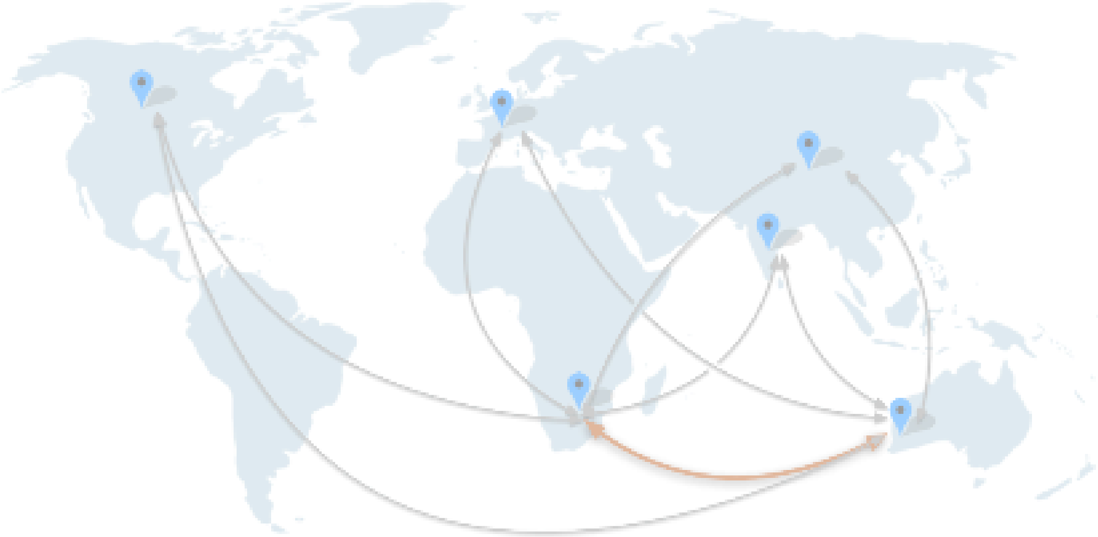
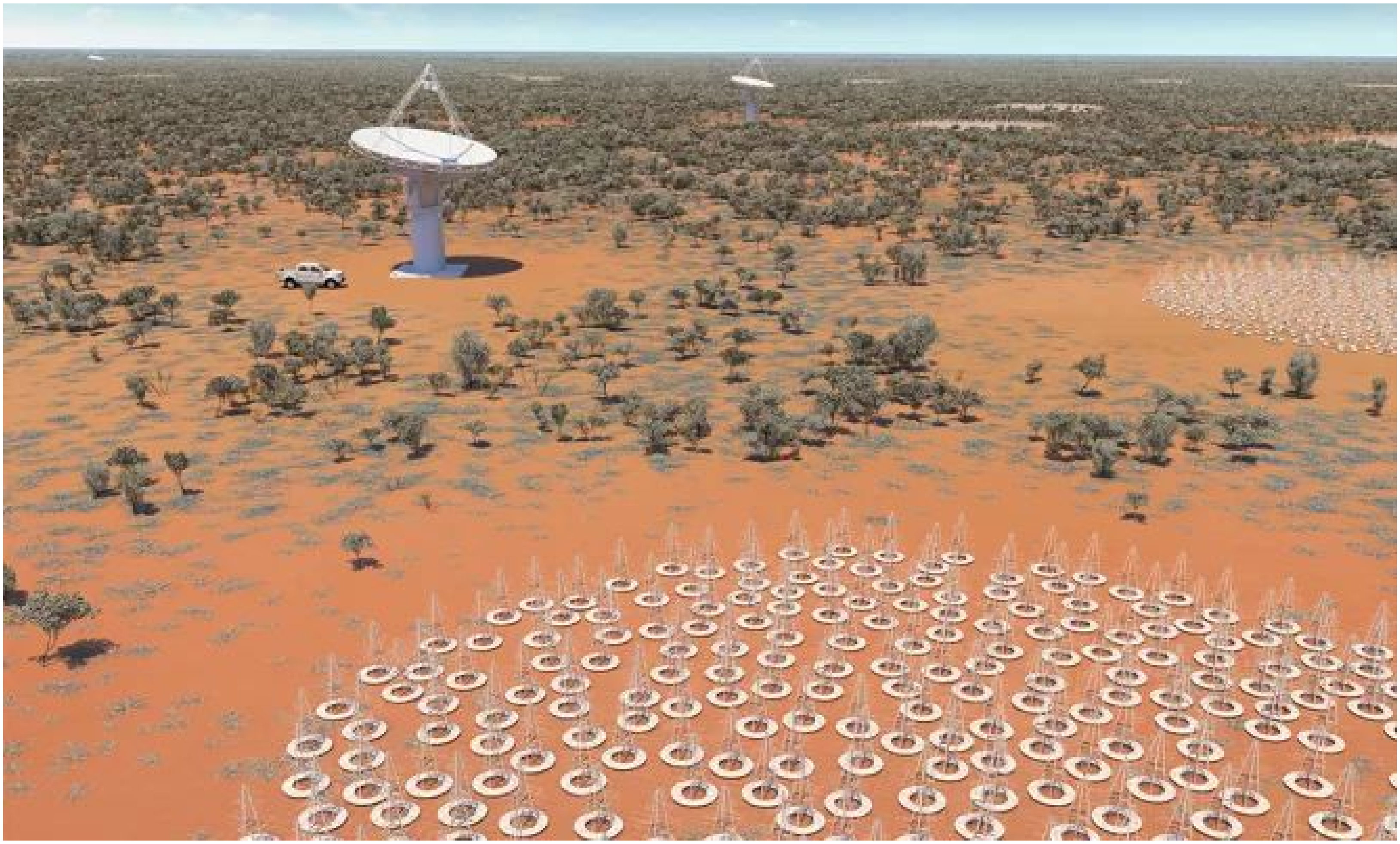
https://aussrc.org.au/

Australian SKA Regional Center

Technology Brief for Industry



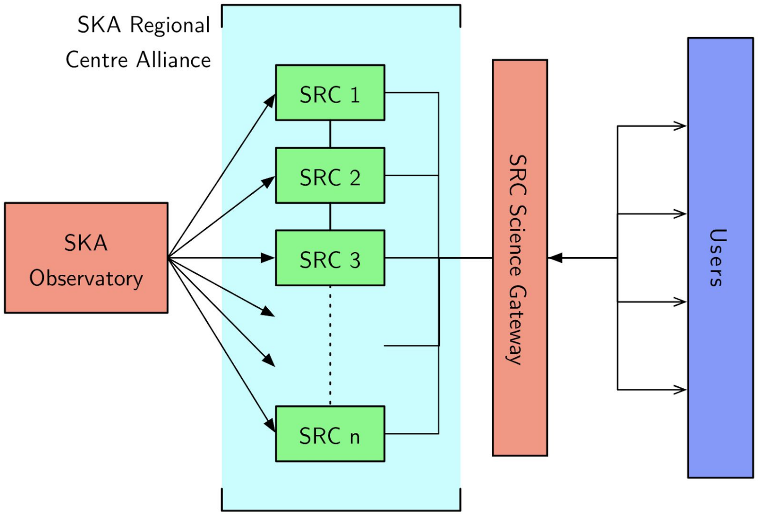


SKA-low and ASKAP artist impression

Square Kilometer Array

The first phase of the SKA Observatory will include two instruments: SKA1-Low that will be constructed in Australia (operating between 50-350 MHz), and SKA1-Mid in South Africa (operating between 0.35-14 GHz).

SKA1-Low will be comprised of 131,072 antennas distributed over 512 stations, while 197 15-m dishes will comprise SKA1-Mid. By the time the SKA Phase 1 is in steady state operations in the mid-2020s it will generate a raw data rate close to 1 TB/s. Although the raw data will be processed by the Central Signal Processing (CSP) and the Science Data Processing (SDP) subsystems of the SKA, the output of data products to be disseminated to science teams are estimated to be 250-300 PB per year for the two arrays combined.

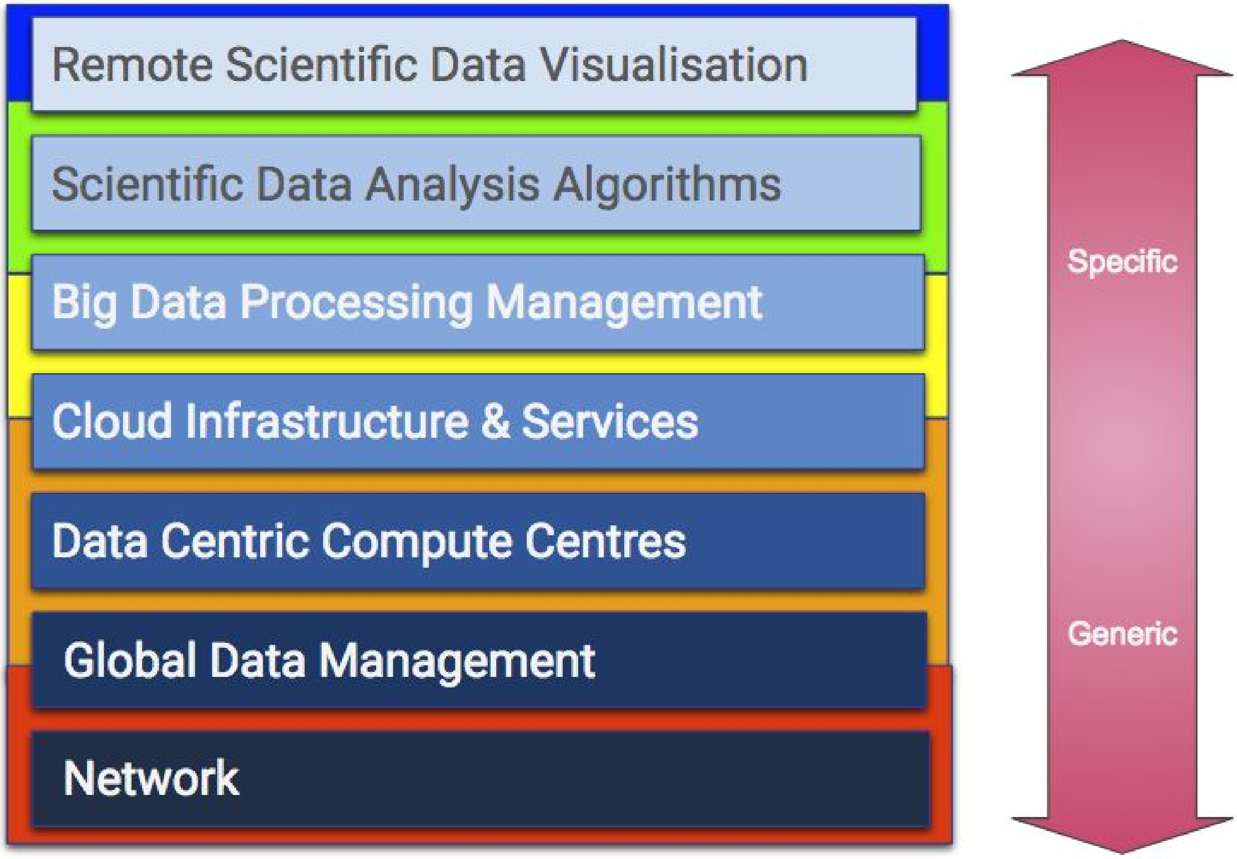


SKA Regional Centres

A network of SKA Regional Centres will receive science data products from the SKA Observatory. Access to SKA science data products, as well as the tools and processing power necessary to fully exploit the science potential of those products, will be provided via a Science Gateway. Access to science data products will be irrespective of a SKA user's geographical location, or whether their local region or country hosts an SRC.

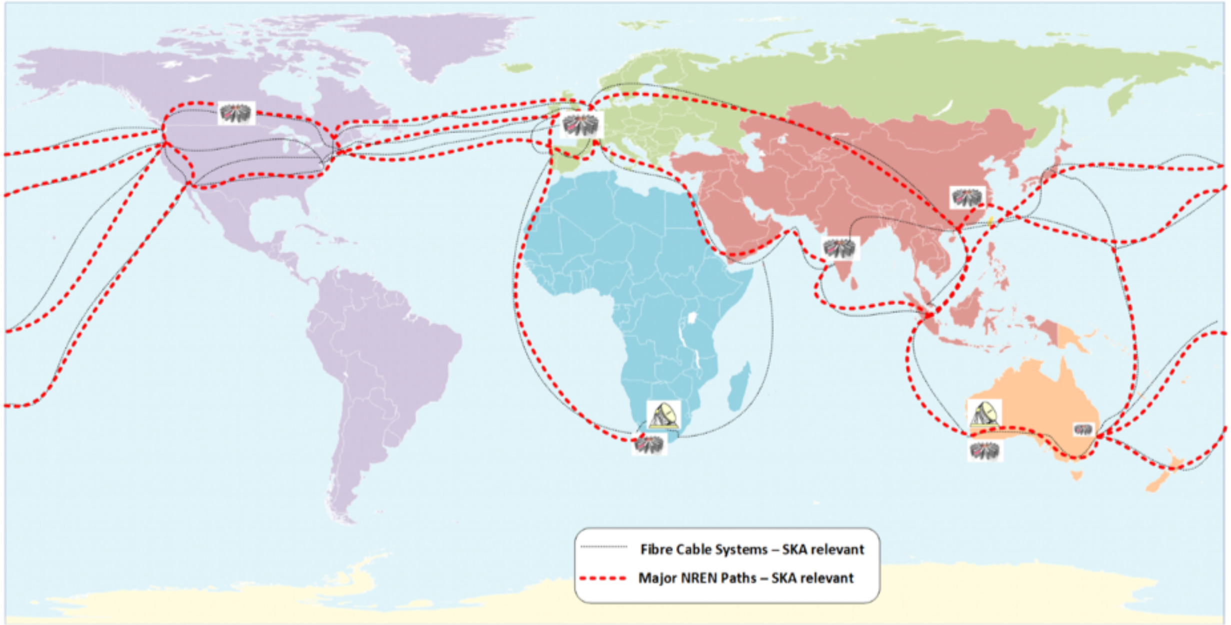
The key objectives of SRCs are to:

- Provide long-term, persistent storage capabilities.
- Provide sufficient computational resources to support processing and analysis of SKA data by the astronomical community at the appropriate scales and with reasonable latency.
- Provide long-term data management and curation including: metadata allowing easy data discovery; examination of data provenance; and combination with other existing, relevant data sets.
- Provide security and data protocols capable of supporting a wide range of access paradigms from fully open access public datasets to proprietary data for individuals or consortia.
- Provide porting and maintenance of the necessary radio astronomy software stack to the cloud platform.
- Provide documentation, training, and user support for SKA researchers.
- Provide the environment that enables innovation in research and successful collaboration.


Remote Scientific Data Visualisation
Scientific Data Analysis Algorithms
Big Data Processing Management
Cloud Infrastructure & Services
Data Centric Compute Centres
Global Data Management
Network
Specific
Generic

SRC Technology Development Themes

Regional Centres will receive 250-300 PB of data per year from the SKA Observatory. This data will need to be: distributed, archived, curated, managed, and processed. The higher level science data products will need to be analysed, visualised, and scientific knowledge will need to be extracted.


Fibre Cable Systems – SKA relevant
Major NREN Paths – SKA relevant

Credit: Richard Hughes-Jones, GEANT

Network

SKA1-Low will send data products to both SRCs in Australia and globally. In addition the Australian SRC will receive data from SKA1-Mid. The SKA SDP subsystem will be sending data to multiple locations simultaneously over long distances at high-data rates (initially < 100Gbit/s but scaling above this into the future). This requires two major components:

- **ingest**: sustained high-bandwidth, multiple streams with access to compute, storage, archive and possibly visualisation.
- **egress**: secure, high-bandwidth remote access to data products, and archive (compute and visualisation).

The network is the underlying fabric will provide access to compute, visualisation, storage and archives. Importantly users should not need to worry about where the data is stored and what computing and related resources are required (i.e. location independent) so they can focus on the science and not the infrastructure.

Consider developing a **Smart** Geo-Aware Compute, Visualisation, Storage and Archive Infrastructure:

- which knows about the network topology (relatively static),
- that has access to real-time fault/outage/downtime notifications (from third-party providers),
- which can monitor the network (plus CPU, archive space etc) in near-real time (initially build on existing SDN and SDS architectures),
- that can make *informed* decisions based on these parameters either automatically and/or via an “advanced” mode to allow the user to choose their preferred option,
- which is likely to have applications in other fields (e.g. geophysics, embedded streaming clients).

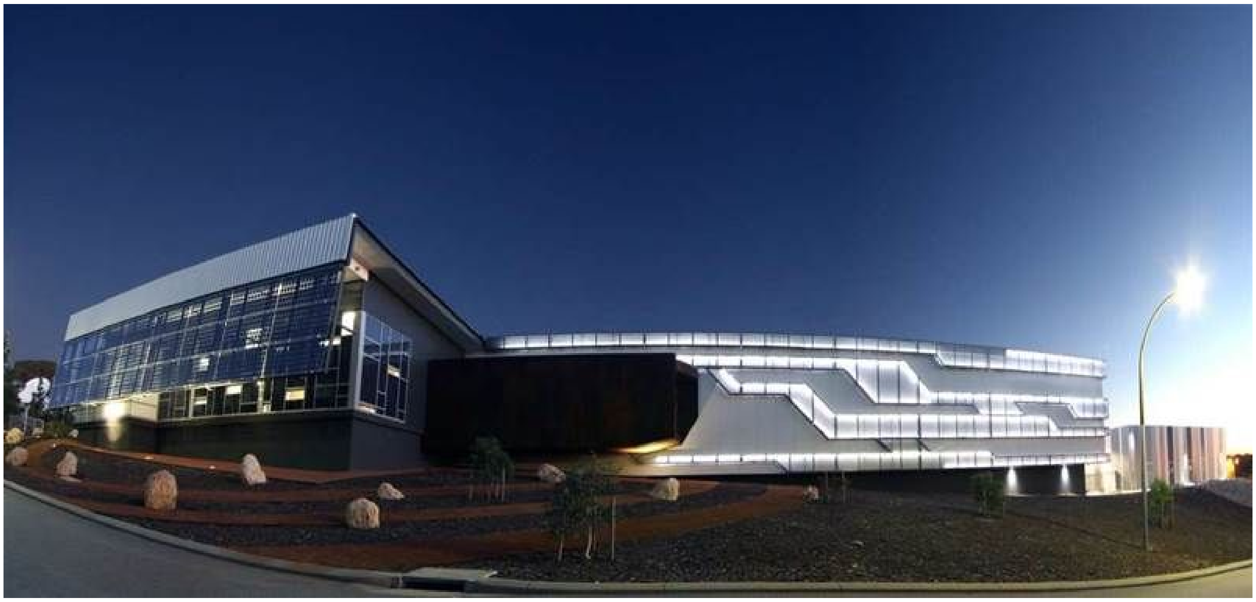
Global Data Management

The SKA data products and thus the Regional Centres will be produced at a rate of up to 300 PB/year. MWA is currently operating at around 5 PB/year and will soon be joined by ASKAP for a total of 10 PB/year.

Full global mirroring of data volumes of this size is not very economic and might even be infeasible due to constraints on availability of sufficient bandwidth. Other schemes, where just certain data sets of interest to the regional community are mirrored are far more complex and would also generate more operational dependencies between the central SKA archives and the various Regional Centres. In addition there are a number of requirements asking to protect the SKA data products from loss due to local disasters and thus a certain level of resilience and thus redundancy will need to be implemented. Using the Regional Centres to provide both access to data sets of interest as well as the required resilience seems to be an interesting alternative to dumb off-site backups, but the implementation and operation of such a scheme adds a lot of complexity.

The issues that need to be addressed are:

- Discovery, retrieval, processing and sharing of data
- Software and/or appliances for data replication/mirroring/distribution
- Global Data Lifecycle Management
- Metadata repositories and global data discovery methods
- Global access control (SSO, AAA)



Data Centric Compute Centres

AusSRC will require a holistic solution to the data intensive computing that addresses the following specific issues:

- Optimized centre for data flowing through the entire system;
- Remote sensor network connected directly to ‘HPC’ centre (memory to memory over WAN);
- Individual sensors in radio astronomy are producing multiple Tb/s;
- Continuous processing (High Throughput Computing, HTC and Many-Task Computing, MTC);
- High efficiency processing to reduce overall costs;
- Support of high input and output data volume and rate (potential up to 1 Exabyte/year);
- Storage tightly integrated with processing;
- Highly efficient cooling and support infrastructure.

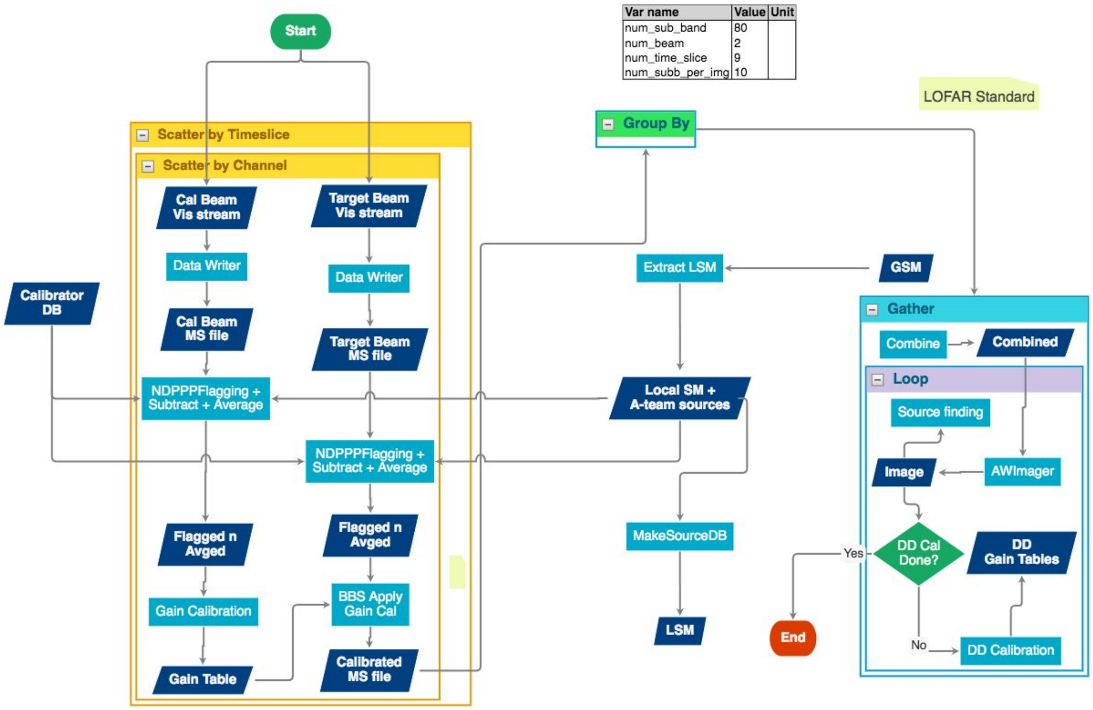
Cloud Infrastructure and Services

The solution for the AusSRC compute infrastructure will be a combination of dedicated HPC and storage resources, and the usage of small and large scale Cloud environments. This leaves the open question, can cloud technologies and providers of services offer a solution that:

- integrates with the HPC and Data Storage;
- provides a simple and effective accounting model;
- and reduces the total cost of ownership for the AusSRC as a whole?

Most large astronomy projects are international collaborations. One team is responsible for one part of the work, whilst other teams are responsible for other parts of the work. A cloud environment allows for flexible e-collaboration with partners able to run their parts of the experiment on dedicated, cloud provided infrastructure through Infrastructure as a Service in geologically disparate locations.

One of the significant problems that technical teams developing the technologies for the future AusSRC is that of legacy code, which has extremely valuable functionally but is “cost prohibitive” to port to different platforms. This requires a new Software as a Service solution utilising the concepts cloud computing and containers with minimum performance penalties.



Var name	Value	Unit
num_sub_band	80	
num_beam	2	
num_time_slice	9	
num_subb_per_img	10	

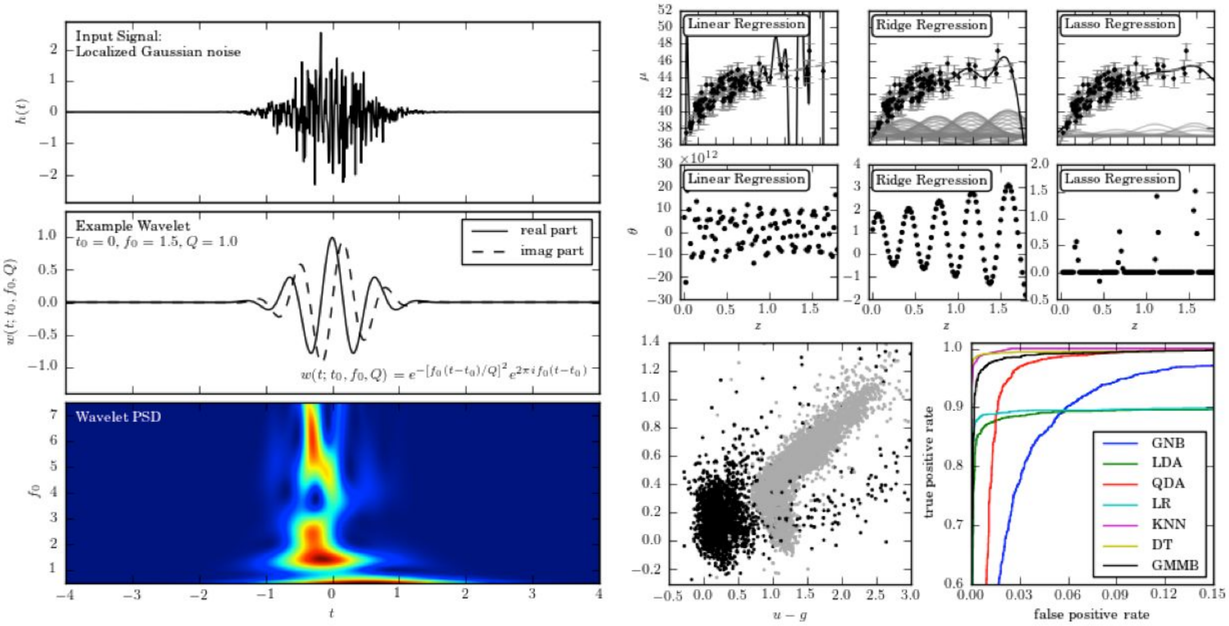
Big Data Processing Management

We are interested in co-designing flexible big data processing (workflow) systems with separation of concerns - processing logic vs physical execution. Such a system would need to support both batch and stream processing, enable managing multiple workflows.

We see this achieved via directed graph-based workflow systems focused on data-centric processing.

For this, we are interested in evaluating existing solutions like and determine whether are fit for purpose or need to be adapted: workflow systems (Taverna, Kepler, Reflex, Yabi…), big data frameworks (SPARK, HADOOP…), real-time streaming frameworks (Storm, AWS Kinetic, Kafka…).

Although web-based scientific workflow management systems would need to satisfy SRC requirements, such a system can also be used in other scientific and industrial domains.

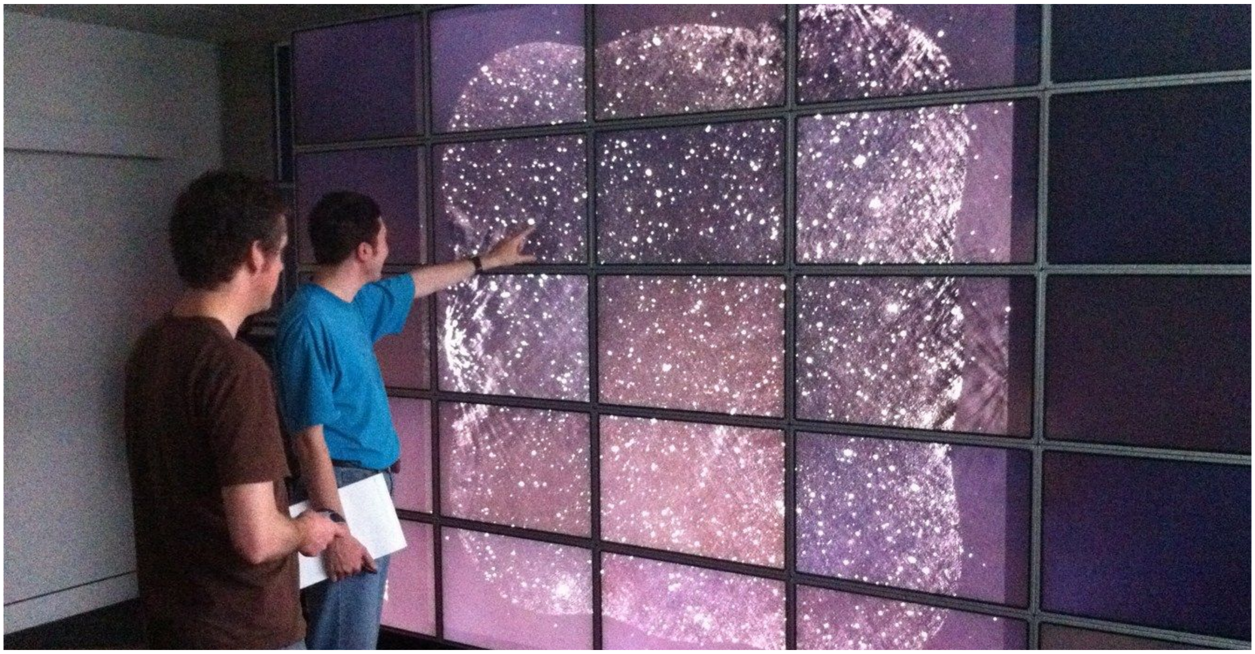


Scientific Data Analysis and Algorithms

Data-intensive algorithms are crucial for the success of the SKA Regional Centre (SRC). Over the years, the (radio) astronomy community has explored these algorithms to solve fundamental problems that are common in many other domains such as medical imaging, geoscience, sensor networks, web intelligence, etc.

Let us examine a few examples of that cross-over. The principle of imaging synthesis in radio astronomy is remarkably similar to Magnetic Resonance Imaging (MRI). More recent radio imaging algorithms exploit the sparsity of signals, and the same algorithms recommend movies to millions of Netflix users. The algorithms used to calibrate antenna gains essentially solve an inversion problem that can be found in many geoscience data pipelines. Multiwavelength data analysis fuses radio emissions with optical, infrared, x-ray images in a way similar to how tweets are correlated with stock prices and other social media to predict economy prospect. The same clustering algorithm that estimates radio galaxy components and peaks may well be adapted for market segmentation and cluster-specific marketing strategies.

This theme aims to establish connections between SRC algorithm capabilities and real-world problems many businesses and verticals are facing. On the other hand, the SRC needs industry contributions to fully realise the model of “Algorithm as a Service” - e.g. how to define and express *Capabilities* and *Limitations* of an algorithm, what is a common user-interaction workflow, how to integrate / leverage data infrastructure, and how to maximise serendipity of scientific / unexpected discoveries from big data.



Remote Scientific Data Visualisation

SKA and its precursor’s scientists need to be able to visualise the data for two main reasons. Firstly to assess the quality of data, and secondly to aid data discovery and knowledge extraction.

However, the data is extraordinarily large with a single multidimensional set of imagery being many TeraBytes in size. This dictates the requirement for visualisation to be remote, streamed, with an advance bandwidth control for multiple clients visualising the same data cube.

Besides, being real data from instruments, the datasets are often noisy, and the most interesting information is often at the noise level. Whatever the technology used to compress the data or provide multiple representations of the same data the impact on quality of data must be measurable.

Most of the commonly used methods to visualise such datasets fail to provide an acceptable solution, and astronomy is not the only case when an effective remote visualisation is required - geological exploration, remote sensing, satellite imagery, medical imaging will face the same problems of increased sizes of imagery.

ICRAR team is currently developing a remote visualisation solution based on JPEG2000, and interested in evaluating other new technologies as well.