SKAO REGIONAL CENTRE NETWORK

Data Movement within the Square Kilometre Array Regional Centre network

Rosie Bolton

Head of Data Operations

20/04/2023

SKAO Mission

"The SKAO's mission is to build and operate cutting-edge radio telescopes to transform our understanding of the Universe, and deliver benefits to society through global collaboration and innovation."

Two telescopes, one observatory, Global Collaboration

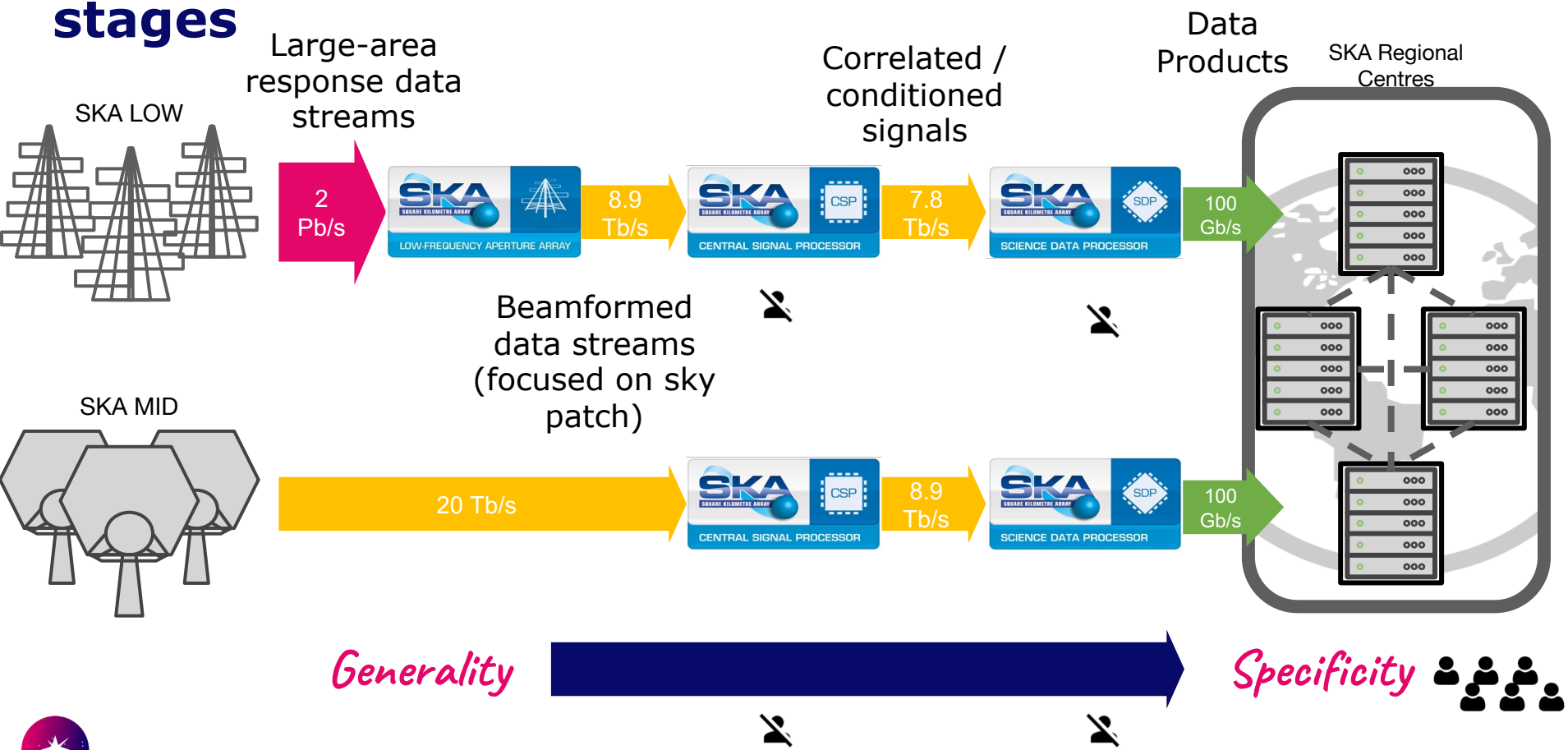
SKA Low

131,000 antennas in the Western Australian desert

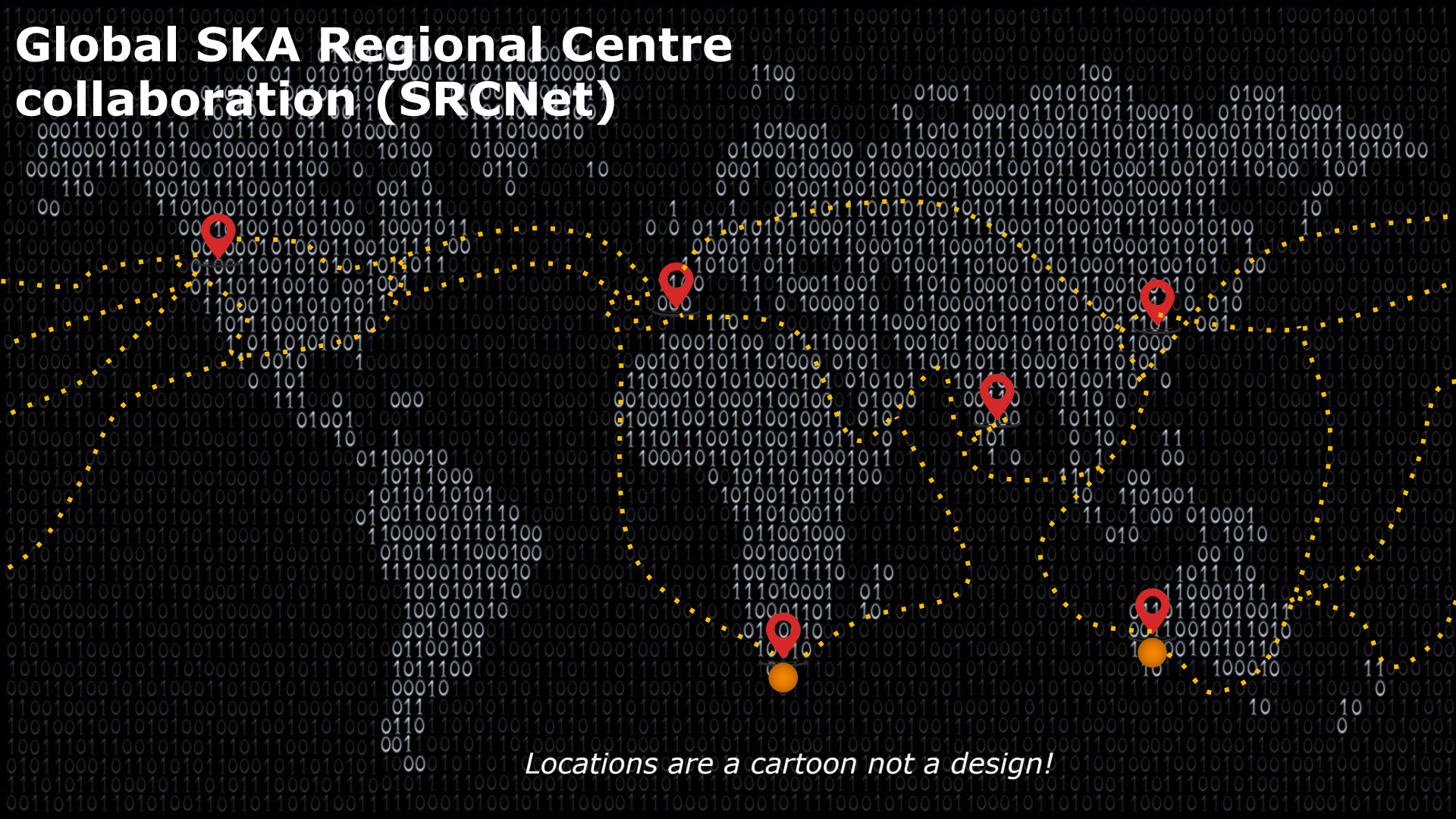
SKA mid

197 dishes in South Africa

SKA Regional Centres: SKAO data processing stages



Global SKA Regional Centre collaboration (SRCNet)



How do users "control" data products?

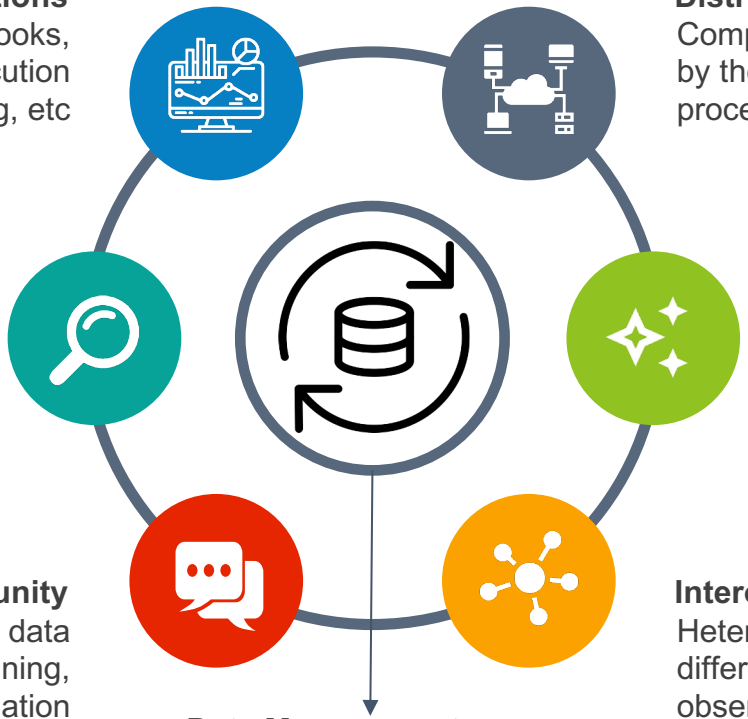
- Need to satisfy users whilst retaining control of SKAO resources
- Science Data Processor is fundamentally part of each SKA Telescope
- Users will specify required data products in Observing Proposals
- All user interaction with data products will be in SRCs

SKA Regional Centre Capabilities

Science Enabling Applications
Analysis Tools, Notebooks, Workflows execution Machine Learning, etc

Distributed Data Processing
Computing capabilities provided by the SRCNet to allow data processing

Data Discovery
Discovery of SKA data from the SRCNet, local or remote, transparently to the user



Visualization
Advanced visualizers for SKA data and data from other observatories

Support to Science Community
Support community on SKA data use, SRC services use, Training, Project Impact Dissemination

Interoperability
Heterogeneous SKA data from different SRCs and other observatories

Data Management
Dissemination of Data to SRCs and Distributed Data Storage

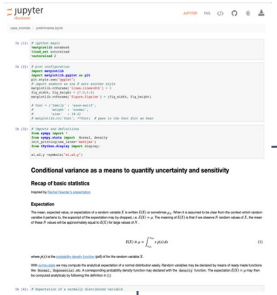
The Role of SRCs: Collaboration platform

SRCs will bridge the gap between the highly data intensive **pre-defined workflows generating SKA data products** in the SDP, and the **iterative flexible, user-led data analysis** required to produce scientific results

SRCs will provide collaborative tools backed up by **powerful compute and data management**

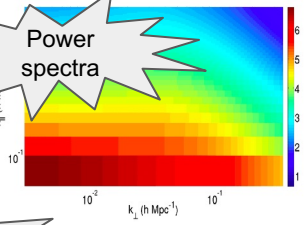
Users will not have access to the SDP or to Raw SKA data!

Workflows



Paul, Sourabh et al. (2016). ApJ 833. 10.3847/1538-4357/833/2/213.

Power spectra



Catalogues / Source List

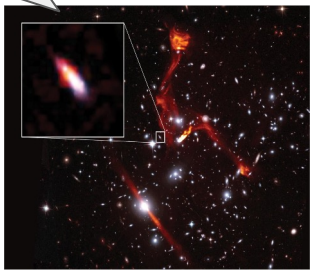
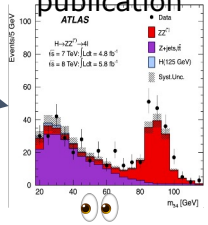


Image cut-outs

Credit: Heywood et al.; Sophia Dagnello, NRAO/AUI/NSF, STScI.

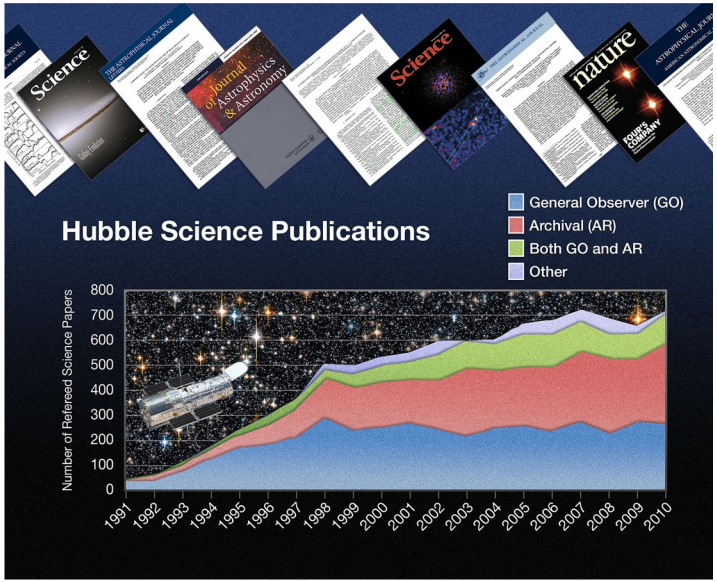
Plots for publication



The Role of SRCs: Support data product (re-)use

Why

- All SKA Data Products will (in time) become public - this is likely to be the biggest science generator (see Hubble)
  - Build SKA science archive around IVOA standards
  - Ensure interoperability with other archives and other experiments



SKA Regional Centers: Data management

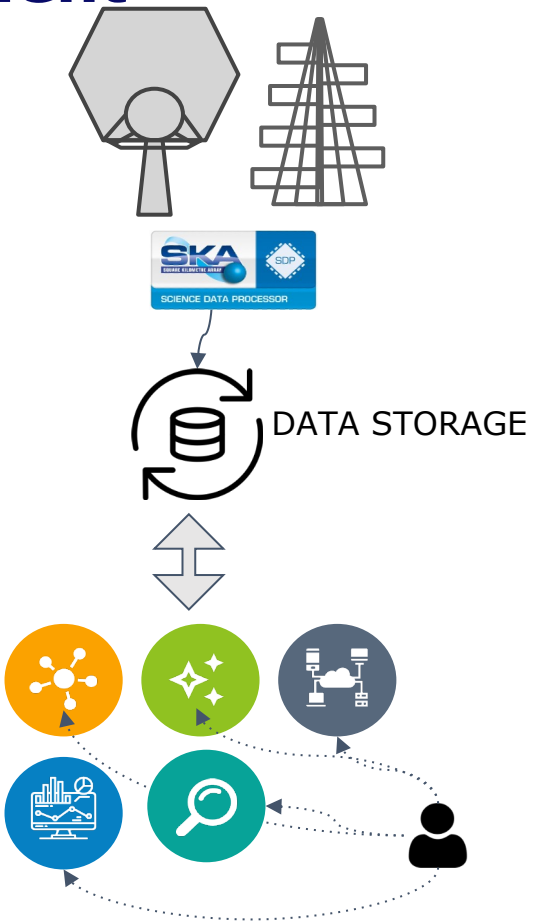
Storing SKAO data growing at up to 700 PBytes each year will be a challenge (plus user-generated data too).

Several million dollars per year in new data, for one copy

Global data management within SRCNet should enable best possible use to be made of available storage resources

Avoid (reduce) unnecessary duplication

Support mirroring of popular data products to enhance user experience

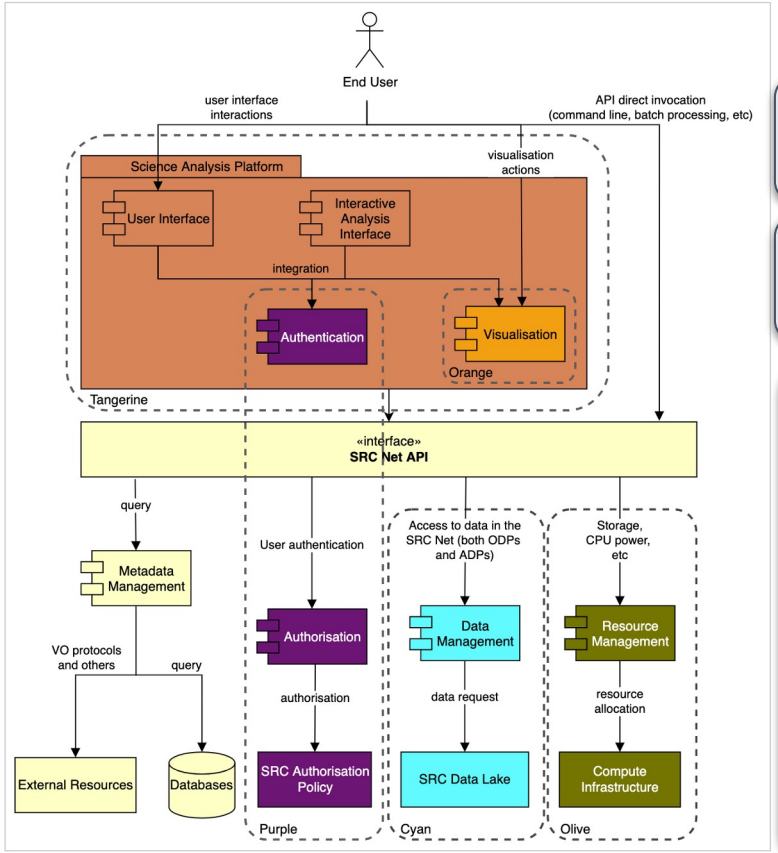


SRC Network global capabilities

Collectively meet the needs of the global community of SKA users

Anticipate heterogeneous SRCs, with different strengths

Defining an architecture (WIP)



Users

Science Platform

Interface (API layer)

Metadata query - Science Data Discovery

Authentication Who? Permissions?

Data Logistics Globally distributed storage sites

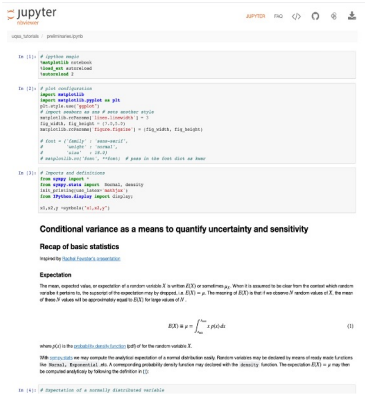
Compute Resource Management Work sharing

Type 1: Planned, Co-located data scenario

Plan data placement ahead of time

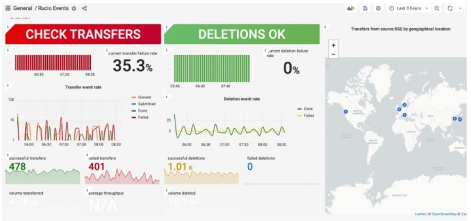
Allocation for compute resources made for projects user group at target SRC

100 Gb/s



SKA Data products generated

Products transferred to target SRC

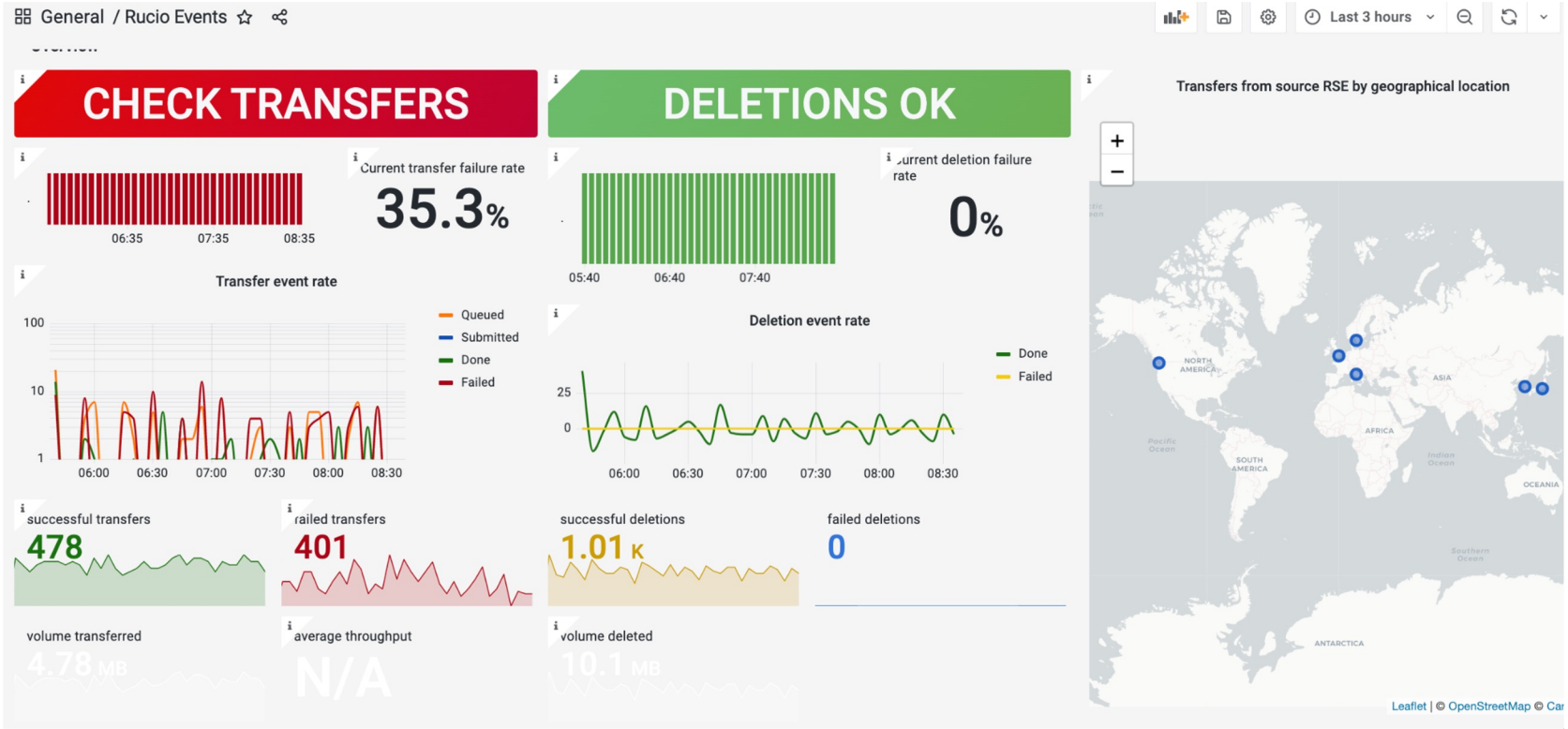


from our live Rucio instance

posed by actors!

Users can be allocated to different SRCs based on data location, HW mapping to compute needs; to even out demand, or other reasons

(Aside) We are prototyping Rucio (continuing our ESCAPE work) for this data placement scenario



Estimates of SKA Data Output... why this is hard!

Co-located scenario is the easiest to predict, but range of SKA science is vast - huge differences in the data volumes required

Examples here are taken from the parametric model used to predict computing needs *within* SKAO, but they also include output data rates

The specific examples are quire out of date - take home message is the range - from <1 to 100 Gbit/s data generation rate, and the fractions of time assumed

SKA1_Low:

HPSO	Time [%]	Tobs [h]	Npix (side)	Channels (DPrepB)	Channels (DPrepC)	Image size [GB]	Non-Vis Rate [Gbit/s]	Visibility Size [TB]	Visibility Rate [Gbit/s]	Total Rate [Gbit/s]
hpso01	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso02a	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso02b	15.6	5.00	18344	500	1500	2.7	8.5	205.8	91.4	99.9
hpso04a	39.8	0.67	-	-	-	-	0.7	-	-	0.7
hpso05a	13.4	0.67	-	-	-	-	2.6	-	-	2.6
Average	-	-	-	-	-	-	4.6	-	42.8	47.4

SKA1_Mid:

HPSO	Time [%]	Tobs [h]	Npix (side)	Channels (DPrepB)	Channels (DPrepC)	Image size [GB]	Non-Vis Rate [Gbit/s]	Visibility Size [TB]	Visibility Rate [Gbit/s]	Total Rate [Gbit/s]
hpso04b	1.0	0.17	-	-	-	-	2.3	-	-	2.3
hpso04c	3.1	0.17	-	-	-	-	2.3	-	-	2.3
hpso05b	2.1	0.25	-	-	-	-	6.9	-	-	6.9
hpso13	6.5	8.00	25339	160	3200	5.1	4.2	-	-	4.2
hpso14	2.6	8.00	18814	300	5000	2.8	2.8	-	-	2.8
hpso15	16.5	4.40	10837	260	2500	0.9	0.8	-	-	0.8
hpso18	13.1	0.02	-	-	-	-	0.1	-	-	0.1
hpso22	7.9	8.00	110601	1000	0	97.9	48.1	-	-	48.1
hpso27and33	13.1	0.12	23549	700	0	4.4	99.3	-	-	99.3
hpso32	13.1	2.20	-	-	-	-	1.3	-	-	1.3
hpso37a	13.1	3.80	94195	700	0	71.0	60.6	-	-	60.6
hpso37b	2.6	8.00	94195	700	0	71.0	28.8	-	-	28.8
hpso37c	2.6	8.00	94195	700	0	71.0	28.8	-	-	28.8
hpso38a	1.3	8.00	113204	1000	0	102.5	50.4	-	-	50.4
hpso38b	1.3	8.00	113204	1000	0	102.5	50.4	-	-	50.4
Average	-	-	-	-	-	-	28.4	-	0.0	28.4

https://ska-telescope.gitlab.io/sdp/ska-sdp-par-model/notebooks/SKA1_System_Sizing.html

Type 2: Distributed data scenario, server side actions

Moving compute to the data

Server-side actions

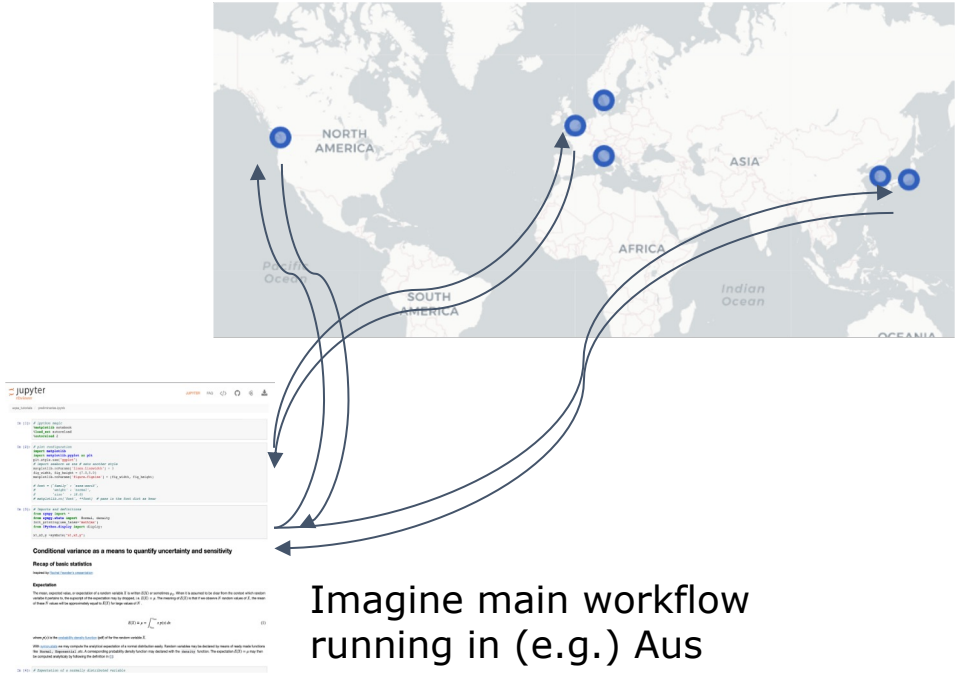
Likely public data

Not SKA Programme Users

Hard to predict

Single workflow running at primary site launching multiple secondary server-side actions at sites holding data

e.g. return image statistics for small area around source location (e.g. re-do source parameter analysis using location from a catalogue), just return the parameters



Imagine main workflow running in (e.g.) Aus

Type 2: Distributed data scenario, server side actions

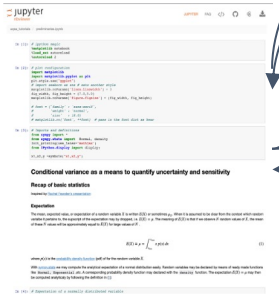
Questions:

How can we predict the network load of this?

Expectation management: can this ever work for interactive jobs (prioritisation??)

Are there similar examples in HEP?

How would this sit alongside the bulk data distribution (planned transfers, known targets, rule-based async)?



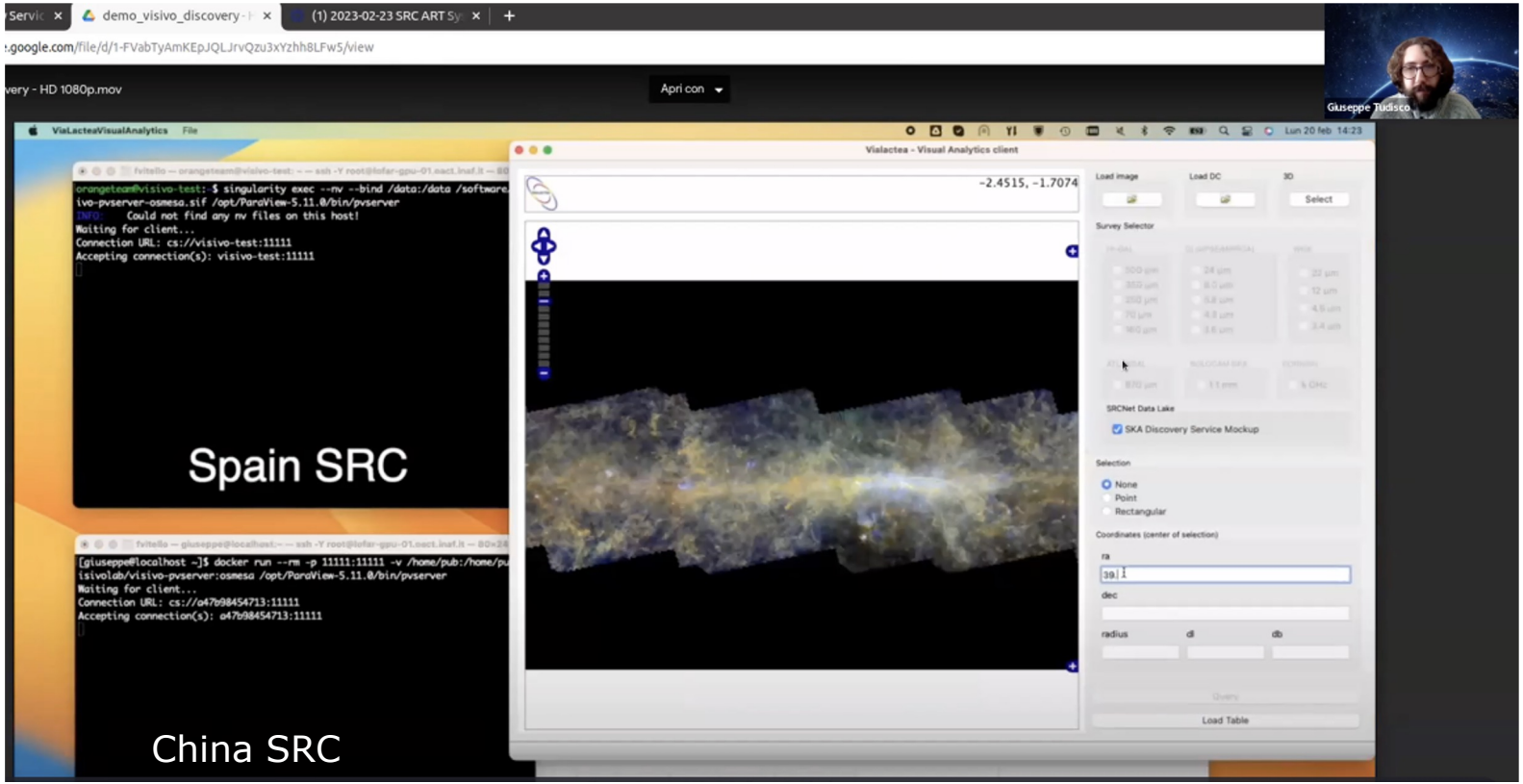
Imagine main workflow running in (e.g.) Aus

Type 3: Remote Data Visualisation

protected and public data

enable users with session in one site to visualise data from multiple sites

interactive session



Questions for SIG

- **SKA full operations ~2028, but data transfers and live system from 2026**
- How do we collectively prepare our networks to support SKA and HL-LHC data transfers?
- Balance between high-level replication (e.g rucio based, "scientific" data transfers) vs detailed tuning (protocols, DTNs, specific paths)?
- What's the hardest problem? (Type 1, 2, 3)
- Can we (do we need to) mimic / run tests with server-side data operations and results/ visualisation data transfer?
- *Mapping to Optical Private Network vs Open Network Environment? What are the advantages of making an SKAONE vs effort required?*
- Which models for data distribution are sensible / need testing? (Degree of coordination of transfers vs asynchronous queueing?)
- So far we working assumption of 100Gbit/s (x2 sites) because looks "big enough" (Science ~25-50Gbit/s) - are we aiming too low? or is this sensible for now?