Australian SKA Regional Centre Design Study Program: Final Report

K. Lee-Waddell, D. Pallot,
G. German, D. Null, A. Shen,
G. Aniruddha, A. Williamson, and K. Holmes

29 November 2022

Introduction to the AusSRC

The SKA Observatory (SKAO) is an intergovernmental project that aims to build the world's largest radio telescopes. In order to fully exploit the scientific output of the immense amount of digital data flowing from the telescopes, the SKAO is working with science communities across the world to establish distributed data computing and networking capabilities. These capabilities will draw on international cooperation through SKA Regional Centres (SRCs), which are nationally lead hubs that form the basis of a global infrastructure.

The Australian SKA Regional Centre (AusSRC) is Australia's portion of this international computing and data delivery network. To establish requirements, develop a SRC prototype and framework, and determine the future (and long-term) direction for the AusSRC, a 3-year Design Study Program (DSP) was launched in 2019 through funding from the Commonwealth, through the Department of Industry, Innovation and Science (DIIS) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). This document provides an overview of the AusSRC DSP.

Background

The SKAO is poised to usher in the next era of astronomical discovery and advanced data processing; however, the resources needed to fully process, distribute, curate, and utilise data flowing from its telescopes are currently beyond the scope of the SKA construction and operations budget. As previously experienced by the Large Hadron Collider (LHC) project, the SKAO and the international SKA science community will need to work collaboratively to establish and build a shared and distributed data computing and networking capability that draws on international cooperation and supports the broad spectrum of SKA Science.

In March 2016, the SKA Organisation Board encouraged SKA member states to form "a collaborative network of SRCs to provide the essential functions that are not presently provided within the scope of the SKA project". By the end of 2018, proto-SRC design and development projects were in advanced stages of planning and initiation across 13 SKA member states. In November 2018, the Board approved the formation of the SRC Steering Committee (SRCSC) with a mission to: "Guide the definition and creation of a long-term operational partnership between the SKAO and an ensemble of independently-resourced SRCs".

In 2019, DIIS – through the Australian SKA Coordination Committee (ASCC) – funded the University of Western Australia (UWA) to organise a community-based DSP for an Australian SRC. CSIRO matched the Government contribution and with a total cash budget of $3.8 million, the 3-year program began.

The AusSRC DSP has been overseen by Management Committee (MC) consisting of representatives from UWA, CSIRO, Curtin University, Astronomy Australia Limited (AAL), the Pawsey Supercomputing Research Centre, the Murchison Widefield Array (MWA), and CSIRO's Australian SKA Pathfinder (ASKAP).

Based on the work of the DSP, in February 2021 the AusSRC team submitted a proposal for long-term funding as part of the overall Department of Industry, Science and Resources (DISR) Federal budget submission covering Australia's contribution to the SKA project. This proposal was successful and will deliver $63 million to build, operate, and maintain the AusSRC over a 10-year period following the completion of the DSP in 2022.

Overall DSP Achievements

General Outcomes

Over the past three years, the AusSRC DSP system has built a fully functional data post-processing system. The frontend interface allows researchers to interact with large quantities of science-ready data from ASKAP and use custom pipelines and workflows (developed by AusSRC DSP personnel) to enhance the quality and usability of current SKA precursor data to achieve scientific objectives. The backend framework is modular (using containerisation), runs on high performance computing (HPC) systems, and can be easily adopted at other facilities/proto-SRCs or completely changed out if a more effective solution is found/developed. DSP personnel have also re-written and better optimised coding to process MWA data. All DSP prototyping results and outcomes are now being shared with the international SKA community as the SRC Network (SRCNet) starts development on the global framework.

The lessons learnt from setting up and operating the DSP greatly informed how to establish a longer term AusSRC capability. A $63 million funding proposal detailing a 10-year AusSRC plan, in line with SKAO timelines and international SRCNet development work, was submitted to the Australian Government and approved in 2021. Grant guidelines were released in March 2022, and the AusSRC is in the final stages of the grant establishment process. A formal Partnership has been formed between CSIRO, Curtin University, the Pawsey Supercomputing Research Centre, and UWA. These four institutions provided significant support of the DSP personnel and activities, establishing that the current Partners have the right level of expertise and support to establish and grow the long-term AusSRC capability.

Direct Benefits

The prototyping work of the AusSRC DSP has enabled scientists currently working with SKA precursor telescopes, such as ASKAP and the MWA, to more efficiently post-process, analyse, and share their data with other team members located around the world.

Several of the ASKAP Survey Science Teams – such as: Deep Investigation of Neutral Gas Origins (DINGO), Evolutionary Map of the Universe (EMU), First Large Absorption Survey in HI (FLASH), Polarisation Sky Survey of the Universe's Magnetism (POSSUM), and Widefield ASKAP L-band Legacy All-sky Blind surveY (WALLABY) – are using AusSRC pipelines and workflows to combine data, run source finding applications, catalogue and archive detected astronomical sources, and perform detailed scientific analysis. Furthermore, EMU is running

enhanced cross-matching algorithms on millions of ASKAP sources with databases from other telescopes currently archived at Data Central (part of the Australian Astronomical Observatory – AAO – at Macquarie University) for streamlined comparison and identification. WALLABY's extracted catalogues and source cutouts are being sent to the Canadian Initiative for Radio Astronomy Data Analysis (CIRADA) at the Canadian Astronomy Data Centre (CADC) so scientists can run kinematic modelling processes, the results of which feed back into the AusSRC archive. FLASH's Consolidated HI Absorption Database (CHAD) is fully integrated into the AusSRC system with full metadata readily linked in the database (rather than being in various external spreadsheets) and the science team is starting to utilise AusSRC-developed cross-matching services.

New "big data" handling techniques and methods are also being tested. Rather than using the traditional Flexible Image Transport System (FITS) format, which was initially released in the 1980s, the AusSRC is helping the POSSUM team test out Hierarchical Equal Area isoLatitude Pixelation (HEALPix; https://healpix.sourceforge.io). HEALPix can not only handle large, multi-TB files, but also enables an efficient division of the files into smaller portions for transmission to other data centres (e.g. other proto-SRCs) where those portions can be quickly stitched back together with no loss of information. Working directly on the Pawsey Supercomputing Research Centre's new Setonix system, AusSRC co-funded DINGO personnel are deploying Data Activated Liu Graph Engine (DALiuGE – a graph-oriented workflow development, scheduling, and execution system; https://daliuge.readthedocs.io/en/latest/) while testing newly developed gridding technology, which can enable the storage of information from TB-sized files into much smaller volume grids that preserve all required astronomical data for advanced processing and future analysis.

For the MWA, AusSRC personnel have been refactoring various components of the Epoch of Reionisation (EoR) pipeline not only to work with the telescope's new correlator, but also to improve the capability of the processing system to continue pushing the limits in this highly detailed field of research. The Birli software library (https://github.com/MWATelescope/Birli), written by an AusSRC developer, has been optimised for heavy data reading and writing operations and can handle various forms of metadata in a more streamlined manner. This new library enables pre-processing of MWA data 2-8 times faster than previous methods and is already integrated into the MWA system. Improvements have also been made to detect and image transient astronomical events, such as fast radio bursts (FRBs), with the MWA.

In preparation for SKA science, the SKAO conducts "data challenges" to invite the global research community to run analyses on mock datasets. The second of these challenges was held in 2021 and involved source detection/extraction from a 1 TB astronomical datacube. The AusSRC (through support from the Pawsey Supercomputing Research Centre) was one of the facility providers that enabled qualified teams ample access to supercomputing resources. We have integrated the workflow of the SoFiA team (the top ranking team to use a non-guided source extraction method during the challenge) into the AusSRC system for all users to access and utilise.

Ongoing impact

The current AusSRC system will continue to be available for science teams to work with SKA precursor data. The user interface already supports many scientists and is connected to not only the AusSRC backend system, but also to other external databases and archives, thereby unifying data and greatly facilitating scientific analysis. Several current science projects have established workflows and pipelines running on the AusSRC that have already produced published results. Other science projects are continuing to expand the functionality of the AusSRC system to handle higher data volumes as well as more diverse data products and post-processing methods. The long-term AusSRC capability will take over these projects and continue testing new ways to handle, store, and transfer data.

Detailed technical reports of all AusSRC DSP activities and projects can be found in Appendices A2-A4. These reports will be shared with the wider SKA community through the SRCNet prototyping program. Our solutions will be tested more wildly and compared to other existing facilities. With a fully working and highly modularised system, the AusSRC framework itself will also be tested by adding software and developments gleaned by other proto-SRCs. With new data actively being ingested to the AusSRC (from ASKAP and MWA), as well as a full complement of active users working on a variety of science projects, the AusSRC will be one of the primary testing centres for the SRCNet and the SKAO workflow system, providing timely and realistic feedback.

The personnel hired and trained during the AusSRC DSP will be transitioned into the long-term project. Their expertise will continue to be developed and they will also help to lead and train new personnel as the AusSRC increases its capability. The AusSRC also continues to engage with the community through events and initiatives to inspire the current and next generation of scientists and engineers.

Finances & historic timeline

The AusSRC DSP was funded by the Australian Government $2M, under grant agreement SKA75597, and CSIRO $1.8M after taxes, under a collaborative agreement, to design a long-term SRC capability and to assist ASKAP and MWA science teams with computing and data challenges. From 2019 - 2022, a total of $2.3M was spent on direct hires and contracted personnel. About $0.3M was spent on materials and other expense items (such as personal computers, travel, training, software licences, and other consumables). An annual breakdown of these expense categories can be found in Figure 1.

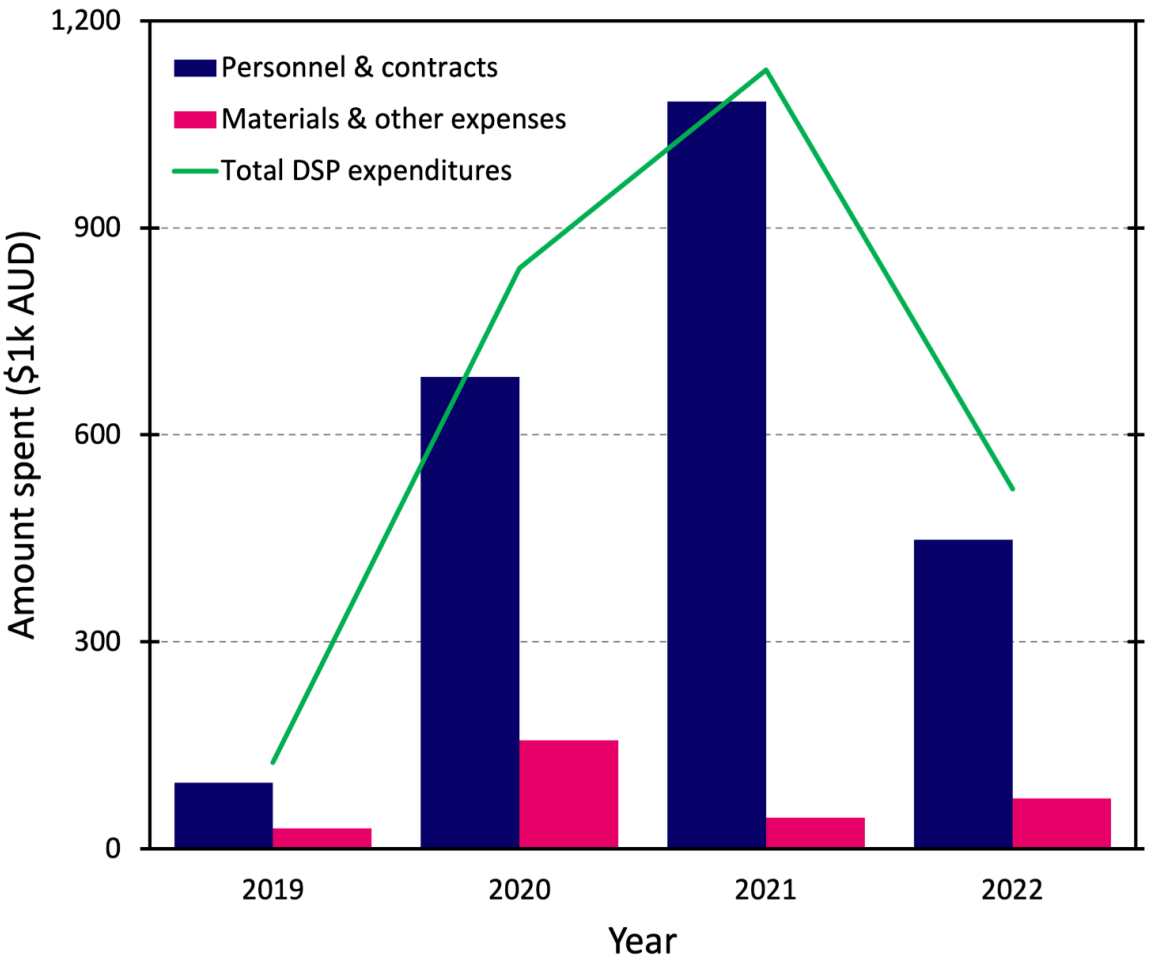

Figure 1: AusSRC DSP spending profile from 2019 - 2022. Personnel were mostly employed using short-term (1-3 year) contracts that were typically paid out on an annual basis to the respective Partner institutions.

At the end of 2022, a planned surplus (contingency funding to transition from the DSP to the long-term program) of about $1.2M will be utilised to ensure the continued employment of DSP personnel and continuity of the science support and development efforts.

Figure 2 shows a timeline of notable AusSRC activities as the DSP progressed. The initial establishment of the program took some time but as paperwork was completed, key deliverables (e.g. hiring personnel, defining requirements for an Australian-based SRC, developing a prototype system that would support SKA precursor science – with the ability to scale up for the SKAO – and establishing a governance structure along with securing funding/resources for a long-term AusSRC capability) were readily achieved.

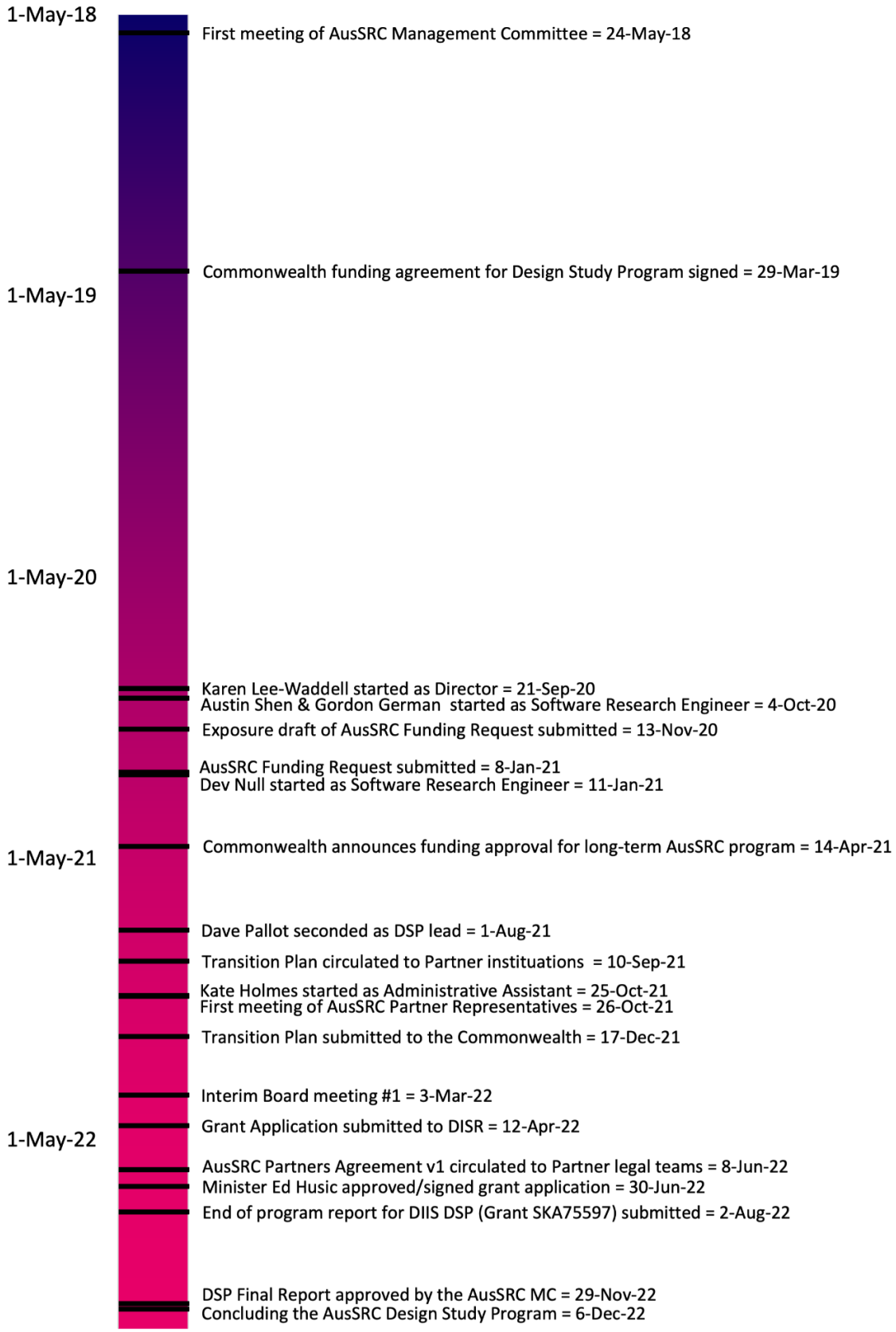

Figure 2: timeline of notable AusSRC events and activities.

Personnel

AusSRC Team

The AusSRC DSP employed 7 full-time personnel, the majority of whom appear in Figure 3, and a handful of part-time contracts/secondments – at CSIRO, Curtin University, and UWA – throughout the duration of the project. Working with various stakeholders (including local and international scientists, engineers, project managers, and government representatives), the AusSRC DSP established a general set of requirements to build a functional SRC that would eventually connect to the international SRCNet to post-process, distribute, and archive SKA science data. With these requirements in mind, the AusSRC personnel worked directly with science teams using SKA precursor telescopes, ASKAP and the MWA, to develop a preliminary framework and prototype of an SRC.

Figure 3: AusSRC DSP Team. L-R: Kate Holmes, Karen Lee-Waddell (front), Dev Null (back), Gordon German, Dave Pallot, Austin Shen. Photo taken December 2021.

Dr Gordon German is a software engineer embedded in the Space and Astronomy team at the CSIRO. He has been with the AusSRC DSP since October 2020. Gordon has a strong background in computational science and engineering, with 20 years of experience in Computational Geoscience and High performance Computing. Gordon works with the FLASH science team, providing computational support for the analysis of ASKAP survey data. He is

also heavily involved in the design and prototyping of the AusSRC data-reduction pipelines, investigating new technologies and resources for the handling and analysis of large datasets.

Kate Holmes commenced her role as Administrative Assistant for the DSP in October 2021, through the UWA. In addition to Kate's strong administration and communications skills, she has a background in astronomy/science outreach, having volunteered at the Perth Observatory and Scitech Discovery Centre. Kate provides administrative support to the Director and DSP team, alongside overseeing long-term projects such as the AusSRC website, asset management system, information management on Confluence. She regularly contributes written content and updates about the DSP to online publications such as the newsletter for the Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory.

Dr Karen Lee-Waddell started as the inaugural Director of the AusSRC in September 2020. Her research background, in radio astronomy (particularly focused on neutral atomic hydrogen (HI) in nearby interacting systems), as well as her logistical experience from 13 years of active service with the Canadian Armed Forces provide a firm foundation for her to lead the AusSRC DSP. Karen was an integral part during ASKAP's commissioning and Early Science phases and is currently the Project Scientist for WALLABY, one of the top ranked survey science projects on ASKAP. She is the Australian representative on the SRCSC and actively engages with the SKAO on various other working groups and teams.

Mx Dev Null is a software research engineer who has been embedded in the Operations team at the Curtin Institute of Radio Astronomy (CIRA) since January 2021. Dev's extensive background in cloud-native and data engineering has been a valuable asset in their work, which has enabled EoR researchers to apply novel analysis techniques at an unprecedented scale with a bespoke Nextflow pipeline. Dev's balanced application of software engineering best practices ensured the successful delivery of a Rust application (Birli) for pre-processing MWA correlator data, which is now the most widely used pre-processor within the MWA's All-sky Virtual Observatory (ASVO). Dev has also contributed towards the development of a graphical processing units (GPU)-accelerated calibration software suite (Hyperdrive; https://mwatelescope.github.io/mwa_hyperdrive/index.html) which has now become an important part of the MWA EoR pipeline.

Dave Pallot is an engineer who started at the International Centre for Radio Astronomy Research (ICRAR) in 2010. He initially joined the CIRA team to develop the MWA operations database, correlator data capture and data archive platform: the MWA ASVO. Dave spent a number of years in the MWA Operations team, where he often travelled to the Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory to help maintain and make upgrades to the instrument. He joined the ICRAR Data Intensive Astronomy team in 2014 and contributed to the Science Data Processors (SDP) and specifically the Summit EoR pipeline, which was nominated for the Gordon Bell Prize in 2020. Dave has contributed to various science and software related projects, including the Next Generation Archive System (NGAS; https://ngas.readthedocs.io/en/master/) and the DALiuGE workflow system. Dave is currently the Technical Lead within the DSP, specifically for the EMU and DINGO projects, while shaping the design for the DSP and SRCNet.

Austin Shen is a research software engineer embedded in the Space and Astronomy team at CSIRO. He has a background in astronomy and astrophysics, completing a Master of Science at UWA, and industry experience working as a data scientist and software engineer. Austin works with the WALLABY and POSSUM science teams to develop computational pipelines for their workflows. He is a co-chair of the WALLABY data archive and cataloguing working group and the AusSRC representative on the SRCSC Science Archive working group. Over the course of the DSP, Austin has contributed to the development of the AusSRC computing platform, assisted WALLABY in post-processing pilot survey phase 2 data, and contributed to three publications for the SKA Data Challenge 2, WALLABY source finding, and kinematic model public data release.

Management Committee

Activities of the AusSRC DSP were overseen by a Management Committee (MC) comprising personnel appointed by each of the parties taking part in the DSP and/or with vested interest in the AusSRC program as a whole. Each member of the MC represented the interests of the party who appointed them (i.e. AAL, ASKAP, CSIRO, Curtin University, MWA, the Pawsey Supercomputing Research Centre, and UWA). As of October 2022, the MC had the following members:

- Eric Bastholm - CSIRO
- Tom Booler - MWA
- Brad Evans - Pawsey Supercomputing Research Centre
- James Murray - AAL
- Peter Quinn - UWA (Chair)
- John Reynolds - ASKAP
- Elaine Sadler - CSIRO (Deputy Chair)
- Steven Tingay - Curtin University
- Andreas Wicenec - UWA
- Guest: Sarah Pearce - SKAO
- Guest: Mark Stickells - Pawsey Supercomputing Research Centre

Science Project Support

The AusSRC DSP provided support for several Survey Science Teams using SKA precursors: ASKAP and the MWA. A brief description of each project and its development goals with the AusSRC is presented in this section with thorough technical details provided in Appendices A2-A4.

EMU

Principal Investigator (PI) = A. Hopkins (http://emu-survey.org)
EMU is a full sky radio continuum survey conducted on ASKAP. The survey requires a post-processing pipeline to produce a detailed radio continuum catalogue with maximum

sensitivity (requiring multi-epoch data combination, referred to as mosaicking), avoiding source detection duplication and segmentation, as well as identifying associated host systems in different wavelengths.

FLASH

PIs = E. Sadler and E. Mahony (https://www.askap-flash.org)
FLASH uses ASKAP to survey the entire southern sky to detect hydrogen absorption line features. The FLASH project requires the development of a post-processing pipeline to execute pre-existing tools and workflows in an efficient and reliable manner. Support for a robust multi-wavelength database of hydrogen absorbers is also needed, as current databases and archives are limited and not well connected.

MWA EoR

PI = C. Trott
The EoR experiment using MWA relies on precision in order to detect and measure the statistical signal from a critical time in the formation of the Universe. This project requires customised processing from the data emerging from the correlator and a careful accounting of all treatments of the data throughout each processing stage.

POSSUM

PIs = B. Gaensler and G. Heald (https://possum-survey.org)
POSSUM uses ASKAP to study magnetic fields in various environments across the Universe, through synchrotron radiation and its associated Faraday rotation. This project requires a post-processing pipeline to further refine observatory-provided data products, convert and combine multi-epoch datacubes using a more modern data format, and segment/“chunk” the combined cubes for transfer to Canada.

WALLABY

PI = L. Staveley-Smith and B. Catinella (https://wallaby-survey.org)
WALLABY will survey the entire southern sky to detect hydrogen emission from hundreds of thousands of nearby galaxies. This project requires a post-processing pipeline that is capable of working with large (~1 TB) datacubes that can combine multi-epoch datacubes, run automated pipelines – for source detection, parameterisation, and kinematical analysis – and transfer/sync advanced data products across archives located in Australia, Canada, and Spain.

PaCER projects

The Pawsey Centre for Extreme Scale Readiness (PaCER) program was established in late 2020 to prepare the Australian research community “to achieve extreme performance on Pawsey’s next-generation supercomputer” by providing an opportunity for researchers to become exascale-ready through the development of new algorithms, workflows, and data

pipelines (https://pawsey.org.au/PaCER/). After a competitive proposal process, two AusSRC co-funded projects were approved for the program and awarded resources: Breakthrough Low-latency Imaging with Next-generation Kernels (BLINK) and HI Visibility Imaging for the SKA (HIVIS).

BLINK

PI = M. Sokolowski
Based on time-domain data to detect transient sources using MWA, this project combines advanced data processing technology offered by next-generation supercomputers and new data processing algorithms that have been optimised for both speed and sensitivity to detect and process transient signals. The developed pipeline will enable real-time image-based transient searches (for pulsars, gamma ray bursts, and FRBs) in MWA and future SKA-Low data.

HIVIS

PI = M. Meyer (https://dingo-survey.org)
Using data from the DINGO survey on ASKAP, this project addresses a significant challenge for the SKA – how to optimally image multi-epoch data sets – by developing a sparse data storage and processing pipeline based on uv-grids. This technology could reduce visibility storage requirements by an order of magnitude and will enable deep direct imaging of hydrogen.

Support Overview

A brief summary of the technical developments for each science project is provided in the table below.

Project	Technical developments	Computing resources	External collaborators
EMU	· Production of continuum 'super mosaics' (tiling adjacent fields) · Continuum source finding · Pipelined workflow: CSIRO ASKAP Science Data Archive (CASDA) download → mosaic large fields → source finding → databasing → multi-wavelength cross-matching with data archived at Data Central	· NCMAS (ASKAP science allocation @ Pawsey) · Pawsey - Nimbus (AusSRC allocation)	· AAO & Data Central (Australia) · NOAO (USA) · NRAO (USA)

Project	Technical developments	Computing resources	External collaborators
FLASH	· Absorption line detection pipeline containerisation and parallelisation · Consolidated archiving of all known HI absorption sources · Pipeline workflow: CASDA download → absorption feature source finding using continuum source catalogues → false source rejection sub-workflow → databasing	· NCMAS (ASKAP science allocation @ Pawsey) · Pawsey - Nimbus (AusSRC allocation)	
MWA EoR	· Development of Birli application (Rust library for processing MWA) · EoR quality analysis pipeline: download data → pre-processing with Birli → flag → calibrate → analyse power spectrum metrics · Development of measurement set support and direction dependent calibration in Hyperdrive	· DUG Technology – commercial supercomputing facility · MWA ASVO and Web Services · Pawsey - Garrawarla	
POSSUM	· YANDAsoft (imaging pipeline) development, HEALPix format implementation · Chunking of continuum cubes for data transfer overseas · Pipelined workflow: CASDA download → mosaic continuum cubes → convert to HEALPix → chunking → data transfer to CADC	· NCMAS (ASKAP science allocation @ Pawsey) · Pawsey - Nimbus (AusSRC allocation)	· CADC/CIRADA (Canada)
WALLABY	· Spectral line source finding with refactored SoFiA-2 application · Data transfer and replication (Australia-Canada-Spain) · Pipelined workflow: CASDA download → mosaic multi-epoch spectral line cubes → source finding → databasing → data transfer/replication across various data centres	· NCMAS (ASKAP science allocation @ Pawsey) · Pawsey - Nimbus (AusSRC allocation)	· AAO & Data Central (Australia) · CADC/CIRADA (Canada) · IAA-CSIC/ SPSRC (Spain)

Project	Technical developments	Computing resources	External collaborators
PaCER: BLINK	· High-time resolution imaging · Time-domain source finding at low frequencies · Development of pipeline to search for transients data in large volumes of MWA voltage capture data and then process the observations for each detection	· Pawsey - Garrawarla · Pawsey - PaCER allocation	
PaCER: HIVIS	· uv-gridding of spectral line data · Implementation large-scale computing applications/tools: DALiuGE execution framework and Adaptable I/O System (ADIOS)	· NCMAS (ASKAP science allocation @ Pawsey) · Pawsey - Nimbus (AusSRC allocation) · Pawsey - PaCER allocation	· ORNL (USA)
All projects	· User interface with NextFlow implementation · Integration of Jupyter notebooks to query Structured Query Language (SQL) databases and perform preliminary analysis · Installation of visualisation tools such as Aladin and the Cube Analysis and Rendering Tool for Astronomy (CARTA)	· Pawsey - Nimbus (AusSRC allocation)	

Community engagement

AusSRC personnel continue to show their passion for the project and SKA astronomy in general at various conferences and events. Even during the COVID-19 era (of limited travel and in-person contact), online media platforms ensure continued engagement with different audiences. From virtual classroom visits to contributions to overseas expos and conferences, AusSRC personnel showcased the activities and future ambition of the AusSRC and SKAO project as a whole. Appendix A4 lists the details of engagement activities over the past few years.

SKAO Data Challenge

As a part of the science preparatory activities, the SKAO runs “data challenges” for the science community. The purpose of these challenges is to inform the development of the data reduction pipelines for the SDPs and SRCs, enabling better support for the intended science that will be achieved with the SKA telescopes. These challenges also allow the science community to

familiarise themselves with the standard products that the SKAO will deliver and optimise their own scientific analysis tools accordingly.

SKAO Data Challenge 2 (SDC2; https://sdc2.astronomers.skatelescope.org/sdc2-challenge) was held between February and July 2021. The AusSRC (with assistance from the Pawsey Supercomputing Research Centre) was an official resource provider, as shown in Figure 4. The competition required researchers to perform scientific source detection and parameter extraction on a 1 TB simulated spectral line datacube. Over 40 teams had originally signed up but only a dozen completed the contest.

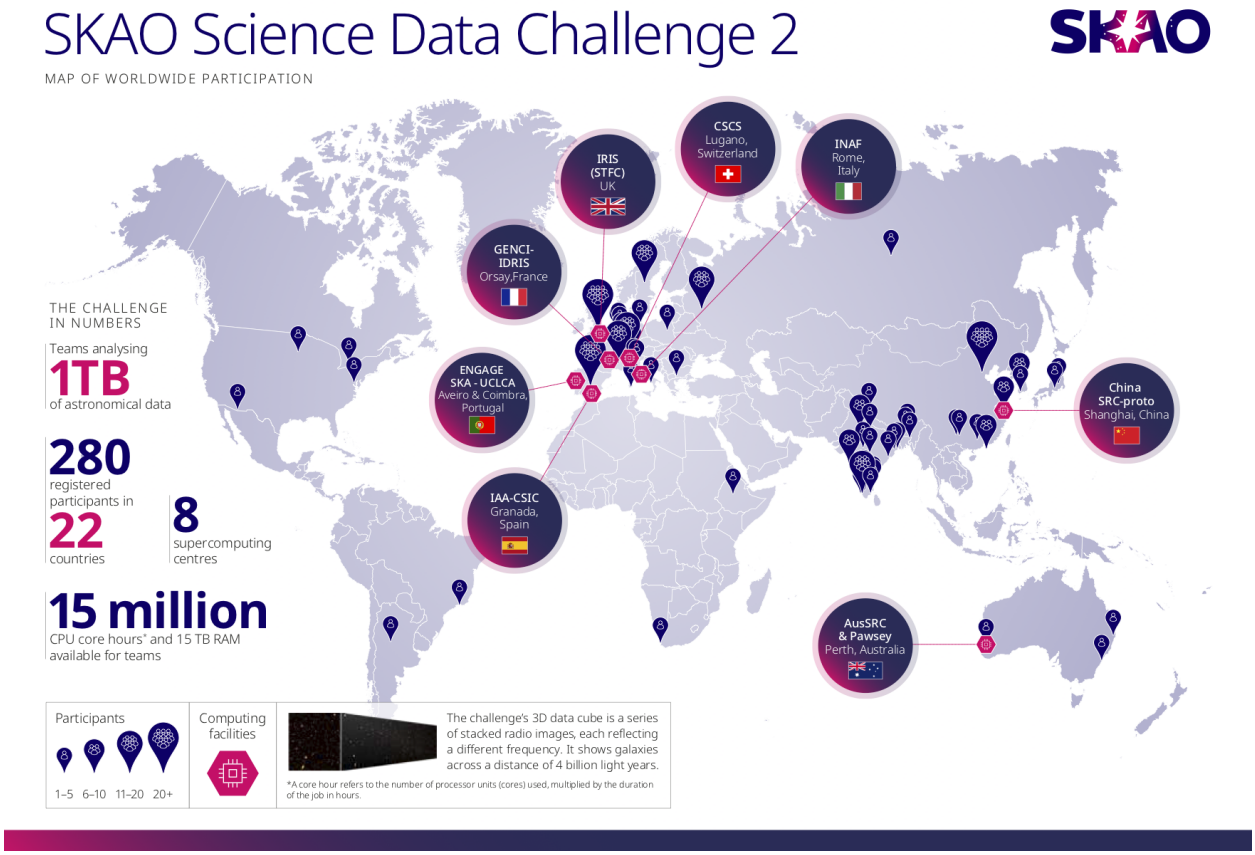

Figure 4: infographic about the SKAO Science Data Challenge 2 showing the locations of the participants as well as the resource providers.

The third place team, “SoFiA”, was the top ranking team to use a non-guided source extraction method during the challenge, the SoFiA-2 application (https://github.com/SoFiA-Admin/SoFiA-2), and was also the team supported by the AusSRC. After the competition had finished, the AusSRC fully integrated SoFiA-2 post-processing methods into various science teams’ workflows.

The top two teams in SDC2 used machine learning (ML) applications. Their high success rate is largely attributed to the completeness of the training set data, replicating the type of detections in the larger datacube (a publication of the exact methods utilised is currently in preparation by

the SKAO). Further research in both ML technologies as well as assembling robust training sets for various science objectives should be incorporated in future challenges and then applied to SRC workflows.

Next steps for the AusSRC

Long-term AusSRC Program

The long-term AusSRC program will be funded by the Australian Government through a DISR Grant (https://www.grants.gov.au/Go/Show?GoUuid=5f8ff1bd-cf63-458f-acda-38dac0d1a665) with in-kind contributions from the four Foundation Partners: CSIRO, Curtin University, the Pawsey Supercomputing Research Centre, and UWA. The program will have two 5-year phases, following the timeline outlined in the international SRC White Paper (V1.0 May 2020). Phase 1 will focus on developing and building up SRC functionality through supporting SKA precursor science from ASKAP and the MWA. Phase 2 will involve rapid increases of capacity and shifting focus towards a service delivery model in order to support full science post-processing of SKA data, as part of the global network of SRCs. The goal is to have the AusSRC supporting ~14% of the capability of the total SRCNet infrastructure by ~2030, with additional resources available for local initiatives.

In order to formally establish the long-term AusSRC program, two key documents need to be completed: the Government Funding and AusSRC Partners Agreements. These agreements should be finalised in (early) 2023. Once that happens, the AusSRC Board will be established to oversee the operation and management of AusSRC.

Handover process

During the transition period between the AusSRC DSP and long-term program, an Interim Board – comprising an independent Chair and one representative from each of the four Foundation Partners – has been established to help guide the development of the long-term program agreements. Once the AusSRC DSP Final Report (i.e. this document) has been finalised and accepted by the MC, high-level responsibility and governance of the carry-over DSP funds, personnel, and science/technical projects will be transferred to the Interim Board.

Appendices

A1. Glossary

Abbreviations and acronyms

AAL	Astronomy Australia Limited
AAO	Australian Astronomical Observatory
ACAMAR	Australia-China Consortium for Astrophysical Research
ADASS	Astronomical Data Analysis Software and Systems
ADP	Advanced Data Products
ADQL	Astronomical Data Query Language
AllWISE	Wide-field Infrared Survey Explorer source catalogue
API	Application Programming Interface
ART	Agile Release Train
ASA	Astronomical Society of Australia
ASCC	Australian SKA Coordination Committee
ASKAP	Australian SKA Pathfinder
ASVO	All-sky virtual observatory
ATNF	Australia Telescope National Facility
AusSRC	Australian SKA Regional Centre
AWS	Amazon Web Services
BDA	Baseline Dependent Averaging
BLINK	Breakthrough Low-latency Imaging with Next-generation Kernels
CADC	Canadian Astronomy Data Centre
CASDA	CSIRO ASKAP Science Data Archive
CHAD	Consolidated HI Absorption Database
CIRA	Curtin Institute of Radio Astronomy
CIRADA	Canadian Initiative for Radio Astronomy Data Analysis
CPU	Central processing unit
CSIRO	Commonwealth Scientific and Industrial Research Organisation

DAG	Directed Acyclic Graph
DBaaS	Database-as-a-Service
DESI	Dark Energy Spectroscopic Instrument
DIIS	Department of Industry, Innovation and Science
DINGO	Deep Investigation of Neutral Gas Origins
DISR	Department of Industry, Science and Resources
DSL	Domain Specific Language
DSP	Design Study Program
EMU	Evolutionary Map of the Universe
EMUCat	EMU Value-added Catalogue
EoR	Epoch of Reionisation
eROSITA	Extended Roentgen Survey with an Imaging Telescope Array
FITS	Flexible Image Transport System
FLASH	First Large Absorption Survey in HI
FoV	Field of view
FRB	Fast Radio Burst
GLEAM	GaLactic and Extragalactic All-sky MWA survey
GLEAM-X	GaLactic and Extragalactic All-sky MWA survey - eXtended
GPU	Graphical Processing Units
HEALPix	Hierarchical Equal Area isoLatitude Pixelation
HI	Neutral atomic hydrogen
HIVIS	HI Visibility Imaging for the SKA
HPC	High performance computing
IAA-CSIC	Institute of Astrophysics of Andalusia - Spanish National Research Council
IaaS	Infrastructure-as-a-Service
IAM	Identity and Access Management
ICRAR	International Centre for Radio Astronomy Research
IdP	Identity provider

I/O	Input/output
JSON	JavaScript Object Notation
JWT	JSON Web Token
LHC	Large Hadron Collider
LoBES	Long Baseline Epoch of Reionisation Survey
LoTSS	LOFAR Two-metre Sky Survey
LSST	Large Synoptic Survey Telescope
MC	Management Committee
MFA	Multi-factor Authentication
ML	Machine learning
MPI	Message Passing Interface
MWA	Murchison Widefield Array
MWAX	Murchison Widefield Array correlator
NCMAS	National Computational Merit Allocation Scheme
NOAO	National Optical Astronomy Observatory (USA)
NRAO	National Radio Astronomy Observatory (USA)
NVSS	NRAO VLA Sky Survey
ODP	Observatory Data Products
OH	Hydroxyl
OIDC	OpenID Connect
ORM	Object-relational mapper
ORNL	Oak Ridge National Laboratories
PaCER	Pawsey Centre for Extreme Scale Readiness
PanSTARRS	Panoramic Survey Telescope and Rapid Response System
PI	Principal Investigator
POSIX	Portable Operating System Interface
POSSUM	Polarisation Sky Survey of the Universe's Magnetism
RAM	Random access memory

RFI	Radio frequency interference
SAML	Security Assertion Markup Language
SDC2	SKAO Data Challenge 2
SDP	Science Data Processors
SIAP	Simple Image Access Protocol
SKAO	SKA Observatory
SMART	Southern-sky MWA Rapid Two-Meter survey
SODA	Server-side Operations for Data Access
SPSRC	Spanish prototype of an SKA Regional Centre
SQL	Structured Query Language
SRC	SKA Region Centre
SRCNet	SKA Regional Centre Network
SRCSC	SKA Region Centre Steering Committee
SSAP	Simple Spectral Access Protocol
SSINS	Sky-Subtracted Incoherent Noise Spectra
SSO	Single sign-on
SUMSS	Sydney University Molonglo Sky Survey
TAP	Table Access Protocol
UWA	University of Western Australia
VM	Virtual Machines
VO	Virtual Observatory
WALLABY	Widefield ASKAP L-band Legacy All-sky Blind surveY
WAVES	Wide Area Vista ExtraGalactic Survey
WCS	World Coordinate System

Software and technologies

	Usage	Website
ADIOS	Framework for scientific data management	https://csmd.ornl.gov/software/adios2
Aladin	Interactive astronomical sky atlas	https://aladin.u-strasbg.fr
AOFlagger	Removing radio-frequency interference from astronomical datasets	https://aoflagger.readthedocs.io/en/latest/
ASKAPsoft	Suite of processing software developed to handle ASKAP data	https://www.atnf.csiro.au/computing/software/askapsoft/sdp/docs/current/pipelines/introduction.html
AusSRC Docker Hub	Docker container repository for AusSRC applications and tools	https://hub.docker.com/u/aussrc
AusSRC Github	Code repository for AusSRC pipelines and workflows	https://github.com/AusSRC
Birli	Library for MWA processing tasks	https://github.com/MWATelescope/Birli
Bucardo	Data replication	https://bucardo.org
CARTA	Image visualisation and analysis tool	https://cartavis.org
CASAcore	Suite of C++ libraries for radio astronomy data processing	https://casacore.github.io/casacore/
Confluence	Collaboration tool	https://www.atlassian.com/software/confluence
CUDA	NVIDIA toolkit for development on GPUs	https://developer.nvidia.com/cuda-toolkit
Data Aggregation Service	Data Central's web-based astronomy data catalogue	https://das.datacentral.org.au/das
DALiuGE	Workflow graph development, management, and execution framework	https://daliuge.readthedocs.io/en/latest/
Django	Python-based web framework	https://www.djangoproject.com
Docker	Containerisation platform	https://www.docker.com
Dropbox	File hosting service	https://www.dropbox.com
EAGLE	Visual workflow editor	https://eagle-dlg.readthedocs.io/en/master/
FLASK	Python framework	https://flask.palletsprojects.com/en/2.2.x/
FreeIPA	Identity management system	https://www.freeipa.org/
GitHub	Version controlled development platform	https://github.com

GitLab	Version controlled development platform	https://www.gitlab.com/gitlab
HEALPix	Pixelation method to subdivide a spherical surface	https://healpix.sourceforge.io
Hyperdrive	Calibration software for the MWA	https://mwatelescope.github.io/mwa_hyperdrive/index.html
JIRA	Collaboration tool	https://www.atlassian.com/software/jira
Jupyter	Interactive Python-based computing platform	https://jupyter.org
MATLAB	Programming and numeric computing platform	https://www.mathworks.com/products/matlab.html
Microsoft Visual Studio	Integrated development environment	https://visualstudio.microsoft.com
MultiNEST	Bayesian inference tool	https://cosmosis.readthedocs.io/en/latest/reference/samplers/multinest.html
Nextflow	Pipeline workflow implementation	https://www.nextflow.io
NGAS	Archive handling and management	https://ngas.readthedocs.io/en/master/
Open OnDemand	Enables remote access to supercomputers	https://openondemand.org
OpenMP	Multi-platform shared memory parallel processing	https://www.openmp.org
OpenStack	Cloud computing infrastructure	https://www.openstack.org
PostgreSQL	Object-relational database	https://www.postgresql.org
PRESTO	Pulsar searching toolkit	http://ascl.net/1107.017
RabbitMQ	Message queuing protocol	https://www.rabbitmq.com
RACS-tools	Processing tasks for ASKAP	https://github.com/AlecThomson/RACS-tools
Rucio	Data management tool	https://rucio.cern.ch
Selavy	Source finding tool	https://www.atnf.csiro.au/computing/software/askapsoft/sdp/docs/current/analysis/selavy.html
Singularity	Containerisation system for HPCs	https://apptainer.org
Slurm	Workload manager	https://slurm.schedmd.com/documentation.html
SoFiA/SoFiA-2	3D source finding application	https://github.com/SoFiA-Admin/SoFiA-2
SSH	Network protocol for remote login and command-line execution	https://www.ssh.com/academy/ssh/protocol
SWIG	Connect programs and libraries written in different languages	https://www.swig.org

TOPCAT	Interactive graphical viewer and editor	http://www.star.bris.ac.uk/~mbt/topcat/
uWSGI	Building hosting services for full stack development	https://uwsgi-docs.readthedocs.io/en/latest/
VisIVO	Integrated visualisation tool	http://palantir7.oats.inaf.it/visivoweb/
WSClean	Algorithm for stacking and convolving radio astronomy images	https://wsclean.readthedocs.io/en/latest/
XID	Unique identification generator library	https://github.com/rs/xid
YANDASoft	Imaging data from radio telescopes such as ASKAP	https://readthedocs.org/projects/yandasoft/

A2. Technical Reports - SRC Prototyping

AusSRC proto-SRC system

Project overview

The Australian-hosted SKA-Low telescope will produce around 300 PB per year of Observatory Data Products (ODP) that science teams around the globe will need to readily access. Another 300 PB will be produced by the SKA-Mid in South Africa. The SKAO has no provision to store this data long term. Instead, SKA member states are forming a collaborative network of SRCs to design, build, deliver, and operate end-to-end support for ODPs, archives, and associated services. SRCs will:

- Store, publish and curate SKA ODP, Advanced Data Products (ADP) and associated metadata for the long-term;
- Provide data workflow and data dissemination solutions;
- Provide compute, data storage and data visualisation resources; and
- Provide science support.

A significant body of work has been undertaken to develop the SRC concept. The program uses a top-down analysis of SRC requirements in global collaboration with other SRCs and the SKAO, and a bottom-up approach to solving practical computational and data problems within the context of the SKA precursor projects. These approaches have led to the design and prototyping of the architecture of the future AusSRC. The DSP has identified, assessed, and tested technical solutions. These solutions will form the basis of SRCs that will help facilitate the discovery of new scientific insights.

Technical developments during DSP

The SRCs will be SKA-specific data science platforms that will provide users with complete software and hardware environments that contain the tools required to interact with ODPs and ADPs within the SRC system. The ODPs produced from the SKA SDP will be ingested and stored within the SRCNet's data archives, ready for discovery and post-processing. Given the data is too large to move to the scientists, it is important that the SRC architecture allows the scientist to bring their code to the data. This makes automation of post-processing significantly more accessible, faster, and overall more efficient as it will minimise expensive data movement operations. The SRCNet will support various Virtual Observatory (VO) compliant services, such as image cutout and catalogue services, that can interoperate seamlessly with the vast ecosystem of astronomy tools that have been developed and matured over many years.

Platform Design

The AusSRC is primarily web based with a wide range of programmatic and graphical interfaces. The AusSRC platform was designed with consistency, reproducibility, usability, reliability, and scalability in mind.

The AusSRC science processing platform contains 3 main components: SRC science interfaces and portal, SRC cloud services, and the SRC HPC processing platform. The SRC science interfaces and portal are a series of web pages, application programming interface (APIs), command line tools, VO tools and applications that enables scientists to interact with the processing platform, services, and data. This interface allows the data within and external to the SRC's to be aggregated and collated to produce science-ready data products and knowledge. It allows scientists to deploy their processing pipelines on the HPC processing platform.

The SRC cloud infrastructure hosts the virtual environment and machines necessary to operate the various databases, science web portals, notebooks, VO compliant science archive services, visualisation platforms, and other SRC long running managed services.

The SRC HPC processing platform allows scientists and developers to deploy their computation workflows that contain the codes, running inside Singularity containers, to reduce the ODPs contained within the SRC archives (ODPs are deposited into the SRC archives by the SDP). These codes contain the algorithms that analyse and extract the science components that are populated in the various databases, images catalogues, and visualisation services for later analysis. Figure A2.1 shows the basic architecture of the AusSRC processing science platform.

Platform Users

There are five main types of users accessing AusSRC resources, each with different requirements and ways of interacting with the systems. They are:

- PIs;
- Scientists;
- Developers;
- SRC engineers; and
- SRC administrators (operations)

A PI is typically a senior scientist that is responsible for a particular science project and other scientists/developers attached to the project. The science projects that are selected as SRC partner projects will typically have a single PI that will work with an SRC engineer to communicate requirements, provide feedback and monitor progress. A science project will typically have developers and or scientists who are working directly with an SRC engineer to maximise the efficiency of their workflows and get the services they require based on the science requirements. i.e. HPC resource and storage allocation, database systems, visualisation applications, VO interfaces etc. Within the context of the SRC, the PI's will be able to grant and revoke access to resources that have been assigned to their science project, for example, adding a new scientist to the group that can access project science data and services.

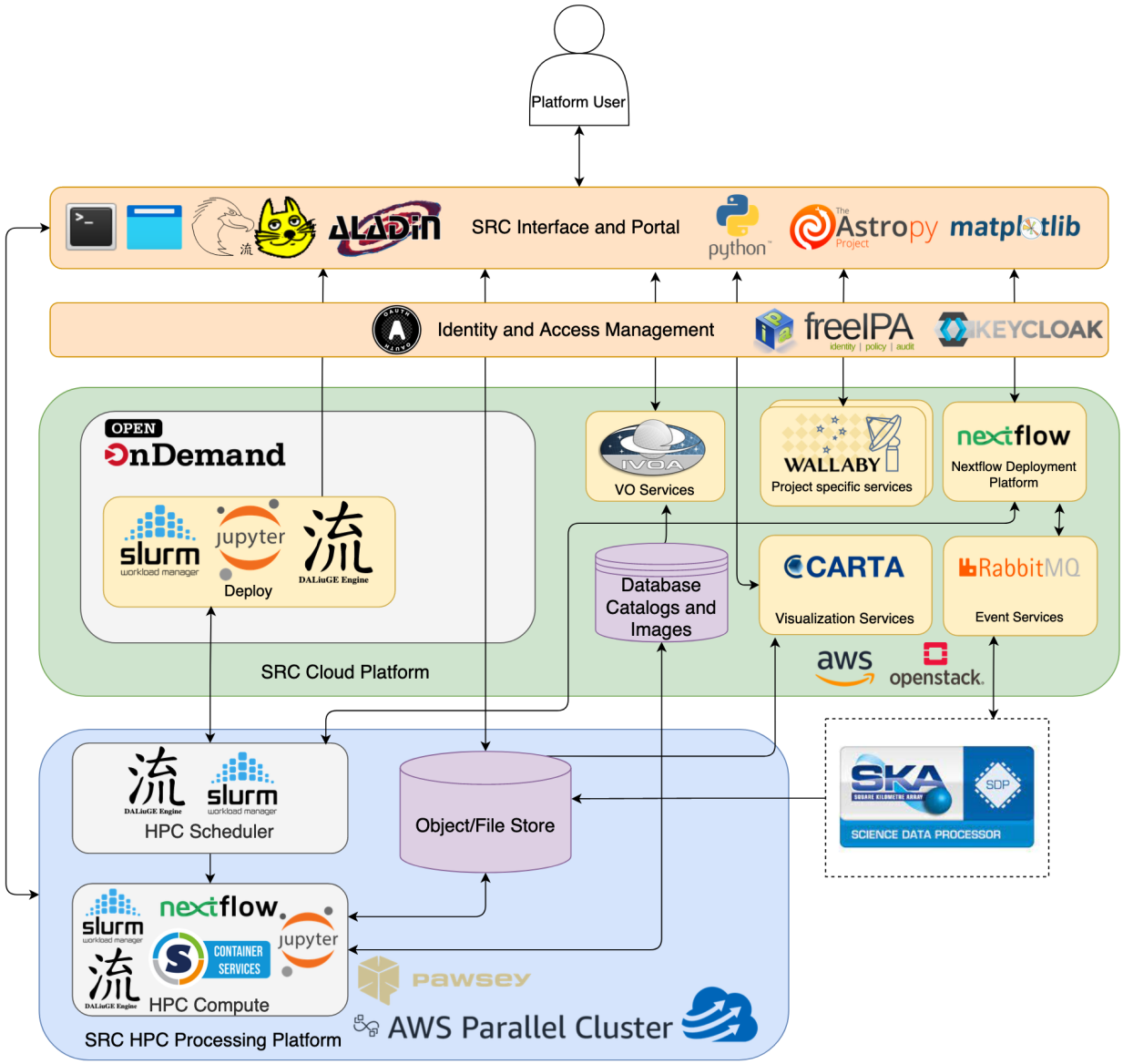

Figure A2.1: AusSRC platform design.

Scientists are users that are interested in particular data sets and services provided by the SRCs. Scientists could act as developers of computational workflows and components and therefore need an SRC allocation of compute and storage with the correct privileges. They could also be exclusive users of the VO, which includes archive and visualisation services used for data discovery, introspection, and science production. It is the intention of the SRCs to help scientists upskill by providing an intuitive set of tools and software practices.

SRC engineers are employees of an SRC and can be experts in science, engineering, and/or system administration. Many of the SRC engineers will be developers who are attached to science projects while others may be more internally facing to make enhancements to the SRC infrastructure.

SRC administrators are responsible for the ongoing maintenance of SRC resources. Some key responsibilities include, but are not limited to, the management of user accounts, resource authorisation and allocation, system maintenance, and deployment of software and hardware upgrades to existing infrastructure (commonly referred to as the DevOps process).

Identity and Access Management

Identity and Access Management (IAM) is a set of procedures and tools that help manage digital identities and how these identities can access critical computing services and information within an organisation. IAM tools allow administrators to assign identities, authentication through a single set of credentials, authorise them to a resource, and monitor identities through their lifetime. IAM provides secure access to devices, customers, workers, business partners, suppliers, mobile users, and infrastructure such as code API and microservices. IAM ensures a seamless and friction-less experience when accessing a multitude of different systems required for a productive and functioning platform.

The process of verifying an entity is who they say they are is called authentication. IAM systems perform authentication via single sign-on (SSO) systems, two- or multi-factor authentication, and privileged access management. These technologies provide the ability to securely store identity and profile data as well as data governance functions to ensure that only data that is necessary and relevant is shared. Authorisation is the process of ensuring that an identity can perform only the tasks they need to or allowed to perform on certain resources.

AusSRC provides a SSO authentication service via Keycloak that relies on the concept of federated identity. Federated Identities allow digital identities on two or more different identity providers (IdPs) to be linked together in a trusted environment. This means that a digital identity can provide their credentials (username and password), ticket, or token obtained via a trusted IdPs used to authenticate to the AusSRC platform. An identity federation enables AusSRC administrators to create, apply, and revoke permissions from a single location making it easier to manage access. Keycloak supports multiple IdPs including, but not limited, to Google, Facebook, Github, Universities, Governments, and other organisations allowing for a multitude of disparate digital identities to access AusSRC resources.

The process of authentication is performed via well known authentication protocols that include Security Assertion Markup Language (SAML) and/or OpenID Connect (OIDC). User access is granted using a least-privilege approach, where an identity is only given those privileges needed for them to complete their task, with best practice including password renewal and Multi-factor Authentication (MFA). Programmatic access that includes API calls to AusSRC services are performed using temporary and limited-privilege credentials such as those issued by a Security Token Service in the form of a JSON Web Token (JWT). Figure A2.2 demonstrates the steps when an unauthenticated identity wants to use an AusSRC service be that a VO service, object store, OpenStack (https://www.openstack.org) provisioning platform, or micro service.

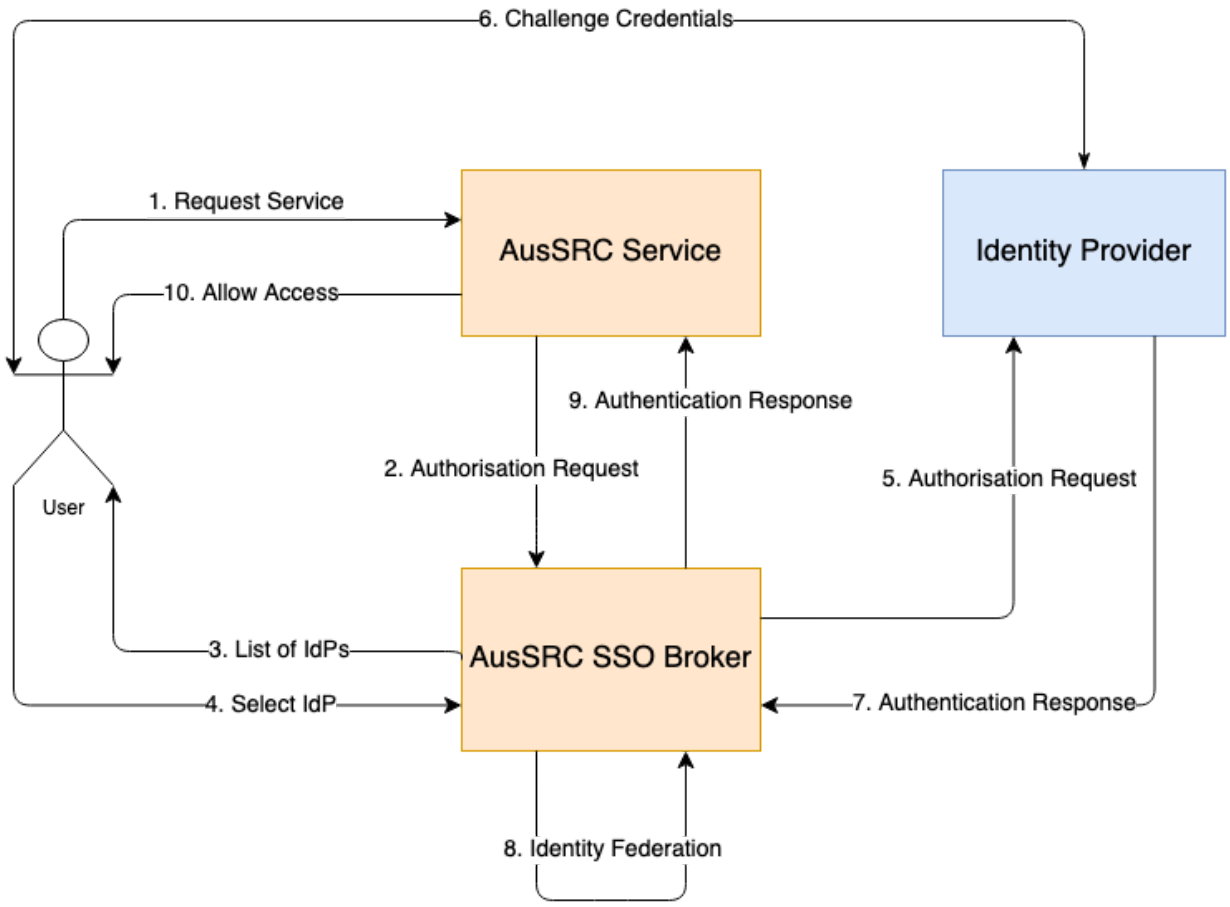

Figure A2.2: Authentication flow.

AusSRC operates an internal FreeIPA (https://www.freeipa.org/) IdP that provides an integrated Identity and Authentication solution for Linux/UNIX networked environments. FreeIPA digital identities with common security requirements are placed in groups or roles that operate similar to Linux groups. AusSRC uses these groups to control access to resources such as files and directories on AusSRC shared file systems. This method allows administrators and PIs to centrally manage access by changing an entity's group membership or attributes once, rather than updating many individual policies when an entity's access needs change. This process is known as Role-based Access Control. Keycloak supports fine-grained authorization policies for OIDC based AusSRC services and is able to combine different access control mechanisms. Figure A2.3 highlights the AusSRC integrated Identity and Authentication service.

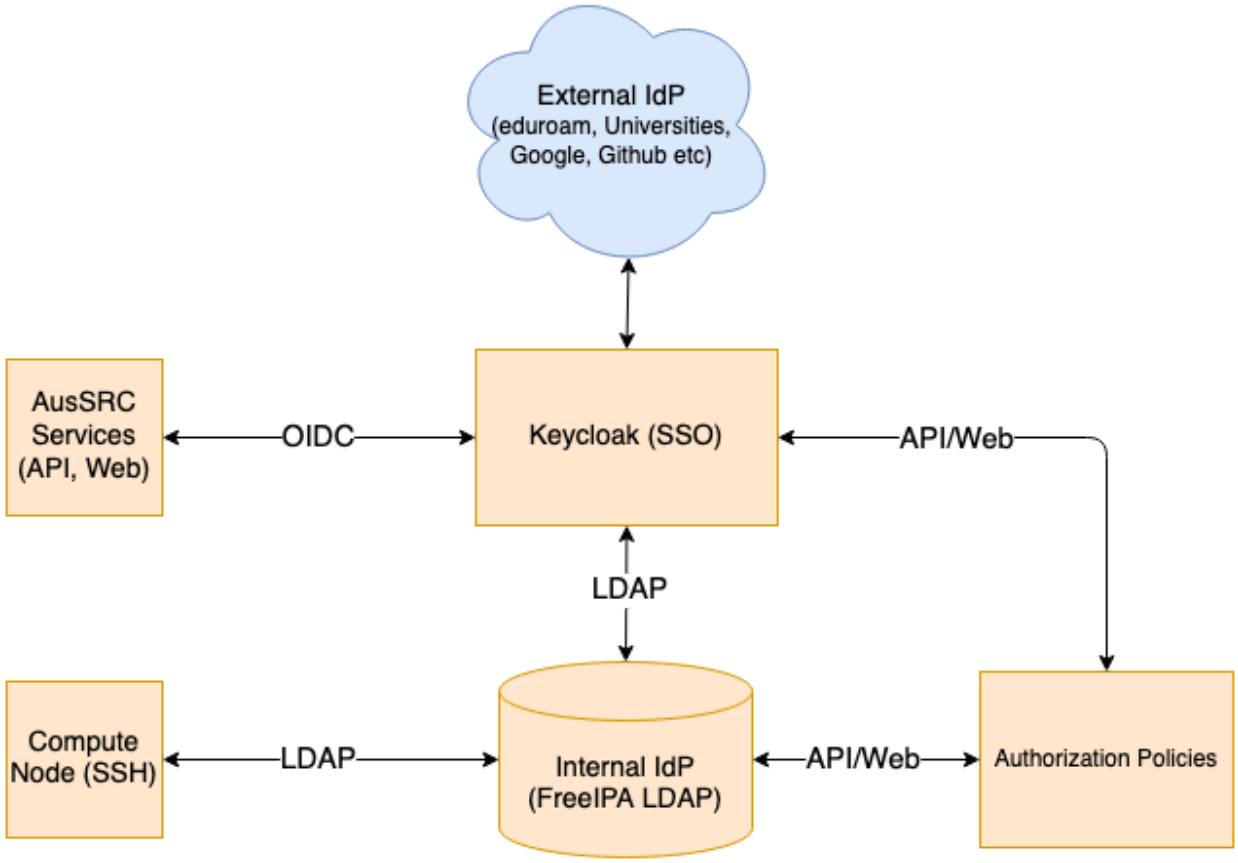

Figure A2.3: AusSRC integrated Identity and Authentication service.

Technology/applications involved and/or tested

During the DSP, various technologies have been investigated and prototyped for the AusSRC proto-SRC system. A brief description of the key (and possibly less common) technologies is provided in this subsection.

The DALiuGE System

DALiuGE is a workflow development, scheduling, and execution system developed by ICRAR to deal with the extreme scale of SKA workflows (Figure A2.4). The system integrates many of the concepts also enabled by Nextflow and other scientific workflow systems, while potentially avoiding scalability bottlenecks. The most visible difference for DALiuGE is its visual workflow editor, EAGLE (https://eagle-dlg.readthedocs.io/en/master/), which allows users to develop complex scientific workflows in a graphical way, while keeping full control of the details of each single component. DALiuGE's other design principle was a complete separation of concerns to allow people with expertise in special aspects to concentrate on those areas without having to bother too much about the rest of the system. In this way, algorithmic development and platform optimisation is completely separate from the development of the workflow and execution control.

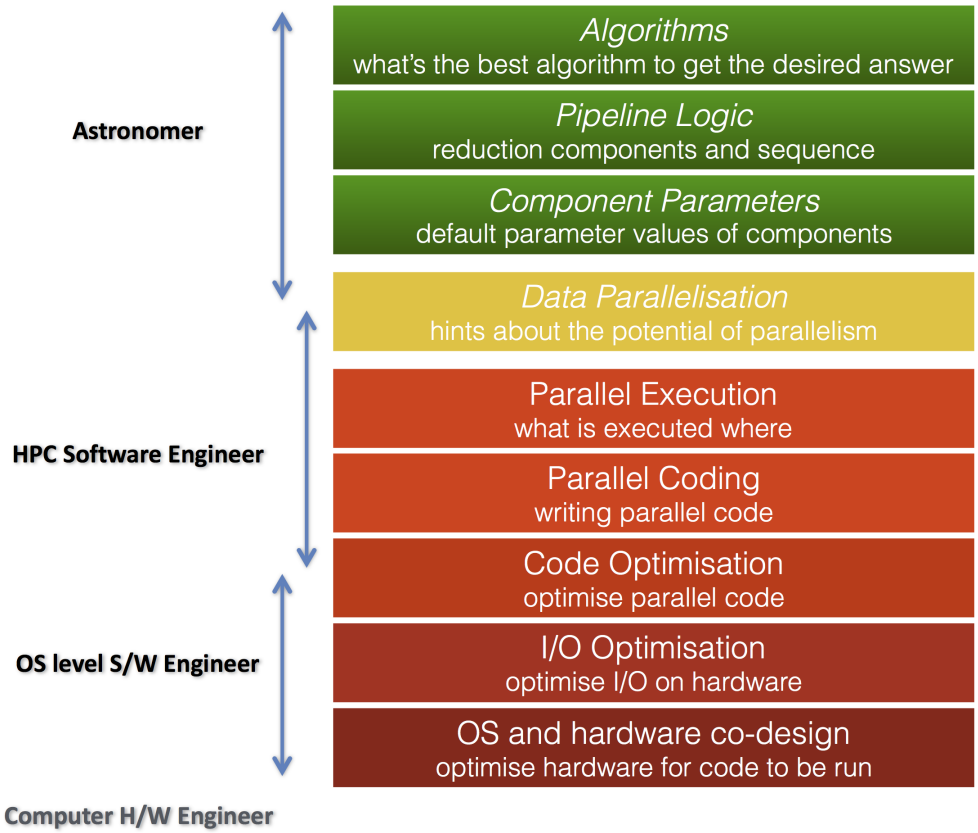

Figure A2.4: Separation of concerns; DALiuGE implements clean and open interfaces to allow people to work exclusively in their area of expertise.

DALiuGE was also designed to enable easy re-use of existing software. It can launch and execute dockerized and shell-based applications. It can execute more integrated applications, which make use of memory rather than files for inter-communication. It is possible to integrate MPI based applications as well as direct shared library applications written in C/C++ or Python. It is also possible to run an integrated shared memory manager across a whole cluster, or use the Apache Plasma Store to provide similar functionality. Existing Python packages can be analysed using an utility tool, which translates the source code into a component palette. These palettes can be loaded into EAGLE and then used to develop workflows with components from those packages. For many existing software packages, this allows integration into the DALiuGE system within a few minutes.

The DALiuGE system, as depicted in Figure A2.5, is integrated in the deployment of the AusSRC. The EAGLE editor is running openly on an AWS instance and then connects to a password protected deployment web service within OpenOnDemand (https://openondemand.org). That web service in turn creates and submits the workflow task script to Slurm, which allocates the required machines. Once activated by Slurm, the workflow script first initialises the DALiuGE workflow managers and then deploys the execution of the actual workflow. The workflow managers trigger and monitor the workflow execution. Once

finished, the Slurm script will shutdown the managers and hand back the allocated resources to Slurm.

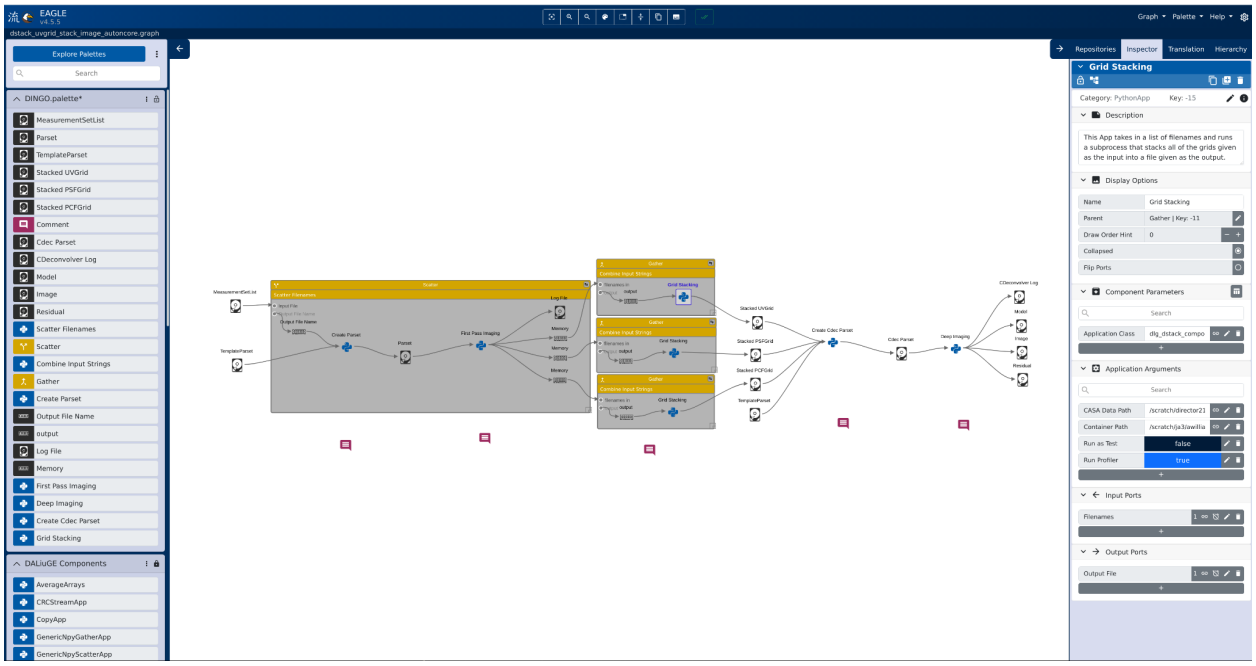

Figure A2.5: The DALiuGE EAGLE editor with the palette pane on the left, the main graph development canvas in the centre and the component inspector on the right. The yellow boxes represent so-called constructs, allowing the users to specify areas of the graph where scattering and gathering is happening.

In order to demonstrate the scalability capabilities, the DALiuGE system had been used to drive the largest scale radio astronomy workflow ever executed. That experiment was nominated for the Gordon Bell prize 2020. The machine used, SUMMIT, was the fastest supercomputer in the world at the time (currently second) and during the experiment, DALiuGE used almost all of the available compute nodes (4560) and a total of 27360 GPUs. On the other extreme end the DALiuGE system can happily be deployed on a single Raspberry Pi or laptop computer, since resource usage is minimal.

Jupyter Notebooks

Jupyter Notebooks are a web-based development environment for data science, scientific computing and machine learning. Notebooks allow users to explore scientific data in an interactive and graphical way, making it easier to communicate findings to others in a convenient manner. The Jupyter Notebook platform supports a myriad of popular scientific languages such as Python, C/C++, Ruby, R, Rust, Java et al.

The AusSRC Open OnDemand portal allows users to deploy notebook instances directly onto a HPC processing cluster where the data can be accessed (see Figures A2.6.1 and A2.6.2). This arrangement allows for the efficient introspection of large scientific data products without having to move it to other platforms for manipulation, which is time consuming and expensive.

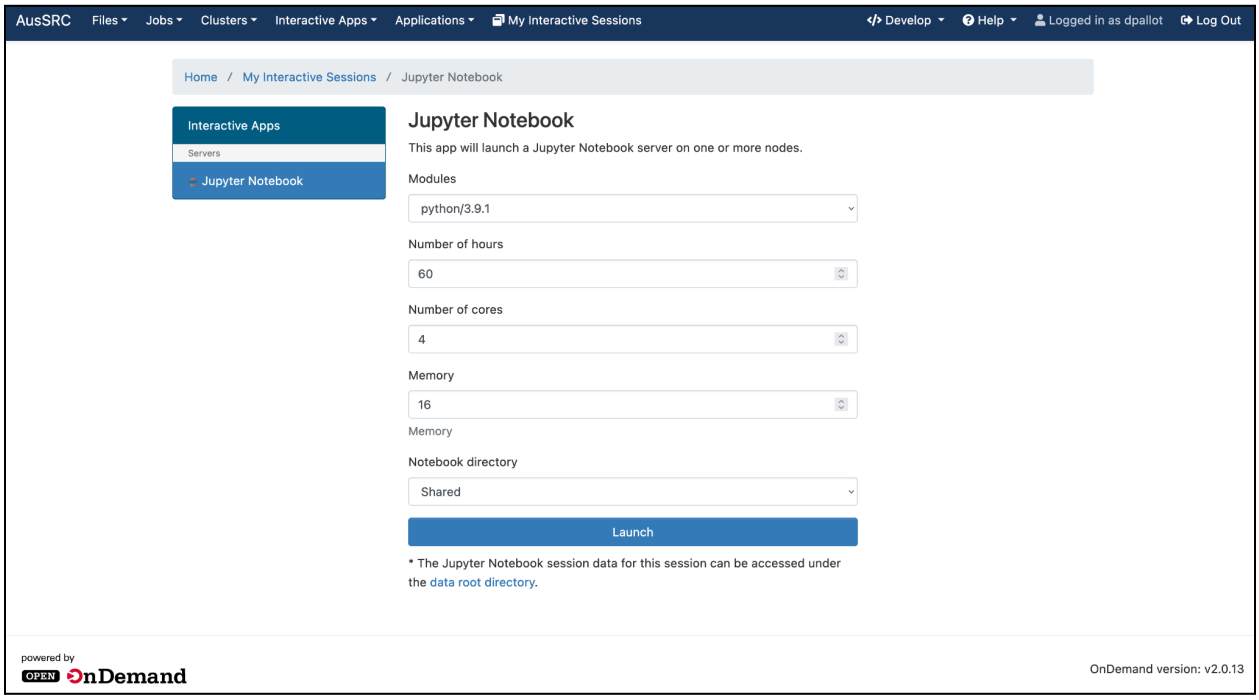

Figure A2.6.1: AusSRC Jupyter Notebook deployment portal

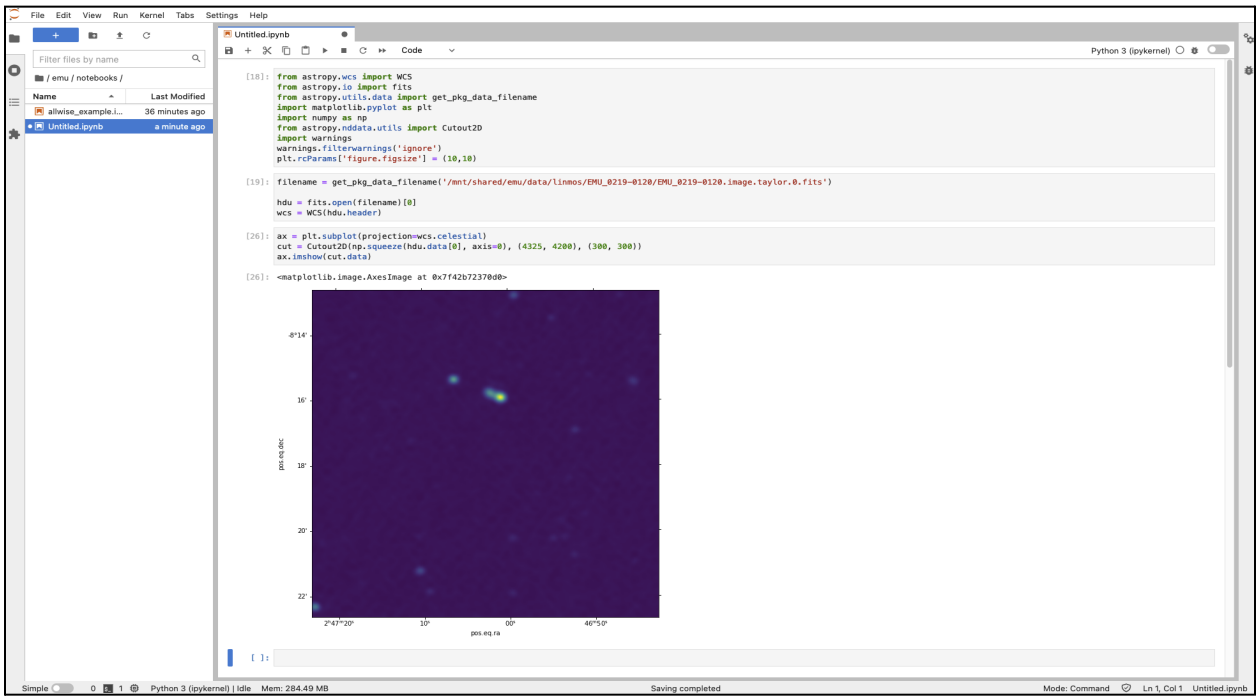

Figure A2.6.2: ASKAP image cutout using a notebook and Python code.

Objectstore Data Integration (and SoFiA-2)

AusSRC provides software components and workflows for the production of value-added data products. These data products are created by ingesting science-ready radio astronomy datasets

(provided by telescope operations) and then applying various post-processing algorithms within the workflows.

Traditionally, most of these (legacy) algorithms assume that their input data resides in a traditional Portable Operating System Interface (POSIX) filesystem, requiring the standard methods of file opening, reading, and writing to process the data. Typically, these algorithms will read in a few tens of Gb of data at a time, normally retrieved from a single radio datacube that may be several terabytes in size.

The radio datacubes are stored within archives hosted by supercomputing or cloud-computing facilities, which are starting to store these large amounts of data in object stores, rather than the POSIX shared file systems. This is the case for the CASDA, hosted at the Pawsey Supercomputing Research Centre, and will likely be the case for data from the upcoming SKAO archives. The Pawsey's Acacia is a 60PB high speed object storage system (for more information, refer to https://projects.pawsey.org.au).

As legacy processes cannot read data directly from an objectstore, the (rather large) datacube must be copied over from the objectstore to a POSIX shared file system prior to starting the pipeline. Additionally this filesystem must then allow efficient and concurrent input/output (I/O) to the same datacube (file) during processing. The pipeline processing quickly becomes an I/O-dominated activity. This bottleneck prompted the AusSRC to investigate and develop an example prototype solution to ingest data directly from the objectstore into the SoFiA-2 source finding workflow.

The SoFiA-2 codebase provides a source finding pipeline, originally designed to detect and characterise galaxies in 3D extragalactic HI datacubes in FITS format. It is written in the C programming language and is typically compiled as a stand-alone executable to run under Linux-based HPC job control systems, such as Slurm. SoFiA-2 however, is constrained by the available physical memory on the machine on which it is running.

Normally, SoFiA-2 assumes that all input data resides in an underlying POSIX filesystem and therefore, cannot access data directly from objectstore. Running a SoFiA-2 pipeline against data stored in an objectstore therefore requires several steps:

1. The entire datacube must be read from the objectstore and stored on a POSIX shared filesystem.
2. Multiple instances of SoFiA-2 are started up on a Slurm job queue (a single instance would take too long to read the entire datacube).
3. Each instance reads a "cutout" or "partition" of the data-cube from the shared file system. The size of this read is limited by the available random access memory (RAM) to the instance. Typically the limit is < 40 Gb (which will use approximately 90 Gb RAM to process).
4. Each instance writes its outputs back to the shared filesystem, and a separate process amalgamates all the required outputs together.

The actual computation time is generally in the order of 2 ~ 3 hours. However, the first step above can take 4 ~ 6 hours for a datacube of around 1 TB. The third step may take another 30 ~ 50 minutes. To try and reduce the amount of time required by I/O bound pipelines, the AusSRC team prototyped a “direct to memory” scheme. A set of python classes were developed that could directly read a specific cutout (as described in step 3 above) from the FITS object directly from the objectstore. This “sub-cube” data is stored in-memory as a flattened Numpy array. A SWIG (https://www.swig.org) “wrapper” class then passes the pointers to this array directly to a modified version of SoFiA-2, running as a shared dynamic library and reading data from memory, rather than from any underlying filesystem.

Overall, the SoFiA-2 codebase was modified as little as possible and the resultant code can be compiled as either a shared dynamic library or as the original standalone executable (both targets are defined in a new Makefile). The SWIG wrapper code handles the conversion between the array-handling within Python and within C. During the investigative phase, several Python3 libraries were tested for efficiency of data transfer from an objectstore:

- urllib (the most common but oldest and updated to urllib2 in Python 2.7)
- requests (the standard library for Python3)
- urllib3 (a third-party library for Python3)

Additionally, three different access patterns for extracting a subcube from a larger datacube were tested. The most efficient combination of library and access pattern was then coded into the SWIG python wrapper that directly injected the data into the memory space of the SoFiA-2 process. This approach shows significant reduction in the time required to ingest the data into the pipelines (Figure A2.7).

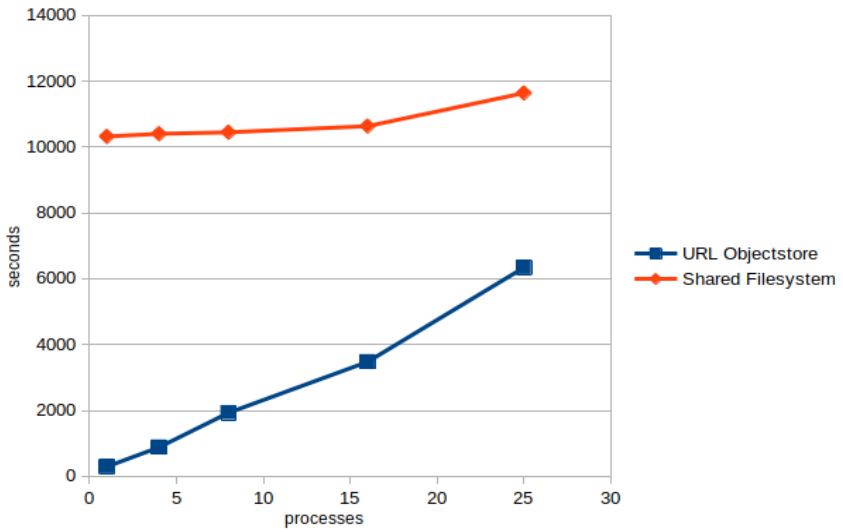

Figure A2.7: Shared POSIX file system ingest time versus direct Objectstore ingest.

Open OnDemand

Open OnDemand helps scientists to use remote HPC resources and applications by making them accessible via a web browser and other devices. Users are able to use these resources

faster and more efficiently than they could using traditional command line tools. Open OnDemand offers easy file management, command line shell access, Slurm job management and monitoring across resource managers, and graphical desktop environments and desktop applications. Popular scientific applications are easily accessible via Open OnDemand such as Jupyter Notebooks, MATLAB, Simulink, or Virtual Desktops.

Open OnDemand allows for applications to be developed and deployed within the context of the portal. Applications can be run in standalone web servers and have the capability to deploy Slurm compatible HPC jobs to the associated HPC platform. AusSRC is using the Open OnDemand platform as the main science portal for the precursor web based applications, as shown in Figures A2.8.1 and A2.8.2.

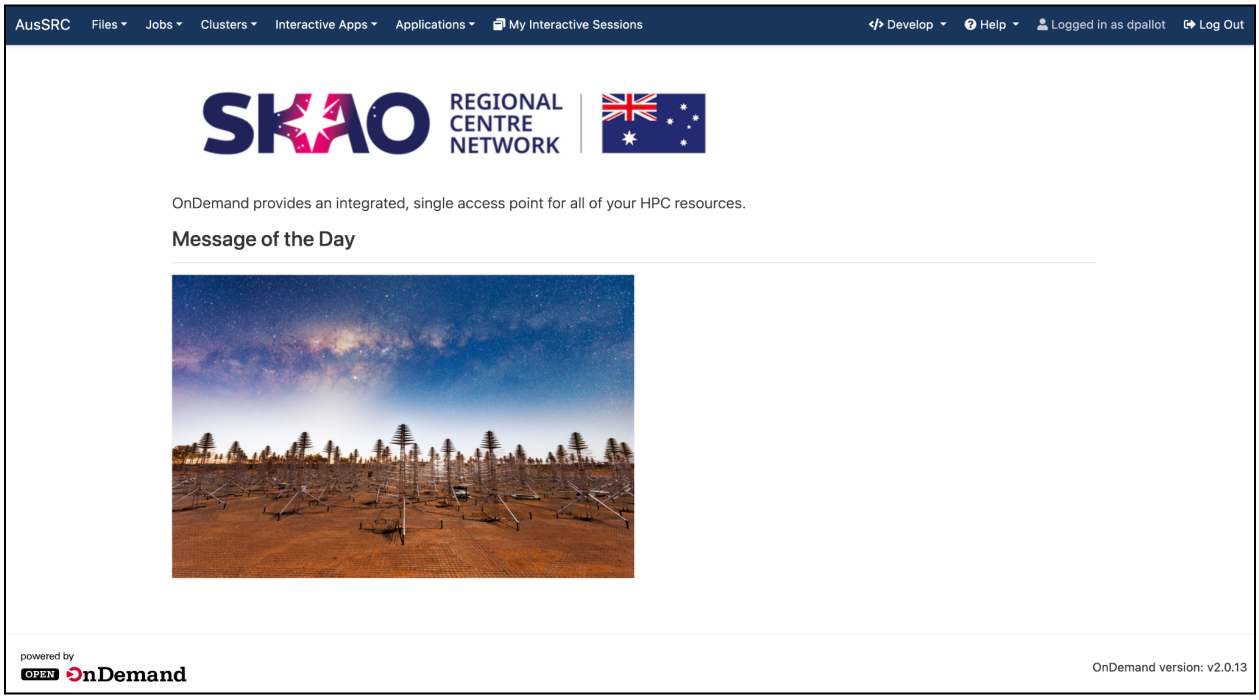

Figure A2.8.1: AusSRC Open OnDemand portal front page.

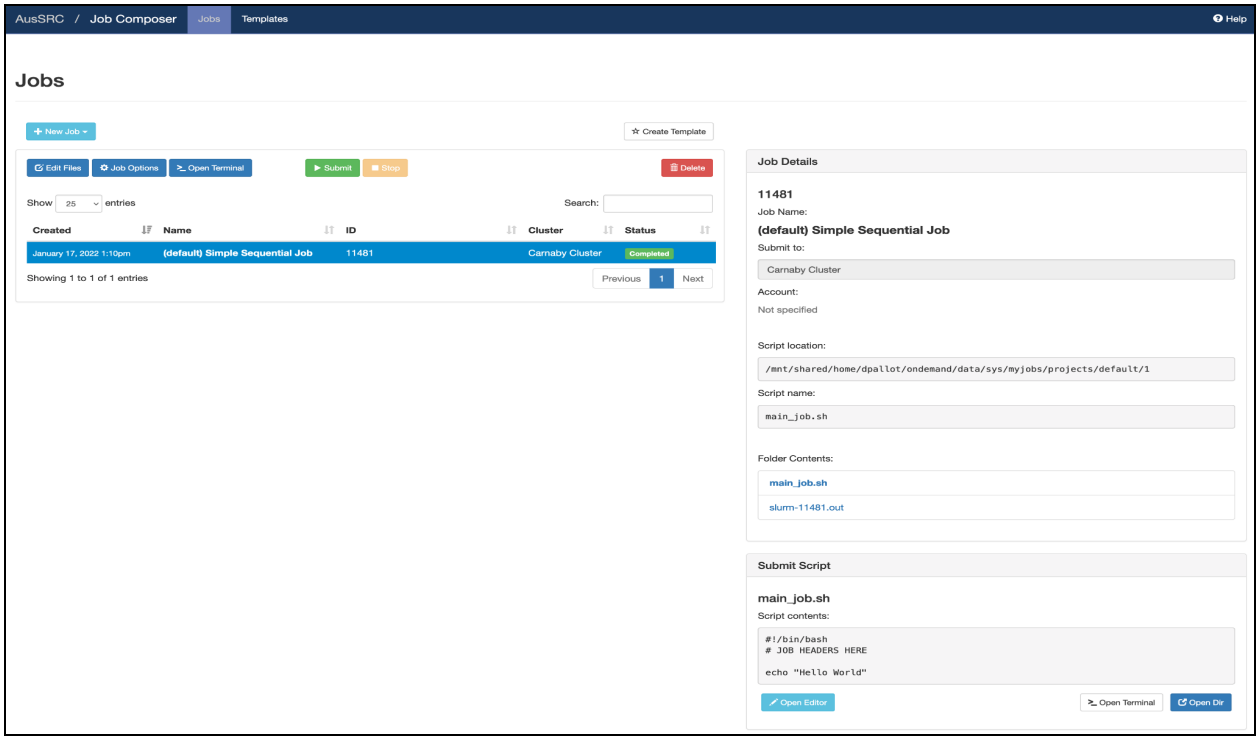

Figure A2.8.2: Open OnDemand Slurm job deployment portal.

OpenStack

OpenStack is an open standard cloud platform that is free. The platform is Infrastructure-as-a-Service (IaaS) that allows users to deploy Virtual Machines (VM) on demand. VM can handle tasks like web services, database engines, processing, storage, identity management and other software services. OpenStack gathers hardware, networking, and storage infrastructure together into resource pools and then allocates virtual resources when needed. OpenStack exposes APIs that allow resources and VM to be allocated programmatically that enhances the flexibility and adaptability of the infrastructure. The OpenStack components include:

- Nova - main computing engineer and acts as the central pool manager.
- Horizon - provides a web-based portal that allows users to administer their virtual cloud environment.
- Neutron - the networking element which allows communication between VMs and externally.
- Cinder - a block storage that provides ephemeral and long term storage to the VMs.
- Swift - object store that is S3 compatible.
- Trove - provide users with Database-as-a-Service (DBaaS) which allows relational databases to be deployed without having to worry about deployment, configuration, backups and monitoring.

Nextflow

Nextflow (https://www.nextflow.io/) is a tool that enables scalable, shareable, and reproducible scientific computational workflows using software containers. It allows the adaptation of pipelines written in the most common scripting languages such as Python and Java. Its fluent Domain Specific Language (DSL) simplifies the implementation and the deployment of complex parallel and reactive workflows on cloud infrastructure and HPC clusters. A Nextflow pipeline is implicitly modelled by a Directed Acyclic Graph (DAG).

A Nextflow script is made by joining different processes (e.g. component science algorithms) together. Each process can be written in any programming language and can be composed directly in the Nextflow script or built into a tagged Docker container, allowing for maximum code portability and reproducibility. Once the process component has been developed and pushed to a code repository (Github, GitLab, etc.), a Docker container is automatically built and pushed to Docker Hub (an online Docker container repository). When a Nextflow script is fully developed, it is deployed onto a Nextflow compatible HPC cluster where the Nextflow driver will pull the script and the relevant component containers and deploy each process as a job based on the DAG instructions encoded into the script. The Docker container can be automatically converted to a HPC compatible Singularity container, if necessary. Singularity aims to provide mobility of compute on HPC clusters and is necessary if the component code depends on the Message Passing Interface (MPI), which is true for many science based packages such as ASKAPSoft.

Nextflow workflows can be deployed on standard HPC, Google Cloud, Amazon Web Services (AWS), or Kubernetes platforms. A Nextflow configuration file can contain multiple platform targets or profiles allowing for maximum portability without changing the fundamental component code. Figure A2.9 outlines the Nextflow development, deployment, and execution process. The AusSRC has developed a series of guidelines for organising and deploying science based computational workflows and their constituent components that will allow for maximum portability between resource providers.

The AusSRC uses Github and Docker Hub for its code and container repositories, respectively. For each AusSRC scientific project, there are typically two repositories, one for the Nextflow workflow scripts and one for the component code (science algorithms). This separation allows components to be developed and evolve independently of the greater workflow logic. It also allows the Nextflow Execution Engine to retrieve the workflow scripts directly from the Github repository without the accompanying component code. If necessary, AusSRC separates repositories for all service code related to databases, web services, and deployment scripts.

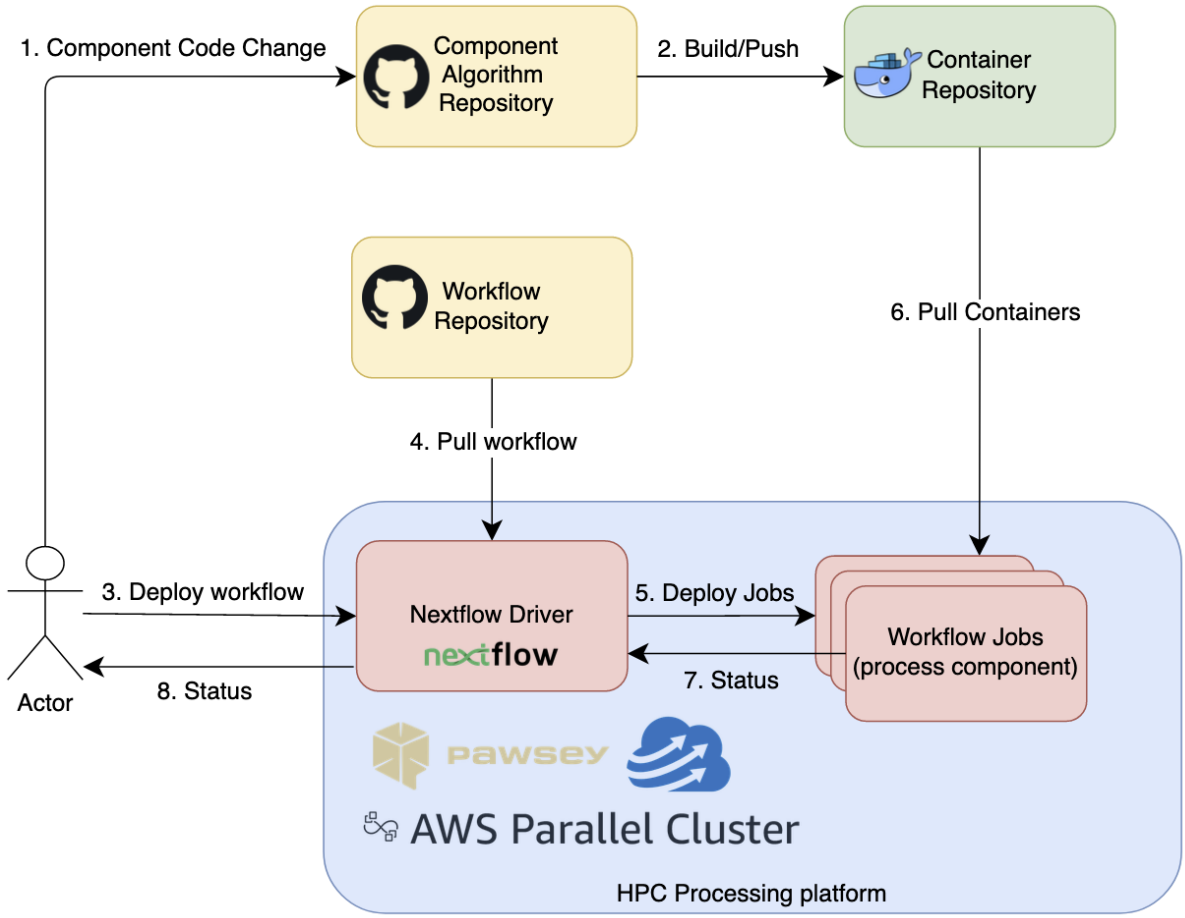

Figure A2.9: Nextflow development, deployment and execution process.

Example repositories and structure:

- *wallaby_services* - databases, VO service deployment scripts etc.
- *wallaby_workflows* - Nextflow workflow scripts.
- *wallaby_components* - science algorithms.
    1. component_1
        - component1_code_directory
        - Dockerfile
    2. component_2
        - component2_code_directory
        - Dockerfile
    3. component_n (not owned by project)
        - Dockerfile

Each workflow component is composed in a Docker container, where appropriate. More than one component can be built into a single container. If the component code is externally owned and a container does not already exist, then all that is required is a Dockerfile to build the component.

Once development of a component is complete, it is built into a Docker container and pushed to Docker Hub. This can be achieved by developing a Github Action so component containers are built and deployed on Docker Hub on a git push, a tagged release, timer (nightly build), or manually. During active development, however, it is expected that the component code is run outside of the context of a container. Integrated development environments such as Microsoft Visual Studio can be configured to directly push the code (workflow and components) to a shared space on a processing cluster so it can be tested and iterated on efficiently.

Nextflow Deployment Platform

The AusSRC has developed a Nextflow Deployment Platform that allows users to deploy Nextflow workflows directly onto a processing platform either manually through a web portal, programmatically through an API, or automatically through an asynchronous event trigger. Nextflow scripts can be scheduled onto a processing platform via an asynchronous event trigger when a subset of observational data meets a particular prerequisite based on a project's requirements. For example, the EMU Value-added Catalogue (EMUCat) workflow combines multiple adjacent tiles (observation), which are independently scheduled/observed, into a single large region. This workflow is triggered when all the prerequisite observations for a specific region have been observed and deposited into the CASDA archive. AusSRC uses RabbitMQ (https://www.rabbitmq.com) as its message broker.

The asynchronous workflow event triggering platform consists of 4 main components, a CASDA/SDP process, science prerequisite processes, a Nextflow workflow scheduler process, and a processing cluster (Figure A2.10).

1. The CASDA/SDP process is a daemon process (i.e. background process that runs unattended) that polls or receives an asynchronous event from CASDA/SDP when a valid observation data product is deposited into the archive. This process emits an event if it is a new observation, which is then captured by the science prerequisite process(es)
2. The science prerequisite daemon process(es) can be one or more daemon processes that consume observation events emitted by the CASDA/SDP process. This process checks if all the observation prerequisites for a particular workflow have been met. If the prerequisites have not been met, the process waits for the required observations to be deposited into CASDA/SDP. If the prerequisites have been met, this process emits an event with the details regarding which predefined Nextflow script to run. The Nextflow script is specified as a preconfigured GitHub repository along with any predefined or dynamic parameters.
3. The Nextflow workflow scheduler is a daemon process that consumes the events emitted from the prerequisite process(es) and schedules workflow jobs onto a remote processing platform. A workflow job is a preconfigured configuration that contains deployments (processing platform host) and pipelines (the Nextflow repository to be run). The Nextflow workflow schedule process has a web interface that allows users to keep track of scheduled jobs and their state (complete, cancelled, failed, etc.).
4. The Nextflow workflow scheduler process monitors the job state and emits an event, which contains its state changes. This event can be consumed by any other system or

user defined processes that may be interested in the state. As a result, this platform could be used as a distributed workflow scheduler across multiple SRCs.

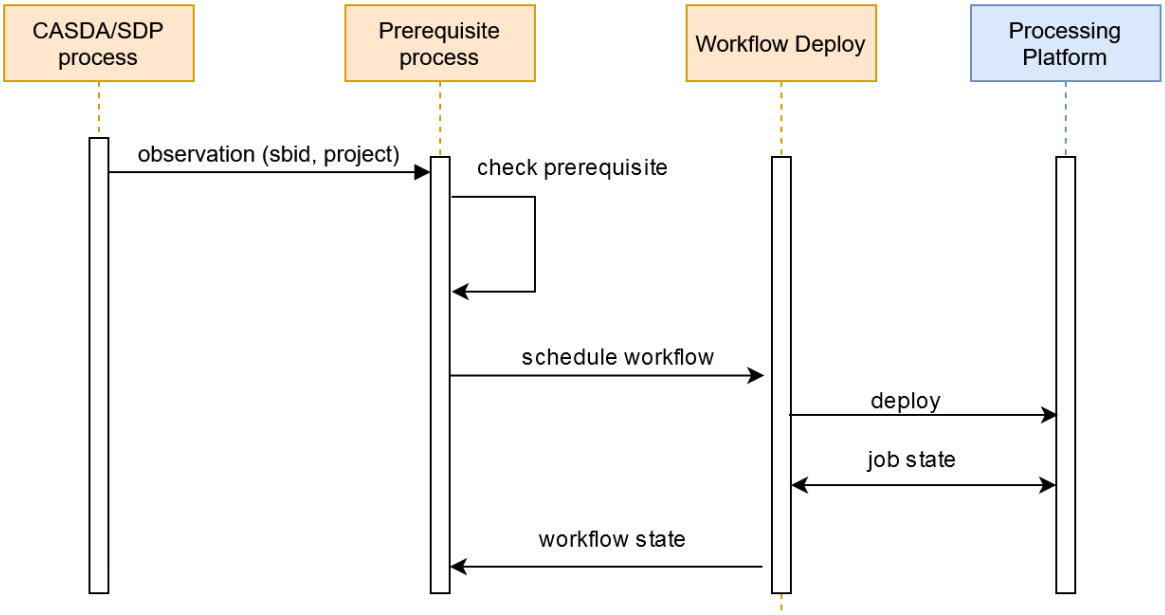

Figure A2.10: Event flow for deploying a workflow based on prerequisites.

VO Services

VO provides scientists a standardised set of tools and archives that use the internet to form a research environment for multi-wavelength astronomy analysis (for examples, see Figures A2.11 and A1.12). AusSRC has deployed Table Access Protocol (TAP) services for the various precursor projects that define a protocol for accessing astronomical catalogue data stored in databases. The TAP supports the Astronomical Data Query Language (ADQL) for catalogue searches and table uploads required for cross-matching with a diverse range of multi-wavelength data sets. AusSRC can also deploy Simple Image Access Protocol (SIAP) services that provide the capability to discover and retrieve image data, Simple Spectral Access Protocol (SSAP) services to access spectral data, and Server-side Operations for Data Access (SODA) services for server-side data processing.

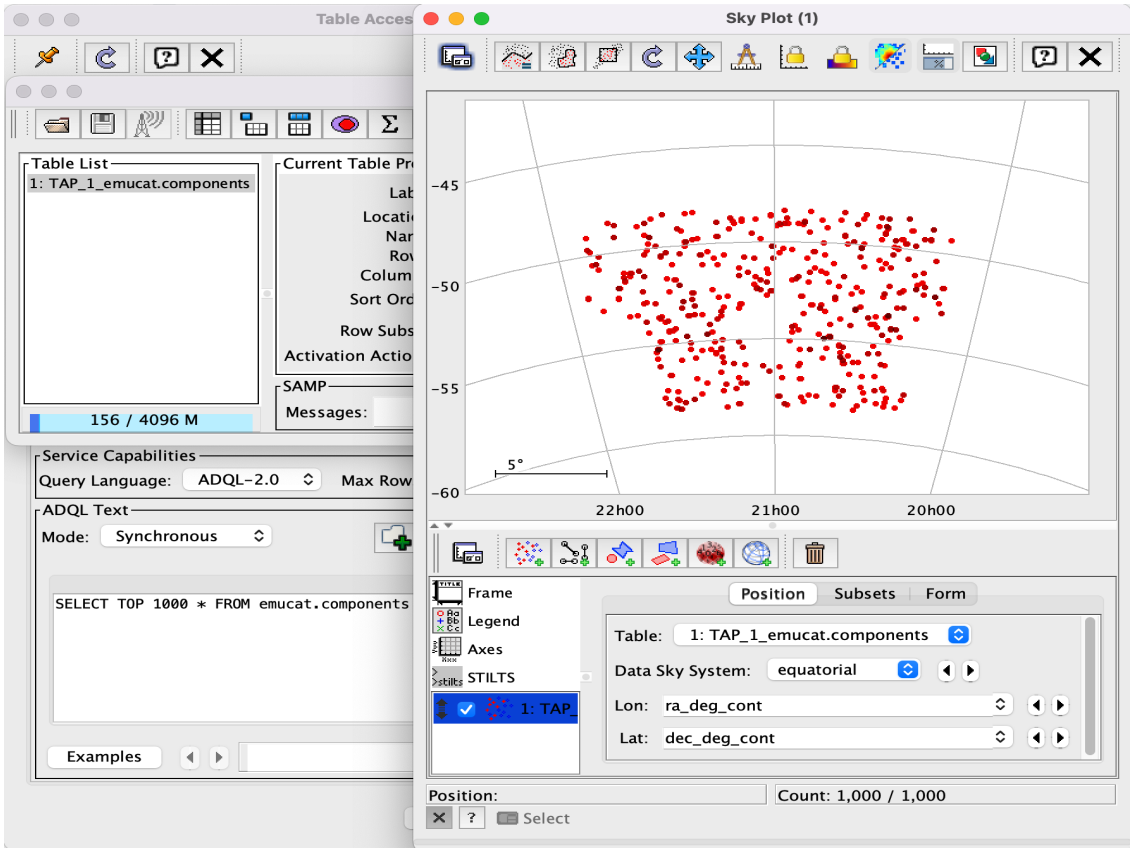

Figure A2.11: Plotting position of EMUCat components via Topcat (http://www.star.bris.ac.uk/~mbt/topcat/) and TAP.

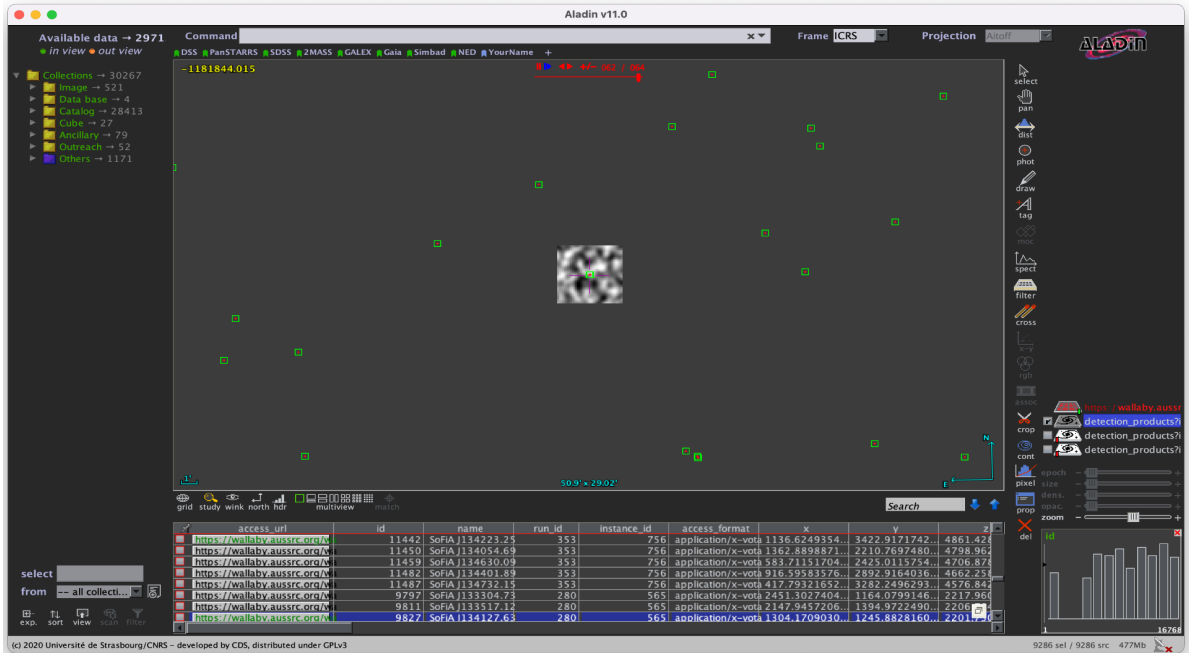

Figure A2.12: Displaying WALLABY image data using Aladin (https://aladin.u-strasbg.fr) and TAP.

Resources/service providers utilised

The AusSRC uses a OpenStack (Nimbus) allocation for provisioning the virtual cloud environment that handles the Open OnDemand service, the precursor catalogue database services, various VO services, authentication services, CARTA visualisation platform (https://cartavis.org), a shared file system, and a Slurm development processing cluster. The AusSRC OpenStack Nimbus allocation pool is:

- 112 VM Instances
- 2040 Virtual CPUs
- 8 TB RAM
- 500 TB storage

The AusSRC deploys precursor science processing workflows primarily on the Pawsey Supercomputing Research Centre's Magnus/Setonix and Garrawarla HPC systems. There is also some usage of the Pawsey's Topaz and Zeus HPC systems. Additionally, the MWA EoR project has deployed processing workflows at DUG Technology, a commercial research and development HPC platform.

External collaborations

- AAO & Data Central
- DUG Technology

Applicability to SKAO

The AusSRC DSP design is a high level blueprint for the greater SRCNet design and was formed by the experience gained while working with various SKA precursor science projects. Many future SKA science projects can be viewed as extensions of the precursor projects that differ in terms of scalability on the processing and storage side. It was therefore important for the design, especially the science workflow frameworks, to be extensible, portable, and platform independent so it can evolve as requirements and conditions change. This ensures the workflows, their components, and the associated services can evolve over time, independent of the hardware and software platforms with which the SRCs need to interoperate. These aspects provide SRC partners with the maximum amount of flexibility that is required for the SKA phased rollout and the ever evolving technology roadmaps and associated costs.

Possible future developments

Data Central is developing a VO compliant TAP service with table uploads making large cross-matching tasks more reliable and efficient for Australian based users and automated science workflows for the SKA. Integrating such services within the AusSRC system should be investigated.

Blue-Lavender prototyping

Project overview

The AusSRC has formed an SRCSC prototyping team named Blue-Lavender, which is a coalition of SRC's from Australia, China, Japan, and South Korea. The team is responsible for contributing to five prototyping themes for the global SRCNet Agile Release Train (ART). The five themes are:

- Data Management
- Authorisation and Authentication
- Science Platform
- Visualisation
- Distribution of Software Tools and Services

The SRCNet ART operates on a 13 week Program Increment timeframe, within which each team selects work from the program backlog and coordinates with other teams to produce the necessary prototyping work.

Science requirements/requested development

The science requirements are developed through collaboration between various SRCSC Working Groups, which provide inputs into the SRCNet ART. It is the responsibility of Working Group 6 (Science User Engagement) to engage with the radio astronomy science community and generate their requirements for the SRCNet ART. This iterative process will maximise the science return and guide the astronomy community towards new end-to-end procedures that will be required to produce science during the SKA era. The work AusSRC developed during the DSP is currently being fed into the SRCNet ART, to shape the international design, which includes the data processing experience in support of the five prototypes.

Technology/applications involved and/or tested

- Deployment of a working Rucio Storage node at JapanSRC, ChinaSRC, and AusSRC platforms in support of Data Management prototype.
- AusSRC IAM integration with SRC IAM in support of the Authorisation and Authentication prototyping important for global IAM compatibility.
- Deployment of Visualisation platforms such VisIVO, CARTA and Aladin at ChinaSRC in support of the Visualisation prototype.

Resources/service providers utilised

- ChinaSRC processing and visualisation cluster
- JapanSRC cloud Infrastructure
- Pawsey Supercomputing Research Centre: Nimbus
- SKA Confluence, JIRA, and Miro for team planning and documentation

External collaborations

- China
- Japan
- South Korea

Applicability to SKA

Blue-Lavender are direct contributors to the global SRCNet effort. The team and its members have extensive experience in the precursor projects and are therefore perfectly placed in helping shape the requirements and the prototype of the overall SRCNet design.

Possible future developments

The SRCNet ART and Blue-Lavender will be aligning with the greater SKA ART in December 2022. Blue-Lavender will co-plan and collaborate with ICRAR's Team Yanda. Team Yanda defines the SDP science data formats in which the SRCNet will be interfacing. This alignment will allow for closer collaboration and communication enhancing the quality of the SRC prototype overall.

A2. Technical Reports - DSP Science Projects

EMU project

Project overview

EMU uses ASKAP to catalogue and make a consensus of radio sources in the southern sky. EMU is expected to detect more than 70 million sources compared to 2.5 million currently catalogued sources in the same region. EMU's main science goal is to understand how galaxies and stars formed and how they evolved to their current state.

EMU will survey the entire southern sky visible to the ASKAP telescope in 30 square degree fields. Each field will be surveyed over an instantaneous 300 MHz band, from about 1110 to 1410 MHz. Radio components are extracted from the ASKAP image data and then cross-matched with multi-wavelength data catalogues. The cross-matched components are added to EMUCat, which is made available to the EMU team and later the public.

EMU will be the touchstone for ~1 GHz radio continuum data for at least a decade after survey completion, and possibly beyond, having the best combination of sensitivity, resolution, and survey area. EMU's surface-brightness sensitivity is considerably better than any existing or planned survey of this scale, at any frequency. With EMU near the confusion brightness limit no future survey can significantly surpass its brightness sensitivity.

Science requirements/requested development

The goal for EMUCat is to provide an all-sky radio catalogue with maximum sensitivity, avoiding duplication, while grouping radio components (which are single Gaussians) into sources, identifying hosts, and compiling properties and tags. This process is colloquially known as data combination and value-add. The assembly of the catalogue consists of a series of steps outlined in Figure A3.1.

Technical developments during DSP

All-sky Catalogue Assembly

EMU's all-sky tiling strategy places adjacent ASKAP fields close enough so as to obtain full sensitivity in the overlapping regions. Therefore, mosaicking of adjacent fields is essential to meeting the survey's sensitivity requirements over the full sky. The goal is to gather a set of adjacent fields from CASDA, merge them together into a larger mosaic or region, run source-finding software on the larger mosaic, and then integrate the results into an all-sky catalogue with no duplication of sources.

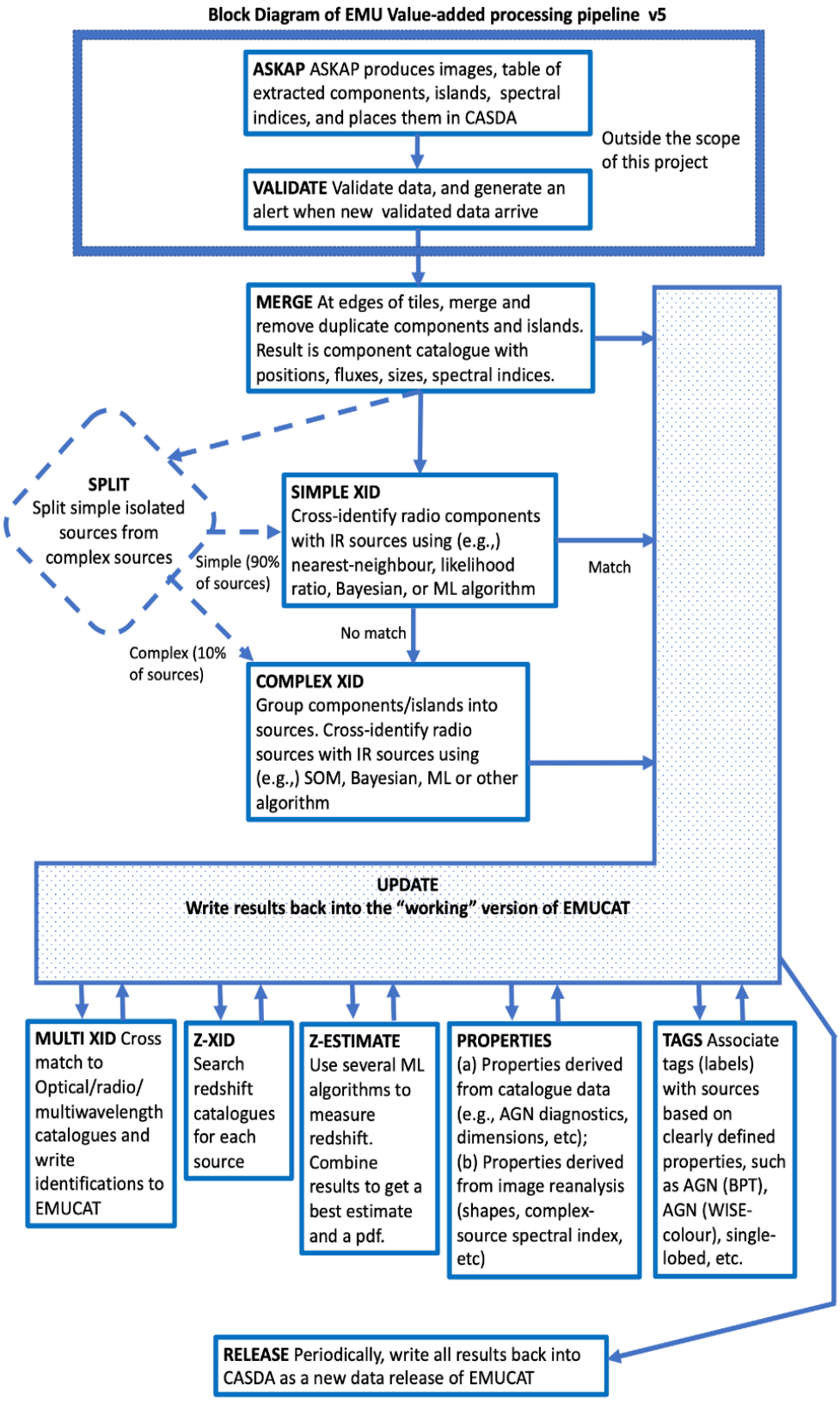

Figure A3.1: EMUCat processing workflow.

The size of the regions are determined based on the performance of the source extraction software, *Selavy*, which is estimated at about 100 square degrees to maintain efficiency. Prerequisites are the set of all ASKAP observations that include any portion of the region (and possibly some additional buffer). Once the processed observations for a full region are completed and uploaded to CASDA, the merge workflow can be executed. The merge workflow consists of the following steps:

1. All prerequisite ASKAP field images downloaded from CASDA
2. Combine tiles into mosaic region using linmos
3. *Selavy* run on mosaic
4. *Selavy* components and islands added to EMUCat database

The merge workflow has been developed to execute once per region, however multiple merge workflows can run in parallel on separate regions. Figures A3.2 to A3.7 illustrate an example of the tiling, region and merge strategy as it applies to the EMU pilot survey data set and is the general strategy for the entire survey. The full EMU survey will eventually cover ~30,000 square degrees, so it is expected that the workflow (e.g. gathering adjacent fields, mosaicking, and running *Selavy*) will need to be run ~300 times over the duration of the full 5-year survey).

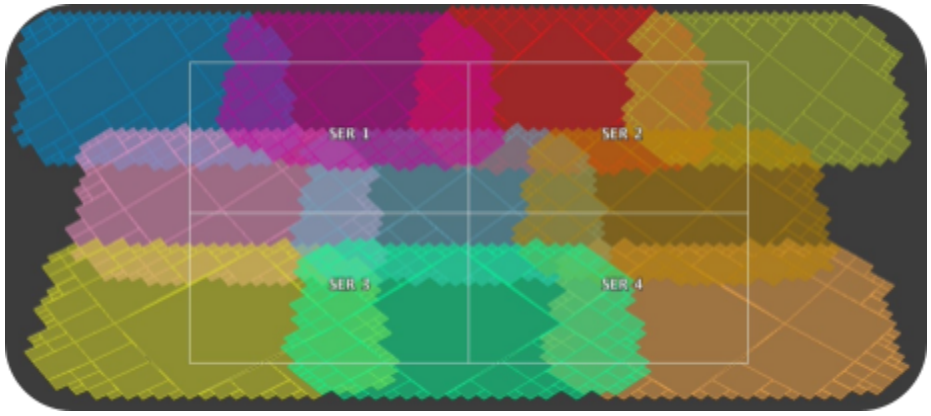

Figure A3.2: Regions 1 to 4 and their 10 overlapping EMU ASKAP fields.

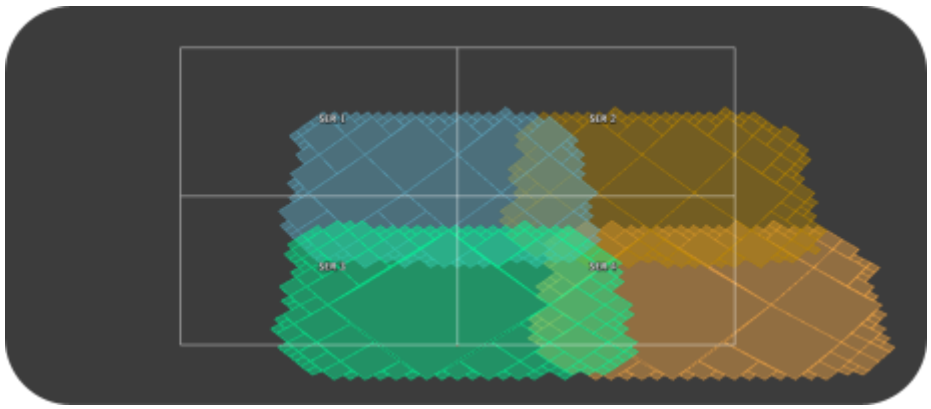

Figure A3.3: A complete region, Region 4, consisting of 4 overlapping ASKAP fields. The prerequisites are met so *merge* can be run.

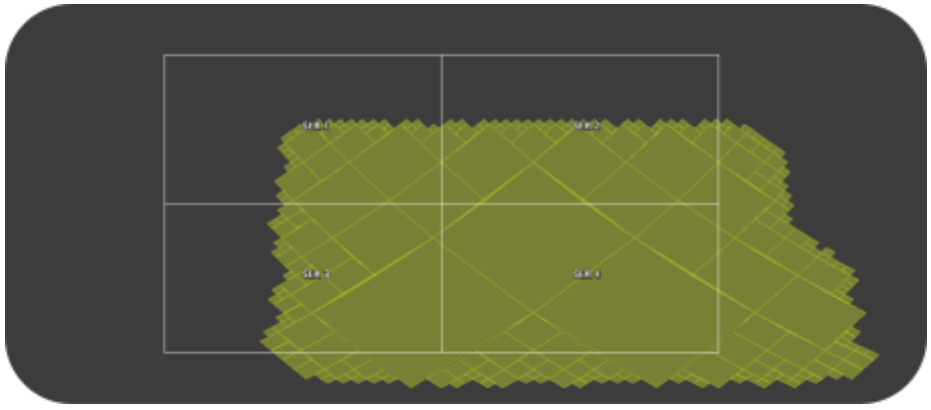

Figure A3.4: A mosaic from their prerequisite fields are created using linmos.

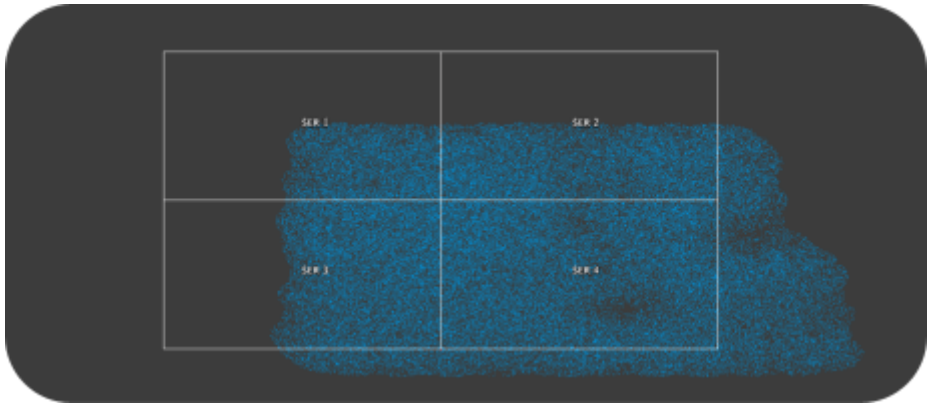

Figure A3.4: *Selvay* is run on the mosaic to extract the radio components.

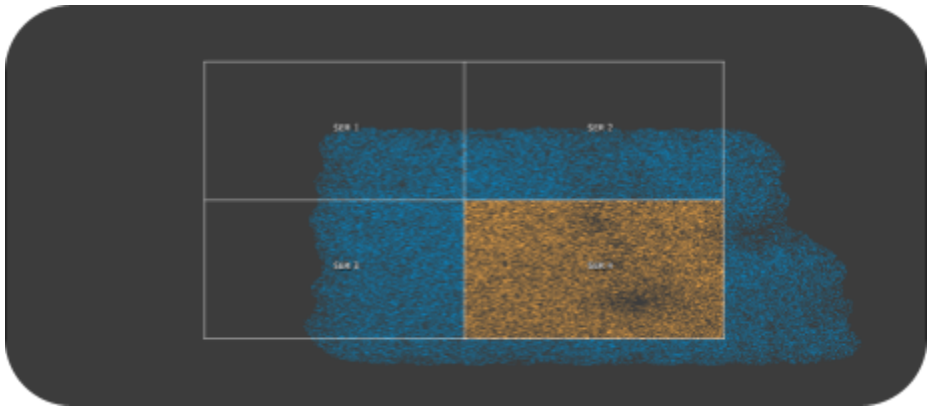

Figure A3.5: Radio components are selected within the boundary region.

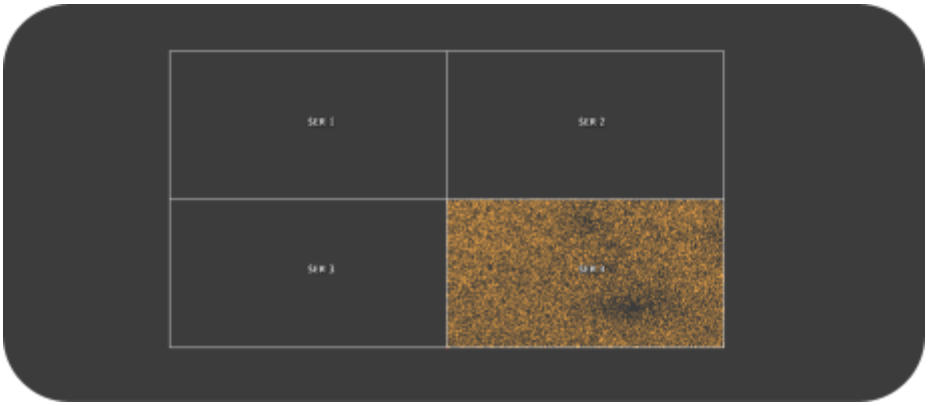

Figure A3.6: Those components that fall outside of the region are virtually discarded.

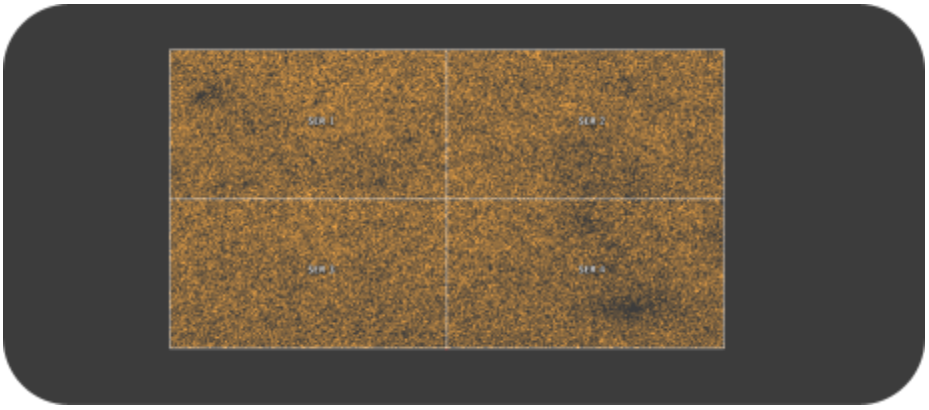

Figure A3.7: Process is repeated until all 4 regions are complete.

ASKAP Data Products

Individual EMU ASKAP observations consist of a single field with 36 simultaneous beams. The beams are arranged in a 6 x 6 square grid. The beam-to-beam spacing (pitch angle) is 0.9 degrees so each field covers 4.5 x 4.5 degrees (~20 square degrees). All 36 beams of each field are combined using ASKAPsoft's linmos utility and these mosaics are delivered to CASDA as a 'level 5' data product ('level 6' after validation). These images all share the same shape (approximately 15k x 15k pixels), are produced using the World Coordinate System (WCS), and each image has a data volume slightly less than 1 GB.

All-sky EMU Catalogue

The final step of the EMU workflow integrates the source-finding from each region into a cumulative all-sky catalogue. The approach uses predefined source extraction boundaries that divide the entire survey into patches of approximately 100 square degrees. All sources having positions within this boundary are added to the cumulative source catalogue. This means that the final step of adding to the cumulative catalogue actually belongs at the end of the EMUCat pipeline, after multi-wavelength cross-matching and the grouping of components into sources has taken place.

Value-Added Workflow

The EMU project groups radio components, which are single Gaussians, into sources using the infrared AllWISE catalogue. A source can be either a single component (Simple XID) or a group of components (Complex XID). The two Simple XID matching algorithms being used are nearest neighbour and likelihood ratio. The Complex XID algorithms being used are the extended doubles and *Selavy* island algorithms. Matches from all algorithms are inserted into the EMUCat database separated by algorithm type. Hostless components are also recorded in the catalogue.

EMUCat Database

The EMUCat database is a PostgreSQL instance, with pgSphere enhancements, that runs on the Pawsey Nimbus Cloud Computing platform. The database schema is subdivided into 4 main categories that reflect the stages of the EMU workflow, these are: Inputs tables, Sources tables, Property and Tag tables (Figure A3.8). Inputs are the radio components that are transformed into sources when they are cross-matched with a specific algorithm and a particular multi-wavelength catalogue. These radio sources are inserted into their respective algorithmic source tables when they are matched e.g. the table *source_lhr_allwise*, which contains all the radio sources that have been matched with AllWISE using the likelihood ratio algorithm. Tables are added to the database when new algorithms are developed or when new multi-wavelength cross-match catalogues are identified. Properties and Tags are metadata elements that describe and categorise the radio sources for each multi-wavelength cross-match and algorithm table pair.

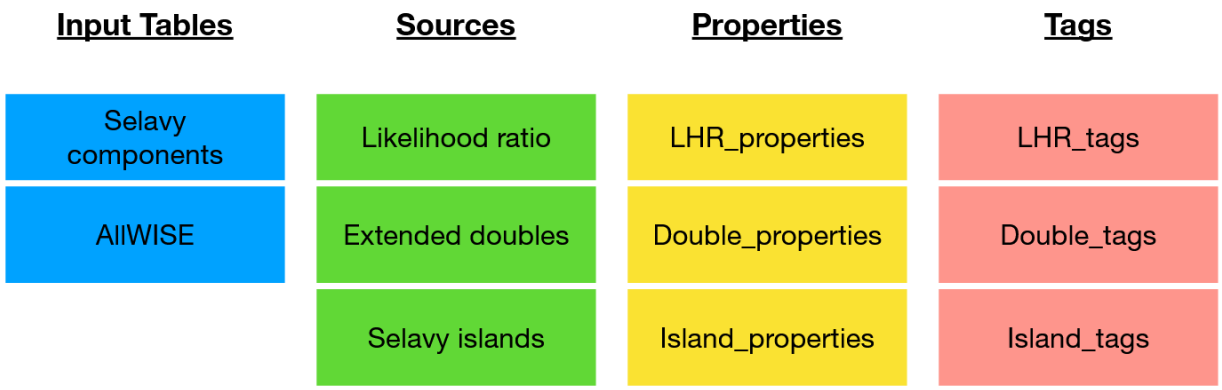

Figure A3.8: The database schema is subdivided into 4 main categories reflecting the stages of the EMU workflow.

Simple XID

The nearest neighbour algorithm cross-matches sources within a 4 arcsec radius and selects the closest potential host based on distance (Figure A3.9). The likelihood ratio algorithm determines matches by distance and AllWISE magnitude. Most sources have at least one source with reliability, $R > 0.8$, but several have first and second highest values of $0.2 < R < 0.8$, these were found to be doubles. Matches with a separation > 4 arcsec are still recorded but the component source is considered unmatched due to high false-ID rate.

For every likelihood match the workflow adds redshift photometry data from the Dark Energy Survey data release 1, 2 and the Dark Energy Spectroscopic Instrument (DESI) Legacy Survey data release 9.

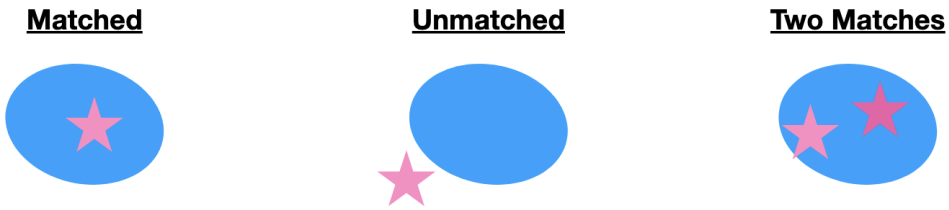

Figure A3.9: illustration of different cross-matching scenarios.

Complex XID

The extended doubles algorithm identifies pairs of extended, unmatched radio components within a maximum separation is 100 arcsec and a minimum deconvolved size 15 arcsec. Pairs that have been identified are recorded in an extended source database table. Future work will involve identifying an extended host using AllWISE (Figure A3.10).

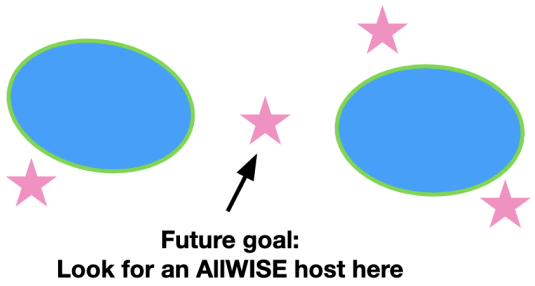

Figure A3.10: illustration of a double radio source (blue) around a likely central host.

The Selavy island algorithm combines all radio components within multi-component islands (flood fill) and an EMU region that provides a unique way to retrieve the associated components (Figure A3.11).

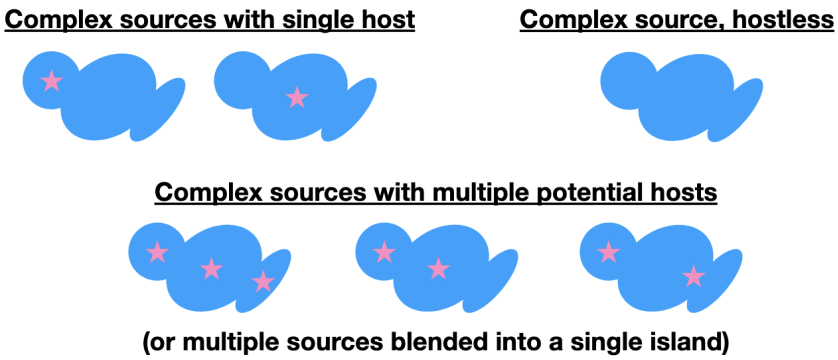

Figure A3.11: illustration of complex sources.

Radio Properties and Tags

Radio properties are a common set of numerical quantities and their uncertainties – such as position, peak and integrated flux, spectral index, deconvolved size, position angle, peak-to-peak separation, and distance to nearest radio source – which are stored in the EMUCat database for each source algorithm such likelihood or nearest neighbour. Radio Tags are a common set of textual labels for a radio source compiled for each source algorithm. For example, given a hypothetical source shown in Figure A3.12, the selected radio tags (in blue) could be:

- single/**double**/triple
- compact/**extended**
- flat spectrum/**steep spectrum**
- **aligned**/misaligned position angles
- **symmetric**/asymmetric flux ratio
- symmetric/**asymmetric shape**

Figure A3.12: hypothetical source.

Technology/applications involved and/or tested

- Docker Hub
- GitHub (https://github.com/ASKAP-EMUCat)
- Jupyter Notebooks
- Nextflow
- PostgreSQL
- VO Services

Resources/service providers utilised

- CASDA
- Pawsey Supercomputing Research Centre: Nimbus, Magnus/Setonix

External collaborations

- AAO & Data Central
- National Optical Astronomy Observatory (NOAO)
- National Radio Astronomy Observatory (NRAO)

Applicability to SKA

EMU will provide the most uniform deep wide-area radio data to inform SKA survey region selection, as well as to complement analyses from recent and upcoming surveys and facilities

such as eROSITA, DESI, GLEAM-X, LoTSS, PanSTARRS, Apertif, 4HS, WAVES, the Rubin Observatory's LSST, Euclid, and more. EMU will serve as a critical and key data resource in machine learning algorithm development for radio astronomy.

The AusSRC and EMUCat teams expect that as EMU progresses it will be used in developing and testing a comprehensive and exemplar AI+radio astronomy application workflow. Preliminary tools for such work are already being explored. The machine learning applications for EMU range from source finding, host or multi-wavelength cross identification, identification of multi-component or giant radio galaxies, image denoising, and more. The algorithms will be applicable to many future datasets, especially those delivered by the SKA. In addition to machine learning, EMU will provide new software developments as well as data storage and computing technology opportunities and advances. The legacy of software and systems developed for the specific needs of EMU is fundamental for the SKA.

Possible future developments

The EMUCat project will continue to identify value-added data products from external multi-wavelength catalogs to enhance the quality of the survey. This includes:

- Radio XID - cross-matching with other radio catalogues such as MWA GLEAM, NVSS, SUMSS catalogues.
- Z-XID - the optical/infrared positions of the radio sources will be used to search spectroscopic and high-quality photo-z catalogues for the redshifts of sources.
- Z-Estimate - for sources that don't have a redshift from Z-XID, machine-learning techniques will be used to estimate photo-z, possibly by classifying them in a small number of redshift bins.

The EMU and AusSRC team will work with the CASDA team to provide an API that will allow the merged mosaics to be ingested as level 7 data products. The EMUCat survey and mosaics will be made available via Aladin Sky Atlas, a multi-wavelength visualisation tool compatible with VO services. Data Central will make the EMU pilot survey phases 1 and 2 mosaic images available for the Data Aggregation Service (https://das.datacentral.org.au/das). This platform provides web-based tools and archive functionality for scientists from a range of disciplines to explore, collaborate, and make new discoveries. The complete EMU survey will be available to astronomers via this VO compliant platform.

FLASH project

Project overview

FLASH is a wide-field survey that analyses distant radio continuum sources to identify intervening foreground hydrogen clouds, the latter of which absorbs the emission of the background source. FLASH aims to target the poorly explored redshift range $0.4 < z < 1.0$, equating to a lookback time of 4 – 8 Gyr. It will cover approximately 34,000 square degrees in 903 pointings of 2 hours per pointing.

FLASH will survey the southern sky at frequencies between 711.5 and 999.5 MHz using ASKAP, and is expected to find several thousand of both intervening and associated HI absorbers. Data from identified intervening absorbers will produce an HI-absorption selected catalogue of galaxies rich in cool, star-forming gas, some of which may be concealed from optical surveys. Similarly, data from associated 21 cm absorbers are expected to provide valuable kinematical information for models of gas accretion and jet-driven feedback in radio-loud active galactic nuclei. FLASH will also detect hydroxyl (OH) 18 cm absorbers in diffuse molecular gas, megamaser OH emission, radio recombination lines, and stacked HI emission.

Science requirements/requested development

The FLASH project requires the development of a data pipeline to enable the existing FLASH post-processing steps (many of which are run manually) to be executed efficiently and reliably. The pipeline will provide an interface with CASDA and reduce the current levels of data handling and format translations.

Support for a robust multi-wavelength data base of HI absorbers is also needed. This resource will naturally expand as the project moves towards the absorption surveys with the SKA that span contiguous redshifts from the nearby Universe to the epoch of reionisation.

Technical developments during DSP

The initial step of the FLASH workflow requires the spectral plots of individual sources of interest to be extracted from the CASDA-stored datacubes (see Figure A3.13). The scripts used by the FLASH science team to accomplish this were optimised by AusSRC to run in parallel on HPC resources and to automatically interface with Pawsey's Acacia storage system. This produced a speedup of 3 - 5 times in processing, depending on the availability of resources.

The next step in the workflow uses the FLASH Linefinder code to identify absorption within the generated spectral plots. The Linefinder code (based on a Fortran implementation of MultiNEST; https://cosmosis.readthedocs.io/en/latest/reference/samplers/multinest.html) was compiled and containerised for use in a Slurm environment on both Nimbus and Setonix. Previous pilot survey phase 1 Linefinder results were archived on Acacia.

The AusSRC also provided a secure platform for the CHAD database and frontend, where cross-matchings could be performed, search queries executed, and results displayed. This was tied to a 10 TB drive and ran on AusSRC's Nimbus cloud allocation.

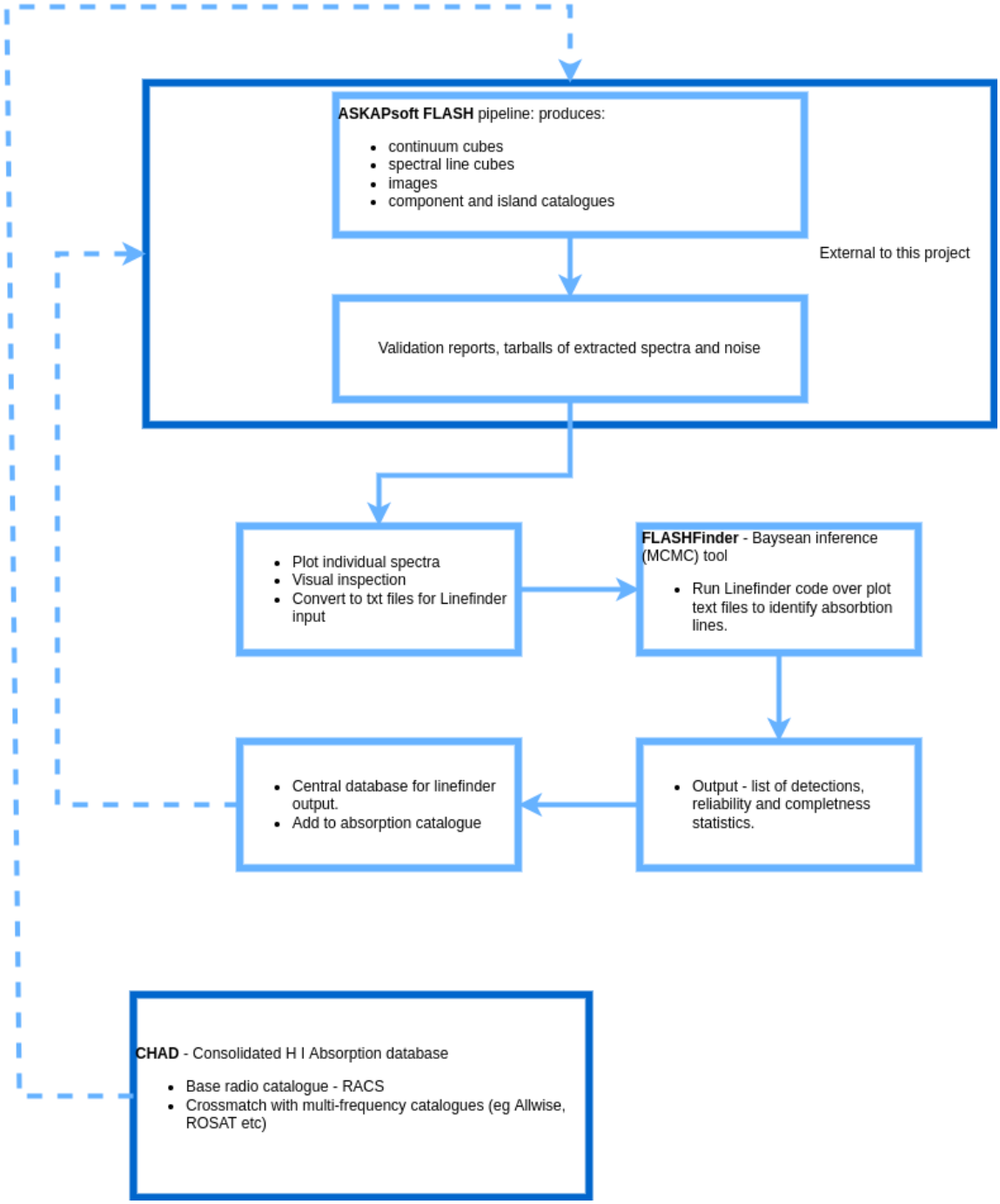

Figure A3.13: the expected workflow for the FLASH science team.

Technology/applications involved and/or tested

- Certification and HTTPS support
- Django/FLASK python web development
- Docker/Docker Hub
- Objectstore/S3 python libraries (Boto3)
- PostgreSQL

- Slurm
- uWSGI web server

Resources/service providers utilised

- CASDA
- Cloudflare
- Pawsey Supercomputing Research Centre: Nimbus, Magnus/Setonix, Acacia

Applicability to SKAO

The FLASH workflow is unlike the other science projects supported by AusSRC (e.g. EMU and WALLABY). During the DSP, the FLASH team has provided valuable additional data-points and insights into user workflow patterns and how they can be modified for use in HPC and automated pipelines. This will be directly applicable to the development of value-add pipelines for the SKAO datasets.

Possible future developments

AusSRC will continue to develop the FLASH absorption-detection pipeline, which will remove several of the manual processing steps that are unlikely able to handle the scale of full survey-size datasets. This will also involve output processing and production of value-add products such as catalogues to store back to CASDA.

MWA EoR project

Project overview

EoR is a precision experiment that requires customised processing and careful consideration of all treatments of the data. The MWA EoR processing pipeline takes raw data from the MWA and the basic metadata provided by the Monitor and Control system. This data is then pre-processed and calibrated for power spectrum analysis. The analysis is sensitive to very faint systematics, which means any data that contains radio frequency interference (RFI) or other data quality issues must be removed before power spectrum analysis is performed. Because of its redundant hexagonal subarray, wide field of view (FoV), and sensitivity to low frequencies, the MWA is unique amongst SKA precursors in its capabilities for delivering EoR power spectrum measurements.

Science requirements/requested development

The existing software used to pre-process and calibrate MWA data did not meet the requirements of the EoR team. New pre-processing and calibration applications were therefore required to process data from the new MWAX correlator. Birli and Hyperdrive were developed, tested, and evaluated to establish confidence in their ability to deliver the level of precision required for EoR science. The MWA EoR Nextflow pipeline was then established using the

Hyperdrive calibration suite to process the large archive of MWA EoR observations that have accumulated over the last decade.

Technical developments during DSP

The AusSRC worked with the MWA EoR and MWA Operations team and developed a brand new pre-processing application, Birli (the Wajarri word for ‘lightning’), for the new MWAX correlator. Birli was written in the Rust programming language using modern software development practices. It was built for the specific demands of the EoR project and to replace the ageing Cotter MWA pre-processor. Birli is significantly faster, supports both MWAX and legacy correlator formats while retaining all the features of its predecessors, such as Cotter. It was deployed to the ASVO, and surpassed Cotter as the most popular pre-processing option.

Hyperdrive is an MWA calibration application, also written in Rust, which provides high-fidelity direction independent calibration for MWA users. The AusSRC helped with the development of Hyperdrive from an early alpha version with the implementation of measurement set support and other features. Current work involves the implementation of direction dependent calibration.

Most observations from radio telescopes contain some faults, and these faults can manifest in diverse and subtle ways that are difficult to detect without a multi-faceted approach. The MWA EoR Nextflow pipeline utilises Hyperdrive alongside Birli through the application of a suite of bespoke quality analysis scripts that look for data quality issues in various facets of the data. The observations are taken through a series of stages that scrutinise different data products for the same observation, preventing an observation from passing to the next stage if it does not meet strict quality requirements.

Metadata

The metadata stage (Figure A3.14) is the first of the stages within the MWA EoR Pipeline. Information is gathered about each observation from the Monitor and Control system via MWA Web Services. This ranges from scheduling information to information about faults that occurred during the observation, as well as information about the files archived for the observation. This information is then used to filter observations based on a set of criteria that prevents observations with too many faults from passing through onto further stages.

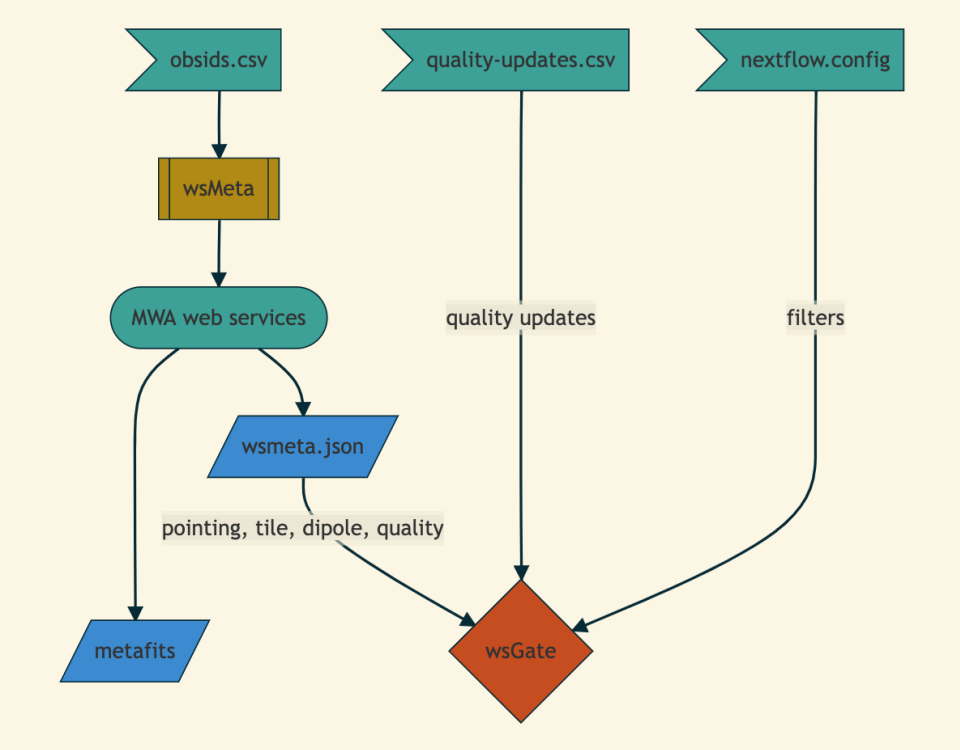

Figure A3.14: The metadata stage of the MWA EoR Pipeline.

Pre-processing

The pre-processing stage (Figure A3.15) produces and analyses FITS visibilities that have been pre-processed, flagged, and averaged by Birli via MWA ASVO. The flag occupancy of the FITS files are analysed for the presence of RFI over each coarse channel and the amplitude of the autocorrelations are measured across frequency and antenna. These measurements are used to produce a list of outlier antennas, which should be flagged in later stages. Observations whose flag occupancy is above the configured threshold are rejected.

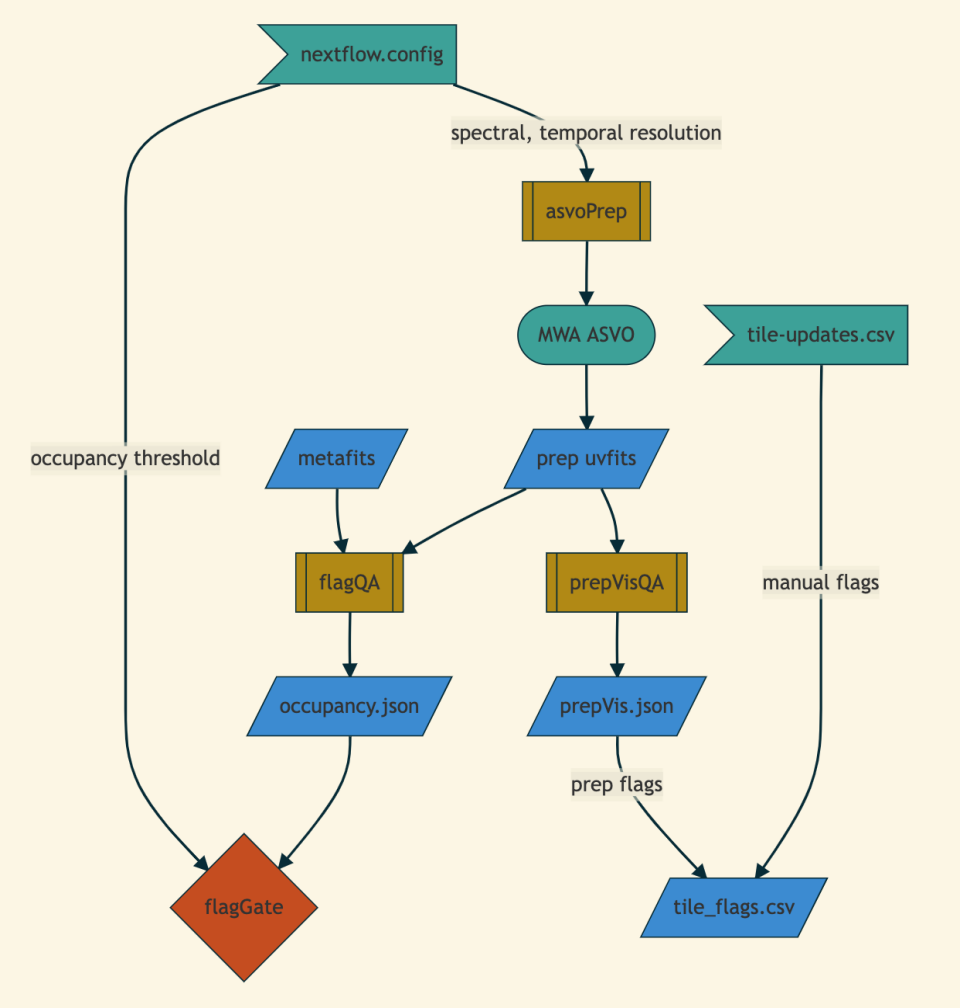

Figure A3.15: The pre-processing stage of the MWA EoR Pipeline.

Direction Independent Calibration

In the direction independent calibration stage (Figure A3.16), Hyperdrive is used to generate one or more direction independent calibration solutions using the MWA Long Baseline EoR Survey (LoBES) sky model. A statistical model of the calibration solution is used to smooth out any noise in the solutions, and the error of the calibration solution model is measured. If the variance or the error of an observation's calibration solution is too high, then the observation is rejected.

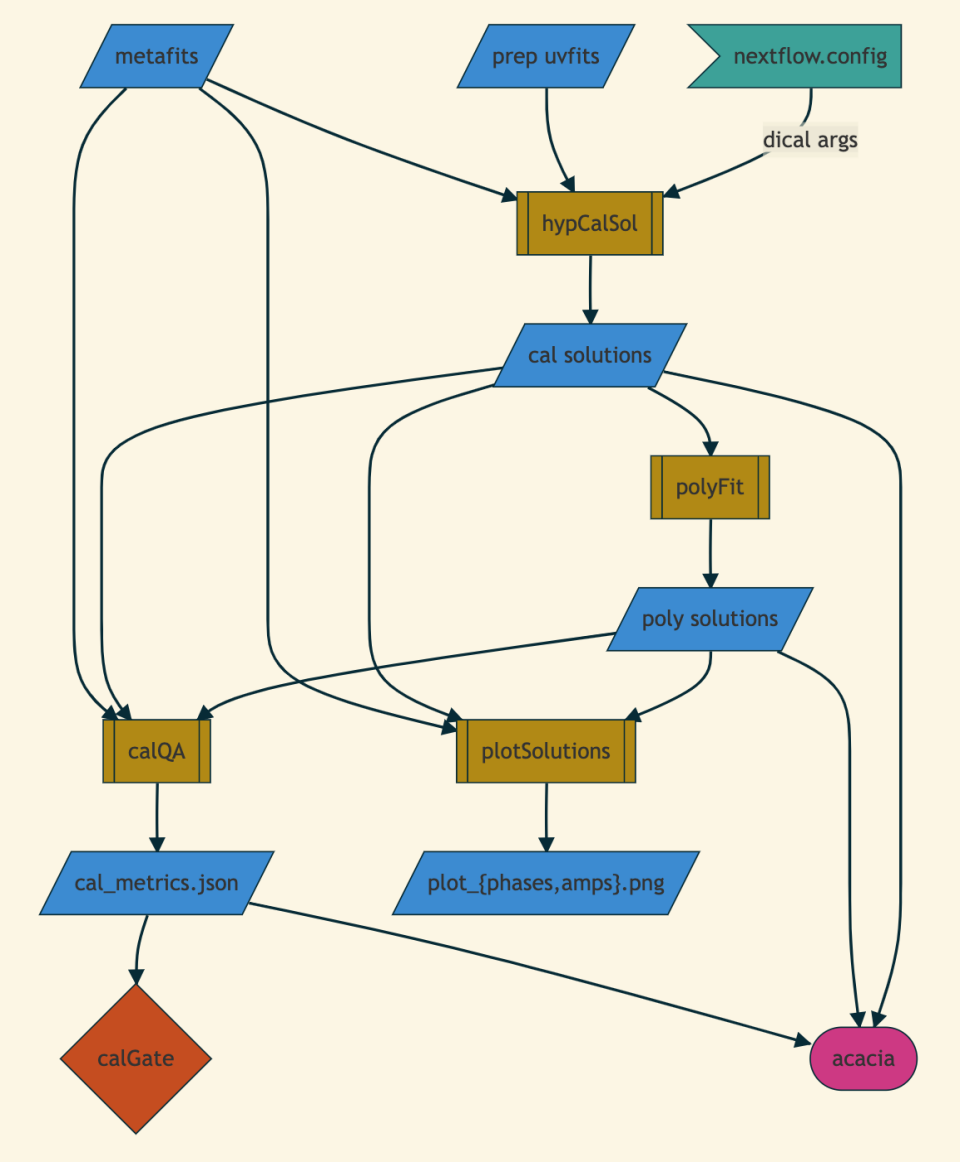

Figure A3.16: Analysis of Direction Independent Calibration solutions in the MWA EoR Pipeline.

Calibrated Visibility Analysis

In the calibrated visibility analysis stage (Figure A3.17), calibrated visibilities in the FITS file format are obtained by applying calibration solutions to the pre-processed visibilities. Baselines that occupy the same position in uv-space should have similar properties and so an analysis of redundant baseline groups is used to detect anomalous baselines that do not conform to the behaviour of their group.

An analysis of power spectrum metrics is performed on these visibilities, as well as their residual, which is obtained by subtracting the simulated sky model from the calibrated visibilities. Power spectrum window contamination measurements are important for detecting where observations are not fit for the final power spectrum integration.

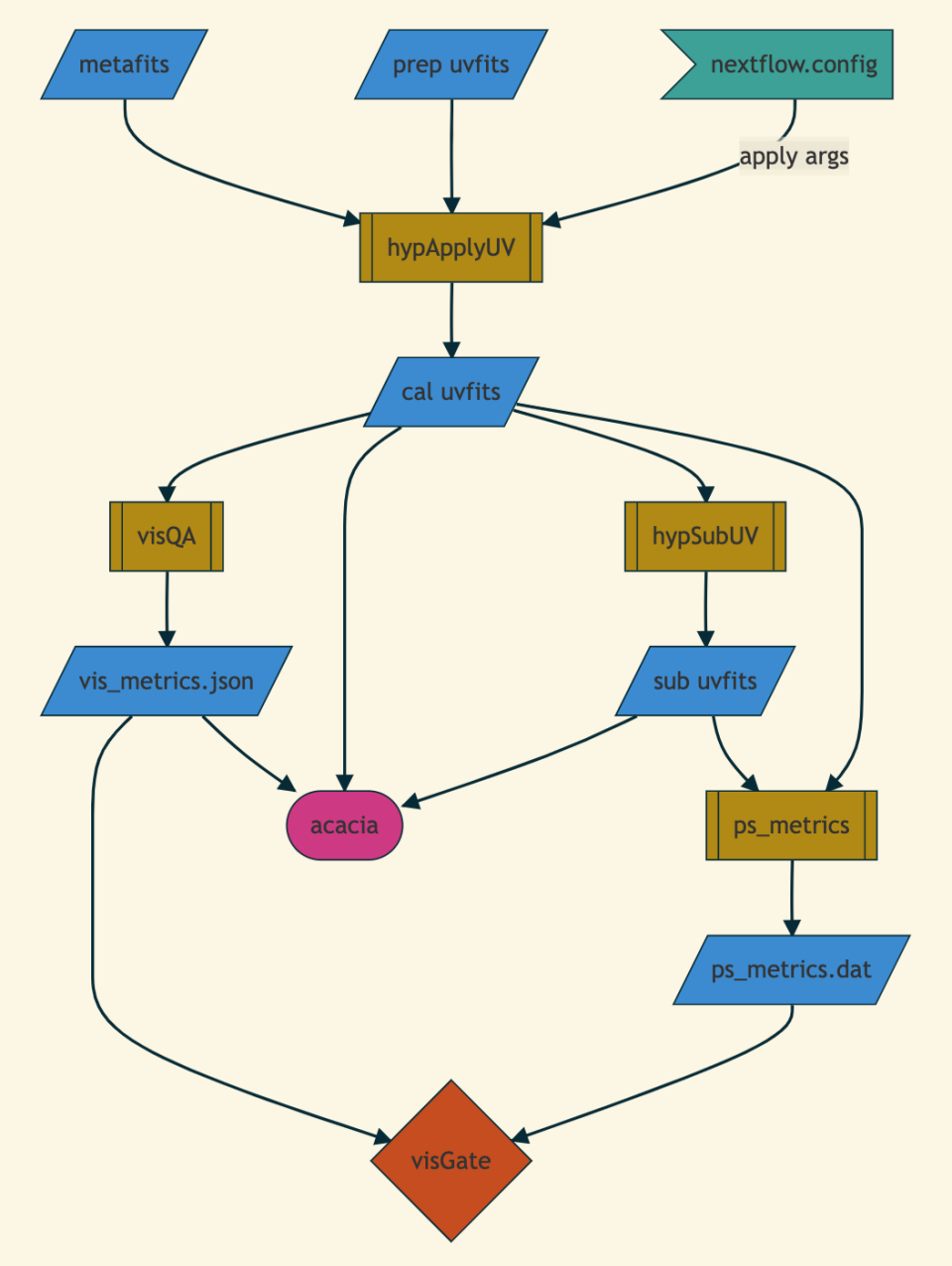

Figure A3.17: Calibrated Visibility Analysis.

Dirty Image Analysis

In the dirty image analysis stage (Figure A3.18), the calibrated and residual visibilities from the previous stage are analysed in measurement set format. WSClean (https://wsclean.readthedocs.io/en/latest/) is used to make dirty (non-deconvolved) images of Stokes XX, YY, and V polarisations, which are used to obtain polarimetric power measurements.

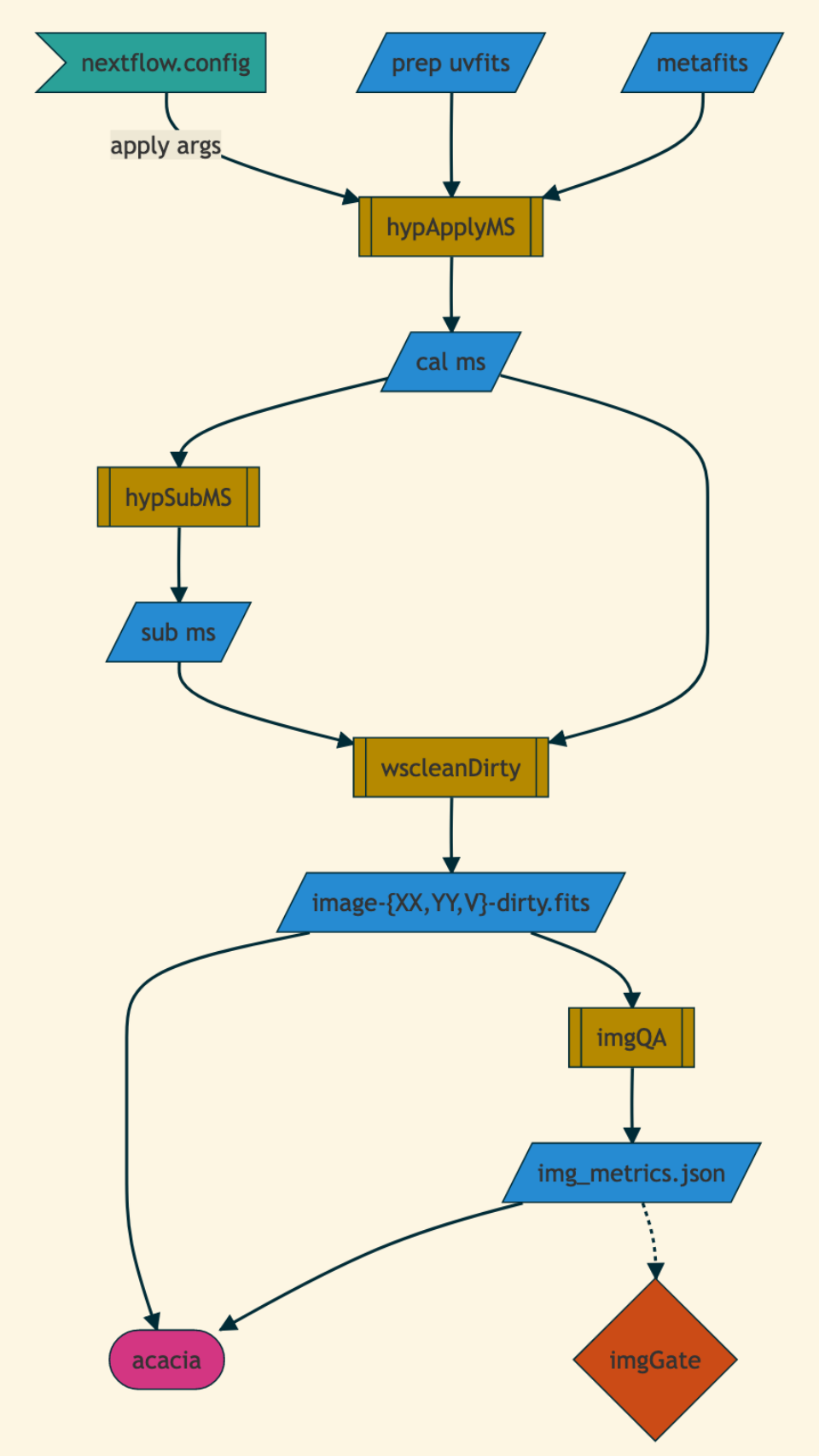

Figure A3.18: Dirty Image Analysis.

Technology/applications involved and/or tested

- AOFlagger
- CASAcore
- CUDA
- Docker
- Nextflow
- Singularity
- Slurm
- WSClean

Resources/service providers utilised

- DUG Technology
- MWA ASVO Services
- MWA Web Services: Detailed Observation Metadata, fault history
- Pawsey Supercomputing Research Centre: Garrawarla, Acacia

Applicability to SKAO

EoR is a primary science case for the SKA. By deeply understanding how the EoR analysis is performed on the precursor telescopes, vital information will be gathered about the data reduction requirements for the SKA.

Possible future developments

The EoR collaboration currently uses the MWA Real Time System for direction dependent calibration, however this is not fit for the purpose of processing data from the MWAX correlator and suffers from several software quality issues. Development of direction dependent calibration in Hyperdrive will be vital for mitigating issues due to ionospheric interference.

Spectral regularisation of calibration solutions will be important in mitigating the effects of spectral structure on calibration performance. Redundant calibration will be required to unlock the full potential of the MWA's hexagonal subarray. Furthermore, a Sky-Subtracted Incoherent Noise Spectra (SSINS) RFI flagging strategy will need to be added to the pre-processing stage to improve detection of faint RFI before the data is averaged.

POSSUM project

Project overview

POSSUM uses ASKAP to study magnetic fields in various environments across the Universe, through synchrotron radiation and its associated Faraday rotation. The novel capabilities of ASKAP will form a magnetic picture of the Universe with unprecedented detail. A key concept is the measurement of Faraday rotation measures, a property of radio sources that shows the strength and direction of magnetic fields along the line of sight.

Science requirements/requested development

Data post-processing

The AusSRC assisted POSSUM through the development and implementation of a data post-processing pipeline. Once POSSUM observations have been processed by ASKAP operations and made available through CASDA, the pipeline is executed to produce HEALPix tiles, which are then shared to CADC for further post-processing. The current POSSUM post-processing pipeline (from the perspective of the AusSRC) follows a series of steps:

1. Download processed observations from CASDA
2. Convolution to a common beam
3. Ionospheric correction on the stokes Q, U cubes
4. HEALPix tiling
5. Mosaicking of HEALPix tiles
6. Upload complete HEALPix pixels to CADC

There were two main components to the development of the post-processing pipeline. First, the implementation of the computational pipeline to automatically execute this workflow for any given observation. Second, the development of components into reusable, containerised, and production-ready code from simple scripts. Some of the components within the workflow had ongoing support, such as common beam convolution with RACS-tools (https://github.com/AlecThomson/RACS-tools). However, other components of the workflow, such as ionospheric correction and HEALPix tiling, required software engineering expertise from the AusSRC to run the codes in a production environment.

Temporary data storage

The mosaicking scheme for HEALPix tiles was complicated and relied on two or more observations before any tile could be completed. The AusSRC provided a temporary storage space for intermediate HEALPix tiles (e.g. between steps 4 and 5 in the pipeline) to support the operation and efficiency of the post-processing pipeline.

Technical developments during DSP

The AusSRC developed a computational pipeline in Nextflow for the POSSUM post-processing workflow, available in the POSSUM workflow repository. The AusSRC assisted POSSUM by providing code contributions to the RACS-tools code base and by containerising and re-writing the ionospheric correction and sky tiling scripts used in the pipeline. Docker images for these codes were published to the AusSRC Docker Hub repository and used in the post-processing pipeline.

Technology/applications involved and/or tested

- Docker
- GitHub (https://github.com/AusSRC/POSSUM_workflow)
- Nextflow

- Singularity
- Slurm

Resources/service providers utilised

- CASDA
- CIRADA VO Space
- Pawsey Supercomputing Research Centre: Nimbus, Magnus/Setonix

External collaborations

- CADC/CIRADA

Applicability to SKAO

Overall, understanding cosmic magnetism/polarisation and the origin of massive magnetic fields throughout the Universe is a key science goal for the SKAO. The developments for POSSUM – from the perspective of proper calibration, efficient data handling, and accurate science extraction – will give significant insight that will help to pave the way for SKA data.

More specifically, the sharing of POSSUM files to CADC is currently done with a VO service implemented by the CADC and is triggered manually when data is available on the AusSRC. Data sharing, in the form of database rows and files, is an important problem that the SRCNet will have to solve. Thus far, the SRCNet has explored the use of Rucio for data management of files. Future work with POSSUM will be to replace the CADC VOspace tools for file sharing with alternative systems, such as Rucio, which will allow the AusSRC and collaborating institutions to test data management tools with real precursor data as part of the SRCNet effort.

Possible future developments

In its current state, the POSSUM post-processing pipeline is relatively slow and inefficient compared to other pipelines the AusSRC has developed for other surveys. The software engineering work for POSSUM and their post-processing components has enabled them to be run in the context of the AusSRC workflow system, which allows the pipeline to be executed in different computing environments. However, there are still computing optimisations that can be applied to parallelise the tiling and ionospheric correction scripts (with both MPI and OpenMP; https://www.openmp.org) to better utilise hardware. In the future, there will be opportunities to profile the execution of the pipeline, identify opportunities for speedup, and re-write components of the pipeline to better use resources.

The AusSRC developed an event system which allows pipelines to be executed in response to messages from CASDA. The event system is also used by WALLABY and EMU for their post-processing pipelines. The POSSUM post-processing pipeline is still manually executed due to focus on development of the pipeline and its components. The AusSRC plans to add the POSSUM pipeline to the event system code so that new observations are automatically post-processed.

WALLABY project

Project overview

WALLABY aims to survey the entire southern sky with ASKAP, detecting HI in and around hundreds of thousands of galaxies. WALLABY will measure the HI properties of each galaxy, derive its distance, total mass, and dark matter content. The 30 arcsec angular resolution of the survey will enable detailed kinematic modelling of thousands of resolved galaxies and greatly facilitate in identifying counterparts in other wavelengths.

Science requirements/requested development

Post-processing workflow

The primary requirement for the AusSRC by the WALLABY team was for the development of computing pipelines to run the post-processing workflows more efficiently. Prior to AusSRC involvement, the process for reducing data to produce catalogues of sources was exceedingly manual, requiring multiple team members to interact with and manipulate the data before catalogues were produced. This workflow was inefficient and not suitable to handle the increased rates of data flow expected for the full 5-year survey. As such, a more efficient method for performing post-processing was required.

It was requested that the AusSRC develop a computational pipeline that captures the science data post-processing workflow that takes footprint pairs, performs mosaicking and source finding, and generates data products. This code needed to be sufficiently modular so that it could be applied to any tile and run on different computing resources. The WALLABY team wanted to use phase 2 of the pilot survey as an opportunity to develop and test this pipeline, so that it could be relied on for full survey operations.

There is complicated logic for determining how to process observations for WALLABY. Each tile in the sky is made up of a pair of observations (also referred to as ‘footprints’) that are offset slightly and need to be mosaicked to produce the high signal-to-noise image required for science (see Figure A3.18). The border of these individual mosaics will be low signal-to-noise regions. To ensure that the entire sky is covered by this high signal-to-noise image, there is an overlapping region between each tile. Therefore, there was a requirement to develop a source finding strategy for executing the post-processing pipelines only on high signal-to-noise regions progressively as fields are observed.

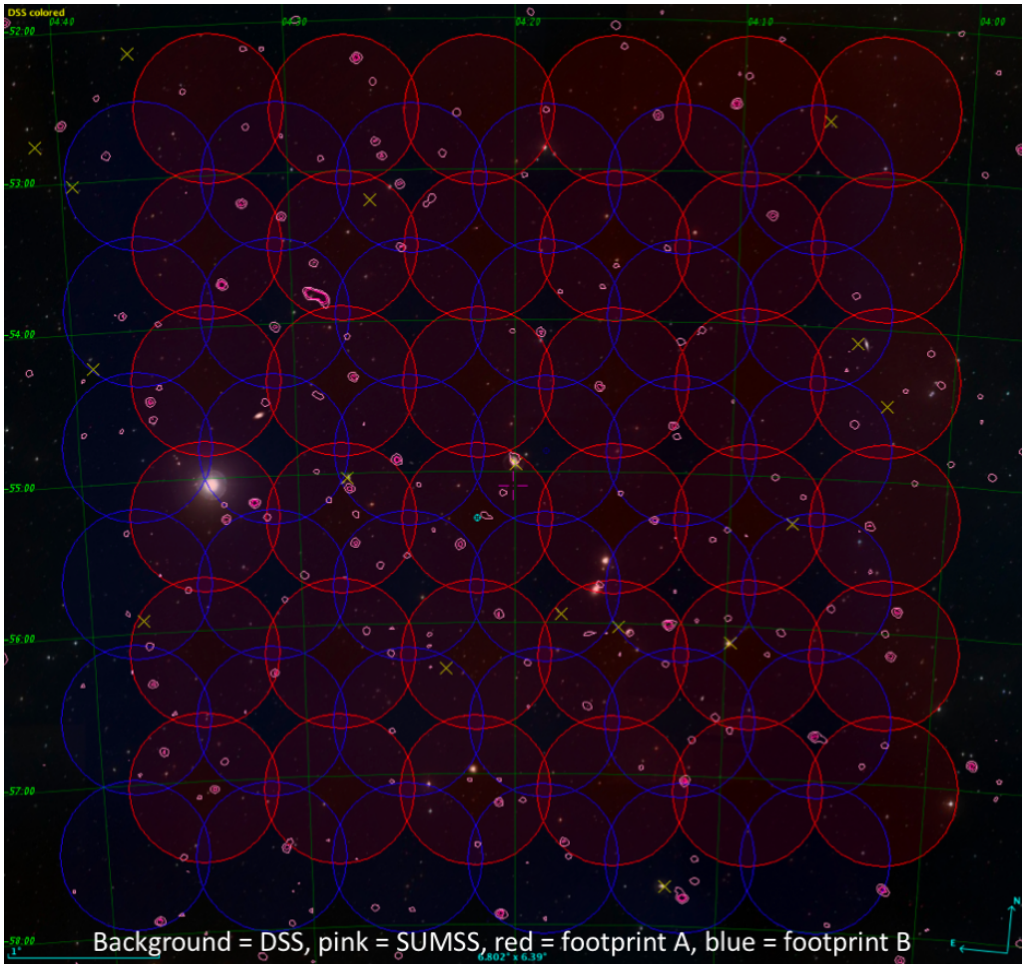

Figure A3.18: Single WALLABY tiles are created from a mosaic of two observations, slightly offset in right ascension and declination, to produce a high signal-to-noise image.

Data storage and access

The data products that WALLABY produce are a catalogue of sources and their properties, associated sub-cubes, moment maps, and spectra. Prior to AusSRC involvement, these products and catalogues were stored as files. This meant that manipulation of the data was done on individual files and distributed to team members was done through tools such as Dropbox data sharing links.

The AusSRC sought to improve the way that the WALLABY team interacted with data by introducing a central database for storing catalogues and data products, and providing ongoing support for maintenance. Web services were developed on top of the database system to provide:

- users with easier access to the data
- project scientists with more convenient methods of data manipulation
- WALLABY members with the ability to contribute to the advanced data products

Technical developments during DSP

The AusSRC supported WALLABY by developing software for processing and distributing pilot survey data in order to prepare an automated system for full survey operations. This includes the development of data post-processing tools for ASKAP observations, the interactive interfaces for quality control of survey products, and systems for data distribution.

Data post-processing

As part of the data post-processing work for the WALLABY survey, the AusSRC developed computing pipelines to perform quality checking of ASKAP observations, processing of quality-checked footprint pairs to produce catalogues of sources, and a system for running these pipelines automatically in response to activity on CASDA. A collection of scripts that are executed in response to messages in the AusSRC event system are used to automatically trigger these computing pipelines in response to completed ASKAP observations.

All post-processing pipelines have been developed in Nextflow and are available in the WALLABY Pipelines repository. The Nextflow main pipelines define the logic for the workflow and the details for the computing tasks that are contained in individual processes. The code for the processes are captured in a collection of reusable Docker image components – stored in the AusSRC Pipeline Components repository – that can be shared between all the pipelines. All Docker images have been released to the official AusSRC Docker Hub registry for public access. Both the WALLABY quality checking and source finding post-processing pipelines are composed from this collection of components.

Quality checks are performed by the WALLABY science team on each ASKAP observation prior to the data being released by CASDA as a ‘level 6’ data product. The AusSRC assists WALLABY in the quality checking process by automatically running codes for performing preliminary source finding (using SoFiA-2) on new observations to generate moment 0 maps. The collection of moment 0 maps for the detections are then combined for viewing, allowing scientists to quickly identify potential issues with imaging. The WALLABY science team is able to quickly inspect the moment 0 map and report the quality of the observation to CASDA. The steps for the quality control pipeline are:

1. Download individual observations from CASDA
2. Source finding to generate detection moment 0 maps
3. Mosaicking of moment 0 maps

The post-processing requirements to produce WALLABY catalogues involve the reduction of a number of observations (at least in footprint pairs to produce single high signal-to-noise tiles) into catalogues. This workflow involves running ASKAPsoft’s mosaicking software, linmos, on observation pairs, running the source finding code (SoFiA-2) on the output tile to identify potential sources in the image, and writing the detections into the database. The AusSRC captured this workflow into a single Nextflow computational pipeline. The steps in the pipeline are:

1. Download footprint pairs from CASDA

2. Mosaicking footprints into a single tile
3. Source finding on mosaic
4. Write source properties to database

In addition to the computational pipelines for performing post-processing, the AusSRC developed a collection of scripts that run in the AusSRC event system for automatically processing high signal-to-noise regions of available fields during full survey. This system relies on messages published to the AusSRC event system from CASDA, as observations are available, and state information about the survey and how it is progressing. Then, based on a pre-defined tiling scheme, the AusSRC worked with WALLABY to define logic for determining which regions of the sky to run the post-processing pipeline.

The source finding code SoFiA-2 is used heavily as part of the post-processing pipelines for WALLABY. The AusSRC has worked closely with the SoFiA team to develop features for the source finding code such as wrappers to use for execution and for interaction with objects stores.

Figure A3.19 shows the tiling and source extraction strategy for full survey, where the tiles are observed in declination bands covering the sky. The strategy assumes that the tiling scheme for the survey is known in advance, which is essential in pre-determining which tiles are adjacent to know what regions to process. The same strategy was employed for phase 2 of the pilot survey, which allowed the post-processing system to handle cases where observations are in right ascension bands rather than declination bands. The regions shown in the diagram are:

1. Central regions of a single tile. The post-processing pipeline for this region can be executed when two footprints for a tile have been observed. The central 4x4 degree region is all high signal-to-noise after the mosaicking.
2. Adjacent regions between two tiles along declination bands. Where there are two adjacent tiles, the border of each tile will be low signal-to-noise and will overlap slightly with the other tile. They must be mosaicked together before the region in between can be processed. As shown in the diagram, the 4x6 degree region between two adjacent tiles within a declination band can be processed.
3. Regions between three adjacent tiles of two different declination bands. Tiles in declination bands will be offset in right ascension such that there are regions between groups of three tiles to mosaic and process. When all three tiles (or two in some edge cases, such as for pilot survey) are available they can be mosaicked and the source finding processing can be applied.
4. Edge regions of tiles that are not covered by the previous three regions. These are intended for completing post-processing on all regions of a tile in parts of the survey where there are no adjacent bands of tiles, or where an isolated region is observed.

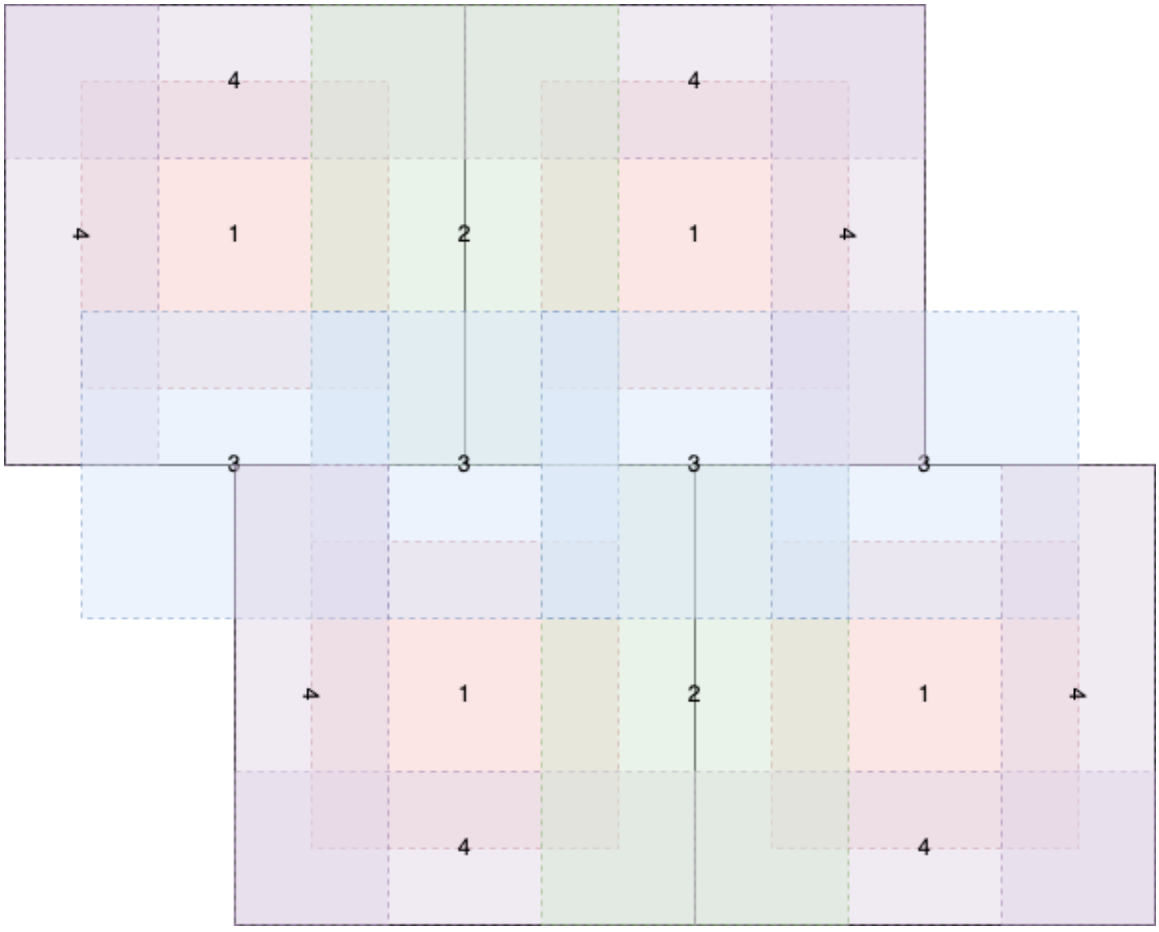

Figure A3.19: WALLABY source finding strategy for covering adjacent fields during phase 2 of the pilot survey. This strategy will be employed during full survey operations.

Database services

The AusSRC developed a schema for WALLABY source finding data products that can be stored in a relational database. The schema contains tables for capturing metadata in the run of the post-processing pipeline, the particular regions on which it is run, the detections produced, the data products that are created by the pipeline, the finalised sources that are defined by the WALLABY science team, and other tags and comments that can be attributed to sources.

The code for deploying a PostgreSQL database for the aforementioned schema is stored in the WALLABY Database repository. A Dockerfile definition and official image that is shared to Docker Hub by the AusSRC makes the deployment of such a database efficient and convenient for collaborators. DevOps practices are utilised to ensure that changes in the schema during development are reflected in the latest officially supported Docker Hub image. This image has been deployed for the WALLABY science team and access is available through SQL queries.

In the WALLABY Database repository, Django code is also used to present an object-relational mapper (ORM) interface. This allows programmatic access to WALLABY data products through Python objects, rather than using SQL queries and therefore provides a more familiar

programmatic interface for astronomers to access data. The ORM has enabled WALLABY team members to write custom queries for accessing the data based on their own science analysis requirements, which have subsequently been shared to the rest of the team.

While the ORM helps significantly with database access, the syntax is still unfriendly due to the length of the code required to perform common but complex queries. The AusSRC worked with WALLABY project members to develop a repository of simple data access functions for the most commonly used queries. These simple Python module functions allow users to access data with single line commands in many cases and simplifies the Python scripts and notebooks for users. This repository is publicly available to allow members of the WALLABY team to contribute to this module and further simplify the data access for the rest of the team and/or to share useful scripts and functionality.

The advanced data products of the WALLABY catalogue are generated by individuals and groups spread internationally. Currently, there are three groups working to generate different data products. The AusSRC is responsible for the generation of source finding products and the storage of that data in a locally hosted database. CIRADA is responsible for generating kinematic models for resolved sources in the source finding database. IAA-CSIC (Institute of Astrophysics of Andalusia - Spanish National Research Council) generates cross-matches between the source finding products and other multi-wavelength sources. To enable this collaboration, the AusSRC worked with these other proto-SRCs for developing a system of database replication to share advanced data products generated between centres.

Bucardo (https://bucardo.org), an open source database replication software, has been implemented at each of the database sites. Each site has certain tables that it is responsible for, where it is the “owner” and where the other two sites are replicas. There is no situation where two databases can own a given table, so it is never a case of multi-master replication, which simplifies the replication process considerably. This scheme establishes automatic sharing of advanced data products to the entire WALLABY science team when they are created by each proto-SRC.

Interfaces for data manipulation

The AusSRC attempted to minimise the number of custom interfaces that need to be developed to provide a wide range of functionality to the WALLABY team by leveraging existing technologies, such as JupyterHub and Django, for presenting a visual interface for data access and manipulation. Template notebooks are used by the science team for the manual inspection and manipulation of source data, downloading and analysing data, and exporting data in appropriate formats for public data release. Where the functionality provided by notebooks is insufficient, the AusSRC has also developed a Django web interface for more visual interaction with the data.

There are template user notebooks for downloading WALLABY advanced data products and for reading the data into memory so that it can be directly manipulated in the notebook. This allows WALLABY users to download data for offline use, or to use on AusSRC resources as the platform for doing their science. The template notebook demonstrates how users could generate

figures from the WALLABY data for their science directly in the notebook. Read-only access to the database is granted to these users and so they are not able to directly manipulate WALLABY data.

The admin notebook provides functionality for WALLABY project scientists to resolve conflicting detections that are generated by the source finding pipeline and to make data publicly available, at the appropriate times. The notebooks enable the use of Python scripts for automatic resolution where possible, but also small widgets within the notebook to visually inspect detections and connect custom functionality with clickable buttons. A WALLABY admin is able to go use a single notebook to filter through all detections from a new pipeline run and select real sources. Additional notebooks are provided for adding comments and tags to the sources that require more information. When the data is ready for public release, WALLABY admins are able to quickly export the catalogue and products into a format required by the external archives (such as CASDA) for public release.

Technology/applications involved and/or tested

- Bucardo
- Docker
- GitHub (https://github.com/AusSRC/WALLABY_database)
- Jupyter Notebooks (https://github.com/AusSRC/WALLABY_notebooks)
- Nextflow
- PostgreSQL
- Python
- RabbitMQ
- Singularity
- Slurm

Resources/service providers utilised

- CASDA
- Pawsey Supercomputing Research Centre: Nimbus, Magnus/Setonix

External collaborations

- CADC/CIRADA
- IAA-CSIC/SPSRC
- SoFiA team

Applicability to SKAO

The development work that the AusSRC alongside WALLABY has done as part of the DSP has led to various learnings that are valuable to the SKAO. By working with real users and attempting to develop an operational system for handling precursor data processing, storage, and archiving, the AusSRC contributed to the requirement generation phase of SRCNet design. The AusSRC has experimented with database replication for collaboration between different proto-SRCs and for sharing survey data products with internationally-distributed scientists.

The WALLABY survey has collaborators from various institutes that work together towards the generation of advanced data products that are shared with the team. Through the work achieved during the DSP, and after establishing database systems for storing these advanced data products, the AusSRC has developed a prototype system for replicating database tables between geographically distributed data centres. Bucardo has been implemented for establishing master-replica sharing of source finding products from the AusSRC to the Spain Prototype of an SRC (SPSRC) and CIRADA, and kinematic products from CIRADA to AusSRC and SPSRC. Figure A3.20 shows how this database replication scheme is used for the distributed generation of advanced data products.

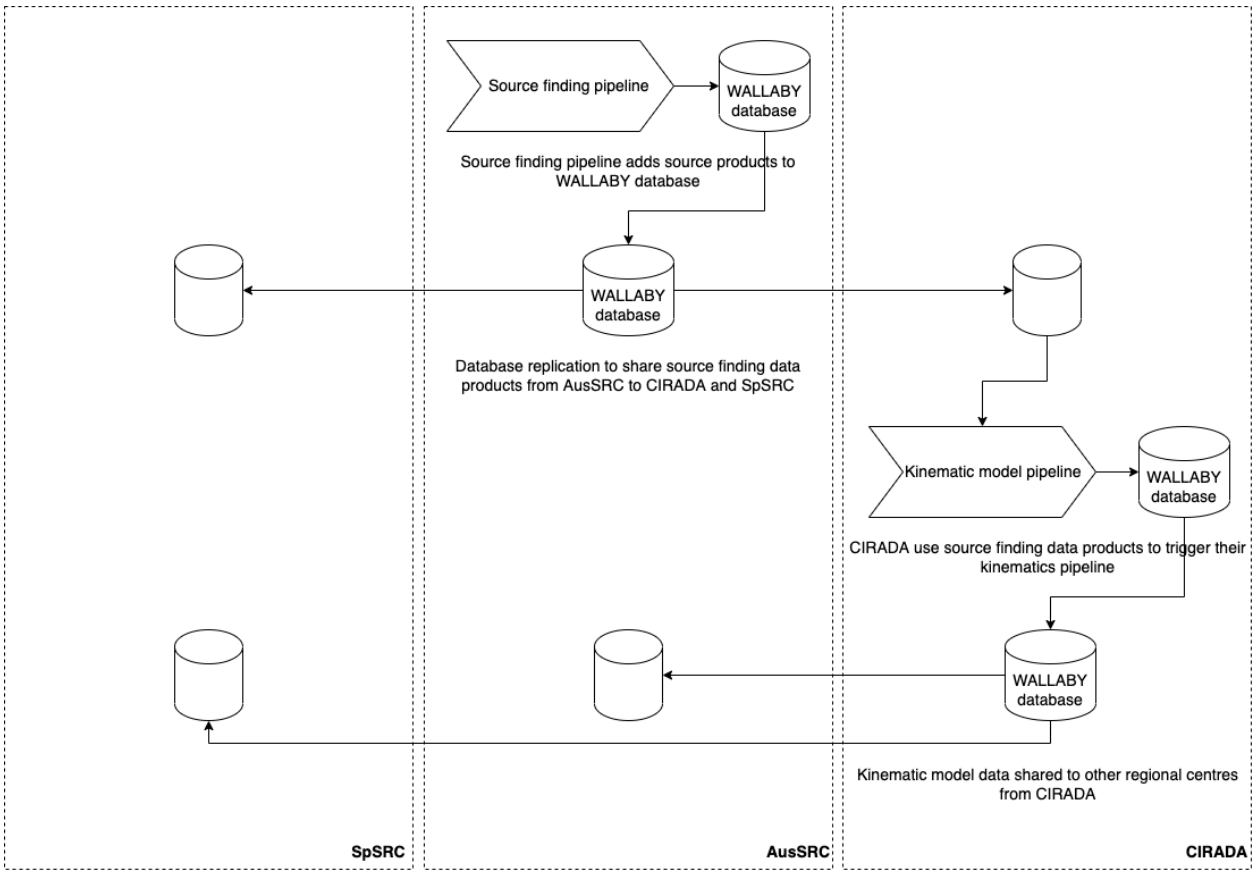

Figure A3.20: Flow diagram showing the distribution of advanced data product generation in WALLABY. Bucardo is used to replicate source finding data from the AusSRC pipelines to CIRADA and the SPSRC. CIRADA then produces kinematic models which are shared to other SRCs.

Sharing and replication of data products between SRCs is a problem that the SRCNet will need to solve. In early program increments of the SRCNet design, developers explored Rucio as a technology for sharing file data products between SRCs. The AusSRC development activities for WALLABY will enable further exploration and testing of various database systems and technologies with real data products across the SRCNet.

Possible future developments

There is ongoing development work that is still required beyond the DSP to develop the prototype AusSRC system into a mature software platform. For WALLABY specifically, this includes development activities to produce a production-ready database replication system with additional services for monitoring the network and operations, and further development of the post-processing event system for supporting other requirements of the science team.

The database replication system is currently in a prototype stage, requiring command line interface setup and maintenance. Updates to database tables, or the addition of new replicas, will require a user to SSH into the computing nodes and update the Bucardo configuration manually. While this achieves the current requirement, in that it allows data products to be shared and enhanced in a distributed manner, it will be important for the replication network to be more reliable going into full survey. This includes DevOps for the database replication code and the development of additional services and systems for monitoring and operating so that minimal system administrator support is required. Such a system would be one of the first instances of distributed enhancement of data products in the SRCNet.

The database systems currently contain all of the information required to determine the state of survey post-processing activities. The database tables include information such as the tiling strategy, which observations have been stored, and which have been processed for WALLABY for any given survey component. Currently, this is presented to the user in a very simple visualisation service through the AusSRC platform. Future development work is required in order to better communicate the post-processing status to WALLABY users, which may include incorporating information about post-processed fields into existing visualisation tools (e.g. Aladin) and presenting that through the AusSRC platform.

The present post-processing framework is only utilised for executing quality control and source finding pipelines, developed by the AusSRC, in response to updates in imaging or post-processing. The system however, has the potential to be used in a more generic manner, allowing other workflows to be executed by scientists in response to WALLABY updates and events in the AusSRC system more generally. We have already identified other workflows that can be executed in this way, such as a bespoke, high spatial resolution imaging pipeline or other quality control pipelines for Milky Way region WALLABY cubes. Future work will likely include updates to the system to incorporate these workflows and eventually to make the AusSRC event system more accessible to scientists.

A4. Technical Reports - PaCER Projects

BLINK

Project overview

FRBs are very bright milli-second radio pulses that originate from distant parts of the Universe. The reduced signal-to-noise ratio at low frequencies, due to long dispersive delays and sky background temperature, make FRB detection at low frequencies challenging. It is computationally expensive, time consuming, and also has large memory requirements. As a result, despite extensive searches below 350 MHz, only one detection, FRB 200125A, was discovered in non-targeted searches. Targeted searches were slightly more successful with many pulses from a repeating FRB 180916B detected. The BLINK project aims to develop an alternative, image-based FRB search capability for low-frequency instruments. Developments have been made on the millisecond imaging software for MWA data and will be further optimised through implementation on GPU to capitalise on parallel programming.

Science requirements/requested development

The BLINK project aims to provide an in-depth study of a large population of low-frequency FRBs. MWA's wide bandwidth, operating at 70-300 MHz, and its huge FoV, makes it suitable for a large-scale survey of the sky for low-frequency FRB searches. Although previous searches for FRBs using the MWA were successful in terms of establishing upper limits on the sensitivity of the searches, no FRB has been discovered by the MWA. Computational costs can be significantly reduced by using image-based approaches to prepare data for FRB searches instead of the traditional approach of tied-array beamforming. The sensitivity of these image-based FRB searches are significantly improved by using the MWA's FoV and allocating greater observation time.

The processing efficiency of the pipeline is enhanced through parallel programming provided by GPUs. These processing efficiencies means that data is processed as close to real time as possible, which will increase the overall number of FRBs detected each year.

Through the pipeline, numerous FRBs will be discovered and many of these will be found for the first time at lower frequencies. This will increase knowledge about the FRB population, where at present, more is known about high-frequency sources. The low-frequency detections will provide strong constraints on the local environment and the emission mechanism of FRBs. They also provide information along the line of sight to the FRB. Tackling the large volumes of data, computational costs of preparing data for FRB searches, and associated processing requirements of this search, along with expanding the understanding of low-frequency FRB population, will also lead the way for future surveys with the SKA telescopes.

Technical developments during DSP

Benchmarking single pulse search using PRESTO

During benchmarking tests for the time taken to search for single pulses – from the pulsar PSR J0027 - 1956, which was discovered by the MWA in one of the observations made as a part of the Southern-sky MWA Rapid Two-meter (SMART) survey (Obs ID 1226062160) – it was noted that beamforming in a single pointing direction was the most time consuming step, taking almost 48 minutes. Figure A4.1 shows the overall time taken to perform the single pulse search in one beam-formed direction. These values clearly suggest that the traditional approach to prepare data for single pulse searches is quite time-consuming, which demonstrates the requirement for image based approaches.

Observation ID	Max DM	prepsubband	single pulse search	Total Processing Time
1274143152	60	00:08:31	00:00:22	00:48:27
	500	00:09:18	00:03:30	00:52:22
	1000	00:17:57	00:02:11	00:59:42
1226062160	60	00:14:25	00:00:29	01:03:40
	500	00:16:43	00:01:56	01:07:25
	1000	00:17:04	00:03:18	01:09:07

Figure: A4.1 The benchmarking of different steps of beamforming, de-dispersion and single pulse search is shown above.

Benchmarking and Optimising the current imager using CUDA

The current version of the imager tool is implemented completely in C++ and generates all-sky images from the data. The most time-consuming sections for this tool were identified with a goal to optimise those sections using GPUs and CUDA. Figure A4.2 shows the imager performance with and without GPU implementation and Figure A4.3 is the image that was produced. The GPU implementation of imager runs almost 2.5 times faster than the original standard central processing units (CPUs) version.

		CPU Version						
		Image Size in Pixels						
		180*180	256*256	512*512	1024*1024	2048*2048	4096*4096	8182*8182
Function Name	Description	(s)	(s)	(s)	(s)	(s)	(s)	(s)
bool CPacerImager::run_imager(){..}	Full execution of imager (including I/O) took :	0.06	0.08	0.13	0.45	2.33	10.43	47.71
bool CPacerImager::run_imager(){..}	full imaging (gridding + dirty image) took :	0.06	0.07	0.13	0.45	2.33	10.42	47.71
if(do_gridding){..}	do_gridding took :	0.02	0.02	0.03	0.05	0.12	0.4	1.64
if(do_dirty_image){..}	do_dirty image took :	0.03	0.04	0.09	0.36	2.05	9.39	43.44

		GPU Version						
		Image Size in Pixels						
		180*180	256*256	512*512	1024*1024	2048*2048	4096*4096	8182*8182
Function Name	Description	(s)	(s)	(s)	(s)	(s)	(s)	(s)
bool CPacerImager::run_imager(){..}	Full execution of imager (including I/O) took :	0.42	0.43	0.46	0.64	1.96	7.13	18.37
bool CPacerImager::run_imager(){..}	full imaging (gridding + dirty image) took :	0.42	0.42	0.46	0.63	1.96	7.12	18.37
if(do_gridding){..}	do_gridding took :	0.02	0.01	0.03	0.04	0.13	0.44	1.59
if(do_dirty_image){..}	do_dirty image took :	0.39	0.41	0.41	0.56	1.72	6.17	14.65
	CPU_time/GPU_time	0.14	0.19	0.28	0.7	1.19	1.46	2.6

Figure A4.2: Shows the imager performance with and without GPU implementation.

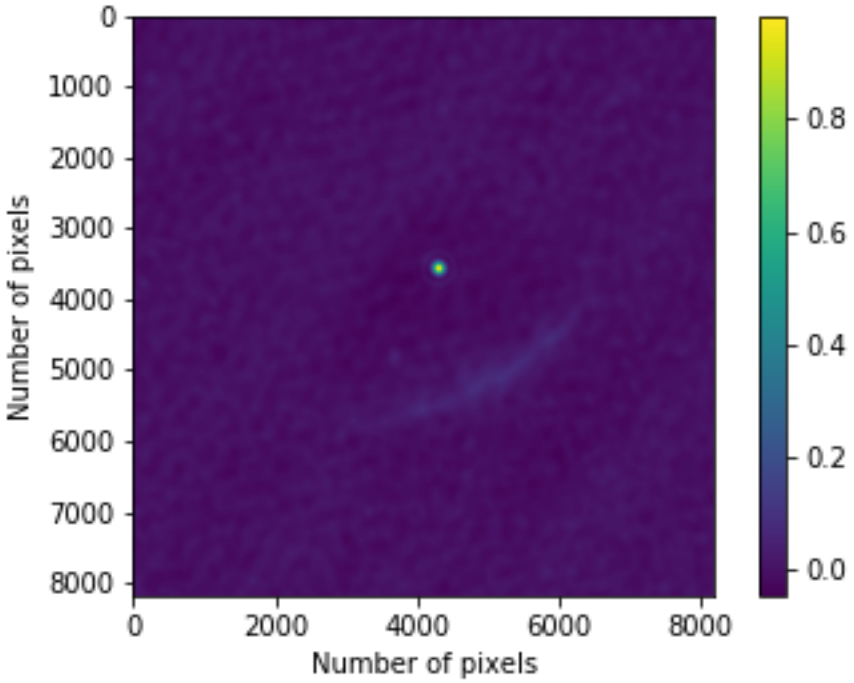

Figure A4.3 The all-sky 8182 x 8182 image produced by the current version of the imager software using visibilities. The bright spot roughly at the centre is the sun and the arc below indicates the galactic plane.

Technology/applications involved and/or tested

- CUDA
- GitHub and GitLab

Resources/service providers utilised

- Pawsey Supercomputing Research Centre: Garrawarla, Setonix, Topaz

Applicability to SKAO

The BLINK project will greatly improve the methods used to handle vast volumes of time-domain data. The optimised FRB search and imaging pipeline developed as part of this project will be applicable to low-frequency data from other telescopes, such as SKA-Low.
The effective utilisation of GPUs will also play a significant role while working with data from the SKA telescopes.

HIVIS

Project overview

The HIVIS project aims to address one of the most significant outstanding Grand Challenge Problems for the SKA – how to optimally image multi-epoch deep datasets. The critical missing gap impacts more than a third of the identified High Priority Key Science projects, yet no solution exists. Through the development of a sparse data storage and processing pipeline based on uv-grids, the HIVIS project aims to reduce the visibility storage requirements for these projects by an order of magnitude. The developed methods will simultaneously enable critically needed reprocessing to optimise the scientific outcomes from these datasets and opens up the possibility for higher resolution spatial and spectral imaging than what is currently achievable.

As a testbed, Pawsey Supercomputing Research Centre will be used to image 250h of ASKAP data from the DINGO pilot surveys, along with 500h of its first ultra deep field. These will yield some of the deepest images ever taken of the HI content in the Universe, enabling groundbreaking new studies of the role this fundamental fuel has played in the ongoing evolution of galaxies and its connection to their dark matter halos. In addition, the results will demonstrate a solution for the SKA data challenges in deep imaging.

Science requirements/requested development

uv-Grid Stacking Pipeline

Producing a pipeline, implemented in DALiuGE, with the following functionality:

- Gridding of ASKAP visibilities for individual observation blocks
- Stacking the gridded visibilities from the multiple observations with correct weighting and flagging.
- Imaging of the gridded and stacked visibilities to produce correctly weighted and RFI free deep images.

uv-Gridding Statistics

Statistics supporting the use of gridded visibilities and RFI flagging are required to assess the fidelity of the data products and the successfulness of the process.

ADIOS Compression investigation

Investigating the potential compression strategies available within the ADIOS-2 software (https://csmd.ornl.gov/software/adios2), in order to reduce the space required for long-term storage of data.

Technical developments during DSP

uv-Grid Stacking Pipeline

The pipeline is visualised in Figure A4.1 using the EAGLE workflow tool, which is the primary means to develop a DALiuGE pipeline. The first step relies on the “scatter” block to produce grids for multiple input datasets, with the “gather” block grouping the gridded datasets together to be stacked. The final step produces the required image cube. This pipeline has been profiled on the Setonix supercomputer managed by the Pawsey Supercomputing Research Centre and the results are shown in Figure A4.2.

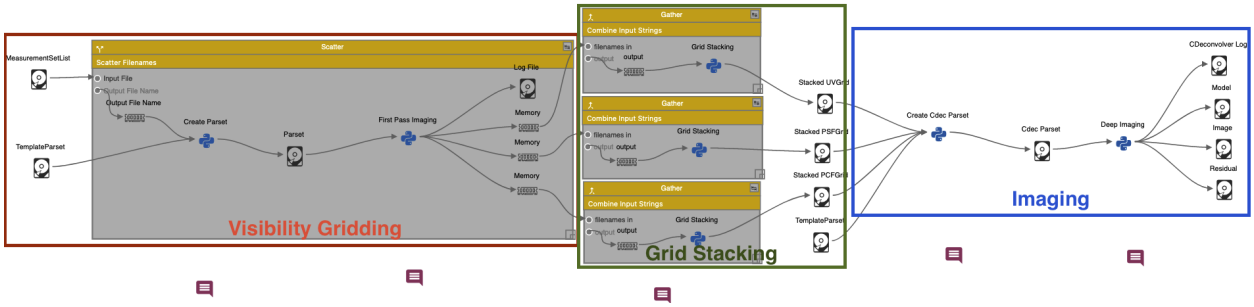

Figure A4.1: pipeline visualised using EAGLE.

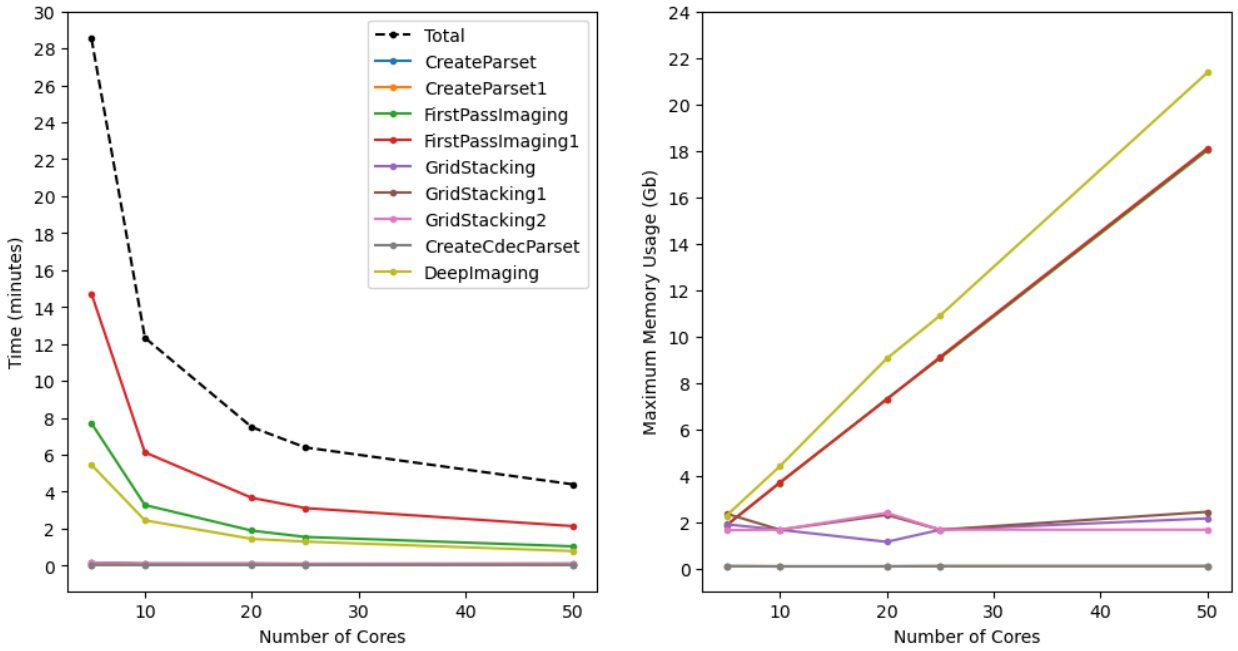

Figure A4.2: performance benchmarking tests on Setonix.

The profiling tests were performed on two datasets that were gridded separately, stacked, and then imaged. The timing results show that the time taken reduces with the number of cores used, as expected. This effect has reduced return as we increase the number of cores as I/O starts to become the dominant contribution to the time taken. The total memory allocated

appears to increase linearly with the number of cores, which is unexpected, as we would expect this to be approximately uniform and independent of the number of cores. This issue seems to appear primarily within the Gridding (FirstPassImaging) and the imaging (DeepImaging) steps.

uv-Gridding Statistics

Various approaches for the statistical analysis of gridded data are being investigated.
These statistics are based on the outputs from the previous stage (i.e. uv-grid stacking pipeline), so they form part of the verification and data quality tests. Preliminary results indicate that the data we selected has been successfully flagged, as there are very few outliers to the residuals, as shown in Figure A4.3.

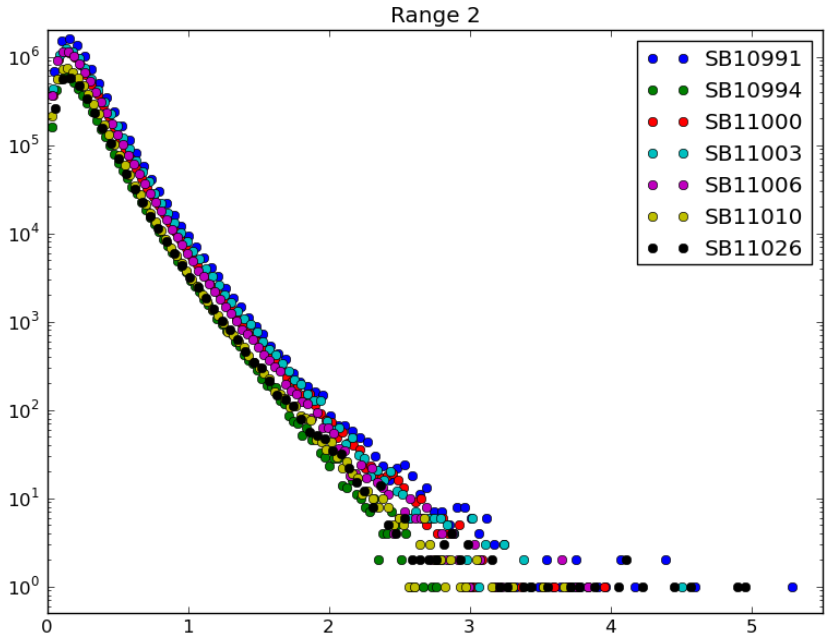

Figure A4.3: Voxel values for data taken in a frequency range known to be contaminated by global positioning satellites (1378-1385 GHz). The seven targeted datasets are overplotted, showing excellent agreement between the statistical distributions of the absolute value of the complex grid cells. We conclude that this data has been fully flagged and there is no opportunity for improvement.

ADIOS Compression investigation

The integration of the ADIOS-2 software into CASAcore is ongoing but not complete. Clashes between nested MPI implementations are likely to be the cause of the problems, and we have a demonstrated path to overcome such difficulties.
The investigation of the various compression options (particularly ‘lossy’ versus ‘lossless’) in the reduction of the space required for long-term storage of data waits on the roll-out of the functionality.

Technology/applications involved and/or tested

- CASAcore
- DALiuGE and EAGLE
- Docker
- Singularity
- Slurm
- YANDASoft

Resources/service providers utilised

- Pawsey Supercomputing Research Centre: Magnus/Setonix, Zeus

External collaborations

- Oak Ridge National Laboratory (ORNL) developed the HPC parallel writing ADIOS application and is leading the integration into CASAcore and the addition of more efficient storage algorithms. HIVIS is one of their prime demonstrations.

Applicability to SKAO

It is clear that groundbreaking SKA science will only be realised if the associated vast computational challenges can be met. The current SKA design, similar to that for ASKAP, only allows the raw spectral-line data products to be stored for a short period of time (~days), after which they will be imaged and discarded. This approach forces long duration projects to be combined in the image domain, with associated systematic risks for ‘baking-in’ errors.
The uv-grid pipeline developed under HIVIS offers a solution that will be as applicable to the SKA as it is to ASKAP. In fact, the SKAO has already expressed interest in DINGO’s approach as they address the otherwise unexplored issue of how to correctly image multi-epoch surveys, where the raw visibilities will not be retained.

The uv-grid pipeline also includes the crucial sparse data storage manager interface with ADIOS to take advantage of the potential data volume savings of the uv-grid format. This approach reduces the amount of I/O and data movement compared to a traditional imaging pipeline, removing bottlenecks and reducing power consumption. All told, the HIVIS program will not only enable the proposed core science and deliver a major legacy dataset for ASKAP, but it will also bridge a critical capability gap for the SKA of relevance to the entire international community, and place Australian researchers in the box seat for future gains in key associated areas of high priority science.

Possible future developments

The advanced developments that storing the daily data on the uv-grids rather than in images was fully discussed in the PhD thesis of Dr Rozgonyi (*Deep interferometric spectral line imaging by gridded visibilities,* 2021). HIVIS is based on the minimal viable model developed during the

PhD. A review of future possibilities was the basis of the final chapter, as summarised below:

- The minimal model has all baselines gridded to one cube. Once sparse storage is implemented, the data can be divided by baseline, and this in turn allows re-application of the station-based calibration. Thus the calibration can be improved with the delivery of improved sky-models after the initial processing. The first steps towards this are included in the HIVIS project, but there will be significant development work to fully explore these capabilities.

- The minimal model is limited to Hogbom-style cleaning, because major cycles are not possible after the application of the W-kernel. As we are limited to residual data products this was not an issue. Storing the W-axis in the datacube becomes possible once sparse storage is implemented. This would allow major-cycle based cleaning (Cotton-Schwab) and accommodate data with strong sources embedded.

- This approach would be also applicable to EoR studies, where the current storage plan is to discard the longer baselines. However, significant effort would be required to translate this approach to the very different data model of EoR and for low frequencies.

- An alternative storage model is to average the short baseline data longer than the long baseline data, referred to as Baseline Dependent Averaging (BDA). BDA retains the data closer to normal visibility data, but must average the data before application of the kernels. This process introduces errors so detailed comparison of the advantages of each approach should be done. A combined strategy, which seamlessly merges the two approaches into one data structure, can then be developed.

A5. Summary of engagement activities

Date	Type	Personnel	Details
8 Mar 2021	Media https://twitter.com/SKA_Australia/status/1368716594155642881?s=20	Karen	Short video clip (for SKA Australia) for International Women's Day showcase
15 Mar 2021	Media https://www.atnf.csiro.au/ATNF-DailyImage/archive/2021/15-Mar-2021.html	All	AusSRC featured in ATNF Daily Astronomy Picture
15 Apr 2021	Media https://cosmosmagazine.com/technology/computing/cosmos-briefing-supercomputing-and-big-data/	Karen	Panel discussion with Cosmos Magazine about super computing & "big data"
16 Apr 2021	Media release http://icrar.org/data-deluge?fbclid=IwAR3zQyNZFkqyyDkB4yQqY2iI2FwXOF5u0-AoJ_PWSmiuFAEibSulFDkpTTk	Karen	Media release about funding announcement. Subsequent interviews with television (channel 9) and radio (ABC WA regional drive, 89.7 FM, 92.1 RTRFM: https://rtrfm.com.au/story/australian-ska-regional-centre/?fbclid=IwAR3sZnRG3wSVMwimatsc37Aaa30d1tCVMpcGSfpMbSDnYYCI5rU4IAtVfSY). Featured on SpaceAustralia.com (https://spaceaustralia.com/index.php/news/australian-government-investing-big-ska-mega-science-project)
17 May 2021	Media https://businessnews.smedia.com.au/HTML5/default.aspx# (issue 17-30 May 2021, p.22-23)	Karen	Article in WA Business News about AusSRC, SKA, computing, and "big data"
5 Jul 2021	Publication (peer-reviewed) 10.1093/mnras/stab1881	Slava, Dave, Karen	First published science results using AusSRC resources
13 Jul 2021	Conference talks https://blogs.unimelb.edu.au/asa2021/#tabmain	Austin, Dev, Karen	3 AusSRC talks presented at ASA 2021

Date	Type	Personnel	Details
31 Aug 2021	Media https://www.innovationaus.com/first-nations-knowledge-and-insights-into-western-astronomy/?utm_content=buffer35a63&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer	Karen	Panel discussion with InnovationAus as part of their “see what you can be” campaign
17 Sep 2021	Outreach https://astrorocks-mtmagnet.com.au	Karen	Invited to talk about SKA and the AusSRC at the Mt Magnet AstroRocks Fest. Included local radio interview.
21 Oct 2021	Outreach/engagement https://www.expo2020dubai.com	Karen	Invited talk about SKAO Data Challenges and the role of SRCs at the Australian Pavilion (lead by SKA Australia)
25 Oct 2021	Outreach	Karen	School talk about astronomy and the SKA at Stirling Public School (Ontario, Canada)
28 Oct 2021	Publication	Austin	Conference proceedings for ADASS XXXI
28 Oct 2021	Conference talk https://www.adass2021.ac.za	Austin	Talk at ADASS XXXI
13 Nov 2021	Outreach/engagement	Kate	Volunteered for ICRAR booth at Astrofest 2021
13 Nov 2021	Outreach/engagement	Karen	Featured talk about “the future of science with the SKA” at Astrofest 2021
18 Nov 2021	Outreach/engagement https://pawsey.org.au/researchers/our-researchers-karen-lee-waddell-deciphering-the-data-of-the-universe-australian-ska-regional-centre/	Karen	Pawsey Supercomputing Research Centre showcase of researchers
17 Dec 2021	AusSRC Website news update https://aussrc.org/december-2021/	Kate, Karen, Dev, Gordon, Dave, Austin	Year in Review AusSRC website post
3 Mar 2022	Outreach	Karen	School talk about astronomy and the SKA at Ridley College (Ontario, Canada)

Date	Type	Personnel	Details
5 Apr 2022	Newsletter https://research.csiro.au/mro/mro-news-autumn-2022-aussrc-update/	Karen, Kate	Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory's Newsletter introducing the AusSRC - Autumn 2022 edition
21 Jul 2022	Conference Talk https://indico.mwatelescope.org/event/8/timetable/#20220721.detailed	Dev	Direction Dependent Calibration techniques
23 Jul 2022	Engagement https://www.globalaustralia.gov.au/aussie-inventors-showcase	Karen	Aussie inventors showcase to attract people in science and technology fields to look for careers in Australia
17 Aug 2022	Outreach	Karen	Multiple school talks about astronomy and the SKA at Mount Hawthorn Primary School during National Science Week
24 Aug 2022	Engagement	Kate	DISR/ASCC meeting, invited to provide a few dot points in a comms update that goes to the entire committee.
24 Aug 2022	Newsletter https://research.csiro.au/mro/mro-news-winter-2022-aussrc-update/	Dev, Kate	Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory's Newsletter explaining Birli - Winter 2022
29 Aug 2022	Outreach/engagement https://spaceaustralia.com/news/women-australian-space-community-dr-karen-lee-waddell	Karen	Showcasing women in Australia's Space/Science/Tech industry
13 Oct 2022	Conference talk https://supercomputing.swin.edu.au/events/acamar/register/acamar8/program/	Karen	Plenary talk about the AusSRC at ACAMAR 8
19 Oct 2022	Publication	Austin	SKA Science Data Challenge 2: analysis and results publication submitted to MNRAS.

Date	Type	Personnel	Details
29 Oct 2022	Outreach/engagement	Kate, Alex, Dave, Dev, Gayatri, Karen	Engaged with general public with AusSRC booth and interactive activity at Astrofest 2022
29 Oct 2022	Outreach/engagement	Dev, Karen	“Big Data” panel at Astrofest 2022
31 Oct 2022	Conference Poster	Austin	ADASS XXXII Conference Poster
8 Nov 2022	Conference talk https://www.atnf.csiro.au/management/atuc/2022oct	Karen	Title = Highlights from the AusSRC Design Study Program
10 Nov 2022	SKA Community Brief	Karen	Event by CSIRO to update the Aus research community
13 Nov 2022	Publications & data release https://arxiv.org/abs/2211.07094 https://arxiv.org/abs/2211.07333	Austin, Karen	WALLABY pilot survey phase 1 source finding and kinematics papers, published with accompanying data release