ACS presents: The Annual
Dennis Moore Oration & 1962 awards

Managing & Processing Astronomical Data in the Square Kilometre Array

with Professor
Andreas Wicenec

Proceedings

Opening and welcome	Dr David Cook FACS CP
Introduction of opening speaker	Mr Jerome Chiew MACS (Snr) CP
Opening address	Hon Stephen Dawson MLC
Introduction	1962 Prize and Medal Finalists Dr Bob Cross FACS CP
Presentation of 1962 Awards	Professor Alex Reid FACS CP
Introduction 1962 Educators	Dr Bob Cross FACS CP
Presentation of Educator Awards	Dr Brian von Konsky FACS CP
Welcome by event partner	DC Alliance
Introduction of 2024 Orator	Dr David Cook FACS CP
Oration delivered	Professor Andreas Wicenec
Vote of thanks	Dr Brian von Konsky FACS CP
Closing remarks	Dr David Cook FACS CP

Oration Organising Committee

Dr David Cook FACS CP

Dr Bob Cross FACS CP

Dr Brian von Konsky FACS CP

Professor Terry Woodings FACS

1962 Awards Judges

Dr Bob Cross FACS CP

Professor Tanya McGill FACS

Professor Tony Watson FACS

Professor Terry Woodings FACS

The Australian Computer Society Annual Dennis Moore Oration Dinner. The University Club of Western Australia, 16 October 2024.

Orator: **Professor Andreas Wicenec**

We wish to thank our exclusive event sponsor

Annual Dennis Moore Oration Past Orators

Since 2012, to commemorate fifty years of digital computing in Western Australia, the WA Branch of the ACS has invited a distinguished scholar and researcher with a connection to WA to present a lecture on the leading edge of an important and emerging area of information and computer technology.

Previous Orators

Year	Orator
2023	Professor Tom Gedeon
2022	Associate Professor Vidy Potdar
2021	Associate Professor Doina Olaru
2020	No Oration held due to Covid restrictions
2019	Associate Professor Rachel Cardwell-Oliver
2018	Professor Jinbo Wang
2017	Professor Matthew Bellgard
2016	Dr Adrian Boeing
2015	Professor Svetha Venkatesh
2014	Professor Craig Valli
2013	Professor Ian Reid
2012	Professor Andrew Rohl

1962 Prize

From a suggestion of Dennis Moore (and with his strong support) 2012 also saw the setting up of an annual prize for the best graduating student in ICT from a WA university. Although the primary criteria are based on academic performance, the candidates are also judged on their ability to promote their ideas in computing and contribution so far.

Previous winners of the 1962 Prize are:

Year	Winner
2023	Shuang Li, Murdoch University
2022	David Adams & Yuval Berman, University of Western Australia
2021	Alistair Martin, Murdoch University
2020	Samual Heath, University of Western Australia
2019	Jarryd Wimbridge, Edith Cowan University
2018	Taaqif Peck, University of Western Australia
2017	Mark Shelton, University of Western Australia
2016	Dalibor Borkovic, Murdoch University
2015	Michael Martis, University of Western Australia
2014	Anthony Long, Curtin University
2013	Laurence Da Luz, Edith Cowan University
2012	Kevin Adnan, Curtin University

The 1962 Prize finalists for 2024 in alphabetical order are:

Disha Grover – Curtin University

Kathryn Morton – Curtin University

Daniel Rodic – Murdoch University

Tom Sargent – Curtin University

Jocelyn Siswanto – Curtin University

Tai Ngoc Vo – University of Western Australia

1962 Medal

In 2019 the 1962 Awards was expanded to include a new award for the most outstanding candidate who completed Doctoral research (eg PhD) in Western Australia in the field of Information Technology and Computer Science.

Previous winners of the 1962 Medals are:

Year	Winner
2023	Dr Manou Rosenberg, University of Western Australia
2022	Dr Uzair Nadeem, University of Western Australia
2021	Dr Naeha Sharif, University of Western Australia
2020	Dr Anupiya Nugaliyadde, Murdoch University
2019	Dr Qiuhong Ke, University of Western Australia

The 1962 Medal finalists for 2024 in alphabetical order:

Dr Hezam A Albaqami – University of Western Australia

Dr Martin Dart – Edith Cowan University

Dr Robert Herne – Murdoch University

Dr Muhammad Ibrahim – University of Western Australia

Dr Muhammad Malik – Edith Cowan University

Dr Sayma Shammi – Murdoch University

Dr Tanmay Singha – Curtin University

1962 Educator Recognition

New in 2023, the 1962 awards will recognise teachers and lecturers that have received awards from other organisations. These are the dedicated people who make the other awards possible.

The 1962 Educator for 2024 being recognised is:

Brett Clarke MACS CP – Catholic Education

Dennis Moore AM MA (Cantab) FACS:

Dennis Moore was born in NSW in 1937. He was educated on scholarships at The King's School, Parramatta where he was captain and dux of the school, and at Queens' College Cambridge where he graduated in 1958 in mathematics.

After a period with commerce and industry in computing and operations research in NSW, he pioneered computing in Western Australia, installing the first computer at UWA in 1962. He introduced WA's first computing qualification – the DipNAAC – at UWA.

In 1965, he was responsible for the purchase and installation of the DEC PDP-6. This was the world's first commercial installation of a time-shared computer and Australia's first high precision graphics device.

He was foundation president of the WA Computer Society, which later merged with the Australian Computer Society, becoming the first WA Branch Chairman. He was Director of the Western Australian Regional Computing Centre in the sixties and seventies. This provided computing services to CSIRO and State Government Departments as well as the University.

He was executive director of Government Computing for WA from 1978 to 1984. During this period he promoted the development of inter-departmental systems and was closely associated with the development of the WA Land Information System and the WA Technology Park. This was followed by a two year stint managing a computer company in Malaysia, including a consultancy to the Sarawak Government.

He then undertook research in RAN DATA, an encryption company which he had helped establish, and was appointed foundation Head of School of Computing at Curtin University of Technology in 1987. From 1998 to 2002 he was Director of Academic Planning at Curtin. From 1995 to 1999 he was Chair of the State Government's Information Policy Council.

Dennis Moore was elected a Fellow of the Australian Computer Society in 1970 and was made a Member of the Order of Australia for services to Information Technology in 1997. He retired in 2002 and was made an Honorary Life Member of the ACS in 2014.

Professor Andreas Wicenec

Professor for Data Intensive Research

Director Data Intensive Astronomy & Astrophotonics

International Centre for Radio Astronomy Research

The University of Western Australia

Biography

Professor at the University of Western Australia since 2010, leading the Data Intensive Astronomy Program (DIA) of the International Centre for Radio Astronomy Research (ICRAR) to research, design and implement Petabyte scale data flows and high-performance scientific computing for the Square Kilometre Array, the Murchison Wide Field Array (MWA), the Australian SKA Pathfinder.

During his graduate, post-graduate, and post-doctoral appointments, he was involved in the software development and reduction of photometric and astrometric Tycho data from the ESA Hipparcos satellite.

He joined the European Southern Observatory (ESO) in 1997 as an archive specialist and was involved in the final implementation of the archive for ESO's Very Large Telescope (VLT) and the ESO Imaging Survey.

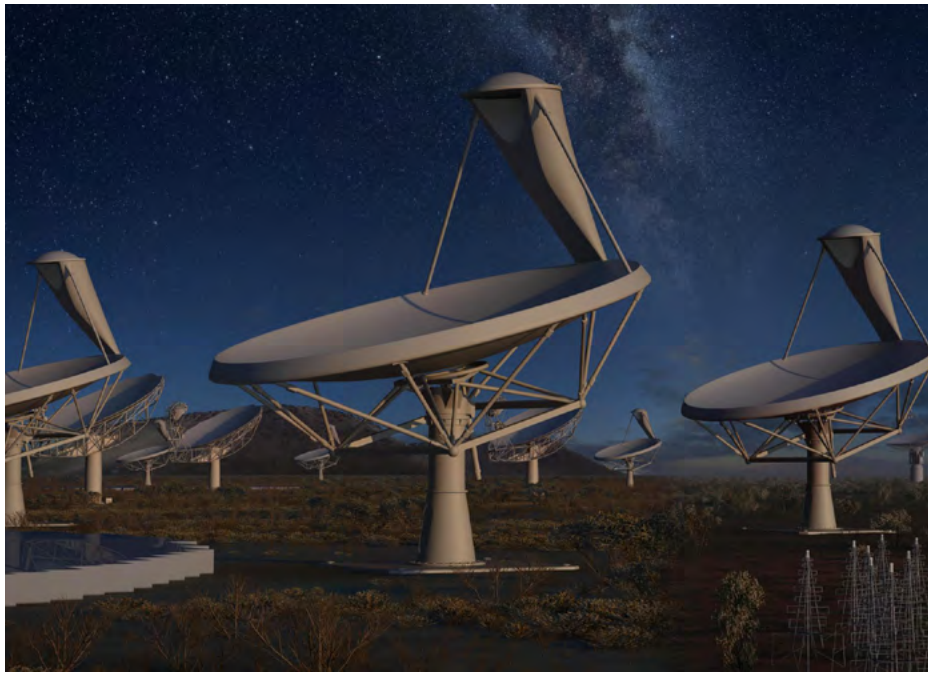
Between 2002 and 2010, he was employed as ESO's Archive Scientist and led the development group of the ALMA archive subsystem. Prof. Wicenec is also involved in the International Virtual Observatory Alliance (IVOA).

His scientific interests and publications include high precision global astrometry, optical background radiation, stellar photometry, dynamics and evolution of planetary nebulae and observational survey astronomy and the related data management, scheduling, and computational concepts.

Astronomical (size) Data Management and Processing in the Era of the Square Kilometre Array

Professor Andreas Wicenec

International Centre for Radio Astronomy Research,
The University of Western Australia



Abstract

One of the biggest science projects ever, the Square Kilometre Array (SKA), is being constructed as we speak. The SKA will consist of two radio telescopes, one here in Western Australia and the other one in South Africa. Every day new antennas are being build, tested, shipped and installed on the two sites in South Africa and Western Australia. Here, in the Murchison Shire, a total of up to 131,072 dipole antennas will be installed, clustered in 512 stations across an area of about 74 km in diameter. In South Africa there will be 197 parabolic dishes distributed across an area of 150 km in diameter.

The science goals for these telescopes are extremely broad and range from observations in our direct vicinity, inside the solar system to studies of the very early universe some 13 billion light-years away. Constructing and operating such delicate scientific instruments in very remote areas is obviously a big challenge, but all of that effort is worth nothing, if we can't analyze and scientifically exploit the data produced by them. Fundamentally, radio antennas are measuring voltages in an extremely high cadence and across many frequency channels and dual polarizations. These measurements are almost immediately digitized and then transferred over fiber links across Wide Area Network links from the observatory sites in South Africa and Western Australia to dedicated High Performance Computing centers in Cape Town and Perth, respectively.

The data rates at this stage of the signal chain are truly eye-watering and reach up to several Terabytes per second (TB/s). This data is directly ingested into bespoke, FPGA based correlators, which perform some averaging and calculate the correlation of the signals from each pair of antennas. The correlators reduce the data rates by about a factor of 10, but the resulting data rates are still of the order of 0.5 TB/s. That data is directly consumed by dedicated High Performance Computing (HPC) clusters.

This is where the job of the so-called Science Data Processor starts and where the contributions and the expertise of my team and myself are concentrated. Correlated voltage measurements are still quite some way from something like an image, which typically is the starting point for actual scientific investigations. Getting from those raw inputs to calibrated data products, ready to be analyzed by scientists requires compute clusters with several hundred Petaflops (PFLOPs) performance (1 PFLOP is 10**15 floating point operations per second) and is thus far out of reach of any normal installation. This compute capacity has to be available continuously, since the telescopes can observe the whole day long, every day and the whole year. The resulting data products are still massive and tick in at up to 1 Petabyte (PB) for every observation lasting between 6 and 12 hours. One PB is the equivalent of 1000 typical laptop storage drives today and thus still far too big to consume on a normal, single computer. In this talk I will present the computational and data management challenges and some of the solutions to maximize the scientific output of the SKA and other large-scale astronomical facilities and outline the limitations and opportunities along the way to support achieving transformational science results and potential Nobel prizes.

1. Introduction

1.1 The Square Kilometre Array Observatory, SKAO

At December 5th, 2022 the SKAO has officially started construction of the two largest radio telescopes in the world. As a brief summary of what the SKAO will be here is a quote from the SKAO web site [1]: *"The SKAO will consist of one global observatory, operating two telescopes, across three sites, for the worldwide scientific community.*

The SKAO is achieved through the committed collaboration of its participating Member States and institutions. Only through this combined capacity in resources, knowledge, and experience (industrial, technical, scientific and at policy level) will the SKA project be realised.

The SKAO was established as an inter-governmental organisation in early 2021. It will undertake the construction, operation and maintenance of the SKA telescopes. The Observatory has a global footprint and will consist of the SKAO Global Headquarters in the UK, the two SKA telescopes at radio-quiet sites in South Africa and Australia, and the associated data processing facilities.

Figure 1: One of the SKA-LOW stations under construction in the Murchison Shire in Western Australia.

While this is a nice and concise summary, it does not provide any detail about what those telescopes actually are, or what they are supposed to observe. Thus, let's dig a little bit deeper and add another quote from the SKAO web site: *"The telescopes will cover two different frequency ranges, and are named to reflect this. SKA-Mid, an array of 197 traditional dish antennas, is being built in South Africa's Karoo region, while SKA-Low, an array of 131,072 smaller tree-like antennas, is being built in Western Australia on the traditional lands of the Wajarri Yamaji. The arrays will both be spread across large distances, with the most distant antennas being separated by 150 km in South Africa, and 74 km in Australia.*

Figure 2: A blend of real SKA-MID antennas on the right and an artist's impression of the more complete array on the left.

Using cutting-edge technology, including some of the fastest supercomputers in the world, they will make it possible to study the Universe in exquisite detail, revealing the inner workings of galaxies, helping us to understand more about the extreme environments around black holes, tracking the journeys of gravitational waves, and enabling a whole host of other ambitious science investigations."

To add a few more details here are a few numbers about the SKA-LOW:

- Size: 131,072 log-periodic antennas grouped in 512 stations
- Collecting area: 419,000 m^2
- Maximum distance between antenna stations: 74 km
- Frequency range: 50 MHz - 350 MHz

Similarly for SKA-MID:

- Size: 197 fully steerable dishes, including the existing MeerKAT radio telescope
- Collecting area: 33,000 m^2
- Maximum distance between dishes: 150 km
- Frequency range: 350 MHz - 15.4 GHz, with a goal of 24 GHz

Going back to the name *Square Kilometre Array,* which in fact means a total of one square kilometre collecting area, you will realize that even both of them together are only less than half of that. They are still pretty massive telescopes but don't quite add up to one square kilometre. The reason for that is simple: That is about what we can expect to be able to build with the budget we have! That also highlights the good thing about radio astronomy: The telescopes will work perfectly fine, even if there are fewer of them. In fact, we started observing already now after having installed just a few antennas. Of course the sensitivity of just half the collecting area will be significantly lower and some of the most demanding observations will not be possible, but both of the telescopes will still be far more capable than any currently existing one. Another very important difference is that the SKA telescopes are being built in the Southern Hemisphere.

The maximum distance between antennas is a very important parameter, since it directly relates to the achievable spatial resolution. Also, just a few words about the frequency ranges: As can be seen from the numbers, the SKA-LOW and the SKA-MID are covering consecutive ranges. Together they will eventually span a huge range between 50 MHz to 24 GHz. Each of the telescopes can split up these frequency ranges into very fine individual channels, providing a very detailed spectral view of the observed signals.
In addition, radio astronomical observations also provide a very high time resolution. For the SKA this will go down to about 100 microseconds (0.0001 seconds). The sensitivity, the spatial, spectral and time resolution are the key parameters for scientists to start dreaming, or rather think in more detail about feasible science projects which could be executed with the SKA, once completed.

2. Scientific Goals

The original SKA Science Book [2], published in 2015, consists of more than 2000 pages in 135 chapters spanning two volumes describing observations covering the whole universe [3]. The science book collected the observing projects researchers from around the world envisioned for the SKA. From the direct surroundings of the earth to the very early universe billions of light-years away, from fundamental physics to search for extraterrestrial life and intelligence [4], everything was covered. The SKA will present such a leap in sensitivity and resolution (Figure 3 and [5]), that the predictions of what we will eventually be able to discover are mere guesses and we will for sure detect new, unexpected and surprising objects and physical effects.

The SKA will truly serve humanity's quest for knowledge and insight about our origin and standing in the universe. The most comprehensive overview of the SKA project can be found in the Construction Proposal [6].

3. Challenges

The main challenge of the SKA project is money, or rather the lack of it! The total budget currently stands at about 2 Billion AUD for the whole construction, including the first two years of operations. As a number that seems a lot, but given that this amount is spread over about 10 years and currently 10 member countries, the actual average contribution per country every year is just 20 million AUD. In addition, there are another 6 prospective member countries observing the project, which will eventually result in additional contributions. It should be noted, that the contributions are different for each of the members. The hosting countries, including Australia, are spending significantly more than other nations.

Why does it cost so much? For one, we are building two extremely sensitive telescopes in very remote and harsh environments on two continents. But in addition, in order to be able to collect and process the data, we also need some of the biggest dedicated scientific high throughput compute facilities in the world.

In some sense this challenge is a good one, since it requires the parties involved in the construction to think about innovative solutions. In particular the processing of the data is certainly technically feasible [7], but would require facilities costing as much as the overall SKAO budget, both in captial and in operational costs (e.g. power). Tackling and overcoming these challenges triggers technological advances and benefits far beyond the science and that is one of the main non-scientific drivers for countries around the world to join the SKAO.

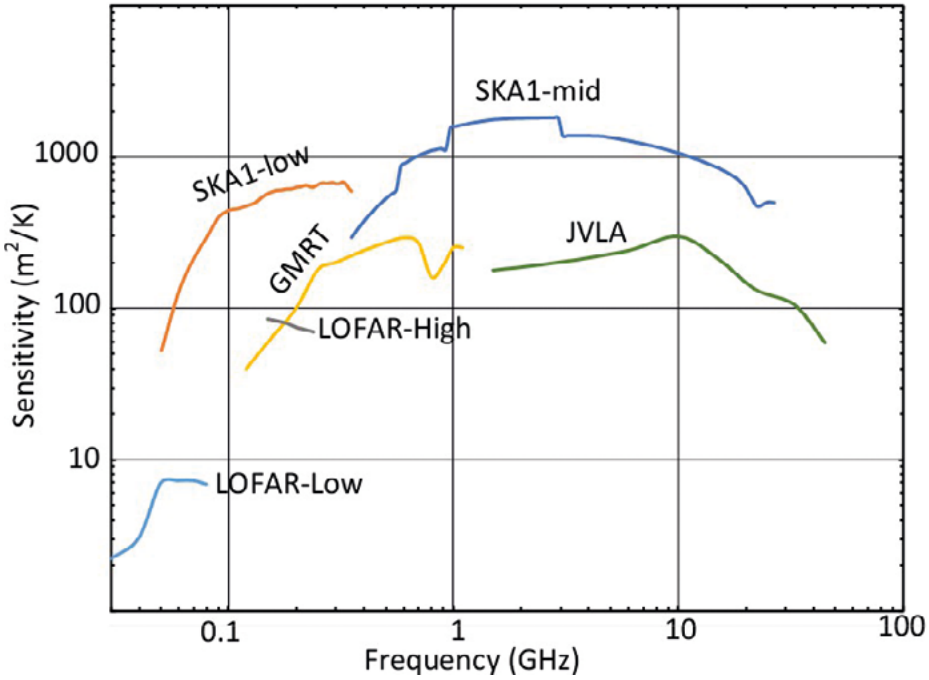


Figure 3: The SKA-LOW and SKA-MID telescope sensitivities over frequency compared with other major existing radio telescopes

Another cost driver originates from setting up and operating a science project on three, globally distributed sites and in collaboration with many diverse countries and individual institutions and companies. This is both a challenge and a rewarding exercise. The SKAO has been established as an inter-governmental organization, and is directly governed by a council with representatives from the governments of the member countries.

Aligning the work of hundreds of people distributed across many time zones and with a very wide range of cultural backgrounds is at the same time quite inefficient, but also very beneficial in terms of societal returns, up to the governmental level. Agreeing on a science project with common ambitious goals is easier than talking about trade protection or diverging regional interests.

4. Solutions

I'm not a politician but a scientist with a very keen interest in technical and algorithmic solutions in the computational domain of the SKA. There are dozens of teams around the world working on various parts of the SKA, spanning from planning and constructing roads to efficiently turning bits into scientific data. All of these parts have to come together and work and function as a whole, else there will be no observatory, or at least no science coming out at the end. Covering even just an overview of all of that would be close to impossible for a single person with limited expertise and insight into many of these parts. Thus, in the remainder of this talk I will mainly concentrate on the areas I know about best. More specifically on the areas my team and me at the International Centre for Radio Astronomy Research (ICRAR) together with two colleagues from CSIRO are actively working on. These include two main focus areas contracted within the SKA construction and two research areas with a very tight relationship with some of the SKA science cases and the computational challenges:

1. Receive and real-time calibration
2. Data lifecycle management
3. Data compression
4. Very deep surveys, combining hundreds or thousands of hours of individual observations.

4.1 SKA Data Flow

The SKA data and processing flow is a complex, distributed data acquisition, transfer, processing and data management problem. On the highest level there are four stages:

1. Raw data collection
2. Beamforming and correlation
3. Real-time processing and calibration
4. Batch/science processing

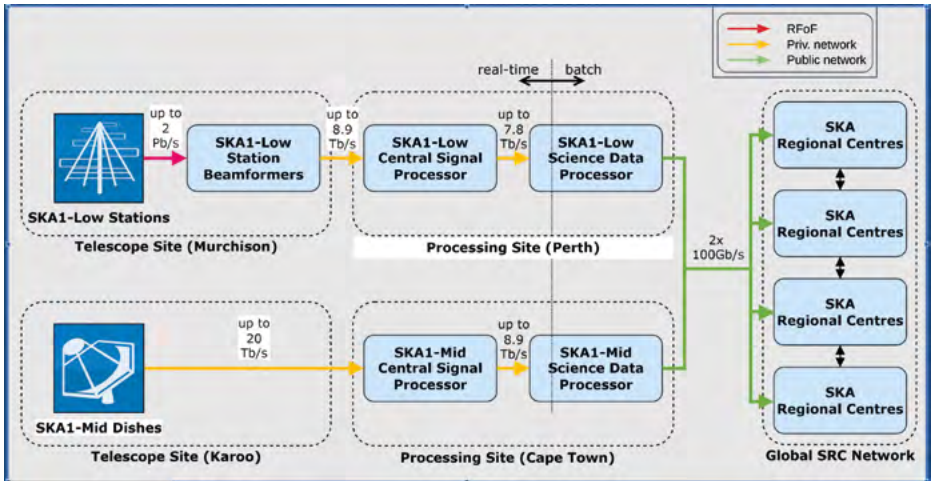


Figure 4: Very high level view of the SKA data flow for both SKA-LOW and SKA-MID. The left part up to the dotted line marks the part of the real-time data flow, the right side will be performed in a more batch-like fashion. The right-most section, 'Global SKA Regional Centre Network', is sitting outside the SKAO and is also funded separately by individual countries.

The first obviously happens at the telescope sites, where the antennas are located. I will not cover this aspect in any extend. The second is actually geographically split, the beamforming is happening on-site, the correlation at the dedicated computing centres. The telescope sites and the compute centres are some 750km apart and that dedicated network link requires special attention as well, but I will not expand on this here either. Once the data has made its way from the sites to the compute centres, it hits the so-called Correlator subsystem, or more specifically a dedicated, bespoke FPGA-based computer. The correlator directly 'listens' to the network streams from the site, very much like watching streaming TV at home, except that the data rate is 100.000 times higher! The correlator will perform it's job, that is computing the correlation between each pair of antennas as well as some significant time and frequency averaging and mitigation of Radio Frequency Interference (RFI). The resulting data stream is reduced by a factor of about ten and will then be sent to the so-called Science Data Processor (SDP) subsystem. More specifically the SDP computing system is a dedicated, very large processing infrastructure consisting of a cluster of more or less standard computers with attached high performance storage. That is where my team comes in, and I will expand on those aspects in more detail below.

4.1.1 Real-time processing and calibration

Note, that at this point in the data flow, the data is still streaming and has never hit any storage device. Here, it is streaming directly into the network interfaces of the receive computers. The developers of the SDP subsystem are working on a set of receive workflows, dealing with multiple different operational modes. All of them will obviously first need to handle the incoming streams, means to unpack, decode and reformat the data. Since the correlator does not have any buffer, this software is highly time-critical, and needs to be able to handle many parallel streams at the same time. Part of my team is working on this key part of the data chain.

The decoded data from the receive software will be forwarded to various real-time processing modules. One of these modules will apply multiple real-time calibration algorithms and one of the firsts of those algorithms will have to deal with distortions of the radio waves caused by the ionosphere of the earth. These distortions come in two flavours, direction dependent and independent, respectively. The direction dependent part is quite complex and articulates itself similarly to watching the floor of a pool through the water waves at the surface: the same patch of the floor appears in multiple directions and is smeared around. In addition this effect is also varying with time as the waves move over the surface.

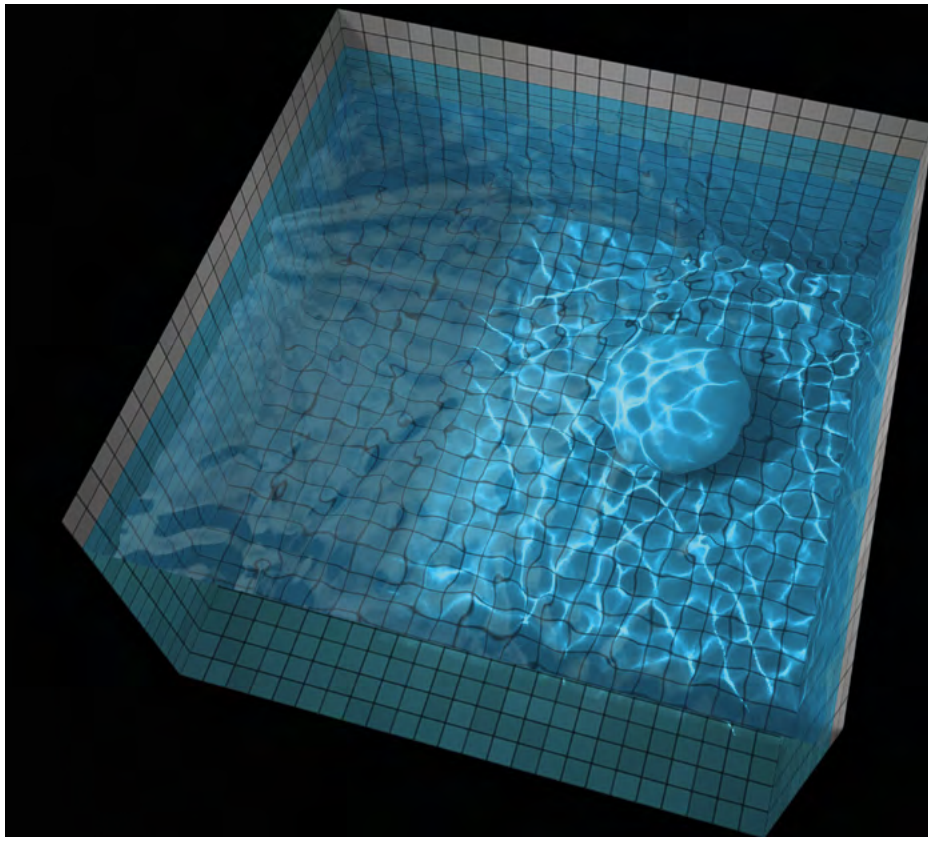


Figure 5: Caustics caused by light passing through the disturbed surface of a pool. The distortions of the regular grid lines of the tiled floor are very obvious, very direction dependent and also changing with time. Looking through the ionosphere into space has a very similar effect and will need to be corrected in real time. Screenshot from https://madebyevan.com/webgl-water.

In order to be able to precisely control the observing directions and the focus of the individual antennas of the radio telescope, we will need to remove these distortions to a very high degree. Part of my team is developing the algorithm doing that [8]. At the end, the real-time processing workflows will write the resulting data to high-speed storage in a format compatible with the batch processing workflows. Usually the data for every single observation of between six and twelve hours will have to be collected, before the batch processing can start.

The kind of processing applied during the execution of the batch processing workflows is very much dependent on the science goals of the observation. That is also reflected in the range of actual storage and processing requirements of each of these workflows, which can span many orders of magnitude [9].

4.1.2 Data Lifecycle Management

One of the seemingly obvious and down-to-earth issues is the basic management of data collected and produced throughout the lifecycle of an observing project. In the case of the SKA, this becomes a task spanning minutes for intermediate data products to half a century, the nominal lifetime of the observatory. It also spans many orders of magnitude of data product sizes and needs to address LAN and WAN distribution of heterogeneous storage technologies. The DIA team is deeply involved in the implementation of a Data Lifecycle Management (DLM) system for the SKAO. This will be the first system from the team being installed operationally for the SKA-LOW initially and then for the SKA-MID as well, in order to keep track of data collected already during the very early array releases. The DLM consists of a set of interconnected, but distributed services running on-top of a high-availability database to keep the metadata. The system provides the FAIR (Findable, Accessible, Interoperable and Reusable [10]) interface to SKAO data, but also ensures integrity and security of the data holdings.

4.2 SKA Related Research Projects

The SKA design and costing model has compromised processing capacity for 'metal on the ground'. Means that the budget for processing and storage is far lower than what would be required to process the full data rate and volume the deployed antennas are capable of producing. This is a sensible decision in that processing and storage costs had continuously been decreasing over the past several decades and that compute equipment has to be replaced and upgraded every five to seven years in any case. On the other hand this also means that it will be very hard or impossible to process the most demanding science observations during the first years of operations. The SKA processing platforms will only be able to process some average data rate and volume, depending on the efficiency of the processing workflows and algorithms and on the

actually affordable capabilities of the processing platform. In order to still get the maximum possible out of the existing deployment, we need to be smarter and that is why we at ICRAR have been involved in a few research projects to tackle some of the bottlenecks.

4.2.1 Workflow Optimization and Scheduling

This research area is a true passion of mine and is based on my belief that large scale data reduction workflows needs to improve in several areas:

1. It needs to become far more efficient and approachable to develop scientific workflows by individuals and teams.
2. Scientific workflows and the execution engines need to be able to scale with the collected data.

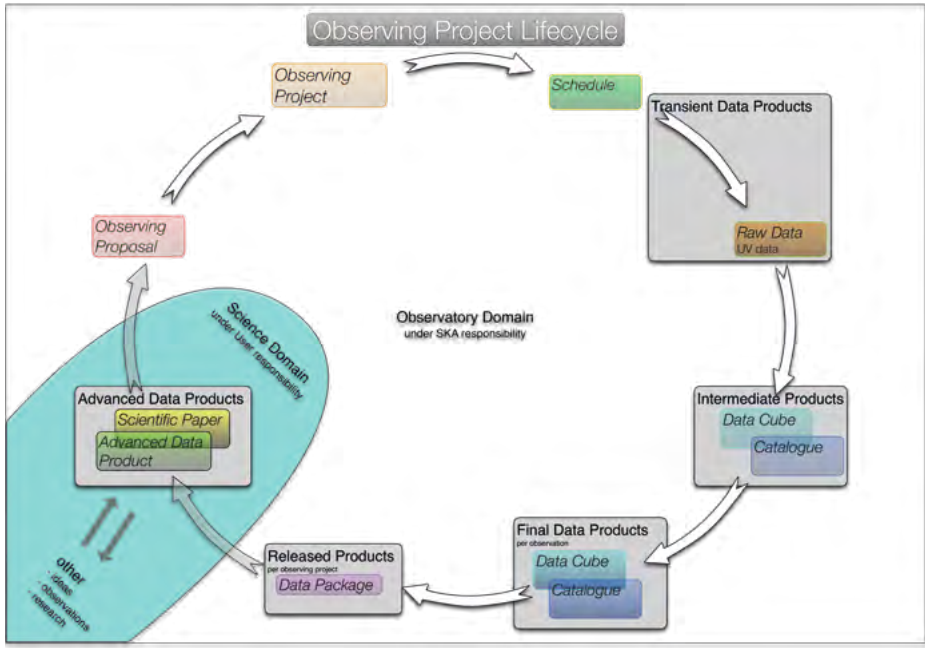


Figure 6: The SKA Observing Project Lifecycle in relation with the various kinds of data products produced, indicated by grey boxes. The 'Transient Data Products' in the upper right corner are not stored anywhere, but only exist in volatile memory. The domain of the SKAO DLM are the three boxes 'Intermediate Products', 'Final Data Products' and 'Released Products'. The 'Advanced Data Products' are responsibility of the users of the data, but will usually require very significant resources to be provided by the SKA Regional Centres.

3. Workflow frameworks need to support the concept of 'separation of concerns'. Means that it must be possible to develop workflows and algorithms as well as workflow systems separately by the respective experts.
4. Workflow frameworks have to support scientific reproducibility.
5. The frameworks need to encourage and support the establishment of well-developed workflows and promote collective development and re-use.

Some of the items above have been successfully addressed in other scientific communities, like e.g. bio-informatics, astronomy and astrophysics is lagging behind in this area. Very likely partially due to the in-grained belief that writing data analysis software or even 'just' some pipelines is straight-forward and an integral part of any data oriented research project. In the best case the software is regarded as a valuable side-product and lives on, in the worst case it sits on some directory on an abandoned account on an institutional server and gets deleted after some time. Being a side-product, the software is very often lacking proper documentation, automated tests, version control and obviously longer term maintenance and support. All of which are standard software engineering practices, but require significant additional effort and are deemed of little to no importance for the original goals of the astrophysical research project. Scientific research builds on-top of peer-reviewed publications, or disproves previous results. In contrary, the associated software in most cases is never really verified, let alone re-used. This leads to enormous amounts of replicated effort and potentially erroneous results.

Our workflow development and execution environment DALiuGE [11] is trying to address these issues by incorporating them into the design of the software system. In the same order as the list above, DALiuGE features:

1. a graphical editor for visual development of logical scientific workflows to improve efficiency and usability.
2. logical workflow constructs to indicate scatters and gathers allowing easy to scalability.The execution engine uses a share nothing approach and scales to millions of individual tasks.

3. complete separation of workflow logic and description from code to the degree that a workflow can be executed in test mode, without any code at all. Conversely, the code behind the individual tasks does not need to be implemented with any knowledge about DALiuGE at all, means that almost any code can be used in a workflow and can be optimized by expert developers.
4. built-in Merkle trees to encode graphs, data and code artefacts to allow comparison of workflow runs to enable scientific reproducibility without the user having to do anything in addition.
5. GitHub and GitLab repositories to keep workflows under version control. This is an integral part of the graphical editor and almost enforces good practice and collaboration around the development of workflows.

DALiuGE had been used to execute a massive test run on Summit, the fastest supercomputer at the time (2019). This run and the associated publication [7] had been nominated for the Gordon Bell Prize in supercomputing in 2020 [12].

Figure 7: EAGLEπ, the graphical workflow editor for the DALiuGE system, developed by the ICRAR team. The very simple example graph shown computes the spherical distance of two coordinates on the sky and also identifies in which constellation one of them is located. This is performed by components from one of the most used astronomical software packages, astropy [13]

4.2.2 Data Compression

One key limitation of radio astronomy processing is the fact that many of the algorithms are I/O bound, means that they are limited by the read and write rates of the various memory and storage layers rather than the actual calculations. For some of these algorithms we would be able to perform up to ten times more calculations on every single byte read and/or written. That means in turn that the processors (CPUs or GPUs) are running idle for significant amounts of time, waiting for data to be available. Moreover, the storage costs to keep and serve the data for the batch processing step, even to just support the average processing requirements, makes up a very significant fraction of the total costs of the processing hardware. It is on this background that we are performing research into reducing the data rate and volume by applying advanced lossless and lossy data compression techniques. Two key questions have to be addressed:

1. How far can we push lossy compression without significantly affecting the scientific results?
2. Where in the data flow can/should we apply compression?

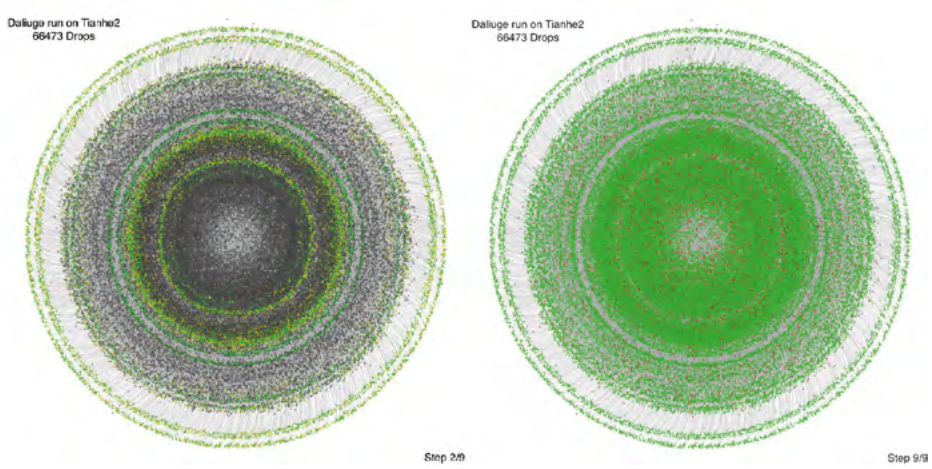


Figure 8: Two snapshots of a test workflow execution on the Chinese supercomputer Tianhe-2 [14]. Each dot represents a task and the connecting lines are indicating the sequence of the execution flow. Grey tasks have not been executed yet, yellow are being executed, green are finished and red have failed.

Lossy data compression is a very well established domain in signal processing and in audio and video streaming in particular. Almost all of the content we are listening or watching every single day is lossily compressed! The reason behind that is quite similar to our argumentation for the SKA: The required data rates for the non-compressed streams are simply too high and would overwhelm the available platforms (mobile phones) and/or make them un-affordable or very bulky. Simply put, the options are, we compress it and loose something, or you can't watch your favourite series on your phone! In the language of a prospective SKA science user that would sound like: We can either lossily compress the data and give that to you now, or you have to wait a decade or more until we will be able to do your science without compression.

Lossy compression can be crude or very sophisticated. As an example consider the following two lossy video compression techniques:

1. lower the frame rate by a factor of two;
2. identify and transmit only changes to the previous frame;

The first one obviously reduces the data by a constant factor of two, but sacrifices the frame rate. The second does not sacrifice the frame rate and it also does not change anything in the perceived quality of the video. However, in the most extreme case, when the whole new frame is different from the previous one, it also would not compress the data at all. In the other extreme case, if a non-changing scene is captured, the compression would be extremely high. In that case the sender only has to tell the receiver that nothing changed. The compression techniques we are employing are far more sophisticated than even the last technique. MGARD [15] for example allows to compress data on multiple levels, separate dimensions and also in a hierarchical fashion, while ensuring that the errors induced by the lossy compression does not exceed a pre-set level. Going back to the change-only compression, you could then tell the compression algorithm to ignore any changes smaller than a certain value and like this increase the compression rate further. With MGARD you can even define regions of interest in the frames, where you don't want to lose anything and thus only apply lossless compression.

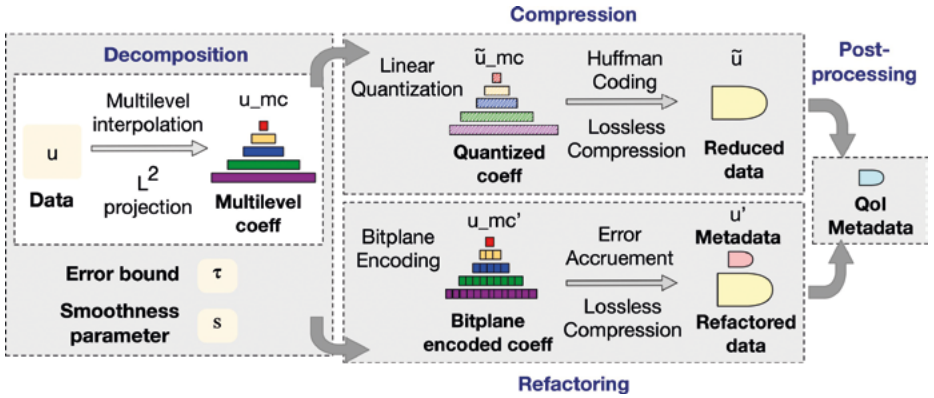


Figure 9: A block diagram of the MGARD compression algorithm, which enables lossy compression with guaranteed error bounds.

The unique problem with radio astronomy data is that most of the data is actually very much just noise (after all, the sky is mostly empty!) and that is fundamentally not compressible. On the plus side, astronomers are not that interested in the noise, thus we could simply throw that away. However, the actual signals are extremely faint and pretty much buried in the noise. That is why astronomical observations tend to accumulate data over long periods of time. Think long exposure time when taking a picture in the dark. When performing that accumulation correctly, the noise stays roughly constant, while the signal adds up with time. That means in turn that after the accumulation, the so-called signal-to-noise ratio (SNR) is much higher, while at every time-step it is usually very low. In other words, the object is only visible after the accumulation and even then only after serious processing.

Thus, if we apply compression to the individual time steps we need to be very careful. How careful we need to be, can be expressed in mathematical formulae and is dependent on the final SNR of detectable objects and the total amount of time we will integrate for. Being able to express that in pure mathematics is a good start, but unfortunately real data is always more complex than the mathematical model we construct. Therefore, we still need to perform many experiments to verify our hypothesis of the applicability of lossy compression.

4.2.3 Deep Surveys

Deep surveys extend the accumulation time mentioned above to hundreds or even thousands of hours in order to be able to detect the faintest of signals. The team at ICRAR have been involved in deep surveys for many years. The CHILES survey [16] collected 1002 hours of data using the Jansky Very Large Array (JVLA) in New Mexico. The goal was to go far deeper than any other survey before.

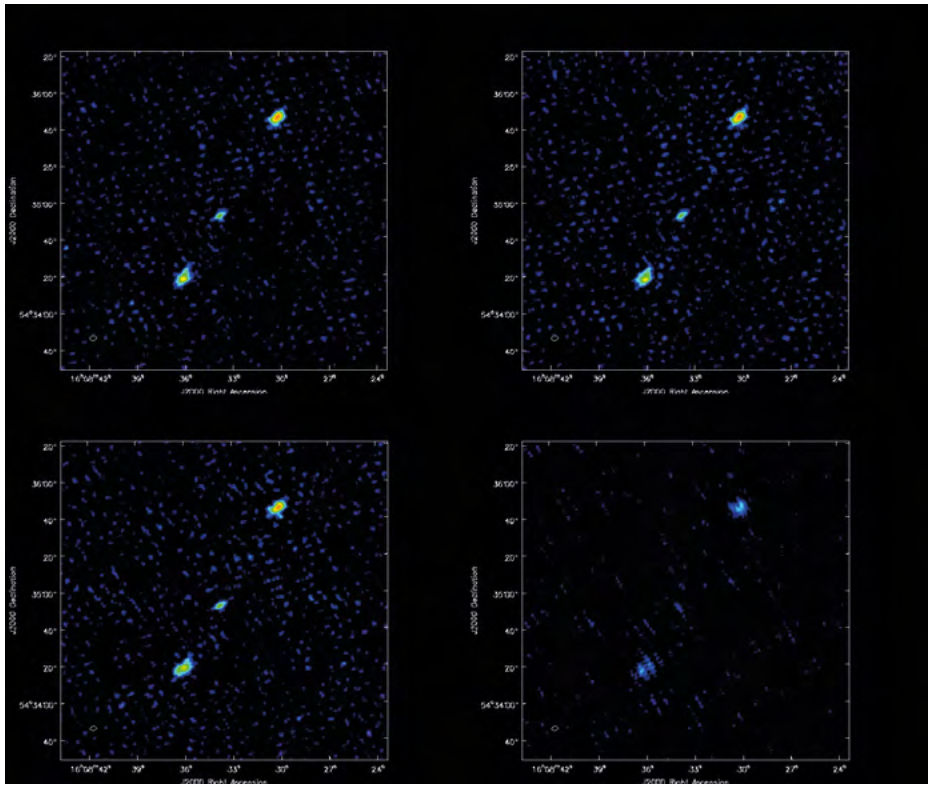


Figure 10: Example of applying compression to a real data sets around a source using varying error bounds of MGARD. The colour scale is radio flux density. The compression was applied with absolute error bounds of 0.6, 1.2, 2.0, and 6.0, respectively (from top left) with data of a sigma of 4 (very non-Gaussian distribution) and a data range of +-100. Visually it is hard to detect a difference between the first two. The lower left one might still be useful for some science, the lower right one, unsurprisingly, not at all. Compression ratios are 11.1, 17.2, 25.0, 75.1, respectively. Means that even the most conservative error bound still achieves a compression of more than a factor of 10.

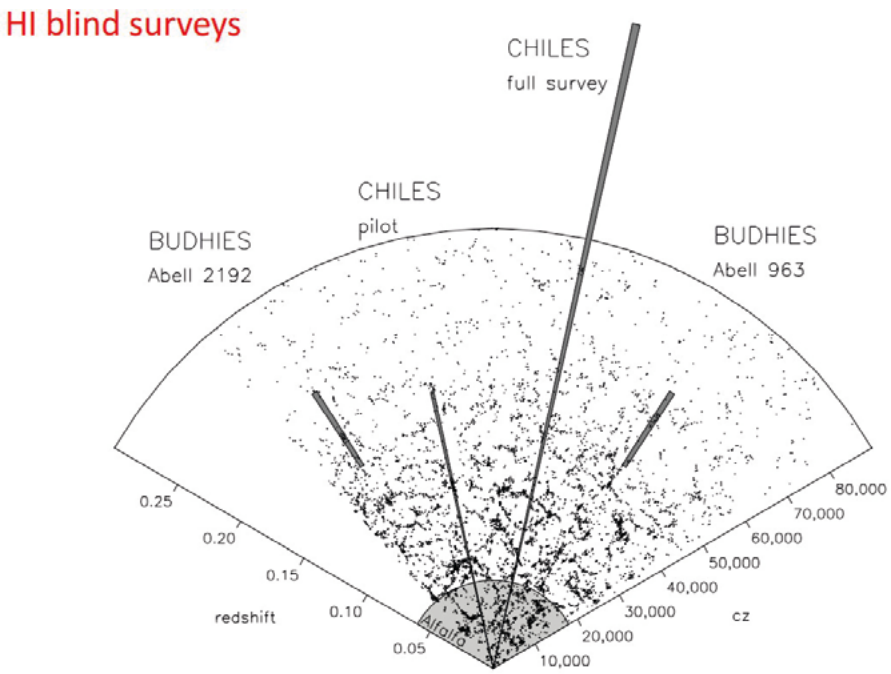


Figure 11: Comparison of CHILES with some other deep surveys. The wedge represents a slice of the sky across a single latitude. The axis on the left is redshift. The axis on the right is speed of light times the redshift in km/s and indicates the apparent velocity with which the galaxies are moving away from us. The tip of the wedge is where Earth is. Each dot inside the pie represents a known galaxy from optical surveys. The pencil beams represent the coverage of a few deep radio surveys. The density of the known galaxies decreases with distance and that is a function of the sensitivity of the surveys. The bubble-like structure of the galaxy distribution is also known as the large-scale structure of the universe. From [17].

Due to the extreme weakness of the signals, radio astronomy data is known for its relatively boring and non-spectacular appearance. This will change with the SKA, since the resolution will reach a level similar to optical telescopes.

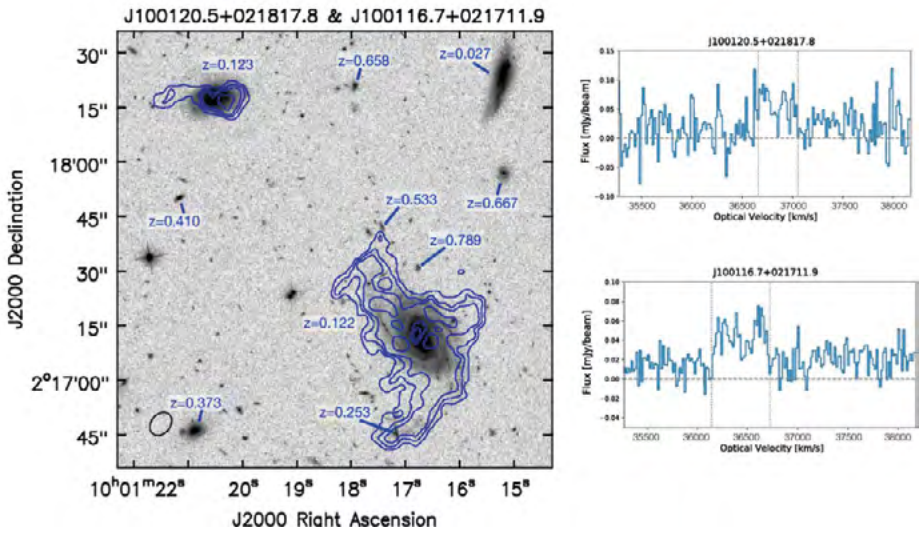


Figure 12: Example of two galaxies detected by CHILES in a small area of the sky. The grey-scale background image in the left panel is an optical image from the Hubble Space Telescope, the blue contours are the radio measurements from CHILES. The z-values provided for some of the galaxies in the image are the measured optical red-shifts. The small black ellipse in the lower left corner of that panel is an indication of the actual size of the spatial resolution of the telescope. That ellipse is over-sampled by a factor of four during the processing, which is about the best reliable super-resolution result you can achieve mathematically. Looking closely, it is very obvious how much better the spatial resolution of the optical image is. Looking at the two plots at the bottom is also quite interesting, since those actually show just how weak the signal even of the two brightest objects in the upper left corner and the big galaxy in the lower right in that picture is. The signals from the galaxies are the sections of the plots between the dotted lines.

References

1. Main SKA Web Site. [Online]. Available: https://skao.int
2. SKA Science Book. [Online]. Available: https://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=215
3. Science Goals. [Online]. Available: https://www.skao.int/en/explore/science-goals
4. SETI. [Online]. Available: https://www.seti.org/csc
5. SKA Telescope Specifications. [Online]. Available: https://www.skao.int/en/science-users/118/ska-telescope-specifications
6. SKA Construction Proposal. [Online]. Available: https://skao.canto.global/s/M8159
7. R. Wang, R. Tobar, M. Dolensky, T. An, A. Wicenec, C. Wu, F. Dulwich, N. Podhorszki, V. Anantharaj, E. Suchyta, B. Lao, and S. Klasky, "Processing Full-Scale Square Kilometre Array Data on the Summit Supercomputer," in SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, 2020, pp. 1–12.
8. M. Rioja, R. Dodson, and T. M. O. Franzen, "LEAP: an innovative direction-dependent ionospheric calibration scheme for low-frequency arrays," Monthly Notices of the Royal Astronomical Society, vol. 478, pp. 2337–2349, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:115512628
9. P. Quinn, T. Axelrod, I. Bird, R. Dodson, A. Szalay, and A. Wicenec, "Delivering ska science," 2015. [Online]. Available: https://arxiv.org/abs/1501.05367
10. M. D. e. a. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 1, p. 160018, Mar 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.18
11. C. Wu, R. Tobar, K. Vinsen, A. Wicenec, D. Pallot, B. Lao, R. Wang, T. An, M. Boulton, I. Cooper, R. Dodson, M. Dolensky, Y. Mei, and F. Wang, "Daliuge: A graph execution framework for harnessing the astronomical data deluge," Astronomy and Computing, vol. 20, pp. 1–15, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213133716301214
12. R. Wang, A. Wicenec, and T. An, "SKA shakes hands with Summit," Science Bulletin, vol. 65, no. 5, pp. 337–339, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095927319307157

13. Astropy Collaboration, A. M. Price-Whelan, P. L. Lim, N. Earl, N. Starkman, L. Bradley, D. L. Shupe, A. A. Patil, L. Corrales, C. E. Brasseur, M. N"othe, A. Donath, E. Tollerud, B. M. Morris, A. Ginsburg, E. Vaher, B. A. Weaver, J. Tocknell, W. Jamieson, M. H. van Kerkwijk, T. P. Robitaille, B. Merry, M. Bachetti, H. M. G"unther, T. L. Aldcroft, J. A. Alvarado-Montes, A. M. Archibald, A. B'odi, S. Bapat, G. Barentsen, J. Baz'an, M. Biswas, M. Boquien, D. J. Burke, D. Cara, M. Cara, K. E. Conroy, S. Conseil, M. W. Craig, R. M. Cross, K. L. Cruz, F. D'Eugenio, N. Dencheva, H. A. R. Devillepoix, J. P. Dietrich, A. D. Eigenbrot, T. Erben, L. Ferreira, D. Foreman-Mackey, R. Fox, N. Freij, S. Garg, R. Geda, L. Glattly, Y. Gondhalekar, K. D. Gordon, D. Grant, P. Greenfield, A. M. Groener, S. Guest, S. Gurovich, R. Handberg, A. Hart, Z. Hatfield-Dodds, D. Homeier, G. Hosseinzadeh, T. Jenness, C. K. Jones, P. Joseph, J. B. Kalmbach, E. Karamehmetoglu, M. Kaluszy'nski, M. S. P. Kelley, N. Kern, W. E. Kerzendorf, E. W. Koch, S. Kulumani, A. Lee, C. Ly, Z. Ma, C. MacBride, J. M. Maljaars, D. Muna, N. A. Murphy, H. Norman, R. O'Steen, K. A. Oman, C. Pacifici, S. Pascual, J. Pascual-Granado, R. R. Patil, G. I. Perren, T. E. Pickering, T. Rastogi, B. R. Roulston, D. F. Ryan, E. S. Rykoff, J. Sabater, P. Sakurikar, J. Salgado, A. Sanghi, N. Saunders, V. Savchenko, L. Schwardt, M. Seifert-Eckert, A. Y. Shih, A. S. Jain, G. Shukla, J. Sick, C. Simpson, S. Singanamalla, L. P. Singer, J. Singhal, M. Sinha, B. M. SipHocz, L. R. Spitler, D. Stansby, O. Streicher, J. Sumak, J. D. Swinbank, D. S. Taranu, N. Tewary, G. R. Tremblay, M. d. Val-Borro, S. J. Van Kooten, Z. Vasovi'c, S. Verma, J. V. de Miranda Cardoso, P. K. G. Williams, T. J. Wilson, B. Winkel, W. M. Wood-Vasey, R. Xue, P. Yoachim, C. Zhang, A. Zonca, and Astropy Project Contributors, "The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package," APJ, vol. 935, no. 2, p. 167, Aug. 2022.

14. Tianhe-2. [Online]. Available: https://www.top500.org/resources/top-systems/tianhe-2-milkyway-2-national-university-of-defense/

15. Q. Gong, J. Chen, B. Whitney, X. Liang, V. Reshniak, T. Banerjee, J. Lee, A. Rangarajan, L. Wan, N. Vidal, Q. Liu, A. Gainaru, N. Podhorszki, R. Archibald, S. Ranka, and S. Klasky, "Mgard: A multigrid framework for high-performance, error-controlled data compression and refactoring," SoftwareX, vol. 24, p. 101590, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352711023002868

16. X. Fern´andez, H. B. Gim, J. H. van Gorkom, M. S. Yun, E. Momjian, A. Popping, L. Chomiuk, K. M. Hess, L. Hunt, K. Kreckel, D. Lucero, N. Maddox, T. Oosterloo, D. J. Pisano, M. A. W. Verheijen, C. A. Hales, A. Chung, R. Dodson, K. Golap, J. Gross, P. Henning, J. Hibbard, Y. L. Jaff´e, J. Donovan Meyer, M. Meyer, M. Sanchez-Barrantes, D. Schiminovich, A. Wicenec, E. Wilcots, M. Bershady, N. Scoville, J. Strader, E. Tremou, R. Salinas, and R. Ch´avez, "Highest Redshift Image of Neutral Hydrogen in Emission: A CHILES Detection of a Starbursting Galaxy at z = 0.376," Astrophysical Journal Letters, vol. 824, no. 1, p. L1, Jun. 2016.

17. CHILES talk. [Online]. Available: https://www.atnf.csiro.au/research/conferences/2016/IDRA16/presentations/VanGorkomJacqueline.pdf

EDGE

SUPPORTING
LOCAL INDUSTRY

EMBRACING AI