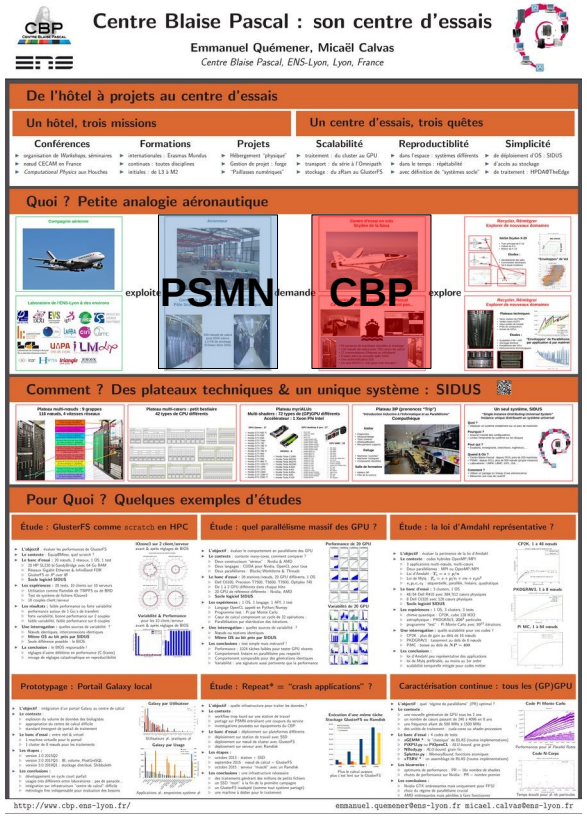
CBPsmn
Centre Blaise Pascal de Simulation et de Modélisation Numérique

JoSy ZFS 2023

Et mon pool, il va tomber en panne ? De la santé des disques durs À la résilience de ZFS

Emmanuel Quémener

CBPsmn avec son centre d'essai : et de production...

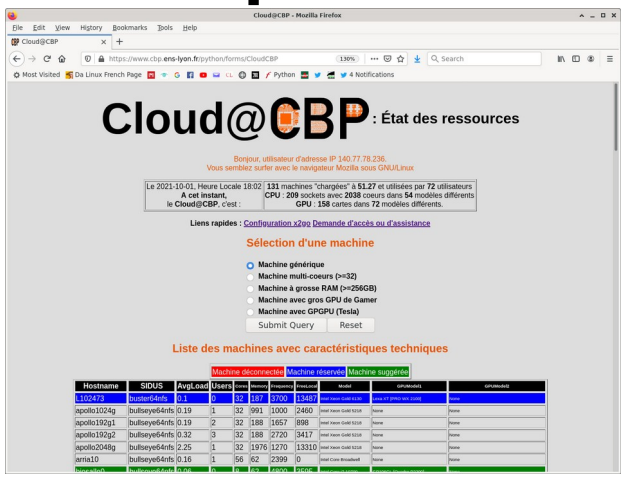


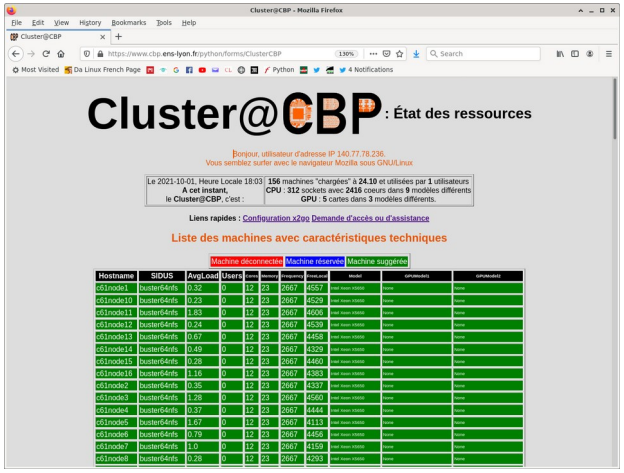
- Nasa X-29
- Cellule de F-5
- Moteur de F-18
- Train de F-16
- Etudes
  - Plans « canard »
  - Incidence >50°
  - « Fly-By-Wire »

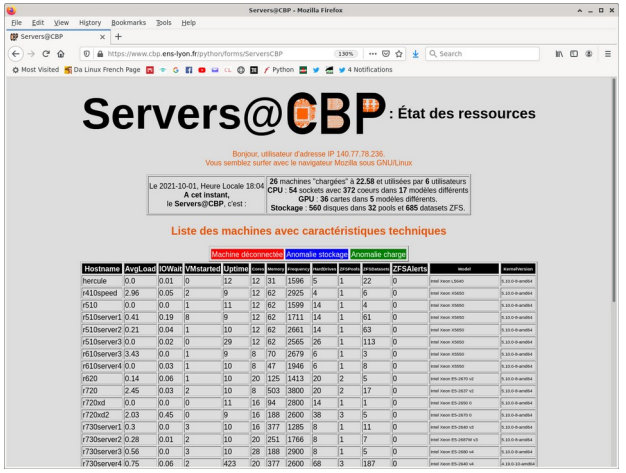
Recycler, réutiliser, explorer de nouveaux domaines…

… Et pour faire cela, il faut du matériel !

Le Centre Blaise Pascal : c'est ... près de 300 machines actives





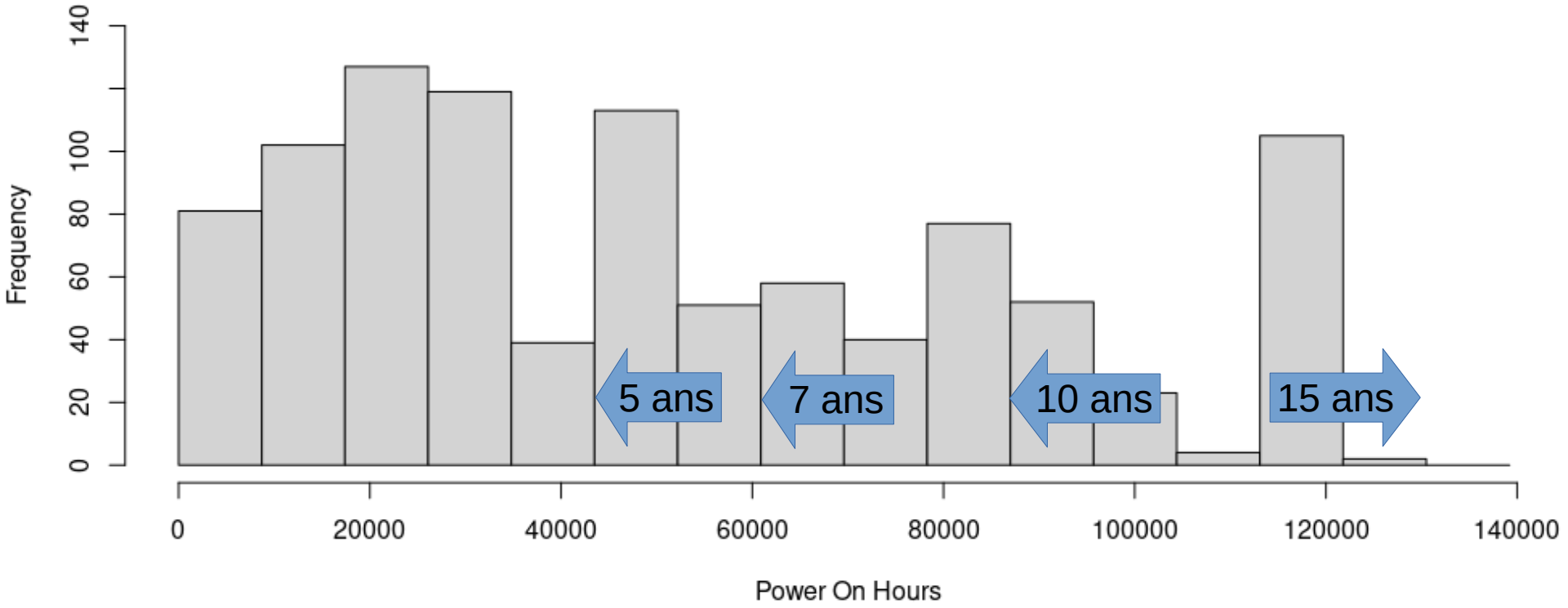


- Les 3 infrastructures Cloud@CBP, Cluster@CBP et Servers@CBP
  - 508 « sockets » cumulant 6218 « vrais » coeurs dans près de 60 modèles
  - Près de 50 TB de RAM ou équivalent (avec la DCPMM, on « triche » un peu)
  - 216 (GP)GPU dans près de 80 modèles (tout ou presque depuis 2008)
  - **1013 disques durs cumulant plus de 50M heures**
  - Un taux de disponibilité supérieur à 99.5 %
  - **Tout ça avec 95 % d'équipements hors garantie**...

Le disque dur : ce pollueur insoupçonné (mais il y a bien pire)

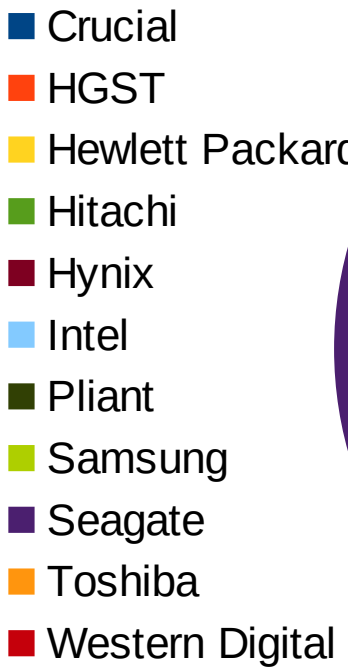
- Un disque dur :
  - À la production : autour de 50 kg de CO_2
  - À l'exploitation (en France) : 10W, 6 kg de CO_2/an
  - Equilibre carbone : 9 ans d'exploitation...
- Déjà au Centre Blaise Pascal, depuis 9 ans :
  - Sauvegarde primaire sur r620 + baies MD1200 de 2TB
  - Sauvegarde secondaire sur x4500 (disques changés de 1TB en 2012)
- Récupération de près de 200 disques sur 2021, 100 en 2022
- Comment estimer leur fiabilité ?
- Comment les réexploiter ?

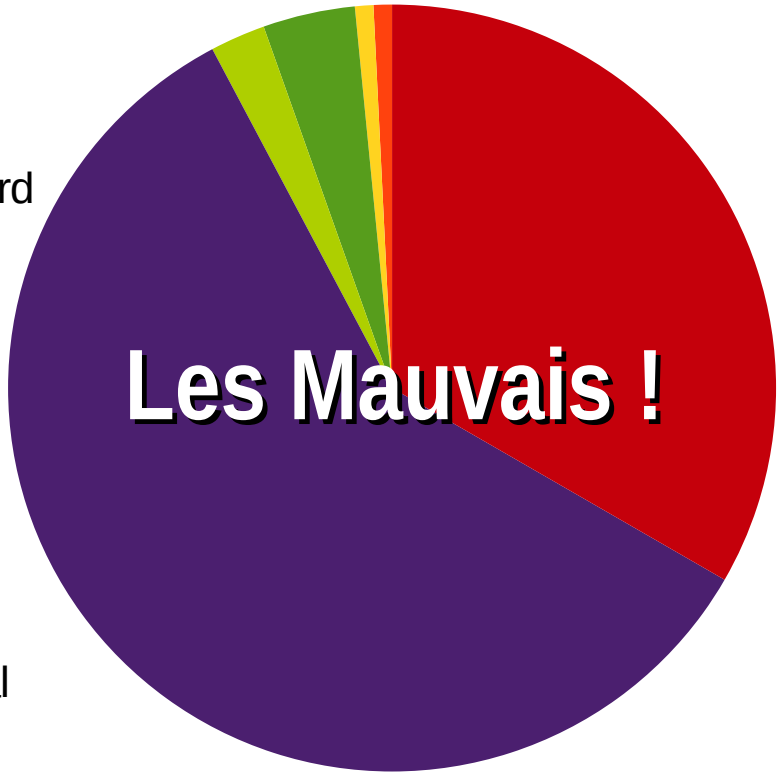
Histogramme des disques actifs : la *Power On Hours*...



- Les disques plus « résistants » que la garantie...
- Mais qu'en est-il des disques déclarés « inaptes »?

HDD : question récurrente… Quelle est la meilleure marque ?

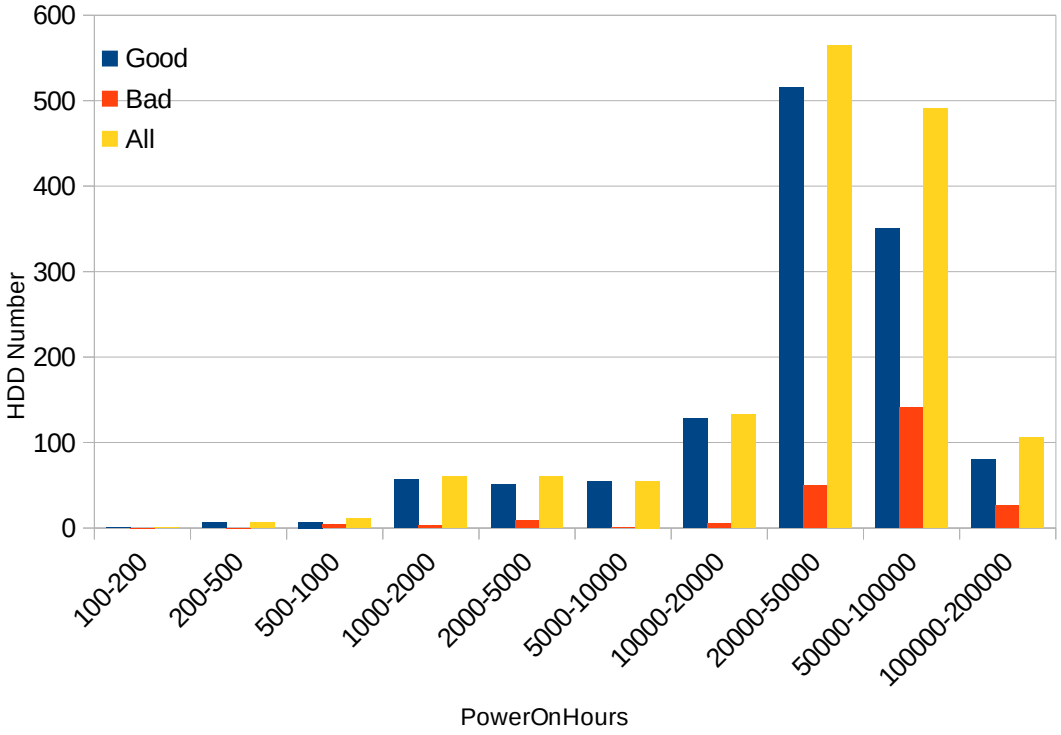


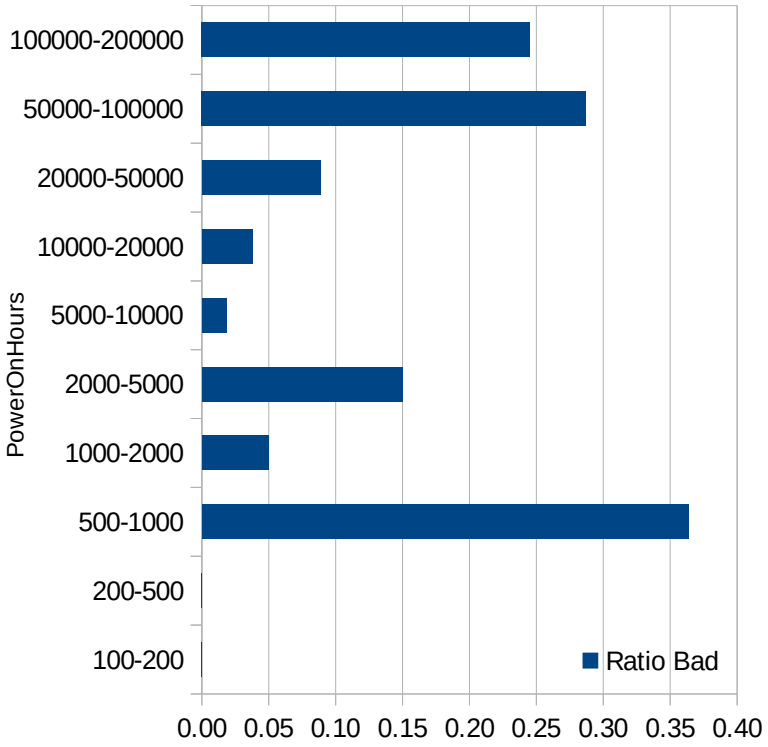


Echantillon de +1500 HDD, extraction des informations SMART

Impossible de statuer...

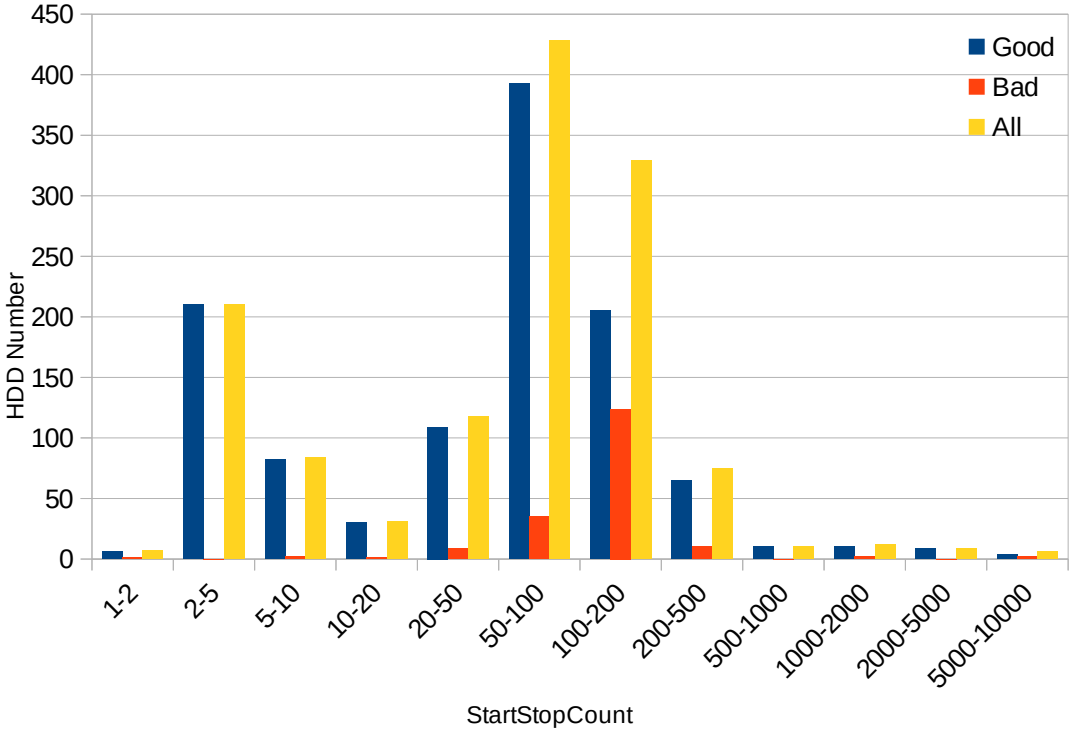
Le retour de *Power On Hours* : Un résultat assez contre intuitif...

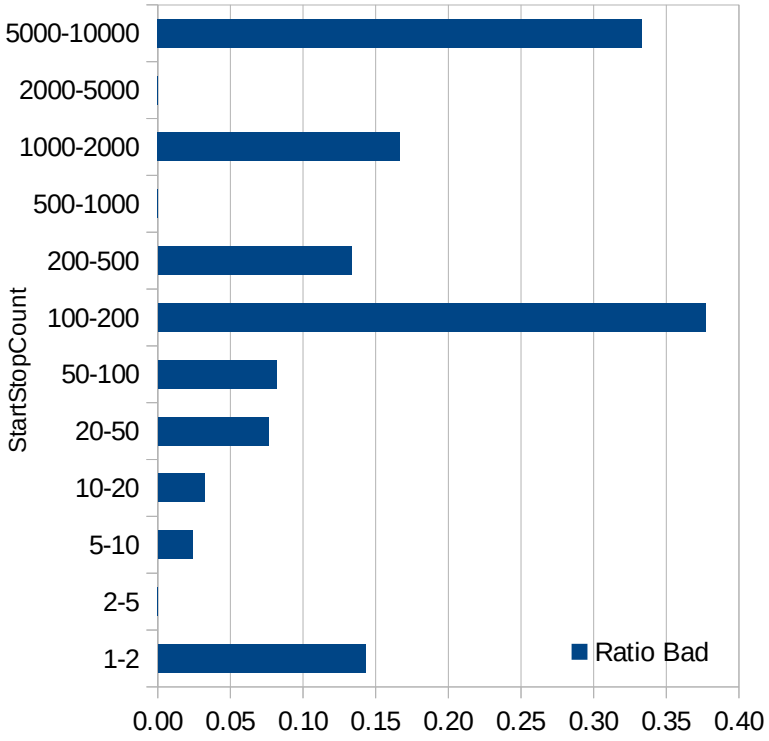




- Un rappel de « la courbe en sourire » des pannes
- Les disques durs plutôt résistants : < 6 ans (<10%)

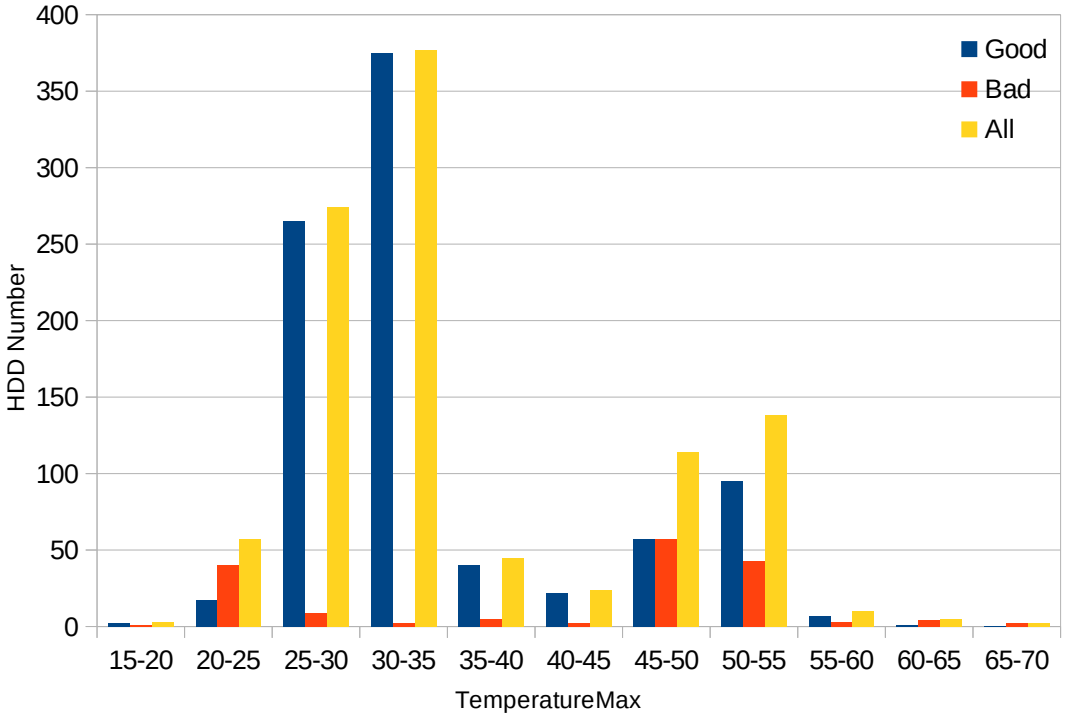
HDD : Les arrêts-redémarrages : le calvaire des disques durs ?

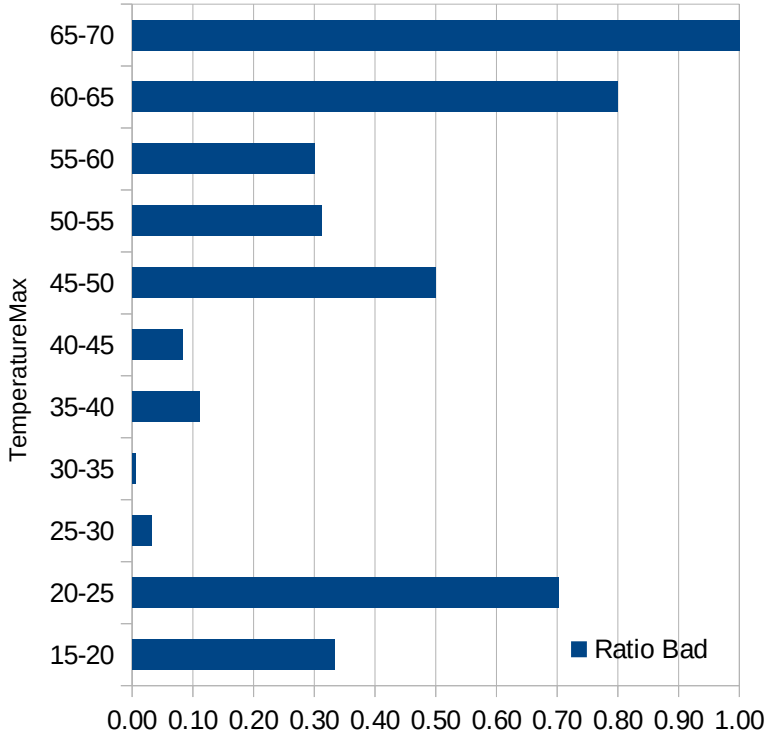




- Sur-représentation des disques HS pour 100-200 A/R
- Économie d'énergie par arrêt des disques : pertinente ?

HDD : la température Maximale Un vrai critère de fiabilisation ?





- Au dessus de 40°C de T° Max, presque 40 % sont HS
- Avoir une séparation entre HDD et CPU pertinente ?

Réexploiter les HDD ?
Oui, mais pas n'importe comment...

- Tout d'abord, s'assurer de leur « santé »
  - TemperatureMax & StartStopCount plus pertinent que PowerOnHours
- Puis, prendre soin de leur fonctionnement :
  - Toujours préserver leur exploitation à une température « raisonnable »
- Ensuite, dans les grappes, bien bétonner le « RAID »
  - Exclure le RAID matériel pour faciliter les changements (→ ZFS :-))
- Enfin, éloigner physiquement CPU et HDD :
  - Normalement, HDD devant CPU, mais seulement dans les serveurs
- Ou alors, exploiter les HDD en relocalisant le stockage

De la pérennité de ZFS un « sauvetage » inespéré...

- Le contexte :
  - Un x4500 sous Solaris datant de 2009,
  - La machine « plantée », OS complètement HS !
  - Mais toutes les données d'un site Web, évidemment non sauvegardé…
  - Pas de CD d'installation ou de clé pour réparer dans l'OS originel
- Le sauvetage en 2020 :
  - Démarrage de la machine sous SIDUS avec Debian Buster (de l'époque)
  - Importation des pools ZFS par : zpool import -a
  - Réalisation d'un snapshot des volumes ZFS par zpool snapshot
  - Duplication par les mécanismes Send/Receive…
- Essayez juste de faire cela avec des contrôleurs matériels…

De la résilience de ZFS : La « grosse panne »

- Le contexte : pour un gros serveur de backup...
  - Une R815 avec 2 MD 3060, 126 disques au total dont 120 de 4TB
    - 10 pools de 12 disques en RAIDz2 pour un « superpool »
  - Backup de stations de travail d'une équipe arrivant : plus de 60 TB
    - Sur la dernière machine, apparition de disques en erreur dans le « superpool »
- La reconstruction : cumul de 7 disques en erreur
  - Pendant la reconstruction de nouveaux HD HS
  - Changement du « socle » matériel : R815 vers R620
- Les enseignements :
  - Privilégier la reconstruction aux « gros » transferts
  - Tester les disques complètement avant exploitation
  - Disposer d'un socle matériel « efficace » pour de gros pools

En conclusion
ZFS robuste, résilient, rapide...

- « Le RAID est une chose trop sérieuse pour la confier à des contrôleurs matériels ! »
-

Appel aux dons !!!
Computhèque comme sanctuaire

- Qui pourrait me fournir les composants suivants :
  - Carte contrôleur IDE sur port ISA 16 bits, disquettes 5.24 pouces
- Autrement, la computhèque du CBP accueille :
  - Tout équipement informatique le plus ancien possible :
    - Les vieux 8 bits des années 1980 : Sinclair ZX, Commodore, Oric, etc...
    - Les vieux PC avec des cartes ISA : 80286, 80386, ...
    - Les vieux périphériques : SCSI, scanner, disques durs, lecteurs de bande, etc...
  - Tout équipement informatique un peu exotique :
    - Machines de technologie : Dec Alpha 21264, HPPA, Sun...
- Merci pour votre générosité : james.mylq@ens-lyon.fr