High density and low energy consumption secondary storage system of petabytes with hard drives for the Alice and HAWC high energy physics experiments.

Eduardo Murrieta
Guy Paic
Lukas Nellen
Instituto de Ciencias Nucleares - Universidad Nacional Autónoma de México

Outline

- Motivation
- Alice and HAWC: HEP experiments
- Massive Storage Classification
- High density and low energy storage system
- Media lifetime, energy consumption and performance tests
- Costs comparaison
- Conclusions

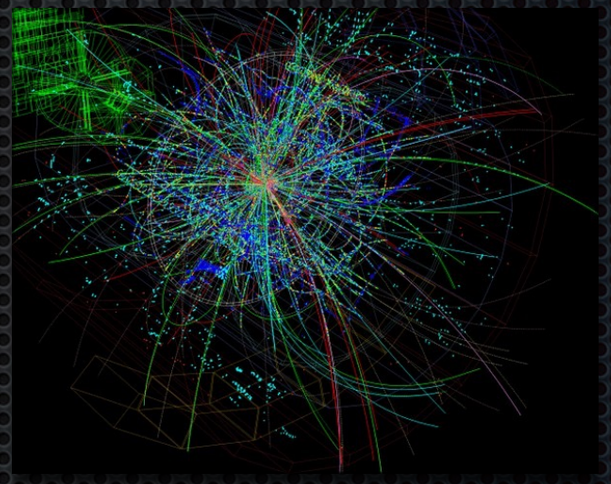
Motivation

- Hundred of Terabytes generated by large HEP experiments must be preserved unless they are infrequently accessed consuming energy and space in a high performance storage system.
- Tapes and Tape Libraries have been used traditionally as a long term storage solution with low energy requirements (The Clipper Group)
  - Big investment

ALICE and the Grid

- **A Large Ion Collider** : LHC CERN's detector
  - Study of the matter in the quark & gluon's plasma state.
  - **Alice RUN-3 (2018) ~ 60 PB/year**
- GRID: Global computing infrastructure, provides resources to storage, distribute and analyze the data generated by the LHC
- Organized by Tiers
  - **UNAM - Tier 2 for ALICE**
    - 1024 cores / 456 TB

Our Objective is to scale to Tier-1
2 - 10 PB on tapes is a requirement !



Worldwide LCH Computing Grid WLCG

HAWC: High Altitud Water Cherenkov Gamma-Ray Observatory

- Study of the origin of the most energetic Cosmic-Rays (0.1-100 TeV)
- 11 TB of raw data per week
- **600 TB per year for 10 years**
- Stored at the ICN-UNAM data center (4.3 PB / lustre)
- Thousands of files remain without access for months after been processed.
- Actual storage use 3 PB (70%)
  - Performance degrades as the filesystem get full.

Massive storage levels

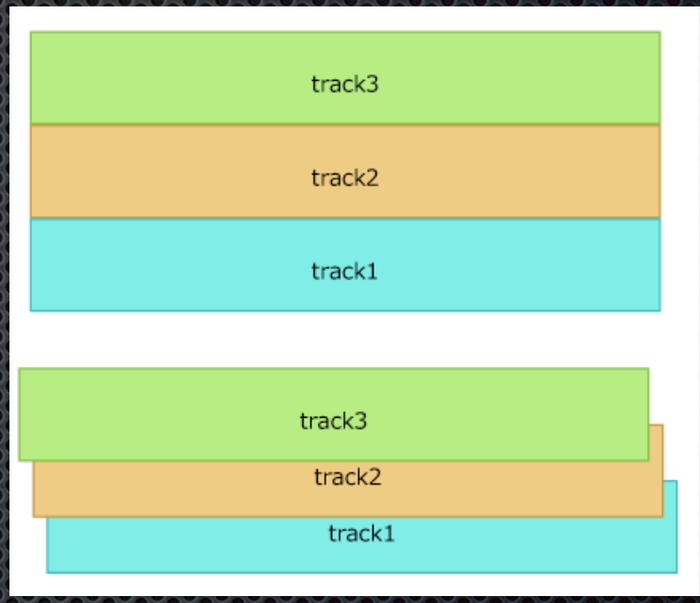
- **HOT-data: Primary storage**
  - **Frequently accessed data**
  - High I/O performance storage
  - **High energy demanding**
  - SAS-III drives / 6-10 TB/drive
  - RAID fault tolerance
- **COLD-data: Tertiary storage**
  - **Infrequently accessed data**
  - Low I/O performance
  - **Lowest energy demanding**
  - Tape Libraries
  - No fault tolerance
- **WARM-data: Secondary storage**
  - **Low frequently accessed data**
    - Alice & HAWC have many TB of data in this class.
  - Low to Medium I/O performance
  - **Can be energy optimized**
  - SATA drives / 8-12 TB/drive
  - RAID or replica tolerance

ALICE and HAWC storage solutions options

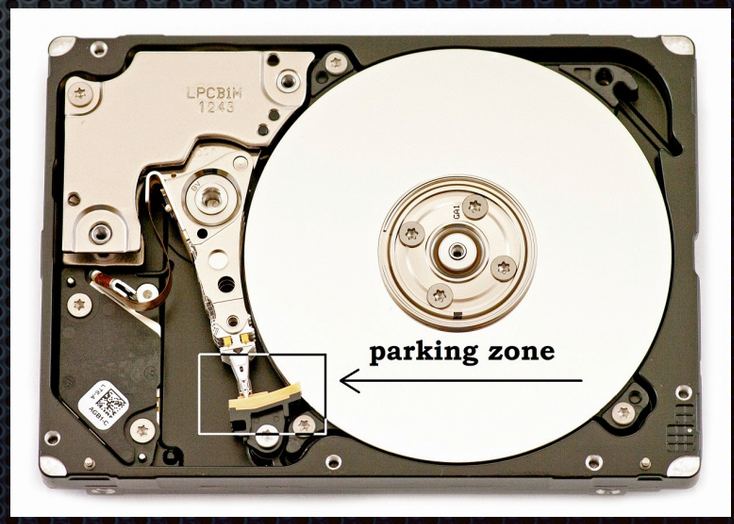
- A) To buy a Tape Library to store the unused data of both experiments.
  - The cost of a tape library for 1 PB is around $160,000 USD.
  - A tape library requires a controlled dust, humidity and temperature environment (more critical than drives), trained personal and maintenance contracts.
- B) To use our experience on storage systems based on drives to implement a low cost and low energy storage system for warm-data.
  - It can be proposed to CERN as a replacement for a Tape Library, as is required for a Tier-1 data center.

SMR drives with Power-Mode control

- Shingled-Magnetic-Recording
  - Writing technic that increase the storage density in a hard drive by overlapping the data tracks.
  - Ideal for **W**rite-**O**nce-**R**ead-**M**any applications
    - i.e. backup systems.



- SCSI standards T10/09-54 and T13/452-2008
  - Permit to change the operating-mode of a hard drive with different energy levels.
  - **`smartctl -s standby,60 /dev/sdX`**



Seagate Archive Drives 8 TB

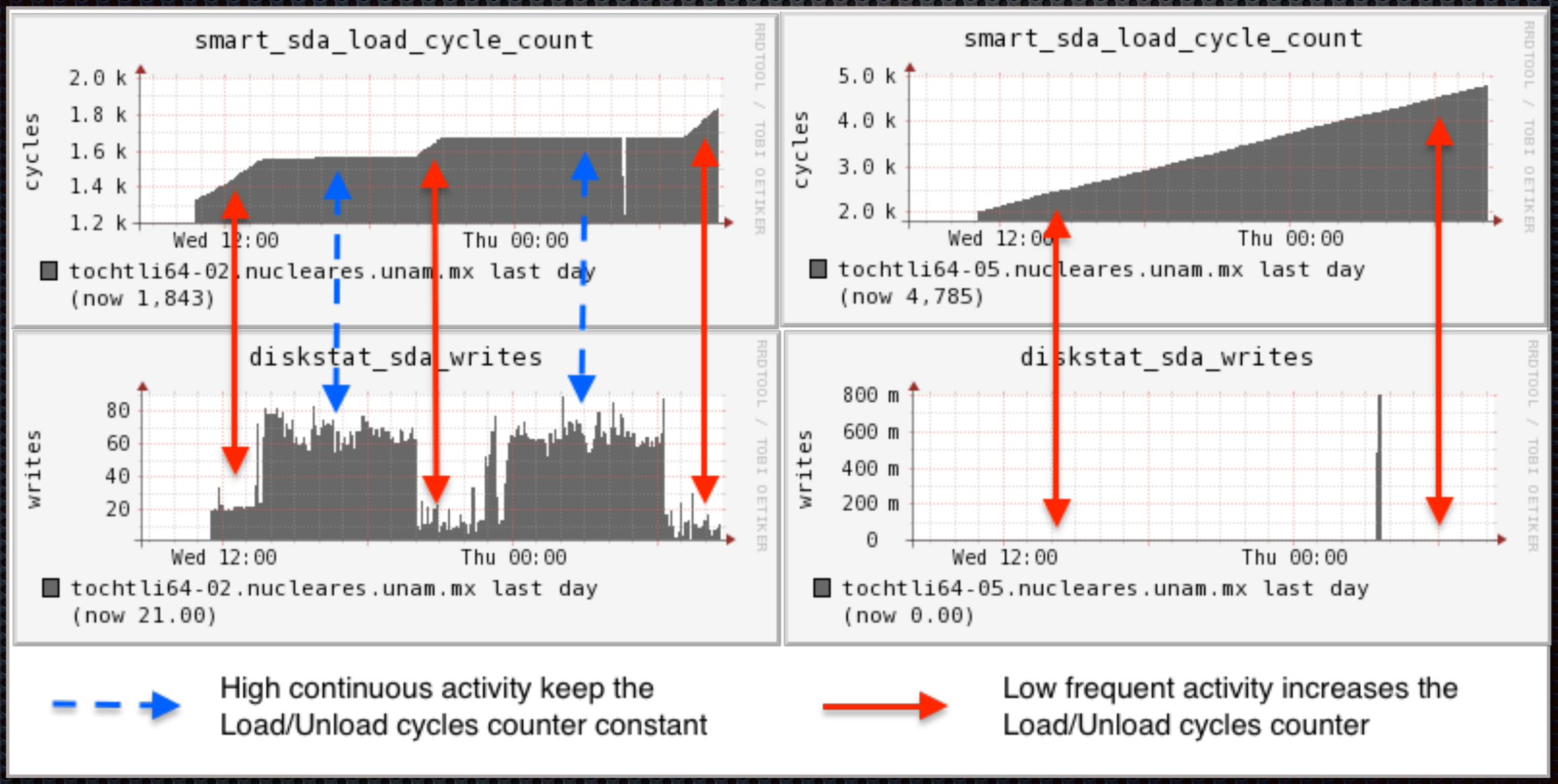
- The Seagate Archive Drives are SMR drives with SCSI operational mode control.

Power_Mode	Power (W)	Timer	Recovery Time	Description
ACTIVE	6	0 s	0 s	Full operational
IDLE_A	4.78	1 s	0 s	Reduced electronics
IDLE_B	4.13	5 m	0.5 s	Heads unloaded, disks at full RPM
IDLE_C	2.4 W	undefined	1 s	Heads unloaded, disks at reduced RPM
STANDBY_Z	**0.5 W**	**disabled**	8 s	**Heads unloaded, motor stopped**

Effect of the commutation mode in a hard drive

- Test Bed
  - Destructive test on two hard drives
    - Maximum number of Load_Cycle_Count = 300,000
    - Time before Idle transition to sleep mode = 8 s
    - Drives used as a filesystem (ext4) on two Linux computing nodes

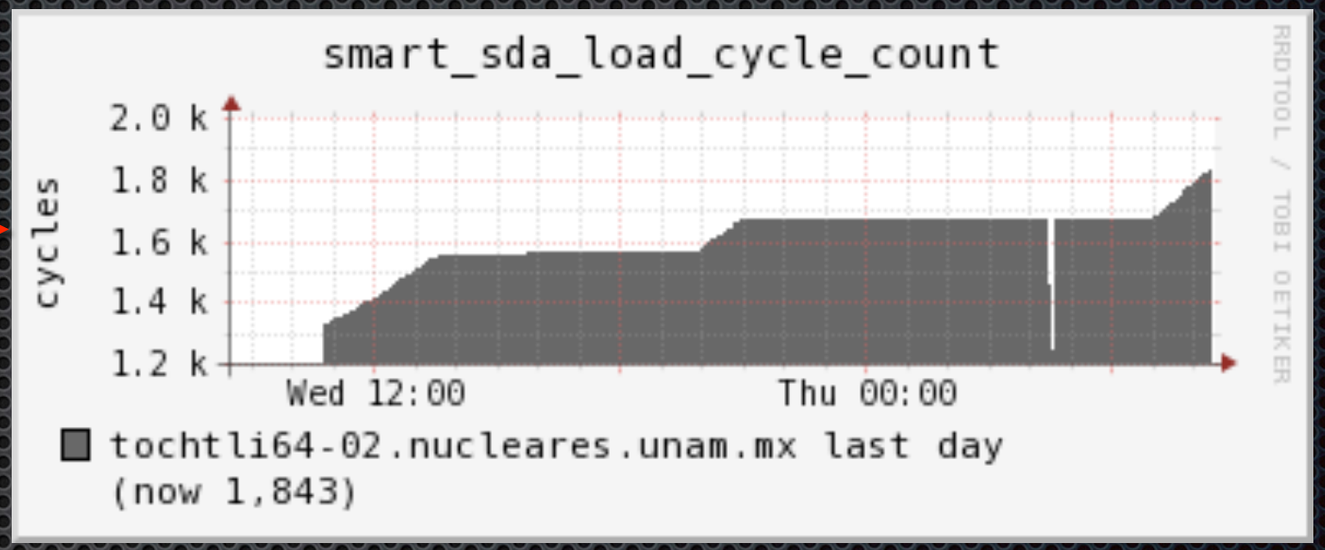
Effect of the access pattern on the Load_Cycle_Count (24h)



HD useful access pattern

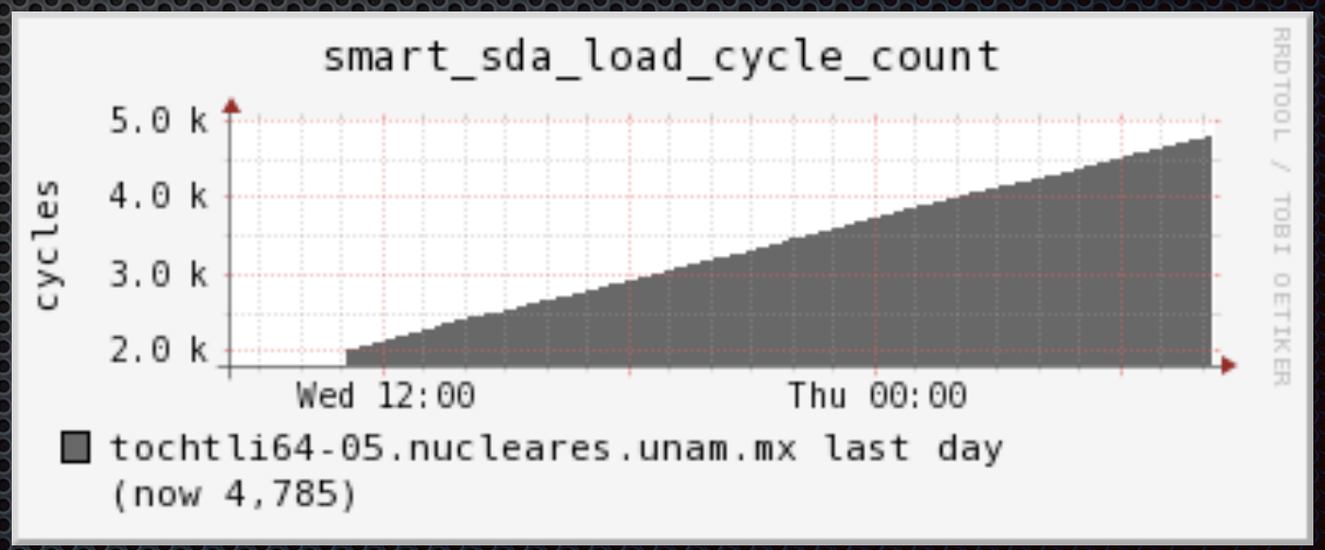
Drives must have periods of high read or write activity mixed with periods of none activity.

The expected behavior for a backup system



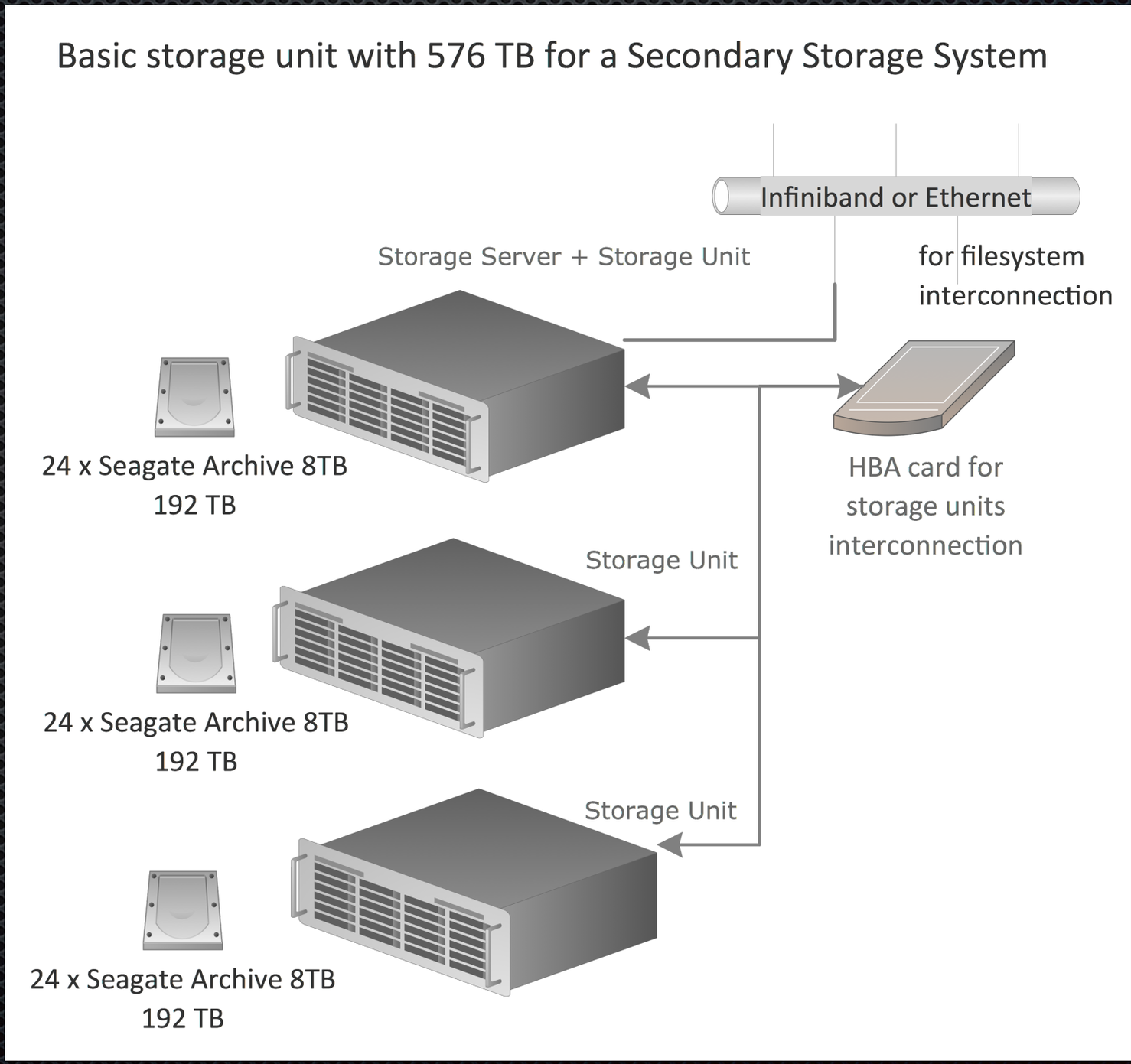
Constant low activity will reduce the expected lifetime of the drive.

Drive died after 151 days
Load_Cycle_Counter = 333,112

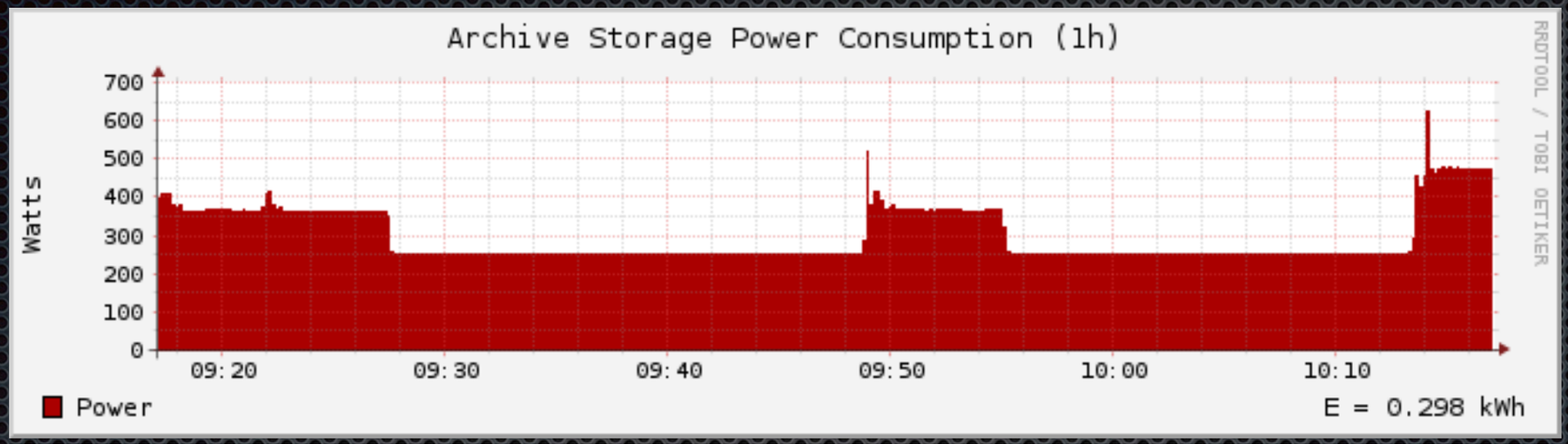


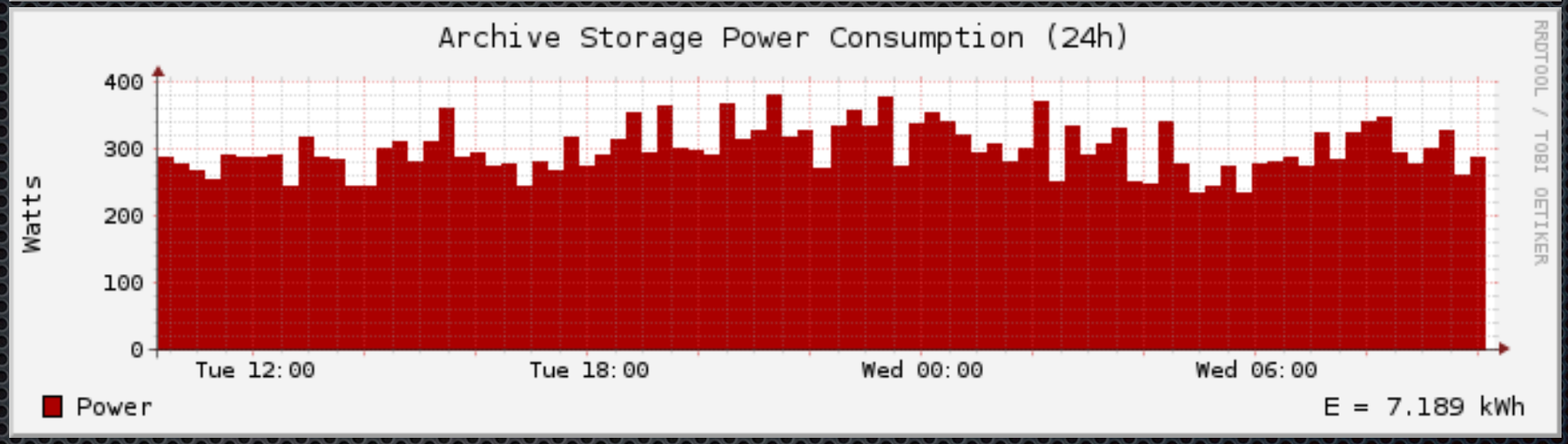
Secondary Storage System of 576 TB

- Basic storage unit of 576 TB.
- One server controls 72 drives with a single HBA card.
- Minimal energy: 295 W
- Maximum energy: 672 W
- Bandwidth: W-800 MB/s, R-1.2 GB/s.
- ~ 45,000 USD



Measured energy for 384 TB

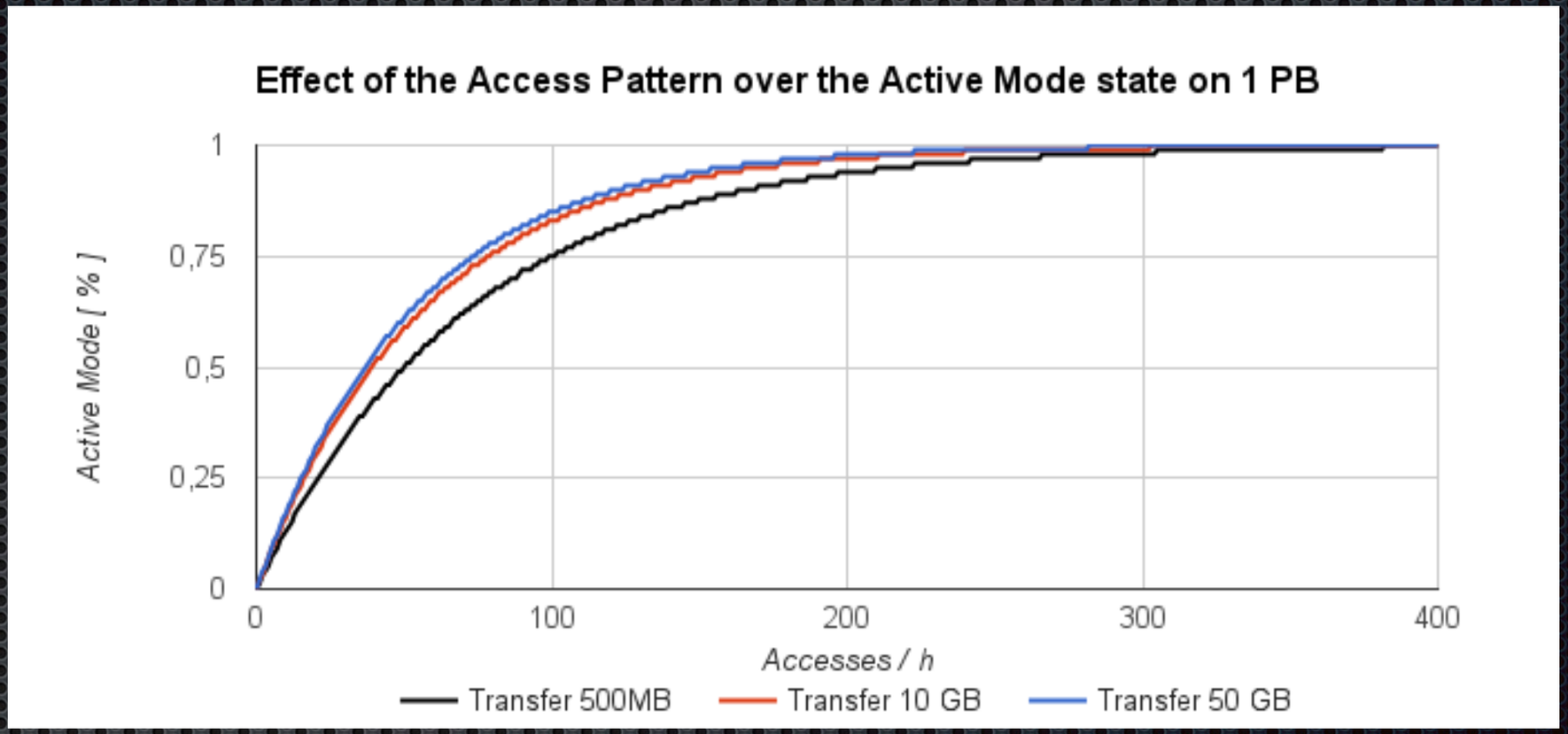




- Simulated access pattern using a Poisson distribution with 6 accesses per hour
- Time before transition IDLE_B->STANDBY: 5 min
- Average Consumed Energy: 300 Wh

Simulated energy consumption for 1 PB

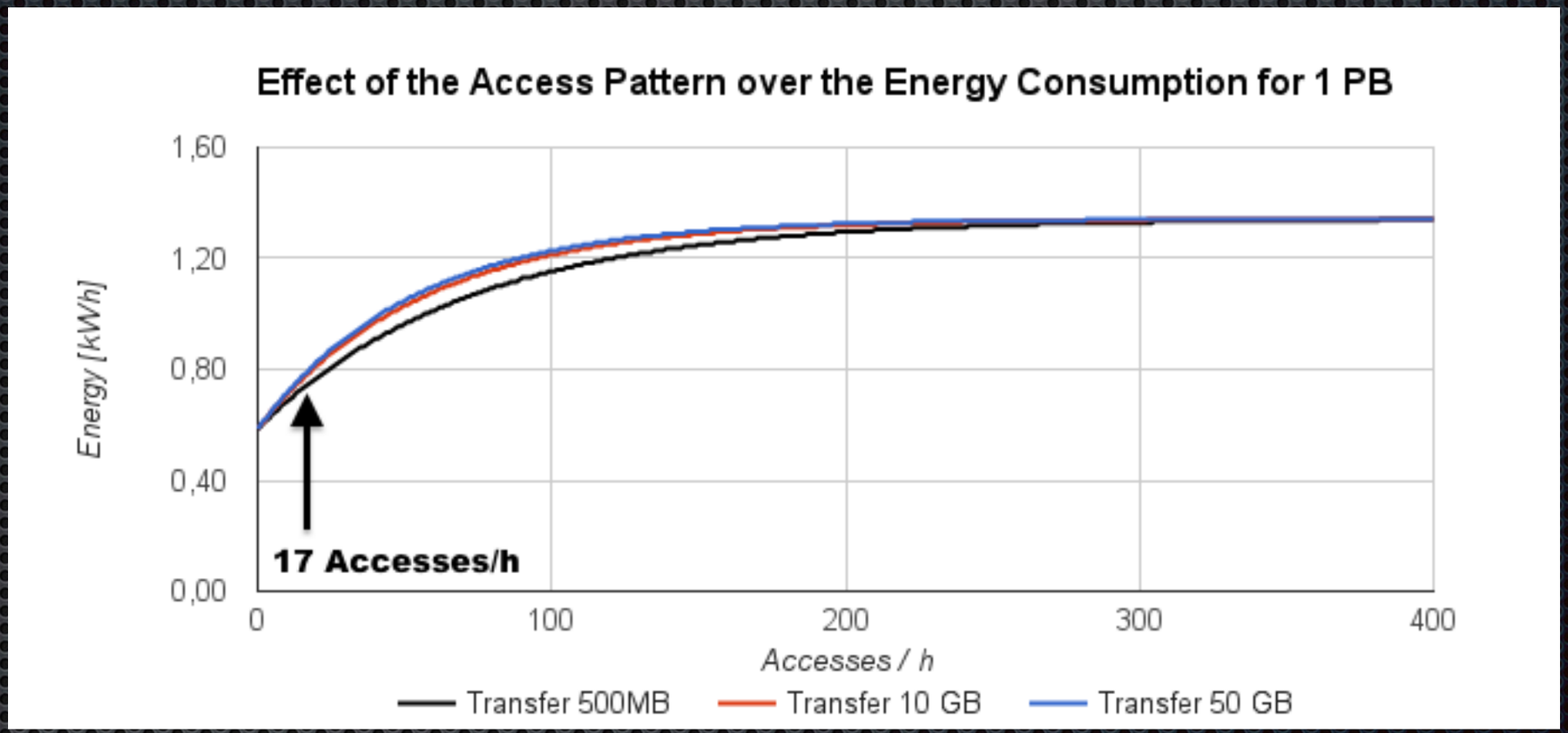
Effect of the access pattern on active mode of the system



- Each curve represents different read/write transfers sizes: 0.5 GB, 10 GB, 50 GB.
- **The graph shows that the transfer size has low effect on the active state of the system.**

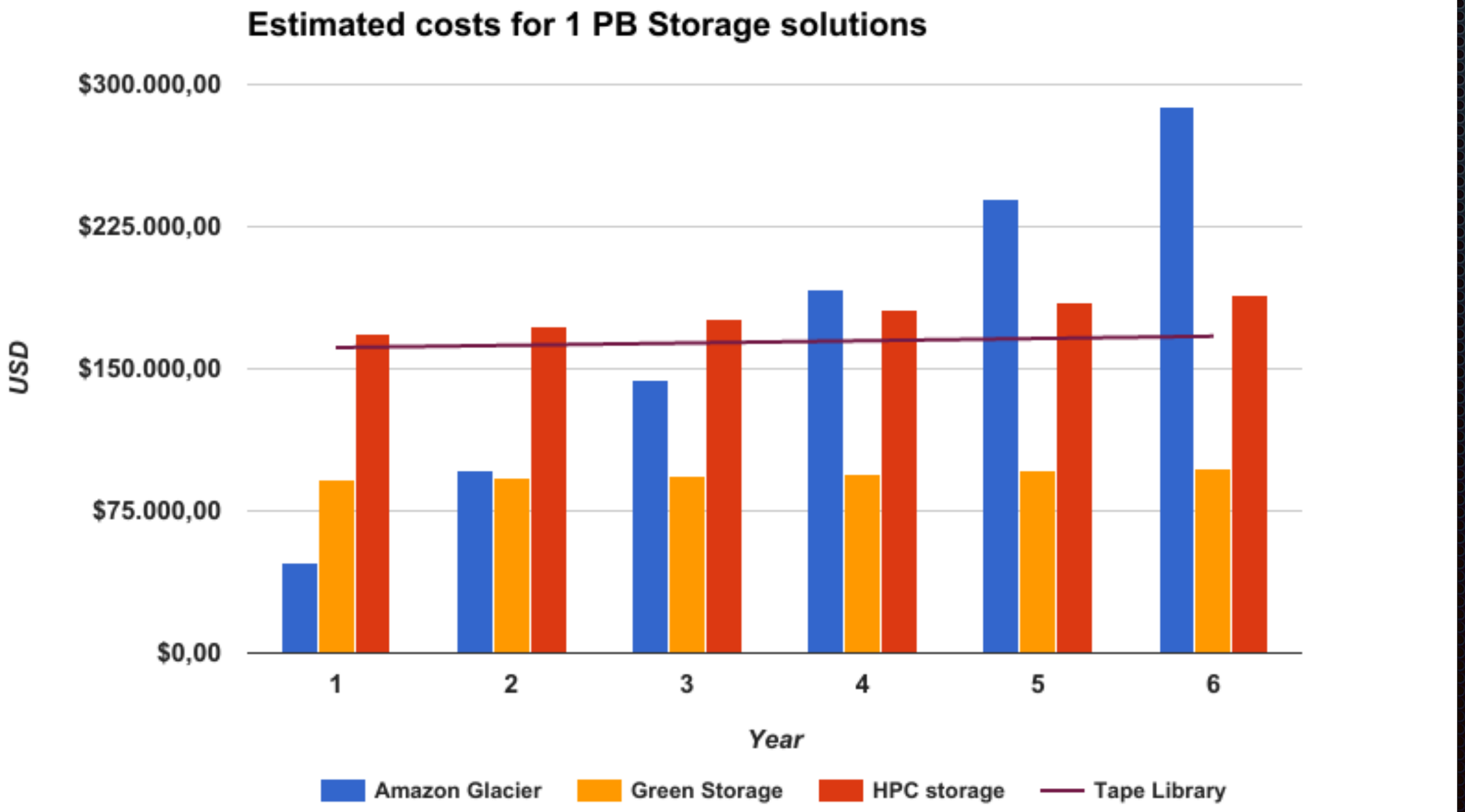
Simulated energy consumption for 1 PB

Effect of the access pattern on the consumption energy for 1 PB

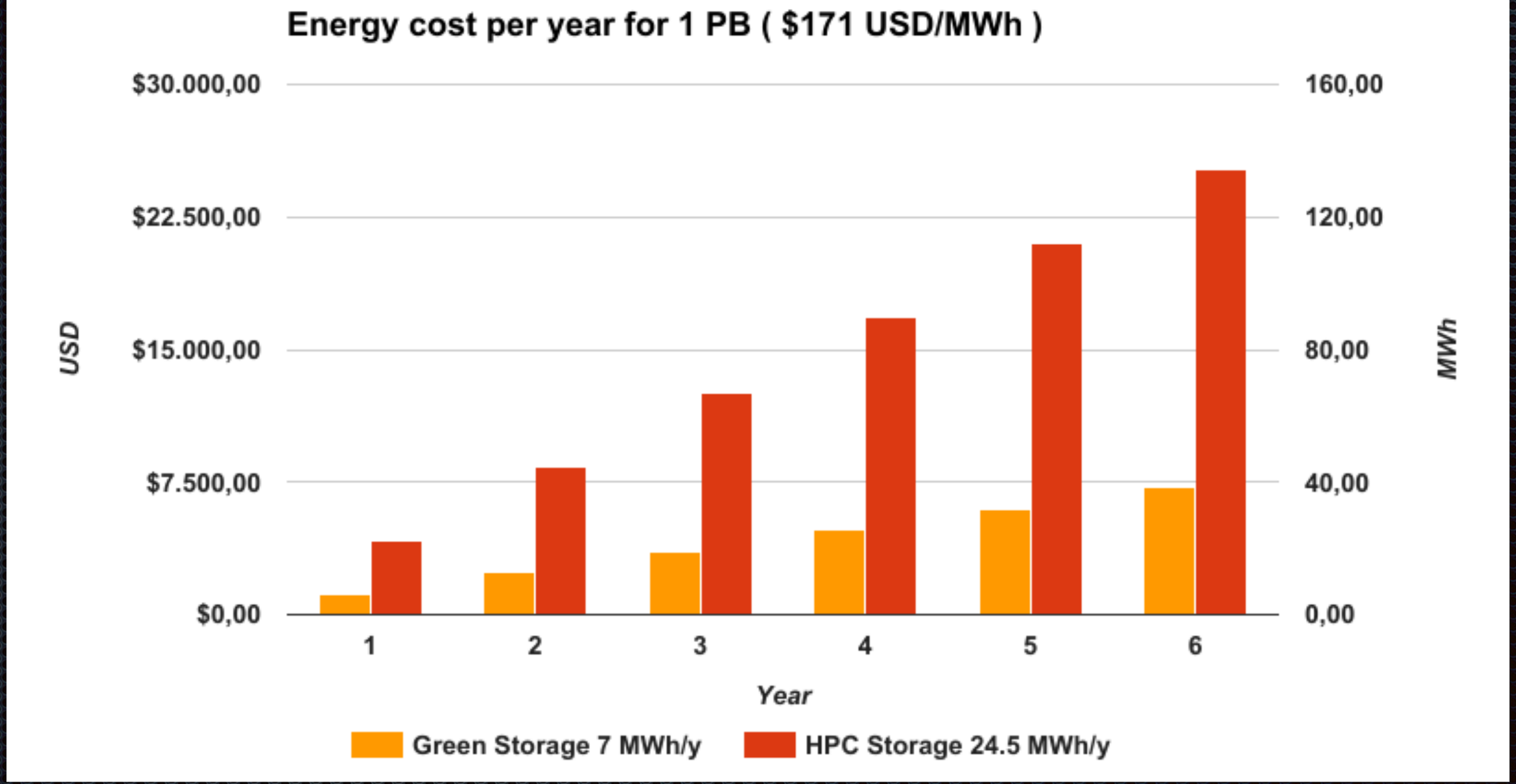


- **The graph shows that under 17 accesses per hour (~400 accesses per day) the system consumes less energy than a Tape Library of 800 Wh.**

Estimated costs for 1 PB



Energy cost per year for 1PB



Supported filesystems

- Any Parallel filesystem can be used with this solution but some tuning can be needed.
  - **Lustre**: lustre pools can be used to control which OST will be used to store the data and keep other OSTs inactive. Fill the OSTs progressively instead of in a balancing mode.
  - **Hadoop-FS**: since the user don't have direct access to the underlying filesystem it can help to keep the disk in the STANDBY mode and activating only the drive with the data
  - **EOS**: the filesystem used by CERN, will be used in our proposal for a Tier-1 data center.
- **Do not use with Hardware RAID**, since them have periods of consistency check, that will kill the drives. **Instead use ZFS** that do not require consistency check.
- **Avoid to activate all the drives at a time**!. A cache server as a fronted for this storage can help on this.

Conclusions

- A solution based on Seagate Archive drives can be a viable solution to store cold/warm data if:
  - The access pattern has periods of high activity mixed with periods without activity.
  - The access pattern is less than 17 accessed per hour to keep the system consuming less energy than a Tape Library.
- The Cloud Storage is costs effective if the storage needs are less than two years.
- The solution is hardware dependent, no filesystem dependent.

Thank you