DM

Data Management

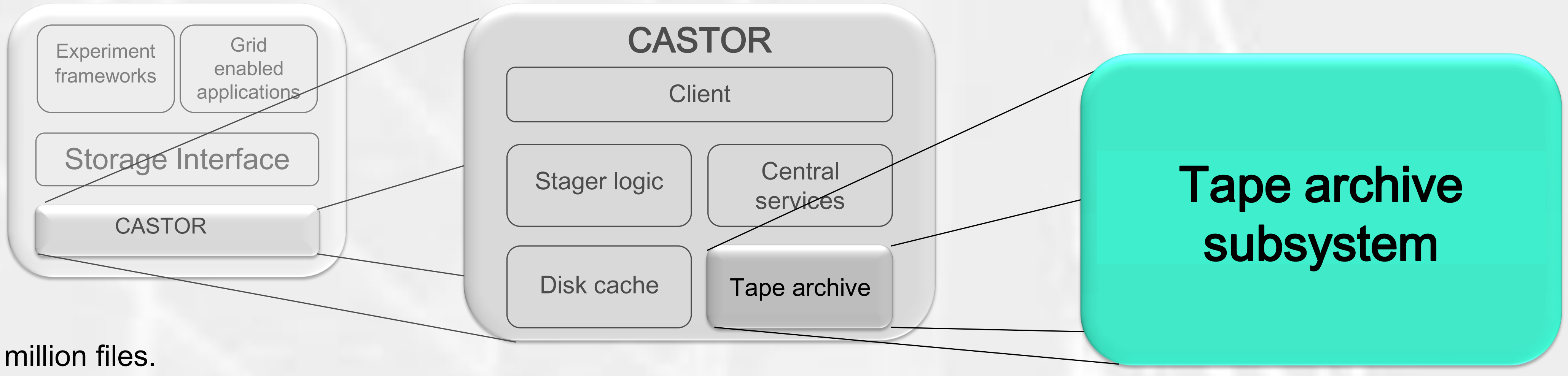
# Increasing efficiency of tape-based storage

Nicola Bessone, German Cancio Melia , Steven Murray, **Giulia Taurelli**

CERN IT Department

**CASTOR**, the CERN Advanced STORage manager, is a hierarchical storage management (HSM) system developed at CERN used to store LHC physics data. CASTOR is in production at CERN and three Tier-1 sites: ASGC, CNAF, RAL

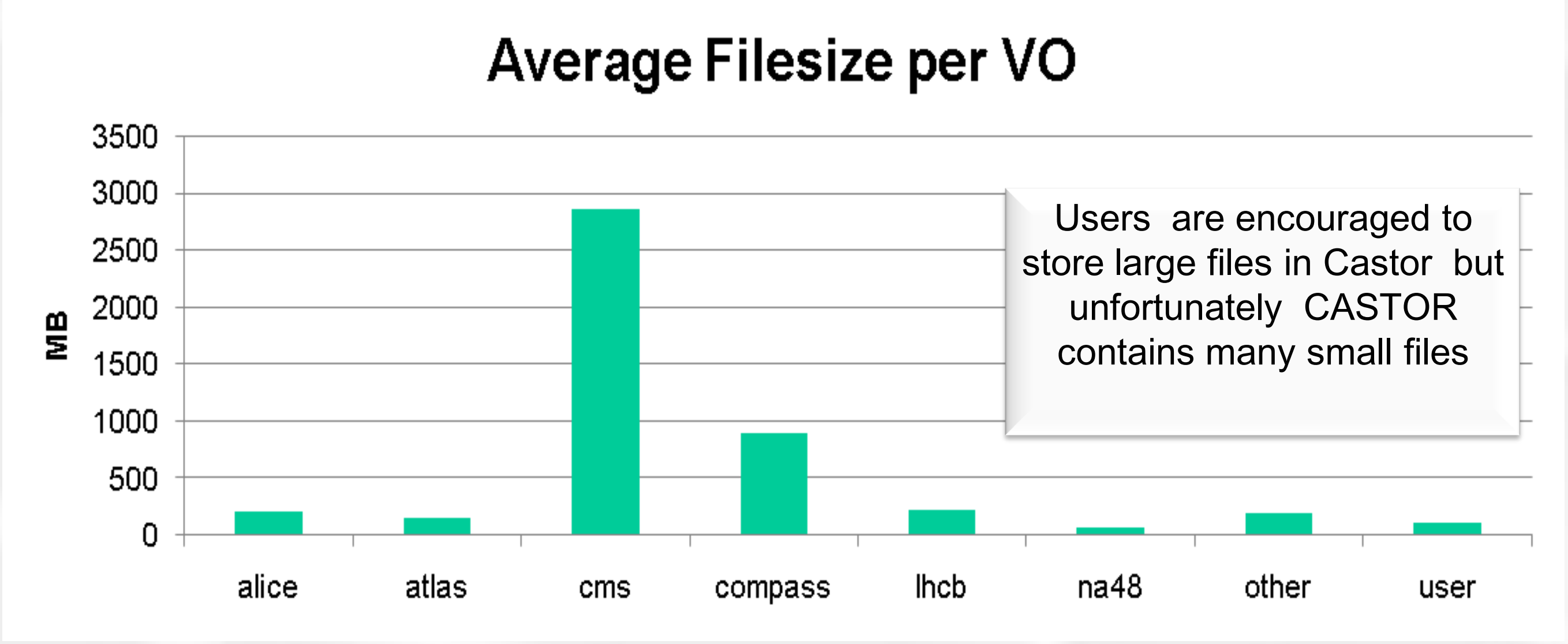
The CASTOR installation at CERN currently stores approximately 17 Petabytes over 114 million files.



All functionalities directly dealing with storage on, and management of tape cartridges, drives, libraries and servers .

# Efficiency challenges…

The tape data format currently used by CASTOR is ANSI AUL and was set in place in the 1990s. The speed and data capacity of tape media has evolved significantly since then. The average capacity of a tape cartridge in the 1990s was between 5 and 10GB, whereas now we are beginning to use 1TB cartridges. In contrast the size of physics data files has not increased by the same magnitude.



There are a number of challenges in terms of CASTOR performance:

## Reads

On average CASTOR reads 1.5 user files per tape mount. This is extremely inefficient considering the fact it takes between 1 and 3 minutes to mount a tape.

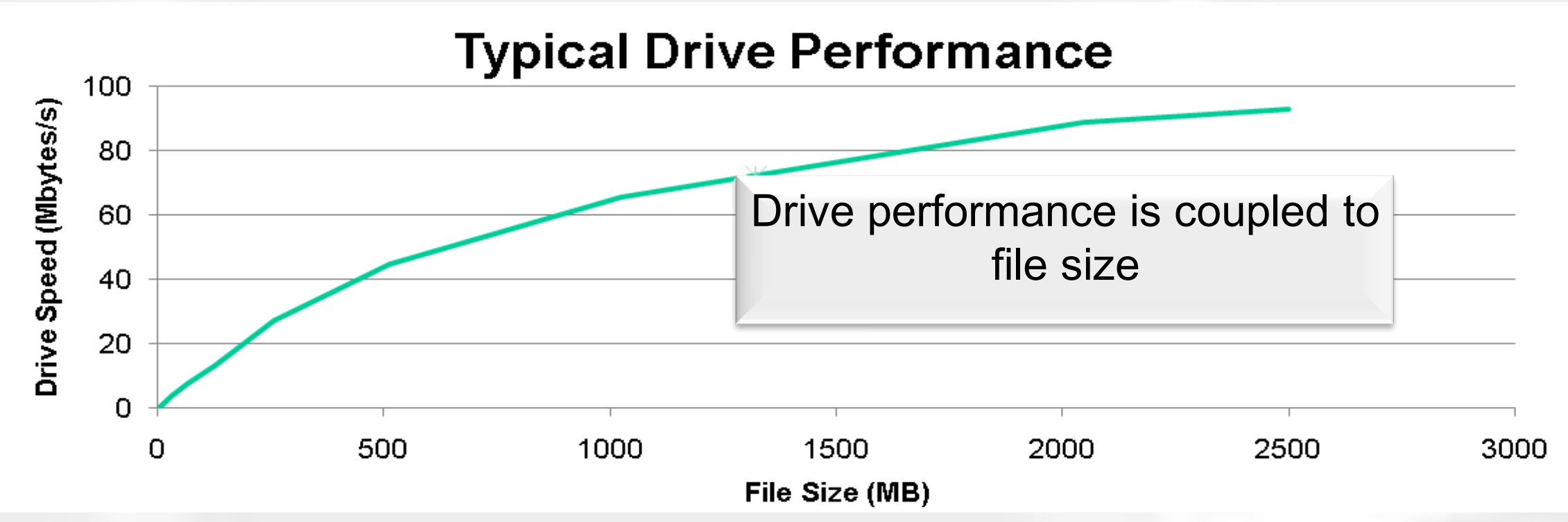
The low number of user files to be read per tape is due to:

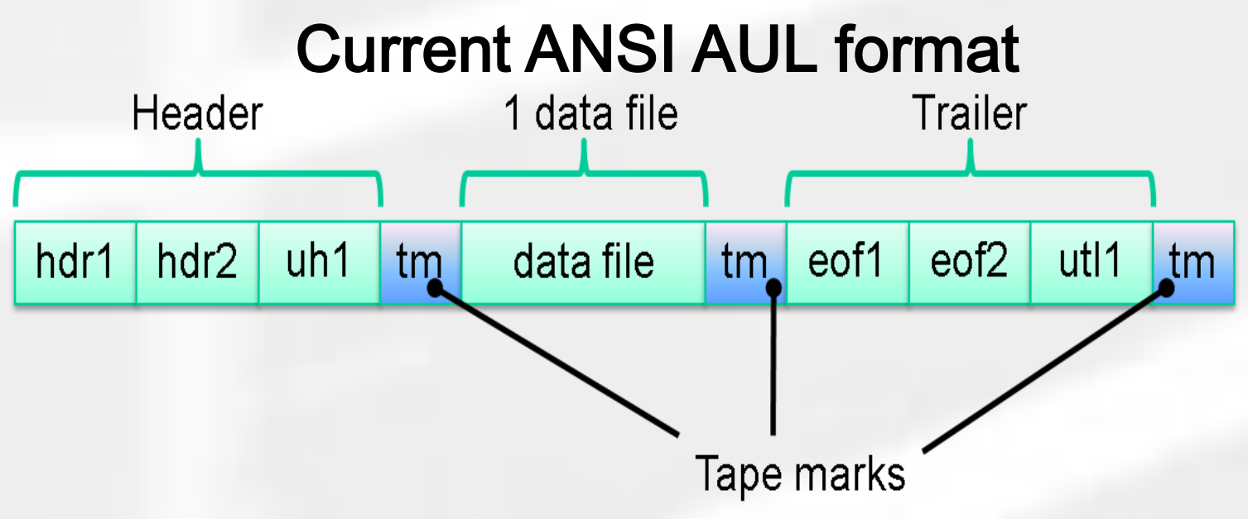
- Related files not being written together on the same tape(s)
- The current low latency requirements of the CERN batch Farms.

**Physics files are relatively small (size matters …)**

## Writes

The usage of migration policies enables building up file streams to be sent to tape. However, the efficiency of writing small files is low due to writing each disk file as a single tape file in the current AUL file format. This format requires writing header and trailer metadata files around the contents of each data file. The writing of tape marks is the most dominant factor in the writing of the header and trailer metadata. A total of ~5-9 seconds is taken per file to write the three tape marks and the metadata (~2-3 per tape mark).





The ANSI AUL format results in a total of ~5-9 seconds overhead per file independent of its size.

The ~5-9 seconds are spent writing 3 tape marks

Recent evolution of tape media at CERN

| Vendor | Current capacity | Future capacity | Nb at Cern | Delta capacity | Cost |
|---|---|---|---|---|---|
| IBM | 700GB | 1000GB | 9692 | 2.9PB | 0.5MCHF |
| SUN 513 | 500GB | 1000GB | 14890 | 7.4PB | 1.3MCHF |
| SUN 613 | 500GB | 1000GB | 15408 | 7.7PB | 1.4MCHF |
| Total | | | | 18.0PB | 3.2MCHF |

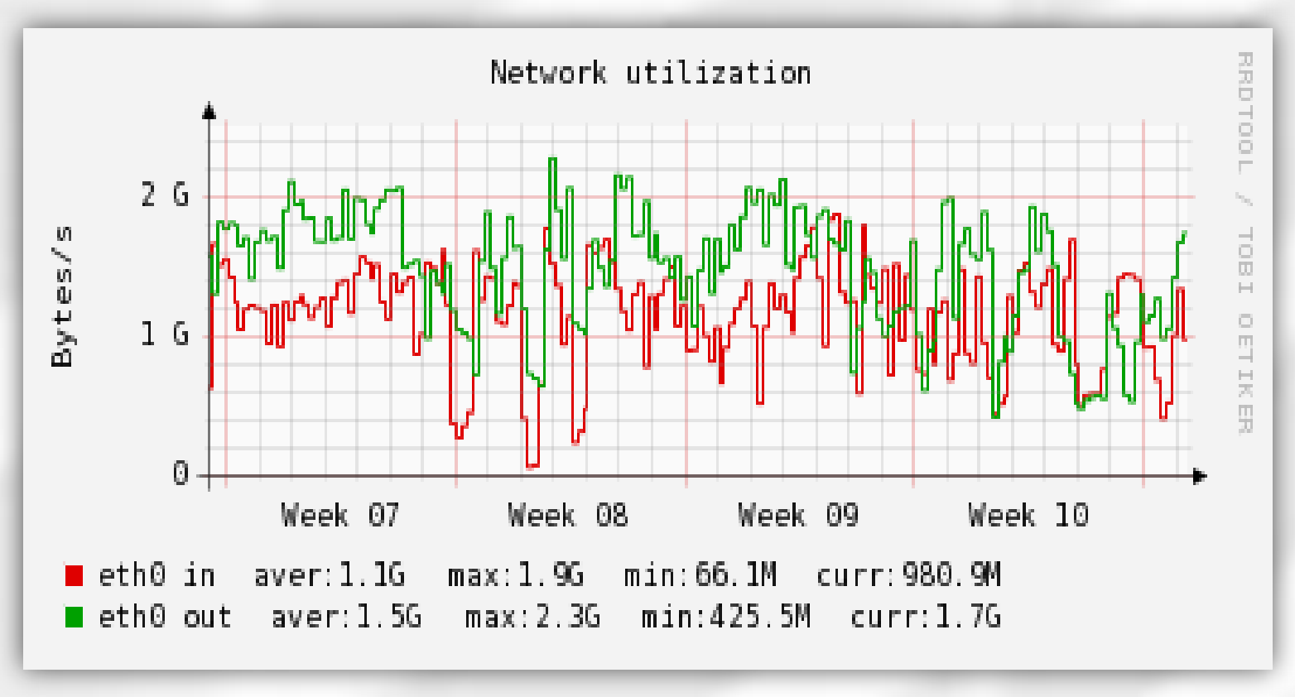
## Media Repack

Media repacking is the copying of data from one set of tapes to another and is done for the following reasons:

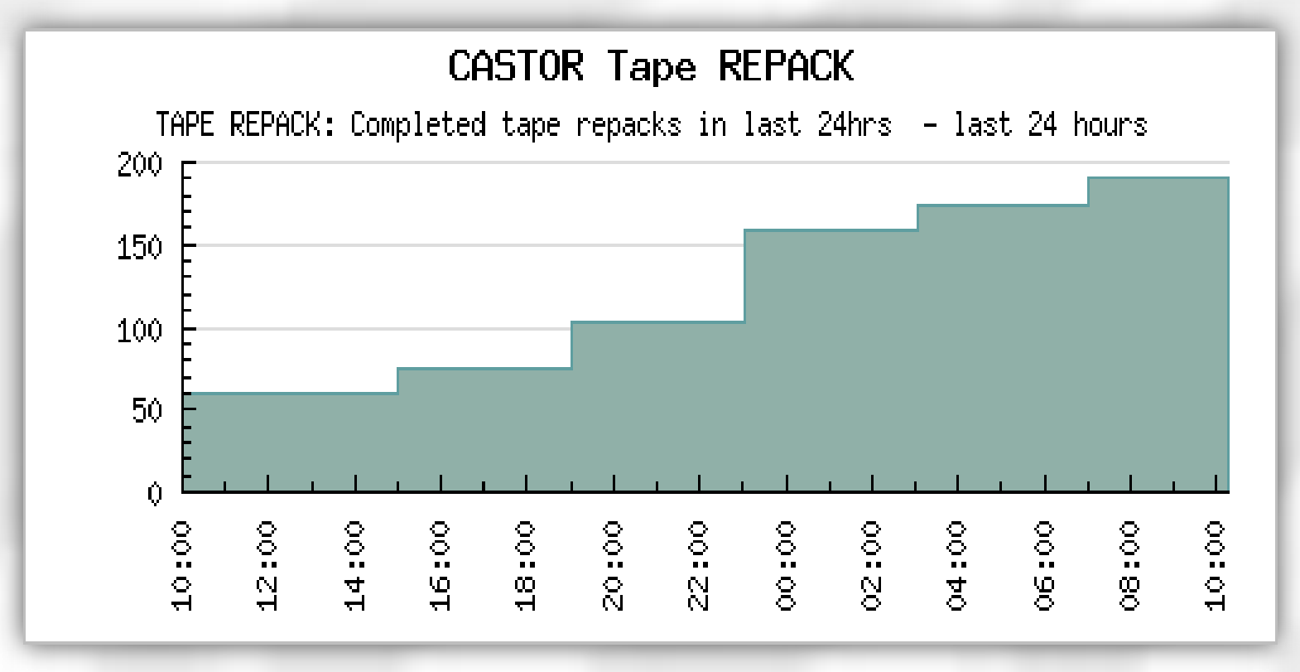
- Data recovery in case of media errors
- Media defragmentation (clean up "holes" after data deletion)
- Media upgrade (old tapes eventually wear out, new tape generations have higher densities)

CASTOR has a repack application to perform this task which reuses the stager layer of the CASTOR architecture.

Repack operations at CERN are done using a dedicated CASTOR instance. This instance has a load equivalent to that of one LHC Experiment!



■ eth0 in aver:1.1G max:1.9G min:66.1M curr:980.9M
■ eth0 out aver:1.5G max:2.3G min:425.5M curr:1.7G



# … and the solutions:

## Increasing operations per mount

In order to minimize the number of accesses to tape media, and consequently to increase the number of read/write operations per tape mount, a number of improvements have been developed:

### Recall and Migration policies

The base concept of both recall and migration policies is holding back the migration and recalls depending on the amount of data and elapsed time. This way, the total count of tape mount operations should be minimized for both reads and writes.

### Prioritization and Access Control

Whenever possible, end users should access data which has already been staged on disk. End users should be encouraged to work in coordination with "alpha" users such as production managers, which are responsible for deciding which data sets are to be staged to or removed from disk. This can be endorsed by defining user and group based access control lists and priorities for initiating tape based recall operations.

**Reduced tape marks = Increased performance**

## New tape format

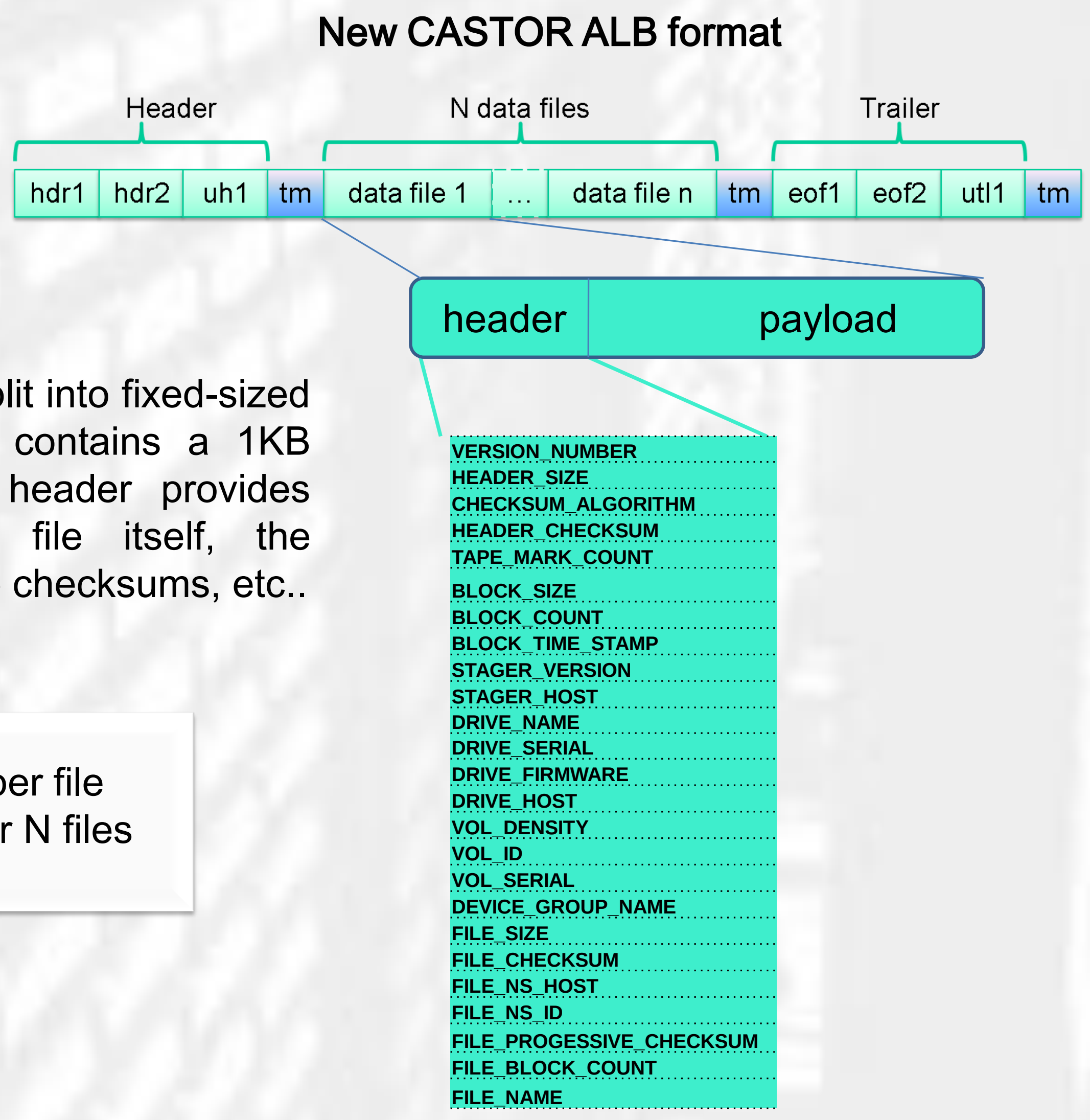
A new tape format (ALB, ANSI Label with Block format) is being developed for CASTOR, with the aim to increase efficiency and redundancy.

### Aggregations

The new ALB format is based on the ANSI AUL format.

While the AUL structure will be kept for this format at least initially, the payload inside each AUL data file will consist of an *aggregation* of multiple CASTOR files, in order to reduce the number of tape marks. The incoming stream (list of files) to be migrated will be aggregated to a configurable maximum total size (e.g. 10GB) and/or configurable maximum number of files (e.g. 1000 files). If a file exceeds the maximum total size it will be written in a separate aggregation consisting of that single file.

On hardware with efficient tape mark handling, the number of files per aggregation can be decreased.



### Block-based format

Every file within an aggregation is split into fixed-sized blocks (e.g. 256KB). Every block contains a 1KB header for self description. This header provides metadata information about the file itself, the aggregation, the tape, the drive, , the checksums, etc..

AUL format = 3 tape marks per file
ALB format = 3 tape marks per N files

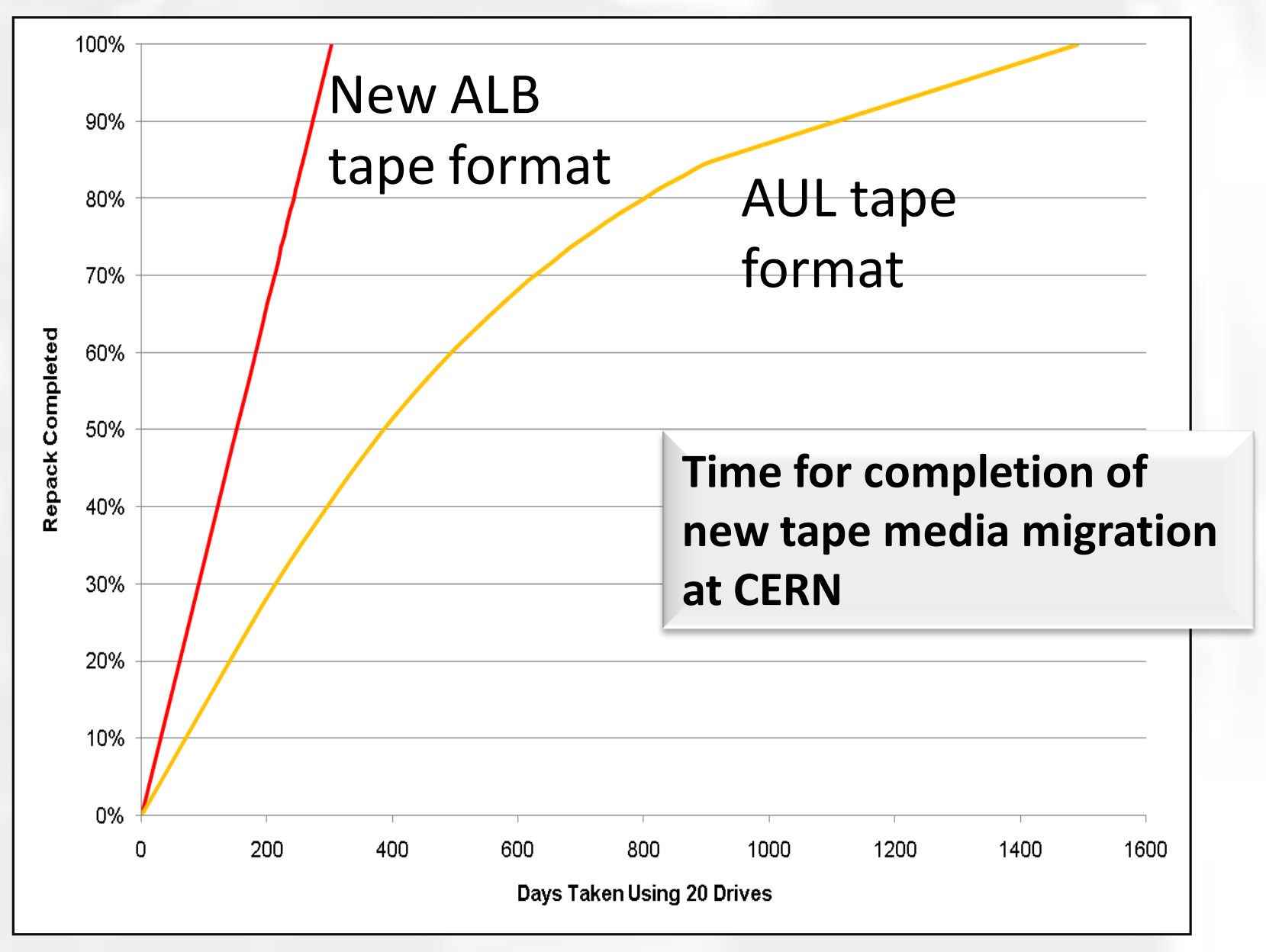
## Benefits for Repack

The new tape format will allow to increase the performance of repacking data from old to newer generation tape media with substantially reduced hardware costs.

Thanks to the new tape format we predict a reduction from 4 years to 1 year for the time needed for repacking all tapes.

**75% saved time**

This time scale will be compatible with increases in tape density which usually occur every 2 years.



Time for completion of new tape media migration at CERN

Background picture by Andras Horvath

CERN IT Department
CH-1211 Genève 23
Switzerland
**http://cern.ch/it-dm**

Many thanks to the CERN tape operations team for their ideas and valuable input

www.cern.ch/castor

CHEP'09, 21 - 27 March 2009, Prague, Czech Republic

CERN