La déduplication extrême d'OS, vous en avez rêvé ? SIDUS l'a fait...

Emmanuel Quémener
Centre Blaise Pascal / École Normale Supérieure de Lyon
46, allée d'Italie
69007 Lyon

Loïs Taulelle
Pôle Scientifique de Modélisation Numérique / École Normale Supérieure de Lyon
46, allée d'Italie
69007 Lyon

Résumé

Offrir à un utilisateur un environnement complet en quelques secondes via COMOD (Compute On My Own Device), simplifier la gestion de centaines de nœuds de calcul, de postes mutualisés ou de stations de travail, limiter l'empreinte du stockage de l'OS sur ses machines, analyser des machines compromises sans démonter quoi que ce soit, tester de nouveaux équipements sans installer de système : tel est l'objectif de SIDUS (pour Single Instance Distributing Universal System). Initié au Centre Blaise Pascal dès février 2010, SIDUS est maintenant le socle de près de 300 nœuds au centre de calcul de l'ENS-Lyon (PSMN). SIDUS n'est ni LTSP, ni FAI ou Kickstart, ni un LiveCD en réseau : c'est une solution basée sur le partage d'une instance unique d'un système d'exploitation par NFS, pour de nombreuses machines. SIDUS a ainsi comme principales propriétés l'unicité de son socle sur le serveur et l'usage des ressources locales sur le client. Hautement scalable, SIDUS est une solution rapide à déployer, légère et n'utilisant que des composants standards, pour ceux qui veulent voir GNU/Linux envahir tous les postes de travail !

Mots-clefs

DHCP, PXE, TFTP, NFSroot, DebootStrap, AUFS, LiveCD

1 Introduction

SIDUS est l'acronyme de *Single Instance Distributing Universal System* et se propose de simplifier à l'extrême l'administration de machines.

Son origine latine « d'ensemble de corps stellaires » est une allégorie : ainsi SIDUS partage le même système d'exploitation avec des machines aux ressources matérielles différentes tout comme les étoiles d'une constellation, de propriétés physiques différentes, appliquent les mêmes mécanismes de fusion nucléaire. SIDUS a donc deux principales propriétés :

- **l'unicité de configuration** : deux machines démarrant sous SIDUS ont exactement le même système d'exploitation ;
- **l'usage des ressources locales** : les processeurs et mémoire vive sollicités sont ceux de la machine locale.

SIDUS n'est donc :

- **ni LTSP** pour *Linux Terminal Server Project* : LTSP propose une gestion simplifiée de terminaux légers en offrant un accès X11 ou RDP à un serveur. Ce dernier supporte ainsi toute la charge de traitement. *A contrario*, SIDUS exploite entièrement (ou à discrétion de l'utilisateur) toute la machine qui s'y raccroche. Seul le stockage du système d'exploitation est déporté sur des machines tierces ;
- **ni FAI** pour *Fully Automatic Installation* : FAI et Kickstart proposent une installation complète simplifiée permettant de limiter voire d'éliminer toute action de l'administrateur. *A contrario*, SIDUS propose un système

unique dans un arbre intégrant à la fois le système de base et toutes les applications installées manuellement ;

- **ni un LiveCD** sur réseau : un *LiveCD* démarre un système minimaliste, nécessairement figé. Il est toujours possible de créer son propre LiveCD mais c'est une opération lourde. Avec SIDUS, il est possible d'installer à la volée sur tous ses clients un nouveau composant instantanément ou de reconfigurer l'instance ;
- **ni monolithique** : dans le cadre de formations en informatique, trois solutions s'offrent aux utilisateurs : exploiter les machines mises à disposition, utiliser leur équipement personnel, installer un environnement virtuel complet figé au téléchargement et donc difficilement modifiable. *A contrario*, SIDUS offre un environnement unique aisément configurable ;
- **ni « original »** : SIDUS exploite des services disponibles sur n'importe quelle distribution : DHCP, PXE, TFTP, NFSroot, DebootStrap, AUFS. Ces quelques mots clés permettant d'installer SIDUS. Il utilise en outre des astuces de distributions de LiveCD et fonctionne sur la distribution Debian depuis sa version Etch.

En définitive, SIDUS est :

- universel : toutes les plates-formes x86 ou x86-64 fonctionnent instantanément ;
- efficace : installation en quelques minutes, démarrage en quelques secondes ;
- économe : à l'origine, 1 cœur, 1Go de RAM, 40Go d'espace disque, et un réseau (GigaBit) Ethernet
- *scalable* : éprouvé sans difficulté sur une centaine de nœuds, en production maintenant sur plus de 300 nœuds ;
- robuste : avec un réseau standard routé, des *uptime* de plusieurs mois dans sa version COMOD ;
- polyvalent : avec la Debian et tout « *science* », toutes les sciences ont leur compte d'Open Source.

2 Où, ou quelles *success stories* ?

- **Sur les postes utilisateur**s : machines mutualisées ou stations de travail individuelles ? Tout a commencé avec une douzaine de clients légers Neoware gonflés en mémoire et *overclockés*. Ce sont maintenant plus de 20 machines équipées de cartes graphiques différentes et offertes à 300 utilisateurs de l'ENS-Lyon ;
- **Sur les nœuds de cluster :**
  - après un démonstrateur de 24 nœuds en mars 2010 au Centre Blaise Pascal, SIDUS sert actuellement 76 nœuds permanents, sur 3 architectures matérielles différentes,
  - après une période de qualification d'une année au Pôle Scientifique de Modélisation Numérique sur quelques dizaines de nœuds, SIDUS équipe maintenant plus de 300 nœuds, dont le nouvel équipement Equip@Meso ;
- **Sur les postes virtuels** : depuis 2011, l'Université Joseph Fourier organise chaque année une école d'été sur le calcul numérique en physique. Au programme, 10 jours intenses ponctués de travaux pratiques : offrir un environnement homogène quasi-instantanément est indispensable. Co-organisateur de ces écoles, le CBP met en place deux images virtuelles de systèmes : l'une autonome utilisable après l'école d'été, l'autre par SIDUS nécessitant seulement une connexion réseau filaire. Ainsi, les professeurs peuvent, quotidiennement, adapter leurs TP. C'est une évolution de cette version qui est utilisée, depuis l'été 2012, par le laboratoire de chimie de l'ENS-Lyon et proposée aux laboratoires de biologie LBMC et IGFL ;
- **Sur les machines suspectes** : le démarrage par le réseau offre une investigation de la mémoire de masse système éteint : inutile d'utiliser un LiveCD sur lequel manque toujours son outil *forensics* préféré ;
- **Sur les machines de prêt** : les fabricants de matériels proposent souvent des équipements d'évaluation. La phase d'installation peut être pénible sur des matériels très, voire trop, récents. Avec SIDUS, le système démarre comme sur les autres équipements déjà en service : quelques minutes pour 20 nœuds.

3 Pour qui : Quels avantages ?

- **Côté utilisateur** : la machine démarre avec seulement les ressources associées. La version VirtualBox

fonctionne au moins sur Linux, Windows et MacOSX : accélération 3D et partage avec l'hôte via un dossier partagé sont disponibles. L'utilisateur retrouve exactement le même environnement que sur les nœuds : l'intégration des codes est donc grandement facilitée. Côté performances, les pertes liées à la virtualisation oscillent entre 10 et 20% (pour VirtualBox) et autour de 5% pour KVM.

- **Côté administrateur** : une opération impacte l'ensemble de l'infrastructure, de l'ordre du simple *sync* sur l'arbre SIDUS. L'installation se déroule en quelques dizaines de minutes pour un système complet. Si des différences entre les systèmes sont minimes, un usage simple de scripts ou de Puppet suffit. Si la différence entre les systèmes est importante, un autre arbre SIDUS est construit, voire cloné instantanément avec des mécanismes de *snapshot* : LVM ou plutôt ZFSonLinux.
- **Côté expérimentateu**r : ingénieur système ou scientifique, l'environnement SIDUS lui offre la reproductibilité. Deux nœuds démarrant sur le même socle SIDUS disposent exactement du même système. Cela permet ainsi, que les machines soient identiques ou pas, de mener des tests vraiment pertinents.

4 Combien : Quelles ressources ?

A titre d'exemple, le serveur des clusters du CBP, également passerelle, héberge les services DHCP, DNS, TFTP, NFS et le serveur de batch OAR. Au démarrage de toute l'infrastructure (76 nœuds), le serveur NFS encaisse sans broncher jusqu'à 900 Mb/s.

Le serveur d'instance SIDUS du PSMN, durant la phase de test, était une machine archaïque (Sun v40z) : elle remplissait sans souci le service de plus de 300 nœuds avant son remplacement par une machine sous garantie.

5 Comment installer le système ?

5.1 En quelques lignes...

Le pré-requis est réduit, à commencer par un réseau idéalement comparable au débit d'un disque dur. Côté services, sont nécessaires : serveurs DHCP, DNS, TFTP et NFS. Les deux derniers vont « porter » SIDUS. La version opérationnelle du CBP exploite en outre des serveurs LDAP (identification/authentification) et NFSv4 (espaces utilisateurs). Côté client, seul suffit un démarrage PXE opérationnel (par la carte, ou par GPXE sur CDROM ou clé USB).

L'installation comporte 8 phases :

1. préparation du système
2. installation de base (socle Debian, debootstrap)
3. installation des paquets complémentaires (TOUT Debian-Science)
4. purge des paquets non désirés
5. adaptation du système à l'environnement local
6. pointage du système vers les serveurs tiers :authentification et partages utilisateurs
7. création de la séquence de démarrage
8. détachement de SIDUS du système hôte

Durant l'installation, les phases coûteuses sont le téléchargement des paquets et le paramétrage de quelques composants (Perl et LaTeX). Elle dure au mieux 45 minutes pour un arbre complet de 32 Go. Quelques précautions sont cependant nécessaires, liées au montage des dossiers systèmes et a l'inhibition du démarrage des services à leur installation dans SIDUS.

Comment maintenant l'offrir sans le dupliquer ? Une première approche a été d'utiliser un cortège de montages volatils (à base de TMPFS) : elle n'est pas viable. La préférence s'est portée sur mécanisme de LiveCD très répandu : la séquence de démarrage intègre ainsi la superposition de deux couches par le liant AUFS (évolution de UnionFS), l'une lecture seule (le CDROM pour le LiveCD et NFS chez nous), l'autre lecture/écriture (en TMPFS)

Mais comment bénéficier de SIDUS et disposer d'un paramétrage conservé d'un démarrage à l'autre ? La première approche avec un montage NFS exclusif pour chaque nœud a été abandonnée, remplacée par un montage iSCSI associé à chaque nœud. Actuellement, au CBP, les machines SIDUS nécessitant une persistance (comme les nœuds Distonet) utilisent le mécanisme NFSroot+iSCSI=AUFS, les autres NFSroot+TMPFS=AUFS.

Pour conclure, les machines mises à disposition sont assez hétérogènes : les nœuds de clusters (disposant d'équipements réseau rapides), les stations de travail (embarquant des cartes graphiques) ou les machines virtuelles (exigeant un partage des données et une accélération graphique) demandent quelques adaptations. Une première solution serait la persistance, mais trop lourde pour les grands parcs de machines : seront alors préférées l'utilisation de scripts de démarrage, l'exploitation d'un arbre SIDUS séparé ou l'installation de composants tierces.

5.2 Préparation du système

Nous devons préparer un peu notre système afin d'accueillir SIDUS. Nous avons la main sur plusieurs services pour déployer nos clients : serveurs DHCP, TFTP, NFS. Nous entretenons de très bonnes relations avec notre service IT ou nous sommes assez libres pour accéder sans contraintes aux serveurs LDAP et DNS bien définis :

- le service DHCP fournit à notre client une adresse IP mais diffuse deux informations complémentaires : l'adresse du serveur TFTP via la variable next-server et le nom du binaire PXE, souvent nommé pxelinux.0.
- le service TFTP entre alors en scène. Il offre par TFTP tout le nécessaire permettant le démarrage du système : le binaire pxelinux.0, le noyau et le démarrage du système du client. Si nous avons besoin d'offrir des paramètres à tel ou tel client, nous construisons un document spécifique dont le nom sera construit à partir de son adresse MAC (préfixé de 01 et dont les « : » sont remplacés par des « - »).
- le serveur NFS s'invite alors dans la boucle : il va offrir la racine du système par son protocole (donc NFSroot). C'est donc dans cette racine, par exemple /srv/nfsroot/sidus que nous allons installer notre système client.

Sur nos configurations nous utilisons respectivement isc-dhcp-server, tftpd-hpa et nfs-kernel-server pour les serveurs DHCP, TFTP et NFS.

5.3 Installation de base par Deboostrap

Debootstrap permet l'installation d'un système dans une racine. Il exige plusieurs paramètres comme la racine d'installation, l'architecture matérielle, la distribution et l'archive FTP ou HTTP Debian à solliciter pour le téléchargement.

Là commence la « spécialisation » de notre installation, à l'origine construite autour d'une distribution Debian. Cet outil est familier de toutes les distributions Debian-like : il sera donc disponible chez les dérivées du système à la spirale (à commencer par Ubuntu). Il sera cependant assez facile de réaliser cela sur les Redhat-like, Fedora intégrant par exemple un clone, Febootstrap, mais que nous n'avons pas testé.

Debootstrap accepte aussi en entrée une liste d'archives (vous savez que Debian est très tatillon sur les licences en séparant les archives en main, contrib et non-free), une liste de paquets à inclure et une liste de paquets à exclure. Nous aurions été ravis de pouvoir, ici, préciser la liste complète des paquets à inclure ou à exclure, mais, malheureusement, cette approche est une voie sans issue : nous installerons donc, dès cette commande debootstrap, un ensemble d'outils indispensables à inclure dès le début (par exemple le noyau, des *firmwares* pour un support étendu de tous les matériels et un ensemble d'outils d'audit).

Nous avons par commodité défini des variables d'environnement correspondant à la racine de notre système $SIDUS et une commande permettant l'exécution d'une commande par chroot avec une option particulière d'installation de paquet.

5.4 Précautions & création d'un « cordon ombilical » avec l'hôte

A la suite de cette commande, nous devons prendre quelques précautions :

- normalement, si le paquet Debian est un service, ce dernier démarre après son installation. Nous devons donc inhiber le lancement de ce service par la définition d'un hook. Ce hook sera supprimé à la fin de l'installation ;

- des paquets exigent l'accès à la liste des processus, du système, des périphériques, de la mémoire virtuelle, des pointeurs de périphériques. Nous devons donc *« binder »* le montage de ces dossiers du système hôte au système SIDUS. Les dossiers concernés sont : /proc, /sysfs , /dev/shm, /dev/pts

5.5 Installation de paquets spécifiques : toute la science !

De manière à simplifier l'installation de paquets appartenant à la même famille, Debian a créé de nombreux meta-paquets, préfixés de « science » : science-chemistry désigne par exemple tous les paquets de chimie. La commande d'installation de tous paquets scientifiques se fait par une seule commande. Comme nous sommes épris de complétude, nous allons aussi rajouter les paquets « suggérés ». Attention ! L'option --install-suggests n'est présente qu'à partir de la distribution Wheezy).

Durant l'installation, les phases les plus longues sont le téléchargement des paquets qui représente plusieurs Go (et dépend donc de la connectivité à Internet et aux miroirs officiels) et la configuration initiale de certains paquets (comme Perl et LaTeX).

Dans le meilleur des scénarii, cela prend 45 minutes pour un arbre complet de 32 Go : expérience réalisée sur un disque en RAM. Sur un disque SSD, cela dure moins d'une heure. Avec des disques mécaniques, la même installation frôle les deux heures du fait de sync permanents forcés par l'installeur.

5.6 Purge des paquets non-sollicités (vente forcée;-))

Malheureusement, cette boulimie de paquets n'est pas sans effet sur la propreté de l'installation : quelques uns s'installent « mal », d'autres exigent une purge, notamment l'installeur Matlab !

5.7 Adaptation de l'environnement local

L'environnement local a une importance et, rien n'est défini : celui par défaut étant l'américain, nous paramétrons les éléments suivants :

- l'environnement local : /etc/locale.gen
- la zone : /etc/timezone
- le clavier : /etc/default/keyboard

5.8 Pointage vers les serveurs tiers NFS et LDAP

Les services NFS et LDAP exigent une petite configuration. L'importation des fichiers directement configurés est préférable.

Dans notre cas, nous avons choisi de les mettre tous sur un serveur Web http://MyServeur.MySite/sidus. Un wget permet directement de les récupérer et les placer au bon endroit avec l'option -O.

Ainsi, voici les fichiers à télécharger :

- l'authentification LDAP : /etc/nsswitch.conf, /etc/libpam_ldap.conf, /etc/libnss-ldap.conf, /etc/ldap/ldap.conf
- le montage des dossiers NFS utilisateurs :
  - pour NFSv4 : /etc/default/nfs-common, /etc/default/idmapd.conf, /etc/fstab
  - pour NFSv3 : /etc/fstab

5.9 Création de la séquence de démarrage

Comment partager SIDUS sans le dupliquer ? Nous allons nous inspirer d'un mécanisme utilisé dans certains LiveCD : le montage de la racine du système consiste en la superposition de deux couches, l'une lecture seule (le système NFSroot) et l'autre en lecture/écriture (un TMPFS dans le cas le plus simple). Les deux couches sont liées par la glue AUFS, le projet successeur de UnionFS.

Tout réside dans un seul et unique *hook* : rootaufs, placé très tôt dans le démarrage initrd. Son principe repose sur cinq étapes :

1. création de dossiers temporaires /ro, /rw et /aufs
2. déplacement de la racine NFSroot du point de montage originel vers dans /ro,
3. montage d'un partage distant, d'une partition locale ou distante, ou d'un volume TMPFS dans un autre point de montage /rw
4. superposition des deux dossiers /ro et /rw dans le dossier /aufs
5. déplacement de /aufs vers le point de montage originel

Ce script, rootaufs se place dans $SIDUS/etc/initramfs-tools/scripts/init-bottom

Le script originel a été inspiré par le projet rootaufs de Nicholas A. Schembri (http://code.google.com/p/rootaufs/). Il a été profondément modifié pour l'adapter à notre infrastructure : une version est disponible sur http://www.cbp.ens-lyon.fr/sidus/rootaufs :

Nous n'en avons pas encore fini pour disposer d'un système fonctionnel : nous allons créer par update-initramfs un initrd spécifique pour notre boot NFS contenant les modifications suivantes :

- aufs dans /etc/initramfs-tools/modules
- eth0 comme DEVICE dans /etc/initramfs-tools/initramfs.conf

Il suffit ensuite de copier le noyau et le boot loader dans la définition du serveur TFTP.

A l'origine, nous avons exploré la possibilité d'offrir un second partage NFS en lecture/écriture pour la persistance des modifications associées aux clients d'un redémarrage à l'autre. Cette version, bien que fonctionnelle, exigeait l'ouverture d'un partage NFS atomique pour chaque client : imaginons la charge supplémentaire pour le serveur !

Nous avons donc préféré une autre approche de la persistance, sous la forme d'un disque réseau de technologie iSCSI. Ainsi, nous créons un volume partagé iSCSI par client. Pour des raisons de simplicité, le volume offert porte l'IP du client et nous ne fournissons que le serveur de volume iSCSI. Les login et mot de passe d'accès par défaut sont dans le fichier rootaufs.

5.10 Rupture du « cordon ombilical » avec le système hôte

De manière à installer correctement SIDUS, nous avons été contraints de lier étroitement système hôte et chroot. Il est donc nécessaire de :

- démonter les dossiers système : /proc, /sysfs , /dev/shm, /dev/pts ;
- effacer les dossiers temporaires ;
- purger des processus liés au dossier d'installation de l'instance SIDUS au besoin.

5.11 Quelques ruses de « vieux coyote »

- pour le paramétrage du nom (hostname) par DHCP, effacer /etc/hostname ;
- paramétrer le /etc/resolv.conf avec une définition persistante ;
- dans /etc/network/interfaces : définir un loopback ;
- modifier le démarrage de GDM3 pour ne le démarrer qu'après le lancement du NSCD ;
- paramétrer /etc/security/limits.conf (indispensable dans un environnement HPC) ;
- paramétrer /etc/fstab avec l'entrée du serveur NFS des comptes utilisateurs ;
- Pour les systèmes virtuels à base de VirtualBox :
  - Installer VBoxLinuxAdditions.run dans le système Sidus

- Pour les systèmes avec une carte InfiniBand :
  - Forcer le chargement des modules dans /etc/modules et régénérer le initrd
  - Exécuter dans /etc/rc.local un script permettant de récupérer l'adresse IP Ethernet et construire une adresse IP pour la carte Infiniband.
- Pour les systèmes avec une carte NVIDIA :
  - Pour la majorité des cartes NVIDIA, les paquets proposés dans la Debian Wheezy permettent une installation complète des pilotes propriétaires, des librairies OpenGL, CUDA et OpenCL. Attention cependant si vous désirez utiliser simultanément l'ICD (Installable Client Loader) OpenCL d'AMD pour exploiter vos processeurs ET votre carte graphique : les paquets AMD et NVIDIA étant mutuellement exclusifs, nous avons dû installer toute l'infrastructure pilote, CUDA et OpenCL *from scratch* pour le permettre.
- Pour les systèmes avec une carte AMD ATI
  - Pour la majorité de cartes ATI, les paquets proposés dans la Debian Wheezy permettent une installation complète des pilotes propriétaires, des libraries OpenGL, CUDA et OpenCL.

6 Comment administrer le système ?

Si l'administration est plus lourde que son installation, le bénéfice du premier efface la perte récurrente du second. Finalement, avec SIDUS, chaque phase d'administration intègre les mêmes mécanismes qu'à l'installation : protection contre le démarrage des nouveaux services et montages de dossiers systèmes. Le reste est identique.

En définitive, une instance SIDUS s'administre comme tout système *chrooté* : il y a cependant des précautions à prendre, toutes les fonctions d'administration n'exigeant pas les mêmes contraintes les unes que les autres. Souvent, un chroot sur la racine de SIDUS suivi de la commande système suffit. Une nouvelle approche, basée sur l'ouverture d'un SSH directement dans l'instance SIDUS, est en préparation.

7 Conclusion

Quelque soit l'environnement (nœud HPC, poste de travail, machine virtuelle), SIDUS apporte une flexibilité inégalée, autant pour l'utilisateur que pour l'administrateur de ces ressources. Sa frugalité, sa rapidité de propagation en font un outil d'une rare polyvalence. L'essayer, c'est l'adopter !