Square Kilometer Array: Ultimate Big Data Challenge

Partners to research the exascale computer systems that are needed for what will become the world's largest radio telescope

Big Data Meets the Big Bang

Every day, the antennas will gather 14 exabytes of data and store about one petabyte.

The Square Kilometer Array (SKA) is one of the most ambitious science projects ever undertaken. A consortium of 10 nations, with the involvement of numerous university scientists and industrial companies, plans on setting up a massive radio telescope made up of millions of antennas spread out across vast swaths of southern Africa and Australia. When it's completed in 2024, the array will give astronomers insights into the evolution of the first stars and galaxies after the Big Bang so they can better understand the history of the universe and the nature of matter.

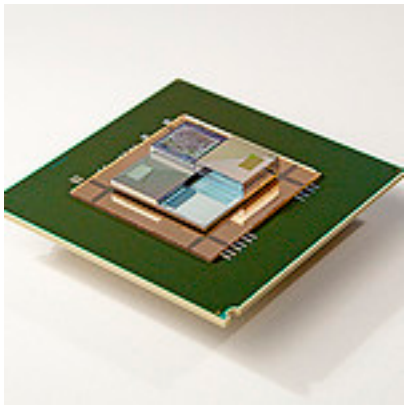
The SKA telescope will collect a deluge of radio signals from outer space. Every day, the antennas will gather 14 exabytes of data and store about one petabyte. (For comparison, one exabyte of digital music would take two million years to play back on an iPod. [1]) Because the telescope is to be made from so many individual antennas, the antennas are to be so widely scattered, and such a large volume of data is being gathered, that a novel computing system must be developed to manage the process of gathering, storing and analyzing data from end to end. The SKA represents the ultimate *Big Data* challenge.

To take on this challenge, IBM and the Netherlands Institute for Radio Astronomy (ASTRON) have created a five-year collaboration, called DOME, aimed at designing an information technology system that could be used for managing the data that the SKA produces. South Africa's National Research Foundation later joined the collaboration as a user platform member, and IBM and ASTRON look forward to other partners joining the initiative.

---

[1] http://www.skatelescope.org/media-outreach/fun-stuff/facts-figures/

Background Information



"This project requires partners who are willing to push the borders of computer science."

Marco de Vos, managing director, ASTRON

ASTRON and IBM have mapped out seven technology projects aimed at dealing with the extreme data-handling requirements of the SKA. The projects cut across a wide array of information technology fields, from information management and analytics to computer system design and chip design. "This project requires partners who are willing to push the borders of computer science," says Marco de Vos, the managing director of ASTRON. [2]

While these technology projects focus on SKA in particular, they could also help solve problems that confront people worldwide who are determined to take full advantage of the *Big Data* phenomenon. The emergence of social networking, sensor networks and huge storehouses of business, scientific and government records create an abundance of information, called *Big Data* in tech-industry parlance. This data comes in a wide variety of forms—not just text and numbers but video, audio, still imagery and, in the case of the SKA, radio waves. And some of it, including the SKA data, must be analyzed in real time. You can think of *Big Data* as a parallel universe to the world of people, places, things and their interrelationships. All of this data creates the potential for people to understand the environment around us with a depth and clarity that was simply not possible before. It's a new natural resource that's available to be mined.

But a natural resource isn't worth much unless you can take full advantage of it. Today, less than 1% of the digital data that has been collected in the world is actually analyzed. [3] All this data is difficult to capture, make sense of and move around. And, unfortunately, today's computing systems aren't up to the task of handling all of this raw information in an efficient and affordable way. That's why we need to develop new systems for managing *Big Data.*

The DOME project doesn't address all of the technology areas that encompass *Big Data*, but, because the requirements of the SKA are so

---

[2] Marco de Vos, managing director, ASTRON, Netherlands Institute for Radio Astronomy, interview, Oct. 16, 2012.

[3] The Digital Universe in 2020, IDG, Dec. 2012.

extreme, the work performed by scientists at ASTRON and IBM and their partner organizations will help the scientific community solve other data challenges. These range from analyzing climate change, genetic information and personal medical data to finding valuable nuggets of insight in vast business databases. “You proceed from just wanting to know what’s out there in space to using the technology for many other purposes. It’s about data storage, and data analysis and data use,” says Chris P. Buijink, secretary-general for the Dutch Ministry of Economic Affairs, Agriculture and Innovation, which financially supports the DOME project. [4]

The DOME technology projects will help build the scientific foundation for the era of cognitive systems.

The DOME research has even broader implications, as well. Taking full advantage of *Big Data* is a key element of what IBM sees as the new era of computing, which we call the era of cognitive systems. We believe that over the coming two decades, fundamental advances in science and technology will alter the relationship between humans and machines—turning computers into intelligent assistants that help people make better decisions and live more successfully and sustainably. The DOME technology projects will help build the scientific foundation for the era of cognitive systems.

How DOME Got Started

Most of us think we know what a radio telescope looks like. It’s a gigantic steel dish pointed toward the heavens. But that’s now old-school thinking. Large telescope dishes are extremely expensive and require a lot of moving parts to follow targets in the sky as the earth moves. It’s impossible to make the gains in data resolution that astronomers want using this technology. So the scientific community has adopted a new paradigm. Instead of using a few very large dishes, they’re using many small antennas with fewer moving parts that together make up a much larger telescope, capable of much higher resolution.

ASTRON was one of the first organizations to wake up to the need for a new approach to radio telescopy. It began testing out the new

---

[4] Chris Buijinks, Dutch Economics minister, interviewed Oct. 17, 2012.

thinking with a project called Low Frequency Array (LOFAR), the largest radio telescope in the world that performs observations in low frequencies, from 10 to 240 MHz. LOFAR, which was officially launched in 2010, is made up of more than 10,000 small antennas distributed across the Netherlands, the UK, Germany, France and Sweden.

The system uses two kinds of antennas. One type is made up of five-foot-tall posts that are held in place with four wires, which are also key elements of the antennas. A small disk at the top of the post contains electronics that amplify the radio waves and then transmit the data along underground cables to a nearby metal box, the size of a small car, where more processing takes place. The other type of antennas, handling the higher frequency signals, are made up of spear-shaped metal pieces arrayed in clusters on electronic circuit boards and covered with plastic tarps. In both cases, there are no moving parts. The system uses algorithms to make allowances in calculations for the movement of the earth. Using these so-called aperture array antennas, a telescope can for the first time view the entire sky at once, not just narrow slices of it.

IBM is one of ASTRON's key technology partners for LOFAR. Once data is collected in the field, it is transported by fiber-optic cables to a data center at the University of Groningen in the north of the Netherlands. There, an IBM Blue Gene supercomputer filters and correlates the data. In addition, IBM and ASTRON collaborated to design specialized data processing chips for the system. The computers and software make it possible for scientists to observe and analyze the raw data as it comes in. Think of it this way: the old-style metal dish telescope is being replaced with a high-tech virtual telescope fashioned from silicon and software.

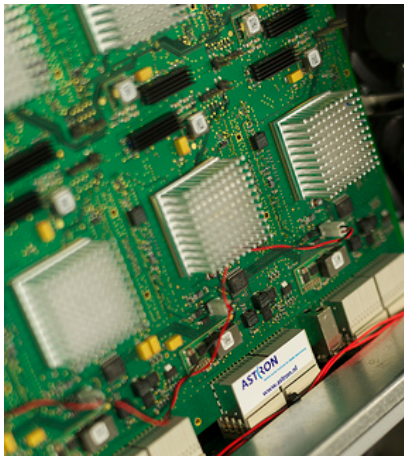
The SKA changes the game yet again. It will involve orders of magnitude larger amounts of data and numbers of antennas. (In addition to more than one half million aperture array antennas, the SKA will have 3,000 traditional dishes to handle higher-frequency signals.) Also, it's possible that thousands of miles will separate the antennas from data centers where much of the data processing is to

"To make it [SKA] affordable, you have to focus on low energy, low energy, low energy in every aspect of the design"

Ton Engbersen, Scientific Director for DOME, IBM Research

be done. All of this will require new thinking in information technology. “There is much more data than can be processed quickly. We must rethink the purpose of computing, rethink the methods by which we compute things and build the fastest computers to deal with this type of data,” says Prof. Arnold Smeulders, an advisor to ASTRON and director of the Informatics Institute of the University of Amsterdam.

The signals from outer space are a combination of valuable data and meaningless noise, so scientists have to process it all to sort out the useful stuff—and store just that. That argues in favor of doing much of the initial processing out in the field, close to the antennas.



The data processing itself will require much more powerful computers capable of 1000 to 10,000 Petaflops.

The data processing itself will require much more powerful computers capable of 1000 to 10,000 Petaflops per second processing power. For comparison, the fastest supercomputer in the world in 2012 is 17.6 petaflops. Achieving that performance level will require new designs for computing systems that are much more capable yet at the same time much more energy efficient.

Since so much of the energy in computing is required to move data around, scientists have to discover ways to move the data as efficiently as possible.

Keeping costs down will be vitally important. “Everything about the SKA is big,” says Ton Engbersen, the scientific director of DOME for IBM. “To make it affordable, you have to focus on low energy, low energy, low energy in every aspect of the design.”

Faced with these challenges, ASTRON engaged IBM Research – Zurich to perform a sort of show-and-tell of emergent technologies. In a series of meetings held at the lab, IBMers presented 20 proposals, of which ASTRON ultimately accepted seven—each intended to overcome a major information technology hurdle posed by the SKA. Those proposals evolved into a public private partnership called DOME and an institution for managing the project—the ASTRON & IBM Center for Exascale Technology, located in the small town of

Dwingeloo, in the Netherlands. ASTRON and IBM are using LOFAR as the test bed for their DOME research, with the hope that eventually the system they design will be used for the SKA.

The DOME Technology Projects

P1: Algorithms and Machines

The most strategic of the DOME projects is called Algorithms & Machines. The SKA challenge is so extreme and nobody has designed a data management system to handle anything like this before. So the goal here is to create an ultra-sophisticated software program that will help the team design the system holistically and optimally—taking into account all of the cost and performance trade-offs. Algorithms & Machines is the brainchild of Ronald Luijten, a Dutchman who has worked at IBM Research – Zurich for 28 years. He envisions it as a virtual R&D skunk works. He and his team are gathering all of the pertinent knowledge in a repository, setting the parameters for the entire computing system and creating optimization algorithms. The software program is cognitive. It will learn what they want and prepare a recommendation on how to fulfill their needs. In addition to LOFAR, the MeerKat telescope in South Africa will be used for development and testing of the Algorithms & Machine software.

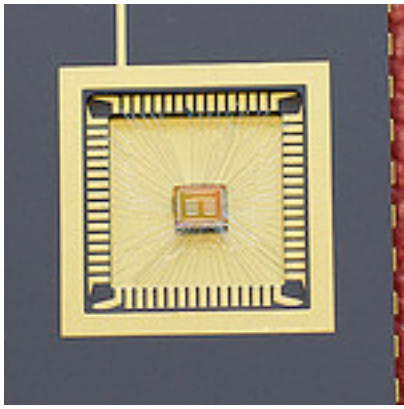
P2: Access Patterns

When the SKA is operating, it's expected to generate as much as an exabyte of data each day that will need to be stored for later analysis using computers. At that scale, storing and moving data will be very expensive. So the Access Patterns team, led by IBM researchers Jens Jelitto and Robert Haas, is developing a *Big Data* depository architecture for optimizing the management of SKA data. In computing, data is stored in a variety of ways, or tiers, depending on how often it is needed and the cost of the storage medium. Magnetic tape is the least costly form of storage, but data retrieval is slow. Disk drives are more expensive, but faster to access. Data storage on memory chips is lightening quick, but extremely expensive. Typically, database

administrators move data from tier to tier based on rigid policies and schedules. The Access Patterns technology will learn from its interactions with the data and parcel it out to the storage medium that's most appropriate for each piece at a particular moment in time.

P3: Nanophotonics

The SKA will contain thousands of antennas with a combined collecting area of about one square kilometre, and the data that's retained to be stored and analyzed will be shipped to far-off data centers, hundreds or thousands of miles away. To transport all of this data, a fiber-optic communications network will have to be built that's capable of moving data at 100 times the rate of today's Internet traffic. [5] But the real bottleneck comes when the data reaches the computers where it's processed and analyzed. The computers transport data internally via electronic bits moving on copper wires. So the SKA will be like attaching a fire hose to a garden sprinkler. DOME's Nanophotonics team, led by IBM researcher Bert Offrein, is taking photonics technology that IBM was already developing for general computing and applying it to the SKA challenge. They're pushing photonics ever further into the center of computing—first into the links that connect one server computer to another, then onto the circuit boards within individual computers, and, finally, on the nanotechnology level, connecting to the microprocessors where the analytics work is done.



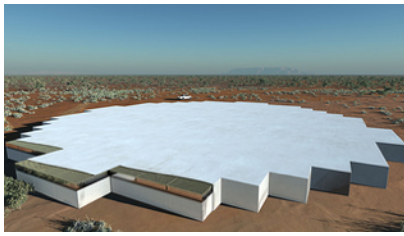
"We'll be able to map the so-called 'dark ages,' the epoch of ionization, when the stars and galaxies formed.

Albert-Jan Boonstra,
Scientific Director for DOME,
ASTRON

P4: Microservers

Today's server computers are expensive and use a lot of energy. They're about the size of a pizza box. It's possible that the designers of the SKA system will decide to perform a first round of data filtering or analysis close to the antennas, or even on them. In order to do so, they would need very small, inexpensive and highly energy efficient servers to perform those processing jobs. The DOME Microservers team, which, like Algorithms & Machines, is led by IBMer Ronald Luijten, is designing a small but powerful server, about the size of a

[5] http://www.skatelescope.org/the-technology/signal-processing/



Scientists from South Africa will work on making the microservers "desert proof" to handle the extreme conditions.

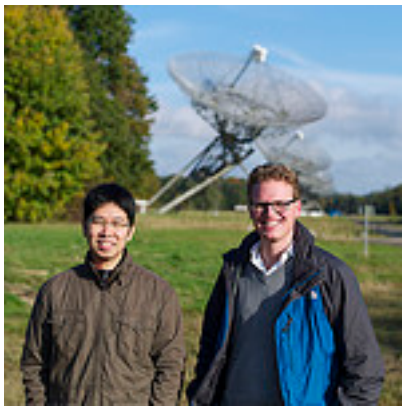
bar of soap, that will contain most of the functions of today's servers. The microprocessor, based on IBM's PowerPC architecture, is inexpensive and extremely energy efficient—which is necessary since the antennas will be in places with no electric power grid. Another possibility is that the SKA organization will want to use a microserver design within data centers. In that case, many of the devices would be packed close together in metal racks. To prevent such dense packages of electronics from overheating, a water cooling technique that IBM has developed could be used. Unchilled water is directed through microscopic channels across the surfaces of chips in a server. Then the hot water is used for some other purpose. In the SKA scenario, it could be used to supply energy for sea water desalination in southern African and Australian deserts.

In addition, computers aren't typically designed to operate in the middle of the Kalahari Desert, which can reach as high as 50 °C (122 °F). Scientists from South Africa will work on making the microservers "desert proof" to handle the extreme conditions.

P5: Accelerators

Today, the most demanding computing jobs are handled by supercomputers that link together thousands of microprocessors so they behave like a single large machine. They use brute force computation to get the job done. But the volume of data produced by the SKA will be too much for even these immensely powerful machines to handle affordably. The DOME Accelerators team, led by IBMers Christoph Haglietner and Jan van Lunteren, is exploring the possibility of creating hybrid systems containing both traditional supercomputer elements and another kind of processor, the accelerator. Accelerators handle certain kinds of computational tasks especially well, including pattern recognition. In one scenario, they might be positioned close to where the SKA data is stored so they can filter the data and send only the useful bits to the main microprocessors for analysis. In another scenario, where data is streaming into the system from the antennas, accelerators might be used to transform it on the fly—for instance removing the distortion to radio waves caused by passing through the earth's atmosphere. These accelerators are programmable, meaning

engineers can use software to change how they operate. No need to install a special accelerator for each particular processing task.



The signals from outer space are a combination of valuable data and meaningless noise.

P6: Compressive Sampling

One way to reduce the sheer volume of data in the SKA data management system is to compress the data as it streams in. This approach could result in significant savings in energy use, storage and processing. The DOME Compressive Sampling team, lead by IBMer Paul Hurley, is developing specialized signal processing and machine learning algorithms for the capture, processing and analysis of radio astronomy data. Conventional compression techniques, such as the JPEG digital image format, gather a tremendous amount of information only to throw most of it away. In contrast, compressive sampling can greatly reduce the number of samples that must be collected. Machine learning algorithms make it possible for the computer to learn through exposure to the data so it can recognize which patterns are significant and retain the data pertaining to them.

P7: RT Communications

Because of the huge amount of data captured by the SKA, it will be vital to move data through the computing system as quickly and efficiently as possible. Traditional communication methods involve repeatedly copying data and descriptive information as it passes through a network from device to device. That causes unwanted communication latencies and may restrict maximum available communication bandwidth. New technology standards have been developed to eliminate all of that unwanted overhead, resulting in a capability called real-time communications. The DOME RT Communications team, led by IBM researcher Bernard Metzler, plans on creating a computing architecture, and, ultimately, a prototype system, for applying real-time communications techniques to the SKA network and supercomputers.

Each of these seven projects aims to solve a critical problem faced by the developers of the SKA system, but it's clear that they will also address challenges facing the science community and technology

industry as we advance into the new era of computing. To make the most of *Big Data*, we need to process data more efficiently, move it faster, store it less expensively and analyze it more effectively--whether it's at rest or on the move.

You can also see the important role that cognitive computing technologies will play in the SKA. The computing system will be too complex to be designed using the old computing paradigms and the SKA data will be too dynamic to be managed using rigid rules and software programming methods. This massive science project requires computers that learn and transform themselves in response to new knowledge and changing requirements.

It's useful to think of the SKA as a scientific sibling to the Large Hadron Collider in Switzerland. There, scientists are studying the tiniest particles for answers to some of the fundamental riddles of existence. Or as Luijten puts it, "At CERN they are creating lots of mini-Big Bang's to understand what happened 13 billion years ago. With the SKA we are detecting signals from the actual Big Bang."

For the people working on SKA, the whole universe is their laboratory. "With the SKA we will be able to fill big gaps in our knowledge of the universe," says Albert-Jan Boonstra, the scientific director of DOME for ASTRON. "We'll be able to map the so-called 'dark ages,' the epoch of ionization, when the stars and galaxies formed." [6]

Big Data offers the potential of tremendous, even earth shattering, advances in the ways humans use information. Global communities of scientists and other experts are gathering huge data sets and building sophisticated models for understanding natural and human phenomena. In a sense, with this data and with the tools to understand it, we're creating a collective intelligence that can be shared and used by many for the betterment of all humankind. That's the ultimate promise of *Big Data*, which the DOME project helps to fulfill.

---

[6] Albert-Jan Boonstra, ASTRON, interviewed Oct. 16, 2012.

Background Information

Articles from IBM experts about the project:

http://www.research.ibm.com/news.shtml

http://ibm.co/HPgK18

Media contacts:

Michael Kiess
IBM Marketing and Communications
Mobile: 49-171-4921178
Phone: 49 7031 16 4051
E-mail: kiess@de.ibm.com

Grit Abe
Media Relations
IBM Research – Zürich
Mobile: +41 77 436 79 91
Phone: +41 44 724 80 60
E-mail: gri@zurich.ibm.com