FlashSort: JouleSort Benchmark Entry, 2010 Daytona 10 GB class

John D. Davis (Microsoft Research, john.d@microsoft.com)
Suzanne Rivoire (Sonoma State University/Microsoft Research, suzanne.rivoire@sonoma.edu)

Submitted to the Sort Benchmark Committee on 8/19/2009

Report Contents

Summary

We ran the 10 GB JouleSort benchmark on a machine with a Fusion-io high-performance SSD drive and sufficient memory to sort the data set in one pass. The system sorts the 10 GB data set in 31.737 ± 0.081 seconds, at an average power of 127.3 ± 1.9 W, for a total sorting energy of 4039.6 ± 61.5 J (24755 ± 377 records sorted per Joule [R/J]). This score more than doubles the previous record of 11,600 R/J.

Hardware description

Box, power supply, and cooling

FlashSort is based on a Supermicro A+ Server 1021M-T2+V server (http://www.supermicro.com/Aplus/system/1U/1021/AS-1021M-T2+.cfm), which is a 1U server that comes with a 560W 80 PLUS-certified power supply. We used the original power supply but eliminated 2 of the 4 cooling fans.

Supermicro A+ Server 1021M-T2+V server cost: $870.67 (not including processor and memory) (June 2009)

Motherboard

The motherboard is a Supermicro Super H8DME-2 (http://www.supermicro.com/Aplus/motherboard/Opteron2000/MCP55/H8DME-2.cfm).

Processor

The processor is a quad-core AMD Opteron 2373 running at 2.01 GHz.

Price at introduction: $377

Memory

The system had 8 DIMMs of Kingston 2GB DDR2-800, ECC DIMMs, KVR800D2D8P6/2G memory, for a total capacity of 16 GB.

Price: $312.19

Storage

- The OS and other system files are stored on a Micron RealSSD P200 drive (http://download.micron.com/pdf/flyers/ssd_p200_flyer.pdf). This drive does not contain any sort data.
    - Price:~$250
- The sort input and output files were stored on a single Fusion-io ioDrive (http://www.fusionio.com/PDFs/Data_Sheet_ioDrive_2.pdf). The ioDrive has a maximum capacity of 80 GB, but it has three possible configurations with different tradeoffs of performance and capacity. We tested all three settings and did not observe any performance differences between the three settings when executing the JouleSort benchmark. We report results using the middle setting, which gives the ioDrive a usable capacity of 37 GB.
    - Price: ~$2500

Overall System Cost:

- Supermicro 1U server (chassis + motherboard) + processor + memory + Micron SSD + Fusion SSD = $870.67 + $377 (estimate) + $312.19 + $250 (estimate) + $2500 (estimate) = ~$4309.86
    - Note: a cheaper OS drive could be used to reduce the storage cost without impacting energy efficiency.

Software

Operating system

The OS was Windows Server 2008. We used cygwin to generate the sort data, run NSort and *time*, and validate the results.

Sorting software

We used a trial version of NSort for Windows x64, downloaded in July 2009. The contents of the nsort.params configuration file are as follows:

```
-filesys:E:\sort, transfer_size=64M
-method=radix
-memory=12000M
-format=size:100
-field=name:randomkey,size:10,pos:1,char
-key=randomkey
-statistics
-processes=4
```

Justification of Daytona designation

Below, we list the criteria set forth for the Daytona category in the Sort FAQ, explaining why we believe FlashSort meets them.

- *Capable of sorting other record and key types besides 100-byte records with 10-byte random keys* – The software (NSort) and hardware used are not limited to the Sort Benchmark record format.
- *[Should] not significantly degrade in performance when sorting other key and record types* – see next point.
- *[Should] not be overly dependent on the uniform and random distribution of key values in the sort input. If the sort data is to be divided into multiple partitions (for instance in a cluster sort), the sort should not rely on any predetermined partition boundaries. The partition boundaries must be determined either by sampling the input data or during the sort.* – We do use a radix sort rather than a merge sort because of our prior knowledge of the data distribution. The boundaries for the partitions are determined during the sort. We have run a merge sort for comparison and determined that the performance degradation is less than 5%.
- *[Should] not overwrite or destroy their input file* – The input file is not destroyed.
- [*Should] be able to run continuously for one hour without a system failure. This requirement is specific to GraySort and MinuteSort benchmarks.* – n/a for JouleSort.

Measurement infrastructure description

Power measurements

We measured power using a WattsUp Pro ES meter (https://www.wattsupmeters.com/secure/products.php?pn=0&wai=57&spec=3), sampling at the rate of 1 measurement per second. The meter was connected via USB cable to a separate Windows machine, which read the data using a program developed in-house based on the WattsUp API (https://www.wattsupmeters.com/secure/downloads/CommunicationsProtocol080620.pdf). The clocks on both machines were synchronized to the second, and power readings were timestamped to allow them to be correlated with the sort run.

Timing measurements

The sort time used in this report is the result of the Cygwin *time* command for each run of NSort.

Measurement Results

In the data below, the timing information is accurate to three decimal places, and the power information is accurate to one decimal place. Errors reported for quantities that are directly measured, including power and time, are the standard deviations of the values reported for the 5 consecutive runs.

The error (E_{err}) reported for the overall average energy (E) is

$$E_{err} = E * \sqrt{\left(\frac{P_{err}}{P}\right)^2 + \left(\frac{T_{err}}{T}\right)^2}$$

where the overall average power is P with error P_{err}, and the overall average time is T with error T_{err}.

	Time	Avg Pwr	Energy	R/J
Run1	31.765	127.3	4043.7	24723
Run2	31.719	126.5	4012.5	24914
Run3	31.609	127.0	4014.3	24914
Run4	31.765	127.9	4062.7	24619
Run5	31.828	127.7	4064.4	24607
AVG	31.737	127.3	4039.6	24755
Error	0.081	1.9*	61.5	377

*Note: Average power error reported based on device measurement accuracy and not deviation in measurements above.

CPU utilization and disk bandwidth information from the NSort statistics is listed below. “Input” and “output” refer to the two phases of the sort.

	In CPU Util	Out CPU Util	Input BW	Output BW
Run1	185	313	809.72	569.15
Run2	187	316	809.06	571.43
Run3	187	315	809.72	576.04
Run4	185	313	811.03	568.83
Run5	187	311	811.03	568.5
AVG	186	314	810.11	570.79
Error	1	2	0.88	3.15

The power factor of the system was at least 0.98 throughout all of our observations.