The storage challenge of SKA's science data processing

Per3S presentation

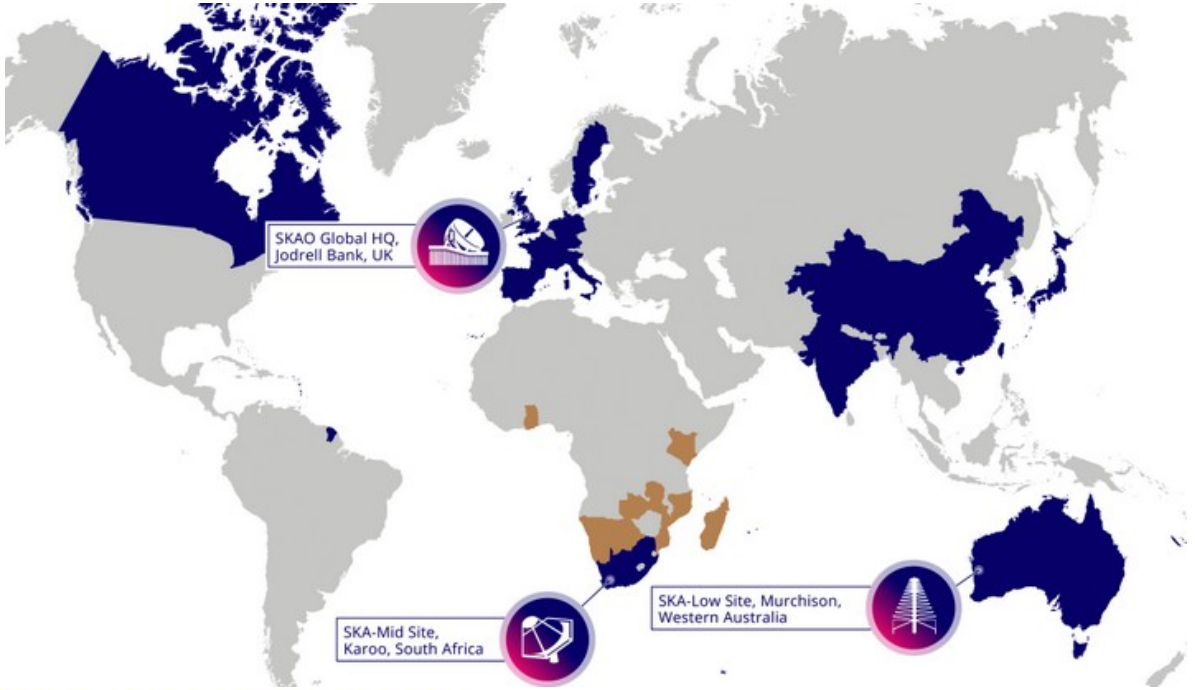
Shan Mignot

2024-05-28

Introduction

SKA Observatory

- Mission statement: “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.”
- Intergovernmental organization



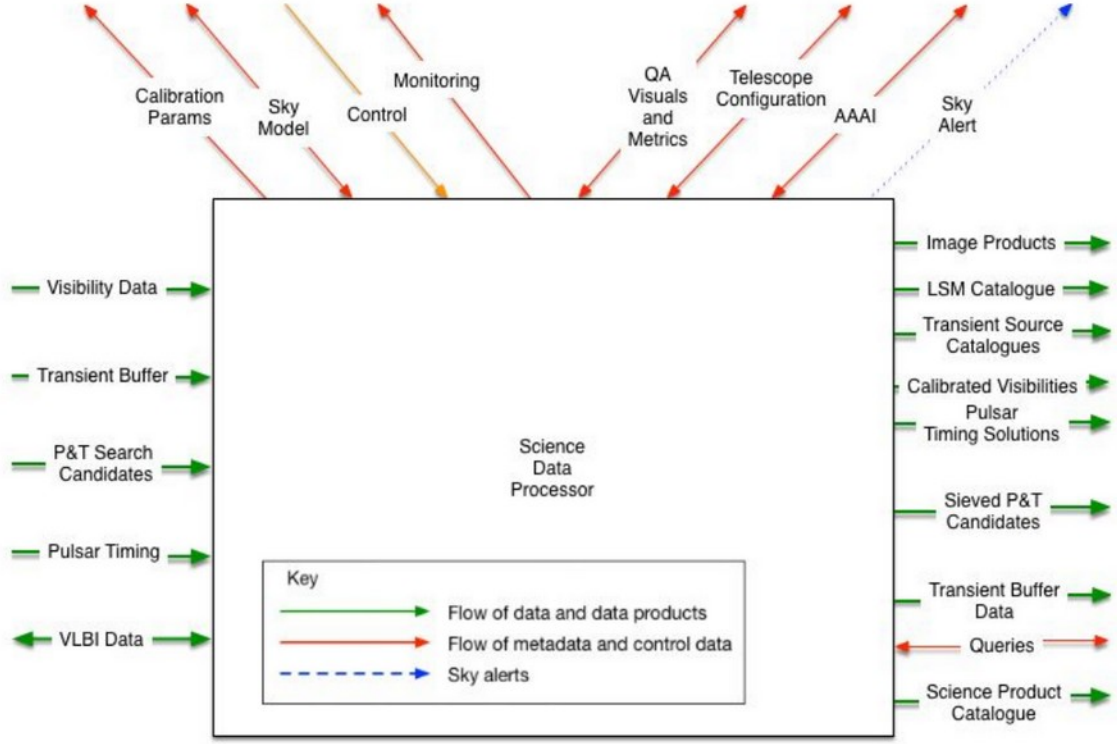
Construction timeline

- Critical Design Review: 2019-2020
- Start of construction: July 2021
- Progressive deployment of antennas
  - AA0.5: minimal array for early de-risking
  - AA1: de-risking via comparison to existing telescopes
  - AA2: already a full size radio-telescope
    - first science with SKA
    - fully functional system which should scale to SKA1
  - AA*: scaled AA2 offered to the community

Schedule Jan 2024	SKA-Low	SKA-Mid
Start of Construction (T0)	1st July 2021	
Array Assembly 0.5 finish (AA0.5) SKA-Low = 4-station array SKA-Mid = 4-dish array	Nov 2024	May 2025
Array Assembly 1 finish (AA1) SKA-Low = 18-station array SKA-Mid = 8-dish array	Nov 2025	April 2026
Array Assembly 2 finish (AA2) SKA-Low = 64-station array SKA-Mid = 64-dish array	Oct 2026	March 2027
Array Assembly 3* finish (AA*) SKA-Low = 307-station array SKA-Mid = 144-dish array	Jan 2028	Dec 2027
Array Assembly 4 finish (AA4) SKA-Low = 512-station array SKA-Mid = 197-dish array	*N/A*	*N/a*
Operations Readiness Review (ORR)	April 2028	April 2028

Data products

- Scientific
  - Alerts
    - detection within a few seconds: high-throughput computing, not demanding computing
    - follow-up: point and observer within a few seconds
  - Pulsar & transient catalogues
  - Image cubes: spatial x frequency channel (x polarisation)
  - Power spectra and line cubes: power versus frequency
  - Raw data (occasionally)



- Internal
  - Calibration: instrument & sky models
  - Closing feedback loops: pointing & beam forming
  - Quality assurance: telescope monitoring & science metrics

Some processing requirements

- Telescope operational availability > 94%
  - Operational Capable: A telescope is operationally capable when it can perform astronomical observations. It is assumed that the telescope will be capable if more than 95% of its collecting area, signal processing and data reduction capabilities are available.
- Asynchronous nature of SDP operation
  - SKAO-SDP_REQ-486: The SDP must achieve an average processing- over observation-time ratio of 1 over a period of 30 days.
  - SKAO-SDP_REQ-495: The SDP shall have an Inherent Availability (Ai) higher than or equal to 99.9%.
  - unavailability of SDP due to saturation of resources not considered in requirements analysis
- Flexible use: up to 16 sub-arrays for Low & commensal observations based on the same data
- 50-year lifespan of SKA: maintainability & modifiability

	MID-Telescope-Time %	LOW-Telescope Time %
Observing (OT)	88%	90%
Standby (ST)	0%	0%
Weather (WT)	3%	1%
Utility (UT) (External: power, cooling water and communications services)	4%	4%
Engineering / Maintenance (MT) (scheduled maintenance, off-line calibration, software updates or testing)	3%	3%
System Fault (DT)	2%	2%
Critical Repair Time (CRT)	1%	1%
Critical Support Delay (CSD)	1%	1%
	100%	100%

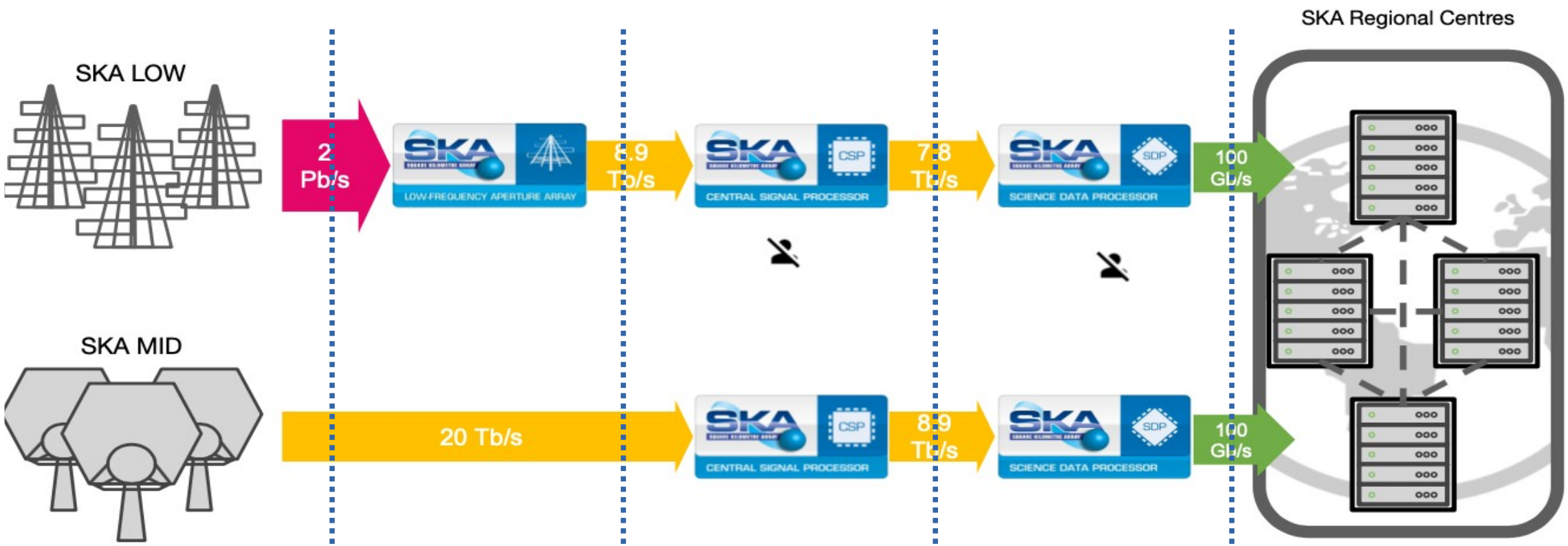
State		Definition
Operational	Observing	At least one sub-array is in the Calibrating or Observing mode, for the purpose of science observations.

from Jama specification

Data paths

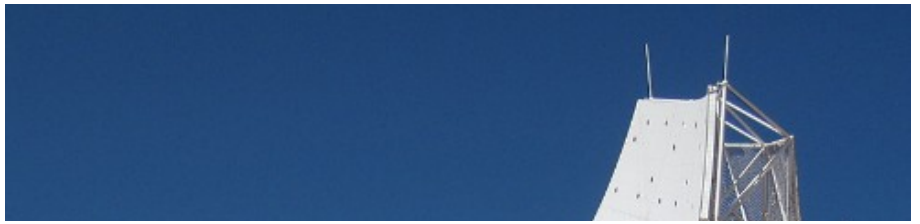
Overview

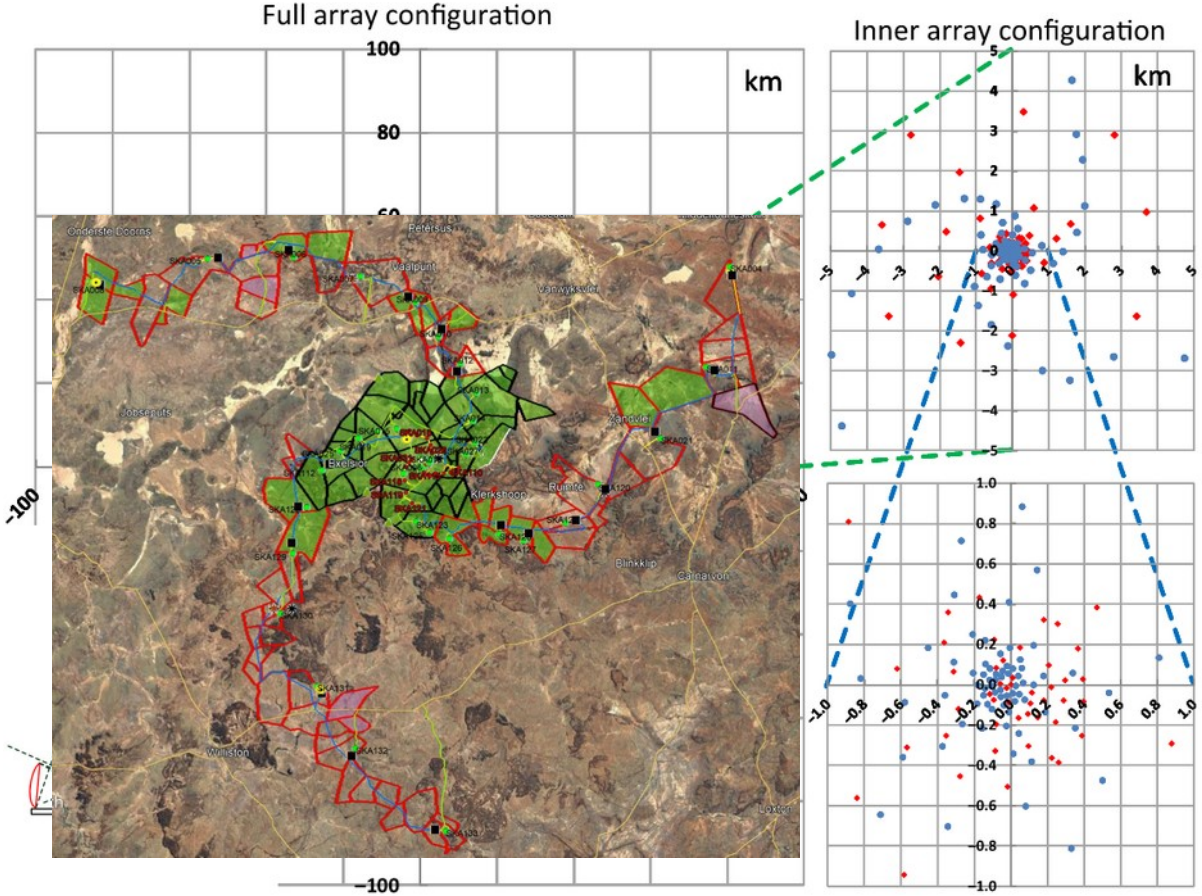
- Edge to cloud architecture to manage data flow and complexity



Mid telescope

- Frequency range: 350 MHz-15.4 GHz
- 197 dishes, including MeerKAT (red dots)
- Dish geometry: 15 m parabolic reflector
  13.5 m for MeerKAT
- Losberg in the South African Karoo region
- Distribution
  - core within ~1km
  - 3 spiral arms
    - up to 150 km baselines
    - logarithmic distribution





Mid data acquisition and transformation

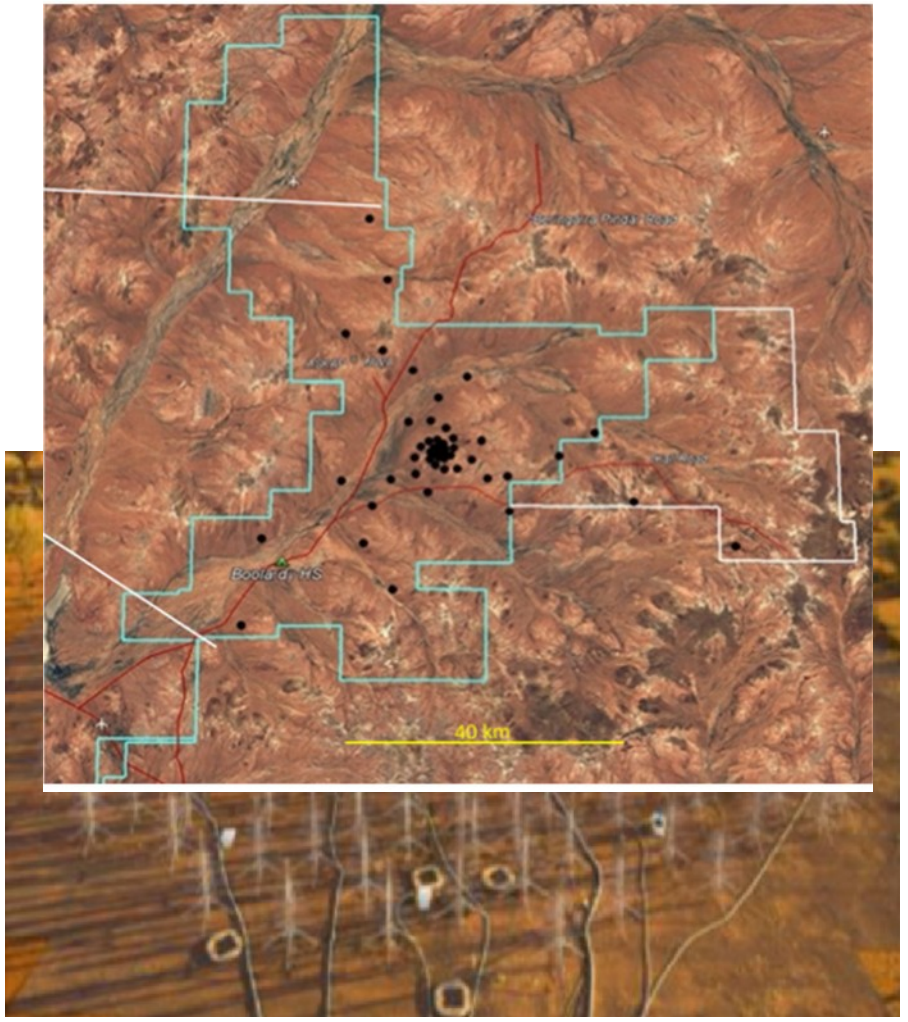
- 5 bands (exclusive) & 2 polarizations
- Digitization on dish: sample RF signal & timestamp data with nanosecond stability over 10 years
- Fiber optics signal transmission to Central Processing Facility (CPF) near core: channelization (<65536 channels)

Band	Frequency (MHz)	Sample rate (GS/s)	ENOB (bits)	Sample (bits)	Data rate (Tb/s)
1	350-1050	3.96	8	12	18.7
2	950-1760	3.96	8	12	18.7
3	1650-3050	3.17	6	12	1
4	2800-5180	5.94	4	8	18.7
5a **5b**	4600-8500 8300-15300	5.94	3	4	18.7

- All other transformations moved to Science Processing Center (SPC) in Cape Town
  - Beam forming (array & < 1500 search beams)
  - Correlation per pair of dishes, per beam, per channel: 12 bytes complex visibility, integration time step (> 0.14 s)
  - Temporal data analysis

Low Telescope

- Frequency range: 50-350 MHz
- 131 072 log-periodic dipole antennas
- Aperture array telescope: 256 antennas per stations, 512 stations
- Boolardy, Western Australia
- Distribution
  - randomised locations
  - core within 4 km (224 stations)
  - 3 spiral arms (288 stations)
    - up to 74km baselines
    - logarithmic distribution



Low data acquisition & transformation

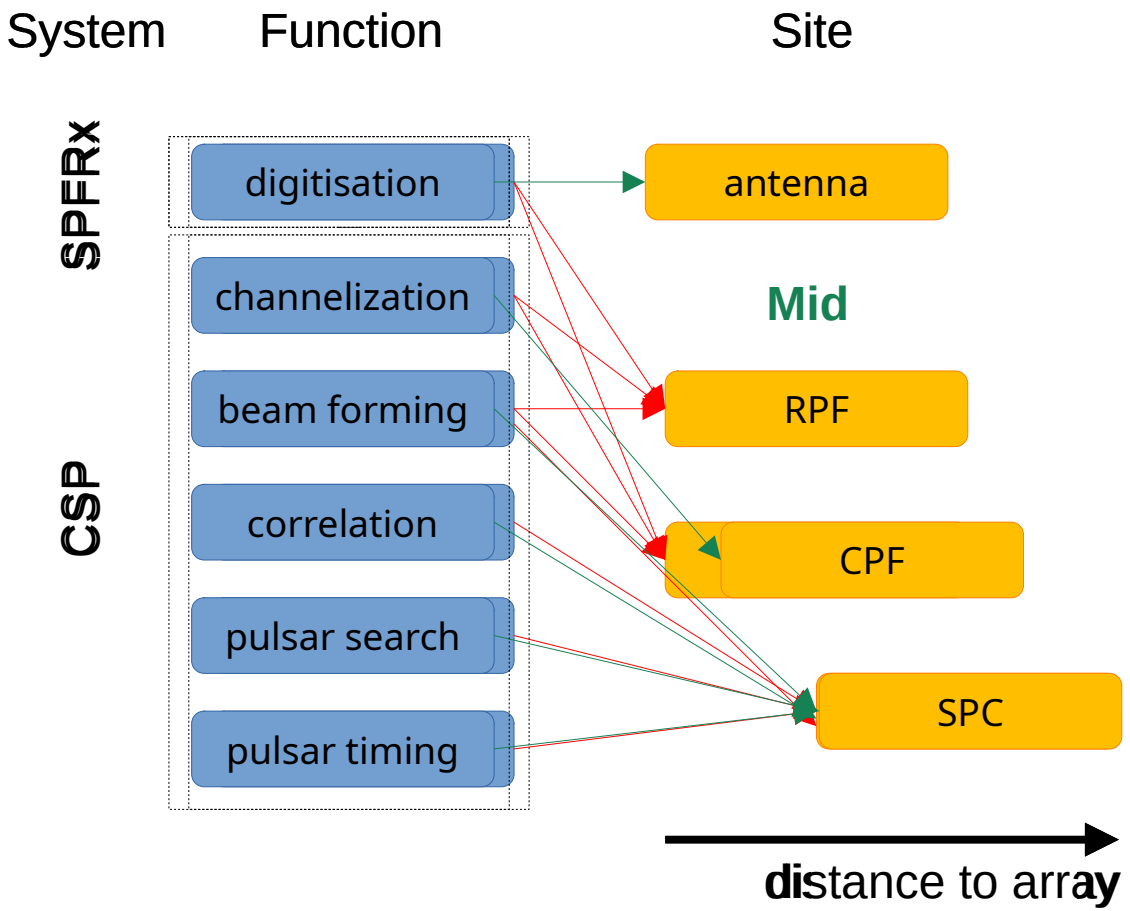
- 1 band & 2 orthogonal polarizations
- Antennas digitized individually: all subsequent operations are digital
  - RF transported by fibre optics (< 4 km): digitization for clusters of antennas
  - CPF for 296 central stations near core of array & 36 Remote Processing Facilities (RPF) for 216 outer stations (clustered by 6)
- Beam forming (48 beams) and channelization (384 over-sampled coarse channels) at station level (CPF or RPF)
- All other transformations moved to Science Processing Center: array beam forming, fine channelization, correlation, temporal data analysis

Architecture (I)

- Early digitization: replicated data streams creates opportunities for using the antennas
  - multiple beams
  - sub-arrays
  - spatial / temporal analyses
- Even with well identified tasks, complexity is highly variable
- Strategies for sizing compute resources:
  - CSP: data flow warrants stream processing, well identified/repetitive tasks and culture point to FPGAs (GPUs are being considered)
  - SDP: some diversity, evolutive, somewhat “specialised” HPC
  - SRC: user-oriented, HPC & HPDA distributed infrastructure

Architecture (II)

- Mapping: delocalize and concentrate
- Criteria
  - accessibility: operation staff, maintenance
  - future extension of infrastructures
  - minimize interferences
  - scale savings
  - energy



Data reduction

- Science Data Processors (SDP) in host countries
  - intermediate between CSP: stream processing, repetitive, strong coupling to antennas
  - and SRC: diversity of investigations and users
- Reduce data flow: process data at the acquisition rate to avoid data loss
- Observatory Data Products
  - science products required to allow transfer and storage
  - ingest: max (single observation) ~0.45 TB/s
  - distribute: average ~10 GB/s (300 PB/year)
- Internal products: feedback to telescope manager and CSP

Data analysis

- SKA Regional Center (SRC) network
  - provided by member countries (worldwide)
- Missions
  - science access
    - archive: 2*300 PB/year
    - in situ visualization and processing
  - multi-epoch data reduction (Project Science Products)
    - storage impedes this at SDP level
    - shared responsibility between SKAO and SRCNet

Science Data Processor

Imaging data management challenges

- Ingest rates for Mid and Low

Mid

12 bytes/visibility
* C^2_{196} = 19110 baselines
* 4 polarizations
* 65535 channels
/ 0.14 s min integration
= 0.43 TB/s

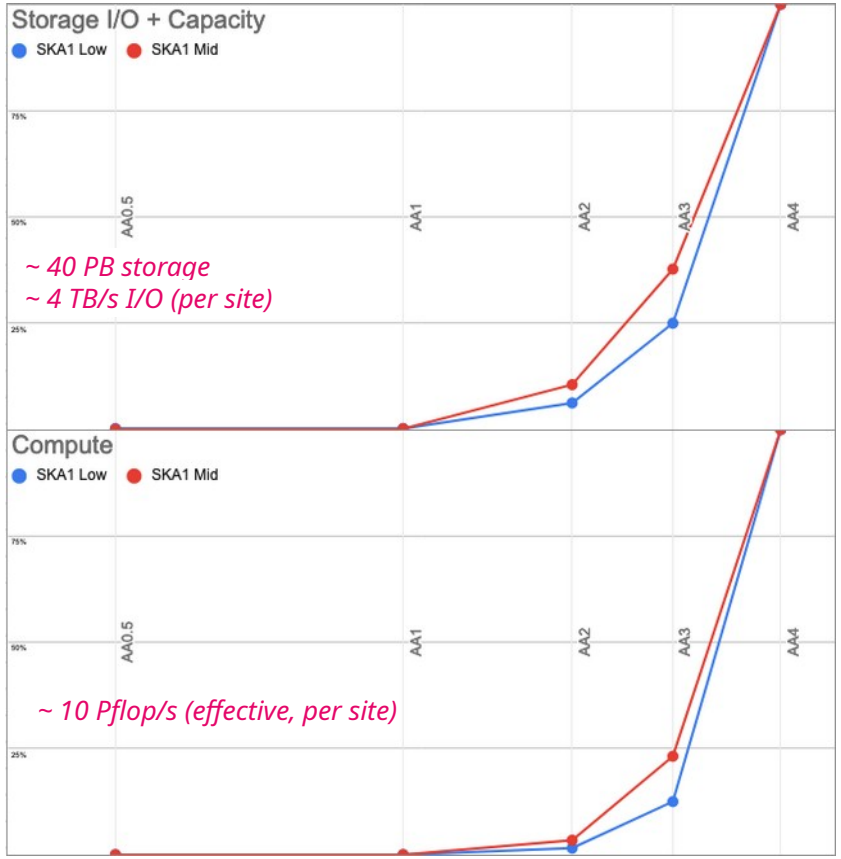
Low

12 bytes/visibility
* C^2_{512} = 130816 baselines
* 4 polarizations
* 65536 channels
/ 0.9 s min integration
= 0.46 TB/s

- Streaming and batch processing (over 24h)
- Baseline: buffer as a central component
  - size (as per CDR): 0.45 TB/s is ~ 40 PB/day
  - random access to data (Fourier transforms!)
  - flexibility in scheduling for higher efficiency

Estimated SDP Scaling: AA1→AA4

(~50x in 17 months! Qualitative only, underestimates the AA2 situation)



courtesy Peter Wortmann

Power

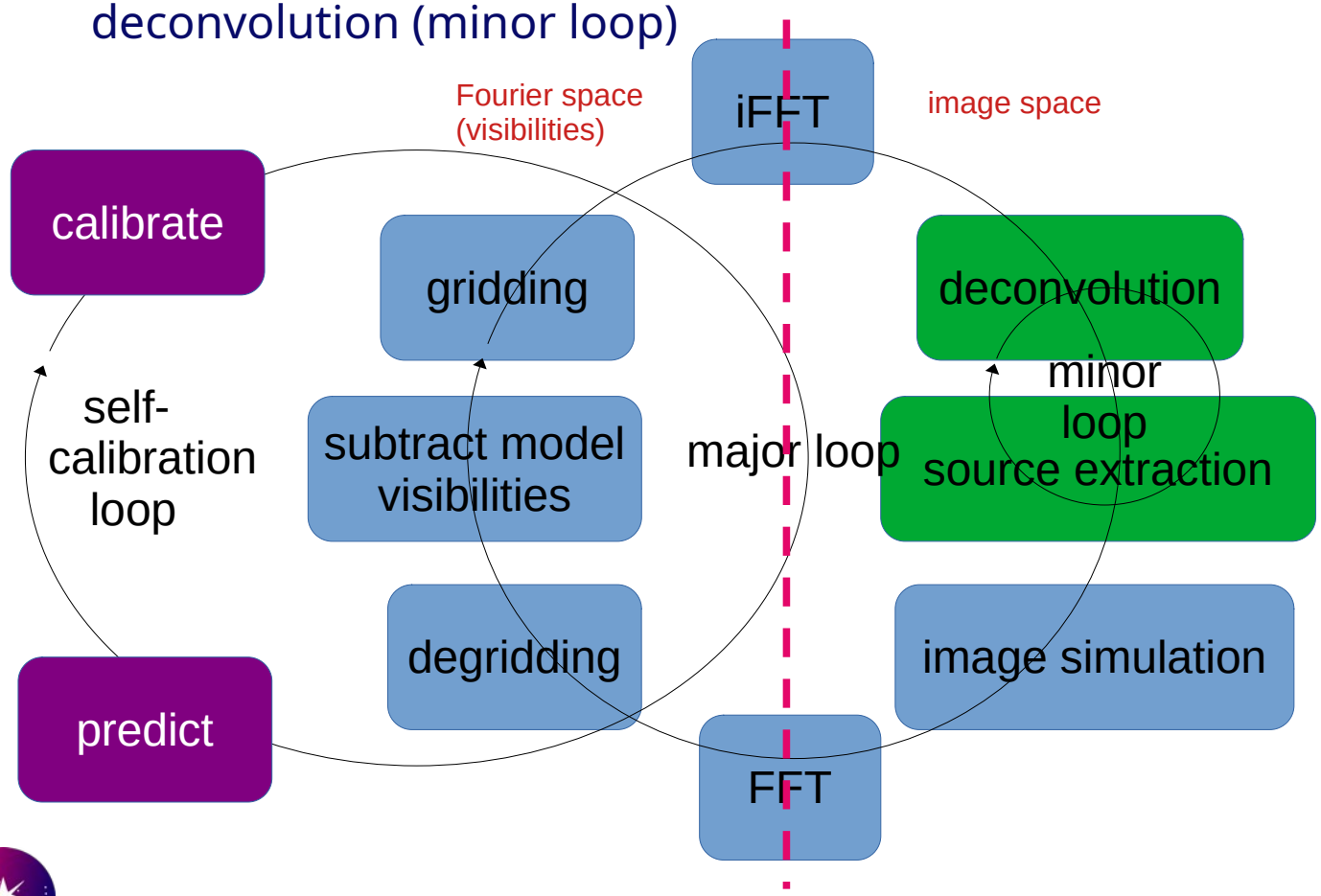
- SPCs in Cape Town and Perth
  - most of CSP (after AA2)
  - SDP
  - Infrastructure & Cooling
- Cost & greenhouse gas issue: multiple power sources (grid/solar/diesel/battery/RUPS)
- SDP allocation
  - average: 1.3 MW Mid / 1.6 MW Low
  - peak: 2.0 MW Mid / 2.23 MW Low
- Green500: Frontier, Lumy, Adastra achieve ~100 PFlops @ 2MW at maximum efficiency

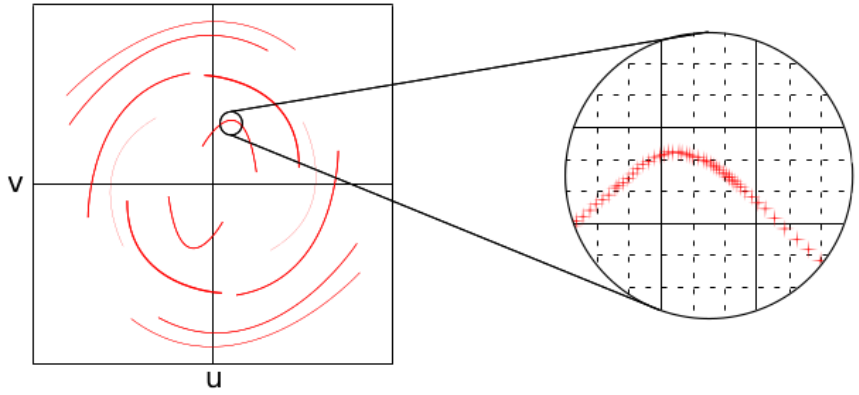
SKA1-Mid SPC/SOC Power Budget in Cape Town				
Products	*AA4 Long Term Average (>30min) [kW]*	*AA4 Peak Instantaneous (<5sec) [kW]*	*AA* Long Term Average (>30min) [kW]*	*AA* Peak Instantaneous (<5sec) [kW]*
PDT4 - MID Digitisation	**230.8**	**323.4**	**230.8**	**323.4**
CSP.CBF	230.8	323.4	230.8	323.4
PDT6 - Network & Computing	**1641.7**	**2481.9**	**589.3**	**872.6**
SDP Hardware MID	1300.0	2000.0	325.0	500.0
PSS Hardware MID	296.0	414.0	222.0	310.5
PST Hardware MID	16.4	26.8	12.3	20.1
OMC Hardware MID	12.9	18.1	12.9	18.1
NSDN MID	5.6	7.8	6.6	9.2
CPF-SPC link MID	8.2	11.5	7.9	11.1
NMGR	2.6	3.6	2.6	3.6
Building losses and cooling	**374.5**	**561.1**	**164.0**	**239.2**
Commissioning Margin	**224.7**	**336.6**	**98.4**	**143.5**
Site Total	**2471.7**	**3702.9**	**1082.6**	**1578.8**

SKA1-Low SPC/SOC Power Budget in Perth				
Products	*AA4 Long Term Average (>30min) [kW]*	*Peak Instantaneous (<5sec) [kW]*	*AA* Long Term Average (>30min) [kW]*	*AA* Peak Instantaneous (<5sec) [kW]*
PDT6 - Network & Computing	**1629.2**	**2270.9**	**429.2**	**598.4**
OMC Hardware LOW	12.9	18.1	12.9	18.1
SDP Hardware LOW	1600.0	2230.0	400.0	557.5
NSDN LOW	6.5	9.1	6.5	9.1
CSP-SDP LOW	7.2	10.1	7.2	10.1
NMGR	2.6	3.6	2.6	3.6
Building losses and cooling	**325.8**	**454.2**	**85.8**	**119.7**
Commissioning Margin	**195.5**	**272.5**	**51.5**	**71.8**
Site Total	**2150.6**	**2997.6**	**566.6**	**789.9**

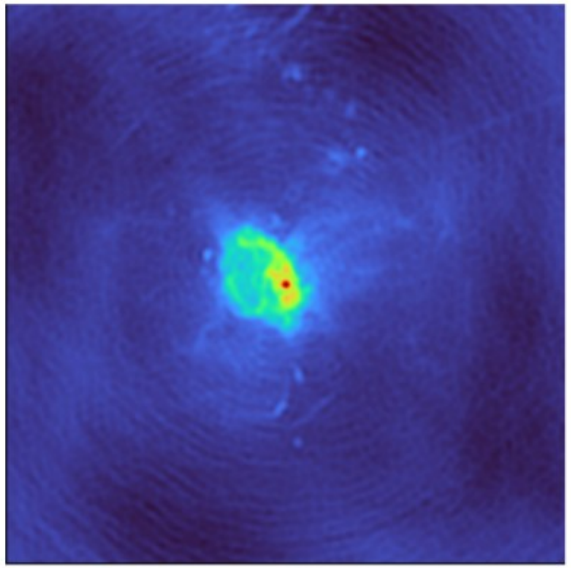
Self-calibration pipeline architecture

- Three nested loops: iterate on calibration products (self-calibration loop) as we iterate on image reconstruction (major loop) with iterative deconvolution (minor loop)





uv coverage resulting Earth rotation (courtesy Nicolas Monnier)



dirty image of Sgr A (courtesy Sunrise Wang)

Software architecture

courtesy Peter Wortmann

- Pressure on resources: SDP sizing assumes 10% efficiency (in the HPCG sense)
  - storing raw visibilities for 20h is ~40 PB: need to keep predicted visibilities in memory
  - assumed 10 major loop iterations exert pressure on nodes: read 4 TB/s over ~500 nodes with 100 GB RAM require processing each visibility in 12.5s
  - image cubes for Mid can be as large as ~14 TB with (theoretically) each visibility affecting all pixels: keep all cube in memory
- Focus on reading visibilities from buffer only once per major loop
- Break the one-to-all correspondence between pixels and visibilities: SwiFTly distributed Fourier transform scheme
- Parallelise over frequency, time, image regions (facets) and baseline length
- Watch reduction steps: storage and network bound

Storage vision for AA2 (I)

courtesy Peter Wortmann

- Performance buffer: support fast processing (hold visibilities accessed multiple times)
  - capacity: 225 TB
    - from 0.2-11.3 TB to 34-120 TB for 4h of observation (depending on averaging)
    - 10 concurrently processed observations
    - 100% margin for reprocessing (pipelines may fail to converge in quality and in time)
  - throughput: 32 GB/s
    - 10% of time accessing data
    - ~80 read/write cycles per observation
    - inefficiency when I/O and compute cannot fully overlap
  - rough capacity/throughput value of 7.2E4 s

Storage vision for AA2 (II)

courtesy Peter Wortmann

- Capacity buffer: support ingestion and hold data until it is processed
  - capacity: 1.5 PB
    - hold data for a week
    - with a 80/20 performance/capacity split
  - throughput: 25 GB/s
    - ingest at 2.3-8.3 GB/s
    - extra data (technical) & data products: add 70%
    - 10 read cycles
  - rough capacity/throughput value of 6.05E4 s

Storage vision for AA2 (III) courtesy Peter Wortmann

- Long term storage: preserve visibilities and data products
  - throughput: 417 MB/s
    - preserve 5% of visibilities
    - infrequent access to data products
  - capacity: 52.6 PB
    - store data until AA* for 2 to 4 years

Conclusion

A (difficult) problem for co-design

- Storage has a significant share of SDP TCO
- Need to validate assumptions and measure inefficiencies (Peter's reasoning)
- Identify access motifs to optimise data model and placement
- Need to characterise resource usage
  - benchmarking at scale is long and costly
  - exclusive access to resource is difficult & permissions are restricted
  - depends on data, software maturity and overall SDP usage

On-going efforts

- SKA France in kind contribution to SKA for sustainable computing
- SKA France & DDN involved in co-design studies with a current focus on procuring AA2 HW
- SKA chosen as an illustrator for Exa-DoST (NumPEx PC3)
  - evaluate I/O and storage management libraries
  - efficiency requires more than just optimising storage (compute, scheduling etc.) find the right balance: coordinate with Exa-Ma, Exa-AToW and energy
- Phase 0 for constructing the French SRC node: report to the MESR in October 2024 for funding
  - at least 3% of SRCNet resources to comply with SKAO rules
  - support the French community's use of SKA data and precursors (LOFAR, NenuFAR)
  - 4 working groups: governance model for a distributed infrastructure, HW, SW and science roadmap

We recognise and acknowledge the Indigenous peoples and cultures that have traditionally lived on the lands on which our facilities are located.

SKAO

www.skao.int