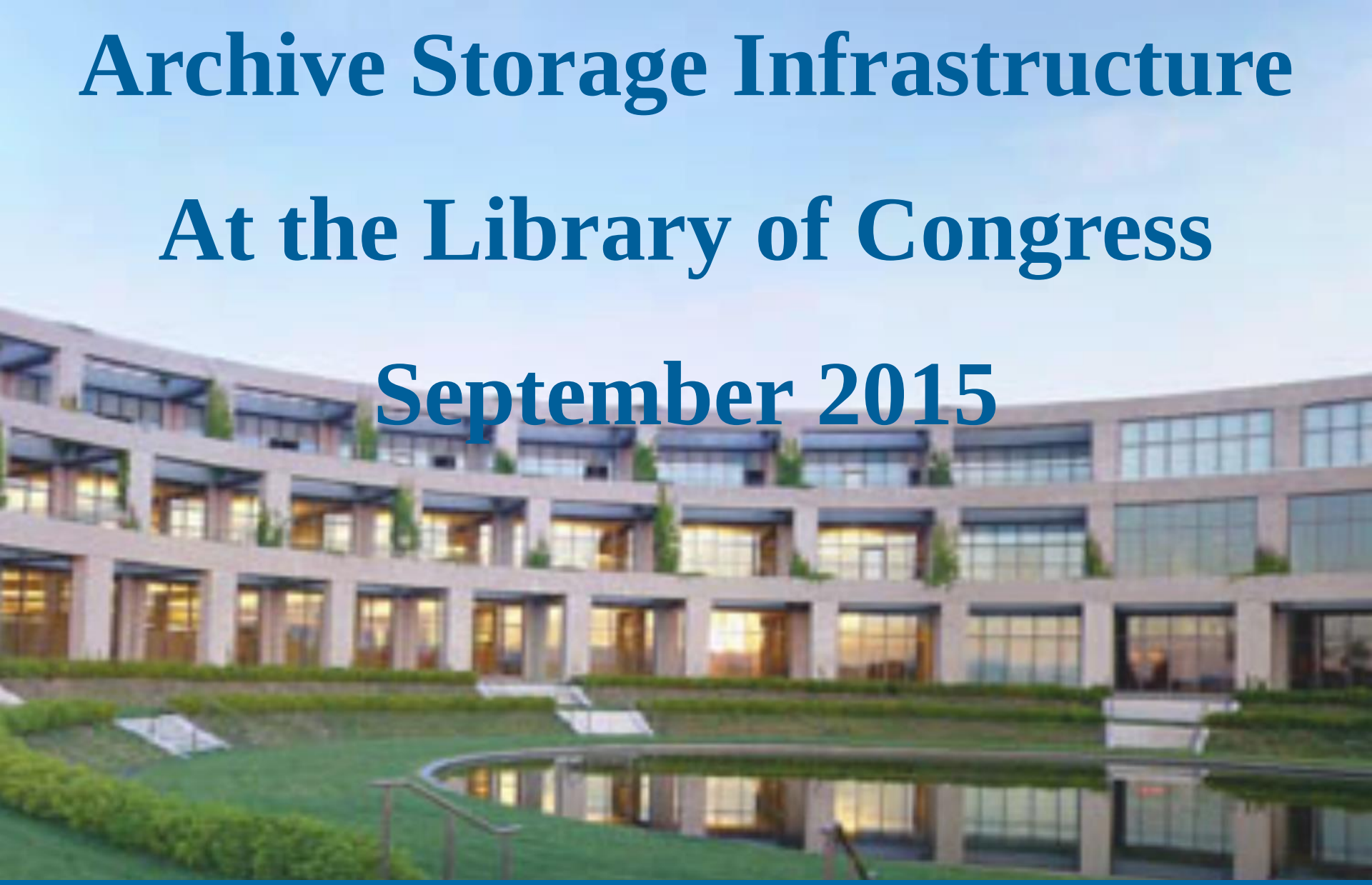
Archive Storage Infrastructure

At the Library of Congress

September 2015



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 National Audiovisual Conservation Center (NAVCC) of the Library of Congress will be the first centralized facility in America especially planned and designed for the acquisition, cataloging, storage and preservation of the nation's collection of moving images and recorded sounds. This collaborative initiative is the result of a unique partnership between the Packard Humanities Institute, the United States Congress, the Library of Congress and the Architect of the Capitol.
- The NAVCC consolidated collections stored in four states and the District of Columbia. The facility boasts more than 1 million film and video items and 3 million sound recordings, providing endless opportunities to peruse the sights and sounds of American creativity.

The Packard Campus – Many Formats

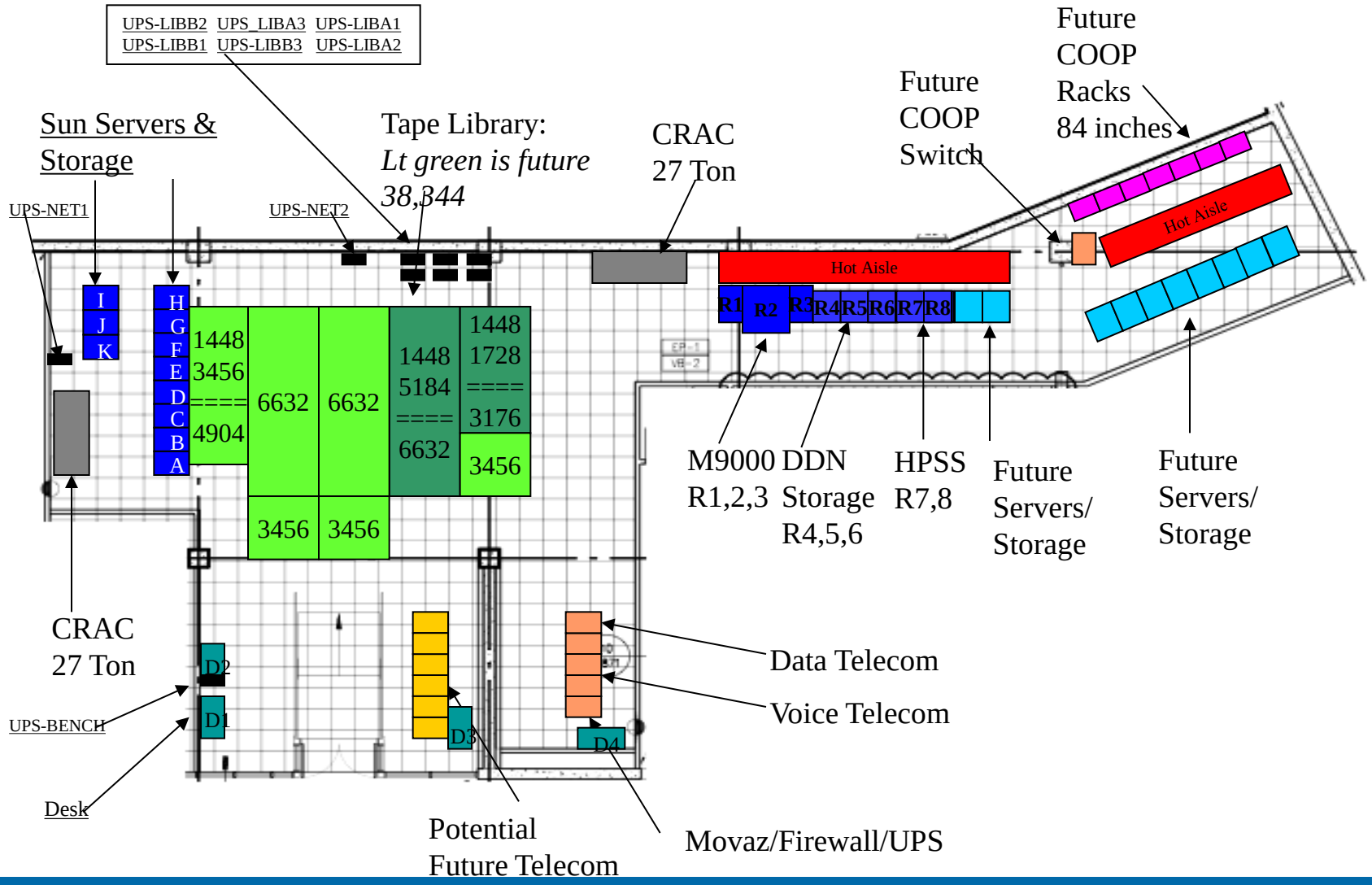
The Packard Campus – Past, Present and Future

- Growth since production
  - February 2009: 10 TB / month
  - February 2010: 45 TB / month
  - February 2011: 91 TB / month
  - February 2012: 118 TB / month
  - *Peak in September 2014: 169 TB / month*
  - February 2013: 71 TB / month
  - February 2014: 40 TB / month
  - February 2015: 45 TB / month
- Current: 6.1 PB and 1.4 Million files replicated in 2 locations. 3 PB and 200 Million files for Newspapers, internet archive, prints and photographs
- 53 Points of Digitization (PODs):
  - 34 Solo (16 in robotic cabinets), 9 Pyramix, 10 Linux(OpenCube,etc)
  - Daily each POD generates: 2GB-150GB for audio and 50GB-1,200GB for video
  - Additional PODs coming in the future include 2K and 4K scan for film, digital submission for Copyright and other (Live capture-264 DVRs, PBS, NBC Universal, Vanderbilt TV News, SCOLA, etc)

The Challenge

- **Projected: 300 TB / week or 1.3 PB / month – at least 5 years off**
- **Counting on doubling of tape density and computing power to keep us in our current 3000 square feet computer room with two 20 ton CRACs and 300 KVA of power**
  - Using 45 KVA

The Packard Campus – Physical Space



Doveryai, No Proveryai

Trust but Verify

Content versus data

- We want to reduce the likelihood of losing content while still recognizing that data loss is inevitable.
- Catch and correct all marginal errors as soon as possible
- Catch and correct all failures as soon as possible
- Some of the regular verification processes that we run:
  - Samfsbackup (meta data backup) 5X/day
  - Verify samfsbackup size and frequency. Send an email if missing.
  - Fix damaged files. Occasionally a file will be marked damaged because it cannot be retrieved from tape. Usually because a tape was stuck in a drive/robot/pass thru port. Find these everyday and attempt to stage. If we can't, then send an email. Send an email when we find damaged files so we know issues are occurring and being corrected
  - Stats: Watch the # and size of files waiting to archive. Warn when the # of files or size of files exceeds thresholds. Usually an indication of some marginal error condition. Fix before file system fills up or we fail to deliver a file for customers.
  - Samfsck: Run this daily with filesystem mounted. Warns when there are marginal conditions with file system before they are catastrophic.
  - # of tapes/TB available: Know when we are running low so we can correct before a failure
  - Tpverify: Verify all tapes with data every 6 months. Verifying all blocks of data on tape with CRC.

The Packard Campus – Status

Current initiatives

- Completed migration of 3.5 PB of content from T10KB to T10KC over a 5 month time frame. Found SHA1 mismatch for 27 files. No content lost
  - One was due to human error. Found through email threads
  - The other 26 appear to be due to errors on the disk between the time the data was written and when it was written to both tapes. Both tape copies' SHA1 values match
    - RAID rebuild?
    - Errors in RAID array?
  - Led us to design a process where we verify the SHA1 digest of the files on tape within 1 week to catch these errors in the future
- Oracle has a roadmap that includes tape to tape migration and storing our SHA1 values in extended file attributes. This will change our verification processes
- First iteration of Archive Integrity Metric (AIM) to improve data informed design
- Piloting a partnership with a local University to provide greater access over Internet II
- Collecting requirements for a storage abstraction layer to simplify customer submission / access and technology maintenance / refresh

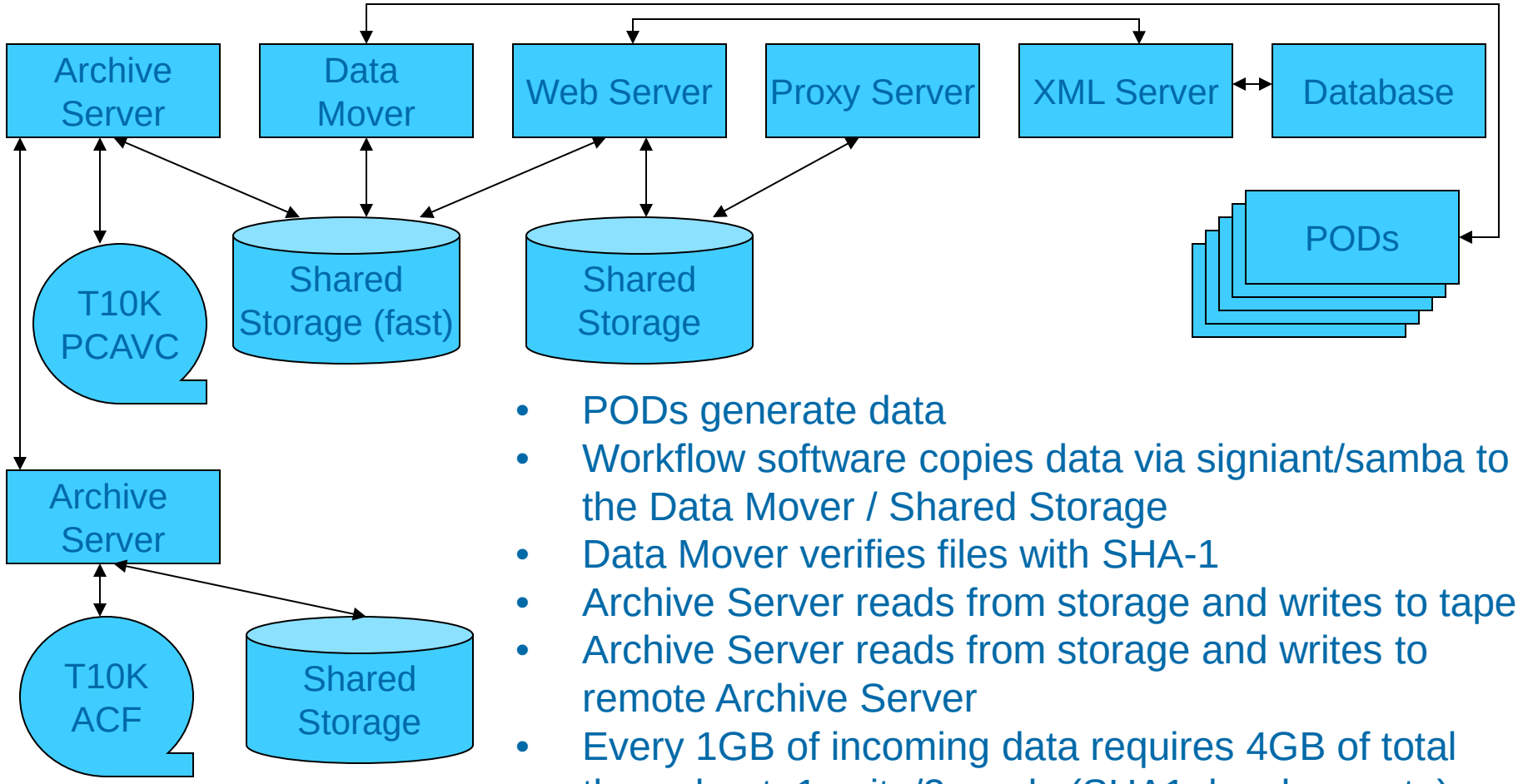
The Packard Campus – Status

Current initiatives

- Orderless ingest maturing and has the potential to fully utilize current configuration
  - History Makers last year ingested 200 TB
  - American Archive this year and next will ingest 1.2+ PB
  - How many people are interested in better understanding MBRS' custom workflow software?
- Less than 20 ingest streams per day last year to almost 30 ingest streams today
  - How does this change our architecture?
    - Digest slow due to small block I/O: tested by running dd … bs=65536K | digest and improved performance. Requested Oracle to improve their digest command. Turned around in a few months
    - File transfer/copy slow: Increased block size at client (win7) and improved throughput. Still experiencing 1-5% failure rates in files every night. New perspective helps.
- NAS taking more of historic SAN load
  - ZS3 with 150 TB and eight 10 Gbe interfaces for high bandwidth throughput
  - Existing 7320 with AD and four 10 Gbe providing easy to deploy and manage storage for smaller (0.1-20 TB) projects

Functional Architecture – Data Movement

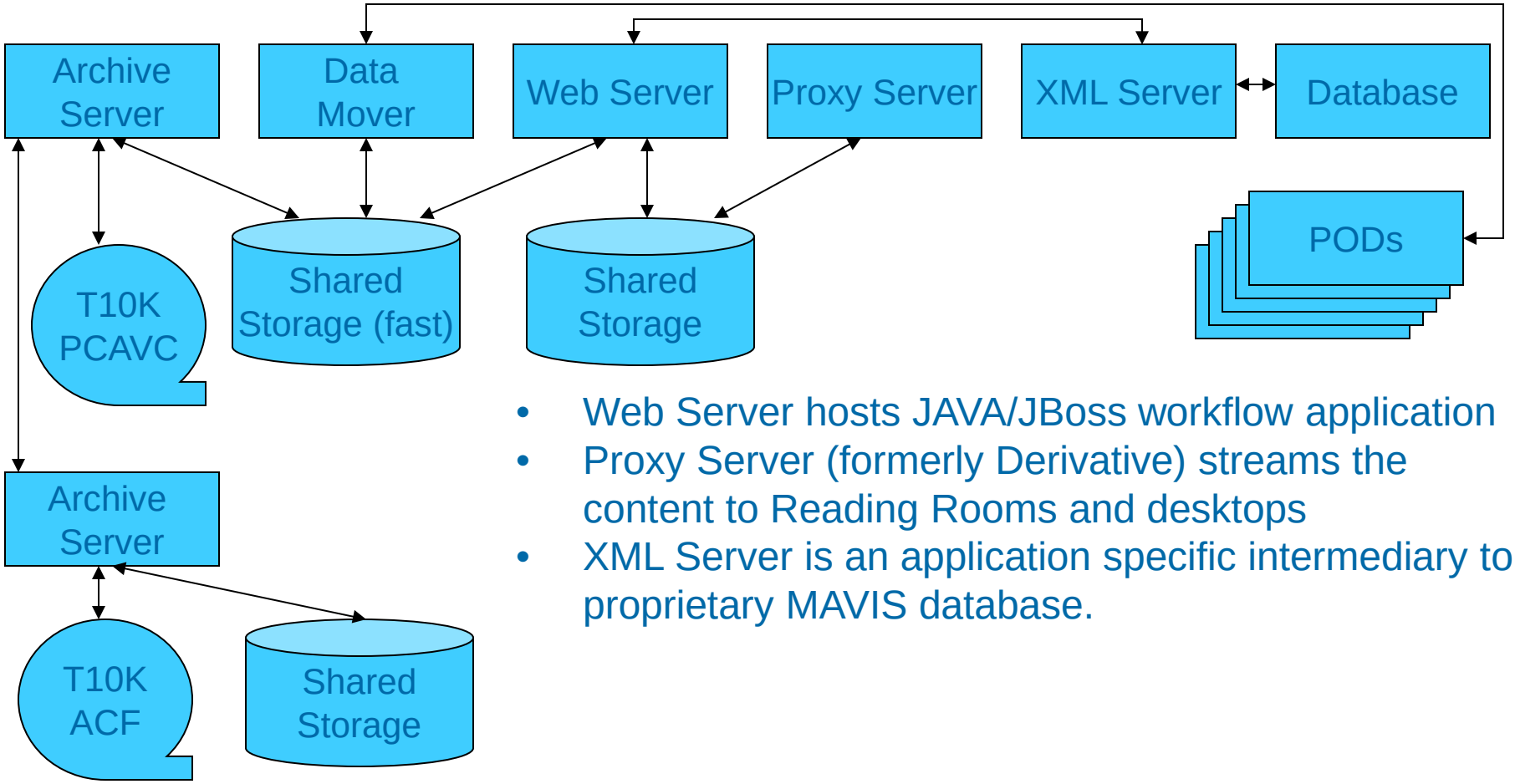
Archive Storage Infrastructure



- PODs generate data
- Workflow software copies data via signiant/samba to the Data Mover / Shared Storage
- Data Mover verifies files with SHA-1
- Archive Server reads from storage and writes to tape
- Archive Server reads from storage and writes to remote Archive Server
- Every 1GB of incoming data requires 4GB of total throughput: 1 write/3 reads (SHA1, local, remote)

Functional Architecture – User Interface

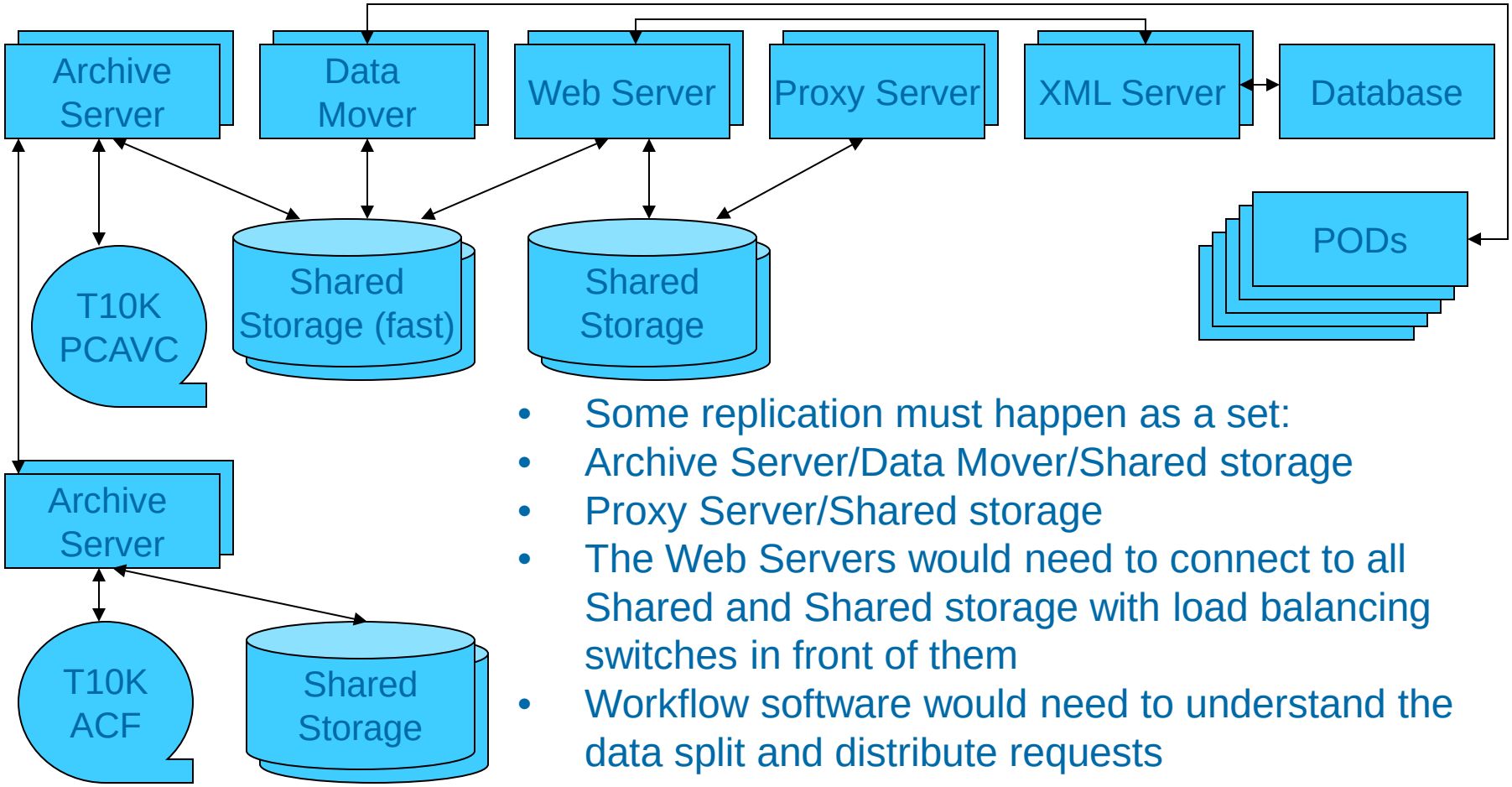
Archive Storage Infrastructure



- Web Server hosts JAVA/JBoss workflow application
- Proxy Server (formerly Derivative) streams the content to Reading Rooms and desktops
- XML Server is an application specific intermediary to proprietary MAVIS database.

Functional Architecture - Scaling

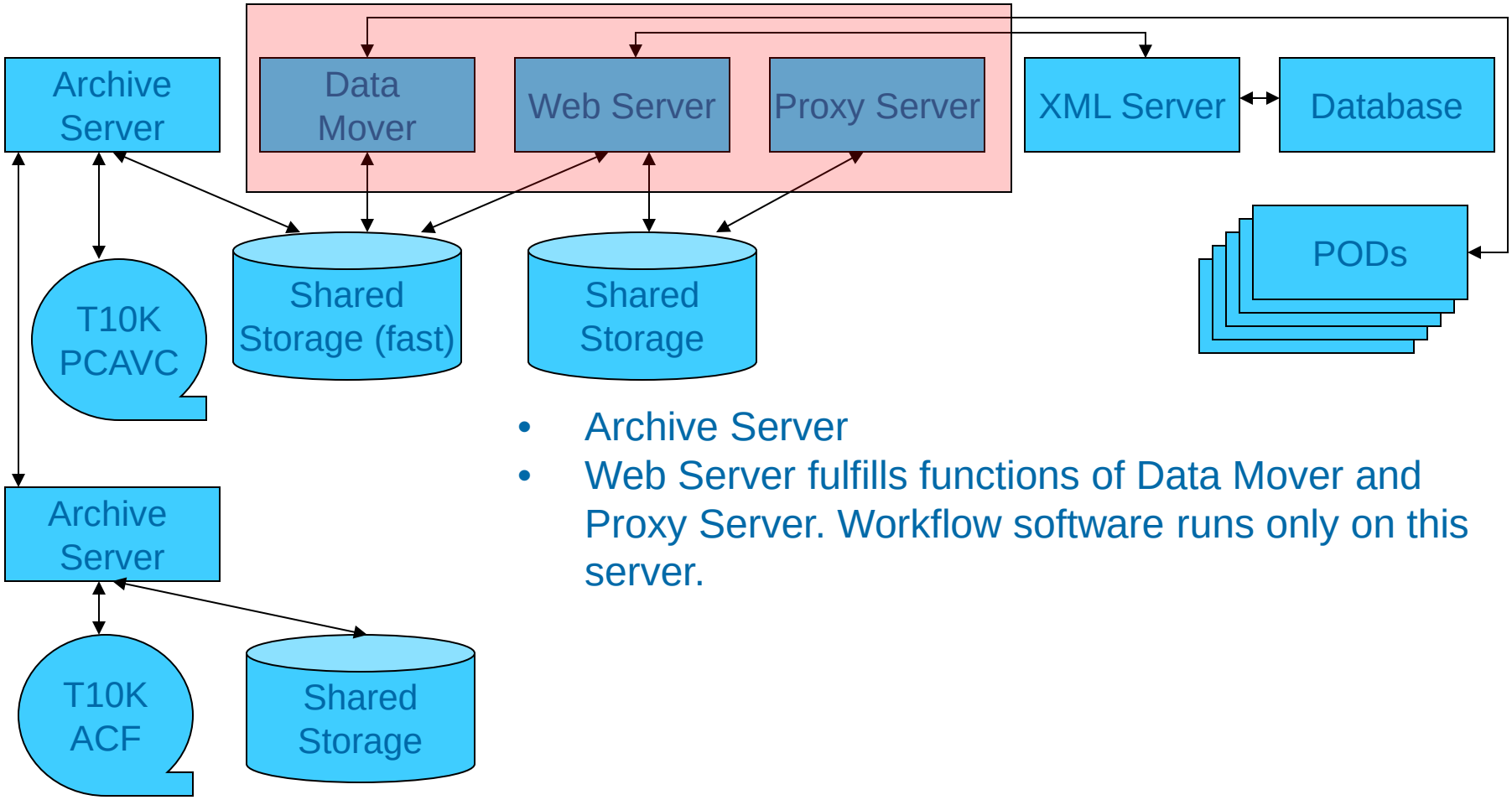
Archive Storage Infrastructure



- 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

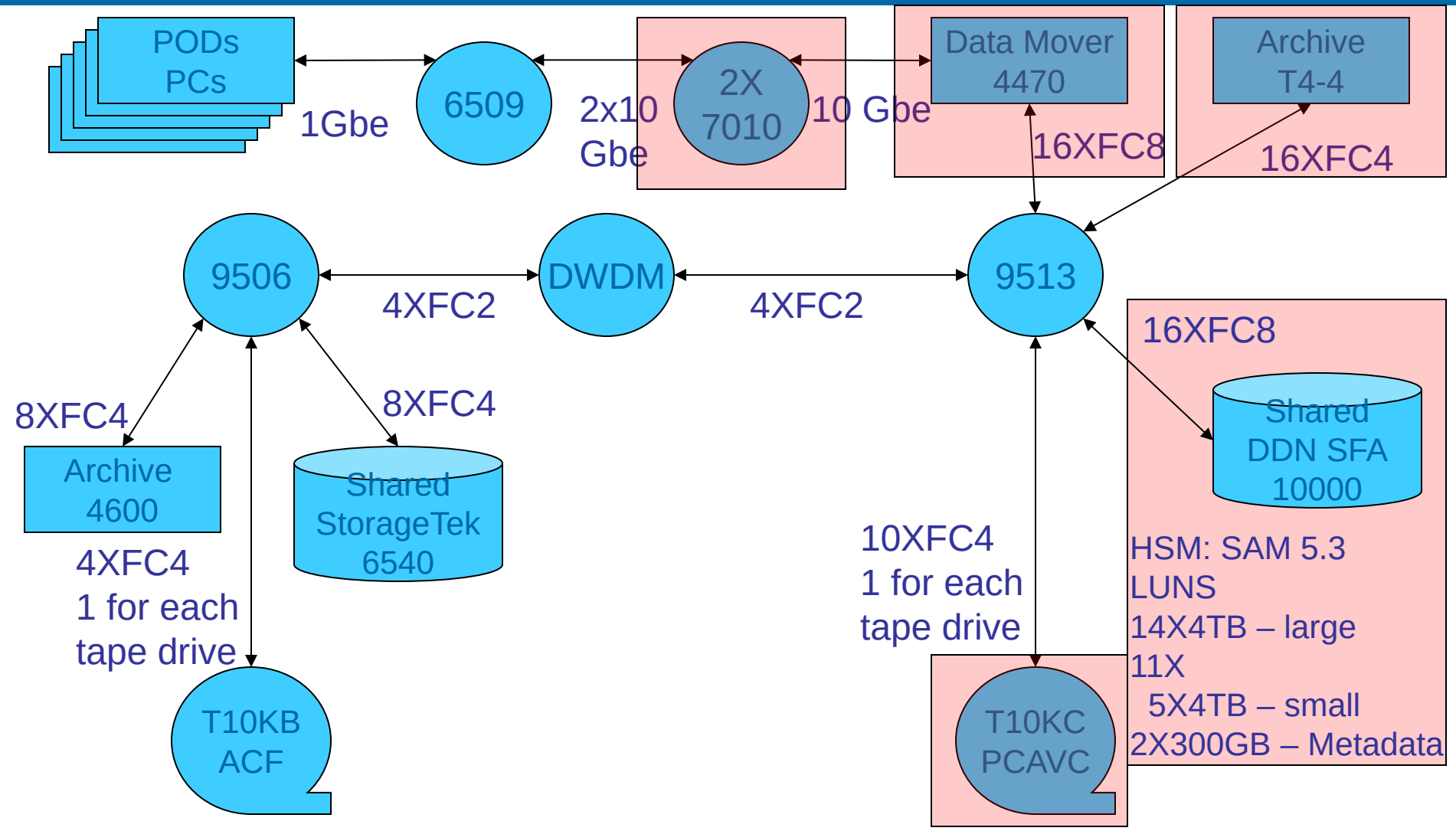
Functional Architecture – Current

Archive Storage Infrastructure

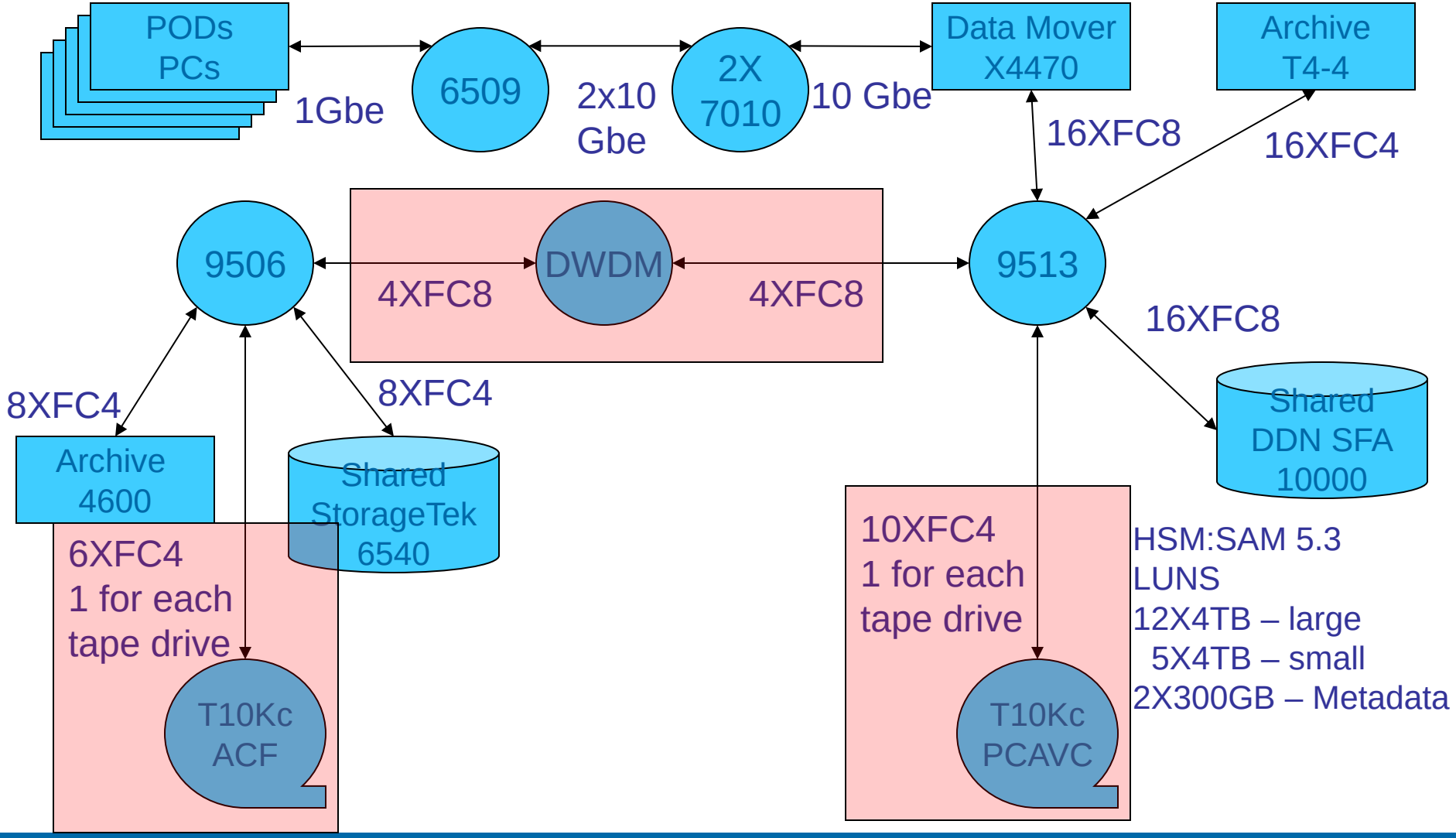


- Archive Server
- Web Server fulfills functions of Data Mover and Proxy Server. Workflow software runs only on this server.

Physical Implementation V2: 6.5 GB/s throughput 2013



Physical Implementation V2+: 6.5 GB/s throughput 2013



Physical Implementation V2.2: 6.5 GB/s Future