25 KJ for 100 GB: Energy-Efficient Sorting.

Ulrich Meyer – Goethe University Frankfurt

MADALGO Review Meeting

Green Computing

Source: http://www.sxc.hu/photo/1255482

Source: www.sxc.hu/photo/107856

Serious Background

- Energy costs account for increasing share within total cost of ownership.
- Environmental impact by energy consumption / air conditioning.

Less Serious Background

- Joulesort-Challenge: Fun, competition, intellectual challenge.

Energy Consumption As New Cost Metric

http://www.sxc.hu/photo/1179339

Traditionally: Pure Computing Time

- The faster the better.
- Press every button:
  fast CPU, huge memory, multicores, . . .

http://www.sxc.hu/photo/98930

Now: Energy Consumption

- shorter computing time $\sim$ less energy.
  $\Rightarrow$ optimize both algorithm and implementation.
- faster machine $\sim$ more energy.
  $\Rightarrow$ find best hardware compromise.

Joulesort Challenge

Part of well-established benchmark: http://sortbenchmark.org/

Task:

- Sort a fixed number of randomly permuted 100-byte records with 10-byte keys.
- The sort must start with input in a file on non-volatile store and finish with output in a file on non-volatile store.
- There are three scale categories for JouleSort: 10^8 (10GB), 10^9 (100GB), and 10^{10} (1TB) records
- The winner in each category is the system with the minimum total energy use.

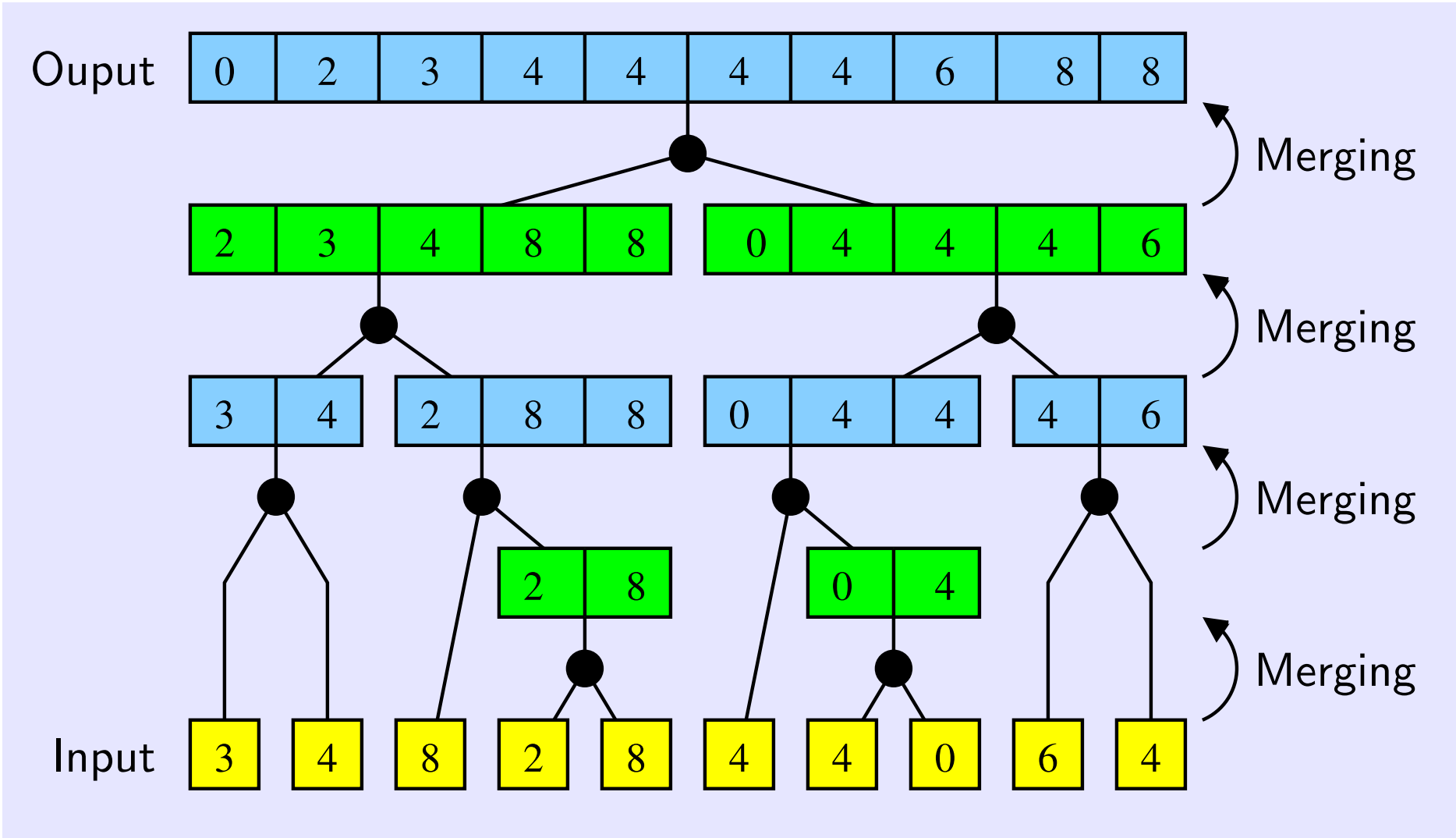
Ex: 30 KJoule = 30 000 watt sec. corresponds to 100 watt for 5 min or 20 watt for 25 min, e.g.

Joule 10^8 recs	2007, 8.6 kJoules **CoolSort** 11,600 records sorted / joule Mobile Core 2 Duo, 13 SATA laptop disks, Nsort Suzanne Rivoire (Stanford), Mehul A. Shah (HP Labs), Partha Ranganathan (HP Labs), Christos Kozyrakis (Stanford)	
Joule 10^9 recs	2007, 88 kJoules **CoolSort** 11,300 records sorted / joule Mobile Core 2 Duo, 13 SATA laptop disks, Nsort Suzanne Rivoire (Stanford), Mehul A. Shah (HP Labs), Partha Ranganathan (HP Labs), Christos Kozyrakis (Stanford)	2009, 87 kJoules **OzSort** 11,600 records sorted / joule 2.6 Ghz AMD Athlon LE-1640, 4GB RA 7x160 GB 7200 RPM SATA, Linux Nikolas Askitis and Ranjan Sinha Univ. Melbourne, Australia
Joule 10^{10} recs	2007, 2920 kJoules **CoolSort** 3,425 records sorted / joule Mobile Core 2 Duo, 13 SATA laptop disks, Nsort Suzanne Rivoire (Stanford), Mehul A. Shah (HP Labs), Partha Ranganathan (HP Labs), Christos Kozyrakis (Stanford)	

Results from 2009.

Note: $\geq$ 25% of compute time in commercial systems is due to sorting.

Sorting with Mergesort



Source: Gerth Brodal

I/O-Performance of Mergesort

Example: $N = 10^9$, $M = 10^7$, $B = 10^4$ elements:

Standard Mergesort: $\mathcal{O}(N/B \cdot \log_2(N))$ I/Os.
30 Merge Phases.

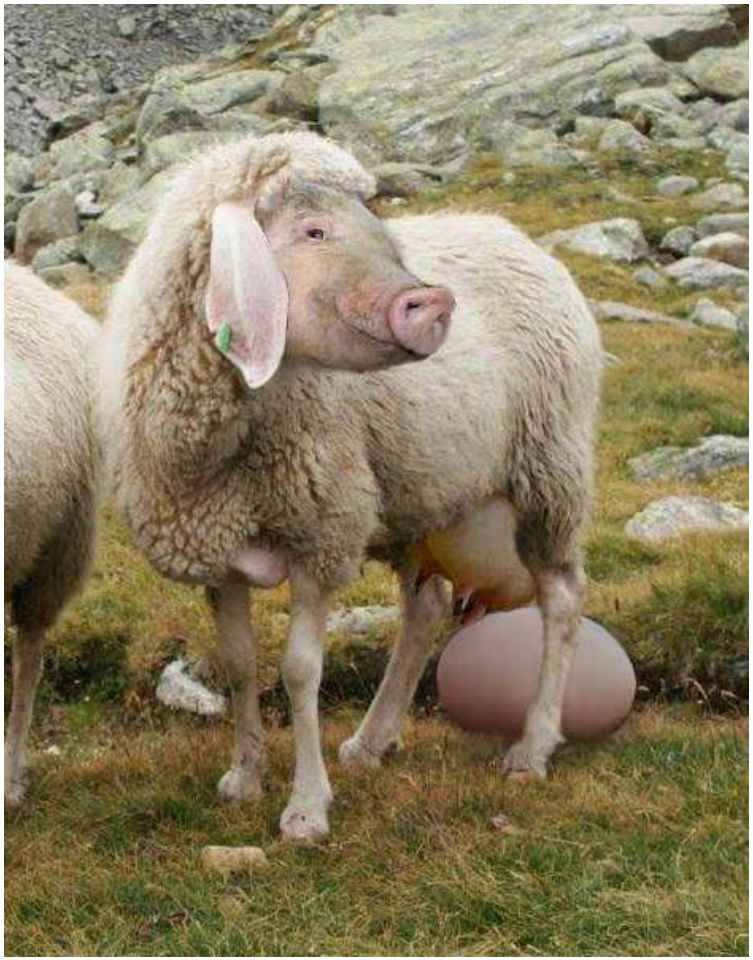
With pre-sorting of $\Theta(M)$-element subsets: $\mathcal{O}(N/B \cdot \log_2(N/M))$ I/Os.
7 Merge Phases.

Additionally with $\Theta(M/B)$-wise merging: $\mathcal{O}(N/B \cdot \log_{M/B}(N/M))$ I/Os.
1 Merge Phase.

Build on our previous achievements to beat the world record:

- Parallel-external sorting algorithms
- Multicore internal-memory sorting algorithms
- Flash disk performance tuning
- ...

Building an Energy-Efficient Sorting Machine



Source: www.zachseinblog.de/wp-content/uploads/2009/10/Wollmilchsau.jpg

Fast but still energy-saving concerning

- CPU/Cores
- Chipset
- I/O-Controller
- External Memory
- Cooling
- . . .

The ‘eierlegende Wollmilchsau’:

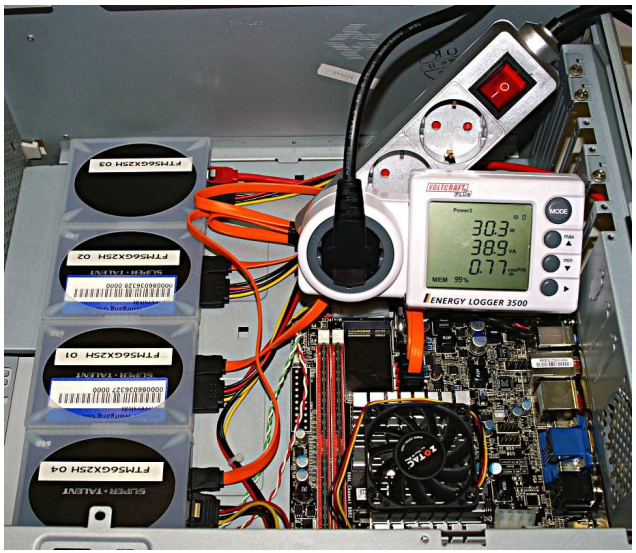
a balanced compromise !!

Side by Side . . .

Source: csl.stanford.edu/∼christos/pics/coolsort.png

Joulesort Winner 2007–2009
[Rivoire etal. (Stanford & HP Labs)]:

- Intel Core 2 Duo Mobile-CPU
- many notebook hard-disks



Our Approach [Beckmann etal.
(GU Frankfurt & KIT Karlsruhe)]:

- Atom N330 CPU
- few flash-disks (SSDs)

Side by Side – Details . . .

	Previous Record [Rivoire etal. 07]		Our Machine	
Component	**Type**	**TDP**	**Type**	**TDP**
Processor	CPU Intel Core 2 Duo T7600 Mobile-CPU	34 W	Intel Atom N330 2 cores, 4 threads	8 W
Memory	Kingston 2x1 GiB	4 W	Kingston 2x2 GiB	4 W
Board	Asus N4L-VM DH	n/a	Zotac IONITX-A	12 W
I/O	2x PCIe to SATA HighPoint Rocket RAID	12 W	4x SATA 3.0 Gibps onboard	–
Disks	13x Hitachi TravelStar 5K160 Notebook HDs	23 W	4x Super Talent FTM56GX25H SSDs	4 W
Fan		n/a		1 W
OS drive		–	USB-Stick	1 W
Estimated Total (net)				30 W
Estimated Total (overall)				37.5 W
Typical Idle (overall)		60 W		25 W
Typical Loaded (overall)		100 W		37 W

The Big Question . . .

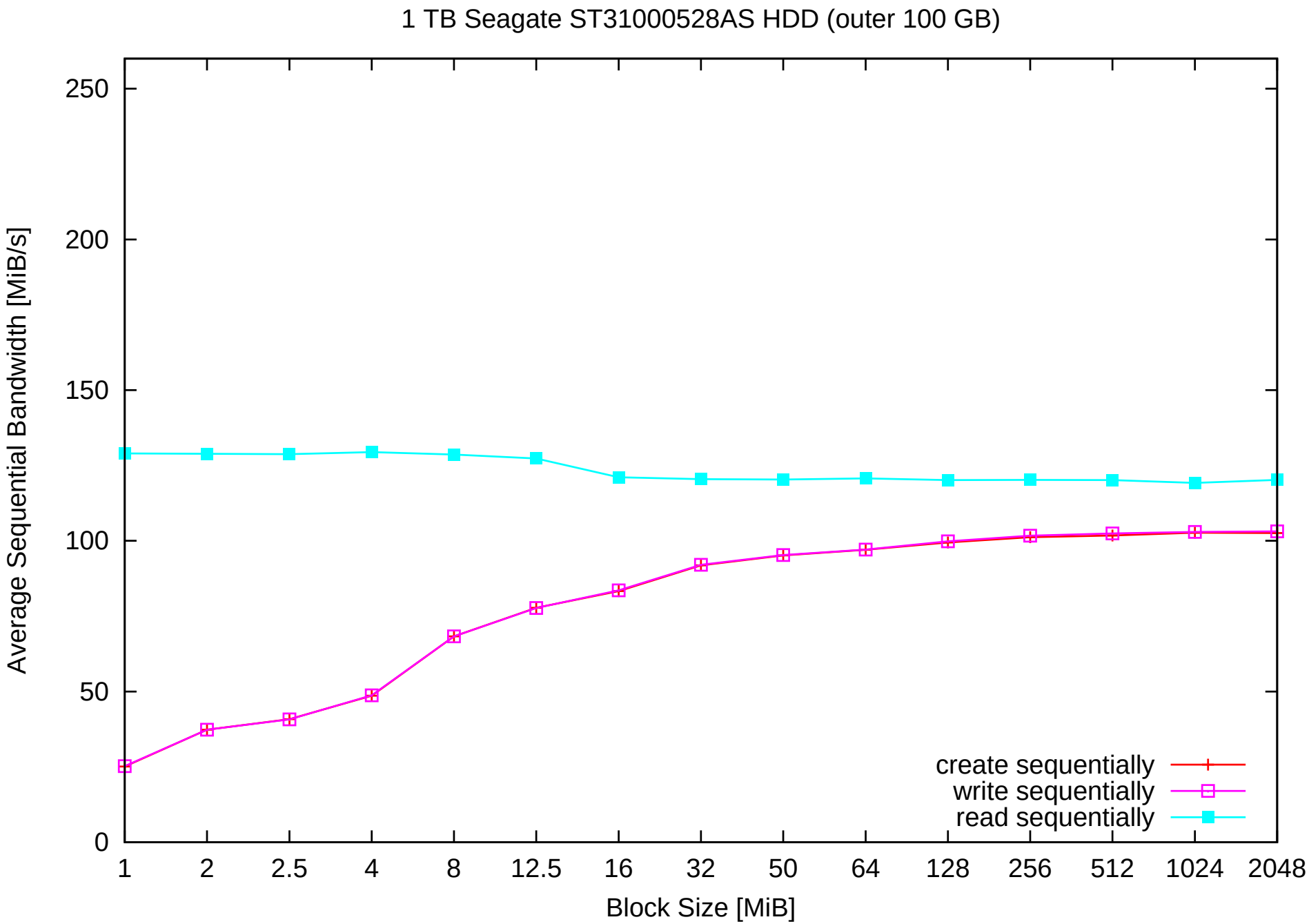
Our machine is much more energy-efficient, but will it also be much slower ???

Source: www.sxc.hu/photo/1155466

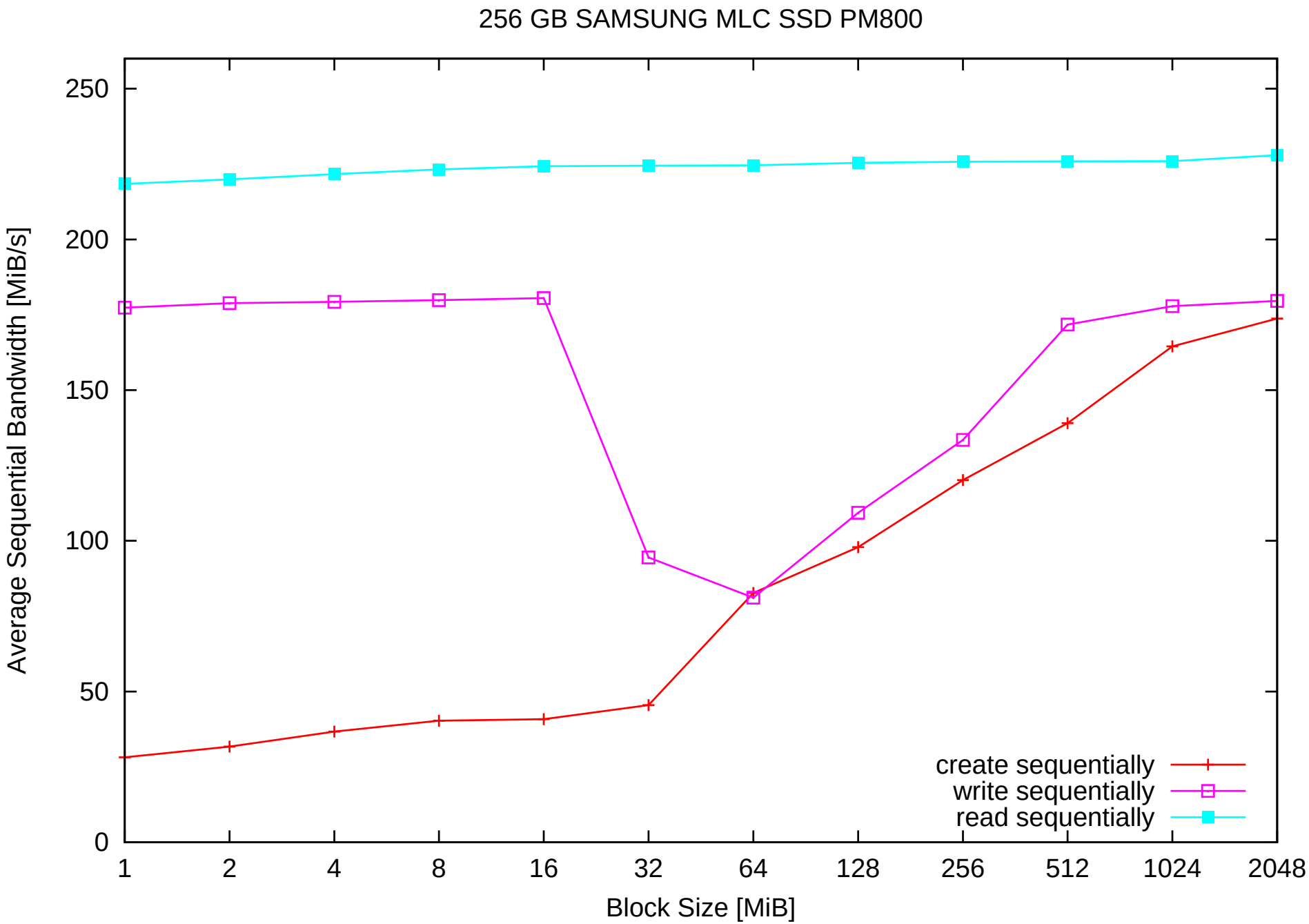
Central: Performance of
Solid State Disks vs. Hard Disks

- SSDs without mechanics
- rapid development cycles
- complicated controllers
- difficult to model/predict

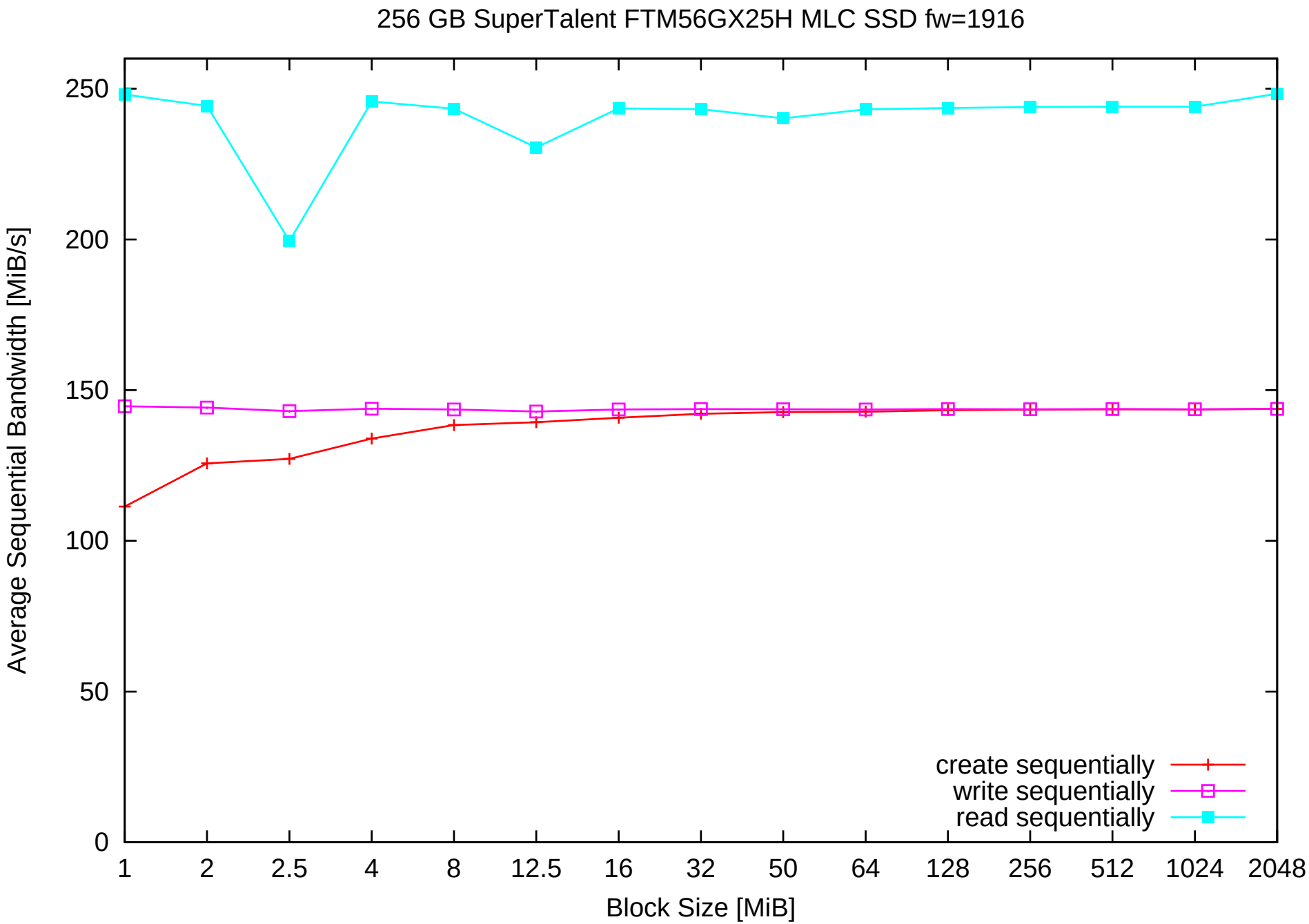
Performance of a Server Hard-Disk



Performance of a ’Problematic’ SSD



Performance of a Good SSD



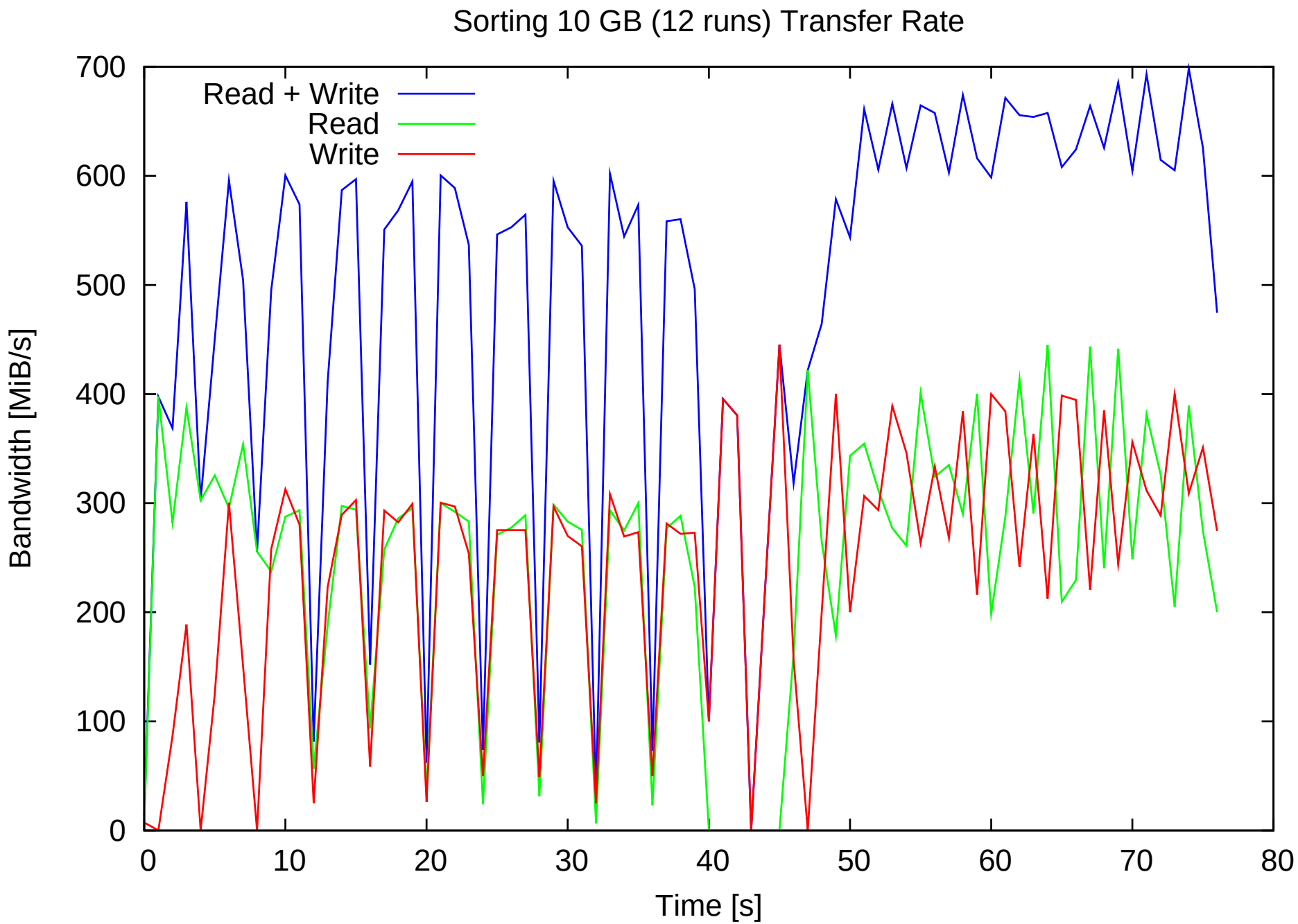
First Results SSDs

With the 'right' controller and under simple access patterns SSDs are much faster than hard-disks.

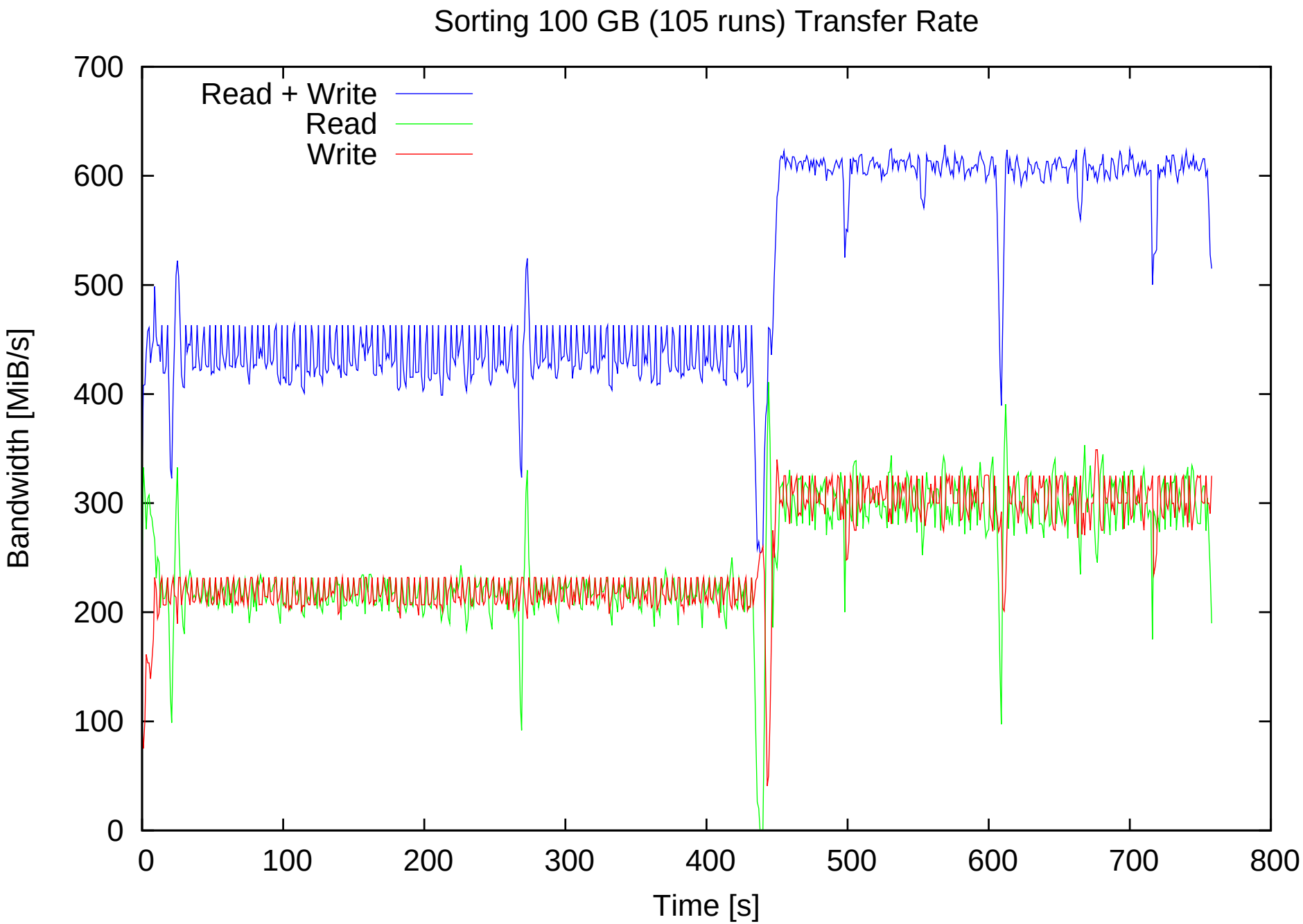
Open questions:

- What happens for mixed read-/write-requests in real algorithms?
- How do SSDs behave in a RAID?

SSD-Throughput for 10 GB Sorting [4-disk RAID, ∅: 1 sec]



SSD-Throughput for 100 GB Sorting [4-disk RAID, ∅: 4 sec]



Joulesort Results and Outlook

Category	Time	∅-Power	Energy	Elements/J	Improvement
10 GB	72 s	32.4 W	2.3 kJ	42 635	≥ 3.7
100 GB	691 s	36.3 W	25.1 kJ	39 853	≥ 3.4
1000 GB	17 026 s	33.6 W	571.8 kJ	17 489	≥ 5.1

We beat previous world records by factors ≥ 3.

Open problems:

- Energy-efficient algorithms for other problems.
- Cost metric 'acquisition + energy-consumption'.

25 KJ for 100 GB: Energy-Efficient Sorting.