Scientific Mass Storage at FNAL

November 2010

Matt Crawford

Mission

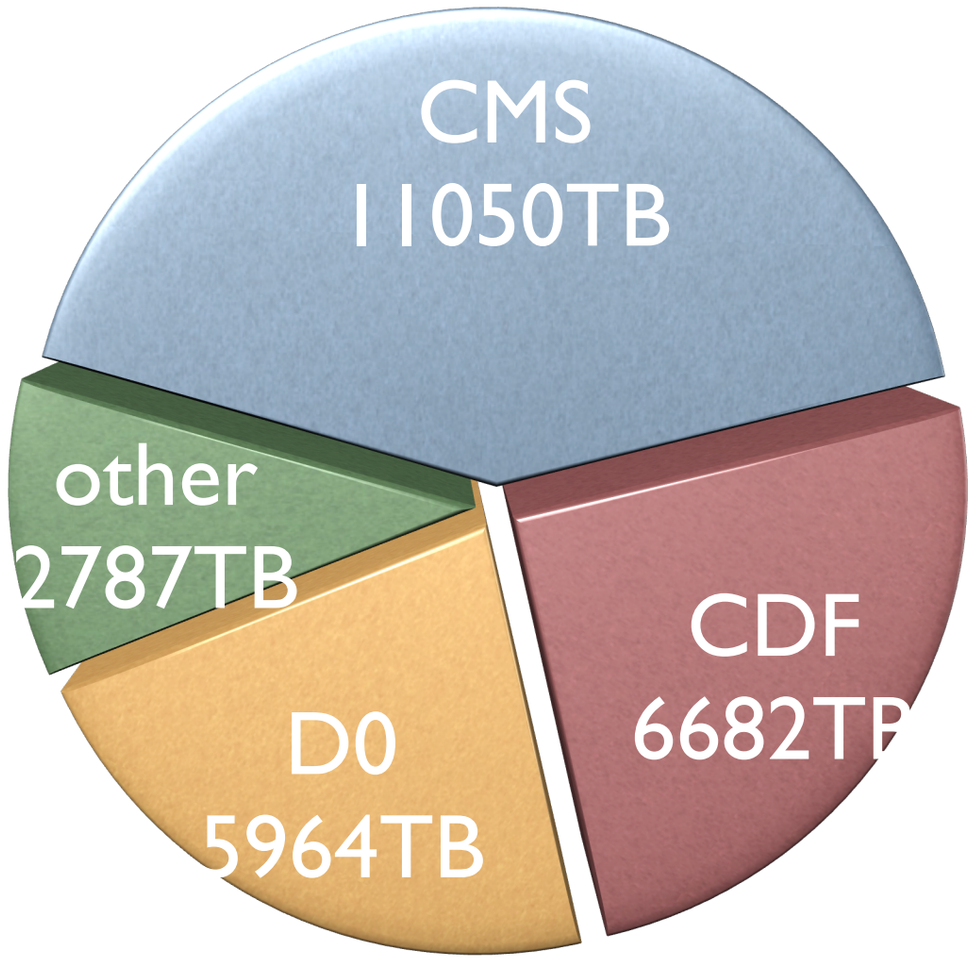
"Data Movement & Storage" operates ...

- mid-sized dCache
- very large-sized tape storage

Develops ...

- Enstore
- dCache, especially SRM, with DESY, NDGF, et al.

Customers



"others" includes: Intensity Frontier experiments, Lattice QCD, some fixed-target data, etc.

Direction: ITIL

Fermilab CD is aiming at ISO20000 certification for central IT services. (Scientific data storage is **not** one).

Also, ITIL (v2) practices for all services.

- {Incident, Problem, Change, Release, Capacity, Availability, ...} Management

It's a lot of new work, but with some value.

Disk storage

	size (TB)	read (TB/d)	write (TB/d)
CMS dCache	8200	150-700	25-200
CDF dCache	444	50-150	---
“pub” dCache	100	2-20	3-7
LQCD Lustre	260	?	?
CMS Lustre	coming soon		
“pub” Lustre	coming later?		

dCache & Lustre

We're using both. Each has virtues—

- dCache has mature load management, many transports, HSM interface.
  - **Strong** support as an EU project.
- Lustre can stripe transfers, may become more "off-the-shelf."
  - Admins still coming up to speed.

Tape Storage

Six Oracle SL8500 libraries, 10,000 slots each, storing 26454 TB (Nov 1) on LTO.

—FCC—

Run 2 processed & other users

—GCC—

CMS ↑

Run 2 raw →

Tape Software

Enstore

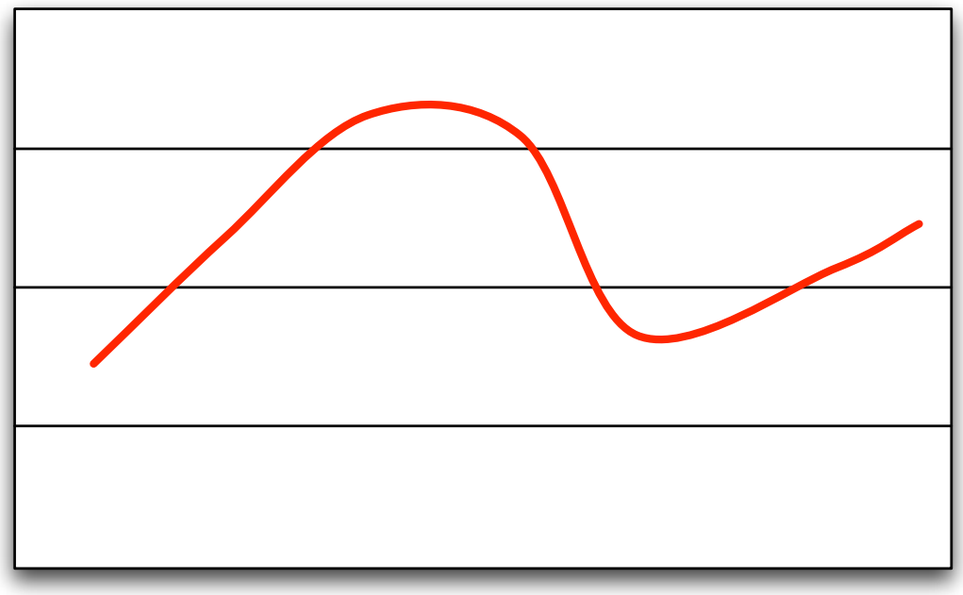
- Automated tab-flipping, periodic random file integrity checks, adjustable read-after-write percentage.
- v2.0's priority system handles much longer queues of tape requests.
- Soon: Chimera compatibility, small file support.
- In use at PIC and LSU.

Tape challenges

Migration to denser media is a normal process, not a special event.

- 9940A/B and LTO1/2 are gone.

Capacity increases beyond LTO5 are expected after CMS data volume forces another library...



Tape challenges

Uncertainty of Tevatron run extension.

- Affects library acquisition (long lead time!) and our choice of final media for Tevatron data.

GCC tape robot room, like all computer rooms, nearing power and cooling limits.

- Will move to multiple drives per host, dual 1Gb or (later) single 10Gb.

SL8500

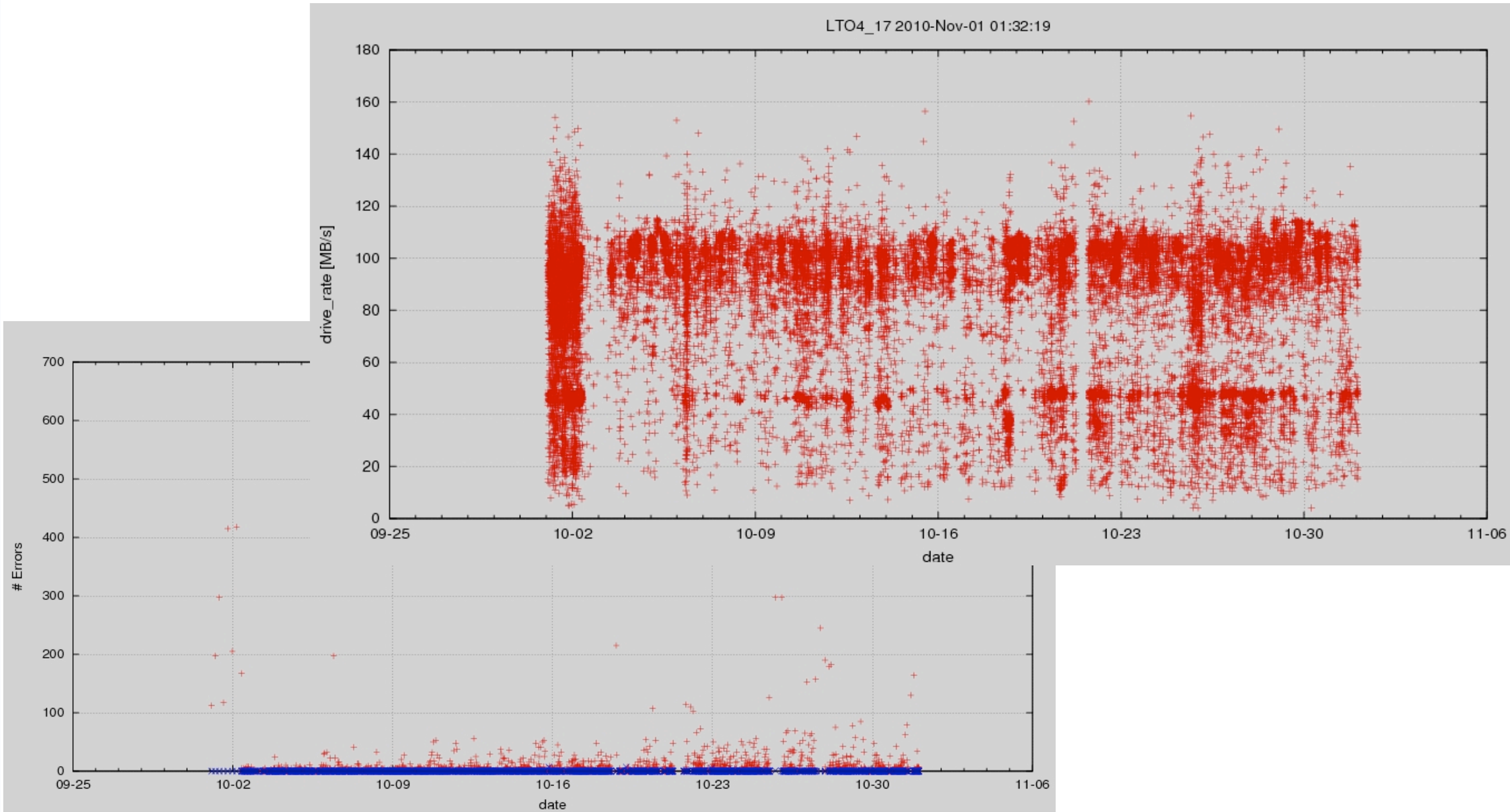
We have had a lot of problems with the SL8500 “handbots” failing in service.

- Diagnosis: long tracks in the 10k slot units are moved by ‘bot acceleration.
  - ECO installed.

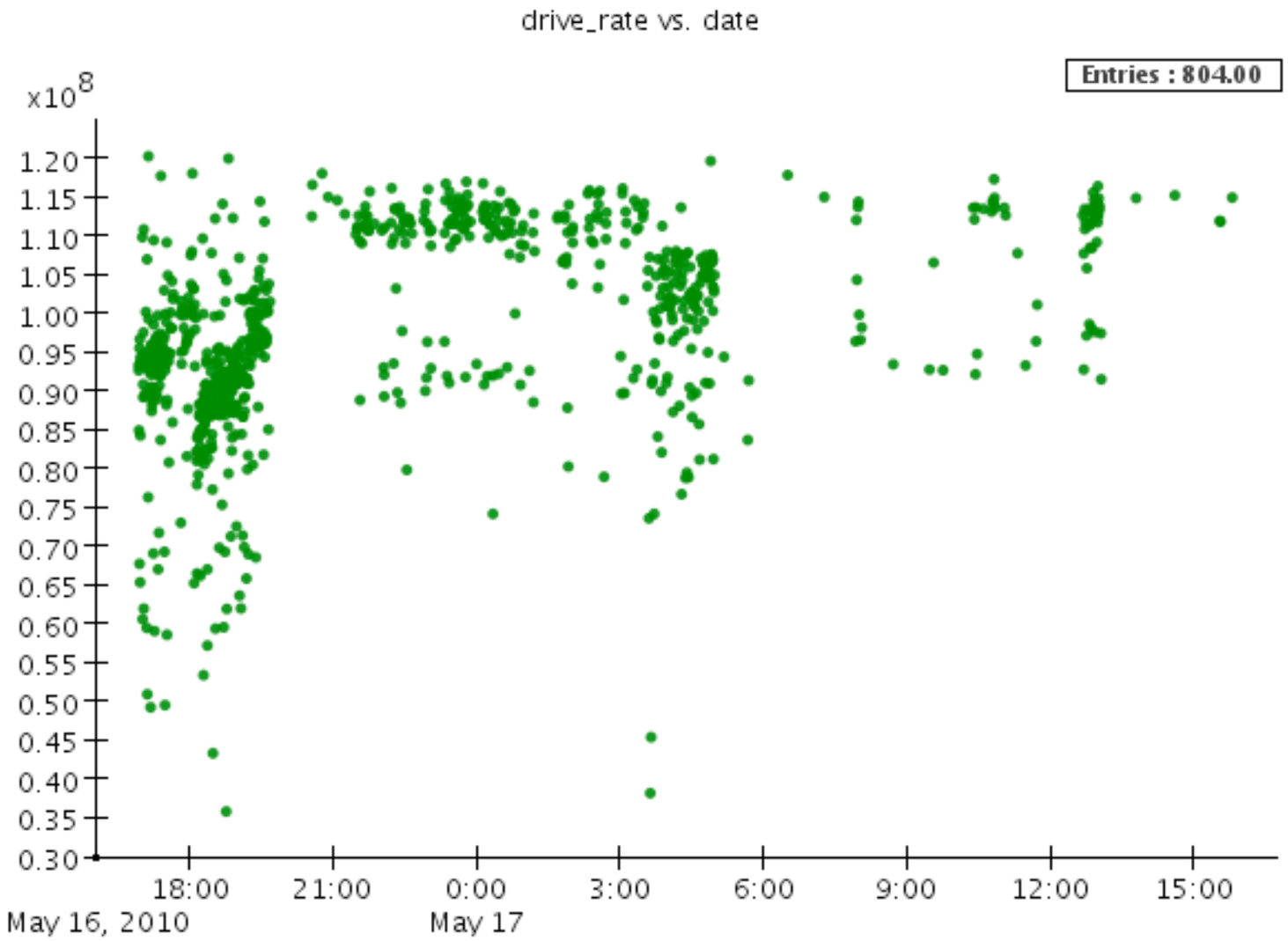
Installed dual bots for CMS on 21 Oct.

- 1 Nov, 3 AM: “saved” from a page & outage due to failure of a bot...
  - ... it was one of the new ones.

LTO4



Drives develop slow transfer rates, fail "capacity test."

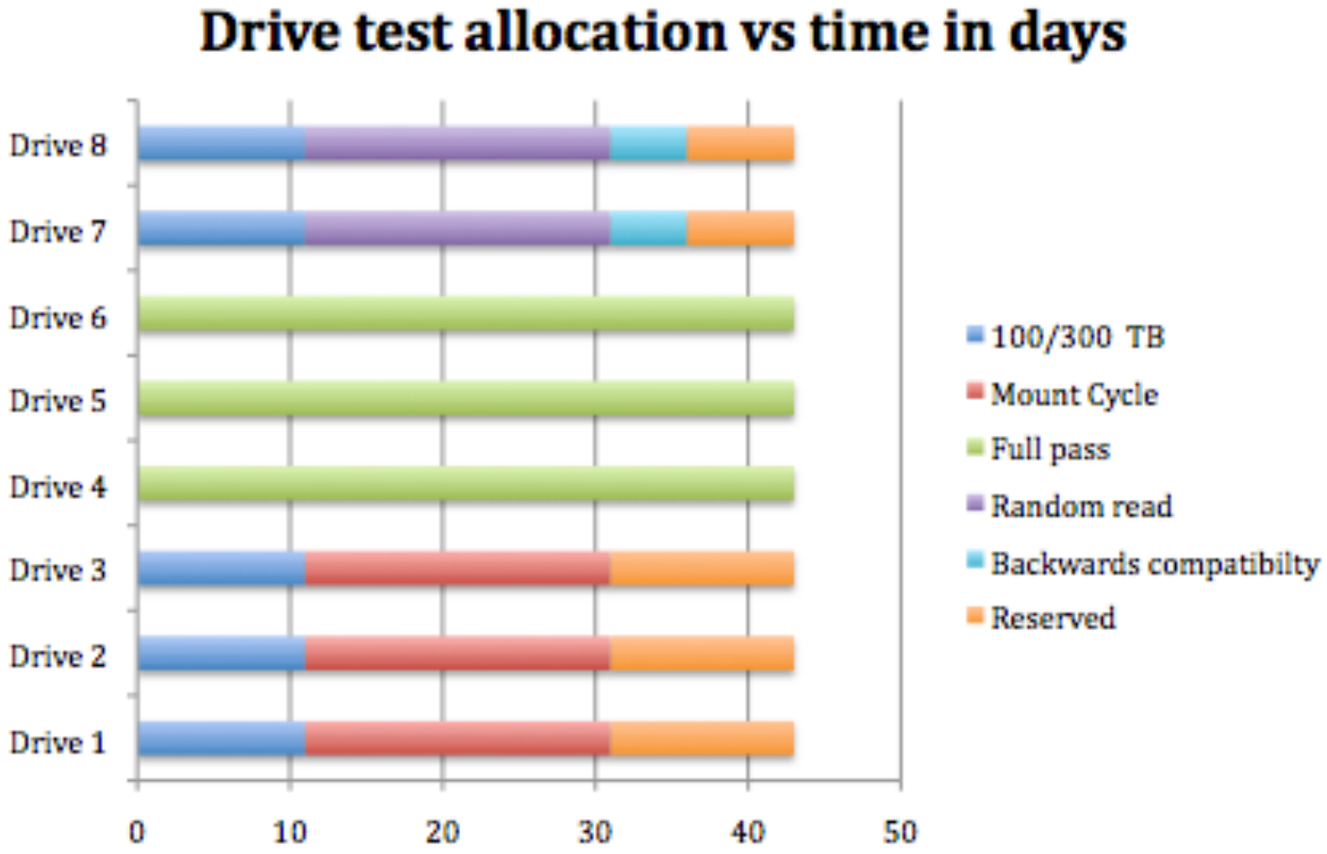


Replaced 52 drives in 5 months, 42 of them for this problem. Another 23 failed test last week.

No data loss, no tapes eaten by drives. Oracle replaces them promptly ... until yesterday.

LTO5 commissioning

8 LTO5 drives are installed. Further purchase order is lined up awaiting acceptance testing, begun last week.



Write 100TB/ Read 300TB

8000 mounts

260 media passes

20d random read

LTO4 compat.

Oracle

After the acquisition of Sun, it was very difficult for a while to get maintenance quotes from Oracle or its maintenance resellers.

Apparent changes of emphasis in disk-based storage product already noted.

Commitment to tape product line seems strong.

Crystal Ball

Tape

- May make Enstore see internals of libraries to increase the request rate, but at the cost of complexity.
- Will take a serious look at “enterprise” drives like T10000C. Drivers may be density and/or LTO reliability.

Crystal Ball

Disk

- ZFS is very attractive at the file system layer. How to get it?
  - OpenSolaris? FreeBSD? Wait for Linux port?
- Expect to see an NFS – xrootd shootout.
- Hope to see a good Lustre consortium.