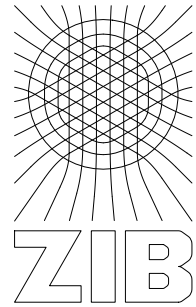
Konrad-Zuse-Zentrum
für Informationstechnik Berlin



Takustraße 7
D-14195 Berlin-Dahlem
Germany

HUBERT BUSCH

Abschlußbericht des DFN-Projekts „Virtueller Supercomputer Berlin – Hannover“

ZIB-Report 04-57 (Dezember 2004)

Abschlußbericht

Virtueller Supercomputer
Berlin – Hannover

Projektbeteiligte:

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)
Regionales Rechenzentrum für Niedersachsen (RRZN)
Fa. IBM Deutschland

Projektlaufzeit:

15.02.2003 – 29.02.2004

Projektleitung:

Dipl.-Math. Hubert Busch (ZIB)

Juli 2004

Dieses FE-Vorhaben wurde vom Verein zur Förderung eines Deutschen Forschungsnetzes e.V. gefördert.

Virtueller Supercomputer Berlin – Hannover

Abschlußbericht

Projektleiter:	Dipl.-Math. Hubert Busch (ZIB)
Projektlaufzeit:	15.02.2003 – 29.02.2004
Hauptantragsteller:	Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB) Prof. Dr. Alexander Reinefeld
Mitantragsteller:	Universität Hannover Regionales Rechenzentrum für Niedersachsen (RRZN) Prof. Dr.-Ing. Gabriele von Voigt
Kooperationspartner:	Fa. IBM Deutschland GmbH
Ansprechpartner:	Dipl.-Math. Hubert Busch, <busch@zib.de> Dipl.-Inf. Thomas Röblitz, <roeblitz@zib.de> Dipl.-Inf. Sebastian Heidl, <heidl@zib.de> Dipl.-Math. Wolfgang Kamps, <kamps@rrzn.uni-hannover.de> Dipl.-Ing. Steffen Heinze, <heinze@rrzn.uni-hannover.de>

1 Beschreibung des Projekts

Gefördert durch den Verein zur Förderung eines Deutschen Forschungsnetzes e.V. (DFN) begann im Februar 2003 unter Federführung des Konrad-Zuse-Zentrums für Informationstechnik Berlin (ZIB) und des Regionalen Rechenzentrums Niedersachsen (RRZN) das Projekt „Virtueller Supercomputer Berlin-Hannover“. Im Mittelpunkt des Projekts steht die Bereitstellung der Ein-System-Eigenschaft für die Nutzer des Hochleistungsrechners HLRN. Zur Erläuterung der Ziele und Aufgaben innerhalb des Projekts ist in diesem ersten Abschnitt ein Ausschnitt (Kapitel zwei) aus dem Projektantrag wiedergegeben. Durch die Integration in dieses Dokument kann die Numerierung der Teilabschnitte von der im Projektantrag abweichen.

1.1 Projektziel

1.1.1 Zusammenfassung

Mit der Installation des neuen Hochleistungsrechners für die norddeutschen Länder (HLRN[1]) steht den Wissenschaftlern ein außergewöhnlich leistungsfähiges System zur Verfügung. Durch die Verteilung der Rechenelemente auf zwei verschiedene Standorte in Berlin (ZIB) und Hannover (RRZN) entstehen jedoch auch neue Herausforderungen für den Betrieb und die effiziente Nutzung des Rechners. Inhalt dieses Projektes sollen die Erforschung und Lösung der durch die Verteilung des Systems hervorgerufenen Probleme (z.B. Scheduling, Kommunikation, I/O) sein. Es sollen effiziente Lösungen zur Bereitstellung eines virtuellen, hoch-performanten und transparenten Systems entwickelt werden, die auf vergleichbare Installationen übertragbar sind.

[1] http://www.hlrn.de/

1.1.2 Angestrebte Ergebnisse

Die Bereitstellung der Ein-System-Eigenschaft aus Nutzersicht steht im Vordergrund des Projekts. Die durch die Verteilung der Rechenressourcen vorhandene Komplexität soll, soweit als möglich, verborgen bleiben und Leistungseinbußen, die zweifellos auftreten werden, sollen möglichst gering gehalten werden. Bestehen jedoch Anforderungen seitens der Nutzer, zu Optimierungszwecken auf Informationen über die Struktur des Systems zuzugreifen, so sollen diese Informationen möglichst einfach verfügbar gemacht werden.

1.1.3 Nutzung der Projektergebnisse

Die Ergebnisse des Projekts fließen direkt in den Betrieb des HLRN ein und unterstützen die Bereitstellung der Ein-System-Eigenschaft des Hochleistungsrechners. Darüber hinaus sollen die entwickelten Komponenten, gemäß dem Projektziel, in vergleichbaren Installationen verwendbar sein.

1.2 Inhaltlicher Hintergund

Auf der Basis der jahrelangen Erfahrungen im Norddeutschen Vektorrechner-Verbund und der Empfehlungen des Wissenschaftsrates aus dem Jahre 1995 zum Wissenschaftlichen Hochleistungsrechnen (High Performance Scientific Computing - HPSC) haben die sechs Bundesländer Berlin, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen und Schleswig-Holstein die gemeinsame Beschaffung eines zur Bearbeitung von Grand Challenges geeigneten Computersystems als Investitionsvorhaben für die Jahre 2001 -2003 beim Wissenschaftsrat angemeldet (Höchstleistungsrechenzentrum Nord (HLRN)). Alle beteiligten Landesregierungen haben einem Verwaltungsabkommen zugestimmt mit dem Ziel der „ *... gemeinsame[n] Förderung des Hoch- und Höchstleistungsrechnens in der Absicht, die bestehende regionale Infrastruktur in Wissenschaft und Wirtschaft durch den Aufbau und Betrieb eines Norddeutschen Verbundes für Hoch- und Höchstleistungsrechnen (HLRN-Verbund) als gemeinsame Verbundaufgabe zu verbessern.*“[2] Nach der Zustimmung des Wissenschaftsrates im Januar 2000 konnte der Hochleistungsrechner im 1. Quartal 2002 beschafft werden.

Das HLRN-System ist im Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB[3]) und am Regionalen Rechenzentrum für Niedersachsen in Hannover (RRZN[4]) installiert. Das System besteht aus 26 IBM p690 Knoten, von denen 13 am ZIB und 13 am RRZN aufgestellt worden sind. Jeder Knoten ist mit 32 Power4 CPUs mit einer Taktfrequenz von 1,3 GHz ausgestattet. Das Gesamtsystem verfügt über 832 Prozessoren mit einer akkumulierten Rechenleistung von 4,3 TeraFlop/s[5], die insgesamt über zwei Terabyte[6] Arbeitsspeicher zur Verfügung haben. Der Zugriff auf die großen Dateisysteme erfolgt über ein SAN-ähnliches (*Storage Area Network*) Magnetplattensystem mit einer Gesamtkapazität von 52 Terabyte. Die beiden Teilkomplexe sind über eine dedizierte Leitung im GWin mit einer Übertragungsrate von 2,4 Gbit/s miteinander verbunden. Abbildung 1 zeigt die Konfiguration des HLRN-Systems.

Auf dem HLRN-System kommt ein breites Spektrum von parallelen Programmen zum Einsatz. Die Hauptanwendungsgebiete für parallele Berechnungen finden sich in der Klima-, Küsten- und Meeresforschung sowie der Grundlagenforschung von Physik, Chemie und den Life-Sciences.

Durch die besondere Situation der HLRN-Installation steht den Projektteilnehmern ein in dieser Form einmaliger Hochleistungsrechner zur Verfügung. Im Rahmen dieses Projekts bietet sich die Chance, bisher getrennt betrachtete Aspekte des parallelen, verteilten Rechnens in ein System zu integrieren. In vier

[2] Quelle: http://www.hlrn.de/verwaltungsabkommen/verwaltungsabkommen.pdf
[3] http://www.zib.de/
[4] http://www.rrzn.uni-hannover.de/
[5] 1 TeraFlop/s entspricht 10^{12} Gleitkomma-Operationen pro Sekunde
[6] 1 Terabyte $= 2^{40}$ Byte

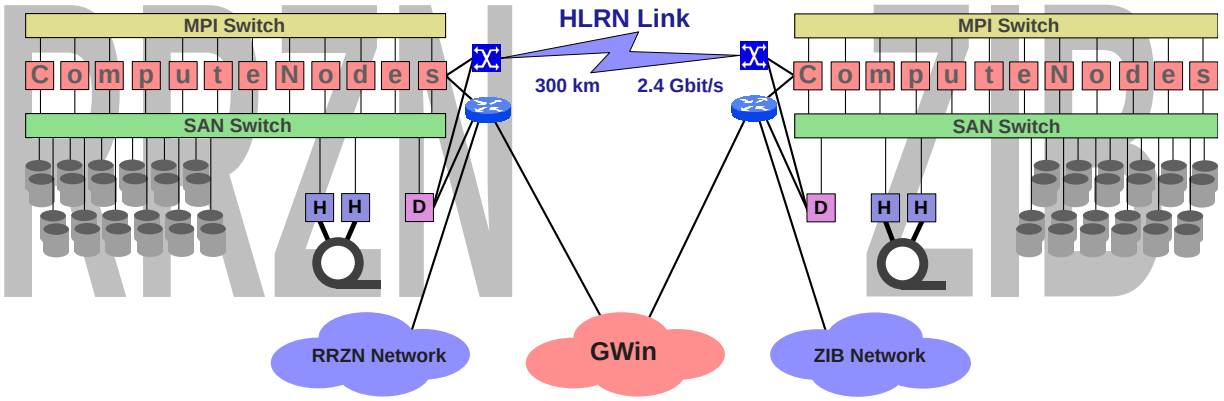


Abbildung 1: Hardware-Konfiguration des HLRN-Systems

verschiedenen Teilprojekten werden Lösungen für Probleme erarbeitet, die in diesem Zusammenhang auftreten können.

Im ersten Teilprojekt geht es um die gemeinsame Nutzung des HLRN-Links durch alle Teilsysteme des HLRN. Hier stehen Anforderungen nach *Quality of Service* (z.B. Bandbreitenbegrenzung, Bandbreitenreservierung) und Fairneß im Vordergrund. Im nächsten Teilprojekt werden Lösungen für die entfernte Nutzung der Dateisysteme gesucht. Ein weiteres Teilprojekt beschäftigt sich mit der effizienten Kopplung von parallelen Programmen, die auf beiden Teilkomplexen verteilt ausgeführt werden. Das vierte Teilprojekt behandelt die gemeinsame Ressourcenverwaltung der beiden Teilkomplexe.

Diese verschiedenen Teilprojekte werden in ein Gesamtsystem zur Verwaltung und effizienten Nutzung eines verteilten Hochleistungsrechners zusammengeführt.

1.3 Informations- und Kommunikationstechnische Beschreibung

Die Konfiguration der HLRN-Knoten ist an beiden Standorten identisch. Es wurden je 13 IBM p690 Systeme installiert. Jedes einzelne System läßt sich flexibel partitionieren, wobei eine bestimmte Anzahl von CPUs, Hauptspeicher und I/O-Steckplätzen einer sogenannten LPAR (Logical Partition) zugeordnet werden. Jedes p690 System des HLRN ist zur Zeit in 4 symmetrische LPARs unterteilt, die mit je acht Power4 Prozessoren, zwei SP Switch2 Adaptern und einem Viertel des Hauptspeichers der jeweiligen Maschine ausgestattet sind. In allen LPARs wird eine eigene Instanz des AIX Betriebsystems ausgeführt. Somit sind sie vollkommen unabhängig voneinander. Die Anbindung nach außen wird über zwei dedizierte LPARs realisiert, die auch als Login-Knoten fungieren. Sie sind über Gigabit-Ethernet an den HLRN-Link angeschlossen und in das GWiN integriert.

Für die Kommunikation während paralleler Berechnungen sind die LPARs in einem Hochgeschwindigkeitsnetzwerk integriert, auf das über die beiden SP Switch2 Adapter zugegriffen wird. Bei der Verteilung des Programms auf beide Standorte müssen die Teilkomponenten über den HLRN-Link kommunizieren.

Der Zugriff auf die Dateisysteme erfolgt an beiden Standorten über ein eigenständiges leistungsfähiges Plattensystem. Die Heimatverzeichnisse der Benutzer werden nur an einem Standort gespeichert und am jeweils anderen über den HLRN-Link zur Verfügung gestellt. Alle anderen Dateisysteme sind nur lokal an einem Standort verfügbar.

Die Ressourcenverwaltung des HLRN-Systems wird durch die LoadLeveler-Software koordiniert, überwacht und durchgeführt. Hier wurden zunächst zwei getrennte, jedoch identisch konfigurierte, LoadLeveler-Instanzen eingesetzt, von denen jede die Systeme eines Standorts bearbeitet.

1.3.1 Beschreibung der Anwendungsszenarien

Die beiden Komplexe des HLRN-Systems sind über eine dedizierte Netzwerkverbindung des DFN miteinander gekoppelt. Diese Verbindung ist über einen transparenten WDM-Kanal mit einer Übertragungsgeschwindigkeit von 2,4 Gbit/s realisiert. Trotz dieser hohen Übertragungsleistung entstehen an dieser Stelle eine Reihe von Problemen, da die interne Vernetzung der Teilkomplexe weitaus leistungsfähiger ist.

Während des Betriebs des HLRN werden folgende Teilsysteme/Komponenten auf die Netzwerkverbindung zugreifen:

1. Scheduler/Load-Balancer
2. I/O-Systeme, insbesondere die Systeme zur Speicherung des Dateisystems `/perm` und die Systeme zur Bereitstellung der Home-Verzeichnisse. Dabei ist zu beachten, daß bei der Speicherung von Daten in `/perm`, die auf Magnetbändern erfolgt, eine Sicherungskopie auf dem Magnetbandsystem des jeweils anderen Komplexes angelegt werden soll.
3. parallele Programme, deren Komponenten auf beide Komplexe verteilt sind
4. Nutzer, die Daten für eine Berechnung auf dem jeweils anderen Komplex lokal zur Verfügung stellen wollen.

In der gegenwärtigen Konfiguration erfolgt die Verbindung der Systeme durch eine Umsetzung von Gigabit-Ethernet (GbE) auf den HLRN-Link. Hierbei wird die zur Verfügung stehende Bandbreite nur zu ca. 40 % ausgenutzt, da bei GbE die Übertragungsrate auf 1 Gbit/s begrenzt ist. Durch diese Einschränkung verringert sich die verfügbare Bandbreite von theoretischen 286 MB/s auf ca. 120 MB/s.

Ausführung paralleler Programme

Bedingt durch die geographische Trennung des Systems entstehen eine Reihe von Abhängigkeiten, die bei der Entscheidung über die Ressourcenzuordnung für ein paralleles Programm zusätzlich zu den üblichen Anforderungen berücksichtigt werden müssen.

So sind zum Beispiel die Daten für eine Berechnung gewöhnlich im Heimatverzeichnis beziehungsweise auf den Dateisystemen des Heimatkomplexes gespeichert. Wenn nun auf dem Heimatkomplex nicht genügend Rechenressourcen verfügbar, jedoch auf dem „entfernten“ Komplex noch Kapazitäten frei sind, dann muß das RMS (Ressource Management System) entscheiden, ob das Programm dort ausgeführt werden kann. Dabei muß berücksichtigt werden, daß die Programmdaten unter Umständen zum Ausführungsort migriert werden müssen, um Verzögerungen während des Zugriffs über das Netzwerk zu vermeiden.

Ein weiteres Problem stellen parallele Programme dar, die auf beiden Komplexen verteilt rechnen. Hier muß die Übertragungskapazität des HLRN-Links zwischen diesen Programmen und den anderen Subsystemen des HLRN aufgeteilt werden.

Eine Reihe weiterer Implikationen und Herausforderungen bei der Ausführung paralleler Programme, die durch die Verteilung der Systemkomponenten entstehen, werden untersucht und im Laufe des Projektes bearbeitet werden.

Datenhaltung

Wie bereits beschrieben, werden die Daten der Benutzer nur an einem der beiden Standorte des HLRN (dem Heimatstandort) gespeichert und am jeweils anderen über den HLRN-Link zur Verfügung gestellt. Dafür muß kontinuierlich ein gewisser Anteil der Bandbreite des HLRN-Links vorgesehen werden, um Verzögerungen beim Datenzugriff zu vermeiden.

Darüberhinaus verfügen beide Standorte über ein Magnetbandsystem zur Datensicherung. Hier soll überprüft werden, ob durch das Ablegen einer Kopie der Daten auf dem Backup-System des anderen Komplexes eine größere Datensicherheit gewährleistet werden kann. Auch dafür muß gegebenenfalls ein gewisser Anteil der Übertragungskapazität des HLRN-Links zur Verfügung stehen.

1.3.2 Projektgliederung

Teilprojekt 1: Begleitende Infrastrukturmaßnahmen

Begleitend zu den drei anwendungsorientierten Teilprojekten 2 – 4 werden im Teilprojekt 1 während der gesamten Projektlaufzeit Arbeiten an der Netzwerkinfrastruktur durchgeführt. Hierzu gehören neben der Inbetriebnahme der Netzverbindung zwischen Berlin und Hannover die jeweilige Integration in die lokalen Teilnetze, die Beobachtung des Betriebs und Führen von Statistiken, auch um Daten für die Entscheidung des Ausbaus der Verbindung auf 10 Gb/s bereit zu stellen.

Nach einer an sich wünschenswerten Entscheidung zu Gunsten einer Aufrüstung auf 10 Gb/s werden in diesem Teilprojekt die genannten Arbeiten an dieser Verbindung fortgesetzt. Sollte die Aufrüstung nicht realisierbar sein, so müssen technische Alternativen zur besseren Ausnutzung der vorhandenen 2,4 Gb/s Verbindung diskutiert und nach Möglichkeit umgesetzt werden.

Teilprojekt 2: Entfernte Nutzung der Dateisysteme

Die effiziente und performante Nutzung der I/O-Systeme des HLRN steht im Mittelpunkt dieses Teilprojekts. Auch hier stellt die Kopplung der jeweiligen Teilsysteme jedes Standorts zu einem Gesamtsystem die größte Herausforderung dar.

Die zu realisierende Verbindung der I/O-Systeme der Teilkomplexe soll für die Nutzer des HLRN und die weiteren Subsysteme transparent sein. Insbesondere folgende Nutzungsszenarien dieser Verbindung sollen unterstützt werden:

1. Da die Daten der Nutzer nur an einem der beiden Standorte gespeichert werden, müssen sie beim interaktiven Arbeiten auf dem jeweils anderen Komplex beziehungsweise bei der Ausführung einer parallelen Applikation dort zugreifbar sein.

2. Zur Sicherung großer Datenmengen sowie zur langfristigen Archivierung von Daten wird auf beiden Teilkomplexen des HLRN ein jeweils lokales Dateisystem `/perm` zur Verfügung gestellt. Diese Dateisysteme sind, ähnlich wie die Heimatverzeichnisse der Benutzer, nur auf dem jeweiligen Teilkomplex verfügbar, der sie bereitstellt, und werden auf Magnetbandsystemen gespeichert. Um eine größere Datensicherheit und gegebenenfalls auch einen lokalen Zugriff zu gewährleisten, ist eine Sicherung des Dateisystems `/perm` sowohl auf dem jeweils lokalen Magnetbandsystem als auch auf dem Magnetbandsystem des anderen Teilkomplexes zu untersuchen.

Teilprojekt 3: Kopplung von parallelen Programmen

Durch die interne Vernetzung der beiden Teilsysteme des HLRN über die SP Switch2 Technologie steht parallelen Applikationen innerhalb eines Komplexes ein Netzwerk mit einer theoretischen Leistung von 400 MB/s pro LPAR bei einer MPI-Latenzzeit von 20 μs zur Verfügung. Im Vergleich dazu bietet die

Netzwerkverbindung der beiden Komplexe zur Zeit eine Bandbreite von ca. 120 MB/s und eine Latenzzeit von ca. 1,7 ms, die darüber hinaus noch mit anderen parallelen Anwendungen und weiteren Teilsystemen des HLRN geteilt werden muß. Selbst bei optimaler Ausnutzung des HLRN-Links erhöht sich die maximal verfügbare Bandbreite lediglich auf ca. 286 MB/s, wobei die Latenzzeit unverändert bleibt. Somit ergibt sich eine asymmetrische Verteilung der verfügbaren Kommunikationsleistung zwischen verschiedenen Komponenten eines verteilt ausgeführten parallelen Programms, die dessen Leistung entscheidend beeinflussen kann.

Die folgenden Arbeiten sind im Rahmen dieses Teilprojekts vorgesehen:

1. Untersuchung der Anforderungen der parallelen Programme, für die eine Ausführung auf dem gesamten HLRN-System „interessant“ ist. Voraussichtlich kommen hier vorrangig lose gekoppelte Systeme in Frage, bei denen die einzelnen Teilkomponenten wenig miteinander kommunizieren. Nichtsdestotrotz sollten auch Programme mit eng gekoppelten Teilsystemen in die Tests aufgenommen werden, um die tatsächlichen Anforderungen an die Optimierung der Kommunikation zu bestimmen.

   Beispielhaft sei hier auf das von der DFG bewilligte Vorhaben[7] „Hochauflösende Grobstruktursimulationen turbulenter Strömungen im Einflussbereich von Gebäuden unter Berücksichtigung thermischer Effekte“ des Instituts für Meteorologie und Klimatologie der Universtität Hannover hingewiesen. Dieses Projekt wird das verteilte HLRN-System als ganzes nutzen. Auf jedem Teilkomplex werden Teilaufgaben mit hohen Anforderungen an die Latenz der Kommunikationsleistung abgewickelt, die Teilaufgaben untereinander können getrennt auf den Teilkomplexen mit der dagegen durch die Entfernung bedingten erheblich höheren Latenzzeit ohne wesentliche Verschlechterung der Performance der Rechnungen abgearbeitet werden.

2. Entwicklung eines flexiblen Systems, das in der Lage ist, die Heterogenität der Netzwerkverbindungen für die Anwendungen vollständig transparent zu halten und das zusätzlich Applikationen die Möglichkeit bietet, Informationen über die Netzwerkstruktur für eine Umstrukturierung zu nutzen. Als Basis für eine Implementation dieses Systems ist MPICH-G2 denkbar.

   Folgende Anforderungen soll dieses System erfüllen:

   (a) Idealerweise sollen parallele Programme, die mehr als einen Teilkomplex des HLRN benutzen, mit akzeptabler Leistung unverändert auf beiden Teilkomplexen ausgeführt werden können.

   (b) Es sollen keine zusätzlichen Prozessoren für die „externe“ Kommunikation beansprucht werden, wie es zum Beispiel bei PACX-MPI[8] der Fall ist.

   (c) Parallelen Anwendungen sollen Informationen über die Netzwerktopologie bzw. über die Verteilung der ihnen zugewiesenen Prozessoren auf die Teilsysteme des HLRN zugänglich sein. So ist ggf. eine Umstrukturierung, bzw. Optimierung des Progamms bezüglich der asymmetrischen Kommunikationsleistung möglich.

Teilprojekt 4: Ein Ressource-Management-System für beide Teilkomplexe des HLRN

Das Ressource-Management-System (RMS) soll den Nutzern die transparente und gradlinige Nutzung aller Ressourcen des Hochleistungsrechners ermöglichen. Die Anforderungen an das RMS stammen sowohl von den Nutzern des HLRN, beziehungsweise den von ihnen gestarteten Applikationen, als auch von den Administratoren des Systems.

[7] DFG-Geschäftszeichen RA 617/6-1
[8] http://www.hlrs.de/organization/pds/projects/pacx-mpi/

Im folgenden werden die Anforderungen beschrieben, die im Rahmen dieses Teilprojekts im RMS zu implementieren sind.

1. Dem Nutzer soll ein einheitliches Verfahren zum Abschicken von Jobs über das gesamte HLRN-System hinweg zur Verfügung gestellt werden. Hier steht die Realisierung der Ein-System-Eigenschaft im Vordergrund.

2. Für parallele Applikationen soll der Zugriff auf Berechnungsdaten, unabhängig von deren Speicherort, vollständig transparent sein. Außerdem sollen den Applikationen Informationen über die reservierten Ressourcen für etwaige Restrukturierungen zur Verfügung stehen.

3. Aus administrativer Sicht und auch aus Performance-Gründen sollen Ressourcen für Applikationen soweit wie möglich auf einem Komplex des Systems lokalisiert werden. So soll die Kommunikation über den HLRN-Link vermieden werden, da sie Leistungseinbußen bei der Programmausführung erwarten läßt. Darüber hinaus können so Probleme vermieden werden, die bei einer Störung der Kommunikation über den HLRN-Link auftreten, wenn ein paralleles Programm über beide Komplexe des Systems verteilt ist.

   Sollte eine Applikation auf beiden Teilkomplexen ausgeführt werden, so müssen im Fall, daß während der Ausführung einer solchen „verteilten“ Applikation der HLRN-Link ausfällt, die ihr zugewiesenen Ressourcen gegebenenfalls automatisch einer Nutzung durch andere Jobs zugeführt werden.

2 Durchführung des Projekts

Dieser Abschnitt beschreibt die Ausgangssituation des Projekts. Es werden die an den Projektarbeiten beteiligten Personen vorgestellt und die Rechner-, Netz-, Dateisystem- und Software-Konfiguration des HLRN-Systems zum Zeitpunkt des Projektstarts erläutert.

2.1 Personal

Die folgenden Personen sind an der Durchführung des Projektes beteiligt:

ZIB
: Hubert Busch (Projektleiter), Sebastian Heidl, Matthias Heyder, Thomas Röblitz

RRZN
: Sebastian Boesler, Steffen Heinze, Dr. Fritz Hüsemann, Wolfgang Kamps, Sieghart Ludwig

IBM
: Dr. Roland Kunz, Jakob Pichlmeier

2.2 Zusammenarbeit mit Fa. IBM

Im Rahmen des Vertrages über den Kauf des HLRN-Systems mit Fa. IBM ist auch die Zusammenarbeit im Rahmen dieses Projekts vereinbart worden. Fa. IBM erbringt Leistungen im Zusammenhang mit erweiterten Features des Resource-Management-Systems LoadLeveler und der Kopplung von parallelen Programmen über den HLRN-Link.

2.3 Systemkonfiguration während des ersten Projektteils

2.3.1 Rechnerkonfiguration

Wie im Abschnitt 1.3 beschrieben, sind am ZIB und am RRZN je 13 IBM p690 Systeme installiert worden. Diese Systeme sind weitestgehend identisch und mit je 32 Power4 Prozessoren und je acht SP

Switch2 Adaptern ausgestattet. Sie unterscheiden sich jedoch in der Menge des installierten Hauptspeichers. Es wurden an jedem Teilkomplex elf Systeme mit 64 GB, ein System mit 128 GB und ein System mit 256 GB Hauptspeicher installiert

Die Gründe für die unterschiedliche Partitionierung sind das Bestreben, den Anforderungen unterschiedlicher Anwendungsprofile gerecht werden zu können und den Betrieb ab Ende 2004 vorzubereiten, in dem es ausschließlich Compute-Knoten mit 32 CPUs geben wird. Abbildung 2 zeigt die unterschiedliche Konfiguration der Partitionen. Zusätzlich ist in dieser Abbildung die Zuordnung der einzelnen Partitionen zu bestimmten LoadLeveler-Klassen farblich dargestellt (siehe Abschnitt 2.3.4). Bei `login` handelt es sich allerdings nicht um eine LoadLeveler-Klasse. LPARs mit diesem Merkmal üben zusätzlich eine Funktion als Login-Knoten aus (siehe Abb. 3).

Jeder Quader im Bild 2 stellt eine logische Partition dar, wobei jeweils vier Quader (LPARs) mit einer kleinen Grundfläche physikalisch in einer p690 Maschine lokalisiert und mit jeweils acht Power4 CPUs und zwei SP Switch2 Adaptern ausgestattet sind. Eine Ausnahme bilden hier die drei großen Quader der Klasse `csmp`. Hierbei handelt es sich um sogenannte *Full-System-Partitions*, bei denen das gesamte System in eine LPAR mit 32 Prozessoren integriert ist. Die Höhe der Quader beschreibt darüber hinaus die Hauptspeicherausstattung der einzelnen LPARs. Da neben dem Großteil dem Maschinen mit 64 GB auch zwei mit 128, beziehungsweise 256 GB Hauptspeicher beschafft wurden, reichen die Werte hier von 16 über 32 bis 64 GB.

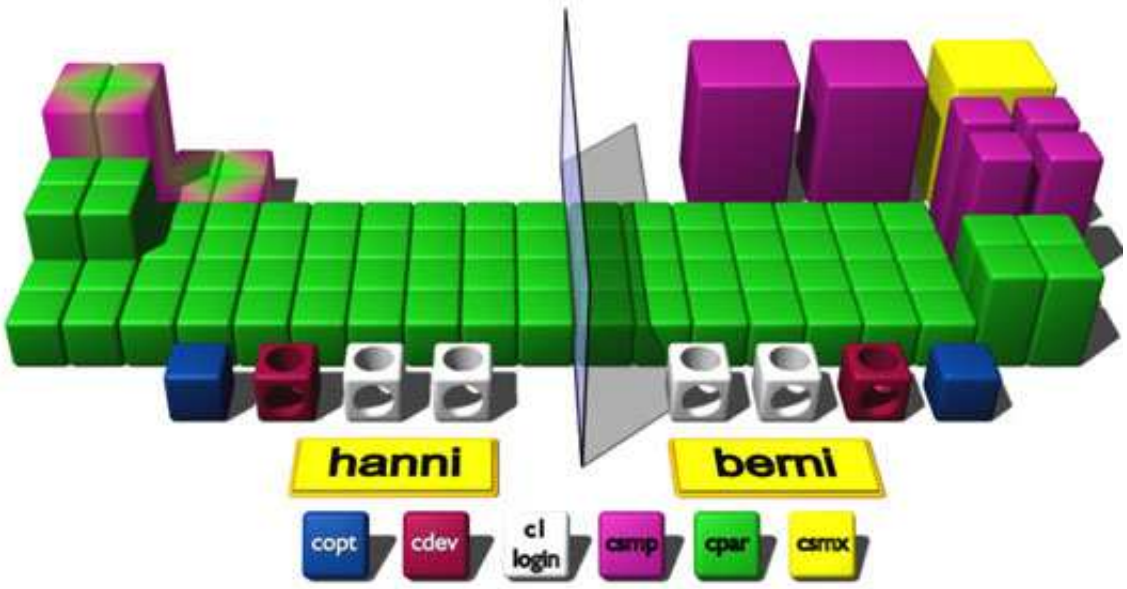


Abbildung 2: Zuordnung der LPARs zu Batch-Klassen

In allen LPARs wird eine eigene Instanz des AIX Betriebsystems ausgeführt, somit sind sie bezüglich des Betriebssystems vollkommen unabhängig voneinander. Die Anbindung nach außen wird pro Site über zwei dedizierte LPARs realisiert, die auch als Login-Knoten fungieren. Sie sind über Gigabit-Ethernet an den HLRN-Link angeschlossen und in das GWiN integriert (siehe Abb. 3).

Datenserver

Neben den Compute-Knoten sind für den Betrieb des HLRN noch weitere Systeme erforderlich, die vorrangig für die Verwaltung der benötigten und der bei Berechnungen anfallenden Daten vorgesehen sind. Hierbei handelt es sich pro Teilkomplex um drei IBM p660 Systeme, die mit je vier Power3 Prozessoren ausgestattet sind. Sie teilen sich auf in einen Datenserver (hdata, bdata) und je zwei *Hierarchical-Storage-Management* (HSM) Server (hhsm, bhsm).

2.3.2 Netzwerkkonfiguration

Sämtliche Rechner eines Teilkomplexes sind über das IBM-proprietäre Netzwerk „SP Switch2“ bzw. „Colony Network“ jeweils mit zwei Adaptern pro LPAR bzw. Server über zwei unabhängige Switch-Ebenen verbunden (siehe Abb. 3). Die Bandbreite zwischen je zwei Switch-Adaptern liegt etwas oberhalb von 300 MB/s, die Latenzzeit beträgt ca. 20 μs. Auf diesem Netzwerk werden sowohl die üblichen IP-Protokolle als auch das firmeneigene Protokoll „User Space“ angeboten.

Die Verbindung zwischen beiden Teilkomplexen erfolgt über den mit Gigabit-Ethernet-Technik betriebenen HLRN-Link, der über die im Rahmen dieses Projekts vom DFN-Verein beigetragenen WDM-Verbindung mit einer möglichen Bandbreite bis zu 2,4 Gb/s zwischen den beiden ca. 300 km entfernten Standorten ZIB in Berlin und RRZN in Hannover realisiert wird. Auf beiden Seiten sind jeweils die Login-Knoten, die Data-Server sowie die Magnetband-Server (HSM-Server) über Gigabit-Ethernet mit dem HLRN-Link verbunden. Eingesetzt wird das IP-Protokoll mit privaten Netzadressen, damit kann der HLRN-Link nur intern vom HLRN-System verwendet werden.

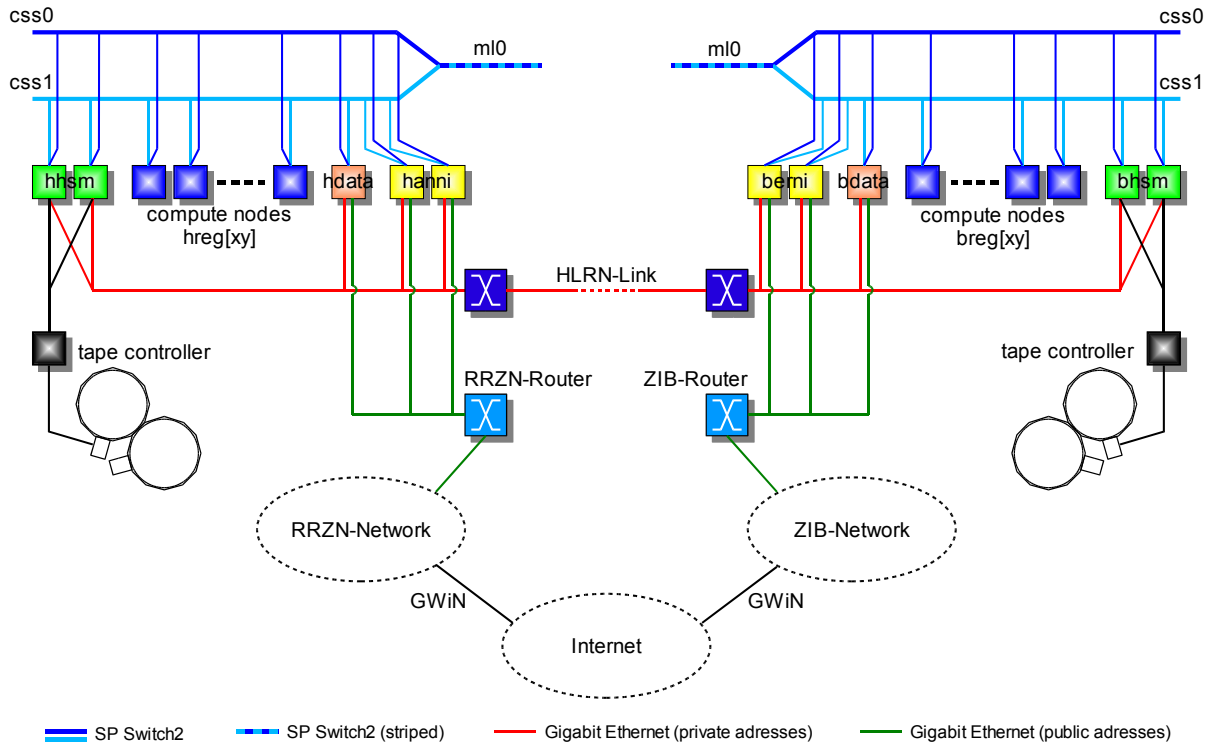


Abbildung 3: Netzwerkkonfiguration des HLRN-Systems

Wie in späteren Abschnitten näher ausgeführt, wird die theoretische maximale Bandbreite von 124,1

MB/s bei speziellen Transfers mit 121,5 MB/s nahezu erreicht. Die Latenzzeit dieser langen Strecke liegt bei 4,5 ms.

Die Verbindung in die beiden lokalen Netze von ZIB und RRZN und darüber dann in das Gigabit-Wissenschaftsnetz des DFN-Vereins erfolgt ebenfalls über die Login-Knoten und die Data-Server jeweils mit Gigabit-Ethernet-Anbindungen. Eingesetzt wird das IP-Protokoll mit Adressen aus dem Adressbereich des ZIB bzw. des RRZN. Die IP-Domain-Namen sind unabhängig von ZIB und RRZN gewählt, verwendet wird die HLRN-eigene Domain hlrn.de. Sicherheitsmechanismen sind auf mehreren Ebenen verwirklicht.

2.3.3 Dateisystemkonfiguration

An beiden Teilkomplexen werden prinzipiell zwei unterschiedliche Magnetplattensysteme betrieben:

- lokale Plattenlaufwerke an den Data-Servern: Diese sind jeweils 2 TB groß und enthalten ausschließlich die üblichen Heimatverzeichnisse, die im Falle des HLRN-Systems mit `/home/b` (Berlin) und `/home/h` (Hannover) bezeichnet werden. Über das NFS-Protokoll stehen diese Filesysteme an sämtlichen Rechnern beider Teilkomplexe zur Verfügung. Die I/O-Bandbreite für einzelne Dateien liegt innerhalb eines Teilkomplexes bei 35 MB/s.
- globale erreichbare Plattenlaufwerke an allen p690-Systemen jedes Teilkomplexes: Diese sind je Teilkomplex 26 TB groß und enthalten Arbeitsverzeichnisse wie `/fastfs/tmp` und `/fastfs/work` sowie den Plattencache für die Nutzung der Magnetbandlaufwerke `/perm`. Zum Einsatz für die globale Nutzung auf allen Systemen eines Teilkomplexes kommt das IBM-Produkt GPFS (Global Parallel File System). Beide GPFS-Filesysteme sind jeweils nur auf dem lokalen Teilkomplex verfügbar. GPFS unterstützt sehr stark die Technik des Platten-Stripings, daher werden I/O-Bandbreiten über 1 GB/s erreicht.

2.3.4 Konfiguration des Resource-Management-Systems

Im HLRN wird das von IBM entwickelte Ressource-Management-System LoadLeveler eingesetzt. Die 26 p690 Systeme wurden in unterschiedlich große logische Partitionen (LPAR, vgl. Abschnitt 2.3.1) unterteilt. Alle LPARs werden von einer LoadLeveler-Instanz, dem Scheduler, verwaltet.

Die für den LoadLeveler relevanten Eigenschaften einer LPAR werden mit Hilfe von sogenannten *Stanzas*, eine Menge von (Attribut,Wert)-Paaren, beschrieben. Jede LPAR ist einer *Klasse* zugeordnet. Jedem Job wird beim Abschicken eine Klasse zugewiesen.

Abbildung 2 stellt alle LPARs und ausgewählte Eigenschaften dar. Jeder Quader entspricht einer LPAR. Während die Höhe der Quader die Größe des Hauptspeichers einer LPAR (16 GB, 32 GB oder 64 GB) veranschaulicht, zeigt seine Grundfläche die Anzahl der Prozessoren (8 oder 32) in einer LPAR. Die Farbe zeigt die Klasse, zu welcher eine LPAR gehört. In Tabelle 1 werden die Klassen und ihre wichtigsten Merkmale, so wie sie während der Laufzeit des Projekts definiert wurden, aufgelistet (April 2004).

Klasse	Beschreibung	Wichtige Merkmale
cpar	parallele Jobs (Standardklasse)	`wall_clock_limit = 2 days`, `total_tasks = 128` Ressourcen stehen dediziert zur Verfügung
cxxl	parallele Jobs (mehr als 128 Tasks)	`wall_clock_limit = 2 days`, Beschränkung auf ausgewählte Nutzer, Ressourcen stehen dediziert zur Verfügung
cmeta	parallele Jobs (über beide Teilkomplexe)	`wall_clock_limit = 2 days`, Beschränkung auf ausgewählte Nutzer
csmp	Jobs, die Shared Memory benötigen	`64 GB Hauptspeicher pro LPAR`, `wall_clock_limit = '2 d (berni), 7 d (hanni)'`, `total_tasks = '32 (berni), 8 (hanni)'`, `node = 1` Ressourcen können gemeinsam genutzt werden
cdev	Programmentwicklung, z.B. Debugging, interaktive Jobs	`wall_clock_limit = 12 h`, `cpu_limit = 2 h`, `total_tasks = 16`, `node = 1`
copt	Benchmarking und Programmoptimierungen	`wall_clock_limit = 15 min`, `total_tasks = 8`, `node = 1`, `node_usage = not_shared`
c1	Jobs, die nur eine CPU benötigen, z.B. tar, zip, Datentransfer	`wall_clock_limit = 12 h`, `job_cpu_limit = 1 h`, `data_limit = 1 GB`, `total_tasks = 1`
	Subklassen für administrative Zwecke (vgl. Abschnitt 3.4.4)	

Tabelle 1: Beschreibung der LoadLeveler-Klassen

3 Teilprojekte

Im Folgenden werden die durchgeführten Arbeiten innnerhalb der einzelnen Teilprojekte beschrieben.

3.1 Begleitende Infrastrukturmaßnahmen

3.1.1 Ergebnisse aus früheren Projekten

Jumbo-Frames

Im Rahmen des DFN-Projektes Tele-Immersion wurden Untersuchungen zur Übertragungsleistung auf der Netzverbindung Berlin-Hannover getätigt. Mit den üblichen Transportprotokollen konnten zunächst nur Datenraten bis etwa 400 Mb/s erreicht werden. Die Ursache hierfür liegt nicht in einer zu geringen I/O-Leistung, sondern in einer zu geringen IP-Paket-Leistung aller heutigen Rechner. Da nach dem Gigabit-Ethernet-Standard die maximale Rahmengröße weiterhin – wie auch bei Standard-Ethernet und Fast-Ethernet – 1.518 Byte beträgt, müssen 10-mal so viele Rahmen pro Sekunde übertragen werden als bei Fast-Ethernet, um die erhöhte Datenrate ausnutzen zu können. Jeder ankommende Ethernet-Rahmen erzeugt einen *packet received interrupt*, der das Betriebssystem zur Verarbeitung veranlasst. Jede Unterbrechung beansprucht einige Prozessortakte. Würde eine Gigabit-Ethernet-Verbindung unter voller Last gefahren, so müsste der betroffene Rechner mehr als 80.000 Interrupts pro Sekunde[9] verar-

[9] Max. Durchsatz / Rahmengröße = max. Anzahl von Rahmen pro Sekunde; 1.000.000.000 bit/s / 1.500 byte/Rahmen / 8 bit/byte = 83.333 Rahmen pro Sekunde

beiten. Das ist erheblich mehr, als heute übliche Rechner verarbeiten können.

Ein Ethernet-Rahmen setzt sich aus einem Ethernet-Header, der Nutzlast und vier Fehlerkorrekturbytes zusammen. Viele Operationen des Protokoll-Stacks werden auf den Header angewandt. Die Analyse und der Aufbau des Paket-Headers kostet, unabhängig von dessen Größe, für jedes Paket gleich viel Zeit. Deshalb erzeugen kleinere Rahmen eine relativ höhere Belastung als wenige große. Im Detail ist die zur Auslastung von Gigabit-Ethernet notwendige Paketrate in Abhängigkeit von der IP-Paket-Größe bereits oben dargestellt.

Ein weiterer Punkt der geringen Effizienz der Nutzung der Gigabit-Ethernet-Leistung besteht darin, daß jedes abgehende oder ankommende Paket eine Speicheroperation verursacht. Memory-Pages sind meist in Blöcken von 4 kB, 8 kB oder 16 kB organisiert. Damit kann ein Standard-Ethernet-Rahmen mit seinen 1.500 Byte Nutzdaten die Speicheroperationen nicht effizient ausnutzen. So erzeugt z.B. ein Transfer von 8.000 Byte sechs Speicheroperationen, obwohl theoretisch bei 4 kB großen Blöcken nur zwei notwendig wären. Für die Nutzlast wäre ein Vielfaches dieser Seitengröße günstig.

Der beliebigen Vergrößerung der Nutzlast sind jedoch enge Grenzen gesetzt. Die von Ethernet eingesetzte Fehlerkorrektur FCS (Frame Check Sequence) nutzt das Fehlerkorrekturverfahren CRC32 (Cyclic Redundance Check). Dieses ist nur effizient bis zu einer Nutzlast von 11.450 Byte.

Der Hersteller von Ethernet-Netzgeräten Alteon hat eine Abweichung vom Gigabit-Ethernet-Standard propagiert, bei der eine Erhöhung der Rahmen-Größe auf 9.018 Bytes die CPU-Auslastung der Rechner verringert. Alteon hat seine Produkte mit diesem Feature, genannt Jumbo-Frames, ausgestattet und viele Hersteller sind diesem Beispiel inzwischen gefolgt. Ein positiver Nebeneffekt der Jumbo-Frames ist, daß der Protokoll-Overhead geringer ist und sich dadurch die Nutzdatenrate erhöht. Jumbo-Frames sind allerdings nicht in jedem Szenario sinnvoll und erfordern besondere Anpassungen des TCP-Stacks, um wirklich Verbesserungen zu bringen. So profitieren Anwendungen, die viele kleine Nachrichten austauschen, nicht von Jumbo-Frames. Nur für Bulk-Data-Transfers, bei denen eine große Datenmenge am Stück übertragen werden muß, sind Jumbo-Frames sinnvoll. Wie in Abschnitt 3.1.5 dargestellt wird, können mit Jumbo-Frames Transferraten von 121,5 MB/s erreicht werden. Bei vielen Systemen wird die Rahmengröße durch den Parameter MTU bei der Konfigurierung des Gigabit-Ethernet-Interfaces festgelegt.

TCP-Fenstergröße

Die TCP-Fenstergröße bestimmt die Anzahl der Bytes, die ein Sender abschicken darf, bevor eine Bestätigung vom Empfänger erwartet wird. Durch sie erfolgt die Flußsteuerung. Wird die Fenstergröße zu klein gewählt, so legt der Sender Pausen ein, wenn entsprechend viele Daten gesendet wurden, ohne eine Bestätigung erhalten zu haben. Die optimale Fenstergröße lässt sich aus Datenrate und Verzögerung ermitteln, wobei die Datenrate beim Gigabit-Ethernet-Netz mit 1 Gb/s fest vorgegeben ist. Die Verzögerung setzt sich aus der Laufzeit der Daten über das Netz und der Verarbeitungszeit bei Sender und Empfänger zusammen.

Die Standard-Fenstergröße ist systemseitig oft mit 16 kB festgelegt. Für Verbindungen wie z.B. der zwischen Hannover und Berlin mit einer Roundtrip-Zeit von ca. 4,5 ms unter Verwendung von GbE ist sie jedoch viel zu klein. Messungen zwischen Primärspeichern in Berlin und Hannover haben ergeben (vgl. Abschnitt 3.1.5), daß eine optimale Fenstergröße etwa 1.024 kB beträgt.

3.1.2 Inbetriebnahme der Netzverbindung

Bereits im April 2002 wurden die lokalen Gigabit-Ethernet-Netze der IBM-Teilkomplexe zwischen Berlin und Hannover über die Gigabit-Ethernet-Verbindung des vom 01.01.2001 bis 28.02.2003 durchgeführten DFN Projekts „Tele-Immersion in Weitverkehrsnetzen“[10] miteinander verbunden. Es wurden seit Bestehen der Verbindung Jumbo-Frames konfiguriert, um die Leistung des Gigabit-Ethernets möglichst gut auszunutzen.

Mit Ende des Tele-Immersionsprojektes Anfang 2003 wurden beide Enden der Gigabit-Weitverkehrsanbindung unmittelbar an die jeweilige Produktionsnetzinfrastruktur in Berlin und Hannover angeschlossen.

3.1.3 Überwachung des HLRN-Links

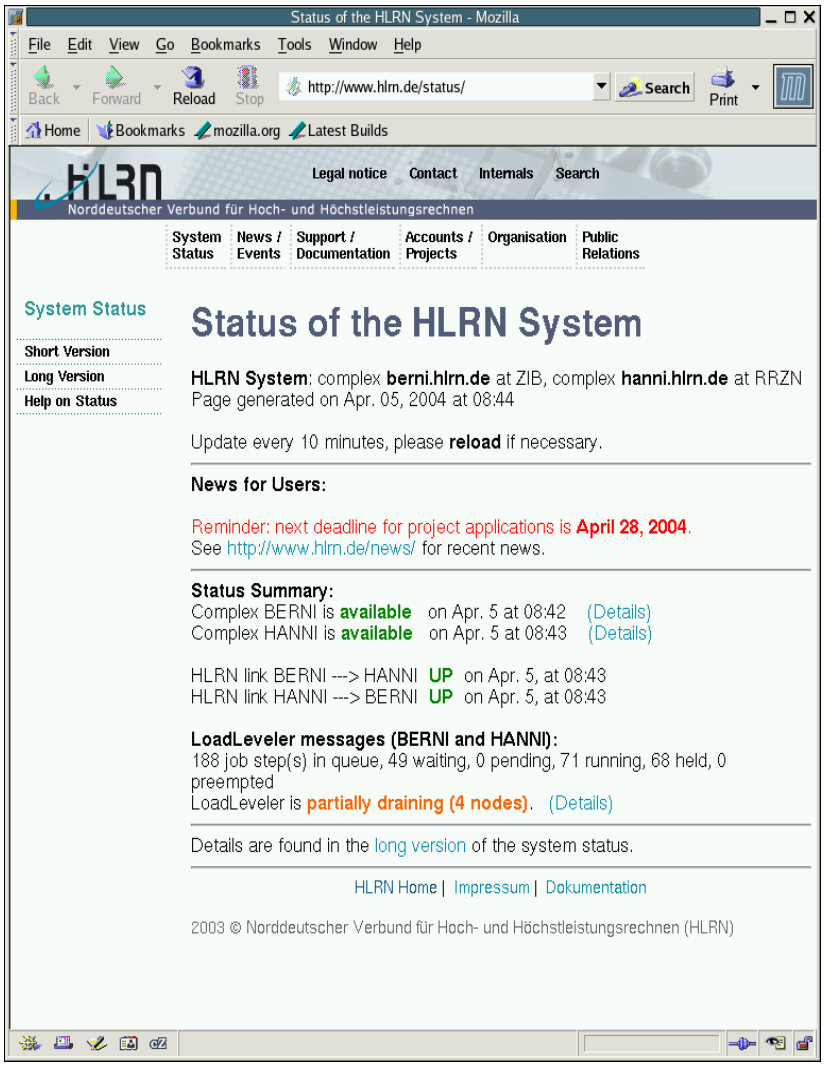


Abbildung 4: Screenshot der HLRN Status Seite (Überblick)

Ermittlung des Link-Status

Um den Status des HLRN-Links abfragen zu können, wurde das TCP-Testprogramm `ttcp` verwendet.

[10] http://www.rvs.uni-hannover.de/projekte/tele-immersion/

Passive Netzkomponenten wie Kabel lassen sich nicht verläßlich testen. Der Test läßt sich also nur zwischen Rechnern durchführen. Das verbreitete Unix-Tool `ping` ist jedoch in diesem Fall ungeeignet, da dessen ICMP-Protokoll nichts über andere Protokolle und Verbindungen wie TCP und UDP aussagt, welche jedoch für Benutzer relevant sind. Das Werkzeug `ttcp` ist eine Client/Server-Anwendung, die im Normalfall über einen vorgebenen TCP-Port eine Verbindung zwischen Client und Server herstellt. Für den reinen Verbindungstest ist das aber nicht erforderlich. Es genügt dabei, nur die Client-Anwendung zu starten und die Fehlerausgabe zu beobachten. Im dem Fall, daß die Leitung fehlerfrei arbeitet, wird `ttcp` melden, daß keine Server-Instanz auf der entsprechenden Maschine läuft. In allen anderen Fällen (inklusive einem Leitungsbruch etc.) wird `ttcp` eine andere Ausgabe bzw. ein Time-Out liefern. Diese Information wird in einer Datei (mit Zeitangabe) für weitere Auswertungen abgelegt.

Abbildung 5: Screenshot der HLRN Status Seite (Details)

Die vom Link-Status-Tool erfaßten Informationen werden aufbereitet und auf den Webseiten des HLRN[11] verfügbar gemacht (siehe Abb. 4 und Abb. 5).

Statistiken der HLRN-Link Auslastung

Mitarbeiter des RRZN haben ein Statistik-System (MRTG) konfiguriert, daß den GbE Switch auf der

[11] http://www.hlrn.de/status/

Seite des RRZN (siehe Abb. 3) überwacht und die Anzahl der übertragenen Bits in verschiedenen Intervallen (5 min, 30 min, 2 h, 1 d) aufzeichnet und die Ergebnisse graphisch aufbereitet. Diese Statistiken sind über die internen Seiten des HLRN verfügbar (siehe Abb. 6). Die Graphen in der Abbildung zeigen die Meßwerte vom Freitag, 28. August 2003.

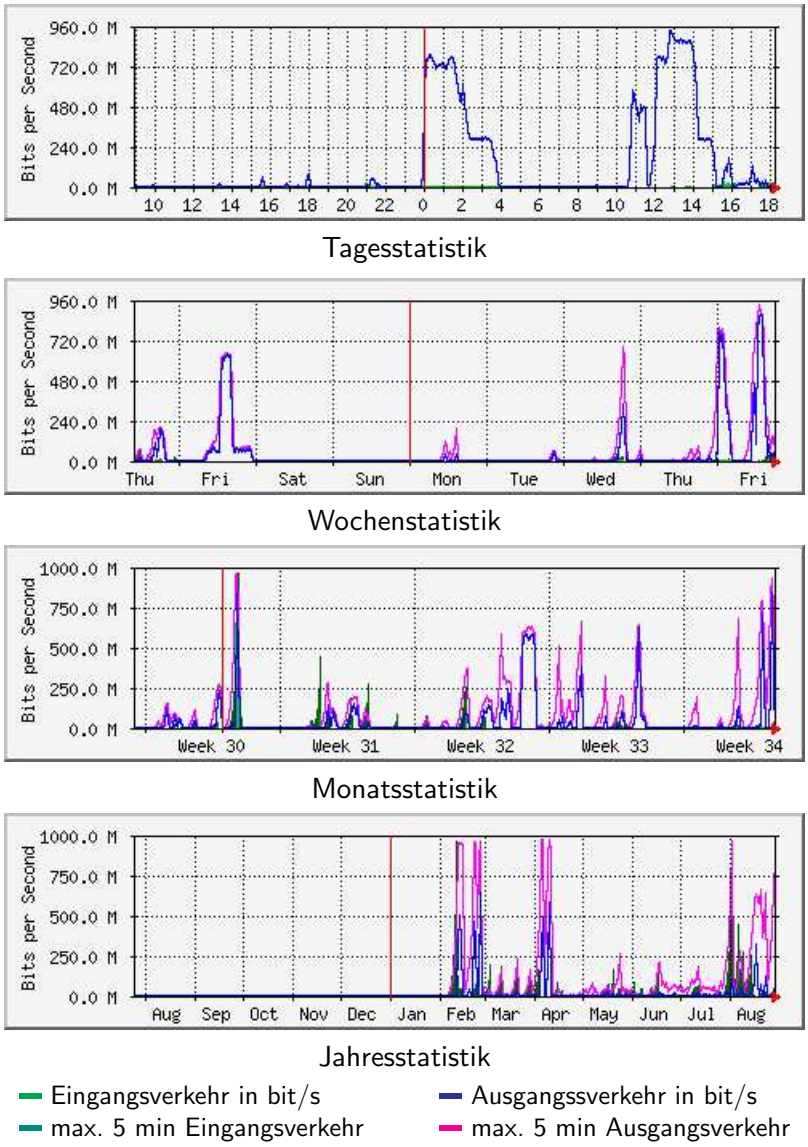


Abbildung 6: Nutzungsstatistik des HLRN-Links

Das für die Generierung der Statistiken verwendete Programm MRTG (Multi Router Traffic Grapher), welches auf einem mit dem zu überwachenden Switch verbundenen Rechner installiert wird, ermittelt in regelmäßigen Abständen mit sogenannten SNMP-Abfragen die momentane Auslastung des HLRN-Links. Diese Werte speichert es ab und generiert daraus Diagramme. Die Abtastrate von MRTG ist $\frac{1}{300}$Hz (zwölfmal pro Stunde).

SNMP (simple network management protocol) ist ein Protokoll zur Steuerung und Überwachung von höheren Netzkomponenten wie Switches und Routern. Die beim HLRN verwendeten CISCO Etherchannel Switches bieten von sich aus die Möglichkeit einer Protokollierung sämtlicher transportierter Pakete an.

Die Daten zur Auslastung des HLRN-Links werden erst seit Februar 2003 erfasst, deshalb wird in den

Monaten vorher (Jahresstatistik) der Transfer wie 0 dargestellt. Tatsächlich wurde in dieser Zeit der HLRN-Link sowohl für das Projekt Tele-Immersion als auch für die Verbindung des HLRN-Systems genutzt.

Die extrem hohen Spitzenwerte in den Monaten Februar, April und Anfang August (30. Woche) resultieren von den Messungen zwischen den Primärspeichern des HLRN-Systems her (vgl. Abschnitt 3.1.5); die um die 500 Mb/s liegenden längeren Nutzungszeiten in der 32. und 33. Woche resultieren aus tatsächlichen Dateitransferoperationen. Ab dem 1. September wurde der Teilkomplex in Hannover für ca. 5 Wochen außer Betrieb genommen. Um den betroffenen Benutzern in dieser Zeit die Nutzung des Teilkomplexes in Berlin zu ermöglichen, wurden auch die kompletten Arbeitsverzeichnisse in /fastfs/work (ca. 6 TB) nach Berlin kopiert. In der Abschlußphase (Donnerstag und Freitag in der Wochenstatistik) wurde in 10 parallelen Strömen Transferraten bis zu 880 Mb/s für jeweils komplette Dateitransfers erreicht.

Während der üblichen Produktionsphase (Mai, Juni, Juli) bleibt die Spitzenbelastung mit bis zu 250 Mb/s noch weit von den theoretischen möglichen 992,8 Mb/s entfernt, allerdings werden auch einige geplante Dienste, z.B. verteiltes Rechnen über beide Teilkomplexe, noch nicht im Benutzerbetrieb eingesetzt.

Diese Erkenntnisse lassen derzeit noch keine Notwendigkeit erkennen, den HLRN-Link in seiner Leistungsfähigkeit aufzurüsten, z.B. durch Nutzung eines zweiten Gigabit-Ethernet-Kanals auf der mit 2,4 Gb/s getakteten Strecke oder durch Wandlung in eine 10 Gb/s-Verbindung.

3.1.4 Ersatzschaltung bei Ausfall des HLRN-Links

Auf Grund der hohen Kosten des leistungsstarken HLRN-Links sieht die Vereinbarung mit dem DFN bzw. mit der T-Systems nur eine jährliche Verfügbarkeit der Leitung von 99,25% vor. Die Fa. T-Systems hat hier im Gegensatz zu sonstigen Leitungen keine „Backup-Verbindung“ vorgesehen, die im Fehlerfall kurzfristig geschaltet werden kann. Zur Überbrückung möglicher Ausfallzeiten des HLRN-Links sollte ein Ersatzweg über das DFN-Wissenschaftsnetz (G-WiN) eingerichtet werden.

Da die umzuleitenden Datenströme nicht im öffentlichen Datennetz geroutet werden können, muß der Datenverkehr durch das G-WiN getunnelt werden. Auf der Basis der vorhandenen aktiven Netzkomponenten standen dafür zunächst nur IP-Tunnel-Techniken zur Verfügung. Die konfigurierten Tunnel verwendeten die „Generic Routing Encapsulation (GRE)“. Für die beiden zu koppelnden VLANs[12] wurde je ein Tunnel eingerichtet. Durch „policy routing“ und Access-Listen wurde sichergestellt, daß nur die vorgesehenen Datenströme die Tunnelverbindungen nutzen konnten.

Da die Wegewahl nicht immer eindeutig war, war der Ersatzweg im Normalbetrieb abgeschaltet und mußte daher im Bedarfsfall aktiviert werden. Nach manuellen Anfängen wurden automatisierte Lösungen geschaffen, die zunächst auf einer Erreichbarkeits-Analyse und später auf SNMP-Traps basierten. Die SNMP-Traps werden von den Netzkomponenten (Switches) bei Link-Ausfall und -Wiederkehr generiert.

Da bei diesem Verfahren die ARP-Tabellen der angeschlossenen Geräte teilweise ungültig werden und daher aktualisiert werden müssen, arbeitet diese Layer-3-Umschaltung nie störungsfrei und eignet sich daher eher für längere Ausfälle.

Seit Anfang März 2004 besteht bei Verwendung der vorhandenen Netzkomponenten die Möglichkeit,

[12] Virtual Local Area Network - virtuelle lokale Netze

Tunnel in IP-Netzen auf Basis des „Layer 2 Tunnel Protocols Version 3 (L2TPv3)“ einzurichten. Damit können komplette Ethernet-Frames getunnelt werden. Die IETF-Spezifikation dieses Tunnelprotokolls befindet sich derzeit noch im Draft-Status.

Das Erkennen des Link-Ausfalls und das Umschalten übernehmen die Spanning-Tree-Instanzen der mit dem HLRN-Link verbundenen Switches. Statt des derzeit noch aktiven Spanning-Tree-Verfahrens nach IEEE 802.1d kann in nächster Zeit eine verbesserte Version nach IEEE 802.1w (Rapid Spanning Tree) eingesetzt werden. Damit verringern sich die Umschaltzeiten von weniger als einer Minute auf weniger als eine Sekunde.

Die L2TPv3-Tunnel unterstützen auch den Transport von Jumbo-Frames. Die vorhandene Hardware läßt allerdings auf dem Ersatzweg teilweise nur eine maximale MTU von 4.470 Byte zu. Damit im Normalbetrieb weiterhin der leistungsmäßig günstigere MTU-Wert von 9.000 Byte genutzt werden kann, ist vorgesehen, eine, wiederum SNMP-Trap basierte, automatische Umschaltung des Interface-MTU-Wertes der angeschlossenen Endgeräte zu implementieren.

3.1.5 Verfahren zum effizienten Übertragen von großen Datenmengen

Im folgenden Teil werden unterschiedliche Verfahren verglichen, um große Datenmengen möglichst effizient und sicher zwischen den beiden Komplexen des HLRN auszutauschen. Diese Verfahren werden später auch Nutzern des HLRN zur Verfügung gestellt, welche häufig mit Dateien von einigen Gigabyte Größe arbeiten.

Transferleistung zwischen Primärspeichern (Hauptspeicher)

Messungen zwischen Primärspeichern eignen sich i.A. dazu, um Charakterisktiken von Netzwerken und -medien festzustellen, da Primärspeicher über um Größenordnungen bessere Zugriffsleistungen (Latenzzeit, Bandbreite) als die jeweiligen zu testenden Netzwerke, insbesondere die hier zu testende Weitverkehrsverbindung, verfügen. Diese Werte stellen damit eine Obergrenze auch im Bezug auf die zu errreichenden Transferleistungen zwischen Sekundärspeichern dar. Die in Tabelle 2 aufgeführten Werte zeigen insbesondere die Abhängigkeit der Transferleistung von der verwendeten Fenstergröße. Für die Messungen wurde das Werkzeug Netperf[13] verwendet.

Fenstergröße	HLRN-Link	lokal
64 KB	14,7 MB/s	115,2 MB/s
128 KB	28,1 MB/s	121,5 MB/s
256 KB	58,1 MB/s	123,5 MB/s
512 KB	111,6 MB/s	123,5 MB/s
1024 KB	121,5 MB/s	123,5 MB/s

Tabelle 2: Gigabit Ethernet Transferleistung auf dem HLRN-Link und lokal (ermittelt mit Netperf)

Transferleistung zwischen Sekundärspeichern (Magnetplatten oder -bänder)

Im realen Anwendungsfall befinden sich die Daten der Nutzer nicht im Hauptspeicher sondern auf Festplatten. Um die für diesen Fall erwartete Leistung festzustellen, wurden folgende Tests unter Verwendung der drei Verfahren NFS, `scp` und `yscp`/`dmscp` durchgeführt. Es spielen dann nicht mehr nur die pure TCP-Transferrate sondern auch I/O-Leistung, etwaige Verschlüsselung und Fehlerkorrektur ein Rolle.

[13] http://www.netperf.org/

Die Verwendung von Programmen der R-Serie (`rsh`, `rcp`) wird beim HLRN-System im Hinblick auf fehlende Sicherheitsmechanismen (wie etwa die fehlende Verschlüsselung von Paßwörtern) nicht zugelassen.

Um die drei genannten Verfahren vergleichen zu können, wurden für die Tests Dateisysteme ausgesucht, auf die alle Verfahren Zugriff haben. Da NFS beim HLRN nur zum Export des Home-Verzeichnisses genutzt wird, wurden auch die Tools `scp` und `yscp`/`dmscp` für Tests mit Daten aus den entsprechenden Home-Verzeichnissen herangezogen. Die mit einem Stream ermittelte Leseleistung der Home-Verzeichnisse liegt dabei innerhalb eines Teilkomplexes bei 34,7 MB/s. Geschrieben wurden die zu übertragenden Daten immer auf das schnelle GPFS, wobei die gemessene Schreibleistung etwa 80,8 MB/s beträgt. Um eine mögliche Verbesserung bei höheren Leseraten zu untersuchen, wurde in einigen Fällen auch vom GPFS gelesen.

Eigenschaften der Übertragungsverfahren

Das ursprünglich von der Fa. SUN entwickelte NFS (Network File System) nutzt zur Übertragung das UDP-Protokoll, eine interne Fehlerkorrektur und keine Verschlüsselung der Daten. Da NFS im HLRN den Nutzern eines Teilkomplexes den schnellen Zugriff auf ihr Home-Verzeichnis ermöglichen soll, wurde es auf die Verwendung im Nahbereich und damit auf viele Threads mit kleinen Fenstergrößen (unter 50 kB) konfiguriert. Die Anpassung der Fenstergröße für transfers über größere Entfernungen ist nicht individuell möglich. Wie sich bei Tests zwischen Primärspeichern gezeigt hat, birgt jedoch gerade eine geringe Fenstergröße ein mögliches Manko gegenüber Techniken mit entsprechenden Einstellmöglichkeiten.

`scp` ist die entsprechende Kopierroutine der Firma SSH (Secure SHell) Communications Security und setzt auf das TCP-Protokoll auf. `scp` nutzt sowohl Verschlüsselung als auch Fehlerkorrektur für die Übertragung des gesamten Datenverkehrs. Eine Änderung der Fenstergröße ist auch bei `scp` nicht möglich.

Die ZIB-Entwicklung `yscp`/`dmscp`[14] verwendet eine gesicherte Sitzungskontrolle unter Nutzung von SSH und eine unverschlüsselte Datenübertragung ähnlich `rcp`. `dmscp` besitzt eine zusätzliche Fehlerkorrektur und nutzt genau wie `scp` das TCP-Protokoll zur Datenübertragung. `yscp` stellt ein Perl-Front-End zu `dmscp` dar, daß in etwa mit dem Funktionsumfang und der Benutzerfreundlichkeit von `scp` aufwartet. `dmscp` nutzt die Vorzüge der oben genannten Verfahren — variable Fenstergröße und Verschlüsselung nur bei Verbindungsaufbau — unter Vermeidung der Nachteile.

Ein viertes Verfahren ist das Erzeugen eines Backups von Daten auf den Magnetplatten eines Teilkomplexes zu Magnetbandlaufwerken im anderen Teilkomplex unter Verwendung des TSM (Tivoli Storage Manager). Dabei wurden drei parallele Streams für das Schreiben auf drei verschiedene Magnetbänder des Typs STK 9940B (maximale Schreibrate 30 MB/s je Magnetband) verwendet.

Durchgeführte Testreihen

Für die Messungen mit dem TSM wurden ca. 2.000 GB an Daten vom GPFS des Teilkomplexes Hannover auf Bänder des Berliner Teilkomplexes (Rechner bhsm) übertragen. Für alle anderen Messungen der Tabelle 3 wurde eine 1 GB große Datei verwendet. Die Datei enthielt den gepackten Inhalt eines Nutzerverzeichnisses, der nicht weiter komprimierbar ist. Es wurden jeweils zehn Meßwerte ermittelt und daraus die Mittelwerte errechnet. In vorbereitenden Tests hat sich erwiesen, daß eine Dateigröße von 1 GB aussagekräftige Ergebnisse liefert, da der Anstieg der Transferleistung darüber nur noch gering ist. Es wurden zehn Kopien der Datei unter anderem Namen verwendet, um mögliche Einwirkungen von schnellen Zwischenspeichern (Cache, Puffer) zu vermeiden. Da die Datenleitung zwischen Berlin und Hannover zum Zeitpunkt der Messungen gleichberechtigt auch anderen Nutzern zur Verfügung stand, gab es äußere Einflüsse auf die Meßergebnisse. Es wurde darauf geachtet, daß Meßreihen nur zu Zeiten,

[14] http://mss.zib.de/

in denen kein hoher Verkehr auf der Datenleitung beobachtbar war, gestartet wurden. Der Verkehr wurde mit den MRTG-Messungen (vgl. Abschnitt 3.1.3) beobachtet. Bei dem mit einem Stern (*) markierten Fall wurde eine 66 GB große Datei übertragen. Die bei `dmscp` eingestellte Fenstergröße war in allen Fällen 1.024 KB (vgl. Abschnitt 2.3.2).

Ergebnisse

Die Ergebnisse für alle vier verwendeten Verfahren sind in der Tabelle 3 zusammengefaßt.

NFS					
Quell-Dateisystem	**Ziel-Dateisystem**	**Quell-Rechner**	**Ziel-Rechner**	**Ergebnis in MB/s**	**# Streams**
`$HOME`	GPFS	bdata	hdata	6,7	1
`$HOME`	GPFS	bdata	hreg01a	6,7	1
scp					
Quell-Dateisystem	**Ziel-Dateisystem**	**Quell-Rechner**	**Ziel-Rechner**	**Ergebnis in MB/s**	**# Streams**
`$HOME`	GPFS	bdata	hdata	5,2	1
`$HOME`	GPFS	bdata	hreg01a	5,2	1
`$HOME`	GPFS	breg01a	hdata	5,8	1
`$HOME`	GPFS	breg01a	hreg01a	9,5	1
yscp(dmscp)					
Quell-Dateisystem	**Ziel-Dateisystem**	**Quell-Rechner**	**Ziel-Rechner**	**Ergebnis in MB/s**	**# Streams**
`$HOME`	GPFS	bdata	hdata	6,5	1
`$HOME`	GPFS	bdata	hreg01a	9,5	1
GPFS	GPFS	bdata	hreg01a	10,5	1
`$HOME`	GPFS	breg01a	hdata	9,5	1
`$HOME`	GPFS	breg01a	hreg01a	17,5	1
GPFS	GPFS	breg01a	hreg01a	19,2	1
GPFS*	GPFS	hreg01a	breg01a	21,1	1
GPFS	GPFS	breg01a	hreg01a	51,5	5
GPFS	GPFS	breg0[1|2]a	hreg0[1|2]a	84,2	10
TSM					
Quell-Dateisystem	**Ziel-Dateisystem**	**Quell-Rechner**	**Ziel-Rechner**	**Ergebnis in MB/s**	**# Streams**
GPFS	STK 9940B	hdata	bhsm	69,7 peak 34,8 sustained	3

Tabelle 3: Ergebnisse der Transferverfahren

Auswertung

Die Transferleistung unter Verwendung von NFS erreicht mit 6,7 MB/s (5,4 %)[15] auf Grund der Optimierung für lokale Zugriffe keine befriedigenden Werte. Sie ist unabhängig von der Leistung des Zielrechners. Die Transferleistung von `scp` ist auf Grund des durch die Verschlüsselung bzw. Entschlüsselung entstehenden Rechenaufwandes abhängig von der Rechenleistung der verwendeten Quell- bzw. Zielsysteme. Aus der Tabelle läßt sich ablesen, daß bei Verwendung der leistungsschwächeren Datenserver (bdata

[15] 100 % $\widehat{=}$ 124,1 MB/s, die maximale Transferleistung des HLRN-Links

bzw. hdata) die Transferleistung niedriger ist als bei NFS. Bei Verwendung der leistungsstarken Login-Knoten (breg01a und hreg01a) beträgt die erreichte Transferleistung 9,5 MB/s (7,7 %). Generell sind die ermittelten Leistungswerte bei `dmscp` höher als die von NFS und `scp`. Gegenüber NFS und `scp` verfügt `dmscp` über variable Fenstergrößen (in diesem Fall 1.024 kB), desweiteren führt es auch keine rechenaufwendige Verschlüsselung durch.

Parallele Transfers

Das Transferverfahren `dmscp` hat sich gegenüber NFS und `scp` als das leistungsfähigste Verfahren herausgestellt. Deshalb wurde untersucht, wie sich die Transferleistung bei Einsatz von mehreren Streams verhält. Es wurden fünf unabhänigige Datenströme gleichzeitig auf einer LPAR gestartet. Dabei wurde eine Gesamtleistung von 51,1 MB/s (41,2 %) erreicht. Anschließend wurde der Test auf zwei unterschiedlichen LPARs mit jeweils fünf unabhängigen Datenströmen gestartet. Die Gesamtleistung betrug hierbei 84,2 MB/s (67,8 %).

3.2 Kopplung der I/O-Systeme

Im Mittelpunkt des Teilprojekts 2 steht die effiziente und performante Nutzung der I/O-Systeme jeweils vom entfernten Standort des HLRN. Die ursprüngliche Planung sah hierfür vor, dies auf der Basis von Fibre-Channel-Kopplungen über den HLRN-Link zu realisieren.

Diese Planung konnte jedoch wegen einer geänderten Anbindung der lokalen Dateisysteme nicht realisiert werden: In jedem Teilkomplex werden RAID-Systeme mit einer Kapazität von je 26 TByte über Fibre-Channel angebunden, jeweils lokal wird zur Anbindung der jeweils 55 separaten Systeme (LPARs) an die hierüber installierten Dateisysteme das proprietäte IBM-Produkt GPFS (General Parallel File System) eingesetzt. GPFS hat ausgezeichnete Eigenschaften für die Bedienung dieser vielen unabhängigen Systeme, allerdings basiert das Konzept auf kurzen Reaktionszeiten, so daß sich Netzverbindungen über 300 km mit Latenzzeiten im Millisekunden-Bereich ausschließen.

Andererseits hat sich im Rahmen des Projekts herausgestellt, daß für die Zielsetzung der effizienten und performanten Nutzung der I/O-Systeme über den HLRN-Link eine direkte Kopplung der Fibre-Channel-Systeme mit den dazu notwendigen Investitionen nicht erforderlich ist. Im folgenden wird dargestellt, wie sich für die beiden Szenarien

- Entfernte Nutzung der Dateisysteme
- Entfernte Nutzung der Magnetbandsysteme

effiziente und performante Lösungen ohne direkte Fibre-Channel-Kopplung realisieren ließen.

3.2.1 Entfernte Nutzung der Dateisysteme

Dem Anwender stehen, wie im Kapitel 2.3.3 näher ausgeführt, für kleinere Dateien die Filesysteme `/home/b` und `/home/h` über das NFS-Protokoll auf beiden Teilkomplexen zur Verfügung. Für jeden Benutzer sind hier nicht mehr als 2 GB Dateiablage vorgesehen. Die erreichbare Transferleistung von 6,7 MB/s (vgl. Tabelle 3) ist hier ausreichend.

Anders ist die Situation für die Dateisysteme, in denen auch beliebig große Dateien abgelegt werden können: Die Dateisysteme /fastfs/tmp, /fastfs/work und /perm sind auf jedem Teilkomplex lokal auf leistungsfähigen Plattensystemen unter dem parallelen System GPSF eingerichtet. Benötigt hier ein Anwender entsprechende Daten auf „dem anderen Teilkomplex“, so muß er selbständig diese Dateien vor dem eigentlichen Rechenlauf zum anderen Teilkomplex kopieren, ggf. auch nach dem Rechenlauf erzeugte Ergebnisse zurück kopieren. Das vom ZIB entwickelte Transferverfahren `dmscp` erreicht, wie im

Kapitel 3.1.5 dargestellt, eine Transferleistung über die Verbindung von knapp 20 MB/s. Allerdings sollte eine für den Benutzer weniger aufwendige und daher mehr automatisierte Methode zur Erkennung der Notwendigkeit und gegebenenfalls Durchführung der Dateioperationen erstellt werden.

In den folgenden zwei Abschnitten wird eine mögliche Lösung für Pre- und Post-Staging Operationen im gemeinsamen Einsatz mit dem Load Leveler und dem GPFS-Filesystem dargestellt.

Pre-Staging

Der Anwender legt in einer Datei in `/home` fest, welche Dateien er für die Abarbeitung seines Jobs aus den großen Dateisysteme `/fastfs/tmp`, `/fastfs/work` oder `/perm` benötigt. Durch Aufruf eines Skripts `get_files` werden sämtliche benötigten Dateien lokal bereit gestellt, unabhängig von ihrem ursprünglichen Ablageort. Dateien, die auf dem anderen Teilkomplex gelagert werden, werden über geeignet parametrisierte `dmscp`-Kopieraufrufe über den HLRN-Link effektiv kopiert und im lokalen `/fastfs/tmp` abgelegt.

Post-Staging

Nach Beendigung des Rechenlaufes werden durch Aufruf eines Skripts `put_files` alle Dateien auf den ursprünglichen Teilkomplex zurück kopiert, bei denen der Anwender dies gewünscht hat.

3.2.2 Entfernte Nutzung der Magnetbandsysteme

Für Backup und Langzeitdatenhaltung wurde von der Fa. IBM das Produkt TSM (Tivoli Storage Manager) geliefert. Es umfaßt die Komponenten Backup, Archive und Migration (Hierarchical Storage Management / HSM). Als Hardwareplattform für die TSM-Server wird derzeit ein p655 Server der Fa. IBM mit 4 Power4 CPUs a 1.3 GHZ und 8 GB Speicher verwendet. Die Anbindung der Magnetbandstationen erfolgt über drei Fibre Channel Leitungen mit je 100 MB/s Transferleistung.

Ziel in dem auf zwei Standorte aufgeteilten Virtuellen Supercomputer sollte es sein, die Datenhaltung derart zu gestalten, daß an beiden Standorten alle Daten prinzipiell verfügbar gemacht werden können. Für die längerfristige Datenhaltung wurde die von der TSM-Software angebotene Möglichkeit der nutzergesteuerten Archivierung von Daten auf Magnetbänder ausgewählt. Diese Methode wird bereits auf Seiten der Rechenzentren sowie von einzelnen Nutzern angewendet.

Die technischen Voraussetzungen für einen Datenaustausch zwischen den Komplexen sind wie folgt gegeben: Die beiden Datenserver (`bdata` und `hdata`) dienen als Benutzerklienten für die auf den Archiv-Servern (`bhsm` und `hhsm`) laufende TSM-Serverdienste. Der Nutzer hat die Möglichkeit, auf einem Server (z.B. `bhsm`) seine Daten von einem client aus (z.B. `bdata`) zu archivieren. Der Zugriff auf die Daten im anderen Teilkomplex erfolgt durch Einräumen der Erlaubnis, die Daten auch auf dem anderen Datenserver (z.B. `hdata`) wieder aus dem Archiv zu holen. Somit hat der Nutzer es selbst in der Hand, seine Daten nach dem längerfristigen Archivieren am anderen Standort wieder aktiv auf Platte zurückzuladen. Dieses Konzept wird am HLRN umgesetzt, allerdings derzeit nur einseitig mit den Magnetbanddiensten im ZIB. Aufgrund technischer Probleme mit der Anbindung der Magnetbandgeräte an den TSM/HSM Server am Standort Hannover wird derzeit mit nur einen aktiven TSM/HSM Server in Berlin erfolgreich gearbeitet.

Im ZIB sind derzeit neun Magnetbandlaufwerke des Typs 9940 B der Firma StorageTek (STK) am Archivsystem angeschlossen. Ein Laufwerk hat eine nominelle Durchsatzleistung (unkomprimiert) von 30 MB/s. Beim Erstellen eines Backups bzw. eines Archivs vom Server `hdata` nach Berlin sind Spitzenwerte von 110 MB/s im laufenden Nutzerbetrieb gemessen worden, was etwa 85 % der über den HLRN-Link möglichen Leistung entspricht. Es gibt derzeit starke Hinweise dafür, daß ein Erreichen des

theoretischen Höchstwertes derzeit durch die Leistungsfähigkeit der Datenserver verhindert wird. Diese Datenserver sind derzeit noch mit 4 Power3 CPUs (375 MHz Taktfrequenz) der Firma IBM ausgerüstet. Mit dem Aufrüsten der Dataserver im 3. Quartal 2004 auf Power4 CPUs ist damit zu rechnen, daß die Durchsatzleistung noch weiter zu steigern ist.

3.2.3 Verlagerung von Terabyte von Daten über den HLRN-Link anlässlich der Umbauarbeiten im HLRN-Komplex Hannover

Die Methode des nutzergesteuerten Archivierens von Daten auf das Magnetbandarchiv der anderen Seite wurde intensiv während der Umbauphase in Hannover genutzt: Die lokalen Dateisysteme `/fastfs/work`, `/aws` und `/perm` in Hannover wurden über den HLRN-Link auf den TSM/HSM Server in Berlin kopiert. Das Datenvolumen lag dabei bei ca. 4 TB. Regelmäßig werden weiterhin die Heimatfilesysteme beider Teilsysteme über `bhsm` in das Berliner Archivsystem gesichert.

Um Benutzern mit Daten in Hannover das Weiterarbeiten in Berlin auch während des Umbaus des HLRN-Komplexes Hannover zu ermöglichen, mußten die benötigten Daten davor von Hannover nach Berlin und danach von Berlin nach Hannover übertragen werden. Es gab (und gibt) zwei unterschiedliche Teilmengen, die auch unterschiedlich gehandhabt wurden.

Die Heimat (`/home`)

Die Heimat liegt in einem JFS-Dateisystem und ist mit „GeoRM“ (IBM Software) nach Berlin gespiegelt, wenn auch die Synchronisation wegen Software-Problemen immmer nur manuell angestoßen wurde. Hier war also nur eine erneute Synchronisation nötig und danach die Aktivierung des Berliner Spiegels. Analog mußte nach dem Umbau verfahren werden.

Die anderen Verzeichnisse (`/aws`, `/homearchive`, `/perm`, `/permarchive`, `/work`)

Diese Verzeichnisse liegen alle in GPFS-Dateisystemen. Nach Rücksprache mit einigen Großnutzern konnten Daten, die während des Umbau-Zeitraums nicht benötigt wurden, auf Magnetbänder archiviert und vom Transfer ausgeschlossen werden. Es blieben knapp unter 1800 GByte Daten, die zu übertragen waren.

Für den Transfer wurde `dmscp` (ZIB Software) verwendet - mit dem UNIX-Tool `pax` zum Einsammeln der Daten (auf der Quell-Seite) bzw. Verteilen (auf der Ziel-Seite). Das UNIX-Tool `cpio` war nicht nutzbar, da es Dateien mit einer Größe über 2 GB nicht bearbeiten kann. Aber auch `pax` zeigte einige Probleme, so daß Nachbearbeitungen nötig waren.

Um (in der Summe) große Transfer-Raten zu erreichen, wurden bis zu zehn parallele Datenströme genutzt, verteilt auf die zwei interaktiven Knoten mit den entsprechenden Knoten als Partnern auf der anderen Seite. Die Daten wurden bei Hin- und Rück-Transfer jeweils von den lokalen Knoten zu den Remote-Knoten übertragen, also nicht von den Remote-Knoten geholt.

Das Transfer-Kommando:

```
( cd ${SRC_PATH} && /usr/bin/pax -x pax -w -t ${DIR} ) |\
  dmscp -checksum -wsz 1024k  -w -s ${REMOTE}-hl    \
        -u root -d ${DST_PATH} -R '/usr/bin/pax -rv -pe'
```

Um diese Parallelität zu ermöglichen, wurden die fünf Verzeichnisse (/aws, /homearchive, /perm, /permarchive, /work) auf der nächsten Ebene unterteilt, d. h. jeder Name (Datei oder Verzeichnis), der hier auftauchte, war eine Teilmenge: z.B. jedes Verzeichnis eines Benutzers. Absteigend nach Datenvolumen sortiert wurden diese Informationen für die Kontrollskripte bereitgehalten, die so beim Start und

nach Transfer einer Teilmenge die nächste unbearbeitete Teilmenge finden konnten. In einem Extremfall wurde auch auf Ebene der Unterverzeichnisse eines Benutzers aufgeteilt. Weiter wurde die Unterteilung aber nicht verfeinert, da der Transfer der Daten und nicht eine optimale Link-Auslastung das Ziel dieser Aktion war und der Aufwand nicht zu groß werden sollte. Es blieben also einzelne Teilmengen so groß, daß sie die Gesamttransferdauer bestimmten und damit die Mittelwerte der Transferraten.

Für einzelne Schritte im Verlauf dieser Transferaufgabe wurden Skripte entwickelt, der Gesamtablauf wurde aber manuell durchgeführt und kontrolliert. Im Rahmen dieser Kontrollen wurden auch in einigen Fällen nach Ende von Transferprozessen, die das zu bearbeitende Verzeichnis vollständig übertragen hatten, manuell neue Prozesse für andere Verzeichnisse gestartet, wenn noch Aufgaben warteten. So sind im Ablauf eines Transfers nach Sinken der Zahl der parallelen Prozesse auch wieder Anstiege zu finden.

Zur Sicherheit (`pax` hatte einige Software-Probleme) wurden auch noch die Quell- und Zielverzeichnisse verglichen (auf Vollständigkeit, Datei- und Verzeichnisattribute wie Länge, Zugriffsrechte, usw.). Die Daten selbst wurden beispielhaft nur im Verzeichnis `/aws` verglichen.

Die Ergebnisse

Der Transferablauf für die hier vorliegende Auswertung wurde aus Startzeiten und den `dmscp`-Meldungen (z.B.: Elapsed time 65,0 secs → 1.383,2 KB/sec) rekonstruiert. Für einen einzelnen Datenstrom wurde eine maximale Transferrate von 19,5 MB/s (156 Mbit/s) erreicht, in der Summe von zehn parallelen Datenströmen 115,4 MB/s (923 Mbit/s). Der letzte Wert entspricht einer Auslastung des HLRN-Links von 90 %.

Hannover → Berlin

Die Daten wurden in mehreren Schritten übertragen, da schon angefangen werden sollte, während die letzten Jobs noch rechneten.

```
2003-08-28 23:55 bis 03:45    830 GB in  3:50 Std.   max. 10 Datenströme
     Maximalwert: 104250.1 KB/sec = 814 Mbit/sec = 79 % Link-Auslastung
     Mittelwert:   63066.5 KB/sec = 493 Mbit/sec = 48 % Link-Auslastung

2003-08-29 11:45 bis 14:08    630 GB in  2:23 Std.   max. 10 Datenströme
     Maximalwert:  95953.4 KB/sec = 750 Mbit/sec = 73 % Link-Auslastung
     Mittelwert:   76993.3 KB/sec = 602 Mbit/sec = 59 % Link-Auslastung

2003-08-29 16:55 bis 17:45     16 GB in  0:50 Std.   max.  6 Datenströme
     Maximalwert:  19499.3 KB/sec = 152 Mbit/sec = 15 % Link-Auslastung
     Mittelwert:    5592.4 KB/sec =  44 Mbit/sec =  4 % Link-Auslastung

2003-08-30 10:41 bis 13:18    310 GB in  2:37 Std.   max.  4 Datenströme
     Maximalwert:  67170.0 KB/sec = 525 Mbit/sec = 51 % Link-Auslastung
     Mittelwert:   34507.3 KB/sec = 270 Mbit/sec = 26 % Link-Auslastung
```

Berlin → Hannover

In diesem Fall wurde vor dem Testbetrieb nur das Verzeichnis `/aws` übertragen, die restlichen Daten nach Ende des Test-Betriebs in einem Block. Hier kam es zu mehr als sieben Stunden Überhang für einen einzelnen Transfer und damit zu dem geringen Mittelwert. Aber die Dauer war angesichts der Tageszeit unwichtig.

```
2003-10-04 12:52 bis 14:20     16 GB in  1:28 Std.   max.  2 Datenströme
     Maximalwert:   8175.5 KB/sec =  64 Mbit/sec =  6 % Link-Auslastung
     Mittelwert:    3177.5 KB/sec =  25 Mbit/sec =  2 % Link-Auslastung

2003-10-10 17:33 bis 05:15   1775 GB in 11:42 Std.   max. 10 Datenströme,
     Maximalwert: 118137.8 KB/sec = 923 Mbit/sec = 90 % Link-Auslastung
     Mittelwert:   44188.6 KB/sec = 345 Mbit/sec = 34 % Link-Auslastung
```

3.3 Kopplung von parallelen Programmen

3.3.1 Das Parallel Operating Environment (POE)

Zur Ausführung paralleler Programme ist auf dem HLRN das *Parallel Operating Environment* (POE) des AIX Betriebssystems installiert, das es ermöglicht, zur Kommunikation zwischen den Programmkomponenten den *Message Passing Interface* (MPI) Standard zu verwenden. Neben der Möglichkeit, die Nachrichten über beliebige IP-fähige Netze zu übertragen, ist das POE in der Lage, die Leistung der IBM-proprietären SP Switch2 Technologie für parallele Programme verfügbar zu machen. Dabei ist es auch möglich, durch *Striping*[16] die Leistung beider SP Switch2 Adapter in einer LPAR zu kombinieren.

Bei der Nutzung von MPI über das SP Switch2 Netzwerk existieren wiederum zwei Möglichkeiten, auf die Hardware zuzugreifen. Im sogenannten *User Space* Modus wird eine möglichst geringe Antwortzeit (Latenzzeit) erreicht, da hier das Betriebssystem umgangen wird. Der User Space Modus ist jedoch nur in Knotenkonfigurationen nutzbar, in denen zwischen allen Kommunikationspartnern eine SP Switch2 Verbindung besteht. Dieser Modus ist somit nicht für parallele Programme geeignet, die auf beiden Teilkomplexen verteilt ausgeführt werden sollen. Demgegenüber wird beim IP-Modus eine IP-Verbindung über das SP Switch2 Netzwerk hergestellt, über die die MPI-Kommunikation ausgeführt wird. In diesem Modus wird nicht die geringe Latenzzeit des User Space Modus erreicht, da hier zusätzliche Arbeit des IP-Protokolls und die Kommunikation über das Betriebssystem die Leistung beeinträchtigen. Ein Vorteil des IP-Modus ist die Möglichkeit, gemischte Konfigurationen von Knoten mit SP Switch2 Hardware und von Knoten mit anderer IP-fähiger Hardware in einem MPI-Programm zu integrieren und so einerseits die Leistung des SP Switches zu nutzen und andererseits durch die Flexibilität des IP-Protokolls auch anders ausgestattete Knoten einzubeziehen. Aufgrund dieser Eigenschaft ist der IP-Modus des POE für die Ausführung von auf beiden Teilkomplexen verteilten parallelen Applikationen geeignet. Dies wurde durch die verteilte Ausführung des Linpack-Benchmarks bereits frühzeitig nachgewiesen. In Tabelle 4 sind die Meßergebnisse bei einer Matrix mit 28000 Zeilen dargestellt, die im Juni 2004 mit der aktuellen Systemsoftware bestimmt wurden. Je Knoten wurden 8 MPI Tasks gestartet.

	Anzahl Knoten			
	2	**4**	**8**	**16**
MPI via IP (Berlin-Hannover)	38,2 GFlop/s	62,0 GFlop/s	-	-
MPI via IP (ein Teilkomplex)	42,3 GFlop/s	78,2 GFlop/s	125,3 GFlop/s	185,6 GFlop/s
MPI via US (ein Teilkomplex)	42,3 GFlop/s	79,0 GFlop/s	124,7 GFlop/s	191,8 GFlop/s

Tabelle 4: Linpack-Meßergebnisse im Juni 2004

Wie im Abschnitt 2.3.2 bereits erläutert wurde, spielt bei TCP/IP Übertragungen über große Strecken die TCP-Fenstergröße eine entscheidende Rolle. Damit auch über große Entfernungen gute Leistungswerte erzielt werden können, muß die Fenstergröße entsprechend der Entfernung erhöht werden, damit die Kommunikationspartner nicht zu lange auf die einzelnen Nachrichten warten. Das POE bietet leider

[16] Datenpakete gehen abwechselnd über einen der Kanäle und werden am Ziel wieder in die richtige Reihenfolge gebracht.

keine Möglichkeit der Anpassung der Fenstergröße im IP Modus und erreicht daher in reinen Übertragungsbenchmarks über den HLRN-Link nur sehr geringe Werte (vgl. Abschnitt 3.3.3). Daher ist im Rahmen des Projekts eine weitere, frei verfügbare MPI Implementation auf ihre Eignung für den Einsatz auf dem HLRN untersucht worden. Hierbei handelt es sich um MPICH[17].

3.3.2 MPICH

In MPICH werden ebenfalls, wie im POE, verschiedene Übertragungsmedien und -protokolle unterstützt, zu denen auch TCP/IP gehört. Es ist jedoch nicht in der Lage, die Möglichkeiten des User-Space-Kommunikation auf dem HLRN zu nutzen. Daher können mit MPICH bei komplexinterner Kommunikation auch nicht so hohe Transferraten erzielt werden, wie mit dem POE (siehe folgender Abschnitt 3.3.3).

Die MPICH Implementation des MPI Standards bietet jedoch, im Gegensatz zum POE, die Möglichkeit, zur Startzeit eines parallelen Programms unter anderem die Puffergröße für MPI-Nachrichten anzupassen. Damit kann bei der Nutzung von MPICH auf die besondere Charakteristik des HLRN-Systems eingegangen werden. Das spiegelt sich auch in den erzielten Meßergebnissen wider.

Für die Tests der MPI-Performance wurde der etablierte PMB Benchmark[18] der Firma Pallas genutzt. Ausschlaggebend war der sogenannte PingPong-Test, der Bandbreite und Latenzzeit eines Kommunikationskanals bei verschiedenen Nachrichtengrößen zwischen zwei MPI-Prozessen bestimmt.

Für die Testläufe wurde eine MPICH-Installation mit den Standardeinstellungen übersetzt. Eine weitere Installation wurde mit einer voreingestellten Puffergröße vom einem MB, die den Erfordernissen der großen Entfernung zwischen den Komplexen entspricht, versehen. Sie ist in den Abbildungen mit *MPICH.bigsock* gekennzeichnet.

3.3.3 Meßergebnisse

Zur Ermittlung der für den Nutzer verfügbaren Kommunikationsleistung bei Benutzung von MPICH bzw. POE und zur Identifikation eventueller Schwachstellen oder Engpässe der Software und der Systemkonfiguration wurden drei verschiedene Testszenarien untersucht. In den ersten beiden Szenarien wird die Bandbreite zwischen den beiden Komplexen des HLRN in unterschiedlichen Konfigurationen untersucht. Diese Messungen erfolgten mit eingestellten Jumbo-Frames auf den Gigabit-Ethernet Verbindungen. Zum Vergleich ist jeweils eine Meßreihe, die ohne Jumbo-Frames bestimmt wurde, hinzugefügt worden. Sie ist mit *MPICH.smallMTU* bezeichnet. Sie wurde mit der gleichen MPICH-Installation erzeugt, die in den Graphen mit *MPICH.bigsock* ausgewiesen ist.

Das erste Szenario testet die MPI-Performance zwischen zwei Login-Knoten des HLRN (`breg01a` und `hreg01a`). In dieser Konstellation wird praktisch die Leistung des HLRN-Links vermessen, da zwischen diesen beiden Knoten nur noch netzwerkinterne Komponenten, wie Switches, liegen. Dieses Szenario bestimmt gleichzeitig die obere Schranke für die auf den Compute-Knoten verfügbare Kommunikationsleistung, da dort noch Routing- und Fragmentierungsprozesse involviert sind. Abbildung 7 zeigt die erzielten Meßwerte für die drei verglichenen MPI-Implementationen. Es zeigt sich deutlich, daß erst die Verwendung der größeren Puffer in *MPICH.bigsock* die Leistung des Links annähernd ausnutzen kann und einen Maximalwert von ca. 90 MB/s erzielt. Das POE stagniert bei ca. 12 MB/s und die unmodifizierte MPICH-Implementation erreicht gerade einmal 3 bis 4 MB/s. Diese Ergebnisse unterstreichen die Bedeutung der Möglichkeit, die Puffergröße anzupassen. Eine weitere wichtigere Erkenntnis liefern die Meßwerte des Plots *MPICH.smallMTU*. Mit Ethernet-Frames der Standardgröße, die hier verwendet werden, ist die erzielte Bandbreite wesentlich geringer als mit Jumbo-Frames. Es werden maximal 20

[17] http://www-unix.mcs.anl.gov/mpi/mpich/
[18] http://www.pallas.com/e/products/pmb/

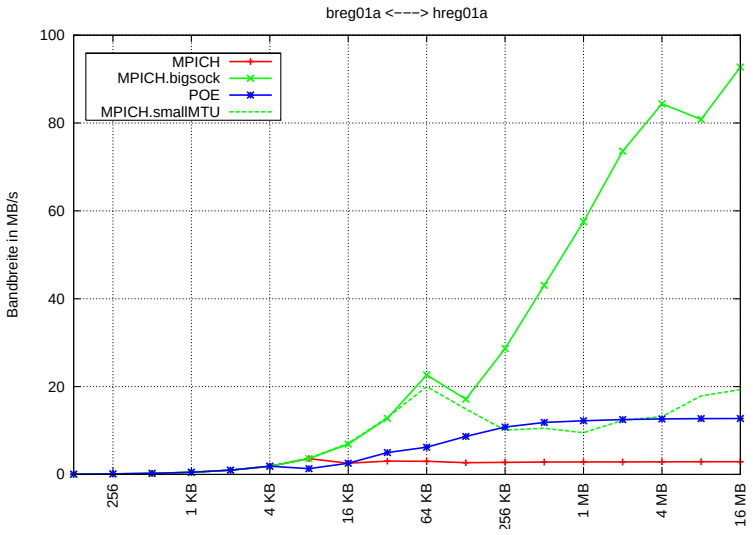


Abbildung 7: MPI-Performance zwischen den Login-Knoten über den HLRN-Link

MB/s erreicht.

Das zweite Szenario testet die tatsächliche Situation für ein verteilt auf beiden Komplexen rechnendes MPI-Programm. Es wird die erreichbare Bandbreite zwischen den Compute-Knoten `breg03a` und `hreg03a` überprüft. Eine Besonderheit dieses Szenarios ist jedoch, daß der Datenverkehr von den Compute-Knoten zunächst über das SP Switch2 Netzwerk zu den Login-Knoten übertragen werden muß. Von dort werden die Daten über Gigabit-Ethernet über den HLRN-Link geschickt, um vom Login-Knoten der anderen Site wieder auf das interne SP Switch2 Netzwerk umgesetzt und zum Compute-Knoten weitergeleitet zu werden. Bei dieser Umsetzung zwischen den verschiedenen Netzwerktechnologien kann es zu erheblichen Verlusten in der Übertragungsleistung kommen, da im SP-Switch2-Netzwerk eine größere Paketgröße (MTU) von 32 KB verwendet wird. Bei der Umsetzung auf Gigabit-Ethernet kommt es dann zu mehrfachen Übertragungsversuchen, bis der Sender die korrekte Paketgröße für Gigabit-Ethernet verwendet. Durch diese Verzögerung, die offenbar bei jedem übertragenen Datenpaket auftritt, sinkt im

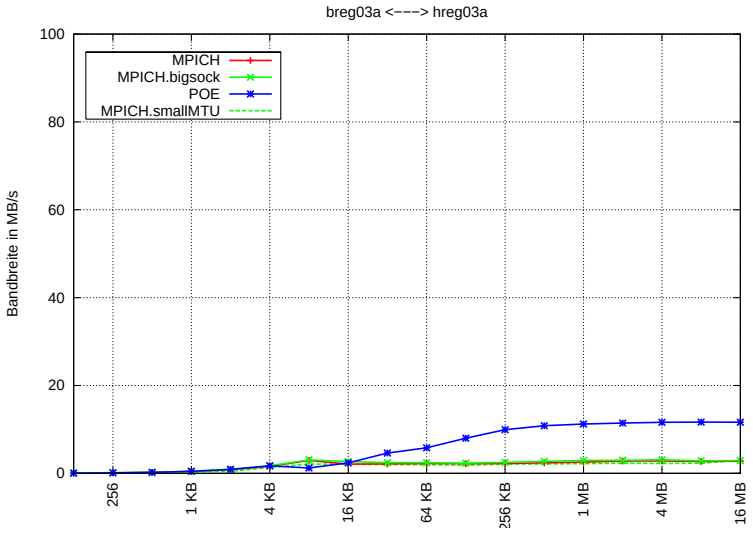


Abbildung 8: MPI-Performance zwischen Compute-Knoten über den HLRN-Link

zweiten Szenario die Leistung der modifizierten MPICH-Installation auf das niedrige Niveau der unmodifi-

zierten Version. Abbildung 8 zeigt, daß einzig das POE hier eine konstante, aber dennoch unbefriedigende Leistung von 12 MB/s erbringt. Die Messung mit Standard-Ethernet-Frames (*MPICH.smallMTU*) zeigt in diesem Szenario identische Meßwerte wie die beiden anderen MPICH-Messungen. Sie wird also ebenso durch die erzwungene Fragmentierung der Pakete behindert.

Die Tests im dritten Szenario wurden zum Vergleich der Meßergebnisse der beiden vorherigen Testumgebungen mit komplex-lokalen Werten durchgeführt. Hier wurde die Übertragungsleistung zwischen zwei Compute-Knoten über das SP Switch2 Netzwerk gemessen. Neben den beiden MPICH-Installationen werden hier POE im IP- und im User-Space-Modus verglichen. Die erzielten Ergebnisse sind in Abbildung 9 dargestellt. Mit ca. 300 MB/s als Maximalwert bestätigt das POE im User-Space-Modus die Erwartung, die MPI-Implementation mit der besten Performance auf dieser Hardware zu sein. Es zeigt sich jedoch auch, daß MPICH mit einer erhöhten Puffergröße (*MPICH.bigsock*) hier ebenfalls akzeptable Leistung für Applikationen bereitstellt. Diese Installation erreichte bei großen Nachrichten ($>$ 1MB) Bandbreiten von bis zu 228 MB/s. POE im IP-Modus bietet mit maximal 157 MB/s nur ca. die Hälfte der Leistung des User-Space-Modus und ist damit für datenintensive, komplexinterne Kommunikation weniger geeignet. Das gilt ebenso für die unmodifizierte MPICH-Installation, die in diesem Szenario maximal 51 MB/s erzielt.

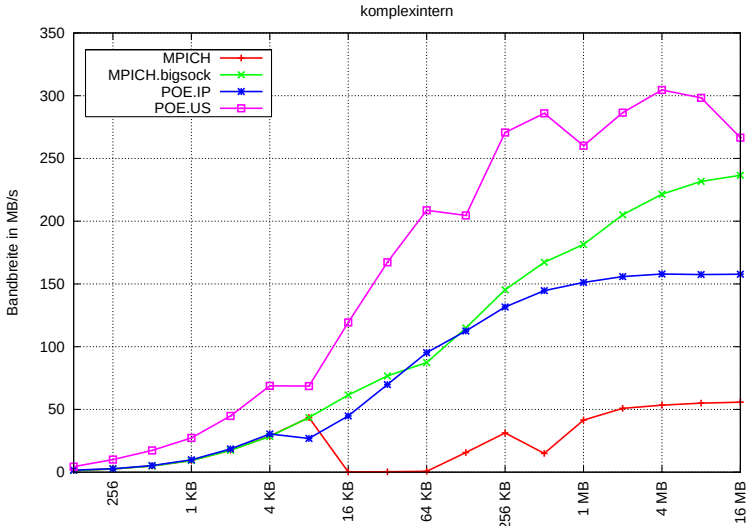


Abbildung 9: MPI-Performance zwischen Compute-Knoten über das SP-Switch2-Netzwerk

3.3.4 Auswertung

Es zeigen sich, insbesondere im ersten und dritten Szenario, daß die MPICH-Installation mit der angepaßten Einstellung für die Puffergröße (*MPICH.bigsock*) dem POE im IP-Modus überlegen ist. Die Meßwerte des Plots *MPICH.smallMTU* unterstreichen außerdem die Bedeutung der Jumbo-Frames für die Datenübertragung zwischen den Teilkomplexen des HLRN. Da diese Daten mit identischen Einstellungen zu *MPICH.bigsock* erzeugt wurden ist klar, daß Jumbo-Frames eine essentielle Voraussetzung für diese Kommunikationsstrecke sind.

Das zweite Szenario zeigt deutlich, das der Medienbruch, das heißt der Übergang vom SP Switch2 Netzwerk auf Gigabit-Ethernet, möglichst vermieden werden muß, um die verfügbare Bandbreite des HLRN-Links ausnutzen zu können. Aus diesem Grund werden im Rahmen der für Mitte 2004 vorgesehenen Umrüstung der beiden HLRN Komplexe auf die High-Performance-Switch-Technologie (HPS) in alle Compute-Knoten ebenfalls Gigabit-Ethernet-Adapter eingebaut. Diese stehen dann ausschließlich

für die komplexübergreifende Kommunikation zur Verfügung. Auf diese Weise wird der Medienbruch verhindert und den Applikationen die volle Bandbreite des HLRN-Links zur Verfügung gestellt.

3.4 Ein Ressourcen-Management-System für beide Teilkomplexe des HLRN

Aus der räumlichen Aufteilung des HLRN-Rechners auf zwei Standorte ergeben sich besondere Anforderungen an das Ressourcen-Management-System. Im folgenden werden für die einzelnen Aspekte dieses Teilprojektes jeweils die Ziele, die aus den Untersuchungen resultierenden Erkenntnisse und Lösungen sowie Optionen für die Zukunft erläutert.

3.4.1 Abschicken von Jobs

Um die Nutzung des HLRN-Rechners möglichst einfach zu gestalten, bestand das wichtigste Ziel des Teilprojekts darin, ein transparentes Verfahren zum Abschicken von Jobs zu entwickeln und dadurch eine Ein-System-Eigenschaft zu realisieren.

Konkret sollte erreicht werden, daß ein Nutzer auf beiden Teilkomplexen gleichberechtigt arbeiten kann. Insbesondere muß die Entscheidung des Ressourcen-Management-Systems, auf welchem Teilkomplex der Job laufen soll, unabhängig von dem Teilkomplex sein, den der Nutzer zum Abschicken eines Jobs ausgewählt hat.

Dazu wurden beide Teilkomplexe unter einer gemeinsamen LoadLeveler-Instanz zusammengeführt, wodurch aus den zwei physischen Teilkomplexen ein logischer Gesamtkomplex wird (vgl. Abschnitt 2.3.4). Das Abschicken eines Jobs wie auch seine Ausführung finden somit stets innerhalb desselben logischen Komplexes statt; das Job-Management ist insofern unabhängig von den beiden Teilkomplexen und der Rolle, die sie für einen Job spielen.

Im Betrieb hat sich die gemeinsame LoadLeveler-Instanz als hinreichend stabil erwiesen: Ausfälle der Verbindung zwischen den Teilkomplexen wurden durch den LoadLeveler so abgefangen, daß beide Teilkomplexe betriebsfähig blieben und, in der Regel, normal weiterliefen. Mit der Verwendung von nur einer LoadLeveler-Instanz wird daher ein robustes und effektives einheitliches Verfahren zum Abschicken der Jobs bereitgestellt. Dadurch konnte das Hauptziel erreicht werden.

3.4.2 Transparenter Zugriff auf Eingabedaten und Konfigurationsinformationen

Die räumliche Entkopplung des Abschickens und der Ausführung eines Jobs hat noch eine weitere Dimension: Die Gewährleistung der Verfügbarkeit der zur Job-Ausführung nötigen Daten. Da die auf dem HLRN-Rechner vorhandenen Dateisysteme im wesentlichen komplex-lokal konzipiert wurden, stehen die Daten nicht zwangsläufig auf beiden Teilkomplexen zur Verfügung. Somit war zu untersuchen, ob vom Job benötigte Dateien automatisch zu dem Teilkomplex, auf dem der Job ausgeführt wird, transferiert werden können.

Es zeigte sich dabei, daß die notwendigen Angaben zu Datenabhängigkeiten schwierig in das Ressource-Management-System zu integrieren sind und zudem die Datenabhängigkeiten teilweise erst nach dem Starten eines Jobs ermittelt werden. Dies und Überlegungen zur Effizienz waren ausschlaggebend, von einer Lösung zum automatischen Dateitransfer wieder Abstand zu nehmen. Stattdessen müssen notwendige Daten (Dateien) gegebenenfalls durch den Job selbst transferiert oder durch den Benutzer vor dem Starten des Jobs auf dem jeweiligen Teilkomplex bereitgestellt werden. Gestützt auf die Untersuchungen in Teilprojekt 1 (vgl. Abschnitt 3.1) wurde ein Werkzeug zum Transferieren von Dateien zwischen den verschiedenen Dateisystemen implementiert (vgl. Abschnitt 3.2.1).

Der transparente Zugriff auf job-bezogene Konfigurationsinformationen wird durch die Verwendung von nur einer LoadLeveler-Instanz sichergestellt (vgl. Abschnitt 3.4.1). Beispielsweise werden die Namen der Knoten, auf welchen ein Job ausgeführt wird, in der Variable `$LOADL_PROCESSOR_LIST` gespeichert.

3.4.3 Einschränkung eines Jobs auf einen Teilkomplex

Durch die Vereinigung beider Teilkomplexe unter einer LoadLeveler-Instanz (vgl. Abschnitt 3.4.1) kann das Ressourcen-Management-System die beiden Teilkomplexe nicht unterscheiden. Daher können Jobs, die über mehrere LPARs verteilt arbeiten sollen, LPARs beider Teilkomplexe zugewiesen bekommen. Dies ist aber in der Regel nicht optimal, da die Kommunikationsmedien innerhalb eines Teilkomplexes (SP Switch2, shared memory) wesentlich leistungsfähiger als der HLRN-Link sind.

Es ist deshalb wünschenswert, Jobs soweit wie möglich nur innerhalb eines Teilkomplexes zur Ausführung zu bringen. Um dies zu realisieren, wurde jeder LPAR ein Attribut zugewiesen, das beschreibt, zu welchem Teilkomplex die LPAR gehört. Enthält eine Jobbeschreibung die explizite Angabe eines dieser Attribute, wird der Job auf dem entsprechenden Teilkomplex ausgeführt.

Durch weitere derartige Attribute ist es zudem möglich, einen Job über beide Teilkomplexe verteilt auszuführen oder dem Ressourcen-Management-System die Auswahl des geeigneteren Teilkomplexes zu überlassen. Eine detaillierte Darstellung der Attribute und ihrer Anwendung sowie der sich daraus ergebenden Möglichkeiten und Probleme findet sich in Abschnitt 3.4.4.

Wie bereits in Abschnitt 3.4.1 dargestellt, hat sich das Ressourcen-Management-System als stabil gegenüber Ausfällen des HLRN-Links gezeigt. Dennoch stellt der Ausfall der Verbindung zwischen den Teilkomplexen eine wesentliche Störung der Jobbearbeitung dar, da der HLRN-Link insbesondere für die MPI-Kommunikations und die Bereitstellung von Dateisystemen (home) verwendet wird.

Da diese Störung nicht nur verteilte Jobs betrifft, wurde keine spezielle Lösung für das Job-Scheduling entwickelt, sondern ein globaler Ansatz verfolgt, der auf dem Umleiten der Datenpakete über das G-WiN basiert (vgl. Abschnitt 3.1.4). Der Vorteil dieses Ansatzes gegenüber dem Anhalten bzw. Beenden von betroffenen Jobs ist, daß diese mit verminderter Effizienz weiter ausgeführt werden können. Dieses Verfahren ist insbesondere bei kurzer Dauer eines Ausfalls sinnvoll. Eine entsprechende Lösung mit automatischer Umschaltung im Fehlerfall ist im Einsatz.

Fällt die dedizierte Netzwerkverbindung für längere Zeit aus, können Jobs manuell angehalten bzw. abgebrochen werden.

3.4.4 Unterscheidung der Teilkomplexe

In Abschnitt 3.4.3 wurde bereits angeführt, daß die gemeinsame LoadLeveler-Instanz zunächst nicht zwischen den beiden Teilkomplexen unterscheiden konnte und — weil diese Unterscheidung für die Einschränkung auf einen Teilkomplex notwendig ist — jeder LPAR ein Attribut zugewiesen wurde, das die Zugehörigkeit zu einem Teilkomplex identifiziert. Das Verfahren soll im folgenden genauer beschrieben werden.

Jeder LPAR können mit dem Attribut `feature` beliebige Eigenschaften (eine oder mehrere Zeichenketten) zugewiesen werden. Die Definition kann von LPAR zu LPAR abweichen und erlaubt somit, die Menge der LPARs in Teilmengen aufzuteilen. Bei der Zuordnung von Jobs zu LPARs überprüft das Ressourcen-Management-System, ob eine LPAR alle Features besitzt, die ein Job erfordert. Jobs werden somit auf die entsprechende Teilmenge eingeschränkt. Dem folgend, wurden alle LPARs mit einem Feature versehen: `berni` bei LPARs in Berlin, `hanni` in Hannover. Wird ein entsprechendes Feature in

der Job-Beschreibung angegeben, wird der Job nur auf dem entsprechenden Teilkomplex ausgeführt.

Das Attribut `feature` ist ein optionaler Parametern einer Job-Beschreibung. Daher werden Jobs, deren Beschreibung kein Feature enthält, nicht auf einen Teilkomplex eingeschränkt.

Es ist daher sinnvoll, ein fehlendes Feature im Nachhinein zu ergänzen. Hierfür stellt der LoadLeveler die Möglichkeit zur Verfügung, die Job-Beschreibung zum Zeitpunkt des Abschickens des Jobs durch ein Filterprogramm zu prüfen und gegebenenfalls zu ändern. Durch den vom HLRN entwickelten Filter wird Jobs, die keine Feature-Angabe in ihrer Beschreibung enthalten, automatisch ein Feature zugewiesen, und zwar entsprechend dem Teilkomplex, von dem der Job abgeschickt wurde.

3.4.5 Lastausgleich zwischen den Teilkomplexen

Das Filterverfahren hat somit zur Folge, daß alle Jobs nach der Prüfung durch das Filterprogramm auf einen Teilkomplex festgelegt sind. Dadurch kann eine Situation entstehen, in der die Teilkomplexe unterschiedlich ausgelastet sind. Wie bereits in Abschnitt 3.4 angedeutet, kann ein Job auch so konzipiert werden, daß er sowohl auf dem einen als auch auf dem anderen Teilkomplex lauffähig ist. Solche Jobs könnten auf dem weniger ausgelasteten Teilkomplex ausgeführt werden. Daraus entstünde zugleich ein Lastausgleich zwischen den Teilkomplexen.

Um dies zu ermöglichen, wurde das Feature `hlrn` definiert. Es ist auf allen LPARs gesetzt und bildet damit die Obermenge von `berni` und `hanni`. Da Jobs, die auf beiden Teilkomplexen lauffähig sein sollen, gewisse Anforderungen hinsichtlich der Verfügbarkeit ihrer Daten erfüllen müssen (vgl. Abschnitt 3.4), darf das Feature `hlrn` nicht vom Filter gesetzt werden, sondern muß explizit angegeben werden.

Für Jobs, die nur ein LPAR benötigen, ist mit diesem Feature eine optimale Möglichkeit des Lastausgleichs gegeben; Jobs, die mehr als ein LPAR anfordern, erwiesen sich hingegen als problematisch: Für diese ist nicht mehr ausgeschlossen, daß sie auf beide Teilkomplexe verteilt werden. Wie in Abschnitt 3.4.3 dargestellt, ist dies aber in der Regel nicht wünschenswert.

Es wurde daher ein Konzept entwickelt, das für Jobs, die über mehrere LPARs laufen, eine automatische Auswahl des Teilkomplexes durch das Ressourcen-Management-System gewährleistet. Hierbei wurde für die Job-Klassen, in denen Jobs mehrere LPARs anfordern dürfen, eine Subklassenstruktur eingerichtet; dadurch umfaßt jede Klasse eine Subklasse, die das Feature `hlrn` repräsentiert, und eine Subklasse, die die Features `hanni` und `berni` repräsentiert. Mit Hilfe des Filterprogramms werden die Jobs entsprechend ihrem Feature in die jeweilige Subklasse eingeordnet. Die Klasse, die das Feature `hlrn` repräsentiert, wird mit einer Methode, die der LoadLeveler bereitstellt, auf einem Teilkomplex geöffnet und auf dem anderen geschlossen, wodurch eine Aufteilung des Jobs auf beide Teilkomplexe verhindert wird.

Darüber hinaus wurde eine Automatik konzipiert, die in gewissen Zeitabständen die Lastsituation beider Teilkomplexe analysiert und gegebenenfalls die Verfügbarkeit der das Feature `hlrn` repräsentierenden Subklasse vom einen auf den anderen Teilkomplex umlegt. Dadurch kann ein Lastausgleich erreicht werden.

Ein allgemeinerer Ansatz zur Beschränkung eines Jobs auf einen Teilkomplex wurde mit der Fa. IBM diskutiert. Hierbei wäre in der Job-Beschreibung in geeigneter Art und Weise anzugeben, daß der Job nur innerhalb eines Teilkomplexes ausgeführt werden soll. Die Umsetzung und Gewährleistung übernähme dann das Ressourcen-Management-System. Nach diesem Ansatz fiele die Entscheidung, auf welchem Teilkomplex ein Job ausgeführt werden soll, direkt durch das Ressource-Management-System und nicht aufgrund externer Lastanalysen. Dies entspräche einer Ausweitung der Funktionalität, wie sie derzeit für Ein-LPAR-Jobs gegeben ist, auf Mehr-LPAR-Jobs. Gegenwärtig stellt das Batch-System diese Funktio-

nalität aber nicht zur Verfügung. Der Ansatz ist jedoch auch für IBM interessant und wird deshalb über das Ende dieses Projektes hinaus weiterverfolgt.

3.4.6 Explizite Verteilung eines Jobs über beide Teilkomplexe

Neben den bereits dargestellten Jobs, die jeweils auf einen Teilkomplex eingeschränkt bleiben sollen, bietet sich durch eine gemeinsame LoadLeveler-Instanz auch die Möglichkeit, Jobs über beide Teilkomplexe verteilt laufen zu lassen. Zwar wird in Abschnitt 3.4.3 nicht ohne Grund darauf verwiesen, daß eine über die Teilkomplexe verteilte Ausführung von Jobs nicht wünschenswert ist; dennoch gibt es Gründe, die eine Verteilung erfordern – beispielsweise wenn der Job für einen Teilkomplex zu groß ist.

Die Basisanforderung besteht hier darin zu gewährleisten, daß ein Job auf beiden Teilkomplexen gleichzeitig laufen kann. Dies wurde durch ein viertes Feature, `meta`, realisiert, das analog zu `hlrn` auf allen LPARs gesetzt ist. Im Gegensatz zu `hlrn` ist `meta` nicht mit einer entsprechenden Subklasse verbunden, so daß dadurch ein Job teilkomplexunabhängig die ersten freien LPARs zugewiesen bekommt.

Die „ersten freien LPARs“ können aus beiden Teilkomplexen stammen. Insbesondere ist die Anzahl von LPARs, die der jeweilige Teilkomplex dem Job zur Verfügung stellt, situationsabhängig völlig variabel und somit überhaupt nicht planbar. Aus dieser Situation ergab sich, über die im Projektantrag beschriebenen Ziele hinausgehend, eine weitere Anforderung: Ein Job soll explizit auf beide Teilkomplexe verteilt werden können, das heißt die Anzahl von Prozessen/Tasks auf jedem Teilkomplex soll explizit spezifizierbar sein.

Eine erste Umsetzung dieser Forderung wurde im Rahmen des Projektes entwickelt und steht den Nutzern – wenn auch mit deutlichen Leistungseinschränkungen – zur Verfügung. Es wurde dafür eine Klasse `cmeta` definiert, der Jobs, die für eine verteilte Ausführung konzipiert sind, zugewiesen werden können. Diese Klasse ist zunächst auf allen LPARs geschlossen, und manuell bzw. durch eine primitive Automatik wird die Klasse gemäß den in der Jobbeschreibung definierten Anforderungen auf geeigneten LPARs geöffnet. Dabei muß dafür gesorgt werden, daß alle anderen Jobs dieser Klasse in einem nicht lauffähigen Zustand (z.B. System-Hold) gehalten werden, um unerwünschtes Überholen zu vermeiden. Nachdem der Job angelaufen ist, wird die Klasse wieder geschlossen und gegebenenfalls der nächste Jobstart vorbereitet.

Da die Methode leistungsschwach und aufwendig ist, bleibt eine Integration der Funktionalität explizit spezifizierbarer Prozesse/Tasks in das Ressourcen-Management-System wünschenswert. Konsultationen mit dem Projektpartner IBM ergaben jedoch, daß die Umsetzung dieser Anforderung nur durch eine Erweiterung der Funktionalität des Ressourcen-Management-Systems implementiert werden kann. IBM hat aber Interesse signalisiert, eine entsprechende Lösung zu entwickeln, so daß dies über das Ende dieses Projektes hinaus weiterverfolgt wird.

4 Zusammenfassung und Ausblick

Seit der Bereitstellung des massiv parallelen Hochleistungsrechners IBM p690 für die norddeutschen Länder im Rahmen des HLRN-Verbunds im 2. Halbjahr 2002 steht den Wissenschaftlern ein leistungsfähiges System zur Verfügung. Durch die Verteilung der Teilsysteme auf die zwei Standorte in Berlin (ZIB) und Hannover (RRZN) waren jedoch auch neue Herausforderungen für den Betrieb und die effiziente Nutzung des Rechners entstanden. Im Rahmen dieses Projektes „Virtueller Supercomputer“ sind wesentliche Fragen eines an zwei Standorten betriebenen eng gekoppelten Hochleistungsrechner beschrieben und weitgehend gelöst worden, die erarbeiteten Lösungen sind komplett in den technischen Regelbetrieb überführt worden.

Die beiden Betreiberzentren ZIB und RRZN danken dem DFN-Verein für die Bereitstellung der dedizierten Datenverbindung zwischen den beiden Teilsystemen in Berlin und Hannover während der Laufzeit des Projekts von Februar 2003 bis Februar 2004. Die Ergebnisse des Projekts haben deutlich gezeigt, daß für den Betrieb dieses gekoppelten Systems eine solche dedizierte Datenverbindung mit garantierten Bandbreiten und insbesondere garantierten, nur durch die räumliche Entfernung gekennzeichneten, geringen Latenzzeiten unabdingbar ist. ZIB und RRZN nutzen diese Datenverbindung über das Projektende hinaus, nunmehr auf der Basis einer kostenpflichtigen Vereinbarung, zunächst bis zum Ende der Laufzeit des Gigabit-Wissenschaftsnetzes zum Jahresende 2005.

Die Bandbreite der Verbindung von 1 Gb/s hat sich zunächst weiterhin als ausreichend gezeigt. Die Betreiber des HLRN-Systems sorgen allerdings auch durch geeignete Parametrisierung der systemseitigen Netzparameter, durch geeignete Werkzeuge für die Benutzer und durch Empfehlungen an die Benutzer für eine effektive Nutzung dieser Datenverbindung. Es hat sich erfreulicherweise gezeigt, daß auch für einzelne Anwendungen sehr hohe Datenraten erreicht werden können. Die erreichten Spitzenwerte liegen bei 123,5 MB/s ($\widehat{=}$ 988 Mb/s, bzw. 99,5 % der theoretisch möglichen Transferrate von 124,1 MB/s) bei synthetischen Werkzeugen (netperf) und bei 92,7 MB/s ($\widehat{=}$ 741,6 Mb/s, 74,7 %) bei realen Anwendungen mittels MPI-Kommunikation — jeweils zwischen den Hauptspeichern übertragen — sowie 21,1 MB/s ($\widehat{=}$ 168,8 Mb/s, 17,0 %) beim Kopieren von einzelnen Dateien zwischen Plattenspeichern.

Die folgenden organisatorischen und technischen Festlegungen waren wesentlich für den Erfolg des Projekts:

- Gemeinsame Benutzerverwaltung: Jedes bewilligte Projekt und jeder zugelassene Benutzer erhalten auf dem Gesamtsystem gültige Kennungen. Jede Änderung einer Eigenschaft, sei es z. B. eine vom Benutzer selbst veranlaßte Änderungen seines Paßworts oder eine von der Benutzerverwaltung veranlaßte Änderung eines Limits, werden umgehend auf allen Rechnern des Gesamtsystems aktiviert.
- Eindeutiges Heimatverzeichnis: Jeder Benutzer kann den Ort seines Heimatverzeichnisses `/home/<benutzer>` nach seinen Erfordernisses selbst wählen. Der Zugriff zu Dateien im Heimatverzeichnis ist von beiden Seiten — außer bei Ausfall des zugehörigen Teilsystems — stets möglich.
- Gemeinsame Batch-Schnittstelle: Das Gesamtsystem verfügt über eine gemeinsame Batch-Schnittstelle, d. h. jeder Benutzer kann unabhängig vom Ort der Abgabe seines Jobs das ausführende Teilsystem wählen oder auch die Auswahl dem System selbst überlassen, so daß ein Lastausgleich erreicht werden kann.
- Jumbo-Frames und TCP-Fenstergröße: Wesentlich für die gute Ausnutzung der verfügbaren Bandbreite einer Datenverbindung mit den speziellen Eigenschaften des HLRN-Links (große Entfernung mit entsprechend hoher Latenzzeit) sind die Eigenschaften des erweiterten Gigabit-Ethernet-Protokolls, sowie die Möglichkeit der Anpassung der TCP-Fenstergröße (siehe Abschnitt 3.1.1). Während der Einsatz von Jumbo-Frames erst die Nutzung der vollen Bandbreite einer Gigabit-Ethernet Verbindung ermöglicht, ist die Veränderung der TCP-Fenstergröße zur Anpassung an die durch die große Entfernung bedingte hohe Latenzzeit erforderlich. Erst die Kombination beider Maßnahmen führte zur erfolgreichen Nutzung der vollen Bandbreite des HLRN-Links.

Während der Arbeit im Projekt hat sich herausgestellt, daß für den realen Betrieb folgende, über die ursprüngliche Planung hinausgehende, Arbeiten zu erledigen waren:

- Zeitnahe Status-Information an die Betreiber und die Benutzer über die Verfügbarkeit der Verbindung und der wichtigsten Dienste (siehe Abschnitt 3.1.3).

- Ersatzschaltung bei Ausfall der direkten Verbindung über einen Tunnel im Gigabit-Wissenschaftsnetz des DFN-Vereins mit geringerer Bandbreite und höherer Latenzzeit (siehe Abschnitt 3.1.4)

Auf Grund äußerer Einflüsse konnten einige Arbeiten nicht wie zunächst geplant ausgeführt werden. Teilweise wurden Alternativlösungen gefunden, teilweise werden diese Arbeitspunkte nach Ablauf des Projekts im Regelbetrieb weiterverfolgt:

- Die ursprüngliche Planung sah vor, für die Kopplung der I/O-Systeme Fibre-Channel-Verbindungen über den HLRN-Link zu realisieren. Hiermit sollte eine direkte Kopplung der lokalen SAN-Dateisysteme erreicht werden. Diese Planung wurde aufgegeben, da die Fa. IBM mit Installation des Systems nunmehr lokal jeweils kein SAN-Dateisystem installierte, sondern mittels ihrer proprietären Lösung „virtual shared disks“ (VSD) nur ein virtuelles SAN bestehend aus vier Teil-SANs. Für die vorgegebenen Anwendungen hingegen — Dateitransfer zwischen Magnetplatten und zwischen Magnetbändern — war dies nicht problematisch, da diese auch effektiv auf der Basis von IP-Protokollen mit Hilfe von Jumbo-Frames durchgeführt werden können.

- Die vorgesehenen Planungen für das gemeinsame Job-Scheduling konnten während der Projektlaufzeit nicht komplett umgesetzt werden, da das von Fa. IBM eingesetzte Produkt LoadLeveler noch nicht über alle notwendigen Features verfügte. Insbesondere ist es nicht möglich, die Tasks eines Jobs in einem bestimmten Verhältnis auf beide Teilkomplexe zu verteilen. Für die zweite Hälfte des Jahres 2004 sind Erweiterungen des LoadLevelers vorgesehen, so daß dann auch die ursprünglich vorgesehenen Leistungen im Benutzerbetrieb erbracht werden können.

- Die unbefriedigenden Leistungen bei MPI-Übertragungen zwischen Compute-Knoten, die während des Projektes festgestellt wurden (siehe Abschnitt 3.3.3), werden im Rahmen der Umrüstung auf den High-Performance-Switch (HPS) Mitte 2004 durch den Einbau eines Gigabit-Ethernet-Adapters in jeden Compute-Knoten behoben. So wird der Medienbruch beim Übergang vom SP Switch2 Netzwerk auf den HLRN-Link vermieden und den Knoten steht die volle Leistung des Links zur Verfügung.

Mit Abschluß des Projekts „Virtueller Supercomputer“ ist im HLRN nachgewiesen worden, daß ein gemeinsamer Betrieb eines verteilten Hochleistungsrechners von zwei unterschiedlichen Organisationen über 300 km hinweg erfolgreich durchgeführt werden kann. Notwendige Voraussetzungen sind dabei neben der dedizierten Datenverbindung mit ausreichender Bandbreite und insbesondere extrem kleiner Latenzzeit, eine gemeinsame Benutzerverwaltung sowie eine hohe Motivation der in beiden Einrichtungen mit dem Betrieb des Systems beauftragten Mitarbeiter.

Das HLRN-System wird von den Wissenschaftlern in Norddeutschland hervorragend angenommen. Das System ist voll ausgelastet. Die Nachfrage übersteigt deutlich die zur Verfügung stehenden Kapazitäten.

Das Konzept der Verteilung des HLRN-Systems auf zwei Standorte, verbunden durch den dedizierten HLRN-Link, hat sich bewährt und wird während der Bereitstellung dieses Systems (mindestens bis ins Jahr 2007) beibehalten. Auch das geplante Nachfolgesystem HLRN II soll wieder über diese beiden Standorte verteilt betrieben werden, der HLRN-Link soll dann mit erhöhter Bandbreite (10 Gb/s) bei gleich guter Latenzzeit betrieben werden.

Insgesamt haben auch die im Rahmen dieses Projekts geleisteten Arbeiten zur guten Akzeptanz des HLRN-Systems geführt.