



Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

# HPSS and tape storage at CC-IN2P3

Pierre-Emmanuel Brinette - Jan 15th 2020  
BELLE II France Computing Workshop

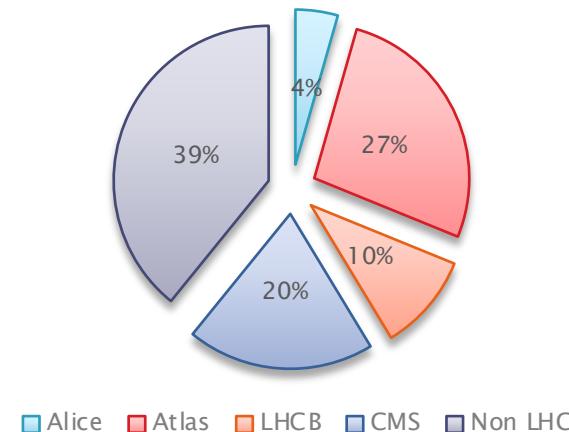
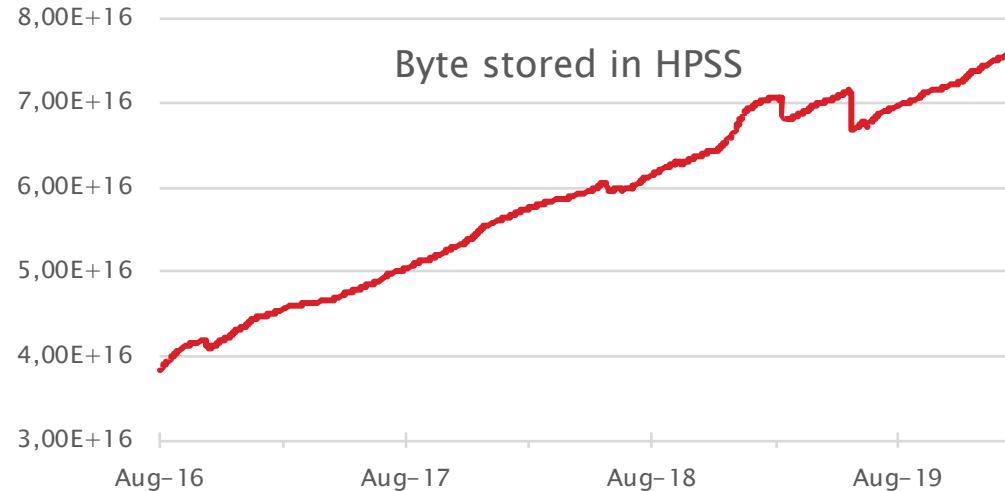
- ▶ HPSS infrastructure
- ▶ dCache configuration for atlas
- ▶ Tape infrastructure evolution



# HPSS Infrastructure

# About HPSS

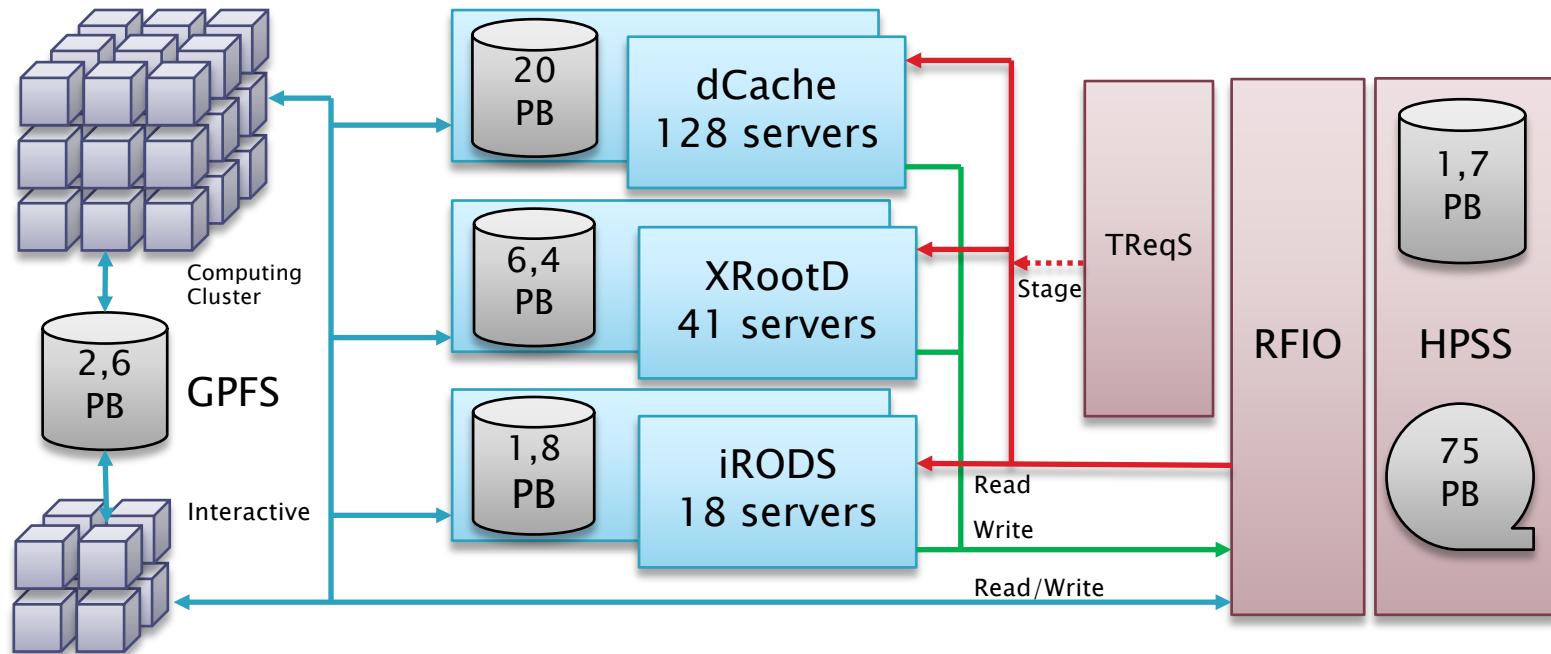
- ▶ HPSS Since 1999
- ▶ Usage (January 2020 )
  - Total : 76 PB
  - Files : 86 M
  - LHC : 46 PB (61 %)
- ▶ Growth last in 2019
  - + 6 Po ( + 8% )
  - Massive clean up last year  
( ~ 7PB deleted from tape)
- ▶ Transferts
  - 105 TB average / day
  - 66 % read access
  - Peak 250 TB / day (~ 3 GB/s)
- ▶ Forecast 2020
  - + 12 to 14 PB





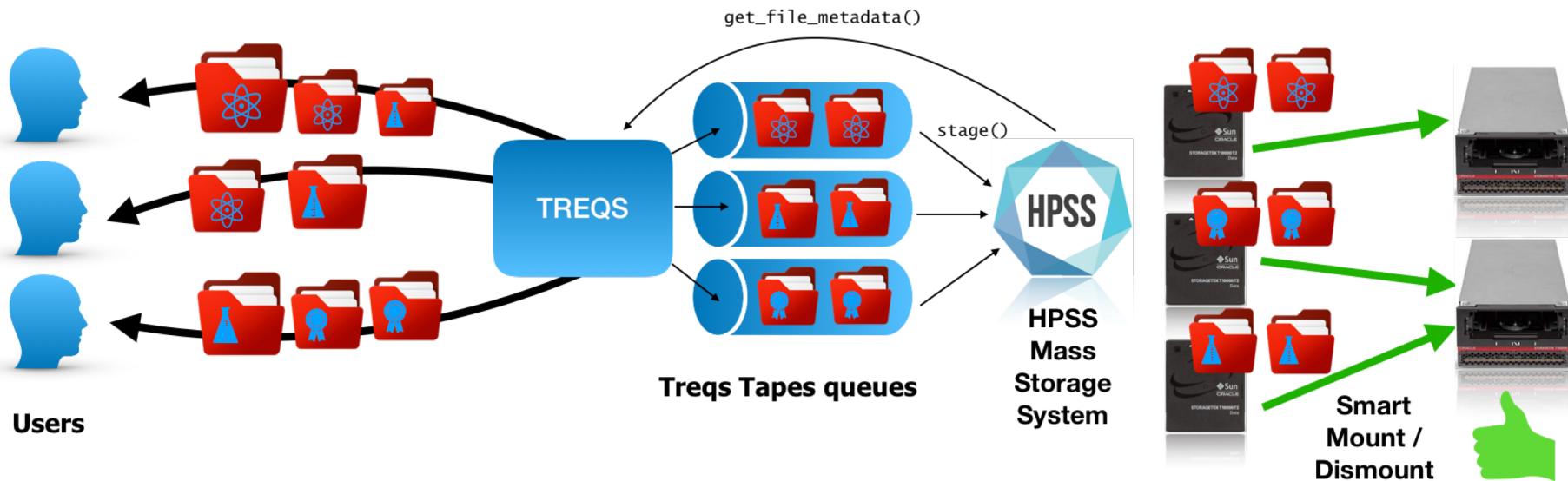
- ▶ HPSS use STK/Oracle Enterprise Technology since 2007
  - 4 libraries Oracle SL8500 / 40000 slots
  - 56 Enterprise drives T10000D
  - 8,5 TB / tape
  - 13000 T10K-T2 tape
  - 20000 empty slots
- ▶ LTO Technology used for backup only (TSM / Spectrum Protect)
  - 20 LTO4 and 17 LTO7 drives

# Storage Overview



- ▶ HPSS hpss-7.5.1.2-20190116.u9
- ▶ 85 % of HPSS access are performed through storage middleware
  - **dCache** (LCG/egee),
  - **XRootD** and **iRods**
- ▶ Still some direct access to HPSS but decreasing
- ▶ **HPSS Servers**
  - 12 Disk mover DELL R720/730 @10 Gbps
    - 1,7 PB of disk cache
  - 9 Tape mover DELL R630 @10 Gbps
- ▶ **HPSS Interface :**
  - RFIO with HPSS extensions
  - Read operations from storage middleware are handled by TREQS 2

- ▶ TREQS 2, the IN2P3 tape scheduler for HPSS
  - *Optimize read operations by grouping files per tapes*
  - *Reduce the number of mounts / dismounts of the same tape.*
  - *Limit the number of drives used for staging*
  - New version in production for 3 years [1] [2]
- ▶ Used for dCache / Xrootd / irods
  - Treqs client “wrap” the HPSS transfer command (rfcp)





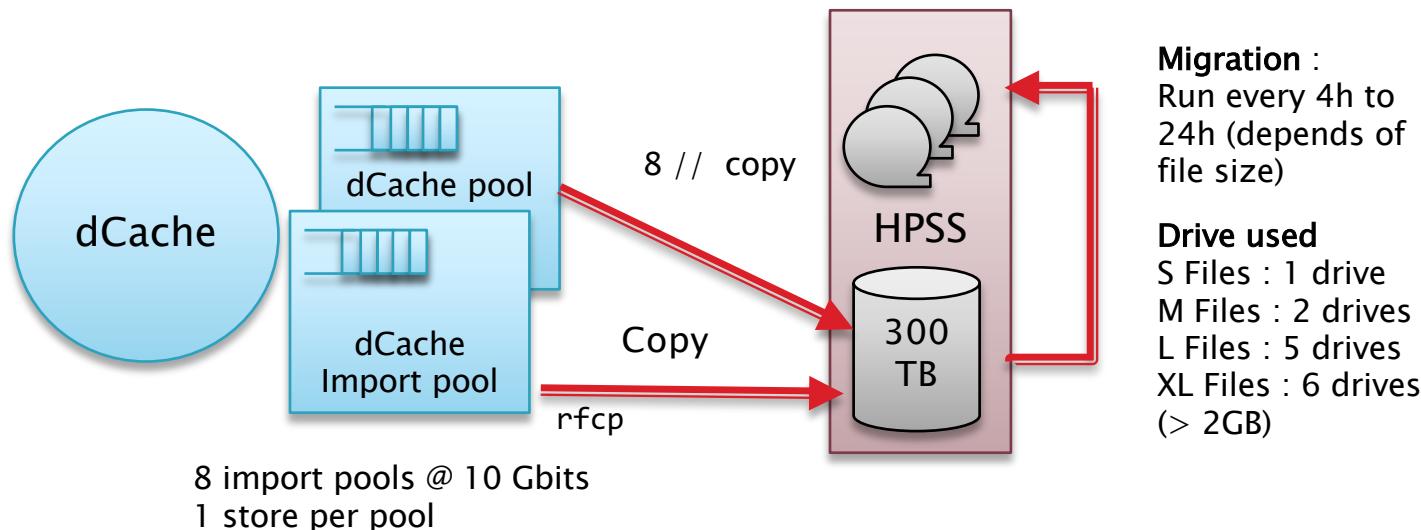
# dCache-HPSS configuration for atlas

# dCache-HPSS Configuration for Atlas (STORE)

- ▶ Store operation
  - 8 import pool , 1 connection per pool
  - 3 Files families (mctape / datatape / archive)
  - 4 storage class (based on file size)
- ▶ Files written in the same time are spread over multiple tapes
  - IE : 6 tapes for XL files.
- ▶ Filename based on pnfsid:

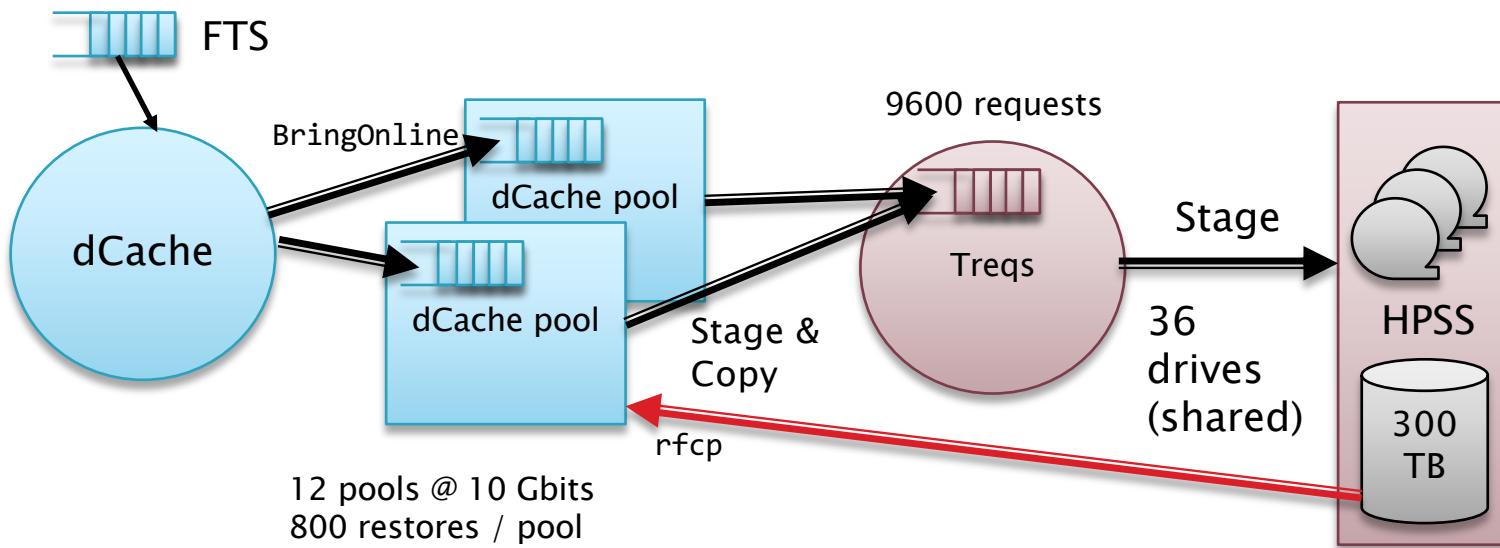
/hpss5/dcache/atlas/**datatape**/2018/10/000055C8CF15B9764B858C974D31928071D3

**datatape | mctape | archive** : dcache Spacetoken → File family



## ▶ Restore operation

- dCache submit requests over 12 pools
- Each pool can handle 800 restores
- Treqs schedule requests and stage files
  - 36 drives configured (shared between Vos)
- HPSS handle only 1 stage requests a time per drive



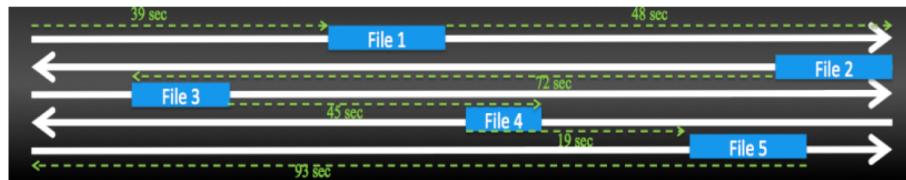


# Tape infrastructure evolution

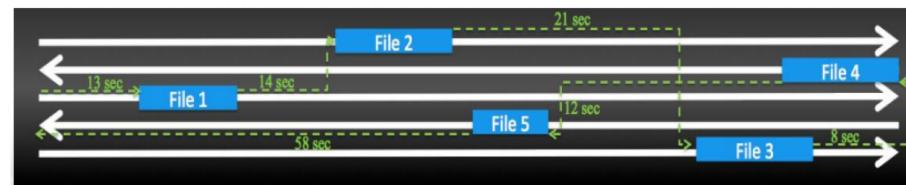
# Tape infrastructure evolution

- ▶ Oracle stopped developing “Enterprise drives” (T10000)
  - T10000-E drives won’t be marketed
  - Need to move to a new technology
- ▶ 2 scenarios :
  - Move to IBM Enterprise class tapes drives (Jaguar)
  - Keep our libraries and use LTO drives.
- ▶ IBM Enterprise tapes (Jaguar) :
  - Native capacity : **20 TB** on a JE cartridge (TS1160)
  - Short media (“Sport” Tape) for storing small files.
  - Drive support latest’s advanced features
    - 64 landing zone allowing fast positioning
    - Tape Ordered Recall (aka Recommended Access Order) and End To End Data integrity
  - Drive is NOT supported on Oracle libraries → **Need to purchase new libraries**
- ▶ LTO 8
  - Native Capacity : 12 TB
  - Media cost 25% lower than Enterprise tape and may decrease quickly.
  - Use the same R/W head than Jaguar (TMR) head and BeFe media.
  - But Only 2 landing zones → Performance lower on random recall.
  - Advanced features not supported (RAO)

- ▶ RAO : Recommended Access oOrdering
  - Drive feature to find the better path to recall a bunch of files.
  - Fully available since HPSS 7.5.1.2
  - Features only supported on « Enterprise » tapes drives



Seqquential Read :  
326 s



RAO Optimized Read :  
126 s  
Gain : 151 %

« Performance Evaluation for Tape Storage Data Recall with TS1150 Drive »  
Guangwei Che - BNL - HUF 2018

# RAO Tests : T10K-D ROA vs T10K-D vs LTO8

- Tests with HPSS 7.5.1
- Recall of tests files with and without RAO

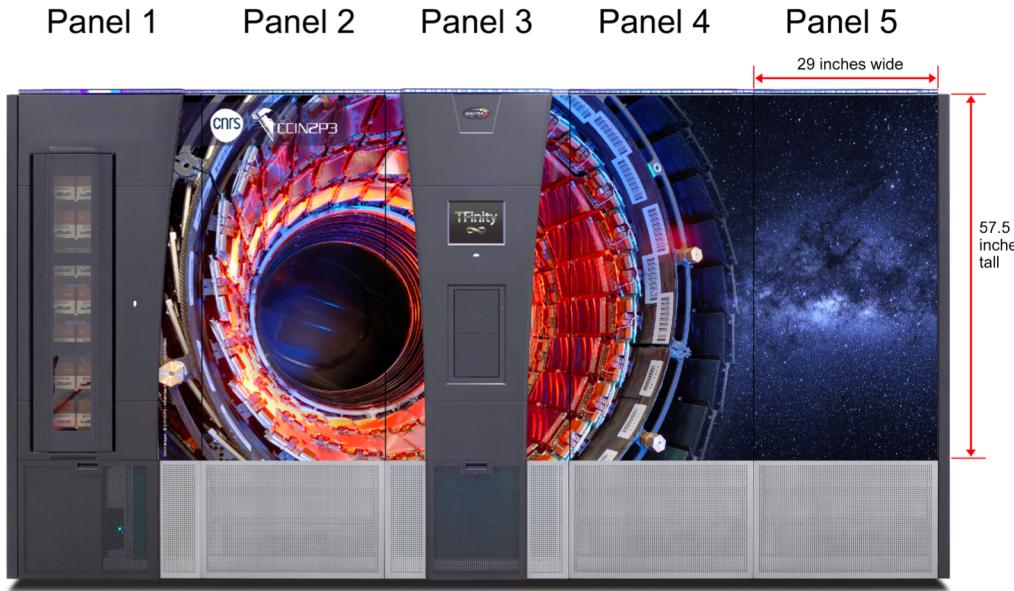
- Tests files : 2200 MB
- 1 T10K-D filled 3646 files
- 1 LTO-8 filled 5205 files
- Samples drawn randomly
- Staging with hpss\_cache (ordered and non ordered ) and hpss quaid (RAO)

T10K-D/RAO vs LTO8  
Gain : ~ 300 % !

T10K-D/RAO vs non RAO  
Gain : ~ 138 %

Test	Sample	Duration	Rate
T10K-D : Unordered read	25 files	19m41s	41 MB/s
T10K-D : Offset ordered read	25 files	19m05s	48 MB/s
	50 files	34m58s	58 MB /s
T10K-D : <b>RAO</b> ordered read	25 files	8m0s	114 MB/s (gain : 137%)
	50 files	13m10s	139 MB /s (gain: 139% )
LTO8 : Unordered read	25 files	23m26s	39 MB/s
LTO8 : Offset ordered read	25 files	24m9s	38 MB/s
LTO8 : <b>Quaid</b> ordered read	25 files	25m5	37 MB/s

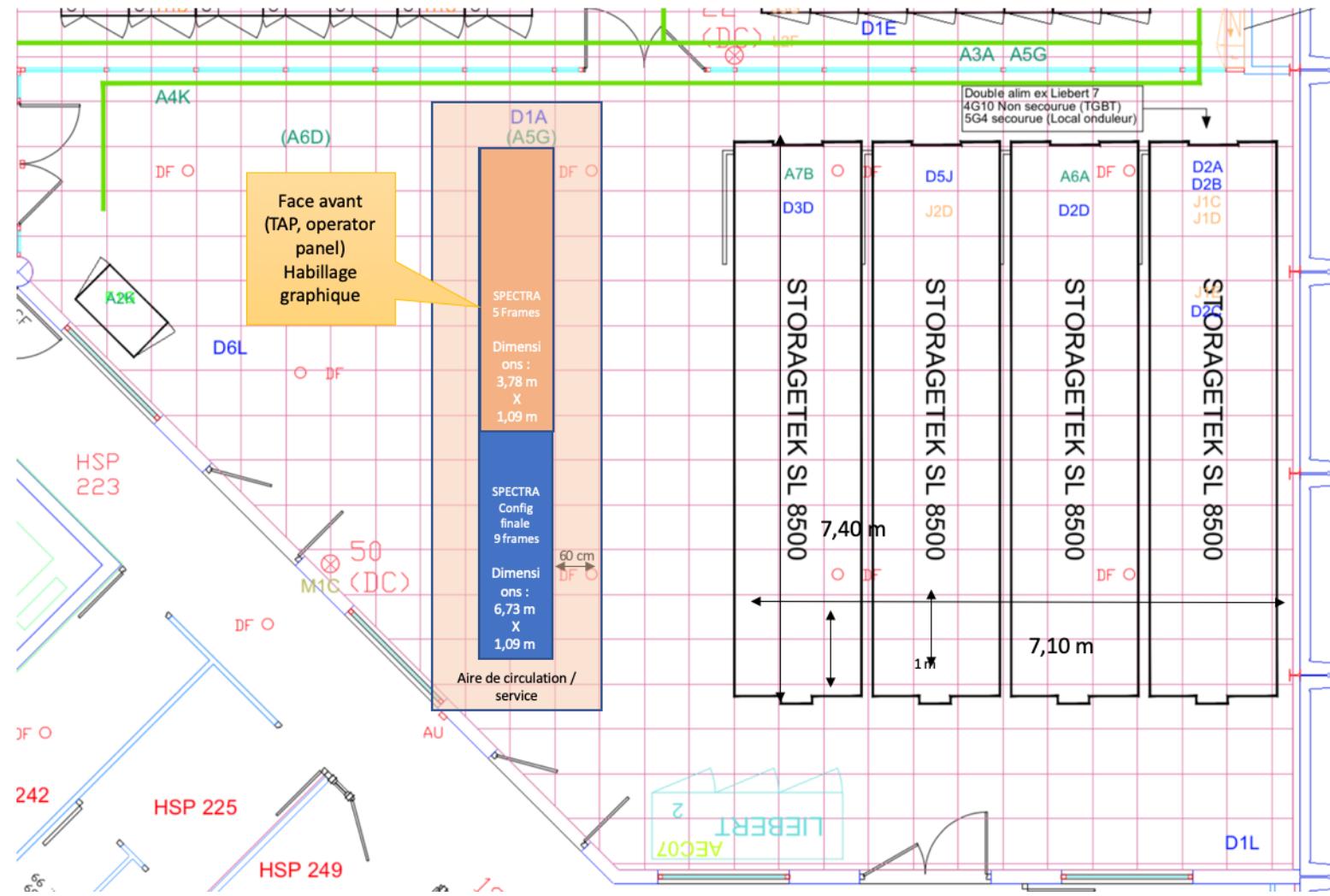
- ▶ For CC-IN2P3 Enterprise drive seems to be the right direction
  - Today's LTO 8 performances are lower than current T10K-D
- ▶ Acquisition soon of a « small » libraries
  - 12 TS1160 drives
  - 20 PB minimum
  - Proposal for tender in Q4 2019
  - Spectralogic Tfinity solution selected
    - Best value for money



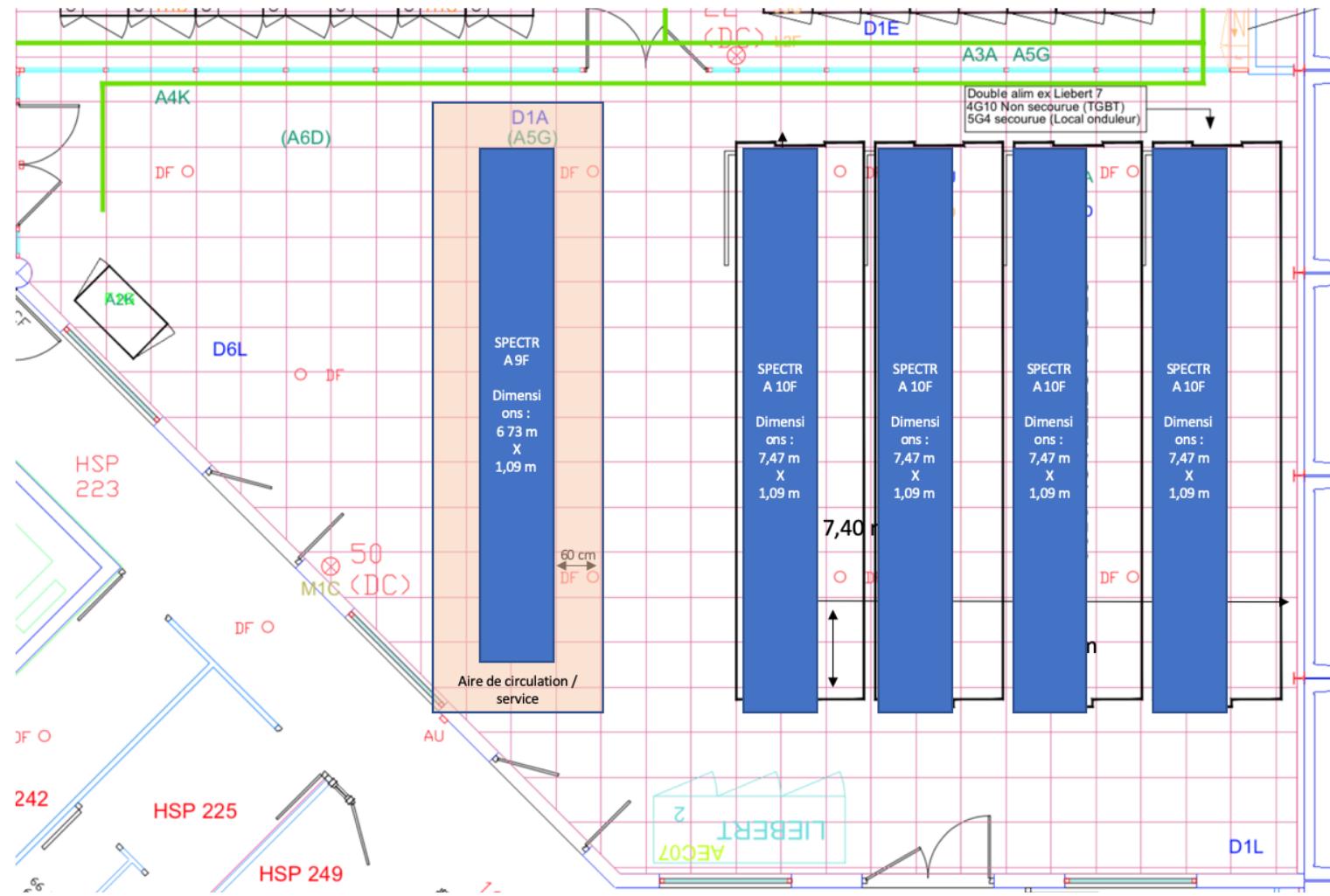
- ▶ New robot
  - SPECTRA LOGIC Tfinity
  - 5 frames
  - 3300 tape slots
- ▶ IBM Enterprise tape drive
  - Jaguar E Tape
    - 20 TB / tape
  - 12 IBM TS1160
    - 400 MB / s
- ▶ Capacity : + 60 PB
- ▶ Expandable ( 2023 ):
  - Up to 9 frames
  - ~ 7000 tapes (140 PB)
  - Up to 48 tape drives max
- ▶ Delivery soon !



# Library Implantation



# 2023 and beyond



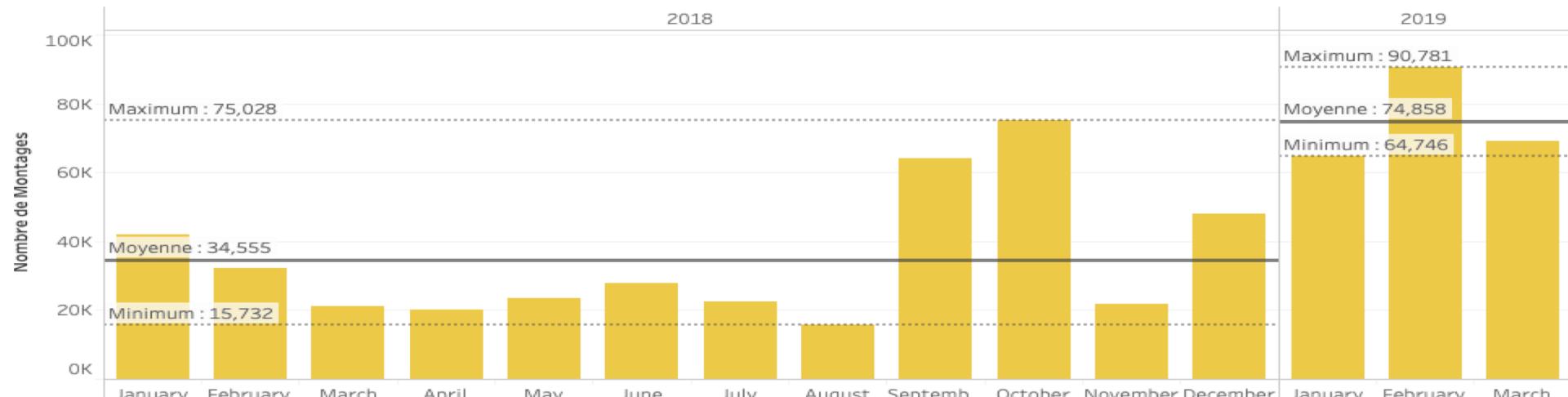
# Thank you

- [1] <https://conference-indico.kek.jp/indico/event/28/session/10/contribution/25>
- [2] <https://indico.cern.ch/event/466991/contributions/1143626/>

# Backup slide

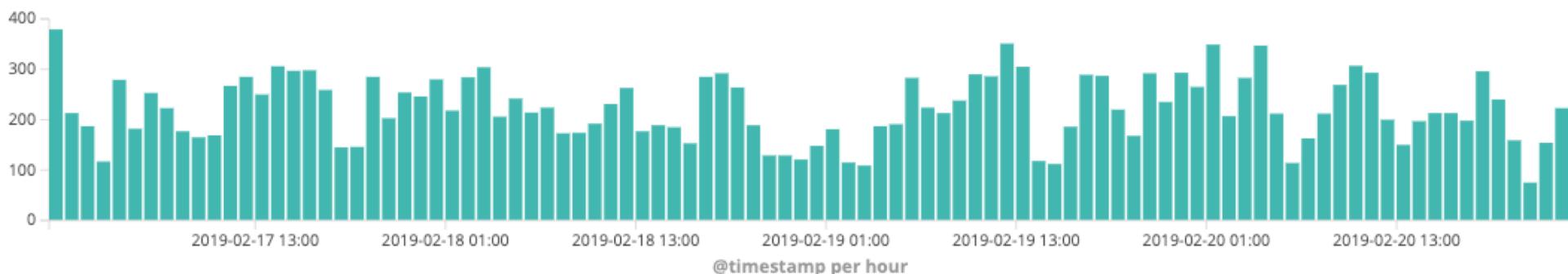
# Mount statistics (T10K)

Montages par modèle et par an (onglet : Montages par an)



16-21 feb 2019

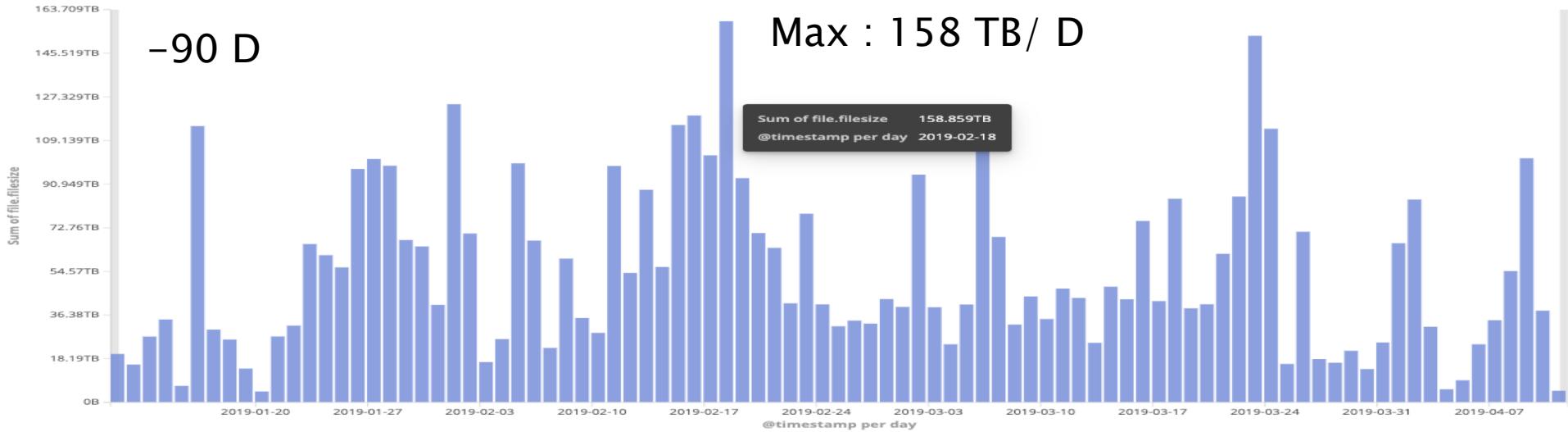
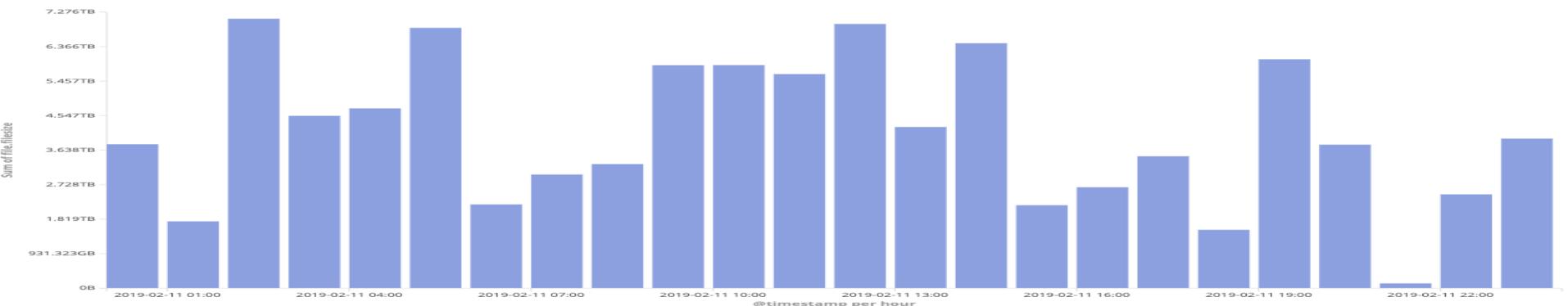
Max : 380 m/h  
Max 5500 /d : (222 /h avg)



# Treqs Staging performances (Q1 2019)

11 feb. 2019

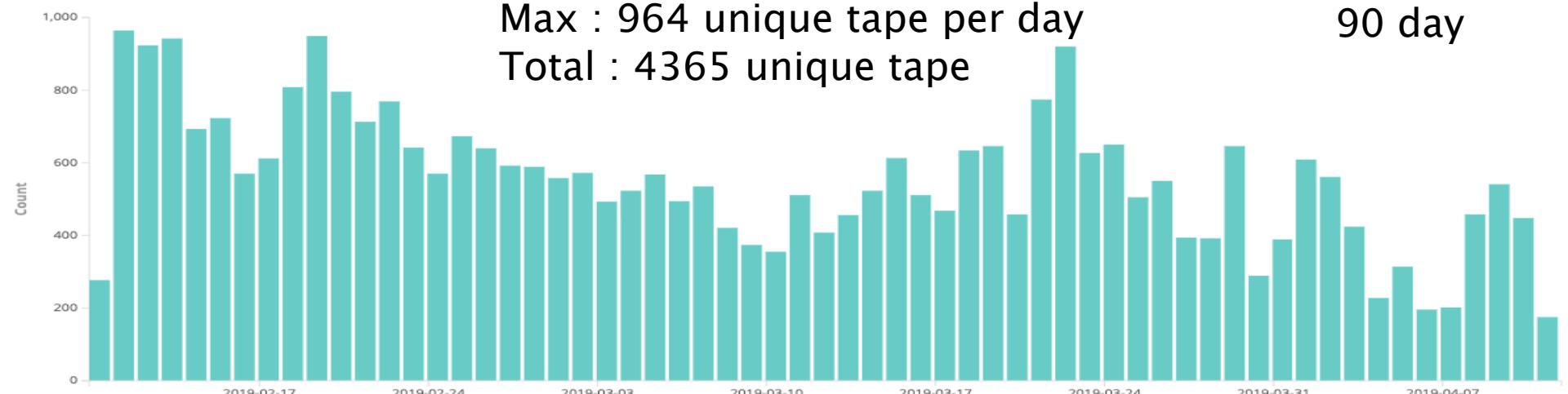
Max : 7,1 TB/h (2 GB/s) (36 Drives) → 55 MB/s/drive (avg)  
98,5 TB/d (4,5 TB/h, 1,14 GB/s)



# Treqs Statistics over 3 months (Q1-Q2 2019)

Max : 964 unique tape per day  
Total : 4365 unique tape

90 day



11 fev 2019

Max : 295 tapes/hour

