

In augustus 2009 ontvingen Peter Boncz, Stefan Manegold en Martin Kersten van het Centrum Wiskunde & Informatica de prestigieuze VLDB 10-year Best Paper Award. Hun winnende artikel beschreef tien jaar geleden de eerste ideeën over ‘hardware-aware’ ontwerpen van databasesystemen. Sindsdien maken zowel database-leveranciers als spin-offs gebruik van de ideeën en de open source code van het CWI. Door Daphne Riksen

# ‘We investeren bewust in technologie’



De winnaars van de VLDB 10-Year Best Paper Award: v.l.n.r. Peter Boncz, Stefan Manegold en Martin Kersten.

In zijn werkkamer in de gloednieuwe aanbouw van het Centrum Wiskunde & Informatica (CWI) op het Science

Park in Amsterdam vertelt Martin Kersten dat 2009 niet alleen vanwege de VLDB 10-year Best Paper Award een mooi jaar was: ‘We kregen ook de Runner-up Best Paper Award van ACM Special Interest Group on Management of Data (SIGMOD) en een uitnodiging om als een van de eersten in *Communications of the ACM* een overzichtsverhaal te schrijven over onze activiteiten. Drie grote erkenningen in het jaar dat ons database-technologieproject MonetDB vijftien jaar bestaat, dat kun je wel een oogstjaar noemen!’, lacht hij.

## Uitgekristalliseerd

De VLDB 10-year Best Paper Award wordt jaarlijks uitgereikt aan het artikel dat tien jaar eerder is gepresenteerd tijdens de International Conference on Very Large Data Bases (VLDB) en sindsdien de meeste invloed heeft gehad. Van de vijftig artikelen uit 1999 werd gekozen voor ‘Database Architecture Optimized for the new Bottleneck: Memory Access’. Martin Kersten, Peter Boncz en Stefan Manegold, alle drie werk-



Dr. Peter Boncz (1970) werkt sinds 2002 bij het CWI en is senior onderzoeker op het gebied van databasearchitecturen en dan vooral voor query-intensieve toepassingen zoals data mining, XML-databases and multimedia retrieval. Hij promoveerde in 2001 bij Martin Kersten aan de Universiteit van Amsterdam. Tijdens zijn promotie-onderzoek raakte hij betrokken bij de CWI-spin-off Data Distilleries waar hij van 1999 tot 2001 fulltime werkte en waarvoor hij de ICTRegie Award 2006 ontving. Hij is medeoprichter van de spin-offs VectorWise en MonetDB BV.

zaam bij het CWI, schreven het veelgeciteerde artikel. Kersten legt uit waarom de ideeën zo invloedrijk zijn gebleken: 'Begin jaren negentig was databasetechnologie een behoorlijk uitgekristalliseerd gebied. Het bestond al twee decennia en er gingen miljarden in om. Het gevolg daarvan was dat het onderzoek verengde: de parameters

waarop werd geoptimaliseerd lagen vast. Indertijd was alles gericht op hardwaretechnologie en functionaliteit uit de tachtiger jaren. Er werd bij de ontwikkeling van complexe software zoals databasesystemen nauwelijks ingespeeld op de enorme vooruitgang in rekenkracht. Men hield alleen rekening met de beperkingen van de harde schijf. Bovendien werd alles geoptimaliseerd voor één specifieke databasetoepassing: het zo snel mogelijk verwerken van administratieve handelingen zoals banktransacties. Wij hebben dat in 1993 omgegooid. Enerzijds hebben we het primaire geheugen centraal gesteld, anderzijds de databasearchitectuur gekanteld. Onze publicatie uit 1999 was het eerste paper waarin stond hoe je datastructuren en -algoritmes zou moeten veranderen om in deze context beter gebruik te maken van moderne complexe hardware. Het was het

begin van een nieuw thema binnen databaseonderzoek: hardware-aware ontwerpen.' 'Wij waren een van de eerste onderzoeksgroepen die hiermee bezig waren en de opgedane kennis namen we mee bij de ontwikkeling van ons databasesysteem MonetDB', vult Boncz aan.

#### VLDB 10-year Best Paper Award

De VLDB 10-year Best Paper Award wordt jaarlijks uitgereikt aan het artikel dat tien jaar eerder tijdens de International Conference on Very Large Data Bases (VLDB) is gepresenteerd en sindsdien de meeste invloed heeft gehad. Martin Kersten, Peter Boncz en Stefan Manegold van het CWI kregen de prestigieuze prijs in 2009 voor hun in 1999 gepubliceerde artikel 'Database Architecture Optimized for the new Bottleneck: Memory Access'. Daarin beschrijven zij hun ideeën over hardware-aware ontwerpen van databasesoftware, waardoor beter gebruik wordt gemaakt van de mogelijkheden van de ontwikkelingen in hardware. De prijs kregen zij ook voor hun baanbrekende werk op het gebied van column-oriented databasetechnologie, wat resulteerde in het open source databasesysteem MonetDB. Deze technologie is vrij beschikbaar voor onderzoekers, gebruikers en databaseontwikkelaars en de achterliggende principes worden gebruikt door onder meer Oracle, Microsoft en SAP.

– Het winnende artikel is te vinden op [www.ictonderzoek.net](http://www.ictonderzoek.net) onder Archief.

– Voor informatie over MonetDB: [www.monetdb.org](http://www.monetdb.org)

#### Slanke code

Traditioneel bestaan relationele databasesystemen uit tabellen met regels, die ook in die vorm in het geheugen worden opgeslagen en verwerkt. Dat is prima voor banktransacties waar individuele klantenrecords worden gelezen en gewijzigd, maar minder logisch wanneer je data wilt analyseren en op zoek gaat naar trends, zoals bij business intelligence gebeurt. 'Hele regels ophalen is dan heel inefficiënt', legt Boncz uit. 'Het is veel slimmer om alleen de kolommen op te halen waarin je geïnteresseerd bent. Kolomsgewijze organisatie is dan veel handiger. We hebben dus de hele architectuur gekanteld en alle algoritmen daarop aangepast. Dat is overigens niet alleen interessant voor business intelligence toepassingen, maar ook voor data-intensieve sciences zoals de biologie en de astronomie. Die disciplines stellen steeds hogere eisen aan dataverwerking en daar zijn we in Nederland goed voor gepositioneerd.' Manegold vult aan: 'In dat soort toepassingen ben je niet uit op zoveel mogelijk transacties in een bepaalde tijd. Omdat je vooral gegevens leest, is dus alle software die transacties moet beschermen overbodig. Met als resultaat slankere code, geoptimaliseerd voor data-analyse.'



Prof. dr. Martin Kersten (1953) is 25 jaar verbonden aan het CWI. Hij is er cluster-leider Informatiesystemen en groepsleider van de groep Database Architectures. Daarnaast is hij één dag per week hoogleraar in multimedia databases bij het Instituut voor Informatica van de Universiteit van Amsterdam. Hij was medeoprichter van Data Distilleries en het recent gestarte MonetDB BV, een spin-off met als taak de open source code van MonetDB te beheren, onderhouden en verspreiden.

## Onderzoeker of entrepreneur

De resultaten van de CWI-onderzoeksgroep blijken ook commercieel interessant. Bedrijven als Oracle, Microsoft, Ingres en SAP gebruiken de principes in hun nieuwe producten en projecten en zijn goed op de hoogte hoe de MonetDB-code in elkaar zit, merken de onderzoekers

tijdens congressen. Daarnaast is vanuit het CWI een flink aantal spin-offs ontstaan. De eerste versie van MonetDB leidde in 1995 tot Data Distilleries. Dit data-mining bedrijf werd in 2003 overgenomen door SPSS, dat vorig jaar door IBM werd ingelijfd. Ook het in 2008 gestarte spin-off VectorWise en het kersverse Spinqe maken gebruik van MonetDB. Speciaal voor de groeiende groep gebruikers van de vrij beschikbare open source code (die 10.000 keer per maand wordt gedownload) werd in 2008 MonetDB BV opgericht, met als taak de code te beheren, onderhouden en verspreiden. Het roept de vraag op of de mensen van het CWI onderzoeker zijn of entrepreneur. 'Er is geen keuze', vindt Boncz. 'Gezond datamanagementonderzoek móet resulteren in vindingen die in de markt worden gebruikt. Leidende universiteiten als Stanford en de University of California in Berkeley maken die keuze ook niet en brachten Google en Ingres voort.

De core business van het CWI blijft wetenschap, maar als je op het terrein van datamanagement als onderzoeker actief bent, dan moet je werk relevant zijn voor het ICT-bedrijfsleven.' Kersten voegt toe: 'Bovendien is statutair

bepaald dat het CWI verantwoordelijk is voor grensverleggend onderzoek én transfer daarvan naar de markt.' Overigens is het CWI altijd aandeelhouder in een spin-off en worden er strikte contractuele afspraken gemaakt met (mede)oprichters over hun tijdsbesteding en eventuele detachering of overstap, legt Kersten uit. 'Elke spin-off vereist de ervaring, kennis en handjes van onze mensen om het van de grond te krijgen. Daar steken we veel effort in.'

## Investeren in technologie

Het is dus geen toeval dat het CWI zo succesvol is met spin-offs en met MonetDB. 'Maar het gaat niet vanzelf', weet Kersten. 'Je hebt een lange adem en veel geld nodig. Om weerwerk te kunnen bieden aan bedrijven als IBM, Google en Microsoft, die miljarden verdienen met datamanagement, moet je bewust investeren in technologie – er zit nu 200 manjaar in het MonetDB-project.' 'Met als consequentie dat we iets minder papers schrijven omdat we investeren in softwareontwikkeling. De MonetDB-software is al een boek van 8.000 pagina's', zegt Boncz. 'Dat is niet zo gebruikelijk in de informatica, maar wel in bijvoorbeeld de astronomie en fysica, waar soms de helft van de gepromoveerden werkt aan bouw en onderhoud van de software voor het experimenteerplatform', zegt Kersten. Slechts een klein gedeelte van de onderzoekers houdt zich daar bezig met het schrijven van papers of het uitvoeren van analyses. 'Als een informaticagroep in Nederland hetzelfde zou doen, word je bij de eerste evaluatie als niet productief beoordeeld', vervolgt Kersten. 'Het risico daarvan is, en dat zie je ook in de praktijk, dat informatica-

**Gezond datamanagementonderzoek móet resulteren in vindingen die in de markt worden gebruikt**



### Visionary outlook CWI researchers rewarded

In August 2009, Peter Boncz, Stefan Manegold and Martin Kersten from CWI (Center for Mathematics and Computer Science) received the prestigious VLDB 10-year Best Paper Award. Their winning article from 1999 was one of the first papers that described how data structures and algorithms would have to be changed to make better use of modern complex hardware. It was the start of a new theme within database research: hardware-aware designs. Since then, both researchers, database suppliers and spin-offs have been making use of the ideas and open source database system MonetDB from CWI. The results from the CWI research group have proven to be very interesting from a commercial viewpoint. During congresses the researchers notice that companies such as Oracle, Microsoft, Ingres and SAP use the principles in their products and projects and that they are well informed about how the MonetDB code works. What's more, numerous spin-offs have emerged from CWI. These successes did not come out of the blue. 'You need long-term commitment and considerable funding and you must deliberately invest in technology', says Kersten. 'To date 200 man-years have been invested in the MonetDB project.' The group has also managed to put together a strong and complementary team that has had a stable core of researchers since 1995. Thanks to this continuity, new PhDs do not have to keep on starting from scratch. Furthermore, PhDs spend an internship of at least 3 months at leading research departments of companies such as IBM or eBay. This interaction is vital, according to Kersten, Boncz and Manegold. 'The outcome is that our PhDs have a broader base: they have not just published articles but are also able to build and use systems. It's therefore hardly surprising that companies are keen to obtain new talent from CWI.' It was not just the VLDB 10-year Best Paper Award that made 2009 a fantastic year: the database architecture group also received the Best Paper Award from the ACM SIGMOD conference and an invitation to be one of the first to write an overview article about its activities for the Communications of the ACM journal.

Dr. Stefan Manegold (1969) is senior onderzoeker op het gebied van databasearchitecturen met een focus op performance en zelfstandige automatische optimalisatie van datastructuren en algoritmen. Na zijn studie Computer Science aan de Technische Universität Clausthal (Duitsland) is hij sinds 1997 verbonden aan het CWI. In 2002 promoveerde hij bij Martin Kersten op het onderwerp 'Understanding, Modeling and Improving Main-Memory Database Performance'. Hij is medeoprichter van MonetDB BV en verantwoordelijk voor het testen en de kwaliteitscontrole van de open source versie van MonetDB.

onderzoek vaak erg microgericht, geïsoleerd en theoretisch van aard is. Dat is jammer, want bruikbare doorbraken realiseer je daarmee niet snel.' Inmiddels is in elk geval de databaseonderzoekswereld volgens Kersten in een volgende fase van volwassenheid beland. Onderzoeksresultaten worden meer en meer van een kwaliteitsstempel voorzien als ze door anderen gevalideerd kunnen worden. Manegold is medeoprichter van dat internationale proces. Hij legt uit: 'Counter-evaluatie van je onderzoek wordt onderdeel van de evaluatiecriteria van grote conferenties zoals ACM SIGMOD. Je kunt je paper en je code inleveren bij een commissie die test of je werk reproduceerbaar is. Daarmee creëer je een onderzoekscultuur die in andere disciplines zoals biologie en medicijnen heel normaal is, maar in de informatica nog ongebruikelijk.'

### Vaste kern

Terug naar het succes van de databasearchitectuurgroep van het CWI. Hoe krijgen ze dat toch voor elkaar? Kersten mag het graag vergelijken met voetbal. 'Wil je in ICT op dit niveau wereldwijd meedoen, dan heb je niet een groep schakers maar een complementair team nodig. Iedereen heeft zijn eigen expertise om het spel te beïnvloeden', legt hij uit. 'Ook heel belangrijk: we hebben al sinds 1995 een beperkte, maar wel vaste kern van onderzoekers, waarbij Niels Nes en Sjoerd Mullender niet onvermeld mogen blijven.' Manegold: 'Die continuïteit is bijzonder, want dan hoeven nieuwe promovendi niet steeds vanaf scratch te beginnen.' Promovendi brengen bovendien een internship van minimaal drie maanden door bij gerenommeerde research-afdelingen van bedrijven als IBM, Microsoft Research, Google of eBay. Die wisselwerking is essentieel, vinden Kersten, Boncz en Manegold. 'Met als gevolg dat onze gepromoveerden een brede basis hebben: ze hebben niet alleen gepubliceerd, maar kunnen ook systemen bouwen en gebruiken. Het CWI is niet voor niets bij bedrijven een gewilde bron van nieuw personeel.'

**I/O**