fF
White Paper

We make sure

FUJITSU COMPUTERS SIEMENS

iSCSI

Internet Small Computer System Interface

Issue February 2005

Contents

1. Summary

iSCSI (Internet Small Computer System Interface) is an extremely interesting technology that could revolutionize some aspects of storage network. iSCSI connects servers with storage systems via TCP/IP networks. This relatively new protocol enhances the range of storage connectivity options. iSCSI doesn´t change the proven SCSI protocol semantics. That means each and every application which is certified for SCSI environments is suited for iSCSI environments as well.The basic idea behind iSCSI is that block-mode I/Os would no longer be transported via direct attached SCSI or specialized Fibre Channel networks (FC networks) but over TCP/IP networks. Block-mode I/Os for example act as the basis for backoffice or database applications.

TCP/IP and FC networks however exhibit different characteristics because they were designed originally with different objectives. If the FC network is mapped to the TCP/IP network, a different overall optimum inevitably results. In plain language this means either cutting back on applications or providing additional resources so that both separate optima can be fulfilled. In end effect, a TCP/IP-based storage network should be defined such that it can offer the FC properties. For example network administrators have to prevent with all means overload situations where already sent data packets can not be handled in time by the receiver and in consequence have to be requested again (so called packet drops). These events have extreme negative consequences on response times especially of backoffice and database applications. Preventing packet drops is not easy in complex, high demanding configurations. One needs technical know-how and considerable additional network resources. In some opinions, the only way to achieve predetermined behavior in TCP/IP networks is so-called over provisioning (designing all network components for peak loads). The best way however is to establish an application-specific dedicated “switched” TCP/IP network with comparable bandwidths as this would be available with FC networks.
However, the greatest challenge lies in permanently adapting a TCP/IP based storage network to changing requirements. Additional load can cause a TCP/IP-based storage network to slip unnoticed into problems (increasingly recurring packet drops). This is excluded in non-cascaded and non-meshed FC networks in all probabilities.

An overall assessment of the above mentioned facts leads to the following conclusion: iSCSI enables a cost-effective migration from DAS model (Direct Attached Storage) to Networked Storage model where e.g. performance requirements, management-effectiveness in large configurations etc. don’t play the dominant role. Furthermore existing SAN (Storage Area Network) environments can be leveraged with iSCSI - it then facilitates the consolidation of "stranded servers". With iSCSI technology, companies can consolidate servers that were not eligible before. In focus are Windows 2000 and Windows 2003 environments.
iSCSI is a cost-effective way to realize networked storage below FC configurations. This is a perfect fit where direct attached storage is used today but for instance higher utilization should be achieved.
iSCSI delivers also a solution to applications which are not supported on NAS systems (Network Attached Storage), e.g. Microsoft SQL Server and Microsoft Exchange Server. iSCSI solves this problem with block-level access.
Customers can now move their storage from direct-attached or internal storage to iSCSI networked storage. Managing the storage network can be done with existing know-how and experience as far as the TCP/IP network is concerned. In the enterprise area, companies can look outside the data center at tier two and tier three applications where cost or technical obstacles were barriers to consolidation.
The enterprise area is in contrast to this with its business critical computing applications (BCC applications). Among other things, this area demands highest throughput, predictable response time behavior, predetermined handling of extreme load situations and efficient management. Tier one application should remain on FC SAN for performance, management-effectiveness etc. reasons.

2. General comments

This White Paper discusses various aspects in relation to requirements for iSCSI solutions. The White Paper does not intend to convey basic technical information on iSCSI. Technical information is available in varying degrees of detail on the Internet, for example. The Storage Network Industry Association (SNIA) provides extensive material at www.snia.org. iSCSI information can of course also be acquired from other sources. Internet search engines offer a suitable means of establishing these sources.
The White Paper is consequently aimed primarily at system architects, system designers, system consultants and decision-makers in the area of IT or people who perform comparable duties.

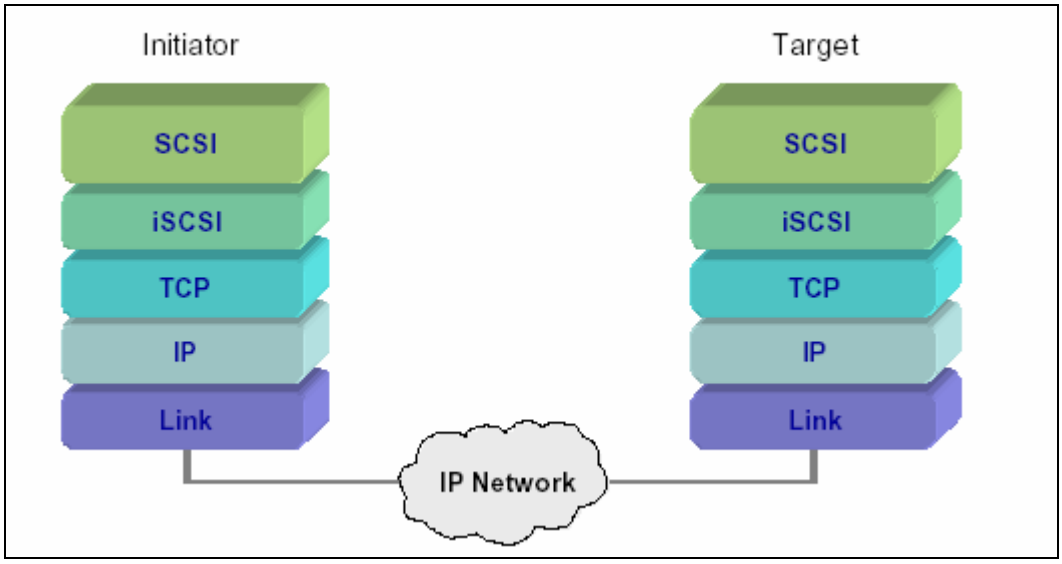
3. Vicinity of iSCSI

iSCSI technology cannot be discussed in isolation. Apart from block-mode access via iSCSI, the file-based access methods, as are used for connecting file servers or NAS systems, likewise use the TCP/IP network. In terms of typical file-based operation (a complete file is loaded onto the server at the beginning of the session, processed and written back to the file server at the end of a longer session), file servers are cornerstone of innumerable applications. It has emerged recently that file-based access methods are increasingly beginning to penetrate the domain of backoffice and database applications. Reasons for this may be the improvement in protocol functionality and the availability of attractively priced, high-performance networks, especially with GigaBit Ethernet (GbE).
We have consequently identified a technological overlap in certain areas between block-mode access methods with FC respectively iSCSI and file-based access methods with file servers. The transparent operating system integration above the device driver layer is a plus for iSCSI. The SCSI model that is used in Direct Attached Storage is preserved. Every application that can presently address Direct Attached Storage is implicitly also suited for iSCSI. Efficient management and mature data management functions (point in time copies etc.) support the extension of the file server application area in backoffice and database applications. The more deeply this application area penetrates the BCC or enterprise area, the more important however is the predictable response times and predetermined behavior in extreme laod situations. Essentially, the more response time critical the applications, the more strictly the above described quality requirements for TCP/IP networks must be fulfilled. The performance tuning in the remaining components is equally as important.
The competitive situation may possibly be resolved through the coexistance of FC respectively iSCSI and file-based access in a general Networked Storage Model or with a Unified Storage Solution. This solution offers iSCSI, file-based and FC interfaces in parallel.
In the short and medium term iSCSI and file-based access will undoubtedly be complementary to FC technology. FC networks, for example, have a number of so-called built-in functions in relation to throughout quality and implement a consistent end-to-end bandwidth of up to 4 GigaBit per second. Conceptually, a more reliable design is produced, especially for the enterprise area. Alongside this, protection of investment for existing configurations plays a key role. It makes little sense to replace an effective and efficient FC network completely with a relatively new technology. However, the result may be different if, for example, there is no FC configuration available as yet in the non-BCC area or if connections are required above and beyond data center boundaries, in which case the laying of a FC connection is considerably more unwieldy than inside a data center.
A strong trend in general is toward a uniform network based on TCP/IP for all types of communication. The convergence will not only affect communication between terminal and server and between server and storage system in this case. A similar development can also be seen in the area of telephony (keyword IP telephony). However, the realization of this vision will not follow in the short term.

4. What is iSCSI?

The well-known Small Computer Systems Interface (SCSI) allows servers, backup servers, Web servers, etc. to process block-mode I/O operations with various peripheral devices. Target devices can be disk storage systems, tape devices, optical devices, printers etc. The traditional SCSI

connection technology uses parallel type cables between the servers and the peripheral devices. This cable type offers restricted coverage and the maximum number of devices that can be connected is very limited. New technologies were therefore developed some years back that considerably ease these limitations. The prevailing technology in the storage area at present is based on SAN with the FC connection protocol. FC connections are typically optical connections and currently offer a bandwidth of up to 4 GigaBit per second (Gbps). In addition to the TCP/IP network[1], however, SANs or FC networks form a separated network which requires different know-how and has to be administered separately.
The logical consequence in terms of further development now lies in enabling the transport of the block-mode SCSI protocol via the generally applicable TCP/IP network rather than using a specialized network technology. The iSCSI protocol therefore defines how block-mode I/O can be transmitted via the TCP/IP network. The iSCSI protocol is based on the SCSI client/server architecture. A distinction is made between iSCSI initiators (generally the server) and iSCSI targets (in general the storage system). Initiator and target can be implemented either in native mode or via upstream routers. A native implementation means that iSCSI functions are implemented directly in the I/O controller area. A router is an additional component that performs protocol conversion outside the initiator or target device. Various approaches are followed in the actual implementation, which differ considerably in terms of the resource allocation. This point is discussed in greater detail further below. The following picture shows the iSCSI protocol stack.



5. iSCSI benefits

If one takes the IT infrastructure of medium sized or large enterprises as a whole, at least three different network types can be found in the enterprise area.
Communication between terminals and servers or communication between the servers is handled via the TCP/IP network. This also includes internal and external communication on the Internet / Intranet or Extranet. File/server communication likewise falls into this category, because from a technical perspective a file server is a specialized server and consequently is based on server to server communication. Storage networks are based on FC technology today. In addition to the two networks mentioned above, we should also look at the telephony network. For telephony network the industry has also the vision using TCP/IP networks instead of specialized networks. Other specialized networks may exist in companies e.g. for process control. Cluster interconnect is another example which makes use of specialized networks and can thus deliver extreme low latency. There are some cluster solutions based on InfiniBand. Specialized networks won't be

[1] The term TCP/IP network should be treated here as a generic term. It covers all layers below TCP. For the sake of simplicity therefore, the term TCP/IP network can also be used to cover the term Ethernet network, although this is not correct from a technical perspective.

looked at any more closely. In the following we will concentrate only on iSCSI, NAS and FC communication.

TCP/IP networks and FC networks at management level can still use the same components in the lower (physical) layers. However, they differ considerably in the higher protocol layers and the switching components (routers, switches [2] etc.).
The fact is, however, that the three network types (TCP/IP network, FC network and telephony network) co-exist in many companies. They comprise different components and demand different know-how and management methods. Management costs exceed the pure procurement costs by a multiple. This applies for all three network types. Having three different network types considerably intensifies this problem.
New developments are now showing how these costs can be reduced considerably. The solution lies in convergence, in other words in the merging of the three separate network types in one single network type. The trend towards IP telephony can be clearly recognized. Several parallel approaches are followed in the storage area in order to use the TCP/IP network for communication between the servers and storage components or between FC network components. This means that there is one uniform network type on the horizontal axis at least, which could offer the following benefits:

- The procurement costs for the individual components should decrease because TCP/IP is used as the common technology. TCP/IP components are produced in large numbers so that scale effects apply.
- Simplified/shorter design, planning, installation and troubleshooting phases based on a uniform network type.
- Management costs are reduced because only one network type has to be managed.
- Ready availability of know-how for consulting, installation, operation and diagnostics. External capacity can be procured relatively easily and hence cheaply on the market.
- TCP/IP as a routed network in principle does not recognize length restrictions. The number of addressable devices is so high in IPv6 that there is no practical limit. Although IPv4 has a much smaller address frame, it should be possible to solve every address problem using private address space.
- No separate concepts for implementing high-availability networks.
- It is to be expected that new functions and particularly higher bandwidths will be made available more quickly for the mass market, i.e. TCP/IP, than for specialized network types.

Unfortunately, every network variant has a different optimization strategy. TCP/IP networks were originally designed for short messages between terminals and servers, where a large number of terminals were assumed. An essential feature is the optimum utilization of all network components in terms of cost efficiency. It is accepted that individual transactions may be disadvantaged unduly here. This so-called best effort approach can be found frequently on the Internet, which is based primarily on TCP/IP mechanisms. The objective in designing FC networks from the outset was to control large volumes of data from relatively few applications with predictable latency in the storage network, even in extreme load situations. An FC network exhibits its strengths particularly in backoffice and database applications, where the overall response time depends essentially on the individual I/O times.
If one combines the different network types, a different overall optimum inevitably results. In plain language this means either cutting back on applications or providing additional resources so that both separate optima can be fulfilled. In end effect, a TCP/IP-based storage network should be defined such that it can offer the FC properties. For example network administrators have to prevent with all means overload situations where already sent data packets can not be handled timely by the receiver and in consequence have to be requested again (so called packet drops). These events have extreme negative consequences on response times especially of backoffice and database applications. Preventing packet drops is not easy in complex, high demanding configurations. One needs technical know-how and considerable additional network resources. In some opinions, the only way to achieve predetermined behavior in TCP/IP networks is so-called

[2] This paper does not distinguish between FC switches and FC directors. FC switch is used synonymously for FC director.

over provisioning. The best way however is to establish an application-specific dedicated "switched" TCP/IP network with comparable bandwidths and as this would be available with FC networks. A united network must be able to fulfill the originally separate load requirements fully. Careful planning is essential here.

6. iSCSI ratification

As in the TCP/IP area, smooth interoperation between components from different manufacturers is also assumed in the storage area. Standards are necessary to ensure this. A distinction must be made between de-facto standards, which develop independently of the standardization organizations and are broadly accepted and standards that are ratified by standardization organizations, so-called official standards. The aim of every official standard is for it also to be broadly accepted within the shortest possible time. The success of TCP/IP and FC is based on a variety of official standards, which have now become broadly accepted. Essential standardization work is therefore performed by the Internet Engineering Task Force (IETF www.ietf.org) and the American National Standards Institute (ANSI www.ansi.org) and their sub-organizations. There are also further standardization organizations, which will not however be listed here individually.
The iSCSI protocol is dealt with by the IP storage sub-group of the IETF (http://www.ietf.org/html.charters/ips-charter.html) and the final official standard has been ratified in 2003.

7. Basic requirements

The only basis in an heterogeneous environment (a iSCSI initiator from vendor A is to cooperate with an iSCSI target from vendor B) can be the finally ratified official standard. This alone is not sufficient for the enterprise area with BCC applications. As part of the qualification process, trouble-free cooperation must also be assured in extreme situations or in non-everyday configurations. Furthermore, suitable support must be provided in order to be able to diagnose and resolve errors that arise despite intensive tests and qualifications.
Apart from qualification and ratification, another essential factor is the loading of the server or also the storage system as a result of protocol handling by the TCP/IP stack. SCSI or FC technology can be viewed as the standard here. In the case of SCSI and FC, the server is only loaded minimally even in the event of an extremely high I/O volume. If the storage is connected directly via SCSI to the server, no complicated routines are needed. The majority of the logic in FC configurations for processing I/O in the storage network is handled in the so-called Host Bus Adapters (HBA). The load increase in this case is not essentially higher than in the SCSI model. However, iSCSI I/O must pass through the TCP/IP stack. Because the TCP/IP stack has to respond to a variety of processing variants due to its historic development and also has to be able to cope with the most varied error situations and data is copied among user space, kernel and driver, the processing of the TCP/IP stack is much more unwieldy than the SCSI stack. If TCP/IP processing is not transferred to a special adapter card (HBA or NIC, Network Interface Card), rather is implemented in the software on the server, a high load can result on the server. An application server could be highly utilized, for example, in handling a GbE data volume. There is then minimal capacity remaining for processing the application. Adapter cards with an offloaded TCP/IP stack (TCP offload engines or also TOEs) are options if not enough CPU power is available. The technical implementations should be studied precisely in the case of TOEs. It is important to investigate whether only core areas of the TCP/IP stack run in the offloaded hardware but error handling continues to be handled in the software on the server, or whether the entire TCP/IP stack is actually handled in the TOE.
The above comments are less critical if the I/O volume to the storage peripherals is low in the entry-level area and sufficient CPU capacity is available (difficulties in offloading the general TCP/IP stack are not discussed here.) However the CPU power of new machines (e.g. based on the Intel-architecture) allows to realize software-based iSCSI implementation in many cases.

8. Throughput

8.1 Data flow control

An important point in the discussion for the enterprise area is how a network handles extreme load situations. It may be necessary, for example, to manage situations where the source sends more data packets than the network can transport or the source can receive. This means that the protocol must inform the sender in some way that the transmission rate is to be reduced or the protocol must exclude the possibility of such situations from the outset. There are different data flow control algorithms that can be applied in such cases. FC networks and TCP/IP networks differ here considerably.
A "credit-based flow control" is found primarily in FC networks. In lay terms, the sender and recipient negotiate the data frames that are to be sent in one batch in advance. The network components and the source must provide suitable resources in the form of buffers in order to guarantee that the incoming data can also be received. A key factor in terms of the number of buffers required is the distance that has to be bridged. The greater the distance, the more buffers are needed to be able to accept all as yet unacknowledged data packets that are still on the connection. Essentially it applies that parallel load, even if it is high priority, may not use resources that have been reserved for some other purpose.
TCP/IP networks, on the other hand, generally use "window-based flow control". The basic principle here lies in the feedback, when a bottleneck occurs anywhere in the network. In this scenario, a component may not be able to receive new packets because the buffer could not be emptied in time. Packets then have to be rejected, resulting in complex recovery measures with renewed packet requests. Recovery measures on the one hand tie up resources and on the other hand increase the dwell time because of the renewed requests for individual packets. Whether this procedure can be regarded as disadvantageous again depends on the load profile. Assuming a data transfer can always be reproduced within a single packet (e.g. less than approx. 1,500 bytes), this procedure is less critical than if enormous data volumes, for example with data blocks of 64,000 bytes or even larger, have to be sent over a network, which uses "window-based flow control" and only realizes relatively later on that the buffer sizes are not sufficient on the network or recipient side. The recovery algorithms do of course differ and have a different effect on the overall runtimes.

Quality of Service (QoS) mechanisms with TCP/IP
The inherent problem described above with packet rejections in the window-based flow control do not have to occur in general. A variety of classification and prioritization mechanisms or general measures for ensuring a specific QoS were proposed and implemented in the TCP/IP area. Unfortunately, no generally applicable procedure, which above all was accepted equally by all vendors, emerged. The most important will therefore be listed briefly below.
Traffic Prioritization according to the 802.1q standard
Data traffic can be split into eight different priority classes based on the bits in a 3-bit field. If only simple algorithms are used, higher priority data traffic may completely overshadow lower priority data traffic. For this reason, a weighted round-robin procedure is also used in some implementations. This ensures that lower priority data traffic is allocated a minimum of resources.
802.1q VLAN Tagging
Virtual Local Area Networks (VLAN) allow the splitting of the data traffic from groups or devices. Members of a VLAN can communicate with one another, but do not see the rest of the network.
Differentiated Services
This allows classification with up to 64 classes. The standard provides the basis for extending from an individual priority assignment to a rule-based system. Decisions on forwarding can be made at every network node in accordance with a rule base.
Resource Reservation Protocol (RSVP)
RSVP uses a differentiated QoS strategy in cooperation with differentiated services. Because RSVP is not used universally in TCP/IP networks, it must be possible to receive the information via non-RSVP segments also. This is achieved on the basis of so-called tunneling via non-RSVP domains.

Multiprotocol Label Switching (MPLS)
Just like RSVP, MPLS establishes a data path through the network in order to speed up the data flow and to enable QoS. In contrast to RSVP, which establishes a reserved route between RSVP-enabled routers, MPLS uses a marking at packet level. MPLS can be used together with RSVP in order to speed up the data flow and to ensure QoS.

Traffic Shaping
Prioritizations can also be ensured using special appliances, which are inserted at central points between two connections. These appliances interpret the incoming data streams and forward them according to predefined rules.

Conclusion
The FC protocol ensures that the incidents of rejected data packets are kept to an absolute minimum. Extensive recovery mechanisms (renewed requests for FC frames) should not arise. I/O times can be predicted reliably. In addition, a reliable data throughput prediction can also be made based on fixed buffer agreements (dynamic or static) depending on the line lengths.
In contrast to the FC protocol, the TCP/IP protocol has a different optimization point. In order to ensure the reliability and predictability of an FC network, a number of mechanisms may have to be applied consistently. Consistent means that a QoS mechanism has to be supported by all components in the network. Deterioration can happen if only a single component does not provide the QoS mechanisms (see RSVP or routers / switches of different vendors). The reliability and predictability of I/O runtimes is a fundamental feature, however, of backoffice and database applications. In extreme cases, the possibility of a backoffice or database application revoking operability if an I/O exceeds a set time limit cannot be excluded. The risk is particularly high in routed networks, because paths may be defined dynamically. Considerable know-how seems to be necessary to emulate the functions integrated in the FC architecture in TCP/IP networks via the different QoS mechanisms. It may be necessary to ensure a considerable over provisioning of resources to counteract the phenomenon of packet dropps and consequently the considerably extended runtimes.

8.2 Bandwidths

The development steps in terms of bandwidths are not synchronous in the TCP/IP and FC area. In the case of Ethernet, a broad availability of 1 GbE (Gigabit Ethernet) can be assumed at present, which impacts the end-to-end view. 2 Gbps technology has been available end-to-end in the FC domain for some time now with the next iteration to 4 Gbps already available. Ethernet has already taken the next step in that 10 GbE has been ratified and is available. Due to costs and the need to ensure protection of investment, 10 GbE technology is concentrated primarily today in the backbone area. Standardization and development work has likewise begun in the FC domain for 10 Gbps FC. First solutions are available but are partly proprietary and have the same cost problems as 10 GbE.
If one compares GbE and 1 Gbps FC in detail, it must also be considered that both have different clock rates. GbE has a nominal bandwidth of 1.250 Gbps, while FC is 1.062 Gbps. This gives GbE a slight edge if the protocol overhead (see below) is ignored. In the case of investments, however, GbE and FC 2 respectively 4 Gbps have to be compared. This means therefore that in the end-to-end analysis, FC offers throughput advantages over TCP/IP networks.

8.3 Protocol overhead

The maximum size of an FC frame is some 2,100 bytes without synchronization and CRC bytes. In TCP/IP networks, the comparable size is typically some 1,500 bytes. If the FC protocol has to be mapped to the TCP/IP protocol in iSCSI routers in particular, it is not possible to exclude the possibility that two TCP/IP packets will have to be generated from one FC frame. iSCSI-HBA and iSCSI router vendors are trying, however, to minimize these effects as far as possible with larger related data volumes using so-called intelligent splitting. A more favorable situation exists with the use of jumbo frames in the TCP/IP network. Jumbo frames are usually 9,000 bytes in size but there may other implementations e.g. with 16,000 bytes. Unfortunately, jumbo frames are not available in all TCP/IP networks or are only used occasionally.

A further point of discussion is how much additional TCP/IP control information is needed with the iSCSI protocol compared with an FC network. The estimated protocol overhead for TCP/IP is just over 4.3 percent, while this figure is just higher than 2.7 percent for FC. Because some parts are optional, the concrete value must be verified in the actual configuration.

8.4 Trunking / link aggregation

By connecting together individual Inter Switch Links (ISL) in a trunk group, a single logical connection can be implemented in certain FC switches with a multiple of 2 or 4 Gbps. Even if one considers a single data stream, the first FC frame can be transferred to the first ISL connection and then the second frame to the second ISL connection and so on for dynamic load balancing. This does not exclude the possibility that FC frames overtake each other. FC switches that support ISL trunking ensure that the FC frames are forwarded to the receiving switch in the sequence they have been received from the source at the sending switch. The function is performed in real-time at the switch protocol level, which means that no major delays result.
This function is referred to as link aggregation in the TCP-IP domain. The higher protocol layers are responsible in this case for ensuring the correct packet sequence. Delays can occur as a result. What must be analyzed however is the probability of out-of-order delivery.

9. Availability

Paths are set up from the initiator to the target via redundant components and connections in order to increase availability. If an interrupt occurs on a path, I/O can be processed via the other path. Multipath software is required on the servers in order to detect problem situations and enable a switch to another path. This software should essentially function in the same way in the iSCSI environment as in FC configurations. However, this software must be tested and qualified as intensively for the iSCSI environment as in the SCSI or FC domain.

9.1 Impact of network configuration

If one considers the core functionality of FC configurations, then without exception they involve a switched network. FC frames are forwarded at deep protocol levels. The latency in the switches is without exception below three microseconds.
Both switched and routed networks can be found in the TCP/IP domain. Switched TCP/IP networks have comparable or to an extent even better properties than FC networks. Routed TCP/IP networks are more critical however in terms of the I/O runtimes. The latency per router is on the scale of a multiple of 10 to 100 microseconds at a minimum. If I/O output has to pass through several routers, this can have a noticeable impact on the response times of an application.
Routers can also be used in the FC network. However, these are used for special functions rather than in the core area. Examples include converting the FC protocol to the SCSI protocol to enable older devices to be operated in FC configurations also, or if greater distances have to be overcome with individual connections and there is no direct FC connection available.
A further issue can arise with signal runtimes because of larger distances. Light has a propagation speed of some 200,000 km per second in fiber optics. This implies approx. 1 millisecond for 100 km, if one also takes the necessary return transportation of the acknowledgement into account. If one assumes an average I/O time of between 0.5 to 6 milliseconds per I/O in normal configurations, this parameter can be ignored in the range of a few kilometers. Because FC configurations are based on independent cabling, the definition of the maximum cable lengths will not pose any problems. Essentially, this should not be a problem in TCP/IP networks. However, it can be problematic if it is attempted to also use historically evolved complex TCP/IP networks where the data flow may not necessarily be conducted over the shortest distance.
It was suggested in the White Paper that the TCP/IP network that controls the iSCSI data traffic should be implemented as a dedicated switched network with an any to any, non blocking bandwidth of at least one Gbps. However, this does not exclude the possibility that because of inadequate design only 100 Mbps or less will be available between two routers. This must be avoided at all costs from the point of view of performance because a bandwidth below 1 Gbps can have negative impacts on the response times of backoffice and database applications.

10. Storage network services

Apart from the pure data switching function, storage networks offer extended functions, such as data mirroring and data replication. The availability and maturity of these functions must be verified when using iSCSI.

10.1 iSCSI boot services

Boot information as well as other data is no longer stored on a local disk in the case of high-availability configurations, rather it is stored centrally on a storage system for a number of servers. This allows defective hardware to be replaced without regenerating the operating system. The new server loads the operating system from the same location as the old server. Other scenarios are configurations where a lot of servers do the same work and one wants to centralize the administration. Examples are blade servers or web servers respectively ERP servers with SAP. For UNIX systems this is relatively easy achievable. One needs an additional server that allocates the IP addresses and tells the clients where the data lies (kernel root drive). The Fujitsu Siemens concepts FlexFrame makes uses of it. For Windows systems one has to establish a so-called PXE Environment (pre execution environment). In principle it works like UNIX but has some restrictions. With the right adapter card it is also possible to boot in SAN environment. This will work with iSCSI too. As the boot process is critical in general (mostly time critical) support for failure diagnosis has to be assured and is a challenge which has to be solved.

10.2 iSNS (internet Storage Name Server)

iSNS extends iSCSI with directory functionality. The directory stores information about iSCSI instances. Corresponding standard services are defined for management and information delivery. This provides comparable functionality to Fibre Channel networks. For instance iSCSI targets can perform a self registering so that iSCSI initiator can find their storage areas. Thanks to iSNS iSCSI initiators and iSCSI targets can check if their counterparts are still active. The session can be terminated if the partner is no longer active. This requires self registering of iSCSI initiators as well otherwise the check process would not work. iSNS is also able to send event-driven messages. E.g. configuration changes in the directory trigger such events. Clients, which have registered for these messages, will see new storage areas immediately. iSNS is the basis for iSCSI gateways by matching iSCSI names into Fibre Channel addresses. iSNS is a central information base for monitoring and management tools. Provided all participants have registered the monitoring tools needs not to run complex discovery processes. All base information is available via iSNS.

iSNS has not proliferated too much today. Besides iSNS there are also other means for locating storage areas. The most important ones are SLP and DNS.

10.3 Function migration

In recent times, mirror and replication functions among others have been moved out to the storage network. The functions run on appliances in the storage network, which behave like storage systems (target functions) with respect to the servers and like servers (initiator functions) with respect to the actual storage. The most important appliance functions include:

- Memory pooling (storage virtualization)
- Local and remote mirroring
- Replication
- Snapshots
- 3rd party copy

These functions do not have to run on dedicated appliances. More recent information shows a further offloading of the above functions to FC switches. There is no evidence at present to suggest that such functions are being provided in TCP/IP networks.

11. Manageability

Manageability refers to the totality of all functions in a centralized end-to-end view:

- Automatic discovery of all subcomponents and connections in a storage network

- Active monitoring of all subcomponents and connections (status, performance etc.)
- Active management of all subcomponents and connections
- Central resource allocation
- Recording and storage of historical data
- Reporting, trend analyses

Methods and resources are currently available in the FC domain for this purpose. Communication with the storage components is handled via different interfaces. Assuming generally accepted interfaces like SNMP are used, no problems should arise. Some functions use hardware-related APIs, such as the HBA API. Adaptations may have to be made in the management tools in order to be able to offer similarly integrated functions. Among other things, these interfaces form the basis for automatic discovery of all subcomponents (vendor, firmware level, allocation to a server, etc.) Based on this, fully automatic storage allocation can be performed using a rule base. The tool has to be able to process at least the following individual steps consistently:

- Configuration of LUNs (Logical Unit Number) in the storage system
- Allocation of the LUN to a port
- Updating of the LUN masking so that only certain servers can access this LUN.
- Definition of zoning (access control and throughput optimization)
- Definition of the LUN in the Volume Manager on the server
- Extension of the file system by the volume and, if appropriate
- Extension of the database tables

A further example is the illustration of a data path with all connections and components starting with the database and ending with the physical disk in the storage system. In our opinion, centralized end-to-end management is absolutely essential for the productive use of a technology in the enterprise domain.

Zoning is also an issue in terms of individual storage allocation functions. Zoning is a central FC component that ensures security and availability of an IT configuration. These functions are mapped in the iSCSI domain either via iSNS (Internet Storage Name Services) or via TCP/IP network functions. VLAN is a suitable resource for this purpose. Because a variety of zones may be essential in the enterprise area, it must be clarified whether the maximum possible number of VLANs is adequate, considering the VLAN for the terminal area must also be covered. According to the official standards, VLANs support at most 4096 definitions. A general response to the question regarding the maximum number of VLANs is therefore difficult if one not only considers the TCP/IP standard but also takes into account proprietary implementations.
The availability of SNMP standard interfaces in the form of MIBs and CIM definitions is a further prerequisite for the enterprise area. This ensures efficient iSCSI integration in standard management applications. Some further standardization work seems to be necessary here also.

12. Security

FC networks are assumed to mean physically segregated networks. The connection cables and switching components are generally laid in specially secured premises to which only authorized persons have access. For this reason, the necessary level of security is assured to the outside world as far as the physical layer is concerned. If an independent, physically separate TCP/IP network were likewise set up for iSCSI configurations and operated with the same care then it reaches the same level.
However, if iSCSI traffic is handled via components that are used together with general communication, or if the components are operated in premises that are not particularly secure, then additional security measures will be required.
The IPsec security protocol, which handles authentication and integrity checking, is therefore provided in the iSCSI standard. The basic mechanisms for authentication and integrity checking are strong encryption methods that tie up considerable resources. The use of IPsec is therefore optional, in other words the user can decide whether or not to use IPsec. The product vendor must offer the function, however, if protocol conformity is to be ensured. Unfortunately, the first iSCSI products do not seem to offer this protocol conformity. Even if IPsec is offered in a iSCSI-HBA, for hardware example, it should be verified whether the encryption functions are being processed on

the HBA (either in the firmware or directly in the hardware) or on the server. In the latter case, the load on the server can be considerable. The problem may be exacerbated even further with the transition to 10 GbE.
In cases where IPsec is not implemented in the iSCSI components, only an external solution remains feasible. Because of the widescale use and standardization of TCP/IP networks, a wide range of security appliances are offered. The most important of these are firewalls and VPN solutions (Virtual Private Networks). Both solutions are often implemented in one combination. From the point of view of storage applications, additional appliances mean increased I/O times.

13. Miscellaneous

13.1 IP V6

The TCP/IP network is continuing to evolve. Some time ago, Version 6 of the IP was standardized (IPv6). This version considerably extends the maximum number of devices that can be addressed among other features. Although the transition from IPv4 to IPv6 seems to be slower, iSCSI products should already support IPv6.
An interesting function would be the compression of user data in real-time. However, the products currently available do not seem to offer such a function.

13.2 FCIP and iFCP

The FCIP protocol (FC over IP) implements a tunneling procedure, whereby FC frames are packed into TCP/IP packets so that they can be transported over a TCP/IP network. This technology allows individual storage network islands to be connected relatively easily over greater distances. In concrete terms this means that a storage router (converter) is connected to an FC switch and this converter packs the FC frames into TCP/IP. These TCP/IP packets are then transported over the TCP/IP network to the recipient. The recipient in turn is a storage router, which is connected to a second FC switch. The storage router unpacks the FC frame from the TCP/IP packet and passes it to the FC switch. The FCIP protocol does not differentiate between FC frames with data and FC frames with control statements for coordinating a fabric. A new FCIP connection set up in this way between two separate fabrics (several switches in one address space) causes the two originally separate fabrics to merge. Each original fabric had a master switch, which was responsible for uniform address assignment. However, only one master is permitted in the newly created fabric. A new master therefore has to be defined for the entire fabric. The definition of the master and the subsequent uniform address assignment is handled automatically in FC networks. This causes an interrupt of a few seconds. The opposite process is run through if the FCIP connection is interrupted for several seconds. Two separate fabrics again result, each of which has to define a new master for itself. Poor quality FCIP connections consequently lead to frequent interruptions not only in an individual connection but also in the fabric as a whole.
The primary goal of iFCP (Internet FC protocol) on the other hand is to connect FC devices via the IP network, whereby IP switching and router components replace the FC fabric services. The iFCP protocol is a gateway-to-gateway protocol. In contrast to the tunneling approach with FCIP, separate fabrics are not merged in a uniform fabric. Error situations, which for example lead to the reconstruction of the fabric, are restricted locally. Two storage routers are also required here as a minimum. One challenge here is to ensure consistent handling of the metadata (access rights in the form of zoning etc.). A standard is emerging with iSNS for solving the problem.
The same comments apply for FCIP and iFCP in terms of bandwidth, latency, and QoS, as were made in relation to the iSCSI connection.

13.3 File-based access methods

In contrast to backoffice and database applications, traditional file-based applications support a different form of access. A large volume of data is typically loaded from the file server onto the server in one action at the beginning of a session and only rewritten to the file server after more extensive processing. The total one-off access time accepted at the beginning and end of the session is in the sphere of seconds. Any overdue delay in individual partial access is relatively unimportant. NAS systems, which use the TCP/IP protocol for communication with the server, are

used productively in innumerable configurations. The TCP/IP network is a reliable basis, even though basic requirements are not fulfilled in some configurations (e.g. insufficient separation of the network between the server and the file server on one side and between the server and terminals on the other side).
File-based access is not limited to these traditional application areas, however. It has emerged recently that file-based access methods are increasingly penetrating the area of backoffice and database applications. This may be because of improvements in protocol functionality and the availability of attractively priced, high-performance networks, especially with GbE. In addition to general improvements in the infrastructure, special hardware components for certain protocols can be used in combination with databases for specific file servers, which considerably reduce the TCP/IP protocol overhead.

The most important standard protocols for processing file-based access include NFS, CIFS, http and FTP. These protocols are being enhanced continuously and consequently open up additional application areas. For example, various functions were introduced with NFS V3 that offer excellent potential for improving access times both for traditional operation and for typical backoffice and database applications with response time-critical requirements. NFS V3, for example, enables more efficient transmission of metadata in that the data no longer has to be requested separately rather can be incorporated in the basic operation. NFS V3 also allows asynchronous write processes. A client typically issues a series of asynchronous write processes followed by a commit call. The commit call causes the file server to write all orders to disk that have not already been written to disk, before the commit request is confirmed. This can considerably enhance the performance. These examples can only stimulate progress and should be interpreted in the same way in the other protocols.
The optional parameters are a further topic for discussion. NFS, for example, can be operated on the basis of TCP or UDP. The respective mode of operation used depends on the quality of the network. A trade-off must be found between the location and type of error handling and the actual errors that occur. Error handling is performed completely in the TCP protocol stack with TCP, whereas with UDP the relevant application has to take over parts of the error handling. If a packet was not transferred correctly within a sequence in TCP the entire sequence does not have to be transmitted again, rather TCP simply requests the relevant packet again. If a packet could not be transmitted properly in a sequence in UDP, the entire sequence has to be transmitted again. UDP is generally only recommended if the network operates very reliably.
Similar considerations are also necessary for mount requests. As a minimum, NFS decides between a hard and a soft mount. A possible subvariant that can also be defined is whether the processes can be interrupted. These parameters define the reaction precisely if a file server is not available and the file system is to be connected to the client.

The advantages and disadvantages of the different protocols could also be discussed in general. For example, is NFS a state-less but CIFS a stateful protocol? State-less essentially offers advantages in error handling (e.g. in the case of a fail-over) but can negatively impact performance. There is relatively little choice here, however, because the protocol to be used is generally determined by the choice of platform (UNIX with NFS, Windows with CIFS).
In summary, file servers can generally be connected relatively easily if one remains within the default parameter range. Default parameters also allow satisfactory results for most applications. However, operation of response time critical, block-mode backoffice and database applications requires tailored optimization measures both on the server and on the file server.

14. Concluding remarks

FC configurations have a number of integrated functions in order to guarantee optimum throughput, predetermined behavior in extreme load situations and maximum security. These functions can be mapped essentially to the TCP/IP network, although this means additional resources and management actions. A fair comparison between FC configurations and TCP/IP configurations must take all costs into account.

The additional bandwidth must be provided on the one hand in the TCP/IP-network area. The basic costs would be defined on this basis. On the other hand, this bandwidth will have to be designed in accordance with the QoS. A suitable over provisioning will presumably be required here in order to avoid critical packet drops. Furthermore, prioritizations and VLANs will also be required. This will increase the costs in comparison with standard TCP/IP implementations. Critical requirements in the enterprise area will generally require installation of a dedicated “switched” network, as is already recommended today for file servers.

At the network components there are savings possible because the TCP/IP components are produced in large numbers so that scale effects apply. Higher costs may have to be expected in the management area because considerable expenditure is needed to ensure predetermined behavior with regard to extreme load situations. Savings will arise in training both in the area of system design and at an operative level. Further positive effects lie in the potential integration into an existing NOC (Network Operation Center). In end effect, a fair and detailed cost comparison will show that many savings propounded by iSCSI advocates are not being achieved.
In the FC area 2 Gbps is the standard today with 4 Gpbs announced and available and a further move toward 10 Gbps is mapped out. A real challenge for FC technology could be the availability of cost-efficient 10 Gbps iSCSI-HBAs with TOEs and cost-efficient 10 Gbps TCP/IP networks.

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/