F

White Paper

We make sure

FUJITSU COMPUTERS SIEMENS

SAN Interworking

Issue December 2004

Contents

1. Introduction

The role of SANs (Storage Area Networks) is becoming increasingly important in storage configurations. Fibre channel switches are the central components of a SAN. The servers and storage systems are connected via fibre channels to these fibre channel switches, thus allowing each server to access each storage area. The switch architecture (non-blocking, any-to-any connections) makes it possible to execute block-oriented accesses very efficiently. According to analyst's oppinion, around 50% of all large companies already have SANs installed. In medium-sized companies, this is around 25%. SANs have considerable advantages over DAS (Direct Attached Storage), the most important of which are:

- SAN provides much greater flexibility and uses the resources more efficiently since any storage area can in principle be assigned to any server. In a fully networked configuration, assignment is made via software tools.
- SANs can be scaled better than DAS configurations. This applies both to simply adding more storage and to increasing the number of addressable storage units. In addition, the distances between a server and a storage unit in SANs can be far larger.
- Administration can be implemented more efficiently from a central point. This appreciably lowers the operating costs.

SANs provide an excellent basis for implementing globalization, disaster recovery and interworking solutions as well as for adding optimization solutions.

2. Global scenarios

Three basic scenarios are described below.

2.1 Interworking between SAN islands

So-called SAN islands came into being in many companies in the past. The reasons for this were either intentional encapsulation of applications, were caused by merging companies with existing SAN configurations or came about in company structures that had one headquarters and a number of independent branches. SAN islands can be found locally in a computer center environment but may also be distributed over different locations. The increase in process integration and improved utilization of resources also raises more need for interworking between the SANs. However, this interworking between the SANs is not intended to merge the SAN islands together to form one big SAN since this could, in certain circumstances, have a negative effect on the flexibility and availability of the separate SANs. This requirement can only be satisfied with protocols such as iFCP that is described briefly in the appendix. Some vendors base their interworking solutions not on iFCP but on special optimized protocols.

2.2 Connections over large distances

SANs have increased the limits of storage access technologies to a very high degree. For example, distances of up to 35 km can be covered today without additional amplification or conversion components, if appropriate fibre optics can be laid between the separate sites. The maximum distance that can be covered is thereby determined by both the quality of the fibre optics and the performance of the access components (the components that generate and receive the laser beam). Distances of 20 or 35 km are not approved for all devices. The so-called DWDM technology can bridge distances up to 100 km. In certain forms of application (e.g. synchronous I/O mirroring) a dedicated fibre optic or DWDM connection may be preferable in terms of latency times and quality-of-service, since protocol conversion is carried out on the lower protocol layers and this only marginally increases the latency times. However, both the accompanying restriction in distance and the costs for such connections may cause problems.

The technology available today provides the means of transferring fibre channel data over the TCP/IP network and offers a solution for both greater distances and for configurations where it is not possible to provide dedicated fibre optics from point A to point B. It also has a positive effect on the costs since widely available TCP/IP infrastructures can be employed.

2.3 TCP/IP networks for block-oriented I/Os (IP storage)

The two scenarios described above were both based on a SAN infrastructure, i.e. a network that differs technically from a client/server network based on TCP/IP. The development and standardization of the iSCSI protocol now

also make it possible to use the TCP/IP network for transporting block-oriented data streams. This results in the following advantages:

- Reduced costs in designing, implementing and analyzing storage networks
- Reduced management effort by using just one type of network
- Widely available know-how with TCP/IP networks
- The TCP/IP and Ethernet market provide better economics-of-scales, since the number of components produced is many times higher than that of fibre channel components
- New technologies are made available much faster in the mass market (TCP/IP and Ethernet technology) than in specialized markets (fibre channel)
- In principle, no distance or addressing limits

3. Usage scenarios in detail

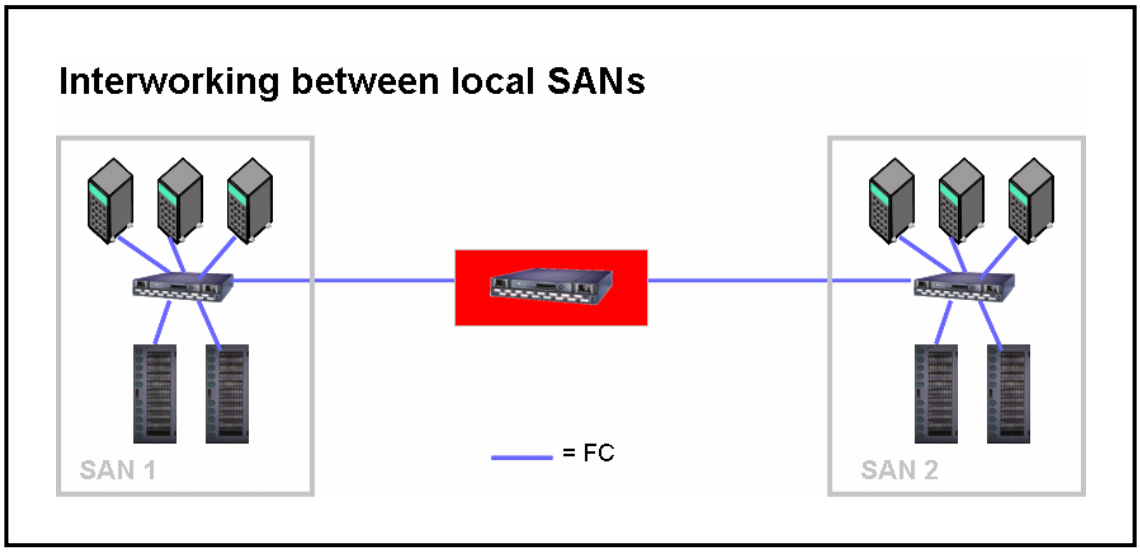
This White Paper is based on the "Storage Connections over Large Distances" and "iSCSI, ready for the Enterprise Area" White Papers, that describe the basic techniques in detail.

3.1 Interworking between SANs

3.1.1 Local interworking

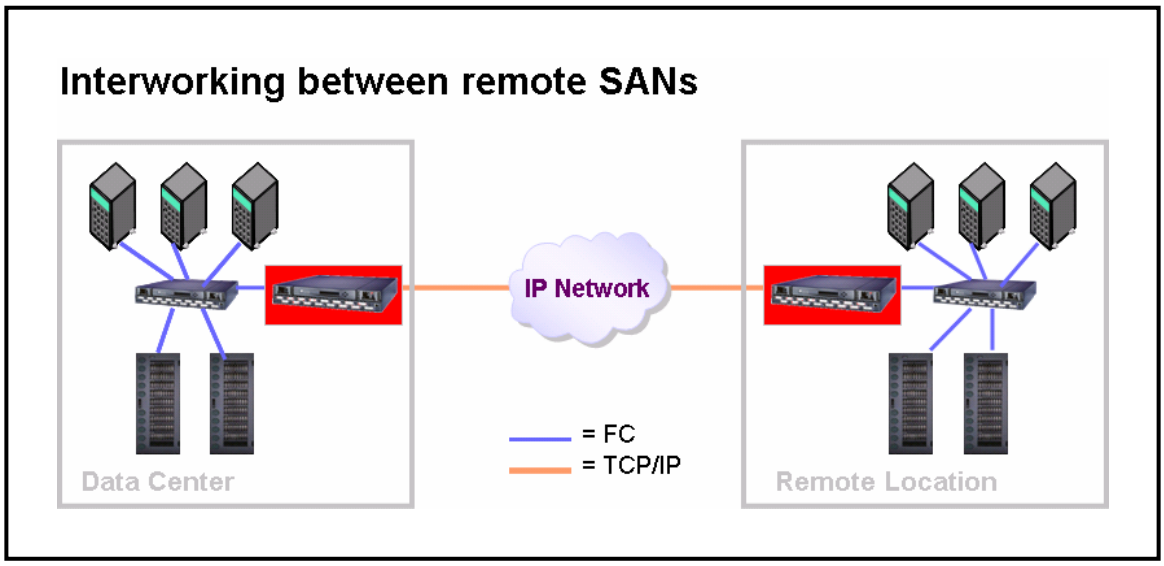
Two separate SANs, with at least one switch each are assumed. The requirement is formulated as follows: separate devices in SAN B must be accessible from SAN A without merging the two separate SANs together into a single fabric. A fabric is a SAN with at least two switches that represent one uniform address space. A real scenario could be, e.g. accessing the tape peripherals in SAN B, which may be in another firewall section. This is possible with the iFCP protocol or with other proprietary protocls. To accomplish this, one fibre channel E port connection is set up between SAN A and a SAN interworking device and one fibre channel E port connection between SAN B and the corresponding SAN interworking device. The autonomy of each fabric is maintained by this SAN interworking devices since the E port functionality over the iFCP protocol is locally terminated. This has the following advantages:

- Errors remain isolated in the fabric concerned. The terminal devices in fabric B, that are visible to fabric A are seen and handled as terminal devices that are connected locally to fabric A.
- Reconfigurations also remain isolated to one fabric. If, for example, an additional switch is installed in a fabric, this may interrupt data traffic for several seconds since each time a switch is added or removed the address assignment has to be renegotiated for any address conflicts that occur. This process is standardized in the fibre channel protocol such that fabric self-configuration can be implemented. This automatic address assignment has the advantage that no manual routing tables have to be maintained, but the disadvantage that the process does not execute without interruption.
- Due to the error and reconfiguration isolation, the number of switches in the fabric can be kept relatively low, thus reducing the time required for a fabric to reconfigure itself.
- The name server tables can be kept relatively small since it is only necessary to enter the zoned devices and not wholesale every device from the other fabrics.



3.1.2 Remote interworking

The same assumptions apply as in the previous scenario except that it is not possible to set up fibre channel connections for the E port connections either due to distance or for other reasons (price, non-availability). However, it is assumed that a TCP/IP connection with comparable quality-of-service and bandwidth (here, e.g. GbE) is available. On the one hand, the data from fabric A is transformed with the iFCP or equivalent protocols such that it can be transferred via TCP/IP to the second device in the vicinity of fabric B. The same advantages are present as in the previous scenario. However, the fault isolation is even more important in this scenario than in the previous one because the error probability increases tendentiously with distance.



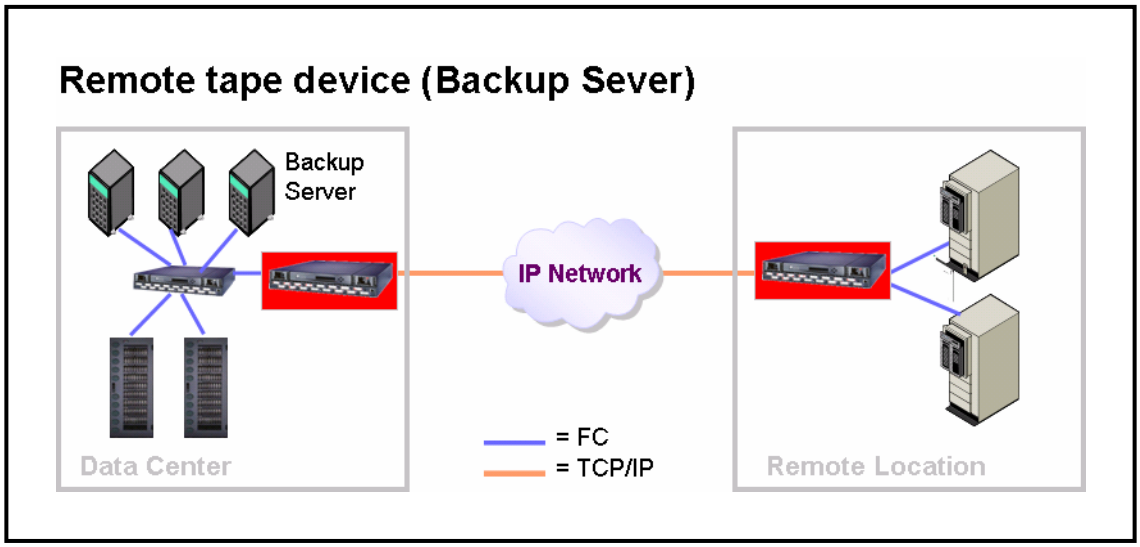
3.2 Connections over large distances (disaster protection)

The following scenarios are based on the principle that the SAN interworking devices convert the fibre channel data stream into a TCP/IP data stream, thus allowing the TCP/IP network to be used for transporting the data. Currently, fibre channel devices can only be supplied with a bandwidth of 1, 2 or 4 Gbps, whereas there is a very wide bandwidth spectrum available in the TCP/IP network that range from just a few Mbps with end-to-end connections up to 10 Gbps, particularly in backbones. This provides the option, where the requirements are low, of also working with bandwidths below 1 Gbps, which has an appreciable effect on the cost structure. On the other hand, backbones can be utilized in an optimum way via link aggregation (parallel usage of several connections).

With all models shown below, it is assumed that the required bandwidth and the quality-of-services is very carefully planned for each scenario. This should be done at the customer, together with the network experts.

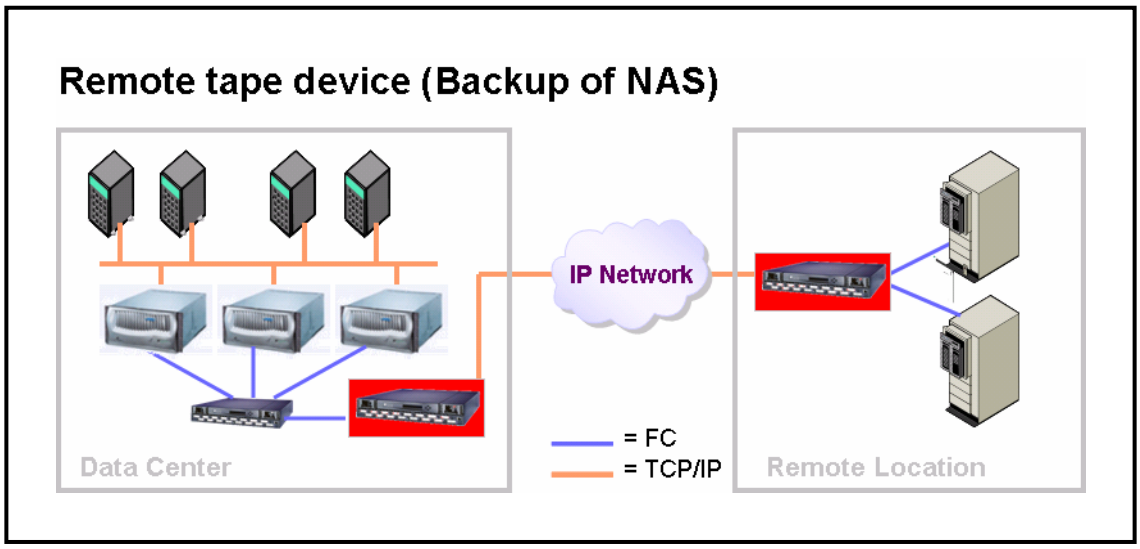
3.2.1 Remote tape peripherals (backup servers)

Disaster protection considerations make it necessary to transfer the backup data to a remote site in addition to local backup data. SAN interworking devices make it possible to use the TCP/IP network to bridge the distances involved. Since writing backup data to tape drives (or other media) is basically an asynchronous process, distances far greater than 100 km can be bridged. Installations exist where the backup data has been transferred efficiently over several thousand kilometers. Throughput can be increased with additional functions for such applications. On the one hand, the data can be compressed before output to the TCP/IP network. This allows smaller bandwidths to be used, thus reducing costs. On the other hand, SAN interworking devices can optimize the necessary acknowledgement traffic. The basic principle lies in just one acknowledgement being required on the network side for a complete SCSI request but in the server direction all sub-requests have already been locally acknowledged, thus practically allowing a "streaming mode" within the overall SCSI request. Due to the protocol characteristics, the effect of such optimizations increases as the distance increases (e.g. greater than 250 km).



3.2.2 Remote tape peripherals (NAS backup)

It is generally of no consequence whether remote backup peripherals are provided for a backup server or an NAS system. However, these must be fibre channel peripherals in both cases.



3.3 Data replication

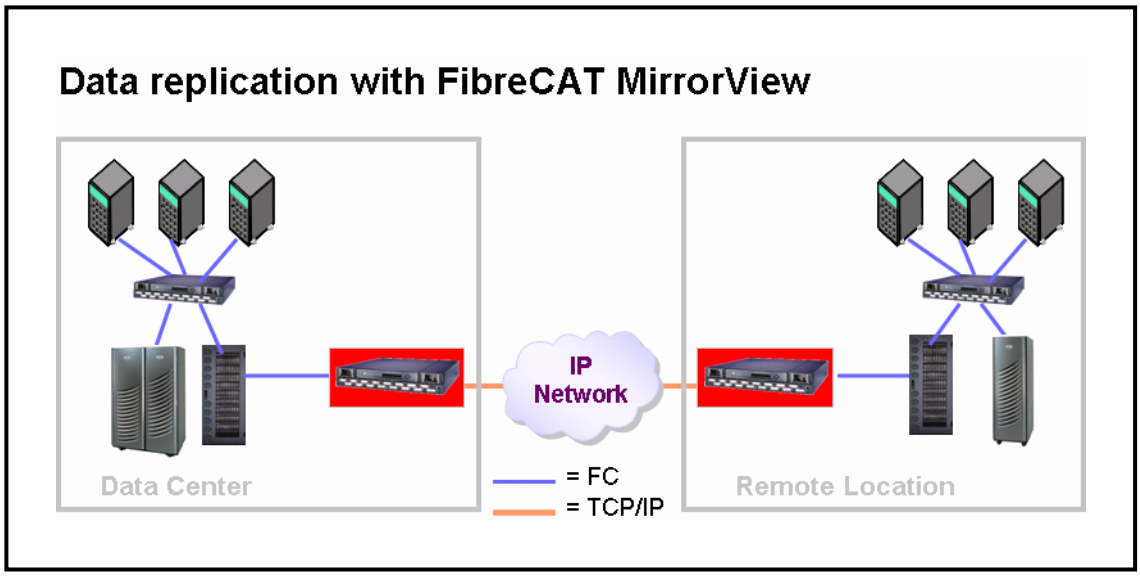
3.3.1 FibreCAT MirrorView

FibreCAT CX Systems offer synchronous data replication with MirrorView. In a local case, the connection is via a fibre channel from system A to system B. There are limits in the maximum distances with fibre channel

connections. Additional components will be needed if distances greater than these limits have to be bridged. In general, connections over greater distances can be made with DWDM.

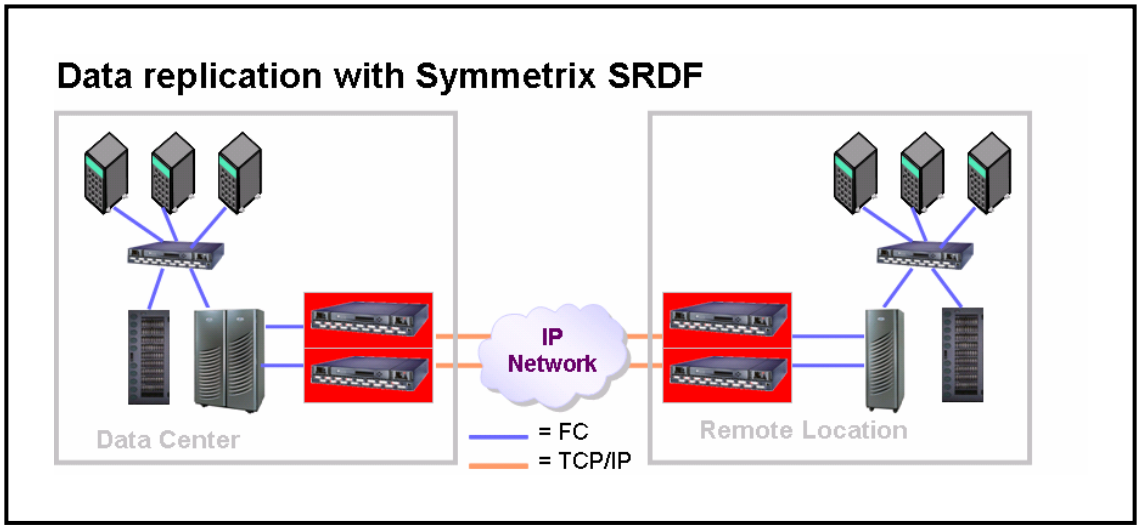
SAN interworking devices can be used to provide a solution that is simpler and generally more cost-effective. SAN interworking devices have to be qualified by EMC for MirrorView and have to be listed in their support tables. SAN interworking devices have two advantages, on the one hand they use the TCP/IP protocol to transport data over large distances, thus allowing an existing network infrastructure to be used. On the other hand, no additional switches are required, if SAN interworking devices themselves provide the necessary buffers.

Latency times play an important role with synchronous data replication. Additional latency times are implicit in the SAN interworking devices since the protocol conversion is at a relatively high level. In addition, TCP/IP connections can be made over various connection types (permanent lines, ATM, direct entry into the carrier backbone, etc.). Dedicated IP links whose connections are made either without or via a low number of high performance routers/switches provide a good basis for satisfying the quality-of-service requirements. However, this requires detailed planning and binding arrangements with the carrier. The general recommendation is to consider native fibre channel or DWDM connections as the preference for response time-critical applications.



3.3.2 Symmetrix SRDF

Synchronous data mirroring is implemented by Symmetrix with SRDF (Symmetrix Remote Data Facility). The same comments made for MirrorView with FibreCAT CX apply for SRDF (also with respect to qualification).

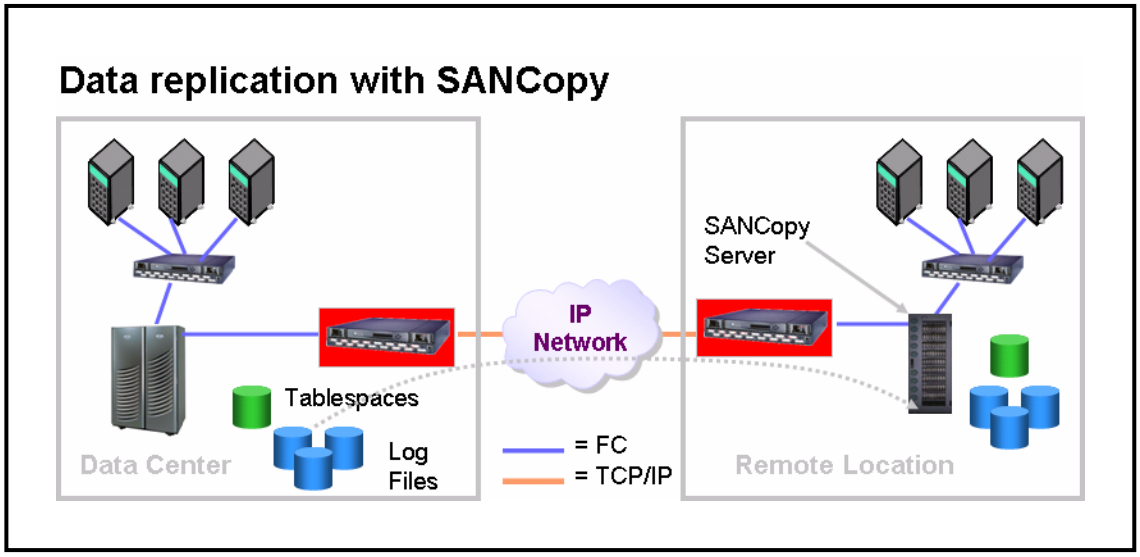


3.4 Data replication with SAN Copy and Open Replicator for Symmetrix

The products EMC SAN Copy and Open Replicator for Symmetrix copies complete LUNs between FibreCATs and Symmetrix systems and vice versa. The SAN Copy server must thereby be installed on a FibreCAT system. SAN Copy and Open Replicator for Symmetrix can be employed efficiently in database environments. The database comprises table spaces (with the actual database) and the log files. Databases provide the option of closing the current log file after a specific time or after a specific number of transactions, while simultaneously opening a new log file. The following steps must be carried out to keep a shadow database up to date over a large distance:

- Initially, the LUNs are transferred once to the remote site together with the table spaces.
- Each time a log file is closed, the relevant LUN is then also transferred to the remote site.
- After the log files have been transferred, the log file information is read into the table spaces at the remote site using database methods. This creates a consistent, always up to date database inventory that can be used immediately to continue working in disaster cases.

SAN interworking devices can also be used here for bridging large distances.



3.5 CentricStor

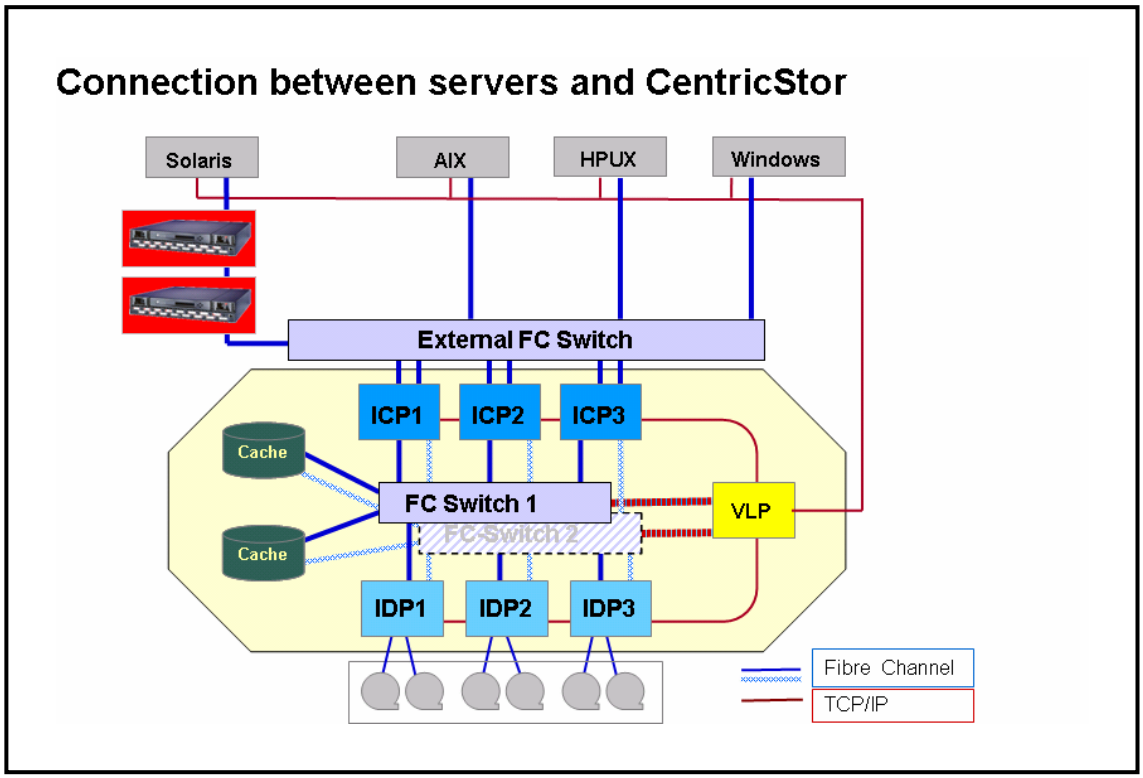
CentricStor is an appliance that virtualizes tape devices and the robotic control of archive libraries. The patented concept guaranties highest performance, scalability and availability. The core components are built of standard Intel servers, standard RAID systems and standard fiber channel products. The tape drives used are LTO, IBM-Magstar and StorageTeck-9940A/B and the robot systems are the widely used products from ADIC and StorageTek and IBM.

The advantages of CentricStor can be summarized as follows:

- It is no longer necessary to maintain different drivers on the application servers. Under UNIX, just one proven, stable tape driver (e.g. generic SCSI driver) suffices for communication with the virtual tape system
- CentricStor realizes an unmatched liability in comparison to server-based solutions (dynamic drive sharing, backup from disk to disk, staging, …).
- Tape drives and media are efficiently used. Especially the streaming mode guaranties highest performance reduces wear on the tapes and the drives
- The high degree of parallelism (up to 512 virtual tape drives), the speed matching function of CentricStor's internal cache and the intelligent cache management reduce dramatically backup and recovery times.
- A number of routine jobs can be carried out autonomously by CentricStor by means of policy defaults. Examples of this are volume replication, media refreshing, compressing tape contents if separate volumes are released due to retention periods, etc.

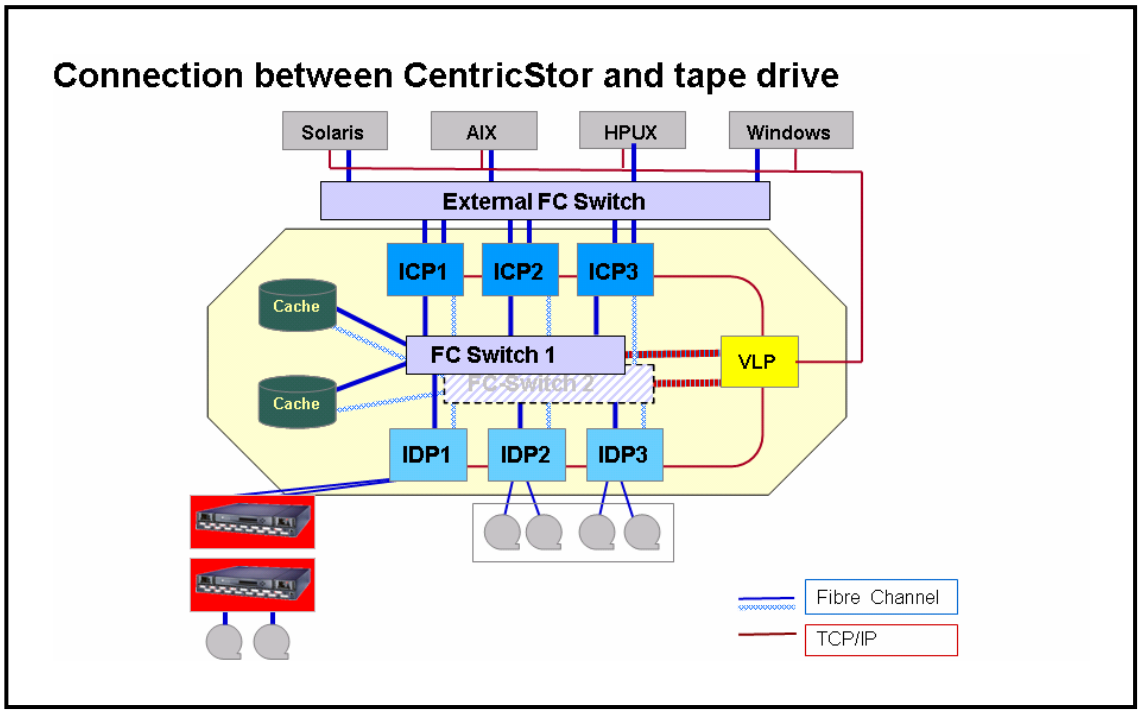
3.5.1 Connection between servers and CentricStor (ICP)

A switched fibre channel can lie between the servers and CentricStor. Each connection within the switched fibre channel can be extended with SAN interworking devices, allowing connections to be made relatively simply over large distances as well as over other private or public grounds. The figure below shows the SAN interworking devices in front of the external CentricStor switch. However, the most practical case is probably where several switches are in front of the CentricStor with one at a great distance or if all connections of the external switches to the ICPs are fed via the SAN interworking devices.



3.5.2 Connection between CentricStor (IDP) and tape drive

A point-to-point fibre channel connection is possible between CentricStor (IDP) and one tape drive. The distance that this connection covers can also be extended with the SAN interworking devices. In this scenario interworking requires however a transparent protocol such as FCIP.



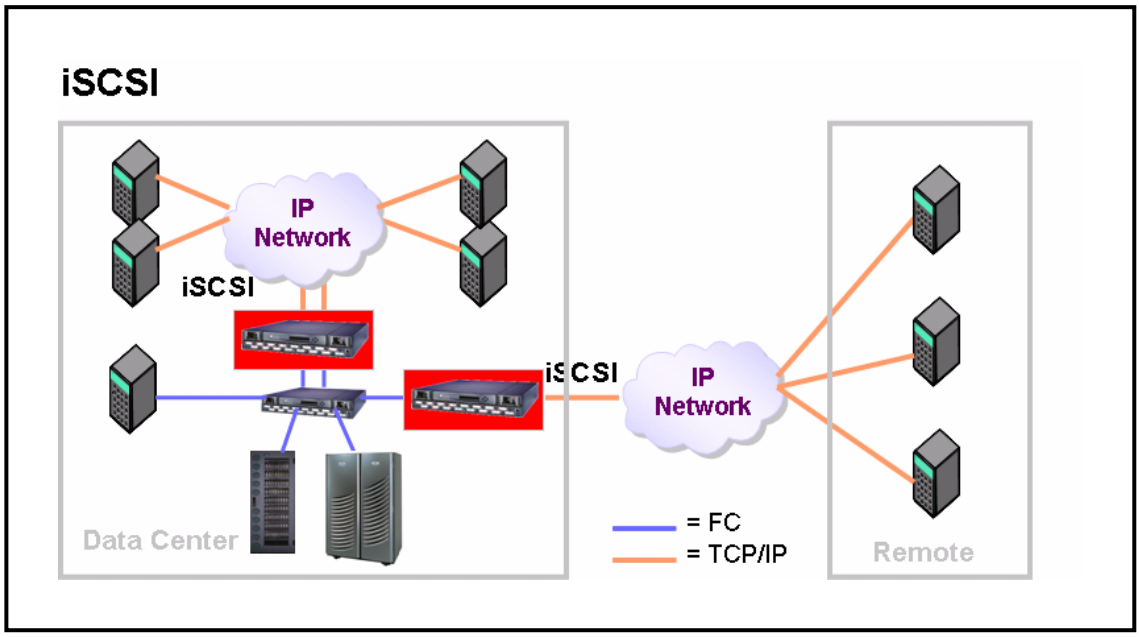
3.5.3 Connection between CentricStor internal switches

In many cases it is desirable to have some CentricStor components locally and some in a remote site. The purpose is to have the data duplicated to the remote site and have access to this data in the event a disaster should happen. The connection among the CentricStor internal switches has to maintain a segregated fabric combining the local and remote CentricStor SAN components. In this case only a transparent protocol such as FCIP will work.

3.6 Block-oriented I/Os via iSCSI (IP storage)

Many SAN interworking devices can also be used as iSCSI routers. Assuming that neither the server nor the storage system has iSCSI functionality, the symmetric model is used. This means that the server is connected via fibre channel to a SAN interworking device that converts the data to iSCSI format and sends it via TCP/IP to the second SAN interworking device. This SAN interworking device receives the data via the TCP/IP interface and transforms it back to the fibre channel format. Since the second SAN interworking device is also connected via fibre channel to the storage system, the reconverted data can be passed to the storage system.

If the server or the storage system already has iSCSI functionality (either software or host bus adapter with TCP/IP offload engine), just one SAN interworking device is needed on the side that does not directly support iSCSI. This can result in high investment protection for the existing hardware.



4. Security

Physically encapsulated fibre channel networks constitute a very high obstacle for intrusive accesses from outside, although the potential risk of internal intrusion should also not be neglected. If such storage networks are provided with additive interfaces via TCP/IP or the complete infrastructure is converted to the TCP/IP network (e.g. with iSCSI), additional security measures have to be employed. Additional security measures in the form of Virtual Private Networks (VPNs) or firewalls are mandatory in particular for connections that leave the physically encapsulated area (remote data replication). But storage networks based on TCP/IP that are only encapsulated logically also have a higher risk potential than fibre channel since a great deal of know-how is widely available on weak points, in particular for the TCP/IP and Ethernet areas.

Appendix
Summary of technologies

5. Inter-SAN connections

At the present time, various methods exist for connecting separate fibre channel-based SANs with each other. Each method has different effects on the overall functionality of a SAN. The various protocols and their differences are discussed briefly below.

5.1 ISL (Inter-Switch Links)

The actual basic method of connecting separate fabrics together is establishing inter-switch links. To do this, the E ports of two switches are connected together with a fibre channel. As a result, the original separate fabrics are merged together to form one overall fabric. The reason for this is the self-configuration function of fabrics that is specified in official standards. For example, each fabric needs a master switch that is responsible for uniform address allocation. However, only one master is allowed in the newly formed fabric. A new master must therefore be defined for the overall fabric. In FC networks, the master definition and subsequent unique address allocation are made automatically. This interrupts the entire I/O traffic for a few seconds. The reverse process executes if the ISL connection is interrupted for several seconds. Two separate fabrics are recreated and each then has to define its own new master.

ISL connections do not primarily need an additional protocol converter, but there must be a fibre optic cable from SAN A to SAN B and there are certain length restrictions (see Global scenarios).

Even if the fibre optics are extended over a greater distance with DWDM (Dense Wave Division Multiplexing), the functionality behavior does not change since the DWDM method is completely transparent to protocols.

5.2 FCIP

The FCIP protocol (Fibre Channel over IP) implements a tunneling method in which FC frames are packed into TCP/IP packets so that they can be transported over a TCP/IP network. With this technique, it is relatively simple to connect separate storage network islands together over large distances. The method entails an FC switch with a protocol converter connected to it that packs the FC frames into TCP/IP packets. These TCP/IP packets are then transported to the recipient over the TCP/IP network. The recipient is again a protocol converter that is connected to a second FC switch. The protocol converter unpacks the FC frames from the TCP/IP packets and passes them to the FC switch. The FCIP protocol does not differentiate between FC frames containing data and FC frames containing control statements for coordinating a fabric. Therefore, FCIP exhibits the same self-configuration function behavior as described under ISL above. Poor quality FCIP connections can lead to frequent interruptions, and not just in an FCIP connection, but throughout the entire fabric.

5.3 iFCP

In contrast, the primary aim of the iFCP (Internet FC-Protocol) is connecting FC terminal devices over the IP network, where IP switching and router components replace the FC fabric services. iFCP is a gateway-to-gateway protocol. In contrast to the tunneling approach with FCIP, the separate fabrics are not merged together to form a single fabric. Error conditions that, e.g. lead to fabric reconfiguration, remain locally isolated since the E port functionality is practically terminated locally. As with FCIP, a minimum of two protocol converters is also required here.

5.4 iSCSI

The iSCSI protocol packs SCSI commands and data into IP packets. Transfer is over the TCP/IP network. In an ideal case, the conversion is carried out native in the application servers and storage systems.

On the application server side, this can either be done in a driver or in an extended host bus adapter. With the extended host bus adapters, the special protocol stack (TCP/IP protocol) required for the iSCSI procedure is processed on the host bus adapter hardware. The conversion in the host bus adapter can either be in the firmware (quasi software solution) or as a pure hardware implementation (in ASICs), which has a considerable influence on the performance. In each case, the application server load is lightened appreciably by not having to process the TCP/IP cycles that are relevant to iSCSI. A similar effect also results from connecting the server to an iSCSI router

via fibre channel. The fibre channel data conversion into the iSCSI protocol and TCP/IP protocol stack handling are then processed in the router. The server is not loaded.

On the storage systems side, the conversion can be carried out directly in the storage system itself or a storage router can be connected before the storage system and this then takes over the task.

6. VSAN (Virtual SAN)

In analogy to VLAN (Virtual LAN), products exist that also implement VLAN functions for SANs. In the most simple case, a large switch is split into several smaller switches via administration interfaces. However, VSANs can also extend over several switches. The basis for assignment to a VSAN is the ports. Splitting into several independent SANs ensures error and reconfiguration isolation. However, compared with the iFCP protocol, there is a severe disadvantage that interworking between the separate VSANs may not be possible.

The VSAN functionality is not standardized yet. As a result, there are only proprietary solutions currently on the market that are only available in special homogeneous environments.

Issue by:

Simon Kastenmüller
Telephone: ++49 (0)89 636 44241
Simon.Kastenmueller@fujitsu-siemens.com
http://www.fujitsu-siemens.de/

Extranet:

http://extranet.fujitsu-siemens.de/