CERN Data Archive

Operational Challenges while going to Exa-Scale

Vladimír Bahyl, IEEE MSST2010

Outline

- Introduction to CERN
- Our Requirements
- Architecture
- Solutions vs. Challenges
  - Disk and Tape details
  - Monitoring
- Exa-Scale and us
- Open Questions
- Conclusion

Introduction to CERN

- **C**onseil **E**uropéen pour la **R**echerche **N**ucléaire
  - European Laboratory for Particle Physics Research
- 20 member states, 8 observers, 36 non-members
- Budget: ~1 billion CHF (~950 million USD)
- Personnel: 2600 Staff, 800 Fellows and Associates, 9000 Users from 562 Institutes in 80 countries

Fundamental Physics Questions

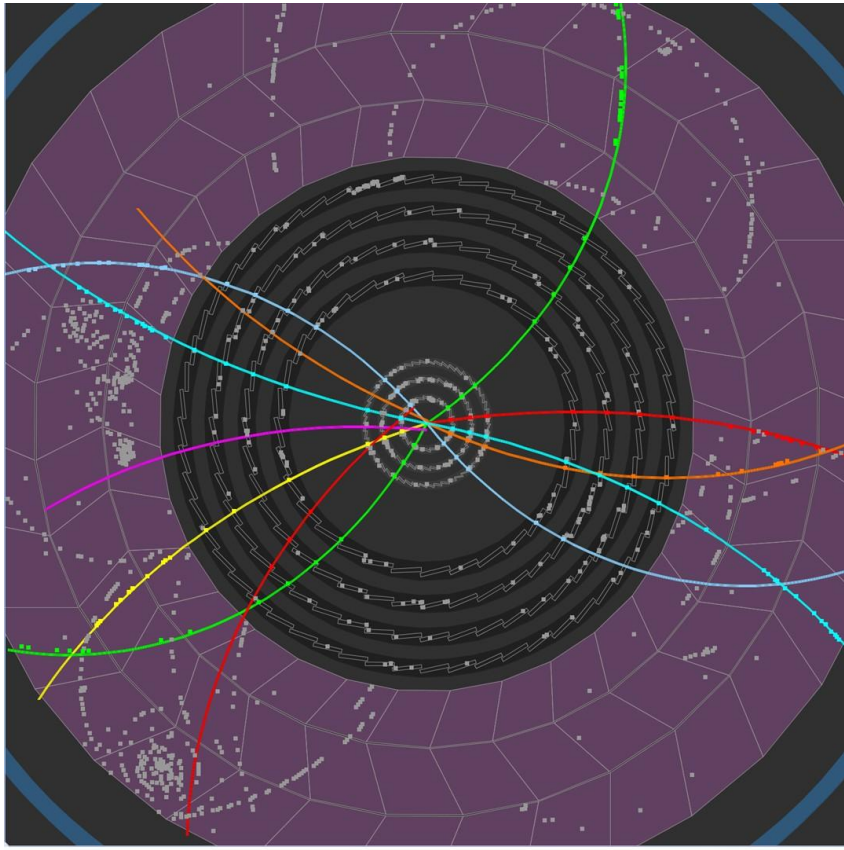
- Why do particles have mass?
  - Newton could not explain it - and neither can we…
- What is 96% of the Universe made of?
  - We only know 4% of it!
- Why is there no antimatter left in the Universe?
  - A proof that Nature is not symmetrical
- What was matter like during the first second of the Universe's life, right after the "Big Bang"?
  - A journey towards the beginning of time

CERN Tools

- The world's most powerful particle accelerator: LHC
  - A 27 km long tunnel filled with high-tech instruments
  - Equipped with thousands of superconducting magnets
  - Accelerates particles to energies never before obtained
- 4 very large sophisticated detectors
  - Hundred million measurement channels each
  - Data acquisition systems processing Petabytes per second
- Top level computing to distribute and analyse the data
  - Sufficient computing power and storage to handle massive amounts of data, making it available to thousands of physicists for analysis
  - A Computing Grid linking ~200 computer centres around the globe

What kind of data ?

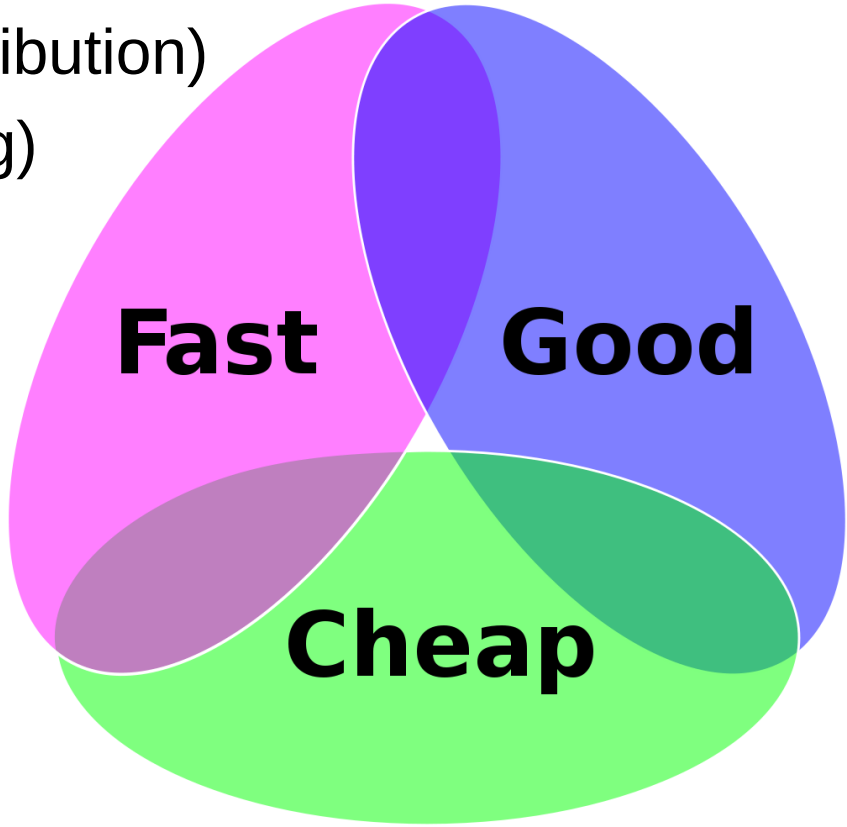
- Digitized tracks of particles in detectors
- Data must be collected as it is generated (~18 months uninterrupted)
- One event similar to the others
- Volume: 15-20 PB/year
- Transfer rates: ~0.5 – 1.5 GB/s
- Keep for > 10 years (forever)



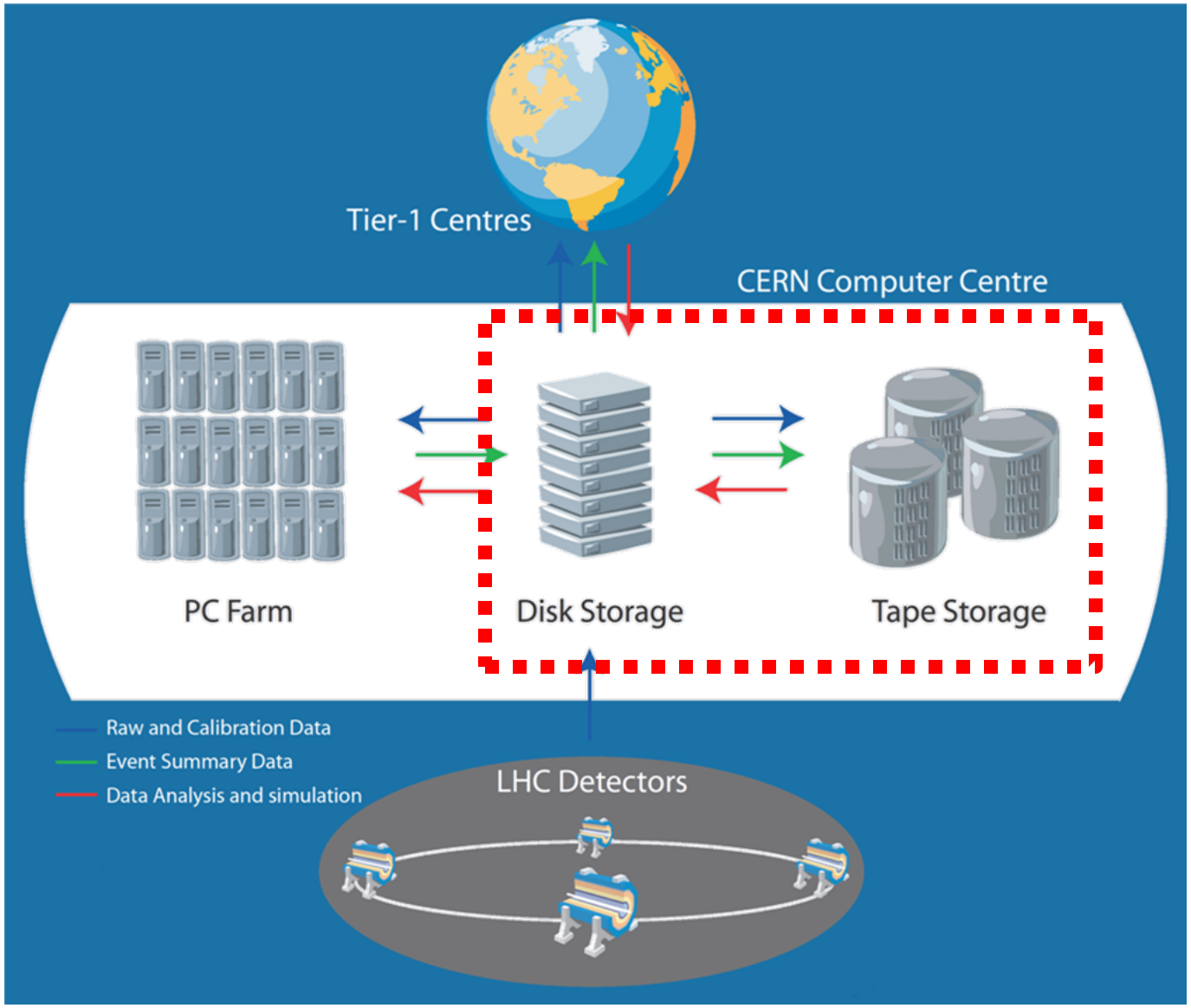
http://atlas.web.cern.ch/Atlas/public/EVTDISPLAY/events.html

Archive Requirements

- Supported access types:
  - Streaming (data recording, distribution)
  - Random (analysis, reprocessing)
- Physical data location hidden
  - Transparently move data from slower media to faster cache
- Use commodity hardware
- Reliability not exaggerated
  - Only 1 copy at CERN
  - Other copies on the grid



Architecture overview

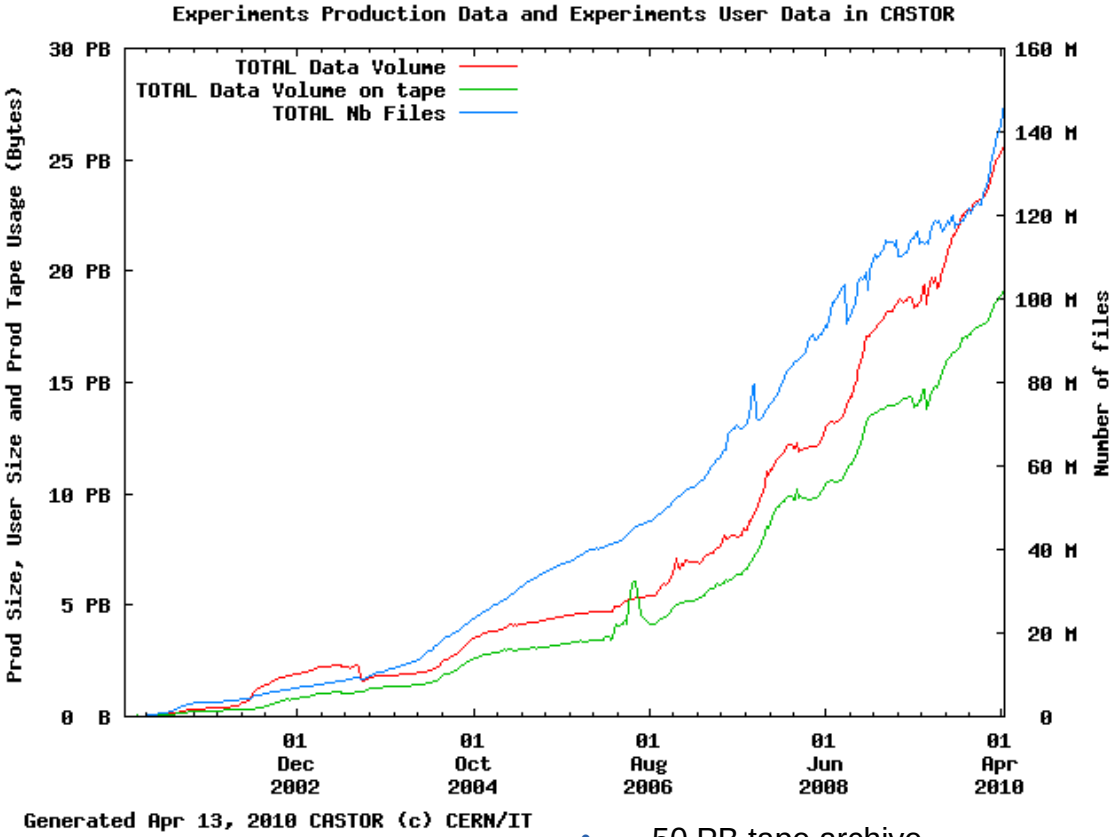


Archive today – CASTOR

- CERN Advanced STORage manager
  - Hierarchical Storage Management system
  - File based with POSIX like hierarchical name space
  - Made @ CERN
    - Used also at other HEP sites
  - Oracle database used for: request queue and metadata repository
  - LSF (from Platform Computing) used for load distribution across disk servers to access files
  - Runs on Scientific Linux

In numbers

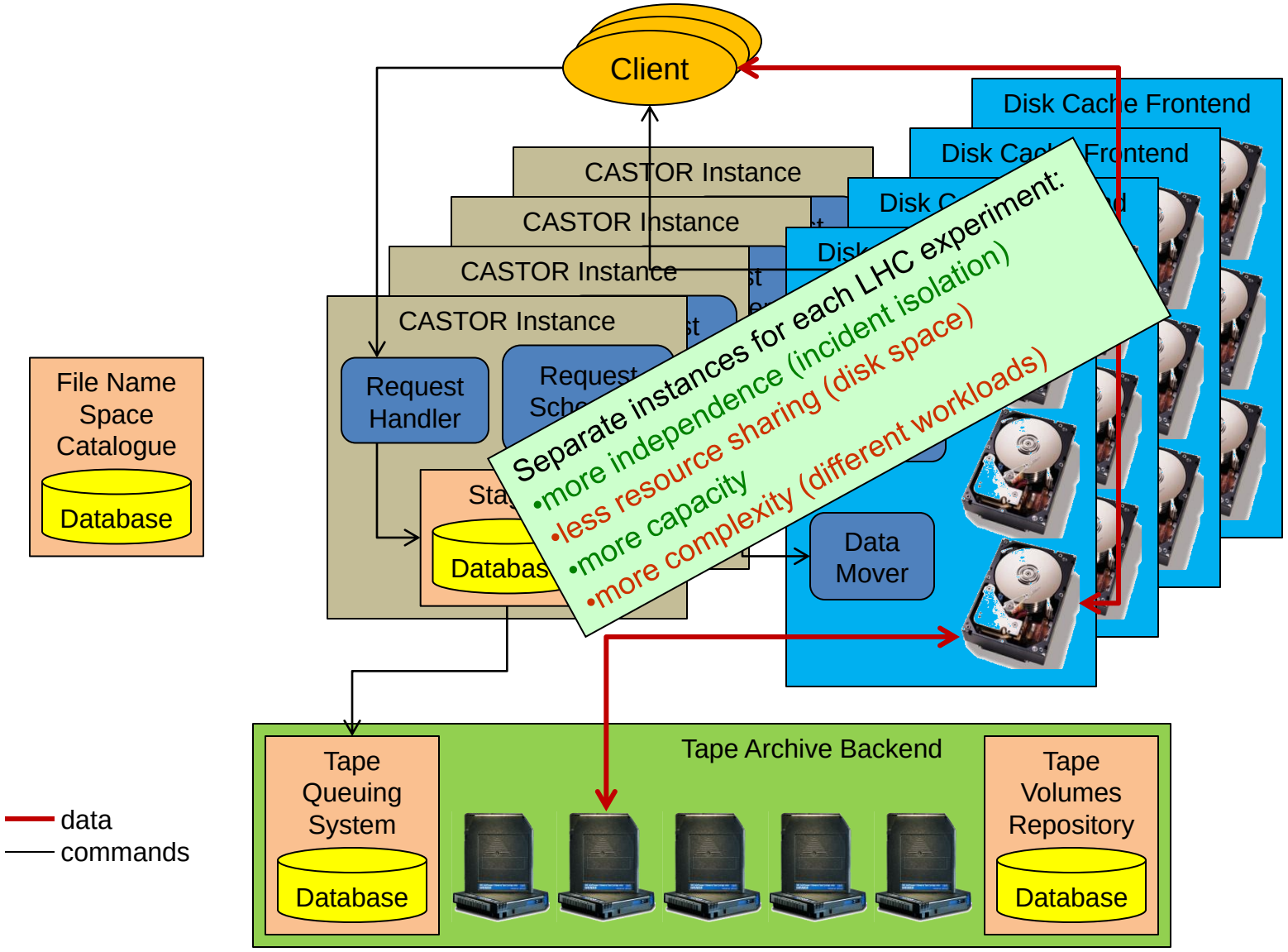
- ~150 M files; ~26 PB data on disk; ~19 PB of data on tape; average file size 150 MB



- 10 PB disk cache
- 1500 (storage in a box) disk servers
- 60 000 SATA disks; 22 disks / server
- Filesystem used: XFS
- Setup evolution
  - Space: RAID5 – with 5 but later with 3 disks;
    Performance: RAID1 today

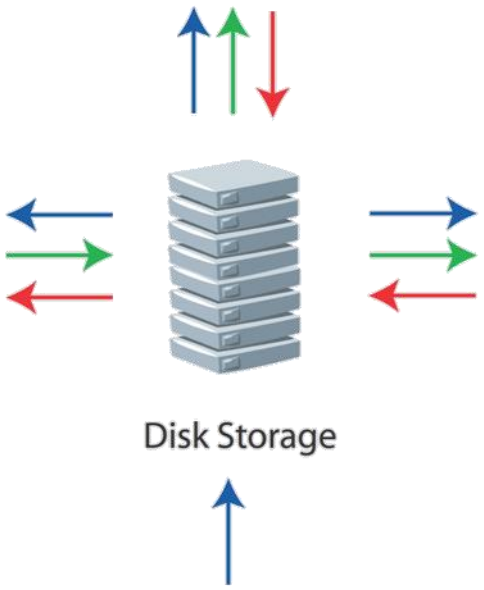
- 50 PB tape archive
- 4 x Sun SL8500, 3 x IBM TS3500
- 70 x Sun T10000B, 60 x IBM TS1130
  - 2 vendors – minimize risk in case of issues
- Using enterprise drives and media
  - Initial cost higher than LTO, but allows media reuse

CASTOR Architecture (simplified)



Disk layer tasks

- Receiving data:
  - Coming from the detectors
  - From the tape layer for analysis
  - Results of the analysis
  - From other centers
- Sending data:
  - From the detectors to the tape layer
  - For analysis at the batch farm
  - To other centers world wide for processing
- Other bulk data transfers
  - Between servers to fight hot spots
  - File merging
  - Draining/Emptying servers for: OS upgrade; Hardware replacement



Disk layer challenges

- Be cheap – use SATA disks – consequences
  - Capacities increase fast while the transfer rate performance is lacking behind
  - Size of user files not changing significantly
  - ☹ Huge slow disk with many small files

- Usage pattern challenge
  - Sparse file access
    - Many clients access remotely few files at the same time
  - Many remote open() files
    - Users keep connections open for long time as they seek within the file and analyze data

- Current limitations
  - Difficulties scaling 1000s connections per disk pool (hundreds per server) to 10x or more
  - Low transfer rate per stream per server
    - Often not higher than 60 MB/s – issue for bulk transfers
  - Unable to prioritize streams, causing e.g. transfers from tape to starve
    - Scheduling already in place, but not enough as jobs look the same
    - Investigating throttling

Tape layer tasks

- Files vs. Data sets
  - Experiments work on data sets corresponding to collisions during certain beam run
  - Current system is file based

☺ WRITE to tape at >3 GB/s? – no problem anymore
  - System designed to split the write stream onto several tapes
  - Waits until enough data to maximize streaming performance

☹ Random READ access on tape
  - With data sets containing 1000s of files, these are spread across many tapes
  - Users asking for files not on disk cause random file recalls
  - Many tapes get mounts but average number of files read is very low

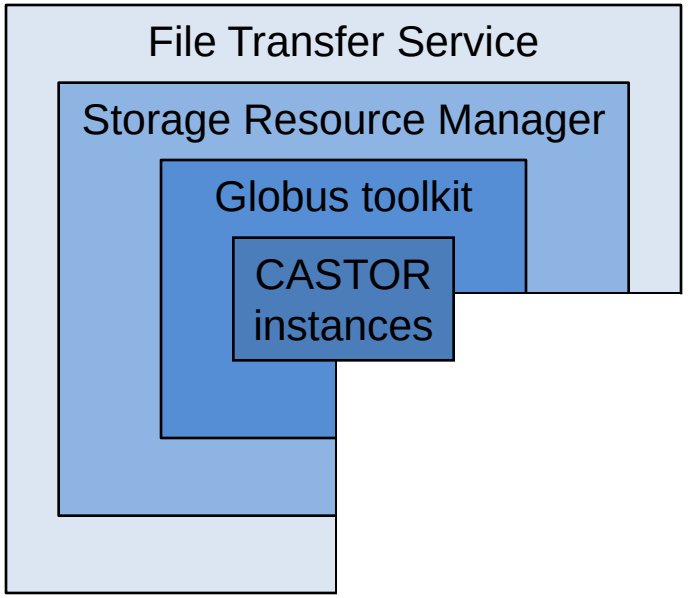
Tape layer challenges

- File collocation
  - Need to find a way how to place related files on the same tape media to improve READ performance
  - Simplest by directory structure
  - Containers offer more functionality but at a cost

- Periodic migration of data to higher capacity media (repack)
  - Costly but necessary operation in order to save on slots and new media
  - Generates significant stress on the system (i.e. repack ~1000 tapes with ~50 000 files)
    - Non trivial to implement as 100% background process because of error handling and resource availability
  - Takes long time
    - Copying 45,000 500GB tapes into 1000GB tapes took around a year using 1.5 FTE and up to 40 tape drives in parallel
    - Next round in 2012 will take ~2 years … but the new drives appear every two years ... we are working on improving the performance
  - 😐 On the positive side, migrating all data is a good way of checking that it is still readable

How is the system used?

- High level tools create several layers on top of CASTOR
  - Low level complexity is hidden from users
  - Users do not need to understand underlying technology
  - Users are not aware of the underlying technology



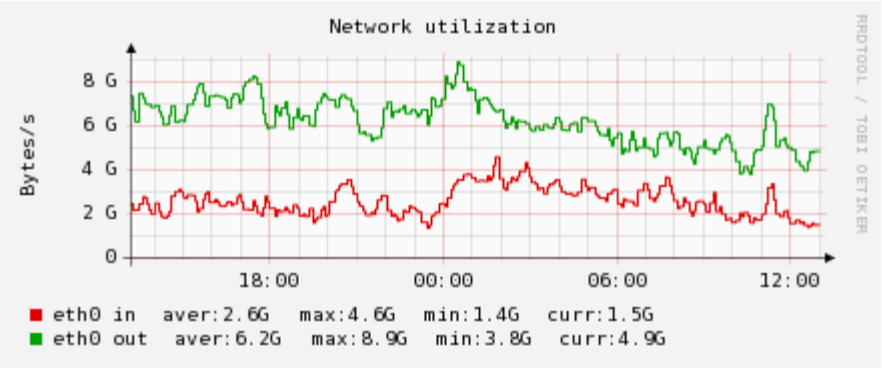
- Complexity is a challenge
  - Misconception that the whole system is down if problems at higher layer
  - Problem solving often requires stack of experts to find the root cause

Monitoring in place

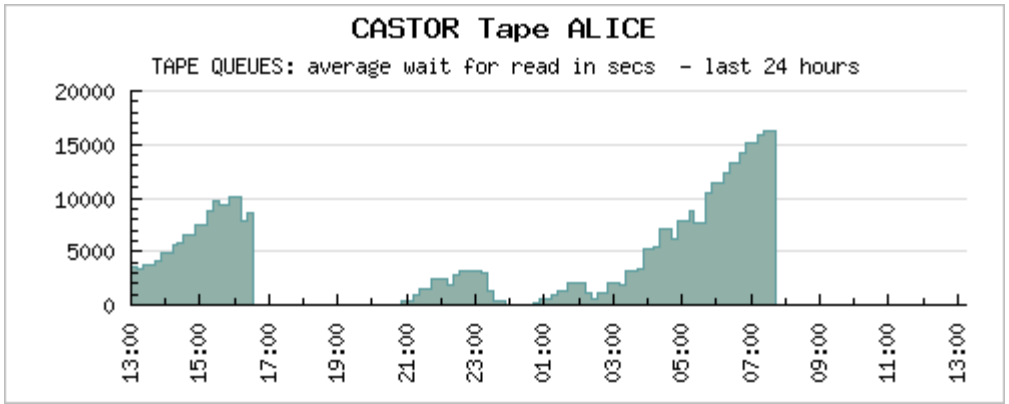
- Crucial part of the system, but pick your viewpoint carefully
- 3 types
  - Developers

11-03-2010 13:02:38.817923	Info	lxfsrc5905	d2dtransfer	15671	15673	Transfer information	castorns	381693276 Stager DB	7fa615e5-8605-3d12-e040-8a8953831cb9	7fa615e5-8606-3d12-e040-8a8953831cb9	N/A	Protocol=rfio SourcePath=lxfsre1108.cern.ch:/srv/castor/01/76/381693276@castorns.6534668538 DestinationPath=/srv/castor/03/76/381693276@castorns.8278959834 ChkSumType= ChkSumValue=
11-03-2010 13:02:38.746727	Info	lxfsrc5905	d2dtransfer	15671	15673	DiskCopyTransfer started	castorns	381693276 Stager DB	7fa615e5-8605-3d12-e040-8a8953831cb9	7fa615e5-8606-3d12-e040-8a8953831cb9	N/A	Protocol=rfio TotalWaitTime=60.746607 JobId=723691 DiskCopyId=8278959834 SourceDiskCopyId=6534668538

  - Operators



  - Users



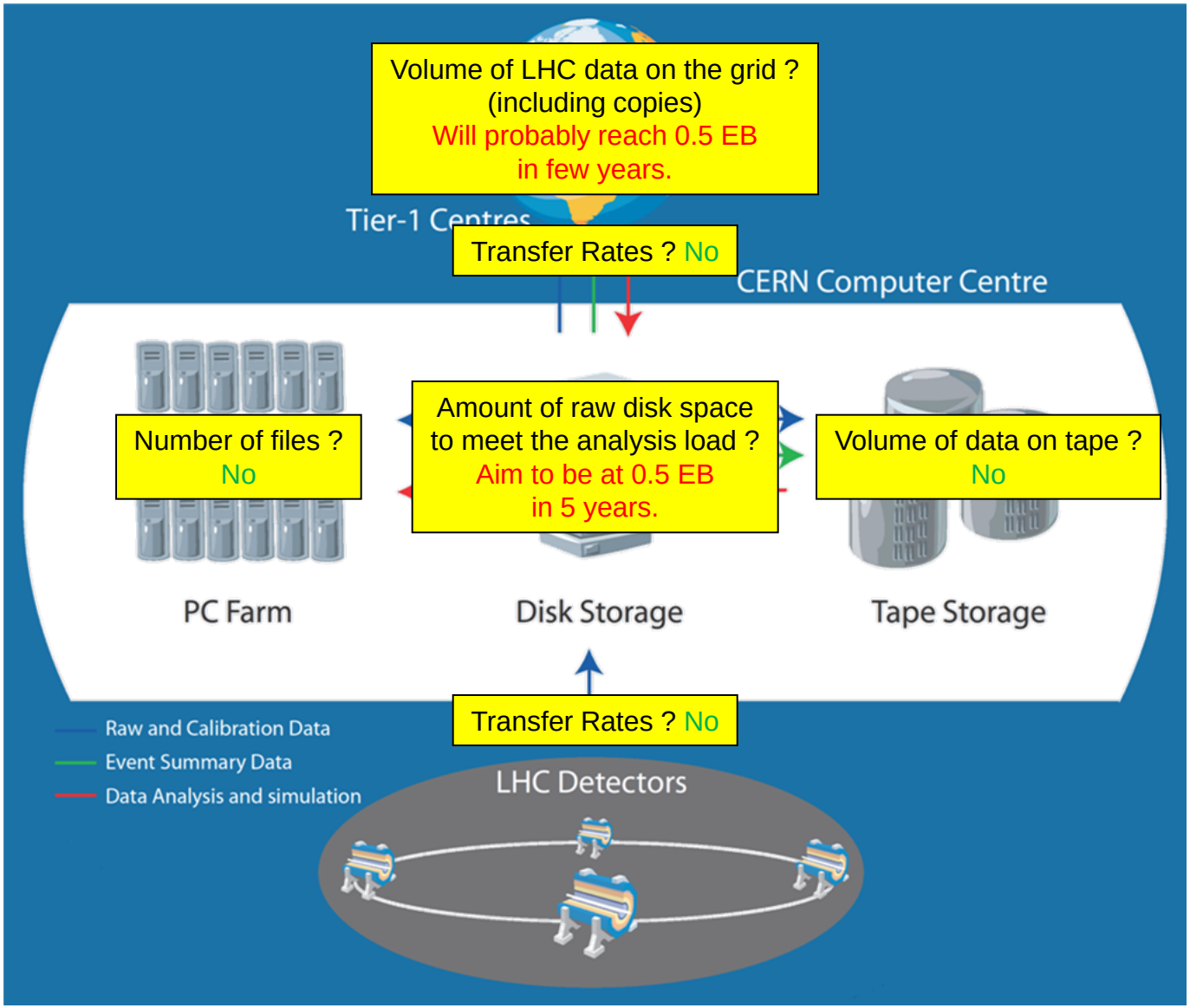
Monitoring challenge

- User: “What happened to my file?”
- Admin: “Your file is lost.”

- How to fill the gap in between?

- Monitoring systems/hardware health is often not enough
- History granularity down to the file level is often needed
  - Should span various applications / systems
    - E.g.: how the file got into the system; when was it last readable
  - Generates huge amount of log data
    - We only keep 3 months of history – insufficient in large archive with lot of inactive data

Exa-Scale Challenge?



(My) Open Questions 1/2

- Efficient data loss handling – quick action can save data
  - Copy still on disk even if tape copy is unreadable
  - Copy on the grid, but it remains difficult to find out who owns the file in several large collaborations
    - Users who do the bookkeeping not us
- Capacity increases faster than filesizes
  - Tape sizes increase and so does the risk of a data loss
  - Data of earlier smaller experiments now fit on fewer tapes; loosing 1 tape can have serious consequences
  - 2nd copies cheaper as small amounts of data
- … and I didn't even mention the data preservation …
  - To store lot of data is technically not so complicated
  - The issue is to understand it decades later …

(My) Open Questions 2/2

- It is sustainable to write and manage the full archive storage software stack in house?
  - The system is already fairly complex and non-trivial to understand
  - Difficult to compete with companies with large client base
  - Need to reuse some well known standard products as building blocks
    - Concentrate only on adding missing features
- Is the current HSM model sustainable?
- Can we afford transparent file level?
  - File granularity overhead is a killer
  - Clear need for bulk only transfers, not individual user files
- Which one of the disk vs. tape in the near line storage scenario are part of a solution and which one is the problem?
  - Both growing in capacities; Not so much transfer rates, seek times
- Need for a solution independent from underlying technologies …

Conclusion

- Current CERN Data Archive solution
  - Is scalable and redundant
  - Satisfies the need of the data recording and data distribution
- Technology evolved, requirements changed
  - Has limitations for analysis
    - Difficulties to support tens or hundreds of thousands remote file I/O operations accessing concurrently small data subset
  - Uses tape technology inefficiently
- CERN
  - Is well aware of the limitations of the current solution
  - Is looking at alternatives to modify the system using industry standard components
  - 2012 – good year to test prototypes when LHC is stopped

Questions ?

- Vladimir.Bahyl@cern.ch