Météo-France Site Report

HPSS-France – Toulouse [2018/04]

Agenda

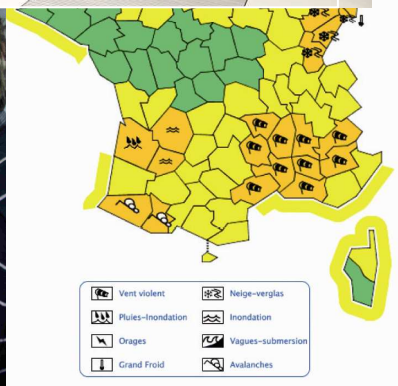
- Météo-France in a few words
- Mass-Storage Growth
- Key features, architecture and evolutions
- Migrate and purge Policies
- Small files,
- HPSS activity and drives statistics
- Gets from tapes
- Future : straight to Mars
- Issues
- Item for discussion

La protection des personnes et des biens : de l'observation à l'avertissement

COLLECTER LES DONNÉES

PRÉVOIR ET EXPERTISER

Le supercalculateur est au cœur du dispositif !



AVERTIR

DIFFUSION SIMULTANÉE

Autorités

- Sécurité civile
  - COGIC
  - Préfet de zone de défense
  - Préfet de département
- Ministère de l'écologie (CMVOA)

Médias

- Présentateurs météo
- Journaux d'information

Population

- Site internet

Contexte général

- Un des objectifs du COP 2017-2021 est l'amélioration des modèles de prévision numérique du temps :

  – augmentation massive des données assimilées, généralisation des assimilations d'ensemble
  – fusion des modèles déterministes et ensemblistes => production purement ensembliste en Métropole à terme
  – augmentation de la résolution des modèles de petite échelle sur certaines zones d'intérêt particulier (aéroport, Sud-Est, ...)
  – augmentation de la résolution des modèles OM (même résolution que les modèles "métropole").

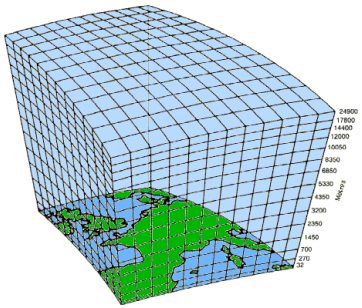
-> La clé du succès : la puissance de calcul

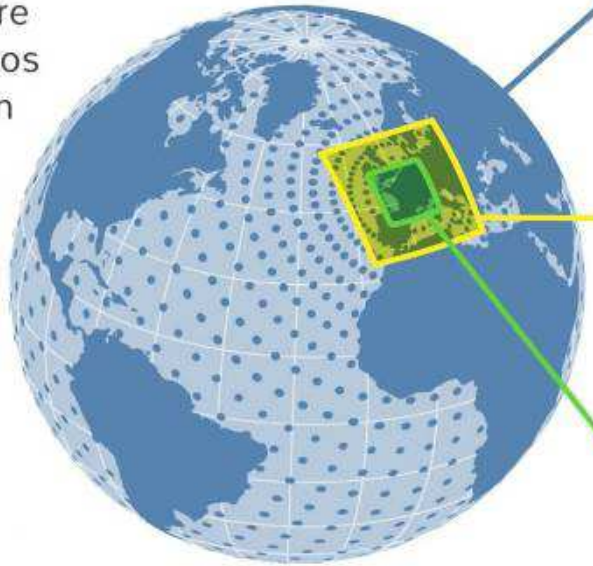
Adapter la maille aux phénomènes à prévoir

Le HPC au cœur du dispositif pour la prévision numérique du temps

DES MODÈLES DE PLUS EN PLUS PRÉCIS

Différents modèles permettent à Météo-France de construire des scénarios de prévision du temps.





Modèle global ARPEGE
Résolution de 7,5 km sur la France, 36 km aux antipodes
105 niveaux verticaux
Prévision jusqu'à 4 jours et demi

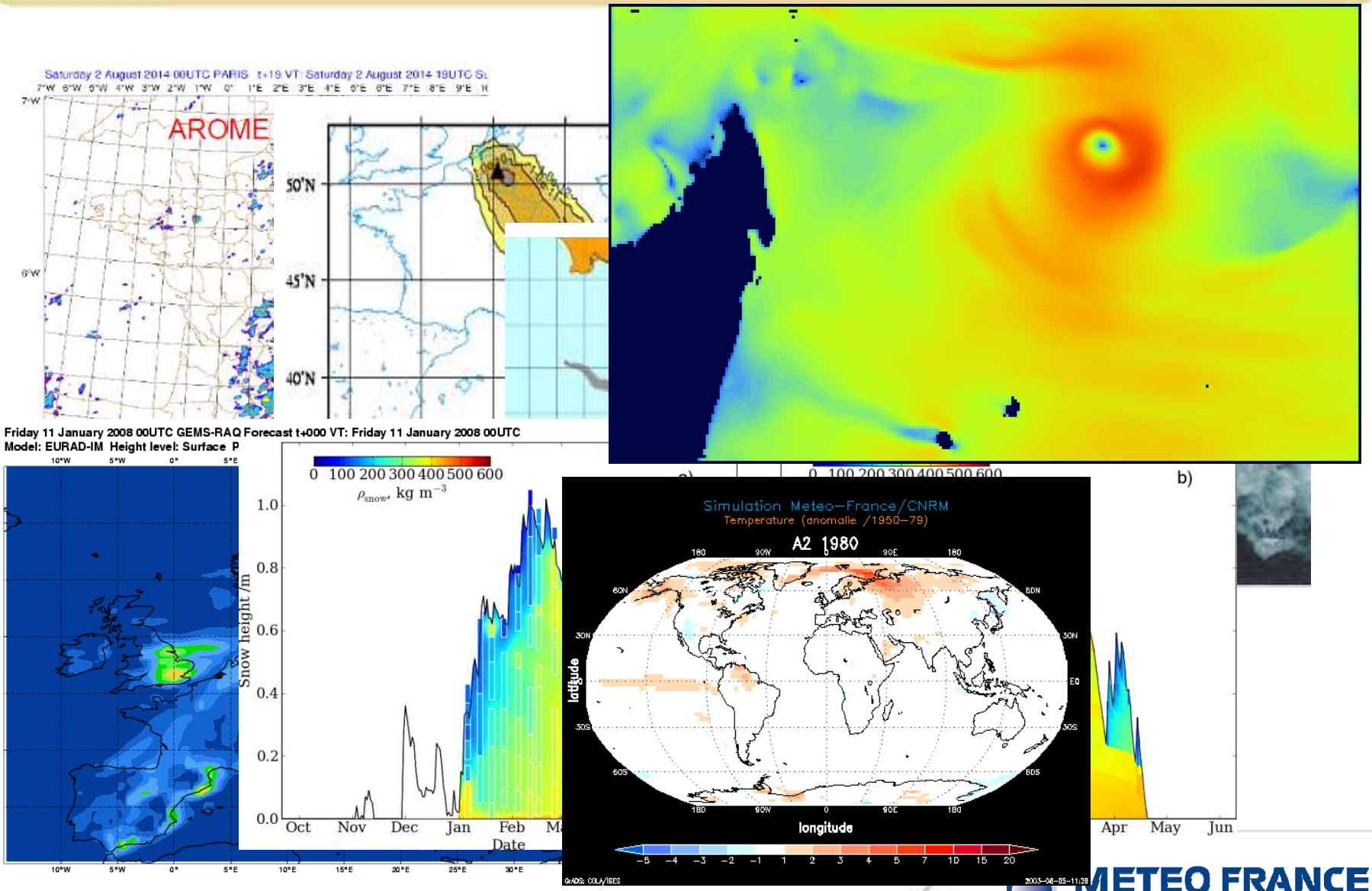
Modèle régional ALADIN
Résolution fixe de 7,5 km,
Centré sur n'importe quel point du globe

Modèle régional à maille fine AROME
Résolution de 1,3 km sur la France métropolitaine
90 niveaux verticaux
Prévision jusqu'à 36 heures d'échéance

Pour la moyenne échéance, utilisation du modèle du Centre Européen de Prévisions Météorologiques à Moyen Terme

Les prévisionnistes ont aussi accès aux prévision des autres principaux services météorologiques (NWS, Met Office, DWD)

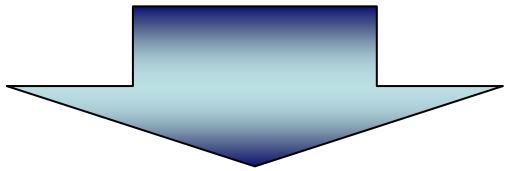
Les applications du HPC



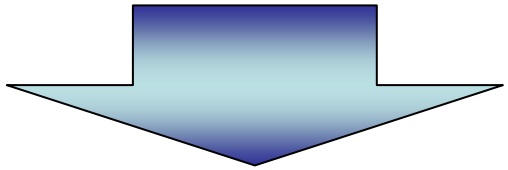
METEO FRANCE
Toujours un temps d'avance

Storage context

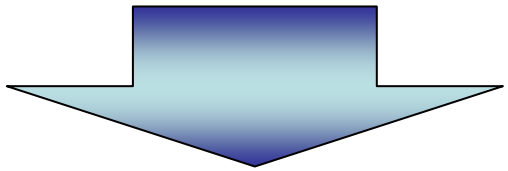
Increase the forecasts quality



Increase the models definition and run frequencies

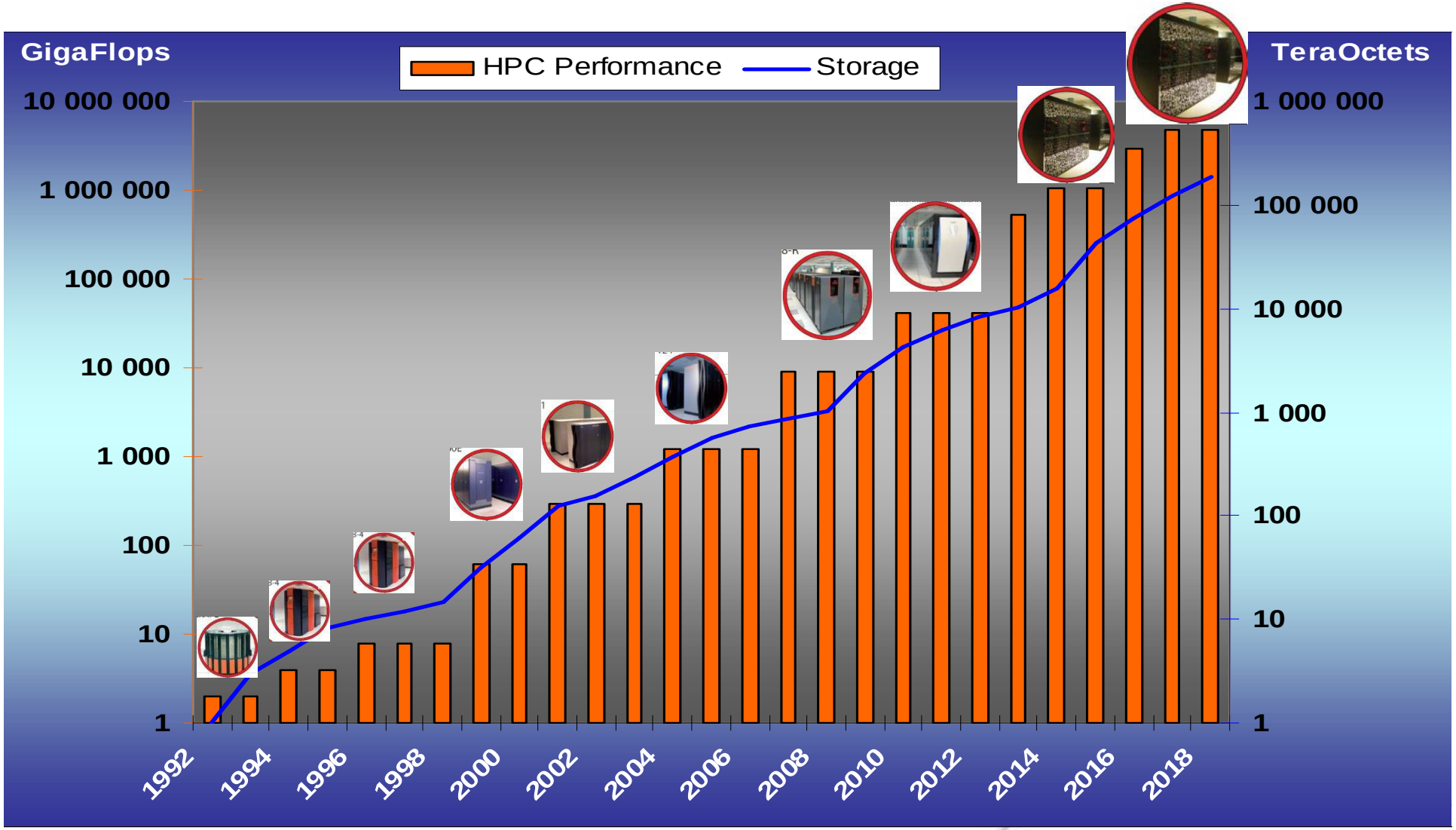


Increase the HPC performance (BULLX 2013 and 2016)



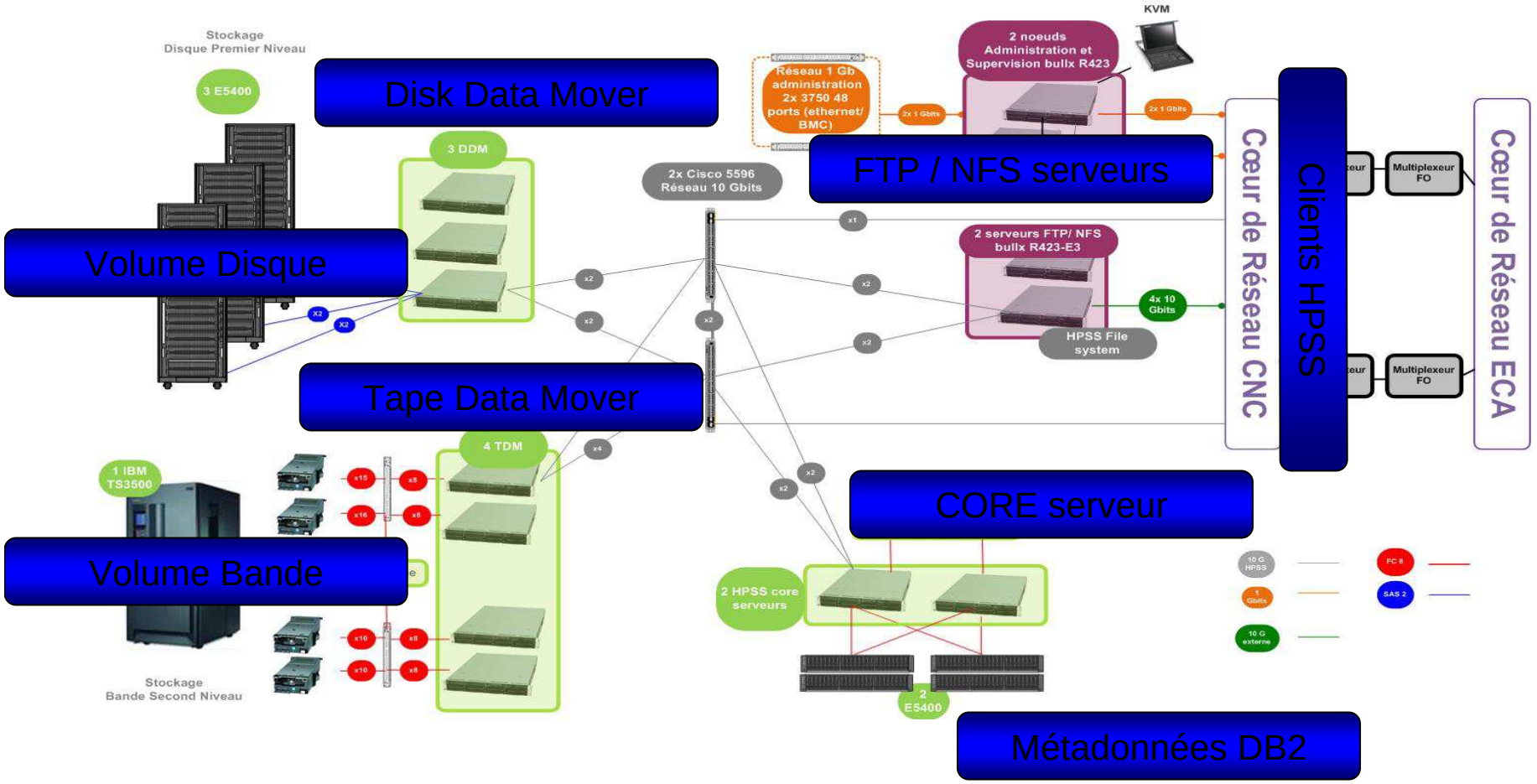
Increase the mass storage capacity and performance (HPSS 2014-2018)

HPSS-Storage Growth



METEO FRANCE
Toujours un temps d'avance

HPSS architecture V 7.4.3



METEO FRANCE
Toujours un temps d'avance

Key features

	2014	2015	2016	2017	2018
Usefull capacity (Pio)	27	43	72	120	180
Number of files (Mfil)	266	345	479	745	1 130
Double copy (Tio)	200	240	280	465	798
Cache disk (Pio)	3	6	12	14	17
Max put / day (Tio)	173	333	665	732	825
Max get / day (Tio)	146	280	532	599	665
Get from disk * / 10 s	200	332	665	731	800
Get from tape */ 600 s	146	173	439	486	519

* : files of 100Mo

Architecture evolution

	2014	2015	2016	2017	2018
FTP/NFS Servers	2	3	6	6	7
DDM Servers	3	6	6	8	8
Netapp E5500 (Datas)	3	5	8	9	10
TS 3500 HA	1	2	2	2	2
TS 4500 HA			1	2	2
TDM Servers	4	6	10	10	12
TS 1150	51	60	140	140	140
TS 1155				16	24
JD	3000	4880	8240	13480	17480

Migrate et Purge Policies

- Migrate
  - All new files are first writen to disk
  - A copy on tape is made 2 hours after creation
- Purge
  - All small files are kept on disk
  - MF requirement :
    - New files stay on disk for a minimum of 20 days after creation
  - Purge statistics :
    - Low latency – Large Files : no purge
    - Production – Large Files : 2 à 3 mois
    - Research – Large Files : environ 1 mois

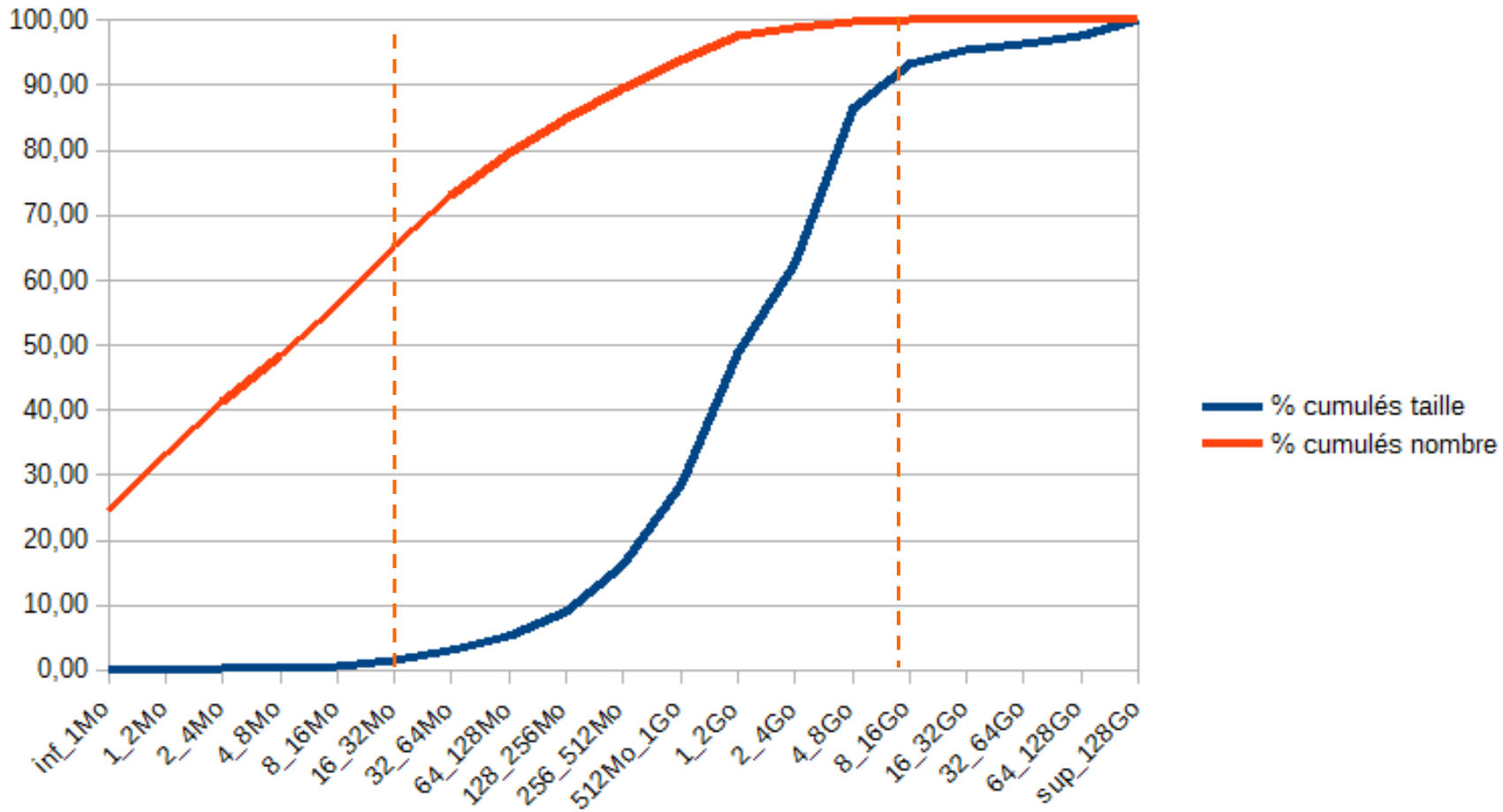
Small files

- Taille moyenne des fichiers = 252 MB
- Taille moyenne des fichiers créés depuis 2016 = 264 MB
- Distribution des fichiers selon leur taille :

Seuil	Nombre (%)	Volume (%)	Volume en To
<32Mo	65,21	0,8	735
<128Mo	79,52	2,23	2000
<256Mo	84,82	3,87	3555

-> seuil conservation des fichiers sur disque augmenté de 8 à 32Mo

Distribution : by range (cumulative)



METEO FRANCE
Toujours un temps d'avance

Mass storage is not a cold archive

- **Data stored 2017/12/31 :** 84 Po for 366 M files
- Data activity during 2017 :
  - Storage growth : +37,9 Po and + 100 M files
    -> + 82% (+64 in 2017)
  - Data put : 46,4 Po (26,3 Po in 2017)
  - Data read : 17 Po ; 36,7 % of put
  - Data deleted 8,5 Po ; 13,8 of put

Mass storage is not a cold archive

- **Data stored 2018/03/31 :** 90 Po for ≈400 M files
- Data activity during 2018/03 :
  - Storage growth : +3,7 Po
  - Data put : 6,11 Po
  - Data read : 2,06 Po
  - Data deleted 2,4 Po

Météo-France specificity : high retreival rate

Statistics for 2018/03:

	Min	Max	avg
Put	95.7TB	275.2TB	200.4TB
Get	151.9TB	30.3TB	67.8TB

METEO FRANCE
Toujours un temps d'avance

BDFH Statistics

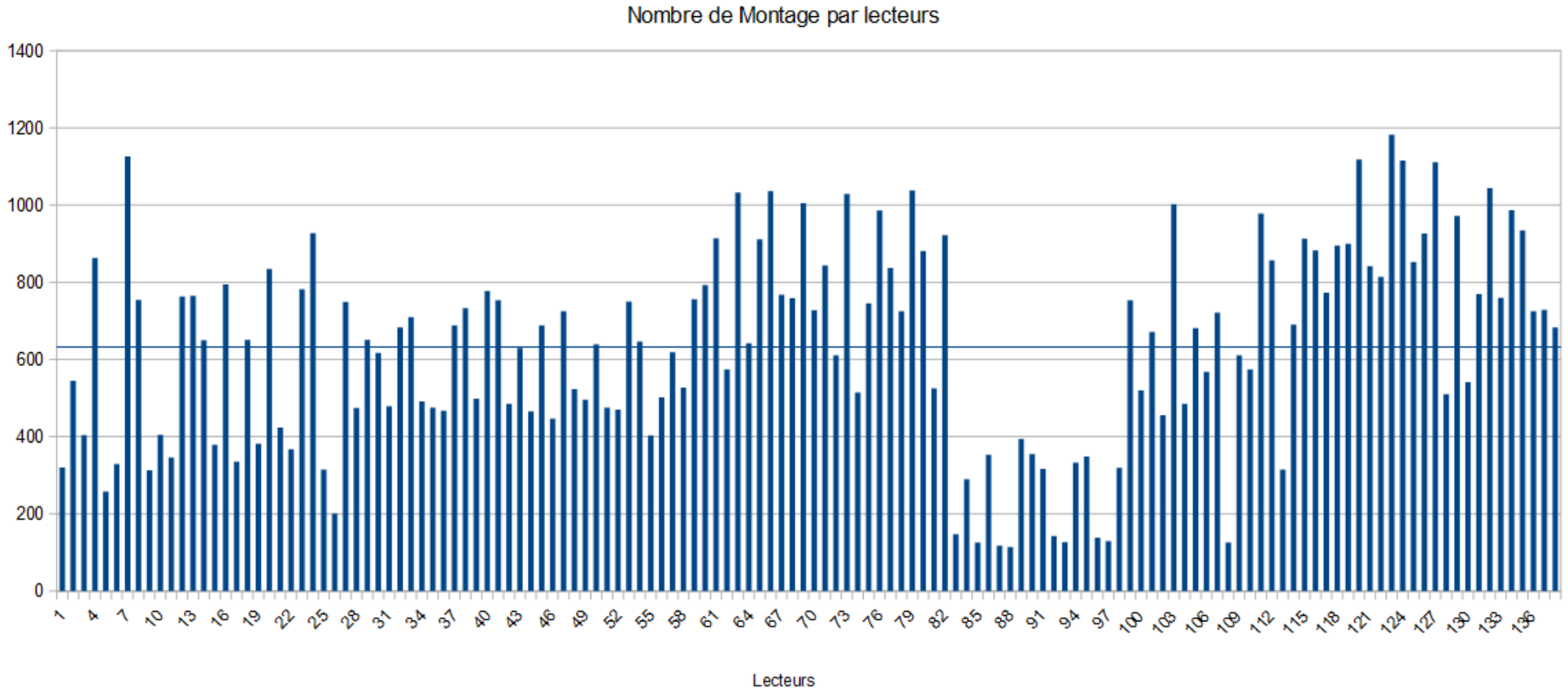
- BDFH in production since 2018/03
  - Very welcome feature : Interceptor
    - Trap all “open file”
    - If file is on disk then continue
    - Else stage file and wait

- Statistics during Mars 2018 :
  - Total request : 4 M
  - Total staged : 840 K ; 426 To
  - Request per day : 27 K
  - Max per day : 127 K

Get from tapes through BDFH

- Mount statitics during Mars 2018 :
  - Total : 32 195
  - Mount per day : 1 038
  - Max : 1 825
- Accessed files profile (Mars 2018).
  - < 100Mo: 56.04 % - Nb request 470 434
  - 100Mo et 150Mo: 13.77 % - Nb request 115 634
  - 150Mo et 200Mo: 1.63 % - Nb request: 13 711
  - 200Mo et 300Mo: 2.84 % - Nb request: 23 869
  - 300Mo et 500Mo: 5.58 % - Nb request: 46 914
  - 500Mo et 1Go: 8.94 % - Nb request: 75 072
  - > 1Go: 11.16 % - Nb request: 93 717

Drives mounts during Mars 2018 (TS3500 P1 & P2 & TS4500P3)



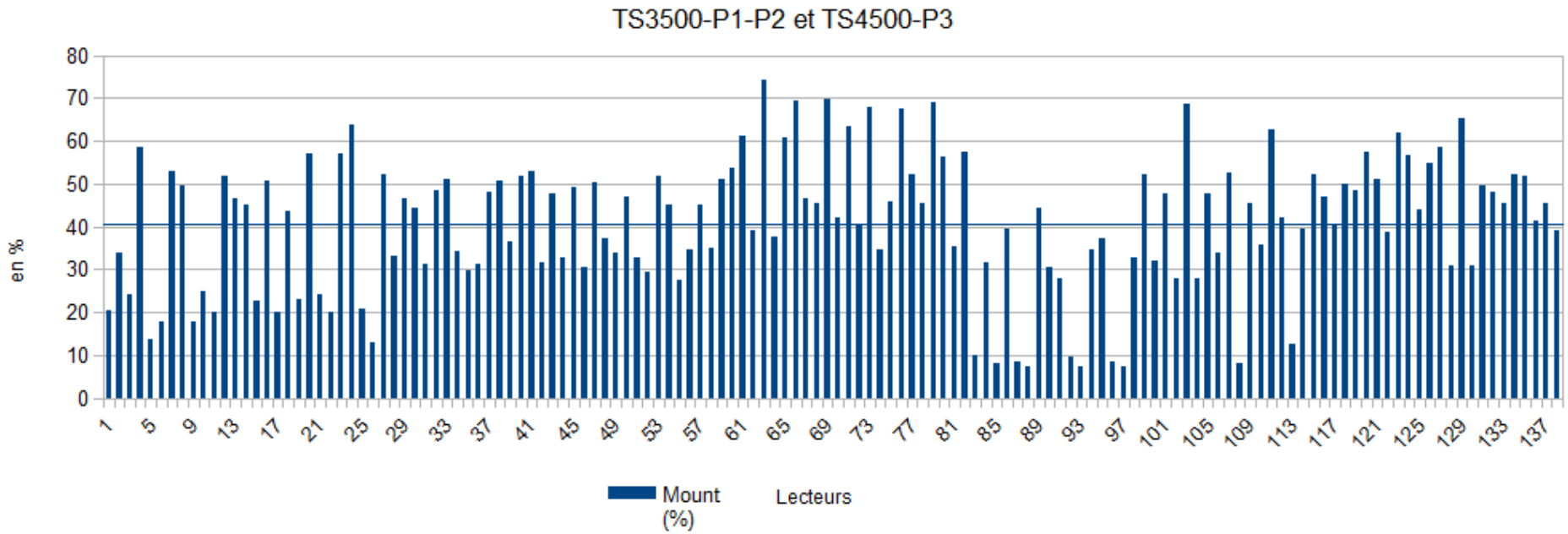
In all : 121 mounts / hour

2909 mounts / day

METEO FRANCE
Toujours un temps d'avance

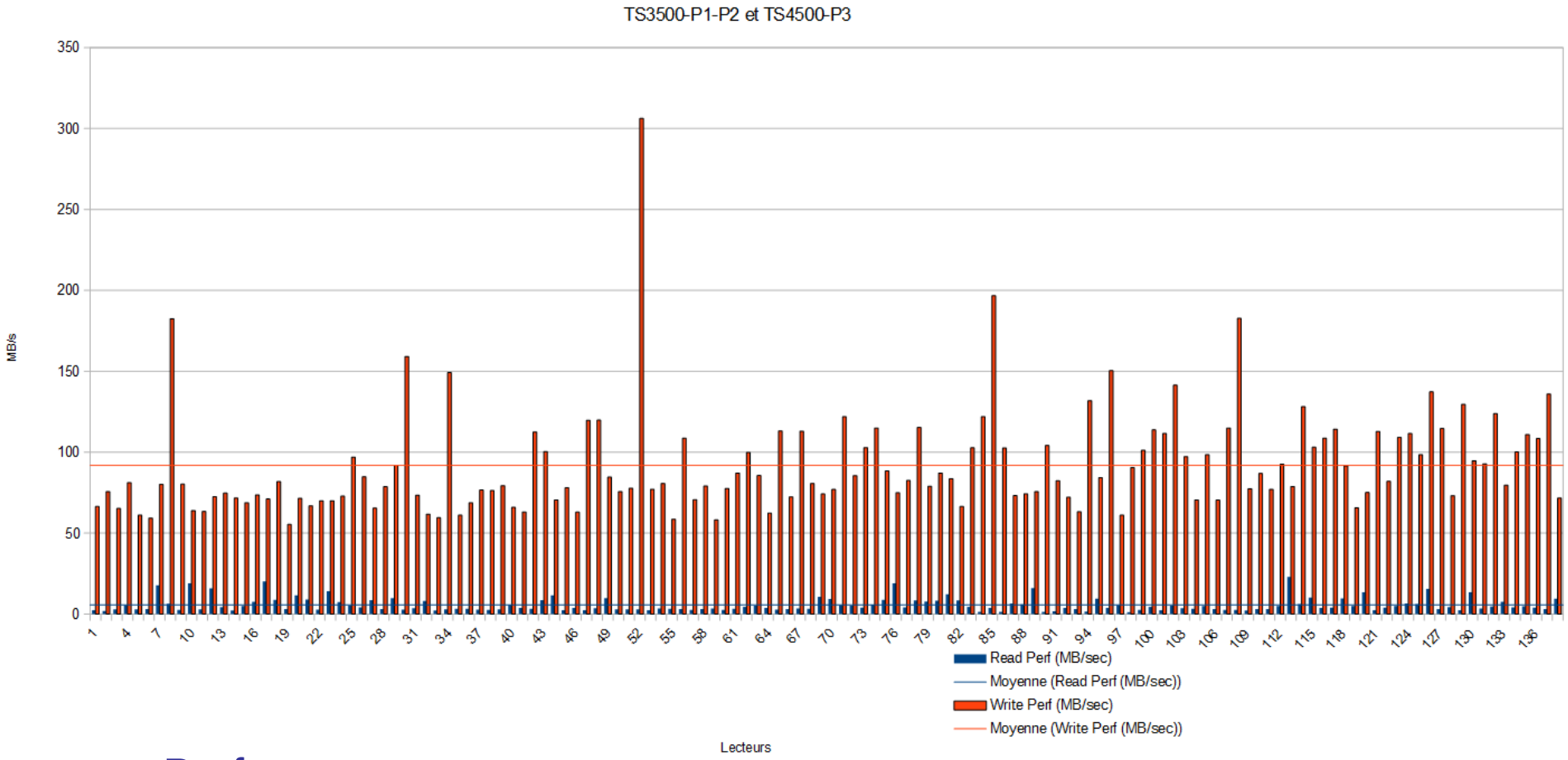
Drive mount time during the last month

- TS3500 P1 & P2 et TS4500 P3 (140 drives)



Average : 40%

Read/Write drives performances



- Performances :
  - Write : 91 Mo/s (in mars 2018)
    - Tape mark every 45s -> 60s
  - Read : 5.6 Mo/s (in mars 2018)

METEO FRANCE
Toujours un temps d'avance

Libraries and Drives statistics

- Compression ratio :
  - TS1150 : 159%
  - LTO6 : 107%

- Robot statistics :

	Accessor	X km	Y km
TS3500P1	A	2 713	1 233
TS3500P1	B	2 236	1 242
TS3500P2	A	205	172
TS3500P2	B	817	420
TS4500P3	A	333	377
TS4500P3	B	312	232

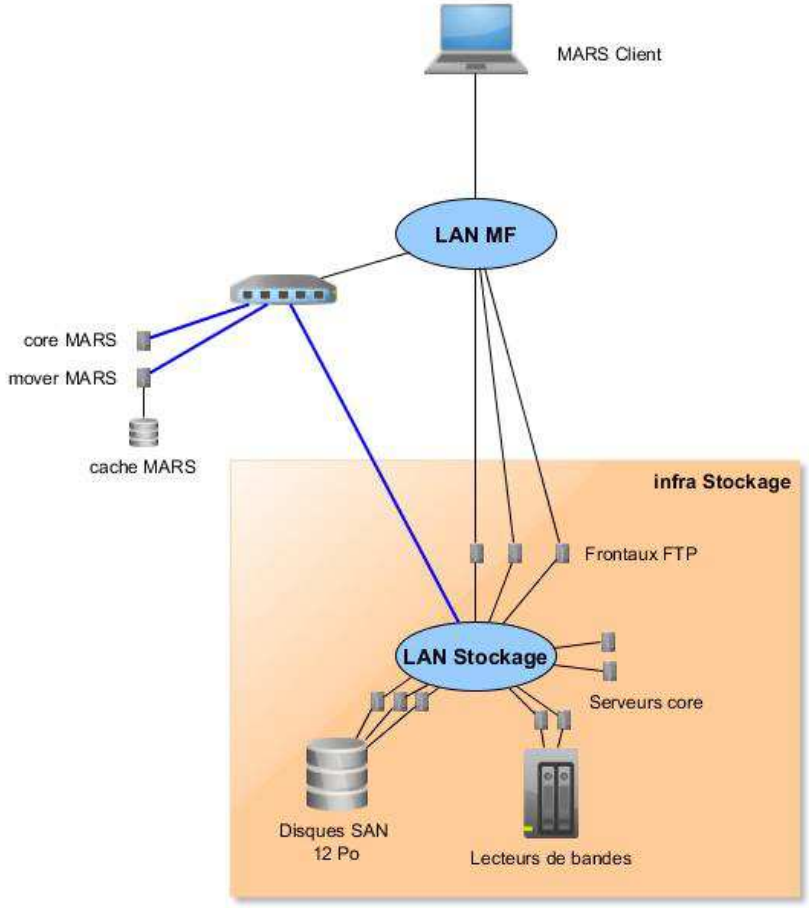
Future evolution (2018-2019)

- Future implementation of an abstraction layer between users and HPSS : MARS
- MARS is an archive software developped and used by ECMWF (European Centre for Medium-Range Weather Forecasts, in Reading – UK)
- Mars will optimize :
  - The datas organization,
  - The datas access,
  - The data sharing
- Mars reduce data flow between HPSS and users by reducing the I/O size to the datas targeted rather than the whole file.

  -> **will reduce the IO size on tape**

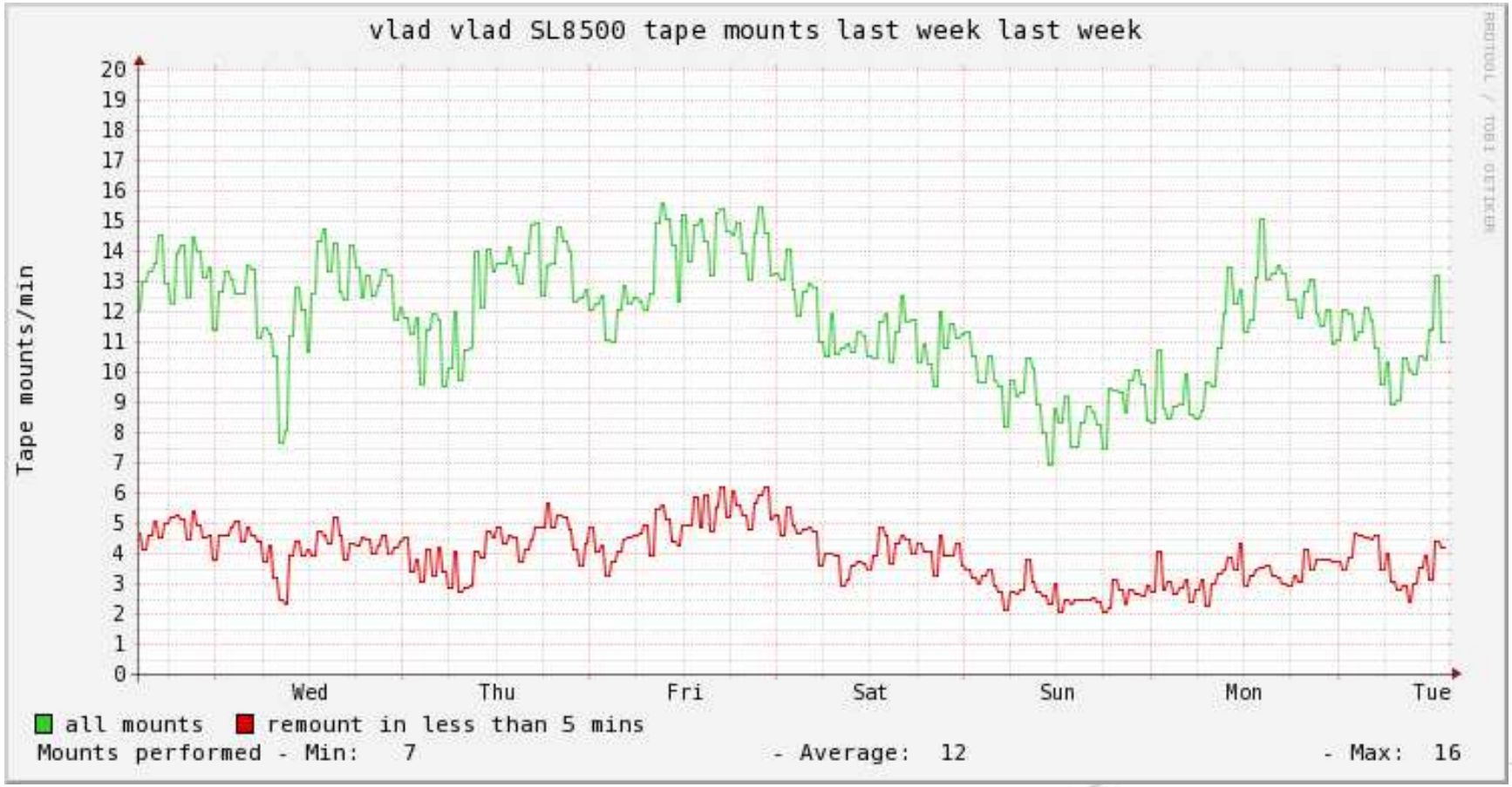
  -> **will increase the drives' sollicitations, with more position-read subcycles**

MARS Architecture



Mount Statistics at ECMWF

- 12 mounts/mn with 200 drives!!



METEO FRANCE
Toujours un temps d'avance

HPSS Perfomance Issues

- Bottleneck on HPSS queue
  - Limited number of HPSS processes // : MAX_CONNECTION
  - No priority to disk cache
    - In case of thousands gets from tapes, the queue cannot accept gets from disks
    - Need a queue management in front of HPSS -> Bull Director
- Latency on hpss connection

  -> bottleneck in case of huge number of // connections
    - Acceptance tests (PV : 800 gets of 100Mo in 10 secondes)

  -> Houston workaround: PFTP -> FUSE + ftp standard

  prohibits the use of default COS depending of the file size

  prohibits interceptor use
- Repack tape to tape
  - HPSS migrate all files from tapes, even if present on disk

HPSS Perfomance Issues

- Bottleneck on HPSS queue
  - Limited number of HPSS processes // : MAX_CONNECTION
  - No priority to disk cache
    - In case of thousands gets from tapes, the queue cannot accept gets from disks
    - Need a queue management in front of HPSS -> Bull Director
- Latency on hpss connection

  -> bottleneck in case of huge number of // connections
    - Acceptance tests (PV : 800 gets of 100Mo in 10 secondes)

  -> Houston workaround: PFTP -> FUSE + ftp standard

  prohibits the use of default COS depending of the file size

  prohibits interceptor use
- Repack tape to tape
  - HPSS migrate all files from tapes, even if present on disk

IBM TS1150 Drive Reliability Issues

- High replacement rate of drives
  - All the drives installed in phase 1 (2015/01) were replaced (51)

- Because of high level of “position-read” cycles and premature wear

  -> IBM action : retrofit of TS1155 (TMR) head on TS1150 (GMR)

  -> No more issue with new TS1150 TMR

Silent corruptions

- A drive made silent corruption during writes (:
  - 73 tapes to repack
    - 16 tapes with partial repack
    - 130 K files not readable
  - 25 tapes sent to Tucson for recover
  - Until now :
    - 80 K files definitively lost
- IBM provided a fix for drive firmware
  - 1 more tape sent to Tucson

Items for discussion

- Monitoring
  - Cache hit ratio (HPSS?BDFH?)
  - Read performance (how to ignore « mount redundancy time » in statistics ?)
- Roadmaps librairies and drive
  - What future for Oracle
  - LTO vs Jaguar
- Technological migration
  - Issues with hierarchy of 3 levels (One disk, one TS1150, one TS155)
    - Migrate from level2 to level3 (TS1150 to TS1155) doesn't process agregats
- HPSS Migration
  - Feedback about Qrep ?

Questions