CSIRO ASKAP Science Data Archive

Minh Huynh (CSIRO, ICRAR/UWA)
James Dempsey and the CASDA team
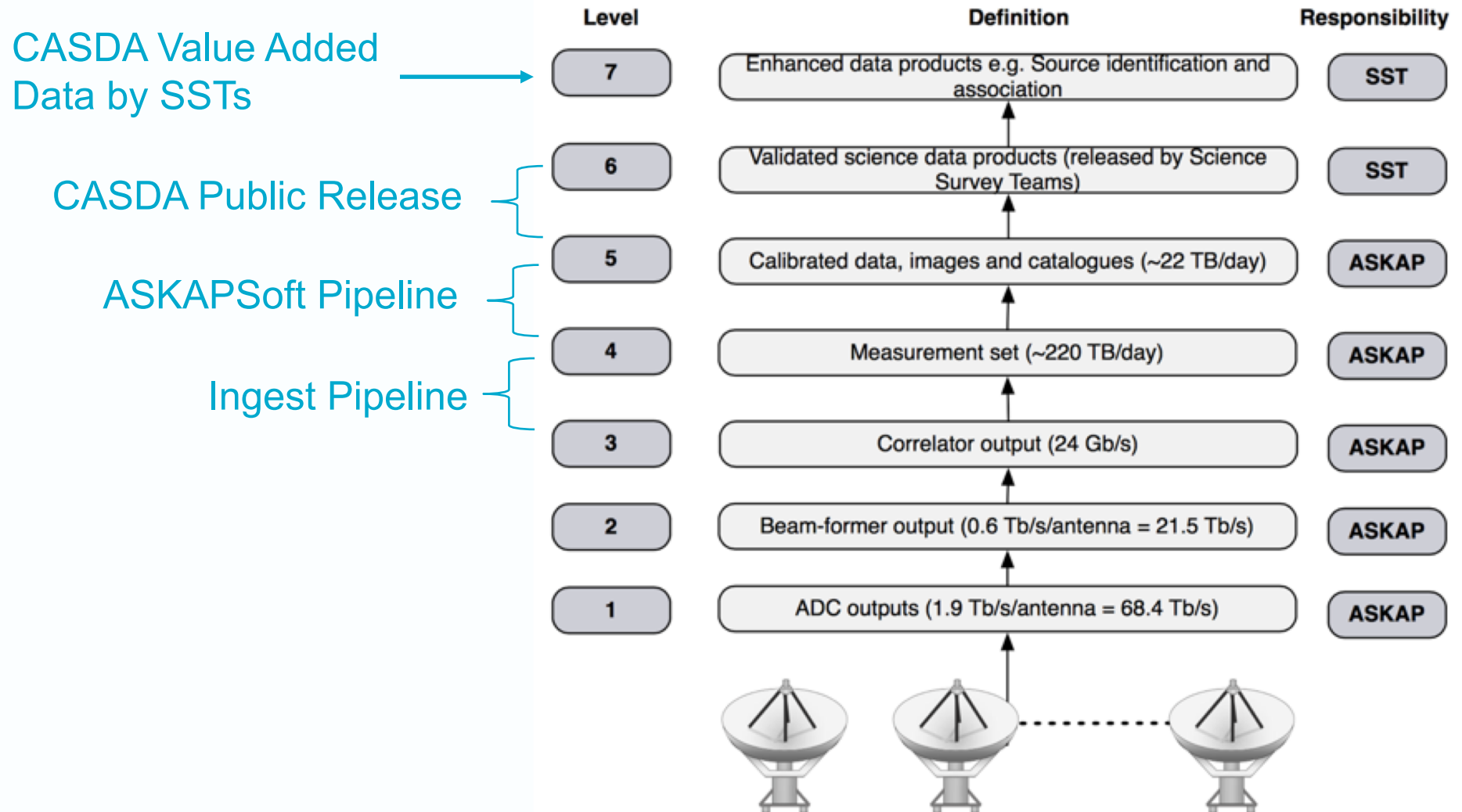
ASKAP and all its wonderful data

POSSUM May 2019
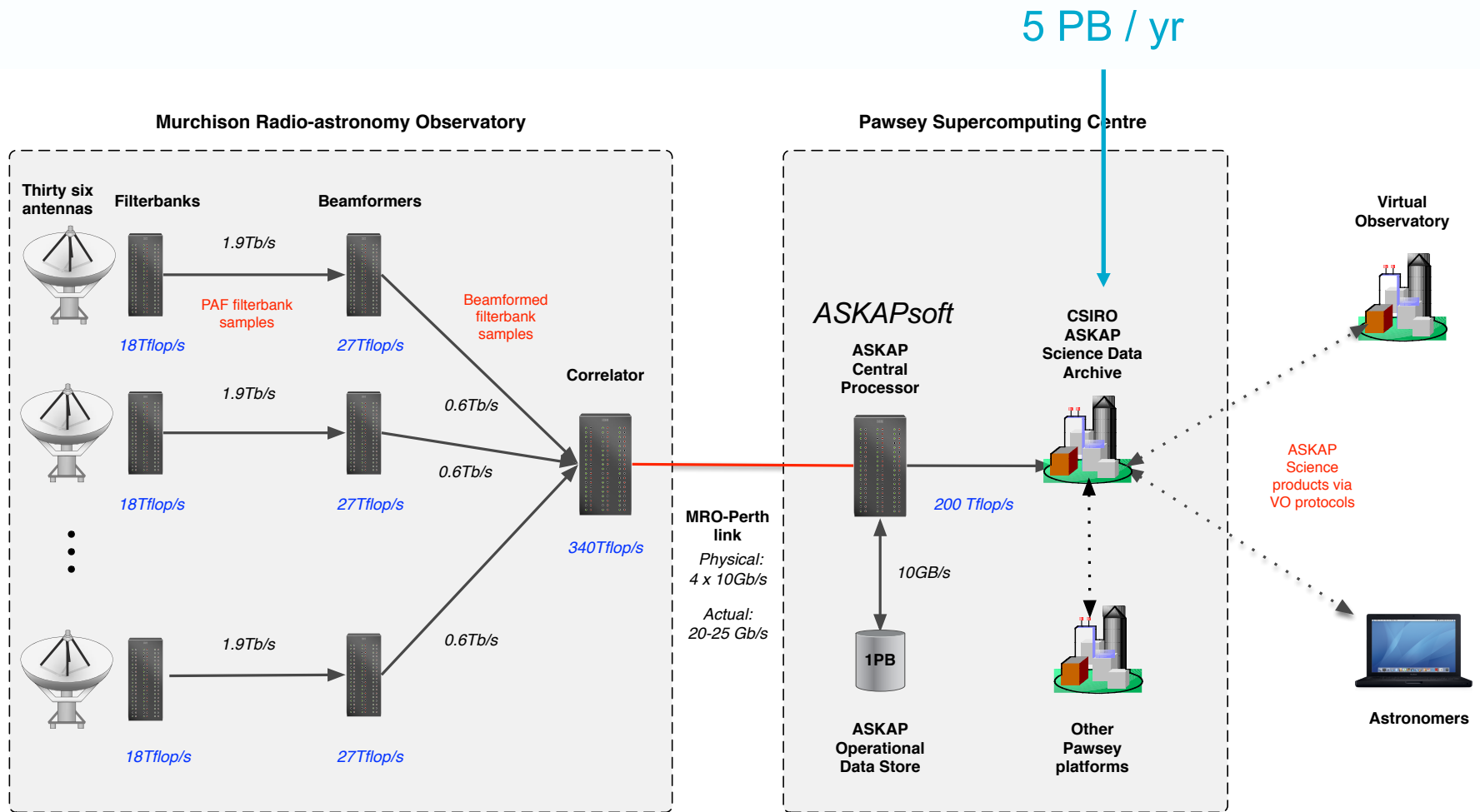
ASKAP and all its wonderful data levels

CASDA Value Added Data by SSTs

CASDA Public Release

ASKAPSoft Pipeline

Ingest Pipeline

Level	Definition	Responsibility
7	Enhanced data products e.g. Source identification and association	SST
6	Validated science data products (released by Science Survey Teams)	SST
5	Calibrated data, images and catalogues (~22 TB/day)	ASKAP
4	Measurement set (~220 TB/day)	ASKAP
3	Correlator output (24 Gb/s)	ASKAP
2	Beam-former output (0.6 Tb/s/antenna = 21.5 Tb/s)	ASKAP
1	ADC outputs (1.9 Tb/s/antenna = 68.4 Tb/s)	ASKAP

CSIRO

CASDA Deployment

	PAWSEY SUPERCOMPUTING CENTRE	CANBERRA
Location	Perth, WA	Canberra, ACT
Functions	• Deposit ASKAP data products • Data access • Virtual Observatory	• Interactive search • Collections • Authentication/Authorisation • Data Validation/Release • User interface
Facilities (Prod)	• 2 dedicated servers • 2 x 10 PB tape library (shared) • Lustre filesystem (shared) • 256 TB disk (initial allocation)	• 5 virtual machines • Integrated with CSIRO Data Access Portal
Facilities (Dev, Test, AT)	• 5 dedicated servers • Lustre filesystem (shared) • 384 TB disk (initial allocation)	• 15 virtual machines • Integrated with CSIRO Data Access Portal

CSIRO

CASDA Functionality

CASDA provides long-term archiving of and access to the large scientific datasets taken by ASKAP.

Functionality includes:

- Long term storage of ASKAP science data products

- Searches and data access via web (CSIRO Data Access Portal) and Virtual Observatory services

- Validation of ASKAP observations

- Upload of value-added science catalogues and image cubes by users

- Digital Object Identification (DOI) for all datasets

- Archive administration, inc. team member access to unreleased data

Approach:

- Agile Scrum software development

CASDA Data Products

CASDA stores the following data products:

Calibrated visibilities

- Archived long-term for continuum data only (frequency averaged)

Images and Image cubes, Spectra and Moment Maps

- Basic image products produced by the pipelines
- Derived image products, such as extracted spectra for detected sources, moment maps, etc

Catalogues

- Detected sources and their parameters

Project and Observation information

Data quality information

- Used to evaluate the quality of an observation
- Evaluation metric files

CSIRO

CASDA and Virtual Observatory

CASDA work is continuing to build CSIRO's skills, knowledge and international reputation in VO protocols and services.

Main Protocols

- Table Access Protocol (TAP)
- Simple Cone Search
- Simple Image Access Protocol (SIA v2)
- Server Operations for Data Access (SODA)
- Simple Spectral Access Protocol (SSAP)

Automated data discovery and retrieval via scripts

Also used as a service layer:

- Cutouts
- File access
- Catalogue Retrieval

CSIRO

Current Datasets in CASDA

- First ASKAP Early Science: ASKAP12 Continuum reduction of NGC 7232 and surrounding area, July 2017

- HIPASS (Parkes, Koribalski et al.) and SGPS (ATCA + Parkes, McClure-Griffiths et al.) HI spectral line cubes

- BETA datasets: PKS2252-089 HI absorption (Allison et al.), Tucana continuum image (Heywood et al.) …

- WALLABY Early Science Spectral Line Cubes (T. Reynolds et al., K. Lee-Waddell et al., Level 7 upload)

- EMU Early Science Cosmology Fields (10 fields)

- ASKAP 36 dish imaging of GAMA G23

- WALLABY Early Science NGC 7232 HI Spectral Line Cubes (2 SBIDs)
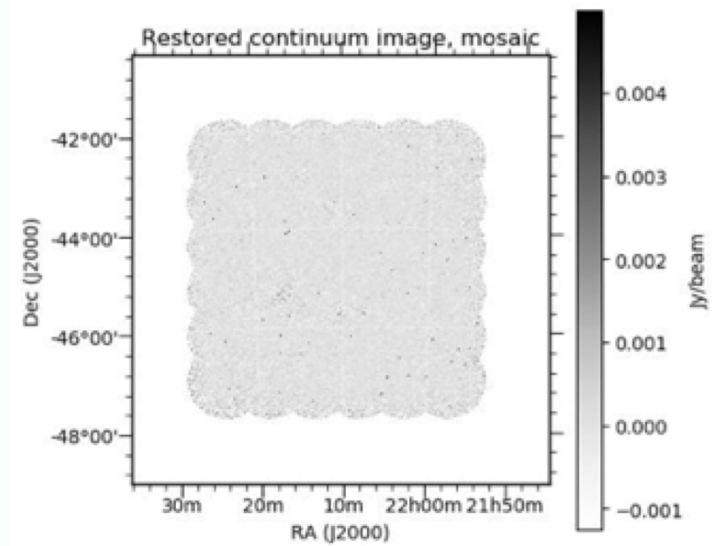
NGC 7232 continuum image

CSIRO

Future Development

Agile Scrum development

CASDA Stage 3 has begun May 2019 – Jun 2019

- 3 sprints in May - June 2019, release at end of June

- Another 3 months in FY 2019/2020, dependent on funding

Highest priority items are:

- Enhanced VO (e.g. TAP uploads, authenticated TAP)

- DAP (UI) Improvements

- More script examples and astroquery/pyvo

- More emphasis on Level 7 data products (?)

- Analysis of consolidation of AT Online Archive (ATCA, Parkes UWL) into CASDA

POSSUM Use Case 1

Story: As CIRADA developers, we want to be able to automatically download polarization data cubes to CADC for level 7 processing.

- Already had discussion with S. Gaudet in Nov last year.

- Can TAP query against obscore and get list of cubes/catalogues and wget the datalinks

- Will work with CADC to get this 'pull' system working

CSIRO

POSSUM Use Case 2

Story: As POSSUM data scientists, we want to have simple and fast (in-browser?) access to validation products/metadata associated with level 5 and/or 6 data products.

- Currently the evaluation files include all ASKAPSoft imaging logs

  - Separating continuum QA pipeline output (Collier) and making it available as a separate download

  - Looking at making the QA pipeline html output available

CSIRO

POSSUM Use Case 3

Story: As POSSUM pipeline operators, we want to be able to automatically upload our level 7 data products back into CASDA and have them properly associated with the corresponding level 6 products from which they were derived.

- Currently, Level 7 uploads are via web UI

    - Looking into a scripted process

- Level 7 uploads are asked for provenance information

    - SBIDs should be listed by uploader during the Level 7 upload

CSIRO