Evaluating RAID in the Real World

Ian Bird, Rita Chambers, Mark Davis,
Andy Kowalski, Bob Lukens, Sandy Philpott,
Roy Whitney
SURA/Jefferson Lab
12000 Jefferson Avenue,
Newport News, VA 23606
{igb,chambers,davis,kowalski,rlukens,
philpott,whitney}@jlab.org
Tel +1-757-269-7514
Fax +1-757-269-7053

Abstract: High energy nuclear physics experiments at the Thomas Jefferson National Accelerator Facility ("Jefferson Lab") will have a data collection rate of 10 MB/second, generating 1 Terabyte (TB) of raw data per day of accelerator running, and a similar amount after processing. The requirement for on-line disk storage for raw and reduced data sets will exceed 1 TB during 1998. This paper discusses the on-line storage strategy that provides both high performance as well as high capacity, and focuses on the in-house evaluation of RAID (Redundant Arrays of Independent Disks) systems to fulfill the needs of both data acquisition and analysis.

1 Introduction

The Thomas Jefferson National Accelerator Facility ("Jefferson Lab") operates a 4 GeV continuous wave electron beam accelerator for nuclear physics research for the U.S. Department of Energy. The largest experiments will generate close to 1 TB of data per day, for some 120-150 days of accelerator running per year. In addition, the reconstruction of this data will yield a similar amount of processed data. Both the raw and analyzed data sets will be stored in an STK silo, using RedWood tape cartridges and drives. The data is collected on data acquisition computers close to the experiments and copied to the Computer Center and into the tape silo over a 1-km long dedicated fibre-channel connection. Once the data is in the silo, it may be copied back out for processing on a "farm" of Unix workstations. Results of this analysis are also stored in the silo. During the last year the Computer Center has gone through 2 major RAID procurements for storage to support different parts of the data path – areas for fast staging of files between tape and the farm, and for large analysis work areas. As part of those procurements we asked the competing vendors to bring to the lab examples of their systems for demonstration and testing. In this paper we present the process we went through to develop the tests that allowed us to determine the behavior of those systems in our environment.

1.1 Why RAID?

The sizes of the data sets that we deal with are largely driven by the data rate. A single hour of running of the largest experiment generates around 50 GB of data. Since there are several things that impose a 2 GB file limit (32 bit OS, tape storage manager software), the current processing is restricted to files of this size. The design goals, however, anticipate the requirement to handle significantly larger file sizes. The RedWood tape cartridges have a capacity of 50 GB and the drives can transfer data at a rate in excess of 10 MB/s. In order to optimize the use of the drives and to allow the tapes to stream, the ideal access method is to stage data from tape to a staging disk and then to transfer from that staging area to the destination. The data processing farm will eventually approach 50-100 CPUs

running in coarse parallelism accessing the data and storing the results via the staging areas. Thus the staging device needs to be capable of reading or writing at 10 MB/s with simultaneous remote accesses (which are at transfer rates limited by the network), so that we require not only a high data stream performance but also a high aggregate performance. In addition we require ideally some 150 GB of staging area per tape drive.

The other main use for high performance redundant storage is in the need of the experiments to have available on-line large data samples for analysis, visualization and algorithm development. Typically a single analysis (of which there may be many simultaneously) will demand some 50-100 GB of data to be on-line and randomly accessible from both the batch processing farms as well as central analytical systems.

In both cases, the need is for both performance and large data set sizes. Only RAID systems provide these capabilities.

1.2 Procurement Process

During the last year, we have undertaken two major procurements. The first was for host attached RAID for use primarily as the fast tape staging space, but also to provide an initial implementation of some of the work areas. Here, performance - both throughput and aggregate rate, was the critical factor. A certain level of configurability was also desirable.

The second procurement was specifically for network attached (NFS) RAID for use as work areas for large data sets. In this case management and configurability of the space was the key. These areas need multiple, reliable, network accesses as well as the ability for the space to be managed with group and directory quotas. Good performance was also an issue.

In addition to standard product research both before, and as part of, the formal procurements, the Jefferson Lab Computer Center chose to include on-site benchmarking as a requirement of the solicitation process. Vendors choosing to participate were required to bring to the laboratory a system closely configured to the system(s) offered in their proposals. Vendors were given advance knowledge of the tests to be performed and the local system and network configuration. Encouraged to pre-configure their RAID systems and to arrive early for system uncrating and installation, all vendors were able to complete the locally administered test suite in well under 4 hours, and generally used the remainder of their scheduled half-day time slot to demonstrate other product capabilities to site staff.

The tests were run in advance by the Computer Center staff on existing hardware in order to understand the infrastructure limitations as well as to determine the time constraints of the tests, calculate baseline results, and ensure that the test results were meaningful.

1.3 Importance of Testing

A relatively large number of manufacturers and integrators now offer RAID products that support storage solutions from the low-end (20 GB) up to multi-Terabyte range, provide a variety of combinations of hardware and software-based RAID architectures, and offer a wide spectrum of native redundancies. The challenge to the consumer in selecting the products best suited in terms of performance, capability, and cost for their specific application is substantial. The market is characterized by product announcements with the next generation always on the horizon, pricing models on a steady downward slope, and vendor capabilities that can in fact vary significantly. A complete evaluation will include

information from vendor-based and trade journal product comparisons as well as recommendations of professional counterparts. Each of these sources has limitations that could be critical particularly given the significant dollar investment (*i.e.*, even at 50 cents/MB, a TB costs $500 K!) and vital role the equipment plays in the business of the institution. Vendor product specifications and technical white papers provide only the vendor's eye view of their product line. Standardized benchmark suites (SPEC LADDIS, etc.) report what in fact are vendor-generated results of "standardized" tests, and allow a test environment and configuration which can vary dramatically from the consumer's intended application. Neither consultation with industry counterparts nor examination of non-biased commercial product evaluations is a guarantee that the equipment will in fact perform suitably in your specific environment and application. On-site testing, with real data and an environment customized to mirror the actual production application, whenever feasible is the ideal culmination of a product evaluation process.

The challenge for the RAID procurements described in this paper, beyond the obvious logistics of performing in-house testing of multiple vendor boxes, was to design tests that could be run in reasonable time frames on existing non-production equipment (in a relatively small computing center), and that in fact provided a realistic picture of how each box would perform in the lab's environment as well as demonstrate the relative capabilities of multiple vendor offerings. While it was important to a fair evaluation process to provide a determinant testing environment (isolated from network broadcasts, for instance), it was equally important to determine how the box would perform in the real world, specifically "our" real world. The tests needed to test for the "right" thing -- would the test in fact be limited not by the vendor's equipment but by the network itself or by the receiving batch node? Would the tests in fact show no measurable differences between the vendor systems or would the differences determined be invalid indicators of the viability of the proposed solution?

2 Analyzing the Data Path and Developing the Tests

The tests that were used in the evaluation were, in the end, relatively straightforward. However, there are a variety of factors that need to be considered in order to arrive at a series of tests that not only demonstrate that the system has the desired capabilities, but are also feasible in a limited time. In the following we give an outline of what those considerations are.

The first step in the analysis is to really understand the data path. This may not be so obvious, especially if the system is new and this analysis is actually part of the initial design process. For example, at first glance the cost/performance ratio for a farm of "pizza box" processors seems to be far better than that of a few large SMP systems. However, taking into consideration the actual data rates and consequent I/O requirements, and then taking into account networking costs, *i.e.*, considering the system as a whole, is that solution still the most cost effective? Consider also, any hidden assumptions. For example, at Jefferson Lab it had long been assumed that ATM would be the only way to deal with the high bandwidth and throughput requirements. However, it became clear that it was much simpler and cheaper to use switched Fast Ethernet, and that the necessary aggregate performance could be easily achieved.

A high level schematic of our actual data path is shown in Figure 1.

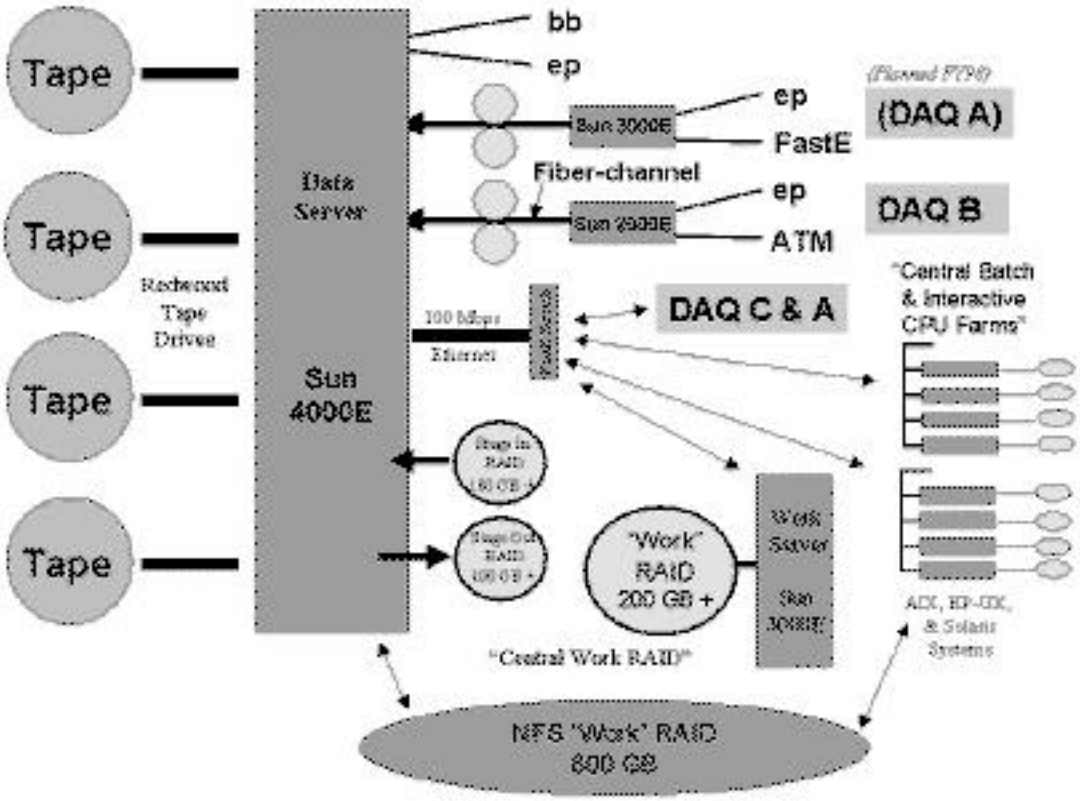

Figure 1. Data Path

Data arrives from the data acquisition systems of the experiments either directly from staging disks connected over 100MB/s Ethernet, or staged over Fiber-Channel and copied into the tape silo via a Sun E4000 data server. The batch processing nodes are dual processor UNIX workstations, with locally attached disks for local processing. This was determined to be the most cost-effective solution. The RAID staging disks attached to the E4000 are used to stage data both into and out of the tape silo. It is these staging areas that were the primary focus of the first procurement. For these areas the requirement on the RAID system was that it should accept data at 10MB/s – matching the capability of the RedWood drives, whilst simultaneously allowing access to and from the farm processing machines at network speeds. Note here, that all these accesses, both to the silo and over the network, are all going through the data server machine. This was a potential bottleneck, although that machine has multiple network and I/O interfaces.

The second procurement was for an NFS RAID server to provide data areas that will be used to hold intermediate processing results from the farm nodes allowing interactive analysis of ongoing processing, as well as for areas to hold large data samples for later analysis by multiple researchers. The requirements for these areas were somewhat different from the staging space. Rather than total throughput, the main demand was that the system be able to provide many simultaneous network accesses with a performance that should be limited by the networks rather than by the RAID system itself.

In order that the tests measure the performance of the RAID systems and not simply measure a bottleneck like network performance, some care must be taken with the design. Some baseline measurements are essential. Such measurements determine exactly where

the bottlenecks are – are the I/O adapters, the remote disk, or the network itself the limiting factors? They also give a base measurement with which to compare the test performance of the new systems. In our case these preliminary tests were made using groups of single disks running software RAID (Solaris DiskSuite) to ensure that the disk performance was really the limiting factor rather than one of the other limits. In that case there would be no reason to use expensive hardware RAID. Furthermore, some tests serve simply to qualify the RAID system – does the equipment pass the test or not? In our case, could the RAID match the 10 MB/s transfer rate of the RedWood tape transport? Would the equipment match or improve on the throughput achieved by the software RAID at each point in the data handling – from DAQ to RAID staging (via Fibre-Channel as well as network), to/from the RedWood tape transports, and collection to/from a network-accessed batch processor? Other tests can be used to discriminate between systems. Tests of transfer rates between memory and RAID or tape, and simultaneous transfers to multiple remote machines will generally provide such data.

In our setup, the parts of the data path that provide real tests were between the RedWoods and the RAID, from the RAID to a remote networked host, and from the RAID to several remote hosts. It is important also to test the data transfer in both directions as the reading and writing performances of the RAID systems are potentially very different. The tests should also be under controlled conditions. This is particularly important in terms of the load on the data server host, the remote hosts, and the network traffic. All the tests that were performed with the RAID systems were with unloaded hosts and quiet networks. However, as part of the test design we also made comparisons between loaded and quiet systems and with real network traffic.

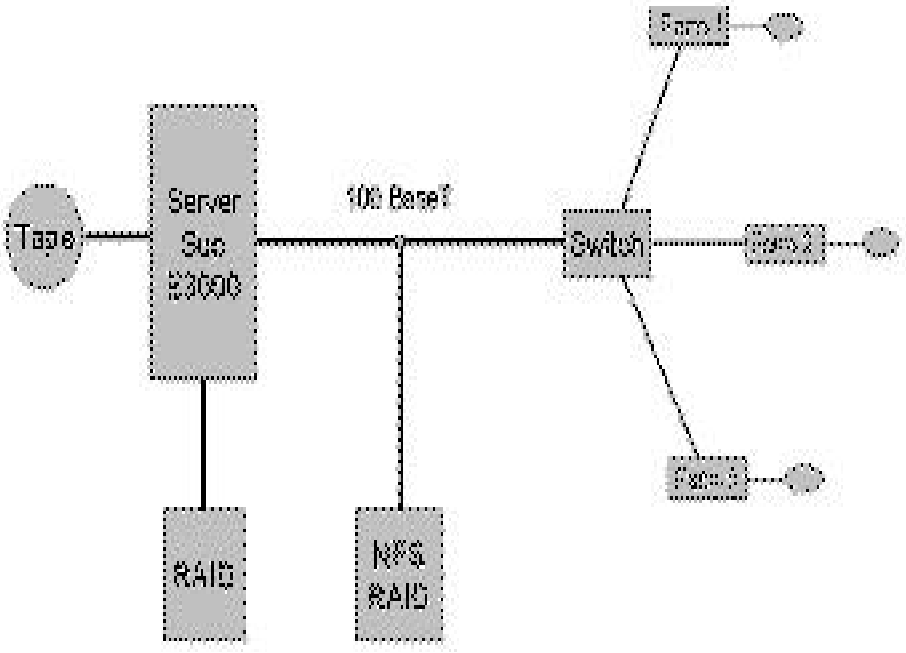

Figure 2. Test setup

For the tests themselves it is essential to limit them to testing what should be tested and not try to do too much. For example, even though in our production environment we will always be handling 2 GB files, we soon realized that for most tests 1 GB files were sufficient, and for other tests involving network transfers 500 MB was enough. Similarly, tests involving several remote hosts give as much information with 3 simultaneous transfers as with 10, since at some point something other than the RAID device becomes saturated. Most of these factors can be estimated ahead of time by running the proposed tests on existing equipment. Once the system has reached an equilibrium state no more information can be extracted – those are the data rates and there is little reason to run the test for a further 20 minutes.

The test programs were Perl scripts which facilitated changing the source on the fly and allowed its execution on multiple platforms without recompiling. For the read tests, the data was transferred to a 256 KB memory buffer and then discarded. For write tests, the same 256 KB buffer was written repeatedly out until the desired file size was reached. The data used was a sample of real physics data. It was realized early on that system generated pseudo-data could not be used as that data was easily compressible, gave very optimistic test results and did not represent the real situation at all. The same script was used for the tape to memory, memory to tape, and tape to disk tests. The network transfers were done with a version of rcp, NFS being far too slow for our needs for the fast staging space. The test system setup (See Figure 2) used a Sun E3000 as the test server for the direct attached RAID, with connections to both the STK RedWood drive as well as a dedicated 100BaseT Ethernet network. Three IBM RS6000 systems were used as the batch farm nodes. A similar environment was used with the Network Attached RAID tests (See Figure 2). You can view a copy of the Perl script we used and descriptions of the specific tests used in the two procurements at the following URL: http://www.jlab.org/ccc/gsfc/raidtest.html.

The test results were all logged automatically to create a permanent record of the test. The individual test results together with an evaluation sheet were returned to the vendor at the conclusion of the test. As part of the evaluation, the staff also considered demonstrated administrative functionality, as well as how easy it was for the vendors to install their systems in our environment, and how long it took to format and build the file systems.

There were some practical logistical considerations. We provided the vendors with the tests ahead of time, so that they could consider how best to configure their systems. However, we were very careful to ensure that the tests were fair, and that comparisons were only made of similar capabilities. The vendors were given a setup time half a day in advance of the testing, and a total test period of 4 hours. In general the tests were run well within that time and they were able to demonstrate other features and enhancements of the systems.

3 Results

The performance test results were highly effective discriminators in the solicitation process we describe for RAID systems and in fact were useful to both the vendors as well as Jefferson Lab. The testing process in the first procurement was completed in advance of the final bid submission date, and based on their results, several vendors opted not to continue in the proposal process based on results that clearly did not meet the minimum specifications of the procurement. On our part, the computing staff who assisted with the setup and configuration of the systems as well as the actual running of the tests, gained an in-depth knowledge of the architecture and capabilities of the systems being proposed. This perspective proved invaluable in the later evaluation of the written proposals.

The results of the tests are shown in Figures 3 and 4. The host-attached RAID tests (Figure 3) were: 1) copying a 1 GB file from memory to the RAID; 2) copying 1 GB from RAID to memory; 3) three simultaneous copies from memory to disk – the results shown are the average rate per process; 4) three simultaneous copies from disk to memory; 5) copy a 1 GB file from disk to tape; 6) copy from tape to disk; 7) copy a 1 GB file from memory to disk and simultaneously copy a file from disk to tape; 8) simultaneous copies from disk to memory and from tape to disk; 9) copy a file from RAID to the local disk on a remote machine and simultaneously copy a file from tape to RAID.

Figure 4 shows the NFS RAID tests: 1) writing to RAID simultaneously from 3 machines; 2) three machines simultaneously reading from the RAID; - in both these cases the results shown are the total throughput; 3) writing a single file to disk; 4) reading a single file from disk; 5) simultaneously 2 machines reading from, and 2 machines writing to the RAID.

The test results demonstrated that there were genuine differences in the performance of the boxes we tested, and that in most cases, the results were significantly different from the published performance specifications claimed by the vendors. Although the equipment we tested was all within the same general class of RAID devices (high performance, high capacity, moderate redundancy), there were wide variations even in the capability of the RAID boxes to meet the qualifying 10 MB/s throughput of the RedWood tape drives. Our tests required the vendors to provide at least SCSI II Fast Wide interfaces and vendors were allowed the opportunity to demonstrate higher performance I/O if they provided all required host adapters. Two vendors provided adapters and demonstrated UltraSCSI connections. On average, this class of RAID equipment was able to provide SCSI II data transfers in the range of the 10 MB/s required, with writes at a slightly lower rate than reads as expected. The UltraSCSI performance was at least 5 MB/s faster, with the winning box (UltraSCSI) achieving 16.5 MB/s with writes (RAM to RAID) and 23 MB/s with reads (RAID to RAM). The throughput for the Network Attached RAID units (NFS servers) averaged around 3-5 MB/s over the 100 Mbit Ethernet test network for a single transfer, with the winning box achieving aggregate data transfers in the 6-8 MB/s range. Only two vendor boxes were tested in the Network Attached RAID procurement due to standardization requirements on the site, with little difference between their performance.

There were several surprises both in the achieved performance as well as the vendor configurations. We found that on several transfers, there were real differences depending on the direction of the transfer, and the differences were not consistent across vendors. Some vendor boxes passed the tape to RAID test but failed the RAID to tape test; for others, the outcome was reversed. The differences are in part due to the fact that in some systems write operations are given higher priority than reads; other systems respond immediately once the data is written to cache and before it is flushed to disk

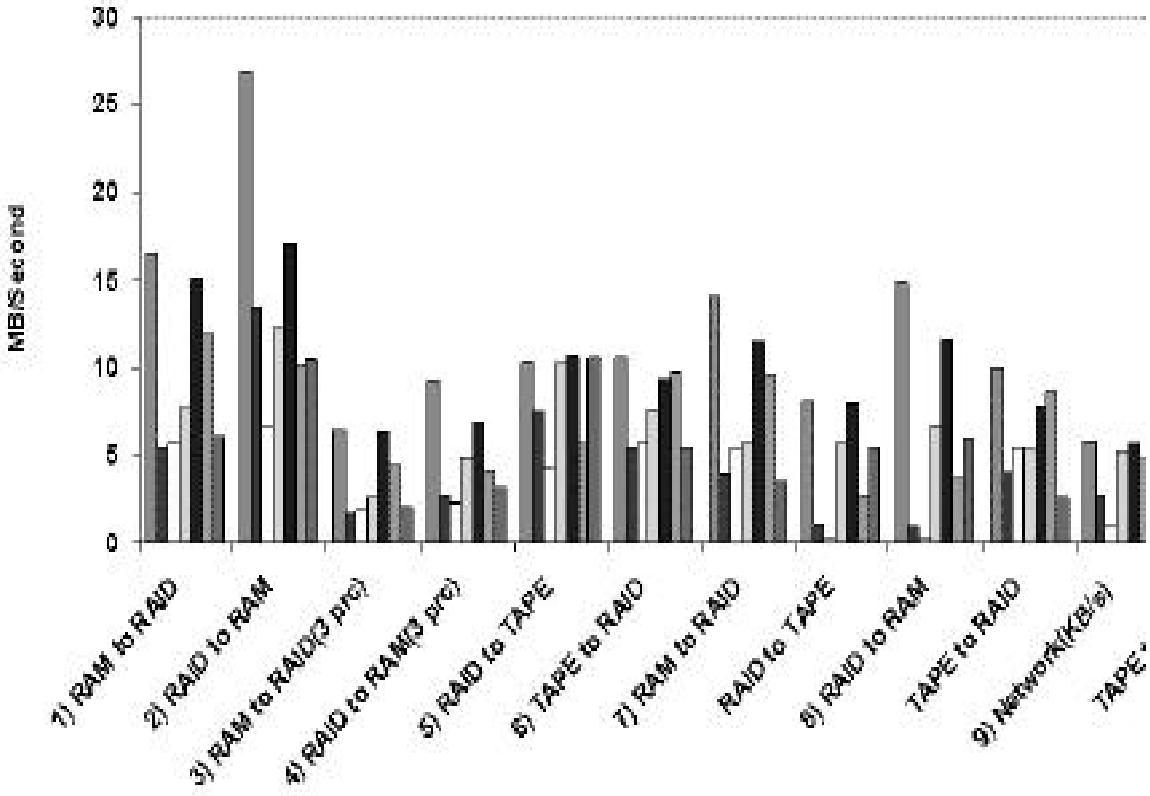

Figure 3. Host-attached RAID results

.

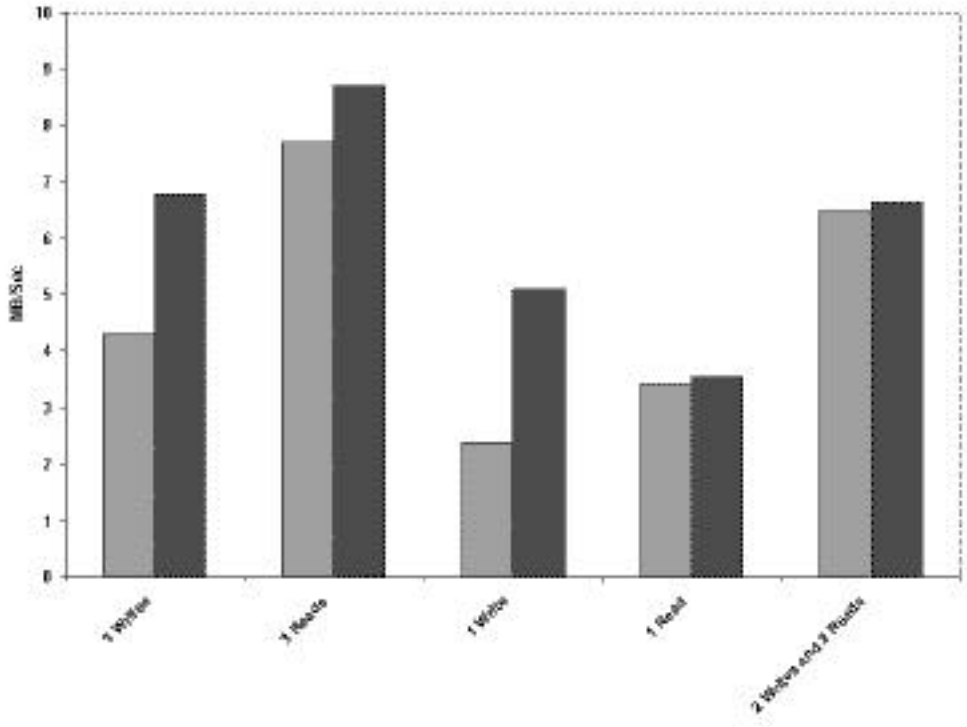

Figure 4. NFS RAID test results

Performance generally scaled well for multiple simultaneous transfers up to approximately 3 nodes, where the limiting factor was the transfer rate to the local disk on the batch node. Beyond 3 nodes, we found that the limiting factor was in fact the RAID system, with some boxes achieving overall higher aggregates than competitor equipment. Another test which provided real discrimination between the vendor boxes was a test that involved simultaneously moving 1 GB of data from a RedWood tape drive to the RAID system while reading a 1 GB file of physics data from the RAID into memory and vice versa. The data rate from the RAID system to memory should be the greater since it is not limited by the tape drive. This test for some equipment showed reductions in the data rates to and from the tape drives if other processes were heavily using the RAID system. The data rates to and from the tape drive were understandably lower than performing a sole tape test, but the better RAID systems had high data rates for both processes. Similar simultaneous tests found that RAM to RAID performance on some of the RAID systems degraded when coupled with network transfers. Our tests also revealed some tuning factors we can use to optimize our data transfers. Repeated testing during test suite development revealed that a blocking size of 256 KB provided the best overall performance on the RedWood tape transports. Furthermore, it was found that the amount of data transferred would affect the overall throughput. Large data sets completely fill buffers and slow the rate. On the other hand, small data sets end up measuring the throughput of the buffer or cache and not the RAID itself.

Vendor configurations proved equally enlightening. We found repeated examples of vendors over-tuning their equipment, which in some cases even resulted in instability and emergency reconfigurations. Our test team discovered that more than one vendor had attempted to skew the results by generating the test file system on only the outer tracks of the disks (i.e. the “sweet spot”) or writing data to the raw device as opposed to a file system, both obviously unattainable in a standard production environment. One further “unattainable” specification was revealed by the vendor who explained that the reason their results did not approach the published specifications was that their in-house testing reported only RAM to cache testing, and not the final data rate achieved when storing to the hard disk. Although many vendors requested to use software RAID across multiple hardware-based RAID volume sets to demonstrate the final transfer rates obtainable, our test plan called for as much apples to apples testing as possible, not to mention the “keep it simple” goal. (As an aside, the use of software RAID in this fashion may be worth considering to achieve the high parallel transfer rates that may be required in the Jefferson Lab environment. At the moment, the increased liability in the event of multiple drive failures, plus the increased complexity of the configuration, advise against this option.)

4 Conclusions

In conclusion, the evaluation of commercial RAID offerings was significantly enhanced by the effort invested to test the equipment in the actual Jefferson Lab computing environment. The testing found real differences between vendor boxes, not only in their architecture and levels of flexibility and redundancy, but also significant variations in the performance of a variety of tests selected to mimic the stress points in our data handling operation. The preparation and logistics, while not insignificant, were well repaid by the knowledge we gained regarding the alternatives proposed in the RAID solicitations as well as potential tuning optimizations realized for our environment. In each procurement, the vendor that won was a clear winner for our environment.

The most important lesson learned is the realization that commercially reported performance results are as a rule highly skewed, the result of optimizing the box in a way few real world applications could, in order to maximize the test. The Jefferson Lab evaluation team uncovered multiple examples even in our own on-site testing process of results invalidated by the vendors’ attempt to “ace the test.” Building file systems on only the sweet spot on the disks, reporting results with redundancies turned off and non-RAID protected configurations, and even reporting transfer rates actually generated by cache-target transfers are but a few of the techniques vendors use to produce standardized results that bear little resemblance to the actual performance the customer will experience. The bottom line for high dollar, high performance RAID systems is that a fairly simple set of tests designed around your specific application may be the best performance indicator available. Testing in the "real world" is the one test you can't afford to forego.

References

[1] This research was sponsored by the Department of Energy contract DE-AC05-84ER40150. Views and conclusions contained in this report are the authors’ and should not be interpreted as representing the official opinion or policies.