Lessons Learned

Upgrading for increased throughput

LIBRARY OF CONGRESS

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

Formerly NAVCC

The Challenge

“This is an Archive. We can’t afford to lose anything!”

- Our customers are custodians to the history of the United States and do not want to consider the potential loss of content that is statistically likely to happen at some point.

Solutions

- Generate SHA1 at source and verify content after each copy
- Store 2 copies of everything digital
- **Test and monitor for failures: Archive Integrity Checker and fix_damaged**
- Refresh the damaged copy from the good copy
- Automate as much as possible
- Acknowledge that someday we’re going to lose something
  - What’s that likelihood?
  - What costs are reasonable to reduce that?

Bigger Problems

File sizes are increasing. We will soon be processing 1 TB files

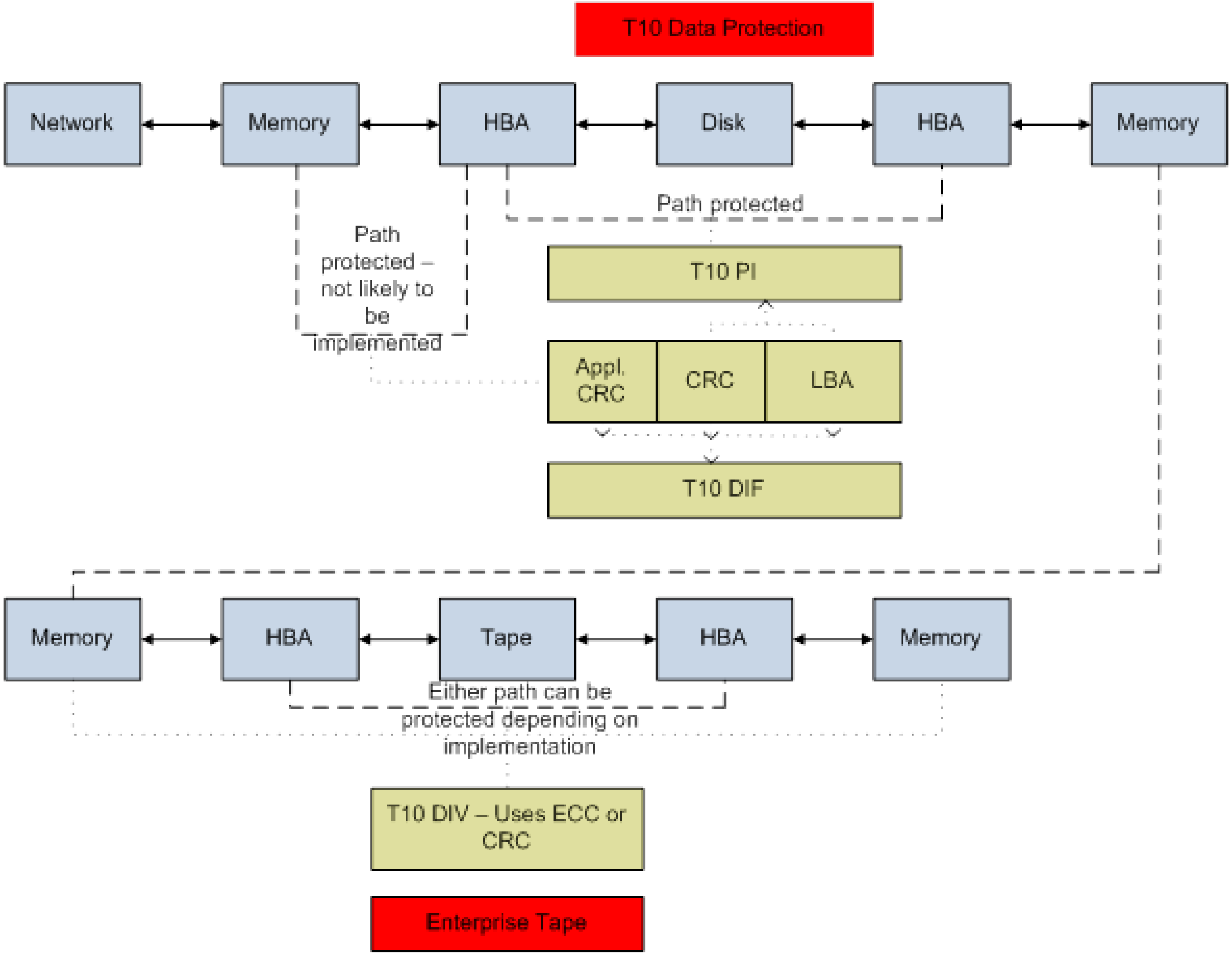
- Audio: 1GB (96/24)
- SD Video: 30 GB (mxf wrapped jpeg2000)
- HD Video: 100’s GB
- 4K scan of film: 1 TB

This means that a failed SHA1 will require retransmission

- Costing hours of time and wasted resources
- Typical individual PC disk speeds are 25 MB/s. Typical times to retransmit
  - Audio: < 1 min
  - SD Video < 30 min
  - HD Video < 3 hours
  - 4K scan of film < 4 hours (assuming faster drives in this environment)

Ensuring data integrity on the data path becomes more crucial as file sizes increase. Monitoring for marginal errors greatly improves reliability and cost effectiveness of technological investments

Datapath Integrity Field (T10-DIF)



T10000C Solution

T10-DIV

- A variation of the T10-DIF (Data Integrity Field) that adds a CRC (SSCA3 not the DIF standard) to each data block from the tape kernel driver to the tape drive
- Verifies the FC path to the tape drive and verifies each write to tape by reading afterward
- Requires Solaris 11.1 and SAM 5.3
- Can be used to verify the tape content without staging back to disk
- We look forward to a tape to tape migration that will use this information to validate the content read from tape during the migration. This assumes very few files are deleted and repacking will provide little value. True for our archive.

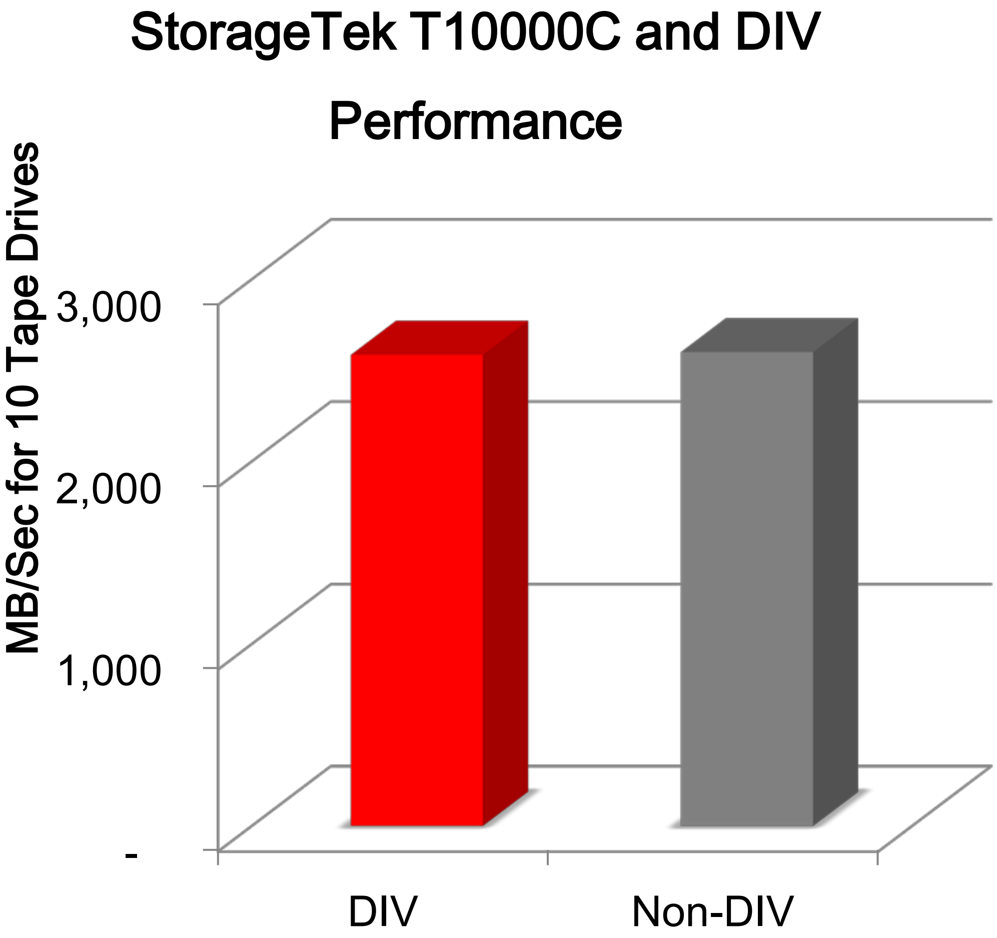
T10000C DIV Performance

T10-DIV

- How does DIV scale? There's a computation involved. Where does that happen?
- Oracle created a test bed and shared the results of their testing. In the process several interesting things were discovered and improvements added
- There are three states for DIV: off, on, verify. These results are for the on setting
- The server used here is a T4-4 with a limited domain of 32 of 64 cores
- Eight FC 8Gb ports, 3 drives zoned per port, 4 ports for tape, 4 ports for disk

T10000C DIV Performance

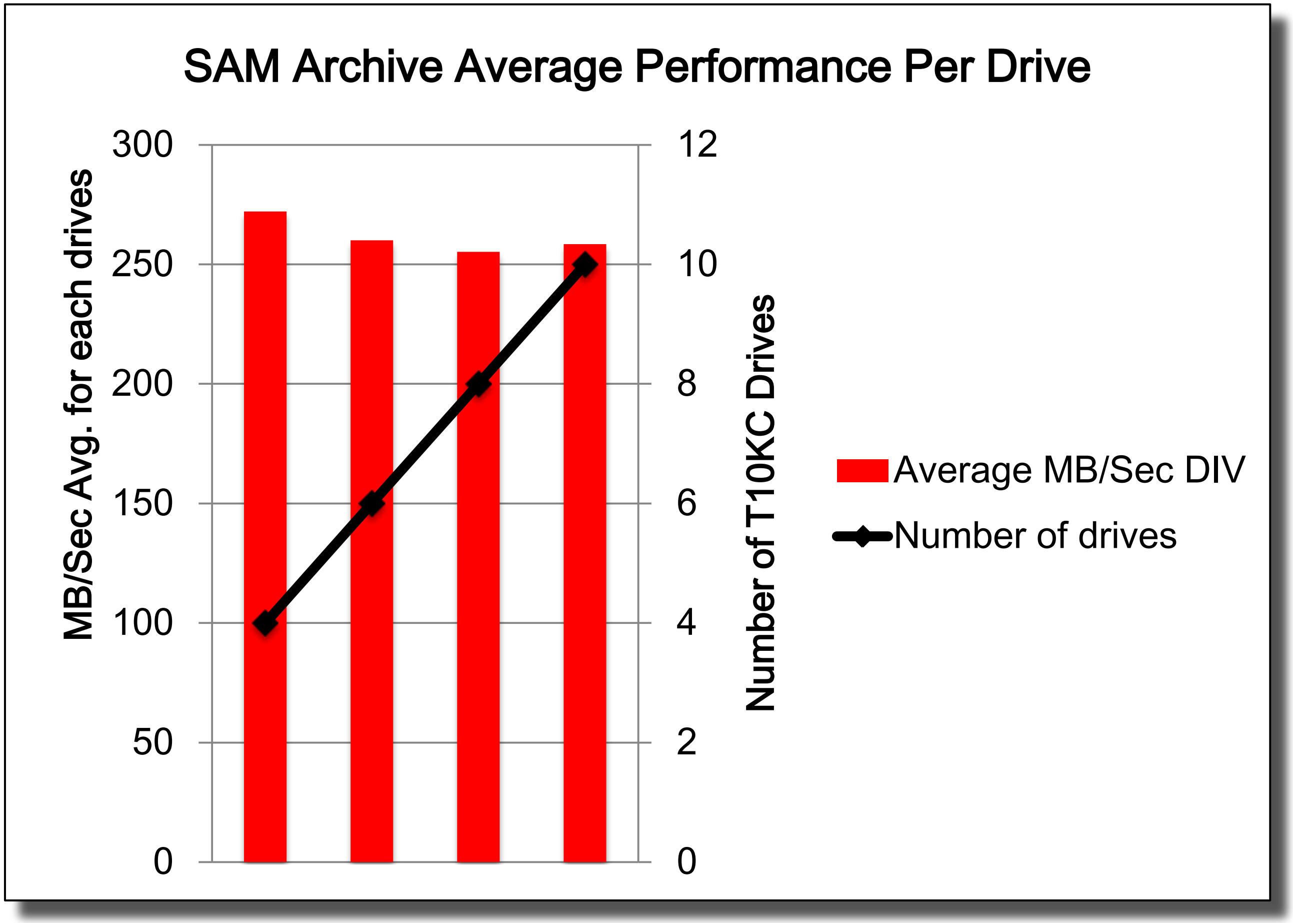
Change in performance



- DIV has no impact on performance
- Maintain 10 StorageTek T10000C drives at optimal speed

T10000C DIV Performance

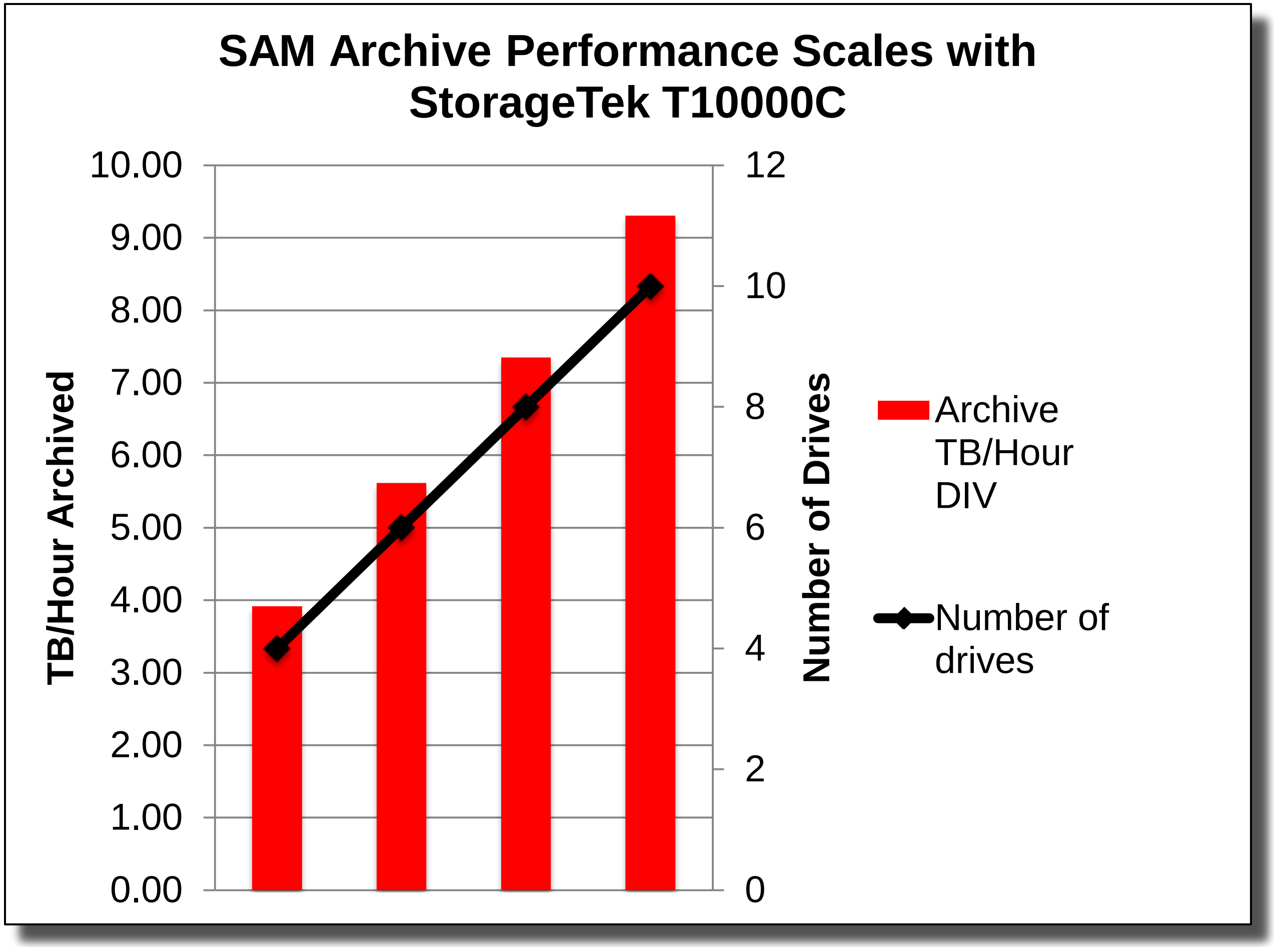
Individual drive throughput



- Average Performance per drive not impacted by additional drives
- Maintain Optimal Transfer rates as drives are added
- Recommend 4 tape drives per HBA port

T10000C DIV Performance

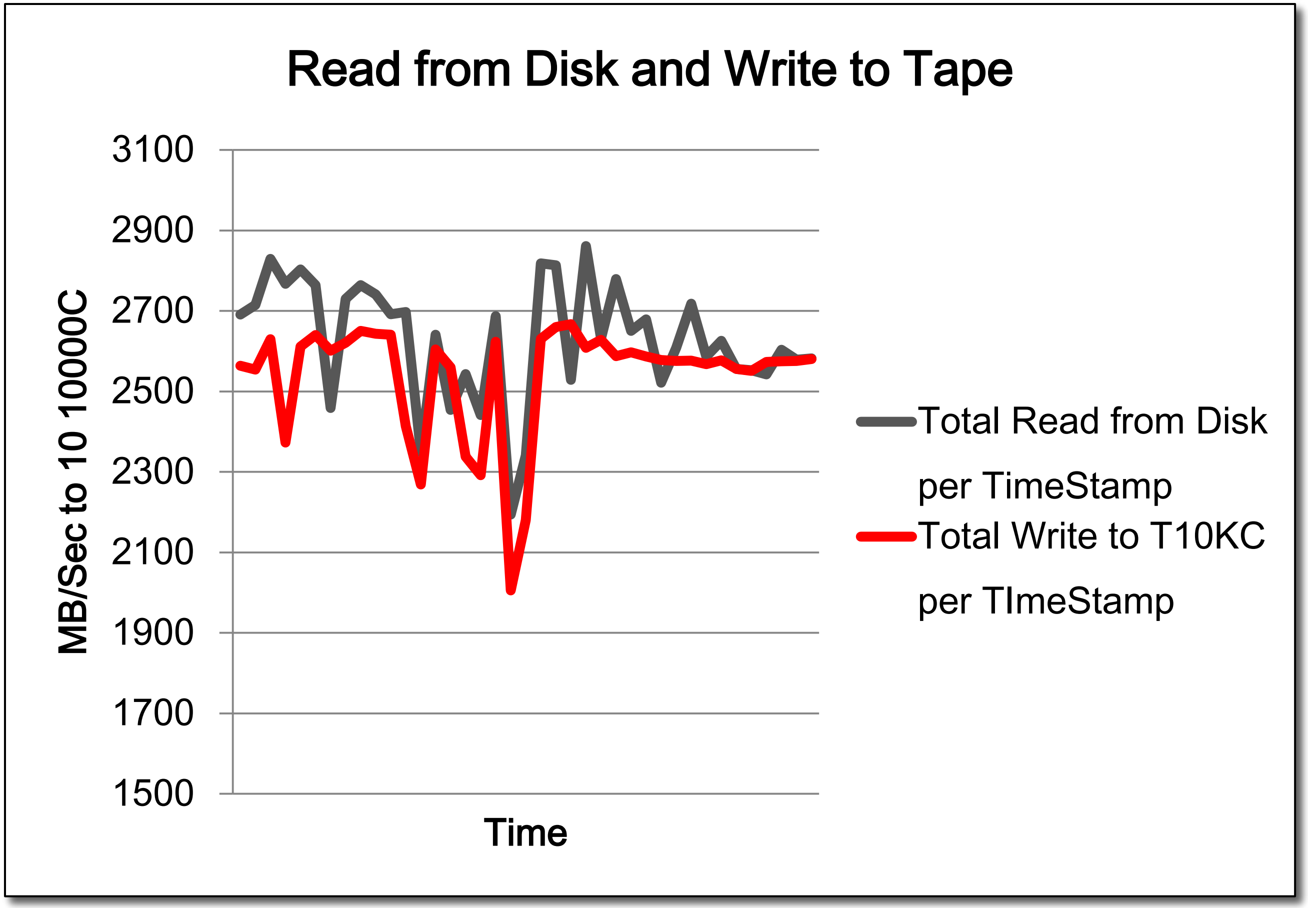
Overall Throughput



- Linear Scale
- 10 drives are maintained at optimum speed

T10000C DIV Performance

Disk performance is key



- 10 StorageTek T10000C drives
- Disk configuration
  - Match disk configuration to archive requirement
  - Reads from disk are as important as writes to tape

Key Architecture Points

- Know the amount of data you need to move over a given time period. 5 TB/12 hours
- Know your disk/tape/server speeds
- We built our cache for 6X our throughput needs because
  - Daily: Write once, read three times. Once for SHA1 check, two more times for each tape copy
  - Testing: Write once, read once
- We built a separate set of LUNs for migration
  - Migration: Write once, read three times. Once for SHA1 check, two more times for each tape copy
- Test and Monitor
  - Standard Operating Procedure for checking network, systems, storage and tape
  - Syslog on all devices that support it. Build pattern recognition for issues whenever they present.
  - HSM filesystem for damaged files, errors
  - Storage Tape Analytics
  - Service Delivery Platform/Automatic Service Request/SDP II

Recent Changes

- 6.5 GB/sec SAN storage for HSM cache – allows us to scale from 5 TB/day to 15 TB/day
- 100 TB NAS storage for proxies and staging instead of SAN – reduce deployment and maintenance complexity
- 7018 switches support needed 10 Gbe (2.6 Tb backplane upgrades will fix oversubscription on 48 port 10 Gbe cards)
- DWDM upgrades from 2 Gbs to 10 Gbs
  - 10 Gbe
  - 8 Gb FC
- Virtualized infrastructure – reduce deployment and maintenance complexity
  - Solaris zones
  - VMWare systems awaiting networking virtualization hardware
- T10KC tapes – provide better $/TB and data integrity
  - Migration beginning in Q4, 2013
  - Full SHA1 verification of content during migration
  - T10 DIV support (SSCA3)
  - Tape Analytics to be implemented to monitor tape storage and drive integrity

Future

- Testing HPSS implementation
- Oracle Tape Analytics to monitor marginal issues with tape drives/library
  - Drive code upgrade, required ACSLS version not available for Solaris 11
  - Older libraries require an HBT upgrade for memory
- Migrating 3.6 PB of T10Kb tapes/data to T10Kc tapes
- Evaluating next generation infrastructure at 15 GB/s
- Continue to monitor HSM solutions for technical excellence
  - Evaluate underlying software and hardware technologies
  - Data integrity
  - Scaling meta data
  - Scaling data throughput
  - Migration strategies

Architecting for Data Integrity

Questions?

- AXF format
- Oracle provides end to end data path integrity in some of its appliances
- Red Hat is in discussions to improve data path integrity
- SHA1 versus SHA2-256/512

Scott Rife

srif at loc dot gov