IEEE MSST

Storage Infrastructure
Design and Management
at Scale

Small Files…too much of a good thing

Small Files / Objects

- Small files create lots of issues
  - Lots of meta data
  - Lots of inodes
  - Indexing causes issues
- Traditional
  - Hash of hashes directory structure
  - Still end up with ~20-30MM files per filesystem
  - Standard commands break things: ls, find, du, etc.

Small Files cont.

- Designing from scratch
  - Switch to an object oriented storage model
  - Example approaches
    - Load the objects as BLOBs into a database field
    - Customize the FS store the inode/extent ID in a DB
    - Pay someone else to do it: Atmos, DDN WOS etc.
  - If you don't need the metadata don't keep it
    - filename, permissions, atime, mtime, ctime, acls all create overhead

Storage at scale

• • •

Lots of Storage
Lots of IO
Lots of Storage Devices
Lots of Data

Small IO…and lots of it.

- Traditional OLTP workloads still exist
  - 95% Read-Miss
  - 80,000 IO/sec
  - 8K request size
  - Latency Requirement <2ms
- What to do…
  - Lots of queues
  - Mix of SSD and Spinning disk based on workload

Telecom Principles for Storage

- Latency
- Jitter
- We've heard data consistency...

performance consistency

The Blend…

Distributed Data Systems / HPCC and Traditional OLTP
Working together

The Blend

- There is a ton of data out there
  - The dataset for analytics is one thing, but what if the customer needs a subset of that too?
  - Running expensive analytics in real-time is way to expense
  - Blend it
- Use HPC to preprocess datasets then do final real-time checks before presentation
  - Analytics are usually batched
  - Customers load in real-time
  - Verify that the data is still valid before it's rendered

What's Next…no one thing

Disruptive technologies/approaches to enterprise arrays & storage transport

Flash

- Manufactures are still duking it out
  - MLC, SLC, post-Flash
- Still a “disk drive”
  - Moving to a cell approach
- Be careful
  - Transport, queues, mgmt hardware, data age

Transport / Convergence

- Is the mainframe back?
  - SAS
  - PCI-Express
- Running more
  - Virtualization platforms
  - Analytics