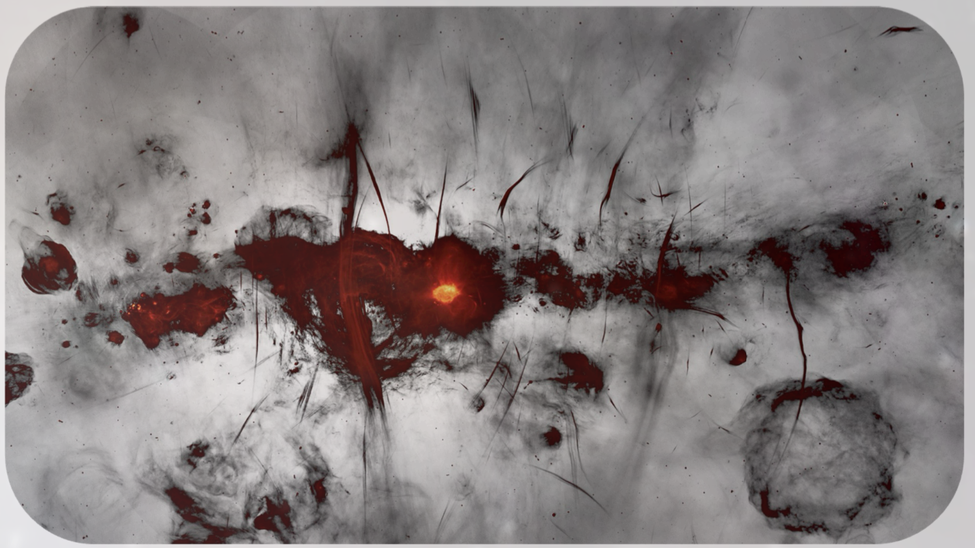
Science and Technology Facilities Council

UKSRC: Data distribution modelling and current testing of the SKA Regional Centres

James Walder
STFC-RAL / UKSRC
IRIS-TWG 20/May/2025
(With inputs from the UKSRC and SRCNet community)

SKA

- The Square Kilometre Array (SKA) is a next-generation radio telescope aimed at being the largest and most powerful ever built
- SKAO (Square Kilometre Array Observatory) is the **governing body** and **Intergovernmental Organisation** responsible for managing the global collaboration, scientific goals, and technical aspects of the SKA project
    - Global project involving over 10 countries, with construction taking place in Australia (SKA-Low) and South Africa(SKA-Mid).
    - **> 50 year lifetime**
- Two Complementary Telescopes:
    - **SKA-Mid** (350 MHz – 15.4 GHz): 197 dish antennas
        - located in the Karoo Desert, South Africa.
          Maximum baseline of O(150) km.
    - **SKA-Low** (50 MHz – 350 MHz): Over 131,000 dipole antennas
        - grouped into 512 stations in Western Australia, covering a maximum distance of O(60+) km.



Credit: I. Heywood, SARAO

- The SKA Regional Centres (SRCs):
    - SRCs will receive data from the SKAO and act as the scientific archive for SKA data.
    - Global Distribution: SRCs will be located across the world, creating a global network that provides the computational power, data storage, and tools needed for international collaboration.

SKA-Low's first glimpse of the Universe

- 17 March 2025
- Image from the Australia
  - First four connected SKA-Low stations
  - 1,024 of expected 131,072 antennas
- 150,000 pixels in current image to ~ 21 million pixels at final construction

"This image was taken using the first four completed antenna stations at Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory. Produced using only 1,024 of the planned 131,072 antennas – less than one per cent of the full telescope – it shows an area of sky equivalent to approximately 100 full moons. 85 of the brightest known galaxies in the region can be seen. It's calculated that the completed SKA-Low will eventually be sensitive enough to show more than 600,000 galaxies in the same frame."

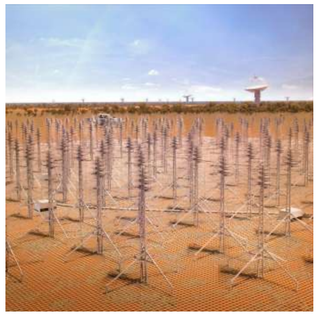
Radio image of ~ 25 square degrees

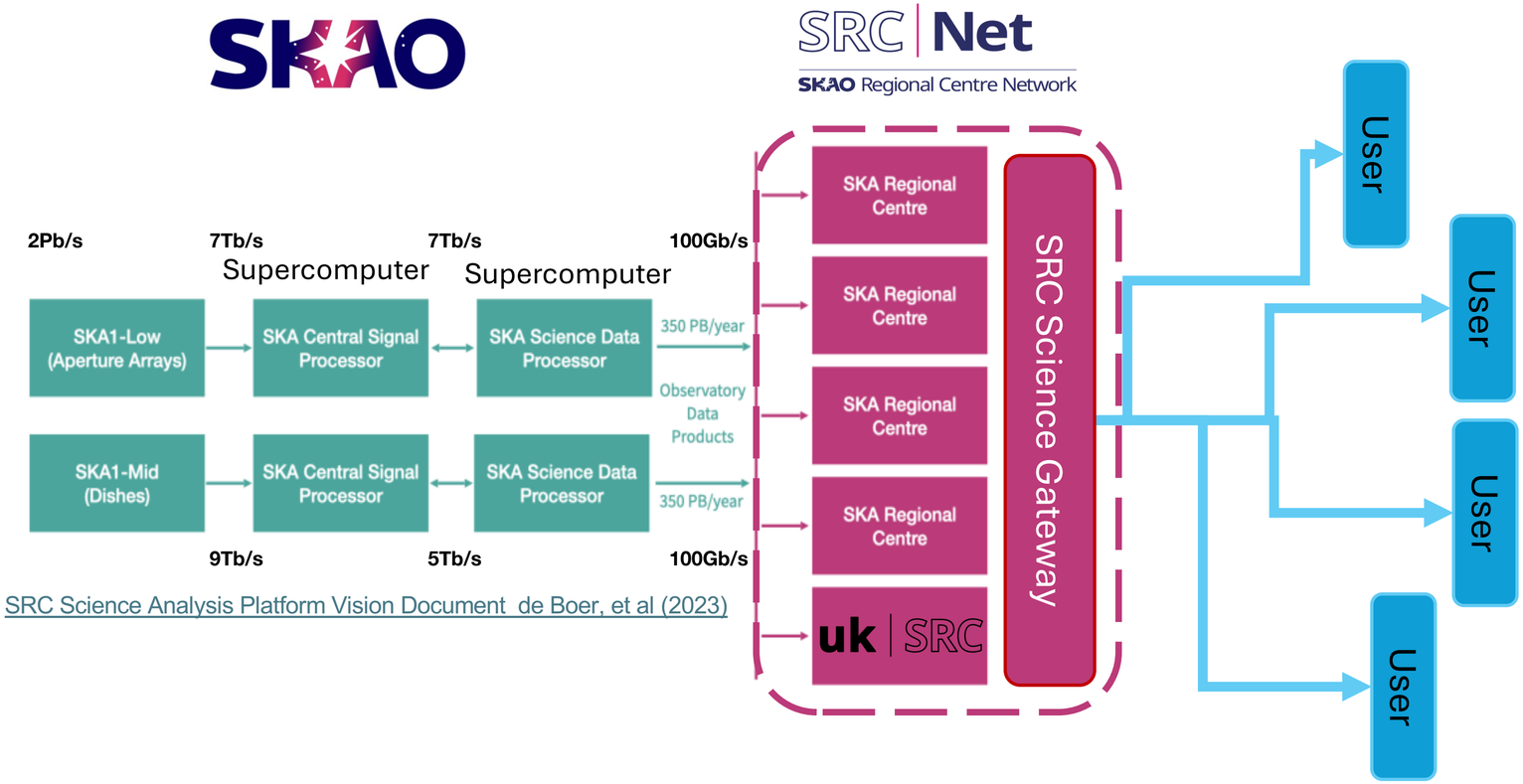
Each 'dot' a Galaxy, contains a supermassive black hole

https://www.skao.int/en/news/621/ska-low-first-glimpse-universe

SRCNet

- SDP (Science Data Processor) to provide calibrated data out to the community
  - SRCnet to archive, process, analyse that data
    - Strong overlaps with Analysis Infrastructure work in HEP





SRC Science Analysis Platform Vision Document de Boer, et al (2023)

....SRCNet is the gateway for the science user communities to access the SKAO data and do science...

UKSRC: Maximising the return on SKAO Investment

UNIVERSITY OF CAMBRIDGE · THE UNIVERSITY of EDINBURGH · Durham University · University of Hertfordshire UH · MANCHESTER 1824 The University of Manchester · UCL · UKRI Science and Technology Facilities Council | Scientific Computing

UKSRC

DRI Infrastructure and Services

Developing digital research infrastructure

Bespoke UK-based computational and data facilities, tools, and services will contribute to the analysis of 700PB of data generated per year by the SKA telescopes.

UK Science community

Strengthening the UK astronomy community

UK astronomers will have opportunities to inform the UKSRC's development and to enhance their skills in preparation for the deployment of the SKA telescopes.

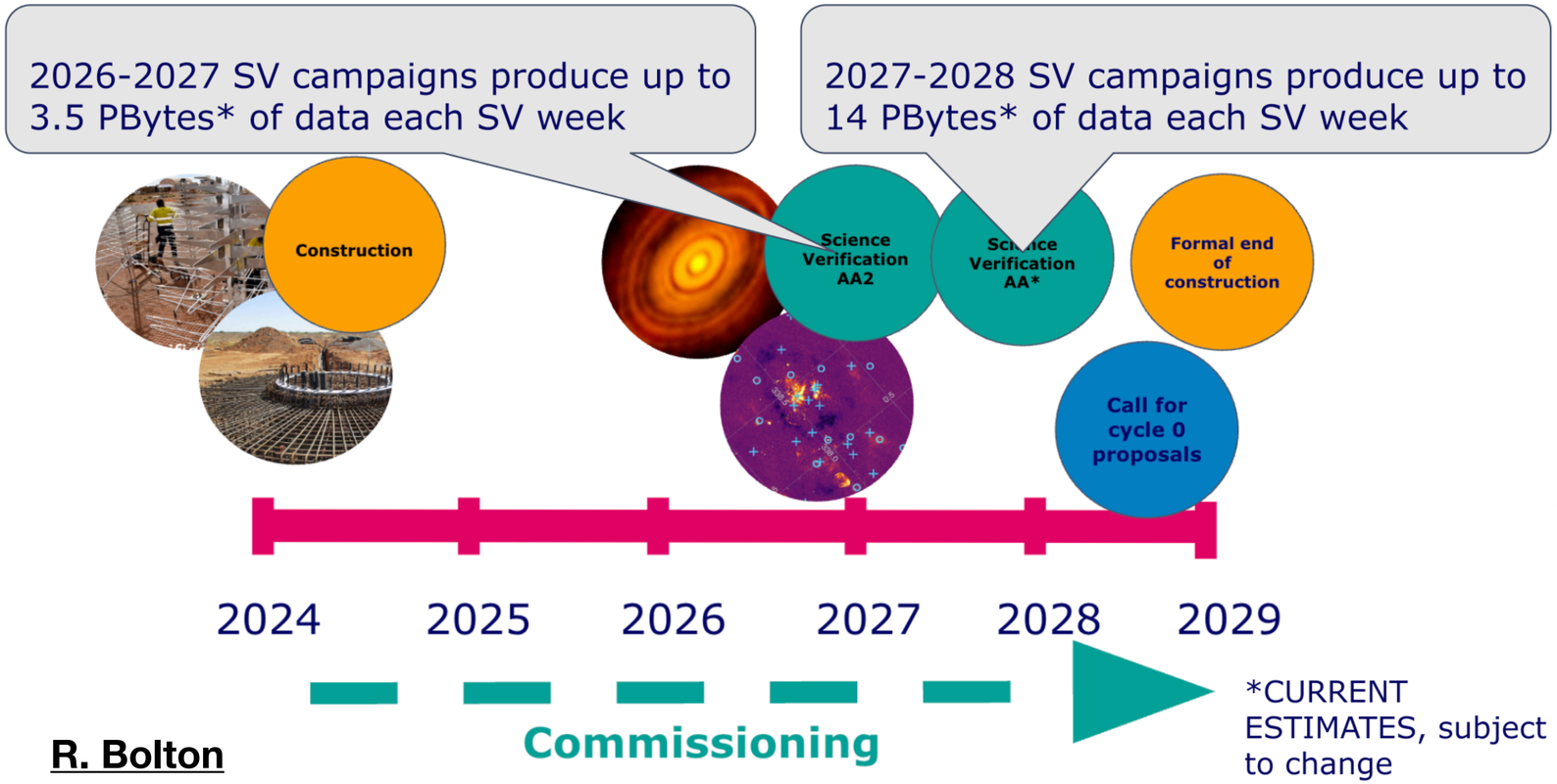
Global SRCNet

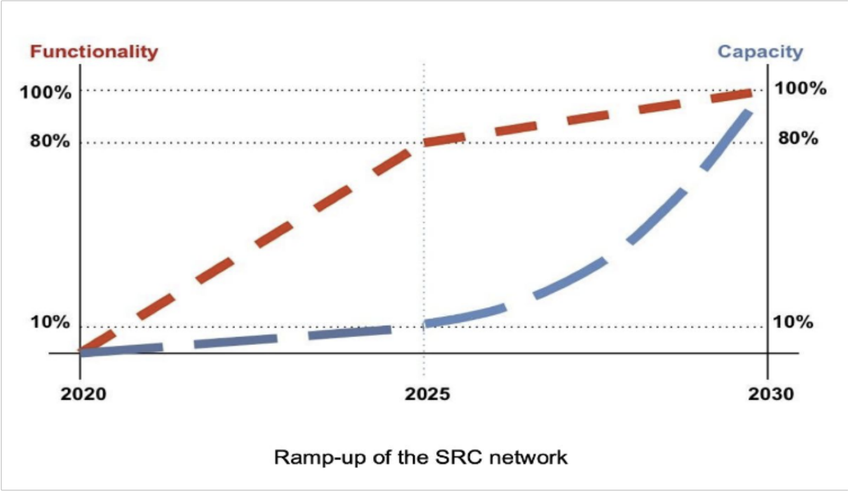
Collaborating internationally

The UKSRC team working with a global network of 14 nations and the SKA Observatory to develop interoperable functionalities to find, access, manipulate and visualise SKA Data products.

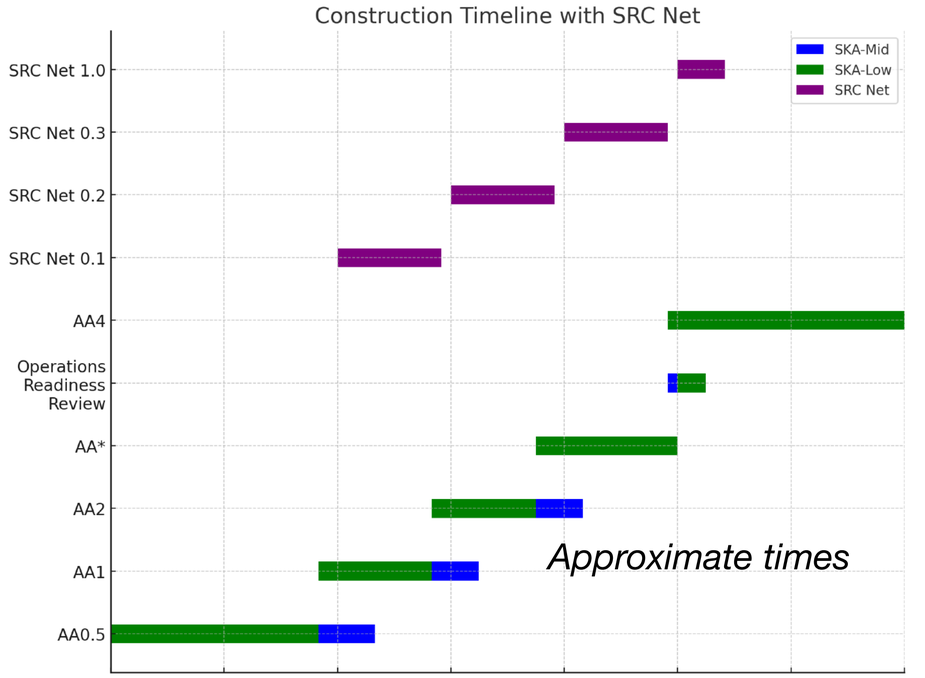
SRCNet Timeline

- Build up capabilities within SRCNet before needing to scale out in capacity





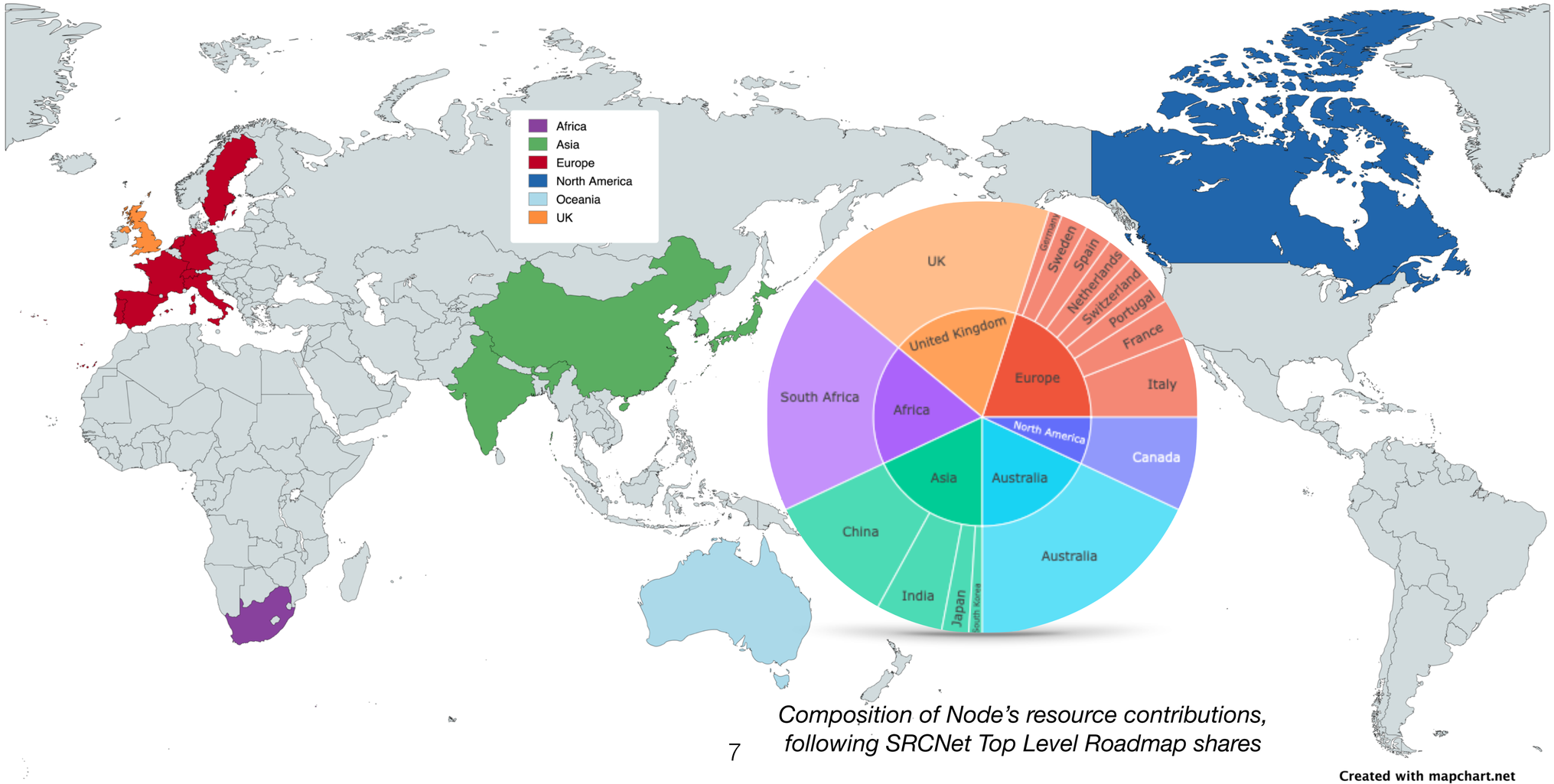
Ramp-up of the SRC network



R. Bolton

Regional Centre shares

- Current modelling assumes a Regional Centre share of storage proportionate to their construction share:
  - ~ 6 similar Global Regions (Canada smaller)
- Australia, South Africa, and the UK dominate as single country entities



Composition of Node's resource contributions, following SRCNet Top Level Roadmap shares

Global Data Challenges

- AA* assumptions for ~2030 operations:
  - ~ 300 PB / year combined telescope output

- Main assumptions for SRCNet:

RAW data too large to keep

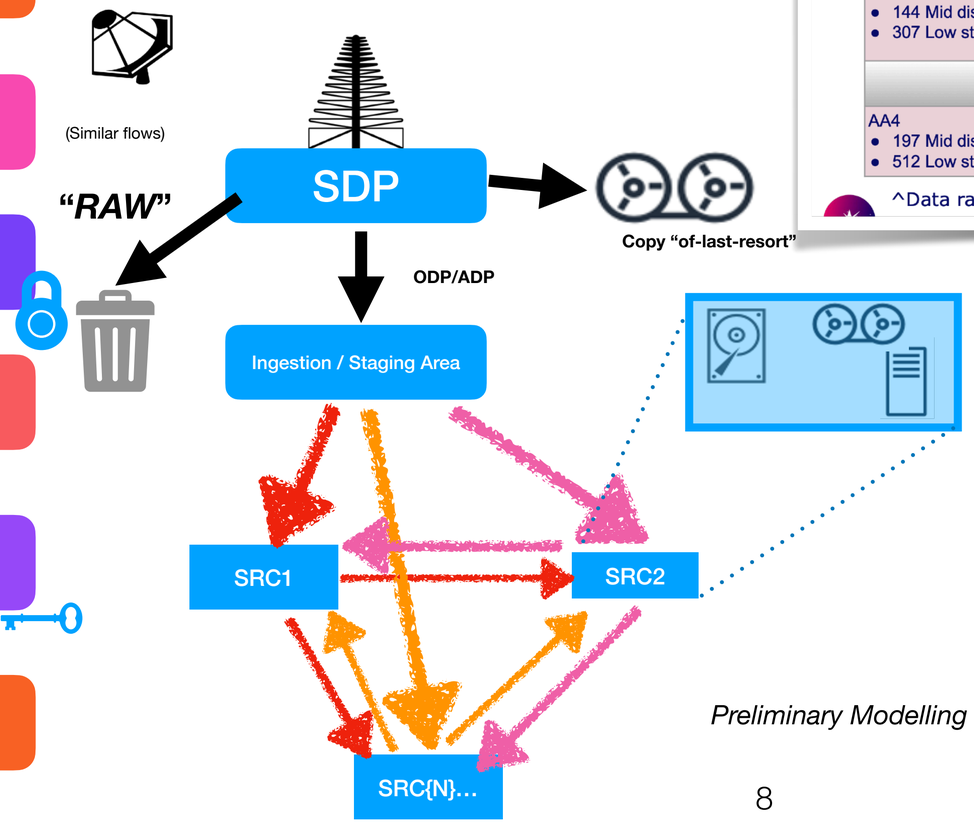
2 Copies on Disk for first year in Datalake

Proprietary access for Project Scientist

Older data archived to TAPE (Nearline)

Open Access to Data after embargo period

ADPs also kept in the datalake



*these numbers should be used as a guide only - email Shari.Breen@skao.int for further information about ongoing work

- Numbers refer to data to be delivered to the science community via the SRCNet
- Data rates calculated assuming a network uptime of 90% with a network overhead of 20%

Milestone	Year	Primary activity	Estimated data rate	
			Low	Mid
AA2 • 64 Mid dishes • 64 Low stations	2026 - 2027	**Science Verification** - observed in dedicated ~week long blocks + single observations interspersed throughout. A higher rate of raw data products will be included at this stage.	1.5 PB/week^ 29 Gbps	2 PB/week^ 39 Gbps
AA* • 144 Mid dishes • 307 Low stations	2027 - 2029	**Science Verification** - observed in dedicated ~week long blocks + single observations interspersed throughout. A higher rate of raw data products will be included at this stage.	5 PB/week^ 96 Gbps	9 PB/week^ 174 Gbps
AA* • 144 Mid dishes • 307 Low stations	2029 +	**Operations** - Observation cycles, starting with shared risk observing, building to successful science observations ~90% of the time	130 PB/year 48 Gbps (1 to 210 Gbps range)	170 PB/year 63 Gbps (0.05 to 300 Gbps range)
Target is to deliver the SKA Baseline Design but the details of this transition between AA* and AA4 are TBD				
AA4 • 197 Mid dishes • 512 Low stations	2030 +	**Operations** - full SKA baseline design	216 PB/year 80 Gbps	400 PB/year 148 Gbps

^Data rates refer to dedicated Science Verification observing weeks, not an average over a year

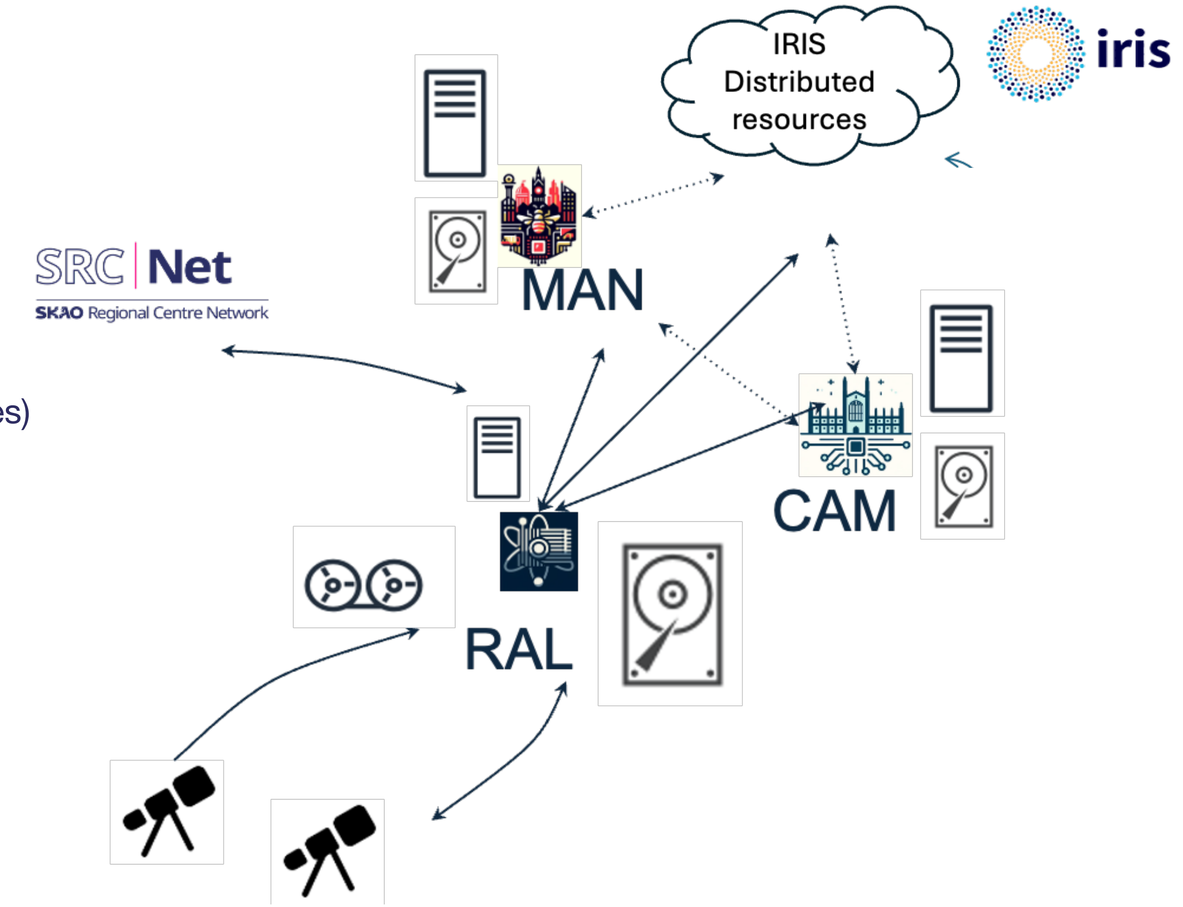
LHCOPN+SKA Meeting March 2025

Preliminary Modelling

- Networks themselves not seen as the major bottleneck:
  - Storage may be the constraining factor.

UK Data Distribution Modelling

- UKSRC will be the largest (non telescope) single country:
  - Expect 100–200 PB/yr at full operations of data ingress (primary + replicas from other SRCs).
  - Tape to grow ~ linearly
  - Disk growth slower (except for AA* to AA4 upgrade)
- Model assumptions:
  - RAL is primary data ingress location in the UK:
    - Majority of bulk data
    - Tape storage
    - Sufficient compute for tasks needing access to bulk data:
      - (e.g. Cutout (data-reduction) services, ‘general purpose’ activities)
  - ‘**core**’ compute sites: e.g. CAM and MAN:
    - compute centres:
      - Sufficient ‘fast’ storage for analysis-level QoS
      - Possibly some ‘bulk’ for staging and resilience.
  - “**Ephemeral**” and opportunistic resources:
    - Access to “compute-only” DRI / e-Infrastructures, e.g. Dirac,
    - Specialised compute needs / boosted resources
    - Mechanisms for data ingress and egress.
  - “**Testbed**” infrastructure (e.g. at UCL). Specialised hardware etc for benchmarking and profiling



Data Ingress

- Simple Notebook created to model data rates:
- Telescope volume as inputs
- Data replication parameters
- Network overheads / retries / peak capacity
- Node size (%)
- **Example**:
  - 300 PB telescope output
  - ~ 180 PB of disk required for a 20% share Node
- Peak Ingress rates of O(200)Gb/s may need to be supported
  - ~ O(100) Gb/s for Egress
- This does not account for workflow and intra-node traffic

	Input
Annual Disk Volume [PB]	900.00
ADP size fraction	0.50
Annual increase fraction	0.00
Year count	0.00
Number of Replicas	2.00
Data Consolidation / retry factor	1.50
Network overheads factor	1.25
Peak capacity factor	2.00
Network Bandwidth factor	2.00

Share [%]	**1**	**2**	**5**	**10**	**15**	**20**
Disk [PB]	9.0	18.0	45.0	90.0	135.0	180.0
Ingress Data [PB]	9.0	18.0	45.0	90.0	135.0	180.0
Egress Data [PB]	4.5	9.0	22.5	45.0	67.5	90.0
Ingress Rate [Gb/s]	4.3	8.6	21.4	42.8	64.2	85.6
Ingress Peak Rate [Gb/s]	8.6	17.1	42.8	85.6	128.4	171.2
Ingress Peak Bandwidth [Gb/s]	17.1	34.2	85.6	171.2	256.8	342.5
Egress Rate [Gb/s]	2.1	4.3	10.7	21.4	32.1	42.8
Egress Peak Rate [Gb/s]	4.3	8.6	21.4	42.8	64.2	85.6
Egress Peak Bandwidth [Gb/s]	8.6	17.1	42.8	85.6	128.4	171.2

The combined expected volume of data output from both telescope sites over one year.

This is assumed over one year of continuous average output.

Annual data [PB] 300

Increase of data (fraction) per year (compounded).

This assumes the total data on disk is increasing slowly over time.

Annual increase in volume fact... 0.00

Number of years from the start.

This allows for extrapolation into future years from a given starting point.

Year increment 0

The average fraction of ADP compared to telescope input.

Assumed average ADP size relative to total telescope data.

ADP size factor 0.50

The number of replicas that SRCNet should hold on disk.

(Note this includes the primary copy and additional copies of data.)

Replicas (Disk) 2

Annual expected data on disk.

Annual Archive Growth [PB] 900

The percentage of data a given SRC should receive.

This is representative of the contribution of the given SRC.

Percent SRC [%] 19.00

Expected volume of data retransmission due to failures and data consolidation activities.

Fraction of re-tries 1.50

Additional overhead of network operations (e.g. packet headers).

Network overhead (8:10 enco... 1.25

Required SRC Node data disk capacity.

SRC Disk [PB] 171

Disk and Compute Modelling

- **Numbers here are indicative only and a WIP:**

- Baseline Storage assumption:
  - ~ 10% of storage should be allocated for fast storage
  - Allow sufficient space for large data cubes *O*(PB) to be processed at single locations, with data for

Date required

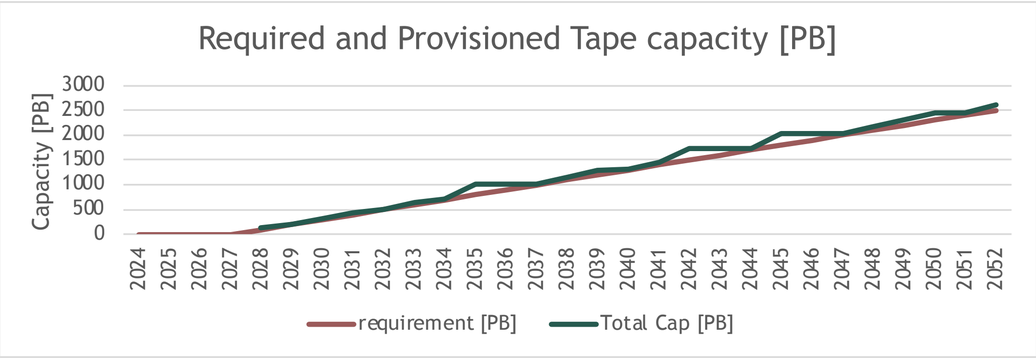
(end of cal. Year)	Bulk Storage [PB]	Fast Storage [PB]	Tape [PB]	Compute [PFLOPS]
2026	4	0.4	–	0.016
2027	18	1.8	5	0.5
2028	80	8	5	1.7
2029	120	12	60	3
2030	180	18	180	7

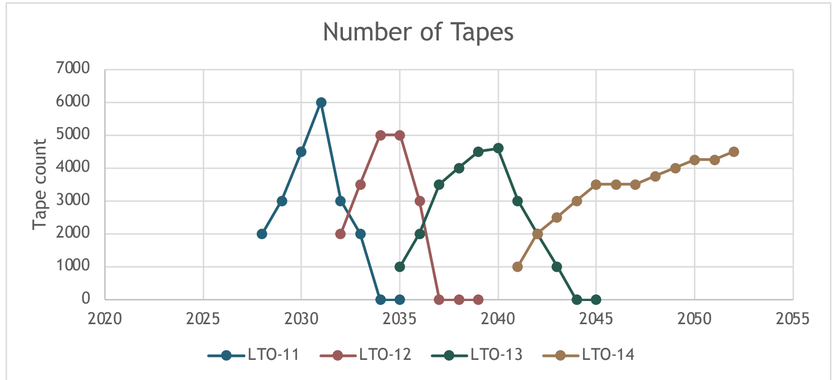
- Compute:
  - Currently based from SRCNet Top Level Roadmap, scaled according to storage needs:
    - Needs refresh - or at least create to justify the process
  - Better mapping of Workflows / science cases to compute needs
  - GPU usage is challenging; more interaction within the science community to understand current and future needs.

Tape Modelling

- Assumption:
  - Write data to tape as soon as it's available
  - (Other models are assessing leaving the tape copy until one disk replica is retired).
  - ADPs also archived:
    - Estimate Mean of 0.5x ODP size:
    - Range could be 0.1 – 3x; workflow dependent
- O(15+) drives likely needed for writes
  - O(2k) (50TB) Tapes per year
  - Recall needs further modelling and SRCNet inputs
  - Assume ~ 1 years worth of recall / year?

	SKA (min)	SKA (max)	UKSRC (min)	UKSRC (max)
Target ODP volume [PB]	300	600	60	120
Avg. ADP Fraction size	0.5			
Archive Volume [PB]	450	900	**90**	**180**
Average Rate [Gb/s]	120	240	24	48
Peak Rate [Gb/s] (assume x2)	240	480	48	96
Drive Speed [Mb/s]	3200			
Drives (Write)	*75*	*150*	**15**	**30**
Tapes [50 TB]	*9000*	*18000*	*1800*	*3600*



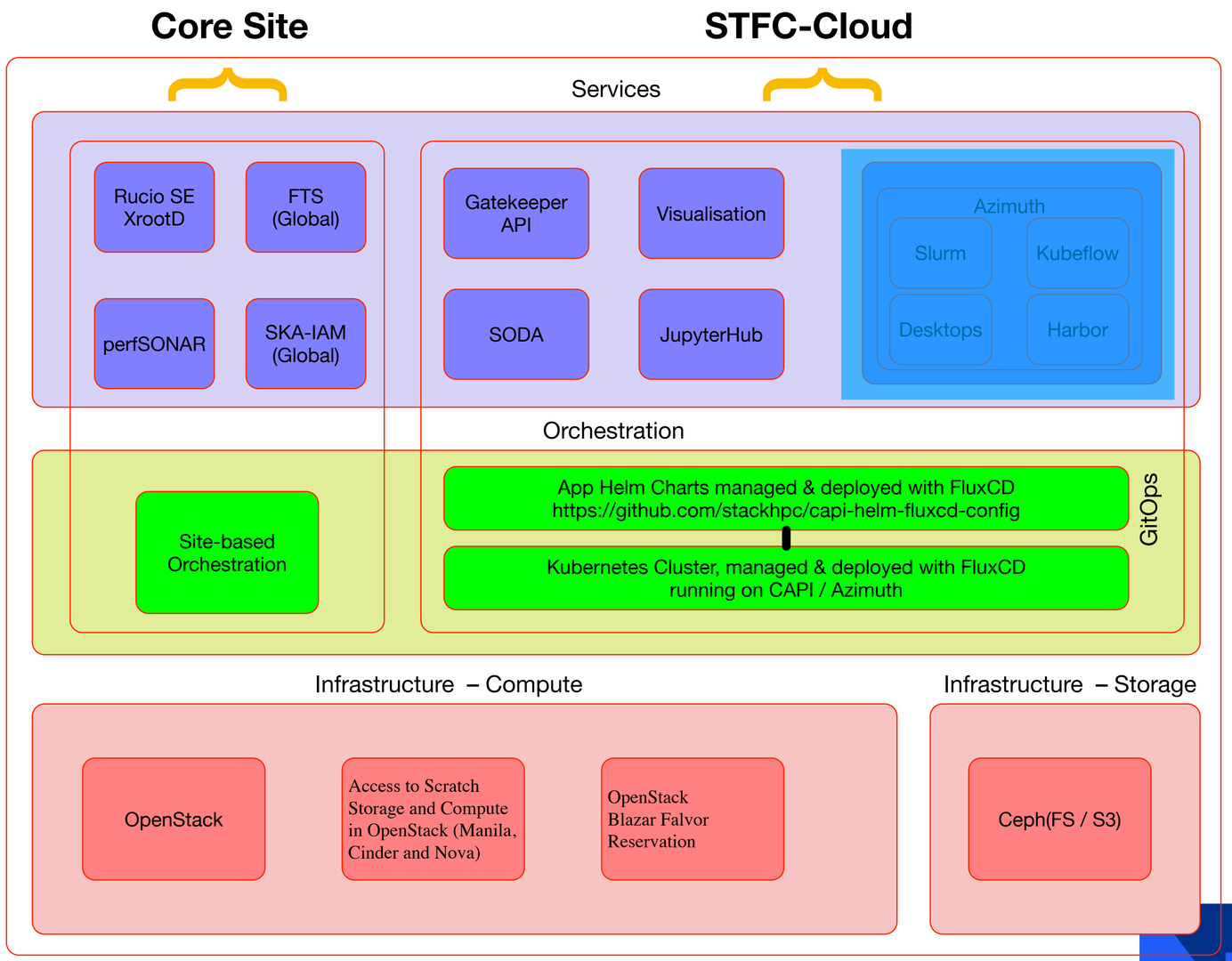


**** Assuming 400MB/s drives.***
Expect 800MB/s available on the timesc

** Tape repacking to be revised*

ukSRC Deployment

- Example from UK;
  - For v0.1, consolidate the deployment at RAL.
  - Teams from across
- Use of existing site deployment tooling for XRootD
  - New, dedicated hardware, Network pod and Ceph cluster
- For local SRCNet Services; Kubernetes based
- Running also Global services, e.g. FTS, SKA-IAM based on site infrastructure
- SKAO also running Rucio on the STFC-cloud



UK Deployment into SRCNet v0.1

- SRCNet v0.1 pledged resources in RAL:
  - 4PB (usable) ceph cluster deployed (commissioning):
    - New network pod:
      - Leaf-spine-supersine topology
      - Active-active 25Gb NICs into Mellanox SN3420, SN3700 switches
    - 13 Storage nodes (22x24TB hard drives)
    - 4xMDS, 3xMons
  - DTN: XRootD servers
    - Active-active 100Gb NICs
    - In the RAL “Science DMZ”
- Jumbo Frames enabled throughput
- IPv4/6 ingress enabled (little ipv4 traffic :( )
- CephFS currently provisioned:
- To explore S3 / Object Store access for the scales required.

- Compute allocation provided via STFC-cloud
  - ~1000 vCPUs

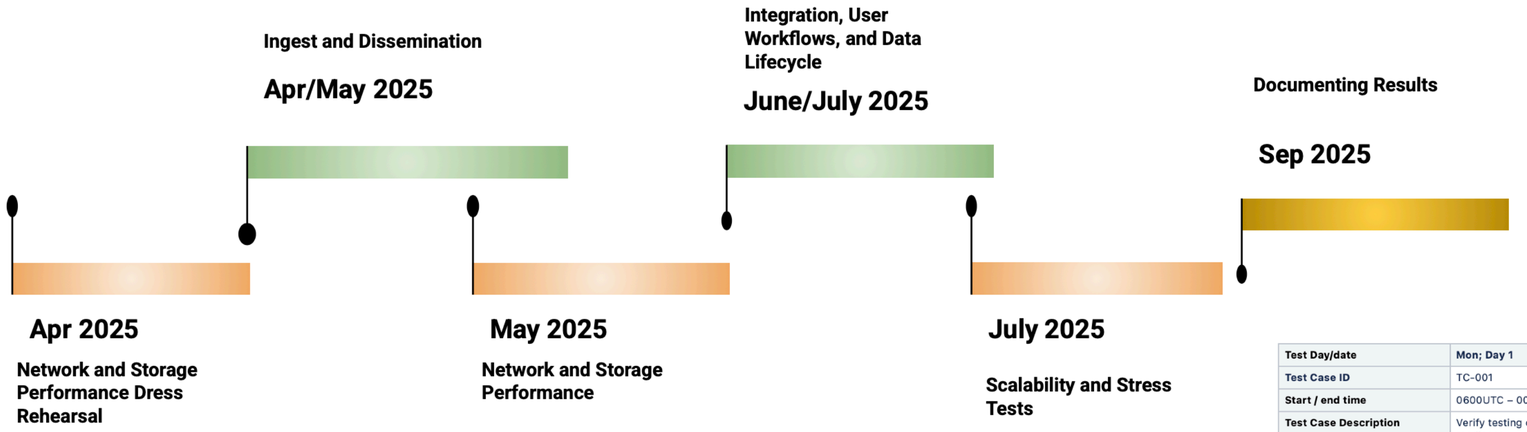
UK developments

- Cambridge:
  - Azimuth platform deployed;
    - Used for much of the current Demonstrator Case work
  - XRootD RSE in commissioning:
    - Testing with VMs using SR-IOV (vs virtIO)
      - (Single Root IO Virtualization)
    - i.e. extracting ~ bare-metal performance from VM infrastructure
    - Could provide scalable infrastructure with resilience cloud environments.
- Manchester:
  - Setting up additional Storage
  - Openstack, similar to CAM and RAL
- SRCNet deployments kubernetes based:
  - Possiblibily to explore kubernetes on bare metal?
- Opportunities:
  - To engage with, e.g. DIRAC, etc to understand data Ingress and Egress solutions to “ephemeral” resource usage.

Start of UKSRC-based perfSONAR mesh-tests

Test Campaigns

- Test Campaigns separated into:



- **Data Movement Challenges**
  - Similar to the (mini) WLCG Data Challenges;
    - Tests of both Capability and Capacity

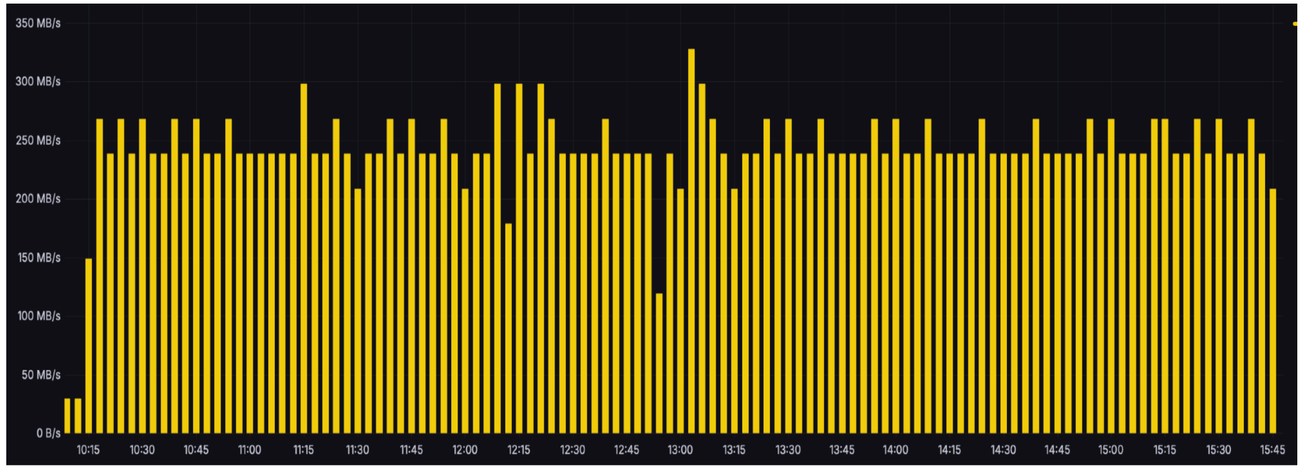
Test Day/date	Mon; Day 1	Tues	Weds	Thurs	Fri
Test Case ID	TC-001	TC-002	TC-003	TC-004	TC-004
Start / end time	0600UTC – 0000UTC (Tues)	0600UTC – 0000UTC (Tues)	0800UTC – 1700UTC	0000UTC – 2359UTC	0000UTC – 2359UTC
Test Case Description	Verify testing code and site readiness	Repeat of Day-1, increased rates	Sustained tests between sites, determined from TP-00{1,2}	Maximise the throughout within SRCNet	Maximise the throughout within SRCNet
Pre-Conditions	• Rucio and critical services running • Data pre-staged at RSEs • Test Plan defined	As Day-1	Rate and matching site characteristics informed from TC-002	As before	As before
Test Steps	Transfers between all links to be used in the tests.	transfers between unique pairs of RSEs for defined rates. Focus on links that can best test each sites Capacity.	Configure full mesh to move data between sites at defined rates	Maximise throughput with strategic set of link transfers	Maximise throughput with strategic set of link transfers (Updated from TC-004)
Expected Results	Successful transfers between participating RSEs, > 70% of possible links covered	Expected volumes of data transfered within the expected period + 20% allowance	Sites can sustain transfers over 3hr periods between links.	75 TB moved in 24/hours	150 TB moved in 24/hours
Actual Results (summary)	SRCNet Test Campaigns - Progress#Monday7April2025: Bad links identified; ~ 75% of remaining links achieved > 85% data transfer volume in the test window. 6.9% transfer failure rate (including bad links).				
Status	Passed	Failed			
Comments/Observations (summary)	Some instability in the morning of sustained rates, improved in afternoon (cause unknown as yet). No reported issues.				
Defects / Issues raised (Jira's, etc.)	SOG-39 - Failing transfers from SWESRC-OSO-T1 and CHSRC_XRD_PROD to STFC_STORM IN PROGRESS				

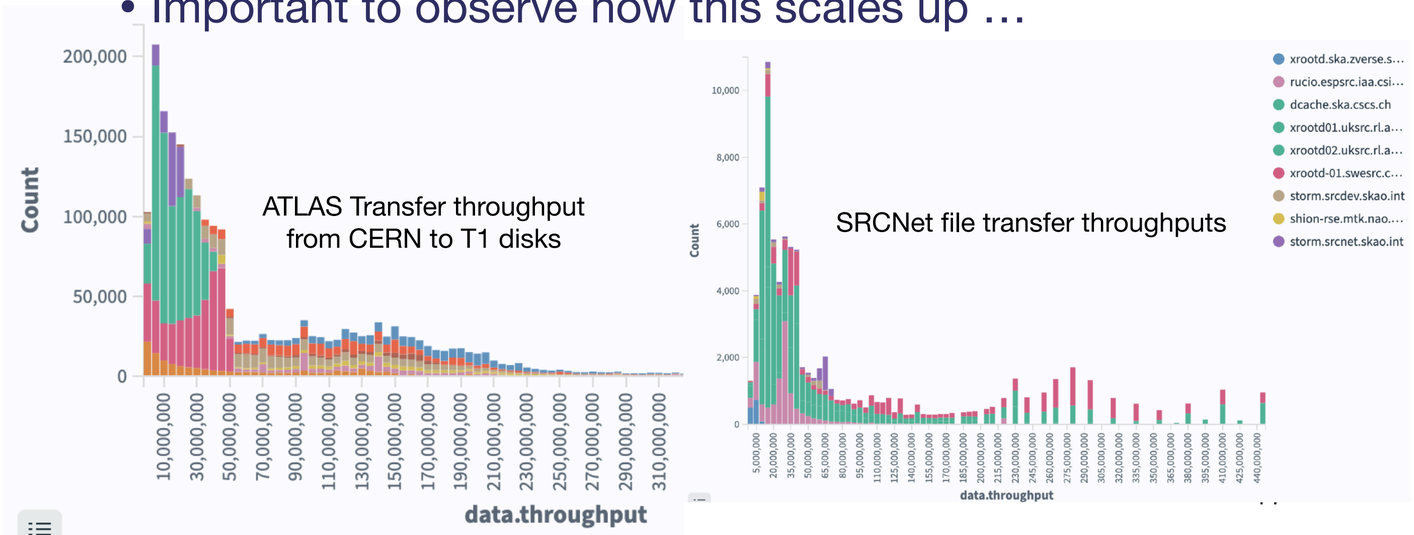
Test Plan outline

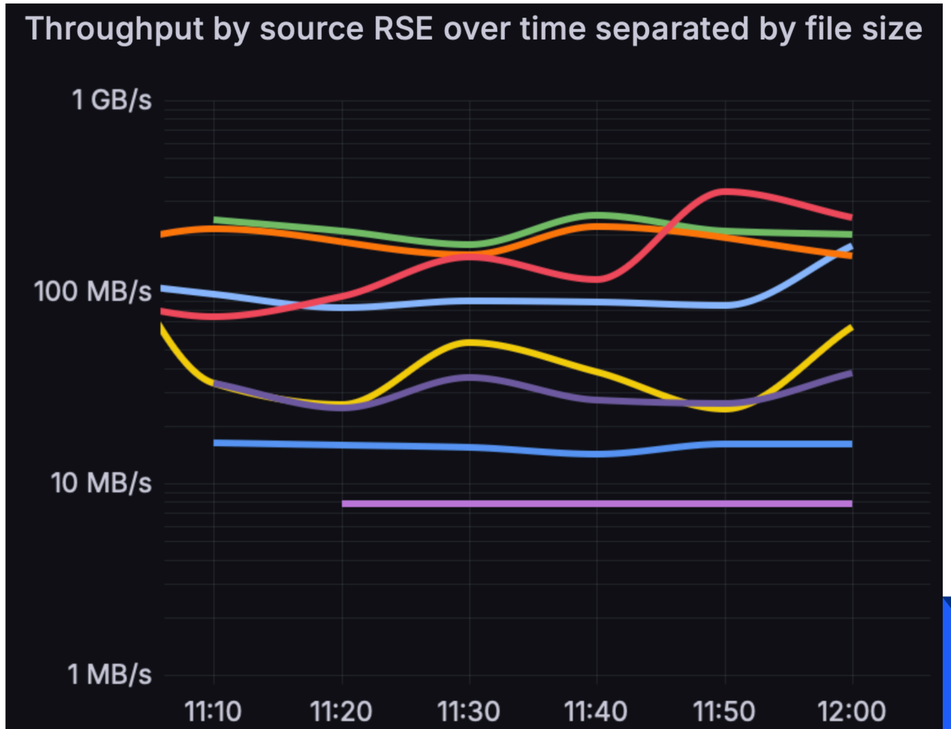
- **Data-Lifecycle and Science Delivery testing**
  - Ingestion of data into the Datalake (including the associated metadata)
  - Use of data in simple ‘analysis’ style environments (e.g. Jupyter notebooks, visualisation tooling).

File throughput testing

- Mesh script in rucio-task-manager able to provide ~ constant rate of transfers (up to site limits):
  - Deployed on k8s with ‘gitops’ control
  - Staggered rules avoid ‘flooding’ FTS (and sites)
- Individual File transfer speeds variation across links
  - O(150-500) MB/s achievable (in some cases)
- Systems are very lightly loaded;
  - Only a few concurrent transfers
- Important to observe how this scales up …

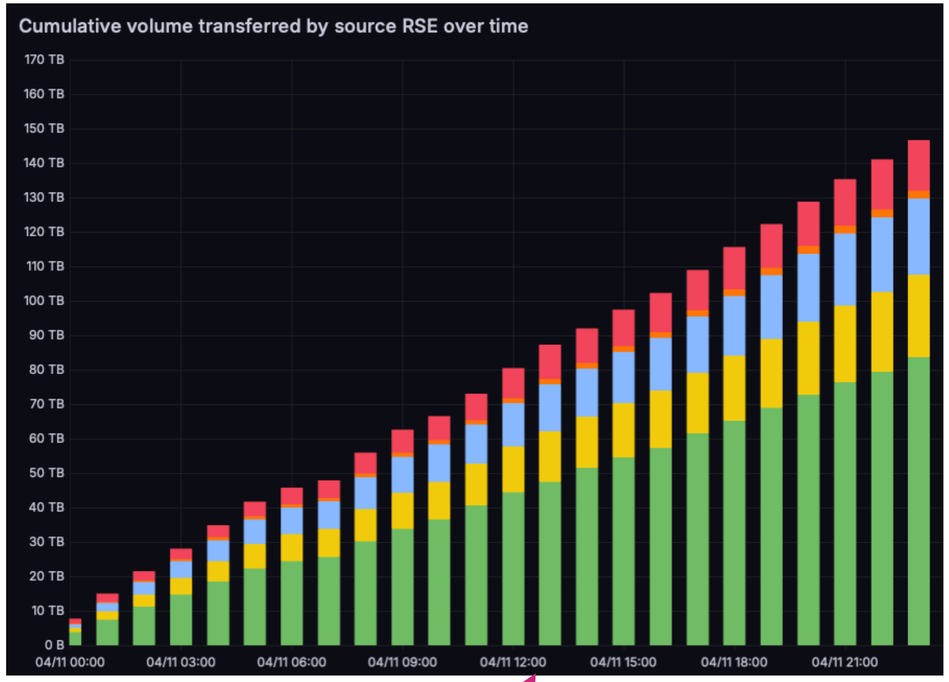




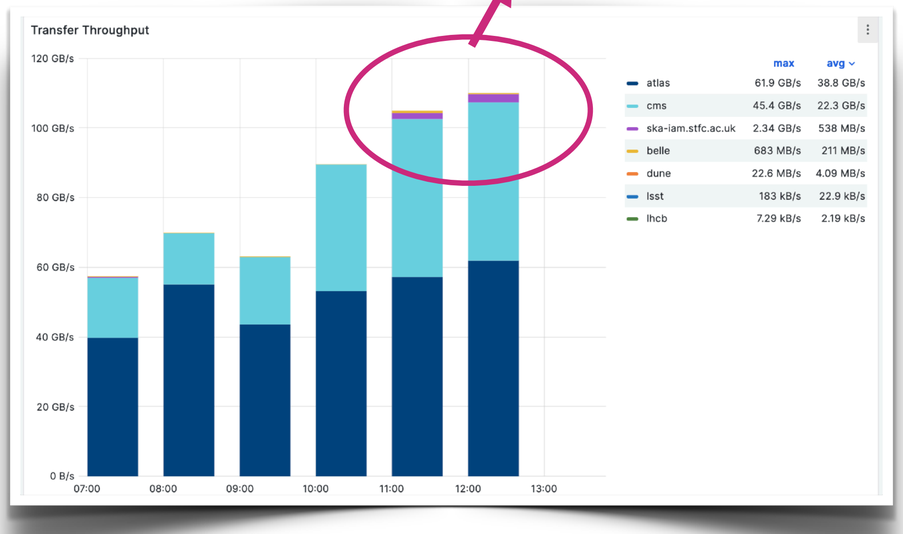


Milestones so far achieved (1)

- Capability tests performed:
  - Replication from Site_A -> Site_B and Site_B->Site_C
    - (Excluding Site_A -> Site_C)
    - Emulating data flow modelling
    - 5GiB files
- Capacity testing:
  - 150TB target in 24hours
    - Target rates for each RSE link-pair
      - Simple modelling emulating movements between larger and smaller RSEs
  - 147TB transferred within the time window



150TB

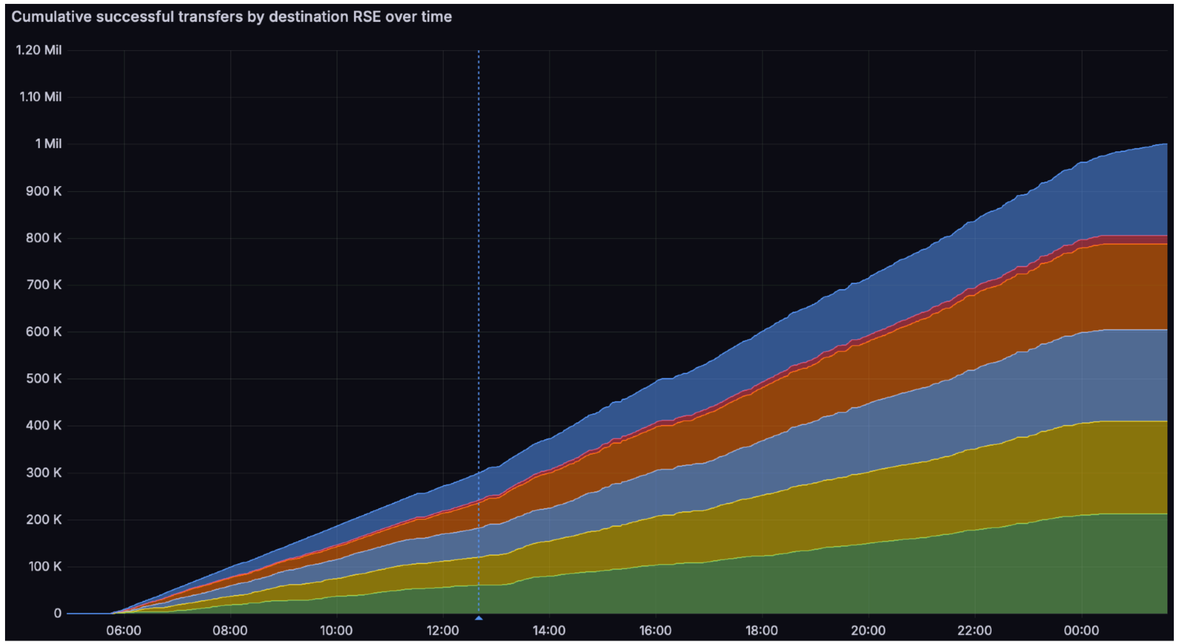


- Improvements in modelling in expected share of resources and long-haul source testing
- SRCNet now visible within the Global FTS Monitoring plots

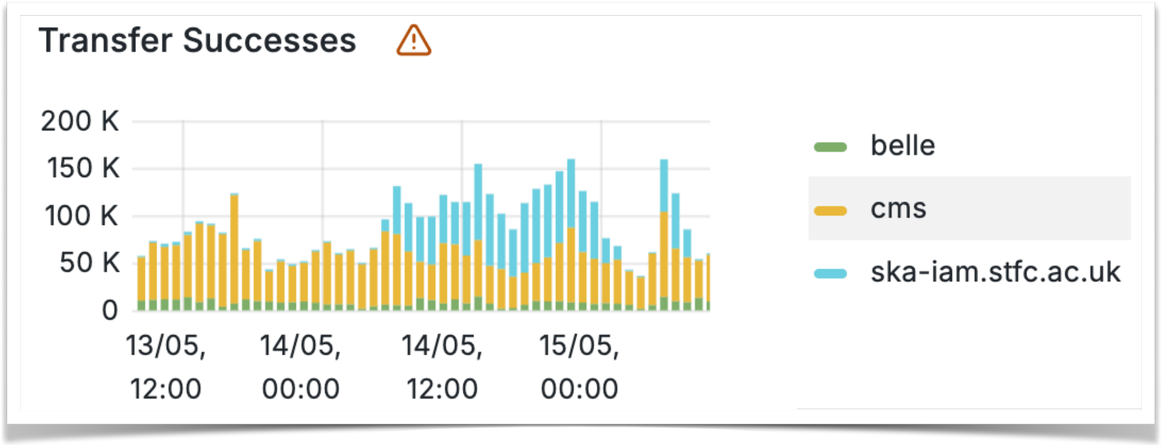
Milestones achieved (2)

- Test of ability of Rucio / FTS to handle higher submission rates, and tests of enabling services (e.g. monitoring)
- 1M x 1MiB files; target to replicate from a single source RSE to other Sites within 24hours
  - Stored as 1k x 1000 file datasets; submitted throughout the test period
  - RAL RSE used as source for all transfers

1M



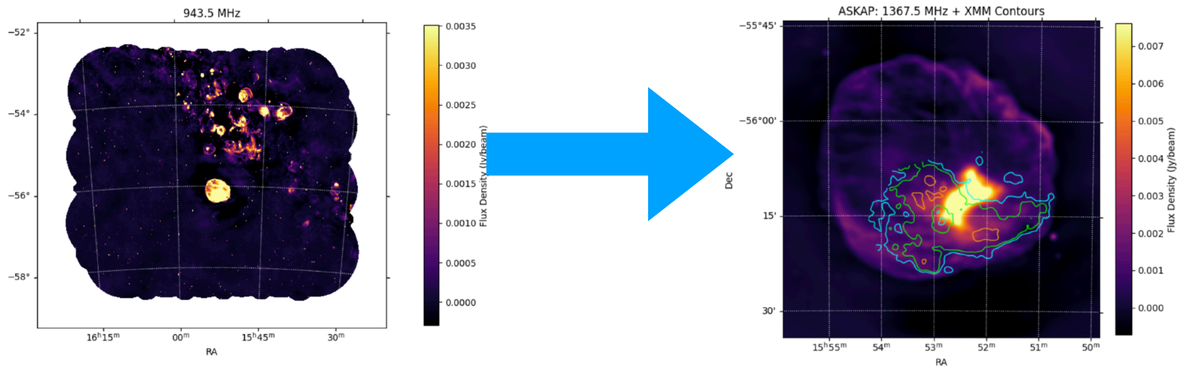
0



- Nominal submission rate (by rucio) ~ 50k/hour, with increase tested to 100k / hour.
- Rate comparable to e.g. CMS during the period.

Data Lifecycle testing (TBD)

- Developing “science test cases”, based on real science workflows
- to be integrated in the v0.1 test campaigns during PI27.
  - Use the Science Gateway interface
  - Data discovery within the SRCNet Datalake (Metadata plugin to RUCIO)
  - Trigger “staging” from bulk storage to local compute
  - Launch Jupyter hub (or CANFAR) instance
  - Run analysis code over dataset
  - Reproducibility on other Nodes



Florent Mertens

https://confluence.skatelescope.org/x/nCAAEw

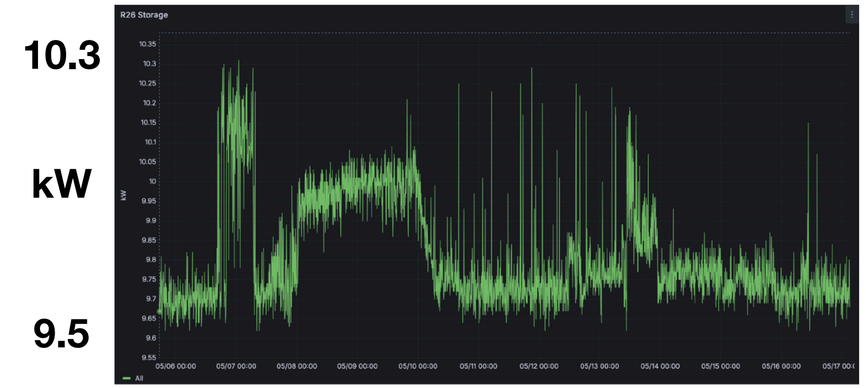
Main Github repo

Science Test cases

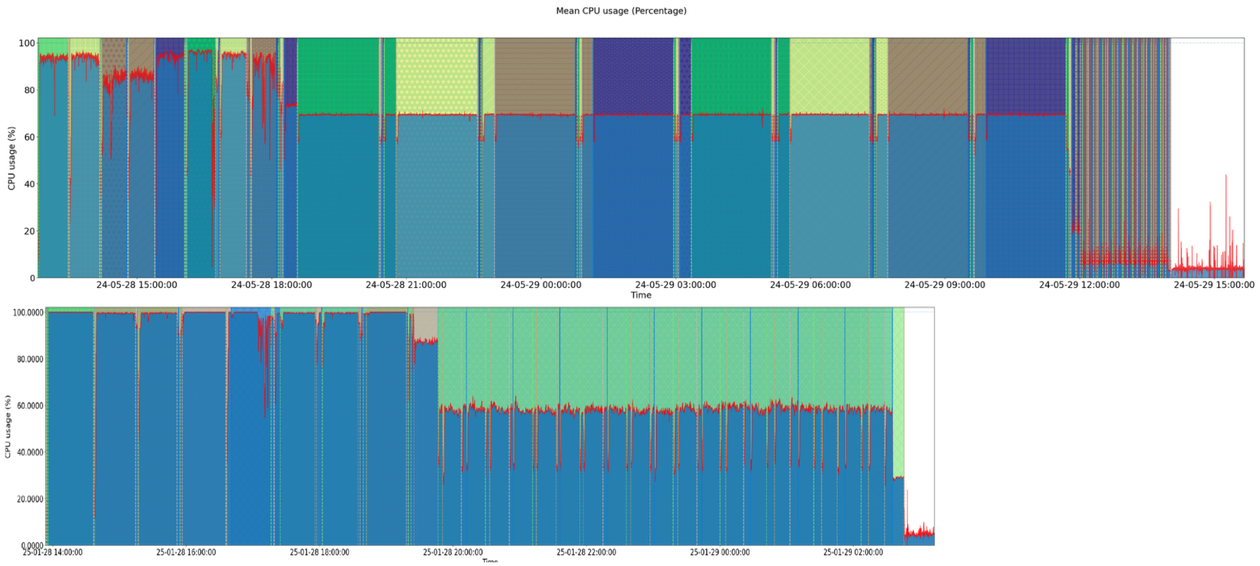
		Validation criteria	
01	**SWF-002-T1 Catalogue Search and Crossmatch** (Bonny)	• TESTED ON AZIMUTH • DATA ON AZIMUTH	SWF-002-T1.ipynb SWF-003-T1.ipynb
02	**SWF-003-T1 Smoothing and Resampling** (Florent + Adélie)	• TESTED ON AZIMUTH UNDER WAY • DATA ON AZIMUTH	SWF-005-T1.ipynb SWF-006-T1.ipynb SWF-008-T1.ipynb
03	**SWF-005-T1 Retrieve and import archival data including postage stamps** (Jacob)	• NOT TESTED ON AZIMUTH • NO DATA DOWNLOAD REQUIRED - DATA QUERY	SWF-010-T1.ipynb
04	**SWF-006-T1 Change sky projection** (Jacob)	• NOT TESTED ON AZIMUTH • NO DATA DOWNLOAD REQUIRED - DATA QUERY	No metadata files were created to any dataset!
05	**SWF-008-T1** Image-based EoR power spectrum (Adélie + Florent)	• NOT TESTED ON AZIMUTH • NO DATA ON AZIMUTH YET • CREATES INTERIM DATA	
06	**SWF-010–T1 Continuum** source finding (Lara)	• TESTED ON AZIMUTH • DATA ON AZIMUTH • CREATES INTERIM DATA	
07	**SWF-008-T2 -** visibility-based power spectrum (Jacob)	• test case was defined • Notebook is being created • No data and not tested on Azimuth	
08	**SWF-012 -** Image cube inspection (Adélie)	• Removed from the list - Needs to be removed or marked as incomplete on the splash page	

Sustainability and Benchmarking

- In infrastructure, work ongoing by Tom Byrne to look at implications of storage types vs performance vs power
- For current cluster setup, we observe ~
  - ~ 2.5 kW / PB(usable) in current load / usage conditions
  - (Includes all mons / mds / leaf switches / mgmt switche)
- For Compute, harder to understand from the infrastructure perspective (per tenant), however
  - Work ongoing into Profiling and Benchmarking of workloads => more efficient use of resources



- Example profiles from Marcus Krell et. al. (UCL)
  - VLBI Workflow example:
  - Old code (top) vs new code (bottom).
    - Mean CPU utilisation in blue
    - Colours represent different workflow steps
  - Optimisations in code leading to better parallelism less wait times and overall shorter execution



- Building monitoring into the platforms to allow users to understand / develop for better use of resources.

SRCNet Timeline

SRCNet v0.1 (Q1 2025)	SRCNet v0.2 (Q1 2026)	SRCNet v0.3 (Q4 2026 to Q4 2027)
SRC ART members	Selected scientists from community	Science Verification community
• **Test data** (and some precursor data) disseminated into a prototype SRCNet • Data can be **discovered** through queries to the SRCNet • **Data dissemination** to SRCNet nodes • Data can be **accessed** through a • prototype data lake • Data **replication.** • **Data can be moved** to a local SRC area where non-connected local interactive analysis portals (notebooks) could allow basic analysis • Unified **Authentication** System for all the SRCs • **Visualisation** of imaging data	• Data **dissemination using telescopes** sites interface • First version of **federated execution**. • Access to **remote operations** on data using services and the possibility to invoke execution into a relevant SRC • Subset of runnable **SDP workflows** • First **Accounting model** implementation. • Storage **User storage areas** • **Visualisation** of imaging and time series data through remote operations • Preparation of **SRCNet User Support**	• Improved data dissemination. Use of available storage • **SKA preliminary data** (and some precursors data) disseminated into a prototype SRCNet • **Upgraded federated computing.** Basic execution planner implementation and move execution to a selected SRC • **Upgrade of subset SDP workflows** runnable in the SRCs • Provide access to the **first set of workflow templates** for science analysis (light ADPs) • **Spectral data visualisation** and manipulation • **Implementation of SRCNet User Support**

Summary

- UKSRC to provide the facility for Radio astronomy research for SKA data and to support the UK radio astronomy community
  - UKSRC successfully accredited as a v0.1 node:
    - Infrastructure based at RAL, expertise from across the UKSRC institutes
  - Building out to stand up services and resources at other core sites

- SRCNet Top level roadmap under refresh with refined set of Storage and Compute requirements expected
  - Together with the implementation plan for v0.2

- Better understanding of compute requirements needed:
  - Work in collecting workflows, benchmarking and profiling to map to SRCNet use cases is ongoing
- Demonstrator and Pathfinder like use cases to inform decisions into UKSRC and deliver tangible outputs into the scientific community.

- Next phase of service deployment to bring federated job execution, requiring good understanding of of data distribution.
  - Demonstrator cases useful
  - Working with Partners in long-haul transfer data movements
  - Within UK, extend work to interact with existing e-Infrastructure, to understand data Ingress/Egress and compute for UKSRC.