Sonexion GridRAID Characteristics

Mark Swan
Performance Team
Cray Inc.
Saint Paul, Minnesota, USA
mswan@cray.com

***Abstract*—This paper will present performance characteristics of the Sonexion declustered parity product, known as GridRAID. Comparisons to MDRAID will be made during the normal operational state of Object Storage Targets (OSTs) as well as the degraded states (reconstruct and rebalance).**

Keywords-Sonexion; GridRAID; declustered parity

I. Introduction

The Cray Sonexion GridRAID product is an implementation of "declustered parity" [1][2][3] by Seagate/Xyratex. This product has several technically intriguing features including reduced time spent in degraded modes, lower OST counts per file system, decreased number of active streams required to reach peak performance, and similar per-building-block performance characteristics to its MDRAID counterpart.

II. Cray Sonexion Overview

The Cray Sonexion 1600 is composed of a single Metadata Management Unit (MMU) and one or more Scalable Storage Units (SSU).

The MMU consists of four servers and either a 2U24 or 5U84 drive enclosure. The four servers are two cluster management servers and two file system metadata servers (i.e. the MGS and MDS). The SSU consists of two Object Storage Servers (OSS), each with 32 GiB of memory, and a 5U84 drive enclosure.

In the MDRAID configuration, of the 84 drives in the 5U84 enclosure, there are two SSDs, 80 spinning disks arranged into eight Object Storage Targets (OSTs) which are RAID 6 8+2 arrays, and two spinning disks as global hot spares. Each OSS has primary responsibility for four OSTs.

In the GridRAID configuration, of the 84 drives in the 5U84 enclosure, there are two SSDs and 82 spinning disks arranged into two OSTs. Each OSS has primary responsibility for one OST.

As a general guideline, the performance of a single SSU is said to be 5 GB/s sustained and 6 GB/s peak. The SSU is the building block of the file system.

III. RAID 6 8+2 Background

In the Sonexion MDRAID configuration, there are ten physical drives associated each OST. The generic notation for this type of configuration is "P drive (N + K)", where P is the number of disks in the array, N is the number of data blocks per stripe, and K is the number of parity blocks per stripe. Therefore, the Sonexion MDRAID is "10 drive (8+2)". The data and parity blocks are not strictly confined to any one of the ten drives but, rather, the data and parity are distributed among the ten drives. The key feature of RAID 6 is that writing and reading of data can continue even when two drives fail concurrently.

Rebuilding a RAID 6 device after the loss of a drive requires regeneration of the data for the missing drive and completely writing the contents of the newly replaced drive. As such, the time to complete this rebuild phase is limited by the speed of a single drive. Given a nominal drive write rate of 50 MB/s, writing an entire 1 TB drive can take approximately 5.6 hours (1,000,000 / 50 / 3600 = 5.55). Consequently, an OST created from ten 4 TB drives would take approximately 22.2 hours to rebuild. During this time, I/O operations are degraded. Furthermore, the loss of a second drive causes the OST to enter what is called "critical mode" since there is no parity protection when two drives are lost.

IV. GridRAID OST Creation

As mentioned previously, in the Sonexion MDRAID product, there are eight OSTs in an SSU and each OSS in the SSU has primary responsibility for four OSTs. In the GridRAID product, an OST is created using 41 of the 82 spinning disks in the standard 5U84 disk enclosure. Therefore, there are two OSTs per SSU and each OSS has primary responsibility for one OST. The generic notation for the GridRAID device is "P drive (N + K + A)", where P is the number of disks in the array, N is the number of data blocks in a stripe, K is the number of parity blocks in the stripe, and A is the number of distributed spare blocks per tile. The specific notation for GridRAID, then, is "41 drive (8+2+2)". Since the RAID 6 stripes are "logical", there are no strict boundaries on the physical devices; all data blocks, parity blocks, and spare blocks are distributed among the physical blocks of the 41 drives. The actual algorithms used to implement the layout of data blocks is considered proprietary information by Seagate/Xyratex and will not be presented in this paper.

V. PERFORMANCE CHARACTERISTICS

The version of NEO software that supports GridRAID supports the same hooks for viewing and changing the OST allocation position pointer as NEO 1.2.3 [4]. Fig. 1 shows the results of running the obdfilter-survey tool from the fast zones to the slow zones of a GridRAID OST.

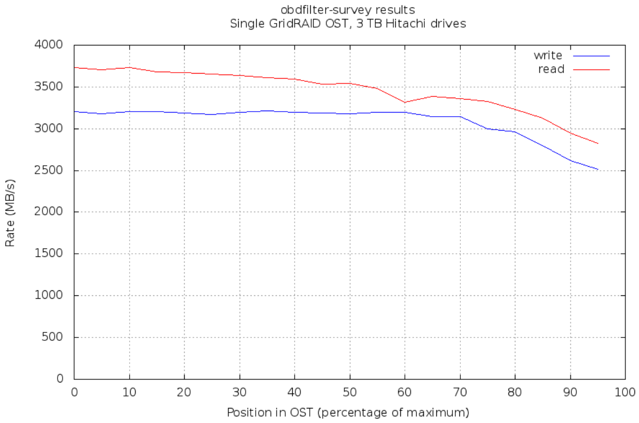

Fig. 1, GridRAID OST speed zone differences

Considering the fact that there are four MDRAID OSTs per OSS and only a single GridRAID OST per OSS, we might wonder if it takes four times as many active streams for the performance of a GridRAID OST to match that of four MDRAID OSTs. Fig. 2 and Fig. 3 show the IOR performance of a single GridRAID OST with 32 MiB transfers.

Fig. 4 and Fig. 5 show comparisons of write and read performance of an MDRAID OSS versus a GridRAID OSS. As we can see, write performance of GridRAID ramps up to the peak with fewer active streams than MDRAID. GridRAID buffered reads take a few more active streams than MDRAID to ramp to the peak. GridRAID direct I/O reads are, overall, not as good as MDRAID but this can be improved with some server-side tuning.

Lustre OSTs can be configured to pre-allocate differently sized chunks of data as files are created [4]. This pre-allocation size can be used to create more contiguous data within files as they are written to the OST. In turn, this larger amount of contiguous data can greatly improve read performance but, sometimes, at the cost of write performance. Fig. 6 and Fig. 7 show the effects of modifying the OST pre-allocation size in order to improve buffered read rates. As we can see, the pre-allocation size of 8 MiB maximizes buffered read rates and causes only a slight buffered write performance decrease.

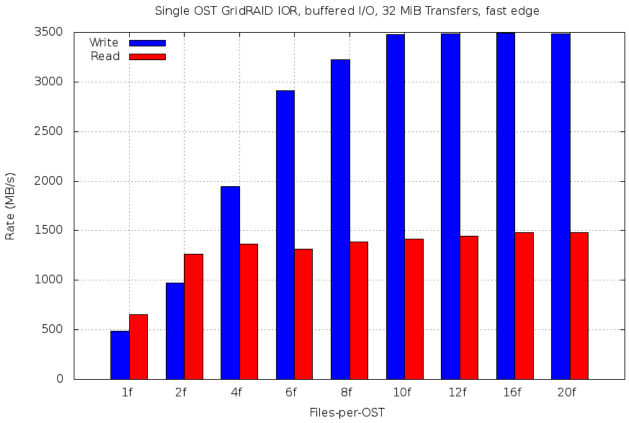

Fig. 2, Performance of buffered I/O, 32m transfers

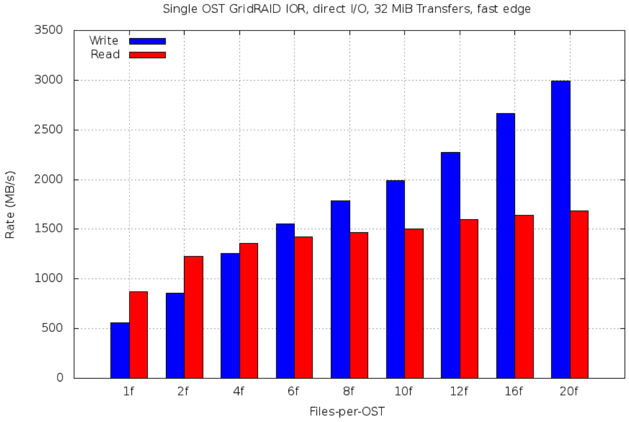

Fig. 3, Performance of direct I/O, 32m transfers

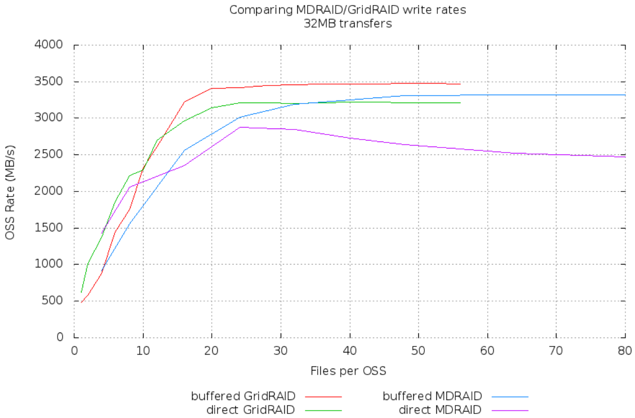

Fig. 4, Comparison of MDRAID and GridRAID write performance, 32m transfers

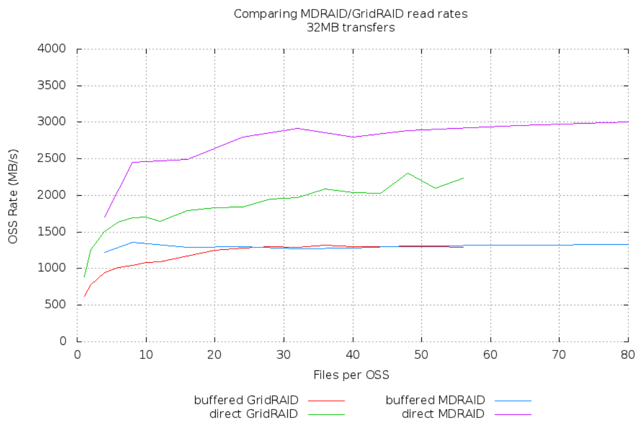

Fig. 5, Comparison of MDRAID and GridRAID read performance, 32m transfers

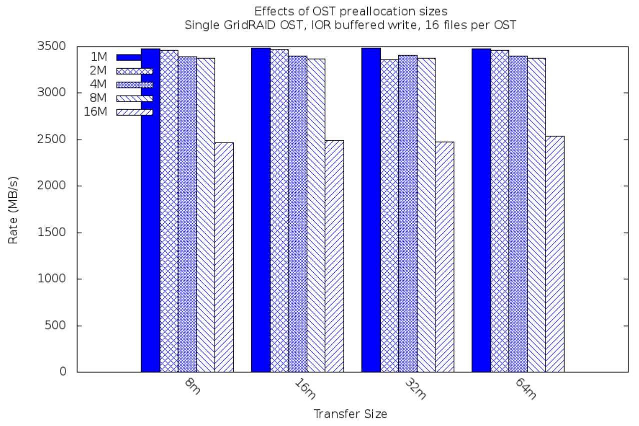

Fig. 6, OST pre-allocation effects, buffered write

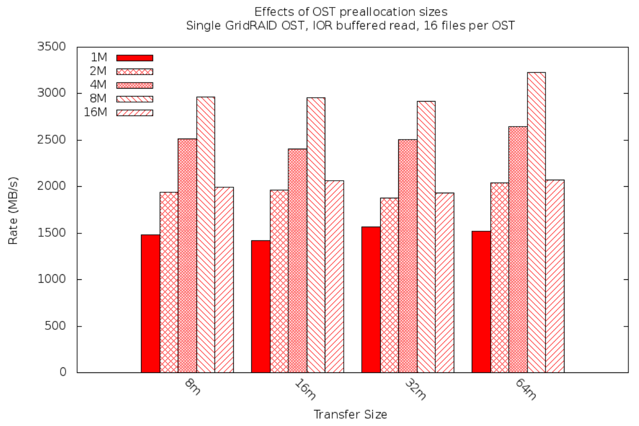

Fig. 7, OST pre-allocation effects, buffered read

VI. GridRAID Degraded Modes

There are two degraded modes of operation for a GridRAID device; reconstruct and rebalance. Like the Sonexion MDRAID product, the Sonexion GridRAID product enables customization of the bandwidth to be allocated to the recovery phases when Lustre client I/O is occurring. In the absence of I/O from Lustre clients, however, the recovery phases will consume the maximum amount of bandwidth available.

A. Reconstruct

When a single drive fails in the 41-drive OST, the data from the failed drive must be reconstructed and distributed to the spare blocks that exist on the remaining 40 drives. There is no single drive that needs to be written since all data for the logical RAID 6 stripes is distributed across all drives. Therefore, this reconstruct phase is not limited by the write rate of a single drive. Instead, the write bandwidth of the remaining 40 drives is available.

Given the same nominal 50 MB/s drive write rate used previously, the data placement is such that the OSS now has 40 * 50 MB/s / 9 (222 MB/s) of bandwidth available to it for reconstructing the OST. For an OST based on 1 TB drives, this is approximately 1.25 hours. For an OST based on 4 TB drives, the time from losing a drive to having a fully functional and redundant OST is a little more than 5 hours. Compare this to the 22.2 hours for the MDRAID product.

Fig. 8 shows the bandwidth over time as an idle GridRAID OST (based on 3 TB drives) goes through the reconstruct phase. Total time is approximately 2.5 hours. The reader might recognize the performance curve as being the same as the performance differences from the fast zones to the slow zones of the disks. According to the data found in “/proc/mdstat” of the OSS, the bandwidth ranged from approximately 98 MB/s per drive at the beginning to approximately 58 MB/s per drive at the end.

Fig. 9 shows two OSTs going through the reconstruct phase at the same time. OST 1 (black) was idle while OST 0 (green) had a series of IOR jobs run against it. The minimum bandwidth allocated to performing the reconstruct was set to 50 MB/s per disk.

Fig. 10 superimposes LMT data from the series of IOR jobs along with the reconstruct rates and times. As IOR begins to move data, the rate of reconstruct for OST 0 decreases to the configured value of 50 MB/s per disk.

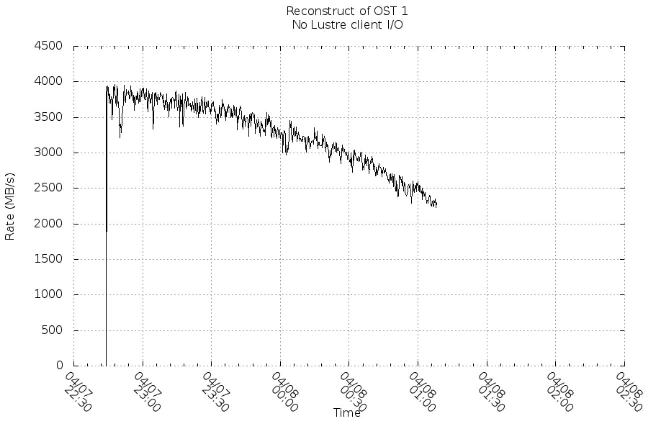

Fig. 8, Reconstruct of idle GridRAID OST

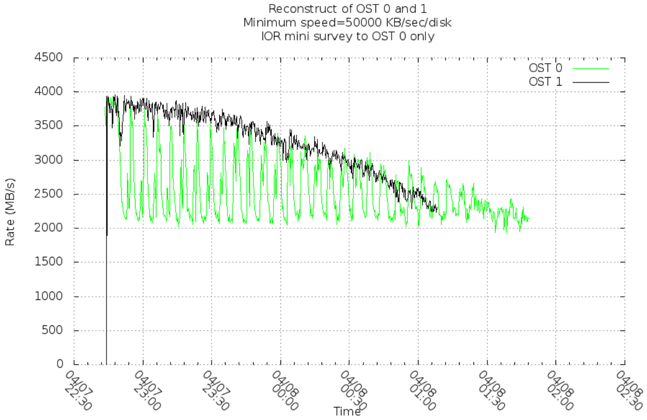

Fig. 9, Reconstruct of two GridRAID OSTs

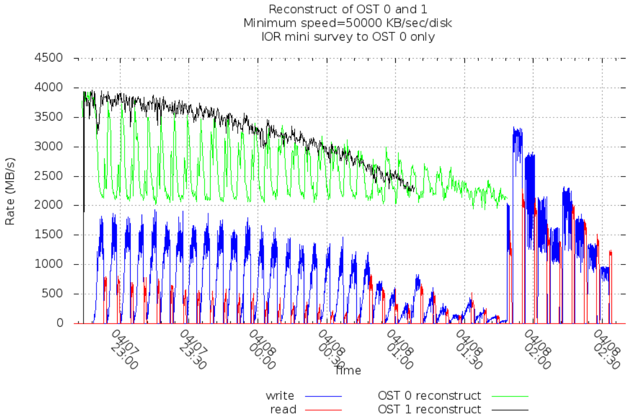

Fig. 10, Superimposing LMT data with reconstruct

B. *Rebalance*

The rebalance phase occurs when a replacement drive is introduced into an OST. The term "rebalance" implies that all the data for the OST (that was on 40 drives after the reconstruct) needs to be rebalanced among the 41 drives that now exist. Similar to the recovery phase of a MDRAID device, this rebalance phase must write the complete contents of the new drive and, as such, is limited by the bandwidth of the device. However, during this period, the OST is completely functional and redundant due to the reconstruct phase that has already completed.

Fig. 11 shows the bandwidth over time as an idle GridRAID OST (based on 3 TB drives) goes through the rebalance phase. Total time is approximately 6.5 hours.

Fig. 12 shows two OSTs going through the rebalance phase at the same time. OST 1 (black) was idle while OST 0 (green) had a series of IOR jobs run against it. The minimum bandwidth allocated to performing the rebalance was set to 50 MB/s per disk.

Fig. 13 superimposes LMT data from the series of IOR jobs along with the rebalance rates and times. As IOR begins to move data, the rate of rebalance for OST 0 decreases to the configured value of 50 MB/s per disk.

Finally, Fig. 14 zooms in on the superimposed LMT data to show how the rebalance bandwidth of OST 0 changes as IOR moves data. It is interesting to note the delay as the rebalance bandwidth attempts to increase as IOR activity decreases and then decreases as IOR activity increases.

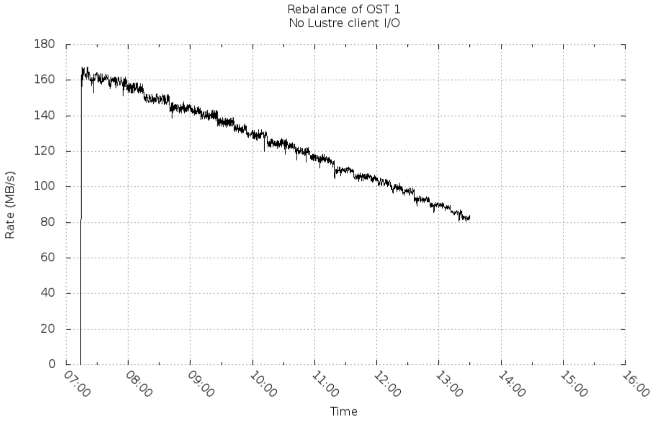

Fig. 11, Rebalance of idle GridRAID OST

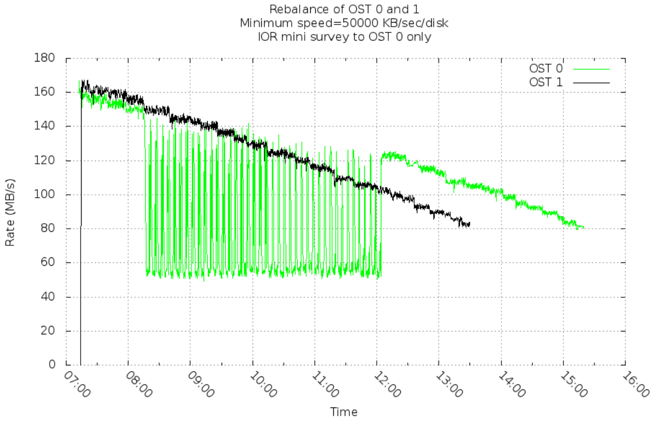

Fig. 12, Rebalance of two GridRAID OSTs

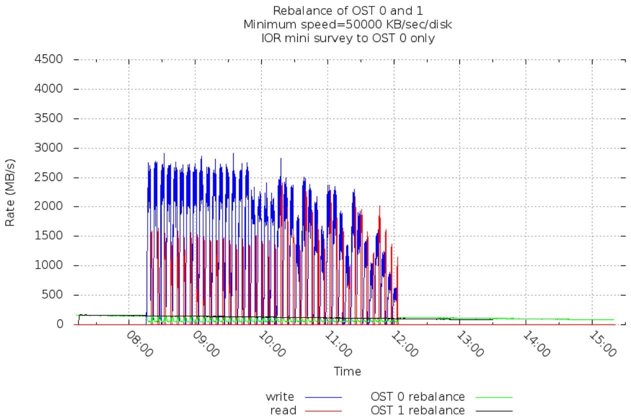

Fig. 13, Superimposing LMT data with rebalance

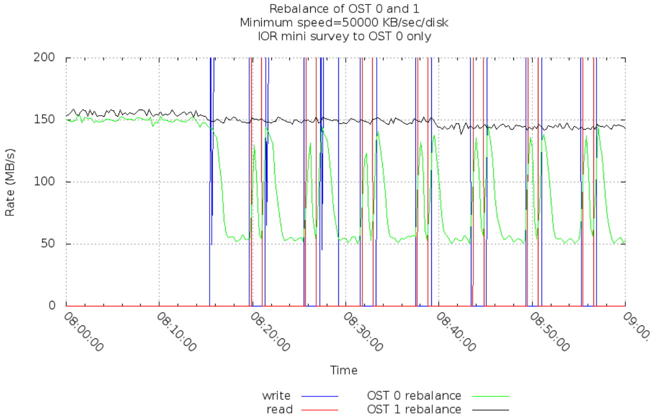

Fig. 14, Superimposing LMT data with rebalance, zoomed in to show fluctuation

C. *Single Drive Failure*

With the MDRAID implementation, losing a single drive causes one of the OSTs to be in a degraded state. That is 25% of the OSTs of that particular OSS. Due to the distributed nature of data in a GridRAID OST, losing a single drive causes 25% of the OST to be in a degraded mode and, since there is only a single GridRAID OST per OSS, this is equivalent to MDRAID. However, as we have previously pointed out, the time it takes to recover from a single drive loss is four times greater with MDRAID than with GridRAID.

D. *Double Drive Failure*

When a MDRAID OST loses a second drive, that entire OST enters into [what is referred to as] “critical state”. The OST is critical because there is no longer any parity protection. Therefore, 25% of the user data for that OSS is at risk. Losing a third drive will cause complete loss of data for that OST. The MDRAID OST remains in critical state for as long as it takes to rebuild the data on a replacement disk (i.e. limited by the speed of a single drive).

In the [proprietary] GridRAID implementation, a second drive loss only puts approximately 5.5% of the OST into a critical state and, therefore, 5.5% of the data for that OSS. The OST will only remain in critical state as long as the reconstruct phase takes which, again, is one-fourth of the time for MDRAID recovery.

E. *Periodic RAID check*

While the periodic RAID check is not strictly defined as a degraded mode of operation, it can have a significant impact on Lustre client performance. The RAID check is an integrity check performed by the OSS on a scheduled basis or can be invoked manually.

Fig. 15 shows two OSTs going through the RAID check phase at the same time. OST 1 (black) was idle while OST 0 (green) had a series of IOR jobs run against it. The minimum bandwidth allocated to performing the RAID check was set to 10 MB/s per disk. The data for OST 0 begins when the RAID check was 26% complete and the data for OST 1 begins when the RAID check was 46% complete.

Fig. 16 superimposes LMT data from the series of IOR jobs along with the RAID check rates and times. As IOR begins to move data, the rate of RAID check for OST 0 decreases to the configured value of 10 MB/s per disk

VII. SUMMARY

The Cray Sonexion GridRAID product, due to the distributed nature of data and parity among 41 drives, offers several benefits over the traditional MDRAID implementation. The most important of these benefits appears to be the lower time spent in degraded and critical states. Performance characteristics of the GridRAID product are on par with, and sometimes surpasses those of, MDRAID.

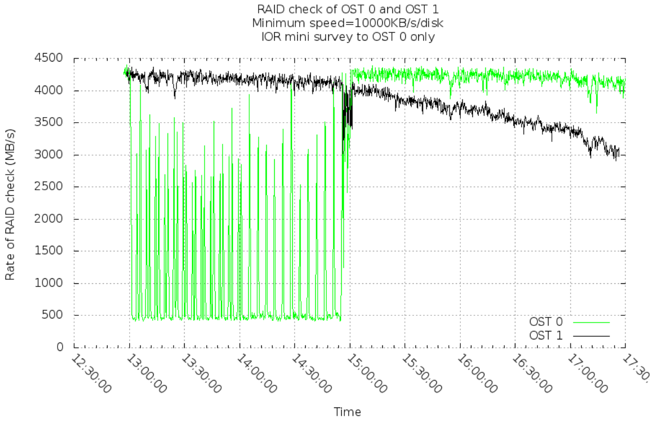

Fig. 15, RAID check of two GridRAID OSTs

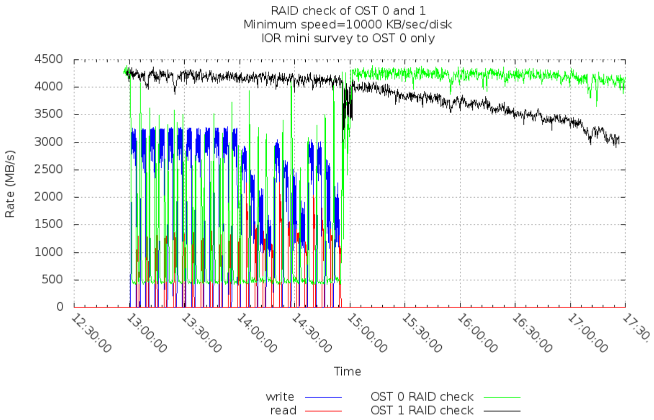

Fig. 16, Superimposing LMT data with RAID check

REFERENCES

[1] Holland, Mark "On-Line Data Reconstruction In Redundant Disk Arrays" CMU 1994.

[2] Merchant, Arif, and Philip S. Yu. "Analytic modeling of clustered RAID with mapping based on nearly random permutation." Computers, IEEE Transactions on 45.3 (1996): 367-373.

[3] Dau, Son Hoang, et al. "Parity Declustering for Fault-Tolerant Storage Systems via t-designs." CoRR:1209.6152 (2012)

[4] Swan, M., “Tuning and Analyzing Sonexion Performance” CUG 2014.