MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Andreas Beckmann
Goethe University


Ulrich Meyer
Goethe University

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
KIT Karlsruhe Institute of Technology
Danmarks Grundforskningsfond
Danish National Research Foundation

Energy-Efficient Sorting using Solid State Disks

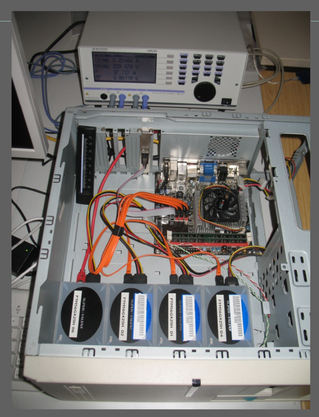
The Sort Benchmark

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
    - 100 TB (2010)
- Classes: Indy (tuned), Daytona (general)

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		2010
Rivoire, Shah, Ranganathan, Kozyrakis Stanford University and HP Labs		Beckmann, Meyer, Sanders, Singler Goethe University and Karlsruhe Institute of Technology
		
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	**OS**	Linux
XFS on Linux Software Raid (Striping)	**File System**	XFS on Linux Software Raid (Striping)
NSort (commercial sorter)	**Software**	EcoSort, DEMsort using STXXL
59 W	**Power Idle**	25 W
100 W	**Power Loaded**	37 W
2007 JouleSort Winner 10 GB, 100 GB		

Algorithms

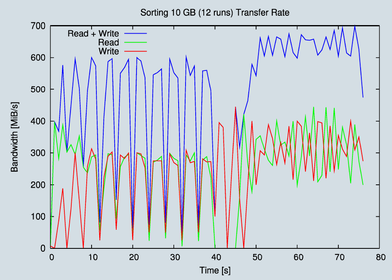
External Memory Multiway Mergesort
- Phase 1: Run Formation
- Phase 2: Merge Runs
- Careful parameter selection for optimal performance while requiring a single merge pass
- Parallel implementations utilize the 4 CPU threads
- Overlapping of I/O and computation
- Run Formation uses key extraction and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

EcoSort (Indy: 10 GB, 100 GB)
- Bring overlapping to the limits
- Allow independent tuning of more parameters

DEMsort (Indy: 1000 GB, 100 TB)
- 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

Nsort (Daytona: 100 GB, 1000 GB)
- Commercial software
- Sorts arbitrary data types

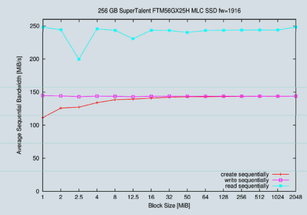
I/O and CPU utilization while sorting 10 GB:

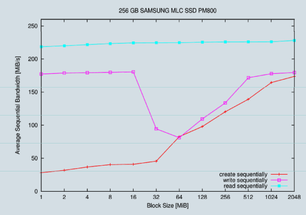


Solid State Disks

Pro
- Built from NAND flash memory chips
- No mechanically moving parts
- Good shock resistance
- Low energy consumption
- Higher throughput than HDDs

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 2010 Sort Benchmark in the JouleSort categories Indy 10 GB, 100 GB and 1000 GB and Daytona 100 GB!

	2007			2010			
Class, Size [GB]	Time [s]	Energy [kJ]	Rec./J	Time [s]	Energy [kJ]	Rec./J	Energy Saving Factor
Indy, 10	86.6	8.6	11628	72.4	2.3	42635	3.7
Indy, 100	881	88.1	11354	691	25.1	39853	3.5
Daytona, 100	881	88.1	11354	756	27.9	35789	3.1
Indy, 1000	7196*	2920*	3425	17026	572	17489	5.1
				2011 (to be submitted)			
Daytona, 1000	7196*	2920*	3425	6486*	1897*	5273	1.5
Indy, 100 TB	-	-	-	9835**	694 MJ**	1441	-

Using low power hardware does not imply an increase in running time: in the 10 GB 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, first 100 TB results to be submitted 2011.

* regular server hardware, not a low energy machine
** 200-node cluster