Study of the viability of a Green Storage for the ALICE-T1

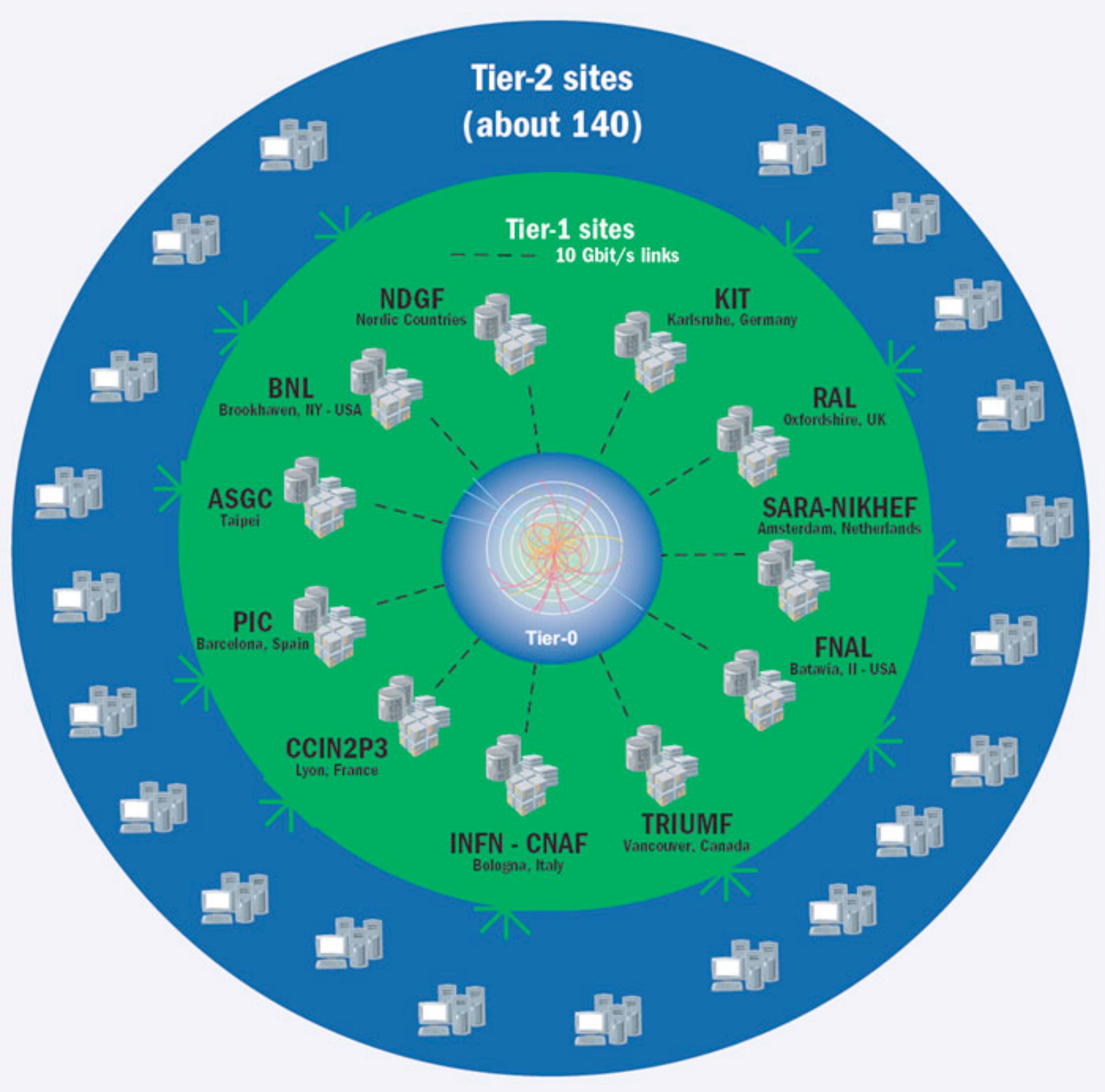
Eduardo Murrieta
Técnico ~~Académico~~: ICN - UNAM

Objective

- To perform a technical analysis of the viability to replace a Tape Library for a Disk Only based solution

Motivation

- CERN uses a hierarchic scheme for the management of its data.
- Classified by different levels of Tiers, each one with different technical requirements and services



Tier-0

- CERN Data Center
  - Collects all data generated by detectors
  - Do a first reconstruction of the data
  - Distribute the raw and reconstructed data to Tier-1
  - It has the custodial of all the information, past, present and future; generated by the LHC and other experiments.
    - Around 100 PB of capacity installed

Tier-1

- Primary responsible for safe keeping a copy of the CERN raw and reconstructed data.
- Distributes data to Tier-2 centers and keep safe a copy of the data generated by the Tier-2.
- Anual availability > 98 %, fast response time on failures.
- **Must have a 10 Gb connectivity to Tier-0**.
- **A combination of mass storage based on hard drives and tapes is a requirement for an optimal performance and long term custodial of the Tier-0 data.**
- 13 sites around the world: only 2 in America in USA.

Tier-2

- Provides sufficient computing power to do tasks of production and reconstruction.
- Provides sufficient storage for temporal and long term backup of data.
- Annual average availability > 95%
- **UNAM has one Tier-2 data center for the ALICE experiment.**
  - **456 TB of storage in disks**
  - **1024 cores**
- **Te objective is to scale this Tier-2 data center to a Tier-1**
  - **10 Gb connectivity to CERN it is now possible from ICN**
  - **The minimal requirements are 2 PB of raw capacity in tapes scalable up to 10 PB for RUN-3**

Tape Library

- A tape library is a robotic system that automates the management of tape cartridges; from hundred to thousands of tapes.
- The most distinctive attribute of a tape is its low energy requirements as a backup system.

Storage media

	MECHANICAL GREEN DRIVES HDD	SOLID STATE DRIVES SSD	MAGNETIC TAPES
CAPACITY	Upto 8TB	1.6TB (->16TB)	8.5 TB (40 TB)
POWER	8 W	5.2 W	0 W (20 W)
PERFORMANCE	190 MB/s	460 MB/s	252 MB/s
RELATION $/TB	~USD $30 / TB	~USD $300 / TB	~USD $8 / TB
GUARANTEE	3 years	5 years	30 years

Power consumption of the media

- **Tapes**
  - The media without access do not require any energy
  - The drive (reader) required to access the media goes from **12 to 20 watts**.
- **Hard drives**
  - Energy requirements varies according to the operation mode of the drive
    - Active mode: ~**9 Watts**
    - Standby mode: ~**7 Watts**
    - Sleep mode: **0.5 Watts**
  - **New technology of “Green Drives” can save energy by commuting between operation modes**
    - **This process reduces the life time of the drive.**

Testbed for Green Drives

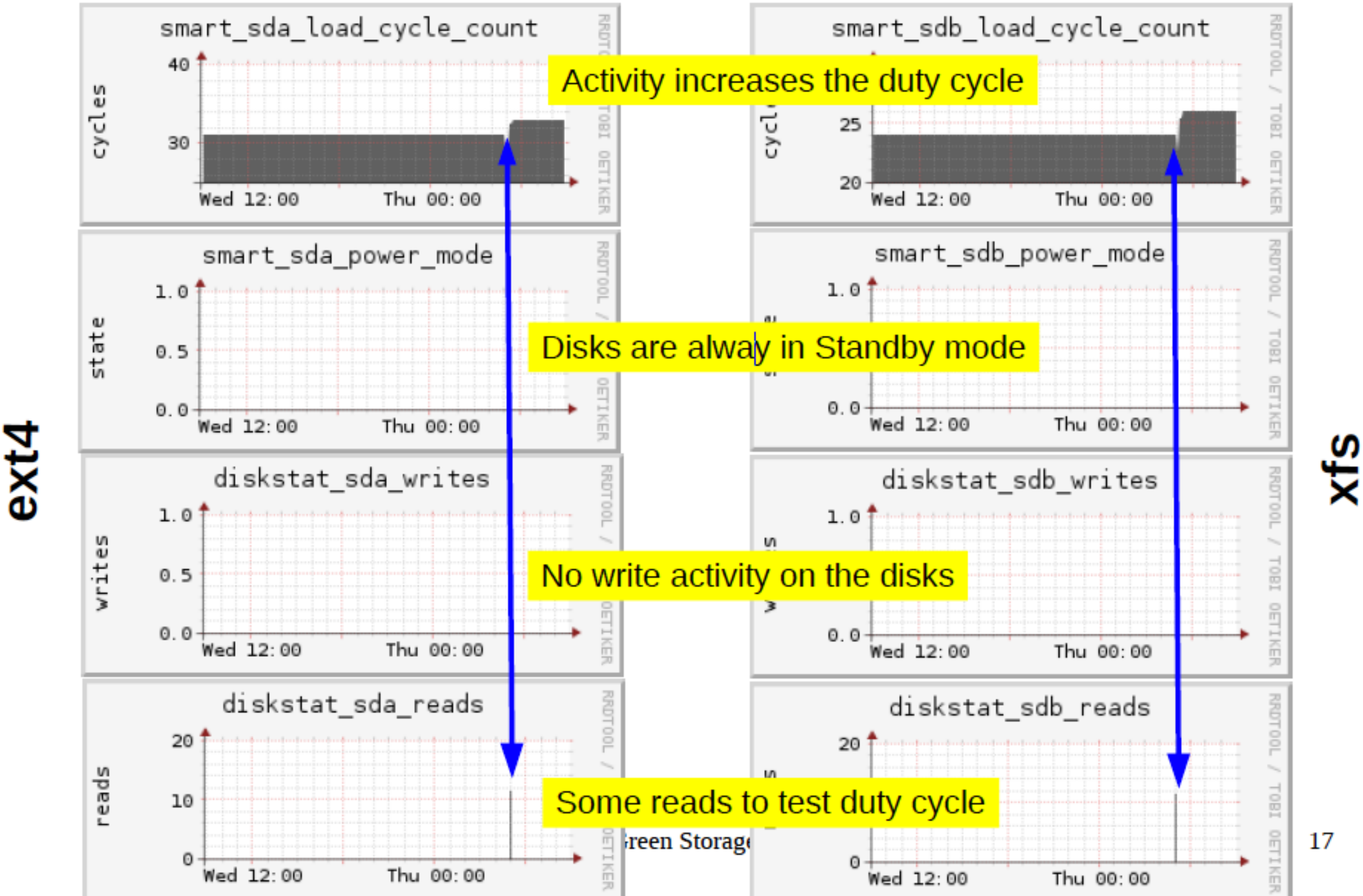
- **4 WD Intellipower hard drives**
  - 3 TB each
  - 5200 RPM
  - 8s idle before switching to Standby power mode
- **1 WD formatted with ext4**
- **1 WD formatted with xfs**
  - Both mounted
  - No access to the drives
- **2 WD as system's drives with ext4**
  - In normal use in a compute node
- **A Ganglia's python module was made to check the smart counters**

Green disk duty-cycle

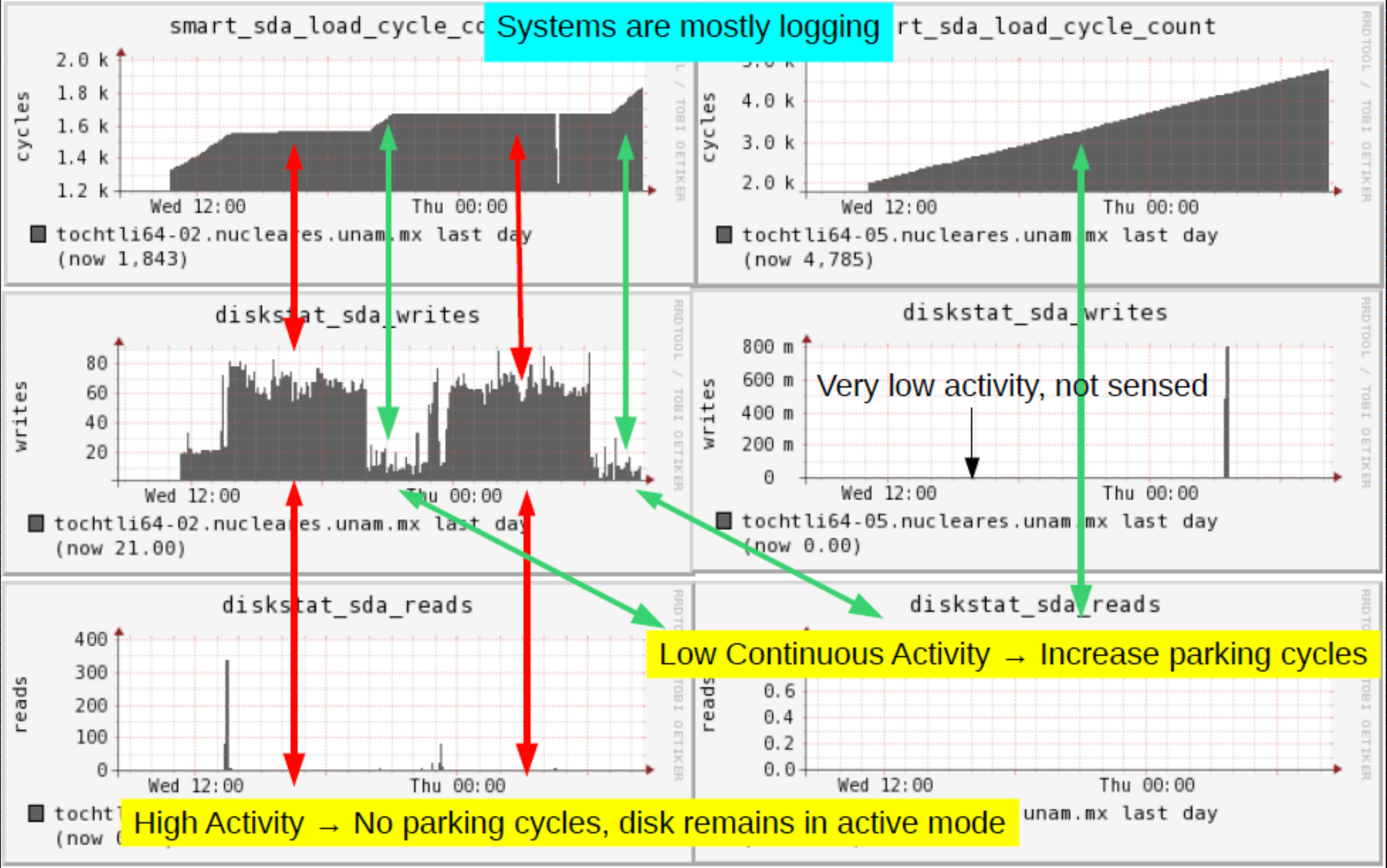
- Maximum Load/Unload (active-sleep mode) cycles: 300,000
  - Default parking time after 8s of inactivity
- Guarantee 2 years => 2 years of work at 24x7
  - 2 y * 365 d/y * 24 h/d * 60 m/h * 60 s / 300,000 =
    - 210 s/cycle ~ **1 parking cycle every 3.5 minutes**

In order to use a Green Drive as a long term backup media, the filesystem synchronization must be very infrequent when there is no read/write operations.

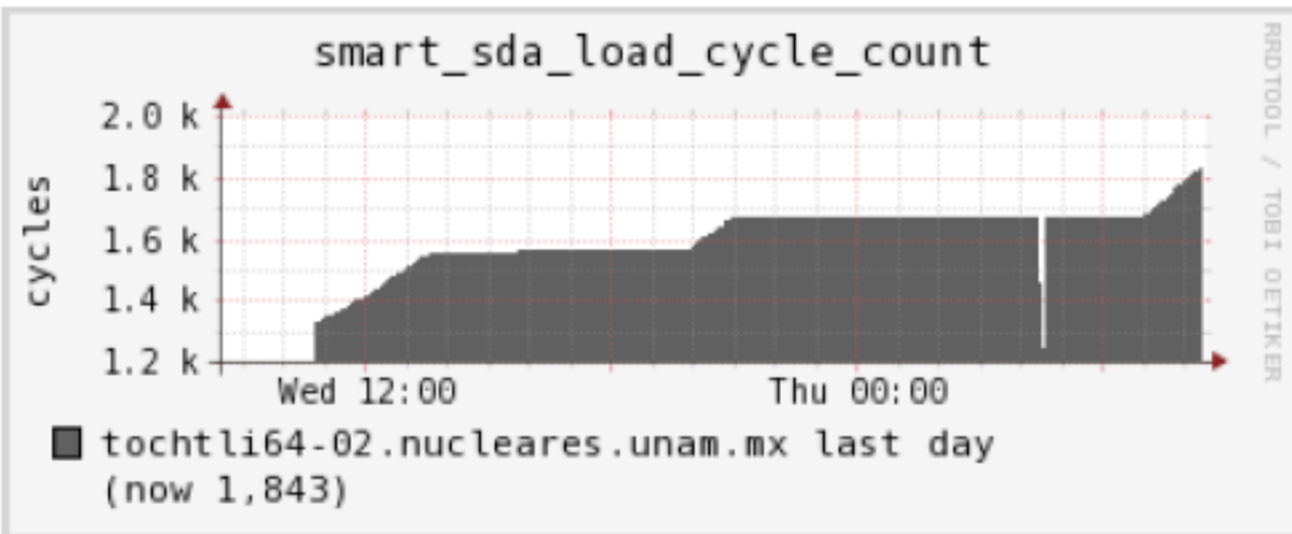
HD duty cycle: inactive filesystem 24 h



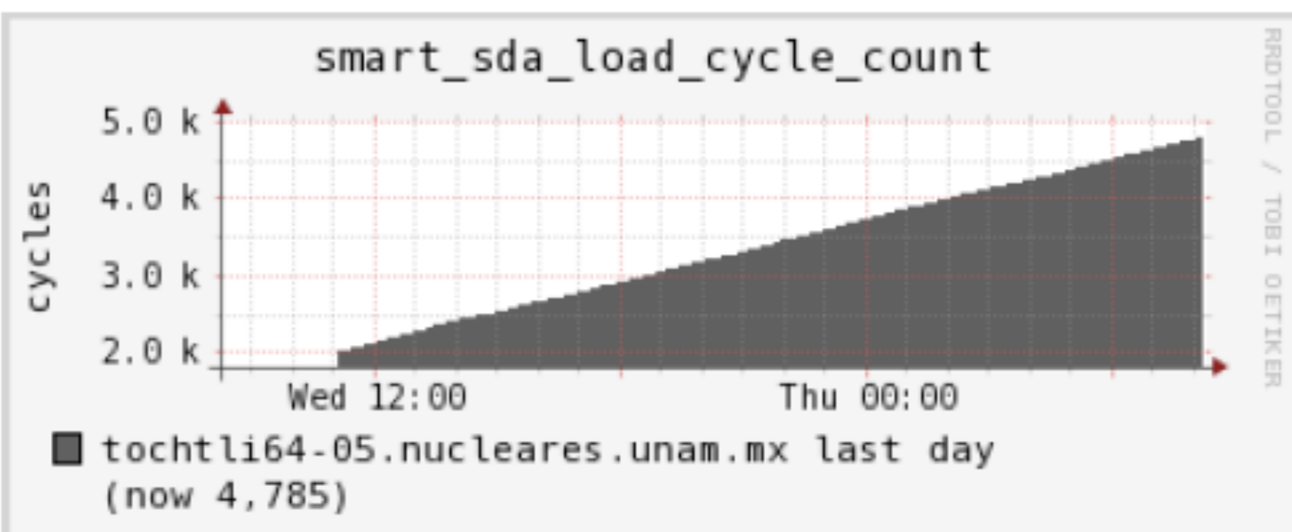
HD duty cycle: Linux OS system on ext4, 2 servers 1 day



Green HD useful duty cycle pattern



- Disk must have periods of high constant write or read activity mixed with periods of no activity in order to preserve its median lifetime.
- **Mostly the expected behavior for a backup system !**



- Constant low activity will degrade the median lifetime of the drive.
- This disk has an access period of 27.5 s.
- It has consumed 1.5% of its total parking cycles in 24 hours of use.
- At this rate it will start failing at 66 days.
  - Some users had reported this short lifetime.
- One solution is to increase the 'idle' timeout to 30s.

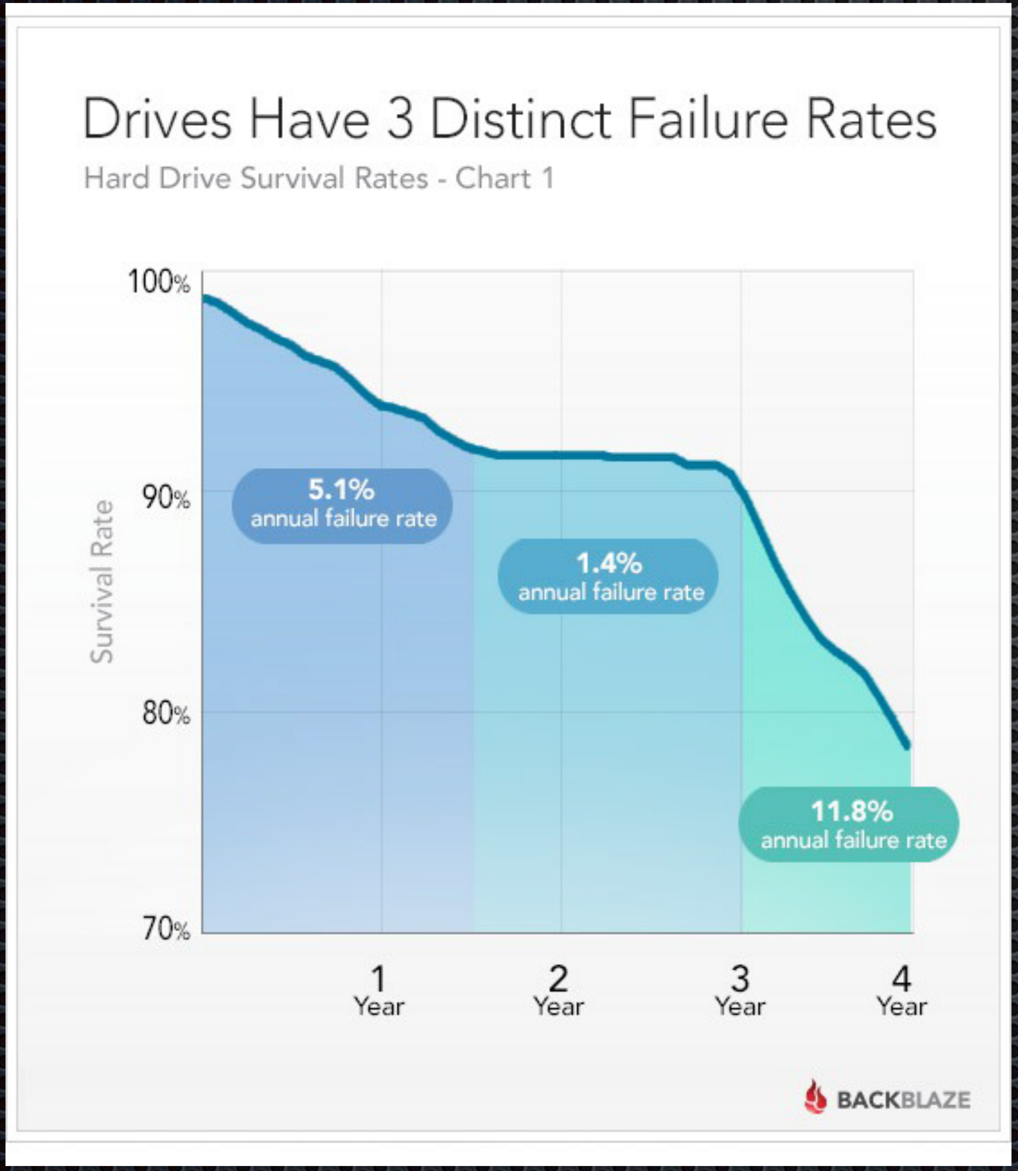
Life time of the media

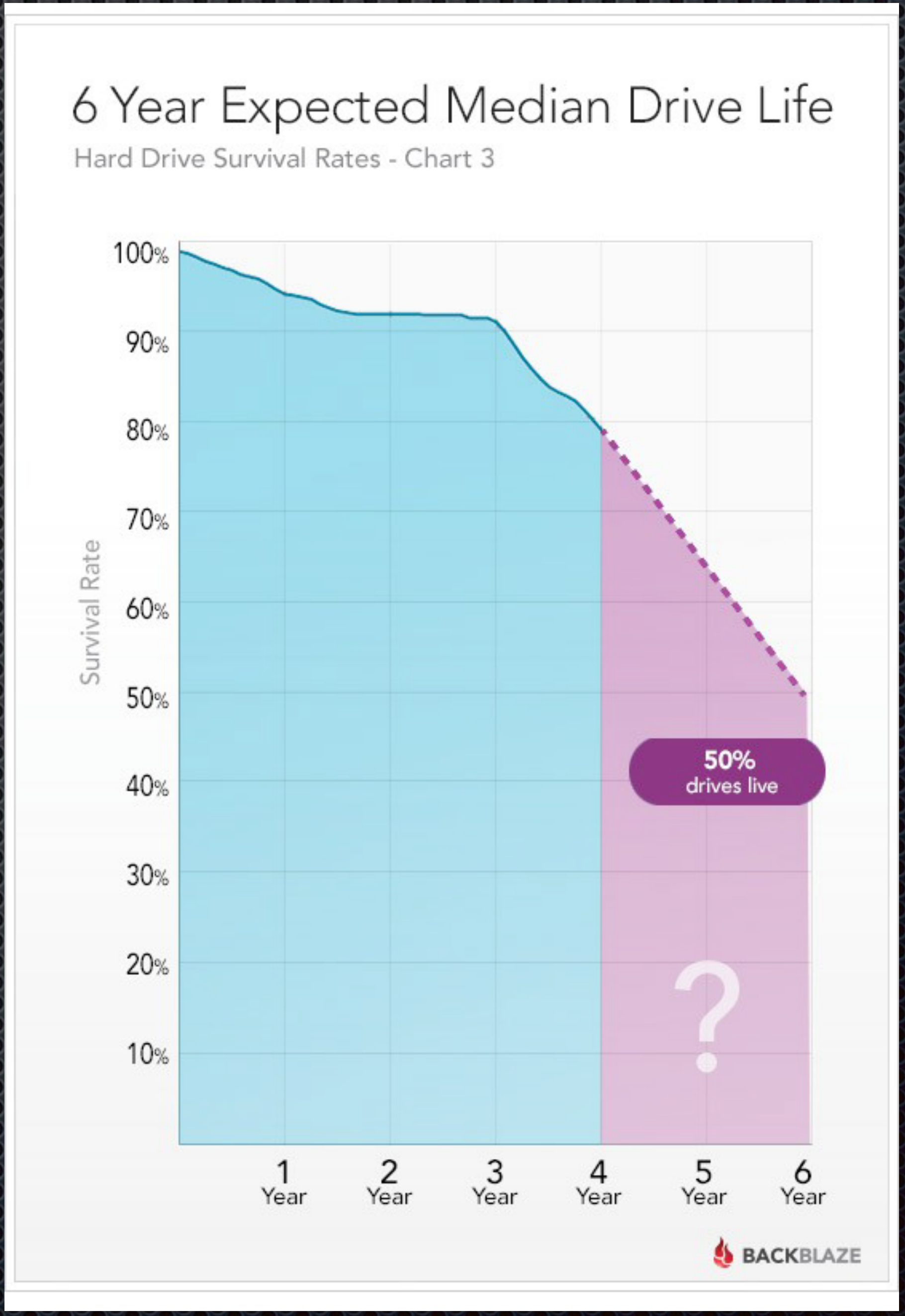
- Drive died after 151 days
- SMART counters:
  - Power On hours : 3644
  - Load_Cycle_Count: 333112
- This doesn't means that drives are defective, but that they are used incorrectly.
- **In order to use a Green Drive as a backup media the access pattern must be the correct.**

Median life time of a hard drive

Backblaze

Study based on 25,000 drives for 4 years
Desktop hard drives with 3 years guarantee





https://www.backblaze.com/blog/how-long-do-disk-drives-last

Power demand

- Appart of the media requirements of energy, each solution has associated an infrastructure that requires energy
  - Temperature, humidity and dust control
    - Tapes are more sensible to this factors than disks
  - UPS Backup energy
  - Platform dependent energy
    - Motherboards, Enclosures, Tape readers, etc.

Scalability

- One important factor when selecting a backup system is its possibility to increase its capacity as the need for space increase in time.
  - For a tape library its scalability possibilities are defined at the beginning.
  - So it is important to define which is the maximum expected space required in the future.
  - How do this impact in the cost at the first purchase?
- For a disk based solution, the limit is imposed mainly by the middleware that support the filesystem.
  - EOS at the moment can manage hundred of Petabytes.
- **For an ALICE-T1 a tape storage of 2 PB are required and 10 PB will be needed for RUN-3**

EOS as a custodial system

- Open source distributed disk storage system
- Use commodity hardware
- Scalability: hundred of Petabytes
- Redundancy by replicas and RAID-like levels (reduce effective total capacity)
- General purpose file system.
- Combined with SMR drives and power-saving capabilities reduce the total cost per terabyte.

Cost estimation

1 PB	EOS with Green Drives	Oracle StorageTek SL3000	IBM TS3500	HP ESL G3	Quantum I6000	SPECTRA T950
Tape Drives +media +library	\$67,000 \$134,000	\$162,354	\$268,477	\$301,563	\$219,344	\$229,470

Tape library costs obtained from an Oracle study of 2013.
EOS solution based on actual quotes.

Conclusions

- A solution based on Green Hard Drives seems to be a good replacement for a Tape Library System if:
  - The access pattern to the drives preserve them with small periods of high activity mixed with long periods of low activity
  - The access pattern also influence on the power consumption on the solution
  - Tape Libraries are cost effective if the amount of storage is in the order of 10 PB according to CERN experts.

Next steps

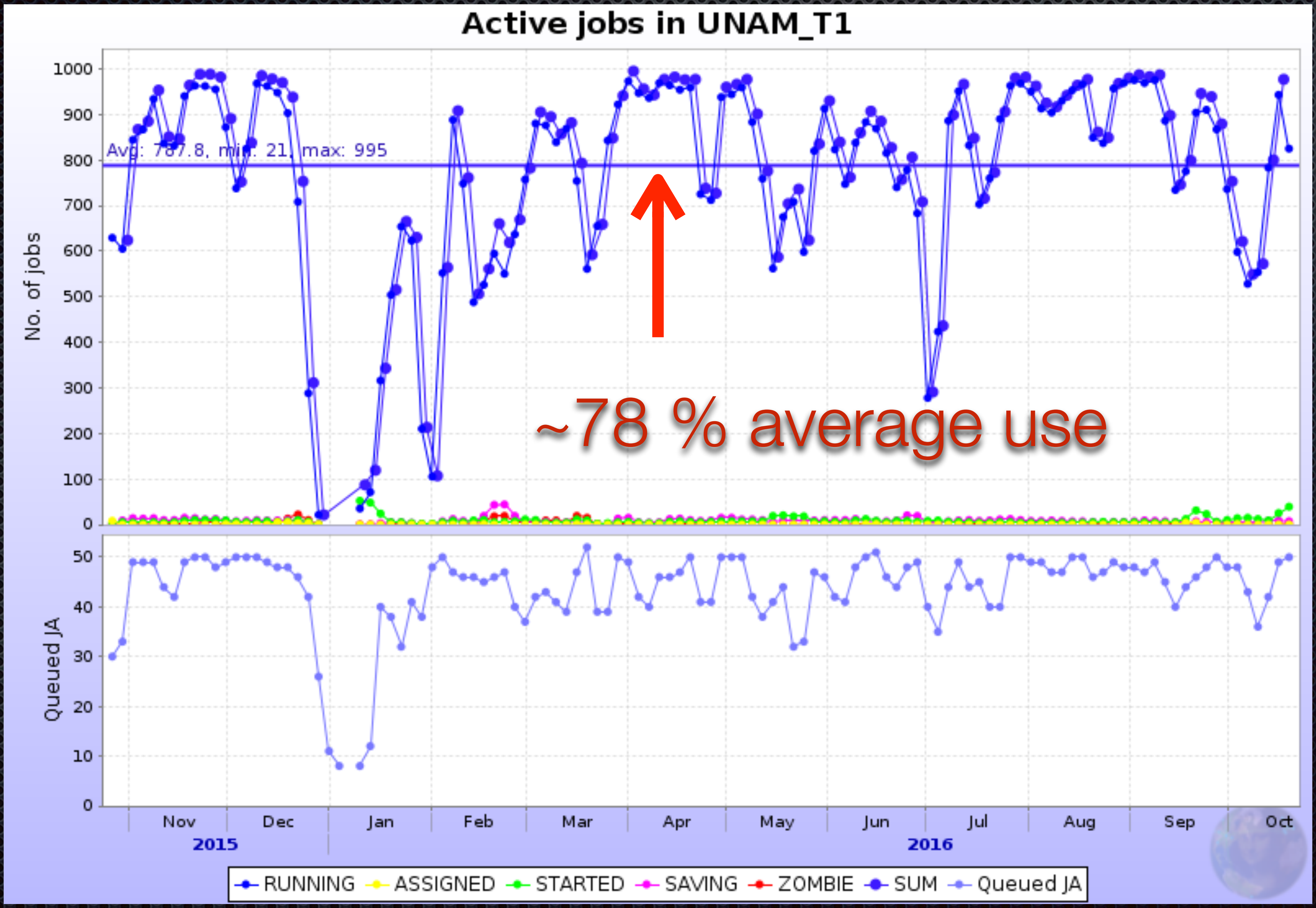
- ICN will soon have a 10 Gb fiber connection to USA
  - This accomplish the network requirement for a T1
  - In the next months ICN will install a small EOS solution based on Green Drives to test the behavior in terms of energy and pattern access of a backup system for an ALICE’s T1
  - Quotes for Tape Libraries for 2PB scalable to 10PB will be requested to providers in order to have a real comparison between solutions.
  - If Green Storage is cost effective, a proposal to the Alice data group will be done in order to replace the Tape Library for a Disk based solution and install a T1 at UNAM.

Report of the production of the ALICE's UNAM-T2

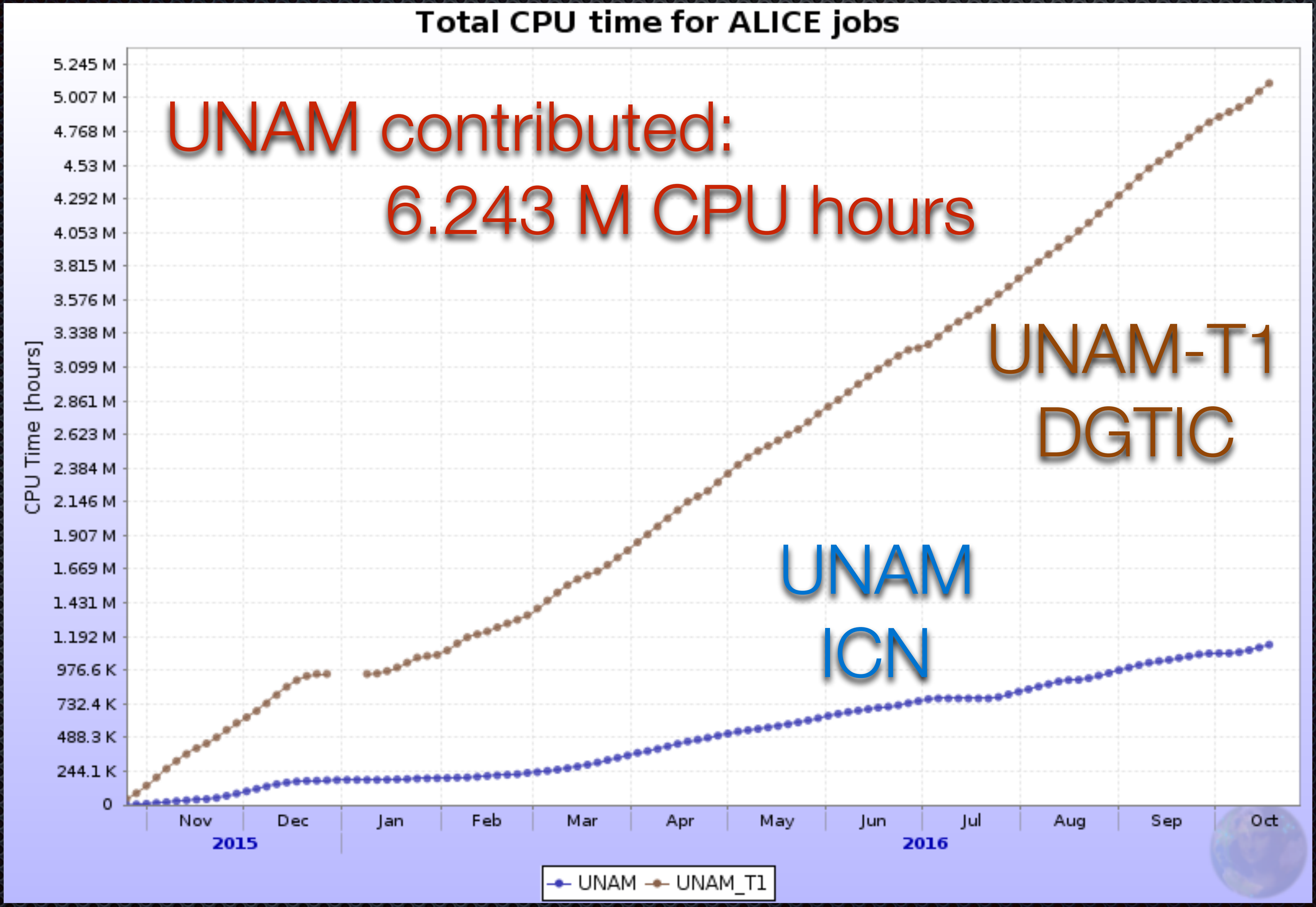
Luciano Díaz
Eduardo Murrieta
ICN - UNAM

Arión Pérez
DGTIC - UNAM

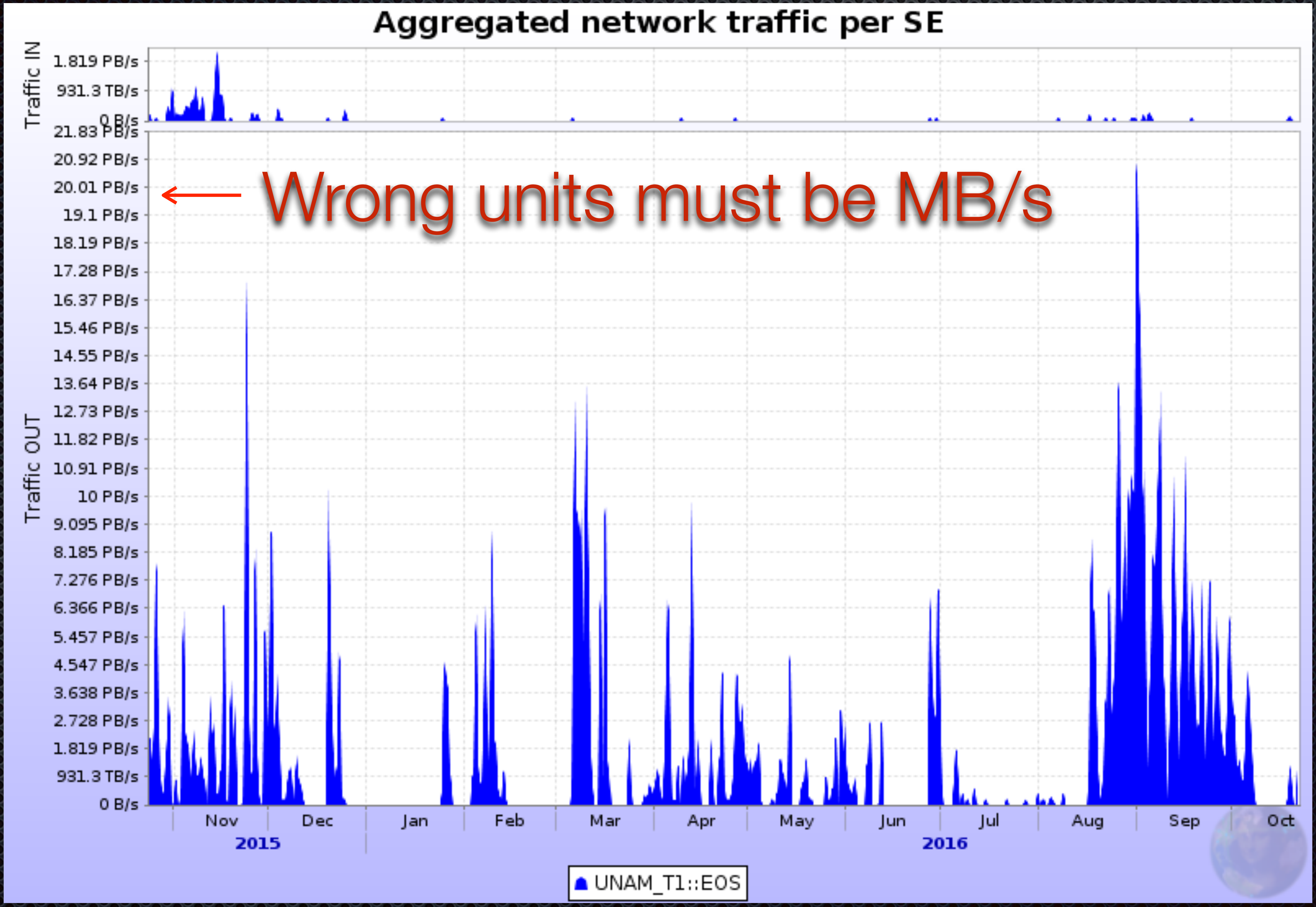
Computing Resources



CPU time contributed last year



Storage resources UNAM-T2



Storage resources UNAM-T2