StatSoft

Business White Paper

STATISTICA Performance with
Fusion ioDrive

Last Update: June 2011

STATISTICA is optimized for processing large amounts of data. Quick access to stored data is an essential part of this processing. Whether processing a large *STATISTICA* Spreadsheet in read-only mode for analysis, or creating temporary objects during data management operations, storage performance directly affects application performance. Fusion-io, developer of an enterprise storage memory platform based on its ioMemory technology, manufactures a line of memory modules known as the ioDrive. Testing concluded that performance from these drives is substantially better than traditional disk drives. This white paper examines the impact of using an ioDrive with *STATISTICA* while analyzing large data sets, and how performance improvements of 300-500 percent can be achieved.

Testing was performed using the ioDrive 160GB Single Level Cell (SLC) card, run on a 2.20GHz AMD Phenom 9550 Quad-Core Processor running 64-bit Vista Business with 8GB of RAM, configured with Samsung HD103UJ SCSI disk rotating at 7200 RPM.

The tests ran in two categories:

1. Analyzing large spreadsheets
2. Extensive TEMP directory access

When working with the large data sources, the size of the files as compared to available physical memory has significant impact on performance. In all cases, files stored on the Fusion ioDrive loaded quickly, even when opening the file for the first time. Files stored on disk had different results depending on their size compared to the amount physical memory used by the OS for file caching. For files that were smaller than the available cache, the disk-based files took longer to load on the first run, but on subsequent runs the times were the same as the ioDrive, indicating that both ioDrive and disk are cached by the OS. The times listed below are for a 2GB random-filled *STATISTICA* Spreadsheet, consisting of 9,000,000 cases by 30 variables, and performing a subset operation to select about 50 percent of the cases into a new spreadsheet.

Subset of V1 < .5 on 9,000,000 cases by 30 variables:

       File on disk:
            1st pass: 72.4 seconds
            2nd pass: 27.5 seconds

       File on Fusion ioDrive
            1st pass: 27.2 seconds
            2nd pass: 27.4 seconds

Once the file is loaded into the OS cache, all operations were about the same between the two.

However, when using files that exceed the OS cache so that the OS cannot cache the entire file, the difference in processing speed is very significant. This test case used a 47GB random-filled *STATISTICA* Spreadsheet of 200,000,000 cases by 30 variables, running summary descriptive statistics on all variables using default settings. Note that this is a parallelized operation, and the system was configured to use all four CPUs. CPU utilization was monitored, with a low utilization meaning more time was spent waiting on disk access. The difference was impressive.

> Basic descriptive stats on 200,000,000-case by 30-variable file:

> > File on disk:
> > > 432 seconds, overall CPU utilization around 32%

> > File on Fusion ioDrive:
> > > 87 seconds, overall CPU utilization around 90%

Putting the file on **the Fusion ioDrive was five times faster** for the simple descriptive statistics, and is confirmed by the process becoming less I/O-bound and more CPU-bound. Note that one would see performance increases for any operation where the disk access time is large compared to the calculation time, but the increase will be less for calculation-intensive operations that are more CPU-bound.

Finally, in the last scenario, the user's TEMP directory was mapped to the Fusion ioDrive. *STATISTICA* makes use of TEMP space to store intermediate objects when they do not fit into local memory. For this test, we scripted several data management operations for a 9,000,000-case by 30-variable file which created several files in the TEMP directory.

> Data management scripting on 9,000,000-case by 30-variable file:

> > TEMP space on disk:
> > > 312 seconds

> > TEMP space on Fusion ioDrive:
> > > 101 seconds

The results show that the performance was **three times faster on Fusion ioDrive**.

In conclusion, *STATISTICA*—already significantly faster than the competition—is even faster paired with the Fusion ioDrive, showing improvements 300-500percent speed improvements.

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.	France: StatSoft France	Italy: StatSoft Italia srl	Poland: StatSoft Polska Sp. z.o.o.	S. Africa: StatSoft S. Africa (Pty) Ltd.
Brazil: StatSoft South America	Germany: StatSoft GmbH	Japan: StatSoft Japan Inc.	Portugal: StatSoft Ibérica Lda	Sweden: StatSoft Scandinavia AB
Bulgaria: StatSoft Bulgaria Ltd.	Hungary: StatSoft Hungary Ltd.	Korea: StatSoft Korea	Russia: StatSoft Russia	Taiwan: StatSoft Taiwan
Czech Rep.: StatSoft Czech Rep. s.r.o	India: StatSoft India Pvt. Ltd.	Netherlands: StatSoft Benelux BV	Spain: StatSoft Ibérica Lda	UK: StatSoft Ltd.
China: StatSoft China	Israel: StatSoft Israel Ltd.	Norway: StatSoft Norway AS		