



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

**Plateau multi-nœuds : 9 grappes
116 nœuds, 4 vitesses réseaux**

- 1 serveur Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 1 serveur Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM
- 16 serveurs Dell R710 avec 2x AMD Opteron 6376 (16 cœurs physiques @ 2.3GHz) et 16 Go de RAM

**Plateau multi-cœurs : petit bestiaire
42 types de CPU différents**

**Plateau myriALUs
Multi-shaders : 77 types de (GP)GPU différents
Accélérateur : 1 Xeon Phi Intel**

- GPU Gamer : 21**
 - Nvidia GTX 560 Ti
 - Nvidia GTX 680
 - Nvidia GTX 690
 - Nvidia GTX Titan
 - Nvidia GTX 780
 - Nvidia GTX 780 Ti
 - Nvidia GTX 750
 - Nvidia GTX 750 Ti
 - Nvidia GTX 980
 - Nvidia GTX 970
 - Nvidia GTX 980 Ti
 - Nvidia GTX 1080 Ti
 - Nvidia GTX 1090
 - Nvidia GTX 1080
 - Nvidia GTX 1080 Ti
 - Nvidia RTX 2070
 - Nvidia RTX 2080
 - Nvidia RTX 2080 Ti
 - Nvidia GTX 1660 Ti
- GPU desktop & pro : 28**
 - NVS 290
 - Nvidia RTX 4000
 - NVS 310
 - Nvidia Quadro 600
 - Nvidia Quadro K2000
 - Nvidia Quadro K4000
 - Nvidia Quadro K4200
 - Nvidia Quadro P1000
 - Nvidia 4800 G5
 - Nvidia 6800 G5
 - Nvidia 8800 G5
 - Nvidia GT 220
 - Nvidia GT 240
 - Nvidia GT 320
 - Nvidia GT 440
 - Nvidia GT 540
 - Nvidia GT 710
 - Nvidia GT 130
 - Nvidia Quadro V1000
 - Nvidia Quadro V2000
 - Nvidia Quadro M2200
 - Nvidia Quadro M5200
- GPU AMD : 19**
 - HD 4850
 - HD 4870
 - HD 5850
 - HD 5870
 - HD 6450
 - HD 6470
 - HD 7470
 - HD 7870
 - HD 7970
 - HD 7990
 - HD 7990X
 - HD 7990X
 - HD 7990X
 - HD 7990X
 - HD 7990X
 - HD 7990X
 - HD 7990X
 - HD 7990X
- GPGPU : 9**
 - Nvidia Tesla C1060
 - Nvidia Tesla M2050
 - Nvidia Tesla M2070
 - Nvidia Tesla M2090
 - Nvidia Tesla K20m
 - Nvidia Tesla K40c
 - Nvidia Tesla K40m
 - Nvidia Tesla K80
 - Nvidia Tesla P100
- Xeon Phi Intel**

**Plateau 3IP (prononcez "Trip")
"Introduction Inductive à l'Informatique et au Parallélisme"
Computèque**

- Atelier**
 - Diagnostics
 - Désassemblage
 - Tests unitaires
 - (Re)Qualification
 - Récupération supports
- Refuge**
 - Machines "ouvertes"
 - Machines "exotiques"
 - Composants obsoletes
- Salle de formation**
 - Ateliers 3IP
 - Fête de la science

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

Cloud@CBP : État des ressources

Bonjour, utilisateur d'adresse IP 140.77.78.236.
Vous semblez surfer avec le navigateur Mozilla sous GNU/Linux

Le 2021-10-01, Heure Locale 18:02 131 machines "chargées" à 51.27 et utilisées par 72 utilisateurs
A cet instant, CPU : 288 sockets avec 3038 coeurs dans 54 modèles différents
le Cloud@CBP, c'est : GPU : 158 cartes dans 72 modèles différents.

Liens rapides : Configuration X2go Demande d'accès ou d'assistance

Sélection d'une machine

- Machine générique
- Machine multi-cœurs (n=32)
- Machine à grosse RAM (n=256GB)
- Machine avec gros GPU de Gamer
- Machine avec CPGPU (Hélio)

Submit Query Reset

Liste des machines avec caractéristiques techniques

Hostname	SIDUS	AvgLoad	Users	Machine réservée	Machine suggérée	Machine déconnectée	Machine éteinte	Machine suspendue	Machine arrêtée	Machine en panne	Machine en maintenance
112473	bulseye4rns	0.1	0	32	187	1300	1349				
apollo1524g	bulseye4rns	0.19	1	32	991	1000	2460				
apollo1501g	bulseye4rns	0.19	2	32	188	1657	989				
apollo1502g	bulseye4rns	0.32	3	32	188	1270	3417				
apollo2048g	bulseye4rns	2.25	1	32	1976	1270	13310				
arma10	bulseye4rns	0.16	1	56	62	2399	0				

Servers@CBP : État des ressources

Bonjour, utilisateur d'adresse IP 140.77.78.236.
Vous semblez surfer avec le navigateur Mozilla sous GNU/Linux

Le 2021-10-01, Heure Locale 18:04 26 machines "chargées" à 22.58 et utilisées par 6 utilisateurs
A cet instant, CPU : 54 sockets avec 372 coeurs dans 17 modèles différents
le Servers@CBP, c'est : GPU : 36 cartes dans 5 modèles différents.
Stockage : 560 disques dans 32 pools et 685 datasets ZFS.

Liste des machines avec caractéristiques techniques

Machine déconnectée	Machine éteinte	Machine suspendue	Machine arrêtée								
Hostname	AvgLoad	CPUWait	VMstat	VMstat	VMstat	VMstat	ZFSAlerts	Alerts	Alerts	Alerts	Alerts
hercule	0.0	0.01	0	12	12	31	1506	5	1	22	0
f410speed	2.96	0.05	2	9	12	62	2925	4	1	6	0
f510	0.0	0.0	1	11	12	62	1589	14	1	4	0
f510server1	0.41	0.19	8	9	12	62	1111	14	1	61	0
f510server2	0.21	0.04	1	10	12	62	2661	14	1	63	0
f510server3	0.0	0.02	0	29	12	62	2565	26	1	113	0
f610server1	3.43	0.0	1	9	8	70	2679	16	1	3	0
f610server2	0.0	0.03	1	10	8	47	1846	6	1	8	0
f620	0.14	0.06	1	10	20	125	1413	20	2	5	0
f720	2.46	0.03	2	10	8	503	3800	20	2	17	0
f720d	0.0	0.0	0	11	16	94	2900	14	1	1	0
f720d2	2.03	0.46	0	9	16	188	2900	38	3	5	0
f730server1	0.3	0.0	3	10	16	377	1265	8	1	11	0
f730server2	0.28	0.01	2	10	20	251	1766	8	1	7	0
f730server3	0.56	0.0	3	10	28	188	2900	8	1	5	0
f730server4	0.75	0.06	2	423	20	377	2600	68	5	187	0

Cluster@CBP : État des ressources

Bonjour, utilisateur d'adresse IP 140.77.78.236.
Vous semblez surfer avec le navigateur Mozilla sous GNU/Linux

Le 2021-10-01, Heure Locale 18:03 156 machines "chargées" à 24.10 et utilisées par 8 utilisateurs
A cet instant, CPU : 312 sockets avec 2416 coeurs dans 9 modèles différents
le Cluster@CBP, c'est : GPU : 5 cartes dans 3 modèles différents.

Liens rapides : Configuration X2go Demande d'accès ou d'assistance

Liste des machines avec caractéristiques techniques

Machine déconnectée	Machine éteinte	Machine suspendue	Machine arrêtée						
Hostname	SIDUS	AvgLoad	Users	Machine réservée	Machine suggérée	Machine déconnectée	Machine éteinte	Machine suspendue	Machine arrêtée
o510od01	bulseye4rns	0.32	0	12	23	2667	4507		
o510od10	bulseye4rns	0.23	0	12	23	2667	4520		
o510od11	bulseye4rns	1.83	0	12	23	2667	4600		
o510od13	bulseye4rns	0.24	0	12	23	2667	4520		
o510od13	bulseye4rns	0.67	0	12	23	2667	4468		
o510od14	bulseye4rns	0.49	0	12	23	2667	4320		
o510od15	bulseye4rns	0.26	0	12	23	2667	4460		
o510od16	bulseye4rns	1.1	0	12	23	2667	4520		
o510od16	bulseye4rns	0.35	0	12	23	2667	4337		
o510od17	bulseye4rns	1.26	0	12	23	2667	4500		
o510od18	bulseye4rns	0.37	0	12	23	2667	4444		
o510od19	bulseye4rns	1.67	0	12	23	2667	4113		
o510od19	bulseye4rns	0.79	0	12	23	2667	4500		
o510od19	bulseye4rns	1.0	0	12	23	2667	4159		
o510od19	bulseye4rns	0.26	0	12	23	2667	4293		

- +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 ordinateurs :
 - *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 CO₂ 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>](mailto:quemener@cri.ens-cachan.fr), [<soullard@cri.ens-cachan.fr>](mailto:soullard@cri.ens-cachan.fr), [<varoqui@cri.ens-cachan.fr>](mailto:varoqui@cri.ens-cachan.fr)

En ce qui concerne la base matérielle, nous avons 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 dans un boîtier amovible, d'un boîtier de 3.5" afin de pouvoir glisser dans une baie de 10" 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 :

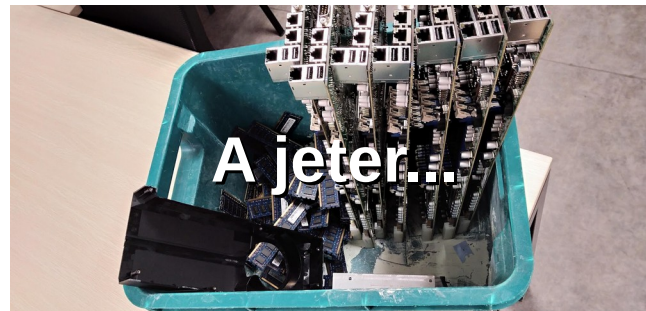
- Le **processeur** : des générations plus récentes : Nehalem vers Westmere
- La **mémoire** : remplir les bancs (et améliorer les traitements)
- Les **périphériques de stockage** : les disques durs, les baies, les cartes
 - Cartes SAS HBA supportant les derniers disques (mais pas les RAID...)
 - Baies 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 :

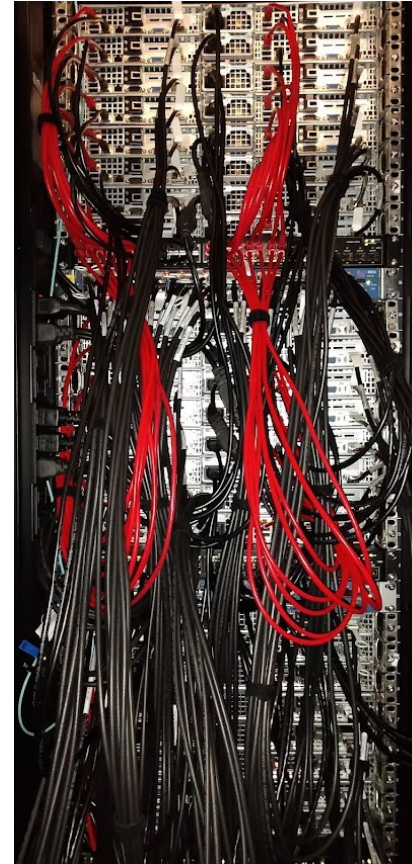
- **Supermicro : 125**
 - Bull (SM) R422 : 116
 - SGI (SM) : 9
- **Dell : 28**
 - R815 : 11
 - R510 : 4
 - R720XD : 1
 - C6100 : 8
 - C8000 : 4
- **Sun : 5**
 - V20z : 2
 - X2200 : 1
 - X4150 : 3
- **HP DL 175 : 10**



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

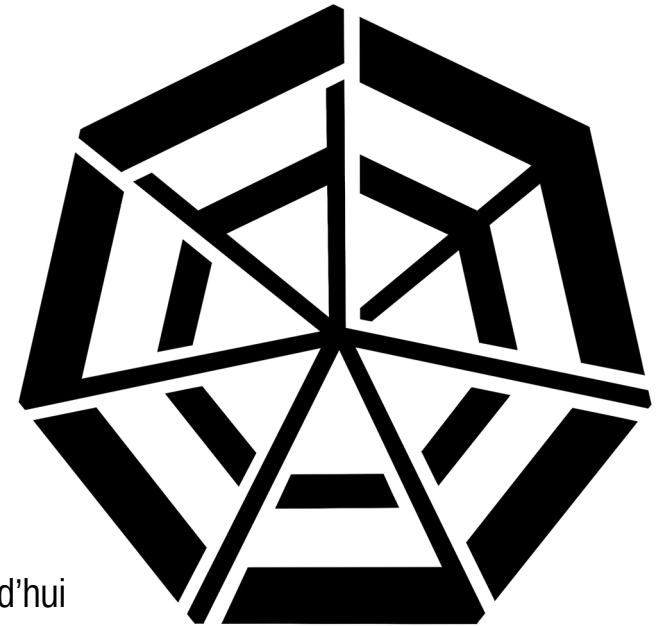


Exigences : *spares parts*, toutes versions FW

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*

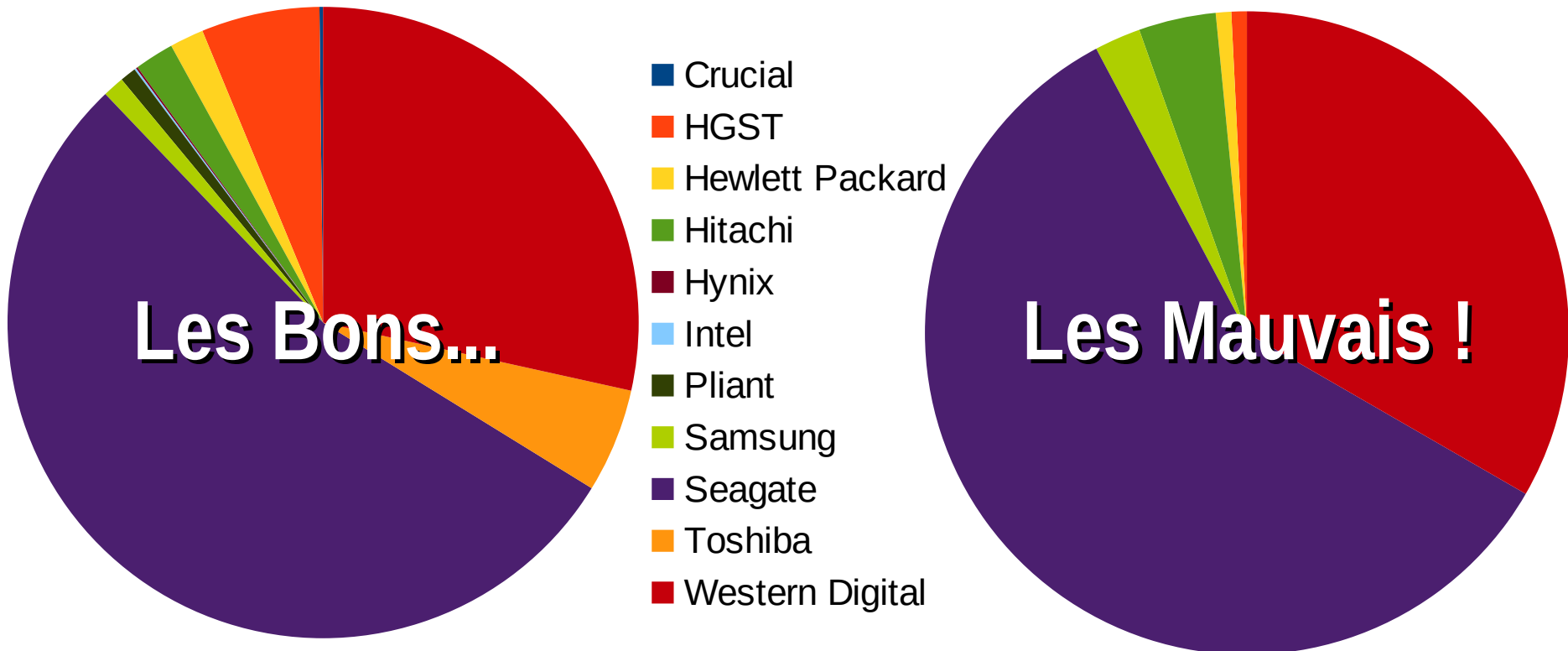


- Machines de 20 kg (dont plus de 10 kg d'aluminium)
 - Bisockets, 4 disques durs, gros GPU possible, alimentation suffisante
 - Mais : carte SAS non supportée (3.1), boot PXE impossible (1.1, 3.1), vidéo
 - Solution : carte SAS, nappe déport PCIe, carte réseau PCIe, nouveau GPU, câble

Un traitement particulier : le HDD un « pollueur insoupçonné »

- Un disque dur :
 - À la production : autour de 50 kg de CO₂
 - À l'exploitation (en France) : 10W, 6 kg de CO₂/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 ?

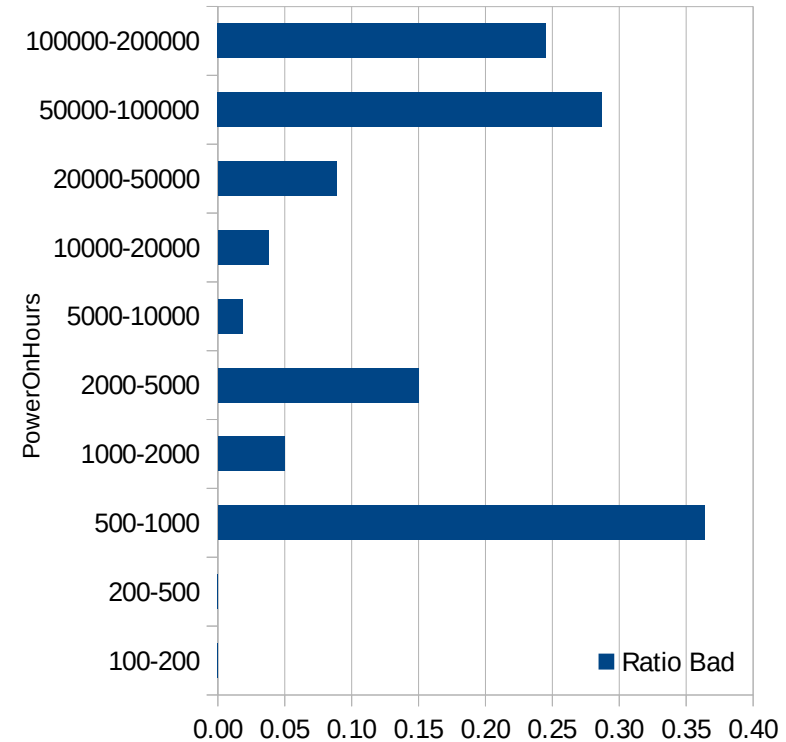
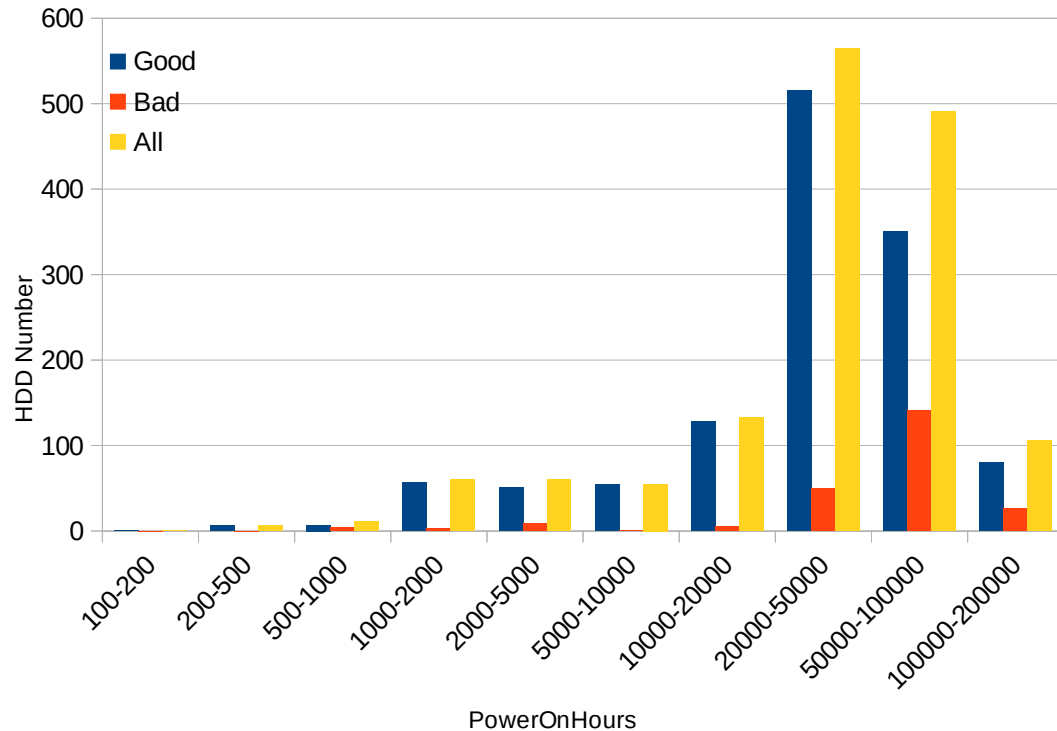
HDD : question récurrente... C'est quoi la meilleure marque ?



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

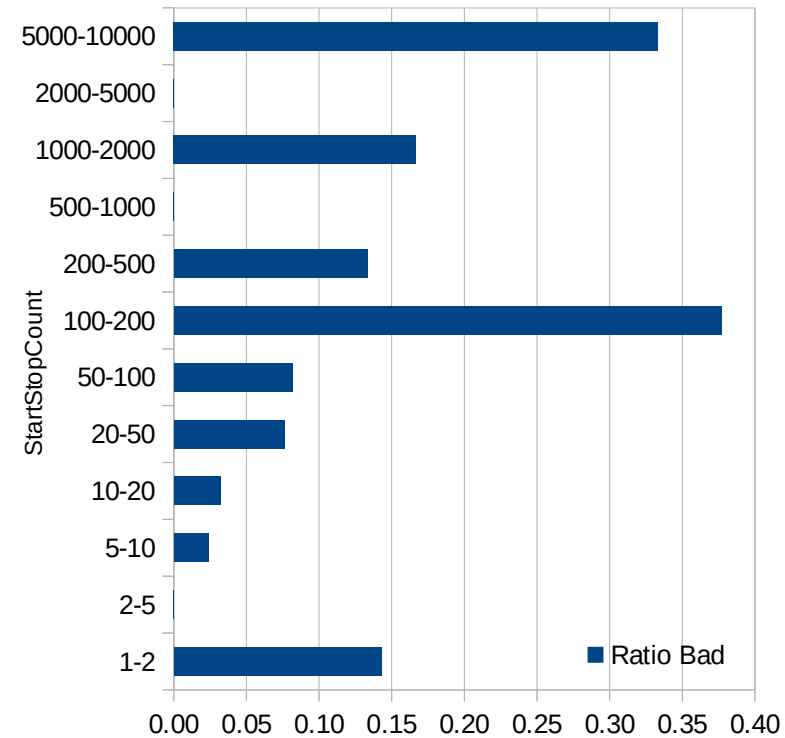
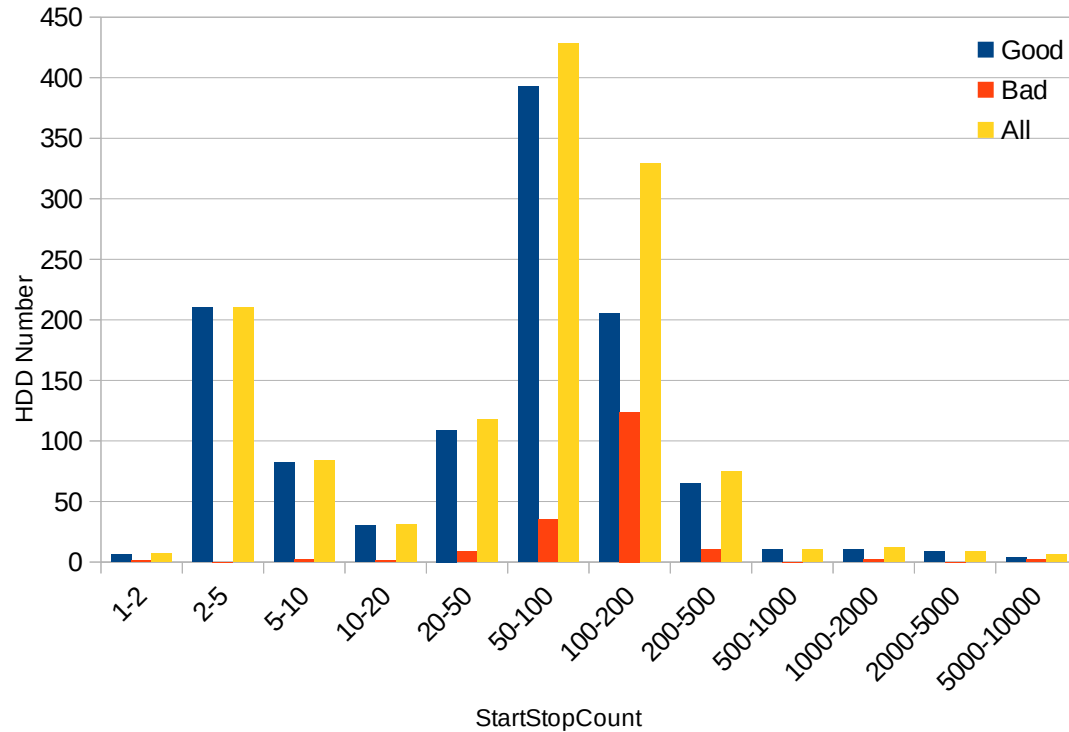
HDD : Heures de fonctionnement

Un résultat contre intuitif...



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

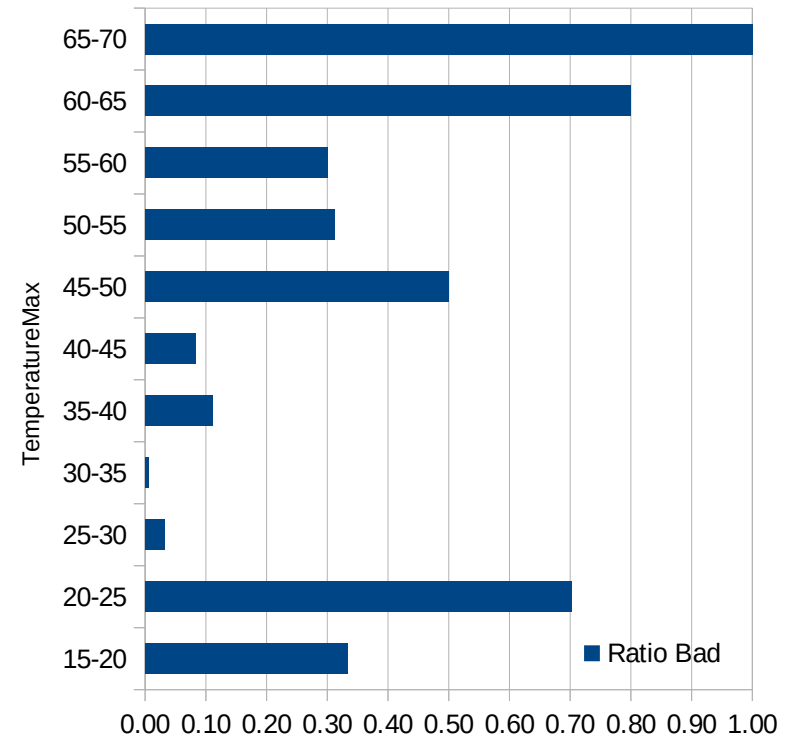
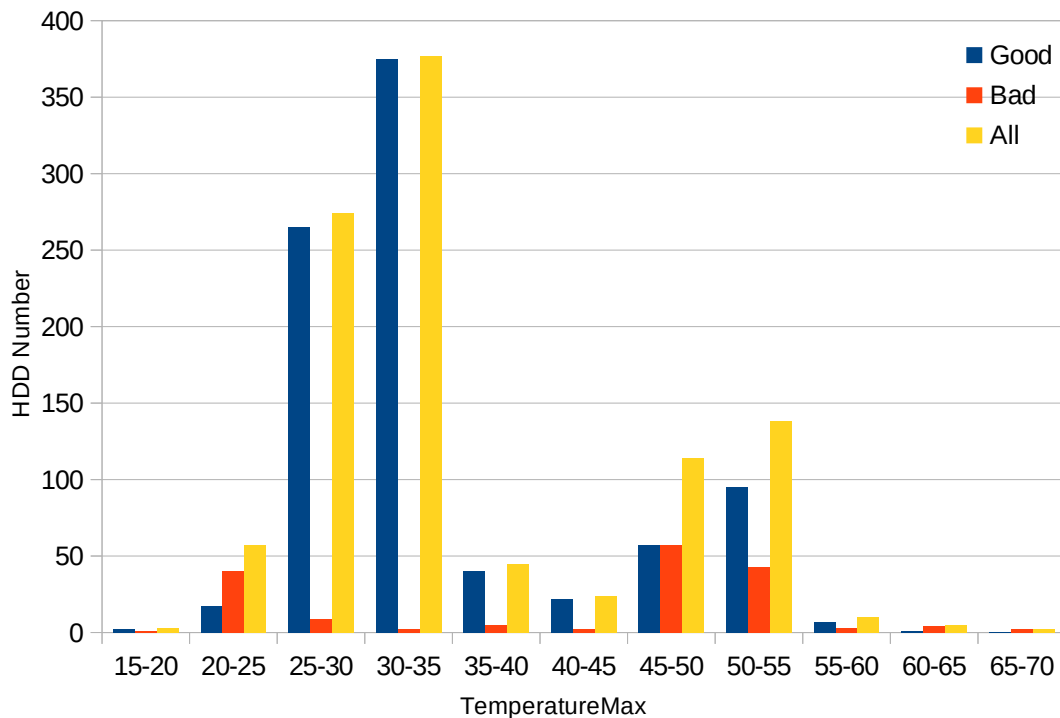
HDD : Les arrêts-redémarrages : le calvaire des disques durs ?



- 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 ?



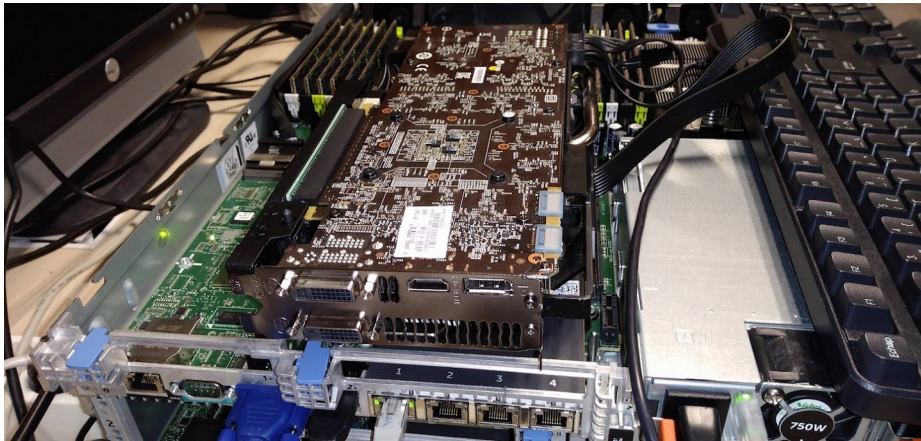
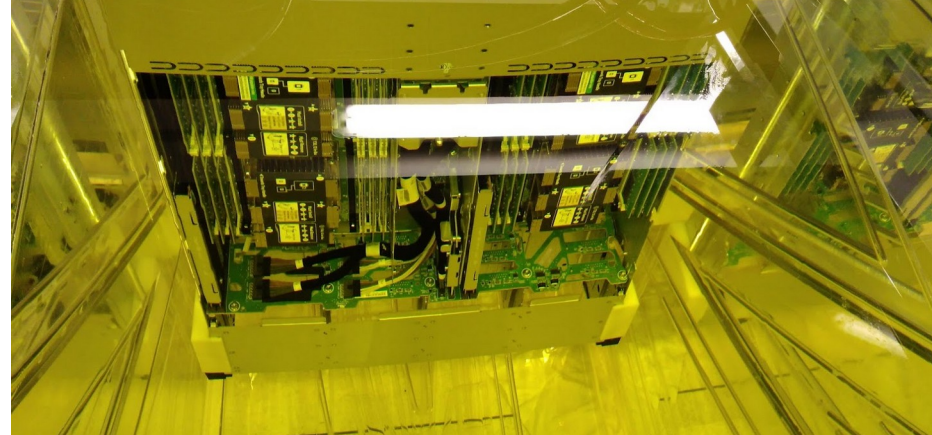
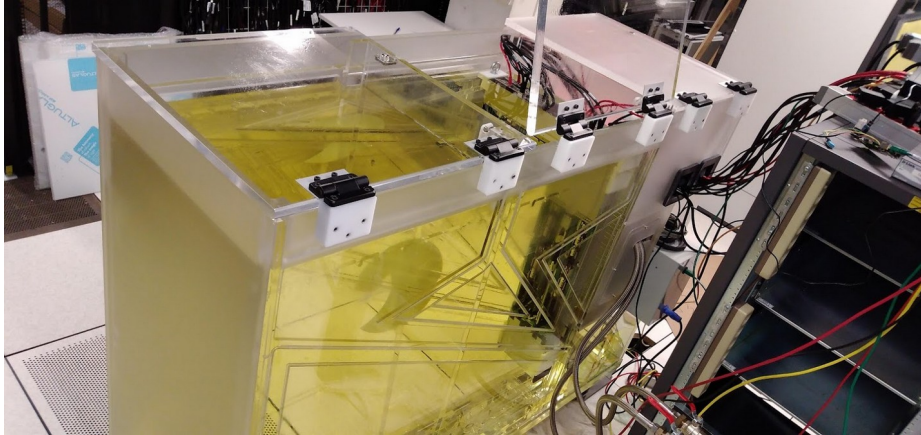
- Au dessus de 40°C de T° Max, presque 40 % sont HS
- Avoir une séparation entre HDD et CPU pertinente ?

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

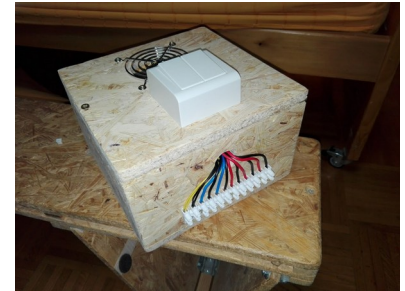
« Immersion Cooling » : réexploiter Quid : réversibilité du processus ?



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 CO₂ :
 - Finalement, en exploitation, empreinte « presque » marginale...
 - Le ratio : CO₂ Production / CO₂ 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...