★ Modern research techniques produce amounts of data far beyond the storage capacities of conventional computing environments. Grid computing offers an effective, efficient and reliable way to manage large volumes of data, say Arjen van Rijn and Dr. Maurice Bouwhuis of the BiG Grid project

The BiG idea

THE emergence of new research methods has united many scientific disciplines in the need to manage large volumes of information. Modern detectors, medical imaging instruments and micro-arrays produce amounts of data far beyond the storage capacities of conventional computing environments, demanding ongoing improvement to research infrastructures. In this context the work of the BiG Grid project, a Netherlands-based initiative which aims to enable access to grid computing infrastructures for scientific researchers, takes on real importance.

"Our goal is to build and roll-out a nationwide, grid-based, e-Science infrastructure," says Arjen van Rijn, the chairman of the BiG Grid executive team. Based himself at Nikhef, the National Institute for Sub-atomic Physics, Van Rijn says the structure of the BiG Grid project reflects the broad relevance of their work. "The BiG Grid project is a collaborative effort between Nikhef, NBIC (Netherlands Bioinformatics Centre) and NCF (the National Computing Facilities foundation)," he explains. "This covers the majority of the ICT-intensive research communities in the Netherlands. So the project is very much user-driven."

Shared requirements

This approach is designed to ensure that the infrastructure meets the needs of scientists from a wide range of disciplines. Understanding researchers shared requirements is a crucial first step. "Typically researchers who want to use grid infrastructures are handling more data than they are used to, and also they want to share it. They want to be able to combine the data, to analyse it, and to engage in innovative research," outlines Van Rijn. However,

while areas like particle physics make clear demands of the infrastructure, scientists from other research communities are often less familiar with large central facilities. The project's consultative approach, taken alongside recognition of researchers common requirements, represents a sound basis for the development of an effective, reliable infrastructure. "We try to discover, through dialogue with the application scientists, what they need for their projects to make maximum use of the infrastructure, which in turn will help them do better science and enlarge their own research potential," says dr. Maurice Bouwhuis, leader of the e-Science support team at SARA, the Netherlands national High Performance Computing and e-Science Support Centre, BiG Grid's principal operational partner. Bouwhuis chairs both the BiG Grid operational steering team and the support & development steering team. Some of these projects don't go past the pilot stage, however others gain real benefits from using grid infrastructure,

particularly in terms of the enhanced opportunities it offers to collaborate with other scientists. This is of particular importance given the broad-based, international nature of much scientific research, which in turn can give rise to new areas of investigation. "The project is not only about enabling access to research infrastructures, but also expanding the research environment," stresses Van Rijn. This work is designed to complement rather than replace existing research infrastructures. "We want the BiG Grid infrastructure to be part of the wider ICT infrastructure for scientific research. As such we need a network – the distributed infrastructure is on top of it, and you have the dedicated, specialist (or 'capability') computing facilities on top of that. So there's a whole eco-system of ICT-based facilities," explains Van Rijn. "Then you have specific application domains, like accelerator experiments at CERN for example. The BiG Grid project acts as an extension of scientific research tools, and

enables scientists to cross between research infrastructures."

Current infrastructures enable complex research across a range of disciplines, including meteorology, life sciences and climate change. However, the dynamic nature of today's research environment means it is important for technological infrastructures to adapt rapidly to emerging priorities, an issue of which Van Rijn is well aware. "New developments emerge continually and you have to incorporate them in your infrastructure," he acknowledges. The project's decision to change part of their hardware infrastructure to be cloud accessible, in response to demand from user communities, provides a clear example. "We have seen that Grid infrastructure is well suited to certain types of usage. But we thought that certain communities, within the life sciences, social sciences and humanities, would want a computing environment that they could further tune and develop according to their needs," says Bouwhuis. "That's why we
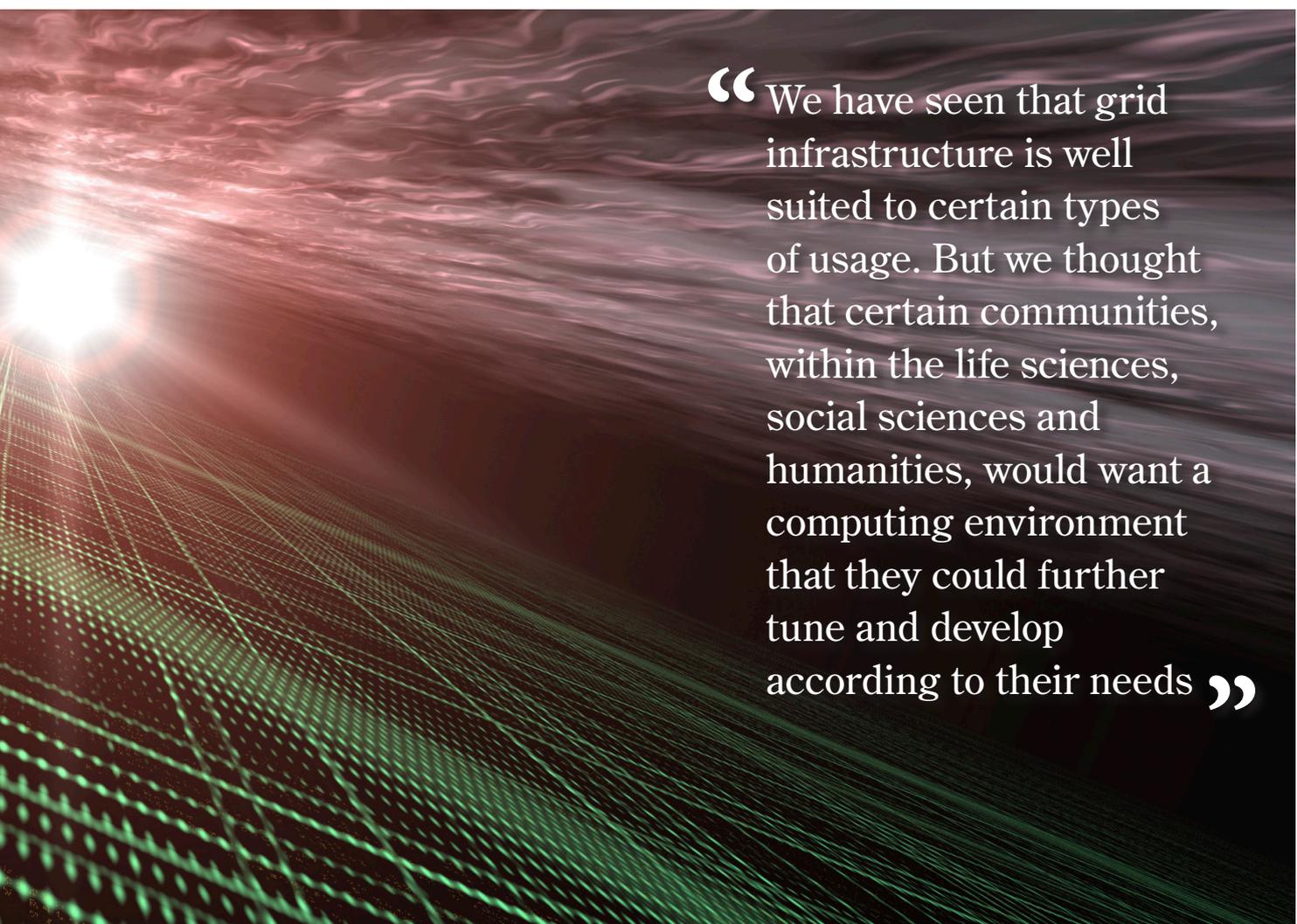
> "We have seen that grid infrastructure is well suited to certain types of usage. But we thought that certain communities, within the life sciences, social sciences and humanities, would want a computing environment that they could further tune and develop according to their needs"

have developed a cloud environment over the last six months, and it's now being used extensively. This wasn't in the original proposal, but we don't want to replace one type of infrastructure with another. There is a raison d'etre for both types and they can serve different communities."

Distributed infrastructure

The distributed nature of the BiG Grid infrastructure allows the project to combine these various aspects. The project's four central facilities (based at SARA, Nikhef, Philips Research and the RUG Centre for Information Technology) each provide both large scale computing and storage capabilities, meaning that in principle scientists are able to work at any of the four sites, while they also allow for more specialised work. "The SARA site is being used for structured data production and long-term archiving - for example data from the Large Hadron Collider experiments - while the Nikhef site caters more for research at the high-energy physics

institute in the Netherlands," says Dr Bouwhuis. This need for both specialised and generic facilities places significant demands on the infrastructure. "It's impossible to have a very large storage environment without a very good network. This is because while you want to get the data in you also want to do things with it," points out Dr Bouwhuis. "That means that you have to get the data out of the storage environment - possibly many times - and as quickly as possible. But if you can do that then you also need to have a very good network, and your computing has to be balanced with the data going through it." Complementing the central facilities, twelve small scale computing and storage clusters are located within centers for life science research throughout the Netherlands (mainly academic hospitals), targeted at the NBIC researchers.

Balancing network computing power and storage is crucial to making effective use of the infrastructure. The very short innovation cycles that characterise the

recent history of grid computing allow for ongoing development of computer power and storage facilities; however, Dr Bouwhuis says there are also other issues to consider. "You can have a perfectly tuned hardware set-up but if the middleware set-up doesn't work well then you are still not reaching the requirements of the community. It also depends on what the community is putting in, on how they use it," he outlines. An active user community that understands the infrastructure and their requirements is best placed to use it effectively. "Building bridges between scientific communities is one of the main things that we do," says Bouwhuis. "Programmers from the BiG Grid project are working together with scientists from the domains themselves to make sure that the scientists are able to run their application on the grid without needing great technical expertise. This is provided by the scientific programmer from the particular scientific domain. So it's a multi-tier approach – there's the BiG Grid
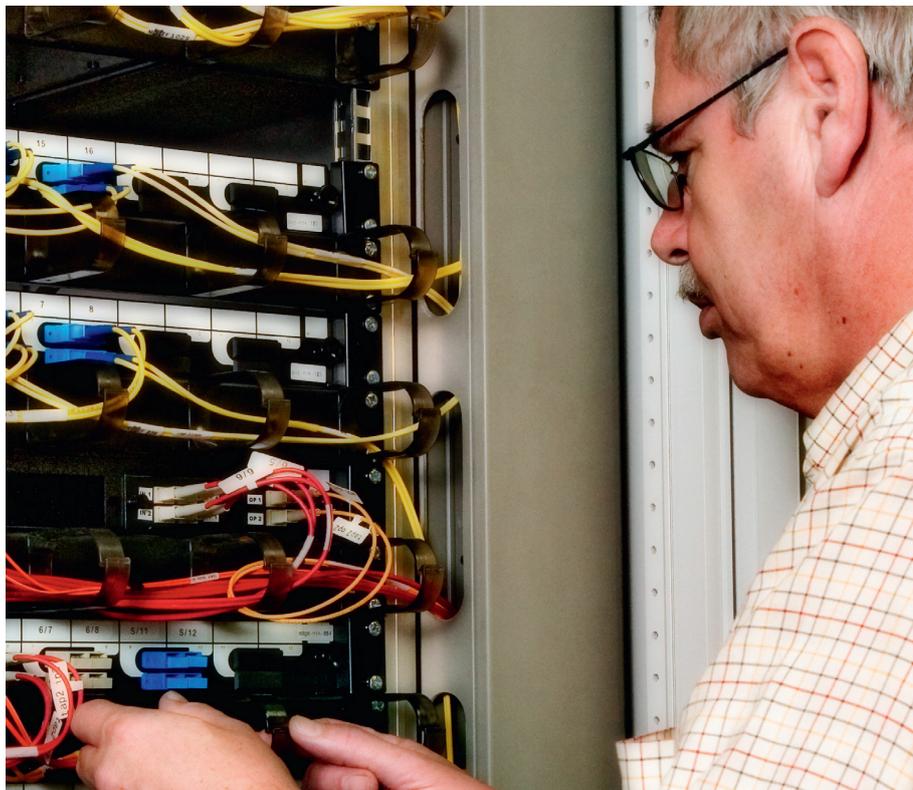
scientific programmer, there's the domain scientific programmer, and at the top there's the scientist."

This work is built on further by domain-specific portals, such as those for nuclear magnetic resonance research and theoretical chemistry simulations.

The ease of using these portals belies the complexity of their structure. "Together with the communities we have put a lot of effort into enabling these portals to use Big Grid facilities.

The users don't know this; they just see the portal and think 'I can run my application really fast through this portal,'" says Bouwhuis.

This is work with broad relevance, and widening the scope of the project further to include other scientific disciplines is a crucial part of the project's future plans.

"Anyone who has a central set of data that he wants to distribute can use the grid – there are many other application domains that are also data-intensive, but that have a slightly different model.

There are many sources and uses of data in bioinformatics, social sciences and humanities, and some researchers may want to combine them," says Dr Bouwhuis. "For example BiG Grid is working with a project looking at the problem of bird strikes on aircraft.

Data from a range of different sensors, from large military radars right through to a small GPS on top of a bird, are combined, and are being used by scientists to make projections on how big the bird migration will be at certain times in certain locations. The outcomes are relevant both for civil and military aviation industry."

Maurice Bouwhuis

Obtained a PhD in physics from the University of Utrecht while working at the National Institute for subatomic physics Nikhef in the area of experimental nuclear physics. He has worked as a post-doc in the US and at a high-energy physics laboratory in Hamburg, Germany, where he was part of the management of the HERMES experiment. Maurice Bouwhuis joined SARA in 2004 and now leads the e-Science Support group within the High Performance Computing Department. This department participates in National projects like BiG Grid and the BioAssist program of the Netherlands Bioinformatics Centre, as well as international projects like EGI, DEISA and PRACE. Within BiG Grid he leads the Operations Steering Team and the Support and Development Steering Team.

At a glance

Full Project Title
BiG Grid, the Dutch e-science Grid

Project Objectives
The BiG Grid project aims to roll out a nation-wide grid-based e-science research infrastructure. The project encompasses data storage facilities and data processing services, enabled by grid services. Human support is
available to many scientific disciplines that will benefit from the project, ranging from particle physics and life sciences to social sciences and humanities.

Project Partners
NCF: Netherlands National Computing Facilities Foundation
Nikhef: National Institute for Sub-atomic Physics
NBIC: Netherlands Bioinformatics Centre

Operational Partners
SARA Computing & Networking Services.
Philips Research Eindhoven.
Donald Smits Center for Information Technology (CIT) Groningen.

Contact Details
Peter Michielse
T: +3170 3440700
E: info@biggrid.nl
W: www.biggrid.nl

Arjen van Rijn

Arjen van Rijn is since 1999 manager of Nikhef, the Dutch National Institute for subatomic physics, which has been involved in national and European grid projects since 2001. He is chairman of the Executive Team of the BiG Grid project, aimed at rolling out an e-Science infrastructure in The Netherlands. He has also chaired the taskforce for the establishment in Amsterdam of the European coordinating grid body, EGI.eu, and is currently a member of its Executive Board.

BiG Grid
the dutch e·science grid