# The Storage Laundrette

**“Taking care of application developers’ dirty laundry…”**

**@TVBurns, CTO Media & Entertainment**

**HPA UK Tech Retreat July 13, 2017**

DELL EMC

# 15 minutes to tell you about 150 days of testing

- All-flash NAS challenges developer assumptions re: storage & networking
  - highlighting issues within I/O middleware layers
  - un-tuned network stack adds ~100 μsec per IOP
  - noticeable in all-flash clusters delivering 250K disk IOPs
- Application performance is the only thing M&E users care about
  - no standard way of measuring or automating test procedures to replicate a given user experience
- Fatter pipes don't help if (e.g.) you have limited workstation PCIe bus bandwidth
- Current OS middleware doesn't address latencies of both flash and object in the same stack
  - Object storage measured in millisec vs. all-flash NAS clusters in μsec
- Multiple tools needed for end-to-end analysis of high-throughput low-latency workflows
  - Separate analysis required for workflows using file-to-object gateways

# How do applications consume storage today?

- Global collaborative M&E {pipeline:VFX :: workflow:Post-production}
  - composed of “best-of-breed” apps in an operational fabric
  - storage architects addressing orchestration layer – “Storage defines workflow”
- Mix of disk IOPs, protocol IOPs and caching affects application performance
- End to end user application experience largely a product of latency
- Packet capture & analysis to uncover app developers’ ~~dirty laundry~~ architecture choices
  - Aging frameworks difficult to re-architect without user trauma
  - (e.g.) QuickTime 7 → QuickTime X

# Application development focus moves up-stack

- Storage has been slow to move up the value chain, now making up for lost time
  - Many storage systems have WAN acceleration and data migration tools
  - 3rd party WAN acceleration & Data movers preferred for vendor neutral pipelines
- Network protocols
  - SMB3.0 with multi-channel & failover
  - NFS v4 with failover
- “Ethernet has eaten the world” (with apologies to Andreessen)
- Trading latency for bandwidth with Software-defined Networking
- Scaling to global collaboration requires MAMs & Apps to factor in “Data Gravity”
  - Point solutions don’t scale all the way from Flash to Object

# Mac Pro 6,1 example

- # system_profiler SPHardwareDataType
  - Model Name: Mac Pro
  - Model Identifier: MacPro6,1
  - Processor Name: 6-Core Intel Xeon E5
  - Processor Speed: 3.5 GHz
  - Number of Processors: 1
  - Total Number of Cores: 6
  - L2 Cache (per Core): 256 KB
  - L3 Cache: 12 MB
  - Memory: 32 GB

# Mac Pro 6,1 iperf test shows PCIe bus limits

- # ./**iperf** -P4 -s &
- # ./**iperf** -c localhost
- ------------------------------------------------------------
- Client connecting to localhost, TCP port 5001
- TCP window size: 2.01 MByte (default)
- ------------------------------------------------------------
- [ 5] local 127.0.0.1 port 50298 connected with 127.0.0.1 port 5001
- [ 4] local 127.0.0.1 port 5001 connected with 127.0.0.1 port 50298
- [ ID] Interval Transfer Bandwidth
- [ 5] 0.0-10.0 sec 34.9 GBytes 30.0 Gbits/sec
- [ 4] 0.0-10.0 sec 34.9 GBytes 30.0 Gbits/sec

# More tools

- **strace** tells you how your application interacts with your operating system
  - -f (follow child processes)
  - -s (see everything)
- **sar** –b provides overall I/O activities
- **iotop** –o shows only processes of threads doing I/O
- **fatrace** Open source tool to trace file access events system-wide
- **wireshark** Open source packet capture & analysis tool
- Use filer statistics as much as possible
  - mixed workloads like 3D animation and render farm hitting the same cluster are tricky to analyze

# Learning from Big Data – “Data Gravity”

- “As the mass of your data increases, the number of services and applications attracted to that data also increases, in direct proportion to the mass of the data”
- Video is so massive you need to keep it in one place, and bring the applications to the data
  - operationally inefficient to do it the other way around
- Most valuable metric you have as a content creator is data usage analytics
  - organize your pipeline to cope with data gravity
  - what data needs to be operant in your hot pool
  - what data needs to be parked in your nearline pool
  - what can be put off in the cloud because you don't think it's ever coming back again

# A tale of three applications – DPX vs. Clip-based

- DPX workflow (“one-file-per-frame”)
  - OpenEXR workflow has more random I/O due to variable depth
- Clip-based workflow (“one file per shot/sequence/episode”)
  - Compression not a bottleneck as GPUs provide real-time de-Bayering
  - CPU decoding for common codecs
- Application cache fighting with storage cache – which becomes invalid first?
  - Phil Karlton: “There are only two hard problems in Computer Science: cache invalidation and naming things.”

# Divergent clip- and frame-based performance

- Application Suite “A”
  - Fantastic DPX performance, terrible OpenEXR performance
  - Testing revealed a bug in the OpenEXR routines, quickly fixed
- Application Suite “B”
  - Fantastic ProRes performance
  - Workarounds for MXF & DNXHD performance
  - Terrible DPX performance, OpenEXR not supported
- Application Suite “C”
  - Excellent DPX & OpenEXR performance
  - Excellent ProRes, MXF & DNXHD performance

# Run-time optimizations

- Some applications support an increase in the number of read threads
  - Tuning required so as not to compete with other processes that need to happen simultaneously
  - Usually a good idea for DPX workflows and a bad idea for clip-based workflows
- Filename Pre-fetch
  - OneFS allows sequential file sequences to be pre-fetched into cache
  - Allows RT playback of DPX, EXR, .ari files, etc.

# Software-defined Networking

- “Leaf-Spine” topology provides fully-meshed Ethernet fabric
- Scalability & agility overcomes inherent latency of TCP/IP protocols
- 80 low-cost switches provides (e.g.) the equivalent of 42,000 3G SDI ports
  - 16 spine and 64 leaf switches running Open OS’s
  - Larger than any existing SDI router
  - Inputs & outputs defined at run-time, not purchase time
- Open switch operating systems allow vendor-neutral meshes
  - Dell OS9
  - Cumulus Linux OS
  - Big Switch Light OS
  - IP Infusion OcNOS
  - Pluribus Netvisor OS


Thanks
Merci
Danke
谢谢
DELL EMC