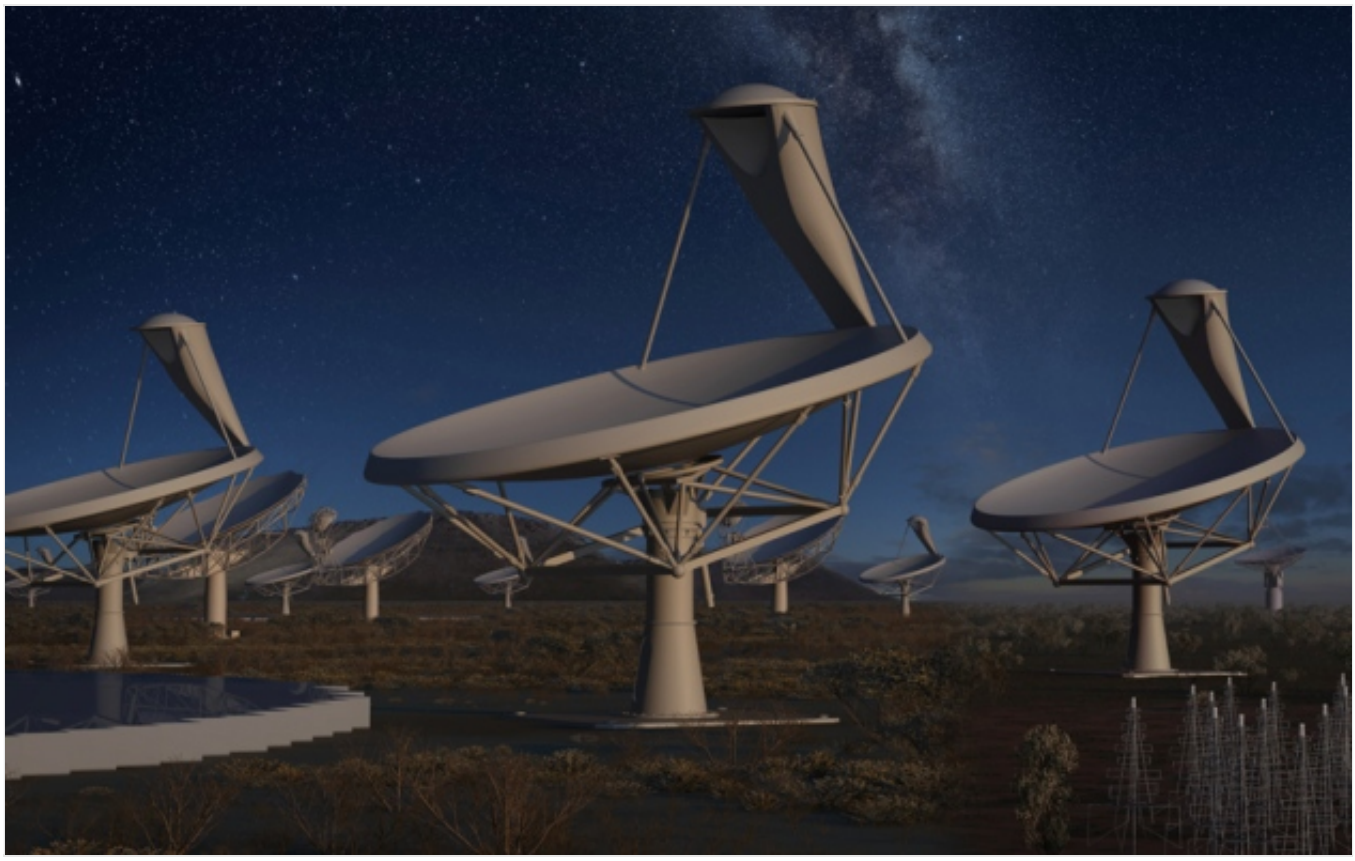
Cloud computing beckons scientists



SKA Organisation

Data from the planned Square Kilometre Array of radio telescopes will require vast computing resources.

Sometime in the next decade, the Square Kilometre Array (SKA) will open its compound eyes — roughly 2,000 radio dishes divided between sites in South Africa and Australia. The radio telescope will then begin staring into supermassive black holes, searching for the origin of cosmic magnetic fields and seeking clues about the young Universe.

Meanwhile, the telescope's engineers are struggling to plan for the imminent data deluge. The photons that will stream into the array's receivers are expected to produce up to 1 exabyte (10^{18} bytes) of data per day, roughly the amount handled by the entire Internet in 2000. Electricity costs for an on-site computing

cluster big enough to process those data could total millions of dollars each year. So the engineers are investigating an increasingly common choice for researchers wrestling with big data: to outsource their computing to the cloud.

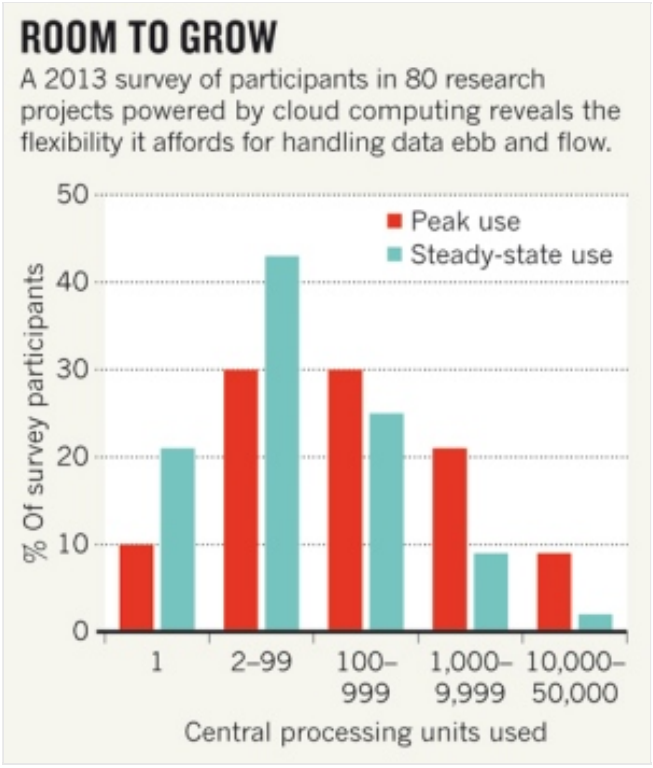
“No one’s ever built anything this big before, and we really don’t understand the ins and outs of operating it,” explains SKA architect Tim Cornwell of the Jodrell Bank Observatory near Manchester, UK. He says that cloud systems — which provide on-demand, ‘elastic’ access to shared, remote computing resources — would provide an amount of flexibility for the project that buying dedicated hardware might not.

Such elasticity can also benefit projects that involve massively parallel data analyses, such as processing and aligning billions of DNA base pairs, or combing through hundreds of photos to identify specific zebras from their stripe patterns. It is also a boon to scientists who require bursts of computing power rather than sustained usage, as do researchers looking at seismic data in the aftermath of an earthquake.

“The rest of the year, when there are no earthquakes happening, they’re just paying for storage,” says David Lifka, director of the Cornell University Center for Advanced Computing in Ithaca, New York, which runs a computing-platform service called Red Cloud.

But the economics of cloud computing can be complex. An ongoing price war between major providers such as Google, Microsoft and Amazon Web Services has cut costs overall, but in many cases, sending data to the cloud, or retrieving them, remains much more expensive than storing them. Amazon’s Elastic Cloud Compute S3 service charges US customers as much as US$0.12 per gigabyte for transfer from its servers, but no more than $0.03 per month to store the same amount of data.

This comes as a surprise to many researchers, according to a 2013 US National Science Foundation survey of 80 science projects that rely on the cloud (see



SOURCE: XSEDE Cloud Survey

'Room to grow'). "Some cloud billing mechanisms are really opaque," says Daniel Perry, director of product and marketing for Janet, a private, UK-government-funded group near Oxford that is working to link British educational facilities to a shared data centre. "Unless you know what you're doing, you may find that you've run out of your grant in three months."

And costs aside, the cloud will probably never suit some computer projects, such as 'deep learning' networks that seek to mimic how the human brain learns. Adam Coates, a computer scientist at Stanford University in California who is involved in such work, says that these systems can require rapid information transfer between billions of connections — something not possible with the cloud. Instead, Coates relies on a dedicated on-site computing cluster. "Having that very high-speed communication is key," he says. "We want vast amounts of computation, but we don't really care about elasticity."

The cloud's dependability is also a concern, says Ken Birman, a computer scientist at Cornell. "It isn't known for being secure, and it isn't known for being extremely reliable." But not all researchers require foolproof data encryption or super-fast, reproducible computations.

For example, CERN, Europe's particle-physics laboratory near Geneva in Switzerland, assembled an in-house cloud to handle the data generated by the Large Hadron Collider. "The CERN data are public data, so we don't have any security concerns," says Tim Bell, who directs the centre's infrastructure and operating services. Instead, CERN focused on providing physicists with an

efficient computing platform. “In the past, when they asked for physical hardware, they were waiting for weeks,” Bell says. “Now they can ask for a virtual machine and get something in the time it takes to have a cup of coffee.”

Universities are also getting into the cloud business. At Cornell, a subscription to Red Cloud costs $400 for 8,585 processing hours; for off-campus scientists, the same subscription is $640. Such on-campus services often appeal to researchers who are not ready for the do-it-yourself nature of commercial providers, which often requires expertise in programming, testing and debugging code. By contrast, Cornell cloud specialists are on site to help researchers using Red Cloud. “The thing you can’t get with commercial clouds is hand-holding,” Lifka says.

Meanwhile, companies such as Microsoft have set up cloud training specifically for academics, addressing issues such as data sharing and security, scientific reproducibility and how funding agencies may view the cloud. “A lot of the training and education content was tuned to a business audience. That meant the on-ramp for researchers was a bit more tricky,” says Daron Green, senior director of Microsoft Research Connections. “We realized there was pretty much a latent demand within the research community.”