*Hintergrund*

---

**28.06.2006 11:02**


## Alexandra Kleijn


# *MareNostrum - Der Supercomputer in der Kirche*

## HPC mit Linux

**HPC-Systeme werden heutzutage oft als Cluster gebaut. Dabei arbeiten viele im Prinzip eigenständige Rechner an der Lösung komplexer Probleme. Ein Linux-Cluster verrichtet in einer spanischen Kapelle seine Dienste für Wissenschaft und Wirtschaft.**

**Die Kapelle auf dem Universitätsgelände in Barcelona beherbergt den Supercomputer MareNostrum**

In einer kleinen Kapelle auf dem Gelände der polytechnischen Universität in Barcelona verbirgt sich eines der leistungsfähigsten Computersysteme der Welt: MareNostrum. Der Name - *unser Meer* - geht auf die von den Römern geprägte Bezeichnung für das Mittelmeer zurück und soll hier im übertragenem Sinne für die zentrale Rolle des Systems für die Wissenschaft stehen.

Auch wenn er sich inzwischen nicht mehr als schnellster Supercomputer Europas bezeichnen darf, so bleibt MareNostrum ein Paradebeispiel dafür, wie erfolgreich sich heutzutage aus Standard-Hardware und freier Software Hochleistungsrechner bauen lassen. Supercomputer, in den 90er Jahren üblicherweise Vektorrechner mit wenigen, dafür aber extrem leistungsfähigen und sehr teuren Prozessoren und die exklusive Domäne einiger wenigen Hardware-Hersteller, kommen inzwischen meist in einem anderen Gewand daher: als Cluster. Das sind Verbünde aus grundsätzlich autonomen Rechnern, die über Hochgeschwindigkeitsnetze miteinander verbunden sind und gemeinsam komplexe Rechenaufgaben bewältigen.

## Das Rechenzentrum CNS

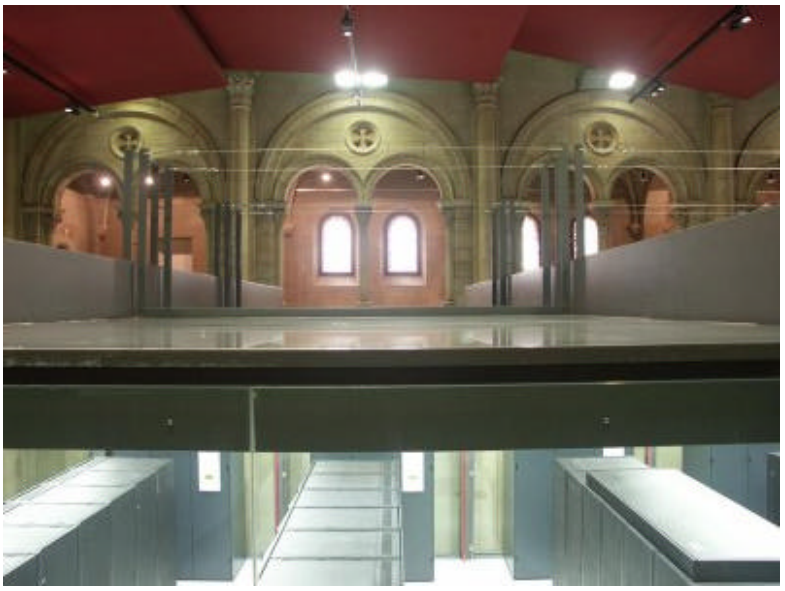
Um eben einen solchen High-Performance-Cluster handelt es sich bei dem Supercomputer in dem ehemaligen Gotteshaus auf dem Universitätscampus. Das System bildet das Herzstück des Centro Nacional de Supercomputación (in der internationalen Literatur auch als Barcelona Supercomputing Center (BSC) bekannt), des 2005 gegründeten spanischen Zentrums für Supercomputer in Barcelona. Das CNS resultierte aus einer gemeinsamen Initiative des spanischen Bildungsministeriums, der katalanischen Regierung (Generalitat de Catalunya) und der polytechnischen Universität Barcelona (UPC) und hat als Ziel, Wissenschaftlern moderne Informationstechnologie zur Verfügung zu stellen. Es gehört zum ebenfalls am UPC beheimateten European Center of Parallelism of Barcelona (CEPBA), dem spanischen Forschungszentrum für Clustertechnologien.

Ein besonderes Augenmerk des CNS liegt auf die Themen Rechner-Architekturen und *Deep Computing*. Dazu gehört Forschung in Bereichen wie Betriebssystemen, Programmiermodellen, Grid Computing, Performance-Analyse und der Anwendungsvirtualisierung. Das Institut möchte sich jedoch explizit nicht nur als IT-Forschungszentrum verstanden wissen, sondern forscht auch intensiv in anderen Bereichen wie Bio- und Geowissenschaften. Das Zentrum führt nicht nur eigene Forschungsaktivitäten aus, sondern stellt auch anderen Wissenschaftlern und Unternehmen aus der Wirtschaft die Rechenkapazitäten des Supercomputers für ihre Projekte zur Verfügung.

Über Projekt-Anträge entscheidet ein CNS-Komitee, in dem auch ein Experte für Supercomputer ausserhalb des CNS Mitglied ist. Dem Gremium steht ein so genanntes Experten-Panel zur Seite, das aus prominenten Wissenschaftlern in den Fachbereichen Geo- und Bio-wissenschaften, Medizin, Physik, Chemie und Ingenieurwissenschaften besteht.

## Aufbau und Architektur

Von der Idee bis zum Bau des Supercomputers verging weniger als ein Jahr. Die tatsächliche Bauzeit, in der die Kapelle ihr kirchliches Interieur preisgeben musste und auf 170 Quadratmetern Fläche eine doch eher profane Ausstattung bekam, betrug vier Monate. MareNostrum kann damit als Vorzeigeprojekt für den zügigen Bau eines Linux-Clusters gelten.



**Die Kapelle mit ihrem neuen Interieur.**

Das vom deutschen IBM-Labor in Böblingen entwickelte System setzt sich aus JS20-BladeCenters mit insgesamt 2406 Einzelblades zusammen, die jeweils mit zwei 64-Bit-PowerPC-Prozessoren bestückt sind. Ein Blade ist ein kompaktes Server-Modul, bestehend aus einer Hauptplatine, CPUs, Arbeitsspeicher und Festplatte(n). Es lässt sich in ein Chassis - hier das BladeCenter, das 14 Blades aufnehmen kann - einstecken und wird darüber mit Strom und Netzwerkanbindung versorgt. Sechs BladeCenter finden wiederum in einem Standard-19-Zoll-Rack Platz. Im Vergleich zu herkömmlichen 1-U-Servern - den sogenannten 19-Zoll-"Pizzaboxen" - ermöglichen Blades eine wesentlich kompaktere Infrastruktur und durch die gemeinsame Stromversorgung auch eine geringere Verlustleistung. Andere Vorzüge dieser Architektur sind die gute Skalierbarkeit und die einfache Austauschbarkeit einzelner Blades während des Betriebs dank der HotSwap-Fähigkeit.

Das gesamte System an diesem einzigartigen Ort umfasst 42 Racks, die in einer 5x8+2-Anordung von einem Glaskasten umgeben werden. 29 Racks beherbergen die Rechenknoten. In sieben Racks verwaltet ein ebenfalls aus dem Hause IBM stammender TotalStorage DS4100-Server 236 TeraBytes an Festplattenspeicher. Die Kommunikation der einzelnen CPUs untereinander und mit dem Festplattenspeicher erfolgt über Myrinet, ein Hochgeschwindigkeitsnetz mit deutlich weniger Overhead als Ethernet. Die Switches dafür nehmen noch mal vier komplette Racks in Beschlag. In einem zusätzlichen Serverschrank kommt die Gigabit-Ethernet-Verkabelung zusammen. Es gibt zwei voneinander getrennte Ethernet-Netzwerke: ein herkömmliches 10/100MBit-Wartungsnetz für das Systemmanagement per SNMP und ein Gigabit-Netzwerk ,über das die Blades gebootet, die System-Images verteilt und die Storage-Server angesprochen

werden. Die strikte Trennung in Service- und Produktivnetz stellt sicher, dass Verwaltung und Anwendungen sich nicht ins Gehege kommen.

Im letzten Rack befindet sich das Zentralhirn des Systems: Hier ermöglicht eine Management-Konsole die Verwaltung und finden auch die Switches für die Außen-Anbindung einen Platz.

MareNostrum punktet mit einem relativ geringen Energieverbrauch und weniger Wärmeentwicklung als viele andere Supercomputer-Anlagen. Das Gesamtsystem verbraucht im normalen Betrieb etwa 520 kW, also pro Stunde etwa so viel wie ein durchschnittlicher Zwei-Personen-Haushalt in zwei bis drei Monaten. Aktuelle PC-Systeme geben sich im Vergleich dazu mit einem Stromverbrauch zwischen 0,15 und 0,3 kW recht genügsam.

Der Cluster ruht auf einem "doppelten Boden", 70 cm über dem eigentlichen Boden der Kapelle. Der Zwischenraum beherbergt die Stromzuleitungen und die Netzwerkverkabelung und spielt eine wichtige Rolle bei der Raumkühlung. In vier externe Klimatisierungseinheiten, eine davon redundant, wird Wasser eingespeist, das diese auf acht bis neun Grad Celsius herunterkühlen und dann an Wassertanks mit einem Gesamtvolumen von 25.000 l weiterleiten. Eine mit der Notstromversorgung verbundene Pumpe befördert das kalte Wasser zu den zehn internen Klimatierungseinheiten (auch hier sind zwei redundant). Die Wassertanks dienen als Puffer und sorgen dafür, dass das System auch bei einem Stromausfall bis zu drei Stunden auf normaler Betriebstemperatur bleibt. Die internen Units fangen die um die 28 Grad warme Raumluft ein, generieren im Boden-Zwischenraum 17 Grad kalte Luft und leiten diese an die Racks weiter zur Kühlung der Blades.

IBM setzt für seine JS20-Blades auf herkömmliche PowerPC970FX-Prozessoren, die mit 2,2 GHz getaktet sind. Die gleichen CPUs verrichten in Apples G5-Desktopsystemen ihre Dienste. Mit 4GByte RAM pro Blade verfügt das System über insgesamt 9,6 TeraByte Arbeitsspeicher. Die theoretische maximale Rechenleistung des spanischen Hochleistungsrechners liegt bei 43,35 TeraFlops. Flops, oder auch Flop/S, bezeichnet die Anzahl an Gleitkomma-Berechungen, die ein System pro Sekunde ausführen kann (FLoating point Operations Per Second) und ist die Standard-Maßeinheit, in der die Leistung von Supercomputern bei wissenschaftlichen Berechungen mit vielen Gleitkomma-Operationen ausgedrückt wird. Zum Vergleich: ein herkömmlicher Desktop-PC mit Pentium 4 oder Athlon 64-Prozessor schafft einige GigaFlops; der momentan schnellste Computer der Welt, ein BlueGene/L-System von IBM im Lawrence Livermore National Laboratory (LLNL) des amerikanischen Ministeriums für Energie, bringt es auf 367 TeraFlops.

Die Top500-Liste, die zweimal im Jahr neu erstellt wird, greift für ihre Platzvergabe auf eine speziell für Hochleistungsrechner angepasste Version des Linpack-Benchmarks zurück. Mit diesem Testprogramm müssen die Systeme ihre Fähigkeiten beim Lösen von umfangreichen Matrizenberechungen in Form linearer Gleichungen unter Beweis stellen. Das Ergebnis, der so genannte Rmax-Wert,

bildet die Grundlage für die Rangordnung, liegt aber meist deutlich unter der theoretisch maximal erreichbaren Rechenleistung, dem Rpeak-Wert. Mit einem Rmax-Wert von 27,9 TeraFlops schaffte es MareNostrum in der letzten Top500, die im November 2005 veröffentlicht wurde, auf Platz 8.

Kritiker des Linpack-Rankings weisen auf die fehlende Berücksichtigung wichtiger Faktoren wie Speicherbandbreite, I/O-Leistung und Latenzzeit hin, wodurch sich die Ergebnisse der einzelnen Systeme nicht wirklich objektiv miteinander vergleichen lassen würden.

Das Betriebssystem für den Hochleistungsrechner liefert Softwarehersteller Novell. Auf dem System läuft Suse Linux Enterprise Server (SLES) 9 in der Version für PowerPC mit einem Standard-Kernel aus der 2.6-Reihe. Einige Module wurden jedoch vom MareNostrum-Team leicht modifiziert, um dem nicht-flüchtigen Speicher (NVRAM-Non Volatile Random Access Memory) mehr Informationen entlocken zu können, die zu Diagnosezwecken eingesetzt werden. Ein Kernel-Patch implementiert zudem das Performance API (PAPI), ein plattformübergreifendes, standardisiertes Interface für den Zugriff auf die im Prozessor integrierten Hardware-Counter. Damit lässt sich die Leistung einer Anwendung oder auch eines ganzen Systems überwachen und analysieren.

Das im IBM-Labor in Böblingen entwickelte Verfahren Diskless Image Management (DIM) sorgt für die Verteilung des Betriebssystems an die einzelnen Rechnerknoten, die mittels des Bootstrap-Protokolls BOOTP über das Netzwerk zum Leben erweckt werden. Dabei wird das root-Filesystem der Blades, das auf den Storage-Servern vorgehalten wird, über das Netzwerk-Dateisystem NFS gemountet. Die Blades verfügen zwar auch über eine eigene Festplatte, diese ist jedoch für spätere Anwendungen reserviert und bleibt zunächst leer.

Den Kontakt zur Aussenwelt stellt das CNS über zwei 1GBit-Verbindungen her. Eine davon führt zum ATM-Backbone Anella Cientifica („wissenschaftlicher Ring“), dem katalanischen Forschungsnetzwerk, und stellt die Standardverbindung zum CNS und MareNostrum dar. Die zweite Verbindung ist für das DEISA-Projekt reserviert und koppelt das CNS an das spanische Forschungsnetzwerk RedIRIS, das mit dem europäischen Forschungsnetzwerk Geant verbunden ist, zu dem auch das deutsche DFN gehört. Sobald Spanien den Umstieg auf Geants noch schnelleren Nachfolger Geant2 geschafft hat, ist ein Upgrade auf 10 GBit geplant. Wenn Projekte es erforderlich machen, lassen sich jederzeit zusätzliche 1GBit-Leitungen direkt zu MareNostrum einrichten.

## Clustertechnik

Für die Kommunikation der auf unterschiedlichen Knoten laufenden Jobteile einer Anwendung setzt MareNostrum eine Implementierung des offenen Message Passing Interface (MPI)-Protokolls ein. MPI stellt den de-facto-Standard für den Datenaustausch in Distributed-Memory-Umgebungen dar. Bei Programmen, die

MPI-Bibliotheken benutzen, verteilt sich die Arbeit über eine Reihe von Prozessen, die alle autonom ablaufen und keinen direkten Zugriff auf die Daten und Variabelen der anderen Prozesse haben. So genanntes Message Passing ermöglicht es den Prozessen, ihre Informationen untereinander auszutauschen.

Für die physikalische Kopplung der einzelnen Knoten ist die bei Clustern häufig zu findende Myrinet-Technik zuständig. Bei diesem ANSI-Standard, dessen Link- und Routingspezifikationen offengelegt sind, werden die Netzwerk-Interfaces der einzelnen Knoten mit speziellem Myrinet-Glasfaserkabel verbunden. Für sehr kurze Latenzzeiten sorgt die direkte Kommunikation der Karten-Firmware mit den Anwendungen und dem Netzwerk unter Umgehung des Betriebssystems. Myrinet schafft dadurch einen Datendurchsatz, der in der Nähe des theoretischen Maximums der physikalischen Schicht, also des Glasfaserkabels, liegt. Durch den Einsatz eines separaten Verwaltungsnetzwerkes auf Ethernet-Basis ist zudem gewährleistet, dass nicht die Netzwerküberwachung die Datenkommunikation zwischen den Blades ausbremst. Für die Cluster-Überwachung greifen die MareNostrum-Spezialisten auf die Open-Source-Management-Software Ganglia zurück.

Als Filesystem für ihren Cluster setzt das Team IBMs POSIX-konforme General Parallel File System (GPFS) ein, das es für Linux und AIX gibt. Unter diesem auf Parallelisierung, Hochverfügbarkeit und einfache Skalierbarkeit ausgelegten Shared-Disk-Filesystem können mehrere Knoten im Cluster gleichzeitig lesend und schreibend auf eine Datei zugreifen, ohne dass diese gesperrt werden muss. Dazu werden die Datenblöcke einer Datei über mehrere Platten verteilt abgelegt.

## Anwendungen

Computer-Cluster wie MareNostrum spielen Ihre Stärken bei der Parallelisierung aus. Dabei wird ein vielschichtiges Problem in kleinere Teilaufgaben zerlegt, die die verschiedenen Knoten des Clusters relativ autonom und simultan abarbeiten. Eine besondere Herausforderung für die Parallelisierung stellt die Skalierbarkeit dar. Bei steigender Prozessorzahl nehmen auch Verwaltungs-und Kommunikationsoverhead im Netzwerk zu. Mehr Prozessoren haben also nicht unbedingt die gewünschte Leistungssteigerung des Gesamtsystems zur Folge. Hochgeschwindigkeitstechniken wie Myrinet und MPI, die speziell für Cluster-Systeme entwickelt wurden, versuchen, den Overhead in Zaun zu halten, doch beliebig erweiterbar sind solche Rechnerverbunde in der Praxis nicht.

Die Systemarchitektur eignet sich gut für die Auswertung von riesigen Datenmengen, wie sie zum Beispiel in der Klimaforschung und der Meteorologie anfallen. Projekte auf MareNostrum beschäftigen sich mit Wetter- und Luftverschmutzungsvorhersagen und mit der Modellierung von Klimaveränderungen in Europa. Auch bei komplexen Sachverhalten, die sich sonst nicht, oder nur mit großem Aufwand, in der Realität abbilden lassen, zeigen Cluster in Simulationsaufgaben, was sie wert sind. Ein prominentes Beispiel hierfür ist die

Proteinfaltung, bei der der Zusammenhang zwischen nicht „richtig“ gefalteten Proteinen und dem Entstehen von Krankheiten wie Krebs und Alzheimer erforscht wird.

Nicht nur für rein wissenschaftliche Projekte steht der katalanische Supercomputer zur Verfügung. Forscher teilen die Rechenleistung mit Unternehmen aus Industrie und Wirtschaft, die auf dem System Anwendungen in Bereichen wie Pharmazeutik, Finanzen und Aeronautik laufen lassen.

## Ausblick

Das MareNostrum-Team experimentiert derzeit mit dem jüngsten Spross in IBMs Blade-Familie: dem Blade-Server mit Cell-Prozessor, dessen offizielle Markteinführung für das dritte Quartal dieses Jahres vorgesehen ist. Cell-CPUs beherbergen mit einem PowerPC-Kern und acht zusätzlichen digitalen Signalprozessorkernen (DSPs) insgesamt neun Prozessorkerne. Der PowerPC-Kern kümmert sich um die Prozessverwaltung und die Datenverteilung und gibt den DSPs ihren Anweisungen. Diese können sich ganz ihrer Hauptaufgabe, komplexen Vektorberechnungen, widmen.

Cell-Prozessoren sind vor allem in grafikintensiven Bereichen wie 3D-Rendering oder Mustererkennung in ihrem Element. IBM hat den Cell in Zusammenarbeit mit Sony und Toshiba entwickelt. Gemeinhin dürfte er auch als das Herzstück von Sonys geplante Spielekonsole Playstation 3 bekannt sein. Jeder Bladeserver ist mit zwei Cell-CPUs bestückt, die sich 1GByte gemeinsames DRAM teilen. Im Unterschied zum Produktivsystem läuft auf den Cell-Blades im Testbetrieb nicht Suse Linux Enterprise, sondern der Red Hat-Ableger Fedora Core für PowerPC.

Als schnellster Supercomputer in Europa wurde MareNostrum inzwischen überholt. So strebten der im März dieses Jahres im Forschungzentrum Jülich eingeweihte Blue-Gene-Superrechner von IBM und Tera-10, ein Itanium2-basiertes SMP-Cluster von Bull , das bei der französischen Kommission für Atomenergie (CEA) in Bruyeres-le-Chatel steht, an ihm vorbei.

Die neue Top500-Liste wird im Rahmen der Internationalen Supercomputer-Konferenz ISC, der vom 27. bis zum 30. Juni in Dresden stattfindet, präsentiert