Files Systems : LVM+XFS un duo gagnant ?

Matthieu NEVEU
neveu@gate.cnrs.fr

Atelier ARAMIS - 21 janvier 2010 - laboratoire GATE-LSE

Que peut on attendre de son FS ?

- File System (wikipedia)

"structure de données permettant de stocker les informations et de les organiser dans des fichiers sur ce que l'on appelle des mémoires secondaires (disque dur, clé USB, disques SSD, etc.).

Une telle gestion des fichiers permet de traiter, de conserver des quantités importantes de données ainsi que de les partager entre plusieurs programmes informatiques. Il offre à l'utilisateur une vue abstraite sur ses données et permet de les localiser à partir d'un chemin d'accès"

Que peut on attendre de son FS ?

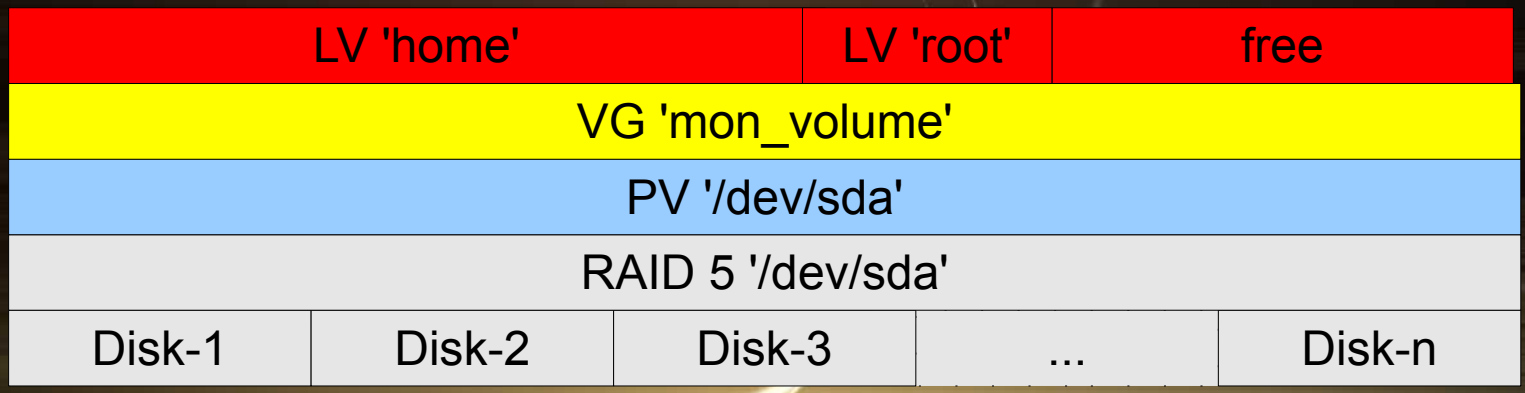
- POSIX
  - "Le standard POSIX impose donc que les fichiers réguliers aient les attributs : La taille du fichier en octets, Identifiant du périphérique contenant le fichier, du propriétaire du fichier, du groupe, Le numéro d'inode, le mode du fichier (RWX), horodatage (timestamp) pour ctime, mtime ou atime, nombre de liens physiques sur cet inode"
- Journalisation
  - *"trace les opérations d'écriture tant qu'elles ne sont pas terminées et cela en vue de garantir l'intégrité des données en cas d'arrêt brutal."*
- Snapshot
  - "est une copie de l'état d'un système à un moment donné du passé"

LVM : principes et présentation

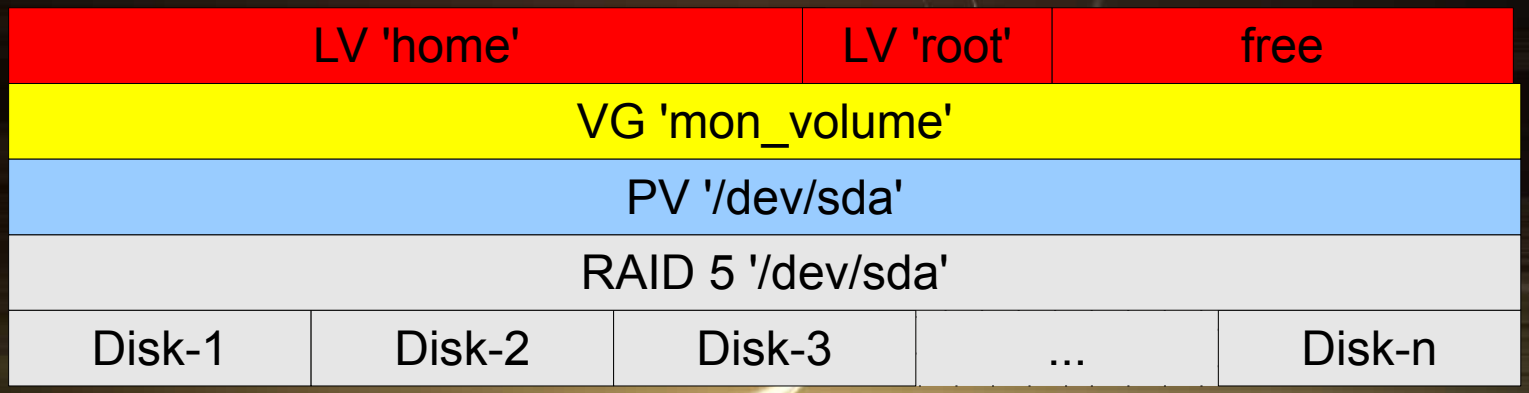
- LVM : Logical Volume Management (Gestion par Volumes Logiques) est une méthode et un logiciel de **découpage**, de **concaténation** et d'**utilisation** des espaces de stockage d'un serveur. Il permet de gérer, sécuriser et optimiser de manière souple les espaces de stockage en ligne dans les systèmes d'exploitation de type UNIX/Linux.

- Physical Volumes : Les disques "physiques" (Disque Dur, partition, ensemble RAID, SAN, DAS, …) sont regroupés dans un/des ***Volumes Physiques*** (PV)
- Volumes Groups : Les PV sont alors placés dans des ***Volumes Groups*** (VG) qui sont un regroupement logiques des PV.
- Logical Volumes : Les VG peuvent ensuite être découpés en autant de ***Logical Volumes*** (LV) que souhaité.
- Le système peut alors utiliser les LV comme des Raw Block Devices à la manière de partitions de disque et donc créer des Files Système "mountable" dessus ou les utiliser pour du Swap

LVM : principes et présentation

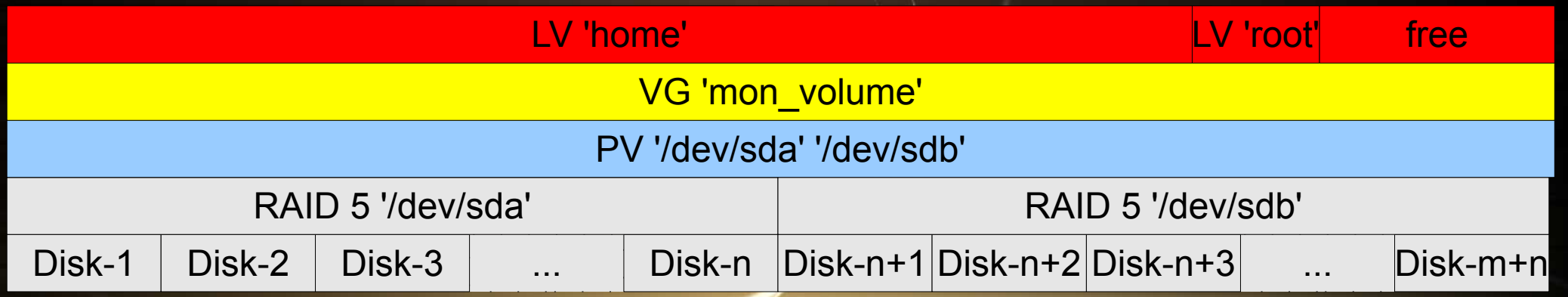


LVM : principes et présentation



- Ajouter, Retirer et Modifier des Volumes
  - On peut ajouter des PV à chaud dans des VG
  - Un PV doit être inutilisé (aucune donnée) pour être retiré d'un VG (possibilité de migrer les données d'un PV vers d'autres)
  - Il est possible d'agrandir ou réduire des LV, mais les filesystems installés dessus doivent prendre en charge cette opération (c'est le cas de XFS).

LVM : principes et présentation



LVM : mise en oeuvre et utilisation

- Google : howto 'ma distro' lvm2 :)

- Editer /etc/lvm/lvm.conf => filter = ["a/.*/"]
- vgscan : scan all disks for volume groups and rebuild caches
- vgchange -a y : change attributes of a volume group, '-a y' active le VG
- pvcreate /dev/sda /dev/sdb …
- vgcreate mon_VG /dev/sda
- vgextend mon_VG /dev/sdb
- lvcreate -L sizeG/M/K -n mon_LV   mon_VG
- lvextend -L +XXG /dev/mon_VG/mon_LV
- FSTAB pour /home ?

  /dev/mon_VG/mon_LV /home xfs noatime 0 0

LVM : extend and reduce

- Pour étendre un LV depuis un espace dispo sur VG
  - lvextend -L +XXG /dev/mon_VG/mon_LV
- Pour changer un disque (non RAID)
  - Un disque (non RAID) /dev/sdb est défectueux, il faut le remplacer par un nouveau /dev/sdc
  - Pvcreate /dev/sdc
  - Vgextend mon_VG /dev/sdc
  - Pvmove /dev/sdb (très long)
  - vgreduce --removemissing mon_VG /dev/sdb
  - Reboot et on retire /dev/sdb et on déplace /dev/sdc à la place

LVM : dump & snapshot

- Dump
  - Vgcfgbackup : backup volume group descriptor area
  - Vgcfgrestore : restore volume group descriptor area
  - Pensez à sauvegarder /etc/lvm/ !
- Snapshot
  - On veut faire un snapshot de 'mon_LV' via l'espace 'free' du VG
  - L'option de lvcreate est : -s (--snapshot)
  - Lvcreate -L xxG -s -n mon_snapshot /dev/mon_VG/mon_LV
  - mkdir /mnt/snapshot
  - mount /dev/mon_VG/mon_LV /mnt/snapshot -onouuid,ro (pour XFS)
  - faites votre rsync, tar, … de /mnt/snapshot vers votre media
  - umount /mnt/snapshot
  - lvremove /dev/mon_VG/mon_snapshot

XFS : présentation

- XFS est un Système de Fichier journalisé de haute performance créé par Silicon Graphics (IRIX OS) porté sur Linux
- XFS est particulièrement efficient dans la manipulation de fichier de très grande taille et les transferts de données.
- Principales caractéristiques
  - GPL depuis 2000
  - Taille de fichier max : 8 Eo (moins 1 octet) sur 64-bit, 16 To binaire sur 32 bit
  - Longueur max de nom de fichier : 255 octets
  - Taille max de volume : 16 Eo
  - Caractères permis dans les noms de fichiers : tous sauf NUL

XFS : journalisation

- Les mises à jour du système de fichier sont d'abord écrites dans le journal avant que la mise à jour du block ne soit faite (taille max du journal 128 Mo)
- Le journal XFS dispose d'entrée "logique" qui offre beaucoup de détails sur les opérations réalisées.
- En cas de crash, les opérations du journal qui précèdent immédiatement le crash peuvent être rejouées pour rétablir la cohérence du système.
- Les récupérations sont réalisés à chaque mount et leur vitesse d'exécution est indépendante de la taille du FS.

XFS : performances

- Groupe d'Allocation : XFS sont partionnés en groupe d'allocation qui sont de taille égale qui gère séparément leur propre inodes et espace disponible. Ceci optimise les performances d'I/O parallèle sur architecture compatible.
- Délais d'allocation : puisque les fichiers sont d'abord bufferisé dans le cache, XFS "réserve" le nombre de blocks contigus pour écrire les fichiers réduisant les problèmes de défragmentation (meilleure performance)
- Gestion des attributs étendus
- Possibilité de gérer les I/O en direct (si l'application le nécessite)
- Snapshots revient à geler le FS (xfs_freeze)
- Online resizing : xfs_growfs permet d'étendre une partition (à utiliser avec LVM). Malheureusement pas de réduction possible
- Utilitaire de backup dédié : xfsdump and xfsrestore. Les backups sont fait dans l'ordre des inodes et peut être fait à chaud ! Backup et restore sont résumable et peuvent être interrompu à tout moment sans difficultés.

XFS : inconvénients

- Il est très difficile de récupérer des fichiers supprimés
- Le FS ne peut pas être réduit
- La création et la suppression d'entrée de répertoire est une opération sur les métadata plus longue que sur d'autres FS

XFS : mise en oeuvre et utilisation

- mkfs.xfs (très rapide)
- fsck.xfs
- Xfsdump & Xfsrestore : puissant !

  le tips : xfsdump -J - / | ssh user@serveur "dd of=/path/to/dump.xfs"

- xfs_admin (change param)
- xfs_freeze
- xfs_growfs (extend)
- xfs_logprint
- xfs_bmap (block mapping)
- xfs_check (consistency)
- xfs_fsr (reorganizer)
- xfs_quota (manage quota)
- xfs_repair (AKA)
- xfs_rtcp (real time copy)
- ...

Exemples d'usage au GATE-LSE CNRS

- Sur serveurs 'normaux'
  - Filers (smb, nfs, ...)
  - Applicatifs WEB et MySql
  - Backup (backuppc)
  - Mailer (cyrus Imap)
- Sur Serveurs 'virtuels' (XEN)
  - DHCP
  - Jabber/XMPP
  - Zabbix
  - Mailer (migration facilité)

Références

- LVM
  - Wikipedia :
    - http://en.wikipedia.org/wiki/Logical_volume_management
    - http://en.wikipedia.org/wiki/Logical_Volume_Manager_%28Linux%29
  - LVM sous Gentoo : http://www.gentoo.org/doc/fr/lvm2.xml
  - EPFL (merci Grégory) : http://sewww.epfl.ch/SIC/SA/SPIP/Publications/spip.php?article816
  - Howto backup xen DomU : http://www.howtoforge.com/back-up-lvm-xen-guest-containing-lvs
- XFS
  - Wikipedia : http://en.wikipedia.org/wiki/XFS_%28filesystem%29
  - SGI : http://oss.sgi.com/projects/xfs/
  - XFS.org : http://xfs.org/index.php/Main_Page
  - RDR for XFS : http://www.ufsexplorer.com/rdr_xfs.php

MERCI !!

Des questions ?