Almacenamiento Orientado a Bloques, Flexible, Escalable y Seguro Sobre Redes IP

Secure Block-Oriented Storage Over IP Networks

Pedro Martínez-Juliá, Antonio F. Gómez-Skarmeta

Resumen

El presente trabajo estudia la viabilidad de una solución basada en IPSec, NBD y RAID para la construcción de nubes de almacenamiento. Esta solución puede ser utilizada tanto por los proveedores como por los consumidores de los servicios de almacenamiento para aprovechar al máximo sus características distribuidas, construyendo sistemas de almacenamiento distribuido de gran capacidad y tolerancia a fallos.

Palabras clave: IPsec, NBD, RAID, nubes de almacenamiento, servidores, proveedores.

Summary

The purpose of the present work is to study the viability of a solution to build storage clouds using IPSec, NBD and RAID. This solution can be used both by the providers and the consumers of the cloud storage services to get the maximum from their distributed features, building high-capacity and fault-tolerant distributed storage systems.

Keywords: IPsec, NBD, RAID, cloud storage, servers and providers.

Gracias a esta solución se pueden construir sistemas de almacenamiento distribuido de gran capacidad y tolerancia a fallos

1. Introducción

Los servicios emergentes en torno a Cloud Computing[1, 2] ofrecen la posibilidad de externalizar diversos recursos que generalmente se encontraban en los servidores propios de una organización y muchas veces localizados en sus propias instalaciones. Entre éstos podemos encontrar almacenamiento ofrecido como servicio, o Cloud Storage, diseñado para complementar a cualquier servicio de la nube que necesite almacenar información, como pueden ser los servicios de plataforma[3].

El Cloud Computing ofrece la posibilidad de externalizar recursos que se encontraban en los servidores propios de una organización

En el presente trabajo analizamos una solución para construir y consumir nubes de almacenamiento de altas prestaciones. Un servicio de almacenamiento distribuido en la nube, debido a los requisitos que tienen los sistemas que necesitan este tipo de servicio, debe ofrecer una serie de características. A continuación se detallan algunas de las más importantes:

- Permitir que la elección y gestión del sistema de ficheros se realice en los equipos del consumidor, pudiendo seleccionar la forma más adecuada de almacenamiento para el objetivo de los clientes del servicio.
- Soportar la distribución y la réplica del almacenamiento entre distintos servidores y proveedores de forma transparente para las aplicaciones que utilizan el servicio.
- Poder ampliar o reducir el espacio de almacenamiento disponible de forma dinámica y así ajustarse a las necesidades de los clientes en cada momento.
- Consumir varios servicios de almacenamiento en paralelo para poder distribuir la información en distintos proveedores y superar el ancho de banda de un único proveedor.
- Que las comunicaciones sean seguras sin provocar un gran impacto en el rendimiento.

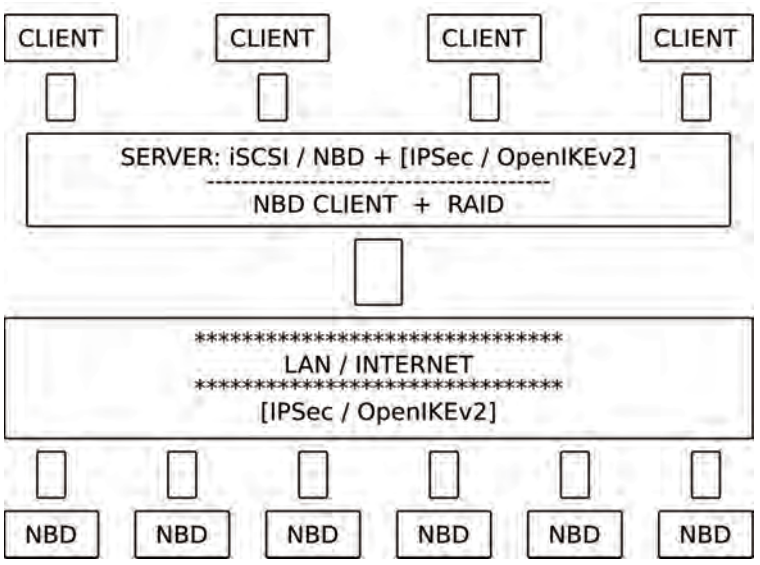
Para proveer las características detalladas proponemos utilizar la tecnología RAID aplicada sobre dispositivos de bloques distribuidos como NBD o iSCSI[4, 5] transportando los bloques de datos sobre redes IP y manteniendo la seguridad de las comunicaciones mediante IPSec.

Esta solución está especialmente enfocada a dos escenarios modelo. Por un lado, mostramos un escenario donde un proveedor de servicios agrega el espacio de varios servidores de almacenamiento y, por otro lado, mostramos un escenario donde un cliente desea utilizar de forma unificada el almacenamiento ofrecido por varias fuentes.

Como se puede apreciar en la **figura 1**, el escenario de producción para proveedores está formado por diversos servidores que ofrecen dispositivos de almacenamiento distribuido (NBD) que, a través de una red local o de Internet, son utilizados por un cliente de almacenamiento distribuido (NBD CLIENT) ubicado en un servidor que gestiona un dispositivo RAID para agregar los distintos dispositivos utilizados. Este servidor es utilizado por los diversos clientes que puedan acceder al servicio de almacenamiento general que ofrece el proveedor.

El escenario para proveedores está formado por servidores con la disposición de almacenamiento distribuido (NBD)

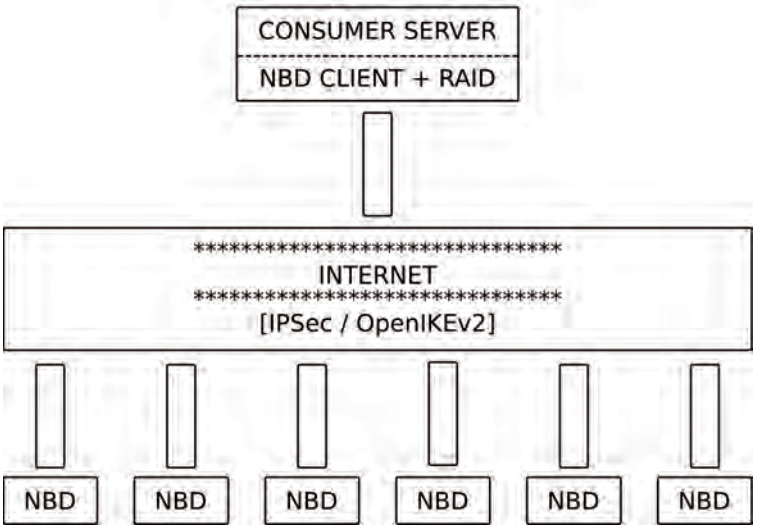
FIGURA 1. ESCENARIO DE PRODUCCIÓN PARA PROVEEDORES



En el escenario de producción para consumidores se aprecian los servidores, el mecanismo de comunicaciones y el servidor consumidor

En la **figura 2** mostramos la estructura del escenario de producción para consumidores en la que se pueden apreciar los servidores (NBD), el mecanismo de comunicaciones (INTERNET) y el servidor consumidor (CONSUMER SERVER). Los proveedores de servicios ofrecen, a través de Internet, su espacio de almacenamiento mediante una tecnología de almacenamiento distribuido (NBD) que a su vez es utilizada por los clientes para acceder al servicio. Una vez que los clientes han accedido a los dispositivos de almacenamiento distribuido, se encargarán de agregarlo mediante RAID y así ofrecer un único dispositivo a las aplicaciones que dispongan en sus servidores.

FIGURA 2. ESCENARIO DE PRODUCCIÓN PARA CONSUMIDORES



El UsenetDHT y el WheelFS son otras soluciones que ofrecen almacenamiento distribuido en la nube

Existen otras soluciones más complejas para ofrecer almacenamiento distribuido en la nube, como podría ser UsenetDHT[6] o WheelFS[7, 8], pero están orientadas hacia su uso directo por las aplicaciones mediante APIs. Este enfoque obliga a modificar las aplicaciones existentes que deseen utilizar el servicio y no ofrece la posibilidad de que un cliente pueda decidir aspectos avanzados sobre la configuración de su espacio de almacenamiento.

A continuación, en la sección 2, detallamos el trabajo realizado, en la sección 3 comentamos los resultados obtenidos y, por último, en la sección 4 exponemos nuestras conclusiones y una breve indicación acerca de las posibilidades futuras de trabajo en este entorno.

2. Trabajo realizado

Se ha realizado un estudio de factibilidad de la solución, construyendo un escenario de experimentación con servidores NBD

Para comprobar la viabilidad y utilidad de la solución que proponemos en el presente trabajo se han realizado experimentos sobre diversos escenarios. En ellos se han utilizado tres equipos, dos actuando de servidores y uno de cliente. Cada servidor exportaba dos dispositivos NBD para tener un total de 4 dispositivos.

Inicialmente se ha realizado un estudio de factibilidad de la solución, construyendo un escenario de experimentación con varios servidores NBD conectados mediante enlaces de 100 Mbits a un switch 100/1000 y un cliente NBD conectado al mismo switch mediante un enlace de 1 Gbit. El switch utilizado estaba siendo compartido con otros sistemas, por lo que los resultados se acercarían a los obtenidos en un escenario de producción real. Los dispositivos remotos se agregaron mediante distintos niveles RAID (0, 1, 5, 6 y 10) y se utilizó el sistema de ficheros EXT3 montado en modo sincronizado. Para comprobar el rendimiento se realizaron varias pruebas para cada nivel de RAID, almacenando múltiples ficheros de diversos tamaños y midiendo el tiempo que tardó cada uno para obtener la velocidad media de cada operación.

Una vez finalizada la prueba inicial, se realizaron pruebas similares utilizando otros mecanismos de interconexión, esta vez dedicados (sin otros equipos conectados), y así poder comprobar el rendimiento de la solución a través de los mismos. Por un lado se realizaron las pruebas utilizando un switch Gigabit

Ethernet donde todos los equipos estaban conectados con enlaces de 1 Gbit y, por otro lado, con los equipos conectados con enlaces de 1 Gbit a una red CWDM para tener una pequña noción del rendimiento de la solución atravesando el backbone de un proveedor.

Por último se realizaron las mismas pruebas utilizando la red de investigación PASITO (Plataforma de Análisis de Servicios de Telecomunicaciones). A diferencia de los escenarios anteriores, en este escenario se han utilizado 4 servidores distribuidos en diversos emplazamientos separados y un cliente situado en nuestras instalaciones. Los servidores han sido provistos en forma de máquina virtual por las siguientes entidades conectadas a la red PASITO:

- Universidad Carlos III de Madrid (UC3M)
- Universidad Politécnica de Valencia (UPV)
- Universidad Autónoma de Madrid (UAM)
- Universidad del País Vasco (EHU)

Este escenario nos da una noción del comportamiento de la solución atravesando multitud de equipos de interconexión distribuidos en un área muy extensa.

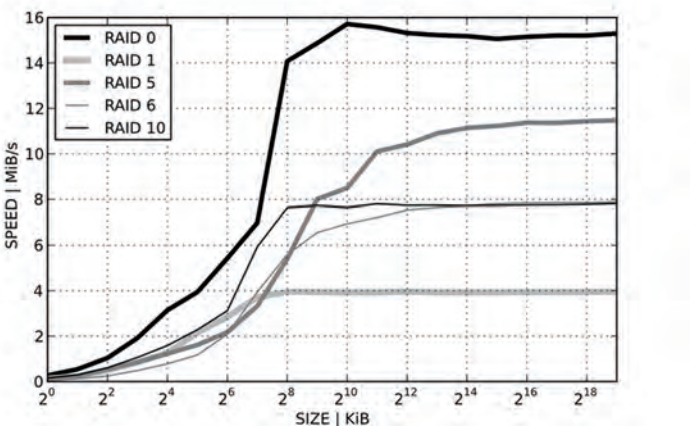
3. Resultados

De las diversas pruebas realizadas se han extraído mediciones para analizar el ancho de banda obtenido en cada escenario y la latencia observada en los escenarios experimentales.

3.1 Ancho de banda conseguido

En la **figura 3** mostramos la evolución del ancho de banda obtenido en el primer escenario, donde los equipos servidores estaban conectados mediante un cable de 100 Mbits y el equipo cliente estaba conectado mediante un cable de 1 Gbit. Se puede apreciar cómo los niveles RAID de mayor redundancia de datos (ya sea por copia-espejo o por cálculo de checksum) consiguen velocidades inferiores, mientras que los niveles más planos (RAID 0 y RAID 5) tienen un mayor aprovechamiento de los recursos de red, aproximándose al máximo teórico de la red que, para este escenario es de 200 Mbits (ya que cada uno de los servidores estaba conectado con un enlace de 100 Mbits). Además, se puede apreciar que los niveles de RAID que no necesitan cálculo de checksum convergen antes que el resto a su velocidad máxima.

FIGURA 3. SWITCH 100/1000, ANCHO DE BANDA TIEMPO DE RESPUESTA DEL MYSQL AUMENTANDO EL NÚMERO DE FILAS



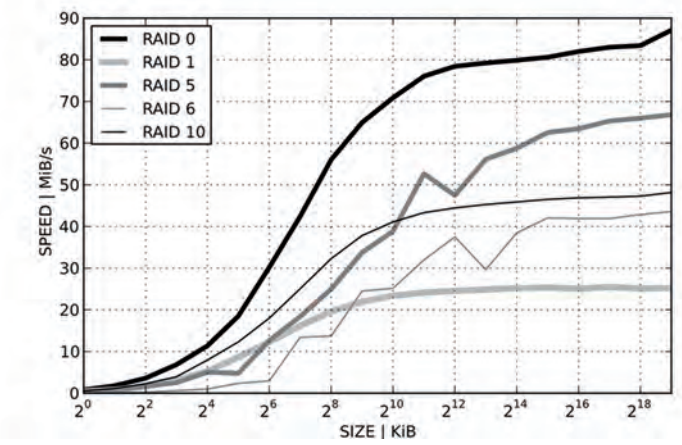
De las pruebas realizadas, se han extraído mediciones para analizar el ancho de banda obtenido en cada escenario y la latencia observada en los escenarios experimentales

Los niveles RAID más planos aprovechan más los recursos de red

En la **figura 4**, se muestran los resultados obtenidos utilizando un switch Gigabit Ethernet y enlaces de 1 Gbit para cada equipo. En estos resultados se pueden apreciar todos los rasgos de la prueba anterior salvo que la convergencia hasta la velocidad máxima de almacenamiento (que son 100 MiBs por el efecto del protocolo TCP/IP sobre el enlace de 1 Gbit) es más suave debido al tiempo inicial necesario para conseguir las velocidades mayores. Además, se puede apreciar que el comportamiento del sistema para los niveles 5 y 6 de RAID es mucho menos estable que para los demás, principalmente por la necesidad de calcular checksum en cada almacenamiento. Por el mismo motivo se puede apreciar que el rendimiento de RAID 10 es mayor que el de RAID 6, teniendo en cuenta que ambos proveen la misma seguridad, RAID 10 sería más recomendable.

El rendimiento de RAID 10 es mayor que el de RAID 6

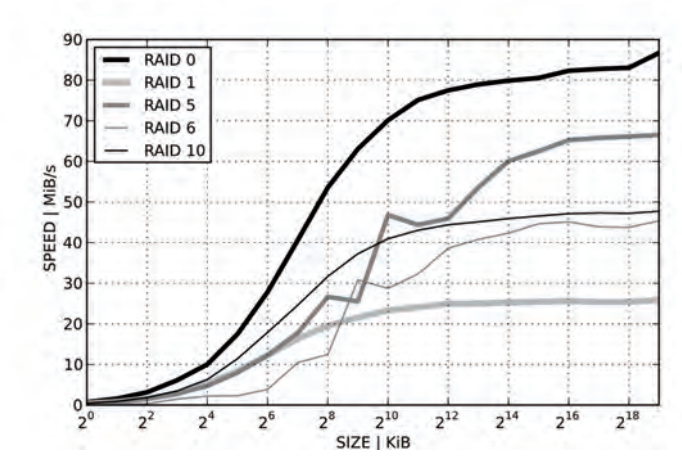
FIGURA 4. SWITCH GIGABIT ETHERNET, ANCHO DE BANDA



En la **figura 5** mostramos la evolución del ancho de banda con respecto al tamaño de fichero para los distintos niveles RAID probados sobre una red CWDM. En este caso, el comportamiento es prácticamente igual al del caso anterior (Switch Gigabit Ethernet). Se puede observar, una vez más, que el ancho de banda está muy cerca del máximo ofrecido por la red, que está al rededor de los 100 MiB/s para el nivel 0 (stripped) de RAID, 66 MiB/s para el nivel 5, 50 MiB/s para los niveles 10 y 6, y de 25 MiB/s para el nivel 1 (Mirror).

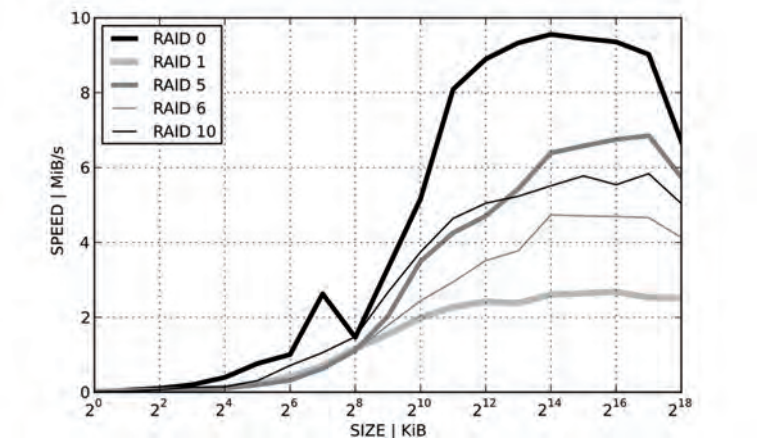
Sobre una red CWDM se puede observar que el ancho de banda está muy cerca del máximo ofrecido por la red

FIGURA 5. RED CWDM, ANCHO DE BANDA



En la **figura 6** mostramos el ancho de banda obtenido en las pruebas sobre la red PASITO. El principal detalle de estos resultados es que, aunque los enlaces de los distintos servidores era grande, la latencia de la red ha provocado una reducción drástica de la velocidad. Aún así, para ciertos tamaños de fichero se llega a velocidades muy cercanas a los 10 MiB/s para RAID 0 y superior a los 2 MiB/s para RAID 1, velocidades muy interesantes para dispositivos de almacenamiento tan distribuidos. Además, a partir de cierto tamaño de fichero se reduce la velocidad de transferencia debido, principalmente, al control de ancho de banda de esa red.

FIGURA 6. RED PASITO, ANCHO DE BANDA



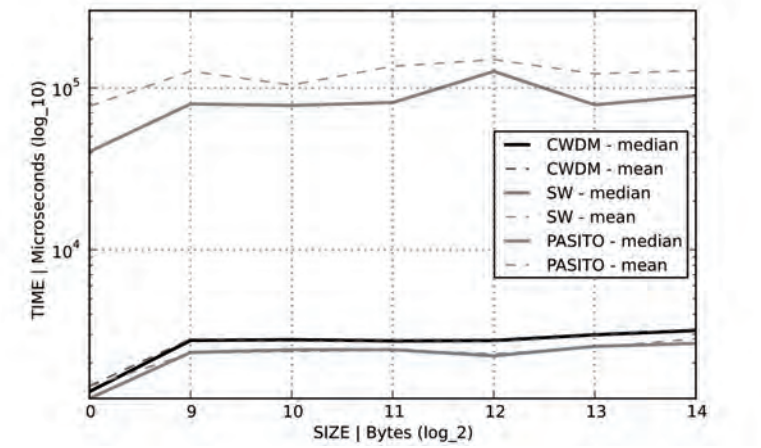
Se han realizado pruebas sobre la red PASITO

3.2 Comparación de latencias

En la **figura 7** mostramos un gráfico comparativo de las latencias observadas en los escenarios analizados en entornos controlados, sin incluir el escenario con el switch 100/1000. De estos resultados cabe destacar la gran diferencia de latencias entre los experimentos realizados sobre mecanismos de interconexión locales y el experimento realizado sobre la red PASITO, a la vez que la similitud de resultados entre la latencia producida por un switch Gigabit Ethernet, que tiene que realizar las tareas de conmutación local, y la producida por la red CWDM, que asignaba una lambda dedicada a cada conexión.

Destaca la gran diferencia de latencias entre los experimentos realizados

FIGURA 7. COMPARACIÓN DE LATENCIAS



4. Conclusiones y trabajo futuro

Debido a la reducida sobrecarga que introduce la capa RAID, la velocidad observada en los experimentos se acerca al ancho de banda teórico de cada escenario, hecho muy favorable para la solución propuesta. Además, la introducción de la red CWDM en las pruebas mantuvo el ancho de banda, provocando únicamente un incremento de 500 microsegundos en la latencia, demostrando que sería viable atravesar el backbone de un operador. Por otro lado, donde sí afecta la latencia provocada por los mecanismos de interconexión probados es en el caso de la red PASITO, aunque no lo suficiente como para invalidar la solución, ya que se consiguen velocidades de almacenamiento elevadas. A la vista de estos resultados, podemos concluir que la solución propuesta es viable y válida para la construcción de nubes de almacenamiento. Una vez confirmada la viabilidad de la solución NBD+RAID+IPSec, el siguiente paso es comprobar y definir formalmente el comportamiento del sistema ante los posibles fallos, ya provengan de la red o de los dispositivos. Otra actividad a tener en cuenta es el diseño de un prototipo de gestor autónomo para servicios de almacenamiento, consumiendo y exponiendo una interfaz de gestión estándar, como la propuesta en la especificación CDMI (Cloud Data Management Interface) de SNIA[9].

Según los resultados, la solución propuesta es viable y válida para la construcción de nubes de almacenamiento

Referencias

[1] Jeremy Geelan. *Twenty-One Experts Define Cloud Computing*. Virtualization, August 2008.

[2] Brian Hayes. *Cloud Computing. Commun*. ACM, 51(7):9 –11, 2008.

[3] G. Lawton. *Developing Software Online With Platform-as-a-Service Technology.* Computer, 41(6):13–15, 2008.

[4] David Sacks. *Demystifying DAS, SAN, NAS, NAS Gateways, Fibre Channel, and iSCSI*. IBM Storage Networking, June 2001.

[5] Wolfgang Singer. *NAS and iSCSI Technology Overview.* SNIA Technical Tutorials, 2007.

[6] Emil Sit, Robert Morris, and M. Frans Kaashoek. UsenetDHT: *A low- overhead design for Usenet. In Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation,* pages 133–146, Berkeley, CA, USA, 2008. USENIX Association.

[7] Jeremy Stribling, Yair Sovran, Irene Zhang, Xavid Pretzer, Jinyang Li, M. Frans Kaashoek, and Robert Morris. *Flexible, wide-area storage for distributed systems with WheelFS. In Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation*, pages 43–58, Berkeley, CA, USA, 2009. USENIX Association.

[8] Irene Zhang. *Efficient file distribution in a flexible, wide-area file system.* Master's thesis, Massachusetts Institute of Technology, 2009.

[9] Storage Networking Industry Association. *http://www.snia.org*

Pedro Martínez-Juliá
(pedromj@um.es)
Antonio F. Gómez-Skarmeta
(skarmeta@um.es)
Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia