Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Projektbericht

Mykhaylo Kabalkin

Lustre File System in Collaborative Workspace

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Mykhaylo Kabalkin

Lustre File System in Collaborative Workspace

Projektbericht eingereicht im Rahmen des Projektes
im Studiengang Informatik Master of Science
am Studiendepartment Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuer : Prof. Dr. Kai von Luck

Abgegeben am 28. Februar 2007

Mykhaylo Kabalkin

Thema des Projektberichtes

Lustre File System in Collaborative Workspace

Stichworte

Collaborative Workspace, Lustre File System, IBM BladeCenter

Kurzzusammenfassung

Mit diesem Projektbericht beschreibt der Autor seine im Rahmen vom „Collaborative Workspace“ Projekt gewonnenen Erkenntnisse. Problematiken, die während der Arbeit aufgetreten sind, werden auch erläutert.

Mykhaylo Kabalkin

Title of the paper

Lustre file system in collaborative workspace

Keywords

Collaborative Workspace, Lustre File System, IBM BladeCenter

Abstract

This paper describes recognitions, which the autor won working on the "Collaborative Workspace" project. Difficulties, which during the work on the project were occured, will be explained too.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Einführung

1.1. Motivation

Das langfristige Ziel des Autors ist es, ein Persistenz-Service System zu entwerfen. Die ersten Ideen des Systems sind im Rahmen von der Seminar-Ringvorlesung [Kabalkin (2007a)] beschrieben. Als einen Teil des Systems sieht der Autor ein verteiltes Dateisystem vor. Im Rahmen der Vorlesung „Anwendungen II“ [Kabalkin (2007b)] wurden von dem Autor dieses Berichtes einige verteilte Dateisysteme untersucht. Ein dabei ausgewähltes verteiltes Dateisystem wird im Rahmen von „Collaborative Workspace“ Projekt detaillierter untersucht.

Dieser Bericht dient dazu, den Inhalt, die Erkenntnisse und die Problematik des gleichnamigen Projektes im Rahmen der Lehrveranstaltung im Informatik-Masterstudiengang an der Hochschule für Angewandte Wissenschaften Hamburg näher zu beschreiben.

1.2. Ziele

Im Laufe des Projektes wurden die Ziele, die sich die Projektbeteiligten stellten, geändert. Die Tabelle 1.1 veranschaulicht die Themen und Ziele der Projektteilnehmer am Startpunkt des Projektes und aktuell.

	STARTPUNKT	AKTUELLER STAND
Christian	Grabstick	Grabstick-Interaktion
Mykhaylo	Verteilter Persistenzservice	Verteiltes Dateisystem
Oliver	Lernmethode für Kollaboration	Barrierefreier, verteilter Desktop
Pascal	Qualitativ hochwertiges Image-Retrieval	Image-Retrieval Webservice
Phillip	Physikbasierte Multitouchinteraktion	Physikbasierter Collaborative Workspace

Tabelle 1.1.: Themen und Ziele der Projektgruppe

Das Ziel des Autors war es, ein ausgewähltes verteiltes Dateisystem zu untersuchen und zu evaluieren.

In [Ramamurthy (2004)] definiert Dr. Bina Ramamurthy ein verteiltes Dateisystem wie folgt:

> „Distributed file systems support the sharing of information in the form of files throughout the intranet. A distributed file system enables programs to store and access remote files exactly as they do on local ones, allowing users to access files from any computer on the intranet.“

Dieses Zitat betrachtet der Autor dieses Berichtes als die Vision seiner Arbeit.

2. Projektübersicht

Das „Collaborative Workspace“ Projekt wurde von fünf Studenten (Christian Fischer, Mykhaylo Kabalkin, Oliver Köckritz, Pascal Pein und Philipp Roßberger) gemeinsam durchgeführt. Alle Teilnehmer verfolgten unterschiedliche Ziele, die aber viele Gemeinsamkeiten hatten. Es wurden einige gemeinsame Entscheidungen getroffen:

- Einigung auf übergreifende Systemarchitektur (Abbildung 2.1)
  - Service-basierte Dienstkopplung ohne Middleware
- Abstimmung der verwendeten Technologien
  - Web-Services, IP-Multicasts (Pascal, Christian)
  - C++ / OpenSG (Christian, Oliver, Philipp)
- Planung der Arbeitsschnittstellen für die Master-Thesis
- Laboraufbau

Es wurde eine übergreifende Systemarchitektur erstellt, die in der Abbildung 2.1 dargestellt ist.

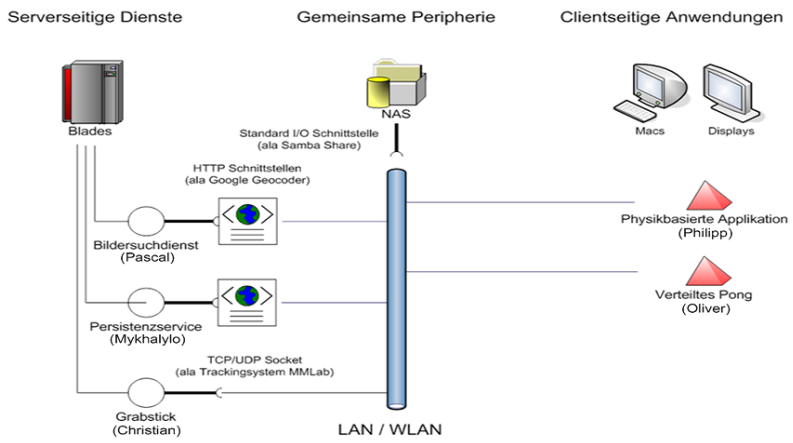

Abbildung 2.1.: Übergreifende Systemarchitektur

3. Erkenntnisse

3.1. Lustre File System

Das Lustre File System (im Folgenden Lustre FS genannt) ist ein skalierbares, sicheres, robustes und ausfallsicheres Cluster Datei System, und wurde von Cluster File System Inc. entwickelt.

Ein Lustre Cluster besteht aus drei Hauptkomponenten:

- Meta Data Server (MDS)
- Object Storage Server (OSS)
- Lustre Clients

Jede dieser Komponenten ist sehr modular in Layout. Die Request-processing Schicht und die Message-passing Schicht werden zwischen allen Komponenten im gesamten System geteilt. Die Abbildung 3.1 stellt die erweiterte Interaktion zwischen den Servern und den Clients in dem Lustre File System dar. Lustre Clients benutzen das Lustre File System. Sie

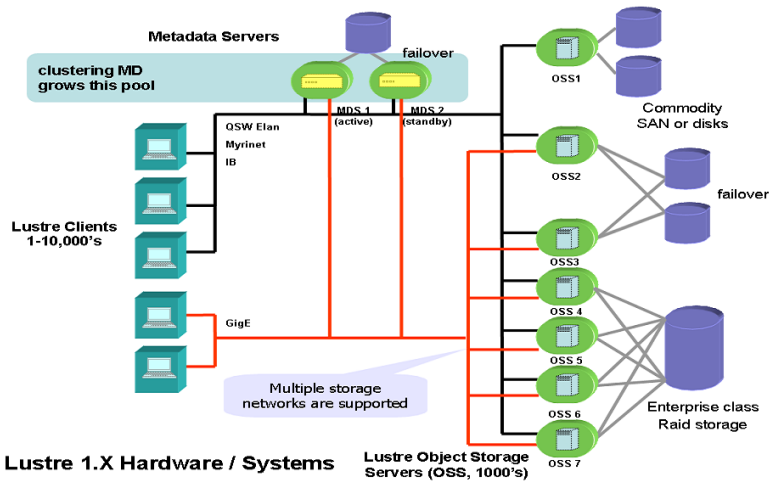

Abbildung 3.1.: Lustre Cluster

interagieren mit Object Storage Servern (OSSs) für Daten I/O und mit dem Meta Data Server (MDS) für den Metadatentransfer.

Wenn die Clients, OSS- und MDS-Systeme getrennt sind, sieht Lustre wie ein Cluster Dateisystem mit einem Dateimanager aus. Es ist aber möglich, dass alle Subsysteme in einem System laufen. Dies führt zu einem symmetrischen Layout. Die Abbildung 3.2 veranschaulicht das Main-Protokoll für die Operationen des Dateisystems.

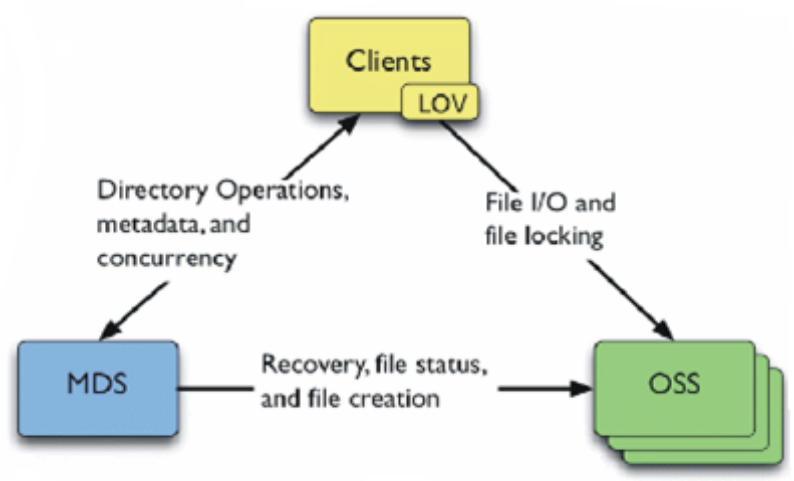

Abbildung 3.2.: Interaktion zwischen Systemen im Lustre File System

Ein großer Vorteil vom Lustre FS ist die Trennung zwischen Metadaten und echten Daten, die das System mit sich bringt, so dass die Übertragung der Nutzdaten direkt zwischen dedizierten Speicherknoten und dem generierenden Knoten abläuft. Die Datenredundanz kann durch Replikationen erreicht werden. Im Grunde genommen ist das Lustre FS nach außen als ein normales Dateisystem sichtbar.

3.1.1. Meta Data Server

Der Meta Data Server (MDS) ist wahrscheinlich das komplizierteste Lustre Subsystem. Es bietet Back-End Speicher für Metadaten Service, speichert Referenzen zu echten Daten und aktualisiert diesen Service bei jeder Transaktion über die Netzwerkschnittstelle. Das Lustre FS beinhaltet geclusterte Metadaten. Die Bearbeitung von Metadaten wird mit Hilfe von der Lastverteilung, die zur Folge hat, dass der gleichzeitige Zugriff auf die Metadaten sehr komplex ist. Die Abbildung veranschaulicht, wie ein Meta Data Server funktioniert.

Ein Meta Data Server benutzt Lockingmodule und die existierende Funktionalität eines Journal Dateisystems (z.B. Ext3[1] oder XFS[2]). Im Lustre File System wir die Komplexität wegen der Anwesenheit eines einzelnen Meta Data Server limitiert. Um single points of failure zu vermeiden, stellt das System einige failover Meta Data Services zur Verfügung, die auf existierenden Lösungen (z.B. Linux-HA[3]) basieren.

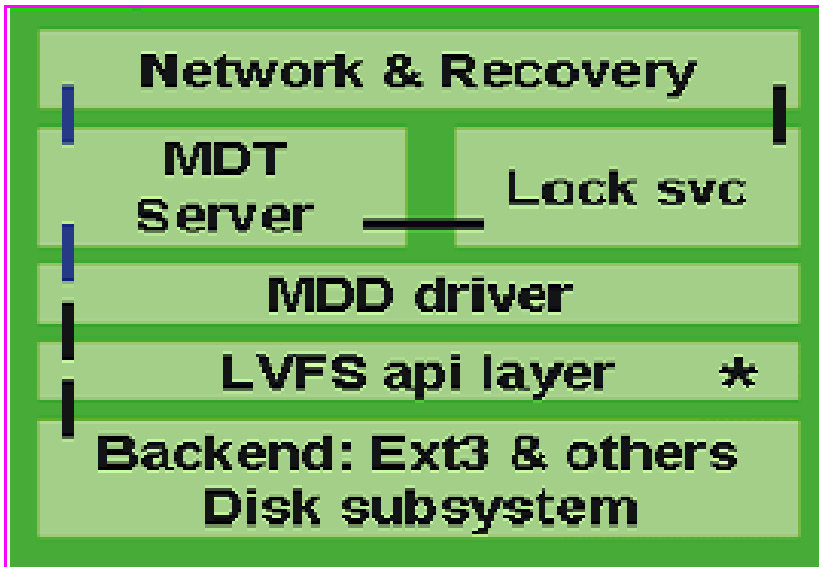

Abbildung 3.3.: Meta Data Server Software Modul

3.1.2. Object Storage Server

Das Hauptkonzept des Lustre File Systems ist Objektspeicherung. Objekte können als I-Knoten gehalten werden. Sie werden verwendet, um die echten Daten zu speichern. Ein Object Storage Server (OSS) ist ein Serverknoten, der den Lustre Software Stack betreibt. Er hat eine oder mehrere Netzwerkschnittstellen und normalerweise eine oder mehrere Festplatten. Jeder OSS exportiert ein oder mehrere Object Storage Targets (OST).

Ein Object Storage Target ist ein Softwareinterface zu einem exportierten back-end Volumen. Es ist konzeptional ähnlich zu dem Export des Network File Systems mit der Ausnahme, dass ein OST die File System Objekte anstelle vom vollständigen Namensraum enthält.

[1] third extended file system
[2] XFS ist ein von der Firma Silicon Graphics (SGI) entwickeltes Journaling-Dateisystem für UNIX-basierte Betriebssysteme wie Linux
[3] The High-Availability Linux Project http://www.linux-ha.org/

Ein OSS stellt einen Datei Input/Output Service in einem Lustre Cluster mit erleichtertem Zugang auf Objekte zur Verfügung. Der Namensraum wird durch einen Meta Daten Service bedient, der den Lustre I-Knoten handhabt. I-Knoten können Verzeichnisse, symbolische Links sein oder spezifische Devices, deren gemeinsame Daten und Metadaten auf dem Meta Data Server gespeichert sind. Wenn ein Lustre I-Knoten eine Datei repräsentiert, halten die Metadaten bloß die Referenzen auf die Datei-Datenobjekten, die auf den OSTs gespeichert werden.

Die OSTs führen die Blockverteilung für Datenobjekte aus. Dies führt zu der verteilten und skalierbaren Zuordnung von Daten. Die OSTs erzwingen auch Sicherheit bei dem Zugriff auf die Objekten von den Clients.

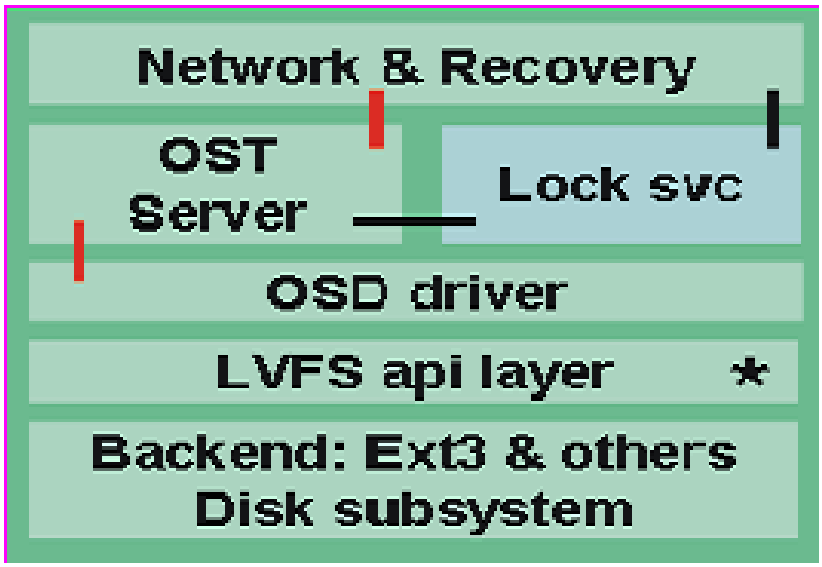

Abbildung 3.4.: Object Storage Server Software Module

Die Software Komponenten in den OSTs sind in der Abbildung 3.4 gezeigt. Diese Abbildung verdeutlicht, wie die Object Storage Targets ein Netzwerkinterface zu anderen Speicherobjekten zur Verfügung stellen.

3.1.3. Lustre FS Tools

Lustre Configuration Maker (LMC)

Mit dem Lustre Configuration Maker (LMC) werden die Konfigurationsdaten in die Konfigurationsdatei geschrieben. In der Zukunft wird es möglich, die Konfigurationsdaten aus der

Konfigurationsdatei zu löschen oder in ein anderes Format umwandeln zu können. Ein Lustre Cluster besteht aus einigen Komponenten: MDS, Mount-Points für Clients, OSTs, LOVs[4] und Netzwerken. Für ein komplettes Cluster wird eine einzige Konfigurationsdatei generiert. LMC ist ein command-line Interface.

Lustre Filesystem Configuration Utility (LCONF)

LCONF ist ein Lustre Tool, das zum Konfigurieren, Starten und Stoppen von dem Lustre File System verwendet wird. Dieses Tool benutzt zum Konfigurieren von einem Lustre Node die Konfigurationsdaten, die in der XML-Konfigurationsdatei definiert sind. Für alle Knoten in dem Cluster existiert eine einzige Konfigurationsdatei. Aus diesem Grund muss diese Datei auf allen Knoten in dem Cluster verteilt werden, oder an solch einer Stelle abgelegt werden, an der auf die Datei von allen Knoten zugegriffen werden kann.

Low Level Lustre Filesystem Configuration Utility (LCTL)

Low Level Lustre Filesystem Configuration Utility ist ein interaktives Tool. Es wird, wie der Name sagt, für die low-level Konfiguration des Lustre File Systems verwendet und kann mit dem Befehl „lctl“ aufgerufen werden. LCTL kann auch als command-line Tool gestartet werden. LCTL wird für die Konfiguration des Netzwerkes, der Devices und den Device-Operationen verwendet.

3.1.4. Sicherheit

Die Sicherheit des Dateisystems ist ein sehr wichtiger Aspekt eines verteilten Dateisystems. Die Standardaspekte der Sicherheit sind Authentifikation, Autorisation und Verschlüsselung. Das Lustre File System unterstützt Object Storage Server mit Network Attached Secure Disks (NASD) [NASD].

Das Lustre FS kann mit einem existierendem Authentifikationmechanismus einfach integriert werden. Das Authentifikationmechanismus benutzt Generic Security Service Application Programming Interface (GSS-API). Das GSS-API ist ein offener Standard, der Sessionmanagement, Authenifizierung, Datenintegrität und Vertraulichkeit von Daten zur Verfügung stellt. Die Kerberos 5 [RFC4120] und PKI Mechanismen werden als Backend für die Authentifizierung in dem Lustre FS benutzt. Der Ablauf eines Lesezugriffes in dem Lustre File System ist in der Abbildung 3.5 dargestellt.

[4] LOV - Logical object volume

In den nächsten Versionen wird das Lustre File System die Autorisierung durch Access Control Lists (ACL) unterstützen, die der POSIX[5] ACL Semantik folgen. Die Flexibilität und zusätzliche durch ACLs zur Verfügung gestellte Fähigkeiten sind in Clustern besonders wichtig, weil auf Grund dessen Tausende von Knoten und Benutzerkonten unterstützt werden können.

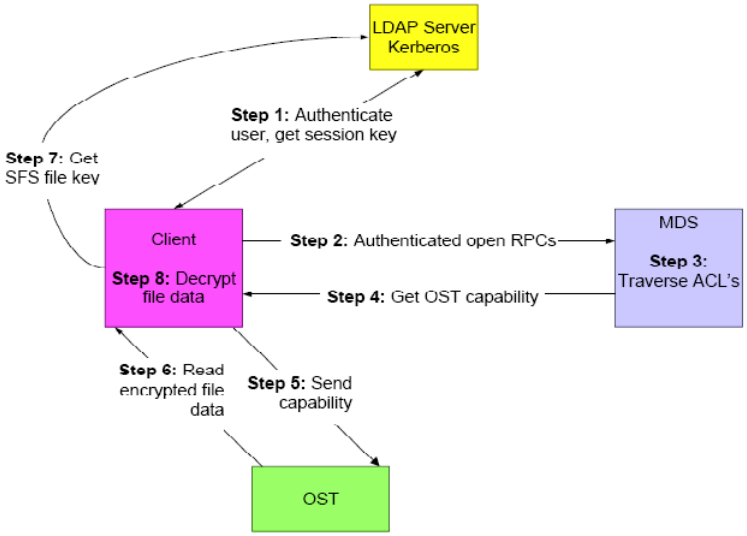

Abbildung 3.5.: Lustre Read File Security

3.2. IBM BladeCenter

In dem „Collaborative Workspace“ Projekt ist ein IBM BladeCenter HS20 [IBMBladeCenterHS20] als Hardware vorgesehen. IBM BladeCenter HS 20 ist keine Standard-Hardwarekomponente. Der Autor dieses Berichtes hatte aus diesem Grund Probleme, die in dem nächsten Abschnitt 3.3 beschrieben werden.

Wie eine IP-Adresse für einen IBM Blade Server vergeben wird, ist im Anhang A.1 beschrieben.

[5] POSIX - portabeles Betriebsysteminterface, das auf UNIX basiert

3.3. Problematik: Lustre File System / IBM BladeCenter

Bei dem ersten Versuch wurde das Lustre File System v.1.4.7.x auf einem Standard-Rechner installiert. Die Installation wurde ohne große Schwierigkeiten auf Suse Linux Enterprise Server (SLES) v.9 (64bit) als Betriebssystem durchgeführt. Das Lustre File System wurde als Single Node Lustre [Cluster File Systems, Inc. (2006)] konfiguriert, und erste Tests wurden erfolgreich absolviert.

Es ist gewünscht, dass IBM Blade Server mit dem Lustre File System benutzt werden. Zuerst sollte natürlich ein Betriebssystem installiert werden. Die Installation von Suse Linux Enterprise Server v.9 schlug fehl und endtet immer mit der Ausgabe: „looking for info files“. Genaue Ursachen des Problems wurden leider auch mit Hilfe von Professoren und Mitarbeitern nicht gefunden. Es wird vermutet, dass der SLES 9 keine passenden Treiber für die neue Hardware des IBM BladeCenters zur Verfügung hat und aus diesem Grund die Installation fehlschlug.

Danach wurde Suse Linux Enterprise Server v.10 (64bit) als Betriebssystem auf den IBM Blade Servern erfolgreich aufgesetzt. Sich darauf zu freuen, war aber zu früh, da sich das Lustre File System v.1.4.7.x darauf nicht korrekt installieren ließ. Es gab folgendes Problemt mit einigen Bibliotheken: Mit den Unixbefehlen wurde festgestellt, dass die betreffenden Bibliotheken 64 bit Versionen sind. Das Lustre FS meldete dagegen, die Bibliotheken seien 32 bit Versionen. Der Autor dieses Berichtes nahm Kontakt mit dem Lustre-Support auf. Deren Aussage war, dass das Lustre File System v.1.4.7.x kein Suse Linux Enterprise Server v.10 unterstützt, die Version 1.5.x aber den SLES 10 unterstützen wird. Die Version 1.5.x soll im April/Mai 2007 erscheinen. Der Versuch, die beta- 1.5.x Version des Lustre File Systems zu erhalten, war leider erfolglos.

Weitere Versuche, das Lustre File System v.1.4.7.x auf anderen Betriebssystemen wie Red Hat Enterprise Linux v.3 und Red Hat Enterprise Linux v.4 zu betreiben, waren leider auch erfolglos. Der Autor dieses Berichtes konnte die genauen Ursachen des 32/64 bit Versionsmismatches einiger Bibliotheken leider nicht feststellen.

Die oben beschriebene Problematik wurde berücksichtigt, und aus diesen Gründen wurde entschieden, das Lustre File System v.1.4.7.x erst auf mehreren Standard-Rechnern zu installieren.

Schon während des Schreibens dieses Berichtes stellte der Autor fest, dass die Version 1.4.9 des Lustre File Systems [LustreFS:1.4.9] zur Verfügung steht. Diese Version sollte den SLES 10 unterstützen. Dies wird im Laufe der Master Thesis von dem Autor dieses Berichtes untersucht.

4. Zusammenfassung und Ausblick

In dem „Collaborative Workspace“ Projekt wurden nicht alle Ziele erreicht, die sich die Projektbeteiligten stellten. Die Gründe dafür sind nicht nur umfangreiche Themengebiete, sondern auch die notwendige Investition der Zeit in die Organisation des Projektes. Da am Startpunkt des Projektes die Projektbeteiligten kein eigenes Labor hatten, wurde sehr viel Zeit in den notwendigen Laboraufbau investiert. Leider wurde die für das „Collaborative Workspace“ Projekt vorgesehene Hardware zu spät geliefert. Dies bedeutete mindestens einen doppelten Zeitaufwand für den Laboraufbau.

Ein verteiltes Dateisystem ist kein System, das man innerhalb von einigen Stunden aufsetzt. Das Lustre File System ist ein komplexes, verteiltes Dateisystem, das viel Funktionalität mit sich bringt. Dies bedeutet einen höheren Konfigurationsaufwand.

Mit diesem Projektbericht beschreibt der Autor seine im Rahmen von dem „Collaborative Workspace“ Projekt gewonnenen Erkenntnisse und Problematiken, die während der Arbeit aufgetreten sind.

Die Erkenntnisse des Projektes sowie die Erkenntnisse aus den Veranstaltungen „Anwendungen II“ [Kabalkin (2007b)] und „Seminar-Ringvorlesung“ [Kabalkin (2007a)] gelten als Grundlagen für die Master-Thesis des Autors.

Die Arbeit mit dem Lustre File System wird im Rahmen der Master-Thesis fortgesetzt. Mit der neuen Version des Lustre File Systems wird ein weiterer Versuch durchgeführt werden, das verteilte Dateisystem auf einem IBM BladeCenter HS20 zum Laufen zu betreiben.

Literaturverzeichnis

[Blade] : *Blade Center*. – URL http://www-03.ibm.com/systems/de/bladecenter/

[NFSv4] : *General Information and References for the NFSv4 protocol*. – URL http://www.nfsv4.org/

[IBMBladeCenterHS20] *IBM BladeCenter HS20*. – URL http://www-03.ibm.com/systems/de/bladecenter/hs20/

[LustreFS:1.4.9] : *Lustre File System v. 1.4.9 Download*. – URL http://downloads.clusterfs.com/customer/public-releases/production/1.4.9

[NASD] : *Network Attached Secure Disks (NASD)*. – URL http://www.pdl.cmu.edu/NASD/

[RFC4120] : *RFC4120, The Kerberos Network Authentication Service (V5)*. – URL http://tools.ietf.org/html/rfc4120

[Cluster 2002] CLUSTER, Filesystem: *Lustre: A Scalable, High-Performance File System*. November 2002. – URL http://www.lustre.org/docs/whitepaper.pdf

[Cluster File Systems, Inc. 2006] Cluster File Systems, Inc. (Veranst.): *Lustre 1.4.7 Operations Manual*. Version 1.4.7.1-man-v36 (11/30/2006). November 2006. – URL https://mail.clusterfs.com/wikis/lustre/LustreDocumentation?action=AttachFile&do=get&target=LustreManual36.pdf

[Kabalkin 2007a] KABALKIN, Mykhaylo: *Persistenz-Service System*. Februar 2007. – URL http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master06-07/kabalkin/abstract.pdf

[Kabalkin 2007b] KABALKIN, Mykhaylo: *Verteilte Dateisysteme*. Februar 2007. – URL http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master06-07-aw/kabalkin/abstract.pdf

[Noveck u. a. 2003] NOVECK, D. ; SHEPLER, S. ; CALLAGHAN, B. ; ROBINSON, D. ; THURLOW, R. ; BEAME, C. ; EISLER, M.: *Network File System (NFS) version 4 Protocol*. April 2003. – URL http://tools.ietf.org/html/rfc3530

[Ramamurthy 2004] RAMAMURTHY, Bina: *Destributed File Systems*. September 2004. – URL http://www.cse.buffalo.edu/gridforce/fall2004/DistributedFileSystemSept29.pdf

A. Anhang

A.1. IP-Adresse für IBM Blade Center vergeben.

Damit ein IBM Blade Server eine IP-Adresse erhalten kann, sind folgende Schritte notwendig:

- Externe Ports müssen über das Advanced Management Modul (AMM) (192.168.70.125) aktiviert und auf VLAN 2 gestellt werden.
- Telnet Verbindung auf 192.168.70.127 als Administrator mit „USERID“ als Benutzername und „PASSWØRD“ als Passwort
- Anschließen folgende Befehle eingeben:
  configure terminal
  interface gigabit 0/17
  switchport access vlan 2 (mit „End“ wird configure beendet)
  write memory

Diese Information stammt von dem Intranet Service Center der Hochschule für Angewandte Wissenschaften Hamburg.