NEW YORK
GENOME CENTER®

30PB Replicated Genomic Data Archive using Spectrum Archive

Spectrum Days 2022
New York Genome Center
Tausif Hasan and Chris Black

Why Tape?

Two main reasons for us right now:

Cost Model

Variable incremental cost (duplicated media) is <$4/TiB for 5 years

Movement

NYGC generates raw data from instruments, nice to avoid internet or interconnect to archive

No fees or limits on recall

Security

Increased data redundancy via off-site storage solutions at a secure location

Theory of Operation - GPFS Pools

External pools and migration scripts

Spectrum Archive software allows adding “tape pools” as external pools to a GPFS filesystem. (*Use can be extended as nearline-storage/hybrid cloud storage)*

Spectrum Scale ILM policy scripts move data from “staging disk” pool to “tape pools”. *(Policies are highly flexible)*

We have two tape pools, A and B - policy writes new data from staging disk to both pools before removing from staging disk. *(Even more pools can be added; in our case, pool B is for offsite redundancy)*

Theory of Operation - GPFS Pools

External pools and migration scripts

Small files (<10MB) and large files handled separately and run at different times

```
#snip…
RULE EXTERNAL POOL 'tapepool'
EXEC '/opt/ibm/ltfsee/bin/eeadm'
OPTS '-p SET_A@lib0 SET_B@lib0'

RULE 'ee_sysmig' MIGRATE FROM POOL 'draid'
TO POOL 'tapepool'
WHERE (FILE_SIZE > 1G)
AND (CURRENT_TIMESTAMP - ACCESS_TIME > INTERVAL '12' HOURS)
AND (is_resident OR is_premigrated)
AND NOT (user_exclude_list)
```

Theory of Operation - Data Recall

Priority of operations and export scripts

Data recalls are designated as high priority tasks in Spectrum Archive and will be serviced by the next available recall-enabled tape-drive in the tape library

Alternatively, a drive can be dedicated to handling recall tasks only to ensure requests will always be serviced regardless of other system operations on archive

```
eeadm recall <fofn.list> -p SET_A,SET_B --async
```

```
eeadm drive list
```

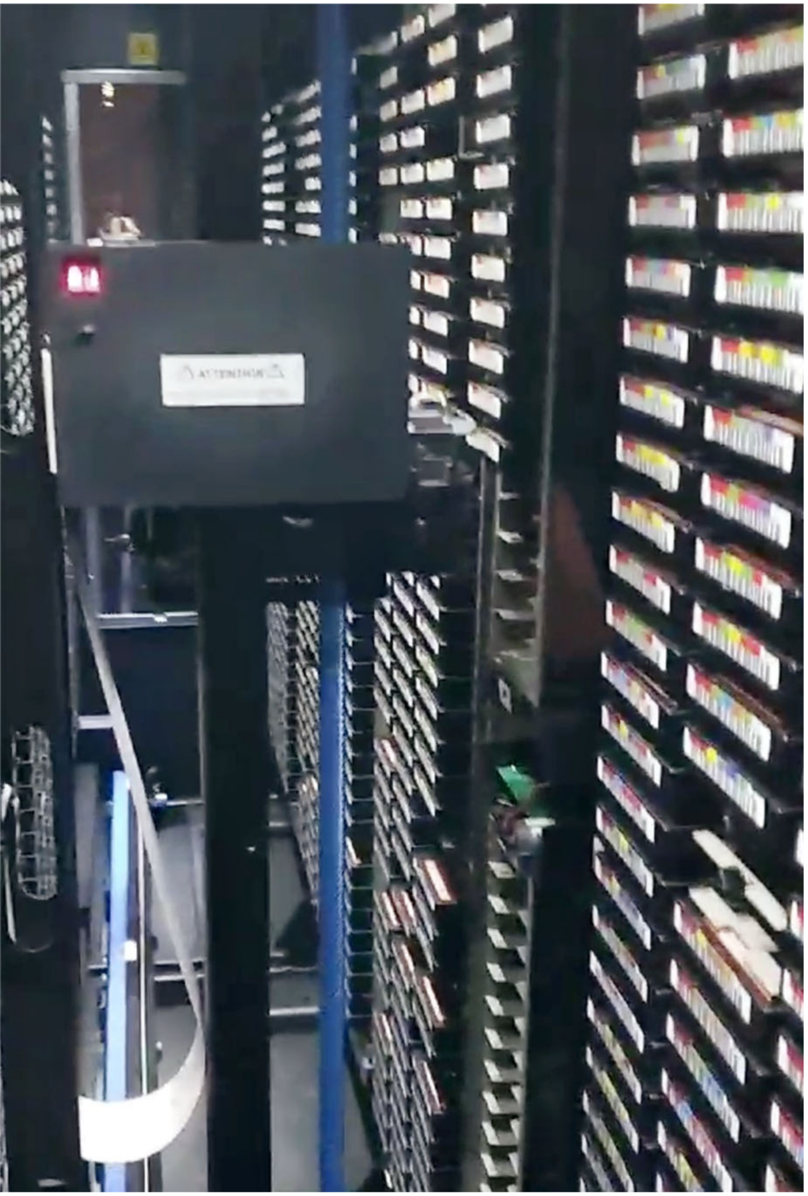
```
DriveS/N Status State   Type  Role Library NodeID  Tape    Task ID
TS1160A01  ok   in_use  TS1160 mrg  lib0       8   A09999JE   1337
TS1160A02  ok   mounted TS1160 -r-  lib0       8   A00001JE   1984
```

The Hardware – TS4500 Tape Library

- The NYGC Tape Library consists of five frames (model 3584-XX)
  - Frame 1 – 3584-D25 is an expansion frame hosting 590 slots and can host 12 tape drives
  - Frame 2 – 3584-L25 is the base frame of the library with two I/O slots for loading and unloading tapes, and can host 16 tape drives
  - Frames 3-5 – 3584-S25 are driveless, storage-only frames with a capacity of 1000 tape cartridges per frame
- There are currently **3315** tapes on-site out of a **4250** total slot limit with 24 tape-drives for read/write and approximately ~30 PiB in each pool
  - We have upgraded tape cartridges from Model 3592-EH8 (TS1150) to 3592-60F (TS1160), jumping from a base capacity of ~9T to ~18T per tape, which prompted the addition of twelve TS1160 Drives
  - A mix of 18 tape drives are available, 12 new TS1160s, and TS older 1150s to work thru reformat backlog

The Hardware (Cont.)

- In addition to the TS4500, there are x5 Lenovo-based archive-nsd nodes (Model 3650-M5) and Four Brocade switches (Model SAN24B-5 Type 2498-X24)
  - These switches handle communication between the archive-nsd nodes and tape-drives, as well as tape-drives and their communication to the TS4500 itself
  - Non-Data Mgmt operations can be carried out from the TS4500 for tape or tape-drive repair and maintenance
- Lastly, the FS7200 staging disk array consists of of x9 enclosures (Model 2076-24G) with one acting as the controller module
  - This array has replaced our previous v7000 array which primarily relied on 7200RPM HDDs and did not offer any flash storage
- Our Archive library has close to 30 PiB of storage capacity per pool, totaling to nearly ~60 PiB of redundant storage between the SET_A and SET_B tape pools



The Hardware (Honorable Mention)

This robot just needs to return some tapes…

Drives at top left, barcoded cartridges …everywhere else:
below drives, 5 deep horizontally
on doors

Picker handles 2 cartridges at a time

Storage Costs

- Annual License Costs
  - GPFS (Spectrum Scale Standard)
  - SKLM (Security Life Cycle Manager, now GKLM)
  - Spectrum Archive (Advanced Edition)
- Iron Mountain Storage
  - Delivery from site to site – Approx. $250 per trip
  - Service/Storage Fee – Approx. $30,000 per year
- Hardware Costs
  - Nodes (x86 w/ FC HBAs)
  - Staging disk: FS7200 System
  - Tape Library
  - Tape Drives
  - Incremental cost for adding one duplicated PB is about $37,000*
  - Low Power, high density

Storage Costs (Cont.)

How much does it cost to store 1 PB for 5 years with tape vs cheap disks vs HPC vs cloud storage?

A single consumer 10TB HDD can cost $200-250, nearly $200k+ for 1PB before factoring-in volume discounts

On-Prem scalable NAS systems can be expensive, easily reaching costs of $300k+ per PB before factoring in maintenance and renewal fees

Over five years, the cost can reach upwards of $2,000,000 without including other licensing/software fees, nor offering off-site redundancy or cloud backups

Tape:
<$40/TiB media (duplicated)
<$6/TiB offside media storage

Where does Iron Mountain come in?

Our tape storage is spilt into two pools,

- SET_A (2750+ tapes)
- SET_B (3000+ tapes)
- SET_A remains at NYGC within the TS4500 and allows for a nearline-like solution for data archiving and retrieval
- SET_B works more as a cold-storage solution, where tapes are sent off-site to IronMountain for storage. They can be recalled in the event of errors in the primary set, or if the primary site is irrevocably lost in a disaster scenario
- We spend ~$6/TB per 5 years for offsite storage

Data Ingestion (Migration)

Spectrum archive uses variations on the format, “eeadm <command>” (previously ltfsee <commands>) to perform data operations on tape

The primary team responsible for ingesting data (migrates) and retrieving data (recalls) from archive is the Operations Team (Taxi)

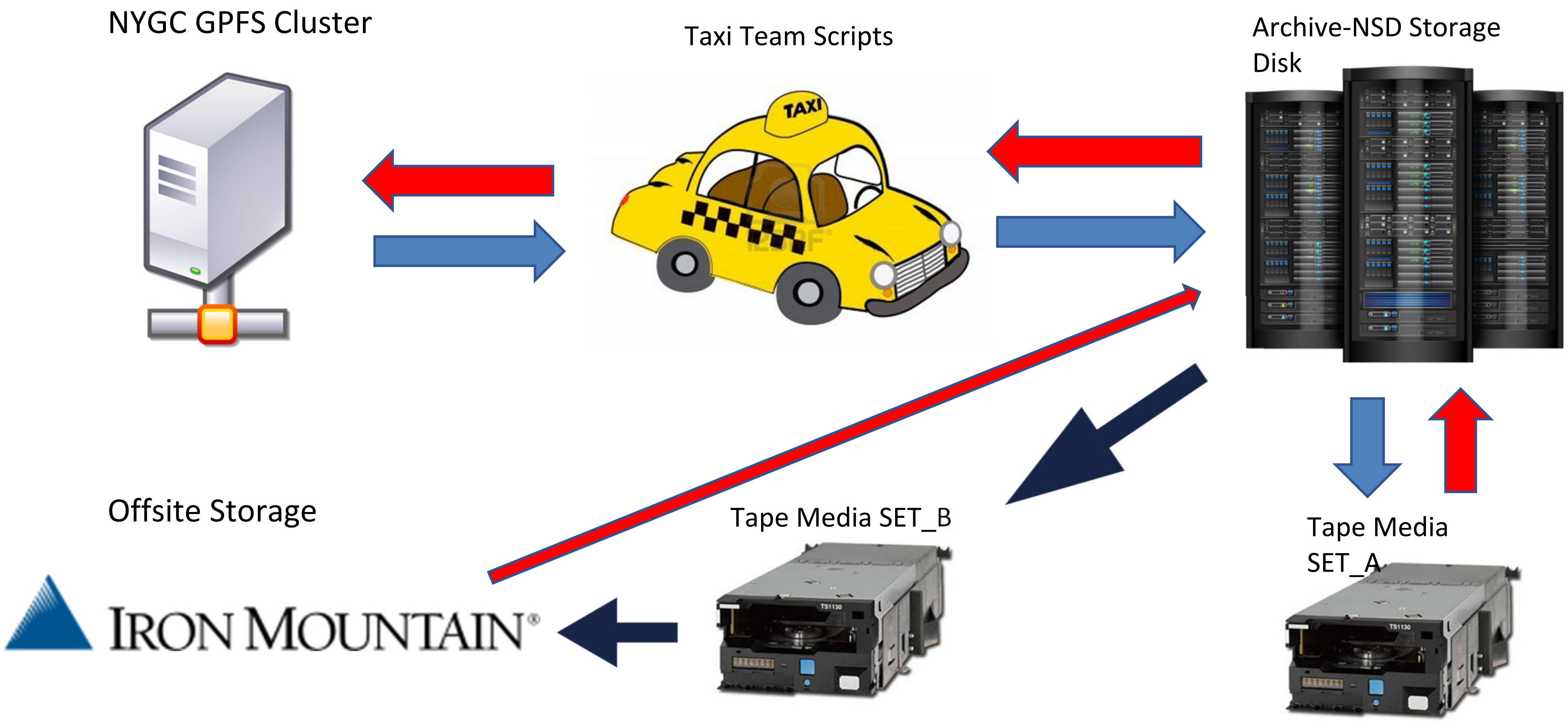
They handle scripting which selects directories on HPC GPFS that are up for archiving, and push these out I/O nodes to the staging disks similar to a regular file-system

From the staging disk, the archive-nsd nodes leverage Spectrum Archive and GPFS policy loops to identify newly eligible candidates on staging disk, and begin writing (migrating) that data to tape on both SET_A and SET_B

Data Recall

- Data already on tape that needs to be retrieved can be recalled by the Taxi team or Spectrum Archive administrators
- Once a fofn (file of file names) is generated, the recall job is submitted as pipeline to an archive-nsd node, which then uses Spectrum Archive and GPFS to begin copying files from tape to staging disk
- After the recall is complete, the data can then be cp'd or rsync'd from /gpfs/archive back to its primary location under /gpfs/otherfs

Data-Flow Diagram



NYGC GPFS Cluster

Taxi Team Scripts

Archive-NSD Storage Disk

Offsite Storage

Tape Media SET_B

Tape Media SET_A

Security & Encryption All tapes are encrypted

Archive uses Security Life Cycle Manager or SKLM for encryption key management between the TS4500, nsd-nodes, and its TS1150/TS1160 drives for end-to-end encryption

SKLM is run on archive-nsd2 & archive-nsd3

These are redundant in an active-passive configuration

This software can be run on a VM, but your mileage will vary...

Common Archive Operations

A Spectrum Archive Administrator carries out the following tasks to assist the Operations Team

- Monitoring space/health on staging disk
- Monitoring migration policies to ingest data from staging to tape
- Running/diagnosing recalls in the event of non-procedural errors in file retrieval
- Investigating GPFS file errors which prevent migrate/recall operations
- Generating tape cartridge candidates for off-site storage from secondary pool, SET_B

- Adding/removing tape and tape-drives for upgrade and maintenance
- Reformatting older JD tapes to increase overall storage capacity via reclamation jobs*
- Optimizing tape-drive roles to serve and ingest data
- Resurrecting Spectrum Archive and GPFS services in event of a failure
- Maintaining GPFS code-base w/ Spectrum Archive, in addition to related hardware firmware on nsd-nodes, Fabric Switches, the TS4500, and its constituent parts

Coming up Next for Archive?

Archive has gone through many changes in a few short years, but more upgrades are soon to come!

Globus endpoint with intelligent staging

Reformatting all remaining JD tapes, increasing capacity from ~9T to ~13.5T

Moving SKLM to a more dedicated system/VM

Generating VMs to run Spectrum Insights for a more in-depth look at data/cost/performance

Re-enabling more tape drives for max bandwidth (currently at 20 drives enabled)

Adding an additional frame for more tape capacity + storage

More Documentation & Knowledge transfer with Daniel Martinez, our newest rescomp and Spectrum Archive admin!

Recall Improvements

Although on-demand “transparent” recalls work, running lists of files (fofn) through “eeadm recall” and *then* starting filesystem reads is much more efficient.

One bit of automation we have around staging from tape pool back to staging disk via eeadm recall is requirement that each line start with “-- /some/path” rather than just “/some/path”.

Migration Automation and Monitoring

Currently a custom script generating and running “mmapplypolicy” commands. Very flexible and works well after initial tuning.

Would suggest more assistance and templates for the “migration loop”.

Globus

We utilize Globus for data transfers to many collaborators and repositories.

We've kept most users away from direct filesystem access to archive, for fear of storms of unoptimized tape activity.

There is now a research implementation of a Globus connector that would efficiently stage data in and out of tape pools prior to starting network transfer - this is exciting for us as it may allow:

More automation for "send archive data to collaborator" use case

User portals to archive data