CERN

Accélérateur de science

Data Storage Technologies

Past, Present, Future

Luca Mascetti
Storage and Data Management Group

A bit of history…

Image from Peggychoucair on Pixabay

A bit of history…

CERN Storage 1990s

SHIFT: Scalable Heterogeneous Integrated Computing Facility

Overview of the SHIFT Architecture

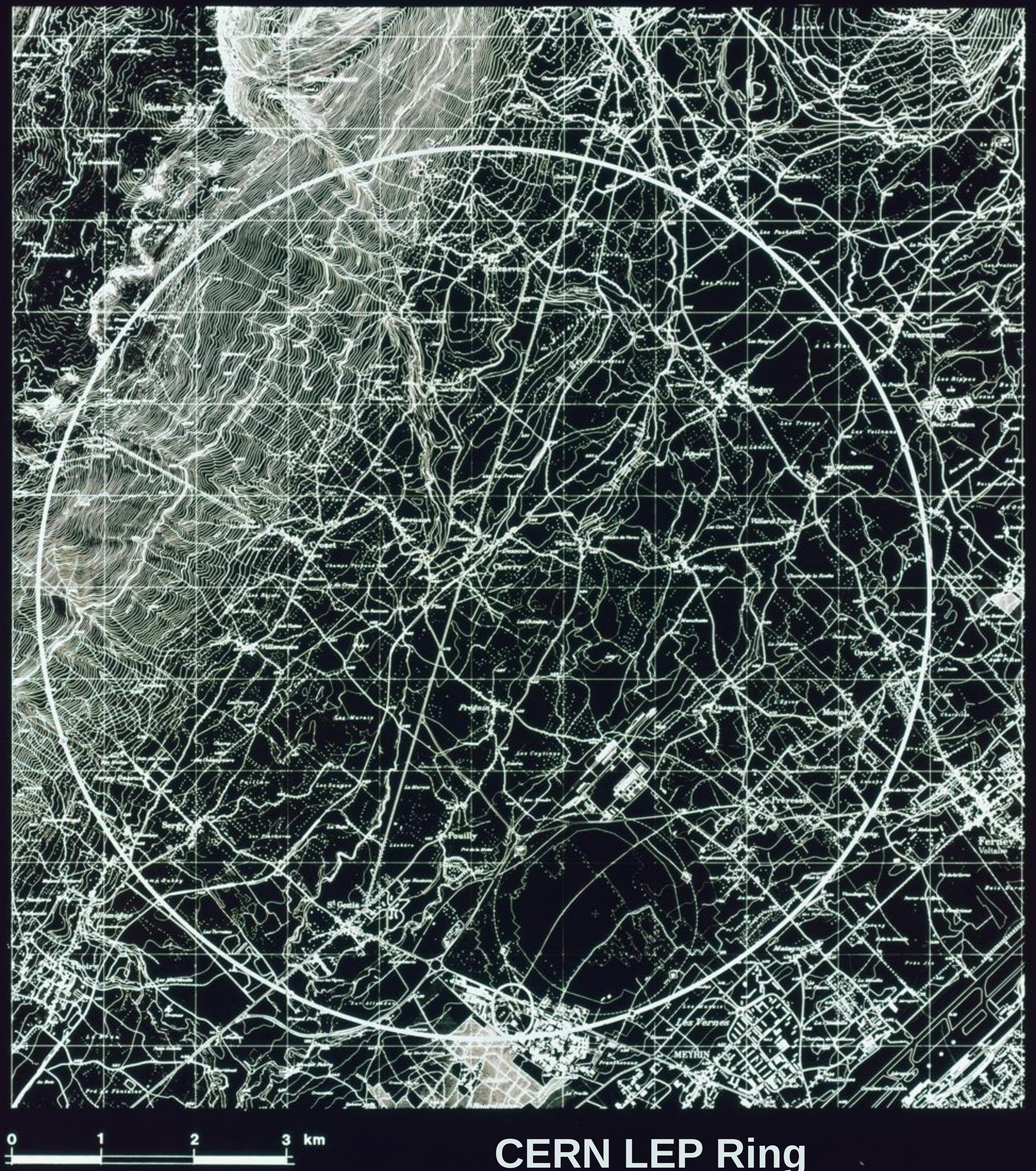
The designers of **shift** were motivated by the appearance on the market of inexpensive processors and storage systems, using technology developed for personal workstations, and which had performance characteristics comparable with those of traditional mainframes.

The goal was to define an architecture which could be used for general purpose High Energy Physics (**hep**) computing, could be implemented to provide systems with an excellent price/performance ratio when compared with mainframe solutions, and could be scaled up to provide very large[1] integrated facilities, or down to provide a system suitable for a small physics department.

Some important characteristics of **offline hep** processing relevant to the the design choices are:

- the volume of data which must be held on online storage (up to the order of 10^{12} bytes);
- the need for access to magnetic tapes, used to store fuller information about "events" (a few terabytes);
- difficulty in finding vectorisable algorithms for a significant fraction of the processing requirements (hard to exploit supercomputers);
- inherent parallelism in much of the processing (events are largely independent);

[1]compared with the capacity of current **hep** Computer Centres.



CERN LEP Ring

SHIFT in 1998


NH NH

CERN Storage 2000s

CASTOR:CERN Advanced STORage manager

1998–2007 CASTOR 1

2005–2022 CASTOR 2

Hierarchical Storage Manager (HSM)

- **Automatic move from disk to tape and vice-versa**
- **Mix of production use-cases and end-users’ analysis**
- **Transfers Scheduling (model as Job-Scheduling)**

CASTOR

CERN Advanced STORage manager


StorageTek
StorageTek
StorageTek

CERN Online Storage 2010s

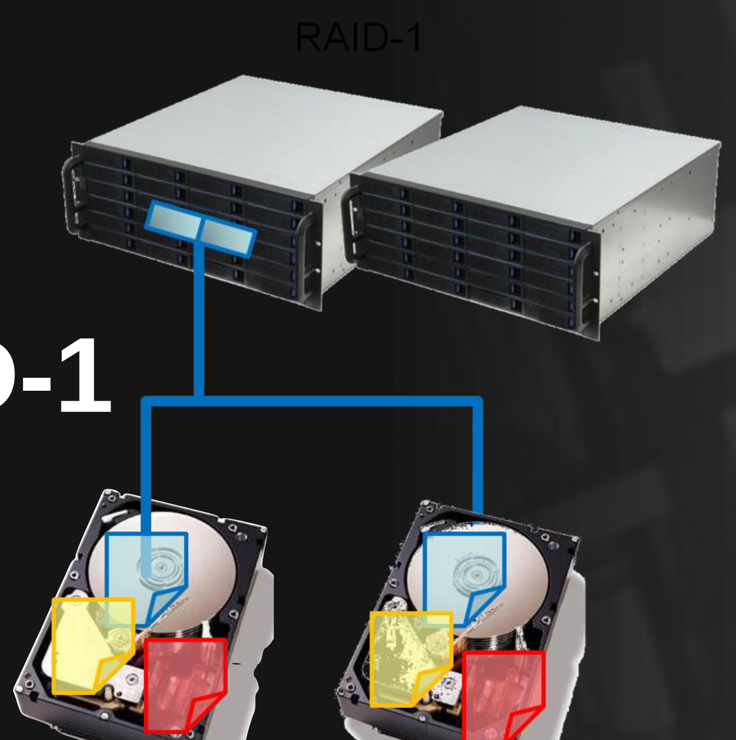
EOS

EOS Open Source

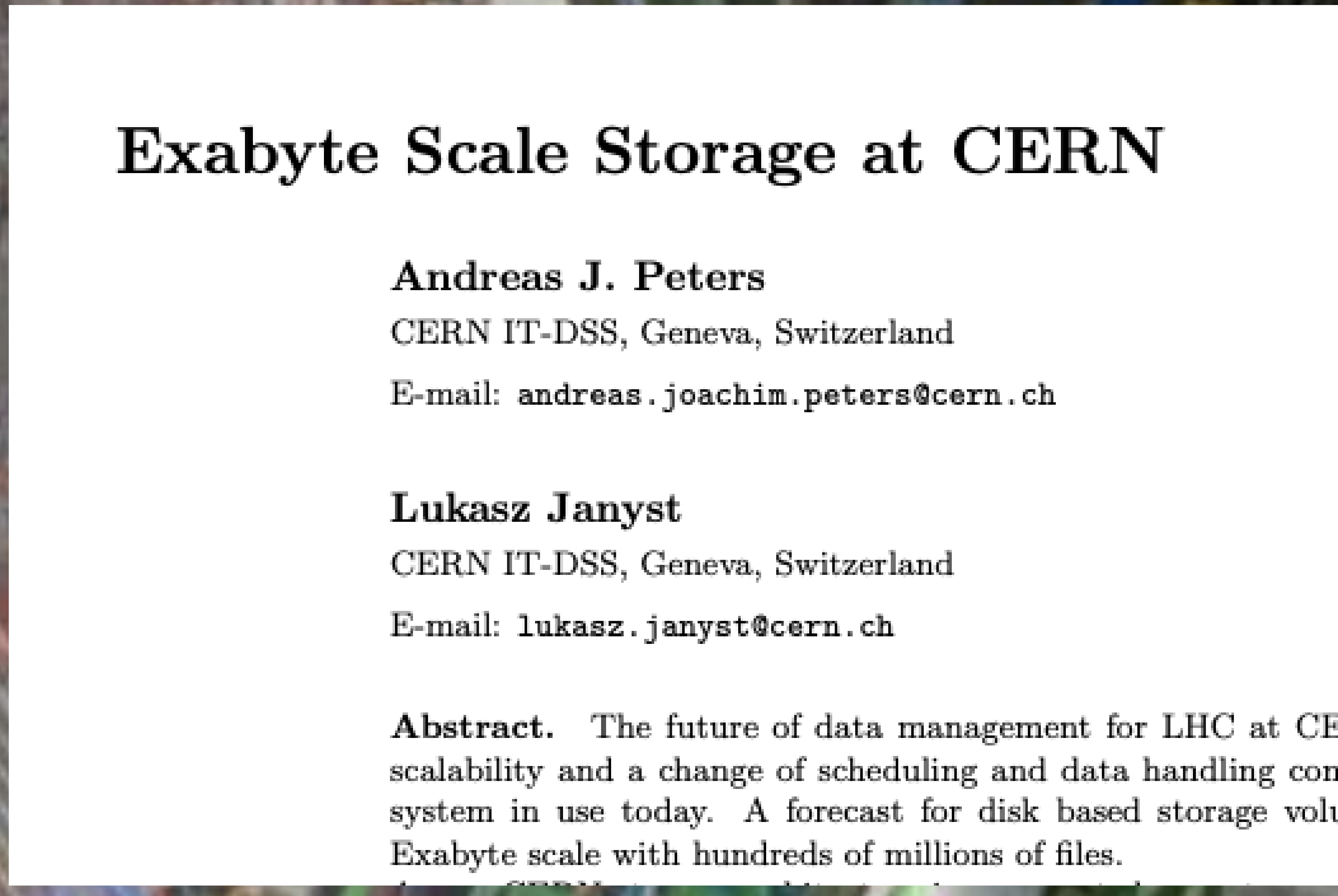
From "Hierarchical" to "Tier" Model

- Dedicated "storage pools" with defined QoS (Analysis, Archive, Tape)
- Experiments' Data Management frameworks manage the transitions
- Low-Latency namespace
- POSIX-like file access
- New data replication paradigm (RAID vs. RAIN and EC)
- Designed to scale at the Exabyte level

RAID-1



RAIN-1



Exabyte Scale Storage at CERN

Andreas J. Peters
CERN IT-DSS, Geneva, Switzerland
E-mail: andreas.joachim.peters@cern.ch

Lukasz Janyst
CERN IT-DSS, Geneva, Switzerland
E-mail: lukasz.janyst@cern.ch

Abstract. The future of data management for LHC at CE
scalability and a change of scheduling and data handling co
system in use today. A forecast for disk based storage vol
Exabyte scale with hundreds of millions of files.

Data Access Patterns in Physics

Data Analysis

- >100k relatively slow streams reading data (almost) sequentially from 60k HDDs
  - 1-100 MB/s - sometimes forward-seeking
    "similar to 100k people watching an individual film on Netflix"

Data Acquisition / Data Taking

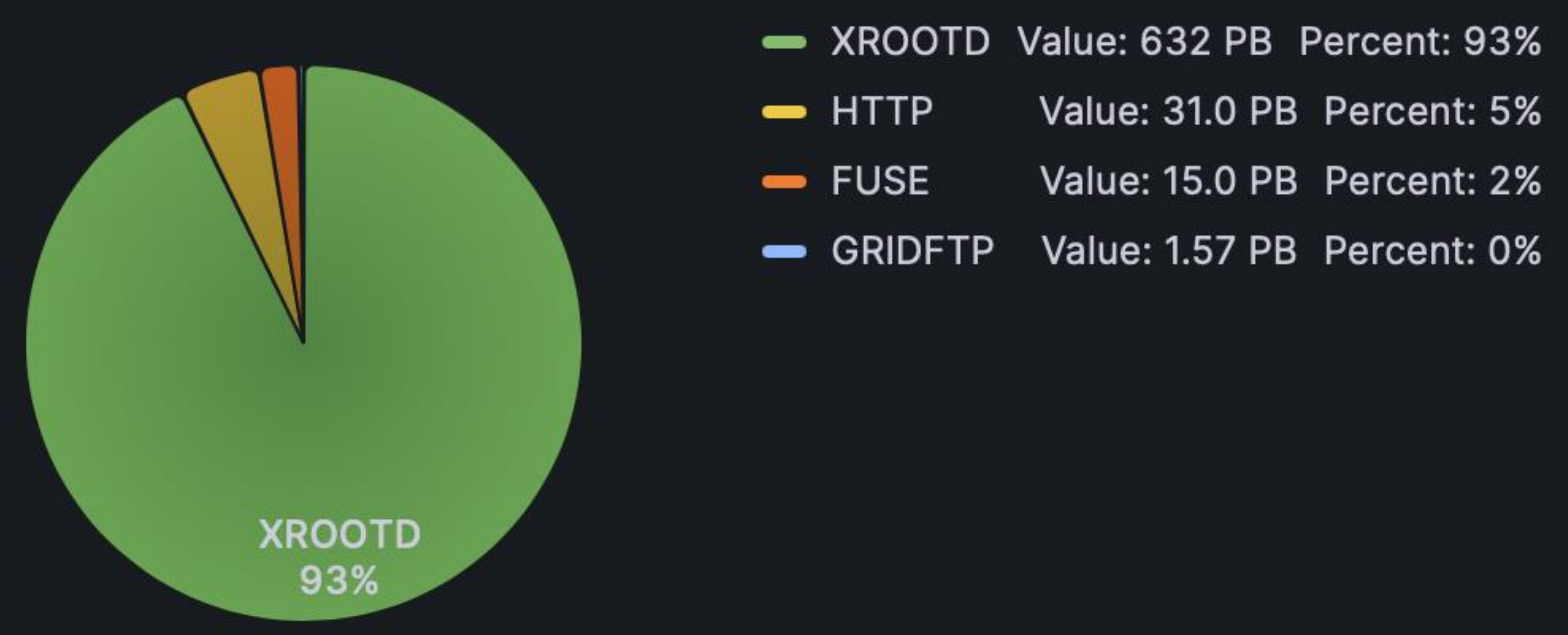
- hundreds of streams possibly as fast as possible
  - 50-250 MB/s with File Replication
  - 400 MB/s-1 GB/s with Erasure Coding

EOS Service (2023)

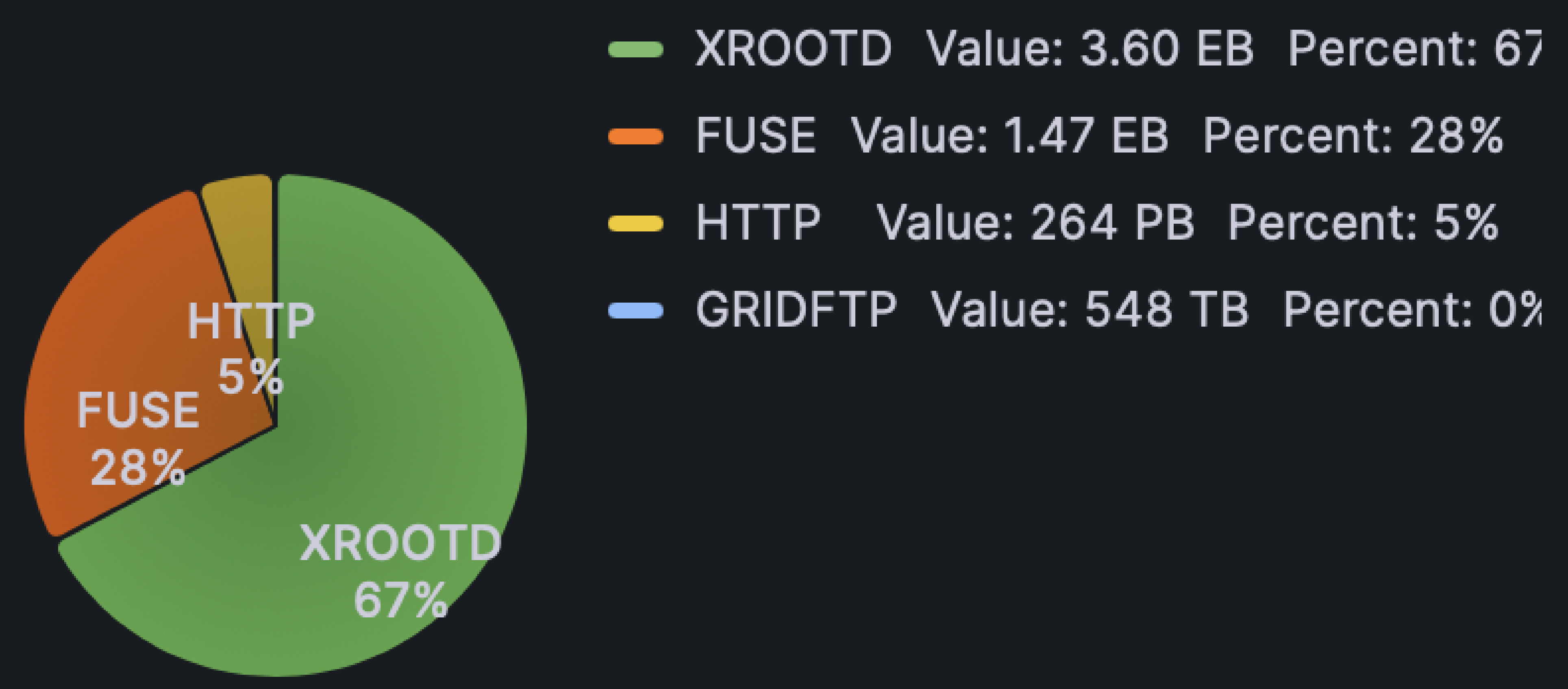
2023
full year

Total amount of files read	Total amount of bytes read	Total amount of files writ...	Total amount of bytes wri...
21.6 Bil	5.34 EB	4.51 Bil	679 PB

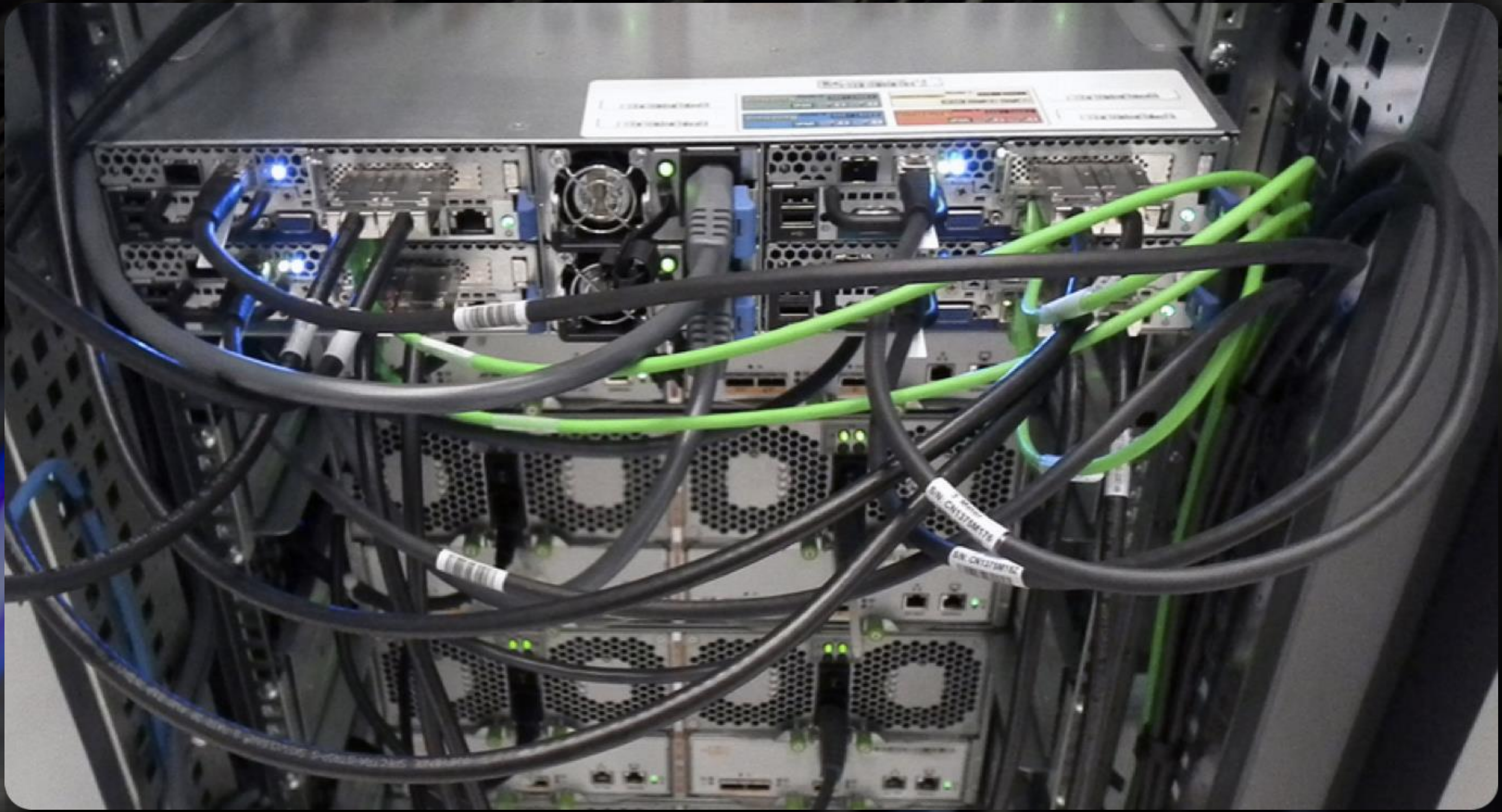
Total data write per protocol and instance: All



Total data read per protocol and instance: All



The Storage "Building Block"



QUAD + SAS Arrays

The Storage “Building Block”

Over the years we commissioned and operate multiple solutions:

- Server + 2x 24-bay SAS Arrays
- Server + 4x 24-bay SAS Arrays
- Server + 8x 24-bay SAS Arrays
- Server + 60-bay SAS Array
- Server + 2x 60-bay SAS Array

Storage Server in 2014: 200 TB
Storage Server in 2023: 1700 TB
Storage Server in 2024: 2300 TB

Networking Evolution in the last 10Y

1Gb ➜ 10Gb ➜ 25 Gb ➜ 40Gb ➜ 100Gb


CERN IT - Operated Disk Storage Capacity
1200 PB
1000 PB
800 PB
600 PB
400 PB
200 PB
0 PB
EOS
others
2010
2012
2014
2016
2018
2020
2022
2024

Dedicated 16 PB EOS Cluster in Point 2 (next to ALICE) hosted in a container

~1/5 of one of an EOS cluster dedicated to one LHC experiment (ALICEO^2)

CERN Archival Storage 2020s

CERN Tape Archive

EOS + CERN Tape Archive (CTA)

- Successor and full replacement of CASTOR2 for Tape Access and management
- Implemented as tape backend to EOS
- Small and fast buffer based on fast SSDs

CERN Tape Archive (CTA) – Tape Library

Physics Storage and Data Management Services

Storage

EOS

eos.cern.ch

Software to manage Disk Storage - **930 PB**

CERN Tape Archive

cta.cern.ch

Software to manage Tape Storage - **730 PB**

DISK TRANSFER

TAPE

Data DISTRIBUTION

Data Management

FTS

File Transfer Service

fts.cern.ch

Middleware to run File Transfers - **1 Billion / year**

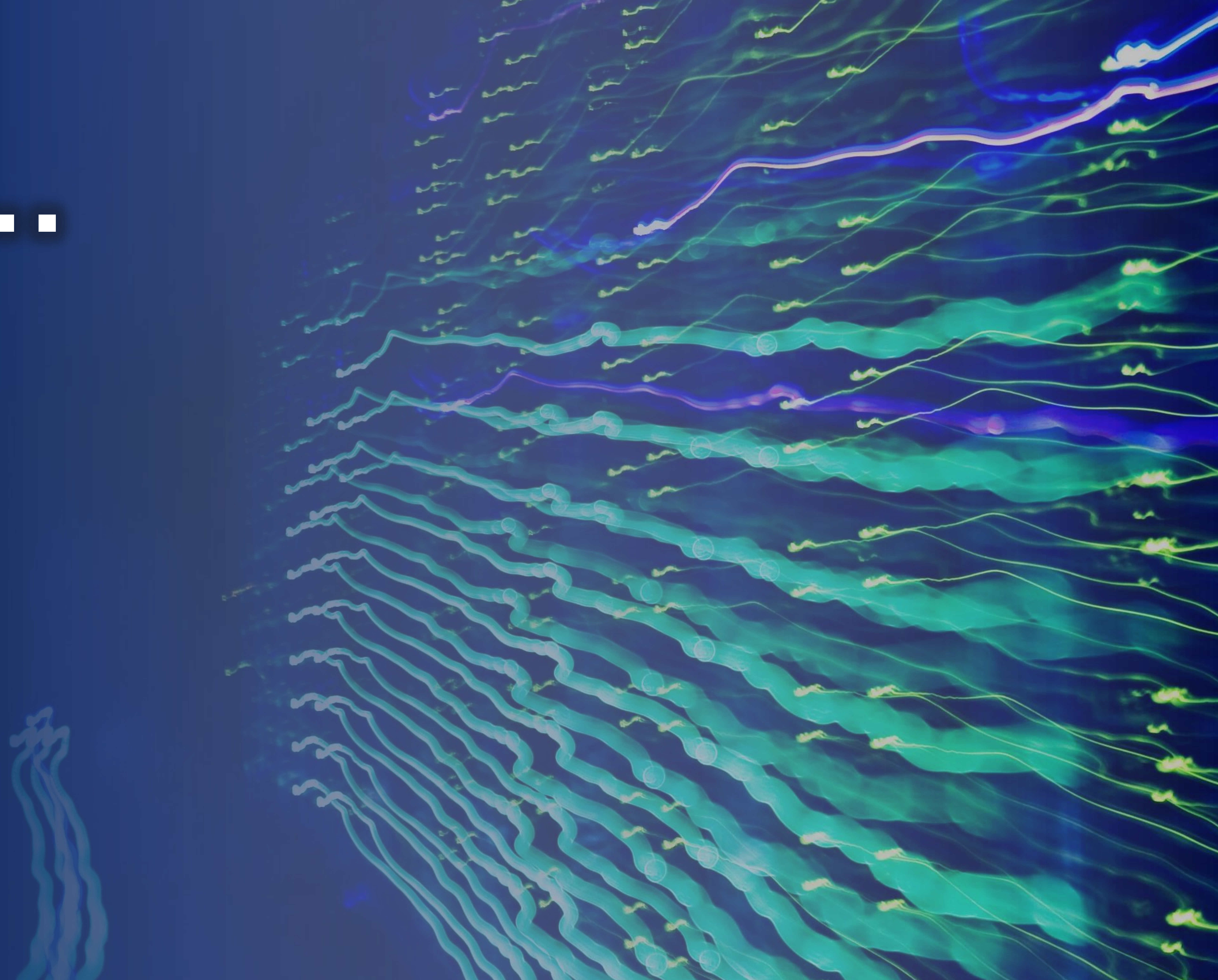
RUCIO

SCIENTIFIC DATA MANAGEMENT

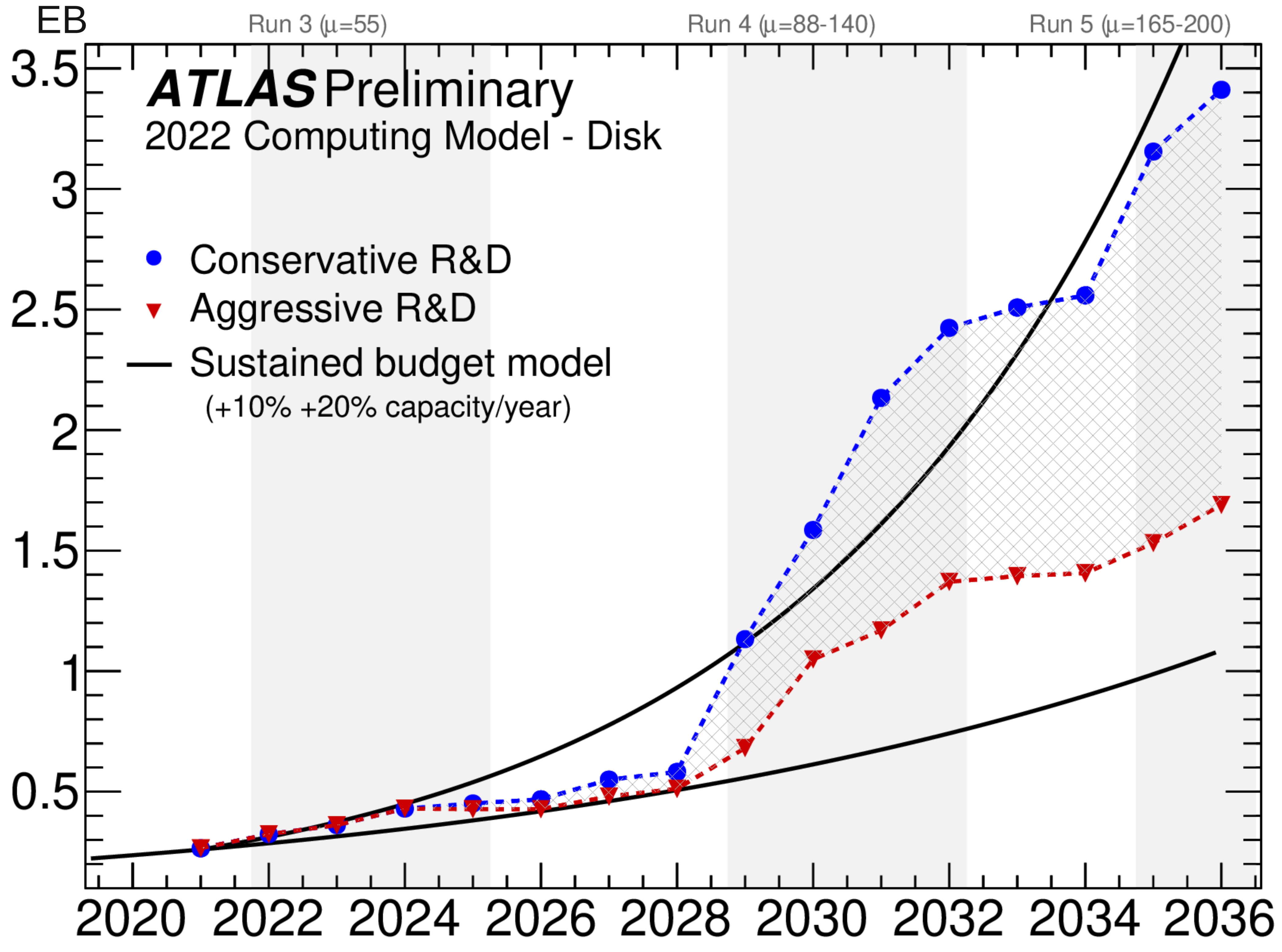
rucio.cern.ch

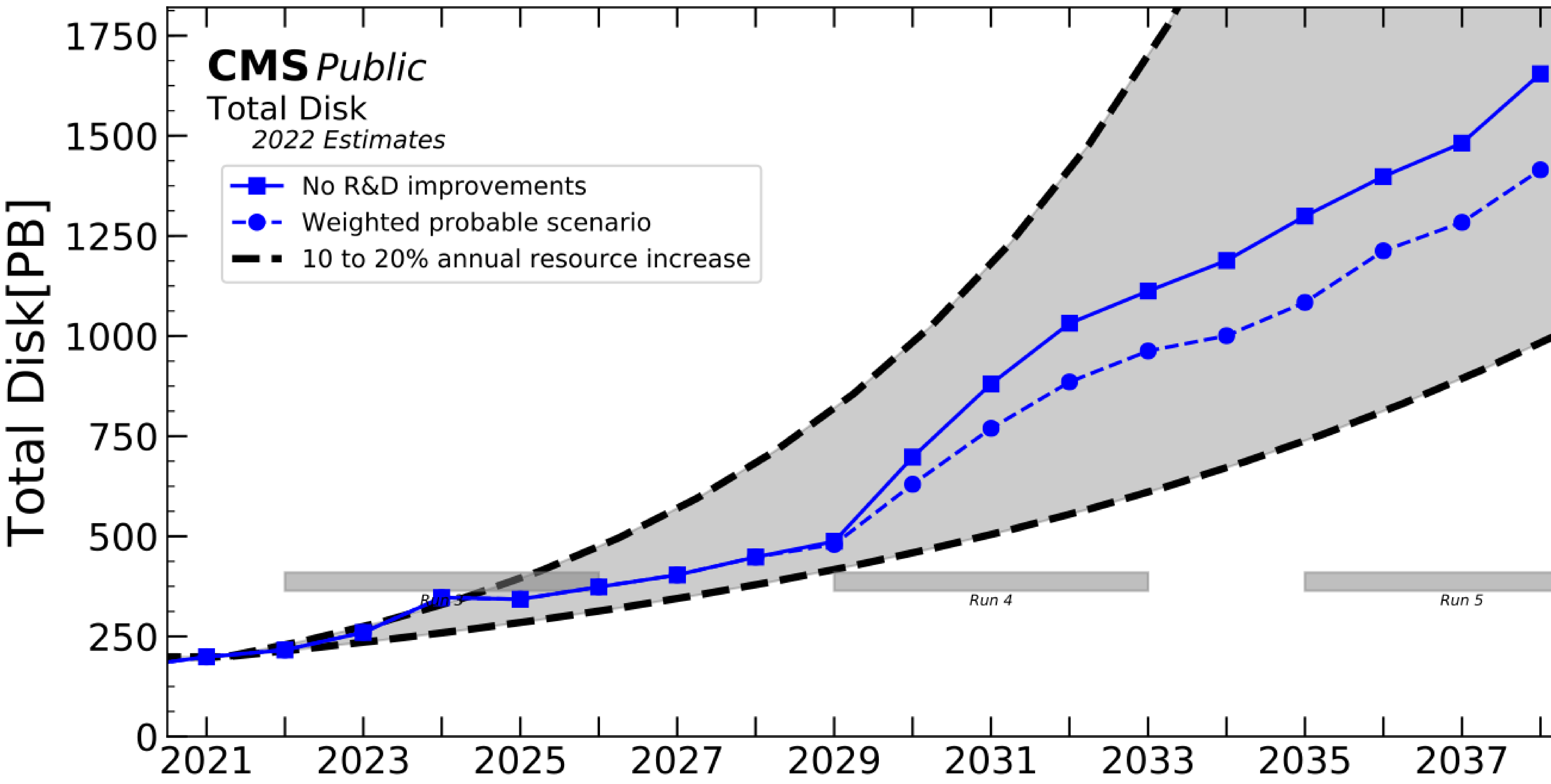
Data Management /
Data Distribution over 162 sites

Future…

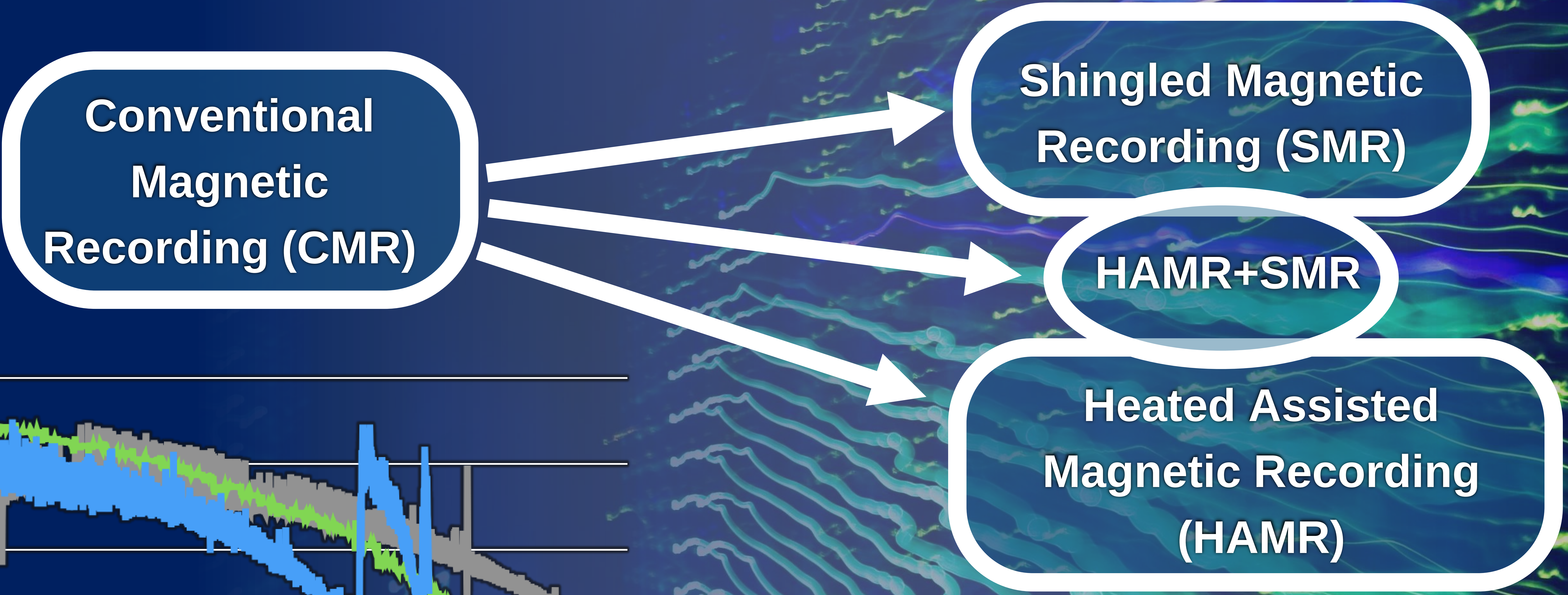


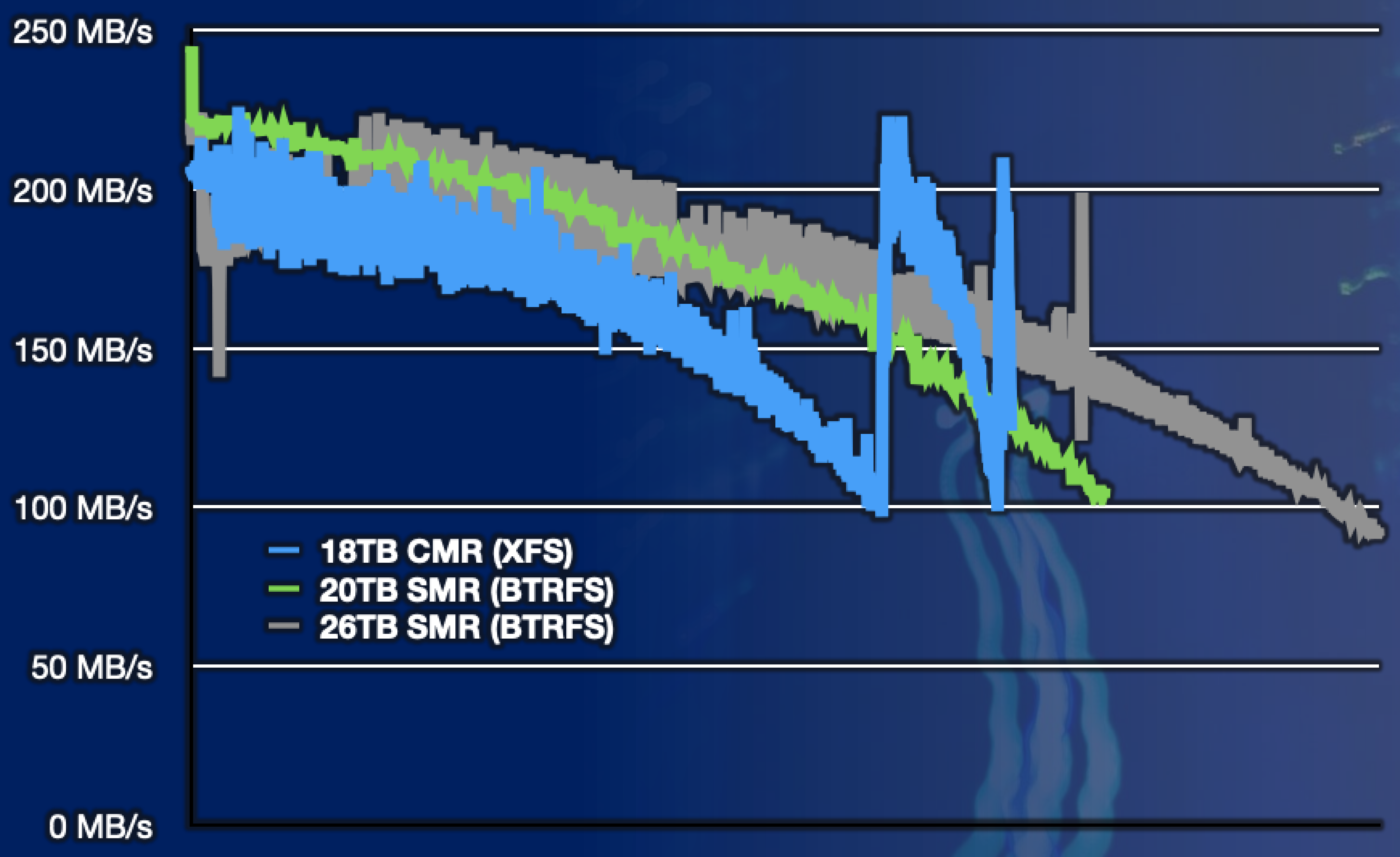
Online Storage
ATLAS and CMS Predictions





Future… Hard Disk Drives





Future Disk technologies increase the bit areal density
There are performance implications:

- Random write access patterns
- Fill and removal cycles of devices (generating “holes”

Future… SSD/NVMe roles…

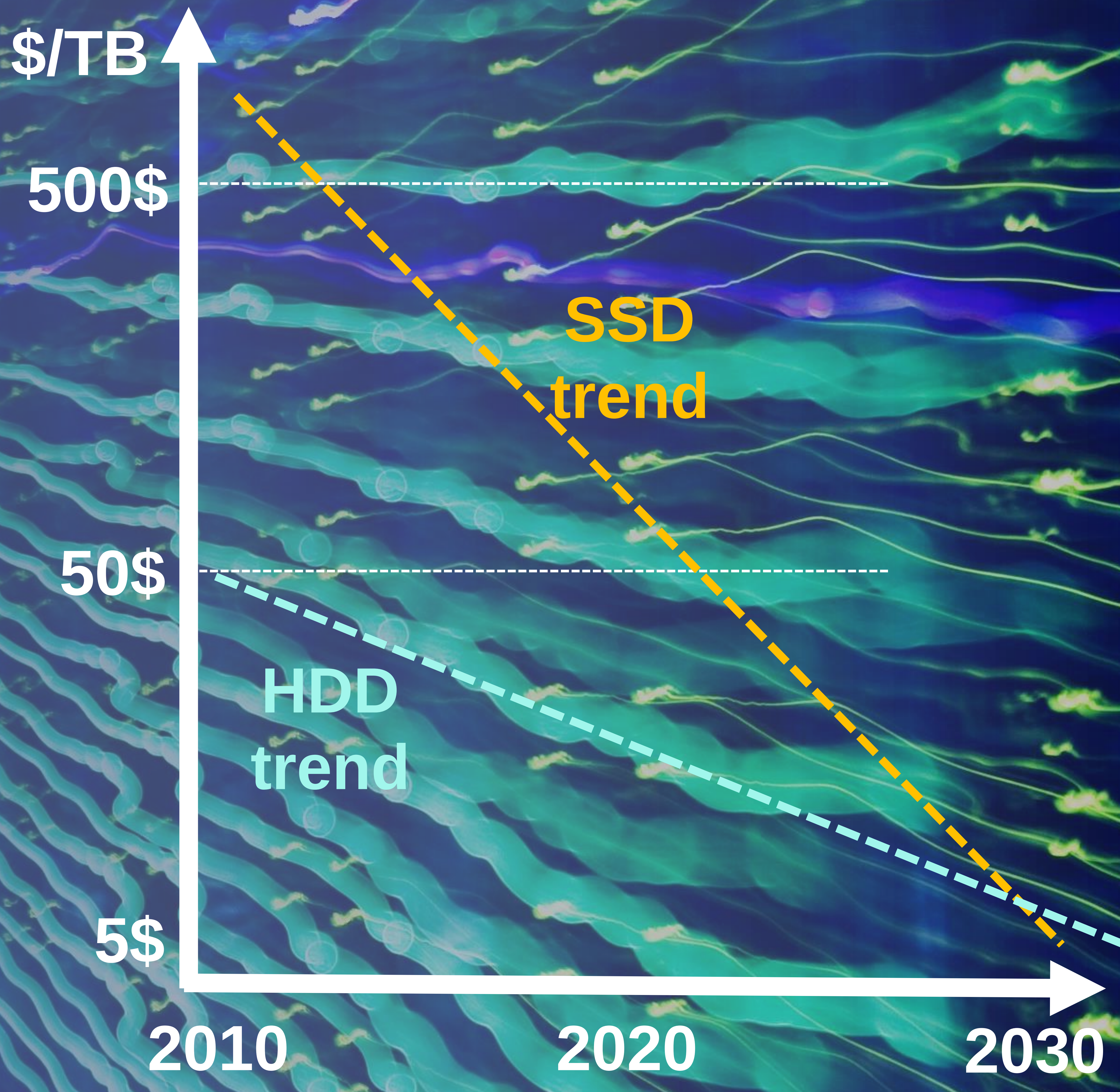
More and more use-cases (e.g. AI, ML, etc..) have different IO patterns compared to physics analysis.

We expect to see in the future more random access (both read and write) to our storage.

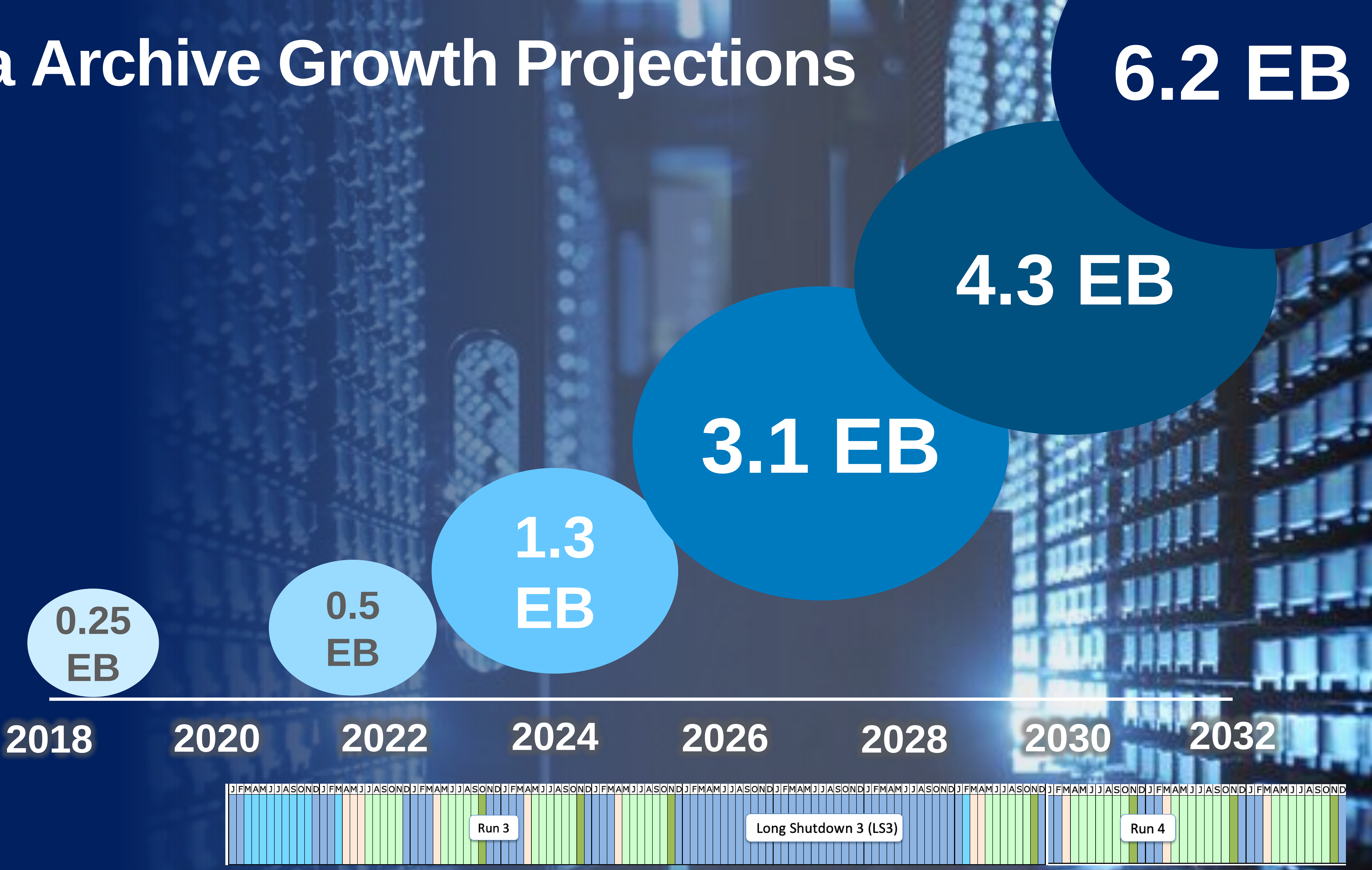
SSD will be fundamental to address these new requirements.

Our storage can “transparently” integrate these technologies.

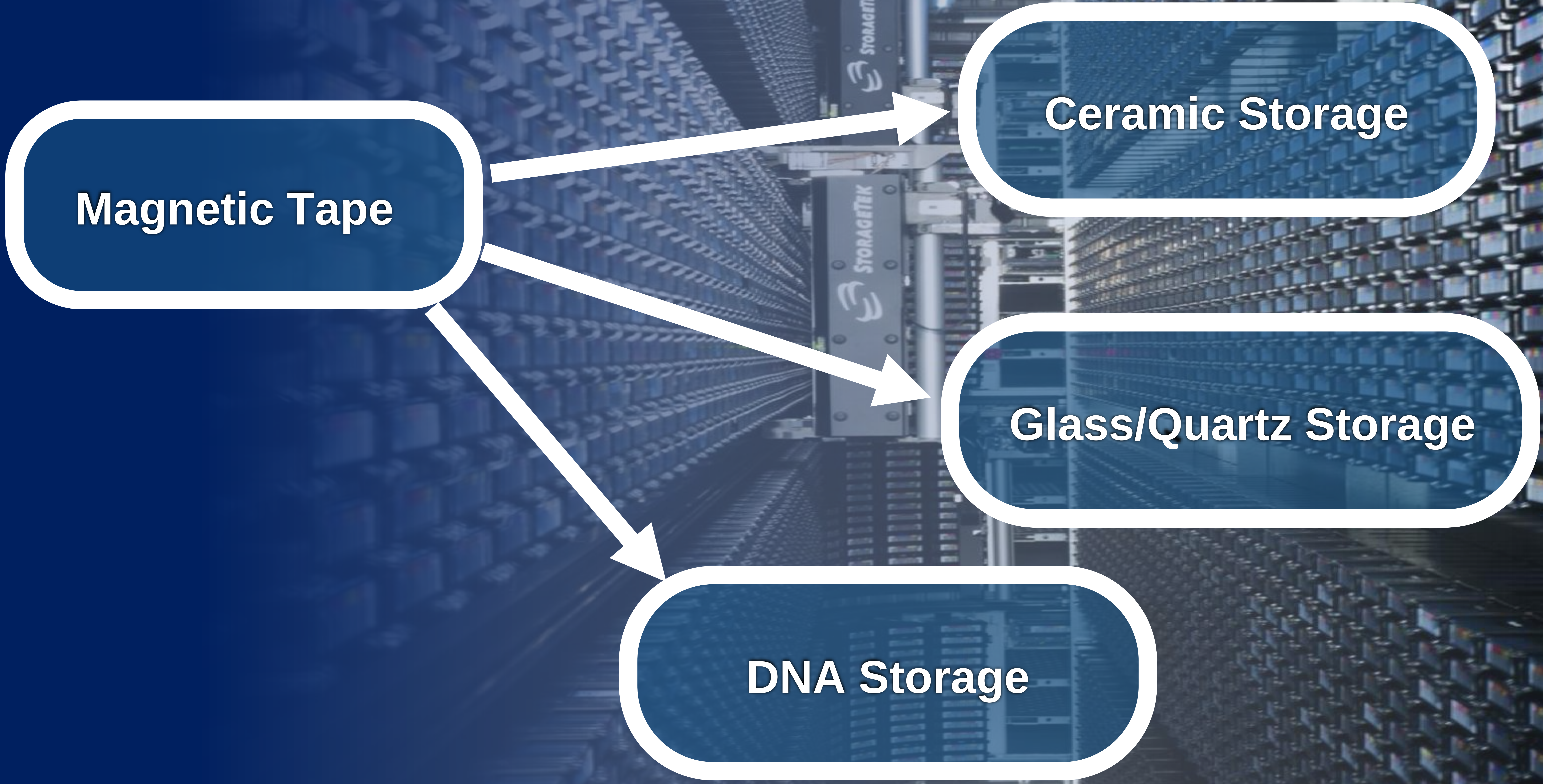
In the future we plan to leverage from built-in trade-offs between performance, reliability, endurance, price and capacity



Data Archive Growth Projections



Future… Data Archive



CERN openlab activities and R&D directions

PHASE VIII

ACCELERATING STORAGE FOR SCIENCE

- **Pioneering sustainable infrastructures**
- **Evaluating emerging storage solutions**

Thanks for the attention!