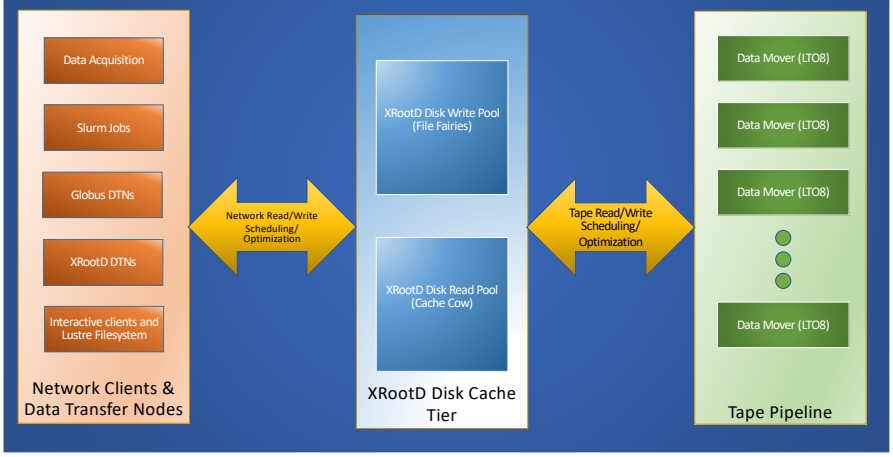
High Performance Tape I/O strategies for Distributed Storage Infrastructure

Bryan Hess and Christopher Larrieu
Thomas Jefferson National Accelerator Facility, Newport News Virginia

U.S. DEPARTMENT OF ENERGY | Office of Science

Jefferson Lab

Jasmine Three Tier Architecture



Mass Storage Evolution

Goal: Mass Storage Infrastructure for a diverse I/O and service level mix, sustaining high throughput to network clients local and remote.

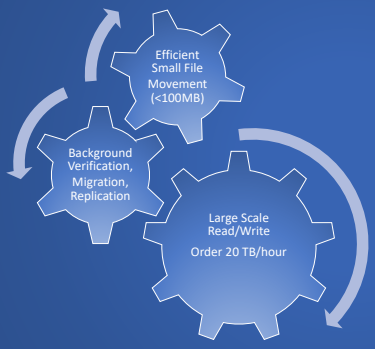
Jasmine, Jefferson Lab's Tape-backed Mass Storage System has been in use since 1999. The system was recently refactored to meet a new generation of workloads.

Redesign Drivers and Goals
- Improve small files operations, a challenge to overall system throughput
- Increase Community Tools Integration; Focus on Tape Performance as core mission
- Focus on Highly Efficient Tape use the core function
- Make daily workflows disk based; hide tape latency and operations from users

Addressing the Challenges
- Central database provides consistent state to all components
- Parallel data transfers are used to fill I/O pipes. Having multiple files ready concurrently allows for a steady stream of data to fill the pipe.
- Caching is used to guarantee a quality of access for files destined for tape
- All files destined for tape go through the file fairies, where they are cached until the scheduler releases them for writing

Service Levels
- Data Capture from Running Experiments
- Background verify of all writes after a tape is full
- Background duplication of high-value data
- Background migration of tape media from generation to generation
- Small files (>100MB) on disk; tape as deep archive
- Store files on disk until duplicates are created
- Store recent "hot data" on disk via policy mechanism

Maximum Tape Performance for Large and Small Files

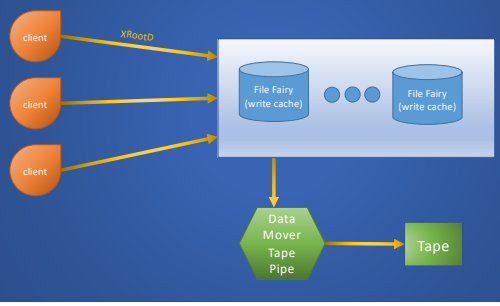


Write Implementation Goals
1. Keep Tape Rolling
2. Do not start tape unless you can stream it
3. Minimize tape load/Unload
4. Avoid costly tape file marks and slow seeks

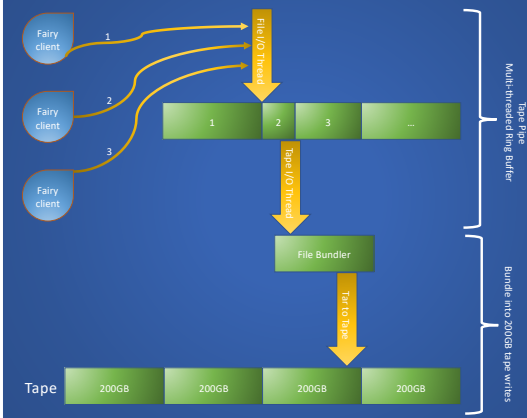
Read Implementation Goals
1. Cache Reads on XRootD Servers
2. Expandable read pool
3. Tunable, Policy-based file retention

Strategy
1. Separate read path and write paths
2. Central scheduling, distributed components
3. Aggregate Streams for High Throughput I/O
4. Small Files archived to tape but always disk resident
5. Integrated SCSI interface

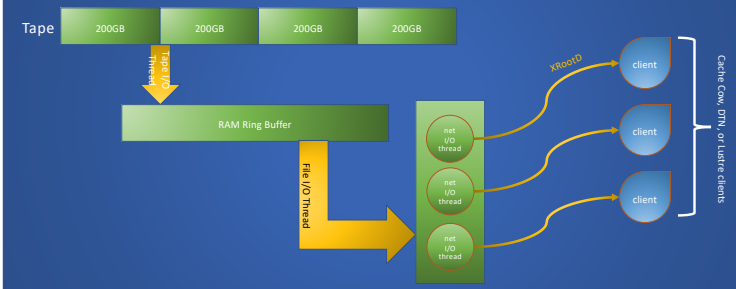
Tape Writes 1: Network to Data Mover



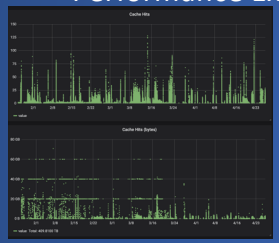
Tape Writes 2: Data Mover Tape Pipe



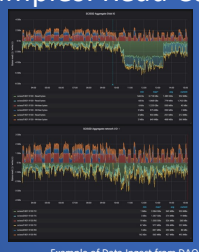
Tape Reads Demultiplexed to Clients



Performance Examples: Read Cache, Write Cache, WAN, and Totals



Example of CacheCow read Cache hits for small files



Example of Data Ingest from DAQ

WAN Network Performance to remote Clients

Storage Summary
- Over 100PB on Tape
- Over 2.5 PB Read cache
- 0.5 PB Write Cache
- 7PB (total) Lustre
- Ingest capability ~10GB/sec (sustained)

Christopher Larrieu

Bryan Hess

JSA Jefferson Science Associates, LLC
Thomas Jefferson National Accelerator Facility