Archive Storage Infrastructure

At the Library of Congress

September 2017

LIBRARY OF CONGRESS

Packard Campus for Audio Visual Conservation
http://www.loc.gov/avconservation/packard/

The Packard Campus

Mission

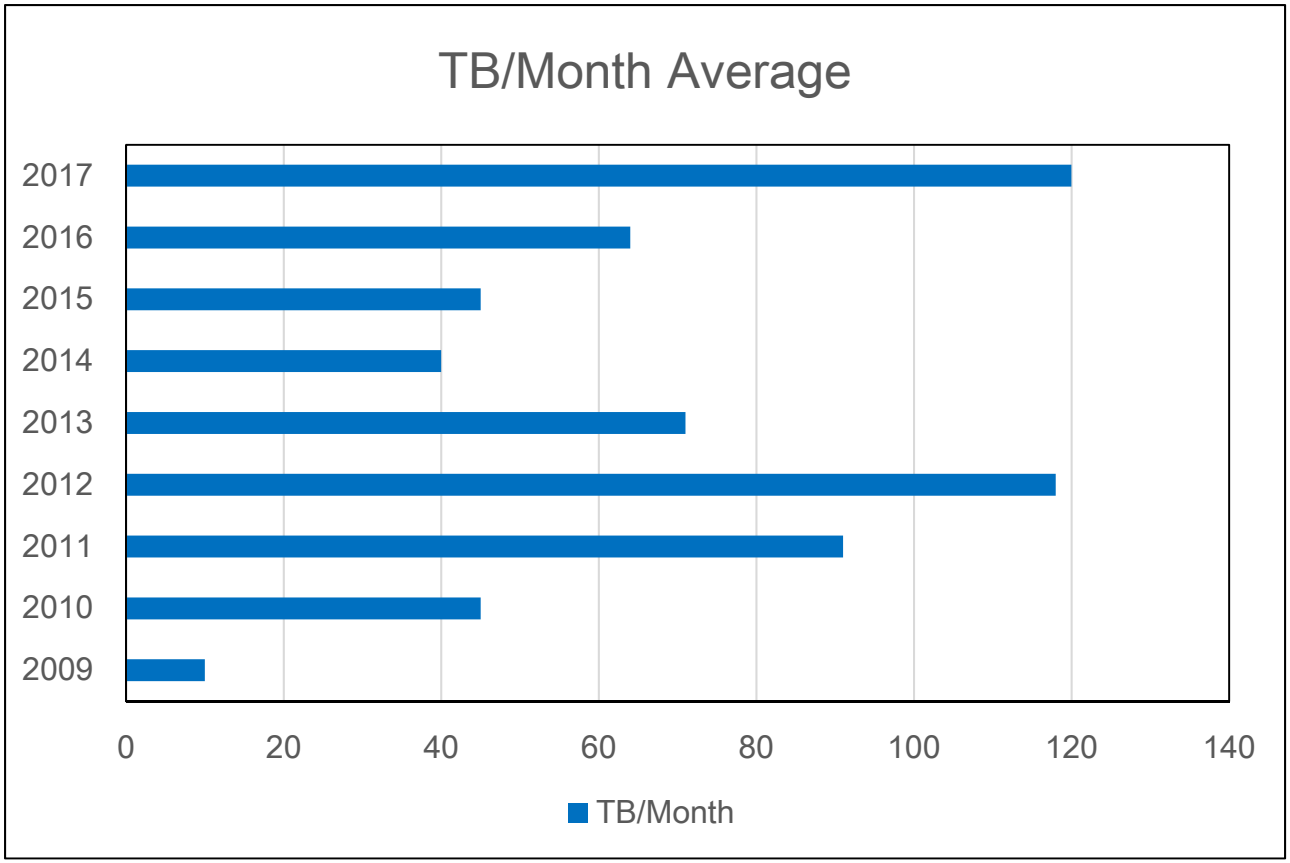
- The National Audiovisual Conservation Center develops, preserves and provides broad access to a comprehensive and valued collection of the world's audiovisual heritage for the benefit of Congress and the nation's citizens.

Goals

- **Collect, Preserve, Provide Access to Knowledge**
- The NAVCC consolidated collections stored in four states and the District of Columbia.
- The facility boasts more than 1.5 million film and video items and 3.5 million sound recordings, providing endless opportunities to peruse the sights and sounds of American creativity.

Formats, Formats and more Formats….

Historic Ingestion Values



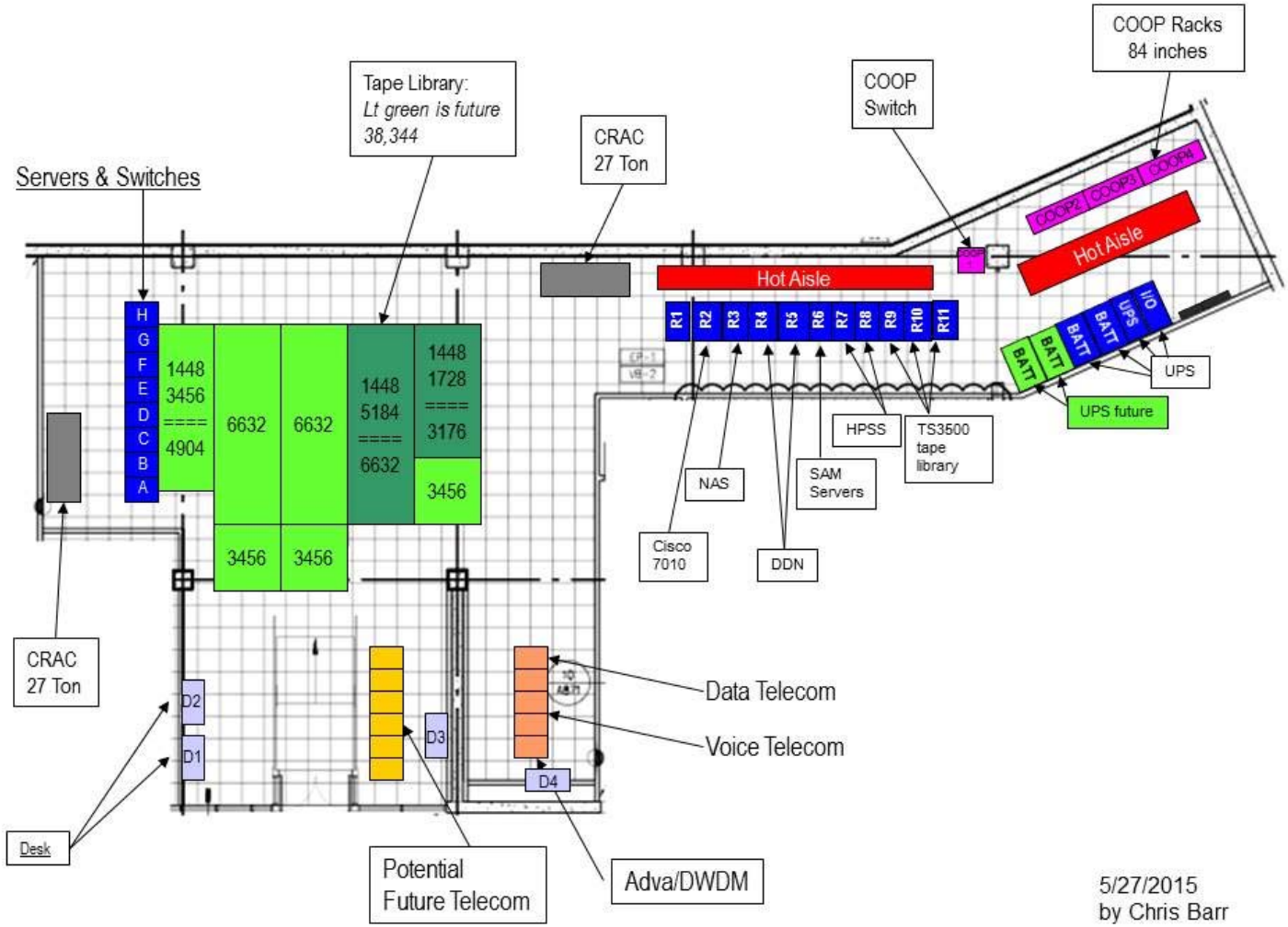
Current: 8.9 PB and 2.1 Million files replicated in 2 locations

Peak in September 2015: 235 TB / month

Challenges

- **Planning Future Storage Needs:**
  - Projected: 300 TB / week or 1.3 PB / month – at least 5 years off
  - Variables inconsistent – personnel, technology, budget
- **Tape Roadmap**
  - *Previously* counted on doubling of tape density & computing power
  - Assumed that this would keep us in our current 3000 square feet computer room with two 20 ton CRACs and 300 KVA of power (Using 42 KVA now)
  - Recent Tape Vendor Announcement – Future unclear
- **Doveryai, No Proveryai (Trust but Verify)**
  - Reduce the likelihood of content loss with inevitable data loss
  - Catch and correct all marginal errors and failures as soon as possible
  - Verify all the content at a regular interval
    - Samfsbackup
    - Samfsck
    - File System Monitoring (Size, # of Files, Damaged Flags, etc)
    - TPVerify

Packard Campus – Data Center



Current Initiatives (OCIO)

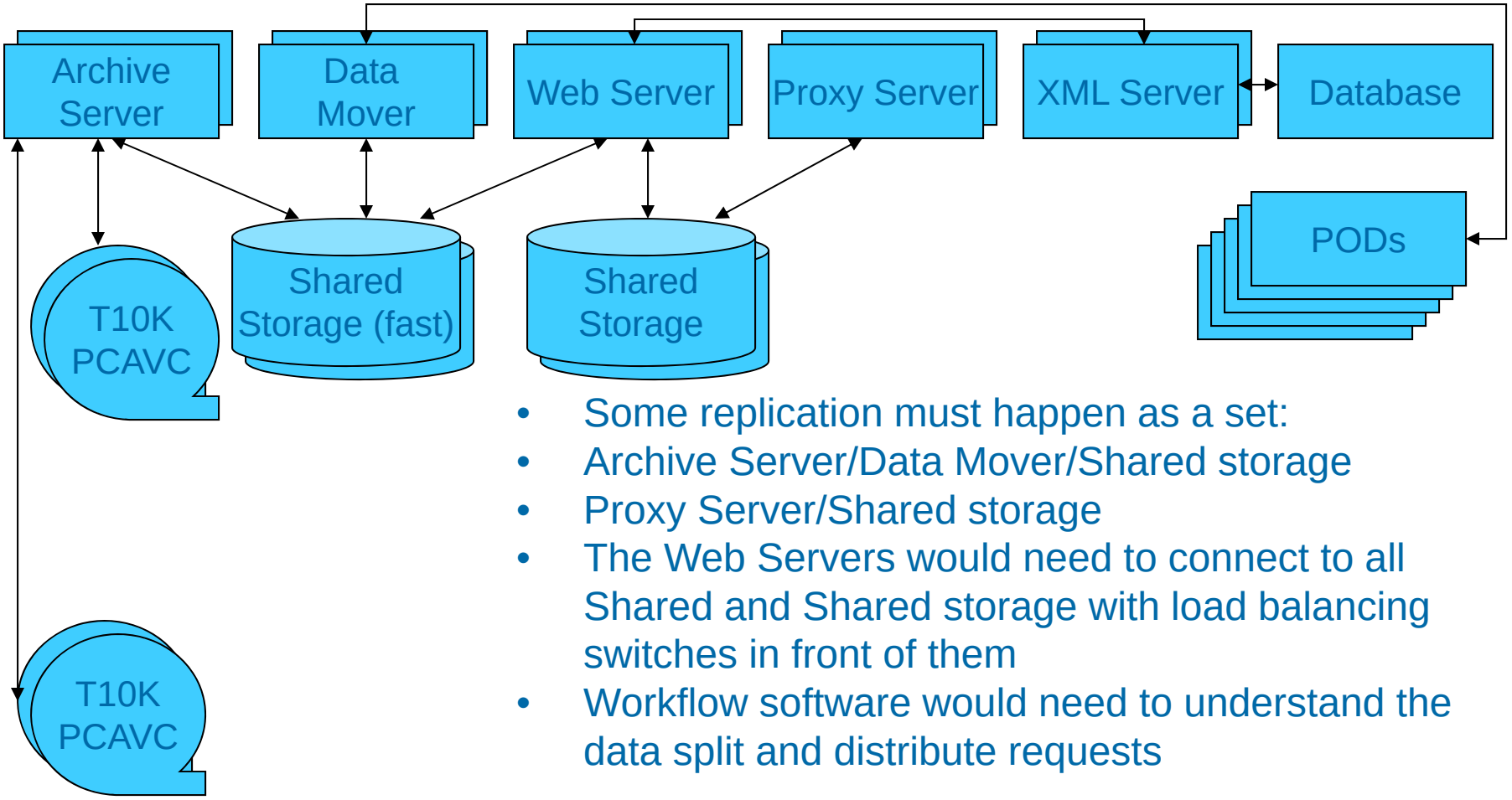
- MIGRATIONS: Completed second migration of 6 PB of content from T10KC to T10KD over a 9 month time frame. No errors this time!
  - Migration as initiated for a tape reformat to T10KD to increase capacity
- HARDWARE REFRESH – Update of the Infrastructure hardware completed in 2017 to modernize the storage and server infrastructure supporting the MBRS Archive process.
- OHSM SIMPLIFICATION- Configuration changes to simplify architecture
  - Removal of local storage cache before the remote copy
  - OSHM move to virtual platform
  - This eliminated complexity in archival asset identification and retrieval
  - Reduction in costs (hardware, power, support)

Current Initiatives (MBRS)

- PCWA DEVELOPMENT - PCWA is the workflow application at the center of our Archival processes at the NAVCC. MAVIS DATABASE Upgrade & Oracle to 12C
- NATIONAL SCREENING ROOM– Web portal for the online publication of rights unlimited Motion Picture Archival content
- FILE TRANSFER SYSTEMS: The NAVCC is leading the way with file transfer systems implementation for sending and receiving Born Digital content within and outside of the Library
  - Media Shuttle – Mostly used for submission and delivery of content between the Library and patrons.
  - Aspera – Mostly used for digital submissions between other Organizations (Sony, NBC, etc)
- FILM PROCESSING - 4K Film scanning has kicked off. This is anticipated to increase content in the archive dramatically.
- LIVE CAPTURE– NAVCC has completed initial testing of Live Capture for Born Digital Content.

Functional Architecture - Scaling

Archive Storage Infrastructure



- Some replication must happen as a set:
- Archive Server/Data Mover/Shared storage
- Proxy Server/Shared storage
- The Web Servers would need to connect to all Shared and Shared storage with load balancing switches in front of them
- Workflow software would need to understand the data split and distribute requests