Storage & FS Challenges

Ric Wheeler
ric@emc.com

Good News, Bad News

- Big files & high bandwidth are the focus of a lot of work
  - When things work, typically we can extract more bandwidth out of a drive than most users need
- Small files, on the other hand, have
  - Poor utilized capacity
  - Degraded performance as object count rises
  - Strange performance bumps

Basic User Requirements

- Complete Set of Data Objects
  - Can you verify that you have all of the files/objects/blocks that I gave you?
- Individual data object integrity
  - Can you validate that the integrity of the objects(files/blocks) is still correct?
- Fully Utilized Capacity
  - How much of the capacity in my storage device can I use for my type of data?

Answering Set Completeness

- Verifying No Lost Files or Object Means
  - Must be able to iterate over all of the stored objects & validate
  - Can we cross check file names without duplicating each dirent or using something like an external DB?
- Challenges
    - 1TB drives can hold 250 Million billion 4k files!
- Key Components
  - Performance of readdir
  - Cross check of readdir against ???

Answering Object Integrity

- Verifying Object/File Integrity Means
  - Read and validate each block of data in the object?
  - Object or page or sector level checks?
- Challenges
  - Handling IO errors when touching every block
  - Where to store the signatures
- Key Components
  - Readdir, read (or read verify)
  - Surface integrity scans
  - Reverse mapping of bad sectors to files?

Answering Capacity Utilization

- Complete Utilization Requires
  - Minimal overhead for answering the first two requirements!
  - Efficient support for high object count – no per directory or file system arbitrary limits
    - Some reduction in performance can be tolerated at the extremes
- Challenges
  - Small file support
  - Repair, backup, enumeration

How the IO Subsystem Can Help

- Robust and Well Understood errors
  - Media errors are common
  - Even Commodity Drives Retry Reads
    - Report errors accurately and with minimal timeouts
  - When a drive is on its last legs, you might need to copy data out to a new disk
    - 30-60 seconds per bad sector is too long for big drives!
- Signing Blocks or Sector Level Block Guard
  - Can flag some errors
  - Not helpful for media level errors which require duplication of data through RAID and so on

Repair Oriented Design

- Minimize “Data Unavailability”
  - Keep in mind worst case fsck time
    - Do we need online fsck for really large devices?
  - fsck bailed – how long to mkfs & restore from a second device?
- Do real IO error handling testing
  - Test with bad drives
  - Test with SW fault injection
  - Test at scale and on aged file systems!

Performance Testing

- Life-time testing
  - Empty to full file system
  - Single and multi-threaded work loads
  - Bulk operations (untar, read all files, etc)
- Test under duress
  - How much performance do you lose during that new online fsck or RAID rebuild?