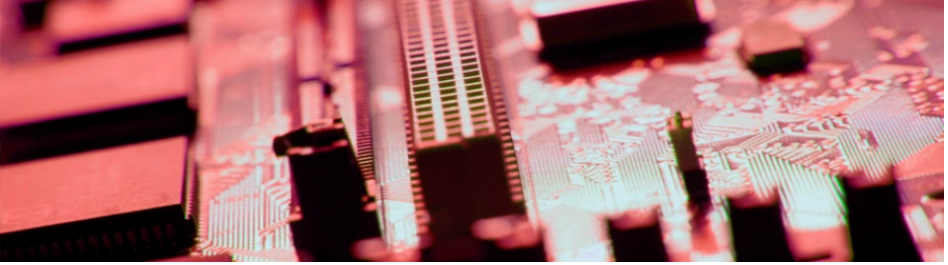


LCI workshop
on high performance clustered computing

Linux Clusters Institute: Survey of Hardware

Georgia Tech, August 15th – 18th 2017

J.D. Maloney | Storage Engineer
National Center for Supercomputing Applications (NCSA)
malone12@illinois.edu

NCSA

Baseline

- Look over slides from Pam Hill's talks at the beginner workshop in May 2017
  - http://www.linuxclustersinstitute.org/workshops/may2017/program.html
  - Segments of some following slides were adapted from her work
- Have an grasp on:
  - Drive classes and their characteristics (HDD, SSD, etc)
  - Storage connectivity options (SAS, Fiber Channel, IB, etc)
  - RAID types and their overhead
  - Difference between vendor marketed space, and what shows up in a 'df'
  - The definitions of bandwidth and latency

Understanding Your Workloads

Finding What You Need

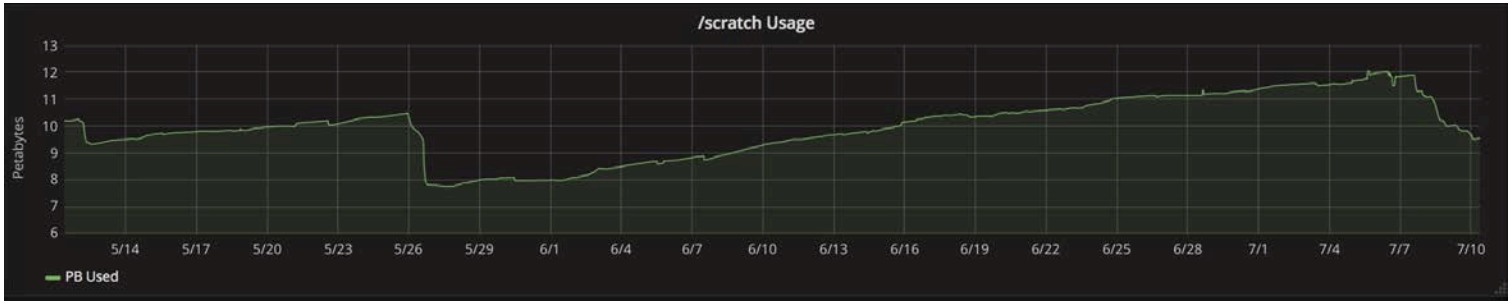
- Some system characteristics are easier to determine
  - File System capacity needed (in TB or PB)
  - Does your system process time critical workloads?
    - eg. weather forecasting, medical analysis
  - Data security requirements (encryption – eg. HIPAA)
  - How much money there is to spend 😉
- Some things are harder
  - Metadata performance needs
  - Streaming I/O performance needs
  - inode capacity

Look at Trends & Metrics

- How fast is data & metadata accumulating
- Average file age
  - How much would weekly purges help scratch
  - Are users already cleaning up after themselves
- Mean and Average File Size
- Load on current system
  - Where is the bottleneck now
  - Where do you want the bottleneck to be?
- Expected life span of the equipment

Trends & Metrics Examples

- Watching Data Growth and Purge Response

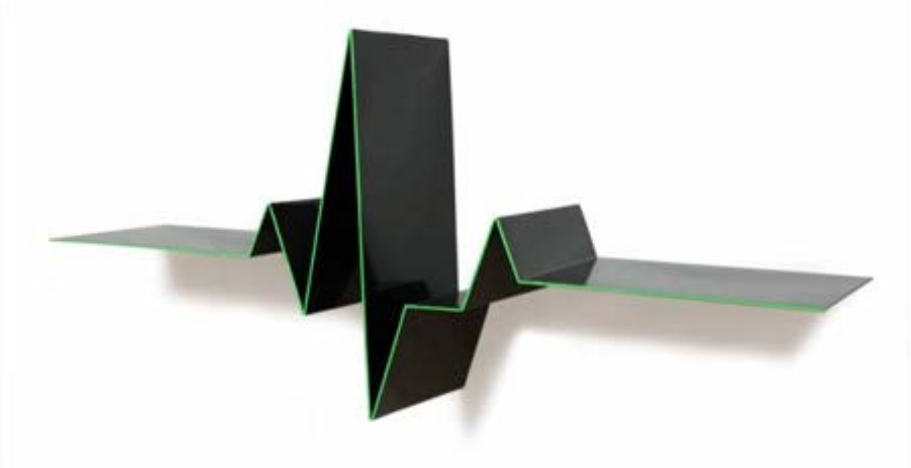


- File Size Distribution

Bucket Size	# of Files	# of Bytes
<4K	154,809,693	225,492,443,961
4K - 64K	12,721,849	166,596,433,302
64k - 1M	1,527,596	390,515,272,937
1M - 25M	6,202,999	74,398,048,664,922
25M - 100M	2,763,015	137,720,231,996,404
100M - 256M	1,935,171	321,317,787,583,963
256M - 512M	139,988	50,892,989,881,670
512M - 1G	102,879	65,079,592,689,059
1G - 5G	22,505	49,855,016,329,885
>5G	8,133	179,744,919,685,555

Talk to Users, Look at the code

- Certain HPC codes have well known I/O profiles
- Many programs don't or depend on user usage
- Some tools available to profile applications:
  - Darshan
  - SAR
  - mmpmon (Spectrum Scale)
  - nmon



Latest Storage Related Hardware Technologies

Drive Technology – Ethernet HDD

- Traditional rotational hard drive that ditches the SATA/SAS connection (electronically…same physically) for an Ethernet connection
- Very light weight ARM based CPU and some memory put onto the hard drive logic board itself
- Built for scale out architectures, OS to serve and store objects runs right on the drive
- Drive chassis functions more as a switch with uplinks than a server itself

Drive Technology - SMR

- Shingled Magnetic Recording
- Allows data to be packed even tighter on a platter at the expense of write performance
- Uses the read-modify-write technique to layer the data on the platter like shingles are laid on a roof
- Used in “archive” drives
- Lowers $/GB cost beneath similar capacity drives that use PMR

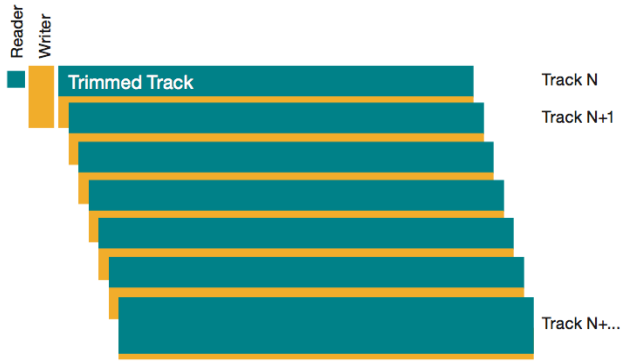


Image Credit: seagate.com

Drive Technology - HAMR

- Heat Assisted Magnetic Recording
- Decreases the grain size of the recording tracks as the heat from a laser heats the section of disk, changing its properties temporarily while the data is written
- No current marketed products, speculation that products will be available in 2018/2019 time frame
- Will be paired with helium technology to provide maximum drive density

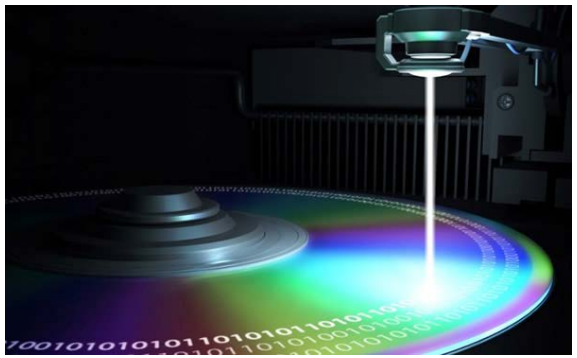


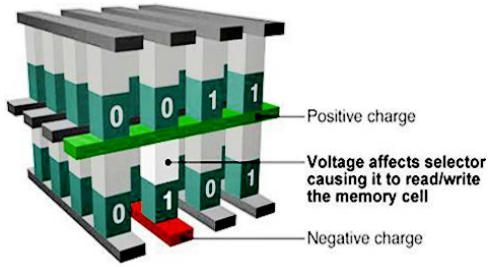
Image Credit: storagenewsletter.com

Drive Technology - NVME

- Non Volatile Memory (NVM Express)
- NAND-based flash that connects over the PCIe Bus
- Interfaces via half-height PCIe AICs or U.2 connected 2.5” drives (sometimes behind a PCIe switch)
- High bandwidth performance, high IOPS, high(ish) price
- Use Cases
  - File system metadata
  - Fast storage pools
  - Compute node LROC or HAWC (Spectrum Scale)

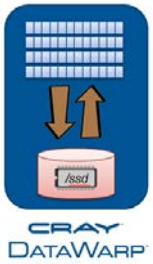
Drive Technology – 3D XPoint

- Flash technology joint developed between Intel and Micron
- New storage medium, replaces NAND gate based flash
- Will come in half-height AICs and memory DIMMs
- Compatible with upcoming Intel CPUs
- Big bonuses
  - High performance at both high and low queue depths
  - Access to memory bus (memory DIMM/Apache Pass version)
- Should be priced higher than NVME SSDs, less than DRAM
- Similar use case to NVME, could be even bigger win for in compute node caching/scratch



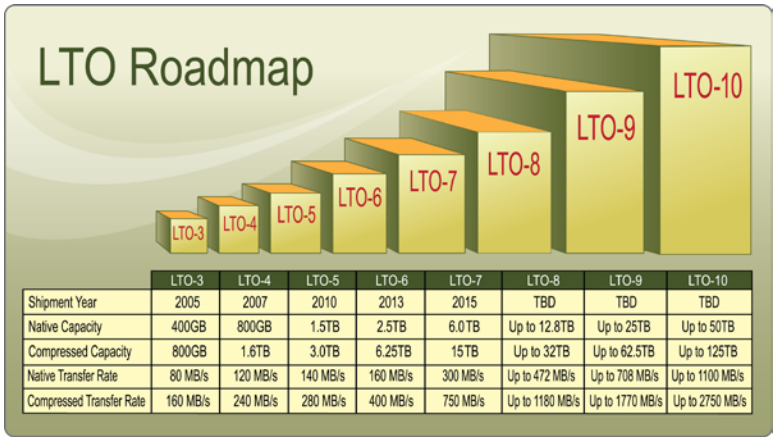
System Technology – Burst Buffer

- Layer of flash above the file system that absorbs high I/O bursts, draining data to disk during lower demand periods
- Either in node (DataWarp) or outside of node (IME, Nytro)
  - DataWarp is specific to Cray machines
- Can act as a high performance tier on top of a slower capacity focused tier
- Take rougher I/O patterns and flush them to disk sequentially which improves disk performance as well



Tape Technology – LTO 7/LTO8

- Mass availability began in Late 2015/Early 2016
- Up to 6TB of data per tape uncompressed
- As standard for tape storage, great for low power archive storage systems
- For tape to be cost effective, one needs to reach a given amount of data for economies of scale to make sense
- LTO 8 is expected to start becoming available late 2017, early 2018



	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10
Shipment Year	2005	2007	2010	2013	2015	TBD	TBD	TBD
Native Capacity	400GB	800GB	1.5TB	2.5TB	6.0TB	Up to 12.8TB	Up to 25TB	Up to 50TB
Compressed Capacity	800GB	1.6TB	3.0TB	6.25TB	15TB	Up to 32TB	Up to 62.5TB	Up to 125TB
Native Transfer Rate	80 MB/s	120 MB/s	140 MB/s	160 MB/s	300 MB/s	Up to 472 MB/s	Up to 708 MB/s	Up to 1100 MB/s
Compressed Transfer Rate	160 MB/s	240 MB/s	280 MB/s	400 MB/s	750 MB/s	Up to 1180 MB/s	Up to 1770 MB/s	Up to 2750 MB/s

Image Credit: spectralogic.com

Fabric Technology - Omnipath

- Intel owned interconnect technology
- Provides high bandwidth (currently 100Gb), low latency interconnects between nodes
- Is now available in an integrated package on Intel Xeon CPUs at good pricing
- Competitor to Infiniband, providing good price/performance and scales well
- Already powering many Top 500 Supercomputers
- Upcoming fabric technology for back-end storage networks

Fabric Technology – EDR/HDR

- Established high bandwidth, low latency fabric delivered by Mellanox
- Speeds of 100Gb (EDR) and upcoming 200Gb (HDR)
- Interface with some major storage appliances today allowing for Infiniband based SAN fabrics
- Supports RDMA (Remote Data Memory Access)

Fabric Technology – RoCE

- RDMA over Converged Ethernet
- Reduces the latency of Ethernet-based fabrics allowing them to perform closer to technologies such as Infiniband or Omnipath
- Useful since Ethernet is both cheaper and easier to deploy than the other more proprietary network technologies

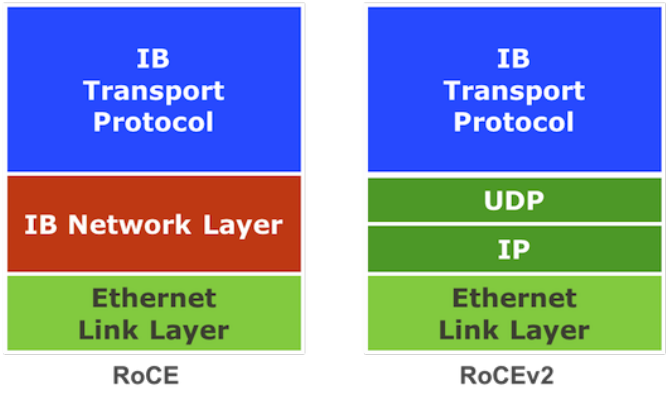


Image Credit: theregister.co.uk

Matching Hardware to Workload

Identify What's Important

- There are always trade offs, find the ones you want to make
  - Are you in a confined space, should you pay for density?
  - Is performance more important than capacity?
  - Is data security (encryption) more important than performance?
  - The classic performance/$ tradeoff
- If you don't know what's important it's hard to choose the right technologies
- Work with relevant stake holders, make sure you understand their needs and expectations

Handling Heavy Metadata Usage

- Common issue in today's HPC environment, file system metadata performance holding back applications
  - Many programs creating tons of small files across the file system
  - Number of file opens and closes grows as well
  - Directories with numerous files in them effected by locking performance
- All flash at the metadata tier to improve performance (NVME especially performant)
- Distribute metadata across many servers (different NSD servers in Spectrum Scale, DNE for Lustre)

Handling Heavy Data Throughput

- Arises when multiple jobs/large jobs go through checkpoints, or when code actually has really good streaming I/O
- Leverage Burst Buffer technology to absorb the short term high throughput load
- Increase spindle count to improve sequential performance, possibly reduce individual drive capacity
  - Caveat lower density drives have worse sequential throughput, find the sweet spot in $/performance
- Balance server capability with spindle performance, some appliances and solutions oversubscribe the server or controller for density purposes

Acknowledgements

- Thanks to Pam Hill from NCAR for content assistance and laying the groundwork for this workshop
- Members of the SET group at NCSA for slide review
- Members of the steering committee for slide review

Questions