SKA: Overview of the Data Processing Challenges and Development of the Science Data Processor

SIPS Keynote

Shan Mignot

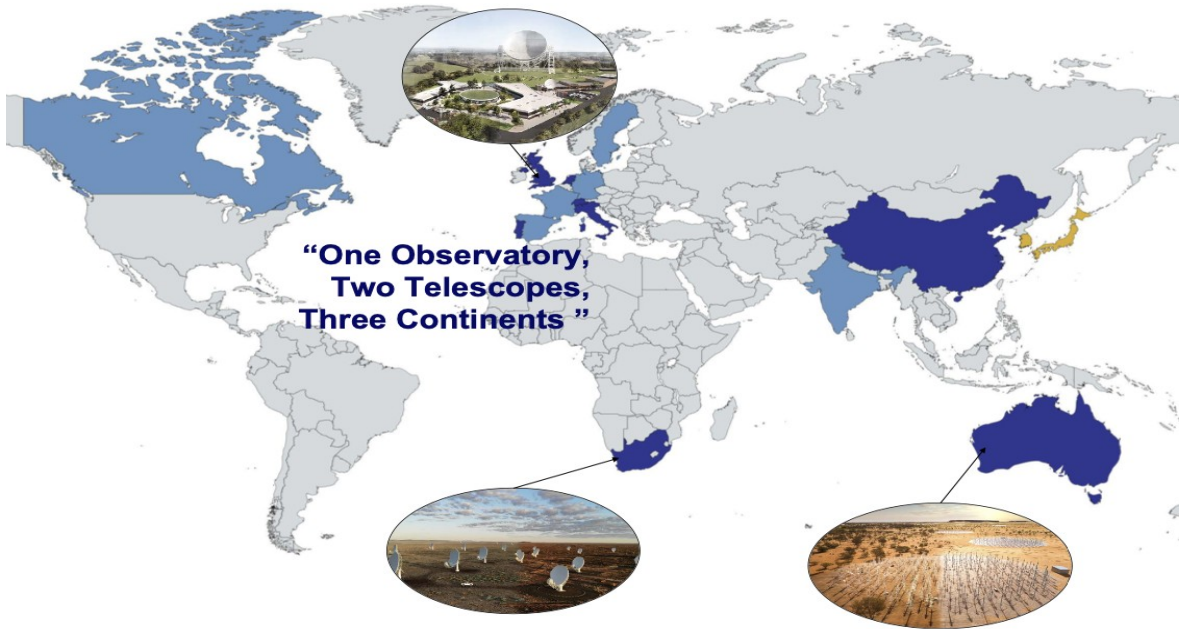
2022-11-02

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



SKAO Member States:

Accession stage:

Membership negotiations:

Interim agreements:

Early stages:

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

	SKA-Low	SKA-Mid
Start of Construction (T0)	1st July 2021	
Array Assembly 0.5 finish (AA0.5) SKA-Low = 6-station array SKA-Mid = 4-dish array	July 2024	May 2024
Array Assembly 1 finish (AA1) SKA-Low = 18-station array SKA-Mid = 8-dish array	September 2025	May 2025
Array Assembly 2 finish (AA2) SKA-Low = 64-station array SKA-Mid = 64-dish array	July 2026	July 2026
Array Assembly 3* finish (AA*) SKA-Low = 307-station array SKA-Mid = 121-dish array	June 2027	August 2027
Array Assembly 4 finish (AA4) SKA-Low = 512-station array SKA-Mid = 197-dish array	*N/A*	*N/a*
Operations Readiness Review (ORR)	August 2027	October 2027
End of Construction (incl. contingency)	July 2028	

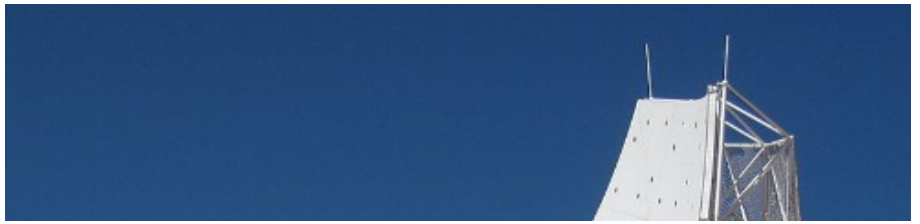
Some processing requirements

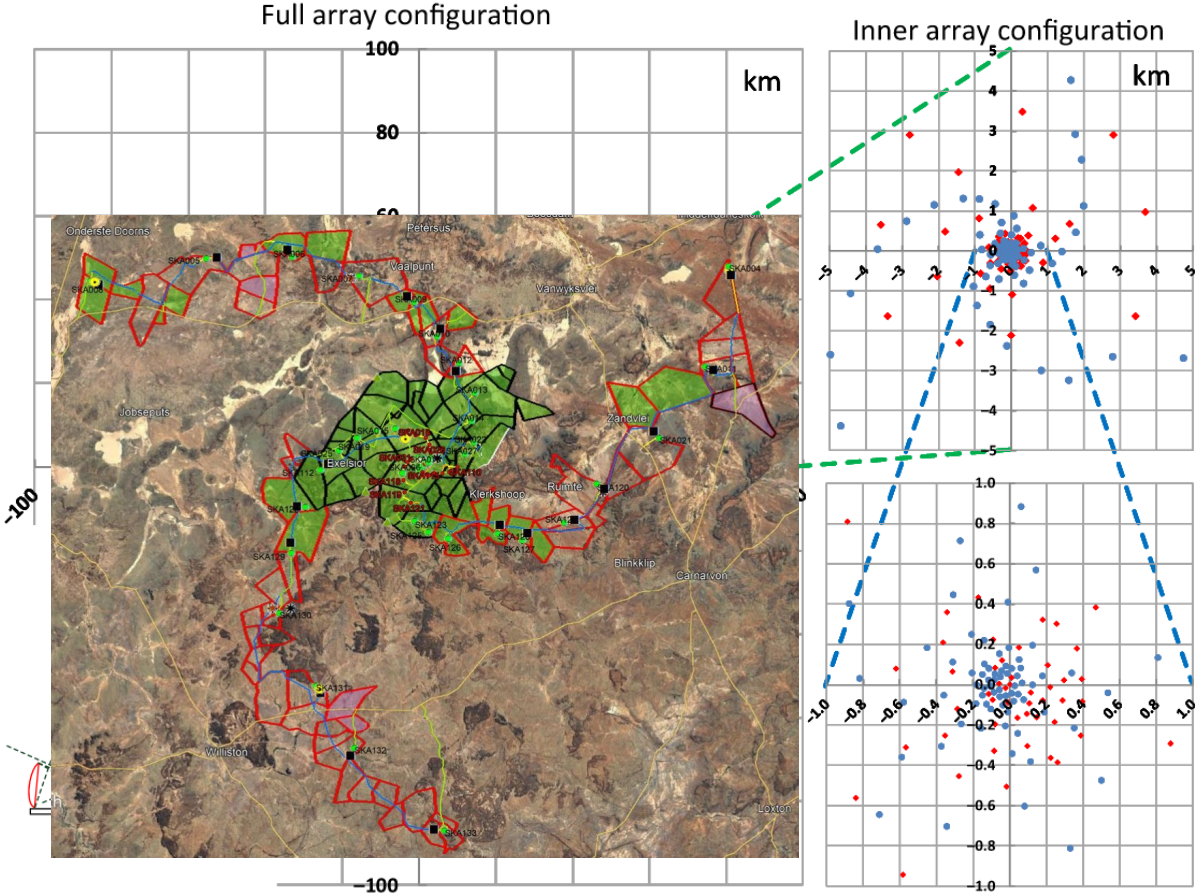
- Observing modes: imaging, non-imaging, VLBI, calibration
- Internal & science data products
- 95% operational availability
- Flexible use
  - operation in independent sub-arrays: up to 16 sub-arrays for Low
  - commensal observations: multiple uses of the same observational data
- Durable system: 50-year lifespan of SKA
  - maintainability
  - modifiability: science, technology, extensions of arrays

Data paths

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

- 5 bands (exclusive)
- 2 orthogonal polarizations
- digitizer
  - sample RF signal
    - 197 dishes*2 polarizations*2 I/Q*sample rate*bits/sample
  - timestamp data
    - per packet
    - nanosecond stability over 10 years

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

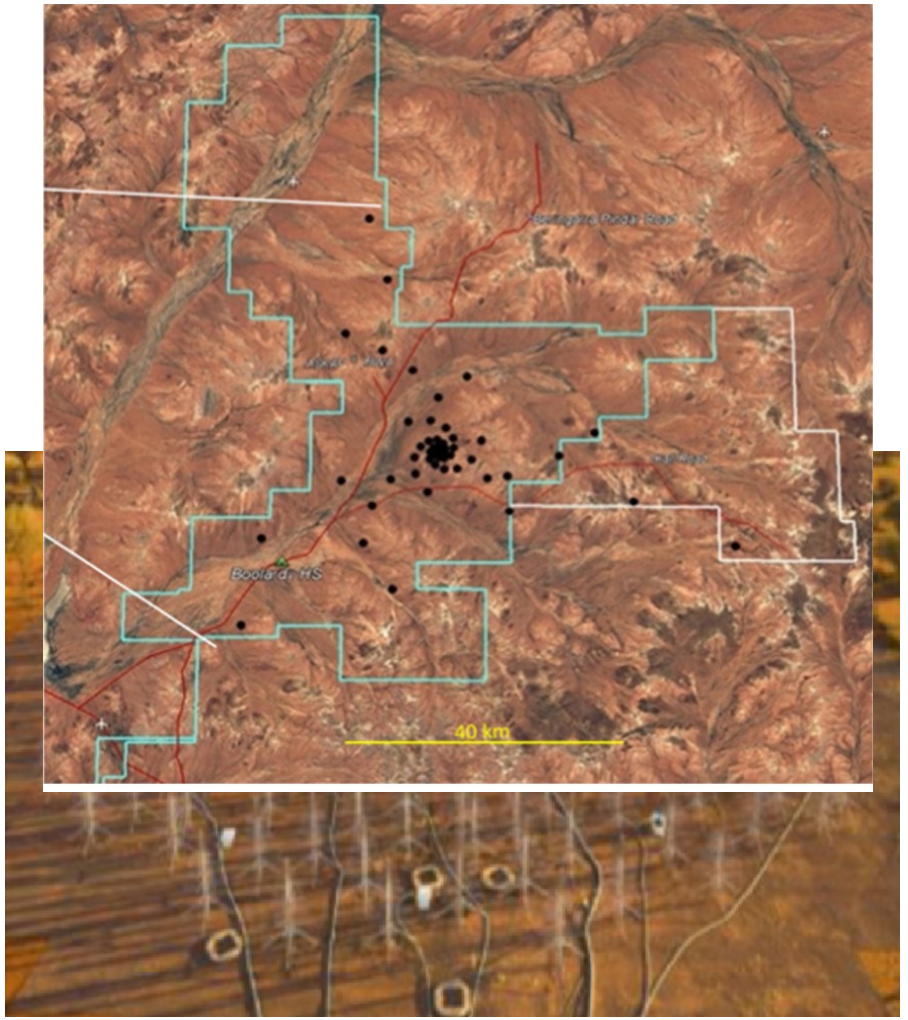
Mid data transform

- Fiber optics signal transmission to Central Processing Facility (CPF) in MeerKAT processor building near core
- Channelization: <65536 channels (spectroscopy)
- Beam forming: array beam & up to 1500 search beams
- Correlation: 12 bytes complex visibility
  - per pair of dishes, per beam, per channel
  - for each integration time step (> 0.14 s)

    12 bytes*19110 baselines*4 polarizations*65535 channels/0.14 s = 0.43 TB/s
- Temporal data anaysis
  - list of potential pulsars (PSS)
  - timing of known pulsars (PST)

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

- 1 band
- 2 orthogonal polarizations
- Antennas digitized individually: all subsequent operations are digital
  - LNA and signal conditioning for each antenna (at ambient temperature)
  - 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): 216 outer stations clustered by 6
    - data flow:

      131072 antennas*2 polarizations*2 I/Q*800 MHz sample rate*8 bits/sample = 1700 Tb/s

Low data transform

- Beam forming at station level (CPF or RPF): 48 beams per station
- Channelization (CPF or RPF): 384 oversampled science coarse channels
- All other transforms moved to Science Processing Center
  - array beam forming
  - fine channelization: 55556 channels, correlator, beam former, PSS, PST moved to SPC
  - correlation: minimum time step 0.9 s

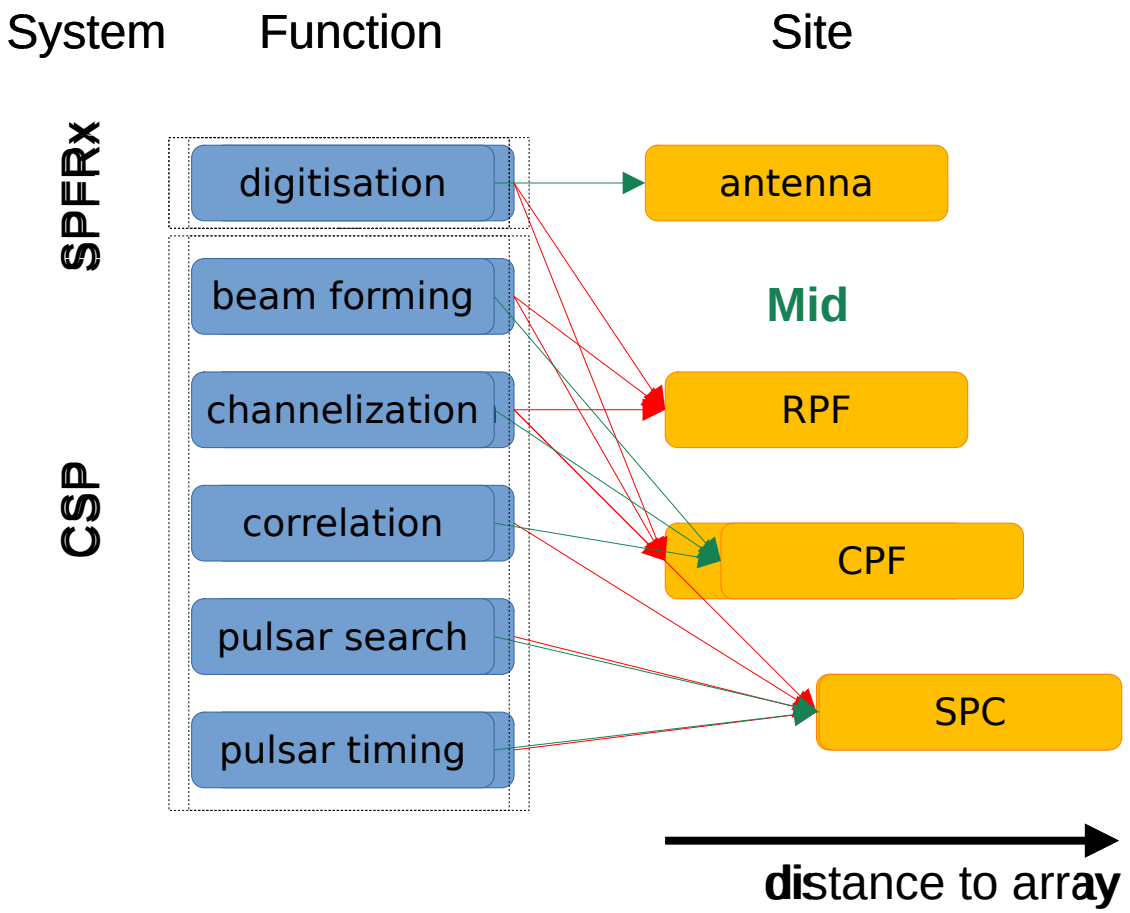
    12 bytes*130816 baselines*4 polar*65536 channels / 0.9 s = 0.46 TB/s
  - temporal data analysis
    - Pulsar search (PSS) in 500 beams
    - Pulsar timing (PST)

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
- Strategy for sizing compute resources (CPF, RPF):
  - data flow warrants stream processing
  - well identified/repetitive tasks and culture point to FPGAs
  - GPUs are being considered

Architecture (II)

- Mapping: delocalize and concentrate
- Criteria
  - accessibility
    - operation staff
    - maintenance
    - extension
  - 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 science: diversity of investigations and users
- 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
  - pointing, beam forming
  - calibration
  - Quality assurance (first look)

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

SDP products

- Pipelines & data products
  - imaging: fast imaging for transient objects, continuum images, calibrated visibilities, de-dispersed cubes
  - spectral analysis: spectral line cubes, resolved spectra
  - temporal analysis: pulsar timing solutions, pulsar candidates
  - calibration: sky model, telescope model
- Internal products
  - closing the control loop for pointing and beam forming
  - maintenance of calibration models
  - live quality assurance for operators

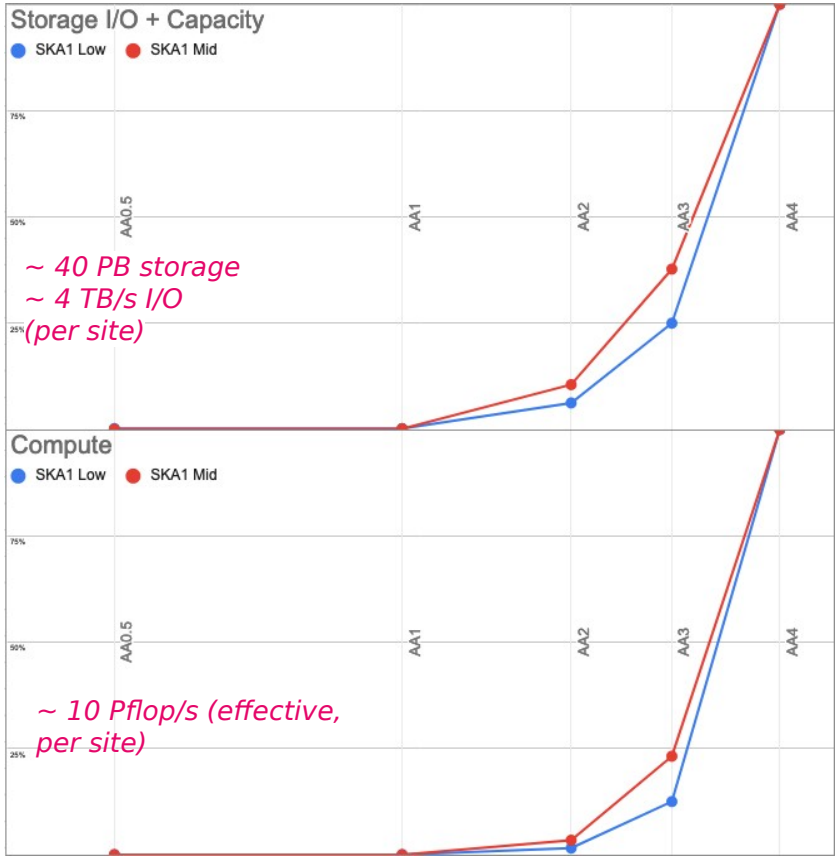
SDP Challenges (general)

- Trustworthy: reduce the volume of data by delivering data products and discarding input data (as 0.45 TB/s is ~ 40 PB/day)
- Reactive: SDP resource availability conditions observation scheduling
- Largely autonomous operation: large and complex HPC system
- Evolutive
  - extension beyond SKA1
  - SW/HW upgrades over 50-year lifespan of SKA: new science, new technologies
- Portability
  - SW able to run on SRC (multi-epoch analysis, reproducibility)
  - HW able to run external SW (commissioning)

SDP Challenges (technical)

- Scale to SKA1: much higher data rates than existing telescopes
  - exponential increase
  - AA2 is already a real challenge: scale to process data within 24h
- Large combination of operational conditions
  - problem sizes: channels, beams, subarrays
  - commensal observations
  - direction-dependent effects: RFI, ionospheric weather, wide field of view
  - combination of real-time (15 s latency) and batch scheduling (max 24 h delay)

Estimated SDP Scaling: AA1→AA4
(~50x in 17 months! Qualitative only, underestimates the AA2 situation)



Sizing SDP (I)

- Compute requirements: count floating-point operations
  - based on parametric model (Python)
  - notably for baseline-dependent visibility averaging, facetting, w-snapshots, w-stacking, IDG
  - 12.5 PFlop/s effective: 125 PFlop/s crest assuming 10% efficiency
    (HPCG 2022 has 1%-3% on major Top500 systems)
- Compute and buffer requirements vs. observation scheduling
  - average effective requirements based on High Priority Science Objectives scenarios built by Science Team
  - Low: 8.47 Pflop/s & 47.4 Gb/s output data rate
  - Mid: 5.09 Pflop/s & 28.4 Gb/s output data rate

Sizing SDP (II)

- Memory and traffic requirements: detailed modeling of constraints imposed by imaging and calibration within allocated time
  - large volume of visibilities & large queries by algorithms: large working sets to store in RAM or transfer on network
  - MeerKAT and Askap require some nodes with TBs RAM: will not scale well for SKA
  - need for careful load balancing and modifiability of execution (high-level execution framework)
  - max requirements identified are: 356 GB per node, 517 GB/s traffic between compute islands for calibration with Mid 1
- Cost
  - SDP capital cost: HW 35 M€ per SDP / SW 55 M€ for data handling and processing
  - SDP power is the single largest operational cost item in the SKA budget

Power

- SPCs in Cape Town and Perth
  - will host 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 effiency

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**

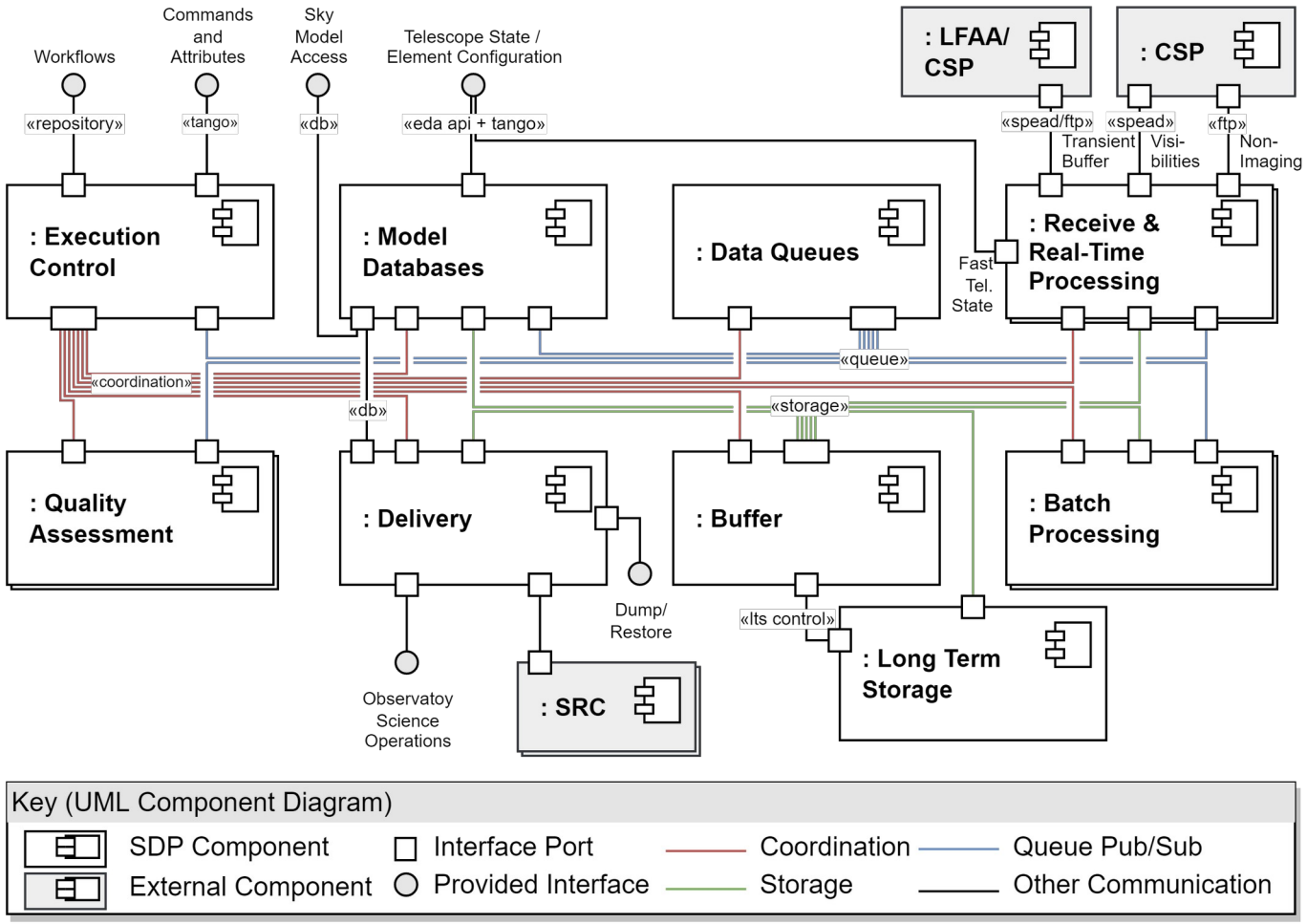
SKA1-Low SPC/SOC Power Budget in Perth				
Products	*AA4 Long Term Average (>30min) [kW]*	*Peak Instantaneous (<5sec) [kW]*	*AA* Long Term Average (>30min) [kW]*	*AA* Peak Instantaneous (<5sec) [kW]*
PDT6 - Network & Computing	**1629.2**	**2270.9**	**429.2**	**598.4**
OMC Hardware LOW	12.9	18.1	12.9	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**

SDP development strategy

- Current status
  - SW actively developed but still very preliminary
  - evolution of SW stack/HW technology/price difficult to predict over coming years
- Builds on existing
  - SW: whole stack from HPC to radio-astronomy algorithms
  - HW: COTS but with accelerators, network etc. quite a diversity (large design space)
- Shorter-term vs longer-term: AA0.5-AA1 to de-risk, AA2+ to scale
- Open-minded
  - BSD 3 license: share findings with communities
  - SAFe Agile approach

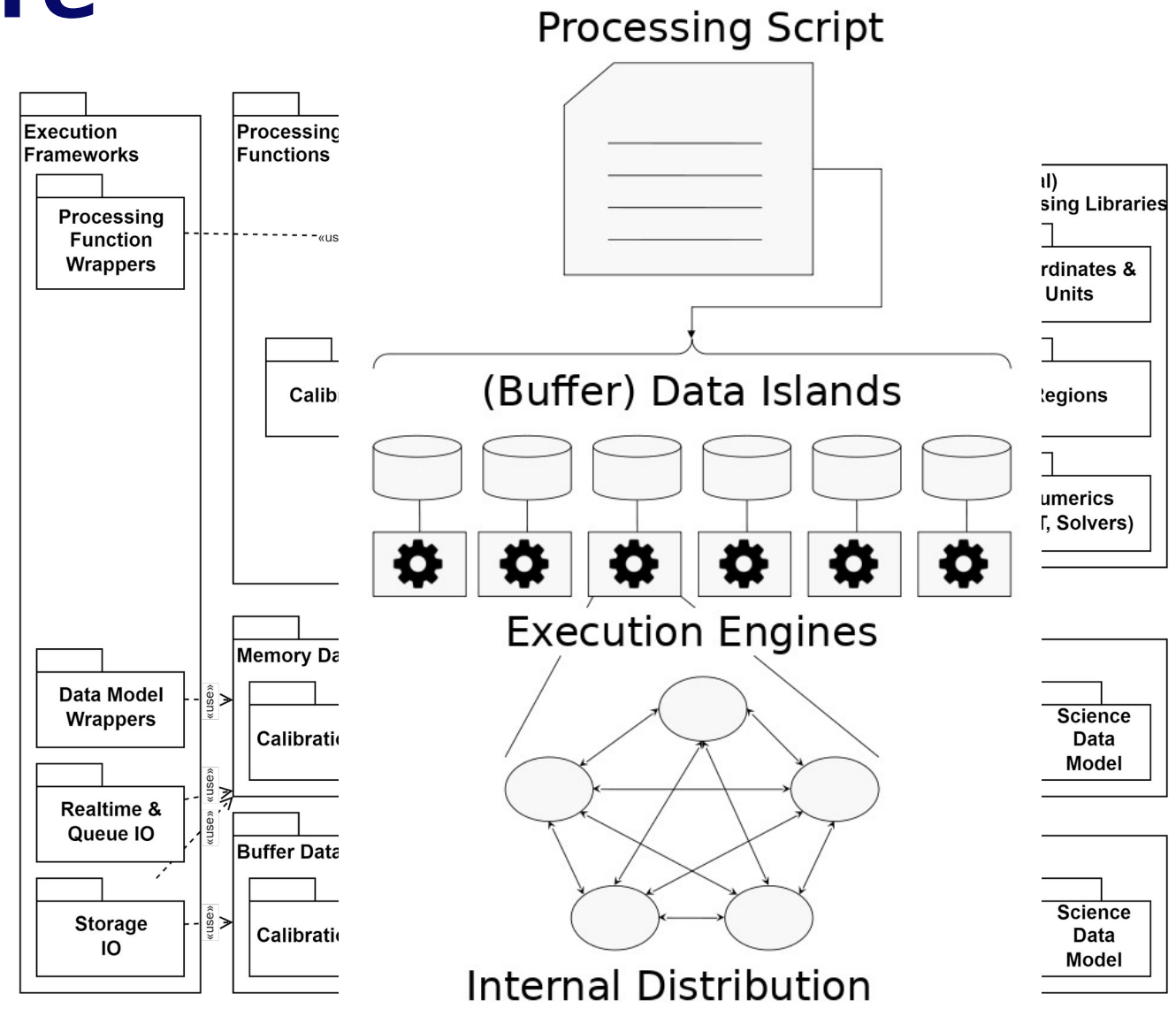
SDP C&C architecture

- Batch & real-time processing
- Central role of buffer
- Data queues: data flow approach to optimizing execution



SDP SW architecture

- Layered scaling concept
  - processing scripts
  - data islands
  - execution engines to schedule within compute islands
  - processing functions
- Data models
  - buffer and memory
  - abstract storage from algorithmic use



SDP HW architecture

- generic versatile servers
  - latency-optimised cores (CPU), throughput-optimised cores (accelerators)
  - latency-optimised network, throughput-optimised network
  - capacity storage, performance storage
  - 4 personalities assigned to servers depending on HW capacity and needs
  - receive
  - processing
  - services
  - storage

Processor Platform			
Compute	**Low**	**Mid**	**Units**
Number of nodes	1,819	1,595	
GPUs Per Node	2	2	
Peak Performance per GPU (project	38	38	DP TFlops
Memory per Node	320	512	GB
Storage per Node	See Below	See Below	
BDN Connection	25 GbE	25GbE	GbE
LLN Connection	100.0	100.0	
DPN Connection	10	10	GbE
Number of compute islands	34	30	
Nodes per compute island	56	56	
Number of management islands per	1	1	
Buffer::Cold	75.00 %		
Total Ingest rate (Required)	0.707	0.775	TB/s
Buffer Size	38	32	PB
Buffer storage per node	21	20	TB
Buffer storage per island	1.5	1.4	PB
Ingest rate per node (Required)	3.1	3.9	Gb/s
Under-subscription at CN	8.0	6.4	
Required Ingest rate per Compute Is	166	207	Gb/s
Buffer::Hot	25.00 %		
Buffer Size	13	11	PB
Buffer Size Per Node	7	7	TB
Total Read Rate (Required)	4.414	4.805	Gbyte/sec
Read Rate per node	4.41	4.81	Gbyte/sec
Read Rate per Compute Island	247.18	269.08	Gbyte/sec
Networking			
Application Network (LLN)	1:1		
Number of 100 GbE Ingress Ports	78	89	GbE
Bulk Data Network (BDN)	Over-subscribed 100GbE:25GbE		
Data Preservation Network (DPN)	Over-subscribed 10GbE:10GbE		
Management Network	1 Gbps	1Gbps	
LLN Bi-Sectional Bandwidth	2.8422	2.4922	
Infrastructure			
Number of Racks	48	43	

Co-design

SDP co-design

- CDR scenario
  - projections based on Moore’s law and past trends: constant improvement of end-user performance at fixed price
  - buy machine as late as practicable
- Computing HW risk mitigation plan: system is feasible but is it achievable by SKAO?
  - potential of emerging technologies (heterogeneous computing)
  - risks vs sizing and SW development strategy
  - procurement strategy & cost (collaborate with suppliers)
  - power
- Agile team leading co-design (SCOOP) mainly for SDP (but can have implications beyond)

SCOOP’s vision for co-design is still work in progress and not yet approved by SKAO

SCOOP co-design roadmap

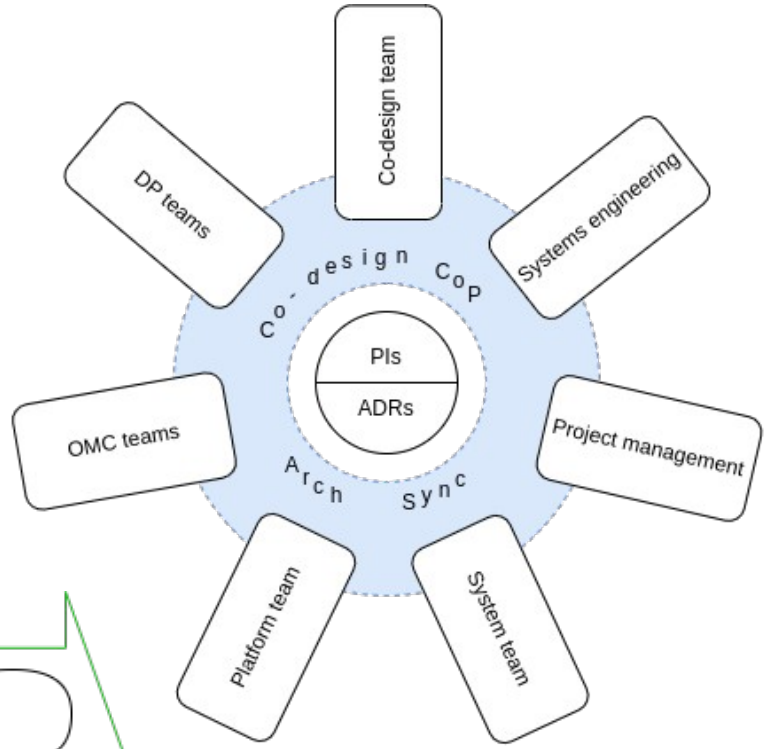
- Contribute to SDP modelling as an extension of existing work
- Evaluation of SW vs. reference platforms: refine/consolidate estimates by benchmarking
- Adaptive management of SW pipelines & SDP resources
  - insufficient resources (sizing and staged deployment of SDP)
  - increase of load (commensal observations, extension of arrays)
- Technology watch & collaboration with industry
  - reference benchmarks for acceptance testing
- Long-term vision of managing SDP SW & HW: procurement risk reduction through AAs, reliability, upgrades, sustainability, heterogenous HW
- Failure management (redundancy, check pointing vs. availability requirement)

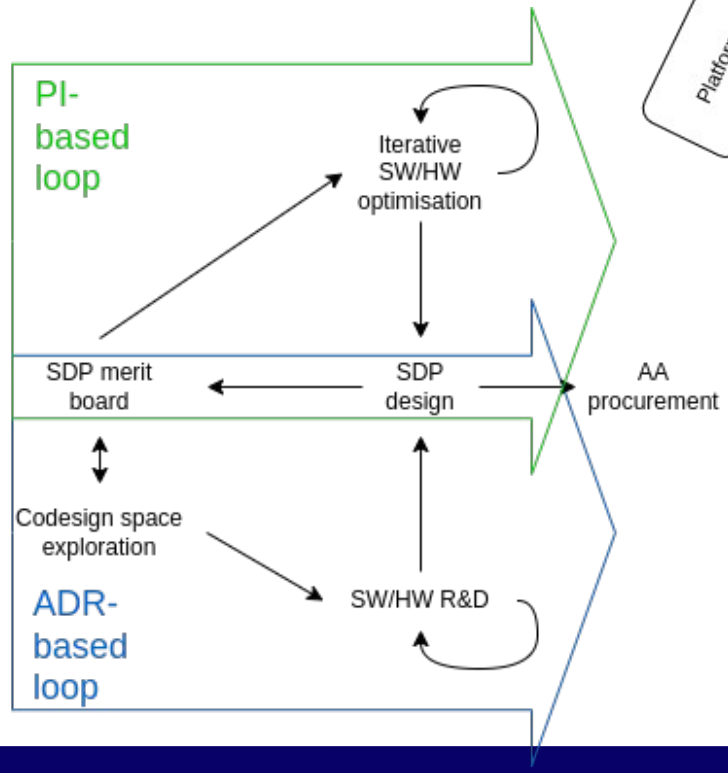
SCOOP co-design technical topics

- Simulation of execution for comparing SW/HW alternatives
- Energy optimisation and management
- Identification and allocation of largest jobs on SDP resources
- Buffer use and organization: volume/latency trade-off, hierarchy, disk/SSD
- Possibility to offload processing from SDP to SRC to optimize availability of SDP via partial data reduction and residuals without compromising data product quality
- Generic/specialized node trade-off (ingest, buffer, streaming, batch)
- SW optimization: workflow organization, algorithms, precision, HW tuning/abstraction
- Impact of scheduling/resource sharing for multiple tasks
- RAM/network topology for distribution of jobs

SCOOP co-design organization

- Co-design within SKAO
  - On-going addition to effort: WBS item and budget
  - Transverse effort within SKAO: co-design team, DP, systems engineering, procurement
- Collaboration with community
  - SKAO based: PI-based inner loop
  - external R&D feeding SKAO effort based on proofs of concept





Perspectives (co-design)

- AA1 in 2025: test co-design alternatives as a secondary objective
- AA2 2026: de-risk co-design solutions
- AA*: integrate AA2 machine into SDP (as AA* would eventually be integrated in larger system)

- Co-design timeline vs. PIs, AAs, procurement, operations/maintenance/upgrades
- ECLAT joint laboratory between CNRS, Inria, Atos for the upstream R&D
- French-Swiss collaboration on co-design

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