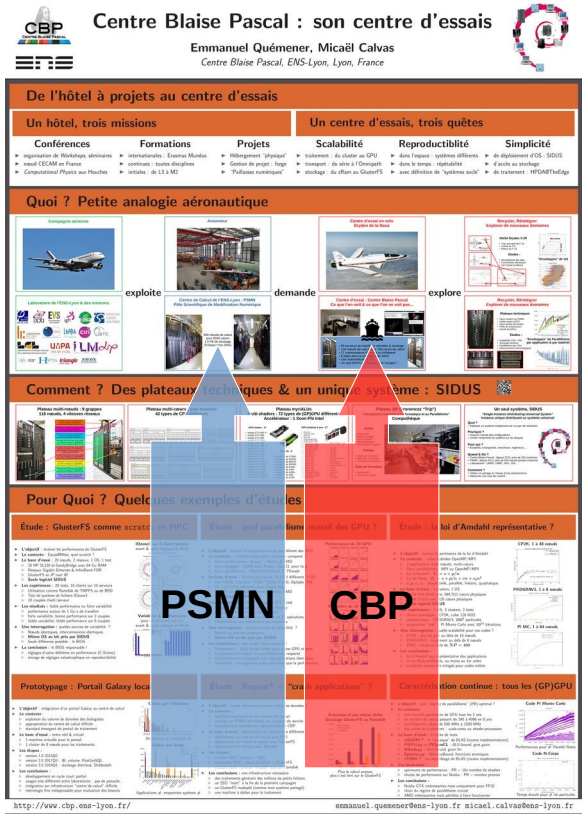
CBPsmn
Centre Blaise Pascal de Simulation et de Modélisation Numérique

JoSy ZFS 2023

Du disque dur à ZFS : 70 ans d'évolution technologique Au service des utilisateurs...

Emmanuel Quémener

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

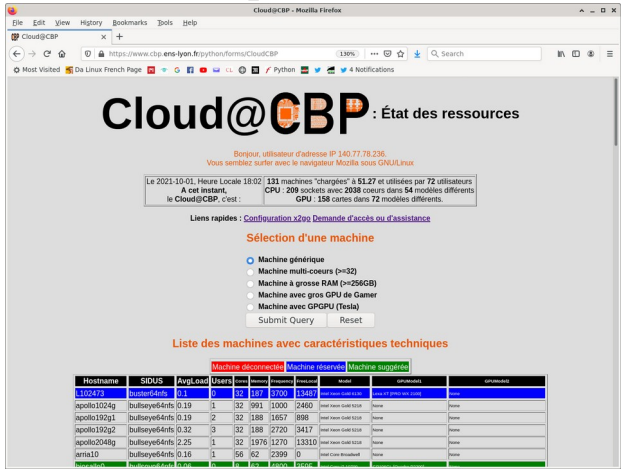


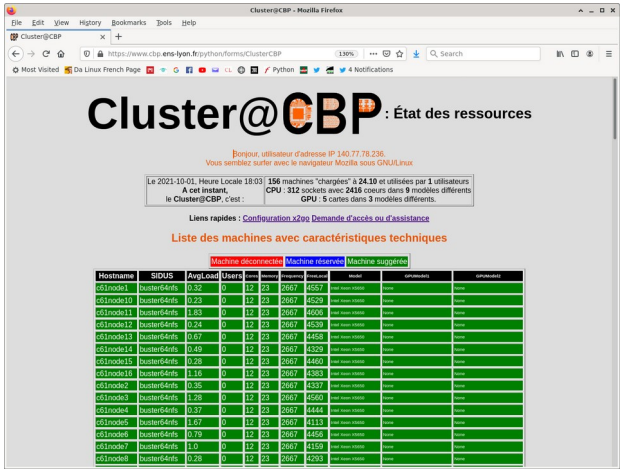
- Nasa X-29
- Cellule de F-5
- Moteur de F-18
- Train de F-16
- Etudes
  - Plans « canard »
  - Incidence >50°
  - « Fly-By-Wire »

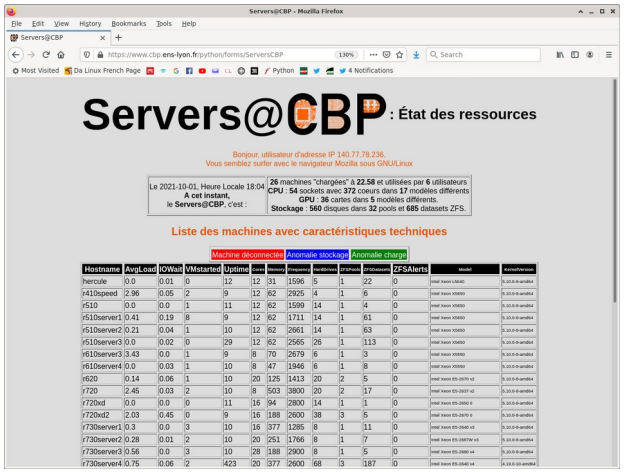
Recycler, réutiliser, explorer de nouveaux domaines…

… Et pour faire cela, il faut du matériel !

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







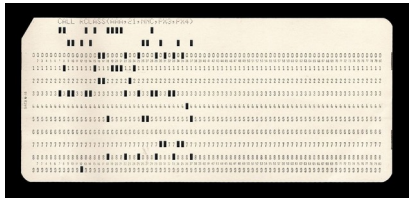
- Les 3 infrastructures Cloud@CBP, Cluster@CBP et Servers@CBP
  - 508 « sockets » cumulant 6218 « vrais » coeurs dans près de 60 modèles
  - Près de 50 TB de RAM ou équivalent (avec la DCPMM, on « triche » un peu)
  - 216 (GP)GPU dans près de 80 modèles (tout ou presque depuis 2008)
  - **1013 disques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

Pourquoi, quoi ou comment un « système de fichiers » ?

- Ce que cela ne traitera pas :
  - Le stockage distribué : pas de Ceph, GlusterFS, ...
- Ce qui sera traité :
  - De la « donnée numérique » à son exploitation et son stockage
  - L'émergence du stockage magnétique
  - L'évolution sur 30 années des disques durs
  - Pour exploiter les données, l'Operating System & ses File Systems
  - Les exigences d'une exploitation de multiples disques
  - ZFS comme système de fichiers « polyvalent »

Au commencement était le binaire... … mais comment le manipuler ?

- Stocker, c'est bien, mais l'écrire et le lire, c'est mieux !
- Stocker des données au format binaire :
  - Des trous dans des fiches (ou des «rubans perforés »)
    - Une fiche équivaut à une ligne (limitée à 72 caractères pour les programmes Fortran)
    - D'ailleurs, un « fichier », c'est un « ensemble de fiches »
  - Des dispositifs d'écriture et de lecture très différents :
    - Des poinçons pour « percer » : écriture
    - Des interrupteurs ou des cellules optiques : lecture



- Mais, deux contraintes cardinales :
  - Le support « lecture seule », la rapidité de lecture...

Un mariage (qui dure…) entre magnétisme & aérodynamique

- Stockage magnétique : un détournement technologique
  - Exploitation pour le stockage du son (1930~) et de l'image (1960~)
  - Numérique : dès les années 50 avec Univac, sur Nickel puis Mylar...
- La physique : $\overrightarrow{rot}\,\vec{B} = \mu_0 \left(\vec{j} + \frac{\partial \vec{P}}{\partial t} + \overrightarrow{rot}\,\vec{M}\right) + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$
  - L'**aimantation à volonté** : pour les lecture/écriture, mais pas que...
  - L'**effet de sol** : pour maintenir la tête à quelques nm du ou des disques
- Le premier : l'IBM 350 (du RAMAC 305)
  - Capacité de 5 Mo, Recherche en 600 ms...

La démocratisation : L'IBM PC/XT ou le Mac SE

- Fin 1983, le PC 5160, deux ans après le PC 5150
  - Même processeur i8088, 128 KB RAM, 64 Ko de ROM,
  - Un FD mais surtout un **disque dur ST506 de 10 MiB** !
  - Et un Microsoft PC DOS 2.0 : support des sous-répertoires et de HD
    - Un système de fichiers FAT12 : le C:\> apparaît mais avec...
      - Une partition maximale de 16 MiB
      - 4068 fichiers au max avec des clusters de 8KiB

- Fin 1987, le Mac SE, trois ans après le Mac 128k
  - Même processeur m68k, 4x pour la RAM, un FD
  - Mais un disque SCSI de 20 MiB au **même format presque 40 plus tard** !

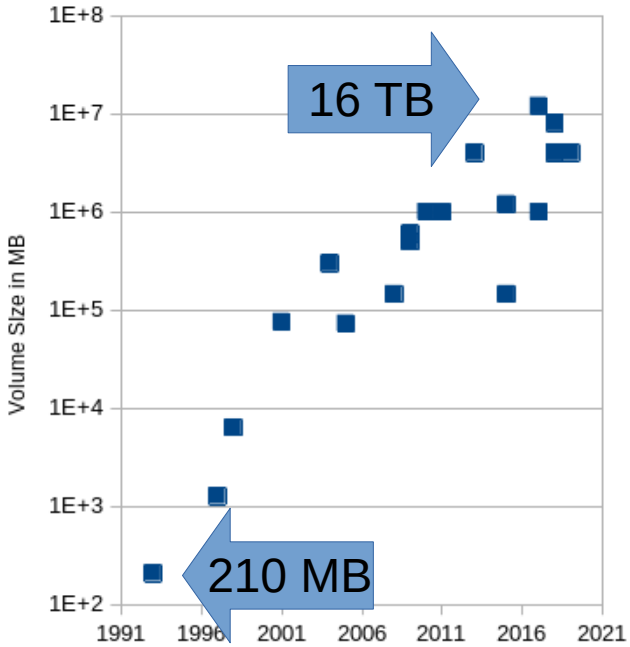
30 années d'évolution de HDD
Petit comparatif entre disques...

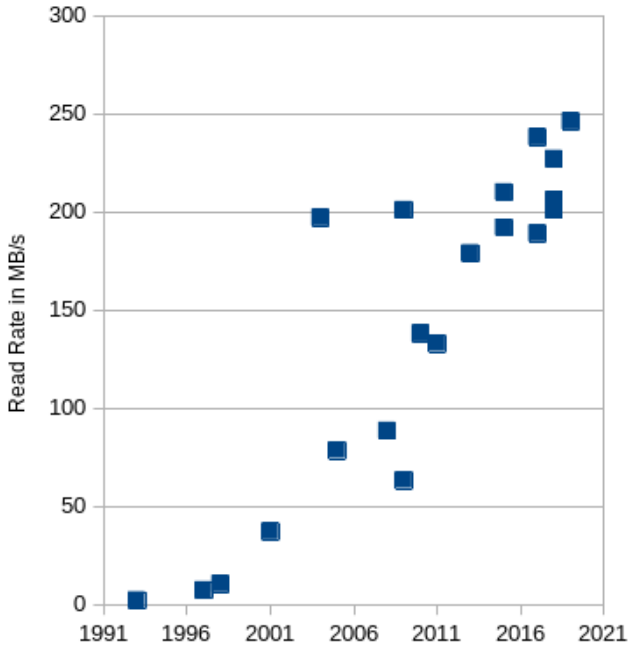
	A	B	C	D	H	I	J	K	L	M	N	O
1	Constructor	Model	Interface	Date	Size in MB	Rate in MB/s	Seek in ms	RPM	Cache in KB	Power Idle in W	Power Max in W	Altitude Max
2	Conner	CFS210A	IDE	1993	210	3	3	3600	32	3	6.2	4500
3	Fujitsu	M1636TAU	IDE	1997	1280	16.6	20	5400	128	4.1	6	3000
4	Samsung	SV0644A	IDE	1998	6400	16.6	11	5400	512	5.8	7.4	3048
5	IBM	IBM-DTLA-307075	IDE	2001	76000	32		7200	2048	6.7	12	
6	Seagate	ST9300653SS	SAS	2004	300000	202	3.1	15000	65536	4.23	7.92	3048
7	Seagate	ST373207LC	SCA	2005	73000	59	5	10000	8192	6.78	12	
8	Seagate	ST914602SSUN146G	SAS	2008	146000		4.5	10000	16384	4.5	4.5	3000
9	Hitachi	HDS7250SASUN500G	SATA	2009	500000	64.8	8.5	7200	16384	7.9	7.9	
10	Seagate	ST3600057SS	SAS	2009	600000	204	3.9	15000	16384	11.68	16.35	
11	Seagate	ST31000524NS	SATA	2010	1000000	140		7200	32768	10	10	3048
12	WD	WD10EALX	SATA	2011	1000000	126		7200	32768	6.1	6.8	
13	Seagate	ST4000DX001	SSHD	2013	4000000	190	12	7200	65536	6.2	7.5	
14	HP	EH0146FBQDC	SAS	2015	146000	300		15000	65536			
15	Seagate	ST1200MM0007	SAS	2015	1200000	204	2.9	10000	65536	4.6	7.5	
16	Seagate	ST1000LM049	SATA	2017	1000000	160	13	7200	262144	0.7	1.9	
17	Seagate	ST12000VN0007	SATA	2017	12000000	210		7200	262144	4.83	9.25	3048
18	Seagate	ST4000DM004	SATA	2018	4000000	190		5400	262144	2.5	3.7	
19	Seagate	ST8000DM004	SATA	2018	8000000	190		5400	262144	3.4	5.3	
20	Toshiba	MG04SCA40ENY	SAS	2018	4000000	205	4.7	7200	131072	6.1	11.8	
21	Seagate	ST4000NM016A	SATA	2019	4000000	215	4.16	7200	262144	2.8	3.2	3048
22	Seagate	ST16000NM010G	SAS	2020	16000000	249		7200	262144	5	10	
23	Seagate	ST1000NM0018	SATA	2020	1000000	194		7200	131072	4.7	7	

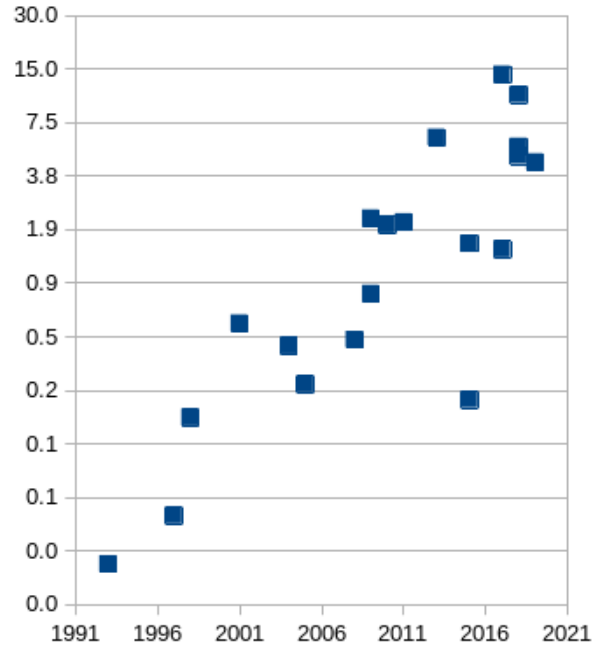
- 22 modèles, 9 marques : IDE, SCA, SATA, SSHD, SAS…
  - Entre 1993 et 2020, petite collection (encore très active).

30 ans d'évolution des HDD
Petit test comparatif...

- De la capacité au débit crête : une évolution différente...



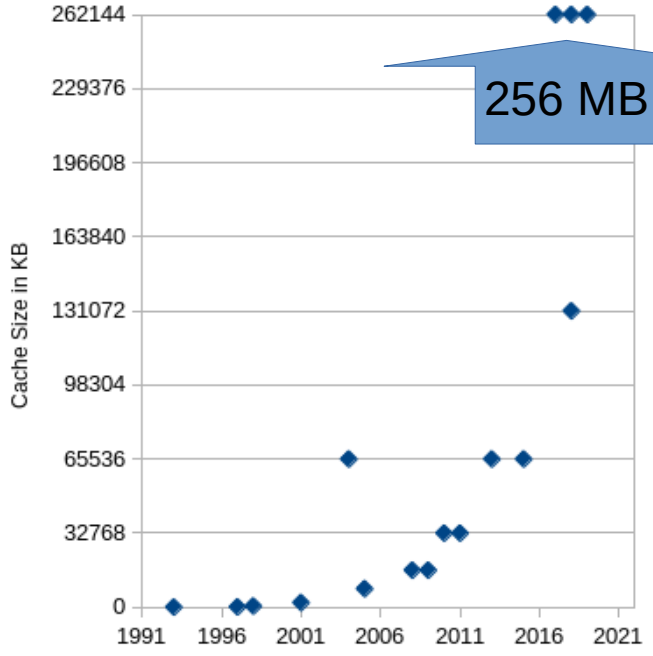


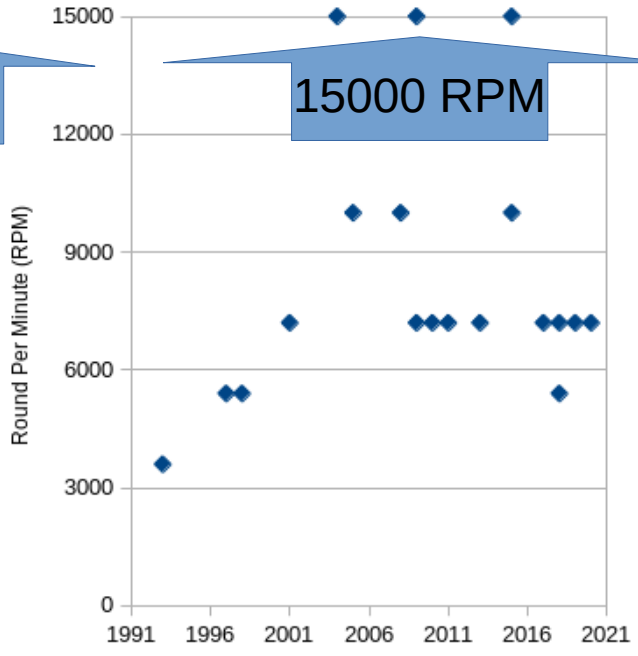


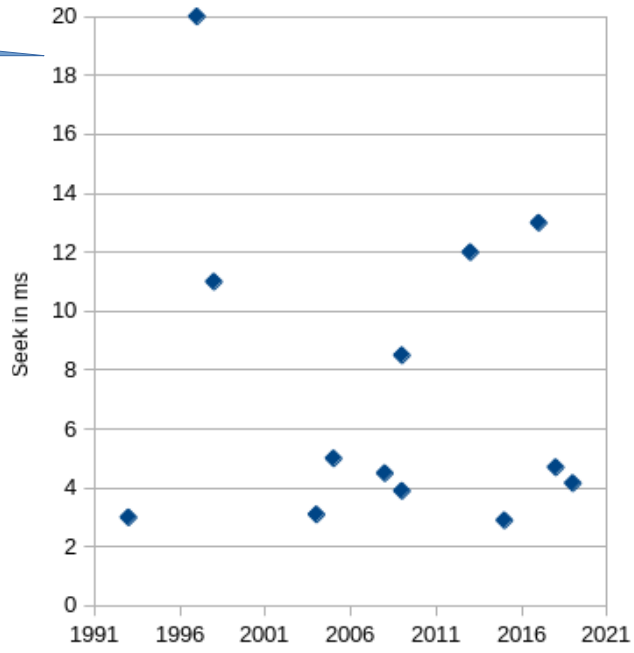
- En 20 ans : un x200 en volume en 20 ans, un x8 en débit
  - De 1/2 heure à 15 heures pour « dumper » un disque (au mieux)...

30 ans d'évolution de HDD : autres métriques significatives...

- Le cache, la vitesse de rotation et la latence...



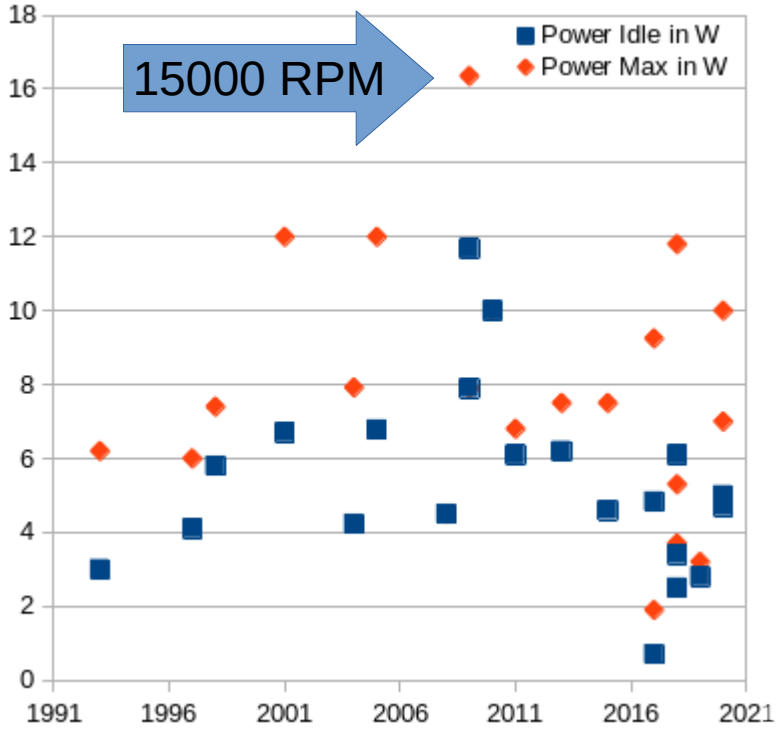


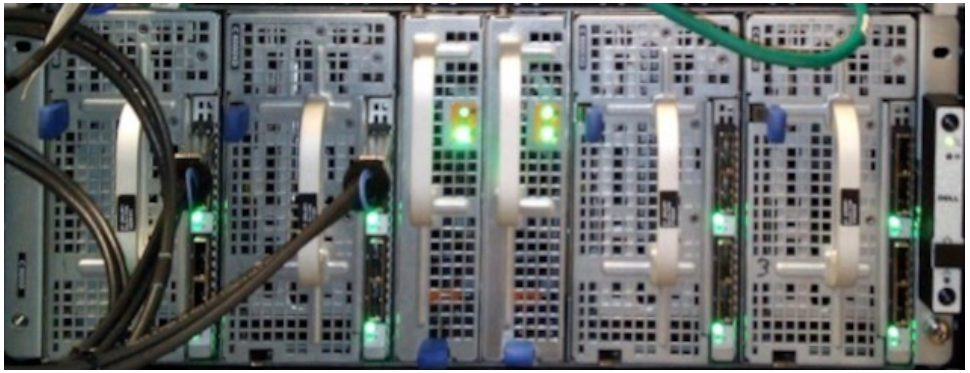


- Peu d'évolution depuis 5 ans sur les disques de serveurs
  - Même vitesse de rotation (et même latence), même cache...

30 ans d'évolution de HDD dernière métrique, pas la moindre...

- La consommation : en baisse avec la baisse des RPM...





- Pour 48 disques, comptez entre 250W et 500W...

Derrière le Hard, le Soft & l'OS
Des besoins cardinaux...

- Lorsque le « soft » aide le « hard » (en fait le BIOS) :
  - Jusqu'à présent, limitation de l'implémentation logicielle : 640kB…
  - Avec MS-DOS 6, « drivespace » et son « compresseur » de disques
- Ça avec des disques aux capacités qui explosent :
  - De 20 MiB en 1984 à 75 GB en 2001 : ~ x2 tous les 17 mois
  - De 75 GB en 2001 à 22 TB en 2023 : ~ x2 tous les 33 mois
- Mais une nécessité triple d'exploiter :
  - Plusieurs HDD simultanément comme un même espace
  - Un « volume physique » unique séparé en plusieurs « volumes logiques »
  - Des fonctions comme les quotas, la journalisation, les instantanés, …

2000’ : des « Operating Systems » aux « File Systems »

- Début 2000 : chaque architecture matériel a son OS
  - AIX pour IBM, HP-UX pour HP, IRIX pour SGI, SunOS/Solaris pour Sun,
  - DigitalUNIX, Tru64, VMS pour DEC, et autres Unix*
  - MacOS pour Apple, AmigaOS pour Amiga, DOS et GNU/Linux pour x86
- Et chaque OS propriétaire a son FS
  - JFS pour AIX, HFS pour HP-UX, XFS pour IRIX, UFS et **ZFS** pour Solaris
  - HFS et HFS+ pour Apple, FAT et NTFS pour M$
  - Et GNU/Linux : ext2/3/4, ReiserFS, BtrFS, mais aussi les précédents...
- Et chacun des OS avec ses méthodes pour le matériel...

Agréger les disques : les 3 « R » rudesse, rapidité, résilience...

- Agréger des disques et les niveaux de RAID
  - *Redondant Array of **Inexpensive** Disks* : la vertu de l'économie
  - RAID Linear : des disques agrégés simplement
  - RAID0 : des disques agrégés avec des accès parallélisés
  - RAID1 : des disques redondés (avec des « spares » possibles)
  - RAID4 : une parité stockée sur un disque des autres disques
  - RAID5 : une parité distribuée sur les disques
  - RAID6 : une double parité distribuée sur les disques
- Sous GNU/Linux, le classique : MD (avec mdadm)

Mais des matériels savent « gérer » le RAID…

- Émergence de contrôleur de disques permettant le RAID
  - Essentiellement des contrôleurs SCSI ou SCA
  - Fonctionnalité frisant l'escroquerie avec la HPT370 pour disques IDE
  - Efficacité sur les cartes 3ware : RAID1 pour IDE
    - Support Debian depuis 2000 dès l'installation de l'OS
    - Disponibilité pour les plus gros disques IDE de l'époque
- Généralisation dans les machines de Matinfo Dell ou HP
  - Cartes MegaRAID ou PERC (série H700), LSI 1068, ...
- Remplacement (avantageux ?) de MDADM/LVM…
  - Quoique : la gestion du remplacement de disque ou la lenteur de création...

Diviser pour mieux exploiter… Des besoins en évolution

- Au début était le partitionnement :
  - Pour contourner les limitations du matériel
  - Pour éviter les « transpirations » de dossiers : croissance incontrôlable...
- Après s'est invitée la virtualisation : de « gros blocs »
  - Des exigences de « fichiers » ou « volumes » depassant TiB
  - Des besoins de gestion des « états » : archivage
- Une solution : LVM pour « Logical Volume Manager »
  - Une partie gestion du matériel avec pv* : pvcreate, pvresize, pvdisplay, ...
  - Une partie gestion des volumes virtuels vg* : vgcreate, vgresize, vgdisplay, ...
  - Une partie gestion des volumes logiques lv* : lvcreate, lvresize, lvdisplay, …

Gérer les « fichiers » ... Et leur environnement utilisateur !

- Le B.A.BA : respectez de « DICT » des RSSI :
  - **Disponibilité** (avec l'accès aux fichiers et l'indexation), **Intégrité** (avec la somme de contrôle),
  - **Confidentialité** (avec la gestion des droits), **Traçabilité** (avec l'horodatage)
  - En gros, tout ce qui concerne l'arborescence, la gestion des droits, ...
- Mais des besoins étendus (utilisateurs ou système) :
  - Quota, archivage, sauvegarde, extensibilité, ...
  - Compression, déduplication, chiffrement, …
- Question de choix, mais avec de nouveaux défis :
  - Fournir une arborescence partagée avec des millions de fichiers
  - Permettre des « roll back » rapides
- Tout ça, la « vache » (britannique) : là le « Copy On Write » se développe...
  - Personnellement : XFS pour le $HOME et BtrFS pour les racines SIDUS

Et ZFS émergea comme une évidence...

- En partant du principe qu'il est nécessaire :
  - Pouvoir agréger des dizaines (centaines) de disques rapidement
  - Imposer une parité simple, double voire triple
  - Disposer d'un mécanisme de réservation en « mode bloc » extensible
  - Proposer la fonctionnalité de « mode fichier »
  - Offrir ces volumes en « mode fichier » par NFS ou SMB
  - Permettre les mécanismes de chiffrement, compression, déduplication
  - Supporter la création d'instantanés et faciliter leur sauvegarde
  - Gérer les quotas, ...
- Alors ZFS remplace : MDADM/LVM et un FS avancé...

ZFS dans deux commandes : « zpool » pour gérer les disques...

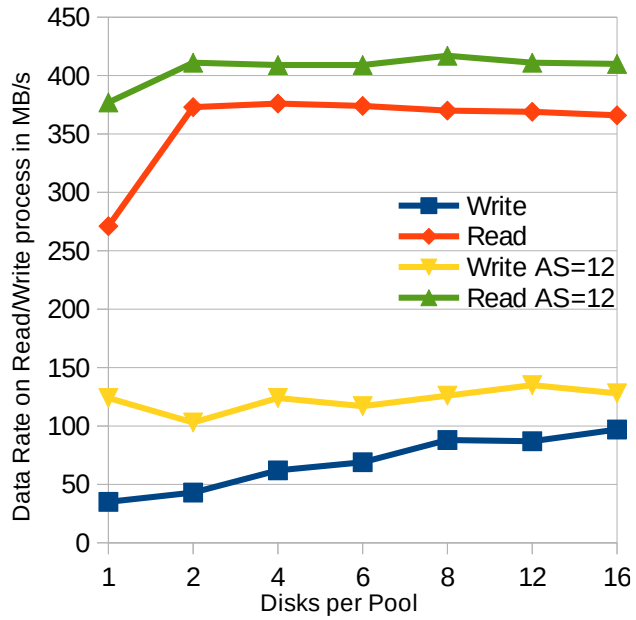
- Seulement 2 commandes à connaître : zpool et zfs
- Créer son « ensemble » de disques : avec zpool
  - zpool create <NomDuPool> <NiveauDuPool> <ListeDeDisques>
  - <NiveauDuPool> : mirror (~ RAID1), raidz1 (~ RAID5), raidz2 (~ RAID6)
  - <ListeDeDisques> : liste des « périphériques » de /dev/
    - privilégier des disques « entiers », mais ça marche avec des partitions !
    - Moi, j'utilise plutôt les wwn (Word Wide Number) que les autres...
  - Des options (souvent) nécessaires : le « -f » et le « -o ashift=12 »
- Visualiser : zpool status <NomDuPool>
- Remplacer : zpool replace <NomDuPool> <Bad> <Good>
- Ajouter un ensemble : zpool add <NomDuPool> <Niveau> <Liste>

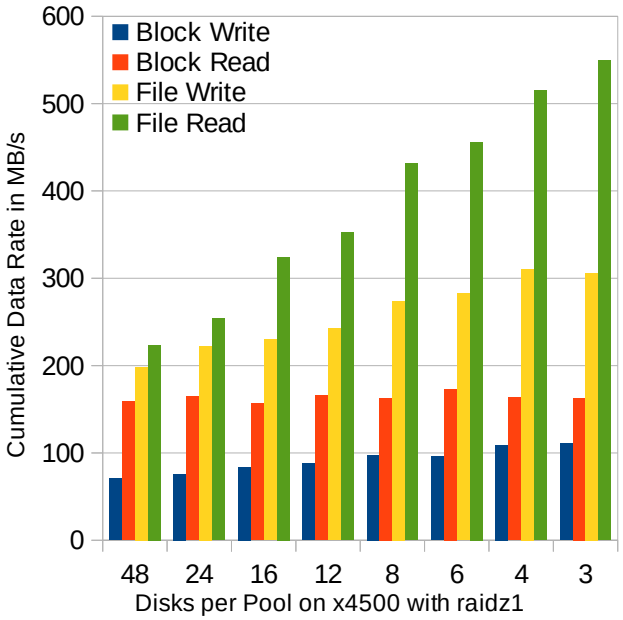
ZFS dans deux commandes : « zfs » pour gérer les contenus...

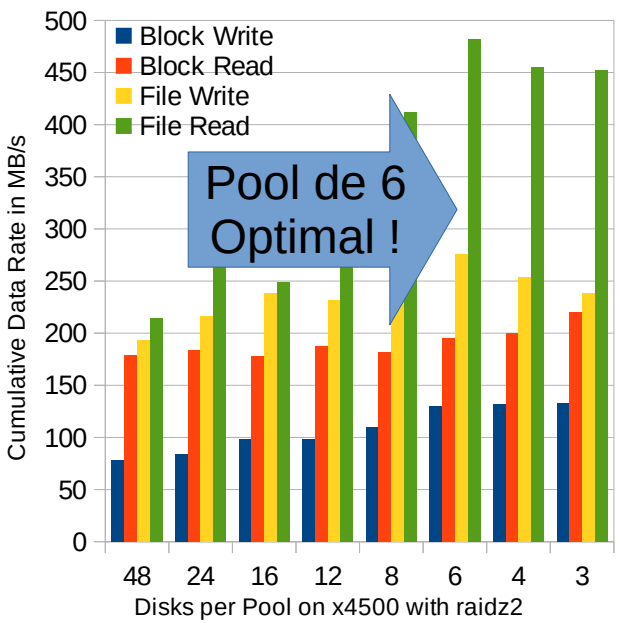
- Pour créer soit un dossier, soit un volume :
  - Pour un dossier : zfs create <NomDuPool>/<Dossier>
  - Pour un volume : zfs create -V <Taille> <NomDuPool>/<Volume>
  - Avec des options à la pelle :
    - Activer la compression : zfs set compress=lz4 <NomDuPool>/<Dossier>
    - Changer le point de montage : zfs set mountpoint=<Montage> <NomDuPool>/<Dossier>
- Pour effectuer un instantané, le fameux « snapshot » :
  - zfs snapshot <NomDuPool>/<Dossier>@<ChaineDeCaracteres>
- Pour envoyer ou recevoir : zfs send / zfs receive

Un peu de métrologie ZFS X4500 : « ordinosaure laboratoire »

- Une machine « autonome » à 48 disques SATA de 2006…
  - En « poussant » ou « tirant » sur tous les disques : 796 MB/s et 1024 MB/s



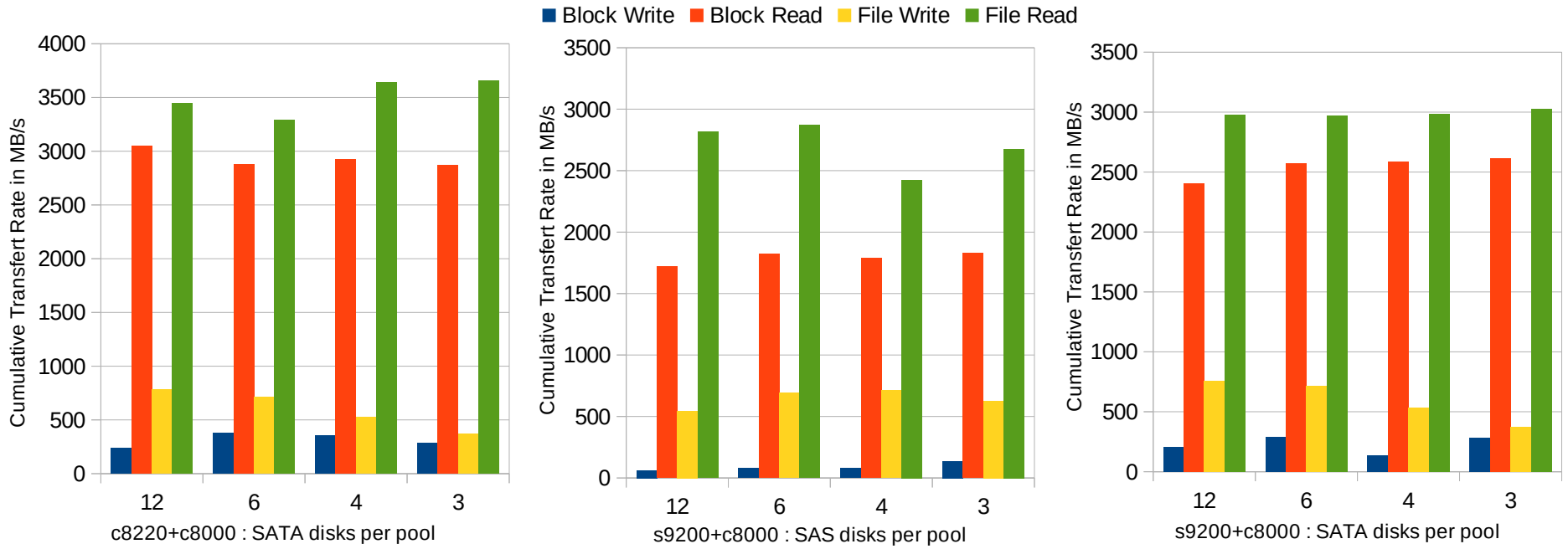




- La X4500 nous enseigne : ashift=12 & pools de 6 disques !

Toujours de la métrologie : Et avec des pools de 12 disques ?

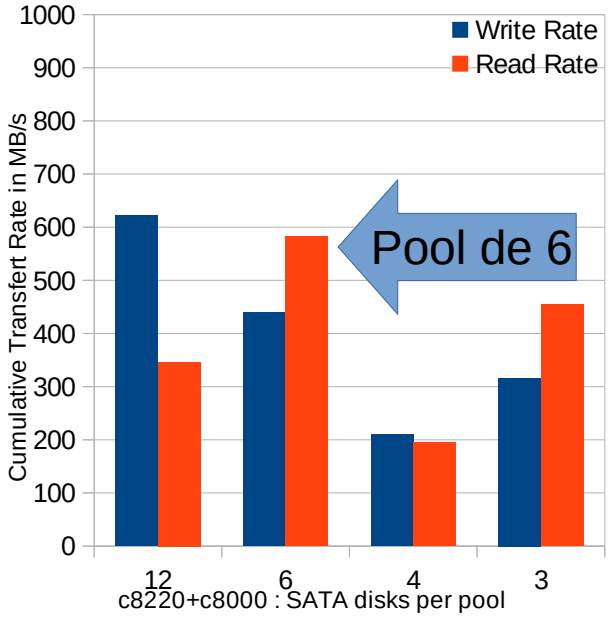
- Des blocs de c8000 de 12 disques, lien SAS 6Gb/s
  - 3 configurations : c8220 + 12 SATA, s9200 + 12 SATA, s9200 + 12 SAS

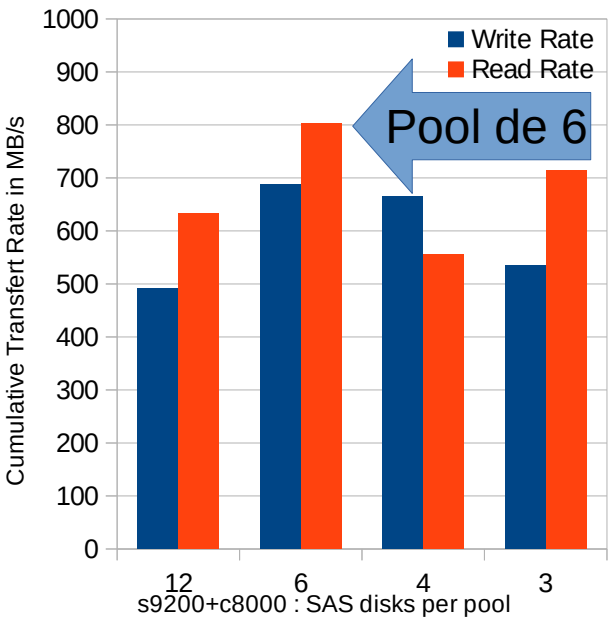


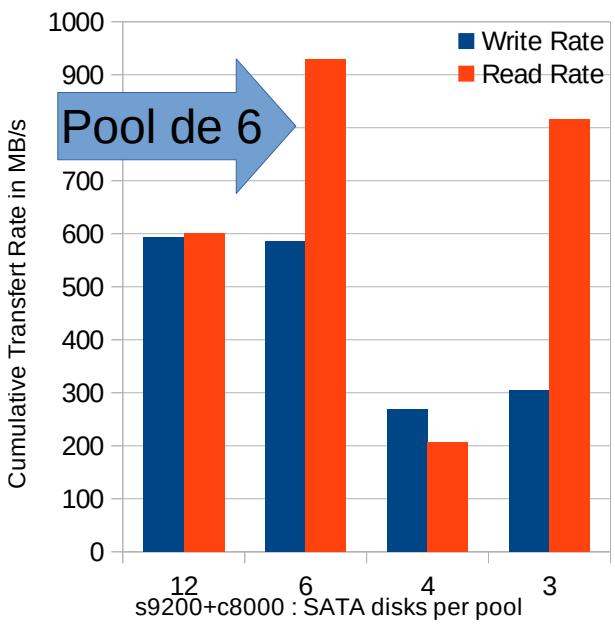
- Des performances « intriguantes », surtout pour les SAS

Et pour des très gros fichiers… Avec ces baies de 12 disques...

- Pour une écriture/lecture de 4 fichiers de 128 GB...
  - 3 configurations : c8220 + 12 SATA, s9200 + 12 SATA, s9200 + 12 SAS



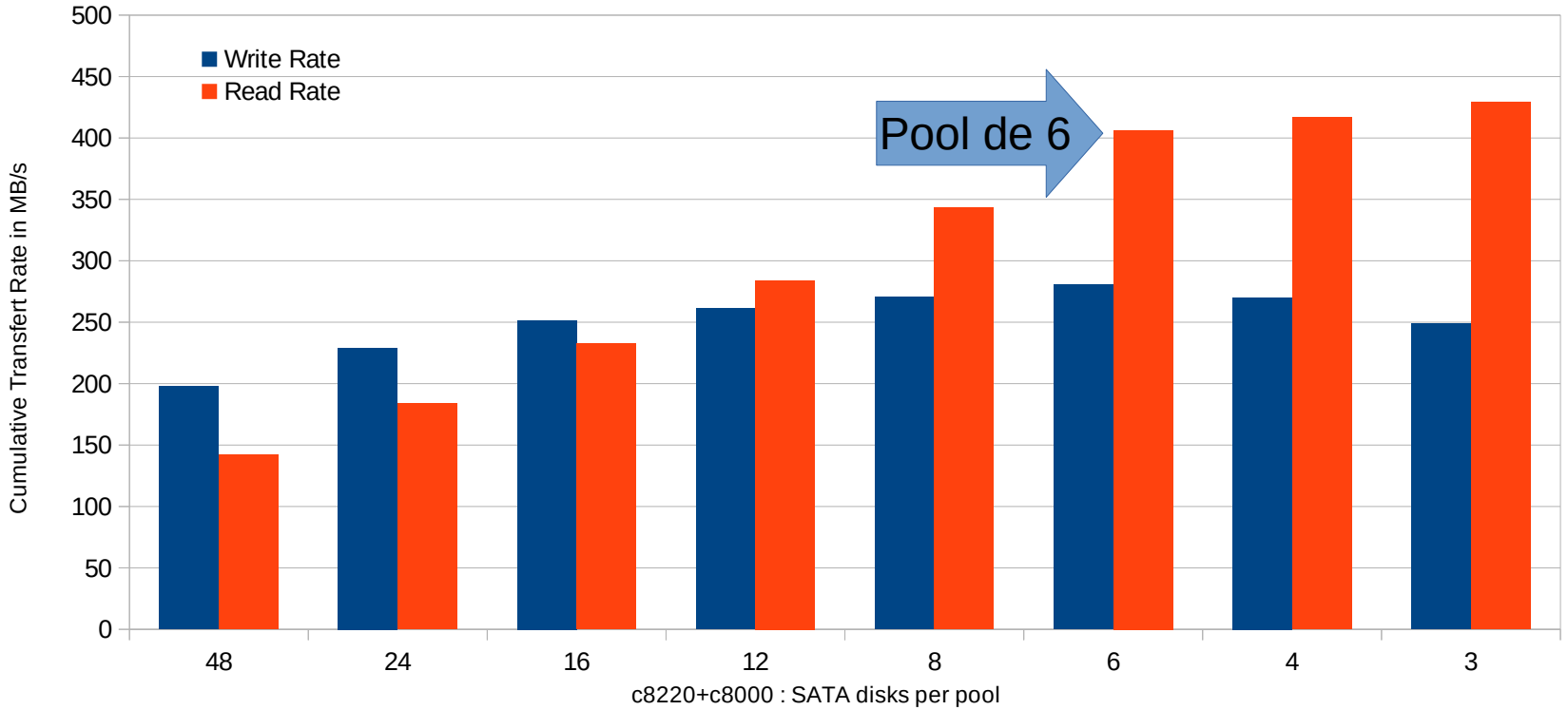




- Des pools de 12 disques, certains identiques, mais...

Toujours sur des « gros fichiers »
Sur la bonne vieille x4500

- Malgré ses « seulement » 4 coeurs et 16 GB de RAM...



- La différence se « joue » sur la lecture : pool > 6 disques

En conclusion
ZFS c'est bien, mais...

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

- ZFS remplace les couches MDADM+LVM+FS évolués
- ZFS permet de s'affranchir du matériel « captif »
- Mais avec quelques inconvénients :
  - Pas d'intégration « directe » dans le noyau (licence)…
  - Pas de possibilité de supprimer des disques d'un pool (comme LVM)
  - Pas de possibilité de fichier non COW (pour du swap par exemple)

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

Iconographie

- Par Mutatis mutandis — Travail personnel, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1406914
- By Jud McCranie - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=116588518
- https://www.refletsdelaphysique.fr/articles/refdp/pdf/2010/01/refdp201018p12.pdf
-