WP 2.6.5
SKA Data Products, storage

Paul Alexander

Task Overview

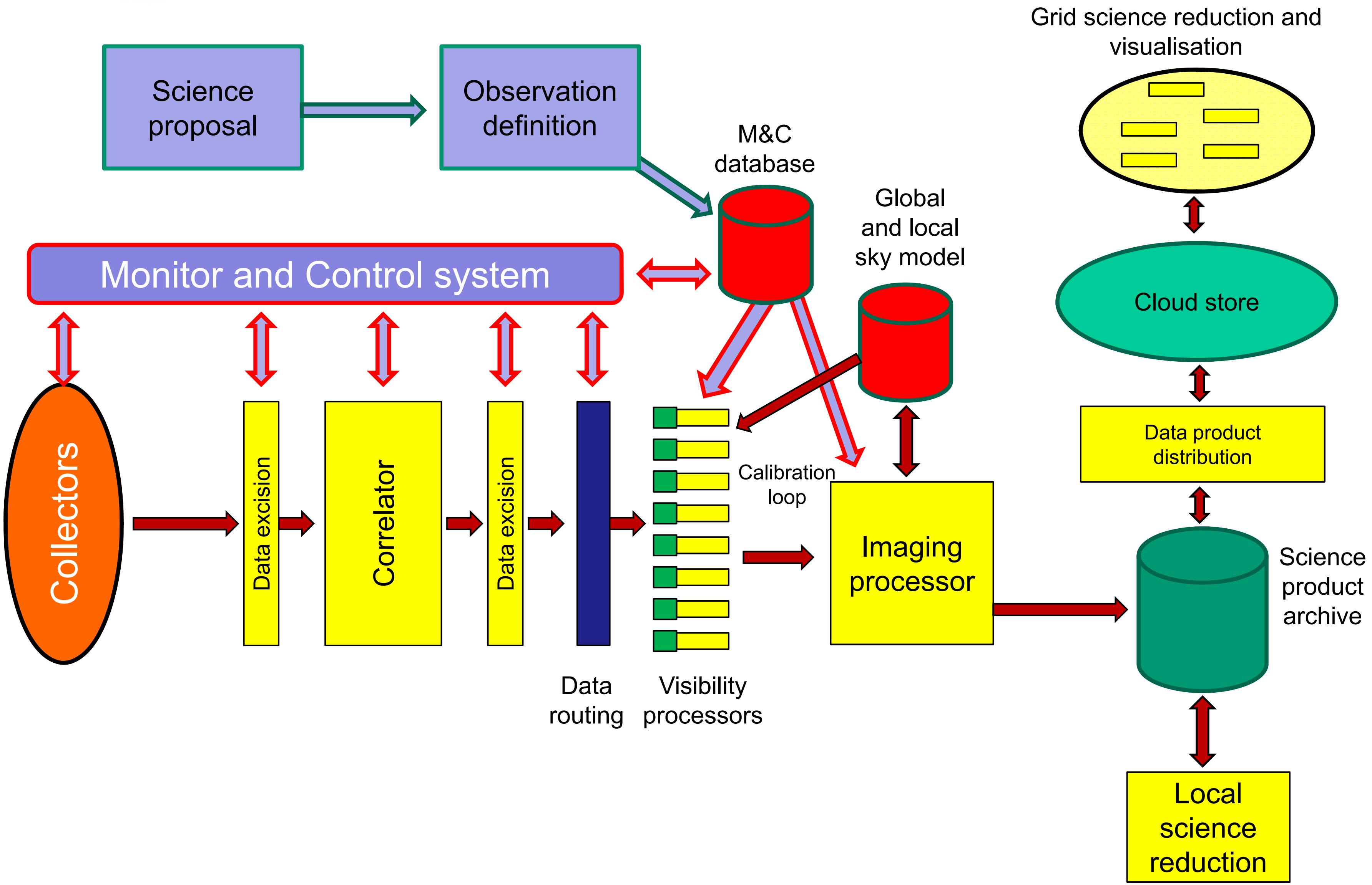
- Addressing data products, data storage and distribution
- Deliverables: elicit and document requirements for
  - Data products including data visualisation
  - Data storage – what data can be stored
  - Data distribution -
- Approach
  - Detailed analysis as required by DoW in the context of

Overall System View of Information and Data Flow – essential to establish requirements

Participants and activities

- University of Cambridge
  - Data system design; data product definition; Hardware/software architecture
- ICRAR
  - Data system design; Database design; Data product definition; Hardware/software architecture
- University of Calgary
  - CyberSKA; data distribution model; data visualisation
- ASTRON
  - Data storage; Hardware/software architecture
- JPL
  - Visualisation and data handling
- SKA-NZ
  - Hardware/software architecture

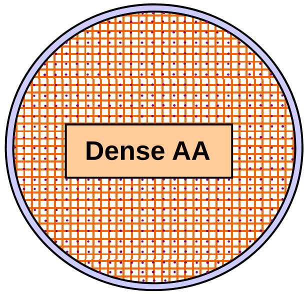
SKA Information and Data System



Data rates to the correlator

Data rate from each collector

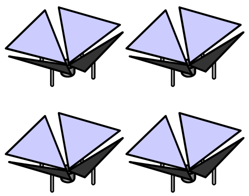
$$G_1 = 2\, N_p\, \Delta f\, N_{bit}\, N_b = 4\, \Delta f\, N_{bit}\, N_b \qquad N_b = \frac{1}{\Delta f}\int_{f_{max}-\Delta f}^{f_{max}} n_b(f)\, df$$



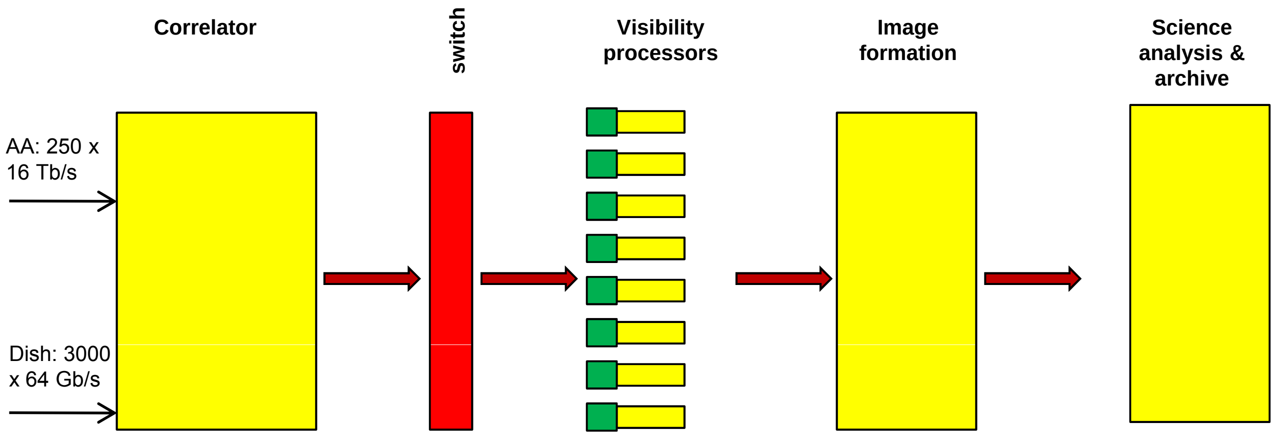
AA, Number of elements N_e ~ 65000; $N_b << N_e$ limited by data rate

250 sq-deg across band N_b ~ 1200 (Memo 100)

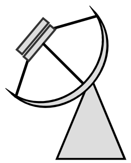
G_1 ~ 16 Tb/s



Sparse AA



12-15m Dishes



Dishes G_1 ~ 64 Gbs

PAFs FoV is constant across the band

N_b ~ 7 to give 20 sq-deg across the band G_1 ~ 60 Gb/s (Memo 100)

Data rates from the correlator

- Standard results for integration/dump time and channel width

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

- Naive data rate then given by

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

- Can reduce this using baseline-dependent integration times and channel widths

$$G = N^2 N_w N_p^2 N_b \frac{1}{a_t a_f}\frac{\Delta f}{f} \int_0^B n(b)\left(\frac{b}{D}\right)^2 \, db$$

$$= N^2 N_w N_p^2 N_b \frac{1}{a_t a_f}\frac{\Delta f}{f}\left(\frac{B}{D}\right)^2 \int_0^B n(b)\left(\frac{b}{B}\right)^2 \, db$$

SKA$_2$ data rates from the correlator

Experiment				3000 Dishes + SPF		1630 Dishes + PAFS		250 AA stations	
Description	B_{max} (km)	Δf (MHz)	f_{max} (MHz)	Achieved FoV[1]	Data rate (Tb/s)	Achieved FoV[1]	Data rate (Tb/s)	Achieved FoV[1]	Data rate (Tb/s)
Survey: High surface brightness continuum	5	700	1400	0.78	0.055	15	0.11	108	0.03
Survey: Nearby HI high res. 32000 channels	5	700	1400	0.78	1.0	15	2.0	108	2.6
Survey: Medium spectral resolution; resolved imaging (8000)	30	700	1400	0.78	1.2	15	2.4	108	5.4
Survey: Medium resolution continuum	180	700	1400	0.78	33.1	15	66	108	14.1
Pointed: Medium resolution continuum deep observation	180	700	1400	0.78	33.1			0.78	0.15
High resolution with station beam forming[2]	1000	2000	8000	0.0015	33.4				
High resolution with station beam forming[3]	1000	2000	8000	0.0015	429				
Highest resolution for deep imaging[2]	3000	4000	10000	0.001	391				

Notes

1. Achieved FoV is at f_{max} and has units of degrees squared. For the AA and PAFs we calculate the data rate assuming it is constant across the band.
2. Assuming that for the dynamic range the FoV of the station only has to be imaged
3. Assuming that for the dynamic range the FoV of the dish must be imaged

SKA1 Data Rates

- AA Line experiment 50 AA-low stations
  - 100 sq degrees
  - 10000 channels over 380 MHz bandwidth
    - ➢ 3.3 GS/s
- Dish Line experiment – 300 15-m dishes
  - 0.5 sq degrees
  - 32000 channels over 1 GHz
    - ➢ 6.1 GS/s

Where does the data rate drop?

For SKA_2
Data rate out of correlator exceeds input data rate for 15-m dishes for baselines exceeding ~ 130km (36km if single integration time)

At best for dishes output data rate ~ input; AA's reduction by $\sim 10^4$

- Image size: $a^2 N_{ch} (B/D)^2 N_b$ Ratio UV to "image" data

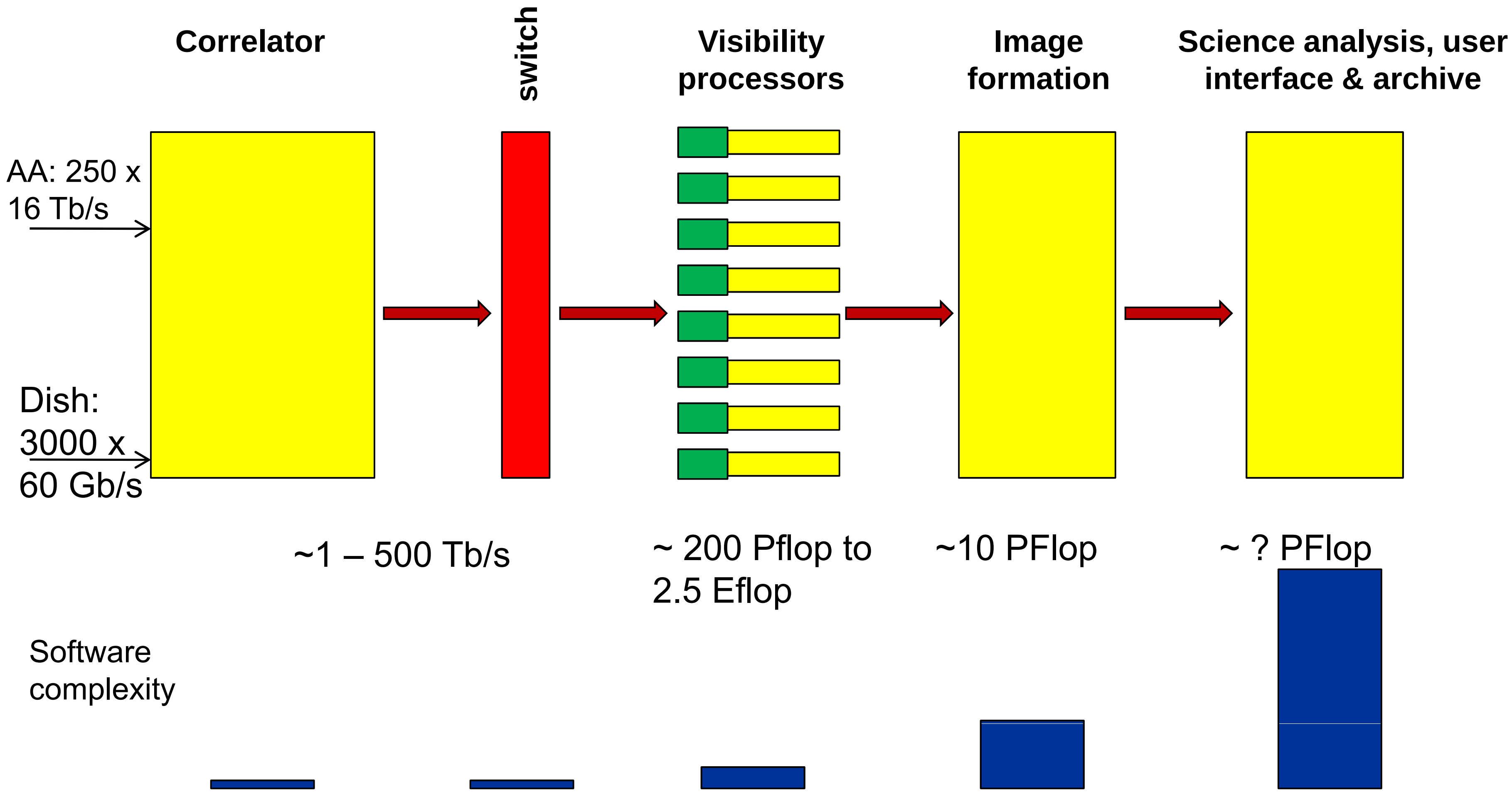
$$\sim 0.06\, T_{obs}\, N^2 g(B) \frac{\Delta f}{f} \frac{1}{a_t a_f} \frac{1}{a^2} \frac{N_p^2}{N_{ch}} \quad \sim 210 \left(\frac{T_{obs}}{1\mathrm{min}}\right) \left(\frac{N}{1000}\right)^2 \left(\frac{N_{ch}}{32000}\right)^{-1}$$

Major reduction in data rate occurs between UV data and image data

Post Correlator UV data Requirements

- Define UV data requirements from DRM
  - Good progress for SKA_2 DRM
  - Need to refine integration times and channel widths based on calibration understanding (↓), FoV shaping etc. (↓), RFI excision (↑)
  - Need SKA_1 DRM
- Persistence of M&C and flagging data associated with UV data
  - Current model quite traditional
  - Need to elicit information on all current approaches
- Embarrassingly parallel – need to consider performance of distributed file systems and data formats to obtain good performance

The SKA Processing Challenge



Model for SKA_1 UV processor

- Highly parallel – consider something achievable – NVIDIA promises 20 TFlop in 2 years – assume 50 Tflop in 2018 timeframe
- Approximate analysis of ops/sample: 200,000/calibration loop, 10^6 total
- 5 calibration loops, 20% efficiency,
- each processor processes ~ 0.01 GS/s of data
- Requirement: ~ 6 PFlop
- Buffer 1 hr of data therefore we need to buffer 100 GB in a fast store
- Require ~ 600 Blades, assume : €2000 per blade

UV processor ~ €1.2m

SKA_1 science product

- AA-low 100 sq degrees spectral line cube
- 20km baseline at 300MHz → resolution ~ 8 arcsec resolution
- ~1.5 $\times 10^8$ pixels; 1000 channels
- Volume size ~ 1.5 $\times$ 0^{11} voxels
- Data set size ~ 1 TB

Final data product ~ 1 TB

Science Data Products

Experiment	T_{obs}	*B*/km	*D*/m	N_b	N_{ch}	N_v	Size / TB
High resolution spectral line	3600	200	15	1	32000	$5\ 10^{13}$	200
Survey spectral line medium resolution	3600	30	56	1000	32000	$8\ 10^{13}$	330
Snapshot continuum – some spectral information	60	180	56	1200	32	$7\ 10^{12}$	30
High resolution long baseline	3600	3000	60	1	4	$7\ 10^{14}$	360

- ~0.5 – 10 PB/day of image data
- Source count ~10^6 sources per square degree
- ~10^{10} sources in the accessible SKA sky, 10^4 numbers/record
- **~1 PB for the catalogued data**

100 Pbytes – 3 EBytes / year of fully processed data

Summary

- Essential to adopt a full systems-based approach to analysing the data flow and requirements
- Good initial progress defining intermediate and science data products, but need SKA_1 DRM to be definitive
- System model assumes highly parallel data model
  - ❑ Work needed to define better this intermediate data model
- Data distribution and visualisation
  - ❑ See separate CyberSKA talk
- No work yet on DB or archive aspects of requirements

Processing model