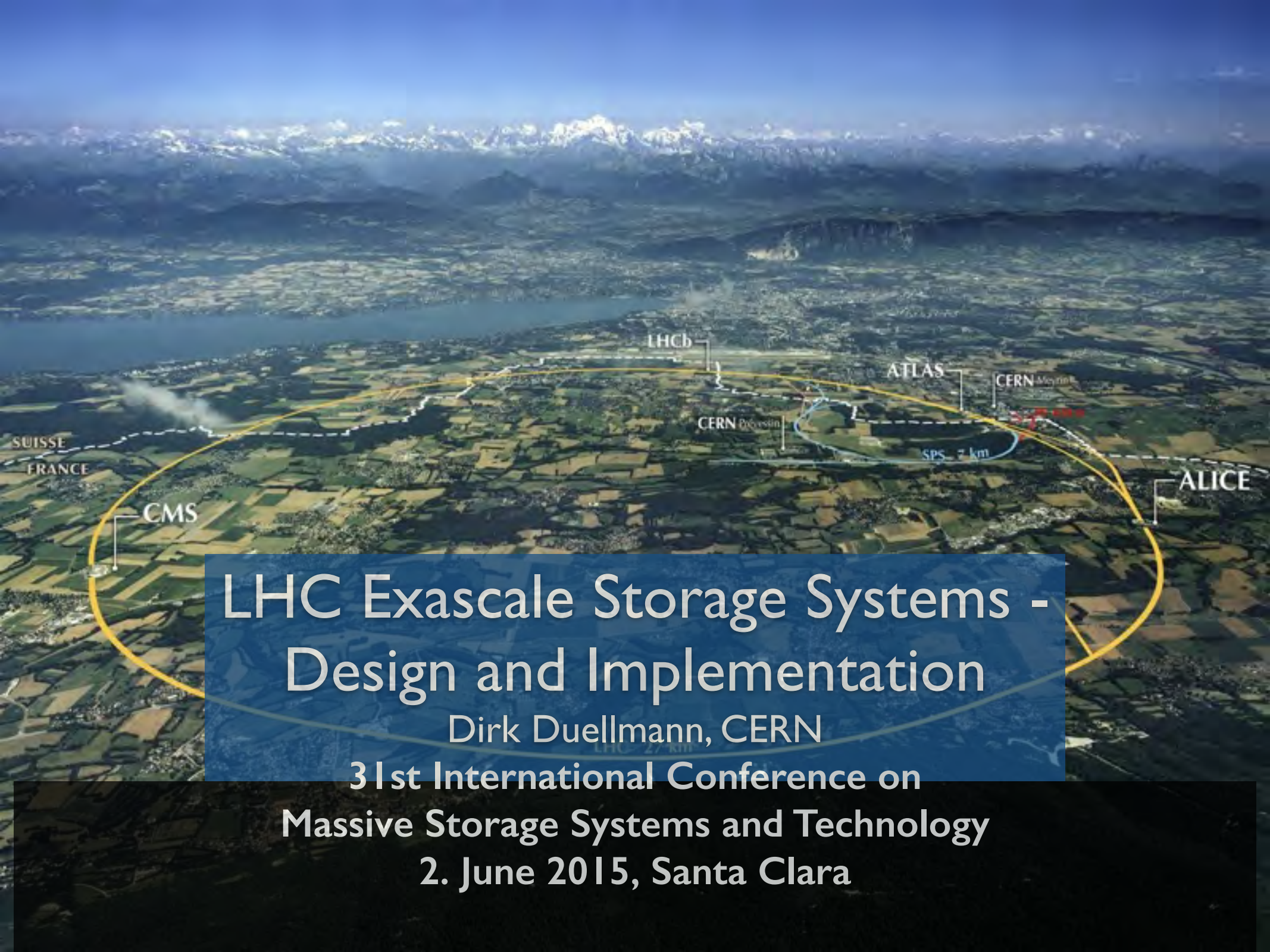


LHC Exascale Storage Systems - Design and Implementation

Dirk Duellmann, CERN

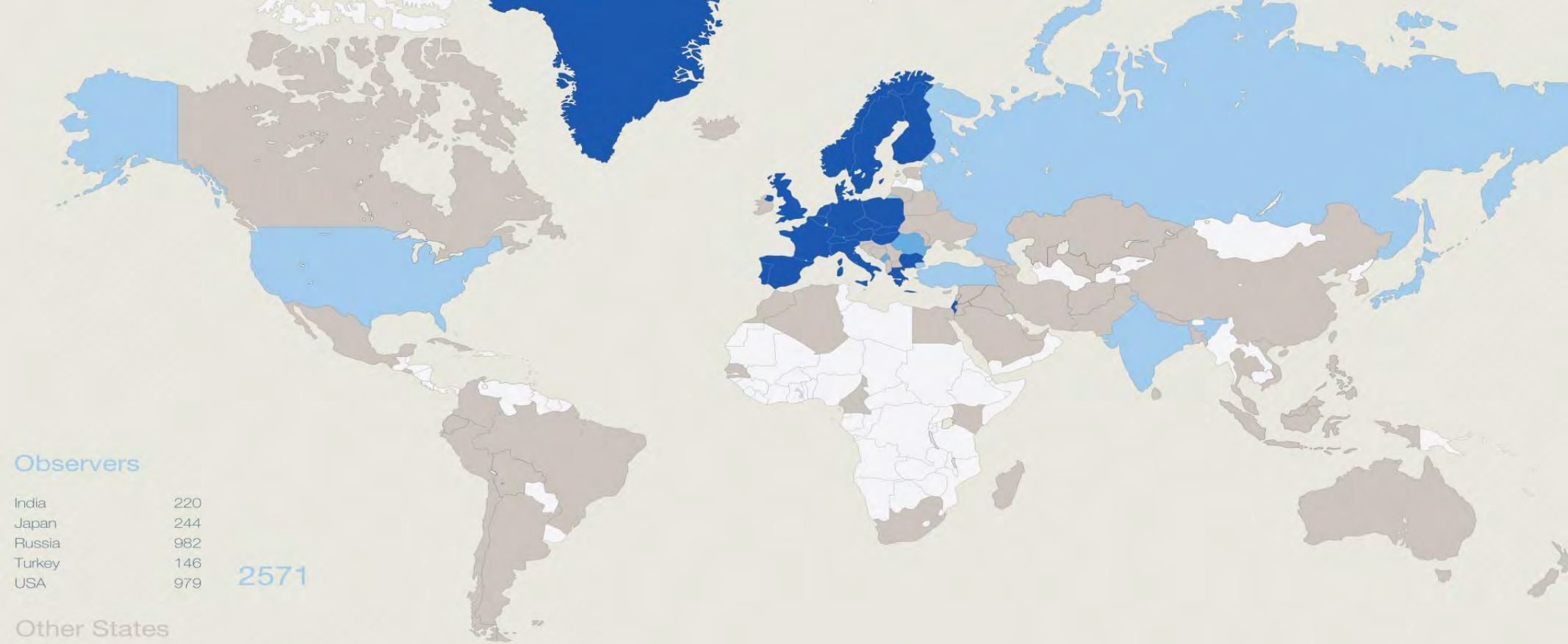
31st International Conference on
Massive Storage Systems and Technology
2. June 2015, Santa Clara

Outline

- CERN environment and challenges
  - volume, distribution, lifetime, community
- Lessons learned - Storage media developments
  - flash and shingled
- big data - more than a buzzword for science computing?
  - new market for methods used in science
  - applied science computing infrastructures
- I included slides from many colleagues in CERN's Data and Storage Services group, including G. Cancio, L. Mascetti and A. Peters

The Large Hadron Collider (LHC)

- Largest machine in the world
  - 27km, 6000+ superconducting magnets
- Fastest racetrack on Earth
  - Protons circulate 11245 times/s (99.9999991% the speed of light)



Observers

India	220
Japan	244
Russia	982
Turkey	146
USA	979

2571

Other States

Afghanistan	1	El Salvador	1	Pakistan	41
Albania	2	Estonia	16	Palestine (O.T.).	4
Algeria	8	Georgia	36	Peru	8
Argentina	11	Gibraltar	1	Philippines	1
Armenia	25	Hong Kong	1	Saudi Arabia	3
Australia	25	Iceland	4	Senegal	1
Azerbaijan	8	Indonesia	1	Singapore	2
Bangladesh	4	Iran	28	Sint Maarten	2
Belarus	47	Ireland	22	Slovenia	27
Bolivia	3	Jordan	2	South Africa	16
Bosnia & Herzegovina	1	Kenya	1	Sri Lanka	5
Brazil	108	Korea, D.P.R.	1	Syria	2
Cameroon	1	Korea Rep.	117	Thailand	12
Canada	134	Kuwait	1	T.F.Y.R.O.M.	1
Cape Verde	1	Lebanon	12	Tunisia	6
Chile	12	Lithuania	19	Ukraine	55
China	280	Luxembourg	4	Uzbekistan	4
China (Tapei)	45	Madagascar	4	Venezuela	9
Colombia	30	Malaysia	15	Viet Nam	9
Croatia	35	Mauritius	1	Zimbabwe	2
Cuba	7	Mexico	64		
Cyprus	16	Montenegro	3		
Ecuador	3	Morocco	12		
Egypt	19	Nepal	5		
		New Zealand	7		

1415

Member States

Austria	99	Greece	152	Slovakia	88
Belgium	106	Hungary	68	Spain	337
Bulgaria	75	Israel	51	Sweden	75
Czech Republic	202	Italy	1686	Switzerland	180
Denmark	53	Netherlands	153	United Kingdom	640
Finland	87	Norway	61		
France	751	Poland	229		
Germany	1150	Portugal	109		

6352

Candidate for Accession

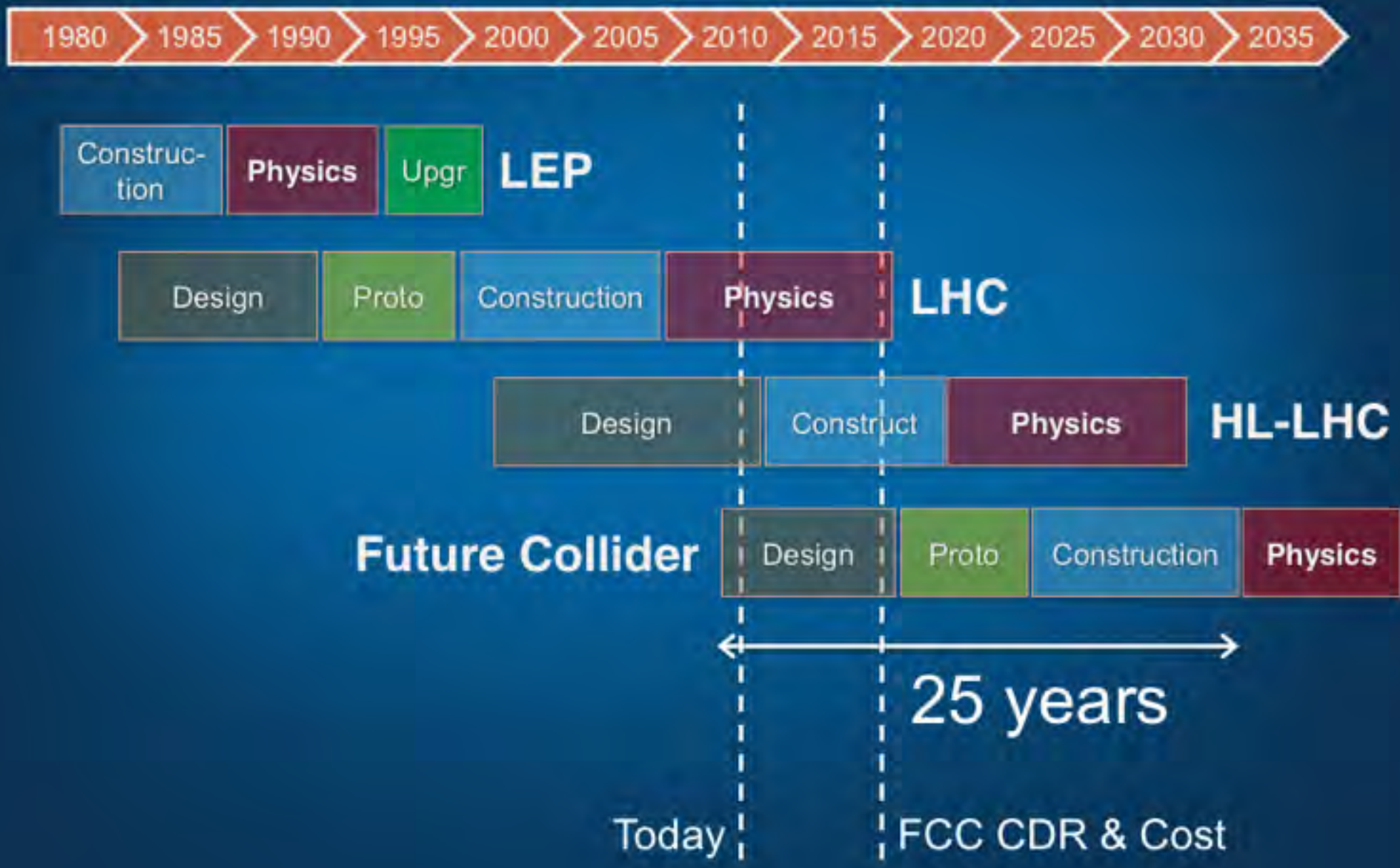
Romania	118

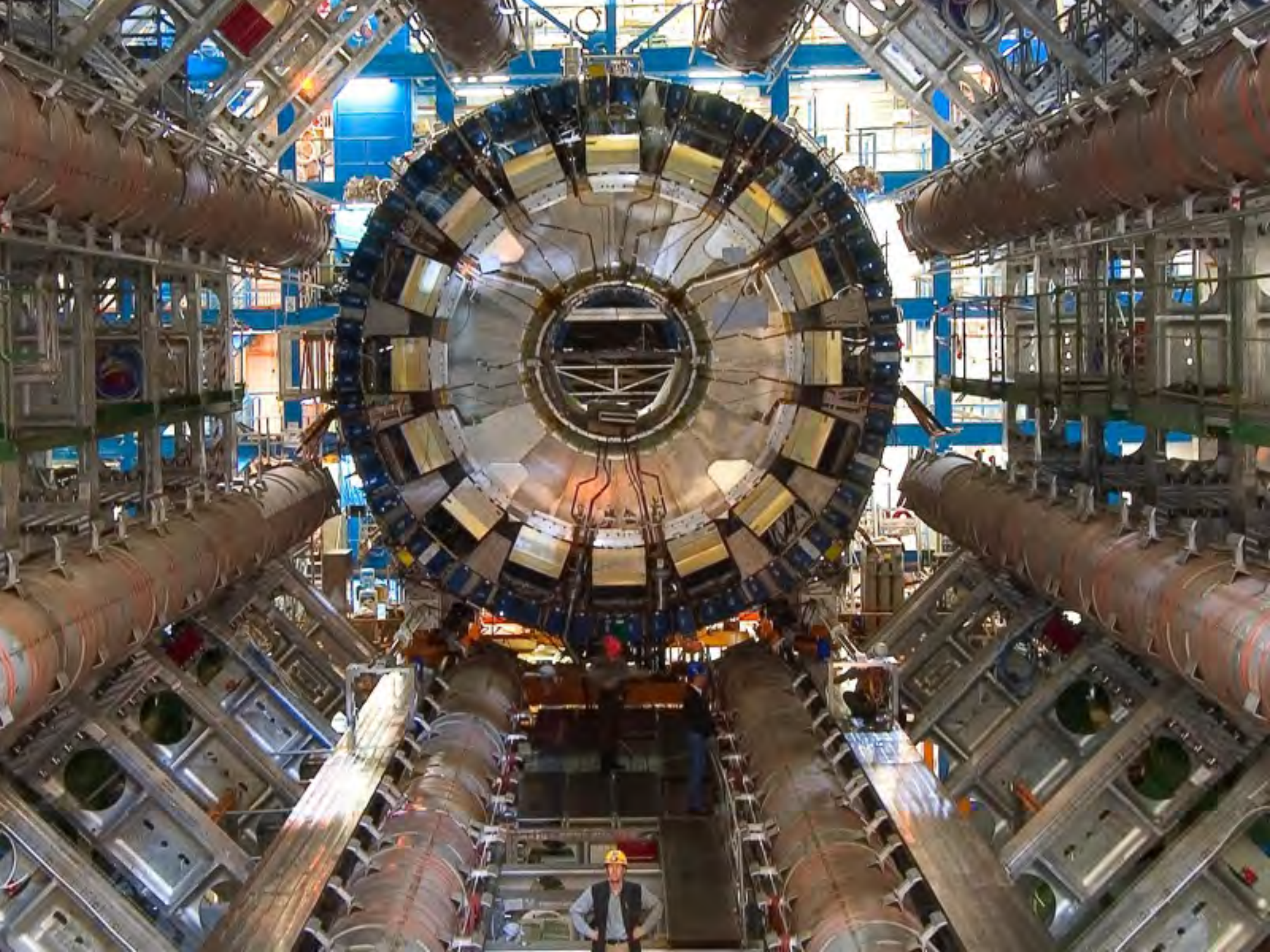
Associate Members in the Pre-stage to Membership

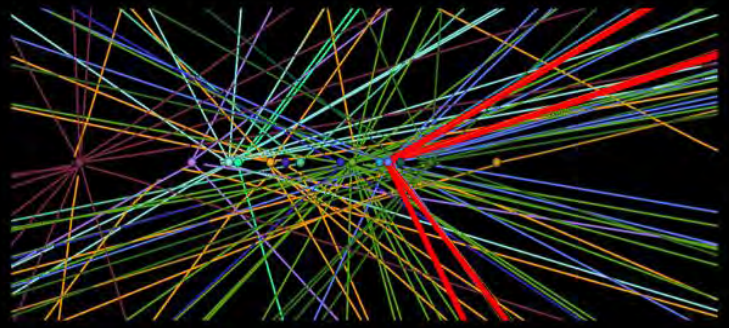
Serbia	41

Distribution of All CERN Users by Nationality on 14 January 2014

HEP Timescale



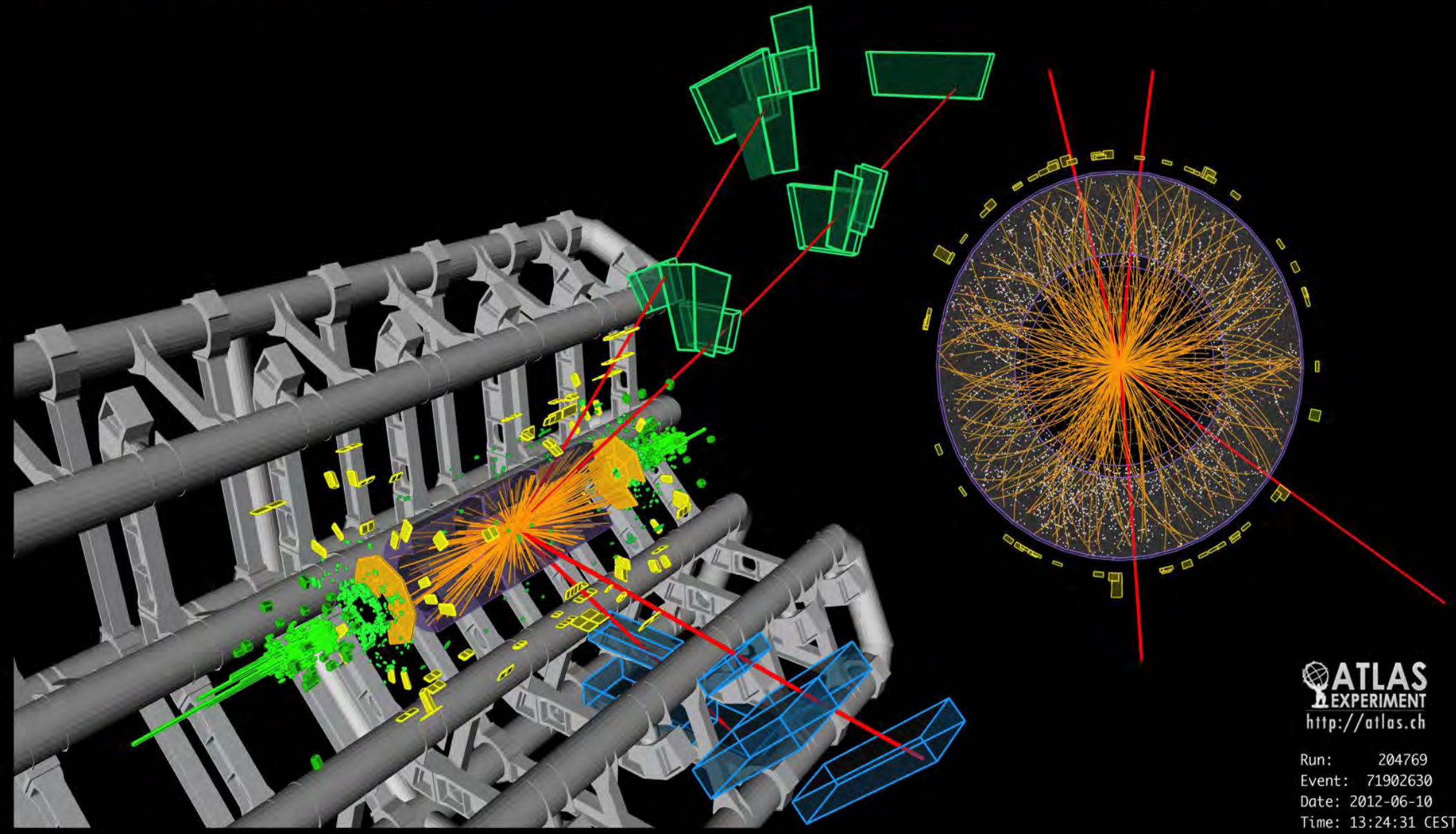




Higgs Boson Discovery

2012

Higgs to 4μ candidate event

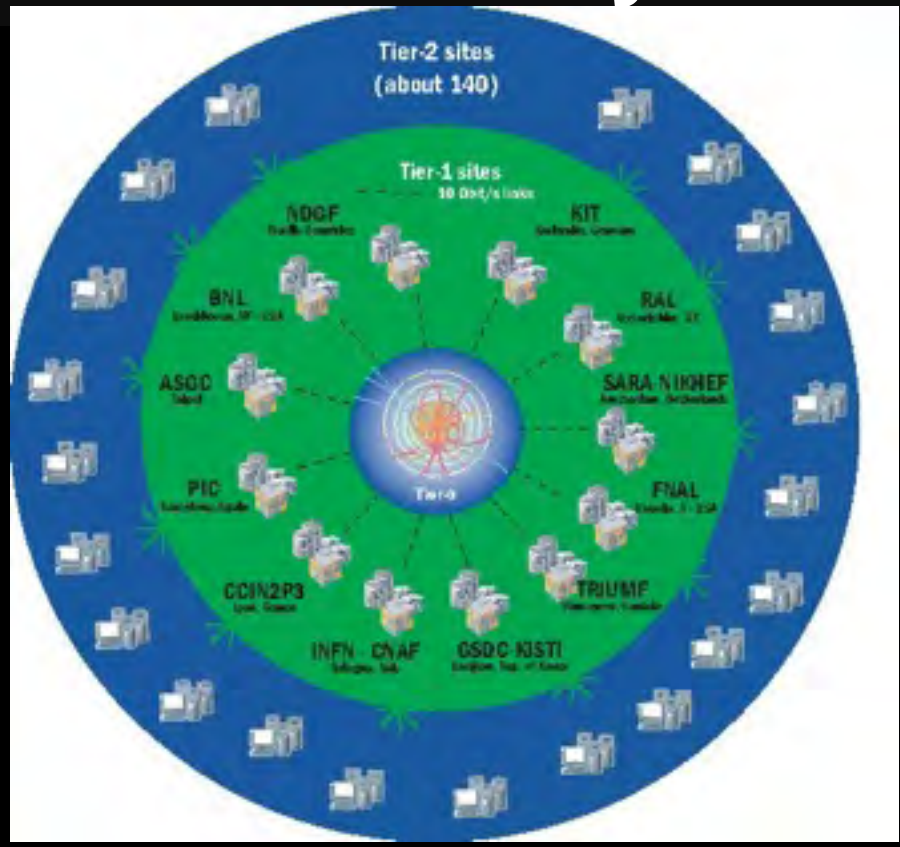


ATLAS EXPERIMENT
http://atlas.ch

Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

WLCG – what and why?

- A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments
- Managed and operated by a worldwide collaboration between the experiments and the participating computer centres
- The resources are distributed – for funding and sociological reasons
- Our task was to make use of the resources available to us – no matter where they are located



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (12 centres + Russia):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~140 centres):
- Simulation
- End-user analysis

- **~ 160 sites, 35 countries**
- **300000 cores**
- **200 PB of storage**
- **2 Million jobs/day**
- **10 Gbps links**

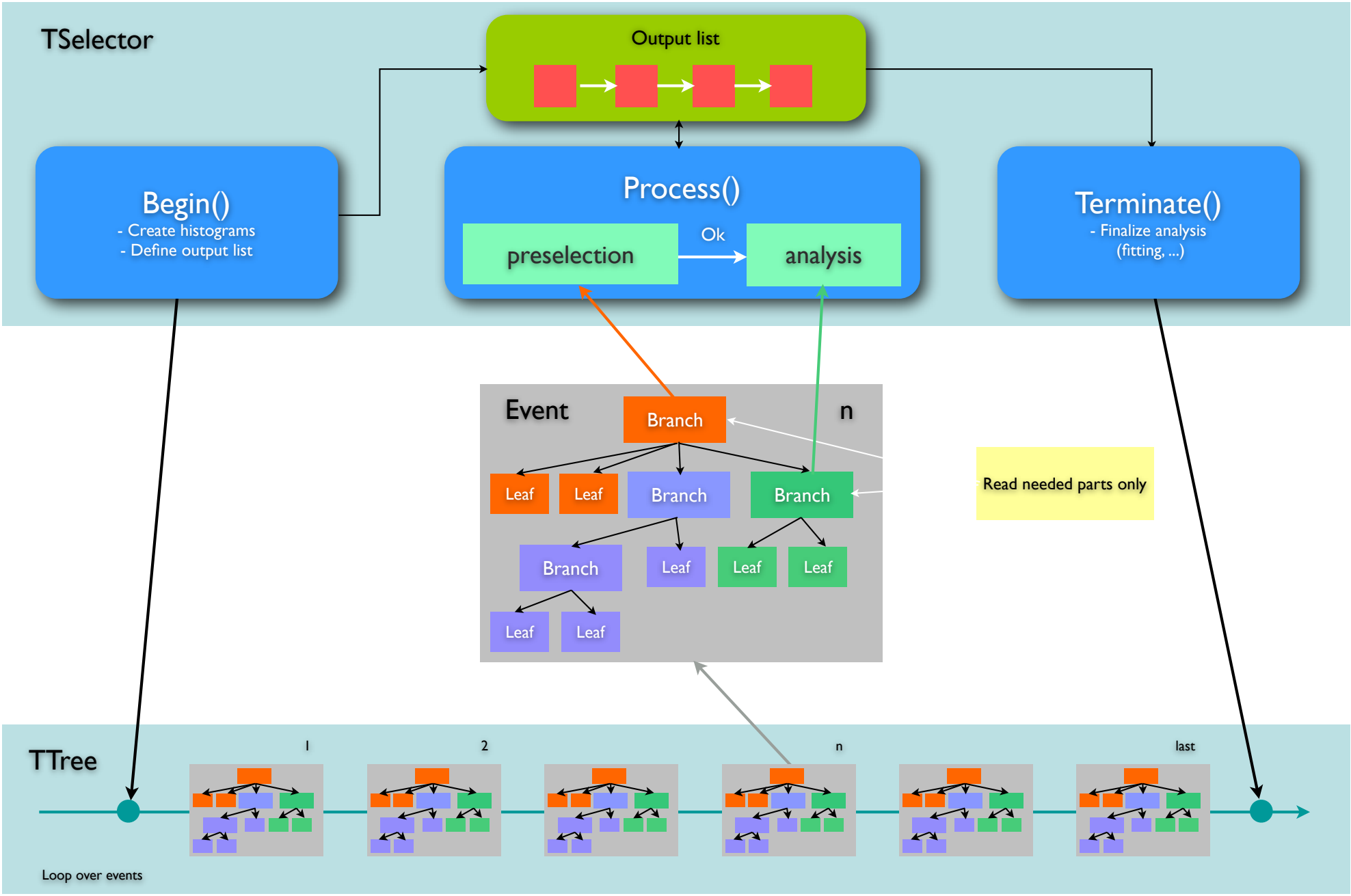
Connectivity (100 Gbps)

How to store/retrieve LHC data models? A short history…

- 1st Try - All data in an commercial Object Database (1995)
  - good match for complex data model and C++ language integration
  - used at PB scale for BaBar experiment at SLAC
  - but the market predicted by many analysts did not materialise!
- 2nd Try - All data in a relational DB - object relational mapping (1999)
  - Scale of PB deployment was far for from being proven
  - Users code in C++ and rejected data model definition in SQL

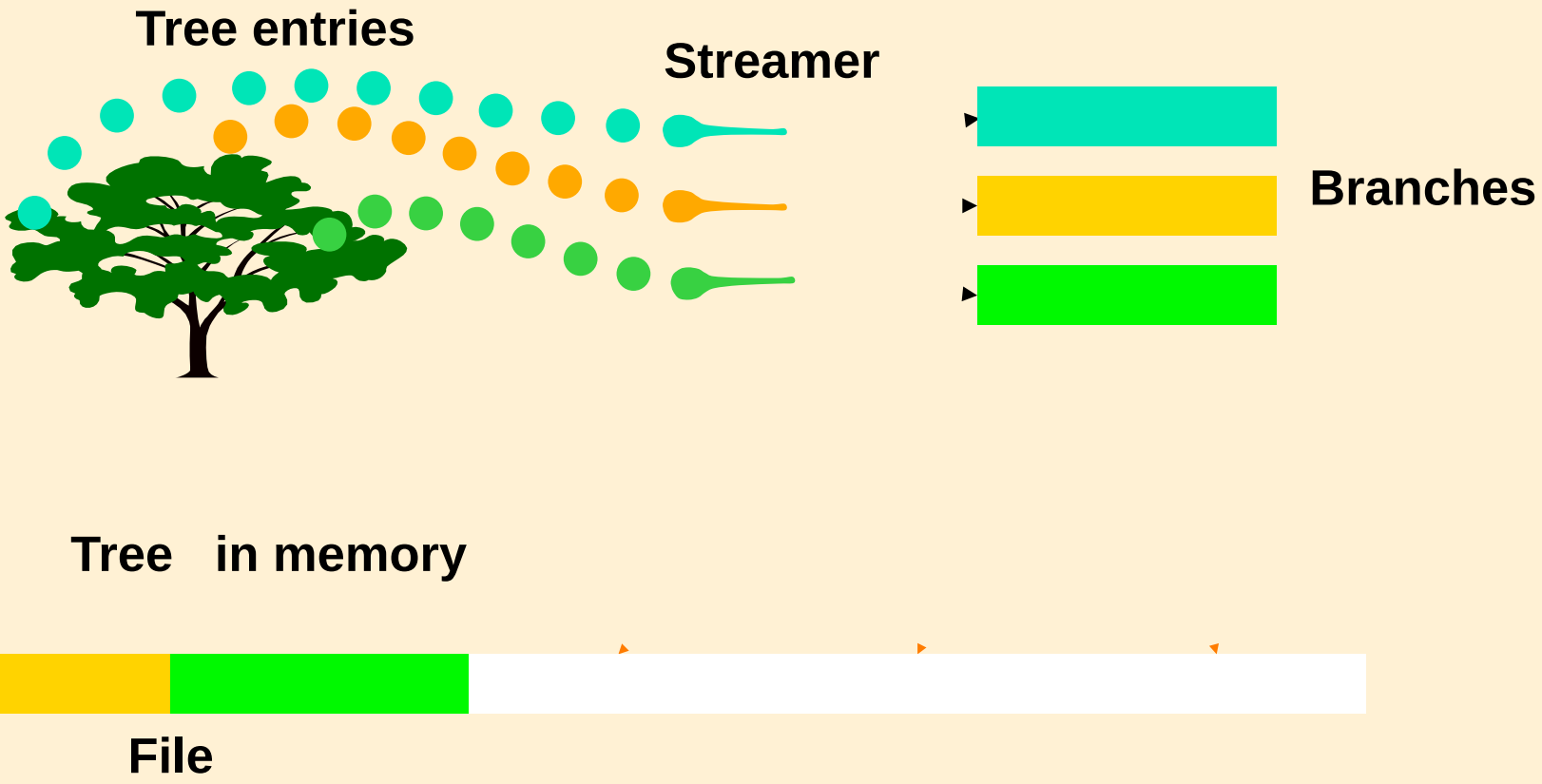
- Hybrid between RDBMS and structured files (from 2001 - today)
  - Relational DBs for transactional management of meta data (TB scale)
    - File/dataset meta data, conditions, calibration, provenance, work flow
    - via DB abstraction (plugins: Oracle, MySQL, SQLite, Frontier/SQUID)
    - see XLDB 2007 talk for details
- Home-grown persistency framework ROOT (180PB)
  - Uses C++ "introspection" to store/retrieve networks of C++ objects
  - Configurable column-store for efficient sparse reading

Processing a TTree



- Scalable, efficient, machine independent format
- Orthogonal to object model
  - Persistency does not dictate object model
- Based on object serialization to a buffer
- Automatic schema evolution (backward and forward compatibility)
- Object versioning
- Compression
- Easily tunable granularity and clustering
- Remote access
  - HTTP, HDFS, Amazon S3, CloudFront and Google Storage
- Self describing file format (stores reflection information)
- ROOT I/O is used to store all LHC data (actually all HEP data)

TTree - Clustering per Object



TTree = container for an arbitrary set of independent event trees

CERN Disk Storage Overview

	AFS	CASTOR	EOS	Ceph	NFS	CERNBox
Raw Capacity	3 PB	20 PB	140 PB	4 PB	200 TB	1.1 PB
Data Stored	390 TB	86 PB (tape)	27 PB	170 TB	36 TB	35 TB
Files Stored	2.7 B	300 M	284 M	77 M (obj)	120 M	14 M

AFS is CERN's linux home directory service

CASTOR & EOS are mainly used for the physics use case (Data Analysis and DAQ)

Ceph is our storage backend for images and volumes in OpenStack

NFS is mainly used by engineering application

CERNBox is our file synchronisation service based on OwnCloud+EOS

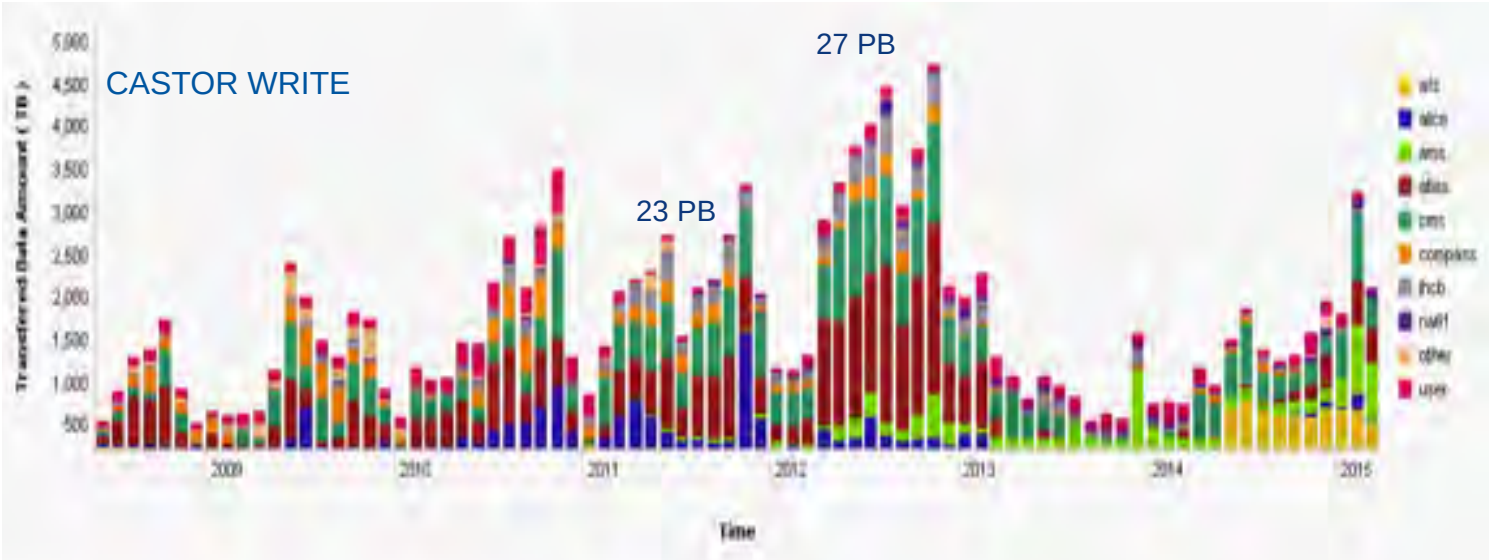
Tape at CERN: Overview

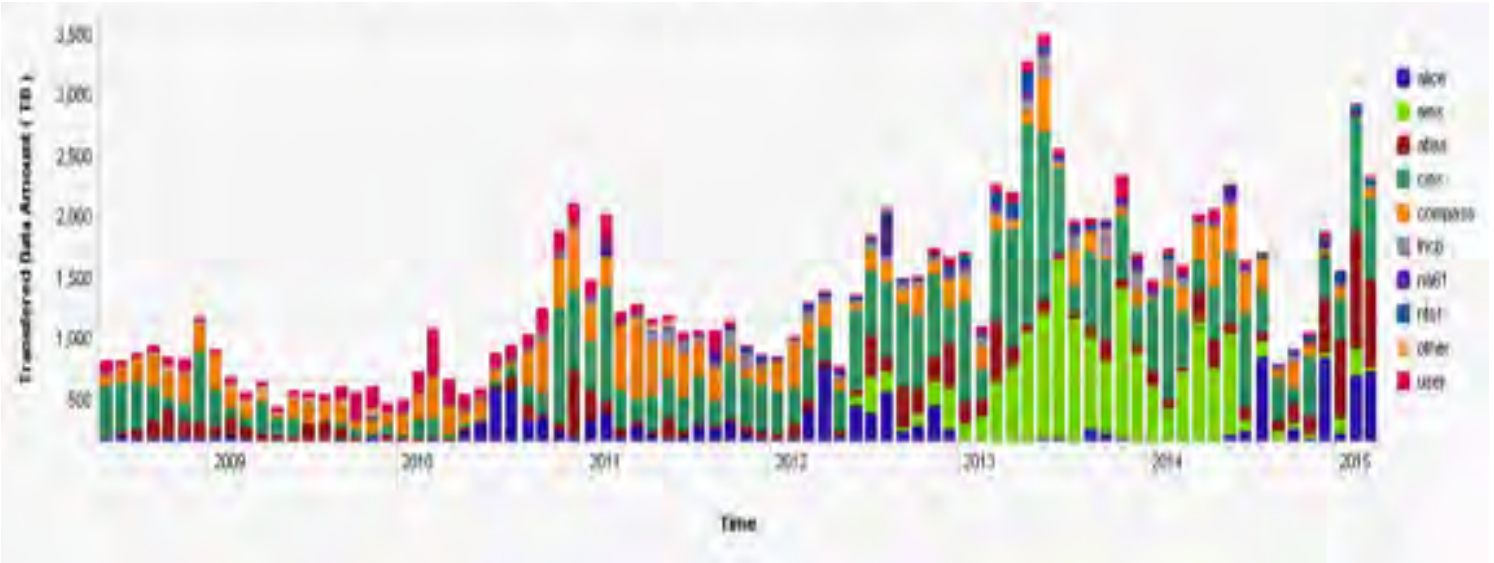
Data:
- ~100 PB physics data (CASTOR)
- ~7 PB backup (TSM)

Tape libraries:
- IBM TS3500 (3+2)
- Oracle SL8500 (4)

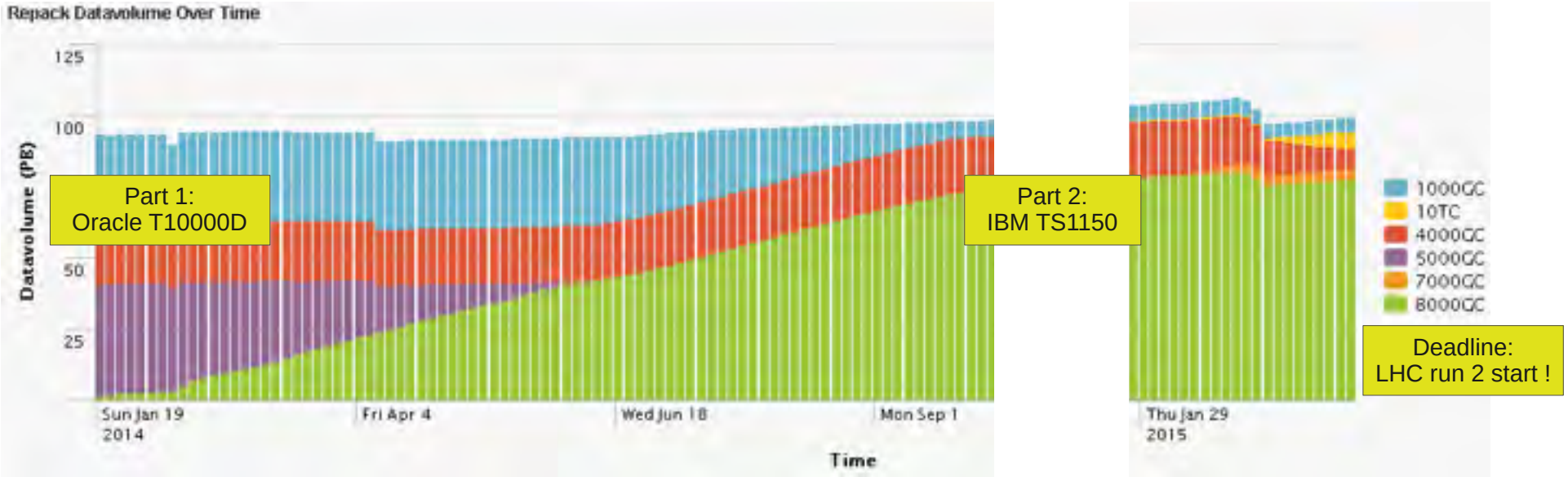
Tape drives:
- 100 archive
- 50 backup

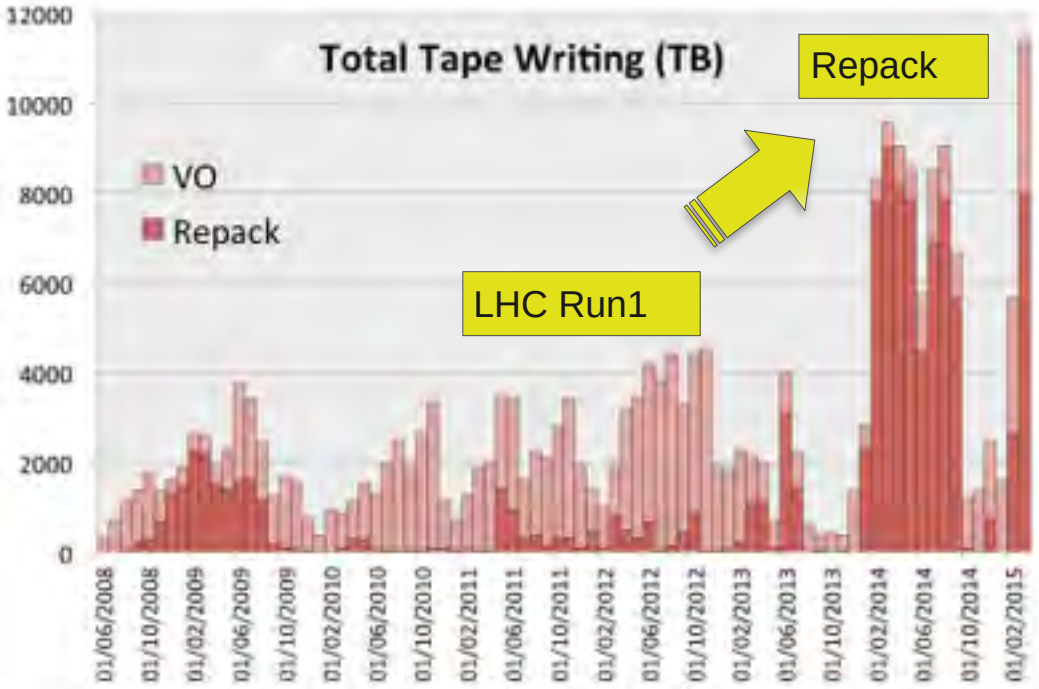
Capacity:
- ~70 000 slots
- ~30 000 tapes

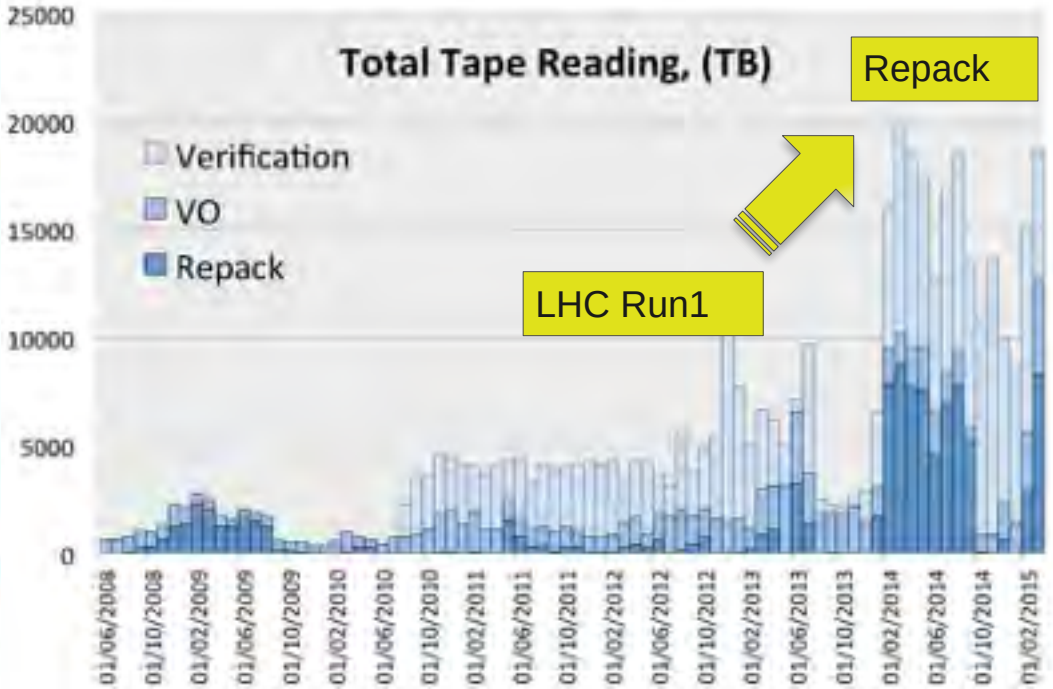




Large scale media migration





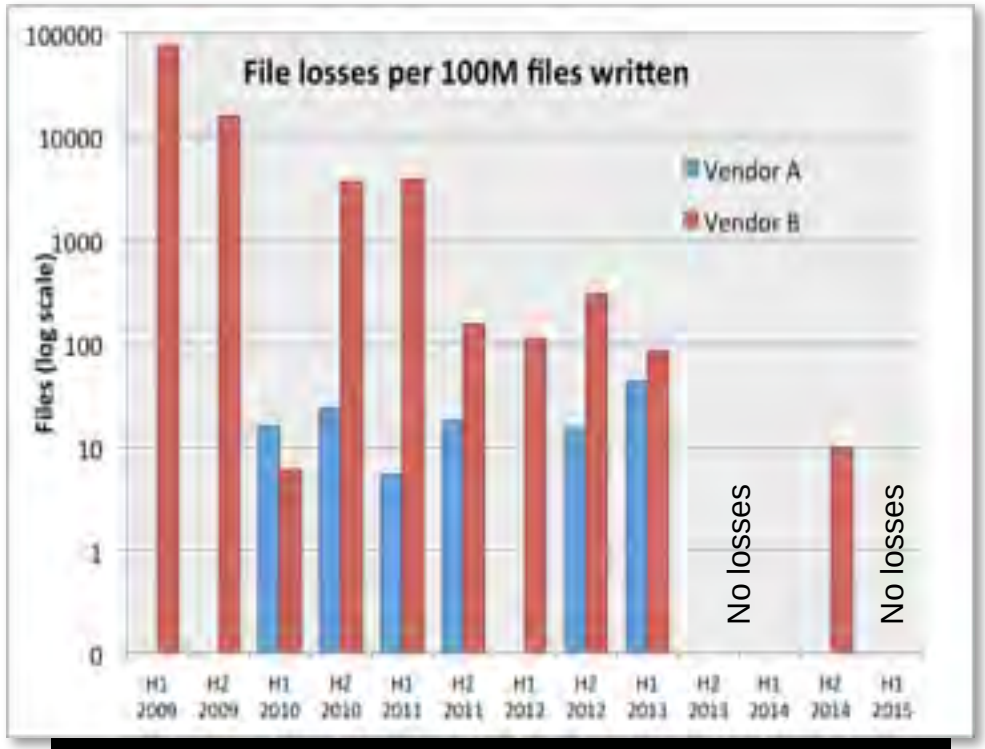


Large scale media migration

- Challenge:
  - ~100 PB of data
  - 2013: ~51 000 tapes
  - 2015: ~17 000 tapes
  - Verify all data after write
    - 3x (300PB!) pumped through the infrastructure (read->write->read)
  - Liberate library slots for new cartridges
    - Decommission ~33 000 obsolete tape cartridges
- Constraints:
  - Be transparent for experiment activities
  - Exploit the high speeds of the new tape drives
  - Preserve temporal collocation
  - Finish before LHC run 2 start

Archive Reliability

- Bit-preservation techniques to improve archive reliability
  - Annual 2012-2015 bit loss rate: O(10^{-16})
  - Systematic verification of freshly written and “cold” tapes
  - Less physical strain on tapes (HSM access, buffered tape marks)
  - With new hardware/media, differences between vendors getting small
  - For smaller experiments, creating dual copies on separated libraries / buildings



- Working on support for SCSI-4 Logical Block Protection
  - Protect against link-level errors eg bit flips
  - Data Blocks shipped to tape drive with pre-calculated CRC
  - CRC re-calculated by drive (read-after-write) and stored on media; CRC checked again on reading. Minimal overhead (<1%)
  - Supported by LTO and enterprise drives

HSM Issues

- CASTOR had been designed as HSM system
  - disk-only and multi-pool support were introduced later
  - requirements for eg aggregate catalogue access and file-open rate exceeded earlier estimates
- At LHC startup also additional conceptual issues with the HSM model became visible
  - “a file” was not a meaningful granule anymore for managing data exchange / staging: each experiment defines data sets
  - Data sets had to be “pinned” by user responsible to cache trashing
    - Large scale users had to trick the HSM logic to do the right thing

New Approach: Separated Archive + Disk Pools

- Result of splitting archive and disk pools
  - reliable user experience on disk
  - reduced archive access for priority work flows, allowed more efficient, less aggressive recall policies -> better efficiency per mount
  - simplified active code base
- Note: above reasoning may not apply to smaller user communities or smaller active data fraction in HSM

RAID Issues

- Assumption of independent drive errors does not hold
  - eg during recovery
  - drives often share also other common failure sources (power supplies, fans, network etc)
- Drive capacity increase and localised (=long) recovery result in probability for 2nd fault during recovery => data loss
- Most large scale systems departed from drive level RAID aggregation
  - but use similar concepts on a different level (eg file or chunk replication)

EOS Project: Deployment Goals

CERN IT Department

- Follow trend in many other large storage systems
  - server, media, file system failures need to be transparently absorbed
  - key functionality: file level replication and rebalancing
- Decouple h/w failures from data accessibility
  - data stays available after a failure - no human intervention
  - this has changed also our approach wrt h/w lifecycle
- Fine grained redundancy options within one h/w setup
  - eg choose redundancy level (and hence storage overhead) for specific data rather than globally
  - initially simple file replica count, more recently we added erasure encoding
- Support bulk deployment operations like retirement and migration building on lower level rebalancing
  - eg retire hundreds of servers at end of warranty period

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

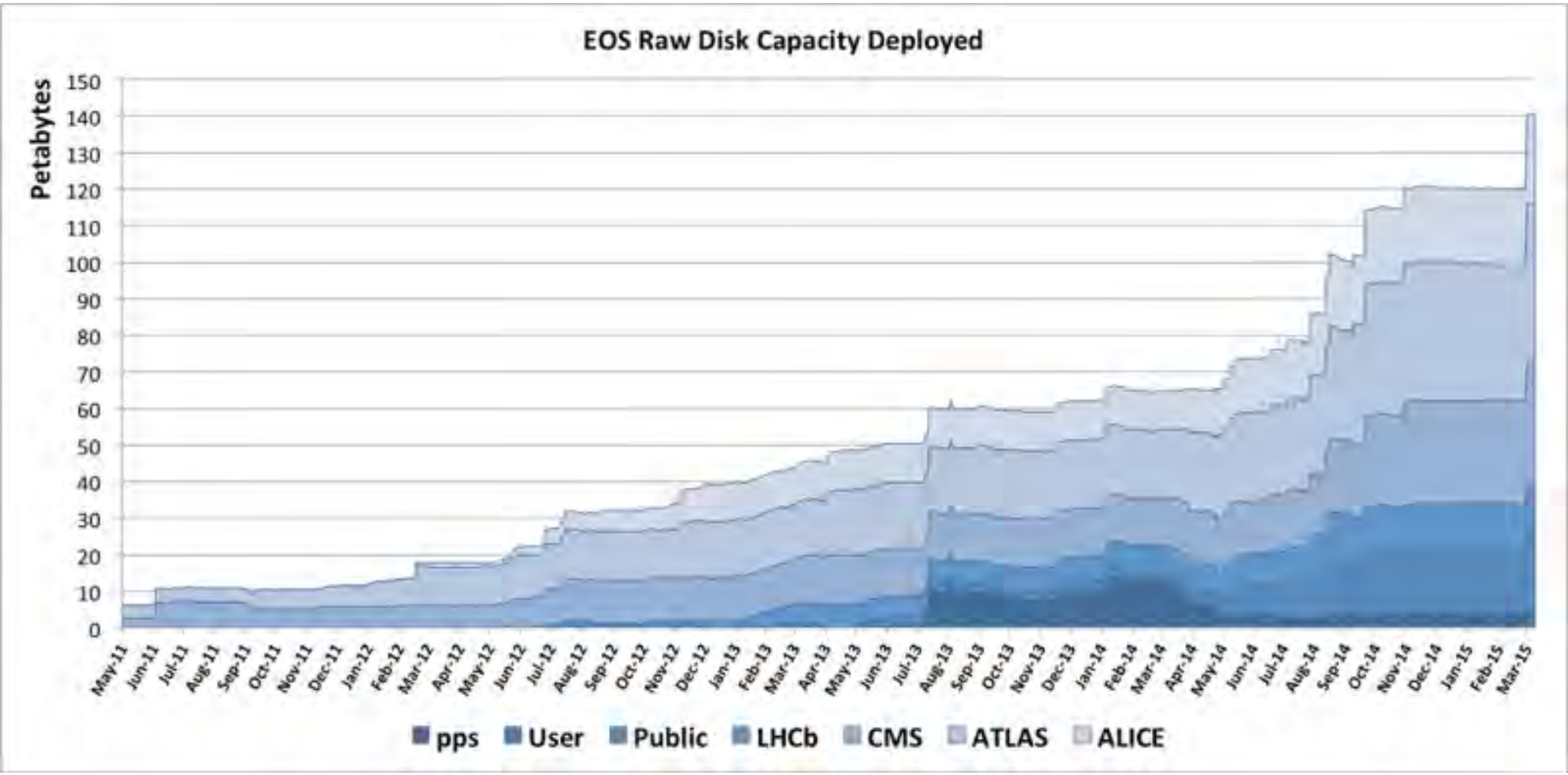
EOS Project: Technology Input

CERN IT Department

- Our archive uses an RDBMS for most internal meta-data
  - reliable - but expensive / not ideal for tree-like data
- EOS: moved to in-memory namespace with sparse hashes per directory
  - file stat calls 1-2 orders faster than on archive namespace
  - write ahead logging with periodic log compaction for persistency
  - {active work item today: namespace cold boot time}
- EOS extends mature XROOTD framework
  - re-uses reliable file access protocol with redirection and federation features
    - eg redirect user transparently if one replica goes down
  - redirection target can be at a different site
    - CERN has part of resources in Budapest (RTT: 22ms)
    - we schedule placement and replica selection for read

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

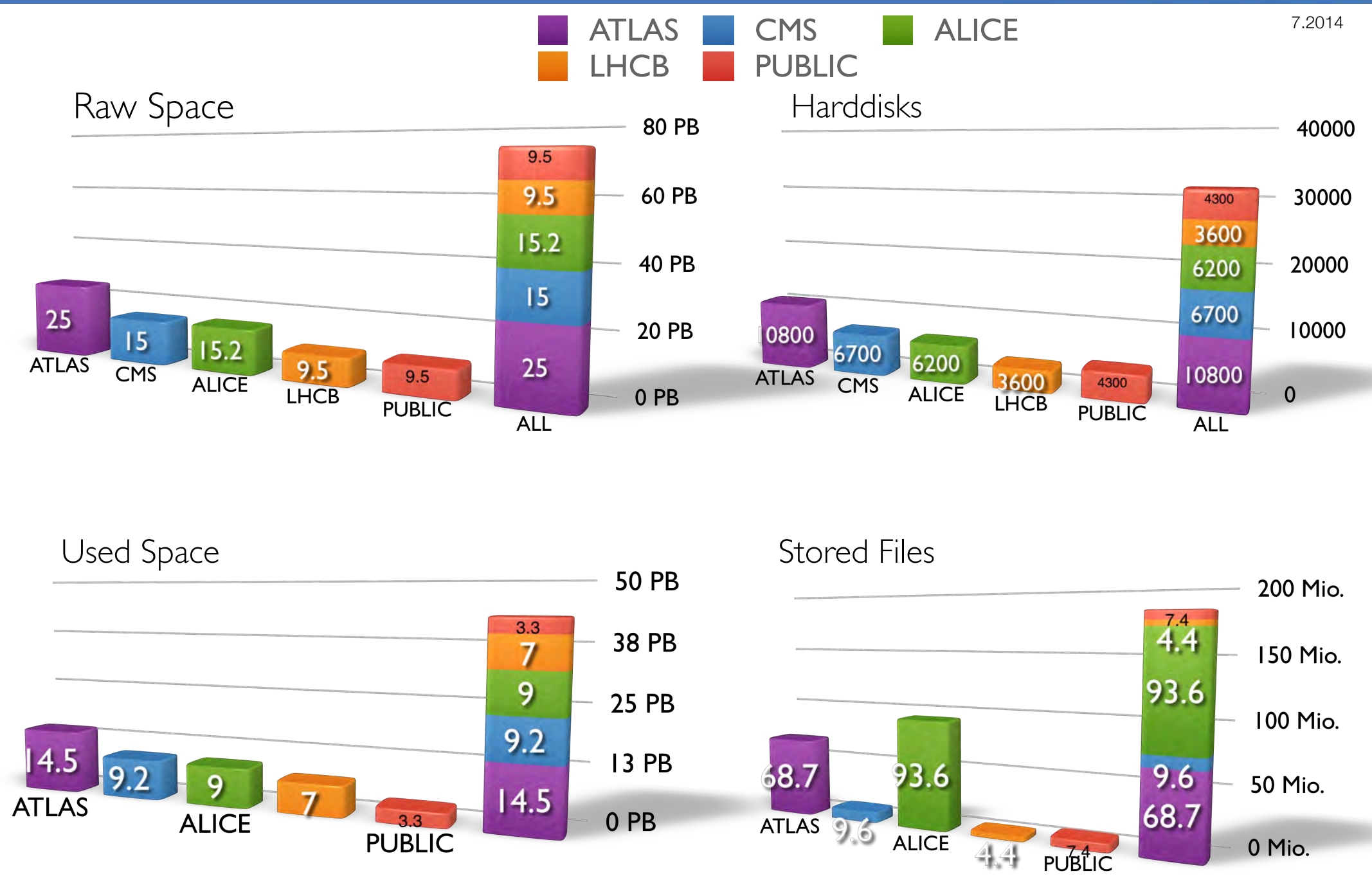
EOS Raw Capacity Evolution



EOS Deployment - Breakdown by Instance (2014)

CERN IT Department

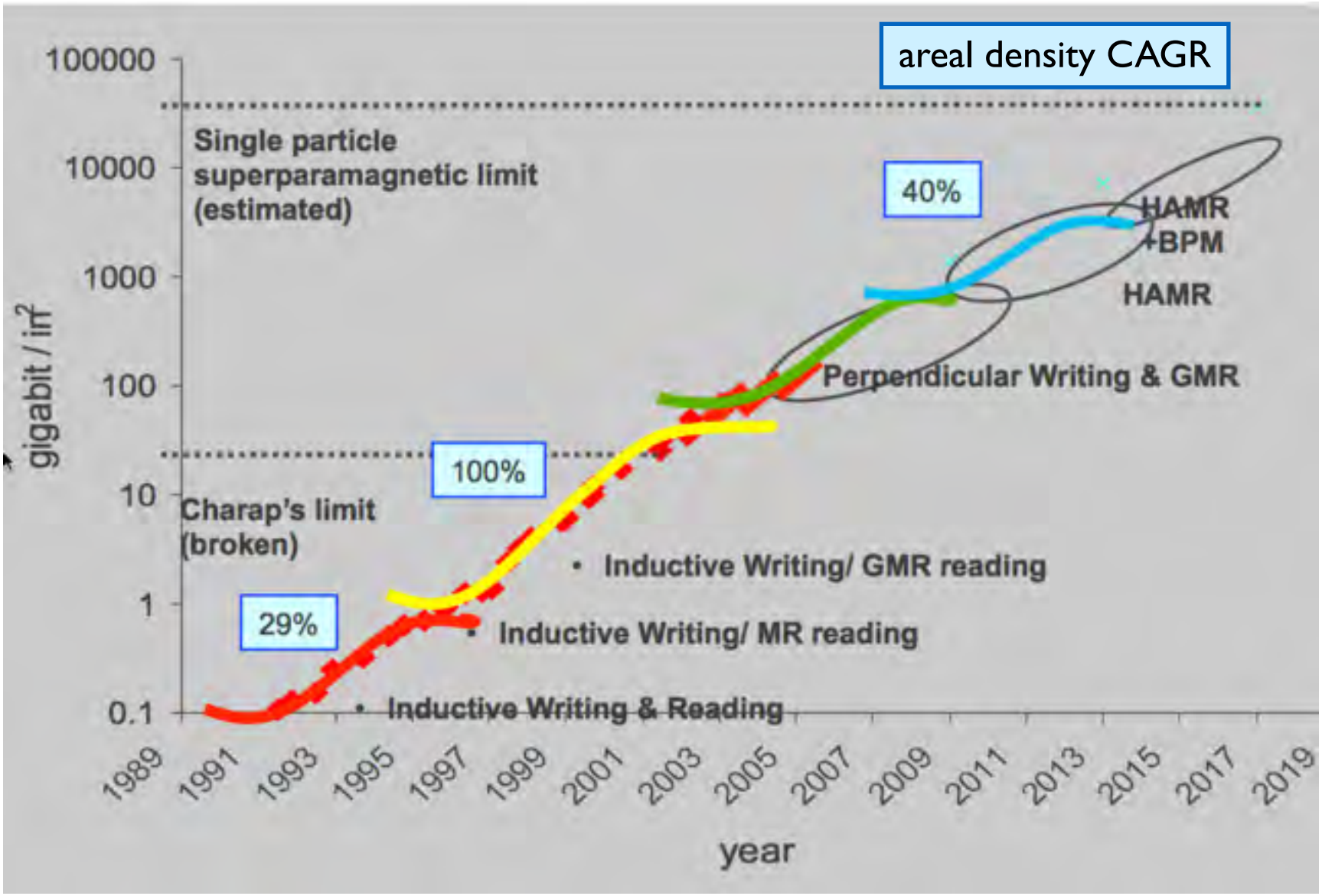
7.2014



Why do we still maintain (some of) our own storage software?

- *Large science community* trained to be effective with an agreed set of products
  - efficiency of this community is our main asset - not the raw utilisation of CPUs and disks
  - integration and specific support do matter
  - agreement on tools and formats matter even more
- *Long term* projects
  - "loss of data ownership" after first active project period
  - change of "vendor/technology" is not only likely but expected
  - we carry old but valuable data through time (bit-preservation)
- We use the same storage system as for current leading-edge projects

Does Kryder's law still hold?
What's next for disk storage?



source: HDD Opportunities & Challenges, Now to 2020, Dave Anderson, Seagate

Impact of Shingled Recording

- Gap between Read and Write performance increases
  - need to check eg if meta data mixing with data is still feasible
- Market / Application Impact
  - Will there be several types of disks?
    - emulation of a traditional disk
    - explicit band management by application
    - constraint semantics (object disk)
- Open questions:
  - which types will reach a market share & price that makes them attractive for science applications ?
  - how can the constrained semantics be mapped to science workflows?
  - CERN openlab R&D area

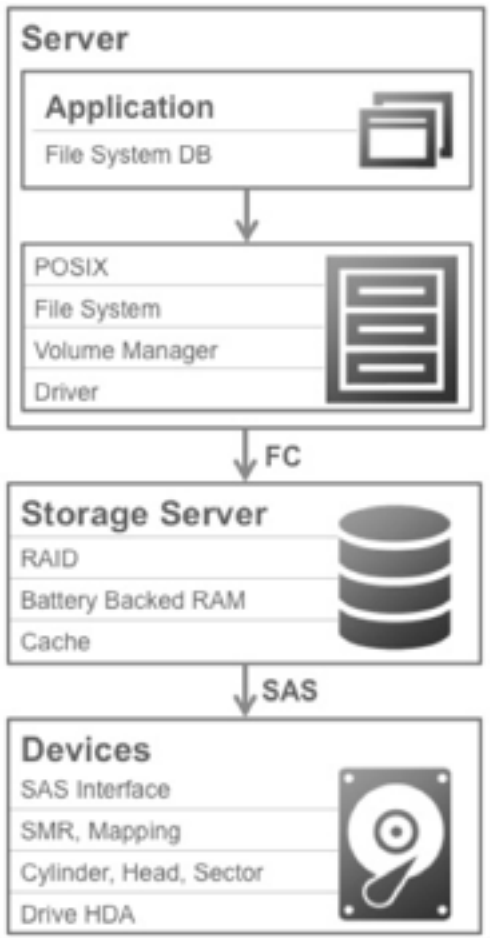
Object Disk

Seagate

- Each disk talks object storage protocol over TCP
  - replication/failover with other disks in a networked disk cluster
  - open access library for app development

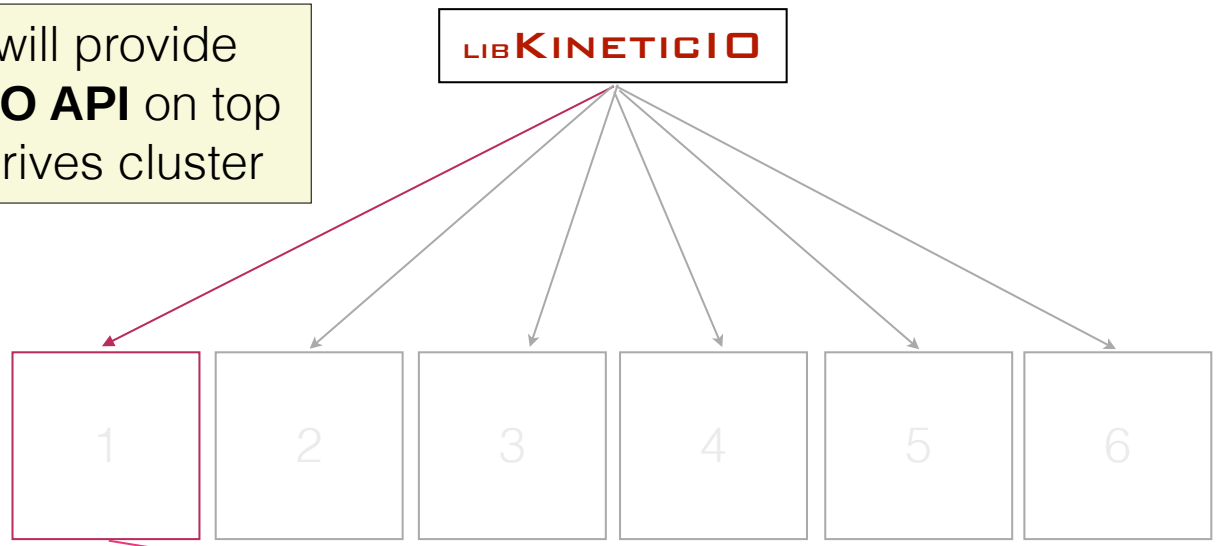
- Why now?
  - shingled media come with constrained (object) semantic: eg no updates

- Early stage with several open questions
  - port price for disk network vs price gain by reduced server/power cost?
  - standardisation of protocol/semantics to allow app development at low risk of vendor binding?



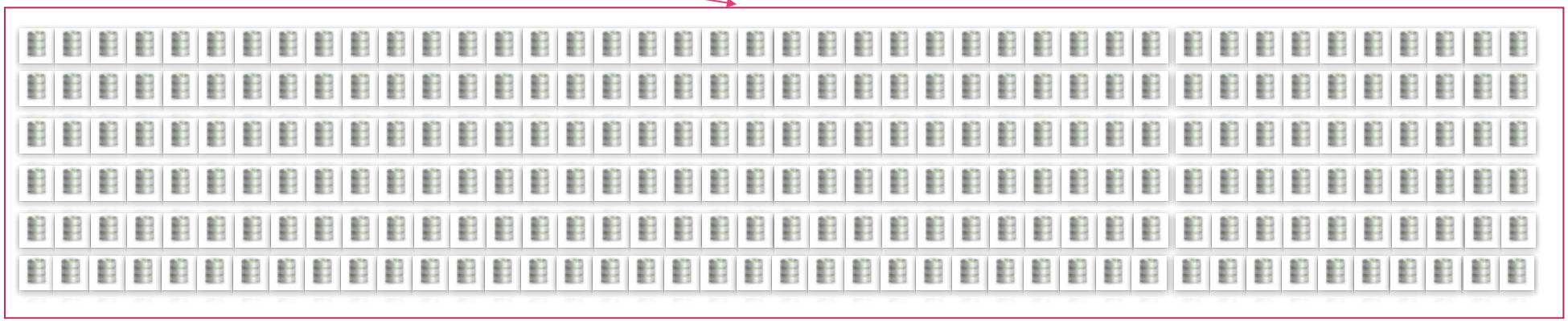
OPENLAB R&D PROJECT

libkineticio will provide a simple **file IO API** on top of a kinetic drives cluster



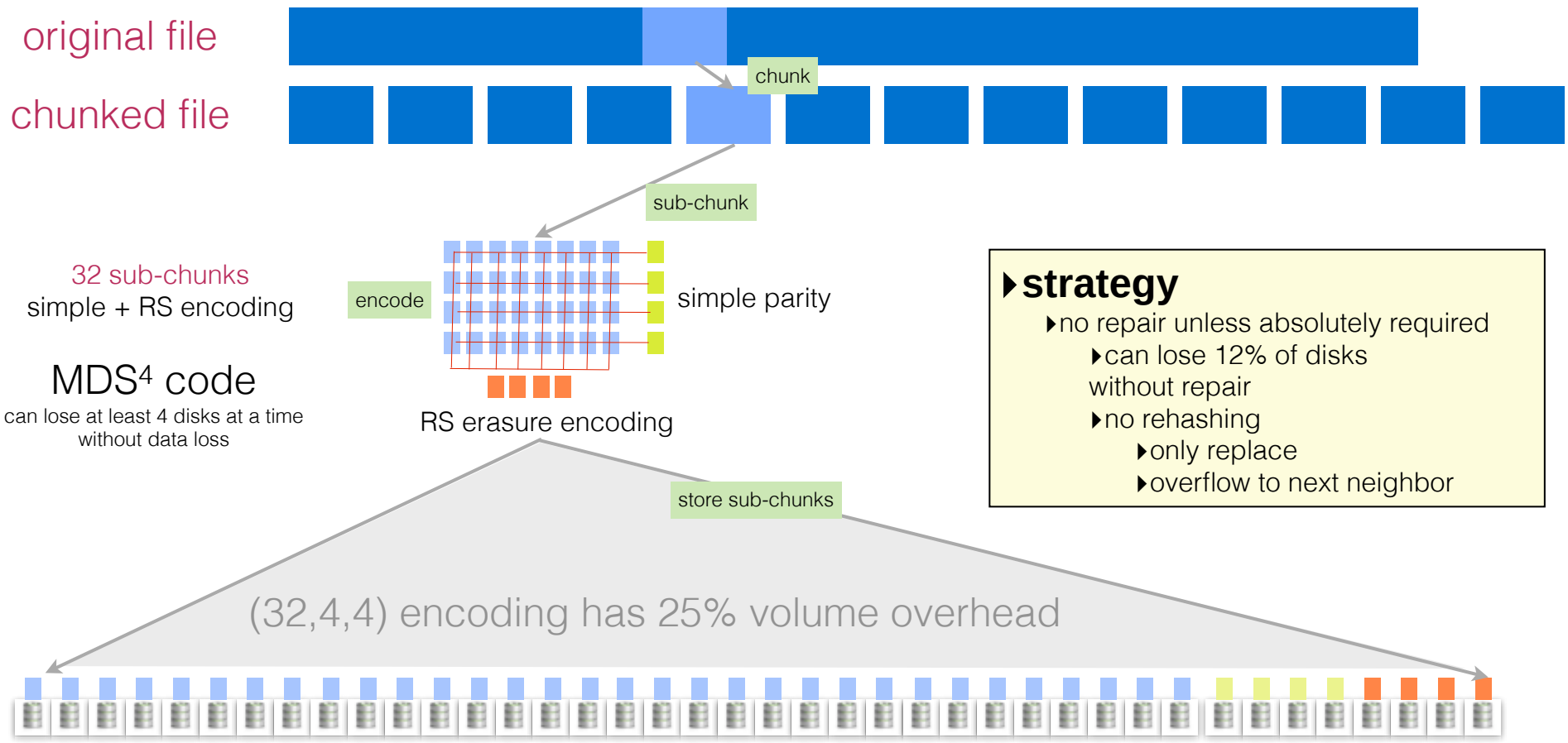
organize disks in kinetic cubes

- use hash ring inside each cube for placement
  - each cube manages redundancy internally
  - file fragments are erasure encoded
- storage index for cube selection - more flexibility - better scalability



latest kinetic drive generation
252 disks - 2 PB = 15U

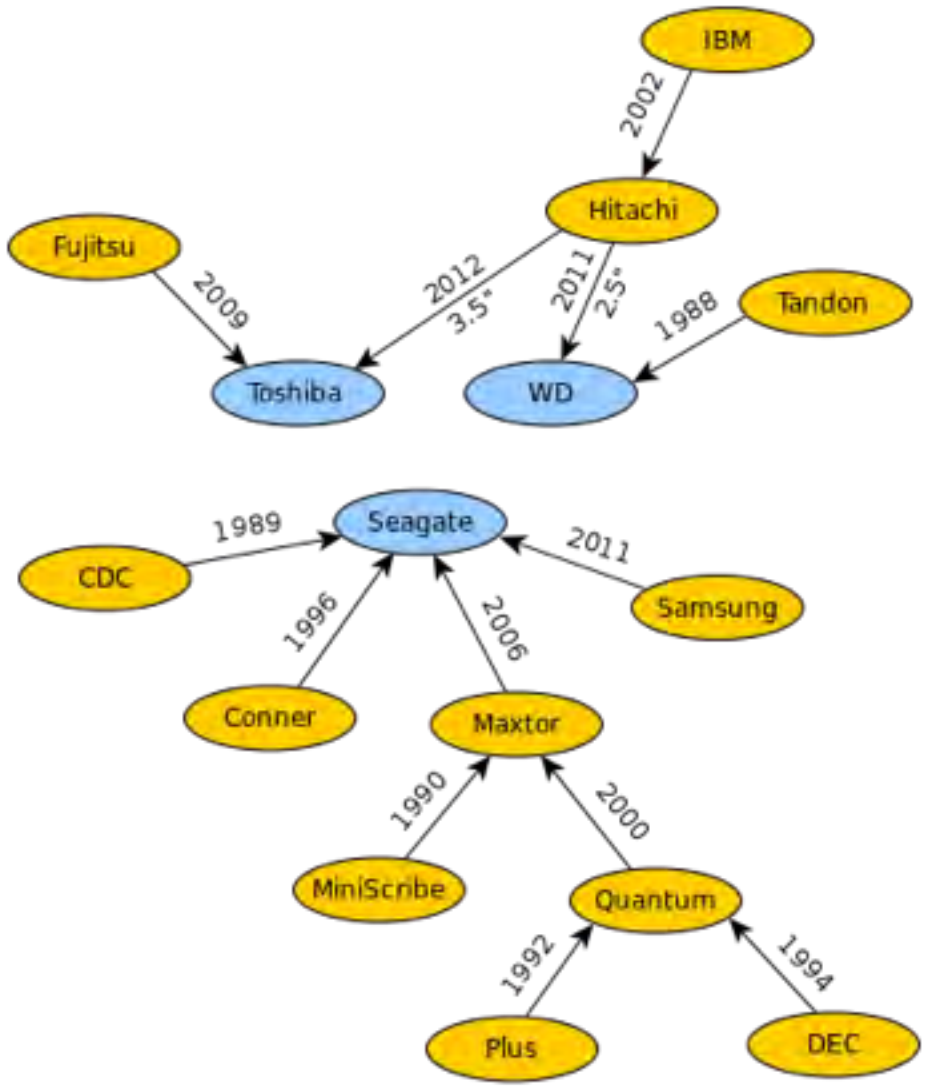
example of file encoding



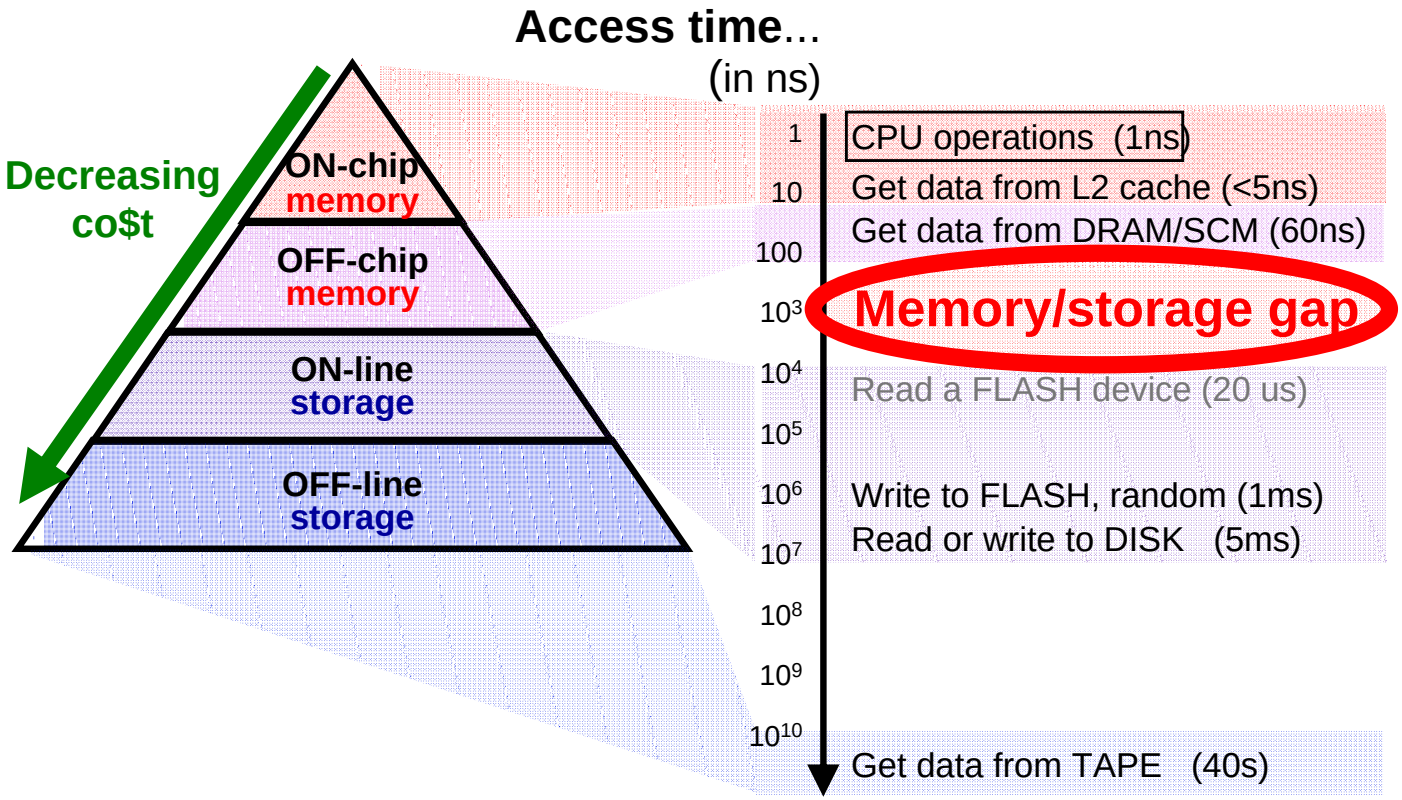
- **strategy**
  - no repair unless absolutely required
    - can lose 12% of disks without repair
    - no rehashing
      - only replace
      - overflow to next neighbor

best kinetic performance for 32M chunks = 1M sub-chunks

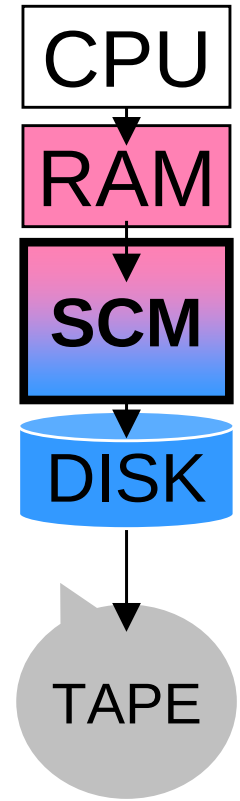
Disk Market Consolidation



Problem (& opportunity): The access-time gap between memory & storage



Near-future



Research into new solid-state non-volatile memory candidates
– originally motivated by finding a "successor" for NAND Flash –
has opened up several interesting ways to change the memory/storage hierarchy...

1) Embedded Non-Volatile Memory – low-density, fast ON-chip NVM
2) Embedded Storage – low density, slower ON-chip storage

3) M-type Storage Class Memory – high-density, fast OFF- (or ON*)-chip NVM
4) S-type Storage Class Memory – high-density, very-near-ON-line storage

* ON-chip using 3-D packaging

Flash: undesired side-effects

- asymmetric read/write performance
- write amplification : factor between user data and resulting flash memory changes
- block recycling : large internal trafic limits client transfers
- past writes influence future performance :
  eg benchmarks on new SSDs have only limited value
- limited durability (!= endurance)

SSD vs HDD

- SSD is less well defined and fragmented market
  - Large (factor 20) spread in performance and price
  - Several orders of magnitude more IOPS
    - current consumer SSDs reach 100k IOPS
  - Still O(10) higher price/GB
  - Better power efficiency - in particular for idle storage
- Still a niche solution in the data centre context
  - "Hot" transactional logs from databases or storage system meta-data (eg CEPH journals)
  - Investigating SSD use as container for EOS namespace

R&D: non-volatile memory.. but how?

- still early days for products, but software integration can already be prototyped
  - transactional memory
  - use an SSD-based filesystem
- discussing CERN openLab project on NV-RAM based catalogue with Data Storage Institute (DSI) Singapore

CERN openlab
HUAWEI
intel
ORACLE
SIEMENS
CISCO
rackspace
SEAGATE
BROCADE
IDT
Yandex
GSI
EMBL-EBI

How can we optimise our systems further?

- Storage analytics project
  - apply statistical analytics to storage (and cpu side) metrics
  - measure quantitative impact of changes on real jobs
  - predict problems and outcome of planned changes
- Non-trivial because
  - needs complete set of system and user side metrics
  - some established metrics are easy to obtain but not suitable for full analysis of a large ensemble with different & varying workloads
    - cpu efficiency = cpu/wall
    - storage performance = GB/s
  - correlation in time does not imply causal relationship

Initial Findings & Surprises

- are hidden / unknown (ab)use patterns
  - really hot files - replicate file and access paths
  - really remote files - some users try working via the WAN (eg 120 ms RTT without using vector reads etc)
  - really broken sw - user writing 1PB a day in two replicas without noticing or running into any quota issues
- Neither users, nor experiments nor system maintainers have an easy time to spot even significant optimisation options in large distributed storage setups
  - Started expert working group across IT services and experiments

Tape libraries

CERN IT Department

CERN IT Department
CH-1211 Genève 23
Switzerland

Tape libraries specificities

CERN IT Department

Tape libraries are highways for airflows:

- Drive 0.57 m³/min
- DC PSU 0.71 m³/min
- Rack module 13.59 m³/min
- Electronics module 4.42 m³/min

Total per SL8500 library:
10x0.57+14x0.71+13.59+4.42=**33.65m³/min**

Operating environment: **ISO 14644-1 Class 8** environment (particles/m³)

Class	>0.5 um	>1 um	>5 um
8	3 520 000	832 000	29 300

CERN IT Department
CH-1211 Genève 23
Switzerland

Prototype 1

CERN IT Department

Using components lying around in my office

Raw sensor in serie with Dylos laser particule counter (same airflow)

Rpi collecting/logging data on SD card. Automatically connects on CERN WIFI when powered

CERN IT Department
CH-1211 Genève 23
Switzerland

Prototype 2

2 Channels RAW sensor

Arduino mega 2560 upgrade

CERN IT
Department
CH-1211 Genève
23
Switzerland

Prototype 3

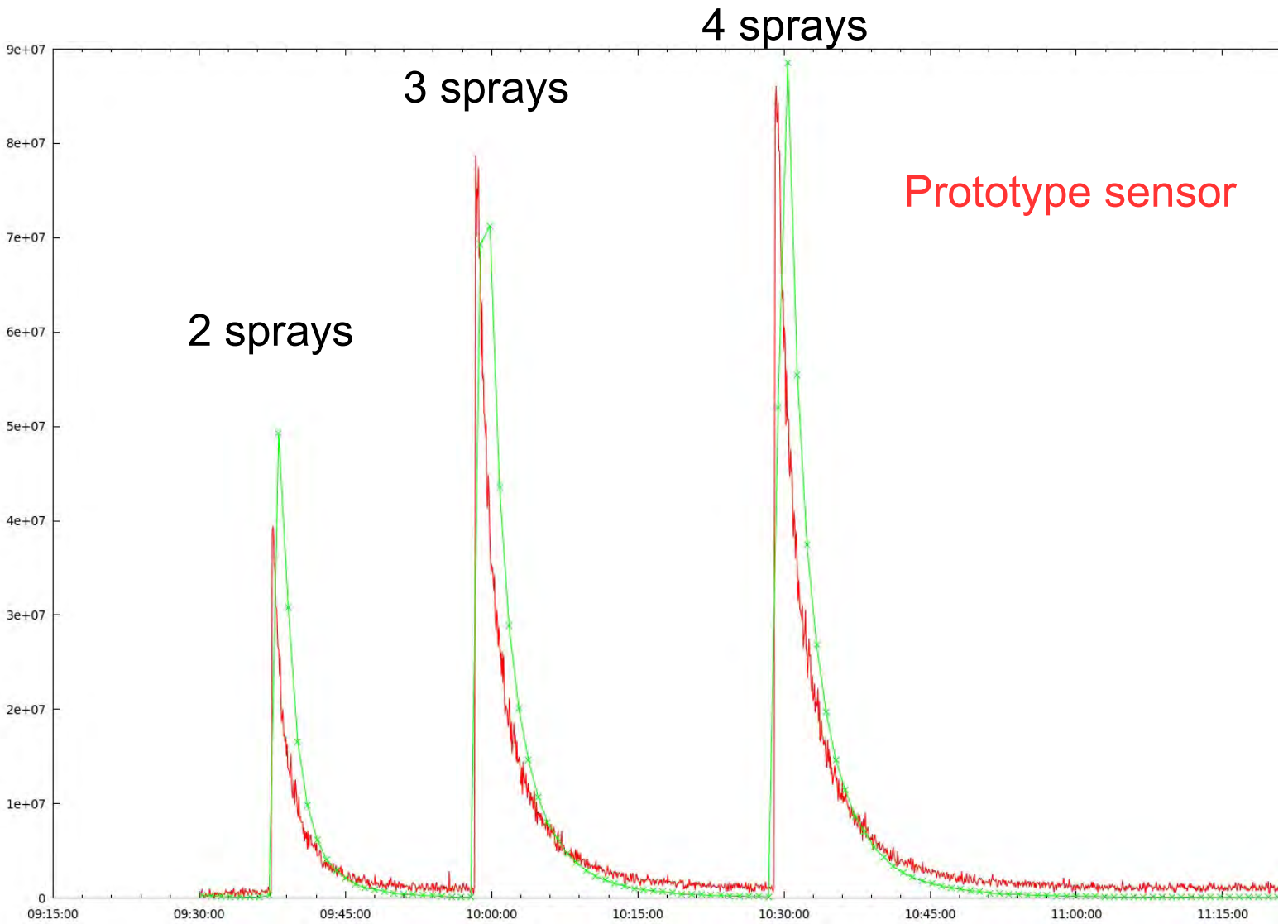
Soldered components on an arduino shield

Everything nicely packaged for rough environments

CERN IT Department
CH-1211 Genève 23
Switzerland

Calibration

CERN IT Department



> 0.5 um particles per m³

CERN IT Department
CH-1211 Genève 23
Switzerland

Future work

CERN IT Department

Integration work in the tape library

- Integrate the sensor in drive trays using onboard connectivity

Integration in collaboration with Oracle:

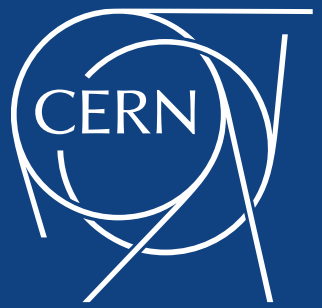
- **Do not want to void warranty**
- Regular technical meetings with Oracle hardware designers

Internship starting on April 1st for 2.5 month (student in applied industrial electronics)

CERN IT Department
CH-1211 Genève 23
Switzerland

Summary

- CERN has a long tradition in deploying large scale storage systems for a large and distributed science community world-wide
- During the first LHC run period we have passed the 100 PB mark at CERN and more importantly have contributed to the rapid confirmation of the Higgs boson and many other LHC results
- During the first deployment phase and the recent shutdown we have adapted the storage and data management models and significantly upgraded & optimised the infrastructure.
  - these changes are only possible as close collaboration between service providers and user responsible
  - the next round of optimisations will need more quantitative analytics
- We are ready for higher demands of RUN2 and confident to keep up with even more demanding LHC-HL upgrades



www.cern.ch

Thank you! Questions?