HOWTO: Les systèmes multi-disques

Stein Gjoen, sgjoen (at) nyx.net
Traduit en français par Patrick Loiseleur loisel (at) lri.fr v0.19b, 10 septembre 1998

Ce document explique comment utiliser au mieux plusieurs disques et partitions avec Linux. Bien qu'une partie de ce texte soit spécifique à Linux, il peut aussi d'appliquer à d'autres systèmes d'exploitation multi-tâches, étant donnée l'approche générale adoptée ici.

Contents

1 Introduction

Cette version a pour nom de code *The newer Generation*

De nouveaux noms de code apparaîtront à mesure des changements pour indiquer l'état du document.

J'ai écrit ce document principalement parce que j'ai hérité de trois vieux disques SCSI pour mettre en place mon système Linux et je voulais savoir comment utiliser au mieux les capacités de parallélisme d'un système SCSI. Et puis j'ai entendu dire qu'il y avait un prix pour les gens qui écrivent des documents...

Ce HOWTO est à lire en parallèle avec le Linux Filesystem Structure Standard (FSSTND): en aucun cas il ne le remplace mais il explique où mettre physiquement les répertoires détaillés dans le FSSTND en terme de disques, partitions, types, RAID, file system (fs), taille physique et autres. Cela aussi bien pour une machine Linux personnelle que pour un gros serveur Internet.

La suite du FSSTND s'appelle Filesystem Hierarchy Standard (FHS) et couvre plus que Linux. FHS 2.0 est sorti mais certains détails restent à préciser et il faudra un certain temps avant que ce nouveau standard ait un impact sur les nouvelles distributions. Le FHS n'est encore utilisé dans aucune distribution, mais Debian a annoncé son intention de s'y conformer à partir de la version 2.1.

(NdT: le FSSTND a été traduit en français et est disponible à l'adresse suivante: `http://www.freenix.fr/linux/fsstnd-fr/`

et la traduction du FHS 2.0 est dans `ftp://lirftp.insa-rouen.fr/pub/linux/french/docs/`

)

Il est également conseillé de lire le Guide d'Installation de Linux et si vous utilisez un PC, ce qui doit encore être le cas de la majorité, vous pouvez trouver des informations plus précises dans les FAQs du forum comp.sys.ibm.pc.hardware.

C'est aussi une expérience pour moi d'écrire ce HOWTO et j'espère qu'il évoluera pour devenir plus détaillé et peut-être même plus correct.

Avant tout quelques rappels légaux. L'actualité a montré combien c'est important.

1.1 Copyright

1.2 Avertissement

Je décline toute responsabilité au sujet du contenu de ce HOWTO. Utilisez les concepts, les exemples et les trucs à vos risques et périls.

Les marques citées dans ce document sont déposées par leurs propriétaires respectifs.

Enfin, vous êtes expressément invités à faire une sauvegarde de tout votre système avant tout grand changement et à intervalles réguliers.

1.3 Nouveautés

Ce HOWTO a maintenant un index et utilise les SGMLtools 1.0.5. Il ne sera donc pas formaté correctement avec une version antérieure.

La nouveauté la plus récente est la section sur le formatage d'un disque unique, étant donné que les disques de 8 Go deviennent abordables. On donne aussi des exemples de configuration RAID avancées. Les gens s'intéressent de plus en plus au VFAT32 et il y a des additions concernant ce système de fichiers.

Le FHS 2.0 est sorti mais aucune distribution ne s'y conforme: lorsque cela arrivera, ce HOWTO changera un peu. Pour l'instant il suit le FSSNTD.

A propos de ce HOWTO justement, j'ai enlevé le préfixe "mini" qui commençait à devenir comique vu sa taille. En fait ce document est si gros que j'ai dû inclure un plan de lecture comme certains lecteurs me l'ont demandé.

Un ajout récent est la section sur la meilleure manière d'obtenir de l'aide face à un problème que vous n'arrivez pas à résoudre, ainsi que d'autres suggestions pour la maintenance. Cette section migrera bientôt vers un autre HOWTO.

A cause des quantités de Spams j'ai dû truquer toutes les adresses électroniques de ce document pour échapper aux robots des spammeurs qui scannent Internet à la recherche d'adresses à rajouter dans leurs listes. Pour m'écrire il faut remplacer les `(at)` par le symbole `@`

Un certain nombre de pointeurs vers des mailing lists ont été ajoutés.

Depuis la version 0.14 il y a eu trop de changements pour les énumérer ici. J'ai reçu beaucoup de remarques et un patch important de kris (at) koentopp.de qui ajoutait de nombreux détails. En fait ce document a grandi au-delà de mes prévisions.

Je suis aussi passé en Debian 1.3 et j'ai remplacé les valeurs d'espace disque de ma vielle Slackware en conséquence. J'utiliserai la Debian comme base pour les discussions et les exemples, mais ce HOWTO s'applique aussi bien à d'autres distributions ou à d'autres systèmes d'exploitation. Au moment où j'écris la Debian 2.0 est sortie en version bêta et elle sera utilisée pour les versions futures de ce document.

Les nouveaux systèmes de fichiers, journalisés, à héritage, ou optimisés pour fichiers à taille variable (comme les fichiers de log) bénéficient d'un nouvel intérêt dans les forums de comp.os.linux. Restez à l'écoute pour les mises à jour. Le vieux programme de défragmentation pour `ext2fs` est en cure de rajeunissement et il y a toujours du travail sur la compression.

La dernière version (en anglais) de de ce document peut être connue avec la commande

finger `finger:sgjoen@nox.nyx.net` sur mon compte Nyx.

On la trouve aussi sur ma page Web:

The Multi Disk System Tuning HOWTO Homepage `http://www.nyx.net/~sgjoen/disk.html` .

La dernière version traduite en français est sur

Freenix `http://www.freenix.org/` .

Ce HOWTO est disponible en plusieurs formats: SGML, HTML, PostScript ou texte simple.

La traduction française que vous lisez est due à Patrick Loiseleur (courrier: loisel (at) lri.fr) et c'est à lui qu'il faut envoyer commentaires, remarques sur la traduction elle-même.

1.4 Remerciements

J'ai le plaisir de remercier les personnes suivantes qui ont contribué à ce HOWTO:

```
ronnej (at ) ucs.orst.edu
cm (at) kukuruz.ping.at
armbru (at) pond.sub.org
R.P.Blake (at) open.ac.uk
neuffer (at) goofy.zdv.Uni-Mainz.de
sjmudd (at) redestb.es
nat (at) nataa.fr.eu.org
sundbyk (at) horten.geco-prakla.slb.com
gjoen (at) sn.no
mike (at) i-Connect.Net
roth (at) uiuc.edu
phall (at) ilap.com
szaka (at) mirror.cc.u-szeged.hu
CMckeon (at) swcp.com
kris (at) koentopp.de
edick (at) idcomm.com
pot (at) fly.cnuce.cnr.it
earl (at) sbox.tu-graz.ac.at
ebacon (at) oanet.com
vax (at) linkdead.paranoia.com
tschenk (at) theoffice.net
pjfarley (at) dorsai.org
jean (at) stat.ubc.ca
johnf (at) whitsunday.net.au
```

Des remerciements spéciaux vont à `nakano (at) apm.seikei.ac.jp` pour avoir fait *la traduction japonaise* `http://jf.linux.or.jp/JF/JF-ftp/other-formats/Disk-HOWTO/html/Disk-HOWTO.html` , contribué au document et donné un exemple de serveur en milieu académique qui est inclus à la fin de ce document.

Si j'ai oublié quelqu'un, faites-le moi savoir. Ils ne sont pas si nombreux, donc lisez attentivement ce document, contribuez à son élaboration et rejoignez l'élite !

Un nouveauté dans ce document est un appendice avec quelques tables que vous pouvez remplir pour simplifier l'élaboration.

Tous commentaires et suggestions (en anglais !) peuvent êtres envoyés à mon adresse: *sgjoen@nyx.net* .

Et maintenant, allons-y !

2 Organisation de ce document

Les HOWTOS sont plus des documents pédagogiques que des manuels de référence. On présentera donc les choses plutôt comme des des problèmes à résoudre et leurs solutions que comme un cours sur la structure des disques durs. Cependant une introduction sur la manière dont un disque dur fonctionne est indispensable.

2.1 Structure logique

Elle est basée sur un empilement de couches avec au sommet le système de fichiers tel que les applications l'utilisent et tout en bas la couche physique.

```
-------------------------------------------------------------
|__     Fichiers, répertoires   ( /usr /tmp etc)          __|
|__     Système de fichiers     (ext2fs, vfat etc)        __|
|__     Gestion du volume       (AFS)                     __|
|__     RAID, concaténation     (md)                      __|
|__     Pilote de périphérique  (SCSI, IDE etc)           __|
|__     Contrôleur              (chipset, carte)          __|
|__     Connection              (cable, réseau)           __|
|__     Disque                  (magnétique, optique etc) __|
-------------------------------------------------------------
```

Dans le diagramme ci-dessus la gestion de volume, le mode RAID et la concaténation sont optionnels. Les trois derniers niveaux sont matériels et les autres logiciels. Chaque niveau sera amplement détaillé ci-dessous.

2.2 Structure du document

La plupart des utilisateurs partent avec un certain matériel et ont des une certaine idée de ce qu'ils veulent faire et de la taille de leur système. Ce sera mon plan: nous parlerons d'abord du matériel, puis des contraintes de mise en place et je détaillerai ma façon de faire. Elle a bien marché chez moi aussi bien que pour des serveurs réseau au travail ou en milieu académique comme me l'a rapporté mon collègue japonais.

Enfin je donnerai certaines tables de valeurs destinées à vous guider dans la mise en place de votre machine. Comme je l'ai déjà dit, tous les commentaires sont les bienvenus.

2.3 Plan de lecture

Bien que n'étant pas le plus gros ce HOWTO est déjà bien gros et on m'a demandé un plan de lecture pour permettre de le lire en diagonale. CHoisissez selon votre niveau:

Expert

Si vous connaissez bien Linux et les technologies des disques durs, consultez seulement les tables en appendice. Eventuellement vous pouvez lire les Questions/Réponses et le chapitre 16 (Bric à Brac)

Expérimenté

Si vous connaissez bien les ordinateurs allez directement au chapitre 4.3 (technologies) et poursuivez.

Débutant

Désolé. Vous devrez tout lire. En plus je vous recommande les autres HOWTOs concernant les disques.

3 Technologies de disques durs

Une discussion très complète sur les technologies des disques durs pour compatibles PC se trouve à:

The Enhanced IDE/Fast-ATA FAQ `http://thef-nym.sci.kun.nl/~pieterh/storage.html`

Elle est aussi régulièrement postée dans les forums Usenet. On ne présentera ici que ce qui est indispensable à la compréhension de la suite.

3.1 Disque

C'est l'appareil où vos données sont physiquement enregistrées, et bien que le système d'exploitation peut les rendre similaires à l'usage, il en existe des types très différents. On ne parlera pas des disquettes, sauf dans une prochaine version si beaucoup de monde le réclame.

3.2 Géométrie

Un disque dur est constitué d'un ou plusieurs plateaux tournants qui contiennent des données lues et écrites par des capteurs. Les capteurs sont fixes les uns par rapport aux autres et les transferts de données ont donc lieu en même temps sur tout les plateaux, ce qui définit un cylindre de pistes. Le disque est aussi divisé en secteurs.

On spécifie la géométrie d'un disque avec trois nombres: le nombre de Cylindres, de Têtes et de Secteurs. En anglais CHS pour cylinders, heads, and sectors.

Il y un certain nombre de conversions entre:

- le CHS physique
- le CHS logique que le disque déclare au BIOS
- le CHS logique utilisé par le système d'exploitation

En pratique c'est une source de confusion importante. Voir le *Large Disk mini-HOWTO*

3.3 Média

La technologie du médium employé détermine des paramètres importants comme le taux de lecture/écriture, le temps moyen d'accès, la capacité et le fait d'être en lecture seule ou non.

3.3.1 Disques magnétiques

C'est le médium le plus courant pour la mémoire de masse. Habituellement c'est la technologie la plus rapide et elle est en lecture/écriture. Le plateau tourne avec une vitesse angulaire constante (CAV) avec une densité physique des secteurs variable. Le nombre de bits par unité de longueur reste constant tandis que le nombre de secteurs logiques par piste varie.

Des valeurs typiques de vitesse angulaire sont 4500 et 5400 tr/min, mais on trouve aussi 7200 et des disques à 10000 tr/min ont fait récemment leur apparition sur le marché. Le temps d'accès est d'environ 10 ms et les taux de transferts entre 4 et 40 Mo/s. Il faut se rappeler que les disques les plus rapides sont aussi ceux qui consomment le plus d'électricité et chauffent le plus. Voir 16.6 (Chaleur et Consommation d'Énergie) à ce sujet.

Notez bien qu'il y a plusieurs types de transferts qui sont mesurés avec des unités différentes. Le premier est le taux de transfert du plateau vers la mémoire cache du disque, mesuré en Mbit/s, qui vaut entre 50 et 250 Mb/s. Le second est entre le cache et l'adaptateur, il est mesuré en Moctets/s et vaut entre 3 et 40 Mo/s. (rappel: un octet = 1 B = 8 bits = 8 b)

3.3.2 Disques optiques

Des disques optiques en lecture/écriture existent mais ils sont lents et peu répandus. Ils étaient utilisés dans les machines NeXT mais très critiqués pour leur faible vitesse. Celle-ci est due à la nature thermique du changement de phase qui matérialise l'enregistrement de données. Même avec des lasers assez puissants, les changements de phase sont plus lents qu'avec un champ magnétique.

Les cédéroms aussi sont de technologie optique, mais comme leur nom (ROM = Read Only Memory) l'indique, ils sont en lecture seule. Leur capacité est d'environ 650 Mo, et le débit peut atteindre 1,5 Mo/s. Les données sont sur une seule piste en spirale, on ne peut donc pas vraiment parler de géométrie pour ces disques. La densité des données est constante donc le lecteur utilise une vitesse linéaire constante (CLV). Le temps d'accès est aussi plus lent, environ 100 ms, en partie à cause de la piste en spirale. Les lecteurs récents utilisent des vitesses angulaires constantes (CAV) à certains endroits du disque: cette technologie mixte CAV/CLV augmente le débit et réduit le temps d'accès car il y a moins besoin d'accélérer et de ralentir la vitesse angulaire (pour garder une vitesse linéaire constante).

Un nouveau type de disque semblable au cédérom (le DVD) permettra jusqu'à 18 Go de stockage.

3.3.3 Disques à états solides

Cette technologie récente est surtout utilisée dans les portables et les systèmes embarqués. Ne contenant aucune partie mobiles ils sont très rapides pour le taux de transfert comme pour le temps d'accès. Le type le plus courant est la mémoire vive "flashable" (flash-RAM) mais d'autre types de mémoire vive sont aussi utilisés. Il y a quelques années de grands espoirs se sont portés sur la mémoire à bulles magnétiques mais elle s'est avérée chère et pas pratique.

En général les disques de mémoire vive sont une mauvaise idée: mieux vaut mettre beaucoup de mémoire sur la carte mère et laisser le système d'exploitation la diviser en fichiers, cache, zone de programmes et de données. Les disques de mémoire vive sont utiles seulement pour des usages très spécifiques, comme des systèmes temps réel avec des délais très courts.

La mémoire flash est aujourd'hui disponible par dizaines de Mo et on pourrait être tenté de l'utiliser pour un stockage temporaire rapide des données. Mais il y a un os: on ne peut écrire sur de la mémoire flash qu'un nombre assez limité de fois. Mettre `swap`, `/tmp` ou `/var/tmp` sur un périphérique de ce genre réduirait drastiquement sa durée de vie. En revanche il peut être intéressant d'utiliser de la mémoire flash pour des données lues souvent et écrites peu souvent.

Pour augmenter la durée de vie il faudra des pilotes spéciaux qui minimisent le nombre de fois où on doit effacer un bloc mémoire.

Cet exemple montre bien l'intérêt qu'il y a à séparer l'arborescence des fichiers entre plusieurs périphériques.

Les lecteurs à état solide n'ont pas d'adressage pas cylindre/tête/secteur mais cette géométrie est simulée par le pilote: ainsi de l'extérieur ils se comportent exactement comme un disque dur.

3.4 Interfaces

Il y a une pléthore d'interfaces dans une gamme de prix très étendue. La plupart des cartes-mères comprennent une interface IDE ou mieux, la puce Triton d'Intel sur bus PCI qui est très répandue aujourd'hui. Beaucoup de cartes-mères ont aussi une puce d'interface SCSI fabriquée par Symbios (nouveau nom de NCR) et directement connectée au bus PCI. Vérifiez ce que vous avec et ce que le BIOS de voter carte-mère supporte.

3.4.1 MFM et RLL

Il fut un temps où c'était la technologie incontournable, un temps où 20 Mo c'était le bout du Monde. Ces interfaces dinausoresques sont d'un lenteur comique comparé à ce qui se fait aujourd'hui. Linux les supporte mais vous seriez bien avisé de vous demander ce que vous voulez mettre dessus. On peut bien sûr penser qu'une partition de secours avec un DOS potable dessus est toujours utile.

3.4.2 ESDI

En fait, ESDI est une adaptation de l'interface SMD, très utilisée sur les "gros" ordinateurs, avec le câblage de l'interface ST506, plus pratique que les 60 + 26 broches du connecteur SMD. L'interface ST506 était très nulle et dépendait complètement du contrôleur et du processeur pour faire les calculs de tête/cylindre/secteur et garder une trace de la position de la tête, etc. L'interface ST506 exigeait du contrôleur qu'il gère de façon détaillée les paramètres physique du lecteur et le formatage des pistes, bit par bit. Ce genre d'interface a vécu 10 ans si on compte les variantes MFM, RLL, ERLL et ARLL. ESDI, d'un autre côté, était "intelligente": le contrôleur avait souvent trois ou quatre puces pour un seul disque, et il y avait un langage de haut niveau pour formater une piste, rechercher et transférer des données. ESDI permettait d'utiliser une densité d'enregistrement variable, ou beaucoup d'autres choses. Bien que pas mal de techniques de ESDI aient été incorporées à IDE, c'est SCSI qui a progressivement détrôné ESDI.

3.4.3 IDE et ATA

Avec les progrès de la miniaturisation, les contrôleurs, autrefois sur une carte ISA, ont été intégrés au disque et IDE (Integrated Drive Electronics) était né. C'était simple, pas cher et assez rapide, si bien que les concepteurs du BIOS ont fixé une de ces limitations arbitraires dont l'informatique est pleine. Avec 16 têtes et 1024 secteurs, la capacité fut limitée à 504 Mo. Dans la plus pure tradition de l'industrie informatique, cette limitation a été ensuite contournée par des bidouilles infâmes dans le BIOS. En clair, vous devez lire très attentivement la documentation de votre BIOS pour savoir de quand il date et quelle taille de disque il autorise. Heureusement avec Linux vous pouvez spécifier directement au noyau (donc sans avoir besoin de passer par le BIOS) les paramètres (CHS) du disque. La documentation de Lilo et de Loadlin détaille comment le faire. IDE est synonyme d'ATA, AT Attachements. IDE utilise un programmes d'entrées-sorties (*PIO-mode*) très gourmand en temps de calcul qui monopolise le processeur principal. Le taux de transfert optimal (théorique) est de 8,3 Mo/s. IDE ne permet pas l'accès direct à la mémoire (DMA)

3.4.4 EIDE, Fast-ATA et ATA-2

Ces trois termes sont à peu près équivalents. fast-ATA et ATA-2 sont synonymes, mais EIDE comprend ATAPI. ATA-2 est ce qu'il y a de mieux actuellement, car plus rapide et autorisant l'accès direct à la mémoire (DMA). Le taux de transfert maximal est 16,6 Mo/s.

3.4.5 Ultra-ATA (ou Ultra-DMA)

Ce nouveau mode DMA est à peu près deux fois plus rapide que l'EIDE PIO-Mode 4. Deux disques avec et sans l'Ultra-DMA peuvent être mis sur la même nappe sans pénalité pour le plus rapide. L'interface Ultra-DMA est compatible au plus bas niveau (au niveau électrique) au Fast-ATA, y compris pour la longueur minimale des nappes.

3.4.6 ATAPI

ATAPI signifie *ATA Packet Interface* et a été conçu mettre des cédéroms sur une interface IDE. Comme l'IDE, il est simple et pas cher.

3.4.7 SCSI

SCSI signifie *Small Computer System Interface* et c'est une interface générique qu'on peut utiliser pour brancher des disques, des plateaux de disques, des imprimantes, des scanners, des graveurs de cédéroms, ... Le nom est mal choisi dans la mesure où c'est utilisé dans les PC haut de gamme et les stations. Elle convient aux environnements multi-tâche.

L'interface standard a 8 bits de large et peut gérer 8 périphériques. L'interface *wide-SCSI* a 16 bits de large (elle est donc deux fois plus rapide à la même fréquence) et peut gérer 16 périphériques. La carte SCSI est toujours comptée comme un périphérique, habituellement avec le numéro 7 (les autres étant numérotés de 0 à 6). Le SCSI 32 bits existe aussi mais il demande en général un ensemble de câbles doubles.

L'ancien standard faisait 5 Mo/s et le nouveau (*fast-SCSI*) 10 Mo/s. L'*ultra-SCSI*, connu aussi sous le nom de *fast 20*, réalise 20 Mo/s sur un bus 8 bits. Des voltages plus bas (LVD, pour *Low Voltage Differential*) permettent d'atteindre de plus grandes vitesses et d'utiliser des câbles plus longs.

Le SCSI est plus rapide, mais plus cher que l'(E)IDE. On ne saurait assez insister sur l'importance de la terminaison et la qualité des câbles. Les disques SCSI sont aussi en général de meilleure qualité que les disques IDE. Souvent on peut les brancher et les débrancher "à chaud" (sans couper l'alimentation), ce qui est surtout utile si on a plusieurs ordinateurs (pour pouvoir transporter les disques d'un ordinateur à un autre).

Parmi les documents à consulter sur le SCSI, le SCSI-HOWTO et la Foire Aux Questions (FAQ) SCSI sont vivement recommandés.

Un autre avantage du SCSI est qu'on peut connecter facilement des lecteurs de DAT pour sauvegarder des données, ainsi que certaines imprimantes ou scanners. Il est même possible de l'utiliser comme un réseau ultra-rapide entre ordinateurs qui partagent des périphérique SCSI. C'est cependant non-trivial en particulier pour assurer la cohérence de la mémoire tampon des deux cartes SCSI.

3.5 Câbles et nappes

Ce n'est pas un cours de hardware mais certaines informations sur les câbles sont nécessaires. Cette pièce si simple de l'équipement est souvent la cause de bien des problèmes. Aux vitesses actuelles il faut tenir compte de son impédance, et sans un minimum de précautions on risque des dysfonctionnement ou bien la panne complète. Certains adaptateurs SCSI sont plus sensibles que d'autres à la qualité des câbles.

Les câbles blindés sont bien sûr meilleurs (ils sont protégés des interférences électromagnétiques) mais beaucoup plus chers. Avec un peu d'habileté vous obtiendrez de bon résultats sur un câble non blindé.

- Pour le Fast-ATA et l'Ultra-ATA, la longueur maximale de la nappe est 45 cm. Les nappes des deux ports IDE sont souvent connectées, donc elle comptent pour *un seul* câble. Dans tous les cas les nappes IDE doivent être aussi courtes que possible. Si vous avez des plantages incompréhensibles ou des changements spontanés de données, examinez votre câblage. Essayer un mode PIO moins élevé (entre 1 et 4) ou déconnectez la seconde nappe si le problème persiste.
- Utilisez le moins de câble possible, mais n'oubliez pas la séparation de 30cm minimum entre deux périphériques ultra SCSI.
- Évitez les empilements entre la nappe et le disque, branchez la prise de la nappe directement sur le disque.
- Utilisez la bonne terminaison pour les périphériques SCSI et à la bonne position, c'est-à-dire aux deux extrémités de la chaîne SCSI. Souvenez-vous que l'adaptateur peut avoir une auto-terminaison: dans ce cas, il suffit de vérifier que l'autre extrémité est bien terminée.
- Ne mélangez pas les câbles blindés et non blindés, n'enroulez pas les câbles autour du métal, évitez de placer les câbles trop près des parties métalliques. Cela peut créer des différences d'impédance qui à leur tour entraînent la réflexion des signaux et augmentent le bruit sur le câble. Avec des contrôleurs multi-canaux le problème se pose de façon plus aiguë encore. On peut essayer de mettre du plastique autour des câbles pour éviter une trop grande proximité avec les éléments métalliques.

3.6 Adaptateurs

C'est l'autre extrémité de l'interface du disque, la partie connectée à un bus de la carte-mère. La vitesse du bus doit être assez élevée pour ne pas être une limitation par rapport à celle du disque. Mettre une rangée de disques RAID-0 sur une carde ISA serait du gâchis (car le bus ISA est trop lent). La plupart des machines actuelles ont un bus PCI 32 bits avec un débit de 132 Mo/s: dans un proche futur au moins, la vitesse du bus ne sera pas un facteur limitant sur ces machines.

Comme l'électronique a migré vers l'intérieur des disques, ce qui reste et qui constitue l'interface E(IDE) est ridiculement petit: souvent c'est intégré au contrôleur du bus PCI. Un adaptateur SCSI est plus complexe et comprend souvent un petit processeur: il est donc plus cher et n'est pas inclus dans le contrôleur PCI. En contrepartie, il décharge le processeur de certains calculs lors des accès disque.

Certains adaptateurs SCSI comportent même une mémoire cache et de l'intelligence pour anticiper les décisions du système d'exploitation. Mais le résultat dépend fortement du système d'exploitation utilisé. Linux a de son côté tant d'optimisations que le gain est souvent assez faible.

Mike Neuffer, qui a écrit les pilotes pour les contrôleurs DPT, assure que ces contrôleurs sont assez intelligents pour obtenir d'excellentes perfomances pourvu qu'ils aient suffisament de mémoire cache, et que les gens qui n'ont pas obtenu de gain de perfomances significatif avec des contrôleurs plus élaborés n'utilisent pas assez bien le contrôleur.

3.7 Systèmes multi-canaux

Pour augmenter les performances globales il faut identifier les facteurs limitants et les éliminer. Dans certains cas, avec un grand nombre de disques connectés, il est intéressant d'avoir plusieurs contrôleurs travaillant en parallèle, aussi bien pour le SCSI que pour l'IDE (les cartes mères ont souvent deux canaux IDE). Bien sûr Linux sait en tirer profit.

Certains contrôleurs RAID offrent 2 ou 3 canaux et c'est intéressant de répartir la mémoire de masse entre plusieurs canaux. Autrement dit, avec deux disques SCSI que vous voulez RAID-er et un contrôleur à deux canaux, placez un disque sur chaque canal.

3.8 Systèmes multi-cartes

On peut avoir du SCSI et du IDE sur la même machine, mais aussi plusieurs contrôleurs SCSI. Vérifiez dans le SCSI-HOWTO quels contrôleurs vous pouvez combiner. Sans doute vous devrez indiquer au noyau qu'il doit juste détecter un contrôleur au démarrage (l'autre contrôleur sera détecté et utilisé plus tard). Voyez la documentation de Lilo et du SCSI pour plus de détails.

Les systèmes à plusieurs contrôleurs peuvent offrir un gain de vitesse appréciable si on configure bien les disques, spécialement en mode RAID0. Pour bien paralléliser les disques et les contrôleurs, ajoutez les disques dans le bon ordre pour le driver `md`. Si le contrôleur 1 est connecté aux disques `sda` et `sdb` et le contrôleur 2 aux disques `sdc` et `sdd`, ajoutez les disques dans l'ordre `sda - sdc - sdb - sdd`, ainsi une lecture ou écriture concernant plus d'un cluster se répartira le plus souvent sur 2 contrôleurs.

La même méthode s'applique aux disques IDE. La plupart des cartes-mères ont 4 ports IDE:

- `hda` maître primaire
- `hdb` esclave primaire
- `hdc` maître secondaire
- `hdd` esclave secondaire

avec les deux disques primaires sur la même nappe, et les deux disques secondaires sur l'autre nappe. Il faut donc les concaténer dans l'ordre `hda - hdc - hdb - hdd` afin de paralléliser au maximum selon les deux canaux.

3.9 Comparatif de vitesse

Les tables suivantes donnent des vitesses indicatives (rappel: il s'agit de vitesses *théoriques* maximales).

3.9.1 Contrôleurs

```
IDE             :       8.3 - 16.7
Ultra-ATA       :       33

SCSI            :
                        Largeur du bus (bits)

Vitesse du Bus (MHz)    |        8       16      32
---------------------------------------------------
 5                      |        5       10      20
10  (fast)              |       10       20      40
20  (fast-20 / ultra)   |       20       40      80
40  (fast-40 / ultra-2) |       40       80      --
---------------------------------------------------
```

3.9.2 Types de bus

```
ISA             :       8-12
EISA            :       33
VESA            :       40    (Parfois poussé à 50)

PCI
                        Largeur de bus (bits)

Vitesse du Bus (MHz)    |       32      64
--------------------------------------------------
33                      |       132     264
66                      |       264     528
--------------------------------------------------
```

3.10 Jeux de tests (Benchmarks)

C'est un sujet très, très délicat et je ne m'engagerai que très prudemment sur ce terrain miné. Il est très difficile de faire des tests comparables et significatifs. Mais que ça ne décourage pas ceux qui voudront essayer ...

On peut utiliser les benchmarks pour un diagnostic du système, pour voir s'il est aussi rapide qu'il le devrait étant donné ses composantes. Ainsi en passant d'un système de fichiers tout simple au RAID, vous attendrez une accélération significative, donc une perte de performances vous informera que quelque chose déco^H^H^H^H ne va pas.

N'essayez pas de bricoler votre propre jeu de test, utilisez plutôt `iozone` et `bonnie`, et lisez la documentation très attentivement. Plus d'info dans la prochaine version du HOWTO.

3.11 Comparaisons

Le SCSI offre de meilleures performances que l'EIDE, mais cela se paye. La terminaison est plus complexe mais rajouter un disque n'est pas très difficile. Avoir plus de 4 (plus de 2 dans certains cas) disques IDE peut être compliqué, alors que le wide-SCSI supporte jusqu'à 15 disques par adaptateur (plus encore pour les contrôleurs multi-canaux).

Vous avez besoin d'un IRQ par contrôleur SCSI, chaque contrôleur pouvant gérer jusqu'à 15 disques. En revanche, vous avez besoin d'un IRQ par disque IDE, ce qui peut générer des conflits.

RLL et MFM sont trop vieux, lents et malpratiques pour être d'un utilité quelconque.

3.12 Perspectives

Le SCSI-3 est en préparation. Des disques plus rapides sont annoncés, et récemment un spécification monstre à 80 Mo/s sur un bus de 16 bits a été proposée.

Certains constructeurs ont annoncé des matériels SCSI-3 mais c'est prématuré car le standard n'est pas encore publié. Le point de saturation du bus PCI se rapproche. Actuellement la limite du bus PCI 64 bits à 33 MHz est 256 Mo/s, mais les futurs bus à 66 MHz grimperont à 528 Mo/s.

Une autre tendance est que l'espace disque est de plus en plus grand. On peut actuellement mettre 55 Go sur un seul disque, mais c'est encore assez cher. Le meilleur rapport espace/prix se situe autour de 8 Go et augmente continûment. L'introduction du DVD aura un grand impact dans un futur proche, avec 20 Go sur un seul disque on peut envisager même l'image intégrale des plus grands sites FTP. La seule chose certaine est que même si les disques ne sont pas mieux, ils seront plus gros.

Note: J'avais écrit dans ce HOWTO que la vitesse maximale des cédéroms était 20x à cause de problèmes de stabilité mécanique, mais peu après le premier cédérom 24x était disponible ... actuellement vous pouvez acheter un 40x et sans aucun doute des vitesses supérieures seront atteintes.

3.13 Recommandations

A mon avis EIDE ou Utra-DMA est mieux pour commencer sur une machine personnelle, spécialement si vous utilisez MS-DOS. Si vous voulez étendre votre système plus tard ou l'utiliser comme serveur, il est fortement recommandé d'utiliser des disques SCSI. Actuellement le wide-SCSI est légèrement plus cher. Le SCSI standard a un bon rapport qualité-prix. Il existe un bus SCSI différentiel qui permet une plus grande longueur de câble, mais il est tellement plus cher qu'on ne doit pas le recommander aux utilisateurs normaux.

En plus des disques vous pouvez ajouter des scanners et des imprimantes sur un bus SCSI.

Gardez à l'esprit que toute extension de votre système augmente la consommation d'électricité, et assurez-vous que l'alimentation et le refroidissement restent suffisants. Beaucoup de disques SCSI ont une option de démarrage en séquence adapté aux grands systèmes. Voir aussi 16.6 (Chaleur et Consommation)

4 Considérations diverses

Avec le PC familial, un utilisateur récemment converti à Linux cherchera surtout à obtenir les meilleures performances pour un matériel donné. Quelqu'un qui achète une machine pour un usage spécifique (comme un fournisseur d'accès à Internet) cherche au contraire à se procurer le matériel en fonction de ses besoins. Ce HOWTO couvre les deux situations.

De manière générale, le mieux est d'avoir autant de disques que possible, mais on ne peut pas en rajouter indéfiniment et le coût est aussi un facteur. A taille totale égale, plus il y a de partitions et de disques, plus la maintenance est compliquée.

4.1 Usage des systèmes de fichiers

Les différentes parties du FSSTND n'ont pas les mêmes exigences en terme de vitesse, de taille et de fiabilité. Casser la racine / est pénible mais peut être facilement réparé, casser `/var/spool/mail` c'est une autre histoire. Voici un bref résumé des principales parties d'un système de fichiers. Notez que c'est indicatif, qu'on peut très bien avoir des binaires dans `/etc` ou `/lib` et des librairies dans `bin`, etc.

4.1.1 Swap

(ndT: le swap est une partie du disque utilisée pour prolonger la mémoire vive de la machine. Il se comporte donc exactement comme de la mémoire vive supplémentaire, mais en 1000 fois plus lent)

Vitesse

Maximum! Si toutefois vous dépendez trop du swap, achetez plus de mémoire vive. Attention au fait que sur la plupart des cartes mères le cache ne marchera pas au-delà de 128 Mo.

Taille

Entre 1 fois et 2 fois celle de la mémoire vive. 4 Mo + 4 Mo (mémoire + swap) suffisent pour un système minimaliste et 16 Mo + 40 Mo permettent d'être à l'aise.

Attention à prendre en compte le type d'applications que vous utilisez. Pour faire du calcul formel ou du ray-traycing il se peut que 128 Mo de mémoire et autant de swap soient nécessaires.

Autre raison de ne pas lésiner sur la taille du swap: certains programmes ne libèrent pas complètement la mémoire qu'ils ont allouée, causant ce qu'on appelle des fuites de mémoire. La mémoire n'est pas libérée, même quand le programme s'arrête. Lorsque la mémoire vive et le swap sont pleins, il n'y a plus qu'à redémarrer. Heureusement ce genre de programmes est peu fréquent, mais avoir beaucoup de swap vous donne de la marge.

Certains programmes bloquent leurs pages en mémoire vive (on ne peut donc pas les swapper). Ce peut être pour des raisons de sécurité ou de performance (par exemple pour un système temps réel). Bien sûr de tels programmes, en occupant de la mémoire qui ne peut être swappée, font que le système commence à utiliser le swap plus tôt que prévu.

Le manuel de mkswap (`man 8 mkswap`) explique que chaque partition de swap ne doit pas excéder 128 Mo sur une machine 32-bit et 256Mo sur une machine 64-bit.

Fiabilité

Moyenne. En cas de problème vous le savez assez vite et vous pouvez perdre le travail en cours. Vous sauvegardez souvent, n'est-ce pas ?

Note 1

Linux permet de bâtir un swap à cheval sur plusieurs disques. Taper `man 8 swapon` pour les détails. Cepandant, un swap réparti sur plusieurs disques est souvent plus lent.

L'entrée dans le fichier `/etc/fstab` doit ressembler à:

```
/dev/sda1       swap            swap    pri=1           0       0
/dev/sdc1       swap            swap    pri=1           0       0
```

Le fichier `fstab` est *très* sensible au formatage utilisé, donc lisez attentivement la page de man et ne copie-pestez pas les lignes précédentes.

Note 2

Certains utilisent un disque de mémoire vive (RAM disk) comme mémoire swap. Mais comme l'usage du swap est d'augmenter la mémoire vive et qu'un RAM disk diminue la quantité de mémoire vive disponible (en particulier pour le cache disque), cette solution est à proscrire.

Note 2bis

Il y a une exception: sur un certain nombre de cartes-mères mal conçues, le cache externe ne peut pas cacher toute la mémoire vive qui peut être adressée. Ces cartes-mères peuvent supporter 128 Mo, mais seuls les premiers 64 Mo bénéficieront du cache. Dans ces conditions, les performances seront améliorées si on utilise les 64 Mo restants comme un RAMdisk pour le swap ou le stockage temporaire.

4.1.2 Stockage temporaire(`/tmp` and `/var/tmp`)

Vitesse

Très élevée. Sur un disque ou une partition séparée, cela réduira la fragmentation, mais de toute façon `ext2fs` fragmente très peu.

Taille

Difficile à dire. A la maison quelques Mo suffisent mais sur un serveur, certains utilisateurs y stockent leurs fichiers de manière à échapper aux quotas et au contrôle, et cette partition peut grandir démesurément. Disons donc: entre 8 et 32 Mo à la maison, 128 Mo pour un petit serveur et jusqu'à 500 Mo (la machine utilisé par l'auteur sert 1100 utilisateurs avec un répertoire `/tmp` de 300 Mo). Gardez un oeil sur ces répertoires, pour les fichiers cachés ou bien trop vieux. Attendez-vous un de ces jours à devoir retailler vos partitions à cause d'un `/tmp` trop petit.

Fiabilité

Faible. Souvent les programmes évitent de planter et produisent le bon message d'erreur quand ces répertoires sont pleins ou provoquent une erreur. Des erreurs de fichiers aléatoires sont bien sûrs plus sérieuses, mais c'est le cas pour toutes les partitions !

Fichiers

Principicipalement de petits fichiers à durée de vie assez courte. Les programmes bien écrits effacent leurs fichiers dans `/tmp` mais si une erreur survient à ce moment-là ils ne plantent pas, donc de "vieux" fichiers peuvent traîner dans `/tmp`. Avec la plupart des distributions, on a la possibilité d'effacer tout le contenu de `/tmp` au démarrage.

Note 1

Dans le FSSTND il y a une note sur la possibilité de mettre `/tmp` dans un disque de mémoire vive. Cependant, pour les mêmes raisons que pour le swap, ce n'est pas recommandé. Comme ça a déjà été dit, n'utilisez pas de flash RAM pour ces répertoires. Gardez en tête que les fichiers de `/tmp` sont effacés au redémarrage, avec certaines distributions.

Note 2

Dans les vieux systèmes on trouve un répertoire `/usr/tmp` mais on recommande de ne pas l'utiliser. Pour les vieux programmes, on en a fait un lien symbolique vers les autres aires de stockage temporaire.

4.1.3 Queues (`/var/spool/news` and `/var/spool/mail`)

Vitesse

Elevée, surtout pour les gros serveurs de news. Pour les queues d'impression: lente. Pour les news on peut envisager du RAID0.

Taille

Pour les seveurs de news et de mail: dépend des besoins. Pour un seul utilisateur quelques Mo suffisent, si on ne part pas en vacances en étant abonné à 10 mailing lists ... (La machine que j'utilise au travail a 100 Mo pour `/var/spool` tout entier)

Fiabilité

Mail: très haute, news: moyenne, queue d'impression: basse. Si votre mail est très important (mais n'est-ce pas le cas ?) songez à une solution RAID.

Fichiers

D'habitude un grand nombre de fichiers de quelques Ko, mais les fichiers d'une queue d'impression peuvent être assez gros.

Note

Certaines documentations des news suggèrent de mettre tous les fichiers `.overview` dans un disque différent de celui des news. Voir les FAQs pour plus d'informations. La taille de ces fichiers est entre 3 et 10 pourcents du total.

4.1.4 Répertoires utilisateurs (`/home`)

Vitesse

Moyenne. Certains programmes (comme les client des news) font de fréquentes mises à jour dans les répertoires des utilisateurs, ce qui peut avoir une importance s'il y a beaucoup d'utilisateurs. Pour les petits systèmes la vitesse n'est pas critique.

Taille

A vous de voir ! Avec certains fournisseurs on paie selon la place disque, donc c'est une question de gros sous. De grands systèmes comme

nyx.net `http://www.nyx.net/`

(service Internet gratuit avec le mail, les news et la Toile) marchent bien avec une taille suggérée de 100 Ko par utilisateur et 300 Ko au grand maximum. Les fournisseurs commerciaux offrent autour de 5 Mo par utilisateur.

Si vous écrivez des livres ou si vous programmez, les besoins augmentent vite.

Fiabilité

Variable. Perdre `/home` sur un système personnel est ennuyeux, mais recevoir 2000 coups de fils d'utilisateurs qui se plaignent que leur répertoire a disparu est plus qu'ennuyeux. Pour certains c'est vital. Vous faites des sauvegardes régulières, n'est-ce pas ?

Fichiers

A vous de voir. Le minimum des fichiers de démarrage de chaque utilisateur est une douzaine de fichiers pour environ 5 Ko.

Note 1

Vous pouvez envisager le RAID pour la vitesse ou la fiabilité. Si vous voulez une vitesse et une fiabilité extrême, vous devriez envisager une autre solution logicielle et matérielle (serveurs haut-de-gamme, système avec tolérance aux pannes, etc).

Note 2

Les brouteurs Web utilisent souvent un cache local qui peut prendre beaucoup de place et provoquer beaucoup d'activité disque. Il y a plusieurs moyens d'éviter cela, voir 8.6.1 (Répertoires Utilisateurs) et 8.6.3 (WWW).

Note 3

La tendance naturelle des utilisateurs est d'utiliser au maximum l'espace disque qu'on leur alloue. Le système de Quotas Linux permet de limiter le nombre de blocs et d'inodes qu'un seul utilisateur peut allouer par système de fichiers. Voir le

Linux Quota mini-HOWTO `http://www.freenix.fr/linux/HOWTO`

4.1.5 Exécutables (`/usr/bin` et `/usr/local/bin`)

Vitesse

Lente. La vitesse de chargement d'un binaire n'est pas critique, j'en veux pour témoin les bonnes performances des systèmes "live" sur un CDROM.

Taille

200 Mo devraient suffire. Un serveur à usages multiples devrait peut-être réserver 500 Mo pour anticiper la croissance.

Fiabilité

Basse. Les binaires essentiels sont en général dans `/bin` et `/sbin`. Si l'on perd tous les binaires, c'est pénible car il faut tout réinstaller, mais pas dramatique.

Fichiers

La plupart entre 10 et 100 Ko. Certains assez gros (emacs ...)

4.1.6 Librairies (`/usr/lib` and `/usr/local/lib`)

Vitesse

Moyenne. On trouve là plein de choses, des fontes comme des librairies dynamiques. Souvent les fichiers sont chargés en entier et donc une vitesse suffisante est nécessaire.

Taille

Variable. C'est là par exemple que les traitement de texte stockent leurs dizaines de mégas de fontes et d'exemples. Le peu de personnes qui m'ont contacté m'ont parlé de 70 Mo, mais une installation Debian 1.2 complète peut prendre plus de 250 Mo. Parmi les plus gros consommateurs de place disque: GCC, Emacs, TeX/LaTeX, X11 et perl.

Fiabilité

Basse. Comme pour les exécutables.

Fichiers

Assez gros avec un odre de grandeur de 1 Mo.

Note

Pour des raisons historiques certains programmes (comme GCC dans `/usr/lib/gcc/lib`) stockent des exécutables dans les répertoires de librairies.

4.1.7 Racine (/)

Vitesse

Assez lent: il n'y a là que le minimum, et la plupart des programmes ne sont lancés qu'au démarrage.

Taille

Assez petit. Cependant c'est une bonne idée de garder quelques fichiers et utilités de dépannage et plusieurs versions du noyau. 20 Mo devraient suffire.

Fiabilité

Elevée. Une panne de la racine peut être relativement coûteuse en temps et en cheveux arrachés. Avec de la pratique vous pourrez faire cela en un heure, mais si vous avez l'habitude de ce genre de choses c'est que quelque chose ne va pas.

Naturellement, vous avez une disquette de secours ? Et vous l'avez mise à jour régulièrement ? Il y a des disquettes toutes faites et des utilitaires de création de disquette de secours. Y passer un peu de temps peut vous épargner de devenit un expert en réparation de la partition racine.

Note 1

Si vous avez plein de disques, pourquoi ne pas mettre une partition de boot de secours sur un disque physiquement différent de celui sur lequel vous démarrez habituellement ? Le peu d'espace que ça vous coûtera sera amplement compensé par le temps gagné en cas de panne.

Note 2

Pour la simplicité comme pour le dépannage, il n'est pas recommandé de mettre la partition racine sur un système RAID niveau 0. Si vous utilisez RAID pour votre partition racine, il faut mettre l'option `md` pour votre noyau de secours.

Note 3

Pour démarrer avec Lilo il est important que les fichiers essentiels au démarrage résident entièrement dans les 1023 premiers cylindres. Ce qui comprend le noyau et les fichiers du répertoire `/boot`.

4.1.8 DOS, etc. <

Au risque de paraître hérétique j'ai inclus ce paragraphe au sujet duquel beaucoup ont des réaction vives. Malheureusement pas mal de disques sont livrés avec des outils d'installation et de maintenance basé sur ces systèmes et il faut en tenir compte.

Vitesse

Très lente. Les systèmes en question ne sont pas réputés pour leur vitesse donc il y a peu d'intérêt à utiliser des disques dernier cri. Le multitâche ou le multithread ne sont pas disponibles, donc les possibilités des disques SCSI ne sont pas pleinement exploitées. Un vieux disque IDE devrait faire l'affaire. Notons que Windows 95 et NT supportent le multi-tâches et devraient donc mieux profiter des caractéristiques du SCSI.

Taille

La compagnie qui produit ces systèmes n'est pas réputée pour écrire des programmes petits et optimisés, attendez-vous à y consacrer plusieurs dizaines de Mo. Avec une vieille version de DOS ou Windows ça peut tenir dans 50 Mo.

Fiabilité

Ha-ha. Comme une chaîne a la force de son maillon le plus faible, vous pouvez utiliser un vieux disque. Comme l'OS est plus facilement susceptible de s'auto-détruire que le disque, vous apprendrez sans doute ici l'importance des sauvegardes de secours.

Dit autrement: "Votre mission, allez-vous l'accepter, est de garder cette partition en état de servir. Cette mise en garde s'auto-détruira dans 10 secondes ..."

On m'a demandé récemment de justifier mes prises de positions dans ce paragraphe. D'abord je m'excuse mais je n'appelle pas DOS et Windows des systèmes d'exploitation. Ensuite il y a des implications légales à prendre en considération. Je ne donnerai au lecteur que quelques mots-clés: DOS 4.0, DOS 6.x et divers utilitaires de compression disque dont le nom devrait rester secret.

4.2 Explication des termes

Bien sûr le plus rapide est le mieux mais souvent le joyeux installateur de Linux a plusieurs disques de vitesse et de qualité variable. Bien sûr ce document pour rester utile à tous doit être général et ne saurait envisager tous les cas particuliers. Même ainsi il y a quelques détails à retenir:

4.2.1 Vitesse

C'est un mélange de plusieurs termes: charge du processeur principal, temps de mise en place du transfert, temps d'accès et taux de transfert. Il n'y a pas d'optimum fixé mais souvent le prix est le facteur déterminant. La charge processeur varie uniquement pour les disques IDE (car c'est le processeur principal qui pilote le disque), et pour les SCSI elle est toujours assez faible. Le temps d'accès est assez petit, quelques millisecondes. Il intervient assez peu si on utilise la queue des commandes SCSI, on peut alors lancer d'autres commandes pendant que les premières attendent leur réponse et le bus est occupé tout le temps au mieux. Dans le cas des serveurs de news, qui ont un grand nombre de petits fichiers, le temps d'accès peut être avoir plus d'influence sur la vitesse globale.

Les deux principaux paramètres sont:

Temps d'accès

Habituellement on donne le temps moyen pris par la tête de lecture pour aller d'une position à une autre au hasard. Ce paramètre a plus d'importance s'il y a beaucoup de petits fichiers. Il y a aussi un petit délai avant que le secteur désiré tourne et se retrouve en face de la tête. Ce délai est proportionnel à la vitesse angulaire. Des valeurs courantes de vitesse angulaire sont 4500, 5400 et 7200 tours/min. Les disques tournant plus vite sont donc plus rapides, mais ils coûtent plus chers, sont parfois bruyants et génèrent de la chaleur, paramètre qui compte si vous avez toute une rangée de disques. Avec les tous récents disques à 10000 tours/min les besoins de refroidissement sont encore plus grands et des schémas d'aération minimale sont donnés.

Taux de transfert

En Mo/s. Ce paramètre est plus important si on a peu de grands fichiers. A densité égale, la vitesse de transfert est proportionnelle à la vitesse angulaire.

Il est important de lire les spécifications des disques très attentivement, et de noter que le taux de transfert maximum est donné comme le taux de transfert entre la mémoire cache du disque et la mémoire principale, et *pas* comme le taux de transfert moyen entre le disque et la mémoire principale. Voir aussi 16.6 (Consommation et Chaleur).

4.2.2 Fiabilité

Bien sûr personne ne voudrait d'un disque pas très fiable. On ferait mieux de considérer les vieux disques comme non fiables. Pour le RAID il est suggéré d'utiliser un ensemble de disques différents de telle sorte que les pannes simultanées soient moins probables.

Autant que je sache, je n'ai connu qu'un cas d'un système de fichiers totalement foutu, mais dans ce cas un matériel instable semblait la cause des problèmes.

Les disques ne sont pas chers de nos jours et les gens sous-estiment toujours la valeur du contenu de leurs disques durs. Si vous avez besoin de matériel fiable, remplacez vos vieux disques et gardez des roues de secours. Un disque peut marcher plus ou moins en continu pendant des années, mais ce qui tue un disque c'est souvent en fin de compte les variations de tension.

4.2.3 Fichiers

La taille moyenne des fichiers est importante pour décider les bons paramètres du disque. Avec beaucoup de petits fichiers c'est le temps d'accès qui compte, et avec peu de gros fichiers c'est plutôt le taux de transfert. La queue des commandes SCSI est très bien adaptée à la gestion de beaucoup de petits fichiers, tandis que pour le taux de transfert EIDE et SCSI sont à peu près équivalents.

4.3 Technologies

Quelles sont les technologies disponibles et qu'est-ce que leur choix implique en terme de vitesse, fiabilité, consommation, flexibilité, facilité d'usage et complexité ?

4.3.1 RAID

C'est une méthode pour augmenter la fiabilité ou la vitesse ou les deux en utilisant plusieurs disques en parallèle. Ainsi les temps d'accès et taux de transferts sont diminués. Avec des miroirs et des vérifications (checksums) on peut améliorer la fiabilité. Ils sont un bon choix pour de gros serveurs mais pour un PC autant tuer une mouche au pistolet laser. Voir les documents et FAQs sur ce sujet.

Avec Linux on peut utiliser un système RAID soit logiciel (le module `md` du noyau) soit matériel avec un contrôleur supporté, qu'il soit PCI-SCSI ou SCSI-SCSI. Une solution matérielle est plus rapide, mais bien sûr plus chère.

Les contrôleurs SCSI-SCSI sont d'habitude réalisés comme un ensemble de disques et un contrôleur, communiquant entre eux par un second bus SCSI et qui se connectent au bus SCSI. De l'extérieur, l'ensemble se comporte comme un seul disque SCSI. Mais cette connexion au bus SCSI peut être un facteur limitant pour les performances. Un avantage significatif de ce genre de matériel est pour les gens qui ont de grands ensembles de disques durs: comme le nombre d'entrées SCSI dans le répertoire `/dev` est limité, cette solution permet d'utiliser plusieurs disques avec un seul fichier de périphérique.

Les contrôleurs PCI-SCSI, comme leur nom l'indique, sont connectés au bus PCI qui est plus rapide qu'un bus SCSI. Ces contrôleurs ont besoin de drivers spéciaux mais ils offrent du coup la possibilité de configurer le RAID à travers le réseau, ce qui simplifie l'administration.

Actuellement seules quelques familles de cartes PCI-SCSI sont supportées par Linux.

DPT

Les plus anciens et les plus matures sont les contrôleurs de

DPT `http://www.dpt.com`

parmi lesquels le SmartCache I/III/IV et le SmartRAID I/III/IV. Ces contrôleurs sont supportés par le pilote EATA-DMA du noyau standard. Cette société a aussi une

page d'information `http://www.dpt.com`

qui décrit certains aspects des technologies RAID et SCSI en plus de l'information sur leurs produits.

On peut consulter les page de l'auteur des pilotes pour contrôleurs DPT sur

SCSI `http://www.uni-mainz.de/~neuffer/scsi`

et sur

DPT `http://www.uni-mainz.de/~neuffer/scsi/dpt` .

Ces contrôleurs ne sont pas les plus rapides mais leur fiabilité n'est plus à prouver.

ICP-Vortex

Très récemment des contrôleurs de

ICP-Vortex `http://www.icp-vortex.com`

offrant jusqu'à 5 canaux indépendants et un matériel très rapide basé sur la puce i960. Le pilote a été écrit par le fabricant lui-même, qui prouve ainsi qu'il soutient Linux.

DAC-960

Encore en bêta-version. Plus d'information dans un futur proche.

Les contrôleurs SCSI-SCSI sont de petits ordinateurs, souvent avec une quantité appréciable de mémoire vive. Ils se présentent du point de vue extérieur comme un disque énorme, rapide et fiable. Ils n'ont donc pas besoin de pilote particulier (en plus de celui de la carte SCSI principale). Certains de ces contrôleurs ont une option pour parler à plusieurs adresses simultanément. D'habitude ils sont configurés grâce à une interface ou à un émulateur de terminal vt100 connecté à leur interface série.

Récemment j'ai appris que Syred faisait aussi des contrôleurs SCSI-SCSI supportés par Linux. Je n'ai pas plus d'information mais on peut regarder sur leur site:

www.syred.com `http://www.syred.com`

Je ne donne ici qu'un rapide aperçu du RAID qui a beaucoup de niveaux et de variantes. Le lecteur intéressé est invité à consulter la FAQ RAID.

- Le mode RAID *0* n'est pas redondant du tout mais offre le plus de vitesse. Les données sont réparties sur plusieurs disques et les opérations de lecture/écriture se font en parallèle. D'un autre côté, si un disque a une panne tout est fichu. Ai-je déjà mentionné les sauvegardes ?
- Le mode RAID *1*. C'est la méthode la plus primitive pour obtenir de la redondance: les données sont copiées sur chaque disque. C'est bien sûr un immense gâchis mais on a un grand avantage: un temps moyen d'accès très court. En effet les ordre de lecture sont envoyés à tous les disques et c'est le premier à répondre qui gagne. Le taux de transfert n'est pas significativement plus élevé qu'avec un seul disque, mais en lisant une piste différente sur chaque disque on peut parfois gagner du temps.

  Si vous n'avez que 2 disques c'est la seule façon d'avoir de la redondance.
- Les modes RAID *2* et *4* ne sont pas très courants et on n'en parlera pas ici.
- Le mode RAID *3* utilise plusieurs disques (au moins 2) pour mettre des données réparties comme en RAID 0. Il utilise aussi des disques redondants pour stocker le OU exclusif des données des disques de données. Si le disque redondant tombe en panne, le système peut continuer à fonctionner sans problème. Si c'est un disque de données qui crashe, le système peut récupérer les données à partir des disques redondants et des autres. Une panne double met le système hors-service.

  Le mode RAID 3 ne fait du sens qu'avec au moins 2 disques de données et un pour la redondance. Il n'y a pas de limite théorique, mais la probabilité de panne augmente avec le nombre de disques. La limite habituelle est de 5 à 7 disques.

  Comme toutes les opérations d'écriture doivent être répercutées sur le disque redondant, la vitesse globale en écriture d'un ensemble RAID 3 est celle de son disque redondant. La vitesse en lecture est celle d'un système RAID 0 ayant autant de disques que le RAID 3 a de disques non redondants. La vitesse chute sévèrement lorsque l'ensemble doit restaurer les données depuis le disque redondant.
- Le mode RAID *5* est comme le RAID 3, mis a part que l'information redondante est répartie sur l'ensemble des disques. Ça augmente la vitesse en écriture, puisque la charge est répartie.

Il y a aussi des modes hybrides basés sur le RAID 0 ou 1, et un autre niveau. Beaucoup de combinaisons sont possibles mais certaines sont assez complexes.

Le RAID *0/1* combine la répartition et la duplication, ce qui donne de très bons taux de transfert et temps d'accès moyen. Le revers de la médaille est que ça requiert beaucoup de disques et que c'est complexe.

Le RAID *1/5* combine la redondance façon RAID 5 et le court temps d'accès du RAID 1. La redondance est améliorée par rapport au RAID 0/1 mais la consommation de disques est significative. Il faudra au moins 6 disques pour mettre en place une telle solution, et peut-être plusieurs canaux ou contrôleurs SCSI.

4.3.2 AFS, Veritas et autres systèmes de gestion de volume

Avoir de nombreux disques et partitions constitue un avantage pour la taille, la vitesse et la fiabilité mais il y a un hic: Si la partition `/tmp` est pleine vous êtes bien embêté même s'il y a de la place dans la partition pour les news, car il n'est pas évident de retransférer les quotas d'une partition à l'autre. Les systèmes de gestion de volume font précisément ce travail. Les plus connus sont AFS et Veritas. Ils offrent aussi d'autres fonctions comme un journal des opérations disque. Veritas n'est pas disponible pour Linux, et il n'est pas certain qu'il puissent vendre des modules du noyau sans publier le code source, il est donc juste mentionné pour information. Pour voir comment ces systèmes fonctionnent vous pouvez consulter *le site de veritas* `http://www.veritas.com` .

Derek Atkins, du MIT, a porté AFS pour Linux et mis en place la

Linux AFS mailing List `mailto:linux-afs@mit.edu` qui est ouverte au public. Pour s'abonner à cett mailing-list il faut envoyer un mail à

linux-afs-request@mit.edu `mailto:linux-afs-request@mit.edu`

et si on trouve un bug

linux-afs-bugs@mit.edu `mailto:linux-afs-bugs@mit.edu` .

Attention: comme AFS utilise du cryptage il est restreint d'usage dans certains pays (ndT: la France par exemple). AFS est maintenant vendu par Transarc et ils ont mis en place un site Web. Voir

le site de Transarc `http://www.transarc.com`

pour des informations générales et une FAQ.

Il y a aussi des développements basés sur la dernière version libre d'AFS.

La gestion de volume est pour l'instant un des gros manques de Linux. Un projet a démarré au sujet d'un système de partitions virtuelles qui réalisera la plupart des fonctions de gestion de volume qu'on trouve dans le système AIX d'IBM.

4.3.3 Le patch `md` pour le noyau Linux

Il y a un projet de la part des développeurs du noyau, `md`, qui fait partie de la distribution du noyau depuis la version 1.3.69. `md` offre diverses fonctions telles que le RAID mais il est envore en phase de développement. Les gens qui l'ont utilisé parlent d'un succès mitigé voire d'un crash total. Bref, soyez prudents.

Actuellement `md` permet le mode linéaire et le RAID niveau 0,1,4 et 5: le plus stable doit être le RAID niveau 0 et 1, le reste est encore en développement. Il est aussi possible d'empiler les niveaux, par exemple de constituer un RAID 1 avec deux paires de disques, chaque paire étant un montage RAID 0.

Il faut bien prévoir quels disques on combine de manière à faire tourner tous les disques en parallèle, ce qui augmente les performances. Pour plus de détails se reporter à la documentation de `md`.

4.3.4 Considérations générales sur les systèmes de fichiers.

Dans le monde Linux `ext2fs` s'est imposé comme le système de fichiers à tout faire. Mais pour certains usages spécifiques, d'autres systèmes de fichiers sont préférables. Pour les serveurs de news un système avec journal (log file systems) est un choix naturel. C'est l'objet de vives controverses et il n'y a que peu de choix actuellement, mais on avance dans ce domaine. Les systèmes de fichiers avec journal ont l'avantage d'une vérification rapide. Un serveur de mail dans la classe 100 Go pourrait souffrir d'une vérification de systèmes de fichiers (avec `fsck`) prenant plusieurs jours au redémarrage.

Le système de fichiers de `Minix` est le plus ancien, très peu utilisé actuellement. Le système `Xiafs` était un candidat sérieux pour devenir le standard de Linux mais il n'a pas vécu.

Adam Richter d'Yggdrasil a posté récemment un message au sujet d'un système de fichiers avec journal et compression, mais c'est encore en développement. Une version qui ne marche pas est disponible sur le *serveur ftp d'Yggdrasil* `ftp://ftp.yggdrasil.com/private/adam`

avec des versions patchées du noyau. Peut-être que ça sera prochainement inclus dans la distribution officielle du noyau.

Le 23 juillet 1997, *Hans Reiser* `mailto:reiser(at)RICOCHET.NET` a publié les sources d'un système de fichiers basé sur la notion d'arbre, *reiserfs* `http://idiom.com/~beverly/reiserfs.html` . Ce système a des fonctionnalités très intéressantes et il est plus rapide que `ext2fs`, mais il est encore expérimental et pas facile à intégrer dans le noyau. on peut attendre d'importants développements dans le futur. Ce projet se distingue du système de fichiers avec journal moyen car Hans a déjà du code qui tourne.

Dans le système `ext2fs` existant, on pourrait ajouter de nouvelles fonctions comme les listes de contrôle d'accès (ACL, Access Control List), là encore dans un proche futur.

Il existe aussi un système de fichiers avec cryptage, mais un fois encore vérifiez qu'il est légal dans votre pays (ndT: rappel: en France c'est illégal pour le moment).

Les systèmes de fichiers sont un champ de recherches académiques et industrielles important, recherches dont les résultats sont souvent accessibles gratuitement (ndT: Il n'y a que les clients d'Apple ou Microsoft qui utilisent des technologies vieilles de 10 ans ...). Linux étant souvent la plate-forme de développement de tels prototypes, on peut s'attendre a des améliorations et des innovations continuelles.

4.3.5 Systèmes de fichiers des cédéroms

Il y a un certain nombre de systèmes de fichiers disponibles pour les cédéroms. Le plus ancien est le format *High Sierra*, nommé ainsi d'après l'hôtel où les accords furent signés par les partenaires industriels. C'était l'ancêtre de l'*ISO 9660*, qui est supporté par Linux (ndT: ce fut le nivellement par le bas: noms de fichiers de 8+3 caractères, majuscules/minuscules confondues, etc). Plus tard une extension *Rock Ridge* fut proposée, ajoutant les noms de fichiers longs et les droits d'accès entre autres.

Le système de fichiers iso9660 de Linux supporte aussi bien le vieux High Sierra que les extensions Rock Ridge.

Cepandant, une fois de plus Microsoft a décidé de choisir une de ces technologies comme nouveau "standard". Leur dernier bébé s'appelle *Joliet* et offre des possibilités d'internationaliation. Ce format est accepté par le noyau Linux depuis la version 2.0.34. Vous devez activer NLS pour l'utiliser.

H. Peter Anvin (hpa (at) transmeta.com) a récemment posté ces lignes:

```
Actually, Joliet is a city outside Chicago; best known for being the
site of the prison where Elwood was locked up in the movie "Blues
Brothers."  Rock Ridge (the UNIX extensions to ISO 9660) is named
after the (fictional) town in the movie "Blazing Saddles."

En fait, Joliet est une cité pas loin de Chicago, surtout célèbre pour
sa prison où Elwood était enfermé dans le film "Blues Brothers". Rock
Ridge (l'extension UNIX de l'ISO 9660) fut baptisé d'après la ville
imaginaire du film "Blazing Saddles."
```

En fait c'était Jake qui était enfermé. Oups !

4.3.6 Compression

Faut-il compresser son disque ou ses fichiers ? Voilà une question âprement débattue, surtout si on prend en compte le danger de perte de fichiers. Il y a pourtant plusieurs options pour les administrateurs aventureux: modules ou patchs du noyau, librairies. La plupart de ces solutions ont de limitations, comme par exemple d'être en lecture seule. Seules quelques références sont données ici; à vous de vous tenir au courant des dernières mises à jour.

- `DouBle` offre la compression de fichiers avec certaines limitations.
- `Zlibc` ajoute la compression au vol des fichiers quand on les charge, de façon transparente.
- Il y a beaucoup de modules qui permettent de lire des fichiers compressés ou des partitions natives de plusieurs systèmes d'exploitation, mais la plupart sont en lecture seule.
- `dmsdos` (actuellement en version 0.9.1.2) offre la plupart des options de compression de DOS et Windows. Il n'a pas encore tout mais de nouvelle fontionnalités sont régulièrement ajoutées.
- `e2compr` étend `ext2fs` avec des fonctions de compression. Il est pour le moment en phase de test donc utilisable seulement pour des hackers du noyau. Voir la

  page de e2compr `http://netspace.net.au/~reiter/e2compr.html`

  pour plus d'information. J'ai eu des rapports selon lesquels c'est assez stable et rapide.

4.3.7 Autres systèmes de fichiers

Il y a le système de fichiers `userfs` qui permet un système de fichiers basé sur FTP, et a entre autres des possibilité de compression. `docfs` est basé sur ce système de fichiers.

Avec les ajouts récents au noyau, on peut mettre un système de fichiers complet dans un seul fichier (appelé *loopback device*). On peut utiliser ça pour concevoir et tester de nouveaux systèmes de fichiers.

Notez que cela n'a rien a voir avec le *network loopback device*.

Il y a aussi un certain nombre de systèmes de fichiers au stade expérimental qui ne sont pas évoqués ici.

4.3.8 Position physique des pistes

Avec les disques petits et lents, certain systèmes de fichiers utilisaient au mieux les caractéristiques physiques lors du placement des données stockées. Cependant, l'augmentation de la vitesse et l'apparition de contrôleurs intégrés avec mémoire cache ont réduit l'effet de ces optimisations.

Néanmoins, on peut toujours gagner un peu avec ce genre d'optimisations. Comme chacun le sait, Linux va un jour *dominer le monde*, mais pour que ce jour arrive plus vite il nous faut employer toutes les ressources.

La plupart des disques tournent à vitesse angulaire constante mais utilisent une densité des données à peu près constante sur toutes les pistes. On a donc un taux de transfert bien plus élevé sur le bord que sur l'intérieur du disque. Mais il y a aussi le fait que le temps d'accès moyen aux données sockées sur le centre du disque est plus court que pour les données stockées au centre ou à l'extérieur.

Mais les disques récents utilisent une géométrie "logique" différente de la géométrie physique, le disque lui-même effectuant la conversion. Trouver le "milieu" du disque est plus difficile dans ces conditions.

Dans la plupart des cas la piste 0 est la plus à l'extérieur mais c'est une convention et pas une norme.

Les pistes intérieures

sont plus lentes pour le taux de transfert comme pour le temps d'accès.

Elles sont plus adaptées à des partitions telles que DOS, la racine ou la queue d'impression, qui ne demandent pas de vitesse élevée.

Les pistes du milieu

sont en moyenne plus rapides que les pistes intérieures pour le taux de transfert comme pour le temps d'accès. Elles sont bien adaptées pour des partitions comme `swap`, `/tmp` et `/var/tmp`.

Les pistes extérieures

ont le taux de transfert le plus rapide mais un temps d'accès moyen aussi faible que les pistes intérieures. C'est là qu'on pourra mettre de gros fichiers comme des librairies.

Le temps d'accès moyen peut être réduit en plaçant au centre les pistes les plus fréquemment demandées. Cela peut être fait avec `fdisk` en découpant un partition dans les pistes du milieu. Ou bien, avec un disque vide au départ, on peut copier un fichier bidon avec `dd` de la taille de la moitié du disque environ; on crée ensuite les fichiers qui ont besoin d'un accès rapide et on efface le fichier bidon.

Le dernier cas sert surtout pour les queues d'impression: on met le répertoire vide de départ au milieu du disque, ce qui réduira aussi la fragmentation.

Avec les systèmes RAID on peut aussi placer des fichiers au centre, mais le calcul est plus compliqué: voir la documentation sur RAID. On peut gagner jusqu'à 50 pourcents.

4.3.9 Vitesse des disques

Le système mécanique est souvent le même dans des disques IDE ou SCSI. Les contraintes mécaniques sont aujourd'hui un facteur limitant même si les progrès continuent. Il y a deux paramètres principaux, habituellement notés en millisecondes (ms):

Mobilité de la tête

La vitesse à laquelle la tête de lecture-écriture peut aller d'une piste à une autre, aussi appelé temps d'accès. Si vous calculez la double intégrale (la moyenne) de la distance sur tous les points de départ et tous les points d'arrivée possibles, vous trouverez que c'est équivalent à 1/3 de l'ensemble des pistes.

Vitesse de rotation

Elle détermine le temps nécessaire pour se placer dans le bon secteur, temps appelé latence.

Quelques valeurs typiques de temps mouvement de la tête:

```
                         Type de disque


Temps d'accès (ms)      | Rapide Moyen Vieux
---------------------------------------------
Pistes voisines             <1      2     8
En moyenne                  10     15    30
Au pire                     10     30    70
```

On voit que les disques dernier cri ont des temps d'accès à peine meilleurs que les disques moyens, mais que les vieux disques sont significativement moins bons.

```
Vitesse de rotation (tr/min)  |  3600 | 4500 | 4800 | 5400 | 7200 | 10000
---------------------------------------------------------------------------
Latence (ms)                  |    17 |    13 | 12.5 | 11.1 |  8.3 |   6.0
```

Comme la latence est le temps moyen pour atteindre un autre secteur, la formule est assez simple:

```
latence (ms) = 60000 / vitesse (tr/min)
```

Ce tableau montre lui aussi que la vitesse des disques progresse moins qu'auparavant. En revanche, la consommation d'électricité, l'échauffement et le bruit augmentent beaucoup.

5 Autres systèmes d'exploitation.

Beaucoup de Linuxiens ont plusieurs systèmes d'exploitation, ce qui est parfois nécessaire ne serait-ce que pour certains programmes de configuration du matériel qui ne tournent que sous DOS ou Windows, pour ne pas les nommer. D'où l'intérêt de cette courte section.

5.1 MS-DOS

Laissons là le débat pour savoir si c'est ou non un système d'exploitation. Ce qui est sûr est que la gestion du disque par MSDOS est très basique. On peut avoir de grandes difficultés avec les gros disques, consulter le *Large Drives mini-HOWTO* à ce sujet. Il est donc plus sage de placer la partition MSDOS au début du disque (sur les numéros de pistes les moins élevés).

Étant conçu pour de petits disque le système de fichier de MSDOS (*FAT*) alloue des blocs énormes sur les grands disques. Il crée aussi pas mal de fragmentation, ce qui ralentit le temps moyen d'accès comme le taux de transfert.

Une solution est d'utiliser le programme de défragmentation mais il est fortement conseillé de faire un sauvegarde des données et de vérifier le disque (avec `chkdsk` ou `scansidk` pour les DOS plus récents) avant de défragmenter.

Mais comme toujours il y a un os, et ici l'os s'appelle *fichiers cachés*. Certains vendeurs les utilisent pour se protéger leurs logiciels. Or un fichier caché ne peut être changé d'endroit sur le disque, même s'il garde la même place dans l'arborescence des répertoires. En conséquence les programmes de défragmentation ne déplacent pas les fichiers cachés, ce qui réduit les effets de la défragmentation.

Étant mono-tâche, mono-utilisateur, mono-tout, il n'y a aucun gain de vitesse à utiliser plusieurs disques sous MSDOS, a moins que vous utilisiez un contrôleur disque qui fait du RAID au niveau matériel.

Les vieilles commandes `join` et `subst` pour gérer plusieurs disques demandaient beaucoup de travail pour un résultat nul. Elles n'existent plus dans les versions récentes.

Bref, il n'y a pas grand chose à faire pour accélérer la gestion disque de DOS. Sauf ceci: beaucoup de programmes ont besoin d'un espace de stockage temporaire rapide et ceux qui sont bien écrits utilisent la variable d'environnement `TEMPDIR` ou `TMPDIR` pour savoir où créer ces fichiers. Vous pouvez faire pointer cette variable vers un autre disque en éditant le fichier `autoexec.bat`:

```
SET TMPDIR=E:/TMP
```

En plus du gain de vitesse, ceci réduira sans doute la fragmentation.

Le programme `fdisk` de MSDOS a du mal parfois a effacer des partitions primaires. On peut utiliser à la place le programme `fdisk` qui vient avec Linux.

N'oubliez pas qu'il existe d'autres alternatives à MS-DOS, la plus connue étant

DR-DOS `http://www.caldera/dos/`

de

Caldera http://www.caldera/ . C'est un descendant direct de DR-DOS de Digital Research. Il a beaucoup de fonctions qui manquent à MS-DOS, comme le multi-tâche.

Une autre alternative, libre, est

Free DOS http://www.freedos.org/

qui est un projet en développement. Un certain nombre d'utilitaires sont aussi disponibles.

5.2 Windows

La plupart des points ci-dessus s'appliquent, mis à part que Windows95 a une meilleure gestion des disques, notamment SCSI.

Pour lire les noms longs vous devrez utiliser le système de fichiers `vfat` (plutôt que `dos`) pour monter ces partitions.

Le nouveau système de fichiers `FAT32` a été introduit à partir de la version OSR2. Il convient mieux aux grands disques. Il n'est encore supporté que par peu de programmes, même par NT 4.0 ou les utilitaires Norton-machin truc. Le noyau Linux supporte le format `FAT32` et aussi le format de cédéroms `Joliet` depuis la version 2.0.35.

La fragmentation est encore un problème. On peut limiter les dégâts en faisant une déframentation avant et après tout gros changement (comme l'installation d'un programme). Enlever les fichiers inutiles et vider la poubelle réduit encore la fragmentation.

Windows utilise aussi un disque pour le swap, et le rediriger peut apporter des gains de performance. Il y a plusieurs mini-HOWTOS qui expliquent comment partager le swap entre plusieurs systèmes d'exploitation.

Très récemment quelqu'un a commencé un projet pour que Win95 reconnaisse le système de fichiers `ext2fs`. Voir

cette page Web http://www.globalxs.nl/home/p/pvs/ pour plus de détails.

Mettre la variable d'environnement `TEMPDIR` est toujours utile mais tous les programmes ne l'utilisent pas. Utilisez `sysedit` pour éditer le fichier `autoexec` comme indiqué plus haut.

Beaucoup de fichiers temporaires sont placés dans `/windows/temp` et changer cela est plus difficile. On peut utiliser `regedit` à cette fin, mais on risque de mettre le système dans un état incohérent; et un Windows en panne est encore moins utile qu'un Windows vivant. Une erreur dans la base des registres peut nécessiter la ré-installation complète de Windows.

De toute façon, beaucoup de programmes ont leurs propres endroits pour mettre leurs fichiers temporaires, il y en a donc un peu partout sur votre disque.

Mettre le swap sur une partition séparée est une meilleure idée, et c'est plus facile à faire. Gardez à l'esprit que la partition swap ne peut être utilisée à rien d'autre, même s'il y a de la place libre.

5.3 OS/2

La seule chose à mentionner ici est qu'on peut ajouter un système de fichiers de manière à pouvoir lire les partitions `ext2fs` depuis OS/2.

5.4 NT

Voilà un système plus sérieux qui proposent la plupart des fonctions dont les noms exotiques composent la publicité informatique.

Voici un bug reporté par acahalan at cs.uml.edu: (traduction d'un extrait de message dans les News)

Le DiskManager de NT a un bug sérieux qui peut corrompre un disque ayant plus d'une partition étendue. Microsoft a mis un fix sur le site *knowledge base* `http://www.microsoft.com/kb/`

(Cela concerne les Linuxiens, car ils ont des souvent des partitions étendues)

5.5 Sun OS

Il y a un peu de confusion entre Sun OS et Solaris. Solaris n'est que Sun OS 5 avec Openwindows et quelques extras. Tapez `uname -a` pour connaître votre version. Parmi les raisons de cette confusion il y a que Sun utilisait un OS de la famille BSD, avec des morceaux de code d'un peu partout et du code propriétaire. Ainsi jusqu'à Sun OS 4.x.y. Puis par une décision stratégique ils ont choisi Unix, System V, Release 4 officiel (SVR4) et Sun OS 5 est né. Ils ont aussi changé de marketing, en vendant d'autres produits en *bundle* avec Sun OS sous le nom de Solaris, actuellement en version 2.6.

5.5.1 Sun OS 4

Sun OS est familier à beaucoup de Linuxiens. La dernière version est 4.1.4 plus un certain nombre de patches. Notez cependant que leur hiérarchie de fichiers est organisée différemment du FSSTND. Taper `man hier` pour un bref topo sur la hiérarchie de fichiers.

5.5.2 Sun OS 5 (i.e. Solaris)

Il y a une procédure d'installation basée sur Openwindows, qui vous aide à partitionner et formater les disques avant d'installer le système à partir du cédérom. Cette procédure plante lamentablement si votre installation est trop exotique, et comme elle cherche à installer tout un système à partir d'un cédérom elle plantera mais pas avant un certain nombre de minutes. C'est l'expérience que j'en ai eu. Pour contourner le problème nous avons tout installé sur une partition et ensuite nous avons déplacé les répertoires aux bons endroits.

Les valeurs par défaut sont bonnes pour la plupart des choses, sauf peut-être pour le swap. Alors que les manuels officiels recommandent d'avoir plusieurs partitions pour le swap, pas défaut une seule partition est utilisée. Il est conseillé de changer cela dès que possible.

Sun OS 5 possède aussi un système de fichiers conçu pour les fichiers temporaires, `tmpfs`. C'est un genre de RAM disk, et comme les RAM disks le contenu en est perdu quand le courant est coupé. Si la mémoire vive manque des parties du pseudo-disques seront déplacés vers la mémoire tampon, il est donc possible d'avoir des fichiers temporaires dans la partition de swap. Linux n'a pas de système de fichiers de ce genre: on en avait parlé mais les opinions étaient partagées. J'aimerais d'ailleurs avoir des commentaires à ce sujet.

Jusqu'ici, le seul commentaire était: non !! Sous Solaris 2.0, créer de trop gros fichiers temporaires dans `/tmp` peut causer une panne dy noyau (*kernel panic*) pour cause de manque de mémoire tampon (ndT: Ce n'est qu'un des milliers de bugs de Solaris 2.0). Le pire est que cette panne complète peut arriver avec des programmes utilisateurs (donc pas seulement avec des programmes en mode noyau) et à moins de savoir contourner le problème le mieux est de ne pas utiliser `tmpfs`.

Voir aussi 16.1 (Combiner le swap et /tmp).

Pour la culture: il y un film appelé Solaris, un film de science fiction très long, très lent et totalement incompréhensible ...

6 Clusters

Je vais brièvement évoquer ici les manières de connecter des machines ensemble, mais c'est un sujet si vaste qu'il pourrait faire l'objet d'un HOWTO. Comme en plus c'est hors-sujet dans ce HOWTO, si vous voulez contactez-moi et prenez cette partie pour en faire un document séparé.

Aujourd'hui les ordinateurs sont obsolètes au bout d'un temps très court. Du vieux matériel peut pourtant se révéler très utile sous Linux. Utiliser un vieux pécé commer serveur réseau a, en plus de la valeur pratique, un certain intérêt éducatif. Je ne parlerai ici que de ce qui concerne les disques.

Plusieurs formes de partage (clustering) sont possibles aujourd'hui, depuis la répartition automatique de la charge entre plusieurs machines jusqu'à des matériels exotiques comme le SCI (Scalable Coherent Interface) qui permet de combiner plusieurs machines en une seule. Il y a eu aussi du partage sur de plus grosses machines, ainsi le VAXcluster en son temps. L'usage habituel du clustering est le partage des ressources comme les disques durs, les imprimantes, les teminaux mais de façon à ce que les ressources soit disponibles à égalité pour tous les noeuds du réseau.

Il n'y a pas de bonne définition du clustering (ndT: ni de bonne traduction ...) mais ici ce mot signifie que plusieurs machines d'un reseau combinent leurs ressources pour servir les utilisateurs.

Linux permet certaines formes de partage mais pour le débutant je décrirai juste un réseau local simple. C'est une bonne manière de profiter de vieux matériel qui serait inutilisable dans ça.

La meilleure façon d'utiliser une vieille machine est d'en faire un serveur de réseau. Dans ce cas, le facteur limitant est plutôt la bande passante du réseau que la vitesse du serveur. A la maison vous pouvez déplacer les fonctions suivantes sur un vieux PC devenu serveur:

- Les news
- Le courrier électronique
- Les proxies Web
- Un serveur d'impression
- Un serveur de modem (PPP, SLIP, FAX, Voice mail)

Vous pouvez aussi monter par `NFS` des disques du serveur. Lisez le FSSNTD pour savoir quels répertoires ne doivent pas être exportés. On exportera `/usr` et `/var/spool`, peut-être aussi `/usr/local` mais sans doute pas `/var/spool/lpd`.

La plupart du temps même de vieux disques offrent des performances suffisantes. Cependant, si vous avez un usage intensif des disques du serveur et un réseau à haut débit, vous aurez sans doute besoin de disques rapides. C'est le cas pour un outil de recherche dans un site Web ou pour une base de données.

Un tel réseau (un *toaster network* comme on l'appelle) peut être une très bonne façon d'apprendre l'administration système. Il y a des HOWTOs sur le sujet mais vous devez garder en tête les choses suivantes:

- Ne choisissez pas les numéros IP n'importe comment. Configurez votre réseau local avec les adresses IP réservées à l'usage privé, et utilisez votre serveur de réseau comme un routeur qui gérera le masquage des adresses IP.

- Si vous configurez le routeur comme un pare-feu (firewall) il se peut que vous soyez incapable d'accéder à vos propres données depuis l'extérieur. Cela dépend de la configuration du pare-feu.

Le réseau *nyx* est un exemple de cluster. Il est constitué de:

nyx
: est l'une des deux machines sur lesquelles les utilisateurs se loguent et assure aussi certaines fonctions réseau

nox
: (ou nyx10) est la machine principale pour utilisateurs et aussi un serveur de courrier électronique.

noc
: est un serveur pour les news. La queue des news est accessible par un montage NFS pour nyx et nox.

arachne
: (ou www) est le serveur Web. Les pages Web sont écrites sur nox à travers un montage NFS

Il y a des projets de clustering assez avancés, notamment:

- *le projet Beowulf* `http://cesdis.gsfc.nasa.gov/linux/beowulf/beowulf.html`
- *le projet GAMMA (Genoa Active Message Machine)* `http://www.disi.unige.it/project/gamma/`

Le partage high-tech demande une interconnection high-tech, et SCI est une des solutions. Plus d'information sur la page Web de

Dolphin Interconnect Solutions `http://www.dolphinics.no/`

ou de *scizzl* `http://www.scizzl.com/` .

7 Points de montage

Il est important de ne pas scinder la hiérarchie des répertoires au mauvais endroits. Cette section dépend fortement du FSSTND et sans doute changera complètement quand le FHS sera utilisée dans une distribution Linux.

Voici donc un liste des répertoires que vous *pouvez* (et non que vous *devez*) mettre sur une partition séparée. Pour indiquer combien il est opportun de placer tel répertoire sur une partition séparée, un échelle de 0 à 5 est adoptée:

```
0=À éviter absolument
1=eventullement
 ...
4=utile
5=recommandé

/
|
+-bin       0
+-boot      0
+-dev       0
+-etc       0
```

```
+-home      5
+-lib       0
+-mnt       0
+-proc      0
+-root      0
+-sbin      0
+-tmp       5
+-usr       5
| \
| +-X11R6     3
| +-bin       3
| +-lib       4
| +-local     4
| | \
| | +bin        2
| | +lib        4
| +-src       3
|
+-var       5
  \
  +-adm       0
  +-lib       2
  +-lock      1
  +-log       1
  +-preserve  1
  +-run       1
  +-spool     4
  | \
  | +-mail      3
  | +-mqueue    3
  | +-news      5
  | +-smail     3
  | +-uucp      3
  +-tmp       5
```

La situation bien sûr peut varier, par exemple sur une machine à la maison il n'est pas très utile de scinder le répertoire `/var/spool` mais pour un fournisseur d'accès à Internet c'est indispensable. Le mot-clé ici est *l'usage*.

QUESTION ! Pourquoi `/etc` ne doit jamais être mis sur une partition séparée ? Réponse: le montage est fait d'après les instructions du fichier `/etc/fstab`, donc si `/etc` n'est pas sur la partition racine, et que cette partition n'est pas montée, aucun montage ne peut être effectué ... c'est comme d'avoir claqué la porte en laissant la clé à l'intérieur.

8 Placement des partitions, des répertoires et des fichiers

Nous en savons maintenant assez pour parler de placement. J'ai mis ma méthode au point après avoir essayé toutes les combinaisons possibles sur mes 3 vieux disques SCSI.

Les tables données en appendice servent à simplifier le processus. Elles vous aideront à optimiser votre système mais aussi à le dépanner éventuellement. Quelques exemples sont donnés.

8.1 Choisir les partitions

Réfléchissez à vos besoins et posez sur le papier une liste de toutes les parties de votre système de fichiers que vous voulez mettre sur une partition séparée. Notez la taille de chacune et triez-les par vitesse décroissante.

La table du chapitre 17 (Appendice A) est utile pour choisir quels répertoires mettre dans quelles partitions. Elle est triée par ordre logique, avec des blancs pour vos notes personnelles et des remarques sur les points de montage. Elle n'est PAS triée par vitesse décroissante, mais les besoins en vitesse sont indiqués par des petits ronds ('o').

Si vous voulez utiliser du RAID notez avec quels disques vous voulez le faire et quelles partitions seront en RAID. Notez que les différents modes RAID offrent une vitesse et une fiabilité variable. Pour simplifier, on suppose dans la suite qu'on a un ensemble de disques SCSI identiques et pas de RAID.

8.2 Répartir les partitions entre les disques.

Il faut maintenant déterminer sur quelles disques physiques seront placées les partitions choisies ci-dessus. Voici un algorithme pour optimiser le parallélisme et l'utilisation du bus. Dans notre exemple les partitions à placer sont 123456789, 9 est celle qui a besoin de la plus grande vitesse et 1 est la plus lente. On les répartit comme suit:

```
A : 9 4 3
B : 8 5 2
C : 7 6 1
```

Cela fait une "moyenne des vitesses" à peu près égale sur chaque disque.

Utiliser la table de l'appendice B pour déterminer quels disques utiliser pour quelles partitions afin de profiter au maximum du parallélisme.

Notez la vitesse de chacun de vos disques dans la bonne colonne. Éventuellement, permutez les répertoires, les partitions et les disques jusqu'à être content du résultat.

8.3 Trier les partitions et les disques

L'étape suivante est de sélectionner les numéros de partition pour chaque disque.

Utilisez la table du chapitre 19 (appendice C) pour sélectionner les numéros de partitions à l'intérieur de chaque disque. Remplissez avec ces valeurs les tables des Appendices A et B. Ces tables vous serviront lorsque vous installerez votre système (étape de partitionnement avec `fdisk` ou `cfdisk`)

8.4 Optimisation

Des considérations spécifiques à un matériel ou à un type d'utilisation peuvent intervenir. Par exemple si le disque C est beaucoup plus lent que les deux autres il vaudra mieux adopter la répartition suivante:

```
A : 9 6 5
B : 8 7 4
C : 3 2 1
```

8.4.1 En tenant compte de spécificité des disques

Des disques de vitesse globale comparable peuvent s'avérer plus ou adaptés à un usage ou à un autre. Comme on l'a déjà dit, les binaires, qui sont nombreux et petits, sont bien à leur place dans un disque de temps d'accès moyen faible et qui gère une queue des requêtes. Les librairies et autres gros fichiers profiteront davantage d'un disque ayant un bon taux de transfert, ce que les disques IDE offrent pour pas cher.

8.4.2 Utilisation du parallélisme

On peut éviter la surcharge du disque en pensant aux tâches. Par exemple si vous exécutez un programme de `/usr/local/bin` il y a des chances que vous accéderez aussi à `/usr/local/lib`; placer ces deux répertoires sur des disques physiquement différents permet de diminuer le temps de recherche et autorise les opérations en parallèle ou l'utilisation du cache. Des gains de performance surprenants peuvent être obtenus ainsi. Identifiez les tâches communes, les partitions qu'elles utilisent et gardez ces partitions sur des disques physiquement différents.

Voici quelques exemples:

Les application bureautiques

comme les traitements de texte ou les tableurs sont des exemples typiques de logiciels peu gourmands en temps CPU comme en accès disque (une fois lancés). Cepandant, ces logiciels ont souvent des fonctions de sauvegarde automatique qui créent du traffic dans les répertoires personnels des utilisateurs. Avoir les répertoires personnels sur plusieurs disques répartira la charge.

Les lecteurs de News

ont aussi des fonctions de sauvegarde automatique, et les fournisseurs d'accès à Internet ont intérêt à séparer les répertoires utilisateurs entre plusieurs disques.

Les queues des serveurs de News (/var/spool/news) sont connues pour leurs grand nombre de répertoires et de fichiers. La perte d'une telle partition n'est pas grave dans la plupart des cas, donc le RAID 0 lui convient parfaitement. Avec beaucoup de petits disques le système pourra supporter un grand nombre de requêtes par seconde. On peut même mettre les news et les fichiers `.overview` sur des disques séparés: voir les FAQs sur les serveurs INN à ce sujet.

Voir aussi la page Web dédiée à

l'optimisation des serveurs INN `http://www.spinne.com/usenet/inn-perf.html`

Les bases de données

sont gourmandes en terme d'accès disques comme de temps de calcul. Cela dépend beaucoup de l'application envisagée. On peut envisager le RAID pour avoir à la fois performance et fiabilité.

Le courrier électronique

met en jeu les répertoires des utilisateurs comme les queues de courrier arrivé/à envoyer. Si possible garder les répertoires des utilisateurs et les queues sur des disques différents. Pour un serveur de courrier on peut envisager de mettre les queues de courrier reçu et à envoyer sur des disques différents.

Perdre du courrier est extrêmement gênant, si vous êtes un fournisseur d'accès ou un routeur. Envisager le RAID et faire des sauvegardes fréquentes.

Le développement de logiciels

peut demander un grand nombre de répertoires pour les binaires, les librairies, les fichiers d'en-tête, les sources et l'archive. Séparer autant que possible tous ces répertoires. Sur des petits systèmes vous pouvez placer `/usr/src` et l'archive sur le même disque que les répertoires personnels.

Surfer sur le Net

est à la mode. Les butineurs ont souvent un cache local qui peut grossir pas mal. Comme le cache est utilisé pour recharger des pages ou retourner à la page précédents, la vitesse compte. Cepandant, si vous êtes connectés à un bon serveur de proxy les utilisateurs n'ont plus besoin de cache individuel. Voir aussi 8.6.1 (Les répertoires personnels des utilisateurs) et 8.6.3 (Le Web).

8.5 Besoins et usage

Lorque vous achetez une boîte de 10 cédéroms avec une distribution Linux et le contenu de gros sites FTP, il peut être tentant de vouloir installer autant de choses que vos disques le peuvent. Cependant, vous ne tarderez pas à trouver que ça vous laisse bien peu de place pour évoluer. Voilà pourquoi je soulignerai quelques points importants.

Tester

Linux est simple et vous n'avez même pas besoin d'un disque dur pour cela. Il sufit d'une disquette de démarrage comme celles fournies avec les distributions. Si vos périphérique ne sont pas supportés, n'oubliez pas qu'il y a souvent plusieurs versions de disquette de démarrage pour les périphériques exotiques qui peuvent vous dépanner jusqu'à la compilation d'un noyau personnalisé.

Apprendre

comment marche un système d'exploitation est très facile avec Linux: c'est un système qui vient avec les sources et une abondante documentation. Un disque de 50 Mo suffit pour avoir un shell et les utilitaires les plus courants.

Si ça devient un hobby

des programmes plus nombreux sont nécessaires, mais 500 Mo sur un seul disque devraient suffire pour les binaires, les sources et la documentation.

Pour un usage professionnel

ou amateur sérieux, il faut encore plus de place, des queues pour le courrier électronique et les nouvelles, etc. Séparer les fichiers entre plusieurs disques peut être bénéfique. La place requise est plus difficile à estimer, mais 2 à 4 Go devraient être plus que suffisants, même pour un petit serveur.

Les serveurs

vont du simple serveur de courrier électronique au gros serveur pour un fournisseur d'accès à Internet. Compter 2 Go pour le système de base, ajouter ensuite de la place (et probablement des disques) pour chaque service proposé. Le coût est ici le facteur limitant mais si on veut justifier le S de Service il faut bien dépenser un peu. J'admets que tous les fournisseurs d'accès ne le font pas.

8.6 Serveurs

Dans les appendices on trouvera les valeurs à employer pour un serveur départemental (de 10 à 100 utilisateurs). Dans cette section on parlera des grands serveurs. De manière générale n'ayez pas peur d'employer le RAID, pas seulement parce qu'il est rapide et fiable mais aussi parce qu'il est un peu plus facile de faire grandir un système RAID. Ce qui est mentionné ici s'ajoute aux remarques précédentes.

Le plus souvent les gros serveurs ne sont pas apparus comme ça, mais ils ont grandi progressivement. Dans la plupart des cas c'est une bonne idée de réserver un ou plusieurs disque SCSI pour chaque tâche. Cela permet de récupérer efficacement les données si le serveur est hors d'usage. Notez que transporter un disque d'une machine à une autre n'est pas si simple, en particulier pour les disques IDE. Et les tours de disques

SCSI ont besoin d'une initialisation correcte pour reconstruire les données, donc vous devez garder une copie papier de votre fichier /etc/fstab comme des numéros de série des disques SCSI.

8.6.1 Répertoires personnels des utilisateurs

Faites une estimation du nombre de disques requis, si c'est plus que 2 je recommande fortement le RAID. Si vous ne l'utilisez pas, vous pouvez utiliser un algorithme de hachage simple pour répartir la charge entre les disques. Par exemple vous pouvez utiliser les deux premières lettres du nom de login, ainsi `jbloggs` est mis sur `/u/j/b/jbloggs` où `/u/j` est un lien symbolique vers un disque physique.

8.6.2 Serveur FTP anonyme

C'est un équipement essentiel si vous attachez de l'importance à la notion de service. Les bons serveurs sont bien maintenus, documentés, à jour, et très populaires où qu'ils soient dans le monde. Le serveur

ftp.funet.fi `ftp://ftp.funet.fi`

(ndT: et en France *ftp.lip6.fr* `ftp://ftp.lip6.fr/`) est un exemple de "gros serveur FTP".

En général c'est plutôt la bande passante du réseau que la vitesse du processeur qui compte. La taille varie beaucoup. Je crois que l'archive de *ftp.cdrom.com* `ftp://ftp.cdrom.com` est une machine *BSD avec 50 Go de disque. La mémoire vive est importante aussi: 256 Mo pour un gros serveur mais de plus petits peuvent se contenter de 64 Mo.

8.6.3 La toile (WWW)

Pour beaucoup c'est la principale raison d'aller sur l'Internet. En plus de consommer de la bande passante, cette activité génère des besoins en cache disque. Garder le cache sur un disque rapide, à part peut être intéressant. Avoir un serveur de proxy est encore mieux. Cela peut réduire la taille du cache pour chaque utilisateur et accélérer le service en diminuant la bande passante utilisée.

Un serveur de cache a besoin d'un ensemble de disques rapides, le RAID0 est idéal dans ce cas car la fiabilité n'est pas primordiale. 2 Go devraient suffire. Ne pas oublier d'adapter la durée de vie des pages dans le cache à la capacité disque et aux besoins. On peut adapter la durée de vie selon les serveurs, voir: `Harvest`,

Squid `http://www.nlanr.net/Squid`

ou le serveur de

Netscape `http://www.netscape.com`

pour plus de détails.

8.6.4 Courrier électronique

La plupart des machines manipulent, peu ou prou, du courrier électronique. Cependant, les grands serveurs de courrier forment une catégorie à part. C'est une tâche très exigeante et même un gros serveur doté de disques rapides et d'une bonne connexion au réseau peut se révéler lent à l'usage. A la différence des news qui sont réparties sur plusieurs serveurs, le courrier électronique est centralisé. La sécurité est donc bien plus importante. Pour un gros serveur envisagez une solution RAID redondante (RAID4 ou RAID5).

8.6.5 News

C'est une tâche qui demande de grands volumes, mais cela dépend beaucoup du nombre de forums où vous souscrivez. Sur Nyx il y a en a 17 Go. Les plus grands groupes sont sans doute dans la hiérarchie `alt.binary.*`, vous pouvez sans doute assurer un bon service avec 12 Go si vous ns vous abonnez pas à ces groupes. Certains que je ne nommerai pas pensent que 2 Go suffisent pour prétendre assurer un "Service d'Accès à Internet". Dans ce cas les news expirent si vite que le mot de "service" se justifie peu. Un vrai serveur de news signifie un trafic de plusieurs Go par jour, et ce nombre ne cesse de croître.

8.6.6 Autres

Il y a plein de services disponibles sur Internet, même si la plupart ont été jeté aux oubliettes par la Toile. Cependant, des services comme *archie*, *gopher* et *wais* existent encore et restent des outils appréciables.

8.7 Pièges

Les dangers de tout scinder entre des partitions distinctes sont mentionnés dans la section sur la gestion de volume. Mais on m'a demandé d'insister sur ce point: quand une partition est pleine, elle ne peut plus grandir, même s'il y a de la place sur les autres partitions.

En particulier, il faut veiller à la croissance explosive de la queue des News (`/var/spool/news`). Pour les machines multi-utilisateurs avec des quotas gardez un oeil sur `/tmp` et `/var/tmp` car certains utilisateurs stockent leurs fichiers là, recherchez seulement les noms de fichiers terminés par gif ou jpeg ...

Il n'y a aucun avantage à tirer d'un seul disque scindé en plusieurs partitions, si ce n'est que ça rend la surveillance des fichiers (avec la commande '`df`') plus facile et que ça permet de mettre les partitions rapides sur le milieu (physique) du disque. Mais ça n'apporte rien en terme d'accès en parallèle à plusieurs partitions

8.8 Compromis

Une manière d'éviter le piège mentionné ci-dessus est de ne mettre que les partitions dont la taille est peu susceptible de varier comme le swap, `/tmp` et `/var/tmp` et de regrouper les autres dans les partitions restantes au moyen de liens symboliques.

Exemple: Soit un disque lent (`slowdisk`), et un disque rapide (`fastdisk`), et une collection de fichiers. Nous mettons `swap` et `tmp` sur `fastdisk`; `/home` et la racine sur `slowdisk`. Et nous avons encore les répertoires (fictifs) `/a/slow`, `/a/fast`, `/b/slow` and `/b/fast` à placer sur les deux partitions `/mnt.slowdisk` et `/mnt.fastdisk` faites avec l'espace restant sur chaque disque.

Mettre `/a` ou `/b` directement sur l'une des deux partitions donnera les mêmes propriétés à tous les sous-répertoires de chacun de ces répertoires, et nous voulons l'éviter. Tailler 4 partitions pour ces 4 répertoires ferait perdre de la flexibilité, nous l'éviterons aussi. La bonne solution est de faire de ces 4 répertoires des liens symboliques vers les bons répertoires de chacun des disques. Ainsi:

```
/a/fast lien symbolique vers /mnt.fastdisk/a.fast
/a/slow lien symbolique vers /mnt.slowdisk/a.slow
/b/fast lien symbolique vers /mnt.fastdisk/b.fast
/b/slow lien symbolique vers /mnt.slowdisk/b.slow
```

Et nous avons tous les répertoires rapides sur le disque rapide sans avoir à faire une partition pour chacun d'entre eux.

Le désavantage est que c'est relativement compliqué et qu'il faut prévoir tous les points de montage, liens symboliques et partitions avant d'installer le système.

9 Implémentation

Les distributions récentes ont des outils qui vous guideront pour le partitionnement et le formatage des disques, et généreront un fichier `/etc/fstab` automatiquement. Mais pour y faire des modifications par la suite, vous devez comprendre les mécanismes que ça met en jeu.

9.1 Disques et Partitions

Avec DOS ou autre vous trouvez toutes les partitions avec des noms comme `C: D:`, sans différenciation pour les disques IDE, SCSI, réseau, etc. Dans le monde de Linux c'est différent. Au démarrage vous verrez un message comme:

```
Dec  6 23:45:18 demos kernel: Partition check:
Dec  6 23:45:18 demos kernel:  sda: sda1
Dec  6 23:45:18 demos kernel:  hda: hda1 hda2
```

Les disques SCSI se nomment `sda`, `sdb`, `sdc` etc, et les disques (E)IDE se nomment `hda`, `hdb`, `hdc` etc. Il y a aussi des noms standards pour tous les périphériques (souris, clavier, disquette, etc), voir `/dev/MAKEDEV` et `/usr/src/linux/Documentation/devices.txt`.

Les partitions sont notées par des numéros sur chaque disque, `hda1`, `hda2`, etc. Sur les disques SCSI il peut y avoir jusqu'à 15 partitions, et sur les disques EIDE drives jusqu'à 63 partitions. Ces deux limites sont bien au-delà de ce qui est utile.

Ces partitions sont montées selon les indication du fichier `/etc/fstab` pour que les fichiers qu'elles contiennent soient accessibles.

9.2 Partitionnement

D'abord vous devez partitionner chaque disque. Sous Linux il y a deux méthodes, `fdisk` et `cfdisk` (plus convivial) (ndT: il y a aussi d'autres outils avec les distributions RedHat ou SuSE). Ces programmes sont complexes, lisez les pages de manuel *très attentivement*. Sous DOS il y a d'autres possibilités, comme `fdisk` ou `fips`. Ce dernier a l'avantage qu'il peut partitionner un disque sans nécessairement écraser toutes les données. Avant de lancer `fips` vous devez défragmenter votre disque. Si vous utilisez `FAT32` vous pouvez utiliser la dernière version de `fips` (à partir de 15c).

Il faudra d'abord défragmenter. Cela mettra toutes les données au début du disque, et l'espace vide restant peut être utilisé pour tailler de nouvelles partitions.

De toute façon, il est indispensable de faire une sauvegarde complète de toutes vos données importantes avant de partitionner.

Il y a trois types de partitions, `primaire`, `étendue` and `logique`. On ne peut démarrer que sur une partition primaire, et le nombre de partitions primaires est limité à 4. Si vous avez besoin de plus de partitions, vous devez définir des partitions étendues, qui contiendront de partitions logiques.

Chaque partition a un numéro qui indique quel système de fichiers elle utilise, pour Linux les seuls types à connaitre sont `swap` et `ext2fs`.

Pour plus d'informations, consuler le fichier README qui vient avec `fdisk` ou le *Partitioning HOWTO*.

RedHat a un utilitaire interactif appelé *Disk Druid* qui est est supposé être une alternative plus conviviale à `fdisk` et automatiser d'autres tâches. Cependant cet outil n'est pas tout à fait mature: s'il ne fait pas ce que vous voulez, utilisez plutôt `fdisk` ou `cfdisk`.

9.3 Disques Multiples (md)

Assurez-vous que vous avez la documentation la plus récente sur cette fonctionnalité du noyau. Ce n'est pas encore stable, vous voilà prévenu.

En bref cela consiste à rassembler des partitions en de nouveaux périphériques `md0`, `md1` etc. en utilisant `mdadd`, puis à les activer avec `mdrun`. Cela peut être automatisé avec le fichier `/etc/mdtab`.

On peut ensuite considérer `md0`, `md1` comme n'importe quel disque. Il y a maintenant un HOWTO sur le RAID avec `md` auquel je vous renvoie pour les détails.

9.4 Formatage

Après le partitionnement vient le formatage, c'est-à-dire l'écriture des structures de données qui permettront de décrire les attributs et la position des fichiers. Si c'est la première fois que vous formatez il est recommandé d'utiliser l'option "verify" ou "check for bad blocks". A strictement parler, c'est inutile, mais cela peut résoudre des problèmes comme la terminaison (pour le SCSI). Voir la documentation de `mkfs` pour les détails.

Linux est compatible avec un nombre impressionnant de systèmes de fichiers. Faire `man fs` pour la liste complète. Notez que votre noyau doit avoir le pilote adéquat pour pouvoir accéder à un système de fichiers. Lors de l'étape de configuration du noyau (`make menuconfig` ou `make xconfig`) vous avez de l'aide en ligne pour chaque système de fichiers et vous pouvez choisir de l'inclure dans le noyau ou d'en faire un module.

Notez que certaines disquettes de sauvetage ont besoin des systèmes de fichiers `minix`, `msdos` et `ext2fs` compilés dans le noyau.

Les partitions de swap (échange) doivent aussi êtres formatées, utilisez `mkswap` pour ça.

9.5 Montage

Les données d'une partitions ne sont pas visibles avant d'êtres montées dans un endroit de l'arborescence appelé point de montage de la partition. Cela est fait à la main avec le programme `mount` ou bien automatiquement durant le démarrage. La liste des partitions avec leur point de montage est dans le fichier `/etc/fstab`. Lisez le manuel de `mount` et faites très attention aux tabulations dans le fichier `/etc/fstab` (elles ne sont pas équivalentes à des espaces).

10 Maintenance

C'est le travail de l'ingénieur système de garder un oeil sur les disques et les partitions. Si une partition est pleine, le système aura des dysfonctionnements, quelle que soit la place libre sur les autres partitions.

Pour voir la liste des partitions actuellement montées, avec le point de montage et le pourcentage de place libre, taper `df`. Cela doit être fait régulièrement, par exemple avec une crontab.

Les partitions de swap peuvent être surveillées avec les outils de statistique de la mémoire comme `free`, `procinfo` ou `top`.

Surveiller l'usage des disques est plus délicat mais c'est important pour les performances. Il faut éviter que le même disque soit sollicité tout le temps quand d'autres sont inactifs.

Il est important quand on installe un logiciel de savoir précisément où vont les fichiers. Ainsi, pour des raisons historiques, GCC qui met des exécutables dans les répertoires de librairie. On peut aussi mentionner X11 dont la structure est très complexe.

Lorsque votre système est au bord de l'asphyxie il est temps de faire la chasse aux fichiers temporaires, fichiers de log, fichiers `core` et autres. Un bon usage de `ulimit` dans les paramètres globaux du shell peut vous aider à éviter d'avoir des fichiers `core` un peu partout.

10.1 Sauvegarde

Le lecteur attentif aura remarqué les allusions répétés à l'utilité des sauvegardes. Les films d'horreur sont nombreux où l'on parle d'accidents et de ce qui est arrivé aux personnes responsables quand la sauvegarde s'est avérée inutilisable, voire inexistante. Il est en général plus simple d'investir dans des moyens de sauvegarde décents que de se trouver une seconde identité ...

Il y a de nombreuses possibilités, et un mini-HOWTO (`Backup-With-MSDOS`) détaille tout ce que vous devez savoir, en plus d'informations spécifiques à MSDOS.

En plus de faire des sauvegardes, vous devez vous assurer que vous pouvez retrouver les données. Les données écrites ne sont pas toujours correctes, et de nombreux administrateurs systèmes ont un jour commencé à restaurer le système après un accident, joyeux à la pensée que tout marchait, lorsqu'ils découvrirent avec horreur que les sauvegardes n'étaient pas utilisables. Soyez prudents.

10.2 Défragmentation

Cela varie beaucoup selon le système de fichiers. Certains souffrent d'une défragmentation rapide et presque paralysante. Heureusement ce n'est pas le cas de `ext2fs` et c'est pourquoi on a très peu parlé des outils de défragmentation. En fait, il en existe, mais il est rare qu'on en aie même le besoin.

Si vous voulez le faire pour une raison ou pour une autre, le moyen simple et rapide est de faire une sauvegarde puis une récupération. Si cela ne concerne qu'une petite partie des fichiers, pas exemple les répertoires utilisateurs, vous pouvez le `tar`-er dans une zone temporaire sur une autre partition, *vérifier* l'archive, effacer l'orignial et le dé-`tar`-er.

10.3 Effacement

Le plus souvent le manque de place est résolu par l'effacement des fichiers inutiles qui s'accumulent. Les programmes qui ne terminent pas normalement laissent toutes sortes de trucs traîner aux endroits les plus bizarres. Normalement un fichier appelé *core* est créé en cas de plantage d'un programme. Il ne sert qu'à deboguer, donc vous pouvez l'effacer si vous ne comptez pas déboguer. Ces fichiers peuvent se trouver n'importe où dont il est recommandé de les chercher de façon globale. (ndT: `find / -name core` devrait marcher)

L'arrêt prématuré des programmes laisse aussi des fichiers temporaires dans des répertoires comme `/tmp` ou `/var/tmp`, fichiers qui auraient été effacés si le programme avait terminé normalement. Ces répertoires sont en général nettoyés au démarrage, mais si vous ne redémarrez jamais ils peuvent finir par être plein de vieux trucs. N'effacez pas les fichiers aveuglément. Des utilitaires comme `find` et `file` peuvent vous servir à localiser les fichiers plus vieux que telle date et à connaître le type d'un fichier.

Beaucoup de choses sont archivés lorsque le système fonctionne, en particulier dans le répertoire `/var/log`. Les messages du noyau sont mis dans `/var/log/messages` qui a une certaine tendance à grossir avec le temps. Il peut être bon d'avoir une petite archive de ce fichier pour pouvoir le comparer avec les messages du noyau si le système commence à se comporter bizarrement.

Si le courrier ou les news ne fonctionnent pas correctement, c'est peut-être dû à une croissance excessive de `/var/spool/mail` et `/var/spool/news`. Faites attention aux fichiers dont le nom commence par ".", il ne sont pas affichés par `ls -l`, c'est pourquoi on recommande d'utiliser plutôt `ls -Al`.

Le dépassement de capacité des répertoires utilisateurs est une question délicate. De véritables guerres ont déjà eu lieu entre utilisateurs et administrateurs à ce sujet. Le tact, la diplomatie et un budget généreux pour de nouveaux disques sont les solutions. En utilisant le mot-du-jour, un petit message dans le fichier `/etc/motd` qui est affiché chaque fois qu'un utilisateur se loggue, on peut sensibiliser les utilisateurs. Mettre les bonnes valeurs par défaut pour empêcher les fichiers `core` d'être produits épargne bien du travail.

Certaines personnes essayent de cacher les fichiers, en utilisant le fait que les fichiers dont le nom commence par un point ne sont pas visibles pour la comande `ls`. Un exemple classique est `...` qui n'est donc pas vu par `ls` et passe inaperçu à côté de `.` et `..` si on fait `ls -al`. La solution est de faire `ls -Al` qui affiche tous les fichiers sauf `.` et `..`

10.4 Mises à jour

Quelle que soit la taille de vos disque, ce sera un jour trop petit. Actuellement ce sont les disques de 6.4 Go qui offrent le meilleur rapport place/prix.

Avec des disques IDE vous aurez peut-être à enlever un vieux disque, le nombre total étant limité à 2 ou 4. Avec le SCSI vous pouvez avoir jusqu'à 7 disques en 8-bit et 15 en 16-bit (wide SCSI) par canal. Mais certains adaptateurs ont plusieurs canaux et qu'on peut mettre plusieurs adaptateurs. Mon point de vue est qu'avec le SCSI on est plus content sur le long terme.

Et maintenant la question bateau, que faire de ce nouveau disque ? Souvent c'est pour étendre les queues qu'on a dû étendre, donc la solution simple est de monter les nouveaux disques dans `/var/spool`. D'un autre côté les nouveaux disques étant plus rapides, c'est peut-être l'occasion de revoir tout en profondeur.

Si la mise à jour est rendue indispensable par le manque de place dans `/usr` ou `/var` elle est un peu plus complexe. Vous pouvez envisager la réinstallation complète de la toute dernière version de votre distribution préférée. Dans ce cas faites très attention à ne pas écraser vos réglages essentiels. Les fichiers de configuration sont pour la plupart dans le répertoire `/etc`. Procéder avec soin, avec une sauvegarde récente et des disquettes de sauvetage qui marchent. Une autre possibilité que la réinstallation est de simplement copier le vieux répertoire vers le nouveau, qui est monté sur un point de montage provisoire. Puis éditer le fichier `/etc/fstab` pour que le chemin du répertoire pointe vers la nouveau, et redémarrez. Si le démarrage échoue, vous pouvez redémarrer avec une disquette de secours, éditer à nouveau `/etc/fstab` et réessayer.

Tant qu'il n'y aura pas de logiciel de gestion de volume pour Linux ça restera à la fois complexe et dangereux. Ne soyez pas surpris si vous découvrez que vous devez restaurer le système d'après une sauvegarde.

Le Tips-HOWTO donne l'exemple suivant pour déplacer toute une structure de répertoire:

```
(cd /source/directory; tar cf - . ) | (cd /dest/directory; tar xvfp -)
```

Ça marchera sur la plupart des systèmes Unix. Attention aux répertoires dont la structure arborescente est trop profonde, elle peut faire échouer un tar autre que GNU tar.

Si vous avez accès à GNU cp (c'est toujours le cas sous Linux) vous pouvez aussi bien utiliser

```
cp -av /source/directory /dest/directory
```

GNU cp sait se débrouiller avec les liens symboliques, les FIFO et les fichiers de périphériques et les copier correctement.

Rappelez-vous que ce n'est jamais une bonne idée de transférer `/dev` ou `/proc`

11 Utilisation avancée

Linux et ses cousins offrent de nombreuses possibilités pour une destruction rapide et efficace du système. Ce document n'y fait pas exception. Avec le savoir vient le pouvoir et donc le danger, et les paragraphes qui suivent présentent des sujets plus ésotériques qui ne devraient pas être abordés avant d'avoir lu et compris la documentation et les pièges. Vous devriez faire une sauvegarde, et essayer au moins une fois d'écraser et de restaurer complètement votre système. Sinon vous ne serez pas le premier à avoir une superbe sauvegarde et rien pour la réinstaller (ou, encore plus gênant, des fichiers essentiels manquent sur la bande).

Les techniques décrites ici sont rarement utiles mais servent à des installations particulières. Pensez sérieusement à ce que vous voulez faire avant d'aller plus loin.

11.1 Paramètres du disque dur

Les paramètres physiques du disque dur peuvent être changés avec l'utilitaire `hdparms`. Le paramètre le plus intéressant est sans doute *read-ahead* qui détermine combien de bits on doit lire d'avance en lecture séquentielle.

Ce qui fait le plus de sens est de sélectionner la longueur moyenne des fichiers. Mais cette moyenne pour tout un disque physique peut être non significative. Probablement cela n'est utile que sur les disques spécialisés dans les news ou le courrier électronique des grands serveurs.

Pour des raisons de sécurité les valeurs par défaut de hdparm sont plutôt conservateurs. L'inconvénient est que vous pouvez avoir des interruptions qui se perdent si vous avez des IRQ à grande fréquence comme lorsqu'on utilise le port série et un disque IDE en même temps, les IRQ du disque vont masquer les autres. Ce qui entraîne des performances tout sauf optimales lors du téléchargement sur Internet. Sélectionner `hdparm -u1 device` enlèvera ce masquage et même améliorera vos performances, ou bien endommagera les données du disque. A essayer avec prudence et avec des sauvegardes récentes.

11.2 Paramètres du système de fichiers

La plupart des systèmes de fichiers viennent avec un utilitaire de configuration: ainsi `tune2fs` pour `ext2fs`. On peut jouer avec plusieurs paramètres, mais le plus utile est peut-être la taille qu'on peut réserver. Cela peut vous aider à avoir plus d'espace utile sur vos disques. En revanche vous aurez moins de place pour réparer le système s'il crashe.

11.3 Synchronisation des axes

Cela ne devrait pas être dangereux en soi, mis à part que les détails exacts des connections ne sont pas bien connus pour beaucoup de disques. La théorie est simple: garder une différence de phase fixe entre les différents disques d'un ensemble RAID. Cela diminue le temps d'attente pour que la bonne piste soit en position pour la tête de lecture/écriture. En pratique , avec de grands tampons pour la lecture d'avance, le gain est négligeable.

La synchronisation des axes ne doit pas être utilisée dans un ensemble RAID0 ou RAID 0/1 car on perdrait le bénéfice d'avoir les têtes de lectures sur des emplacements différents.

12 Pour plus d'information

Il y a pas mal d'information disponible pour ceux qui mettent en place un grand système, par exemple les fournisseurs d'accès à Internet. Les FAQs des forums suivants sont utiles:

12.1 Forums

Parmi les plus intéressants:

- *Storage* news:comp.arch.storage .
- *PC storage* news:comp.sys.ibm.pc.hardware.storage .
- *AFS* news:alt.filesystems.afs .
- *SCSI* news:comp.periphs.scsi .
- *Linux setup* news:comp.os.linux.setup .
- *Linux (francophone)* news:fr.comp.os.linux .

La plupart des forums ont leur propre FAQ destinée à répondre aux questions les plus courantes, comme le nom de Foire Aux Questions l'indique. Si vous ne les trouvez pas dans la queue des news vous pouvez aller directement à

l'archive FTP des principales FAQs ftp://rtfm.mit.edu . La version hypertexte se trouve à

l'archive HTTP des principales FAQs http://www.cis.ohio-state.edu/hypertext/faq/usenet/FAQ-List.html .

Certaines FAQs ont leur propre site, en particulier

- *la FAQ SCSI* http://www.paranoia.com/~filipg/HTML/LINK/F_SCSI.html et
- *la FAQ de comp.arch.storage* http://alumni.caltech.edu/~rdv/comp_arch_storage/FAQ-1.html .

12.2 Mailing lists

Ce moyen de communication destiné aux développeurs a un bon rapport signal/bruit. Repensez-y à deux fois avant de poser des questions sur les mailing-lists car le bruit ralentit l'effort de développement. Parmi les listes qui nous concernent, `linux-raid`, `linux-scsi` et `linux-ext2fs`. La plupart des mailing lists intéressantes sont sur le serveur `vger.rutgers.edu`, mais il est vraiment surchargé, essayez plutôt un mirroir. Il y a un miroir de quelques listes sur *le site de Redhat* http://www.redhat.com . La plupart des listes sont aussi accessibles sur le site

Linux HeadQuarters http://www.linuxhq.com/lnxlists , et le reste de la toile est une mine d'or pour les informations.

Si vous voulez en savoir plus sur les listes existantes vous pouvez envoyer un message au *serveur de listes de vger.rutgers.edu* mailto:majordomo@vger.rutgers.edu

donc le corps contiendra le seul mot "lists". Si vous voulez savoir comment marche une mailing list envoyez un message avec le seul mot `help` à la même adresse. A cause du succès de ce serveur il est possible que la réponse prenne un certain temps.

Il y a aussi un certain nombre de serveurs majordomo intéressants, comme

la liste des pilotes EATA mailto:linux-eata@mail.uni-mainz.de

et la

liste des entrées/sorties intelligentes mailto:linux-i2o@dpt.com .

Les mailing lists évoluent rapidement mais un certain nombre de listes intéressantes sont sur

la page du Linux Documentation Project http://sunsite.unc.edu/LDP .

12.3 HOWTO

Ce sont les premières sources d'information générale, mais on y trouve aussi la solution à bien des problèmes spécifiques. Les HOWTOs apparentés à celui-ci sont `Bootdisk`, `Installation`, `SCSI` et `UMSDOS`. le site principal en anglais est

l'archive du LDP sur sunsite `http://sunsite.unc.edu/LDP` . Le mirroir en France (qui contient aussi la traduction des HOWTOs en français) est

Freenix `http://www.freeenix.fr` .

Il y a un nouveau HOWTO qui parle de la mise en place d'un système RAID DPT, voir

the DPT RAID HOWTO homepage `http://www.ram.org/computing/linux/dpt_raid.html` .

12.4 Mini-HOWTO

Parmi ceux qui nous concernent: `Backup-With-MSDOS`, `Diskless`, `LILO`, `Linux+DOS+Win95+OS2`, `Linux+OS2+DOS`, `Linux+Win95`, `NFS-Root`, `Win95+Win+Linux`, `ZIP Drive`.

On les trouve aux mêmes endroits que les HOWTOs.

Le vieux `Linux Large IDE mini-HOWTO` est obsolète, lisez plutôt `/usr/src/linux/drivers/block/README.ide` ou `/usr/src/linux/Documentation/ide.txt` (ces fichiers font partie de la documentation des sources du noyau).

12.5 Documentation locale

Le plupart des distributions Linux ont un

répertoire de documentation `file:///usr/doc`

où l'on trouve souvent un sous-répertoire

un sous-répertoire pour les HOWTOs `file:///usr/doc/HOWTO`

Les fichiers de configuration mentionnés plus haut sont dans le répertoire */etc* `file:///etc` . En particulier

/etc/fstab `file:///etc/fstab`

pour les points de montage et

mdtab `file:///etc/mdtab`

qui est utilisé pour la configuration du RAID.

La documentation des

sources de linux `file:///usr/src/linux`

est bien sûr la source ultime d'information. Pas seulement avec les commentaires qui sont dans le code mais aussi avec le

répertoire de documentation `file:///usr/src/linux/Documentation` . Si vous vous posez une question au sujet du noyau vous devez d'abord chercher là.

Les fichiers où sont stockés

les messages du noyau `file:///var/log/messages`

permettent de savoir ce qui se passe, en particulier si les messages ont défilé trop vite au démarrage. Avec la commande `tail -f /var/log/messages` dans une fenêtre ou un écran séparé, vous aurez une information toujours à jour sur ce qui se passe dans votre système.

Vous pouvez aussi utiliser le système de fichiers

/proc `file:///proc`

qui donne de l'info en temps réel sur le système. Utiliser `cat` plutôt que `more` pour voir ces fichiers car leur longueur déclarée est zéro.

Tout est basé ici sur le Filesystem Structure Standard (FSSTND). Il est en train de changer de nom pour devenir File Hierarchy Standard (FHS) et être moins propre à Linux. Il y a une

page Web du FHS `http://www.pathname.com/fhs`

qui explique comment rejoindre la mailing list privée des développeurs.

12.6 Pages WWW

Il y a un grand nombre de pages Web intéressantes, et elles bougent beaucoup, ne soyez pas étonnés si ces liens deviennent obsolètes.

Un bon point de départ est sur Sunsite: c'est

l'archive du Linux Developpement Project `http://sunsite.unc.edu/LDP/`

- Mike Neuffer, l'auteur du cache caching et des pilotes pour contrôleurs RAID, a des pages intéressantes sur

  SCSI `http://www.uni-mainz.de/~neuffer/scsi`

  et

  DPT `http://www.uni-mainz.de/~neuffer/scsi/dpt` .

- Sur le développement du RAID 1 logiciel, voir la

  page des développeurs RAID 1 `http://www.nuclecu.unam.mx/~miguel/raid` .

- Sur (entre autres) la mesure de performances, le RAID, la fiabilité, voir la page du projet

  Linas Vepstas `http://linas.org` .

- Il y a aussi un HOWTO sur

  comment avoir en RAID la partition racine `ftp://ftp.bizsystems.com/pub/raid/Root-RAID-HOWTO.html` .

- Voir enfin ici pour la documentation détaillée de

  ext2fs `http://step.polymtl.ca/~ldd/ext2fs/ext2fs_toc.html` .

- Mark D. Roth a une page sur

  VPS `http://www.uiuc.edu/ph/www/roth`

- Un projet similaire:

  Enhanced File System `http://www.virtual.net.au/~rjh/enh-fs.html`

- Il y a un projet de compression qui s'intégrerait à `ext2fs` et s'appelle `e2compr`. Voir

  la maison-page de e2compr `http://netspace.net.au/~reiter/e2compr.html` .

- Pour plus d'information sur le démarrage et sur BSD voir

  ici `http://www.paranoia.com/~vax/boot.html`

  page.

On trouve des tableaux sur les disques, les contrôleurs, etc. à la page appelée *The Ref* `http://theref.c3d.rl.af.mil` . On peut l'interroger en ligne ou télécharger la base de données par

FTP `ftp://theref.c3d.rl.af.mil/public` .

12.7 Moteurs de recherche

N'oubliez pas que vous pouvez utiliser les moteurs de recherche, comme:

- *Altavista* `http://www.altavista.digital.com`
- *Excite* `http://www.excite.com`
- *Hotbot* `http://www.hotbot.com`

Il y a aussi

Dejanews `http://www.dejanews.com` , dédié à la recherche dans les news, qui archive les forums depuis 1995.

Si vous voulez de l'aide vous posterez sans doute dans le forum

Linux Setup `news:comp.os.linux.setup`

(ndT: Pour les francophones consulter plutôt le

forum français sur Linux `news:fr.comp.os.linux`)

13 Comment obtenir de l'aide

Il se peut que, dans l'incapacité à résoudre vos problèmes par vous-même, vous ayez besoin d'aide. Le moyen le plus sûr est de demander à quelqu'un dans le groupe d'utilisateurs Linux le plus proche de chez vous.

Une autre possibilité est de poster dans les news. Le problème est que le rapport signal/bruit des newsgroups est parfois faible et votre question peut très bien passer inaperçue.

Quel que soit l'endroit où vous demandez, il est important de bien poser la question. Dire juste *mon disque dur ne marche pas* ne risque pas de vous aider: au mieux, quelqu'un vous demandera d'être plus précis.

Il est recommandé de décrire le problème avec assez de détails pour permettre aux gens de vous aider. Il peut se produire là où vous vous y attendez le moins. Voilà pourquoi il faut décrire:

Matériel

- Le Processeur
- Le chipset (LX, BX, etc)
- Le bus (ISA, VESA, PCI etc)
- Les cartes d'extension (carte graphique, etc.)

Logiciel

- La version du BIOS (Pour la carte-mère et éventuellement les adaptateurs SCSI)
- LILO, s'il est utilisé
- La version du noyau et les patchs ou modifications éventuels
- Les paramètres du noyau (s'il y en a)

- Les programmes qui font apparaître l'erreur (avec numéro de version)

Périphériques

- Type du disque, fabriquant, version et modèle.
- Les autres périphériques présents sur le même bus.

Un exemple d'inter-relation de ces différents éléments: on a déjà vu un vieux chipset qui cause des problèmes si on utilise certaines combinaisons de carte graphique et d'adaptateur SCSI.

Joindre à votre message un extrait (bref) du contenu de `/var/log/messages` peut être utile (mais parfois regarder ce contenu suffit à détecter la source du problème). Bien sûr si le disque est en panne il est possible que ces messages ne soient pas enregistrés, mais on peut au moins scroller en arrière avec les touches `SHIFT` et `PAGE UP`.

14 Remarques en guise de conclusion

La configuration des disques et le choix des partitions sont difficiles, et on n'a pas donné de règles fixes ici. Cependant, y travailler un peu peut apporter des gains considérables. Maximiser l'usage d'un seul disque quand les autres sont inactif est loin d'être optimal, regardez les LED, elles ne sont pas là que pour la décoration. Avec un système bien fait, les petites diodes qui indiquent l'activité des disques doivent clignoter comme des lampes de discothèque. Linux permet le RAID au niveau logiciel mais supporte aussi quelques contrôleurs RAID SCSI. Vérifiez ce qui est disponible. Plus tard, si vous re-partitionnez votre système, vous pourrez jetez à nouveau un oeil à ce document. Les commentaires et les contributions sont bienvenus.

14.1 En préparation

Il y a encore quelques sujets qui vont apparaître ici. En particulier je vais ajouter d'autres exemples de tables pour la configuration de grands réseaux. Des exemples de réseaux marchant sans problème sont les bienvenus.

Il reste aussi un peu de boulot dans ce HOWTO sur les systèmes de fichiers et utilitaires.

Une grande section sera ajoutés sur les technologies de disque dur ainsi qu'une meilleure description sur l'utilisation de `fdisk` or `cfdisk`. La section sur les systèmes de fichiers se remplira au fur et à mesure que les nouveautés sortiront.

J'ai reçu récemment une plaquette de DPT, qui fabrique le premier système RAID hardware supporté par Linux. Leurs feuillets portent maintenant le petit pingouin Linux. Bientôt plus d'information à ce sujet.

Il y a quelques petits passages qui font double emploi avec le Filesystem Hierarchy Standard. Les enlever signifiera probablement un remaniement complet des tables de la fin de ce document.

J'envisage aussi d'écrire un programme qui automatiserait le processus de décision, en donnant un point de départ simple et plus complet.

14.2 Demande d'information

Ecrire ce document a pris un certain temps et bien qu'il commence à ressembler à quelque chose, ce document a encore besoin d'information que seul vous, précieux lecteurs, pouvez m'apporter.

- Plus d'information sur la taille de swap nécessaire et la plus grande taille de swap autorisée avec les différentes versions du noyau.

- Est-ce qu'il est fréquent qu'un disque soit abîmé ou qu'un système de fichier soit corrompu ? Autant que je me souvienne, je n'ai jamais connu que des problèmes dûs à du matériel défectueux.
- J'ai aussi besoin de documentation sur la vitesse comparée des disques.
- Y a-t-il d'autres contrôleurs RAID compatibles avec Linux ?
- Des pistes quant aux systèmes de fichiers, à la gestion de volumes et assimilés sont bienvenues.
- Quels utilitaires dignes d'intérêt sont disponibles ?
- Il faudrait aussi une liste complète des sources d'information. Peut-être sur un document séparé ?
- L'usage de `/tmp` et `/var/tmp` est difficile à déterminer, en fait savoir quels programmes utilisent quel répertoire n'est pas évident, plus d'information à ce sujet est bienvenue. Cependant, il reste clair que ces deux répertoires doivent être sur des disques différents pour profiter du parallélisme.

14.3 Suggestions pour participer à un projet.

Sur les forums comp.os.linux.* on trouve plein de bonnes idées. Je vais en lister ici quelques-uns en rapport avec notre sujet. Les projets ambitieux comme un nouveau système de fichiers doivent toujours être postés soit pour trouver des collaborateurs soit pour voir si quelqu'un ne travaille pas déjà dessus.

Des outils de Planning

qui automatisent la conception d'un système constituent un projet de taille moyenne. Une sorte d'exercice en programmation par contraintes.

Des outils de partitionnement

qui acceptent en entrée le résultat du programme mentionné ci-dessus et formatent les disques en parallèle puis créent l'arborecence de fichiers avec les bons liens symboliques. Ce serait encore mieux si on intégrait ça à des programmes d'installation existants. Le programme d'installation de Solaris est un bon exemple à méditer.

Des outils de surveillance

qui surveillent les partitions et tirent la sonnette d'alarme avant qu'elles soit pleines.

Des outils de migration

qui permettent de déplacer sans danger des arborescences entières (par exemple pour migrer vers un système RAID). Ce serait par exemple un script shell assez simple contrôlant un programme de sauvegarde. Cependant, veillez à ce qu'il soit sécurisé et qu'il permette de revenir en arrière.

15 Questions / Réponses

Voici quelques questions fréquentes et leur réponse.

- Q: De combien de disque dur Linux a besoin ?
- R: Linux marche très bien avec un seul disque dur. Avoir assez de mémoire vive (32 ou 64 Mo) est un meilleur choix point de vue performances que d'acheter un second disque. Les disques IDE sont moins chers, mais aussi moins rapides que les SCSI.
- Q: J'ai un seul disque, est-ce que ce HOWTO est fait pour moi ?

- R: Oui, mais en partie seulement. Voir la section sur le 4.3.8 (Positionnement physique des pistes).
- Q: Y a-t-il des désavantages dans ce cas ?
- R: Un seul petit désavantage. Si une partition n'a plus de place libre, le système peut se bloquer ou se comporter bizarrement. La gravité dépend bien sûr de la partition affectée. Cependant, ce n'est pas difficile à contrôler, avec la commande `df` qui donne une vue générale de la situation. Utiliser aussi la commande `free` pour s'assurer que la mémoire virtuelle (c'est-à-dire la mémoire vive + le swap) est suffisante.
- Q: OK, je dois donc séparer mon système entre autant de partitions que possible pour un seul disque ?
- R: Non, car cela a plusieurs désavantages. D'abord la maintenance est plus complexe et le gain peut être mineur. Des partitions trop grandes ne sont pas non plus l'idéal. Il y a un juste milieu, qui dépend du nombre de disques que vous avez.
- Q: Est-ce que cela veut dire que plus de disques permettent d'avoir plus de partitions ?
- R: En un sens, oui. Cependant, certains répertoires ne doivent pas être séparés de la racine (voir le FHS pour les détails)
- Q: Et si j'ai beaucoup de disques ?
- R: Si vous avez plus que 3 ou 4 disques vous devriez penser à utiliser un des modes RAID. Cependant, il vaut mieux garder la partition root sur une partition simple (sans RAID), voir la section sur le 4.3.1 (RAID) pour les détails.
- Q: J'ai installé le dernier Windows95 mais je n'arrive pas à accéder aux partitions Windows depuis Linux.
- R: Sans doute votre partition Windows est formatée en `FAT32`. C'est le cas pour Windows 95 OSR2 et pour Windows 98. Linux a un support pour ce système de fichiers depuis le noyau 2.0.35.
- Q: Je n'arrive pas à faire correspondre la somme des taille de mes disques et celle de mes partitions.
- R: Il est possible que vous ayez monté une partition en un point qui n'était pas un répertoire vide. Le point de montage est un répertoire et s'il n'est par vide le montage masquera son contenu. Et en faisant la somme vous verrez qu'il manque la place occupée par le contenu de ce répertoire avant montage.

  Pour résoudre ça vous pouvez démarrer depuis une disquette de sauvetage et voir ce qui se cache derrière les points de montage. Vous pouvez ensuite effacer ou transférer ces données en montant la partition en question sur un point de montage temporaire.
- Q: Qu'est ce que ce nyx qui est mentionné plusieurs fois dans ce HOWTO ?
- R: C'est un grand système utilisant les Unix libres et avec 10000 utilisateurs. Je m'en sers pour héberger mes pages Web mais aussi comme source d'inspiration pour ce HOWTO, en ce qui concerne la configuration de réseaux assez vastes. Voir la

  page d'accueil de Nyx `http://www.nyx.net`

  qui indique aussi comment obtenir un compte gratuit.

16 Bric-à-brac

C'est une section où vont tous les paragraphes que je n'ai pas pu caser ailleurs: ils y restent plus ou moins longtemps.

16.1 Combiner le swap et /tmp

On a discuté dans les forums linux au sujet de systèmes de fichiers spécialisés pour le stockage temporaire. Un peu comme `tmpfs` sur les machines *BSD et Solaris, et `swapfs` sur les machines NeXT.

Combiner le `swap` et la partition `/tmp` permet de gagner de la place. Ce système de fichiers spécialisé n'est rien d'autre qu'un RAM disk qu'on peut swapper, et qui n'est mis sur le disque que lorsque la place est limitée, ce qui revient à metter les fichiers temporaires sur la partition de swap.

Il y a pourtant un hic. Ce schéma interdit d'agir en parallèle sur le `swap` et sur la partition `/tmp` ce qui peut effondrer les performances. Autrement dit, on échange de la place disque contre de la vitesse.

Il y a aussi un problème de sécurité vis-à-vis des utilisateurs qui tentent d'effondrer une machine en remplissant le répertoire `/tmp`.

16.2 Disques de swap entrelacés.

Les partitions de swap sont accédées par la méthode du colibri (c'est-à-dire dans le désordre), afin de répartir grosso modo la charge entre plusieurs disques. Linux offre en plus la possibilité d'attribuer des priorités aux disques, ce qui est utile si on a des disques de vitesse différente. Voir `man 8 swapon` et `man 2 swapon` pour les détails.

16.3 Faut-il avoir ou non une partition de swap ?

Dans de nombreux cas vous n'avez pas besoin d'une partition, par exemple si vous avez beaucoup de mémoire vive, mettons 64 Mo, et si vous êtes le seul utilisateur de la machine. Dans ce cas vous pouvez essayer de tourner dans partition de swap et voir (par exemple avec les rapports du système ou avec la commande `top`) s'il y a des moments où vous n'avez plus de mémoire libre.

Enlever la partition de swap a deux avantages:

- Vous gagnez de la place disque
- Vous gagnez sur le temps moyen d'accès car la partition de swap aurait occupé le milieu du disque (qui est plus rapide)

Au total, avoir une partition de swap est comme avoir des toilettes chauffées: on n'en a pas besoin la plupart du temps, mais c'est bien agréable parfois. (ndT: Ah qu'en termes galants ces choses-là sont mises !)

16.4 Points de montage et /mnt

Dans une ancienne version de ce document, je proposais de mettre toutes les partitions montées sur des sous-répertoires de `/mnt`. C'est cependant une mauvaise idée car `/mnt` lui-même peut être utilisé comme point de montage, ce qui rend toutes les autres partitions inaccessibles. (voir Questions et Réponses). Je propose plutôt de monter les partitions directement dans la racine avec des noms comme `/mnt.nom-bien-choisi`.

Certaines distributions Linux utilisent des points de montage comme `/mnt/floppy` et `/mnt/cdrom` ce qui montre bien combien les choses sont peu claires. Espérons que le FHS mettra de l'ordre dans tout ça.

16.5 SCSI: numéros et noms symboliques

Les partitions sont nommées dans l'ordre où elles sont trouvées, et ne dépendent pas du numéro SCSI. Cela signifie que si vous ajoutez un disque avec un numéro intermédiaire, ou si vous changez les numéros d'une autre manière, les noms de partitions sont intervertis et ne correspondent plus à rien. C'est important si vous utilisez des disques amovibles. Dans ce cas il faut réserver les premiers numéros aux disques fixes et les derniers pour les media amovibles.

Beaucoup se sont fait avoir par cette "feature" et réclament qu'on fasse quelque chose. Personne ne sait quand ce sera fixé. Pour l'instant, donc, il faut faire avec. Par exemple c'est une bonne idée de mettre le disque contenant la partition racine au premier numéro SCSI. Ainsi il ne sera pas re-numéroté si un autre disque a une panne.

Le coeur du problème est le nombre limité de bits disponibles pour les numéros majeurs et mineurs des fichiers du répertoire `/dev` utilisés pour décrire le device lui-même. Voir `man MAKEDEV`. Actuellement deux solutions sont envisagées:

scsidev

crée une base de données avec les disques et l'endroit où ils sont, voir `man scsifs`.

devfs

est un projet à plus long terme, qui veut contourner tout la numérotation des fichiers de périphériques en faisant du répertoire `/dev` un répertoire dy noyau tout comme `/proc`. A suivre.

Les numéros SCSI sont aussi utilisés pour l'arbitrage. Si plusieurs disques demandent un service, le disque qui a le numéro le plus faible a la priorité.

16.6 Consommation et Chaleur

Il n'y a pas si longtemps, une machine de puissance équivalente à un PC d'aujourd'hui consommait du courant triphasé, et exigeait un refroidissement à air ou même à eau. La technologie a progressé très vite, offrant des composants rapides mais aussi peu gourmands en énergie. Cependant, il y a des choses qu'on doit garder en tête avant d'ajouter à l'ordinateur un disque ou une carte PCI. Gardez à l'esprit que l'énergie consommée va bien quelque part, et que la plupart est transformée en chaleur. Si la chaleur n'est pas dissipée, il en résultera une surchauffe qui diminue la fiabilité et la durée de vie des composants. Les constructeurs ont de exigences de refroidissement, en termes de mètres cubes par minute, et on ne saurait trop conseiller d'en tenir compte.

Gardez des passages pour l'air, nettoyez la crasse et vérifiez la température des disques. S'il sont brûlants au toucher, c'est sans doute qu'ils sont en surchauffe.

Si possible utilisez l'accélération séquentielle (*sequential spin-up*) pour les disques. C'est l'accélération qui consomme le plus d'électricité et si tous les disques démarrent en même temps vous risquez de dépasser la puissance fournie par votre alimentation.

16.7 Dejanews

C'est un système que la plupart conaissent déjà. Il permet d'effectuer des recherches parmi les articles postés dans les *forums Usenet* depuis 1995 jusqu'à maintenant, et offre aussi une interface Web pour lire et poster des articles. Voir

Dejanews `http://www.dejanews.com`

Ce qui est sans doute moins connu est qu'ils utilisent 120 stations Linux parallèles, la plupart utilisant le module `md` pour gérer 4 et 24 Go d'espace disque (plus de 1200 Go au total). L'ensemble grandit sans cesse mais actuellement il est essentiellement constitué de Bi-Pentium Pro 200MHz et de Bi-Pentium II 300 MHz avec 256 Mo de mémoire vive ou plus.

Une machine de la base de données a normalement 1 disque pour le système d'exploitation et entre 4 et 6 disques gérés par `md` où les articles sont archivés. Les disques sont connectés à un adaptateur PCI SCSI (BusLogic Modèle BT-946C ou BT-958), un par machine.

Les erreurs disque ne constituent que 0,25% des indisponibilités du système.

Enfin, ce n'est pas de la publicité que je fais, mais juste un example de ce qu'il faut pour mettre en place un service Internet majeur. (ndT: le site `voila.fr` de France Télécom utilise un nombre comparable de stations Linux pour un moteur de recherche)

16.8 Structure de la hiérarchie des fichiers

Il y a beaucoup de schémas pour les hiérarchies de fichiers, qui diffèrent du FHS par la philosophie, la stratégie et l'implémentation. Il n'est pas possible de les détailler ici, le lecteur est renvoyé à `man hier` qui est disponible sur beaucoup d'architectures.

16.9 Numérotation des pistes et optimisation

Autrefois les systèmes de fichiers utilisaient les paramètres physiques du disque pour optimiser les transferts, par exemple en essayent de metter tout un fichier dans la même piste afin d'économiser les temps du changement de piste. Aujourd'hui avec les paramètres logiques, le cache et les schémas pour éviter les secteurs défectueux, ce genre d'optimisation ne fait plus de sens et peut même coûter plus cher qu'elle ne rapporte. Certains systèmes d'exploitation utilisent encore ce genre d'algorithmes, mais plus Linux.

17 Appendice A: Partitionnement: points de montage et liens symboliques

La table suivante fait de la conception un simple exercice avec un crayon et un papier. Il est conseillé de l'imprimer (avec des fontes à casse fixe) et d'ajuster les nombres jusqu'à obtenir satisfaction.

Le point de montage est le répertoire sous le nom duquel vous voulez accéder à une partition ou périphérique. Cette table est aussi l'endroit idéal pour noter les liens (ou raccourcis) que vous établirez. La taille correspond à une installation assez complète de Debian 1.3.

```
Répertoire       Point de montage  vitesse temps   taux       taille
                                           moyen   de
                                           d'accès transfert



swap             __________        ooooo   ooooo   ooooo      (32) ____

/                __________        o       o       o          (20) ____

/tmp             __________        oooo    oooo    oooo            ____

/var             __________        oo      oo      oo         (25) ____
/var/tmp         __________        oooo    oooo    oooo            ____
/var/spool       __________                                        ____
/var/spool/mail  __________        o       o       o               ____
/var/spool/news  __________        ooo     ooo     oo              ____
/var/spool/____  __________        ____    ____    ____            ____
```

```
/home            __________        oo      oo      oo               ____

/usr             __________                                   (500)____
/usr/bin         __________        o       oo      o          (250)____
/usr/lib         __________        oo      oo      ooo        (200)____
/usr/local       __________                                        ____
/usr/local/bin   __________        o       oo      o               ____
/usr/local/lib   __________        oo      oo      ooo             ____
/usr/local/____  __________                                        ____
/usr/src         __________        o       oo      o          (50) ____

DOS              __________        o       o       o               ____
Win              __________        oo      oo      oo              ____
NT               __________        ooo     ooo     ooo             ____

/mnt._________   __________        ____    ____    ____            ____
/mnt._________   __________        ____    ____    ____            ____
/mnt._________   __________        ____    ____    ____            ____
/_____________   __________        ____    ____    ____            ____
/_____________   __________        ____    ____    ____            ____
/_____________   __________        ____    ____    ____            ____
/_____________   __________        ____    ____    ____            ____


Espace disque total :                                              ____
```

18 Appendice B: Partitionnement: emplacement des partitions

Ici vous choisirez dans quel disque va chacune des partitions de la table précédente, en gardant à l'esprit les remarques dans la section 4.3.8 (postition physique des pistes).

```
  Disque           sda     sdb     sdc     hda     hdb     hdc     ___

No SCSI          |  __   |  __   |  __   |

Répertoire

swap             |       |       |       |       |       |       |

/                |       |       |       |       |       |       |

/tmp             |       |       |       |       |       |       |

/var             :       :       :       :       :       :       :
/var/tmp         |       |       |       |       |       |       |
/var/spool       :       :       :       :       :       :       :
/var/spool/mail  |       |       |       |       |       |       |
/var/spool/news  :       :       :       :       :       :       :
/var/spool/____  |       |       |       |       |       |       |

/home            |       |       |       |       |       |       |
```

```
/usr            |       |       |       |       |       |       |
/usr/bin        :       :       :       :       :       :       :
/usr/lib        |       |       |       |       |       |       |
/usr/local      :       :       :       :       :       :       :
/usr/local/bin  |       |       |       |       |       |       |
/usr/local/lib  :       :       :       :       :       :       :
/usr/local/____ |       |       |       |       |       |       |
/usr/src        :       :       :       :

DOS             |       |       |       |       |       |       |
Win             :       :       :       :       :       :       :
NT              |       |       |       |       |       |       |

/mnt.___/_____  |       |       |       |       |       |       |
/mnt.___/_____  :       :       :       :       :       :       :
/mnt.___/_____  |       |       |       |       |       |       |
/_____________  :       :       :       :       :       :       :
/_____________  |       |       |       |       |       |       |
/_____________  :       :       :       :       :       :       :


Place totale:
```

19 Appendice C: Partitionnement: numérotation

Cette troisième table sert juste à trier les partitions en attribuant un numéro à chacune, sous la forme attendue par `fdisk`. Ici vous pouvez tenir compte de la position physique des pistes pour l'optimisation.

Ces numéros seront utilisés pour mettre à jour les tables précédentes: les trois tables sont très utiles pour la maintenance.

En cas de crash disque, vous trouverez utile de savoir quel numéro SCSI correspond à quel disque, gardez en conséquence une copie papier de cette information.

```
        Disque:  sda     sdb     sdc     hda     hdb     hdc      ___

Taille totale:  |  ___  |  ___  |  ___  |  ___  |  ___  |  ___  |  ___
No SCSI         |  __   |  __   |  __   |

Partition

1               |       |       |       |       |       |       |
2               :       :       :       :       :       :       :
3               |       |       |       |       |       |       |
4               :       :       :       :       :       :       :
5               |       |       |       |       |       |       |
6               :       :       :       :       :       :       :
7               |       |       |       |       |       |       |
8               :       :       :       :       :       :       :
9               |       |       |       |       |       |       |
10              :       :       :       :       :       :       :
11              |       |       |       |       |       |       |
12              :       :       :       :       :       :       :
13              |       |       |       |       |       |       |
```

```
14               :       :       :       :       :       :       :
15               |       |       |       |       |       |       |
16               :       :       :       :       :       :       :
```

20 Appendice D: Exemple 1: serveur généraliste

La table suivante montre la configuration d'un serveur généraliste de taille moyenne. C'est un serveur réseau (DNS, courrier électronique, FTP, news, imprimante partagée, etc.), un serveur X pour plusieurs programmes de CAO, un serveur de cédérom et de bien d'autres choses. Les fichiers sont sur 3 disques SCSI d'une capacité de 600, 1000 and 1300 Mo.

On pourrait augmenter la vitesse en séparant `/usr/local` de `/usr` mais on a supposé ça n'en valait pas la peine vue la complexité de gestion que cela entraîne. Avec 2 disques de plus ça serait plus envisageable. `sda` est vieux et lent et pourrait aussi bien être remplacé par un disque IDE. Les deux autres disques sont assez rapides. On répartira la charge principale entre ces deux-là. Pour réduire le déséquilibre on a mis `/usr/bin` et `/usr/local/bin` sur un disque et `/usr/lib` et `/usr/local/lib` sur un autre.

Avec du RAID on pourrait gagner en fiabilité mais on a jugé que le patch de `md` n'était pas assez fiable et qu'un contrôleur RAID matériel était au-delà du budget.

20.1 Points de montage et liens

```
Répertoire       Mount point      speed    seek     transfer        size     SIZE


swap             sdb2, sdc2       ooooo    ooooo    ooooo           32       2x64

/                sda2             o        o        o               20        100

/tmp             sdb3             oooo     oooo     oooo                      300

/var             __________       oo       oo       oo                       ____
/var/tmp         sdc3             oooo     oooo     oooo                      300
/var/spool       sdb1                                                         436
/var/spool/mail  __________       o        o        o                        ____
/var/spool/news  __________       ooo      ooo      oo                       ____
/var/spool/____  __________       ____     ____     ____                     ____

/home            sda3             oo       oo       oo                        400

/usr             sdb4                                               230       200
/usr/bin         __________       o        oo       o               30       ____
/usr/lib         -> libdisk       oo       oo       ooo             70       ____
/usr/local       __________                                                  ____
/usr/local/bin   __________       o        oo       o                        ____
/usr/local/lib   -> libdisk       oo       oo       ooo                      ____
/usr/local/____  __________                                                  ____
/usr/src         ->/home/usr.src  o        oo       o               10       ____

DOS              sda1             o        o        o                         100
Win              __________       oo       oo       oo                       ____
NT               __________       ooo      ooo      ooo                      ____
```

```
/mnt.libdisk    sdc4            oo      oo      ooo                         226
/mnt.cd         sdc1            o       o       oo                          710


Espcace disque total: 2900 MB
```

20.2 emplacement des partitions

```
Répertoire        sda     sdb     sdc


swap            |       |   64  |   64  |

/               |  100  |       |       |

/tmp            |       |  300  |       |

/var            :       :       :       :
/var/tmp        |       |       |  300  |
/var/spool      :       :  436  :       :
/var/spool/mail |       |       |       |
/var/spool/news :       :       :       :
/var/spool/____ |       |       |       |

/home           |  400  |       |       |

/usr            |       |  200  |       |
/usr/bin        :       :       :       :
/usr/lib        |       |       |       |
/usr/local      :       :       :       :
/usr/local/bin  |       |       |       |
/usr/local/lib  :       :       :       :
/usr/local/____ |       |       |       |
/usr/src        :       :       :       :

DOS             |  100  |       |       |
Win             :       :       :       :
NT              |       |       |       |

/mnt.libdisk    |       |       |  226  |
/mnt.cd         :       :       :  710  :
/mnt.___/_____  |       |       |       |


Place totale:   |  600  | 1000  | 1300  |
```

20.3 Numérotation

```
         Disque:     sda      sdb    sdc

Capacité totale: |   600  |  1000 |  1300  |

Partition
```

```
1                |   100  |   436  |  710  |
2                :   100  :    64  :   64  :
3                |   400  |   300  |  300  |
4                :        :   200  :  226  :
```

21 Appendice E: Exemple 2: serveur en milieu universitaire

L'exemple suivant est dû à `nakano (at) apm.seikei.ac.jp`, et montre la configuration d'un serveur en milieu universitaire.

`/var/spool/delegate` est un répertiore pour les fichiers de log et de cache d'un serveur de proxy Web qui s'appelle "delegated". Il y a 1000 à 1500 requêtes pas jour, et le disque est rempli en moyenne à 15 ou 30 poucents.

`/mnt.archive` est utilisé pour les gros fichiers qui ne sont pas souvent utilisés, comme les données expérimentales (et spécialement les images), les sources de programmes et les sauvegardes de Win95.

`/mnt.root` est une copie de sauvegarde de la racine contenant des utilitaires pour le dépannage. Une disquette de démarrage est faite pour démarrer sur cette partition.

```
=================================================

Répertoire                 sda      sdb     hda

swap                    |    64  |    64  |       |
/                       |        |        |   20  |
/tmp                    |        |        |  180  |

/var                    :   300  :        :       :
/var/tmp                |        |   300  |       |
/var/spool/delegate     |   300  |        |       |

/home                   |        |        |  850  |
/usr                    |   360  |        |       |
/usr/lib                -> /mnt.lib/usr.lib
/usr/local/lib          -> /mnt.lib/usr.local.lib

/mnt.lib                |        |   350  |       |
/mnt.archive            :        :  1300  :       :
/mnt.root               |        |    20  |       |

Espace total :             1024    2034    1050


=================================================

        Disque :          sda      sdb     hda

Place totale :          |  1024 |  2034 |  1050 |

Partition
1                       |   300  |    20  |   20  |
2                       :    64  :  1300  :  180  :
```

```
3                        |   300 |    64 |   850 |
4                        :   360 :   ext :       :
5                        |       |   300 |       |
6                        :       :   350 :       :


Filesystem         1024-blocks  Used Available Capacity Mounted on
/dev/hda1              19485   10534     7945     57%   /
/dev/hda2             178598      13   169362      0%   /tmp
/dev/hda3             826640  440814   343138     56%   /home
/dev/sda1             306088   33580   256700     12%   /var
/dev/sda3             297925   47730   234807     17%   /var/spool/delegate
/dev/sda4             363272  170872   173640     50%   /usr
/dev/sdb5             297598       2   282228      0%   /var/tmp
/dev/sdb2            1339248  302564   967520     24%   /mnt.archive
/dev/sdb6             323716   78792   228208     26%   /mnt.lib
```

Apparemment `/tmp` et `/var/tmp` sont trop grands. On pourrait les regrouper sur la même partition si l'espace disque vient à manquer.

`/mnt.lib` semble aussi trop grand, mais je prévois une nouvelle installation de TeX et de ghostcript, ce qui prend 100 Mo avec les fontes japonaises !

Le système est sauvegardé sur un Seagate Tapestore 8000 (Travan TR-4, 4G/8G).

22 Appendice F: Exemple 3: SPARC Solaris

L'exemple suivant montre la configuration d'un serveur SPARC sous Solaris 2.5.1 en milieu industriel. En plus des services comme le courrier électronique, c'est un serveur pour des applications de CAO et de bases de données.

La simplicité prime ici, donc `/usr/lib` n'a pas été séparé de `/usr`.

C'est une configuration classique, prévue pour 100 utilisateurs.

```
Disque:       SCSI 0                      SCSI 1

Partition     Taille(Mo)  Montée sur     Taille (Mo) Montée sur

  0           160         swap           160         swap
  1           100         /tmp           100         /var/tmp
  2           400         /usr
  3           100         /
  4            50         /var
  5
  6           le reste    /local0        le reste    /local1
```

A cause des besoins spécifiques à ce serveur, il est parfois nécessaire d'avoir de grandes partitions disponibles. On met tout ce qu'on peut sur le disque, en laissant une grande partition `/local1`.

Cette configuration a été utilisée un certain temps avec succès.

Pour un système général et plus équilibré il faudrait échanger `/tmp` et `/var/tmp` puis déplacer `/var` vers le disque 1.

23 Appendice G: Exemple 4: Serveur avec 4 disques

Cet exemple illustre tous les conseils de ce HOWTO, sauf le RAID. Il est assez compliqué, je l'admets, mais offre de grandes performances avec un matériel moyen. La taille des partitions n'y figure pas mais on peut trouver des valeurs typiques dans les autres exemples.

```
Partition       sda             sdb             sdc             sdd
                ----            ----            ----            ----
        1       root            overview        lib             news
        2       swap            swap            swap            swap
        3       home            /usr            /var/tmp        /tmp
        4                       spare root      mail            /var
```

La configuration est optimsée vis-à-vis du positionnement des pistes mais aussi pour diminuer le temps d'accès moyen.

Si vous voulez DOS ou Windows vous devrez utiliser `sda1` et décaler les autres partitions. Il serait intéressant d'utiliser le swap de `sdb2`, `sdc2` et `sdd2` pour le swap de Windows et pour le répertoire temporaire de Windows. Voir les HOWTOs qui expliquent comment faire cohabiter plusieurs systèmes d'exploitation.

Un exemple avec 4 disques utilisant plusieurs types de RAID est donné ci-dessous:

```
Partition       sda             sdb             sdc             sdd
                ----            ----            ----            ----
        1       boot            overview        news            news
        2       overview        swap            swap            swap
        3       swap            lib             lib             lib
        4       lib             overview        /tmp            /tmp
        5       /var/tmp        /var/tmp        mail            /usr
        6       /home           /usr            /usr            mail
        7       /usr            /home           /var
        8       / (root)        spare root
```

Ici toutes les partitions en double exemplaire sont combinées en RAID 0 avec deux exceptions, le swap qui est entrelacé, et et les partitions `home` et `mail` qui sont réalisées en RAID 1 pour des raisons de sécurité.

Notez que les fichiers de démarrage et la racine sont séparés: seuls les fichiers de démarrage doivent être placés en-desous de la limite du 1023-ième cylindre. Le reste de la racine peut être placé n'importe où, et ici ils sont placés sur la partition la plus lente et la plus à l'extérieur. Par simplicité et pour la sécurité, la partition racine n'est pas un système RAID.

24 Appendice H: Exemple 5: Avec 2 disques

Avec deux disques on peut faire moins de choses compliquées mais le schéma ci-dessous devrait donner un point de départ:

```
Partition       sda             sdb
                ----            ----
        1       boot            lib
        2       swap            news
        3       /tmp            swap
        4       /usr            /var/tmp
        5       /var            /home
        6       / (root)
```

25 Appendice I: Exemple 6: Avec un seul disque

Même si ça tombe hors du champ de ce HOWTO, il est indéniable que les très grands disques deviennent abordables. On voit maintenant des disques de 10 à 20 Go, et la question est alors: comment tirer profit de tels monstres ? Il est intéressant de constater que les gens n'ont auun problème à remplir de tels disques, et l'avenir semble très rose pour les fabricants qui prévoient des disques encore plus gros.

Bien sûr on peut faire moins d'optimisations qu'avec deux disques mais on peut utiliser quelques trucs pour optimiser la position des pistes et minimiser les mouvements de la tête.

```
Partition       hda             Size estimate (MB)
                ----            ------------------
         1      DOS             500
         2      boot            20
         3      Winswap         200
         4      data            Selon la taille du disque
         5      lib             50 - 500
         6      news            300+
         7      swap            128     (maximum avec une puce 32 bits)
         8      tmp             300+    (/tmp et /var/tmp)
         9      /usr            50 - 500
        10      /home           300+
        11      /var            50 - 300
        12      mail            300+
        13      dosdata         10      ( Windows bug workaround!)
```

Souvenez-vous que `dosdata` est un système de fichiers DOS qui doit être sur la toute dernière partition, sinon Windows plante.