http://membres-liglab.imag.fr/donsez/cours

Supports Physiques de Stockage

Didier DONSEZ

Université Joseph Fourier (Grenoble 1)

Polytech'Grenoble LIG/ADELE

Didier.Donsez@imag.fr

Didier.Donsez@ieee.fr

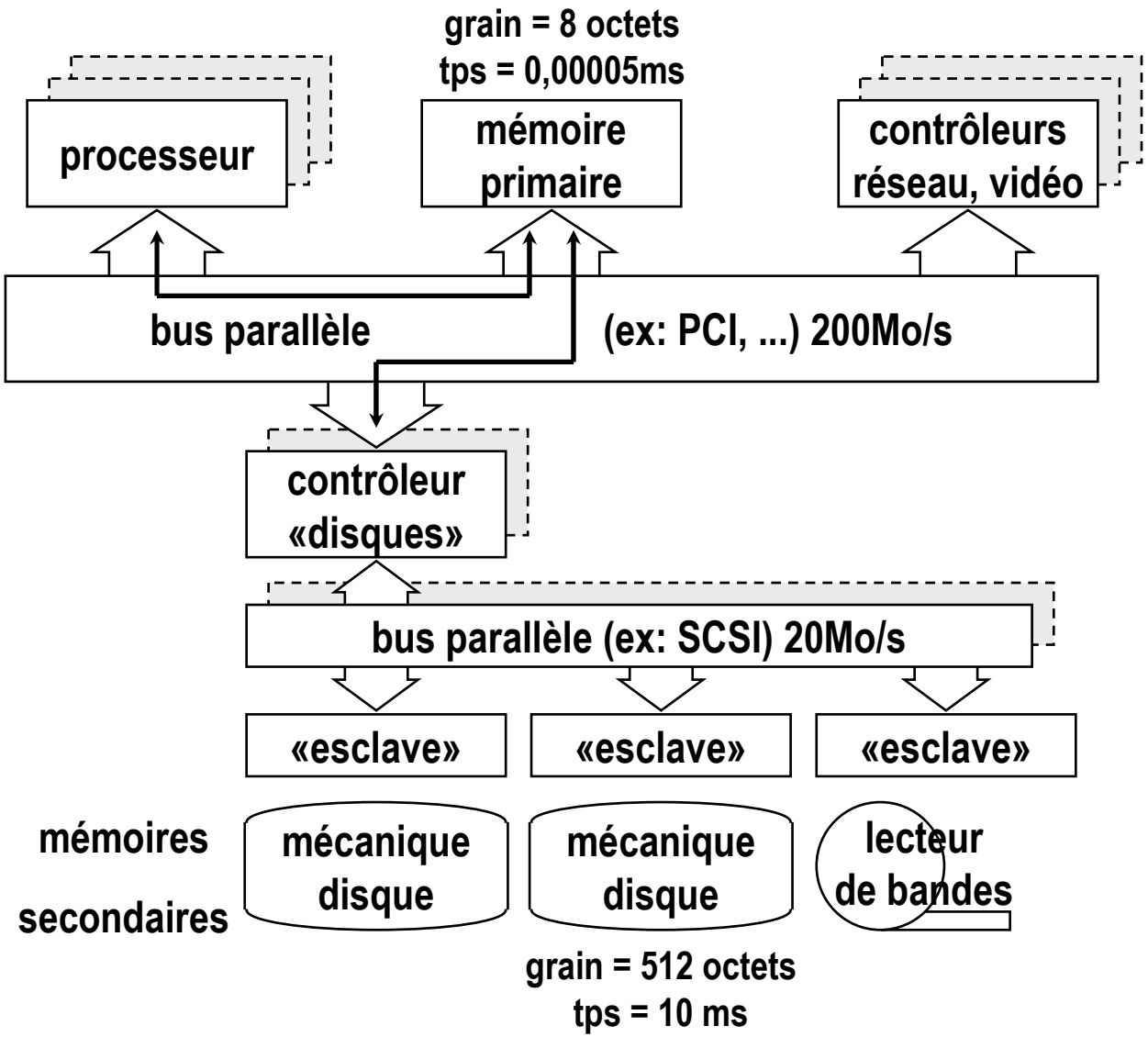
Motivations

- Propriété de persistance
  - à garantir
  - dans les systèmes de fichiers
  - dans les SGBDs
  - …
  - en cas de problème
    - Perte de l ’alimentation électrique, ...

Mémoires Persistantes vs Mémoires Volatiles

- Mémoire Primaire
  - zone de stockage des données
    en vue d'un calcul dans le processeur.
- Mémoire Secondaire
  - zone secondaire de stockage des données
    - Généralement de très grande capacité, résistant aux pertes de courant
- Mémoire Volatile
  - les données disparaissent en cas de perte de courant
- Mémoire Persistante ou Non Volatile
  - Une fois écrites, les données sont conservées

Architecture d'un ordinateur



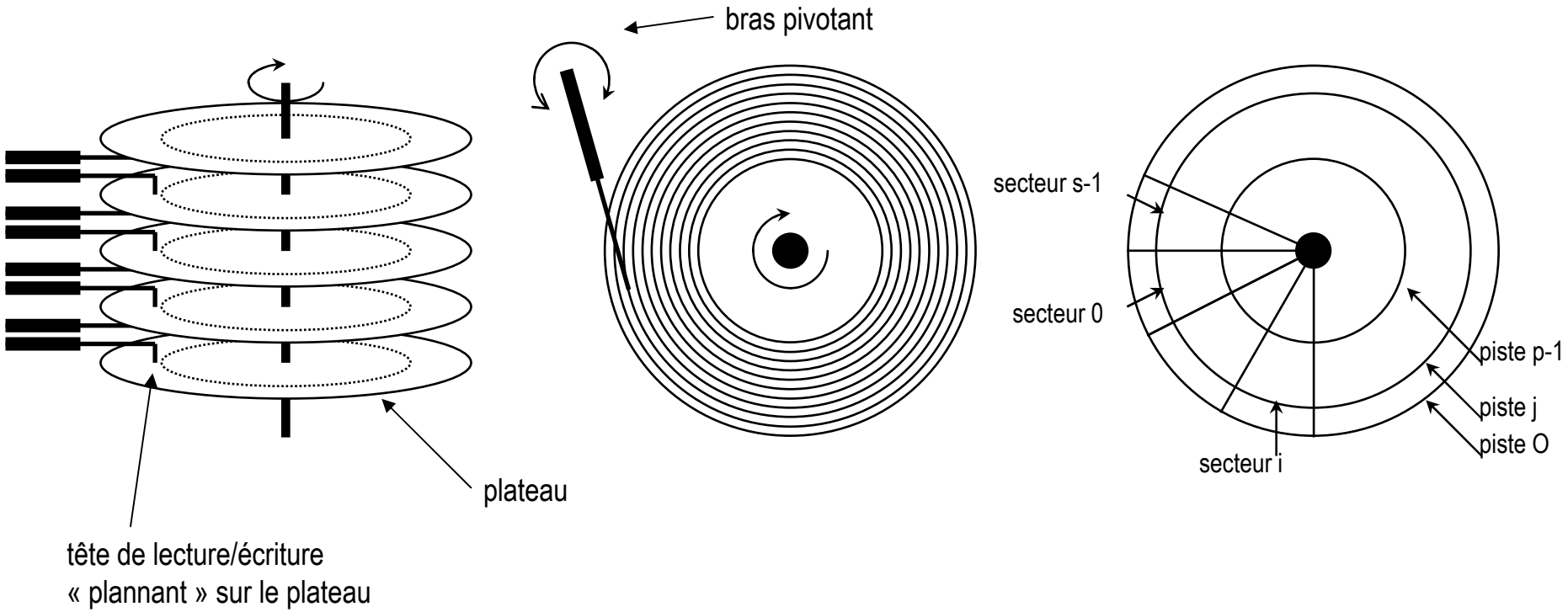
Mémoire volatile

- RAM (Static) RAM
  - On-Chip Cache
  - Processeur 8 et 16bits (et carte à puce)
- Dynamic RAM
    - Module
  - Acrononyme Techno bits Mhz Go/s
  - DRAM (A)synchronous
  - SDRAM Synchronous 64b 133*
  - RDRAM RamBus 16b 800 1.6
  - DDR DRAM Double Data Rate 64b 266 2.1

  - *100, 133, 266, 333 et 400 MHz (2005)

Disques Mécaniques (i)

- Organisation
  - blocs de mémoire de taille fixe (N octets)
    disque organisé en Tête, Cylindre, Secteur



Disques Mécaniques (ii)

- Accès Aléatoire aux blocs
  - Temps d'accès TA =
    Temps de Déplacement du bras (Seek Time)
    (positionnement sur la piste= 1-10 ms pour un HD)
  - + Temps de Latence (Rotational Latency)
    (positionnement sur le début du secteur)
    + Temps de Transfert (Data Transfer Time)
    (parcours du secteur)
  - Temps d'accès moyen
- Facteur de Blocage (blocking factor)
  - accès mémoire primaire : par bloc de 8 octets
  - accès mémoire secondaire : par bloc de N octets (1 secteur)
    N=512-1024
- MTBF (Mean Time Before Failure)
  - temps moyen avant une panne (500000-1000000 heures pour les HD)

Disques Mécaniques (iii)

- Type
  - Amovible : Le support peut être retiré du lecteur
  - Solidaire: Le support et le lecteur sont solidaires
- Disquettes (Floppy Disk)
  - 1 plateau de plastique souple recouvert de particules magnétiques
  - Mécanique basse vitesse, Amovible
    - Usage: PC mais en voie d' extinction
    - Remarque : comment faire pour reprendre des documents archivés depuis 10 ans ?

Disques Mécaniques (iv)

- Disques Durs (Hard Disk)
  - Plusieurs plateaux d'aluminium recouverts de particules magnétiques. Les têtes sont très proches des plateaux
  - Mécanique à haute vitesse (5000 à 15000 tours / minute)
  - Facteur encombrement (01/2004)
    - 3 ½ : <300 Go, 0.8 Go/cm2, 1.2€/Go en ATA
    - 2 ½ : <80 Go, 1.45 Go/cm2, 4€/Go en ATA
    - Mais aussi 1.8 pouce, 1 pouce et même 0.85 pouce
  - Disques et Lecteur solidaires
    - Usage : PC, Serveurs, Mass Market (Apple iPod™, …)
    - Remarque:7% de DD dans le Mass Market en 2003, en augmentation en 2004
      - (Magnétoscope numérique, console de jeux, récepteur iTV, \

Disques Mécaniques (v)

- Disques Durs (Hard Disk) suite
  - Autres critères
    - Résistance aux accélérations (chocs, …)
    - Bruit en fonctionnement (critère Windows XP Media Center)
    - …
  - Acteurs: Western Digital, Maxtor, Hitachi, Toshiba, …

Disques Mécaniques (vi)

- CD-ROM(Compact Disk) , DVD (Digital Versatile Disk)
  - 1 plateau de plastique recouvert de particules réfléchissantes
  - La tête de lecture est un couple faisceau laser et cellule et l'écriture est unique (Gravage) ou multiple (CD-RW)
    - Usage : PC, Mass Market
- CD Opto Magnétique
  - Un champ magnétique permet d'orienter les particules réfléchissantes à l'écriture

Disques Mécaniques (vii)

- Juke-box ou Robot
  - archive « ternaire» : 3 niveaux de mémoire (DRAM/Disques Durs/Bandes ou CD) fonctionnant en cache
  - Pre-fetching (pré-chargement), …
  - Remarque : Imprimante parfois intégrée pour les robots graveur (CD/DVD)

Mémoire « électronique » non-volatile

- Principaux avantages : compacité et résistance aux chocs
- Inconvénients : faible ratio capacité/prix

- UPS RAM
  - RAM alimentée par une pile (batterie)
  - Usage : Serveur
    - (Format : Banc de DRAM alimentée et batterie sur carte fille d' un serveur)
    - rendre les écritures de journalisation asynchrones
    - Performances: 90 % pour les SGF (serveur NFS)
  - Usage : Embarqué (TINI, iButton)
    - Temps d'accès uniforme, évite le risque de crash en cas d'écriture en FlashRAM
- EEPROM
  - Cible : Carte à puce, carte mère (Setup BIOS, …)
  - Nombre limité d'écritures (10^5)

Mémoire « électronique » non-volatile

- FlashRAM (NAND)
  - Cibles nomades, mobiles ou embarqués
    - Carte à puce, PC industriel embarqué, PDA, Téléphone portable (handset), Console de Jeux, Appareil Photo Numérique, Lecteur vidéo portable, STB, Passerelle domotique, Récepteur GPS, Voiture, Lecteur MP3, « Clé » USB, …
  - Grande capacité
    mais temps d'écriture lent et grain d'écriture >> grain lecture
  - Nombre de réécritures limitées (1000000 par point mémoire)
    - Inconvenients pour les swap et les journaux (log)
    - algorithmes de « wear levelling » (étalement de l'usure)
  - Plusieurs « form factors » et interfaces
    - SmartMedia, Compact Flash, PCMCIA, Memory Stick, SD
    - USB Memory Key, Disque 2,5 pouce SSD …
- *Magnetoresistive RAM*
  - Très Grande capacité par unité de surface
  - Encore en laboratoire (à surveiller)

Contrôleur Disque

- Pilote une ou plusieurs unités de stockage reliées par
  - un bus spécialisé parallèle (SCSI, ATA)
  - un bus série (USB, FireWire, SATA,…)
  - un réseau IP (iSCSI sur GigaEthernet)
- Requête Synchrone / Requête Asynchrone
  - un contrôleur asynchrone est capable d'émettre plusieurs requêtes vers les périphériques différents sans attendre la réponse d'une requête précédente.
    - Gain de performance
- Hot Plug
  - Possibilité d'ajouter/de retirer un disque « à chaud » (c.a.d. sans interrompre le fonctionnement des autres disques)

Contrôleur Disque sur Bus parallèle (i)

- IDE/ATA (Advanced Technology Attachment)
  - ATA 8,3 Mo/s
  - ATA-2 16,6 Mo/s
  - Ultra DMA : 33,3 Mo/s
  - Ultra-DMA2: 66,6Mo/s
  - 100, 133 Mo/s
- Synchrone, Non Hot-Plug
- Destiné les PC et stations d'entrée de gamme
- Domine le marché
  - 172 millions de disques ATA vendus en 2001

Contrôleur Disque sur Bus parallèle (i)

- SCSI (Small Computer System Interface)
  - SCSI: 5 Mo/s (Macintosh 1984)
  - SCSI2: 10 Mo/s
  - Wide SCSI: 20 Mo/s
  - Ultra Wide SCSI: 40 Mo/s (sur 3 mètres)
  - Ultra2 Wide SCSI: 80 Mo/s (sur 12 mètres)
  - Ultra3 Wide SCSI: 160 Mo/s
- Asynchrone, HotPlug
- Disques hautes performances (swap)
- Niche de marché: serveurs, RAID, SAN

Contrôleur Disque : Les tendances

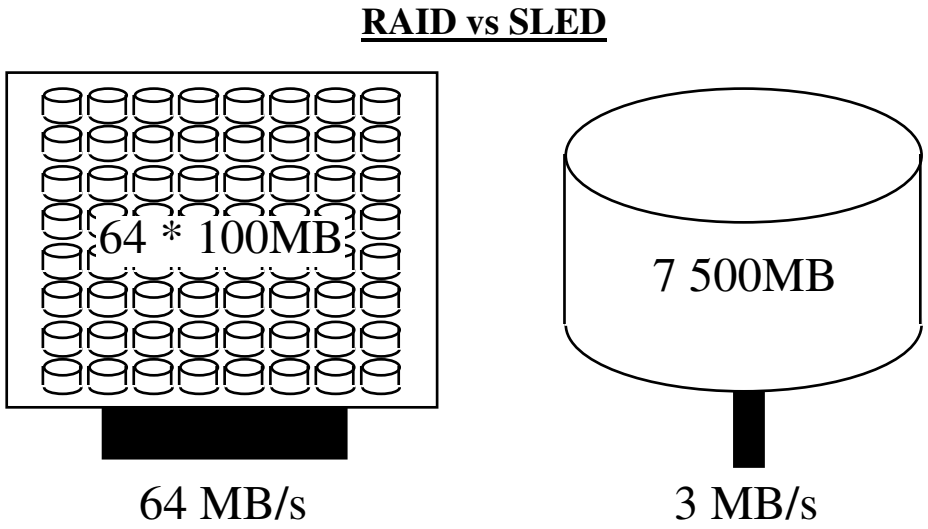
- SATA (Serial ATA) http://www.serialata.org
  - ATA sur bus série
    - Connectique simplifiée (pas de natte parallèle)
    - Allonge la distance entre hôte et équipemement
  - Débits: 150Mbps (2001) à 300 Mbps (2007)
  - Mais non hot plug
- eSATA (external SATA)
  - hot plug, disque externe
- Serial Attached SCSI http://www.scsita.org
  - SCSI sur bus série
- iSCSI (internet SCSI)
  - Trames SCSI tunnellé sur TCP/IP
    - Les unités de disques et les clients sont reliés par du GigaEthernet par exemple

Bus IO hautes performances

- InfiniBand (www.infinibandta.com)
  - Serveurs Haut de Gamme
  - Réseau commuté jusqu'à 64000 nœuds (hôtes et périphériques)
    - 2.5, 10, 30 Gbits/s
    - 17 m sur cuivre à 10 Kms sur fibre mono-mode
  - Adressage IPv6
  - Interfaces
    - HCA (Host Channel Adapter)
    - TCA (Target Channel Adapter)
- RapidIO (www.rapidio.org)
  - 256 à 65536 périphériques
  - 8 à 64 Gbits/s
  - Serveurs embarqués (switch TelCo, …)

RAID Redundant Array of Inexpensive Disks [Patterson 88]

- En 1988
- SLED : Single Large Expensive Disk



- RAID : Redundant Array of Inexpensive Disks
  - tableau de disques peu coûteux pour améliorer les débits I/O
  - cependant il faut introduit de la redondance
    à cause de la MTBF qui suit une loi de poisson

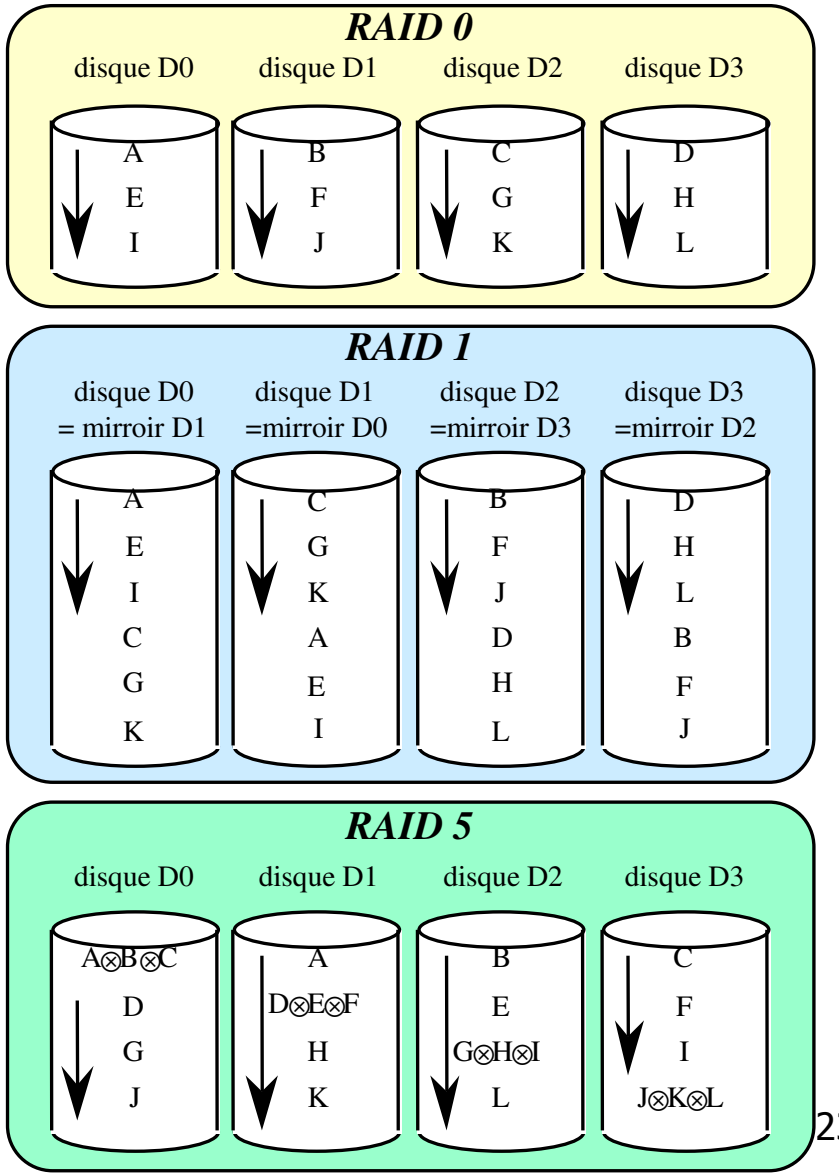
Niveaux de fonctionnement des RAID

- Redondance
  - Niveau 1 : Les Disques Mirroirs
    - TANDEM Mirrored Disks
  - Niveau 2 : Code de Hamming pour ECC
    - CM2 DataVault
  - Niveau 3 : Un Seul Disque de Contrôle par Groupe de Disque
  - Niveau 4 : Lectures et Ecritures indépendantes
  - Niveau 5 : Contrôle réparti sur les disques du Groupe

- Sans redondance
  - Niveau 0 : Stripping
    - répartition des blocs contigus d'un fichier entre les disques mais pas de redondance (ie AID)

Redondance et Balayage (scan) parallèle

- RAID 0
  - répartit les données sur plusieurs disques pour améliorer les performances.
- RAID 1
  - effectue des copies miroirs de disques pour survivre aux pannes.
- RAID 5
  - utilise la correction d'erreur et la répartition des données pour fournir performance et sécurité de manière économique et efficace.
  - La correction est basée sur la propriété du XOR (⊗) :
    (X ⊗Y) ⊗ Y = X



Nouveaux Niveaux de fonctionnement des RAID

- Redondance
  - Niveau 0/1 ou « 10 » : Stripping sur des disques miroirs
  - Niveau 0/5 ou « 50 » : Stripping sur des groupes 5
  - Niveau 6 : Contrôle redondant
    - Remarque
      - 2,3,4,5 ne résiste qu'à la panne d'un seul disque dans le groupe
    - Le niveau 6 ajoute du contrôle pour la redondance
    - Plusieurs propositions
      - N+Q : autre fonction que XOR
      - 2D
        - Soit une matrice de disques
        - les lignes forment des groupes RAID5
        - les colonnes forment des groupes RAID5

Contrôleur RAID

- Caractéristiques
  - Niveaux de RAID géré 0, 1, 5, 10 (0/1)
  - Hot Sparing : Secours permutable (disques de réserve)
  - Hot Swapping : Echange à chaud (par un opérateur humain)
- Contrôleur RAID
  - Disques SCSI surtout et maintenant ATA
  - De + en + intégré sur la carte mère des PCs
- Interfaces
  - Low Cost : PCI + Disques SCSI
  - UW-SCSI
  - FiberChannel (25 Mo/s)
    pour des Architectures en Clusters
- Exemple de Configuration (2002)
  - disques SCSI de 80 Go 15000 Tr/min
    (organisés par groupe de 5 pour RAID 5)

Contrôleur Disque sur bus série

- Principaux avantages
  - intégré à de nombreux ordinateurs (enfouis : TV, Camera, …)
  - Connectique hotplug, faible coût et « longue » distance
- USB
- FireWire
- FibreChannel
- IrDA
  - sans fil (faisceau infrarouge)
  - débit jusqu ’à 100 Ko/s

Contrôleur Disque sur bus série *USB (Universal Serial Bus)*

- Chaîne de périphériques (jusqu'à 127)
- Hot Plug-and-Play (matériel et logiciel système)
  - 1,5 Mb/s et 12 Mb/s (1.1), 480 Mb/s (2.0)
- Exemple de Périphériques :
  - USB1.x : ZIP, LS120, système de fichiers de balladeur MP3, appareil photo, memory key, scanner, webcam, Pointage (Souris, Joystick …), Bluetooth …
  - USB2.0 : imprimante, disque dur externe, graveur externe, lecteur DVD, Scanner, WiFi, Appareil photo numérique,, Memory Key, …
- Parc
  - 1 Milliard en 2005, 3,5 Milliard en 2006
  - 95% des dispositifs de stockage, 95% des webcam, 90% des appareils de photo numérique, …
- La suite : WUSB (Wireless USB)
  - S'appuiera sur la couche UWB (Ultra Wide Band)
  - 480 Mbits/s à 3 mètres, 110 Mbits/s à 10 mètres

Contrôleur Disque sur bus série Firewire IEEE 1394

- appelé également iLink (Apple), ou DV
- Domaine de la vidéo numérique surtout
- Bus série jusqu'à 63 périphériques
- Isochrone (temps réel)
- Support plusieurs débits simultanés: 100, 200, 400, 800 Mbits/s
- Remarque
  - Disques externes FireWire pour caméscopes numériques
    - 1Go produit par un caméscope numérique pour 5 minutes de vidéo !
- http://www.1394ta.org

Contrôleur Disque sur bus série

- FibreChannel http://www.fibrechannel.org

Connexions Série (ii)

- Les « plug-and-participate »
  - Simplification de la connexion domotique (home network environnement)
  - HAVi Home Audio/Video Interoperability – www.havi.org
    - GUNDIG A.G., Hitachi, Ltd., Matsushita Electric Industrial Co., Ltd. (Panasonic), Philips Electronics N.V., Sharp Corporation, Sony Corporation, Thomson Multimedia S.A., Toshiba Corporation.)
  - Bluetooth
    - (IBM Corporation, Intel Corporation, Nokia Corporation, Telefon AB L.M. Ericsson, Toshiba Corporation.)
  - JetSend™ (HP)
  - UPnP
    - Universal Play And Play

Pérennité du stockage

- Durée de vie des données
  - 1 an, 10 ans, 50 ans, 100 ans, 1000 ans, 10000 ans, …
  - Cf perte des dossiers « magnétiques » du FBI
- Différents facteurs
  - Support (Média)
    - *Papier (carte perforée)*
    - Magnétique, Optique, Opto-Magnétique …
    - Plastique, Métal, Silicium
  - Lecteurs
    - Matériel
      - exercice : retrouver un lecteur disquette 5 ¼ ou 8 et bientôt 3 ½
    - Pilotes pour OS (driver pour lecteur SyQuest sur WinXP)

Pérennité du stockage

A lire : Gordon Bell and Jim Gray, Digital Immortality, 1 October 2000, Technical Report, MSR-TR-2000-101, Microsoft Research

Data-types	Rate (Bytes/hour)	Per day / per 3 year	Lifetime amount
read text, few pictures	200 KB	2 –10 MB / GB	60-300 GB
Email, papers, written text		0.5 MB / GB	15 GB
photos w/voice @100KB	200 K	2 MB / GB	60 GB
photos @200 KB	Ten images/day	2 MB / GB	150 GB
spoken text @120wpm	43 K	0.5 MB / GB	15 GB
spoken text @8Kbps	3.6M	40 MB / GB	1.2 TB
music or high quality sound	60 M	60 MB / GB	5.0 TB
video-lite 50Kb/s POTS	22 M	0.25 GB/TB	25 TB
video 200Kb/s *VHS-lite*	90 M	1 GB/TB	100 TB
DVD video 4.3Mb/s	1.8 G	20 GB/TB	1 PB

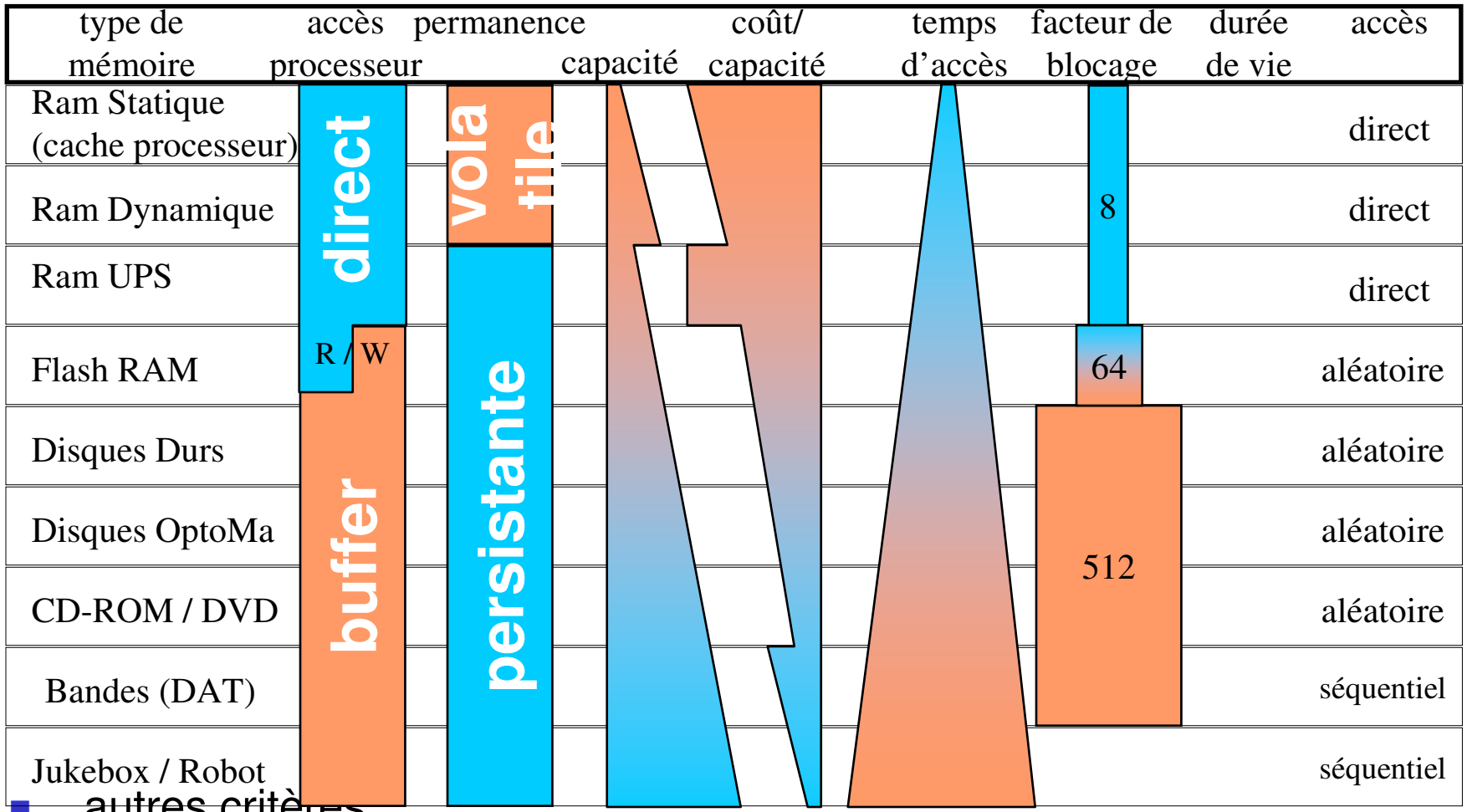
Le futur des disques

- Le contrôleur esclave du disque
  est un processeur de plus en plus puissant
  avec de plus en plus de RAM (Cache)

- Fonctions de recherche sur le disque
  - Plein texte, structuration en enregistrement (BDR), …

- Fonctions de filtrage de données sur le disque

Récapitulons : Mémoires Persistantes vs Mémoires Volatiles



- autres critères :
  - l'encombrement et la résistance aux chocs/températures/bombardement X…

Bibliographie

- John M May, « Parallel io for high perf computing », ed Morgan Kaufmann, 1-55860-664-5, 2001
- Jacques Peping, " Solutions de stockage ", Ed Eyrolles & CMP - 01/1999, 320 pages, ISBN: 2-212-09057-9
  - problématique du stockage des données d'entreprise et aborde les technologies des périphériques et l'évolution des modèles d'architectures de stockage. QA 76.9 A73 PEP
- Chris Date, "Introduction aux Bases de Données", 6ème édition, Ed Intl Thomson Publ. ISBN 2-84180-964-1
  - Voir le chapitre (6ème édition)
- Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom, "Database System Implementation", 2000, Ed Prentice Hall, ISBN 0-13-040264-8

Bibliographie

- A lire absolument
  - J. Gray, G. Graefe, “The 5 minute rule, ten years later,” SIGMOD Record 26(4): 63-68, 1997
    - http://www.sigmod.org
    - http://research.microsoft.com/~gray/5_min_rule_SIGMOD.doc

Bibliographie

- W.Ng. Spencer, "Advances in Disk Technology: Performance Issues", IEEE Computer, May 1998,Vol 31 No 5, pp75-81.
  - Des grandeurs pour les DD
- G. Lawton, "Storage Technology Takes Center Stage", IEEE Computer, November 1999,Vol 32 No 11, pp10-12,16.
  - Les tendances 1999 dans le stockage
- "Storage Systems", Special Issue, IEEE Computer, December 2002.
  - Les tendances 2002 dans le stockage
- « Storage », ACM Queue, Volume 1, Issue 4, June 2003
- E. Grochowski and R. D. Halem, Technological Impact of Magnetic Hard Disk Drives on Storage Systems, IBM Systems Journal 42 (2003), no. 2, 338–346.