Aufbau des großen GridKa-Clusters am Forschungszentrum Karlsruhe

M. Alef, B. Hoeft, H. Marten, J. van Wezel, IWR

Einleitung

Teilchenphysiker aus aller Welt arbeiten daran, ihre Meßergebnisse in einem extrem leistungsfähigen Daten-Grid auszuwerten. Die riesigen zu verarbeitenden Datenmengen stellen enorme Anforderungen an die beteiligten Rechenzentren. Das „Grid Computing Centre Karlsruhe“ (GridKa), das seit Herbst 2001 im Institut für Wissenschaftliches Rechnen (IWR) des Forschungszentrums Karlsruhe als deutsches LHC-Tier1-Regionalrechenzentrum aufgebaut wird [1, 2], hat diese Herausforderung angenommen (Abb. 1, 2). Dort wird von den 4 LHC-Experimenten Alice, Atlas, CMS und LHCb in weltweiten so genannten „Data Challenges“ die Funktionsfähigkeit, Stabilität und Skalierbarkeit der Software zur Detektorentwicklung, zur Simulation und zur Datenauswertung getestet. Gleichzeitig steht das GridKa auch den 4 weiteren schon heute aktiven Hochenergiephysik-Experimenten BaBar (SLAC, Stanford), CDF und D0 (Fermilab) und Compass (CERN) für deren Datenanalyse zur Verfügung. Mit den daraus gewonnenen Erfahrungen entwickeln GridKa-Mitarbeiter zusammen mit den Nutzern sowie den anderen Grid-Zentren die Grid-Infrastruktur weiter.

Abb. 1: Blick in das GridKa-Rechenzentrum.

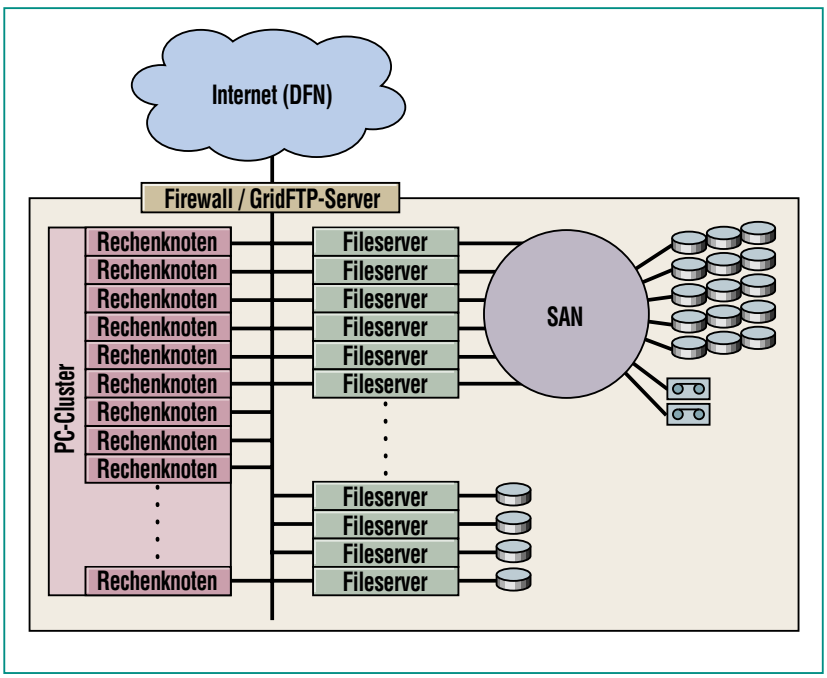

Abb. 2: Struktur des GridKa-Clusters (Rechenknoten, Fileserver, Netzwerkanbindung).

Die Ressourcen des GridKa werden, entprechend den Anforderungen der beteiligten Experimente, zweimal im Jahr erweitert. Zum „Meilenstein Oktober 2004“ umfaßt das GridKa rund 1060 CPUs, 220 TB (netto) Online-Platz auf Magnetplattensystemen sowie mehrere 100 TB im Magnetbandarchiv. Diese Kapazitäten werden bis zum Jahr 2007 auf rund das 4-fache anwachsen.

In mehreren Disziplinen – Netzwerkanbindung, Datenhaltung und Clusterbetrieb/-Kühlung – zählt GridKa inzwischen schon zur Riege der weltweit führenden Rechenzentren.

PC-Cluster

Für die Auswertung der Experimentdaten, sowie für Simulationsrechnungen, ist ein PC-Cluster installiert, der im Oktober 2004 etwa 1.000 Intel-Prozessoren (Pentium 3, 1,26 GHz, bis Xeon, 3,06 GHz) sowie rund 70 AMD-Opteron-CPUs enthält [3]. Jeder einzelne Rechenknoten dieses Clusters ist als Doppelprozessorsystem mit jeweils 1 bis 2 GB Hauptspeicher sowie einer lokalen Festplatte ausgelegt. Als Betriebssystem ist Linux installiert. Es ist bisher nicht geplant, parallele Anwendungen laufen zu lassen. Deshalb sind die meisten Rechner lediglich per Fast-Ethernet angebunden. Die neueren Knoten haben eine Gigabit-Ethernet-Anbindung.

Um eine solch große Anzahl an Rechnern effizient verwalten zu können, sind leistungsfähige, standardisierte und skalierbare Installations-, Administrations- und Überwachungsverfahren unbedingt notwendig. Für die Linux-Betriebssysteminstallation wurde das Rocks-Toolkit des San Diego Supercomputer-Zentrums [4] auf einem zentralen Installationsserver bereitgestellt und an die Erfordernisse des GridKa angepaßt. Bei der Anlieferung neuer Rechner kann ein kompletter Schrank mit 36 Rechnern innerhalb von 1 Stunde (!) installiert werden.

Das GridKa-Cluster wird im sogenannten Batchmodus betrieben. Dabei übergibt der Benutzer dem System vordefinierte Prozeduren und erhält einige Zeit später die Ergebnisse. Im Rocks-Toolkit ist das Batchsystem OpenPBS [5] als Standard enthalten; aus Stabilitäts- und Supportgründen wird im GridKa zur Zeit jedoch die kommerzielle Weiterentwicklung PBS-Pro verwendet [6] (Abb. 3a).

Schwachstellen und Engpässe im System können mit dem Auslastungsmonitor Ganglia erkannt werden [7], und zur „Gesundheitsüberwachung“ wird Nagios verwendet [8] (Abb. 3b).

Kühlung

Ein nicht zu unterschätzendes Problem beim Aufbau kompakter Clustersysteme ist die Kühlung. Der Stromverbrauch – und damit die Wärmeabgabe – der installierten Rechner ist in den vergangenen Jahren parallel zur Steigerung der Leistungsfähigkeit enorm angewachsen (Abb. 4). Ein zuverlässiger Betrieb des Clusters ist nur möglich, wenn die Rechner ausreichend gekühlt werden.

Schon zu Beginn des GridKa-Aufbauprojekts war abzusehen, daß die im Rechenzentrum vorhandene Klimaanlage nicht ausreicht. Eine Erweiterung wäre aufwendig geworden und hätte sehr hohe Luftströmungsgeschwindigkeiten notwendig gemacht, die die Aufenthaltsqualität im Rechnerraum und damit die Konzentrationsfähigkeit der Mitarbeiter zum Beispiel bei einer Störungssuche empfindlich eingeschränkt hätte.

Bei den früheren Großrechnern, auch bei denen des IWR, war eine Wasserkühlung üblich. Im PC-Bereich sind zwar wasserdurchströmte CPU-Kühlkörper auf dem Markt und werden inzwischen auch schon von einem Hersteller von Rechnerschränken angeboten. Für den Aufbau des GridKa-Clusters kam eine solche Technik aber nicht infrage: Einerseits dürfen die durch die Nähe von Wasser und Elektrizität denkbaren Sicherheitsprobleme nicht vernachlässigt werden [9]. Andererseits erzeugen nicht nur die Prozessoren Wärme, sondern auch die anderen Rechnerkomponenten (Hauptspeicher, Mainboard, Festplatte, Netzteil, ...). Schätzungen sagen, daß deshalb auf diese Art nur rund 60% der anfallenden Wärme per Wasser abgeleitet werden können [10].

Im GridKa wurden als praxisgerechte Alternative – in Zusammenarbeit mit dem Elektronikschrankhersteller Knürr – ge-

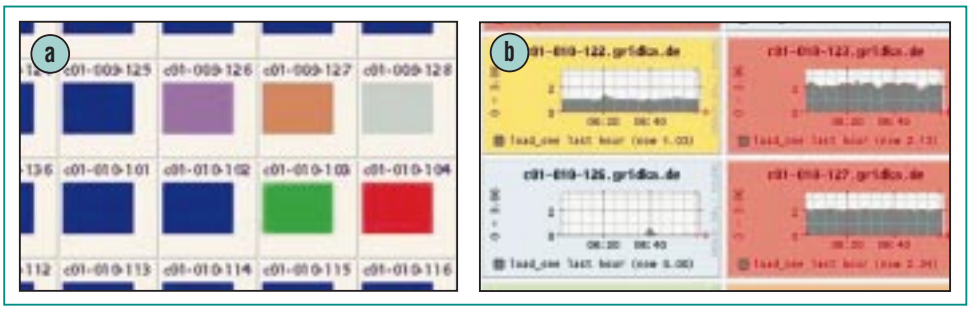

Abb. 3a+b: Grafische Benutzungsoberflächen helfen bei der Bedienung und Überwachung der GridKa-Rechner: Übersicht über die Auslastung des Clusters a) im Batchbetrieb (PBS-Monitor), b) detailiertere Ansicht (Ganglia).

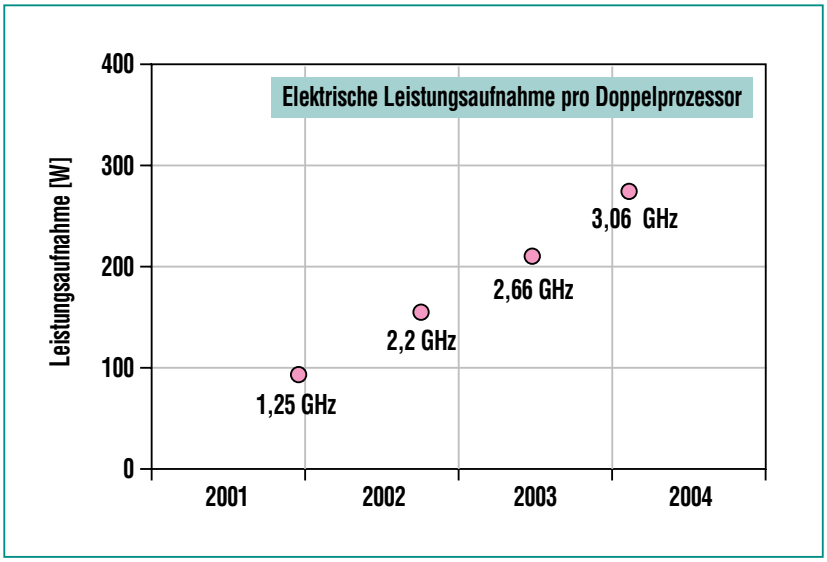

Abb. 4: In den letzten Jahren ist die elektrische Leistungsaufnahme – und damit auch die Wärmeabgabe – von PCs drastisch angestiegen. Die Grafik zeigt die Leistungsaufnahme der im GridKa eingesetzten Doppelprozessorsysteme in Abhängigkeit vom Zeitpunkt der Inbetriebnahme (bzw. dem Prozessortakt).

schlossene Rechnerschränke konzipiert, in denen ein zentraler Luft-Wasser-Wärmetauscher die von Lüftern innerhalb des Schranks umgewälzte Luft abkühlt (Abb. 5a+b). Im Oktober 2002 wurde ein Prototyp geliefert und in Betrieb genommen, ohne daß bisher grundsätzliche Probleme aufgetreten sind [3]. Mit den inzwischen knapp 30 weiteren solcher Schränke ist das GridKa des Forschungszentrums Karlsruhe weltweit das erste Rechenzentrum, das konsequent auf eine solche eigentlich simple, aber wirkungsvolle Technik zur vollständigen Ableitung der Rechnerabwärme setzt!

Lokale Datenhaltung

In einem Grid kann im Prinzip jeder Rechner auf die weltweit verteilten Daten zugreifen. Im Prinzip könnte ein Programm seine Eingabedaten bei ausreichend performanter Internet-(WAN-)Anbindung aus einem entfernten Rechenzentrum abholen und die Ergebnisse dort wieder ablegen. In der Praxis ist das jedoch nicht realisierbar: Die Antwortzeiten wachsen mit der Entfernung, bei einem Datentransfer zum Beispiel vom Fermilab in Chicago liegen diese bei mindestens etwa 50 Millisekunden. (Ein 3-GHz-Prozessor verbringt derweil 150 Millionen Taktzyklen mit Warten!) Die Zugriffszeiten auf eine lokale Festplatte sind um ein Vielfaches kleiner. Ein weiteres Argument ist, daß zur direkten Versorgung der vielen Clusterknoten eine WAN-Anbindung zu langsam und auch zu teuer wäre – das heutige GridKa-Cluster mit seinen rund 530 Knoten könnte theoretisch immerhin über 5 GBytes/s verarbeiten. Deshalb werden auch vor Ort weiterhin Datenspeicher benötigt.

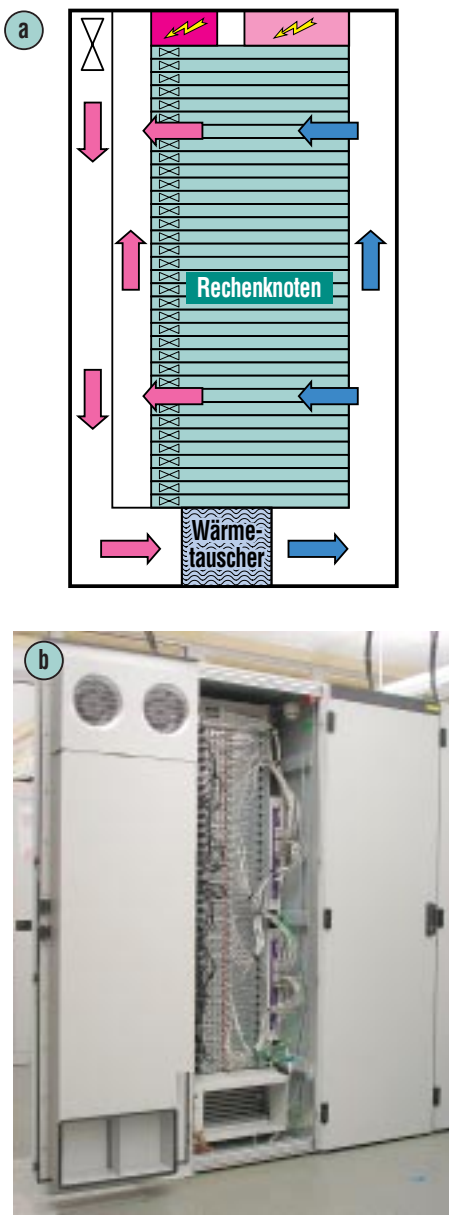

Abb. 5a+b: GridKa ist weltweit das erste Rechenzentrum, das die Abwärme der Rechner vollständig per Wasser ableitet. Die Luft wird von Ventilatoren, die oben in der rückseitigen Schranktür eingebaut sind, umgewälzt und durch einen Luft-Wasser-Wärmetauscher (unten) gedrückt. Die Rechner saugen vorne die gekühlte Luft an. Modifikationen an den Rechnern sind nicht notwendig.

Dennoch verschwimmen die Grenzen zwischen direkt oder per WAN angebundenem Speicher immer mehr. Die Gigabit-Strecke zwischen Hannover und Berlin wird zum Beispiel auch für den direkten Datenzugriff zwischen den dortigen Hochleistungsrechnern benutzt [11]. Es werden auch immer mehr Produkte entwickelt, mit denen Online-Daten verteilt und in sicherer Entfernung zu den Rechnern gehalten werden, um einen Datenverlust im Katastrophenfall zu minimieren. Und bei der jährlichen Supercomputer-Konferenz findet seit einigen Jahren der „High Performance Bandwidth Challenge“ statt, bei dem Firmen ihre Software für die direkte Speicheranbindung über WAN ins Rennen schicken [12]. Im folgenden sind aber mit „Online“ nur solche Daten gemeint, die lokal vor Ort gehalten werden.

Um allen Knoten des Clusters eine ausreichend hohe Bandbreite zu den Daten zu garantieren, wird für die Online-Datenhaltung im GridKa ein skalierbares paralleles Filesystem eingesetzt. Mehrere Rechner können damit gleichzeitig auf dieselbe Festplatte zugreifen. Zugleich können in einem derartigen System Daten schneller gelesen oder geschrieben werden, indem die Festplattenzugriffe parallelisiert werden. Dazu wird jede Datei über mehrere Platten verteilt („Striping“).

In Ethernet-Netzwerk sind mehrere gleichzeitige Zugriffe auf Dateien schon seit vielen Jahren möglich; NFS, SMB und AFS sind einige etablierte Protokolle zur gemeinsamen Nutzung von Dateisystemen in einem Netzwerk. Ein paralleles Filesystem bietet im Prinzip die gleiche Möglichkeit, hier jedoch sind die Speicher z. B. mittels SCSI-, ATA- oder Fibre-Channel-Verbindungen direkt angebunden. Ein im Filesystem enthaltener, aufwendiger Locking-Mechanismus verhindert, daß mehrere angeschlossene Rechner gleichzeitig versuchen, dieselbe Datei zu verändern.

Für die Online-Datenhaltung wurde im GridKa das „General Parallel File System“ (GPFS) von IBM gewählt [13]. GPFS wird in Computer-Clustern unter AIX schon lange verwendet und ist seit rund 2 Jahren auch unter Linux verfügbar. Das parallele Filesystem läuft im GridKa auf einem eigenen, kleineren Fileserver-Cluster, das die Daten per NFS für die übrigen Rechner bereitgestellt. Die GPFS-Fileserver sind weitgehend redundant, der Ausfall maximal knapp der Hälfte dieser Server oder deren Plattenanbindungen würde lediglich zu einer verringerten Datenübertragungsrate, aber nicht zu einem Totalausfall von Daten führen. Ein weiterer Vorteil ist, daß viele Wartungsarbeiten wie der Austausch von Platten oder Fileservern, oder das Vergrößern oder Verkleinern von Filesystemen, im laufenden Betrieb erfolgen können.

Datenaustausch und Sicherheit im Grid

Die Grid-Zugangsrechner, die von außen direkt erreichbar sein müssen, sind durch eine Firewall vor potentiellen Angriffen aus dem Internet geschützt. Jedoch werden Firewall-Konzepte gerade im Grid-Zusammenhang kontrovers diskutiert, weil die erreichbare Bandbreite traditioneller Firewalls begrenzt ist und weil die Konfiguration der offenen Ports in einer weltweiten, dynamischen Kollaboration mit einem hohen Administrationsaufwand und vielen Fehlermöglichkeiten verbunden ist.

Eine sichere und skalierbare Lösung dazu können dedizierte Server sein, die direkt am Internet-Backbone angeschlossen sind und auf denen das „Grid File Transfer Protocol“ (GridFTP) läuft [14].

Um die Performance von GridFTP zwischen den beiden Grid-Partnern GridKa und CERN zu messen, wurde der WAN-Anschluß des Forschungszentrums Karlsruhe von ursprünglich 34 Mbit/s Ende 2001 auf derzeit 1 Gbit/s aufgerüstet (Abb. 6). Messungen haben ergeben, daß nicht das GridFTP-Protokoll selbst, sondern die darunter liegende TCP/IP-Schicht ein begrenzender Faktor ist. Während mit Standard-Implementierungen von TCP/IP über eine einzelne Gbit-Leitung eine mittlere Transferleistung von nur 470 Mbit/s erreicht wurde, konnten mit weiterentwickelten TCP/IP-Implementierungen bis zu 980 Mbit/s übertragen werden[15]. Allerdings können derartige Geschwindigkeiten nur beim Senden von Daten erzielt werden, die vom Hauptspeicher des Senders direkt in den Hauptspeicher des Empfängers geschrieben werden. Sobald der Transfer von/auf Festplatten oder gar Magnetbändern verläuft, ist deren Mechanik der begrenzende Faktor. Eine der Herausforderungen für die Zukunft wird darin bestehen, die Datenübertragungen

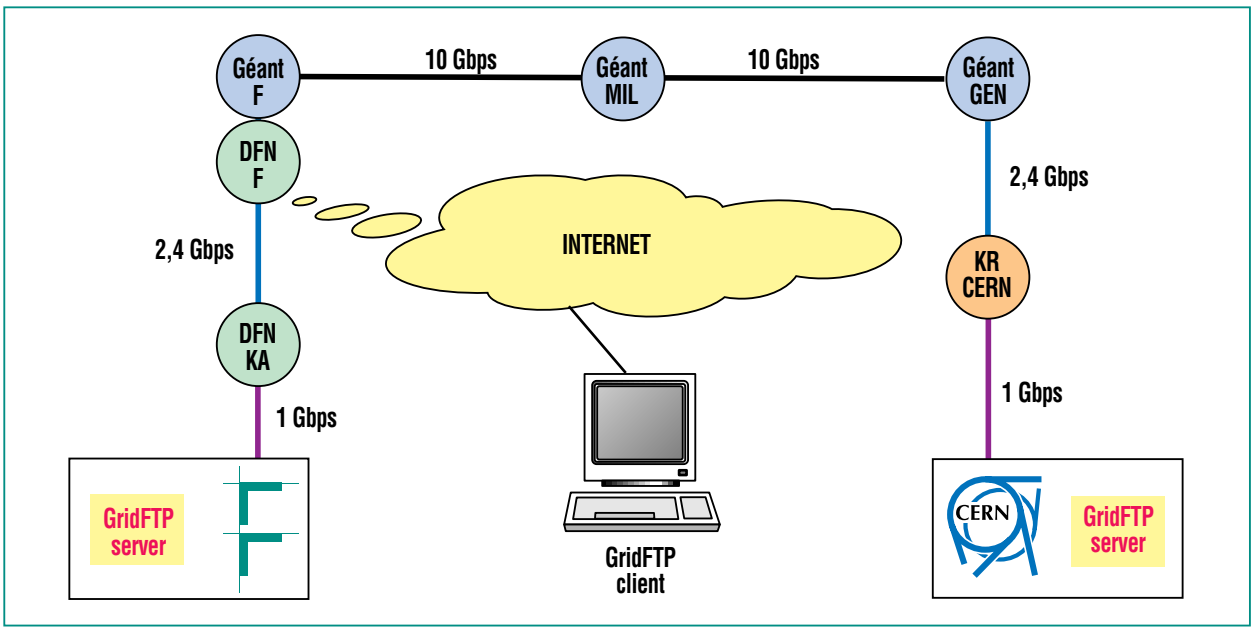

Abb. 6: GridFTP zwischen dem Forschungszentrum Karlsruhe (GridKa) und CERN. Externe Nutzer haben über das Internet weltweiten Zugriff auf ihre Daten.

bis auf die Speichermedien hinab zu parallelisieren. Die GridFTP-Spezifikation berücksichtigt bereits nicht nur die Parallelisierung eines Datenstromes über mehrere Prozesse einer Server-Client-Kommunikation, sondern auch die Aufspaltung eines Datentransfers über mehrere parallel geschaltete Server. Letzteres ist allerdings zur Zeit noch nicht implementiert.

Während die beschriebene Gigabit-Internet-Anbindung im GridKa schon genutzt wird, werden ab Herbst 2004 Tests mit der 10-Gigabit-Technologie beginnen.

Ausblick

Das Grid wird über enorme Rechner- und Speicherkapazitäten verfügen, die über die weltweiten Rechenzentren verteilt werden. Im Grid Computing Centre Karlsruhe (GridKa) entsteht ein PC-Cluster, das zur Zeit mehr als 1.000 CPUs, über 200 TB Plattenplatz und mehrere 100 TB Magnatbandkapazitäten enthält. Diese Installation entspricht einem im Jahre 2001 von 8 deutschen Hocherergiephysik-Gruppen aufgestellten Plan, der vom GridKa in jeweils 2 Erweiterungsstufen (Meilensteinen) pro Jahr bisher pünktlich erfüllt werden konnte. Der weitere Ausbau sieht bis zum Jahr 2007 über 4.000 Prozessoren und mehrere Petabyte Speicherplatz auf Magnetplatten und -bändern vor.

Literatur

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

[2] H. Marten, *Beitrag in dieser Ausgabe der Nachrichten*

[3] http://www.gridka.de/hardware/hardware.html

[4] http://rocksclusters.org/Rocks

[5] http://www.openpbs.org

[6] http://www.pbspro.com

[7] http://ganglia.sourceforge.net

[8] http://nagios.org

[9] C. Windeck, *c't 7/2004, S. 150-161*

[10] http://www.mpibpc.gwdg.de/inform/MpiNews/cientif/jahrg9/1.03/1a.03/scta.html

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

[12] http://www.sc-conference.org/sc2003/infra_bwc2.html

[13] http://www.ibm.com/servers/eserver/clusters/software/gpfs.html

[14] http://www.globus.org/datagrid/gridftp.html

[15] R. Stoy, B. Hoeft, *DFN-Mitteilungen 64 (2004) 15-17*