CSIRO ASKAP Science Data Archive

CSIRO ASTRONOMY AND SPACE SCIENCE (CASS)
www.csiro.au

CSIRO ASKAP Science Data Archive (CASDA)

Talk outline

A: CASDA overview

B: Requirements and use cases

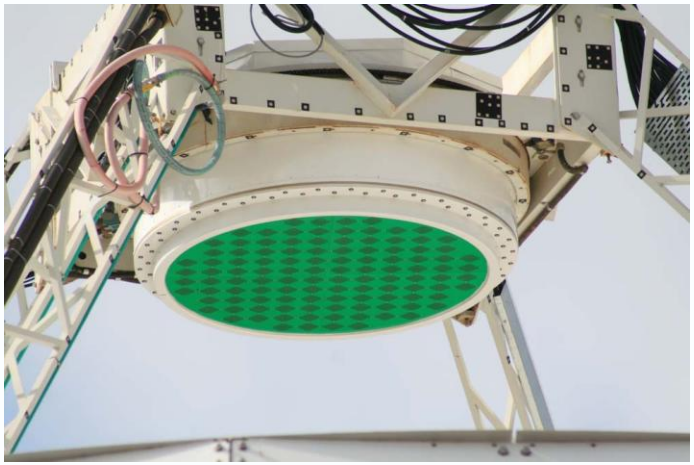
C: Data access and data volumes

D: Communications with science users

E: High performance computing

Australian SKA Pathfinder (ASKAP)

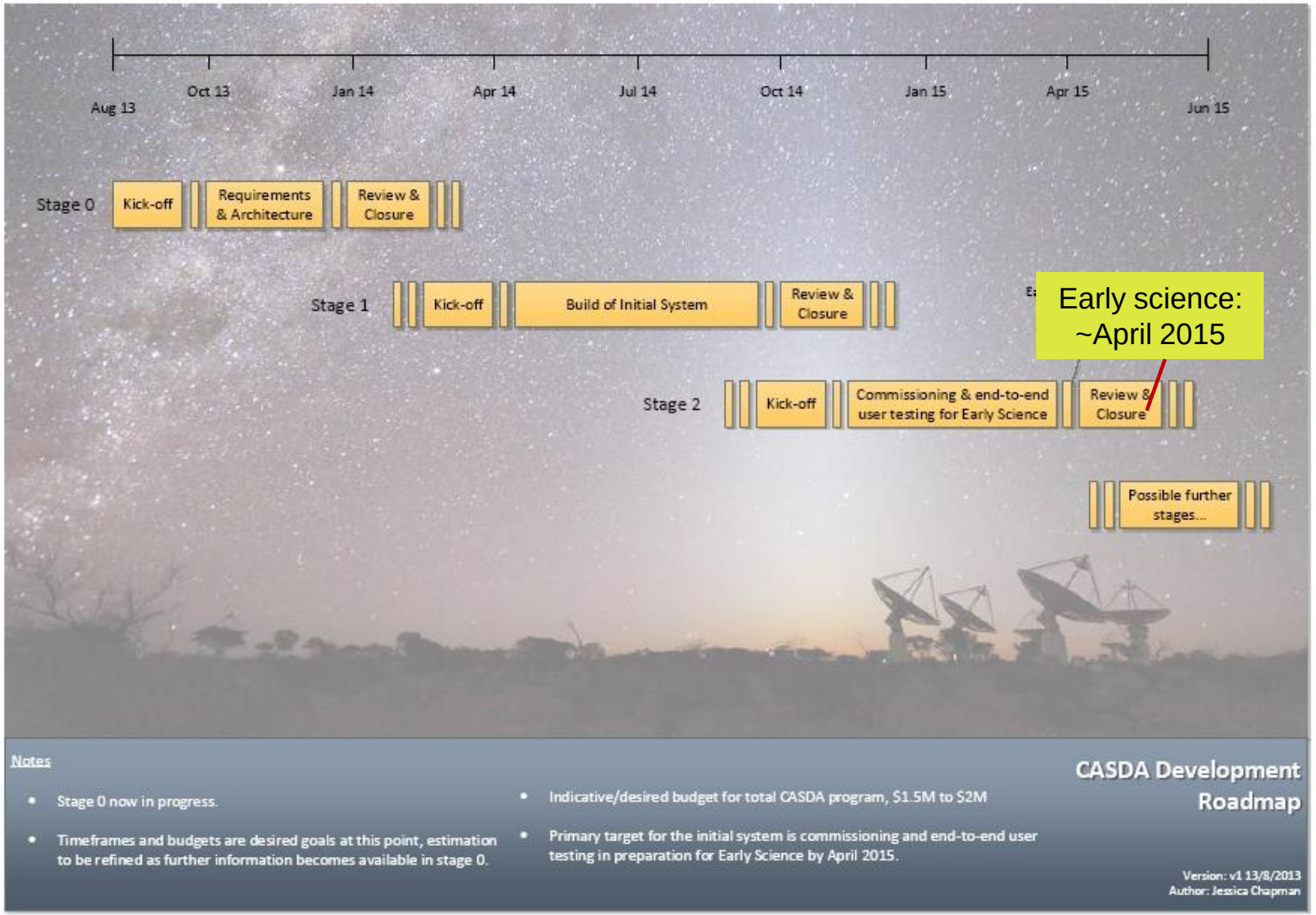
A fast survey instrument



- 36 × 12 m dishes
- Max baseline = 6 km
- Phased array feeds – 188 elements
- Digital beamforming
- 30 deg^2 FOV

- 700 – 1800 MHz
- 300 MHz Bandwidth
- 16,384 frequency channels

CASDA: Project roadmap



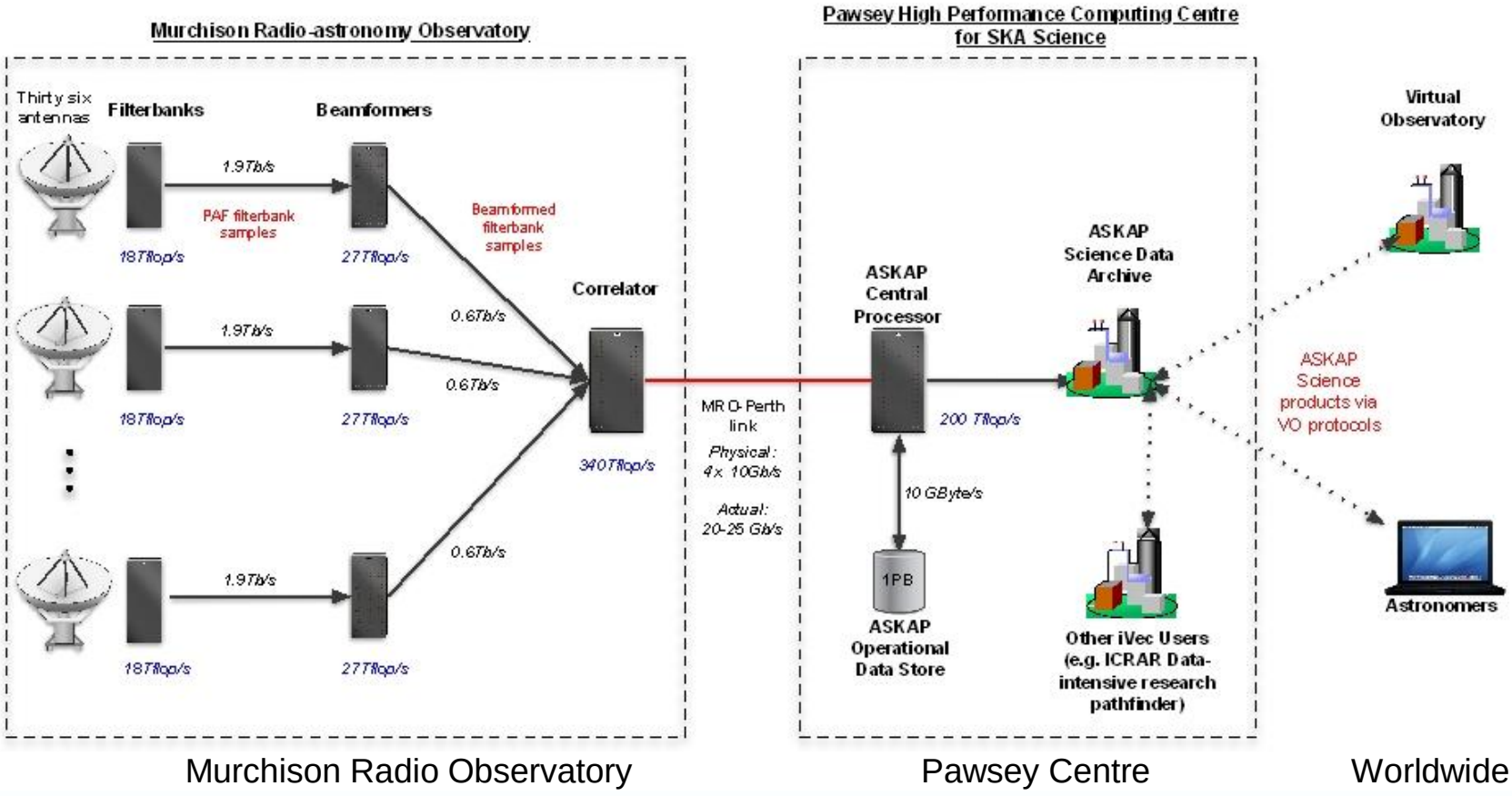
CASDA Stage 0 development team

IM&T:	CASS:
Euan Sangster (Project leadership) Angus Vickery (Project leadership) Dan Miller (Project Manager) James Dempsey (Project engineer) Jared Pritchard (Business analyst) Dave Morrison (Infrastructure specialist) Simon Bear (Software developer) Adam de Laine (Testing) Bradford Greer (Software/architecture)	Jessica Chapman (Project Leader) Ian Heywood (Project Scientist) Arkadi Kosmynin (Software developer) Matthew Whiting (Science data processing) Ben Humphreys (Science data processing)

Team uses an agile approach to project mgt and software development

Stage 0 activities

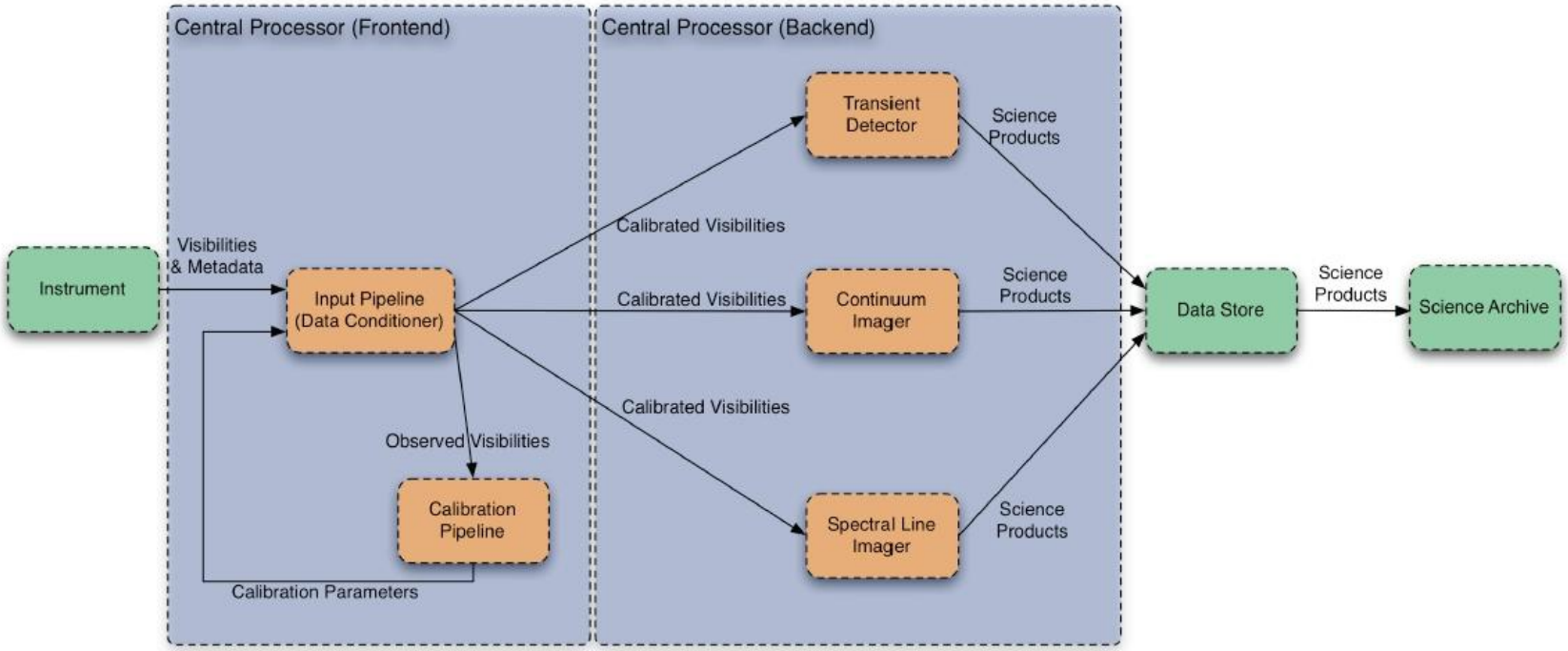
- Requirements description
- Use case model
- Workflows model
- System Architecture description
- Decisions on VO, database and middleware solutions
- Getting started at Pawsey Centre
- DB and VO prototypes for test purposes
- Preliminary operations and support model
- Preliminary Design Review



Central processor: Processes the raw visibilities and outputs science data products.

Science Data Archive Facility: Data storage and access to science data products

ASKAP Central Processor



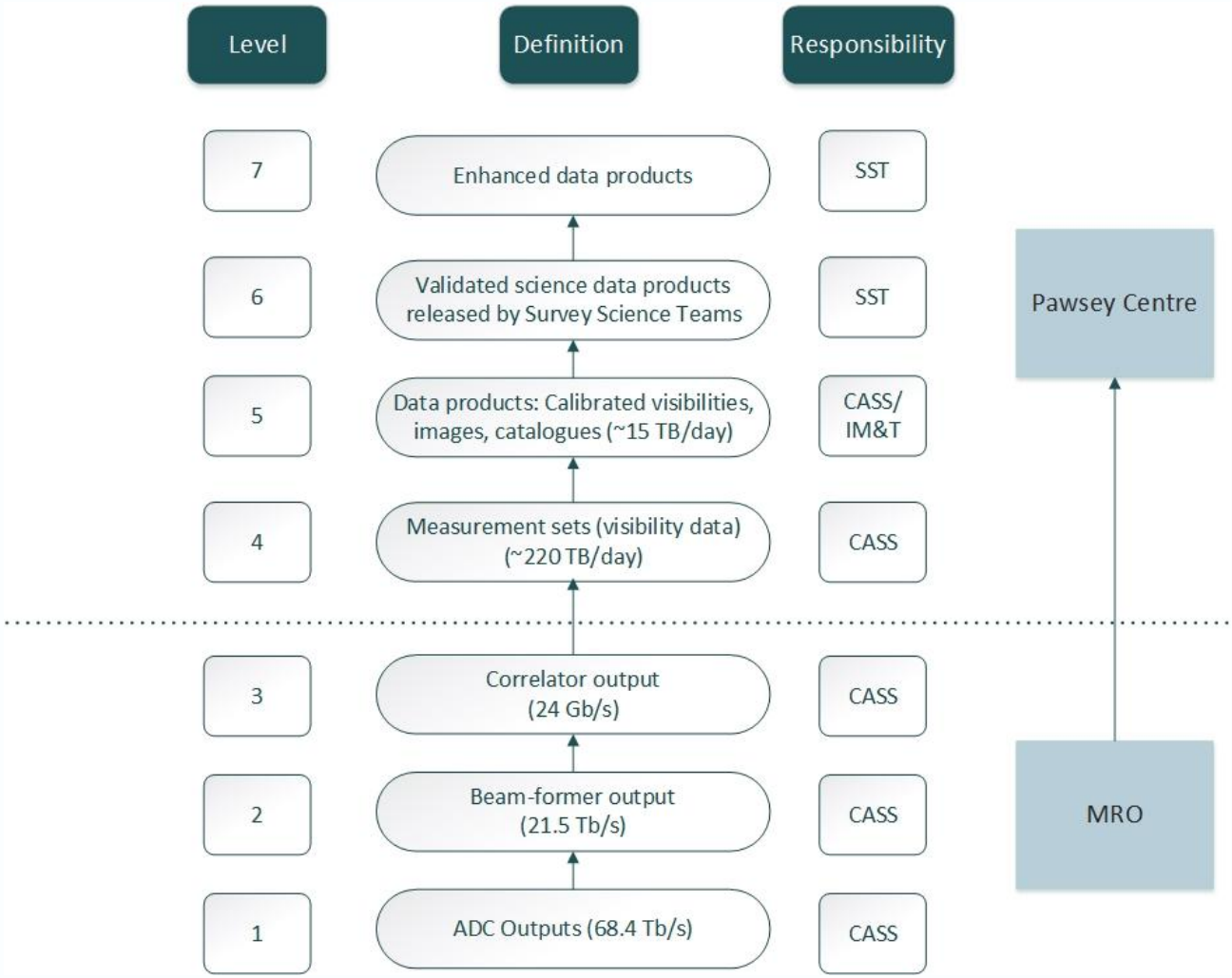
Data from same set of visibilities can be passed through three pipelines for **Transient, Continuum** and **Spectral Line** imaging.

In principle this allows the data for up to three different projects to be observed ‘commensally’.

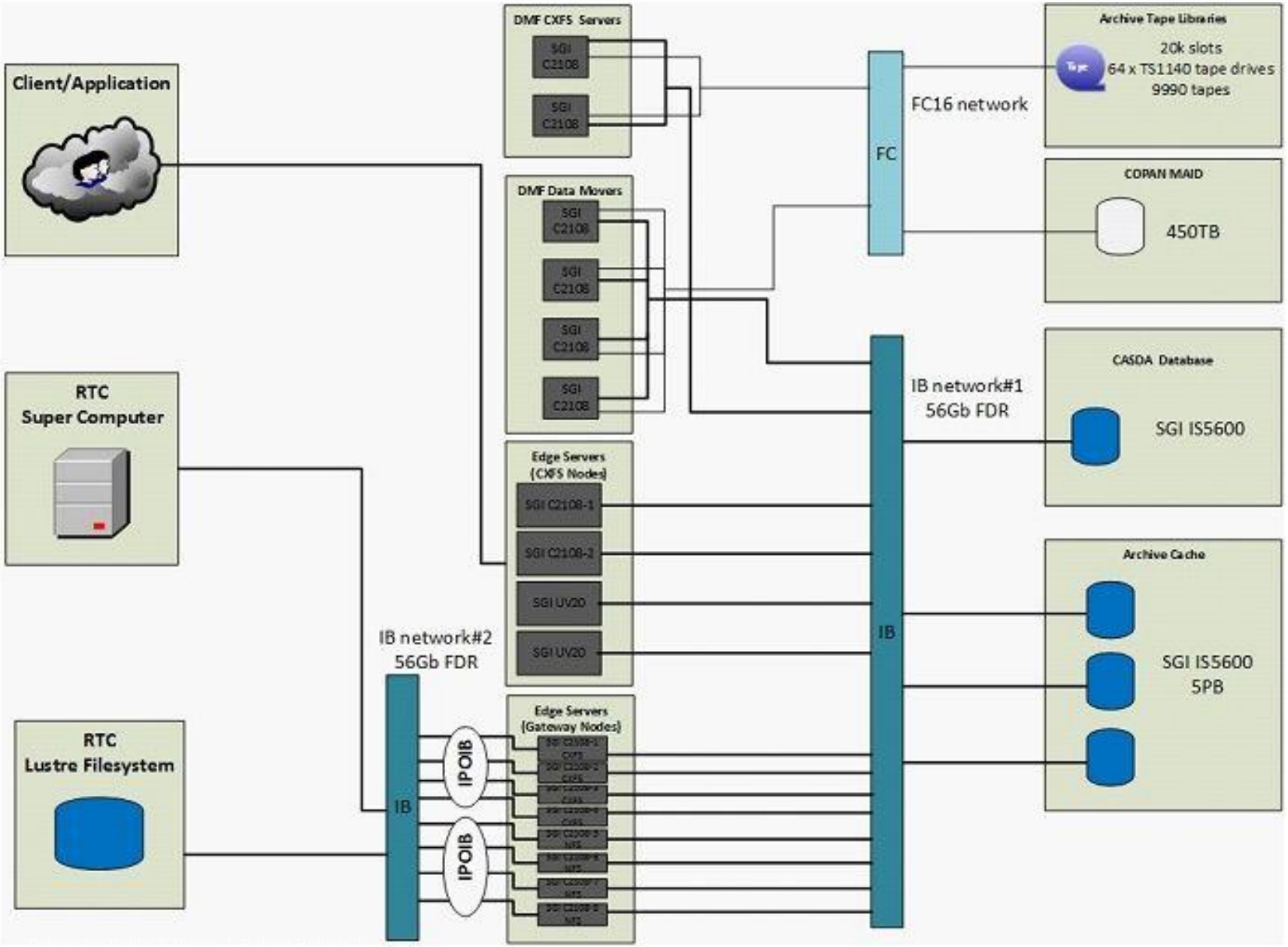
CASDA Data Products

Product	Data type
Calibrated continuum visibility data (stored as ‘measurement sets’)	CASA
Continuum image cubes (small number of frequency channels)	FITS
Spectral line image cubes (large number of frequency channels)	FITS
Postage stamp image cubes	FITS
Continuum source detections	Catalogue
Spectral line source detection catalogues	Catalogue
Transient source detection catalogues	Catalogue
Transient light curves – properties	Catalogue
Bright Source Catalogue (global sky model)	Catalogue

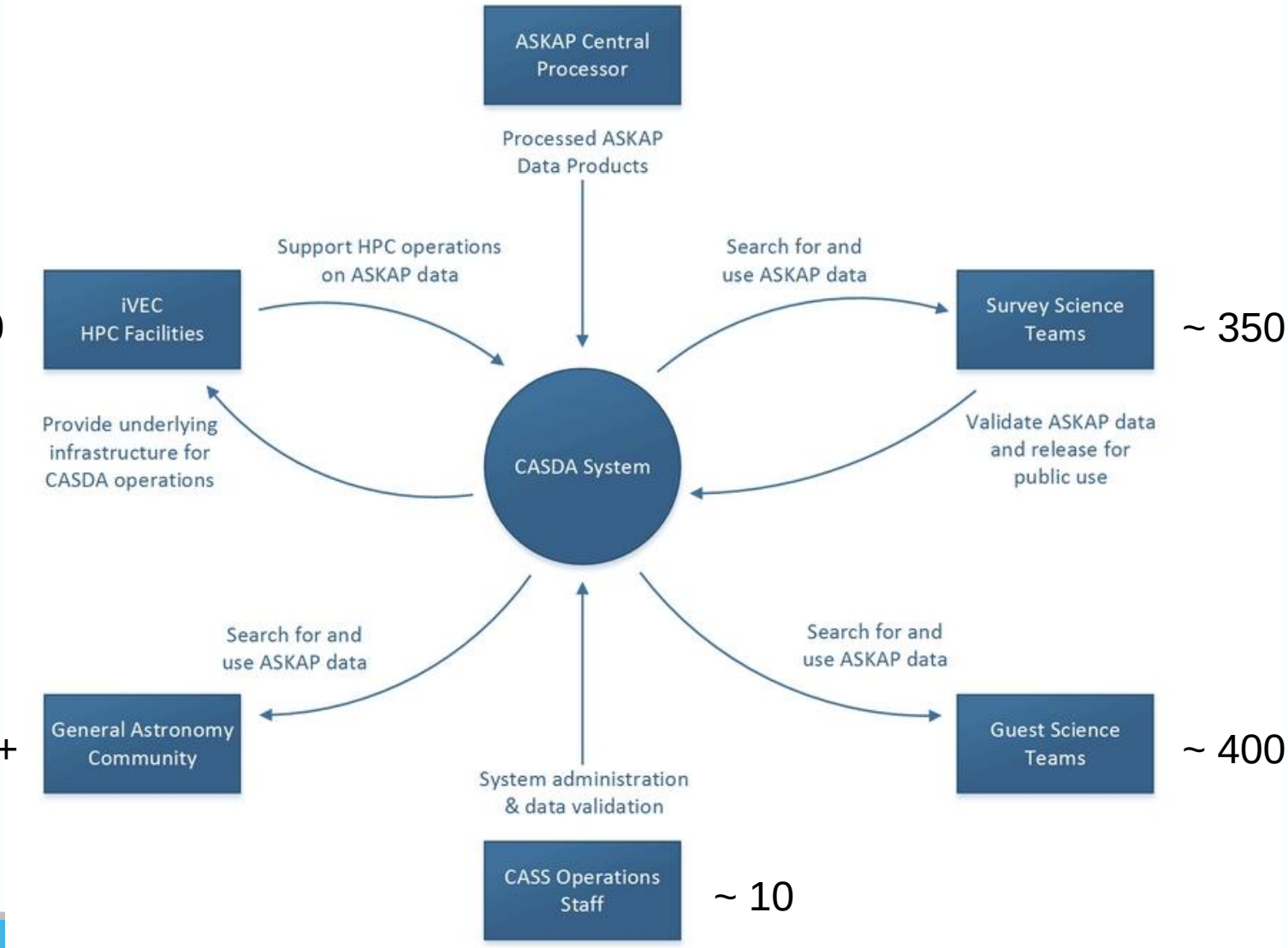
ASKAP Data Processing levels



CASDA use of Pawsey Centre infrastructure

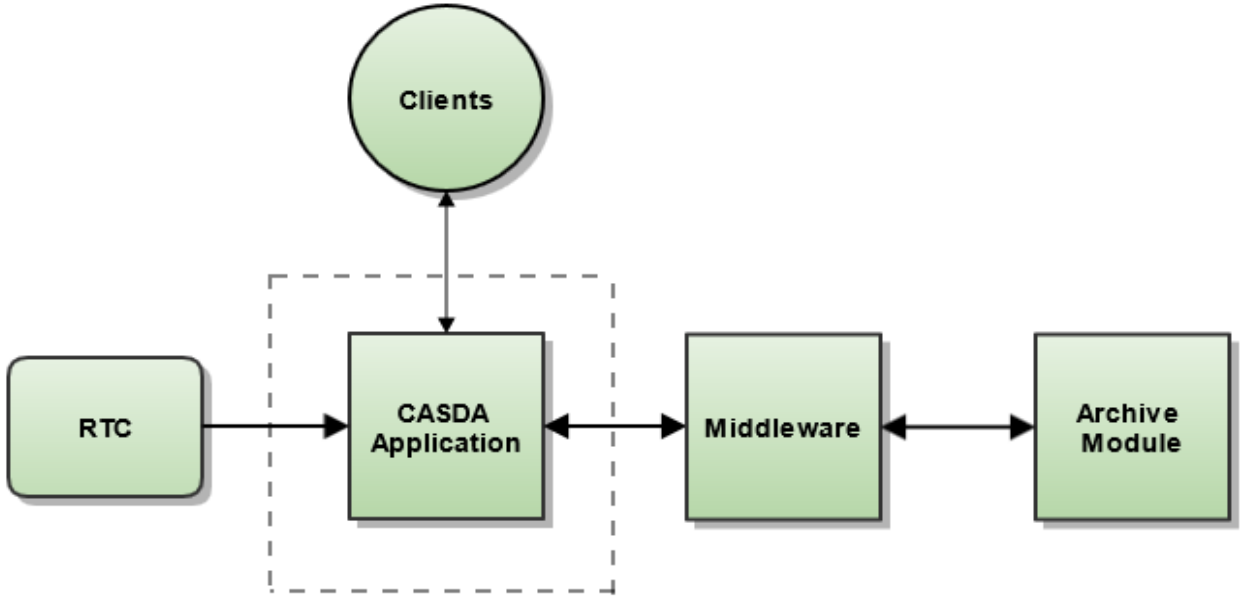


CASDA connections



CASDA analysis and planning

(decisions in progress)

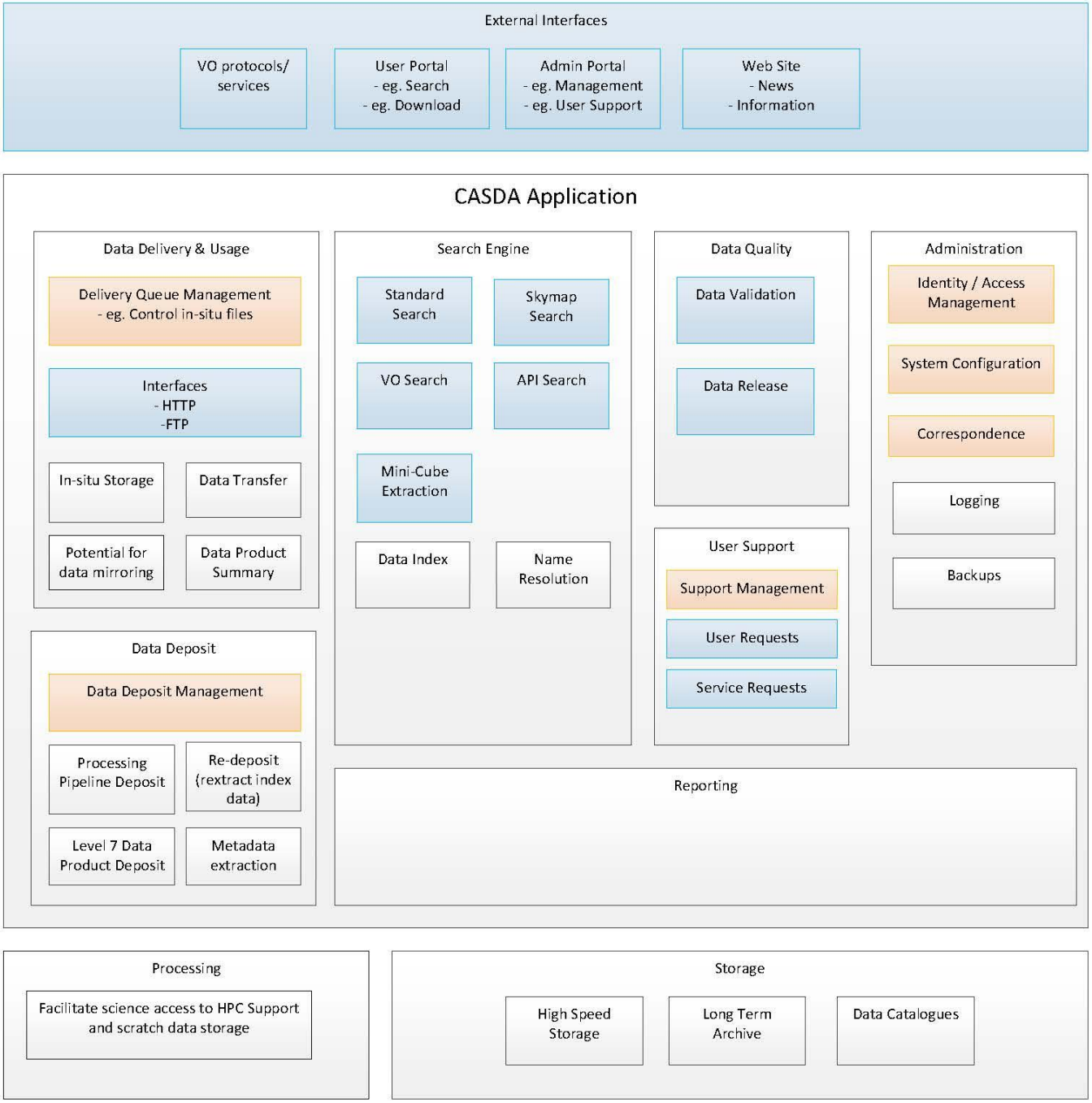


Middleware: NGAS

Database backends: PostgreSQL

VO tools: Based on CDS VO libraries (looking at ADQL, UWS and SAVOT llibraries) with additional implementation

CASDA application: Will include components from CSIRO DAP



CASDA Application capabilities

CASDA (Some) high-level requirements

Ref: ***CSIRO ASKAP Science Data Archive: Requirements and Use Cases***

Essential Requirements (a subset as examples)
ASKAP data products are open access and made publically available as soon as possible.
CASDA will provide access to images, image cubes and catalogues using VO protocols
Long term data storage will be provided at the Pawsey Centre.
The CASDA design will not restrict the potential future requirement for one or more copies to be stored at other locations.
CASDA will provide a repository for Survey Science Teams to upload predefined and VO-compatible science catalogues and will provide search tools for such catalogues (under negotiation).

Survey Science Teams: Example use cases

- Run query to obtain a listing of the visibility files archived and sky regions observed for project.
- Set data validation flags following review of image quality reports
- Simple cone searches
- Complex catalogue queries
- Download image cube ‘cut-outs’
- Download selected image cubes for further analysis
- *Upload ‘final’ science catalogues into archive and make available for general use.*

Survey Science Projects: File-based data sizes in CASDA

SSP	Type	Nfields	Time per field	Visibility data size per field (TB)	Image data size per field (TB)
EMU	C	1200	12	2.4	0.003
POSSUM	C	1200	8	1.5	1.0
WALLABY	S	1200	8	Not archived	1.8
DINGO	S	966	8	Not archived	1.3
FLASH	S	850	4	Not archived	0.5
GASKAP	S	644	12	Not archived	[0.5]
VAST	T	1200	8	<1	Probably not archived

For full ASKAP, CASDA will archive on average about 15 TB per day

Science Data Access

Access to images and catalogues will be provided through Virtual Observatory services:

Simple Image Access Protocol returns link to images/cubes identified for a given position. CASDA will provide tools to generate image ‘cut-outs’.

Cone searches: Returns table results such as positions and fluxes for sources detected within an area around a given position.

Table Access Protocol: Allows for complex querying of tables. For example – could return a list of detections for sources above a given flux density with negative spectral indices (slopes).

We are setting up a VO demonstrator to trial and test VO implementations for ASKAP data.

Need to build VO expertise in radio astronomy and engage community.

D: Communications with science users

- CASDA Science Reference Group (phase 0) ~ 15 members. Regular meetings since Oct 2013
- User Requirements draft document (v0.8) distributed 4 Nov 2013 for comments
- User Requirements document (v1.0) will be released in early Jan 2014
- CASDA monthly newsletter distributed by email
- Articles in ASKAP newsletter (but more needed for CASS website)

E: HPC data processing / high volume data

Issue	Notes
Survey Science teams – will require access to temporary data storage and HPC – for post processing of ASKAP data products	Difficult / impossible to transfer ‘large’ amounts of data over networks.
Radio astronomy community is unfamiliar with applying for HPC and data storage on other national facilities	Information /education for Australian astronomy community is needed
Australian application processes for HPC and data storage on different facilities are not well integrated	

iVEC Facilities

iVEC manages several high performance and data storage facilities including the Real Time Computer and Magnus supercomputers in the Pawsey Centre.

All iVEC facilities are allocated as follows:

Category	Allocation (compute + storage)	Notes
Radio astronomy	25%	Shared between ASKAP and MWA
Geoscience	25%	
Partners	30%	CSIRO, UWA, Curtin Uni, Murdoch Uni and Edith Cowan Uni
National Researchers	15%	Managed through the NCMAS
Director's time	5%	

iVEC Magnus Supercomputer

Magnus – CRAY supercomputer. Currently has 69 TF compute power installed. Final configuration will be at least 1 PF.

HPC allocations on Magnus include provision of scratch data storage.

Major application rounds for Magnus in Q4 each year for following calendar year:

Science teams with Australian researchers may be able to apply:

- to NCMAS [ncmas.nci.org.au]
- to iVEC for the Major Partner share allocation [portal.ivec.org/ivecallocation/]

The iVEC Director's Share scheme (5% share) is always open.

Thank you

Jessica Chapman

CSIRO Astronomy and Space Science

t +61 2 9372 4196
e Jessica.Chapman@csiro.au
w atnf.csiro.au