

# Draft 1.0

## Some calculations for the sizing of tape storage performance

Bernd Panzer-Steindel, CERN/IT  
07.03.2005

### Introduction

This note tries to give some guidelines on how to estimate the number of tape drives one needs for a given performance requirement based on measurements, experience and some mathematical formulas.

One can essentially distinguish three different points which define the effective performance of tape drives :

1. There is the measured maximum performance value per drive (Maxspeed) which is today in the area of 20-40 MB/s.  
(LTO-2 = 20 MB/s, STK 9940B = 30 MB/s, IBM 3592A = 40 MB/s)
2. The access pattern defined by the user application and the 'translation' into tape transfer requests by the mass storage system.
3. The coupling to the disk storage (cache) and the performance of it.
4. The failure rate of tape drives, the corresponding servers and the software.

### Data Rates

There are several data rates to be considered in a Tier 1 :

- the data rate from the T0  
for the tape rate this has actually be doubled as there will be situations where the T0 export couldn't take place for some time (days ?) and then to cope with the backlog the data → the current data flow + the backlog to empty the T0 buffers.
- the data rate from the continuous processing of RAW and ESD data which overlaps with the RAW+ESD data input from the T0
- the more un-organized data rates coming from the T2 input, the T2 requests for ESD/AOD copies, the analysis/calibration of specific RAW data samples, user analysis, etc.

The first two can be handled in a well described and organized manner, where one can achieve high tape drive efficiencies ( ~ 50 % ), while the latter is more the ‘chaotic’ case where efficiencies can drop to below 5 %.

## Disk pool performance

The performance of the used disk pool depends on quite a number of different configuration and performance values like :

- performance per server / disk array
- RAID5 versus mirrored file system performance
- How many streams read/write are concurrently running ?
- Guarantee the ‘full’ tape drive speed, the OS can only in a very limited way do a ‘quality of service’ for the disk I/O streams. Thus a disk server can run at 100 MB/s with 50 streams, but each of them equally only at 2 MB/s.
- How does the disk pool manager arrange for load balancing and ‘preferring’ tape streams ?
- .....

This is an example of a setup we just tested in the framework of the ALICE-IT data challenge :

20 disk server, each with 11 times 200 GB file systems (mirrored)  
on a 10 GB switch, each node with GB connections

total disk bandwidth : 8.8 GBytes/s  
total server bandwidth : 2.2 GBytes/s  
limited by the switch : 1.1 GBytes/s  
system running with 130 concurrent input (writing) streams and 25 output (reading) streams to tape drives : 0.5 GBytes/s + 0.5 Gbytes/s  
load balanced streams over all file systems

→ average tape stream speed ~ 20 MB/s == 66% efficiency for 9940 B, 50 % for IBM 3592A

Despite the fact that we have in principle very large available bandwidth, this alone does not help to guarantee full tape speed.

Another important point is the consideration of certain thresholds. New announced tape drives will be able to run in principle at > 100 MB/s speeds. The efficiency now becomes non-linear, because to reach such speeds very special tuning/investments have to be made for the disk pools :

e.g.

- One has to use RAID5 + striped setups to reach single stream performance > 100 MB, but this couples the disks strongly, thus stronger performance degradation for multi stream access.
- With a limit of number of parallel accesses to the now large file systems, the total amount of equipment increases.
- There are network bandwidth limitations in the server. Even with multiple GB interfaces, then single stream is limited to one Gbit (Linux issues). Plus the whole issues about the support and configuration of load balancing and connectivity in the switches.
- Another alternative is to move to high end servers with 10 Gbit interfaces which would require a more expensive network infrastructure. Or use a complete 2 Gbit SAN infrastructure (tape and disk).

Thus a whole lot of issues arise when high performance tape drives are crossing certain bandwidth thresholds. This is then a TCO calculation → tape drive cost and efficiency versus infrastructure costs.

The threshold is here probably in the 50 MB/s area. CERN is currently conservatively assuming a maximum speed for ALL future drive of 50 MB/s.

## Access pattern and tape drive efficiency

To calculate the efficiency of a drive several input parameters need to be considered :

- fSize = size of the file
- nRequest = number of different requests (files) per command
- tMount = time to mount and un-mount a tape
- tOverhead = overhead time per file processing (tape-marks)
- tSpeed = native transfer speed per drive

**Throughput [MB/s]** =  $nRequests * fSize / (tMount + nRequests * tOverhead + nRequests * fSize / tSpeed)$

**Efficiency** = Throughput / tSpeed

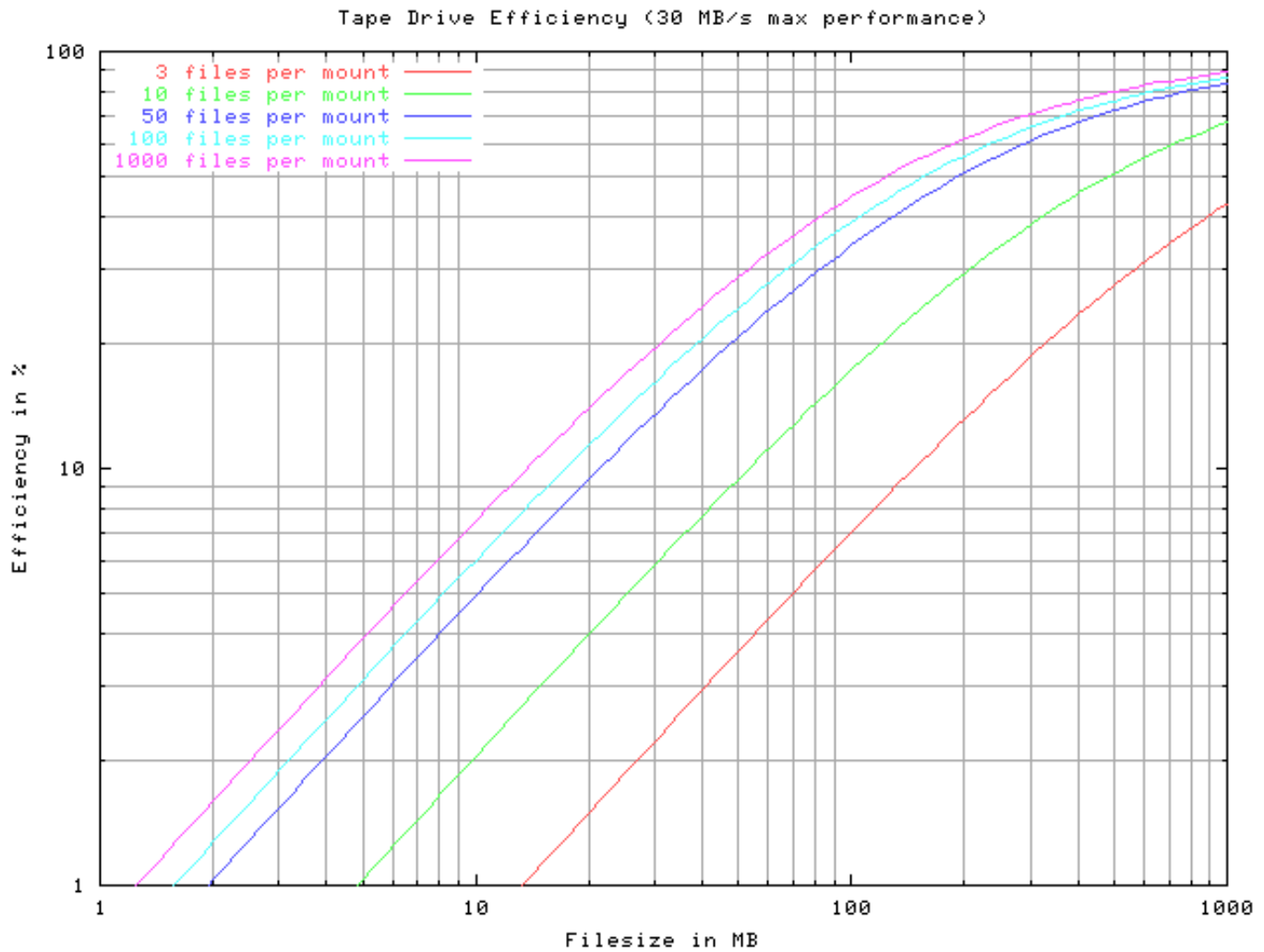
Input numbers for 9940 type drives (applies more or less also for the IBM 3592):

**tOverhead** The overhead time per file on tapes comes from multiple tapemarks which have to be written, a measurement with 1000 files per migration to one tape (different file sizes: 1, 10, 100, 1000 MB) show this time to be about 4.4 seconds.

**tMount** this time has actually several components

- time to move one tape from its slot to the tape drive;  $3600 / 160 = 22.5\text{s}$   
(the normal STK Powderhorn Silo can do 160 robot arm movements per hour, the upgraded one 240)
  - time to mount the tape in the tape drive ; 15 s
  - average time to positioning the tape to file to be read ; 45s
  - time to rewind and unload; 45 s
- the total tMount time is about 120s

The following plot show the efficiencies for a tape drive with 30 MB/s maximum speed in dependence of the number of files written/read in one go to/from tape and the file size.



## Summary

If everything has been tuned correctly one can assume a tape drive efficiency of 50% , but one should add a factor 2 for the overall performance to cope with peaks, unexpected backlogs, failures, etc.

If one moves from the organized production environment (storing the data coming from the network, re-processing of raw data) into the more 'chaotic' usage of tapes , e.g. analysis with frequent flushes of disk caches, the efficiency drops quite quickly. And if the pattern is random over a large data set, the size of the disk buffer becomes unimportant for the sizing of the tape system.

Thus a sample calculation for the number of tape drives needed could look like the following :

Tape drive maxspeed = 40 MB/s  
Data rate T0 → T1 = 100 MB/s  
Data rate reproc = 150 MB/s  
Data rate t2,analysis,etc = 50 MB/s

Number of needed tape drives = 50  
=  
(100 MB/s \* 2 (backlog) / 0.5 (efficiency) / 40 MB/s)  
+  
(150 MB/s \* 2 (peaks) / 0.5 (efficiency) / 40 MB/s)  
+  
(50 MB/s \* 2 (peaks) / 0.1 (efficiency) / 40 MB/s)