

# PROJEKTWISSEN: DATEIVERWALTUNG AM LEIBNIZ-RECHENZENTRUM

Ein Cluster-Filesystem mit Dateivirtualisierung verwaltet am Leibniz-Rechenzentrum die Datenbestände der Supercomputer-Nutzer. Die Lösung bietet hohe Skalierbarkeit, ohne dabei die einheitliche Sicht auf alle Files zu opfern.

Datei-Virtualisierung lässt Wissenschaftler schnell auf ihre Anwenderdaten zugreifen – Management und Konsistenz sind weitere Pluspunkte

## Supercomputer-Nutzer speichern im Cluster

Dateien extrem unterschiedlicher Größe und Anforderungen fallen am Leibniz-Rechenzentrum (LRZ) in Garching an. Bei der Verwaltung hilft das Cluster-Filesystem Data Ontap GX von Network Appliance (Netapp).

Seit Juli rechnet im LRZ-Neubau in Garching ein neuer Supercomputer: Ein SGI-System mit 4096 Intel-Itanium-2-CPU hat den betagten SR8000-Rechner von Hitachi abgelöst. Mit 24,36 Teraflops (Billionen Fließkommaoperationen pro Sekunde) Linpack-Rechenleistung belegt er Platz 18 im Top-500-Ranking der Numbercruncher.

Diese Rechengewalt stellt hohe Anforderungen an die Speicherinfrastruktur. „Wie bei vielen Hochleistungsanwendungen gibt es eigentlich zwei Typen von Dateien“, erläutert Systembetreuer Christoph Biardzki: „Zum einen die Programmcodes und die Eingabedaten der Benutzer, die verhältnismäßig klein sind. Zum anderen die Zwischenergebnisse der Rechnungen, die viele Terabytes groß sein können.“

Diese Spanne stellt auch komplett unterschiedliche Anforderungen an die Dateisysteme. „Schon bei der Beschaffung war klar, dass wir zwei unterschiedliche Arten von Storage brauchen“, betont Biardzki – und zwar einen für die temporären Dateien, die wirklich superschnell verarbeitet werden müssen, dafür aber nicht so hohe Anforderungen an Zuverlässigkeit und Datenkonsistenz stellen wie der andere Teil, der die

Eingaben, Programmcodes und Endergebnisse der Benutzer speichert.

Wird die erste Gruppe von SGLs parallelem Dateisystem CXFS verwaltet, so wählte das LRZ für das Management der Benutzerverzeichnisse Netapps Cluster-Filesystem Data Ontap GX. Hervorstechendes Merkmal: Mehrere individuelle Dateisysteme lassen sich in einem gemeinsamen Namensraum zusammenfassen und einheitlich verwalten.

### File-Virtualisierung hilft bei Verwaltung

Stößt ein gängiger NFS-Server an seine Leistungsgrenze, dann wird ein neues Dateisystem aufgemacht. Man muss dann zwei verwalten und stets schauen, wer wo seine Daten liegen hat. Bei der Netapp-Variante hingegen braucht sich der Benutzer nicht darum zu kümmern, auf welchem der Filer – der Data-Ontap-GX-Cluster des LRZ umfasst deren sechs – seine Dateien liegen.

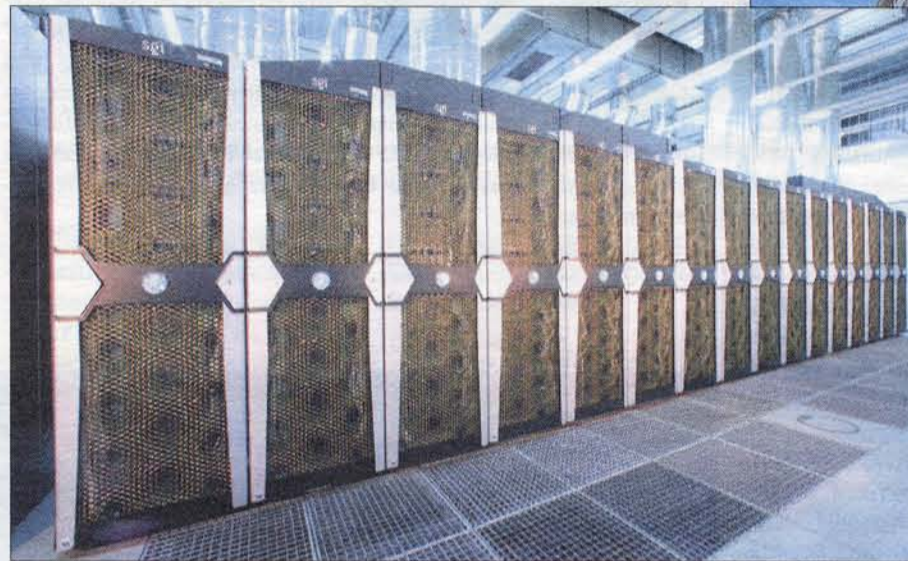
„Der Administrator kann Daten transparent im Hintergrund migrieren, ohne den Benutzer damit behelligen zu müssen“, ergänzt Biardzki. Eine wichtiger Punkt sei schließlich auch die Snapshot-Funktion, die andere Hochleistungs-Filesysteme nicht angeboten hätten – oder wenn, dann nur mit großen Performanceeinbußen.

Dass man den Sprung auf ein brandneues Produkt wie Data Ontap GX gewagt hat, liegt unter anderem an dessen kon-

zeptioneller Verwandtschaft zum AFS (Andrew Filesystem), das schon seit Jahren am LRZ eingesetzt wird. „Es ist keine komplette Neuentwicklung“, unterstreicht Biardzki – „ein bedeutender Faktor, bedenkt

bearbeiten. „Ansonsten war eigentlich nicht mehr nötig“, zieht Biardzki Bilanz.

Die Performance entsprach auf Anhieb den Vorgaben. Zur Bewertung wurde zum einen der sequenzielle Durchsatz von



Im Sommer ist der Bundeshöchstleistungsrechner in Bayern II (links) im Rechenzentrumsneubau in Garching (oben) in Betrieb gegangen. Die Nutzer profitieren nicht nur von der gewaltigen Rechenpower, sondern auch von der im Vergleich zum Hitachi-Vorgängersystem drastisch verbesserten Metadaten-Performance durch das Cluster-Filesystem Data Ontap GX. Fotos: Payer/Gsicom, Kieß

man, dass die komplette Neuentwicklung eines Dateisystems bis zum stabilen Betrieb in der Regel etwa fünf Jahre dauert“. So ging denn auch die Installation gemessen an der Neuheit des Produkts sehr leicht vonstatten. Zwar gab es zu Implementierungsbeginn noch keinen offiziellen Support – erst im Juni 2006 startete die allgemeine Auslieferung von Data Ontap GX –, doch erhielt das LRZ Unterstützung aus der Netapp-Vertretung im nahen Grasbrunn. Außerdem wurde für einen Tag ein Mitarbeiter der Special Deployment Teams aus Israel eingeflogen, um einige spezielle Anforderungen zu

der Platte gemessen – für den Fall, in dem ein Nutzer eine relativ große Ergebnisdatei in seinem Home-Verzeichnis ablegt. Zum anderen kam ein spezieller Metadaten-Benchmark zum Einsatz. Dieser testete dediziert, wie schnell das System neue Dateien erzeugt, verändert und löscht.

### Metadaten-Leistung ist kritische Größe

Bei der Arbeit mit sehr vielen kleinen Dateien spielt Data Ontap GX seine architekturbedingten Stärken aus: Es gibt keinen dedizierten Metadatenserver – Daten und zugehörige

Metadaten liegen stets auf demselben Server.

„Für kleine Dateien ist das natürlich sehr günstig“, so Biardzki. Bei getrennter Metadatenhaltung, wie bei CXFS oder beim quelloffenen Cluster-Filesystem Lustre, sei der Aufwand höher, um von den Metadaten zu den Daten zu kommen. „Das fällt bei einer 500-Gigabyte-Datei nicht so ins Gewicht, bei einer kleinen aber sehr wohl.“ Deshalb sei das Netapp-System bei der Arbeit mit vielen kleinen Dateien, etwa beim Kompilieren von Programmen, um den Faktor Zehn schneller als das CXFS auf der SGI. „Das ist ein Unterschied,

den die Nutzer schon merken.“

Bei riesigen Files hingegen stößt die Netapp-Lösung an ihre Grenzen. Hier spielen die schnellen parallelen Dateisysteme ihre Stärken aus. Dafür sind sie aufgrund der schieren Zahl beteiligter Komponenten aber auch anfälliger für die Ausbreitung von Fehlern.

Bei Data Ontap GX wiederum bleiben Fehler isoliert; es sind vielleicht einige Volumens betroffen, aber nie das ganze Dateisystem. Eine notwendige Konsistenz im Hinblick auf den Nutzerkreis – bundesweit greifen über 100 Projekte mit jeweils mehreren Usern auf den Rechner zu. fm

LRZ-Speichergruppe muss 1,7 Petabyte verwalten – Datenvolumen verdoppelt sich jährlich

## „Die Anwender können die Archivierung selbst steuern“

Als Storage-Verantwortlicher am Leibniz-Rechenzentrum (LRZ) kümmert sich Bernd Reiner nicht nur um die Sicherung der hauseigenen Daten, sondern auch um die aller Münchner Hochschulen. Als Archiv dienen performante Tape-Systeme.

Sie haben am LRZ das erste Data-Ontap-GX-Produktivsystem eingerichtet. Wie lief denn die Installation?

Eigentlich erstaunlich leicht. Wir hätten mehr Probleme erwartet.

Und wie sind Sie mit der Performance zufrieden?

Wir hatten klare Benchmark-Anforderungen, und das System hat sie im ersten Anlauf geschafft – wobei das schon im Vorfeld überprüft worden war.

Wie sieht die Speicherorganisation insgesamt aus?



Betreut die Datei- und Speichersysteme am Leibniz-Rechenzentrum: Bernd Reiner. Foto: Kieß

Die Home Directories der Anwender werden zunächst noch einmal gespiegelt auf andere Netapp-Filer. Diese Daten werden dann zusätzlich auf Magnetbänder gesichert.

### Und was geschieht mit den vom Supercomputer gerechneten Daten?

Die werden direkt von den Knoten auf unsere Systeme abgelegt. Zudem wird eine Zweitkopie am Rechenzentrum Garching der Max-Planck-Gesellschaft erstellt, um auch eine örtliche Trennung zu haben.

### Wer regelt die Archivierung?

Das Einlagern und Auslagern ins Archiv können die Nutzer ganz einfach selbst steuern. Das ist gekoppelt mit dem Speicher-Verwaltungssystem IBM Tivoli Storage Manager (TSM).

### Wie entwickelt sich das Datenvolumen?

Momentan haben wir 1,7 Petabyte Daten hier liegen, und dieses Volumen verdoppelt sich eigentlich jedes Jahr. Deshalb haben wir auch mit den Storage-Libraries performante Tape-Systeme beschafft, die eine

Schreib-Lese-Performance von 120 Megabyte pro Sekunde für unkomprimierte Daten erreichen. Das müssen die Platten erst einmal liefern können.

### Nutzen Sie Caching?

Zum Teil. Der TSM verfügt über etwa 160 Terabyte Plattenplatz. Und bei Backups wird zunächst da reingeschrieben, zusammengefasst und anschließend in einem Rutsch auf die Tapes übertragen. Sonst würde man im Falle eines Netzwerkproblems diesen Stop-and-Go-Betrieb bekommen, den die Bänder nicht mögen.

### Und für die Archivierung?

Hier geht die Datei direkt aufs Band, wenn sie eine bestimmte Größe hat – das kann man einstellen. Kleinere Dateien bis 50 Megabyte gehen ebenfalls in einen Platten-Cache, werden dort gesammelt und dann komplett rausgeschrieben. fm

## Datenhaltung läuft zweigleisig

Beim Höchstleistungsrechner in Bayern (HLRB II) am Leibniz-Rechenzentrum handelt es sich in der ersten Ausbaustufe um ein Linux-basiertes SGI-Altix-System mit 4096 Intel-Itanium-2-CPU und mehr als 17 Terabyte Hauptspeicher. Für die Ablage und Weiterverarbeitung von Rechenergebnissen stehen 340 Terabyte an Plattenplatz zur Verfügung. Die Daten können dort mit einer aggregierten Bandbreite von 20 Gigabyte pro Sekunde gelesen oder abgelegt werden. Als Managementsoftware dienen SGI's XFS-Dateisystem und XVM Volume Manager sowie deren Cluster-Erweiterungen CXFS und CXVM.

Für Backup und Archivierung kommuniziert der Superrechner über einen TSM-Server (IBM Tivoli Storage Manager) mit dem Tape-System – einer modularen SL8500-Bandbibliothek von StorageTek. Diese nutzt T10000-Laufwerke mit 500 Gigabyte nativer Kapazität pro Cartridge und einem Datendurchsatz von 120 Megabyte pro Sekunde (unkomprimiert). Die Benutzerverzeichnisse mit Programmquellen, Konfigurationsdateien et cetera liegen auf einem Data-Ontap-GX-Cluster mit sechs Knoten – jeweils Netapp-Filer vom Typ FAS3050 (Foto). Er bietet 40 Terabyte Nettokapazität auf Fibre-Channel-Platten, die auch von außerhalb zugänglich sind. Zusätzlich sind etwa zehn Terabyte auf SATA-Platten als Nearline-Speicher integriert, um auch innerhalb des Clusters replizieren zu können. Zusätzlich werden alle Daten noch auf T10000-Magnetbändern gesichert. fm

