Study of Solid State Drives performance in PROOF distributed analysis system

S.Y. Panitkin[1], M. Ernst[1], R.Petkus[1], O. Rind[1] and T. Wenaus[1]

[1]Brookhaven National Laboratory, Upton, NY 11973, USA

Abstract. Solid State Drives (SSD) is a promising storage technology for High Energy Physics parallel analysis farms. Its combination of low random access time and relatively high read speed is very well suited for situations where multiple jobs concurrently access data located on the same drive. It also has lower energy consumption and higher vibration tolerance than Hard Disk Drive (HDD) which makes it an attractive choice in many applications raging from personal laptops to large analysis farms. The Parallel ROOT Facility - PROOF is a distributed analysis system which allows to exploit inherent event level parallelism of high energy physics data. PROOF is especially efficient together with distributed local storage systems like Xrootd, when data are distributed over computing nodes. In such an architecture the local disk subsystem I/O performance becomes a critical factor, especially when computing nodes use multi-core CPUs. We will discuss our experience with SSDs in PROOF environment. We will compare performance of HDD with SSD in I/O intensive analysis scenarios. In particular we will discuss PROOF system performance scaling with a number of simultaneously running analysis jobs.

1. Introduction

The Parallel ROOT Facility - PROOF [1] is a distributed analysis system that allows to exploit inherent event level parallelism of high energy physics data. PROOF is a part of ROOT [2] framework, it is distributed and supported by the ROOT team.

It can work efficiently on different types of hardware and scales well from a multi-core laptop to large computing farms. From that point of view it is well suited for both large central analysis facilities and Tier 3 type analysis farms.

In its current implementation it is tightly integrated with Xrootd [3] and allows to fuse distributed computing and distributed storage for effective use of resources. Using grid terminology, in PROOF a computing element is a storage element. Hence local data processing is encouraged, with PROOF performing automatic matching of code with data. In such an architecture the local disk subsystem I/O performance becomes a critical factor, especially when computing nodes use multi-core CPUs. The purpose of this study was to explore disk sub-system performance in realistic physics analysis scenarios in PROOF.

2. Tests setup

2.1. General description

For our tests we have chosen two analysis scenarios representing different use cases and different type of data access. The first one emulates a typical interactive analysis of data at ROOT prompt, with single variable plotting. This scenario exercises a very sparse access to data, with

Table 1. Some parameters of Mtron Solid State Disk model MSP-SATA7035064.

Average access time	Sustained Read/Write	IOPS (Sequential/Random)	Write Endurance
0.1 ms	120/80 MB/s	81K/18K	> 140 years at 50 GB per day

Table 2. Some parameters of Seagate HDD model ST3750640NS.

Average seek time Read/Write	Average Track-to-Track Seek Time. Read/Write	Sustained transfer rate	Average Latency
8.5/9.5 ms	0.8/1.0 ms	105 MB/s	4.16 ms

only one variable being read in. Processing is minimal since only filling and plotting of a histogram is involved.

For this scenario we used PROOF Bench test suite available in the ROOT distribution. It allows to generate quickly an arbitrary number of events resembling a typical event structure in high energy and heavy ion physics, with global event information and collections of tracks in ROOT tree format. In these tests we generated 10^7 events with average size of about 1 kB per event.

The second scenario was intended to be more realistic. This test is running a typical Atlas Higgs search analysis on a dataset in Derived Physics Data (DPD) [4] format. The dataset had 200 files with about 3.4 million simulated Monte-Carlo events with total volume of about 46.4 GB. This scenario is both CPU and I/O intensive. Every variable in the event was used for analysis that involves complex calculation and multiple histogram fitting.

2.2. Hardware details

All machines used in the tests had dual quad core Intel ”Kentsfield” CPUs, running at 2.0GHz, with 16 GB of RAM. For these tests we have used 64 GB SSDs manufactured by the Mtron Corporation, model number MSP-SATA7035064 . Some of this model parameters relevant for our tests are presented in the Table 1.

For comparison we utilized 750 GB (7200 rpm) hard drives manufactured by Seagate Technology Corporation, model number ST3750640NS. This type of disk is typical for the Atlas Tier 1 site at BNL as of 2009. Some of the drive parameters relevant for our tests are presented in the Table 2.

Most of the tests were done with the PROOF farm configured with one redirector node and one worker node. This configuration was intended for study of local I/O performance. Another configuration with one redirector and eight worker nodes was used for study of the farm-wide scaling performance. All machines used in these tests were running Scientific Linux 4.2 operating system, with default settings for I/O. In all cases we used ext3 file system. The tests were performed in a single user mode, with no ambient load on the farm. In order to avoid effects of data memory caching the farm was rebooted before every measurement.

Information about analysis rates were obtained using tools provided by PROOF. Additionally we monitored the farm's hardware parameters during the tests using Ganglia [5] monitoring system.

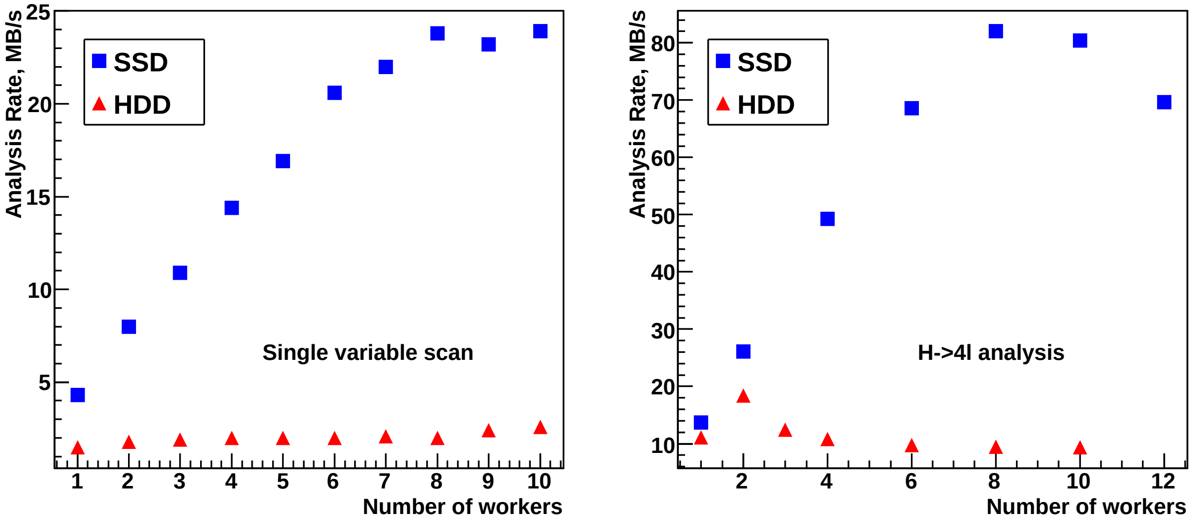

Figure 1. Comparison of SSD vs HDD single drive performance vs worker load. See description in the text

3. Results

3.1. Single drive performance

Figure 1 shows the measurements of the single drive performance. The analysis rate, expressed in MB per second, is shown as a function of the number of workers. A worker in this context is an analysis job running in PROOF. It is obvious that in this scenario SSD holds a clear advantage in read rate over the HDD.

Left Panel in Figure 1 shows comparison between HDD and SSD in single variable scan scenario. At full CPU occupancy, with 8 workers running, SSD shows order of magnitude higher analysis rate. The HDD performance never exceeds 3 MB/s rate, while the SSD reaches almost 25 MB/s, with 8 workers accessing the data concurrently. With a number of workers larger than eight the SSD performance does not increase any more but levels off. At this point the system seems to be CPU limited since number of workers is equal to number of CPU cores.

On the right panel in Figure 1 the comparison is shown for the realistic physics analysis. In terms of I/O this analysis is more challenging, with analysis rates reaching more than 80 Mb/s. Qualitatively the situation is similar to the one in single variable scan case. The SSD demonstrates up to factor of ten better analysis rate with 8 running workers. The HDD achieves its maximum performance of about 18 MB/s with two running workers. After that HDD performance rapidly deteriorates to below 10 MB/s. It simply can not keep up with I/O demand.

The reason for this performance disparity is clear from Tables 1 and 2. The poor performance of HDD in multi-worker scenario can be explained by relatively high latencies associated with concurrent data access. The SSD random access latency is about 100 times better.

3.2. Single drive vs software RAID performance

We compared performance of a single drive with one of the software RAID 0 array. For the SSDs we used two disk array and for the HDDs a 3 disk array. The results of this comparison are shown in Figures 2 and 3.

Figure 2 presents summary of the measurements done for the single variable scan case. In this

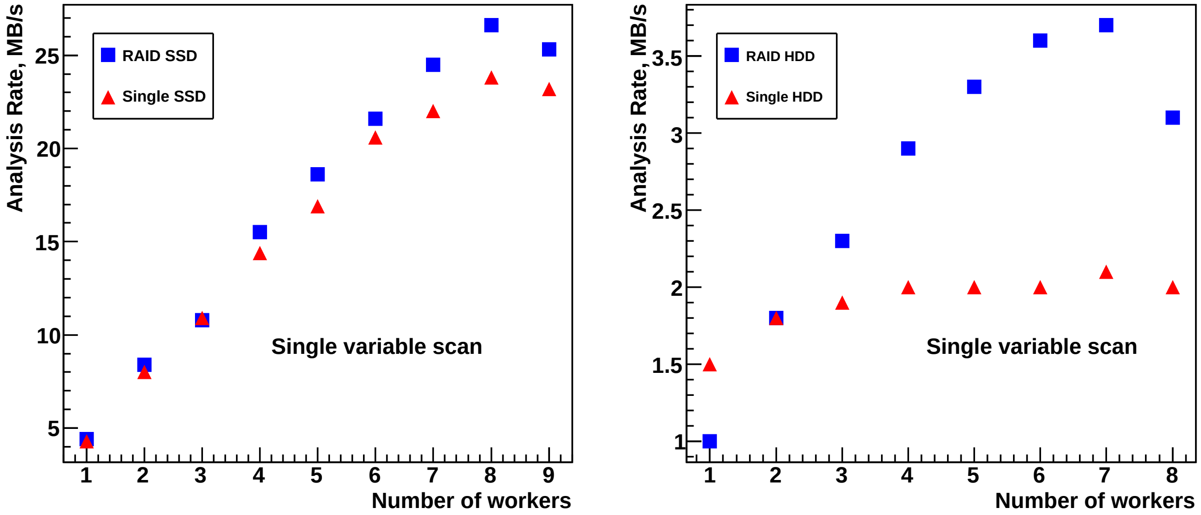

Figure 2. Single drive vs RAID performance in the single variable scan type of analysis. See description in the text

case the SSD RAID array does not offer significant performance advantage compared to a single SSD. This can be seen in the left panel of the Figure 2. The analysis rate for the RAID is pretty much identical to a single drive for up to 4 worker load and is only 10% higher with 8 workers. On the other hand the HDD RAID shows clear advantage over a single hard drive in this scenario. This can be seen in the right panel of Figure 2. The HDD RAID performance keep rising, almost linearly, reaching about 3.7 MB/s rate with 7 workers, though performance increase from 6 to 7 workers is negligible. With 7 workers the RAID array shows almost factor of two better performance than single drive. The HDD RAID analysis rate drops by about 20% with 8 workers.

We also made a similar comparison for realistic analysis. The results are shown in Figure 3. The SSD comparison is shown in the left panel. Since this analysis is much more input intensive the absolute performance numbers are much higher in this case than for a single variable scan case. The SSD RAID performance exceeds 100 MB/s with 8 workers. The single SSD reaches about 80 MB/s with 8 workers. Both configurations do not show increase in performance beyond 10 worker case. In this scenario the HDD RAID seems to be more beneficial for performance. It offers a factor of three increase in the analysis rate compared to a single drive. The HDD RAID rate achieves almost 30 MB/s with 3 workers after which it started to deteriorate. This deterioration is clearly I/O related since at this number of workers the system's CPU is still under-utilized. Please note that, even even at its peak, the HDD RAID performance is about 5 times worse than that of the single SSD disk. A single hard drive peaks at about 18 MB/s with 2 analysis workers. Then rapidly drops below 10 MB/s. Clearly, a single hard drive can not cope with multiple workers in this scenario. Figure 4 illustrates PROOF scalability. Since input/output is mostly localized in PROOF the aggregate PROOF farm performance is just a sum of single node performances.

4. Conclusions

From our tests it is evident that the Solid State Drive technology offers significant performance advantage in parallel analysis environment, such as PROOF. We observed up to an order of

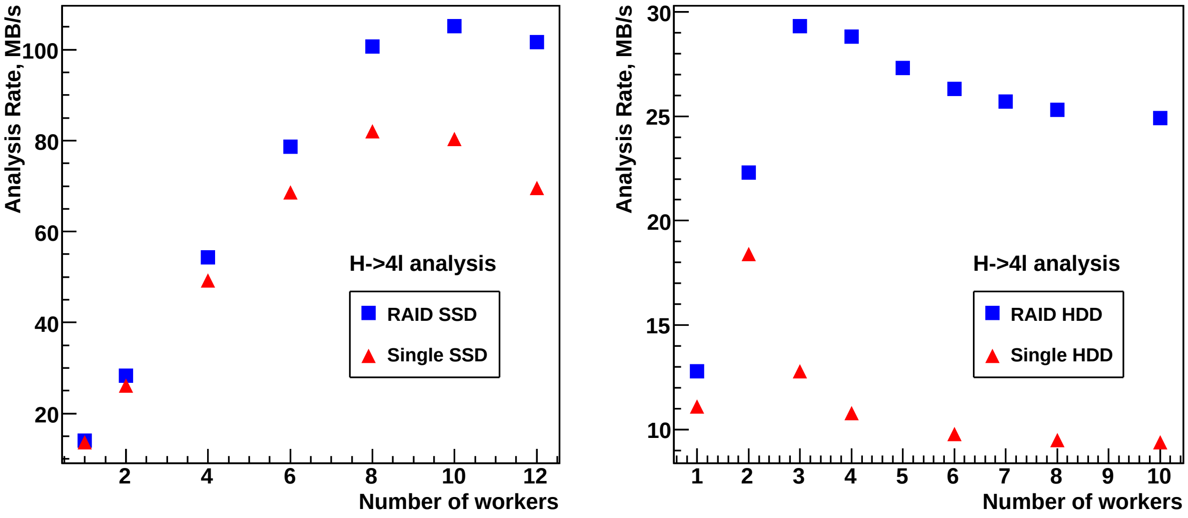

Figure 3. Single drive vs RAID performance in H→4l analysis. See description in the text

magnitude better analysis rates with SSD compared to HDDs.

With current multi-core CPUs utilized in analysis farms the task of providing running jobs with data at a optimal rate becomes a challenge. A single ROOT analysis job running on a modern CPU can develop 10-15 MB/s read rate. With 8 cores this adds up to 80 to 120 MB/s read rate - a very significant load, capable of saturating a Gigabit network link or bringing down a hard drive. In our tests we observed that single hard drive reaches maximum performance with two concurrent workers and becomes a bottleneck with larger number of workers. A Hard drive RAID helps somewhat, but fails to provide adequate performance with number of workers larger than four. On the other hand a single SSD is capable of providing most of the needed data rate in all cases. The only cases where we observed performance deterioration of SSDs were related to CPU limited regime. Our tests of a two disk software RAID on SSD showed an increased analysis rate of 10 to 20% over a single drive.

Currently, the main drawbacks of the SSD technology are relatively high price and smaller capacity. We hope that both of these drawback will disappear in the near future. That will allow significantly increase performance of physics analysis farms.

References

[1] Ballintijn M *et al.* 2006 *Nucl. Instrum. Meth.* **A559** 13–16

[2] Brun R, Rademakers F, Canal P and Goto M 2003 (*Preprint* `cs/0306078`)

[3] Dorigo A, Elmer P, Furano F and Hanushevsky A 2005 *TELE-INFO'05: Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics* (Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS)) pp 1–6 ISBN 960-8457-11-4

[4] Adams D, Barberis D, Bee C, Hawkings R, Jarp S, Jones1 R, Malon D, Poggioli L, Poulard G, Quarrie D and Wenaus T 2005 The atlas computing model Tech. rep. CERN cERN-LHCC-2004-037/G-085

[5] Massie M 2004 *Parallel Computing* **30** 817–840 ISSN 01678191 URL `http://dx.doi.org/10.1016/j.parco.2004.04.001`

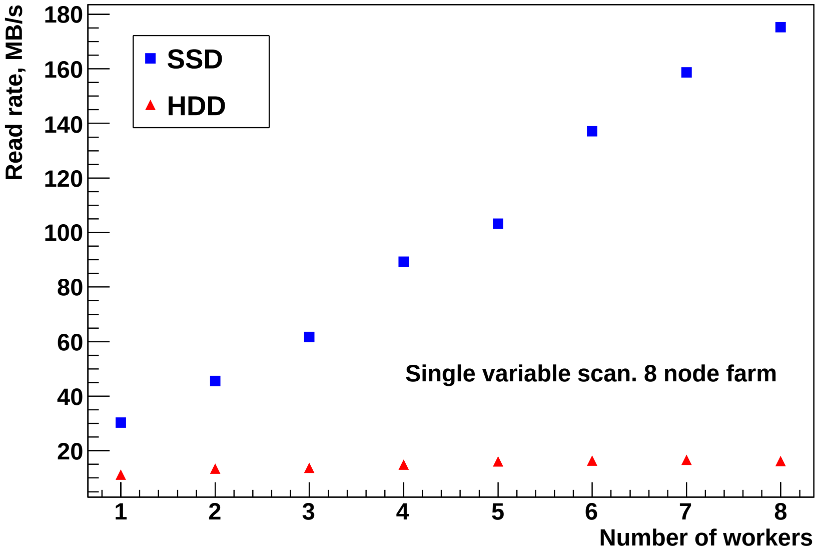

Figure 4. Aggregate analysis rate as a function of number of workers per node for an 8 node farm. See description in the text