CSIRO ASKAP Science Data Archive

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

Jessica Chapman, ATUC Meeting, 5 December 2013

CSIRO ASKAP Science Data Archive (CASDA)

Talk outline

A: CASDA overview and timeline

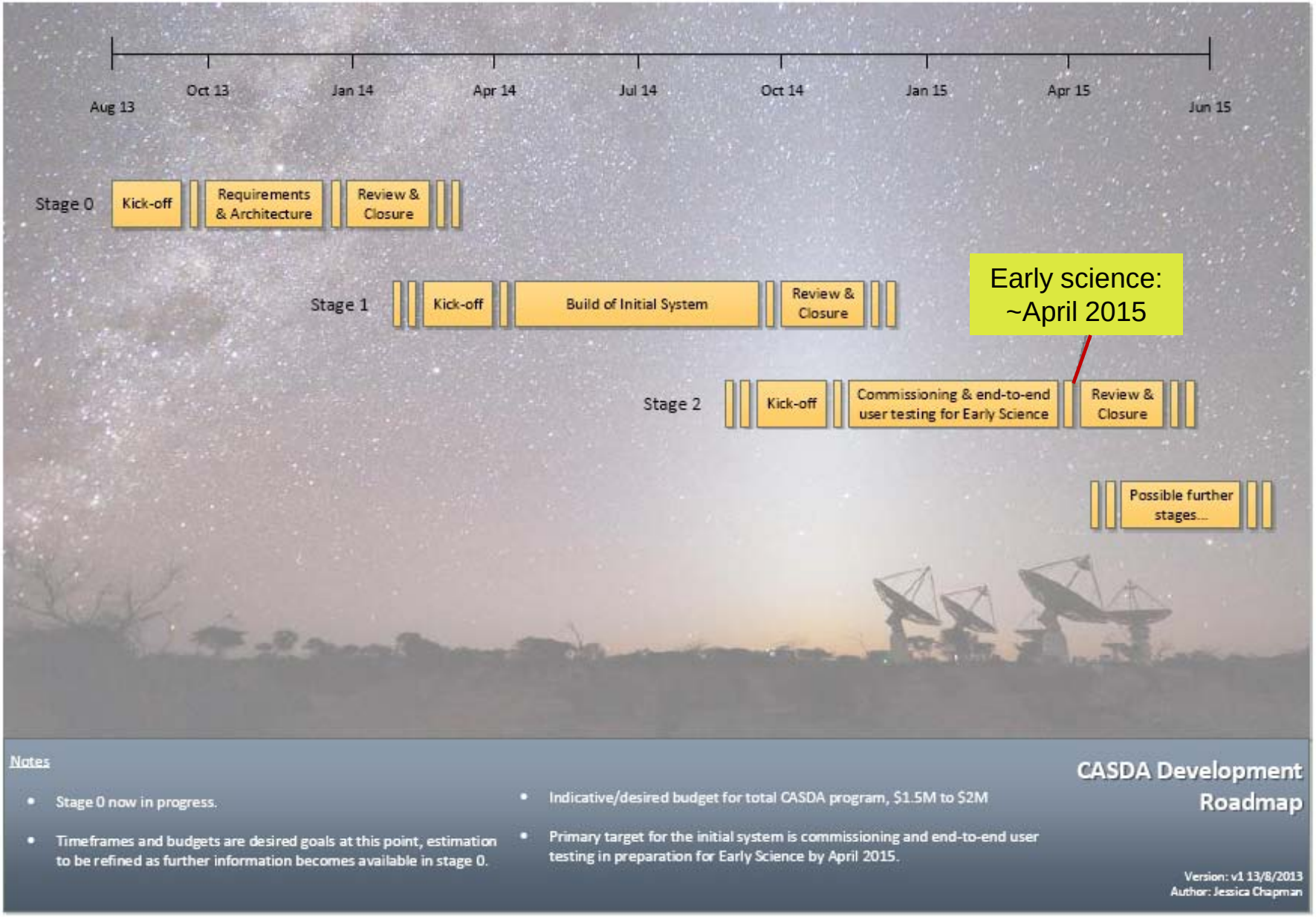
B: Requirements and use cases

C: Data access and data volumes

D: Science communications

E: High performance computing

CASDA: Project roadmap

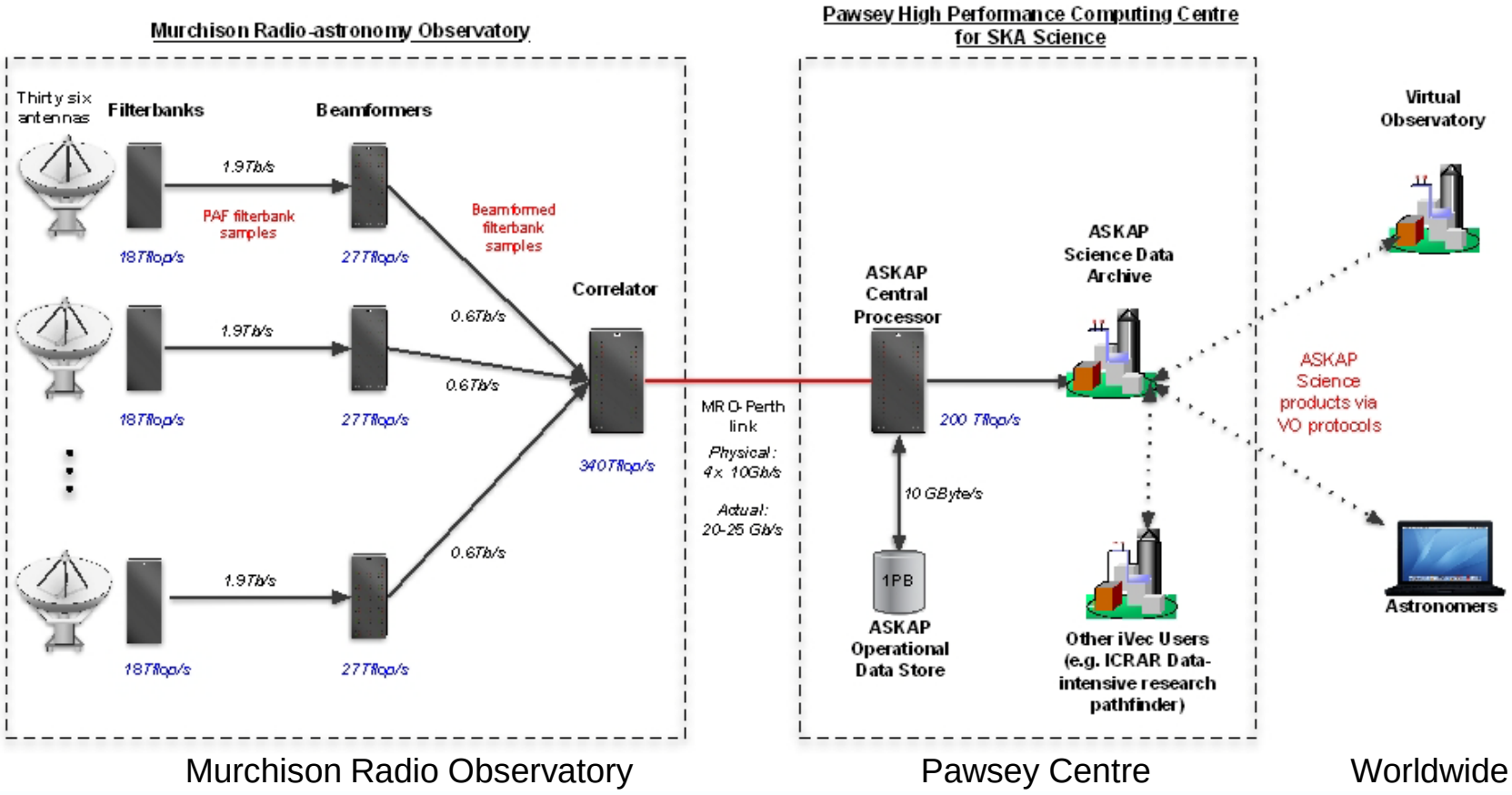


CASDA Stage 0 development team

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)	Jessica Chapman (Project Leader) Ian Heywood (Project Scientist) Arkadi Kosmynin (Software developer) Matthew Whiting (Science data processing) Ben Humphreys (Science data processing)

Stage 0 deliverables

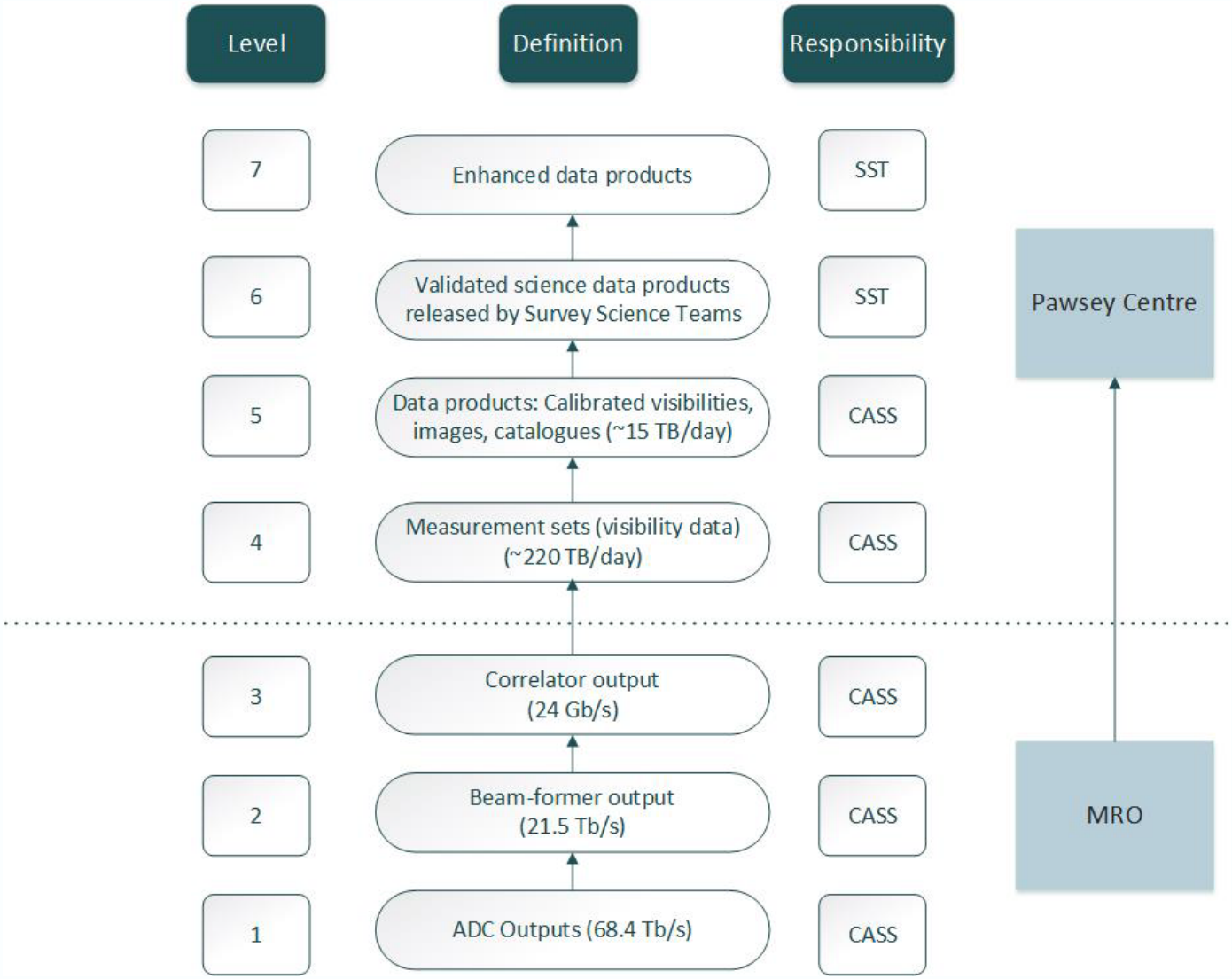
- Requirements description
- Use case model
- Workflows model
- System Architecture description
- Middleware design (database solution)
- Preliminary support model (support from iVEC, IM&T & CASS)
- Preliminary Design Review – probably December 2013



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

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

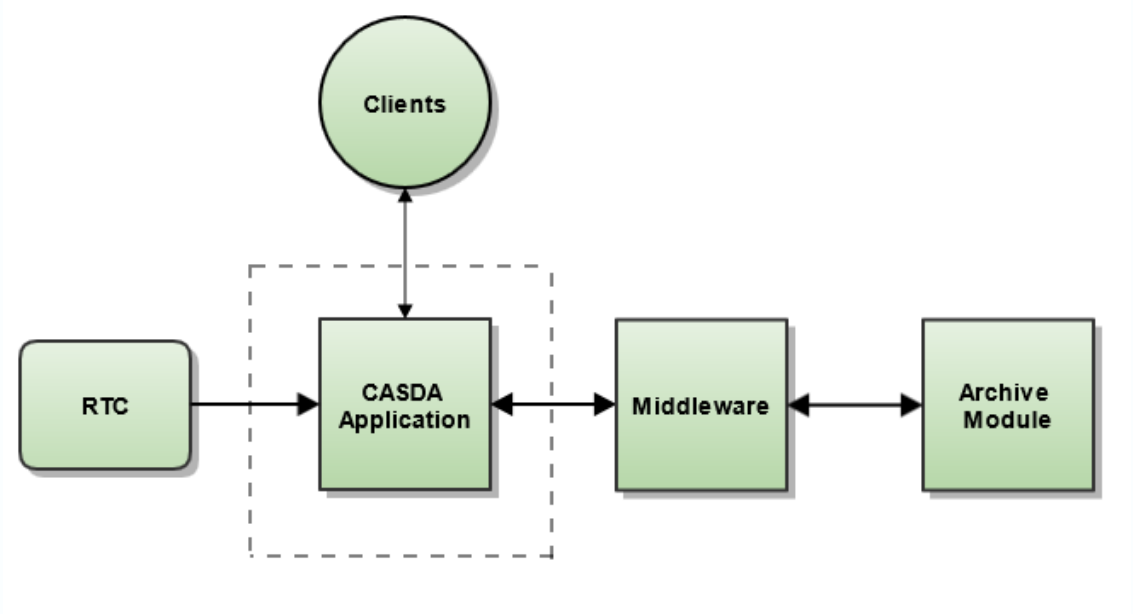
ASKAP Data Processing levels



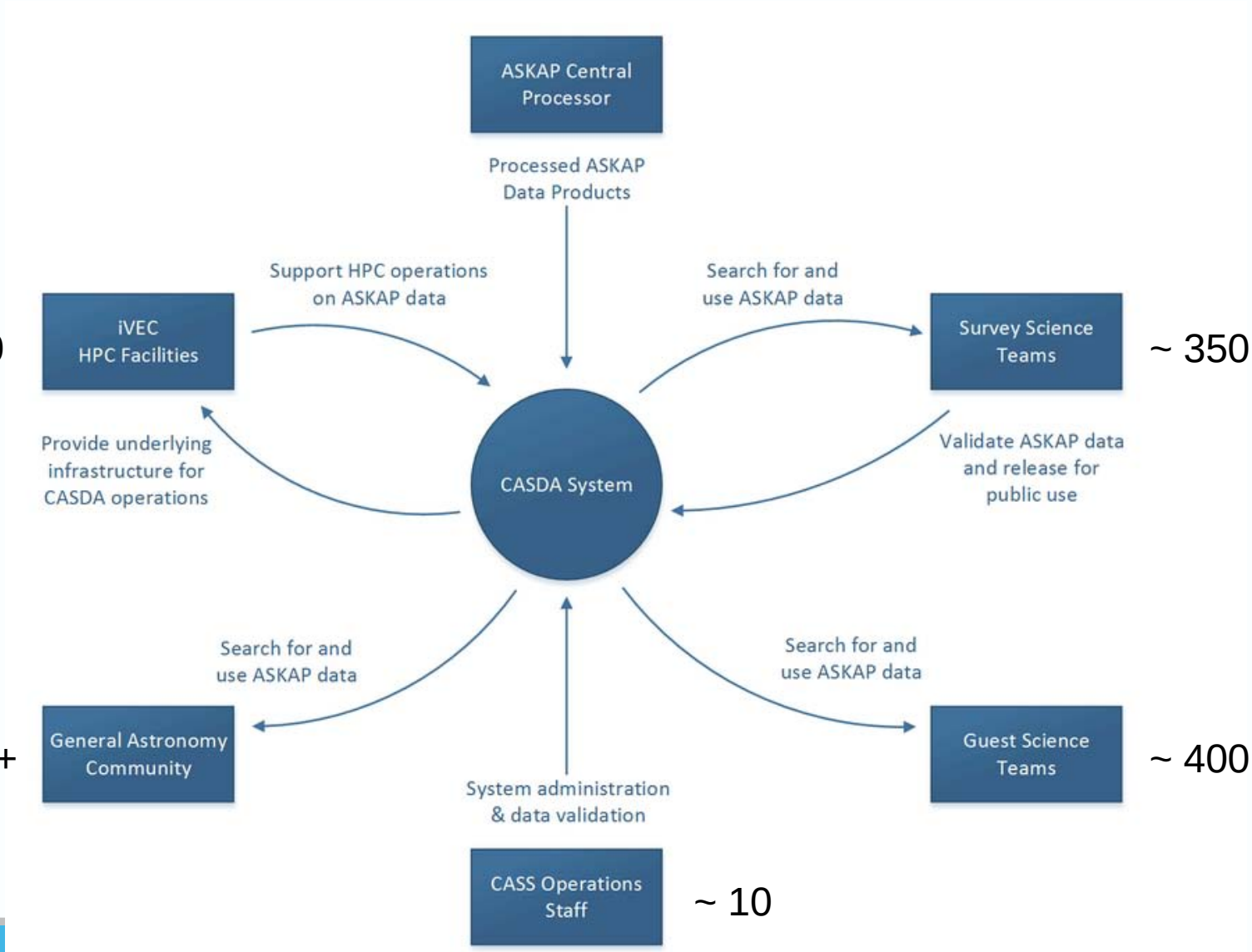
CASDA Data Products

Product	Data type
Calibrated continuum visibility data (stored as ‘measurement sets’)	File
Continuum image cubes (small number of frequency channels)	File
Spectral line image cubes (large number of frequency channels)	File
Postage stamp image cubes	File
Continuum source detection catalogue(s)	Catalogue
Spectral line source detection catalogue(s)	Catalogue
Transient source detection catalogue(s)	Catalogue
Transient light curves – properties	Catalogue
Bright Source Catalogue (built up for a global sky model)	Catalogue

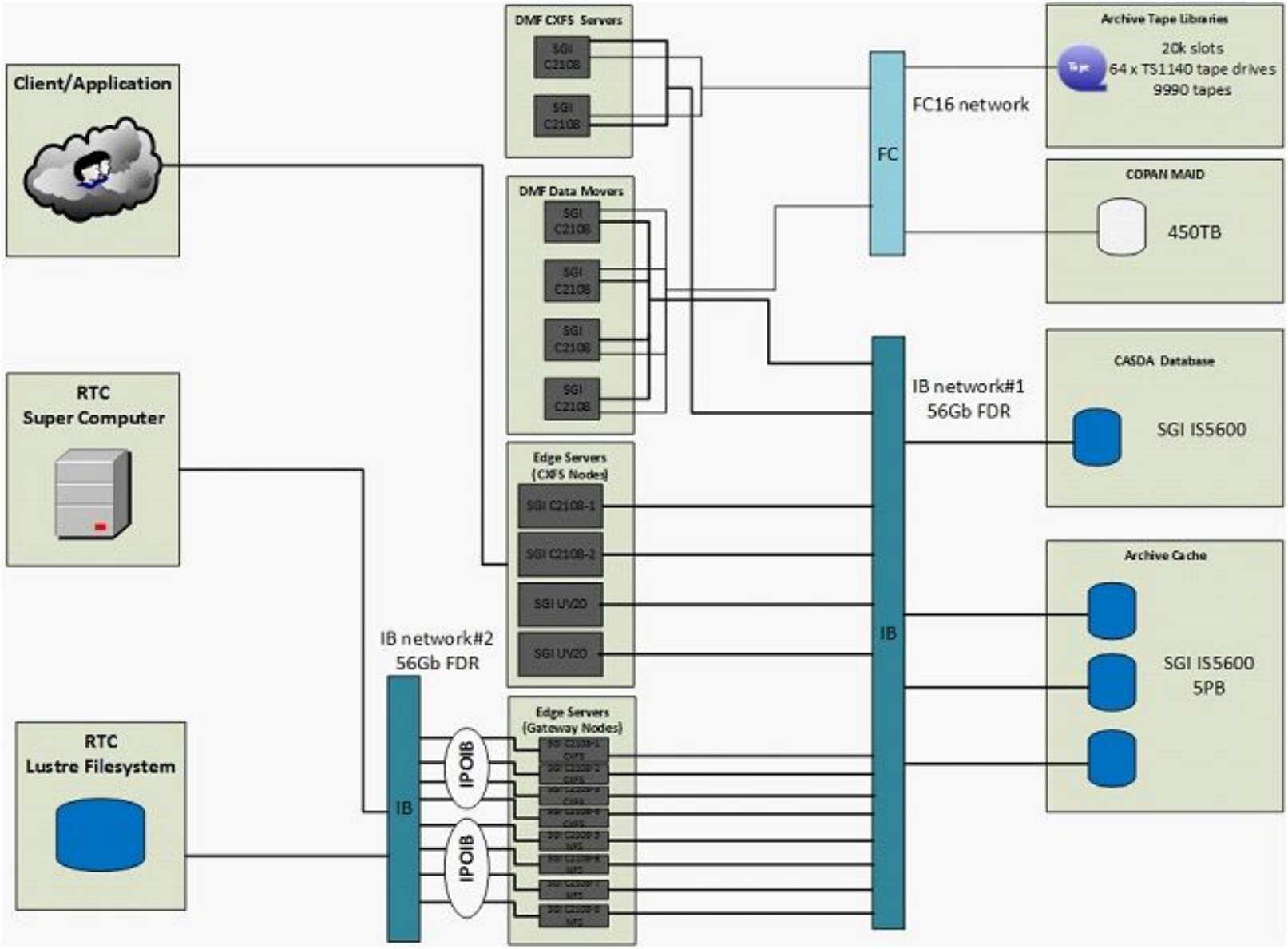
CASDA context



CASDA stakeholders



Physical infrastructure



CASDA High-level requirements

Ref: CSIRO ASKAP Science Data Archive: Requirements and Use Cases
(in ATUC doc repository)

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
CASDA will handle the ingestion of data products generated by ASKAP Survey Science Projects, Guest Projects and ToO observations in a timely and efficient way
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.
- Access automated information on image quality, RFI system performance etc.
- Cone searches
- More complex catalogue queries
- Set data validation flags following review of image quality reports
- Generate and download image cube ‘cut-outs’
- Download selected image cubes for further analysis
- *Load ‘final’ science catalogues into archive and make available for general use.*

The science survey teams will have at least some high-volume data requirements

General Users: Example use cases

- Cone searches – identify sources of interest
- Download small number of full-sized images
- Download set of postage stamp images
- Carry out more complex table searches
- Download light curve information for given source(s)
- Comparison of ASKAP data with data from other facilities

In most cases the data volume requested by individual or small groups of external users is likely to ‘reasonably’ low.

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 will be developing a VO demonstrator to trial VO protocol implementations and to test these against the CASDA requirements.

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	2.6E-3
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

D: Science Communications

- CASDA Science Reference Group (phase 0) – 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)
- Presentations – e.g. Astroinformatics workshop next week
- Advice from ATUC?

Does ATUC have any advice for requirements of guest science projects and/or the general astronomy community?

CASDA Science Reference Group

Ian Heywood (**Chair**: Project Scientist)

Dan Miller (Project Manager), Jared Pritchard (Business Analyst)

James Allison, Nathan Clarke

Richard Dodson, Phil Edwards

Lisa Harvey-Smith, Ian Heywood

George Hobbs, Andrew Hopkins

Baerbel Koribalski, Tara Murphy

Naomi McClure-Griffiths, Martin Meyer,

Vanessa Moss, Ray Norris,

Lisa Harvey-Smith, Cormac Purcell,

Elaine Sadler, Paolo Serra

Nick Seymour, Lister Staveley-Smith

Steven Tingay, Matthew Whiting

The SRG has provided valuable input towards the specification of use cases.

HPC data processing / high volume data

Issue	Notes
Survey Science teams – likely to require access to temporary data storage and HPC – for post processing	Cannot transfer very large amounts of data over networks. CASDA will not provide postal services.
Radio astronomy community is unfamiliar with applying for HPC and data storage on other national facilities	Information /education for Australian astronomy community is needed

iVEC computing 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 with several thousand CPU cores.

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

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

Researchers can apply for

i) to NCMAS [ncmas.nci.org.au]

ii) to iVEC for consideration under the Major Partner share allocation [portal.ivec.org/ivecallocation/]

NCMAS applications are reviewed before the iVEC partner share applications.

Where applications are not successful, or are only partly supported, they are forwarded to iVEC for consideration through the partner scheme.

Use of Magnus (ctd)

The Director's Share scheme (5% share) is always open. Researchers may apply at any time. Response from iVEC within 6 weeks.

HPC allocations on Magnus include provision of scratch data storage.

It will be the responsibility of science teams to apply for HPC facilities as needed

There are also application systems for archival type data storage.

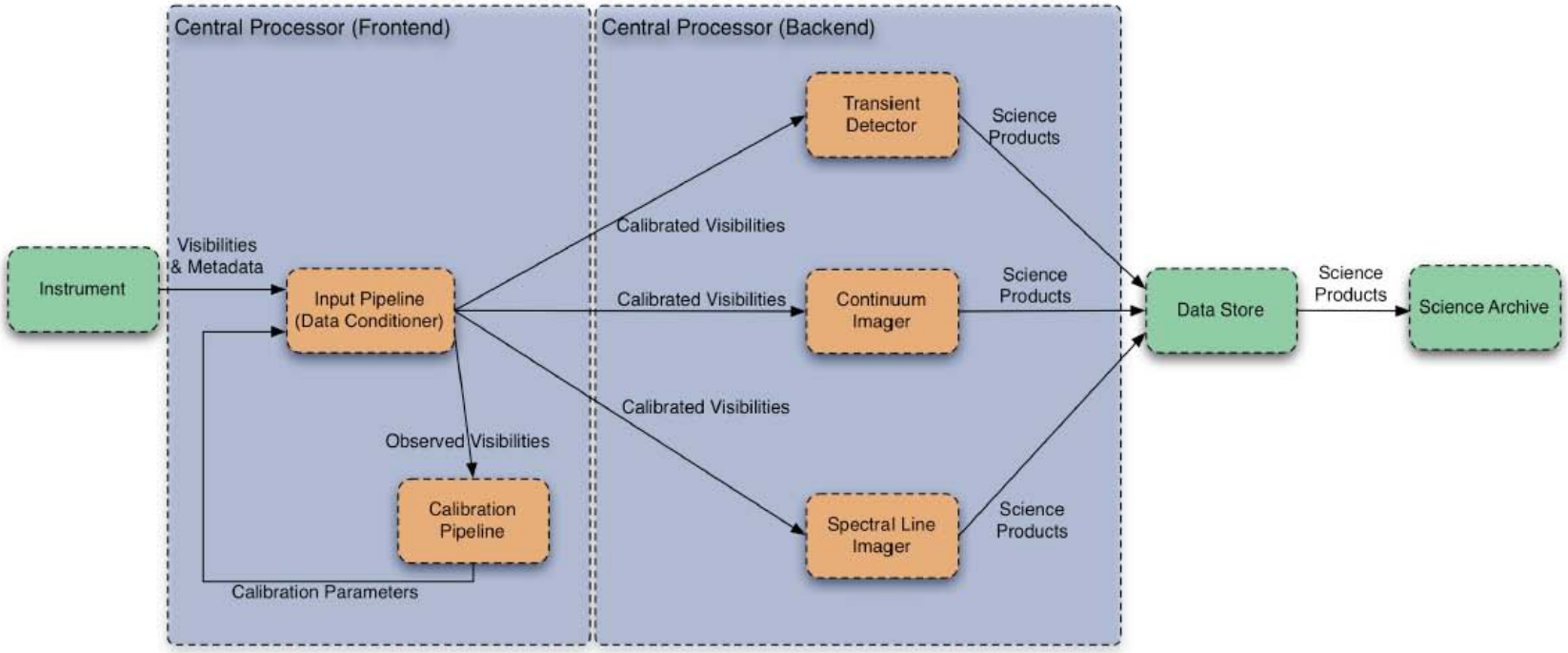
Thank you

Jessica Chapman

CSIRO Astronomy and Space Science

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

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’.