Analysis of the ECMWF Storage Landscape

Matthias Grawinkel*, Lars Nagel*, Markus Mäsker*, Federico Padua*, André Brinkmann*, Lennart Sorth#

* Johannes Gutenberg University Mainz, Germany

European Centre for Medium-range Weather Forecasts, Reading, UK

2015-02-17

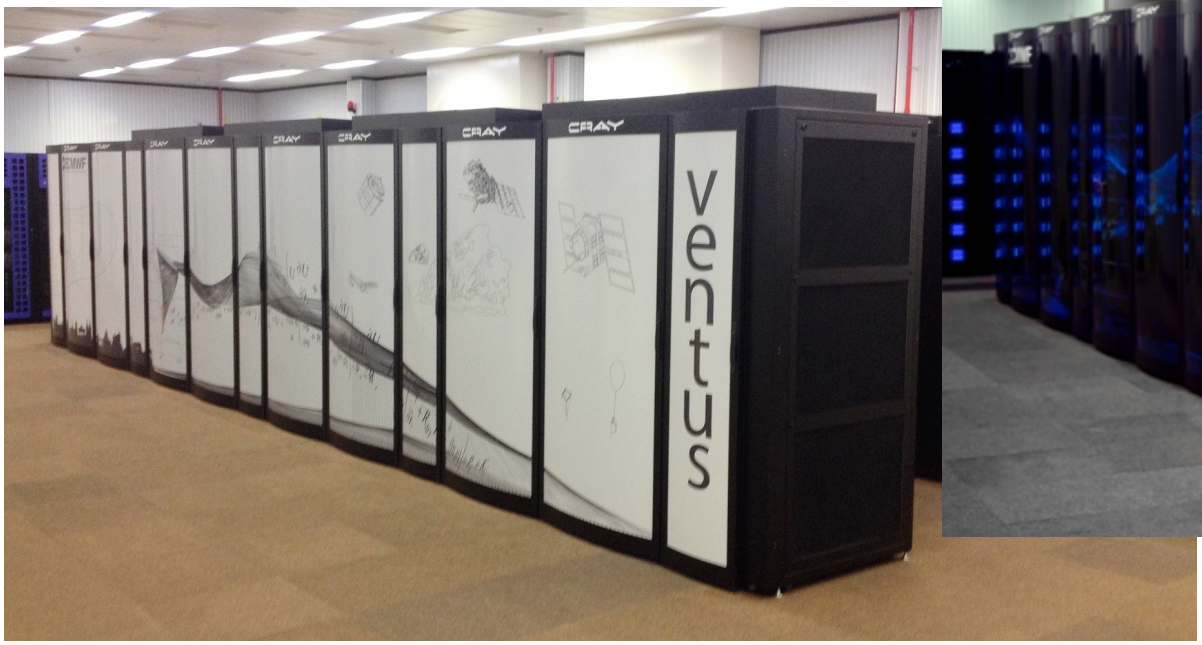
ECMWF

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

European Centre for Medium-range Weather Forecasts

- Global weather forecasts for up to 15 days and seasonal forecasts for up to 12 months
- Multiple supercomputers (Top 500 Nov. 2014: 28, 29, 82, 83)
- ~100 PB total storage capacity in 2014/09
- Two in-house developed data handling systems: ECFS, MARS
- Compound annual growth rate (CAGR) > 50%



Motivation

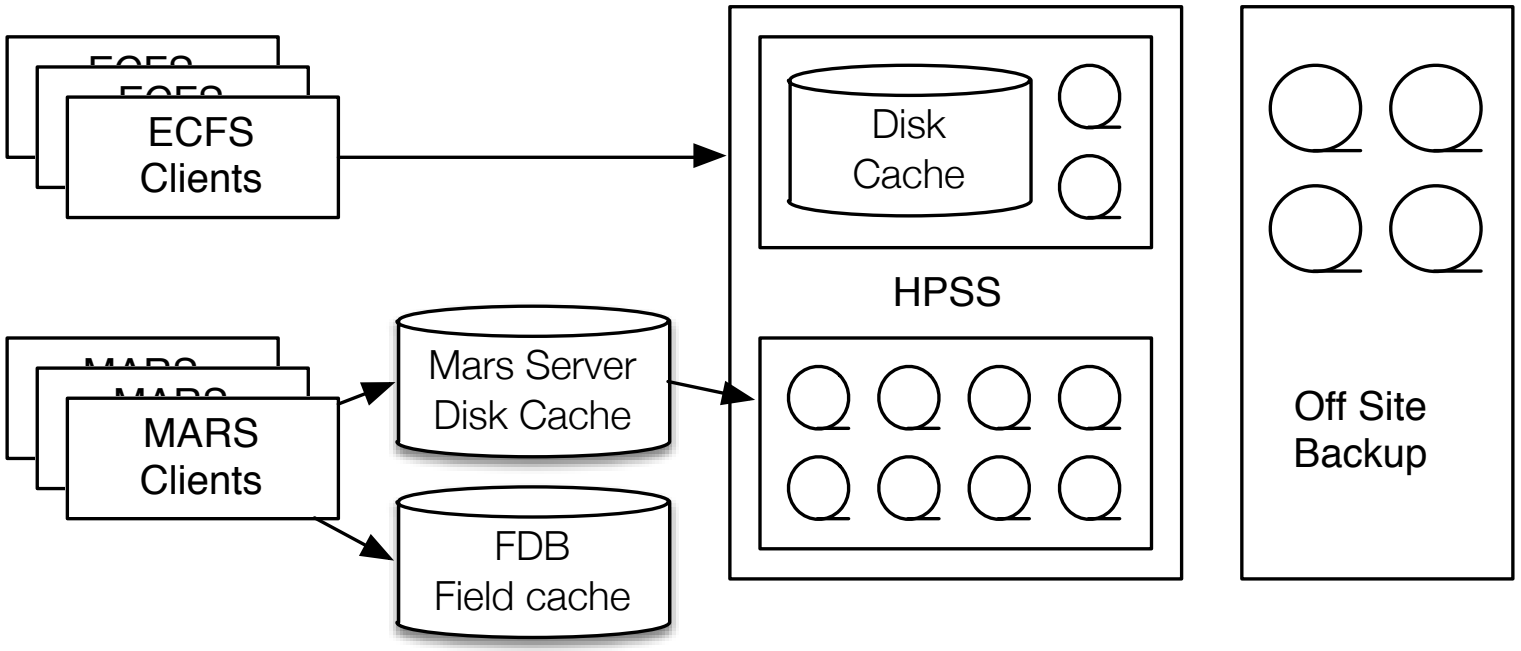
- How to build (active) archives?
  - Content & behavior of existing systems
  - Current problems?
  - Future challenges?
- Only a few studies and traces available
  - Low coverage of the research topic
  - Required to design and evaluate systems

→ First study of large-scale active archive

  - In depth-analysis of two systems
  - Characterization of content and usage
  - Analysis of caching behavior
  - Study of tape backend
  - Release of scripts & trace files

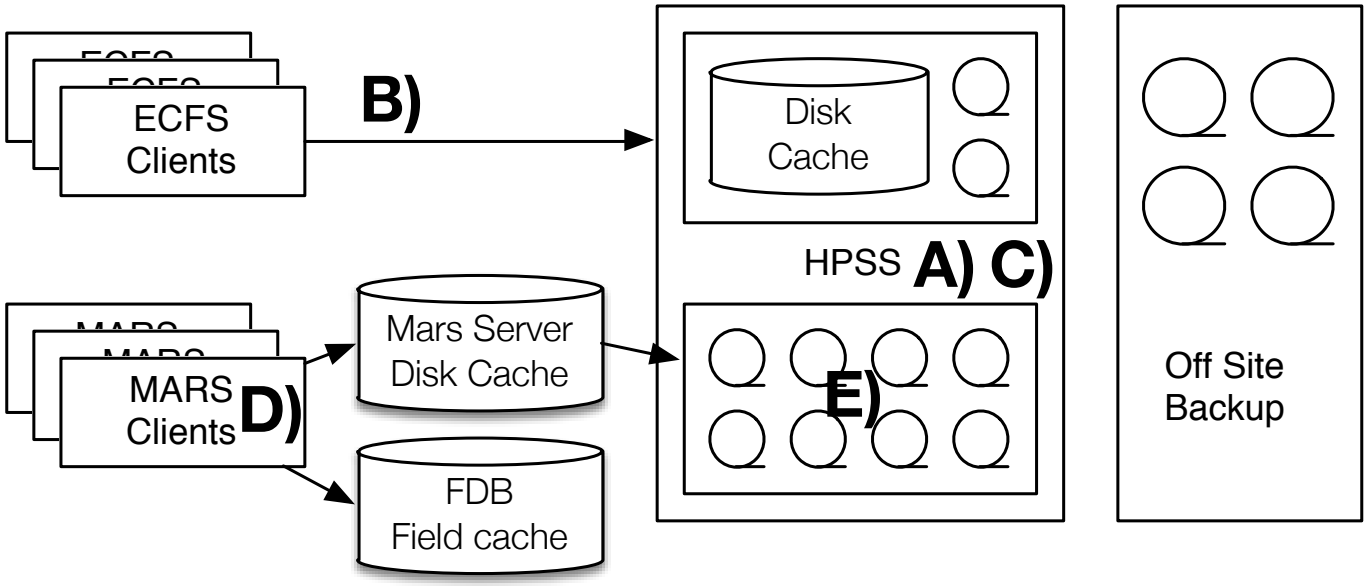
ECMWF Storage Landscape

- *ECFS* is a general purpose user accessible archive for intermediate and long-term file storage
- *MARS* is an object database for meteorological data



- Files are staged and cached on disk drives
- Every file eventually has a primary copy on tape
- Important files have secondary tape copy

Investigated Trace Files



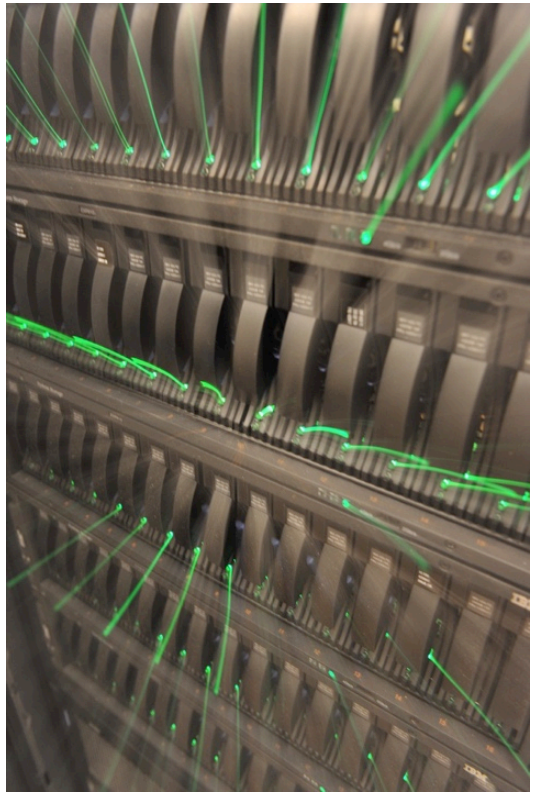
- A) ECFS / HPSS database snapshot of 2014/09
- B) ECFS access trace: 2012/01 - 2014/05
- C) MARS / HPSS database snapshot of 2014/09
- D) MARS feedback logs: 2010/01 - 2014/02
- E) HPSS WHPSS logs / robot mount logs: 2012/01 - 2013/12

- Extracted, sanitized, and obfuscated traces available now

Data Handling System: ECFS

- Client tools for PUT, GET, DEL, RENAME on full files
- 14.8 PB of primary data
- 137.5 mil. files in 5.5 mil. directories
- 0.34 PB disk cache (disk/tape ratio: 1:43)
  - Cache categories defined by file size

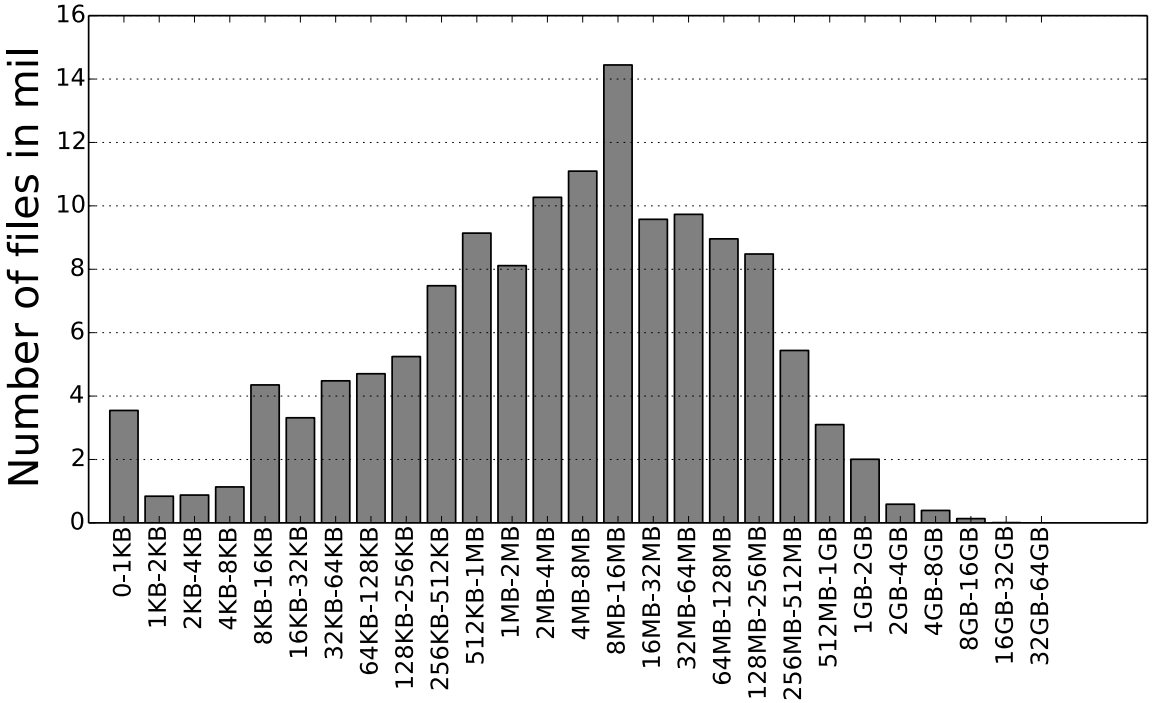
Group	From	To (incl.)	Count	Used Capacity
Tiny	0	512 KB	36.0 mil.	4.4 TB
Small	512 KB	1 MB	9.1 mil.	6.3 TB
Medium	1 MB	8 MB	29.5 mil.	101 TB
Large	8 MB	48 MB	30.0 mil.	585 TB
Huge	48 MB	1 GB	29.7 mil.	6.2 PB
Enormous	1 GB	32 GB	3.1 mil.	8 PB



ECFS Content Characterization

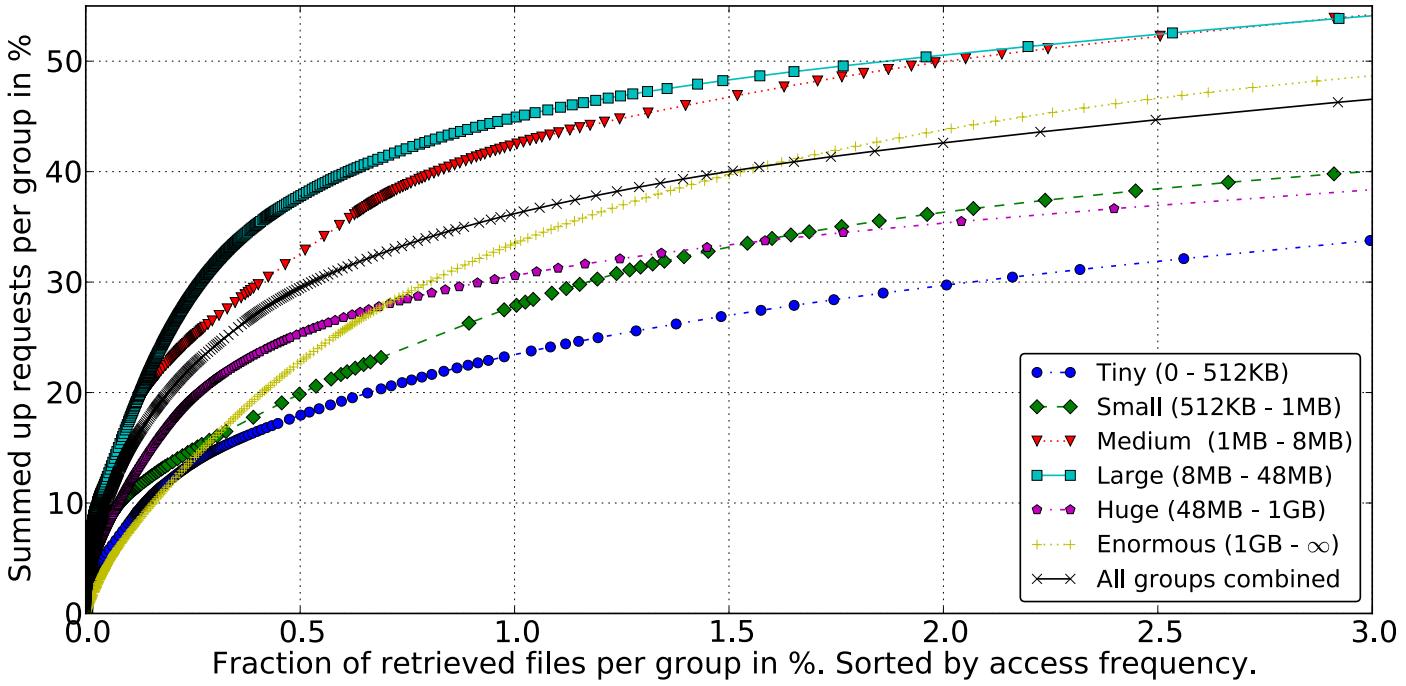
- Based on HPSS database snapshot – 2014/09
  - Only 26.3 % of files on tape were ever read ≥1 times

By file count	By used capacity
unknown (27.8%)	unknown (39.3%)
.gz (20.4%)	.tar (21.3%)
.tar (7.8%)	.gz (12.5%)
.nc (7.6%)	.nc (7.9%)
.grib2 (1.9%)	.lfi (2.2%)
.raw (1.7%)	.pp (1.0%)
.txt (1.5%)	.sfx (0.9%)
.Z (1.5%)	.grb (0.8%)
.bufr (1.4%)	.grib (0.4%)
.grb (1.4%)	.bz2 (0.3%)



ECFS Workload Characterization

- Timespan 2012-01-01 to 2014-05-20
  - 78.3 mil. PUT requests → 11.8 PB
  - 38.5 mil. GET requests → 7.2 PB
    - 12.2 mil. unique files (9% of full file corpus)
    - Cache hit ratio by requests: 86.7%
    - Cache hit ratio by bytes: 45.9%



ECFS User Sessions

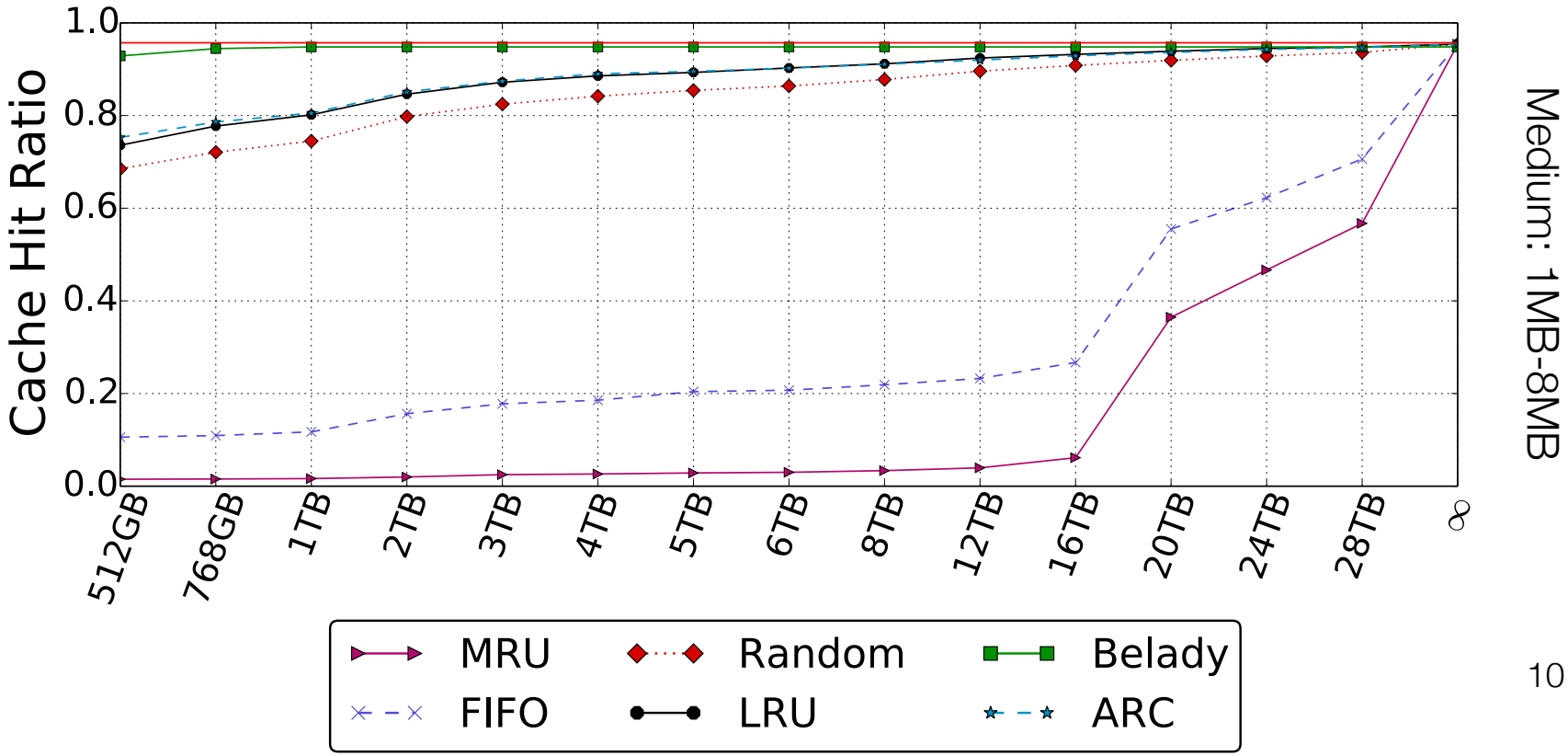
- Identified 1,190 users, 2.7 mil. sessions
- Session lifetime from seconds up to 10 hours of constant traffic

Key	Count	5th P	mean	99th P
Total #actions per session	2.7 mil.	2	47	579
Sessions with GET requests	1.1 mil.	1	36	571
- Retrieved data		0.6 MB	7.2 GB	86 GB
- #ReGET requests	0.13 mil.	1	32	442
Sessions with PUT Requests	2.3 mil.	1	34	373
- Uploaded data		0.02 MB	5.6 GB	65 GB

11% of GET requests within a session are re-retrievals of a file

What is the impact of smaller or bigger caches?

- Developed modular cache simulation environment
  - MRU, FIFO, RANDOM, LRU, ARC, Bélády, ECMWF baseline
  - Cache per size-category (capacity + strategy)
- Replayed ECFS access trace
  - 12 months warm up, measured following 17 months

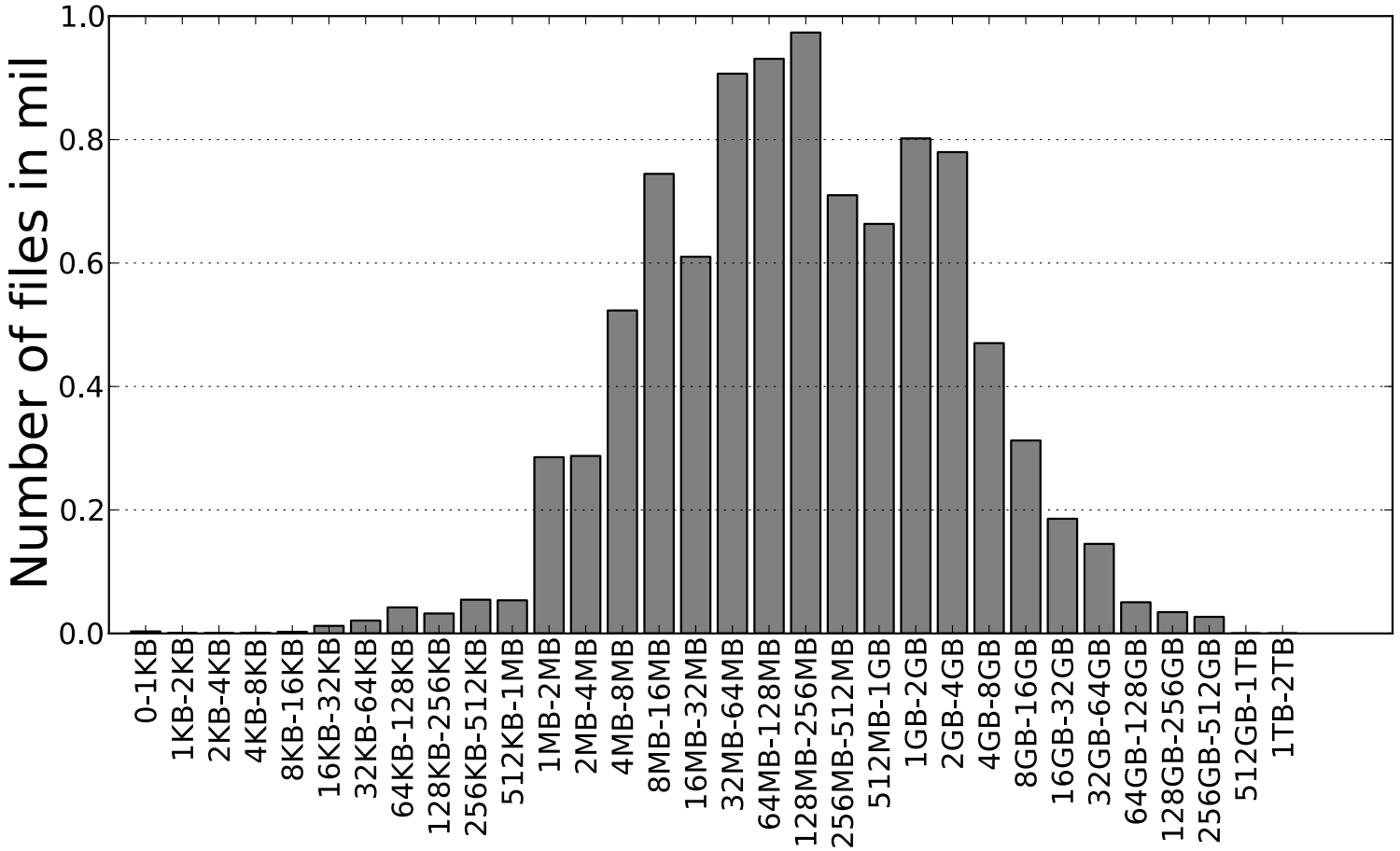


Data Handling System: MARS

- Query: “Get temperature & humidity for Santa Clara from $date till $date with a 5 minute resolution”
  - MARS then assembles and writes out a results file
- 170 bil. fields in 9.7 mil. files
  - 200 mil. new fields each day (i.e. sensor data, model output)
- 37.9 PB of primary data, 800 GB metadata
- 3-tiered caching hierarchy
  - Field database (FDB) on HPC storage: Variable size <1 PB
  - 1 PB disk cache on MARS servers
    - 250 TB reserved for manual optimizations
  - HPSS/tape

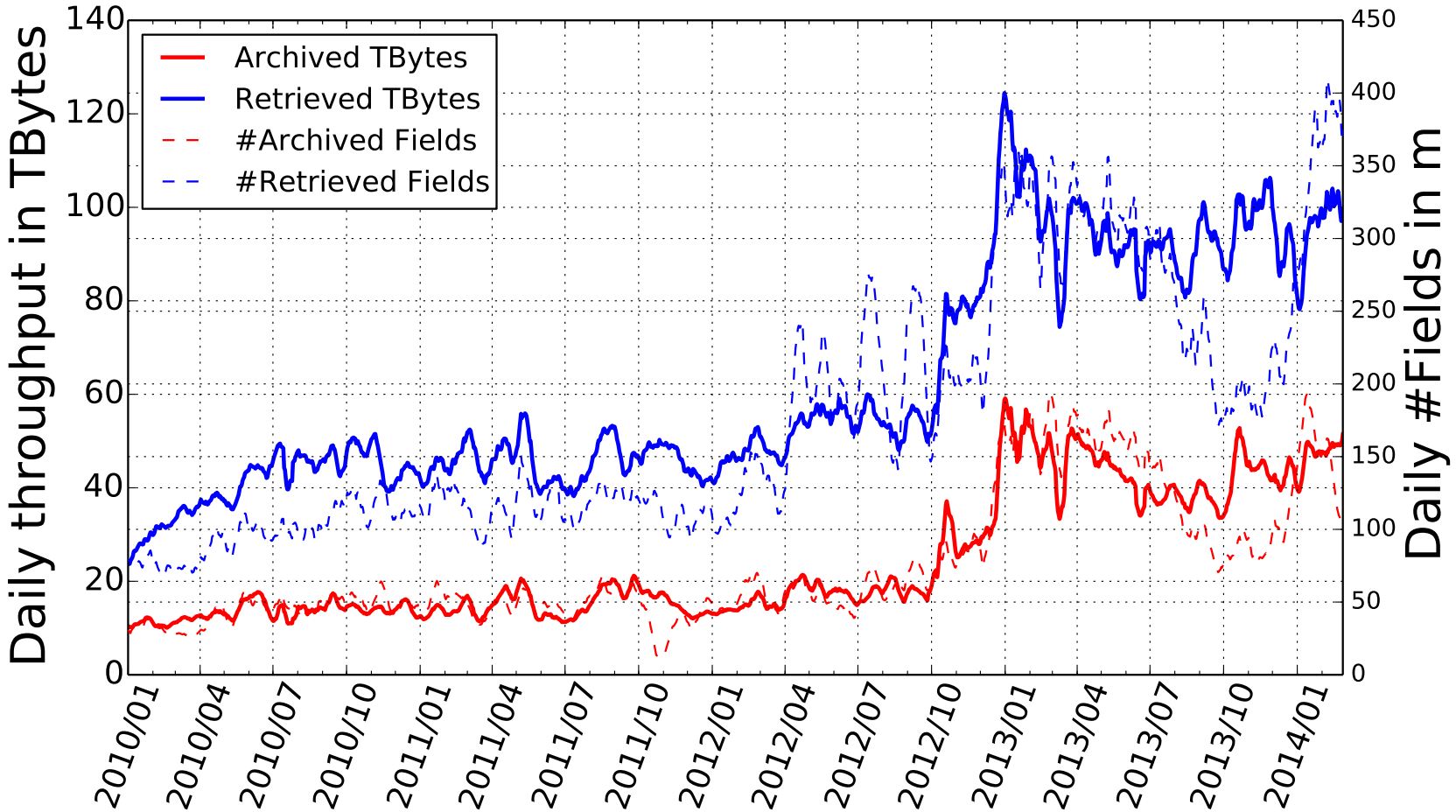
MARS Content Characterization

- Based on HPSS database snapshot 2014/09
  - Only 23 % of files on tape were ever read ≥1 times



MARS Workload Characterization

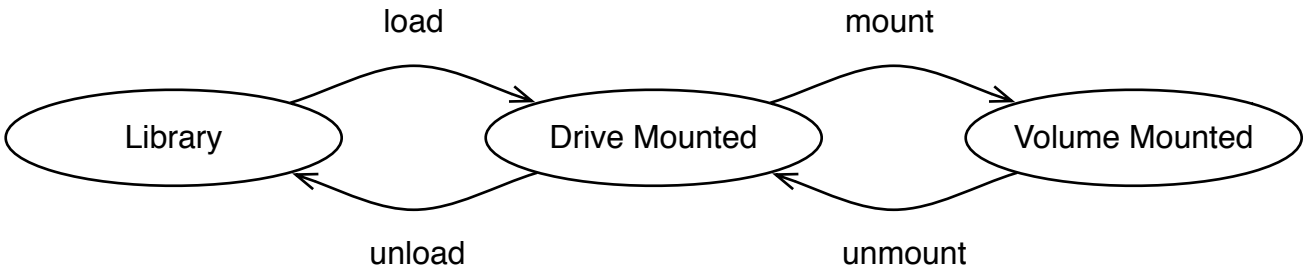
- MARS feedback logs from 2010-01-01 till 2014-02-27
  - Contain queries and description of results (#fields, bytes, source)

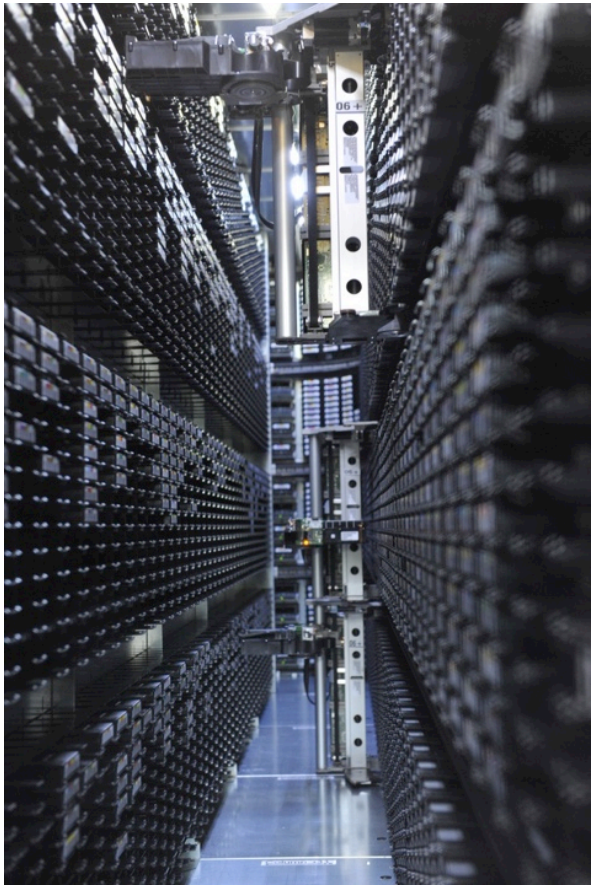


Totals over Observed Timeframe (1518 days)

Total archive requests	115 mil.
Total archived bytes (fields)	35.9 PB (114.7 bil.)
Total retrieve requests	1.2 bil.
- involving ≥ 1 tape loads	25.3 mil. (2.2%)
- from HPSS/tape only	16 mil. (1.4%)
Total retrieved bytes (fields)	91.6 PB (269 bil.)
- from FDB bytes (fields)	54.2 PB (212 bil.)
- from MARS/disk bytes (fields)	29.4 PB (43.3 bil.)
- from HPSS/tape bytes (fields)	8 PB (13.3 bil.)

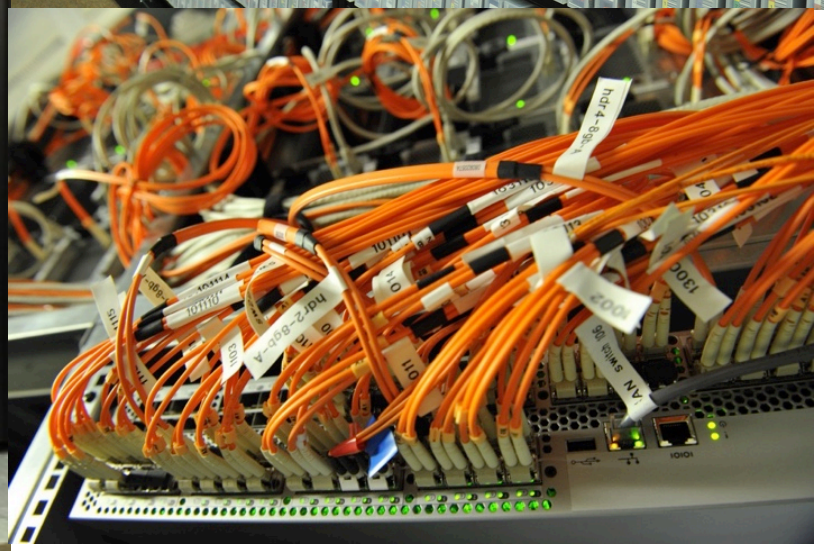
Tape Mount Logs





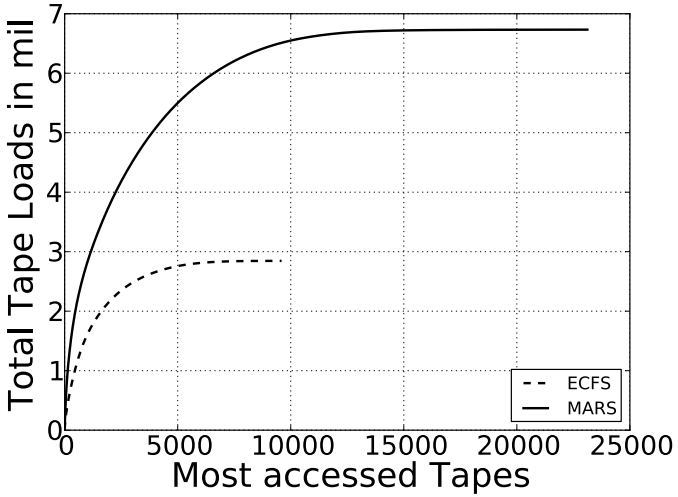
Tracked Tapes & Drives: 2012+2013

32,712 tape identifiers
231 drive identifiers
9.6 mil. tape loads
~9 loads per minute



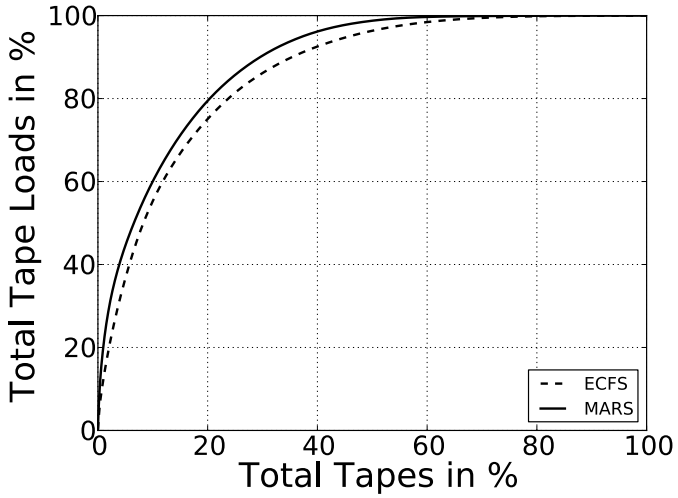
Tape is Actively Used

Tape mount frequencies				
	#tapes	median	mean	99th P
MARS	23,118	46	291	3,351
ECFS	9,594	85	297	2,470

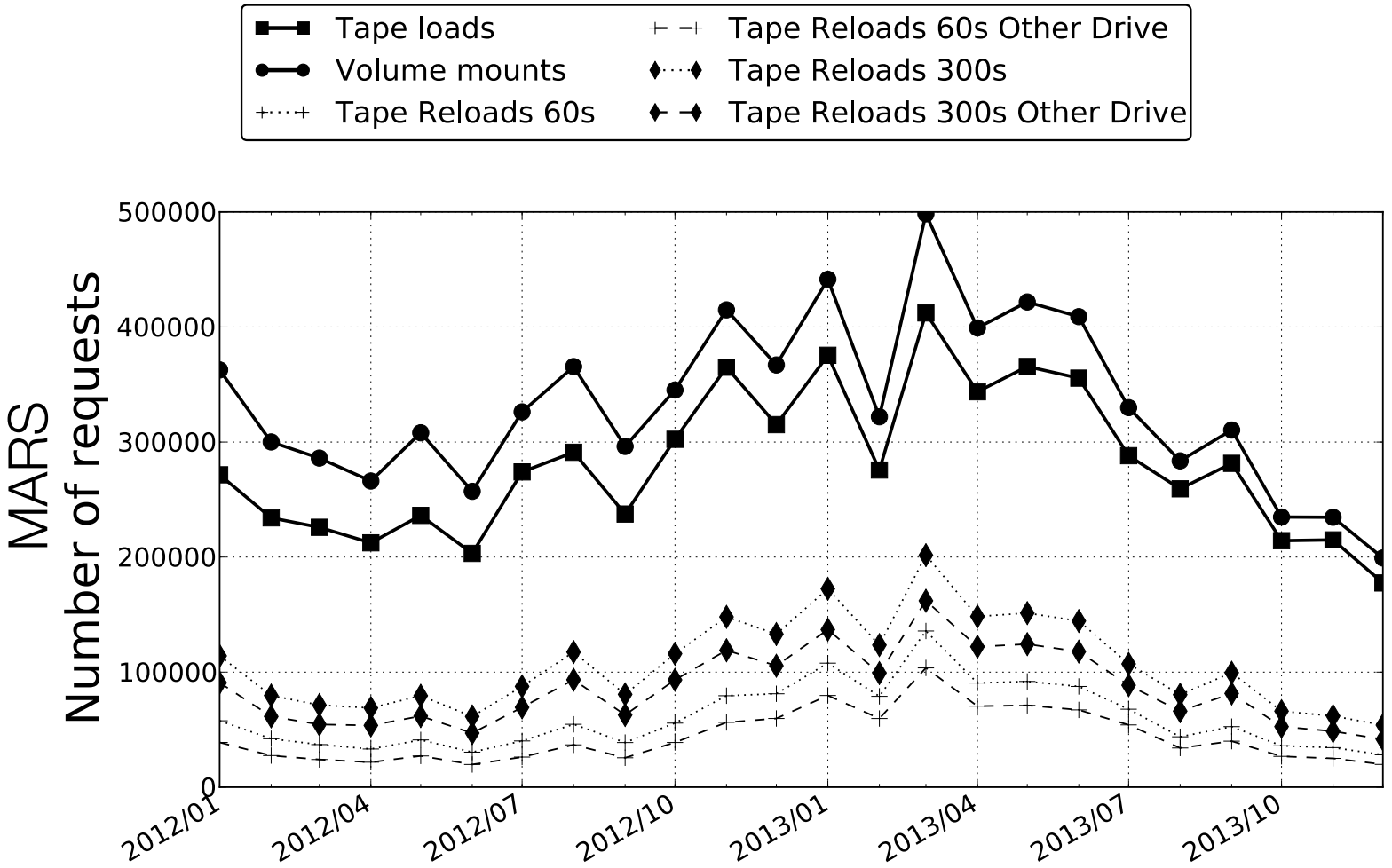


Mount requests per cartridge.
Absolute (top) + normalized (bottom)

Tape mount latencies in seconds				
	#mounts	median	mean	99th P
MARS	6.7 mil.	35	54.4	262
ECFS	2.8 mil.	32	48.2	257



Remounts and Reloads

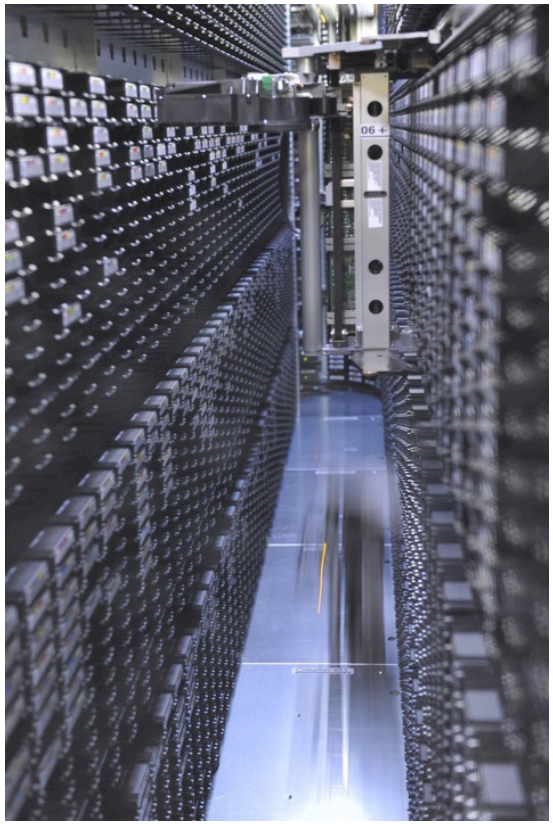


14.8% of all loaded tapes were unloaded from another drive less than 60 seconds ago

Improve Tape (Un)loading?

- Goals: Minimize #drives, mean time to mount, tape mounts
  Maximize tape re-use
- Identified hot tapes: 20% of tapes account for 80% of mounts
- Analysis of drive utilization showed exploitable idling times
- Optimistic preloading?
  - Correlation analysis showed potential
- High tape reload rates suggest to keep (certain) tapes in the drives

→ Further investigation required



Conclusion

- ECMWF in operation since 1975
  - Lots of hands-on experience
  - Predictable production workloads
    - Manual optimizations
  - Chaotic research workloads …
- ECFS resembles archives investigated in related work
- MARS opens a new category of archives
- Tape + disk caches can be used to build efficient non-interactive systems
- Heavy use of tape has drawbacks
  - High wear-out
  - Unpredictable, stacking latencies
    - MARS-Error: Query requires too many tapes
- Potential for smarter tape (un)loading strategies

github.com/zdvresearch/fast15-paper-extras/

We‘re hiring: research.zdv.uni-mainz.de

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

ECMWF