

# Technical Report: Samsung HD103SJ Spinpoint F3

Paolo Bertasi, Federica Bogo, Martina Muscarella  
University of Padova

April 24<sup>th</sup>, 2011

## 1 INTRODUCTION

Simulation of complex mathematical methods plays an increasingly important role in a wide range of research fields: galaxy evolution study, proteins analysis, meteorology are just a few examples of such areas. Clearly, performing these simulations requires higher and higher computational power: this trend has encouraged the design, development and deployment of computer clusters for computational purposes. “Eridano”, set up at the Department of Information Engineering at the University of Padua, is an example of such a cluster.

Since high performance scientific computing corresponds to CPU-intensive applications, designers usually pay most of the attention to the the highest layers of the memory hierarchy (i.e., caches and main memory); lower layers, such as external storage, are thought to play a secondary role. However, scientific applications often involve the management of very large datasets and, consequently, an intensive use of hard drives. If not carefully considered, I/O efficiency can become the actual bottleneck for the system as a whole.

Hence, our goal here is to benchmark the Samsung HD103SJ Spinpoint F3 hard drives our cluster uses, generously donated by Samsung, in order to accurately evaluate the performance impact they may have.

The rest of this paper is structured as follows: Section 1 provides an overview of the structure of “Eridano”, pointing out, in particular, configuration parameters and performance specifications of hard drives; Section 2 details the tests we carried out and the results we obtained; finally, Section 3 sums up some conclusions.

## 2 CLUSTER OVERVIEW

“Eridano” is made up of sixteen nodes, joined by two interconnection networks: a dedicated 10 Gb/s Ethernet, relying on a Fujitsu XG2600 switch,

and a 1 Gb/s Ethernet. While the former is used for high performance communications among compute nodes, the latter makes the cluster accessible from remote in order to perform administration and monitoring tasks.

Each node is equipped with an Intel i7 950 @ 3.07 GHz, 12 GB of DDR3 RAM @ 1600 MHz in tri-channel configuration and six HDDs SAMSUNG HD103SJ Spinpoint F3 (for a total of 96 disks). We remark that these drives were the fastest 7200rpm drives available on the market when the Eridano cluster was set up (late summer 2010). Drive configuration of the disks is shown in Table 1, while their performance specifications are listed in Table 2 (for further details, see [2]).

Drive Configuration	
Capacity	1 TB
Interface	Serial ATA 3.0 Gbps
Buffer DRAM Size	32 MB
Byte per Sector	512 B
Rotational Speed	7,200 RPM

Table 1: Samsung HD103SJ Spinpoint F3 HDD: drive configuration.

Performance Specifications	
Average Seek time (typical)	8.9 msec.
Data Transfer Rate Media to/from Buffer (Max.)	250 MB/sec.
Data Transfer Rate Buffer to/from Host (Max.)	300 MB/sec.
Average Latency	4.17 msec.
Drive Ready Time (typical)	11 sec.

Table 2: Samsung HD103SJ Spinpoint F3 HDD: performance specifications.

Currently, each node runs GNU/Linux Gentoo (kernel 2.6.36).

### 3 TESTS

Our tests aimed at evaluating, in particular, homogeneity and I/O efficiency of hard drives under different workload conditions.

In order to perform an evaluation as accurate as possible, we developed an ad-hoc benchmark suite. The code, written in C and in Bash scripting, adopts some careful design choices: transfer of data from and to disks is made by means of *direct* I/O, bypassing kernelspace buffers; dummy reads and writes of data invalidate disk buffer entries before measurement; finally, raw disk access avoids the performance impact of filesystems.

This section provides a brief overview of the tests: for each of them, we describe the goal it addresses, the methodology it adopts and the results it produced.

### 3.1 Bandwidth Test

Modern disks are logically organized as linear arrays of data blocks: the access to a sequence of contiguous blocks requires a fixed seek time, independent of the amount of data, plus a transfer time that is directly proportional to it; the nearer are the accessed data to the outer rim of the disk, the lower is the transfer time. This crucial (but often disregarded) aspect must be taken into account when performing disk I/O operations.

**Goal** The Bandwidth test pursues two goals: measuring the read and write average bandwidth achieved by the disks and evaluating the spread of performance across different devices.

**Methodology** In order to perform an exhaustive evaluation, we measured the peak read (write) bandwidth achieved by each drive when reading (writing) different amounts of data, placed at different offsets along the disk logical linear array; in the following, we consider that offset 0 (i.e., the beginning of the array) corresponds physically to the area of the disk closer to the outer rim.

We considered 15 different offsets, interleaved by approximately 62.5 GB. Starting at each one of these offsets, we accessed sequences of contiguous data blocks of different size: more precisely, we considered 22 increasing powers of two as block size, from a minimum of 512 B until a maximum of 1 GiB.

This methodology provided us with a collection of 352 samples per disk, each one corresponding to a distinct “offset-block size” pair; for each sample, we calculated mean and standard deviation across all 96 disks.

**Results** Figure 1 and Figure 2 show how read and write bandwidth average values change as a function of both transferred block size and offset along the disk.

Write and read throughput values turn out to be similar and follow the same trend. Indeed, as expected, performance heightens as the distance from the outer rim decreases and the block size increases (for larger amount of data, transfer time dominates seek time). The throughput peaks (143 MB/s) occur at transfer sizes of 1 GiB and offsets equal to 0.

Read and write bandwidth standard deviation values are shown in Figure 3 and Figure 6. (Low) variability in performance is an extremely important characteristic for disks, particularly when used in large clusters – since the performance of the whole system is typically dictated by that of the slowest disk(s) is typically. The variability between the disks we are testing is

relatively small, with a standard deviation always less than 7%. By means of comparison, we reproduce the same performance curves measured on 13 “identical” Western Digital WD1600AAJS drives used by our group for competing in the PennySort benchmark 2 years ago [4].

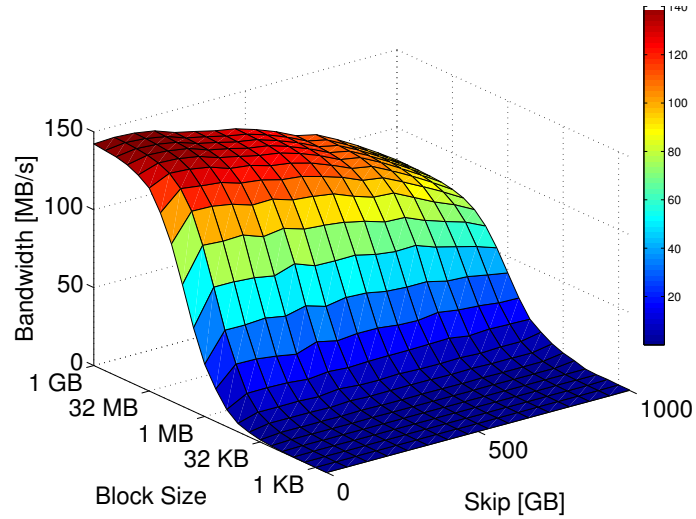


Figure 1: Average read bandwidth as a function of both read block size and offset along the disk.

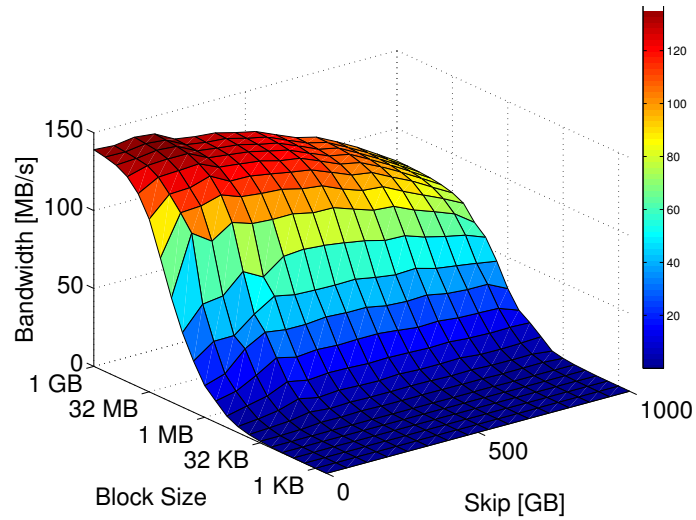


Figure 2: Average write bandwidth as a function of both written block size and offset along the disk.

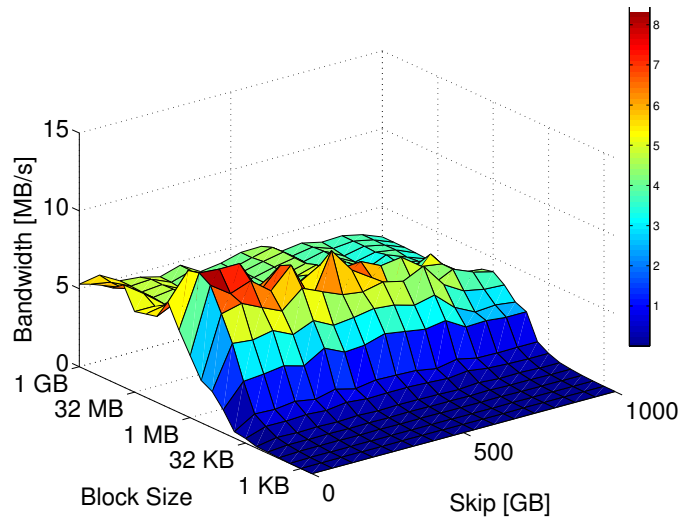


Figure 3: Read bandwidth standard deviation as a function of both read block size and offset along the disk.

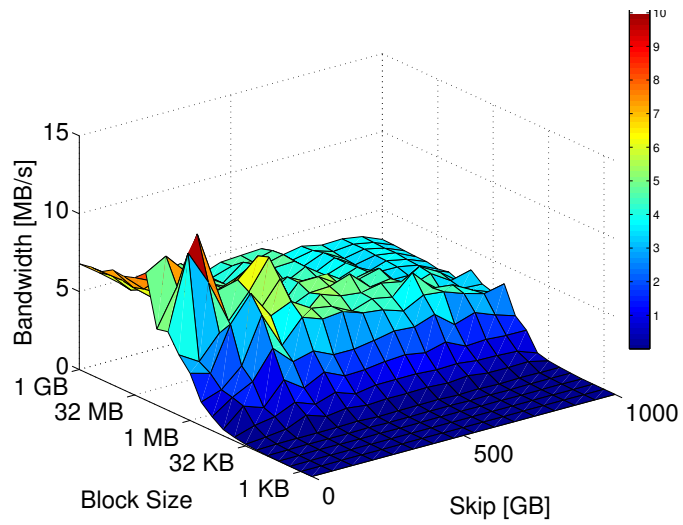


Figure 4: Write bandwidth standard deviation as a function of both written block size and offset along the disk.

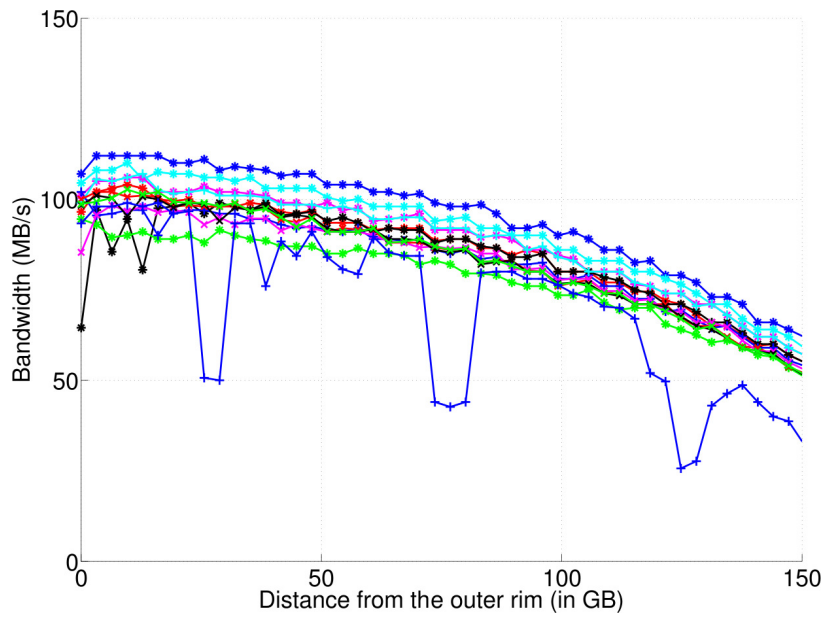


Figure 5: Read speed of 13 “identical” Western Digital WD1600AAJS drives, as a function of the distance in gigabytes from the outer rim of the disk.

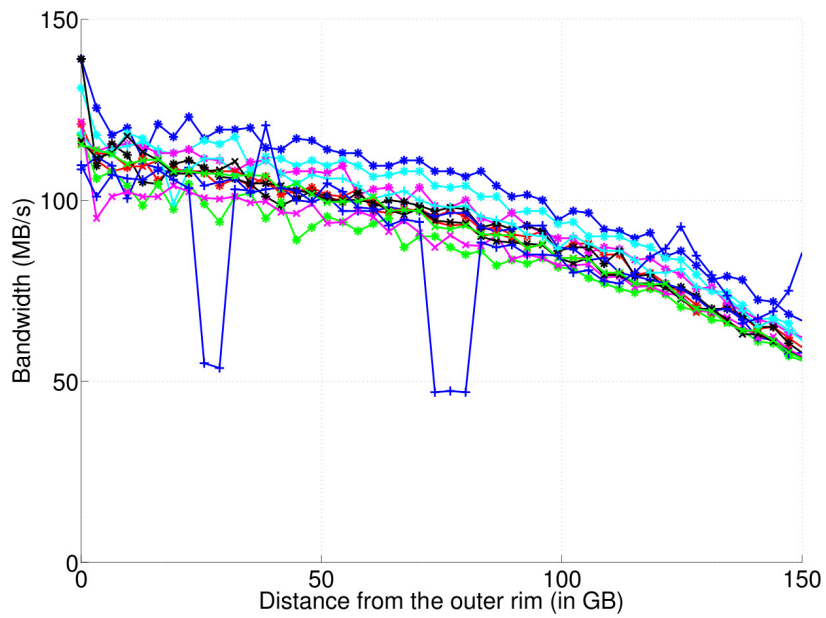


Figure 6: Write speed of 13 “identical” Western Digital WD1600AAJS drives, as a function of the distance in gigabytes from the outer rim of the disk.

### 3.2 Time fluctuations Test

The Bandwidth test adopted a “synchronic” approach, comparing the measurements carried out for different devices in a single instant; in contrast, the Time fluctuations test assesses how the I/O efficiency achieved by a single drive may change over time.

**Goal** We evaluated the performance fluctuations in read (write) bandwidth exhibited by the same disk over a time interval of at most one day.

**Methodology** As in the previous test, we accessed data placed at 15 different offsets along the device; the read (written) block size was kept fixed at 128 MB. For each different offset we took 24 measurements, calculating their mean and standard deviation. We performed two different experiments: in the first we paused for 1 second before the end of a measurement and the beginning of the next (yielding a total time of approximately 15 minutes), in the second we paused for 1 hour (yielding a total time of approximately 1 day).

**Results** Figure 7 and Figure 8 show how read (write) bandwidth average and standard deviation vary as a function of the distance from the outer rim, for one of the examined disks, over a period of 15 mins and one day, respectively.

In both cases, standard deviation increases as the transfer rate increases; but in this case, deviation values remain relatively small at no more than 7-8%.

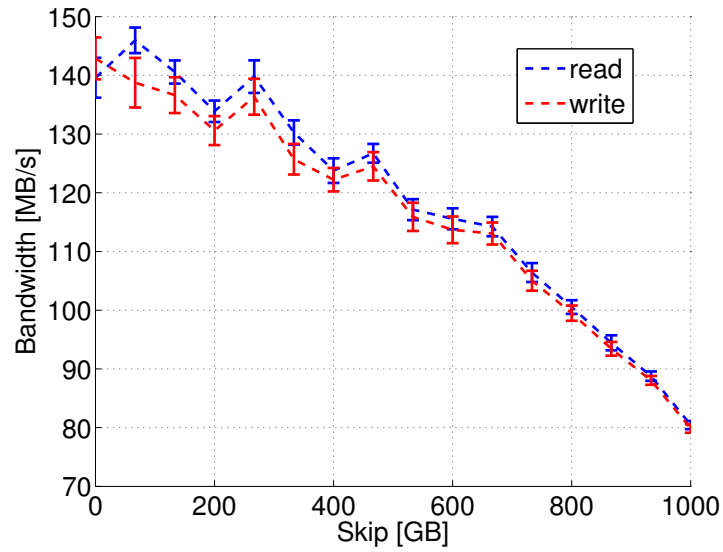


Figure 7: Bandwidth average and standard deviation as a function of the distance from the outer rim (“Skip”, measured in gigabytes), over a period of 15 mins, with a block size of 128 MB.

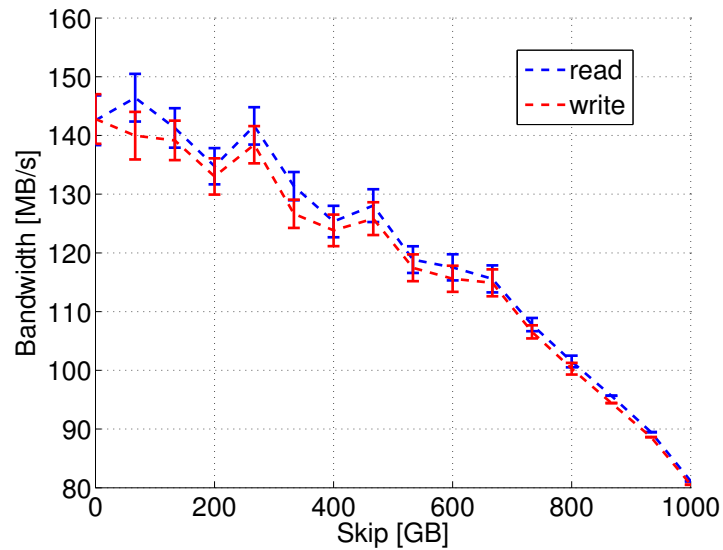


Figure 8: Bandwidth average and standard deviation as a function of the distance from the outer rim (“Skip”, measured in gigabytes), over a period of one day, with a block size of 128 MB.

### 3.3 Temperature Test

In general, temperature can noticeably affect the performance of various computer components. While cold seems to be suitable for RAM and CPU, excessively low temperatures are known to deteriorate disk health [6] (but not other performance metrics, at least in the short term).

**Goal** Our purpose is to evaluate how disk I/O efficiency may vary under different temperature conditions, ranging – approximately – from 15 °C to 35 °C.

**Methodology** In this test, we transferred data blocks of fixed size (128 MiB) from and to the high performance portion of the drive (i.e., the outer rim). Measurements are taken every hour, along a one-week period; during all this period, we changed the environment temperature. At the beginning of each sample interval, the device temperature was estimated by means of the SMART monitoring system<sup>1</sup> provided by the hard drives themselves: such values are plotted in Figure 9.

**Results** Results obtained for one of the disks in the test are shown in Figure 10. Read bandwidth appears completely independent of temperature (at least within the measurement interval). In contrast, write bandwidth exhibits a binary behavior – remaining almost unchanged as long as temperature remains above 19 °C, and dropping to less than 30% (from approximately 130MB/s to slightly more than 30MB/s) as soon as temperature drops to 19 °C or below. Probably, this anomalous behavior can be ascribed to a design choice, introduced by the hard drive manufacturer to avoid damage to the disk [6].

### 3.4 Buffer Test

One of the most interesting features of the hard drives we are benchmarking is the Buffer-to-host hit transfer rate, i.e. the maximum rate at which data can be transferred from the 32 MB disk buffer to the Serial ATA bridge. Its peak value is (theoretically) equal to 300 MB/s. We decided to evaluate how much the rates achieved between buffer and host deviate from that maximum when transferring *different* amounts of data.

**Goal** The Buffer test measures the actual read bandwidth achieved by the disk buffers, as a function of the size of the transferred blocks.

---

<sup>1</sup>via the hddtemp [1] tool

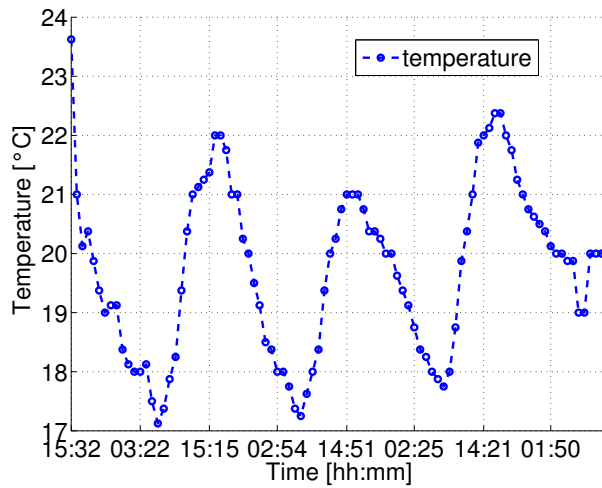


Figure 9: Measured device temperature as a function of time.

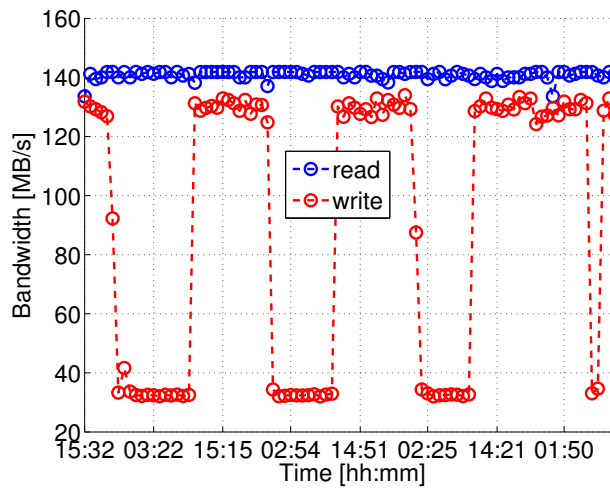


Figure 10: Read and write bandwidth achieved when transferring data blocks of 128 MiB from and to the outer rim of the disk.

**Methodology** As a first thing, we carried out some tests for determining the minimum value for the amount of read data, above which performance measurements remain unchanged; we empirically set this threshold equal to 1 GiB. Hence, in order to measure the buffer-to-host bandwidth, we read that preestablished quantity of data transferring different amounts of bytes (i.e., block sizes) at a time: the values we considered range from 512 B to 64 MiB (note that the disk buffer size is equal to 32 MB).

**Results** Buffer-to-host bandwidth measurements are shown in Figure 11. As expected, the bandwidth increases as the read block size increases, until reaching a maximum of over 250 MB/s at a block size of 512 KiB; this peak transfer rate is maintained by all block sizes in the interval from 512 KiB to 8 MiB. For larger blocks, performance dramatically decreases (note how the bandwidth halves), probably due to phenomena similar to cache thrashing. This suggests that the end user should avoid “maxing out” the disk cache when designing efficient code.

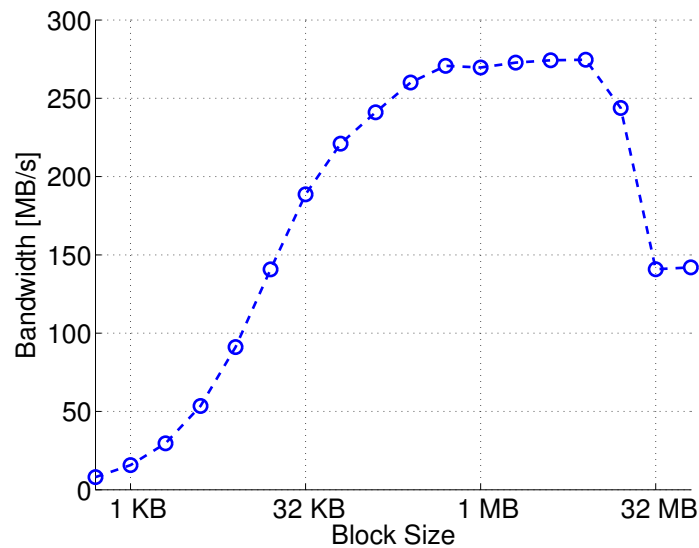


Figure 11: Buffer-to-host bandwidth as a function of the size of the read data blocks.

## 4 CONCLUSIONS

In this paper, we have benchmarked the hard drives used by the cluster “Eridano”, set up at the Department of Information Engineering at the University of Padua. In particular, we have tested some crucial aspects like read and write average bandwidth, changes in I/O efficiency over time, spread of performance among different disks, transfer rate degradation under different temperature conditions, average bandwidth of disk buffers.

The Samsung HD103SJ Spinpoint F3 drives appear to offer excellent I/O performance (sporting a sustained peak throughput of over 140 MB/s) with relatively small fluctuations over time and across disks (a crucial factor on large clusters). They also seem to be quite insensitive to temperature fluctuations, except for a drastic drop in write performance when the temperature falls below 19 °C.

The most significant results we have obtained are summarized in Table 3. Some of these values can be directly compared to those listed in Table 2, showing good correspondence with the specifications provided by the manufacturer.

<b>Measured Performance</b>	
Data Transfer Rate Media to Host	143 MB/sec.
Data Transfer Rate Host to Media	140 MB/sec.
Standard Deviation of Data Transfer Rate Media to Host	8 MB/sec.
Standard Deviation of Data Transfer Rate Host to Media	10 MB/sec.
Data Transfer Rate Host from Buffer (Max.)	275 MB/sec.

Table 3: Samsung HD103SJ Spinpoint F3 HDD: measured performance.

## 5 ACKNOWLEDGEMENTS

We are indebted to Samsung Europe for the generous donation of 100 of their hard drives for these and other tests (including our successful attempt to break the Datamation record [3,5]), for a total market value of 6000-8000 Euros – and in particular to Massimo Germanò, for pushing this project along.

## References

- [1] hddtemp. <http://www.guzu.net/linux/hddtemp.php>.
- [2] Samsung Spinpoint F3 Specifications sheet. [http://www.samsung.com/global/system/business/hdd/prdmodel/2010/2/11/510607F3\\_Spec\\_sheet\\_rev1.pdf](http://www.samsung.com/global/system/business/hdd/prdmodel/2010/2/11/510607F3_Spec_sheet_rev1.pdf).
- [3] Marco Bressan, Paolo Bertasi, Federica Bogo, and Enoch Peserico. psort 2011, winner of pennysort benchmark 2011. [http://www.sortbenchmark.org/psort\\_2011.pdf](http://www.sortbenchmark.org/psort_2011.pdf).
- [4] Marco Bressan, Paolo Bertasi, and Enoch Peserico. psort, yet another fast stable external sorting software. In *Proceedings of the 8th Symposium on Experimental Algorithms (SEA2009)*, 2009.
- [5] Marco Bressan, Michele Bonazza, Paolo Bertasi, and Enoch Peserico. Datamation: a quarter of a century and four orders of magnitude later. In *IEEE Cluster*, 2011. To appear.
- [6] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX conference on File and Storage Technologies*, pages 2–2, Berkeley, CA, USA, 2007. USENIX Association.