Transitioning Legacy Digital Libraries

Merritt E. Jones

The MITRE Corp.
1829 Dolly Madison Blvd
McLean VA 22102-3481
Phone: +1-703-883-5471, FAX: +1-703-883-3320
merritt@mitre.org

THIC meeting at NSWC Carderock
Oct. 14, 1998

MITRE

BACKGROUND

- There are existing digital libraries that have in excess of 100,000 pieces of media.
- Digital libraries in the hundreds of terabytes to petabytes are routinely discussed and are being procured and developed.
- There are at least two programs in the 6+ petabytes of storage range.

There are major issues associated with transitioning and technology insertion for libraries of this size.

This presentation looks (somewhat loosely) at these issues and may present more questions than answers.

A Few References

- Some programs and facilities relevant to this topic include
  - The National Center for Atmospheric Research
  - Lawrence Livermore National Lab
  - NASA Earth Observing System
  - National Imagery and Mapping Agency Libraries
- For the purposes of this presentation, we will mostly consider the experiences of NCAR and LLNL.

NOTE: For this presentation, *transition*, *migration* and *technology insertion* have essentially the same meaning.

National Center for Atmospheric Research

- Some facts
  - 153 terabytes of data
  - Data as far back as the 1970's and kept "forever."
  - 5.1 million files
  - 6.0 terabytes/month net growth
  - 32 terabytes/month served
  - 166,200 tape cartridges
    - 8,600 of them are in robots

The size has doubled and the data rate has quadrupled over the last 2 years.

National Center for Atmospheric Research (cont.)

- Some observations about the NCAR approach
  - There have been several migrations, mostly to newer versions of same technology.
  - It takes as long (actually longer) to migrate data as it does to write it.
  - NCAR uses an approach, “data ooze,” which migrates files to new technology as they are rewritten.
  - Files not migrated by the owner are migrated by a background utility.
  - Tapes are migrated in the background by tape sequence number (basically the oldest first).
  - The current migration to a new and different tape technology began about 3 months ago. A few hundred tapes have been migrated.

Lawrence Livermore National Lab

- Some facts
  - About 48 terabytes of data
  - Data goes back more than a decade and is kept “forever.”
  - Current growth is 1 terabyte/month and increasing rapidly.
  - About 80,000 3490 and 3590 tapes.
    - About 30,000 of them are in robots.
  - There have been 3 migrations in 7 years (18 track to 36 track to extended length to 3590).

Lawrence Livermore National Lab (cont.)

- Some observations about the LLNL approach
  - New data are written to new media.
  - Existing files are rewritten to new media.
  - A utility is used to repack badly fragmented tapes (which writes to new media).
  - Anything not rewritten within 7 to 10 years will be forced to new media.
  - Files in the robots migrate but files in the vaults lag (perhaps far) behind.
  - Some file migrations skip one or more generations of technology.
- *With this approach, backward read compatibility is very important.*

Two Large Systems On The Way

- National Imagery and Mapping Libraries
  - The largest of these libraries will
    - Ingest about 5 terabytes/day
    - Grow to about 7 petabytes
    - Keep data for 5 years
    - Service thousands of simultaneous users
    - Provide access to significant existing data
  - The development effort is under way
- NASA Earth Observing Satellite
  - Data storage to begin mid 1999
  - Data growth in the 1 terabyte per day range

The message is that the problem is already here.

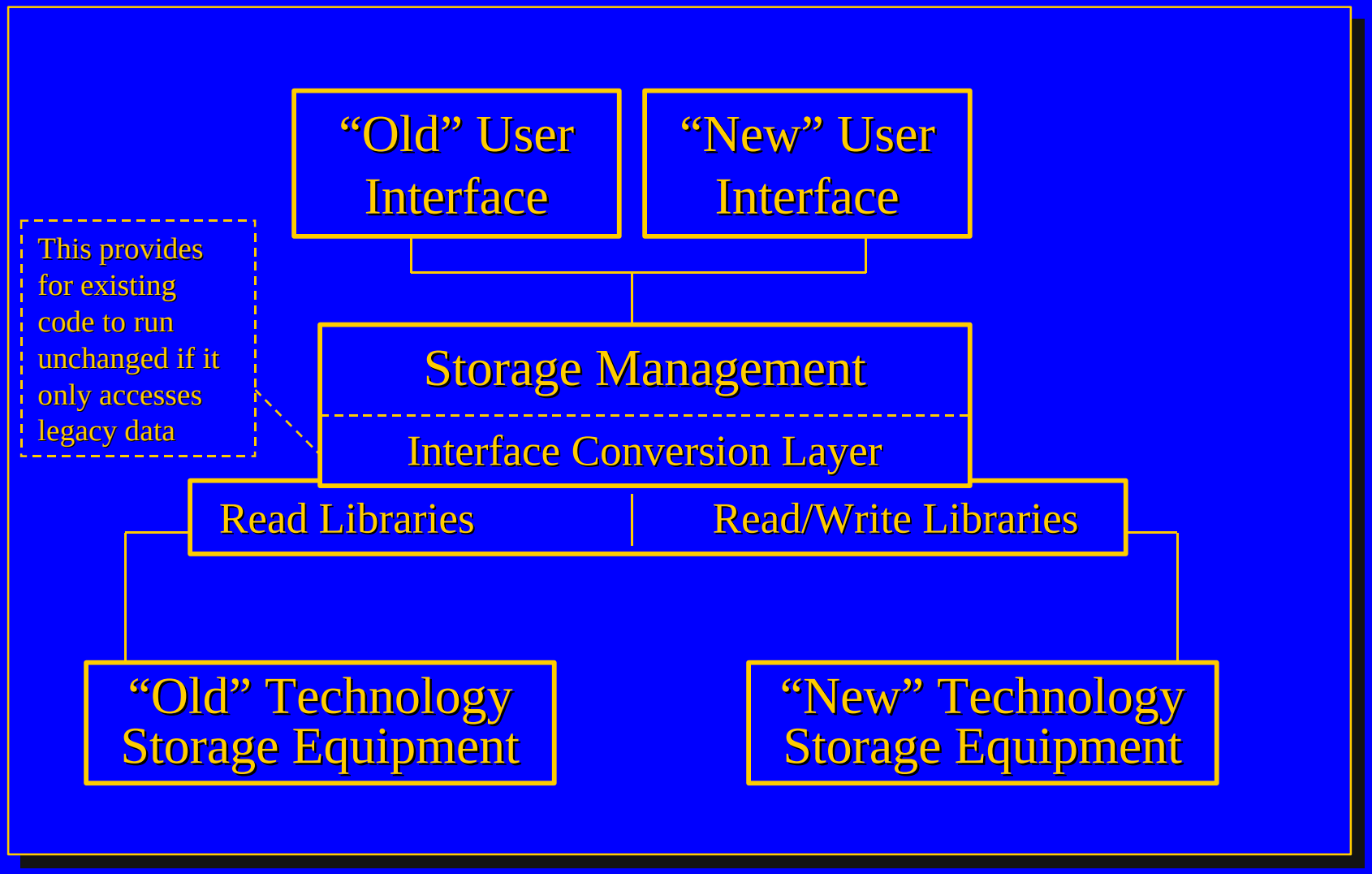
A Few Thoughts and Observations

- Strong support of standards efforts would help in transitioning to new technology.
  - IEEE Storage Systems Standards Working Group
  - AIIM FMP (portable metadata) working group
  - ISO International Archiving Workshop
- A noteworthy portion of the system resources needs to be allocated to the transition process.
- Older media need to be shepherded (sticky tapes, etc.).
- Backward write compatibility provides a means to skip technology generations with seldom used files.

A Few Thoughts and Observations (cont.)

- Most facilities will not allow any significant down time (hours) for transitioning.
- If storage management software is changed
  - It is necessary to do a bulk metadata update so access can be quickly provided to the old and new data.
  - It may be necessary to “acquiesce” the system for some period of time (a day?).

A Transition Architecture



Some Closing Thoughts

- It seems less and less likely that migration for large digital libraries can be done as a “point in time” or a bulk process.
- It is clear that transition planning should be a part of the initial planning for new digital libraries.
- It appears that concurrent access to data on old technology and new technology for an extended period may be required.
- It appears that a “continuous” migration following the introduction of new technology is advantageous.

My personal opinion is that “continuous” migration for large digital libraries is required.