Muon g-2 and Slow Tape Staging

Adam Lyon

Fermilab

Discussion about Slow Tape Staging 2022-10-03

Slow tape access

If you look at the Active volumes page you may see something like (at 10:30pm last night)...

```
CD-LTO8G2.library_manager
label      mover              tot.time status              system_inhibit      rq. host         updated           volume family
FM3974L8   L8G2_032.mover     246      DISMOUNT_WAIT (8    ) (none         none) stkendca1907     10-02-22 22:27:39 nova.raw2root_fd_ddenergy.cpio_odc
FM3973L8   L8G2_008.mover     242      HAVE_BOUND    (47   ) (none         none) stkendca1906     10-02-22 22:28:02 nova.raw2root_fd_t05.cpio_odc
FM5704L8   L8G2_002.mover     3567     SEEK          (63   ) (none         full) stkendca1813     10-02-22 22:27:38 icarus.icarus.cpio_odc
FL7882L8   L8G2_007.mover     525      SEEK          (91   ) (none         full) stkendca61a      10-02-22 22:27:10 GM2.gm2_daq.cpio_odc
FM4752L8   L8G2_014.mover     1557     SEEK          (0    ) (none         full) stkendca1814     10-02-22 22:27:58 GM2.gm2_5220A.cpio_odc
FL7886L8   L8G2_030.mover     174      HAVE_BOUND    (8    ) (none         full) stkendca67a      10-02-22 22:28:01 GM2.gm2_daq.cpio_odc
FM6920L8   L8G2_033.mover     8992     ACTIVE-WRITE  (0    ) (none         none) stkendca2025     10-02-22 22:27:31 GM2.gm2_5310A.cpio_odc
FL7892L8   L8G2_012.mover     39       MOUNT_WAIT    (14   ) (none         full) stkendca1818     10-02-22 22:27:38 GM2.gm2_daq.cpio_odc
FM5586L8   L8G2_018.mover     4014     ACTIVE-READ   (0    ) (none         full) stkendca66a      10-02-22 22:27:51 icarus.data_artroot_raw_offbeambnbminbias.cpio_odc
FM4779L8   L8G2_006.mover     19599    SEEK          (7    ) (none         full) stkendca2218     10-02-22 22:27:59 GM2.gm2_5220A.cpio_odc
FM3975L8   L8G2_009.mover     88       SEEK          (49   ) (none         none) stkendca1906     10-02-22 22:28:02 nova.raw2root_fd_t05.cpio_odc
FM5204L8   L8G2_031.mover     1560     SEEK          (0    ) (none         full) stkendca55a      10-02-22 22:27:57 icarus.data_artroot_raw_offbeambnbminbias.cpio_odc
FL7915L8   L8G2_010.mover     574      SEEK          (0    ) (none         full) stkendca2018     10-02-22 22:27:51 GM2.gm2_daq.cpio_odc
FM4107L8   L8G2_001.mover     373      SEEK          (35   ) (none         full) stkendca68a      10-02-22 22:27:20 mu2e.phy-sim.cpio_odc
FL7807L8   L8G2_025.mover     19       MOUNT_WAIT    (0    ) (none         full) stkendca1814     10-02-22 22:27:44 nova.nova_production.cpio_odc
FL7912L8   L8G2_013.mover     435      SEEK          (0    ) (none         full) stkendca2045     10-02-22 22:28:01 GM2.gm2_daq.cpio_odc
FL7867L8   L8G2_026.mover     16978    SEEK          (28   ) (none         full) stkendca58a      10-02-22 22:27:22 GM2.gm2_daq.cpio_odc
FL7735L8   L8G2_019.mover     505      SEEK          (14   ) (none         full) stkendca2019     10-02-22 22:27:22 GM2.gm2_daq.cpio_odc
FM4762L8   L8G2_021.mover     21440    SEEK          (21   ) (none         full) stkendca2218     10-02-22 22:27:26 GM2.gm2_5220A.cpio_odc
FM5438L8   L8G2_035.mover     21690    ACTIVE-READ   (35   ) (none         full) stkendca62a      10-02-22 22:27:55 icarus.data_artroot_raw_offbeambnbminbias.cpio_odc
FL7903L8   L8G2_024.mover     491      SEEK          (21   ) (none         full) stkendca2022     10-02-22 22:27:07 GM2.gm2_daq.cpio_odc
FM5632L8   L8G2_017.mover     9398     SEEK          (0    ) (none         full) stkendca1819     10-02-22 22:28:02 icarus.data_artroot_raw_offbeambnbminbias.cpio_odc
FL7917L8   L8G2_015.mover     565      SEEK          (0    ) (none         full) stkendca2019     10-02-22 22:27:00 GM2.gm2_daq.cpio_odc
FM4803L8   L8G2_003.mover     26713    SEEK          (42   ) (none         full) stkendca1814     10-02-22 22:27:20 GM2.gm2_5220A.cpio_odc
FL7878L8   L8G2_023.mover     393      SEEK          (0    ) (none         full) stkendca70a      10-02-22 22:27:43 GM2.gm2_daq.cpio_odc
FM5483L8   L8G2_020.mover     60       SEEK          (7    ) (none         none) stkendca1908     10-02-22 22:27:48 nova.raw2root_fd_ddnnbar.cpio_odc
FM4621L8   L8G2_005.mover     21747    SEEK          (28   ) (none         full) stkendca1911     10-02-22 22:27:48 GM2.gm2_5220A.cpio_odc
FM5683L8   L8G2_011.mover     3679     ACTIVE-READ   (7    ) (none         none) stkendca57a      10-02-22 22:27:45 minerva.mc_reconstructed.cpio_odc
FM6919L8   L8G2_034.mover     6554     ACTIVE-WRITE  (0    ) (none         none) stkendca55a      10-02-22 22:27:50 GM2.gm2_5310A.cpio_odc
FL7858L8   L8G2_028.mover     16882    SEEK          (7    ) (none         full) stkendca55a      10-02-22 22:27:31 GM2.gm2_daq.cpio_odc
FL7741L8   L8G2_022.mover     19878    ACTIVE-READ   (7    ) (none         full) stkendca65a      10-02-22 22:28:00 GM2.gm2_daq.cpio_odc
FL7907L8   L8G2_029.mover     429      SEEK          (28   ) (none         full) stkendca2022     10-02-22 22:27:32 GM2.gm2_daq.cpio_odc
FM3969L8   L8G2_016.mover     189      SEEK          (98   ) (none         none) stkendca1908     10-02-22 22:27:05 nova.raw2root_fd_ddslowmono.cpio_odc
FL7879L8   L8G2_027.mover     23       MOUNT_WAIT    (7    ) (none         full) stkendca1818     10-02-22 22:27:47 GM2.gm2_daq.cpio_odc
FM4813L8   L8G2_036.mover     12438    SEEK          (0    ) (none         full) stkendca2010     10-02-22 22:27:01 GM2.gm2_5220A.cpio_odc
FM5390L8   L8G2_004.mover     1842     SEEK          (0    ) (none         full) stkendca68a      10-02-22 22:27:34 icarus.icarus.cpio_odc
```

Muon g-2 is using 22 out of 36 tape drives (~60%) at that moment.

Muon g-2 using a lot of tape drives seems to be common.

Why is Muon g-2 using so many tape drives?

- The Fermilab director has told the collaboration to complete the experiment and analyses as soon as possible.
- We are two years behind in our production.
- When she asked us what we would do to fulfill her mandate, we replied that we would double our production rate.
- We are successful with a lot of help from SCD (now CSAI) (thanks!)

What are we doing now?

- We are currently running full production on our Run 4 data (taken in FY21) as fast as the system will let us.
  - Run 4: Prestaging raw data from tape and writing processed files to tape
- We are currently running pre-propduction on our Run 5 data (taken in FY22), but more slowly.
  - Run 5: Prestaging raw data from tape (processed files go to tape, but we do the analysis before they leave the cache)
- We are completing our Runs 2/3 data analyses for publication
  - Re-prestaging some Run 3 data that we reprocessed earlier that fell off the cache
- We still have Run 6 data to collect and process (FY23).

We've gotten good at production!

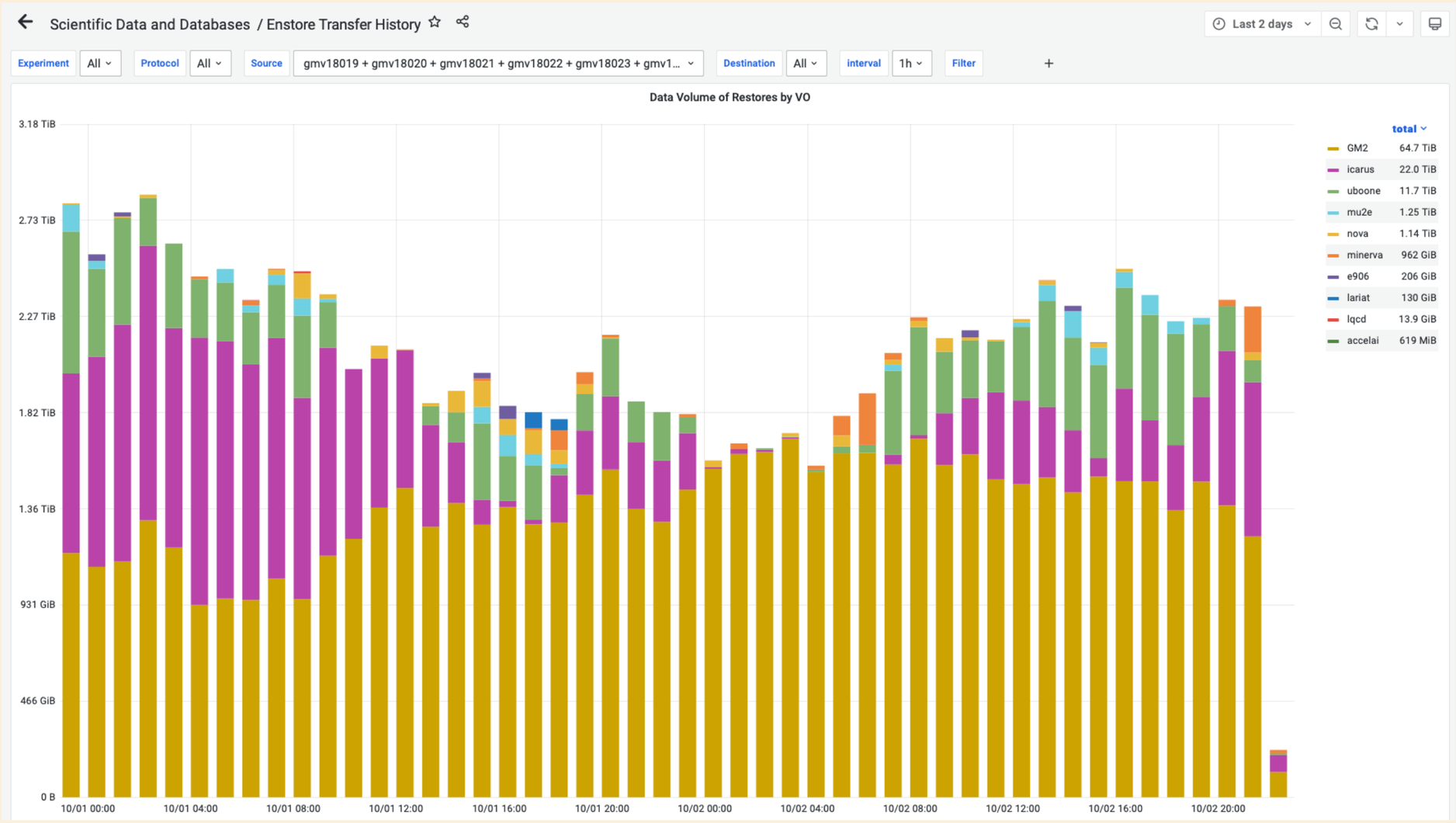
- We make very efficient use of POMS
- We have production shifters to keep jobs going all the time
- We do a rolling production strategy
  - A run is split into datasets (Run 4 has 22 datasets; Run 5 21 datasets)
  - While running jobs on a dataset, we are prestaging the raw data files for the next one
  - We effectively prestage all the time
- We are careful about what we prestage (remember that killing a prestaging job does *not* remove the requests from dCache)

- Run 4 is 2.25 PB of raw data (1.1M raw data files) (~4 × Run2, ~2 × Run 3)
- Run 5 is 3.4 PB of raw data (1.5M raw data files)

Other experiments are not shut out

See https://lsvip.fnal.gov/monitor/goto/eZ6U-d4Vk?orgId=1 for data volume by IF VO for the LTO8G2 library.

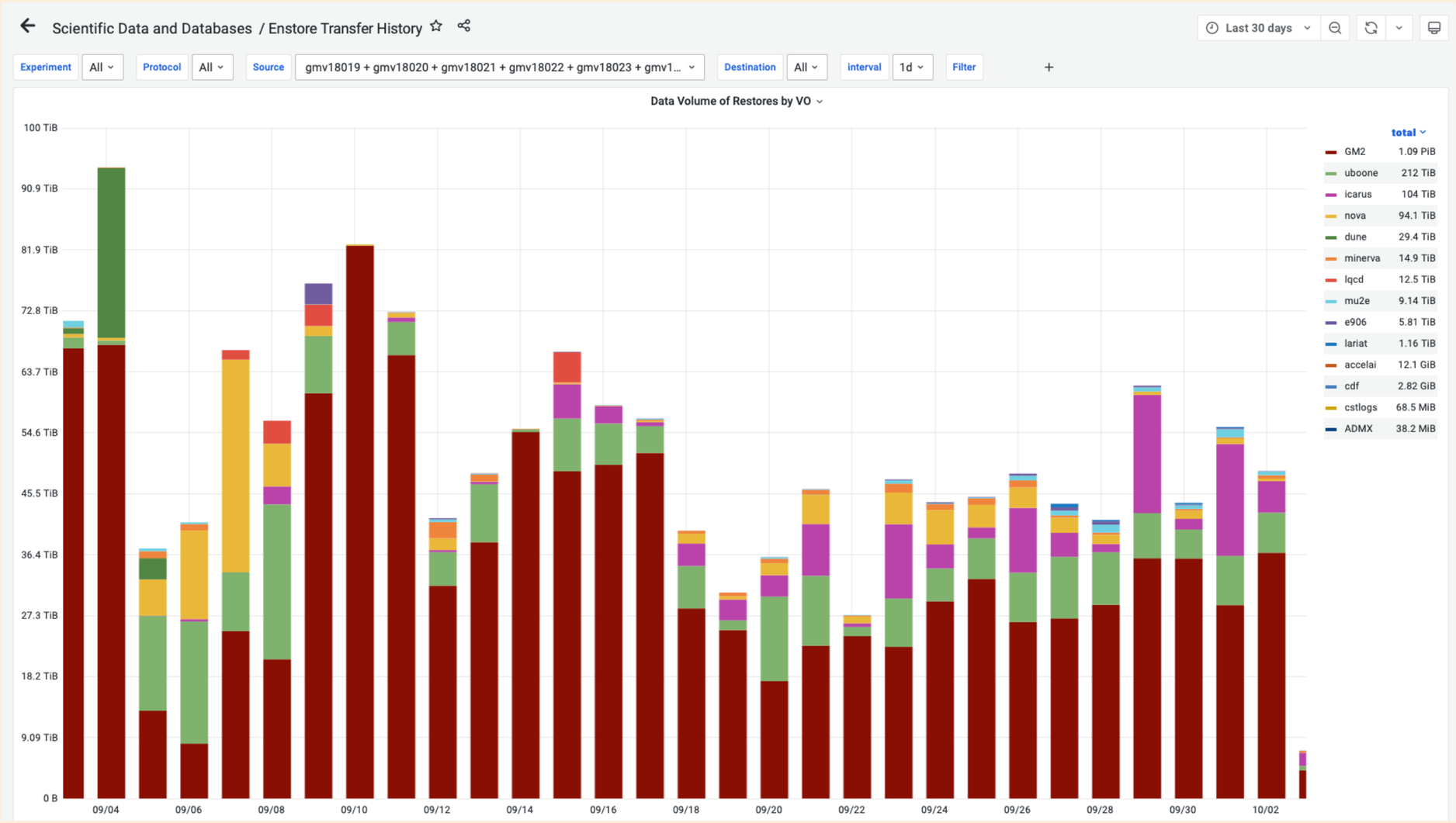
Here's this weekend (as of 10:30pm Sunday night)



Other experiments are not shut out

See https://lsvip.fnal.gov/monitor/goto/eZ6U-d4Vk?orgId=1 for data volume by IF VO for the LTO8G2 library.

Here's the past month

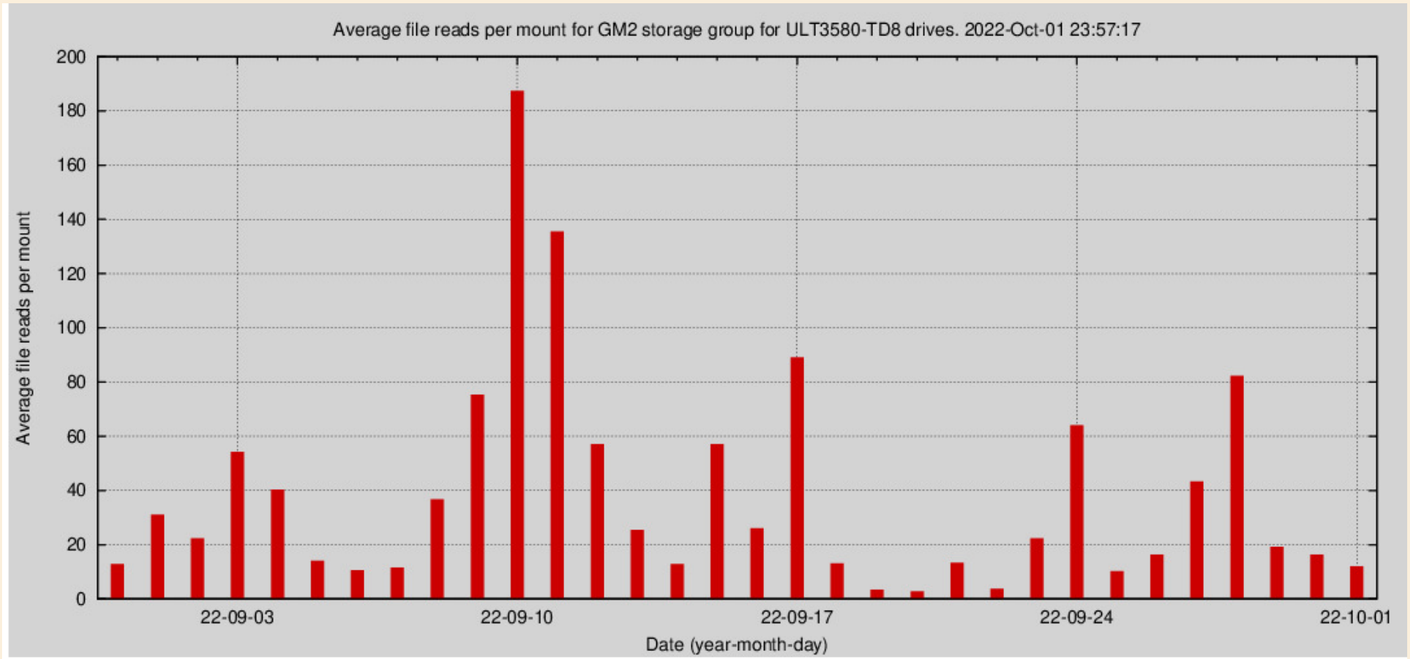


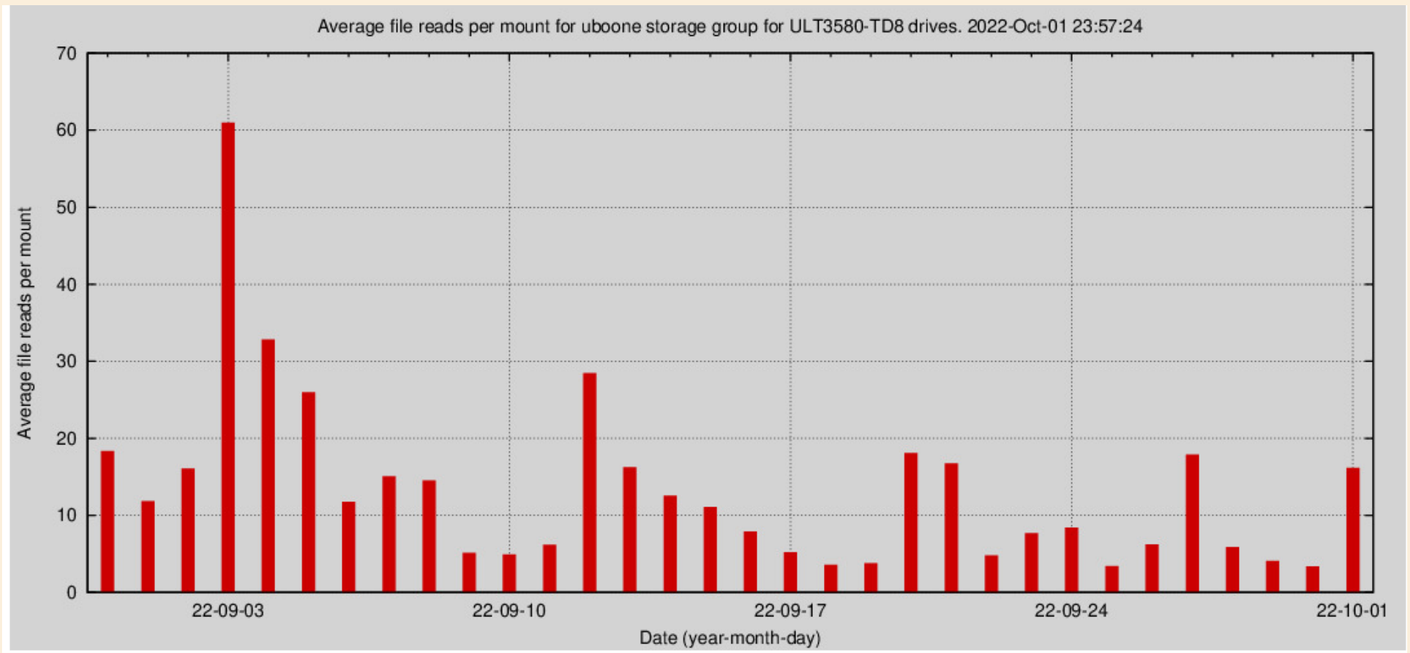
Remember the limitations of the system

- The Enstore queue is **much** smaller than the dCache queue
- Only the Enstore queue groups files by tape volume (so they are read together)
- Small Enstore queue leads to inefficient use of tapes, many re-mounts, and lots of seeking

- Hopefully this can be fixed by the new CTA system and we'll use tape drives much more efficently

Poor tape drive efficiency





Things we are doing now to help

- We know your pain - we've been there (often)!

- We expect to finish Run 4 full production by the end of October
  - Prestaging Run 4 raw will complete a week or two before then
- We're prohibiting analyzers from prestaging large datasets
- We're delaying prestaging of Run 4 *analysis* datasets
  - Though many Run 4 production files are likely still on disk

Things we're doing soon to help

- Writing of our production files will soon move to the new library (TFF2)
  - Will help to alleviate the huge backlog of tape writes
  - Reduce pressure on the LTO8G2 library since writes compete with reads
- CSAI will move our Run 5 raw data tapes to the old LTO8F1 library
  - LTO8F1 holds older data and is used much less than LTO8G2
  - Our Run 5 prestaging will no longer compete with your tape reads
- We will institute a "real-time" pre-production strategy for Run 6
  - We will process Run 6 raw data for pre-production before files leave our write pool
  - No prestaging necessary for Run 6 pre-production
  - Prestaging will be necessary for full production, but we expect Run 6 to be a smaller run