

Neuer Parallelrechner bei der GWDG: AMD-Magny-Cours-Cluster der Firma MEGWARE

Ende Februar 2011 ist ein neuer Parallelrechner-Cluster mit 64 Knoten und 3.072 Cores in den Benutzerbetrieb gegangen. Der Schwerpunkt dieses Clusters liegt in der Bearbeitung von mittelgroßen shared-memory-parallelen Jobs. Die Hardware basiert auf Vier-Prozessor-Boards mit AMD-Magny-Cours-Prozessoren. Untereinander sind die Systeme mit einem QDR-Infiniband-Netz gekoppelt, wodurch auch größere hybrid-parallele Jobs möglich werden. Der Cluster wird über den Frontend *gwdu106* und die LSF-Queue *gwdg-smp* bedient.

Einleitung

Wegen des steigenden Bedarfs nach Rechenleistung für shared-memory-parallele Anwendungen bei den Nutzern der GWDG wurde im Jahr 2010 die Erweiterung der Parallelrechnerkapazität mit SMP-Rechnern geplant.

Die GWDG und das Institut für Astrophysik der Georg-August-Universität Göttingen, das im vergangenen Jahr ebenfalls eine Beschaffung von Parallelrechnerkapazität eingeplant hatte, haben ihre für 2010 zur Verfügung stehenden Mittel in eine gemeinsame Beschaffung eines SMP-Rechenclusters gebündelt, der bei der GWDG betrieben wird und von den Partnern anteilig entsprechend ihrer finanziellen Beteiligung genutzt werden kann. Der zentrale Betrieb anteilig finanzierter und genutzter IT-Ressourcen hat sich in der Vergangenheit bereits vielfach wegen der dabei erreichbaren Synergieeffekte bei Beschaffung, Administration und Auslastung bewährt.

Die GWDG hat in Abstimmung mit dem Institut für Astrophysik Anfang September 2010 im Rahmen einer beschränkten europaweiten Ausschreibung für ein Clustersystem mit Hochgeschwindigkeitskommunikations-

netz acht Unternehmen zur Abgabe eines Angebotes aufgefordert, von denen drei bis zum Stichtag Anfang Oktober ein Angebot abgegeben haben. Nach der Bewertung der Angebote, bei der neben der durch Benchmarks ermittelten Rechenleistung auch der zu erwartende Stromverbrauch berücksichtigt wurde, lag die Firma MEGWARE vorne – mit mehr als 10 % Abstand zum nächsten Angebot.

Die Lieferung des Systems erfolgte Ende Januar 2011, die Abnahme dann nach dem erfolgreichen Durchlaufen eines vierwöchigen Probetriebs mit eingeschränktem Nutzungszugang Ende Februar 2011. Seitdem läuft der neue MEGWARE-Magny-Cours-Cluster im Regelbetrieb. Magny-Cours ist der Name, der von der Firma AMD für seine neuen Zwölf-Core-Prozessoren mit 64-bit-Unterstützung verwendet wird, die in den Rechenknoten des Clusters zum Einsatz kommen.

Beschreibung des MEGware-Magny-Cours-Clusters

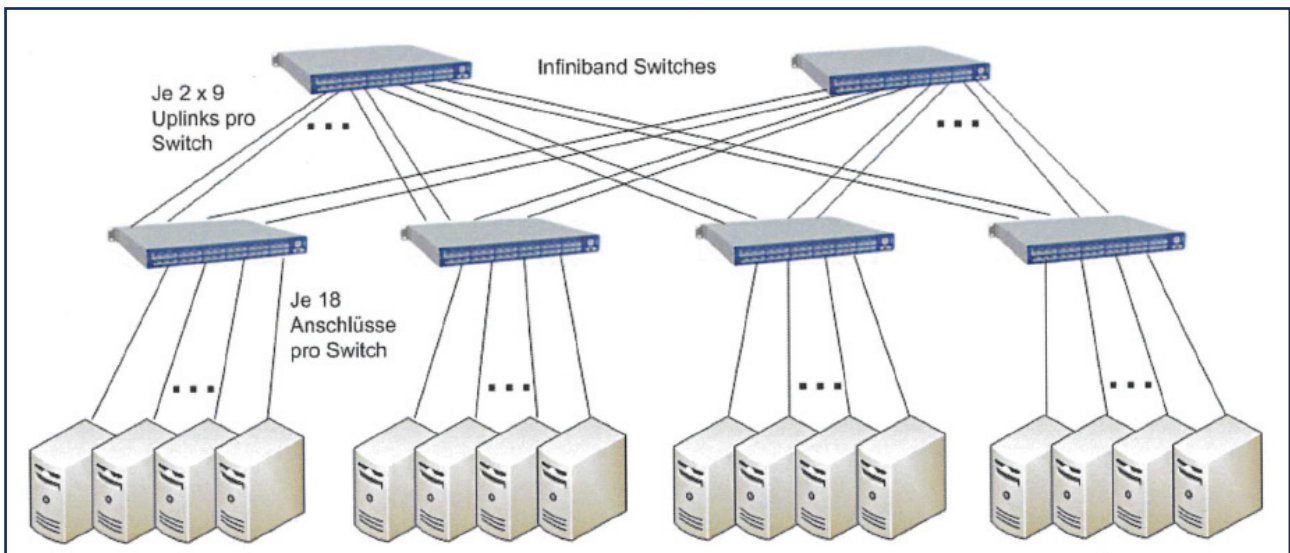
Der MEGWARE-Magny-Cours-Cluster enthält insgesamt 64 Rechenknoten (*gwdp001,...,gwdp064*) mit je vier Zwölf-Kern-Prozessoren vom Typ AMD Magny-Cours

6174, 128 GByte Hauptspeicher und 1 TByte Plattenspeicher, von denen 945 GByte für lokalen /scratch-Bereich zur Verfügung stehen. Der Zugangsrechner *gwdu106* hat zwei Acht-Kern-Prozessoren vom Typ AMD Magny-Cours 6136 mit 32 GByte Hauptspeicher und 1 TByte Plattenspeicher.



1 MEGWARE-Magny-Cours-Cluster

Auch wenn die Einzelleistung mit 8,8 GFlop/s pro Core etwas hinter der Leistung der Vorgängersysteme der x86_64-Architektur zurückbleibt, kommt der neue Parallelrechner mit seinen insgesamt 3.072 Rechenkernen auf eine Gesamtleistung von 27,878 TFlop/s, einem verteilten Hauptspeicher von 8,192 TByte und einem verteilten lokalen Plattenspeicher von 64 TByte. Damit ist der MEG-



2 Infiniband-Netzwerk im MEGWARE-Magny-Cours-Cluster

WARE-Magny-Cours-Cluster das zur Zeit leistungsstärkste Rechensystem bei der GWDG.

Die Gesamt-Peak-Leistung der Parallelrechnersysteme der GWDG wurde durch die Anzahl der neuen Cores mehr als verdoppelt (siehe auch http://www.user.gwdg.de/~parallel/parallelrechner/Hardware_Ueberblick.html), und das System hat die bisher bei den Parallelrechnersystemen der GWDG vorhandenen Cores und den Hauptspeicher beinahe verdoppelt.

Die Kopplung der Rechenknoten erfolgt durch ein Infiniband-Kommunikationsnetz und ein Gigabit-Ethernet-Netz sowie ein Fast-Ethernet-Netzwerk für Servicezwecke. Das Infiniband-Netz ist entsprechend Abb. 2 hierarchisch aufgebaut, wobei in der unteren Schicht vier sogenannte Edge-Switches, Mellanox 50x0 QDR-Infiniband-Switches mit je 36 IB-Ports, die direkte Kopplung jedes einzelnen Knoten mit 40 Gbit/s mit dem Infiniband-Netz gewährleisten. Darüber liegt eine Schicht von zwei baugleichen, aber wegen ihrer Funktion Spine-Switches genannte QDR-Infiniband-Switches. Diese sind mit

jedem der vier Edge-Switches mit einem Trunk von neun Infiniband-Kabeln verbunden und schaffen so ein durchgängiges „fully-non-blocking“-QDR-Infiniband-Netzwerk.

Energieverbrauch

In der Ausschreibung des Systems war eine Vollkostenrechnung über fünf Jahre Laufzeit gegenüber der Gesamt-Rechen- und -Kommunikationsleistung des Systems Grundlage für die Zuschlagserteilung. Ganz im Sinne von „Green IT“ spielten dabei die Energiekosten

und damit der Energieverbrauch eine wesentliche Rolle für die Kaufentscheidung. Das einzelne System hat unter Vollast eine Leistungsaufnahme von ca. 840 Watt, was sich auf 54,6 kW für das Gesamtsystem aufsummiert. Dies ist gegenüber dem zuvor beschafften schon besonders energieeffizienten NEC-System wiederum mehr als eine Verdopplung der Effizienz bezogen auf die Cores. Bezogen auf die Peak-Leistung pro Core, was die wichtigere Bezugsgröße ist, sind die Cores des neuen Systems immer noch mehr als 60 % energieeffizienter als die Cores des NEC-Systems.



3 Cluster-Verkabelung (blau: Service-Netzwerk, gelb: 1 Gbit/s-Ethernet, schwarz: Infiniband)

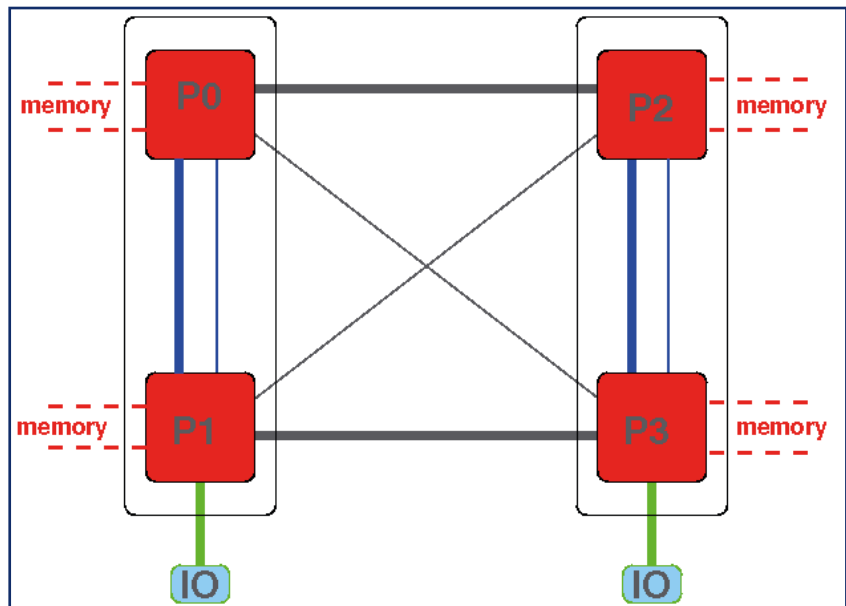
Der Prozessor

Auf dem Prozessorchip 6174 der vierten Generation des AMD® Opteron® aus der Prozessorseerie mit der Kurzbezeichnung Magny-Cours sind je zwei native Sechskernprozessoren Lisbon auf einem Multi-Chip-Modul (MCM) untergebracht. Der gesamte Chip ist in 45nm-Technologie gefertigt und bringt es damit auf 900 Millionen Transistoren und hat immerhin 1.944 Pins, was zum Design einer neuen Fassung für den Prozessor namens G34 führte.

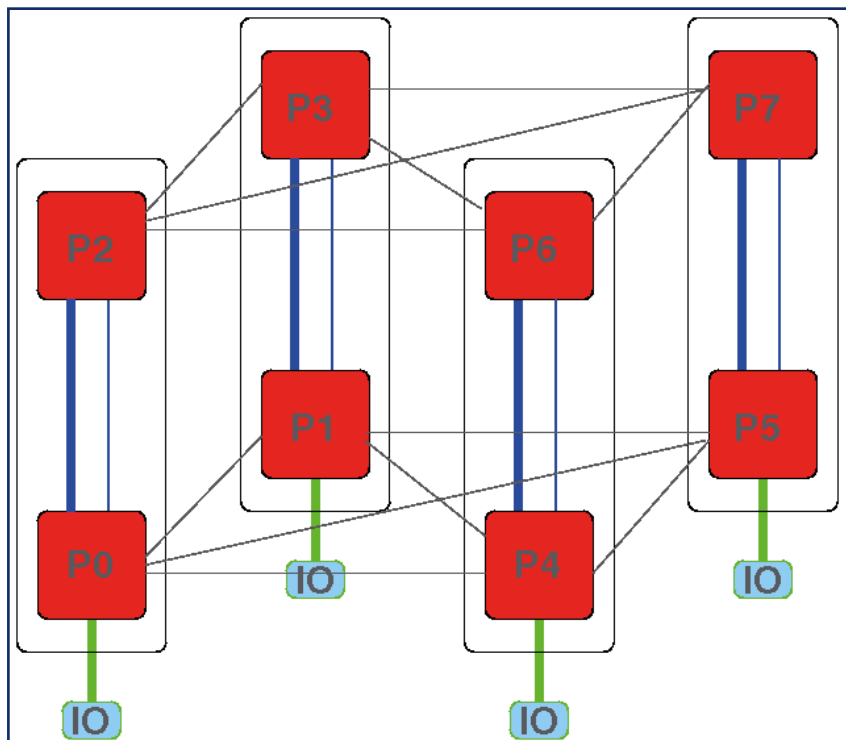
Die Prozessoren sind mit 2,2 GHz getaktet, pro Prozessorsockel gibt es vier HyperTransport-Links (jeweils 6,4 GT/s), und zwei mit 3,2 GHz getaktete Pfade zum Hauptspeicher, womit jeder Lisbon-Chip im MCM seine Speicherbandbreite behält.

Von besonderer Bedeutung für die Rechengeschwindigkeit bei wissenschaftlichen Anwendungen sind die SSE-Verarbeitungseinheiten, die mit ihrer Datenbreite von 128 bit gleichzeitig zwei Fließkomma-Operationen mit 64-bit-Operanden bearbeiten können und pro Takt zwei Ergebnisse liefern. Da von den vier gleichzeitig möglichen Befehlen zwei vom SSE-Typ sein können, liefert jeder Magny-Cours-Kern pro Takt maximal vier Resultate von Fließkomma-Operationen, was eine theoretische Spitzenleistung eines Kerns von 8,8 GFlop/s erlaubt. Tatsächlich wurden in unseren Benchmarks bei der realistischen Anwendung einer Matrix-Multiplikation 8,0 GFlop/s gemessen.

Der Magny-Cours-Prozessor besitzt eine dreistufige Cache-Hier-



4 Die Speicheranbindung und die HT-Verbindungen zwischen den Dies im Prozessor und im Zwei-Prozessor-Board (dicke Striche: x16, dünne Striche x8)

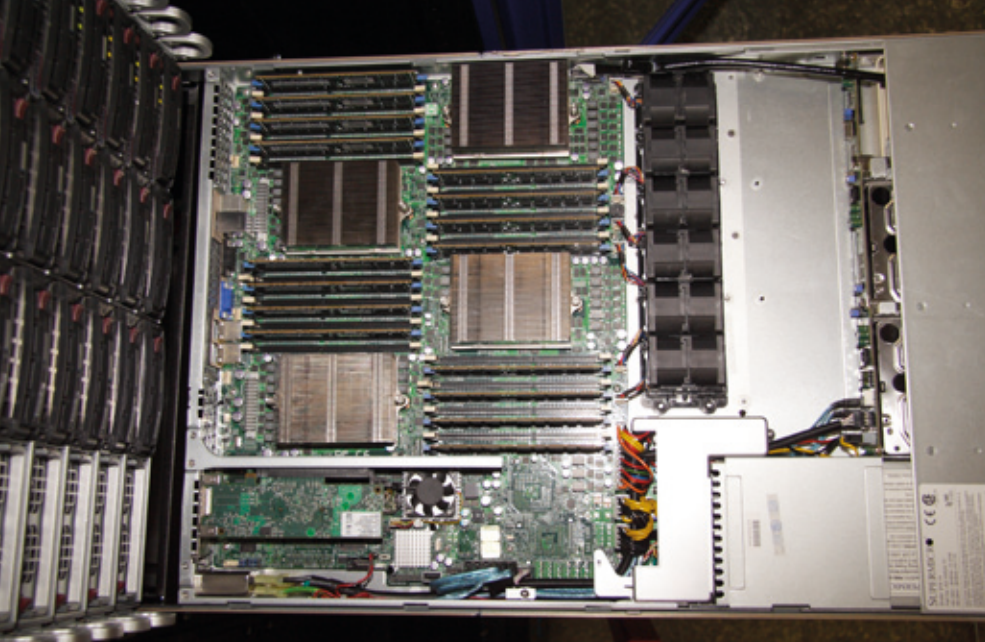


5 Die HT-Verbindungen im Vier-Prozessor-Board (dicke Striche: x16, dünne Striche x8)

archie, deren erste beide Stufen, L1- und L2-Cache, im Wesentlichen gleich geblieben sind. Der L1-Cache hat je Kern 64 + 64 KB (Daten + Instruktionen). Der L2-Cache besitzt 512 KB je Kern und wird, wie der L1-Cache, mit Prozessortakt getaktet. Dahingegen ist der im Takt des Hypertransport getaktete L3-Cache pro Multi-Chip-Modul zweimal 6 MB groß.

Jeder der beiden wird von einem Die mit jeweils sechs Cores genutzt.

Mit seinen zwei vollständigen Sechskern-Lisbon-Prozessoren bildet aber auch der Prozessor selbst ein Core-Netzwerk mit nicht mehr symmetrischen Punkt-zu-Punkt-Verbindungen.



6 Ein geöffneter Clusternode

Das Multi-Core-Modul des Magny-Cours kommuniziert mit vier HyperTransport-Links (6,4 GT/s) nach außen. Dabei ist allerdings ein (nicht kohärenter) x16-Link pro Prozessor für IO Richtung Chipsatz reserviert.

Innerhalb des MCM stehen ein kohärenter x16-Link und ein kohärenter x8-Link für die Kommunikation zwischen den Dies im MCM zur Verfügung. Untereinander sind die beiden Dies also mit eineinhalb HT-Links verknüpft. Wie außerdem in Abb. 4 grafisch dargestellt, besitzt jedes Lisbon-Die einen kohärenten x16-Link und einen kohärenten x8-Link für die Kommunikation über die MCM-Grenzen hinaus für die Zwei-Prozessor-Boards. Bei den Vier-Prozessor-Boards wird die eine x16-Link wiederum aufgespalten in zwei x8-Links, so dass pro Magny-Cours-Chip drei kohärente x8-Links zur Anbindung der weiteren CPUs bereitstehen. Zählt man die beiden oder sechs x8-Links jeweils als halbe Links, so kommt man jeweils auf drei volle HT-Links pro MCM für die Inter-Prozessor-Kommunikation auf dem Board. Zusammen mit dem IO-HT pro MCM zum Chipsatz sind dies dann vier HT-Links.

Der Rechenknoten

Die 4-Wege-Magny-Cours-Knoten sind jeweils in Ein-Höheneinheiten großen Serversystemen (umgangssprachlich auch Pizzaboxen) untergebracht. Das verwendete Board, Supermicro H8QGI-F, mit den AMD-Chipsets SR5690, SR5670 und SP5100 unterstützt die HyperTransport-Links mit jeweils 6,4 GT/s zwischen den Sockeln und ist mit jeweils 16 Speichermodulen á 8 GByte, insgesamt 128 GByte, ausgestattet.

Wie oben beschrieben, kann in einem Zwei-Sockel-System wie dem Frontend jedes Die mit jedem anderen direkt verbunden werden, auch wenn nicht alle Links gleich breit und damit gleich schnell sind. In einer Vier-Wege-Maschine entsteht ein Würfel (siehe Abb. 5), bei dem mit maximal zwei Schritten jeder mit jedem kommunizieren kann.

Der Anschluss an das Infiniband-Netzwerk erfolgt über einen an PCI-Express (PCI-E x8) angeschlossenen Mellanox ConnectX QDR HCA (Host Channel Adapter) mit einer theoretische Bandbreite von 40 Gbit/s. Im Bechmark gemessen wurden 21,46 Gbit/s.

Schnelle Anbindung der Knoten an das StorNext-Speichersystem der GWDG

Auf diesem Cluster wie auch dem NEC-Nehalem-Cluster werden die Home-Verzeichnisse der Benutzer nun über eine schnelle StorNext-Anbindung auf den Cluster-Knoten als LAN-Klienten zur Verfügung gestellt. Die Nutzung entspricht der bisherigen Regelung für die Benutzer im Parallelrechner-Filesystem, wo jedem Benutzer temporär jeweils 300 GB zusätzlich an Plattenplatz bereitstehen mit einer Grace Period von vierzehn Tagen.

Nutzungshinweise

Zugang

Als Frontend für den MEGWARE-Cluster dient der Rechner *gwdu106*, auf dem Sie sich aus dem GÖNET heraus per ssh mit *ssh gwdu106.gwdg.de* einloggen können. Ihr Homeverzeichnis ist dort das gewohnte UNIX-Home wie auf den anderen Dialog-Maschinen der GWDG. Für den Login muss Ihr Account für den Zugriff auf die Frontends der Parallelrechner freigeschaltet sein; falls dies noch nicht der Fall ist, schreiben Sie bitte eine E-Mail an den GWDG-Support (support@gwdg.de).

Programmierungsumgebung

Der Frontend-Rechner hat dieselbe Umgebung installiert wie die einzelnen Knoten des Clusters und kann daher zur Kompilierung und zum kurzen Funktionstest von Programmen verwendet werden.

Es stehen verschiedene Compiler und Bibliotheken zur Verfügung,

die über das sog. Modules-System in der Shell ausgewählt werden können; alle Operationen laufen hierbei über den Befehl *module*. Eine Übersicht über die vorhandenen Module erhalten sich durch Aufruf des Befehls *module avail*.

Die allgemeine Struktur ist unterteilt in

- *comp* für Compiler
- *lib* für Libraries
- *mpi* für MPI-Umgebungen
- *tools* für einzelne Programme

Eine Kurzbeschreibung eines bestimmten Moduls erhalten sie mit *module help <MODULNAME>*. So steht beispielsweise das Modul *comp/intel/11.0* für den Intel-Compiler in Version 11.0; geladen wird dies durch *module load comp/intel/11.0*.

Durch Laden des Moduls wird die Umgebung in Ihrer Shell für diesen Compiler gesetzt, d. h. Umgebungsvariablen wie PATH, LD_LIBRARY_PATH, MANPATH und auch CC, FC etc., die auf *icc* bzw. *ifort* gesetzt werden. Natürlich können Sie mehrere Module hintereinander laden. So lädt ein anschließendes *module load mpi/intelmpi/4.0.1.007* zusätzlich die Intel-MPI-Umgebung. Auch wurde versucht, Abhängigkeiten der Module zu berücksichtigen. Wenn Sie zuerst versuchen, das Intel-MPI-Modul zu laden, werden Sie darauf aufmerksam gemacht, dass zunächst das Modul für den Intel-Compiler geladen werden muss. Um gleich die komplette Intel-Umgebung zu laden, können Sie auch das Module *intel-compiler-suite-v11* verwenden, dass die Module für Compiler, Intel-MPI und MKL lädt.

Ein Vorteil des Modul-Systems ist, dass Sie diese Module auch wieder aus der Umgebung entfernen können, z. B. durch *module unload comp/intel/11.0*.

So können Sie schnell zwischen verschiedenen Compilern und Libraries wechseln. Hierfür gibt es speziell auch den Befehl *switch*. Weitere Dokumentation hierzu finden Sie in der man-page von *module*.

Interaktiver Programmstart

Kurze Funktionstest Ihrer Programme können Sie direkt auf dem Frontend *gwdg106* mit dem Kommando *mpirun* durchführen: *mpirun -n 2 ./mpi_exec*. Bitte achten Sie darauf, dass diese Funktionstest nicht zu viele Ressourcen des Frontends verbrauchen, da sonst der Dialog-Betrieb für andere Nutzer gestört werden könnte.

Batch-Betrieb

Wie auf den anderen Clustern, sorgt das Batch-System LSF über einen Fair-Share-Mechanismus für die gerechte Aufteilung.

Zur Zeit ist im Batch-System die Warteschlange *gwdg-smp* für den MEGWARE-Cluster konfiguriert. Wie üblich ist die maximale Ausführungszeit (Walltime) 48 Stunden, es können bis zu 1.008 Cores verwendet werden.

Das MPI-Programm *mpiprogram* kann z. B. mit folgendem Kommando submittiert werden:

```
bsub -q gwdg-smp -a intelmpi -n <nproc> -W <hh:mm> mpirun.lsf <path_to_mpiprog>
```

Soll statt Intel-MPI Open-MPI verwendet werden, sieht das Kommando so aus:

```
bsub -q gwdg-smp -a openmpi -n <nproc> -W <hh:mm> mpirun.lsf -x LD_LIBRARY_PATH <path_to_mpiprog>
```

Bitte beachten Sie den für Open-MPI notwendigen ausdrücklichen Export der Variable *LD_LIBRARY_PATH* durch die *-x* Option.

Wie gewohnt, können Sie natürlich auch entsprechende Skripte unter Verwendung der *#BSUB*-Zeilen einsetzen. Z. B. kann ein Job für das vorinstallierte Quantenchemie-Paket Gaussian09 mit folgendem Skript gestartet werden:

```
#!/bin/sh
#BSUB -q gwdg-smp
#BSUB -a openmp
#BSUB -n 48
#BSUB -W 24:00
#BSUB -C 0
export g09root="/usr/product/gaussian"
.$g09root/g09/bsd/g09.profile
export GAUSS_SCRDIR="/scratch"
g09 input.com output.log
```

Eine ausführliche Beschreibung des Batch-Systems LSF bei der GWDG finden Sie unter <http://gwdg.de/index.php?id=1334>.

Schwardmann
Boehme
Engster

Kontakt:

Dr. Ulrich Schwardmann
uschwar1@gwdg.de
0551 201-1542

Dr. Christian Boehme
cboehme1@gwdg.de
0551 201-1839

David Engster
dengste2@gwdg.de
0551 201-1559