Les "nouveaux" systèmes de fichier

ARAMIS

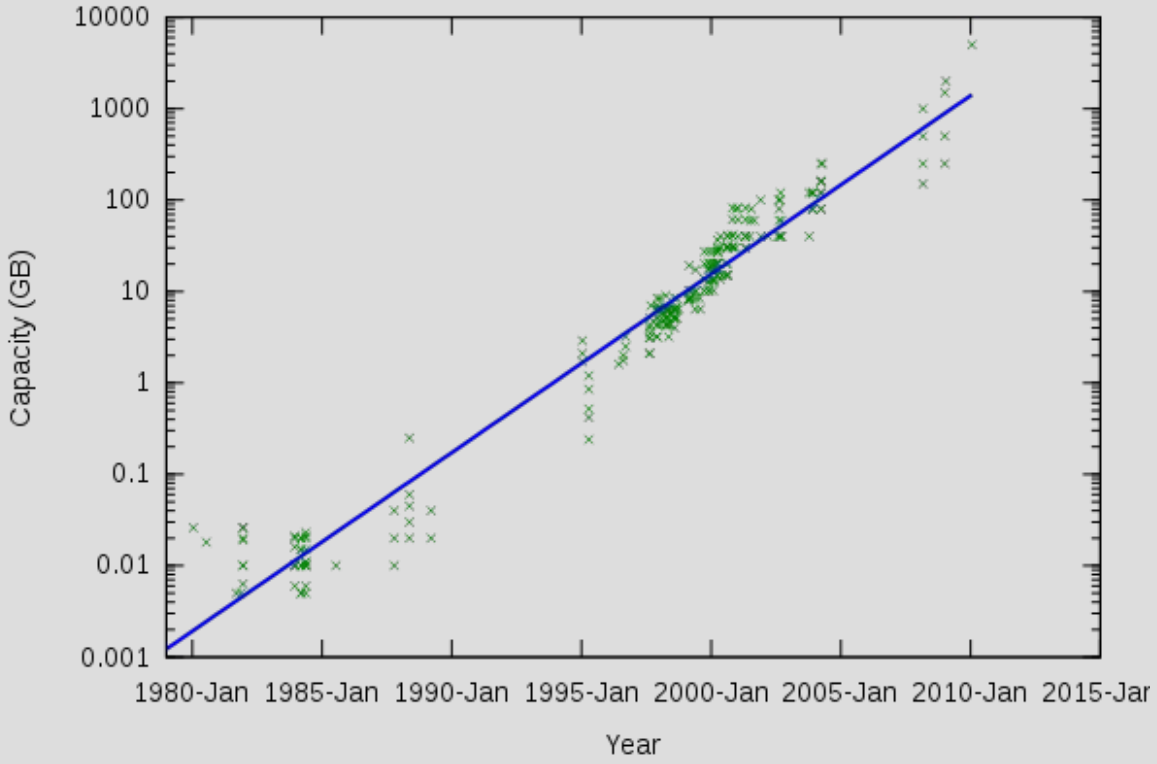
Se rappeler de FFS, ZFS,XFS,NTFS...

Un peu de vocabulaire

Découvrir BTRFS, EXT4, ExoFS, NilFS2

Pour rattraper un café ARAMIS !

Une loi de Moore pour les disques durs ?



Évolution de la taille des disques

Petit rappel sur la mécanique

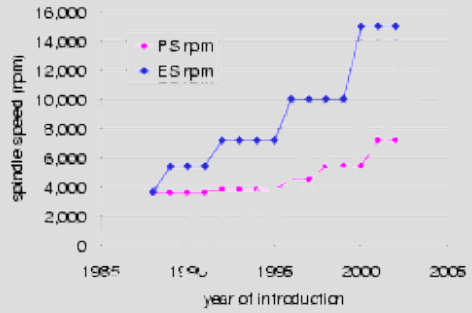


Figure 1: The adoption of higher rotational speeds. Data from Control Data product guides, 1988-1989 and Seagate product guides, 1990-2002.

Aujourd'hui : Vitesse de rotation moyenne autour de 10000RPM impliquant un temps d'accès autour de 5ms

Si 200 I/O. Si le bloc adressable fait 512o alors 100ko/s de bande passante au mieux.

SSD : Une petite révolution

- Temps d'accès
- Très haute densité volumétrique
- Faible consommation
- Fiabilité

Inconvénient : le prix au gigaoctect

```
aragorn:~# du -sh /home/maurin/Maildir/
2.7G    /home/maurin/Maildir/
aragorn:~# sync ; echo 3 > /proc/sys/vm/drop_caches ; time find /backup-
   online/home/maurin/Maildir/  -type f -exec head -1 \{\} \; -ls | wc -l
69706
real    1m36.433s
...
730 => I/O par seconde sur un RAID5 3HD Sata 7.2k 160Go
...
aragorn:~# sync ; echo 3 > /proc/sys/vm/drop_caches ; time find /home/maurin/Maildir/  -type f
   -exec head -1 \{\} \; -ls | wc -l
69771
real    0m31.452s
...
2250 => I/O par seconde sur un RAID5 3SSD Sata 7.2k 160Go
...
```

Problème de fiabilité

La pannes matérielle :

- >>1% an [B.Schroeder&GGibson Fast07]
- Facile à détecter
- Partie ou intégralité du disque affecté
- Exemple de donnée constructeur : MTBF > 1,2.10e6heures

La corruption matérielle :

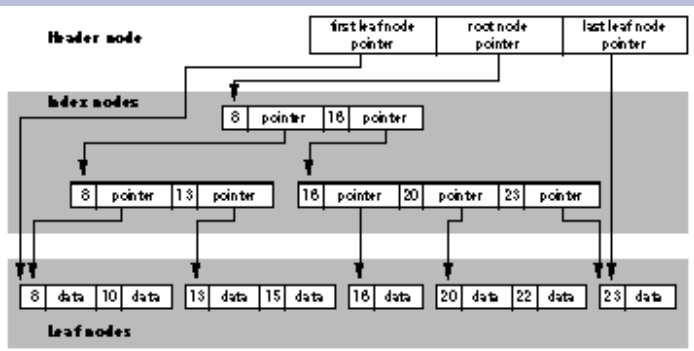
- Difficile à quantifier ~1‰ sur du SATA [L.Bairavasundaram,Login06]
- Pas détectée ou à posteriori sur des reconstructions RAID
- Exemple de donnée constructeur : 1 secteur (512bits) pour 10e15bits avec un code de correction d'erreur de 10b

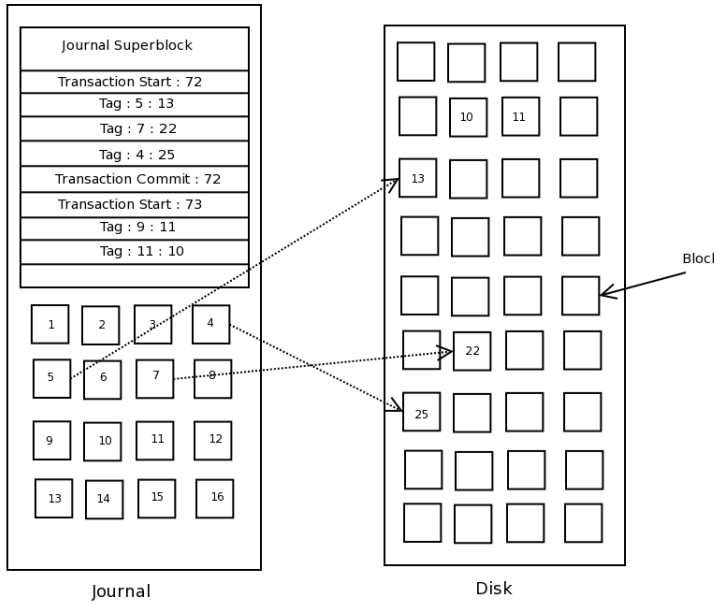
Les systèmes de fichiers actuels

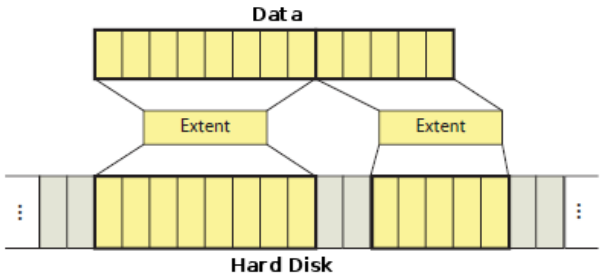
NTFS
Ext2/3/4
XFS
HFS
FFS
FAT
... etc.

Vocabulaire

- B+tree
- Journal
- extents







Histoire d'un système de fichier

Fast File System

1979 : Prototype BSD 3.0-4.0
1982 : Naissance BSD 4.2 – une optimisation de l'espace
1986 : Abstraction des calculs d'adresse par le matériel
1987 : Empilement de couche de système de fichier virtuel
1988 : Doublement de la taille des blocs (8ko) et des fragments (1ko)
1990 : Allocation dynamique des blocs
1996 : Écritures asynchrones
1999 : Clichés disque
2001 : Doublement de la taille des blocs (16ko) et des fragments (2ko)
2002 : Vérification en ligne
2003 : Adressage 64 bits
2004 : Acces Control Lists
2005 : Mandatory Acces Controls
2006 : Code multi-processeur

Bilan de 25 ans d'évolution FFS

- Augmentation de l'espace d'adresse
- Écritures séquentielles et atomiques
- Ajout de fonctionnalités

NTFS

- Arrive en 1993 avec NT3.1
- Évolution majeure avec NT5
- Système de fichier de la majorité des utilisateurs de micro-informatique

Port NTFS-3G stable depuis 2007 !

EXTended FS

- V4 Passe en stable en 2010
- Taillé pour attendre la suite...

XFS

- Arrive en 1994 avec IRIX 5.3
- Passe en GPL avec port linux en 2000
- Conçu pour fonctionner en environnement multiprocesseurs, avec un gestionnaire de volume pour de grandes partitions de données

ZFS

- Arrive en 2005 sur Solaris, code libéré en 2006 en CDDL
- Système de nouvelle génération

Nouveaux systèmes de fichiers Linux...

- BTRFS
- NilFS2
- EXOFS

Nouveaux concepts

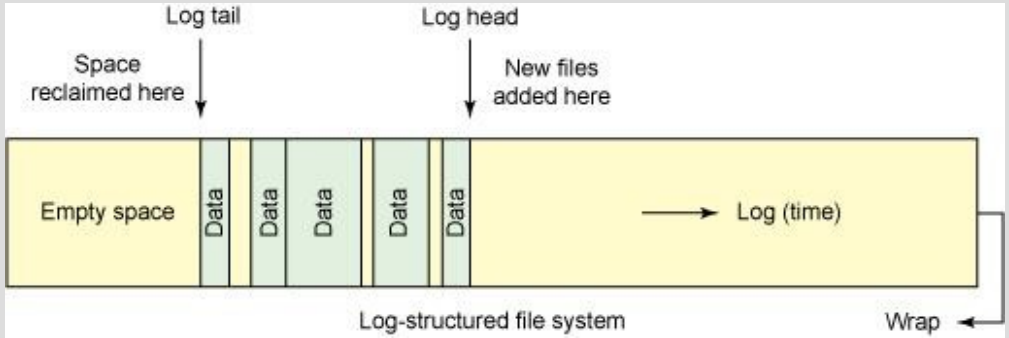
- Gestion du stockage
- Sécurisation de l'information
- Extension du modèle objet

BTRFS

- Proposé en 2007, intégré en 2009 au kernel, en expérimental
- Supporté par Oracle

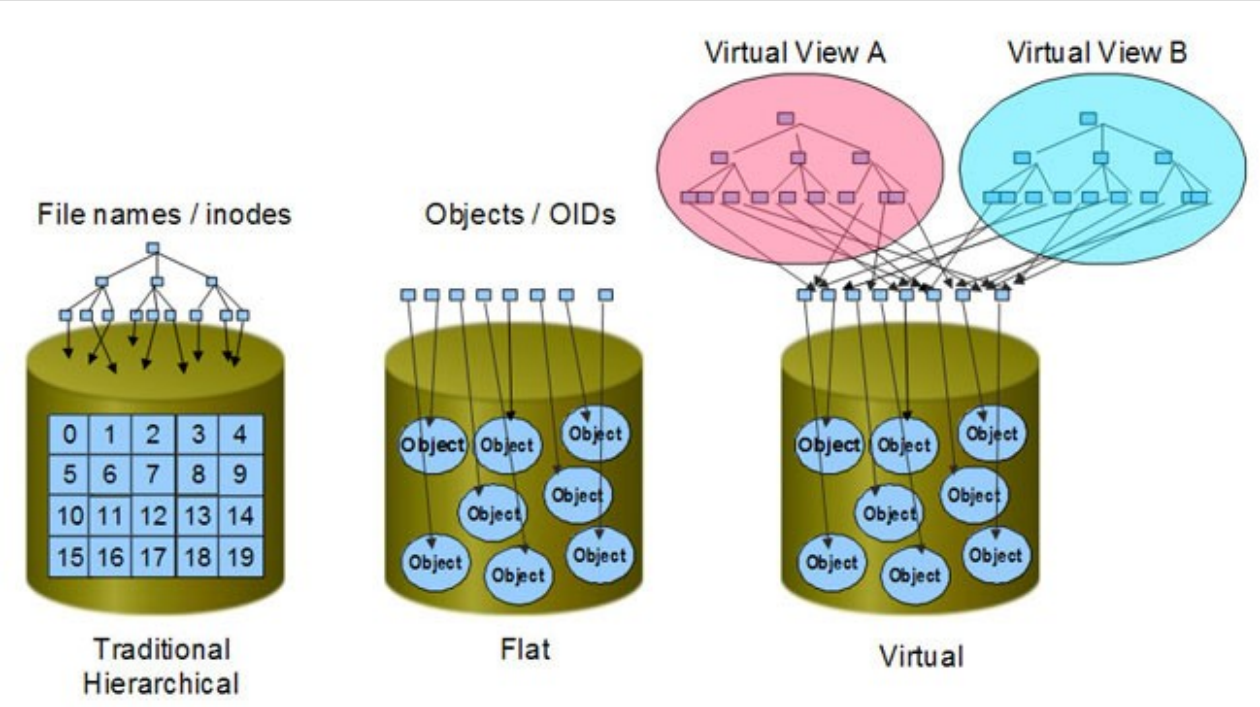
NILFS(2)

- NilFS(v1) apparaît en 2005, NILFS(2) est intégré à Linux en 2009
- Supporté par Nippon Telegraph and Telephone (NTT)

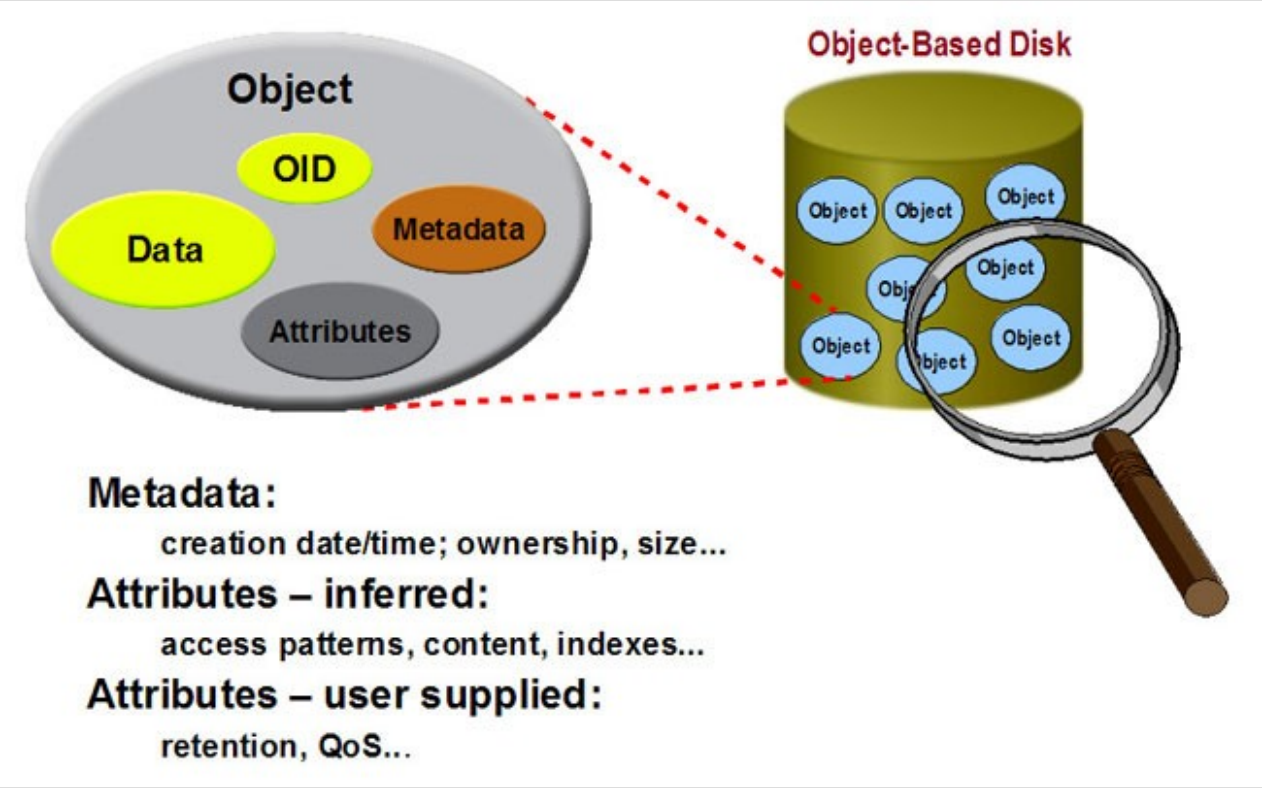


EXOFS

- Dérive de ext2 (Extended Objet File System)
- Introduit par IBM, intégré au kernel en 2009
- Orienté périphérique de support objet

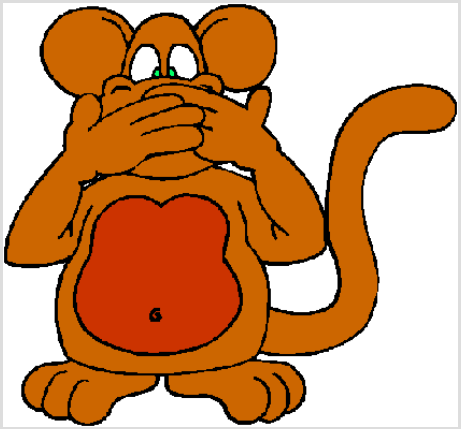


Périphérique de support objet



Ici on ne parlait pas...

- Des systèmes de fichiers distribués (Google File System, General Parallel File System, Lustre, Ceph le DFS objet, ...)
- Des gestionnaires de volumes (XVM,LVM, ...)
- Des gestionnaires de blocs (Soft/Hard RAID, DRBD, ...)



Références

1. "5.4 B-Trees," n.d., http://lcm.csa.iisc.ernet.in/dsa/node121.html.
2. Kerner Sean Michael, "A Better File System for Linux? - InternetNews.com," 10, 2008, http://www.internetnews.com/dev-news/article.php/3781676.
3. "Anatomy of ext4," n.d., http://www.ibm.com/developerworks/linux/library/l-anatomy-ext4/index.html.
4. "Ceph: A Linux petabyte-scale distributed file system," n.d., http://www.ibm.com/developerworks/linux/library/l-ceph/index.html.
5. "Ext4 Wiki," n.d., https://ext4.wiki.kernel.org/index.php/Main_Page.
6. "Hard Drive Performance Characteristics," n.d., http://www.redhat.com/docs/manuals/linux/RHL-9-Manual/admin-primer/s1-storage-perf.html.
7. "Linux: The Journaling Block Device | KernelTrap," n.d., http://kerneltrap.org/node/6741.
8. Jone Tim, "Next-generation Linux file systems: NiLFS(2) and exofs," Next-generation Linux file systems: NiLFS(2) and exofs, 10, 2009, http://www.ibm.com/developerworks/linux/library/l-nilfs-exofs/index.html.
9. "NTFS - Wikipedia, the free encyclopedia," n.d., http://en.wikipedia.org/wiki/NTFS.
10. "NTFS Technical Reference" http://technet.microsoft.com/en-us/library/cc758691(WS.10).aspx
11. "Object-Based Storage Devices," n.d., http://developers.sun.com/solaris/articles/osd.html#History.
12. "Seek time vs througput - 2nd USENIX Conference on File and Storage Technologies — Technical Paper," n.d., http://usenix.org/events/fast03/tech/full_papers/anderson/anderson_html/.
13. "APPLE Technical Note TN1150: HFS Plus Volume Format," n.d., http://developer.apple.com/mac/library/technotes/tn/tn1150.html#BTrees.
14. Diedrich Olivier, "The Btrfs file system - The H Open Source: News and Features," 17, , http://www.h-online.com/open/features/The-Btrfs-file-system-746597.html.
15. Boaz Harrosh and Benny Halevy, "The Linux Exofs Object-based pNFS MetaData Server," 10, 2009, http://www.open-osd.org/bin/viewfile/Main/WebHome?rev=2;filename=exofs-pnfs-mds-design-2009-10-15.html.
16. "ubidesign.pdf," n.d., http://www.linux-mtd.infradead.org/doc/ubidesign/ubidesign.pdf.
17. "relatime as default since 2.6.30" http://valhenson.livejournal.com/36519.htmlhttp://kernelnewbies.org/Linux_2_6_33
18. "XFS Wiki," n.d., http://xfs.org/index.php/Main_Page.
19. " Introduction to ZFS" by Tom Haynes Adding Full-Text Filesystem Search to Linux by Stefan Büttcher and Charles L.A. Clarke
20. "ZFS (Community Group zfs.WebHome) - XWiki," n.d., http://hub.opensolaris.org/bin/view/Community+Group+zfs/.

Usenix :
HDFS Scalability: The Limits to Growth by K. Shvachko http://www.usenix.org/publications/login/2010-04/openpdfs/shvachko.pdf
XFS: The Big Storage File System for Linux by C. Hellwig http://www.usenix.org/publications/login/2009-10/pdfs/hellwig.pdf
Data Corruption in the Storage Stack: A Closer Look by L. Bairavasundaramhttp://www.usenix.org/publications/login/2008-06/openpdfs/bairavasundaram.pdf
The Present and Future of SAN/NAS: Interview D. Hitz, B. Pawlowsky by M. Seltzer http://www.usenix.org/publications/login/2008-06/openpdfs/netapp.pdf
A Brief History of the BSD Fast File System by Marshall K. McKusick http://www.usenix.org/publications/login/2007-06/openpdfs/mckusick.pdf
Porting the Solaris ZFS File System to the FreeBSD Operating System by P. Jakub Dawidek and Marshall Kirk McKusick
Ext4: The Next Generation of the Ext3 File System by A. Mathur, M. Cao, and A. Dilger http://www.usenix.org/publications/login/2007-06/openpdfs/mathur.pdf
A Comparison of Disk Drives for Enterprise Computing by K. Chan http://www.usenix.org/publications/login/2006-06/openpdfs/chan.pdf
Disks from the Perspective of a File System by M. McKusick http://www.usenix.org/publications/login/2006-06/openpdfs/mccusick.pdf