Erfolgreicher Einsatz des parallelen Dateisystems Lustre

Roland Laifer

Das neue parallele Dateisystem *Lustre* wird vom Rechenzentrum seit Anfang des Jahres erfolgreich auf dem Landeshöchstleistungsrechner HP XC6000 eingesetzt. Dabei handelt es sich unseres Wissens sowohl um den ersten Produktionseinsatz im deutschsprachigen Raum als auch um den weltweit ersten Einsatz für Home Directories. Somit ist nicht verwunderlich, daß diverse Vorträge des Rechenzentrums über die Erfahrungen mit Lustre auf großes Interesse stießen - zuletzt auf der International Supercomputer Conference ISC2005 in Heidelberg.

Ein paralleles Dateisystem zeichnet sich dadurch aus, daß Daten zwischen mehreren Rechnern, auf denen sie genutzt werden, und den Festplatten, auf denen sie gespeichert werden, vollständig parallel übertragen werden. Durch die Parallelität ergibt sich in der Regel ein Geschwindigkeitsvorteil beim Datenzugriff. Parallele Dateisysteme werden häufig auf Clustern eingesetzt, weil durch die ständige Erhöhung der CPU-Leistung, der Knotenanzahl und des verfügbaren Hauptspeichers die I/O-Anforderungen der auf den Clustern laufenden Anwendungen immer weiter steigen. Steht auf einem Cluster beispielsweise nur das serielle Dateisystem NFS zur Verfügung, kann der Datenzugriff schnell zum Flaschenhals werden.

Das Dateisystem Lustre[1] wurde von Cluster Filesystems Inc. entwickelt, siehe www.clusterfs.com und www.lustre.org. Wesentliche Designziele waren eine hohe Leistung, d.h. ein möglicher Durchsatz von mehreren Gigabyte pro Sekunde, und eine hohe Skalierbarkeit, d.h. die mögliche Nutzung mit mehreren tausend Clients. Ein wichtiges Konzept zum Erreichen dieser Ziele ist die Trennung zwischen den Metadaten - das sind z.B. Verzeichnisse oder Dateiattribute wie Name und Zugriffsrechte - und den eigentlichen Daten, d.h. dem Inhalt der Dateien. Ebenso wichtig ist ein ausgeklügeltes System zur Verwaltung der Locks, womit u. a. das gleichzeitige Schreiben von verschiedenen Rechnern auf den gleichen Datenblock verhindert wird und womit die Konsistenz des Dateisystems gewährleistet bleibt. Abb. 1 zeigt die wichtigsten Lustre-Komponenten und -Protokolle. Die Datenübertragung zwischen Client und Server wurde optimiert und erfolgt über die in Clustern typischerweise existierenden schnellen Netzwerke, wobei derzeit Gigabit Ethernet, Myrinet, Quadrics und Infiniband unterstützt werden. Aktuelle Lustre-Versionen werden derzeit nach spätestens einem Jahr als Open Source kostenlos zur Verfügung gestellt.

[1] Der Name *Lustre* wurde aus einem Amalgam der Worte Linux und Cluster gebildet.

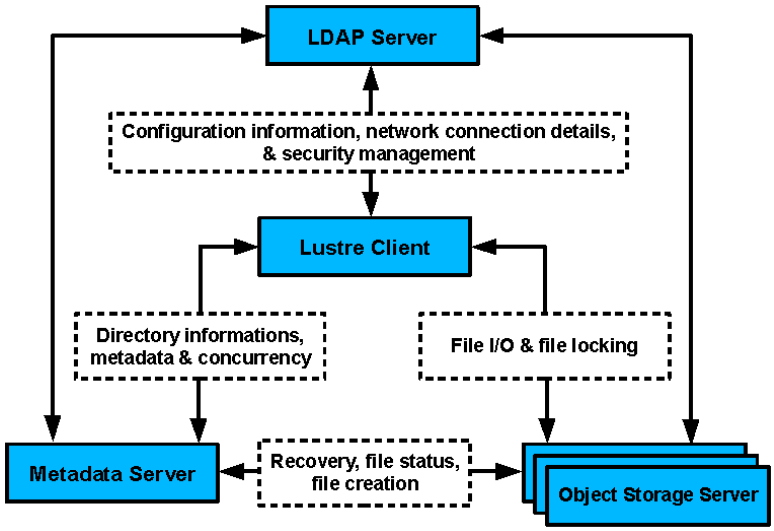

Abb. 1: Komponenten und Protokolle des Dateisystems Lustre

Lustre wird bereits auf vielen der weltweit größten Rechner eingesetzt. Außerdem kann in den nächsten Jahren auch mit einem verstärkten Einsatz auf kleineren Linux-Clustern gerechnet werden. Die stärksten Hindernisse sind dabei wohl, daß derzeit nur bestimmte Linux-Kernelversionen und -Derivate unterstützt werden, sowie die Komplexität der Konfiguration und Administration.

HP arbeitet mit Cluster Filesystems Inc. zusammen und hat ein eigenes Lustre-Produkt erstellt, was unter dem Namen *HP StorageWorks Scalable File Share (HP SFS)* vertrieben wird. HP SFS enthält zusätzliche Software zur Ausfallsicherheit und vereinfachten Administration. Außerdem wird HP SFS nur zusammen mit bestimmter Hardware von HP geliefert, beispielsweise kommen als Server die Xeon-basierten HP ProLiant G3 oder G4 und als Speichersysteme die sog. EVA3000[2] oder SFS20 zum Einsatz. Abb. 2 zeigt das HP SFS System des Landeshöchstleistungsrechners HP XC6000, wobei im oberen Teil des Bildes die Itanium2-basierten Knoten des Clusters und somit die Clients (C) des Dateisystems abgebildet sind.

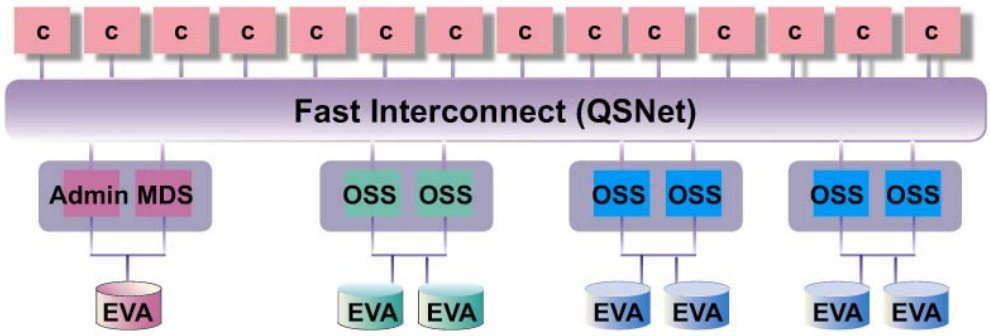

Abb. 2: HP SFS am Landeshöchstleistungsrechner HP XC6000

Die Server sind in ausfallsicheren Paaren konfiguriert, d.h. wenn beispielsweise der Metadata Server (MDS) ausfällt, übernimmt der Admin Server dessen Rolle und Dateisystemzugriffe können nach einer kurzen Pause ohne Abbruch weiterlaufen. Die sog. Object Storage Server

[2] Das Rechenzentrum nutzt EVA5000 Speichersysteme.

(OSS) halten die eigentlichen Daten. Diese sind in der Regel mittels Striping verteilt, d.h. große Dateien sind in Blöcke zelegt und diese werden gleichmäßig verteilt über die OSS gespeichert. Dadurch läßt sich schon beim Zugriff auf *eine* Datei eine sehr hohe Leistung erzielen. Der Admin Server erlaubt die Verwaltung des Systems von einem zentralen Punkt, beispielsweise werden von ihm aus die anderen Server gebootet oder Dateisysteme konfiguriert.

Auf dem Landeshöchstleistungsrechner HP XC6000 gibt es 8 HP SFS Server und 2 Dateisysteme, siehe Abb. 2. Die Dateisysteme heißen *data* und *work*. Das Dateisystem *data* wird für Home Directories und Software genutzt, hat eine Speicherkapazität von 3.8 TB und nutzt 2 dedizierte OSS. Das Dateisystem *work* und wird für Scratch-Daten genutzt, z.B. zum Speichern der Zwischenergebnisse von Batchjobs. Es hat eine Speicherkapazität von 7.6 TB und nutzt 4 OSS.

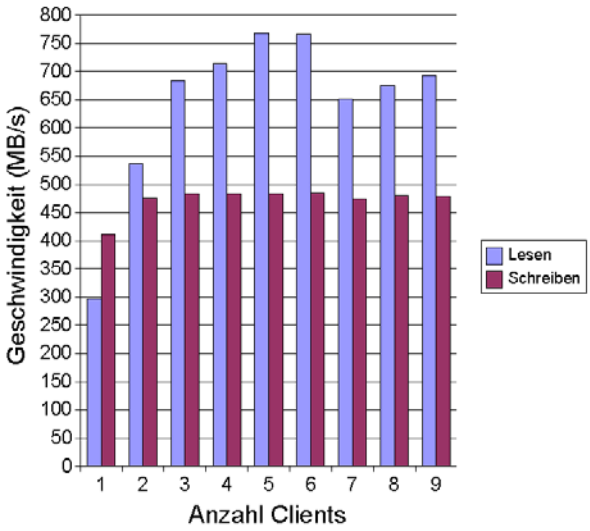

Abb. 3: Leistung des Dateisystems work beim sequentiellen Zugriff

Aus Abb. 3 erkennt man die Leistung beim sequentiellen Lesen und Schreiben im Dateisystem *work*. Von einen Itanium2-Client läßt sich über den Quadrics-Switch bereits eine Schreibgeschwindigkeit von 400 MB/s und eine Lesegeschwindigkeit von 300 MB/s erreichen. Der Flaschenhals auf der Serverseite ist beim Schreiben das Plattenspeichersystem EVA5000 und beim Lesen der 2 Gbit/s FC-Adapter, über den die Speichersysteme angeschlossen sind. Mit dem Dateisystem *data* läßt sich in etwa die halbe maximale sequentielle Leistung erreichen, denn der wesentliche Skalierungsfaktor ist die Anzahl der OSS. Neben der Datenübertragungsleistung besticht Lustre oder HP SFS auch durch eine sehr gute Metadatenleistung. So können auf der HP XC6000 bis zu 5000 Dateien pro Sekunde erzeugt bzw. gelöscht werden.

Um den Einsatz auf dem Landeshöchstleistungsrechner vorzubereiten, hat das Rechenzentrum bereits im Frühjahr 2004 mit einem Betatest des HP SFS begonnen. In Zusammenarbeit mit HP wurden viele Softwarefehler beseitigt, wobei das Rechenzentrum laut Aussage von HP der aktivste Betatestkunde war. Durch diese Vorarbeiten konnte zum Ende des letzten Jahres gleich nach der Verfügbarkeit die erste offizielle HP SFS Version auf der HP XC6000 eingesetzt werden. Zu Anfang des Produktionsbetriebs gab es von Zeit zu Zeit weitere Probleme, die aber in Zusammenarbeit mit HP ebenfalls gelöst werden konnten. Inzwischen ist der Betrieb sehr stabil; beispielsweise gab es von Mitte Mai bis Mitte August nur ein ernsthaftes Problem im Bereich des Anschlusses der Server an den Quadrics-Switch. Bei

diversen Hardware-Ausfällen zeigte sich, dass die Mechanismen zur Ausfallsicherheit funktionieren; die Anwendungen laufen dabei nach einem kurzen Hänger einfach weiter.

Weitere Informationen erhalten Sie beim Autor unter Tel. -4861 bzw. E-Mail laifer@rz.uni-karlsruhe.de.