Energy-Efficient Sorting using Solid State Disks

Andreas Beckmann   Ulrich Meyer
Goethe University Frankfurt am Main

Peter Sanders   Johannes Singler
Karlsruhe Institute of Technology

The Sort Benchmark

The Benchmark

- Sort 100 byte records with a 10 byte key
- Introduced 1985, starting with 100 MB
- New categories added targeting
  - Speed/Size/Throughput (GraySort)
  - Time (MinuteSort)
  - Cost Efficiency (PennySort)
  - Energy Efficiency (JouleSort, 2007)
    - 10 GB, 100 GB, 1000 GB

Sorting large data sets

- Is easily described
- Has many applications
- Stresses both CPU and the I/O system

Energy Efficiency

- Energy (and cooling) is a significant cost factor in data centers
- Energy consumption correlates to pollution

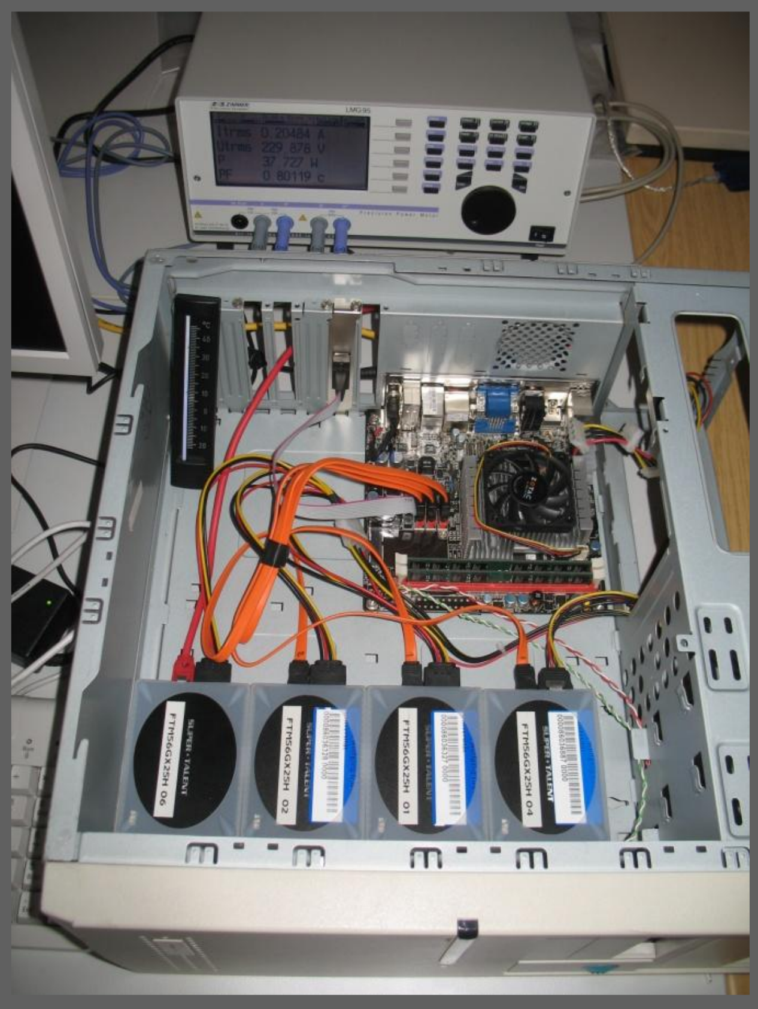
JouleSort Hardware Selection

2007

Rivoire, Shah, Ranganathan, Kozyrakis
Stanford University and HP Labs

2010

Beckmann, Meyer, Sanders, Singler
Goethe University and
Karlsruhe Institute of Technology



2007		2010
Intel Core 2 Duo T7600 (Mobile CPU) 2 cores, 2 threads, 1.66 GHz	**Processor**	Intel Atom 330 2 cores, 4 threads, 1.6 GHz
2 GB	**Memory**	4 GB
2 PCI-e Disk Controllers (8+4 SATA) 1 SATA (onboard)	**I/O**	4 x SATA 3.0 Gb/s (onboard)
13 x Hitachi Travelstar 5K160 160 GB Notebook HDD	**Disks**	4 x SuperTalent FTM56GX25H 256 GB SSD
Linux XFS on Linux Software Raid (Striping)	**OS** **File System**	Linux XFS on Linux Software Raid (Striping)
NSort (commercial sorter)	**Software**	EcoSort, DEMsort using STXXL
59 W 100 W	**Power Idle** **Power Loaded**	25 W 37 W
2007 JouleSort Winner 10 GB , 100 GB		

Algorithms

External Memory Multiway Mergesort

- Phase 1: Run Formation
- Phase 2: Merge Runs
- Careful parameter selection for optimal performance while requiring a single merge pass
- Parallel implementations utilize the 4 CPU threads
- Overlapping of I/O and computation
- Run Formation uses key extraction and radixsort
- Two implementations:

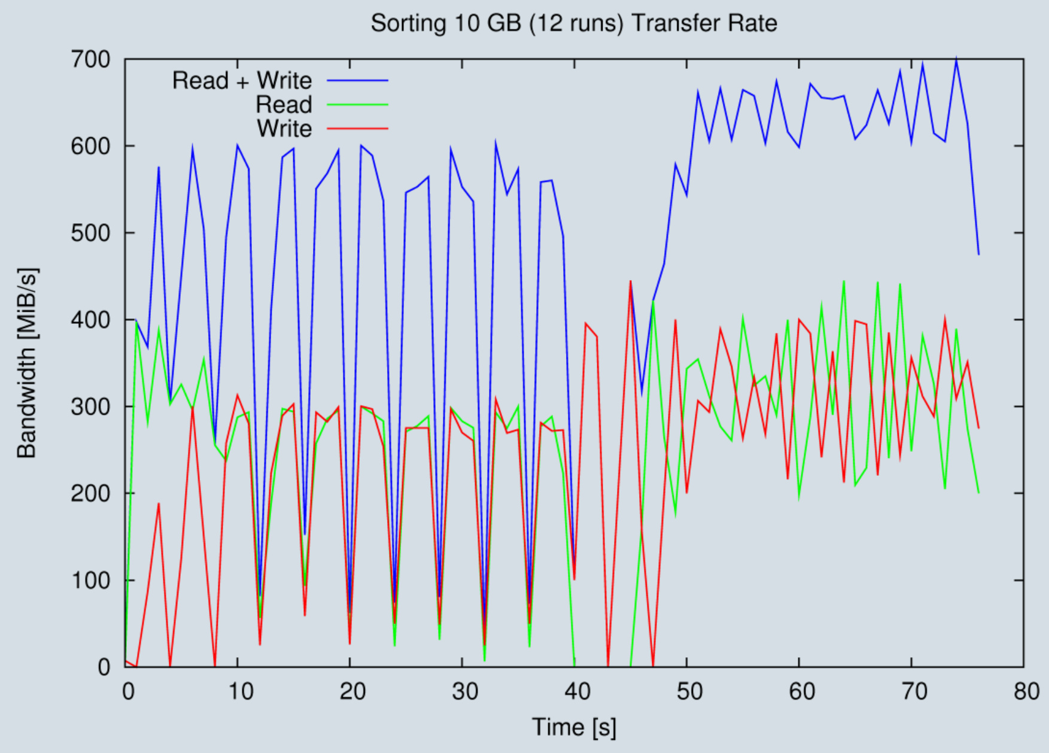
EcoSort (10 GB, 100 GB)

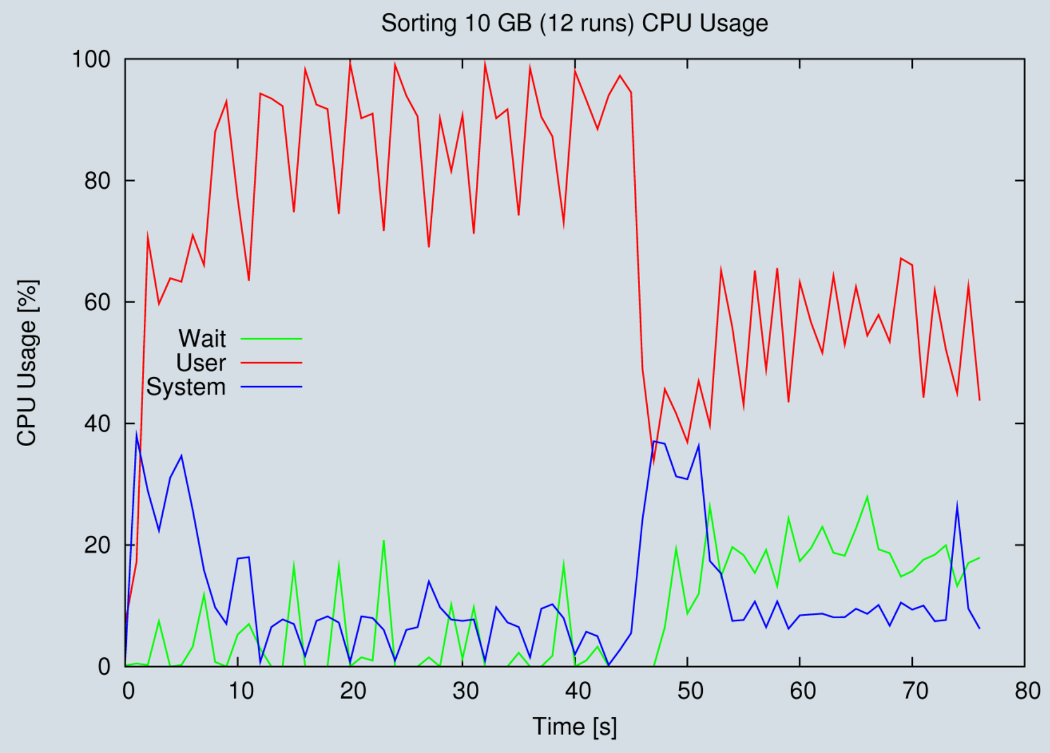
- Bring overlapping to the limits
- Allow independent tuning of more parameters

DEMsort (1000 GB)

- Developed by Sanders, Singler et al. at the Karlsruhe Institute of Technology
- Won the 2009 Sort Benchmark in the categories MinuteSort and GraySort using a 200-node cluster
- Efficient also on a single node
- Allows in-place sorting, needed to sort 1000 GB with just 1024 GB of storage

I/O and CPU utilization while sorting 10 GB:





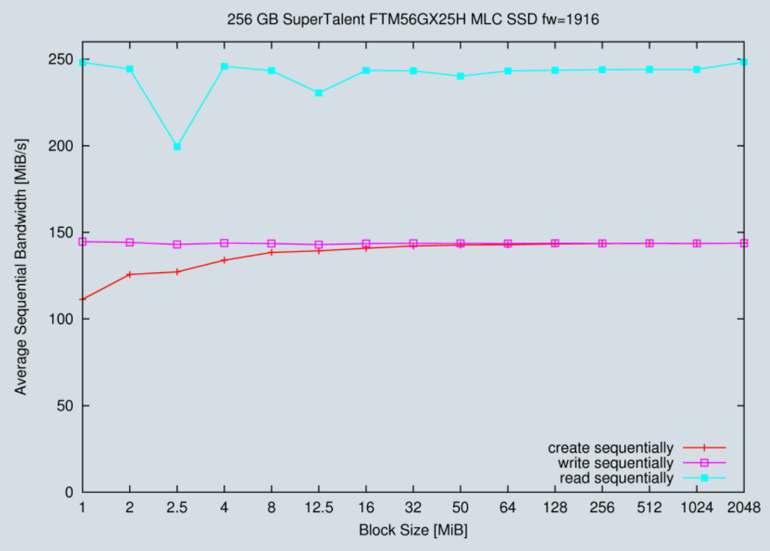
Solid State Disks

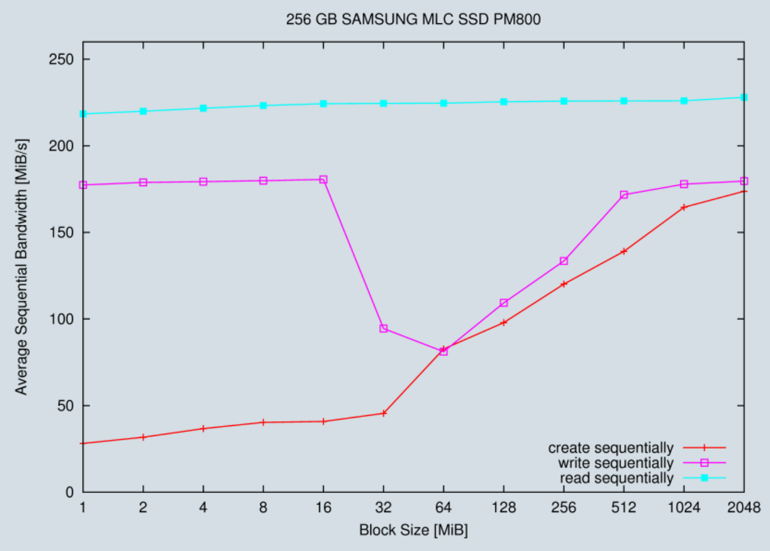
Pro:

- Built from NAND flash memory chips
- No mechanically moving parts
- Good shock resistance
- Low energy consumption
- Higher throughput than HDD
- Support for ATA TRIM command (few models)

Con:

- Higher price and less capacity than today's HDDs
- Small block random writes are slow
- Performance may degrade depending on access pattern
- Properties vary depending on manufacturer, model, firmware





Results

Winner of the Sort Benchmark 2009/2010 mid-year round in the JouleSort Indy categories 10 GB, 100 GB and 1000 GB!

	2007			2010			
Size [GB]	Time [s]	Energy [kJ]	Rec./J	Time [s]	Energy [kJ]	Rec./J	Energy Saving Factor
10	86.6	8.6	11628	76.7	2.8	35453	3.0
100	881	88.1	11354	756	27.5	36381	3.2
1000	7196*	2920*	3425	21906	723.7	13818	4.0

Using low power hardware does not imply an increase in running time: in the 10GB and 100 GB category we beat previous results both in terms of energy consumption and running time.
As a consequence of winning all three categories using a single machine, a new 100 TB JouleSort category was introduced for the 2010 Sort Benchmark.

* The 2007 results for the 1000 GB category were achieved on regular server hardware, not a low energy machine. So we cannot compete in terms of running time, only in energy consumption.