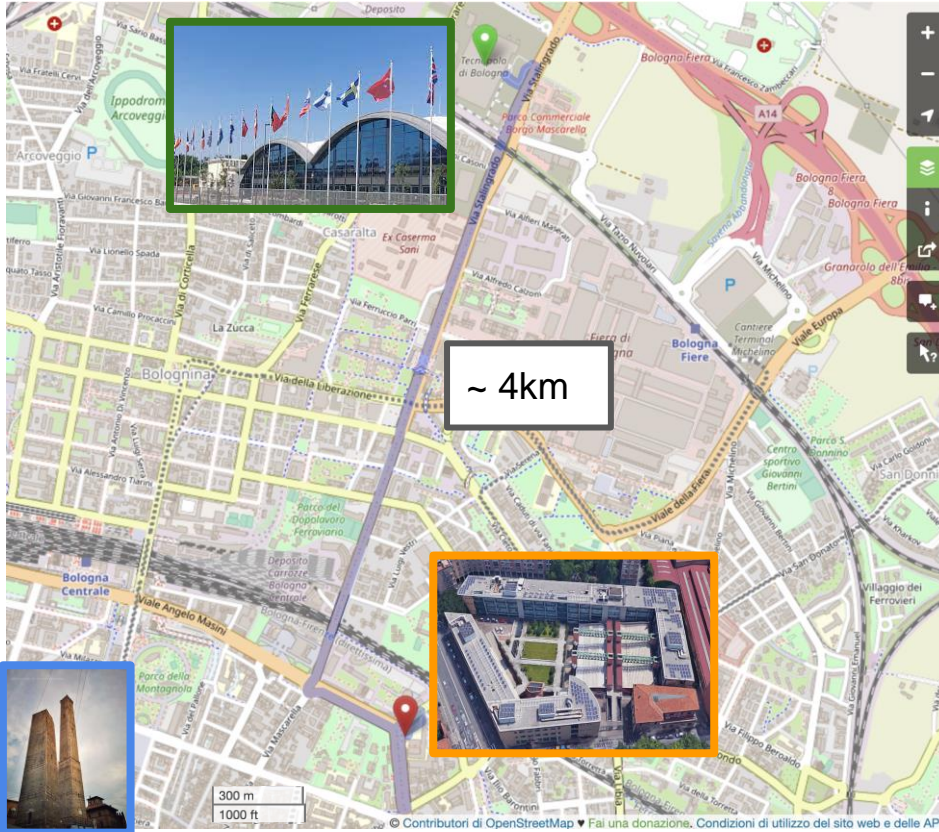


Storage relocation

INFN-CNAF experience

The new INFN-CNAF Data Center at Bologna Tecnopolo



- To be able to operate on two different locations using the same servers we established Long Range FC link (**7Km**) using Dark Fibers
- FC switch: Brocade G620
 - 32Gb/s FC LWL 10 km SFP+

Storage relocation: overview and goal

- Make the storage relocation as transparent as possible for end users
- Storage resources at source to relocate:
 - Disks
 - Tapes
 - On two tape libraries:
 - SL8500 - 80PB installed, 50PB USED
 - TS4500 - 102PB
 - Data Management and Gateway servers
 - 12 StoRM WebDAV, 24 XrootD and 5 HSM servers

Disk-based-storage relocation: the strategy

- Delay in planned purchase until new site is ready (power and network fully functioning)
- Install new equipment at the new site
- Migrate data using GPFS functionality:
 - add new disks to existing FS;
 - migrate the data (without interruption in data access);
 - remove old disks
- Dismiss the end-of-life systems
- Empty, disassemble and relocate the more recent system, then reassemble and re-initialize by the vendors

Disk storage resources before relocation

Model	Net capacity, TB	End of support
Huawei OS5800v5	8999	2027
DDN SFA 7990	5840	2025
DDN SFA 2000NV (NVMe)	24	2025
Huawei OS6800v3	3400	07/2024
DDN SFA12k	10120	12/2022
Dell MD3860f	2308	12/2024
Dell MD3820f	50	11/2023
Huawei OS18000v5	6520	7/2024

The strategy for disk-based-storage relocation: what happened in reality

- Delays in infrastructure readiness: delivery, installation and testing of new storage systems (64PB) was done with great hurry from mid of June to mid of July 2024
- For acceptance tests, we filled the storage with synthetic data up to the 50% of the available space. At that point the storage showed very good performance, even higher than requested (3.5MB/s/TB)
- On July 17th, the first 6 (out of 8) storage systems were put into PROD for starting data migration of ATLAS, ALICE and LHCb
 - Average rate of data migration was in 10-20 GB/s range for each exp

The strategy for disk-based-storage relocation: what happened in reality (2)

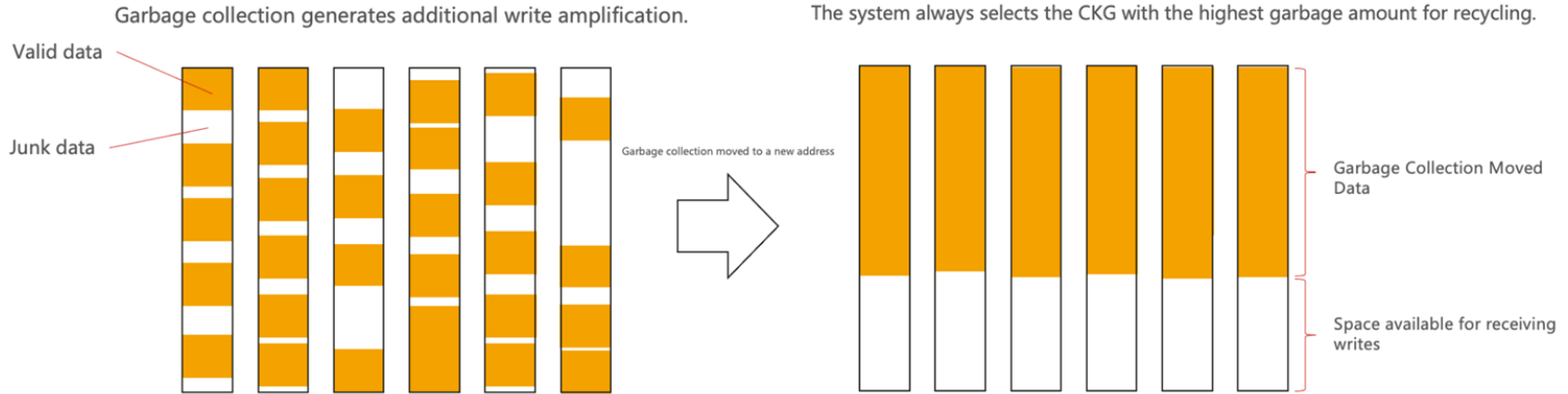
- At the same time LHCb started (very) data-intensive analysis which saturated capabilities of storage systems
- As result - in accordance with LHCb - we declared two weeks of downtime for LHCb to complete data migration
- By the end of data migration - when almost 95% of available space on new storage was filled up - we observed a huge drop in performance (down to 20-30% of measured during acceptance tests)
- The vendor was called in for troubleshooting and it turned out that the storage system works only in “Thin Provisioning” mode

How thin provisioning works

- All new writes go to not-yet-used blocks:
when a file is deleted, the corresponding file system block is marked as “freeable”, but for storage system it is still “in use”
- To free up space from deleted data, an admin needs to run the "reclaim space" procedure, followed by invoking the "garbage collector"
- This leads to the write amplification phenomenon

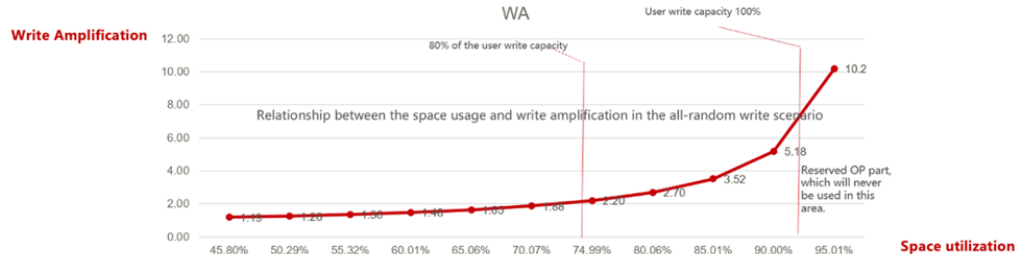
Write amplification in Thin Provisioned systems

Always New Write Cost: Garbage Recycling



$$\text{Write Amplification} = \frac{\text{Total amount of data written to disks}}{\text{Total amount of data written to hosts}}$$

$$= \frac{(\text{Number of written data on the host} + \text{Number of moved data for garbage collection})}{\text{Number of written data on the host}} = 1/\text{Ratio of garbage in recycling}$$



Taken from a private communication with the vendor

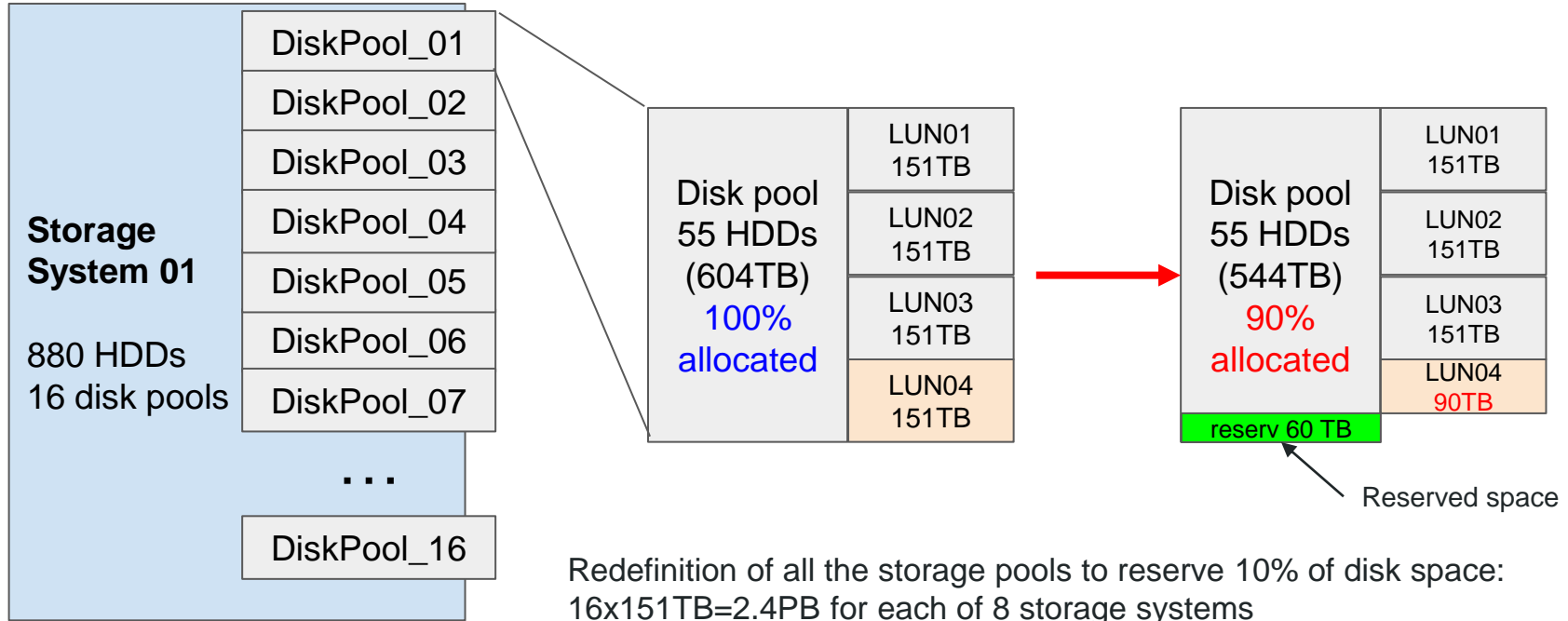
Solutions proposed by the vendor

- Enable support for “Thin Provisioning” at the file system level
- Do not use more than 90% of the “usable space”

Enable support for “Thin Provisioning” at the file system level

- Currently our file system **does not support** thin provisioning functionality
- Drawback of enabling support for “Thin Provisioning”:
 - Necessary to re-format all the disks → move all data
 - It involves significant performance degradation, as the "space reclamation" procedure that writes zeros to the space to be freed (which creates traffic and takes a significant amount of time)
 - It requires having approximately 5% reserved space (not accessible to users) [1]
 - Supported only on certified hardware (IBM flash)

Reduce storage allocation to 90% of “usable space”



Redefinition of all the storage pools to reserve 10% of disk space:
 $16 \times 151\text{TB} = 2.4\text{PB}$ for each of 8 storage systems
We moved ~15PB of data

Tape-based-storage relocation: the strategy

- Install a new tape library (IBM TS4500) at the new location and connect it via Long Distance FC link to the servers at the old location
- Switch production (writing) to the new library, disassemble and move one of the two old libraries (the other IBM TS4500) to the new location (one-week downtime for tape data access)
- Do repack from SL8500 to TS4500 to complete data migration via LD FC link within 12 months
 - Bandwidth 2x32Gbit/s (~ 7GB/s)
 - Migration rate limited by number of tape drives in new library and max transfer rate of old tape drives (250MB/s), so that with 8 tape drives dedicated to data migration we could do 2GB/s, thus moving 50PB of data in ~12 months
 - Of course, stop writing on SL8500

The strategy for tape-based-storage relocation: what happened in reality

- Because of delay in procurement (no new tapes available on the new IBM TS4500) we had to stop repack and begin to write data again to SL8500, postponing the SL8500 decommissioning to the end of 2025 at least
- Also, unexpectedly we were asked to free up space occupied by the tape libraries by the end of Jan 2025
- Thus, we were forced to relocate SL8500 to Tecnopolo as well

Relocating tape library (twice)

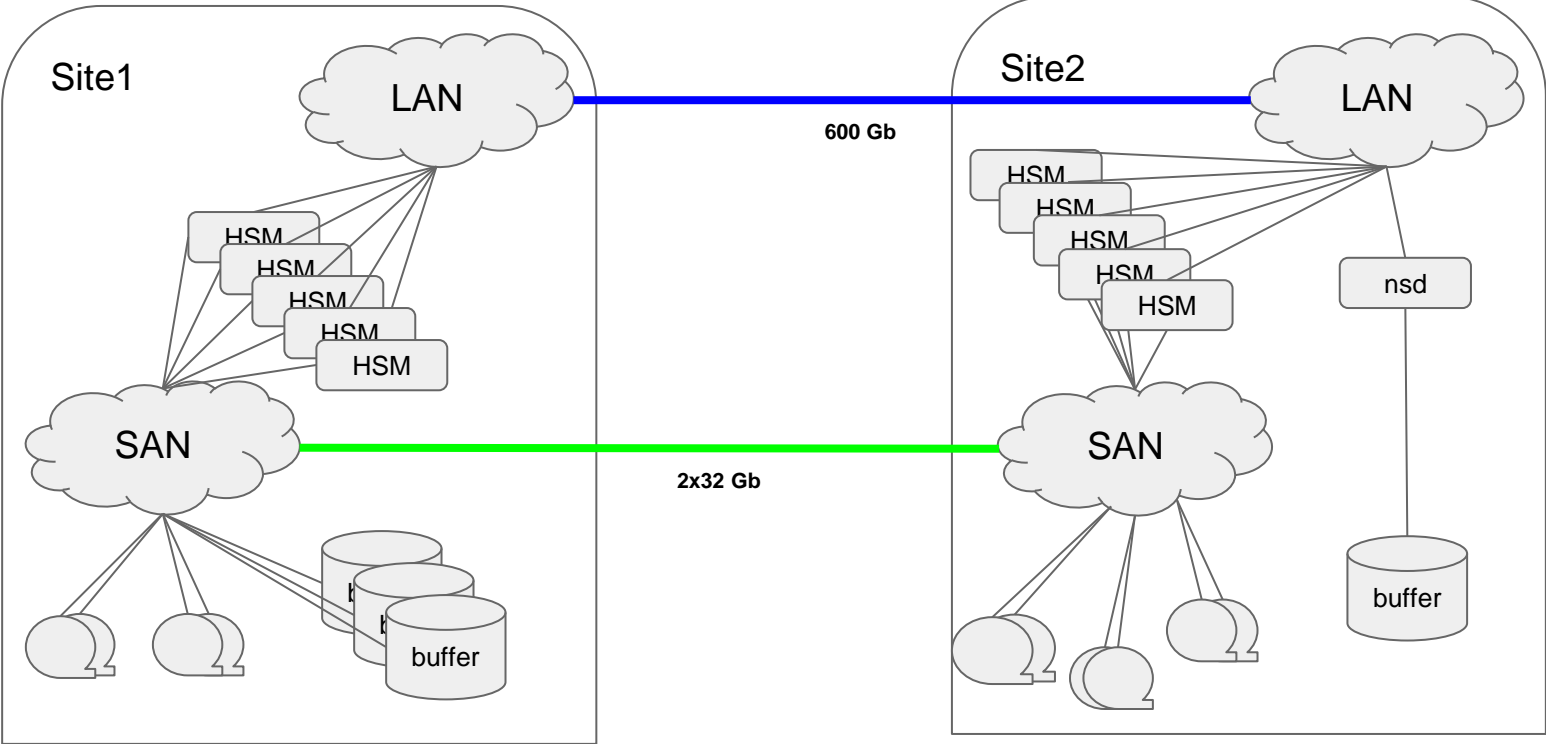
- Declared 1 week of downtime in tape data access, but the process of relocation took a little more than one week for a library
 - Unload tapes from library - 1 day
 - Disassemble library - 1 day
 - Transportation - 1 day
 - Re-assembly and cable tape drives - 1 day
 - Realignment of library and reload tapes - 1 day
 - After-relocation troubleshoot - 2 days
 - Some issues with FC cabling, tape drives, robotics
- Relocation and reconfiguration of tape servers have been done in parallel with the tape libraries movement



Relocating tape libraries



Setting up a metropolitan SAN



Relocation of data management servers

- Thanks to the unified network infrastructure over the two sites, the new servers at the new facility can access the file systems at both the locations
- We only needed to install the servers and put them in production by using the hostname alias used by the experiments

Lessons learned

- Avoid “Thin Provisioned” storage systems for multi-PB and frequent rewriting use cases
 - Need to better check storage system performance at the higher level of occupancy during acceptance test
 - As an option: procure and keep as “reserved” at least 10% of disk space respect to the pledged space
- Remote location for tape infrastructure (<10km) is not a problem at all
 - Streaming data is not affected by a slightly higher latency
 - It may be an advantage to keep tape infrastructure on a different location also considering different environment requirements for tapes and CPU/disks

Thanks

References

1. <https://www.ibm.com/docs/en/storage-scale/5.1.8?topic=devices-storage-scale-thin-provisioned>
2. <https://www.dell.com/support/kbdoc/it-it/000123351/powerstore-alerts-capacity-utilization?lang=en>
3. <https://docs.netapp.com/us-en/active-iq-unified-manager-97/online-help/concept-what-performance-capacity-used-is.html>
4. <https://indico.cern.ch/event/1225131/contributions/5587454/attachments/2747666/4781495/StoRM%20Tape%20status%20-%20PreGDB%202023-11-07.pdf>