Limits of Scalability for Large Archives

April 18, 2012

Henry Newman

Instrumental Inc

CTO/CEO

hsn@instrumental.com

Copyright 2012, Instrumental, Inc.

The Archive Problem

Instrumental
HIGH PERFORMANCE INNOVATION

- What archivists expect:
  - The expectation of archival data is that what you put is what you get out and that performance is completely scalable no matter which archival software is used
- What archivists find:
  - There are limits to the scalability of data due to bottlenecks to performance

Copyright 2012, Instrumental, Inc.

What I Want to Cover

- Scalability of data integrity based on hardware construction, CRCs and ECC
- POSIX standards and the impact on scalability
- How expectations are likely not well understood by management
- What needs to be done to get us where we need to be

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

Based on hardware construction, CRCs and ECC

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

- Hardware construction
  - Performance bottlenecks in memory, PCIe, switches, and storage
  - Linear scalability is not occurring
  - Migration times are not scaling
  - Rebuild times are increasing
  - No per file checksum standard
  - Computing software checksums is CPU intensive
  - Better to have hardware checksums, but no end-to-end data integrity at this time
    - The promise of T10 PI has been a promise for a long time

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

Instrumental
HIGH PERFORMANCE INNOVATION

- Increases in rebuild times
  - Higher capacities but transfer rates not increasing as fast
  - RAID-6 is reaching the end of its useful life

Year	Drive Size in GB	Drive Type	Max Transfer Rate in MB/sec	Estimate Time to Read the disk	Drives to Saturate Channel	Estimate Time to reconstruct RAID-6 8+2 at 90% of full rate	Total Time in hours RAID-6 reconstruct
1994	4	SCSI	9	556	2.2	4444	1.23
1998	18	USCSI	29	776	1.4	10000	2.78
2002	146	FC	89	2051	2.2	16222	4.51
2005	300	FC	119	3151	3.4	16667	4.63
2009	450	FC	125	4500	6.4	12500	3.47
2009	1500	SATA	105	17857	7.6	41667	11.57
2011	3000	SATA/SAS	112	33482	7.1	83333	23.15
2012 est	4000	SATA/SAS	129	38820	6.2	111111	30.86

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

- ECC
  - Error Correcting Code – traditionally used for memory but can also be used for storage
  - 8 bits for 64-bit paths; can detect and automatically correct errors of 1 bit and can detect, but not correct, errors greater than 1 bit
- CRC
  - Cyclic Redundancy Check - used in networks and storage devices
  - Usually 16 or 32 bits in length using various algorithms
- Different data integrity methods use either ECC or CRC
- There has not been much change in a long time

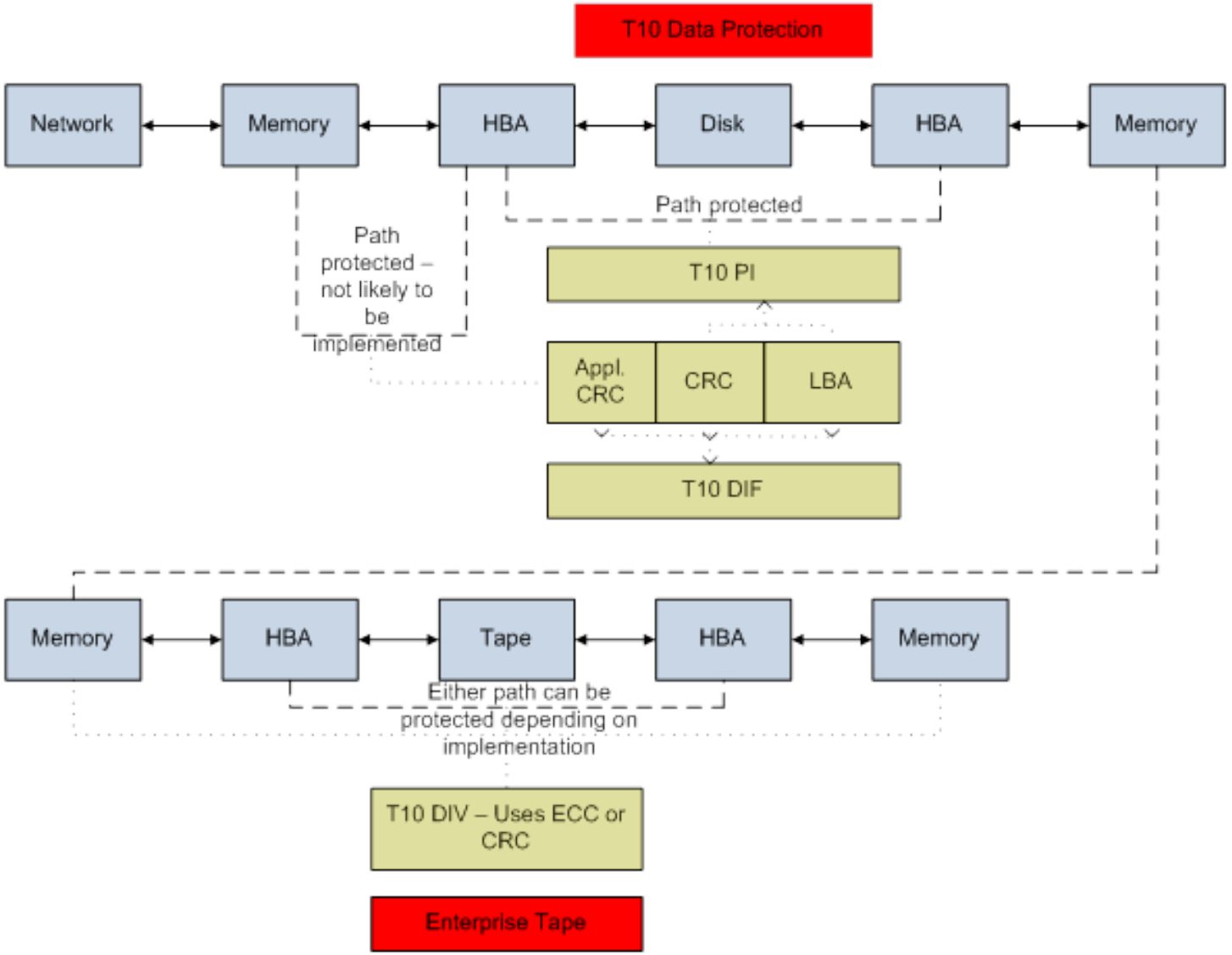
Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

- Attempts to address data integrity via hardware
  - T10 DIF (Data Integrity Field disk)
  - T10 PI (Protection of Information disk)
  - T10 DIV (Data Integrity Validation tape)
- No full implementation of standards among vendors
- No coverage of the entire data path as application CRC not implemented in VFS layer
- Users forced to develop their own data integrity checks and run them periodically

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity



Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

- Hard error rates:

Device	Hard Error Rate (1 bit in this number of bits moved)	PB Equivalent Data Moved Before Error
Consumer SATA	1.0E+14	0.01
Enterprise SATA	1.0E+15	0.11
Enterprise SAS/FC	1.0E+16	1.11
LTO	1.0E+17	11.10
T10000A/B/C IBM TS11XX	1.0E+19	1110.22

- Increased rebuild times and increased I/O increase the likelihood that another hard error will occur during the rebuild

Copyright 2012, Instrumental, Inc.

Scalability of Data Integrity

- Undetectable bit error rates; a.k.a. silent data corruption:

Type	Channel Error Rate	Undetectable Error Rate
Ethernet	1.0E-12	1.0E-21 estimated
SATA	1.0E-12	1.0E-17
Fibre Channel	1.0E-12	1.0E-21
SAS	1.0E-12	1.0E-21
FC/SAS with T10 PI	1.0E-12	1.0E-28

- Some form of data integrity checking is crucial since the transfer of data into and out of large archives will encounter these errors; **it is not if these errors will occur but when**.

Copyright 2012, Instrumental, Inc.

POSIX Standards

The impact on scalability

Copyright 2012, Instrumental, Inc.

POSIX Standards

- Issue for applications where there are many threads writing to a common shared file system
- POSIX atomic behavior is an issue for databases
- POSIX extended attributes - limited or no information on data provenance, backup and archiving, user metadata, file reliability and other attributes
  - No standardization in this area and no plans

Copyright 2012, Instrumental, Inc.

POSIX Standards

- Changes to the POSIX I/O standards have been proposed to address some of the issues
  - None of these have been ratified and ratification is not expected
  - Proposed set of extensions for POSIX for HPC
    - When file opened by multiple clients, metadata server revokes any read caching and write buffering capabilities to force consistency and this greatly reduces performance
    - Cluster file systems that enforce POSIX consistency require stateful clients with locking subsystems

Copyright 2012, Instrumental, Inc.

Just a few POSIX Standards Extensions

- Possible POSIX I/O that would help significantly
  - Needed to be more friendly to HPC, clustering, parallelism, and high concurrency applications
  - Ordering – Replace streams of bytes with more applicable method for distributed memories mapped to many storage devices
  - Coherence – Overhead of cache invalidation for reads is high; block boundaries can present coherence issues for application
    - Method for applications to assume all responsibility for coherency is needed
  - Metadata – Standard support of “lazy” attributes needed along with portable bulk metadata interface for file system metadata
  - Agreement on POSIX extended attributes for data integrity, archiving and other areas

Copyright 2012, Instrumental, Inc.

POSIX Standards

- Possible POSIX I/O extensions (cont.)
  - Extensions for archive applications – To use POSIX API
  - Assumption by most UNIX utilities that all data is on disk
  - T10 DIF support Per file checksum such as a SHA256
  - Provenance to ensure integrity of chain of custody of file
  - Lots more
- None of this is going to happen as the vendors that control POSIX do not want the extra work
  - And if I were a cynical person, they do not want standards so they can sell proprietary products

Copyright 2012, Instrumental, Inc.

How Expectations Are Not Understood by Management

It's hard to do it the right way and it is easy to do it the wrong way

Copyright 2012, Instrumental, Inc.

Management Expectations

Instrumental
HIGH PERFORMANCE INNOVATION

- More hardware will solve the problem but:
  - Lack of scaling in HSM and file system
  - Migration times are increasing
  - More time required to validate data in the archive
- The existing personnel can handle it
  - Lots of planning required and not given needed emphasis
  - Monitoring and correcting errors takes lots of time with existing tools

Copyright 2012, Instrumental, Inc.

Management Expectations

Instrumental
HIGH PERFORMANCE INNOVATION

- Migration times:
  - Assumes use of 4 Oracle/StorageTek T10000C drives for migration
  - Shows migration time in days with various bandwidth percentages assumed for system and tape
  - Migration with assumed minimal impact to production

PiB of Data in Archive	Migration Time (Days) - 30% of the system bandwidth	Migration Time (Days) - 30% of tape drive bandwidth of 4 T10000C drives	Migration Time (Days) - 50% of the system bandwidth	Migration Time (Days) - 50% of tape drive bandwidth of 4 T10000C drives	Migration Time (Days) - 70% of the system bandwidth	Migration Time (Days) - 70% of tape drive bandwidth of 4 T10000C drives
2	18	86	11	52	8	37
5	45	216	27	129	19	92
10	90	432	54	259	39	185
20	180	863	108	518	77	370

Copyright 2012, Instrumental, Inc.

Management Expectations

- With large archives, migration will become a continual process
- Validation of data while migration is occurring can be problematic
  - Was data invalid prior to migration?
  - Did migration change the data?
  - Did something change the data in the path?
  - Need lots of CPU cycles to drive I/O for migration while also performing data validation
- Tape channel speeds are not scaling as fast as capacities – similar to problem with disks

Copyright 2012, Instrumental, Inc.

What Needs to be Done

How do we get where we need to be?

Copyright 2012, Instrumental, Inc.

What Needs to be Done

- Users need to be able know what they have and manage their data
- Administrators need to be able to implement Information Lifecycle Management (ILM) policy based on business rules
- Current frameworks have:
  - No ILM support for archival systems
  - Only have limited ILM support that is proprietary
- ILM benefits include:
  - Users can manage their data with a UNIX file system framework
  - System Administrators can implement system wide policy
    - No longer be specific to a single vendor software
    - Very likely reducing the growth rate of data

Copyright 2012, Instrumental, Inc.

What Needs to be Done

- With the current framework:
  - All ILM functions must be done in user space
  - Each vendor is doing something different
  - Many vendors are looking at the regulatory issues with very different requirements than archival ILM requirements
  - None of what is being done moves from system to system as there are no standards
  - What the user can do from an application in a standard way is limited by open()
  - Per-file metadata does not exist in system space
  - ILM is not part of the standards process except to a limited degree by SNIA

Copyright 2012, Instrumental, Inc.

What Needs to be Done

- ILM requirement is across agencies
  - Per-file metadata, reliability and policy is needed by many agencies and is implemented currently in user space by vendors
  - Industry needs the same thing
  - If applications cannot communicate down the data path then scaling, reliability and ILM are not possible
  - Agency after agency is reinventing the preservation archive wheel

Copyright 2012, Instrumental, Inc.

What Needs to be Done

- Use POSIX extended attributes and define a common set of attributes that move with a file
  - File systems will need to support common set along with UNIX commands (ls, tar, etc.) and file transport (ftp, pNFS, rcp, etc.) will need to support framework
  - As a start, some ILM suggested attributes would address:
    - Archive/backup
    - OSD framework
    - Data integrity (PI)
    - Performance hints
  - Current T10 OSD and PI need to be addressed as part of this
    - But sadly it looks like OSD

Copyright 2012, Instrumental, Inc.

What Needs to be Done

Instrumental
HIGH PERFORMANCE INNOVATION

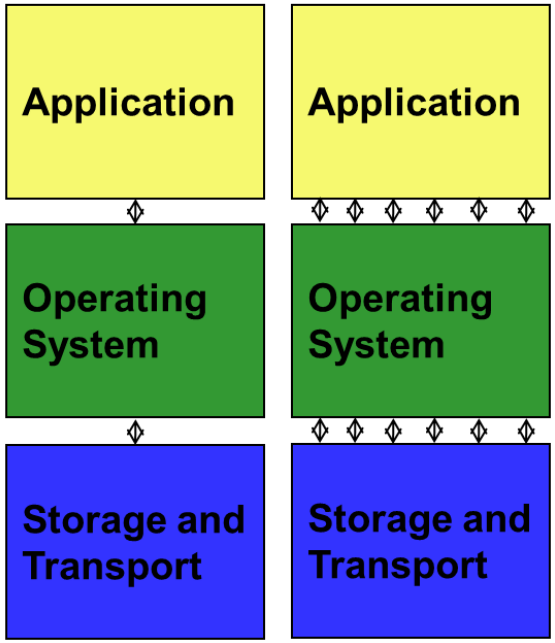
Datapath Changes are Needed

Current

Current POSIX system calls open/read/write/aio. Limited communication with OS layer

POSIX Atomic operations open/read/write/aio. No communication with physical layer

Block based storage and limitations of 30+ year old technology



Future Needs

ILM framework that applications can interface. Information is passed on through generations of systems. User searchable

ILM archival and backup policies based on per file information or defaults.

More intelligent storage to address issues such as fragmentation, encryption, OSD and reliability to name a few

Data path today is the same as the data path 20 years ago

Focus on end-to-end view of I/O in whole datapath. Moved from 3 separate boxes to 1. The benefits of this framework extend to the entire datapath, not just ILM.

Copyright 2012, Instrumental, Inc.

Thank you for listening

Copyright 2012, Instrumental, Inc.