Storage-Area-Network an der Humboldt-Universität

Das Rechenzentrum der Humboldt-Universität bietet seit acht Jahren im Rahmen des universitätsweiten File-Service die Verteilung und Pflege von Plattenplatz in den Instituten sowie einen Backup-Service. Wachsender Bedarf an Plattenplatz, immer schneller wachsende Plattenkapazitäten und abnehmende Toleranz gegenüber Stillstandszeiten stellen den File-Service vor qualitativ neuartige Aufgaben. In diesem Papier wird gezeigt, wie das RZ mit Hilfe des Konzeptes des Storage-Area-Network *(SAN) diese Aufgaben zu lösen versucht.*

Thesen zum File-Service

Um den Gegenstand dieses Artikels zu veranschaulichen, seien an dieser Stelle Thesen vorangestellt, wie es nach Meinung des Autors um den File-Service heute bestellt ist.

- Es gibt einen Bedarf an einem zentralisierten File-Service für
  - Home-Directorys für Studierende, genutzt im Pool-Betrieb,
  - Austausch von Daten in Arbeitsgruppen,
  - zentralisierte Software-Pflege,
  - effizienteres Speichermanagement durch weniger Personal für Planung, Aufbau, Erweiterung und Pflege.
- Festplatten werden immer größer (derzeit 180 GByte), und gleichzeitig sinkt der Preis pro GByte deutlich.
- Durch dieses günstige Preis-Leistungs-Verhältnis können sich selbst kleine Arbeitsgruppen deutlich mehr als 100 GByte Festplattenkapazität leisten.
- Jede Festplatte geht irgendwann kaputt. Es muss mit Datenverlust auch dann gerechnet werden, wenn vermeintlich sichere Festplattensysteme wie RAID-5 oder Mirrors eingesetzt werden. Eine Begründung findet sich in
  - unsicherer Hard-, Firm- und Software,
  - Desastern wie Wasser, Feuer, Überspannungsspitzen (Blitz), Überhitzung durch Ausfall der Klimaanlage,
  - mangelnder Managementkompetenz vor Ort.
- Ein Restore von mehr als 100 GByte Daten dauert mehr als einen Tag, mehr als 1 TByte vermutlich mehr als eine Woche – eine Belastung für den Backup-Service, das LAN und die Nutzer.
- Standard sind derzeit File-Server mit lokal angeschlossenen SCSI-Disks, die nicht nur anfällig für Datenverlust sind, sondern deren Reorganisation zudem umständlich, zeitaufwendig und meist mit Stillstandszeiten verbunden ist.
- Nutzer hassen Stillstandszeiten. Die Erwartung an die Verfügbarkeit des File-Service geht in Richtung 100 %.

Allgemeine Anforderungen

Zentralisierte File-Server haben nur dann eine Zukunft, wenn sie eine Verfügbarkeit von mehr als 99,9 % besitzen, d. h. ihnen wird maximal acht Stunden Down Time pro Jahr für Wartung und für Unvorhergesehenes eingeräumt. Außerdem müssen sie über eine hohe Flexibilität verfügen, um mit den Bedürfnissen der Nutzer wachsen zu können. Für den Administrator sind noch solche Eigenschaften wie übersichtliche und einheitliche Verwaltbarkeit und Funktionssicherheit wichtig. Für Verfügbarkeit und Flexibilität bedeutet das konkret:

- Verfügbarkeit durch Redundanz (kein Single Point of Failure)
  - Mehrfach vorhandene Server (räumlich getrennt) können gegenseitig ihre Plattensysteme austauschen.
  - Redundante Datenpfade führen über doppelt vorhandene Host-Bus-Adapter auf duale RAID-Controller (gegenseitig ersetzbar) in RAID-5-Plattenstrecken.
  - Daten müssen räumlich getrennt gespiegelt werden.
  - Netzteile und Lüfter sollten redundant ausgelegt sein.
  - Alle Storage-Komponenten benötigen eine USV-Unterstützung.
  - Platten, Netzteile und Lüfter müssen im laufenden Betrieb austauschbar sein (Hot Swap).
  - RAID-5-Plattenstrecken sollten sich bei Plattenausfall automatisch und im laufenden Betrieb reorganisieren (Hot Spare).
- Flexibilität durch dynamisch konfigurierbare Betriebssysteme (Konfiguration ohne *Reboot*)
  - Server sollten in der Lage sein, physische Platten und logische Plattenpools dynamisch zu importieren und zu exportieren (inklusive Zwangsimport von „gestorbenen“ Servern).
  - Ein dynamisches Volume-Management und die darauf liegenden File-Systeme sollten eine Vergrößerung der File-Systeme zulassen.
  - Die Server müssen von sich aus in der Lage sein, Volumes auf Anforderung zu spiegeln.

Mögliche Lösungen

Um den heutigen Anforderungen an den File-Service gerecht zu werden, bietet der Markt im Wesentlichen drei Konzepte, die hier kurz dargestellt und bewertet werden.

Übersicht

NAS (Network Attached Storage)
NAS ist ein auf File-Server basierendes Konzept, in dem spezialisierte Standalone-Geräte ihre Daten via NFS und CIFS im LAN verteilen. Diese Server sind sehr einfach zu verwalten, besitzen durch interne Redundanz eine hohe Verfügbarkeit und in der Regel lokal (SCSI, FC-AL) angebundene Plattensysteme.

Zur Desastervermeidung ist eine Datenreplikation zu anderen NAS-Servern über das LAN möglich. Drei Hauptnachteile sind:

- Durch proprietäre Replikationsprotokolle können Daten nur zu NAS-Servern desselben Herstellers übertragen werden (Herstellerbindung).
- An einem ausgebauten NAS-Server können nur noch schwer Änderungen vorgenommen werden (Disk oder Volumes reorganisieren oder hinzufügen).
- Sie sind teuer.

SAN (Storage Area Network)
SAN ist ein Konzept, über ein zweites, vom LAN unabhängiges Netzwerk Storage-Nodes (Disk-Systeme, Tape-Librarys) mit SAN-Clients (: LAN-Server, Hosts) zu verbinden. Das Transportprotokoll im SAN ist SCSI-3. Den Servern werden von den Storage-Nodes Festplatten und Tape-Geräte als SCSI-Geräte zugeordnet. Ohne Unterbrechung des Server-Betriebes ist es möglich, Datenbereiche im SAN hinzuzufügen, herauszunehmen, zu verschieben und zu vergrößern. Durch Switch-Technologie werden hohe Bandbreiten erzielt. Positiv ist die Möglichkeit, die SAN-Endpunkte (Storage-Server, SAN-Clients) praktisch vollständig herstellerunabhängig in das SAN einzufügen. Da derzeit noch an den Standards der Switch-Protokolle gearbeitet wird, erzwingt die Beschaffung größte Aufmerksamkeit bei der Auswahl der Switch-Technik. Das SAN-Konzept ist ebenfalls teuer.

iSCSI (SCSI over IP)
Ähnlich wie beim SAN werden beim iSCSI-Konzept Storage-Einheiten über ein Netzwerk verbunden, nur dass dafür das klassische LAN und IP-Protokoll benutzt wird. Dabei werden SCSI-Blöcke in TCP/IP-Pakete gekapselt und so zwischen Storage-Nodes und den Host-Systemen transportiert. iSCSI ist ähnlich flexibel wie SAN und bietet außerdem die Gelegenheit, die vorhandene LAN-Infrastruktur kostengünstig zu nutzen. Derzeit sind aber noch folgende Nachteile zu verzeichnen:

- Durch TCP-Kapselung und IPsec (für sichere Verbindungen) erhält man eine sehr viel geringere Performance als im SAN/NAS.
- An den Standards wird noch gearbeitet.
- Es gibt dafür kaum Produkte.

Bewertung

Eine subjektive Bewertung ist der Tabelle 1 zu entnehmen.

Kriterium	**NAS**	**SAN**	**iSCSI**
Verwaltbarkeit	++	0	0
Interne Redundanz	++	+	+
Externe Redundanz	+	++	++
Flexibilität	0	++	++
Performance	+	++	--
LAN-Entlastung	-	++	--
Interoperabilität	--	+	+
Marktverfügbarkeit	++	++	-
Kosten	--	--	+

Tabelle 1: Bewertung NAS/SAN/iSCSI

Die Stärken von NAS liegen eindeutig in der einfachen Verwaltbarkeit und in der hohen internen Redundanz der Systeme. SAN und iSCSI haben, bezogen auf die Verwaltung durch ihren heterogenen Ansatz, naturgemäß ein Problem. Auch muss man selbst auf die interne Redundanz der Systeme bei SAN/iSCSI achten. Dagegen haben SAN/iSCSI eindeutig Vorteile bei der Herstellung von Duplikaten ihrer Datenbestände und im flexiblen und dynamischen Umgang mit Storage-Nodes in ihren jeweiligen Netzen. Dank des breitbandigen Netzwerkes hat SAN eindeutig in den Punkten Performance und LAN-Entlastung die Nase vorn. Aber dass NAS/iSCSI das LAN belasten, dürfte in naher Zukunft mit der Einführung von 10-Gbit-Ethernet kaum noch eine Rolle spielen.

Es gibt durch den proprietären Ansatz von NAS überhaupt keine Interoperabilität mit anderen Herstellern (in der Regel müssen sogar die Festplatten für den drei- bis vierfachen Marktpreis vom NAS-Anbieter bezogen werden). Das Verbindende in den Netzwerkkonzepten SAN/iSCSI ist das SCSI-Protokoll, das die Interoperabilität garantiert. Dort können sich Third-Party-Hersteller gut einbringen. Im Gegensatz zu iSCSI gibt es im SAN-Bereich schon eine Vielzahl von Firmen, die plattformübergreifend den Markt bedienen.

Zusammenfassend kann man feststellen:

- NAS ist einfach in der Handhabung, aber vom Ansatz her proprietär und inflexibel.
- SAN ist enorm flexibel und performant – leider nicht ganz billig.
- iSCSI ist genauso flexibel wie SAN, hat aber derzeit deutlich Nachteile in der Performanz und in der Marktverbreitung (mit einer Produktvielfalt ist erst mit weitestgehender Standardisierung, so ab 2003 zu rechnen).

Für die Ablösung des alten File-Service ist die Entscheidung zugunsten des SAN-Konzepts gefallen, das bestens geeignet ist, zukünftigen Herausforderungen an den File-Service gerecht zu werden. Es eröffnet erst-

malig die Möglichkeit, einen Non-Stop-File-Service zu installieren, der selbst im völligen Desasterfall die Arbeitsfähigkeit garantieren kann.

Einführung in die SAN-Technologie

SAN ist eine Technologie für Massenspeicher, in der bis 16 Mio. Server und Storage-Nodes durch ein Netzwerk verbunden werden können (Fabric). Es gibt auch die Möglichkeit der Punkt-zu-Punkt-Vernetzung und Loop-Topologien (FC-AL). Diese werden aufgrund ihrer Beschränkungen hier nicht weiter betrachtet. Wie jede Netzwerk-Technologie wird auch diese, wie in Abbildung 1 zu sehen, in Schichten spezifiziert. Interessant ist, dass als Übertragungsprotokolle alle möglichen Standards verwendet werden können. Reale Produkte implementieren jedoch standardmäßig nur SCSI und meistens IP. Die typischen Transferraten liegen derzeit bei 1 oder 2 Gbit/s.

Eigenschaften eines SAN:

- Das Transportprotokoll im SAN ist SCSI-3. Den Servern werden von den Storage-Nodes Festplatten als SCSI-Geräte zugeordnet.
- Das vorherrschende, physikalische Transportmedium sind Glasfasern (Fibre-Channel: FC). Kupferverdrahtung ist zwar möglich, wird aber wegen starker Längenbeschränkungen selten eingesetzt. Durch die FC-Verkabelung kann Plattenkapazität an jeden Ort gebracht werden (Disk aus der Dose).
- Hinzufügen und Herausnehmen von Storage-Nodes ist ohne Unterbrechung des Server-Betriebes möglich.
- Datenbereiche können im SAN ohne Unterbrechung des Server-Betriebes vergrößert oder verschoben werden.
- Es werden hohe Bandbreiten durch Switch-Technologie erzielt. Durch redundante Datenwege ist ein dynamischer Lastausgleich im SAN möglich, zugleich werden durch intelligentes Rerouting der Switches fehlerhafte Komponenten umgangen.
- Es existieren Möglichkeiten für Server-Clustering, Data-Sharing und Remote-Mirroring.
- Backups können im SAN abgewickelt werden. Damit wird das LAN entlastet.
- Durch ein vereinfachtes Management kann administratives Personal entlastet werden.
- Es gibt ausgefeilte Sicherheitskonzepte im SAN (LUN-Masking und Zoning).

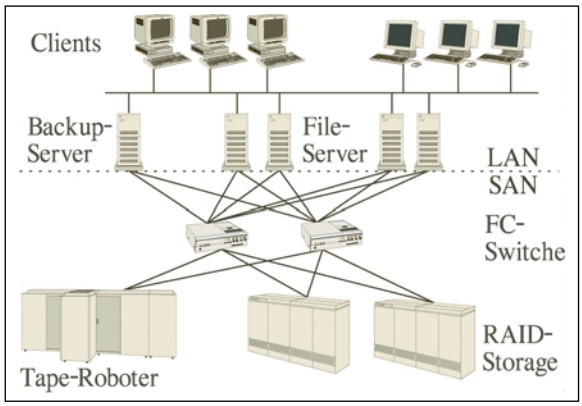

Abb. 2: FC-Fabric

Eine typische SAN-Vernetzung ist aus Abbildung 2 ersichtlich. Wie dem Bild entnommen werden kann, sind die meisten Komponenten im SAN redundant. Über zwei Host-Bus-Adapter (HBA – FC-Controller-Karten in den Servern) führen die Datenwege über getrennte Switche zu redundanten Disk-Arrays. In den Disk-Arrays werden Festplatten zu RAID-5-Strecken zusammengefasst, die im SAN als LUNs (virtuellen Festplatten) verkauft werden. Jeder Server besitzt einen Originaldatensatz in einer LUN eines Disk-Arrays und eine Kopie in einer zweiten LUN des zweiten Disk-Arrays. Fällt irgendeine Storage-Komponente (HBA, Switch, Disk-Array) in diesem Ensemble aus, so gibt es dazu immer ein Pendant. Fällt der Server selbst (auf Dauer) aus, dann kann ein Geschwister-System seinen Datenbestand zwangsimportieren und ihn im LAN verteilen. Zudem sind alle im Bild ersichtlichen Komponenten in getrennten Räumlichkeiten aufgehoben, besitzen redundante Stromversorgungen und Lüfter und USV-Unterstützung. Der Totalverlust eines Datenbestandes ist damit ziemlich unwahrscheinlich.

Eine typische Anwendung im SAN ist der LANfree Backup, wie er in Abbildung 3 veranschaulicht wird. In einem ersten Schritt verständigen sich der Backup-Server und der File-Server darüber, dass ein Backup fällig ist. Der Backup-Server und Server A machen daraufhin via SAN

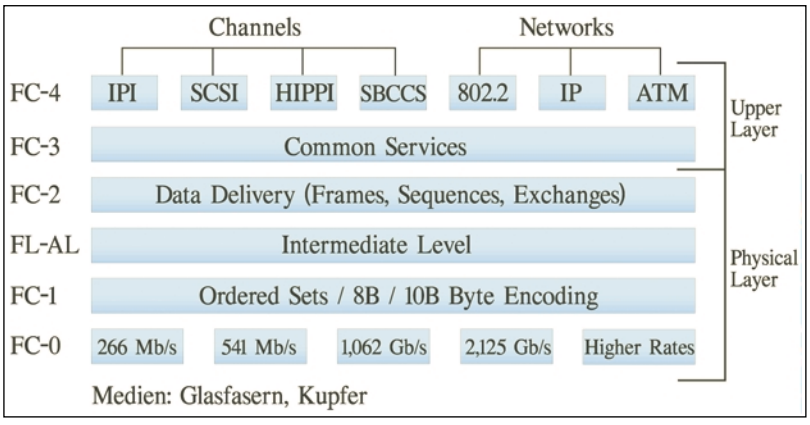

Abb.1: SAN-Schichtenmodell

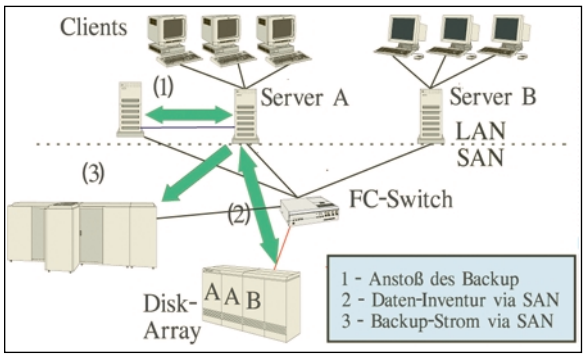

Abb. 3: LANfree Backup

eine Inventur der zu sichernden Daten im Disk-Array. Ist Server A damit fertig, fordert er vom Backup-Server ein Tape-Laufwerk im Tape-Roboter an. Nachdem ihm der Backup-Server dieses zugewiesen hat, schreibt der Server A selbständig den Backup aufs Tape. Der Datenweg führt vom Disk-Array über das SAN zum Server A und von dort über das SAN auf das Tape im Roboter – völlig ohne Beteiligung des LAN. Eine verbesserte Abwandlung des Verfahrens ist das Serverfree Backup. Dort werden vom Server A nur noch Steueranweisungen an das Disk-Array gegeben, das dann selbsttätig die Daten zum Tape schickt.

Als letztes Szenario veranschaulicht Abbildung 4 die Flexibilität des SAN-Konzeptes. Im August/September 2001 bezog die Chemie ihre neuen Räumlichkeiten auf dem Gelände in Adlershof. Ihre alten File-Server wurden zu diesem Zeitpunkt ausgemustert und durch neue, SAN-fähige Server ersetzt. Ihre Festplattenkapazität beziehen diese Systeme aus einem etwa 700 Meter entfernt stehenden Plattenturm im Johann von Neumann-Haus (JvN-Haus). Wenn im Frühjahr 2002 das IKA-Gebäude und mit ihm der neue Standort des Rechenzentrums funktionsbereit sind, dann sollen die Chemie-Server ihre Plattenkapazität in die dort bereitstehenden Festplattentürme verlagern. Zu dem Zweck werden ihnen zwei LUNs passender Größe aus dem RZ zugewiesen. Die Server erzeugen dann nacheinander zweite und dritte Kopien ihres Datenbestandes in diese LUNs. Wenn die Kopien aller Datenbestände synchronisiert sind, kann der Originaldatensatz (erste Kopie) im Turm des JvN-Hauses abgehängt werden. Die nun freien Festplatten im JvN-Turm können den Instituten für Informatik und Mathematik als Wachstumspotential dienen. Diese Verlagerung der Chemiedaten geschieht ohne Unterbrechung des Server-Betriebes. Alle Daten liegen dann abschließend gespiegelt in RAID-5-LUNs vor.

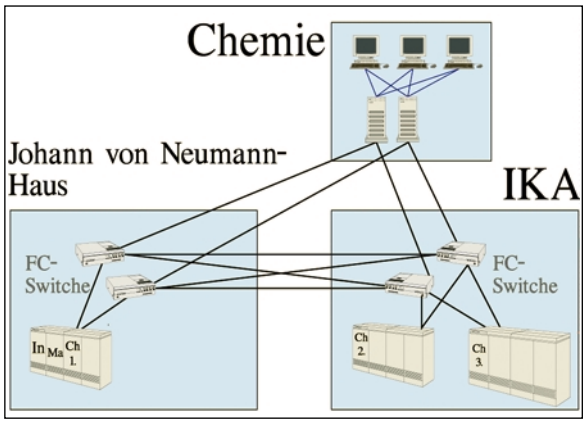

Abb. 4: Datenbewegung im SAN

SAN an der Humboldt-Universität

Als es 1999 darum ging, ein Nachfolgesystem für den seit 1993 arbeitenden dezentralen File-Service auszuwählen, fiel die Entscheidung zugunsten eines SAN-Konzeptes. Seitdem sind in den Standorten Adlershof und Mitte je drei RS/6000-Server mit je 1,5 TByte angeschlossener Netto-Diskkapazität in Betrieb gegangen. Pro Standort steht derzeit nur ein 16-Port-Switch zur Verfügung, d. h. es gibt noch keine Redundanz. Wenn das RZ im Frühjahr 2002 das IKA-Gebäude in Adlershof bezieht, dann soll für die dort bereits ansässigen Institute und das Institut für Physik vollständige Redundanz bezüglich der Server-, Switch- und Storage-Technik hergestellt werden. Die Planung für den

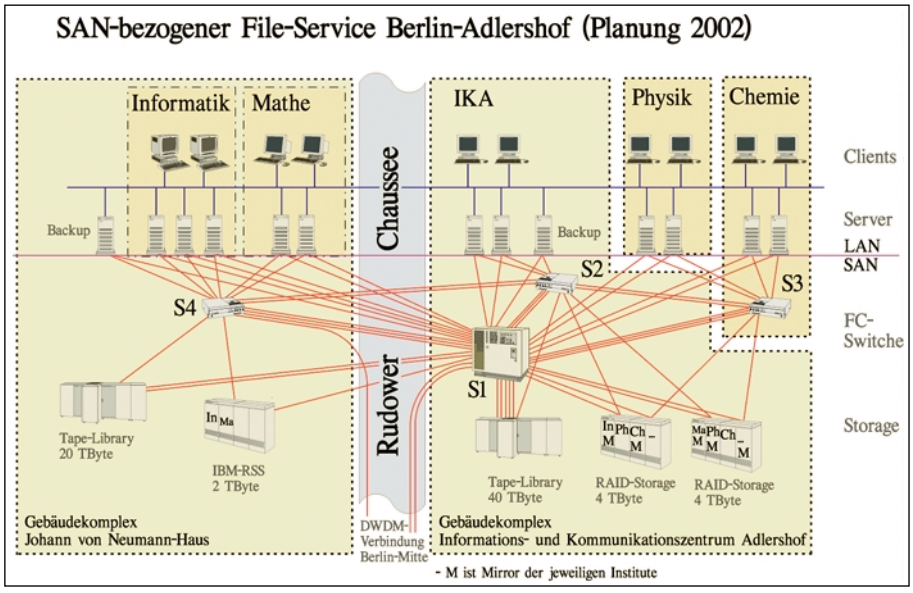

Abb. 5: SAN-Ausbau Adlershof (Planung)

Ausbau ist aus Abbildung 5 ersichtlich. Hauptbestandteile sind dort ein 64-Port-Switch (ausbaubar auf 128 Ports), zwei Storage-Einheiten mit je 4 TByte Disk-Volumen und ein neuer Tape-Roboter mit 40 TByte (unkomprimiertem) Aufnahmevermögen.

Über DWDM-Technik werden die SANs in Adlershof und Mitte verbunden. Der Standort Mitte wird in etwas kleinerem Rahmen ebenfalls in Richtung Redundanz ausgebaut und erweitert.

TLAs

CIFS	Common Internet File System (Windows-Netzwerk-Protokoll)
DWDM	Dense Wavelength Division Multiplexing
FC	Fibre Channel
FC-AL	Fibre Channel/Arbitrated Loop
HBA	Host Bus Adapter
IKA	Informations- und Kommunikationszentrum Adlershof
iSCSI	Small Computer Systems Interface over IP
JvN	Johann von Neumann
LAN	Local Area Network
LUN	Logical Unit Number
LWL	Lichtwellenleiter
NAS	Network Attached Storage
NFS	Network File System (Unix-Netzwerk-Protokoll)
RAID	Redundant Array of Independent Disks
SAN	Storage Area Network
SCSI	Small Computer System Interface
TLA	Three Letter Acronym
USV	Unterbrechungsfreie Stromversorgung

Frank Sittel
sittel@rz.hu-berlin.de

So war es zu lesen in den RZ-Mitteilungen Heft Nr. 5/1993

Sekundärspeicher: VHS-Robotersystem RSS-48 ...; dadurch wird eine Speicherkapazität von 1,04 TByte erhalten.

So war es zu lesen in den RZ-Mitteilungen Heft Nr. 7/1994

Es muß das Ziel sein, einen Computerverbund aufzubauen, der es weltweit jedermann gestattet, an jedem Ort das Verbindungskabel seines Computers in die „Steckdose" zu stecken und damit auf die Computerressourcen zugreifen zu können, die er gerade zur Bewältigung seiner Problemstellung benötigt.