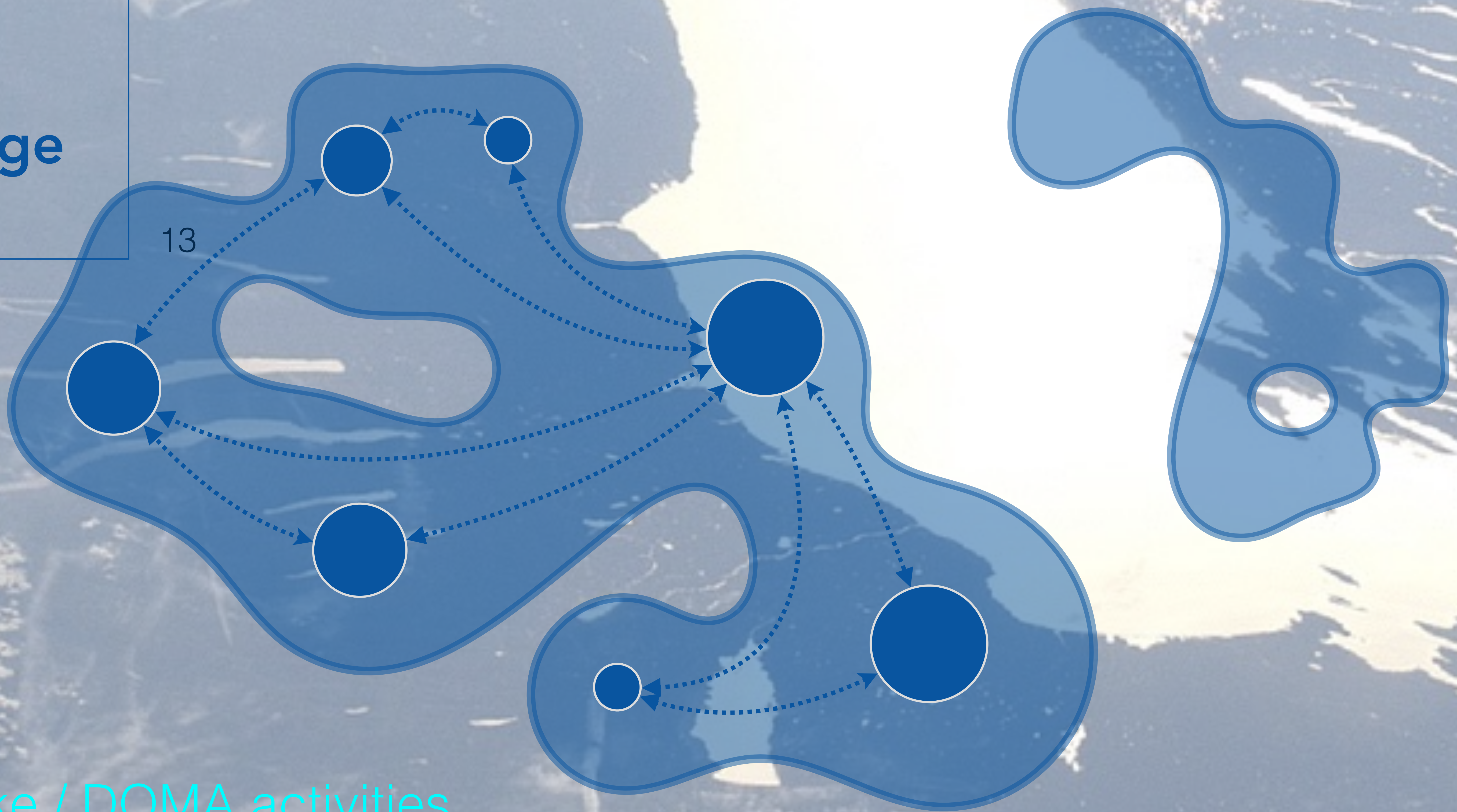
“Data Lake”

exploring geographical distributed storage systems

High-Luminosity LHC
CERN-SKA partenship

- n sites
- k replicas
- k<<n
- latency > 1ms
- caches + persistent storage

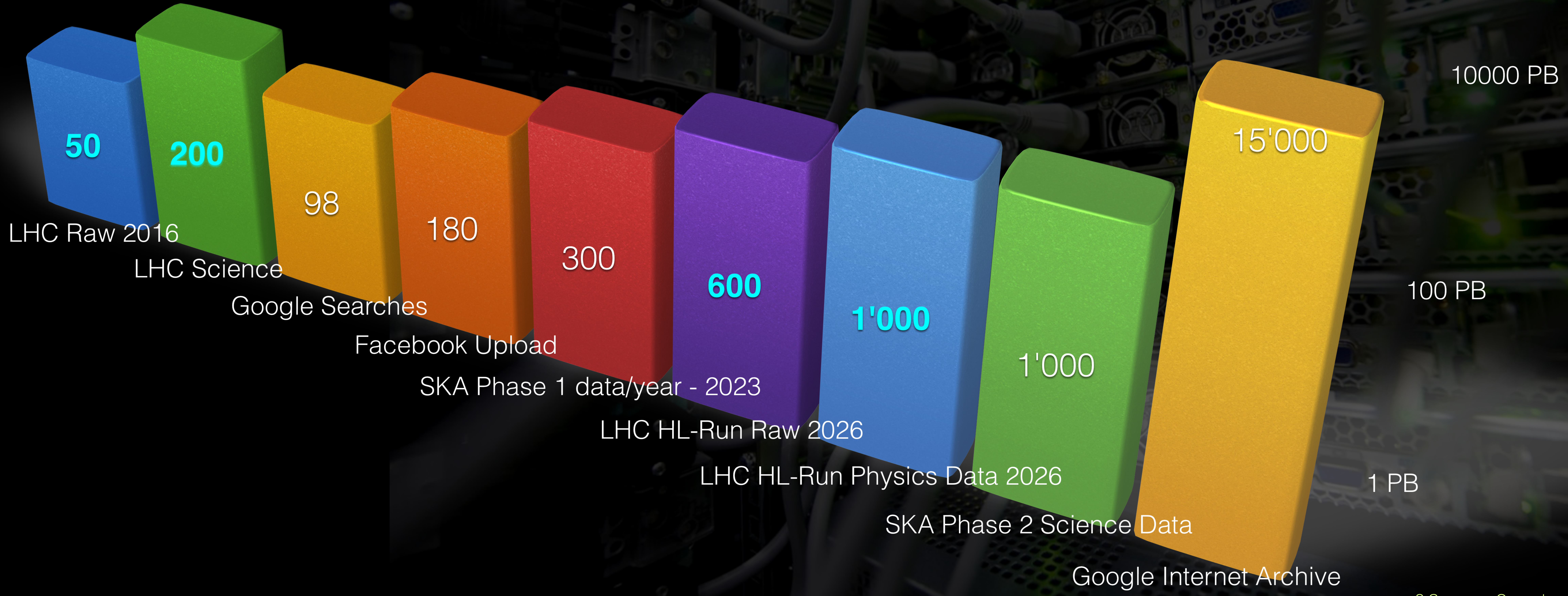
13



[More background to Data Lake / DOMA activities]

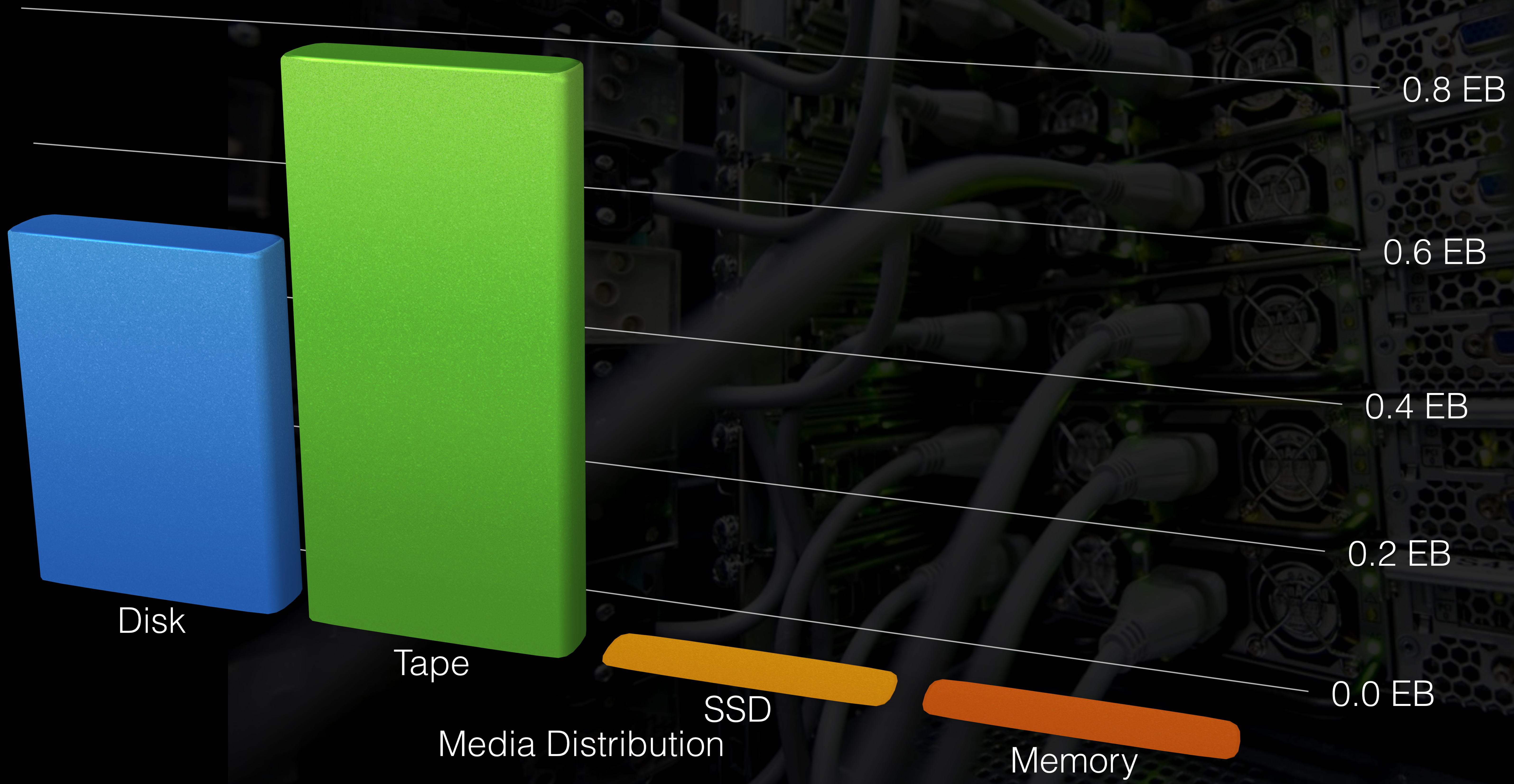
WLCG Report 2019

HEP Data Volume in perspective



S.Campana Granada

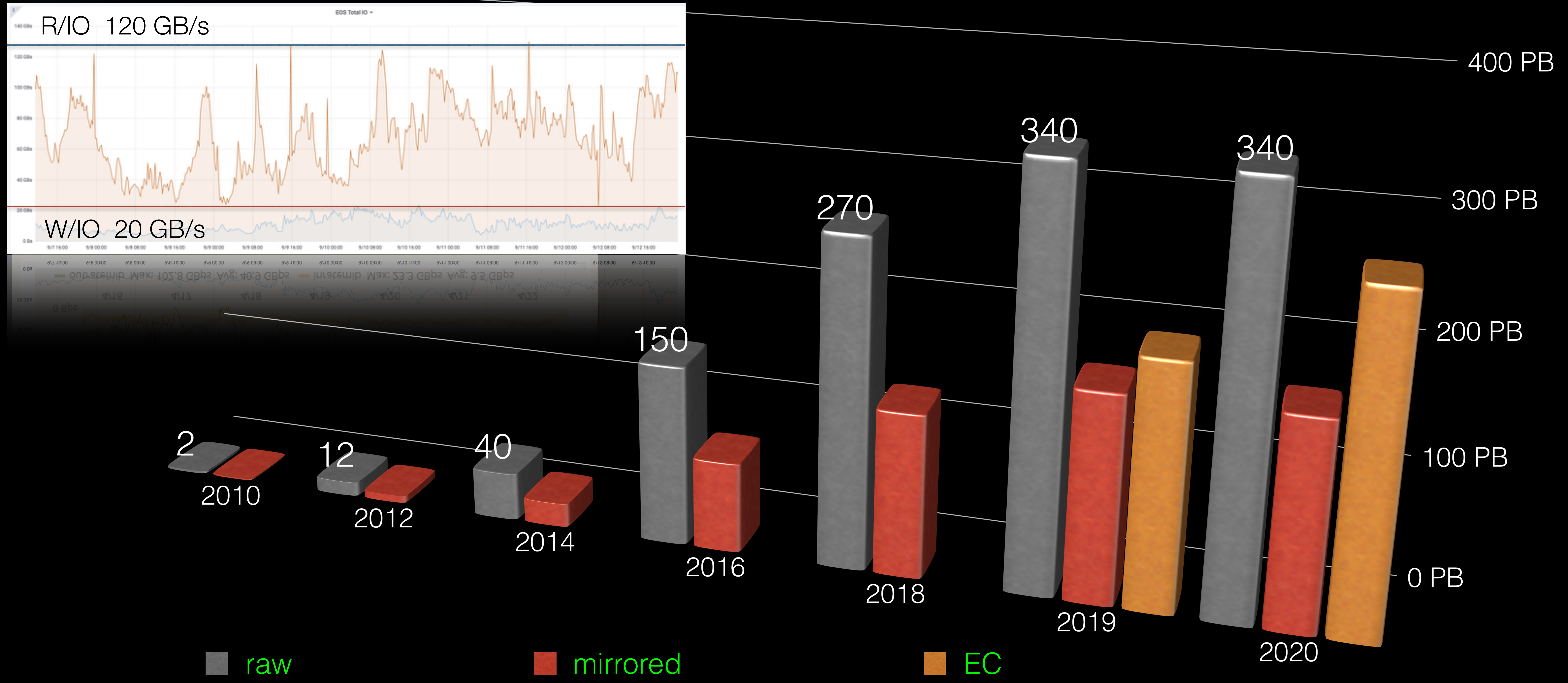
2019 Storage Media Pledges LHC



WLCG Rebus DB

Disk Storage at CERN

EOS



Files Stored 4.92 Bil

Storage Nodes 1500

Hard Disks 60k

Raw Space 340 PB

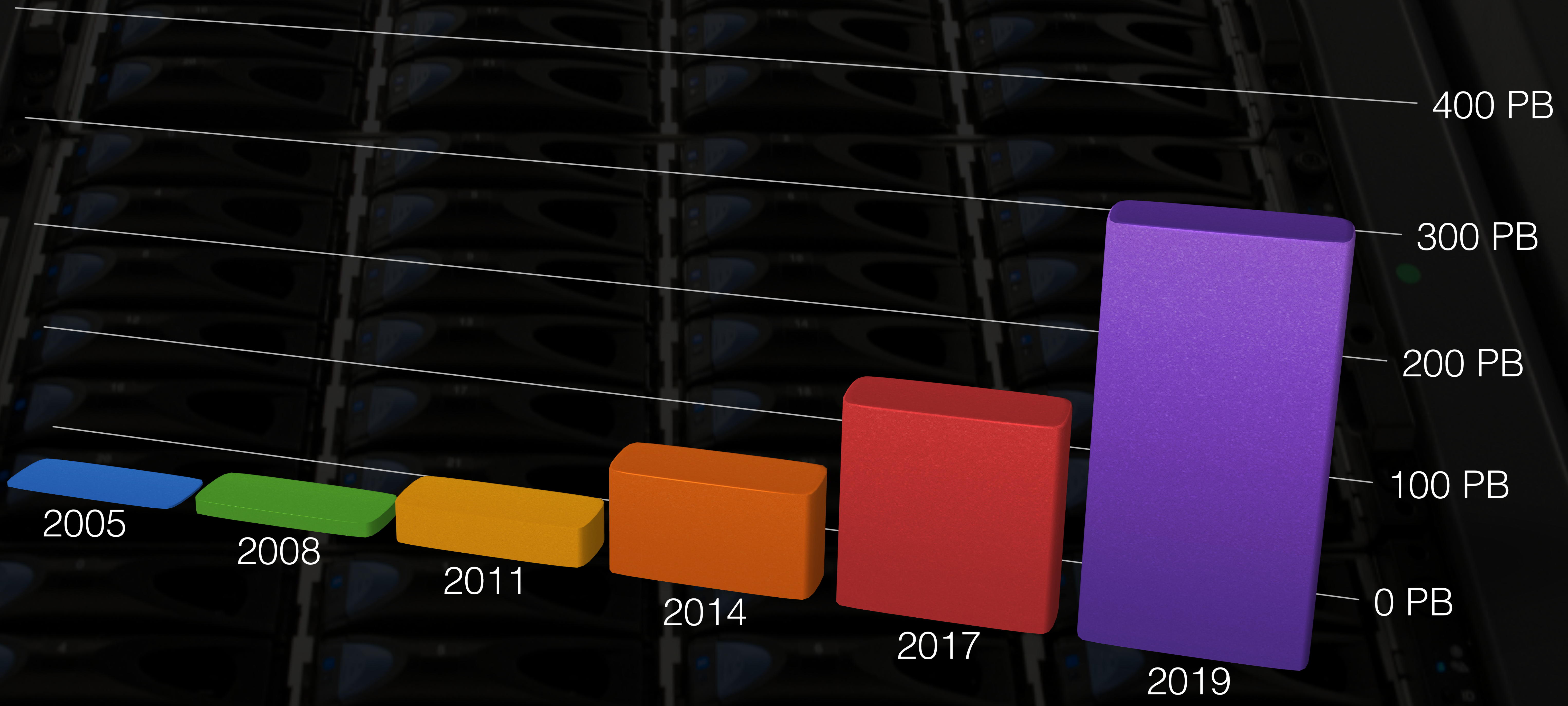
Single Instance 60 PB

IO Streams >100k

IT-ST

Tape Archive at CERN

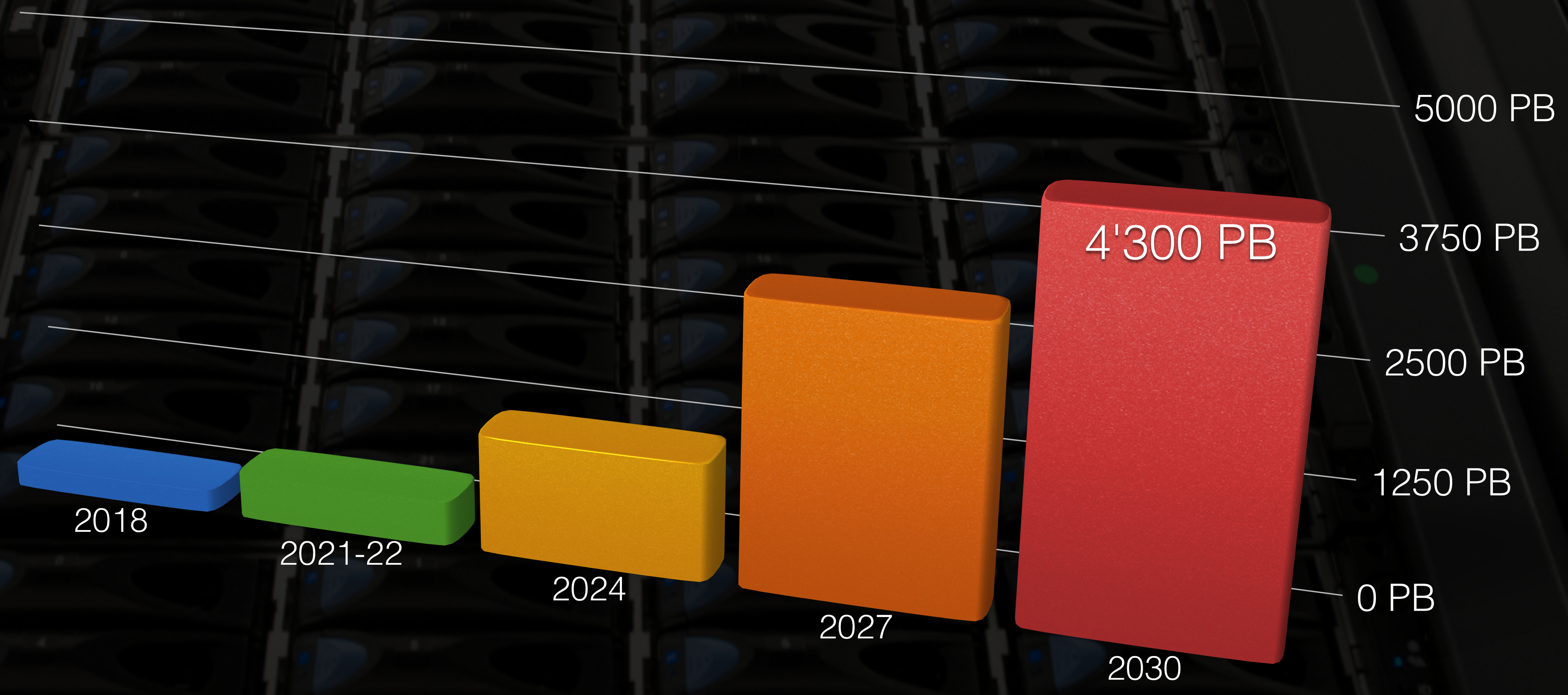
past to presence



G. Cancio - IT-ST

Tape Archive at CERN

prediction

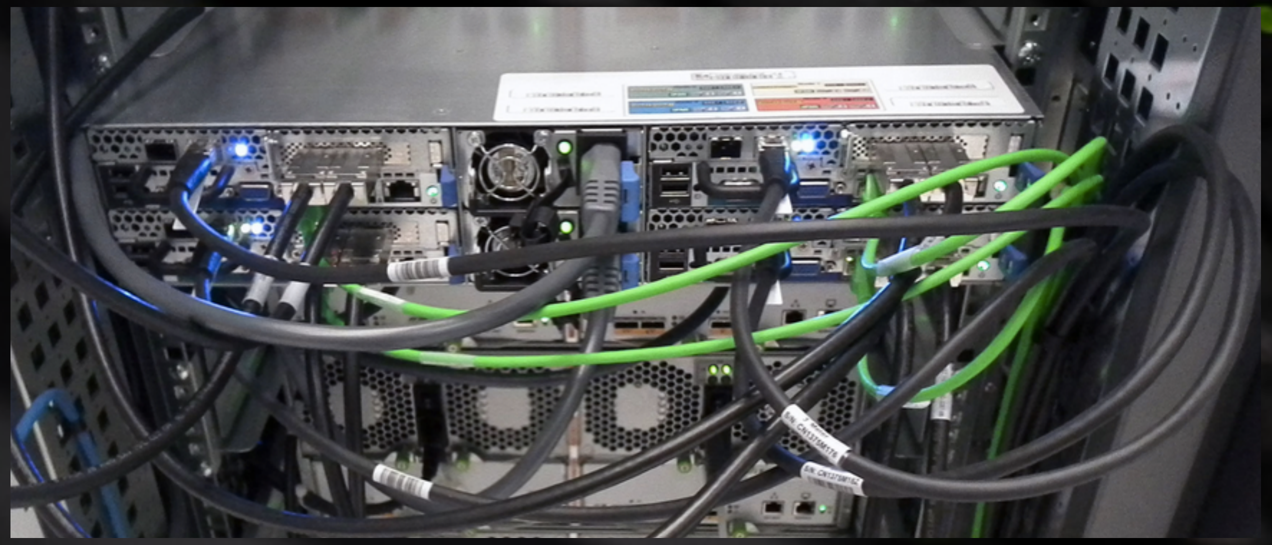


G. Cancio - IT-ST

Storage Hardware

Physics Data Storage Hardware CERN

- Profiting from ***economy of scale***
  - minimise price per GB
- System Unit:
  - 8 physical cores (16 virtual) 64-128GB RAM
  - disk-tray of 24x 4-6-10-12TB HDDs



- Running different generations
  - 2 trays per system unit - 48 disks
  - next gen.: **2 dense trays** per system unit - 120 disks
  - **4 trays per system unit - 96 disks**
  - 8 trays per system unit - 192 disks

up to 2.4 PB per disk server

Tapeless Archive Project

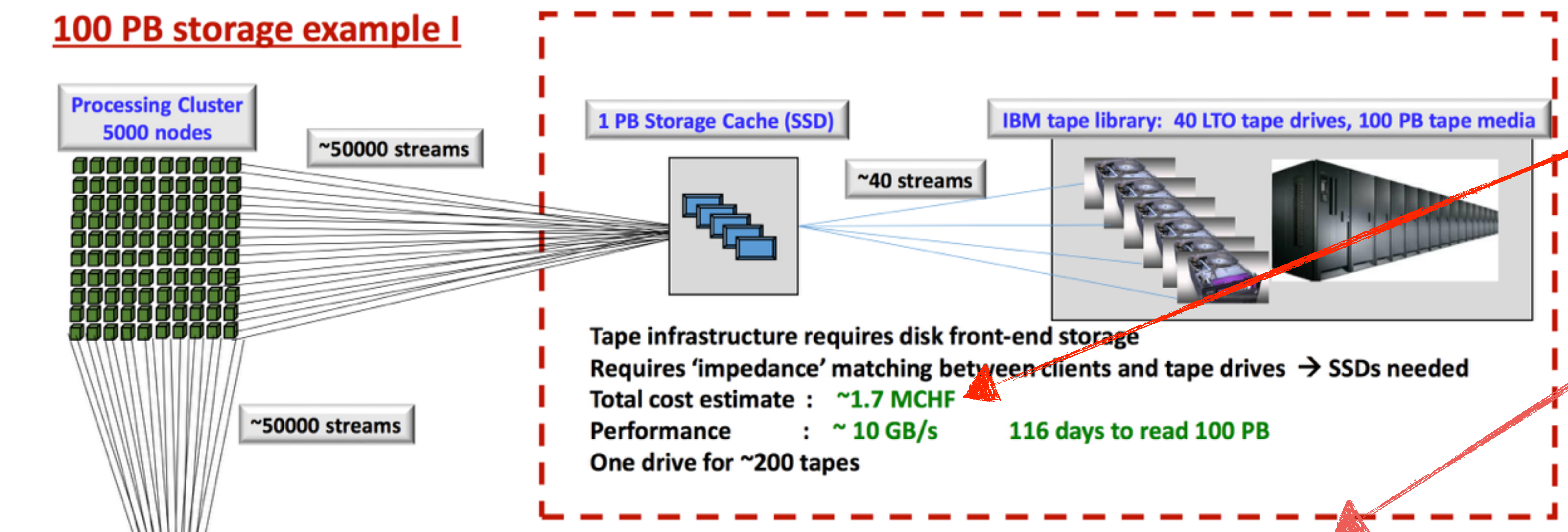
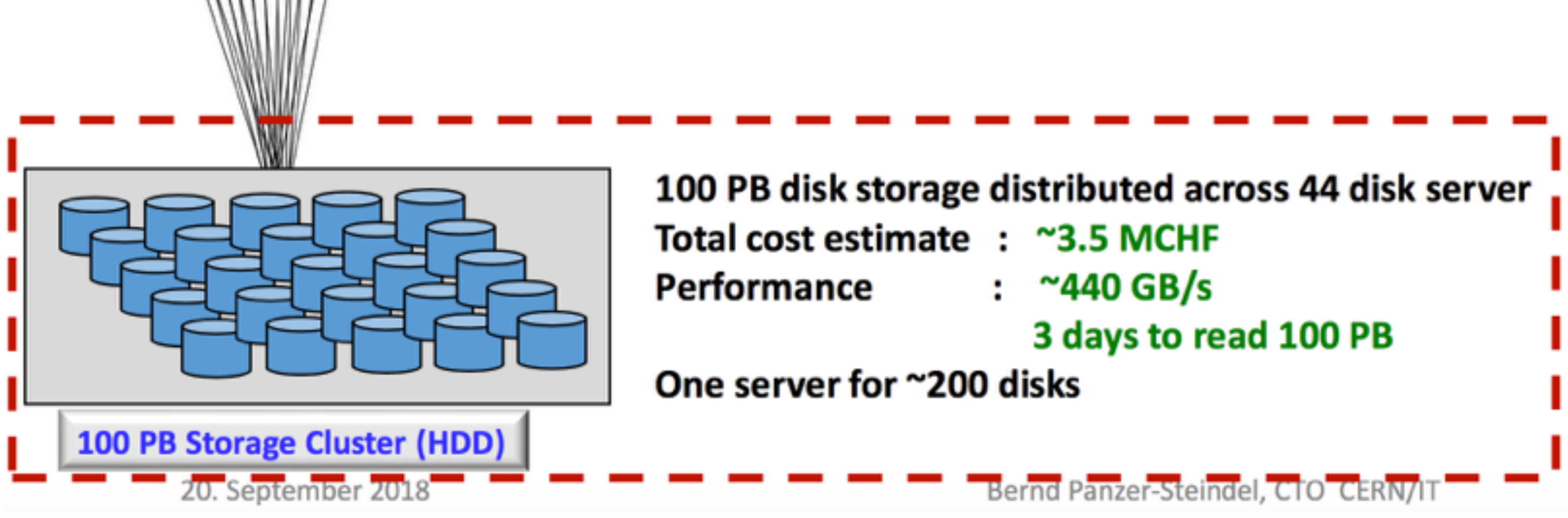
@ KISTI / S. Korea

Goal:
low-cost **disk-only erasure encoded archival storage** requiring deletion/integrity safety features

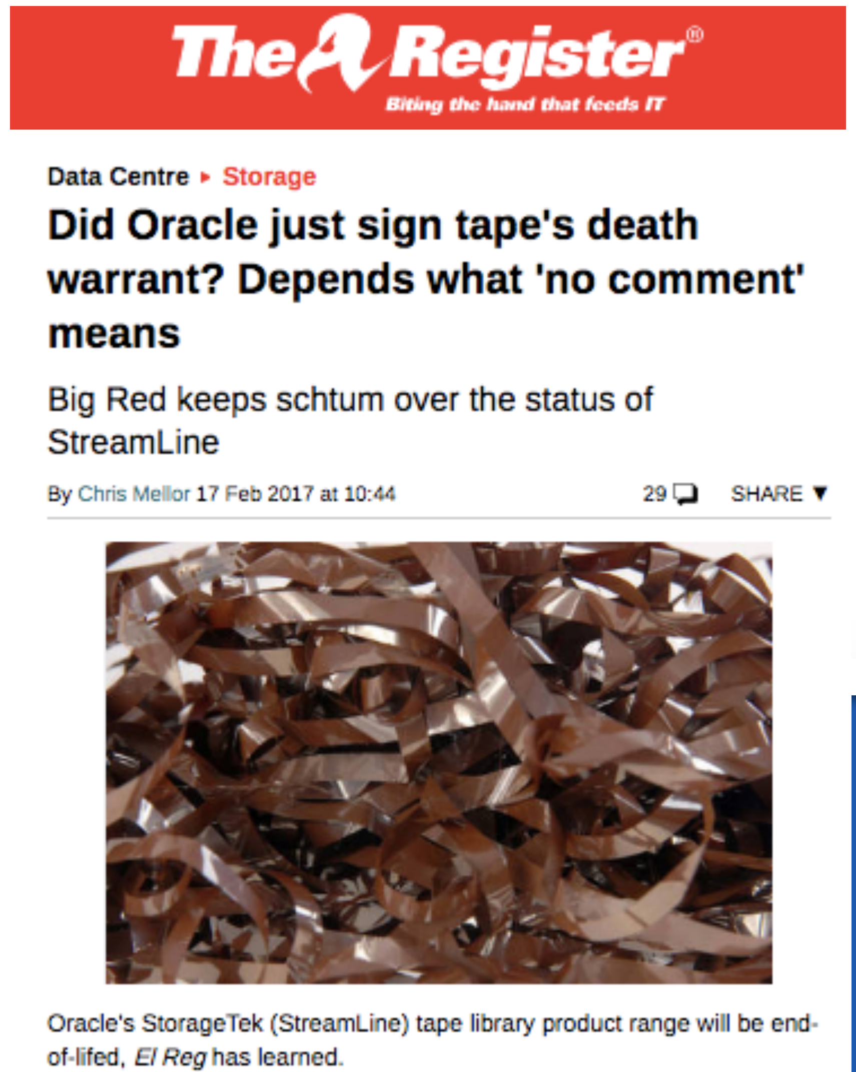
10 nodes - 15 PB usable space

Tape: Lower Cost & Data Safety

B. Panzer-Steindel CERN IT CTO

- still (50%) cheaper
- physical deletion is slow
- however
  - single vendor problem (enterprise)
  - media shipments shrinking


The Register
Biting the hand that feeds IT

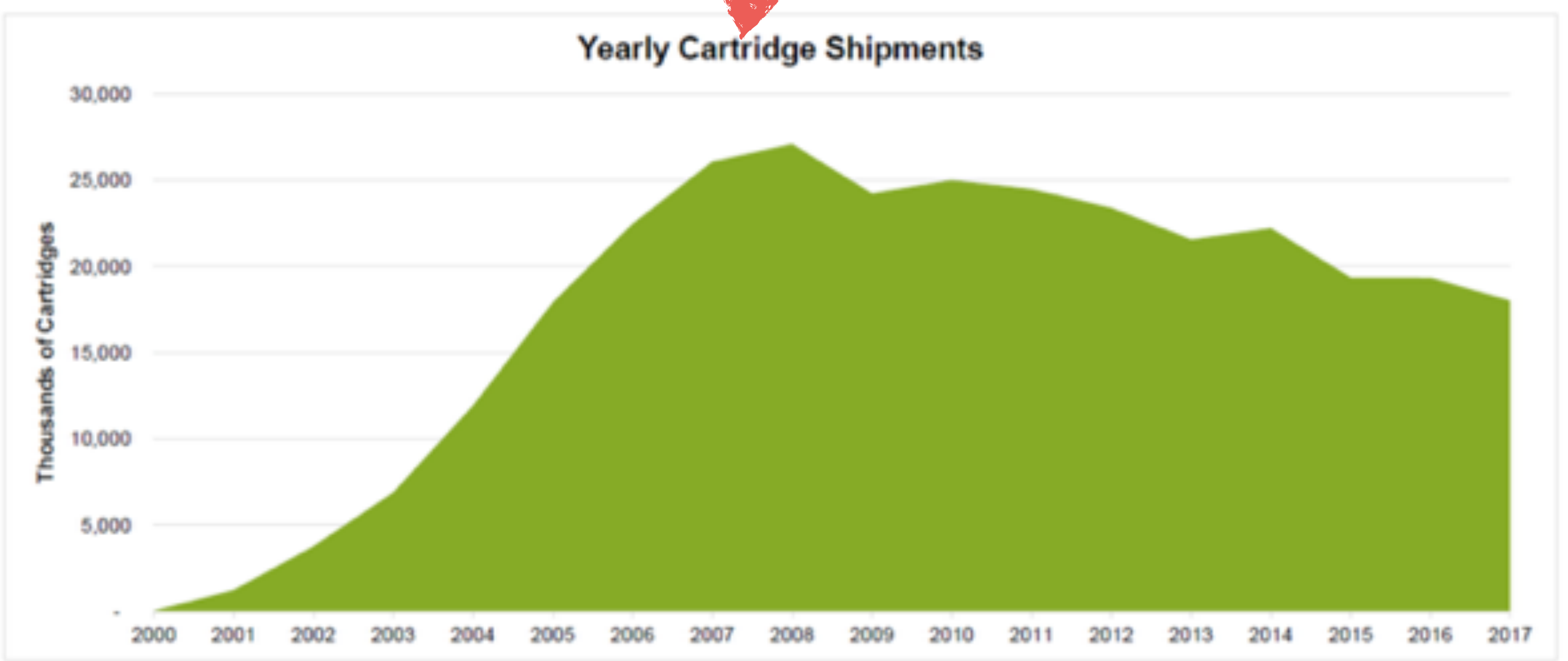
Data Centre ▸ Storage

Did Oracle just sign tape's death warrant? Depends what 'no comment' means

Big Red keeps schtum over the status of StreamLine

By Chris Mellor 17 Feb 2017 at 10:44 29 SHARE ▼

Oracle's StorageTek (StreamLine) tape library product range will be end-of-lifed, *El Reg* has learned.



Storage Software Development Projects in HEP

https://rucio.web.cern.ch

Data Management

File Transfer

https://fts.web.cern.ch

EOS CTA

https://eos.cern.ch

Disk Tape Storage

CS3

Home | Programme | Community | Abstracts | Committees

www.cs3community.org

Cloud Services for Synchronisation and Sharing

28 - 30 January 2019, Roma

Previous Workshops

Krakow 2018 - Amsterdam 2017 - Zurich 2016 - Geneva 2014

HEP Storage Software Development

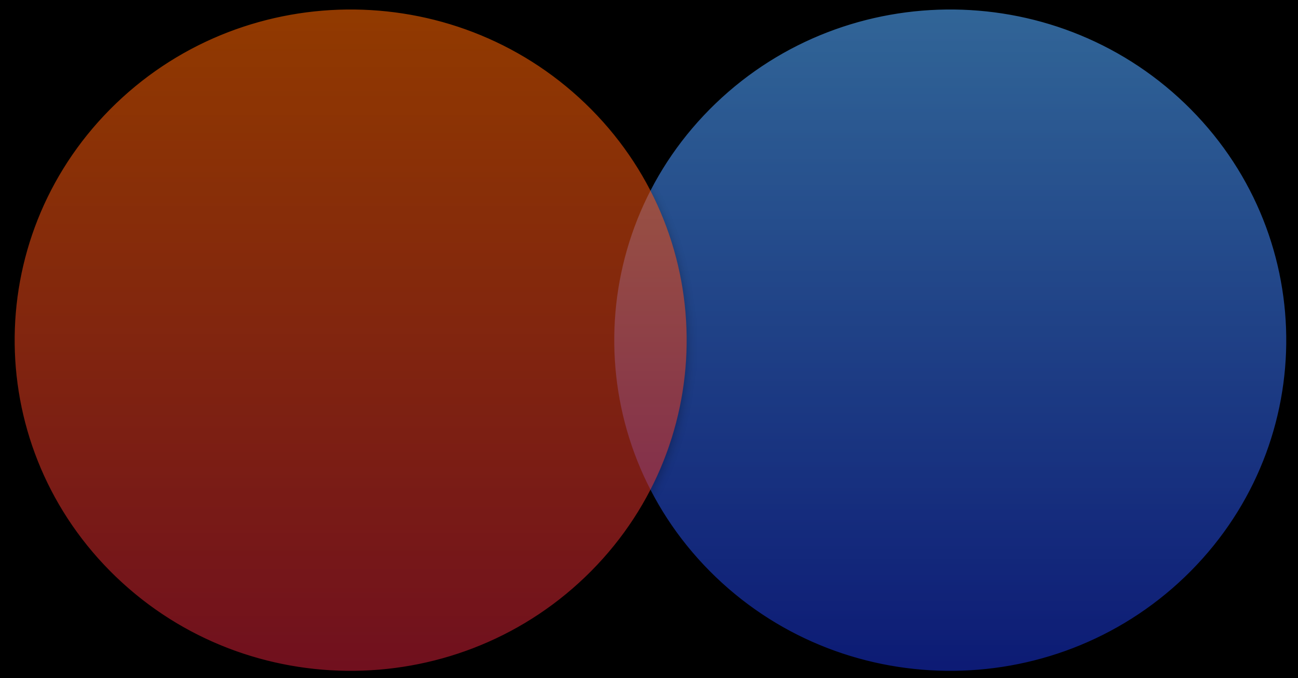
Why is there so much HEP Storage related Software Development?

HEP

- Amga
- AliEn
- Castor
- CERNBOX
- CTA
- DAVIX
- dCache
- Dirac
- Dynafed
- DPM
- EOS
- FTS
- GFAL
- Phedex
- ROOT
- Rucio
- VOMS
- XCache
- XRootD
- …

BigData

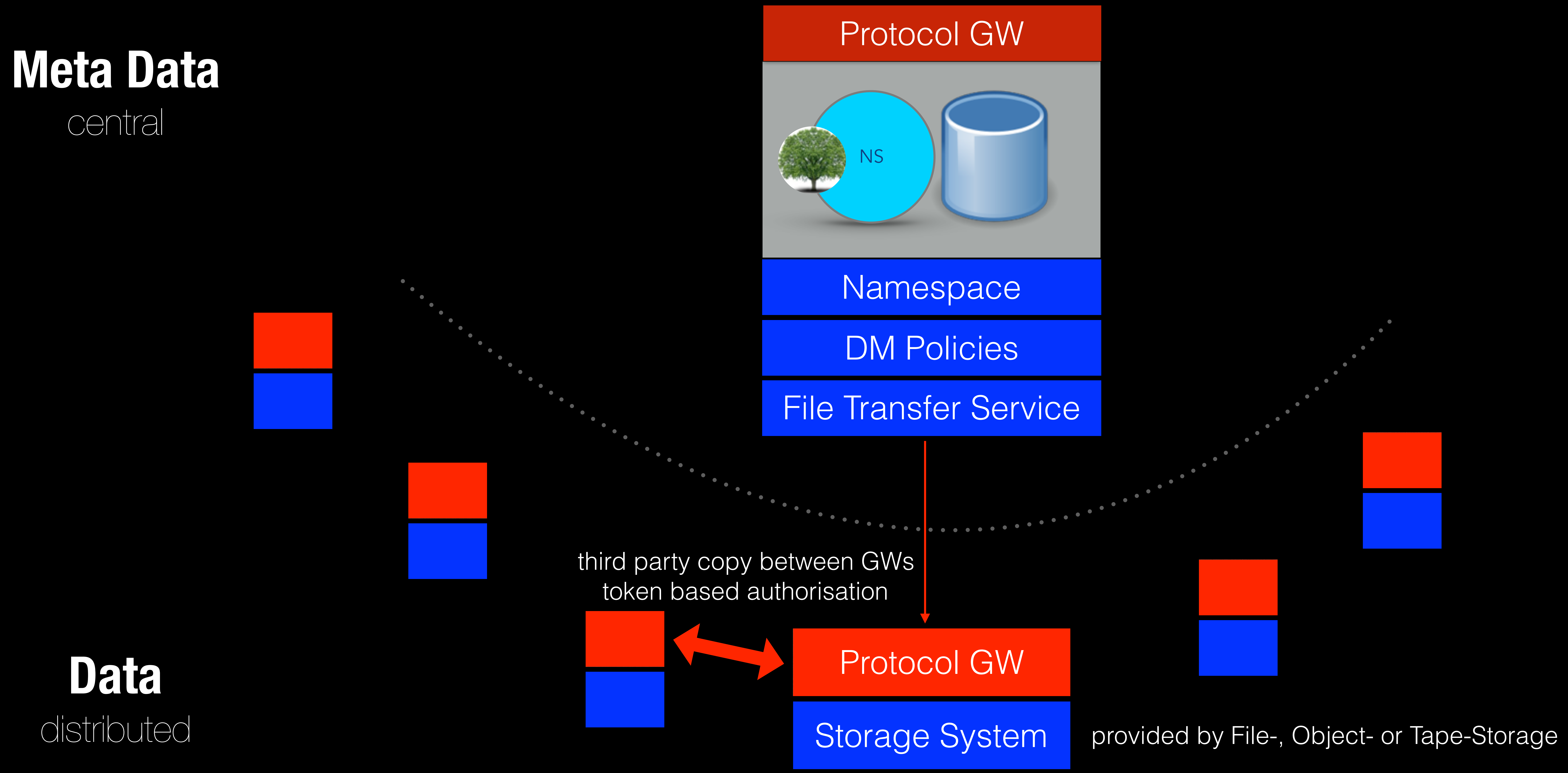
- AI Store
- AlluxIO
- AWS
- CEPH
- GCS
- Hadoop
- SkyllaDB
- Spark
- …



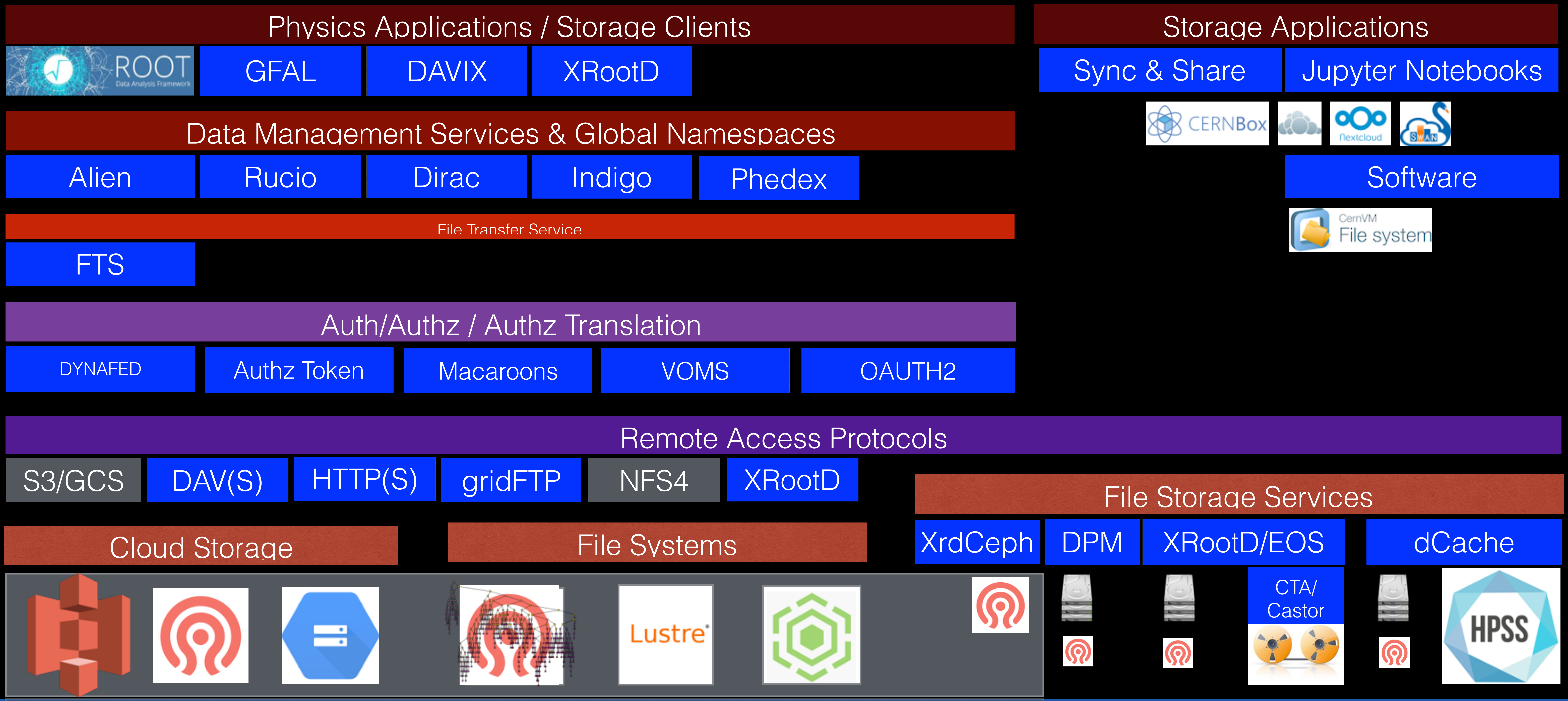
minimal overlap in development projects

many projects date back to the GRID area before BigData skyrocket

A simplified GRID Storage Architecture



HEP GRID Storage Ecosystem

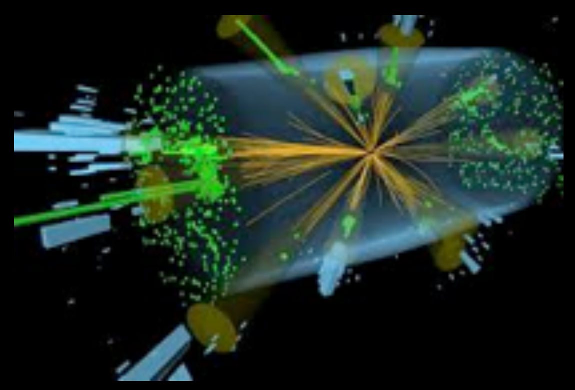


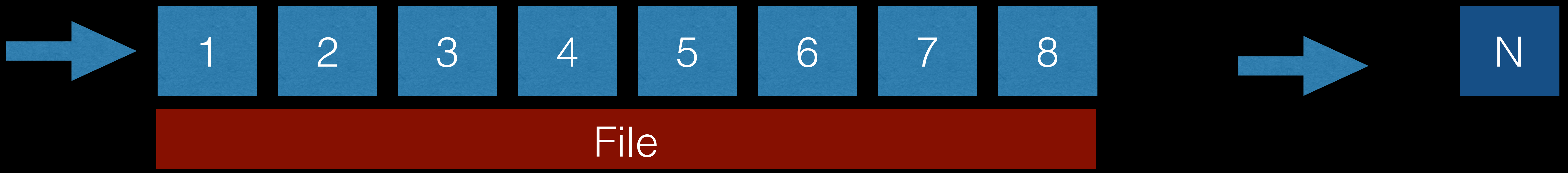
HEP & BigData Technology

- If we would use BigData Analytics in physics, we could profit from all the existing BigData storage technologies, protocols & analytics frameworks
- Why is that not yet mainstream?

Physics Data Formats

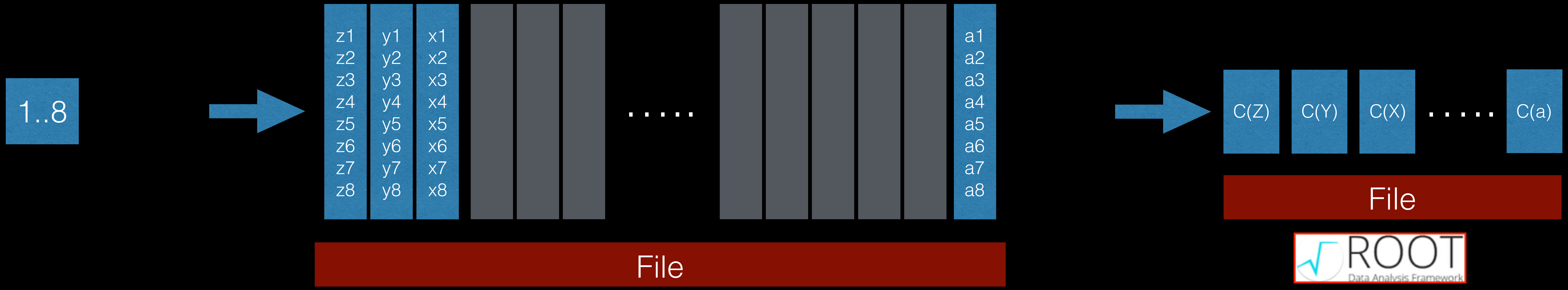
unstructured raw data - each physics event is stored in a compound block - events are assembled during data taking from many detector systems





compressed raw data with individual events

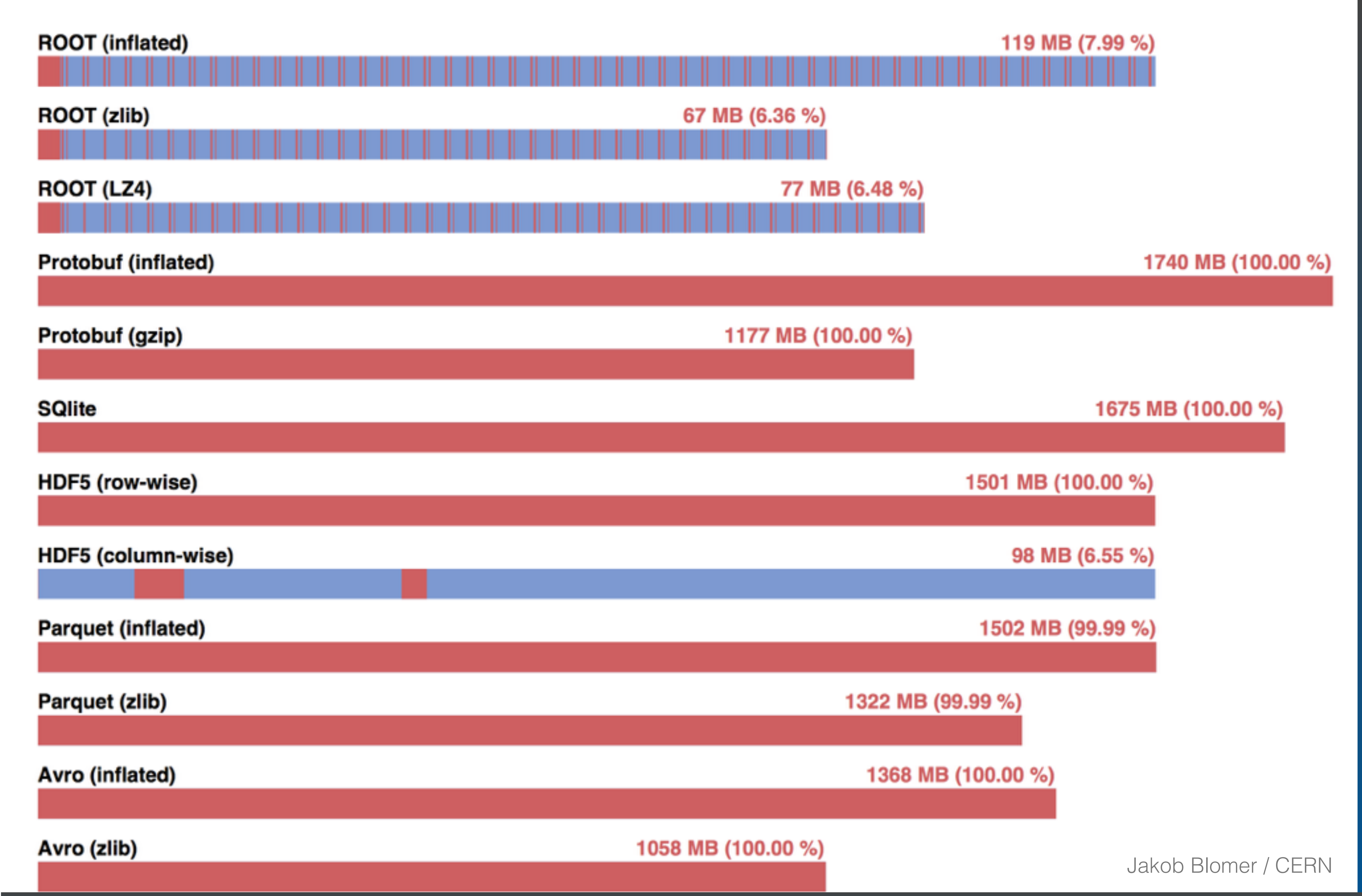
structured data - data is stored optimised for volume and access patterns



ROOT Data Analysis Framework

Data Formats & Storage Access Patterns
in selective analysis use cases

read pattern (read) in a selective physics analysis workflow



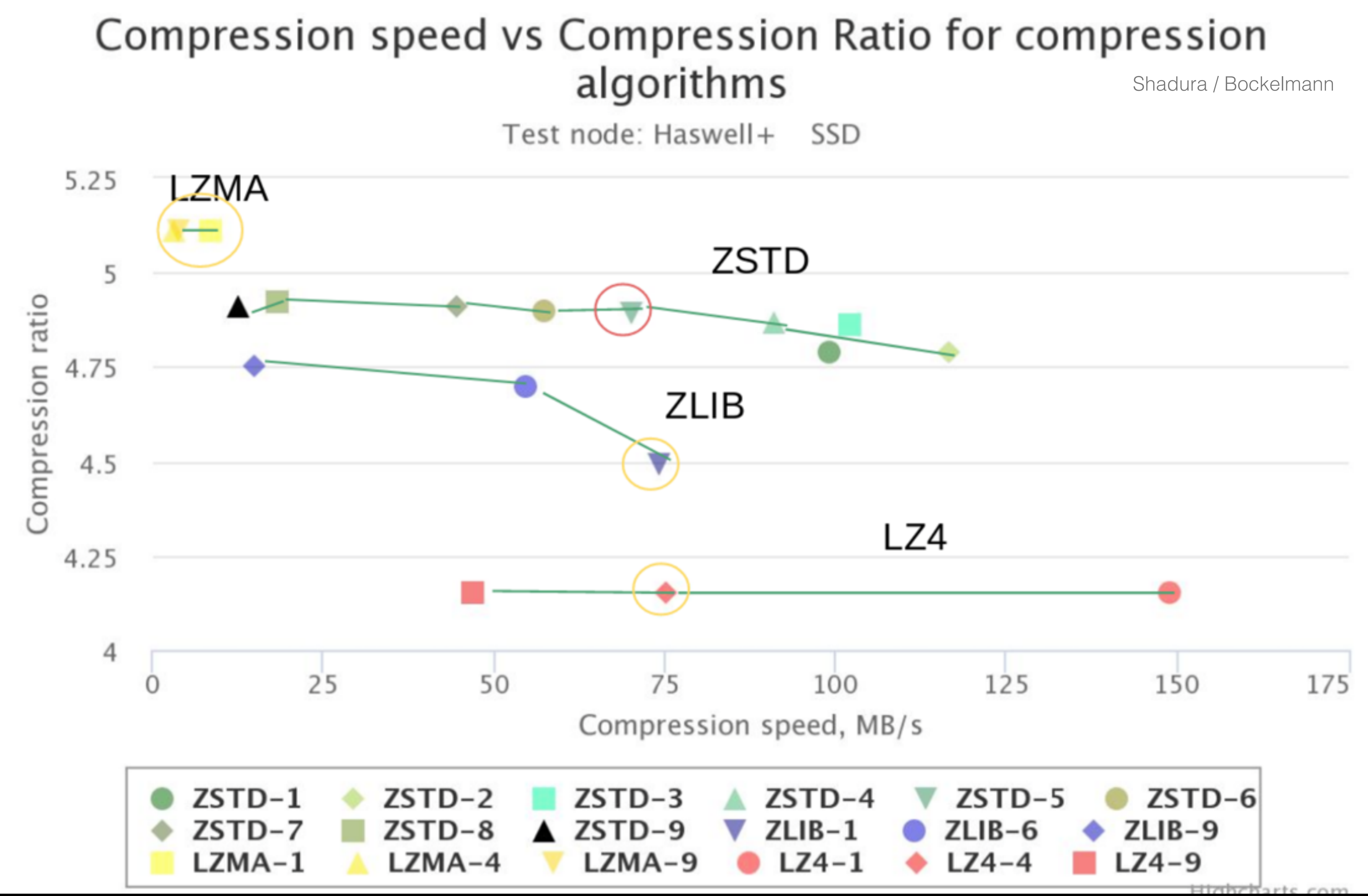
- ***sparse access*** pattern
  cry for certain access protocol capabilities in LAN & WAN environments

- predictable read patterns allows to use ***asynchronous multi-byte-range read*** requests to to compensate latencies

- good news: most of traffic in HEP is still mainly sequential forward-seeking IO
  jobs@CERN like 100.000 people watching all a different movie with 1 MB/s streaming average

physics analysis uses high parallelism with relatively slow streams (tens of MB/s) - no need for high throughput clients in the GRID

Data Formats & Compression Algorithms

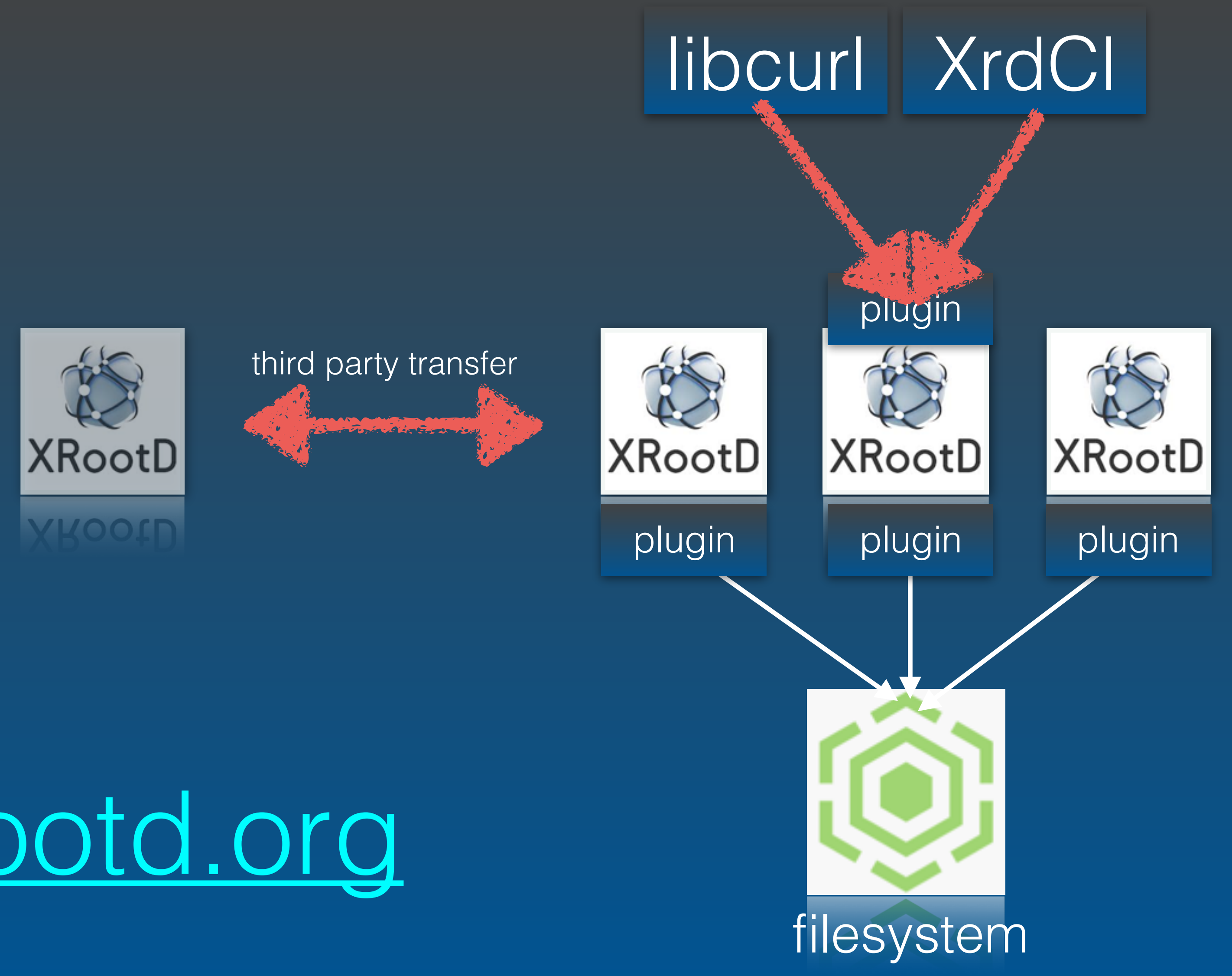


- **compression** done on **application side**
- **LZMA** cheapest for storage, most expensive for CPU
- best **algorithm** has to be selected **per use case** (de-/compression speed)
- compression inside storage systems rarely a benefit for physics data
- de-duplication marginal

HEP Data Access Protocols

prot/usage	LAN	WAN	WAN Transfer third party transfers
Mounted FS	high	-	-
XRootD	high	high	medium
HTTP(S)	low	low	comissioning
S3	low	low	-
gridFTP	-	-	

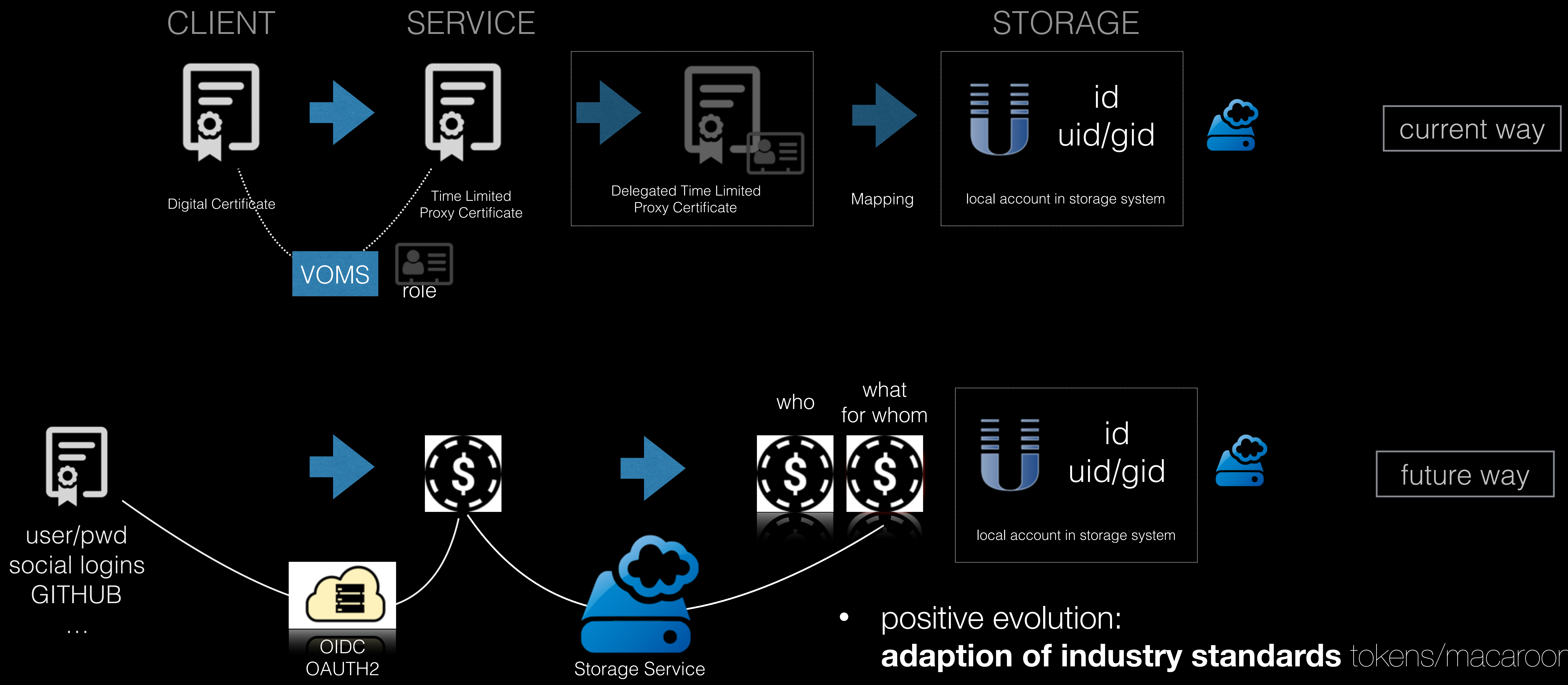
XRootD



http://xrootd.org

- data **client/server framework**
  think of NFS or HTTP server written in C++ with own protocol
- optimised for remote access in **LAN & WAN**
  - arbitrary request redirection
  - third-party transfer between XRootD server with credential delegation
- front-end **protocol plugins** XrootD & HTTPs
- **storage** back-end **plugins** XRootD & HTTPs & S3
- **authentication** plugins (krb5, x509, sss, unix)
- **authorization** plugins (rule-based, tokens, macaroons)
- **proxy** & **cache** plug-ins, **clustering** support

HEP Authentication & Authorization



- positive evolution:
  adaption of industry standards tokens/macaroons

Storage for future LHC Online Systems

Storage for future LHC Online Systems

- high **capacity** & high **IO** requirements - confined environment
- wide range of solutions possible: from distributed high performance **parallel filesystems** to **object storage** - key issue **cost**

TDR	ALICE	ATLAS	CMS	LHCB
IO Rate	200 GB/s	60 GB/s	61 GB/s	100 GB/s
Capacity	60 PB	36 PB	5.7 PB	100 PB

Storage for HPC

Storage for HPC

- typical use case: **MPI applications** requiring low latency access & high stability
  - most of LHC related computing done with **HTC** (trivial parallelism via batch jobs)
- playground of high performance filesystems Lustre, Spectrum Scale and others
  - e.g. NERSC Lustre 700 GB/s
  - CERN 'exotic' pioneering with CephFS

SCs significant resource for opportunistic computing

a common problem is the availability of storage clients for these platforms (e.g. FUSE based filesystems, services for data injection and extraction) and the external connectivity of HPC facilities

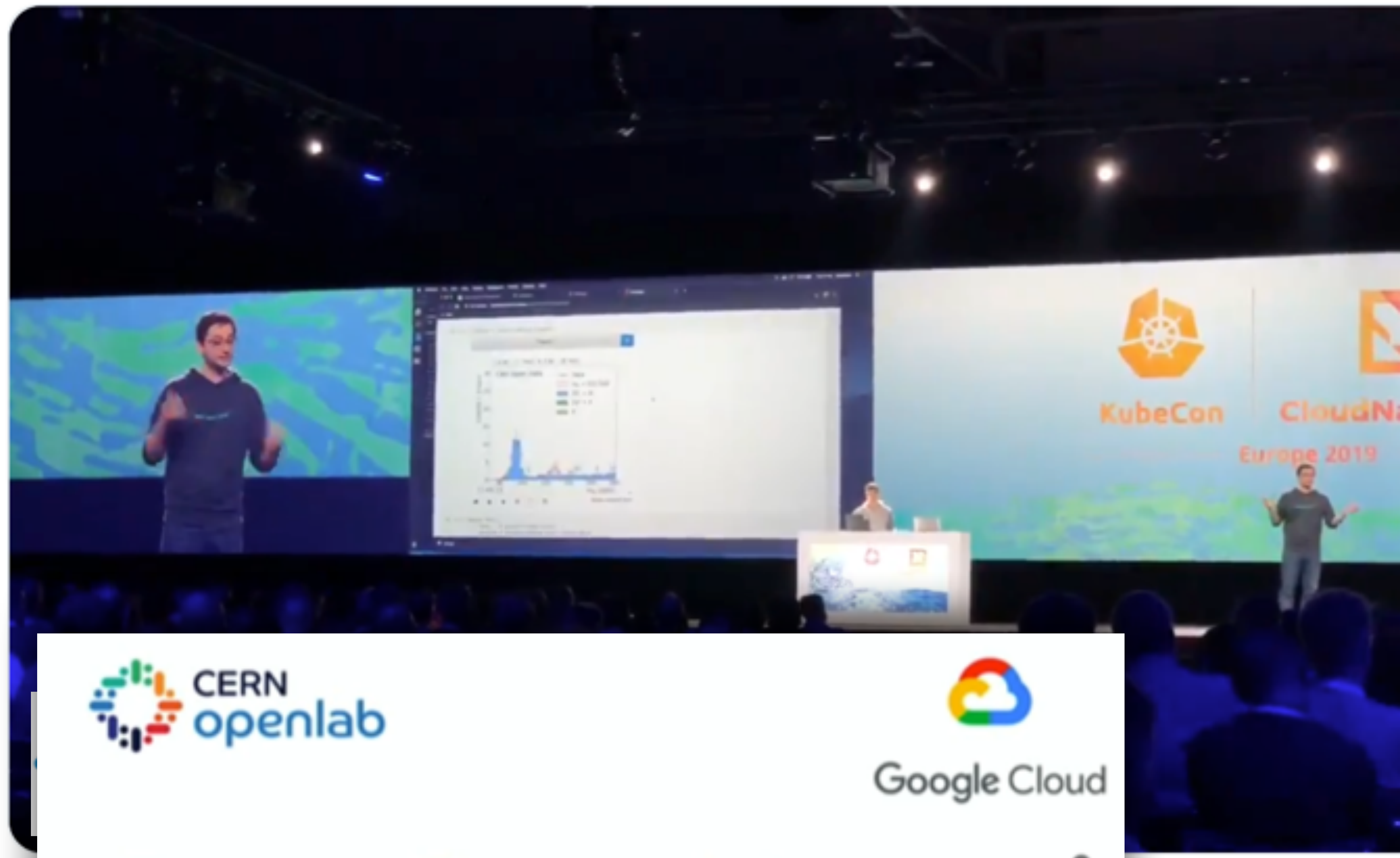
Storage in public clouds

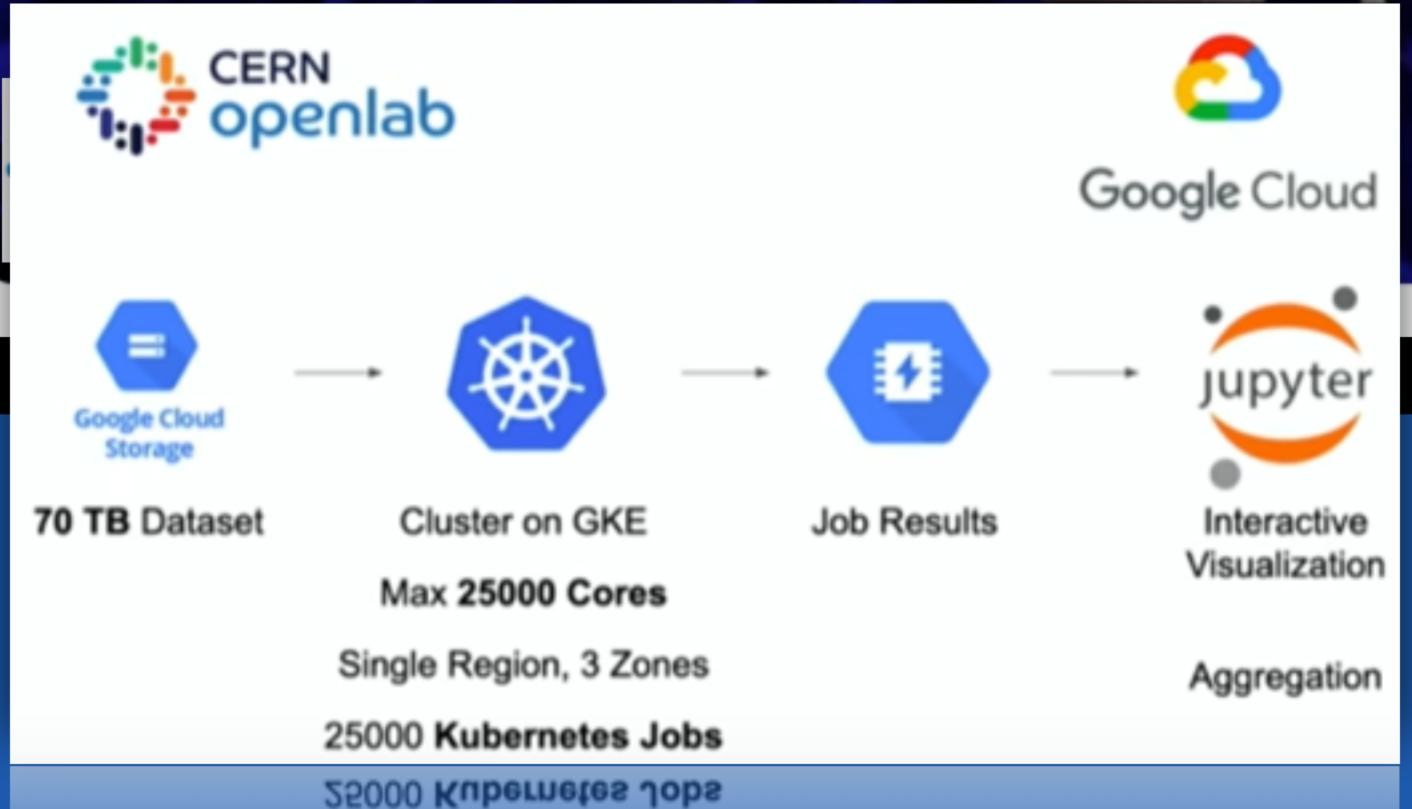
Storage in Public Clouds

- Public Cloud Services like AWS or GCS allow **time limited access to CPU resources**
  in times of high computing demands
  - **simplest use case simulations** mainly producing data
- Public Cloud Storage S3-like **easily integrated** as GRID resource
  pricing for storage and data access **not competitive** to replace HEP storage systems
- CERN successfully demonstrated **feasibility** of physics workflows in public clouds   CERN openlab collaborations

CERN scientists “rediscover” the Higgs boson live on stage at KubeCon using Google Cloud. Solution used Google Kubernetes Engine, Memorystore, and Storage (with network traffic peaking at 175G/s)! #k8s5

CERN R. Rocha et al





Storage for HOME directories

Storage for HOME directories

- several centres in HEP use commercial solutions like NetApp, Spectra Scale, DFS
  hit by unexpected increase in license costs
- CERN started replacing **DFS**
- **MALT** project: CERN strategy to decrease risk of vendor lock-in
- CERN also looking into long-term alternative for ***AFS*** *future unclear*

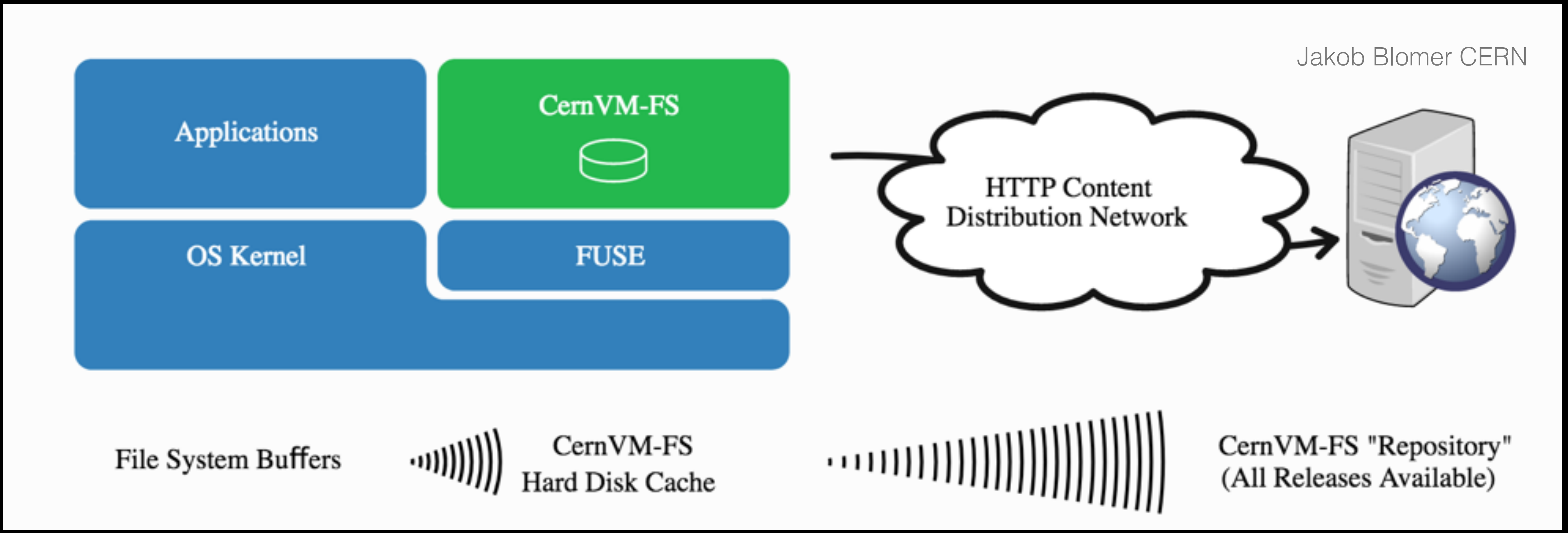
https://malt.web.cern.ch/malt/

Storage for Software Distribution

CernVM File System is a network file system based on HTTP and optimized to deliver experiment software in a fast, scalable, and reliable way

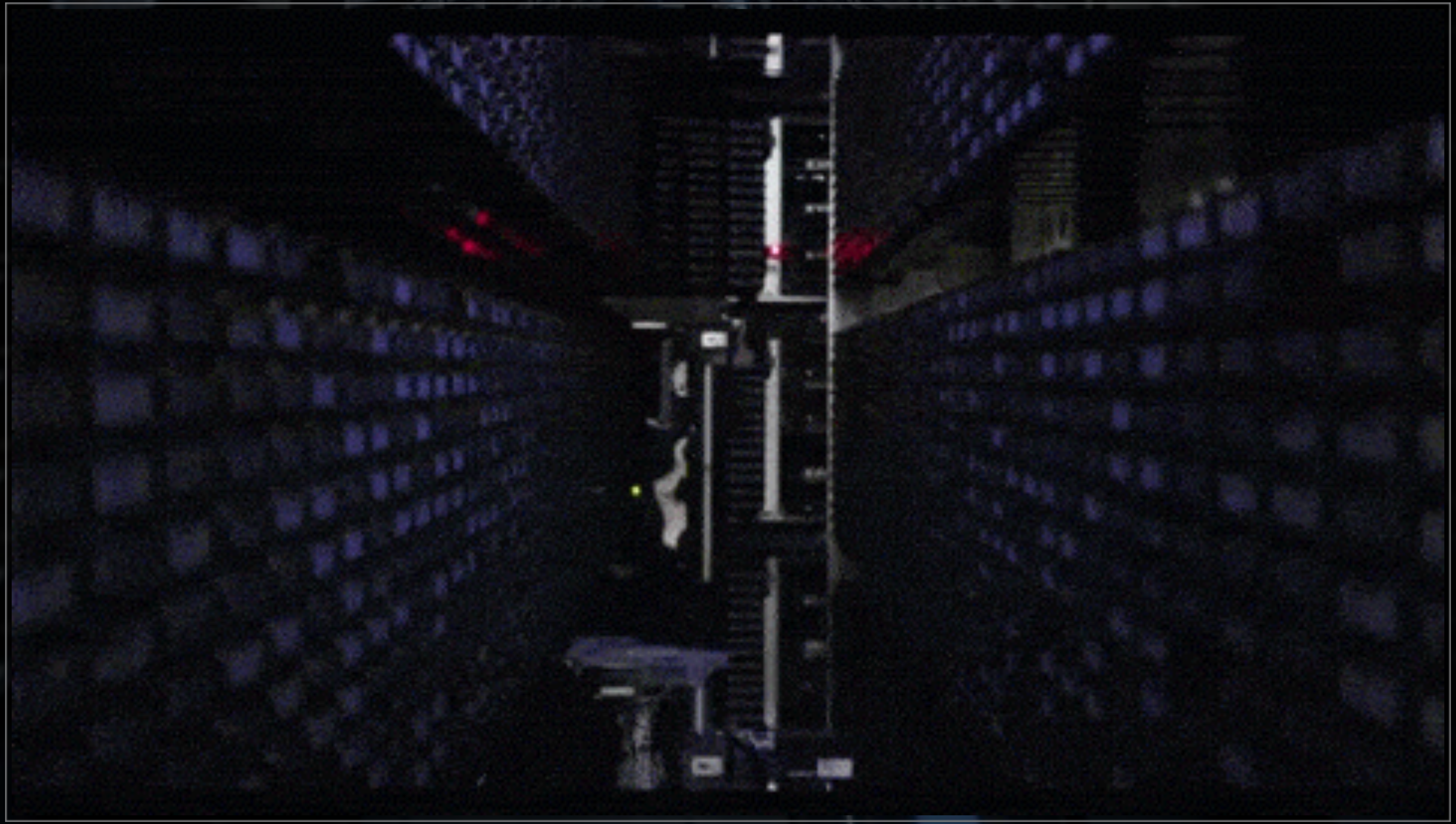
- typical use case: need to start any kind of software in 100k batch jobs at the same time

https://cernvm.cern.ch/portal/filesystem



- client is implemented as a FUSE based filesystem
- works like a content delivery network
- the central repository is published in Ceph S3 at CERN
- very popular and widely adapted
- already more than a read-only filesystem for software

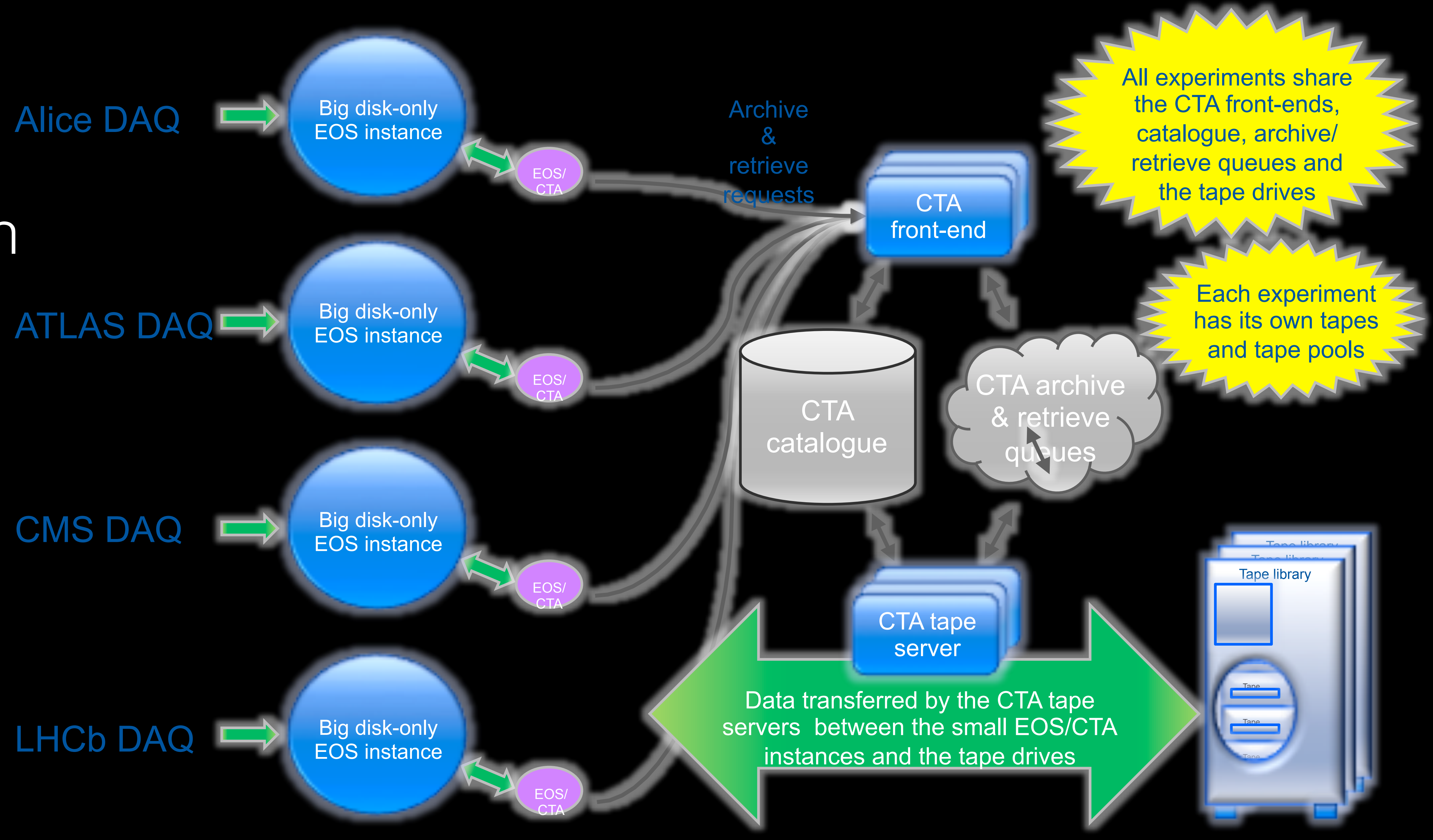
CERN Storage Software for Tape



CERN Storage Software for Tape

CTA is the next generation storage software for tape archiving and backup use cases developed by the CERN storage group

- replacing CASTOR software after 20 years
- design is decoupling disk pool implementation from tape storage
- flat namespace
- no HSM model, no complex GC



Inventory

- **HEP community** has built a very **modular stack of storage related software components** => allowed to **integrate** more or less **any storage solution** into the GRID
    - in the context of the **HEP Software Foundation** and Initiatives like the **XDC, IRIS** and **ESCAPE** projects many of these components are made **available to other sciences**
- **integrate-everything approach** is **not** the **most efficient** => inline with the community diversity
- **changes** in storage service implementations/technology **are slow** but happening => stateful nature of the service and limited resources to adapt new technologies
    - lifespan of storage software in HEP is decades

Inventory

- **file storage** is common denominator for most HEP storage systems
  - **site storage** systems have to fit also **requirements** of **local community**
- direct use of **object storage** appealing for physics analysis
  - diversity of the infrastructure is not compatible with global enforcement of object storage
  - **object storage** good match as a generic backend-solution - hide behind files
    find the economic/performant approach to gateways/security mechanisms
- Ancient community driven products/protocols **dying off**
  - **SRM** storage resource manager to handle tape storage access - lack of success - decommissioning ongoing
  - **gridFTP** - globus filte transfer protocol - replacement in testing
  - **rfio**- remote access protocols initially used in **Castor** - decommissioned

Inventory

- Storage Tiering: **HSM*** model for GRID DM has died *hierarchical storage management
  - no storage system can predict better access to data than the community who wants to use the data
    - manual(user driven) storage tier migrations have proven very successful and are part of data management frameworks used in HEP
- **Modern authentication & authorisation mechanisms** are slowly adapted by the community - they are not usable directly in filesystems - only via gateways
  - OIDC/OAUTH2
  - Macaroons
  - Tokens

Summary & Outlook

- **HEP storage** is a **diverse universe** of commercial & open source storage components
  - diversity is a **blessing & a curse** at the same time
- **HEP storage delivered** required functionality & infrastructure for LHC & other experiments
- In the future **cost** is becoming a **hard limitation** on *what is doable* and at the same time more competition on resources
- **Tape** is the strategic media *still*
- **Network** is a strategic resource to enable remote access for storage & caching
- **HEP** community has ability to **shape the future** and aim for simpler & more efficient storage within budgets
  - 😉 open source technology helps in achieving this goal
  - 😨 amount of data produced is not influenced by storage technology but physics & computing models …

Thanks for the attention!
Questions?