Portable workflows on interoperable infrastructures for astrophysics

Rob Simmonds

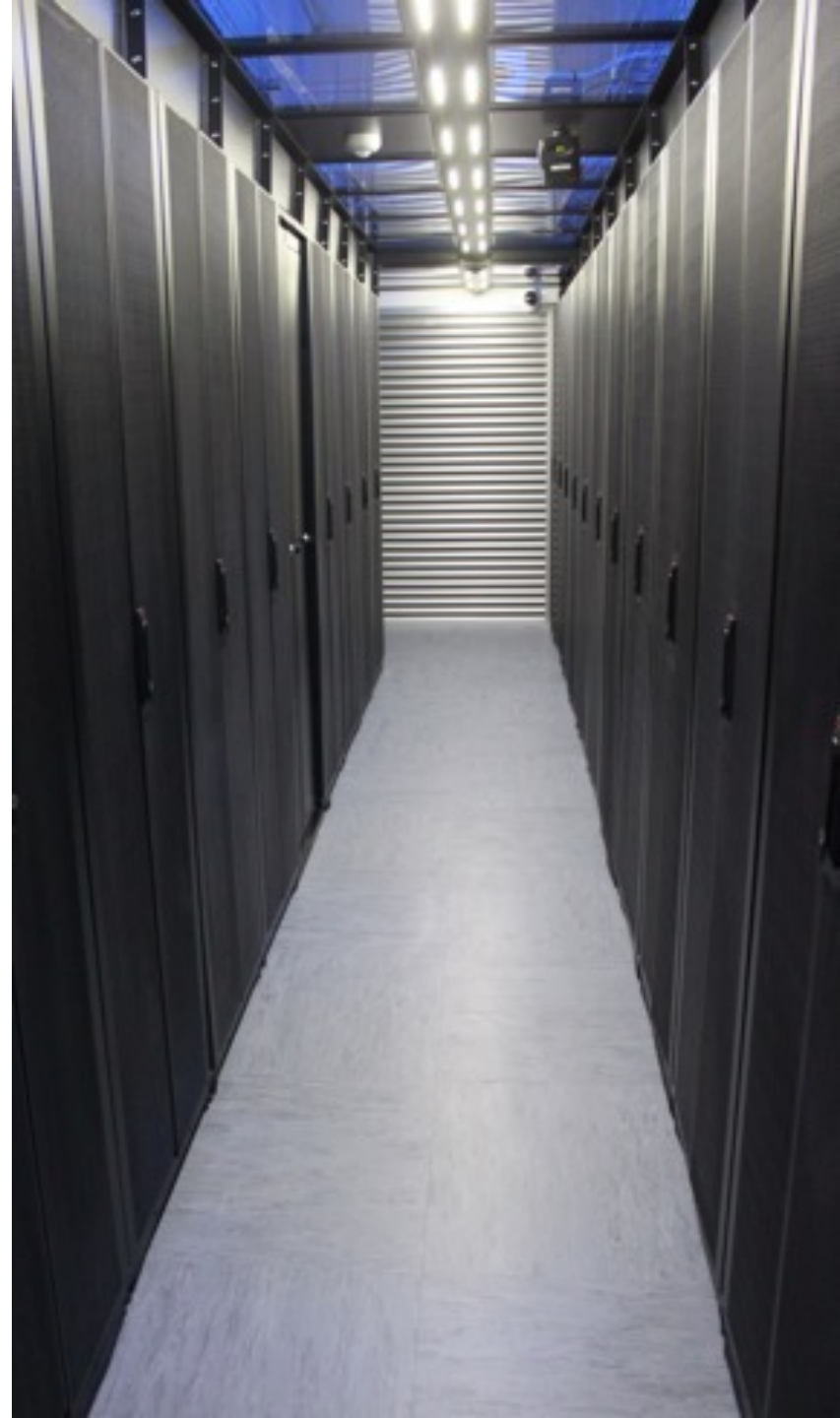
Professor UCT Computer Science

Acting Director IDIA

CTO Ilifu

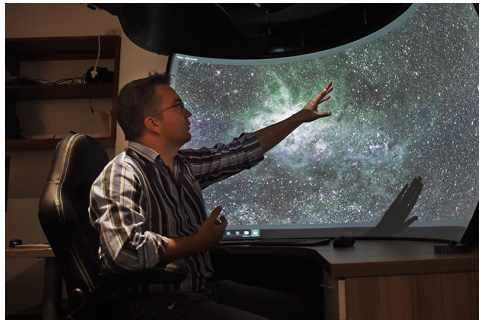
Talk Overview

- Astronomy in ZA
- Radio astronomy workflow
- The IDIA processing pipeline
- The CARTA visualization system
- Deploying specialized environments
- Future



ZA Astronomy

- MeerKAT – Radio telescope leading to the Square Kilometer Array (SKA) run by SARAO
- MeerLICHT – Optical telescope that tracks MeerKAT based at Sutherland
- SAAO runs many other optical telescopes at Sutherland
- IDIA/ilifu users primarily working with MeerKAT and MeerLICHT data



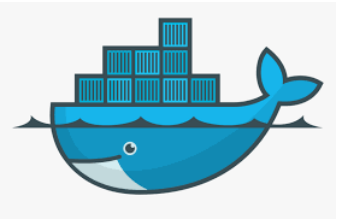
MeerKAT

Dishes	**64**
Dish Diameter	**13.5 m**
Total collecting area	**9161 m^2**
Pairs of dishes	**2016**
Maximum baseline	**8 km**
Resolution	**6” (robust = -0.5)**
A_e / T_{sys} (per dish)	**~6 m^2 K^{-1} (1.7 over spec!)**
Observing frequency	**580 – 3500 MHz**
Bandwidth	**856 MHz**
Spectral Channels	**32,768**

The path to the SKA

- A large number of technical challenges to overcome before being able to deliver SKA
    - Addressed in part by solving those related to its pathfinders, including MeerKAT that will form the centre of the SKA1-mid telescope
- The Inter-University Institute for Data-Intensive Astronomy (IDIA) focussed on the computational challenges
    - **IaaS cloud**
    - Data transfer
    - Software stack / containers
    - **Data processing (pipelines)**
    - **Visualisation**
    - Data quality

Why IaaS Cloud?



- Many users happy with generalized environments
  - Can be provided using VMs or bare metal deployment
- Users do want specific software stacks including specific versions
  - Can often now be provided by containers that run in the generalized environments
- Need for interactive web facing applications
  - VMs and/or Docker type containers and flexible network routing and proxying makes these easy to provide
- Some users want environments that they self manage for specialized applications
  - These final two points are simplified with IaaS cloud systems
  - IDIA runs an OpenStack IaaS system called ilifu

MeerKAT Large Survey Projects (LSPs)

Imaging

- LADUMA (Deep neutral hydrogen)
- MIGHTEE (Deep continuum imaging of the early universe)
- Fornax (Deep HI Survey of the Fornax cluster)
- MHONGOOSE (targeted nearby galaxies HI)
- MeerKAT Absorption Line Survey (extragalactic HI absorption)

Time domain

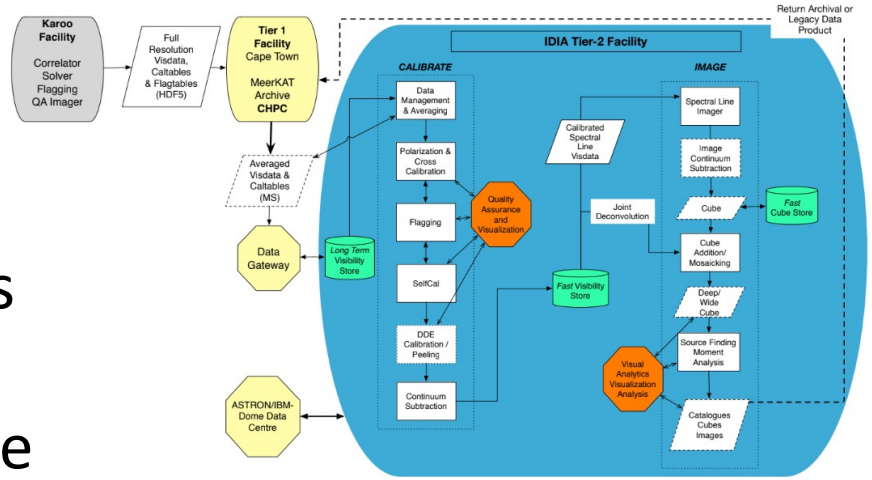
- ThunderKAT (exotic phenomena, variables and transients)
- TRAPUM (pulsar search)
- Pulsar Timing (MeerTIME)

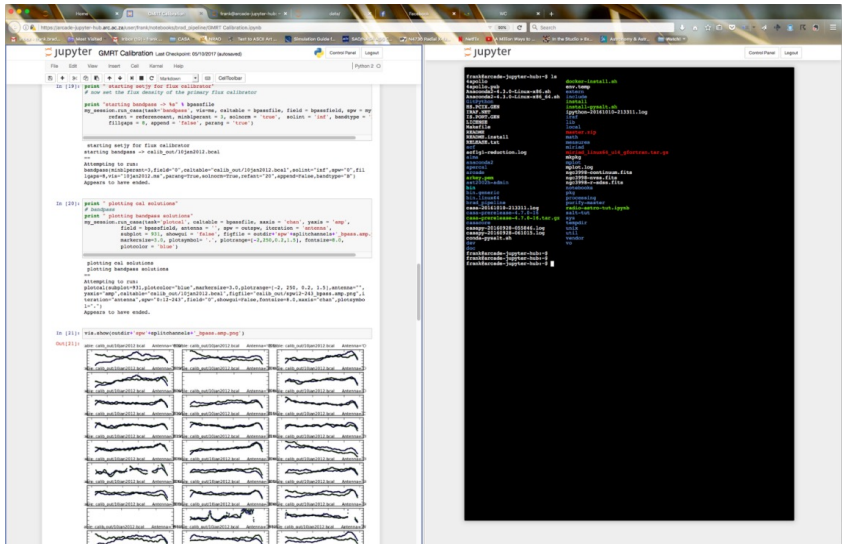
http://public.ska.ac.za/meerkat/meerkat-large-survey-projects

Radio Astronomer Workflow (simplified example)

- Data fetched from Observatory / Archive
- Pipeline software creates data cubes
- Data cubes examined using visual analytics tools
- Pipelines reparametrized, rerun, and cubes reexamined to expose “interesting” features (the science bit…)
- Camara ready images sliced from data cubes

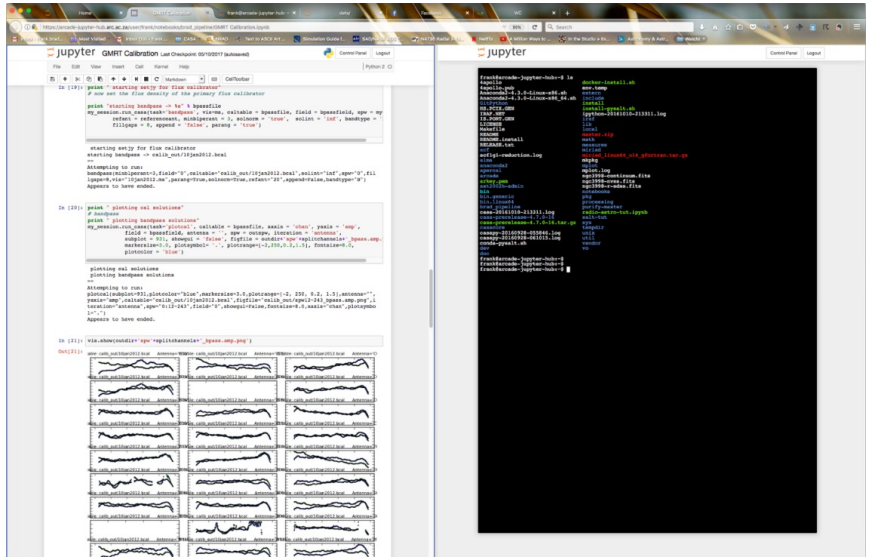
  - Aim to have Pipeline available through EGI software distribution services
  - Also interested in having our visual analytics tool available

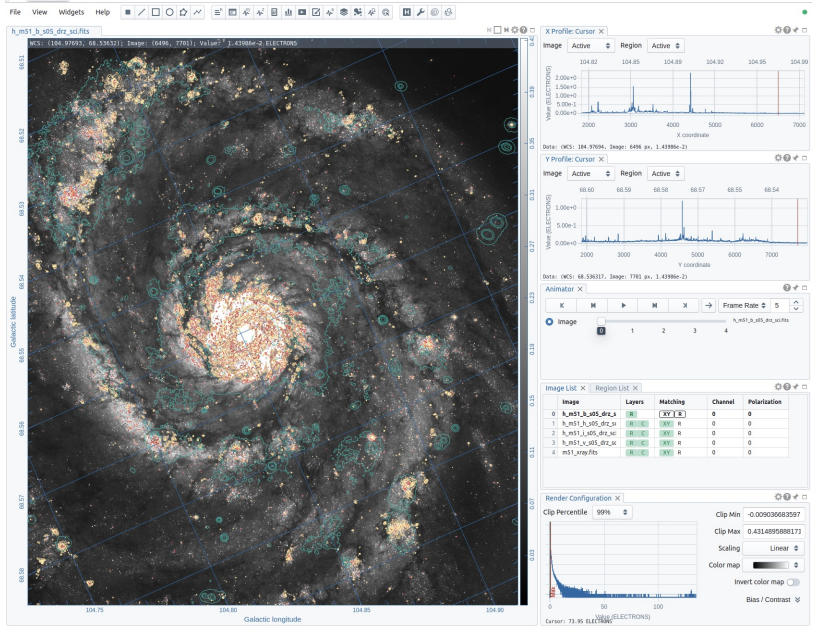


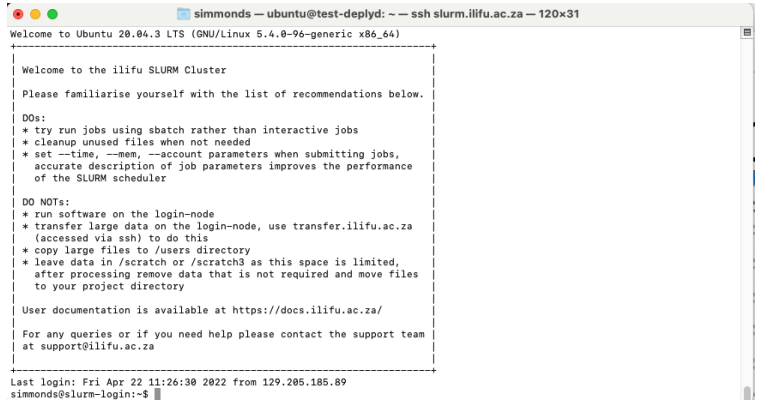


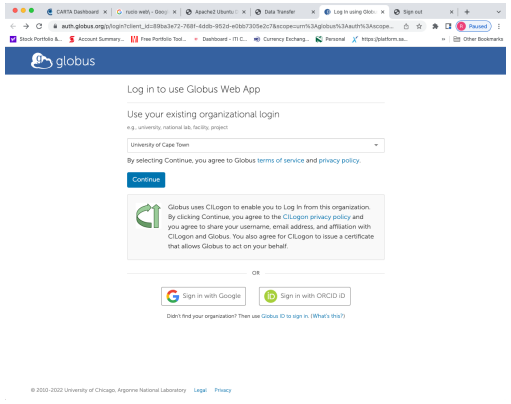
Tools deployed at ilifu

-









IDIA MeerKAT pipeline (processMeerKAT.py)

- Brad Frank, Jordan Collier, Srikrishna Sekhar, Russ Taylor
  - v2 released Sept 2022, used by LSPs (LADUMA, MIGHTEE, more)
- Parallelised package for HPC processing
  - Deployed at IDIA on OpenStack system ilifu
- Similar approach to two-round CASA-based crosscal, **but**
  - Fast, necessary for 32k MeerKAT data (~12 TB raw per 8 hours)
  - Full Stokes polarisation calibration with output polarisation cubes
  - Adds our own automated (PyBDSF) selfcal masking algorithm + AWP
- Builds and submits pipeline jobs to resource managers
  - Input measurement set, build / run your config file, request resources
  - Optionally insert your own scripts, specify containers and MPI wrappers
- Goal: T(cal) ~ T(obs); result: T(crosscal) << T(obs)

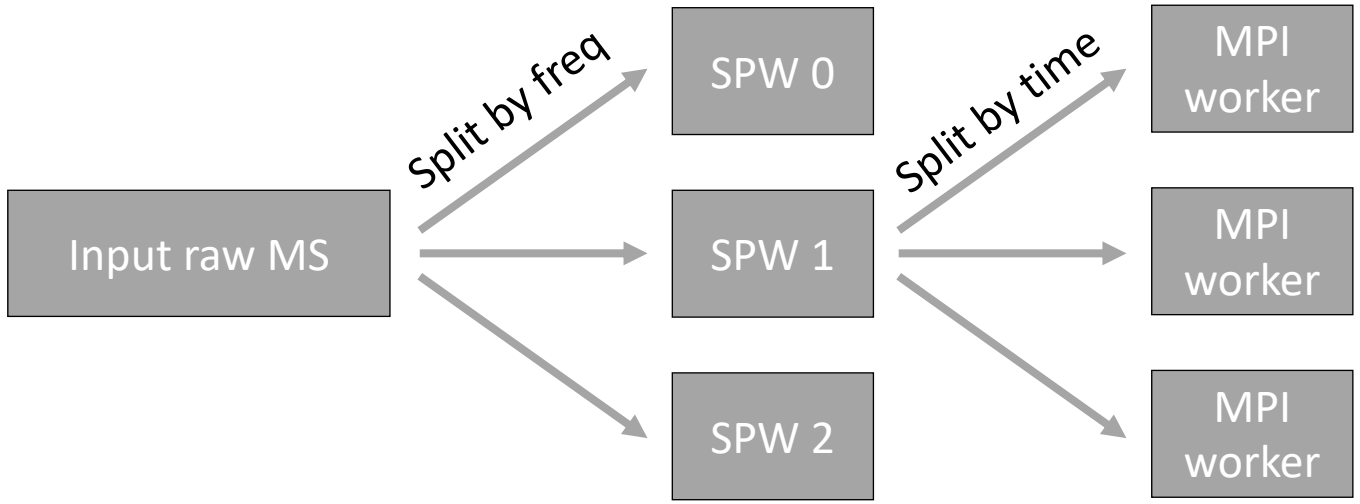
IDIA MeerKAT pipeline implementation

- Packaged as set of Python scripts
  - https://github.com/idia-astro/pipelines/releases/tag/V2.0
- These interact with singularity containers:
  - casa-6.5.0-modular.sif
    - Contains the modular CASA 6.5 install
  - kern7.simg
    - Contains the software packages provided by the Kern repository.
  - ASTRO-PY3.simg
    - Commonly used source finding packages.

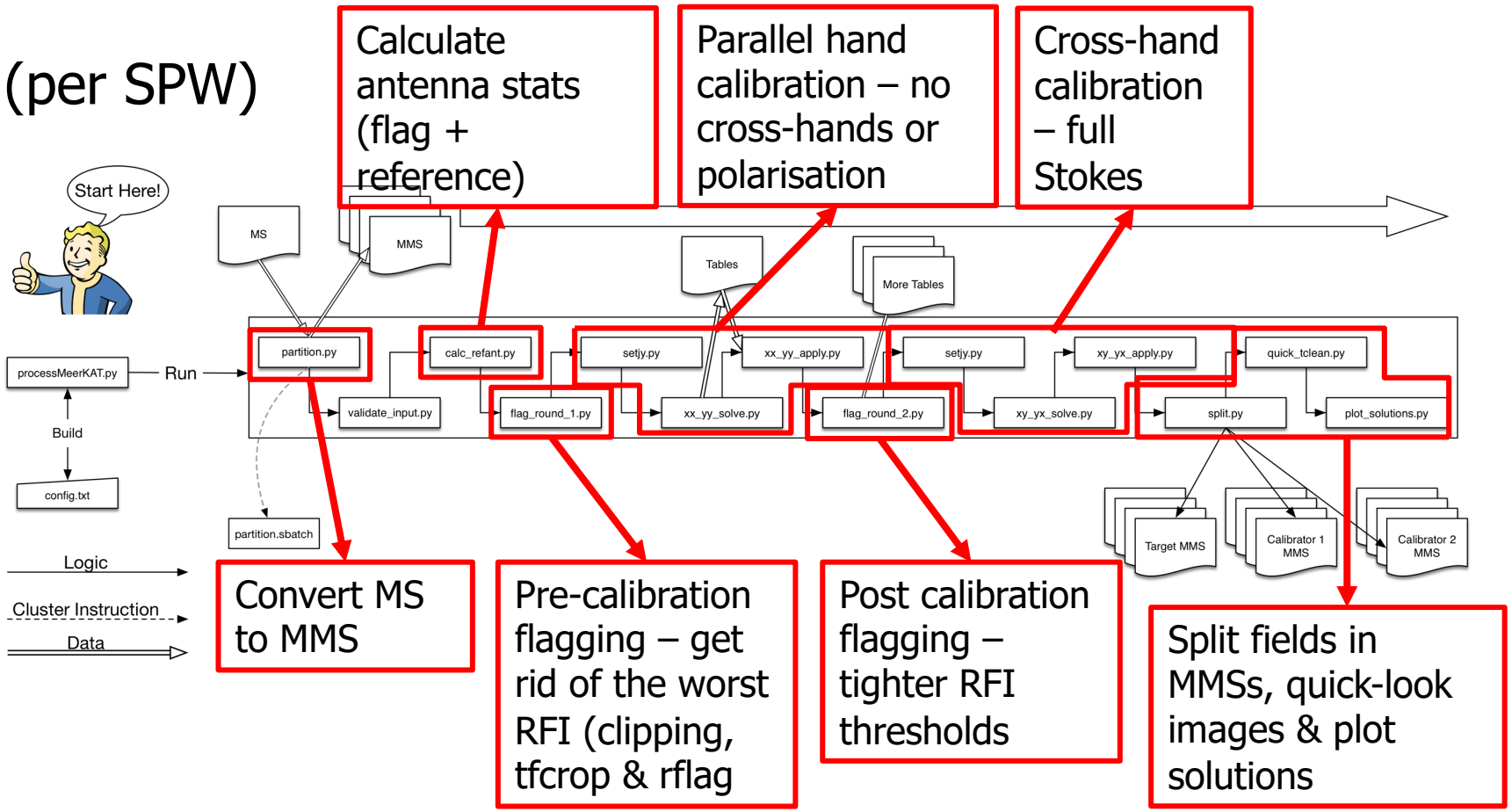
These can be run on basic HPC systems that have singularity installed

IDIA MeerKAT pipeline parallelism

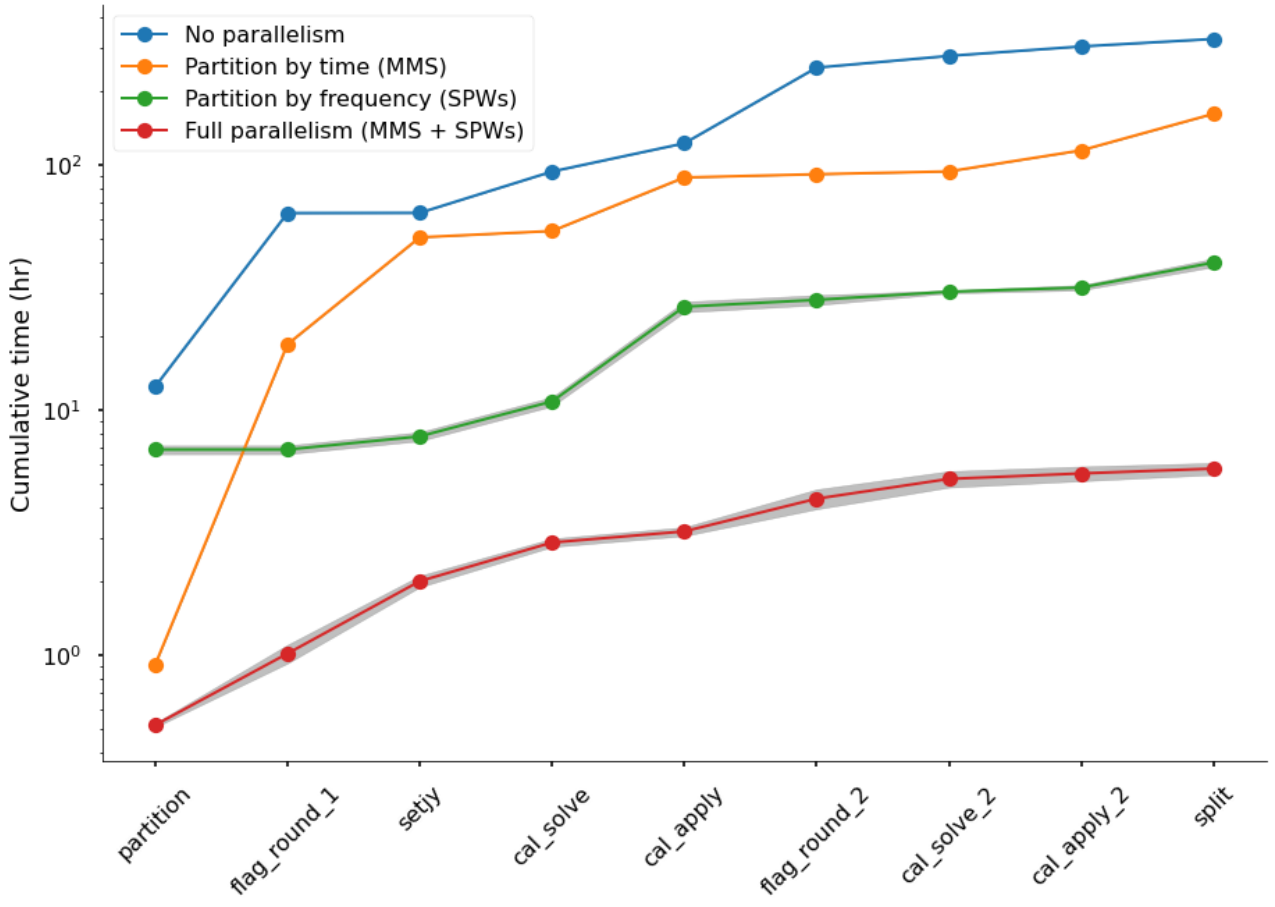
- Two parallelism schemes across cluster
  - Partition in frequency (SPW): IDIA pipeline scheme
  - Partition in time (scan): CASA multi-measurement set (MMS)



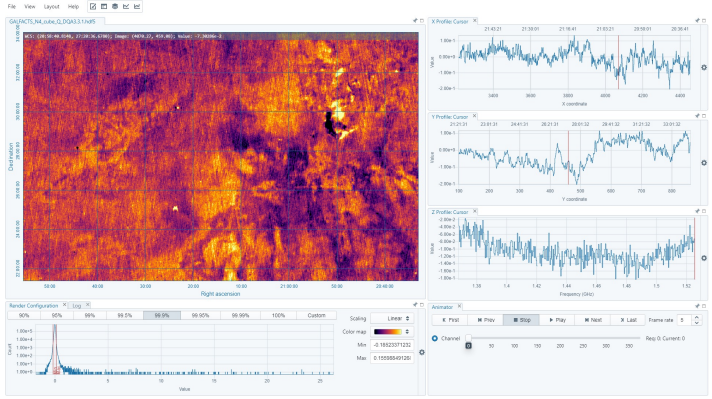
IDIA pipeline schematic



IDIA MeerKAT pipeline performance

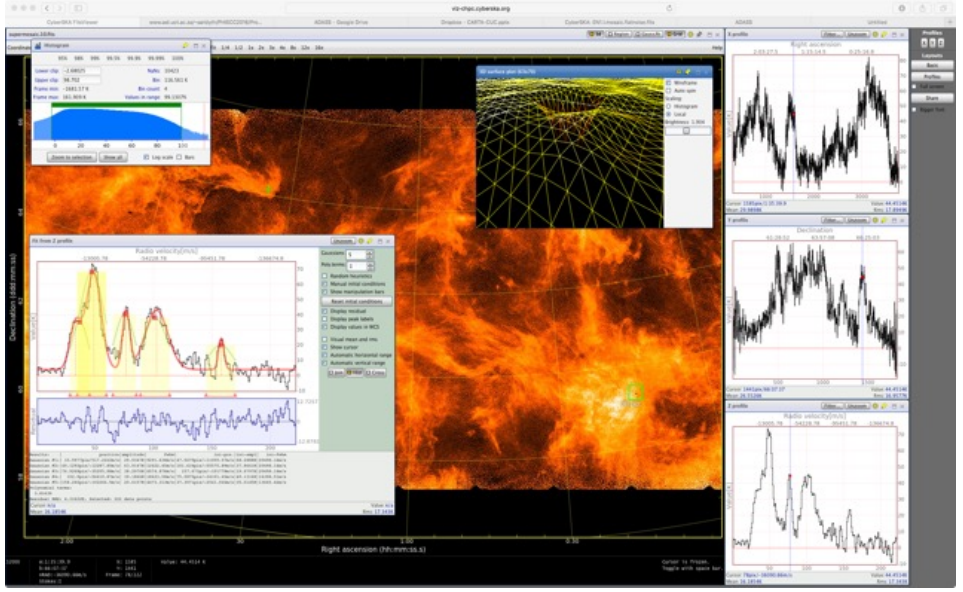


CARTA



- Cube Analysis and Rendering Tool for Astronomy (CARTA)
  - IDIA (South Africa) – NRAO (US) – ASIAA (Taiwan) – UofA (Canada)
  - A remote visual analytic tool for large image cubes
  - Performs latency hiding by doing large scale processing on remote servers / the Cloud, and interactive processing in a web client on the user's local computer
  - Supports many image formats: FITS, CASA, Miriad, and HDF5
  - v3.0 released August 2022, available here: https://cartavis.org
  - Deployed at ALMA Regional Centres and servers around the world
  - Being explored as a system to use at the SKA Regional Centres

CARTA Deployment

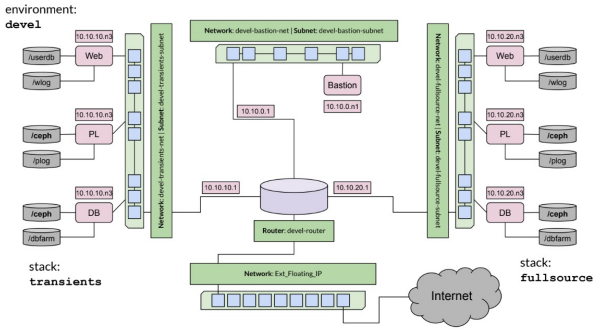


- CARTA can be run in serval ways. The main types:
  - Site Deployment Mode (SDM)
  - User Deployment Mode (UDM)
- SDM is preferred for integrated web environments since it integrates will with tools such as Jupyter Hub and provides more interesting options for user collaboration
- Works fine on IaaS system such as OpenStack, since users have control of network services
  - Not clear how to setup correct reverse proxies without this on systems where users don't have root access
- SKA Regional Centre testing sites that don't have OpenStack are tending to use Docker / Kubernetes but these need root access

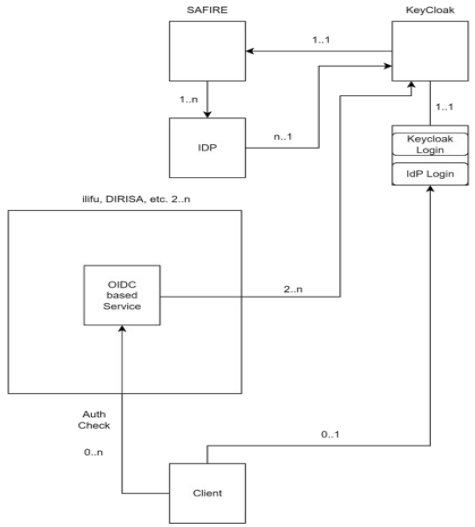
- This highlights that while software such as the IDIA pipeline code can be run on most clusters access to IaaS clouds makes deployment much easier for other software.

Specialized environment example

- Real-time processing for MeerLICHT performed in environment run by the research team
- Consists of VMs and virtualized networking components that run their workflows, with much image processing performed overnight
- Image data transferred from Sutherland to UCT/ilifu as it is captured by MeerLICHT
- Processing performed to search for transients
- Alerts issued and corresponding MeerKAT data examined



Cloud Federation



- Have federated ilifu with EGI cloud as part of EGI-ACE that simplifies gaining access to cloud resources
- Starting work with DIRISA on federating ZA systems
  - Start with making sure all web-based systems setup to use EduGain/SAFIRE federation
  - Then plan to add role-based authorization system
  - Experience gained withing with EGI key to this activity
- South African Open Science Cloud (SAOSC)
  - In planning with guidance from EOSC
  - Infrastructure plans include federation of software tools and repositories
- African Open Science Platform (AOSP)
  - About to start deploying systems.
- Global Open Science Cloud (GOSC)
  - Focusing on how to link regional OSC activities.

EUROPEAN OPEN SCIENCE CLOUD

Summary + Future

- The OpenStack cluster hosted at UCT, has been in production since 2015 which supports large community of radio astronomy researchers
- Have much better understanding of requirements than when we started
- Have developed Pipeline and Visual Analytics system aimed at next generation radio telescopes
- Aim to work with EGI to simplify access to these tools on EGI resources as agreed to has part of the EGI-ACE project
- Working with EGI has provided us with access to best-in-class federation tools, and experience of operating systems using these
- Working with SKA Regional Centre planners on what tools should be supported at their sites
- Starting work with DIRISA to provide federation of South African services


IDIA
Inter-University Institute
for Data Intensive Astronomy


ilifu