HPSS Archive Storage Technologies for Easing the Friction of Data Movement.

Jim Gerry
jgerry@us.ibm.com
IBM HPSS Senior Architect & Consultant

Easing the friction of moving data at extreme-scales.

Top Publicly Disclosed HPSS Sites as of 4Q2019		PB	M Files	Since
ECMWF	European Centre for Medium-Range Weather Forecasts	451.14	312.95	2002
UKMO	United Kingdom Met Office	356.75	538.83	2009
SSC	Shared Services Canada	210.76	25.95	2017
NOAA-R&D	National Oceanic and Atmospheric Administration Research & Development	181.71	98.31	2002
LBNL-User	Lawrence Berkeley National Laboratory - User	181.53	233.03	1998
BNL	Brookhaven National Laboratory	172.73	191.40	1998
Meteo-France	Meteo France - French Weather and Climate	162.69	516.97	2015
MPCDF	Max Planck Computing and Data Facility	142.25	278.78	2011
CEA TERA	Commissariat a l`Energie Atomique - Tera Project	130.37	24.65	1999
DKRZ	Deutsches Klimarechenzentrum	116.28	23.25	2009
ANL	Argonne National Laboratory	105.76	463.12	2008

- Sample HPSS production readiness workload demonstration (November 2019)
  - In 24-hour we ingested 1 PB to tape while simultaneously recalling 791 TB from tape.
  - HPSS pushed FOUR 13-frame IBM TS4500 tape libraries (scheduled to house over 500 PB of tape media) to 2,168 mounts per hour.

Easing the friction of moving large and small files on tape.

- Friction:
  - At 400 MB/s, it takes 42 minutes to read-or-write a 1 TB file.
  - Small files are not tape friendly, so they kill tape drive performance.
- Technologies & Advantage for large files
  - HPSS RAIT (tape stripe with rotating parity blocks).
  - Scales tape transfers and cuts redundant tape cost.
- Technologies & Advantage for small files
  - Automatic small-file aggregation on tape
    - Enables near-native small-file tape writes.
  - Automatic logical grouping of files by directory.
    - Reduces tape seeks on recall.
  - Full-aggregate recall (FAR).
    - Enables near-native small-file tape reads.

9x

Easing the friction reading files on tape.

- Friction:
  - Recalling files from tape takes so long.
  - Lots of tape movement and very little data movement.
  - Same tape gets mounted over and over.
  - Single tape is in the drive for a long time.
- Technology: HPSS Tape Ordered Recall (TOR)
  - Sort tape recalls by cartridge to minimize tape mounts.
  - Sort the files being recalled from each cartridge.
    - Offset Ordered – BETTER than random.
    - Recommended Access Ordered (RAO) – BEST
- Advantage:
  - Reduce the number of tape mounts.
  - Spend less time with each tape.
  - Recall files faster and with fewer tape drives.

File Count	Offset Ordered (minutes)	RAO (minutes)
50	30.43	6.22
100	52.55	10.27
150	81.27	16.97
200	96.71	17.32
250	122.29	20.10

Offset ordered vs RAO recalls compared by Cristina Gabriela Moraru, Master Thesis

Offset Ordered 2x

RAO 5x

Easing the friction for data integrity.

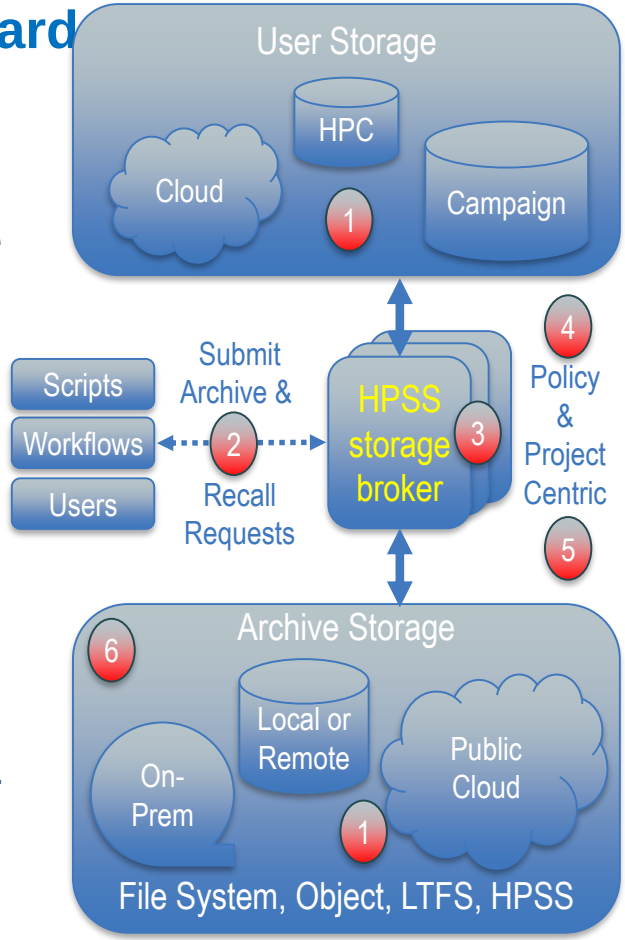
- Friction:
  - Files must be recalled from tape and validated to ensure they were properly written to tape – this can be costly.
- Technology: HPSS end-to-end data integrity.
  - When a file is written to HPSS tape:
    - The file checksum is calculated and verified.
    - An HPSS CRC for each tape block is inserted into the data-stream.
    - Using the HPSS CRC, the tape drive performs a read-after-write validation of each block written to tape.
    - Checksum or CRC write error results in an HPSS tape write failure.
- Advantages:
  - No longer necessary to recall files from tape to ensure they were properly written.
  - Full-tape data re-validation is done by the drive at drive speed.

File Checksum
+
HPSS CRC
+
T10-LBP Read-After-Write
=
True Data Integrity

Low Overhead
Data Re-Validation

Easing the friction for data movement going forward

- Friction: numbers and volumes of HPC storage systems.
  - Storage often changes with each new HPC procurement.
  - A blend of file system, object and tape storage that spans multiple generations is often used to support HPC.
  - No single interface to efficiently archive and recall datasets.
- Technology: HPSS storage broker feature
  1. Support for local & remote, on-premises & cloud, HPSS, LTFS, File Systems, and Object storage using the jcloud toolkit.
  2. Interface for submitting & monitoring archive & recall requests.
  3. Requests are sorted and optimized.
  4. Policy-centric automation for dataset grouping, placement, protection (multi-copy or parity) and access preferences.
  5. Project-centric ownership and management of archived datasets.
  6. Intentional separation of User Storage and Archive Storage to simplify access and management of archived datasets.



Thank You!

Jim Gerry

jgerry@us.ibm.com

IBM HPSS Senior Architect & Consultant