Storage Architecture of IU’s Media Digitization and Preservation Initiative

Brian Wheeler
Senior System Engineer
Indiana University Libraries

MDPI:
Media Digitization and Preservation Initiative

- Goal: “To digitize, preserve and make universally available by IU’s Bicentennial—subject to copyright or other legal restrictions—all of the time-based media objects on all campuses of IU judged important by experts.“

- 280,000+ audio and video items

- ~7PB over 4 years

- 9TB per day peak

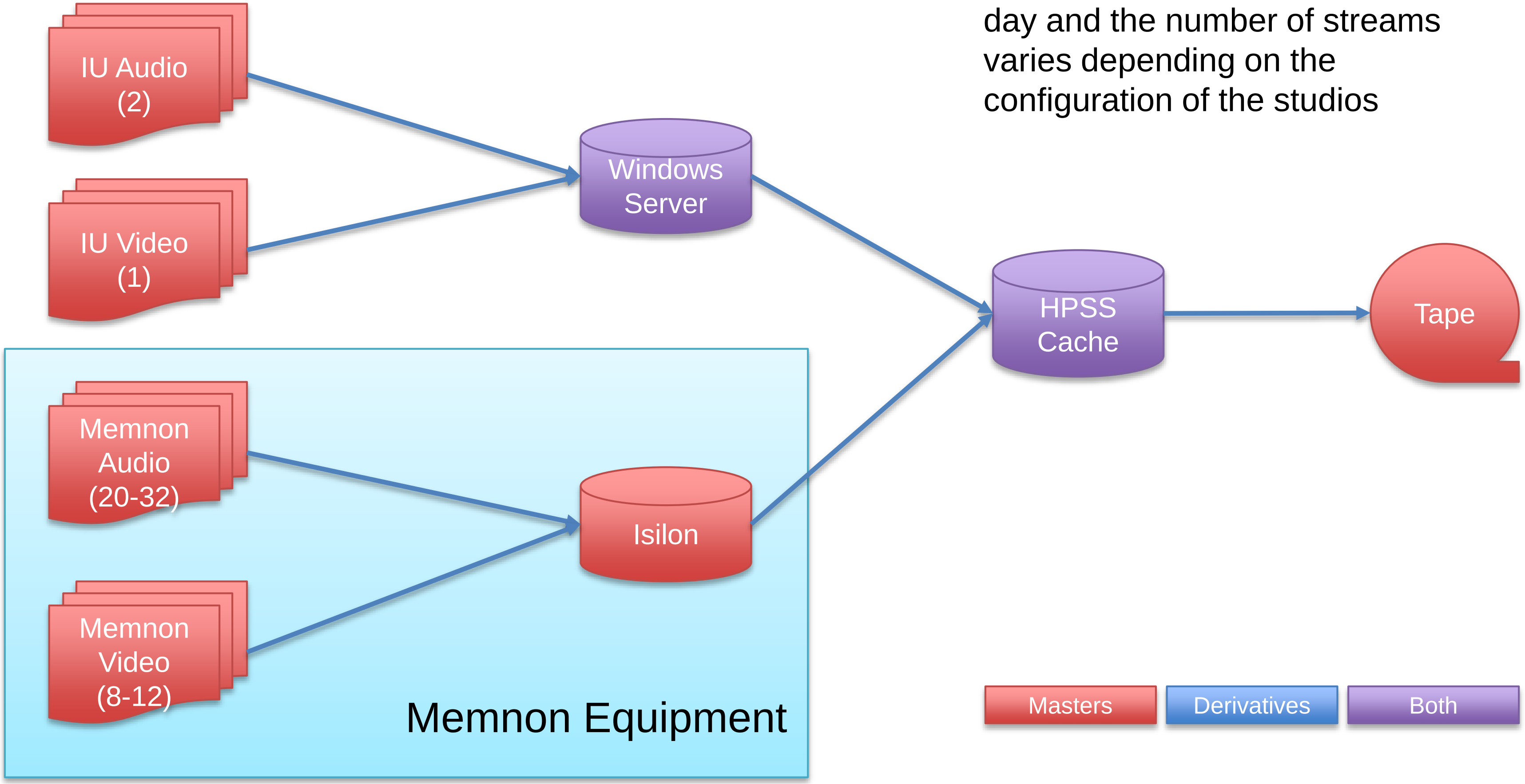
- https://mdpi.iu.edu/

Background

- Utilizes IU’s HPSS-based Scholarly Data Archive service for storage.
  - HPSS is tape-based with large disk caches.
  - Data is mirrored between the Bloomington and Indianapolis campuses
  - IBM TS3500 Library, TS1150 drives, IBM 3592 JD tapes (10T native)
- Two digitization sources:
  - For bulk digitization, a partnership with Memnon Archiving Services, a Sony company based in Belgium. Memnon has a digitization facility on campus.
  - An IU facility for delicate or damaged material. Also processes formats not suitable for factory-style digitization: wax cylinders, wire recordings, etc
- Initial production batches started 6/2015.

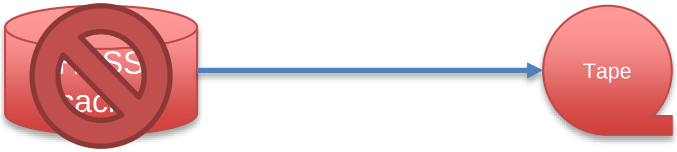
Initial Ingest

- IU runs one shift per day
- Memnon runs two shifts per day and the number of streams varies depending on the configuration of the studios

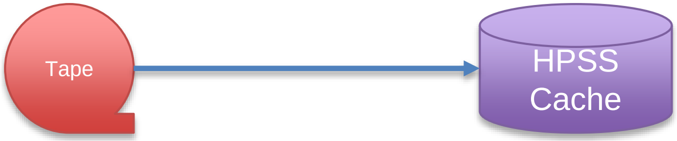


Tape Copy Validation / Transfer to Xcoder

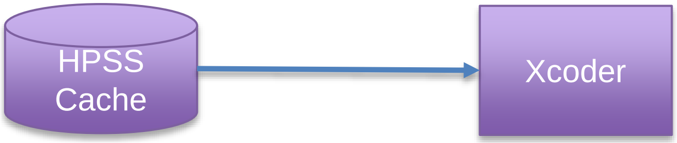
- Disk cache copy is purged after data is written to masters tape pool



- Data is re-read from tapes, after incoming data stream is idle



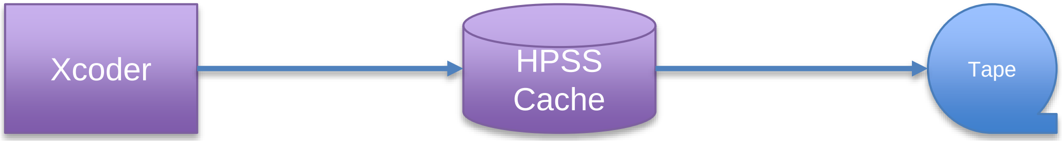
- Checksums validated on transcoder during download



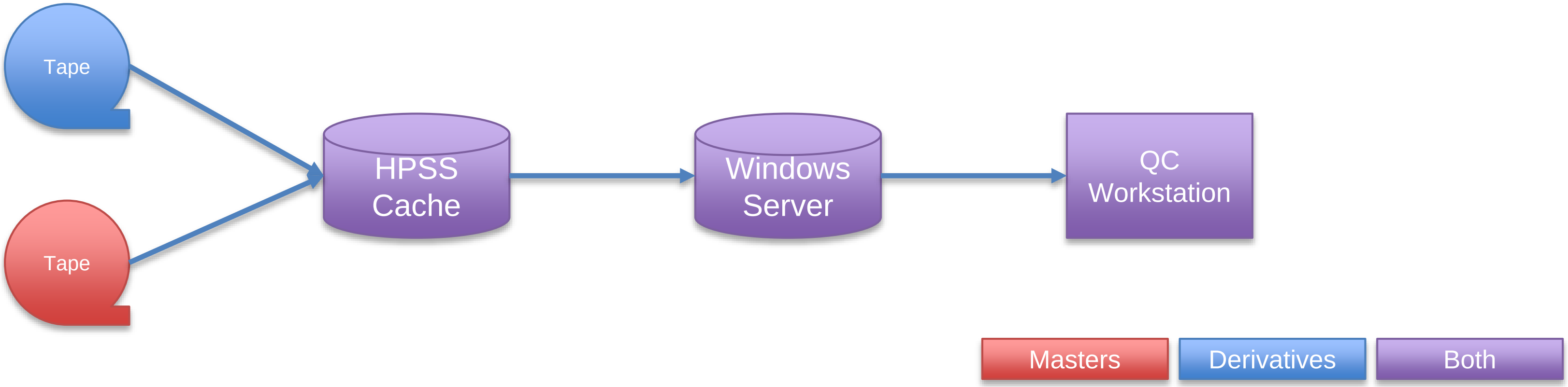
🟥 Masters 🟦 Derivatives 🟪 Both

Derivative Creation / Manual QC

- Derivatives created and transferred to a derivatives tape pool (to avoid contention with incoming masters)

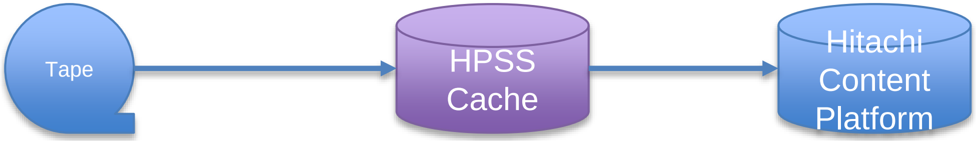


- Objects may be retrieved for manual QC by staff

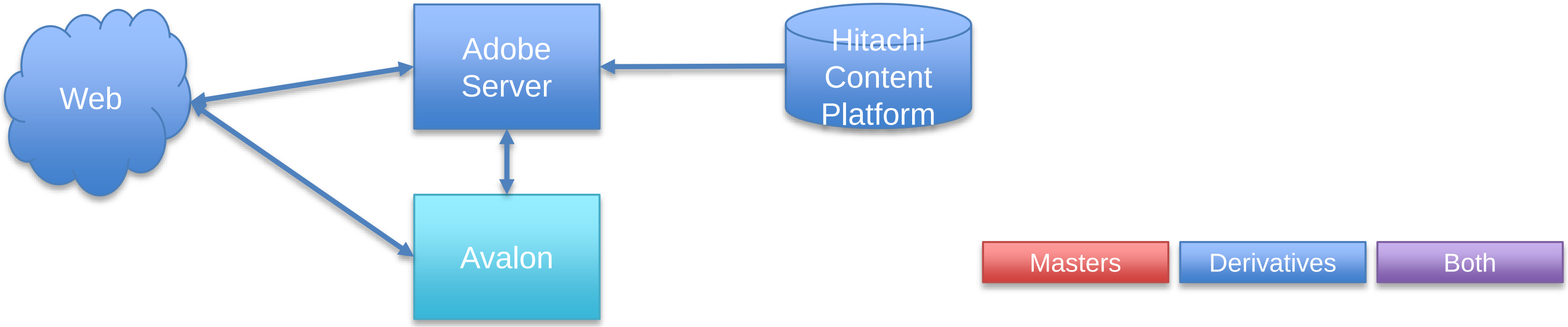


Access Distribution

- Derivatives retrieve from tape and sent to Hitachi Content Platform



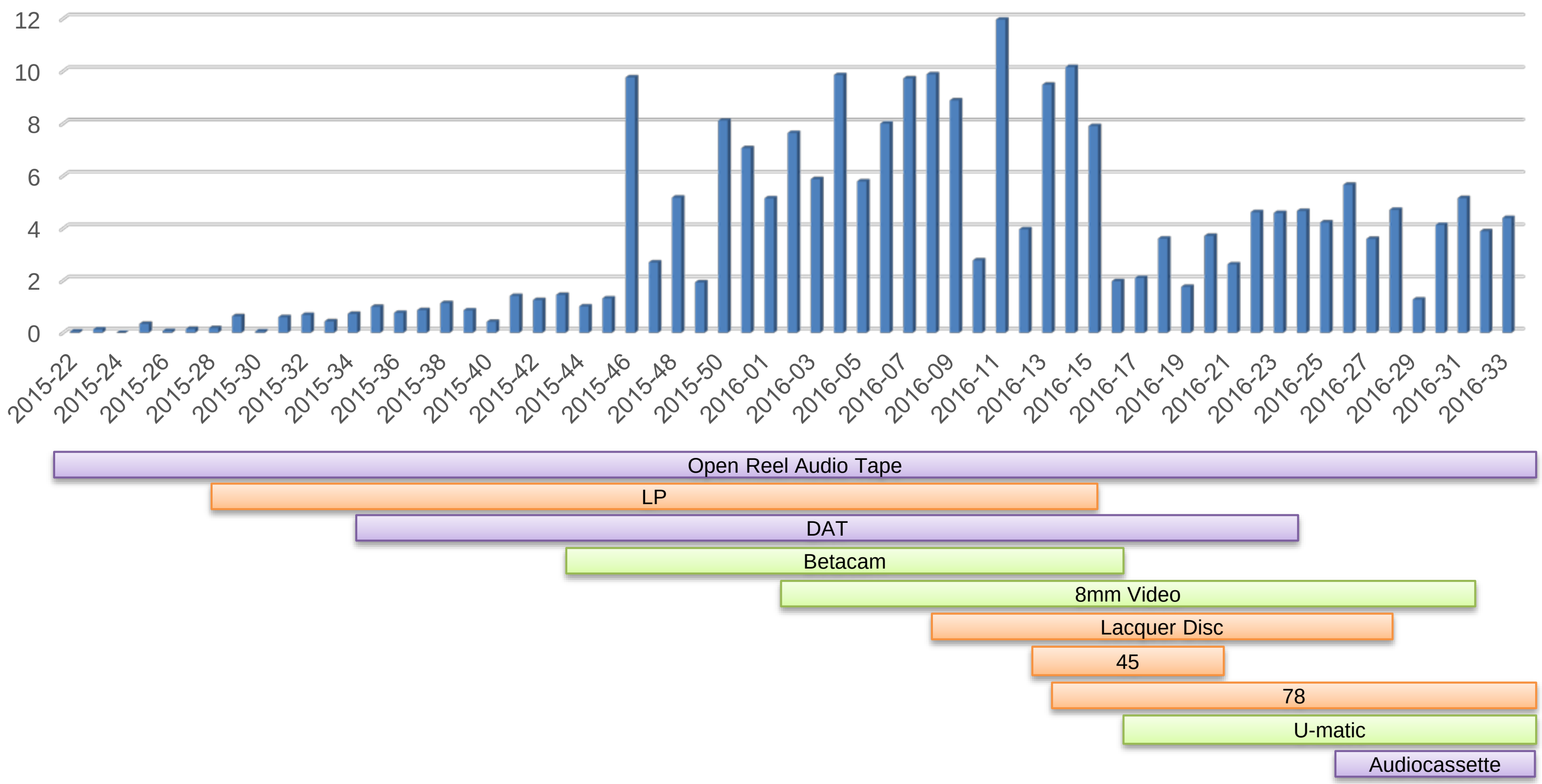
- Users access the content via Avalon Media System and Adobe Media Server



Actual Throughput

- Estimated 9.8T/day peak
- 20 days with > 12T
- 4 days at 20T

Average Per Day (TB)



Open Reel Audio Tape

LP

DAT

Betacam

8mm Video

Lacquer Disc

45

78

U-matic

Audiocassette

Future Directions

- Investigating Film Scanning
  - Potentially 20TB/day in addition to the current throughput
- Digital preservation functionality based on HydraDAM2 Project
  - A collaboration between Indiana University Libraries and WGBH Educational Foundation Media Library and Archives supported by a grant from NEH
  - An AV digital preservation repository implemented as a Hydra head.
  - IU's implementation will use HPSS as back end storage
  - Will provide easy access to master files, fixity checking, and migration.
- Exploring options for out-of-region storage
  - IU is a member of the Digital Preservation Network (DPN) and Academic Preservation Trust (APTrust). DPN was designed for content in TB range vs PB

Questions

Links/Resources:

- Media Digitization and Preservation Initiative – https://mdpi.iu.edu
- Avalon Media System – https://avalonmediasystem.org
- HydraDAM2 – https://wiki.dlib.indiana.edu/display/HD2/HydraDAM2