"IMPLEMENTACIÓN DE UN CLUSTER ALTAMENTE DISPONIBLE, UTILIZANDO SERVIDORES IBM PSERIES CON SISTEMA OPERATIVO AIX, COMPARTIENDO UN SISTEMA DE ARCHIVOS CONCURRENTES"

Paul Melitón Vergara Granda
Facultad de Ingeniería en Eléctrica y Computación
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
pvergara@espol.edu.ec

Resumen

El presente trabajo describe paso a paso la ejecución de un proyecto de instalación de servidores UNIX y sus respectivas aplicaciones en Banco DelBank. Este proyecto abarcó una amplia gama de factores que lo hicieron único, tanto por el cambio total de plataformas y aplicaciones, así como también por la participación de diferentes proveedores que interactuaban con soluciones completamente nuevas en el mercado. La compra realizada por Banco DelBank tenía como objetivo principal el implementar una solución de alta disponibilidad que proteja y garantice el servicio tanto de servidores como del software instalado en ellos, para luego instalar la aplicación GPFS para que la base de datos Oracle pueda disponer de un sistema de directorios y archivos en paralelo. Con este esquema de archivos se facilitó de gran manera la administración de la base datos y el manejo de respaldos en contraste con el manejo de datos crudos (raw). Al finalizar el proyecto se hizo una validación de los servidores, verificando que éstos se protegerían mutuamente y con tiempos de recuperación menores a los solicitados por el Cliente.

Palabras Claves: *Alta disponibilidad, HACMP, GPFS, base de datos, recuperación, Oracle RAC.*

Abstract

This present work describes step by step the implementation of the project which consisted of the installation of UNIX servers and their applications in DelBank Bank. This project covered a wide range of factors that made it unique, not only as a complete change of platforms and applications, but also as the involvement of different providers who interacted with new and similar solutions on the market. The purchase made by DelBank Bank had as main objective to implement a high availability solution that protected and guaranteed the service of both, servers and software installed on them; afterwards, installing GPFS in order to permit to Oracle database accesses and manages a system of directories and files in parallel. With this new scheme of files, it was much easier the administration and management of Oracle database and its backup, in contrast to the raw data (raw) used previously. At the end of the project, servers were validated in order to verify that they would protect each other with recovery times shorter than those requested by the Client.

Keywords*: High Availability, clustering, HACMP, GPFS, database, recovery, Oracle RAC*

1. Introducción

Antes de convertirse en una entidad bancaria, BANCO DELBANK empezó sus operaciones como Delgado Travel, una empresa cambiaria y de envío de dinero con sucursales en diferentes países. El colapso económico ocurrido a finales de los años noventa y posteriormente la dolarización adoptada en 1999, ocasionó que la tasa de desempleo se eleve considerablemente ocasionando una falta del poder adquisitivo, razón por la cual muchos de nuestros compatriotas tuvieron que emigrar a otros países en busca de un mejor porvenir económico. Tiempo después, los migrantes empezaron a enviar dinero a sus familias en Ecuador, y eligieron a Delgado Travel como la empresa indicada para realizar dichas transacciones. En vista al notable incremento de clientes que solicitaban nuevos servicios, Delgado Travel se vio en la obligación de convertirse en un banco, pero esta conversión implicaba actualizar todo su personal y sus sistemas de bases de datos y servidores, acorde a la nueva realidad.

Para el año 2000 el sistema financiero y contable de Delgado Travel se ejecutaba sobre sistema operativo SCO Unix y base de datos FoxPro en

computadores Intel. En vista al incremento considerable en el número de transacciones y de clientes, los accionistas de Delgado Travel tuvieron que buscar una solución que se acople a la nueva realidad actual de aquel entonces, tanto a nivel de hardware como de software.

A nivel de Solución de software Delgado Travel adquirió el sistema Abanks a la compañía Arango Software International, el cual se adaptaba a sus requerimientos; pero a nivel de funcional, Abanks requería que sus archivos de base de datos y aplicaciones sean accesados concurrentemente por más de un usuario desde varios servidores Unix, con un rango de transferencia elevado y cuya base de datos sea administrada con comandos de sistema operativo.

2. Antecedentes y Justificación

Se realizaron varias pruebas con diferentes servidores, tanto a nivel de alta disponibilidad como a nivel de aplicaciones y fue IBM con sus servidores pSeries la seleccionada como el proveedor de la solución, dejando atrás a competidores como SUN Microsystems y HP. IBM ofertó GPFS (General Parallel Filesystem) y a la larga fue la mejor solución que soportaba el manejo de un sistema de archivos en filesystems concurrentes. A diferencia de otras aplicaciones, GPFS permitía una tasa de transferencia mucho mayor a soluciones similares y además contaba con la certificación de Oracle 9iRAC, seleccionada como la base de datos de Abanks.

3. Especificaciones del Proyecto

El proyecto DelBank se realizó durante los años 2003 y 2004 e IBM se adjudicó el negocio luego de participar en una licitación contra varios proveedores, entre los cuales estaban empresas que cotizaban servidores Hewlett Packard, Sun Microsystems y Unisys; finalmente Banco DelBank decide contratar a IBM del Ecuador como su proveedor de servicios tanto a nivel de servidores con sus equipos pSeries como a nivel de networking y almacenamiento.

Se ofertaron dos equipos pSeries p630 como servidores de base de datos y dos pSeries p615 como servidores de aplicaciones; además de un sistema de almacenamiento Storage Fastt600 y como sistema para alta disponibilidad entre los servidores, la aplicación tipo cluster HACMP. Luego de realizar pruebas de rendimiento, Banco DelBank finalmente decide aceptar la propuesta de IBM Ecuador, en base a la relación precio/rendimiento y crecimiento futuro, teniendo en cuenta que este tipo de soluciones era la primera en ser instalada en Ecuador y pocas similares en Sur América.

4. Herramientas y tecnologías utilizadas

El cliente Banco DelBank optó por adquirir una tecnología tipo cluster en los tres niveles de arquitectura: Hardware con servidores IBM pSeries; Base de Datos con Oracle y su versión de cluster 9iRAC y la Aplicación Abanks propuesta por Arango Software. El cluster estaba formado dos pares de servidores independientes pero interconectados. El cluster estaba configurado de modo tal que podía proveer alta disponibilidad y permitir que la carga de trabajo sea transferida a un nodo secundario si el nodo principal deja de funcionar.

Una característica importante de la solución tipo cluster presentada a banco DelBank, es que se presentaban a las aplicaciones como si fueran un solo servidor, permitiendo que la administración de diversos nodos del cluster con servidores AIX se la realice desde un solo punto. El software de administración del cluster provisto por el sistema operativo AIX permitió proveer este nivel de transparencia, logrando así que todos los nodos puedan actuar como si fueran un solo servidor, los archivos fueron almacenados de modo tal que podían ser accesados por todos los nodos del cluster y por todos los usuarios del sistema.

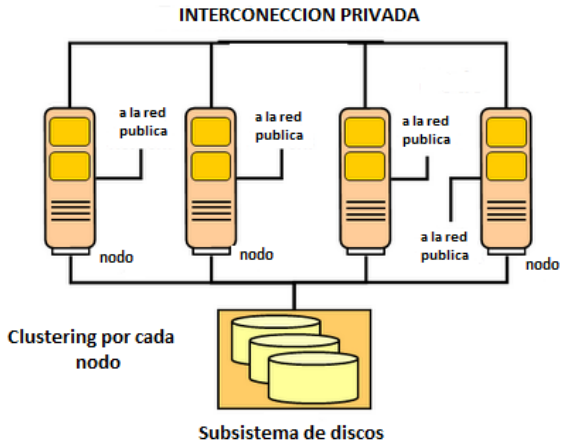

Figura 1. *Servidores compartiendo recursos*

5. Análisis del Procesos de Instalación

Al momento de la adquisición de la solución con IBM del Ecuador, la configuración inicial consistía de dos clusters diferentes. El primer cluster era del tipo GPFS que abarcaba dos nodos pSeries 630 y correspondía a los servidores de base de datos; el segundo cluster del tipo HACMP con dos nodos pSeries 615 y correspondía a los servidores de aplicaciones.

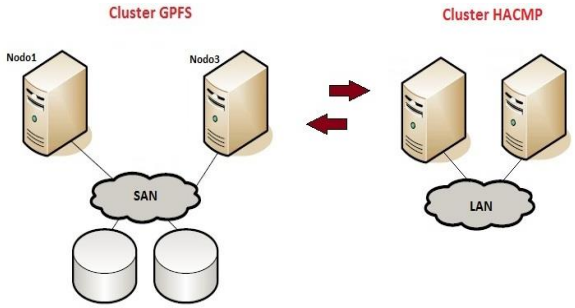

Figura 2. *Bosquejo inicial de la solución planteada*

Se realiza la instalación planteada inicialmente, es decir con dos nodos interconectados por una red local y compartiendo discos externos, sin ningún problema. Pero al momento de hacer las pruebas, el cluster no recuperaba a ninguno de los servidores, quedando en estado de falla. En vista de que este tipo de instalación era completamente nueva en Sur América no se contaba con ningún soporte local en la región. Localmente se encuentra errores en la propuesta ofrecida al cliente y es luego de la firma del contrato que se determina que el cluster GPFS no puede mantenerse con dos nodos, sino con tres nodos para mantener el quorum o instalando la versión 8.4 del software de administración de los discos FastT600 *IBM Storage Manager*. Esta versión permitía cambiar un parámetro llamado **Persistent Reservation** con el que se lograba administrar un cluster de dos nodos.

La versión del Storage Manager que arribó para Banco DelBank era la 8.3, la misma que no contaba con la opción descrita anteriormente. A esa fecha el caso problema continuaba y había que decidir si comprar la versión actualizada de Storage Manager v8.4 por USD 25.000 o instalar un tercer nodo únicamente para quorum.

Finalmente, IBM tuvo que asumir el costo de adicionar un nodo pSeries p615 para poder cumplir con la solución exigida por el cliente y para lograr el requerimiento mínimo de quórum para GPFS. Este cambio implicaba adicionar tarjetas Gigabit Ethernet y switch adicional con al menos 8 puertos Gigabit Ethernet TX para la interconexión de los nodos del cluster con HACMP y GPFS.

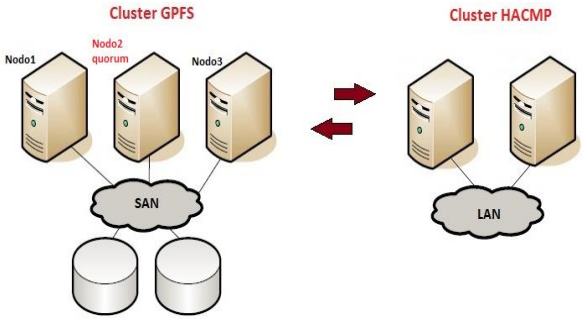

Figura 3. *Bosquejo final de la solución en DelBank.*

6. Diseño de la Solución GPFS

El diseño de la solución implicaba la respectiva verificación de hardware y software, tanto a nivel de aplicaciones, bases de datos y software de alta disponibilidad. Se procedió a la instalación física y eléctrica de los servidores dentro de un rack drawer tipo y modelo 7014-T42. Los dos servidores de base de datos y los dos de aplicaciones fueron configurados de una manera similar para permitir una mejor y fácil administración tanto a nivel de software como de hardware. En cambio, el quinto servidor pSeries 615, fue utilizado para lograr el quórum.

Luego se procedió a la Definición de arreglo de discos con su correspondiente protección. La definición del arreglo de discos debió ser configurada para sacar el mejor provecho al servidor de almacenamiento y adaptarse a los requerimientos del cliente. A nivel de discos, se utilizó RAID 1 (espejamiento de discos); aunque era la más costosa de todas por que utiliza solo el 50% de la real capacidad de todos los discos, fue la elegida por el cliente por ofrecer mejor protección contra fallas de disco, mejor performance y mejor protección de datos. Una ventaja adicional de tener el espejamiento realizado por el subsistema de discos, era que se liberaba de esta carga al sistema operativo AIX.

Se definieron las redes internas y externas y luego de esto se procedió a la asignación de direcciones a cada uno de los adaptadores de red, tanto internas como externas y diferenciando los servidores de base de datos con los de aplicaciones. Los dos servidores pSeries 615 de aplicaciones utilizaban HACMP para proteger posibles fallas de adaptadores Ethernet, fallas de procesador, fuentes de poder o falla de los mismos servidores.

Figura 4. *Redes públicas y privadas en servidores de base de datos*

No estaban conectados a las redes GPFS y Oracle RAC, pues únicamente ejecutaban las

aplicaciones de aBanks, solicitando información y requerimientos a los servidores de base de datos vía red externa, es decir la misma red utilizada por los usuarios del sistema. Las redes privadas en los servidores de aplicaciones eran las definidas por HACMP e incluían las redes seriales rs232 y eran las encargadas de identificar como última instancia si un servidor estaba caído o no; en general, estatus actual. La red de aplicaciones incluía los dos puertos seriales por servidor pSeries y dos adaptadores ethernet, el uno para servicio a usuarios y el segundo para adquirir una dirección temporal al momento de reiniciar.

Para la Interconexión entre GPFS y Oracle 9iRAC, GPFS permitió a la base de datos distribuirse en filesystems concurrentes, dejando en el pasado la obligatoriedad de crear la base de datos en almacenamiento crudo (RAW), cuya administración era demasiado complicada y desperdiciaba espacio en disco. GPFS hacía un mapeo de todo el almacenamiento como un espacio dividido en pequeños bloques (*striping*) mejorando así el acceso a información de una manera considerable. La interconexión de Oracle y GPFS implicaba conocer algunos componentes de Oracle RAC, su interconexión con las redes internas y la definición de un conjunto de parámetros para que la interacción entre estos dos niveles de software sea el óptimo.

7. Diseño de Pruebas

Para las diferentes pruebas de simulación de fallas, se crearon tres diferentes instancias de base de datos, una por cada nodo con el fin de verificar como una sentencia SQL cambiaba de instancia para terminar la operación indicada. Para las pruebas de simulación con Oracle, se utilizó dos tipos diferentes de SQLPLUS lanzados desde sesiones clientes:

Sesión Inactiva – Se tomó nota de fecha, hora y el nombre de la instancia antes y después de la falla. No había consultas activas pero si clientes conectados a una instancia en el momento de la falla. Se requería verificar si la sesión podía volver a conectarse a otra instancia y si era posible realizar consultas desde la misma sesión después de la falla.

Sesión Activa – Se tomó nota de la fecha y hora y el nombre de la instancia antes y después de la falla. Después de la primera marca de tiempo se lanzaron varias instrucciones **SELECT** y luego se inició la condición de falla. Se necesitaba verificar que la consulta que se estaba ejecutando no se afectaba por la falla y comprobar si que ésta consulta continuaba procesándose en otra instancia como resultado de una recuperación (failover).

8. Implementación de la Solución GPFS

La implementación de la solución GPFS implicaba la realización de un conjunto de procesos tanto de hardware, sistema operativo y aplicativo.

Instalación de servidores y dispositivos

Se configuraron los tres servidores pSeries del cluster GPFS, con 4GB de memoria RAM y con sistema operativo AIX; a nivel de sistema operativo, se definió un espacio para paginamiento equivalente al doble del tamaño de la memoria RAM y finalmente se procedió a realizar el espejamiento de los dos discos de sistema operativo.

Configuración de Redes

Aunque las redes internas GPFS y Oracle 9iRAC no fueron particionadas por ser solo para la interconexión y transporte de datos entre nodos y el subsistema de discos, a nivel de usuarios se instaló un firewall para proteger la red interna que compartía información con otras sucursales en otros países y mantenían accesos a la internet.

La red de usuarios no fue duplicada por el costo que significaba instalar otro switch para definir la nueva red de respaldo. En caso de falla de red de usuarios, el cliente remplazaría el switch con problemas por uno de contingencias. Pero entre la red de usuarios locales y los servidores IBM pSeries, el personal de Banco se instaló un firewall Cisco para prevenir posibles accesos no deseados de usuarios a los servidores de base de datos o aplicaciones.

Ejecución de aplicaciones Oracle

Para instalar la base de datos Oracle 9iRAC fue necesario instalar varios paquetes de software. La ejecución del software de Oracle se la hizo desde un sistema de archivos GPFS, creando así una sola copia de la imagen binaria.

Eliminación de puntos de falla

En vista a que HACMP fue instalado en únicamente en los dos servidores de aplicaciones, se utilizó un solo comando para su configuración: **xclconfig**, que era una aplicación de X-Windows que simplificaba la tarea de configurar un clúster HACMP de dos nodos. Con esta herramienta pudimos automatizar una de las cinco configuraciones predefinidas para un clúster de dos nodos.

9. Ejecución de Pruebas y Resultados

El resultado de la ejecución de pruebas incluía una lista de fallas controladas que afectarían la conexión en tiempo de balanceo de carga y failover,

así como la conmutación de la aplicación luego del error.

Las pruebas ejecutadas en este capítulo se llevaron a cabo de diferente manera con la idea de lograr el mismo resultado, simulando un ambiente real de contingencia.

Simulación de fallas de adaptadores de la red interna

La simulación de error estaba relacionada con la interconexión del clúster utilizado por Oracle9i RAC y por GPFS y la comunicación entre instancias de Oracle, comprobando cómo se produce la recuperación de instancias una vez que la interface de red asociada con la interconexión del clúster falla y como el tráfico cambia a la interconexión secundaria. Además se comprueba el comportamiento de la sesión cuando su interface falla y como sigue activa una vez que se interconecta a la red secundaria.

Simulación de fallas de adaptadores de la red de usuarios

Se deshabilitó la interface de red en0 en un nodo de aplicaciones. Un usuario ejecutaba la consulta en una instancia desde ese nodo. Luego se desconecta físicamente el cable del adaptador en0 para simular la falla de caída del adaptador de red.

En vista que en los nodos de aplicaciones existían dos tarjetas de red de 10/100 Mbps (en0 y en1) incluidas en el mainboard de los pSeries, HACMP manejaba la conmutación entre los dos adaptadores rápidamente (en cuestión de uno o dos segundos), restaurando las sesiones de usuarios casi inmediatamente sin causar mayor problema a las consultas hechas a las instancias de la base de datos.

Simulación de caídas de nodos del cluster

En esta prueba, se simulaba la caída del nodo mientras la instancia ejecutaba la respectiva consulta de prueba En vez de apagar directamente el nodo utilizando el comando **rsh shutdown**, se utilizó un archivo de activación en el nodo que iba a fallar.

10. Análisis de resultados

Una vez realizadas las pruebas de alta disponibilidad con los cinco servidores IBM pSeries, se pudo poner en práctica mucho de los conocimientos teóricos que se tenían hasta ese momento. Para el respectivo análisis de las pruebas ejecutadas, se dividieron los resultados obtenidos de las mismas en tres diferentes áreas: hardware, software operacional y base de datos.

11. Conclusiones

El proyecto en Banco DelBank fue el pionero de la instalación de clusters en Ecuador y en Sur América en general. Fue uno de los más completos no solo por la complejidad del mismo, sino por el reto que el mismo implicaba pues en su momento, siendo Ecuador el encargado de orientar a otros países en la preventa y postventa de soluciones similares.

Banco DelBank mantuvo la infraestructura tecnológica de alta disponibilidad para proteger los sistemas de información y datos críticos por muchos años.

Con la utilización de clusters altamente disponibles se logra estabilizar los servicios informáticos de una empresa en constante crecimiento. Una demostración de lo estable que ha quedado esta instalación, es que ha seguido funcionando ininterrumpidamente desde el año 2004. En el lapso de estos años, se han realizado cambios de discos, fuentes de poder y se han incrementado memoria a los servidores, pero nunca han tenido de paralizar la operación, demostrando en la práctica lo que en teoría debería pasar con este tipo de instalaciones, obteniendo valores mayores al 99,99% de disponibilidad del sistema.

12. Recomendaciones

La tecnología de clusters con GPFS ha mejorado tecnológicamente en el transcurso de los años y sigue manteniéndose líder en el mercado mundial como responsable de la disponibilidad, seguridad y confiabilidad de la infraestructura en muchos clientes. Por lo anteriormente expuesto se recomendaría que:

En universidades y centros educativos superiores, podrían dictarse materias relacionadas con este tipo de soluciones, pues involucra un amplio grupo de proveedores tanto a nivel de sistemas operativos (AIX, Solaris, Linux, HPUX), base de datos (Oracle, DB2, Informix) y aplicaciones (Baan, JD Edwards, SAP, People Soft).

La instalación de laboratorios con clusters similares es posible, pues el costo del mismo se reduce considerablemente si se instala Linux en vez de AIX y finalmente empresas proveedoras de bases de datos están abiertas a ofrecer sus productos para pruebas en centros educativos.

13. Referencias

[1] Wan Hee Kim, Paulo Queiroz, Andrei Vlad. 2011. SG24-7844-00. Implementing the IBM General Parallel File System (GPFS) in a Cross-Platform Environment. First Edition (June 2011). This edition applies to IBM AIX 6.1 TL05, IBM Virtual IO Server 2.1.3.10-FP23, IBM General Parallel File System 3.4.0.1, RedHat Enterprise Linux 5.5. ibm.com/redbooks.

[2] Octavian Lascu, Vigil Carastanef, Lifang (Lillian) Li, Michel Passet, Norbert Pistoor, James Wang. 2003. SG24-6954-00 Deploying Oracle 9*i* RAC on IBM Eserver Cluster 1600 with GPFS. First Edition (October 2003); this edition applies to Version 5, Release 2, Modification 01 of AIX and Version 9.2.0.x of Oracle9i Real Application Clusters. ibm.com/redbooks.

[3] Pedro Diaz Robles. 2004. Oracle Real Application Cluster 10g. https://sites.google.com/site/mysitepeter/oracle-real-application-cluster. First Edition (March 2004).

[4] Octavian Lascu, Zbigniew Borgosz, Josh-Daniel S. Davis, Pablo Pereira, Andrei Socoliuc. 2003. SG24-6978-00. An Introduction to the New IBM Eserver pSeries High Performance Switch. First Edition (December 2003) applies to Version 5, Release 2, Modification 2 of AIX 5L (product number 5765-E62. ibm.com/redbooks.

[5] Jorge García Delgado. 2010. UT-06 Implantación de Soluciones de Alta Disponibilidad. Primera Edición (Agosto 2010).