Energy-Efficiency of Long-term Storage

Irina Tolokonnikova

Seminar "Energy-Efficient Programming"
Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

2015-01-14

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

informatik
die zukunft

Agenda

Archive

- storage of digital data for many years
- requirements:
  - preservation
  - retrieval
  - auditing
- archival data $\neq$ backup data
- needs to be cheap to obtain, cheap to operate, easy to expand
- high costs for energy consumption
  $\rightarrow$ room for improvement

Google

- How much data are we talking about?

Google

- How much data are we talking about?
  - DKRZ: $>$ 100 PetaBytes total capacity [1]

Google

- How much data are we talking about?
  - DKRZ: $>$ 100 PetaBytes total capacity [1]
  - Google: $\sim$ 15 ExaBytes (in 2013) $=$ 15000 Petabytes (only estimation)

Google

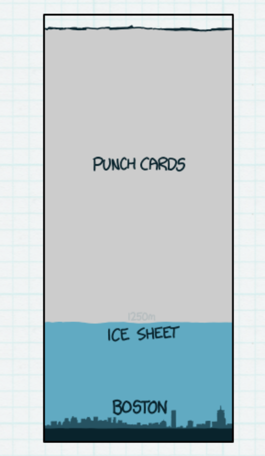


Figure: 15 ExaBytes of punch cards would be enough to cover New England, to a depth of about 4.5 kilometers

not this

Figure: LP [wikipedia.org]

not this

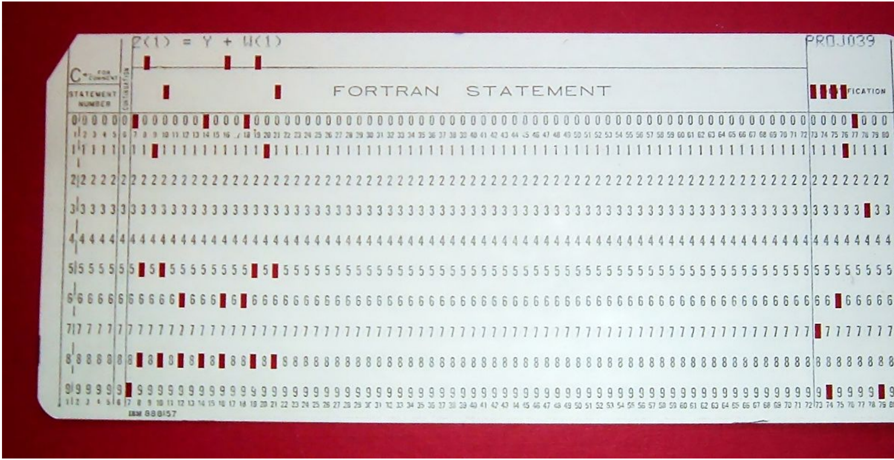


Figure: punch card [wikipedia.org]

not this

Figure: a United States National Archives Records Service facility in 1959. Each carton could hold 2000 cards [wikipedia.org]

not this

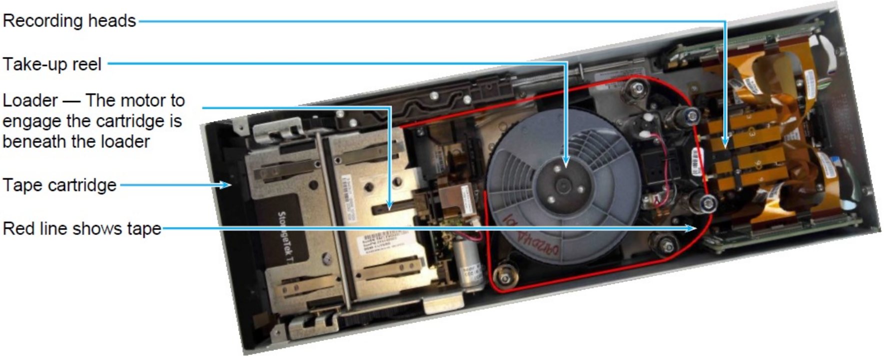
Figure: 3,5-inch floppy disk

not this?

Figure: compact cassette [wikipedia.org]

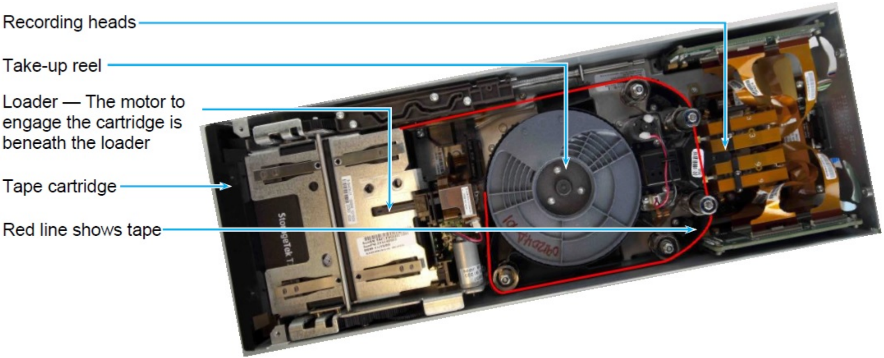
Tape

- used as a cartridge with a single reel
- holds several tens to thousands of GB (state wikipedia.org 13.01.15)



Tape

- used as a cartridge with a single reel
- holds several tens to thousands of GB (state wikipedia.org 13.01.15)
- Oracle StorageTek T10000 T2 hold 8,5 TB



DKRZ

- 7 automated Oracle/StorageTek SL8500 tape libraries
- 8 robots per library
- over 67000 slots for magnetic tape cassettes

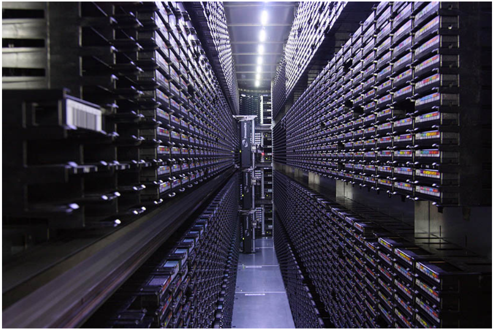


Figure: Inside the Tape library of DKRZ [1]

lifetime and costs

- lifetime: 30 years
- costs: less than 1 cent per GB
- 238X less energy over 12 years than HDD

Figure: upgrades save money and space [2]

pros and cons

Pros	Cons
cheap	needs special expensive equipment
long lifetime	sequential access pattern
no power needed when not accessed	

Hard Drives

- easy and fast to access data storage
- searching, consistency checking and inter-media reliability operations
- costs: 0.07 $per GB and falling
- lifetime: 10 years, but easy to break mechanics

pros and cons

Pros	Cons
easy access, simply system	needs much power, even when turned off
matches requirements of big data	easy to break
higher bandwidth (200X)	needs extra space for redundancy

Colarelli, Grunwald et al.(2002)

- massive array of idle disks = MAIDs
- aim: storage densities matching those of tape, with reduced energy consumption
- but operating same data volume in disks costs 10X more than in tape
- idea: use a cache manager to keep only part of disks in an array powered up
- varying spin-down delays

Results

- good trade off in performance and energy efficiency
- read performance still effected by the spin-down delay
- but 82% of read requests were satisfied by the cache
- least energy consumed with 4 sec spin-down delay

SSD

- costs: 0.66 $per GB , yet too expensive
- lifetime depends on usage, $\sim$10 years
- yet unclear, how unused data behaves on SSD
- coming soon?

Pergamum tomes by Storer et al. (2008)

- interfaces and protocols change slowly
- using inter- and intra-device redundancy
- work energy efficient, by not spinning up idle disks

→ intelligent, self managing storage device

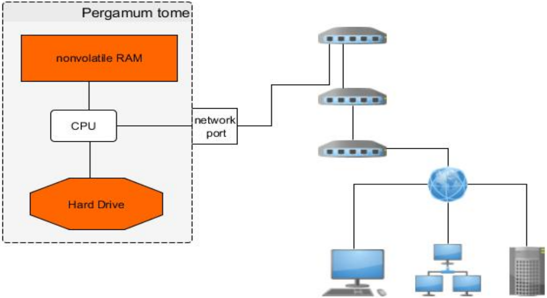


Figure: Pergamum tome, redrawn

Results of Pergamum

- size of the hard drive
- nonvolatile RAM handles many types of requests(e.g. hashes) without spinning up the disk
- using signatures for redundancy checking in entire inter-disk group
- using trees of hash values to reduce signature data
- once added to the network, the tome automatically joins a redunancy group or builds new one

→ makes storage management easier

- using intra-device redundancy, recovering from small errors without other devices
- aim to be price-competitive with tape

Problems and improvments

- still not included in data archives(?)
- redundancy overhead, but much energy saved
- "disposable" tomes
- encoding time 10X longer than on laptop processor BUT 10X less power consumed
- future work:
  - better algorithms
  - parallel processes (distributed searching)

A Spin-Up Saved is Energy Earned, Greenan et al.(2008)

- idea: use redundancies on active devices instead of waking up inactive ones

$\rightarrow$ Power aware coding

- three conditions needed:

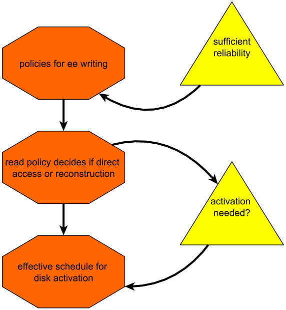


Figure: Three conditions for a power-aware system

Power Aware Techniques

- rules known from Pergamum tome
- Power Schedule
  - each code instance should have own write policy
  - write parallel across disk groups
- Power-Aware Read Algorithm
  - minimize the number of disk activations
  - first find out, if lost data is recoverable
  - like solving a matrix where inactive devices are treated as erased
- Disk Activation Algorithm
  - perform search to find best activation
  - how and when is a spin-down performed?

observation while testing

- mind the trade-off trilemma!

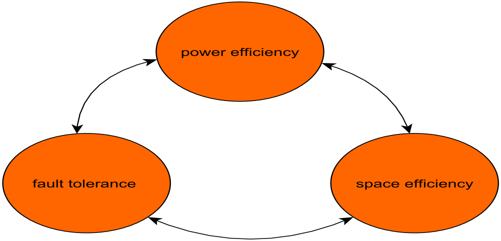


Figure: the trade-off when trying power aware coding

- open questions:
  - which enviroments will benefit from power aware coding?
  - how to find optimal policies?
  - robust metrics have to be developed for evaluation the power-reliability-performance trade-off

Conclusion

	Disk	Tape
Max shelf life (bit rot)	10 years	30 years
Best practices for data migration to new technology	3-5 years	8-12 years
Uncorrected Bit Error Rate, Probability (avg 1 error in x TB)	10^{-14} (~10's of TB)	10^{-19} (~1 million TB)
Power and cooling	238X	X

Figure: Disk compared to Tape [3]

Conclusion

- Pergamum tomes by Storer et al.
  - Pergamum tomes added to networks
  - redundancy overhead used to recover errors
  - energy saved by not spinning up other disks
  - self managing system with "disposable" nodes
- Power Aware Programming
  - try to use less disks as efficient as you can
  - mind the trade-off trilemma between fault tolerance, space efficiency and power efficiency
  - *"Initial results show that power-aware coding may be well suited for the write-once, read-maybe workload of long-term archival storage systems."*

How would you store...

- ...(your own) private medical data?
- ...research data of a medical study?
- ...data of all patients of a hospital?

References

[1] https://www.dkrz.de/Klimarechner-en/datenarchiv (13.01.2015)
[2] https://what-if.xkcd.com/63/ (13.01.2015)
[3] Dr. Mark L Watson: *Advanced Tape Technologies for Future Archive Storage Systems*. MSST - Media II (Tape Media and Libraries), 2013
[4] Colarelli, Dennis, Dirk Grunwald, and Michael Neufeld. *The case for massive arrays of idle disks (maid)*. The 2002 Conference on File and Storage Technologies. 2002.
[5] Storer, Mark W., et al. *Pergamum: Replacing tape with energy efficient, reliable, disk-based archival storage.* Proceedings of the 6th USENIX Conference on File and Storage Technologies. USENIX Association, 2008.
[6] Greenan, Kevin M., et al. *A Spin-Up Saved Is Energy Earned: Achieving Power-Efficient, Erasure-Coded Storage.* HotDep. 2008.