Best Practices in Accessing Tape-Resident Data in HPSS

David Yu, Guangwei Che, Tim Chou, Ognian Novakov

Brookhaven National Laboratory

Abstract

Tape is an excellent choice for archival storage because of the capacity, cost per GB and long retention intervals, but its main drawback is the slow access time due to the nature of sequential medium. Modern enterprise tape drives now support Recommended Access Ordering (RAO), which is designed to improve recall/retrieval times.

BNL's mass storage system currently holds more than 100 PB of data on tapes, managed by HPSS. Starting with HPSS version 7.5.1, a new feature called "Tape Order Recall (TOR) has been introduced. It uses the RAO mechanism for improving the access times over sequential media claiming a performance improvement by 30% to 60%.

Prior to HPSS 7.5.1, we have been using an in-house developed scheduling software, called ERADAT. ERADAT accesses files based on the order of the logical position of each file and consequently has demonstrated great performance at BNL. We have compared the new TOR method to the method of using logical position under different conditions such as number of access requests.

In this presentation we will demonstrate a series of tests, which indicate how effective the TOR (RAO) is under different scenarios and what are the best methods in restoring data from tape storage under different conditions.

Purpose

As the amount of expected data generated from scientific experiments are increasing, the need of utilizing tape storage has becoming more cortical. We need to re-evaluate the new technologies introduced by vendors, and find the best practice of utilizing the available options and use them effectively in production.

Starting from HPSS 7.5.1 release, a new feature named Tape Ordered Recall (TOR) has offered a great advantage for large amount of data recall. It is designed to reduce the file access time from tape media. Additionally, Quaid is a smart recall tool introduced in HPSS7.5.1 as well.

To better understand and evaluate the data recall performance in HPSS7.5.1 release, we have performed tape recall test with different features and tools. The test results and conclusion for each individual tool under certain test setting and configuration are discussed in this paper.

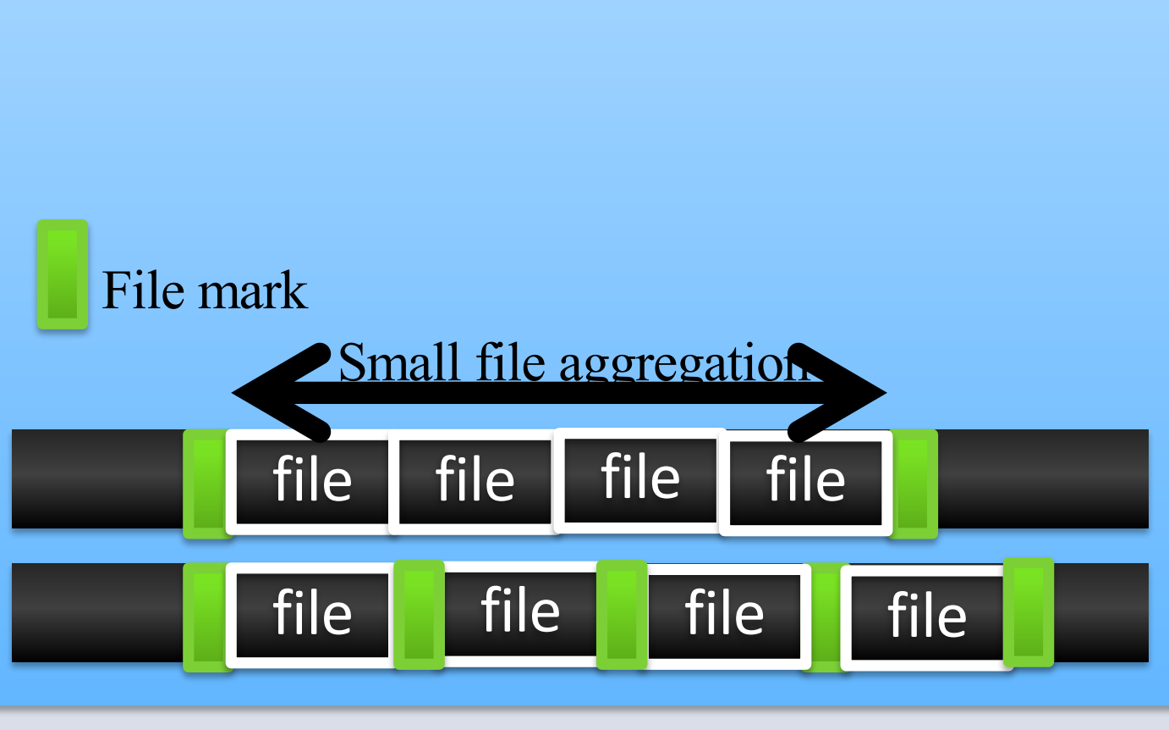
Environment

Software and Equipment

- HPSS core and gateway server: RHEL6.9, HPSS7.5.1p2u1, DB2v10.5p8
- HPSS mover and pftp client server: RHEL6.9, HPSS7.5.1p2u1
- Tape Library: Oracle StorageTek SL8500 Modular Library System
- Oracle Automated Cartridge System Library Software (ACSLS): 8.4.0
- Tapes and drives:
  - Oracle T10K-D: 8 TB Capacity, Max claimed native data rate: 250 MB/sec[1]

Data sets

- Large File: 10 G per file. 800 files per tape
- Small files: 1 G per file, 8000 files per tape
  - With "Small files aggregation"
  - Without "Small files aggregation"



Staging Methods and Comparisons

Staging modes:

Default

by default TOR (Tape Ordered Recall) is enabled, HPSS will use the enterprise drive built in feature RAO (Recommended Access Order, if available from the drive) or make linear offset ordering (if RAO not available), to schedule and submit the staging requests.

Noschedule

TOR is disabled, staging request will not be ordered before submitting to HPSS.

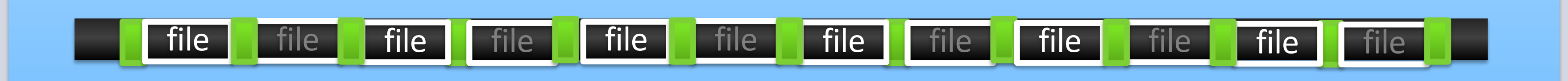
Nodrive

RAO is disabled, staging request will be processed based on linear offset ordering before submitting to HPSS.

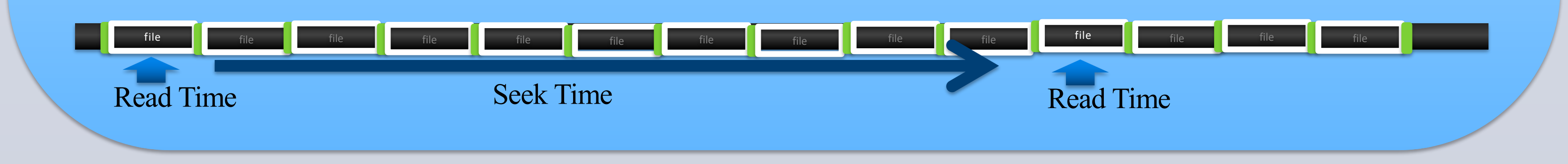
Staging coverage :

100 % stage: Every files on the tape will be recalled.

50% stage: Half of the files on the tape will be recalled – Every other file



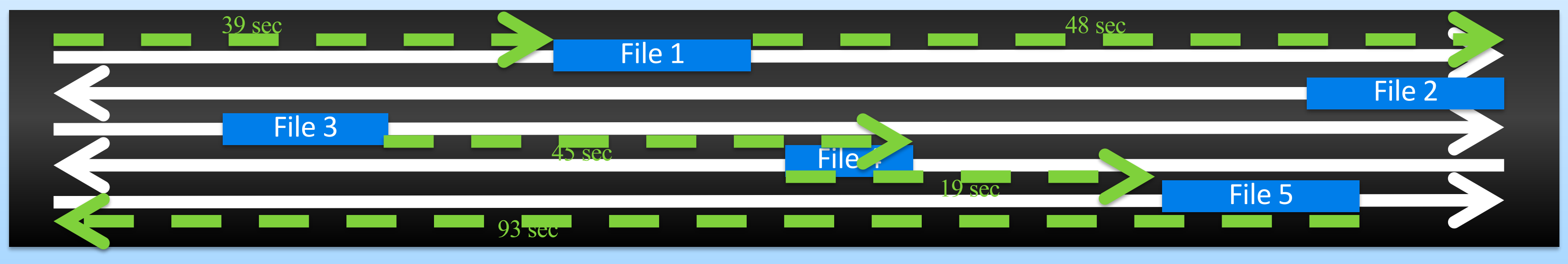
10% stage: 10% files on the tape will be recalled – Skip every 10% -1 files.



Accessing Tape-Resident Data

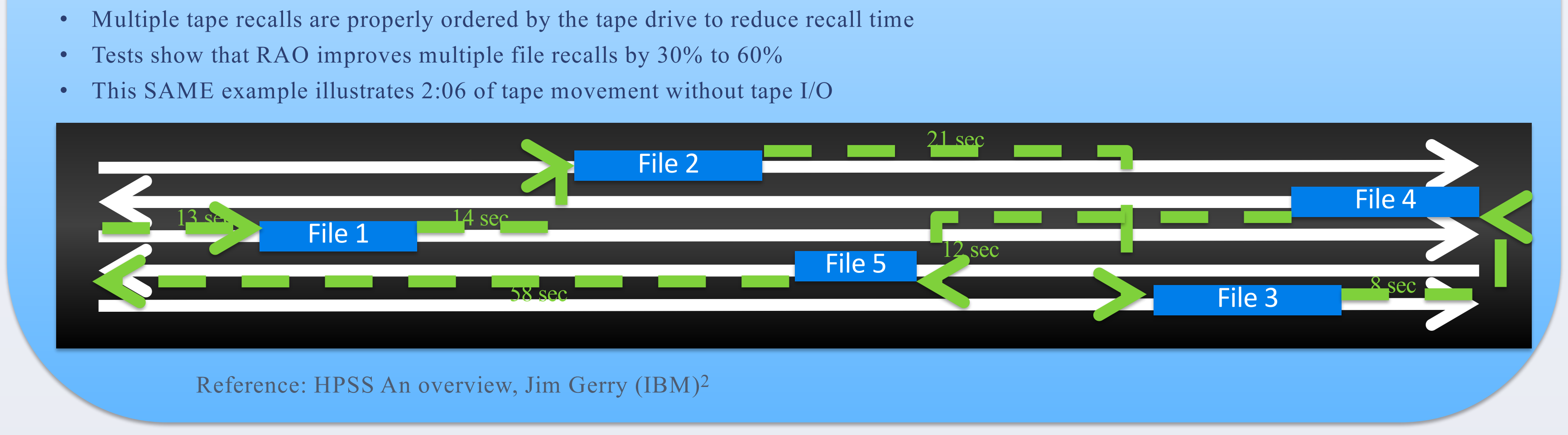
LTO Tapes drives supports offset ordered recalls (Sequentially Access)

- The serpentine nature of tape may still result in long seek times between file reads
- This example illustrates 5:16 of tape movement without tape I/O



Enterprise tape drives now support Recommended Access Ordering (RAO)

- Multiple tape recalls are properly ordered by the tape drive to reduce recall time
- Tests show that RAO improves multiple file recalls by 30% to 60%
- This SAME example illustrates 2:06 of tape movement without tape I/O

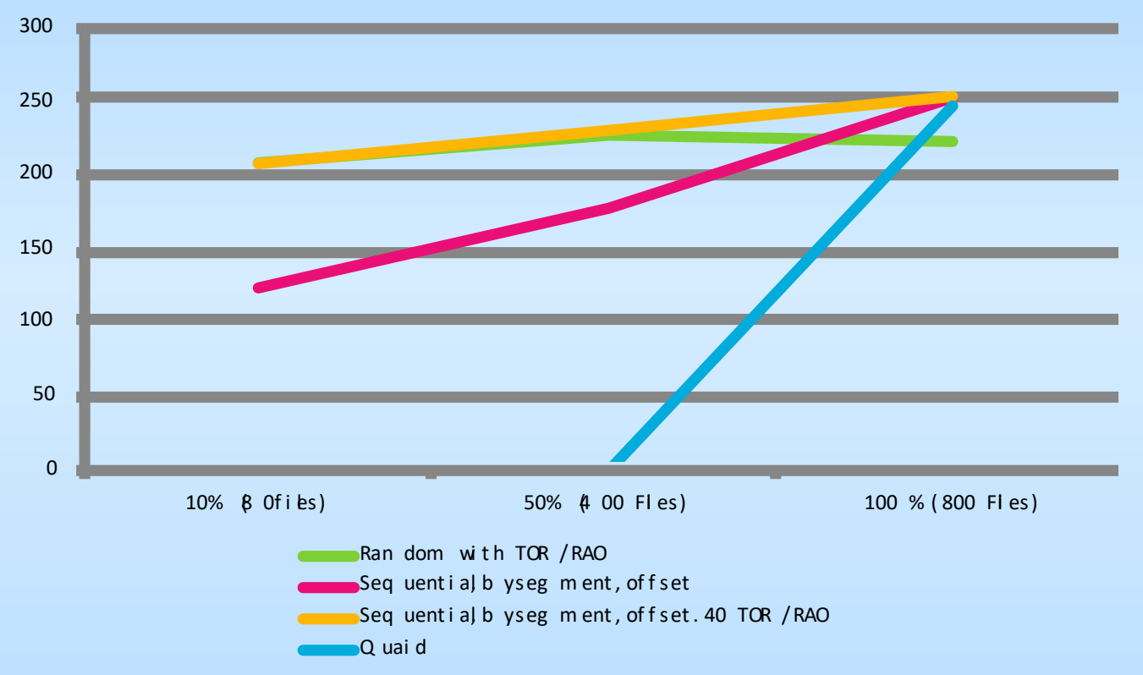


Reference: HPSS An overview, Jim Gerry (IBM)[2]

Results

Large file: 10 G files (No file aggregation), in MB/s

	10% (80 files)	50% (400 Files)	100% (800 Files)
Random submit with TOR/RAO	208.87	227.64	224.46
Sequential, by segment, offset	123.96	180.17	253.18
Sequential, by segment, offset. With 40 Threads for TOR/RAO	208.31	228.39	253.91
Quaid	N/A	N/A	247.53



Small file: 1 G files (No file aggregation), in MB/s

	10% (80 files)	50% (400 Files)	100% (800 Files)
Random submit with TOR/RAO	N/A	N/A	229.3
Sequential, by segment, offset	52.29	140.29	252.14
Sequential, by segment, offset. With 40 Threads for TOR/RAO	96.38	136.37	249.01
Quaid	110.87	107.06	184.97

Small file: 1 G files (with small file aggregation), in MB/s

	10% (80 files)	50% (400 Files)	100% (800 Files)
Random submit with TOR/RAO	N/A	N/A	20.45
Sequential, by segment, offset	50.47	139.73	255.0
Sequential, by segment, offset. With 40 Threads for TOR/RAO	19.21	105.92	224.49
Quaid	N/A	21.6	74.86

Conclusion

For none-aggregated files, best practice is to use sequential submission via ERADAT with HPSS TOR/RAO turn on. In this example, we used 40 Threads submission.

For files in "small files aggregated" segment, RAO is not recommended.

Traditional sequential access based on segment position and offset is the best practice for files within "small file aggregation" block.

RAO feature did not do better than sequential staging when recalling 100% files back (Full restore).

RAO is designed to reduced tape seeking time, we also proves that seeking time is the main factor that impacts the staging performance.

RAO is not supported in LTO environment, sequential accessing by segment position and offset is the best solution for LTO drives.

To avoid unnecessary seeking time, we should try to write data with same dataset, use file family. Do not mix unrelated data within the same tape.

References

1. StorageTek T10000D Tape Drive, ORACLE DATA SHEET
http://www.oracle.com/us/products/servers-storage/storage/tape-storage/t10000d-ds-1991052.pdf

2. HPSS An overview, Jim Gerry (IBM)
http://konferenz-nz.dlr.de/pages/storage2016/present/2.%20Konferenztag/06_17_06_16_ibm.pdf

BROOKHAVEN NATIONAL LABORATORY | Scientific Data and Computing Center

U.S. DEPARTMENT OF ENERGY | Office of Science