Andreas Stiller

# Dresden macht Dampf

## Neuer Petaflops-Rechner an der TU Dresden

**Woher stammt die erste deutsche Lokomotive? Nu gloa, nicht aus Leipzig, sondern aus Dresden (Übigau) von Johann Andreas Schubert, der an der dortigen TU lehrte, die damals noch Technische Bildungsanstalt zu Dresden hieß. Und dort macht man nun mit einem neuen Rechner (Wasser-)Dampf.**

Zwar hat es etwas länger gedauert als geplant – unter anderem verzögerte zwischenzeitlich sogar ein Brand im neugebauten Lehmann-Gebäude den Fortgang –, doch nun ist er im Vollausbau da und wartet auf Arbeit: der neue Hochleistungsrechner/Speicherkomplex HRSK II am Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden.

Rund 60 Millionen Euro haben der Bund und der Freistaat Sachsen in Rechner und Gebäude investiert und so erschienen denn auch zur feierlichen Einweihung die Geldgeber in Gestalt der Bundesministerin für Bildung und Forschung, Prof. Dr. Johanna Wanka, und des sächsischen Ministerpräsidenten Stanislaw Tillich.

### Rangfragen

Während der (längeren) Wartezeit auf die Bundesministerin rätselten einige der Gäste, wie wohl die korrekte Rangfolge bei der Begrüßung ist, erst Bundesministerin, dann Ministerpräsident oder umgekehrt? ZIH-Direktor Prof. Dr. Wolfgang E. Nagel löste das chevaleresk, indem er die Dame zuerst begrüßte – wiewohl das streng protokollarisch eigentlich die falsche Reihenfolge war. Zwar rangiert eine Bundesministerin in der Top-70-Liste der offiziellen Rangfolge auf Position 23 um einen Platz über einem Ministerpräsidenten – es sei denn, der ist gleichzeitig Bundesratspräsident –, aber bei Veranstaltungen im eigenen Bundesland geht dann eben doch der Ministerpräsident vor. Und so oblag es dann auch Stanislaw Tillich, mit den Festreden der Gäste zu beginnen. Er lieferte zunächst eine galante Entschuldigung für die Verspätung der aus Berlin angereisten Bundesministerin: „Wir haben hier in Sachsen zwar fixe Rechner, aber noch nicht so fixe Autobahnen …“

Beide Politiker erinnerten in ihren Grußworten an eigene Computererfahrungen aus früheren DDR-Jahren. Tillich, der an der TU Dresden Konstruktion und Getriebetechnik studiert hatte, berichtete unter anderem über eine irgendwie in die DDR eingeschmuggelte HP 9000 mit 24 KByte Speicher (okay, die Kiste von 1987 hatte eigentlich schon 24 MByte …). Professorin Wanka hatte nicht weit weg von Dresden, in Leipzig ihr Diplom in Mathematik erhalten. Sie plauderte über ihre ersten Programmiererfahrungen mit Algol 60 und erwähnte in diesem Zusammenhang auch den bekannten DDR-Informatiker Nikolaus Joachim Lehmann, den Namenspatron des neuen Rechnergebäudes.

Die TU Dresden ist außerdem einer der ganz großen Gewinner der vom Bund und den Ländern ausgeschriebenen Exzellenzinitiative 2012 mit einer Graduiertenschule, zwei Excellenzclustern und einem besonders lukrativen Zukunftsprojekt mit einer bewilligten Gesamtförderung von 135 Millionen Euro.

Rund 400 Millionen Euro steckt der Bund laut Prof. Wanka in der laufenden Förderperiode bis 2017 in den Bereich HPC, davon 300 Millionen in Gebäude und Forschungsanlagen – das meiste davon dürfte bereits verbraucht oder verplant sein. Ob das Konzept der jeweils auf fünf Jahre befristeten Exzellenz-Förderung nach 2017 fortgeführt wird, ist nach ihren Worten indes offen. Wie es weitergehen soll, dafür wurde eine Expertenkommission mit Beteiligung internationaler Wissenschaftler berufen. Weniger Geld soll es hernach für Exzellenzforschung (derzeitiges Budget 2,7 Milliarden Euro) nicht geben, aber womöglich ein anderes Konzept, unter anderem mit unbefristeten Förderungen.

Spannend wird für viele eher eine andere Rangfolge als die oben genannte Top-70-Liste sein, nämlich wie sich der von Bull mit direkter Warmwasserkühlung aufgebaute Rechner im deutschen und internationalen Vergleich so schlägt. Ob er mit seinen nunmehr rund 1,6 PFlops theoretischer Spitzenleistung unter die Top-70 der weltweiten Supercomputer wird einziehen können? Das wird man Anfang Juli wissen, wenn zur ISC'15 in Frankfurt am Main die nächste Top500-Liste der Supercomputer veröffentlicht werden wird. In der aktuellen Liste vom November 2014 dürfte er irgendwo um Platz 50 herum rangieren.

Im bundesdeutschen Vergleich gibt es die ganz großen drei (JSC Jülich, HLR Stuttgart und das Leibniz-Rechenzentrum in München/Garching), die sich im Gauss-Centre for Supercomputing zusammengeschlossen haben, um so im europäischen Konzert PRACE mit einer gewichtigen Stimme sprechen zu können.

### Ligafragen

Hinter der ersten kommt die zweite Liga, und zwar mit 16 Zentren, die sich auf Initiative von Prof. Nagel als Gauß-Allianz konstituiert haben. Ganz vorne rangiert hier der Rechner des Rechenzentrums Garching (RZG) der Max-Planck-Gesellschaft, das sich nur einen Steinwurf weit weg vom Leibniz-Rechenzentrum befindet. Dahinter folgen Konrad und Gottfried des norddeutschen Verbunds für Hoch- und Höchstleistungsrechnen (HLRN) in Berlin und Hannover. Wenn man die zusammenrechnet, liegt das HLRN vorne. An den beiden Rechnern sind aber gleich sieben Bundesländer beteiligt – in diesem Rahmen steht der neue Rechner in der sächsischen Hauptstadt für ein einzelnes Bundesland mit seinen 1,6 PFlops recht gut da. Er befindet sich vermutlich sogar auf einem Aufstiegsplatz, denn der Wissenschaftsrat hat erst Ende April der Bundesregierung empfohlen,

**Modisch in schwarz gekleidet: ZIH-Mitarbeiter Danny Rotscher vor dem HRSK II mit den BL720-Blade-Racks von Bull**

„ein Nationales Hoch- und Höchstleistungsrechnen aufzubauen, das aus den bestehenden Zentren der Ebene 1 und einigen Zentren der Ebene 2 besteht." Und da hat die TU-Dresden als Exzellenz-Uni und als vorrangiger Vertreter der neuen Bundesländer sehr gute Chancen, vorne mit dabei zu sein.

Aber die oben genannten deutschen Spitzenteams legen nach und wollen die Zweitligisten auf Abstand halten. Schon in wenigen Wochen wird das Leibniz-Rechenzentrum, deren Direktor Prof. Dr. Arndt Bode ein freundliches Grußwort in Dresden vortrug, noch vor der ISC'15 die jetzt von Lenovo gelieferte Ausbaustufe 2 des SuperMUC offiziell in Betrieb nehmen. Das sind immerhin zusätzlich 6144 Haswell-Prozessoren (Xeon E5-2697v3), die zusammen mit den vorhandenen Sandy-Bridge-Prozessoren den Rechner auf rund 7 Petaflops (ohne Rechenbeschleuniger) hieven sollen. Damit dürfte der SuperMUC der schnellste Rechner in Deutschland sein, jedenfalls ein paar Monate lang. Dann nämlich will das HLRS in Stuttgart die nächste Ausbaustufe des Cray-XC40-Rechners Hornet in Betrieb nehmen und womöglich noch einige TFlops drauflegen.

Der aktuelle deutsche Tabellenführer JSC wird mit seinem JuQueen (5,9 PFlops Spitzenleistung) dann auf den dritten Platz verdrängt werden. In Jülich wartet man mit dem Nachfolger für den JuQueen erst einmal ab. IBMs BlueGene-Line läuft ja aus, OpenPower wäre die nächstliegende Option. Und so schickt das Jülicher SC mit dem Jureca erst einmal eine zweite Mannschaft ins Petaflops-Rennen. Der bei der russischen Firma T-Platforms in Auftrag gegebene Rechner soll in erster Ausbaustufe möglichst ebenfalls noch vor der ISC'15 in Betrieb gehen. Er ist wie der SuperMUC und der HRSK II von Bull mit direkter Wasserkühlung versehen. Seine im Vollausbau etwa 40 000 Haswell-Kerne sollen rund 1,8 PFlops abliefern.

## Energiefragen

Energieeffizienz ist einer der Schwerpunkte in Dresden, gefördert sowohl in dem Exzellenzcluster „Center for Advancing Electronics Dresden (cfaed)" als auch in dem Zukunftsprojekt „die synergetische Universität". Schon zuvor hatte das ZIH in einem vom BMBF geförderten Projekt „Cool Silicon" mit ortsansässigen Forschungseinrichtungen und Firmen zusammengearbeitet, um Basistechnologien für energieeffiziente Rechner zu erforschen.

**Der gemeinsame Druck auf den grünen Knopf, in korrekter Rangfolge von links: Ministerpräsident Stanislaw Tillich, Bundesministerin Prof. Dr. Wanka, der frisch gewählte Rektor der TU Dresden Prof. Dr. Hans Müller-Steinhagen und ZIH-Direktor Prof. Dr. Wolfgang E. Nagel**

Der neue Supercomputer soll nun nicht nur rechnen, sondern gleichzeitig als Forschungsobjekt dienen. Nicht nur, dass sein Warmwasser energieeffizient zur Heizung des geplanten Neubaus des benachbarten Physik-Instituts dienen soll. Man schrieb auch ein hochfiligranes Energiemonitoring mit bis zu 1000 Samples/s bei 2 Prozent Genauigkeit fürs Blade und 100 Samples/s für CPU und Speicher in die Ausschreibung hinein, zehnmal mehr, als es der ohnehin sehr hohe Anforderungskatalog der Energy Efficient HPC Working Group im höchsten Level 3 verlangt. Und schon dieses Level 3 schafft derzeit nur eine Handvoll Rechenzentren weltweit, darunter das LRZ in München, das CSCS in Lugano und der L-CSC-Rechner der GSI in Darmstadt.

Offenbar konnte nur Bull mit speziell entwickelten FPGAs im BL720-Blade diese Anforderung wie gewünscht erfüllen. Das FPGA hängt am Baseboard Management Controller, dessen Werte der Job-Manager Slurm auslesen und der laufenden Software zuordnen kann. Gemeinsam mit Bull gründete das ZIH die Forschungskooperation HDEEM: High Definition Energy Efficiency Monitoring.

Eines der Probleme, an dem man hier arbeitet, ist es, den Jitter zwischen den zahlreichen Messstellen so klein wie möglich zu kriegen. Ziel ist es, ein präzises Energieprofil als eine Art Fingerabdruck von verteilt laufenden Applikationen zu gewinnen. Mit diesen Kenntnissen, so hofft man, kann man dann Prozessorkonfigurationen, Taktfrequenzen, Kernzahlen, MPI, Load-Balancing und so weiter lastabhängig auf Energieeffizienz optimieren. Das ZIH entwickelt außerdem für das hauseigene Monitoring-Tool Vampir entsprechende Energieschnittstellen.

Aber nicht nur die Energieaufnahme wird filigran überwacht, sondern auch das I/O-System – schließlich steht das SK im Rechnernamen für Speicherkomplex. Diese Performance-Analyse allein liefert 1 GByte Daten pro Stunde.

## Taktfragen

Performance soll bei dem Monitoring natürlich nicht auf der Strecke bleiben. In die berichtete Gesamtleistung des HRSK II fließt mit 137 TFlops auch die vergleichsweise bescheidene Leistung der ersten Ausbaustufe mit ein, die mit Sandybridge- und Westmere-Prozessoren bestückt ist. Mit dieser Leistung konnte die TU Dresden noch nicht bei den Top500 mitspielen, aber nun hat man ja die zehnfache Performance. Die neuen schwarzen Racks in den beiden Compute-Inseln (zu je 612 Knoten) und der Throughput-Insel (232 Knoten) mit insgesamt 2908 Haswell-EP-Prozessoren (Xeon E5 2680v3) und 34896 Kernen erbringen allein bereits 1,2 PFlops Spitzenleistung – und zwar korrekt berechnet mit dem niedrigeren AVX-Basistakt von 2,1 GHz. Viele andere Rechenzentren schummeln hier und legen stattdessen den um 20 Prozent höheren Grundtakt ohne AVX zugrunde (so stehen sie dann leider auch in der Top500-Liste). Letztlich zählt aber die real gemessene Performance. Im Linpack müssten die drei Inseln bei guter Wasserkühlung schon um 1 PFlops herum schaffen – und da kommen dann noch die alte Ausbaustufe sowie insgesamt 108 GPU-Knoten mit 216 Nvida-Tesla-K80-Karten hinzu, die theoretisch 300 TFlops Spitzenleistung zuliefern.

(as@ct.de)

**Steckbrief HRSK II**

| Inseln | Knoten | Prozessoren/Knoten | CPU-Kerne insg. | Speicher | lok. SSD (Festplatte) | Kühlung |
|---|---|---|---|---|---|---|
| **Phase 2** | | | | | | |
| HPC 1 | 612 | 2 × Xeon E5-2680v3 12C | 14688 | 64 GByte | 128 GByte | Wasser |
| HPC 2 | 612 | 2 × Xeon E5-2680v3 12C | 14688 | 64 GByte | 128 GByte | Wasser |
| Throughput | 232 | 2 × Xeon E5-2680v3 12C | 5568 | 64–256 GByte | 128 GByte | Wasser |
| GPU | 64 | 2 × Xeon E5-2680v3, 2 × Nvidia Tesla K80 | 1536 | 64 GByte | 128 GByte | Luft |
| SMP | 2 | 4 × Xeon E7-4850v3 14C | 112 | 2 TByte | 128 GByte + 1 TB (Disk) | Luft |
| **Phase1** | | | | | | |
| Insel 1 | 270 | 2 × Xeon E5-2690 8C | 4320 | 32–128 GByte | 128 GByte | Wasser |
| Insel 3 | 180 | 2 × Xeon X5660 6C | 2160 | 48 GByte | 128 GByte | Luft |
| Insel 2 (GPU) | 44 | 2 × Xeon E5-2450v2, 2 × Nvidia Tesla K20 | 704 | 32 GByte | 128 GByte | Luft |
| SMP | 2 | 4 × Xeon E5-4650L 8C | 64 | 1 TByte | 128 GByte + 1 TB (Disk) | Luft |
| Dazu 6,6 PByte Festplatten, 40 TByte SSD, InfiniBand FDR, 4 × 10 GbE zum Campus | | | | | | |