What Is Wrong With the Data Path

Henry Newman

10 June 2008

Agenda

What is broken and why

How we got here

Some Thoughts

Instrumental
www.instrumental.com

WHAT IS BROKEN AND WHY

What the user can and can't do and the impact

Instrumental
www.instrumental.com

What Is Broken

- *The open() system call has not changed in the last 20 years*
- *What the user can do from an application is limited by open()*
- *Scaling and reliability require more information down the data path*
  - *True for all I/O not just HPC*
- *What about archive policy*
  - *Per-file policy must managed in user space*
    - *People are reinventing this wheel time and time again*
- *Per-file metadata does not exist in user space*

Our Broken I/O World

- ***T10 DIF allows a checksum to be transmitted from the application to the disk drive***
  - ***What standard framework exists to allow the application to do this?***
- ***T10 OSD allows a variety of I/O hints and security features per object***
  - ***What framework is available to access?***
- ***Per-file metadata, reliability and policy is needed by many, and is implemented currently in user space time and time again***
  - ***LOC, NARA, DoD HPCMP just to name a few***

HOW WE GOT HERE

A trip down memory (I/O) lane

Instrumental
www.instrumental.com

The I/O Stack

The USL/XOpen/OpenGroup (POSIX), SNIA(no standards yet)

Application/User space

T10, T13, IETF (NFS), SNIA (no standards yet) OpenFabrics, T11 (FCOE)

Drivers, OS and File System Interface

T11, OpenFabrics, IEFT, Ethernet, T10, T13, SNIA (no standards yet)

Storage and Transport

The data path today is basically the same as the data path of 20 years ago. Limited improvement for management (errors or system)

Instrumental
www.instrumental.com

Standards Process

- ***User space is controlled by the OpenGroup***
  - ***When did the open() last change***
    - ***I think 1990 for O_Direct***
- ***SCSI and ATA are controlled by T10 and T13***
  - ***New Standards for DIF and OSD, but no changes above in the stack***
- ***Channels are controlled by T11 (FC), IEEE for ethernet, OpenFabrics Alliance for IB, IETF for IP and NFS***
- ***Seemingly not a lot of intergroup communication***
- ***The USG worked on standards in the 70s and 80s but no more***
  - ***DARPA, NSF, DoD and others***

Standards Process Seems Disjointed

- ***No standards for***
  - ***File systems other than the interface***
  - ***HSM policies***
  - ***Per-file metadata***
- ***Lots of different standards bodies as mentioned***
  - ***Some have competing goals***
  - ***SNIA is industry controlled with little to no outside input and little to no agreement on standards***
    - ***Only one standard so far SMI-S***
- ***No standards for error correction for each file***
  - ***In the file system nor for archives***
  - ***DIF addresses the packet issue but not the file***
    - ***DIF is not support on tape or SATA***

The Problem

If applications cannot communicate down the data path then full POSIX scaling, reliability, ILM are not possible

Agency after agency is reinventing the preservation archive wheel

- ***NSF, LOC, NARA, DoD HPCMP just to name a few***
  - ***User space applications without standard to manage collections***
    - ***These are impossible to port***
- ***Preservation is just one example of wheel re-invention***

SOME THOUGHTS

Here are some ideas that Gary and I have discussed that would improve things

Framework

Use POSIX extended attributes and define a common set of attributes that move with a file

- ***File systems will need to support the common set and ftp, pNFS, rcp, etc will need to support the movement***
- ***As a start, some suggested attributes would address:***
  - ***ILM***
  - ***Archive***
  - ***OSD framework***
  - ***Data integrity***
  - ***Performance hints***

Of course, file systems will need to deal with some of the T10 DIF fields

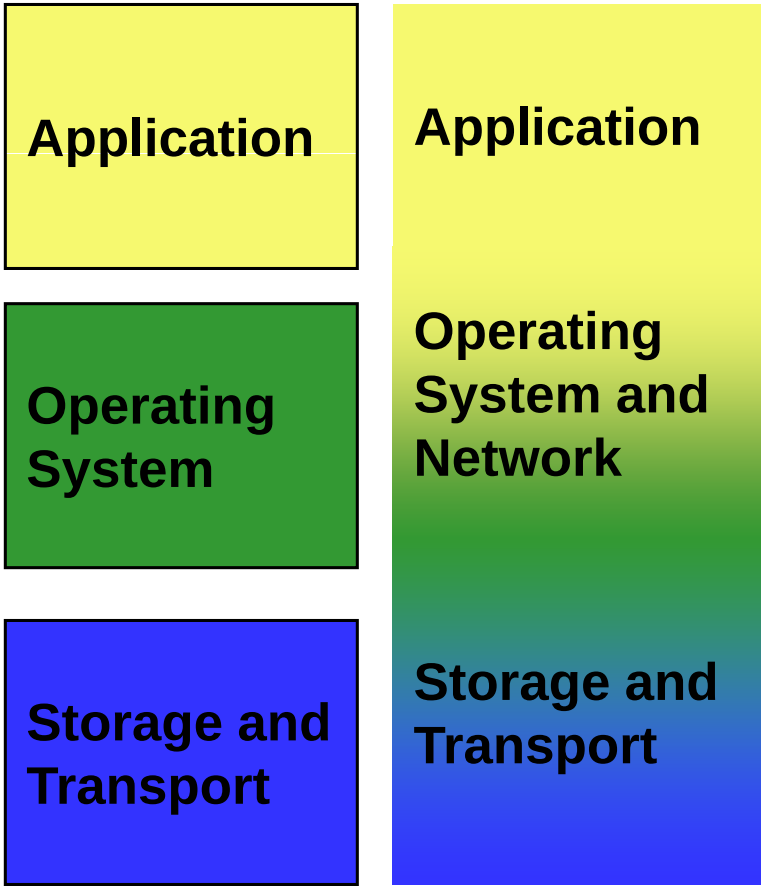
Data Path Research is Needed

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

Changes to support new constructs for different types of latencies for data

OSD combined with networking constructs could address different latencies

Given physical limitations of storage, optimizations must be done at higher level to impact the technology

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

Need to focus on an end-to-end view of I/O

Instrumental
www.instrumental.com