

***What Is Important and  
Wrong with the Datapath for  
Petascale Storage Systems?***

**Henry Newman  
Digital Technology Center  
14 January 2008**

# Agenda

***Hardware Components***

***Software Components***

***Areas of Concern***



# *Hardware Components*

**What are the hardware components  
and the potential area of risk and  
change**

# What Hardware is in the datapath

## **Server**

- ☐ *CPU's, memory, PCIe bus, NIC, HBA/HCA*

## **Switch**

- ☐ *Interconnect to storage and external network*

## **Disk based storage**

- ☐ *RAID, Just a Bunch Of Disks (JBOD) connectivity issues*

## **Tape**

- ☐ *Tapes and Libraries*

## **Future hardware storage technologies**

**All of these impact cumulative service cost and replacement cost cycles -End of Life/Service (EOS/L)**

**What about software?**

# *Software Components*

**What are the software components  
and the potential area of risk and  
change**

# What Software is in the datapath

## ***Operating System***

- ▣ ***Operating systems change and are not necessarily compatible***

## ***Per File Metadata***

- ▣ ***Home grown or package***

***File systems affect storage management and performance (some of which affect hardware)***

- ▣ ***Storage HSMs, hardware impacts***

***Device Drivers and firmware all over the place***

***Future software storage technologies***

***All of these impact cumulative costs based on service costs and replacement cost cycles - EOS/L***

***What about hardware?***



# *Areas of Concern*

**Lots of Worries**

# Migration

## ***How do we migrate hardware and software***

- ▣ ***Everything has a life; and technology changes requires almost constant migration***
- ▣ ***Constant migration to new more compact media is required-given floor space and data growth***

## ***More important - What about our data and formats?***

- ▣ ***If PDF is only good for 25 years, how can we migrate to new formats without major work in software?***
  - ***No framework as part of technology migration to migrate formats***

***There seems to be a lack of integration between all of layers of software***

***How do we change one component or (worse yet) move to a new system?***

# Data Reliability

## ***Silent data corruption***

- ▣ ***Undetected errors***
- ▣ ***Mis-corrected errors***
- ▣ ***Is it hardware is it software***
  - ***Where is it***
  - ***What caused it***

***Per-file error detection and correction is required from file creation through the life of the file***

- ▣ ***Common and open error correction algorithms throughout the system***

# Undetectable Bit Error Rate (UDBER)

UDBER	Sustain Transfer Rate Per Second for a Year						
	0.5 GB/sec	1 GB/sec	10 GB/sec	100 GB/sec	1 TB/sec	10 TB/sec	100 TB/sec
1.E-21	0.0	0.0	0.0	0.0	0.3	2.7	27.1
1.E-20	0.0	0.0	0.0	0.3	2.7	27.1	270.9
1.E-19	0.0	0.0	0.3	2.7	27.1	270.9	2708.9
1.E-18	0.1	0.3	2.7	27.1	270.9	2708.9	27089.2
1.E-17	1.4	2.7	27.1	270.9	2708.9	27089.2	270892.2
1.E-16	13.5	27.1	270.9	2708.9	27089.2	270892.2	2708921.8
1.E-15	135.4	270.9	2708.9	27089.2	270892.2	2708921.8	27089217.7

This does not include errors as hardware degrades  
such as a failing drive or controller  
Bit error rates of most channels are 10E12

# Standards Process Seem Disjointed

## ***No standards for***

- ▣ ***File systems***
- ▣ ***HSM policy***
- ▣ ***Per-file metadata***

## ***Lots of different standards bodies:***

- ▣ ***T-10, T-11, T-13, IETF, SNIA, OpenGroup etc etc***
- ▣ ***T-10 DIF: Data Integrity Field can improve end-to-end reliability***
  - ↗ ***If used***

## ***No standards for error correction for each file***

- ▣ ***In the file system nor an archive***

# **Operational Error Management Needed**

***Collection, coordination and  
management of errors***

***Proactive error management***

- ▣ ***For example many RAID vendors hide SMART data  
from system manager***

  - ***Self-monitoring data provided by hard disks***

***Frameworks to track and manage errors  
and warning throughout the system***

- ▣ ***Tracking things like data corruption and DIF errors  
(for example)***

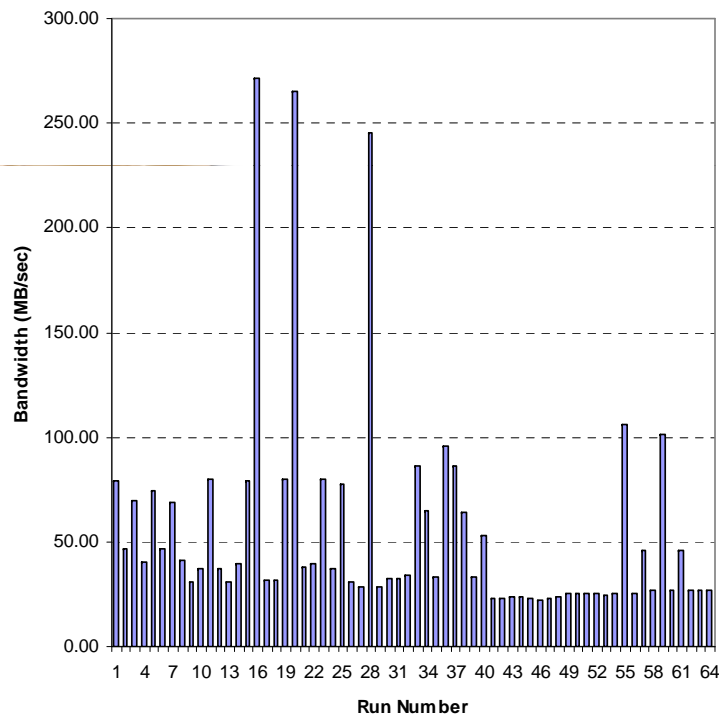


*What is wrong with all of this*

**The Datapath is a mess**

# Petascale I/O Challenges

Repeated Runs of Same Benchmark  
on 64-processor IBM System  
Reasons for Large Variance  
Not Understood  
(DoD HPCMP I/O Benchmark)



## No commonly accepted parallel file I/O benchmarks

- Strategic problem for technology, no way to inform decision makers on need for investments
  - No GUPS benchmark equivalent
  - HPCS currently developing I/O usage scenarios, not benchmarks

## Storage usage patterns are not understood

- Critical for designing high performance storage systems

## Tuning and analysis tools are crude

- Very difficult for I/O experts to find/understand/fix errors
- Tuning requires extensive experimentation

## Programmability is very poor

- Data moved in blocks by programmer
  - Approach is very difficult and prone to errors
  - No high-level representations for disk storage
- UNIX I/O standards require atomic operations
- Limited information can be based to the file system in a standard way

**Achieving high performance is extremely difficult ... and performance challenges are not well understood**

# Latency Tolerance CPUs Last 35 Years

***Hardware evolution based on memory component latencies have caused dramatic shifts in design***

- ▣ Vectors started it all because of memory performance limitations moving to***
- ▣ Multiple levels of memory (L1, L2, L3, NUMA) moving to***
- ▣ Multi-threaded CPUs***

***All of these changes are based on the need to hide latency when accessing memory as latency has increased as a function of CPU performance***

- ▣ This latency trend is not going to change***

# Latency Tolerance Data Path Last 35 Years

***Application data path has not changed in a similar way to address latency***

- ▣ ***Storage latency has changed only a little bit as compared with CPU performance***
- ▣ ***File systems do not pass topology to block devices to impact latency***
  - ***This has a major impact on RAID storage and readahead***
  - ***You cannot readahead if you do not know where the data is***

***Without addressing this latency HPCS systems cannot scale if they require access to the storage when they are running***

# It's all about Latency

***Application changes such as multithreading are the same techniques used to address latency issues in the late 70s and 80s with vector based computers***

- ▣ ***Systems today are efficient if they hide memory latency***
- ▣ ***I/O is efficient today if they can hide latency by multithreading the I/O***
- ▣ ***I/O can be efficient by making large I/O requests***

***The latency in the data path is growing as a function of computation***

- ▣ ***True for memory***
- ▣ ***True for I/O***

# *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

Application

Operating System

Storage and Transport

Application

Operating System and Network

Storage and Transport

## 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 a higher level to impact the technology

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

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

# Conclusions

***The current system and methods are broke***

- ▣ ***Standard bodies look at their single area and no one looks end-to-end***
  - ***Take T10 DIF to add the application checksum correctly you will not to make a change to the open() system call***
    - ***Are they talking to the OpenGroup/POSIX-No***

***Storage performance is limited by latency end-to-end***

***Researchers and industry must begin looking at end-to-end solutions for products and standards not just a single standards body***

- ▣ ***What is happening with T10, IEFT, Linux community, SNIA and all of the groups is not providing and end-to-end solution but in my opinion making it very messy***

***The good news is storage people will have jobs***