Improving Lustre® OST Performance with ClusterStor GridRAID

John Fragalla
Principal Architect
High Performance Computing
Legacy RAID 6 No Longer Sufficient

- 2013 – RAID 6 data protection challenges
  - Long rebuild time per 4TB drive
  - Impact to large systems
    - Multiple ongoing rebuilds across system
    - Diminished performance

- 2014 - 6TB and larger capacity drives
  - Excessive rebuild time using RAID 6
  - Unacceptable impact to large systems
  - Prolonged risk window

New Approach Required to Support Large Capacity Drives
ClusterStor GridRAID

- ClusterStor GridRAID restores full redundancy up to 400% faster
- Consolidates 4 to 1 reduction in Object Storage Targets

Mandatory to effectively implement high capacity drives and solutions
What is ClusterStor GridRAID?

GridRAID is ClusterStor’s Implementation of Parity Declustered RAID

“Parity Declustering is designed to balance cost against data reliability and performance during failure recovery. It improves on standard parity organizations by reducing the additional load on surviving disks during the reconstruction of a failed disk’s contents. And it yields higher user throughput during recovery, and/or shorter recovery time.”

Mark Holland, Garth A. Gibson (1992)
ClusterStor GridRAID - Introduction

Block Device: Logical Address Space

GridRAID Lives Here

RAID MATH

Storage Device Abstraction

Redundancy Encoding

Unit Mapping Strategy
ClusterStor GridRAID - Tile Definition

Data is organized into “tiles”:

- Stripes laid out in typewriter fashion

Smallest number of rows that contain an integral number of stripes

Stripes in this example are defined as RAID6 (3+2)
Notice, each remaining drive receives less recovery work. Therefore, each disk has more bandwidth for client IO. Also, when extended across the full disk, all drives participate in the recovery increasing the available source bandwidth.

Notice, in a single tile several disks are not utilized. For this reason a unique permutation is used for each subsequent tile.
ClusterStor GridRAID

**Distributed Sparing**

Distributed Sparing:

Augment tile matrix with *spare columns*:

![Image of matrix]

Permute all the columns:

![Image of permuted matrix]
ClusterStor GridRAID

Reconstruction – Recovery Phase 1

Data reconstruction to the distributed spare

- Begins automatically if a distributed spare is available
- Source data distributed across all remaining drives
  - Default recovery bandwidth of 50 MB/sec / drive
  - \( 40 \times 50 \text{MB/s} = 2 \text{GB/sec} / 9 = 222 \text{MB/sec} \) recovery rate
  - \( 4 \text{TB} / 222 \text{MB/sec} = \sim 5 \text{ hours} \)

- Reconstructed data written to all surviving drives
- Full redundancy restored in \( \sim 5.5 \text{ hours} \) with default setting
- \(< \frac{1}{4} \text{ time of RAID 6} \)
GridRAID Improves Lustre® Performance

- Traditional RAID 6 with 50MB/s rebuild rate of a 4TB Drive take approximately 22 Hours to complete Rebuild
  - Translates: A OST performance will degrade by 45% of peak performance under heavy I/O load during rebuild for 22 Hours
    - Note 4 OSTs per OSS with RAID 6
- GridRaid using the same 50MB/s, but now it’s a reconstruction rate, of that same 4TB Drive takes 5.5 Hours to completely recover
  - Translates: The performance impact to an OST under heavy I/O load during reconstruction is the same 45% degradation, but only for 5.5 hours instead of 22 hours.
    - Note 1 OSTs per OSS with GridRaid
Slowest OST impacts a Lustre® FS

• When MPI applications using a Lustre FS, overall I/O is impacted by the slowest running OST during run-time
• GridRaid, 4x less OSTs per OSS, performance impact for application runtime lasts ¼ as long compared to RAID 6
  • RAID 6 and 4x more OSTs at a smaller size per OST, application performance impact will exist for 22 hours, and performance impact during rebuild is 4x longer than GridRaid
• Applications performance impact using GridRaid for each Lustre OST during reconstruction will only exist ¼ of the time compared to traditional RAID 6
ClustreStor Tunable Settings

• Customers may choose to change the default reconstruction rate that favors more front-end I/O performance, or vice versa
  • Same command can be used for traditional RAID 6 rebuild rate as well
• Default value is 50MB/s and higher in the rare case of a double disk failure per OSS controller
  • Default values can be reduced or increased depending on application performance requirements
• Settings can be dynamically set globally for all OSTs or per OSS controller to optimize application performance during run-time
Benefits of ClusterStor GridRAID for Lustre®

- **Improved MTTR**
  - Reconstruction >4X faster than RAID 6 @ 50MB/s per drive

- **Distributed Recovery Workload**
  - Less disruption to client IO
  - Overcomes the single drive bandwidth bottlenecks

- **Reduced Lustre OST Configuration**
  - Reduces OST count by a factor of 4

- **4 Distributed Spare Volumes per SSU**
  - RAID 6 provides 2 Global Hot Spares per SSU
Thank You

John_Fragalla@xyratex.com