MANCHESTER 1824

The University of Manchester
Jodrell Bank Observatory

IRIS SKA Update

T3.1 ESDC Processing: Inventory of SKA science cases and post-SDP computing requirements

- 11 Science Working Groups (SWGs)
- 13 High Priority Science Objectives (HPSOs)
- 15 SKA SDP Data Products

Use Cases were selected:

1. to be representative of a wide range of processing models;
2. to cover a range of SKA science working groups;
3. to use a variety of SDP data products as input data;
4. to be high usage cases, i.e. they will need to be run as standard processing on the majority of datasets.

Table 1: List of SKA Science Working Groups.

SKA Science Working Groups			
1	Extragalactic Spectral Line	7	Our Galaxy
2	Solar, Heliospheric & Ionospheric Physics	8	Epoch of Reionization (EOR)
3	Cosmology	9	Extragalactic Continuum
4	Cradle of Life	10	HI Galaxy Science
5	Magnetism	11	Pulsars
6	Transients		

Table 2: Summary table of processing Use Cases used within WP3.

No.	Name	Input Data	SWGs
1	Calibration & Imaging	Calibrated Visibilities	**1**, 3, **8**
2	Pulsar Re-folding	Pulsar Candidates	**11**
3	Rotation Measure Synthesis	Image Cube [4]	**5**
4	Object Detection and Classification	Image Cube [1]	1, 3, 4, 5, 6, 7, **9**, 10
5	Automated Object Classification	LSM Catalogue	1, 3, 4, 5, **6**, 7, 9, 10

T3.2 ESDC Data storage: Inventory and sizing of SKA science data products and ESDC user-derived products

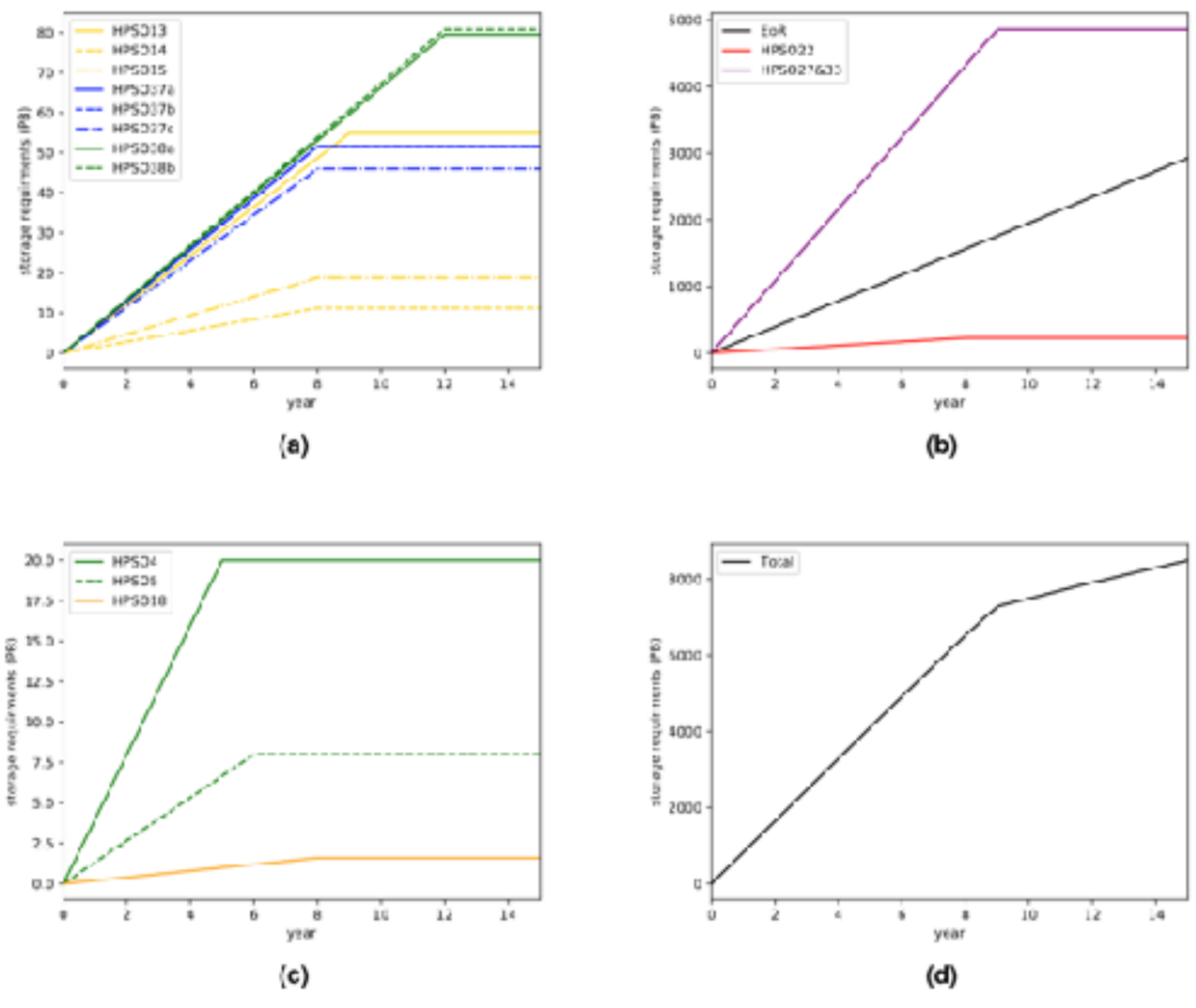


Figure 1: (a) Data storage requirements at SRCs for the HI and continuum HPSO. (b) Data storage requirements at SRCs for the EoR, magnetism and cradle of life HPSOs. (c) Data storage requirements at SRCs for the pulsars and transients HPSOs. (d) Data storage requirements at SRCs for all HPSOs.

Storage **volume** requirements depend on:

- Rate of ingest of SDP data products;
- Rate of production of advanced data products.

Rate of ingest of SDP data products was based on the SDP System Sizing, assuming that the ESDC holds a complete copy of all products.

Rate of production of advanced data products was determined using the AENEAS processing Use Cases and was determined to be **3:1 in volume** (output:input).

Also examined the **number** of data products expected as a function of time. Relevant for the choice of Data Management System (DMS).

T3.4 Design and costing for distributed ESDC computing architecture

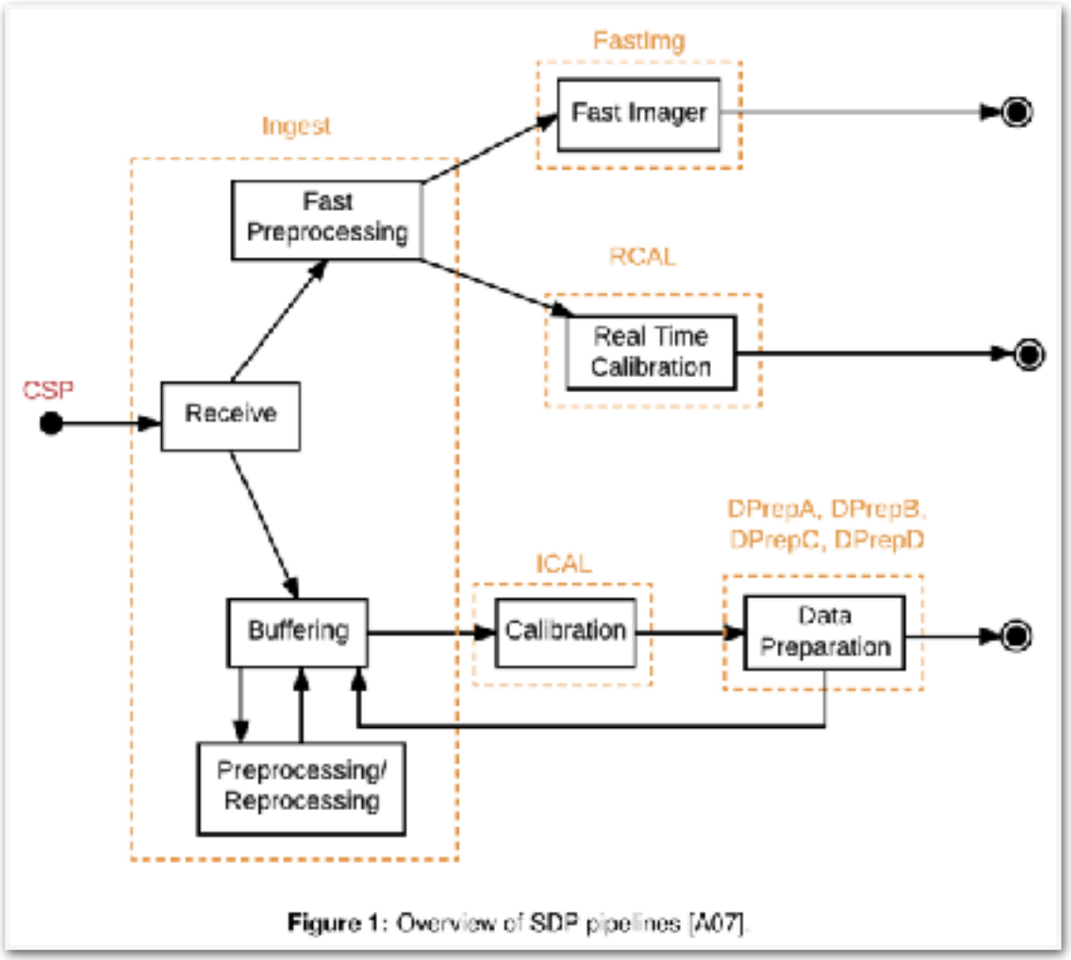
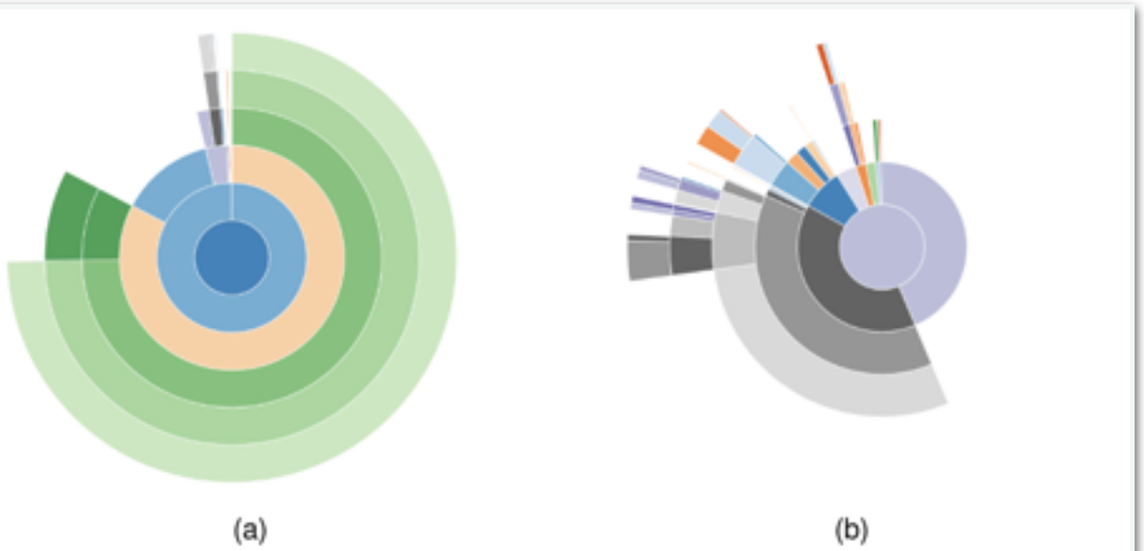


Figure 1: Overview of SDP pipelines [A07].

Processing is being examined in terms of (1) compute load; (2) memory requirements; (3) potential for distribution; (4) suitability of platform.



(a) (b)

Figure 9: Visualisation of processing time breakdown for PyRMSynth. (a) The overall breakdown, dominated by `write_output_files` (green); (b) the breakdown beneath `compute_dirty_image`, dominated by the FFT (grey). See text for details.

Re-processing + Post-processing

For more detail on use cases / compute models, the deliverable document is publicly available:

D3.1: https://www.astron.nl/aeneas2020/doku.php?id=intra:deliverables

For an initial costing, the deliverable document is publicly available:

D3.3: https://www.astron.nl/aeneas2020/doku.php?id=intra:deliverables

T3.6 Validation, Verification & Proof of concept activities utilizing SKA pathfinder and pre-cursor facilities

Table 2: Summary table of processing Use Cases used within WP3.

No.	Name	Input Data	SWGs
1	Calibration & Imaging	Calibrated Visibilities	**1**, 3
2	Pulsar Re-folding	Pulsar Candidates	**11**
3	Rotation Measure Synthesis	Image Cube [**4**]	**5**
4	Object Detection and Classification	Image Cube [1]	1, 3, 4, 5, 6, 7, **9**, 10
5	Automated Object Classification	LSM Catalogue	1, 3, 4, 5, **6**, 7, 9, 10

Table 6: Summary table of reference data sets used within WP3.

Dataset No.	Dataset Name	Equivalent SDP Product	SWG	Size (% SDP)
1	LOFAR1	Calibrated Visibilities	8	10.5 TB (15%)
2	LOFAR2	Image Cube [4]	**1,9**	36 GB (7%)*
3	MWA	Image Cube [1]	9	2.2 GB
4	SciServer	LSM Catalogue	6,9	0.34 GB (0.03%)
5	GBT	Pulsar Candidates	11	1 MB

* Image cube parameters for HPSO-27.

Table 6: Summary table of Use Case instances.

			Environment				
Use Case No.	Use Case Name	Data Set	1	2	3	4	5
1	Calibration & Imaging	LOFAR1	N	Y	Y	–	Y
2	Pulsar Re-folding	Pulsar	Y	Y	Y	–	N
3	Rotation Measure Synthesis	LOFAR2	Y	Y	Y	–	N
4	Object Detection and Classification	MWA	Y	Y	Y	–	N
5	Automated Object Classification	SciServer	Y	Y	Y	–	Y

T3.6 Validation, Verification & Proof of concept activities utilizing SKA pathfinder and pre-cursor facilities

Table 2: Summary table of processing Use Cases used within WP3.

No.	Name	Input Data	SWGs
1	Calibration & Imaging	Calibrated Visibilities	**1**, 3
2	Pulsar Re-folding	Pulsar Candidates	**11**
3	Rotation Measure Synthesis	Image Cube [4]	**5**
4	Object Detection and Classification	Image Cube [1]	1, 3, 4, 5, 6, 7, **9**, 10
5	Automated Object Classification	LSM Catalogue	1, 3, 4, 5, **6**, 7, 9, 10

Table 3: Summary table of working environments used within WP3.

Env.	Name	Processor	Memory	Cores	Represents
1	MacBook Pro	3.5 GHz Intel Core i7	16 GB DDR3	4	Basic User
2	Linux Box	Intel Xeon E5-2640 v4	256 GB DDR4	40(*)	Advanced User / HPC
3	GridPP1		16 GB	8	Grid (Standard PP)
4	GridPP2	Tesla V100 GPU	16 GB HBM2	640	Grid (Accelerated)
5	SurfSARA	Intel Xeon Gold 6148	32 GB	4	Grid (Fat)

* 2×10 hyper-threaded

Table 6: Summary table of reference data sets used within WP3.

Dataset No.	Dataset Name	Equivalent SDP Product	SWG	Size (% SDP)
1	LOFAR1	Calibrated Visibilities	8	10.5 TB (15%)
2	LOFAR2	Image Cube [4]	1,9	36 GB (7%)*
3	MWA	Image Cube [1]	9	2.2 GB
4	SciServer	LSM Catalogue	6,9	0.34 GB (0.03%)
5	GBT	Pulsar Candidates	11	1 MB

* Image cube parameters for HPSO-27.

Table 6: Summary table of Use Case instances.

			Environment				
Use Case No.	Use Case Name	Data Set	1	2	3	4	5
1	Calibration & Imaging	LOFAR1	N	Y	Y	–	Y
2	Pulsar Re-folding	Pulsar	Y	Y	Y	–	N
3	Rotation Measure Synthesis	LOFAR2	Y	Y	Y	–	N
4	Object Detection and Classification	MWA	Y	Y	Y	–	N
5	Automated Object Classification	SciServer	Y	Y	Y	–	Y

Machine learning applications

All AENEAS use cases are containerised and running on IRIS/GridPP

- Also eMERLIN processing using Jupyter on IRIS;
- DIRAC TS tested for all use cases - works well, but needs some tweaking for general astro processing;
- Rucio instance running @ RAL;
- IRIS workshops held @ Manchester - gradually moving users onto IRIS but x509 is causing delays

Currently:

- Procurement of IRIS fat nodes - including sandbox node;
- Setting up fast (“hot”) buffer tests;
- Rucio-DIRAC collaboration for Rucio catalog “DIRAC-mode”;
- Multi-user tests on new Rucio instance @ RAL - would be good to have ElasticSearch capability with Rucio & more generic resource tagging for DIRAC