IT from A-Z

We have updated the most important key words from our comprehensive IT compendium especially for this magalogue.

Computer Architectures

Operating Systems

Clusters

Storage Networks

Magnetic Tape Storage

The complete reference material can be found in the Internet. The table of contents and index help you find the desired pages quickly and easily.

1. Computer Architectures

The following sections contain a brief overview of 64-bit computer systems, as well as the most important current CPUs.

Currently, the most important subject for processors is the transition from 32 to 64-bit architectures. Most RISC architectures have already taken this step, for example SPARC. The Intel® IA64 architecture went to 64-bit with the Itanium in the summer of 2001. The follow-up model Itanium 2 has been available since the end of 2002. AMD has been supplying the Athlon 64 (desktop version) and the Opteron (server version) since the 2nd quarter of 2003. Intel® licensed AMD's 64-bit technology and integrated it in new processors and chip sets under the name Intel® EM64T (Extended Memory 64 Technology). The first products have been available since the third quarter of 2004. One of the outstanding features for the AMD 64-bit architecture and Intel® EM64T is complete 32-bit compatibility. 32-bit operating systems and applications thus run with very good performance. An important factor here is not only the availability of the hardware, but also the right operating systems. Microsoft's Windows Server 2003 is available in the Enterprise version with support for the IA64 architecture. A version for AMD's 64-bit architecture and Intel® EM64T is in development. There will also be a corresponding version for Windows XP. Kernels with corresponding 64-bit support are already available for both architectures in Linux. However, only when there are applications available with 64-bit architecture, can the user take full advantage of the performance of the 64-bit support.

1.1 Intel® Processors

1.1.1 Intel® Server Products with 64-Bit

Intel® has two complementary 64-bit architectures that offer an excellent value for all areas of business application: The new 64-bit platforms based on Intel® Xeon™ processors with Intel® EM64T offer the largest selection of applications for mainstream server and workstation solutions. They make an outstanding performance available for 32-bit applications and more reserves available for 64-bit applications, and can be used flexibly and inexpensively.

The platforms based on Intel® Itanium® 2 processors were developed specifically for extremely demanding and critical business 64-bit capacities. They offer the best performance and scalability, as well as the best RAS (Reliability, Availability, Manageability), but at much lower costs than proprietary RISC and mainframe solutions.

1.1.2 Intel® Itanium® 2 Processors

The Itanium® architecture has been expanded by less expensive and energy-saving systems now that the existing Intel® Itanium® 2 processors have been joined by the Itanium® 2 processor 1.60 GHz with 9 MB L3-cache and the low voltage version of the Itanium® 2 processor – optimized for dual processor systems.

1.1.3 Intel® Xeon™ Processor Family

The Intel® Xeon™ processor family continues to offer the processing capacity and variability necessary for front-end and small-business servers, as well as high performance computing systems in cost-sensitive environments. The Intel® Xeon™ processor with 800 MHz system bus, designed for dual processor server and workstation platforms, is now available with clock speeds between 2.80 GHz and 3.60 GHz. The processor is manufactured with the newest Intel® 90-nanometre process technology. It offers features such as Hyper Threading technology, Demand Based Switching (DBS) with expanded Intel® SpeedStep® technology, Intel® Extended Memory 64 technology and Streaming SIMD Extensions 3 (SSE3).

1.1.4 Intel® Pentium® 4 Processor Extreme Edition

The Intel® Pentium® 4 processor with HT technology makes a processor appear to be two processors to software programmes. The result is that programmes can be executed more efficiently. Performance is improved in today's multi-tasking environments, and the system's response speed is increased, since the threads, or programme instructions, can be executed simultaneously by the processor.

1.1.5 Intel® Pentium® 4 Processors

Intel® Pentium® 4 Processor with advanced 800 MHz system bus, suitable for Hyper Threading technology, is available with clock speeds of up to 3.80 GHz, 2 MB L2 cache (Intel® Pentium® 4 670 processor), equipped with 90-nanometre process technology. It offers better transfer rates and response times, which are necessary to process demanding applications smoothly, such as 3-D visualization or operating systems from the next generation.

1.1.6 Intel® Pentium® M Processors

The Intel® Pentium® M processor is one of the most important components of the Intel® Centrino™ mobile technology. It is based on an architecture developed especially for mobile computer usage.
Advantages: Excellent mobile performance and energy-saving features for thinner, lighter notebooks.

1.1.7 Intel® Celeron® D Processors

The Celeron® D processor is the successor to the Celeron® processor line. Celeron® D is based on the Prescott Core with a maximum clock speed of 533 MHz and 256 MB L2 cache (as compared to 400 MHz and 128 MB L2 cache in Celeron®).

The Intel® Celeron® D processor offers a balance between proven technology and a reasonable price for desktop and notebook computers that are expected to fulfil basic tasks. The Intel® Celeron® D processor is available with a clock speed of up to 3.20 GHz. It is designed to support elementary computing demands, such as exchanging emails with friends and family or managing finances.

1.1.8 Intel® Celeron® M Processors

The Intel® Celeron® M processor is based on the micro-architecture developed by Intel® especially for mobile computing. It offers users a balance between processor technology designed for mobility, good mobile performance and exceptionally reasonable prices in notebooks that are more elegant and lighter.

The Intel® Celeron® M processor line represents a new generation of Intel® technology that offers the most reasonably-priced mobile PC for the market segment.

1.2. AMD Processors

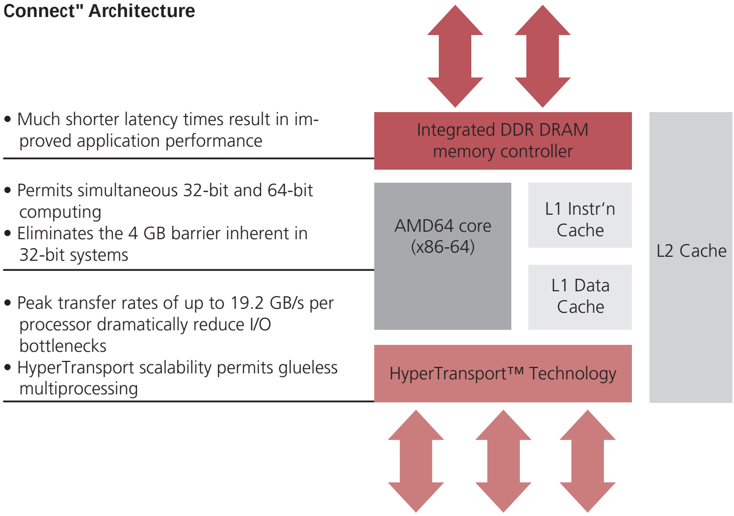
1.2.1 AMD Opteron™ Processors

The AMD Opteron™ processor, enabling simultaneous 32- and 64-bit computing, represents the landmark introduction of AMD64 technology with Direct Connect Architecture. The AMD Opteron™ processor is designed to run existing 32-bit applications with outstanding performance and offers customers a seamless migration path to 64-bit computing. These processors provide a quantum leap in compatibility, performance, investment security and a reduced total cost of ownership (TCO).

The AMD Opteron™ processor is offered in three series: the 100 series (1-way), the 200 series (up to 2-way) and the 800 series (up to 8-way).

The AMD Opteron™ processor integrates the following key system elements:

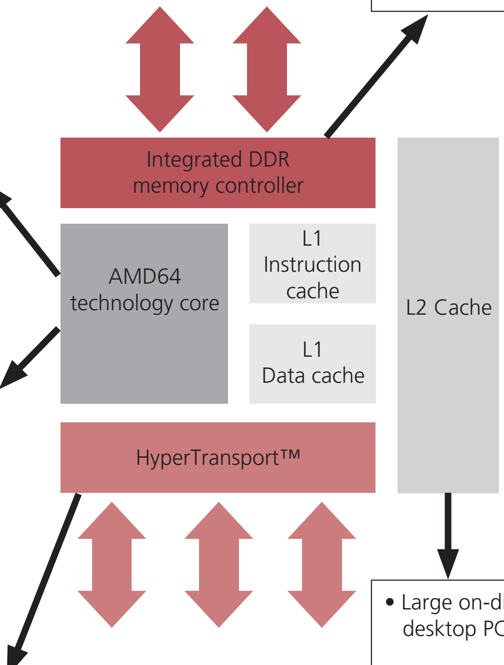
The AMD Opteron™ processor with "Direct Connect" Architecture



- Much shorter latency times result in improved application performance
- Permits simultaneous 32-bit and 64-bit computing
- Eliminates the 4 GB barrier inherent in 32-bit systems
- Peak transfer rates of up to 19.2 GB/s per processor dramatically reduce I/O bottlenecks
- HyperTransport scalability permits glueless multiprocessing

Integrated DDR DRAM memory controller
AMD64 core (x86-64)
L1 Instr'n Cache
L1 Data Cache
L2 Cache
HyperTransport™ Technology

AMD AthlonTM 64 architecture



- Leading-edge software performance
- Enables high-performance 32-bit and 64-bit computing
- Double the number of internal registers for improved performance
- An enormous memory addressable to well over 4 GB opens up a host of possibilities that are impossible with current 32-bit technology

- High performance and high bandwidth
- Drastically reduced DRAM latency
- Improves the performance of many applications, in particular those with large memory requirements
- Supports PC3200, PC2700, PC2100 and PC1600 DDR SDRAM memory
- Unbuffered DIMMs
- 64-bit or 128-bit interface
- Up to 6.4 GB/s memory bandwidth

- Windows XP SP2 enhanced virus protection reduces the risk of certain viruses and worms, for example, MSBlaster and Slammer, thus reducing downtime and the associated costs, as well as protecting user privacy against the threats posed by worms and spyware.

- HyperTransport technology increases system I/O bandwidth
- Up to 8 GB/s system bandwidth
- System bus speeds of up to 2000 MHz

- Large on-die cache memory system for desktop PCs
- 64 KB LI execution cache
- 64 KB LI data cache
- 512 or 1024 KB L2 cache
- 640 or 1152 KB total usable cache
- Provides significant performance boosts to many applications, especially memory hoggers such as graphics programs

1.2.2 AMD Athlon™ 64 Processors

The AMD64 technology provides unprecedented application performance and opens up new computing experiences and possibilities. Specific features of the AMD64 technology:

- Highest performance level for many existing 32-bit applications without changes.
- Simultaneous 32-bit and 64-bit high performance computing.
- Doubles the number of internal registers to increase performance.
- Expands the storage address space beyond 4 GB and makes realistic visual images for applications that use graphics intensively, such as 3-D games, and real time results for applications with extreme memory requirements, such as Digital Media.

>> *See schematic figure p.102:*
AMD Athlon™ 64 architecture

The most important advantages of the AMD64 architecture are:

HyperTransport™ technology for I/O high-speed communication

- A 16-bit link up to 2000 MHz
- Up to 8 GB/s HyperTransport™ I/O bandwidth
- Up to 14.4 GB/s processor-to-system total bandwidth

The HyperTransport Technology™ contributes to increased system performance, since the previous I/O bottlenecks have been eliminated by increasing I/O bandwidth and reducing the I/O latency times. This means a better total performance, faster application loading and outstanding multimedia capabilities.

Integrated DDR memory controller

The integrated Double Data Rate (DDR) memory controller effectively reduces one of the most serious and common system bottlenecks of previous platform designs: memory latency. The integrated DDR memory controller of the AMD Athlon™ 64 processors offers the following advantages:

- Performance increase through the direct connection of the processor with the RAM: Memory latency is dramatically reduced. The performance of many applications is then significantly improved, especially applications with intensive memory usage, such as digital media and 3-D games.
- Supports DDR memory technology for high performance systems with low total cost of ownership and low development costs.
- Uses ECC security for greater system stability and helps assure that systems work reliably.

Features and advantages of the AMD64 architecture

Feature	Advantage
Simultaneous 32- and 64-bit computing capability	Users can operate 32-bit and/or 64-bit applications and operating systems as they desire - without compromising performance
Direct Connect architecture overcomes the real challenges and reduces the bottlenecks of system architectures	Increases the memory latency performance, offers more balanced data throughput and I/O and makes an expanded "linear symmetrical multiprocessing" possible
Support of up to three coherent HyperTransport™ links, whereby top bandwidths of up to 19.2 GB/s per processor are possible	Makes considerable I/O bandwidths available for current and future application demands
256 Terabyte storage address space	Creates significant performance advantages for applications for which large (or numerous) data sets must be held in the memory
Scalable from 1 to 8 processors over the entire data or computer centre with identical hardware and software infrastructure	Makes maximum flexibility of the IT infrastructure possible and contributes fundamentally to the success
Integrated memory controller reduces latency times for memory access in SMP server systems	Allows fast data processing with better performance and increased productivity
Low power processors in HE (55 watts) and EE (30 watts) offer uncompromising performance	Increased computing density, lower operating costs for data centres with a limited energy consumption budget

1.2.3 AMD Sempron™ Processors

The AMD Sempron™ processor was developed to fulfil the daily needs of home and business PC users. These fully equipped processors offer customers concerned with the price of desktop PCs the best performance.

2. Operating Systems

The following sections provide a brief overview of the operating systems from Microsoft and Linux.

2.1 Windows

2.1.1 Windows 2000

Microsoft Windows 2000, referred to up until now as Windows NT 5.0, has been expanded by the addition of several new features and functions as compared with Windows NT. These concern the areas of administration, scalability and expandability, as well as storage and hardware management. Microsoft offers Windows 2000 in four versions:

Windows 2000 Professional corresponds to Windows NT Workstation and supports up to 4 GB main memory and two processors.
Windows 2000 Server is the successor of Windows NT Server and offers hardware support for max. 4 GB main memory and four processors. The Windows Terminal services, which replace the Windows NT 4.0 Terminal Server edition, are already contained in this server version.

The Windows NT Enterprise Edition will continue as **Windows 2000 Advanced Server**. Up to 8 GB main memory and eight processors are supported. In addition to the Windows 2000 Server functions, the IP load balancing (with up to 32 servers) and failover clustering for two servers are also available.
The Windows 2000 Datacenter Server represents the top end. It supports up to 32 processors and 64 GB of main memory, and offers the following additional functions beyond those of the Windows 2000 Advanced Server: failover clustering for 4 servers and process control for work load management. Another important feature is the support of virtual servers. Several instances of the operating system can be run on multi-processor servers, for example 2 virtual Servers with 4 processors each can be set up on an 8 processor server.

Installation of Windows 2000

Windows 2000 is installed in the computer without an operating system using a bootable CD. Plug & Play is a new feature of Microsoft Windows 2000, which simplifies the installation process. Another improvement made to Windows 2000 is that it must be rebooted very seldom, in comparison to Windows NT. The USB support has also been implemented in Windows 2000. Unlike the Windows NT server, when installing Windows 2000 it is not necessary to determine if the Windows 2000 server should be used as the domain controller, or not. With the help of the Configure Your Server wizard, the service for the Active Directory (directory service especially for user administration) can be installed at a later stage.

Repair mechanisms

Windows 2000 is equipped with an improved, protected boot mode. An additional, improved repair mechanism has been implemented in the command line.

Administration

Microsoft Windows 2000 implements Active Directory as a central platform that simplifies access to the management of network and system resources. Unlike in the User Manager for Windows NT, users can be organised, configured and managed hierarchically in containers in the Active Directory. With Windows 2000, the user administration is not just more structured, there is no longer a limit of approx. 20-40,000 users per domain as there is under NT. Further features include centralised configuration management and the configurable and expandable Microsoft Management Console (MMC).
IntelliMirror technology allows Windows 2000 workplaces to be centrally configured. With the help of the Active Directory, the configuration settings for users or groups can be centrally filed. This means that the user will have exactly the same configurations at all Windows -2000 workplaces and the user's software will be automatically installed at the respective workplaces. The configurations can also be set so that the user can not change them.

Scalability and extendibility

Windows 2000 supports up to 64 GB of physical memory. With the Microsoft Cluster Server, two or more servers can operate interlocked. Thus, the devices can monitor each other so as to maintain operation without interruption if one of the servers fails. During normal operation, the servers can share the workload, thereby increasing productivity.

Storage management

NTFS now also implements quotas to define the maximum amount of disk space available to the users. The NTFS expansion EFS (Encryption File System) allows the encryption of sensitive data on file level or directory level.
With the distributed DFS file system, the distributed structures of directories and files on Windows 2000/NT, NetWare and Unix servers can be summarized and presented in an organized manner. In this way, users can find files in the network much more easily.

Hardware management

Plug & Play allows PC cards to be used with portable computers without any difficulty. In addition, the expansion of the Windows Driver Model (WDM) is intended to enable identical driver software to be used in Windows 98 and Windows 2000.

Security functions: In order to increase the operational security, Windows 2000 prevents deleting critical operating system files. In addition, it only allows the installation of certified drivers.

Network security: The Security Service Provider Interface (SSPI) is already in use in Microsoft Windows NT 4.0, for example in the NT LAN Manager and the Secure Sockets Layer (SSL). In Windows 2000, the SSL is expanded and Kerberos authentication is introduced in accordance with Kerberos 5. Furthermore, Windows 2000 supports Smart Cards, which increases the security in user log-ons and digital e-mail signatures.

2.1.2 Windows XP

Windows XP impresses with its significantly improved, clear design, permitting intuitive operation even for users with relatively little practice. In addition to the new design, a multitude of additional capabilities have been developed.

The redesigned Start menu groups the most frequently used applications for easy access. The five most-used programmes are displayed first, and the default e-mail programme and Web browser are always available. The new taskbar grouping keeps the taskbar clean and organized. Open files are now grouped according to application type.

In addition to the design, Windows XP has a series of further novel features enhancing user friendliness. Windows wakes from hibernation far faster, i.e. the batteries of the laptop no longer have to run for nothing, because all the applications are immediately ready for use again when called up. It is also possible for several persons to use one PC together or side by side. A separate account is created for each user. The users can be changed quickly and simply while the other person's applications keep running. This means, for example, that downloading a file from the Internet does not have to be discontinued when there is a change in user, because all the open applications continue to run in the background.

Hardware support was greatly improved. Devices such as digital cameras, scanners or DV cameras are automatically recognised and all the recording, processing and playback functions Windows XP provides can be universally used.

The performance has also been significantly enhanced. The first improvement under Windows XP becomes clear the moment the computer is started up. After several starts with the same software and hardware, Windows XP arranges the start files on the hard disk for rapid access. Together with the so-called Prefetching feature and an optimised network login, the system starts up to 34% faster than with earlier Windows versions. The same function also means programmes start faster.

The enhanced multitasking performance is supported by measures including utilizing downtime for system activities, adapting the user interface to the computer's potential and efficient RAM management.

The stand-by and hibernate modes help for laptops. In stand-by, the system powers down the monitor, hard disk and other devices, but continues to supply the main memory where the open documents and programmes are stored. After the system has been re-activated, work can be resumed where it left off in less than 2 seconds with relatively new laptop models. If the laptop enters the Hibernate mode, the data from the main memory are saved to the hard disk in compressed form and the computer is shut down completely. Startup from this mode takes a little longer, but the desktop state is as it was before Hibernate was entered. A further improvement has been made to the power management functions introduced in Windows 2000. Hibernate and Standby now operate considerably faster and more reliably, saving Notebook batteries and increasing mobile operating time.

With Remote Desktop, Windows XP allows users remote access to all their applications, documents and network connections on their office computer's desktop. Through a secure, encrypted Internet or dial-up connection, users can link to the company PC and create a virtual session there with the customary Windows desktop on the remote computer. Since all the programmes run on the company PC and only keyboard and mouse inputs and display outputs are transferred, the remote desktop is also ideally suited for modem or ISDN connections.

Wireless networks (e.g. 802.11b/g-WLANs) support Windows XP Professional as standard. Windows XP Professional recognises and configures everything automatically as soon as a WLAN connection is established or cell-to-cell roaming takes place. Certificates that are distributed or stored on smart cards as well as 802.1X authentication ensure the utmost transfer security (both in cable Ethernet and in wireless networks).

Windows XP Professional recognises networks and their TCP/IP settings, so that the network connectivity of the PC is established automatically and is maintained while roaming. If the network does not provide for automatic configuration, Windows XP Professional allows working with alternative TCP/IP settings. In this way a notebook user can draw the TCP/IP configuration dynamically from a DHCP server in the company network and work with static TCP/IP addresses at home.

Windows XP is an uncompromising further development of the proven Windows 2000 technology and therefore operates even more stably than its predecessors. The new Windows engine is based on the dependable 32-bit architecture of Windows 2000, featuring a fully protected memory model. During installation, the setup programme of the operating system can incorporate driver, compatibility and security updates from the Internet if the user chooses to install them, even if such have become available after the Windows XP CD-ROM was issued. Windows Update keeps the system abreast of the latest changes. Service packs, troubleshooting and drivers for new hardware devices can be downloaded from Microsoft on the Internet easily and automatically if the user so wants.

A large number of programmes run on Windows XP that failed to run on Windows 2000. If a programme can be run only on Windows 95 or Windows 98, most likely it can now also be run on the new operating system. Should an older application nevertheless not be supported by Windows XP, such can be performed in a special compatibility mode presenting the characteristics of earlier Windows versions.

Windows XP gives top priority to the quality of device drivers since it contributes significantly to system stability. When a new component is installed, an indication is given whether Microsoft has tested and certified the driver. If a non-certified driver has been installed and fails to operate correctly, the system can revert to the earlier operational driver within a matter of seconds. By means of the side-by-side DLL support, multiple versions of individual components can run side by side, so that every application uses the version best suited to it.

If failure is experienced despite the significantly enhanced driver handling, the System Restore feature can restore the PC to a previous state and thus cancel all configuration changes made in the meantime, even after weeks have elapsed.

Great importance is attached to security in Windows XP. Sensitive files can be stored encrypted, passwords and user names in special areas are protected from unauthorised access. The Encrypting File System (EFS) permits files containing particularly sensitive information to be stored in encrypted form on the hard disk of the PC or on network servers. This dependably protects data including offline files and folders from unauthorised access. The new multi-user support of EFS allows multiple (authorised) users to encrypt and jointly inspect sensitive files.

The Internet firewall (Internet Connection Firewall - ICF) integrated in Windows XP automatically protects data from unauthorised access from the Internet. The function just has to be activated for the data communication or LAN connection wanted. Windows XP Professional uses 128-bit encryption for all encryption operations. In addition, there are many further security features in Windows XP, such as Kerberos V5 authentication support, Internet Protocol Security (IPSec), centrally defined security guidelines and a lot more.

A series of features in Windows XP simplifies installation and administration of the operating system and helps cut the costs for these tasks within the company. The following is just a selection of these features.

For companies wanting to install XP Professional on several PCs, Microsoft has provided sophisticated, tried and tested mechanisms for automatic installation. These allow unattended Windows XP installations to be preconfigured, enabling the setup routine to run reliably and without user interaction. Operating system images can be created with ease, even for PCs with different hardware, and can be installed automatically across the network with the aid of the Windows 2000-compliant Remote Installation Service (RIS).

Of interest to international companies and educational establishments: Windows XP Professional supports the Multilingual User Interface (MUI). A PC can have a number of interfaces at a time (English, German, Spanish etc.), providing every user with the Windows desktop in the language he or she prefers.

Administrators can deploy software restriction policies to prevent unwanted applications from running on Windows XP Professional PCs, thus restricting workstations to the programmes important for the company.

In addition to the Group Policy settings provided for Windows 2000, there are numerous new ones provided for Windows XP Professional for even more comprehensive, policy-based management through the Active Directory. With the new Resultant Set of Policy (RSoP) tool of Windows XP Professional, administrators have a powerful tool to plan, monitor and troubleshoot the impact of various group policies on a specific user or computer. The location-related group policies prove to be exceedingly useful for users often en route with their notebook. If the user is at the office, the company group policies apply. When the user is away or at home, however, he can employ the Windows XP Professional functions useful for single PCs or in small LANs, without this requiring reconfiguration by an administrator or the user himself.

The USMT (User State Migration Tool) allows easy migration of a user's data and personal settings from the original system to Windows XP. Preferred desktop settings, folder options of the Windows Explorer, Internet configurations and favourites, e-mail access, messages and addresses, network drives and printers, specific folders etc. can be migrated with ease.

Windows XP offers consistent support and end-to-end-integration for all digital media. Transferring and recording, display and playback, archiving and sending: Windows XP offers context-sensitive support through every step and automatically proposes all pertinent tasks. Windows XP opens up audio, photo and video capabilities as never experienced before. The

Media Player can create MP3 files and play DVD videos via the PC through separate plug-ins. Windows XP automatically recognises if a digital camera or scanner is connected to the PC.

Windows XP identifies what files are in a folder and not only displays a thumbnail, but also proposes the most likely options for the respective media type in the new taskbar: audio data, for example in the "My Music" folder, is played at a click of the mouse.

The Remote Assistance integrated in the Help and Support Centre of Windows XP enables colleagues or an in-house support centre to be called for help at any time. Instead of the problem being described to a specialist in words over the phone, the expert can show the solution in a clear and intelligible manner on the user's own screen by remote control. If the expert's aid is enlisted via the online contacts of the Windows Messenger or via e-mail, he receives the authorisation to view the user's screen on his own PC to help find or even implement the solution. The user can follow every step and retains full control at all times. The expert's Remote Assistance authorisation for the user's PC expires automatically after the session ends. This comprehensive and central support tool of Windows XP allows the number of calls for support addressed to the company's own support centre to be reduced and hence cuts the costs for support with hardware and software problems.

Windows Longhorn

Microsoft has announced that Windows Longhorn will be the definitive successor to Windows XP for the consumer sector. There will not be a Longhorn server variant.

Up to now the operating systems from Microsoft were always launched in two different versions for the client and for server administrators. This tradition started with Windows 2000, since the Professional version was designed for workstations and the Server for high-end networks and Web hosting. The procedure was similar with Windows XP, where the Windows Server 2003 variant was only available more than 14 months after the first publication of XP.

Microsoft is now taking a completely new direction with Windows Longhorn, since this version will only be optimised for clients and workstations. Key elements for the developer platform Windows WinFX under Windows XP and Windows Server 2003 will be available at that point.

Longhorn offers users fundamental improvements in productivity, further developments in security and reliability and important new possibilities for software developers.

The successor with the code name Blackcomb will only be used on servers and high-end systems.

According to current information, Microsoft will keep to its Longhorn date, which plans for client operating systems for early 2006.

2.1.3 Windows Server 2003

The new Microsoft Windows Server 2003 has been available since April 2003. As compared to the Windows 2000 Server, many features and functions were developed further or from scratch. The following versions are available:

Windows Server 2003, Standard Edition

- 2-way symmetric multiprocessing (SMP)
- 4 GB RAM

Windows Server 2003, Enterprise Edition

- 8-way symmetric multiprocessing (SMP)
- Clustering with up to eight node
- 32 GB RAM in 32-bit versions and 64 GB RAM in 64-bit versions
- Hot Add Memory
- Non-Uniform Memory Access (NUMA)

Windows Server 2003, Datacenter Edition

- 32-way symmetric multiprocessing (SMP)
- 64 GB RAM in 32-bit versions and 128 GB RAM in 64-bit versions
- Windows Sockets: Direct access for SANs

This version will only be available under the Windows Datacenter programme, offering a package of hardware, software and services.

Windows Server 2003, Web Edition

- 2-way symmetric multiprocessing (SMP)
- 2 GB RAM
- Windows Server 2003, Web Edition is specially designed for use as a Web server. Servers with this operating system can be members of an Active Directory domain, even though they cannot offer the Active Directory service themselves.

  This version will be available only through special partners.

An overview of the principle features of Windows Server 2003

XML Web services

The XML Web services provide reusable components built on industry standards that invoke capabilities from other applications independent of the way the applications were built, their operating system or platform, or the devices used to access them. The IIS 6.0 security settings are locked down during setup to ensure that only required services can be run. Using the IIS Security Lockdown wizard, server functionality is enabled or disabled based on the administrator's requirements.

Directory services

Active Directory security settings for users and network resources span from the core to the edge of the network, helping to make a secure end-to-end network. Active Directory is now faster and more robust, even over unreliable WAN connections, thanks to more efficient synchronisation, replication, and credential caching in branch office domain controllers.

Update management

The automatic update provides the ability to systematically download critical operating system updates, such as security fixes and other patches. Administrators can select when to install these critical operating system updates.

Internet firewall

The Internet firewall of the server makes Internet connection more secure.

Server hardware support

Driver verifiers check new device drivers to help keep the server up and running.

Application verification

Applications executed on Windows Server 2003 can be tested in advance, for example for heap corruption and compatibility.

Server event tracking

Administrators can record the operating time exactly with the help of the new protocol. It writes Windows events for server shutdowns to a log file.

Configure your server wizard

The Configure Your Server wizard leads administrators through the process of setting up various server roles such as a file server, print server, or remote access server, ensuring that components are installed and configured correctly the first time.

Manage your server wizard

The Manage Your Server wizard provides an interface for ongoing management of the server, making it easy to perform such common tasks as adding new users and creating file shares.

Remote server administration

With Remote Desktop for Administration (formerly known as Terminal Services in Remote Administration mode), administrators can manage a computer from virtually any other computer on the network.

Shadow copy

This feature provides time-based network sharing. Administrators can save network folder contents and later determine the status of the folders as they existed at this time. End users can recover accidentally deleted files or folders on network shares without requiring system administrator intervention.

Terminal server

When using Terminal Server, a user can access programmes running on the server. For example, a user can access a virtual Windows XP Professional desktop and x86-based applications for Windows from hardware that cannot run the software locally. Terminal Server provides this capability for both Windows and non-Windows-based client devices.

Additional functions of the Enterprise Edition include

Cluster service: The cluster service for Windows Server 2003, Enterprise Edition and for Datacenter Edition supports up to eight-node clusters. This provides increased flexibility for adding and removing hardware in a geographically dispersed cluster environment, as well as providing improved scaling options. Various cluster configurations with dedicated storage are possible:

- Multiple cluster configurations in a Storage Area Network (SAN)
- Clusters spanning multiple sites, geographically dispersed clusters

Metadirectory services

Microsoft Metadirectory Services (MMS) helps companies to integrate identity information from multiple directories, databases and files with Active Directory. MMS provides a unified view of identity information, enables the integration of business processes with MMS, and helps synchronise identity information across organisations.

Hot add memory

Hot Add Memory allows memory to be added to a running computer and made available to the operating system and applications as part of the normal memory pool. This does not require re-booting and involves no downtime. This feature currently will operate only on servers that have the respective hardware support.

Non-Uniform Memory Access (NUMA)

System firmware can create a table called the Static Resource Affinity Table that describes the NUMA topology of the system. Windows Server 2003, Enterprise Edition uses this table to apply NUMA awareness to application processes, default affinity settings, thread scheduling, and memory management features. Additionally, the topology information is made available to applications using a set of NUMA APIs.

Terminal services session directory

This load balancing feature allows users to reconnect to a disconnected session. Session Directory is compatible with the Windows Server 2003 load balancing service and is supported by third-party external load balancer products.

The additional functions of the Datacenter Edition include an expanded physical memory space: On 32-bit Intel® platforms, Windows Server 2003, Datacenter Edition supports Physical Address Extension (PAE), which extends system memory capability to 64 GB. On 64-bit Intel® platforms, the memory support increases to a maximum of 16 terabytes.

Windows Sockets: Direct access for SANs

This feature enables Windows Sockets applications that use TCP/IP to obtain the performance benefits of storage area networks (SANs) without making application modifications. The fundamental component of this technology is a multilayer Windows Socket service that emulates TCP/IP via native SAN services.

2.1.4 Windows Small Business Server 2003

Windows Small Business Server 2003 is a server solution offering small and medium-sized companies features such as e-mail, secure Internet connectivity, business Intranets, remote access, support for portable devices, as well as an application platform for collaboration. Windows Small Business Server 2003 is so user-friendly that it carries out the network configuration and server installation virtually automatically. Windows Small Business Server 2003 is a fourth generation product. With the Microsoft Windows Server™ 2003 operating system at its core, it incorporates Microsoft Exchange Server 2003 and Microsoft Windows® SharePoint™ Services. Using this combination of technology, and by incorporating innovative management tools, small and medium-sized companies can create a faster and more efficient business environment. Windows Small Business Server 2003 is available in two versions.

The Standard Edition includes:

- Windows Server 2003
- Exchange Server 2003
- Outlook 2003
- Windows SharePoint Services

The Premium Edition also includes:

- FrontPage 2003
- SQL Server 2000 SP3
- ISA Server

The principle features of the Windows Small Business Servers 2003 are briefly described here.

E-mail, networking and Internet connectivity

With Windows Small Business Server 2003 a company has everything it needs to effectively use the Internet. It offers a solution for common access to the Internet that is easy to manage, a firewall to protect the local network, Internet
e-mail based on the Exchange Server and productivity tools, e.g. Microsoft Outlook® Web Access und Remote Web Workplace.

2.2 Unix Operating Systems

Unix is still the leading operating system in the workstation world. In fact, it is a family of operating systems because practically all workstations manufacturer supply their own version of Unix, which at least as far as the user interface is concerned, differs considerably from the other versions. However, there is a tendency to overcome this wide variance of interfaces, as several manufacturers have begun to port their system to alien architectures.

The Unix implementations can be categorised under two standards: Berkeley Unix (BSD) and AT&T's System V Release 4 (SVR4). At present, the SVR4 is ahead of its rival - new versions of Unix follow its standard. As a general rule, if a programme is written for one of these two standards, it can be ported to another system of the same standard without major difficulty.

Different standards are also employed for the user interfaces (GUI - Graphical User Interface). However, the more recent ones all follow the X11 definition. For several years, the MOTIF definition - which is also based on that of X11 - has clearly been progressing. More and more Unix implementations are using this interface, while the use of the competitor interfaces, like OPENLOOK, have been on the decline.

Linux

Linux is a freely available multi-tasking and multi-user operating system. Linux was invented by Linus Torvalds and developed further by a number of other developers throughout the world. From the outset, Linux was placed under General Public License (GPL). The system can be distributed, used and expanded free of charge. In this way, developers have access to all the source codes, thus being able to integrate new functions easily or to find and eliminate programming bugs quickly. Thereby drivers for new adapters (SCSI controllers, graphics cards, etc.) can be integrated very efficiently.

Presently, Linux is successfully being used by several millions of users worldwide. The user groups vary from private users, training companies, universities, research centres right through to commercial users and companies, who consider Linux to be a real alternative to other operating systems.
The extensive network support of Linux, including different servers such as Appletalk, Netware or LAN Manager servers as well as the multitude of supported network protocols, makes Linux a secure and reliable network server system.

There are two different ways of obtaining Linux: All necessary parts can be downloaded free from the Internet. This means that an individual operating system can be assembled almost free of charge. The use of a so-called distribution is easier. This is offered by various companies and includes a wide range of applications and installation programmes, that significantly simplify the installation of Linux.
The distributions differ especially in terms of the enclosed components, such as programming environments, network software and graphical user interfaces. We recommend distributions from SuSE or Red Hat. Both these Linux distributions are very sophisticated and include a wide range of documentation as well as a graphically supported installation. All transtec Linux systems are certified and offered with the current versions of SuSE and Red Hat.

In addition to their distributions for PCs and workstations, both SuSE and Red Hat offer special packages for server operation. With SuSE these are the SuSE Linux Enterprise Server and the SuSE Linux Standard Server. Apart from the composition of the package specifically for server operation, it is distinguished from the "normal" distribution by the following points. For one thing, SuSE carries out extensive tests to ensure that the individual packages are compatible with one another and with important business applications. What is more, SuSE guarantees up to 2 years' support for the package, even after the respective version has been superseded. Equally important for the business environment is the provision of important patches and updates.

SuSE LINUX Enterprise Server

The SuSE LINUX Enterprise Server is one of the leading server operating systems for professional use in heterogeneous IT environments in all sizes and orientations. It is available for all relevant hardware platforms: for the 32 and 64-bit processors from AMD and Intel®, as well as for the entire eServer series from IBM, including mainframes.
The SuSE LINUX Enterprise Server can be used for all company-relevant server and network services: from file, print, Web and security, through to application and middleware solutions, particularly for servers with high availability, data and network services of critical corporate importance. A single server operating system with a uniform Linux code basis!

advantage

- Considerable savings in cost
- Enhanced investment security through system maintenance and open source code
- Higher scalability and stability
- High Performance Cluster
- High availability cluster

SuSE LINUX Standard Server

The SuSE LINUX Standard server wit its graphical administrator interface is oriented towards small organisations and departments that want to make their Internet access, as well as e-mail, printing and data services, reliable and secure. The Standard Server is available for AMD and Intel® 32-bit processors (x86) and supports up to two CPUs. Support and maintenance are included in the package with the SuSE Maintenance Programme.

The only information needed for setup and management of SuSE LINUX Standard Server is the network address, the information of the Internet provider and the company/user name. By means of the graphical interface, the Internet access can be configured to buffer and filter Web contents, increasing security and saving provider costs.

The assignment and administration of file and user rights have been fully revised and substantially improved. In SuSE LINUX Standard Server user data is administered in a central directory service (LDAP). Security of data has been enhanced by using Access Control Lists (ACL) that can be configured and modified easily and quickly via a Web interface even from remote hosts.

Graphical configuration wizards allow a user-friendly configuration of server services: for instance, use as a Windows domain controller, file and print server in Windows environments or as an Internet gateway and name server (DNS). In addition, a complete e-mail server for small organisations is integrated in SuSE LINUX Standard Server.

Clients can be easily and quickly connected to SuSE LINUX Standard Server. The automatic allocation of network IP addresses (DHCP) can be configured entirely graphically. Windows domain accounts are set up automatically. In this way external staff members are provided with secure access to company data via VPN.

Apart from numerous server services for secure and efficient utilisation of the Internet and file/print services under Windows and Unix, SuSE LINUX Standard Server can also be used as application server. Thus, small organisations and departments are equipped with a versatile server operating system for the establishment of a powerful Intranet and a reliable basis for the operation of client/server applications.

Furthermore, SuSE offers complete packages for special applications, such as e-mail servers or firewalls.

The development of these packages reflects the general trend: the use of Linux in commercial systems as well is on the increase, where it provides a cost-effective and dependable basis for many applications.

2.3 Computer Viruses/Types of Viruses

A sad story: There are computer viruses on nearly all computer platforms and operating systems. These are nowadays practically only spread via e-mails and the Internet.

A computer virus is an instruction sequence which, if executed, changes a memory area by copying itself there. This memory area can be an executable file, or a programme stored on floppy disk, hard disk, etc., or also in RAM.

File viruses

File viruses attach themselves to selected programmes, employing various techniques to do so. They are spread whenever an already infected programme is called up. Resident file viruses represent an intensified version in as much as they take root in the memory after they are called up. Thus each time a programme is executed infection is repeated. This type of virus is harder to eliminate; in some cases, it can even survive a warm start of the computer. Most of the viruses known today belong to this category. So-called stealth viruses were developed to go undetected by virus scanners, checksum programmes, or protection programmes. Similar to the U.S. Stealth bomber, they possess a camouflage mechanism that enables them to conceal their existence. Polymorph viruses are the latest and most dangerous kind of file viruses. This category consists of viruses, which alter their appearance for each new infection. Normally, they do this by re-encoding the virus code with a key which varies with each infection.

System viruses use particular programme components, like the master boot record, partition boot record, diskette boot sector, FAT, and the operating system's root directory as the infection medium. These system areas are connected to the respective data media, and infection occurs when the respective data media is used.

Boot sector viruses attack exclusively the boot sector on floppy disk or hard disks. The infection spreads slowly, since it can only take place during booting up from an infected floppy disk.

transtec recommends that systems be continuously checked with the help of suitable virus software so as to prevent extensive damages.

3. Cluster

3.1 High Performance Computing Clusters

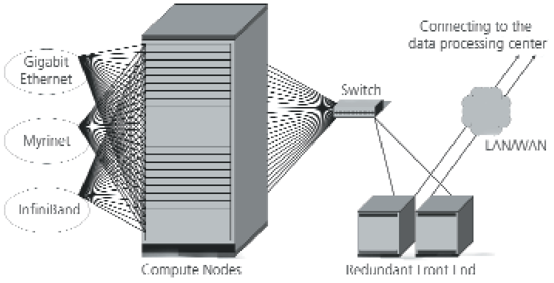
In the mainframe sector (High Performance Computing HPC) diverse mainframe solutions are classically encountered. Due to their vastly improved price/performance ratio, however, HPC clusters on IA32 basis have in recent years conquered an appreciable market share.

Use of a mainframe is generally considered for three reasons:

- Due to its complexity, the problem is in principle unsolvable on single systems.
- The user wants to have the result calculated with higher accuracy.
- The user wants to save time and arrive faster at a result.

Apart from the price advantage, factors in favour of an HPC cluster as an alternative are the good scalability and simple administration.

Ideal setup of an HPC cluster



From the user's viewpoint, a cluster consists of a software interface distributing its application to the resources. These programmes, also called middleware, are superposed on the operating system. The computing jobs are then distributed to the processors via a dedicated network.
The front end computer administers the compute nodes to be reached via a private network. They do the actual work and send the partial results to the front end. Inexpensive Common Off The Shelf (COTS) hardware is generally used. In practice the number of nodes is usually between 8 and 256, although they can certainly reach into the thousands.
The task the front end performs is to administer the jobs of the users in a queue, to supply the nodes with jobs, to archive the results following receipt and monitor the operating status of the nodes. The management software allows the administrator to replace defective nodes quickly or to expand clusters as needed.
Front end failure leads to total cluster failure. That is why it should have a redundant power supply unit, be connected via a UPS, the boot hard disk should be mirrored in RAID 1 and the data distributed over multiple hard disks in RAID 5.
The number and type of nodes as well as the network type determine the performance of the cluster. For optimisation, different architectures should be examined for a specific application. For the processor variants, comparison should be made particularly between the Intel® Xeon EM64T and AMD Opteron CPU. The size of the main memory should be selected such that swapping is not required. The demands made on the speed of the hard disk locally integrated in the nodes are frequently met by SATA hard disks.

A cluster can be used in both batch and parallel mode. In the former case the same programme runs on all the nodes, such as a Raytracer. The front end sends various jobs or start values to the nodes. After processing through compute nodes, the results are sent back to the front end, which activates them after a quick plausibility check. The next task is sent as soon as the results come in. When calculating a movie a cluster typically works in batch mode, for example. Every node calculates a different section of the image. A data exchange with other images of the film is not necessary. The more nodes participating in a computing job, the quicker the film will be finished.

In parallel mode, the compute nodes operate simultaneously on a common end result. In this case the results have to be matched, the data has to be synchronised. Unlike a mainframe, which is an SMP (symmetric multiprocessor system) and has a central shared memory, an HPC cluster has memories locally distributed in the computing nodes. A high-performance network is necessary in order to achieve rapid access to the data in an adjacent memory area in the parallel mode.

The range of network cards for this is diversified. Gigabit Ethernet is meanwhile usually included in the nodes as standard, only the costs of a GigE switch are still significantly higher than in the Mbit/s range. For parallel operation in the high-performance segment, however, the latency of this technology, the coordination time until the beginning of data transfer, is often too long. InfiniBand and Myrinet, SCI Dolphin or Quadrics networks can be up to ten times faster. Apart from an optimised hardware, this is achieved by the use of a proprietary protocol. The question of which technology to go for has to be resolved on a case-to-case basis. It is helpful to compare experimental results in a test environment. InfiniBand usually has the best price/performance ratio.

Extreme node expansion for increasing performance in the parallel mode is usually pointless. The speedup, or the performance gained, depends on the degree to which the application is parallelised. If, for example, when 10 nodes are used the cluster achieves nine times the speed of a single node, the speedup has the factor 9. If 100 notes were used the same application would achieve only 48 times the speed of a single node.
There are several ways to set up and configure the operating system on the cluster. Clusters are typically operated with Linux. Use of Windows is uncommon and demands special knowledge. Microsoft has announced a Windows version especially for HPC clusters for the 2nd quarter of 2005. The choice of the Linux distribution depends on the administrator's preferences. The classic version is the local installation of the operating system on the front end and on the nodes. The increased administration overheads can be a drawback with this solution. For instance, maintenance of the operating system requires individual kernel update on all nodes.
The alternative is an installation according to the boot-from-LAN concept. Here the respective images are deposited on the front end, which are loaded during booting via the network card in the main memory. The local hard disk consequently serves only for external storage of intermediate results.

The set up of an HPC cluster in 19" technology is advisable. The packing density increases from 0.5 CPU/U with standard enclosures to up to 2 CPU/U. Furthermore, the cooling of the nodes can be improved and maintenance is facilitated. Blade systems increase the packing density to 4 CPU/U. The waste heat of a cluster should not be underestimated. 10 kW per 19" rack can be reached with 40 dual DPU nodes with 250 watts each, for example. Reliable operation often requires climate control. Using a completely water-cooled server rack is an economical option.

The failure of a computing node is not critical as a rule. It reduces the cluster performance in proportion to the total number of nodes. On-site service is therefore generally not required, which reduces the upfront costs. The administrators can easily exchange defective nodes themselves. Similarly, subsequent extension can also be carried out without any difficulty. Additional nodes do not necessarily have to have the same performance. With a system used in a batch mode, the native extra performance is integrated without loss. In a parallel mode, a faster node will not operate with full performance in practice, but only with the effective speed of the slowest node in the system. In practice therefore the administrator will in this case be more likely to define subclusters of uniform performance.

HPC clusters are meanwhile put to a wide variety of uses. Typical examples are the crash test simulation, aerodynamics calculation, drug research, data mining in banking and insurance business, meteorology, 3-D animation, astronomical calculations, scientific analysis, simulations of passenger or vehicle traffic or creating film sequences. Because of the stability of the solution and the sophisticated technology, HPC clusters will surely be used even more in the future.

Further information can be found online in the HPC whitepaper from transtec AG.

3.2 High Availability Clusters

A high availability cluster refers to the close connection of several systems to form a group which is managed and controlled by a cluster software. Physically, a cluster is a group of two or more independent servers that serve the same group of clients and can access the same data. Putting this into practice with the latest technologies means, in general, that servers are connected via the usual I/O buses and a standard network for client access. From a logical point of view, a cluster represents a single administrative unit, in which any server can provide any authorised client with any available service. The servers must have access to the same data and be equipped with a common security system. This means that according to the current technical standard, all servers in a cluster tend to have the same architecture and run on the same version of the same operating system.

Although clusters can be structured in different ways, all types of clusters share the same three advantages:

- Error-tolerant high availability of applications and data
- Scalability of hardware resources
- Easier administration of large or rapidly growing systems

Higher availability

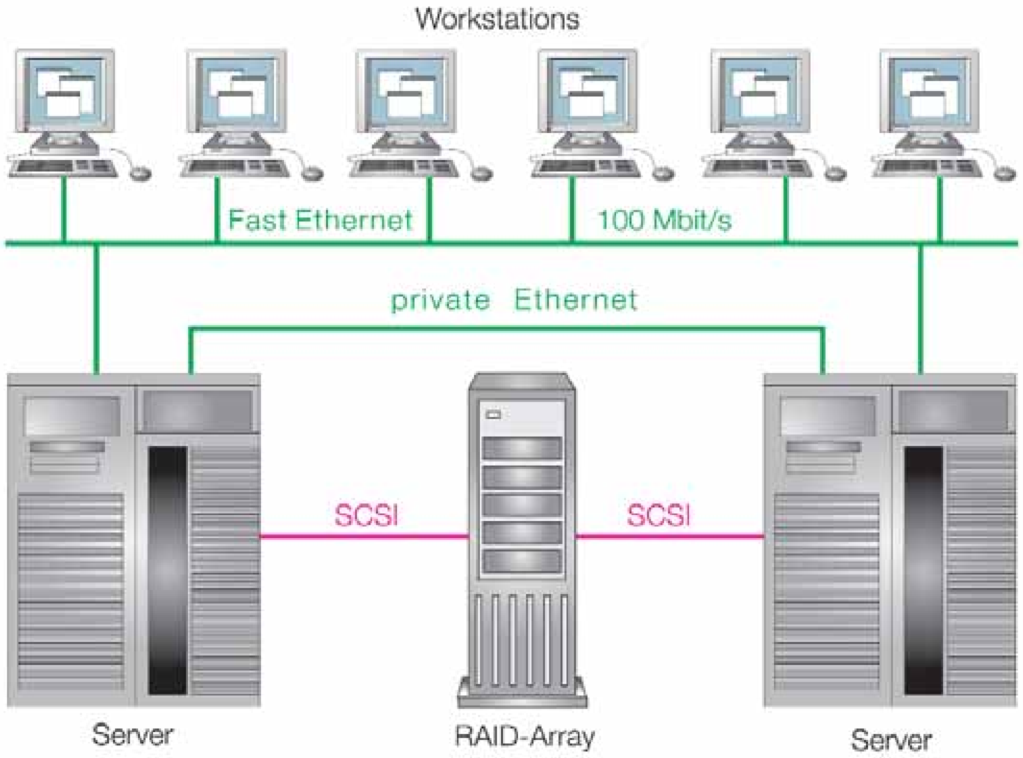
In company critical environments, the high availability of services (e.g. Web servers, databases, or network file systems) forms the factor for success of a company. Common reasons for problems with the availability of services are the different forms of system malfunctions. A malfunction can be due to the hardware, the software, the application or the system execution. Protection against these malfunctions can be provided using a cluster as an application server or a data server. Besides providing redundancy for computing performance, I/O and storage components, clusters also enable one server to function for another server without interruption, if the latter breaks down. In this way, clusters allow high availability of server services even when malfunctions occur:

- When hardware or software malfunctions occur or when a server fails, the other server can take over its functions.
- When the network interface of a server fails so that the connection to the clients is interrupted, the clients still have the possibility of using the second server.
- When an I/O bus or an I/O adapter fails, the data of the RAID system can be accessed via an alternative path.
- In case of a disk failure, the data can still be accessed via the RAID system.

In all cases, both the servers and the clients must identify and deal with the malfunction. It is true that the cluster service performance is interrupted, but with clusters, this interruption normally lasts only a few seconds and not several minutes or hours as is the case with conventional recovery methods.

The illustration depicts a simple HA cluster, which consists of two

Microsoft cluster server configuration



networked servers and a RAID system for data storage.

Scalability

Another advantage of some cluster architectures is their scalability, which permits applications to be expanded beyond the capacity of a single server. Many applications have several threads of relatively delimited activities, which interact only occasionally. The applications with several threads can run as pseudo-parallel processes on one server with one processor or as genuine parallel processes in symmetrical multi-processor systems (SMP). In a cluster, groups of the application threads can be executed on different servers because they all have access to the same data. When an application becomes too large for a server, a second server can be installed, in order to create a cluster and increase the application's capacity.
Permitting servers to access the same data requires co-ordination. This co-ordination can be accomplished by a database manager or a distributed data system. In a cluster with no co-ordination, simultaneous and direct access by several servers to any data file is not possible. However, some applications can be scaled even with this limited accessibility. The applications can be distributed so that the individual programmes use different data.

Easier administration

The third advantage of clusters is their easier administration. Clusters simplify the complexity of system administration by enhancing the scope of the applications, data and user domains administered by a single system. The system comprises the following areas, among others:

- Operating systems
- Middleware
- Application maintenance
- Administration of user accounts
- Configuration management and data backup

The complexity and cost of system administration depend on the size and especially the number of the included systems. For instance, running daily data backups is necessary for all servers that store important data, independent of the amount of data to be secured. In addition, user-account modifications must be updated on all servers accessed by the user.
Clusters reduce the number of individual systems and thereby also the cost of system administration, by integrating a large number of applications, data and users into one computer system. One of the advantages is that a cluster system must contain only one set each of user accounts, data access authorisations, data backup rules, applications, data base managers, etc. Although the individual systems may vary due to the different cluster architectures used, it is normally more efficient to administer one cluster rather than the corresponding number of unconnected server systems.
The load distribution between the individual cluster servers depends on the operating system. Both Unix and Open VMS clusters offer automatic load distribution and scalability. Windows 2000 clusters currently offer only better availability. With the Cluster Server of the Windows 2000 Advanced Server, two servers can be connected within a cluster, with the Windows 2000 Datacenter Server, four servers can be connected. The automatic IP load distribution, which is independent from the Cluster server, allows the distribution of Web applications on up to 32 systems.

High availability clusters with Linux Failsafe

The Linux Failsafe cluster software is a universal, freely configurable, and error-tolerant high-availability system, with which individual services can be redundantly set up. This software enables a service to be automatically or manually migrated to another node (server) in case of an error. An error is not necessarily due to defective hardware: application errors, desolate processes, and critical system conditions are also recognised as such and handled accordingly. In most cases the individual nodes of the cluster must be able to access common data sections in order to take over the data in case of an error. An SCSI RAID or SAN architecture can be used here, which allows the servers to take over data sections in case of an error. If the data sections were successfully taken over, then the active cluster server can restart the services and make them available using the same IP.
Thus, the downtime of a service can be calculated and even controlled (within certain limitations), since various criteria can be configured for the migration of services. With the corresponding redundant server systems, an availability of at least 99.999% can be achieved with this solution.

Further information can be found online in the HPC whitepaper from transtec AG.

4. Storage buses

5. Hard disk and RAID

>> Chapters 4 and 5 can be found in the Internet

6. Storage Networks

6.1 Storage Area Networks

Mainframe solutions have been increasingly replaced with more economical, decentralised server and workplace computers ever since the introduction of the first IBM PC on August 12th, 1981. They normally administer only one application, which accesses local memory via SCSI protocol and supplies several clients via Ethernet.

A reverse trend has been felt since the end of the Nineties, however. The more and more complicated administration and increasing complexity from the numerous server systems have led increasingly to bottlenecks for IT departments. Additionally, the dynamic usage of CPU and storage resources adapted to current business needs is only possible to a certain degree.

The development of central storage systems and a more effective protocol than SCSI goes back to 1988, when Fibre Channel (FC) from ANSI (American National Standards Institute) was first defined as a standard. It took more than a decade though, until storage networks based on FC able to reach a broad market segment and could be considered mainstream technology. Two international organisations have done the most for the standardisation and dissemination of SANs:

- Storage Networking Industry Association (SNIA)
  www.snia.org
  www.snia-europe.org
- Fibre Channel Industry Association (FCIA)
  www.fibrechannel.org

Storage Area Networks (SAN) were the foundation of storage centralisation. A secondary, high-speed network by which the server and central storage systems communicate, and which exists parallel to a LAN for client/server data transfer.

The changeover from server-centred to storage-centred infrastructure requires fundamentally new approaches to high availability and data security in business. A failure affects several servers and their applications at the same time, no longer just a single server and its applications. As a result classic mainframe features such as redundant storage systems and data paths, time-controlled data copies (snapshots) and synchronous/asynchronous replication were integrated into the Open Systems world.

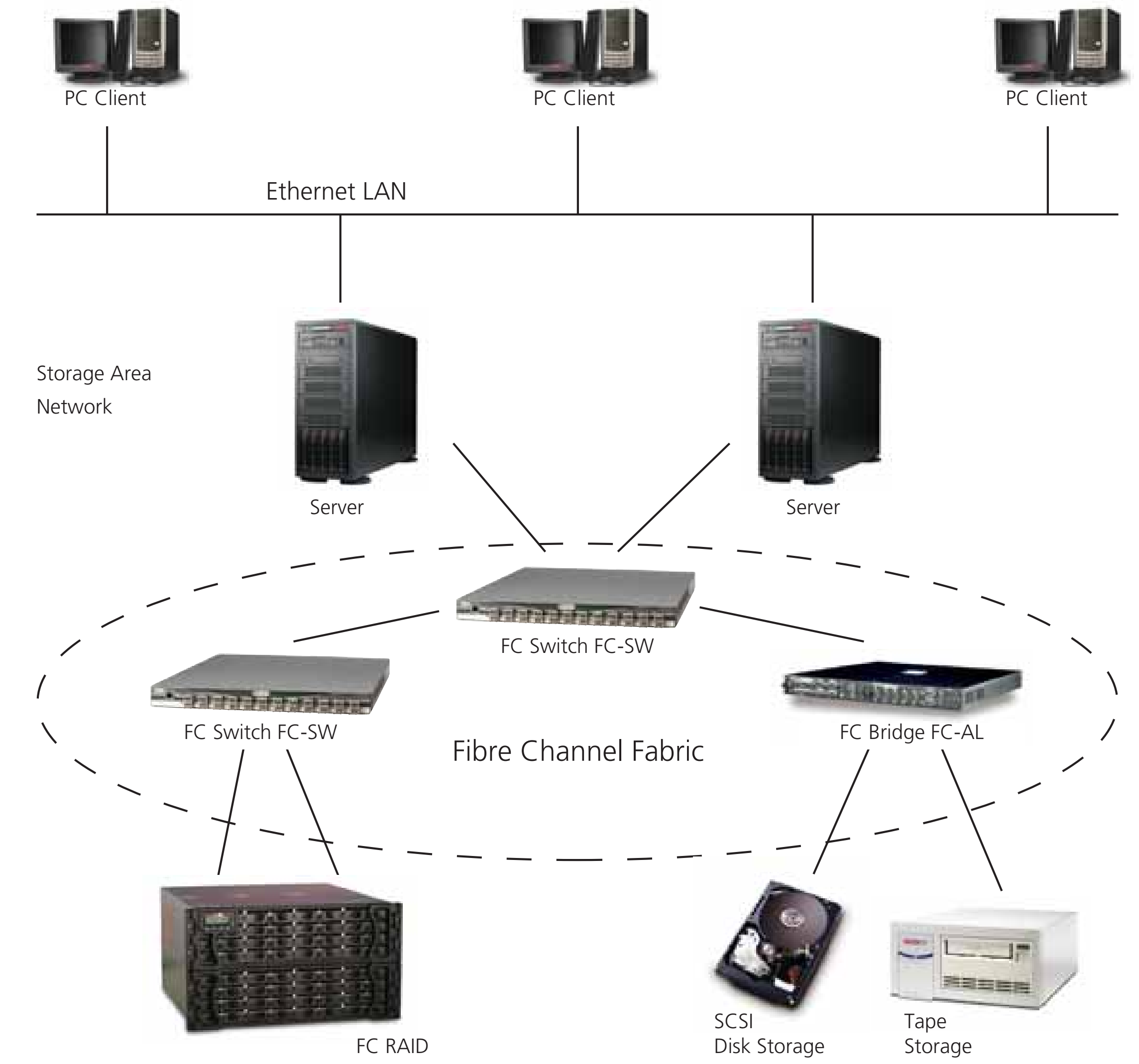
The introduction of the SAN is not done without problems. The user can look forward to very high performance, as well as common usage and flexible assignment of storage resources between server systems and a less complicated management. Some of these promises have meanwhile become reality. The higher bandwidth and I/O performance of FC storage systems are unquestioned. Studies from market researchers such as IDC and Gartner regularly have proven that IT administrators can manage up to seven times the amount of storage after incorporating a SAN. Faster data backup and relieving the LAN through FC networks are also welcome side effects.

The reality as regards flexibility and management was disillusioning, however. The automated, dynamic assignment of storage resources was only possible in the last two years with increased market maturity in virtualisation solutions. The shared use of a logical SAN volume by several, non-clustered servers remains a challenge. A lack of compatibility for standard products, many proprietary solutions and lacking management standards were commonplace for FC SANs.

The lack of interoperability has been overcome for the most part through user pressure. Proprietary solutions are limited to the high-priced Enterprise storage systems from companies such as EMC, IBM or Hitachi. In April 2003 the SMI-S 1.0 specification (previously known as BlueFin) was agreed by the SNIA and accepted by the IT industry, so that there is now a standard for managing the storage systems. Furthermore, iSCSI was defined as a standard for more reasonably priced, Ethernet-based storage networks. The SANs will continue to develop and we can look forward to progress in the next few years.

>> *See schematic figure p.118:*
Structure of an FC-SAN

Structure of an FC-SAN



6.2 Fibre Channel

6.2.1 Fibre Channel Fundamentals

Fibre Channel (FC) is the general term for a standard array that was developed, and is being developed further, by ANSI to establish new protocols for a flexible information transfer. This development began in 1988, as an extension to the standard Intelligent Peripheral Interface (IPI) Enhanced Physical and branched into several directions.

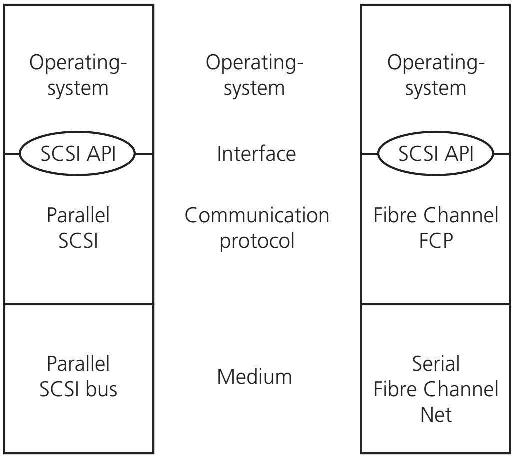
The principle goals of this development are:

- High-speed transfer of large amounts of data
- Separating the logical protocol from the physical interface
- New definition and implementation of interfaces
- Implementation of different protocols and the simultaneous transfer to the same base, if possible.
- Standardisation of interfaces and reduction of interface formats.

For bit-parallel signal transmission the following challenges have emerged at high clock speeds. Signals that leave the sender at the same time must reach the receiver at nearly the same time, which turns out to be problematic with increasing transmission rates. Serial data transport does not have this kind of limitation on the phase shift. As a matter of fact, the SCSI instruction sets are placed on a serial interface in order solve the problem of the phase shifts with high-speed data transmission.
FC makes a protocol interface available on serial hardware only in order to simultaneously transport other protocols – mostly SCSI, but also IPI-3, IP etc. This is possible because FC serves as a carrier for these instruction sets in a way that the recipient can distinguish between the two. This separation of the I/O operation from the physical interface is an important performance feature. All applications running on a server have a view of devices connected via FC as normal SCSI devices and send or receive SCSI I/O requests. Few or no software or driver changes had to be done.

The protocol overhead is also significantly reduced in comparison to parallel SCSI and TCP/IP in order to increase the transmission rate. The FC SAN can achieve an information data load of over 90 percent at a given clock speed, while conventional networks in actual operation can only reach 20 to 60 percent of the maximum transmission rate possible.

SCSI vs. FC interface



Data transmission and frame structure

The entire connection process between two FC devices is called exchange. An exchange consists of sequences that are divided into frames, transferred and reconstituted again in reverse order on the distant end.

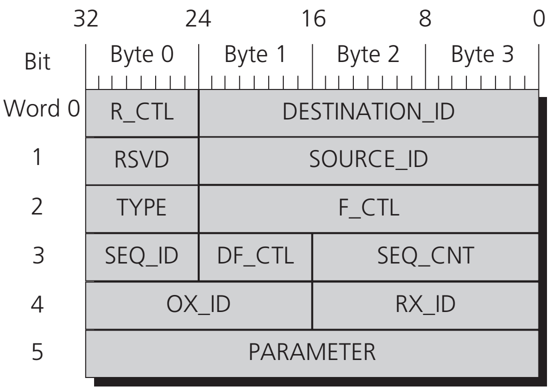
Exchange									
Sequence					Sequence				
Frame	Frame	Frame	Frame	Frame	Frame	Frame	Frame	Frame	Frame

An FC frame consists of so-called transmission words (TW) that are 4 bytes long. The frame size is variable with maximum 2148 bytes. The frame contains information data, a frame header, optional headers, CRC information, SOF (start of frame) and EOF (end of frame). After the removal of all additional information, only 2048 bytes remain for information data per frame. The length of the shortest FC frame (a frame without information data) is thus 36 bytes. This corresponds to a maximum information data share of 95.3 percent.

SOF 1 TW	Frame Header 6 TW	Otionale Header (64 Bytes)+Daten 0-528 TW	CRW 1 TW	EOF 1 TW

The frame header with a size of 6 TW (24 bytes) is divided into fields with control information. It also contains the source and target address of the FC frame.

Structure of an FC frame header



32 24 16 8 0

Bit	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	R_CTL	DESTINATION_ID		
1	RSVD	SOURCE_ID		
2	TYPE	F_CTL		
3	SEQ_ID	DF_CTL	SEQ_CNT	
4	OX_ID		RX_ID	
5	PARAMETER			

The goal for the development of Fibre Channel was also to bring the two established methods of data transmission (channel technology, network) together incorporating the advantages of each technology. Fibre channel was to combine the best aspects of these two opposing communication methods in a new I/O interface, which offers block-based data access with typical network characteristics.

Traditional I/O channels

- **Block-based data access**
- **Works in a closed, structured environment with a predefined configuration**
- **Peripheral devices are directly connected to the host system**
- **Any change causes changes to the configuration**
- **The host system contains all configuration information**
- **Transfer of large amounts of data between host and devices**
- **Processing overhead is reduced to a minimum and requires little to no software support**
- **The most important requirement for data transmission is error-free delivery, in which transmission delay time is of secondary importance.**

Networks

- **File-based data access**
- **Typically functions in an open, unstructured environment, which can be changed independent of servers**
- **Each host can communicate with each device at any time (peer-to-peer)**
- **Increased software support is needed for access control, establishing and closing sessions, packing data into packets, routing etc.**
- **Generally larger protocol overhead, i.e. poorer ratio between overall and information data**
- **There are also applications that do not necessarily require error-free data transmission. For example for video/audio streaming the continuity of data transmission is usually more important than error-free transmission**

6.2.2 Topologies for Fibre Channel Networks

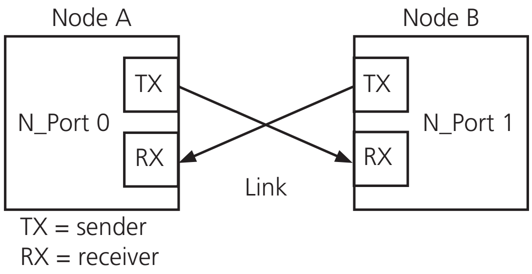
Fibre Channel devices are also called nodes; each has at least one port to permit access to the outside world (in other words, to another node). Each Fibre Channel port has a send and a receive channel. The connection between the send and receive channels of two ports is made via electric or standard fibre optic cables, and is called a link.

The way in which two or more ports are connected with each other is summarised under the term topology.
The Fibre Channel standard defines three topologies: point-to-point, arbitrated loop and fabric.

Point-to-point

This is the simplest way to connect two FC ports with one another. The two ports are connected via a crosslink, so that the entire bandwidths of both devices are available. The ports are called N_PORTs.

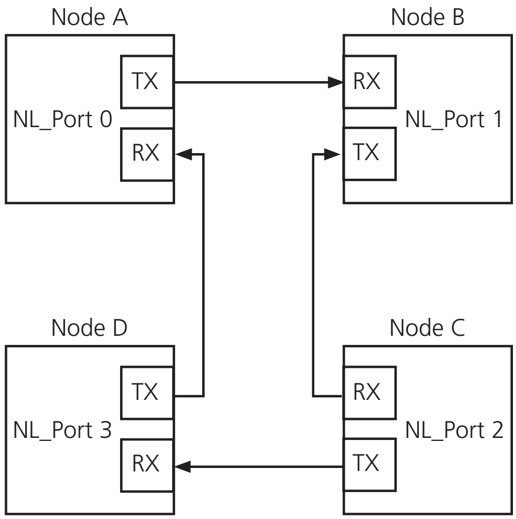
Point-to-point topology



Arbitrated loop

In an arbitrated loop (AL) the ports are connected in a ring. As many as 126 ports can be connected to each other. The data packets are transferred from port to port. They are only accepted by the ports for which they are meant. Packets not meant for a certain port will simply be transferred on. All ports share the bandwidth. Two ports are active at the same time. An arbitrated loop is also created by a loop switch, however without the limitation of a reduced bandwidth due to several devices.

Arbitrated loop topology



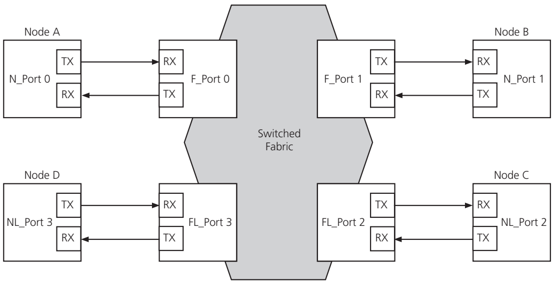
In order to be able to handle RAID memory systems more easily, the FC-AL supports the backplane architecture, as well as the normal cable link. The hard disks are connected to the backplane via a 40-pin SCA (Single Connector Attachment) plug, which includes both data and power supply lines. If there is no drive in one of the ports, the backplane logic bypasses the empty slot, and the circuit remains closed. A further task of the backplane is the automatic configuration of the drive and also to support the hot-plug function, which is the exchanging of a drive during operation. The same principle is also used by Fibre Channel hubs. In a Fibre Channel loop, a device failure or cable defect will break the circuit, blocking the entire bus, so the hub bypasses any port that is not in use or blocked by interference. Consequently, the flow of data to other devices is not interrupted and the bus continues to operate normally. FC-AL products have been available since 1996, but they are currently only used in some low cost products. Nearly all products support Fibre Channel fabrics in addition to FC-AL.

Fabric

A fabric allows dynamic linking between nodes via the ports connected to the network. Please keep in mind that this application of the term "Fabric" may also be used as a synonym for the terms "Switch" or "Router".

Each port in a node, a so-called N_Port or NL_Port, is connected to the fabric by means of a link. Each port in a fabric is called an F_Port. Each node can communicate with every other F_Port connected to the same fabric with the help of the fabric network service (peer-to-peer principle). With this kind of topology the fabric, not the ports, routes and determines the transport route for the individual FC frames to be transferred. Another connection element in fabrics are E_Ports (extension ports), through which several nodes (usually FC switches) can be connected to each other directly, in order to maximise available bandwidths or gain higher availability in the network.

Fabric topology



The fabric and Arbitrated Loop topologies can be combined with each other in order to add a variety of service grades and performance rates to the node. Other networks such as SONET, ATM or IP (also known as FCIP or FC-over-IP) can be used in a fabric between individual fabric elements, in order to bridge physical distances between nodes that are too great to be covered by the FC fibre optic cable connection between N_Ports. These special connections can exist among fabric elements that are spread out over a larger geographical area and that are not directly connected to nodes.

World Wide Name and FC Address

The function of a fabric can be compared to that of a telephone system. We dial a number and the telephone system finds the path to the requested target. If a switch or link crashes, the telephone company routes the calls via other paths, which the caller rarely notices. Most of us are unaware of the intermediate links which the telephone company employs in order to make our simple phone call a success. TCP/IP uses the same basic idea, which is demonstrated again in the above-mentioned double function of the Fibre Channel as an I/O as well as network standard.

Fibre channel needs participant recognition, just as with a telephone system the caller identifies the destination with separate components such as country code and area code.

FC nodes and ports are globally clearly identifiable by their World Wide Name (WWN or WWP). The WWN/WWP is usually written as a hexadecimal and has a length of 64 bits. WWN/WWP are device addresses, comparable to MAC addresses, and make it possible to identify the nodes/ports themselves.

To address data packets in an FC environment a Fibre Channel address or PORT_ID of 24 bits is used, which leads to a total number of more than 16 million connectable ports. This address space is available only in real fabric topologies with the respective protocols.

Bit 23-16	Bit 15-8	Bit 7-0
Domain address	Area address	Loop address

In the FC Arbitrated Loop topology only the lowest byte is used by the FC-AL protocol for addressing. There are only 126 addresses - the loop IDs - available. This loop address is called AL_PA (Arbitrated Loop Physical Address). In order to differentiate between source and destination, they are referred to as AL_PD (destination) and AL_PS (source).

Whether the AL_PA is dynamically or statically assigned can be determined individually for most devices. Dynamic assignment of the AL_PA is not recommended for operating systems and applications that access storage devices via device paths. In this case, a different AL_PA will be assigned through a new process, the device paths will change, and the device can no longer be addressed. In the worst case a wrong device might even be used.

6.2.3 Cabling of Fibre Channel Networks

The two basic possibilities are electric data transmission via copper wires and optical transmission via glass fibre cables, although the optical cabling is used almost exclusively in FC devices such as switches and host bus adapters.

Copper wires

There are versions for 1 Gbit and 2 Gbit with plug formats DB9, HSSDC and HSSDC-2. The maximum cable length is 20 meters. There are special "equalised"
cables for 1 Gbit/s with a length of up to 40 meters.

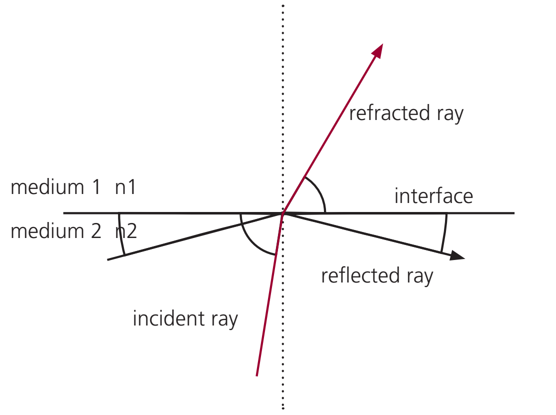
Fibre optic cable/glass fibre cable

Fibre optic cables (FOC) or glass fibre cables use total reflection of light waves from an optically tighter to an optically thinner medium as the fundamental physical principle behind data transmission. The signal transfer takes place via light pulses from a laser or LED, which are transmitted within the glass fibre.

The FOC with the simplest construction consist of a concentric optical core with a high refractive index covered by an optical cladding with a lower refractive index. Light that enters the FOC at a certain angle is conveyed continuously by total reflection at the border between core and cladding.

The advantages of FOC are high transmission rates into high Gbit ranges, the possibility to bridge large distances and resistance to interference.

Total reflection of light waves in optical fibres



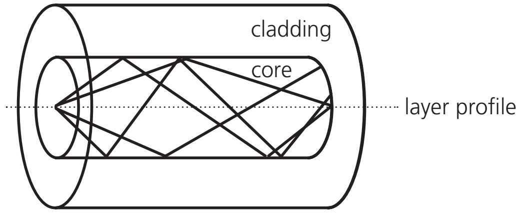
The signal quality, and thus the maximum transmission distance, is determined in principle by the distortion of the signal during transmission. The following factors lead to a poorer output signal.

- **Attenuation:** The interaction of the light with the medium causes the signal level to get weaker over increasing distance. After a certain distance it is no longer possible to clearly distinguish between level 0 and 1.
- **Dispersion:** This term refers to the dispersion of the signal in the transmission direction. As a result of dispersion effects it becomes impossible to differentiate between neighbouring signal peaks. Chromatic dispersion is usually differentiated from modal dispersion.
  - **Chromatic dispersion:** Differences in run time due to wavelength. The propagation speed of the light depends upon the wavelength. Since the light sources used are never 100 percent monochromatic, signals will always be dispersed over long transmission distances by the difference in run time and due to the different wavelengths.
  - **Modal dispersion:** Shifts due to the different optical path lengths in the fibre.

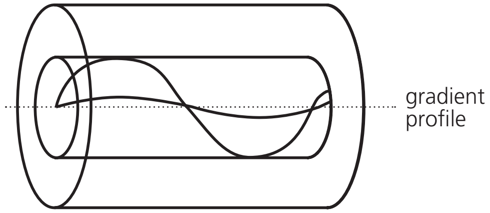
The light signal can enter the glass fibre at different angles. In geometric-optical proximity, light with a large angle to the optical axis is called high order mode. Light near the optical axis is accordingly called low order mode. Light of a high order mode must travel longer distances than a signal of a low order mode. This leads to differences in run time and thus to dispersion effects in the output signal. This effect can be countered with special glass fibres:

- **Gradient-profile fibres:** The refraction index decreases continuously from inside to outside, minimising signal dispersion. The term for this is graded-profile.

Step-profile fibre

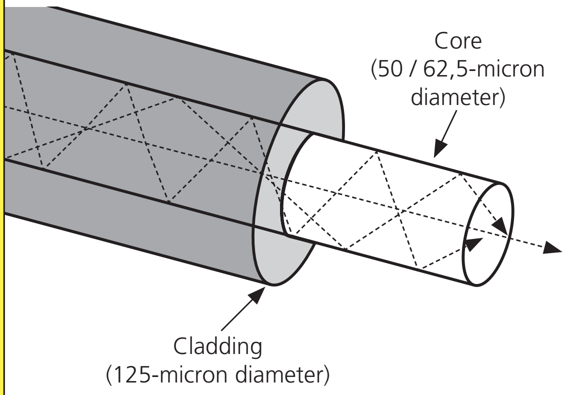


Gradient-profile fibre

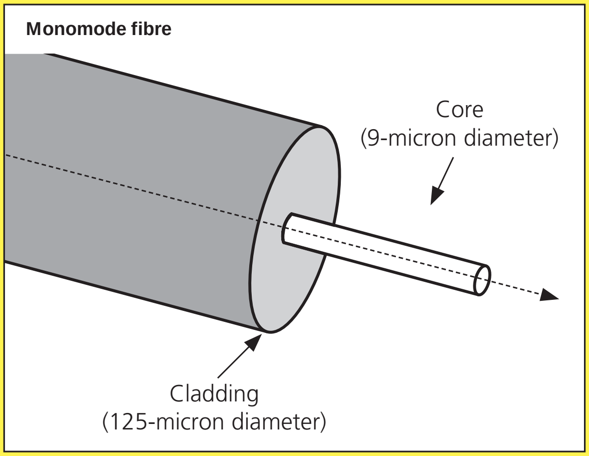


- **Monomode fibres:** In contrast to the multimode fibres they only conduct light pulses of a certain wavelength. Their selected core diameter is small enough to allow light propagation almost exclusively along the longitudinal axis.

Multimode fibre

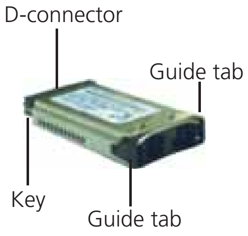


Monomode fibre



SFP in duplex

Schematic diagram of a GBIC



In summary, the maximum cable length for FOC is dependent upon fibre diameter, transmission rate, wavelength of the light pulse used and the spectral quality of the light source. Lasers are used mainly in two different wavelength ranges: 770–860 nm (short wave) and 1270–1355 nm (long wave). Multimode/monomode indicates the way that the laser pulses are guided in the glass fibre core. Multimode means that many different modes are running simultaneously. This is because of the large diameter of the cable core. Monomode means that only one wave mode can run through the fibre at a time. Because of the glass fibre core's small diameter there is only one possible optical path. Multimode is used for short wave lasers, single mode for long wave lasers.

Signal conversion in the Fibre Channel

Nearly all devices now use so-called GBIC (Gigabit Interface Connector) or SFP (Small Form factor Plugable) for the translation of electric signals in the server and storage systems into optical or electric signals for Fibre Channels. These are a combination of transmitter and receiver, or a so-called transceiver. The GBICs originally had the connection possibilities of DB9, HSSDC, SC-longwave and SC-shortwave.
In the course of the interface miniaturisation, in order to achieve a higher packing density on the units, and to move to a transmission rate of 2 Gbit/s, the GBICs were replaced with SFPs, which have now become the standard. Versions with electric interfaces in the HSSDC-2 format are usually available, as are optical interfaces with LC-longwave or LC-shortwave connections. SFPs are transceivers that are inserted into the corresponding slot on the devices. Devices with SFP slots are thus very flexible in relation to the interface format. GBICs and SFPs are hot-swap capable, and can thus be installed and removed during running operation.

6.3 SMI-S – The New Standard for SAN Management

There are ANSI, IETF and ISP/IEC norms for the management of the lower levels in SAN. It is advantageous that Fibre Channel and SCSI have remained compatible with each other through their continued development. The applications affected by the migration from parallel SCSI bus to a serial FC link do not have to be re-programmed. Services such as housing monitoring (SAF-TE, SES) and error handling are also standardised for both technologies. What had been missing however, were API standards for the cooperation between storage systems, server and software in a heterogeneous network at the level above, regardless of whether the SAN was based on Fibre Channel or Ethernet and iSCSI-TCP/IP.

Today there are manufacturer-specific, proprietary management software packages for each individual SAN device (HBAs, switches, storage systems, tape systems). Management in a SAN that spans various manufacturers, such as McDATA's SANavigator, requires very time-consuming programming of the interfaces for every individual manufacturer-specific API that must be supported.

SMI-S in contrast, defines a general, modular interface for the management of storage networks. As an object-oriented specification it makes coherent criteria available in order to identify and classify objects in a SAN, as well as monitor real and virtual resources, and transfer information using a commonly available transport mechanism. An object can be a complete RAID or just a defect sector on a hard disk. The goal of SMI-S is to provide a web-based management interface for all devices in a SAN independent of the manufacturer.

The Storage Management Interface Specification (SMI-S) emerged from the IBM BlueFin initiative and was adopted as the new standard in April 2003 by the SNIA Distributed Management Task Force (DMTF). Based on the Common Information Model (CIM) it represents a fundamental expansion of WBEM technology (Web-based Enterprise Management), which is based on three open standards:

- DMTF's CIM
- XmlCIM – an XML code specification defined especially for CIM
- HTTP – as the information transport mechanism between applications and systems conform with CIM

The WBEM architecture was adapted in consideration of the special demands of storage operation. In the first generation SMI-S supports and standardises the following SAN management functions:

- Discovery – automatic identification and registration of devices (HBAs, switches, RAIDs, tape systems) in a SAN
- Monitoring – continuous checking of the state of the SAN fabric and the operation of each device in the SAN (so-called health monitoring)
- Device management – active control, configuration and re-configuration of devices in the SAN

SMI-S is clearly more effective than SNMP (Simple Network Management Protocol), which already supports several storage systems with a series of MIBs (Management Information Bases), in order to communicate with framework applications (IBM TSM, HP OpenView, CA UniCenter, etc.) via Ethernet. All parameters are actively requested individually from all components in the network with exact IP addresses and passwords under SNMP.

With a unified SMI-S form the necessary parameters are automatically requested from the existing components, and those to be added (called client) and made available to the storage management software (called provider). Clients can be monitored, configured and reconfigured. SMI-S also supports the simultaneous communication of a provider with several clients via so-called lock managers.

The transition of the products to SMI-S by the manufacturers is still in the early stages. By the end of 2005 over 80 percent of all SAN devices should be equipped with either a translation interface of the native API to SMI-S Standard (proxy model) or already have an SMI-S compatible API (native model). SMI-S will then doubtlessly simplify the management of products within a SAN, reduce training expenditures and replace SNMP as the management standard.

6.4 iSCSI – the IP-Based Storage Network

The internet SCSI protocol (iSCSI) defines a standard for the transfer of SCSI requests on the basis of a TCP/IP communication and thus represents an alternative to Fibre Channel for the set-up of storage networks. Just like Fibre Channel, iSCSI allows a block-based communication in a network.

SCSI and FC are US standards of the ANSI committees T10 (SCSI) and T11 (Fibre Channel), whereas all guidelines connected to the IP protocol family, and thus for iSCSI as well, are created by the IETF (Internet Engineering Task Force), which is an informal association of experts. Besides iSCSI they also include FCIP, iFCP, mFCP and iSNS.

You can find additional information at:

- www.ietf.org
- www.snia.org

The design guidelines for the IETF for iSCSI are the following:

- SCSI devices must be able to communicate with each other over an IP connection
- TCP should serve as transport protocol
- TCP connections should be used sparingly
- Ethernet and IP standards may not be modified for iSCSI
- The server systems in the IP network must be able to access several storage systems simultaneously
- Performance enhancements of SCSI devices in the IP network must be allowed by iSCSI specifications

The advantages of iSCSI are in the use of TCP/IP capable infrastructure components to create a storage area network (SAN). Administrators can use their experience with the management of IP LANs and WANs in the establishment of a SAN and therefore do not need to learn any other technologies such as FC. iSCSI SANs can also be connected with each other via broadband connections across locations, i.e. asynchronously replicated storage systems can be created. The disadvantage is in the additional protocol overhead in IP networks in comparison to FC storage networks. TCP/IP sends small data packets and basically assumes an unreliable transport path. There is no pre-existing, fixed transmission path, such as it exists for SCSI or Fibre Channel. Transferred packets can arrive on different paths, delayed, with transmission errors or not at all. The receiver has the task of putting the packets back into the correct order and requesting any missing information. An FC SAN on the other hand, transfers data in blocks over the shortest distance and keeps the correct sequence between server and storage systems. These fundamental differences demonstrate that iSCSI storage networks will not be comparable to FC SANs in the near future as regards the potential bandwidth and latency. iSCSI has gained an increasing number of users however, as a reasonably-priced, central storage network with medium performance and the option of connecting easily to other locations.

iSCSI – SCSI over IP

The SCSI protocol family differentiates between two SCSI device conditions: initiators and targets. These conditions must be considered separately from the underlying hardware. Every SCSI device can be the initiator or the target for SCSI communication. The initiator mostly gives commands, which are then executed and acknowledged by the target. As a rule, the SCSI controller (host bus adapter) takes on the role of the initiator, while SCSI end devices (hard disks, storage systems, tape libraries etc.) primarily act as targets.
Comparable to the Fibre Channel, in which SCSI commands are given a Fibre Channel header, included in FC frames and transferred serially via copper or fibre-optic cables, iSCSI also provides this kind of basis for communication via a TCP/IP infrastructure. iSCSI can thus be understood from the server side as an interface between the TCP/IP interface of a network card and the SCSI subsystems of the operating system.

>> *See schematic figure top of p. 127:*
Storage network technologies in comparison

An iSCSI SAN can be made up of multiple iSCSI initiators (usually servers) and some iSCSI targets, such as RAID systems or tape libraries, which is comparable to a standard FC SAN. Each peripheral which is part of the iSCSI SAN must have access to iSCSI protocol support via at least an Ethernet connection, in order to be connected with an infrastructure component (switch, router, etc.). Ideally, a Gbit Ethernet port exclusively assigned to the iSCSI communication should be available to connect to the iSCSI. Existing SCSI or FC devices
can be integrated in the iSCSI SAN via iSCSI gateways.
In order for an initiator to find a storage system it needs a list of the IP addresses of iSCSI devices which provide a target portal. These address can be manually entered, determined via broadcasts or requested via the Internet Storage Name Service Protocol (iSNS), which makes it easier to locate devices for iSCSI initiators.

Addressing and naming in the iSCSI protocol

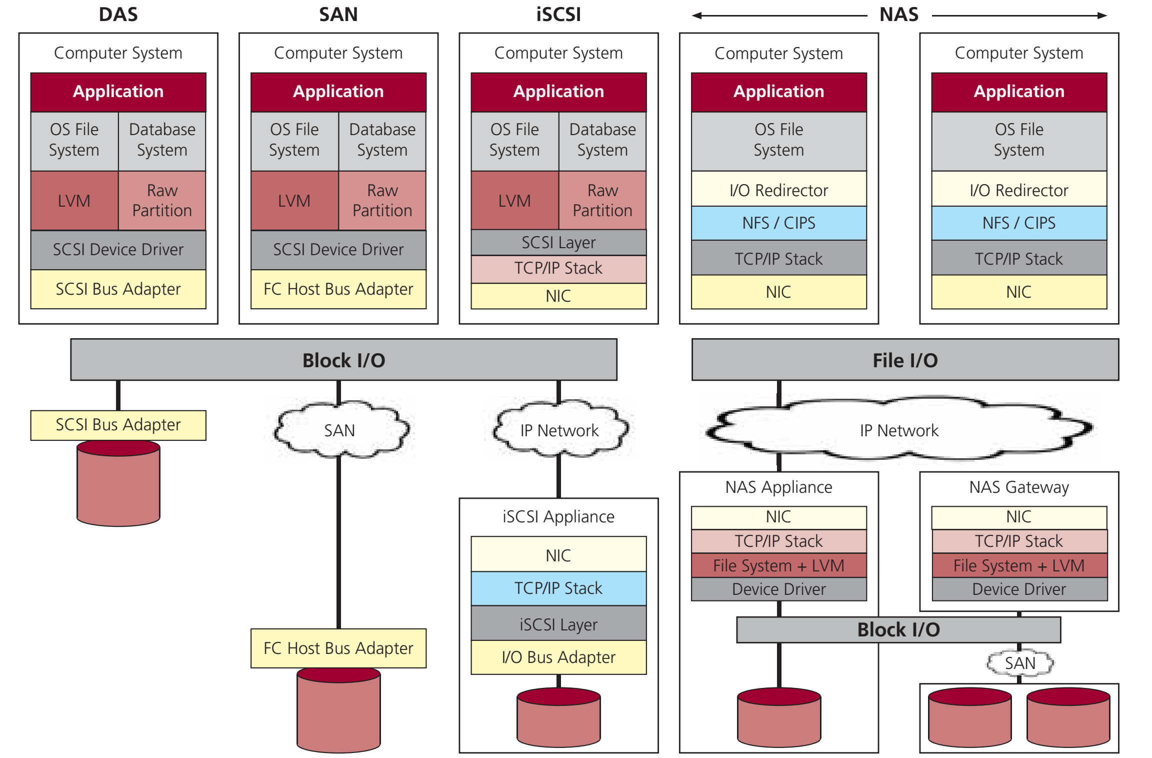
iSCSI uses TCP/IP for a reliable data transmission via a potentially unreliable network. The iSCSI layer receives the SCSI commands from operating system interfaces and packs them into TCP/IP compatible packets or frames for transport. The iSCSI protocol monitors the integrity and verification of the respective read/write operations requested. In practice, multiple accesses on the initiator and target sides must be processed.

>> *See schematic figure on bottom left of p.127:*
iSCSI protocol layer

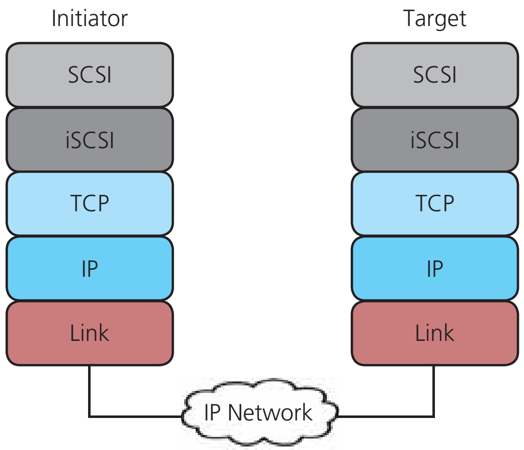
Initiators and targets have a network ID as participators in an IP network, which corresponds to the assigned IP address. As is shown in the next figure, a network identity can have one or more iSCSI nodes.

>> *See schematic figure on bottom right of p.127:*
Structure of an iSCSI network entity

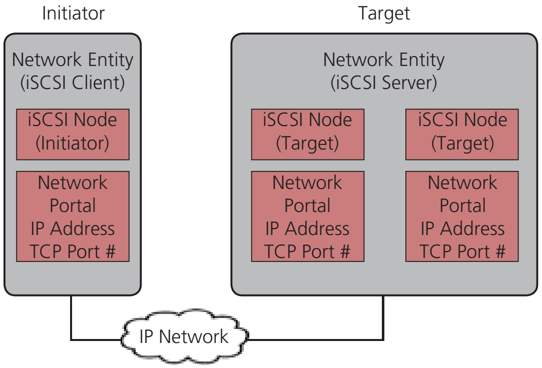
Storage network technologies in comparison



iSCSI protocol layer



Structure of an iSCSI network entity


Initiator
Target
Network Entity (iSCSI Client)
Network Entity (iSCSI Server)
iSCSI Node (Initiator)
iSCSI Node (Target)
Network Portal IP Address TCP Port #
IP Network

An iSCSI node identifies an SCSI device which is accessible via the network. With a RAID system, these iSCSI nodes are typically the logical drives (LUNs). Each iSCSI node is identified by a dedicated iSCSI name, which can be as long as 255 bytes.

The combination of IP address and TCP port creates a unique address for an iSCSI device within a network, whereas the 255 byte iSCSI name provides the user with a readable alpha-numerical ID. Separating the iSCSI name from the iSCSI address assures, for example, that a storage system will always have the same ID, regardless of its location. While the IP address changes with a relocation into a different network segment, the iSCSI name remains the same and allows the initiators to recognize the device quickly. Thus, a redundant RAID system with multiple network paths to the server can still be identified as a single device by its iSCSI name. An iSCSI name consists of three parts: a type ID, the identifying position and a unique ID determined by the identifying position.

In Fibre Channel the 64-bit long, hexadecimal World Wide Name (WWN) provides for the global identification of a device, whereas the 24-bit long Fibre Channel address serves as a network ID. For the simple administration of heterogeneous FC and IP networks, a WWN can be used as an iSCSI name. The type identification is then the "eui" (IEEE EUI) format, for "extended unique identifier".

In addition to the iSCSI name, the iSCSI protocol supports another supplementary alias. This can be used by IT administrators if the iSCSI name was predefined by the manufacturer or another source and thus has little or no relevance. The alias can also have up to 255 bytes and
be used for login, for example. The iSCSI protocol causes management software to display the aliases to the administrator either via a Command Line Interface (CLI) or a GUI.

iSCSI session management

An iSCSI session between an initiator and a target must be authenticated through an iSCSI login, just like the Fibre Channel port login (PLOGI). The iSCSI specification supports the following authentication methods:

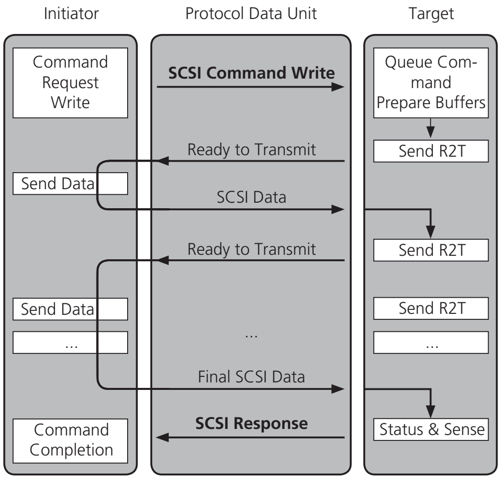
- KRB5 – Kerberos V5
- SPKM1 & SPKM2 – Simple Public-Key Generic Security Service (GSS) API
- SRP – Secure Remote Password
- CHAP – Challenge Handshake Authentication Protocol

None – no authenticationDuring the login iSCSI names, aliases and variable parameters, the selection of security protocols to be used, the number of supported parallel TCP connections, timeout values or the maximum data payload size supported are negotiated between devices. If there are different values, the largest common value is used for the iSCSI session. The parameters are exchanged as text field entries:

Originator sends <key> = <value>
Responder replies <key> = <value> | None | Reject | NotUnderstood | Irrelevant

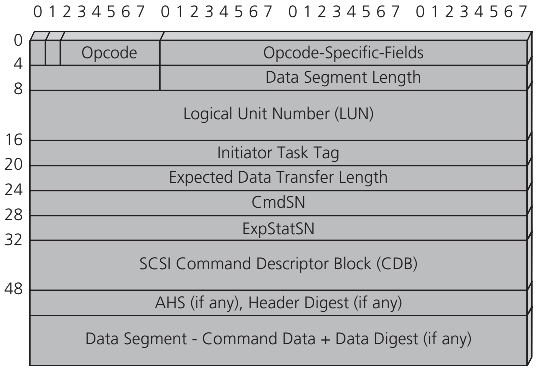
As soon as the login process has been completed, SCSI commands can be transferred between initiator and target. If multiple TCP connections were made for one session, each command/result pair must be transferred via the same connection. This is known as connection allegiance and assures that reading and writing operations can be executed without additional overhead from monitoring several TCP connections.

Example of an iSCSI write operation



Processing an SCSI command between iSCSI initiator and target includes the transfer of several subordinate commands, status and data packets, as is shown in the figure above. They are called iSCSI Protocol Data Units (PDUs).

Structure of an iSCSI Protocol Data Unit (PDU)



0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7
0
4
8
16
20
24
28
32
48
Opcode
Opcode-Specific-Fields
Data Segment Length
Logical Unit Number (LUN)
Initiator Task Tag
Expected Data Transfer Length
CmdSN
ExpStatSN
SCSI Command Descriptor Block (CDB)
AHS (if any), Header Digest (if any)
Data Segment - Command Data + Data Digest (if any)

The first PDU in our example transmits an SCSI command to write certain blocks from the initiator to the target. Since TCP/IP was developed for the transmission of continuous byte streams, there is no mechanism for the recognition of the block limits of the SCSI command in the byte stream. iSCSI uses a field in the PDU header instead, which contains the command block length. And whereas a byte stream in LAN operation can have an endless length, a maximum length of 232-1 is defined for iSCSI. After receiving the SCSI command the iSCSI target affirms its readiness to receive data together with the maximum buffer size available with a "ready to transmit" R2T PDU. The speed of the data flow is actively controlled by the iSCSI target device, in contrast to normal TCP/IP operation.

If an initiator does not receive an answer from the target to waiting commands, it can request or check the target status. This iSCSI-NOP-Out command with set P-bit (Ping) can also include test data that the target is to return intact. If the target does not answer or the returned test data is corrupt or incomplete, the initiator can close the connection and start a new session.

While connections between servers and storage systems are usually strictly defined in an FC SAN, and can only be interrupted for maintenance or booting the server, in an iSCSI SAN more or fewer TCP connections may be necessary, depending upon the load created by the initiator. An initiator can therefore independently establish additional connections to the target or close individual TCP connections. A complete logout is usually only necessary in case of connection errors or maintenance on the servers or storage devices.

iSCSI error handling

The traditional SCSI architecture assumes a mostly disturbance-free environment. SCSI devices directly connected use a dedicated, parallel bus to transfer data, quite separated from any network disturbances. An iSCSI network on the other hand, can be established via an error-prone WAN and Internet connection as well as a stable, local gigabit Ethernet. This is why the iSCSI specifications contain adaption and error handling routines for various possible problems. In order to provide reliable error handling, the initiator and target must have a buffer to store commands and answers until confirmed by the other side.

An individual PDU can have missing or inconsistent data within the iSCSI frames. This is called a format error and causes the iSCSI PDU to be rejected by the receiver. It returns a Reject PDU to the sender, which indicates where the first corrupt byte was discovered. Another category of iSCSI errors are corrupt data in the data payload (data digest error) or header (header digest error). The affected PDU is also rejected in this case and a Reject PDU is sent.

The error correction mechanisms in an iSCSI network provide detection and resending for corrupt or missing iSCSI frames as well as testing, termination and restarts for inactive or interrupted TCP connections. Finally, the iSCSI protocol provides for the termination of a complete session with the cancellation of all tasks and subsequent new login, if other attempts to correct an error have been unsuccessful.

iSCSI performance

In our experience the best achievable transfer rates in a 1 Gbit Ethernet iSCSI network are between 30 and 60 MB per second. There are several factors to consider, however.

Video-streaming or similar constant data traffic will not work with iSCSI, since the quality of the data transfer cannot be guaranteed as compared to an SCSI bus or FC network. The direct communication between two iSCSI targets, e.g. between a hard disk system and a tape library, for serverless backup is not possible either.

One of the most important and frequently mentioned factors is the expected CPU load. The packing and unpacking of SCSI commands in TCP/IP packets limits not only the performance of the servers connected in an iSCSI network, but also that of the storage systems. Several suppliers therefore offer TCP/IP Offload Engines (TOE) or special iSCSI host bus adapters instead of a normal gigabit Ethernet network card, which execute these operations on their own CPUs and thus reduce the load on the central system processor. The costs for this are close to those for Fibre

Channel host adapters and reduce the cost advantages for iSCSI compared to FC significantly. The alternative is to use common, inexpensive Ethernet network cards together with so-called iSCSI software initiators (Microsoft, Cisco etc.) or special iSCSI target software for storage systems.

Fujitsu Siemens established exact figures for the CPU load: A common server with Intel® architecture needs around 5000 CPU cycles for a cycle through an SCSI driver, while at least 50,000 cycles are necessary for a TCP/IP stack. Not to forget the iSCSI stack overhead. LAN experts generally use the following rule of thumb: The transfer of 1 bit requires 1 Hz clock frequency from the processor. A TCP/IP connection with 1 Gbit/s therefore uses the full capacity of a 1 GHz processor. Since many external RAID systems use even weaker CPUs, software controlled iSCSI storage systems can suffer from performance drops with large data loads. With current single processor servers the load from iSCSI is thus approx. 30%, with a dual processor system it is only 15%. Whether this is tolerable for the user must be decided case by case. The availability of inexpensive TOE cards or iSCSI host bus adapters will become increasingly important with the expected expansion of 10 Gbit networks, however.

7. Magnetic Tape Storage

7.2.4 SAIT Drives

Milestones in Sony's Development (S)AIT tape technology and tape drives

1996 – Sony introduces Advanced Intelligent Tape (AIT) technology

1996 – AIT technology using AME tape introduced

1996 – AIT-1 wins "Best New Technology Award" from Byte Magazine

1997 – AIT-1 approved as standard by European Computer Manufacturers Association (ECMA)

1998 – AIT-2 technology announced

1998 – AIT-1 enhanced from 25 GB to 35 GB

1999 – AIT-2 begins customer shipments

1999 – AIT Forum created in Denver

1999 – AIT prototype demonstrates more than 1 MB/in2 areal density

1999 – AIT-2 format approved as an industry standard by ECMA

2000 – AIT-2 won "Storage and Peripherals Award of the Year" by Imaging & Document Solutions Magazine

2000 – AIT-3 technology announced

2000 – AIT-1 value series introduced and data transfer rate enhanced to 4 MB per second

2000 – AIT prototype demonstrates 6.5 Gbit/in2 areal density

2001 – Sony introduces AIT WORM (Write Once, Read Many) drives and media

2001 – Sony delivers AIT-3 drive and storage media to the market

2001 – The Information Storage Industry Consortium (INSIC) approves R-MIC specification (Remote Memory In Cassette)

2001 – Sony introduces SAIT technology

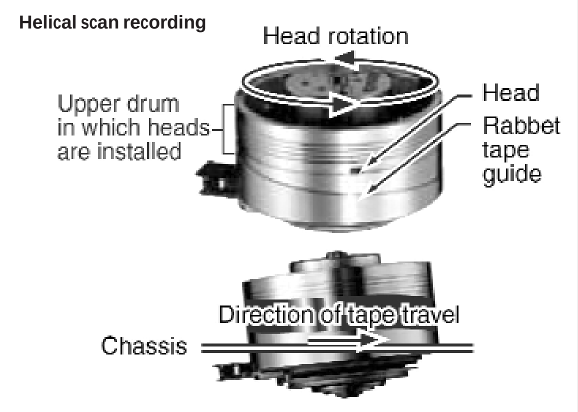
2002 – Sony breaks areal density record by demonstrating: 11.5 Gbit/in2

2002 – The first models of SAIT-1 drives and media are shipped to OEMs

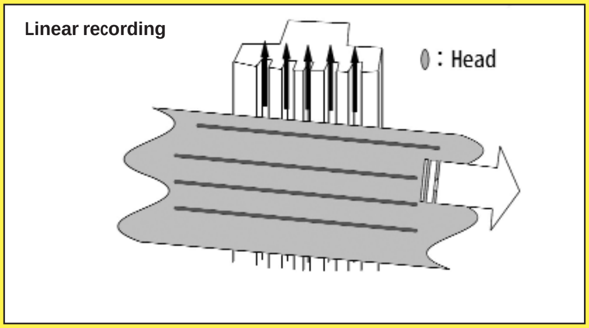
2003 – Production SAIT-1 drives and media start shipping to OEMs

Helical scan recording

The core technology that enables SAIT to offer unsurpassed capacity is the helical scanning system. The read-write head rotates at a specific angle, while the tape is wrapped around it, achieving up to two times the density compared to linear recording. The benefits of helical scan recording go beyond higher density recording, however. Benefits also include reduced wear-and-tear and improved overall drive and media durability.



Helical scan recording



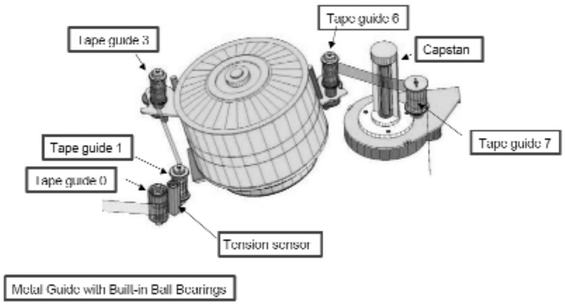
Linear recording

Phase servo capstan

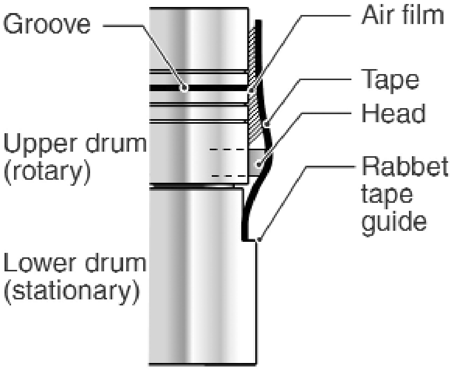
The data written on the SAIT tape forms helical tracks, and is read out from the tape by tracking along the helical tracks. Tape speed must remain constant during recording to record accurate track width. Micron-level tape tracking is reliably achieved using the phase servo motor of the capstan. Error signals precision-control the capstan speed allowing it to accurately position the tracks on the head.

This system provides high data integrity, ensuring that data is readable after backup. It also enables upgrading to a higher capacity.

Diagram of tape threading, metal guide with ball bearings



Schematic figure of an air film



Reduced tape load

SAIT utilises a rotational metal guide with built-in ball bearings to support a stress free environment for the tape and drive mechanism, reducing tape load.

Tape tension

Variance in tape tension can cause read-write errors and wear-and tear to the media and hardware. SAIT incorporates a tension sensor that provides feedback on system conditions to realise stability and optimise stress free tension. By permanently controlling the tape tension, overall reliability is drastically increased.

Tape tension comparison

	SAIT	LTO2	SDLT600
Grams	10	100	100

Optimal air film

Adding a thin, carefully calculated groove to the drum enables a constant micron-depth film of air to be created and a structure whereby the tape is evenly attached to the drum in spite of the low tension, reducing potential damage to the tape and increasing overall durability of the SAIT.

Low tape speed

Friction caused by the movement of the tape causes heat, which can lead to the destruction of tapes in extreme cases. In SAIT, the upper part of the drum containing the head is rotated steadily at high speed in order to maintain a sufficient head-to-tape relative speed, while the tape runs at low speed in low-friction contact with the head. In addition to minimising friction, this also results in narrow track pitch, enabling high-density recording.

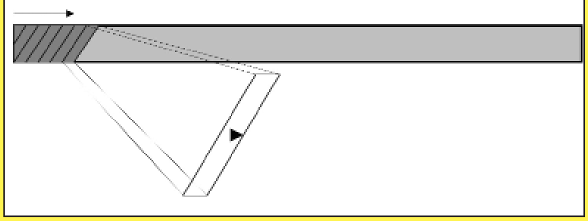
Tape speed comparison (inches per second)

	SAIT	LTO2	SDLT600
Read/write	1	217-232	108

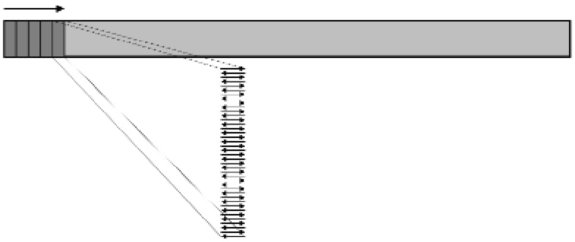
Single pass recording

Unlike linear serpentine recording technologies, which record up to 56 tracks in a back-and-forth motion before filling up the tape, helical scanning records the SAIT tape in one pass. This not only reduces the wear-and-tear on the drive and media, but also improves overall performance by eliminating the "shoe-shining" effect.

One pass recording in helical scan recording



Linear recording



Passes required to fill tape

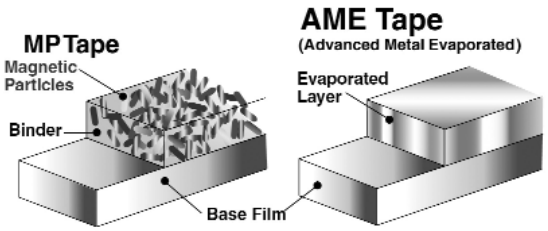
	SAIT	LTO2	SDLT600
Passes	1	64	40
Tracks	133 per group	512	640
Channels	8	8	16

Advanced Metal Evaporated (AME) media

Metal Particle (MP) media, commonly used by competing technologies, uses binder polymer to adhere the magnetic material to the base film of the media. There are several disadvantages to the use of this binder material: Binder degradation occurs over time, usually from humidity and temperature changes. The shed particles build up on the head, resulting in the need for frequent cleaning. Over prolonged periods of time, oxide shedding, a reaction referred to as "sticky shed syndrome" can occur, where the surface of the media becomes softer than normal and can create a gummy residue. This can cause increased friction, instant head clogging, tape seizing, and even damage to the hardware and/or media.

SAIT utilises Advanced Metal Evaporated (AME) technology. AME has already performed very well in Exabyte Mammoth and Sony AIT technologies. Pure cobalt is used as the magnetic medium in AME. It is vacuum-evaporated onto the base film. As a result, AME media can achieve higher recording density that that of existing MP media. Deposits on the head are also substantially reduced as no binder polymer is used to secure the magnetic substance, thus eliminating sticky shed syndrome. To further enhance the durability of the AME media, a Diamond-Like Carbon (DLC) coating is applied to offer superior wear resistance.

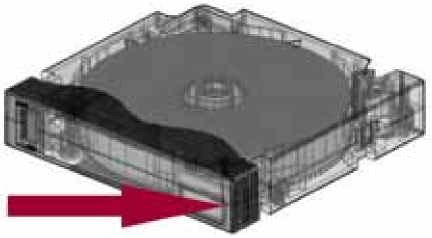
Comparison of MP tape and AME tape



Remote Sensing-Memory-in-Cassette (R-MIC)

SAIT and AIT media incorporate an 8-kByte solid-state memory chip (R-MIC) mounted on the data cartridge. The system log and data position information is stored on the R-MIC, significantly reducing load and file access time, improving performance. In addition, because the tape itself does not have to be accessed for this information, there is less wear-and-tear, prolonging the life of both the media and hardware.

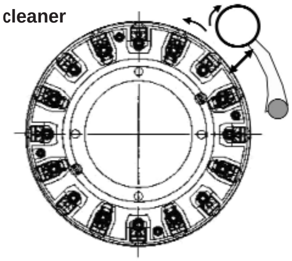
R-MIC



Built-in head cleaner

SAIT drives incorporate an internal head cleaner to help prevent head contamination, which can obstruct the reading or writing of data and cause wear on the head and media. The automatic head cleaner dramatically reduces head contamination, improving data integrity, and increasing overall drive and media reliability.

Function of a head cleaner



Error correction

SAIT has an exceptionally reliable error correction system. Capacity loss per track is smaller than with existing rewrite functions, since rewriting is possible at different heights per block, producing extraordinarily accurate re-recording levels. Furthermore, burst errors of up to 16% of the overall track can be corrected and reproduced normally using the triple-step error correction code. And by using 3-level error correction codes, data reliability is significantly improved.

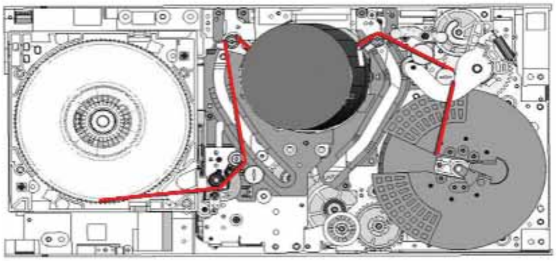
ECC – Error Correction Code

	SAIT	LTO-2	SDLT600
ECC levels	3	2	Not published

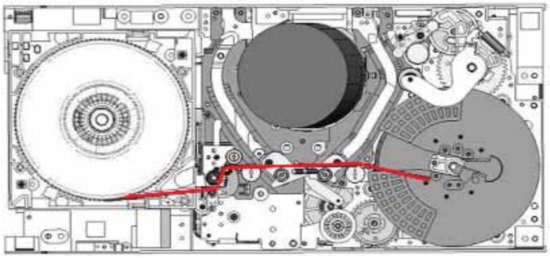
Dual mode tape path

The SAIT mechanism utilises a dual-mode tape path delivering smooth operation and minimal tape friction. Only during read, write and low-speed tape positioning operations is the tape wrapped around the drum. During high-speed searches the SAIT tape path is kept very short and simple, and utilises the contents of the R-MIC to provide reliable positioning information.

Tape path



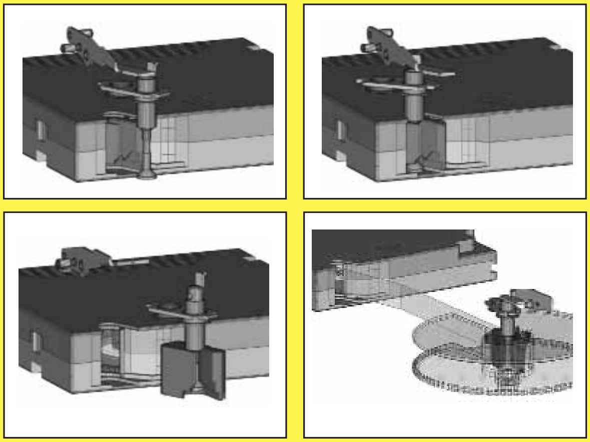
Tape path when searching



Simple loading mechanism

The automated process of loading single-reel tape cartridges into a drive has not always been dependable. Sometimes the drive might not catch the leader, the leader may become dislodged, or the leader may break altogether.

Loading mechanism



SAIT's leader block chucking system enables reliable loading of the tape, and eliminates the possibility of faulty chucking. Three non-contact optical sensors monitor the leader chucking operation. If mis-chucking occurs, the sensors will signal the drive to automatically retry.

Optical sensors

Non-contact optical sensors have been adopted to monitor the operation of the tape and drive, such as the leader block chucking operation. This further enhances accuracy. And by eliminating contact with the mechanism, reduced wear-and-tear is achieved for the mechanism and the sensors, improving overall durability.

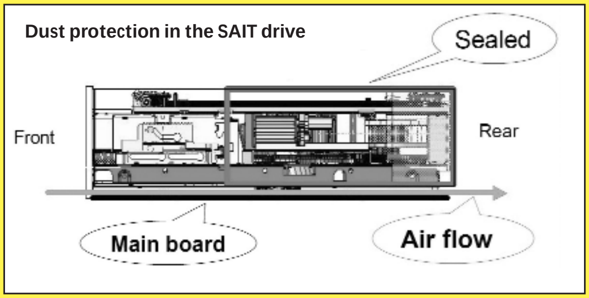
Dust resistant design

Dust and other airborne contaminants can be detrimental to both tape drives and media, causing operational problems, including read-write errors, and could potentially result in drive and/or media failures. SAIT incorporates several design features to reduce dust and other airborne contaminants from entering the mechanism.

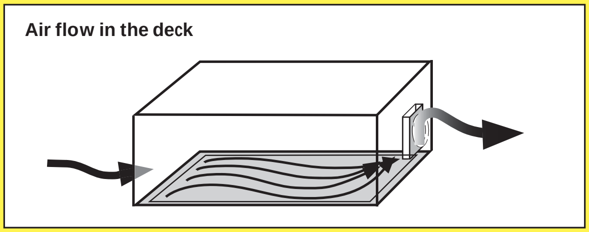
The bezel protects the drive from dust contamination. The bezel acts as a barrier, reducing dust intake through the media slot. SAIT drives also have an internal shutter to offer increased protection from dust particles within the drive mechanism.

Sealed deck

The entire unit has a tightly enclosed sealed structure in which the heat-generating circuit boards are kept separate from the mechanical workings. Air flows under the deck to provide effective cooling of the electronics. In addition to providing superior durability, it also offers effective dust protection: The device is virtually dust-proof.



Dust protection in the SAIT drive



Air flow in the deck

Conclusions

Sony's SAIT tape drive technology, in addition to the existing market standards of LTO Ultrium and SuperDLT, is a valuable addition to the market and reflects Sony's years of experience in research and development for magnetic recording. SAIT's technological advancements, including AME tapes, R-MIC, 3-level ECC, leader block chucking system, and the dust proof design, all add up to a highly reliable tape drive, ideal for use in demanding storage environments.

Overall reliability specifications

	SAIT	LTO-2	SDLT600
MTBF	500,000 hrs.	250,000 hrs.	250,000 hrs.
Duty Cycle	100%	100%	100%
Average Head life	50,000 hrs. At 25°C	60,000 hrs.	50,000 hrs.

>> Chapters 8 to 24 can be found in the Internet