BIG TELESCOPE, BIG DATA: TOWARDS EXA-SCALE WITH THE SKA

Anna Scaife
Jodrell Bank Centre for Astrophysics
Advanced European Network of E-infrastructures for Astronomy with the SKA (AENEAS - 731016)

CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE

Royal Society 8-9 April 2019 Numerical algorithms for high-performance computational science
Advanced European Network of E-infrastructures for Astronomy with the SKA

AENEAS - 731016

11/04/17

CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE

Royal Society 8-9 April 2019 Numerical algorithms for high-performance computational science
Future SKA Science Archive

- Google: 98PB
- Facebook: 180PB
- CERN: 73PB
- YouTube: 15PB
- LOFAR Long Term Archive: 25PB

SKA Phase 1 Science Archive: 300PB

Per year: 1 Petabyte
Shorter baseline
= smaller Fourier frequency
= larger image scale

Longer baseline
= larger Fourier frequency
= smaller image scale

\[ V_{ij} = \langle E(\vec{r}_i, t) E^*(\vec{r}_j, t) \rangle_t \]
Fourier Transforms

Astronomy

Interferometers

Anna Scaife University of Manchester

Introduction to Radio Imaging

October 16, 2017

Royal Society
8-9 April 2019
Numerical algorithms for high-performance computational science
\[ V(u, v, w) = \int \frac{I(l, m)}{\sqrt{1 - l^2 - m^2}} e^{-2\pi i(ul + vm + w(\sqrt{1 - l^2 - m^2} - 1))} \]

\[ I_{\text{meas}}(l, m) = \iint S(u, v)V(u, v)e^{2\pi i(ul + vm)} \, du \, dv \]

This sampling function identifies the values of \((u, v)\) that we sample according to our baseline distribution.

\[ S(u, v) = \sum_{i=1}^{M} \delta(u - u_i, v - v_i) \]

Where \(M\) is the number of different visibilities that we have:

\[ M = N_{\text{ant}}(N_{\text{ant}} - 1) / 2 \times N_\tau \times N_f \]
\[ V(u, v, w) = \int \frac{I(l, m)}{\sqrt{1 - l^2 - m^2}} e^{-2\pi i (ul + vm + w(\sqrt{1 - l^2 - m^2} - 1))} \]
\[ V_{\text{grid}}(u_k, v_k) = \left[ [V(u, v) \cdot S(u, v)] \ast C_{aa}(u, v) \right] \cdot III(u_k, v_k) \]

The gridded visibility data on a grid with $\sqrt{k} \times \sqrt{k}$ pixels.

The input visibility data, sampled at a number of times and frequencies.

The convolution/gridding kernel function.

Sample onto regularly spaced grid using the Shah function.
\[ V_{\text{grid}}(u_k, v_k) = \left[ [V(u, v) \cdot S(u, v) \cdot W(u, v)] \ast C_{aa}(u, v) \right] \cdot III(u_k, v_k) \]

To make an image we can now simply FFT our UV grid, but we must also correct for the gridding function that we have introduced and normalise the weights:

\[ I_{\text{meas, dirty}}(l, m) = \frac{\text{FT}^{-1}[V_{\text{grid}}(u, v)]}{\left( \sum W_{\text{grid}}(u, v) \right) \text{FT}^{-1}[C_{aa,\text{grid}}(u, v)]} \]

The image that we have made is known as the DIRTY IMAGE, because we have not made any correction for the weighted sampling \( S(u, v)W(u, v) \).
Because we are multiplying our continuous visibilities by $S(u,v)W(u,v)$ the DIRTY IMAGE shows us the convolution of their Fourier transforms.

$$V(u,v) \cdot [S(u,v) \cdot W(u,v)] \Leftrightarrow I(l,m) * b_{PSF}(l,m)$$

Where the point spread function, or synthesized beam, or dirty beam, is defined as

$$b_{PSF}(l,m) = FT^{-1}[S(u,v) \cdot W(u,v)]$$

It would be nice if we could just divide out this multiplication directly in Fourier space, but we can’t because it has zero-valued components.
Challenge 1: Undersampling
Challenge 2: Spectral behaviour

This is because astrophysical radio sources have frequency dependent behaviour. We can exploit this by taking a Taylor expansion of the frequency dependence.

Gridding separate frequencies together results in image plane aberrations.
Challenge 3: Non-coplanarity

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

We think of images as being 2D, but in reality our antennas are distributed in 3 dimensions and the sky is curved.

This is known as the \textbf{w-effect}.

It introduces a direction dependent \textbf{phase} that is different for each pair of antennas.

Effectively, each antenna pair sees a different sky.
Challenge 3: Non-coplanarity
Challenge 4: Direction-dependent effects
SDP will deliver **standard data products**

For imaging observations these are image data products

A standard SKA1-MID image data product has **30k x 30k pixels**

SKA1 will have up to **65k frequency channels** and **4 polarisations**

At 4 Bytes per voxel that equates to

\[30k \times 30k \times 65k \times 4 \times 4 = 936 \text{ TeraBytes}\]

- even for a snapshot image
The five stages of learning about SKA data products:

(1) **Denial**
   
   “What do you mean I can’t have the visibilities?”

(2) **Anger**
   
   “That’s crazy! I need the visibilities!”

(3) **Bargaining**
   
   “What if I help with commissioning? Can I have the visibilities then?”

(4) **Depression**
   
   “It’s never going to work if I can’t have the visibilities.”

(5) **Acceptance**
   
   “...”