Technischer Überblick

SuperMUC: ein neuer Höchstleistungsrechner für Europa

Mit mehr als 3 Petaflops Rechenleistung ist SuperMUC, der neue Rechner des Leibniz-Rechenzentrums, einer der leistungsfähigsten und universell nutzbarsten Computer in Europa und weltweit. Aber wie funktioniert eigentlich ein solcher Rechner?

Von Matthias Brehm und Reinhold Bader

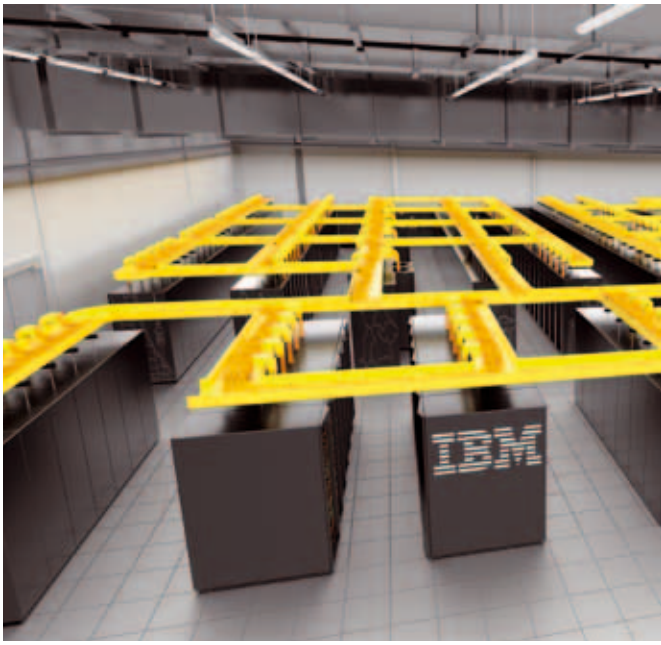


Abb. 1: Computergeneriertes Bild des neuen Höchstleistungsrechners SuperMUC, Juni 2012.

NACH EINER BETRIEBSZEIT von fünf Jahren wurde der Höchstleistungsrechner in Bayern (HLRB II), eine SGI Altix 4700, Ende Oktober 2011 außer Betrieb genommen und durch ein wesentlich leistungsfähigeres System mit dem Namen SuperMUC ersetzt (Abb. 1). Bei SuperMUC handelt es sich um die erste Ausbaustufe eines Clustersystems der Firma IBM, das aus 19 miteinander gekoppelten Rechnerinseln besteht. Das System wurde im obersten Stockwerk des erweiterten Rechnerkubus des LRZ installiert. Die für 2014 geplante zweite Installationsstufe wird dann zusätzlich den Platz des bisherigen Rechners einnehmen.

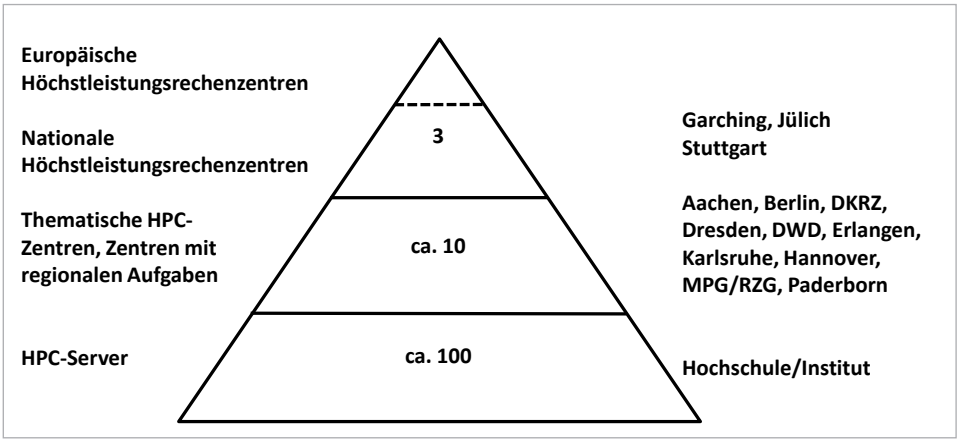
Die Leistungsdaten des neuen Systems sind bereits in der ersten Ausbaustufe imposant: Mit einer Spitzenrechenleistung von etwas mehr als 3 Petaflops (also drei Billiarden Rechenoperationen pro Sekunde oder eine 3 mit 15 Nullen), mehr als 150.000 Prozessorkernen und 300 Terabyte Arbeitsspeicher wird es zum Beschaffungszeitpunkt einer der leistungsfähigsten Rechner in Europa und der Welt sein, der auf Grund der Art der Prozessoren zudem universell nutzbar ist. Während der bisherige Rechner vor allem für Projekte aus Wissenschaft und Forschung innerhalb Deutschlands genutzt wurde, ist der neue Rechner auch Teil des deutschen Beitrags zur europäischen Höchstleistungsrechner-Infrastruktur innerhalb von PRACE (Partnership for Advanced Computing in Europe). Der deutsche Betrag wird vom Gauss Centre for Supercomputing e.V. (GCS) erbracht.

Auf dem Weg nach Europa: das Gauss Centre for Supercomputing

Die drei nationalen Höchstleistungsrechenzentren, das Höchstleistungsrechenzentrum Stuttgart (HLRS), das Jülich Supercomputing Centre der Forschungszentrum Jülich GmbH (JSC) und das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ), haben ihrer langjährigen Zusammenarbeit im Jahr 2007 mit der Gründung des Gauss Centre for Supercomputing eine organisatorische Basis gegeben. Das GCS stellt die nachhaltige Versorgung der computergestützten Wissenschaften in Deutschland und Europa mit Höchstleistungs-Rechenkapazität der obersten Leistungsklasse (Capability Computing) sicher (Abb. 2).

Bereits im Sommer 2008 begannen innerhalb des GCS umfangreiche Aktivitäten zur Koordinierung der

Abb. 2: Die Versorgungspyramide des High Performance Computing (HPC) in Deutschland.



BEIDE ABB.: LRZ/H. SATZGER; LRZ

Beschaffung, insbesondere zur Auswahl der jeweiligen Rechnerarchitektur, des Beschaffungszeitpunktes und der Finanzierung. Gleichfalls wurden die Koordinierung der Nutzerbetreuung, die Definition gemeinsamer Nutzungsrichtlinien und eines abgestimmten Zugangs- und Review-Verfahrens sowie gemeinsame Schulungs- und Trainingsmaßnahmen vorangetrieben. Der daraus entstandene Förderantrag führte schließlich dazu, dass der Bund (50 %) und die Länder Baden-Württemberg, Nordrhein-Westfalen und Bayern (jeweils 16,6 %) insgesamt 400 Millionen Euro für die Beschaffung und den Betrieb von Höchstleistungsrechnern bis zum Jahr 2017 bereitstellten. Der für SuperMUC verfügbare Anteil beträgt ein Drittel des Gesamtbetrags.

Zur Architektur des Rechners

Bei der Beschaffung des Rechners SuperMUC standen drei Aspekte im Vordergrund: eine breite und einfache Nutzbarkeit des Systems für verschiedene Wissenschaftsdisziplinen, hohe Zuverlässigkeit sowie eine möglichst hohe Energieeffizienz. Diese Ziele diskutierte das LRZ ab Mai 2009 mit Herstellern im Rahmen einer Markterkundung. Nach dem europaweiten Teilnahmewettbewerb wurde ab März 2010 in einem Wettbewerblichen Dialog mit vier Firmen intensiv verhandelt und eine umfangreiche Leistungsbeschreibung erstellt, die auch Benchmark-Programme beinhaltete. Die Entscheidung fiel im November 2010 zugunsten von IBM mit dem System X iDataPlex, das auf 64 Bit Intel Standard-Prozessoren der neuesten Generation basiert. Die wichtigsten Charakteristika des neuen Rechners sind:

	Thin Node-Insel	Fat Node-Insel (zugleich Migrations-system SuperMIG)
Anzahl Inseln	18	1
Anzahl Cores	147.456	8.200
Anzahl Knoten	9.216	205
Prozessor	Intel Sandy Bridge-EP	Intel Westmere-EX
Peak-Rechenleistung (PFlop/s)	2,94	0,078
Gesamter Hauptspeicher (TByte)	288	51
Gemeinsamer Hauptspeicher pro Knoten (GByte)	32	256
Bandbreite zum Hauptspeicher pro Core (GByte/s)	6,4	4,3
Verbindung innerhalb einer Insel	FDR10	QDR
InfiniBand-Verbindungstopologie innerhalb der Inseln	Non-blocking Fat Tree	Non-blocking Fat Tree
Verbindung zwischen den Inseln	FDR10	
Verbindungstopologie zwischen Inseln	Ausgedünnter (4:1) Fat Tree	
Bisektionsbandbreite des Verbindungsnetzwerkes	35,6 TByte/s	
Größe und Bandbreite des parallelen Dateisystems GPFS	10 PByte mit 200 Gbyte/s	
Größe und Bandbreite des Home Dateisystems	1,5 PByte mit 10 GByte/s	
Stromverbrauch des Systems (MW)	<3	

Das Systemkonzept des SuperMUC

Das Gesamtsystem ist in 19 Compute-Inseln mit jeweils etwa 8.200 Rechenkernen unterteilt (Abb. 3). Eine dieser Inseln ist mit Knoten mit besonders viel Hauptspeicher ausgestattet (eine sog. Fat Node-Insel mit 205 Rechenknoten). Die hierbei verwendete Prozessortechnologie ist Intel Westmere-EX. Ein Rechenknoten besteht aus 40 Cores, die auf einen gemeinsamen Hauptspeicher von 256 Gigabyte zugreifen. Diese Insel wurde schon 2011 vorab als Migrationssystem („SuperMIG“) geliefert; sie soll nach der Integration ins Gesamtsystem durch Programme benutzt werden, die extrem viel gemeinsamen Hauptspeicher benötigen, etwa für Pre- oder Postprocessing.

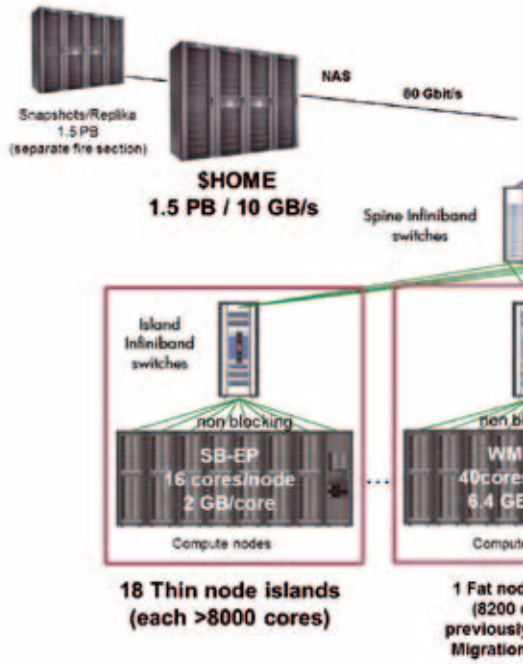


ABB.: LRZ; IBM

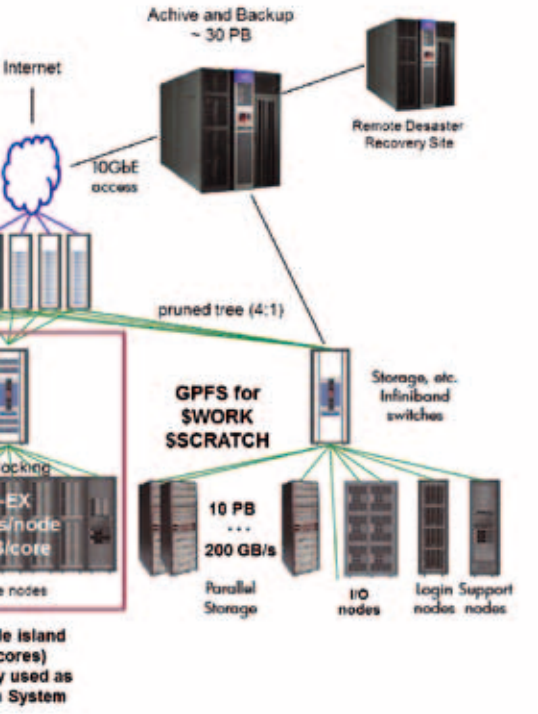
Abb. 3: Die Architektur des SuperMUC.

Der Großteil des Systems besteht aus deutlich schlankeren Rechenknoten mit je 16 Rechenkernen und 32 GByte Hauptspeicher (sog. Thin Node-Inseln); hoch skalierbare Programme sind in der Lage, ihre Daten über eine Vielzahl solcher Knoten zu verteilen und dennoch effizient auf diesen Daten zu operieren. Jede Insel besteht aus 512 Knoten (zuzüglich Ausfallreserve und Serviceknoten). Die Besonderheit hierbei ist, dass alle Knoten wassergekühlt sind, sowohl für die Prozessoren als auch die Speicherbausteine (Abb. 4). Ein Knoten besteht aus jeweils zwei 8-Core Sandy Bridge-EP Sockeln, deren Eigenschaften besonders für das Hochleistungsrechnen geeignet sind.

Jeder Sandy-Bridge Sockel besitzt einen für die acht Cores gemeinsamen Level 3-Cache von 20 Megabyte, die L1- und L2-Caches sind dagegen jedem Core dediziert zugeordnet. Die theoretische Bandbreite zum Hauptspeicher liegt bei 6,4 Gigabyte/s pro Core. Ein entscheidendes Architekturmerkmal der Prozessoren ist ihr erweiterter Befehlssatz (Nachfolge der SSE4-Befehle) mit der Bezeichnung Advanced Vector Extension (AVX). Für rechenintensive Aufgaben erlaubt AVX, pro Takt acht Gleitkommaoperationen mit 64 Bit Genauigkeit auszuführen, was einer Verdopplung der Leistung pro Takt gegenüber bisherigen Intel-Architekturen bedeutet. Eine weitere Besonderheit der Sandy-Bridge Prozessoren ist die Hochgeschwindigkeitsverbindung zwischen den Cores und den L3-Caches. Hiermit wird die Skalierbarkeit von Applikationen im gemeinsamen Hauptspeicher deutlich verbessert und zusammen mit der Quickpath-Technologie, die Zugriffe von einem Sockel auf den anderen ermöglicht, auch das bekannte NUMA-Problem (non-uniform memory access, d. h. ungleiche Speicherzugriffszeit je nach Lage der Daten) abgemildert, das für bestimmte Anwendungsklassen erhebliche Leistungseinbußen nach sich ziehen konnte. Schließlich gibt es für die Hyperthreading-Technologie zwei vollständige Registersätze, die eine Trennung von Betriebssystem und Rechenaufgaben ermöglichen oder eine bessere Auslastung der Rechenwerke gestatten.

Die Knoten einer Insel sind in Form eines sog. Fat Trees mit einem nicht-blockierenden InfiniBand-Netzwerk in FDR10-Technologie (QDR bei der Fat Node-Insel) miteinander verbunden. Die Latenz für den Nachrichtenaustausch zwischen Knoten beträgt weniger als zwei Mikrosekunden. Da erwartet wird, dass die meisten Applikationen nur eine oder wenige Inseln benutzen werden bzw. sich durch das Volumen-/Oberflächenverhältnis bei Gebietsaufteilung auch der Datentransfer über Inselgrenzen hinweg reduziert, ist aus Kostengründen zwischen den Inseln nur ein um den Faktor 1:4 ausgedünntes Netzwerk vorgesehen. Im Prinzip ermöglicht dies einer Applikation, das gesamte System zu nutzen, vorausgesetzt die Bandbreitenanforderungen sind nicht extrem hoch. Applikationen bis zu einer Größe von ca. 37.000 Rechenkernen können prinzipiell sogar die volle Bandbreite nutzen.

Hohe Bandbreite, hohe Zugriffsraten und Datensicherheit sind für heutige Hochleistungsrechensysteme entscheidend, denn seit Jahren wird ein Trend zum datenintensiven Rechnen beobachtet. Um die erwarteten enormen Datenmengen effizient speichern und verarbeiten zu können, wurde ein 10 Petabyte großes paralleles Speichersystem beschafft, das mit dem IBM General Parallel File System (GPFS) als bewährter und bekannter Software betrieben wird. Als Hardware wird ein System von DDN eingesetzt. Die aggregierte Bandbreite beträgt 200 Gigabyte/s. GPFS dient vor allem für die Speicherung von großen Ergebnisdatensätzen. Für die Speicherung der Benutzerdaten wie Programmquellen, Eingabedatensätze und Daten zur Jobsteuerung dient ein anderes Network Attached Storage (NAS) System von NETApp, das eher für viele



kleine Dateien optimiert ist und ein hohes Maß an Zuverlässigkeit bereitstellt. RAID, End-to-End Data Integrity, Snapshots und asynchrones Mirroring auf ein zweites System in einem anderen Brandabschnitt des Rechnergebäudes ermöglichen ein hohes Maß an Sicherheit, nicht nur gegen Plattendefekte oder physische Zerstörung, sondern auch gegen unbeabsichtigtes Löschen oder Überschreiben von Daten durch den Benutzer selbst, was in der Praxis durchaus häufig vorkommt. Die Größe des Plattenplatzes für den NAS-Speicher beträgt 1,5 Petabyte (plus 1,5 Petabyte für die gespiegelten Daten) bei einer aggregierten Bandbreite von 10 Gigabyte/s.

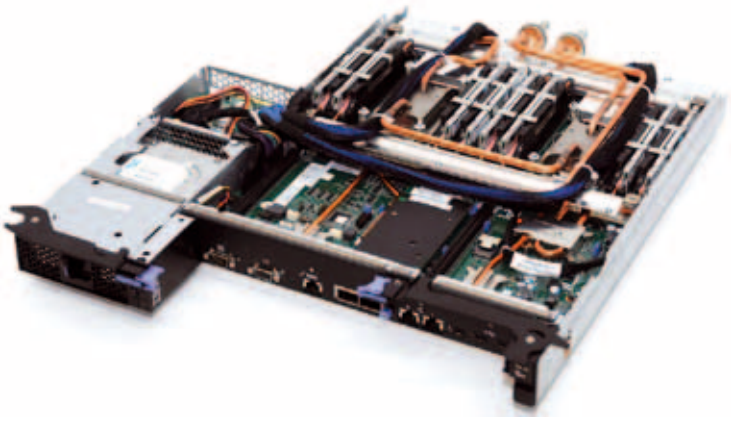
Neben der Speicherung auf Platten können die wertvollen Ergebnisdaten auch auf Bändern archiviert werden. Hierfür hat das LRZ in einer ersten Ausbaustufe zwei Bandroboter in zwei räumlich getrennten Rechenzentren vorgesehen. Insgesamt steht auf 11.000 Bändern eine Speicherkapazität von 16,5 Petabyte zur Verfügung. 22 LTO-5 Laufwerke und vier Hochleistungsserver ermöglichen einen Gesamtdurchsatz von 3 Gigabyte/s. Um das Archivieren zu beschleunigen, wird zusätzlich ein 8 Terabyte großer SSD-Speicher für die Metadaten und ein 2 Petabyte großer Plattenspeicher als Cache verwendet. In einer zweiten Ausbaustufe soll die Gesamtkapazität auf 44 Petabyte und der Durchsatz auf 6 Gigabyte/s erhöht werden.

Zusätzlich zu den Rechenknoten gibt es noch zahlreiche Service- und Managementknoten, z. B. für das Login zum interaktiven Arbeiten, Archivieren und Backup oder Monitoring.

Energieeffizienz durch völlig neue Warmwasserkühlung

Der Energieverbrauch des neuen Rechners von etwa drei Megawatt unter Volllast, zu dem noch der Aufwand für die Kühlung hinzukommt, stellte das LRZ vor hohe finanzielle und technische Probleme und prägte ganz entscheidend die Verhandlungen mit den Herstellern. So wurde beschlossen, bei der Kühlung des Rechners einen völlig neuen Weg zu beschreiten. Die meisten Knoten des SuperMUC nutzen eine Warmwasserkühlung, die aufgrund hoher Vorlauftemperaturen gleich mehrere Vorteile verbindet: Etwa 10 % Energie werden so gespart, da auf den Knoten weniger bis gar keine aktiven Lüftungskomponenten mehr benötigt und Leckströme verringert werden. Außerdem braucht das Rechenzentrum keine energieintensiven Kältemaschinen, was den Energieverbrauch des Gesamtsystems erheblich reduziert. Die Wasserkühlung bringt zudem wertvolle Wärmeenergie zurück, die sich vielfältig verwenden lässt. Im Vergleich zu konventionellen, mit Kaltluft gekühlten Systemen reduzieren sich die CO_2-Bilanz und auch der Lärmpegel im Rechnerraum signifikant. Die Warmwasserkühlung, die die Chips des Systems direkt kühlt, wurde eigens durch IBM entworfen und implementiert. Der SuperMUC kombiniert diese Kühlung mit den energieeffizienten Intel Xeon Prozessoren und einer anwendungsorientiert arbeitenden Systemsoftware. Durch all diese Maßnahmen soll der gesamte Energieverbrauch um 30 bis 40 % gesenkt und ein wesentlicher Beitrag zum Klimaschutz geleistet werden (s. auch S. 24–26).

Abb. 4: Ein Knoten des Rechnersystems mit der innovativen Warmwasserkühlung.



Die Software-Umgebung des SuperMUC

SuperMUC wird mit dem Betriebssystem SUSE Linux Enterprise Server (SLES 11) betrieben. Zusätzlich werden die folgenden Softwarekomponenten verwendet:

- IBM Tivoli Workload Scheduler LoadLeveler als Batchsystem, um Jobs, Jobklassen und Computerressourcen zu verwalten,
- IBM General Parallel File System, um die vielen Einzelplatten zu einem parallelen Filesystem zu koppeln,
- IBM Parallel Environment mit einer hochoptimierten MPI-Implementierung und Tools für die Performanceanalyse,
- Mellanox Unified Fabric Manager (UFM), um die InfiniBand-Infrastruktur zu managen und zu überwachen und
- Extreme Cluster Administration Toolkit (xCAT), mit dem die Installation, die Softwareprovisionierung und das Management eines solch großen und komplexen Rechensystems vereinfacht werden.

Um optimalen Code aus Fortran, C- oder C++ Quellen zu generieren, kommen die Produkte von Intel zum Einsatz, die in der Lage sind, die besonderen Eigenschaften der Prozessoren (z. B. AVX) optimal auszunutzen. Darüber hinaus wird auch die OpenMP-basierte parallele Programmierung innerhalb eines Knotens von diesen Compilern bereitgestellt. Obwohl die C/C++ Compiler von Intel mit den GNU-Compilern weitgehend kompatibel sind, sind auf dem System auch alle GNU-Compiler vorhanden. Der Intel VTune Amplifier liefert Angaben über das Performanceverhalten und Schwachstellen von Applikationen aufgrund von Countermessungen innerhalb der Prozessoren, der Intel Inspector kann Fehler bei der Speicherverwaltung oder beim Threading herausfinden. Der Intel Trace Analyzer and Collector oder Vampir NG der Technischen Universität Dresden ermöglichen es, das Skalierungsverhalten von MPI-Programmen besser zu verstehen und zu optimieren. Schließlich steht mit Intel MPI eine zweite MPI-Implementierung zur Verfügung, die es erlaubt, Applikationen zu entwickeln, die auch außerhalb einer IBM-Umgebung ablauffähig sind bzw. von Drittherstellern gelieferte Programme auf dem SuperMUC auszuführen. Die Produkte von Intel werden durch eine Vielzahl von Open-Source Tools zur Performanceanalyse ergänzt (PAPI, Likwid, Scalasca, IPM u. v. a. m.). Für die Fehlersuche steht der DDT Debugger mit einer graphischen Benutzerschnittstelle bereit.

Das LRZ arbeitet im Rahmen diverser Drittmittelprojekte an der Weiterentwicklung und Anpassung von Programmierwerkzeugen für den SuperMUC.

Um eine hohe Applikationsleistung zu erreichen, sind speziell optimierte mathematische Bibliotheken für die Lineare Algebra oder Fouriertransformationen entscheidend. Hierzu werden unter anderem die Intel Math Kernel Library (MKL), das Portable Extensible Toolkit for Scientific Computations (Petsc), FFTW, die NAG Bibliotheken sowie die GNU Scientific Library bereitgestellt. Zahlreiche Applikationen aus den Bereichen Strömungsdynamik, Strukturmechanik, Elektromagnetik, Chemie und Festkörperforschung stehen Anwendern, die keine eigene Programmentwicklung betreiben, für anspruchsvolle Simulationen zur Verfügung.

Rechenbetrieb

Der größte Teil des SuperMUC wird über das Batch-System LoadLeveler zugänglich sein, jedoch wird eine Thin Node-Insel vorrangig für den interaktiven Zugriff bereitgestellt. Damit sollen die Programmentwicklung und die Fehlersuche in hochskalierbaren Anwendungen beschleunigt werden. Die Maximallaufzeit großer paralleler Programme wird im Normalfall auf zwei Tage begrenzt sein; der Anwender muss daher selber dafür sorgen, dass die für den Neustart des Programmes notwendigen Daten in regelmäßigen Abständen auf das parallele Dateisystem hinausgeschrieben werden.

Das LRZ ermöglichte den Nutzern einen nahtlosen Übergang auf das neue System, indem es eine zum bisherigen Höchstleistungsrechner äquivalente Softwareausstattung auf dem Migrationssystem SuperMIG bereitstellte. Von nun an steht am LRZ eine durchgängige und weitgehend binär-kompatible Softwareumgebung zur Verfügung, beginnend beim Linux-Desktop über das Linux-Cluster bis hin zur höchsten Leistungsklasse des SuperMUC. Die Herausforderung der nächsten Jahre wird es sein, für ein so großes System die Parallelität der Applikationen weiter zu erhöhen und das System effizient zu nutzen.

DIE AUTOREN

Dr. Matthias Brehm leitet die Gruppe Applikationsunterstützung am Leibniz-Rechenzentrum, Dr. Reinhold Bader die Gruppe HPC Server und Dienste.