Surviving the flood

Planned big-science facilities are set to generate more data than all the global Internet traffic combined. **Jon Cartwright** finds out how scientists will deal with the data deluge

SKA Organisation

Data hungry The Square Kilometre Array, currently being built in southern Africa and Australasia, will produce more than 250 000 petabytes of data every year – enough to fill 36 million DVDs.

When the €2bn ($2.6bn) Square Kilometre Array (SKA) sees first light in the 2020s, astronomers will have an unprecedented window into the early universe. Quite what the world's biggest radio telescope will discover is of course an open question – but with hundreds of thousands of dishes and antennas spread out across Africa and Australasia, you might think the science will be limited only by the enormous extent of the telescope's sensitivity, or its field of view.

But you would be wrong. "It's the electricity bill," says Tim Cornwell, the SKA's head of computing. "While we have the capital cost to build the computer system, actually running it at full capacity is looking to be a problem." The reason SKA bosses are concerned about electricity bills is that the telescope will require the operation of three supercomputers, each with an electricity consumption of up to 10 MW. And the reason that the telescope needs three energy-hungry supercomputers is that it will be churning out more than 250 000 petabytes of data every year – enough to fill 36 million DVDs. (One petabyte is approximately 10^{15} bytes.) When you consider that uploads to *Facebook* amount to 180 petabytes a year, you begin to see why handling data at the SKA could be a bottleneck.

This is the "data deluge" – and it is not just confined to the SKA. The CERN particle-physics lab, for example, stores around 30 petabytes of data every year (and discards about 100 times that amount) while the European Synchrotron Radiation Facility (ESRF) has been annually generating upwards of one petabyte. Experimental physics is drowning in data, and without big changes in the way data are managed, the science could fall far short of its potential.

Data drizzle

Cornwell says that he can remember the start of his astrophysics career at the UK's Jodrell Bank Observatory in the late 1970s, when staff could print out every data point from an experimental run on a sheet of paper six metres long. As sample sizes have inflated, however, experimental groups have been forced to overcome problems associated with taking the data – whether storing it, processing it or transferring it from one place to another. "Over 30 years, each of those has been a factor at some point," Cornwell says.

At the SKA today, being able to process data without blowing the energy budget is the key concern. While the actual number of data falls with each processing step, Cornwell and his colleagues still have to perform careful computer modelling in order to determine exactly which experiments will be possible. Some will not – and electricity costs will be to blame. "This is what people predicted five years ago – that capital costs would be exceeded by the running costs," Cornwell notes.

The problem at the SKA is not simply down to the number of data being generated. Unlike many other experimental facilities, the SKA's data will be coming from disparate sources – dishes and antennas – that are spread over much of the southern hemisphere. As a result, the data must be collated before anything else can be done with them. If the data originated at roughly the same place, on the other hand, other possibilities for streamlining would have opened up. That is true at CERN, which in 2006 launched a special computing network to farm out data from its Large Hadron Collider (LHC) to labs around the world for processing, thereby avoiding the need for costly, on-site number crunching. Today, the Worldwide LHC Computing Grid consists of more than 170 computing centres in 40 countries and played a vital role in the discovery in 2012 of the Higgs boson.

Despite the success of the Grid, CERN is concerned about what the future holds for data-intensive computing. In May, a public–private partnership between CERN and various computing companies called CERN openlab produced a white paper, "Future IT Challenges in Scientific Research". The paper outlined six main challenges: how to extract data, and how to initially filter them; the best types of computing platforms and software to handle the data; how to store the data; where to find the computing infrastructure, whether it is on-site, over a CERN-type grid, or in a Internet-shared "cloud"; how to transmit data; and how to analyse them efficiently.

Physics institutions feel the pressure of these challenges differently. At the ESRF, where X-ray data must be recorded within a confined region around a sample, engineers have to extract one gigabyte of data per second – a tiny fraction of what is possible at CERN or the SKA. However, even that relatively small amount is tricky to handle. The synchrotron was originally mandated to give visiting scientists all the raw data that they generate during their experimental runs, but Andy Grotz, the group leader of software at the ESRF, says that aim is no longer realistic, and that they must reduce it. “We suffer from the data deluge, in that we almost cannot keep up,” he adds. “We have to change the way we work radically.”

Often little is lost by reducing raw data. For instance, thousands of X-ray images may be needed to define a crystal according to the most common parameters – orientation, strain and so on – but, once those parameters have been calculated, the raw data are, for many visiting scientists, superfluous. The trouble is how to reduce the raw data when the requirements of visiting scientists can be so varied. Grotz says the ESRF has in the past allowed its computer scientists to help visiting scientists reduce data “on a goodwill basis” so that the files are small enough to be taken home on a USB stick or any other convenient medium. Now, he says, the lab is looking at ways of formalizing the process because the raw data sets are always too unwieldy.

One option – which Grotz and others have submitted as a research and innovation project to the European Commission (EC) under its Horizon 2020 funding programme – is to set up a cloud-computing facility with other European science institutions for the express purpose of data reduction. The key requirement of such a system would be that it is easy to use, even for those scientists who are not computer-savvy. “More and more users want to be able to use this as a turnkey system, where they can provide a sample and get out data that they understand,” says Grotz.

Different requirements

If it goes ahead, Grotz and colleagues’ Horizon 2020 project would follow on from Cluster of Research Infrastructures for Synergies in Physics (CRISP). This project, which has run for three years under backing from the EC, brought together 11 European research facilities, including the SKA, the ESRF and CERN, to tackle all aspects of the data deluge. CRISP has had some successes, such as finding new ways to extract data quickly from detectors, but Grotz says other targets – such as automatically storing the contextual data (or “metadata”) from experiments – has proved difficult because of the innate differences between research institutions.

Differences between hosting institutions may not only be practical. Bob Jones, the head of CERN openlab, believes the recent scandals of how governments can tap into private data have galvanized people into thinking about who should have access to what data. Science is competitive, he says, and groups that have helped fund an experiment may be concerned if that experiment’s data are farmed somewhere else for processing, because it might allow non-participating groups to sneak access. Some data could even carry a political or security risk, he says – a satellite’s image of a war zone, for instance.

Legislative answers to such problems could hinder collaborative computing efforts or they could streamline it, says Jones. But whatever happens, he says, there needs to be a collective decision. “There are a number of interests, but really it boils down to Europe deciding what the rules are for accessing data, rather than having them imposed on it by a third party.”

When it comes to the data deluge, it seems, staying above water will not be easy. Grotz says that a more general problem is financial, in the sense that computing is often bottom of the list for managers who are budgeting experimental infrastructure. “By the time we get to the software and computing infrastructure, the money has usually run out,” he says. A change of mindset is needed, but Grotz thinks that we are still in that antediluvian world where just generating the data is the priority. “It’s like we’re still working with slide rules,” he says.