Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

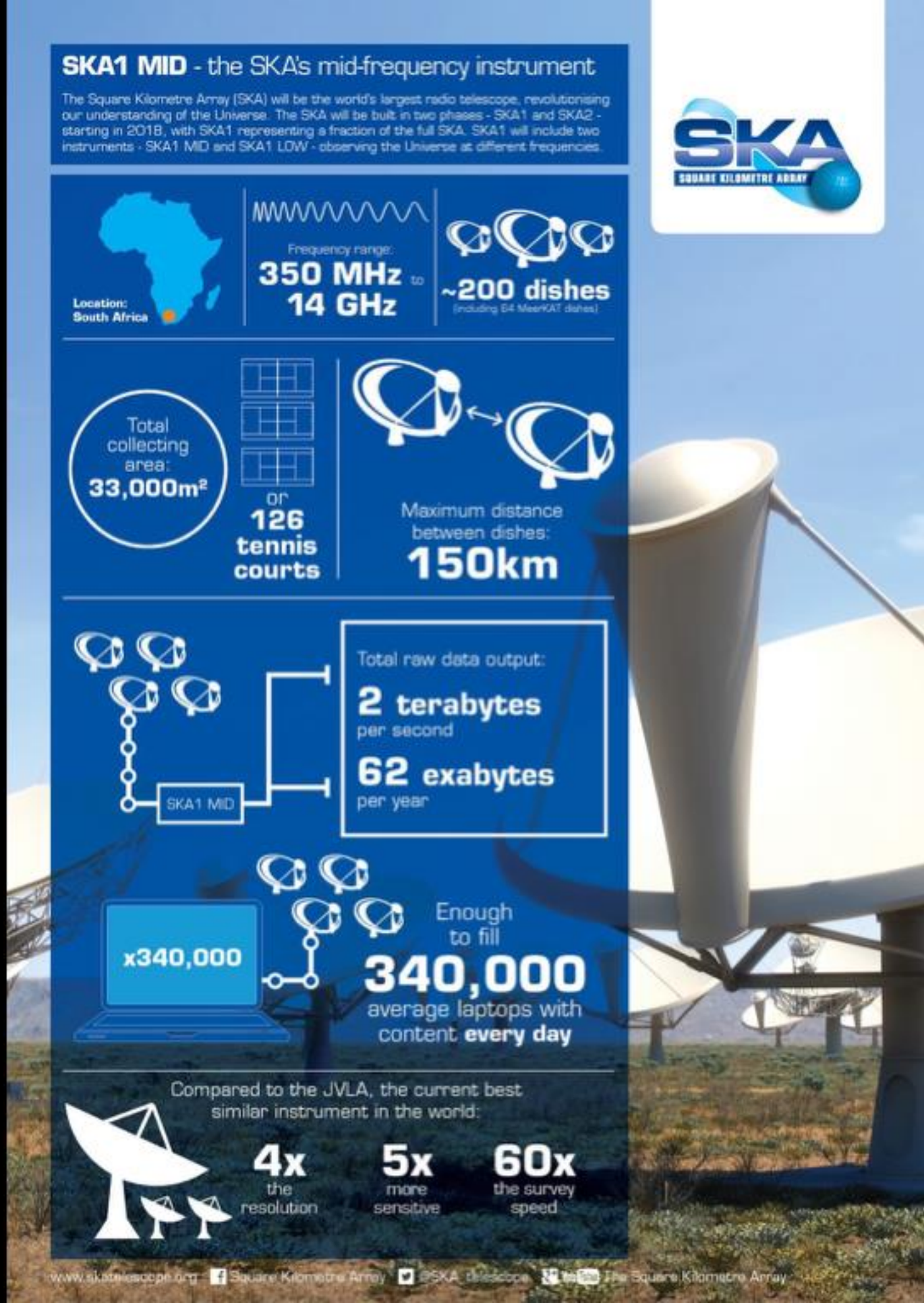
AENEAS: An SKA Regional Centre for Europe

Rohini Joshi
University of Manchester

Develop a concept and design for a distributed, federated European Science Data Centre (ESDC) to support the astronomical community in achieving the scientific goals of the Square Kilometre Array

DISH ARRAY

SKA1 MID

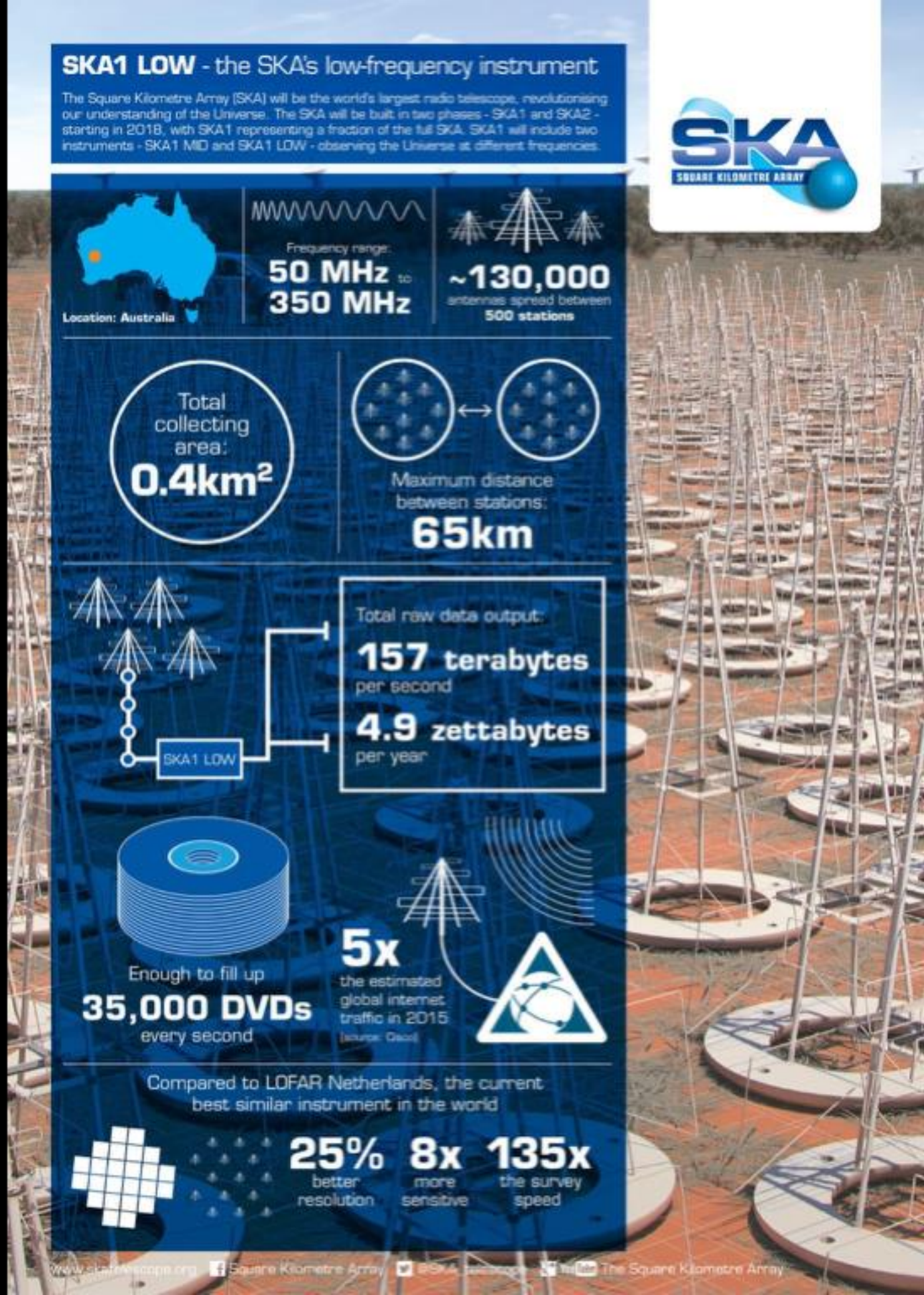


APERTURE ARRAY

SKA1 LOW

TERABYTE = 10^{12} BYTES

ZETTABYTE = 10^{21} BYTES

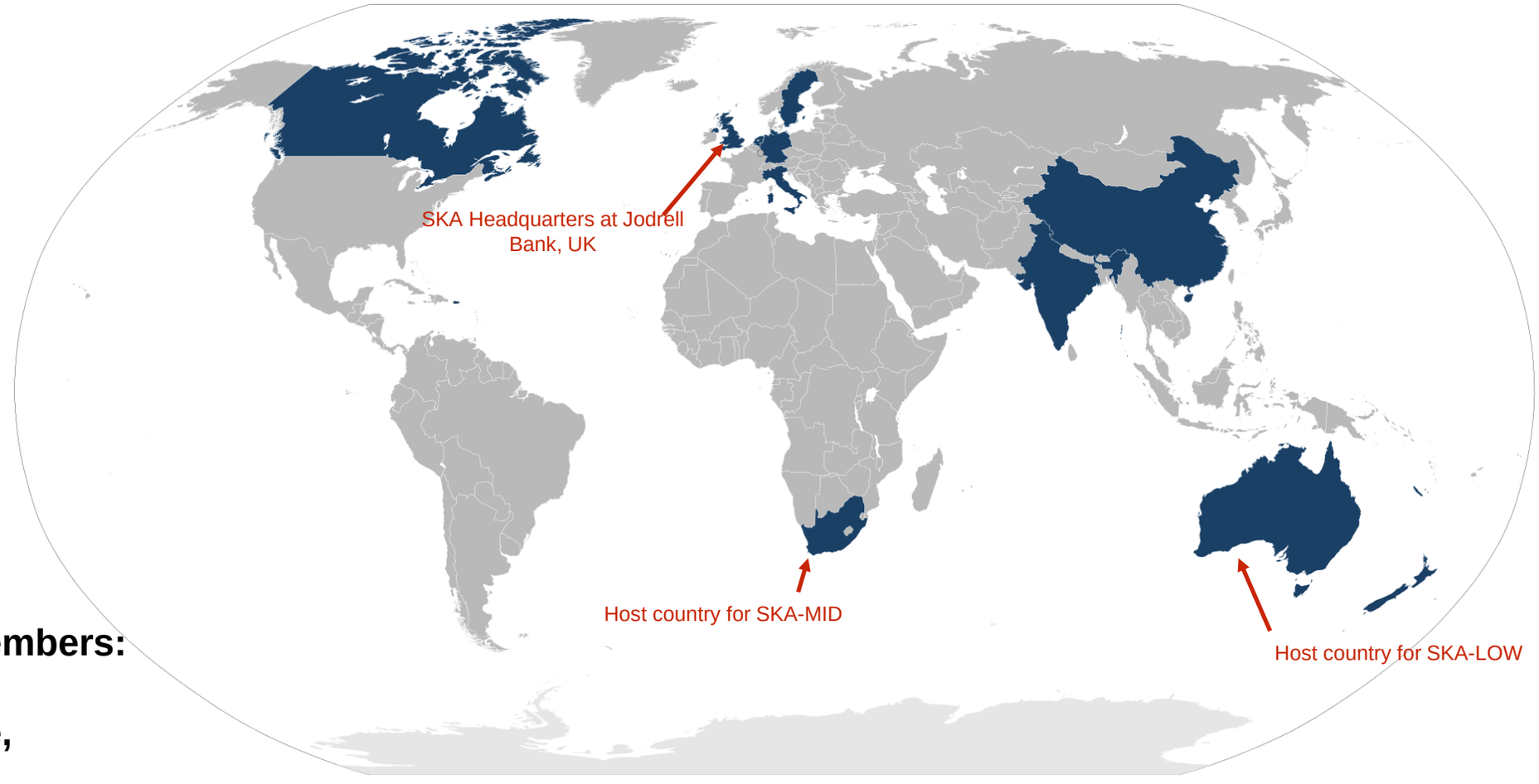


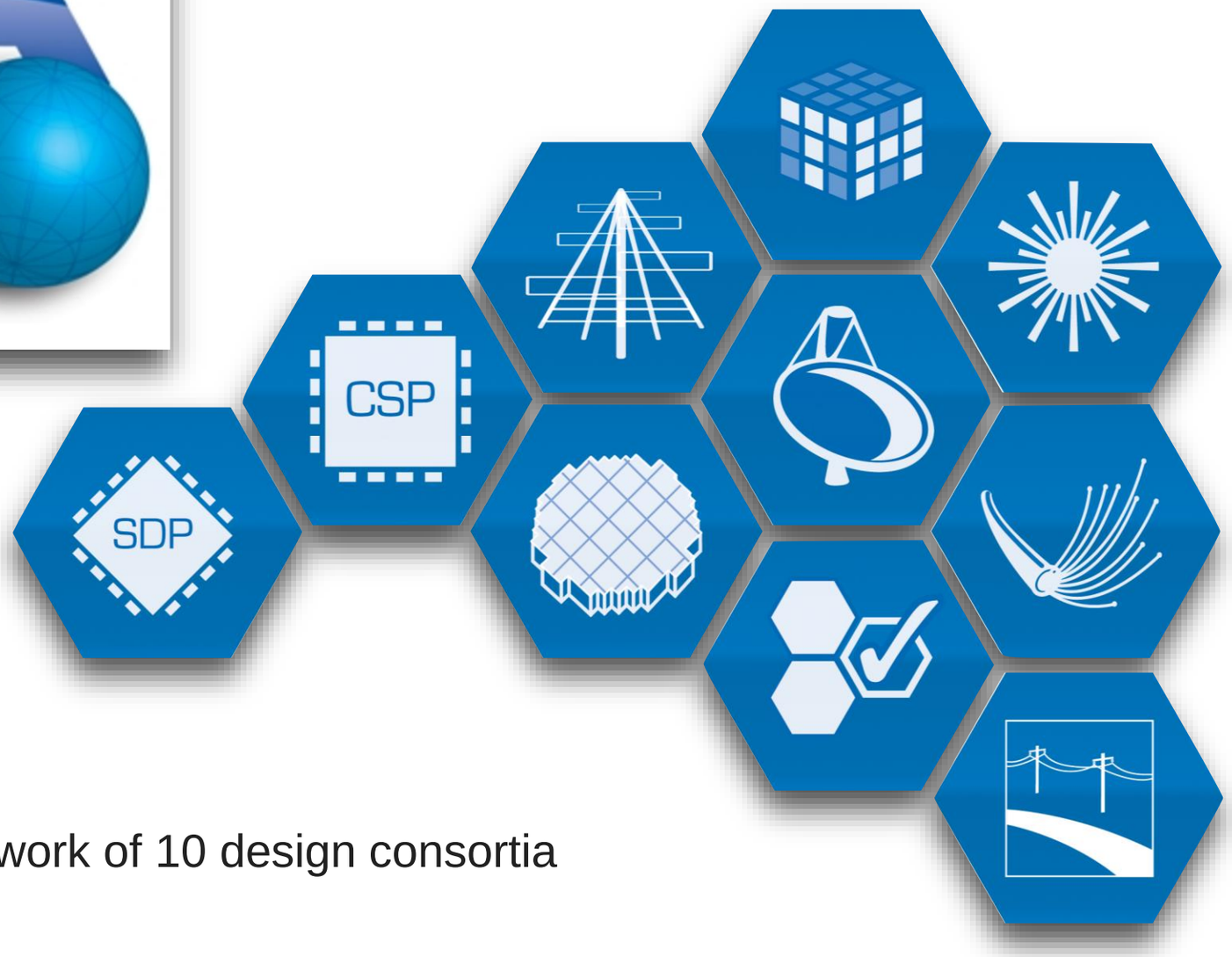
The Square Kilometre Array International Organisation (SKAO)

The Square Kilometre Array

- **Australia**
- **Canada**
- **China**
- **India**
- **Italy**
- **Netherlands**
- **New Zealand**
- **South Africa**
- **Sweden**
- **UK**

Potential new members: Spain, Portugal, Germany, France, others





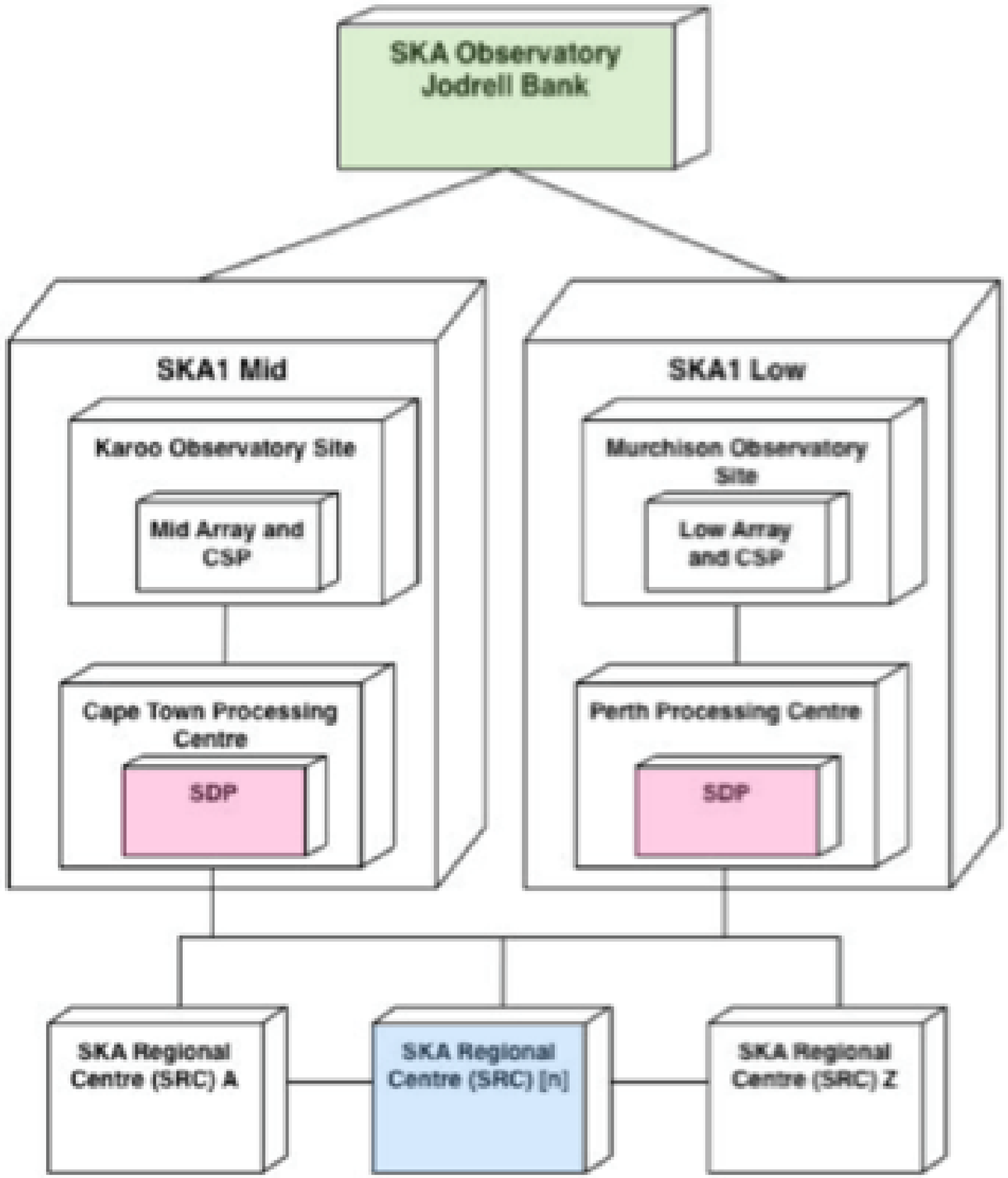
The SKAO oversees the work of 10 design consortia

Global Network of Centres


SKA
SQUARE KILOMETRE ARRAY


SKA
SQUARE KILOMETRE ARRAY
SDP
SCIENCE DATA PROCESSOR

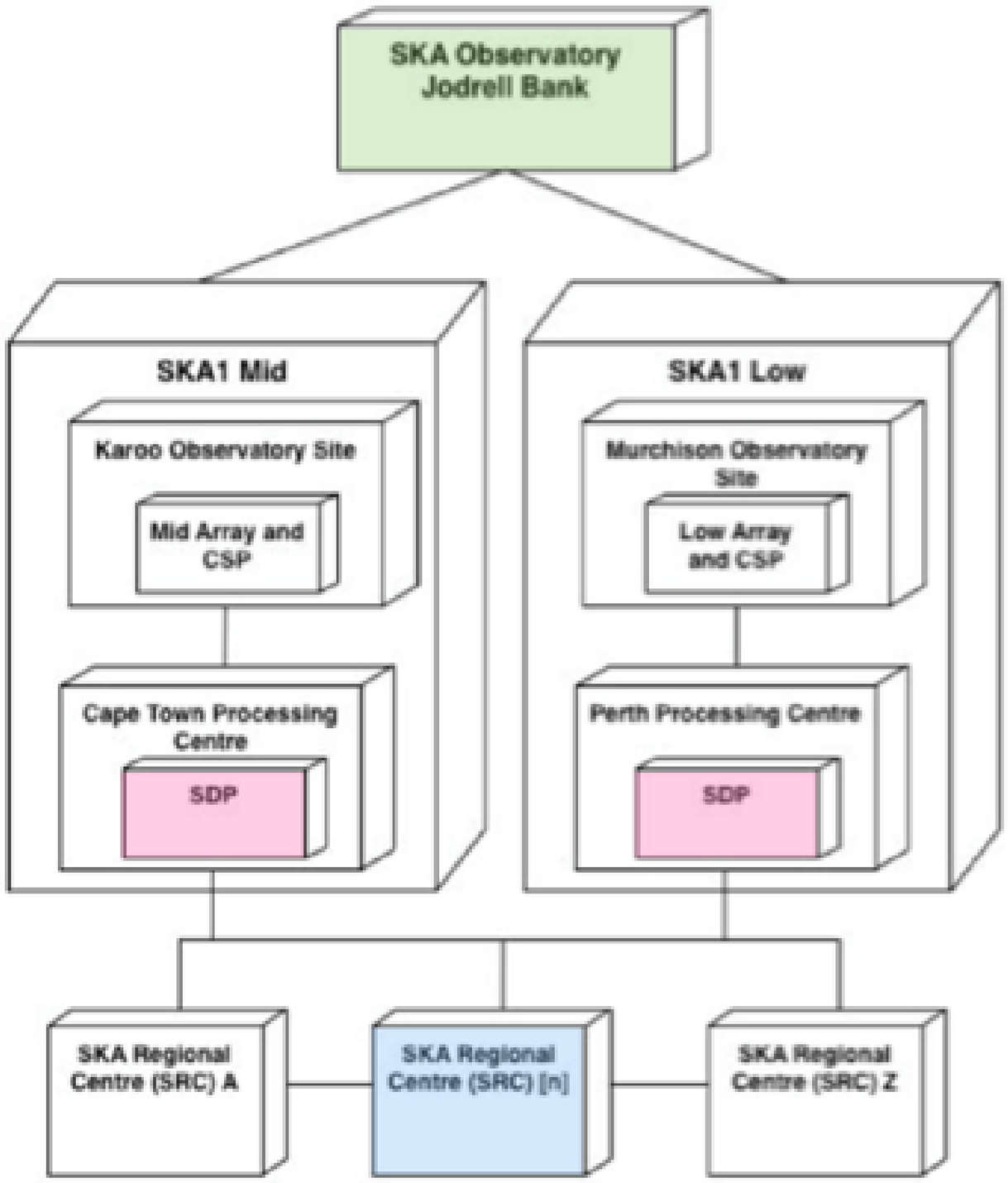

Aeneas
Advanced European Network of E-infrastructures
for Astronomy with the SKA



CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE



CENTRAL SIGNAL PROCESSOR

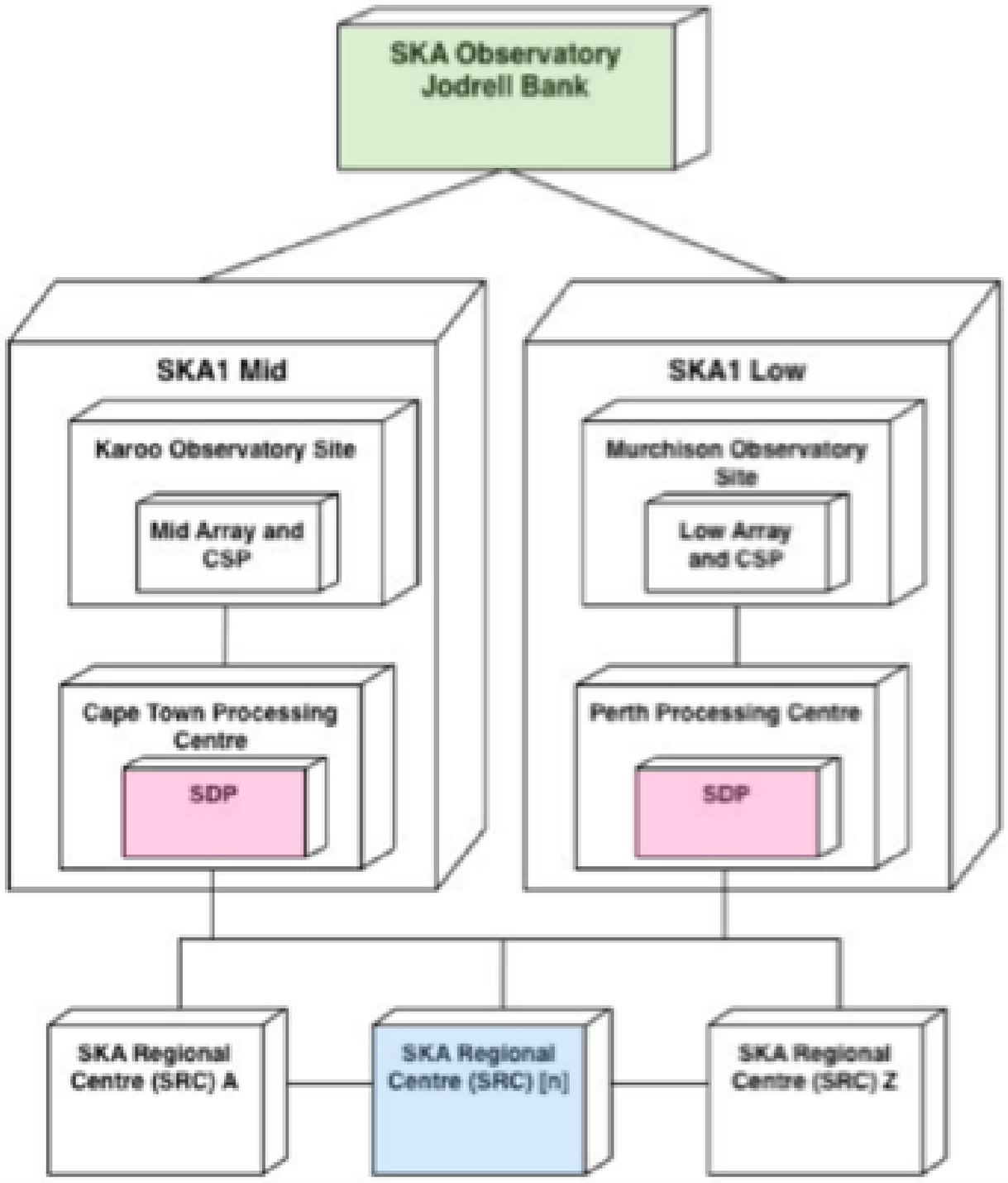
CENTRAL SIGNAL PROCESSING

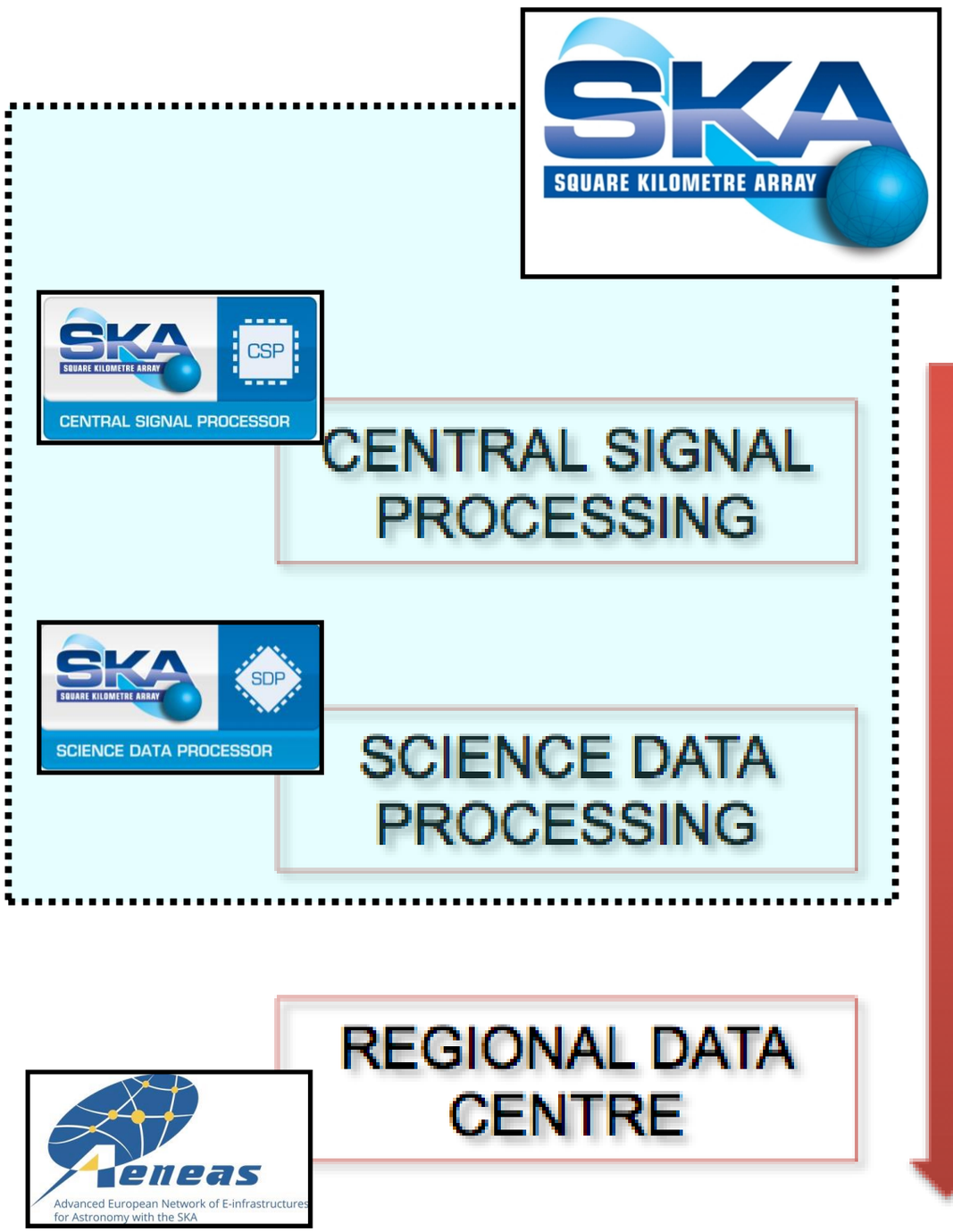
SCIENCE DATA PROCESSOR

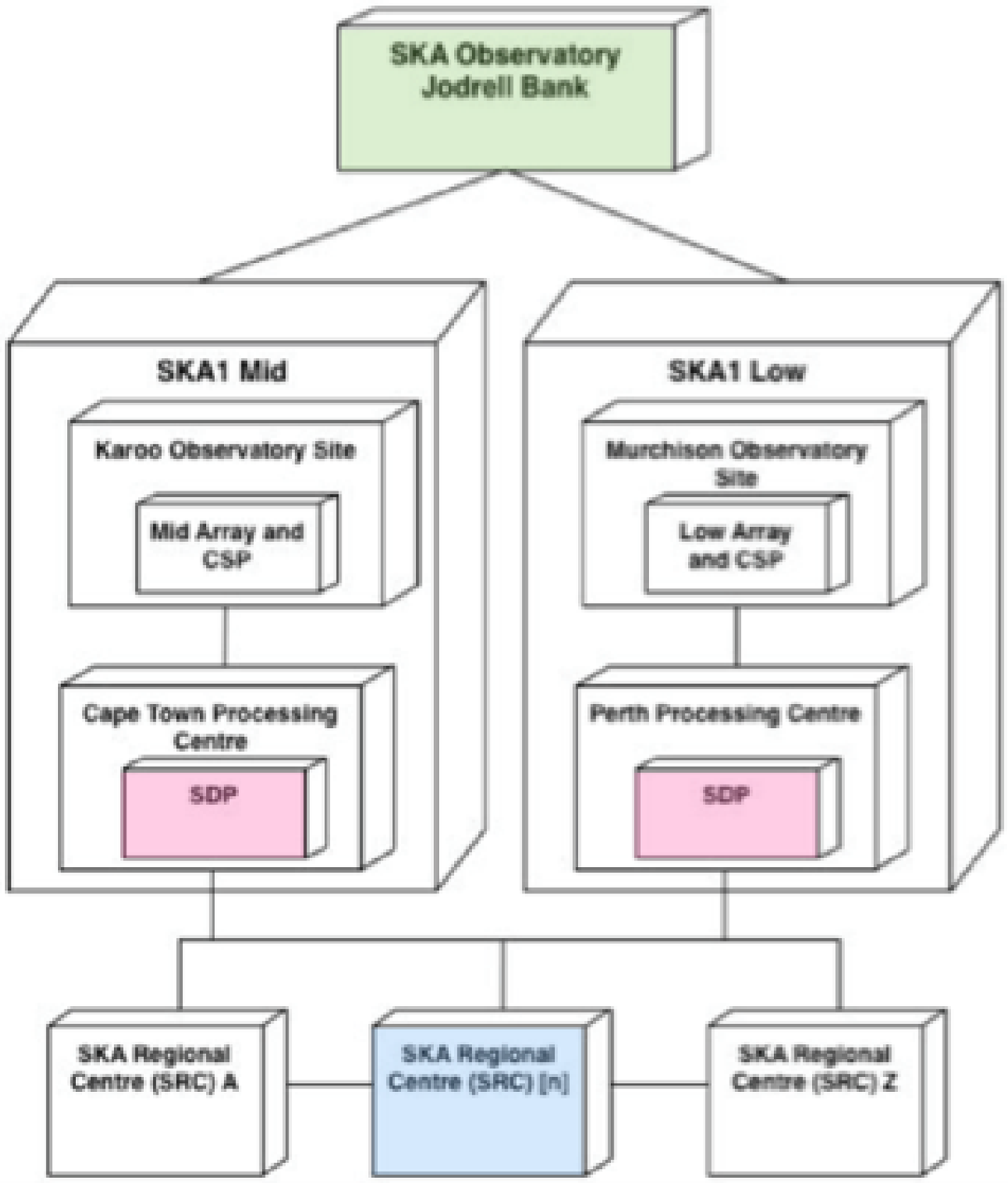
SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE

Advanced European Network of E-infrastructures for Astronomy with the SKA


SKA Observatory
Jodrell Bank
SKA1 Mid
Karoo Observatory Site
Mid Array and CSP
Cape Town Processing Centre
SDP
SKA1 Low
Murchison Observatory Site
Low Array and CSP
Perth Processing Centre
SDP
SKA Regional Centre (SRC) A
SKA Regional Centre (SRC) [n]
SKA Regional Centre (SRC) Z

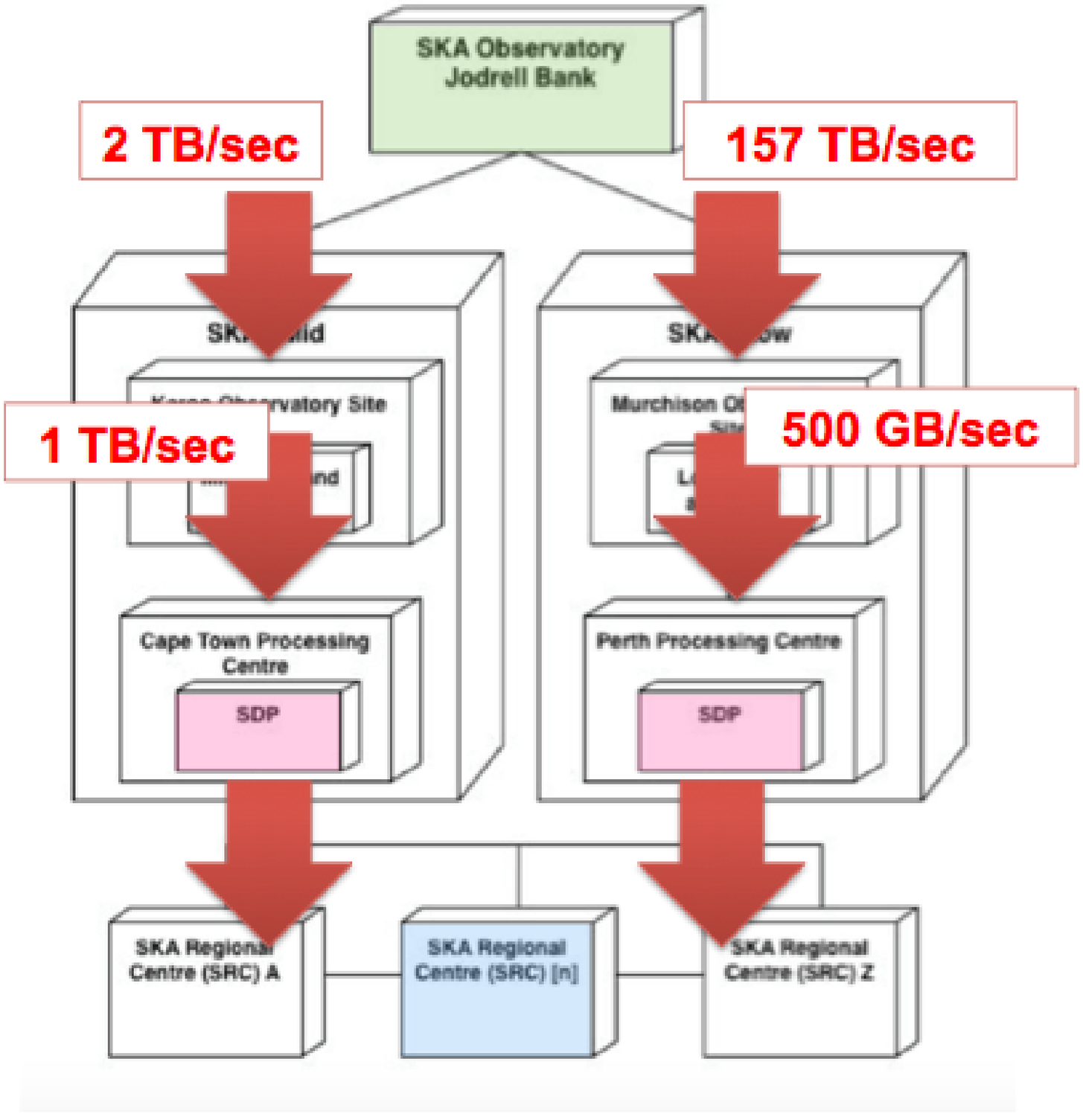

SKA
SQUARE KILOMETRE ARRAY
CSP
CENTRAL SIGNAL PROCESSOR
CENTRAL SIGNAL PROCESSING
SDP
SCIENCE DATA PROCESSOR
SCIENCE DATA PROCESSING
REGIONAL DATA CENTRE
Advanced European Network of E-infrastructures for Astronomy with the SKA



CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

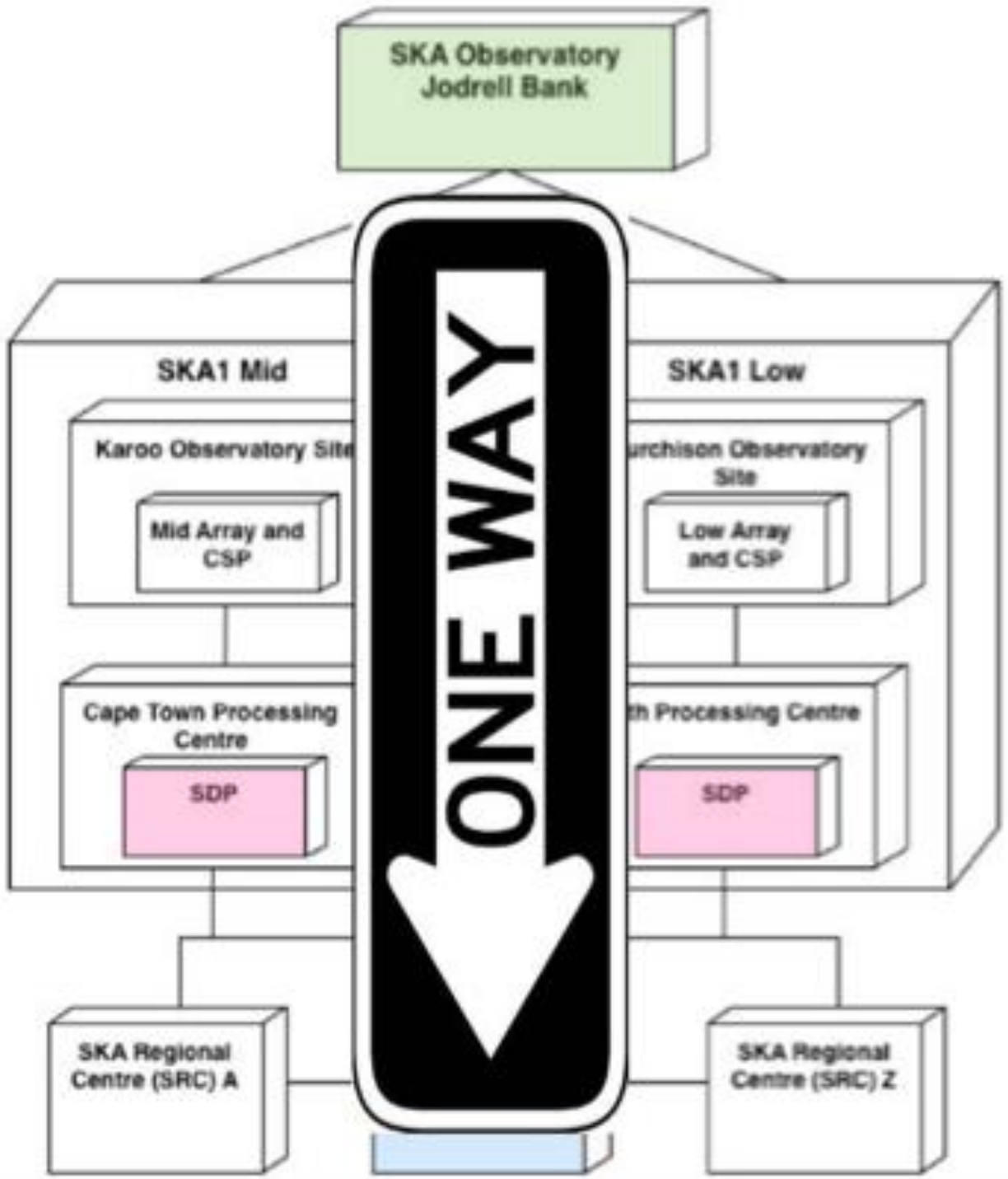
REGIONAL DATA CENTRE



CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE



Standardized data products

A standard SKA1-MID image data product has
30k x 30k pixels

SKA1 will have up to **65k frequency channels**
and **4 polarisations**

At 4 Bytes per voxel that equates to
30k x 30k x 65k x 4 x 4
= **936 TeraBytes**

Advanced European Network of E-infrastructures
for Astronomy with the SKA AENEAS - 731016

Future SKA Science Archive

2017
2023

uploads to
facebook.
180PB

searches on
Google
98PB

CERN
73PB

LOFAR
Long Term Archive
25PB

YouTube
15PB

SKA
Phase1 Science Archive
300PB

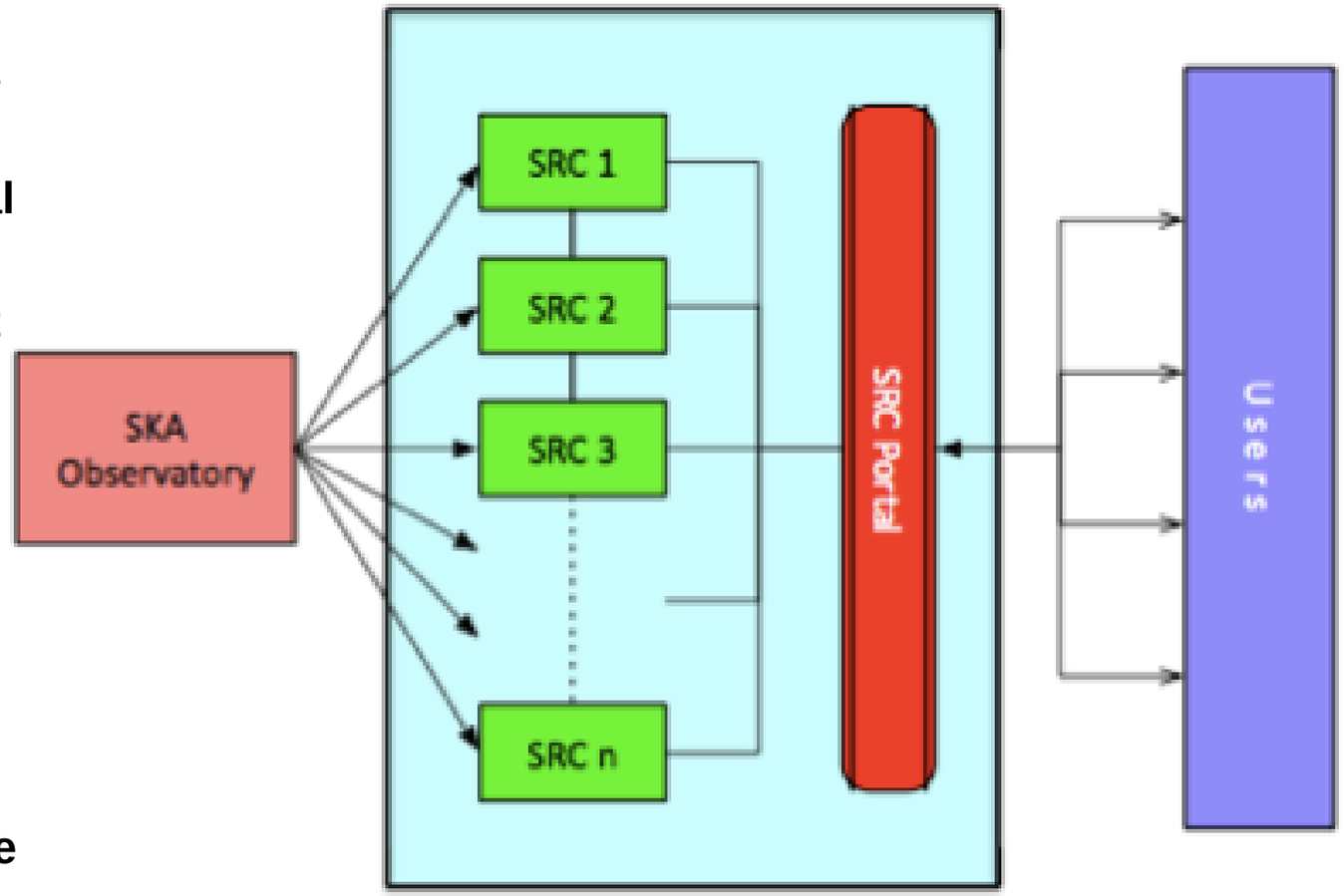
PER YEAR
- 1 Petabyte

SKA Regional Centres

- **Provision long-term SKA Science Archive**
- **Provide access and distribute data products to users**
- **Provision and Management of computational resources for post-processing**
- **Provide platform for continued development of software**
- **Provide user support for SKA Science Archive data products and analysis**
- **Multiple regional SRCs, locally resourced but interoperable**

Joint SKAO/SRC functions

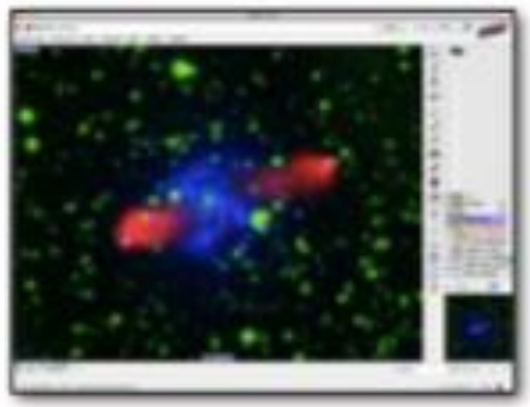
- **User support for SKAO data products**
- **User support for SKAO provided software and tools**
- **Distribution of SKA data packs to users (SDP or SRC)**



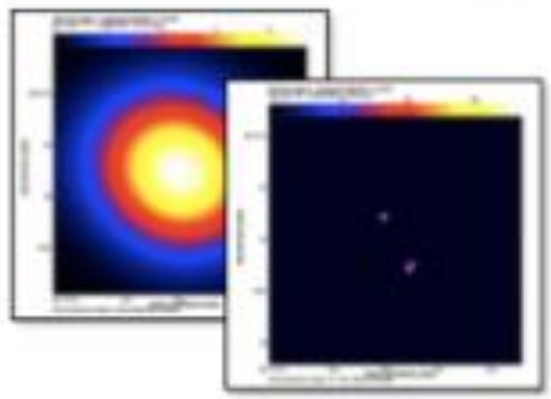
Regional Centre Functionality

Data Discovery

- Observation database
- Quick-look data products
- Flexible catalog queries
- Integration with VO tools
- Publish data to VO



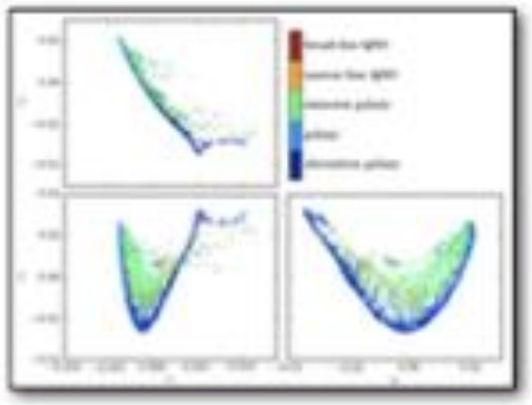
Data Processing



- Reprocessing
- Calibration and imaging
- Source extraction
- Catalog (re-)creation
- DM searches

Data Mining

- Multi-wavelength studies
- Catalog cross-matching
- Transient classification
- Feature detection
- Visualization



Boundary conditions

- **SKA Regional centres must adhere to the data policies as defined by the SKA**
- **SRCs must meet minimum requirements to join the network**
- **An accreditation process for SRCs in the network will be defined by the SKAO**
- **SRCs may be heterogeneous in nature but with common core functionality**
- **Some SRCs may provide additional community specific functionality**
- **SRCs must support the key science project teams as well as general users**
- **Finding the balance between keeping stakeholders happy and keeping data access simple**

Data Products at the Regional Centre

- **Image type data products**
  - **Image cubes**
    - **Continuum Survey, Magnetism, Hi Kinematics, ISM**
    - **Data archive for these experiments would range from a fraction of a PB to 120 PB**
    - **Since hours of telescope time differ, it is useful to look at data generated per 6 hour observation. This will range from 0.1 to 100 GB**
  - **U-V Grid – calibrated visibilities**
    - **EoR experiments on SKA1 LOW**
    - **Data archive of almost 220 PB**
    - **Per Observation ~270 GB**
- **Non-image data products**
  - **Pulsar search and timing experiments**
    - **Data archive of 250 GB to a few PB, per observation less than 3 GB**
    - **LSM Catalogue, Transient catalogue, Pulsar timing solutions, Transient buffer data, Sieved pulsar and transient candidates**

What do we know about data and its consumption

- **Key science based namespaces – further extended to be PI-based**
- **Hierarchical data structure within each experiment**
- **Granularity at which data is managed varies from experiment to experiment**
- **Pre-determined ”push” mechanism from the SDP to Regional centres**
- **Data storage will be at various locations possibly under different administrative domains**
- **Likely lifecycle will be 3 stages:**
  - **Initial flurry of activity when observations come in, steady state as observations are monitored**
  - **Second stage of activity as data is being analysed and secondary data products are being created**
  - **Scientific results published, no more analysis of raw data, but outputs of analysis are available and raw data moved into long term storage**

Who is consuming this data?

- **Number of users would be a few thousand**
- **“Passive users” consuming data will also be generating secondary data which may not be smaller than raw data**
- **Likely X.509 certificates for authentication**
- **Commonly used tools are CASA, AIPS, Miriad, PRESTO, SIGPROC**

Initial Prototyping on GridPP using DIRAC

- **LOFAR data – GOODS-N survey. One observation is 3.5 TB**
- **Uploaded to the Manchester storage nodes, with a replica of a sub-set of the data at Imperial**
- **Accessed using Logical File Names (LFNs)**
- **Calibration using LOFAR software on CERNVM-FS being tested on GridPP**
- **Using DIRAC’s DMS and WMS**
- **GridPP liaisons - Andrew McNab and Alessandra Forti**

Open Questions

- **How will the SKA science archive be distributed across regional centres?**
  - **Defining what constitutes a dataset will help answer this question**
  - **How will the compute requirements for analysis of various experiments determine where/how the corresponding data is stored?**
  - **Where are the users located?**
- **How will data be transported within different locations of the European regional centre?**
- **How can we take optimal advantage of existing infrastructures?**
- **What measures will need to be in place for smooth interoperability across SRCs?**
- **Will there be a global namespace or regional centres will maintain their own?**
- **What replica management policy should be enforced?**
- **What metrics can we use to define data importance (last accessed, compute effort required)**

Slide credits:

Anna Scaife

Thank you!