The Repack Challenge

Daniele Francesco Kruse

(IT Department – DSS Group)

The Repack Plan

*What does it mean to **repack** a tape?*

To move data from that source tape to one or more destination tapes.

*Why is it **important**?*

1) To reclaim space of fragmented tapes
2) To move data out of problematic tapes
3) To reuse tapes in other applications
4) To move data onto next generation tapes

Repack reduces risk. It saves time, money, resources.

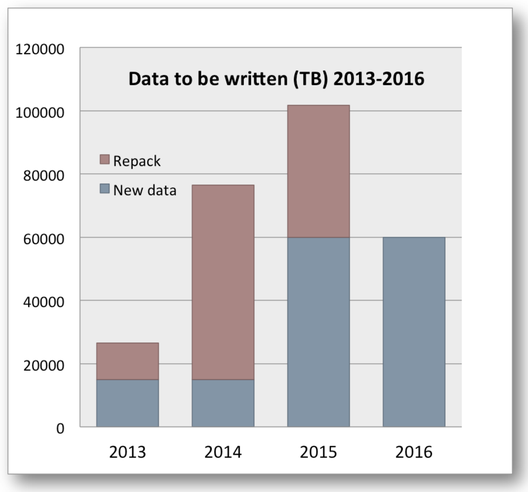
In 2014 - We are expecting the *next tape drive generation*

>90 PB - Amount of physics data stored on tape at CERN

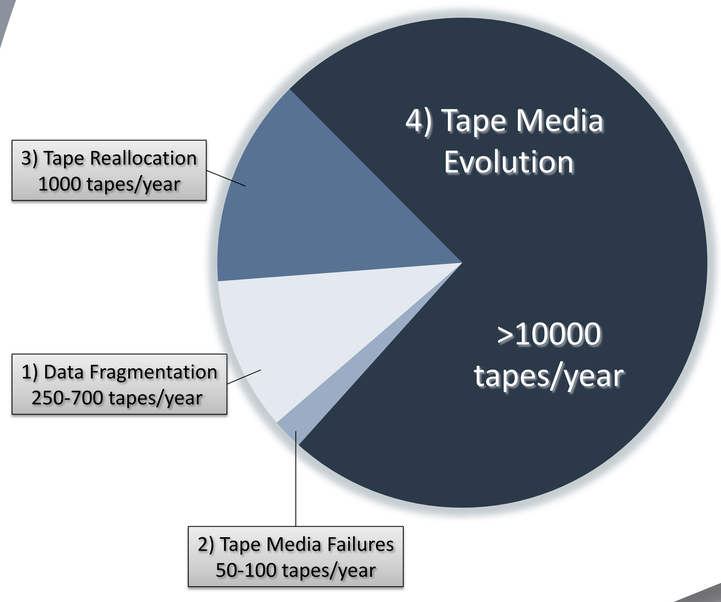
3.8 GB/s - Throughput needed to complete repack before the LHC run

56,000 - Tapes to be repacked

This will be the largest repack exercise so far, and the most throughput demanding. The challenge is to make it also the most ***transparent for physicists***.



How many Tapes we Repack?



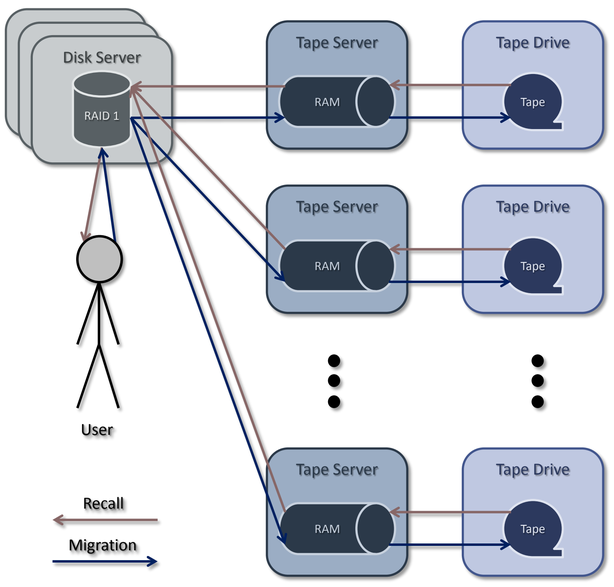
How CASTOR works

Physics data at CERN is stored using a homegrown HSM tool called **CASTOR**.

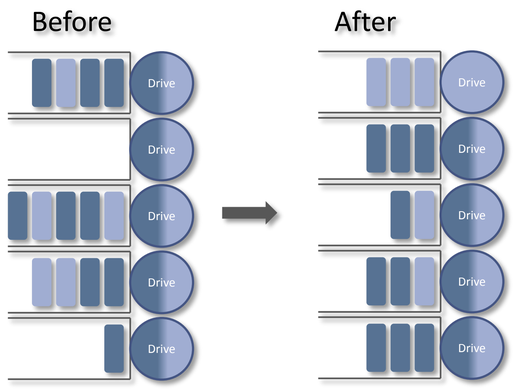
Data coming from experiments is temporarily stored on a disk cache made up of hundreds of disk servers, and then flushed to tapes. This is what we call ***migration***.

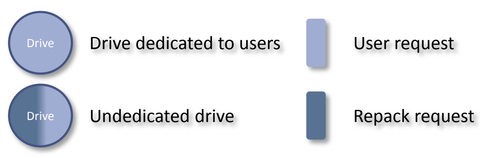
When this data is eventually needed by physicists to perform their analysis, it is first read out of tapes and then staged again on the disk cache, this process is called ***recall***.

The ***repack process*** acts just like a normal user, in that it requests the recall of all data from the tape it wants to repack, then the data gets copied onto the disk cache, and eventually migration rules make sure that the recalled data is correctly migrated to new tapes.



Improving Library Utilization





To be as **transparent** as possible for the user community the first thing to do is to give ***user recalls higher priority*** than repack.

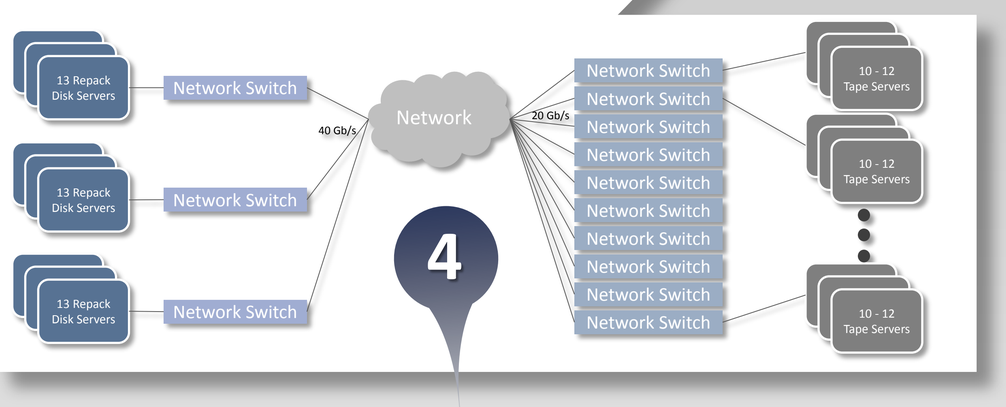
Secondly, to make sure that users do not have to wait for a full repack to finish before getting hold of a drive, we dynamically and automatically ***dedicate a number of drives per library to user-only activities***.

Repack needs to be not only transparent, but also ***efficient***. We have several libraries, and within those libraries there are different types of drives able to read different sets of tapes. Each pair *<library, drivetype>* has thus a different request queue. In many cases we have some queues that are overloaded while some others are rather empty.

While we cannot do much for repack recalls, because a certain order needs to be kept (to maintain data *temporal collocation*), we can certainly do a wiser choice while deciding on-the-fly which empty tapes to use for repack migrations. This is done by choosing the destination tapes based on which pair *<library, drivetype>* has the least load.

Disk and Network Optimization

RAID10 (Repack)	RAID1 (CASTOR default)
Max controller throughput: ~400 MB/s	Max controller throughput: ~350 MB/s
Synch time: ~3-4 seconds	Synch time: ~1-2 minutes
1 stream 395 MB/s 100%	1 stream 130 MB/s 100%
2 streams 304 MB/s 77%	2 streams 107 MB/s 82%
4 streams 272 MB/s 70%	4 streams 95 MB/s 73%
8 streams 240 MB/s 61%	8 streams 87 MB/s 67%



Old repack setup

17 disk servers, 24 disks each, split in 12 RAID 1's and 1 Gb ethernet connectivity. Back in 2009 during the last large repack exercise, the average recall speed was only **56 MB/s** (half of the drives' potential throughput).
*Overall repack speed **1.8 GB/s**.*

New repack setup

39 disk servers, 24 disks each, organized in 3 RAID 10's, and 10 Gb ethernet connectivity with dedicated network switches. Average recall speed: **205 MB/s** (150 MB/s for old media). Average migration speed: **220 MB/s**.
*Overall repack speed **6 GB/s**.*