Virtualización Xen libre sobre hardware redundado

Free Xen virtualisation on reusable hardware

Virginio García López, Jose Fco. Hidalgo Céspedes

Resumen

A partir de un chasis con electrónica redundada en medios planos, vamos a diseñar una solución Xen adaptada a una organización que trabaja con servidores de desarrollo y explotación en VLANs separadas. El anfitrión manejará la complejidad de los conmutadores de fibra y de red en alta disponibilidad, y permitirá que los servidores virtualizados manejen hardware sencillo. El diseño llevado a cabo permitirá la migración en vivo de máquinas virtuales.

Palabras clave: virtualización, alta disponibilidad, chasis, Xen, Multipath, Bonding, Ebtables.

Summary

Using a chassis with midplanes redundant electronics, we are going to design a Xen solution for a corporation which has both explotation and development systems in different VLANs. The Hypervisor host manages the fiber switches complexity and the high availability for the network. Therefore, it allows simple virtualized servers hardware handling. The design makes live migration of virtual machines easiest.

Keywords: Virtualization, High availability, Xen, Chassis, Multipath, Bonding, Ebtables.

El diseño Xen permitirá la migración en vivo de máquinas virtuales

1. Introducción

Cuando se dispone de hardware redundado en un entorno de servidores de desarrollo y explotación, se plantea cómo aprovechar dicha redundancia de forma que las máquinas virtuales se beneficien de la alta disponibilidad.

Expondremos la experiencia llevada a cabo con máquinas Xen sobre blades ensamblados en un chasis común, donde han de mantenerse separadas las máquinas en VLANs distintas. Veremos cómo la redundancia de almacenamiento y de red se llevará a cabo en el anfitrión, y se propagará a las distintas máquinas virtuales.

El chasis divide su electrónica en medios planos horizontales, así el plano superior funciona independiente del inferior

2. Hardware

2.1. Arquitectura del chasis

El diseño del hardware permite alojar hasta catorce blades – de arquitecturas PowerPC e Intel –, integrar módulos de gestión remota, conmutadores ethernet y fibre channel, así como compartir unidades de DVD o puertos USB.

Esta organización ahorra espacio físico que permite consolidar servicios existentes con los nuevos en el mismo centro de proceso de datos. En el interior de cada blade, la electrónica se encuentra redundada, de forma que cada conjunto de elementos ofrece un conector único hacia cada una de las partes redundantes del chasis: los medios planos.

El chasis divide su electrónica en medios planos horizontales, de forma que el medio plano superior funciona independientemente del inferior. A cada medio plano conectaremos una fuente de alimentación, un conmutador fibre channel y dos conmutadores ethernet. Ambos planos funcionan en activo-activo, ofreciendo dispositivos duplicados al sistema operativo de cada blade.

2.2. Conectividad ethernet

El chasis tiene alojados cuatro conmutadores ethernet (dos en cada medio plano). Cada uno de estos switches tiene seis puertos externos que permiten conectar el chasis con la infraestructura de red de la UMU. La conexión de cada switch a los blades se hace a través de puertos internos.

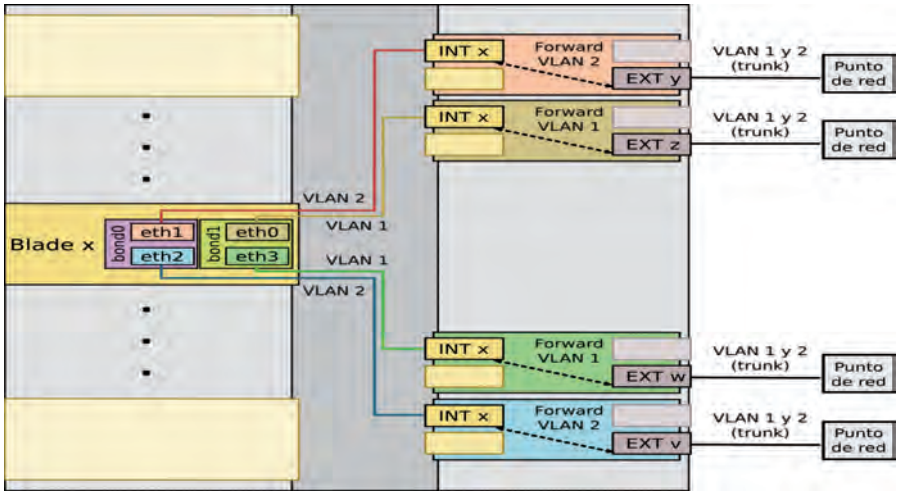
Una forma sencilla de garantizar la redundancia sería conectar varios cables de red desde los puertos externos a los conmutadores y realizar una agregación de enlaces en el switch. Sin embargo, contamos con ciertas limitaciones en el diseño, y una de estas limitaciones es la cantidad puertos de conmutación disponibles en la sala. Perseguimos el objetivo de reducir las conexiones a un cable por switch. Otro requisito viene impuesto por la organización de los servidores: queremos mantener los servidores de desarrollo y de explotación en distintas VLANs.

Teniendo en cuenta estos requisitos se ha implantado una solución que consiste en conectar cuatro cables –uno por cada switch–, y configurarlos en modo trunk: de forma que haya tráfico etiquetado tanto de la VLAN de desarrollo como la de explotación. Con este mecanismo todos los switches reciben el mismo tráfico, sin embargo, para que los blades reciban el tráfico por sus interfaces, es necesario activar el reenvío (forwarding) entre los puertos internos y los externos. En este proceso de reenvío se filtran las VLAN de desarrollo y de explotación, de forma que cada blade reciba la VLAN de desarrollo por dos interfaces, cada una correspondiente a un medio plano redundante, e igualmente la VLAN de explotación se vea en dos interfaces correspondientes a los dos medios planos.

El objetivo es reducir las conexiones a un cable por switch

Una vez que cada blade ve cada VLAN por dos interfaces, el sistema operativo es el encargado de realizar una agregación de enlaces o bonding **(figura 1)**. El bonding permite que los dos interfaces se manejen como un único interfaz. La ventaja es que si el sistema operativo detecta la caída del enlace de un interfaz físico, puede utilizar el otro interfaz físico del agrupamiento. Esta es una buena forma de garantizar la alta disponibilidad, pero la arquitectura del BladeCenter[1] presenta otro problema adicional: los interfaces no pueden perder el enlace accidentalmente. Esto es así porque los interfaces de red están conectados internamente a los puertos internos de los switches. Para que el sistema operativo advierta de la caída de un enlace, debemos desactivar su puerto interno manualmente en el switch. En el caso de que uno de los cables se desconecte o falle accidentalmente, el sistema operativo no podrá darse cuenta, y por lo tanto no utilizará la agregación de enlaces correctamente.

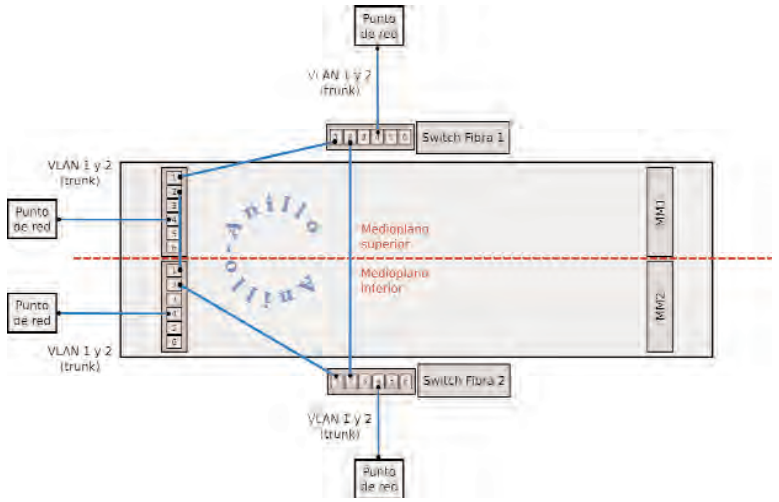
FIGURA 1. ESQUEMA DE CONEXIÓN DE LOS SWITCHES CON UN BLADE



Si el sistema operativo detecta la caída del enlace de un interfaz físico, puede utilizar otro interfaz físico del agrupamiento

Este inconveniente obliga a modificar el diseño del hardware para asegurar la alta disponibilidad tras un fallo de cableado. La solución adoptada es conectar unos switches con otros formando un anillo (**figura 2**). Esto es, estamos duplicando el tráfico para garantizar la redundancia. Lógicamente el tráfico no debe llegar duplicado a los blades, por lo que es esencial activar el protocolo Spanning Tree[2] en los puertos externos de los switches. Con esta configuración, si un cable que se está usando fallara, el Spanning Tree buscaría otro camino rompiendo el anillo. Aunque el sistema operativo nunca detecte la caída de un enlace, el tráfico seguiría llegando sin interrumpir el servicio.

FIGURA 2. ANILLO DE CONEXIÓN ENTRE CONMUTADORES.



Cada switch de fibra permitirá que cada disco (LUN) sea accesible por un camino distinto

2.3. Almacenamiento

Para conseguir la alta disponibilidad ante un fallo eléctrico ubicamos dos switches de fibra, cada uno en un medio plano del chasis. Conectaremos cada uno de los switches de fibra a la infraestructura de almacenamiento previamente existente compuesta por switches de fibra y la SAN.

Una vez configurados los switches de fibra, cada uno permitirá que cada disco (LUN) sea accesible por un camino distinto. De esta forma el sistema operativo ofrecerá un dispositivo que agrupe todos los caminos posibles a un mismo disco.

El módulo bonding crea un nuevo interfaz que agrupa varios interfaces físicos

3. Sistema Operativo

Por su orientación a servidores, escogemos la distribución de Linux CentOS. Esta distribución deriva del código fuente de Red Hat y su descarga es gratuita. Entre las ventajas que aporta destacamos la facilidad para instalar servidores idénticos con kickstart[3,4], la posibilidad de utilizar caminos múltiples de fibra (multipath) desde la instalación del sistema, y el soporte de Xen, tanto en la parte del kernel como en herramientas específicas como virt-manager[5].

3.1. Bonding

La agregación de interfaces ethernet se hace a través del módulo bonding[6]. Este módulo crea un nuevo interfaz que agrupa varios interfaces físicos. Puesto que tenemos una red de desarrollo y otra de explotación, crearemos dos dispositivos: bond0 y bond1 en activo-pasivo. Cada uno de estos dispositivos

agrupa a dos interfaces de medios planos distintos, de forma que se garantiza la alta disponibilidad a la vez que se mantiene la separación de VLANs.

3.2. Multipath

La alta disponibilidad en el almacenamiento se consigue a través de multipath[7]. Se trata de software libre basado en el device-mapper[8] de Linux. En el ciclo de vida de un dispositivo multipath, las operaciones más comunes son: crear un nuevo dispositivo a partir de dos o más caminos, redimensionar el dispositivo, y finalmente eliminarlo. Para realizar estas operaciones es necesario actuar sobre tres subsistemas: el interfaz sysfs, el sistema de mapeo de bloques device-mapper y el software multipath.

Las operaciones sobre los discos (LUN) las llevaremos a cabo a través del interfaz sysfs montando sobre /sys [9]. Este interfaz permite escanear en busca de nuevos discos, detectar cambios de tamaños en discos existentes, y eliminar discos. Las operaciones sobre los dispositivos agrupados se realizan con las herramientas de multipath, de forma que podremos crear y eliminar nuevos dispositivos redundantes en caliente. El subsistema device-mapper es útil para redimensionar el dispositivo multipath en un sistema en funcionamiento.

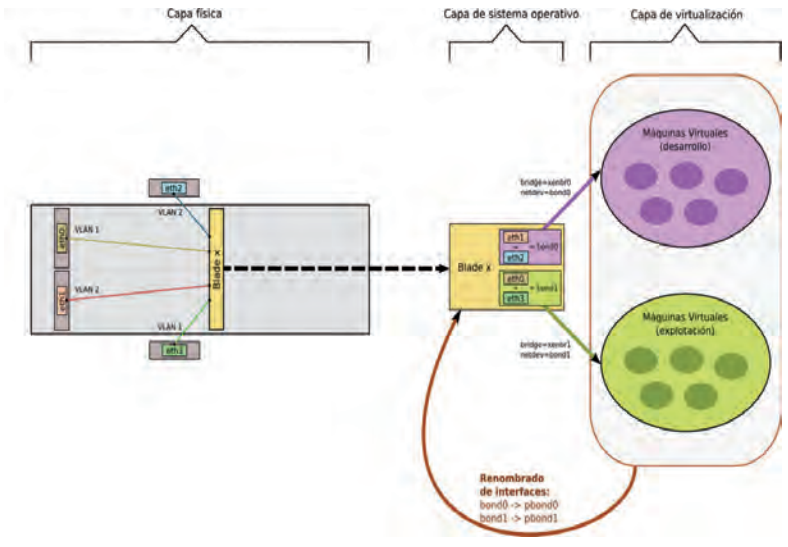
La configuración de multipath puede realizarse de varias formas: una de las más sencillas es utilizar la configuración por defecto para que detecte automáticamente el fabricante de la SAN. Utilizaremos una estrategia de listas negras y excepciones: todos los dispositivos estarán en una lista negra, y sólo aquellos dispositivos cuyo WWID escojamos, los pondremos como excepción a la lista negra, añadiéndole un alias para que el nombre del dispositivo sea representativo.

Se puede configurar el multipath por defecto para que detecte automáticamente el fabricante de la SAN

4. Xen

Los servidores de desarrollo y de explotación serán máquinas virtuales Xen[10]. El diseño y las configuraciones del hardware y del sistema operativo anfitrión deben transmitirse a las máquinas virtuales, de forma que éstas vean un hardware sin duplicidades, pero se apoyen en un sistema redundado que proporciona alta disponibilidad.

FIGURA 3. ESQUEMA DE ALTA DISPONIBILIDAD PROPAGADA A LAS MÁQUINAS XEN



El diseño y las configuraciones del hardware y del sistema operativo anfitrión deben transmitirse a las máquinas virtuales

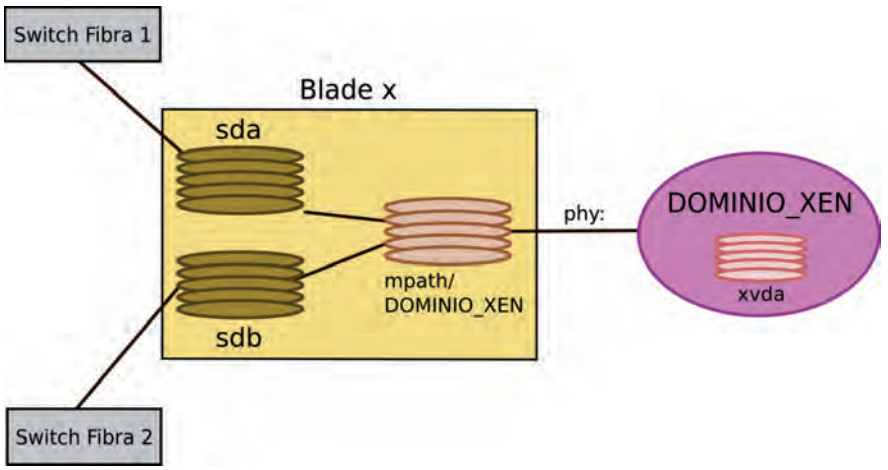
Para configurar la red, se crearán dos puentes (bridges), cada uno a un interfaz de bonding. De esta forma el puente xenbr0 corresponderá al bond0 y el xenbr1 al bond1 **(figura 3)**. Dependiendo si la máquina virtual está en la VLAN de desarrollo o de explotación, la asociaremos a un puente u otro.

Desde dentro de la máquina Xen, tan sólo se verá un único interfaz eth0. Al tener en todos los anfitriones ambos puentes, podremos alojar en un mismo anfitrión máquinas virtuales de ambas redes.

El almacenamiento se realizará directamente sobre dispositivos multipath **(figura 4)**. Los alias de multipath permiten adoptar criterios homogéneos de nombrado, por ejemplo, podemos llamar a cada dispositivo multipath con el mismo nombre que la máquina virtual. El uso de multipath no sólo añade redundancia, sino que ofrece un excelente rendimiento en comparación con almacenar las imágenes Xen en ficheros o servirlas a través de NAS. Además, como veremos, facilita la migración en vivo entre máquinas virtuales. Puesto que el anfitrión es capaz de añadir o modificar dispositivos multipath en caliente, estos cambios pueden propagarse a las máquinas virtuales: es posible añadir nuevos discos, o modificar el tamaño de los existentes. Si en la máquina virtual utilizamos LVM con un sistema de ficheros que permita redimensionado (como ext3), obtenemos una gran flexibilidad para manipular el espacio de almacenamiento.

El uso de multipath ofrece un excelente rendimiento

FIGURA 4. LOS DISPOSITIVOS MULTIPATH CORRESPONDEN A DISCOS XEN



Para llevar a cabo migraciones en vivo hay que habilitar el rango de IPs confiables

La migración en vivo de máquinas virtuales permite mover una máquina en ejecución de un anfitrión a otro. Esto es útil para equilibrar la carga de máquinas virtuales entre distintos anfitriones. El primer paso para llevar a cabo migraciones en vivo es habilitar el rango de IPs confiables, esto es, las IPs de los anfitriones en los blades. Para realizar la migración en vivo, es necesario que ambos anfitriones (origen y destino), tengan idénticas configuraciones de dispositivos. Por una parte, la configuración del bonding resulta homogénea en todos los blades. Por otra, la configuración del multipath debe ser común para todos los anfitriones. Conseguir una configuración de multipath homogénea resulta sencillo: basta con tener un fichero común con todos los WWID y los alias. Cada anfitrión sólo creará los dispositivos multipath de los discos que detecte, ignorando las configuraciones que no le correspondan.

Una vez llevada a cabo la migración, es deseable que la nueva ubicación de la máquina virtual permanezca tras posibles reinicios del anfitrión. Para esto es necesario compartir los directorios que contienen los perfiles de las máquinas virtuales Xen, así como editar los ficheros de configuración para que cada anfitrión arranque automáticamente las máquinas que les correspondan, basándose en el

hostname de cada anfitrión. Es recomendable que los ficheros compartidos no establezcan dependencias con otras máquinas -por ejemplo, con un NFS externo-, ya que esto puede entorpecer el restablecimiento automático en caso de reinicio inesperado. Una solución sencilla para mantener sincronizados los ficheros de configuración, y a la vez mantener un histórico de las configuraciones del multipath y de los perfiles de las máquinas virtuales, es utilizar Subversión.

Trabajar con servidores en máquinas virtuales Xen permite filtrar el tráfico de red en el propio anfitrión. Si no tenemos plena confianza en una máquina virtual, podemos aplicar este filtrado con el objetivo de que no sea efectivo un cambio de MAC o IP dentro de la máquina virtual, así como para evitar que sea posible espiar (sniffer) el tráfico de red. Para esto utilizaremos la herramienta ebtables (Ethernet Bridge Tables)[11], que tiene un funcionamiento similar a iptables, y permite filtrar la MAC. Esta herramienta también permite filtrar el tráfico ARP, ya sea para filtrar la MAC, la IP o por ambos campos del paquete. La particularidad que tenemos que tener en cuenta es que los paquetes de entrada al puente por el interfaz vifX.0 (para la máquina virtual con identificador *X*) son los de salida desde dentro de la máquina virtual, y viceversa.

Linux con Xen permite construir máquinas virtuales sencillas, delegando la complejidad del hardware al anfitrión

5. Conclusiones

El uso de blades es en la actualidad una buena herramienta para consolidar servidores y gestionarlos de forma centralizada. Los sistemas operativos están dotados de elementos lógicos que permiten aprovechar los elementos hardware redundantes para mayor tolerancia a fallos. Linux con Xen nos permite construir máquinas virtuales sencillas, delegando la complejidad del hardware al anfitrión. Es posible adoptar Xen libre para trabajar con máquinas virtuales en entornos separados de desarrollo y explotación. Implantando un diseño homogéneo y siguiendo unas pautas comunes de configuración, se posibilita la migración entre anfitriones sin corte de servicio.

Referencias

[1] BladeCenter H Chassis
http://www-03.ibm.com/systems/bladecenter/hardware/chassis/bladeh/index.html

[2] Spanning Tree Protocol
http://en.wikipedia.org/wiki/Spanning_tree_protocol

[3] Kickstart Installations
http://www.redhat.com/docs/manuals/linux/RHL-9-Manual/custom-guide/ch-kickstart2.html

[4] Anaconda
http://fedoraproject.org/wiki/Anaconda

[5] Virtual Machine Manager
http://virt-manager.et.redhat.com/

[6] Linux Ethernet Bonding Driver
http://www.kernel.org/doc/Documentation/networking/bonding.txt

[7] Using Device-Mapper Multipath
http://www.redhat.com/docs/manuals/csgfs/browse/4.6/DM_Multipath/index.html

[8] Device-mapper Resource Page
http://sources.redhat.com/dm/

[9] SAN Persistent Binding and Multipathing in the 2.6 Kernel
http://dims.ncsa.illinois.edu/set/san/src/linux-mpio.pdf

[10] The Xen Hypervisor
http://www.xen.org/

[11] Ebtables
http://ebtables.sourceforge.net/

Virginio García López
(virginio@um.es)
Jose Fco. Hidalgo Céspedes
(jhidalgo@um.es)
Universidad de Murcia