FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Konzept für das zentrale Datamanagement im Forschungszentrum Jülich

Otto Büchner, Ulrike Schmidt,
Wolfgang Gürich, Lothar Wollschläger

FZJ-ZAM-IB-2005-19

Dezember 2005

(letzte Änderung: 31.12.2005)

Einleitung

Gesamtkonzept für das zentrale Datenmanagement

Bei der Begutachtung des Programms „Wissenschaftliches Rechnen“ wurde von den Gutachtern die Weiterentwicklung eines Gesamtkonzepts für das Datenmanagement gefordert, das sowohl die Supercomputer als auch die zentralen Dienste für die wissenschaftliche Infrastruktur umfassen soll.

Das zentrale Datenmanagement lässt sich untergliedern in

A. Online Storage für zentrale Fileserver und Datenbanksysteme in der wissenschaftlichen Datenverarbeitung (Workstation-Gruppen, zentrale Dienste)
B. Online Storage für Supercomputing und Clustersysteme
C. Band-Robotersysteme für Backup- und Archiv-Dienste sowohl für zentrale als auch dezentrale Aufgaben für Wissenschaft und Verwaltung.

Eine weitere Dimension in der Betrachtungsweise unterscheidet:

i. Benutzerbezogene Datenhaltung
ii. Projektbezogene Datenhaltung

Das Gesamtkonzept soll die technischen, wirtschaftlichen und organisatorischen Rahmenbedingungen beschreiben, Entwicklungstendenzen aufzeigen und Vorschläge für die Weiterentwicklung in den nächsten Jahren erarbeiten.

Der Bereich Online Storage für zentrale Fileserver und Datenbanksysteme wurde im ZAM im Rahmen von F&E - Arbeiten untersucht. Neben der Analyse von Nutzungsmodellen war die praktische Erprobung von SAN-Komponenten unterschiedlicher Hersteller in einem Verbund von besonderer Bedeutung, da es nur zu wirtschaftlich tragbaren Lösungen kommen kann, wenn die Abhängigkeit von Herstellern und Lieferanten breit gestreut ist. Anfang des Jahres 2005 wurden die Ergebnisse dieser Arbeiten der DV-Kommission vorgestellt und in diesem Papier in Teil 1 präsentiert.

In den IBM-Supercomputer-Komlpex ist ein modernes, flexibles und leistungsfähiges Plattenspeichersystem auf der Basis eines Storage Area Networks (SAN) vorhanden. Zusammen mit den Band-Robotersystemen und der HSM-Funktion steht den Benutzern nahezu ein fast unendlich großer, virtueller Online-Speicherraum zur Verfügung. Mit dem Global Parallel Filesystem (GPFS) ist ein modernes paralleles Filesystem im Einsatz, das im europäischen Grid-Projekt DEISA den Kern eines verteilten Datenhaltungssystems bildet. Auf

lokaler Ebene können mit GPFS Cluster von Supercomputern und Linux–basierten Rechnern gebildet werden. Das Papier beschreibt im Teil 2 Strategien zum Ausbau des Online Storage im Supercomputing und die Ausweitung des Datenverbundes auf Linux-basierte Clustersysteme.

Bei den Band-Robotersystemen sind im Jahr 2004 Beschaffungen zur Erhöhung der Bandkapazität vorgenommen worden. Das war nötig, um die hohen Nachfragen nach Speicherkapazität für HSM und für Archivierung zu befriedigen.
Im 3. Teil des Papiers werden die Dienste Backup und Archivierung konzeptionell dargestellt.

Eine mögliche Lösung, die derzeit untersucht wird, für die projektbezogene Datenhaltung wird im Teil 4 skizziert.

Teil 1

Online Storage für Datenbanken und Fileserver und Workstationgruppen

STATUS

Das ZAM hält für die wissenschaftliche Datenverarbeitung im Forschungszentrum Jülich Datenbanken auf mehreren Servern bereit. Auf die Datenbanken wird in der Regel mit Datenbanken-Clientsoftware zugegriffen.

Weiterhin betreibt das ZAM für die Institute IME, IPP und ICG Fileserver. Die Daten werden den Benutzern über NFS und HTTP zur Verfügung gestellt.

Beim Workstationgruppen-Konzept werden den Instituten wissenschaftliche Arbeitsplätze in Form von Linux-Klienten angeboten. Die Daten für die Arbeitsplätze verwaltet ein Workstationgruppen-Server per NFS und Samba. Derzeit sind über 15 Gruppen mit mehr als 300 Klienten im Einsatz.

Die genutzte Festplattenperipherie aller genannten Server besteht aus lokal angeschlossenen SCSI-Plattensubsystemen, SCSI-RAIDs und SCSI/SATA-RAIDs. Der Anschluss der Plattenperipherie mittels Kupferkabel an die Hostbus Adapter ist unflexibel und bei einigen Servern nicht erweiterbar. Ein Großteil der SCSI-Plattensubsysteme und der SCSI-RAIDs sowie der Server-Hardware muss wegen Überalterung ausgetauscht werden.

SERVERKONSOLIDIERUNG

Im Jahr 2004 wurde begonnen, die Anzahl der Workstationgruppen-Server zu reduzieren. Gruppen mit wenigen Klienten wurden zu einer großen Gruppe zusammengefasst. So konnten bis Anfang 2005 sieben vorwiegend kleinere Gruppen mit über 100 Klienten auf einen campusweiten Server migriert werden. Das Datenvolumen, das auf den Server übertragen wurde, liegt bei wenigen hundert Gigabyte. Der Server hat in 2005 noch ein paar kleine und mittlere Gruppen übernommen. Für Workstationgruppen mit großem Datenvolumen trägt die Lösung nicht. Mittelfristig sind maximal drei Server jeweils mit mehr als hundert Klienten und Datenvolumen im Tetrabyte-Bereich geplant.

Bei den Fileservern für Institute sind jeweils ein oder zwei Server im Einsatz. Hier steht der Austausch der Hardware auf der Agenda. Bei den Datenbanken sind AIX- und Linux-Systeme im Einsatz. Für das erste Halbjahr 2006 ist ein Austausch des primären Datenbankservers durch eine leistungsfähigere Hardware in Vorbereitung.

Ein wichtiger Gesichtspunkt ist die Auswahl der Serverplattform. Für das ZAM gibt es folgende Alternativen:

- Linux auf PC-Hardware
- AIX auf pSeries Hardware
- Solaris auf Sun Sparc

Für die Betriebssysteme ist Expertise im ZAM vorhanden. Linux-Rechner werden als Desktopsysteme und als Applikation-Server in der DV-Landschaft des Forschungszentrums Jülich breiten Raum einnehmen. AIX-Systeme werden im Supercomputing des ZAM auf absehbare Zeit eingesetzt. Sun-Systeme mit dem Betriebssystem Solaris werden nur noch vereinzelt genutzt, wohl aber Opteron-Rechner unter SuSE, geliefert von SUN. Die Migration von Solaris-Systemen nach Linux wird angestrebt. Ein campusweiter Workstationgruppen-Server mit Solaris als Betriebssystem wird auf absehbare Zeit im Einsatz bleiben.

SPEICHERKONSOLIDIERUNG

Bei der Anschaffung neuer Speichermedien sollten folgende Prinzipien beachtet werden:

- Hohe Verfügbarkeit
- Flexible Ausbaufähigkeit
- Wartbarkeit bei geringen Personalmitteln
- relativ kostengünstige Lösungen

Aus Kostengründen kann keine Hochverfügbarkeits-Lösung (HA) für alle Fileserver angestrebt werden. Dennoch sollten Ausfallzeiten minimiert werden. Die Länge der akzeptablen Ausfallzeiten muss sich an der Anzahl der betroffenen Benutzer orientieren. Da bei einem Ausfall eines Workstationgruppen-Servers mehr als hundert Benutzer nicht mehr arbeiten können, muss die Ausfallzeit auf ein bis zwei Stunden begrenzt sein. Sind nur wenige Benutzer betroffen, sollte der Ausfall nicht mehr als einen Tag dauern. Die Speichermedien müssen so gewählt werden, dass sie in den nächsten zwei Jahren durch den Zukauf von Komponenten erweitert werden können. Die Systeme müssen einfach konfigurierbar und automatisch zu überwachen sein. Die Institute, die für die Kosten ihrer Fileserver aufkommen, achten besonders auf preiswerte Lösungen.

UNTERSUCHUNGEN ZUR DATENHALTUNG

Um fundierte Aussagen für ein Konzept zur Datenhaltung machen zu können, wurde 2004 mit Untersuchungen zu dem Thema begonnen, die Anfang 2005 abgeschlossen wurden.

Folgende Alternativen waren gegeneinander abzuwägen:

- Direct Attached Storage (DAS) über SCSI-Kabel gegenüber Fiber Channel (FC): Für DAS spricht der günstige Preis, insbesondere fallen keine hohen Kosten für den FC Hostbus Adapter an. Die FC Schnittstelle bietet mit Switch praktisch unbegrenzte Erweiterbarkeit.
- Einheitliches SAN gegenüber verteilten Lösungen: Für SAN spricht ein effizienter Einsatz der Platten und flexible Ausbaufähigkeit. Bei verteilten Lösungen entfallen Kosten für FC-Switches und die Komplexität der Lösung wird geringer.
- SCSI-Platten gegenüber SATA-Platten: Für SCSI-Platten spricht die längere Haltbarkeit und eine höhere Performance. Die SATA-Platten sind deutlich günstiger im Preis.

Mitte 2005 wurden erste Untersuchungen zu iSCSI-Lösungen vorgenommen. Die vorläufigen Ergebnisse sind in einem eigenen Abschnitt dokumentiert.

AUFBAU EINER TESTUMGEBUNG

Um Entscheidungen zu künftigen Anschaffungen zu ermöglichen, wurde eine Testumgebung aus AIX- und Linux-Servern mit FC- und SATA-RAIDs aufgebaut.
Als Testsysteme standen zur Verfügung:

zwei Transtec 1002 Flatrack Server mit

- je 2 Intel Xenon 2.8 GHz Prozessoren,
- 4 GB ECC DDR RAM,
- je 2 36 GB Ultra 320 SCSI-Platten,
- je 2 Intel Dual Port Gigabit Server Adapter,
- je 2 PCI-X FC Adapter.

ein Transtec 6600 FC Premium RAID mit

- Dual Controller mit je 1 GB Cache,
- 16 hot-swap 146 GB Fiber Channel Festplatten.

ein Transtec 6100 SATA Premium RAID mit

- Single Controller mit 512 MB Cache,
- 16 hot-swap 250 GB SATA Festplatten.

zwei McData Spheron 4500 Switches 8 Port.

Für die Untersuchung der Einsatzmöglichkeiten von AIX-Systemen als Fileserver wurde

ein IBM pSeries 615 Server mit

- Power 4 1.2 GHz Prozessor,
- 1 GB ECC RAM,
- 36 GB LVD SCSI Disk,
- 100 MB Ethernet Adapter,
- Cambex FC Hostbus Adapter,

als Testmaschine genutzt.

Kurzzeitig konnten ein Transtec 1002 Flatrack Server mit Ultra 320 Hostbus Adapter und ein Transtec 6100 SATA RAID mit U320 SCSI-Schnittstelle (DAS) erprobt werden, um mögliche Performance-Unterschiede zwischen Fiber Channel und Direct Attached Storage zu untersuchen.

Die Messungen wurden komplettiert durch Untersuchungen an einem existierenden, campusweit genutzten Solaris-Workstationgruppen-Server:

eine Sun Sparc Fire 280R mit

- 2 x 900 MHz Prozessoren,
- 1 GB RAM,
- 1 GB Ethernet Netzwerkadapter.

Die Fiber Channel Komponenten wurden redundant ausgelegt. Mit zwei Linux-Servern sollte eine HA-Lösung mittels des Programmpakets „heartbeat“ implementiert werden. IBM bietet HA-Lösungen für seine Systeme an. Diese Lösung konnte aus Kosten- und Zeitgründen nicht getestet werden. Die Linux-Rechner waren mit SuSE-Linux-Enterprise-Server 8 (SLES8) vorinstalliert. Während des Tests wurden die Rechner auf SLES9 aufgerüstet. Das IBM-System hatte die AIX Version 5.2. Es gibt mitgelieferte Verwaltungssoftware für die RAID-Systeme (RAIDWatch), für die Fiber Channel Switchs (SANPilot) und für die Fiber Channel Karten im Linux (SANblade).

Besonderes Augenmerk muss auf die Auswahl der benutzten Filesysteme gelegt werden. Für Fileserver kommen aus Stabilitätsgründen nur Journal-Filesysteme in Betracht. Nach einem Crash des Servers müssen die Filesysteme schnell und ohne Datenverlust wieder herstellbar sein. Für AIX ist das Filesystem JFS2 (eine Erweiterung von JFS) die einzige sinnvolle

Alternative. Für Linux sind als Journal-Filesysteme Reiserfs und EXT3 seit längerem verfügbar. Als neuere Journal-Filesysteme sind auch das Filesystem XFS von SGI und das JFS von IBM, eine Variante des JFS Filesystems im OS/2, in den gängigen Linux Distributionen zu finden. Alle Filesysteme wurden in Hinsicht auf Performance und Stabilität getestet. Für die Durchsatzmessungen wurde bonnie++ sowohl auf AIX sowie auf Linux benutzt.

FUNKTIONSTESTS DER KOMPONENTEN

Die Transtec-RAIDs entsprechen EonStor-RAIDs von Infortrend und sind unter anderen Namen von mehreren Anbietern erhältlich. Die RAIDs waren mit 2 RAID5-Sets über je 8 Platten konfiguriert. Nachteil dieser Konfiguration ist das Fehlen einer Hot-Spare-Platte. Um die Sicherheit der RAID-Sets gegenüber Plattenausfällen zu erhöhen, wurden 3 RAID-Sets mit je 5 Platten und einer globalen Hot-Spare-Platte angelegt. Beim Ausfall einer RAID-Platte wird automatisch die Hot-Spare-Platte in das betroffene RAID-Set integriert und die Parity des RAID-Sets in 3 – 4 Stunden neu berechnet.

Das FC-RAID ist im Gegensatz zum SATA-RAID mit redundantem Kontroller ausgestattet. Ein Kontroller Ausfall wird vom angeschlossenen Hostsystem nicht bemerkt. Das RAID arbeitet ohne Unterbrechung mit dem zweiten Kontroller weiter.
Beim Transtec 6600 FC Premium RAID traten anfangs Fehler auf, die durch Austausch von FC-Platten und Kontroller-Memory beseitigt wurden. Die RAIDs besitzen eine Netzwerkschnittstelle, über die Konfiguration und Steuerung mittels Telnet, die Überwachung und Steuerung über Web-Services vorgenommen werden kann. Die Java-basierte Web-Schnittstelle ist unhandlich und entbehrlich. Das RAID kann so konfiguriert werden, dass RAID-Events (Hardwaredefekt, Reboot) über Mail gemeldet werden.

Die McData Spheron 4500 Switches sind nur per Web-Interface oder serieller Schnittstelle konfigurierbar. Überwachung ist in der gelieferten Softwareversion nur über SNMP möglich. Die Konfiguration bestand im Wesentlichen aus dem Einschalten des Default Zoning. Dabei wird der Switch als Blackbox benutzt. Alle RAID-Sets sind von allen angeschlossenen Servern als logische Disks zu sehen.

Bei einem der beiden Linux-Server waren die Treiber für die FC Hostbus Adapter integriert. Beim zweiten Server konnte das problemlos nachgeholt werden. Da die Server jeweils 2 FC Hostbus Adapter besitzen, die aus Redundanzgründen jeweils an unterschiedliche Switch-Systeme angeschlossen sind, wurden alle logischen Platten doppelt von den Rechnern erkannt. Der Grund war, dass das Failover, das eine automatische Übernahme des zweiten Adapters bei Ausfall der Verbindung des ersten Adapters gewährleistet, nicht eingeschaltet war. Mit dem Comand-Line-Interface (CLI) der SANblade-Software ist ein Einschalten der Failover Funktion nicht möglich. Mit der Java-Schnittstelle von SANblade ist Failover konfigurierbar, allerdings funktionierte die Software nicht korrekt. Recherchen im Netz ergaben eine dritte unkomplizierte

Lösung. Das Failover kann als Option beim Laden des Kernelmoduls dem Linux bekannt gemacht werden und funktioniert unter SLES8.
Nach Aussage von Qlogic wird das Funktionieren ihrer FC-Karten für SLES9 nicht garantiert. Die Karten funktionieren für SLES9, sogar mit der Failover Option. In den Release Notes von SuSE zu SLES9 wird aber vom Failover abgeraten.

Bei dem AIX-System müssen nach Einbau der FC-Karte die Device-Treiber für die Karte und die FC-Platten nachinstalliert werden. Failover ist im AIX möglich, wurde aber nicht getestet.

STABILITÄT DER FILESYSTEME

Im ZAM sind verschiedentlich Probleme mit Linux-Servern aufgetreten, so dass nach einem Crash des Servers Filesysteme nicht mehr brauchbar waren. Für einen Fileserver mit Filesystemen im Bereich von mehren hundert Gigabyte Größe könnte ein solcher Vorfall einen Ausfall von mehreren Tagen bedeuten. Es ist auch nicht völlig sichergestellt, ob alle Daten aus dem Backup fehlerfrei rekonstruiert werden können. Das Wissen um die Schwachstellen im Linux, hat uns bewogen, AIX-Systeme in die Untersuchungen mit einzubeziehen.

Getestet wurden das JFS2 auf dem AIX-System unterhalb des Logical Volume Managers sowie EXT3, Reiserfs, JFS und XFS auf Linux. Für die Stabilitätsuntersuchungen wurden die Schreib-Operationen von bonnie++ und einem eigenen Testprogramm, das viele kleine Dateien erzeugt durch

- Unterbrechung der Datenverbindung auf dem FC-Switch oder
- durch Reset des RAID-Controller oder
- durch Shutdown des RAID-Controller

unterbrochen. Nach Wiederherstellung der Datenverbindung wurde versucht, das Filesystem ohne Datenverluste zu reaktivieren.

Problemlos ist ein Reset des RAID-Controllers, wenn die Systeme über Fiber Channel angeschlossen sind. Der IO wird für die Dauer des Reboot des RAIDs (1-2 Minuten) unterbrochen. Fehlermeldungen gibt es keine. Wenn das RAID wieder verfügbar ist, wird der IO wieder aufgenommen. In den anderen getesteten Varianten, Shutdown des RAID-Controllers , Unterbrechung des FC-Pfades am Switch oder Reset des RAID-Controllers, der über SCSI-Kabel angeschlossen war, traten SCSI-Fehler auf. Der IO wurde mit einer Fehlermeldung abgebrochen.

Nachdem das RAID wieder verfügbar war, waren im Fall EXT3, Reiserfs und JFS die Filesysteme readonly verfügbar. XFS macht vorsorglich ein Shutdown des Filesystems. Ein Umount und ein Mount auf das XFS-Filesystem behebt das Problem. Alle Dateien auf dem XFS-Filesystem waren noch vorhanden. Eine Datei die permanent im Schreibzugriff war, war nicht mehr lesbar. Bei den anderen Linux–Filesystemen muss das Filesystem mit Umount

abgehängt werden und mit einem Filesystem-Check (fsck) überprüft werden. Nach dem fsck waren mehr Datenverluste zu beobachten als beim XFS-Filesystem.

Positiv ist das leichte Reaktivieren des XFS-Filesystems nach einer Unterbrechung der Datenverbindung zu vermerken. Der automatische Shutdown des Filesystems bei Problemen sorgt dafür, dass mit Umount/Mount das Filesystem im laufenden Betrieb reaktiviert werden kann. Bei Reiserfs und JFS kam es teilweise zu Blockierungen, die ein Reset des Linux-Systems nötig machten.

Das AIX-System reagierte bei Unterbrechung des Fiber Channel Pfades am Switch mit Fehlern des LVM (Logical Volume Manager). Das Filesystem blieb gemountet und war nach Verfügbarkeit des RAID wieder aktiv. Datenverluste konnten nicht festgestellt werden.

PERFORMANCE MESSUNGEN

Um Aussagen zur Performance von SATA- und FC-Platten, der verschiedenen Filesysteme und der Betriebssysteme zu machen, wurden Messungen mit bonnie++ vorgenommen (siehe Anhang 2). Das Programm besteht aus einem IO-Test und einem File-Creation-Test. Der IO-Test besteht aus sequenziellem Output (Character, Block und Rewrite), sequenziellem Input (Character, Block) und zufälligem Suchen. Die Messungen bezüglich Character-IO sagen wegen der CPU-Auslastung von nahezu 100 % nichts über die Geschwindigkeit der RAIDs, sondern beschreiben die Performance von Hardware und Betriebssystem des Servers. Deshalb wurden zur Bewertung der Filesysteme im Wesentlichen die Werte bei Block-IO (Input und Output) betrachtet.
Um zu verhindern, dass der Hauptspeicher des Servers eine Rolle spielt, wurde die Größe der zu lesenden und zu schreibenden Dateien auf das zweifache der Hauptspeichergröße gesetzt. Unter Linux wurde ein deutlicher Unterschied zwischen den Betriebssystemen SLES8 und SLES9 festgestellt. Mit SLES9 stieg der Character-IO um bis zu 15 % bei gleicher Hardware gegenüber SLES8. Alle referierten Linux-Zahlen sind unter SLES9 gemessene Zahlen.

Der maximale Durchsatz für FC-Platten liegt bei 140 MByte/sec für Schreiben und 115 MByte/sec für Lesen. Er übertrifft die Werte für SATA1-Platten beim Schreiben und Lesen, die beide bei 85 MByte/sec liegen, um 65 % bzw. 35 %. Bei den Zahlen für den Schreibzugriff muss in Betracht gezogen werden, dass das FC-RAID mit 1 GB doppelt so viel Cache zur Verfügung hat, wie das SATA-RAID. Die Werte für den Lesezugriff ließen sich verbessern, wenn man die RAID-Sets größer als 5 Platten dimensioniert. Dies ist aber aus Gründen der Stabilität nicht erwünscht. Im Laufe des Jahres 2005 wurden RAIDs mit SATA2-Platten verfügbar. Diese SATA2-RAIDs sind in der Geschwindigkeit dem RAID mit FC-Platten aus der Testumgebung gleichwertig.
Die Performance-Unterschiede zwischen den Filesystemen waren unter SLES8 noch deutlicher als unter SLES9. Messbare Unterschiede gibt es bei der Schreib-Performance für FC-Platten.

Das schnellste Filesystem XFS war um mehr als 30 % schneller als in diesem Bereich das langsamste EXT3. Beim Lesen von FC-Platten und Lesen und Schreiben von SATA-Platten lagen die gemessenen Werte fast im Bereich der Messgenauigkeit zwischen 95 % und 102 % der EXT3-Werte. Dennoch kann man XFS, das in drei Bereichen (Schreiben/Lesen FC und Schreiben SATA) das schnellste Filesystem war, zum Sieger erklären.

Die Werte für das SATA-RAID mit Fiber Channel Anbindung konnten durch Messungen mit einem fast identischen RAID mit SCSI-Anbindung über Kupferkabel verifiziert werden. Die gemessenen Daten sind fast identisch. Der FC-Anschluss eines SATA-RAIDs bringt keinen Performance-Gewinn.

Die Zahlen, die für AIX gemessen wurden, sind nicht vergleichbar mit den unter Linux gemessenen Zahlen. Die Hardware des AIX-Testsystems ist älter und schwächer ausgebaut. Der maximale Wert für Character-IO liegt bei 34 MByte/sec. Beim Lesen von FC- und SATA-Platten können fast die Werte wie unter Linux erreicht werden. Beim Schreiben auf die Platten sind die Werte niedriger (max. 48 MByte/sec).

Die Workstationgruppen- und Fileserver werden vornehmlich als NFS-Server eingesetzt. Deshalb wurden bonnie++ Messungen sequenziell und parallel über das Netz vorgenommen. Der NFS-Server war mit einer 1 GBit Verbindung am Ethernet angeschlossen, zwei Klienten hatten eine 100 MBit Schnittstelle, ein Klient (der 2. Server) war auch mit einem GBit angeschlossen. Für die Klienten mit 100 MBit Anschluss ist der Netzwerkdurchsatz der begrenzende Faktor. Beim Lesen und Schreiben über NFS sind ca. 10.5 MByte/sec (84 MBit/sec) erreichbar. Im Fall des Klienten mit 1 GBit Anschluss können bei sequenziellem NFS-Zugriff (nur ein NFS-Klient) bis 24 MByte/sec geschrieben und 42 MByte/sec gelesen werden. Bei parallelem Zugriff aller 3 Klienten wurden akkumulierte Werte von 35 MByte/sec für das Schreiben und 55 MByte/sec für das Lesen erreicht.

HOCHVERFÜGBARKEIT FÜR LINUX

Eine Hochverfügbarkeitslösung (HA) muss einen Serverausfall überbrücken. Ein Ersatzserver übernimmt die Dienste des defekten Servers, ohne dass die Klienten längere Unterbrechungen haben oder booten müssen. Das Programmpaket heartbeat bietet eine HA Lösung für Linux an, die am Beispiel des Dienstes NFS getestet wurde. Zwei Server - ein aktiver und ein passiver - überwachen sich gegenseitig. Ist der aktive Server vom passiven Server aus nicht mehr erreichbar, wird der bislang passive Server zum Aktiven und übernimmt die Dienste des bisher aktiven Servers.

Problematisch wird es dann, wenn durch ein Kommunikationsproblem zwischen den Systemen beide Server aktiv werden. In diesem Fall könnte ein gleichzeitiger Schreibzugriff auf das gleiche Filesystem zu Datenverlust führen. Die Lösung für dieses Problem bietet eine

Komponente von heartbeat mit Namen stonith (shut the other node in the head). Der passive Server macht den aktiven Server über Zugriff auf eine mit Web-Interface konfigurierbare Steckerleiste vorübergehend stromlos und wird dann erst aktiv.
Die Kommunikation zwischen den Servern wird redundant durch ein serielles Kabel und eine zusätzliche Netzwerkverbindung sichergestellt. Für den NFS Export werden nicht die originären Netzwerkadressen der beiden Server, sondern eine dritte, virtuelle Adresse genutzt. Beim Start von heartbeat vollzieht der aktive Server folgende Aktionen:

- Aktivieren der virtuellen Adresse,
- Mounten von Filesystemen,
- Starten des NFS-Server.

Beim Stopp von heartbeat werden die Aktionen in umgekehrter Reihenfolge deaktiviert (Stopp NFS, Umount Filesystem, Deaktivieren der virtuellen Adresse).
Der Dienst NFS ist nicht ganz unproblematisch für eine Übernahme eines von Klienten gemounteten Filesystems. Klienten und Server unterhalten während eines NFS Mounts Verbindung über den rpc.statd, zum anderen hält der Server aktuelle Statusinformationen zu den Mounts innerhalb des Dateisystems /var/lib/nfs. Die erste Hürde kann man übergehen, in dem man den rpc.statd auf dem Server durch eine Option clusterfähig macht. Die zweite Problematik wird dadurch umgangen, in dem man die Daten von /var/lib/nfs auf einen Bereich eines Benutzer-Filesystems legt, das vom aktiven Server mit übernommen wird. Die Daten werden durch symbolische Links auf dem jeweils aktiven Server unter /var/lib/nfs verfügbar gemacht.
Mit diesen Konfigurationen wurde verifiziert, dass auf Klienten ein simulierter Servercrash nur durch eine kurzzeitige Hang-Situation während der Übernahme wahrgenommen wird.

iSCSI – Untersuchungen

SAN-Lösungen über IP sind in der letzten Zeit stark in der Diskussion. Ein Vorteil von iSCSI gegenüber FC ist größere räumliche Distanz zwischen Datenspeicher und den Serversystemen, die die Daten verwenden. Mit iSCSI könnten beispielsweise Rechner direkt am Experiment in den Instituten betrieben werden, die Daten aber auf einem RAID-System im ZAM liegen. Um die generelle Einsatzfähigkeit von iSCSI-Systemen in unserer Umgebung zu evaluieren, wurde ein Infortrend-iSCSI-RAID mit 12 x 250 GByte SATA-Platten und 2 x 1 GBit-Ethernet Schnittstellen angeschafft. Die iSCSI-Initiator-Software, die auf den Client-Systemen, die Umsetzung von IP-Paketen ins SCSI-Protokoll vornimmt, ist auf allen gängigen Betriebsystemen vorhanden. Dagegen sind iSCSI Hostbus Adapter im Wesentlichen nur für Linux und Windows erhältlich. Das RAID wurde mit Linux-, AIX- und Windows-Clients erfolgreich jeweils mit Initiator-Software getestet. Für einen produktiven Einsatz sollte mindestens eine separate GBit-Ethernet-Verbindung zwischen RAID und Server zur Verfügung stehen. Bei den getesteten RAID-SATA2-Platten ist der Ethernetdurchsatz mit ca. 100 MByte/sec für Lesen und Schreiben eine obere Grenze, die für den Betrieb durchaus akzeptabel ist. Ein Verzicht auf iSCSI Hostbus Adapter muss man mit höherer Systembelastung

bezahlen. Als Faustregel wird 1 Gigahertz eines Prozessors für 1 GBit-Ethernetdurchsatz angegeben. Diese Regel ließ sich durch Messungen bestätigen. Von den 2 x 2.8 GHz Prozessoren des Linux-Testsystems war ein Prozessor zu gut 30% mit der Umsetzung der IP-Pakete beschäftigt.

EMPFEHLUNGEN

Die Untersuchungen sowie der derzeitige Einsatz von Infortrend RAIDs zeigen, dass diese Systeme als Speichermedien für Fileserver geeignet sind. Die RAIDs sind kostengünstig und werden von mehreren Anbietern unter verschiedenen Namen vertrieben. Konfigurierbarkeit und Performance können positiv bewertet werden. Zur langfristigen Haltbarkeit der Systeme kann naturgemäß noch keine Aussage gemacht werden. Die einheitliche Überwachungssoftware samt der E-Mail-Alarmierung ist positiv zu bewerten.

Die Untersuchungen verifizierten, dass der Anschluss eines SATA-RAIDs über Fiber Channel kein Performance-Gewinn gegenüber dem Anschluss des gleichen RAID über SCSI-Schnittstelle bringt. Deshalb können SATA RAIDs mit SCSI-Schnittstelle für kleine Lösungen (Profil 1) betrieben werden.

Eine SAN-Lösung, die Fileserver mehrerer Institute umfasst, ist aus finanztechnischen Gründen nicht zu realisieren. Für die Workstationgruppen wird ein SAN eingesetzt werden, das mittelfristig alle Gruppenserver umfasst.

Aus Kostengründen ist eine Lösung mit SATA-Platten angemessen. Für ein SATA-RAID sind ca. 1.500 Euro je TByte Bruttokapazität zu veranschlagen. Beim FC-RAID bewegt sich der Preis für 1 TByte Kapazität je nach Ausbaustufe des RAIDs zwischen 5.000 Euro und 8.000 Euro (siehe Anhang 1). Wenn die Fileserver vornehmlich als NFS-Server im JuNet agieren, sind SATA-RAIDs ausreichend. Die Vorteile von FC-RAIDs können nur bei hohen Anforderungen an die lokale Performance oder bei schnellen Netzwerken im Cluster-Umfeld ausgenutzt werden.

Bei den Linux-Filesystemen hat bezüglich Performance und Stabilität XFS am besten abgeschnitten. An Stabilität steht, soweit getestet, das XFS-Filesystem dem JFS2-Filesystem unter AIX kaum nach. Das ZAM wird bei seinen Linux-Servern XFS als Filesystem für Massendaten einsetzen.

Beim Vergleich der Server-Architekturen bei Linux und AIX sprechen alle messbaren Kriterien für den Einsatz von Linux. Insbesondere verbietet sich wegen der hohen Kosten ein Einsatz von

AIX-Servern im Profil 1. Dennoch gibt es eine Reihe von Gründen, die für den Einsatz von AIX-Servern sprechen:

- stabilere Hardware,
- bessere und schnellere Wartung durch Einsatz von Wartungstechnikern vor Ort,
- langjährig erprobte, stabile Software und Filesysteme,
- Kontinuität von AIX (Linux ist in einem Jahr ganz anders),
- niedrigere Ausfallzeiten wegen geringerer Softwarewartungsfrequenz,
- weniger Gefahr vor Hackern (AIX ist ein Exot).

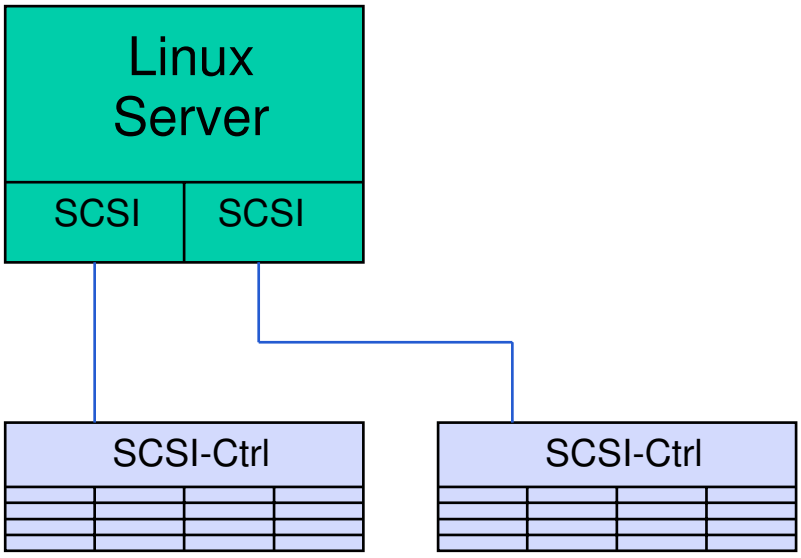
Für Anforderungen, die ein hohes Maß an Sicherheit, Stabilität und Verfügbarkeit beinhalten, ist im Einzelfall eine Lösung mit AIX-Servern zu erwägen. Im Rahmen der Konzentration auf wenige Betriebssysteme und aus Gründen begrenzter Solaris-Expertise im ZAM werden Solaris-basierte Systeme bei Neuanschaffungen als Fileserver nicht in Erwägung gezogen.

Um auf die verschiedenen Anforderungen der Institute eingehen zu können werden zwei Profile für Fileserver mit Plattenperipherie vorgeschlagen.

PROFIL 1

Eine preiswerte Lösung für Anforderungen von kleinerem bis mittlerem Datenvolumen und normalen Verfügbarkeitsansprüchen bestehend aus:

- einen Linux-Server mit einem oder zwei SCSI-Adaptern,
- ein oder zwei SATA-RAIDs mit SCSI-Adapter (max. 9.6 TByte Nettokapazität),
- Ausfallzeiten bis 24 Stunden sind tolerabel,
- die Lösung ist nicht erweiterbar.

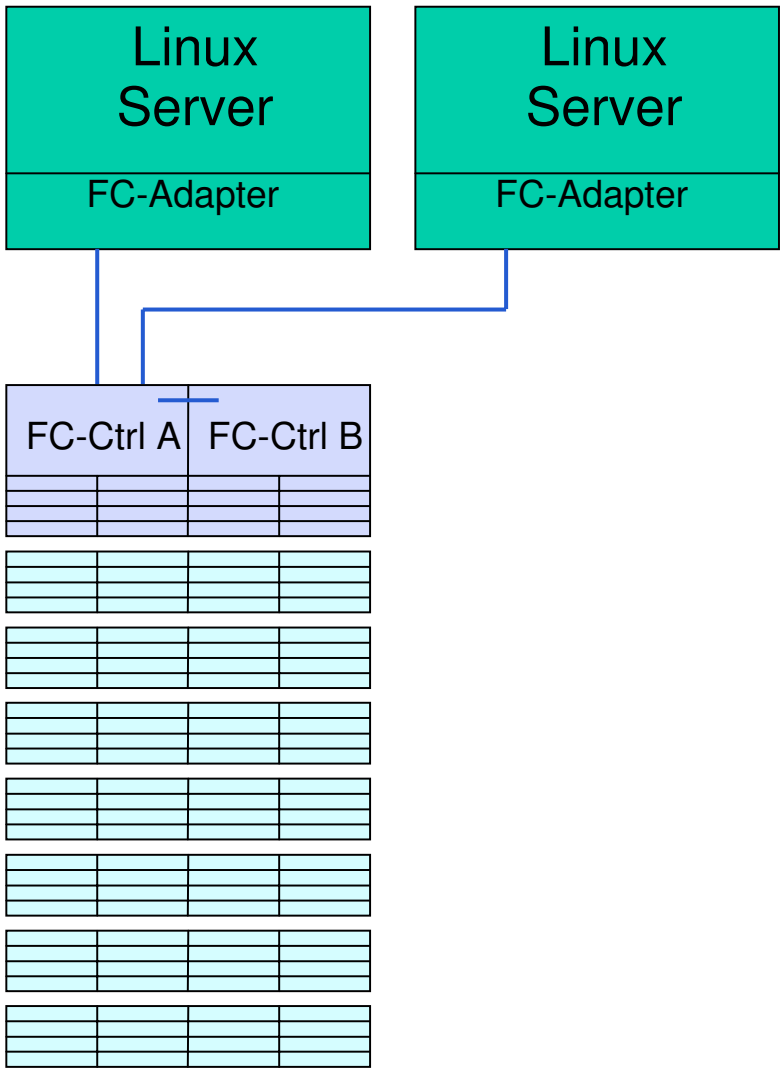


Beispiel Profile 1: 1 Server, 2 Raids

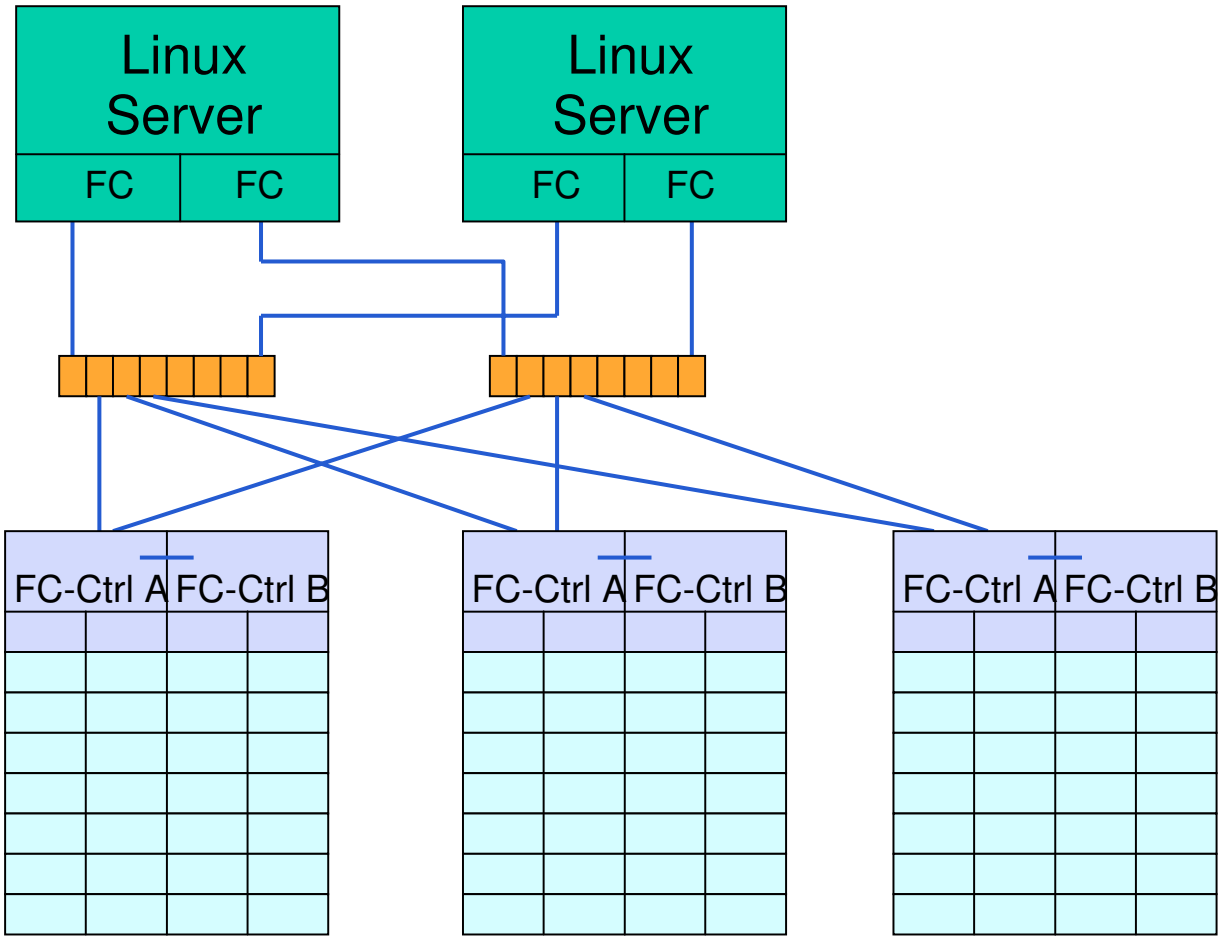
PROFIL 2

Eine Lösung für Anforderungen von großem Datenvolumen und (oder) hohen Verfügbarkeitsansprüchen bestehend aus:

- einem oder mehreren Linux-Servern mit FC-Adapter,
- HA-Lösung mit FC-Switch für Linux möglich,
- alternativ einem AIX-Server mit FC-Adapter,
- ein erweiterbares SATA/FC-RAID mit redundantem Kontroller (max. 38.4 TByte Nettokapazität)
- alternativ mehrere SATA/FC-RAIDs mit FC-Switch.



Beispiel Profile 2: 2 Server, 1 Raid



Beispiel Profile 2: 2 Server, n Raids mit Switch

Die höhere Verfügbarkeit im Profil 2 gegenüber Profil 1 wird erreicht durch:

- redundanten Kontroller im RAID,
- störunanfällige FC-Verbindungen gegenüber SCSI-Kabel,
- redundanten Linux-Server,
- oder robuste Hardware und Software von AIX-Servern gegenüber Linux–Servern.

Das ZAM wird für die Workstationgruppen die getestete SAN-Infrastruktur einsetzten. Wegen derzeit noch moderaten Datenmengen (unter 2.5 TByte) werden die Filesysteme auf dem FC-RAID zusätzlich auf das SATA-RAID gespiegelt. Damit sollen im Fall von korrupten Filesystemen lange Restore-Zeiten verhindert werden.
Eine Empfehlung zu iSCSI kann noch nicht gegeben werden. Es müssen vorher noch weitere Erfahrungen gesammelt werden.

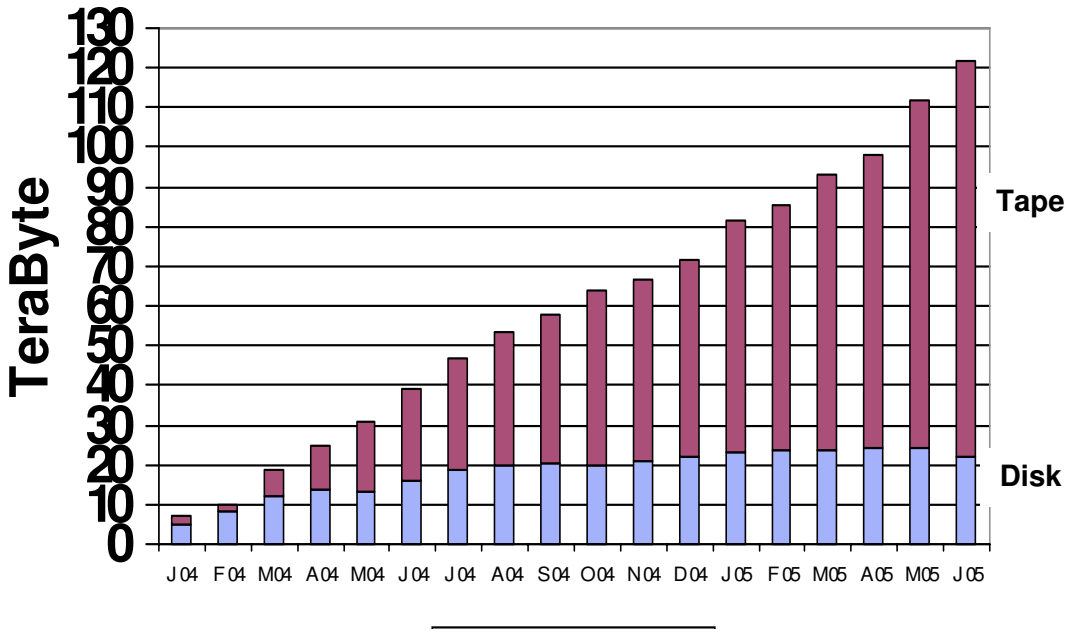
Teil 2

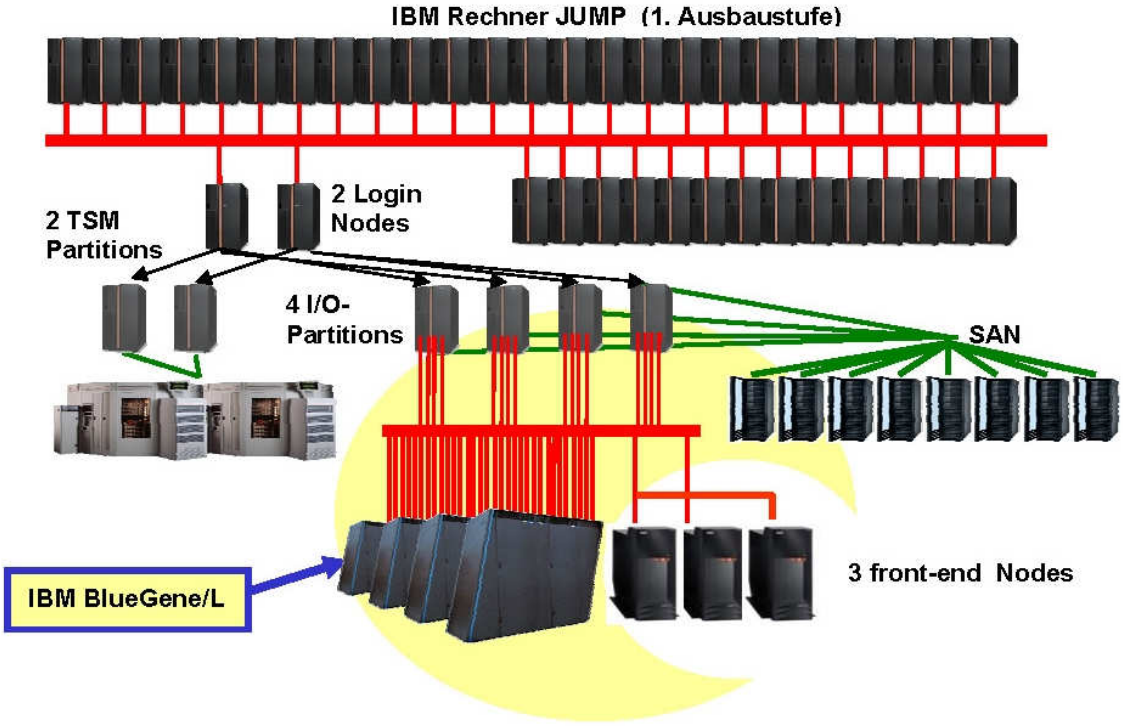
Storage für Höchstleistungsrechner und Cluster-Computing

STORAGE DES IBM SUPERCOMPUTERS

Die Daten des IBM Supercomputers JUMP liegen auf Online Storage und auf Bändern im Roboter. Der Online Storage besteht aus IBM FastT RAID Systemen mit ca. 56 TB Plattenplatz für Benutzerdaten. Die FastT sind über ein Fiber Channel SAN mit vier IO-Knoten auf dem Jump verbunden. Die IO-Knoten stellen allen anderen Jump Knoten den Speicherplatz über „virtuell shared disk (VSD)“ und den „High Performace Switch (HPS)“ zur Verfügung. Alle VSD-Clients sehen virtuelle Platten. IO-Zugriffe eines VSD-Clients werden durch den HPS zum VSD-Server geleitet und dann vom IO-Knoten bedient. Als Filesystem wird das „Global Parallel Filesystem (GPFS)“ von IBM genutzt. GPFS sorgt durch Locking-Mechanismen dafür, dass parallele Zugriffe auf die gleiche Datei sequentiell verarbeitet werden und die Konsistenz des Filesystems jederzeit gesichert ist. Das GPFS Filesystem wird mit der „Hierarchical Storage Manager (HSM)“ Komponente des „Tivoli Storage Manager (TSM)“ verwaltet. Diese Funktion lagert automatisch Dateien vom Plattenspeicher auf Roboterbänder nach bestimmten Kriterien aus. Die Kriterien sind z.B. Alter und Größe der Dateien. Nach der Auslagerung werden die eigentlichen Daten einer Datei auf dem Plattensystem gelöscht, d.h. es entsteht wieder freie Plattenkapazität. Im Dateiverzeichnis des Filesystems wird vermerkt, dass sich die Datei auf Bändern befindet. Die Daten einer migrierten Datei werden automatisch und unsichtbar für den Benutzer wieder vom Band zurückgeladen, wenn er diese in Programmen verwendet. Zur Datenspeicherung wird ein Robotersystem mit einer Kapazität von z.Zt. 2,8 PetaByte verwendet. Damit kann das Datenwachstum abgefangen werden, ohne regelmäßig neue Plattenspeicher kaufen zu müssen. Die folgende Grafik zeigt das Wachstum der Benutzerdaten seit Inbetriebnahme des JUMP-Systems. Zurzeit werden monatlich 10 TB neue Daten in das System eingespeist. Das Zusammenspiel zwischen GPFS und HSM funktioniert problemlos.

Datenspeicherung bei JUMP





Zur Anbindung weiterer Supercomputer und Cluster ist geplant, das GPFS als zentrales gemeinsames Filesystem für alle Systeme einzusetzen. Dieser Plan wurde im Rahmen des DEISA Projektes mit dem Multicluster GPFS (GPFS Version 2.3) erprobt. Mit dieser GPFS Version ist es möglich sowohl von AIX-Clustern als auch von Linux-Clustern auf dasselbe Filesystem zuzugreifen. Die Geschwindigkeit hängt von dem verwendeten Netzwerkmedium ab. Beim Einsatz des HPS im AIX ist ein Durchsatz von 1.5 GB/sec gemessen worden, beim Einsatz von Gigabit Ethernet ist entsprechend der Netzwerkgeschwindigkeit 120 MB/sec gemessen worden.
Zurzeit wird dieses Konzept für den neuen Supercomputer JUBL untersucht. Dabei sind die Lizenzkosten ein wichtiger Punkt, da beim derzeitigen Lizenzmodell die Kosten proportional mit der Anzahl der angeschlossenen CPUs anwachsen.
Weitere Untersuchungen mit einem Linux-Cluster sollen folgen.
Sollte sich dieses Konzept verwirklichen lassen, haben durch den Einsatz des Multicluster GPFS die Benutzer der Höchstleistungsrechner auf jedem System die gleiche Datensicht und parallelen Zugriff auf ihre Daten von allen Supercomputern.
Das GPFS wird mit der TSM/HSM Software verwaltet, so dass der verfügbare Platz für die Benutzer nur durch die Kapazität der Bandroboter eingeschränkt ist. Geplant ist der Anschluss eines Multi-Petabyte Systems um den großen Anforderungen der Benutzer gerecht zu werden.

Teil 3

Band Storage

KONZEPT FÜR DEN EINSATZ VON BANDROBOTER-SYSTEMEN

Die Bandroboter-Systeme werden im Forschungszentrum für drei zentrale Dienste eingesetzt:

1. Backup/Restore
   Es werden die wichtigen Daten der Rechner in den OE, die Daten der zentralen und dezentralen Server und die Daten der Supercomputer-Nutzer regelmäßig gesichert. Die Daten werden, wenn ein Mal eingerichtet, automatisch, d.h. ohne weiteren Aufwand durch den Nutzer, gesichert. Diese Sicherung schützt den Nutzer bei Ausfall des lokalen Speichermediums, auf dem seine Originaldaten liegen und es schützt den Nutzer bei unbeabsichtigtem Löschen von Dateien. Für den Einsatz im wissenschaftlichen Umfeld ist es notwendig, dass die eingesetzte Sicherungssoftware Funktionen zur Verfügung stellt, die es dem Nutzer selbst gestatten, seine Daten wiederherzustellen ohne Inanspruchnahme eines Administrators. Zum Schutz vor logischen Fehlern, wird mehr als eine Generation von Originaldaten im Robotersystem gespeichert. Bei fehlerhaften Applikationen, die die Originaldaten verfälschen, können daher zeitlich zurückliegende Generationen durchsucht werden, bis eine zeitnächste, gültige Version der Sicherungsdaten gefunden wird.

2. Archivierung
   Im Gegensatz zur Datensicherung werden nach dem vollständigen Kopieren der Originaldaten vom lokalen Datenträger in das Robotersystem die Daten auf dem lokalen System gelöscht. Die Archivierung von Daten wird daher für große, nur noch selten benutzte Dateien oder für Dateibäume (eine hierarchisch angeordnete Menge von Einzeldateien) von abgeschlossenen Projekten sinnvoll eingesetzt. Die Archivierung bremst die Nachfrage nach Kapazitätserweiterung auf den lokalen Plattensystemen und nimmt wichtige Datenbestände aus dem sich schnell ändernden Umfeld der DV des täglichen Betriebs.
   Bei der Archivierung nehmen die Daten im Robotersystem den Status von Originaldaten an, daher müssen diese Daten zusätzlich zur Archivierung mit der Backup-Funktion (s.o.) in ein, meist zwei Generationen gesichert werden.

3. Migration von aktiven (online) Daten auf preisgünstigere Speichermedien
   Seit der Inbetriebnahme des Supercomputers JUMP wird im Forschungszentrum Jülich auch die HSM-Funktion (Hierarchical Storage Manager) erfolgreich bei der Speicherung

von Daten eingesetzt. Diese Funktion lagert automatisch Dateien vom Plattenspeicher auf Roboterbänder nach bestimmten Kriterien aus. Die Kriterien sind z.B. Alter und Größe der Dateien. Nach der Auslagerung werden die eigentlichen Daten einer Datei auf dem Plattensystem gelöscht, d.h. es entsteht wieder freie Plattenkapazität. Im Dateiverzeichnis des Betriebssystems wird vermerkt, dass sich die Datei auf Bändern befindet. Die Daten einer migrierten Datei werden automatisch und unsichtbar für den Benutzer wieder vom Band zurückgeladen, wenn er diese in Programmen verwendet. Die migrierten Daten sind Originaldaten und müssen daher immer noch in vier Generationen mit der Backup-Funktion gesichert werden. Die HSM-Funktion stellt mit der Verbindung von Plattensystemen und Roboterbändern einen fast unendlich großen, virtuellen Online-Speicherraum zur Verfügung. Die HSM-Funktion greift stark in das jeweilige Betriebssystem und seines Filesystems ein; der Hersteller (IBM) arbeitet derzeit auch an einer Linux-Version, so dass HSM-Funktionen mittelfristig auch für die vom ZAM betreuten Workstation-Gruppen eingesetzt werden können.

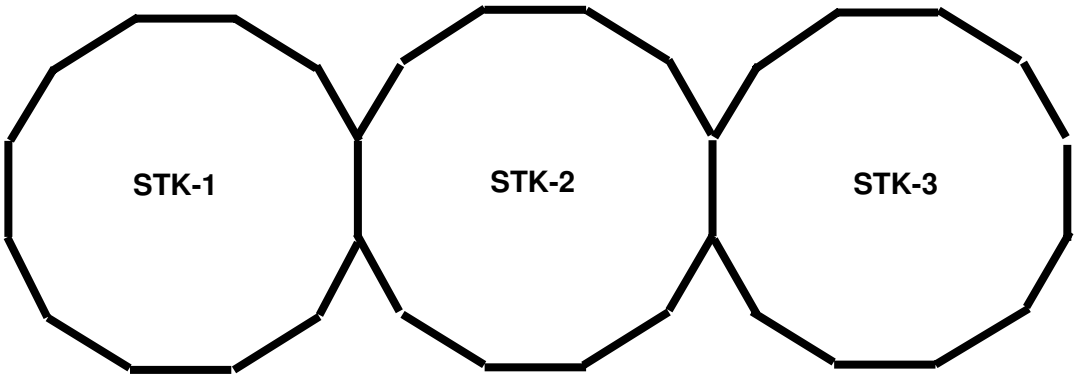
Das angestrebte Ziel für den Einsatz von Bandroboter-Systemen ist, für die beschriebenen Funktionen (Backup, Archiv, HSM) ausreichend Kapazität zur Verfügung zu stellen und durch Speicherung von Originaldaten und Backup-Kopien in getrennten Gebäuden eine hohe physikalische Sicherheit zu erreichen.

IST-STAND DER ROBOTERSYSTEME

Derzeit betreibt das ZAM vier Robotersysteme. Ein System ist vom Hersteller IBM mit 150 TByte Kapazität und dient für Backup und Archivierung der Dateien von Rechnern auf dem Campus und von Dateien der Server-Systeme, die z.T. im selbem Maschinenraum untergebracht sind (s. Risikobewertung des ZAM vom 1.7.2004). Die Kapazität dieses Systems wird im September 2004 vollständig aufgebraucht sein.

Die drei Robotersysteme der Firma STK (s. Abb.) haben sich in der Zeit ihres Einsatzes als sehr flexibel erwiesen. So konnten in Abständen von ca. 4 bis 6 Jahren einzelne Silos mit neuer Hardware (Mechanik und Laufwerke) und neuen Bändern ausgestattet werden, so dass in einem rollierenden Verfahren eine laufende Erneuerung nach dem Stand der Technik möglich war.

Die Silos STK-1, STK-2 und STK-3 wurden ab 2000 für DMF, eine CRAY-Variante des HSM, für die Benutzer des CRAY-Komplex eingesetzt.



Letzter Umbau:	2004	Letzter Umbau:	2002	Letzter Umbau:	2006
Laufwerke:	8 (9940b)	Laufwerke:	8 (9940b)	Laufwerke:	0
Bänder:	5.500	Bänder:	5.500	Bänder:	3.000
Kapazität:	1,1 PByte	Kapazität:	1,1 PByte	Kapazität:	0,6 PByte
Nutzung:	FZJ/SC	Nutzung:	SC	Nutzung:	FZJ

Mit dem IBM Supercomputer JUMP wurde ab 2002 schrittweise der CRAY-Komplex (2 x T3E, 1 x SV1, 1 x J90) zurückgebaut. Die früheren CRAY-Nutzer haben ihre langfristig benötigten Daten zum IBM Supercomputer transferiert. Bis zum Ende der Lebensdauer des CRAY-Vektorrechners SV1 wurde daher nur noch das Silo STK-3 genutzt.

Das Silo STK-2 wurde mit der Installation des IBM Supercomputers JUMP auf den neuesten Stand gebracht und mit Laufwerken und Bändern höchster Speicherdichte ausgestattet. Genutzt wird das Silo für HSM und Backup der Daten des Supercomputers.

Nachdem das Silo STK-1 durch Kopieren der CRAY-Dateien zum IBM Supercomputer freigemacht worden war, wurde dieses Silo ebenfalls auf den neuesten Stand gebracht und die vorhandenen Laufwerke mit höchster Speicherdichte auf beide Silos verteilt. Zusätzlich wurde das Silo mit neuen Bändern bestückt, so dass insgesamt 11.000 Bändern mit einer Gesamtkapazität von 2.2 PByte zur Verfügung standen.

Nach Abbau des CRAY-Vektorrechners SV1 wurde auch das Silo STK-3 mit 3.000 Bändern mit einer Speicherkapazität von je 200 GByte bestückt, so dass zurzeit im STK Komplex 14.000 Bänder mit einer Kapazität von 2.8 PByte vorhanden sind.

KAPAZITÄTSBEDARF

Im Backup wird der wachsende Bedarf durch die immer größer werdenden Plattenlaufwerke bei den Arbeitsplatzrechnern und Abteilungsservern bestimmt. Hier ist im FZJ (wie aber auch weltweit) ein Faktor von 2 pro Jahr sichtbar, also eine Verdoppelung der notwendigen Kapazität pro Jahr. Es hat sich gezeigt, dass für unser Profil von Daten eine Komprimierung der Daten bei der Speicherung keinen Gewinn bringt.

Die Archivierungsfunktion belegt 50 TByte und wird von einigen Großbenutzern aus IME, IKP und IPP dominiert. Zur Zeit verdreifacht sich dieser Bestand pro Jahr. Neben dem „kontinuierlichen“ Wachstum wird der Kapazitätsbedarf aber auch durch neue Projekte, wie das EU-Projekt I3HP, das International Lattice DataGrid (ILDG) und die geplante Speicherung von Experimentdaten von COSY, bestimmt.
Die Komprimierung der Daten kann hier das Wachstum etwas abmildern.

Bei HSM und Backup für den IBM Supercomputer JUMP kann seit der Produktionsfreigabe Ende Januar 2004 ein Wachstum von 40 TByte pro Monat beobachtet werden. Diese Zahl setzt sich aus 10 TByte Originaldaten, 10 TByte für die Spiegelung der Daten, 10 TByte für den Voll-Backup und noch mal ca. 10 TByte für die 4 Generationen des inkrementellen Backup zusammen.

MITTELFRISTIGE KAPAZITÄTSERWEITERUNG UND ERHÖHUNG DER SICHERHEIT

Für einen erhöhten Schutz gegen Katastrophen, zumindest für die Archivdaten, werden in Zusammenarbeit mit dem Rechenzentrum der RWTH Aachen Sicherheitskopien der Archive von FZJ und RWTH wechselseitig beim jeweiligen Partner gespeichert. Wegen der großen Datenmenge ist dies aber für die Backup-und Migrations-Daten nicht möglich. Diese werden derzeit in einem Robotersystem gespeichert.

Es ist geplant, das Großgerät „IBM Supercomputer“ im Jahre 2007 zu erweitern. Im Zuge dieser Investition wird sich die Online-Plattenkapazität von derzeit 75 TByte wahrscheinlich verachtfachen verbunden mit den entsprechenden Kapazitätsanforderungen für HSM und Backup. Hier bietet sich dann die Beschaffung eines neuen Robotersystems an, das dann räumlich getrennt von den anderen in der Maschinenhalle des ZAM-Neubaus installiert werden kann. Bei einer entsprechenden Verteilung von HSM (Originaldaten) und Sicherungskopien kann eine notwendige Verbesserung der physikalischen Sicherheit erreicht werden. Gleichzeitig

hat das derzeitige IBM Robotersystem das Ende seiner Lebensdauer erreicht. Hierbei ist dann besonders der Verlust der 16 Bandlaufwerke für den Gesamtdurchsatz schädlich, so dass insbesondere der Bedarf an neuen Bandstationen groß ist.

TSM BACKUP KONZEPT

Die Datensicherung mit dem Programm Produkt *Tivoli Storage Manager* (TSM) ist ein zentraler Dienst für Rechner im JuNet des Forschungszentrum Jülich. Unter der Datensicherung versteht man das Anlegen von Kopien der Daten eines Systems, wobei die Originaldaten auf dem System erhalten bleiben. Diese Maßnahme soll den Benutzer vor dem Verlust seiner Daten schützen, sei es durch versehentliches Löschen oder Überschreiben, oder durch Hardware-Ausfall der Platte, auf der die Daten gespeichert sind.

Prinzipiell werden alle lokalen Daten des Systems (Workstation/PC) gesichert, um Fehler durch den Administrator/Benutzer zu verhindern (falsche Exclude Angaben, fehlende Domain Angaben). Der Administrator/Benutzer hat zusätzlich die Möglichkeit den Umfang der zu sichernden Daten seinen Bedürfnissen anzupassen. Die dazu notwendige Kapazität an Sicherungsspeicher konnte bis dato immer zur Verfügung gestellt werden, da die Hardware-Entwicklung der Bandspeicher synchron mit der Entwicklung bei Platten z.B. bei PCs verlief. Es wird davon ausgegangen, dass das auch mittelfristig noch der Fall sein wird.

Die Sicherung der Daten erfolgt auf Dateiebene, wobei pro Datei eine Kopie angelegt wird, die an zentraler Stelle außerhalb der Workstation/PC gehalten wird. Je Datei werden zurzeit im Normalfall bis zu vier Kopien gehalten. Wird eine weitere Kopie angelegt, so wird jeweils die älteste Kopie überschrieben (Generationenkonzept). Die Kopien können inhaltlich identisch oder verschieden sein, was von dem gewählten Sicherungsmodus abhängt.

Man unterscheidet zwischen *inkrementeller Sicherung* und *expliziter Sicherung* der Daten.

- Bei der *inkrementellen Sicherung* werden bei jedem Sicherungslauf Kopien von denjenigen Dateien angelegt, die neu sind oder sich gegenüber der Kopie vom vorherigen Sicherungslauf verändert haben. Die zuletzt angelegte Kopie (aktive Kopie) einer Datei enthält somit den aktuellen Stand der Daten zum Zeitpunkt der Sicherung; die zuvor angelegte Kopie den vorherigen Stand der Daten.
- Bei der *expliziten Sicherung* wird von allen zu sichernden Dateien eine neue Kopie angelegt. Dabei spielt es keine Rolle, ob die Datei neu ist, verändert wurde oder gegenüber einer vorherigen Sicherung unverändert ist. Somit können die Sicherungskopien der Daten identisch sein.

Von Dateien, die auf der Workstation/PC gelöscht wurden, werden die Sicherungskopien entsprechend markiert (inaktive Kopien) und nach 4 Wochen ebenfalls gelöscht.

Die Datensicherung kann automatisiert werden, wobei der TSM Backup Server innerhalb eines definierten Zeitintervalls (Schedule) die Sicherungskopien von der Workstation/PC zieht. Dieser Vorgang wird zentral überwacht und der eingetragene Administrator/Benutzer per Email benachrichtigt, wenn eine Datensicherung nicht erfolgreich gelaufen ist.

Die Datensicherung ist immer rechnerbezogen. Der Rechner muss in der zentralen JuNet Datenbank angemeldet sein. Die Installation und die automatische Datensicherung verlangen Root- bzw. Administrator-Rechte auf dem Unix- bzw. Windows-System. Der Administrator/Benutzer muss eine offizielle Email-Adresse des Forschungszentrums besitzen.

SPEZIELLE EMPFEHLUNGEN FÜR FILESERVER

Der wesentliche limitierende Faktor beim Restore von Sicherungsdaten ist die Zeit, die zur Erstellung der Inodes benötigt wird (Faustregel: 100 Inodes pro sec, abhängig von Prozessor, Memory und Netzanbindung). Für Systeme, die als Fileserver eingesetzt werden und daher im Fehlerfall die Daten möglichst schnell wieder herstellen müssen, sollten bzgl. der Datensicherung folgende Empfehlungen beachtet werden.

- *Collocation by Filesystem* sollte aktiviert sein. Die Sicherungsdaten eines Filesystems werden dabei auf möglichst wenigen Bändern zusammengefasst. Somit wird die Anzahl der Tape Mounts beim Restore der Daten minimiert. Die Eigenschaft *Collocation by Filesystem* lässt sich nur TSM serverweit[1] nicht individuell für einen Teilnehmer an der Datensicherung einstellen.
- *Paralleles Restore*[2] der Daten eines Filesystems sollte genutzt werden, wenn sich die Sicherungskopien auf mehr als einem Band befinden. Der TSM Server ermittelt unter Berücksichtigung der Verteilung der Sicherungsdaten auf die Bänder und der übrigen Systemlast die Anzahl der parallelen Restores.
- Registrierung der Fileserver immer auf den neuesten TSM Servern, um die maximale Performance bzgl. Software- und Hardware-Technologie zu erhalten.

1 Diese Eigenschaft wird als Option des Platten Cache Bereichs des logischen TSM Servers definiert (DEFine STGpool ... Collocate=Filespace).

2 Die Berechtigung zur Nutzung dieser TSM Funktion muss bei der Registrierung des Systems mit einem oberen Threshold bzgl. der Anzahl paralleler Restores (MAXNUMMP=*n*) eingetragen werden. Aktiviert wird ein paralleles Restore durch Nutzung der *Resourceutilization*-Option in der TSM Kontrolldatei des Klients.

Die Datensicherung mit TSM ist in diesem Zusammenhang nur ein Hilfsmittel. Die Hardware von Server-Systemen muss entsprechend der Wichtigkeit ausgelegt sein (RAID, Spiegel, redundanter Kontroller, …).

Kriterien für einen Server sind:

- Anzahl der unterstützten Benutzer
- Tolerable Ausfallzeit
- Wichtigkeit des Rechners für das Unternehmen

ZUGRIFFSRECHTE AUF SICHERUNGSDATEN

Auf Sicherungsdaten beim TSM Backup Server sollte generell nur lesend zugegriffen werden können. Die Löschberechtigung wird daher zum Selbstschutz nicht erteilt.

Bei Unix Systemen kann der Benutzer root auf alle Sicherungsdateien aller Benutzer zugreifen, ein Benutzer jedoch nur auf seine eigenen und solche für die er Leserecht hat. Bei Windows Systemen kann jeder auf alle Sicherungsdateien des Systems zugreifen.

Sollen Mitbenutzer Zugriff erhalten, wird empfohlen die TSM Funktionen zur Definition von Rechten[3] zu nutzen. Diese erlauben die Rechte granular zu vergeben (rechnerbezogen, benutzerbezogen und dateibezogen). Soll der Zugriff von einem anderen Rechner aus erfolgen, muss dieser Rechner ebenfalls an der Datensicherung teilnehmen. Dieses Verfahren wird beispielsweise in den vom ZAM administrierten Workstationgruppen praktiziert.

Die Freigabe von Sicherungskopien sollte keinesfalls durch Weitergabe des Rechnernamens und des TSM Backup-Passwords an andere erfolgen.

LÖSCHEN DER SICHERUNGSKOPIEN ALTER FILESYSTEME/FILESPACES

Die Sicherungskopien alter Unix Filesysteme oder Windows Filespaces bleiben nach TSM Policy unbefristet im TSM Sicherungsspeicher liegen, da TSM zum Zeitpunkt der Datensicherung nicht weiß, ob diese Filesysteme/Filespaces nur temporär nicht verfügbar sind oder gar nicht mehr existieren. Sicherungskopien von Filesystemen/Filespaces, die länger als ein Jahr nicht mehr gesichert wurden, sollen regelmäßig gelöscht werden.

3 Mittels *User Access List* (GUI) bzw. *dsmc set access* (Command).

VERSCHLÜSSELUNG VON SICHERUNGSDATEN

Sollen personenbezogene Daten auf dem System verarbeitet werden, so müssen diese laut *IT-Sicherheitsregeln für den Grundschutz* - Regel D4 - verschlüsselt werden. Diese Daten müssen daher schon im Original auf dem Rechnersystem verschlüsselt gespeichert sein, damit sind dann auch automatisch die Daten der Sicherungskopien verschlüsselt.

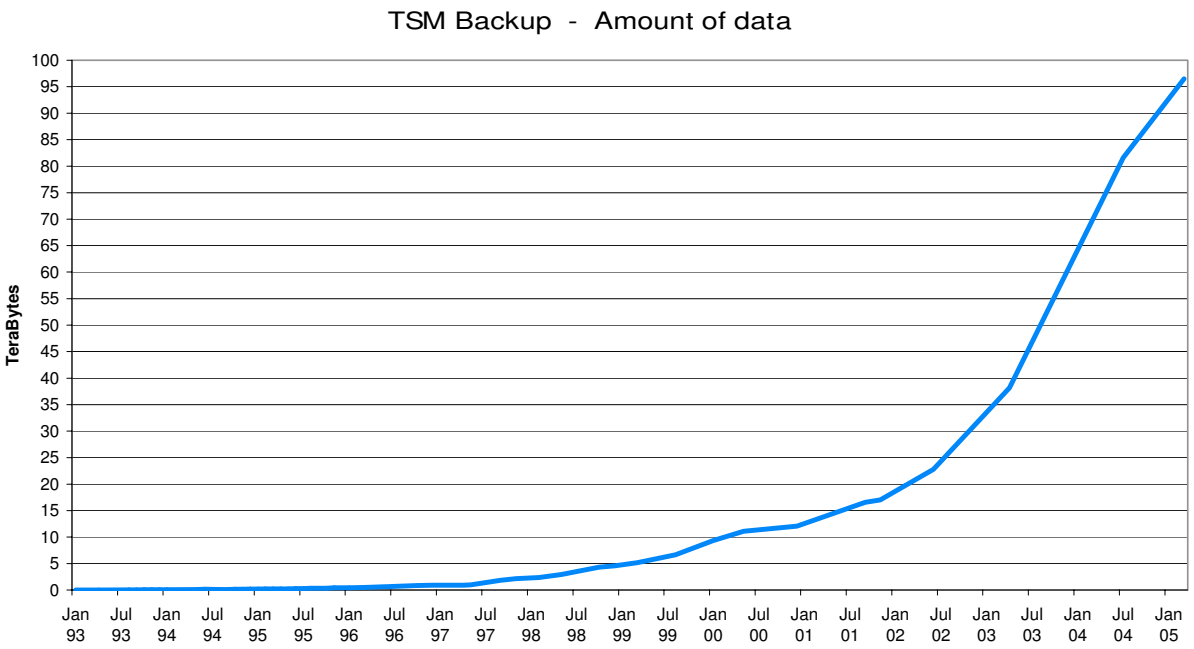
ERGÄNZENDE ZUKÜNFTIGE AUFGABEN

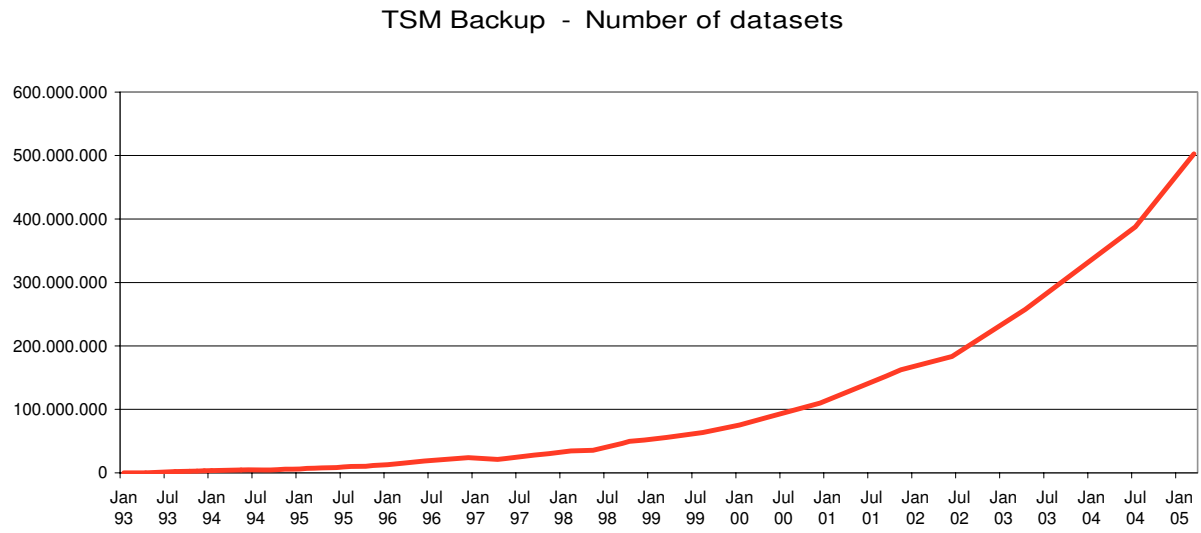
Für die vollständige Realisierung des TSM Backup Konzeptes müssen ab 2006 die folgenden Funktionen in die zentrale Benutzerverwaltung integriert werden:

- Beantragung der Datensicherung
- Backup-Password zurücksetzen
- Löschen von Berechtigungen zur Datensicherung
- Sperren/Entsperren einer Berechtigung zur Datensicherung
- Übertragen von Berechtigungen zur Datensicherung

Details: siehe Anhang 3.

ENTWICKLUNG DER NUTZUNG DES BACKUPS





TSM ARCHIV KONZEPT

Das im Forschungszentrum eingesetzte Programm Produkt *Tivoli Storage Manager* (TSM) bietet neben der Funktion der Datensicherung die Möglichkeit Daten zu archivieren. Im Gegensatz zur Datensicherung muss die Archivierung einer Datei vom Benutzer selbst durchgeführt werden. Unter Archivierung wird in der Regel eine Langzeitspeicherung der Daten im Archiv-Speicher verstanden, wobei die Originaldaten weiterhin existieren können, nicht mehr existieren müssen oder einen geänderten Zustand haben können. Bei der Archivierung laufen also keine automatischen Prozesse. Die Kontrolle bei der Archivierung unterliegt prinzipiell dem Benutzer des Archives. Dieser entscheidet wann welche Daten in das Archiv abgelegt oder daraus gelöscht werden.

Das Konzept unterscheidet zwischen *persönlichen Archiven*, die an einen Benutzer gebunden sind und *Langzeitarchiven* z.B. von Projektdaten oder Messdaten eines Instituts und/oder einer Gruppe. Da TSM nicht gestattet, ein Archiv in ein bereits existierendes Archiv zu übertragen, gibt es zusätzlich *persönliche übernommene Archive*. Diese sollen jedoch nur zeitlich begrenzt zur Verfügung stehen.

- *Persönliche Archive* sind an einen Benutzer gebunden und werden mit der offiziellen Email-Adresse des Benutzers (z.B. `G.Mustermann`) eingerichtet. Das Archiv ist somit auf Grund der Vergabe-Policy der Email-Adresse im Forschungszentrum Jülich eindeutig.

- *Langzeitarchive* werden mit einem durch den Betreuer des Archives frei wählbaren Namen, an den als Endung das Institutskürzel (z.B. `projekt1_daten.zam`) angehängt wird, eingerichtet. Dieser Name muss im TSM Archiv des Forschungszentrum Jülich eindeutig sein. Dies wird bereits bei der Antragstellung überprüft.
  Neben diesen Langzeitarchiven wird die Funktion des TSM Archivs im Zusammenhang mit der Projektbezogenen Datenhaltung (siehe auch Teil 4) genutzt.
- *Persönliche übernommene Archive* sind persönliche Archive, die bei Ausscheiden des Nutzers nicht aufgelöst werden konnten und dem DV-Ansprechpartner zeitlich begrenzt zur Verfügung gestellt werden, damit dieser die Archivdaten sichtet, die benötigten Daten zurückholt und in seinem persönlichen Archiv speichert. Diese Archive werden durch eine angefügte Nummerierung an den Namen des persönlichen Archives (z.B. `G.Mustermann_1`) gekennzeichnet.

ZUGRIFFSRECHTE AUF ARCHIVE

Der Benutzer/Betreuer eines Archivs soll im Allgemeinen alle Zugriffsrechte auf das Archiv besitzen. Er kann alle Daten lesen und auch löschen. Zurzeit gibt es auf Grund von Softwaremängeln aus sehr alten ADSM/TSM Versionen die Einschränkung, dass einige Benutzer ihre Archivdaten zum Selbstschutz nicht löschen dürfen. Dies soll mit Hilfe der Benutzer bereinigt werden.

Der Verlust von Archivdaten durch versehentliches Löschen der Archivdaten soll nicht durch eine zusätzliche Backup-Kopie abgesichert werden. Ein explizites Löschen von Archivdaten bewirkt, dass alle Kopien (3) gleichzeitig gelöscht werden. Abgesichert wird durch die Anzahl der Kopien nur der Verlust der Daten durch Media-Fehler (Bänder).

Bei Langzeitarchiven kann es sinnvoll sein, dass neben dem Betreuer weitere Personen auf die Archivdaten zugreifen dürfen. Sollen die Mitbenutzer nur lesenden Zugriff erhalten, wird empfohlen die TSM Funktionen zur Definition von Rechten[4] zu nutzen. Diese erlauben die Rechte granular zu vergeben (benutzerbezogen und dateibezogen). Die Mitbenutzer benötigen für diese Zugriffsart ein eigenes persönliches Archiv. Es wird nicht empfohlen das Langzeitarchiv generell durch Weitergabe des Archivnamens und des Archiv-Passwords an andere freizugeben.

VERSCHLÜSSELUNG VON ARCHIVDATEN

Sollen personenbezogene Daten im Archiv abgelegt werden, so müssen diese laut *IT-Sicherheitsregeln für den Grundschutz* - Regel D4 - verschlüsselt werden. Je nach Sensitivität kann es auch aus anderen Gründen bei Langzeitdaten sinnvoll sein, die Daten zu

4 Mittels *User Access List* (GUI) bzw. *dsmc set access* (Command).

verschlüsseln. TSM bietet keine Möglichkeit Archivdaten auf Benutzerebene individuell zu verschlüsseln, sondern nur auf Administratorebene des TSM Client Systems[5]. Der Benutzer verpflichtet sich in diesen Fällen personenbezogene Daten und sensitive Daten vorab zu verschlüsseln und erst dann ins Archiv zu schreiben.

ERGÄNZENDE ZUKÜNFTIGE AUFGABEN

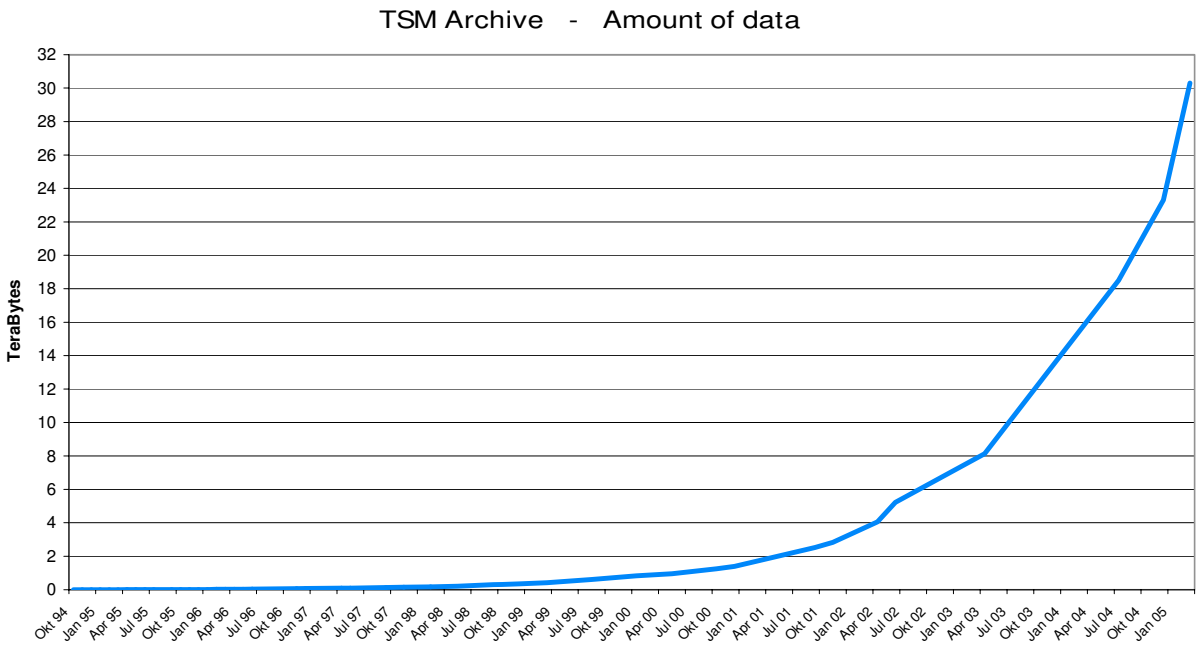
Für die vollständige Realisierung des TSM Archiv Konzeptes müssen ab 2006 die folgenden Funktionen in die zentrale Benutzerverwaltung integriert werden:

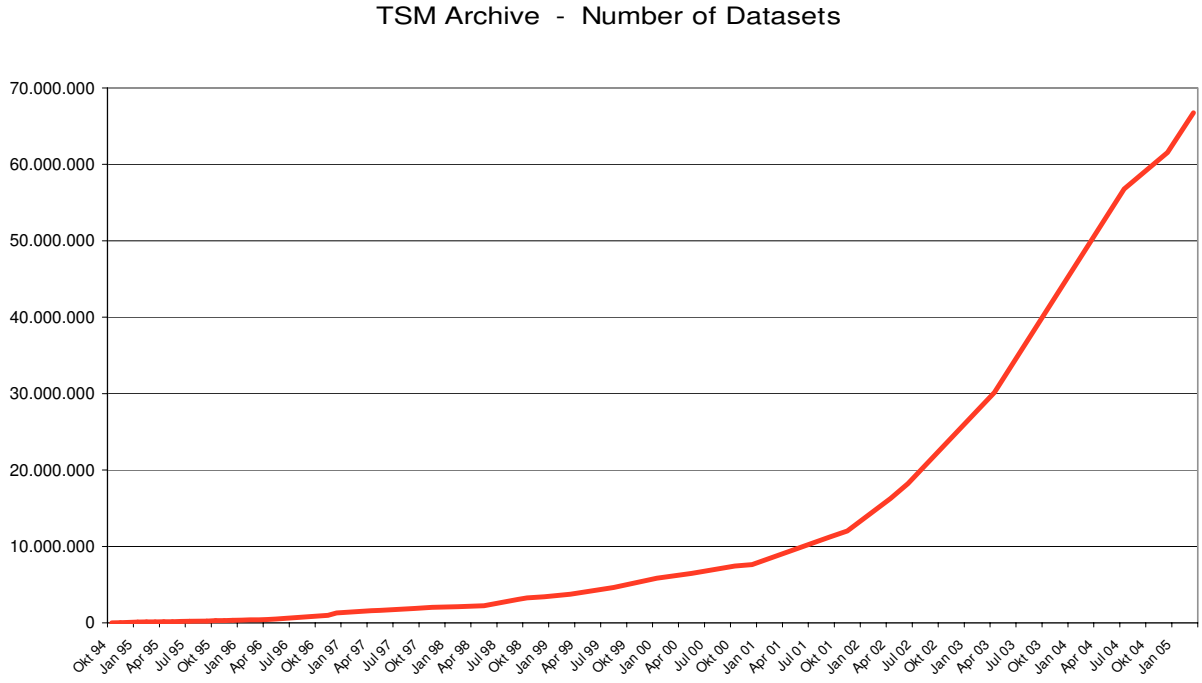
- Beantragung von Archiven
- Archiv-Password zurücksetzen
- Löschen von Archiven
- Sperren/Entsperren eines Archivs
- Übertragen von Archiven

Details: siehe Anhang 4.

5 TSM sieht keinen zentralen Schlüssel vor, mit dem die Daten verschlüsselt ins Archiv geschrieben werden können. TSM bietet Verschlüsselungsmechanismen im Prinzip nur auf System-Administrator Ebene an. Bei Windows Systemen ist in der Regel der System-Administrator auch der Benutzer. Bei Unix-Systemen ist das in der Regel nicht der Fall. Der System-Administrator kann für das System die Verschlüsselung aktivieren (*encryptkey* Option). Dabei kann festgelegt werden, ob der Verschlüsselungs-Key wiederum verschlüsselt auf dem System abgelegt (*save*) oder immer danach gefragt (*prompt*) werden soll. Falls beim Save-Verfahren die Datei mit dem verschlüsselten Key (*TSM.PWD*) verloren geht, z.B. durch Reinstallation des Systems oder Plattenfehler ohne Sicherung, und der Verschlüsselungs-Key zwischenzeitlich vergessen wurde, sind die verschlüsselten Archivdaten nicht mehr nutzbar. Gleiches gilt, falls der Benutzer/Betreuer aus irgendwelchen Gründen beim Prompt-Verfahren den Verschlüsselungs-Key nicht mehr weitergeben kann. In diesen Fällen besteht auch für den TSM Administrator keine Möglichkeit die Daten in entschlüsselter Form wieder zur Verfügung zu stellen. Weiterhin besteht beim Prompt-Verfahren die Möglichkeit für verschiedene Dateien bei der Archivierung unterschiedliche Verschlüsselungs-Keys anzugeben. Diese Funktionalität ist durchaus in bestimmten Fällen sinnvoll, erhöht aber die Gefahr, die Daten durch Verlust der Verschlüsselungs-Keys zu verlieren.

ENTWICKLUNG DER NUTZUNG DES ARCHIVES

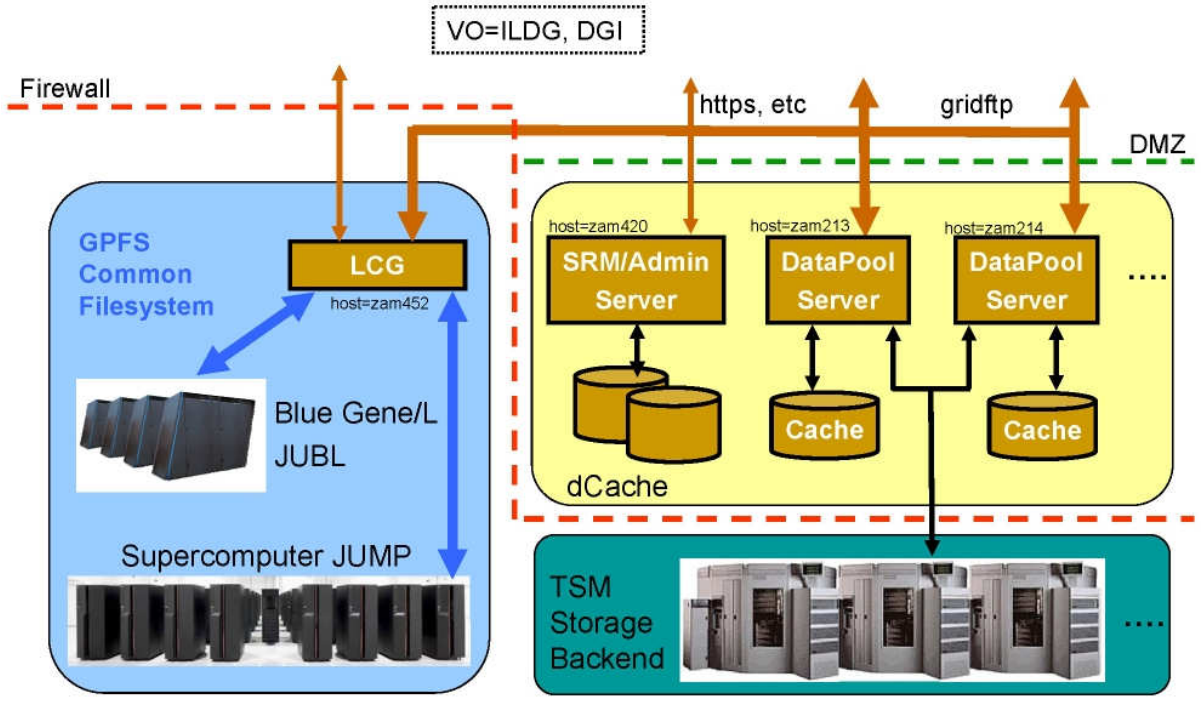




Teil 4

Projektbezogene Datenhaltung

Im Gegensatz zur benutzerbezogenen Datenhaltung, bei der die Existenz und die Lebensdauer von Daten immer an die Existenz und die Lebensdauer eines Benutzer-Accounts gebunden sind, wird die projektbezogene Datenhaltung in Zukunft eine neue Aufgabe für die wissenschaftlichen Rechenzentren darstellen. Benutzer stellen Ergebnisse ihrer Rechnungen einer Benutzergemeinschaft (Community) zur weiteren langfristigen Auswertung zur Verfügung. Die Daten können aus Messungen bei Experimenten stammen (im Forschungszentrum z.B. bei Textor oder dem Scout-Projekt).
Die Datenhaltung einer Community benötigt andere Lösungen für die Katalogisierung der Daten, die Zugriffsberechtigungen und die Zugriffstechniken als benutzerbezogene Daten.
Im Rahmen des BMBF-geförderten Projektes D-Grid erprobt das ZAM das Grid-fähige Datenhaltungssystem dCache. Als eine erste Anwendergruppe wird das System in 2006 von der ILDG (Int. Lattice Data Grid)-Gruppe zur Speicherung und Auswertung von Simulationsdaten aus dem Bereich QCD genutzt werden.



Das dCache-System, das seit einigen Jahren in der Hochenergiephysik eingesetzt wird, besteht aus einem Disk-Pool-Management-System, das mit einem einheitlichen Filesystem den Zugriff auf über mehrere Server verteilte Daten ermöglicht. Die Auslastung der Disk-Pools wird automatisch geregelt.

Das Storage-Element besteht aus drei Linux-Servern - einem dCache-Management-Server und zwei dCache-Pool-Servern. Die Pool-Server sind mit jeweils 1.5 TB-Plattenbereich als Cache ausgestattet und über eine HSM-Komponente mit dem Band-Robotersystem (von STK) verbunden. Die Implementierung der HSM-Komponente besteht auf der Client-Seite aus der API-Schnittstelle der Linux-TSM-Software. Auf der Server-Seite wurde ein virtueller TSM-Server auf einem bestehenden realen TSM-Serversystem unter AIX aufgesetzt.

Anhänge

ANHANG 1: KOSTEN FÜR FILESERVER-SYSTEME

KOSTEN FÜR PLATTENSPEICHER

Speicher	Brutto Kapazität	Netto Kapazität	Kosten Euro
RAID FC/FC (*)	2,3 TB (16x140GB)	1,7 TB (12x140 TB)	20000
Erweiterung FC/FC	2,3 TB (16x140GB)	1,7 TB (12x140 TB)	8500

Speicher	Brutto Kapazität	Netto Kapazität	Kosten Euro
RAID SATA/FC	6,4 TB (16x400GB)	4.8 TB (12x400 TB)	28800
Erweiterung	6.4 TB (16x400GB)	4.8 TB (12x400 TB)	10800
RAID SATA/FC	4.0 TB (16x250GB)	3.0 TB (12x250 TB)	25100
Erweiterung	4.0 TB (16x250GB)	3.0 TB (12x250 TB)	7100

Speicher	Brutto Kapazität	Netto Kapazität	Kosten Euro
RAID SATA/FC	6,4 TB (16x400GB)	4.8 TB (12x400 TB)	10000
RAID SATA/SCSI	6.4 TB (16x400GB)	4.8 TB (12x400 TB)	9000
RAID SATA/FC (*)	4.0 TB (16x250GB)	3.0 TB (12x250 TB)	7000
RAID SATA/SCSI	4.0 TB (16x250GB)	3.0 TB (12x250 TB)	6000

SATA-RAIDs, nicht erweiterbar, ein Kontroller, 3 Jahre Garantie

SATA-RAIDs, erweiterbar bis 128 Platten, redundanter Kontroller, 5 Jahre Garantie

FC-RAIDs, erweiterbar bis 128 Platten redundanter Kontroller, 3 Jahre Garantie

KOSTEN FÜR SERVER

Linux Xeon 2.8 GHz (*)	2 Proc, 4 GB Hauptspeicher		3000
Linux Qlogic FC-Karte (*)	2 Gbit		1000
AIX IBM P615 (obsolet) (*)	1 Proc, 1 GB Hauptspeicher	(Preis 2003)	4000
AIX IBM p5 520	2 Proc, 4 GB Hauptspeicher		8000
AIX Cambex FC-Karte (*)	2 Gbit		2000
Sun Sprac IIIiV420	2 Proc, 4GB Hauptspeicher		4000
Sun FC Karte	2 Gbit		400

KOSTEN FÜR FC-SWITCH

McData Sphereon 4500 (*)	8 Port	5500

Die mit (*) bezeichneten Komponenten wurden für die Untersuchungen benutzt.

ANHANG 2: ERGEBNISSE DER MESSSUNGEN BEI FILESERVERN

Performance-Messungen bei Infortrend-Raids Infortrend mit bonnie++

Machine	-Per Char- K/sec	%CPU	--Block--- K/sec	%CPU	-Rewrite-- K/sec	%CPU	-Per Char- K/sec	%CPU	--Block--- K/sec	%CPU	--Random- /sec	%CPU
Linux zam986 mit FC/FC-Raid												
XFS												
zam987fx	39805	99	139320	33	50948	14	42726	98	120753	14	744.5	1
zam987fx	40580	99	143418	34	49481	14	41901	98	114908	15	790.3	
JFS												
zam987fj	38170	99	121658	32	46684	10	42022	98	108783	12	630.0	1
zam987fj	37920	99	117888	30	46537	10	41740	97	107487	12	582.3	
Reiser												
zam987fr	34678	99	110539	46	40635	14	39907	96	109968	18	626.0	2
zam987fr	35346	99	111027	48	41089	14	41005	96	110873	18	641.9	2
EXT3												
zam987f3	34080	97	100869	48	39305	12	40553	95	113878	13	579.3	1
zam987f3	33929	98	106944	49	39288	11	41308	95	112250	13	589.0	1
Linux zam986 mit SATA/FC-Raid												
XFS												
zam987sx	38158	99	81136	19	33816	9	40665	98	86077	10	526.0	1
zam987sx	39650	99	83653	20	34074	9	42045	98	84270	12	527.2	1
JFS												
zam987sj	38169	99	77830	19	34285	7	42830	98	88320	10	453.5	0
zam987sj	38317	99	75957	19	34137	8	42950	98	86762	11	425.2	0
Reiser												
zam987sr	35267	99	80016	33	31992	10	41453	96	85201	14	460.5	1
zam987sr	35093	99	73744	31	32880	11	41106	96	84025	15	445.5	1
EXT3												
zam987s3	34291	97	84191	45	34333	10	41501	94	88102	11	423.9	1
zam987s3	34782	99	77798	35	32911	10	41905	95	87920	11	416.7	0
Linux ibi300 mit SATA/SCSI-RAIDs												
XFS												
ibi300sx	38564	99	82321	22	31542	9	40545	97	74044	9	526.6	1
ibi300sx	38607	100	80596	22	31473	9	40805	96	71729	9	533.0	1
Reiser												

```
-------------------------------------------------------------------------------
ibi300sr 33511 100  70736  32  30060  11   37760  92   68427  11   578.5   1
ibi300sr 34205  99  79747  36  28471  10   31177  75   49349   8   574.5   1

-------------------------------------------------------------------------------
Solaris zam188 mit SATA/SCSI-RAIDs

zam188sv 25308  97  59001  40  16485  22   26806  93   60171  32   306.2   2
-------------------------------------------------------------------------------
Solaris zam188 mit SATA/SCSI-RAIDs

zam188fv 24816  96  67816  47  23437  32   27551  96   76194  41   452.7   4


AIX zam271 mit FC/FC-Raid
-------------------------------------------------------------------------------
zam2718f 34076  93  42929  18  29166  14   33407  90   97142  24   273.1   1
zam2718f 34153  93  42959  18  28677  14   33415  90   98159  26   271.2   1

AIX zam271 mit SATA/FC-Raid
-------------------------------------------------------------------------------

zam2718s 34461  93  46036  19  29832  14   33671  90   87162  20   212.2   1
zam2718s 34402  94  46171  19  29825  14   33915  91   88227  21   211.3   0

-------------------------------------------------------------------------------
Messungen über NFS Server zam987 sles9 FC/FC XFS 1 Filesystem
-------------------------------------------------------------------------------
1 Klient
zam986f1 18843  63  24225  14  16743  15   30517  90   42563  12   329.9   1

2 Klienten
zam986f2 17698  59  21051  12  14935  13   30039  90   25795   7   224.2   0
zam420f2  9982  26  10371   2   2953  67   11631  29   11123   3   1193    6

3 Klienten
zam986f3 16107  54  18497  10  11720   9   24927  75   33768  10   206.9   0
zam420f3  9568  24  10094   2   3267  61   10113  25   10317   3   343.0   1
zam271f3 10143  30   9924   4   5061   2   10830  31   10994   2   191.3   0
```

ANHANG 3: ERGÄNZENDE ZUKÜNFTIGE AUFGABEN ZUR DATENSICHERUNG

BEANTRAGUNG DER DATENSICHERUNG

Die Teilnahme an der Datensicherung muss beantragt werden. Für die Anmeldung werden folgende Informationen benötigt:

1. der Rechnername im JuNet
   (primärer Name oder Alias)
2. der Rechnertyp
   (PC :Windows, PC: Linux, PC: sonst, Unix Workstation: AIX, Unix Workstation: SUN, Unix Workstation: SGI, Unix Workstation: HP)
3. die Rechnerfunktion
   (ZAM-Server, WSG-Server, WSG-Klient, Instituts-Server, Stand-Alone-System)
   Der Sonderfall des TSM Servers für Unix Systeme des IFF, der noch aus FDDI-Zeiten stammt, soll aufgelöst werden. Noch genutzte Registrierungen sind in Absprache mit den Benutzern auf die anderen TSM Server zu verteilen.
4. ein initiales TSM Backup Password
   (max. 63 Zeichen bestehend aus A-Z, 0-1, +._-&)
5. der Name des Administrators/Benutzers
   (Email-Adresse des Administrators/Benutzers)
6. automatisches Schedule
   (ja/nein, falls ja Auswahl aus verfügbaren Schedules – Tag, Nacht, Wochenende, Server, Client)
7. Kennzeichnung der Speicherung personenbezogener Daten
   (ja/nein, falls ja muss der Benutzer einen zusätzlichen Hinweis über die Notwendigkeit der Verschlüsselung der Daten im Sinne der *IT-Sicherheitsregeln für den Grundschutz* - Regel D4 - unterschreiben)

Die Informationen aus 2 und 3 werden benötigt, um den zuständigen logischen TSM Backup Server zu bestimmen. Der TSM Backup Servername und die Portnummer auf dem TSM Backup Server müssen dem Benutzer mitgeteilt werden, da sie zur Konfiguration des TSM Klienten benötigt werden.

Zur tatsächlichen Registrierung beim TSM Backup Server werden die Informationen 1,4,5,6 aus der Anmeldung benötigt.

Mit der Registrierung zur Datensicherung bekommt der Administrator/Benutzer einen Hinweis, dass unter folgenden Umständen die Registrierung aufgelöst und die Sicherungsdaten gelöscht werden.

1. Wenn der Dienst von dem registrierten Rechner ein Jahr nicht in Anspruch (Lesen oder Schreiben) genommen wurde.
2. Wenn ein Administrator/Benutzer seine offizielle Email-Adresse zurückgibt und das Backup leer ist.

➔ Die Antragstellung für die Datensicherung muss in die zentrale Registry des ZAM (Dispatch) integriert werden, d.h. die entsprechenden Online Formulare müssen entwickelt und Routinen implementiert werden.

BACKUP-PASSWORD ZURÜCKSETZEN

Das TSM Backup-Password wird verschlüsselt auf dem TSM Backup Server und auf dem registrierten System abgelegt. Es kann nicht vom TSM Administrator ausgelesen werden. Bei Verlust des TSM Backup-Passwords muss der Administrator/Benutzer das TSM Backup-Password zurücksetzen lassen.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

TSM bietet die Funktion des Password-Expiring. Dabei tauschen TSM Server und TSM Klient nach definierter Anzahl Tage neue Passwords aus. Bei Unix Klienten geht das in der Regel problemlos mit Ausnahme von Linux Systemen. Windows Systeme, die das automatische Backup nutzen, haben das verschlüsselte Password in der Windows Registry stehen, wo es nicht automatisch geändert wird. Das Ändern dieses Password ist nicht intuitiv und benutzerfreundlich. Daher ist das Password-Expiring für Windows- und Linux-Systeme nicht aktiviert.

LÖSCHEN VON BERECHTIGUNGEN ZUR DATENSICHERUNG

Berechtigungen zur Datensicherung und zugehörige Sicherungskopien werden auf Antrag des eingetragenen Benutzers/Administrators gelöscht.
Gelöschte Sicherungskopien können nicht durch den TSM Backup Server wieder hergestellt werden. Wenn die Daten noch auf dem Klienten existieren, können sie durch einen erneuten Sicherungslauf neu erstellt werden.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

Wenn ein Administrator/Benutzer das Forschungszentrum Jülich verlässt, ohne sich um die noch reservierten und belegten IT-Ressourcen zu kümmern, gelten folgende Regeln, die spätestens dann aktiv werden, wenn die Email-Adresse des Administrators/Benutzers gelöscht werden soll.

- Wenn das Backup zum Zeitpunkt des Löschens der Email-Adresse leer ist, wird es direkt gelöscht.
- Wenn das Backup Daten enthält, aber der Dienst von dem registrierten Rechner seit einem Jahr nicht mehr genutzt wurde (Lesen und Schreiben), werden die Berechtigung zur Datensicherung und die Sicherungskopien gelöscht.
- Wenn das Backup aktuelle Daten enthält, wird die Berechtigung zur Datensicherung zwischenzeitlich auf den DV-Ansprechpartner übertragen und dieser informiert mit der Bitte einen neuen Administrator/Benutzer des Rechners zu benennen.

SPERREN/ENTSPERREN EINER BERECHTIGUNG ZUR DATENSICHERUNG

Berechtigungen zur Datensicherung werden 10 Tage nach dem Sperren einer Email-Adresse z.B. nach Überschreiten der Frist der Gültigkeitsverlängerung ebenfalls gesperrt.

Zum Sperren/Entsperren stehen Lock/Unlock Mechanismen zur Verfügung. Das Backup-Password sollte nicht umgesetzt werden, da bei Windows Systemen, die das automatische Backup nutzen, das verschlüsselte Backup-Password in der Windows Registry steht. Das Ändern dieses Backup-Passwords nach dem Entsperren ist nicht intuitiv und nicht benutzerfreundlich.

ÜBERTRAGEN VON BERECHTIGUNGEN ZUR DATENSICHERUNG

Berechtigungen zur Datensicherung können vom Administrator/Benutzer des Systems auf Antrag auf einen anderen Administrator/Benutzer übertragen werden. Der neue Administrator/Benutzer hat dann Zugriff auf alle Sicherungskopien einschließlich der inaktiven Kopien gelöschter Dateien und ebenfalls auf persönliche Daten z.B. Sicherungskopien von E-Mails.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Administrators/Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

AKTIONEN

1. Entwurf von Online Formularen und Entwicklung von Routinen für die zentrale Registry (Dispatch) - siehe Verweise mit ➔ in Text (in 2006)
2. TSM Server des IFF in Absprache mit den Benutzern migrieren und auflösen (in 2006)

3. Neues Angebot zum Download und zur Installation der TSM Client Software anstelle des NFS-Zugriffs zur Verfügung stellen (in 2005 erfolgt)
4. Überarbeiten der TKI-0368 *Datensicherung mit TSM/ADSM* (in 2005 erfolgt)
5. Überarbeiten der Web-Seiten (in 2005 erfolgt)

ANHANG 4: ERGÄNZENDE ZUKÜNFTIGE AUFGABEN ZUR DATENARCHIVIERUNG

BEANTRAGUNG VON ARCHIVEN

Archive müssen beantragt werden. Für die Anmeldung werden folgende Informationen benötigt:

1. eine Kennzeichnung des Archivtyps
   (persönliches Archiv, Langzeitarchiv)
2. der Archivname
   (Email-Adresse oder Langzeitarchivname)
3. ein initiales TSM Archiv-Password
   (max. 63 Zeichen bestehend aus A-Z, 0-1, +._-&)
4. der Eigentümer bzw. der Betreuer
   (bei persönlichen Archiven die offizielle Email-Adresse des Benutzers,
   bei Langzeitarchiven die Email-Adresse des Betreuers oder eine Funktions-Email)
5. im Fall von Langzeitarchiven eine Projektbeschreibung
   (vom Institutsleiter unterschrieben)
6. Kennzeichnung der Speicherung personenbezogener Daten
   (ja/nein, falls ja muss der Benutzer einen zusätzlichen Hinweis über die Notwendigkeit der Verschlüsselung der Daten im Sinne der *IT-Sicherheitsregeln für den Grundschutz* - Regel D4 - unterschreiben)

Zur tatsächlichen Registrierung beim TSM Archiv Server werden die Informationen aus 2,3,4 aus der Anmeldung benötigt.
Mit der Registrierung eines persönlichen Archivs bekommt der Benutzer einen Hinweis, dass ein persönliches Archiv aufgelöst wird, wenn die Gültigkeit der offiziellen Email-Adresse abgelaufen ist und nach einer Wartefrist gelöscht werden soll. In diesem Fall werden die Archivdaten unwiderruflich gelöscht. Mit der Registrierung eines Langzeitarchivs erhält der Betreuer einen Hinweis, dass er vor dem Löschen seiner Email-Adresse für eine ordnungsgemäße Übertragung des Archivs auf eine andere Person zu sorgen hat.

➔ Die Antragstellung für ein Archiv muss in die zentrale Registry des ZAM (Dispatch) integriert werden, d.h. die entsprechenden Online Formulare müssen entwickelt und Routinen implementiert werden.

ARCHIV-PASSWORD ZURÜCKSETZEN

Das TSM Archiv-Password wird verschlüsselt auf dem TSM Archiv Server abgelegt. Es kann nicht vom TSM Administrator ausgelesen werden. Bei Verlust des TSM Archiv-Passwords muss der Benutzer oder Betreuer das TSM Archiv-Password zurücksetzen lassen.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

LÖSCHEN VON ARCHIVEN

Daten aus einem gelöschten Archiv können nicht wieder hergestellt werden. Persönliche Archive werden auf Antrag des Eigentümers gelöscht. Für Langzeitarchive ist zum Löschen eine Bestätigung durch den Institutsleiter erforderlich.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

Wenn ein Benutzer/Betreuer das Forschungszentrum Jülich verlässt, ohne sich um die noch reservierten und belegten IT-Ressourcen zu kümmern, gelten folgende Regeln, die spätestens dann aktiv werden, wenn die Email-Adresse des Benutzers/Betreuers gelöscht werden soll.

- Persönliche Archive
  Wenn das persönliche Archiv zum Zeitpunkt des Löschens der Email-Adresse leer ist, wird es direkt gelöscht.
  Wenn das persönliche Archiv Daten enthält, wird es in ein *persönliches übertragenes Archiv* für den DV-Ansprechpartner umgewandelt und ein entsprechender Brief an den DV- Ansprechpartner geschickt mit dem Namen des alten und neuen Archivs, sowie der Bitte um Aufräumen und der Angabe einer Frist, bis wann das Archiv noch zugreifbar ist. Wenn der DV-Ansprechpartner sich ebenfalls nicht kümmert, wird das Archiv nach einem Jahr gelöscht.
- Langzeitarchive
  Lanzeitarchive werden nicht gelöscht. Wenn die Email-Adresse des Betreuers oder die Funktions-Email eines Langzeitarchivs aufgelöst werden soll, wird ein Brief an den Institutsleiter verschickt mit der Bitte um Nennung eines neuen Verantwortlichen oder der Erlaubnis zum Löschen des Archivs. Die Email-Adresse des Betreuers wird zwischenzeitlich auf den DV-Ansprechpartner umgetragen.

SPERREN/ENTSPERREN EINES ARCHIVES

Archive werden 10 Tage nach dem Sperren einer Email-Adresse z.B. nach Überschreiten der Frist für eine Gültigkeitsverlängerung ebenfalls gesperrt.

Zum Sperren/Entsperren stehen Lock/Unlock Mechanismen zur Verfügung. Das Password sollte nicht dazu umgesetzt werden.

ÜBERTRAGEN VON ARCHIVEN

Persönliche Archive können nur dann auf einen anderen Mitarbeiter übertragen werden, wenn dieser noch kein eigenes persönliches Archiv besitzt. Andernfalls müssen die Daten zurückgeholt und vom neuen Besitzer erneut unter seinem Namen archiviert werden. Langzeitarchive können durch den Institutsleiter auf andere Verantwortliche übertragen werden.

➔ Ein entsprechendes Online Formular mit Prüfung der Authentizität des Benutzers muss erstellt und in die zentrale Registry (Dispatch) integriert werden.

AKTIONEN

1. Entwurf von Online Formularen und Entwicklung von Routinen für die zentrale Registry (Dispatch) - siehe Verweise mit ➔ in Text (in 2006)
2. Neues Angebot zum Download und zur Installation der TSM Client Software anstelle des Zugriffs über NFS zur Verfügung stellen (in 2005 erfolgt)
3. Überarbeiten der TKI-0261 *Archivierung mit TSM/ADSM* (in 2005 erfolgt)
4. Überarbeiten der Web-Seiten bzgl. Archivierung (in 2005 erfolgt)