UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

DISEÑO DE ALMACENAMIENTO CON REPLICACIÓN REMOTA PARA BUSINESS CONTINUITY Y DISASTER RECOVERY

INFORME DE SUFICIENCIA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE TELECOMUNICACIONES

PRESENTADO POR:

DARLING MANET OSCANOA ACUÑA

PROMOCIÓN
2003- I

LIMA – PERÚ
2008

DISEÑO DE ALMACENAMIENTO CON REPLICACIÓN REMOTA PARA BUSINESS CONTINUITY Y DISASTER RECOVERY

Dedico este trabajo a:
Mi madre, por sus sabios consejos y apoyo
siempre incondicional,
Mi padre, por el ejemplo de perseverancia y que
con sacrificio se logran los
objetivos en la vida,
A mis hermanos por el apoyo.
Y a Carol, por estar siempre ahí
cuando más la necesitaba

SUMARIO

En el presente informe se expondrá las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre. Se explicarán los fundamentos teóricos y protocolos relacionados a la implementación de un almacenamiento externo con alta disponibilidad. Se realizará la evaluación de alternativas, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

Se expondrá un caso de estudio donde se mostrará las técnicas explicadas, se mostrarán estadísticas de entrada y se calcularán los recursos de hardware y transporte necesarios para el caso de estudio. Finalmente se presentarán conclusiones y recomendaciones.

ÍNDICE

CAPÍTULO II

FUNDAMENTOS TEORICOS

CAPÍTULO III

DESARROLLO DE LA PROBLEMÁTICA

CAPÍTULO IV
CASO DE ESTUDIO

PRÓLOGO

Desde finales de la década de los años 1970s cuando se inició el concepto de business-recovery o recuperación del negocio éste ha continuado creciendo y desarrollándose, moviéndose originalmente desde sus aplicaciones originales sobre mainframes hasta incluir recuperación de desastres para telecomunicaciones, procesamiento distribuido y más reciente, recuperación de desastres en el área de redes y en el área personal de trabajo.

Cualquiera sea la razón: accidentes, desastres, atentados, eventos naturales que interrumpan las actividades de cualquier negocio, una cosa es cierta: El negocio y la corporación pierden dinero. Generalmente el monto de este dinero depende cuan preparado está el negocio para afrontar con estos eventos. Consecuentemente un actualizado, bien planificado y bien practicado plan de recuperación de desastres permitirá al negocio retornar rápidamente en operación a comparación a los meses o más aún años de repercusión del desastre en los negocios que no cuentan con este plan de recuperación de desastres.

En el presente informe se expondrá las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre. Se explicarán los fundamentos teóricos y protocolos relacionados a la implementación de un almacenamiento externo con alta disponibilidad. Se realizará la evaluación de alternativas, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

Se expondrá un caso de estudio donde se mostrará las técnicas explicadas, se mostrarán estadísticas de entrada y mediante software de EMC se calcularán los recursos de hardware y transporte necesarios para el caso de estudio. Finalmente se presentaran conclusiones.

El presente informe no pretende ser una guía para desarrollar un Business Continuity Planning (Planeamiento de Continuidad de negocios), pues solo abarcará la fase 4 de su desarrollo centrándose exclusivamente en calcular y dimensionar los recursos de Hardware, Software y Transporte necesarios para una solución de almacenamiento replicado basado en storage.

En el Capitulo I se expondrán los antecedentes históricos de perdidas de grandes cantidades de dinero con la caída de los sistemas de una empresa, se explicarán las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre, se identificará la problemática a tratar, se mostraran estadísticas globales de las estrategias y los tipos de replicación que las áreas de IT están implementando, asimismo se definirán los objetivos, justificación, alcances y limitaciones del presente informe.

En el Capitulo II se expondrá el marco teórico, los protocolos y tecnologías relacionados a la implementación de un almacenamiento externo con alta disponibilidad.

En el Capitulo III se realizará la evaluación de alternativas, opciones de conectividad, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

En el Capitulo IV se expondrá un caso de estudio donde se mostrará las técnicas explicadas anteriormente, se mostraran estadísticas de entrada y se calcularán los recursos de hardware y transporte necesarios para este caso de estudio.

Finalmente se presentaran conclusiones y recomendaciones.

CAPÍTULO I
DESCRIPCION GENERAL

En el presente capítulo, se expondrán los antecedentes históricos de perdidas de grandes cantidades de dinero con la caída de los sistemas de una empresa, se explicarán las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre, se identificará la problemática a tratar, se mostraran estadísticas globales de las estrategias y los tipos de replicación que las áreas de IT están implementando, asimismo se definirán los objetivos, justificación, alcances y limitaciones del presente informe.

1.1 Antecedentes

Desde el inicio de la era de las computadoras y la centralización de la información que ésta trajo consigo, se vio la necesidad de contar con mecanismos de recuperación de la información y el acceso a la misma ante cualquier evento o falla.

De ese modo apareció un primer mecanismo de recuperación ante desastres: los backups, en este campo el punto de partida data de 1951 con las ahora obsoletas tarjetas perforadas como una manera de respaldar la información generada por la primera maquina de computación digital la UNIVAC I (Universal Automatic Computer) [1.1]

En la década de los años 1960s aparecen las cintas magnéticas como un método de backup más eficiente que las tarjetas perforadas, pues lograron almacenar en una sola cinta el equivalente a 10000 tarjetas, lo cual llevo a su rápida aceptación en el mundo de la informática.

[1.1] Referencia: http://www.backuphistory.com/

En 1956 IBM introduce el primer disco duro HD, en las décadas de los años 1960s y 1970s el backup en disco no fue posible debido a su alto costo, gran tamaño y poca capacidad de almacenamiento en comparación con las cintas, a mediados de los 1980s los avances en estos HDs hacen posible considerar backups a disco, sin embargo no es hasta la década de los años 1990s que los HDs se vuelven una real alternativa para backup.
Cronológicamente en el tiempo aparecen también otras tecnologías como los floopy disk, los CDs, DVDs, Flash Drives, Blue Ray Disk HD-DVD en el orden dado como alternativas para respaldar información.
Desde finales de la de década de los años 1970s cuando se inició el concepto de business-recovery o recuperación del negocio este ha continuado creciendo y desarrollándose, grandes ejemplos de eventos que afectaron la disponibilidad de los servicios informáticos de las empresas pueden ser listados, a continuación se presentaran solo alguno de ellos:

1.1.1 Ataque Terrorista Torres Gemelas NY 11 de Setiembre del 2001

Uno de los desastres más grandes causados por ataque terrorista recordados en la historia es el del reciente 11 de Septiembre del 2001, cuando un grupo de terroristas secuestraron 4 aviones comerciales e hicieron que se estrellaran 2 de ellos con las torres gemelas del World Trade Center en Manhattan, los resultados fueron cerca de 3000 personas muertas y más de $70 billones de dólares en perdidas [1.2]

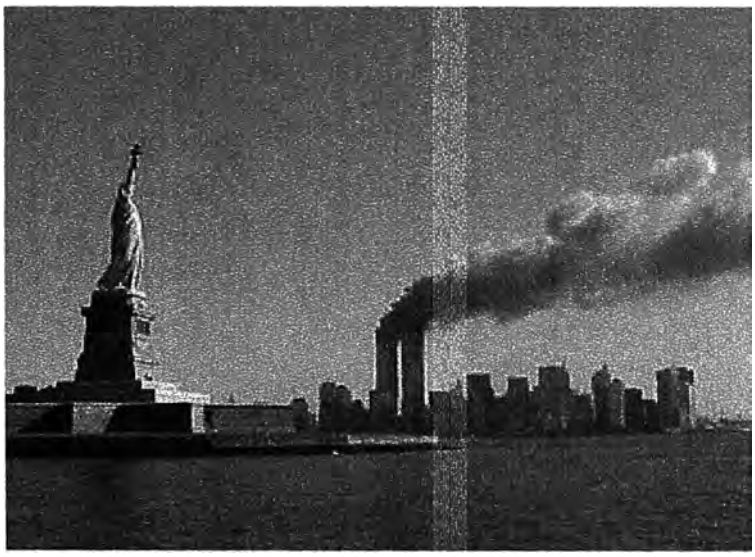


Figura 1.1: Estatua de la Libertad y Torres Gemelas, 11 de Septiembre

[1.2] Referencia:
http://www.bledconference.org/ECBledHome.nsf/5f5370162bfefab8c12565ef00600a0c/0aabf713c26915c0c1256bba002c73bc?OpenDocument

En la tabla siguiente se muestra en resumen el efecto producido por este ataque terrorista a las empresas que operaban en las torres gemelas.

Company/ Building/floor	Industry	# victims	# employees at ground zero	IT damage
Cantor Fitzgerald 1/101-105	Investments	730	1000	Resumed on line trading on 13 Sept with eSpeed and Tradespark
Marsh McLennon Tower 1/93-100 Tower 2 47-54	Consulting Financial	245	1,900	Lost most of IT staff, recovered all systems
AON Tower 2/92-99, 100	Insurance	165	1,100	No data loss, But high speed access down until Oct 14
Morgan Stanley Tower 2 5 WTC	Investments	6	3,700	$150M No data lost, email restored in 72 hours
Merrill Lynch	Investment Services	3	9,000	Relocated 8,000 employees, Restored sytems same day
American Express Tower 1, 3WFC, /WTC	Financial Services	11	3,200	$140 M no interuption in customer Service

Tabla 1.1: Resumen impacto ataque terrorista del 11 de Septiembre [1.3]

Debido a la experiencia dejada por un atentando anterior al mismo edificio (coche bomba en 1993) [1.4] y la planificación para la llegada del año 2000 la mayoría de estos negocios contaba con un plan de recuperación de desastres, Infraestructura de Sites Secundarios que permitieron al negocio recuperarse del desastre en tiempos razonables.

Cantor Fitzgerald, firma de negocios financieros en línea que factura $50 trillones anuales pudo recuperar su sistema principal de transacciones financieras a las 48 horas del atentando gracias a que cuenta con sites secundarios en New Jersey y London.

[1.3] Referencia http://www.bledconference.org/ECBledHome.nsf/5f5370162bfefab8c12565ef00600a0c/0aabf713c26915c0c1256bba002c73bc?OpenDocument

[1.4] http://en.wikipedia.org/wiki/World_Trade_Center_bombing

Marsh, firma financiera, consultora y broker de seguros perdió 129 miembros de su staff de IT, también perdió 2 datacenters, 255 servidores, aplicaciones y data por cerca de $70 millones. Los proveedores de IT en este caso fueron los artífices para que Marsh recupere sus sistemas a la medianoche del atentando, a pesar de todo esto Marsh reportó un crecimiento total anual del 4% para el 2001

Un firma de seguridad de la Bolsa de NY (NYSE) anónima, a pesar de estar ubicada a 2 cuadras del Trade Center tuvo que mover sus aplicaciones al site de su proveedor de Recuperación de Desastres, la cual le tardó 1 semana en completarse para poder atender a los más de 1 millón de clientes con las que cuenta. Sin embargo debido a las condiciones del site del proveedor no tenían la seguridad de poder siquiera operar por más de 6 semanas, A partir de ese momento la firma empezó con la construcción del site de contingencia en New Jersey [1.5]

El 11 de Septiembre demostró la importancia estratégica del planeamiento de recuperación de desastres y la administración de crisis, la habilidad de restaurar/recuperar los sistemas, el e-commerce, operaciones aun con las perdidas catastróficas de las instalaciones físicas fue clave para la supervivencia de los negocios de inversión y financieros ahí establecidos. Una lección aprendida es que aunque la mayoría estaba preparada para perder sus sistemas, no estaban preparadas para perder gente capacitada justamente en la recuperación del desastre.

1.1.2 Aerolínea Mesaba Minneapolis caos en los vuelos 2000[1.6]

La historia de Mesaba empieza en Diciembre del 2000, durante la temporada más alta del año, una falla de Hardware ocurrida en un servidor Novell Netware donde residía la aplicación "Fligth Track", este sistema critico hace todo desde juntar pilotos con sus asistentes de vuelos a los vuelos, mapeo de rutas, cálculo de condiciones climatológicas, seguimiento de carga y pasajeros en los vuelos.

El sistema cayó desde el mediodía hasta las 6pm, resultando en 350 vuelos cancelados y más de 300 retrazados: "Tuvimos que hacer modificaciones de la restauración por unas 3 semanas" Mesaba perdió más de $1 millón y molestos clientes.

"Nosotros no podríamos superar que esto nos pase otra vez" dijo Scott Ficet Director de IT de Mesaba

[1.5] http://www.informationweek.com/news/showArticle.jhtml?articleID=6507729
[1.6] http://www.snwonline.com/case_studies/disaster_planning_11-17-03.asp

Mesaba aprendió la lección y hoy cuenta con un DataCenter alternativo donde replica su data de producción de manera Sincrónica vía una conexión IP arrendada entre ambos sites. A pesar de esto Mesaba no ha podido probar en la realidad sus sistemas de Recuperación de Desastres, aunque una vez estuvo muy cerca de hacerlo, tuvieron un problema con su sistema de aire acondicionado y la temperatura subió a más de 40C, sin embargo como eran las 9:30pm Mesaba optó por apagar servidores menos críticos disminuyendo la temperatura y controlando el problema, "Esta fue una decisión por la vía más fácil, en vez de mover todos los sistemas al site de contingencia"

1.1.3 Northgate Information Solutions Incendio en UK el 2005[1.7]

Northgate Information Solutions es un proveedor de aplicaciones de software y soluciones de outsourcing de los servicios públicos, RRHH y es también el más grande proveedor de HR y sistema de pago de planillas en el Reino Unido (Tiene más del 30% de todos los empleados de UK).

Figura 1.2: Incendio en UK 2005, destruye la oficina principal de Northgate

¿Que paso?

El incendió ocurrió el Sábado 11 de Diciembre del 2005 a las 06:00am en la oficina principal de Northgate ubicada en Boundary Way, Hemel Hempstead UK

Deshabilitó 212 sistemas de producción, relacionados a 209 clientes, incluyendo el sistema de pago de planillas (payroll), sistemas de admisión de pacientes y beneficios

Destruyó comunicaciones de voz y data, no hubo pérdida de vidas gracias a la hora en que ocurrió el incendio. Fue catalogado como el más grande incidente de este tipo en Europa en tiempos de paz.

1.7 http://www.availability.sungard.com/United+Kingdom/Resources/Case+Studies/Northgate+Information+Solutions.htm

La respuesta fue:

Llamar inmediatamente a la compañía de planeamiento de BC, DR[1.8] y Administración de crisis. La primera reunión del Equipo de emergencia fue la misma tarde del 11 de Diciembre en la oficina de Northgate en Londres. Distribuyendo más de 100 personas del staff técnico trabajando por más de 12 horas diarias los primeros 10 días. Finalmente todos los sistemas protegidos con Recuperación de Desastres fueron rápidamente restaurados, el resto de sistemas fueron restablecidos en su totalidad para la Navidad de ese año.

¿Qué trabajó bien?

Los planes de BC, DR y Administración de crisis. Respuesta rápida de todas las partes, staff, clientes, partners y proveedores. Gran entrega del personal para recuperarse del desastre. Minimizar el corte de servicio a los clientes. Procedimientos de Seguridad existentes y buenas relaciones con las autoridades locales. La organización revisaba y testeaba continuamente sus planes de DR para los demás locales.

Lecciones aprendidas:

Los planes de BC, DR y Administración de crisis son esenciales para el negocio y deberían ser revisados, testeados y mantenidos periódicamente. Proveer a los empleados claves un efectivo despliegue del lugar de trabajo, su localización física es crítica. Se deben tener claramente definidos los roles, tareas, prioridades y la forma en que deben reportar. Documentación comprensible y actualizada es clave pero no es suficiente para recuperar rápidamente, se necesita que el personal técnico conozca bien estos procedimientos. Ensayar múltiples escenarios de falla, no solo planear para la caída de un sistema o todo el datacenter. Una efectiva priorización es la clave para una efectiva recuperación, se debe tener claro como el negocio debe priorizar. Los sistemas de Control de Cambios deberían incluir además el impacto en el DR/BC Planning. Rotación de los respaldos o Backups y Procesos de Storage también deberían ser auditados.

Finalmente el desastre ocurrido a Northgate demuestra que el Business Continuity Planning o Planeamiento de la continuidad de los negocios es vital para cualquier negocio.

En General cualquiera sea la razón: accidentes, desastres, atentados terroristas, eventos naturales que interrumpan las actividades de cualquier negocio, una cosa es cierta: El negocio y la corporación pierden dinero. Generalmente el monto de este dinero depende cuan preparado está el negocio para afrontar con estos eventos. Consecuentemente un actualizado, bien planificado y bien practicado plan de recuperación de desastres permite al

[1.8] BC: Business Continuity (Cotinuidad de Negocios), DR: Disaster Recovery (Recuperación de Desastres)

negocio retornar rápidamente en operación a comparación a los meses o más aun años de repercusión del desastre en los negocios que no cuenta con este plan de recuperación de desastres.

1.2 Necesidades de las Empresas por contar con soluciones de alta disponibilidad

En el caso que una empresa experimente una interrupción de sus operaciones, pues simplemente no le es posible entregar el producto que acordó con sus clientes para los cuales esta empresa cobra, si esta interrupción tiene un "gran tiempo de duración" [1.9] esta empresa se convierte en un proveedor no confiable.

Según un estudio realizado por recovery-disaster.net existen más de 35 posibles eventos que podrían causar cortes de servicios catastróficos al negocio[1.10]. Ejemplos como:

- Una organización financiera experimenta una gran pérdida cuando un programador corrompió una Base de Datos actualizando un programa de producción sin seguir los procedimientos estándares
- Un Hospital metropolitano pierde irrevocablemente toda su BD de Fármacos, incluyendo información de los pacientes, cuando un disco se corrompe lleva a encontrar que los backup en cinta que sacan cada noche por más de 2 meses tenían archivos dañados. No hay backup de la BD perdida.
- Una compañía de seguros experimenta un corte debido a la caída de energía , aunque UPS y Generadores alimentaron al datacenter, otras operaciones en el mismo edificio carecieron de energía
- Una organización de servicios fue puesta fuera del negocio cuando se inundó toda el área alrededor de sus oficinas, bloqueando el acceso al edificio, aunque esta tenía un plan de Recuperación de Desastres en otro site, los únicos backup fueron almacenados en el cuarto de computadoras inaccesibles en ese momento.

En general los obstáculos de la disponibilidad pueden ser clasificados como [1.11]:

Desastres (<1%), naturales o causados por el hombre: inundaciones, sismos, incendios, huracanes y atentados terroristas son ejemplos de eventos calamitosos.

Ocurrencias no planificas (13%), fallas: corrupción de BDs, fallas de componentes, errores humanos.

[1.9] Los tiempos aceptables de corte de servicio, dependen del rubro de la empresa y de los objetivos de la empresa para ser una empresa fiable.
[1.10] http://recovery-disaster.net/it-disaster-recovery/disaster-recovery-threat.htm
[1.11] Source: Gartner Inc.

Ocurrencias planificadas (87%), backup, reporting, extraer data de warehouse, aplication y data restore.

El Costo del DownTime o Tiempo de Caída del servicio.

El costo del tiempo de caída de servicio variará de industria a industria de negocio a negocio y dentro de un negocio de aplicación a aplicación.

Los costos del downtime incluyen costos tangibles que pueden ser medidos en dólares, tales como pérdida de ganancias (ganancia actual, también ganancia futura), pérdida de productividad, pérdida de inventario, cargos tardíos y penalidades, costos legales, pérdida de inversión, pagos de compensación.

Los costos del downtime incluyen costos intangibles que no son fácilmente medidos, pero ellos impactan el negocio en la misma manera. Ejemplos incluyen: daños en la reputación ante clientes, partners, proveedores, bancos, financieros; deteriora el rendimiento financiero o valor compartido y decrementa la satisfacción del cliente.

Existen investigaciones realizadas por Gartner, IDC y AMR por calcular el promedio en dólares por hora que pierde las empresas por industria. A continuación se muestra el resultado del estudio realizado por AMR.

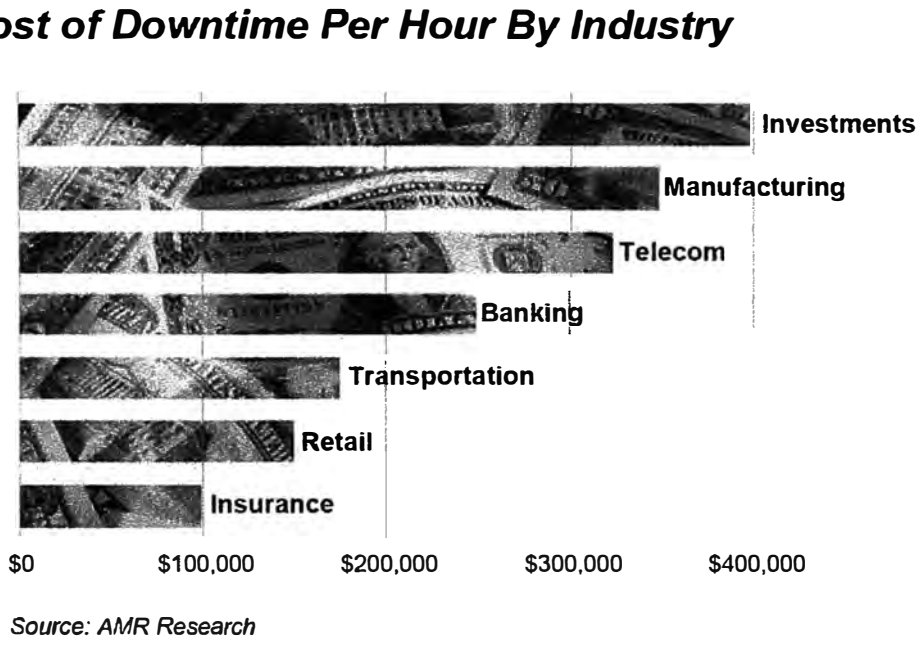


Figura 1.3: Costo del Downtime por hora según industria

En este estudio se puede visualizar claramente que las empresas que por la naturaleza misma del rubro dependen de sistemas automatizados, tales como

telecomunicaciones, financieras manufactureras y de energía. Por ejemplo el estudio muestra que una empresa de inversiones pierde $400 mil por hora de caída de servicio.
De esa manera no es de extrañar que justamente las industrias que se verían más afectadas de perder dinero con las caídas de sus sistemas sean las que más invierten en mantener un Business Continuity Planning actualizado. Tal como lo muestra otro estudio de la misma AMR.

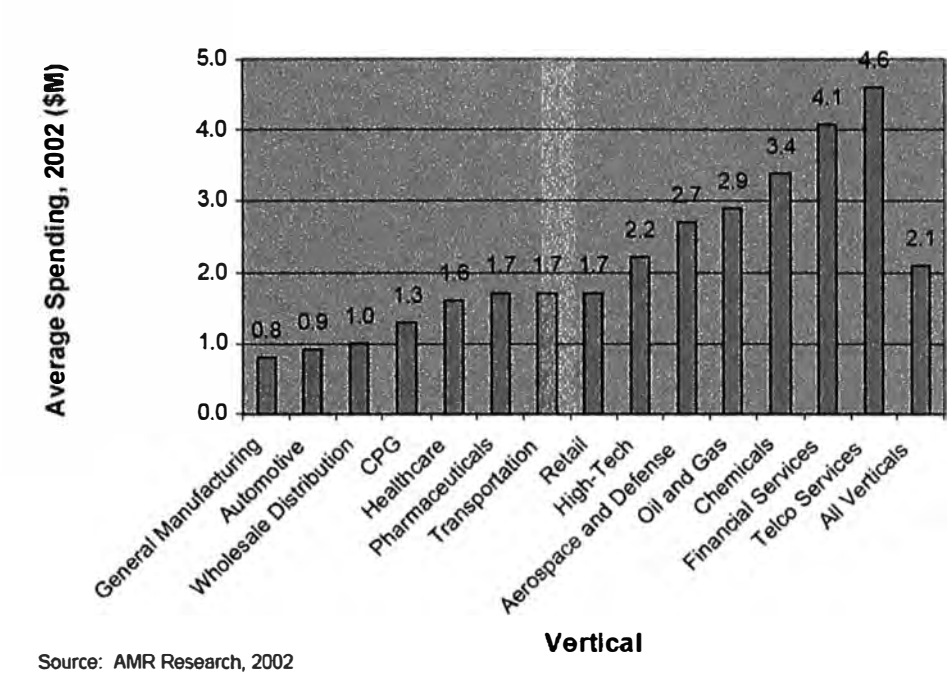


Figura 1.4: Gastos promedio en Business Continuity según industria

Adicionalmente vale la pena mencionar las siguientes estadísticas:

- Un incendio provoca el cierre de aproximadamente el 44 por ciento de los negocios afectados. 2 de 5 compañías sufrirá un siniestro en los próximos cinco años.[1.12]
- Compañías que experimentan corte de servicio por más de 10 días nunca se recuperarán[1.13]
- 80% de los negocios sin un bien estructurado plan fallarán en los próximos 12 meses.
- 90% de los negocios que pierden data fallarán en los próximos 2 años
- 43% de las compañías nunca se recuperarán de un desastre.
- En 2003, los desastres naturales tuvieron un costo de unos 60 mil millones de dólares a nivel mundial (fuente: Naciones Unidas). "Es más sencillo y económico construir una reputación desde cero que reparar una dañada" (fuente: Reputation Institute & Harris Interactive Inc.).

[1.12] Gartner/Dataquest
[1.13] London Chamber of Commerce & Industry, Information Centre Guide, May 2003

- Menos que el 20% del Global 2000 tienen un bien estructurado y actualizado Plan de BC, más aun menos del 10% de las empresas tienen una amplia y actual estrategia de Business Continuity.

Finalmente para justificar el costo o la inversión de un "Business Continuity Plan" es que este costará menos que el costo de tener los sistemas caídos en caso de desastre durante el periodo de tiempo que dure la caída.

1.3 Planteamiento del Problema

El Business Continuity Planning (BCP) es un proceso que prepara a una organización a seguir operando cuando un desastre ocurre. Este documento contiene los procedimientos y lineamientos generales con el fin de reestablecer las operaciones interrumpidas en un tiempo prudencial.

Este BCP tiene un ciclo de vida, según diversos autores (O´Hehir, 1998) (Hamilton, 1999) (Cornish, 1999) (Howe, 1999) (Craig, 2000). Las mejores prácticas de Business Continuity sugieren 10 áreas o fases[1.14]:

1. Inicio del Proyecto y Administración
2. Evaluación de Riesgos y Control
3. Análisis del Impacto al Negocio
4. Estrategias desarrollando el BC
5. Respuesta a Emergencias y Operaciones
6. Desarrollando y Planificando el BC
7. Conocimiento y Programas de entrenamiento
8. Manteniendo y Ejecución del BCP
9. Relaciones Públicas y Coordinación de Crisis
10. Coordinación con Autoridades Públicas

En la fase 4 El negocio debe reconocer y escoger la estrategia de replicación de información a los sites secundarios.

Una encuesta realizada por Gartner [1.15] a 120 CIOs y Directores de IT de sus respectivas empresas acerca de las diferentes estrategias usadas en sus compañías, los resultados arrojaron que no existe una única solución de un proveedor que satisfaga todas las necesidades de las empresas.

[1.14] Generally Accepted Practices for BC Practitioners by Disaster Recovery Journal and DRI International.
[1.15] Garnet ID: G00126421 Survey Confirm There are Many Effective Disaster Recovery Strategies.

A la pregunta Arquitectura que usan para replicación de data, se dieron 2 opciones:

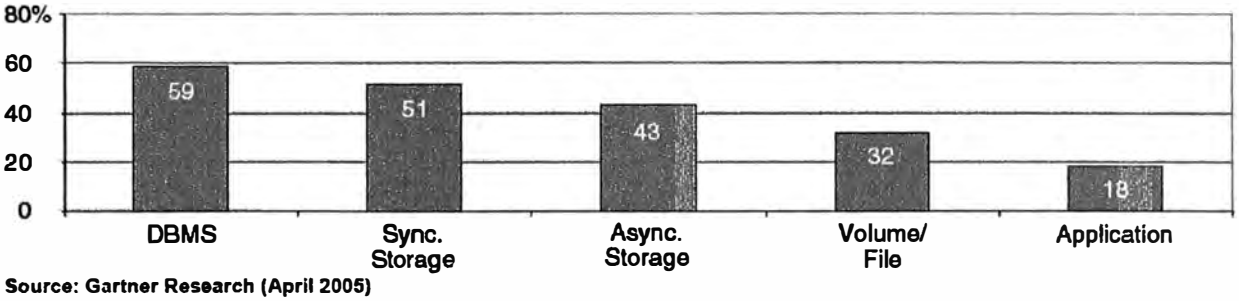


Figura 1.5: Arquitectura de replicación de data usada

Este estudio muestra que actualmente la más popular de las implementaciones son DBMS log-shipping (DataBase managament system). La segunda más popular son soluciones sincrónicas basadas en storage, seguidas por las asíncronas también basadas en storage.

A la pregunta si usa soluciones sincrónicas basadas en storage:

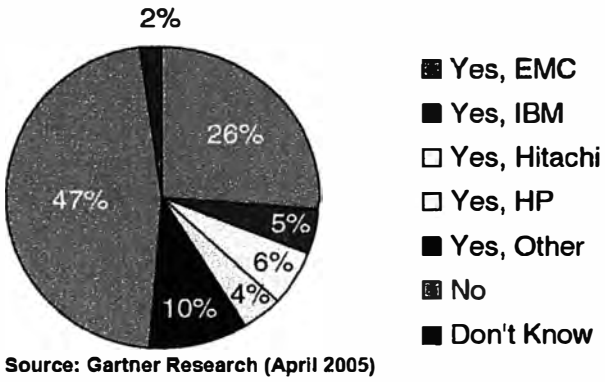
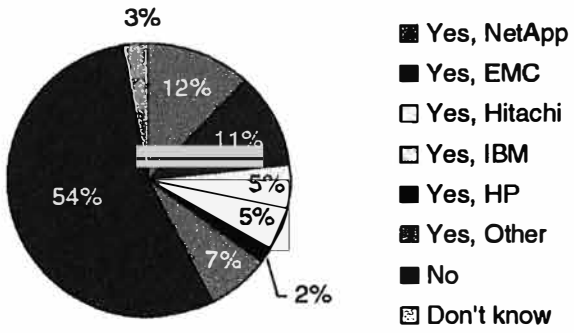


Figura 1.6: ¿Usa Arquitectura de replicación de data sincrónica?

Se encontró que el líder mundial en este tipo de soluciones es EMC Corporation[1.16]
A la pregunta si usa soluciones asíncronas basadas en storage:

[1.16] EMC Corporation: www.emc.com



Due to rounding, percentages total 99 percent.

Source: Gartner Research (April 2005)

Figura 1.7: ¿Usa Arquitectura de replicación de data asíncrona?

Adicionalmente un estudio estadístico de EMC encontró que de las replicaciones basadas en Storage (Sincrónicas o Asíncronas), el tipo de conectividad más común es Fibre Channel, aunque el protocolo GigabitEthernet viene creciendo en estos últimos años.

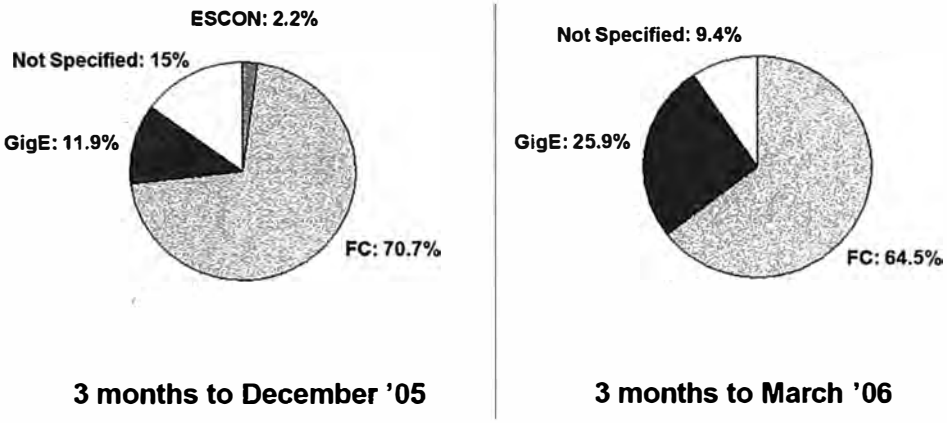


Figura 1.8: Distribución de protocolos en las implementaciones [1.17]

El presente informe se centrará básicamente en las fases 4 y 6 concernientes al área de replicación de data basada en Storage.

1.4 Objetivos de la investigación

Explicar los fundamentos teóricos y protocolos relacionados al almacenamiento externo con replicación remota para Continuidad de Negocios, listar y evaluar los parámetros y técnicas para el diseño y dimensionamiento de una solución de

[1.17] SRDF Best Practices – Extended Distance Networking – EMC World Boston 2006

almacenamiento replicado basada en storage que pueda brindar continuidad al negocio en caso de desastre.

1.5 Alcances

En el presente informe se expondrá las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre. Se explicarán los fundamentos teóricos y protocolos relacionados a la implementación de un almacenamiento externo con alta disponibilidad. Se realizará la evaluación de alternativas, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

Se expondrá un caso de estudio donde se mostrará las técnicas explicadas, se mostrarán estadísticas de entrada y mediante software de EMC se calcularán los recursos de hardware y transporte necesarios para el caso de estudio. Finalmente se presentaran conclusiones.

1.6 Limitaciones del Trabajo

El presente informe no pretende ser una guía para desarrollar un Business Continuity Planning, pues como se menciona solo abarcará la fase 4 de su desarrollo centrándose exclusivamente en calcular y dimensionar los recursos de Hardware, Software y Transporte necesarios para una solución de almacenamiento replicado basado en storage.

almacenamiento replicado basada en storage que pueda brindar continuidad al negocio en caso de desastre.

1.5 Alcances

En el presente informe se expondrá las necesidades de las empresas por contar con soluciones y datacenters de alta disponibilidad que puedan brindar continuidad al negocio en condiciones de desastre. Se explicarán los fundamentos teóricos y protocolos relacionados a la implementación de un almacenamiento externo con alta disponibilidad. Se realizará la evaluación de alternativas, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

Se expondrá un caso de estudio donde se mostrará las técnicas explicadas, se mostrarán estadísticas de entrada y mediante software de EMC se calcularán los recursos de hardware y transporte necesarios para el caso de estudio. Finalmente se presentaran conclusiones.

1.6 Limitaciones del Trabajo

El presente informe no pretende ser una guía para desarrollar un Business Continuity Planning, pues como se menciona solo abarcará la fase 4 de su desarrollo centrándose exclusivamente en calcular y dimensionar los recursos de Hardware, Software y Transporte necesarios para una solución de almacenamiento replicado basado en storage.

CAPÍTULO II
FUNDAMENTOS TEORICOS

En el presente capítulo se expondrá el marco teórico, los protocolos y tecnologías que hacen posible una implementación de un almacenamiento externo con alta disponibilidad.

2.1 Bases Teóricas

Para realizar un correcto diseño y dimensionar una solución de almacenamiento externo replicado para brindar continuidad de negocios es importante conocer a fondo los conceptos y funcionamiento de cada parte de la tecnología que hace posible este tipo de soluciones. Por ello, a continuación se detallan los conceptos y el funcionamiento de la tecnología envuelta con el propósito del informe.

2.1.1 Almacenamiento Externo

En la historia de los ordenadores se han usado varios métodos para el almacenamiento de datos. Al principio se recurrió a tarjetas perforadas. A continuación se pasó al soporte magnético, empezando por grandes rollos de cintas magnéticas abiertas.
Con la aparición de los discos magnéticos ésta técnica llegó a su sentido más amplio. En los discos es más sencillo acceder a cualquier punto de la superficie en poco tiempo, ya que se accede al punto de lectura y escritura usando dos coordenadas físicas. Por una parte la cabeza de lectura/escritura se puede mover en el sentido del radio del disco, y por otra el disco gira permanentemente, con lo que cualquier punto del disco pasa por la cabeza en un tiempo relativamente corto. Esto no pasa con las cintas, donde sólo hay una coordenada física.

Tradicionalmente los sistemas operativos se han comunicado con dispositivos de almacenamiento a través de canales, tales como bus paralelo, ESCON y SCSI. Éstas tecnologías de canal brindaron conexiones físicas fijas entre los sistemas operativos y sus

dispositivos periféricos. Debido a la gran integración entre los protocolos de transmisión y las interfaces físicas, éstas minimizan el overhead[2.1] requerido para establecer comunicación y transportar grandes cantidades de data a los dispositivos definidos estáticamente.

Las características de este tipo de conexión: Gran rendimiento, Bajo Overhead, Configuración física estática, Cortas distancias, Conectividad a un sistema simple

Con el tiempo se evaluaron otras alternativas como la tecnología de redes para comunicarse con dispositivos de almacenamiento, estas brindaban ciertas mejoras como la flexibilidad para acceder al almacenamiento pero a la vez tenían problemas, sus características: Bajo rendimiento, Alto Overhead de protocolo, Configuración Dinámica, Grandes distancias, Conectividad a través de diferentes sistemas.

Finalmente Fibre Channel recoge los mejores beneficios de ambos conexiones de canal y de redes, convirtiéndose de este modo Fibre Channel en la evolución natural de las conexiones directas SCSI, consiguiendo actualmente velocidades de transferencia de 400MB/s (comúnmente conocido como 4Gbps) [2.2]. Debido a esto es que es actualmente el protocolo por defecto usado para las conexiones a un almacenamiento externo es Fibre Channel.

En términos simples un almacenamiento externo o storage puede ser visto como una caja negra que entrega capacidad (en GB, GigaBytes normalmente) a los servidores Open Systems [2.3] conectados vía Fibre Channel en modo directo o vía switches Fibre Channel.

Un ejemplo sencillo y directo es que al termino de una implementación de Almacenamiento que se conecta a un servidor Windows, el servidor Windows descubra y utilice un disco G:\ de 300GB en su consola de Administración de discos provenientes del Storage, para las aplicaciones Windows que almacenaran información en este disco G:\ se vuelve completamente transparente que este disco no sea interno al servidor, si no más bien resida en almacenamiento externo o Storage como se le conoce en el medio.

[2.1] Overhead: Capacidad adicional o costo indirecto necesario para un fin.
[2.2] Esta por aprobarse la velocidad de 1600MB/s en Fibre Channel, ya existen vendedores ofreciéndolas.
[2.3] Servidor Open Systems: Servidor de Arquitectura abierta: Windows, IBM-AIX, HP-UX, Linux, VMWARE, etc.

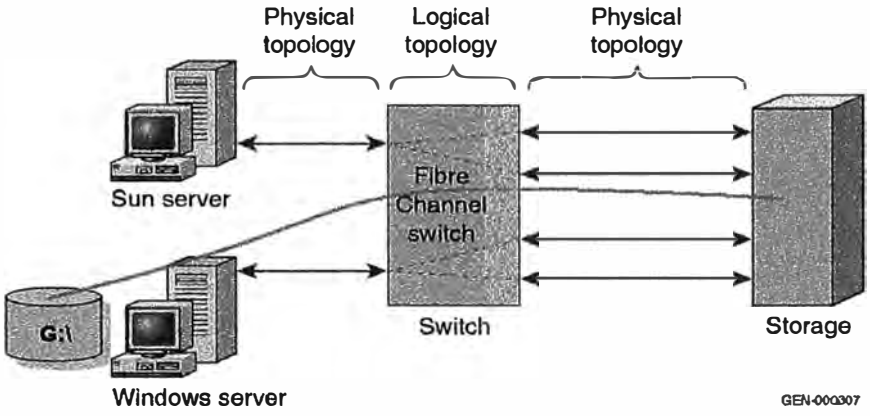


Figura 2.1: Ejemplo Servidor Windows Almacenamiento Externo

De este mismo ejemplo podemos extender la idea y dirigirnos al propósito del presente informe, consideremos que ahora poseemos un site secundario y que este primer storage tiene la capacidad de conectarse a un segundo storage del mismo fabricante de tal modo que pueda copiar (replicar) la información que el servidor Windows escribe y reescribe en su DriveLetter G:\ al segundo storage y con ello podríamos conectar un segundo servidor Windows con el fin de reestablecer las operaciones accediendo a la data copiada en este site alternativo en casos de desastre, ese es el concepto de replicación basada en storage discutida en el capitulo I.

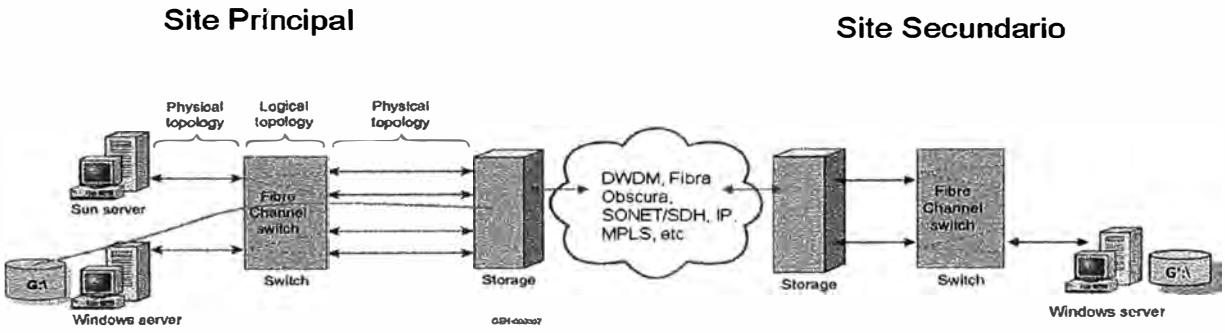


Figura 2.2: Replicación Basada en almacenamiento externo

Estos almacenamientos externos hacen uso de la memoria volátil de estado sólido denominado "cache" para todos los procesos de lectura/escritura solicitados por los servidores conectados a ellos. De esta manera logran reducir considerablemente el tiempo de respuesta de estas peticiones de lectura/escritura. La cache es el corazón de un sistema

de almacenamiento externo, y cada I/O [2.4] en el sistema de almacenamiento externo debe pasar por la cache, ya sea lectura o escritura. La cache sirve para varios propósitos, sin embargo desde el punto de vista del usuario este tiene principalmente 2 funciones.

Primero, La cache mantiene la data recientemente accedida disponible para lectura, los usuarios accederán a la data que ellos usaron más reciente. Los datos que no son tocados, son menos probables a ser accedidos de nuevo. Debido a este patrón la cache en el sistema de almacenamiento usa el algoritmo Least Recently Used (LRU) o más recientemente usado para determinar cual slot de cache reemplazar y cuales mantener.

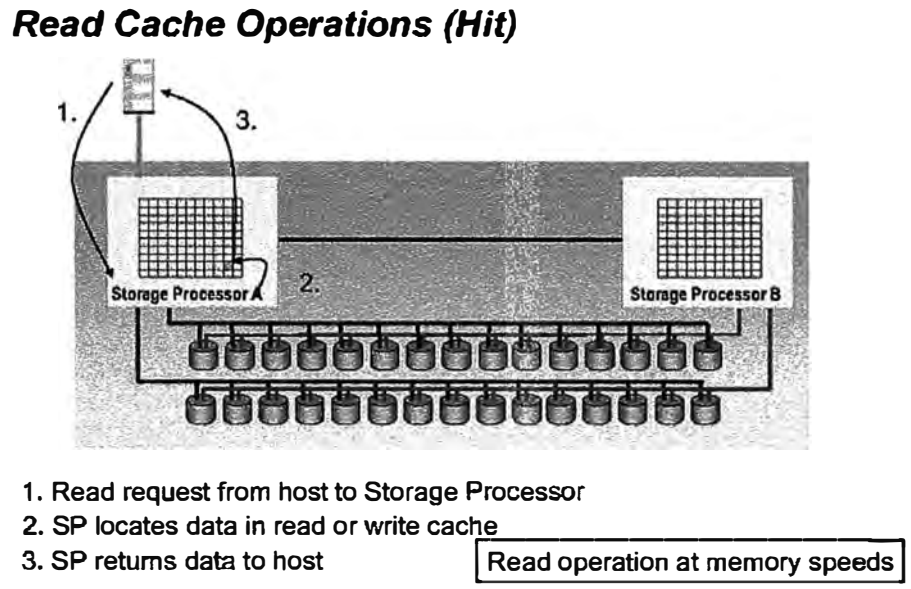
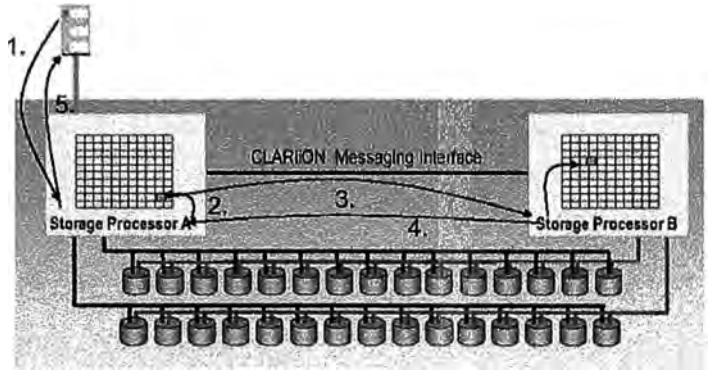


Figura 2.3: Operación de Lectura directamente desde cache a la velocidad de la memoria

Segundo, La cache es usada como una memoria temporal de escritura hasta que ellos puedan ser pasados a disco propiamente[2.5]. Cuando un servidor envía una solicitud de escritura ésta es almacenada en la cache y el servidor es notificado que la data ya está almacenada. Ésta data luego es "destaged" a disco como un proceso en background cuando el sistema se encuentre con baja utilización.

[2.4] I/O Solicitud de Lectura o Escritura desde el servidor.
[2.5] El proceso de pasar la data en la cache a disco físico es llamado "destage"

Write Cache Operations



1. Write request from host to SP-A
2. SP-A stores data in Cache
3. SP-A uses CMI to send data to SP-B
4. SP-B stores data in cache and sends complete to SP-A
5. SP-A sends complete to host

Figura 2.4: Operación de Escritura en Storage de 2 Controladoras

2.1.2 Modelos de Almacenamiento externo de EMC

EMC Corporation [2.6] el líder mundial en almacenamiento externo[2.7] tiene 2 líneas de almacenamiento externo: Clariion ® para el segmento mediano y Symmetrix ® [2.8] para el segmento alto.

2.1.2.a Clariion ®

La plataforma de almacenamiento externo #1 en el mercado mediano [2.9]. La serie CX3 es la última generación de los sistemas de almacenamiento Clariion. Cada generación adiciona mejoras de rendimiento, disponibilidad y escalabilidad, mientras la arquitectura de alta disponibilidad se mantiene constante.

Clariion fue el primer sistema de almacenamiento Fibre Channel en la industria en ofrecer una solución extremo a extremo de 4Gbps.

Clariion ofrece una arquitectura completamente redundante: fuentes, baterías, data paths o rutas de datos, operaciones como upgrade de hardware y software en línea. Permite el uso de software adicional como Snapview ® (copias locales), SANCopy ® (copias remotas a diferentes fabricantes) y MirrorView ® (copias remotas a otro Clariion con propósitos de Recuperación de Desastres y Continuidad de Negocios). Clariion permite escalar desde una configuración de 5 discos hasta 480 discos. Clariion permite combinar RAID1, 1/0, 3, 5 y tipos de discos SATA y Fibre Channel.

[2.6] www.emc.com
[2.7] Gartner Cuadrante Mágico 2007
[2.8] Clariion ®, Symmetrix ®, Snapview, SanCopy, MirrorView son marcas registradas de EMC Corporation.
[2.9] IDC Q2-2006

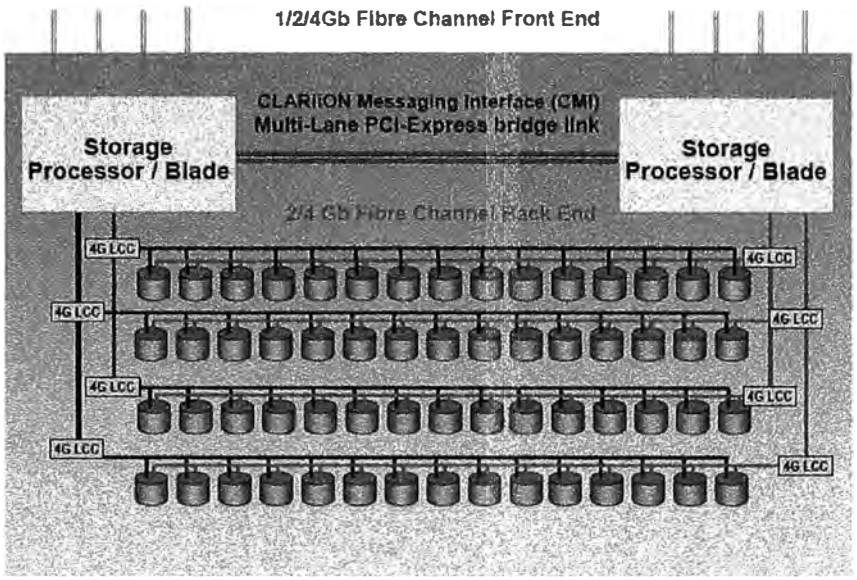


Figura 2.5: Arquitectura de Clariion

En la figura 2.5 se puede apreciar la arquitectura de Clariion que cuenta con 2 Tarjetas Procesadoras llamadas SP (Storage Processor) denominadas SPA y SPB. Cada uno de estas SPs cuenta según el modelo con: 1, 2 y 4 Procesadores; 1, 2, 4 y 8 GB de memoria Cache; 1, 2 y 4 puertos Back-End que sirven de conexión hacia las bandejas de discos; 1, 2 y 4 puertos Front-End para la conexión de los servidores; algunos modelos cuentan con puertos para la conexión iSCSI [2.10]

La arquitectura Clariion trabaja los SPs en forma activo-activo y que en el caso de falla de una de estas (la SP sobreviviente) puede asumir toda la carga del storage. En la figura 2.5 podemos visualizar que cada una de las bandejas de discos (DAEs o Disk Array Enclosures) aloja 15 discos físicos y que cada unos de estos DAEs puede ser accedido por ambos SPs.

En las figuras 2.3 y 2.4 explica exactamente como Clariion debido a su arquitectura de 2 tarjetas Controladoras SPs maneja los procesos de escritura espejada en cache y lectura por medio de la cache.

2.1.2.b Symmetrix ®

"El Storage más confiable del mundo" Actualmente se viene produciendo la serie DMX3 y DMX4.

[2.10] iSCSI, Es una tecnología de conectividad de bajo costo, la cual encapsula el protocolo SCSI sobre IP.

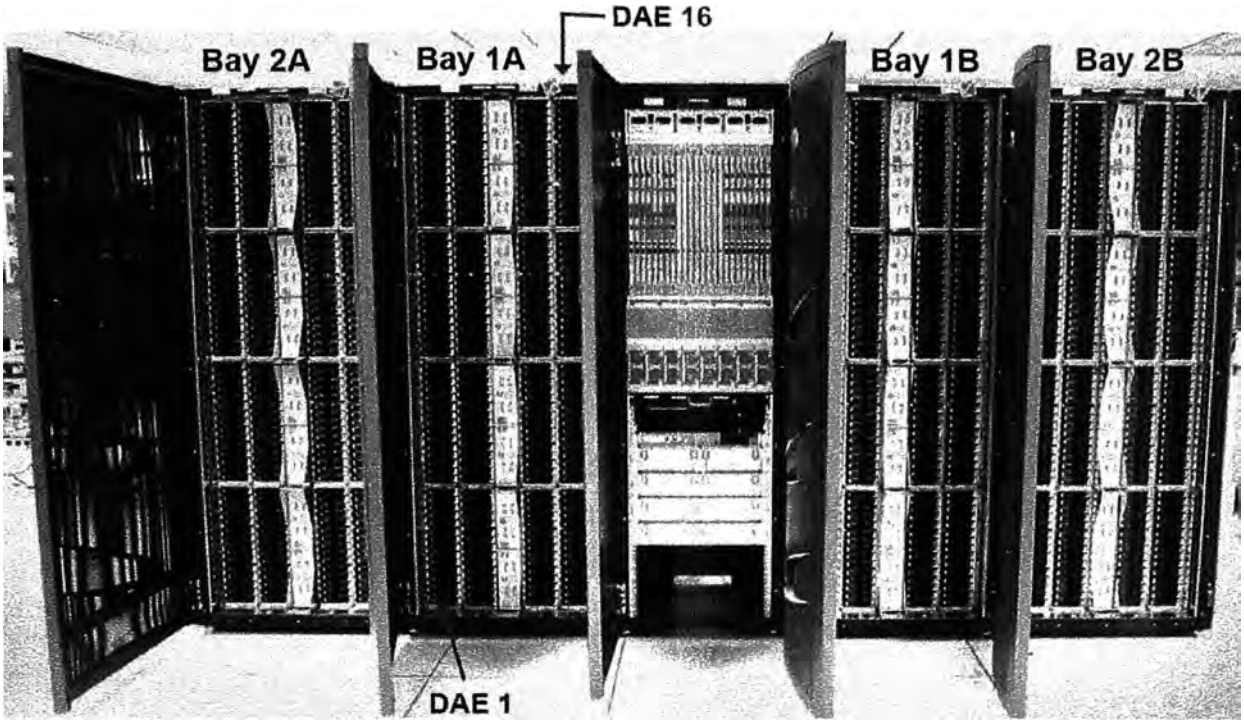


Figura 2.6: Vista delantera Del DMX3-4500

En la figura 2.6 se puede apreciar el DMX3 modelo 4500, el cual cuenta con 1 System Bay (Gabinete Central) y 4 Storage Bays (Gabinete laterales). El DMX3 soportará configuraciones de hasta 2400 discos y 160 DAEs. Cada Storage Bay puede contener hasta 16 DAEs con hasta 15 discos cada uno, para un total de 240 discos por Storage Bay
Cada Gabinete tiene 2 zonas de energia 2N redundantes

El System Bay aloja las baterías, las fuentes de poder, el servidor Service Processor donde corre el sistema operativo del Symmetrix llamado Enginuity, además tiene un midplane de 24 ranuras. Detallados como:

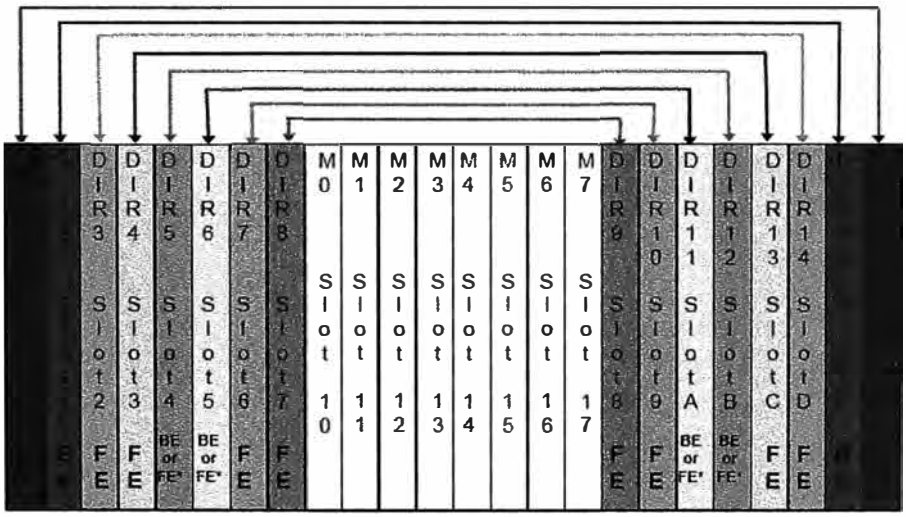


Figura 2.7: Midplane DMX, DMX2, DMX y DMX4

La figura 2.7 muestra que el DMX3 puede tener hasta 8 Directores de Discos (llamados también Back-End o BE, que son los que conectan los DAEs o Bandejas de Discos y por lo tanto los discos al sistema, el cual incluye a la cache). También muestra que se pueden tener hasta 12 Directores de Canal (llamados Front-End o FE, los cuales dan acceso a los servidores).

En el DMX3 se pueden configurar desde 4 hasta 8 tarjetas de memoria Cache, rotuladas del M0 al M7 en el midplane, éstas tarjetas vienen de distintos tamaños: 8GB, 16GB, 32GB, 64GB y son agregadas en parejas. El Storage DMX3 soporta hasta un máximo de 2 tipos de tamaño de cache en el mismo Storage. La cantidad de cache normalmente viene especificada por la cantidad de discos físicos que tendrá el storage. Sin embargo, existen procesos opcionales del storage que requerirán de más memoria cache (más tarjetas) como la replicación remota SRDF ® [2.11], el determinar cuanta cache adicional se necesitará es fundamental para que el SRDF funcione correctamente, este es uno de los objetivos del presente informe.

El Storage Bay contiene 16 DAEs, cada uno de éstos tiene capacidad de alojar 15 discos, resultando la capacidad del Storage Bay en 240 discos. Los puertos Directores de Discos o Back-End son conectados a estos DAEs vía sus tarjetas LCC (Link Controller Card) que viene en estos DAEs. Adicionalmente los Storage Bays consecutivos son cableados en cadena vía estas mismas LCCs.

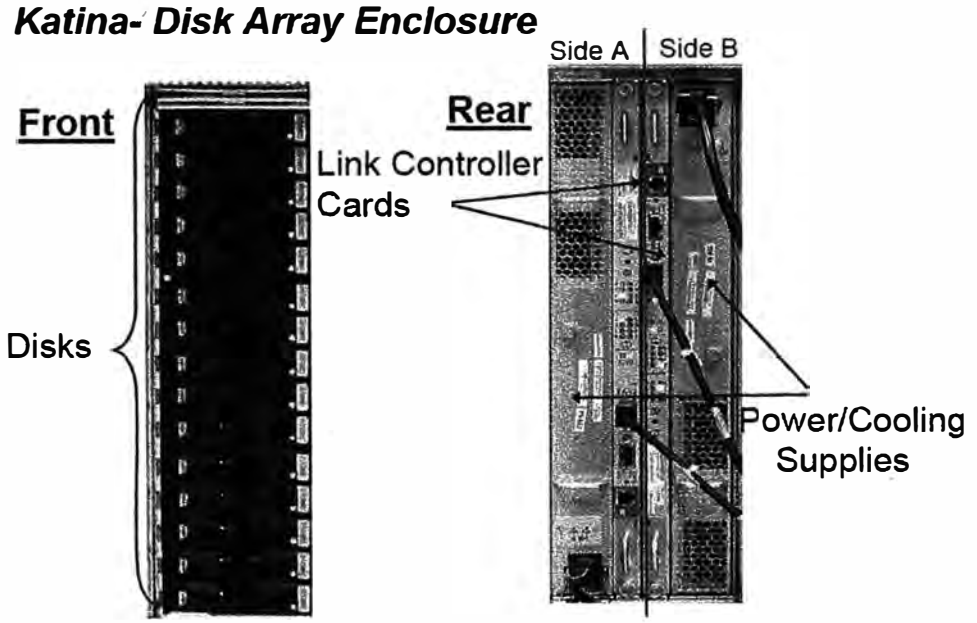


Figura 2.8: Katina Disk Array Enclosure

[2.11] SRDF ® marca registrada de EMC Corporation

La cache tiene 2 propósitos principales: Mantener la data recientemente accedida disponible para lectura desde cache y la otra para ser usada como una memoria temporal de escritura hasta la data pueda ser pasada a disco propiamente, proceso en background conocido como "destage". DMX3 permite alta disponibilidad y velocidad de acceso a los datos debido a que cuenta en el midplane con una matriz de conexión todos contra todos.

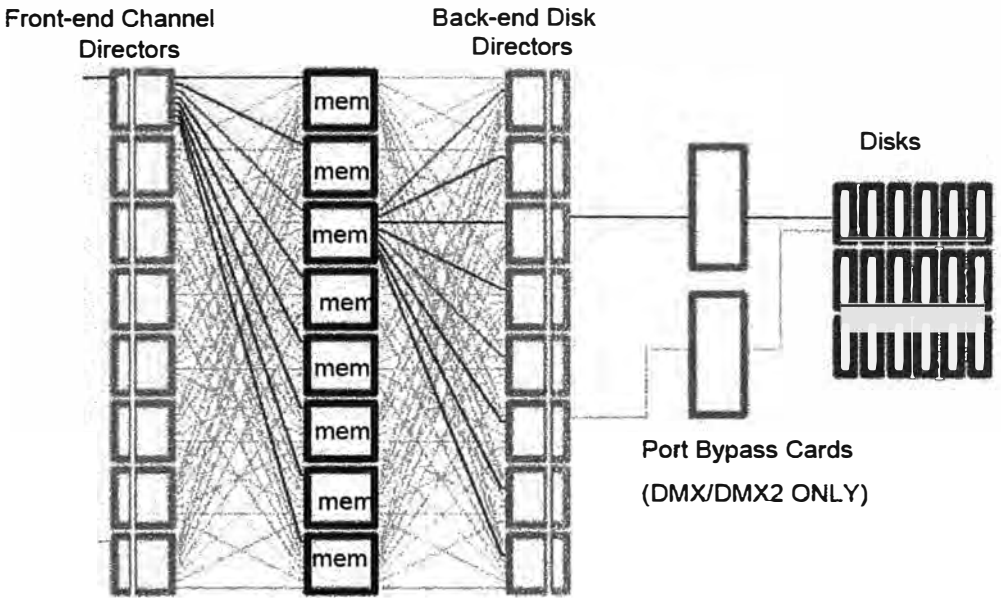


Figura 2.9: Diagrama Funcional de las series DMX, DMX2, DMX3 y DMX4

Es decir las 8 posiciones de tarjetas para memoria cache tienen conexiones directas hacia los servidores o puertos Front-End y conexiones directas hacia los DAEs (LCCs) o discos o puertos Back-End como se muestra en la figura 2.9

En el nivel más bajo, la data existe en discos físicos y en el sistema. En el nivel de data existen 2 arquitecturas: FBA usado para Open Systems y CKD usado por Mainframes. El presente informe se centrará en el formato FBA, pues analizaremos la recuperación de desastres para Sistemas abiertos u Open Systems.

La estructura de la data FBA es organizada jerárquicamente desde bloques pequeños llamados "sectors". La mayoría de la gente asociará el término sector con un bloque de 512-bytes en disco. Sin embargo, en Symmetrix un sector es formado por 16 bloques de 512-bytes dando un total de tamaño de sector de 8K. Un sector es el I/O más pequeño que el sistema puede procesar y consecuentemente el bloque más pequeño que tiene CRC para la revisión de la integridad de la data. Si la petición de escritura desde el servidor es solo una parte de este sector, el DA lee el resto del sector desde el disco y

recalcula el CRC. Tracks son formados con 8 sectores, para un total de 64K, Finalmente un cilindro es 15 tracks, haciendo 960K. Figura 2.10 muestra esta jerarquía.

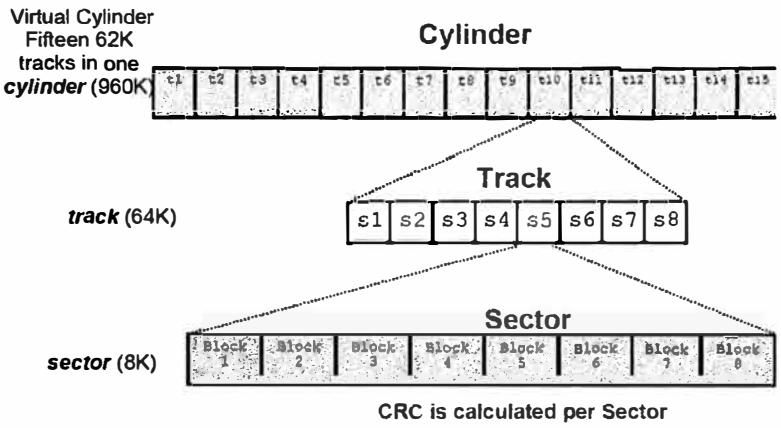


Figura 2.10: Jerarquía de datos DMX3

En el sistema toda la data en la cache debe ser del mismo tamaño básico denominado slot de cache, el cual es de 64K para Open Systems

Actualmente un DMX3 completamente cargado discos y tarjetas directores back-end, front-end y tarjetas de memoria cache puede albergar 1920 discos de capacidad en 9 racks (1 System Bay y 8 Storage Bays), luce de la siguiente manera:

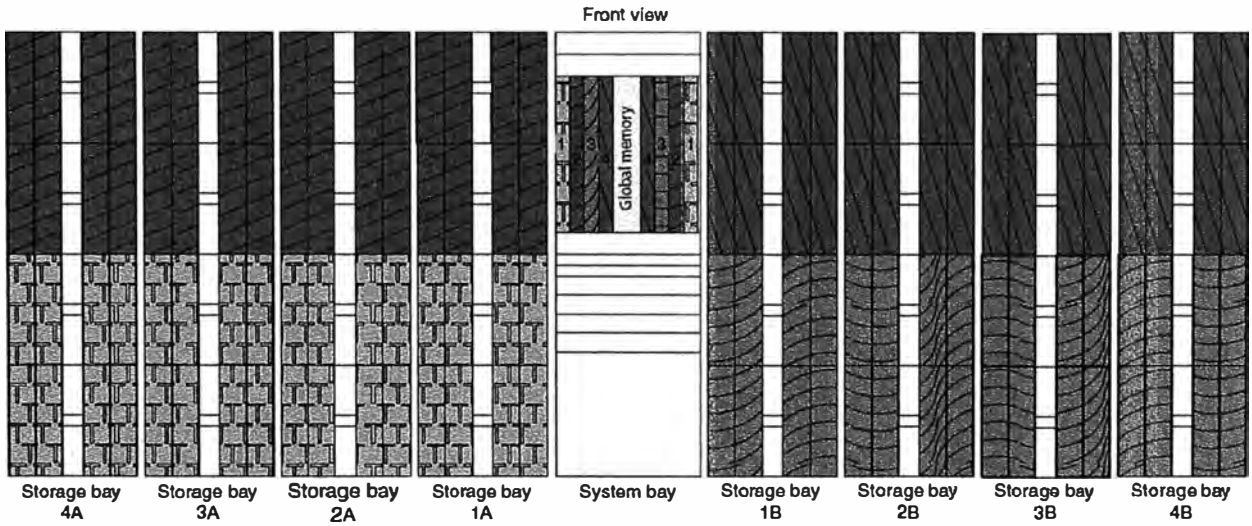


Figura 2.11: DMX3, 4 DA Pair, 1920 discos de capacidad.

El DMX3 define cuadrantes manejados por un par de directores back-end. Según la figura 2.11 el Par de Directores back-end de color amarillo maneja todos los discos en los DAEs de color amarillo, de la misma manera con los otros colores.

2.1.3 Fibre Channel

El Canal de fibra, del inglés Fibre Channel, es una tecnología de red utilizada principalmente para redes de almacenamiento, disponible primero a la velocidad de 1Gbps y posteriormente a 2, 4, 8 y 10Gbps.

El Canal de fibra está estandarizado por el Comité Técnico T11 del Comité Internacional para Estándares de Tecnologías de la Información, comité acreditado por el Instituto de Estándares Nacional Americano.

Nació para ser utilizado principalmente en el campo de la supercomputación, pero se ha convertido en el tipo de conexión estándar para redes de almacenamiento en el ámbito almacenamiento empresarial. A pesar de su nombre, la señalización del Canal de Fibra puede funcionar tanto sobre pares de cobre, como sobre cables de fibra óptica.

El Canal de fibra fue especialmente interesante para simplificar las conexiones y aumentar las distancias, más que para aumentar la velocidad. Más tarde amplió su aplicación al almacenamiento en disco SCSI, permitiendo velocidades más elevadas y un número mucho más elevado de dispositivos.

También aportó soporte para un número elevado de protocolos de nivel superior, incluyendo SCSI, ATM e IP, siendo para SCSI (SCSI-3) su uso más frecuente (FCP).

Capas Fibre Channel

Similar al protocolo OSI Fibre Channel define capas

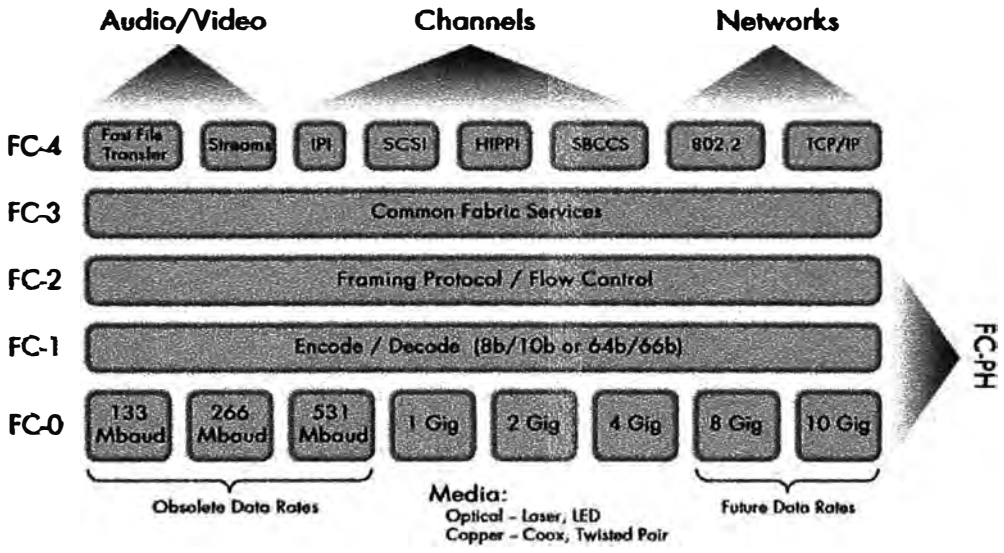


Figura 2.12: Capas Protocolo Fibre Channel

- FC0 La capa física, que incluye los cables, la óptica de la fibra, conectores, velocidades de TX y comercialmente en la actualidad tenemos 4Gbps, etc.

- FC1 La capa de enlace de datos, que implementa la codificación y decodificación de las señales 8b/10b, esto significa que de cada 10 bits transmitidos 8 son de data. Para las velocidades de 8Gbps y 10Gbps se cambia el esquema a 64/66b
- FC2 La capa de red, definida por el estándar FC-PI-2, que constituye el núcleo de Fibre Channel y define los protocolos principales: métodos de trama, métodos de secuencia de tramas y control de flujo.
- FC3 La capa de servicios comunes, una fina capa que puede implementar funciones como el cifrado o RAID.
- FC4 La capa de mapeo de protocolo, en la que otros protocolos superiores, como SCSI, se encapsulan en unidades de información que se entregan a la capa FC2. Comúnmente estos protocolos son SCSI-3 Serial e IP

Topologías del Canal de fibra, un enlace en el Canal de Fibra consiste en dos fibras unidireccionales que transmiten en direcciones opuestas. Cada fibra está unida a un puerto transmisor (TX) y a un puerto receptor (RX). Dependiendo de las conexiones entre los diferentes elementos, podemos distinguir tres topologías de Canal de fibra principales:

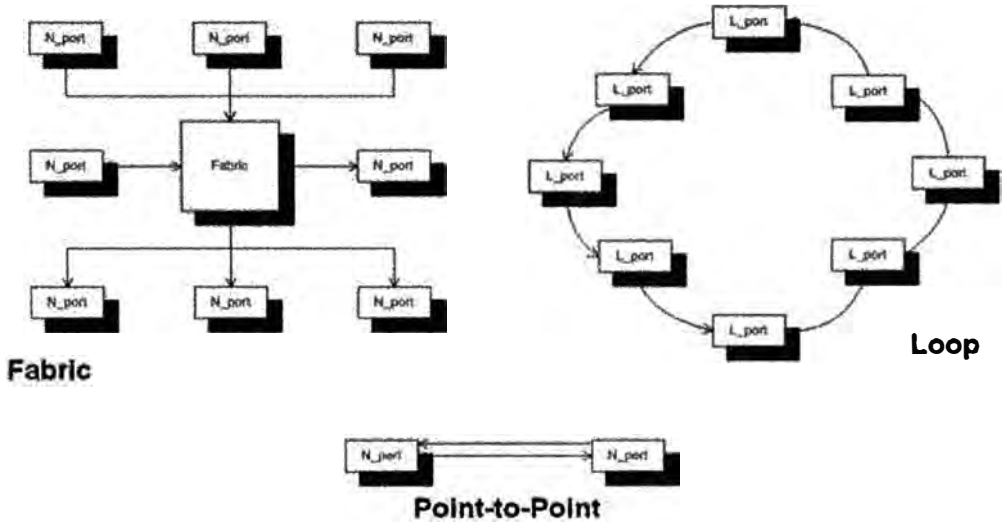


Figura 2.13: Topologías en Fibre Channel

- **Punto a punto (FC-P2P),** Dos dispositivos se conectan el uno al otro directamente. Es la topología más simple, con conectividad limitada a dos elementos.
- **Anillo arbitrado (FC-AL),** En este diseño, todos los dispositivos están en un bucle o anillo, similar a una red Token ring.
- **Medio conmutado (FC-SW),** Todos los dispositivos o bucles de dispositivos se conectan a conmutadores (switches) de Canal de fibra, conceptualmente similares a

las modernas implementaciones Ethernet. Los conmutadores controlan el estado del medio físico, proporcionando interconexiones optimizadas.

Direcciones Fibre Channel, Se usan 2 tipos de direcciones análogos al mundo IP sobre Ethernet.

- Dirección Fibre Channel, usado para enrutamiento de tramas desde el origen al destino, direcciones de 24 bits, son asignadas dinámicamente cuando los puertos Fibre Channel se registran a la Fabric.
- Dirección World Wide Name (wwn), Son grabadas desde fabrica [2.12] por los fabricantes de HBAs, Switches y Storage. Esta dirección es de 64 bits y esta compuesta por el wwpn y wwnn

Formato de Trama Fibre Channel

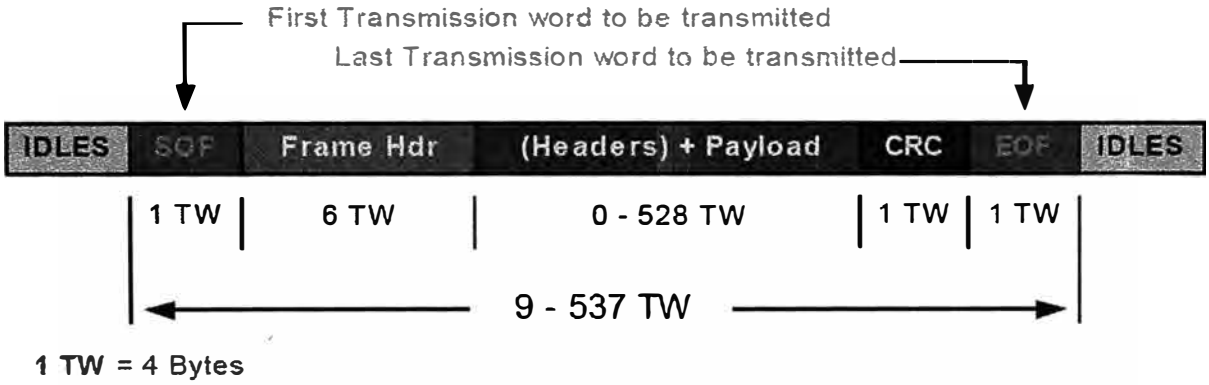


Figura 2.14: Estructura de trama Fibre Channel

El SOF es un indicador del inicio de la trama, la trama puede tener una longitud desde 9 TW hasta 537 TW = 537 x 4B = 2148 Bytes.

El CRC como siempre viene a ser la suma de paridad de toda la trama.

[2.12] Análogas a la MAC del mundo Ethernet

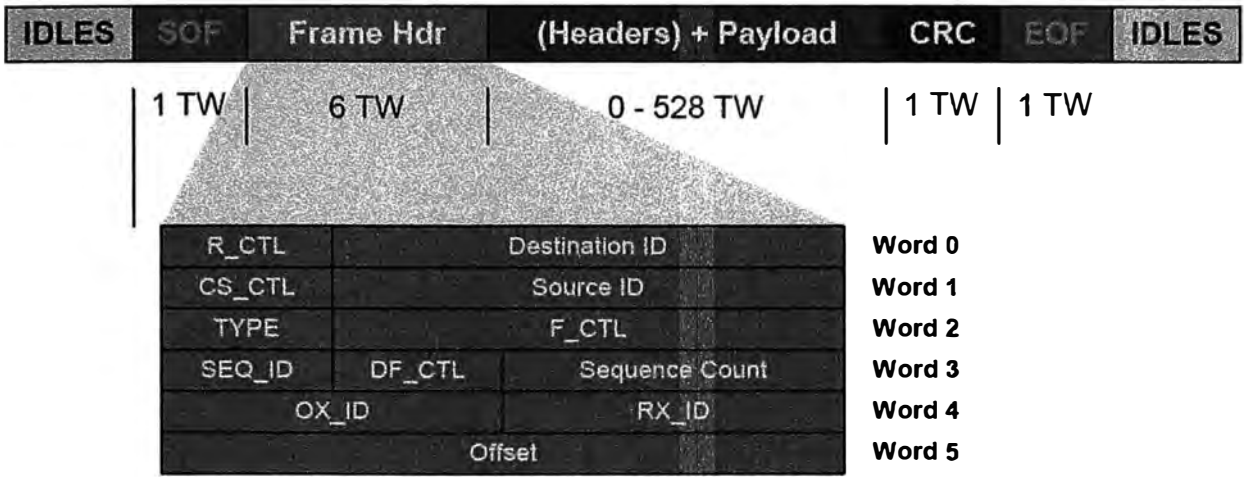


Figura 2.15: Cabecera de la trama Fibre Channel

- "Routing Control (R_CTL)", campo para indicar requerimientos de enrutamiento
- "Destination ID (D_ID)", Dirección Fibre channel destino
- "Source ID (S_ID)", Dirección Fibre channel origen
- "Class Specific Control (CS_CTL)", Campo Clase de servicio
- "Frame Type (TYPE)", Para definir el tipo de frame, si es de data para indicar el tipo de carga de protocolo de nivel superior, una valor de 08 indica FCP (SCSI sobre FC)
- "Frame Control field (F_CTL)", Control de intercambio, control de secuencia, y política de reconocimiento.
- "Sequence ID (SEQ_ID)", usado para rastrear todas las tramas en una secuencia particular entre 2 puertos.
- "Data Field Control (DF_CTL)", Define si existirán cabeceras adicionales.
- "Sequence Count (SEQ_CNT)", indica el orden secuencial dentro de una trama dentro de una secuencia.
- "Originator Exchange ID (OX_ID)", Usado para identificar todas las tramas que son parte de una intercambio.
- "Responder Exchange ID (RX_ID)", Usado para identificar tramas que son parte de intercambio especifico.
- "Offset/parameter field", Tiene diferentes usos dependiendo del tipo de frame.

Host Bus Adapters (HBA) en Fibre Channel, para la conexión de los principales sistemas, arquitecturas de ordenador y buses se necesitan estas tarjetas, disponibles en PCI, PCI-X y recientemente PCI Express. Cada HBA tiene un identificador único (World Wide Name), similar a la dirección MAC en Ethernet, utiliza un identificador único repartido por rangos entre los fabricantes (reparto realizado por IEEE), y que le sirve al switch del Fibre Channel para identificar las tarjetas (HBA) que tiene conectadas. Sin embargo, los WWNs son más largos (8 bytes). Además, se distinguen dos tipos de WWNs en un HBA: WWN de nodo, compartido por todos los puertos de un adaptador de host, y un WWN de puerto, único para cada puerto. Ejemplo de fabricantes de HBAs: Emulex, LSI Logic, QLogic

Figura 2.16: HBA

2.1.4 FICON

FICON (Fibre Connectivity) es un canal de I/O diseñado para soportar baja latencia, conexiones de alta ancho de banda entre mainframes de IBM y un Storage Controller (Controlador de Almacenamiento). FICON fue realizada sobre la tecnología Fibre Channel, compartiendo los niveles bajos, incluyendo FC-0 (Interfaz física), FC-1 (8b/10b Codec), FC-2 (Entramado y señalización) y FC-3 (Servicios Comunes). FICON es un tipo de FC-4. Su lugar en la arquitectura Fibre Channel es análogo a FCP, el cual es un tipo FC-4 usado para Open Systems. FCP fue diseñado para soportar SCSI.

FICON fue diseñado para remplazar ESCON y soportar sistemas de almacenamiento mainframe con formato CKD.

FICON es desarrollado y mantenido por el comité técnico T11 de la "International Committee for Information Technology Standards". Borradores y otra documentación técnica puede ser encontrada en www.t11.org bajo FC-SB

Attribute	FICON	ESCON
Link rate	212 MB/s	20 MB/s
Effective max data rate	200 MB/s	17 MB/s
Duplex	Full duplex	Half duplex
Type of switching	Packet switching with frame-by-frame routing using FSPF and Classes 2 and 3	Virtual Circuits (Class 1)
Multiple concurrent I/O	Yes	No
Maximum distance without droop	Depends on BB_Credit and link speed (typically 100 km at 100 MB/s).	9 km

Tabla 2.1: FICON comparado a ESCON

2.1.5 SAN, DAS

SAN (Storage Area Network) es una Tecnología que resuelve 2 problemas de conectividad de storage:

-Conectividad Host-to-Storage, un Servidor puede acceder y usar almacenamiento provisto por este storage.

-Conectividad Storage-to-Storage, para replicación de data entre arreglos de almacenamiento, el propósito del presente informe.

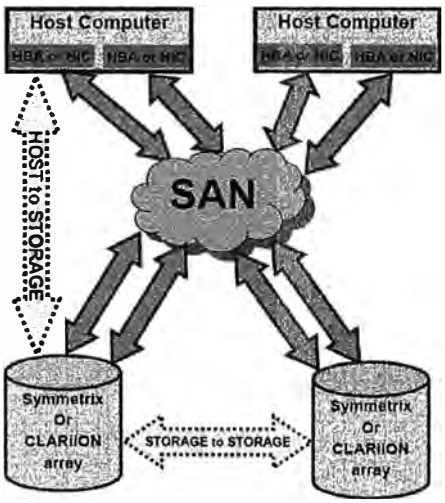


Figura 2.17: FICON comparado a ESCON

La tecnología SAN usa protocolos de I/O a nivel de bloques (block-level) a diferencia de las redes NAS que usa protocolos de I/O a nivel de archivos (file-level).

Con tecnología SAN el servidor recibe los dispositivos de almacenamiento en formato raw o crudo, justamente como los almacenamientos directamente conectados tradicionales. Con SAN el servidor puede construir cualquier filesystem nativo al Sistema Operativo en cualquiera de estos dispositivos raw presentados.

Los dispositivos de conectividad de una SAN soportan múltiples protocolos como: Fibre Channel, iSCSI, FCIP, iFCP. Estos 2 últimos son usados para hacer extensiones a la Fabric.

Los tipos de dispositivos de conectividad Fibre Channel son switches y directores, también están los Router multiprotocolo y Gateways. Los tipos de switches son Departamentales (hasta 32 puertos) y los Directores (Modulares desde 32 hasta 384 puertos [2.13]).

Una Fabric es un espacio virtual usados por los nodos para comunicarse entre si, esta Fabric está compuesta por un switch o un grupo de switches interconectados, La Fabric brinda servicios para administrar la comunicación entre los nodos.

Los switches Fibre Channel similarmente a los del mundo Ethernet necesitan una configuración para optimizar las comunicaciones entre los nodos conectados a este switch. Esta configuración para unir 2 nodos conectados a un switch se llama zonas y el conjunto de zonas activas en una Fabric se denomina ZoneSet. El concepto de zoning es análogo al de VLANs en el mundo Ethernet. Existen Zonas por puertos y por wwn, siendo esta ultima la recomendada para las implementaciones.

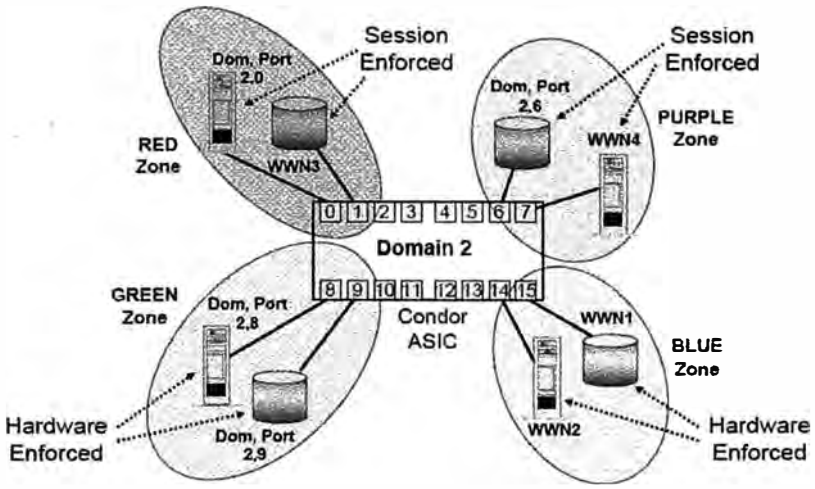


Figura 2.18: Concepto de Zonas en un switch

DAS (Direct-Attach Storage), Es la arquitectura propiamente dicha para la conexión host-to-storage. Posee un canal dedicado, transporte paralelo, ejemplos tenemos: SCSI y ESCON. Sus ventajas: Tiene bajo overhead, alta Tazas de transferencia. Sin embargo sus

[2.13] Director EMC Connectrix ED-48000B®, es una marca registrada de EMC.

desventajas son: Configuración estática, distancia limitada, limitaciones de topología, limitaciones de escalabilidad

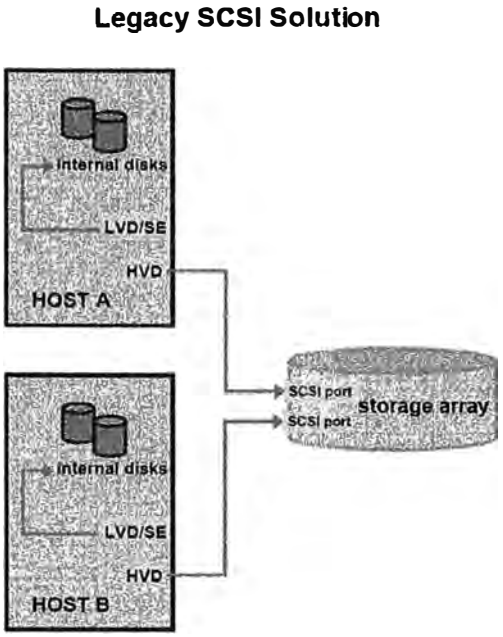


Figura 2.19: Solución SCSI

2.1.6 Gigabit Ethernet, IP, MPLS

GibabitEthernet, GbE es la evolución de la tecnología Ethernet y FastEthernet usada para la comunicación de ordenadores en redes LAN. GibabitEthernet ofrece la velocidad de transferencia de 1024Mbps (1Gbps). GigabitEthernet es definido por la especificación IEEE 802.3z, la cual fue liberada en Junio de 1998.

Gibabitethernet es implementado en Fibra Óptica, pero también puede ser implementado en pares trenzados como el UTP CAT6.

Esta tecnología al igual que su antecesor Ethernet es una tecnología que reside en la capa 1 y 2 del modelo OSI.

IP, El TCP/IP es la base del Internet. TCP/IP establece su propio modelo de capas, similar al estándar OSI, en ambos modelos en la capa 3 se ubica el protocolo IP el cual sirve para direccionar un nodo IP en el mundo de tal modo que se puedan enlazar computadoras que utilizan diferentes sistemas operativos, incluyendo PC, mini computadoras y computadoras centrales sobre redes de área local y área extensa e incluso a nivel mundial, ésa es la definición del Internet.

MPLS, Se podría denominar una tecnología hibrida de transporte y enrutamiento, pues rescata las bondades de ATM e IP obteniendo con ello reducir el tiempo de procesamiento de cada paquete que atraviesa por la red de transporte de un operador. En síntesis MPLS combina: Capa2, define intercambio de etiquetas y Capa3, Enrutamiento en la capa de red.
MPLS ofrece QoS, Ingeniería de Trafico, VPNs y soporte para multiprotocolo. De este modo los operadores de telecomunicaciones han implementado estas redes como una evolución natural de ATM o simplemente para contar con una red que permita VPNs y con ello poder brindar servicios de datos IP que permitan a sus clientes conectar varios sites entre si a nivel IP.

2.1.7 DWDM

DWDM es un método de multiplexación muy similar a la Multiplexación por división de frecuencia que se utiliza en medios de transmisión electromagnéticos. Varias señales portadoras (ópticas) se transmiten por una única fibra óptica utilizando distintas longitudes de onda cada una de ellas. De esta manera se puede multiplicar el ancho de banda efectivo de la fibra óptica, así como facilitar comunicaciones bidireccionales. Se trata de una técnica de transmisión muy atractiva para las operadoras de telecomunicaciones ya que les permite aumentar su capacidad sin tender más cables ni abrir zanjas. Para transmitir mediante DWDM es necesario dos dispositivos complementarios: un multiplexador en lado transmisor y un demultiplexador en el lado receptor.
Usando DWDM varias longitudes de onda separadas[2.14] (o canales, un diferente color, denominados lamba, simbolizados como "λ") pueden ser multiplexados en una luz multicolor transmitidos sobre una simple fibra óptica. Cada longitud de onda puede transportar una señal a cualquier "bit rate" [2.15] menor que el límite superior definido por la electrónica, típicamente hasta varios Gigabits/s
Diferentes formatos de data a diferentes "bit rate" pueden ser transmitidos juntos. Específicamente data IP, ESCON SRDF, Fibre Channel SRDF, SONET y ATM pueden todos viajar dentro de la misma fibra óptica.
Los sistemas DWDM son independientes del formato del protocolo y no existe impacto de performance introducidos por el sistema DWDM así mismo.

[2.14] DWDM puede multiplexar hasta 160 señales o λ, fuente wikipedia
[2.15] "bit rate", es la tasa de transferencia medida usualmente en Megabits/s, GigaBits/s

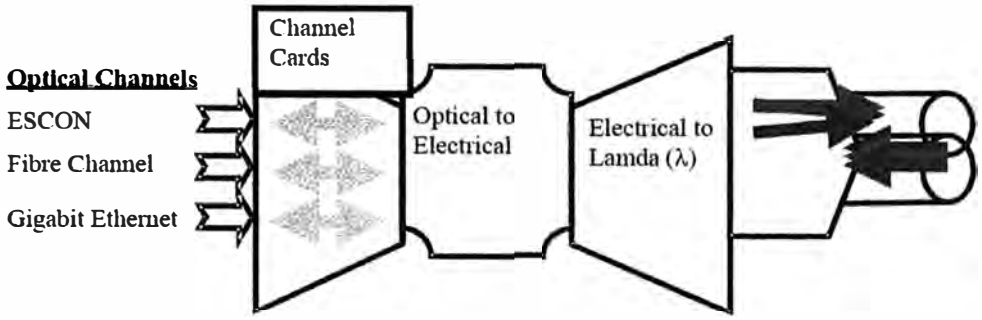
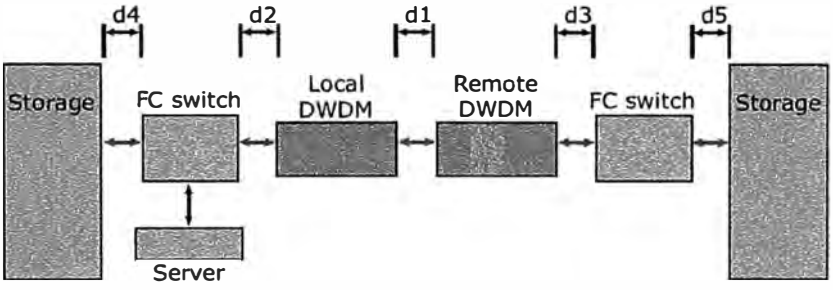


Figura 2.20: Concepto Tecnología DWDM

La figura 2.20 muestra que en el mundo del almacenamiento múltiples canales de replicación como SRDF y Fibre Channel ISL (Inter Switch Links) [2.16] pueden ser transferidos sobre un mismo par de enlaces de fibra. Estos es importante para clientes que arriendan una fibra oscura, pues con DWDM pueden ahora pasar más canales de data. Con las tecnologías actuales la capacidad de un simple par de fibras ópticas es virtualmente ilimitada, la limitación viene en el DWDM por la conversión Óptico-eléctrico.



d1 = DWDM signal over dark fiber medium.
d2 and d3 = Local ISL connections between switches and DWDM input. Can be SM or MM depending on DWDM and switch interfaces or local distance requirements.
d4 and d5 = Local storage or server connections into the fabric.

Figura 2.21: Concepto General de extensión Fibre Channel usando DWDM

La figura 2.21 muestra el concepto por el cual los clientes pueden extender sus SAN a cientos de Km. mediante DWDM, la cual es una de las soluciones típicas para soluciones de replicación para Recuperación de Desastres y Continuidad de Negocios.

Topologías DWDM disponibles incluyen point-to-point, lineales y anillos con esquemas de protección y sin protección ante fallas en enlaces entre dispositivos

[2.16] ISL, Es la conexión entre 2 switches Fibre Channel.

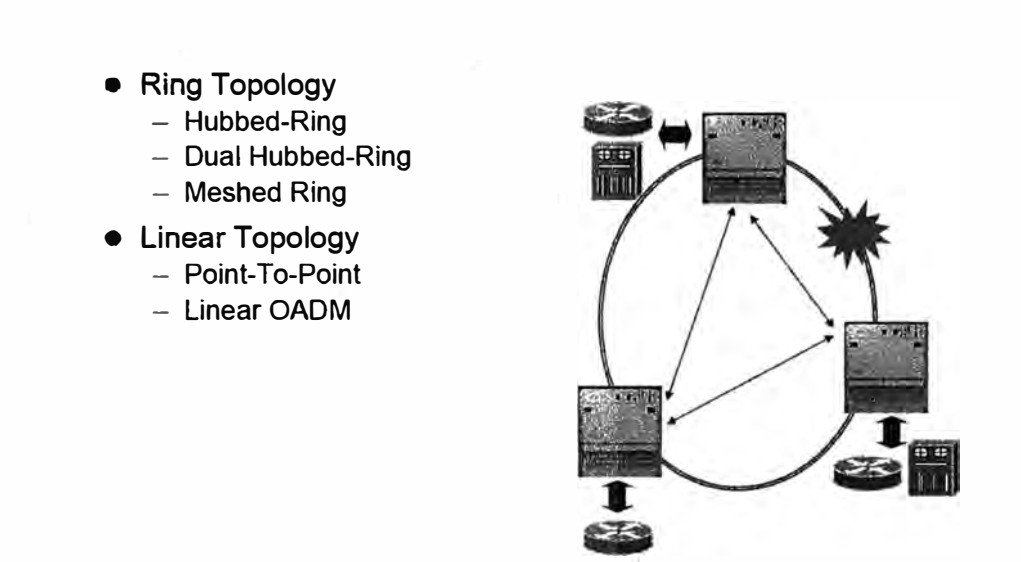


Figura 2.22: Topologías DWDM

Fabricantes de equipos DWDM homologados para una solución de replicación SRDF son: Nortel, Akara (Ciena), Lucent, Cisco, Huawei y Alcatel.

2.1.8 SONET/SDH

SONET (Synchronous Optical NETwork), es un estándar para transporte de telecomunicaciones ópticas, desarrollado por la ANSI. SONET define una tecnología para señales de diferentes capacidades de transporte a través de la red sincrónica óptica. El estándar define un octeto interpolado multiplexado ocupando la capa física del modelo OSI. La Sincronización es provista por un elemento principal en la red con un reloj muy estable (Stratum3), el cual es transmitido a través de sus enlaces OC-N, este reloj es luego usado por otros elementos de red para sus relojes.

SONET es generalmente usado en SAN para consolidar múltiples canales de baja frecuencia (Cliente ESCON y 1, 2, 4 Gbps Fibre Channel) en una sola conexión de alta velocidad. Esto puede reducir los requerimientos de longitudes de onda DWDM en una SAN extendida.

Esta también puede ofrecer soluciones de conexión a grandes distancias brindadas por las empresas de telecomunicaciones a un menor costo. En comparación a soluciones DWDM y más aun Fibras Oscuras.

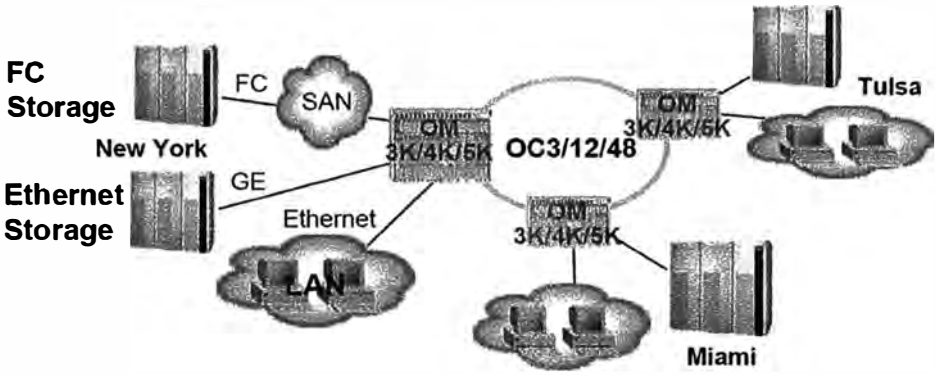


Figura 2.23: Storage sobre SONET/SDH

SDH es el estándar mundial, y como tal SONET puede ser considerado como un subconjunto de SDH. ANSI TDM combina 24 DS0s (64Kbps) en una señal DS1 de 1.54Mbps. ITU TDM combina 32 canales E0s (64Kbps) en una señal E1 de 2.048 Mbps, estas diferencias no permitían la compatibilidad entre estos estándares. El acuerdo alcanzado especifica una tasa de transmisión básica de 52Mbps (STS-1) para SONET y una tasa de transmisión básica de 155Mbps (STM-1, para SDH).

STS	Optical Carrier	Optical Carrier Rate (Mb/s)
STS-1	OC-1	51.840
STS-3	OC-3	155.520
STS-12	OC-12	622.080
STS-48	OC-48	2488.320
STS-192	OC-192	9953.280

Tabla 2.2: SONET/SDH Velocidades de Transmisión.

2.1.9 Definición de Términos

Latencia, Es el tiempo que demora en viajar una señal por el medio de transmisión.

Niveles de Protección RAID (Redundant Array of Independent Disks), Es un mecanismo utilizado en los sistemas de almacenamiento que usa múltiples discos duros donde distribuye la data y generalmente realiza una suma paridad para de esa manera este grupo de discos ofrezca cierta tolerancia a la falla de un disco o múltiples discos sin perdida de data, mayor integridad, mayor rendimiento y mayor capacidad. Existen diversos

niveles de protección siendo los más comunes: RAID0, RAID1, RAID3, RAID5, RAID6, etc.[2.17] Cada uno de los cuales ofrecen diferentes características mencionadas

[2.17] http://es.wikipedia.org/wiki/RAID

CAPÍTULO III
DESARROLLO DE LA PROBLEMATICA

En el presente capítulo, se realizará la evaluación de alternativas, opciones de conectividad, métodos de diseño y dimensionamiento de soluciones de almacenamiento replicado.

3.1 Alternativas de Solución

Como se vio en el Capitulo I, replicaciones basadas en storage existen 2 grandes tipos replicación sincrónica y asíncrona, cada una de las cuales ofrece beneficios y limitaciones, cada una de éstas son aplicables a ciertos ambientes.

Para empezar con la definición de cada una de ellas debemos empezar con la definición de RPO y RTO.

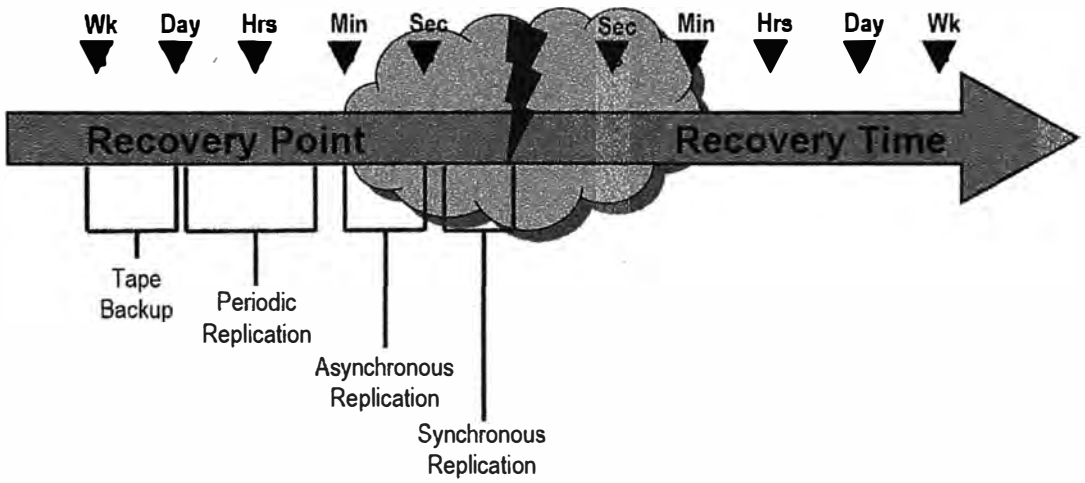


Figura 3.1: Definición RPO, RTO

RPO, Es el monto máximo de data que una aplicación puede perder medido en tiempo, en otras palabras es el monto de data que el negocio puede tolerar perder (costo de transacción vs. el riesgo).

Cada negocio y cada aplicación necesitan escoger un RPO específico, aquí las características de la data que influyen en determinar el RPO adecuado son: legislación, el cual puede requerir por cuanto tiempo la data debe ser accedida y por quien; procesos del negocio pueden estar determinado por niveles, los cuales pueden estar amarrados un punto en el tiempo (cierre de periodos, reportes trimestrales, plazos de impuestos, ciclos de facturación, etc.)

Procesos de los negocios que son amarrados a los niveles de satisfacción de los clientes asociados con la data como sus propósitos son cambiantes por periodos. Por ejemplo, la data puede empezar como transaccional, luego migrar a facturación, luego como un reporte a servicio al cliente, luego al sistema de marketing para finalmente ser colocado al histórico.

RTO, Éste refiere al máximo tiempo que una compañía costea para regresar una aplicación en línea, luego de una caída. En otras palabras el tiempo que toma recuperar la data una vez que un desastre u otro evento de recuperación han sido declarados.

3.1.1 Replicación Sincrónica (SRDF/S® [3.1] de EMC)

En el modelo de replicación Sincrónica el RPO es cero, es decir el almacenamiento mantiene imagen espejada en tiempo real de la data a los devices espejados remotos.

El software de replicación entre almacenamientos Symmetrix es llamado SRDF (Symmetrix Remote Data Facility), el SRDF en su modo Sincrónico es llamado SRDF/S

Su funcionamiento es descrito en la figura 3.2

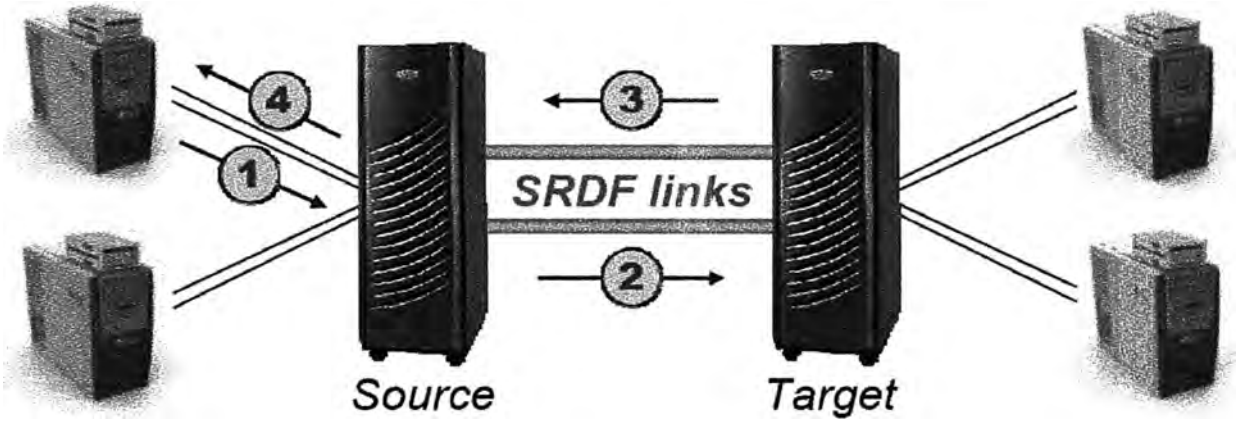


Figura 3.2: SRDF modo Sincrónico (SRDF/S)

1.- Un I/O de escritura (petición) es recibido desde el servidor en el Origen

2.- El I/O es transmitido al Destino

[3.1] SRDF/S, SRDF/A, Symmetrix son marcas registradas de EMC Corporation.

3.- Un Acknowledgment o reconocimiento es enviado desde el Destino al Origen

4.- El I/O es atendido finalmente al Servidor

En SRDF los volúmenes en el Origen son denominados R1 y los volúmenes en el Destino R2, estos volúmenes R1 y R2 están siempre totalmente sincronizados a la culminación de una secuencia de I/O de escritura. Debido a esta adición de tiempo para la atención de una escritura es que una solución sin diseño o mal diseñado de SRDF/S puede tener un gran impacto en la calidad del servicio de las aplicaciones.

Debido a que con la distancia se incrementa la latencia de la señal, es que SRDF/S es limitado en Distancia, el campo de acción de SRDF/S es hasta 200Km

Si el paso 3 nunca sucede, el Origen atenderá el I/O de escritura del servidor luego de predeterminado timeout [3.2] con el fin de mantener los sistemas funcionando.

Debido al funcionamiento de SRDF/S la característica principal de serialización es obtenida por defecto, es decir las solicitudes de escritura son enviadas al almacenamiento destino en el mismo orden como ellos son escritos en el origen con el fin de tener una instancia de tiempo, es decir tener una copia consistente y recuperable en el destino.

Si el Symmetrix remoto es no accesible por algún motivo, entonces las escrituras son acumuladas como "invalid tracks"[3.2] en el registro o "invalid tracks table", cuando el Symmetrix remoto se vuelve disponible nuevamente estos invalid tracks son enviados a este Symmetrix remoto.

Como el RPO de este modo es cero, SRDF/S soporta los niveles tiers de más alta de calidad para las aplicaciones de los negocios

3.1.2 Replicación Asíncrona (SRDF/A® de EMC)

En el modelo de replicación asíncrona el RPO puede ser desde minutos hasta horas. El software de replicación entre almacenamientos Symmetrix es llamado SRDF (Symmetrix Remote Data Facility), el SRDF en su modo Asíncrono es llamado SRDF/A.

SRDF/A es una solución de espejamiento remoto que no tiene impacto en las aplicaciones de producción aun a grandes distancias debido a que las solicitudes de escritura I/O desde el servidor son reconocidas localmente. Los cambios hechos a los mismos blocks de data son enviados periódicamente solo una vez al Symmetrix remoto. De este modo SRDF/A reduce y optimiza el uso del ancho de banda necesario para este tipo de implementaciones. Más aún SRDF/A provee una alternativa de Recuperación de Desastres en adición a

---

[3.2] timeout es el tiempo fuera máximo a espera para recibir el acuse de recibo desde el Almacenamiento Destino

[3.2] Tracks, como se vio en el Capitulo II, es una de las unidades básicas de almacenamiento en disco de 64K

SRDF/S, pues SRDF/A también mantiene siempre una imagen consistente sobre los R2 todo el tiempo.

SRDF/A puede satisfacer un amplio rango de RPOs y RTOs, SRDF/A puede compartir los puertos de replicación del Symmetrix con el SRDF/S

La arquitectura del funcionamiento de SRDF/A es descrita en la figura 3.3

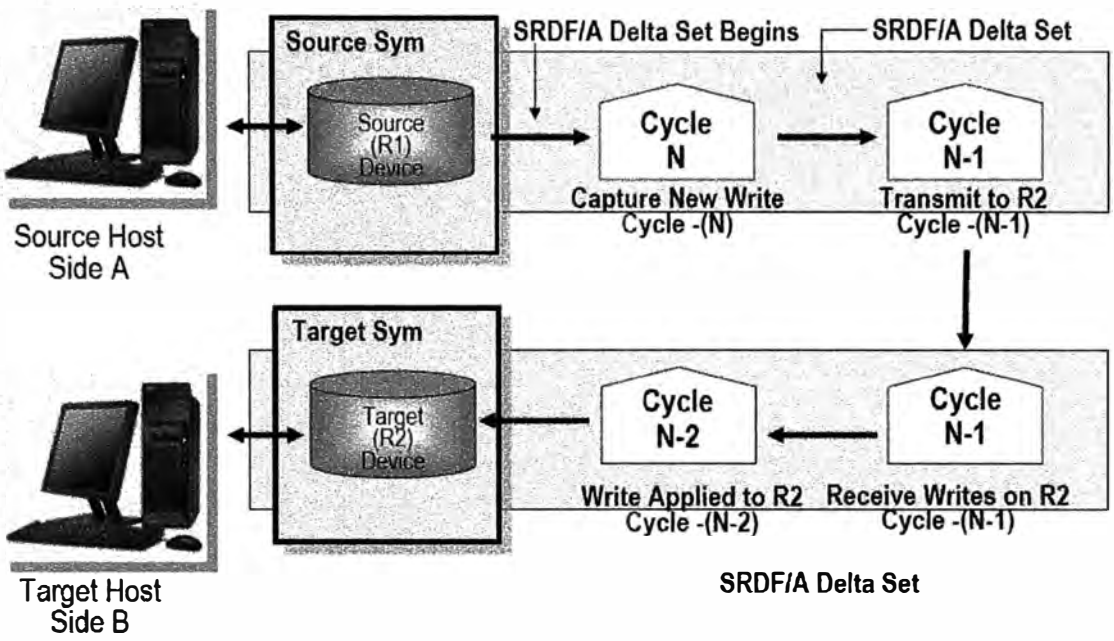


Figura 3.3: SRDF Asíncrona (SRDF/A)

En SRDF/A se definen ciclos de tiempo, siendo un ciclo la unidad mínima y 2 veces el ciclo equivalente al RPO de la solución, cuando el ciclo "N" SRDF/A es activo en el Symmetrix Origen, este recolecta todas los nuevas escritura en el R1 en la cache del Symmetrix, sobrescribiendo cualquier track duplicado previstos para transferirse por el enlace. El ciclo se encuentra activo por un determinado monto de tiempo que puede ser configurado en el Symmetrix cuando se inicia la implementación de SRDF/A; el valor por defecto es de 30 segundos. Luego que este tiempo ha sido alcanzado los datos del deltaset heredan la siguiente posición de ciclo (N-1) y con ello inicia la transferencia del deltaset sobre los enlaces al R2. Luego un nuevo ciclo N inicia recolectando nuevas escrituras de nuevo para ser transferidos en el siguiente deltaset.

En el ciclo (N-1), el deltaset es temporalmente recolectados en el R2 para el destage a discos físico. Cuando el ciclo (N-1) ha terminado de transferir toda la data en el R2 y el mínimo tiempo ha sido alcanzado, el deltaset ahora hereda la posición de ciclo siguiente

(N-2) e inicia el destage de la data a los discos. El deltaset es considerado completado al R2 en el inicio del ciclo (N-2).

De este modo toma 2 ciclos para que los cambios realizados al R1 sean alcanzados y copiados al R2, el cual determina que el mínimo RPO de una solución SRDF/A es 2 veces el ciclo. Así, con los valores por defecto de tiempo de ciclo de 30 segundos, el RPO mínimo es de 60 segundos.

SRDF/A permite replicación a distancias ilimitadas, es decir el site principal podría estar en Lima y el site de Recuperación de Desastres podría estar en Inglaterra con SRDF/A.

SRDF/A soporta múltiples niveles tiers de calidad para las aplicaciones de los negocios, es decir con SRDF/A es posible tener varios niveles de RPOs para las aplicaciones, pues mantiene el concepto de SRDF Groups donde es posible asociar los devices de ciertas aplicaciones con cierto RPO.

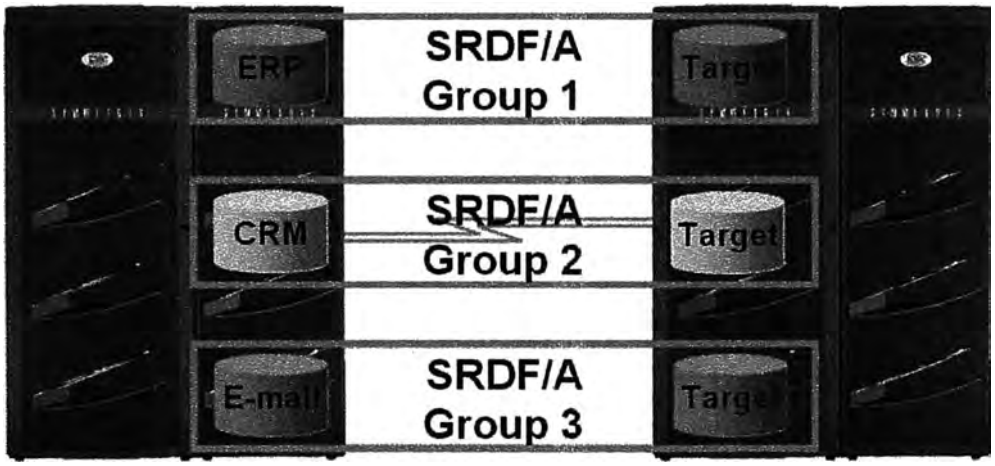


Figura 3.4: Múltiples grupos SRDF/A

En la figura 3.4 se muestra que los devices de la aplicación ERP que pertenecen al SRDF Group 1 pueden ser administrados independientemente de los otros grupos, logrando un RPO objetivo para esta aplicación.

Debido a que una solución SRDF/A bien diseñada no tiene impacto en el rendimiento de los sistemas, algunos cliente optan por cambiar el modo de SRDF/S a SRDF/A durante los periodos de alta cantidad de escritura, este ambiente es normalmente encontrando en las madrugadas durante la ejecución de los procesos batch [3.3], de este modo un cliente podría

[3.3] Proceso batch, procesos que normalmente se ejecutan nocturnamente con la finalidad de procesar la data y obtener otra, esto se hace por lotes (batch). Ejemplos de esto es sacar un reporte diario de la cantidad de productos vendidos en ese día o la cantidad de llamadas recibidas por reclamos.

cambiar de SRDF/S a SRDF/A a la 01am de cada día y volver de SRDF/A a SRDF/S a las 7am del mismo día.

A su vez los volúmenes de un Symmetrix se pueden replicar al mismo tiempo a 2 Symmetrix (en 2 datacenters remotos distintos), concurrentemente hacia el datacenter cercano en SRDF/S y hacia el datacenter más distante en SRDF/A

Figura 3.5: Concurrente SRDF/S y SRDF/A

Del ejemplo se consigue protección ante eventos catastróficos regionales como un terremoto que sacuda a toda la costa este de USA, pues la data esta replicada en otro continente incluso. Este tipo de replicación se esta volviendo un requisito para las entidades financiera que quieran contar con certificación internacional ISO.

En ambos modos SRDF la conexión entre los Symmetrix se realizan mediante tarjetas "Remote Director Adapter" ubicadas en las ranuras delanteras o Front-End del Symmetrix, estas vienen en diversos protocolos como: Fibre Channel Remote Adapter "**RF**", ESCON Remote Adapter "**RA**" y tarjetas multiprotocolo como: MPCD-GigaBitEthernet Remote Adapter "**RE**", iSCSI y FICON. El grafico 3.6 muestra los 2 principales Remote Adapters.

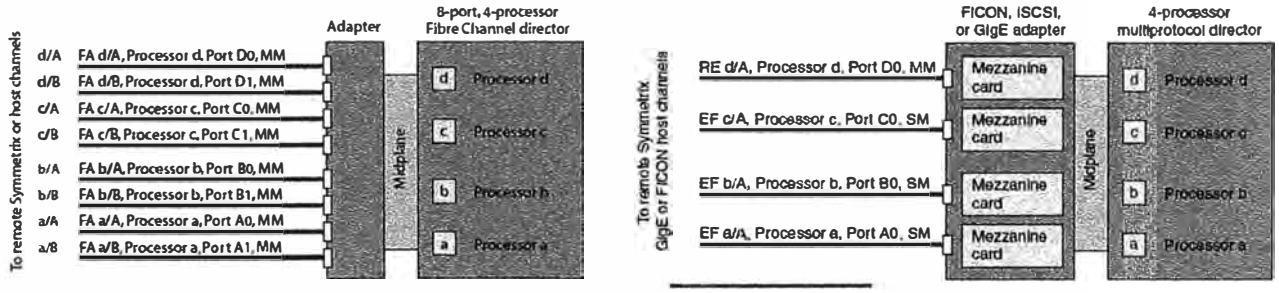


Figura 3.6: Remote Director Adapter: RF (izquierda) y MPCD (derecha)

Dimensionar la cantidad de estos enlaces, su ancho de banda y su calidad es unos de los propósitos del presente informe.

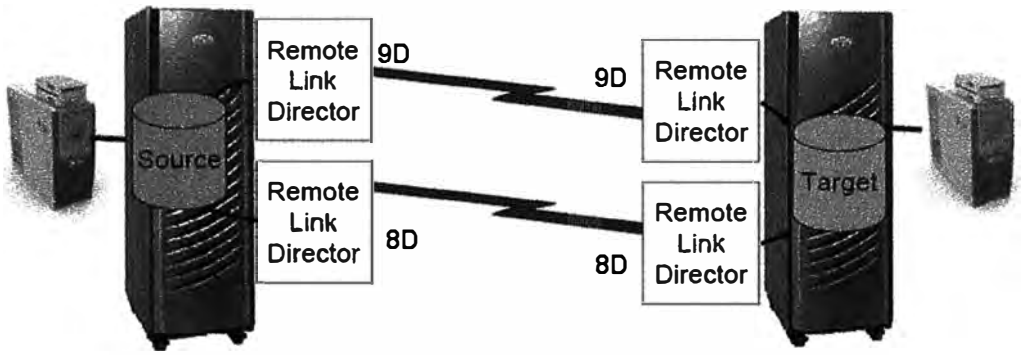


Figura 3.7: Remote Link Director (RLD)

El caso más típico a analizar es el de la conexión vía Fibre Channel, por ejemplo se coloca 2 Tarjetas Front-End en las ranuras 8 y 9 del Symmetrix, cada una de estas Tarjetas viene con 4 procesadores etiquetados como A, B, C y D, cada uno de estos procesadores maneja 2 puertos Fibre Channel con conectores LC de Fibra óptica. De este modo según la figura 3.7 se tienen 2 conexiones RLDs en los puertos 8D y 9D. Los SRDF Groups, también conocidos como RDF Groups o grupos RDF, define lógicamente las relaciones entre los Symmetrix. Un SRDF Group es un juego de conexiones de puertos directores configurados para comunicarse con otro juego de conexiones de puertos directores de otro Symmetrix. Los volúmenes lógicos (devices) R1 son asignados a estos SRDF Groups.

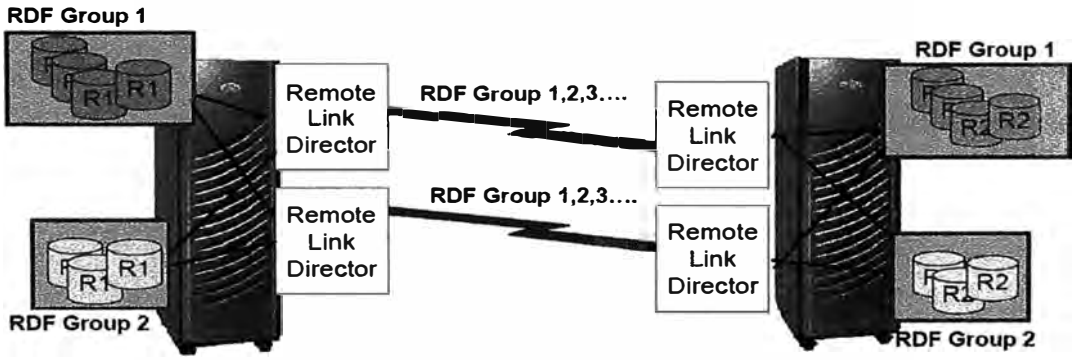


Figura 3.8: SRDF Groups

Varios SRDF Groups pueden compartir un mismo enlace físico entre RLDs.

3.1.3 Soluciones de Conectividad para Almacenamiento Replicado

3.1.3.a Enlace Fibre Channel Topología punto a punto

En este modo de configuración punto a punto los 2 Symmetrix son conectados vía un enlace de fibra óptica de menos de 3Km de distancia, un Symmetrix podría ser colocado en un edificio y el segundo en otro edificio. Mientras físicamente están muy cercanos, esta

configuración permite una alimentación de energía y consideraciones de seguridad por separado.

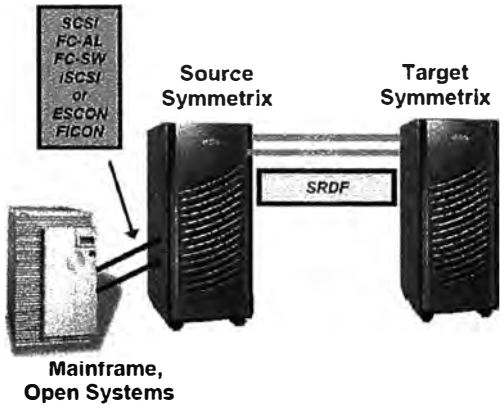


Figura 3.9: SRDF Enlace Fibre Channel Topología punto a punto

Esta es la mejor opción cuando los datacenters primario y secundario están relativamente muy cercanos, además de ser la opción más simple.

Puede ser ESCON o Fibre Channel, si es SRDF/FC se puede incluir también switches cuando los datacenters están más que 3Km separados.

3.1.3.b Enlace Fibre Channel Topología Switched

La solución SRDF Campus conmutada permite a los datacenters origen y destino estar hasta 60Km aparte. La solución campus usa una fibra monomodo privada o una línea arrendada, link extenders y/o switches dinámicos. Este tipo de soluciones requiere link extenders o repetidores.

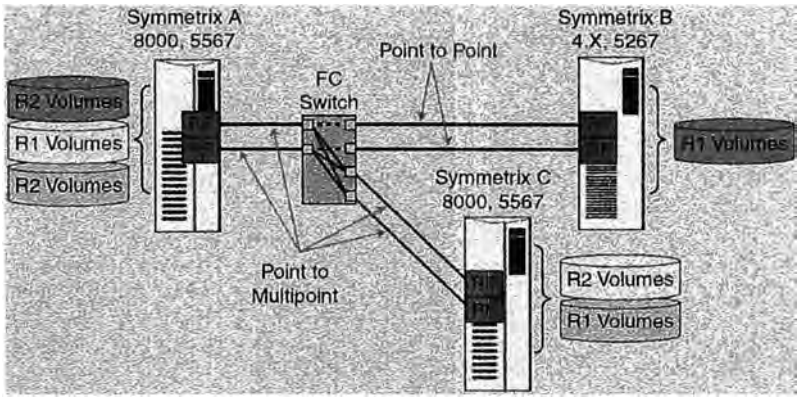


Figura 3.10: SRDF Enlace Fibre Channel Topología Switched

Este tipo de solución es siempre realizado sobre una infraestructura Fibre Channel. Usa Switches FC obligatoriamente. La latencia que incluye el switch es típicamente insignificante. Este tipo de solución es muy usual en clientes que requieren consolidación de muchos storages primarios en unos pocos storages secundarios.

Los sistemas de almacenamiento Symmetrix DMX3 soportan configuración punto-multipunto, permitiendo así múltiples grupos SRDF (SRDF Groups) ser asociados con un

mismo puerto. Esta función es llamada "switched SRDF". Configuraciones Switched SRDF son permitidas por infraestructuras "conmutadas" (Fibre Channel o IP). Un director Fibre Channel o GigE puede ser configurado para soportar comunicación SRDF con más que un Symmetrix en simultáneo.

Un completo entendimiento de switched SRDF requiere una explicación de SRDF Groups, un SRDF Group es definido como un juego de devices (primarios o secundarios) configurados para comunicarse con otro juego de devices en un Symmetrix remoto. Un Director RF en un Symmetrix DMX3 puede ser configurado para soportar hasta 64 SRDF Groups, por lo que una conexión lógica puede ser compartida por múltiples Symmetrix vía el mismo puerto SRDF. Alternativamente múltiples Symmetrix pueden "fan in" (congregarse) a un solo Symmetrix, lo cual significa que el mismo Director RF puede ser configurado para soportar comunicación SRDF con más que un Symmetrix en simultaneo a través de un Fibre Channel switch, proveyendo una verdadera capacidad conmutada.

Según la figura 3.10 SymmA se esta comunicando con el SymmB y el SymmC en simultaneo a través de 2 Directores RF. Cada director RF es configurado para soportar hasta 3 SRDF Groups. Cuando el SymmA esta comunicándose con el SymmC, cualquiera de los Directores RF del SymmA puede comunicarse con cualquiera de los directores en el SymmC.

Hay 4 caminos lógicos desde el SymmA al SymmC (2 caminos lógicos en cada director RF a cada director RF en el SymmC). Cuando el SymmA se esta comunicando con el SymmB, hay un solo camino lógico desde cada director RF a cada uno de los RF director. Esta es una conexión punto a punto. La habilidad de soportar conectividad punto-multipunto no esta relacionada a ser primario (origen) o secundario (destino).

3.1.3.c Enlace Fibre Channel Topología Distancia Extendida

Distancia extendida del enlace FC, usando FC permite a las redes SRDF aprovechar los beneficios de una red óptica a distancias de hasta 120Km usando tecnología DWDM (configuraciones ESCON pueden alcanzar hasta 200Km). Configuraciones SRDF basadas en Fibre Channel deben ir siempre a través de switches primero antes de conectar al multiplexor DWDM, esto con el fin de poder abastecer el incremento de Buffer to Buffer Credits (BB_Credits) [3.4] requeridos por el propio funcionamiento de Fibre Channel.

[3.4] Buffer to Buffer Credits (BB_Credits) son usados por Fibre Channel para manejar el control de flujo por Hardware, son memoria temporales, el switch asigna una cantidad a cada puerto esta cantidad se podría llenar según la distancia del enlace debido al retardo de la transmisión, cada fabricante de switch ofrece una tabla

La multiplexación brinda la oportunidad de concentrar múltiples señales de sobre un mismo enlace óptico. Usando esta tecnología, múltiples Symmetrix pueden mover data SRDF a través de un simple multiplexor a múltiples Symmetrix en diferentes lugares. Las distancias que han sido actualmente probadas/testeadas por los laboratorios de EMC son de hasta 200Km. Sin embargo, se debe revisar la matriz de soporte de EMC por los últimos multiplexores y las distancias soportadas con estos.

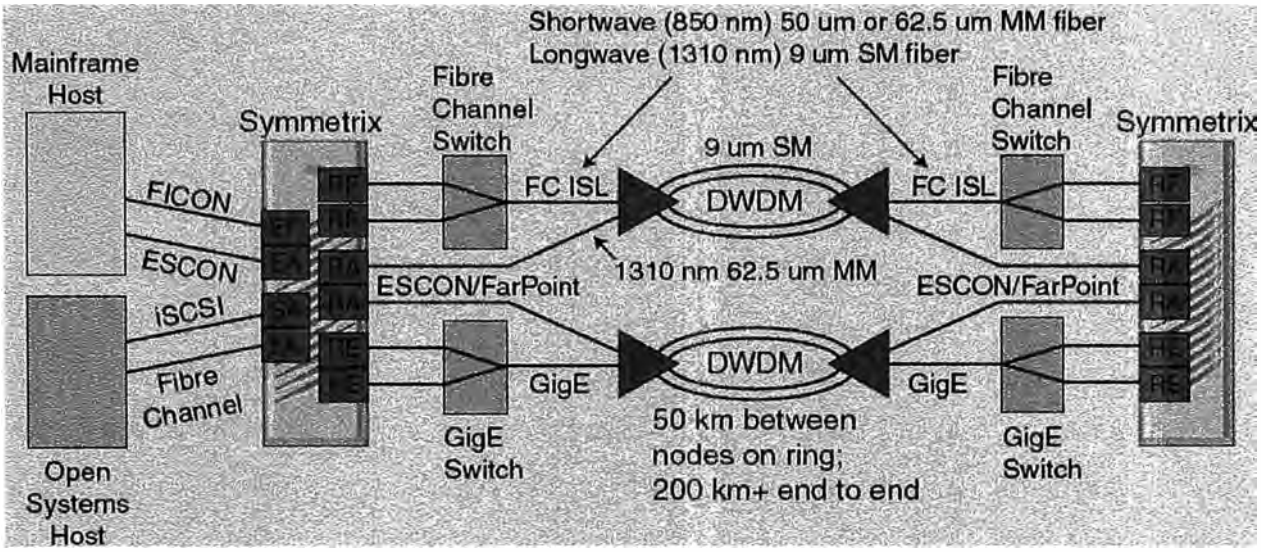


Figura 3.11: SRDF Enlace Fibre Channel Topología Distancia Extendida

El impacto de la longitud de un enlace DWDM sobre el retardo en la propagación (1 mseg por cada tramo de 200Km) debe de ser considerado también.

3.1.3.d Infraestructura típica de SRDF Extendida

La solución de SRDF distancia extendida usa enlace arrendados T1, E1, T3, E3, OC3, OC12, OC48 o líneas de alta velocidad ATM en vez de cables de Fibra ESCON. La máxima distancia entre el origen y el destino es dictada por las limitaciones del proveedor del enlace de telecomunicaciones. Esta solución de distancia extendida es usada comúnmente en distancias mayores a 60Km. Sin embargo, esta puede ser usada en distancias menores de 60Km si la línea de fibra óptica privada o arrendada (fibra oscura) es muy cara o por alguna razón no esta disponible.

---

mostrando las distancias y los requerimientos de BB_Credits para considerar este efecto al diseñar la SAN extendida.

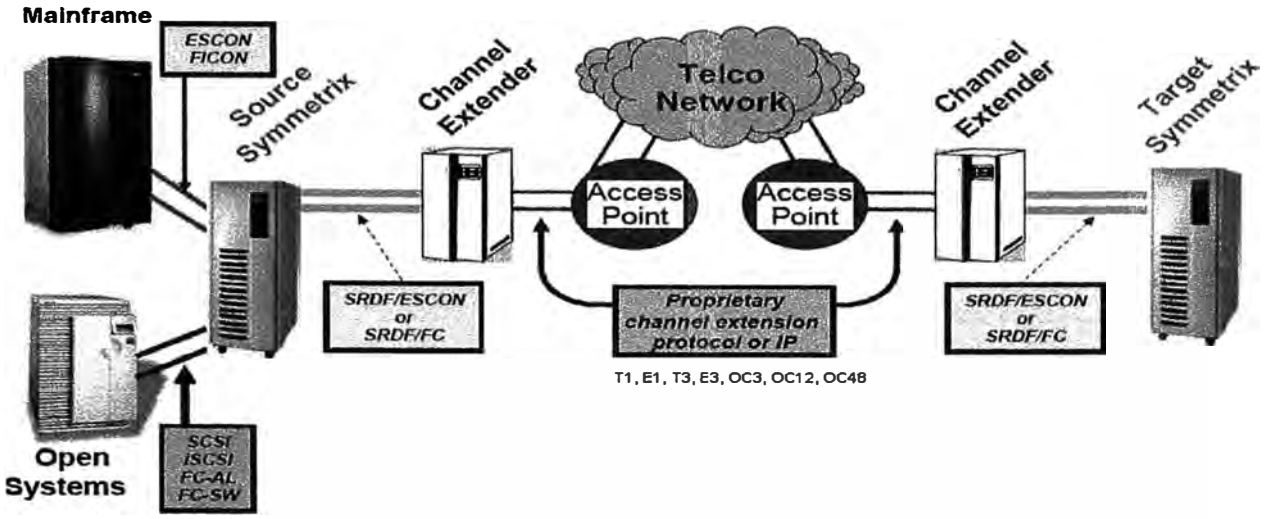


Figura 3.12: SRDF Extended Distance Infraestructure

3.1.3.e GigabitEthernet Nativo y soporte iSCSI

El soporte para IP nativo de los productos SRDF en Symmetrix es basado en tecnología GigabitEthernet o GigE, esta habilita conectar directamente un Symmetrix a una red IP y usar esta como transporte. Ésta incrementa la distancia de conectividad entre Symmetrixs y permite además conectar un Symmetrix a una infraestructura Ethernet existente y permite acceso directo a líneas de alta velocidad vía IP.

El soporte para IP nativo es través de las tarjetas multiprotocolo (MPCD, Multiprotocol Channel Director), estas MPCD proveen adicionalmente compresión a diferentes tazas.

Todas las implementaciones SRDF sobre IP deben incluir un estudio del dimensionamiento y planeamiento de la misma, así como una evaluación de la calidad de la red (network assessment) para poder asegurarnos una implementación exitosa.

La figura 3.13 ilustra la naturaleza conmutada de GigE y su soporte punto-multipunto, la figura también muestra el cálculo de las conexiones TCP y un rendimiento estimado por puerto RE sobre la distancia extendida para requerimientos del dimensionamiento de los enlaces.

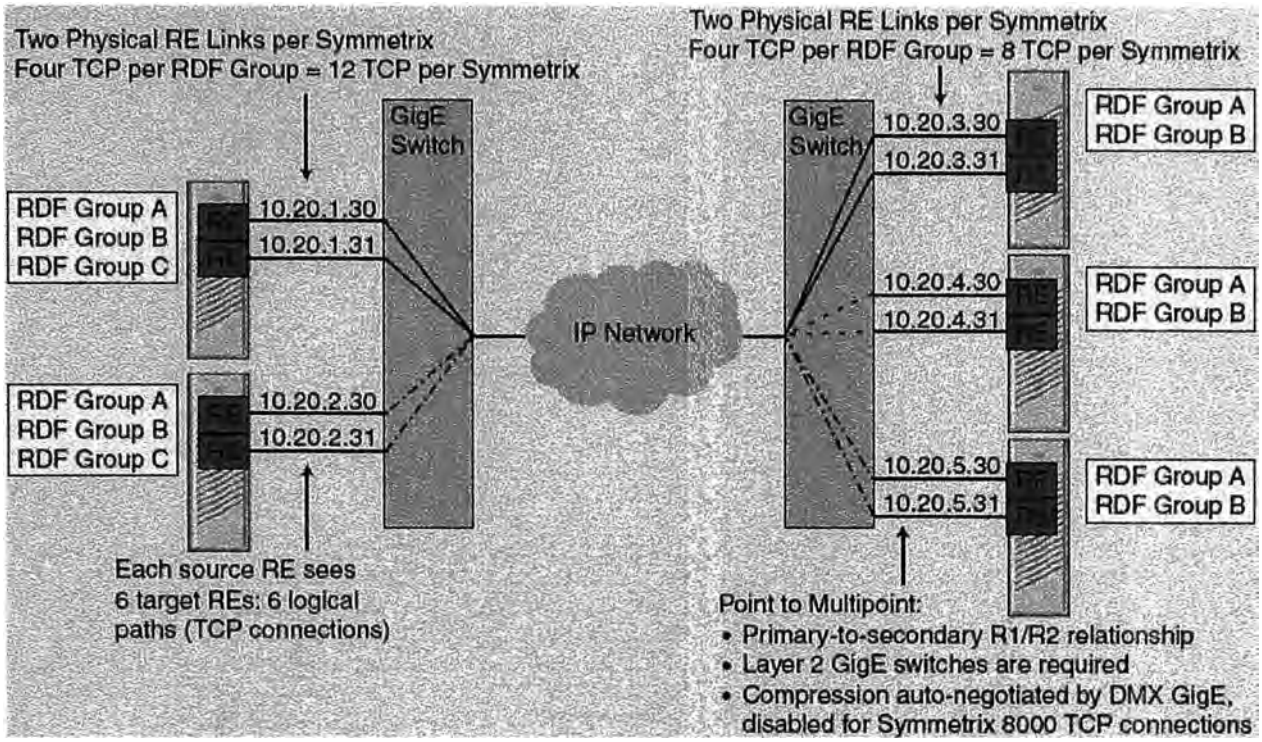


Figura 3.13: SRDF con transporte GigabitEthernet Nativo

Limite de velocidad, las mejores prácticas recomiendan el uso de un ancho de banda dedicado para SRDF y una VLAN exclusiva para la misma. Así la especificación "speed limit" de symmetrix permite configurar el máximo flujo de trafico a enviar al enlace IP y no congestionarlo

MTU, por defecto esta configurado en 1500 Bytes. Sin embargo, puede ser incrementado hasta a 9000 bytes, se debe asegurar que toda la red permita estos tamaños largos antes de configurar este parámetro en los Symmetrix.

Escalamiento de las conexiones TCP, los puertos GigE del Symmetrix crea una conexión TCP entre los puertos pares del origen y destino para transferir data SRDF, cada conexión TCP corresponde a un enlace lógico SRDF. Cuando los puertos GigE son conectados a switches o una red enrutada, cada puerto origen puede conectarse a múltiples puertos destinos, creando múltiples conexiones TCP. Múltiples conexión TCP son deseables pues incrementan el throughput.

Compresión, implementando compresión puede reducir los requerimientos de ancho de banda. DMX GigE directors soportan compresión on-board usando el algoritmo de compresión ILZS. Data SRDF es comprimida antes de encapsularse en paquetes IP. Debido a ello es altamente recomendable usar compresión. Compresión y descompresión adiciona latencia al tiempo de transmisión del paquete a través de la red, usualmente 0.125mseg por salto y 0.5 mseg por round trip. Algunos dispositivos podrían ser más aun.

Se puede tomar como línea base una compresión 2:1 y luego medir el impacto y encontrar la mejor taza de compresión.

Las soluciones de conectividad para almacenamiento replicado también se dividen según el área de cobertura en las cuales trabajan, así se describe 3 grandes áreas de implementación de SRDF mostradas en la figura 3.14

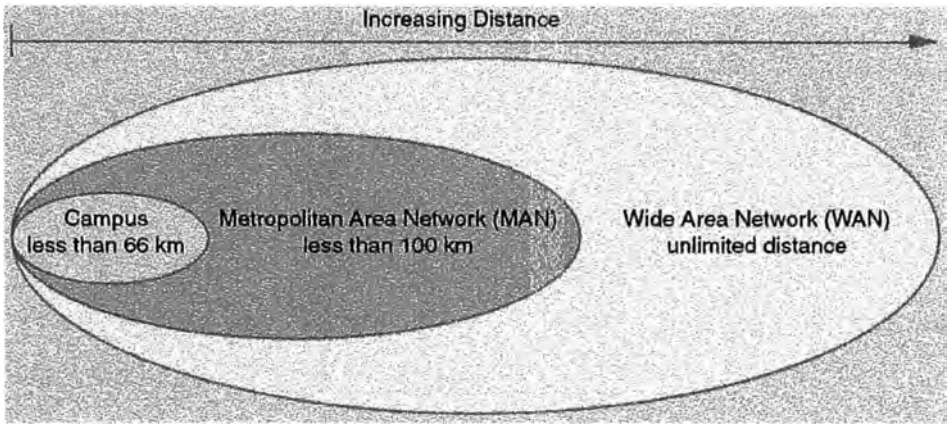


Figura 3.14: Áreas de cobertura implementaciones SRDF

Campus, provee conexión directa limitadas en distancia como alternativa para ambientes con fibra óptica oscura. Normalmente, redes metropolitanas son usadas para configuraciones de muy cortas distancias.

Implementaciones Campus incluyen: conexión directa GigE (point-to-point), conexión directa Fibre Channel (point-to-point), Switched SRDF, conexión directa ESCON.

Metropolitan Area Network (MAN), Provee conectividad SRDF para distancias típicamente menores a 100Km. Sin embargo, hasta 200Km o más, utilizando fibra oscura, DWDM o SONET/SDH. Estos ambientes son caracterizados por tazas de error extremadamente pequeñas y un gran ancho de banda de configuraciones de fibra, de modo que recuperación de errores y compresión son menos significantes al escoger opciones de conectividad SRDF.

Algunas configuraciones disponibles para distancias MANs son:

- Switched Fibre Channel/DWDM: RF-FC Switch-DWDM-DWDM-FC Switch –RF
- SRDF/GigE via DWDM: RE-DWDM-DWDM-RE
- SRDF/GigE sobre SONET: RE-router-anillo SONET-router-RE
- Switched Fibre Channel sobre SONET: RF- FC Switch – Dispositivo FC/SONET-anillo SONET – Dispositivo FC/SONET – FC Switch - RF

Wide Area Network (WAN), Provee conectividad SRDF/ESCON para grandes y muy grandes distancias (transoceánicas a transcontinentales) usando las redes de telecomunicaciones como: IP, SONET/SDH o ATM. WANs son diferenciadas de las MANs por ser ambientes con pérdidas o errores y limitados en ancho de banda, donde recuperación de errores, buffering de la data y capacidades de compresión son opciones de conectividad extremadamente importantes.

Opciones de configuración WAN: T1/T3, E1/E3, SONET/SDH, ATM OC3, IP.

Las distancias listadas anteriormente son las más usuales. Sin embargo, existen exoneraciones a estas, por ejemplo costos muy altos, uno o más operadores, limitado ancho de banda, único servicio disponible en el lugar, etc.

3.2 Solución del problema

Como se ha visto en los capítulos anteriores es realmente importante y vital realizar un diseño, planeamiento y dimensionamiento de los componentes claves de una solución de almacenamiento replicado (distancia de la solución SRDF/S, Ancho de Banda y calidad de los enlaces, Cantidad de Remote Adapters, posibles volúmenes o LUNs con problemas de performance por la replicación SRDF/S, Cache adicional necesaria y RPO en el caso de SRDF/A), la cual permita poder lograr una implementación exitosa.

Para lograr un buen diseño nos valemos de estadísticas recolectadas en los sistemas de almacenamiento Symmetrix [3.5] antes de implementar la replicación en el momento de la evaluación de la misma, estas estadísticas de entrada estandarizadas para cualquier sistema de almacenamiento y servidores conectados a los mismos son: la cantidad de reads, los tamaños de estos reads, cantidad de writes, los tamaños de estos writes, tiempo de respuesta, etc. Recolectados como muestras cada 10min por una cantidad de días (pueden ser 3 ó 5 días). Estos días deben incluir los días más significativos de la operación del negocio en el mes, días tales como: facturación, cierres contables, fechas límite de pagos, etc.

El grafico 3.14 muestra la cantidad total de IO/s (read y writes) que es consumido por todo un sistema de almacenamiento Symmetrix priori a la implementación de replicación SRDF.

[3.5] En Symmetrix los archivos de estadísticas que recolectan todo el status del mismo son denominados ".btp"

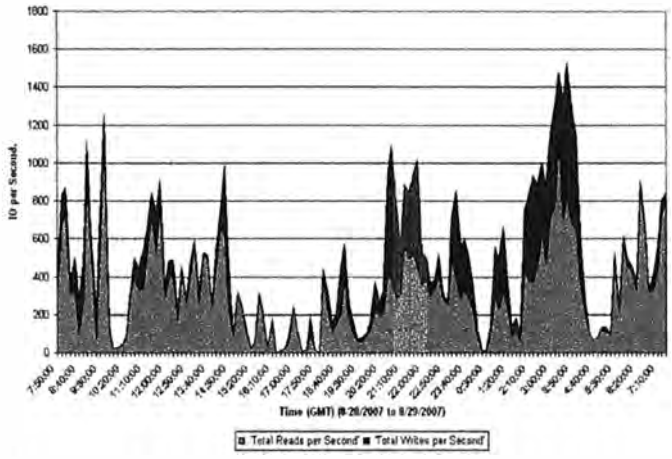


Figura 3.15: Perfil del IO total

La forma de enfrentar el problema de dimensionamiento y planeamiento es modelar el funcionamiento del almacenamiento cuando este es replicado.

¿Y por qué necesitamos modelar?

- Ayuda a definir los componentes de la solución
- Ayuda a definir el costo de la solución
- Reduce riesgos en la implementación
- Ayuda con la capacidad actual y planeamiento del rendimiento
- Planeamiento de la capacidad (Storage, switch, circuitos y crecimiento del storage)
- Rendimiento (Critico para soluciones Sincrónicas, medir el crecimiento de la carga de trabajo en el storage)
- Ayuda a comprender las opciones donde ubicar el datacenter secundario.
- Un apropiado planeamiento y mantenimiento permitirán lograr los SLAs (Service Level Agreements) comprometidos

El objetivo del modelamiento para el dimensionamiento es identificar los requerimientos técnicos claves que podrían impactar críticamente en los requerimientos del negocio:

- ¿Cuanto ancho de banda es necesario?
- ¿Cuales fabricantes de switches y protocolos ayudaran a incrementar la distancia que nosotros podamos replicar?
- ¿Cual será la latencia esperada para soluciones sincrónicas?
- ¿Cuanto buffer adicional (cache y disco de paginación) será necesario para operaciones asíncronas?
- ¿Cuantos Fibre o GigE Remote Adapter se necesitaran?

- ¿Existen volúmenes (LUNs) con problemas potenciales (escritura intensa) que podrían ser impactados en rendimiento?
- ¿Qué RPO puede ser alcanzado?

El modelamiento necesita validar áreas/puntos de la conectividad para garantizar el correcto dimensionamiento.

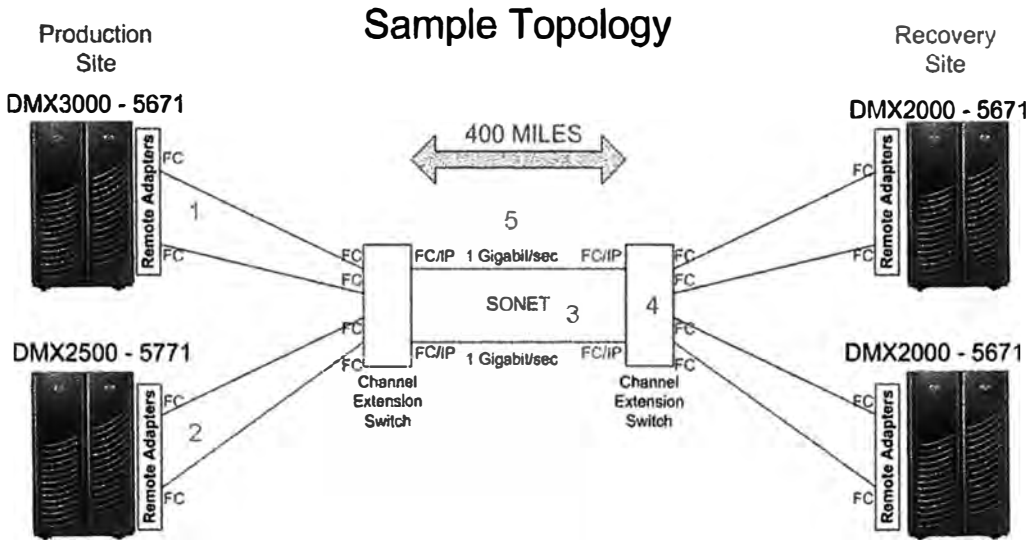


Figura 3.16: Áreas a validar de la conectividad

De la figura 3.16 se ilustran las áreas que se necesitan validar de la conectividad para el dimensionamiento de la replicación:

1. Tipo de Storage, Nivel de micro código, tipos de RA afecta la taza de transferencia, tiempo de respuesta para SRDF/S y cálculos de cache para SRDF/A
2. Remote Adapters afectan el monto de de data que puede ser colocada en la red en adición a definir las opciones de conectividad. Estos también ayudan a determinar el protocolo y la encapsulación de red que reduce el ancho de banda total.
3. Definiendo los Remote Adapters, conectividad entre extensiones y tipos de circuitos es critico para comprender el efecto de la encapsulación, por ejemplo:

FC -> FC/IP -> SONET = ~32-44% overhead

Gig-E ->SONET = ~8-10% overhead

FC -> FC/IP -> ATM = ~45-57% overhead

4. Channel Extensión switches tienen una variedad de latencias pass-thru [3.6] que adicionan latencia al tiempo de respuesta sincrónico (en conjunción con el protocolo)
5. El ancho de banda disponible determina si la carga de trabajo de la operación propia de los servidores replicados pueden ser soportados y es una variable critica para determinar la cache adicional para replicación Asíncrona

Comprender los componentes y como estos podrían afectar la viabilidad de la solución es crítico, más aún cuando se trata de determinar costos y los SLAs (tales como el RPO) que pueden ser logrados.

Las estadísticas que se usaran como entrada del modelo pueden ser recolectadas tanto en el almacenamiento (como archivos ".btp" del WLA/STP), como en los servidores (Linux: IOSTAT, HP-UX: SAR, Solaris: IOSTAT, Microsoft: PERFMON, IBM-AIX: IOSTAT, IBM iSeries: WRKDSKSTS, IBM z/OS: SMF/CMF records).

Como se mencionó anteriormente se necesita estandarizar estas estadísticas de entrada con los valores necesarios para los cálculos. De lo anterior obtenemos el siguiente patrón estandarizado para los registros de las muestras cada 10min.

CSV - 13SEP06, 8:00, 9464, 9000, 12369, 9000, VOLSER, 24,06, 0:15:00

- Date & Time — Fecha y hora de la observación
- Reads — Cantidad de lecturas para el intervalo de observación
- Read block size — Tamaño promedio de las lecturas en ese intervalo
- Writes — Cantidad de escrituras para el intervalo de observación
- Write block size — Tamaño promedio de las escrituras en ese intervalo
- Lun/Volume — Volumen ID o nombre del dispositivo visto desde el S.O.
- Response time — Tiempo de respuesta promedio para el dispositivo
- Connect time — Tiempo promedio para transferir un bloque de data
- Interval length — Tiempo transcurrido en el intervalo.

3.2.1 Calculando Requerimientos Replicación Sincrónica SRDF/S[3.7]

- Requerimientos de ancho de banda (comprimido y sin compresión)
  - Sin compresión (para cada muestra)
    - $Writes\_per\_\sec ond(WPS) x Write\_Blocksize = Volumen\_MB/S$ .....(3.1)

[3.6] pass-thru, tiempo que demora el dispositivo en procesar y conmutar un paquete
[3.7] Formulas obtenidas del whitepaper de EMC: EMC Replication Modeling Tool (ET) – SRDF 2007

- $Suma\_de\_observaciones\_Volumenes\_MB/S = MB/S\_\text{int}ervalo$ .....(3.2)

  - Compresión, dividir lo anterior por la taza definida

- Impacto de la latencia de la red en el tiempo de respuesta sincrónico
  - La formula a utilizar:
    - $SRT = STD + (PTDxNPx2) + (DLxNP/2)$ .....(3.3)
    - SRT = Synchronous Response Time
    - STD = Standardized Response Time o Tiempo de respuesta estandarizado (basados en el tipo de storage, nivel de micro código, Tipos de RA y el tamaño de bloque de escritura)
    - PTD = Network Switch Pass Through Delay (ms), tiempo de atención en el switch respectivo
    - NP = Number of Network Passes (2 o 4 dependiendo del protocolo /Fastwrite Fibre Channel), numero de veces que necesita el protocolo atravesar la red para su propio funcionamiento.
    - DL = Distance Latency (round trip), latencia debida a la distancia ida y vuelta. Esta se debe a las limitaciones de luz al viajar por un medio, esta adiciona un retardo en la transmisión de 1mseg por cada 200Km de recorrido.
- Cantidad de Remote Adapter requeridos:
  - RA Time calculado por la observación en un intervalo
    - $(WPS/Max\_Link\_WPS)x1000mseg$ .....(3.4)
    - Max_Link_WPS varia dependiendo de los resultados de las pruebas para diferentes familias de micro código y tipos de RA (GigE, FC, Escon)
    - 1000mseg de uso equivale a 1 Enlace RA físico de ese tipo a ese nivel de micro código, esto quiere decir que por cada 1000mseg de Remote Adapter throughput, un puerto RA es requerido.
  - Uso del RA por intervalo es la suma de todas las observaciones en el intervalo
- Potenciales Volúmenes/LUNS con posibles problemas de performance, antes de la implementación es requerido un análisis para determinar que volúmenes (si

existiesen) tendrán un inaceptable retardo debido a la latencia inducida por el espejamiento SRDF/S

Comparamos el actual WPS para cada volumen contra su calculado "Volume Write Limit under Synchronous" (VWL) o Limite de escritura del volumen bajo Sincrónico.

- $VWL = 1000 mseg / SRT$ .....(3.5)
- De modo que si un volumen tiene actualmente 400 WPS y se calcula un Synchronous Response Time (SRT) de 4mseg producirá un VWL de (1000/4) = 250 WPS
- El modelo marcará todo volumen donde el VWL (dado que es el limite máximo) es menos que el actual WPS

- Riesgos que no pueden ser tratados por el modelo:
  - Latencia variable asociada con degradación de la red o enrutamiento diverso
  - Reducción del ancho de banda asociado con el encapsulamiento propio del protocolo usado.

3.2.2 Calculando Requerimientos Replicación Asíncrona SRDF/A[3.8]

- Requerimientos de ancho de banda (comprimido y sin compresión)

  Este cálculo es el mismo que para el caso Sincrónico. No se considera que en algunas ocasiones pueda existir Localización de paginas en cache en el mismo track, el cual es un beneficio del SRDF/A que reducirá un el ancho de banda.

- Cache requerida para SRDF/A
  - En términos simples, este es el área bajo la curva del ancho de banda para cada ciclo de 30seg (x2).

    Es decir la **mínima cache requerida** es determinada sumando la data total cambiada en cada incremento de 30seg. Para asegurar que el estimado de cache se acomoda aun en el peor escenario, este es asumida que cada write ocupa un slot de cache separado. **El máximo monto de cache** es determinado por la diferencia entre la taza de-stage de transmisión WAN (debido a la taza de transferencia disponible) menos el monto total de data cambiada mientras dure el ciclo actual.

[3.8] Formulas obtenidas del whitepaper de EMC: EMC Replication Modeling Tool (ET) – SRDF 2007

- Cuando la taza de cambios (Write MB/S) excede el ancho de banda disponible, esa diferencia es absorbida dentro de la cache (deltasets SRDF/A) hasta el tiempo en que el consumo del ancho de banda baje de la utilización máxima por un tiempo suficiente que permita a la cache ser de-staged a través del enlace. Esta dará como resultado que el tiempo del ciclo y consecuentemente el RPO se incremente.
    - Cache = (Write MB/s x Cycle Size x 2) x Cache Slot Size (ajustado a MB).....(3.6)
    - Una nota importante: Cuando usamos data recolectada con intervalos de recolección mayores que el ciclo del SRDF/A (el cual es casi siempre), la medida del ancho de banda (MB/S) usado para calcular la cache SRDF/A será más bajo mayor sea la diferencia entre el tiempo del ciclo y el intervalo de recolección, este es comúnmente llamado como tomar "el promedio del promedio"

En los gráficos siguientes se muestra el efecto de no contar con suficiente ancho de banda para poder transmitir los cambios y como afecta al RPO y la cantidad de cache adicional requerida para poder atender este incremento.

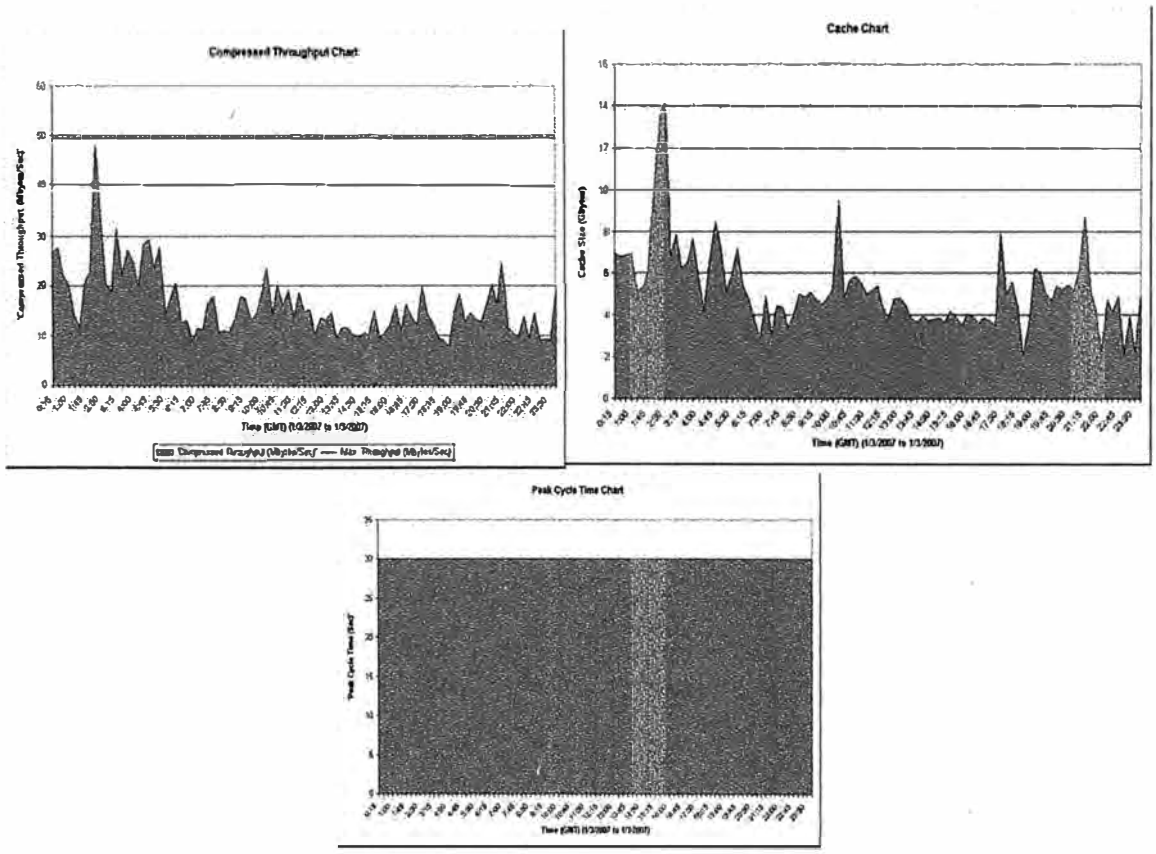


Figura 3.17: SRDF/A Con suficiente ancho de banda.

En la figura 3.17 observamos una solución SRDF/A bien diseñada, según la carga de trabajo de los volúmenes replicados es necesario 50MByte/s con lo cual solo es necesario 14GB de cache adicional, con lo que la cache a configurar seria 16GB (2 tarjetas adicionales de 8GB), con ello también se puede visualizar que el tiempo de ciclo es alcanzado, es decir los 30segundos por defecto.

A continuación se muestra que pasaría si en vez de proveer los 50MByte/s necesarios para la replicación, solo se provee de 28MByte/s.

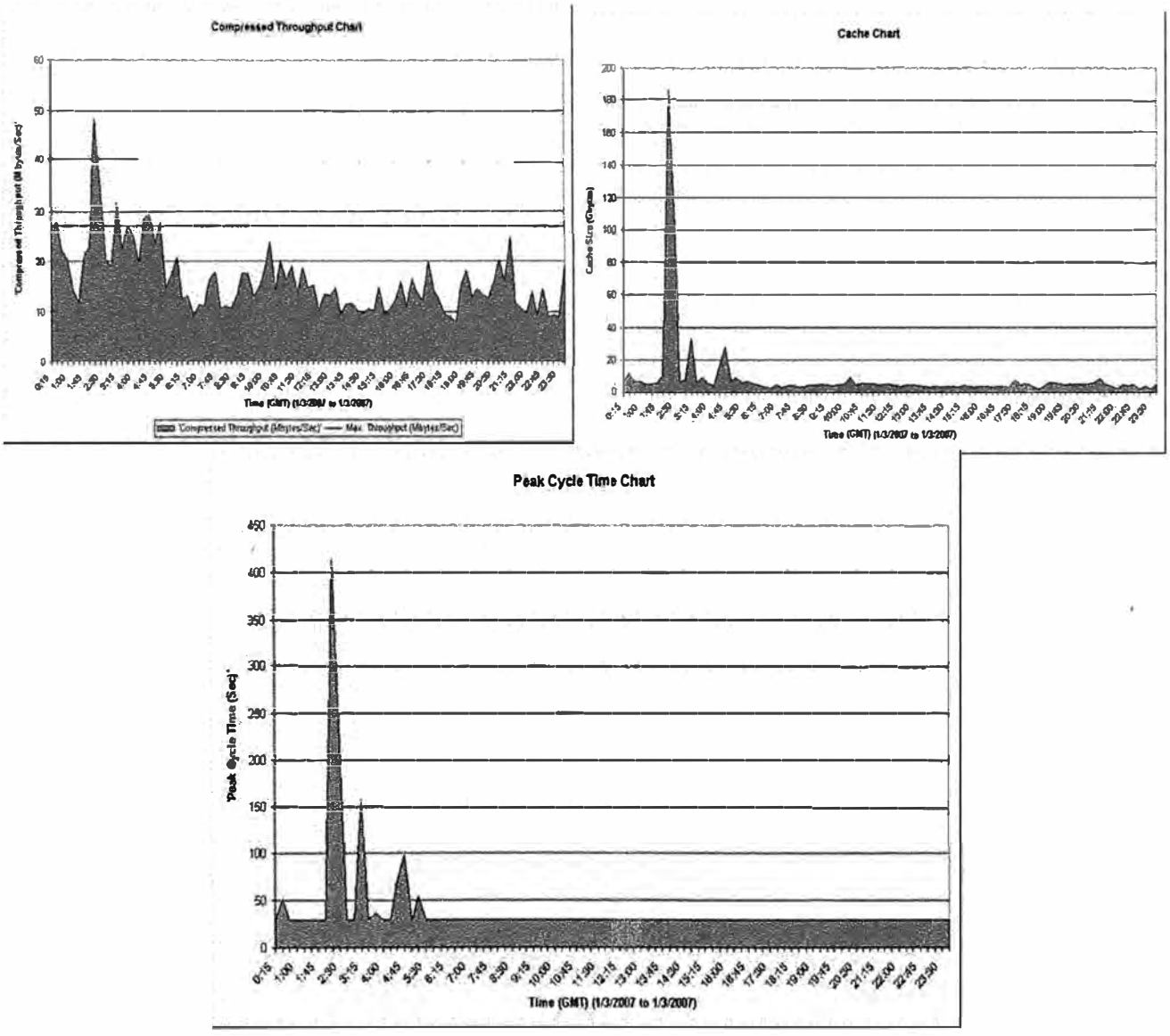


Figura 3.18: SRDF/A Con insuficiente ancho de banda.

De la figura 3.18 se puede visualizar que por varios periodos de tiempo en el día la carga de trabajo WPS (MB/S) sobrepasa el ancho de banda disponible de 28MByte/s, esta sobrecarga es asumida por la cache, debido a ello se puede visualizar que en la hora de mayor pico el requerimiento de cache necesaria para soportar este adicional se eleva a 180GB de cache adicional (lo cual se vuelve inviable según los tamaños de las tarjetas cache y precios actuales de las mismas[3.9]).

Sin embargo, el peor efecto es que el RPO en la hora pico se eleva a 2x400 seg. (13minutos aprox.), con lo cual en esta solución no se puede especificar en un SLA que el RPO ofrecido es de 60 segundos siempre.

- Tiempo de ciclo SRDF/A y RPO (30 segundos por defecto)
  - Si el WPS (MB/s) es menor que el ancho de banda provisto, el tiempo de ciclo será 30 segundos sin problemas. A fin de mantener consistentemente el tiempo de ciclo en 30 segundos este ancho de banda debe ser mayor que la tasa de cambios WPS (MB/S).
  - Si por el contrario el WPS (MB/s) es mayor que el ancho de banda provisto, el tiempo de ciclo será incrementado para contabilizar los cambios acumulados.
    - $$PCT = \left[\frac{DCT}{(CWPS / BWL)}\right] + INT\left[1 - \frac{1}{(CWPS / BWL)}\right] \text{.....(3.7)}$$
    - Si PCT es menor que 30 seg., el PCT será establecido a 30 segundos.
    - PCT = Peak Cycle Time
    - DCT = Desired Cycle Time (Tiempo de ciclo deseado, por defecto es 30 segundos)
    - CWPS = Compressed Write MB/s o Taza de escritura comprimida.
    - BWL = Bandwidth Limit, ancho de banda definido por el usuario.
    - INT = Data Collection Interval Length, longitud del intervalo de recolección de data.
  - El RPO estimado para SRDF/A será el tiempo de ciclo actual más el previo tiempo de ciclo aun aunque en realidad el real RPO podría ser menos.
- Cantidad de Remote Adapter requeridos:

[3.9] La cantidad máxima de cache soportada por un Symmetrix DMX4 es 512GB, los tamaños de las tarjetas modulares disponibles son de 8, 16, 32 y 64GB y sus costos oscilan en las decenas de miles de dólares.

- Es la misma que para SRDF/S. No se considera que en algunas ocasiones pueda existir Localización de paginas en cache en el mismo track, el cual es un beneficio del SRDF/A que reducirá un el ancho de banda y con ello la cantidad de Remote Adapter.

CAPÍTULO IV
CASO DE ESTUDIO

En el presente capítulo se expondrá un caso de estudio donde se mostrará las técnicas explicadas anteriormente, se mostrarán estadísticas de entrada y mediante software de EMC se calcularán los recursos de hardware y transporte necesarios para este caso de estudio. Finalmente se presentarán conclusiones.

4.1 Caso de Estudio

4.1.1 Descripción del ambiente y las necesidades

El caso de estudio propuesto muestra al cliente Banco UNI que cuenta con un DMX3 en su oficina principal en el Distrito del Rímac en Lima (origen), donde el principal sistema y en donde radica el negocio del Banco se encuentra en una Base de Datos Oracle 10g ® [4.1]de 10TB de capacidad. Dada la importancia vital de este sistema para la existencia del Banco y por estar próximos a obtener las certificaciones ISO es necesario contar con Datacenters alternativos que puedan permitir Continuidad de Negocios y Recuperación de Desastres. Es decir seguir operando aun en condiciones de desastre.

- Esta recuperación debe procurar ser lo más rápida posible (RTO deseado 2horas) y con la menor perdida de data posible (RPO deseado 0).
- Adicionalmente el ente de certificación ISO esta solicitando al Banco poder tener la capacidad de recuperarse ante desastres regionales como un terremoto en Lima.

Por lo que el Banco UNI opta por implementar 2 datacenters alternativos para poder cumplir con ambos requerimientos.

El primer datacenter lo ubica a 30Km del Rímac en el distrito de Chaclacayo con el fin de cumplir los requerimientos más exigentes que su negocio le exige (RTO deseado 2 horas, y RPO=0), el cual realizará una replicación SRDF/S para ello tiene implementado una red

[4.1] Oracle Database 10g es una marca registrada de Oracle Corporation.

DWDM de propiedad del Banco UNI, con ello extenderá su SAN vía switches Fibre Channel los cuales tendrán 2 conexiones entre switches denominada ISLs a 1Gbps (100Mbyte/s) cada una.

El segundo datacenter lo ubica a 560Km en la ciudad de Trujillo para cumplir con el 2do requerimiento de un tener un datacenter fuera de la región, para ello arrendara enlaces OC3s a las operadoras de Telecomunicaciones, mediante estos enlaces transportará protocolo IP. El detalle se ilustra en la figura 4.1 a continuación.

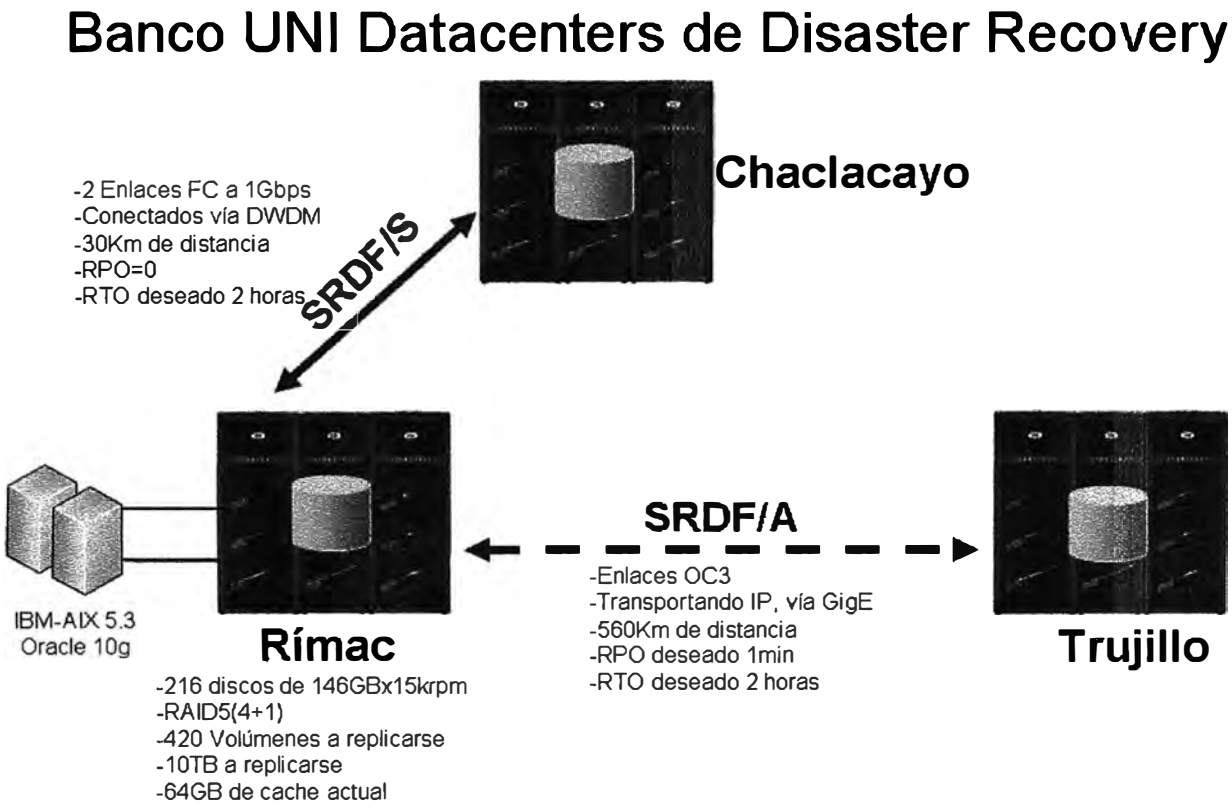


Figura 4.1: Banco UNI Datacenters de Recuperación de Desastres

4.1.2 Estadísticas de entrada

Se recolectaron muestras cada 10min de la carga de trabajo de los volúmenes de producción que serán replicados, esta incluyeron los días de mayor carga de escritura en los volúmenes a replicarse, esto para incluir en el análisis estos picos de escritura. Es muestras se recolectaron en el DMX3 como archivos en formato ".btp". Éstas cuando se exportan a un archivo en formato .csv lucen como (solo se muestran los primeros registros a manera ilustrativa):

Fecha	Hora	# reads	tamaño reads	# writes	tamaño writes	Lun/ Volume	tiempo respuesta	tiempo conexión	intervalo
27-Aug	12:00:00	2560	59146	0	0	021	0	0	00:10:00
27-Aug	12:00:00	2624	56546	0	0	022	0	0	00:10:00
27-Aug	12:00:00	2671	56179	205	4174	023	0	0	00:10:00
27-Aug	12:00:00	2429	59843	3	4096	024	0	0	00:10:00
27-Aug	12:00:00	7084	75767	147	9853	025	0	0	00:10:00
27-Aug	12:00:00	7570	71972	44	6411	026	0	0	00:10:00
27-Aug	12:00:00	7123	75007	44	15656	027	0	0	00:10:00
27-Aug	12:00:00	6834	76792	142	11538	028	0	0	00:10:00
27-Aug	12:00:00	12927	69085	1	16384	029	0	0	00:10:00
27-Aug	12:00:00	13382	68185	57	9887	002A	0	0	00:10:00
27-Aug	12:00:00	16661	55599	164	9078	002B	0	0	00:10:00
27-Aug	12:00:00	15806	58607	2	32768	002C	0	0	00:10:00
27-Aug	12:00:00	27325	79011	3	16384	002D	0	0	00:10:00
27-Aug	12:00:00	23999	85430	0	0	002E	0	0	00:10:00
27-Aug	12:00:00	24448	85541	0	0	002F	0	0	00:10:00

Tabla 4.1: Estadísticas de entrada

4.1.3 Cálculos y resultados

4.1.3.a Resultados SRDF/S a Chaclacayo

Con las formulas expuestas en el capitulo III, EMC desarrolló herramientas internas para realizar estos cálculos, esta herramienta denominada ET [4.2] nos permite calcular los requerimientos. De este modo obtenemos:

Resumen de la carga de trabajo:

Tamaño de Bloque de Lectura Promedio	61.0KB
Tamaño de Bloque de Escritura Promedio	29.0KB
Máximo IOs por segundo	16,757
Máximo IOs de Escritura por Segundo	2,695

Tabla 4.2: Resumen carga de trabajo SRDF/S

[4.2] ET o EMC Remote Replication Designer es software propiedad de EMC Corporation

Resumen de los resultados:

Máximo Throughput Comprimido Requerido	117.48MB/seg.

Tabla 4.3: Resumen de resultados SRDF/S

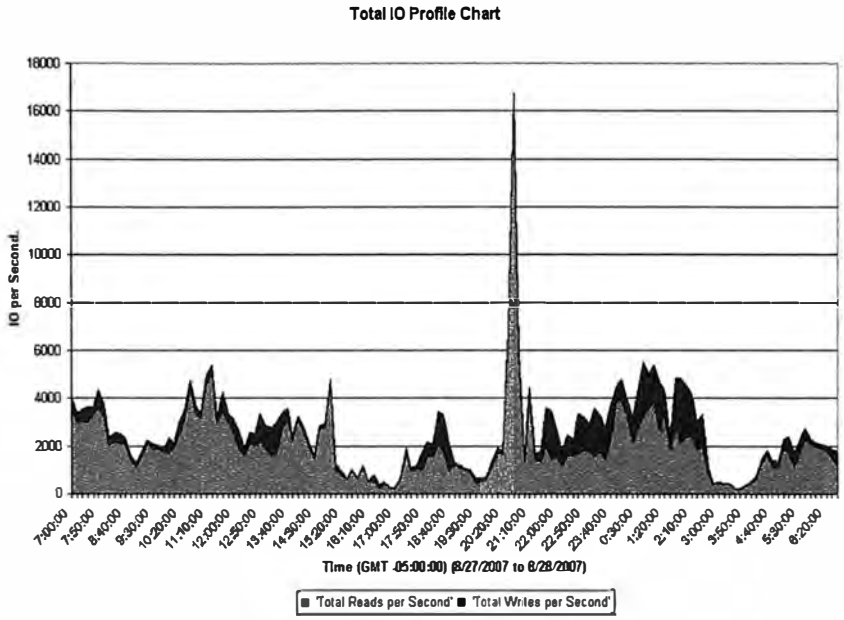


Figura 4.2: SRDF/S IO Profile

En la figura 4.2 se puede visualizar los 16000 IOs máximo de carga de trabajo y que el storage soporta más reads (azul) que writes (guinda).

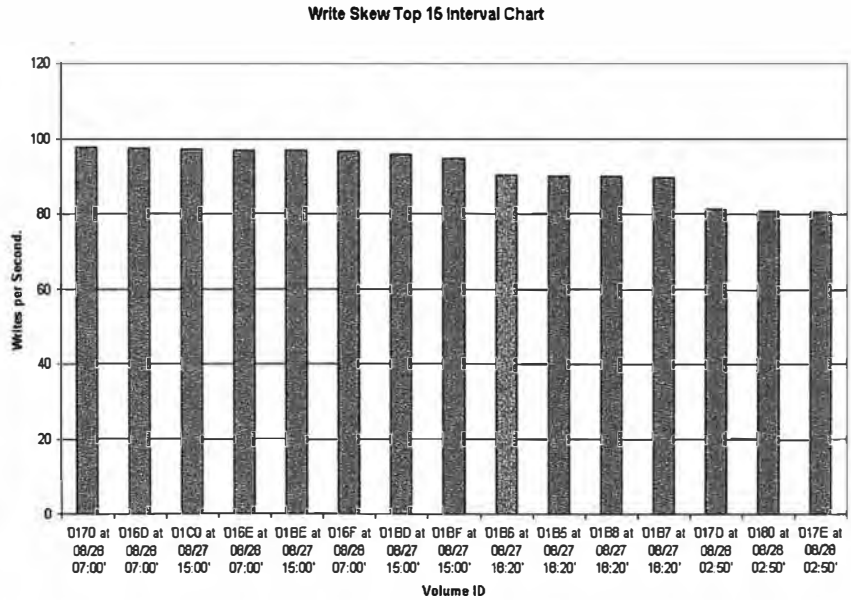


Figura 4.3: SRDF/S Write Skew Top 15 por Intervalos

La figura 4.3 muestra los 15 volúmenes con mayor actividad de escritura de todo el storage, si es que existiese algún problema de performance luego de la replicación es más que seguro que uno de estos volúmenes estará envuelto en el problema. Aun aunque no se pueda asegurar que estos volúmenes tendrán problemas de performance por la replicación, para esta lista de los 15 volúmenes top es recomendado:

- Identificar las aplicaciones escribiendo en esos volúmenes
- Distribuir la carga de trabajo en una mayor cantidad de volúmenes ya sea a nivel de SO (host striping) o a nivel de storage mediante metaluns (array striping) o distribuir esta en más luns.

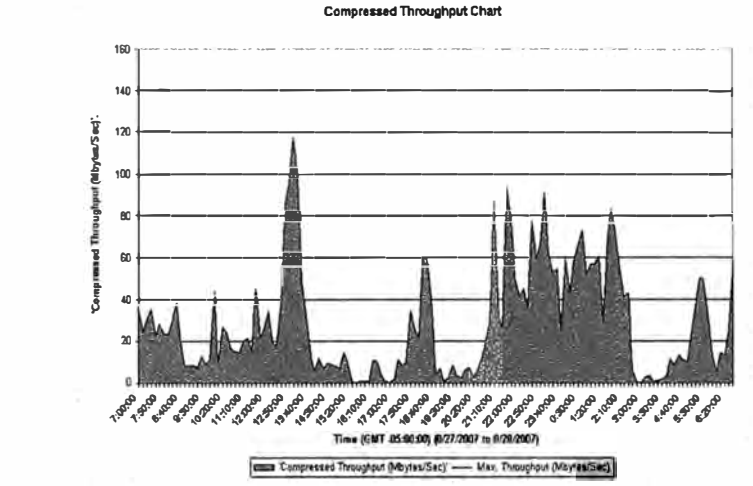


Figura 4.4: SRDF/S Compressed Throughput

La figura 4.4 muestra el throughtput comprimido (taza de compresión 1:1) donde planificamos según las mejores practicas utilizar solo el 70% de cada enlace, es decir, 70%(2x100MB/s)=140MB/s. Estos 140MB/s son suficientes para la replicación SRDF/S hacia Chaclacayo.

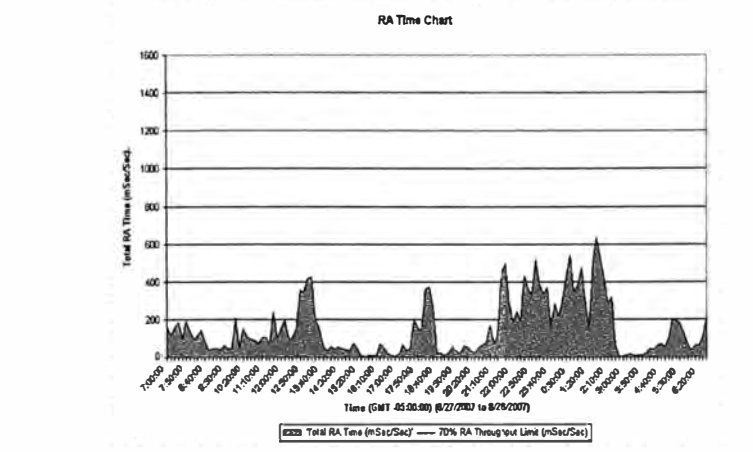


Figura 4.5: SRDF/S RA Time

La figura 4.5 nos muestra que los milisegundos de ocupación de las 2 RAs (nuevamente 2x1000mseg x 70%=1400mseg) que se posee hacia Chaclacayo son suficientes. En general se puede concluir que la replicación Sincrónica hacia Chaclacayo no tendrá problemas y esta correctamente dimensionada.

4.1.3.b Resultados SRDF/A a Trujillo

El caso del SRDF/A hacia Trujillo debido al gran tamaño 10TB de la Base de Datos y los cambios que esta tiene durante el día, son necesarios enlaces OC3 (155Mbps o 15.5MB/s). En este caso el manipular la cantidad de OC3 moverá también el RPO alcanzado y la cache adicional necesaria, por lo que a partir de los resultados que obtendremos se debería hacer un estudio de costos, lo cual no es uno de los alcances del presente informe. Finalmente el Banco UNI deberá decidir haciendo un balance basados en los SLAs (RPOs) que necesita y los costos que incurren adquiriendo más cache o arrendando más enlaces OC3.

Análisis para una Conectividad de 6 enlaces OC3

Empezamos por la solución más cara, pero que a la vez ofrecerá el mejor RPO aun a los 560Km de distancias entre datacenters.

El resumen de la carga de trabajo viene siendo la misma que la del SRDF/S, pues esta es dependiente solo del storage origen.

Resumen de la carga de trabajo:

Tamaño de Bloque de Lectura Promedio	61.0KB
Tamaño de Bloque de Escritura Promedio	29.0KB
Máximo IOs por segundo	16,757
Máximo IOs de Escritura por Segundo	2,695

Tabla 4.4: Resumen carga de trabajo SRDF/A

El Resumen de resultados para SRDF/A con 6 enlaces OC3:

Máximo Throughput Comprimido Requerido	64.08MB/seg.
Máximo Tamaño del RPO	3.75 Gbytes
Máximo Tiempo de RPO	00:01:00
Máxima Cache de Paginación de Disco Requerida	0.00 Gbytes
Máxima Cache Total Requerida	8.23 Gbytes

Tabla 4.5: Resumen de resultados SRDF/A con 6 enlaces OC3

Aquí el IO Profile y la lista de los 15 Volúmenes top de escritura son las mismas que las del SRDF/S, pues igualmente estas son dependientes solo de la carga de trabajo aplicada en el storage origen.

Por el contrario para los gráficos de Throughput varia en relación al SRDF/S, pues esta solución SRDF/A hacia Trujillo es realizada vía los Remote Adapters MPCD con protocolo GigE, es decir transporta IP, la cual permite planificar y esperar una compresión de 2:1. En la figura 4.6 en el lado izquierdo se muestra el throughput sin compresión para los 6 enlaces OC3, planificados al 70% de su capacidad (6x15.5MB/sx70%=65.1MB/s), se puede ver claramente que sin compresión los 6 enlaces OC3 no son suficientes.

el RPO de 1minuto es alcanzado durante todo el día sin inconvenientes, a pesar de los 560Km de distancia entre datacenters.

Finalmente la figura 4.8 muestra que la cache adicional requerida para esta solución es de 8.23GB adicional, por lo que se tendrá que adquirir 2 tarjetas de 8GB para el storage origen.

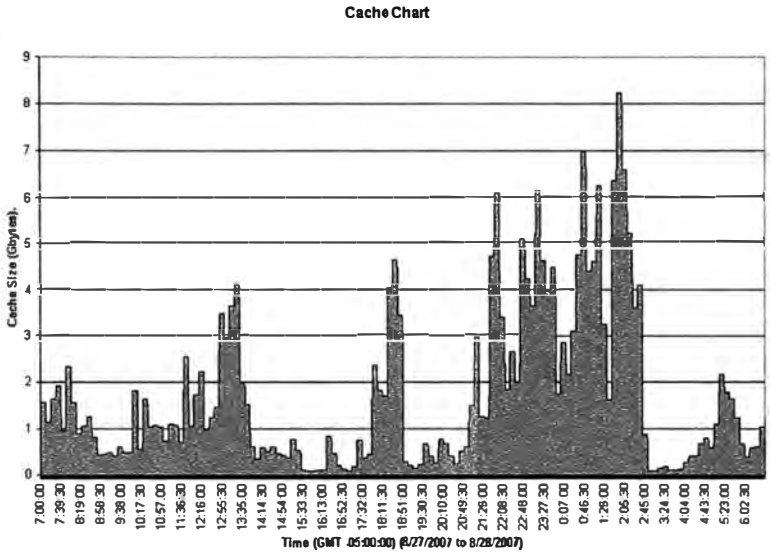


Figura 4.8: SRDF/A Cálculo de cache adicional con 6 enlaces OC3

Análisis para una Conectividad de 4 enlaces OC3

Continuamos con la solución más equilibrada y factible a la vez

El resumen de la carga de trabajo viene siendo la misma que la del SRDF/S, pues esta es dependiente solo del storage origen.

El Resumen de resultados para SRDF/A con 4 enlaces OC3:

Máximo Throughput Comprimido Requerido	64.08MB/seg.
Máximo Tamaño del RPO	41.44 Gbytes
Máximo Tiempo de RPO	00:15:56
Máxima Cache de Paginación de Disco Requerida	0.00 Gbytes
Máxima Cache Total Requerida	47.45 Gbytes

Tabla 4.6: Resumen de resultados SRDF/A con 4 enlaces OC3

El resumen nos muestra que se necesitan 47.45GB de cache adicional y que el máximo RPO durante los picos de carga de trabajo se incrementa a 15:56min (16minutos).

En la figura 4.9 se puede visualizar que estos 4 enlaces OC3 no son suficientes inclusive con la taza de compresión 2:1 para soportar los picos de la carga de trabajo, según el

funcionamiento y como se vio en el capitulo III estos picos que sobrepasan el máximo throughput son asumidos por la cache, es por ello que el adicional de cache se incrementará.

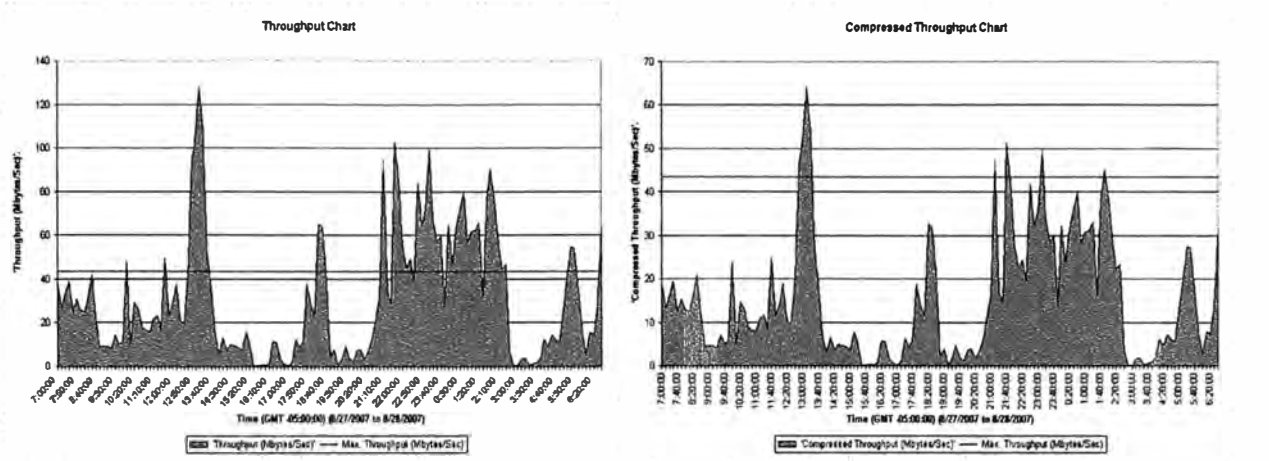


Figura 4.9: SRDF/A Throughput sin compresión (izq) y comprimido razón 2:1(der) con 4 enlaces OC3

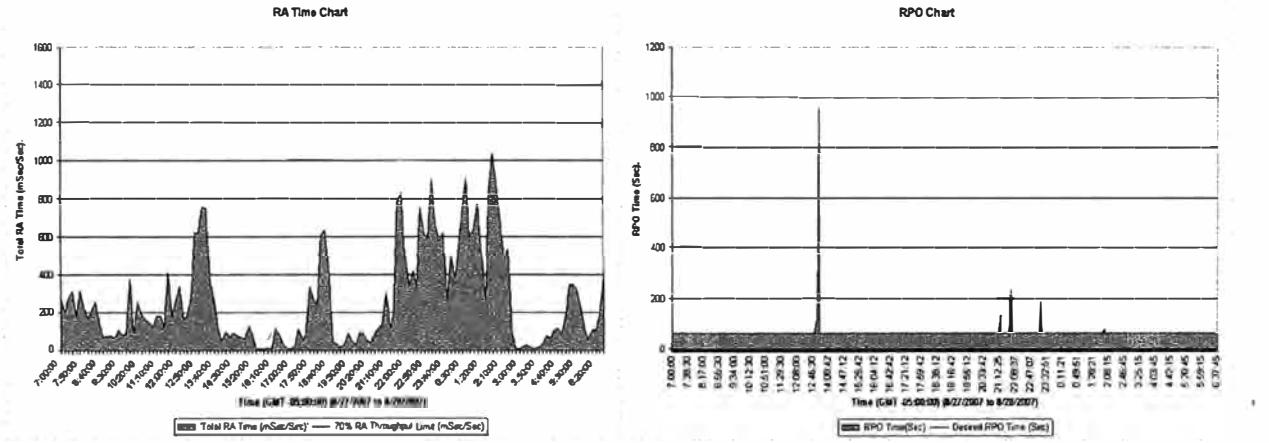


Figura 4.10: SRDF/A Tiempo RA (izq) y RPO obtenido (der) con 4 enlaces OC3

La figura 4.10 lado izquierdo muestra que los 2 Remote Adapter son suficientes para la solución y en la misma figura en el lado derecho se puede observar que el RPO durante los picos de carga de trabajo se incrementa a 956segundos (15:56min), es decir el peor RPO que ofrece la solución es de 16minutos.

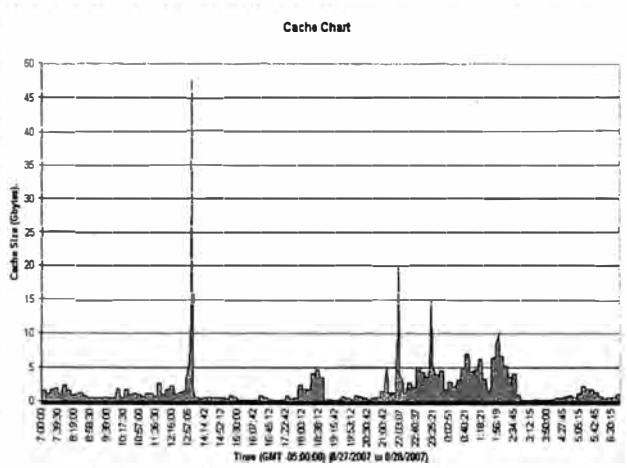


Figura 4.11: SRDF/A Cálculo de cache adicional con 4 enlaces OC3

La figura 4.11 muestra que la cache adicional necesaria durante los picos es de 47.45GB (2tarjetas de 16GB+2tarjetas de 8GB=48GB Total adicional). Es decir esta es la máxima cantidad de cache adicional necesaria para la solución.

Análisis para una Conectividad de 2 enlaces OC3

Continuamos con la solución más económica, pero a la vez no viable.

El resumen de la carga de trabajo viene siendo la misma que la del SRDF/S, pues esta es dependiente solo del storage origen.

El Resumen de resultados para SRDF/A con 2 enlaces OC3:

Máximo Throughput Comprimido Requerido	64.08MB/seg.
Máximo Tamaño del RPO	292.94 Gbytes
Máximo Tiempo de RPO	03:45:09
Máxima Cache de Paginación de Disco Requerida	0.00 Gbytes
Máxima Cache Total Requerida	731.65 Gbytes

Tabla 4.7: Resumen de resultados SRDF/A con 2 enlaces OC3

Si bien algunos negocios pueden tolerar perder 3 o 4 horas de data (RPO=3h o RPO=4h), un negocio del tamaño del Banco UNI con 10TB de tamaño de Base de Datos no es permisible perder tanto tiempo de data.

Más determinante aun es el hecho que esta solución necesita de 731.65GB de adicional lo cual es imposible, pues como se indicó el DMX3 soporta como r 512GB en total para todo el storage.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1.- Como se vio para el diseño de un almacenamiento con replicación remota para Continuidad de Negocios y Recuperación de Desastres tanto para replicación Sincrónica como Asíncrona es un factor importante y determinante para conseguir una implementación exitosa el realizar un buen diseño, planeamiento y dimensionamiento de los factores claves (Ancho de banda, cantidad de puertos Remote Adapter, latencia, posibles Volumenes con posibles problemas, Cache adicional, RPO, etc.) según el tipo de replicación. El no dimensionar correctamente estos factores claves puede traer diferentes problemas en la implementación y provocar la caída del proyecto. Con el modelo expuesto se pueden dimensionar las características principales y necesarias para el correcto funcionamiento de la replicación, así mismo el modelo advierte aquellas características que no puede dimensionar.

2.- Un análisis adicional del estado actual del rendimiento de todo el almacenamiento debe llevarse a cabo conjuntamente con el planeamiento propuesto en el presente informe.

3.- Como se vio en el caso de estudio del SRDF/A se puede lograr un balance entre el RPO requerido y los costos que implican la cache adicional y los enlaces de telecomunicaciones para las distancias que se manejan en este tipo de replicación.

4.- Si durante el análisis del planeamiento con el modelo propuesto se hallase volúmenes con alta carga de escritura, se requerirá que estos sigan un proceso de reacomodo a nivel de servidor o a nivel de almacenamiento.

5.- Para conexiones IP arrendadas a operadores de telecomunicaciones es altamente recomendado realizar una evaluación de la calidad de la red (network assessment).

RECOMENDACIONES

6.- Recomendaciones y/o advertencias para el diseño de ambas replicaciones:

- Si el storage origen esta configurado para ser "más rápido", o más grande que el destino, puede causar problemas de performance.
- Si existen menos discos físicos en el destino esto puede originar un "cuello de botella" en el storage destino. La cantidad de discos físicos replicados en el storage destino deberían ser iguales o mayores a la cantidad de discos físicos replicados en el storage origen. Esto aplica solo para los discos replicados dentro del almacenamiento.
- Evitar tener un tipo de protección "más lenta" en el storage destino, ya que esto puede causar cuellos de botella en el storage destino.
- Si se tienen muchos storage origen congregándose a un solo storage destino se pueden originar cuellos de botella en este.
- Puertos adicionales RDF podrían ser necesitados para manejar las cargas de trabajo pico.
- Si se tienen discos más rápidos en el storage origen esto puede originar cuellos de botella en el storage destino.
- Los niveles de micro código en los storages deben utilizarse los últimos aprobados y del mismo nivel en ambos.
- GigE Throttle es el valor que ayudara a prevenir que los puertos GigE sobresaturen la red IP, este se ajusta a la cantidad máxima de la conexión IP arrendada.
- El Throughput depende del tamaño del IO, nunca se debe planificar al 100% de utilización del enlace, la mejor práctica indica que se debe planificar al 70% de utilización de la capacidad del enlace.

7.- Recomendaciones y advertencias para el diseño, planeamiento e implementación para replicación sincrónica SRDF/S son:

- Si el ancho de banda WAN es insuficiente el rendimiento sufrirá
- Para soluciones SRDF/S sobre FC se recomienda planificar hasta 4 enlaces (de ser posibles) al 50% de utilización, esto reducirá el tiempo de respuesta que las aplicaciones obtendrán de los volúmenes cuando ya estén replicándose.
- Las diferencias resaltantes al momento de escoger FC o GigE: FC brinda mucha menor latencia, pero FC hace 2 round-trips en comparación a 1 solo round-trip que hace GigE, FC además tiene más alta capacidad para soportar IO (4Gbps vs 1Gbps

de GigE), adicionalmente la curva de comportamiento es mucho más plana a niveles muy altos de IO/s.

- Se recomienda no mezclar trafico SRDF/S con SRDF/A en los mismo puertos RA (Remote Adapters), ya que esto impactaría negativamente el rendimiento del SRDF/S

8.- Recomendaciones y advertencias para el diseño, planeamiento e implementación para replicación sincrónica SRDF/A son:

- La correcta implementación de una solución SRDF/A requiere planeamiento, este planeamiento comprende: evaluar la carga de trabajo desde los servidores a replicarse, se necesita comprender las características de esta carga trabajo, siendo relevante identificar las horas pico y su duración, perfil de IO, etc. Los volúmenes lógicos con mayor data a transmitir serán los que determinaran el RPO.
- Para poder garantizar "no más que 60 segundos de RPO" se necesita realizar una evaluación fina con las consideraciones listadas anteriormente de incluir en el análisis días pico del negocio en una semana, en un mes, en todo el año.
- El ancho de banda SRDF/A provisto debe ser mayor o igual en promedio a la carga de trabajo de escritura; caso contrario la replicación SRDF/A posiblemente se corte.
- Si en la solución el storage origen tiene insuficiente cache, el SRDF/A se caerá debido a que se alcanzará el limite "Write Pending Limit" y con ello la replica será cortada. Para una sincronización después del corte se necesitaría una copia completa full, la cual tomara unas horas según la cantidad de data a replicar.

9.- Después de implementada la solución es necesaria una continua y periódica revisión de los factores claves previamente diseñados, estos deberían mostrar los elementos claves en niveles saludables, caso contrario se debe tomar medidas correctivas.

10.- El BCP o Business Continuity Planning contienen las mejores practicas globales para implementar Recuperación de Desastres en todas las Fases, estas fases se enumeran en el capitulo I, entre las cuales se pueden destacar que no solo son suficientes contar con procedimientos escritos que permitan reestablecer la data replicada, si no también testearlos, así como tener el personal capacitado que pueda ejecutar estos procedimientos, a su vez la locación de este personal no debería ser el site principal.

11.- Finalmente se recomienda tener en consideración los crecimientos esperado de la carga de trabajo y el tamaño de información replicada al momento de realizar los cálculos en la fase del planeamiento con el modelo descrito, estos márgenes de seguridad considerados adicionalmente nos permitirán soportar circunstancias atípicas de la operación, las cuales no hallan sido incluidas en las estadísticas recolectadas.

BIBLIOGRAFÍA

1. Eileen Colkin, "Partners you can count on", Informationweek.com – USA, 2001

2. John R. Harrald, "What Have We Learned from The September 11th - Panel contributions", Bledconference.org – USA, 2002

3. Mary Brandel, "Mesaba learns that disaster planning never really ends", snwonline.com – USA, 2003

4. Mark Farrington, "Business Recovery Following Buncefield Oil Fire", Symphony Business Solutions – UK, 2007

5. Disaster and Recovery Journal and DRI International, "Generally Accepted Practices For Business Continuity Practitioners", Disaster and Recovery Journal and DRI International – USA, 2007

6. Gartner, "Survey Confirm There are Many Effective Disaster Recovery Strategies IDG00126421", Gartner – USA, 2005

7. EMC Corporation, "EMC Networked Storage Topology Guide", EMC Corporation - USA, 2007

8. EMC Corporation, "Clariion Foundations", EMC Corporation – USA, 2007

9. EMC Corporation, "Symmetrix Foundations", EMC Corporation – USA, 2007

10. EMC Corporation, "SAN Foundations", EMC Corporation – USA, 2006

11. EMC Corporation, "Speed Guru Qualification Documentation", EMC Corporation – USA, 2007

12. EMC Corporation, "Business Continuity over Optical Networks", EMC Corporation – USA, 2004

13. EMC Corporation, "Symmetrix Business Continuity - SRDF Solutions", EMC Corporation – USA, 2005

14. EMC Corporation, "EMC Symmetrix Remote Data Facility (SRDF) Connectivity Guide", EMC Corporation – USA, 2007

15. EMC Corporation, "SRDF Connectivity Solutions", EMC Corporation – USA, 2004

16. Mike Lawrence, "EMC Replication Modeling Tool (ET) – SRDF, EMC Corporation – USA, 2007