South Africa:
SKA Regional Centre Activity

Dr. Rob Simmonds
Associate Director
IDIA

IDIA Inter-University Institute for Data Intensive Astronomy

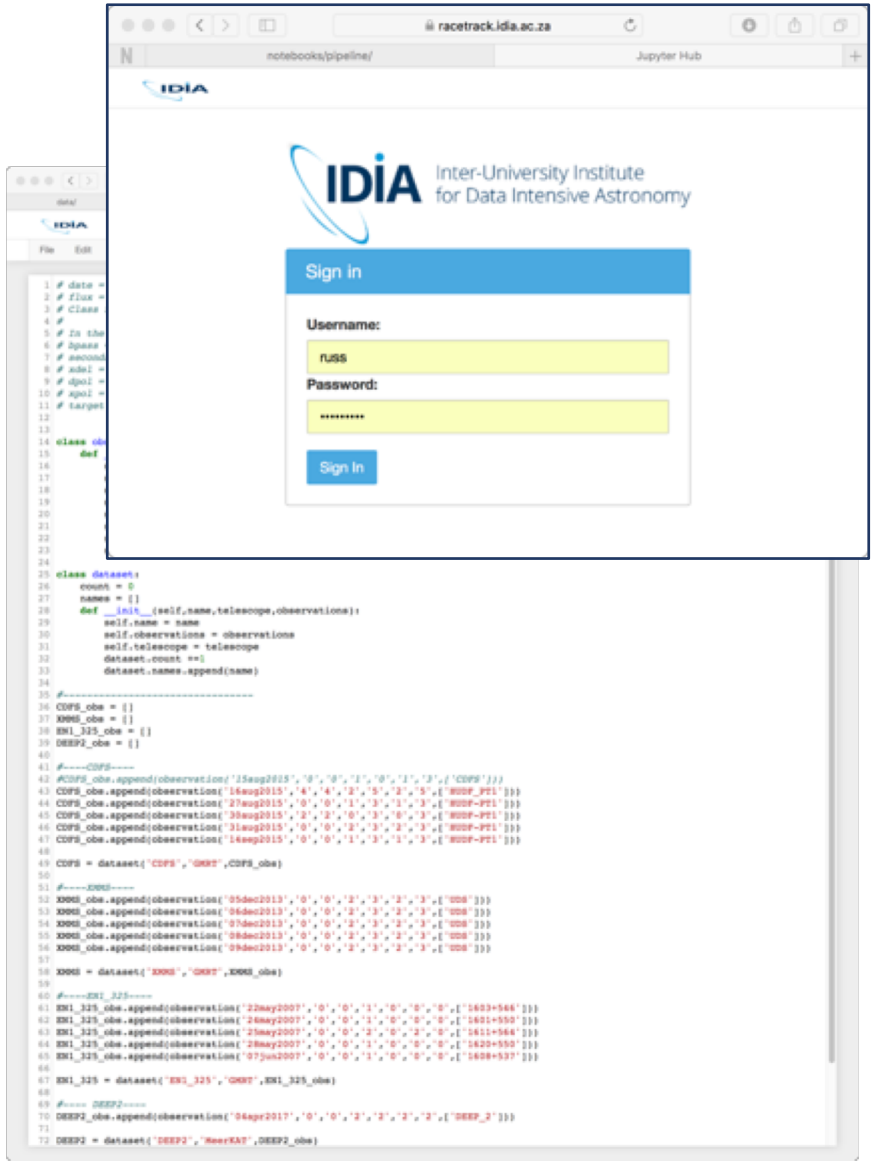
University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

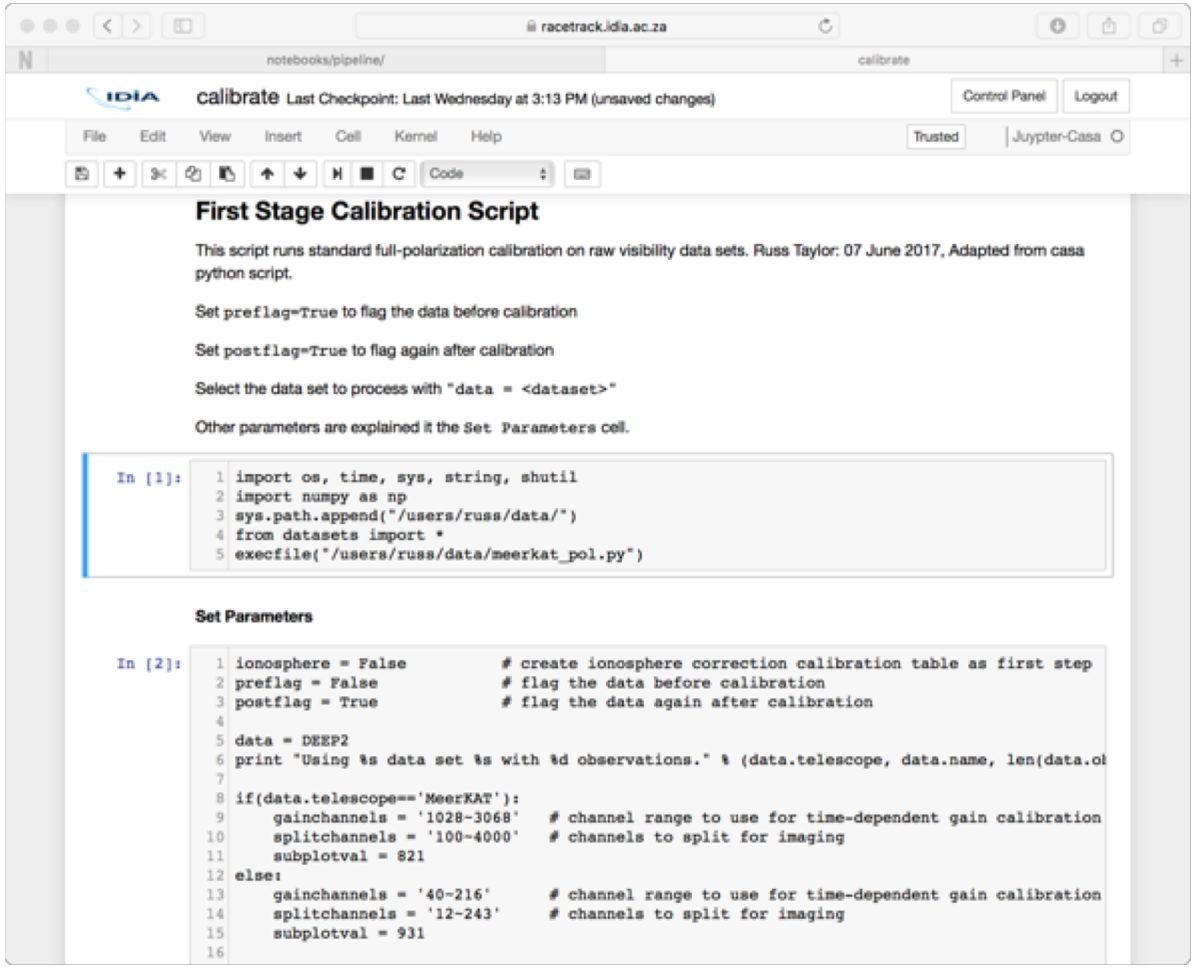
UNIVERSITY of the WESTERN CAPE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

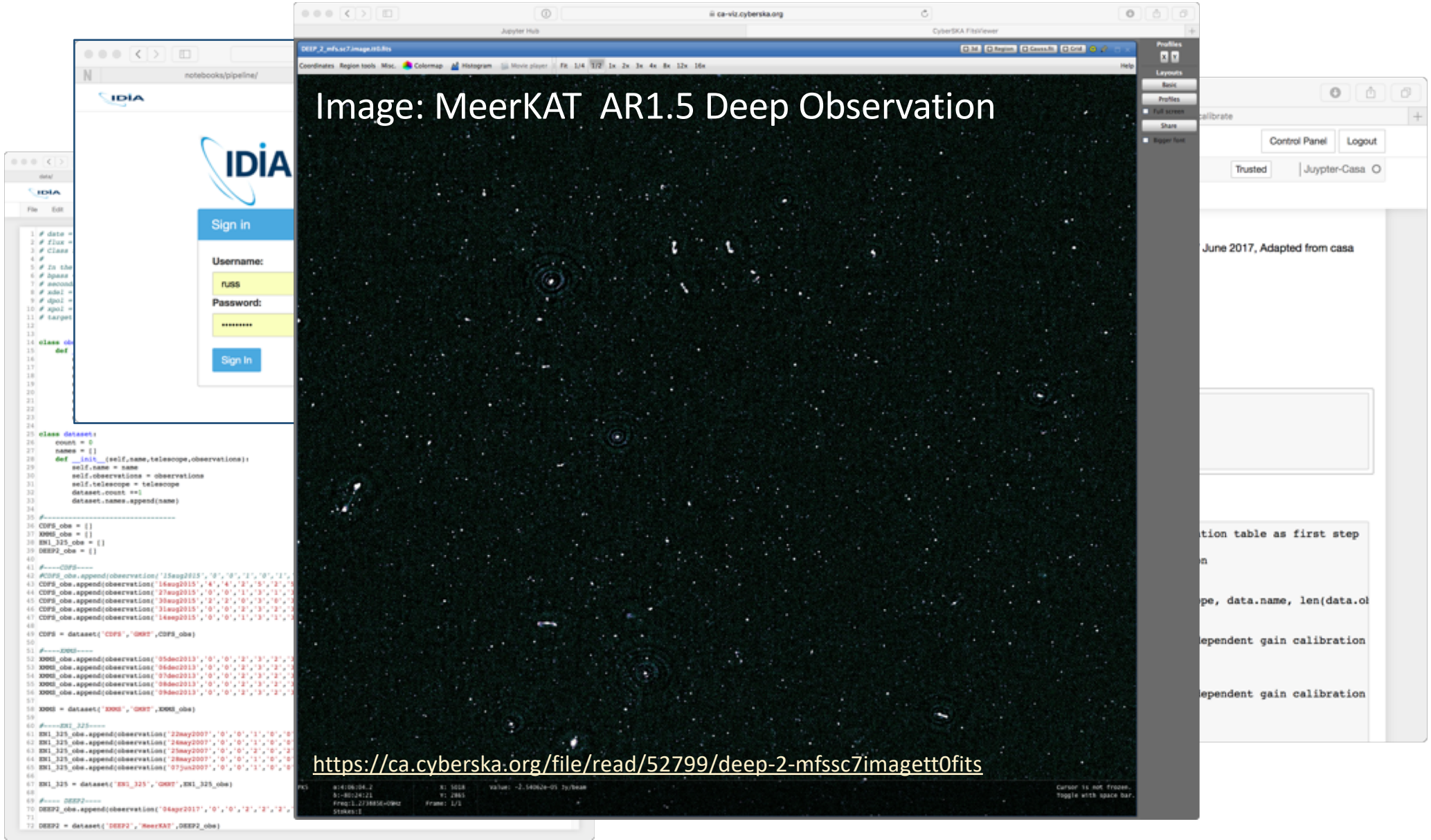
SAP®

Data Intensive Astronomy Cloud



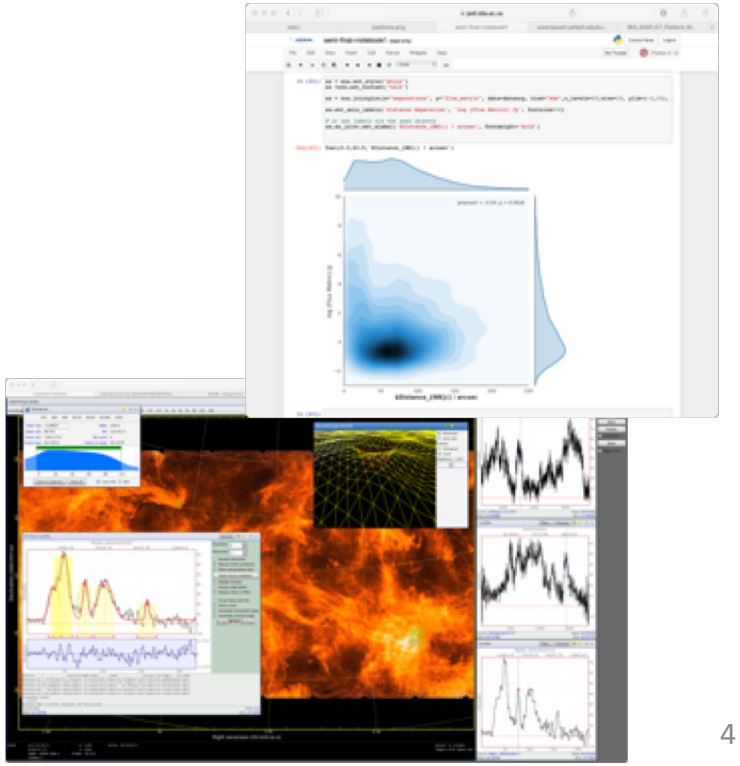


Data Intensive Astronomy Cloud



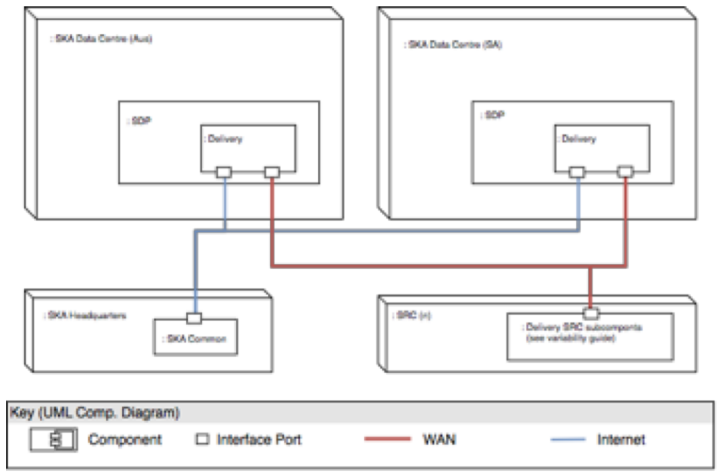
IDIA MeerKAT Large Science Projects

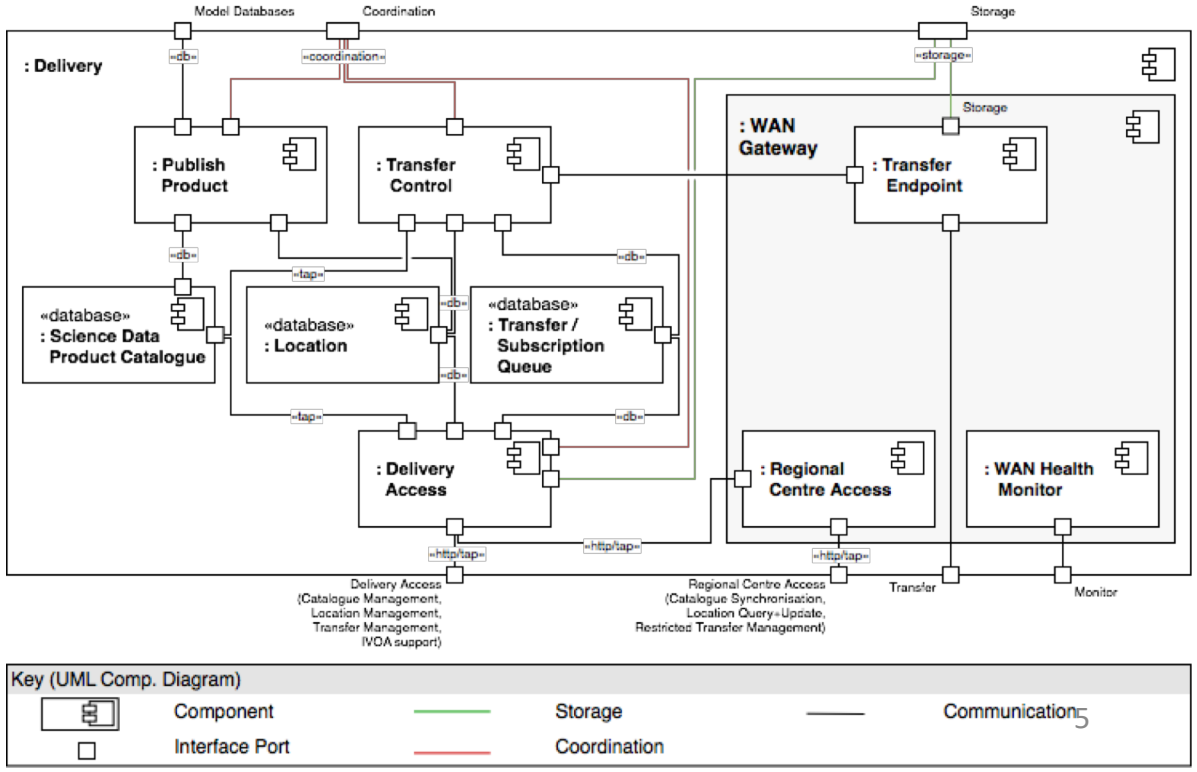
- **A Transient IDIA**
  - Pipelines processing and joint post-processing analytics for **ThunderKAT** radio and **MeerLICHT** optical observations
- **Pipeline Development for the MeerKAT Imaging Large Survey Projects**
  - Collaboration among **5 MeerKAT LSPs** to jointly develop pipeline processing of MeerKAT data
- **IDIA Visualization Toolkit: Converting Data Into Discoveries**
  - Development of visualization and visual analytics for **MeerKAT big image data sets** and use cases.
- **HIPPO: HELP-IDIA Panchromatic Project**
  - Multi-wavelength data fusion and analysis
  - Machine learning for classification from multi-wavelength data
- **Data Intensive Astronomy with LADUMA**
  - analytics and simulations for **LADUMA** HI science
- **How do Galaxies Form and Evolve**
  - Analytics and simulations for **MONGHOOSE** study of nearby galaxies
- **HI Intensity Mapping**
  - **MeerKLASS** preparatory studies
- **Very Long Baseline Interferometry**
  - Calibration, imaging and analytics of VLBI data sets
- **Open time science projects**
  - E.g. **MHISHAPS, VELA**,…



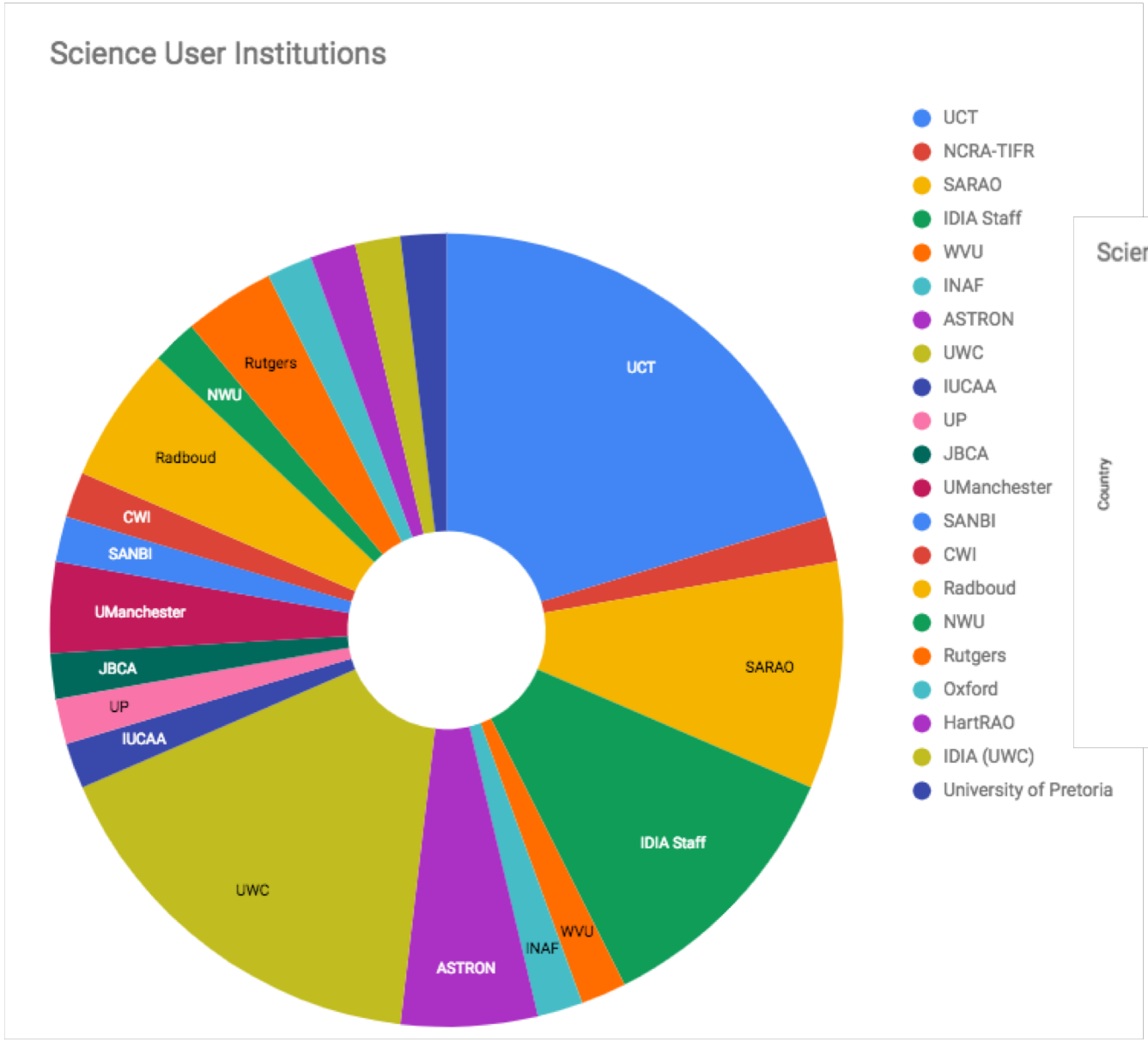
SDP Delivery Design

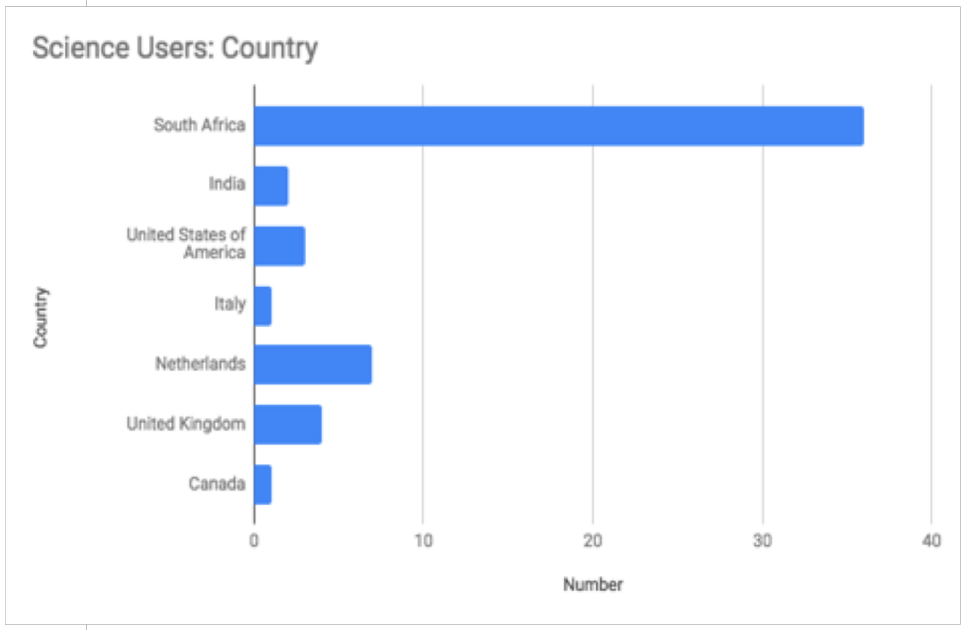
- IDIA leading SKA Data delivery design
  - Group members from IDIA, CADC, ASTRON and IAA
- CDR passed in Jan 2019
- Currently completing OAR updates to documents that are due at end of March.
- Parts being prototyped on IDIA cluster and supporting MeerKAT LSPs
  - Data delivery in place
    - Transfers from SARAO node, IDIA and ASTRON





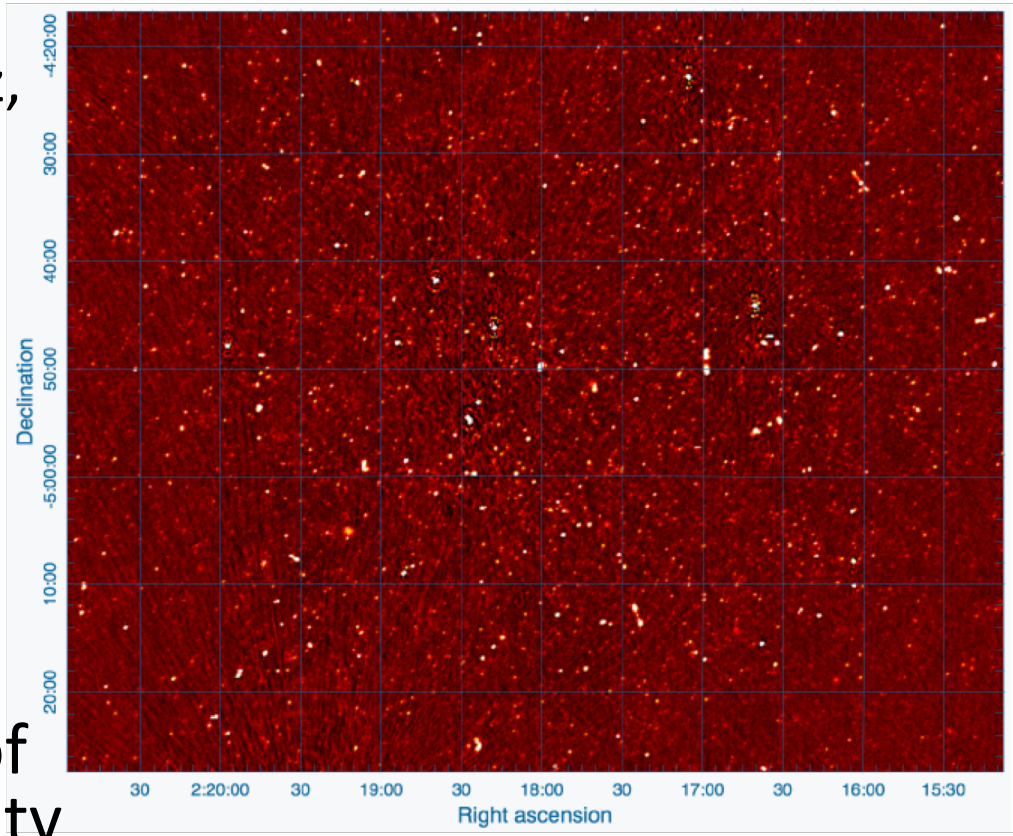
IDIA Cloud MeerKAT LSP Users





Data

- Example: MIGHTEE data
  - Currently 55-dishes, 4096 channels over 856MHz, ~6hr duration
    - 1.5-2.0 TB datasets
  - Soon moving to 32k channels
    - > 40 TB datasets
- Looking at different models for processing
  - Initially all visibility processing at IDIA
  - Aim to move initial visibility processing to SARAO during this year
- IDIA starting to receive data that is not part of the LSPs associated with targets of opportunity

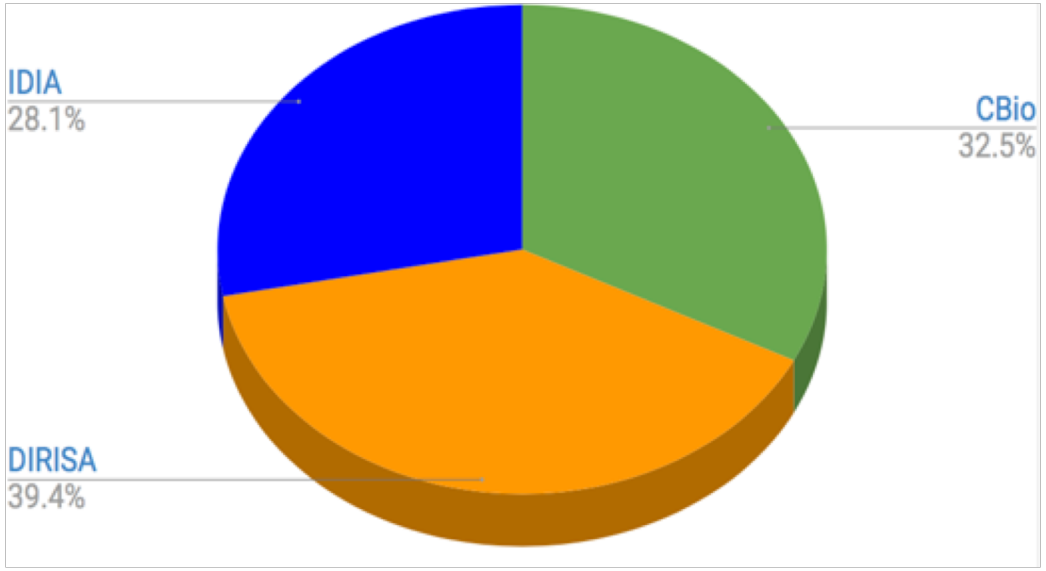


ILIFU: Tier 2 Data Intensive Research Facility

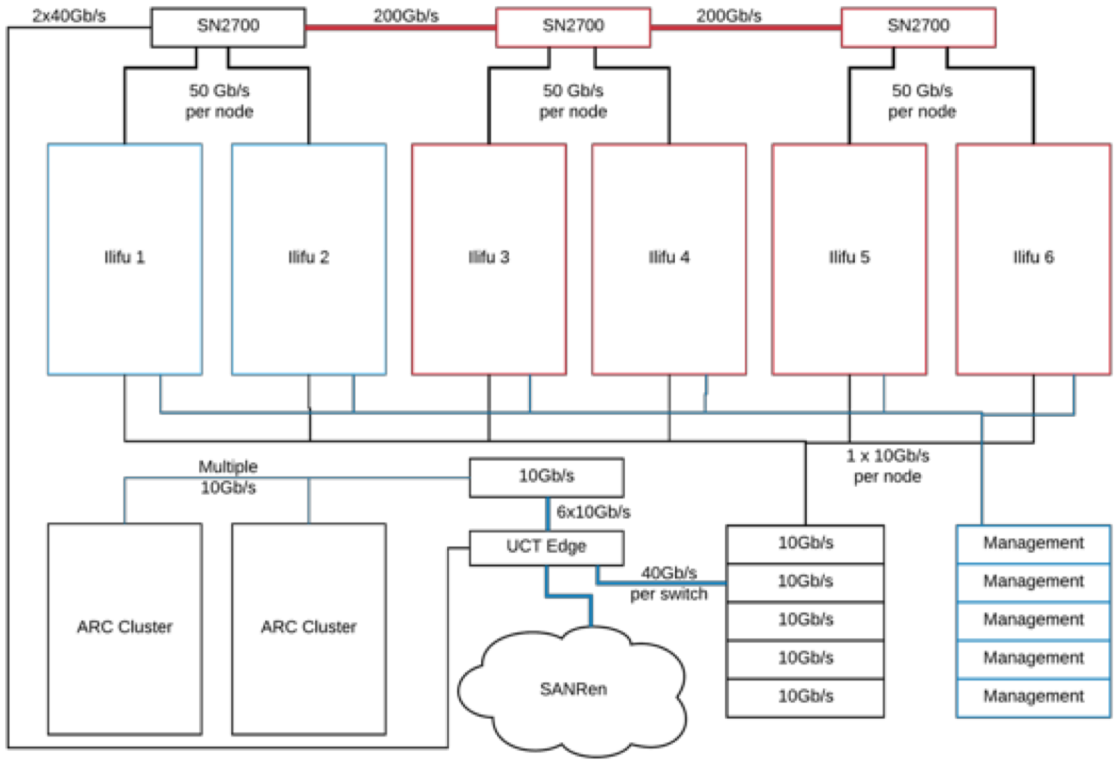
Joint investment DST/DIRISA, IDIA, Computational Biology (NIH)

- Astronomy (IDIA, SARAO)
  - Data Intensive Astronomy with priority on MeerKAT Large Survey Programs
  - Precursor SKA Regional Science Centre
- Data Intensive Bioinformatics
  - Tuberculosis Surveillance in Africa (UWC)
  - Imputation service for African human genetics (UCT)
  - Omics for Precision Medicine (SU)
- Research Data Management (CPUT)
- South African Data Intensive Research Cloud federation with T1 and T3 infrastructure

ILIFU Cloud



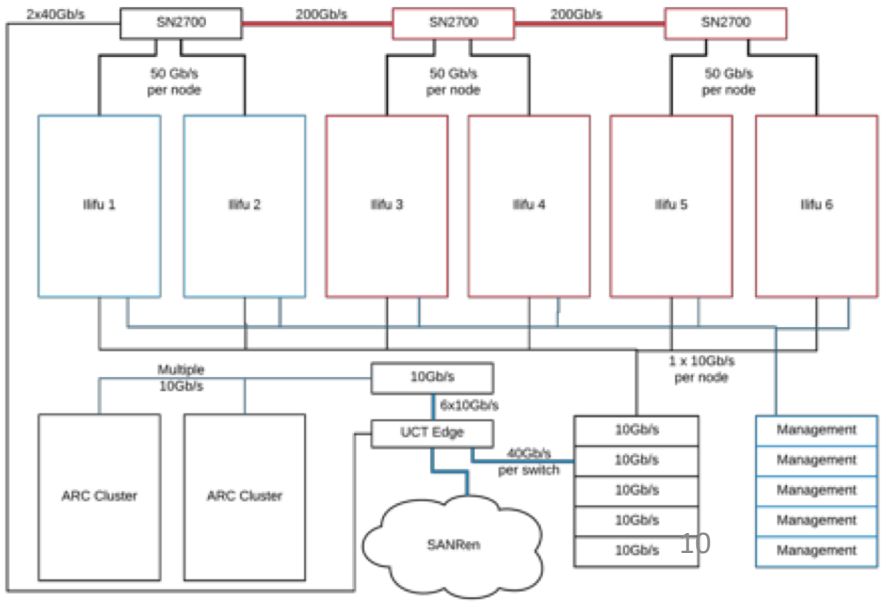
Infrastructure contributions 2019



- System scheduled using fair share aided with appropriate limits to guild this
- IaaS system managed by OpenStack
- 2/3 of DIRISA funds still to be spent – will make additional purchase this year

Ilifu cloud components (in place so far)

- ARC nodes provide spare management, 8 compute nodes and 300TB (usable) from CEPH targets
- ILIFU racks 1-2 provide 40 (Intel E5-2697A) compute nodes, 8 GPUs (Nvidia p100) and 0.5PB (usable) storage (BeeGFS)
- ILIFU 4-6 provide 80 (Intel gold 6142) compute nodes and over 2PB (raw) of disk storage (CEPH) and management nodes
- ILIFU 7 (not shown) provides 0.5PB of off site backup storage
- Currently connected to SANReN at 10 Gb/s
  - Will upgrade to 100Gb/s when SANReN core upgrade takes place
- Storage a mix of CEPH and BeeGFS, with Manila used for user level file-system provisioning

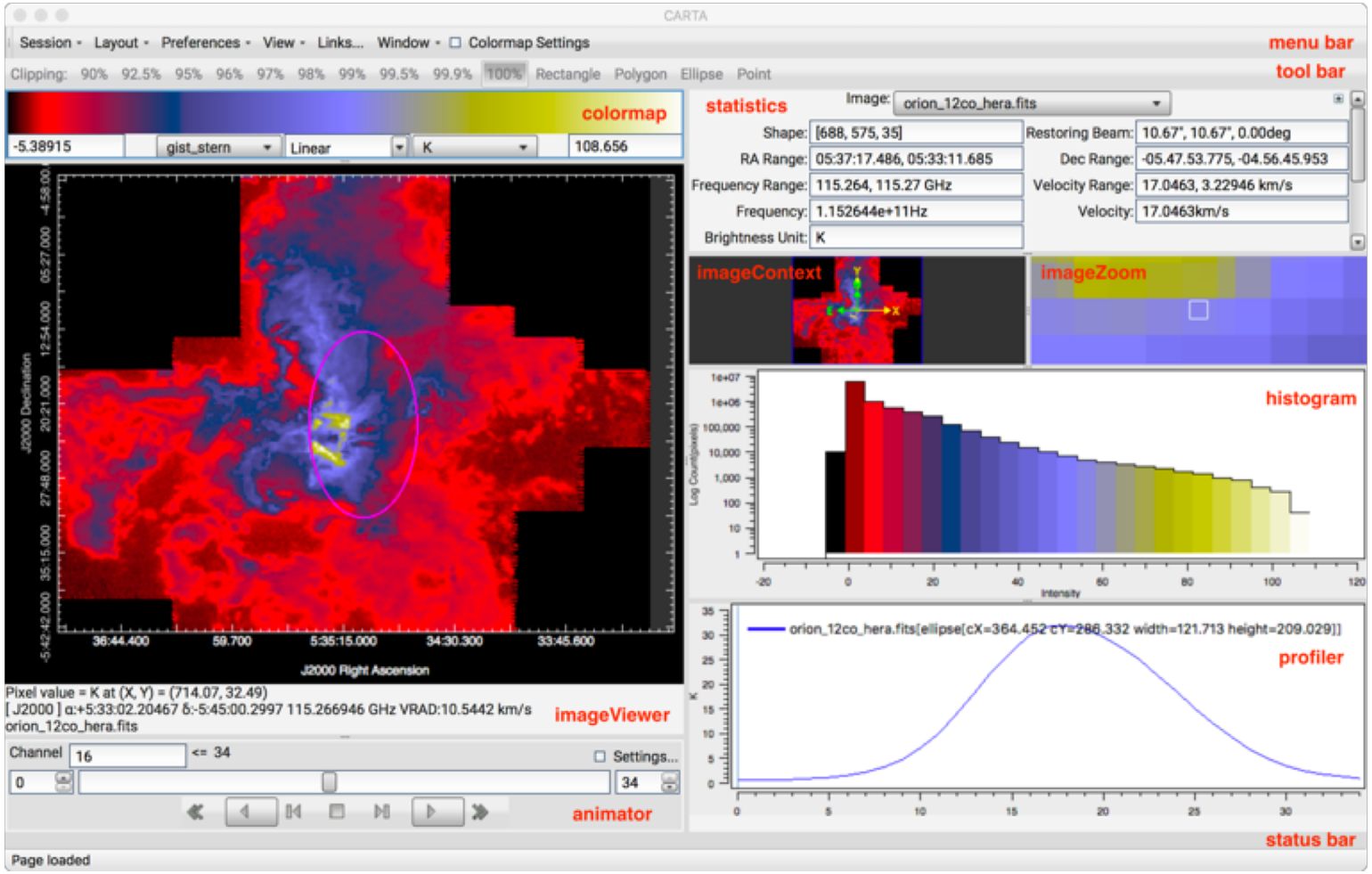


Pipelines

- IDIA pipeline framework being used for MeeKAT LSP processing released to community last week
- Presented in Singularity containers with:
  - CASA, drive-casa, python libraries, JupyterHub/JupyterLab Notebooks
- Will be combined with data transport system to provide automated archive to product execution
- Using elastically constructed cluster for batch and spawned lab instances

CARTA viewer

- Developing new viewer
  - NRAO collaboration
  - Will replace CASA and cyberSKA viewers
- Aim to scale visualisation and analytics to multi-terabyte cubes with remote viewing
- Now using HDF5 to better support parallel I/O
  - See ADASS 2018 paper on new schema
- Initial release at end of 2018. 1.1 due in April.



SRC operations planning

- Discussions going on between stake holders on how SA SRC will be operated
- Major stakeholders include:
  - DST and Meraka (CSIR) from government
  - SARAO (SKA-SA)
  - IDIA
- Documentation on how to distribute work between partners nears completion

Summary

- SDP has developed baseline architecture for delivering data to SRCs. Updated SRC interface definition document in progress but was not part of CDR submission.
- MeerKAT Regional Centre framework being developed in multi-partner collaboration and data arriving at IDIA daily.
- CARTA viewer development is ongoing with CARTA architecture update to unify GUI across platforms
- South Africa planning to have SRC in addition to SKA1 Mid Processing Centre; partner contribution plan nearing completion