Conférence JCAD2021

~~“Digital Staging”~~

“ Digital re* & de* ”

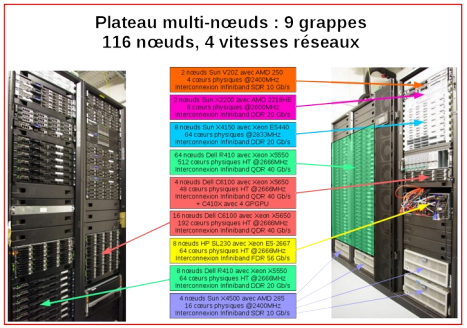
(prononcez “Digital restar’n’destar”)

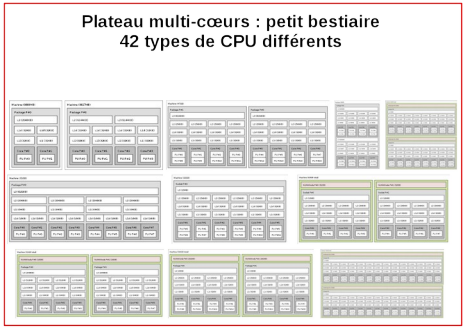
… ou comment offrir de nouveaux services (ou approches) avec du vieux matériel ?

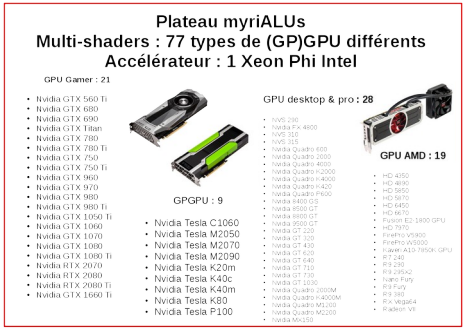
Emmanuel Quémener

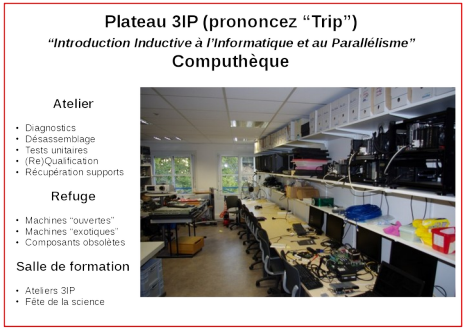
Centre Blaise Pascal et son centre d'essais...

- Centre Blaise Pascal : 3 hébergements
  - Hôtels à conférences
  - Hôtel à formations
  - Hôtel à projets
- Centre d'essais : 3 quêtes
  - Reproductibilité
  - Scalabilité
  - Simplicité
- Ses propres plateaux techniques

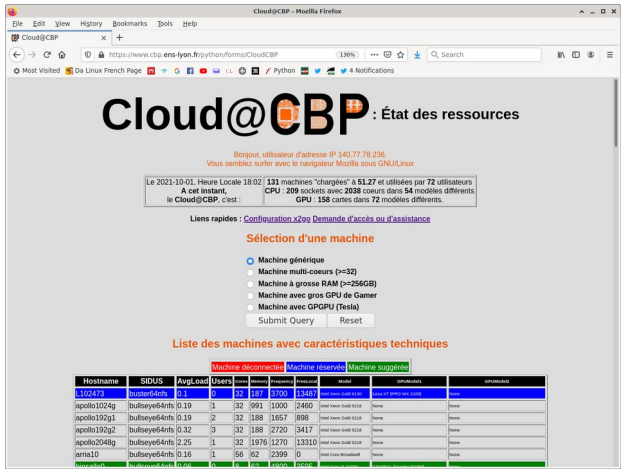


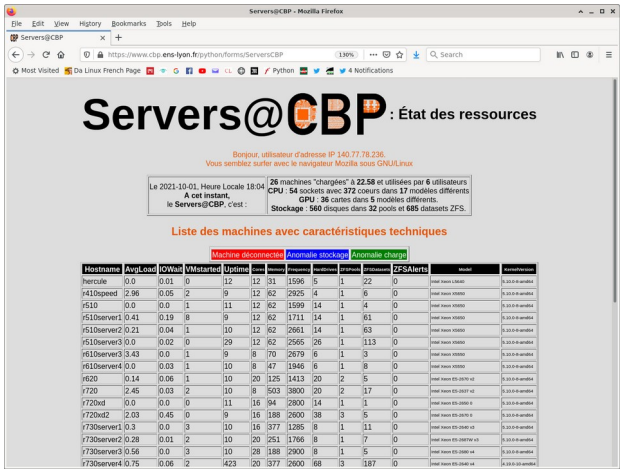


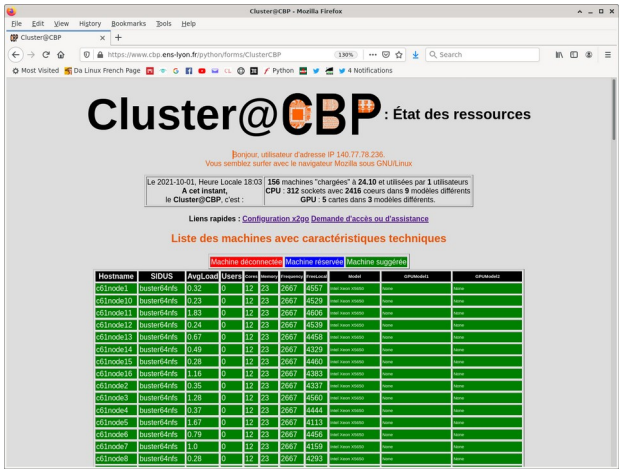




Le Centre Blaise Pascal : c'est aussi ... plus de 300 machines actives







- +4600 (vrais) coeurs, 28 TiB RAM, +1200 HDD, ~4PiB
- 90 % de machines « hors garantie constructeur » !
- Récupération : PSMN, DSI, labos internes & externes...

… Et une centaine d'autres « prêtes » à repartir...

« Home Staging » personnel, Mais pas pour vendre...

Requalifier, Réactualiser, Recycler.

« Digital staging » vs « Home Staging »

- « Home staging » pour Google :
  - *« the activity or practice of styling and furnishing a property for sale in such a way as to enhance its attractiveness to potential buyers. »*
- Mais sur Wikipédia, à « staging », pour les computeurs :
  - *Staging (cloud computing), a process used to assemble, test, and review a new solution before it is moved into production and the existing solution is decommissioned*
  - *Staging (data), intermediately storing data between the sources of information and a data warehouse*
  - *Disk staging, using disks as an additional, temporary stage of backup process before finally storing backup*
  - *Staging site, a website used to assemble, test, and review its newer versions before it is moved into production*
- Finalement, c'est plutôt des actions préfixées de « ré » ou « dé »...

« de-* » & « re-* »
Entre idéologie & pragmatisme

- Déqualifier
- Détoxifier
- Déspécialiser
- Décortiquer
- Détourner

- Requalifier
- Redéployer
- Réaménager
- Réassigner
- Réfugier

Une recherche du ratio CO_2 fabrication/usage faible

Une volonté d'exploiter le + et le + longtemps possible...

Une approche « écologique » à cycle court !

Déjà, il y a 20 ans, à Lyon les JRES 2001

De l'intérêt d'utiliser la même plate-forme système sur un campus en général au radical déploiement de GNU/Linux sur l'ENS-Cachan en particulier

Emmanuel Quémener, Pascal Soullard, Pascal Varoqui
<quemener@cri.ens-cachan.fr>,<soullard@cri.ens-cachan.fr>, <varoqui@cri.ens-cachan.fr>

En ce qui concerne la base matérielle, nous avions remarqué qu'un PC équipé d'un 486DX était suffisant pour router 10 Mb/s. Ayant pris le parti d'équiper nos routeurs génériques d'interfaces FastEthernet et de placer plus de 2 interfaces dans un PC, le choix d'un Pentium de fréquence supérieure à 133 MHz semblait opportun. De plus, les cartes devaient être interchangeables à volonté sans configuration dans le BIOS de chacune d'elles : l'interface PCI offrait cette fonctionnalité. Nous avons ainsi fixé notre choix sur une carte 3Com 3c905. Les cartes Pentium ne disposant généralement que de quatre ports PCI, nous nous sommes décidés sur des routeurs Pentium à 4 ports PCI. La chasse aux cartes vidéo ISA fut lancée. De petits disques durs de capacité inférieure à 1 Go feraient largement l'affaire, même devant la nécessité de recompiler un noyau. Ceux-ci seraient situés dans un boîtier amovible permettant l'échange d'un routeur par un autre par le simple échange de disques durs.

Ainsi, le routeur générique était un PC de bureau déclassé, équipé généralement d'une carte mère ASUS T2P4 ou TXP4, d'un processeur Pentium cadencé autour de 200 MHz, de 32 à 64 Mo de mémoire vive, d'un disque dur da[illegible] 1 Go, d'un boîtier [illegible] afin de [illegible] dans une baie 19", d'une carte vidéo ISA et enfin de 4 cartes réseaux 3Com 3c905. L'investissement se limitait donc, la machine étant déclassée pour une utilisation bureautique, à l'acquisition des quatre cartes réseaux. Les installations logicielles des machines se réalisèrent également progressivement, au gré des migrations. Elles se basaient sur une distribution Debian 2.2 et des versions de noyau 2.2.x avant que le noyau 2.4 ne se stabilise.

Détournement d'un vieux poste de travail
En un firewall à routage dynamique

Sur une machine, la qualification : Fiabiliser, consolider, pérenniser...

Derrière le modèle de Von Neumann :

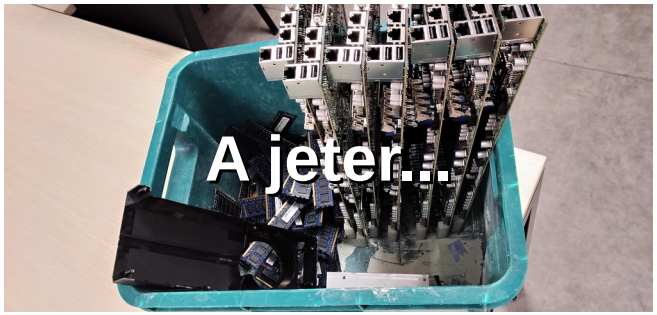
- Le **processeur** : des générations plus récentes : Nehalem vers Westmere
- La **mémoire** : remplir les bancs (et améliorer les traitements)
- Les **périphériques de stockage** : les disques durs, les baies, les cartes
  - Cartes SAS HBA supportant les derniers disques (mais pas les RAID...)
  - Baies SAS de 2009 supportant des disques de 16 TB !
  - Disques durs encore exploitables
- Les **périphériques de communications** :
  - De l'infiniband pour du « SAN-like » : une manière de séparer les flux
  - Des cartes Infiniband pour du 10G : « juste » un adaptateur à acheter...

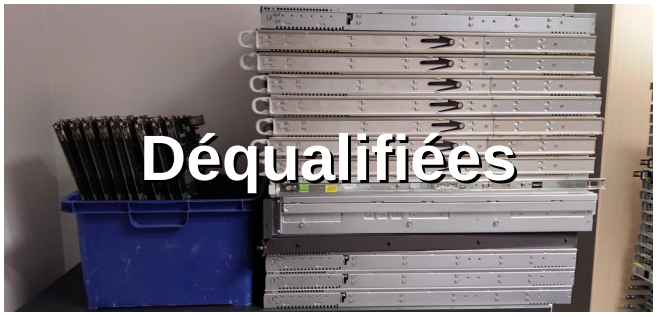
Les **3A** : test, assemblage, exploitation

Activités 2020-2021 : passage dans une autre dimension, 168 machines

Les machines à qualifier :

- **Supermicro : 125**
  - Bull (SM) R422 : 116
  - SGI (SM) : 9
- **Dell : 28**
  - R815 : 11
  - R510 : 4
  - R720XD : 1
  - C6100 : 8
  - C8000 : 4
- **Sun : 5**
  - V20z : 2
  - X2200 : 1
  - X4150 : 3
- **HP** DL 175 : **10**

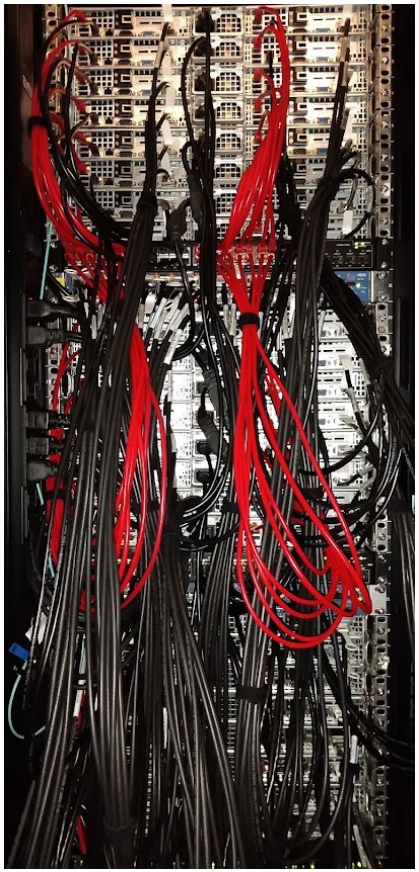




Au final, **134** machines opérationnelles, en l'instant...

Requalifier & Redéployer Cluster « classique »

- Supermicro R422 : 112 nœuds
  - Seuls 98 rescapés, 10 *spares* à ce jour
- Requalification :
  - **InfiniBand** : ibv_srq_ping_pong
  - **IPMI** : module & réseau
  - **Mémoire & CPU** : mbw parallélisé par xargs
  - **Alimentation**
- Réexploitation :
  - 64 nœuds dans 42 U, 3x32A
- Souci : version FW QDR

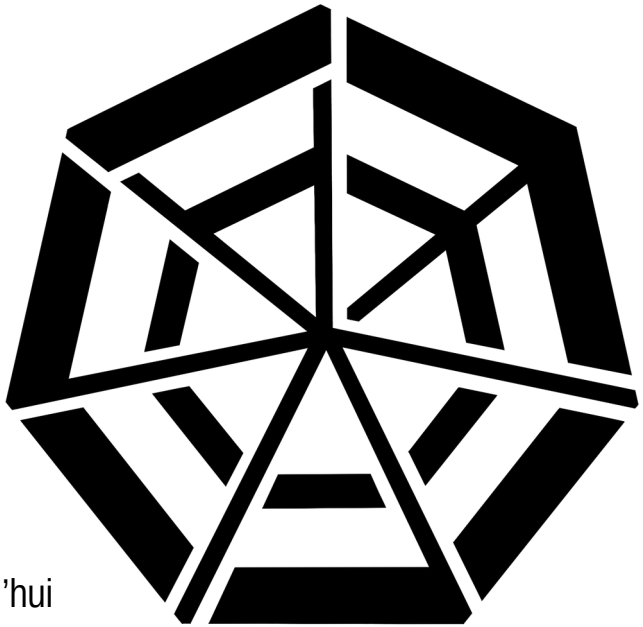


Exigences : *spares parts*, toutes versions FW

Sur les Machines du CBP : SIDUS

Je n'installe pas, je démarre !

- **Quoi ?**
  - Déployer un système simplement sur un parc de machines
- **Pourquoi ?**
  - Assurer l'unicité des configurations
  - Limiter l'empreinte du système sur les disques
- **Pour qui ?**
  - Étudiants (vous quoi!), enseignants, chercheurs, ingénieurs, ...
- **Quand & Où ?**
  - Centre Blaise Pascal : depuis 2010, plus de 280 machines aujourd'hui
  - PSMN : depuis 2011, plus de ~800 nœuds (sa propre instance) aujourd'hui
- **Comment ?**
  - Utiliser un partage en réseau d'une arborescence
  - Détourner le mécanisme de LiveCD



« Deux machines ayant démarré SIDUS ne peuvent pas ne pas avoir le même système ! »

Requalifier & Redéployer Cluster ailleurs : ISA à Lyon

- Cluster SIDUS depuis 2016
  - 8 vieux nœuds SGI Supermicro
- Cluster début 2021 :
  - 8 nœuds Sun X4150
  - 8 nœuds HP DL175
  - 4 nœuds Bull Supermicro R422
- Cluster fin 2021 :
  - 8 nœuds DL175
  - 10 nœuds Dell R410

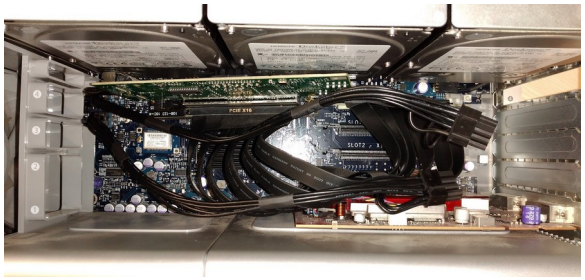
Conclusion : contre-productif de déployer des antiquités...

Déqualifier & Détourner
Ateliers 3IP & Fête de la Science

- Postes de travail Optiplex
  - Salles de formation, ...
- Dépeçage complet :
  - Il ne reste que la carcasse...
- Réassemblage pour atelier
  - Découverte de l'ordinateur
  - Machines exploratoires...

Une exigence : des espaces de stockage pour pièces !

Requalifier & Détourner : Macpro X *Machine Learning* Machines

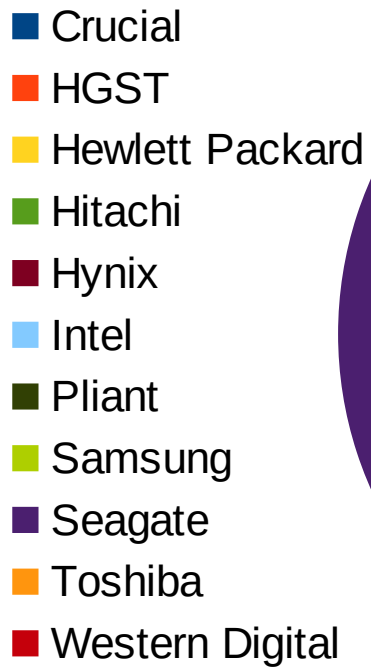


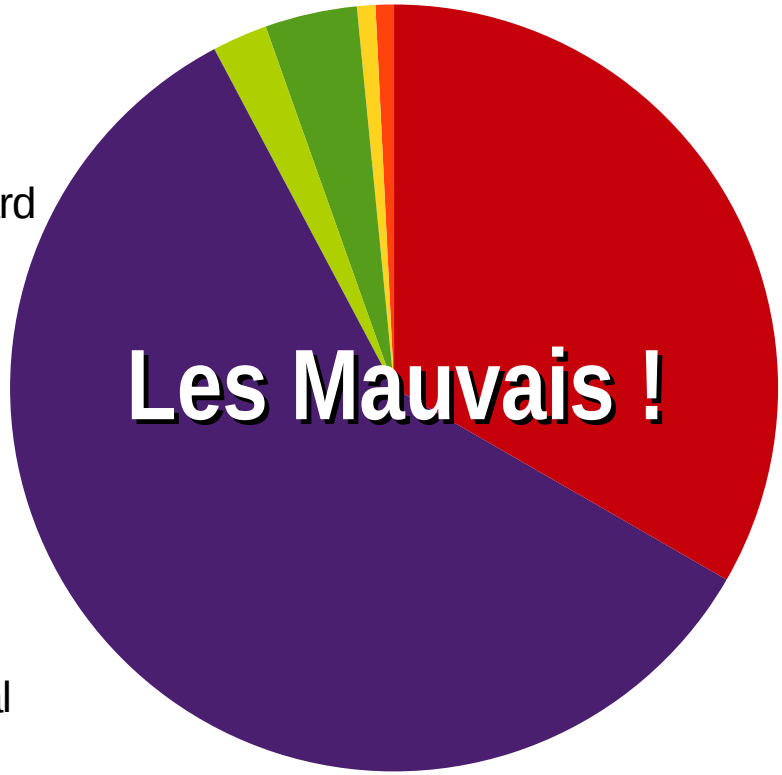
- Machines de 20 kg (dont plus de 10 kg d'aluminium)
  - Bisockets, 4 disques durs, gros GPU possible, alimentation suffisante
  - Mais : carte SAS non supportée (3.1), boot PXE impossible (1.1, 3.1), vidéo
  - Solution : carte SAS, nappe déport PCIe, carte réseau PCIe, nouveau GPU, câble

Un traitement particulier : le HDD un « pollueur insoupçonné »

- 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 principale sur x4500 (disques changés de 1TB en 2012)
- Récupération de près de 200 disques sur 2021
  - Comment estimer leur fiabilité ?
  - Comment les réexploiter ?

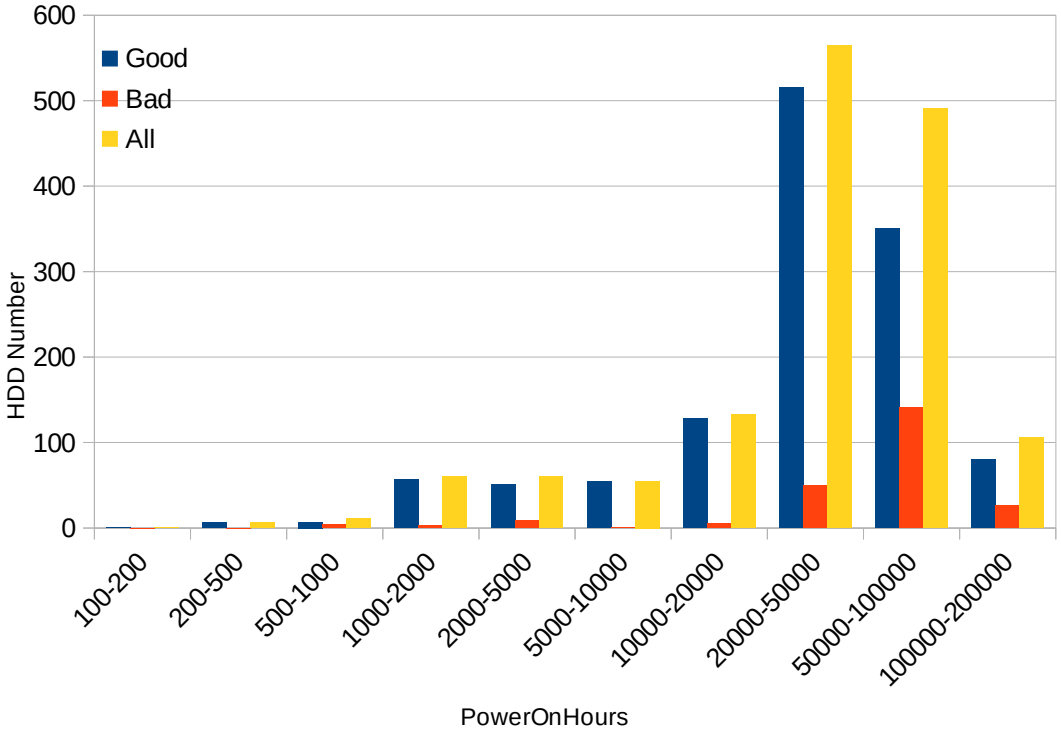
HDD : question récurrente…
C’est quoi la meilleure marque ?

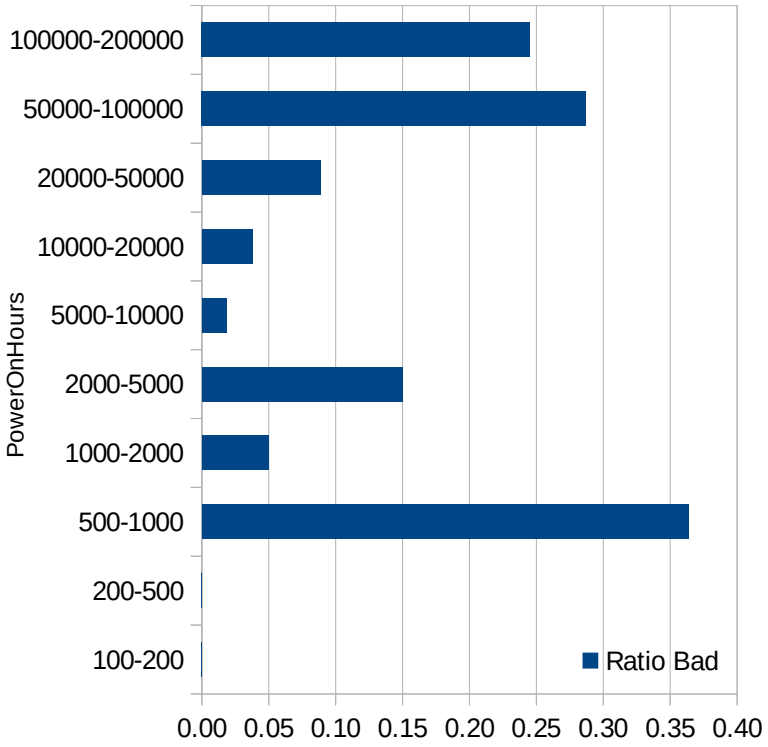




Echantillon de +1500 HDD, extraction des informations SMART
Impossible de statuer...

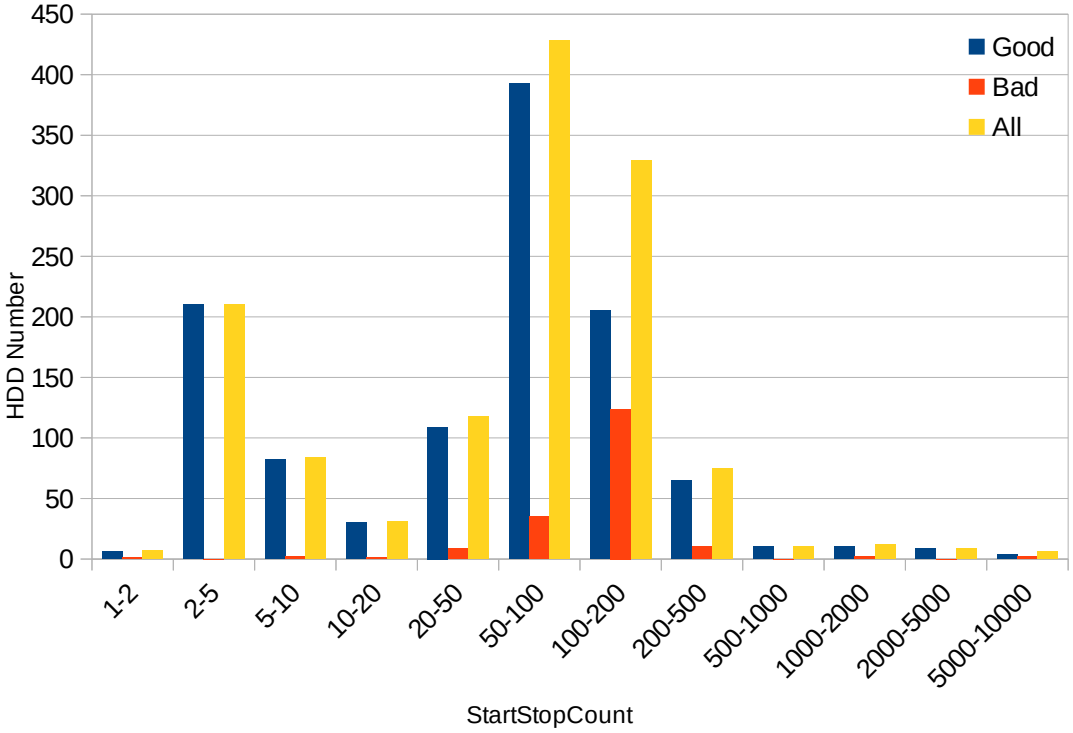
HDD : Heures de fonctionnement
Un résultat contre intuitif...

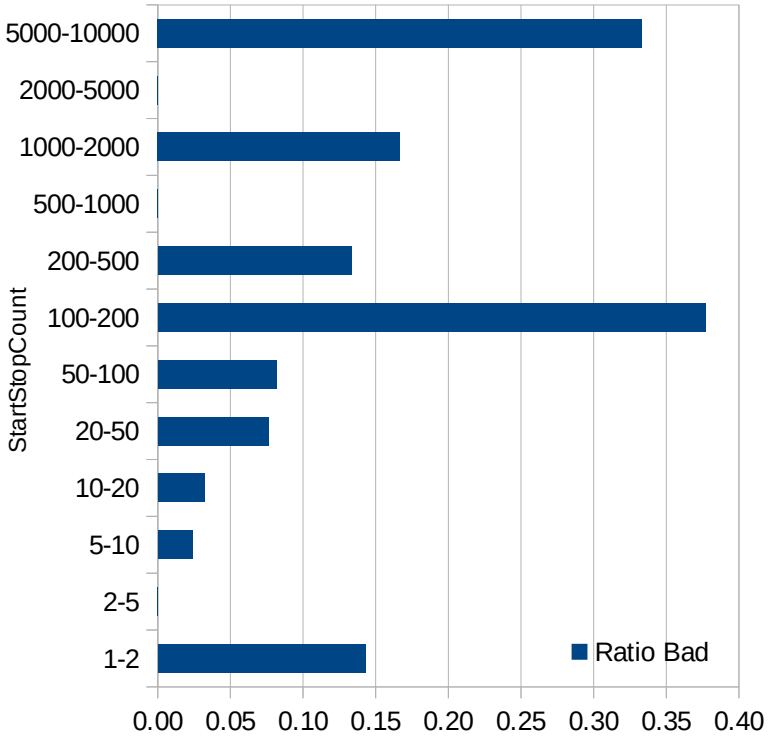




- Un rappel de « la courbe en sourire » des pannes
- Les disques durs plutôt résistants : < 6 ans (<10%)

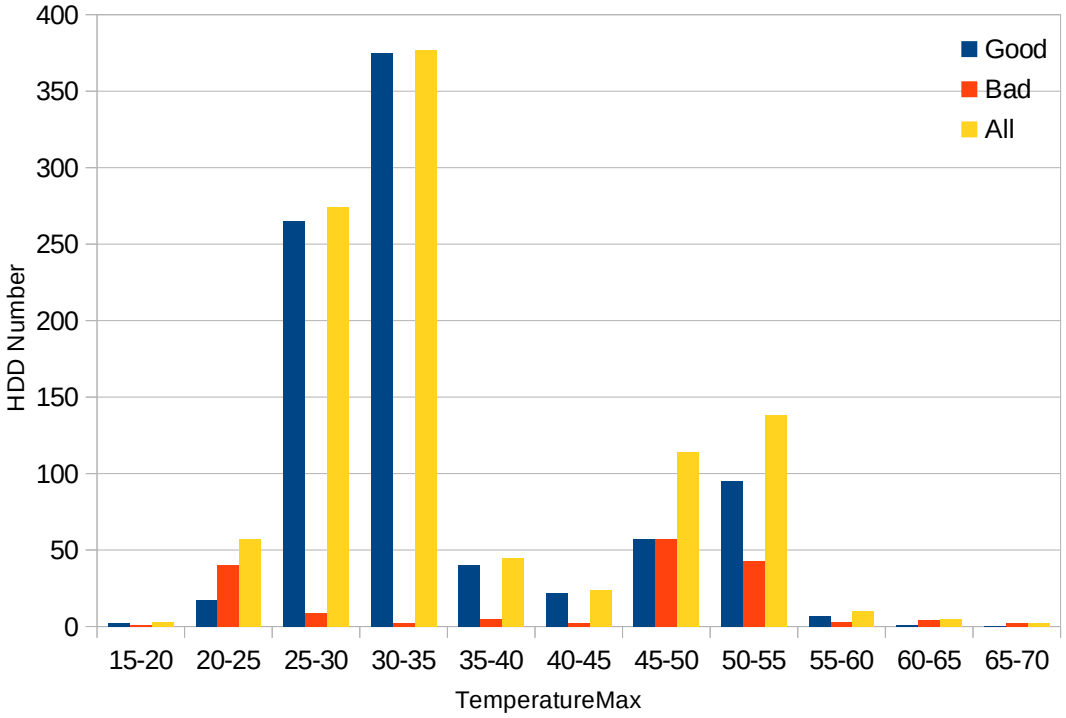
HDD : Les arrêts-redémarrages : le calvaire des disques durs ?

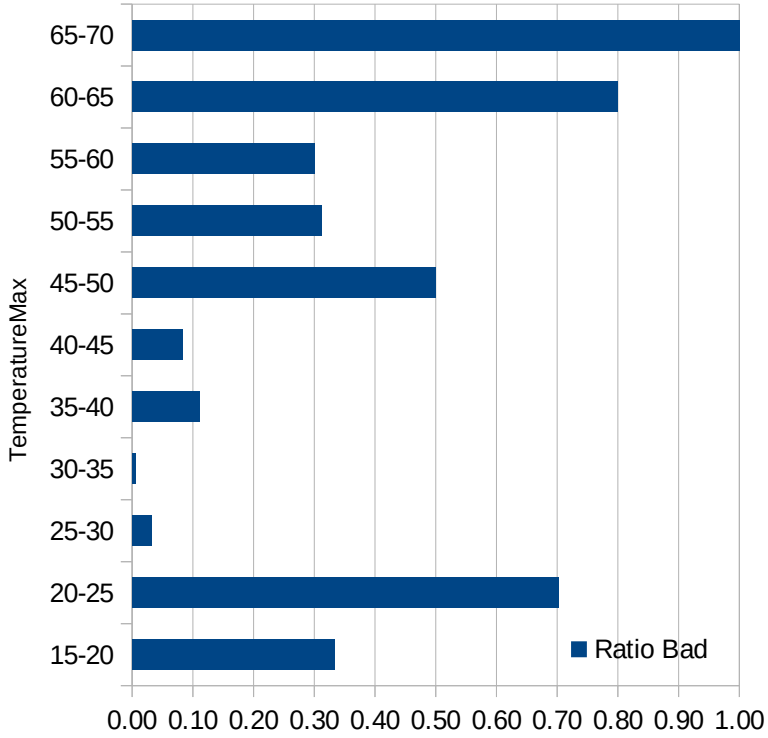




- Sur-représentation des disques HS pour 100-200 A/R
- Economie d'énergie par arrêt des disques : pertinente ?

HDD : la température Maximale Un vrai critère de fiabilisation ?



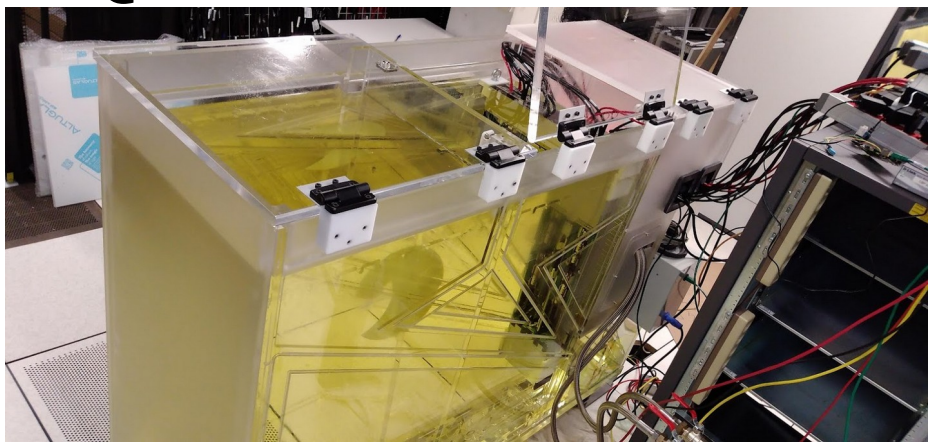


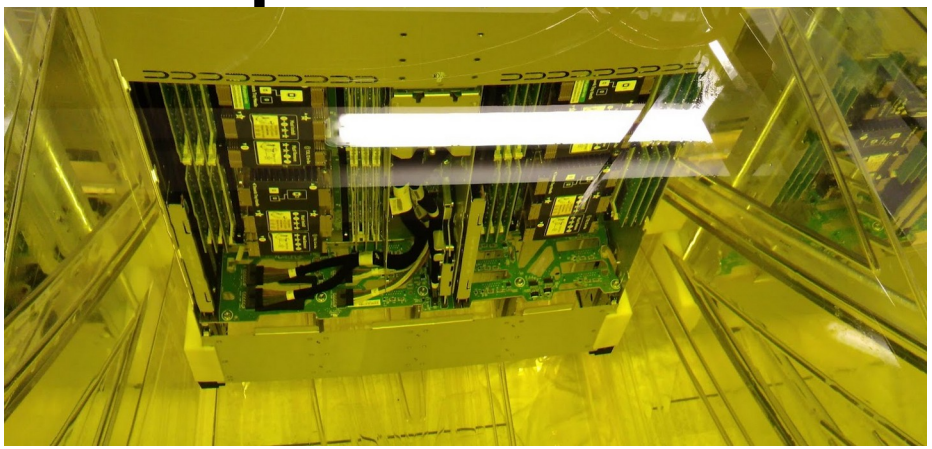
- Au dessus de 40°C de T° Max, presque 40 % sont HS
- Avoir une séparation entre HDD et CPU pertinente ?

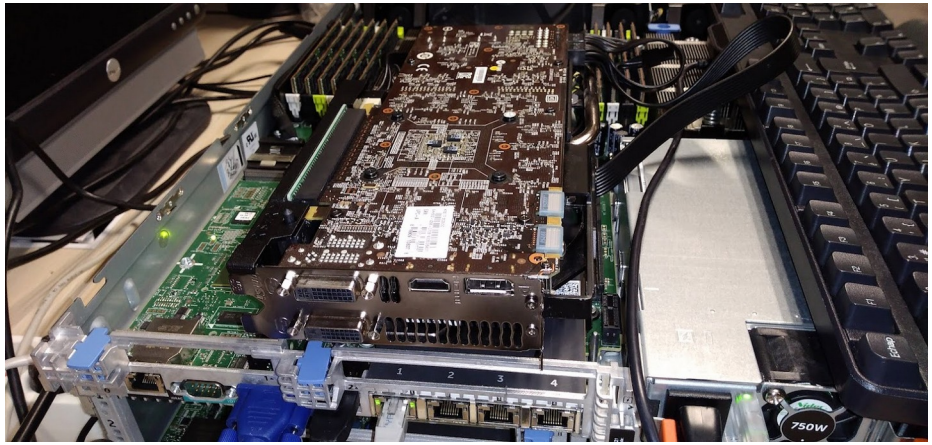
Réexploiter les HDD ?
Oui, mais pas n'importe comment...

- 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

« Immersion Cooling » : réexploiter Quid : réversibilité du processus ?



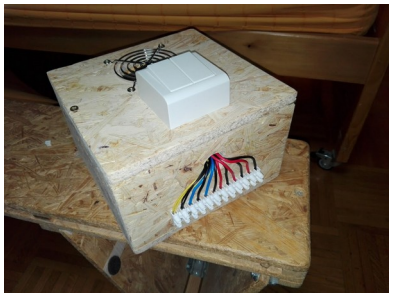


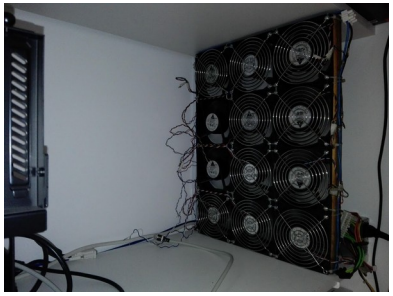


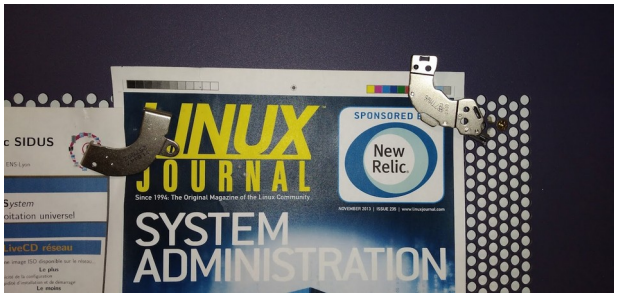
Évaluer l'immersion, mais sans « casser » de matériel...

« *Rien ne se perd, rien ne se crée. Tout se transforme !* » Lavoisier

- Au delà des composants IT :
  - Des composants électriques : une vraie mine !
    - Alimentations : du 12V, du 5V, du 3.3V
    - Des ventilateurs : du 4 cm au 12 cm
  - Des boîtiers : capot pour support de carte mère
  - Des portes de baie, des aimants de HDD
  - Des palettes de transport, des planches de protection







En conclusion

- La réduction de l'empreinte CO_2 :
  - Finalement, en exploitation, empreinte « presque » marginale...
  - Le ratio : CO_2 Production / CO_2 Exploitation si possible < 1
- Matériel informatique : matière (presque) première
  - Atouts : composants génériques, adaptation aux besoins
  - Objectifs : réassignement de missions, utilisation le + longtemps possible
  - Exigences : une gouvernance attentive procurant...
    - Du stockage pour entreposer les équipements
    - Du budget (et des procédures simplifiées) pour des pièces détachées
    - Du temps et des actions valorisées
    - De la volonté...