Institute ans SAN

Frank Sittel
sittel@cms.hu-berlin.de

SAN, Disk Management, Virtualisierung

Seit der Einführung des SANs (Storage Area Network) an der Humboldt Universität vor mehr als drei Jahren hat sich einiges getan. Besonders nach dem Umzug des CMS nach Adlershof haben Umfang und Funktionalität eine beträchtliche Erweiterung erfahren. Stand und Perspektiven unseres SANs und wie der geneigte Leser davon profitieren kann, werden in diesem Artikel beschrieben.

Einleitung

Zu den Gründen, ein SAN aufzubauen, wurde bereits ausführlich in [1] und [2] eingegangen. Um es kurz zu sagen, es geht in erster Linie um sicheren Speicher – er sollte physisch sicher sein und wenn möglich, hoch verfügbar. Da jede Art von Hardware ausfallen kann, gilt es also Vorkehrungen zu treffen, mit einem solchen Ereignis umzugehen. Wir lassen uns dabei von den folgenden Prinzipien leiten:

- dezentrale Redundanz
- zentrale Verwaltung

Eine interne Redundanz besitzen viele Geräte (doppelte Netzteile und Lüfter, RAID) – sie nützt nur nichts, wenn Strom oder Klimaanlage ausfallen oder gar ein Wasserrohr platzt. Redundanz muss mit Dezentralisierung kombiniert werden, damit man im Desasterfall auch wirksam vor solcher Unbill geschützt ist. In einem Netzwerk ist eine dezentralisierte Redundanz leicht zu bewerkstelligen – sie muss allerdings auch beherrschbar sein. Ideal wäre eine zentrale Management-Konsole, von der aus sowohl das Netzwerk als auch der Zustand der Speicher und Clients kontrolliert werden können. Wie eine solchermaßen zentral verwaltete, dezentrale Redundanz vom CMS realisiert wurde, kann im Folgenden nachgelesen werden.

SAN und seine Struktur

Das Rückgrat des SANs ist ein Netzwerk aus Switches. Sie kommen von der Firma Brocade und haben 8, 16 oder 2 x 64 Ports. Es sind durchgängig 2 Gbit/s Ports. Diese können derart in Gruppen zusammengefasst werden, dass ein Port nur die Mitglieder seiner Gruppe sehen kann, aber keine außerhalb – diese Gruppen werden Zonen genannt. Das einzige in unserem SAN zugelassene Protokoll ist SCSI.

Die physische Struktur unseres SANs lässt sich durch zwei grundlegende Merkmale beschreiben (vergleiche auch Abb. 1):

- Wir betreiben zwei unabhängige Netzwerke.
- Diese Netze sind nach dem Core-Edge-Prinzip gegliedert.

Zwei Fabrics

Eine Fabric ist ein Netzwerk, das durch miteinander verbundene Switches erzeugt wird. An ein solches Storage Area Network werden dann Festplatten, Bandlaufwerke, Virtualisierungsserver (siehe unten) und die Clients angeschlossen. Die Clients im Sinne des SAN sind all die Maschinen, die Speicher-Kapazität vom SAN importieren und dann ihrerseits im LAN als Server für File Space, Mail, Datenbanken usw. agieren. Im Sinne des oben deklarierten Prinzips der Redundanz haben wir zwei unabhängige Fabrics installiert, d. h. es existiert keine direkte Verbindung zwischen ihnen. In ihrer technischen Ausstattung sind sie identisch. Alle Komponenten (Storage, Server, Clients) haben mindestens zwei Fiber-Channel-Adapter, über die sie mit beiden Fabrics verbunden sind.

Core – Edge

Die Idee ist, von einem sicheren Kern aus die Ränder zu bedienen. Unser Kern besteht aus je einem Silkworm 12000 Switch pro Standort und Fabric. In diesen Switches ist neben allen anderen Komponenten auch das Controller-Board doppelt vorhanden, so dass auch Firmware-Updates und Reboots eines Controller-Boards im laufenden Betrieb durchgeführt werden können. In jeder Fabric werden die SW12k von Adlershof und dem »alten« Standort in Mitte per DWDM verbunden. An den Cores hängen dann sternförmig die Edge-Switches. Diese 8- und 16-Port-Switches bringen das SAN in die Fakultäten und Institute und an sie werden dann die lokalen

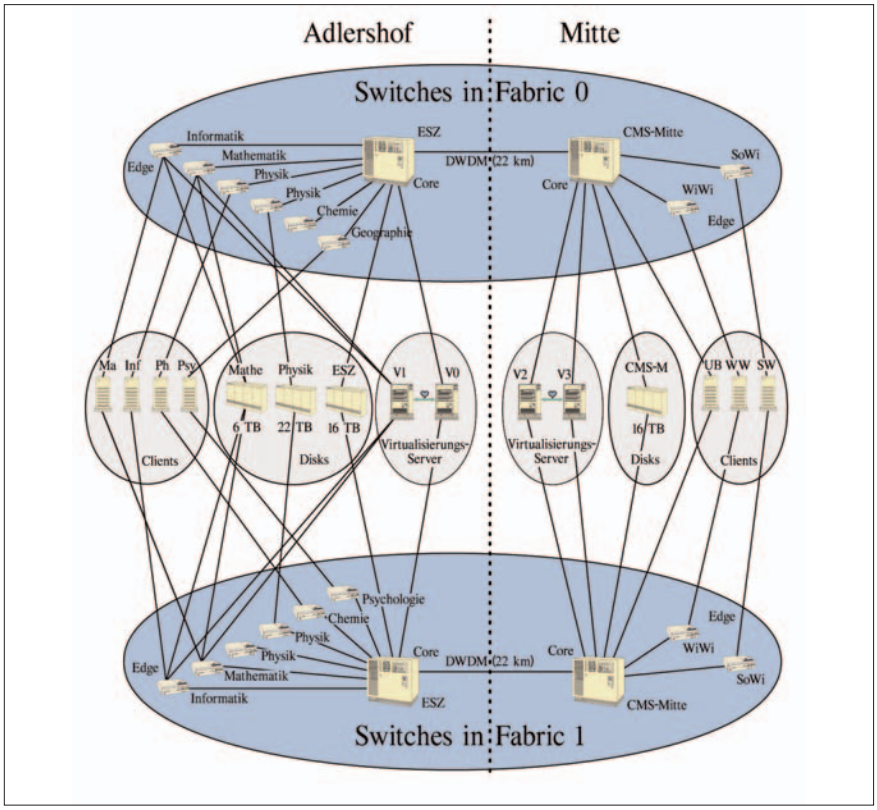

Abb. 1: Prinzipieller Aufbau des SAN

Clients, aber auch dezentral installierte Disk-Speicher und Virtualisierungsserver angeschlossen.

Wenn es möglich ist, werden die Clients gebäudeübergreifend ans SAN gebracht – z. B. wird ein Informatik-Server in der einen Fabric mit einem Switch in seinem Gebäude verbunden und in der anderen Fabric mit einem Switch im Nachbargebäude der Mathematik (siehe Abb. 1). So soll eine hohe Netzverfügbarkeit erzielt werden – denn fällt im Keller der Informatik der Strom aus (oder es brennt oder …), so ist mit hoher Wahrscheinlichkeit der Switch in der Mathematik immer noch erreichbar.

Unsere Speichertürme

Während das SAN und die weiter unten beschriebene Virtualisierung und Client-Anbindung lediglich laut Lehrbuch installiert sind, haben die von uns verwendeten Speichertürme schon eine sehr eigene Note. Sie sind nicht von der Stange gekauft, sondern in ihrer Gesamtheit von uns entworfen worden. Der Speicher in einem Turm besteht aus 12 Infortrend-RAID-Systemen, die intern 8 oder 12 IDE-Platten besitzen. Diese werden in ein RAID-5-Set plus Reserve-Platte verwandelt und entweder als Ultra-SCSI oder Fiber-Channel herausgereicht. Im Falle des Ultra-SCSI-Ausgangs kommen noch Chaparral-Router zum Einsatz, die Ultra-SCSI in Fiber-Channel-SCSI konvertieren. Die RAID-Sets haben Netto-Größen zwischen 560 GByte und 1,9 TByte. Eine SUN-Workstation übernimmt die Kontrollaufgaben über diesen Speicherturm. Wir betreiben derzeit sechs solche Türme an vier Standorten. Diese geben etwa 60 TByte Plattenkapazität über 28 2-Gbit-Kanäle an das SAN heraus.

Nun sehe ich den einen oder anderen Leser schon die Nase rümpfen – IDE-Platte?

Die gängigsten zwei Vorurteile gegenüber IDE-Platten lauten:
– IDE-Platten gehen schnell kaputt
– IDE-Platten sind langsam

Gehen IDE-Disks schnell kaputt?

Die ersten zwei IDE-Türme laufen jetzt seit nahezu drei Jahren, in ihnen rotieren 192 Festplatten mit je 100 GByte Kapazität – davon sind 8 real auf der Strecke geblieben. Von 288 IDE-Disk, die seit einem Jahr bei uns Dienst tun, ist noch keine kaputtgegangen, von 144 Platten, die Anfang diesen Jahres in Betrieb gingen, ist nach zwei Wochen eine ausgestiegen. Von 624 Platten 9 Stück verloren – wer Plattenstatistiken kennt, wird zugeben, dass dies für einen ununterbrochenen Dauerbetrieb eine sehr geringe Verlustquote ist.

Sind IDE-Disks langsam?

Im Prinzip njein – das soll heißen, dass moderne Ultra-ATA-100/133-Disks eine nahezu geniale Geschwindigkeit beim sequentiellen Datentransfer hinlegen und es in dieser Kategorie durchaus mit viel teureren SCSI-Disks aufnehmen können, aber ziemlich erbärmliche Resultate beim Random Access (zufälliger Lese-/Schreibzugriff auf kleine Datenportionen) hervorbringen. Das liegt vor allem an dem sehr schlichtem IDE-Protokoll, das billige Controller im Rechner und auf der Platte ermöglicht. Nun werden unsere Platten niemals solo im SAN angeboten, sondern prinzipiell als RAID-5-Set organisiert und vom RAID-Controller per SCSI herausgereicht. Dieser RAID-Controller beherrscht natürlich auf der SCSI-Seite all die Vorzüge, die eine halbwegs vernünftige Random Access Performance ermöglichen. Mit Hilfe seines eigenen Caches und des Caches der Platten erbringt er dann trotz der IDE-Kanäle eine gute Leistung. Zudem halten wir Plattenzahl und –größe, bezogen auf einen RAID-Controller, klein, um möglichst wenig Clients auf ein RAID-Set zu ziehen. Tatsächlich hat die Erfahrung gezeigt, dass wir nur zwei Clients in unserem System haben, die eine hohe Platten-Performance benötigen. Für diese werden wir spezielle Lösungen finden.

Virtualisierungsserver

Die Virtualisierungsserver stellen (wie bereits in [2] beschrieben) das Bindeglied zwischen dem realen Speicher und den Clients dar (siehe auch Abb. 2). Neben dem SAN-Management (im Wesentlichen Zonen-Konfiguration) stellen sie das wichtigste Tool zur Speicherverwaltung im SAN dar. Sie haben die Aufgabe, auf der einen Seite den realen Disk-Speicher zu importieren, mit ihm einige Zauberei anzustellen (in die richtige Größe bringen, auf unterschiedliche RAID-Sets spiegeln und zeitgesteuert Schnappschüsse anfertigen) und auf der anderen Seite unseren Clients als maßgeschneiderte, gespiegelte Platte anzubieten. Für die Funktion des Virtualisierungssystems werden auch hier wieder zwei Grundsätze formuliert:

- Speicher und Clients werden so einfach wie möglich gehalten.
- Die gesamte Intelligenz des Systems liegt in den Virtualisierungsservern.

Einfacher Speicher

Um das Management übersichtlich zu halten, darf es nicht zu viele Punkte geben, an denen man herumfummeln muss. Die Infortrendgeräte beherrschen eine Menge Features zur Strukturierung der zu verwaltenden Platten (Logical Drives, Logical Volumes, Partitioning, LUN-Masking …). Aber je mehr man diese Mittel benutzt, um so mehr muss man sich merken, was, wie, wo und aus welchem Grund konfiguriert wurde, damit im Fehlerfall eine Ursachenforschung überhaupt möglich wird. Bei uns laufen derzeit 72 RAID-Arrays plus 10 Chaparral-Router – da kann man sich schnell mal irren. Viel besser ist es, die Arrays alle völlig identisch zu konfigurieren (RAID-5 + Reserve-Disk), sie ohne jede Restriktion (kein LUN-Masking) ins SAN zu exportieren und eine weiter hinten liegende Instanz entscheiden zu lassen, was daraus wird. Dort packt man sie in eine Zone mit den Adaptern der Virtualisierungsserver, die für den physischen Plattenimport zuständig sind. Solchermaßen ihrer Intelligenz beraubt, gibt es für die Disk-Arrays nur noch zwei Zustände: geht oder geht nicht – weitergehende Analysen kann man sich dann dort sparen.

Einfache Clients

Ähnlich verhält es sich am anderen Ende der Zuordnungskette – bei den Clients. Wie schon in [2] erwähnt, werden wir den RAID-5-Sets nicht trauen und prinzipiell 2 RAID-5-Sets aufeinander spiegeln. Das könnten natürlich auch die Clients mittels Logical Volume Managements oder ähnlicher Methoden. Aber bei einer Vielzahl von Clients, Architekturen und Betriebssystemversionen ist eine zentrale Kontrolle mehr als schwierig und dezentrale Administratoren sind damit oft überfordert. Auch hier ist es besser, Einfachheit walten zu lassen. Der Client erhält genau eine Platte und wer die spiegelt, kann ihm egal sein. Auch hier kennt der Client nur zwei Zustände: Er hat die Platte oder er hat sie nicht – hat er sie nicht, dann entfallen schon viele Fragen, die man ans Betriebssystem richten würde.

Die Zentrale: Virtualisierungsserver

Wir verwenden als Virtualisierungssystem IPStor der Firma Falconstor. Die Virtualisierungsserver sind nun die Instanz, in der alle Informationen über Zuordnung und interne Aufteilung des realen Speichers, über Größe, Handhabung und Zugriffsrechte der virtuellen Platten an die

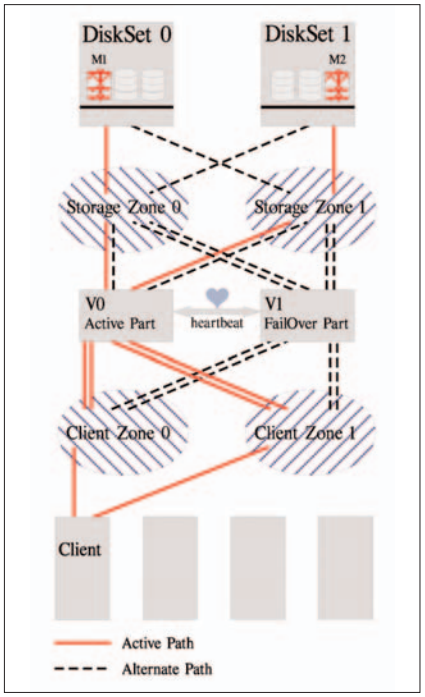

Abb. 2: Arbeitsweise der Virtualisierungsserver

Clients gebündelt sind. Sie geben eine maßgeschneiderte Disk an genau den Client, der bei ihnen registriert ist und spiegeln diese auf zwei Arrays in verschiedenen Gebäude. Wir lassen sie zeitgesteuert für bestimmte Clients Snapshots anfertigen. Sie arbeiten als Failover-Pärchen an den Standorten Adlershof und Mitte. Über eine administrative grafische Oberfläche oder per Kommandozeile erhält man schnell einen Überblick über den Gesamtzustand des Systems. Da wir prinzipiell spiegeln, schmelzen die oben erwähnten 60 TByte Netto-Disk auf 30 TByte zusammen. Die gute Nachricht ist aber – es ist noch einiges frei … und zwar für Sie, ja genau für Sie, die Leser (zumindest wenn Sie Angehörige der Humboldt Universität sind). Sie erhalten den Speicher dann über lokale oder zentrale Fileserver.

Clients

Wie weiter oben ausgeführt, sind unsere SAN-Clients ihre LAN-Server. Sie werden in der Regel mit zwei Fiber-Channel-Adapters ausgerüstet, die mit jeder Fabric verbunden werden. Wie der Abb. 2 entnommen werden kann, verfügt der zugeordnete Virtualisierungsserver in den Client-Zonen über je zwei Adapter. Typischerweise gibt er die virtuelle Platte an den Client über alle seine Adapter heraus, so dass der Client die gleiche Platte viermal zu Gesicht bekommt. Eine Loadbalancing Software stellt einen Treiber zur Verfügung, der dafür sorgt, dass aus den vier Platten wieder eine wird und dass alle vier Ziele gleichmäßig genutzt werden. Die Platte ist für den Client zugreifbar, solange auch nur ein Pfad existiert.

Kleine Statistik

Aus den Anfangszeiten des SAN haben wir noch einige Systeme, die direkt zugewiesene SAN-Speicher besitzen:

Mathematik	0,5 TB
Informatik	0,5 TB
Chemie	0,5 TB
Kultur- und Kunstwissenschaften	0,5 TB

Über die Virtualisierung verteilen wir im Moment folgende Festplattenvolumen (ohne Snapshot-Bereiche):

Physik	4,0 TB
Universitätsbibliothek	1,6 TB
Psychologie	1,6 TB
Mathematik	1,0 TB
Informatik	1,0 TB
Geographie	0,8 TB
Sozialwissenschaften	0,5 TB
Geschichte	0,2 TB
CMS	3,8 TB

Wie geht's weiter?

Wir werden das hier vorgestellte System weiter ausbauen, d. h. weitere Clients integrieren und zusätzliche Standorte hinzugewinnen. Da wir in der Vergangenheit wirklich gute Erfahrungen mit IDE-Disks gemacht haben, werden wir diesen Weg weiter beschreiten und uns demnächst für S-ATA-Systeme interessieren. Unsere Virtualisierungsserver beherrschen »SCSI over IP« – eine Technologie, die wir uns auf jeden Fall näher ansehen werden. Außerdem stehen noch Tests für den koordinierten Zugriff mehrerer Clients auf die gleiche Disk aus.

Literatur

[1] Sittel, F., Weickmann, C.: Fileservice für die Institute auf Basis eines Storage Area Network bzw. http://edoc.hu-berlin.de/e_rzm/22/sittel-frank-2001-11-01/PDF/4.pdf

[2] Sittel, F.: Storage-Area-Network an der Humbold-Universität. *RZ-Mitteilungen* 22/Nov. 2001, S. 10-14 bzw. http://edoc.hu-berlin.de/e_rzm/24/sittel-frank-2003-04-17/PDF/17.pdf