Forschungszentrum Jülich

ZAM aktuell

Höchstleistungsrechner
Datenkommunikation
Kooperatives Computing
Mathematik

Nr. 117 • August/September 2003

Auf dem Weg zum neuen IBM-Supercomputer: Teilsystem mit 1,3 TFLOPS in Betrieb

Wie geplant wurde am 28. Juli im Zentralinstitut für Angewandte Mathematik der Benutzerbetrieb auf einem Teilsystem des neuen IBM-Supercomputers „Jump“ (Jülich multiprocessor) aufgenommen. Der Rechner steht für Projekte des Wissenschaftlichen Rechnens zur Verfügung, die nach positiver Begutachtung durch die Rechenzeitkommissionen des John von Neumann-Instituts für Computing bzw. des Forschungszentrums Jülich ein Rechenzeitkontingent auf dem System erhalten haben. Das System umfasst sechs Knoten IBM p690, jeweils ausgestattet mit 32 Prozessoren des Typs POWER4+ mit 1,7 GHz Taktrate und 64 GByte gemeinsamem Hauptspeicher. Es erreicht damit eine Peak-Leistung von 1,3 TFLOPS. Fünf der sechs Knoten sind als Compute-Knoten konfiguriert und für die Ausführung paralleler Programme vorgesehen; der sechste Knoten (Login-Knoten) besteht u.a. aus einer Login-Partition mit mehreren Prozessoren und einer Partition für Datenmanagement-Aufgaben (I/O, Backup, etc.).

Die Knoten des Teilsystems sind über ein Gigabit-Ethernet miteinander verbunden. Da dies bei der knotenübergreifenden Ausführung paralleler Programme zu erheblichen Kommunikationsengpässen und damit zu Performance-Verlusten führen würde, können Programme derzeit nur bis zu 32 Prozessoren anfordern. Mit der Endausbaustufe des Rechners Anfang des kommenden Jahres wird auch ein schnelles Verbindungsnetzwerk (4-GByte-Switch) verfügbar sein, mit dem eine knotenübergreifende Programmausführung sinnvoll ist. Mit 37 Knoten wird dieses System eine Spitzenleistung von 8 TFLOPS erreichen.

Interaktiver Zugang und Dateisysteme

Das Teilsystem wird unter dem Betriebssystem AIX 5.1 betrieben, sein Host-Name ist ***jump.fz-juelich.de***. Der Zugang erfolgt über Secure Shell (*ssh*) oder UNICORE; der Benutzer gelangt damit automatisch auf einen Prozessor der Login-Partition. Diese Partition steht für die Vorbereitung, die Submission und die Nachbearbeitung von Batch-Jobs sowie für die Ausführung interaktiver Anwendungen zur Verfügung.

Alle Benutzerdaten werden in HOME-Filesystemen unter Kontrolle von GPFS (General Parallel Filesystem) gespeichert. Auf die Daten kann von jedem Knoten aus in gleicher Weise zugegriffen werden. Länger nicht genutzte Daten werden durch die Hierarchical Storage Management Software (HSM) in den Kassettenroboter migriert und bei Bedarf automatisch zurückgeholt. Daher wird kein ARCHIVE-Filesystem, wie es von den Cray-Systemen her bekannt ist, benötigt. Zur temporären Speicherung von Dateien existiert ein WORK-Filesystem, das im Gegensatz zum HOME-Filesystem nicht gesichert wird und regelmäßig bereinigt wird. Das WORK-Filesystem ist wie die HOME-Filesysteme Teil des GPFS. Die Übertragungsgeschwindigkeit einer Anwendung zum GPFS kann zur Zeit bis zu 30 MByte/s betragen.

Nutzung der Compute-Nodes

Auf dem IBM-Supercomputer werden Batch-Jobs und interaktive parallele Programme vom IBM-eigenen Batch-System LoadLeveler verwaltet. Batch-Jobs werden mit dem Aufruf *llsubmit <script>* submittiert, wobei die angegebene Script-Datei außer den auszuführenden Anweisungen Steueranweisungen und Ressourcenangaben an den LoadLeveler enthält. In Abhängigkeit von der angeforderten Anzahl von Prozessoren und der spezifizierten Wall-Clock-Zeit wird im LoadLeveler automatisch eine geeignete Job-Klasse ausgewählt; eine eventuell vorhandene Klassenangabe des Benutzers wird dabei nicht beachtet. Maximal können derzeit pro Job 32 Prozessoren und eine Wall-Clock-Zeit von vier Stunden angefordert werden. Statusabfragen der Jobs und des LoadLevelers sind möglich mit *llq, llqall, llclass, llstatus* oder *xloadl*. Interaktive parallele Programme (bis zu 30 Minuten) oder Debugging-Läufe können mit dem FZJ-spezifischen Kommando *llrun* gestartet werden. Diese werden ebenfalls unter der Kontrolle des LoadLevelers auf den Compute-Knoten ausgeführt.

Forschungszentrum Jülich GmbH · D-52425 Jülich · Zentralinstitut für Angewandte Mathematik
Telefon: +49 (2461) 61 - 64 00 · Telefax: +49 (2461) 61 - 2810 · E-mail: zam@fz-juelich.de · URL: <http://www.fz-juelich.de/zam/>

Datenkonvertierung und -transfer von Cray zu IBM

Bei der Migration von Cray-Anwendungsprogrammen auf den IBM-Supercomputer müssen auch die Eingabedaten der Anwendungen migriert werden. Hierbei ist zu berücksichtigen, dass die internen Zahlenformate der beiden Systeme nicht bzw. nur teilweise kompatibel sind. In der Regel wird es also notwendig sein, numerische Eingabe-Datensätze vom Cray- in das IBM-Format zu konvertieren. Hierfür stehen verschiedene Hilfsmittel auf den Cray-Systemen zur Verfügung:

- Foreign File Conversion: Ein spezieller Cray I/O Layer, der es Fortran-Anwendungen erlaubt, Daten in Fremdformaten zu lesen bzw. zu schreiben. Beispielsweise können hiermit Ausgabedaten auf einem Cray-System im IBM-Format erzeugt werden.
- FDCP: Eine Utility zum Konvertieren des Fortran Record-Formats. Diese Utility ist besonders für die Migration von Daten der CRAY T3E-Systeme geeignet.
- Konvertierungsroutinen: Sie erlauben die explizite Konvertierung von Cray-Daten in das IBM-Format und sind vor allem für die Benutzung in C- oder C++-Anwendungen gedacht.

Eine interessante Alternative zur Datenkonvertierung stellt die Benutzung maschinenunabhängiger Datenformate dar. Solche De-Facto-Standards, wie XDR und NetCDF, werden in einigen Wissenschaftsbereichen seit langem für den Datenaustausch genutzt.

Die Übertragung von Dateien von den Cray-Systemen zum IBM-Supercomputer kann vom IBM-System aus mit *ftp*, *xftp* oder *scp* unter Angabe des Rechnernamens „craydata“ (schnelle HiPPI-Verbindung) erfolgen, z.B. *ftp craydata*; die Authentifizierung geschieht über Cray-Benutzernummer und Passwort. Da es sich hier um eine interne, gesicherte Verbindung handelt, entfallen die üblichen sicherheitstechnischen Beschränkungen für *ftp*.

Parallele Programmierumgebung

Die Architektur und Software des IBM-Supercomputers unterstützen eine Vielzahl von parallelen Programmierparadigmen: Message-Passing mit MPI (basierend auf MPI 1.2 und den in MPI 2 definierten Funktionen „einseitige Kommunikation“ und „parallele Ein-/Ausgabe“), einseitige Kommunikation mit IBM LAPI oder Cray SHMEM (basierend auf IBM TurboSHMEM), Multi-Threading mit OpenMP oder Posix Threads sowie beliebige Kombinationen derselben (hybride Programmierung). Neben den dazu notwendigen Compilern und Bibliotheken stehen folgende Werkzeuge zur Verfügung: der parallele Debugger TotalView, das MPI-Leistungsanalysewerkzeug Vampirtrace/Vampir, das OpenMP-Leistungsanalysewerkzeug GuideView, das OpenMP-Verifikationstool Assure sowie verschiedene Tools der Firma IBM (Callgraph-Profiler gprof/Xprofiler, Hardware-Counter-Profiling-Tools hpmcount und hpmlib, MPI-Profiling-Werkzeug MP_profiler, Cache-Simulator Sigma). Die IBM-Tools müssen über das *module*-Kommando eingebunden werden.

Mathematische Software und Anwendungssoftware

Derzeit sind auf dem IBM-Supercomputer unter anderem die folgenden Pakete installiert:

- *ESSL*: sequentielle Basisbibliothek; enthält *BLAS*-Routinen (auch für dünnbesetzte Matrizen), FFT, etc.
- *ESSLsmp*: Inhalt wie bei *ESSL*; einige Routinen sind parallelisiert mit Hilfe von OpenMP (Shared-Memory)
- *PESSL*[*smp*]: MPI-parallele Version der *ESSL*, deutlich geringerer Umfang als *ESSL*; enthält parallele *BLAS*-Routinen, Teile von *ScaLAPACK 1.5*, FFT etc.
- *BLACS*[*smp*]: Kommunikationsbibliothek zu *PESSL* und *ScaLAPACK*. Die Versionen mit *smp* im Namen können in Verbindung mit Posix-Threads oder OpenMP verwendet werden. Wird die 64-Bit-Adressierung genutzt, müssen die *smp*-Versionen verwendet werden.
- *WSMP* Watson Sparse Matrix Package: Parallelismus teils mit Posix-Threads, teils mit MPI implementiert
- *LAPACK 3.0* Linear Algebra PACKage: nur sequentielle Routinen, braucht *ESSL*[*smp*] für *BLAS*
- *ScaLAPACK 1.7* Scalable Linear Algebra PACKage: braucht *ESSL*[*smp*] für *BLAS* und *BLACS*[*smp*] zur Kommunikation oder *BLACSpd*, Public-Domain-Version der *BLACS*, da die *PESSL*-Version nicht vollständig ist
- *ARPACK*, ARnoldi PACKage: iterativer Löser für dünnbesetztes Eigenwertproblem
- *PARPACK*, Parallel *ARPACK*: MPI-parallele Version
- *NAG Fortran 77 Library Mark 19*: nur Double Precision
- *ANSYS*: sequentielles Finite-Elemente-Paket
- *LS-DYNA*: paralleles Finite-Elemente-Paket
- *Gaussian03*: Ab-initio Chemieprogrammpaket, Threads (shared memory)
- *CPMD 3.7*: MPI-paralleles Car-Parrinello MD-Paket, das auch im hybriden Modus genutzt werden kann

Folgende Softwarepakete erlauben auf Jump graphisches Post-Processing: *IDL*, *AVS/Express*, *RAPS*, *GLview*, *VMD* und *MOLMOL*.

Weitere Informationen

Dokumentation zur Nutzung des IBM-Supercomputers finden Sie unter ***http://jumpdoc.fz-juelich.de***. Ein Handbuch ist in Vorbereitung. Bei Problemen mit dem Betriebssystem, der System-Software, Compilern, Bibliotheken, etc. kontaktieren Sie bitte die zentrale Beratung (Tel. 6400, *beratung.zam@fz-juelich.de*), bei Fragen zur Anwendungssoftware, Parallelisierung oder Optimierung von Programmen wenden Sie sich bitte an die Beratung Supercomputing (Tel. 4416, *sc.zam@fz-juelich.de*).

Redaktion: Dr. Sabine Höfler-Thierfeldt, Tel. 6765