Para quoi ? Paradoxe du parachute parabolique ou paradigme au paracétamol, à quoi bon paravirtualiser ? Xen [Xen], ce nom ne doit pas être inconnu des personnes suivant l'évolution ou l'actualité du noyau Linux, ni par celles portant attention à la virtualisation des systèmes d'exploitation (OS).

Car voilà bien le but de Xen, permettre à de multiples OS invités de fonctionner simultanément sur une même machine hôte. Ces OS invités et leurs applications fonctionnent dans des machines virtuelles, isolées et sécurisées et partagent les ressources de la machine hôte.
Xen est un hyperviseur de machine virtuelle, aussi appelé moniteur de machine virtuelle, en anglais Virtual Machine Monitor (VMM).
Xen offre les fonctionnalités suivantes :

- Une isolation complète et sécurisée entre les machines virtuelles hébergées, un contrôle des ressources et une garantie de qualité de service.
- Des performances pour les machines virtuelles proches d'un OS natif.
- La possibilité de migrer des machines virtuelles entre des serveurs Xen, sans interruptions de service.
- Un très bon support du matériel, Xen utilise principalement les pilotes du noyau Linux.

- Surveiller les machines virtuelles, leur allouer des ressources et détecter des comportements inhabituels ou suspects.
- Réaliser des plans de reprise d’activité ou de continuité d’activité.
- Tester le fonctionnement d’une application sur plusieurs OS, sans disposer d’autant de machines.
- Le développement, test et debogage d’un noyau fonctionnant dans une machine virtuelle isolée.
- Tester, sur une seule machine, différents OS.
- Tester une architecture réseau.
- Le développement de nouveaux OS en profitant du large support matériel des pilotes du noyau Linux fournis par la machine virtuelle.

Et pour parfaire le tout, Xen est un Logiciel libre publié sous la GNU GPL.

Déroulement de l’article

Au cours de cet article, nous découvrirons les techniques et technologies qui sont liées à Xen et à la virtualisation.
Nous commencerons par un bref retour sur les débuts de la virtualisation, puis l’environnement dans lequel est né le projet Xen, ainsi que ses fonctionnalités.
Nous poursuivrons par l’architecture d’un système Xen, les apports de Xen 3.0 et ce que l’on peut attendre de la version 3.1.

Une présentation de Xen 3.0 a été faite lors de FOSDEM 2006. Vous trouverez la vidéo sur ftp.belnet.be dans mirror/FOSDEM.

La plupart des grandes sociétés historiques de l'informatique ont développé leurs mainframes, machines virtuelles et OS. Puis, est venu le temps de la micro-informatique, de son processeur unique, de son OS unique et mono-tâche.
Au fil du temps, l'OS mono-tâche est devenu multitâche, les processeurs s'apparient et gagnent des cœurs. Et, ironie de l'histoire, nous voilà revenu au temps héroïque, avec l'éclosion de multiples projets de machines virtuelles.
Il existe deux grands types de virtualisation. Elle peut être soit logicielle :

- QEMU, émulateur de plate-forme x86, PPC et SPARC.
- Plex86, Bochs, WMware (1999), VirtualPC (1997), émulateurs de plate-forme x86.
- PearPC, émulateur de plate-forme PPC pour machine x86.
- Java Virtual Machine (JVM), machine virtuelle pour le byte-code Java. Elle existe pour différentes architectures comme, x86/GNU/Linux, x86/Windows, SPARC/Solaris...
- User Mode Linux, noyau Linux fonctionnant en espace utilisateur.
- RT-Linux, micro noyau temps réel dur faisant fonctionner le noyau Linux en espace noyau non temps réel.
- Virtuozzo, partitionnement du système au niveau noyau pour Linux et Windows 2003.
- Linux VServer, BSD Jail, isolation du système de fichier et des processus

« jails » des systèmes BSD. Une jail est un environnement d'exécution isolé et sécuriser. Une seule instance de Solaris est utilisée.

- IBM Logical Partitioning (LPAR), solution de partitionnement pour les ordinateurs pSeries. LPAR permet le fonctionnement de partitions AIX et GNU/Linux.
- IBM Dynamic Logical Partitioning (DLPAR), solution de partitionnement incluse dans AIX 5L version 5.2. DLPAR permet d'ajouter ou de retirer dynamiquement des ressources d'une partition.
- La famille des iSeries d'IBM supporte le partitionnement logique. La première partition OS/400 charge l'hyperviseur appelé « the Hypervisor », celui-ci permet le contrôle, la gestion et l'isolation entre les partitions. Celles-ci peuvent contenir le système OS/400 ou GNU/Linux.
- z/VM est un OS permettant la virtualisation, il est le successeur de VM/ESA. z/VM supporte plusieurs OS invités comme GNU/Linux, OS/390, TPF, VSE/ESA, z/OS et lui-même. Les ressources de la machine hôte sont gérées par le programme de contrôle appelé « CP ».

Soit matérielle :

- Mainframe IBM (zSeries), ils permettent de faire fonctionner simultanément plusieurs OS comme z/OS, z/OS.e, z/VM, VSE/ESA, TPF et GNU/Linux.
- Sun E10k et E15k, partitionnement matériel statique pour Solaris en 1996, le partitionnement matériel devient dynamique en 1999 (Dynamic System Domains).

de serveur, la gestion des ressources, l'authentification, l'autorisation et la comptabilisation des usages.

Le projet Xen est né à l'Université de Cambridge en 2001 sous l'impulsion de Ian Pratt, leader et architecte en chef du projet. La première version publique de Xen (version 1.0) est publiée en octobre 2003.
Elle est stable et est utilisée en environnement de production. La version 2.0 est publiée en octobre 2004. Elle apporte son lot d'améliorations, un support matériel plus large, une configuration plus souple. La version 3.0 est publiée le 5 décembre 2005.
Les nouveautés notables sont : le support des architectures 32 bits SMP, des architectures 64 bits pour les OS invités et la technologie Vanderpool d'Intel.
Le projet Xen bénéficie de la collaboration depuis 2003 de l'Intel Research Cambridge et depuis 2004 de l'HP Labs.
En 2004, Ian Pratt, Nick Gault (fondateur de Network Physics) et Moshe Bar (leader d'OpenMosix) créent XenSource [XenSource], une société basée à Palo Alto (Californie) dédiée au développement de Xen, à sa promotion et à son support.
Aujourd'hui, Xen est activement supporté par une vingtaine de grandes entreprises dont AMD, Hewlett-Packard, IBM, Intel, Network Appliance, Novell, Red Hat, Sun Microsystems ainsi que l'OSDL et une dizaine de grandes universités. La version 3.0 de Xen a vu la participation de plus de 150 développeurs open source, répartis tout autour du globe.

Les grandes lignes de Xen

la version 3.0 de Xen. IBM travaille au portage vers le PowerPC 970, HP vers l'IA64 et Sun vers le Sparc.
Le multiprocesseur (SMP) est supporté par Xen depuis la version 1.0, la version 3.0 permet à l'OS invité d'utiliser le SMP pour les architectures 32 bits.

Le SMT, maintenant appelé hyperthreading, n'est pas une nouveauté. Le Simultaneous Multi Threading est une technique datant des années 1950.

L'hyperthreading (SMT) est supporté par Xen et l'OS invité depuis la version 3.0. L'équipe de Xen rapporte qu'ils ont obtenu de très bons résultats sur des systèmes octo-processeurs avec la version 2.0. Ils ont aussi de bons retours d'utilisations de Xen sur des systèmes 32 voies.
Le développement initial sur l'architecture x86 est en partie à l'origine du choix de la paravirtualisation. Dans une virtualisation complète, la machine virtuelle simule à l'identique la machine hôte, ce qui permet de faire fonctionner un OS invité non modifié.
L'architecture x86 n'ayant pas été prévue pour gérer la virtualisation, elle rend complexe le développement d'une solution de virtualisation complète.
Le choix de la paravirtualisation a aussi été motivé par la possibilité d'obtenir des performances proches d'un OS natif (cf. § Architecture d'un système Xen), performances difficiles à atteindre avec une virtualisation complète sur l'architecture x86. Xen pénalise les performances de 2 à 3% en moyenne.
Intel avec l'extension d'architecture appelée « Vanderpool » ou « VT » et AMD avec « Pacifica » ont inclus dans leurs processeurs le support de la

La famille des processeurs compatibles x86 possède un modèle de protection composé de quatre niveaux d'exécution aussi appelés « rings » et numérotés de 0 à 3, respectivement du plus privilégié au moins privilégié.
Le ring 0 est couramment dédié à l'exécution de l'OS et le 3 aux applications de l'espace utilisateur. Les rings 2 et 3, originellement prévus pour la virtualisation, sont rarement utilisés. Les OS actuels sont conçus pour fonctionner sur le ring 0 et utilisent des instructions uniquement disponibles sur celui-ci. OS/2 est l'un des rares OS à utiliser les rings 2 ou 3. Ces deux rings ont été supprimés de l'architecture x86_64.
Dans le cas d'un système Xen sur l'architecture x86, l'hyperviseur est exécuté dans le ring 0, les OS invités dans le ring 1 ou 2 et les applications dans le ring 3. Pour l'architecture x86_64, les OS invités et les applications sont exécutés dans le ring 3.
Pour faciliter la virtualisation sur l'architecture x86, Intel avec l'extension Vanderpool ajoute deux nouveaux modes d'exécutions appelés VMX root et VMX non-root. Ces deux modes supportent les quatre niveaux d'exécutions (ring 0 à ring 3).
Avec Vanderpool, les OS utilisent le ring 0 et n'ont plus besoin d'être modifiés. Les applications restent en ring 3. OS et applications fonctionnent dans le mode d'exécution VMX non-root, sans requérir de modifications. L'hyperviseur utilise le mode d'exécution VMX root, accédant ainsi à un niveau de contrôle et de privilège plus important. Seul l'hyperviseur doit être modifié pour pouvoir utiliser l'extension Vanderpool et gérer les nouveaux modes d'exécution VMX.
Au mode d'exécution VMX root, Vanderpool associe des nouvelles instructions

changement de contexte, comme c'est le cas sur les processeurs x86. Les TLB étiquetés est une technologie issue des processeurs RISC comme les processeurs Alpha, MIPS ou SPARC.
Comme Intel avec Vanderpool, AMD avec Pacifica ajoute deux nouveaux modes d'exécutions appelés « Host Mode » et « Guest Mode ». L'hyperviseur fonctionne dans le Host Mode et l'OS invité dans le Guest Mode. Comme dans le schéma d'Intel, le Host Mode permet à l'hyperviseur l'accès à un ensemble d'instructions pour gérer les machines virtuelles. Ces nouvelles instructions utilisent une structure dédiée appeler VMCB (Virtual Machine Control Block) fournie par le processeur.
Coté sécurité, Pacifica isole, au niveau matériel, les OS invités les uns des autres, en filtrant les instructions et les événements, comme les accès DMA et en permettant de gérer les interruptions des processeurs virtuelles.

Architecture d'un système Xen

Xen est un hyperviseur. Son rôle est d'ordonnancer le fonctionnement des différentes machines virtuelles, en les perturbant le moins possible, pour approcher les performances d'un OS natif. Pour cela, Xen se doit d'être le plus léger et rapide possible.
L'architecture d'un système Xen est composé de l'hyperviseur Xen et des différentes machines virtuelles sécurisées, appelées « domaine » dans la terminologie Xen.

par le daemon Xend qui fonctionne dans l'espace utilisateur du domaine 0.
Le daemon Xend est responsable de la gestion des domaines et fourni un accès à leurs consoles. Il est accessible à la fois par un client en ligne de commande et par un navigateur web. Dans les deux cas, le protocole HTTP est utilisé pour l'échange des commandes et réponses.
Cette architecture garantit à l'hyperviseur une protection contre les bugs et les plantages des pilotes. Elle décharge aussi l'équipe de Xen du développement des pilotes, pour se concentrer uniquement sur le rôle principal de l'hyperviseur. Elle assure de plus au système Xen un large support matériel via les pilotes du noyau Linux. Cette architecture simple et légère permet à Xen d'offrir de hautes performances.

La virtualisation des pilotes

Dans l'architecture d'un système Xen, seul l'hyperviseur et le noyau Linux du domaine 0 ont un accès direct au matériel et ont une connaissance des caractéristiques de celui-ci. Ils utilisent, pour chaque périphérique, le pilote et la configuration idoines fournis par le domaine 0. Ce dernier exporte une version virtualisée des périphériques vers les autres systèmes invités. Cette approche permet aux périphériques d'être contrôlés uniquement par le système du domaine 0 et d'être disponible pour tous les OS invités.
Ce qui signifie que le système Linux du domaine 0 doit être configuré à la fois pour supporter le matériel sous-jacent, l'hyperviseur Xen et fournir des périphériques virtuels. Cette approche permet aussi de limiter le crash d'un

Il est en grande partie écrit en Python (version 2.3) et son interface HTTP est accessible par défaut sur le port 8000. Toutes les interactions entre Xend et Xen transitent par le serveur XCS (Xen Control Switch).

Migration des machines virtuelles

Xen permet de migrer une machine virtuelle d'un serveur Xen vers un autre sans quasiment arrêter la machine virtuelle. Durant cette procédure, la machine virtuelle en fonctionnement est copiée sur le serveur Xen final. Un arrêt de 60 à 300 ms est nécessaire pour synchroniser les machines virtuelles avant le démarrage de la machine virtuelle finale.
Lors du Symposium Linux d'Ottawa (OLS) de 2005, Ian Pratt a présenté les traces de la migration d'un serveur Quake. Celles-ci ont montré une interruption du serveur d'à peu près 50 ms. Les joueurs n'ont rien vu.

Xen 3.0

La version 3.0 de Xen est sortie le 5 décembre 2005. Cette version amène de nombreuses nouveautés et améliorations tant du point de vue de la stabilité que des performances, de la sécurité et du contrôle de l'usage des ressources. Elle vise le marché des data centers et des serveurs d'entreprises avec le support du SMP, l'adressage de plus de 4 Go de mémoire et la possibilité de virtualiser tout OS sans les modifier (cf. Vanderpool). Il est à noter que cette

x86, x86_64, IA64. Le port vers l'architecture PowerPC est quasi complet. Le port de Xen vers l'architecture x86_64 supporte les OS invités compilés pour la cible x86_64 et les applications 32 et 64 bits. Cette version inclut aussi le support de la technologie Intel Vanderpool. L'intégration de la technologie AMD Pacifica est prévue au cours l'année 2006. A ce titre, XenSource a annoncé à l'Intel Developer Forum d'août 2005 [IDF-2005], avoir réussi à faire fonctionner le système paravirtualisé GNU/Linux et le système totalement virtualisé Windows XP SP2 sur des plateformes composées de pré-version de Xen 3.0 et de processeurs Intel avec l'extension Vanderpool.

Xen 3.0 apporte le support du SMP pour l'OS invité et le support de l'hyperthreading pour l'hyperviseur et l'OS invité. Le SMP est supporté jusqu'aux architectures 32 voies. Le daemon Xend a été modifié pour régulièrement rééquilibrer la charge, en déplaçant périodiquement les machines virtuelles sur les CPU supportés (hyperthreading compris).

Deux nouveaux modes d'adressage de la mémoire sont inclus dans Xen 3.0 : le support du mode PAE sur x86 pour passer la barrière des 4 Go et le support de l'adressage mémoire sur 64 bits pour les architectures qui le supportent.

Xen 3.0 supporte aussi les modules « Trusted Platform ». Ce support est issu de l'initiative « Secure Hypervisor » d'IBM.

Le daemon Xend a été amélioré et offre la possibilité d'ajouter ou d'enlever à la volée un processeur virtuel à une machine virtuelle et améliore la migration des machines virtuelles entre serveurs. Ce qui a pour effet de fluidifier et de faciliter la répartition de charge des machines virtuelles composant un cluster. L'une des améliorations notables de cette troisième mouture de Xen

d'architecture Vanderpool et Pacifica, pour permettre une virtualisation complète et sans perte de performance sur les architectures x86 et AMD64 et ainsi permettre à Xen la virtualisation de tout OS sans devoir les modifier. Sont aussi au programme de la version 3.1, des améliorations des outils de contrôle et de configuration, le support du SMP sur les architectures 32 et 64 bits à la fois pour l'hôte et les OS invités, le support des systèmes NUMA, des périphériques InfiniBand, le support du hotplug CPU et du hotplug RAM. Ces dernières évolutions, liées au support matériel, passent aussi par l'évolution du noyau Linux. Des discussions sont en cours sur la liste de développement du noyau Linux et la liste de développement de Xen, concernant l'inclusion, dans la version stable du noyau Linux, d'une couche (API) dédiée à la virtualisation. Cette API dédiée permettra d'utiliser Xen sans devoir patcher le noyau. Elle sera aussi utile aux autres solutions de virtualisation comme UML…

Mise en œuvre

Selon vos goûts, il existe plusieurs façons d'installer et de tester Xen avec vos OS favoris :

- Un CD live de démonstration d'un système Xen est disponible [Demo CD]. Il permet de tester Xen sur une machine sans l'installer sur le disque dur. Ce CD est basé sur la distribution Debian.
- La distribution Xenophilia est une distribution GNU/Linux basée sur Xen [Xenophilia].

Pour aller plus loin sur la toile :

- Il existe trois mailing-lists officielles pour suivre l'évolution du projet Xen :
- xen-devel@lists.sourceforge.net
- xen-announce@lists.sourceforge.net
- xen-changelog@lists.sourceforge.net
- Les utilisateurs de Xen peuvent poster sur la liste xen-devel.
- Le portail francophone sur la virtualisation : http://xenfr.org
- Conception de Xen et benchmarks : http://www.cl.cam.ac.uk/netos/paper/2003-xensosp.pdf
- Performances comparées de GNU/Linux et NetBSD sur Xen : http://users.piuha.net/martti/comp/xendom0/xendom0.html

Références

- [Brainshare] NetWare sur l'architecture Xen : http://www.novell.com /brainshare
- [Deb] Page de suivi des paquets Xen 3.0 : http://packages.qa.debian.org /x/xen-3.0.html et dépôt des paquets Xen 3.0 : ftp://ftp.fr.debian.org /debian/pool/main/x/xen-3.0
- [Demo CD] Demo de Xen sur un CD live : http://www.cl.cam.ac.uk /Research/SRG/netos/xen/download.html et http://www.xensource.com
- [Fedora] Projet de virtualisation pour les distributions de Red Hat :