Simulation and Performance Analysis of the ECMWF Tape Library System

JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Markus Mäsker, maesker@uni-mainz.de
Johannes Gutenberg University Mainz
15.11.2016

SC16
Salt Lake City, Utah | hpc matters.

JG|U

Misconception No. 1: Tape is irrelevant!

Misconception No. 1

Tape is irrelevant!

> *Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King*
> *Jim Gray, Dec. 2006*

Misconception No. 1: Tape is irrelevant!

Tape advantages:

- 30+ years media life
- Lowest bit error rate
- Capacity
- Sequential I/O
- Energy efficiency
- Total Cost of Ownership (TCO)

Tape disadvantages:

- Random Access:
  latency of multiple **minutes** common
- WORM media
  (write once, read many)

Tape is reliable, cheap, but very slow at random reads.

Use case: cold data, such as backup and archives

JG|U

Misconception No. 2: Tape $\Leftrightarrow$ slow $\neq$ fast $\Leftrightarrow$ HPC

Misconception No. 2

Tape is far too slow for supercomputing environments.

Misconception No. 2: Tape $\Leftrightarrow$ slow $\neq$ fast $\Leftrightarrow$ HPC

Ken Batcher

"A supercomputer is a device for turning **compute-bound** problems into **I/O-bound** problems."

The HPC community always had to deal with slow I/O devices

- Batch processing systems
- Tiered storage architectures

$\Rightarrow$ Tape-based cold storage tier widely used

Misconception No. 3: Research has been done!

Misconception No. 3:

There is nothing left to discover for academic research.

JG|U

Misconception No. 3: Research has been done!

Most recent academic research papers on robotic tape libraries:

- Performance Analysis Of Tape Libraries For Supercomputing Environments, Ilker Hamzaoglu and Huseyin Simitci, **HPCS 1999**
  - ⇒ Seek time on tape is the bottleneck
  - ⇒ Robot fetch time with little impact
- Performance Measurements of Tertiary Storage Devices, Theodore Johnson and Ethan L. Miller, **VLDB 1998**
  - ⇒ Fetch times small and nearly deterministic
  - ⇒ Complex optimizations of tape placement not necessary

Misconception No. 3: Research has been done!

Although Johnson and Miller speculate:

> *"More complex systems involving multiple tape racks, robot arms, and pass-through slots might show a more complex behavior."*

Today:

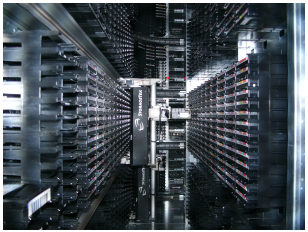
⇒ Robot fetch time is a relevant factor.
⇒ Considerable improvements can be achieved.

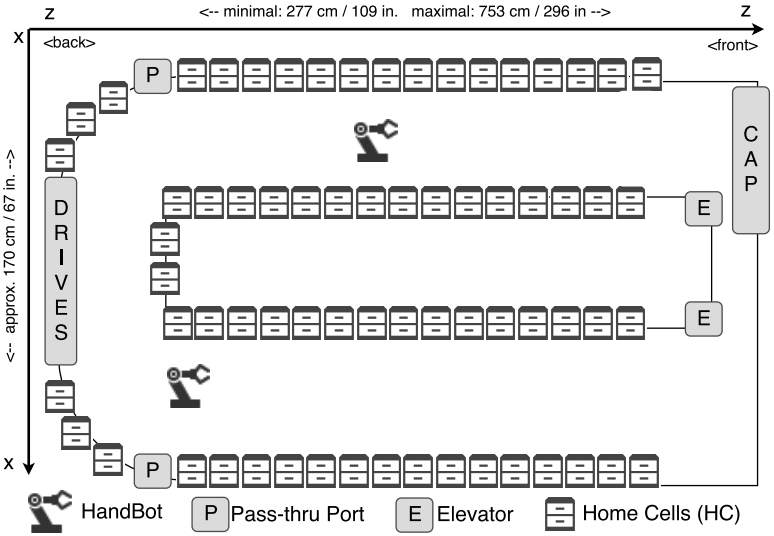
Robotic Tape Libraries

A Quick Introduction to Robotic Tape Libraries ...

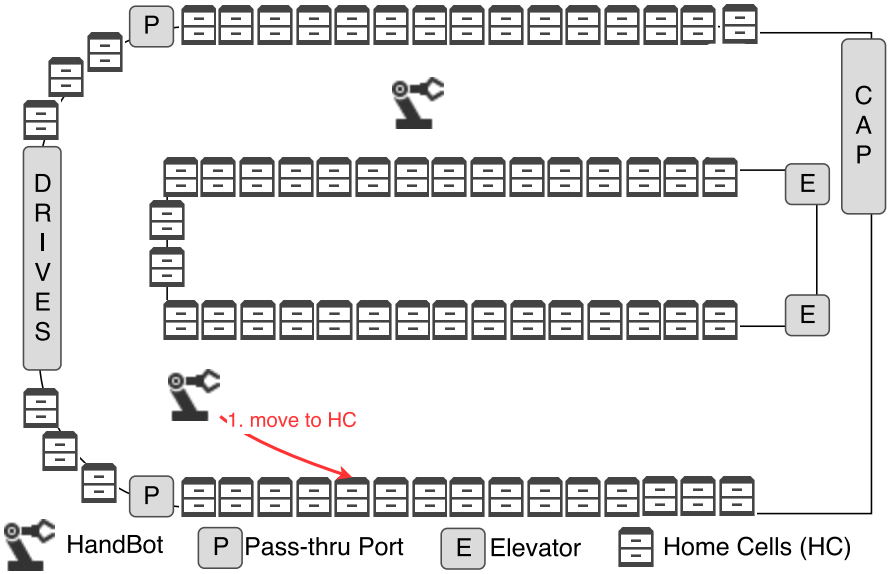
... based on the Oracle/StorageTek StreamLine 8500

Oracle/StorageTek SL8500 and a Library Storage Module (LSM) sketch (topview)

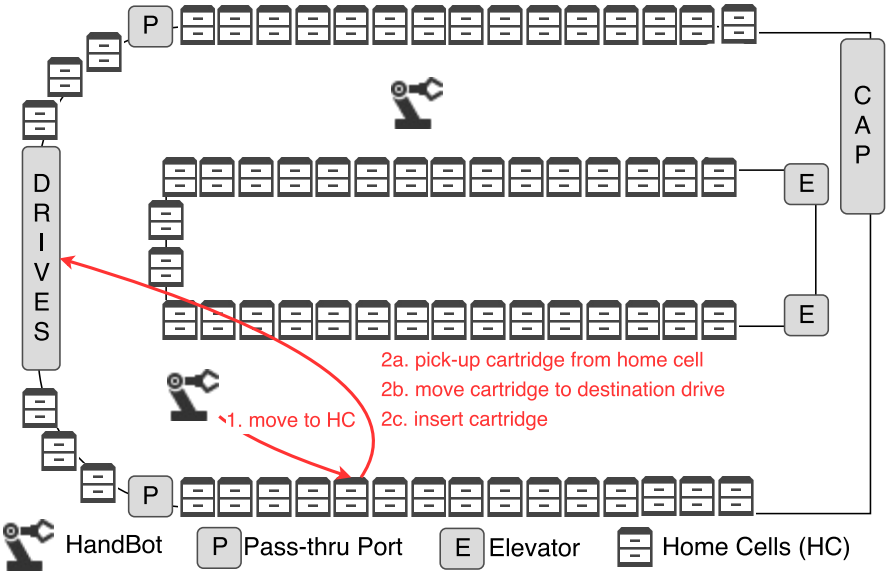




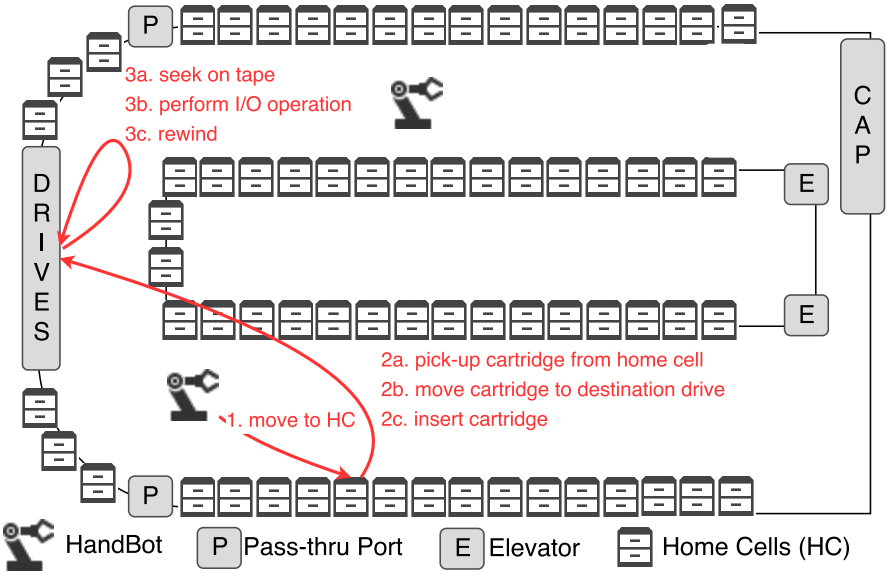
Phases of one cartridge I/O request



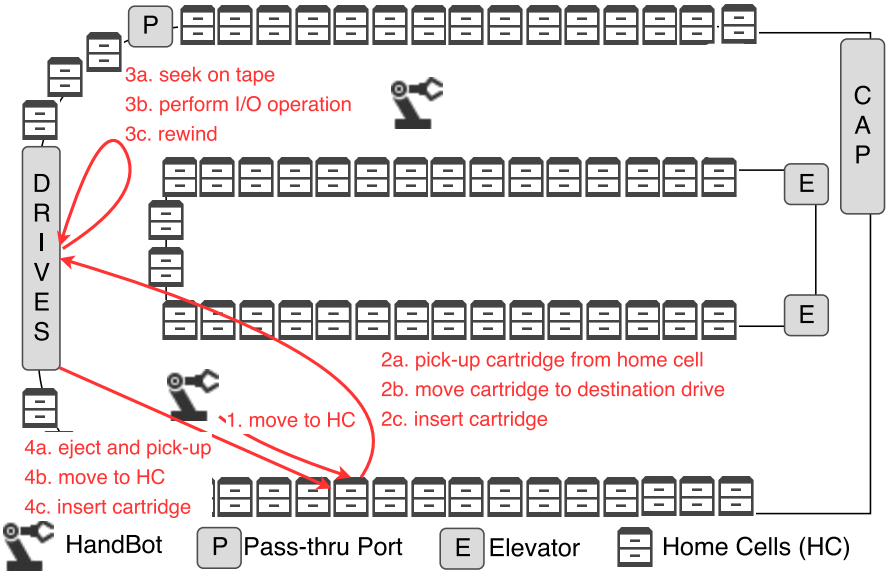
Phases of one cartridge I/O request



Phases of one cartridge I/O request



Phases of one cartridge I/O request

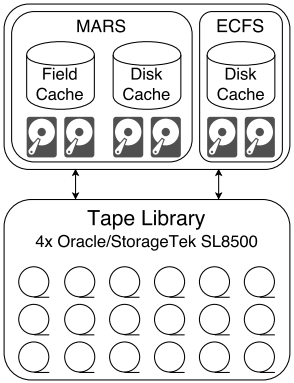


JG|U

ECMWF Storage Landscape (2014)

European Centre for Medium-Range Weather Forecasts (ECMWF)

- 4x Oracle/StorageTek SL8500 Tape Libraries
- HPSS - Hierarchical Storage Management
- 100 PB storage capacity
  - ECFS: user archive
    14.8 PB on tape, 340 TB disk cache
  - MARS: meteorological data
    37.9 PB on tape, 1 PB disk cache
- 45 % compound annual growth rate (CAGR)
- So-called *active archive*



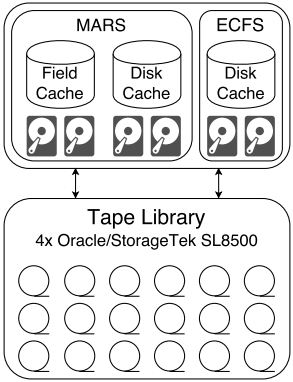
For more details:

Analysis of the ECMWF Storage Landscape, FAST 2015,
Matthias Grawinkel, Lars Nagel, Markus Mäsker, Federico Padua, André Brinkmann, Lennart Sorth

ECMWF Storage Landscape (2014)

ECMWF robotic tape library trace:

- September 2011 until April 2014
  Warm-up: until Dec. 2012
  Evaluation: starting Jan. 2013
- 15.2 million cartridge accesses:
  Cartridge id, time stamp, duration
- 231 unique tape drive identifiers
- 32,712 unique cartridge identifiers



For more details:
Analysis of the ECMWF Storage Landscape, FAST 2015,
Matthias Grawinkel, Lars Nagel, Markus Mäsker, Federico Padua, André Brinkmann, Lennart Sorth

Evaluation Sections

Evaluation:

Part 1: Analysis of the strategies currently in place
Part 2: Two new strategies
Part 3: Analysis of two different tape library setups

Part 1: Defer Dismount (DD) parameter

Defer Dismount (DD) parameter:

- Delay cartridge eviction for **X seconds** after I/O operation completes
  - Earlier eviction permitted if necessary
- Core parameter to tweak robot behavior in HPSS

Part 1: Effects of the Defer Dismount parameter

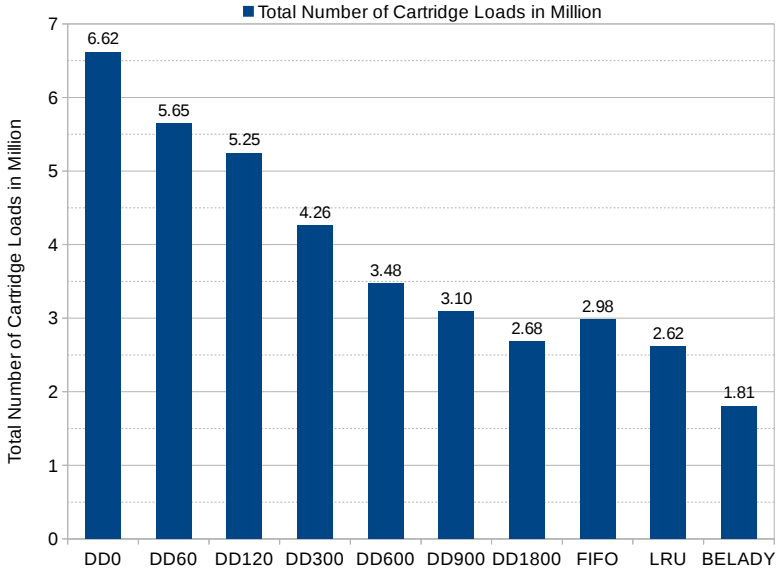
Simulation strategies

DD<X>: Defer dismount parameter of X seconds

FIFO: Evict cartridges in first-in / first-out order

LRU: Evict cartridges in least recently used order

BELADY: Evict that cartridge who's next access is the farthest in the future (theoretical optimum)



Part 1: Effects of the Defer Dismount parameter

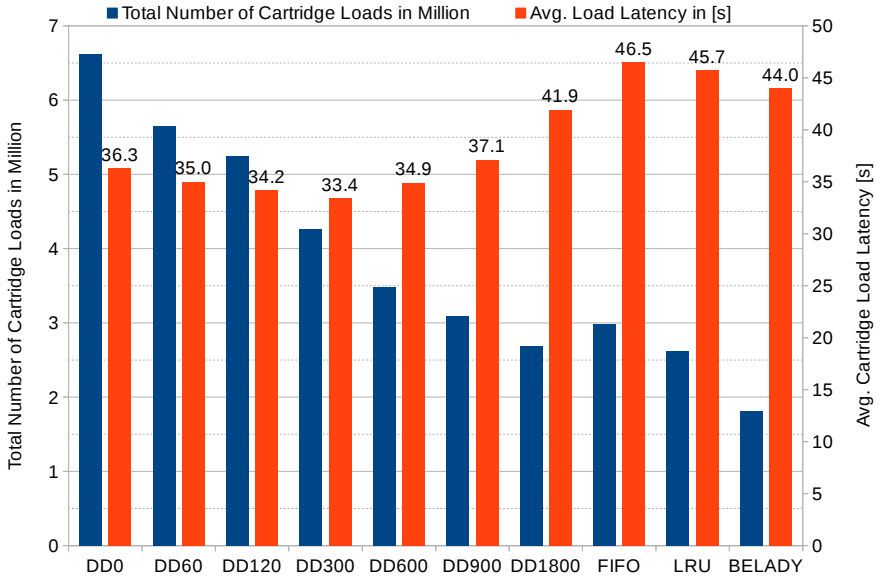
Simulation strategies

DD<X>: Defer dismount parameter of X seconds

FIFO: Evict cartridges in first-in / first-out order

LRU: Evict cartridges in least recently used order

BELADY: Evict that cartridge who's next access is the farthest in the future (theoretical optimum)



Part 1: Effects of the Defer Dismount parameter

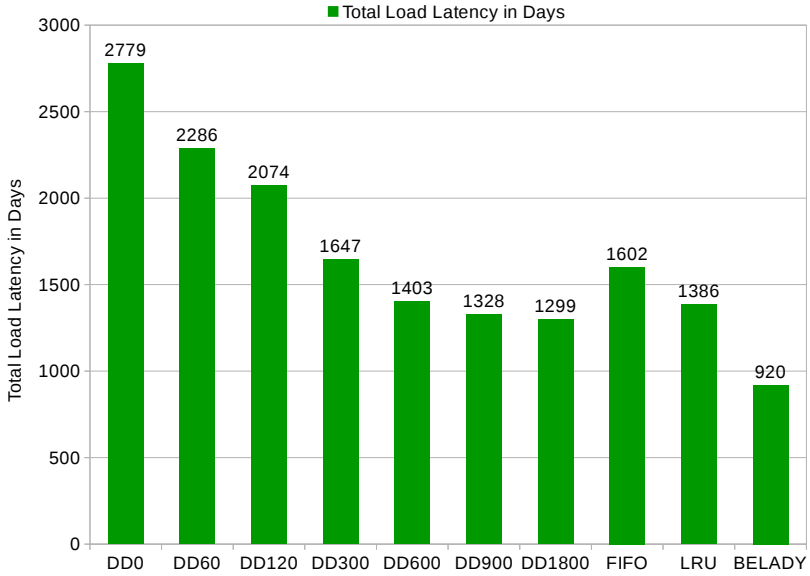
Simulation strategies

DD<X>: Defer dismount parameter of X seconds

FIFO: Evict cartridges in first-in / first-out order

LRU: Evict cartridges in least recently used order

BELADY: Evict that cartridge who's next access is the farthest in the future (theoretical optimum)



Part 2: New strategies

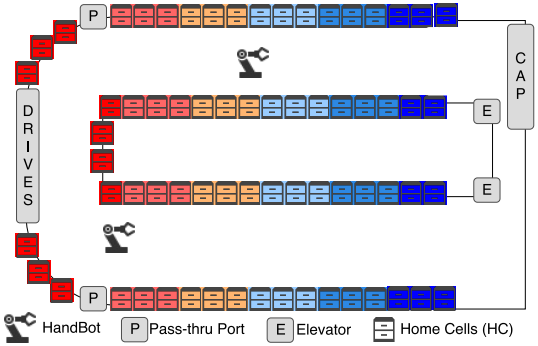
1st: New cartridge placement strategy
2nd: New cartridge eviction strategy

Part 2: New cartridge placement strategy

Strategy in Oracle/StorageTek ACSLS[1]:

- Home cell assignment static within a LSM

Pareto principle applies:
80% of the load is caused by
20% of the cartridges



New cartridge placement strategy:

Dynamic home cell assignment based on access frequency

Place hot cartridges into home cells close to the tape drives.

[1] Automated Cartridge System Library Software (ACSLS)

Part 2: New cartridge eviction strategy

New cartridge eviction strategy:

Try to keep exactly one drive empty in every LSM

1. Load a requested cartridge into a drive
2. Eject the least recently used cartridge from another drive
3. Place it into a home cell

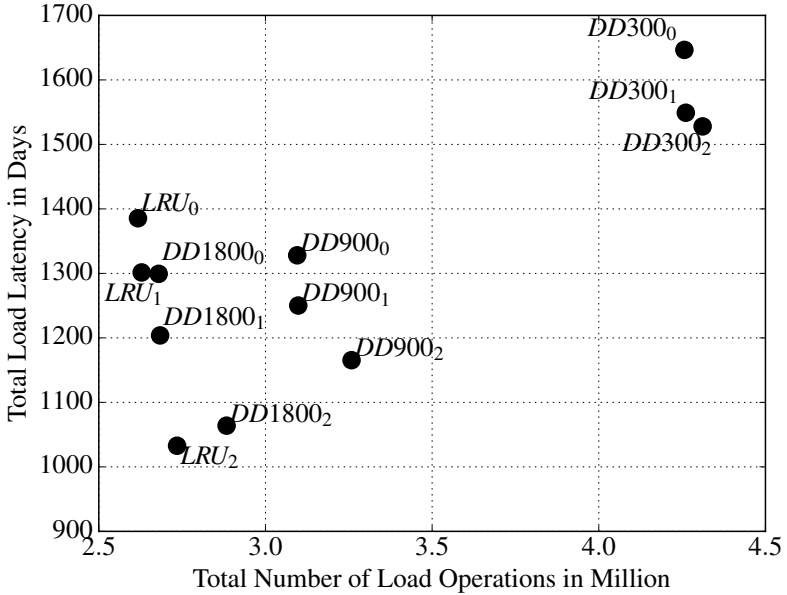
   $\Rightarrow$ Unload process not part of the critical path

Part 2: Evaluation results

Subscript-0:
Default strategy

Subscript-1:
New cartridge placement strategy

Subscript-2:
New cartridge placement strategy
and
new eviction strategy

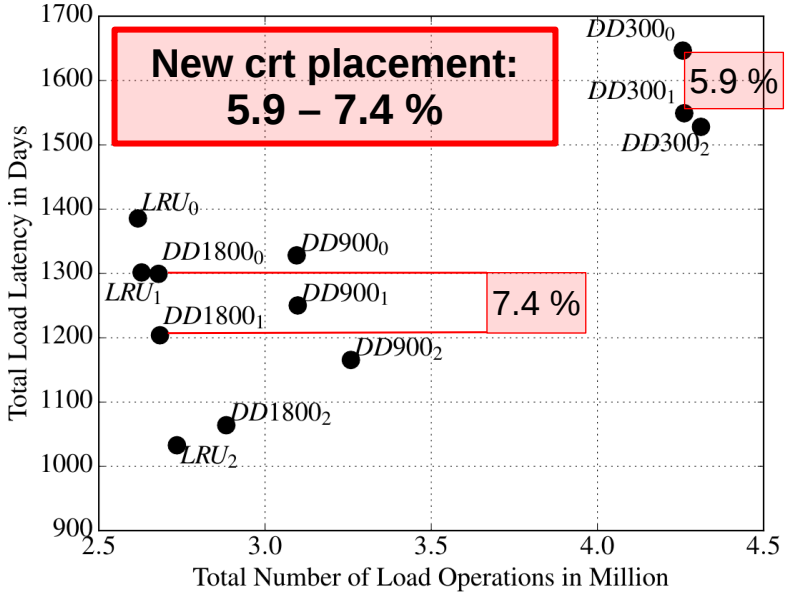


Part 2: Evaluation results

Subscript-0:
Default strategy

Subscript-1:
New cartridge placement strategy

Subscript-2:
New cartridge placement strategy
and
new eviction strategy

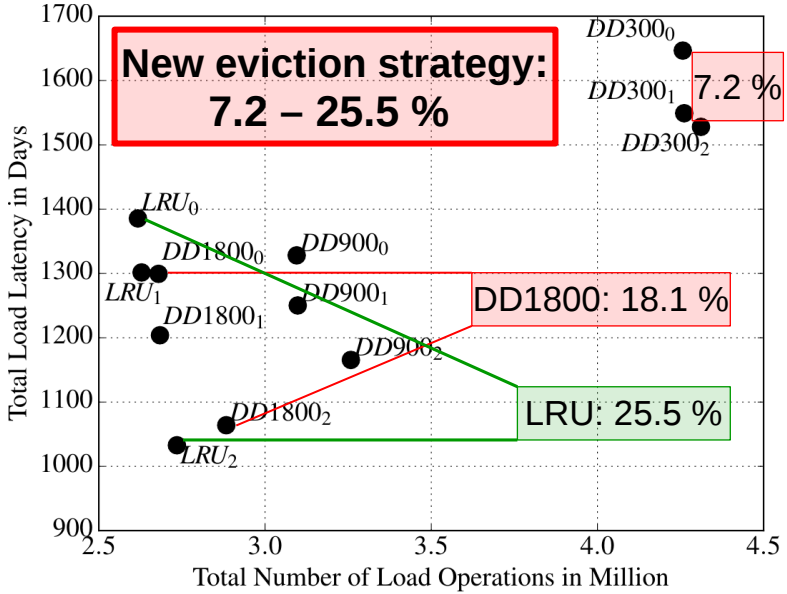


Part 2: Evaluation results

Subscript-0:
Default strategy

Subscript-1:
New cartridge placement strategy

Subscript-2:
New cartridge placement strategy
and
new eviction strategy

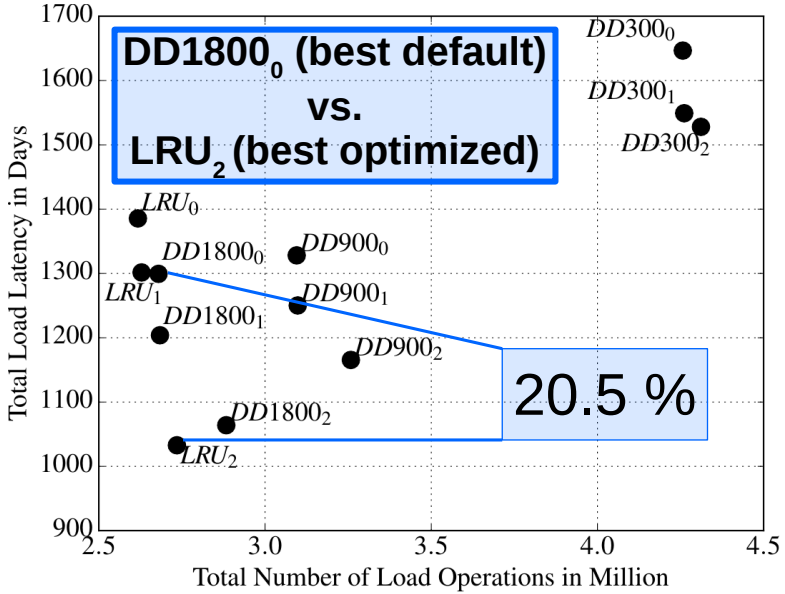


Part 2: Evaluation results

Subscript-0:
Default strategy

Subscript-1:
New cartridge placement strategy

Subscript-2:
New cartridge placement strategy
and
new eviction strategy



Part 3: Different number of tape drives

Part 3: Different tape drive count

Evaluate a different number of tape drives.

Part 3: Different number of tape drives

Small Configuration

- Four SL8500 tape libraries
- 12 tape drives per LSM
  16x12 = 192 drives

vs.

Large Configuration

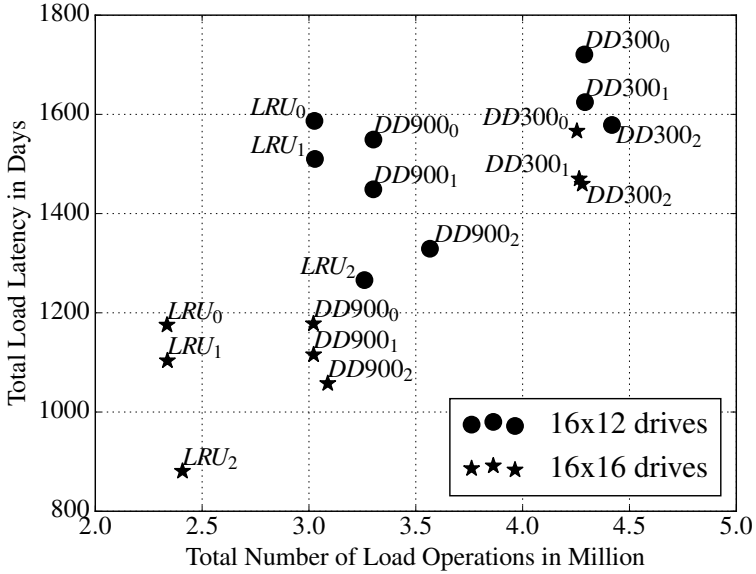
- Four SL8500 tape libraries
- 16 tape drives per LSM
  16x16 = 256 drives

Part 3: Different number of tape drives

Notable:

$\bullet - LRU_2$ almost as good as $\star - DD900_0$

$\star - DD300$ worse than $\bullet - DD900$



Conclusion

Key takeaway

The cartridge load latency of large robotic tape libraries

- . . . is neither constant nor irrelevant
- . . . can be optimized greatly

Reduced load latency:

- Defer Dismount value is crucial ⇒ vary by 2.1x
- Hot/Cold classification of cartridges ⇒ 5.9 - 7.4 %
- Improved load/unload strategy ⇒ 20.5 %

Special thanks to Lennart Sorth and the ECMWF
Q & A

Backup: Future Work

Ongoing work:

- Extend the simulator by a disk cache tier
- Support of read/write operations on tape
- Support of different tape drive and cartridge types/generations

Particularly interesting:

- Traces from other domains than weather forecasting
- Traces of different robotic tape library than the SL8500