The Square Kilometre Array Computing Challenge

Tim Cornwell, Head of Computing, Square Kilometre Array
Great Observatorys for the coming decades

E-ELT optical/IR

Construction approved
Great Observatories for the coming decades

Atacama Large Millimetre Array (ALMA): mm/submm
Chajnantor Plateau @ 17,000 ft
Early science now
Inaugurated on 13th March 2013
Great Observatories for the coming decades

James Webb Space Telescope: due for launch in 2018
Great Observatories for the coming decades

Square Kilometre Array: radio
Construction start 2017/18
SKA Phase 1 (SKA1)
Cost: €650M, construction start 2017

SKA1_MID
254 Dishes including:
64 x MeerKAT dishes
190 x SKA dishes

SKA1_LOW
Low Frequency Aperture Array Stations
96 Dishes including:
36 x ASKAP
60 x SKA dishes

Southern Africa

Australia
SKA Phase 2 (SKA2)
Cost: TBD; construction start 2022

Southern Africa
- SKA2_MID
  - 2500 Dishes
  - Mid Frequency Aperture Array Stations

Australia
- SKA2_AA
  - Mid Frequency Aperture Array Stations
- SKA2_LOW
  - Low Frequency Aperture Array Stations
Design process

- Design of SKA to be undertaken in global consortia, which act as contractors to the central office.
- SKA Office will run system engineering, receive and review designs from consortia, monitor progress, analyse Earned Value.
- SKA Office issued a baseline conceptual design to serve as starting point for design, based on previous work and CoDRs.
- 10 consortia formed to undertake the design.
- SKA Office holds the design authority for the project.
Log-Periodic Array: ~250,000 antennas
Offset-Optics Antenna Design

- Multiple Feeds to cover multiple bands
- 15-m Effective Diameter
- 13.5-m Diameter

- MeerKAT dish design

Measuring the Universe with the world’s largest radio telescope
SKA is driving development of new science & technical solutions

• Dishes, feeds, receivers (N=2500)
• Low and mid aperture arrays (N=250, n>1,000,000)

2013: global internet traffic ~ 100 ExaByte/month (Exa is 10^{18})

SKA Phase 1 will be ~1.3 ZettaByte/month (Zetta is 10^{21})

So, SKA Phase 1 will require dates rates ~13 times current global internet traffic

SKA Phase 2 will have dates rates 10 times larger.

This is in South Africa only. Australian data rates will be larger!!

**INDUSTRY ENGAGEMENT IS CENTRAL TO THE SKA**
What Happens in an Internet Minute?

- 639,800 GB of global IP data transferred
- 20 New victims of identity theft
- 204 million Emails sent
- 47,000 App downloads
- 1300 New mobile users
- 61,141 Hours of music
- 20 million Photo views
- 320+ New Twitter accounts
- 61,141 App downloads
- $83,000 In sales
- 320+ New Twitter accounts
- 100,000 New tweets
- 277,000 Logins
- 6 million Facebook views
- 2+ million Search queries
- 1.3 million Video views
- 30 Hours of video uploaded
- 639,800 GB of global IP data transferred

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2 times the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second

Exploring the Universe with the world’s largest radio telescope
Science Data Processing

- Lead by U. Cambridge (Paul Alexander)
  - Core team: ASTRON (NL), CSIRO (AU), MeerKAT (SA), Cambridge (UK), ICRAR (AU), CHPC (SA), Hartree Centre (UK)
- Responsible for processing of visibility data into images
- Strong industry engagement
  - IBM, INTEL, NVIDIA, etc.
- Heavy compute load
  - SKA1: 30 to 100PF
- Massive data rate
  - SKA1: 0.3 to 3 TB/s

![Element Concept Diagram]

The level 1 Data Flow Diagrams are shown below
Science flow

Telescope Manager

Meta Data

Science Data Processor Local M&C

Master Controller

Local M&C Database

Science Data Processor

Data Routing → Ingest → Data Routing

Visibility processing → Image Plane Processing

Sky Models, Calibration Parameters...

Data Buffer

Multiple Reads

Data Products

Time Series Search → Time Series Processing

Tiered Data Delivery

Exploring the Universe with the world's largest radio telescope
Australian component
Exploring the Universe with the world’s largest radio telescope
Data delivery to RSECs

Elements to note in this data flow include:
- Data routing is considered a key aspect of the overall data flow and is shown explicitly. By appropriate data routing, the design aim is to make the data embarrassingly parallel where this is achievable.
- The initial ingest phase will provide data editing and flagging, possible phase rotation, assimilation of the metadata into the SDP data structures and averaging thereby providing experiment selection depending on the averaging performed.
- The Data Buffer is explicitly included to perform two functions. The first is to enable iterative algorithms to be included in the workflow. This requires a write once read N buffer where N is the number of iterations. The second function to be performed by the buffer is load-balancing. It gives the ability to load-balance the processing system by scheduling a long low data-rate, low computational-cost experiment following a high data-rate, high computational-cost experiment and buffering the former until the processor has completed the second operation. The data buffer is shown after a data routing step so that the data are optimally organised for the processing step.
- Two main pipelines are shown—visibility and time-series processing. Other pipelines are variants on these (transient detection via fast imaging) or not shown explicitly (direct analysis of visibility data for EoR power spectrum measurements).
- Three databases are shown, one for sky-models, one for calibration parameters etc., and an archive of final data products and a database for local monitor and control information.
- A master Controller is connected to all components of the system as well as providing hardware and HPC monitoring and control functionality. The Master Controller provides a single interface to the Telescope Manager.
- The Master Controller also provides various expert-level views into the system and manages overall control.
Regional Science and Engineering Centres (RSECs)

- National/international science and engineering centres
- Similar to CERN Tier1
- May specialize in science areas e.g. Epoch of Reionization
- Local archive of science data products
- Post-processing of science data products
  - e.g. multi-EB cube
- Funded separately from SKA

- Board has yet to fully define RSECs
Challenges

Data rate

Software

Systems

Algorithms

Compute

Architecture

7.6.5.4.3.2.1

ST

ST

ST

ST

SKA

SKA

SKA

SKA

Definition

Level

Responsibility

Candidate of ExaScale Architecture

Four types of architectures are considered

General Purpose (GP)

Ordinary CPU-based MPPs

e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters

Capacity-Bandwidth oriented (CB)

With expensive memory-I/F rather than computing capability

e.g.) Vector machines

Reduced Memory (RM)

With embedded (main) memory

e.g.) SoC, MD-GRAPE4, Anton

Compute Oriented (CO)

Many processing units

e.g.) ClearSpeed, GRAPE-DR

Fresnel diffraction

A

B

A'
## Data flow

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADC outputs</td>
<td>SKA</td>
</tr>
<tr>
<td>2</td>
<td>Beam-former output</td>
<td>SKA</td>
</tr>
<tr>
<td>3</td>
<td>Correlator output</td>
<td>SKA</td>
</tr>
<tr>
<td>4</td>
<td>Visibility data</td>
<td>SKA</td>
</tr>
<tr>
<td>5</td>
<td>Calibrated data, images and catalogues</td>
<td>SKA</td>
</tr>
<tr>
<td>6</td>
<td>Validated science data products (released by Science Teams)</td>
<td>ST</td>
</tr>
<tr>
<td>7</td>
<td>Enhanced data products e.g. Source identification and association</td>
<td>ST</td>
</tr>
</tbody>
</table>

- **0.3 to 3 TB/s**
  - Dish arrays
  - Low frequency aperture array

- **10 - 500 TB/s**
  - ~ 100 PB data set read multiple times over several days
  - e.g. 1 year Redshifted Hydrogen survey ~ 4EB

- **Enhanced data products e.g. Source identification and association**

- **Validated science data products (released by Science Teams)**

- **Calibrated data, images and catalogues**

- **Visibility data**

- **Correlator output**

- **Beam-former output**

- **ADC outputs**

- **STK**

- **SKA**
Challenges

Data rate

Software Systems

Compute Architecture

Algorithms

Four types of architectures are considered:

1. **General Purpose (GP)**
   - Ordinary CPU-based MPPs
   - e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters

2. **Capacity-Bandwidth oriented (CB)**
   - With expensive memory-I/F rather than computing capability
   - e.g.) Vector machines

3. **Reduced Memory (RM)**
   - With embedded (main) memory
   - e.g.) SoC, MD-GRAPE4, Anton

4. **Compute Oriented (CO)**
   - Many processing units
   - e.g.) ClearSpeed, GRAPE-DR

Fresnel diffractive fraction

A

B

A'

v

w

IESP Meeting@Kobe (April 12, 2012)
Compute architecture

- Kondo et al requirements analysis

![Diagram with axes labeled as Requirement of Memory Capacity (PB) and Requirement of B/F, showing different capacity and BW requirements with open square markers and arrows indicating high and low BW for small, middle, and high capacity requirements.](attachment:image.png)
Three axes of computing

- **General Purpose (GP)**: Ordinary CPU-based MPPs, e.g., K-Computer, GPU, Blue Gene, x86-based PC-clusters.
- **Capacity-Bandwidth oriented (CB)**: With expensive memory-I/F rather than computing capability, e.g., Vector machines.
- **Reduced Memory (RM)**: With embedded (main) memory, e.g., SoC, MD-GRAPE4, Anton.
- **Compute Oriented (CO)**: Many processing units, e.g., ClearSpeed, GRAPE-DR.

IESP Meeting@Kobe (April 12, 2012)
Kondo et al projection

Performance Projection

- Performance projection for an HPC system in 2018
  - Achieved through continuous technology development
  - Constraints: 20 – 30MW electricity & 2000sqm space

<table>
<thead>
<tr>
<th>Node Performance</th>
<th>Total CPU Performance (PetaFLOPS)</th>
<th>Total Memory Bandwidth (PetaByte/s)</th>
<th>Total Memory Capacity (PetaByte)</th>
<th>Byte / Flop</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose</td>
<td>200~400</td>
<td>20~40</td>
<td>20~40</td>
<td>0.1</td>
</tr>
<tr>
<td>Capacity-BW Oriented</td>
<td>50~100</td>
<td>50~100</td>
<td>50~100</td>
<td>1.0</td>
</tr>
<tr>
<td>Reduced Memory</td>
<td>500~1000</td>
<td>250~500</td>
<td>0.1~0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Compute Oriented</td>
<td>1000~2000</td>
<td>5~10</td>
<td>5~10</td>
<td>0.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>Injection</th>
<th>P-to-P</th>
<th>Bisection</th>
<th>Min Latency</th>
<th>Max Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-radix (Dragonfly)</td>
<td>32 GB/s</td>
<td>32 GB/s</td>
<td>2.0 PB/s</td>
<td>200 ns</td>
<td>1000 ns</td>
</tr>
<tr>
<td>Low-radix (4D Torus)</td>
<td>128 GB/s</td>
<td>16 GB/s</td>
<td>0.13 PB/s</td>
<td>100 ns</td>
<td>5000 ns</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Storage</th>
<th>Total Capacity</th>
<th>Total Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 EB</td>
<td>10TB/s</td>
</tr>
<tr>
<td></td>
<td>100 times larger than main memory</td>
<td>For saving all data in memory to disks within 1000-sec.</td>
</tr>
</tbody>
</table>
Technology readiness

Projected vs. Required Perf.

Gap between requirements and technology trends

Requirements (PfLOPS)

CP | RM | GP | CB

Exploring the Universe with the world’s largest radio telescope
Challenges

Data rate
Compute Architecture
Software Systems
Algorithms

Four types of architectures are considered:
- General Purpose (GP)
  - Ordinary CPU-based MPPs
    - e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters
- Capacity-Bandwidth oriented (CB)
  - With expensive memory-I/F rather than computing capability
    - e.g.) Vector machines
- Reduced Memory (RM)
  - With embedded (main) memory
    - e.g.) SoC, MD-GRAPE4, Anton
- Compute Oriented (CO)
  - Many processing units
    - e.g.) ClearSpeed, GRAPE-DR

Fresnel data fraction
A
B
A'
v
w

Exploring the Universe with the world’s largest radio telescope
Energy required to move a word

Access vs Reach

Source: Energy at ExaFlops, Peter M. Kogge, SC09 Exa Panel
Challenges

Data rate

Compute Architecture

Software Systems

Algorithms

Four types of architectures are considered:

- General Purpose (GP)
  - Ordinary CPU-based MPPs
  - e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters

- Capacity-Bandwidth oriented (CB)
  - With expensive memory-I/F rather than computing capability
  - e.g.) Vector machines

- Reduced Memory (RM)
  - With embedded (main) memory
  - e.g.) SoC, MD-GRAPE4, Anton

- Compute Oriented (CO)
  - Many processing units
  - e.g.) ClearSpeed, GRAPE-DR
Key challenge is scaling to 100PFlop
- ASKAPsoft scales to about 100TFlops
- For scaling need:
  - Flexible but high performance software framework
  - Tool support
  - Access to top end processing systems
  - Data simulator
- Algorithm people, HPC people, data people
Scaling ASKAP Processing to ~10,000 cores, SKA1 must scale to ~10,000,000 cores
Astroinformatics challenges

• Use cases and data products are very similar to ASKAP, MeerKAT, and LOFAR
• Solve those and you’ve solved SKA1
• Except for scale
Astroinformatics challenges

• SKA1_Survey data cube very large
  – ~ 100 - 1000 size of ASKAP cubes
  – Massive visualisation challenge
• Reanalysis is time consuming
  – Entire sky: 10,000s @ 10TB/s
  – One pointing: 100s @ 10TB/s
• Download from archive is prohibitive
• Remote execution (VOSpace) will be vital
Timeline

- Request for Proposals
- Proposal responses
- Proposal evaluation
- Cost ceiling established
- Design consortia start
- Preliminary Design Review
- Prototype systems deployed
- Critical Design Review
- Seek SKA1 funding
- Develop SKAO governance
- SKA1 construction approved
- Tender & procure construction
- SKA1 construction
- Detailed design of SKA2
- SKA1 early science

3 years