SDP Data Products and Rates

Rosie Bolton

University of Cambridge
Project Scientist for Science Data Processor Consortium

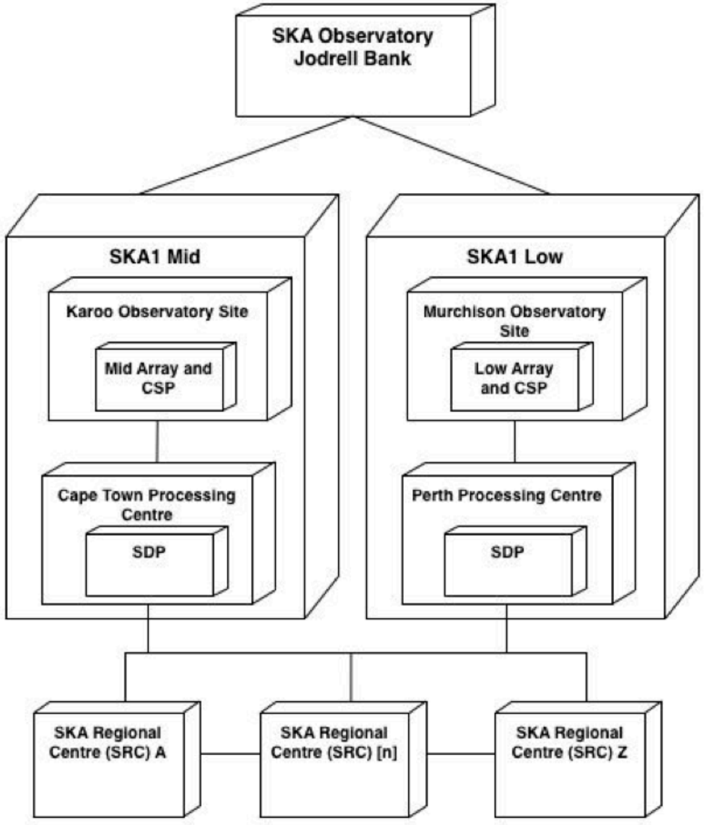
SKA SQUARE KILOMETRE ARRAY SDP

SCIENCE DATA PROCESSOR

UNIVERSITY OF CAMBRIDGE

Summary

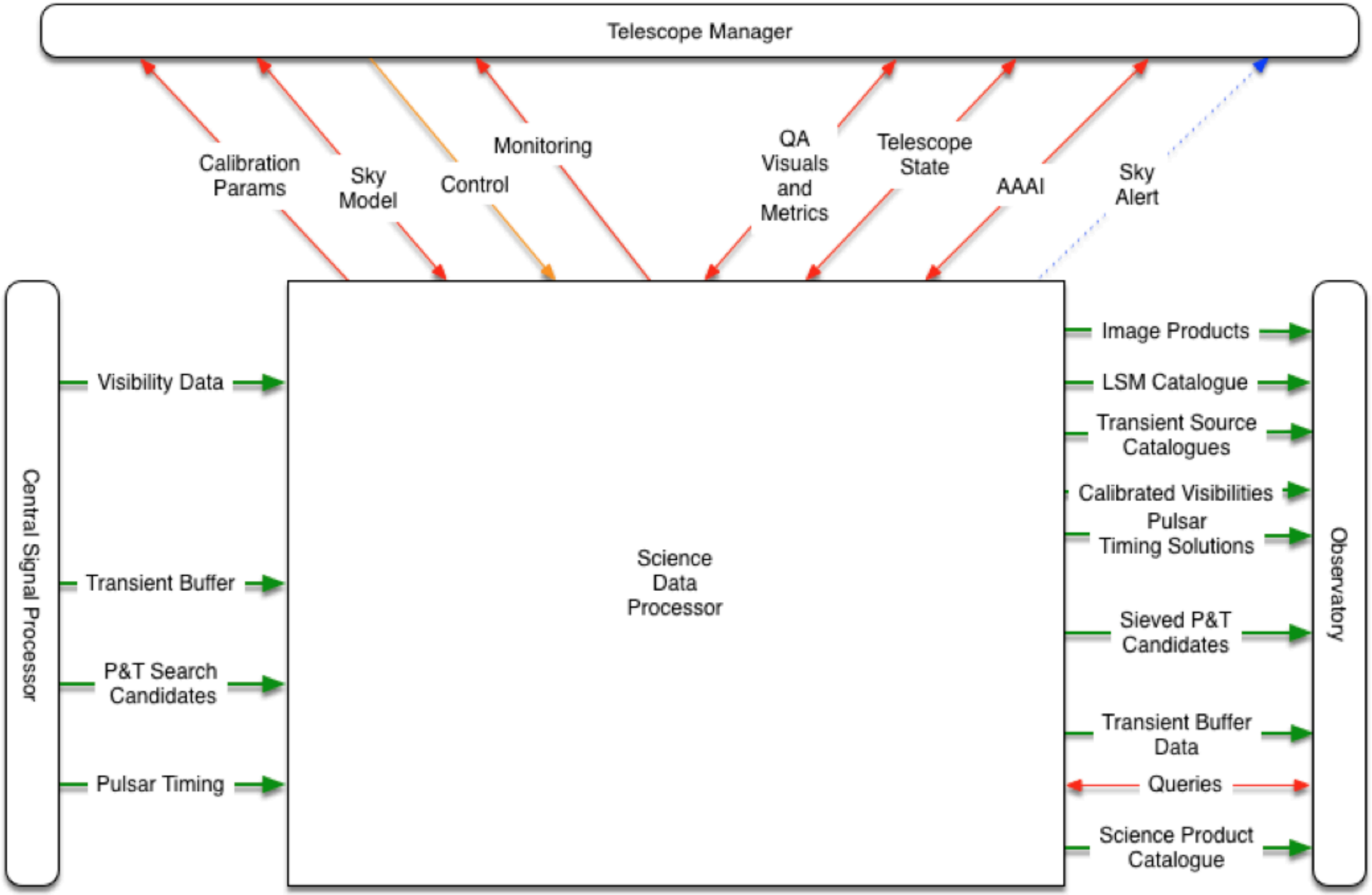
- SKA Scope: Is limited!
- SKA will deliver only from a subset of standard **data products**.
- Paradigm shift: The SDP compute is a resource, in an identical way to the "on-sky" time.
- No option for reprocessing data in current design.
- But the board has adopted regional centres…



SKA Observatory Jodrell Bank
SKA1 Mid
Karoo Observatory Site
Mid Array and CSP
Cape Town Processing Centre
SDP
SKA1 Low
Murchison Observatory Site
Low Array and CSP
Perth Processing Centre
SDP
SKA Regional Centre (SRC) A
SKA Regional Centre (SRC) [n]
SKA Regional Centre (SRC) Z

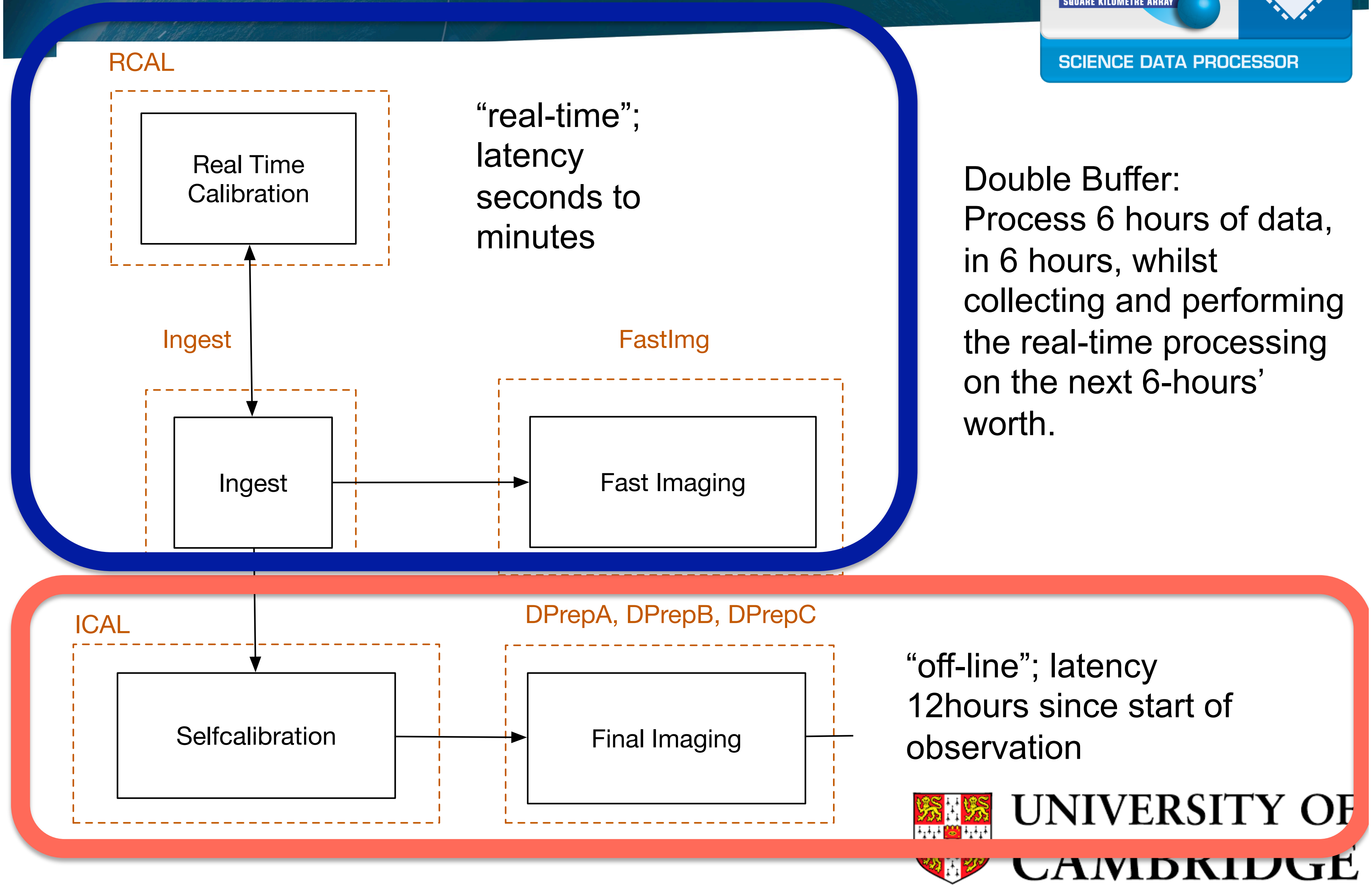
SKA Data Products from Science Data Processor

SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR



UNIVERSITY OF CAMBRIDGE

SDP Visibility processing



“real-time”; latency seconds to minutes

Double Buffer: Process 6 hours of data, in 6 hours, whilst collecting and performing the real-time processing on the next 6-hours’ worth.

“off-line”; latency 12hours since start of observation

Image products

- **Continuum Taylor Term** Images

- Coarsely channelised "continuum" full Stoke's images (~500 channels)

1. **Residual image (continuum & spectral line)**

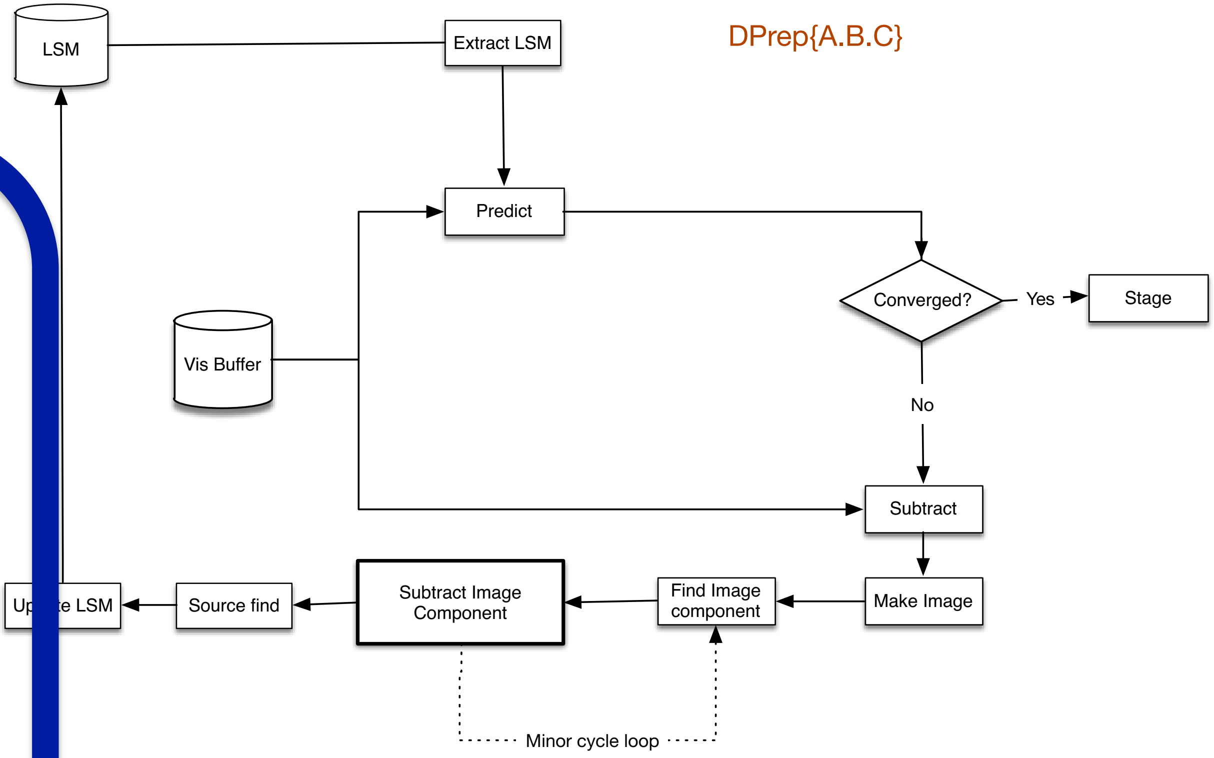
2. **Clean component image (continuum and spectral line)**

3. **Spectral line cube** (continuum subtracted)

4. **Calibrated gridded visibilities**, at the spatial and frequency resolution required by the experiment.

5. **Accumulated Weights** for each uv cell in each grid.

- Representative Point Spread Function for observation



DPrep{A.B.C}

Auxiliary products multiply up the required archive size (by a factor of 5-10) but they are necessary to enable science product generation

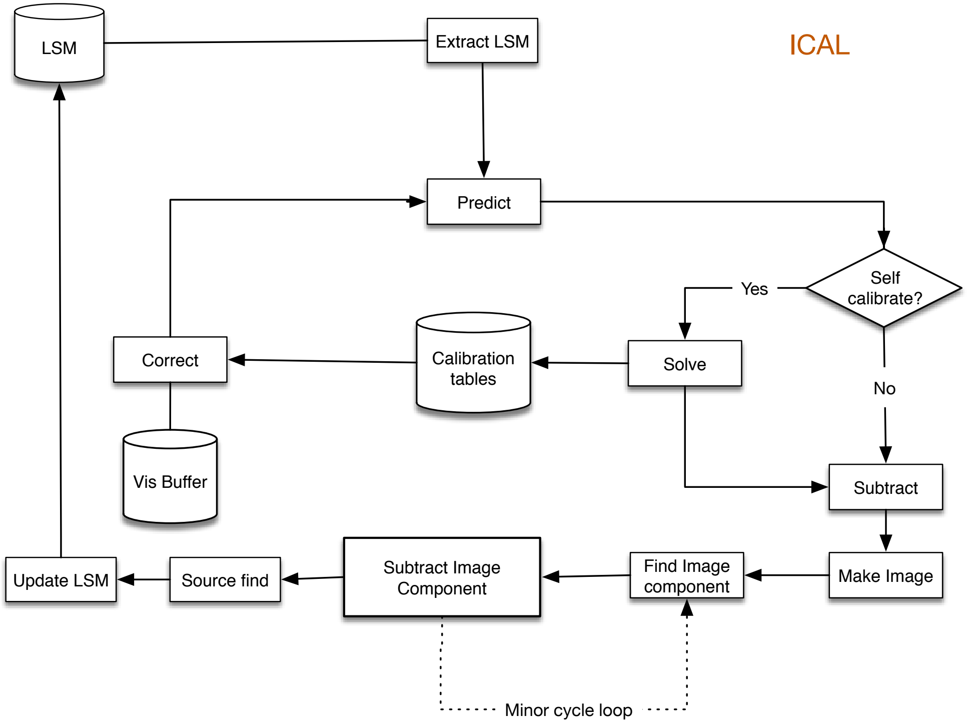
SDP Visibility processing: Parametric model

SQUARE KILOMETRE ARRAY

SCIENCE DATA PROCESSOR

Pipeline: ICAL, Self-calibration
Extract LSM; Update LSM; Notify LSM . . . (not yet modelled)
Predict via DFT $C_{\mathrm{Predict}} = (1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} 32 N_{\mathrm{source}} \times R_{\mathrm{vis,full\,FoV}}$
Solve $C_{\mathrm{Solve}} = (1 + N_{\mathrm{SelfCal}}) N_{\mathrm{beam}} N_{\mathrm{pol}} \left(\frac{1}{t_{\mathrm{ICAL,G}}} + \frac{f^B_{sol}}{t_{\mathrm{ICAL,B}}} \right) \times 48 N_{\mathrm{it}} N^2_{\mathrm{a}}$
Subtract $C_{\mathrm{Subtract}} = 16 \times ((1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}}) \times R_{\mathrm{vis,full\,FoV}}$
Correct $C_{correct} = 224 \times (1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} \times R_{\mathrm{vis,full\,FoV}}$
Weight visibilities $C_{\mathrm{Weight}} = 16(1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} R_{\mathrm{vis,full\,FoV}}$
Phase rotation $C_{\mathrm{phrot,total}} = 2 \times (1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N^2_{\mathrm{facet}} \times 28 R_{\mathrm{vis,full\,FoV}}$
Coalesce $C_{\mathrm{Coalesce}} = 8(1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N^2_{\mathrm{facet}} R_{\mathrm{vis,full\,FoV}}$
Grid $C_{\mathrm{grid}} = (1 + N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N^2_{\mathrm{facet}} N_{\mathrm{mm}} \times \sum_{\mathrm{i,j}}^{\mathrm{baselines}} R_{\mathrm{vis,facet\,FoV}}(B_{\mathrm{i,j}}) 8 (N_{\mathrm{pix,kernel}}(B_{\mathrm{i,j}}))^2$
De-grid $C_{\mathrm{de-grid}} = C_{grid}$
Gridding kernel update $C_{\mathrm{Kernels}} = \frac{(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N^2_{\mathrm{facet}} N_{\mathrm{pol}} N_{\mathrm{beam}}}{t_{\mathrm{kernel,update}}} (\sum_{\mathrm{i,j}}^{\mathrm{baselines}} 5 N^2_{\mathrm{cvff}} \log_2(N^2_{\mathrm{cvff}}) N_{\mathrm{f,kernel}})$
FFT $C_{\mathrm{FFT}} = \left(\frac{(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N^2_{\mathrm{facet}} N_{\mathrm{pol}} N_{\mathrm{beam}} (N_{\mathrm{subbands}} N_{\mathrm{Tt}})}{t_{\mathrm{snap}}} \right) 2.5 N^2_{\mathrm{pix,facet}} \log_2(N^2_{\mathrm{pix,facet}})$
IFFT $C_{\mathrm{IFFT}} = C_{\mathrm{FFT}}$
Re-project $C_{\mathrm{Reproj}} = \frac{(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} r^2_{\mathrm{facet}} N_{\mathrm{pol}} N_{\mathrm{beam}} (N_{\mathrm{subbands}} N_{\mathrm{Tt}})}{t_{\mathrm{snap}}} 50 N^2_{\mathrm{pix}}$
Image spectral fitting $C_{\mathrm{spectralfit}} = 2 \frac{(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N_{\mathrm{pol}} N_{\mathrm{beam}} (N_{\mathrm{subbands}} N_{\mathrm{Tt}}) N^2_{\mathrm{pix}}}{t_{\mathrm{obs}}}$
Identify Component $C_{\mathrm{Identifycomponent}} = \frac{2(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N_{\mathrm{minor}} N_{\mathrm{beam}} N_{\mathrm{pol}} N_{\mathrm{subbands}} N_{\mathrm{Tt}} N^2_{\mathrm{pix}}}{t_{\mathrm{obs}}}$
Subtract Image component $C_{\mathrm{Subtractimagecomponent}} = \frac{2(1+N_{\mathrm{SelfCal}}) N_{\mathrm{major}} N_{\mathrm{minor}} N_{\mathrm{pol}} N_{\mathrm{beam}} (N_{\mathrm{subbands}} N_{\mathrm{Tt}} N^2_{\mathrm{patch,pix}})}{t_{\mathrm{obs}}}$
Source Find Insignificant

- We continue to develop a parametric model for the SDP compute.
- From this we estimate the required compute rate for each pipeline.



ICAL

UNIVERSITY OF CAMBRIDGE

SDP will be cost constrained

- There is always more processing one *could* do on the data
- Limited budget puts constraints on both the hardware built and the use made of it (power)
- SKAO have provided example power caps to consider

Power caps: 4MW, 10MW

SDP will be cost constrained

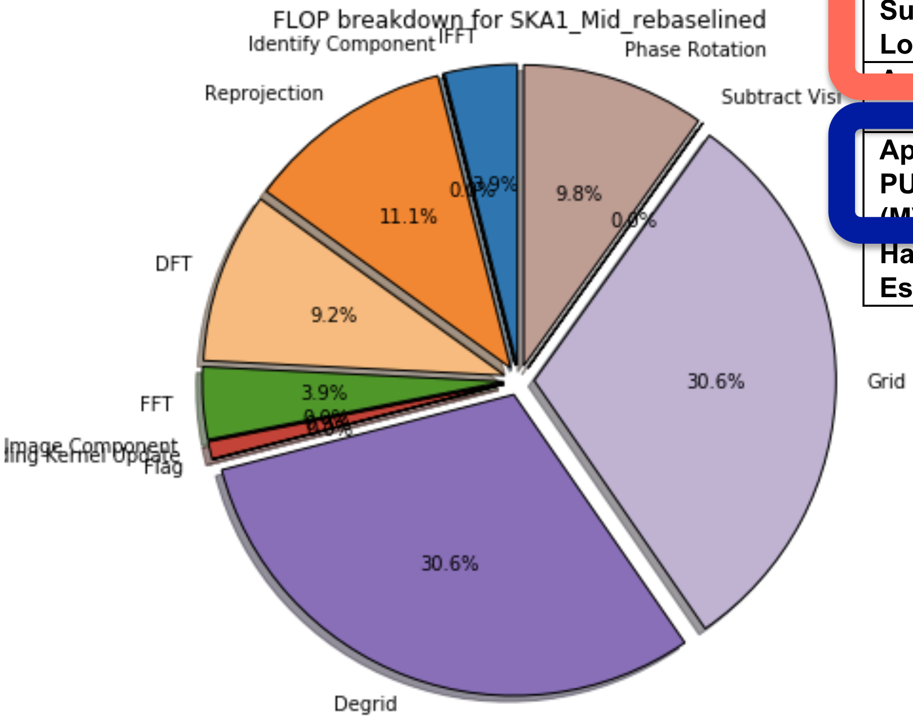
SQUARE KILOMETRE ARRAY

SDP

SCIENCE DATA PROCESSOR

SKA1 SDPs are ~200-300 PFLOPS systems (25% efficiency)

	LOW (50-350MHz)	MID Band 1 (350-1050 MHz)	MID Band 2 (950-3050 MHz)	MID Band 5 (4.6 to 9.6 GHz)
DD CAL (not in iPython)	18.3	17.4	17.4	17.4
ICAL	4.9	9.5	7.5	6.3
DPrep A+B	4.8	10.8	9.2	6.8
DPrep C	12.0	30.4	23.0	17.4
[illegible]	[illegible]	[illegible]	3.0	2.5
Sustained Compute Load Total (PFLOPS)	41.5 PFLOPS	72.1 PFLOPS	60.2 PFLOPS	50.5 PFLOPS
[illegible]	[illegible]	[illegible]	5.8	4.9
Apparent power, with PUE and power factor (MVA)**	5.8	9.9	3.3	6.9
Hardware CAPEX Estimate (M€)***	57	110	92	77



Need 6MVA for LOW, 10VA for MID
c.f.
4MVA for LOW, 10MVA for MID

UNIVERSITY OF CAMBRIDGE

SDP will be a one way street

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

- Power caps are at the right sort of level to enable processing, according to our model

However, there's no "slack" here

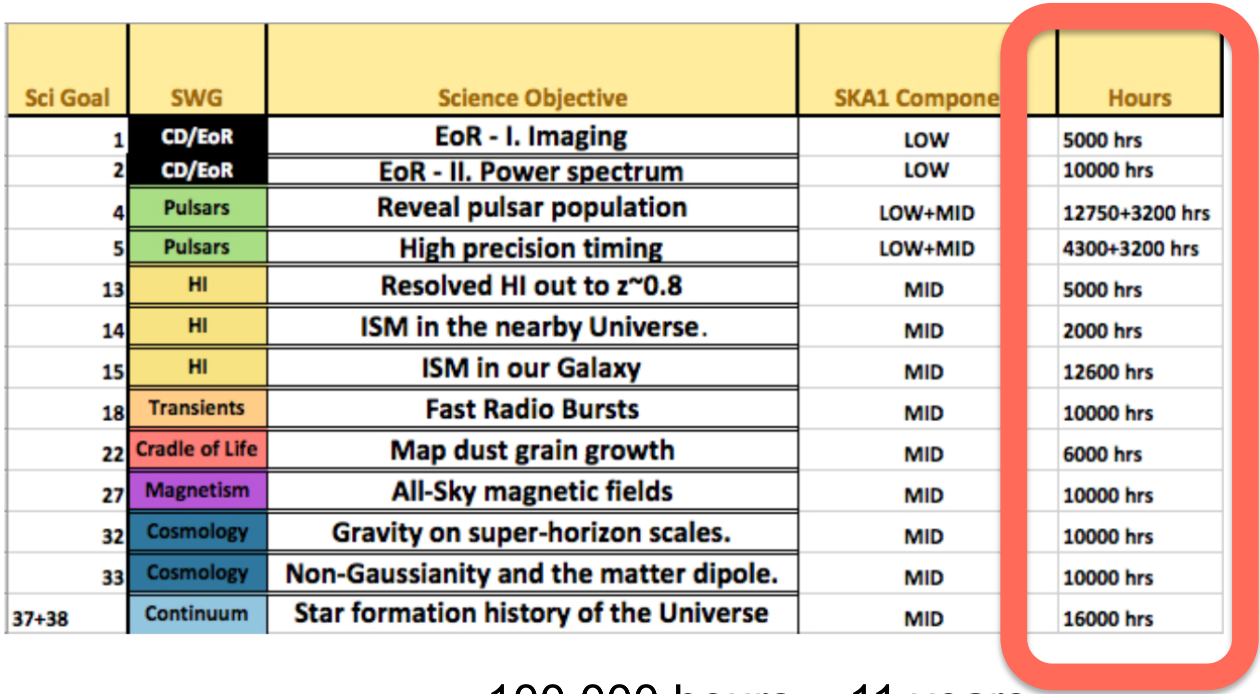
OK?

- No re-processing of data
- Each 6 hour observation is passed through the pipelines only once
- Each 6 hour observation has data products pushed through to the archive about 12 hour after the observation started.

UNIVERSITY OF CAMBRIDGE

High Priority Science Objectives

- SKAO has developed a list of HPSO experiments – programmes targeting specific scientific goals and taking long periods (~5000-16000 hours) of telescope time.
- Draft schedule for these taking 5-15 years to complete
- Just an example – *NOT* to be taken as "real" time allocation



Sci Goal	SWG	Science Objective	SKA1 Compone	Hours
1	CD/EoR	EoR - I. Imaging	LOW	5000 hrs
2	CD/EoR	EoR - II. Power spectrum	LOW	10000 hrs
4	Pulsars	Reveal pulsar population	LOW+MID	12750+3200 hrs
5	Pulsars	High precision timing	LOW+MID	4300+3200 hrs
13	HI	Resolved HI out to z~0.8	MID	5000 hrs
14	HI	ISM in the nearby Universe.	MID	2000 hrs
15	HI	ISM in our Galaxy	MID	12600 hrs
18	Transients	Fast Radio Bursts	MID	10000 hrs
22	Cradle of Life	Map dust grain growth	MID	6000 hrs
27	Magnetism	All-Sky magnetic fields	MID	10000 hrs
32	Cosmology	Gravity on super-horizon scales.	MID	10000 hrs
33	Cosmology	Non-Gaussianity and the matter dipole.	MID	10000 hrs
37+38	Continuum	Star formation history of the Universe	MID	16000 hrs

100,000 hours = 11 years

We can use these to generate example SDP use cases and archive growth rates.
Also could enable load balancing if we relax latency requirement of off-line processing.

High Priority Science Objectives

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

HPSO	Total (PFLOPS)	Hours of telescope time	Fraction of time
U.HPSO-4a Pulsar Search MID SPF1	~0	800	0.01
U.HPSO-4b Pulsar Search MID SPF2	~0	2400	0.04
U.HPSO-5a Pulsar Timing MID SPF2	~0	1600	0.02
U.HPSO-5b Pulsar Timing MID SPF3	~0	1600	0.02
U.HPSO-13 Hi Kinematics and Morphology	25.6	5000	0.07
U.HPSO-14 Hi MID	32.7	2000	0.03
U.HPSO-15 Studies of the ISM in our Galaxy	26.2	12600	0.19
U.HPSO-18 Transients MID	~0	10000	0.15
U.HPSO-22 Cradle of Life MID Band 5	25.4	6000	0.09
U.HPSO-27 All Sky Magnetism	26.3	10000	0.15
U.HPSO-37a Continuum Survey MID band 2	28.1	2000	0.03
U.HPSO-37b Continuum Survey MID band 2 (deep)	28.1	2000	0.03
U.HPSO-37c Continuum survey, band 2 wide	28.1	10000	0.15
U.HPSO-38a Continuum Survey MID band 5	26.1	1000	0.01
U.HPSO-38b Continuum Survey MID band 5	26.1	1000	0.01
Weighted average FLOPS value for MID HPSOs			20 PFLOPS
Approximate AVERAGE Apparent power requirement[2]			~2.7 MVA

Average FLOPS values 3.5x lower for MID than maximal case.

Could still build system for maximal case but not power it up all the time

or

Relax latency requirement and save both capital cost and power cost

UNIVERSITY OF CAMBRIDGE

Data rates: be careful what you wish for!

SKA
SQUARE KILOMETRE ARRAY
SDP
SCIENCE DATA PROCESSOR

- SKA images ~30k pixels on a side for MID, ~8k for LOW
  - (Low images have 20x fewer pixels than MID from simple B_{max}/D ratio)
- Each instrument is capable of 65k channels, 4 stokes parameters in output images
- Auxiliary products give 7x increase in data size c.f. single image cube
- 4 Bytes per pixel (32 bit)
- Data creation rate scales as number of pixels (2D), number of channels and inverse of observation length
- If observations are too short (or too many channels), data rate going out is larger than visibility data rate coming in!

So much for "data reduction"

UNIVERSITY OF CAMBRIDGE

HPSO analysis of archive growth rate

MID:

Weighted average archive growth rate for all experiments is

9 Gbits/s

Much lower than the once per 6 hours maximal rate of

3 Tbits/s* (or 30x 100Gbits/s)

25% of time is “allocated” to Non-imaging experiments

Many imaging experiments are continuum experiments

The spectral line experiments (so far) do not need all 65k channels

*Yes, the observant of you will have calculated that the incoming data rate for both SKA instruments is around 40G visibilities per second, or ~3.8 Tbits/s.

But an ultra deep blind spectral line (Hi) survey at full spatial resolution is on the list of " things which might kill us "

HPSO analysis of archive growth rate

LOW:
~50% of time “allocated” to EoR experiment.
Rate of generating image data is relatively modest (only ~200 channels) – **5 Gbits/s**
~50% of time “allocated” to NIP experiments
Relative to the imaging experiments, archive is zero

HPSO average image archive growth rate 3 Gbits/s
Full resolution cube every 6 hours would give ~200 Gbits/s (with auxiliary products)

HPSO analysis of archive growth rate

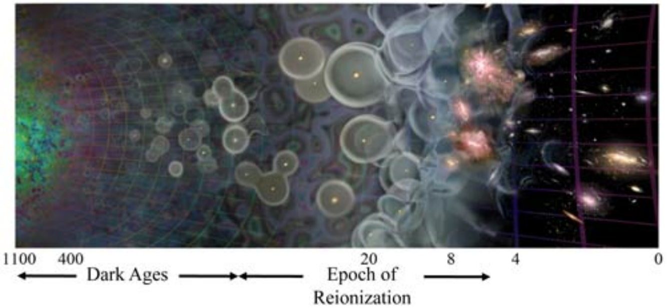
LOW:

EoR Science is expected to need access to ungridded visibility data.

Even if we average this (post-calibration) in frequency (<10kHz) and time (<10s) we estimate that each of the 3 EoR experiments will generate data rate of averaged visibilities is 200Gbits/s.

There are 3 EoR experiments, each 5000 hours – total for all three is 1.4 Exabytes!

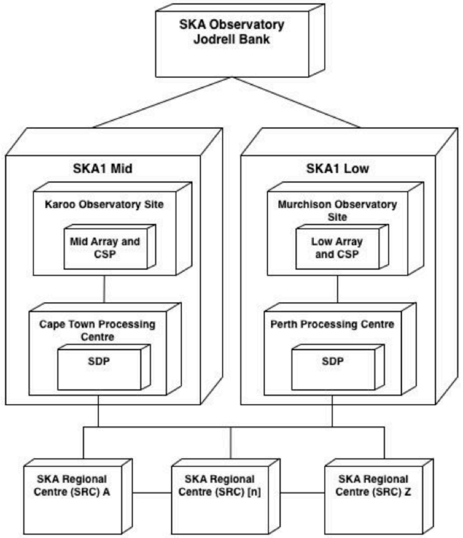
SKAO estimate only 11% (1000 hrs pa) of time likely "good" enough for EoR… (so combined EoR experiments take 15 years)



She might look cute, but she's going to eat a hell of a lot of peanuts!

Summary

- SKA Scope: Is limited!
- SKA will deliver only from a subset of standard data products.
- Paradigm shift: The SDP compute is a resource, in an identical way to the “on-sky” time.
- No option for reprocessing data in current design.
- But the board has adopted regional centres
- Collaborations with UK science teams are needed to foster better understanding of requirements of SDP and SKA regional centres.


SKA Observatory Jodrell Bank
SKA1 Mid
Karoo Observatory Site
Mid Array and CSP
Cape Town Processing Centre
SDP
SKA1 Low
Murchison Observatory Site
Low Array and CSP
Perth Processing Centre
SDP
SKA Regional Centre (SRC) A
SKA Regional Centre (SRC) [n]
SKA Regional Centre (SRC) Z

Don't be afraid to talk about the elephants!

END