Inrichting database heeft invloed op prestaties van cluster

De Clusters komen naar het bedrijfsleven

Teus Molenaar

Zij leverden het eerste Infiniband-cluster in Europa en het eerste AMD-Opteron cluster ter wereld. Het Hoofddorpse ClusterVision staat graag vooraan. De twee directeuren menen dat, na de wetenschappelijke wereld, nu het bedrijfsleven rijp is voor de hedendaagse supercomputers.

De grid- en cluster-onderneming uit Hoofddorp bestaat nu twee jaar, maar heeft al een aardige lijst met klanten weten op te bouwen. Zoals in Groot-Brittannië het Rutherford Appleton Laboratory, universiteiten in Sheffield, Manchester en Bristol en het Institute of Pharmaceutical Innovation in Bradford. In de Benelux prijken de universiteiten uit Leiden, Groningen, Eindhoven en Utrecht en de Vrije Universiteit Amsterdam op de klantenlijst, evenals TNO Instituut voor Toegepaste Aardwetenschappen, Cenaero (het Belgische centrum voor luchtvaartonderzoek), het Academisch Ziekenhuis Utrecht, NIKHEF en Delft Hydraulics. En in Duitsland het Max Planck Instituut.

Minder exotische omgevingen

Toepassingen zijn onder meer onderzoek naar nieuwe medicijnen, eigenschappen van elementaire deeltjes, medische studies, klimaatmodellen, scheikunde en mechanisch engineering. Een heel hoog academisch gehalte, dat wel. Maar de clusters gaan

> Beheerders die gewend zijn met Red Hat te werken kunnen vrij snel aan de slag

naar minder exotische omgevingen toe, zo is Matthijs van Leeuwens overtuiging. Hij is verkoopdirecteur bij ClusterVision. "We gaan ons de komende tijd juist meer richten op het bedrijfsleven." Technisch directeur Alex Ninaber valt hem bij: "De technologieën die nu worden ontwikkeld op de High Perfomance Computing (HPC)-platformen zijn over vijf tot zes jaar *mainstream*." Intel en AMD zien de HPC-omgeving als test voor hun nieuwe producten. "Als wij iets niet voor elkaar krijgen of bepaalde haperingen ontdekken, dan worden er meteen een paar ingenieurs vanuit Amerika overgevlogen om het probleem te verhelpen en tegelijkertijd na te gaan wat er aan de hand is," vertellen zij. Zo was ClusterVision de eerste die een AMD Opteron cluster bouwde. Bij de universiteit in Manchester gebruiken wetenschappers zeventig van deze 64-bits processoren tegelijkertijd om chemische reacties en complexe moleculaire systemen te simuleren.
En ClusterVision was in Europa de eerste die een supercomputer cluster bouwde op basis van de Infiniband-technologie. Dit netwerk-communicatieprotocol heeft een bandbreedte die bijna tien keer zo groot is als Gigabit Ethernet en *latency* die tien keer zo laag ligt. Dit cluster (met 66 Intel Xeon-processoren) staat te stampen in de universiteit van Utrecht en ondersteunt onder meer het onderzoek naar de evolutie van sterren.

"Geen wetenschap meer voor mij"

Dr. Alex Ninaber heeft chemie gestudeerd, dr. ir. Matthijs van Leeuwen civiele techniek. Beiden zijn gepromoveerd en hadden voor hun studie enorm veel rekenkracht nodig. 'CPU-junkies', zo omschrijven zij zichzelf in die tijd. Tijdens de promotie hebben ze een cluster gebouwd, omdat er toen gewoonweg geen te koop was. Of ze de wetenschap niet missen? "Niet echt," reageert Ninaber meteen. "Daar ben je tegenwoordig te veel bezig voorstellen te schrijven voor mogelijk onderzoek om geld los te krijgen en artikelen te schrijven om te laten zien dat je dat geld waard bent. Aan echt onderzoek kom je te weinig toe. Geen wetenschap meer voor mij." Van Leeuwen is minder resoluut in zijn oordeel: "Ach, af en toe mis ik het toch wel. Maar dit is toch wel leuker.

Foto Harry Otto

Matthijs van Leeuwen, verkoopdirecteur bij ClusterVision.

Clusters beginnen al wel door te sijpelen naar het bedrijfsleven, bijvoorbeeld bij Canaero. Daar wordt de giga-rekenkracht benut om luchtstromingen rond landingsgestel, romp en motoren te berekenen. Doel is nieuwe ontwerpen te kunnen maken die leiden tot stillere vliegtuigen. Het gaat hier over een groep van 85 servers en drie redundante controle- en opslagservers. Als een server het laat afweten, neemt een ander het werk onmiddellijk over. Met zijn 176 Intel Xeon 3.06 GHz-processoren, 176 Gigabyte werkgeheugen en 15 Terabyte opslagcapaciteit kan de computer een theoretische piek halen van maar liefst 1 miljard berekeningen per seconde. En natuurlijk is het onderzoek naar nieuwe medicijnen uiteindelijk ook een ondernemingsactiviteit.

Goedkoper

Dergelijke rekenwonders waren tot voor kort altijd supercomputers van illustere fabrikanten als Cray, IBM, Hitachi, Bull en Fujitsu. De apparaten gaan voor miljoenen euro's over de toonbank (en doen na hun levenscyclus dienst als ingewikkelde fontein, zoals de watergekoelde Cray bij Météo France in Toulouse) en de kopers dienen nog eens jaarlijks tien tot twintig procent van de aanschafprijs over te maken aan de fabrikant voor onderhoud. De grootste reden om tot een cluster over te gaan, volgens Van Leeuwen en Ninaber, is kostenbesparing. Een volledig dubbel uitgevoerd cluster met Myrinet voor de netwerkverbindingen biedt volgens hen dezelfde prestaties tegen een aanschafprijs die zes tot acht keer lager ligt dan die van een supercomputer (los van de jaarlijkse onderhoudsafdracht).
Een belangrijk overweging vindt Van Leeuwen de schaalbaarheid van een cluster. "Je kunt heel klein beginnen en servers toevoegen naarmate je daar behoefte aan hebt. Dat is erg kosteneffectief. Voor capaciteitsproblemen of kostenbesparing is een database-cluster het antwoord. Daarmee kun je veel kosteneffectiever werken dan met één grote machine."

Besturingssysteem

De parel van ClusterVision is, volgens de directeuren, diens besturingssysteem: het ClusterVisionOS. De onderneming bouwt Beowulf-clusters, een architectuur die zo langzamerhand een soortnaam is geworden. De computer-groepen worden gemaakt voor rekenintensieve taken, opslag of database-gebruik. Bij de laatste ligt het accent meer op de benaderbaarheid van de

Supercomputers

Voor een definitie van 'supercomputer' is een kijkje op de Free On-line Dictionary of Computing (Foldoc) nuttig: "Een brede term voor een van de snelste computers die tegenwoordig beschikbaar zijn. Dergelijke computers worden gewoonlijk gebruikt om getallen te kraken, zoals voor wetenschappelijke simulaties, (geanimeerde) grafische weergaven, analyses van geografische data (bijvoorbeeld bij petrochemisch bodemonderzoek), structuuranalyses, vloeistof-stromingen, natuurkunde, scheikunde, elektronisch ontwerp, onder-zoek naar kernenergie en meteorologie. De bekendste fabrikant van supercomputers is wellicht Cray Research."

Clusters van computers hebben zich nog niet opgewerkt in deze definitie. Ze komen al wel voor in de Top 500 Supercomputers die jaarlijks verschijnt. Aad van der Steen, hoofd van de HPC Group aan de Universiteit van Utrecht, is een van de samenstellers hiervan. De lijst van 2003 is de dertiende op rij. De meeste nieuwe systemen, zo meldt het rapport, zijn variaties op een groep van RISC-gebaseerde symmetrische multiprocessor-knooppunten die via een snel netwerk met elkaar zijn verbonden.

"De markt voor supercomputers is erg dynamisch en dat geldt speciaal voor Beowulf-clusters die de laatste jaren bijzonder snel in zwang zijn geraakt. Het aantal leveranciers dat voorgeconfigureerde clusters verkoopt is dientengevolge erg snel gegroeid. Daarom heb-ben we, althans voor dit rapport, besloten dergelijk configuraties niet op te nemen: de snelheid waarmee cluster-bedrijven en -systemen komen en gaan maakt dat haast onmogelijk." Er is overigens wel een sectie in het rapport waarin de cluster-karakteristieken en hun positie worden vergeleken met andere supercomputers.

Foto Harry Otto

Matthijs van Leeuwen: "We gaan ons de komende tijd juist meer richten op het bedrijfsleven."

gegevens. In beginsel is de architectuur hetzelfde. Het besturingssysteem is gebaseerd op de Red Hat Linux-distributie. Het gebruikt ook de Red Hat Package Manager voor de installatie, opwaardering en verwijdering van software-pakketten.
"Gebruikers en beheerders die gewend zijn met Red Hat te werken, kunnen vrij snel aan de slag," zegt Van Leeuwen. "Database-beheerders die niet gewend zijn aan Linux zullen zich dat toch eigen moeten maken."

> De te gebruiken database dient uiteraard geschikt te zijn voor een cluster

ClusterVisionOS gaat uit van een meester/slaaf-opstelling. De 'meester'-server heeft volledige controle over de 'slaaf'-servers. Het centrale knooppunt slaat alle informatie over de 'slaven' op (het heeft een directory met alle 'slave-images') en voorziet elke 'slaaf' van een IP-adres en Linux-kernel op het moment van opstarten. Als een 'slaaf' opstart, synchroniseert hij ook zijn hard disk image met een virtueel image die op de 'meester' staat. Gewoonlijk zijn alle 'slaven' hetzelfde en delen ze dezelfde virtuele image op de 'meester'. Het is overigens mogelijk hiervan af te wijken.

Het voordeel van de gebruikelijke opstelling is dat de 'meester' het zenuwstelsel vormt van het cluster en het dus volstaat alleen dit knooppunt te beheren. Andere pluspunten zijn dat 'slaven' gemakkelijk kunnen worden toegevoegd of verwijderd, dat alleen een reservekopie nodig is van de harde schijf van de 'meester' en dat een enkel commando volstaat om een 'slaaf' te herstarten en terug te laten keren naar een bekende configuratiestatus.
Ook de stroomtoevoer naar het cluster loopt via het centrale knooppunt. Alle 'slaven' kunnen individueel of groepsgewijs aan en uit worden geschakeld via de 'meester'.

Jobs

Voor een DBA is het van belang te weten hoe het cluster omgaat met job-afhandeling. Het is mogelijk bepaalde query's voorrang te geven boven andere of een tijdvenster te maken waarin helemaal geen query's mogen worden gedaan, omdat het cluster dan bijvoorbeeld zijn rekentaken moet uitvoeren.
Het besturingssysteem gebruikt voor job queuing de Sun Grid Engine (SGE) of het Portable Batch System (PBS), De Maui Scheduler is een optie en dient voor een doelmatiger *job scheduling* binnen het wachtrijensysteem. PBS en SGE accepteren batch- en parallelle jobs, bewaren en beschermen de job tot hij aan de beurt is en zorgen ervoor dat de opdrachtgever zijn antwoorden krijgt. Enkele eigenschappen zijn automatische *loadlevelling* (bijvoorbeeld door processen te migreren als dat nodig mocht blijken), *file staging* (het beschikbaar stellen van

bestanden voordat een job die deze files nodig heeft in gang kan worden gezet), *job interdependancy*, beveiliging en autorisatie, in kaart brengen van gebruikersnamen, het bijhouden van de afhandeling van jobs en ten slotte het bieden van een grafische schermopbouw. De Maui Scheduler is een dienst aan SGE of PBS en definieert de job scheduling afspraken en zorgt voor naleving ervan. Maui bepaalt welke jobs aan de beurt zijn en waar ze worden afgehandeld, en in welke volgorde.

Database

De te gebruiken database dient uiteraard geschikt te zijn om in een cluster te worden gebruikt. Er is overigens in de front end wel een SQL interpreter die bepaalt waar de data naar toe moeten. "Het is handig als een database-beheerder de onderliggende architectuur van het cluster begrijpt. Dat is nuttig bij de inrichting van de database. De ordening van het cluster heeft invloed op de prestaties van de database. Daar moet dus rekening mee worden gehouden. De database zelf grijpt trouwens ook behoorlijk hard in op de hardware. Dat geldt voor Oracle 9i RAC (Real Applications Cluster, TM) en voor MySQL," zegt Van Leeuwen.

De meeste clusters die ClusterVision heeft gebouwd, hebben Oracle 9i RAC als database. In 2003 is de onderneming officieel tot Worldwide Oracle Partner benoemd. Het bedrijf levert turn key clusters af. In nauw overleg met de klant bouwt het specifieke clusters. Het ontwerp ervan hangt af van het gebruik. Een cluster waar vooral veel wordt gerekend, heeft andere eigenschappen nodig dan eentje waarop een zware database draait. Het samenstellen, configureren en afregelen van een cluster dat gemakkelijk is te gebruiken, stabiel en veilig is, vereist diepgaande kennis van Linux en kost aanzienlijk veel tijd. Vandaar dat ClusterVision computergroepen aflevert die meteen aan de slag kunnen met de applicaties waarvoor zij zijn gebouwd.

Juist vanwege die volledigheid wil ClusterVision zijn clusters leveren met MySQL Cluster. "De parellele, open database MySQL is toch een stuk goedkoper dan Oracle," licht Ninaber toe. "Omdat voor MySQL geen licentie nodig is, geeft dit ons de mogelijkheid een compleet aanbod te doen. Bij Oracle moet je toch altijd een licentie hebben."

Teus Molenaar is freelance journalist.

Europese ambities

De twee drijvende krachten achter ClusterVision hebben elkaar ontmoet bij de Engelse onderneming Compusys HPC, een divisie van Compusys die gespecialiseerd is in clusters. Het onderdeel is ongeveer zeven jaar geleden opgericht door Alex Ninaber; Matthijs van Leeuwen kwam er later bij werken. De divisie was de eerste die commerciële clusters aanbood. Het tweetal kreeg binnen het bedrijf, naar eigen gevoel, te weinig ruimte om hun ideeën uit te werken. Het gevolg is dat zij twee jaar geleden een eigen bedrijf zijn begonnen: ClusterVision.

"Om op volledige snelheid te beginnen, hadden we partners nodig," zegt Van Leeuwen. "Dankzij de Universiteit van Amsterdam, ECL en NEC HPCE hebben we een vliegende start kunnen maken. Omdat Engeland een belangrijke markt is voor ons, hebben we meteen ook een kantoortje in Londen geopend. Maar het hoofdkantoor staat in Hoofddorp. Niet alleen omdat wij Nederlanders zijn, maar vooral ook omdat wij Europese ambities hebben en dan heb je toch meer aan een vestiging op het continent. Bovendien is het makkelijker hier mensen te vinden, is de hardware hier iets goedkoper en is er in Nederland toch meer gevoel voor kwaliteit," meent Van Leeuwen.

ECL Computers is een distributeur in computer-componenten en multimedia-producten. Magazijn en kantoorruimte zijn gevestigd in Hoofddorp. In hetzelfde bedrijfspand heeft ClusterVision zijn onderkomen; bovenop de hardware die nodig is om clusters te bouwen. Power Computing & Communications van de Universiteit van Amsterdam (PPC UvA) is eind 1995 opgericht door de sectie Computersystemen van de faculteit Natuurwetenschappen, Wiskunde en Informatica. Prof. dr. Bob Herzberger, de Nederlandse HPC- en grid-specialist, is wetenschappelijk directeur van PPC UvA. De samenwerking met ClusterVision (onder andere door een belang van vijftien procent) geeft de universiteit een weg naar het bedrijfsleven en geeft ClusterVision toegang tot de knapste koppen op het gebied van HPC en grid-computing.
De samenwerking met NEC HPCE (High Performance Computing Europe) houdt onder meer in dat ClusterVision *cluster solution provider* is voor NEC HPCE voor Noord-Europa.