



¹ Erasmus Computing Grid & Hogeschool Rotterdam, Instituut voor Communicatie, Media en Informatietechnologie, G.J. de Jonghweg 4-6, 3015 GG Rotterdam.
Email: l.v.de.zeeuw@hro.nl

² Erasmus Computing Grid & Erasmus Medisch Centrum, Instituut voor Celbiologie en Genetica Dr. Molewaterplein 50, 3015 GE Rotterdam.
Email: t.a.knoch@erasmusmc.nl

³ Biophysical Genomics, Kirchoff Institute of Physics, University of Heidelberg, Im Neuenheimer Feld 227, D-69120 Heidelberg, Duitsland.
Email: ta.knoch@taknoch.org

⁴ Hogeschool Rotterdam, Instituut voor Communicatie, Media en Informatietechnologie, G.J. de Jonghweg 4-6, 3015 GG, Rotterdam.
Email: bervd@hro.nl

⁵ TU Delft, B2-300, TBM-building, Jaffalaan 5, 2628 BX Delft
Email: j.vandenberg@tudelft.nl

⁶ Erasmus Medisch Centrum, Instituut voor Celbiologie en Genetica, Dr. Molewaterplein 50, 3015 GE Rotterdam.
Email: f.grosveld@erasmusmc.nl

* Gelijkwaardige inbreng. postadres ^{1,2}

Erasmus Computing Grid Het bouwen van een 20 Tera-FLOPS Virtuele Supercomputer

Luc V. de Zeeuw ^{1*}

Tobias A. Knoch ^{2, 3*}

Jan van den Berg ^{4, 5}

Frank G. Grosveld ⁶

Samenvatting

Het Erasmus Medisch Centrum en de Hogeschool Rotterdam zijn in 2005 een samenwerking begonnen teneinde de ongeveer 95% onbenutte reken capaciteit van hun computers beschikbaar te maken voor onderzoek en onderwijs. Deze samenwerking heeft geleid tot het Erasmus Computing GRID (ECG), een virtuele supercomputer met na voltooiing een reken capaciteit van 20 Teraflops. Dit artikel schetst enige achtergronden van grid computing, beschrijft een aantal toepassingen die mogelijk zijn met een grid infrastructuur en geeft de wijze weer waarop het ECG wordt vormgegeven. In het verlengde hiervan bevat het een pleidooi om grid computing binnen het onderwijs een betere basis te geven om op die manier vanuit het onderwijs een substantiële bijdrage te leveren aan het versterken van (rekenintensief) onderzoek.

Keywords

GRID computing, supercomputing, high-throughput distributed computing.



Luc de Zeeuw



Tobias Knoch

Overcapaciteit van pc's

Het Erasmus Medisch Centrum (Erasmus MC) en Hogeschool Rotterdam (HR) beschikken samen over zo'n 13000 pc's, verdeeld over al haar locaties in Rotterdam. Regelmatig staan deze pc's wél aan maar zijn niet in gebruik. Uit het oogpunt van energiebesparing zou het goed zijn als de pc's dan worden uitgeschakeld. Erg praktisch is dat niet omdat het opstarten tamelijk lang duurt. Ook voor de pc zelf is het niet erg goed: de levensduur van de pc neemt wat af. Gebruikers kiezen er daarom vaak voor de pc de gehele werkdag of voor altijd aan te laten staan. Vele organisaties gebruiken hun pc's slechts voor 25% van de tijd en dan meestal maar gedeeltelijk: de Pentium 4 CPU van de pc waar op dit moment deze tekst op wordt getypt wordt voor maximaal 4% belast. Het zou meer rendabel zijn als we de beschikbare pc's écht aan het werk zouden zetten en op deze wijze een bijdrage zouden leveren aan het oplossen van relevante (rekenintensieve) maatschappelijke problemen. De meeste in de toekomst benodigde computercapaciteit is in principe al beschikbaar op miljoenen desktop computers bij particulieren en tal van

bedrijven en organisaties. Adequate 'resource management' van de bestaande computer infrastructuur zou kunnen helpen om de 'total cost of ownership' van pc's te verlagen: computer resources zijn immers in principe verkoopbaar.

Grid computing

Basisidee

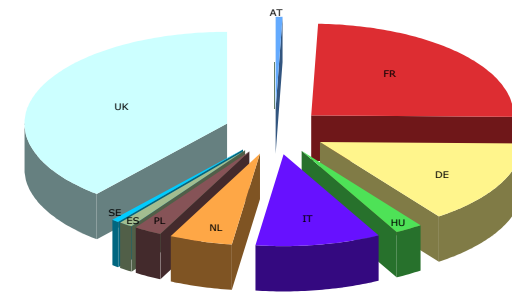
Zo gemakkelijk als elektrische stroom beschikbaar vandaag de dag is, zo eenvoudig zou ook de rekenkracht en randapparatuur van computers te ontsluiten moeten zijn. In 1998 hebben Ian Foster en Carl Kesselman dit idee uitgewerkt en hier aan de naam 'grid computing' (I) gegeven. Ter toelichting: het Engelse woord voor elektriciteitsnet is 'power grid'.

De enorme groei van het internet heeft in de afgelopen jaren tot grote snelheidsverbeteringen geleid in de datacommunicatie. Als gevolg hiervan bestaan er voor het ontsluiten van computer resources vanuit het oogpunt van datacommunicatie geen principiële belemmeringen meer en kan de visie van Foster en Kesselman in principe nu realiteit worden.

Doekle Terpstra
keynotespeaker



Country	million EUR
AT	2,7
FR	98,3
DE	58,8
HU	8,6
IT	40,7
NL	20,0
PL	9,1
ES	4,5
SE	2,3
UK	155,0
Total	400,0



Grid Research Funding (2002-2006)
in the 10 GridCoord Countries
Total ≈ 400 million EUR

Figuur 1

EU subsidies 2002-2006 (II)

Grid computing wordt gezien als een belangrijke nieuwe toepassing op ICT-gebied. Dat ook Europa het belang van grid computing inziet blijkt o.a. uit de subsidie die voor grid research projecten wordt uitgegeven. (Fig. 1).

Virtuele organisaties en 'resource sharing'

Grid computing gaat over meer dan alleen het ontsluiten van rekencapaciteit. Grid computing gaat over 'resource sharing', over het delen van computer capaciteit, het toegankelijk maken van meetapparatuur (sensor grids), het uitwisselen van bestanden, over virtuele organisaties en over het samenwerken van mensen die hun kennis delen. Sommige auteurs zien grid computing als de 'Next Generation Internet'. In ieder geval is duidelijk dat grid computing een enorme invloed zal hebben op de manier waarop rekenintensieve problemen zullen worden benaderd.

Erasmus Computing Grid

Bij het Erasmus MC en het Instituut voor Communicatie, Media en Informatietechnologie (CMI) van de HR is enige jaren terug het belang van grid computing erkend en is

men begonnen om de concepten rond grid computing in onderwijs en onderzoek een plaats te geven. Zo zijn er sindsdien verschillende (afstudeer-)projecten bij zowel het CMI als het Instituut voor Celbiologie en Genetica van het Erasmus MC uitgevoerd. Beide instellingen beschikken over een behoorlijk aantal computers en hebben belang bij een zo optimaal mogelijk gebruik ervan. Daarnaast hebben beide instellingen vanuit hun bedrijfsdoelstellingen (onderzoek en onderwijs) een helder doel voor ogen voor het gebruik van een computing grid. Een realistische infrastructuur is nodig om deze uitgangspunten in de praktijk te brengen.

Voor het verder welslagen van het initiatief is een projectorganisatie gerealiseerd met twee projectleiders en een stuurgroep, met vertegenwoordigers uit beide organisaties. Het ECG wordt nu bedrijfsmatig gerund vanuit het 'ECG office' (ECGO). Het ECGO, nu nog een werkverband, zal naar verwachting binnenkort in een stichting worden ondergebracht.

Na de zomer van 2007 zijn er bij het Erasmus

MC én de HR 7000 hosts beschikbaar voor het ECG, waaronder een aanzienlijk aantal 'dual core' machines. Daarmee is het ECG dan al één van de grootste autonome computing grids in de wereld. Als het ECG met alle Erasmus MC en HR pc's medio 2008 klaar zal zijn, is een 20 Teraflops virtuele supercomputer gerealiseerd.

Hierbij staat Flops (III) voor Floating Point Operations Per Second (drijvende komma operaties per seconde). Alle pc's van HR en het Erasmus MC samen zijn in staat om maximaal 20 Teraflops uit te voeren waarbij Tera staat voor 10^{12} . Ter vergelijking: de snelste 'general purpose' supercomputer (IV) in de wereld, de IBM Blue Gene/L (V), haalt met 131072 processoren 367 Teraflops (maar die kost dan ook vele miljoenen dollars).

Beschikbare middleware

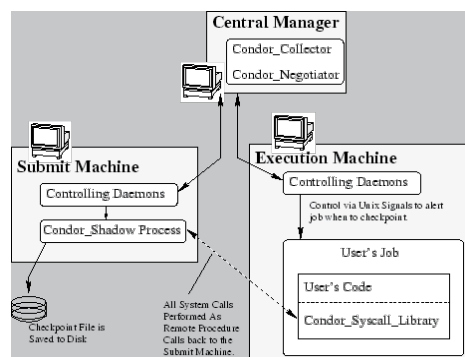
Het ECG is een computing grid. Voor het bouwen van een grid is uiteraard middleware nodig waarmee desktop resources beschikbaar komen. Er is in het public-domain onder andere keuze uit:

- ♦ **Boinc** (VI) - Geschikt voor het ontsluiten van reken capaciteit op privésystemen ten behoeve van specifieke researchprojecten.
- ♦ **Condor** (VII) - Geschikt voor het bouwen van autonome computing grids. Condor is de de facto standaard voor grid middleware.
- ♦ **Globus Toolkit** (VIII) - De de facto standaard. Geschikt voor 'grid computing research' (standaarden) en het koppelen van autonome grids.

De keuze voor het ECG is gevallen op Condor.

Condor is relatief eenvoudig te installeren. Condor beschikt over een uitstekende documentatie en goede tutorials wat de omgeving ook erg geschikt maakt om te worden toegepast om de concepten rond grid computing te onderwijzen. Dit neemt niet weg dat Condor een professioneel en zeer uitgebreid systeem is met enkele honderden configuratie opties en vele gebruiksmogelijkheden.

Er is verder een behoorlijke 'installed base' en een actieve gebruikersgroep. Verder zijn er enkele bedrijven die Condor support commercieel aanbieden. Condor is in wezen een geavanceerd batch systeem. De batch jobs worden gerund op individuele pc's. Condor zorgt voor het beheer van wachtrijen, planning van rekentaken, bepaling van prioriteiten en classificatie van de mogelijke hulpbronnen. (Fig. 2).



Figuur 2:
De basis architectuur van grid computing in termen van Condor (Bron: Condor Research Project at the University of Wisconsin-Madison).

Toepassingen

Onderwijs aan de HR

Vanuit de HR wordt het ECG gebruikt om het onderwijs in grid computing meer inhoud te geven. Studenten kunnen opgedane theorie over computersystemen in de praktijk brengen. Sommige projecten richten zich op de verbetering van de grid infrastructuur. Daarnaast wordt ook aan problemen rond het programmeren van een computing grid aandacht besteed.

Op dit moment lopen er twee projecten. Eén project gaat over het ABC vermoeden (IX), een wiskundig probleem uit de getaltheorie. Het tweede project gaat over het maken van een 'fractal movie' (voor dit begrip, zie video.google.com) beoogd. Een computing grid is uitermate geschikt om een fractal movie in relatief korte tijd te produceren. Een dergelijk project geeft studenten een goede ervaring met de mogelijkheden en beperkingen van een computing grid.

Behalve dat we nu beschikken over een zeer goede infrastructuur voor het onderwijs rond grid computing bij het CMI kunnen ook andere HR-instituten voor hun onderzoek gebruik gaan maken van het ECG. Het grid maakt op deze wijze multidisciplinaire HR-projecten mogelijk.

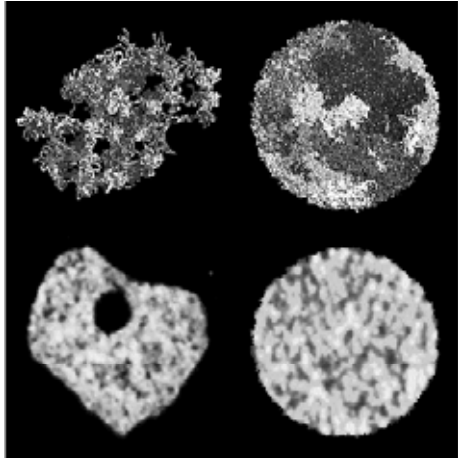
In het verlengde hiervan faciliteert het ECG het zogenoemde 'outside-in-inside-out' beleid van de HR waarbij de HR aansluiting zoekt bij dagelijkse praktijk in tal van bedrijven en instellingen in de regio.

Onderzoek bij het Erasmus MC

Het Erasmus MC zal in tal van projecten gebruik gaan maken van de grid infrastructuur

ten behoeve van onderzoek en patiëntdiagnostiek. We geven hiervan enige voorbeelden:

- ♦ In het kader van het Erasmus Rotterdam Gezondheid en Ouderen (ERGO) onderzoek worden 15.000 gezonde mensen gevolgd gedurende een lange reeks van jaren. Tweemaal per jaar wordt een MRI-scan van de hersenen van deze mensen gemaakt en vergeleken met eerdere scans. Deze taak is bijzonder rekenintensief. Een standaard pc is met het vergelijken van één scan circa 7 uur bezig. Per persoon moeten er 20 scans worden vergeleken. Voor de analyse van deze scans wordt nu het ECG ingezet om zo inzicht te krijgen in aandoeningen zoals bijvoorbeeld Alzheimer.
- ♦ Een andere Erasmus MC-afdeling wil door middel van simulaties inzicht krijgen in de verspreiding van besmettelijke ziekten, bijvoorbeeld influenza.
- ♦ Bij de afdeling genetica probeert men meer te weten te komen over de structuur van DNA. Het menselijk DNA (X) is te beschouwen als een lange codestring van zo'n 3.500.000.000 karakters waarbij elk karakter wordt voorgesteld als een A, C, G of T. Per persoon verschilt deze codestring. Met behulp van het ECG probeert men de structuur van het DNA beter te begrijpen. (Fig. 3).



Figuur 3
Het gebruik van grid computing in genoom onderzoek: het computer model van een enkel chromosoom (links boven) en een complete celkern (rechts boven); gesimuleerd microscoop beeld van het model van een complete celkern (rechts onder); microscoop beeld van een echte celkern in vivo (links onder).

Andere toepassingen

Een computing grid kent uiteraard nog vele toepassingen buiten medische research. Hierbij kan gedacht worden aan de volgende voorbeelden:

- ♦ Auto-industrie: simulaties om auto's veiliger te maken.
- ♦ Architectuur: levensecht lijkende simulaties waarbij door een gebouw kan worden gelopen zonder het eerst te bouwen.
- ♦ Entertainment-industrie: geavanceerde 'rendering' om digitale karakters echter te maken.

- ♦ Financiële industrie: Monte Carlo-simulaties om nieuwe zakelijke kansen te ontdekken.
- ♦ Klimaatonderzoek: het door middel van simulaties vergelijken van verschillende scenario's voor de opwarming van de aarde.

Toegankelijkheid voor studenten

Studenten van de HR kunnen via de bestaande computing grid infra-structuur ervaring opdoen met dit type computersysteem. Het Erasmus MC op hun beurt geeft de hbo-studenten van HR toegang tot haar onderzoek waarbij de opgedane ervaring nuttig kan worden gemaakt, veelal in samenwerking met studenten van de Erasmus Universiteit. Op deze wijze leren alle studenten gebruik te maken van de aanwezige grid infrastructuur voor het uitvoeren van relevant onderzoek, onderzoek dat tot voor kort niet gedaan kon worden. Bijgevolg profiteren zowel HR als het Erasmus MC van deze unieke samenwerking.

Conclusies

In dit artikel hebben we inzicht proberen te geven in hoe bij het Erasmus MC en de HR grid computing is opgepakt en wordt vormgegeven in aanvulling op de bestaande ICT-voorzieningen. De in allerlei opzichten verschillende organisaties, Erasmus MC en HR, hebben elkaar gevonden in dit initiatief en naar verwachting zullen nog diverse instellingen en bedrijven binnen en buiten Rotterdam (en mogelijk zelfs buiten Nederland) zich aansluiten en een bijdrage gaan leveren aan de ontwikkeling en het gebruik van het ECG'.

Grid computing maakt het mogelijk zeer rekenintensieve problemen op te lossen. In

allerlei domeinen komt dit soort van problemen voor zoals in de wiskunde, gezondheidswetenschappen, economie, technische wetenschappen, en meer. In dit artikel is hiervan een aantal concrete voorbeelden gegeven. In veel gevallen kan het oplossen van deze problemen als zeer maatschappelijk relevant worden gekenschetst.

Het moge verder duidelijk zijn dat we pas onze eerste schreden hebben gezet op het pad van de professionele ontwikkeling van het ECG. De toekomst is erop gericht dit pad verder te vervolgen. Dit houdt veel onderzoeks- en ontwikkelingswerk in teneinde in staat te zijn de beschikbare capaciteit optimaal te benutten. Verder zullen we andere organisaties moeten gaan wijzen op de unieke reken capaciteit die via het ECG beschikbaar is en nog zal komen. We zijn dan ook erg benieuwd naar welke toepassingen de komende jaren op het ECG zullen gaan draaien.

Ook is duidelijk geworden dat er de komende jaren een groep van grid computing specialisten opgeleid zal moeten worden om de diverse gebruikers op allerlei wijzen te faciliteren bij het oplossen van hun problemen met behulp van het ECG. De bestaande infrastructuur biedt reeds voldoende basisvoorzieningen om genoemde opleiding mogelijk te maken.

'Het Erasmus Computing Grid is niet zonder reden vernoemd naar Desiderius Erasmus, de grote Europese geleerde en netwerker: het laat de ambitie zien om de gestarte ECG grid computing activiteiten ver uit te breiden, tot over de grenzen van de regio Rotterdam.

Referenties

- I Ian Foster en Carl Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, (1998)
- II Grid Coord (European Commission), Grid Research In Europe, survey summary, (2006)
- III <http://en.wikipedia.org/wiki/FLOPS>
- IV <http://www.top500.org/list/2007/06/100>
- V http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html
- VI <http://boinc.berkeley.edu/>
- VII <http://www.cs.wisc.edu/condor/>
- VIII <http://www.globus.org/>
- IX <http://nl.wikipedia.org/wiki/ABC-vermoeden>
- X <http://nl.wikipedia.org/wiki/DNA>

