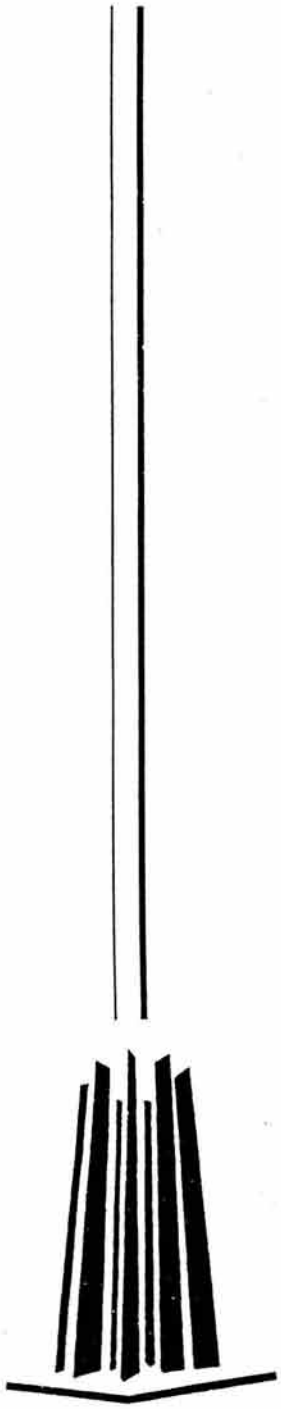
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

CAMPUS ARAGON



"Técnica RAID 0 + 1 para el Almacenamiento de Datos en Servidores Unix"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
INGENIERO EN COMPUTACIÓN

PRESENTA:
MARÍA DEL ROCIO ARELLANO PÉREZ

ASESOR: ING. JUAN GASTALDI PÉREZ

MÉXICO, D.F. 2004

A mis padres que siempre me acompañan en mi corazón.

AGRADECIMIENTOS

Agradezco a mis padres su apoyo incondicional durante todos estos años, su ejemplo, su dedicación, su fortaleza para sobrellevar los malos momentos y darme las bases sólidas para obtener los logros y satisfacciones que hasta ahora he tenido, a ellos debo mis logros y el compromiso con mis ideales, mi prójimo, mi futuro, pues supieron darme el empuje y el consejo necesarios en todo momento.

También agradezco a Álvaro, mi esposo, que ha sido guía y base para hacer las cosas de manera correcta, con calidad y compromiso, además de ser mi apoyo moral en todo momento y ayudarme a sobrellevar los momentos amargos con optimismo pero siempre con la cabeza fría. Su amor y amistad han sido muy valiosos en todos estos años felices que hemos compartido juntos.

Mis profesores han sido base en mi formación, desde los que ya no recuerdo su nombre, pues fue hace muchos años cuando me enseñaron educación básica, como los más recientes a quienes admiro no solamente por su experiencia y amplios conocimientos; sino también por la labor y misión tan importante que tienen en la vida académica de todos nosotros.

Mis amigos también han jugado un papel importante, pues compartimos experiencias invaluables que han formado nuestro carácter pero que a la vez nos permitieron disfrutar de nuestro paso por la Universidad creando lazos que aún perduran con el tiempo.

Al Ing. Juan Gastaldi por hacer aceptado participar en este proyecto tan importante para mi, aportando sus conocimientos y vocación de enseñanza, así como a los profesores Manuel Quintero, Marcelo Pérez, Ricardo Gutiérrez y Enrique García por su colaboración y buena disposición para el término satisfactorio de mi trabajo de tesis.

Gracias a todos ellos por su apoyo invaluable.

Rocío.

CONTENIDO

INTRODUCCIÓN

Una de las tareas más esenciales en el campo del cómputo es el almacenamiento de datos en dispositivos diseñados para este fin, los cuales han estado en constante desarrollo y han tenido avances significativos a lo largo de la historia del cómputo.

Los dispositivos de almacenamiento han permitido reutilizar los datos almacenados para diversas aplicaciones y han permitido avances en otras áreas del cómputo como lo son la programación o desarrollo de sistemas para una gran gama de aplicaciones que soportan procesos de la vida ordinaria del ser humano.

El contar con estos medios de almacenamiento ha ocasionado una dependencia a ellos cada vez mayor, algunas de las necesidades que tenemos cuando hablamos de almacenamiento son, por mencionar algunas :

- Mayor seguridad en los datos, es decir, evitar las pérdidas de información debido a fallas en los dispositivos de almacenamiento.
- Mayor tolerancia a fallas en los equipos de almacenamiento.
- Mayor disponibilidad de los datos que permitan a las empresas seguir con su operación diaria no importando si se ha tenido una falla en el equipo o si existen elementos en el ambiente operativo que no permitan hacer una explotación a los datos de una manera eficiente y rápida.
- Mayor capacidad en los dispositivos de almacenamiento que permitan ir creciendo conforme las necesidades lo van requiriendo.
- Mayor desempeño de los dispositivos de almacenamiento que permitan obtener la información de manera eficiente.

Debido a que los medios de almacenamiento como los discos duros, son los dispositivos más lentos dentro de los sistemas de cómputo, se han tenido que desarrollar alternativas que nos acerquen a cubrir estas necesidades, una de ellas, ampliamente utilizadas y conocidas, son las *Técnicas RAID* ó *Niveles de RAID.*

Las técnicas RAID (*Redundant Array of Independent Disks)* fueron desarrolladas para administrar y utilizar los recursos de almacenamiento de la empresa de manera más flexible y lograr ante todo mayor disponibilidad de los datos, mejor desempeño y mayor capacidad.

Los niveles de RAID son convenientes en cualquier sistema y/o aplicación que sufra crecimiento de datos almacenados y sea necesario proveerle disponibilidad, crecimiento y desempeño.

Las técnicas RAID, sin embargo deben ser aplicadas de acuerdo a las necesidades específicas de la empresa ya que una técnica puede ser conveniente para ciertas aplicaciones y para otras no, además debe considerarse un factor muy importante cuando se elija un nivel de RAID a implantar: el costo.

Este trabajo tiene como objetivo el mostrar los diferentes niveles de RAID que pueden ser aplicados, así como sus ventajas y desventajas. También un ejemplo en el cual se utilizó un nivel de RAID para solucionar el crecimiento de datos del servicio de correo electrónico de un servidor.

El capítulo 1 es una introducción a la tecnología más común para el almacenamiento de datos: los discos duros. Se expone su historia, su funcionamiento general y la manera en que interactúan con los sistemas de cómputo a través de interfaces como SCSI ó fibra óptica, por ejemplo.

El capítulo 2 muestra los niveles de RAID más conocidos, los más comerciales, los anidados y de los que solamente conocemos sus características más generales,

pues no han sido implementados comercialmente. Se exponen también, sus ventajas y sus desventajas.

En el capítulo 3 se muestran las características de una de las aplicaciones más conocidas que implementa los niveles de RAID más comerciales en el mercado: VERITAS Volume Manager. Se exponen las características del equipo de Sun Microsystems Sun StorEdge A5200 y las de Volume Manager que nos permiten realizar las técnicas RAID comentadas.

En el capítulo 4 se muestra el empleo del nivel de RAID 0+1 en un servidor de correo electrónico a través de Volume Manager y el arreglo de discos A5200, se muestran las características del equipo y las de otros que pueden también funcionar con este tipo de implementación.

Finalmente en el capítulo 5 se muestran las características más generales y beneficios de las *Redes de Almacenamiento (SAN – Storage Area Network),* las cuales representan la tendencia en cuanto a almacenamiento distribuido de datos se refiere. Este capítulo fue introducido debido a los comentarios que se realizan a lo largo del trabajo acerca de los beneficios que una SAN puede traer a un ambiente de cómputo donde el crecimiento de datos debe ser lo más flexible, disponible y crítico se refiere.

La mayoría de los términos técnicos de este trabajo se han mantenido en inglés como son conocidos en el ambiente computacional, además de no existir, para muchos de ellos, un equivalente en español.

Capítulo 1. EL ALMACENAMIENTO EN SERVIDORES UNIX.

1.1 ANTECEDENTES.

En un inicio con el desarrollo de las primeras computadoras personales e inclusive con las computadoras más grandes, no se tenia la noción de almacenamiento como lo conocemos hoy en día. Si se pretendía usar las computadoras en alguna aplicación o para correr algún programa, se tenia que escribir el programa una y otra vez para obtener los resultados, es decir, no existía la posibilidad de almacenar por un tiempo relativamente largo el mismo programa y poderlo utilizar cuantas veces fuera posible sin la necesidad de rescribirlo.

Posteriormente el medio de almacenamiento que proveyó esta funcionalidad fue el papel. Era posible tener el código de un programa a través de papeles o tarjetas perforadas que aseguraban que el programa estuviera permanentemente almacenado en un medio.

Los inconvenientes de este tipo de almacenamiento son vastos ya que no solamente representaban un método complicado de almacenar información, sino que también corrían los riesgos naturales de un papel: fragilidad, desgaste, flamable, etc. Claro, muchos de los medios de almacenamiento anteriores y actuales no son tampoco ajenos a estos riesgos.

El escribir un programa en tarjetas era una tarea nada trivial, hacerlo funcionar lo era mucho menos. Otro inconveniente de las tarjetas era precisamente que no presentaban flexibilidad en la corrección de errores, si se cometía algún error en el programa había que tomar una tarjeta, o varias, perforarlas y probar de nuevo el programa.

Posteriormente surgieron los discos flexibles en los que se puede almacenar datos por un tiempo relativamente largo. El proceso de almacenar la información

es transparente para el operador y una característica importante y que a mi juicio le ha dado validez durante tanto tiempo, es la facilidad de transportar estos mismos datos de un lugar a otro de una manera por demás sencilla.

La desventaja de los discos flexibles es su poca capacidad de almacenamiento; solamente hay unidades de unos cuantos megabytes cuando en algunas ocasiones contamos con archivos de hasta gigabytes de tamaño.

Gracias a los discos flexibles o disquetes como muchos los conocen, fue posible el almacenar programas de computadora que eran intercambiados según las necesidades. Cada vez que se quería iniciar la computadora con el sistema operativo, era necesario primero introducir el disco correspondiente y posteriormente introducir los programas a ejecutar.

Más adelante se desarrolló un dispositivo de almacenamiento denominado disco duro. Este dispositivo es conectado directamente a la computadora y permanece fijo almacenando la información tanto del sistema operativo, configuración de la maquina y datos del usuario.

Estos discos duros empezaron la pesada carga de almacenamiento de datos electrónicos procesados por las computadoras. La mayor ventaja fue la capacidad de almacenamiento y la relativa confiabilidad de los datos almacenados ya que estos discos son más resistentes a factores ambientales y al estar fijos se reduce el riesgo de traslaparse, quemarse, perderse, mojarse, etc.

Actualmente prácticamente todas las computadoras cuentan con un disco duro desde donde iniciar el sistema operativo y desde donde tomar su configuración y programas necesarios para operar.

En el caso de los servidores unix, el disco duro es un factor importante no solamente como almacenamiento, sino también como factor indispensable para

el inicio (booteo) del equipo.

1.2 LOS DISCOS DUROS COMO MEDIO DE ALMACENAMIENTO DE INFORMACIÓN.

1.2.1 HISTORIA

Los primeros discos duros que se fabricaron fueron experimentos que nos llevaron a los modernos discos actuales. Estos discos posicionaban sus cabezas lectoras directamente sobre el disco para escribir y leer los datos. Esta característica hacía a estos dispositivos demasiado lentos sobre todo cuando se trataba de accesos randómicos a la información, ya que no era posible mover las cabezas a gran velocidad.

Posteriormente en 1950 IBM creó los primeros discos duros en los que las cabezas lectoras no tenían contacto directo con los discos, al contrario, las cabezas podían estar suspendidas a una pequeña distancia del disco. Con este diseño revolucionario se puede considerar el nacimiento de los discos duros modernos.

El primer disco duro fue el IBM 305 RAMAC (Random Access Method of Accounting and Control), el cual salió al mercado en 1956. Estos primeros discos duros eran demasiado grandes y de poca capacidad de almacenamiento (apenas 5 megabytes de 7 bits), así como también su *densidad* era muy reducida.

En 1973 IBM introdujo el modelo 3340 el cual es considerado como el padre de los discos modernos. Esta unidad tenia dos *spindles* , una permanente y la otra móvil, cada una con una capacidad de 30Mb.

Otra característica de este último disco, fue la reducción en la distancia entre las cabezas lectoras y la superficie del disco. En este modelo esta distancia se

redujo a 17 micro pulgadas.

Posterior a estos modelos se fabricaron otros con nuevas características que mejoraban las de los primeros.

A continuación se listan algunos de estos y otros discos duros que tuvieron algún impacto importante.

- ***Pimer disco duro (1956):*** El disco de IBM RAMAC. Su capacidad era alrededor de 5 Mb, almacenados en discos de 50.24 pulgadas. Su densidad era de 2 000 bits por pulgada cuadrada y su velocidad era de 8800 MB/s.
- ***Pimer disco de cabezas flotantes (1962):*** El modelo 1301 de IBM reduce la distancia entre las cabezas lectoras y el disco a 250 micro pulgadas. Su capacidad era de 28 MB. Supero al RAMAC tanto en densidad como en velocidad en un 1000 %.
- ***Primer disco removible (1965):*** El modelo 2310 de IBM fue le primer disco con empaque removible.
- ***Primer disco con cabezas de ferrita (1966):*** El modelo 2314 de IBM fue el primer disco duro que uso cabezas de ferrita, posteriormente las computadoras personales (PC) lo adoptaron.
- ***Primer diseño moderno de disco duro (1973):*** El modelo 3340 de IBM, apodado "Winchester", fue introducido. Su capacidad era de 60 MB y fue considerado el antecesor de los discos modernos.
- ***Primeras cabezas de película delgada (1979):*** El modelo 3370 de IBM fue el primero en utilizar cabezas de película delgada, los cuales fueron un estándar en computadoras personales por varios anos.
- ***Primer disco de*** *form factor* ***de 8 pulgadas (1979):*** El modelo 3310 de IBM fue l primer disco de *platos (platters)* de 8 pulgadas. Su importancia radica en reducir de 14 pulgadas a 8.
- ***Primer disco de*** *form factor* ***de 5.25"(1980):*** El modelo ST-506 de Seagate fuel el primer disco de este form factor y fue utilizado en las primeras computadoras personales.

- ***Primer disco de** form factor **de 3.5" (1983):*** El modelo RO352 de Rodime fue el primero en utilizar el form factor de 3.5.[1]
- ***Primer tarjeta de expansión de disco (1985):*** Quantum introduce la *Hardcard*, un disco de 10.5 MB montado en una tarjeta de expansión ISA para computadoras personales.
- ***Primer disco de 3.5"con** actuador **de bobina** voice coil **(1986):*** El modelo CP340 de Conner Peripherals fuel el primero en utilizar este tipo de actuador.
- ***Pimer disco de 3.5" "Low-Profile" (1988):*** El modelo CP3022 fue el primer disco de 3.5" de 1" de altura que se conoce actualmente como "low-profile" y es un estándar para los discos de 3.5" modernos.
- ***Primer disco de** form factor **de 2.5" (1988):*** PrairieTek introdujo el disco platos de 2.5". [2]
- ***Primer disco en usar cabezas magneto resistivas y descodificación de datos PRML (1990):*** El modelo 681 de IBM fue un disco de 857 MB, el primero en utilizar cabezas y descodificación de este tipo.
- ***Primer disco de película Delgada (1991):*** IBM fuel el primero en sustituir el oxido por película delgada como medio en la superficie de los platos de los discos. Fueron utilizados en su mainframe "Pacifica".
- ***Primer disco de** form factor **de 1.8" (1991):*** El modelo 1820 de Integral Peripherals fue el primer disco duro con platos de 1.8".
- ***Primer disco de** form factor **de 1.3" (1992):*** El modelo C3013A de Hewlett Packard es el primer disco de 1.3".

1.2.2 OPERACIÓN DE LOS DISCOS DUROS.

Un disco duro utiliza discos planos, redondos llamados *platters* (platos); estos están cubiertos de ambos lados con una materia especial llamada *media* la cual está diseñada para almacenar la información a manera de patrones magnéticos. Los platos (platters) tienen un orificio al centro que permite sean montados en lo

[1] Este modelo fue uno de los principales estándares en el mercado.
[2] Este tamaño fue tomado como estándar para computadoras portátiles.

que llamamos *spindle*. Estos platters giran a altas velocidades y son manejados por un motor especial conectado al *spindle*. Unos dispositivos de lectura y escritura electromagnética llamados *heads* (cabezas), son montados en *sliders* y usados para grabar información en los discos o bien leerla desde ellos. Los *sliders* son montados en *arms* (brazos), los cuales están a su vez conectados a un mismo punto y son posicionados sobre la superficie del disco por un dispositivo llamado *actuator* (actuador). Una tarjeta lógica controla la actividad de los componentes y se comunica con el resto de la computadora.

Cada superficie de los platos del disco puede almacenar billones de bits de datos; sin embargo estos son organizados en pedazos (*chunks)* para permitir el acceso rápido y eficiente a la información. Cada plato tiene dos cabezas lectoras, una por encima y otra en la parte de abajo, de tal modo que un disco de 3 platos tiene seis superficies y por tanto seis cabezas lectoras. Cada plato tiene la información almacenada en círculos concéntricos llamados *tracks* (pistas). Cada track a su vez es dividido en pequeños pedazos llamados *sectores*, cada uno de los cuales puede almacenar 512 bytes de información.

Los discos son aislados del exterior para asegurar que no sea contaminado con partículas que puedan estar en el aire y que puedan dañar las cabezas lectoras.

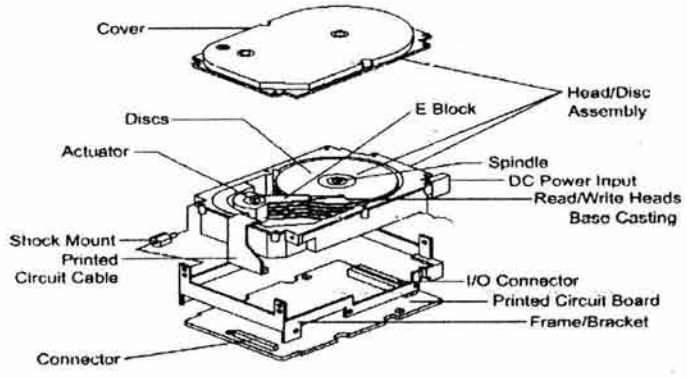


Figura 1.2.1 Un disco duro.

Platters y Media

Cada disco duro contiene uno o más platos o platters que son utilizados para almacenar la información. Estos están compuestos de dos sustancias principales: un *substrato* que forma el plato en si y le da su estructura y rigidez; y un recubrimiento magnético el cual se denomina *media* y que es precisamente el encargado de registrar los impulsos magnéticos que representan los datos. De hecho los discos duros toman su nombre precisamente de la rigidez de los platos utilizados, comparados con los discos flexibles o *floppy* los cuales son elementos, como su nombre lo indica, flexibles.

Los platters son el componente en el disco duro encargado propiamente de almacenar la información, por tanto la calidad de estos y de su recubrimiento es critica. La superficie de los platos debe ser fabricada con precisión y también debe de ser tratada para evitar contaminación dentro de los discos.

Un factor determinante en cuanto a las capacidades y performance de los discos es su tamaño, el cual es conocido generalmente como el *form factor* del disco. Existen tamaños estándar para la fabricación de los discos duros, por lo que en muchas ocasiones hemos de oír mencionar tamaños similares o estándar de estos.

Por lo regular se menciona el tamaño de un disco haciendo referencia a su *form factor,* por ejemplo, podemos hablar de un disco duro de 3.5", lo que hace referencia por lo regular al *form factor* del disco.[3]

El tamaño del platter es usualmente el mismo para discos de un form factor especifico; sin embargo, como habíamos mencionado, podemos encontrar que

[3] Normalmente el *form factor* tiene que ver con el tamaño del *platter*, aunque en algunas ocasiones esto no siempre se cumple.

los discos más actuales no siempre cumplen con esta regla.

Algunas computadoras personales, no hace mucho tiempo, usaron discos de 5.25". Actualmente este tamaño ya no es encontrado en las computadoras modernas, ahora se utilizan discos de 3.5". Cabe aclarar que tanto los discos de 5.25" y los de 3.5" son en realidad de 5.12" y 3.74" cada uno, pero por convención se utilizan los tamaños antes mencionados.

Conforme la tecnología ha evolucionado hemos visto como algunos dispositivos cambian para brindar mayor eficiencia o rapidez, o capacidad. En el caso de los discos duros se ha notado una tendencia a fabricar discos duros cada vez más pequeños.

Lo anterior es debido a que mientras el disco sea más pequeño, se obtiene un mejor desempeño debido a los siguientes factores:

- Mejor rigidez : Los platters que son más rigidos cuentan conmayor capacidad para soportar golpes y vibraciones además de que pueden ser utilizados con *spindles* de gran velocidad y otros componentes de alto desempeño.
- Facilidad de fabricación: Los mejores discos son los más lisos e uniformes por lo que su fabricación debe ser cuidadosa de estos dos factores. Los platters pequeños son más fácilmente provistos de estas dos características.
- Reducción de proporciones: La tendencia de las *spindles* es de aumentar su velocidad. Los discos pequeños son más fáciles de girar sin la necesidad de utilizar motores muy potentes. También es más rápido el girarlos desde una posición de reposo.
- Ahorro de energía: Los discos pequeños utilizan menos energía en su funcionamiento.
- Mejor desempeño en la búsqueda de información: Reduciendo el tamaño de los platters también se reduce la distancia que las cabezas lectoras tiene que recorrer para hacer una lectura o escritura randómica.

La siguiente tabla muestra algunos ejemplos de discos duros, el tamaño de sus platos así como su form factor.

Diámetro del platter	*Form factor*	*Aplicación*
5.12	5.25"	Primeras PCS, usados en algunos servidores a mitad de los anos 1990. Ya obsoletos.
3.74	3.5"	El tamaño estándar par la mayoría de las PCS.
3	3.5"	Discos de 10 000 RPM (Revoluciones por Minuto)
2.5	2.5", 3.5"	Discos para Laptop (2.5" form factor); discos de 15000 RPM (3.5" form factor)
1.8	PC Card (PCMCIA)	Tarjetas PCMCIA para laptops
1.3	PC Card (PCMCIA)	Originalmente utilizados en PCS pecunias (ya no se fabrica)
1	CompactFlash	Cámaras digitales, PC de mano (hand-held), y otros dispositivos electrónicos.

Tabla 1.2.1 Características de discos duros.

Los discos duros también pueden estar formados por uno o más platters, según su diseño. Los discos duros estándar para PC por lo regular tienen de uno a 5 platters por disco. En computadoras más grandes, especialmente en servidores, los discos utilizados tienen 12 o más platters por disco.

De cualquier modo los platters dentro de un disco están separados por una especie de anillos también ensartados en la *spindle*. Todos los platters son conectados a esa misma *spindle.*

Como habíamos mencionado, cada platter tiene dos superficies en las que es posible almacenar los datos, sin embargo no todos los fabricantes diseñan los

discos para almacenar datos en ambas, hay algunos que tienen modelos en los que se utilizan ambas superficies, pero otros en los que solamente usan una de ellas para proveer a sus discos de diferentes capacidades.

El form factor también influye en el número de platters de un disco. Los discos estándar para PC están limitados a un form factor de 1", limitando por esto el número de platters a utilizar en esos discos. Existen también los discos de 1.6", denominados "half height" los cuales son utilizados por lo regular en servidores y tienen más platters que los estándar para PC.

Como ya se mencionó, los patrones magnéticos que representan los datos, son grabados en un material muy delgado llamado *media* el cual se encuentra en la superficie de los platters de los discos, sin embargo el platters en sí también está compuesto de otro material llamado *substrato*, el cual no tiene ninguna función más que la de darle rigidez y forma al platter y contener en su superficie la capa del material que llamamos *media.*

El substrato debe ser un material rígido, fácil de trabajar, ligero, estable, inerte a campos magnéticos, barato y disponible. El material que se usa más comúnmente para fabricar los platter es el aluminio, el cual cumple con todos los factores anteriores.

Debido a la velocidad en que los platters giran, éstos deben ser ligeros y totalmente lisos. Conforme la tecnología en cuanto a los discos duros va avanzando, vamos encontrando que las cabezas lectoras son diseñadas para giran cada vez más rápidamente por lo que también es una exigencia el contar con platters hechos de algún material que soporte estas grandes velocidades sin que ocasione errores en su funcionamiento. Por esto se han hecho estudios a los substratos con que son hechos los platters y se ha encontrado como alternativa el cristal para fabricar los platters.

El cristal es más ligero y mucho más liso, además de proveer mayor rigidez a los

platters y ser más resistente a las altas temperaturas provocadas por la gran velocidad en que giran éstos.

La principal desventaja de utilizar cristal en lugar de aluminio, es precisamente su fragilidad, particularmente cuando se trata de platters muy finos o delgados. Por esto aún se investiga en otro tipo de substratos o bien, componentes de éstos. Podemos mencionar por ejemplo, un compuesto de cristal y cerámica para fabricar platters el cual tiene las ventajas mencionadas para los platos de cristal, pero la cerámica le provee de mayor resistencia.

El otro material, como mencionamos, que interviene en la fabricación de los platters para discos duros es precisamente lo que llamamos *media*, el cual es un substrato con el que se cubre la superficie del platter para grabar propiamente en él los datos.

Esta capa de *media* es un recubrimiento muy delgado de material magnético en el que se graba la información, típicamente esta capa es de apneas unas cuantas millonésimas de pulgada de grosor.

En los primeros discos duros de utilizó una *media* de óxido. Éste óxido en realidad era una especie de "moho" de óxido de hierro. Este moho de óxido de hierro era adicionado al platter a través de alguna otra sustancia para fijarlo[4].

El óxido en realidad no es caro para usarse en los platters, pero tiene algunos inconvenientes:

- El material es fácilmente dañado por el contacto de las cabezas lectoras, y
- Era útil cuando se trataba de almacenar datos con baja densidad en el platter.

Este último punto se contrapone con las tendencias tecnológicas que hemos

observado, cada vez se requiere almacenar más cantidad de información en el mismo espacio.

Hoy en día los discos duros utilizan una capa delgada de *media (thin film media)*. Una vez más esta capa delgada es un material magnético adherido a la superficie de los platters.

Existen algunas técnicas para adherir esta capa a los platters. Uno de ellos es el denominado *galvanoplastía (electroplating)*, el cual deposita el material en los platters usando un método similar al utilizado en joyería. Otro método es el denominado *chisporroteo (sputtering)*, el cual usa un proceso en el que se deposita el material a través de vapor, similar a los métodos empleados en la fabricación de semiconductores para depositar una capa extremadamente delgada en la superficie de los platters.

El segundo método tiene la ventaja de dejar una capa mucho más uniforme y lisa del material sobre los platters, siendo éste método el más utilizado por los fabricantes de discos duros hoy en día no importando su alto costo.

En comparación, la capa delgada de media (thin film media), es mucho más delgada e uniforme que el óxido mencionado en un principio; además de tener propiedades magnéticas superiores, permitiendo almacenar mayor cantidad de datos en el mismo espacio. También es mucho más resistente.

Después de aplicar esta capa de media en la superficie de los platos, éstos son cubiertos con una capa muy fina de carbón para protegerlos de contactos accidentales con las cabezas lectoras.

Actualmente se investiga en nuevas técnicas y materiales para utilizarse en la superficie de los platos y hacerlos más resistentes y con mayor capacidad de almacenamiento. Uno de éstos métodos es en el que se utiliza una solución

[4] De hecho los discos más antiguos tenían un color café en la superficie gracias al color del óxido.

química conteniendo moléculas y partículas de hierro y platino de tal modo que cuando se aplica esta solución a los platter y mediante un método de calentamiento, se provoca que ambas moléculas se mezclen formando un "enrejado" de cristales[5].

Esta nueva tecnología podría aumentar las capacidades de almacenar mucha más información en los discos en un 10 hasta un 100 por ciento, sin embargo, todavía se encuentra en investigación.

Otra característica de los platters a mencionar es que éstos están divididos en estructuras específicas para permitir el almacenamiento organizado de los datos. Cada platter está dividido en *tracks (pistas)* formando círculos concéntricos. Estas son similares a los "anillos" de los troncos de los árboles y, contrario a lo que se podría pensar, diferentes de los *tracks* de un disco musical de vinilo, ya que éste último es en realidad una espiral que recorre todo el plato y por tanto no podemos decir que sus pistas sean "anillos" concéntricos.

Cada *track* a su vez está dividida en varios pedazos llamados *sectores*. Se podría decir que un sector es la mínima unidad direccionable de información almacenada en un disco duro, y normalmente su tamaño es de 512 bytes.[6] Hoy en día los discos contienen cientos de sectores por una sola pista.

Otra característica importante de los discos duros y que a menudo nos da la referencia para saber de qué capacidad es un disco, es precisamente lo que se llama su *densidad areal (areal density)*. Esta es llamada algunas veces *bit density*, y se refiere a la cantidad de datos que pueden será almacenados en un espacio específico del platter de los discos.

La *densidad areal* es una medida "bidimensional y es obtenida a través del producto de dos medidas de densidad "unidimensionales":

---

[5] Este método es llamado por IBM como "súper enrejado de nanocristal"
[6] Los primeros discos para PC tenían 17 sectores por pista.

- *Track Density (densidad de pistas):* Esta medida hace referencia a qué tan juntas están las pistas concentricas unas de otras dentro de un platter. Por ejemplo, si tenemos un platter de 3.74" de diámetro, su radio sería de aproximadamente de 1.87". Hay que considerar que la porción más interior del plato es donde se encuentra la *spindle (espina)*, y la porción más exterior del plato tampoco puede ser utilizada. Entonces consideremos un platter de aproximadamente 1.2" de radio usable para almacenamiento. Si en ese espacio tenemos 22000 pistas, entonces la densidad de pistas (*track density)* es de aproximadamente 18333 tracks por pulgada (TPI).
- *Linear ó Recording Density (Densidad linear):* Esta es una medida que hace referencia a qué tantos bits pueden ser almacenados dentro de una pista. Si en una pulgada de una pista podemos grabar 200 000 bits de información, entonces la densidad linear para esa pista es de 200 000 bits por pulgada por pista (BPI). Cada pista en el platter es de diferente longitud, por tanto no todas las pistas son grabadas con la misma densidad linear.[7].

Tomando estas dos medidas nos lleva a determinar la *densidad areal*, medida en bits por pulgada cuadrada. Si la máxima densidad linear del disco antes mencionado (18333 TPI) es 300000 bits por pulgada por track, entonces su máxima densidad areal será de 5,500,000,000 bits por pulgada cuadrada, o bien, 5.5 Gbits/in^2.

La razón por la que la densidad no es igual en todas las pistas de un platter, es porque las pistas del exterior tienen mayor longitud que las pistas interiores. Esto significa que si grabáramos la misma cantidad de información tanto en las pistas interiores como en las exteriores, éstas últimas tendrían una densidad linear menor.

El incrementar la densidad areal supone el incrementar ambos factores mencionados, la densidad linear y la densidad de pistas; sin embargo el hacer esto involucra un mayor cuidado de otros factores debido a que el tener mayor

densidad aumenta el riesgo de interferencias entre bits, así como también la sensibilidad de las cabezas lectoras para obtener los datos respectivos.

Cabezas lectoras/escritoras

Las cabezas lectoras son dispositivos que transforman las señales eléctricas en señales magnéticas y viceversa. Las cabezas son en si pequeños electro magnetos que realizan esa conversión.

Las primeras cabezas lectoras utilizadas en los primeros discos duros (ferrita, metal y capa delgada) usan los dos principios básicos de la fuerza electromagnética:

- Aplicando una corriente eléctrica a una bobina produce un campo electromagnético, lo que es utilizado para escribir sobre el disco.
- Aplicando un campo eléctrico a una bobina provoca una corriente eléctrica, lo que es utilizado para leer información del disco.

Las cabezas lectoras en los discos actuales ya no utilizan estos principios ni esta tecnología; mas bien utilizan el principio de magnetorresistencia, en la cual algunos materiales pueden cambiar su resistencia cuando son expuestos a campos magnéticos.

Las cabezas lectoras/escritoras en sus inicios solían utilizar el mismo dispositivo tanto para escribir como para leer; actualmente se utilizan dispositivos distintos para cada una de estas operaciones.

Con la separación de estos dos elementos se tiene más cuidado en las características y capacidades de cada uno de estos, de tal modo que es posible mejorar su desempeño por separado.

---

[7] Los fabricantes por lo regular hacen referencia a la máxima densidad obtenida en sus discos.

Otra de las características de los discos duros, y que los hacen diferentes de otro tipo de almacenamiento de información, es que las cabezas lectoras/escritoras no tienen contacto directo con la superficie del disco. La razón de esto es debido a las altas velocidades en que giran los discos y la necesidad de las cabezas de ir de un lado a otro con gran rapidez para hacer una lectura o escritura randómica.

Los discos actuales utilizan cabezas que no hacen contacto con la superficie que están magnetizando. Este espacio entre las cabezas y el platter es conocido como *altura flotante (floating height o flying height).*

Las cabezas en realidad están diseñadas de tal manera que cuando el disco no está girando lo presionan; pero cuando empieza a girar a grandes velocidades ocasiona una corriente de aire que levanta las cabezas y las hace "flotar" de tal modo que se encuentren a una distancia extremadamente pequeña del disco, pero nunca lo tocan.

Debido a esta pequeña distancia se debe cuidar la fabricación de los discos ya que cualquier partícula de polvo podría dañar el mecanismo o interferir con las operaciones de lectura y escritura sobre los platters.

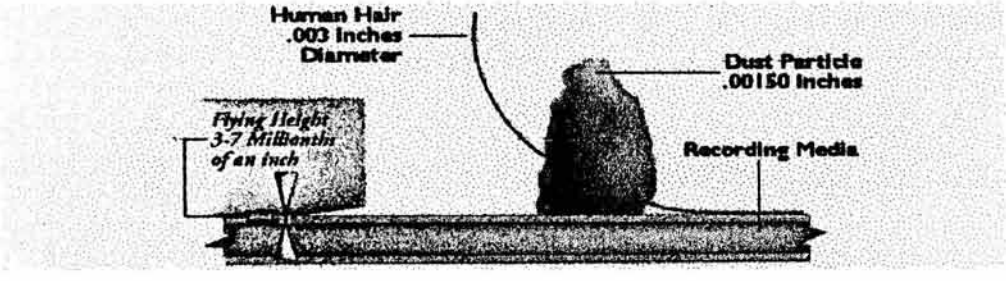


Figura 1.2.2 Comparación de dimensiones entre una partícula de polvo y la distancia entre las cabezas y los platters.

Algunas veces que los discos duros se encuentran aislados para evitar que el polvo interfiera en su funcionamiento; sin embargo esto no es del todo cierto ya

que para su propio funcionamiento, el disco necesita de aire, que al ser provocado por el girar del disco, haga que las cabezas se levanten.

La distancia entre las cabezas y el platter es calculada cuidando que no sea tan pequeña que permita choques entre estas y el disco, ni tan grande que las cabezas no logren leer o escribir adecuadamente por su sensibilidad. En un disco moderno esta distancia es de apenas 0.5 micro pulgadas.[8]

Con distancias tan pequeñas, la ingeniería debe tomar en cuenta aun mucho más factores que afectan el desempeño del disco, no solamente la distancia entre los platters y las cabezas, sino también la velocidad del disco, la sensibilidad de las cabezas, la textura de los platters y su resistencia, etc.

A lo largo del tiempo se han utilizado diferentes tipos de métodos y tecnología para la fabricación y funcionamiento de las cabezas lectoras. A continuación se menciona brevemente los tipos de cabezas lectoras/escritoras que se han ido utilizando desde los primeros discos duros hasta la actualidad.

- **Cabezas de ferrita**: Fueron las primeras utilizadas y las más simples. Una cabeza de ferrita estaba compuesta por un centro de hierro en forma de U al que se le enredaban alambres eléctricos, lo que formaba un clásico electro magneto, pero muy pequeño. Su funcionamiento se basaba en campos magnéticos; la dolarización de este campo magnético resultaba en la lectura o escritura de datos en el platter, la dolarización inversa también resultaba en la operación de lectura o escritura inversa.
- **Cabezas *Metal In Gap (MIG)* :** Su funcionamiento es similar al de las cabezas de ferrita, solamente que se anexo a este tipo una aleación metalica a la cabeza lo que mejoraba su capacidad de magnetizacion permitiendo ser utilizada con platters de mayor densidad.
- **Cabezas de película delgada *(Thin Film Head):*** Estas cabezas son totalmente diferentes a las de ferrita. Utilizan un proceso de fabricación

llamado fotolitográfico, similar al utilizado para fabricar procesadores. Se utiliza una aleación con la que se cubren las cabezas con un patrón especifico. El producto son cabezas pequeñas y precisas.

- **Cabezas magneto resistivas (MR/AMR):** Estas cabezas utilizan un material conductor especial que cambia su resistencia en la presencia de un campo magnético. Mientras la cabeza pasa sobre la superficie del disco, este material cambia su resistencia debido a los cambios en el campo magnético que corresponden a los patrones de datos almacenados en el disco. Una de las ventajas de este tipo de cabezas es que para la escritura se utiliza un dispositivo (cabeza) diferente que el utilizado para la lectura, permitiendo así eficientar por separado cada una de estas operaciones.
- **Cabezas *Giant Magnetoresistive (GMR):*** Estas cabezas funcionan bajo el efecto *magnetoresistivo gigante (giant magnetoresistive)*, utilizando un campo magnético grande y capas delgadas de material magnético se puede observar grandes cambios en la resistencia de estos materiales. Este tipo de cabezas están compuestas de 4 capas de materiales diferentes y su diseño favorece la sensibilidad de las cabezas y por tanto la posibilidad de utilizar discos de mayor densidad areal.

Sliders, brazos y actuadores.

Las cabezas lectoras/escritoras hoy en día son tan pequeñas que necesitan estar fijas a otro dispositivo. El *slider* es precisamente el dispositivo que sostiene a la cabeza en su posición correcta respecto al platter.

A su vez los brazos sostienen a los sliders dentro del disco. Estos están hechos de un material ligero y de forma triangular para poder "flotar" sobre la superficie del disco durante su operación.

El actuador es un dispositivo importante en la operación del disco duro. Su función es la de posicionar las cabezas lectoras/escritoras en la pista (track)

[8] Como comparación podemos mencionar que el cabello humano es de un grosor de 2000 micropulgadas.

adecuada en donde se escribirán o leerán los datos.

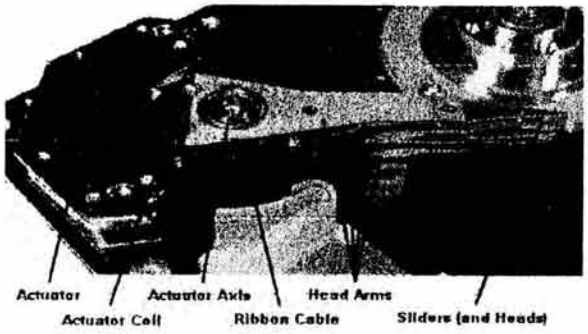


Figura 1.2.3. Un disco duro con sus cabezas y sliders.

Los primeros actuadores funcionaban con motores de pasos que posicionaban las cabezas en una pista especifica. La inconveniencia de este tipo de dispositivos era que los motores de pasos no son lo suficientemente sensibles a movimientos finos, así como también eran vulnerables a cambios en las pistas debido a la contracción o expansión de los discos por calor.

Posteriormente los actuadores que se utilizaron y que aun siguen vigentes son los que utilizan bobinas o dispositivos llamados *voice coil*, los cuales utilizan atracción electromagnética para posicionar las cabezas en la pista adecuada.

En la operación de los discos también interviene la técnica utilizada para encontrar una posición determinada en el disco. Esto se hace a partir de mecanismos llamados *servo sistemas* en donde un dispositivo es controlado a través de alguna acción, se mide el resultado obtenido, se hacen los ajustes necesarios y se repite la acción. Esto es un sistema con retroalimentación.

En el caso de los discos duros también se aplican técnicas parecidas para controlar el movimiento de los componentes involucrados en la lectura y escritura de información en los discos duros.

Conectores

Los discos duros son conectados a las computadoras o servidores a través de algún tipo de conector o interfaz. Los discos modernos utilizan alguna de las siguientes dos interfaces: IDE (ATA) y sus variantes, o SCSI.

Si inspeccionamos físicamente el equipo podemos distinguir fácilmente que tipo de interfaz utiliza:

- ***IDE / ATA:*** Se trata de un conector rectangular de 40 pines.
- ***SCSI :*** Un conector en forma trapezoidal de 50, 68 u 80 pines. En el caso de un conector de 50 pines se dice que es un *narrow scsi*; en caso de 68 pines se trata de un *wide scsi* y 80 pines hace referencia a *wide scsi* con conector simple (*SCA - Single Connector Attachment)*.

La interfaz IDE/ATA es utilizada principalmente en computadoras personales y máquinas pequeñas, SCSI es utilizado para la conexión de dispositivos de almacenamiento o discos duros a máquinas más robustas como servidores y equipos unix.

En lo consecuente nos enfocaremos a la interfaz SCSI, que es precisamente la mayormente utilizada en servidores unix.

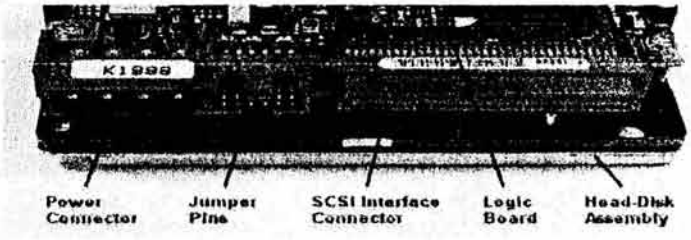


Figura 1.2.4 Interfaz SCSI de un disco duro.

Jumpers

Los dispositivos de disco también cuentan con lo que se conoce como *jumpers*, los cuales son unos pequeños orificios utilizados para darle cierta configuración al disco a través de la inserción de algún pin o aguja en ellos para hacer una especie de puente y lograr que el dispositivo tenga alguna configuración extra.

En el caso de los discos duros SCSI, estos cuentan con una mayor cantidad de jumpers que los de tipo IDE/ATA debido a que son más sofisticados y tienen más funciones para controlar su operación.

Estos jumpers pueden variar de un fabricante a otro, sin embargo a continuación se lista la función de algunos de jumpers más importantes y su función a grandes rasgos:

- *SCSI Device ID:* Todo dispositivo SCSI debe contar con identificador que permita saber al equipo hacia donde dirigir sus señales. Funciona como una dirección. Por lo regular los *narrow scsi* cuentan con 3 pines para establecer identificadores del 0 al 7. Los *wide scsi* cuentan con 4 pines para un direccionamiento de 0 a 15.
- *Termination Activate:* Dentro de los equipos existe un bus scsi al cual se conectan los dispositivos scsi. Los dispositivos conectados deben "terminar" el bus para su buen funcionamiento.
- *Disable Auto Start:* Si este jumper esta presente, el disco no girara automaticamente en cuanto se le provea energía, por el contrario esperara una orden de la maquina para empezar a girar.
- *Delay Auto Start:* Este jumper indica al disco girar automaticamente pero esperar un número determinado de segundos antes de hacer esta acción.
- *Stagger Spin:* Muy parecido al anterior. En este caso se implementa una

función en la cual se prevé la inicialización de dos discos al mismo tiempo a través de una operación matemática que controla el momento o tiempo en que cada disco iniciara sus operaciones.

- *Narrow/Wide:* Permite la configuración del disco ya sea como narrow scsi o como wide scsi.
- *Disable Parity*: Deshabilita el chequeo de paridad en el bus SCSI.

En un principio los discos duros funcionaban dependiendo de las instrucciones que la computadora le mandaba, pero en sí no eran capaces por ellos mismos de efectuar alguna operación automáticamente o por sí mismos; en cambio los discos duros actuales ya cuentan con circuitería que les permite controlar sus funciones, liberando a la computadora de todo el trabajo. [9]

A pesar de lo anterior, aún así contamos con circuitos o interfaces dentro de la computadora que permiten la comunicación entre el disco duro y la computadora. Estas interfaces van desde las más simples IDE/ATA hasta las más complejas como las del tipo SCSI.

También los discos duros actuales contienen no solamente circuitería que permite realizar ciertas funciones, sino también cuentan con procesadores y código residente en memoria ROM que determina sus funciones.

A este código en ROM (Read Only Memory), se le llama el *firmware* del disco y en el podemos encontrar información acerca de la inicialización del disco y su comportamiento. En la mayoría de los casos este *firmware* puede actualizarse a través de programas o datos provistos por sus fabricantes y que están disponibles en muchas ocasiones a través de Internet.

Una parte interesante e importante en el desempeño de los discos es la manera en que procesan una petición.

Se dice que la interfaz IDE/ATA, utilizada principalmente en las computadoras personales, solamente procesa una petición a la vez; en cambio la interfaz SCSI tiene la habilidad de manejar varias peticiones de lectura y escritura a la vez, aunque con un límite. A esta característica se le conoce como *command queuing and reordering* ó *multiple command queuing*, y es aprovechada principalmente por servidores y computadoras mayores que cuentan con varios usuarios accediendo a los datos al mismo tiempo.

Lo anterior significa que la lógica del disco recibirá varias peticiones al mismo tiempo y deberá procesarlas y encontrar dónde se encuentran los datos dentro del disco de alguna manera. Algunas de estas peticiones serán satisfechas por los datos que aún se encuentran en el caché del disco; sin embargo muchas de ellas habrá que obtenerlas del disco directamente.

Teniendo varias peticiones la interfaz o controlador del disco deberá decidir en qué orden procesará las peticiones de lectura y escritura. Recordando que el disco duro tiene funciones no solamente electrónicas, sino que también mecánicas, éste es un elemento dentro de la computadora de comportamiento sumamente lento comparado con los demás componentes electrónicos (memoria, procesador), por tanto la manera en que procese las peticiones va ligado directamente al desempeño del mismo.

Hay muchos algoritmos y métodos de establecer el orden en que serán atendidas las peticiones de lectura y escritura del disco, sin embargo la mayoría caen en las tres siguientes categorías, o bien son derivadas de las mismas:

- ***First In, First Out:*** Este es el mpetodo más simple en el que las peticiones son atendidas en el orden en que éstan llegaron. Si tomamos un elevador como caso práctico para ejemplificar esto, tenemos que, suponiendo que cuatro personas toman el mismo elevador y la primera en llegar va al piso 77, la segunda al 31, la tercera al 94 y la última al 20, tomando este método

[9] A esta circuitería se le llama *tarjeta lógica* del disco duro o *logic board.*

significaría que el elevador tendría que ir a cada uno de estos pisos en ese orden. Es clara la pérdida de tiempo que éste procedimiento involucraría y en términos de cómputo, el mal desempeño que resultaría de la aplicación de éste algoritmo.

- ***Seek Time Optimization:*** En este caso se analizan las peticiones y se ordenan dependiendo del número de cilindro en el que se encuentra el dato en relación con el número de cilindro en el que se encuentren las cabezas en ese momento. A éste método también se le conoce como *elevator seeking* (elevador), porque está diseñado para evitar los movimientos innecesarios de las cabezas sobre la superficie del disco, análogo a lo que hace un elevador para eficientar su funcionamiento.
- ***Access Time (Seek and Latency) Optimization:*** El problema con el algoritmo anterior, es que éste no toma en cuenta el movimiento rotacional o latencia rotacional del disco. En éste algoritmo (Access Time Optimization), se toma en cuenta no solamente el número de cilindro en donde se encuentra un dato y su relación con el número de cilindro en donde se encuentran las cabezas; sino también se trata de eficientar el orden en que son procesadas las peticiones tomando en cuenta también el tiempo que tardaría la cabeza lectora en llegar a ése dato debido a la rotación del disco (recordemos que las cabezas lectoras inciden sobre la superficie del disco que está rotando por debajo de ellas). A este algoritmo también se le conoce como *Optimized Reordering Command Algorithm (ORCA).* Como ejemplo tómese la figura siguiente en donde se compara el algoritmo anterior de elevador contra el algoritmo ORCA. El orden de proceso de las peticiones cambia debido a que el algoritmo de elevador no toma en cuenta el tiempo que necesita la cabeza lectora para llegar a un dato por la rotación del disco.

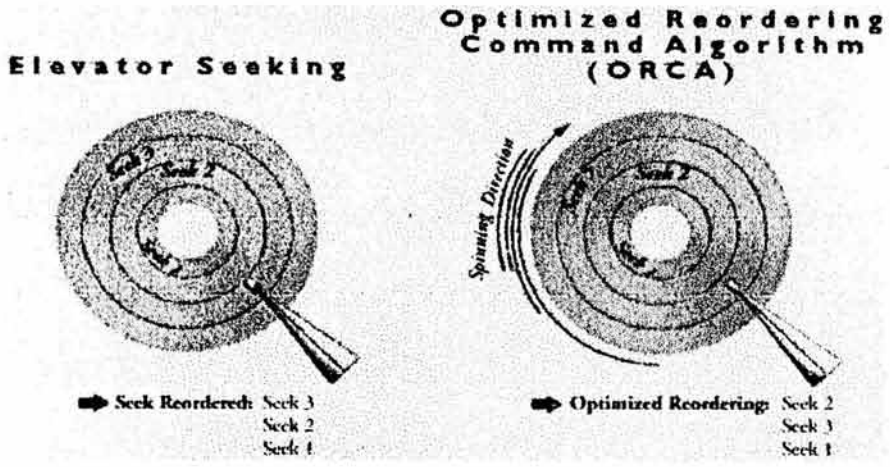


Figura 1.2.5 Algoritmos de búsqueda para lectura/escritura.

Pistas y Cilindros (Tracks y Cylinders)

Como sabemos, un disco está formado de varios platters, los cuales utilizan 2 cabezas para leer y escribir los datos, una cabeza situada por encima del plato y la otra por debajo; aunque en algunos casos esto no sea totalmente cierto debido a que ciertos fabricantes tienen algunos tipos de discos que solamente cuentan con sólo una de las cabezas.

Las cabezas que acceden los platters están unidas en un sólo ensamblaje de brazos, por lo que todas las cabezas dentro del disco se mueven en unísono hacia una dirección. Esta característica hace que todas las cabezas de un disco estén siempre posicionadas en el mismo número de pista (track).[10]

De lo anterior resulta que por lo regular se hace referencia a la posición de las cabezas por su número de cilindro y no por su número de pista. Un cilindro es básicamente el conjunto de pistas en las cuales todas las cabezas del disco están posicionadas en un momento específico. Por ejemplo, si un disco tiene cuatro platters, tendrá ocho cabezas, y un cilindro 720, por ejemplo, estaría compuesto por el conjunto de ocho pistas, una por cada superficie de los

[10] No es posible tener una cabeza situada en el *track* 0 y otra en el *track* 1000, por ejemplo.

platters, en la pista 720.

El nombre de cilindro viene del hecho de que si mentalmente visualizamos estas pistas, forman un esqueleto en forma de cilindro a lo largo de todos los platters del disco. Como referencia de esto observemos la figura siguiente en donde se hace referencia a la estructura del disco y a las pistas y cilindros.

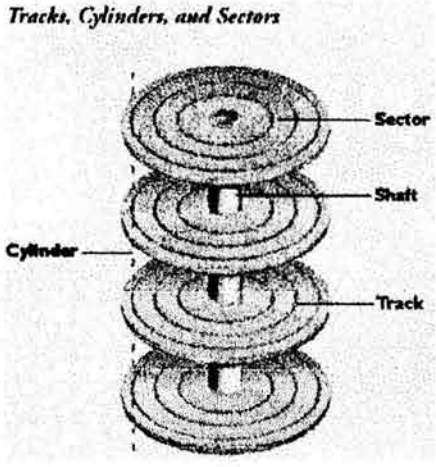


Figura 1.2.6 Pistas, cilindros y sectores de un disco.

Formato de discos

Darle formato a un disco significa hacerle los ajustes necesarios para poder alojar la información con una estructura definida (pistas y sectores), la cual pueda ser reconocida para su lectura y escritura.

Los primeros discos tenían definidos pistas y sectores de igual tamaño, de tal modo que las pistas más alejadas del centro del disco también contenían el mismo número de sectores que las pistas más cercanas. Este tipo de formato ocasionaba el desperdicio del disco puesto que las pistas alejadas con más extensas y pueden ser estructuradas con un mayor número de sectores.

Actualmente esta característica ha desaparecido, de tal modo que se ha dado una nueva estructura a los discos en la que las pistas más alejadas contengan más sectores que las pistas más cercanas.[11]
De cualquier modo, el proceso para darle formato a un disco y poder utilizarlo se hace en tres pasos:

1. Formato a bajo nivel.
2. Particionado.
3. Formato a alto nivel.

Contrario a lo que pudiera pensarse, el formato de un disco no es un proceso aislado, sino que comprende tres procesos importantes para su realización.

Formato a bajo nivel (Low-Level Formatting)

El formato a bajo nivel es el proceso que dicta la posición de las pistas y los sectores en el disco, y escribe las estructuras de control que definen dónde están dichas pistas y sectores. Este proceso realmente crea el formato físico que define dónde se almacenarán los datos en el disco.

Durante este proceso el disco es sometido a un formato físico por lo que si anteriormente contenía datos, estos se perderán y el disco quedara totalmente limpio con la estructura apropiada de pistas y sectores.

Con los primeros discos era posible hacer un formato a bajo nivel desde una computadora personal, por ejemplo, debido a que las pistas y sectores por pistas siempre eran los mismos; sin embargo, actualmente, esta operación es realizada de fábrica de tal modo que un usuario convencional no puede realizar un formato a bajo nivel desde su computadora, esto es por la complejidad de la estructura actual de los discos hablando en cuanto al número de sectores por pista y pistas con que se les da formato.

[11] Este tipo de estructura se le conoce como *Zone Bit Recording*

Particionado

Este es un proceso ejecutado a través del sistema operativo de la computadora que se esté utilizando. A través del particionado se define la estructura lógica del disco duro para ser utilizado a la conveniencia del usuario final. Se definen varias "piezas" o "pedazos" del disco que son vistas como volúmenes del disco.

Formato en alto nivel

Una vez que se ha hecho el formato a bajo nivel y que se han definido los pedazos lógicos que conformaran al disco, lo que resta por hacer para poder utilizarlo es un formato en alto nivel, o bien que el sistema operativo que se esté utilizando cree las estructuras de control o las estructuras del sistema de archivos (file system) para almacenar los programas y datos dentro del disco.
En el caso del sistema operativo DOS se utiliza el comando FORMAT para hacer este proceso; en el caso de UNIX Solaris se utiliza el comando newfs o mkfs.

Si quisiéramos limpiar totalmente el disco, no es necesario un formato a bajo nivel, el formato a alto nivel pude realizar este trabajo.

Geometría

Muchas ocasiones nos referimos a los discos duros indicando su geometría. Ésta es simplemente la descripción general de las características del disco en términos de como estructura sus datos en sus platters, pistas y sectores.

La geometría física de un disco hace referencia a el número de cabezas, cilindros y sectores utilizados por el disco.

En algunas ocasiones también se hace referencia a la geometría lógica del disco, la cual puede variar a la geometría física debido que en la segunda se toma no sólo en cuenta las características físicas, sino también algunas otras características como los sectores realmente utilizables, entre otras cosas.

1.3 INTERFACES DE LOS DISCOS DUROS

La interfaz con que el disco se conecta al resto de la computadora es un factor importante en cuanto al desempeño de los discos duros.

La interfaz es el canal de comunicación a través del cual todos los datos fluyen para ser leídos o escritos desde o hacia el disco duro. La interfaz puede ser uno de los factores que limiten el desempeño del disco de manera importante.

A lo largo de la historia de los discos duros se han utilizado diferentes tipos de interfaces las cuales han sido mejoradas por otras con nuevas características y mayor velocidad de transferencia de datos.

Algunas de estas interfaces son IDE/ATA, USB, PCI y SCSI. La primera y segunda son utilizadas principalmente en computadoras personales, las dos restantes pueden ser utilizadas tanto en computadoras personales como en servidores; especialmente SCSI ha sido la interfaz predominante en el campo de servidores, estaciones de trabajo y equipos grandes, por tanto se describirá más a fondo.

Cualquier interfaz que se trate, ésta se comunica con la computadora a través de lo que conocemos como bus de comunicaciones. A lo largo de este bus se transfieren datos entre todos los componentes de la computadora.
Las computadoras más recientes utilizan un bus PCI, en el caso de computadoras personales, en el caso de servidores y estaciones de trabajo,

éstas utilizan PCI o SBUS, entre otros[12].

También se ve que no solamente la interfaz es importante para el desempeño del disco duro, sino también lo es el bus al cual está conectada esa interfaz y la velocidad a que éste opera.

Entre el disco duro y la interfaz existe una serie de comandos que son utilizados para indicarle al disco duro que realice alguna lectura o escritura. La computadora, a través de la interfaz, manda comandos al disco, y el disco, a su vez, manda datos a través de ésta hacia la computadora.

Algunas interfaces son más simples que otras en su funcionamiento y en la cantidad de comandos que manejan, impactando de algún modo la respuesta hacia alguna acción de lectura o escritura de datos hacia el disco.

La cantidad de tiempo que la interfaz, su controlador, y el disco duro, requieren para procesar un comando es llamado *command overhead*.

La interfaz que se utiliza para conectar un disco con la computadora también puede soportar otro tipo de dispositivos que también son conectados al equipo. En el caso de la interfaz IDE/ATA, se dice que soporta diversos tipos de dispositivos, incluyendo varios modelos de discos duros e inclusive discos ópticos; sin embargo es utilizada principalmente por usuarios de computadoras personales, además de ser una opción más o menos económica comparada con SCSI.

En el caso de SCSI, es una interfaz más compleja y es utilizada principalmente en servidores, siendo una opción para aplicaciones comerciales o de negocio en donde se busque un mejor desempeño.

[12] Los primeros buses utilizados en computadoras personales eran ISA o VESA.

Otras Interfaces

A pesar de que las interfaces más utilizadas y difundidas son IDE/ATA y SCSI, también hay otras que son utilizadas principalmente para brindar cierta facilidad al conectar dispositivos a la computadora o servidor.

Entre estos se encuentran: *Parallel Port, PCMCIA, USB, FireWire, i.Link, Fibre Channel, USB2 y Serial ATA.*

Todas estas interfaces fueron diseñadas para brindar facilidades que las interfaces convencionales IDE/ATA o SCSI no tienen.

Estas interfaces son utilizadas en aplicaciones especificas, siendo Fibre Channel una de las empleadas para conectar dispositivos de fibra a servidores o computadoras grandes.

1.3.1 SCSI (SMALL COMPUTER SYSTEM INTERFACE)

En 1979 Shugart Associates crearon la *Shugart Associates Systems Interface (SASI)*, la que fuera predecesora de SCSI y que se trataba de una interfaz rudimentaria, con pocas capacidades, contando solamente unos cuantos comandos comparado con lo que ahora soporta SCSI.

Shugart Associates quisieron hacer de SASI un estándar de ANSI por lo que en 1981 se asocian con NCR Corporation para estandarizar la interfaz. En 1982 se empezó a trabajar en la estandarización de SASI hasta lograrlo y entre las cosas que se cambiaron fue precisamente el nombre, dejándolo como SCSI.

SCSI fue creado en un principio para discos duros, posteriormente se le

desarrollo para soportar toda una variedad de dispositivos, de hecho fue creado para ser una interfaz de alto nivel, expandible y de alto desempeño. Por esta razón, es la opción para usuarios de servidores o de computadoras de alto nivel.

SCSI comenzó como una interfaz paralela, permitiendo la conexión de varios dispositivos a un sistema, los cuales transmiten información a lo largo de múltiples líneas de datos.

SCSI-1

SCSI fue aprobado como un estándar en 1986 y posteriormente, cuando surgieron subsecuentes estándares de SCSI, se le denomino SCSI-1 por ser la especificación original.

SCSI-1 define las bases de los primeros buses SCSI, incluyendo longitud del cable, señalización, comandos y modos de transferencia.

SCSI-1 utiliza un bus de 8 bits, con una transferencia de 5 Mb/s. Sólo soportaba transmisión a un sólo punto, con terminación pasiva.

SCSI-2

SCSI-2 es una mejora muy amplia del SCSI original. El estándar define las siguientes características nuevas a la especificación original:

- ***Fast SCSI:*** La velocidad de transferencia se duplico a 10 Mhz.
- ***Wide SCSI:*** El ancho del bus SCSI fue incrementado a 16 o hasta 32 bits.
- ***Mas dispositivos por bus:*** para los buses con Wide SCSI, hasta 16 dispositivos son soportados, a diferencia de los 8 dispositivos que en SCSI-1 fueron soportados.
- ***Mejores cables y conectores:*** SCSI usa un extenso número de cables y conectores diferentes. SCSI-2 extiende los conectores básicos de 50 pines

definidos en SCSI-1.

- ***Señalización Diferencial:*** Para permitir el uso de cables de mayor longitud se introdujo la señalización diferencial.
- ***Comandos Adicionales:*** SCSI-2 permite el manejo de una mayor diversidad de dispositivos como CR-ROMs, scanners y otros dispositivos removibles. SCSI-1 se enfocaba solamente a discos duros.

SCSI-3

La nueva versión de SCSI, llamada SCSI-3, empezó a desarrollarse en 1993 cuando SCSI-2 tenía ya ocho años en la industria. Con SCSI-3 se consideraron una gran cantidad de tecnologías, comandos y características para formarlo.

Por esta gran cantidad de factores involucrados en el nuevo estándar SCSI-3, se decidió no realizar un sólo documento relacionado al estándar SCSI-3, sino hacer varios documentos con cada uno de los estándares relacionados con SCSI-3.

Entre los estándares, protocolos y comandos de SCSI-3 o que tienen que ver con éste, encontramos interfaces como IEEE-1394 y Fibre Channel. Así también encontramos la forma más implementada de SCSI-3, que fue conocida en sus comienzos solamente como SCSI, fue *SCSI-3 Parallel Interface (SPI)*. Sin embargo ,también se desarrollaron varias versiones de SPI, así por ejemplo, tenemos *SCSI-3 Parallel Interface-2*, lo cual ya resulta un tanto confuso.

Adaptadores SCSI

La mayoría de los discos IDE/ATA son controlados por controladores IDE ya integrados en la tarjeta madre de las computadoras. En el caso de SCSI esto por lo regular no se cumple, por lo que la mayoría de los sistemas requieren de la inserción de una tarjeta especial que sirva como la interfaz entre el bus SCSI y la computadora.

Este dispositivo es llamado *SCSI host adapter*, o en algunos casos *host bus adapter (HBA)*, o *SCSI controller* o simplemente *SCSI Card*. De cualquier modo su trabajo consiste en actuar como el *gateway (interfaz)* entre el bus SCSI y el bus interno de la computadora. Manda y responde comandos y transfiere datos hacia y desde los dispositivos en el bus y dentro de la computadora en sí.

Configuración de dispositivos SCSI

Topología de BUS SCSI.

La palabra topología hace referencia a la figura o estructura de las cosas. En cuestiones computacionales, ésta hace referencia a la manera en que los dispositivos están conectados o dispuestos. La interfaz SCSI utiliza una topología de bus. Esto significa que todos los dispositivos están conectados como una cadena, linealmente en una línea larga. Esto es análogo a la configuración de una red Ethernet en donde todos los dispositivos van conectados a la misma línea de comunicación.

Cuando se trata de un bus SCSI, es imperativo que cada dispositivo esté conectado en una misma línea. Esto significa que cada dispositivo (incluyendo el host adapter) está conectado a uno o máximo dos dispositivos, no más. Los dos dispositivos al final del bus deben ser terminados, o bien, deben terminar el bus, ya sea internamente o externamente. El bus nunca debe ser conectado en un loop, estrella u otro tipo de configuración.

Para un bus con dos dispositivos, por ejemplo, la topología luciría como sigue:

Terminador – Dispositivo A – DispositivoB – Terminador

Para un bus con cuatro dispositivos sería como sigue:

Terminador – DispositivoA – DispositivoB – DispositivoC – DispositivoD – Terminador

No importa que dispositivo está localizado en qué lugar de la cadena, y cualquiera de los dispositivos puede ser externo o interno, pero lo que sí es requerido son los terminadores a los extremos del bus. Los terminadores pueden ser dispositivos dedicados a esta función, o bien pueden estar integrados en los dispositivos que se encuentren a los extremos. Tomando el segundo ejemplo anterior, si el DispositivoD tuviera un terminador interno, podría ser habilitado para terminar el bus en lugar de utilizar un dispositivo por aparte como se ve en el ejemplo.

No importa en donde se localicen los dispositivos dentro del bus como ya dijimos, inclusive podemos situar al propio adaptador a la mitad del bus; sin embargo hay que deshabilitar su terminador, de lo contrario los dispositivos conectados a un lado del adaptador podrían quedar sin comunicación hacia el bus.

También es posible conectar dispositivos sin respetar un estricto orden, es decir, podemos dejar algunos conectores libres, sin ningún dispositivo conectado y aún así nuestra configuración funcionará.

Para ejemplificar lo anterior, tomemos la siguiente configuración legal:

Terminador—DispositivoA—(sin conectar)--DispositivoB--DispositivoC--(sin conectar)--Terminador

Sin embargo la siguiente configuración no funcionará para los siguientes dispositivos que se quisieran conectar en los conectores libres:

Terminador—DispositivoA—DispositvivoB—DispositivoC—Terminador—(sin conectar)--(sin conectar)

Número de dispositivos

Una de las características de SCSI que lo hacen ser una interfaz poderosa, es precisamente el número de dispositivos que pueden ser conectados al bus y el hecho de poder realizar diferentes "conversaciones" entre dispositivos simultáneamente.

Los buses SCSI pueden soportar hasta 16 dispositivos diferentes.

Sin embargo hay que tomar en cuenta algunos aspectos que tienen influencia en cuanto al número de dispositivos que pueden ser usados en un sólo bus SCSI:

- ***Ancho del bus:*** En sus orígenes SCSI fue definido como un bus de 8 bits, con soporte para 8 diferentes dispositivos. Cuando se creó wide SCSI de 16 bits, el número de dispositivos también se extendió a 16. Hay que tomar en cuenta que el Host Adapter es un dispositivo SCSI también, por lo que tenemos la posibilidad de conectar hasta 7 dispositivos diferentes en un narrow SCSI, o bien, 15 si se trata de wide SCSI.
- ***Modo de transferencia y método de señalización:*** Algunos modos de transferencia y métodos de señalización limitan el número máximo teórico de dispositivos a conectar debido a consideraciones en cuanto a la integridad de las señales eléctricas.
- ***Longitud del cable:*** Muchos modos de transferencia SCSI involucran el realizar un balance entre la longitud del cable y el número de dispositivos que pueden soportar. Para Ultra SCSI en particular, reduciendo el número de dispositivos en el bus, permite el uso de cables de mayor longitud y viceversa.
- ***Consideraciones prácticas:*** En teoría podríamos decir que seríamos capaces de usar hasta 15 discos duros en nuestro bus SCSI, sin embargo

ello requeriría de un sistema con espacio para éstos (si son internos) y una fuente de poder muy poderosa, valga la redundancia. En el caso de dispositivos externos, esto involucraría el uso de muchos cables y la necesidad de localizar estos dispositivos cerca uno de otro.

Identificadores de dispositivos SCSI

Cada dispositivo SCSI tiene una dirección en el bus SCSI la cual está determinada por un número específico. Para narrow SCSI, que soporta hasta 8 dispositivos, éstos están numerados del 0 al 7; para wide SCSI, soportando 16 dispositivos, la numeración corresponde del 0 al 15. La prioridad que un dispositivo tiene en el bus SCSI está basado en su identificador SCSI (SCSI ID). Para los primeros 8 ID, los números mayores tiene la más alta prioridad, por tanto el 7 tiene la más alta prioridad y el 0 la más baja. Para wide SCSI, los ID adicionales del 8 al 15 otra vez aumentan su prioridad en cuanto aumentan su numeración; sin embargo la secuencia completa tiene menor prioridad que la secuencia inicial del 0 al 7. Lo anterior se resume al siguiente orden en cuanto a prioridad se refiere : 7, 6, 5, 4, 3, 2, 1, 0, 15, 14, 13, 12, 11, 10, 9, 8.

La intención de la prioridad en los dispositivos es para usarse como guía en un proceso de "arbitraje". En general el proceso de arbitraje determina cual dispositivo puede tener el control del bus. Si más de un dispositivo quiere el control al mismo tiempo, el dispositivo con mayor prioridad lo obtendrá, en tanto que el dispositivo de menor prioridad tendrá que esperar. Una de las restricciones en cuanto a los identificadores, es que dos dispositivos no deben tener el mismo número o se creará confusión dentro del bus y no operarán adecuadamente. En configuraciones en donde se tenga un alto tráfico de datos sobre el bus, se asignan los ID de mayor prioridad a los dispositivos más lentos como scanners o dispositivos de cinta.

Una configuración común es asignar el identificador 7 , el más alto, al host adapter.

El método por el cual se asignan los identificadores depende de cada dispositivo. Hay algunos que cuentan con pines o jumpers, otros tiene switches y algunos más lo hacen a través de software.

En el caso de lo jumpers es necesario revisar detenidamente la configuración de éstos ya que por lo general se habilitan varios pines para asignar un identificador en específico.

1.3.2 FIBRE CHANNEL

Este tipo de interfaz es una alternativa de las interfaces de alto desempeño como lo es SCSI. La interfaz es llamada *Fibre Channel (Fibro Canal)*, y su nombre proviene del hecho que originalmente fue diseñada para operar sobre canales físicos de fibra óptica; posteriormente el cobre fue soportado sobre esta interfaz.

Inclusive *Fibre Channel* está definido como parte de la familia de estándares SCSI-3. así como SCSI, Fibre Channel es una colección de protocolos y opciones, actualmente lo que se implementa es lo que se llama *Fibre Channel Arbitrated Loop (FC-AL)* que significa que varios servidores y dispositivos de almacenamiento sean conectados a lo que sería una red de almacenamiento, la que se discutirá en el último capítulo de este trabajo.

Las características generales de FC-AL son su velocidad y su ancho de banda de hasta 4Gbits/s, con la posibilidad de crecer en un futuro cercano. Uno de los principales beneficios de utilizar Fibre Channel, es la posibilidad de conectar dispositivos hasta 10 kilómetros de distancia uno de otro.

A pesar de todas las bondades que nos otorga el uso de Fibre Channel, es una tecnología utilizada primordialmente para aplicaciones grandes y servidores, no para usuarios finales ni computadoras personales debido a su gran costo;

además de que si se utilizara en aplicaciones o computadoras pequeñas se vería desperdiciado su poder.

Algunas de las características de Fibre Channel son :

- Desempeño desde 266 megabits/segundo hasta 4 gigabits/segundo
- Soporta distancias de hasta 10 km.
- Conectores pequeños
- Utilización de un gran ancho de banda
- Mayor conectividad
- Gran disponibilidad
- Soporte a múltiples niveles, desde pequeños sistemas hasta supercomputadoras
- Habilidad de manejar múltiples comandos, incluyendo Internet Protocol (IP), SCSI, IPI, HIPPI-FP, y audio / video.

Todo lo que un puerto de Fibre Channel tiene que realizar es manejar una conexión simple punto a punto.

Fibre Channel no solamente es utilizado para almacenamiento, también es utilizado en redes de computadoras y en redes de almacenamiento. Fibre Channel es ideal para las siguientes aplicaciones:

- Gran desempeño en los sistemas de almacenamiento
- Bases de datos extensas
- Respaldos de sistemas y recuperación de los mismos
- Clusters
- Almacenamiento a través de redes
- Alto desempeño en trabajos en grupo
- Backbone de redes
- Redes de audio / video digital.

En el caso de querer conectar algún dispositivo de almacenamiento a un sistema o servidor, se debe de contar con un adaptador e instalarlo en el sistema, igual a como se haría con un adaptador SCSI.

Algunos de los productos o dispositivos que manejan Fibre Channel son:

- Cables de cobre: Cuatro tipos de cables de cobre son definidos por el estándar de Fibre Channel.
- Cajas de discos: Las cajas de discos que usan Fibre Channel contienen un back plane construido son un loop de Fibre Channel, lo que permite el retirado o instalado de discos en caliente.
- Drivers: Los drivers de Fiber Channel soportan múltiples protocolos, típicamente SCSI e IP. Los sistemas operativos soportados más populares son Windows NT, AIX, Solaris, IRIX y HPUX.
- Extenders: Son usados para proveer distancias mayores de conexión.
- Discos de Fibre Channel: Manejan mejores capacidades de transferencia y mayor cantidad de almacenamiento. Usan comandos SCSI.
- Conectores de Fibra óptica: Se trata de un conector que puede ser conectado en caliente.
- Adaptadores (Host Bus Adapters): Son similares a los adaptadores SCSI y Tarjetas de Red. Estos adaptadores están disponibles tanto para fibra óptica como para cobre. Hay adaptadores disponibles para Sbus, PCI, MCA, EISA, GIO, HIO, PMC y Compact PCI.

También hay dispositivos utilizados para redes de almacenamiento (SAN) como hubs, analizadores, ruteadores, switches, bridges, gateways, etc.

1.4 TECNICAS RAID.

En algunos casos el uso de un sólo disco o de varios discos aislados limita la capacidad de almacenamiento o el desempeño de los servidores, y por ende, de

los servicios provistos.

En el caso de los servidores unix, estos pueden contar con muchos usuarios haciendo uso de una parte del disco, de tal modo que podrá haber un momento en el cual ya no sea suficiente el espacio con el que se cuenta y se tenga que involucrar la adquisición de otro disco y así sucesivamente conforme se vayan agotando los recursos de almacenamiento disponibles. Se podrían ir conectando discos y discos para satisfacer estas necesidades.

Como se mencionó en el capitulo anterior, el hardware nos impone límites en cuanto a los dispositivos que podemos conectar al sistema, por lo que habrá un momento en que la posibilidad de ampliar nuestros recursos de almacenamiento pudieran alcanzar este límite y vernos imposibilitados de hacer un crecimiento de nuestros recursos de almacenamiento.

Otra de las desventajas de ir adicionando discos para ampliar nuestros recursos, es el hecho de que el conectar un nuevo disco al sistema y dejarlo listo para utilizarse involucra el dar de baja los servicios con los que cuentan nuestros servidores. En ese caso, la mayoría de las organizaciones que proveen servicios cruciales para la empresa, no pueden darse el lujo de tener tiempos fuera de servicio ya que esto repercute directamente en las ganancias monetarias, prestigio o imagen que se desea tener.

Debido a las desventajas que este esquema representa, en Berkeley, se desarrollaron unas técnicas llamadas *Redundant Arrays of Inexpensive Disks* o más comúnmente llamadas RAID o niveles de RAID.

Estas técnicas RAID son aplicadas a varios discos físicos, muchas veces conectados a un mismo dispositivo llamado arreglo o array, para proveer ya sea de mayor capacidad o seguridad o desempeño en comparación con la utilización de un sólo disco o de varios discos por separado.

Los principales beneficios que se obtienen con esta técnica son:

- **Mayor seguridad en los datos.** A través de lo que llamamos *redundancia*, la mayoría de los niveles de RAID proveen la protección de los datos almacenados en el arreglo. Esto significa que una falla en un disco no involucra la pérdida de la información, y por tanto tampoco la recuperación de la misma a través de un respaldo.
- **Mayor tolerancia a fallas**. Ligado con el punto anterior, una falla en algún disco la mayor de las veces no significará una pérdida de información. Podemos decir que algunos tipos de configuraciones son más tolerantes que otras.
- **Mayor disponibilidad.** Las características anteriores también ocasionan que la recuperación a una falla sea más rápida e inclusive en algunos casos, sin necesidad de detener el servicio.
- **Mayor capacidad.** Otra de las ventajas de utilizar las técnicas RAID es la posibilidad de agrupar varios discos pequeños en un arreglo o disco lógico que funcione como un disco de mucha mayor capacidad (suma de las capacidad de cada disco pequeño), de tal modo que si se necesitan, por ejemplo, 300GB de almacenamiento en disco, podemos agrupar varios discos de 72GB para satisfacer esa necesidad, de lo contrario sería prácticamente imposible contar con un disco de dicha capacidad total.
- **Mejor desempeño.** Con las técnicas RAID también puede mejorarse el desempeño del sistema en cuanto a las lecturas y/o escrituras dependiendo del nivel que se implemente. Lo anterior se debe a la posibilidad de escribir simultáneamente en varios discos al mismo tiempo.

Existen varias técnicas o niveles de RAID, los cuales tienen o carecen de ciertas características, por lo que siempre es necesario tomar en cuenta todo el entorno del sistema como lo son tipo de aplicaciones que se corren, tipo de accesos que se realizan a la información, hardware con que se cuenta y sobre todo, el costo. Una configuración que tenga la mayor cantidad de beneficios será la configuración más costosa también.

En muchos casos lo que habrá que hacer para saber cual es la técnica de RAID más conveniente a implementar, será haciendo un balance entre desempeño, redundancia y menor costo. No siempre una técnica de RAID será la más conveniente para todos los sistemas, ni tampoco hay una técnica que abarque por completo estas tres características, por lo regular la de mejor desempeño no siempre es la que tiene mejor redundancia y viceversa.

En el siguiente capítulo, se revisarán cada una de estas técnicas RAID, sus ventajas y desventajas, ya que como se comenta, no existe una técnica buena y otra mala, todo depende de las características que deseamos tener en nuestros sistemas para que una técnica sea más conveniente que otra.

Capítulo 2. NIVELES DE RAID.

2.1 CONCEPTOS GENERALES DE RAID

2.1.1 Discos físicos y lógicos.

Un *disco físico* es el disco duro real que se utiliza junto con otros para formar algún nivel de RAID. Estos discos por lo regular forman parte de un arreglo de discos físicos, de tal modo que en lugar de tener varios discos individuales conectados cada uno a una interfaz o conector diferente cada uno hacia el servidor, se cuenta con una *caja* de discos que puede contener en sus *bahías o slots* varios discos duros. La *caja* o arreglo de discos, como se le conoce, es conectada al sistema a través de una sola interfaz. Hay también equipos que utilizan más de una interfaz para conectarse, pero este diseño más bien responde a la necesidad de tener conexiones redundantes en caso de una falla en las interfaces.

Un *disco lógico* es la agrupación y estructuración de varios discos físicos que se ven y utilizan como si fueran un sólo disco pero de mayor capacidad y mejor desempeño. No se trata de un equipo de hardware que podemos conectar y desconectar, o que podemos palpar directamente. El sistema simplemente lo ve como un disco normal y es utilizado como cualquier otro disco físico. Algunos programas de administración de discos le llaman *volumen* o *metadispositivo*[1] por tratarse de una estructura lógica realizada a través de software especial para ello, o bien a través de equipos muy especializados.

[1] El término metadispositivo es utilizado por la herramienta de administración de discos Legato Disk Suite, que Sun Microsystems distribuye junto con su sistema operativo Solaris.

2.1.2 Mirroring (Espejos)

Una de las configuraciones posibles con las ténicas RAID, es precisamente la de espejo, en donde se tienen n copias de la misma información, por lo general en diferentes discos físicos. Todo esto con el fin de proveer una redundancia total de la información. En caso de que una de las copias de la información llegara a dañarse o perderse completamente, las demás copias siguen funcionando.

Este tipo de configuración nos da la mayor redundancia posible, sin embargo una de sus desventajas es la lentitud con que son realizadas las escrituras ya que se deben de realizar varias veces dependiendo del número de copias (espejos) que se tengan. También el costo es una desventaja; se debe de contar mínimamente con el doble de capacidad para realizar esta configuración.

2.1.3 Duplexing

Duplexing es una extensión de *mirroring* ya que a través de esta configuración no solamente se cuentan con varias copias de la información formando un espejo, sino que también se realiza redundancia de conexiones utilizando varias controladoras, de tal modo que la falla en una controladora tampoco signifique la pérdida de la información.

Duplexing es superior que *mirroring* en cuestiones de disponibilidad de datos, pero también lo es en costo.[2]

2.1.4 Striping

Una de las configuraciones más comúnmente utilizadas es precisamente el *striping,* en donde, basados en la manera en que están distribuidos los discos

[2] Ambas configuraciones, Duplexing y Mirroring son complemetarias y no contrarias como algunos autores sugieren.

dentro de un arreglo físico de discos o caja de discos, se toman pequeñas partes de cada uno de éstos para realizar las escrituras sobre todos al mismo tiempo o en paralelo, incrementando así el desempeño en la escritura de datos debido a que todos los discos son impactados al mismo tiempo.

El striping puede realizarse por bytes de datos o por sectores de 512 bytes, dependiendo del nivel de RAID que se configure será el tipo de striping que se realice.

Striping es utilizado en varias configuraciones de RAID, que dependiendo de éstas, pueden agregar, además del acceso en paralelo a los discos, la posibilidad de implementar cierta redundancia. Esto se verá más claramente cuando se discuta cada uno de los niveles de RAID.

2.1.5 Paridad

La paridad es otra función que es integrada en muchos de los niveles de RAID para proveer redundancia a los datos sin la necesidad de "desperdiciar"[3] la mitad de nuestros recursos de almacenamiento totales implementando *mirroring*.

La manera en como se realiza la paridad es a través de una operación lógica sobre los datos existentes y guardando el resultado en otro disco dedicado enteramente a esta función, o bien, distribuyendo este resultado a lo largo de todos los discos que forman el disco lógico.

Con esto lo único que se desperdicia es 100/N de espacio dedicado a almacenar paridad, en donde N es igual al número total de discos físicos que conforman el disco lógico o volumen. Si tenemos, por ejemplo, un disco lógico formado por 4 discos físicos, entonces el espacio necesario para almacenar la paridad sería de 100/4 , o bien , del 25% (1 disco) del total de los discos utilizados. En una

configuración de *mirroring* se necesitarían de otros 4 discos adicionales para proveer redundancia.

La paridad es calculada a través de la función lógica XOR, la cual resulta con un valor positivo si uno y sólo uno de sus operandos tiene un valor positivo, en cualquier otro caso será negativo su valor. La siguiente tabla muestra los valores obtenidos a través del operando XOR.

A	B	XOR
0	0	0
0	1	1
1	0	1
1	1	0

Tabla 2.1.1 Operando XOR

Una característica importante de la función XOR, es la posibilidad de obtener un dato faltante a través de la paridad y un dato ya existente aplicando la misma función, es por esto que esta función es utilizada por los niveles de RAID para brindar redundancia.

Por ejemplo, si calculamos A XOR B obtenemos la paridad P, y si calculamos la misma B XOR P, entonces obtendremos el dato que nos falta, o sea, A. Lo mismo pasará si tomamos ahora al dato A XOR P para obtener el dato B.

Lo anterior significa que en el caso de una falla en alguno de los discos que conforman nuestro disco lógico, los datos contenidos en éste podrán ser recuperados a partir de la paridad previamente calculada, y los datos sobrevivientes.[4]

---

[3] Algunas de las configuraciones de RAID requieren un espacio del volumen para datos de control.

[4] Si bien esta técnica nos provee los datos perdidos por la falla, las lecturas y escrituras son evidentemente lentas por el cálculo de la paridad.

La ventaja de la paridad es el hecho de tener redundancia sin sacrificar tantos recursos de almacenamiento, sin embargo su desventaja más grande es el tiempo que involucra el realizar las operaciones necesarias para calcular la paridad, y sobre todo, en caso de la pérdida de un disco, el tiempo necesario para recuperar los datos perdidos a través de los datos aún existentes y la paridad.

2.2 NIVELES DE RAID.

Existen varias maneras de implementar RAID en nuestros sistemas, ya sea a través de lo que se conoce como *striping* o *mirroring* o *concatenation*, etc., de cualquier forma todos estos métodos fueron definidos en 1988 en Berkeley y desde entonces los investigadores los bautizaron con el nombre de *niveles de RAID*, aunque para ser precisos éstos no simbolizan niveles en ningún sentido, o sea que no quiere decir que el nivel 1 sea precisamente el sucesor del nivel 0.

Podríamos imaginarnos que la palabra *nivel* hace referencia a alguna versión, o que los *niveles de RAID* están construídos unos sobre otros y que el nivel N+1 es mejor, más completo o tiene alguna particularidad adicional con respecto al nivel N, lo cual no es completamente cierto.

En sí los niveles de RAID no son mejores unos que otros, todo depende de la utilización que se le quiera dar y el sistema del cual estén formando parte; por tanto un nivel de RAID puede o no ser tan conveniente para una persona u otra dependiendo de las necesidades específicas de cada sistema.

Todos los niveles tienen ventajas y desventajas; sin embargo alguno de ellos podrá ser utilizado con éxito en algún sistema dependiendo de las aplicaciones que éste corra.

A continuación se describen los principales niveles de RAID existentes en el mercado.

2.2.1 NIVELES SIMPLES.

RAID 0 - Concatenation

Este nivel de RAID es el más simple, fácil y menos costoso de implantar. Consiste en la asociación, a través de software o hardware, de varios discos físicos en uno virtual de igual tamaño a la suma de la capacidad de cada uno de los discos físicos que lo conforman. Así, por ejemplo, si tomamos un disco de 4 GB, otro de 6GB y otro de 2GB para formar un disco virtual a través de concatenación, éste último tendrá una capacidad total de 12GB.

Las escrituras se realizan de manera secuencial, es decir, el primer disco físico es llenado con datos antes de pasar a escribir en el segundo disco, y así sucesivamente.
La siguiente figura ejemplifica el nivel RAID 0 en concatenación.

RAID 0 - CONCATENATION

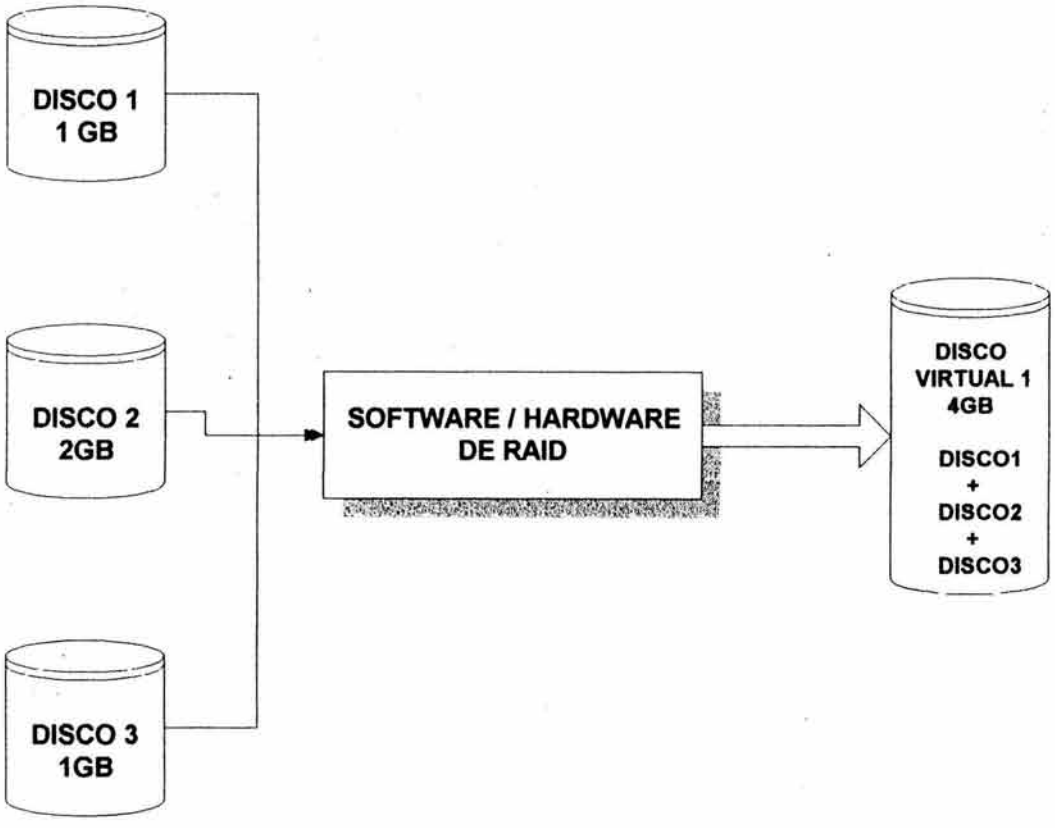


Figura 2.2.1 RAID 0- Concatenation

Ventajas:

- Permite la creación de un disco virtual (volumen o metadispositivo) de mayor capacidad que un disco físico
- Permite la utilización del espacio total del disco virtual, no hay desperdicio por manejo interno del software o hardware utilizado para realizar el arreglo o configuración.
- No requiere un mínimo de discos a utilizar, se pueden ocupar desde uno hasta n discos físicos.
- Al tener disponible para datos de usuario el 100% del espacio total del disco virtual, es más barato que cualquier otra configuración

Desventajas:

- La principal es la ausencia de cualquier tipo de redundancia en los datos, la pérdida de un disco físico por alguna falla significa la pérdida total de los datos
- Este tipo de configuración no mejora el desempeño en cuanto a la lectura ni a la escritura de datos, el desempeño es el mismo al que se tendría utilizando un sólo disco aislado (sin configuración RAID)

RAID 0 - Striping

Este es el segundo caso de RAID 0. En esta configuración se toman pequeños "pedazos" de igual tamaño de varios discos físicos que formarán un arreglo o disco virtual. Se requiere un mínimo de 2 discos físicos para poder crear una configuración de este tipo.

Los pedazos tomados de cada disco son llamados *chunks*, *interlace value ó stripes*[5], dependiendo del software o hardware que se esté utilizando para formar este tipo de RAID. Las escrituras se realizan simultáneamente en varios de los discos que conforman este arreglo, y siendo los *chunks* de igual tamaño, la cantidad de bytes escritos en cada disco es la misma en un tiempo.

El siguiente figura muestra este tipo de RAID.

[5] El término cambia dependiendo del software que se utilice, pero el concepto es el mismo.

RAID 0 - STRIPING

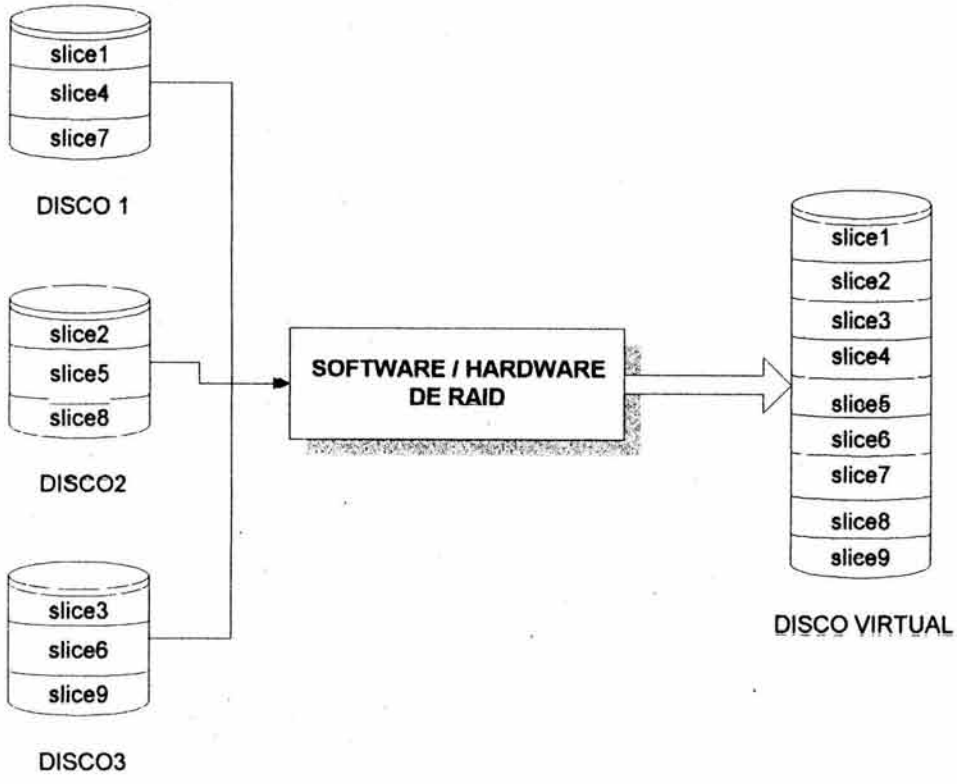


Figura 2.2.2 RAID 0 - Striping

Como se ve en la figura, el disco virtual está compuesto de varios pedazos tomados de cada uno de los discos físicos que participan en la configuración. Las escrituras se realizarían en el primer slice antes de pasar al segundo y así sucesivamente. En un momento podría existir una escritura de datos lo suficientemente grande como para impactar al mismo tiempo el slice1, slice2 y slice3 de tal modo que las escrituras se estarían realizando en 3 discos físicos al mismo tiempo.

Ventajas:

- Permite la formación de un disco virtual con mejor desempeño (eventualmente, dependiendo del tamaño de los datos a escribir) en las

escrituras. Puede ser inclusive de mayor dimensión a los discos físicos que lo conforman.

- Por su naturaleza, permite distribuir las cargas de trabajo entre varios discos.
- Al realizar las escrituras simultáneamente en varios discos, su desempeño en esta operación se ve favorecido. En el caso de las lecturas, el desempeño es el mismo.
- El espacio configurado para este arreglo está disponible en un 100% a los datos del usuario, es decir, no existe en este nivel, el uso de información de control almacenada por el hardware o el software en el disco virtual.
- Permite la utilización balanceada de todos los discos que conforman el arreglo.

Desventajas:

- Al igual que con *concatenation*, el *striping* no cuenta con redundancia, esto significa que la falla en alguno de los discos físicos que conforman el arreglo involucra la pérdida de toda la información ahí contenida.
- Para poder realizar striping es necesario contar con al menos 2 discos físicos, de lo contrario solamente podríamos realizar *concatenation.*

RAID 1 - Mirroring

Este nivel de RAID implementa redundancia de los datos contenidos en un disco virtual haciendo una copia idéntica de los datos en cada uno de los *submirrors*[6] que conforman el arreglo o *mirror.*

[6] Un submirror es un disco virtual en sí que después es relacionado con otro para formar lo que se conoce como mirror.

En esta configuración se crea un disco virtual llamado *mirror*, conformado por dos o más copias idénticas de la información las cuales son llamadas *submirrors*. Las escrituras son realizadas en todos los *submirrors*, y en cuanto un nuevo *submirror* es anexado a la configuración, se le hace un proceso de copiado de los datos, el cual es llamado *sincronización*.

Cada submirror puede ser formado por un disco virtual también, ya sea en concatenation o en striping, y asociado con otro disco virtual de igual tamaño para formar el mirror[7].

Las lecturas pueden ser realizadas desde un sólo *submirror* o desde varios dependiendo de la política de lectura que sea designada para ese disco virtual. Las políticas más comúnmente utilizadas son:

- *Submirror preferido*. Las lecturas son realizadas desde un sólo *submirror* del *mirror*.
- *Round Robin*. Las lecturas son realizadas en todos los submirrors, tomando cada uno de estos su lugar para la operación de acuerdo a un algoritmo definido por el hardware o software utilizado para el manejo del arreglo.

La siguiente figura ejemplifica este tipo de RAID.

[7] Un RAID 0+1 ó 1+0 involucra tener ambos submirrors en striping.

RAID 1 - MIRRORING

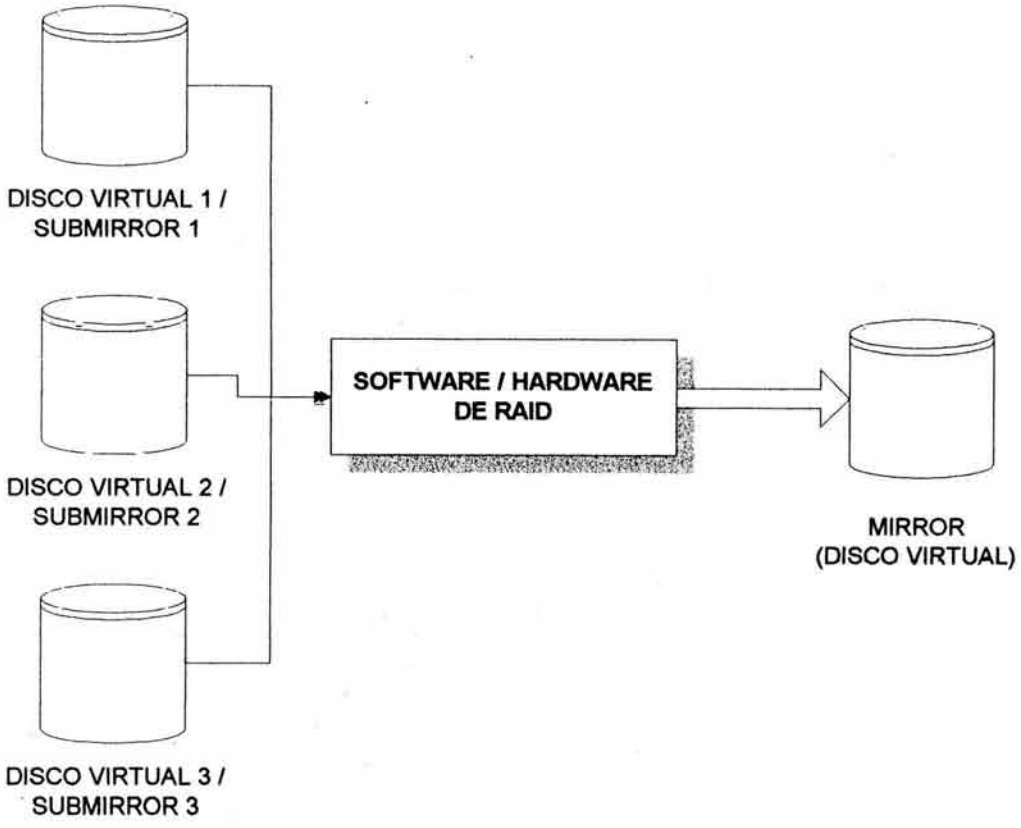


Figura 2.2.3 RAID 1 - MIrroring

En esta figura el disco virtual1, disco virtual2 y disco virtual3 contienen la misma información y son llamados cada uno "submirrors". Juntos forman el mirror en sí, que es precisamente lo que un administrador manejaría directamente y desde donde extraería la información. El software o hardware que hace esta configuración es el encargado de realizar las operaciones en todos los submirror para actualizar al información.

Ventajas:

- La principal es la implementación de redundancia al tratarse de un arreglo 100% redundante

- La pérdida de cualquiera de los disco físicos no provoca la pérdida de la información. Varias copias (submirrors) están disponibles.
- El desempeño en las lecturas puede ser incrementado si se tiene una política de *Round Robin*.
- En algunas ocasiones los submirrors pueden ser utilizados para hacer respaldos de información.
- La recuperación a desastres es relativamente rápida y muy sencilla.

Desventajas:

- Al tratarse de un arreglo 100% redundante se necesita mínimamente el doble de espacio o discos físicos para mantener al menos dos copias o submirrors de la información.
- Lo anterior lo hace ser una solución altamente costosa.
- Al realizarse las escrituras sobre todos los submirrors que conforman el arreglo, la operación necesita más tiempo para ser realizada. En un sistema que realiza muchas escrituras de datos, éste tipo de configuración no es la adecuada.
- El desempeño en las escrituras es altamente deteriorado. Con un mirror de 2 submirrors, el desempeño es decrementado hasta en un 40%.[8]

RAID 2 - Bit-level striping with Hamming code ECC

Este nivel nunca fue implementado comercialmente debido al costo que representaba.

El RAID 2 no implementa un esquema de redundancia de datos como lo hacen los demás niveles, ya sea con paridad o con espejos, éste a diferencia de aquellos lo

[8] Una escritura involucra escribir en todos los submirror de la configuración.

hace repartiendo los datos al nivel de bits en varios discos de datos y varios discos de redundancia. Los bits de redundancia son calculados a través de un código parecido al ECC (Error Correcting Code). Cada ves que un dato es escrito, estos códigos son calculados y escritos también junto con los datos. Cuando los datos son leídos, también lo son lo códigos de tal modo que se garantiza que los datos no hayan cambiado desde que fueron originalmente escritos por primera vez. Si un error de un bit ocurre, los datos son corregidos al momento.

Este nivel nunca fue implementado debido a su complejidad y costo ya que se necesitaría una gran cantidad de discos para poder realizarlo.

RAID 3- Striping with dedicated parity.

Varios niveles de RAID implementan alguna forma de redundancia. En el caso de RAID 3, y otros, lo hacen a través del cálculo de paridad entre datos realizando la operación lógica XOR como se mencionó anteriormente.

Con RAID 3 se dice que la paridad es dedicada porque se asigna todo un disco físico del arreglo para contener la paridad, resultado de la operación XOR de varios datos. Este disco por lo tanto no podrá estar disponible para contener datos del usuario .

La paridad en estos niveles de RAID sirve como redundancia, de tal modo que al perder alguno de los discos físicos que forman el arreglo, los datos aún pueden ser recuperados utilizando esta paridad y los datos de los discos sobrevivientes.

El RAID 3 no es una implementación comercial por lo que no es soportado por muchos productos, sobre todo si se trata de realizarlo a través de software.

La siguiente figura ejemplifica este nivel de RAID

RAID 3 - MIRRORING WITH DEDICADED PARITY

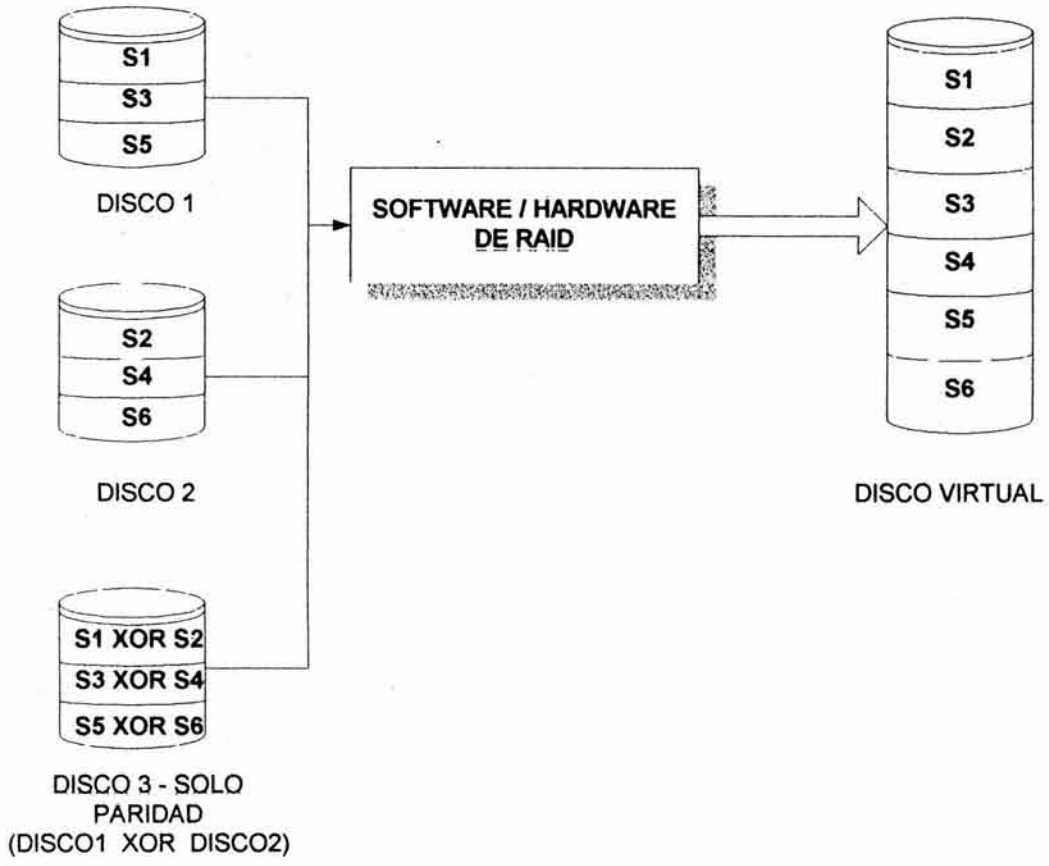


Figura 2.2.4 RAID 3

En esta figura, el DISCO 3 sólo almacena la paridad, o resultado de la operación XOR entre 2 datos, de la información del arreglo.

Ventajas:

- Se tiene redundancia en los datos a través de la paridad.

- Nos protege de la falla de un disco (sólo uno).
- Es más barato que configurar Mirroring

Desventajas:

- Una falla en dos o más discos significa la pérdida de toda la información
- El desempeño en las escrituras es afectado por el tiempo que le lleva al sistema calcular la paridad si se trata de RAID por software.
- En caso de la pérdida de un disco, los datos son leídos a través de la paridad, lo que hace el proceso sumamente lento.
- No es recomendable para sistemas cuyas operaciones de I/O son en un 25% o más, operaciones de escritura.
- Un disco está completamente dedicado a paridad por lo que no es posible utilizarlo para alojar datos del usuario. Existe un desperdicio.
- El disco de paridad constituye un cuello de botella en el desempeño general del disco virtual.
- Se requiere un mínimo de 3 discos físicos para realizar esta configuración.

RAID 4 - Block - level striping with dedicated parity

Este nivel incrementa el desempeño como lo hacen el RAID 3 y el RAID 5 por tratarse todos ellos de una forma de striping; sin embargo, éste difiere del RAID 3 en que hace el striping por bloques y no por bytes, y del RAID 5 en que la paridad la guarda en solamente un disco.

A pesar de que el desempeño es mejorado, el hecho de guardar la paridad en un sólo disco contribuye a formar un cuello de botella en este factor, este disco invariablemente va a ser escrito cada vez que se realice una escritura de datos.

Este nivel no es muy utilizado ya que es preferido el RAID 3 o el 5 en su lugar, por tanto la mayoría de los proveedores de software para configuraciones RAID, no soportan este tipo de configuración.

Las ventajas y desventajas son las mismas que las de RAID 3.

RAID 5 - Striping with distributed parity

Este nivel también implementa el mismo mecanismo de paridad como redundancia, pero la diferencia primordial con RAID 3 es la distribución de ésta a lo largo de cada uno de los discos que conforman el arreglo.

Así por ejemplo, si un disco virtual en RAID 5 está conformado por un disco1, disco2 y disco3, los datos se estarán escribiendo en stripes (chunks) tomados de cada disco, primero se escribirá en el stripe del disco1, luego en el del disco2 y finalmente se calculará una XOR entre los datos de disco1 y disco2 para guardar el resultado (paridad) en el disco3. Posteriormente se continuará escribiendo en el disco1, luego el disco3 y la paridad en el disco2. Después corresponderá escribir en el disco2 y disco3 y guardar paridad en el disco1. Y así sucesivamente.

La siguiente figura muestra este tipo de configuración.

RAID 5 - MIRRORING WITH DISTRIBUTED PARITY

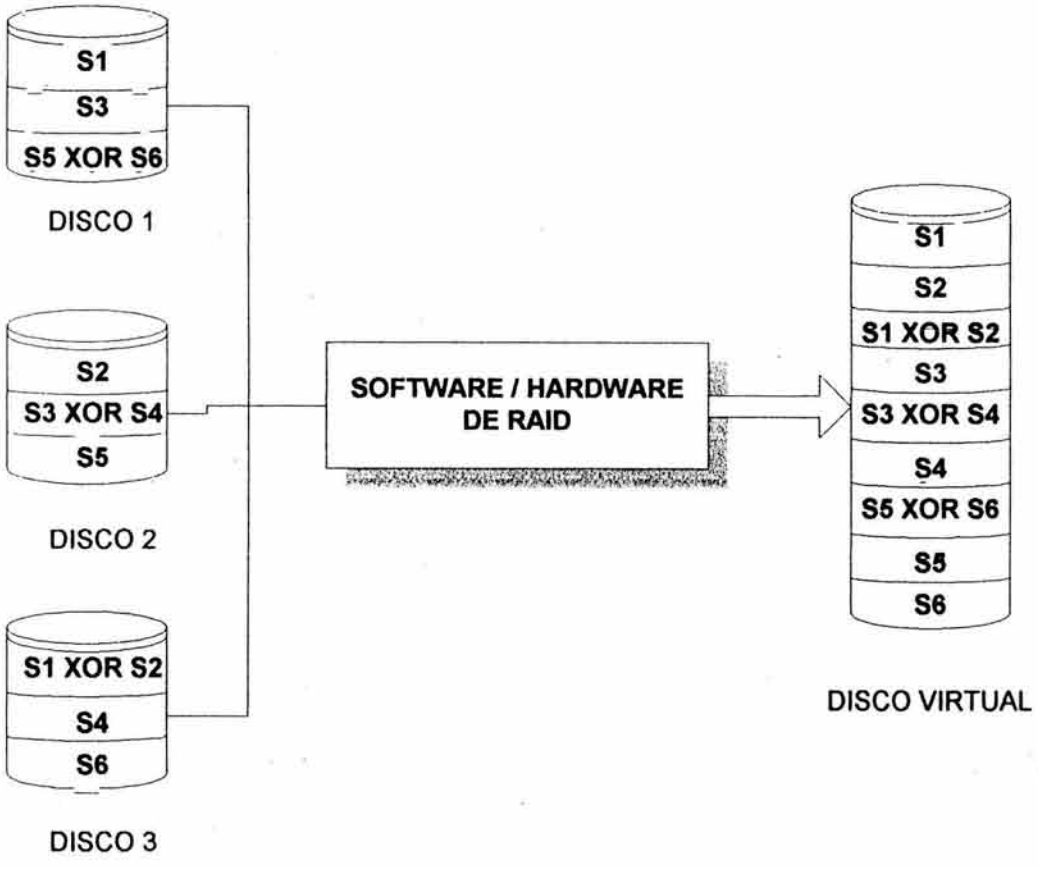


Figura 2.2.5 RAID 5

Como se puede apreciar en la figura, la paridad se distribuye en cada uno de los discos que constituyen el arreglo o disco virtual eliminando el cuello de botella de la configuración de RAID 3.

Ventajas:

- Se tiene redundancia en los datos a través de la paridad.
- Nos protege de la falla de un disco (sólo uno).
- Es más barato que configurar Mirroring.

- Soportado por la mayoría de los proveedores tanto de software como de hardware.
- Es el nivel de RAID más popular en las configuraciones por hardware.

Desventajas:

- Una falla en dos o más discos significa la pérdida de toda la información
- El desempeño en las escrituras es afectado por el tiempo que le lleva al sistema calcular la paridad si se trata de RAID por software.
- En caso de la pérdida de un disco, los datos son leídos a través de la paridad, lo que hace el proceso sumamente lento.
- No es recomendable para sistemas cuyas operaciones de I/O son en un 25% o más, operaciones de escritura.[9]
- Un tercio del espacio utilizado para esta configuración es utilizado para almacenar la paridad, por lo que no se podrá utilizar para los datos del usuario. Existe un desperdicio.
- Se requiere un mínimo de 3 discos físicos para realizar esta configuración.

RAID 6 - Block -level striping with dual distributed parity

Este nivel es muy parecido al RAID 5 pero implementa el cálculo de doble paridad. Mientras RAID 5 calcula paridad de datos para distribuirlos a lo largo de todos los discos que conforman el arreglo, RAID 6 hace dos veces esta operación y distribuye de igual manera el resultado.

Otra diferencia con respecto a RAID 5 es el hecho de que RAID 6 hace striping con bloques de información en lugar de bytes.

---

[9] Según documentación de Sun Microsystems

La gran desventaja de este tipo de configuración es la afectación al desempeño en cuanto a las escrituras se refiere por el hecho de tener que calcular la doble paridad.

Ventajas:

- Se tiene redundancia en los datos a través de la paridad.
- Soporta la falla de hasta dos discos.

Desventajas:

- Una falla en tres o más discos significa la pérdida de toda la información
- El desempeño en las escrituras es afectado por el tiempo que le lleva al sistema calcular la doble paridad si se trata de RAID por software.
- Se requiere de un mínimo de cuatro discos para poder formar esta configuración.
- No todo el espacio está disponible para datos de usuario.
- Representa un costo extra

2.2.2 NIVELES ANIDADOS

Se les llama "niveles anidados de RAID" a aquellas configuraciones que integran o combinan las características de 2 niveles de RAID para generar un nuevo nivel con mejor tolerancia a fallas en general que aquellos que lo conforman.

Este tipo de configuración por lo regular es creado tomando un conjunto de discos físicos y dividiéndolos en dos grupos. Con cada grupo se crea un disco virtual

utilizando un nivel simple de RAID, posteriormente se toman estos discos virtuales para formar otro disco virtual utilizando otro nivel simple de RAID.

El hecho de que se tengan dos niveles simples de RAID diferentes aplicados a los discos, significa que hay dos maneras de aplicarlos, uno primero que el otro o viceversa. En general la forma en que están aplicados estos niveles, no agrega mejor desempeño o rapidez al disco virtual final. En general un nivel de RAID X +Y y un nivel Y + X no se diferencian por desempeño, pero sí por la tolerancia a fallas que presenta cada uno.

El orden en que son aplicados los niveles anidados son determinantes para la redundancia o tolerancia a fallas de todo el arreglo. Así por ejemplo un nivel RAID 0 + 1 es menos tolerante a fallas que un nivel RAID 1 + 0.

A continuación se explican las características de estos niveles y se mencionan otros niveles anidados que no son muy conocidos ni utilizados en el mercado y por tanto no existe documentación al respecto.

RAID 0 + 1 – Mirrored Stripes (Espejos en striping)

Este nivel combina las características del striping en cuanto a la mejora del desempeño por la posibilidad de realizar escrituras en paralelo, y las características del mirroring por realizar una copia o espejo de la información adicionando redundancia a toda la configuración.

En esta configuración se realizan dos o más discos virtuales en RAID 1 o striping, posteriormente se toman estos mismos para formar una configuración en espejo, es decir, cada una de las partes del mirror (lo que serían los submirrors), está formada de un disco virtual en striping.

Para ejemplificar lo anterior veamos la siguiente figura.

RAID 0 + 1 - Espejos en striping

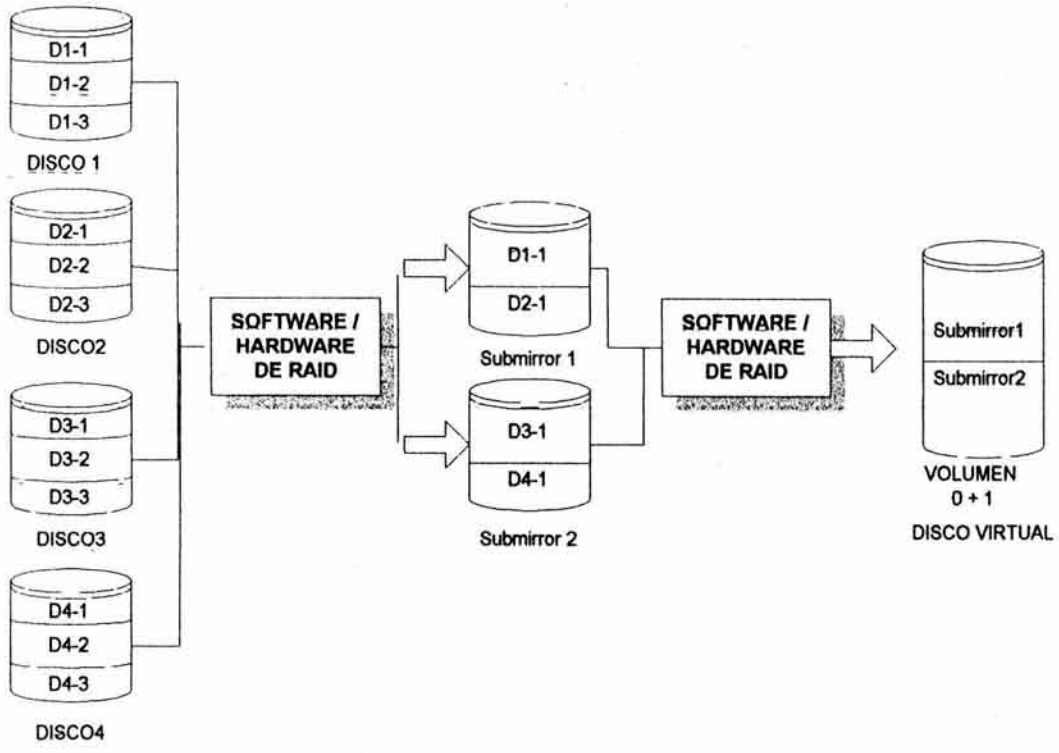


Figura 2.2.6 RAID 0+1

En esta figura se muestra como se forman primero los submirrors con discos físicos utilizando una configuración de striping, posteriormente estos mismos son utilizados para formar el mirror como tal.

La principal ventaja de esta configuración es precisamente la conjunción de las características del striping y las del mirroring: desempeño y redundancia.

A pesar de lo anterior, podemos decir que esta configuración es menos confiable que RAID 1+ 0, ya que si alguno de nuestros discos físicos llegara a fallar, entonces toda una configuración de striping se vería afectada por tanto nos quedaríamos sin un submirror de nuestra configuración final.

La siguiente figura muestra este caso de falla.

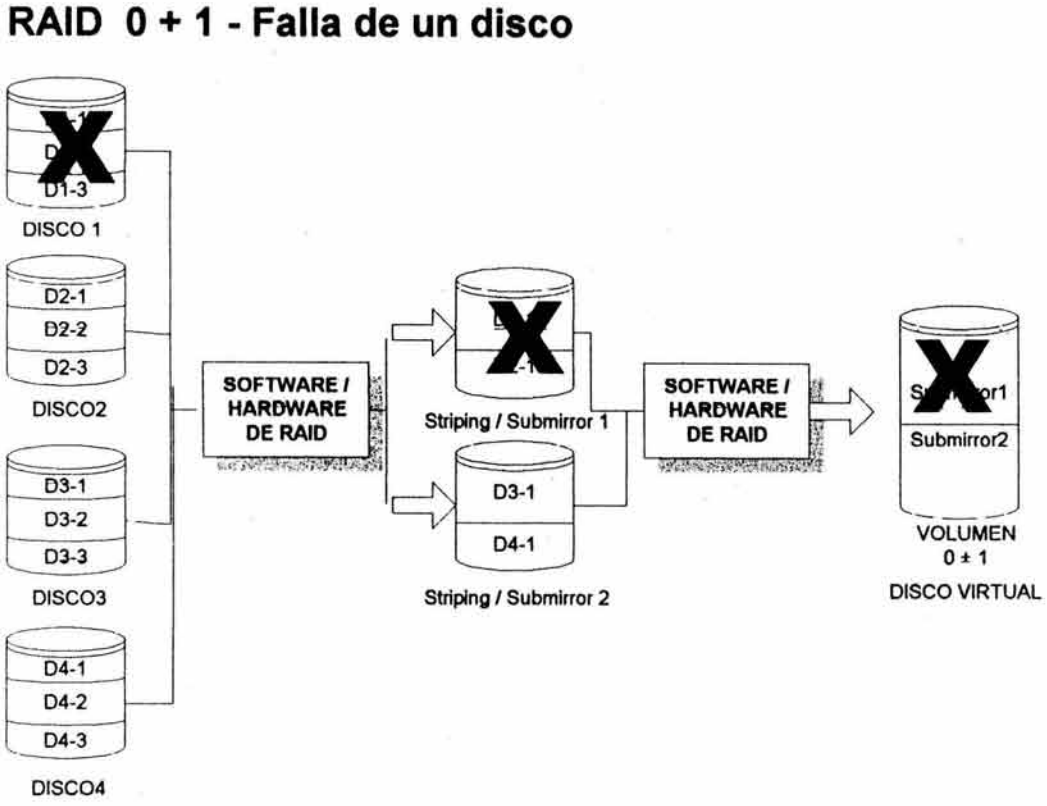


Figura 2.2.7 Falla en RAID 0+1

RAID 1 + 0 – Striped Mirrors (Striping de espejos)

Este nivel es muy parecido al anterior por combinar las características de striping y mirroring, pero implementa un mayor grado de tolerancia a fallas.

En esta configuración se toman varios discos físicos y se configuran en espejo, por ejemplo, en pares formando mirrors de 2 submirrors.

Posteriormente todos estos mirrors formados de submirrors son tomados para formar un disco virtual en striping, lo que significa que cada stripe de esta configuración está espejeado.

El siguiente diagrama muestra esta configuración.

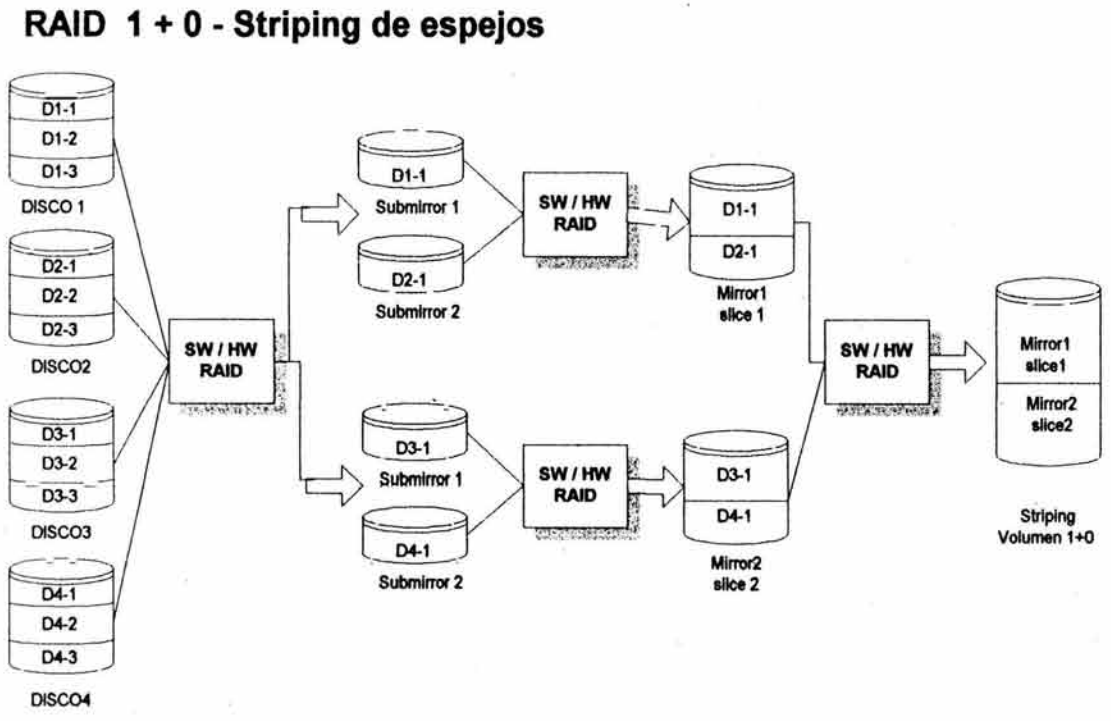


Figura 2.2.8 RAID 1+0

La falla en un disco físico no provocará la pérdida de todo el stripe que forma la configuración final, ya que existe un submirror con esa misma información que podrá seguir siendo utilizado a pesar de la falla.

La siguiente figura muestra este caso.

RAID 1 + 0 - Falla en un disco

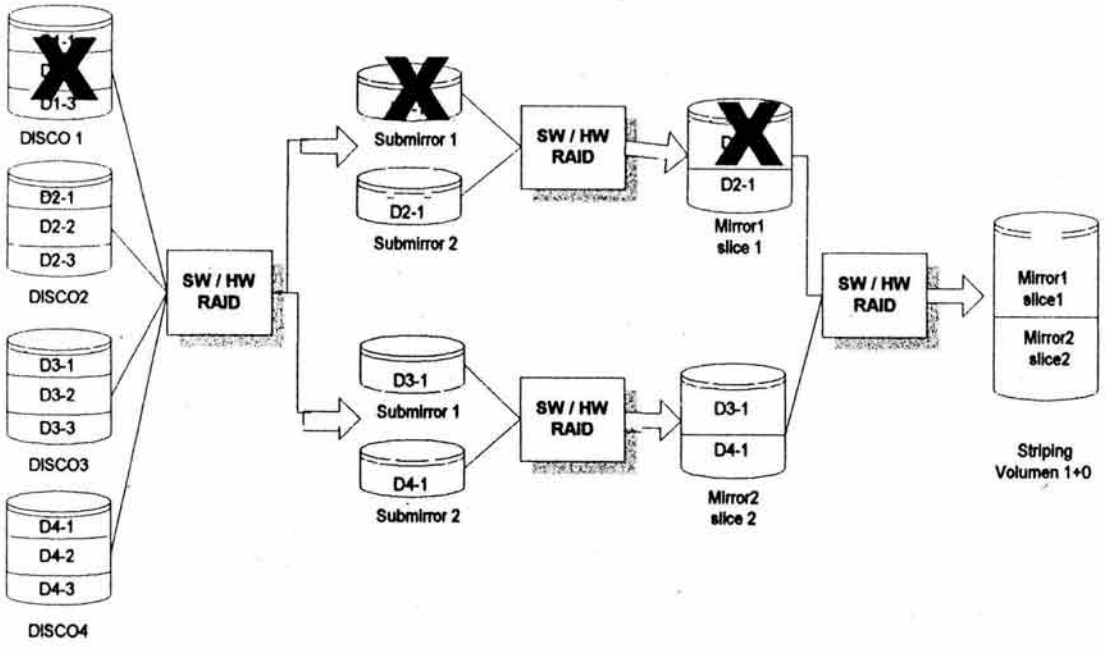


Figura 2.2.9 Falla en RAID 1+0

Otros niveles anidados

- RAID 0 + 3 – Se trata de una configuración en RAID 3 (Striping con paridad dedicada) cuyos stripes están formados a su vez por discos virtuales en striping simple.
- RAID 3 + 0 – Se trata de una configuración de RAID 0 (striping simple) cuyos stripes están formados por discos virtuales en RAID 3.
- RAID 0 + 5 – Se trata de una configuración de RAID 5 (striping con paridad distribuida) cuyos stripes están formados por discos virtuales en striping simple.
- RAID 5 + 0 – Se trata de una configuración de RAID 0 (striping simple) cuyos stripes están formados por discos virtuales en RAID 5.

Los niveles de RAID más comerciales, 0, 1, 5 y 0+1, pueden ser implementados por varias herramientas de software, entre ellas se encuentran VERITAS Volume

Manager y Legato Disk Suite, las cuales Sun Microsystems comercializa para utilizarlas junto con sus equipos de almacenamiento y servidores.

En este trabajo de tesis nos enfocaremos solamente a VERITAS Volume Manager y en el siguiente capítulo revisaremos sus características más sobresalientes y las del equipo de almacenamiento Sun StorEdge A5200 de Sun Microsystems. Ambos productos, software y hardware, son utilizados en el capítulo 4 para ilustrar un ejemplo de la utilización de los niveles de RAID en un caso práctico.

CAPÍTULO 3. VOLUME MANAGER

3.1 INTRODUCCIÓN A LOS ARREGLOS DE DISCOS A5200 DE SUN MICROSYSTEMS.

Las técnicas RAID pueden ser aplicadas a través de una gran variedad de productos de diferentes proveedores, ya sea a través de hardware o software.

En el caso de tratarse de RAID por hardware, nos referimos a dispositivos que integran dentro de su circuitería las funcionalidades de algún(os) niveles de RAID, o bien, necesitan de una tarjeta electrónica controladora que realice esta misma función.

El otro caso es donde se emplea un software para realizar esta misma funcionalidad y proveer de varios niveles de RAID que pueden ser elegidos según sea el propósito del sistema en donde se utilizará.[1]

En general realizar RAID a través de hardware resulta mucho más atractivo cuando se requiere de un mejor desempeño y rapidez en las operaciones del sistema al que se le aplicarán estas técnicas debido a que todas las operaciones relacionadas a mantener toda la funcionalidad de RAID son realizadas a través del hardware en sí, y no a través del servidor o procesador del sistema destinado a dar servicio a usuarios finales, lo que sería el caso de realizar RAID por software.

A pesar de lo anterior, realizar RAID por software resulta mucho más económico y fácil de implantar, ya que como es común en cualquier aplicación, se dispone de interfaces gráficas y una variedad más extensa de comandos para lograr al configuración deseada.

---

[1] Dentro de esta categoría se encuentra VERITAS Volume Manager

Volume Manager cae dentro de esta categoría, ya que se trata de un software útil en la configuración de niveles de RAID.

Como hemos de notar, para realizar los niveles de RAID, se requiere por lo regular más de un disco físico de almacenamiento. Existen algunos software, como Solstice Disksuite de Legato, que permiten realizar configuraciones con un sólo disco; sin embargo, este tipo de configuraciones resultan poco comunes debido a que no tiene sentido realizarlas por lo limitadas que resultan en cuanto a confiabilidad y disponibilidad de datos.

Cuando un sistema o servidor empieza a crecer en cuanto a los datos que alberga, o bien, ya es grande desde su surgimiento, no se utilizan discos aislados para su funcionamiento; mas bien se utilizan dispositivos especiales que contienen una gran variedad de discos los cuales son comúnmente llamados *arreglos de discos*.

Hay mucha variedad de arreglos de discos de varios proveedores, es más, cada proveedor integra varios modelos de arreglos que pueden abarcar un mercado más extenso. Existen arreglos de discos para aplicaciones o servidores pequeños, medianos y moderadamente grandes, cuando se trata de sistemas demasiado extensos, este panorama suele cambiarse hacia otro esquema: las redes de almacenamiento (SAN).[2]

Las redes de almacenamiento no serán tratadas en este trabajo de tesis, solamente se mencionarán algunos conceptos generales y sus beneficios en el capítulo 5; sin embargo cabe mencionar que la tendencia de muchos sistemas es el crecimiento exponencial de recursos de almacenamiento. Cuando resulta insuficiente la utilización de arreglos de discos y técnicas RAID, el siguiente paso podría ser una red de almacenamiento (SAN) que no solamente integre esos dos

[2] SAN proviene de Storage Attached Network.

factores; sino que provea de una mayor flexibilidad de crecimiento y una mayor disponibilidad de datos.

Mencionaremos las características generales de una familia de arreglos de almacenamiento de Sun Microsystems, Sun StorEdge A5000 (A5100, A5200), debido su utilización en el caso práctico del siguiente capítulo.

Sun StorEdge A5200

El arreglo A5200 de Sun comparte las mismas características generales de la familia de A5XXX, es decir, el A5000 y el A5100 tienen las mismas características que el A5200.[3]

Existen dos variantes de estos arreglos, los que integran 14 discos y los que integran 22. La diferencia entre estos es precisamente la cantidad de discos que albergan y las capacidades de estos. En general una caja o arreglo de 22 discos contendrá discos de menor capacidad que los que alberga una caja de 14 discos.

Los arreglos A5XXX utilizan conexiones de fibra óptica y el estándar de ANSI/ISO FC-AL (Fibre Channel Arbitrared Loop), ó SCSI-3 como también se le conoce.

Se dice que estos arreglos son de alta disponibilidad por las características que poseen, además de ser posible expandir su capacidad de almacenamiento conectando cajas en cadena. Cada caja es capaz de almacenar hasta 1606 GB de almacenamiento por cada una con 22 discos de 73 GB cada uno. Existen otras configuraciones, aunque actualmente en el mercado solamente se venden los arreglos A5200, sus antecesores ya no son comercializados por Sun.

Algunas características generales de estos equipos son las siguientes:

---

[3] A la fecha Sun Microsystems solamente comercializa el A5200, los anteriores han salido de venta.

- Convertidores FC-AL, los cuales se les conoce como GBIC (Gigabit Interface Converters)[4]
- Una tarjeta *backplane* al frente y otra en la parte trasera en donde van insertados los discos. Cada tarjeta puede albergar 7 u 11 drives.
- Hasta 4 cajas pueden ser conectadas en cadena o a través de HUBs, dependiendo de la configuración deseada.
- Para conectar el dispositivo al servidor que lo utilizará, se utiliza una tarjeta que es insertada en el servidor la cual es llamada *FC-AL Host Adapter* y contiene dos interfaces de conexión de fibra (FC-AL) para implementar redundancia.
- Se utiliza para conectar el dispositivo al servidor un cable de fibra óptica con un distancia máxima de 500 metros.
- Dentro de la caja o arreglo se cuenta con hasta dos tarjetas a las cuales van conectados los cables de fibra óptica que se utilizarán para la conexión con el servidor. Estas tarjetas son llamadas *Interface Boards*, y contienen dos interfaces cada una en donde se insertan dos GBICs para proveer redundancia y dualidad de puertos.
- El arreglo integra componentes redundantes como por ejemplo, fuentes de poder, ventiladores y *leds* de diagnóstico que nos indican cuando alguno de los componentes tiene un error.
- Se cuenta con 14 ó 22 discos de FC-AL los cuales tienen dos puertos de conexión para proveer dos rutas hacia los datos almacenados. Estos discos son *hot-pluggable*, es decir, pueden ser insertados o retirados en "caliente", sin necesidad de dar de baja el equipo o realizar alguna operación especial. Claro, todo lleva un procedimiento, a pesar de ser hot-pluggable, debe tenerse cuidado al insertar o retirar discos debido a que estas operaciones inciden directamente en la información contenida y que le servidor podría estar utilizando en ese preciso momento.

[4] Estos GBIC también son utilizados en componentes para SAN como switches, por ejemplo.

- Estos arreglos cuentan con una pantalla electro-luminiscente (electro-luminiscent display) ó *touch screen* que es llamada *Front Panel Module*, mediante la cual puede realizarse la configuración de la caja a través de sus menús. Se puede acceder a ellos tocando la pantalla.

Las siguientes figuras muestran algunas de estas características.

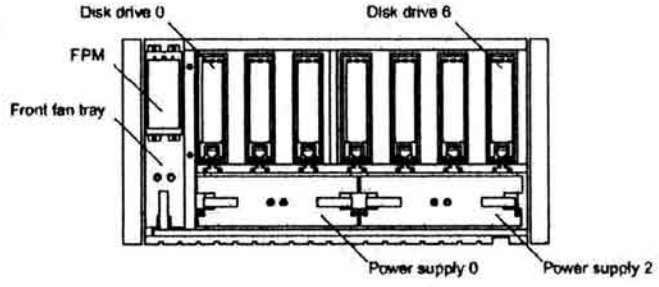


FIGURE 1-1 Front Components (14 slot)

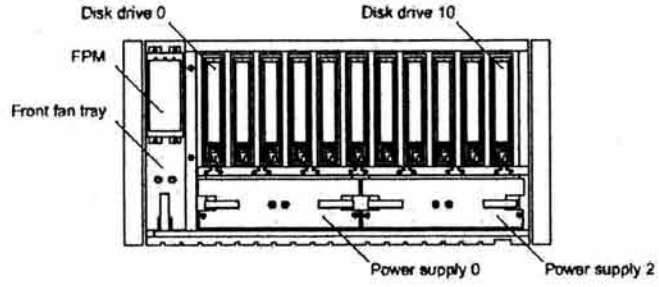


FIGURE 1-2 Front Components (22 slot)

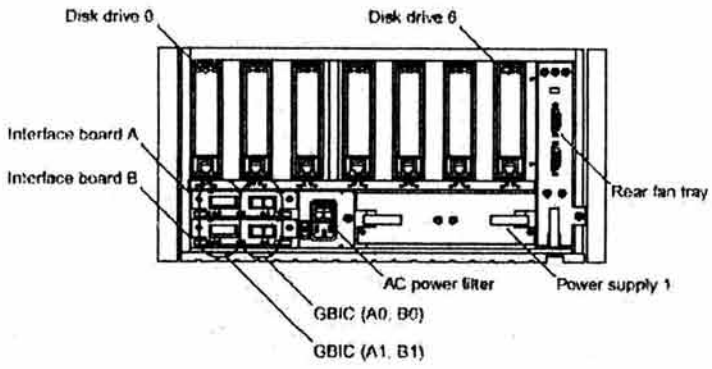


FIGURE 1-3 Rear Components (14-slot)

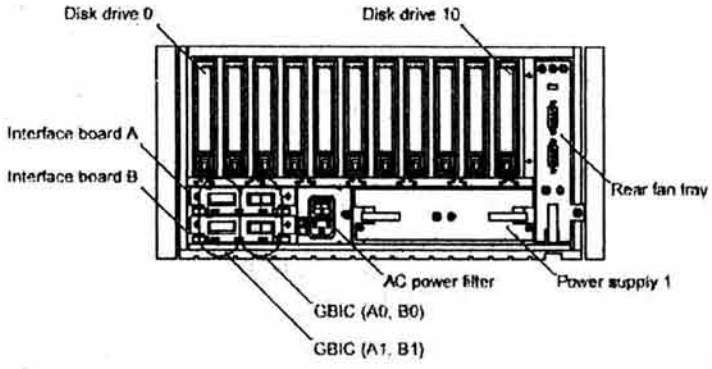


FIGURE 1-4 Rear Components (22-slot)

Figura 3.1.1 Características Externas del A5XXX de 14 y 22 drives.

Más específicamente, el interior del arreglo tiene las siguientes características generales:

- No existen cables de interconexión de los componentes, toda la comunicación y conexión de componentes se hace a través de la llamada *Interconnect Asembly*, la cual está formada por una motherboard y un Centerplane al que van insertados los demás componentes.
- Dos bandejas de ventiladores están disponibles dentro del arreglo. Ambas son necesarias para la operación de la caja.
- El arreglo cuenta con tres fuentes de alimentación. Solamente dos son necesarias para la operación de la caja, la tercera es redundante.
- Un filtro de corriente alterna.

- Dos *interface boards*, para proveer la conexión de la caja con el servidor que la utilizará. Solamente una es requerida para proveer la comunicación servidor-arreglo.
- En caso de contar con ambas *interface boards*, se puede contar con hasta 4 GBICs por arreglo de discos. Solamente un GBIC es necesario para la conexión servidor-arreglo.
- Dos backplanes de discos, uno en la parte frontal y el otro en la parte trasera. A través de estos backplanes se hace la inserción y manejo de los discos de FC-AL que contiene el arreglo.
- Dependiendo del tipo de arreglo que se trate, puede albergar 14 ó 22 discos de FC-AL, los cuales son *hot-pluggable*.

Una característica importante de estos dispositivos es la existencia de varios componentes *hot-pluggable*, es decir, no es necesario dar de baja el equipo para poder retirarlos o insertarlos. Entre ellos se encuentran las fuentes de alimentación (*power supplies – PS)*, bandejas de ventiladores *(Fan Tray –FT)*, Front Panel Module (FPM), las *interface boards*, los GBICs y por supuesto, los discos FC-AL (DD).

La siguiente figura muestra la interconexión de cada uno de estos componentes dentro del arreglo.

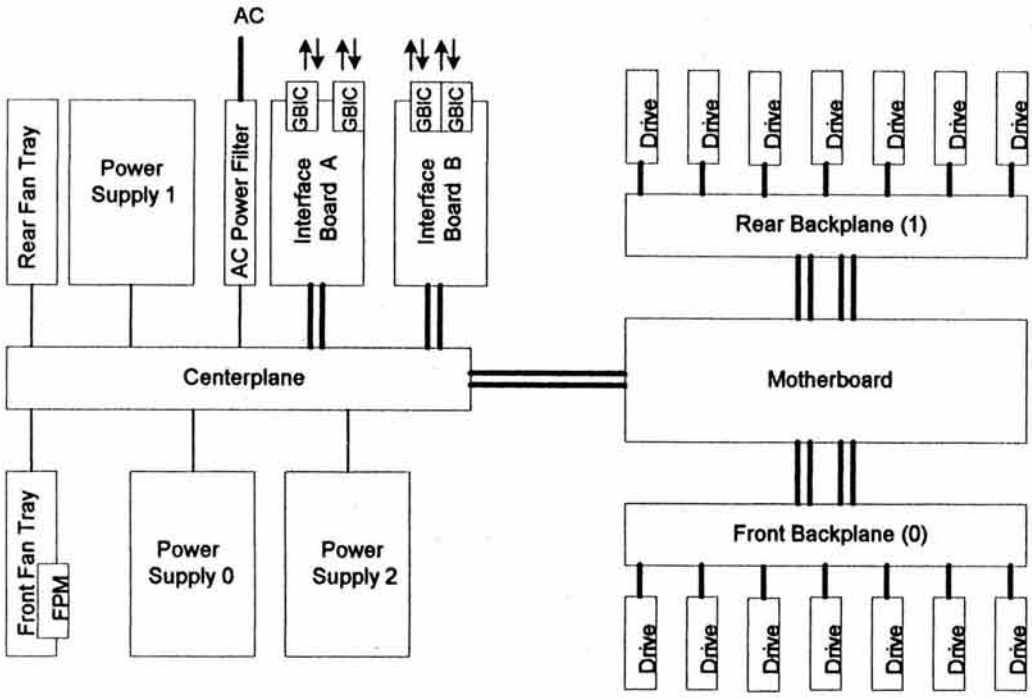


Figura 3.1.2 Componentes dentro del arreglo.

La función general de algunos de estos componentes es la siguiente:

FC-100 HOST ADAPTER. Si bien esta tarjeta no está dentro del arreglo, es una pieza fundamental para la conexión de éste al servidor. Esta tarjeta debe estar insertada dentro del servidor, en este caso un servidor de Sun Microsystems también[5]. Esta tarjeta es el punto de conexión dentro del servidor con el arreglo, su nombre indica que la velocidad a la que puede transmitir datos es a 100 Mb/s. Contiene dos compartimentos donde pueden ser insertados 2 GBICs o convertidores con los que se provee una dualidad de conexiones hacia el arreglo, una característica importante cuando se trata de dispositivos de alta disponibilidad. Estos GBICs son también hot-pluggable lo que incrementa su disponibilidad también. La fibra óptica que es insertada en estos GBICS del lado del Host

[5] Cualquier dispositivo de almacenamiento requiere de una tarjeta adaptadora ya sea SCSI, FC-AL u de otro tipo.

adpater encuentra también su conexión en los GBICS de las Interface Boards de la figura anterior. Es así como se conecta el arreglo al servidor.

GIGABIT INTERFACE CONVERTER (GBIC). Dos de estos convertidores van insertados en el FC-100 Host Adapter y otros dos en la Inerface Board dentro del arreglo de discos para proveer la conversión de señales eléctricas a ópticas y viceversa. Los arreglos A5200 utilizan fibra óptica para su conexión con el servidor o entre ellos cuando se trata de arreglos en conectados en cascada; sin embargo, ya una vez dentro del arreglo, las señales que utiliza son eléctricas, el GBIC hace la conversión de señales que permiten la comunicación entre servidor – arreglo. Una de las características más sobresaliente de estos componentes es que son hot-pluggable, o que pueden conectarse o retirarse en "caliente".

INTERFACE BOARD.

Esta tarjeta es la más importante del arreglo debido a que a través de ella se realiza todo el control y manejo de toda la caja de discos, desde configuración hasta monitoreo de las condiciones ambientales del equipo. Un A5200 puede contener hasta 2 interface boards, aunque para la conexión con el servidor solamente es necesaria una. Existen configuraciones en las cuales 2 servidores o hosts utilizan un mismo arreglo A5200, y cada uno de ellos se conecta al arreglo a través de una de las dos interface boards de la caja. Cada Interface board contine dos slots para insertar hasta 2 hot-pluggable GBICs. Una interface board va a proveer soporte a los puertos A de cada uno de los discos duros dentro del arreglo; y una interface board B proveerá soporte al puerto B de los discos. En caso de contar con solamente una interface board, solamente se tendrá soporte para un puerto de los discos de FC-AL del arreglo, aunque en realidad ésta no es una configuración común.

Los A5200 pueden ser configurados de dos maneras a través del Front Panel Module integrado: con un *split loop*, o con un *full loop*.

El split loop es una configuración en la que los discos de enfrente van a separarse de los discos de la parte trasera. Esta configuración es útil cuando se requiere conectar un sólo A5200 hacia dos servidores diferentes o cuandò se requiere de una mejor rapidez de transferencia de datos ya que, si bien el arreglo de discos es sumamente rápido, con transferencias grandes podríamos generar un poco de latencia al tratarse de demasiados discos conectados sobre un mismo loop o camino.

El full loop es una configuración en donde tanto los discos de enfrente como los de atrás están conectados al mismo loop o camino. Esta configuración es más usada cuando se trata de un arreglo usado por solamente un servidor y cuando se requiere mayor cantidad de almacenamiento ya que las conexiones en cadena (daisy – chain) solamente son soportadas cuando a cada arreglo dentro de la cadena se le configura con un full loop.

DISCOS DUROS FC-AL (Drives).

Dentro del arreglo vamos a contar con hasta 14 o 22 discos duros de FC-AL que son hot-pluggable y que manejan el estándar de FC-AL, aunque internamente entienden los mismos comandos que los discos SCSI. Estos discos son de 3.5 pulgadas, cuentan con dos rutas por los cuales pueden tener comunicación para la escritura y lectura de datos y con un identificador conocido como World Wide Number que es un número muy extenso que se designa a cualquier dispositivo de fibra óptica, parecido a lo que sería una dirección MAC ó Ethernet de una tarjeta de red.

Este WWN es único para cada dispositivo y a través de él puede también hacerse una distinción entre cada disco. El arreglo en sí también cuenta con un identificador de este tipo.

FRONT PANEL MODULE (FPM).

Un arreglo de este tipo se distingue por contar con una pequeña pantalla luminiscente conocida como Front Panel Module y desde la cual puede consultarse el estado general de los componentes del arreglo y hacer la configuración del mismo estableciendo un split loop o un full loop.

Esta pantalla es un *touch screen*, es decir, con sólo tocarlo se puede navegar a través de sus menús y submenús hasta llegar a obtener la información del status del equipo o bien lograr la configuración idónea.

Las operaciones que se pueden realizar a través de esta pantalla son:

- Desplegar el estado de la caja, los discos, los loops y los errores en general que han ocurrido
- Desplegar información como el, WWN, el nombre del arreglo, su identificador, entre otros.
- Configurar la caja con su nombre, el tipo de loop y su identificador.

El identificador de la caja es un número del 0 al 3 que se le asigna al arreglo para identificarlo en caso de que varios arreglos sean conectados en cadena, el nombre de la caja tiene la misma función.

A continuación vemos una figura que nos muestra los menús desplegados en el Front Panel Module.

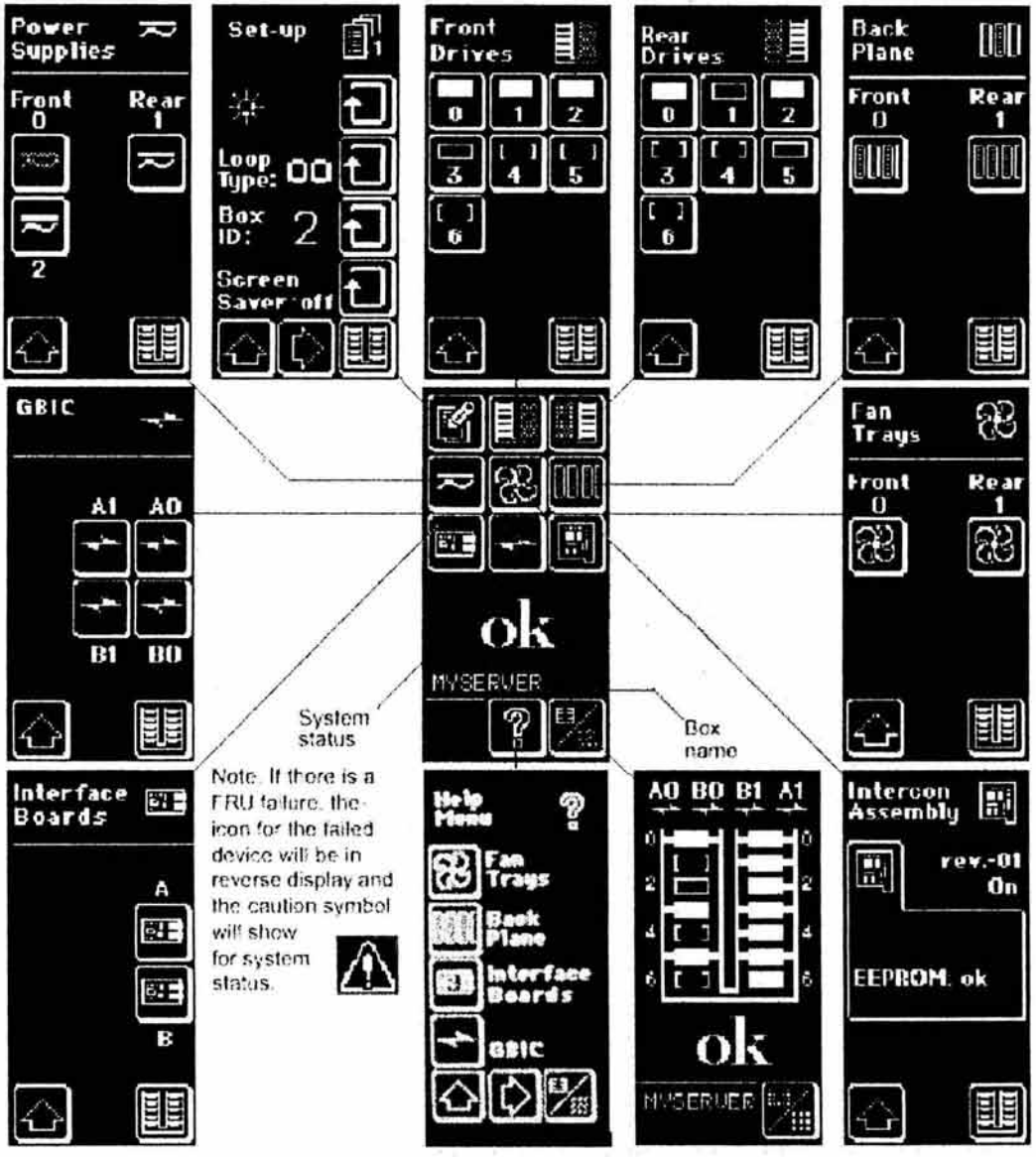


Figura 3.1.3 Front Panel Module.

Los arreglos A5200 son equipos con cierta inteligencia para prevenir la falla en alguno de sus componentes, ya que al detectar un error grave, el mismo equipo se dará de baja.

A pesar de lo anterior, estos arreglos no tienen la capacidad de realizar RAID por hardware, por lo que esa tarea es realizada a través de un software que provee esta funcionalidad. Una de las herramientas ampliamente utilizadas en el mercado y en nuestro ejemplo práctico es VERITAS Volume Manager, el cual se explica a grandes rasgos.

3.2 CARACTERÍSTICAS GENERALES DE VOLUME MANAGER.

Volume Manager es un software realizado y comercializado por VERITAS Software Corporation[6].

3.2.1 VERITAS SOFTWARE CORPORATION.

VERITAS Software Corporation es una empresa dedicada a proveer software para la administración del almacenamiento y protección de datos, la disponibilidad de aplicaciones y la recuperación a desastres (pérdida de datos).

Según datos publicados en su sitio, el 86% de la inversión de 500 empresas en el mundo, recaen en las soluciones de almacenamiento de VERITAS Software Corporation, además de estar situada entre las primeras 10 empresas más importantes del mundo que se dedican a la producción de software.

VERITAS Software fue fundada en 1989 para desarrollar y vender productos de alta disponibilidad. En Abril de 1997, se unió con Open Vision Technologies para expandir sus propias soluciones de respaldo y almacenamiento jerárquico con el sistema operativo propio de Open Vision.

[6] Existe una alianza entre Sun Microsystems y VERITAS para que el primero comercialice también l.os productos del último.

En mayo de 1999, VERITAS se extendió aún más con la adquisición del Network and Storage Management Group de Seagate Software. Este paso le permitió diversificar sus soluciones que solamente existían para el mercado de servidores UNIX al de Windows NT y NetWare.

Esta diversificación le permitió a VERITAS poder utilizar un producto común para varias plataformas, sistemas operativos y ambientes de cómputo y facilitar la administración y bajar los costos.

El corporativo de VERITAS reside en Montain View, California.

Los productos de VERITAS se enfocan principalmente al manejo de volúmenes de información y almacenamiento de datos, así como aplicaciones para respaldos y recuperación a fallas a través de *Clusters*.

El software para manejar almacenamiento es precisamente Volume Manager, el cual está disponible para varias plataformas, entre ellas Sun Solaris; de ello depende la manera de instalar la herramienta y en algunas ocasiones, la manera de utilizarse ya que cada plataforma tiene sus propias peculiaridades.

Tomaremos las características particulares de Volume Manager instalado en una plataforma Sun Solaris porque esa es la plataforma que usamos en nuestro caso práctico.

3.2.2 CARACTERÍSTICAS DE LA INSTALACIÓN DE VOLUME MANAGER.

La instalación de Volume Manager en un equipo Sun Solaris consta de dos pasos importantes: la adición de los paquetes que forman la aplicación como tal y, la configuración de la herramienta.

La instalación de los paquetes se realiza como cualquier otro dentro de una plataforma Solaris, es decir, a través de los comandos bien conocidos por los administradores de estos equipos (*pkgadd)*[7].

La configuración de la herramienta es el paso de la instalación que realiza ajustes importantes a nuestra información en caso de ser necesario. Para que Volume Manager pueda administrar nuestros discos, es necesario que les adicione datos de control, o, como suele llamarse a esta operación, los tome bajo su control.

Cuando se configura Volume Manager, éste revisa el tipo de caja o arreglo de discos que tiene conectado el servidor al cual se le está instalando el producto, también revisa el uso que se le está dando a los discos y al sistema en general.

La configuración de Volume Manager puede hacerse escogiendo una de dos opciones:

- *Quick Installation*
- *Custom Installation*

Quick Installation.

Al realizar una configuración rápida, Volume Manager decide las acciones que habrá de realizar sobre los discos con datos ya existentes y con los discos que no tienen información aún. Sus decisiones se resumen a lo siguiente:

1. Volume Manager revisa si el disco de inicio (boot) del sistema está ya bajo su control, en caso de que no sea así, intenta *encapsular* o adicionar sus datos de control sobre él.
2. Volume Manager revisa cada uno de los discos del arreglo de discos. Si el disco ya tiene datos, entonces lo encapsula también.

[7] *Pkgadd* es un comando de Solaris para adicionar software en formato de "paquete" al sistema.

3. Volume Manager inicializa cada uno de los discos del arreglo que no tenga datos, dejando en ellos sus datos de control..

Todas estas operaciones son realizadas sin interacción del administrador, Volume Manager las realiza sin preguntar ningún detalle.

Se dice que este tipo de configuración no es recomendable debido a que el administrador no tiene control de las acciones a realizarse, además que el encapsular el disco de inicio es una tarea complicada y que no muchas veces se realiza con éxito, inclusive se comenta que el encapsular el disco de inicio solamente debe realizarse en el caso de que se desee hacer un espejo (mirror) del disco de inicio, en ningún otro caso es recomendable debido a que la recuperación a una falla con el disco de inicio utilizando Volume Manager no es una tarea trivial y en algunos casos tampoco tiene éxito ocasionando la pérdida total del sistema.

Custom Installation.

La configuración personalizada es la más recomendable de realizar ya que el administrador tiene el control sobre las acciones que habrán de realizarse a los discos dentro del sistema y dentro del arreglo de discos en sí. Las acciones que realiza Volume Manager en este caso se resumen a lo siguiente:

1. Volume Manager pregunta si debe encapsular el disco de inicio.
2. Volume Manager detecta las controladoras de discos y pregunta lo que habrá de realizar con cada disco, ya sea una encapsulación o una inicialización.
3. Volume Manager encapsula o inicializa los discos que le fueron indicados.

En esta ocasión el administrador podrá realizar una encapsulación o una inicialización de discos a sus necesidades, si los discos ya tienen datos y se

decide hacer una inicialización, se perderán totalmente los datos ya existentes. En caso de realizarse una encapsulación, los datos serán preservados.

Cabe aclarar que la encapsulación se utiliza cuando deseamos conservar datos ya existentes en un disco y al mismo tiempo deseamos que ese disco sea manejado por Volume Manager; sin embargo hay restricciones para poder realizar esta operación, ya que Volume Manager necesita de un poco de espacio para adicionar sus datos de control al disco.

En cualquiera de los casos, la instalación de Volume Manager coloca los discos que toma bajo su control en un grupo por omisión llamado *rootdg*. En cada sistema debe haber por lo menos este grupo.

Un grupo es una colección de discos de Volume Manager que comparten una configuración común. Típicamente el grupo de discos contiene volúmenes que tienen relación de alguna manera como por ejemplo, volúmenes de sistemas de archivos (file systems) que pertenecen a un departamento particular ó volúmenes de bases de datos particulares.

Otros grupos de discos pueden ser creados posteriormente e incluir los discos inicializados o tomados bajo el control de Volume Manager hacia cualquier grupo, la única restricción es que debe haber al menos un grupo llamado rootdg y en cada grupo debe haber al menos un disco.[8]

La idea de hacer grupos de discos surge de la necesidad, en muchas ocasiones, de discriminar o separar la información en grupos de trabajo o cuando un conjunto de discos van a ser utilizados por un host y otro conjunto por otro host diferente.

La configuración de grupos, hosts, discos, etc, es guardada internamente en el disco, de tal modo de que cada disco contiene información importante del host al

[8] Al quitar el último disco de un grupo se borra el grupo por completo.

cual pertenece, entre otras cosas. Esta información es guardada por Volume Manager al momento en que toma control sobre cada disco.

Formato físico de los discos con Volume Manager.

Volume Manager necesita poner bajo su control los discos que habrá de administrar. Cuando decimos que la herramienta los pone bajo su control, lo que realmente está realizando la aplicación es dar un formato especial a los discos adicionando datos de control.

El formato que da Volume Manager se distingue por tener dos especie de "particiones" las cuales son llamadas regiones. Así pues hay una región privada y una región pública.

La región privada alberga los datos de control que utiliza Volume Manager para manejar los discos y la región pública es el resto del disco que será el espacio total de ese disco que estará disponible para que el usuario pueda realizar los discos virtuales o volúmenes que necesite.

Cabe destacar que todos los discos que se ponen bajo el control de Volume Manager van a contener este formato, por tanto cuando se realiza una encapsulación de discos para conservar datos existentes, se debe tomar en cuenta que esta operación solamente va a ser posible cuando en el disco haya espacio suficiente para albergar la región privada del disco.

La región pública del disco será utilizada por Volume Manager para formar los Volúmenes o discos virtuales que el administrador le indique, lo que hay que

destacar es que cuando se realizan volúmenes se utilizan pedazos de cada uno de los discos físicos, pero estos pedazos no siempre van a ser del mismo tamaño, por lo que Volume Manager va tomando pequeñas porciones de la región pública conforme las vaya utilizando. Estas porciones o pedazos son llamados subdiscos.

La siguiente figura ilustra el formato que Volume Manager da a los discos:

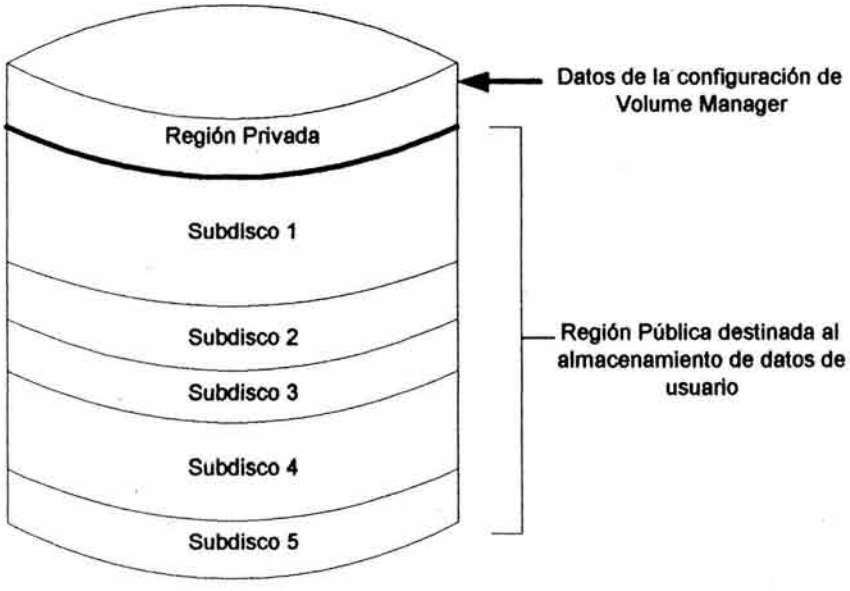


Figura 3.2.1 Particionamiento de discos a través de Volume Manager.

Dentro de la región privada Volume Manager guarda los siguientes datos:

Disk Header. Esta cabecera contiene el identificador (ID) del host al que el disco pertenece, de tal modo que solamente un host pueda tener acceso a la información contenida en el disco. También se guarda un identificador del disco de 64 bytes.

Configuration Database. La base de datos de configuración guarda información acerca de la configuración de un grupo de discos particular. Por omisión, Volume Manager guarda cuatro copias de esta base de datos para evitar el perder toda la configuración en el caso de un error en alguno de los discos.

Cada copia de la base de datos guarda la siguiente información:

- *Dgname.* El nombre del grupo de discos al cual pertenece al disco y que le es asignado ya sea por Volume Manager o bien por el propio administrador.

- *Dgid.* Un identificador de 64 bytes que es único universalmente y que le es asignado al grupo de discos al cual pertenece el disco.

- *Registros.* Existe un registro por cada objeto de Volume Manager. Un objeto de Volume Manager, como se verá después, es alguna de las instancias con las cuales podemos formar Volúmenes de Volume Manager y que tienen características específicas dentro de la configuración. Podemos decir que los objetos de Volume Manager son análogos a un ícono dentro de un ambiente Windows. Cada ícono puede ser manipulado y tiene características especiales.

Kernel Log. En esta parte se guarda información relacionada con las acciones que se han realizado sobre los objetos pertenecientes a ese grupo y si han ocurrido errores durante esas operaciones. Este kernel log es utilizado cuando se requiere recuperar el estado de grupo de discos después de un *crash* o un reinicio del sistema.

3.2.3 CARACTERÍSTICAS GENERALES DE LA INTERFAZ GRÁFICA DE VOLUME MANAGER.

Volume Manager, como cualquier otra aplicación gráfica, cuenta con la facilidad de menús y submenús, así como botones de acceso rápido, para realizar la mayoría de sus tareas, así como la utilización del ratón para seleccionar, de-seleccionar, expandir y hacer drag-and-drop.

Al ejecutar su interfaz gráfica lo primero que tenemos que ingresar es la contraseña de root[9] del sistema en el cual está instalado Volume Manager, posteriormente veremos una pantalla como la siguiente.

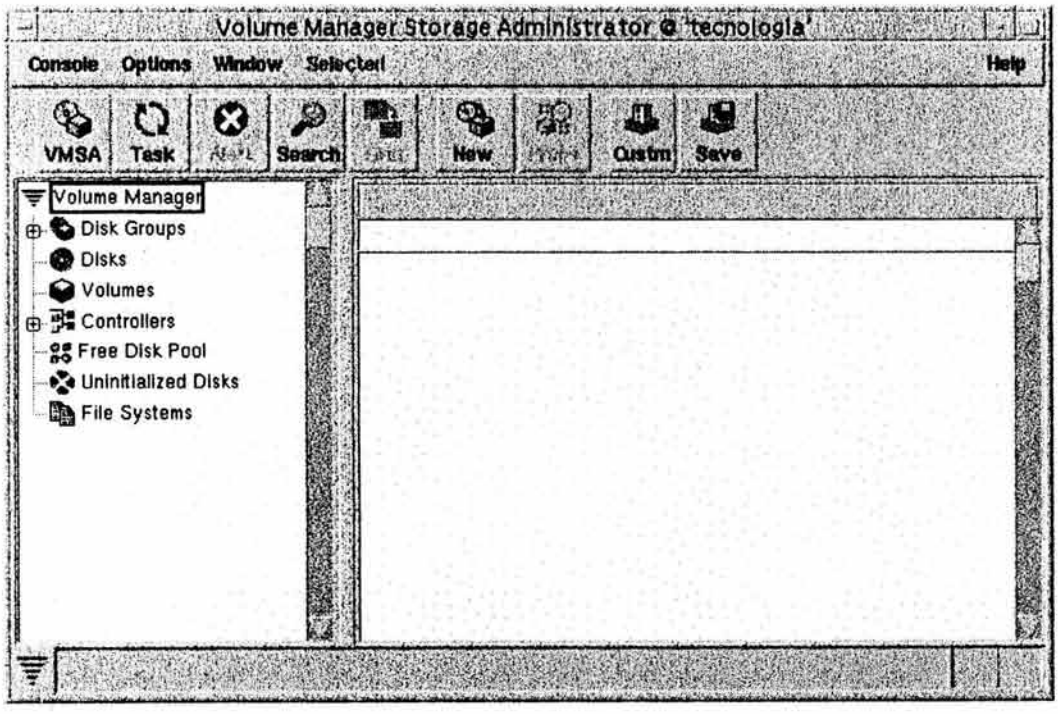


Figura 3.2.2 Pantalla del cliente.

En esta pantalla tenemos los siguientes elementos:

[9] *Root* es el súper usuario o usuario administrador de un sistema UNIX por omisión.

Tool Bar. Esta barra de herramientas provee acceso directo a algunas las funciones más utilizadas de Volume Manager.

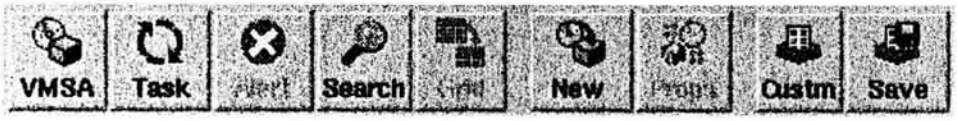


Figura 3.2.3 VMSA Tool Bar

Algunos de estos botones más utilizados son:

TASK: Al momento de realizar alguna acción sobre los objetos de Volume Manager, podemos ver o monitorear el resultado de esa operación a través de este botón.

GRID: Nos permite ver a detalle la configuración o disposición de los objetos de Volume Manager.

NEW: A través de este botón podemos realizar nuevos Volúmenes con cualquiera de los niveles de RAID que Volume Manager soporta (RAID 0, 1, 0+1, 1+0,5).

Menu Bar. Como en cualquier aplicación gráfica, la barra de menú va a contener todos los menús y submenús que contienen toda la funcionalidad del software.

Figura 3.2.4 VMSA Menu Bar

Object Tree. Parecido a Windows, la interfaz gráfica de Volume Manager integra un árbol de objetos desde el cual podemos verificar las propiedades de cada uno

de los objetos que conforman nuestra configuración, modificarlos o aplicarles una acción en específico, expandir el árbol, ver su interdependencia, etc.

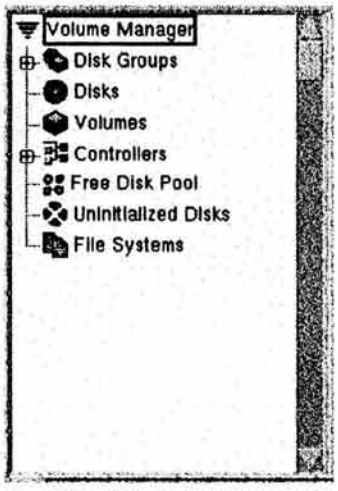


Figura 3.2.5 VMSA Object Tree

Command Launcher. A través de los menús podemos habilitar una ventana más, la llamada command launcher. Esta ventana nos mostrará los comandos o acciones que podemos realizar con Volume Manager a través de su interfaz gráfica; pero ordenados por tipo de objeto, es decir, podemos encontrar primero todos los comandos que tienen que ver con los volúmenes, después los que tienen que ver con los grupos y así sucesivamente para tener una visión clara de lo que podemos hacer con cada objeto dentro de nuestra configuración.

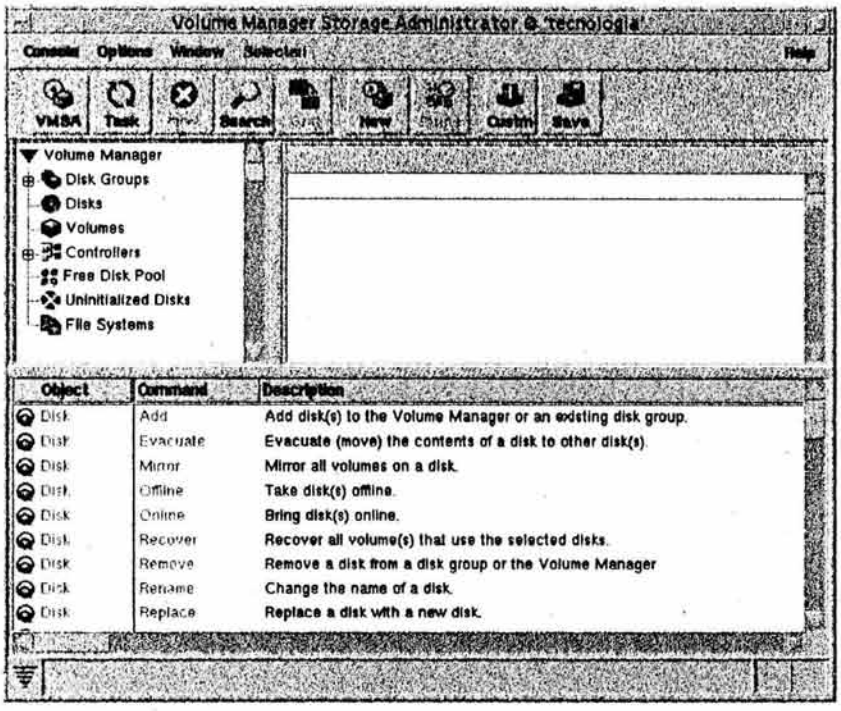


Figura 3.2.6 VMSA Command Launcher

En términos generales, la interfaz gráfica de Volume Manager es bastante intuitiva por lo que hace muy sencilla la tarea de administración.

3.3 LOS OBJETOS DE VOLUME MANAGER.

Se ha hecho referencia a los objetos de Volume Manager en varias ocasiones, sin embargo es necesario explicar a detalle a qué se refieren estos objetos y cuales son sus propiedades.

Observemos las siguientes figuras.

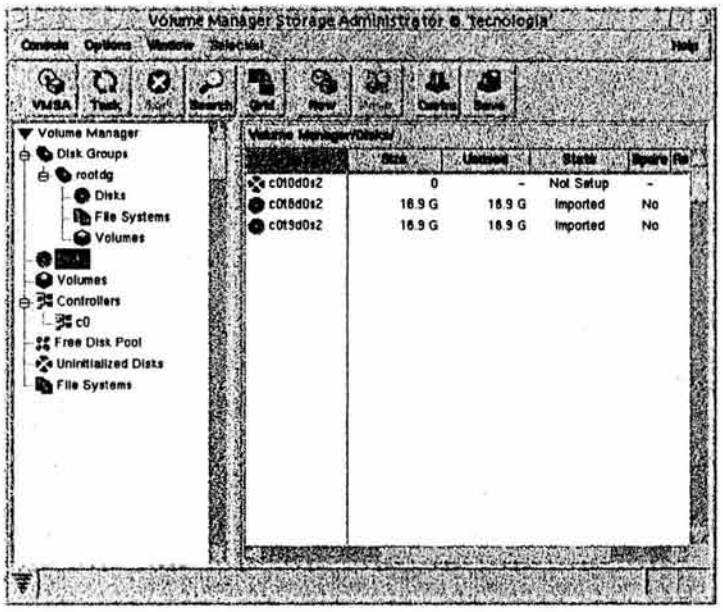


Figura 3.3.1 Vista de los discos en el sistema.

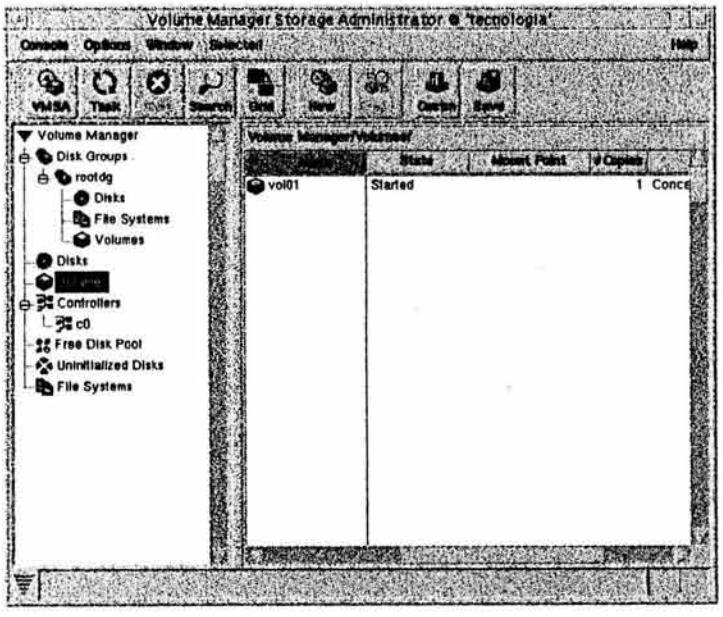


Figura 3.3.2 Vista de los volúmenes existentes en el sistema.

En las figuras anteriores se puede ver la interfaz de Volume Manager mostrando alguna vista de sus componentes.

Cada uno de estos componentes como lo son, *disks, volumes, controllers, etc.* Integran íconos con ciertas características.

Si vemos más a detalle cómo es mostrado un volumen dentro de la interfaz gráfica de Volume Manager, podemos ver que está integrado por varios elementos como en la última figura.

Cada uno de estos elementos se les llama objeto y tiene características propias, así como desempeña una labor en específico; sin embargo podemos decir que todos estos objetos parten de la idea o concepto inicial de los niveles de RAID que se revisaron anteriormente.

Para poder realizar un volumen o disco virtual es necesario tomar pedazos de varios discos para ser conjuntados lógicamente a través de un software dedicado a la administración de volúmenes, en este caso Volume Manager. Estos pedazos son reconocidos lógicamente por la aplicación.

La siguiente figura muestra la representación ó *layout*, de un volumen en RAID 0 (striped) a través de la interfaz gráfica de Volume Manager.

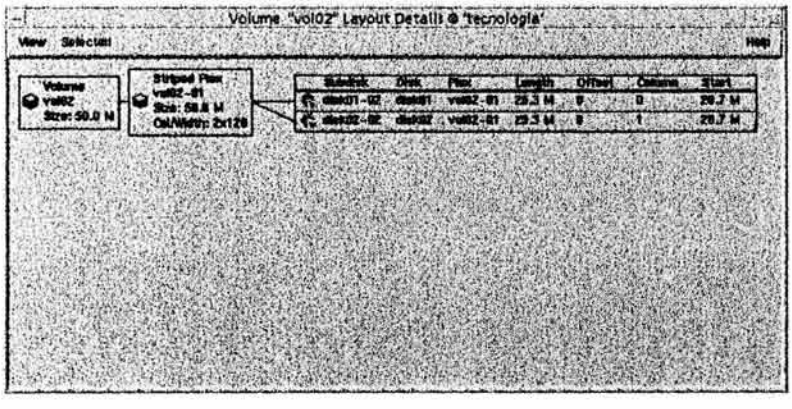


Figura 3.3.3 Vista del *layout* de un volumen en striping (RAID 0).

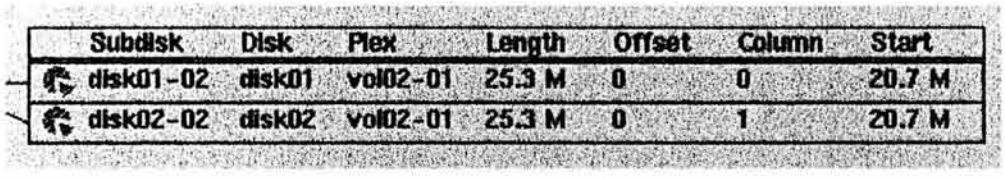
Esta figura nos servirá para hacer referencia a los objetos más importantes de Volume Manager que son los siguientes:

Subdisks (subdiscos).

Un subdisco es un conjunto de bloques contiguos de un disco. Un subdisco debe residir enteramente en un sólo disco físico.

También puede pensarse como la representación de esos pedazos de disco que son tomados para formar un volumen o configuración de RAID.

En la figura de referencia del volumen en striping podemos ver los subdiscos representados de la siguiente manera:



Subdisk	Disk	Plex	Length	Offset	Column	Start
disk01-02	disk01	vol02-01	25.3 M	0	0	20.7 M
disk02-02	disk02	vol02-01	25.3 M	0	1	20.7 M

Figura 3.3.4 Vista de los subdiscos de un volumen.

La figura anterior se encuentra recortada, pero podemos apreciar los subdiscos o pedazos de disco con que está formado este volumen.

El volumen está formado por dos subdiscos, requisito mínimo para formar un RAID 0 en striping, llamados disk01-02 y disk02-02. El nombre de los subdiscos puede ser alterado, pero Volume Manager les asigna un nombre por omisión formado

por el nombre del disco al que pertenecen (disk01 y disk01 en este caso), y un número consecutivo.[10]

El número consecutivo de los subdiscos se refiere al número de pedazo que fue tomado de ese disco físico, en nuestro ejemplo este volumen fue tomado de los segundos dos pedazos del disco disk01 y disk02, lo que nos hace pensar que los primeros pedazos de estos discos ya fueron ocupados previamente para formar otro volumen.

El objeto muestra otros datos que tienen que ver con su tamaño y la relación que guarda con otros objetos del volumen que revisaremos a continuación.

Plexes.

Volume Manager asocia los subdiscos dentro de otro objeto llamado *Plex*. Un plex puede contener varios subdiscos y un volumen puede contener varios plexes, hasta 256.

Así pues, tomando la figura anterior, podemos observar el plex al que estan asociados los subdiscos del volumen en stripping.

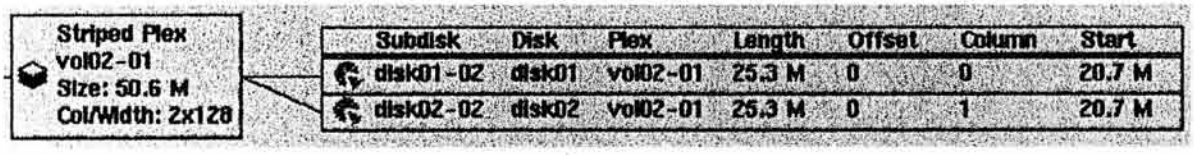


Figura 3.3.5 Vista del *plex* de un volumen.

Se dice que los plexes son copias de información almacenada por el usuario, de tal modo que cuando adicionamos un espejo para formar un RAID 1, lo que

[10] Cuando se renombra un volumen, no se renombran los subdiscos automáticamente.

estamos realizando es adicionar un plex más al volumen que ya tengamos realizado previamente.

La siguiente figura muestra el *layout*[11] de un volumen en espejo, notemos que está formado por más de un plex.

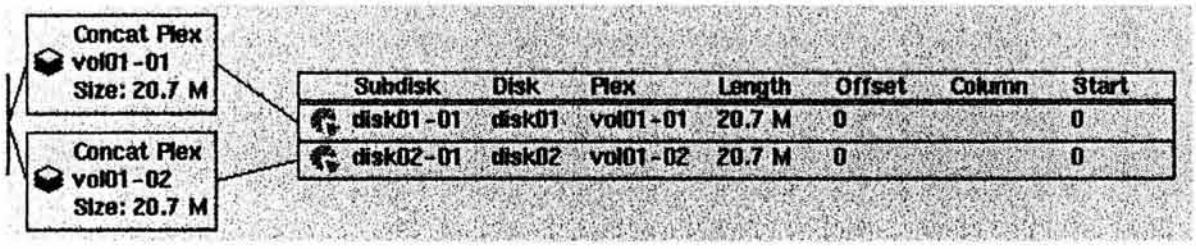


Figura 3.3.6 Vista de los *plexes* de un volumen en espejo.

Esta figura nos muestra un volumen RAID 0 concatenado, con un espejo también concatenado, es por eso que cada plex solamente tiene asociado un subdisco, de haber sido un volumen en espejo de tipo striping, cada plex tendría asociado varios subdiscos.

Volumen.

Un volumen consiste en uno o más plexes. Como mencionamos anteriormente, un volumen con varios plexes la mayoría de las veces significa que se trata de un volumen en espejo.

Algunas de las características más generales de los volúmenes con Volumen Manager son las siguientes:

- Podemos realizar volúmenes en RAID 0, 1, 1+0, 0+1 y 5.

[11] Por *layout* nos referimos a la distribución de elementos.

- Los volúmenes pueden tener más de dos mirrors (o más correctamente dicho, submirrors, ya que el mirror se le llama a la asociación lógica de varias copias de información o submirrors).
- Un volumen en RAID 5 no se le puede asociar un espejo.
- Un plex también puede ser el objeto correspondiente a una estructura llamada *log*, la cual no es utilizada para almacenar datos y aunque es representada gráficamente como un plex, no representa ni tiene la misma función que un espejo. El *log*, se explicará más adelante.

La siguiente gráfica muestra el *layout* completo del volumen en espejo visto en partes anteriormente.

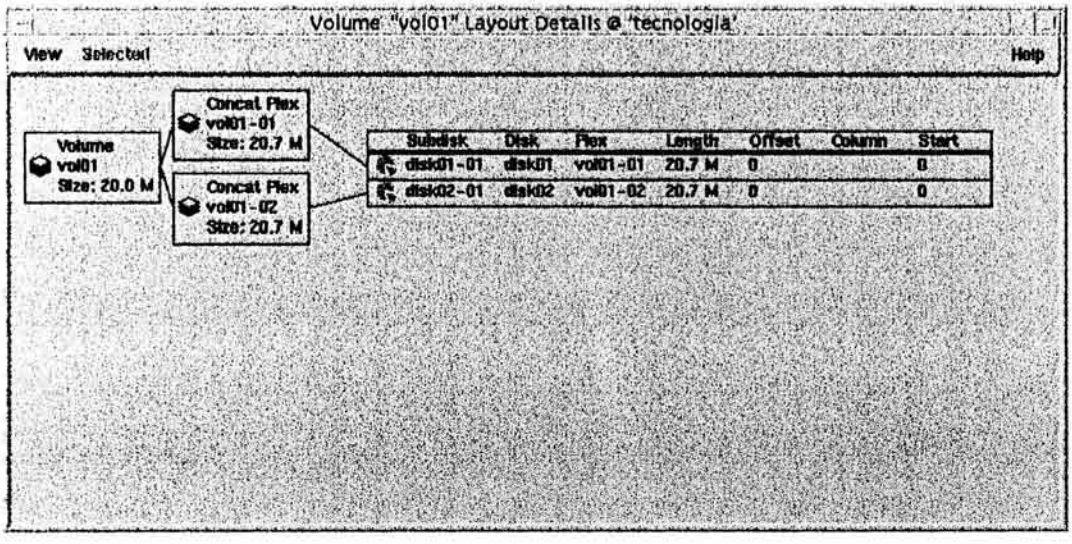


Figura 3.3.7 Vista de un volumen concatenado con un espejo concatenado también. (RAID 1).

Todos lo objetos de Volume Manager pueden ser vistos a través de la interfaz gráfica, pero también existen comandos que pueden mostrarnos la misma

información, y en algunos casos más completa, que pueden ser tecleados a través de la línea de comando de UNIX.[12]

La línea de comando no será cubierta en este trabajo.

3.4 OPERACIONES CON VOLÚMENES.

Los volúmenes son los objetos que se utilizan directamente en el sistema, es decir, son los objetos que almacenarán los datos, a los cuales les crearemos un file system[13] y que el sistema podrá utilizar a las necesidades.

La creación de los volúmenes a través de Volume Manager es una operación sumamente sencilla, no obstante es importante saber que muchos de los errores que obtenemos de Volume Manager al momento de realizar volúmenes, son resultado de la "protección" que provee la aplicación a realizar actividades que por su configuración o naturaleza representen un peligro en la integridad de los datos.

Así pues Volume Manager nos prevé de realizar un volumen con espejo, por ejemplo, utilizando un mismo disco físico. Esto es porque una configuración de espejo está diseñada para prevenirnos a la falla de un disco, por lo tanto, el realizar una configuración con pedazos de un sólo disco físico no tendría ningún sentido, en el momento en que ese disco físico llegue a fallar, ambas partes del mirror quedarían fuera del funcionamiento, esa configuración sería absolutamente absurda.

Así como el ejemplo anterior, también hay algunas otras configuraciones que Volume Manager nos prevee de realizar porque representan un absurdo o una situación peligrosa.[14]

---

[12] Estos comandos son instalados con Volumen Manager, no son propios del sistema operativo Solaris.
[13] Por *file system* entendemos la configuración que le da el sistema operativo a una partición para poderla utilizar con datos de usuario y de sistema.

A pesar de lo anterior, debemos tener en cuenta que podemos realizar configuraciones aún más robustas conociendo las limitaciones y características de Volume Manager; así por ejemplo debemos de tomar en cuenta que la creación de grupos de discos tiene una influencia importante en la creación de volúmenes y la disponibilidad de datos.

Hemos dicho anteriormente que los grupos de discos en Volume Manager nos sirven para administrar de una mejor manera la información que almacenaremos en nuestro sistema y que por omisión Volume Manager crea un grupo llamado rootdg.

No es recomendable dejar todos lo discos que administrará Volume Manager en este grupo ya que todos los discos y volúmenes son representados por Volume Manager como objetos y cada grupo tiene una limitante en cuanto al número de objetos que puede manejar (2048), por tanto, si alcanzamos ese límite podríamos tener dificultades para hacer alguna configuración en especial.

Otra razón para no dejar todos los discos de Volume Manager sobre el grupo rootdg es porque cuando un host[15] sufre una caída o algún daño, se puede delegar, o forzar a delegar sus permisos de acceso a la los datos administrados por Volume Manager a otro host que tenga instalado también Volume Manager y que le sean conectados directamente los discos o arreglos de discos que contienen la información. Estas operaciones se hacen a nivel de grupo de discos, es decir, se delega el acceso a la información de un cierto grupo de discos a un host, por lo tanto este grupo debe de tener una característica en especial: debe tener un nombre distinto a cualquier grupo existente en el nuevo host.

En general se recomienda lo siguiente con respecto a los grupos de discos:

---

[14] Por ejemplo, poner los dos submirror de un mirror en un sólo disco fisico, o hacer un striping con un sólo disco también.

[15] Por *host* nos referimos a un servidor.

- Todos los grupos de discos de todos nuestros sistemas deben tener nombres distintos
- Todos los grupos que residen en un sólo sistema deben tener nombres diferentes. De hecho es una restricción de Volume Mnager.
- Cada host debe tener un grupo llamado rootdg. Una restricción del sistema también.
- En general el grupo rootdg debe permanecer con pocos discos.
- Todos los grupos de discos deben contener al menos un disco.

Observando las recomendaciones y restricciones anteriores podemos crear los volúmenes que necesitemos.

Para crear un volumen podemos utilizar la interfaz gráfica o la línea de comando. Si se utiliza la primera se debe utilizar el botón de *NEW* de la barra de herramientas (*tool bar)* vista anteriormente.

A continuación se llena la forma siguiente dependiendo de las características del volumen que se desee crear.

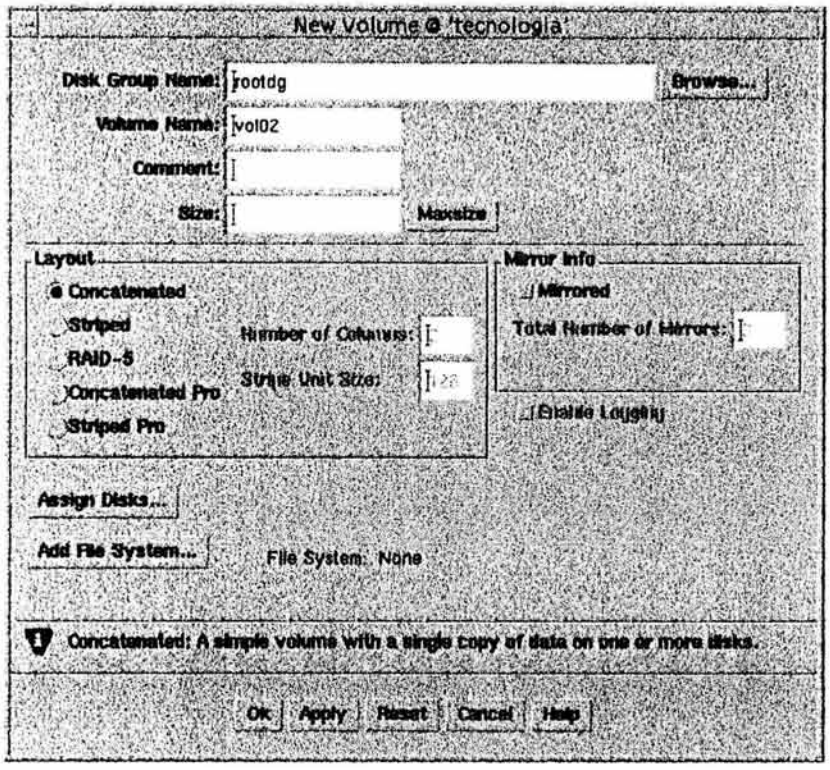


Figura 3.4.1 Pantalla de creación de un volumen.

A través de esta forma podemos realizar volúmenes de tipo concatenación, striping, RAID5 y RAID 0+1 o concatenacion con mirror. Adicionalmente a esto, podemos posteriormente un mirror a un volumen que no sea de tipo RAID 5.

Debemos integrar el grupo de discos al que pertenecerá el volumen, un comentario, el tamaño del volumen, el número de columnas o discos que formarán ese volumen, el tamaño del stripe unit (pedazo o chunk), adicionar uno o varios mirror, etc.

Una característica de Volume Manager es que puede tomar algunas decisiones por nosotros o dejar que nosotros le indiquemos cómo proceder. Así pues para la creación de un volumen podemos asignar los discos de los cuales Volume

Manager tomará los pedazos de disco para formarlo, o bien, puede decidir cuáles discos formarán ese volumen de acuerdo al espacio disponible en ellos.

Al momento de crear un volumen también podemos crearle un sistema de archivos (file system) e inclusive montarlo en el sistema para usarlo inmediatamente.

También podemos cear un volumen de un tamaño tal que ocupe todo el espacio disponible en los discos a través del botón *maxsize*.

En general se hacen las siguientes consideraciones en cuanto a la creación de nuevos volúmenes.

- Volume manager maneja nombres para los volúmenes por omisión. El nombre por omisión está formado por el nombre del grupo de discos al que pertenece el volumen, más un número consecutivo dependiendo del orden en que fueron creados los volúmenes.
- Este nombre por omisión puede no ser significativo en relación al uso o tipo de información que almacenará, por lo que en ocasiones será recomendable cambiar el nombre del volumen a una más representativo

- El utilizar el botón de *maxsize* para la creación de los volúmenes puede resultar en un volumen compuesto de varios pedazos pequeños o subdiscos que se encuentran dispersos entre varios discos físicos lo que podría generar un desempeño muy pobre en las escrituras y/o lecturas de datos en el volumen. Para mejorar el desempeño no solamente basta con establecer apropiadamente el tipo de RAID que se utilizará dependiendo de las aplicaciones a ejecutar; sino también en planear adecuadamente la distribución del volumen entre los discos e inclusive entre las controladoras del sistema.

- El campo de *number of columns*, solamente aplica a los volúmenes de tipo striping incluyendo RAID 5 y es equivalente al número de discos físico de los cuales se tomarán subdiscos del mismo tamaño para formar el volumen.
- Los botones de *Assign disks* y *Add File System* son útiles en caso de que se quieran asignar discos físicos específicos para formar el volumen y cuando se le quiera formar un file system al volumen, incluyendo la creación del punto de montaje, si éste no existe, y la edición de archivos de configuración del sistema necesarios para montar automáticamente el volumen al momento del inicio del sistema, entre otras cosas.

Otro punto importante en la creación de los volúmenes con Volume Manager es la posibilidad de asignar a los volúmenes de tipo mirror o RAID 5 un objeto llamado *log*.

Este log puede ser un *Dirty Region Logging* en caso de asociarse a un volumen de tipo mirror, o bien un *RAID 5 logging* en caso de tratarse de un volumen en RAID 5.

En general ambos log proveen la misma funcionalidad final aunque trabajan de diferente forma.

El *Dirty Region Logging* (DRL) es un archivo de log que reside en un subdisco extra a los que ya forman un volumen de tipo espejo o mirror.

Éste log registra los cambios que han sido realizados a un volumen de tipo espejo con la finalidad de proveer de más rapidez al momento de una falla en un volumen de este tipo ya que al momento de sincronizar la información a un submirror que sufrió un daño desde uno que esté sobreviviente, el DRL tendrá registrados los datos que habrá de sincronizar de tal modo que no toda la información de sincronizará, solamente la que haya cambiado.

Un DRL es un plex agregado a un volumen de tipo mirror que tiene las siguientes características:

- Es un log que registra las regiones del volumen que han cambiado debido a las escrituras sobre el mismo, manteniendo un mapa de bits que indica estos cambios; pero no se almacenan datos de usuario.
- Después de la falla del sistema, solamente las regiones marcadas como "dirty", o que sufrieron cambios que no fueron escritos en el volumen, serán sincronizadas o actualizadas.

Al no almacenar datos de usuario propiamente, el DRL es un subdisco muy pequeño, de aproximadamente 1 bloque por cada Gb de información. El DRL no excederá de 5 Kb, en caso de volúmenes muy grandes, Volume Manager reacomoda la información del log para almacenar el mapa de bits en un espacio no mayor a 5 Kb.

El *RAID 5 logging* previene la corrupción de datos que podría generarse cuando se presenta una falla y los datos, o la paridad, de un volumen en RAID 5 no han sido aún escritos al momento del fallo.

Cuando se utiliza un log de este tipo, tanto los datos como la paridad son escritos primero en el log y posteriormente en el volumen como tal.

Por omisión Volume Manager crea volúmenes en RAID 5 con un log que le provee de más seguridad en los datos al momento de una falla.

Para ambos tipos de log, lo que se recomienda es que estén alojados en discos físicos diferentes a los que conforman el volumen al cual pertenecen[16].

[16] Este es un requisito de Volume Manager.

3.5 OPERACIONES ESPECIALES CON VOLUME MANAGER

Además de la creación y borrado de volúmenes, existen varias operaciones que podemos realizar sobre los mismos a través de Volume Manager. Estas operaciones son las que realmente distinguen un administrador de discos de otro.

En general, lo que he denominado como "operaciones especiales", no es más que la administración de los volúmenes y las características de Volume Manager que lo hacen un software versátil por la flexibilidad con la que podemos administrar la información.

Las siguientes operaciones pueden ser realizadas con Volume Manager a través de su interfaz gráfica o a través de línea de comando en UNIX.

Evacuación de un disco.

Como es sabido todos los volúmenes que formamos con Volume Manager están compuestos de varios "pedazos" de discos diferentes que pertenecen a un mismo grupo. Estos discos pueden contener varios subdiscos de volúmenes diferentes, por o que una falla en un disco puede afectar a varios volúmenes.

Cuando un disco empieza a comportarse de manera anómala, es posible evacuar o cambiar todos los subdiscos que contiene a otro disco que esté libre y que pueda contener los subdiscos que estarían afectados por la falla en ese disco. A esta operación se le llama *evacuación.*

Para que la evacuación pueda llevarse a cabo se necesita lo siguiente:

- El disco destino debe pertenecer al mismo grupo

- El disco destino no debe contener objetos que puedan entrar en conflicto con los que albergará.
- El disco destino debe tener suficiente espacio libre para recibir los subdiscos del disco que empieza a fallar.

Para evacuar un disco se debe seleccionar el disco a evacuar e ir posteriormente al menú de *Disks* y seleccionar *Evacuate*.

Movimiento de grupos de discos (Deport / Import).

Es posible mover grupos enteros de discos entre hosts o servidores que tengan instalado Volume Manager y a los que sea posible conectar físicamente el arreglo de discos o discos que son administrados por el software.

Este tipo de operaciones se conocen como *Deport* e *Import*.

Un grupo de discos es asociado con un host en particular. Algunas veces el administrador pudiera necesitar asociar un grupo de discos con otro sistema. Cuando esto se realiza bajo el control del administrador, el proceso involucra la *deportación* del grupo del sistema original y posteriormente la *importación* del mismo en el nuevo host.

Las razones por las cuales pudiera ser necesario realizar una deportación son las siguientes:

- Deshabilitar el acceso a todos los volúmenes en ese grupo de discos.
- Preparar el grupo para ser importado por otro sistema.

Como hemos visto, en cada disco se guarda información importante que determina la pertenencia de un grupo a un sistema, esta información es

Falta página

N° 109

Debido a lo anterior, en el caso de rootdg, cuando se realiza la deportación o la importación, debe cambiarse el nombre del grupo para que éste pueda ser importado por un nuevo host,

Hot Spare/ Hot Relocation.

Existen dos variantes por las cuales se protege la información de nuestros volúmenes contra la falla de algún disco, independientemente de la redundancia provista por volúmenes en espejo o en RAID 5. Estas son, en primer lugar, la designación de *Hot spares*, y en segundo lugar un proceso llamado *Hot relocation*.

Un *Hot Spare* es un disco que el administrador designa para ser utilizado únicamente como reemplazo de algún otro que llegue a fallar[17]. Este disco debe tener las siguientes características:

- No debe contener ningún objeto, es decir, no debe formar ningún volumen.
- Debe pertenecer a un grupo de discos al cual será asignado, es decir, un hot spare no puede sustituir un disco contenido en otro grupo de discos.

Un host spare moverá todos los subdiscos u objetos contenidos en un disco hacia él cuando una falla ocurra.

Hot Relocation, también funciona en caso de una falla en los discos, solamente que en este caso, se mueven los objetos o subdiscos que se detecten como fallidos, es decir, solamente mueve parcialmente el contenido de un disco hacia él.

[17] En equipos que realizan RAID por hardware también existe esta configuración. El arreglo de Sun Microsystems StorEdge T3 es un ejemplo.

Hot relocation no es definido por el administrador, ya es una función de Volume Manager; sin embargo puede ser deshabilitado.

Hot Relocation no es posible si lo siguiente ocurre:

- Los subdiscos que pertenecen al disco que falló no son parte de un volumen redundante como un espejo o un RAID 5.
- No hay espacio suficiente disponible para mover los subdiscos contenidos en el disco que falló.

Definir un hot spare solamente involucra cambiar sus propiedades a través de la interfaz gráfica o la línea de comando.

Snapshot

Cuando se necesita hacer un respaldo de la información de un volumen que esté en producción, se pude utilizar la función de Snapshot de Volume Manager para crear una copia del volumen a respaldar . Después con la nueva copia, se puede respaldar a cinta la información contenida sin necesidad de detener el servicio[18].

Los requisitos para realizar un Snapshot son los siguientes:

- Se debe conocer el nombre del volumen a respaldar.
- Se debe escoger un nombre apropiado para el nuevo volumen copia o snapshot.
- Se debe designar un disco para ser usado por la copia o snapshot.
- Se debe tener suficiente espacio sin utilizar para realizar el snapshot.

[18] Un *snapshot* es un mirror de tipo temporal de un volumen.

Un snapshot no es más que un espejo especial que contiene toda la información de un volumen. Este espejo se separa y se trata como un volumen aparte el cual puede ser respaldado a cinta.

El único objetivo del snapshot es realizar un respaldo de información, por lo que una vez terminada esta operación, el snapshot puede ser eliminado.

Volume Relayout.

A través de volume Manger también se puede cambiar en línea la estructura o distribución de un volumen, es decir, es posible cambiar un volumen de tipo striping a concatenation, por ejemplo.

Algunos de los requisitos para realizar este tipo de operaciones son :

- Escoger la nueva estructura o nivel de RAID (concatenated, striped, RAID 5, etc)
- Especificar espacio adicional a ser usado por la nueva estructura del volumen.

Cabe mencionar que estas operaciones pueden realizarse en línea, es decir, no es necesario suspender el servicio para poder realizarlas.

Resize

Otra de las operaciones que pueden realizarse sobre los volúmenes de Volume Manager, es la alteración en el tamaño de un volumen, ya sea para hacerlo más grande y para disminuir su tamaño.

Cuando un volumen ya tiene integrado un file system, éste puede ser alterado en su tamaño para crecerlo, pero no siempre para disminuirlo ya

que sufriría daños en los datos que almacena, solamente cuando se utiliza Veritas File System puede realizarse ambas operaciones. Todas las operaciones pueden ser realizadas en línea, es decir, sin suspender el servicio.

Este tipo de operaciones son esenciales en un software de este tipo, ya que es frecuente el necesitar crecer un volumen debido a su uso.

En el siguiente capítulo se revisará un ejemplo de la utilización de RAID en un servidor de correo electrónico con la ayuda de Volume Manager, aunque puede utilizarse cualquier otro producto de administración de discos.

CAPITULO 4. CASO PRÁCTICO: RAID 0+1 EN UN SERVIDOR DE CORREO ELECTRÓNICO

4.1 ANTECEDENTES

Uno de los principales servicios en RedUNAM ha sido por mucho tiempo el correo electrónico que han usado principalmente investigadores y académicos de la UNAM y que posteriormente fue extendido hacia la comunidad universitaria.

Este servicio estuvo desde sus inicios montado sobre plataforma UNIX, de hecho, prácticamente desde su creación, sobre Solaris de Sun Microsystems.

El esquema de servicio fue por mucho tiempo simple, es decir, los usuarios se conectaban al servidor como usuarios de UNIX y utilizaban las herramientas disponibles para mandar y recibir sus correos. Esto no incluía ninguna herramienta gráfica, solamente línea de comando o programas que permitían interactúar con el servicio a través de menús.

En sus inicios el servicio no era muy demandado debido a que la población universitaria que hacía uso de estos recursos no era extensa, pero con el auge de los servicios de conexión a Internet comercializados por varias empresas particulares, e inclusive por la UNAM, el servicio empezó a crecer.

Uno de los principales problemas de cualquier servicio de correo electrónico, y sobre todo de aquellos en los que se tiene comunicación directa con el usuario y existe un nivel de servicio específico, es el crecimiento en almacenamiento que representan los buzones de cada usuario.

En el caso del servidor de correo electrónico de RedUNAM (el cual ha sido alojado en diversos equipos de cómputo a través de los años), los buzones de correo de cada usuario no tienen un límite estricto de tamaño o *quota*, como suele llamarse en términos computacionales. Esto ha representado un problema constante

debido a que los usuarios pueden recibir todo el correo que deseen, máxime si se trata de investigadores o funcionarios que por lo regular reciben muchos mensajes y están suscritos a varias listas de discusión, esto ha orillado a los administradores a crear programas que mueven los mensajes hacia un lugar con mayor espacio dentro del equipo cuando se alcanza un cierto tamaño de buzón[1].

Este problema involucra que los recursos de disco destinados en un inicio para almacenar el correo de los usuarios, resulte insuficiente al cabo de poco tiempo, o de una manera exponencial si el crecimiento de usuarios no es controlado. Además de que el servicio no puede ser interrumpido frecuentemente para hacer labores administrativas que corrijan la falta de espacio.

Las empresas dedicadas a otorgar este tipo de servicios de manera gratuita cuentan con mecanismos más sofisticados que permiten absorber este crecimiento sin problemas, por ejemplo el uso de varios servidores de correo, redes de almacenamiento, etc; además de que los niveles de servicio con el cliente son muy limitados, es decir, no hay garantía al usuario de que sus mensajes no serán perdidos o borrados.

Este tipo de mecanismos no pueden ser puestos en marcha en la UNAM debido a los recursos con que cuenta la institución y al enfoque no comercial que se le ha dado al servicio por muchos años. Sin embargo, lo único que ha podido hacerse es planear cada vez de una manera más eficiente la manera en que son configurados los servidores y los servicios, y es aquí donde las técnicas RAID juegan un papel sumamente importante , sino vital, en el servicio.

Para entender los beneficios de la configuración RAID que se hizo sobre el servidor de correo electrónico, es necesario recordar los siguientes aspectos técnicos y funcionales del servicio:

---

[1] El programa revisa que cada buzón sea menor o igual a 3 Mb de espacio, si esto no se cumple se manda un aviso al usuario para que depure su correo, de lo contrario sus correo son movidos a otro lugar.

1. El servicio de correo electrónico está montado sobre equipos de Sun Microsystems, utilizando su sistema operativo nativo: Solaris.
2. El almacenamiento de buzones está montado también sobre equipo de Sun Microsystems, el más novedoso hasta hace dos años : Sun StorEdge A5200.
3. El servicio de correo electrónico es provisto a través de *sendmail* extraído y compilado con herramientas libres de *GNU* de acuerdo a las necesidades de la Universidad.[2]
4. *Sendmail* por omisión utiliza agentes de correo internos que almacenan los mensajes sobre el directorio (o partición) */var/mail* del sistema.[3]
5. Se tiene alrededor de 23,000 usuarios de correo electrónico, dados de alta como usuarios de sistema con privilegios muy limitados.[4]

4.2 ELEMENTOS DE LA CONFIGURACIÓN.

Los siguientes elementos forman parte de la configuración del servicio de correo electrónico:

Hardware

- Ultra Enterprise 3500 de Sun Microsystems
  - 4 procesadores Ultra SPARC II a 450 MHz
  - 2 GB de memoria RAM
  - 2 discos de almacenamiento interno
  - 1 tarjeta adaptadora (Host Adapter) para fibra con 2 GBICs
- Sun StorEdge A5200 de Sun Microsystems

---

[2] Sendmail fue compilado con características de "anti-SPAM".
[3] Es posible utilizar agentes de dominio público como *qmail* para almacenar los correos en un lugar distinto.
[4] La mayoría de los comandos del sistema operativo fueron modificados en sus premisos para que no fueran ejecutados por la mayoría de los usuarios.

- 14 discos FC-AL (Fibre Channel Arbitrated Loop) de 9GB
- 2 tarjetas (Interface Board) con 2 GBICs cada una

Software

- Solaris 7
- Veritas File System
- Veritas Volume Manager
- Sendmail 8.10
- Pop3
- Herramientas de seguridad (TCP-Wrappers, SSH, Portsentry, etc.)

A continuación se describe de manera general algunas características de estos elementos y el porqué fueron elegidos para formar parte de la configuración.

Ultra Enterprise 3500 – Este equipo fue adquirido por la Dirección General de Servicios de Cómputo Académico (DGSCA) en el año de 1998 para sustituir el equipo anterior en donde se encontraba alojado el servicio de correo electrónico.

Este equipo está clasificado por Sun Microsystems como un servidor "Midrange" porque sus características lo hacen ideal para montar servicios que necesitan de alta disponibilidad como lo son Bases de Datos o ERP (*Enterprise Resource Planning)* para empresas grandes.

El equipo es óptimo para:

- Servicios de Internet/Intranet.
- Bussiness Intelligence.

- Enterprise Resource Planning (ERP).
- Gráficos.
- Aplicaciones de Negocio.
- Bases de Datos.

Sun StorEdge A5200 - Mucho se ha hablado acerca de estos equipos por lo que no se redundará más sobre sus características, solamente cabe indicar que el equipo fue adquirido por la misma Dirección junto con el equipo anterior para brindar el almacenamiento requerido por los servicios alojados en la Enterprise 3500.

El utilizar arreglos de disco de este tipo facilitó, por sus características, la planeación y reestructuración de la información contenida en los viejos servidores que prestaban el servicio, no solamente de correo electrónico, sino también de web y bases de datos. Fue posible la reinstalación de los servidores y una reestructuración que permitiera absorber el crecimiento del servicio.

Solaris 7 - La única razón, por la cual se utilizó este sistema operativo, es porque los equipos arriba mencionados solamente pueden trabajar con él, además de que al ser un sistema operativo comercial, el soporte es amplio y su dominio por parte del administrador es completo.

Veritas Filesystem - La utilización de los productos de Veritas (Veritas Filesystem y Veritas Volume Manager), es principalmente por su robustez, desempeño y confiabilidad, aún cuando se trate de productos

de terceros, es decir, estos productos no son de Sun Microystems, pero son tan populares en el mercado que el mismo Sun los comercializa y recomienda a sus clientes. Veritas y Sun son socios de negocio.

La adquisición de los productos de Veritas para los servidores de RedUNAM respondió a la necesidad de contar con herramientas más robustas que permitieran tener mayor disponibilidad del servicio y confiabilidad de que la información en caso de un fallo sea más fácilmente recuperable, además de reducir tareas de administración.

El administrador puede utilizar el filesystem nativo de Sun Microsystems (UFS), o utilizar Veritas Filesystem para disponer de mayores ventajas.

Veritas Volume Manager –Mucho se ha hablado acerca de las características de Volume Manager y sus ventajas por lo que lo único que se mencionará aquí es la razón por la cual fue adquirido este software en DGSCA.

En el año de 1998, DGSCA adquirió el equipo más grande con que Sun Microsystem contaba, la Enterprise 10000. Este equipo sería utilizado para las necesidades de la Universidad además de estar disponible para los cursos que se imparten en esa Dirección desde el año 1999 para clientes en Latinoamérica de Sun que también adquirieron este equipo.

Además de los cursos de Enterprise 10000, se impartieron cursos de Volume Manager, por lo que el software llegó a la institución a través de todo este convenio entre Sun y la

UNAM, quedando también disponible para los servidores de correo y web que ya se tenían en DGSCA.

Sendmail, pop y - herramientas de seguridad

La mayoría de las herramientas y aplicaciones que se tienen corriendo en los servidores de RedUNAM, forman parte de lo que se llama software libre o no comercial, debido a los recursos con que cuenta la institución. El caso de sendmail y las demás herramientas no hacen la excepción, por varias generaciones se han utilizado compilándolas e instalándolas de acuerdo a las necesidades imperantes del momento.

4.3 CONFIGURACIÓN DE HARDWARE.

Los elementos de hardware arriba mencionados (Enterprise 3500 y A5200), trabajan juntos a través de la configuración ilustrada en la figura 4.1.

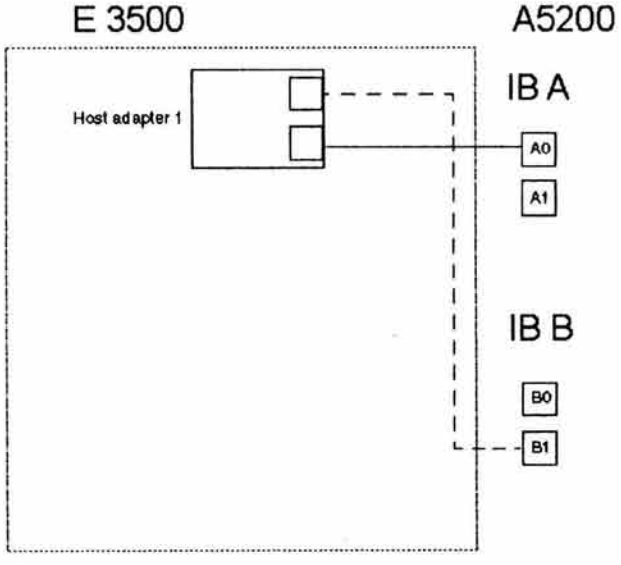


Figura 4.3.1 Diagrama de conexión de elementos.

En la figura 4.1 se muestra la conexión que existe entre el arreglo A5200 y la Enterprise 3500 a manera de diagrama de bloques. Dentro de la E3500 se cuenta con una tarjeta adaptadora (Host Adapter) con dos GBICS que se conectan al A5200 a través de sus GBICS.

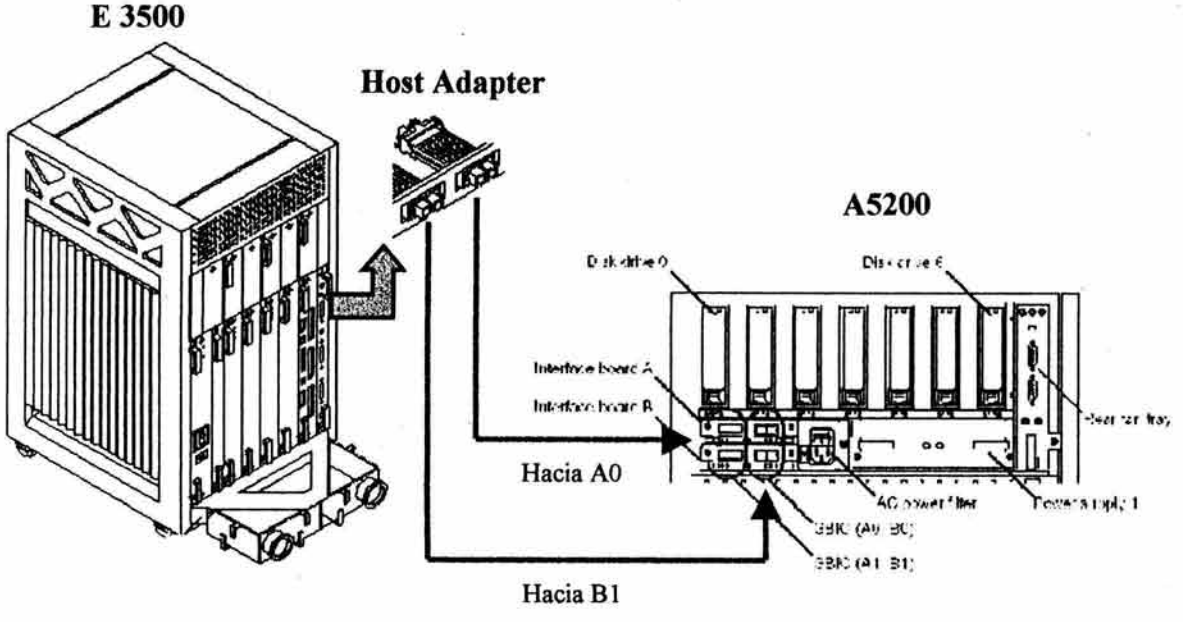


Figura 4.3.2 Conexión de GBICS.

La figura 4.2 muestra en el sitio en que se encuentran la tarjeta adpatadora y los GBICS para ser conectados con los del A5200.

Las características de esta configuración son las siguientes:

1. El E3500 tiene una tarjeta adaptadora con 2 GBICs (Gigabit Interface Converter) que le permiten acceder al arreglo desde 2 rutas físicas distintas (aunque parten de la misma tarjeta adaptadora). Los datos son mandados a través de las fibras ópticas que conectan la tarjeta del E3500 a las tarjetas

del arreglo. Ambas rutas son utilizadas, de manera que se hace un balanceo de cargas y en caso de tener un error en alguna de ellas, toda la comunicación se realiza a través de la que esté sin falla.

Los GBICs son los encargados de convertir las señales eléctricas que produce el servidor a señales ópticas que viajan a través de las fibras hacia el arreglo.

2. EL arreglo A5200, cuenta con 2 tarjetas de interfase (*Interface Boards)*, la cuales pueden alojar hasta 2 GBICS cada una. Cada GBIC representa un puerto de la tarjeta, por lo que los datos pueden ser accedidos por un total de 4 puertos. En el caso de nuestra configuración, el arreglo cuenta con 2 interface boards con solamente un GBIC cada una, de tal modo que las dos fibras que provienen del E3500 se conectan en cada GBIC de estas dos tarjetas.

   Ésta configuración es muy popular por ser de las más simples y económicas, y si bien no provee una redundancia total, sí nos protege de una falla en una ruta física, ya sea por un error en una fibra, en un GBIC o en una tarjeta del arreglo.

   Los datos viajan del E3500 hacia los discos del arreglo accediéndolos a través de dos rutas distintas; así mismo cada disco dentro del arreglo cuenta con dos rutas de acceso, A y B, siendo cada una de éstas atendida por una Interface Board A o B, según sea el caso.

3. El arreglo está configurado como *Full Loop*, de tal modo que el servidor puede tener acceso a todos los discos de arreglo, además de acceder a cada uno de ellos a través de dos rutas físicas distintas, producto de las conexiones arriba mencionadas.

4.4 CONFIGURACIÓN DE SOFTWARE.

La configuración de hardware es esencial para proveer soporte a las configuraciones del software, ya que al tratarse de niveles de RAID a través de

software, la parte más interesante y compleja de la configuración del servidor no radica en el hardware. Podemos hacer una configuración tan robusta como queramos con el software, pero si no planeamos a la par nuestras conexiones de hardware, de nada nos servirán las complejidades y ventajas que nos brinda la aplicación.

Ésta es la parte medular del capítulo, pues describe la forma en que está configurado el software utilizando un nivel de RAID para proporcionar ventajas en cuanto a la disponibilidad y confiabilidad de los datos y el servicio.

Particionamiento.

El servidor de correo, alojado en el E3500, tiene la siguiente distribución:

Partición	Tamaño (Mb) *	Recurso	Tipo	Utilidad general
/	500	Disco interno	Sistema operativo/ Metadispositivo de DiskSuite	Raíz del sistema operativo, almacena archivos de configuración y de inicialización del servidor
/usr	2000	Disco interno	Sistema operativo/ Metadispositivo de DiskSuite	Almacena todos los programas y comandos del sistema operativo y de algunas aplicaciones.
/var	1000	Disco interno	Sistema operativo/ Metadispositivo de DiskSuite	Almacena todas las bitácoras del sistema y algunos archivos temporales.
/opt	1000	Disco interno	Sistema operativo/ Metadispositivo de DiskSuite	Almacena archivos ejecutables de algunas aplicaciones que son

				instaladas como "paquetes" al servidor.
/home	2000	Disco interno	Directorios de usuarios/ Metadispositivo de DiskSuite	Almacena archivos de usuarios-adminsitradores del servidor.
/tmp	5000	Disco interno	Sistema operativo	Almacena archivos temporales además de ser utilizado como "swap" (memoria virtual) del sistema.
/var/mail	60000	A5200	Sistema operativo/ Volúmen de VxVM	Almacena todos los archivos-buzones de los usuarios del sistema que contienen el correo electrónico.
/var/spool	2500	A5200	Sistema operativo/ Volúmen de VxVM	Almacena archivos relacionados con la cola de distribución del correo electrónico.
/home/users00 - /home/users19	50000	A5200	Directorios de usuarios/ Volúmen de VxVM	Almacena toda la información de los usuarios. Almacena todos los directorios de trabajo de los usuarios.
/home/log	20000	A5200	Sistema de archivos para uso interno del administrador	Almacena bitácoras y resspaldos de la información de los usuarios.
* Tamaños aproximados				

Tabla 4.4.1 Particionamiento del sistema.

De estos filesystems o particiones, solamente algunos están alojados directamente en el disco interno del E3500, los demás están montados en volúmenes realizados con Volume Manager.

Todos los filesystems contenidos en el arreglo, son manejados a través de Volume Manager, por lo que representan discos virtuales donde está contenida la información. La principal razón de esto, es porque esos filesystems tienden a crecer desmesuradamente, en especial /var/mail, por lo que es necesario contar con la ventaja de poder crecerlos de manera dinámica, sin dejar de dar servicio y sin necesidad de destruir y volver a crear toda la configuración. Estas características no podrían tenerse sin la utilización de Volume Manager u otro producto semejante de administración de discos.

Volume Manager.

Dentro de Volume Manager, estos filesystems tienen la configuración RAID 0+1 porque se hizo de cada uno un espejo de dos partes, en donde cada parte está formada por un volumen en RAID 0 en striping.

Para ejemplificar esto, tomemos la partición más importante de nuestro servidor, /var/mail.

/var/mail es un filesystem de 60 GB (aunque los discos del arreglo son cada uno de 18 GB), por lo que para formarlo se tomaron un pedazo de cada uno de los 7 discos de la parte frontal del A5200, estos pedazos no son todos del mismo tamaño. La siguiente figura ejemplifica esto.

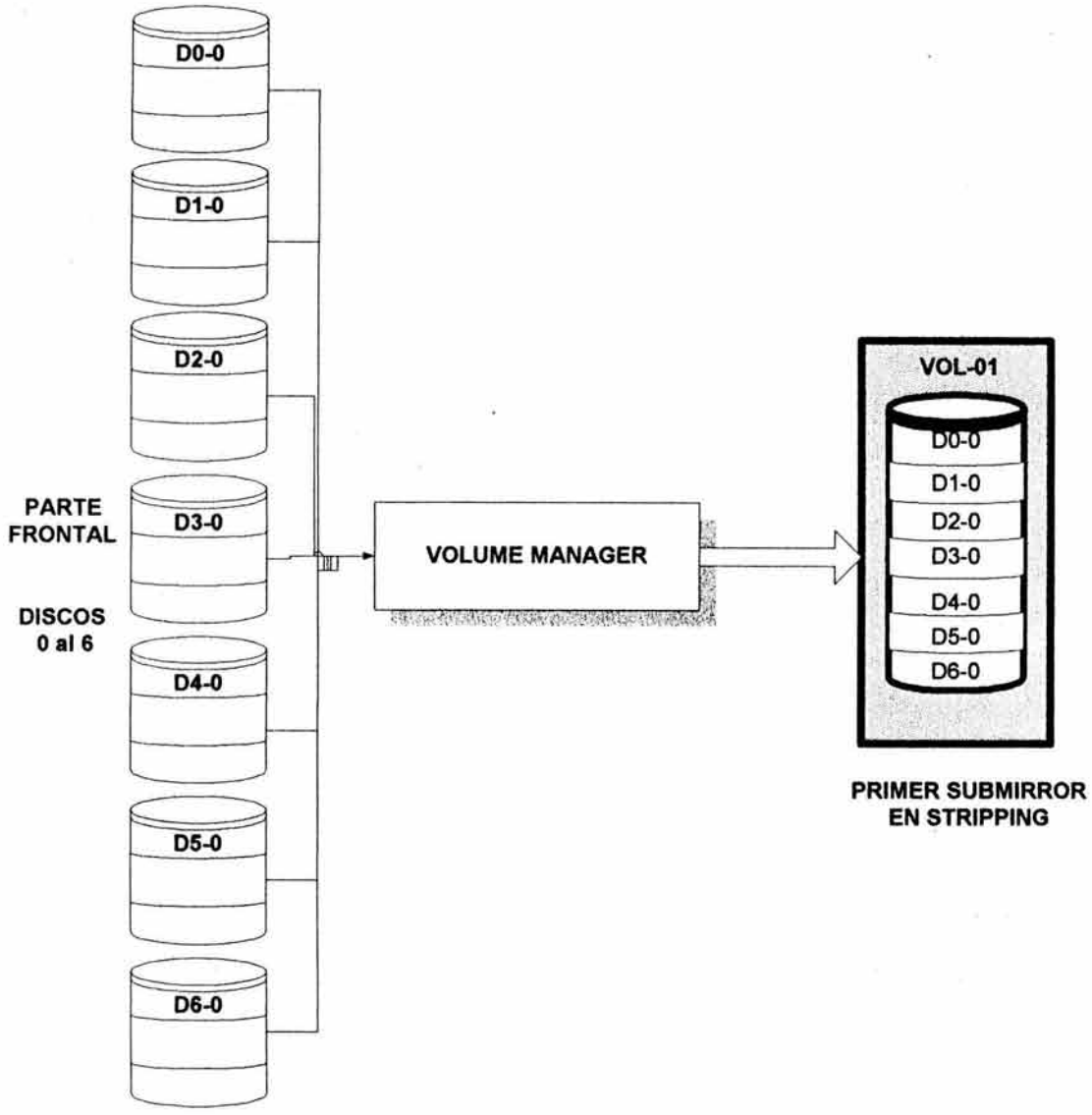


Figura 4.4.1 Primer submirror de /var/mail

El volumen resultante tiene una configuración de RAID 0 en striping, por lo que ofrece todas las ventajas de este nivel de RAID, pero no provee redundancia en los datos, por lo que a este volumen se le adicionó otro, de la misma configuración, para que ambos formaran un espejo (un espejo de dos partes).

La siguiente figura ejemplifica la configuración.

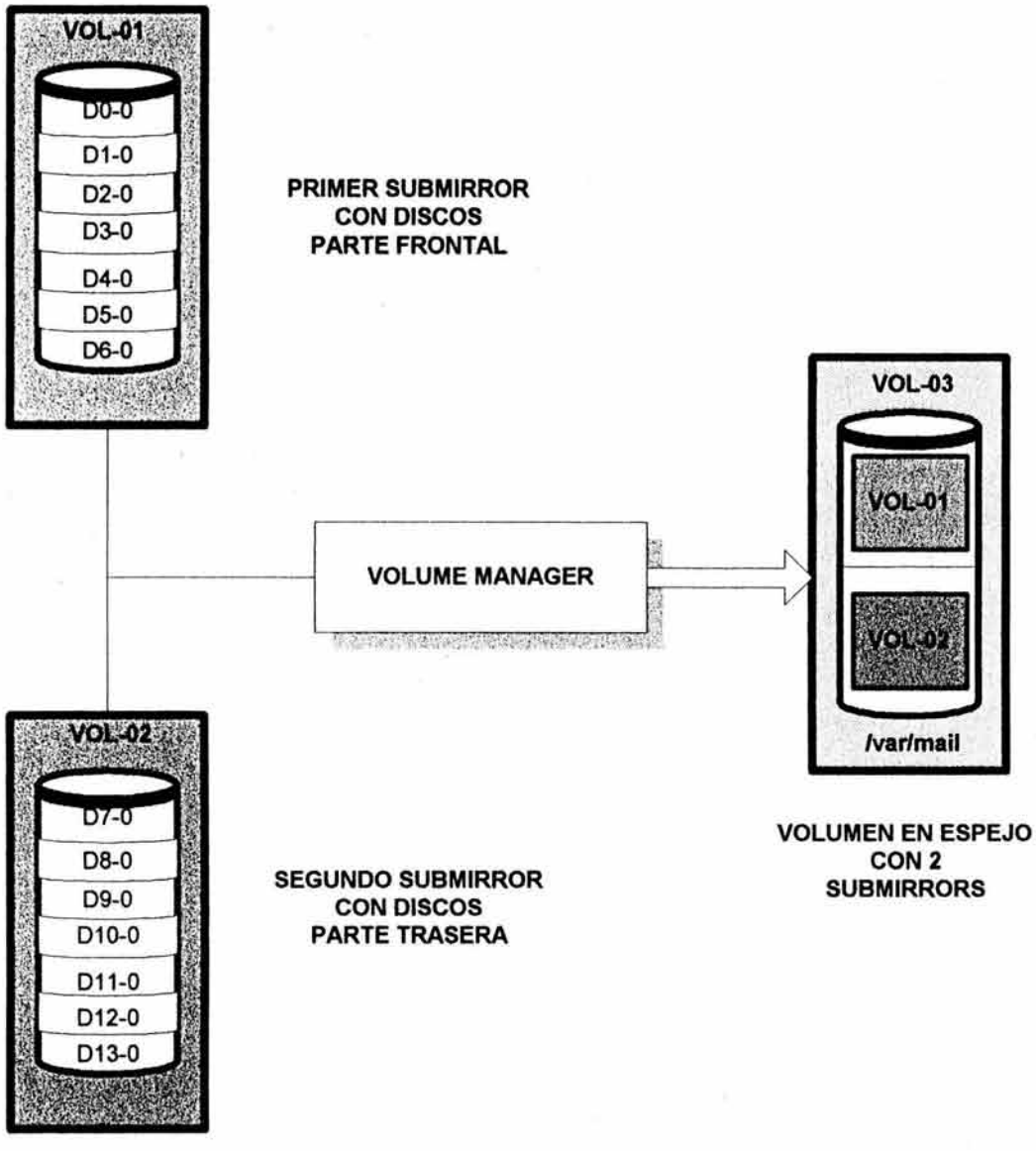


Figura 4.4.2 Volumen en espejo de /var/mail

El sistema hace referencia al espejo, no a los volúmenes que lo forman, de hecho, Volume Manager, es el encargado de administrar estos Volúmenes y saber dónde está la información, el administrador, el sistema y las aplicaciones hacen

referencia al espejo en cualquier momento. El dispositivo físico a que hace referencia /var/mail, es precisamente al volumen en espejo de Volume Manager.

4.5 VENTAJAS DE LA CONFIGURACIÓN.

Las ventajas que ofrece esta configuración son las siguientes:

Redundancia.

La principal ventaja de utilizar RAID 0 + 1 en esta configuración es el aprovechamiento del desempeño brindado por el striping, el cual es el mejor entre todos los niveles de RAID, y la redundancia provista por el espejo.

Los filesystems como /var/mail se ven beneficiados al tratarse de sistemas de archivos muy accesados (cada usuario accede a su buzón varias veces al día) y con la información más importante del servidor, ya que es aquí en donde radican los mensajes de todos los usuarios (razón principal del servidor de correo electrónico).

En caso de que alguno de los discos que conforman esta configuración llegara a fallar, el filesystem no se ve afectado, debido a que cada una de las partes del espejo tiene la misma información y los usuarios pueden tener acceso inclusive cuando el disco fallido esté siendo reemplazado.

Desempeño.

Como se ha mencionado, la configuración aprovecha las ventajas del striping, que posee el mejor desempeño en cuanto a escrituras se refiere, de entre todos los niveles de RAID.

La ventaja radica en que, en situaciones ideales, la escritura de los datos puede impactar a los 7 discos simultáneamente, realizando más rápidamente la escritura de los datos en comparación a realizar esto mismo con un sólo disco.

Para ejemplificar esto, podemos decir lo siguiente. Supongamos que tenemos un disco de 35GB que es utilizado como /var/mail. Si se realizara una escritura de un tamaño x, en donde x es igual al *stripe unit* (véase el capítulo 3) de nuestra configuración en striping mostrada anteriormente para el servidor de correo, la escritura de esos datos tardaría más por tratarse de un sólo disco que tiene que mover sus cabezas para realizar las escrituras; en cambio con una configuración en striping, la escritura impactaría a los 7 discos al mismo tiempo, por lo que el escribirlos le llevaría al sistema 7 veces menos el tiempo que tardaría con un sólo disco.

Crecimiento controlado.

Al decir crecimiento controlado, no me refiero a limitar el crecimiento, sino simplemente al hecho de que el crecimiento en cuanto a información contenida en los volúmenes de Volume Manager puede ser administrada de una manera más organizada y que brinda confiabilidad al administrador debido a que se puede adicionar "en caliente" más espacio de disco a los volúmenes cuando éstos hayan consumido todo su espacio disponible.

Habíamos mencionado que uno de los principales problemas de un servidor de correo es el espacio que cad usuario consume son su buzón, y que muchas veces no importaba qué tanto espacio se dejaba disponible para esto, simplemente el espacio será consumido.

Con este tipo de configuración puede crecerse el tamaño de los filesystems que ya no tengan espacio disponible, sin necesidad de dar de baja el equipo, y por consiguiente dejar al usuario sin servicio, y mucho menos involucrando tareas de rediseño de las particiones y volúmenes del servidor para dejar más espacio a esos filesystems que con facilidad se llenan.

La única limitante de esto es el espacio disponible en el arreglo, si ya no contamos con más, es necesario conectar y configurar un arreglo extra con espacio disponible.

Disponibilidad.

Como resultado de todo lo anterior y sumado a las características del arreglo, la disponibilidad de los datos y del servicio se ve ampliamente incrementada, debido a que los datos son redundantes y a que se cuentan con rutas físicas redundantes a través de fibra óptica y en caso de una falla en un disco o en alguno de los componentes del arreglo, éstos pueden ser reemplazados "en caliente" (hot swap) sin necesidad de dar de baja el equipo o tener tiempos muertos para labores de mantenimiento.

En suma, tanto el equipo como el software y su configuración, hacen que el servicio cuente con mayor disponibilidad y confiabilidad.

4.6 RECOMENDACIONES.

4.6.1. CONECTIVIDAD CON EL A5200

En general cualquier configuración puede tener puntos de falla o mejoras que pueden ser implantadas para proveer a la configuración de mayor disponibilidad y/o redundancia.

Uno de los puntos débiles de esta configuración es el uso de un sólo Host Adapter en el servidor, si en algún momento esta tarjeta sufriera algún daño que le impidiera tener comunicación con el arreglo de discos, no tendríamos acceso a la información.

Una configuración mucho más completa, y más costosa por supuesto, es la que se muestra en la siguiente figura.

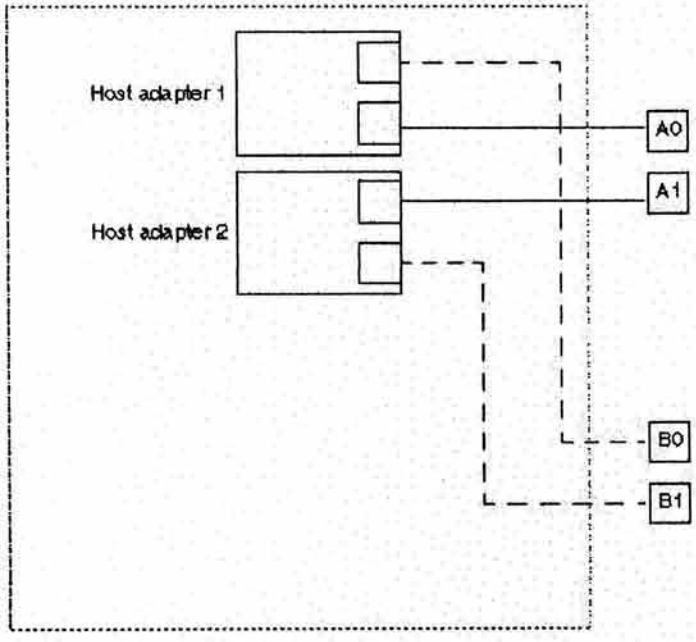


Figura 4.6.1 Configuración redundante.

En la configuración de la figura 4.6 se cuenta con otra tarjeta Host Adapter para proveer de mayor redundancia al acceso a los datos por rutas distintas.

Si una tarjeta llegara a fallar, es posible acceder a los discos desde la otra para ambas Interface Boards del A5200, además de que todos los puertos de las IB-A e IB-B son utilizados.

4.6.2. CONFIGURACIÓN DE /VAR/MAIL

Si bien /var/mail está configurado con un adecuado nivel de RAID, este filesystem crecerá hasta donde los recursos de hardware le permitan y en algún tiempo no será suficiente el espacio que se planeó en un inicio /var/mail utilizaría.

En situaciones como estas la solución a largo plazo es la utilización de una NAS o una SAN, dependiendo de las necesidades particulares del servicio.

4.7 Otros equipos para la configuración.

Los equipos utilizados en la configuración mostrada del servidor de correo electrónico de RedUNAM, ya no son comercializados por Sun Microsystems, ya que han sido sustituidos por otros con mejores características.

Por lo anterior, se presentan los equipos que pueden sustituir a los mostrados aquí para esta configuración, ya que son equivalentes en cuanto a las características y las aplicaciones para lo que son utilizados.

Los equipos expuestos en esta parte son los siguientes:

1. Sun FIRE 3800 Server.

Este equipo es un servidor de la nueva generación de servidores de Sun Microsystems que integra procesadores Ultra Sparc III. Este equipo es el que sustituye al E3500 de la configuración del servidor de RedUNAM.

La SunFire 3800 también está enfocada a .

- Servicios de Internet/Intranet
- Bussiness Intelligence
- Enterprise Resource Planning (ERP)
- Servicios de email
- Web Hosting
- Aplicaciones de Negocios

Las capacidades máximas de este equipo son:

- 2 Dynamic System Domains
- Dynamic Reconfiguration
- Sun Fireplane Interconnect

- Hasta 8 procesadores UltraSPARC® III Cu.
- Hasta 64 GB de memoria.
- Hasta 2 tarjetas Uniboard CPU/Memory.
- Hasta 12 slots cPCI hot-swappable.

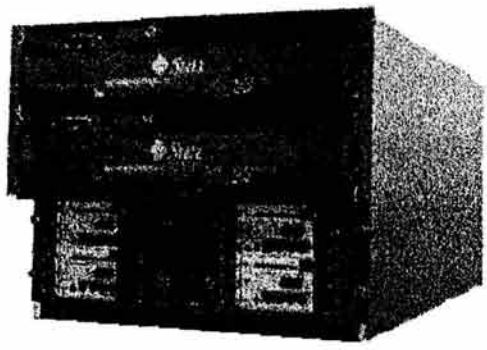


Figura 4.7.1 Sun Fire 3800

2. Sun StorEdge T3

El arreglo T3 es un dispositivo de Fibre Channel que utiliza cobre como medio, es controlado a través de una interfaz de red y una consola de administración. Es posible formar una SAN con varios de estos arreglos.

Algunas de sus características son las siguientes:

- Es un almacenamiento económico, potente y escalable.
- Es multiplataforma.
- Realiza RAID por hardware.
- El software de administración para estos equipos permiten su manejo, configuración y monitoreo a través de un puerto Ethernet que se configura separado de la red de datos de usuario.
- Posee baterías redundantes.

La capacidad máxima de un sólo arreglo T3 son de 660 GB ó 9 discos de 73 GB; sin embargo, como se mencionó, varios de estos arreglos son combinados para formar una SAN y proveer de mayor espacio de almacenamiento.

Figura 4.7.2 Sun StorEdge T3

En el siguiente capítulo se revisarán algunos conceptos generales de SAN y sus beneficios debido a que la tendencia, en cuanto a crecimiento en el almacenamiento se refiere, son las redes de almacenamiento.

CAPÍTULO 5. INTRODUCCIÓN A SAN (STORAGE AREA NETWORK)

5.1 CONCEPTOS

Este capítulo es una introducción a los conceptos generales de SAN debido a que a lo largo de este trabajo se le han hecho referencias y porque su utilización va adquiriendo más fuerza entre empresas que requieren de grandes capacidades de almacenamiento cada vez más crecientes.

Los problemas relacionados con el almacenamiento tienen que ver con el constante crecimiento de los datos, el gasto que involucra el crecer a su vez los dispositivos de almacenamiento conforme van creciendo e inclusive para prevenir un crecimiento futuro, organizar grandes cantidades de datos y en lo posible tenerlos en línea o disponibles y las implicaciones que trae al administrador las tareas de respaldo y mantenimiento.

Para proveer a un sistema de la capacidad de almacenamiento que requiere, podemos conectar discos o arreglos de discos hasta el límite que el mismo equipo nos permita, a este esquema de conexión se le llama *Almacenamiento Directamente Conectado (DAS – Directly Attached Storage)*; sin embargo este modo de utilizar los recursos de almacenamiento presenta varias limitaciones:

- Cada servidor controla el almacenamiento que tiene conectado directamente.
- La cantidad de dispositivos conectados a un servidor está limitada por las tarjetas de I/O que disponibles..
- No existe flexibilidad de utilizar de manera dinámica por otros servidores el espacio disponible.

ALMACENAMIENTO DIRECTAMENTE CONECTADO (DAS)

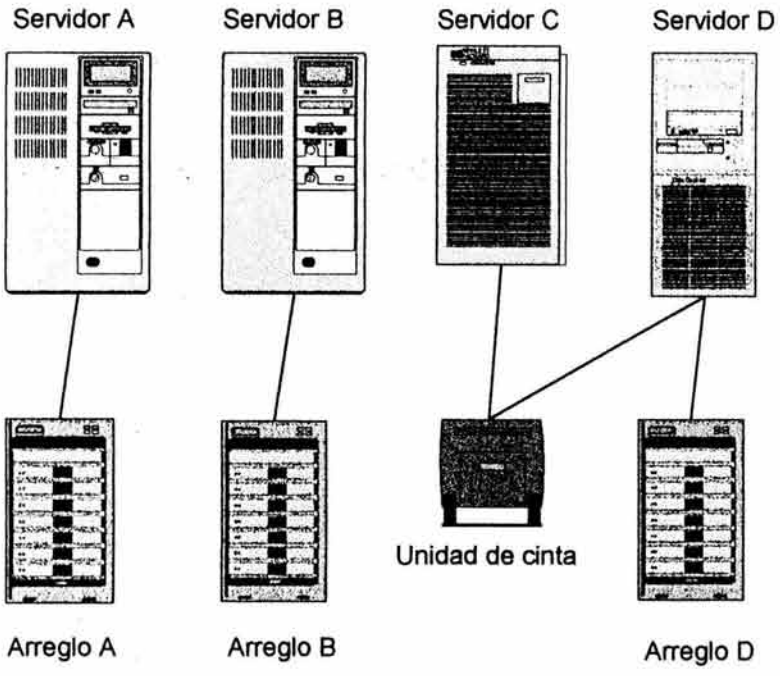


Figura 5.1.1 Direct Attached Storage (DAS)

Así por ejemplo si tomamos como referencia la figura 5.1.1, el servidor A puede tener disponible 10 GB de espacio en sus discos, espacio que los servidores B y C no pueden utilizar de manera sencilla.

Es por esto que SAN surge como respuesta a estas limitaciones, ya que ofrece un modelo de red que hace más eficiente el uso del almacenamiento.

Las SAN utilizan Fibre Channel por poseer características superiores a las tecnologías comunes como lo son SCSI, algunas de estas son:

- Amplia conectividad.
- Largas distancias.
- Alta velocidad de transmisión.
- Baja latencia.
- Integridad de los datos.

- Grandes transferencias de datos.
- Detección de errores.

Una *Storage Area Network* es una infraestructura de red que conecta cualquier servidor con cualquier dispositivo de almacenamiento a través de una red de alta velocidad formada por switches de *Fibre Channel* interconectados. Estos switches interconectados forman el corazón de la SAN: el Fabric.[1]

Un Fabric es una red cuidadosamente diseñada de switches de Fibre Channel inteligentes que provee escalabilidad a nivel empresarial, desempeño, manejabilidad y disponibilidad.
La figura 5.1.2 muestra un modelo genérico de una SAN.

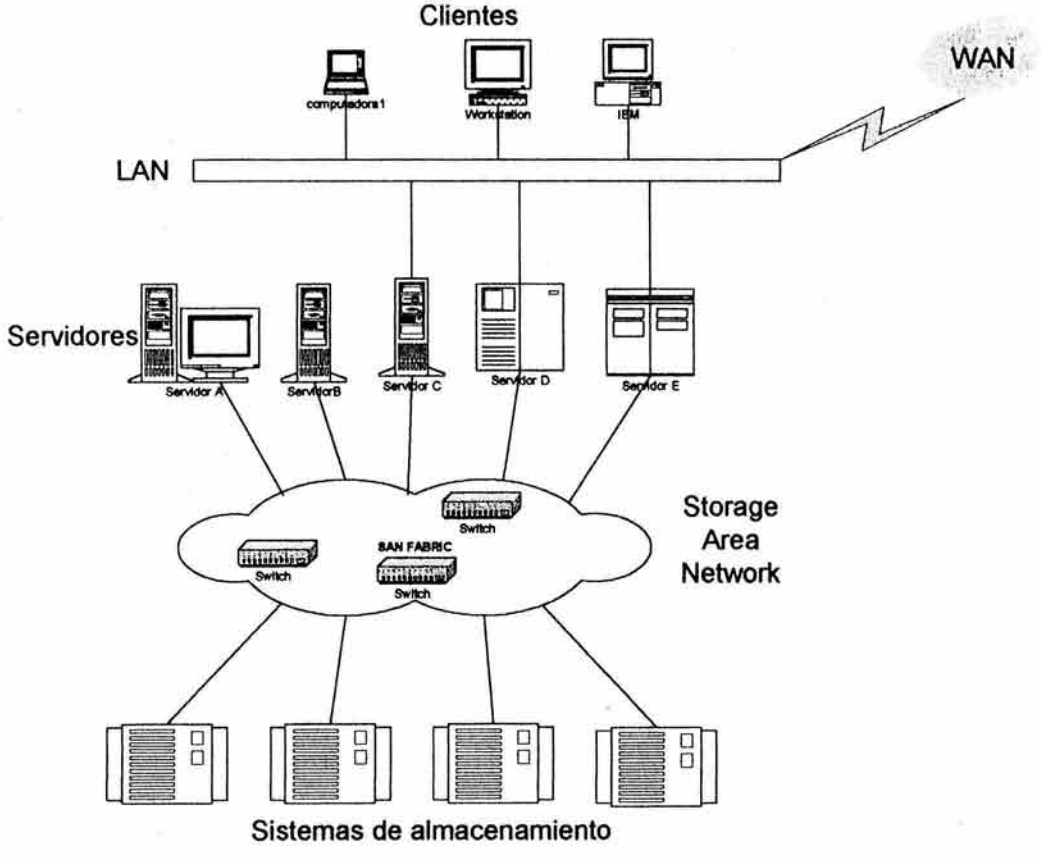


Figura 5.1.2 Storage Area Network (SAN Fabric)

[1] Brocade Communications Systems, Inc. llama SAN Fabric a la red que es formada por equipos especializados para SAN. www.brocade.com.

En la figura 5.1.2 se muestran varios servidores que están conectados a una red de almacenamiento que está formada por switches de Fibre Channel y otros elementos de red que utilizan el mismo protocolo.

Los sistemas de almacenamiento también son especializados ya que deben soportar la utilización de Fibre Channel así como contar con dispositivos de conexión que utilicen señales ópticas. Por lo regular estos dispositivos cuentan con software y terminales que permiten realizar su administración [2]

Los clientes que utilizan estos servidores sí hacen uso de la red convencional (LAN), diferente de la utilizada para la SAN.

Los elementos básicos de una SAN son:

1. Host Bus Adapter (HBA).- Este elemento, como lo hemos mencionado en los anteriores capítulos, es una tarjeta de I/O en el servidor que deseamos que esté utilizando el almacenamiento de la SAN. Es una interfaz que comunica al servidor con los demás elementos de la SAN, por lo regular utilizando Fibre Channel.
2. Fibre Channel Cable. – Provee la conexión física entre el servidor y los demás elementos de la SAN. Puede utilizar cables ópticos o de cobre.
   La siguiente tabla muestra las características de los cables utilizados en una SAN.

Tipo de Medio	Velocidad	Distancia (m = metro)
Fibra Mono-Modo de 9 μ (Largas distancias)	100 Mb / sec 200 Mb /sec	2 m – 10 km
Fibra Multi-Modo 50 μ (Distancias cortas)	100 Mb /sec 200 Mb / sec	2m – 500 m
Eléctrica (cobre)	100 Mb / sec	0m – 24 m

[2] De los principales proveedores en México de este tipo de dispositivos son EMC y Hitachi, quienes además de proveer de los dispositivos, proveen el análisis, diseño, implantación de la SAN; así como de los dispositivos de red que son soportados por sus equipos.

3. GBIC (Gigabit Interface Converter).- Al igual que con el A5200 de los capítulos anteriores, este dispositivo convierte las señales luminosas utilizadas por Fibre Channel, en señales eléctricas que utilizan los switches que conforman la SAN. También los switches de Fibre Channel integran este tipo de dispositivos para realizar la conexión.
   Los GBIC son pequeños módulos seriales, reemplazables en caliente que proveen la interface de la media (fibra o cobre).
   Los dispositivos de 2 Gbit/seg usan GBICs llamados *Small form Factor Plug (SFP).* La siguiente tabla muestra las distancias alcanzadas por estos dispositivos.

Tipo	Distancia
SFP Short Wave Laser	Hasta 500 metros
SFP Long Wave Laser	Hasta 10 kilómetros
Extended Long Wave Laser	Hasta 80 kilómetros

4. Switch de Fibre Channel.- Forman la red de almacenamiento y proveen la comunicación entre los dispositivos de almacenamiento y los servidores que los utilizarán.

   Estos dispositivos proveen las siguientes características:
   - Conectividad multipuerto (8, 16, 64, 128 puertos)
   - Soporta dispositivos Fabric y Loop.
   - 1 Gbit/seg o 2 Gbit/seg de velocidad (auto-sensable).
   - Desempeño Full Dúplex con ruteo dinámico.
   - Trunking lógico de 8 Gbit/seg para conectividad entre switches.
5. Sistema de Almacenamiento.- Proveen el almacenamiento como tal, por lo regular son dispositivos de gran capacidad que están formados por muchos discos y a su vez proveen de características como *hot pluggability (conexión en caliente),* niveles de RAID por hardware, redundancia, etc.

También se consideran a los dispositivos de cintas como parte de la SAN ya que también pueden ser utilizados por varios servidores a través de la SAN.[3]

También se pueden utilizar en una SAN, como en redes TCP/IP, ruteadores, puentes y extensores que permiten extender sus capacidades.

La figura 5.1.3 ejemplifica la conexión entre estos componentes.

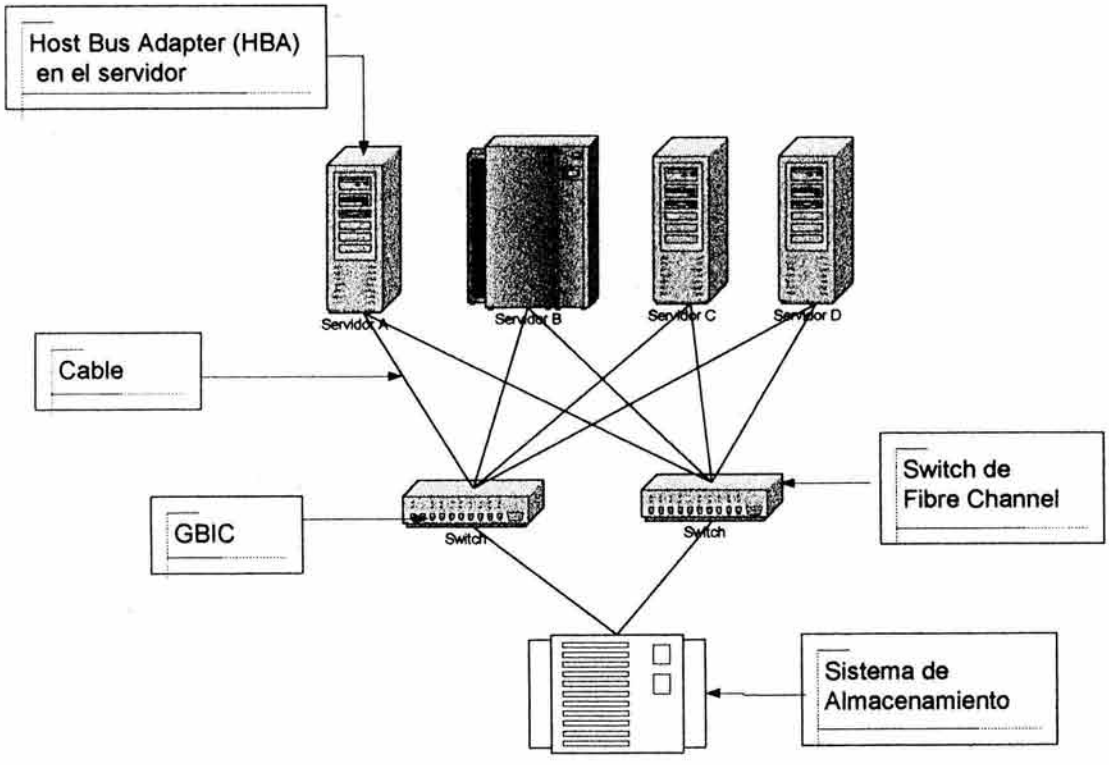


Figura 5.1.3 Elementos que intervienen en la SAN.

[3] A pesar de que pueden ser integradas como parte de la SAN, los respaldos por lo regular son realizados dentro de los mismos dispositivos de almacenamiento de disco a disco, lo que disminuye tanto el tiempo de respaldo como el de recuperación de datos.

Algunos de los beneficios de la utilización de Fibre Channel en una SAN son:

- Una nueva infraestructura de Red Multipropósito para conectar dispositivos de almacenamiento de *open systems*[4], redes, cadenas de video, y *clústers* de servidores.
- Provee un medio de hardware de transporte general para protocolos superiores como SCSI, IP, etc.
- Alta velocidad: tasa de 2 Gbit/seg., conexión dedicada full-dúplex.
- Hasta 10 km. (más extensiones para soportar distancias de miles de kilómetros, excelentes para recuperación de desastres (DRP).
- Soporte a sistemas heterogéneos (incluyendo AIX, NT, Solaris, LINUX, Novell y otros)
- Adopta entornos y aplicaciones actuales.

Otra de las ventajas los beneficios de la utilización de SAN es la mejora en la utilización del almacenamiento.

A este respecto, la figura 5.1.4 muestra cómo el almacenamiento puede ser compartido por varios equipos no importando qué tipo de sistema operativo utilicen, de esta manera se desperdician menos recursos de almacenamiento.

[4] *Open Systems* o sistemas abiertos, se refiere a aquellos sistemas, ya sean aplicaciones o equipos, que pueden ser utilizados en varios ambientes o sistemas operativos.

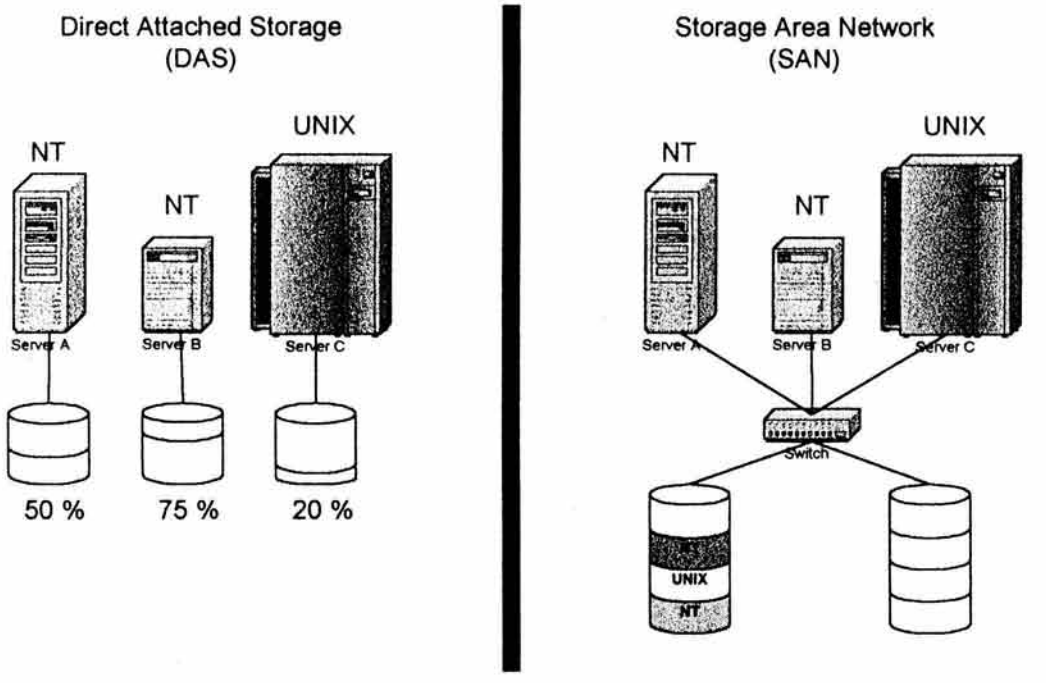


Figura 5.1.4 Utilización del almacenamiento

También con la utilización de SAN se puede mejorar la disponibilidad de la información al tenerla ya sea como espejo dentro del mismo dispositivo de almacenamiento o realizando respaldos a disco que además de ser realizados más rápidamente, son recuperados casi al instante.

La figura 5.1.5 muestra un esquema tradicional de respaldos en contraposición con el mismo esquema, pero en una SAN.

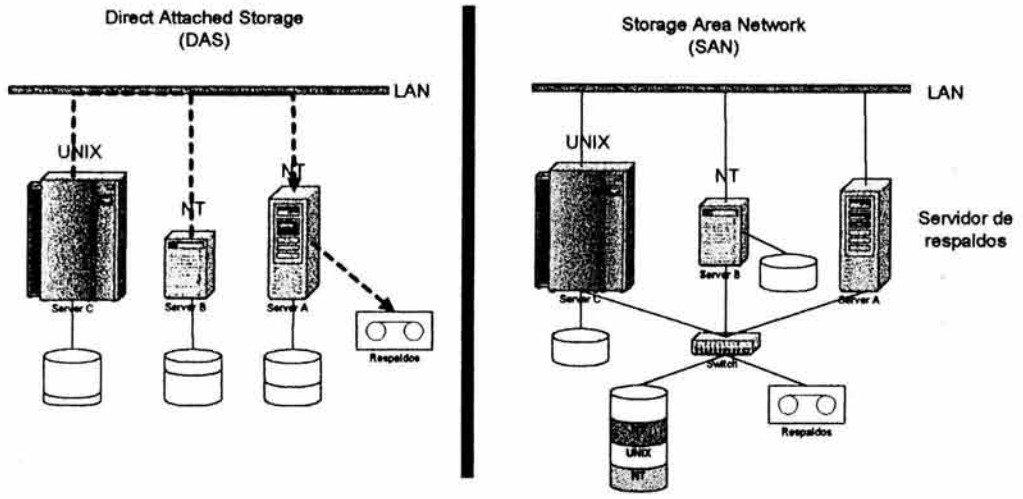


Figura 5.1.5 Disponibilidad de la información.

Hablando nuevamente de disponibilidad, pero a nivel de servidores, también a a través de una SAN podemos mejorar este aspecto.

La figura 5.1.6 muestra un esquema tradicional de servidores principales con servidores en espejo, así como el mismo esquema pero en una SAN.

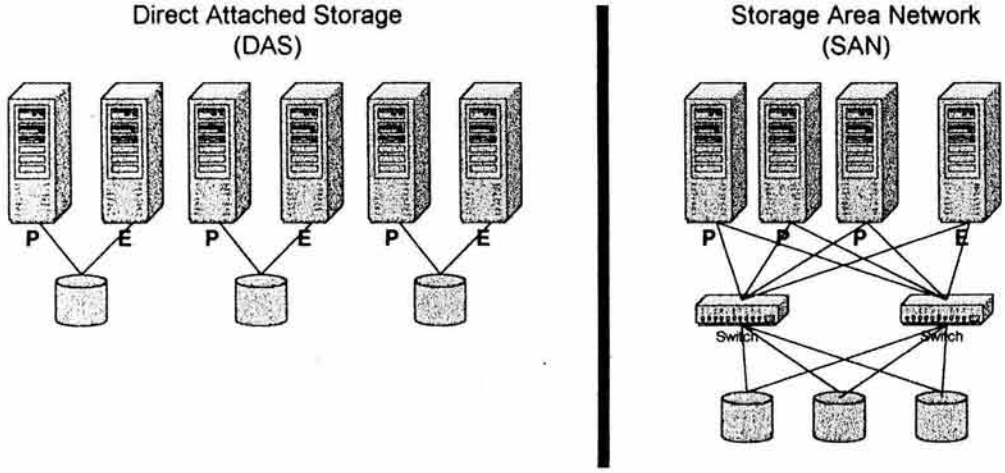


Figura 5.1.6 Protección de aplicaciones mejorada.

Las redes de almacenamiento (SAN) utilizan equipos especializados para brindar la conexión entre los servidores y los dispositivos de almacenamiento, todos ellos utilizando Fibre Channel; sin embargo también es posible utilizar como solución al problema del almacenamiento creciente, lo que se conoce como NAS (Network Attached Storage).

La principal diferencia entre una SAN y una NAS es que la segunda utiliza las redes existentes de TCP/IP, así como estos mismos protocolos, para proporcionar la conexión entre los servidores y los dispositivos de almacenamiento.

Tanto SAN como NAS proporcionan una solución a los problemas de almacenamiento; sin embargo hay que hacer notar que ambas pueden ser la solución correcta dependiendo del tipo de necesidades y el costo.

Evidentemente una SAN es mucho más costosa al necesitar la construcción de una red independiente que trabaja bajo Fibre Channel, contrariamente a NAS que puede utilizar la infraestructura ya existente.

En la figura 5.1.7 puede observarse como la red de datos, identificada como LAN, no es la misma que la red de almacenamiento, identificada en este caso como Brocade SAN Fabric.

El Brocade[5] es un proveedor de switches y elementos de red para SAN (SAN Fabric) muy conocido en el mercado.

[5] Brocade ha contribuido a la generación de estándares para la construcción de SAN y de consolidación de almacenamiento en las empresas (www.brocade.com).

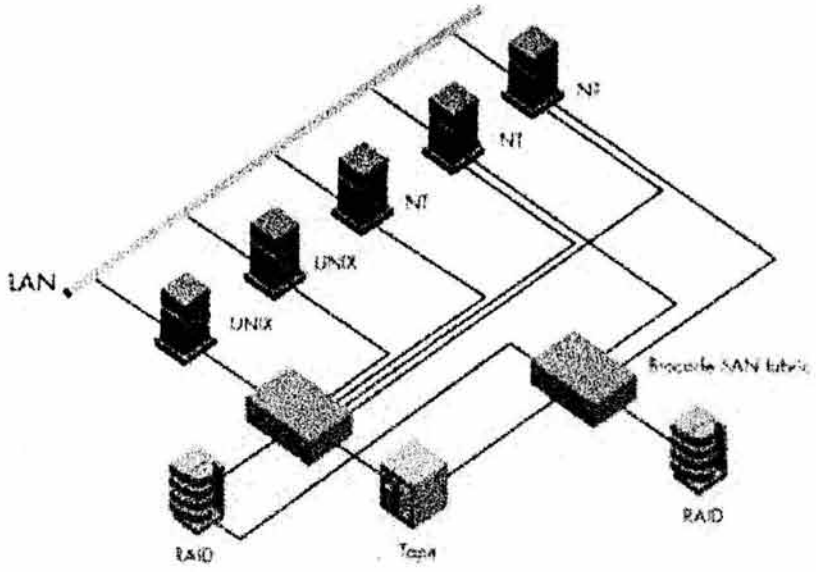


Figura 5.1.7 SAN Fabric.

5.2 PRINCIPALES APLICACIONES

5.2.1 RESPALDO Y RESTAURACIÓN DE DATOS.

Muchas organizaciones han tenido que enfrentar el reto de respaldar más datos con ventanas de respaldo más cortas y han tenido que encontrar estrategias de respaldo para varios tipos de datos que dependen qué tan críticos son.

Tradicionalmente, los modelos de respaldo y restauración comprenden el uso de discos y unidades de cinta dedicadas para cada servidor en particular, bajo este esquema cada servidor respalda su información localmente en las unidades de cinta que tiene conectadas. Este diseño es relativamente pobre en cuanto al uso de los recursos de cinta se refiere ya que si la unidad de cinta de un servidor no se encuentra en uso, otro servidor no puede utilizarlo de manera sencilla. Además de lo anterior, cada plataforma de sistema operativo tiende a utilizar aplicaciones propias para el respaldo y restauración de datos.

La figura 5.2.1 muestra este esquema tradicional de respaldo de datos.

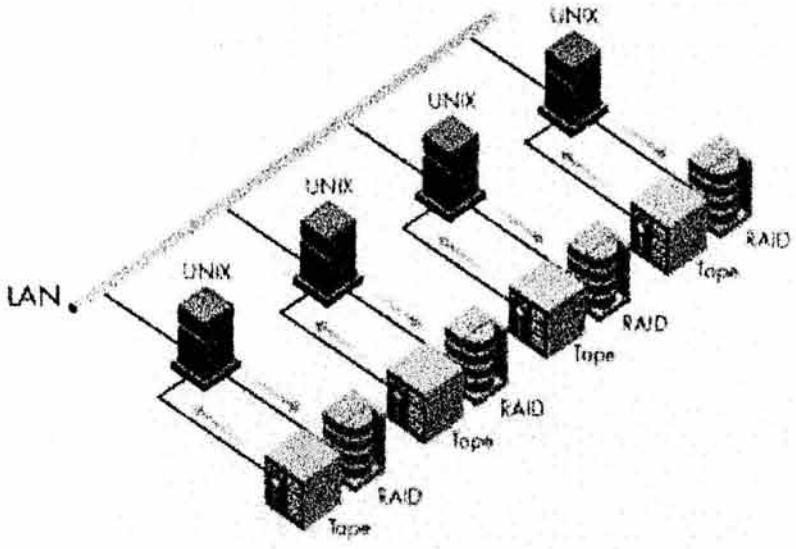


Figura 5.2.1 Esquema de respaldo centralizado.

Otro modelo de respaldo y restauración más orientado a centralizar el recurso destinado a esta labor es aquel en el que se tiene un servidor primario de respaldos que controla los recursos de cinta. Algunas aplicaciones sofisticadas como VERITAS Netbackup, Legato Networker y Tivoli Storage Manager, por ejemplo, son los encargados de controlar el proceso de respaldo.

Bajo este esquema, el servidor de respaldo recibe datos de otros servidores a través de una LAN (Local Area Network) o una WAN (Wide Area Network), para almacenarlos en los discos o cintas que tiene directamente conectados.

Este esquema centralizado provee mejor utilización de los recursos; sin embargo, su desventaja son los cuellos de botella que la red puede introducir al proceso, lo que puede generar un impacto en la habilidad del sistema de coincidir con las ventanas de respaldo y restauración de la empresa.

La figura 5.2.2 muestra este esquema centralizado.

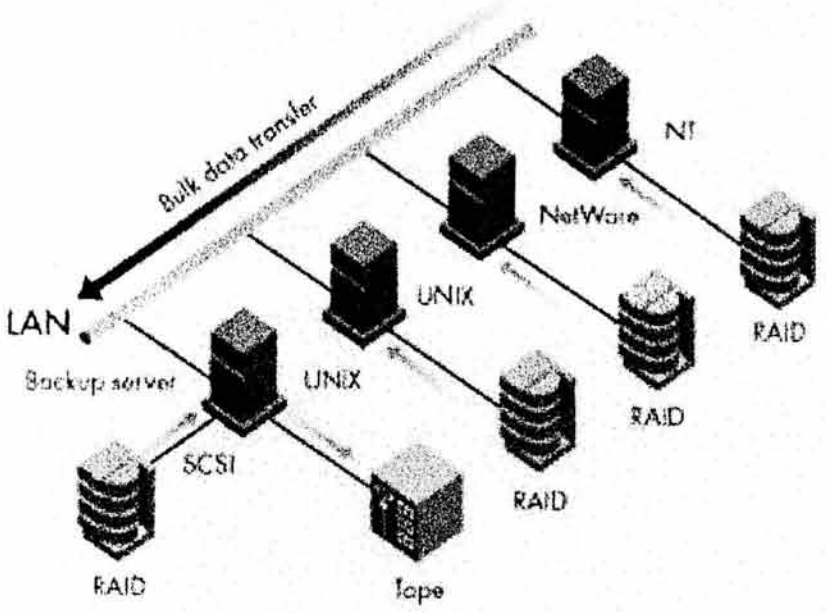


Figura 5.2.2 Esquema centralizado de respaldos a través de LAN.

En contraste con todo esto, una SAN puede acelerar y simplificar el proceso de respaldo y restauración de datos. La ventaja que se tienen es la utilización de una red independiente de Fibre Channel o lo que se llama Fibre Channel Fabric. Las capacidades de 2 Gbit/sec. Full duplex del Fibre Channel Fabric, puede mejorar significativamente el desempeño del proceso de respaldo. Además, Fibre Channel en sí, está diseñado para transportar grandes bloques de datos con mayor eficiencia y confiabilidad que las redes basadas en el protocolo de red IP.

Los dos esquemas más populares de SAN para el respaldo y restauración de datos son los modelos conocidos como "LAN-free" y "server-free" .

LAN-free

Bajo este modelo, la LAN no es utilizada para transportar los datos a respaldar sino que se utiliza la red de Fibre Channel para este propósito, con la particularidad de que cada servidor manda sus datos directamente a los

recursos de cinta y con aplicaciones sofisticadas de respaldo se lleva su control.

La figura 5.2.3 muestra este modelo de respaldo.

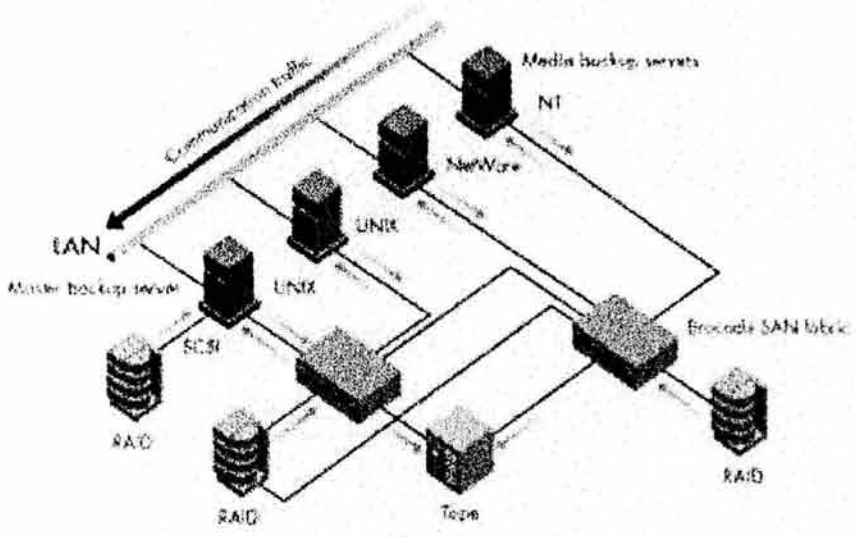


Figura 5.2.3 Modelo de respaldos *LAN-free*

Server-free

Este modelo es más reciente y su funcionamiento radica en que los datos son transferidos directamente entre los dispositivos de almacenamiento (por ejemplo, de disco a cinta) sin el uso de servidores. Este proceso es realizado mediante una tecnología llamada *Third-Party Copy*, que es implementada en los dispositivos SAN o en los dispositivos de almacenamiento directamente.

El modelo *server-free* reduce significativamente ciclos de CPU en el servidor, por lo tanto liberando de carga de trabajo al equipo que puede ser utilizado para mejorar el desempeño de las aplicaciones que tiene corriendo.

La figura 5.2.4 muestra este modelo.

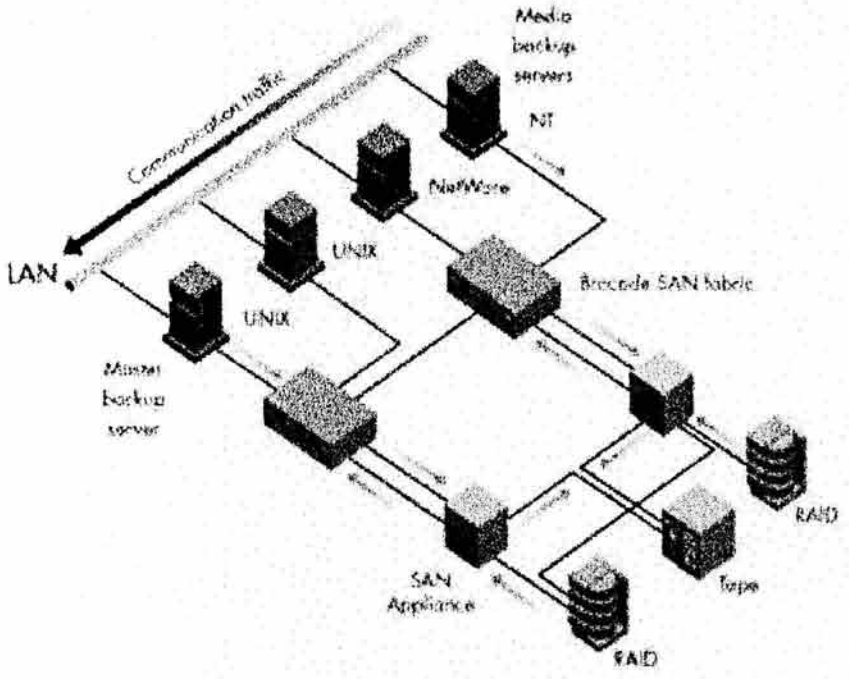


Figura 5.2.4 Modelo *Server-free*.

5.2.2 CONTINUIDAD DEL NEGOCIO.

Las redes de almacenamiento proveen la habilidad para recuperar datos de manera rápida y recuperarse después de un desastre. Todo esto es debido a las características distribución de almacenamiento que se tiene con una SAN.

Para protegerse de tiempos fuera en los servidores y reducir el riesgo al negocio, una SAN debe satisfacer requerimientos como los siguientes:

- Eliminar puntos de falla para incrementar la disponibilidad a los datos.
- Incorporar software que permita tener una mejor tolerancia a fallas.

- Contar con un esquema de respaldo y recuperación adecuado que reduzca el tiempo de recuperación.
- Utilizar respaldos remotos y espejos de los centros de datos separados por grandes distancias para minimizar el impacto de un desastre mayor.

El factor más importante de utilización de SAN para proveer la continuidad del negocio es provisto por la factibilidad de acceder datos a grandes distancias gracias a la tecnología de Fibre Channel que permite una conectividad de hasta 120 km.

Esta distancia permite a los usuarios mantener separados los centros de datos destinados para la recuperación de desastres.[6]

5.2.3 ALTA DISPONIBILIDAD

Algunos de los beneficios clave de SAN para la alta disponibilidad incluyen redundancia, protección dinámica a fallas y capacidades de re-enrutamiento. También provee capacidades de *hot – pluggable*[7] que permiten a las empresas instalar, configurar y poner el almacenamiento disponible sin experimentar tiempos fuera de servicio.

SAN también puede soportar operaciones de alta disponibilidad a través de configuraciones de *clúster.*[8] Los clústers son normalmente utilizados para asegurar que las aplicaciones continuarán dando servicio en el caso de la falla en un servidor.

Los ambientes tradicionales de clúster que no integren el uso de SAN, incluyen dos servidores compartiendo el almacenamiento en disco. Si un servidor falla, el otro asume la carga del servidor que falló y continúa dando el servicio. El

---

[6] A pesar de que los Planes de Recuperación de Desastres han sido reconocidos como una práctica importante en la empresa, no cualquiera destina recursos para su implementación. La SAN facilita esta práctica.

[7] El término *hot-pluggable* se da a los equipos o componentes que pueden ser reemplazados sin necesidad de apagar el equipo.

[8] Por *clúster* nos referimos a una configuración redundante diseñada con dos elementos: uno activo y otro pasivo. Si uno falla el otro se activa para dar servicio.

servidor "sobreviviente" accede a los datos a través del disco compartido. Esto representa un diseño relativamente poco flexible porque está limitado a dos servidores compartiendo el almacenamiento y uno de ellos permanece pasivo hasta que la falla ocasiona que se active. Además de esto, un factor importante es que los servidores en este tipo de configuración son dispuestos uno cerca del otro, lo que resulta en poca protección contra desastres.

En contraste con lo anterior, con SAN, más servidores pueden compartir o utilizar el almacenamiento de la SAN, además de que es posible manejar mayores distancias entre servidores contribuyendo a un plan de recuperación de desastres más efectivo.

5.2.4 CONSOLIDACIÓN DE SERVIDORES Y ALMACENAMIENTO.

El modelo tradicional de almacenamiento DAS (Direct Attached Storage) tiene la desventaja de hacer la utilización de los recursos menos dinámica, por lo que usualmente las empresas necesitan comprar más servidores para manejar el almacenamiento o comprar más almacenamiento para sus servidores.

Con el esquema de SAN se provee una conectividad más flexible, el uso de los recursos de manera más eficiente y escalable. Todo esto por la habilidad de conectar cualquier servidor hacia cualquier recurso de almacenamiento a través de switches de SAN, promoviendo la utilización de los recursos de manera compartida. Con este modelo, también es posible contar con un ambiente heterogéneo en cuanto a los equipos que adquiere la empresa.

Las anteriores fueron algunos de los principales usos de las redes de almacenamiento; sin embargo, cabe mencionar que la adquisición y construcción de una SAN debe responder a los requerimientos de la empresa y a una planeación de capacidad muy detallada debido a lo costosa que puede resultar este tipo de tecnología.

CONCLUSIONES

En algunos casos el uso de un sólo disco o de varios discos aislados limita la capacidad de almacenamiento o el desempeño de los servidores, y por ende, de los servicios provistos.

En el caso de los servidores unix, estos pueden contar con muchos usuarios haciendo uso de una parte del disco, de tal modo que podrá haber un momento en el cual ya no sea suficiente el espacio con el que se cuenta y se tenga que involucrar la adquisición de otro disco y así sucesivamente conforme se vayan agotando los recursos de almacenamiento disponibles. Se podrían ir conectando discos y discos para satisfacer estas necesidades.

Para este tipo de problemas, las Técnicas RAID pueden ser una solución conveniente.

Existen varias técnicas o niveles de RAID, las cuales tienen o carecen de ciertas características, por lo que siempre es necesario tomar en cuenta todo el entorno del sistema como lo son tipo de aplicaciones que se corren, tipo de accesos que se realizan a la información, hardware con que se cuenta y sobre todo, el costo. Una configuración que tenga la mayor cantidad de beneficios será la configuración más costosa también.

En muchos casos lo que habrá que hacer para saber cual es la técnica de RAID más conveniente a implementar, será haciendo un balance entre desempeño, redundancia y menor costo. No siempre una técnica de RAID será la más conveniente para todos los sistemas, ni tampoco hay una técnica que abarque por completo estas tres características, por lo regular la de mejor desempeño no siempre es la que tiene mejor redundancia y viceversa.

No existe una técnica buena y otra mala, todo depende de las características que deseamos tener en nuestros sistemas para que una técnica sea más conveniente que otra.

En el caso de la técnica RAID 0 +1, ésta puede ser utilizada en los servidores unix no importando, la mayor de las veces, la aplicación que ejecuten, ya que esta es precisamente una de las técnicas más completa en cuanto a desempeño y redundancia se refiere.

Cuando nuestras necesidades de almacenamiento son más especializadas y requerimos compartir datos entre sistemas heterogéneos o multiplataforma, e inclusive nuestras necesidades de disponibilidad son mayores, la tendencia, si el costo lo permite, son las Redes de Almacenamiento.

BIBLIOGRAFÍA

Sitios en Internet

- **STORAGEREVIEW**
  http://www.storagereview.com

- **BROCADE**
  http://www.brocade.com

- **VERITAS**
  http://www.veritas.com

- **SUN MICROSYSTEMS**
  http://www.sun.com
  http://docs.sun.com

Libros y manuales

- **Sun StorEdge Volume Manager Administration**
  Sun Microsystems, Inc,
  Oct. 1999.

- **COCKCROFT, Adrian.**
  Sun Performance and Tuning.
  Sun Microsystems Press.
  1998.