The Square Kilometre Array and Data Intensive Radio Astronomy

Erik Rosolowsky[1] (U. Alberta)
Bryan Gaensler (U. Toronto, Canadian SKA Science Director)
Stefi Baum (U. Manitoba)
Kristine Spekkens (RMC/Queen's)
Jeroen Stil (U. Calgary)

Summary

Canada is a partner in the Square Kilometre Array (SKA) project, which aims to build the largest and most powerful radio telescope yet constructed. The telescope will consist of multiple antennas spread across South Africa and Australia. The Canadian Astronomical Society (CASCA) recently identified securing SKA construction funding as the top priority ground-based effort for the community[2].

Modern radio telescopes are no longer being driven by the telescopes/antennas that collect the signal but rather by the hardware and software that correlate and analyze the signals received. Telescopes are evolving to become huge correlation and data mining machines, which push computer science, data storage, visualization techniques, and analysis algorithms to their limits. The SKA project poses a challenge: how can we manage the rich data stream from the telescope and extract scientific meaning from the exabytes of data produced every year by the facility. Precursor facilities to the SKA will start producing data over the next year and SKA early science will begin while the array is still under construction as early as 2020. Phase I of the SKA (Figure 1, below) is expected to be completed in 2023, where it will produce science-ready data for analysis at rates of up to 3 PB/day. Since data management is an integral part of the SKA design, the telescope will be designed to deliver these huge data sets to the research groups of astronomers. Beyond the basic data delivery however, there is still a large gulf between the calibrated data and scientific breakthrough. The tools and facilities required to mine these data sets for scientific insight still need to be developed. Without such investment, Canada will see little scientific return on a projected $60 million investment. In this whitepaper, we describe how Canadian users will rely on Compute Canada resources for scientific analysis and storage of SKA data at rates of tens of PB per year. The SKA thus represents a long-term strategic consideration for Compute Canada.

Scientific Motivation

SKA science is built around 5 main themes, where a generational improvement in telescope sensitivity will lead to groundbreaking discoveries.

---

[1]We are a group of university scientists, each engaged with different aspects of SKA planning and development.

[2] http://casca.ca/wp-content/uploads/2016/02/MTR_draft_v1.12.pdf

- *Galaxy evolution, cosmology and dark energy* -- Using a spectral line of atomic hydrogen, the SKA will map billions of galaxies across a large fraction of the visible Universe. With a rich, complete map of galaxies now and into the distant past, we will

Figure 1: Artist impression of a few of the thousands of antennas that will make up the SKA. Stations of these dishes will be sited across southern Africa. They will be linked with telescopes across the continent and to Australia, making a single, globe-spanning radio telescope that is orders of magnitude more sensitive that current facilities. Image credit: SKA Organization

gain new insight into the nature of dark matter and dark energy. We will also understand how the broader evolution of the Universe shapes the evolution of galaxy populations.
- *Strong-field tests of gravity using pulsars and black holes* -- One of the major discoveries of radio astronomy was finding *pulsars*, the ultracompact remnants of massive stars. Pulsars have characteristic masses about that of our Sun packed into a size 20 km across. They are often spinning, giving rise to pulses of radio emission. This combination of good timing from the spin and the extreme densities means they can be used as excellent clocks, probing the nature of space and time. The SKA will the first few systems in exotic configurations such as a pulsars orbiting black holes, providing tests of the theory of relativity and hopefully the first clues of the physics beyond relativity.
- *The origin and evolution of cosmic magnetism* -- Magnetic fields permeate the Universe, holding a sizeable fraction of energy in galaxies. These fields are difficult to study because their observational signatures are weak. With the exquisite sensitivity of the SKA, we can finally address open questions about magnetic field origin and its influence on the structure and ongoing evolution of the galaxies and galaxy clusters.
- *Probing the cosmic dawn* -- The SKA will provide a unique observational view on the time between the Big Bang and the first generation of stars, the so-called Cosmic Dark Ages. By studying how the Universe lights up with the first stars and galaxies, we will finally see the missing link between the origin of the cosmos and the systems we observe around us. In that link lies the answers to why we do not see the remnants of

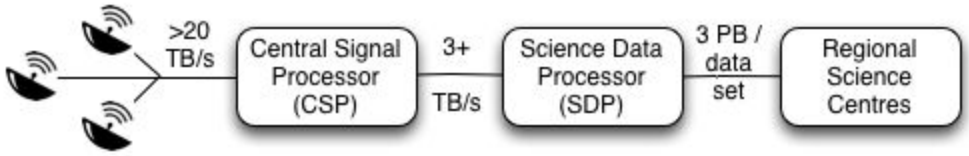
the first generations of stars around us today and what was driving the first light in the Universe.

- *The cradle of life* -- The past two decades have revealed planets to be ubiquitous through the Milky Way Galaxy. We will use the SKA to map thousands of forming systems, tracking planet formation through critical phases of assembly. The SKA will also be sensitive to heavy molecules, exploring the cosmic chemistry as vehicle for bringing pre-biotic molecules to planets. Finally, the SKA can search for leakage radio signals from extraterrestrial civilizations

While these five areas are the design targets of the instrument, the role of serendipity cannot be ignored. Experience has shown that every generational improvement in telescope technology has led to unintended discoveries that usher in a deeper understanding of the Universe. The scientific promise of the facility is great, but the SKA is fundamentally driven not by astronomical instrumentation but by computation. This driver makes meeting the SKA data challenge one of the critical design considerations for the success of the facility.

The SKA Data Problems and Solutions

While the total data rate within the telescope will exceed current data rates for the internet (> 20 TB/s), this data will be assembled into a correlation system which will “only” produce scientific data rates of 3 TB/s. These data will then be run through a standardized imaging pipeline to produce scientific studies with volumes on the 1-10 PB scale. In the language of the SKA, the Central Signal Processor (CSP) digests the raw data flow, which is then calibrated by the Science Data Processor (SDP) into scientifically useful data. The SDP is responsible for providing the final data products to the astronomical users. The figure below shows a schematic of the SKA data flow.



Canadian Participation

Canada is already heavily invested in the SKA data challenge. The Central Signal Processor effort is being led out of the National Research Council facilities at the Dominion Radio Astrophysical Observatory (DRAO) in Penticton, BC. Canadian universities and the Canadian Astronomy Data Centre (CADC) are members of the Central Signal Processor and Science Data Processor Consortia, with efforts on developing data processing and distribution methods, science portals, and remote visualization software.

Current Use of Research Computing

Since the SKA is still under development, the full scale of the SKA data challenge has not been borne by the research community or Compute Canada. However, CC infrastructure is still an integral part of SKA preparation activities as well as analysis activities on precursor facilities. As part of the RPP programme, CC is supporting the CyberSKA, an online portal that serves access to data and analysis tools for radio astronomy researchers. The CyberSKA was developed in Canada and is being deployed in a federated system globally. The Canadian node will be hosted on the CC cloud systems. The Central Signal Processor team actively uses the CyberSKA portal to coordinate their efforts and the CyberSKA serves as one of the technologies used for delivery by the Science Data Processor team.

CC HPC resources are also used for the processing of radio astronomy observations. For example, astronomers use the general purpose machines (e.g., orcinus, jasper on WestGrid) to process image data and remove imaging artifacts from interferometry data. Typically, processing a large (though not exceptional) observational data set as would be used in a regular observing program can take 1-4 core-years of time with typical data impacts at the TB scale. However, the current processing is bottlenecked by several single-threaded stops and relies on large amounts of RAM per core (32 GB is recommended). Processing is further restricted by I/O with transfer to/from (lustre) storage occupying up to half the processing time. This processing is being used to guide SDP development.

Future Requirements

Since the principal technology challenges of the SKA lie in the exceptionally high data rates, the data management plan is a central part of the telescope design. However, most of data management focuses on how to provide calibrated imaging and time-series studies of the radio sky with a quality that suits scientific needs. Beyond the basic data delivery, there is still a large gulf between the calibrated data and scientific breakthrough. The astronomical community will need to digest the planned dataset sizes in the 10s of PB. Moreover, a given data set can be subject to a multitude of different processing approaches, each targeting a different scientific outcome. It seems likely that the community will organize into large scale collaborations to exploit these studies for a variety of different outcomes, converging toward an organization structure like that already used on the largest experiments in particle physics.

To facilitate the optimal management of these resources, the SKA project and the Canadian SKA users advocate the development of a (potentially virtual) Canadian SKA Data Centre, ideally partnering with Compute Canada to deliver functionality to suit the scientific community. Creating such a data centre would take place as part of the Canadian efforts on the SDP. Chris Loken (SciNet) serves on the SKA Organization's Data Flow Advisory Panel, providing input on these strategic directions. The SKA data distribution plan may also require developing

interoperability between HPC resources in different regions, sharing the computational burden across the globe using multiple agencies.

Outside the plans for any future data centre, the Canadian community would certainly still rely on CC resources for processing survey data. In particular, access to the HPC resources would serve as integral parts of the science plans for smaller research teams and in HQP training. Even projects using small fractions of survey data sets create computational needs that outstrip institutional resources.

Data

Operating in its most data intensive mode, the SKA Phase 1 can deliver 3 PB/day of fully processed observations that require scientific analysis. It is anticipated that this flow will be stopped down to whatever level will be manageable downstream. However, at the outset of operations (2023), scientific collaborations will need to plan on the analysis of ~1-10 PB size datasets. Even anticipating for continuous growth of computing and storage capabilities, these 3 PB scales will be challenging for individual researchers to manage. Thus, the data sets will likely be analyzed using shared, national-scale computing resources. Precursor facilities and early science will begin producing data sets at the 1/10 scale as early as 2018.

Software

There is a large need to improve the quality of the scientific software required to analyze SKA data sets. However, for many applications, the computational problem is easily parallelized. The SKA data sets typically represent images of the (polarized) radio emission over the sky (two dimensions) as functions of frequency and time. These data are produced at high resolution, leading to large data sets. However, these domains can be divided and processed independently (a MapReduce model) to take advantage of parallel processing.

Substantial efforts are needed to develop domain specific software for large-scale radio astronomical data sets that operate on HPC. The astronomical community will also need to move into hierarchical data formats, which allow for efficient access of proximal data in any of the domains. These efforts must take place within the astronomical community with partnerships in other disciplines and industry, but a close partnership with Compute Canada is required for testing and deployment.

Processing

The amount of processing required to exploit the SKA data volume is difficult to estimate since the current software available for that analysis is not designed with large data volumes in mind. However, extremely rough estimates can be gleaned by exploring a simple analysis algorithm (such as source identification and characterization) on a small data set and assume that the time scales linearly with data volume. From this scaling, a single analysis of a 10 PB study will take 100 core years with 2016 processors. We anticipate 10s of comparable data sets being of

interest to the Canadian community and the need for processing the same data set multiple times. A typical year of SKA data could require 10,000 to 30,000 core-years of processing for the simple scientific analyses.

Networking

Transport of the data to the scientists remains an open question for the SKA telescope, but falls into the domain of the Science Data Processor. Minimizing redundant data transfer is part of the design goal for the telescope, but the regional data centre model being explored would require transport of the scientific data to the appropriate centre. If there is a Canadian regional centre developed, it must be in collaboration with Compute Canada. As such, the facility would require intake of ~100 PB of data per year and furnish processing as described above. However, once transported and analyzed, only a small fraction of the incoming data volume would move back out to astronomers. Canadian teams are also developing remote visualization solutions to allow users to explore large data sets within the regional centres. We anticipate partnerships with both Compute Canada and CANARIE to meet networking challenges for the SKA.