Frank den Hollander f.j.den.hollander@rug.nl
Kristien Piersma k.i.piersma@rug.nl

Fotografie: Gerhard Lugard

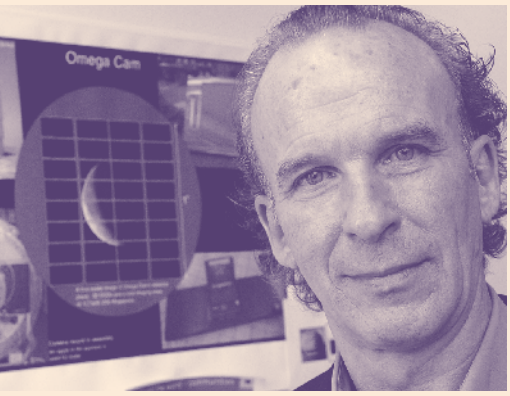
Nieuwe methode voor grootschalige dataopslag en verwerking

In het TARGET-project werken verschillende onderzoeksgroepen van de Rijksuniversiteit Groningen samen aan de oprichting van een science operation centrum voor grootschalige dataopslag en verwerking. Hoogleraar astronomische informatietechnologie prof. dr. Edwin Valentijn leidt de afdeling OmegaCEN en is bezig samen met drs. Robert Janz van het CIT het TARGET-project op te zetten.

het interview

Edwin Valentijn

Ik heb Sterrenkunde gestudeerd in Leiden en na mijn promotie heb ik vijf jaar in het buitenland gezeten, bij de ESO (European Southern Observatory) in München. Vervolgens ben ik naar het Kapteyn-instituut gegaan en heb de La Palma-sterrenwacht helpen opzetten. Dat was heel spannend, ik zat de helft van de tijd hier en de helft van de tijd in La Palma. Daarna heb ik acht jaar bij SRON gewerkt, Stichting Ruimteonderzoek Nederland, waar ik me vooral heb beziggehouden met de ISO-satelliet.
Uiteindelijk ben ik weer teruggegaan naar het Kapteyn-instituut waar we het OmegaCEN-datacentrum hebben opgericht. Dat is een onderneming van NOVA,



de nationale onderzoeksschool voor Astronomie, een van de zes toponderzoeksscholen die Nederland heeft.

Sterrenkunde is een vakgebied waarin met heel veel data wordt gewerkt en in het kader van OmegaCEN hebben we een informatiesysteem ontwikkeld waarmee dataverwerking, opslag, analyse en verspreiding inzichtelijk gemaakt kan worden. Met het TARGET-project moet dit systeem ook toegankelijk worden voor andere onderzoeksgroepen binnen de universiteit.

Wat is de historie van OmegaCEN?

We zijn begonnen met een camera, de OmegaCAM. Dat is een zeer geavanceerde camera met 32 chips (CCDs) en in totaal 256 megapixels. Eén foto is al een gigabyte. De camera maakt iedere twee minuten een foto. Als de telescoop van de camera klaar is, komt de camera in Paranal in Chili naast de VLT-telescopen te staan, de grote achtmetertelescopen van de ESO.

De OmegaCAM heeft een eigen *dedicated* telescoop van 2,6 meter, een zoektelescoop met een groothoekcamera. Het bijzondere is dat hij heel groot kijkt, maar hij heeft zo gigantisch veel pixels dat hij ook nog heel scherp kijkt. Je doet dus beide dingen tegelijk.

De gigabyte-images die iedere paar minuten worden gemaakt, acht tot tien uur per nacht, 300 nachten per jaar, en dat tien jaar lang, leveren honderden terabytes aan data op die verwerkt moet worden. Er moet een heleboel mee gebeuren: het moet onder andere schoongemaakt worden, de satellieten en vliegtuigen die langskomen moeten verwijderd worden. Voor het verwerken van de data zijn grote clusters nodig, een computer doet een minuut of drie over het verwerken van het beeld van een chip. Ga maar na, als je 32 chips hebt, drie maal 32 minuten, dan is ie al anderhalf uur bezig, terwijl je maar een paar minuten bezig bent geweest om de foto te nemen.

Daar is dus moderne technologie voor nodig, anders ga je op een gegeven moment achter jezelf aanrennen. Op parallelle clusters kunnen alle chips tegelijk worden bewerkt. Maar op een gegeven moment moet alles ook weer worden gesynchroniseerd. Bij OmegaCEN hebben we een systeem ontworpen voor de verwerking van die grote hoeveelheid data. Niet alleen voor het omgaan met zulke grote databestanden, maar ook voor het oplossen van administratieve problemen van gebruikers. Want iedere keer wordt de informatie bewerkt en soms moet de operatie over omdat er een betere methode is, een betere aanpak of een beter script of inzicht. Daar is administratie voor nodig. Als dat allemaal via verspreide werkstations gebeurt op verschillende plekken met verschillende mensen, dan is het overzicht binnen de kortste keren ver te zoeken. Met het systeem dat wij hebben gebouwd, kan dat allemaal worden bijgehouden.

Centraal vs. decentraal

Met dit project komen we eigenlijk weer in een nieuwe fase terecht. In de zeventiger jaren waren er alleen centrale computers en moest alles op grote centrale faciliteiten gebeuren, daarna kwam er een periode waarin alles werd gedecentraliseerd door de opkomst van krachtige werkstations. De individuele werkstations zijn nog steeds heel belangrijk, maar niet meer voldoende. Voor grote projecten met hele grote databestanden is er een grote behoefte om met centrale systemen te werken.

Het Centrum voor Informatie Technologie (CIT) van de RUG heeft een cluster van 200 nodes, waarvan wij zestien nodes hebben voor het verwerken van onze cameragegevens. En als we in de toekomst achterlopen in tijd doordat er meer data wordt waargenomen dan wij kunnen verwerken, dan kunnen we de boel opschalen naar meerdere nodes van het cluster. Die schaalbaarheid lukt niet met individuele werkstations.

Daarnaast moet de data worden opgeslagen. We praten over honderden terabytes aan data, dat moet ergens professioneel staan. Het moet veilig staan en dat vergt onderhoud. Dat zijn allemaal dingen die je graag weer centraal wilt regelen. Bovendien wil je dat gebruikers op ander locaties, ook in het buitenland, bij de data kunnen. Dat is in mijn ogen de kentering van werkstation-science naar e-science met een gecentraliseerde maar ook gedistribueerde infrastructuur. Die behoefte zie je op meerdere plekken binnen de universiteit.

Informatiesysteem

We sluiten ook aan bij GRID, de internationale beweging die computerkracht aan elkaar knoopt. GRID is heel mooi voor sommige toepassingen, maar als ik mijn honderdduizenden bestanden ergens neerzet, weet ik nog niet wat wat is. Naast opslag moet je ook nog weten wat je gedaan hebt.

Je hebt een informatiesysteem nodig dat je vertelt wat er is gebeurd met de data. Omdat zoiets er nog niet was, hebben we dat zelf ontworpen en gebouwd. Zo'n systeem koppelt processen en opslag aan informatie. Daar zijn we ontzettend ver in gegaan, zover als je kunt. Alles wat gedaan is, wordt bijgehouden; als je een product (een *target*) opvraagt, geeft het systeem aan of het nog steeds actueel is, of dat er sinds dat het uitgerekend is nieuwere methodes zijn bedacht. Of dat er nieuwe kalibraties zijn bedacht, bijvoorbeeld om de atmosfeer er af te trekken, of de gevoeligheid van de telescoop. Er is wel een dozijn kalibraties die moeten gebeuren, die veranderen in een research-omgeving. Het systeem kijkt dan of alle kalibraties nog steeds up-to-date zijn. Als dat niet zo is, krijg je een soort *roadmap* van wat er moet gebeuren. Door op een knop te druk-

> Sterren-kundigen zijn gek op veel data <

ken ga je naar het cluster waar het opnieuw wordt uitgerekend (*target-processing*). Het systeem is nu volop in werking en we hebben het gepubliceerd onder pakkende titels als: 'The Universe as a spreadsheet' en 'Chaining to the Universe'.

Het is in een groter kader ook heel erg gericht op de samenwerking tussen de verschillende onderzoeksgroepen. Het gaat vaak om kostbare projecten waar meerdere onderzoeksgroepen op verschillende plekken in Europa aan werken. Ze willen met elkaar samenwerken, maar hebben allemaal een heel klein beetje geld en maar een paar mensen om zo'n systeem te onderhouden. Door alles aan elkaar te knopen, kunnen die verschillende onderzoeksgroepen met elkaar samenwerken. Als iemand in München een bepaald stuk van de kalibraties heeft gedaan, bijvoorbeeld de atmosfeer gedurende een paar maanden in het syteem heeft bekeken en berekend, dan ziet iemand op een andere plek meteen dat dat is gebeurd. Daar kan hij dan van profiteren.

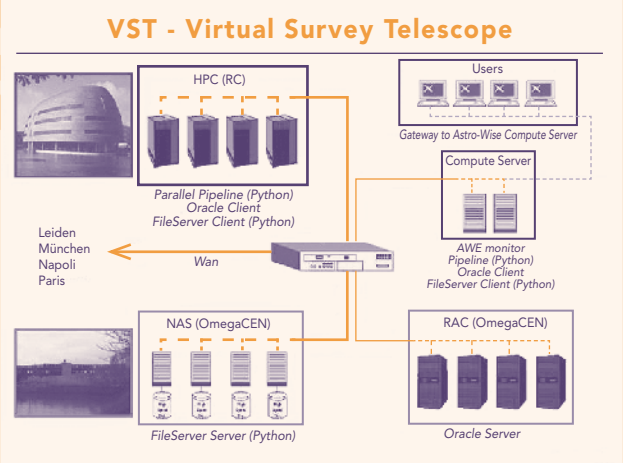
Dat doen wij dus in het kader van OmegaCAM en dat is uitgegroeid naar Astrowise, het Europese samenwerkingsverband waar ik Europees geld voor heb gekregen.

Wat is het product van de camera?

Het zijn individuele projecten, van klein tot groot. Het kan een project zijn van een nachtje maar het kan ook een project zijn van 200 nachten. Men moet inschrijven voor tijd die na een competitie wordt toegewezen. Als je die tijd hebt gekregen, wordt de data in Chili genomen en ontvang je vervolgens de ruwe plaatjes. Er zijn allerlei verschillende onderzoeken. Met zo'n groothoek kun je natuurlijk heel mooi zeldzame objecten vinden, maar je kunt de gegevens ook gebruiken voor statistische studies, bijvoorbeeld het in kaart brengen van de donkere materie door middel van vervorming van sterrenstelsels.

Je hebt een systeem waarbij enorm veel data moet worden verwerkt en waarin je in die zee van data gaat zoeken naar iets dat bijzonder is. Dat gebeurt door onderzoeksgroepen op allerlei verschillende plekken en promovendi die je niet wilt overladen met ICT-ellende.

We hebben een infrastructuur gemaakt om de opslag en de research aan elkaar te knopen door middel van een database en de verschillende databasegroepen ook nog eens aan elkaar te koppelen. De infrastructuur bestaat uit vier componenten: de opslag van de gegevens (i) gebeurt in de fileservers, die voor de verwerking naar het cluster (ii) van het CIT gaan. De gebruikers (iii), kunnen zich over de hele wereld bevinden en er is een database (iv). Die vier componenten zijn door glas aan elkaar verbonden. Het is een peer-to-peer-netwerk, dus de data kan alle kanten opgaan. Diezelfde infrastructuur hebben we gespiegeld naar Leiden, München, Napels en Parijs.



Als iemand in München iets doet, zien wij het ook meteen en vice versa.

De infrastructuur is er dus al, waar gaat het TARGET-project nog verder?

De Sterrenkunde loopt wellicht wat voorop omdat wij zo graag met zo ontzettend veel data werken. Maar andere groepen hebben natuurlijk dezelfde problematiek. Het idee was om dat te gaan poolen, naar elkaar over te laten gaan. Dat is best lastig vanwege de weerstand tegen centralisatie. Dat wordt over het algemeen als fnuikend gezien voor de creativiteit, en wellicht terecht- echter in onze ontwerpen hebben we zoveel mogelijk rekening willen houden met de veranderende inzichten van de onderzoeker.

Dat kun je alleen doorbreken door de onderzoeksgroepen te betrekken in het centrale proces. Dat is eigenlijk wat we met TARGET willen en doen. We hebben een projectgroep die bestaat uit verschillende afdelingen die de hardware-expertise van het CIT met elkaar deelt. Ook de actieve en betrokken rol van database administrators is heel belangrijk voor dit project.

Universiteitsbreed

We hebben binnen de universiteit rondgekeken welke groepen met een vergelijkbare problematiek te maken hebben. Met de groep van Kunstmatige Intelligentie hebben we het Cultural Heritage-project als pilot gekozen. Dit betreft het automatisch herkennen van handgeschreven tekst. Daar zijn we nu ongeveer een half jaar mee bezig en dat is inmiddels operationeel.

Het gaat om de archieven van het Kabinet van de Koningin, zo'n duizend boeken, allemaal door een klerk volgeschreven met een prachtig handschrift, maar lastig te ontcijferen en tijdrovend om iets op te zoeken. De KI-groep probeert dat met behulp van de computer te doen. Ze hebben een methodiek ontwikkeld om teksten te ontcijferen en uiteindelijk een Google van handgeschreven teksten te maken.

Als je het hele proces bekijkt: er wordt een digitale *image* van een scan gemaakt, de digitale image moet uitgelijnd worden, er moeten vlekjes uitgehaald worden, er moeten allerlei bewerkingen uitgevoerd worden. Het lijkt eigenlijk heel erg op Sterrenkunde. Er worden allerlei *events* uitgehaald. Bij hun is dat stukjes tekst, bij ons melkwegstelsels. De hele productiemethode lijkt erg op dat van Sterrenkunde. We hebben het daarom als pilotproject uitgekozen voor de TARGET-groep.

Ook vanuit de industrie is er belangstelling. We krijgen ondersteuning van Oracle voor de databases en Atos Origin Groningen doet ook mee. Zij hebben te maken met hetzelfde probleem: zij moeten ook met steeds meer data omgaan, informatie moet steeds langer bewaard worden. Probeer maar eens voor te stellen hoe dat in een bedrijf moet met allerlei verschillende mensen en verschillende systemen. Met terabytes aan data is het voor de industrie ook heel interessant hoe je informatiesystemen kunt onderhouden die met variabele code werken. Onze database houdt bij met welke code welk object is gemaakt, dat kan bijna tot in het oneindige worden doorgevoerd.

In welk stadium bevindt het TARGET-project zich op dit moment?

De TARGET-projectgroep is formeel sinds 1 mei opgericht, maar we zijn al zes maanden bezig met het opzetten van de pilot. De huidige projectgroep bestaat naast het CIT en OmegaCEN uit de Kunstmatige Intelligentie-groep en LOFAR doet ook mee. Het LOFAR-project heeft natuurlijk met dezelfde problematiek te maken: ontzettend veel data en een gedistribueerde gebruikersgemeenschap. Er zijn nog een vijftal groepen binnen de universiteit die ook meedoen, maar nog geen deel uitmaken van de huidige pilot, waaronder Alfa-informatica, het UMCG met het Lifelines-project en Bioinformatica. Wij kunnen pas aan de slag als onze regering richting gaat geven aan de grote gelden voor kennismaatschappij-onderzoek. Dat zou nog wel eens een paar maanden kunnen duren.

U bent behalve met besturen nu heel veel bezig met al die datastructuren, heeft u niet weer eens zin om gewoon door een sterrenkijker te kijken?

Ik heb natuurlijk m'n promovendi en verder nog een beperkt aantal onderwijstaken, ik geef het college 'Virtual observatories', virtuele waarneemtechnieken. Daarnaast is er best nog wel tijd voor wat onderzoek. Melkwegstelsels is mijn onderwerp, *morphological transitions*, hoe melkwegstelsels van morfologie veranderen. Dat gebeurt ergens heel ver in buitengebieden van clusters. Daar worden melkwegstelsels vervormd en evolueren ze van spiraalstelsels naar zogenaamde SO-stelsels. Dat is een heel zeldzaam proces, dat gebeurt heel kort en is heel moeilijk te vinden. Met een widefield-telescoop is het mogelijk hele grote gebieden af te zoeken, en onlangs vonden we een aantal stelsels in transitie die branden als een fakkel. Dat is mijn eigen onderzoek, maar het andere onderzoek, het ICT-stuk, het extreem linken van data-items, dat is ook een uitdaging en eigenlijk liggen die twee vakgebieden in elkaars verlengde: *the universe as a spreadsheet.*

Links

- Website OmegaCEN: **www.astro.rug.nl/~omegacen**
- De persoonlijke homepage van Edwin Valentijn: **www.astro.rug.nl/~valentyn**
- Astronomical Wide-field Imaging System for Europe (Astro-WISE), met mooie sterrenfoto's: **www.astro-wise.org**