AA Processing Requirements for SKA1

Paul Alexander

Background

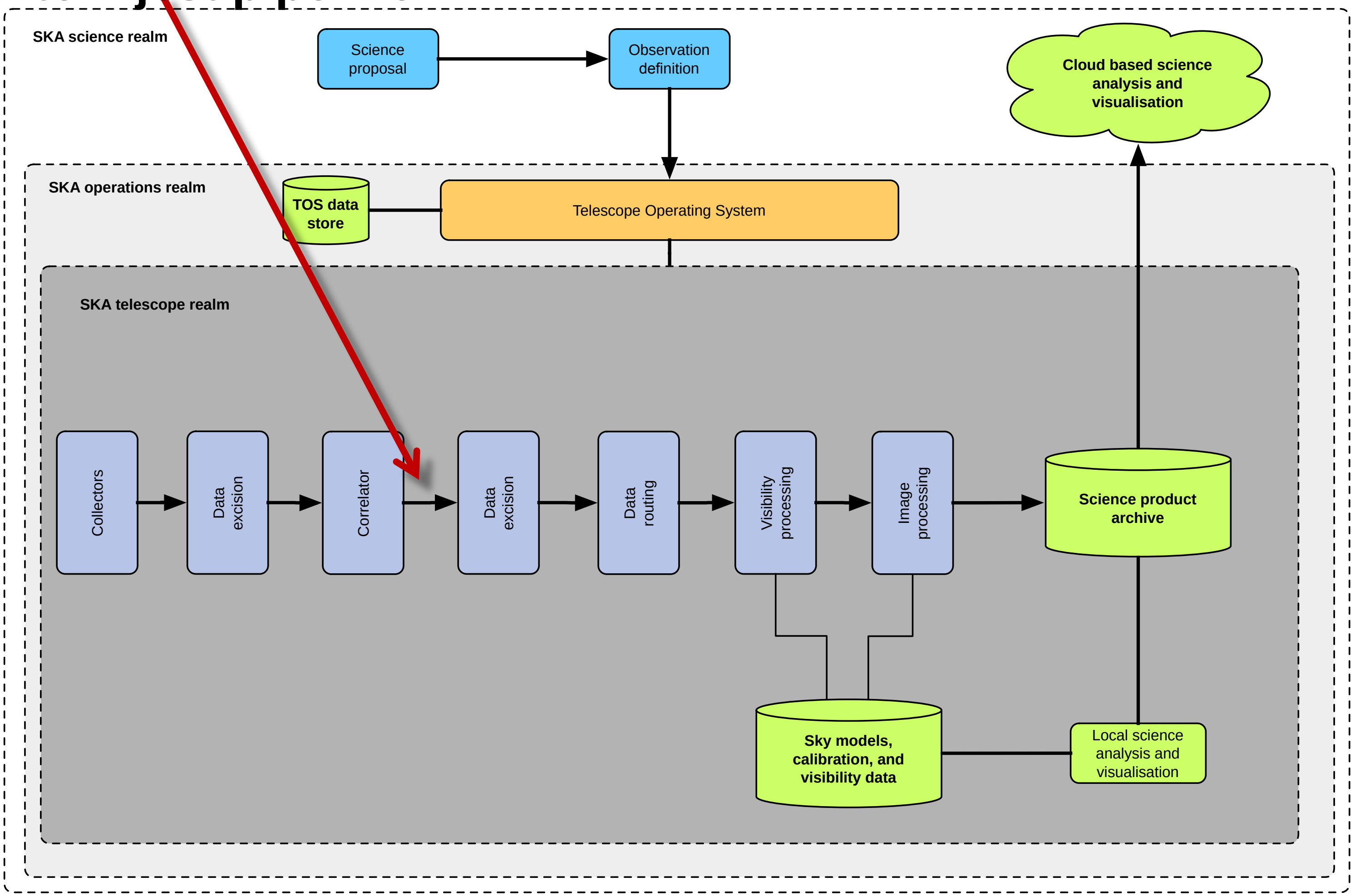
Based on work for the SKA S&C CoDR

- Analysis is to derive detailed input requirements to the S&C domain
- Analysis is based on DRM 1.3 and 2.0
  - requirements driven by EoR experiment
- Some discussion on pipeline requirements – drawing heavily on recent work by JB and TC
- Overall computational load depends on
  - Data rate
  - Complexity of imaging or analysis problem
  - Algorithm and its implementation

Plus some questions and input from you!

System concept for analysis

Data rates will be calculated here from correlator to Injest pipeline



EoR Imaging DRM Requirements

Science Requirements from the DRM		
Parameter	Value	Comment
Redshift coverage	6 – 30	
Brightness temperature sensitivity	1 – 3 mK	
Angular resolution	2' – 5'	
Radial resolution	2 Mpc	
Field of view	> 5 deg	Set by cosmic variance

EoR Imaging DRM Requirements

Technical Requirements from the DRM		
Parameter	Value	Comment
Frequency range	50-240MHz	
Critical frequency	100 MHz	
Frequency resolution	100 kHz	RFI excision is critical and may need high resolution ~ 1 kHz
Bandwidth	$\Delta f/f \sim 1$	Cover complete frequency range in each observation
Maximum baseline (core)	5km	To provide angular resolution
Baseline source subtraction	~200km	
Integration time	>1000 hrs	Set by cosmic variance
A/T	$>1000\ m^2K^{-1}$	
Antenna diameter	7m – 30m	
Core UV coverage	$N_d > 160$	

Analysis

Channel requirements

- Straight forward
  - 1.7×10^5 at 1 kHz resolution for RFI excision
  - 1.7×10^3 in the final data products
  - Data rate drops by this factor after the injest pipeline

Sensitivity and Collector distribution

- **Requirement:**

  10mK in a 5' beam and 3.3mK in a 2' beam

- From SYS_REQ_1310 the requirement is that A/T = 1000 m^2K^{-1} across the 70-450 MHz band of the AA-low.

- Translated in Memo 130 as a total collecting area of 1.25×10^6 m^2 distributed in 50 180-m stations with a distribution of:

Core (r <0.5 km)	~50% (25 stations)	6.25×10^5 m^2	**f = 0.81**
Inner (1< r<2.5 km)	~20% (10 stations)	2.5×10^5 m^2	
Mid (2.5<r<100 km)	~30% (15 stations)	3.75×10^5 m^2	

Analysis

Sensitivity and Collector distribution

- High filling factor in core means ***flexibility*** in logical configuration
  - Very important to meet EoR requirement
  - ***Extensibility*** to SKA2 gives filling factor ~1 in inner region
- **Resolution**:

  2' corresponds to ~ 6km at 70MHz
  2' corresponds to ~ 2.5 km at 240MHz

N.B. would still need beam forming across the full band

- DRM1.3 matches "station" diameter to 5 degree FoV giving D = 30m
  - In Inner region:
    N ~ 1200, but data rate scales as N^2
    Adopt instead requirement on UV coverage and take 200 75m stations
    Beyond 2.5km 85 70m stations or 15 180m stations

Dynamic Range

N.B. may need to consider more sophisticated definition of dynamic range

DRM1.3 gives the flux densities of the faintest EoR structures to be imaged:

- ~0.3mJy/Beam (1σ) at 100 MHz.
- Jonathan made the point yesterday, source contamination is worse than smooth foregrounds
- in a 25 sq-degree field an order of magnitude estimate would suggest that we would expect to find a 3C brightness object
- even by selecting a region with no 3C-like source, consideration of the source counts suggest it seems very likely that the field will still be contaminated by a number of sources with a flux > 1Jy
- This implies a dynamic range requirement of >65dB.

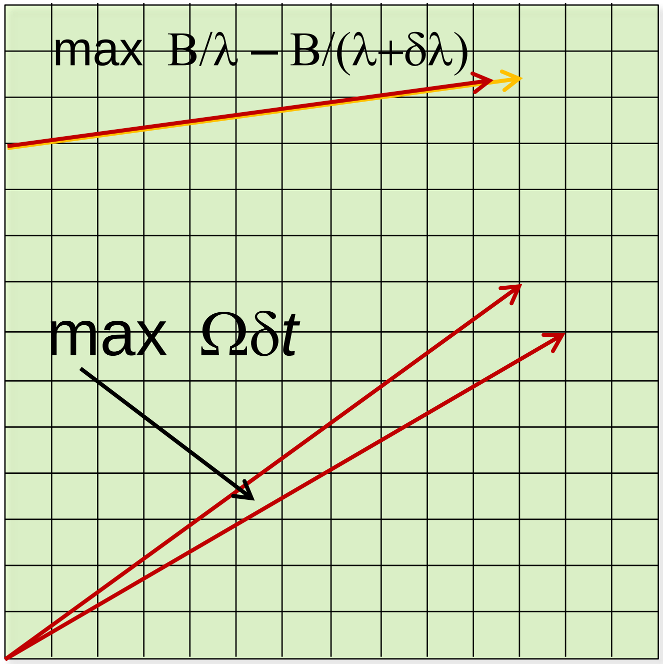
Correlator Output Data Rates

- For imaging, after correlation the data rate is fixed by straightforward considerations
  - Must sample fast enough (limit on integration time) δt
  - Baseline $\propto$ B/λ
  - UV (Fourier) cell size $\propto$ D/λ

$$\Omega \delta t \frac{B}{\lambda} < \frac{1}{X} \frac{D}{\lambda}$$

  - Must have small-enough channel width to avoid chromatic aberration

$$\delta \left(\frac{B}{\lambda} \right) < \frac{1}{X} \frac{D}{\lambda}$$



Correlator Output Data Rates

- Adopt criteria similar to EVLA but using isotropic smoothing kernel in UV-plane

$$\frac{\delta t}{s} = a_t \frac{D}{B} \sim 1200 \frac{D}{B} \qquad\qquad \frac{\delta f}{f} = a_f \frac{D}{B} \sim \frac{1}{10} \frac{D}{B}$$

- Data rate then given by

$$G = g(B) \frac{1}{2} N^2 N_p^2 N_b \frac{1}{\delta t} \frac{\Delta f}{\delta f} 2 N_w \qquad\qquad G = g(B) N^2 N_w N_p^2 N_b \frac{1}{a_t a_f} \frac{\Delta f}{f} \left(\frac{B}{D}\right)^2$$

antennas # polarizations # beams word-length

- Can reduce this through the injest pipeline using additional smoothing or baseline-dependent integration times and channel widths

S&C Requirements

N_D	=	200
N_{ch}	=	1.7×10^5
N_b	=	16
B_{max}	=	5km

G_{out}(RFI)	=	27.5 GB/s
G_{out}	=	275 MB/s
δt	=	18 s
N_{ch}	=	1.7×10^5

N.B. 200 times larger for 30-m logical stations

- AA with 45 degree scan allows 5hr track per day
- 1000 hr total integration gives 200 days
- Each observation is 500 TB UV data (1kHz) reducing to 5 TB
- 150 GB per day of processed data cube at 100 kHz channels

Do we need to store UV data until complete 1000 hr integration complete?

YES Some analysis approaches will require this

S&C Requirements – long baselines

75-m station

G_{out}(RFI)	**=**	**2500 GB/s**
G_{out}	=	250 GB/s
dt	=	0.45 s
N_{ch}	=	1.9×10^4

180-m station

G_{out}(RFI)	**=**	**187 GB/s**
G_{out}	=	8.9 GB/s
dt	=	1.08 s
N_B	=	67
N_{ch}	=	7.9×10^3

- Even for 180m station with 200 km baselines full imaging
  - 160 TB of UV data per 5-hr track
  - Image product 16k x 16k x 6k (24 TB per field) with 133 fields
- For 25 km baseline
  - 2k x 2k x 1k (64 GB) per field 133 fields

Precise requirements for the calibration and source subtraction need careful consideration as they could drive requirements for S&C domain and hence SKA

Wide Field Imaging

Standard problem

$$V(u,v,w) = \int \left[\frac{I(l,m) e^{j2\pi w\left(\sqrt{1-l^2-m^2}-1\right)}}{\sqrt{1-l^2-m^2}} \right] e^{j2\pi(ul+vm)} dl dm$$

Fresnel number $R_F = wl^2/2 = B\lambda/D^2$

Some Approaches

- 3D-imaging — l,m,n space with $n = \mathrm{sqrt}(1-l^2-m^2)$
- Faceting — Image in regions where small-angle approximation applies (or UV facets)
- Snapshot imaging — Instantaneously project array onto a single plane
- Aw-projection — Use a modified convolution kernel to

Wide-Field Imaging?

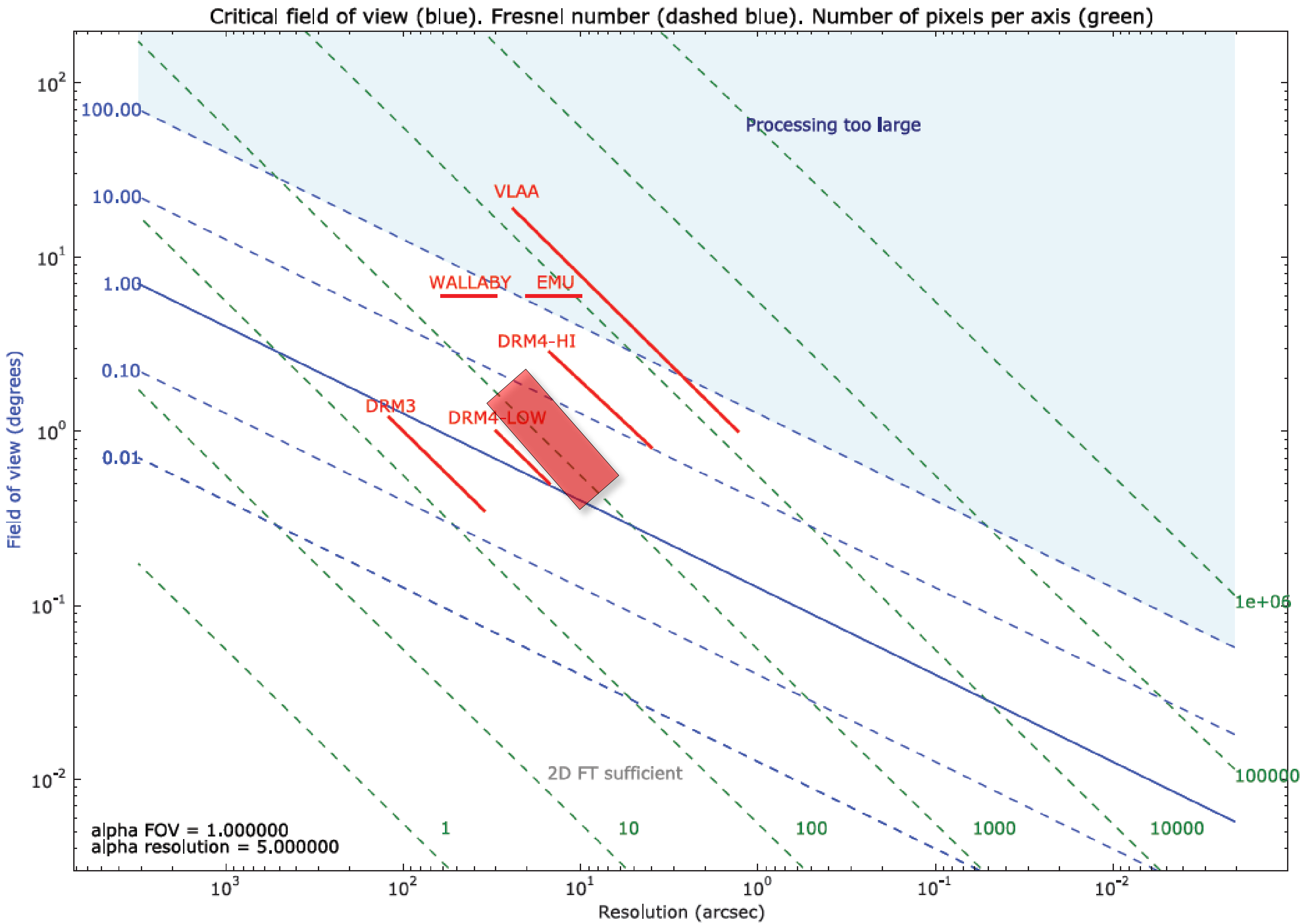
Fresnel number $R_F = B\lambda/D^2$

consider critical frequency 100 MHz

75-m station	
B	**R_F**
5 km	2.7
25 km	13
200 km	107

180-m station	
B	**R_F**
5	0.46
25 km	2.3
200 km	18.5

- The imaging problem is a wide-field problem but not severe
- For 30-m logical stations 6.25 larger

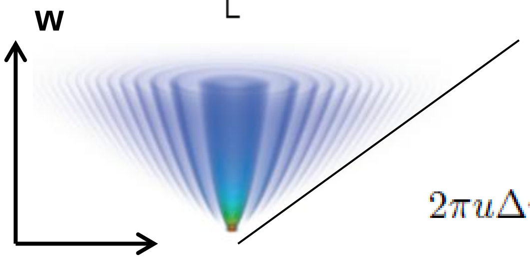


Tim Cornwell

Consider w-projection

$$V(u,v,w) = \int \left[\frac{I(l,m) e^{j2\pi w\left(\sqrt{1-l^2-m^2}-1\right)}}{\sqrt{1-l^2-m^2}} \right] e^{j2\pi(ul+vm)} dl dm$$

$$V(u,v,w) = \tilde{G}(u,v,w) * V(u,v,w=0)$$



Convolution in data space

$$G(\ell,m,w) = e^{-i\pi[w(l^2+m^2)]}$$

$$2\pi u \Delta u / w \geq 2\pi$$

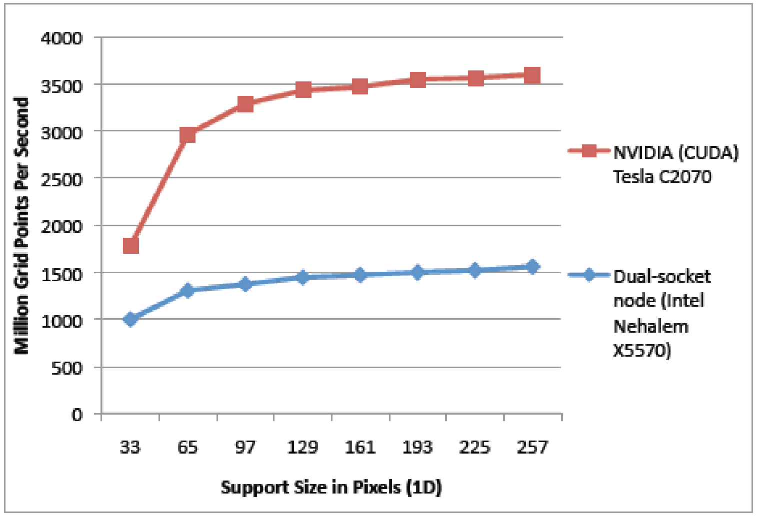
$$\tilde{G}(u,v,w) = \frac{1}{iw} e^{-i\pi \frac{(u^2+v^2)}{w}}$$

$$N_{supp} = 8 \frac{w}{(\Delta u)^2} = 8 R_F$$

oversampling

Number of operations proportional to number of gridded points

$$N_p = N_{vis} \times \overline{N_{supp}}^2$$



- N_{vis} = 30 GS/s
- Mean(N_{supp}^2) ~ 338
- N_p ~ 10^{13} s^{-1}
- Achievable with 3000 Tesla or 6000 cores

But is this the best approach?

W-projection

- Good: Produces single field; accesses each sample once therefore efficient
- Bad: Need to recalculate kernel often to allow for changing beam in Aw projection

Faceting not good for these

$$V_{A'B} = \int I(l,m) e^{-2\pi j\left(u\left(l - a\sqrt{1-l^2-m^2}\right) + v\left(m - b\sqrt{1-l^2-m^2}\right)\right)} \frac{dldm}{\sqrt{1-l^2-m^2}}$$

Snapshot imaging

Instantaneous reprojection onto distorted tangent plane – stack of 2D FT's in time which are combined in image plane

- Good: Produces single field; small kernel; calculating A-projection goes in sync with timing of snapshot; single access to each visibility
- Unknown: Cost of doing image-plane reprojections;

Quantitative simulations and real results – input from LOFAR?

For Pre-construction

- Detailed analysis and development of algorithms and implementations
- Continuous cycle of develop – test – deploy – analyse
- Expect algorithm development and implementation on HPC platforms to go hand-in-hand
- One solution very unlikely will need a tool box for different experiments, but also hybrid approaches with balance changing dynamically

For Pre-construction

- Detailed analysis and development of algorithms and implementations
- Continuous cycle of develop – test – deploy – analyse
- Expect algorithm development and implementation on HPC platforms to go hand-in-hand
- One solution very unlikely will need a tool box for different experiments, but also hybrid approaches with balance changing dynamically

Other issues

- Until proven otherwise will assume a multi-pass approach is required to achieve high dynamic-range imaging
  - Sequoia will have a Lustre file system with a 50 PB Lustre File store with an access of 1 TB/s provided by NetApp
  - This is what we need to deliver the above
  - Imaging pipeline is not all that is required – direct statistical processing of the UV data will also be needed (c.f. CMB analysis)

Conclusions

- AA processing for SKA1 is achievable
- EoR experiment dominates the requirements
- Know what we have to do for the PEP phase