PLATAFORMA CLÚSTER BASADA EN CENTOS

Área de conocimiento: Redes y Telecomunicaciones

Raúl Hernández Palacios, Felipe de Jesús Núñez Cárdenas, Javier Hervert Hernández, Miriam De la Cruz Bautista.

Área académica de Sistemas Computacionales. Escuela Superior de Huejutla, Universidad Autónoma del Estado de Hidalgo.

Contacto: raulhp@ugr.es – rapalacios81@hotmail.es Acceso al Corredor Industrial S/N, Parque de Poblamiento, Huejutla de Reyes, Hidalgo. Teléfono: 7717172000 Ext. 5880 y 5881.

Resumen---- El diseño y la implementación de un clúster para aplicaciones de alto rendimiento y disponibilidad, que trabaje con sistema de ficheros paralelo utilizando PVFS, mediante un RAID sobre una plataforma Linux hace que la distribución de los procesos sea más confiable y de bajo costo, ya que al combinar la utilización de un RAID-10 por software y PVFS2 hace más tolerante a fallos el sistema de almacenamiento, evitando al mínimo la pérdida de datos ya que trabajan mediante replicación dentro del sistema RAID. Con esto se pretende cubrir cualquier expectativa sin importar el tipo de proceso que se ejecute ya que será capaz de adaptarse a sus necesidades, así como poder ampliar el número de nodos en el clúster para poder obtener una escalabilidad, un mejor rendimiento y mayor desempeño de la plataforma.

Palabras clave ---Sistema de ficheros paralelos, RAID, replicación, clúster.

I. Introducción

Actualmente las aplicaciones necesitan un procesamiento alto que las computadoras personales actuales no cubren. La computación de alto rendimiento basada en clúster ofrece capacidades de cómputo a gran escala, en la cual involucra desde un sistema de archivos que se encarga de almacenar los datos en los nodos del clúster, hasta la comunicación de los nodos del cluster. La E/S se ha convertido en un aspecto importante en la computación de alto rendimiento ó altas prestaciones. En años recientes, las computadoras paralelas, incluyendo clústers en particular, se están incrementando en cantidad de nodos de procesamiento como forma de conseguir mayores prestaciones globales del sistema. Una de las investigaciones actuales para este proyecto radica en la mejora de las comunicaciones entre los nodos del cluster, empleando protocolos de comunicación más ligeros y redes Gigabit Ethernet o Infiniband, con la finalidad de que las comunicaciones en un sistema clúster no se convierta en el cuello de botella del sistema global. Otra de las investigaciones que surgirán es el de mantener un sistema, además de alto rendimiento, que el mismo ofrezca tolerancia a cualquier fallo, principalmente en los servidores de datos para que las aplicaciones estén siempre disponibles para los usuarios, a nivel de red es posible

también dar esta disponibilidad al sistema, primeramente mediante la técnica Channel Bonding *[1]* del núcleo de Linux que permite dar soporte de disponibilidad y balanceo de carga en las interfaces de red de cada nodo del sistema clúster, se obtiene de esta manera una redundancia en el sistema global y las aplicaciones.

El resto del artículo esta organizado de la siguiente manera: la sección II describe el estado del arte de la temática afrontada; la sección III es la descripción metodología; la sección IV muestra la experiencia obtenida y los primeros resultados de la plataforma; la sección V algunas conclusiones y el trabajo futuro, y en la sección VI algunas referencias bibliográficas.

II. Estado del Arte

Un clúster se puede definir, según lo expresado en *[2]* como "*la agrupación de ordenadores (generalmente computadoras personales) conectadas mediante una red y trabajando en un problema de tamaño considerable que ha sido dividido para ser procesado de forma paralela*".

Otra definición que puede ser considerada lo bastante concreta en *[3]* que define clúster como "*Un conjunto de máquinas unidas por una red de comunicación trabajando por un servicio en conjunto"*, teniendo en cuenta que el término máquina hace relación únicamente a las computadoras personales.

II.1 Sistema RAID

RAID (Redundant Array of Independent Disk) *[4]*. La idea central en RAID es replicar datos sobre varios discos de manera que los datos no se pierdan si alguno de los discos falla. Existen diversas configuraciones RAID con diferentes características en rendimiento y formas de replicación de datos. RAID 10 es una combinación de los RAID 0 y 1, que proporciona velocidad y tolerancia al fallo simultáneamente. El nivel de RAID 10 fracciona los datos para mejorar el rendimiento, pero también utiliza un conjunto de discos duplicados para conseguir redundancia de datos. Al ser una variedad de RAID híbrida, RAID 10 combina las ventajas de rendimiento de RAID 0 con la redundancia que aporta RAID 1. Sin embargo, la principal desventaja es que requiere un mínimo de cuatro unidades y sólo dos de ellas se utilizan para el almacenamiento de datos. Las unidades se deben añadir en pares cuando se aumenta la capacidad, lo que multiplica por dos los costes de almacenamiento. La elección de esta versión de RAID se ha dado por la capacidad de discos de almacenamiento que tienen los nodos para la plataforma desarrollada.

II.2 Sistema PVFS2

PVFS *[5]* es un sistema de ficheros paralelo gratuito para Linux que actualmente se encuentra en su segunda versión. Para implementar alta disponibilidad en PVFS2 es necesario utilizar Heartbeat *[6]* que permite verificar el estado de cada uno de los servidores de datos o metadatos del sistema de almacenamiento. Pero esta

alternativa requiere que el almacenamiento esté compartido (con una SAN). Para permitir tolerancia a fallos, evitando compartir el almacenamiento, se ha agregado a PVFS2 replicación de datos en el lado de los servidores como se describe en *[7]*. De esta forma se evita el coste de usar una SAN y se pueden aprovechar los discos incluidos en los nodos del clúster. En la implementación actual de replicación un bloque de datos de archivo es almacenado en dos servidores diferentes con su respectiva copia de bloque.

PVFS en su primera versión implementa una alternativa para lograr la tolerancia a fallos, en *[8]* son consideradas las escrituras grandes con RAID5 y escrituras pequeñas con RAID1.

PVFS2 utiliza, como capa de comunicaciones, la librearía BMI *[9]* (Buffered Message Interface) que soporta diversas tecnologías de red, dentro las que se encuentran Infiniband y Mirynet, que permiten alcanzar grandes anchos de banda y baja latencia en las comunicaciones del sistema, pero a un costo elevado debido a la infraestructura física necesaria para mantener todo el sistema de comunicaciones de un clúster. En cambio la tecnología Ethernet, también soportada por BMI, que son las interfaces de red que comercialmente se encuentran a nuestro alcance y a un ancho de banda conveniente de 1000Mbps.

II.3 Comunicación con SSH

SSH ó Secure SHell *[10]* es un protocolo que facilita las comunicaciones seguras entre dos sistemas usando una arquitectura cliente/servidor y que permite a los usuarios conectarse a un host remotamente. A diferencia de otros protocolos de comunicación remota tales como FTP o Telnet, SSH encripta la sesión de conexión, haciendo posible la integridad y seguridad de los datos en la comunicación.

III. Descripción de la metodología

El esquema que se muestra en la Fig. 1 consiste en seguir una serie de etapas; la etapa 1 consiste en analizar los requerimientos tanto de hardware y software que se tiene y que se necesite, en la segunda etapa se realiza un diseño global en el cual nos basaremos posteriormente, en la etapa 3 se desarrollara el prototipo mas adecuado a las necesidades en el cual se pondrá en práctica todo lo que se analizó al principio para hacer el prototipo, en la etapa 4 se realizarán las pruebas necesarias para diagnosticar su funcionalidad, y de este punto depende si se concluye o no la última etapa ya que debe de cumplir con todo lo requerido para dar un funcionamiento de calidad o de lo contrario se regresará a la etapa 1, si cumple con los requerimientos planteados se pasa a la etapa 5 que es la entrega final.

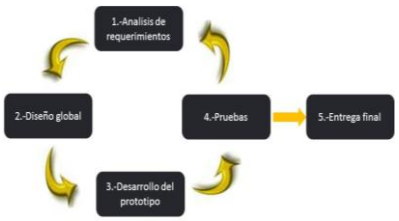

Fig. 1 Esquema grafico sobre la metodologia evolutiva empleada

IV. Resultados experimentales

IV.1 Evaluación de funcionamiento del sistema.

Para la primera fase, se realizaron las configuraciones necesarias en únicamente dos de los nodos que formarán el clúster, en primer lugar la instalación del sistema operativo Linux CentOS 5.8 con kernel 2.6.18-308.13.1.el5, además de las configuraciones de la conectividad de las tarjetas de red y de la comunicación segura con SSH, para la interconexión de los nodos se usa un Switch con capacidad de soportar Gigabit Ethernet (1000Mbps). De la misma manera se han realizado las configuraciones necesarias de la librería MPI para dar soporte al sistema de ficheros PVFS2.

IV.2 Experimentación con RAID-10

En la Fig. 2 muestra el panorama de funcionamiento lógico del sistema RAID-10, todas las réplicas del Disco1 y Disco3, se hacen en el Disco2 y Disco4 respectivamente. Esta misma configuración se mantiene en cada uno de los nodos del clúster que tienen el rol de servidor de datos de la configuración de PVFS2.

La experimentación se obtuvo en cada uno de los nodos de clúster para verificar que el sistema RAID que se ha configurado durante la instalación tenga el funcionando adecuado, esta prueba se realizo desde la consola de CentOS. Para ello se utilizo el comando cat /proc/mdstat el cual muestra el estado del RAID, esto se hace solo para verificar que este se muestre como activo.

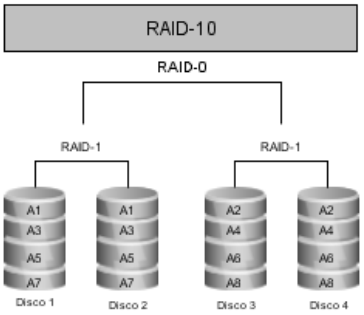

Fig. 2 Funcionamiento lógico de RAID-10

IV.3 Uso de PVFS2

En PVFS2 mantiene una estructura dónde un nodo es configurado como servidor de datos o servidor de metadatos o ambos roles al mismo tiempo y en el mismo nodo. El servidor de metadatos o metaservers almacenan los metadatos de los archivos, además de los atributos de los archivos creados, la distribución dentro del sistema global. Los nodos clientes del clúster se comunican en primer lugar con el servidor de metadatos para saber la ubicación del archivo que se quiere leer o escribir. Los servidores de datos o dataservers, comúnmente llamados servidores de E/S se encargan de almacenar trozos de los ficheros. Estos trozos se almacenan de una manera siguiendo la técnica Round Robin que se muestra en la Fig. 3. El archivo original se divide en tamaños iguales para hacer la

distribución en cada uno de los servidores de E/S.

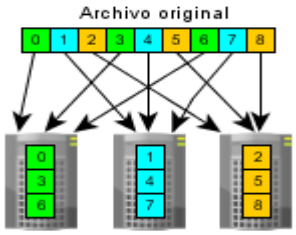

Fig. 3 Esquema de distribución Round Robin.

En la configuración inicial del sistema de alto rendimiento se define como primer prototipo utilizar dos servidores con la configuración de RAID10 en cada nodo, un nodo es configurado como servidor de E/S y servidor de metadatos, y el segundo servidor sólo como E/S. La ventaja del primer servidor es que el servidor de metadatos sólo se tendrá que comunicar con un servidor, con esto se ahorra el tráfico en la red y se evita la congestión en el caso de transferencias grandes de datos. En la Fig. 4 se observa la configuración del sistema global, cada uno de los servidores con su correspondiente configuración RAID de cuatro discos duros, accediendo a este sistema por dos nodos clientes.

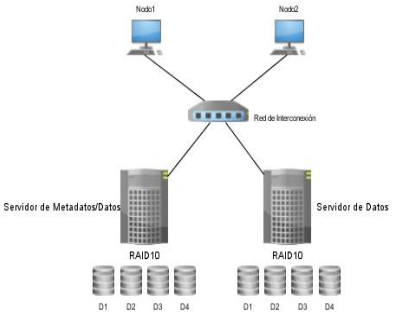

Fig. 4 Esquema de configuración lógica del sistema de almacenamiento.

V. Conclusiones y trabajo futuro.

En nuestro primer prototipo de implementación se ha podido comprobar la disponibilidad de los datos mediante la configuración RAID10, debido a que si un disco falla, existe el disco de reemplazo que almacena las copias de los datos. El sistema de archivos PVFS2 da soporte a un acceso a los datos de una manera rápida por la distribución de los mismos en todos los servidores de E/S y con el uso de Gigabit Ethernet permite alcanzar los 1000Mbps en las transferencias.

En el trabajo futuro se incrementará el número de servidores de E/S para dar un mejor rendimiento al almacenamiento de los datos que permitirá a los clientes un acceso más rápido a los archivos. De la misma forma, se mejorará el ancho de banda con el uso de la técnica Channel Bonding que viene nativa en el núcleo de Linux y en cada nodo se pretende tener dos tarjetas de red Gigabit para alcanzar un ancho de banda teórico de hasta 2000Mbps. Se implementará mayor disponibilidad de los datos con la herramienta DRBD (Distributed Replicated Block Device) que permite la distribución de bloques de datos en distintos servidores del clúster.

VI. Referencias

[1] C. University, «Parallel Virtual File System,» 2011. [En línea].

Available: www.pvfs.org.

[2] «Heartbeat,» [En línea]. Available: http://www.linux-ha.org/.

[3] E. Nieto, R. Hernández, H. Camacho, A. Díaz, M. Anguita Y J. Ortega, «Replicación de Datos en PVFS2 para Conseguir Tolerancia a Fallos,» XX Jornadas de Paralelismo, *2007.*

[4] P. Carns, W. I. Ligon, R. Ross y P. Wyckoff, «BMI: a network abstraction layer for parallel I/O,» Parallel and Distributed Processing Symposium. Proceedings. 19th IEEE International, *2005.*

[5] D. Barrett, R. Silverman y R. Byrnes, SSH, The Secure Shell: The Definitive Guide. Second Edition, O'Reilly Media, 2005.

[6] D. A. Patterson, G. Gibson y R. H. Katz, «A case for redundant arrays of inexpensive disks (RAID),» ACM SIGMOD international conference on Management of data, *vol. 17, nº ISBN:0-89791-268-3, pp. 109 - 116, 1988.*

[7] A. F. Diaz, J. Ferreira, J. Ortega, A. Cañas y A. Prieto, «CLIC: Fast Communication on Linux Clusters,» Second IEEE International Conference on Cluster Computing (CLUSTER'00), *p. 365, 2000.*

[8] D. H. Milone, A. A. Azar y L. H. Rufiner, «Supercomputadoras basadas en "Clusters" de PCs,» Revista Ciencia, Docencia y Tecnología, *vol. XIII, nº 25, pp. pp. 173-208, 2002.*

[9] M. Lauria y M. Pillai, «A high performance redundancy scheme for cluster file systems,» Proceedings of the 2003 International Conference on, *pp. 216-223, 2003.*

[10] C. V. Juan Esteban y M. A. Cristian Alejandro, Implementación de un Servidor Web Apache, *Talca, Chile, 2007.*