ICRAR

International
Centre for
Radio
Astronomy
Research

Scheduling, deploying and monitoring 100 million tasks

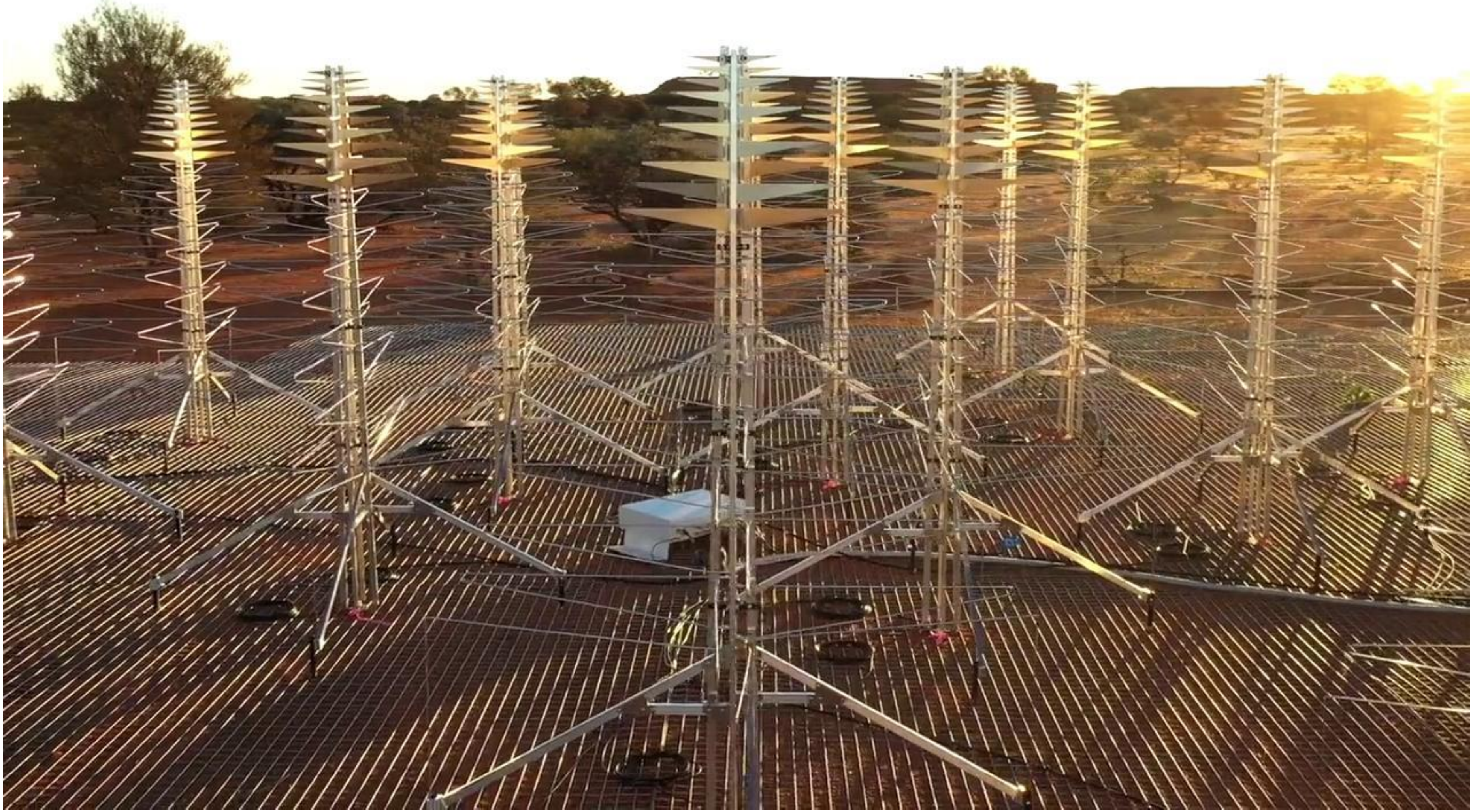
Andreas Wicenec

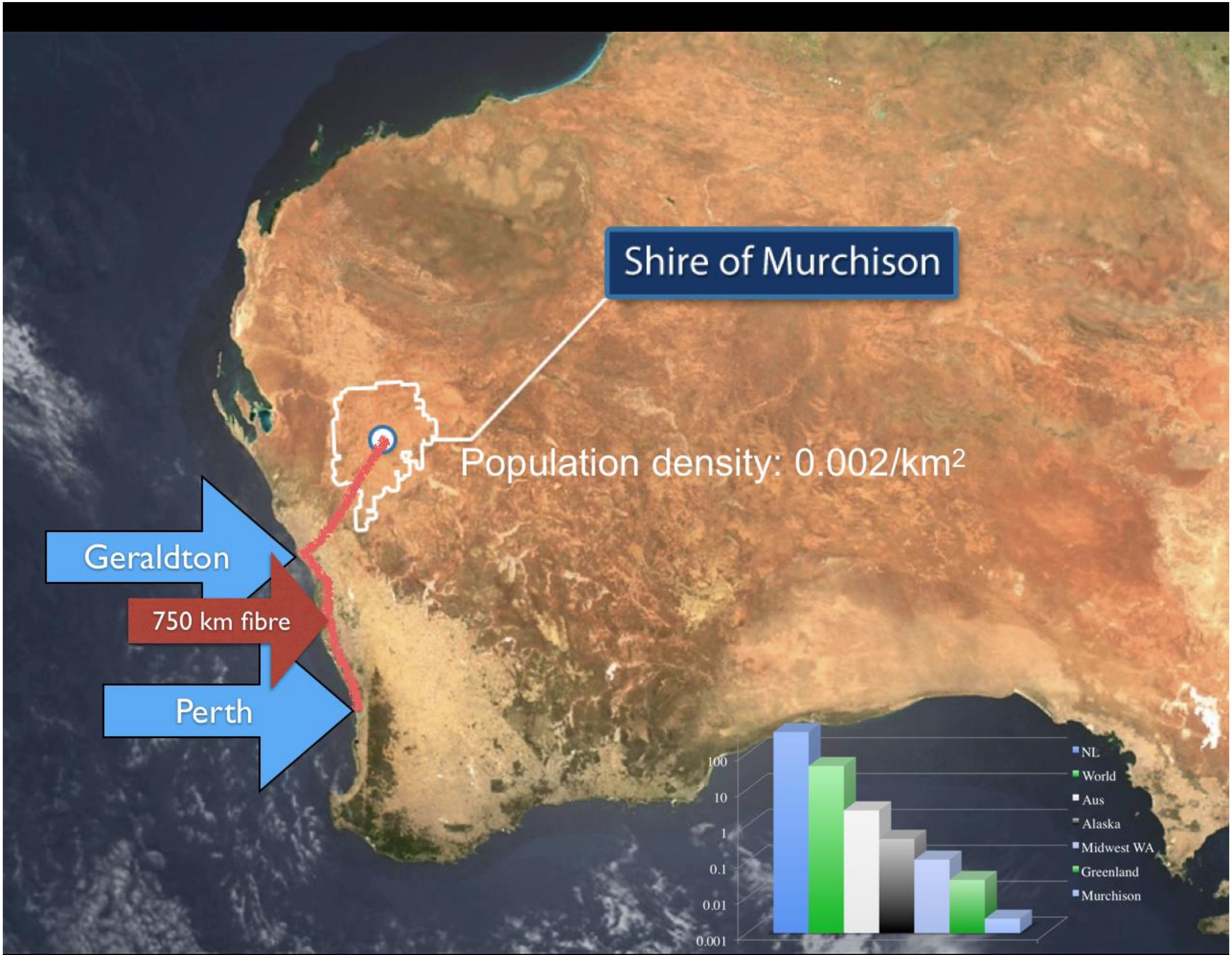
& The ICRAR team

Curtin University

THE UNIVERSITY OF WESTERN AUSTRALIA

Government of Western Australia
Department of the Premier and Cabinet
Office of Science





Shire of Murchison

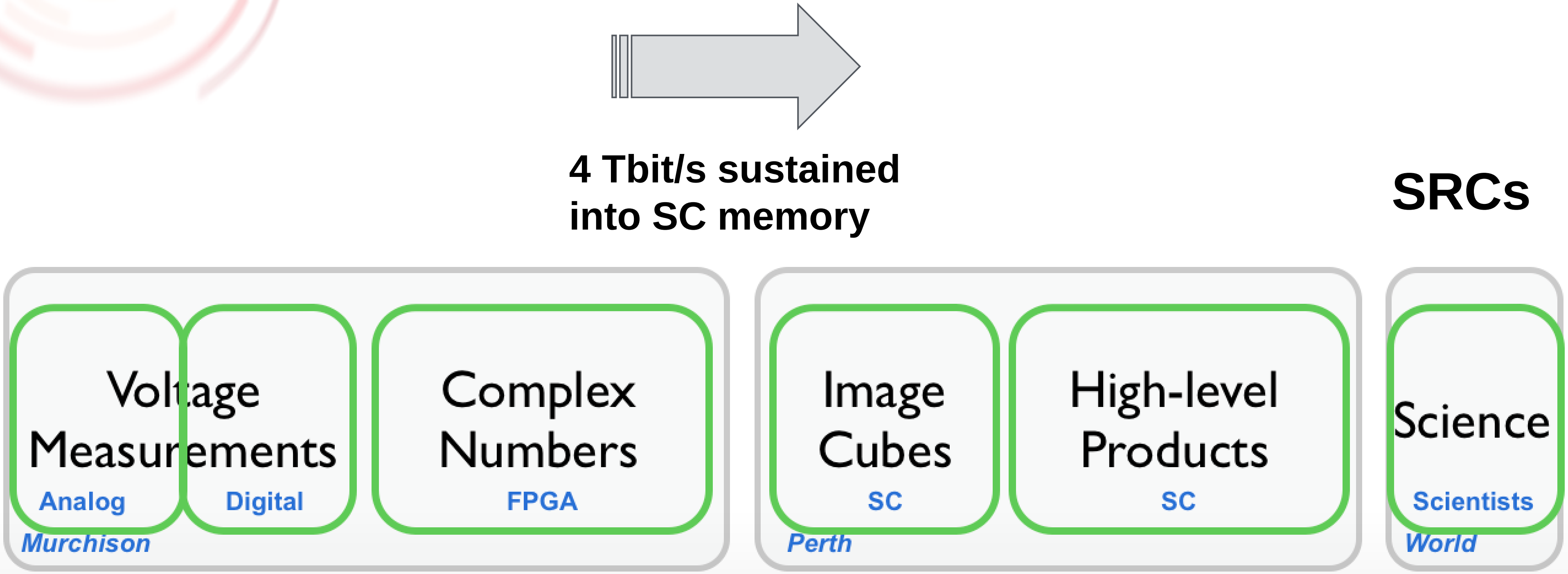
Population density: 0.002/km²

Geraldton

750 km fibre

Perth

SKA Data Flow



4 Tbit/s sustained into SC memory

SRCs

Voltage Measurements

Analog

Digital

Complex Numbers

FPGA

Murchison

Image Cubes

SC

High-level Products

SC

Perth

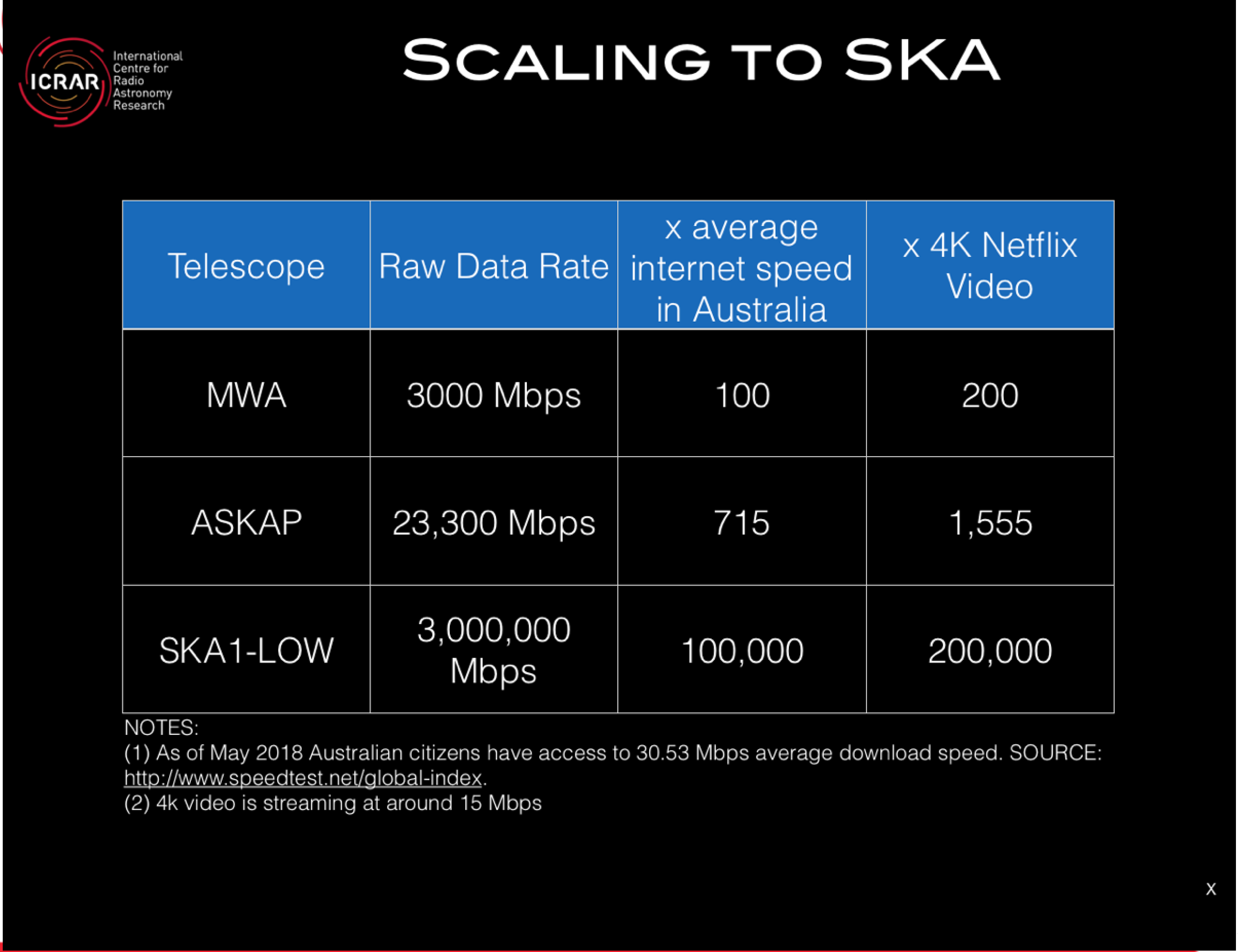
Science

Scientists

World

RT calibration solutions

24/7 operations



Scaling to SKA

Telescope	Raw Data Rate	x average internet speed in Australia	x 4K Netflix Video
MWA	3000 Mbps	100	200
ASKAP	23,300 Mbps	715	1,555
SKA1-LOW	3,000,000 Mbps	100,000	200,000

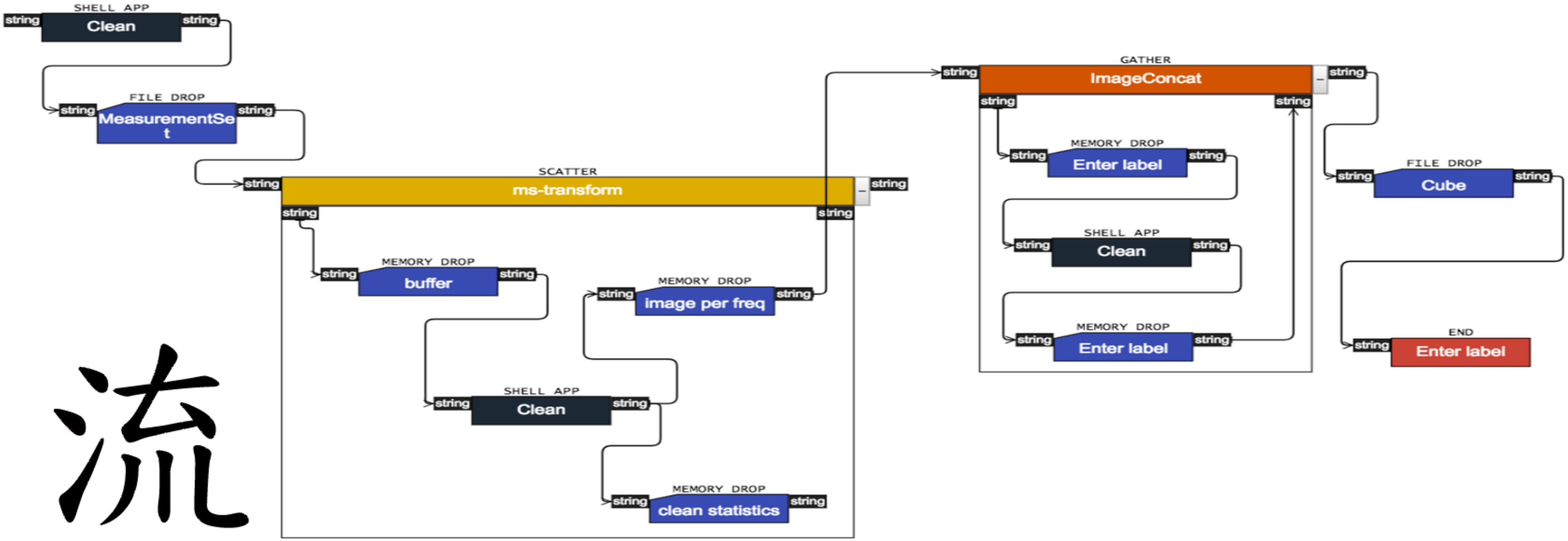
NOTES:
(1) As of May 2018 Australian citizens have access to 30.53 Mbps average download speed. SOURCE: http://www.speedtest.net/global-index.
(2) 4k video is streaming at around 15 Mbps

x

The
DALiuGE System

A scalable, graph oriented workflow development, scheduling and execution system.

Execution System: DALiuGE



流

- In-house development: **Data Activated Flow Graph Engine**.
- Manage tasks and data locality and execution monitoring.
- Unique feature: Data is represented by active tasks during execution.
- ⇒Scalability verified up to tens of millions of tasks.
- ⇒Applications and Data are tasks collectively called ‘Drops’.
- DALiuGE is a graph based system, using graph optimisation and scheduling.

DALiuGE: Motivation 流

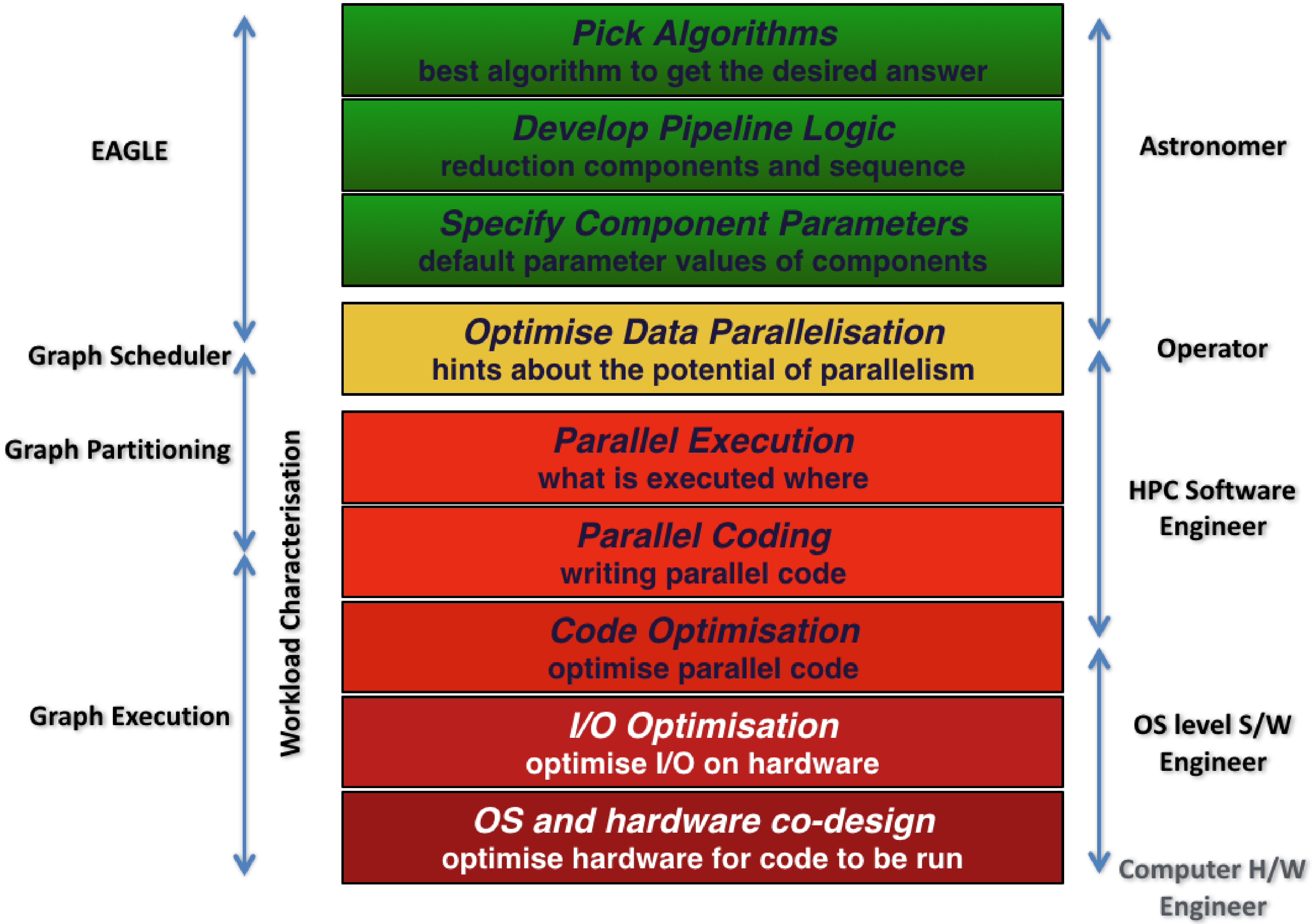
Lower boundary
for maximum
number of tasks

$$2^{16} * 300 * 4 = 78,643,200$$

Channels Facets Polarisations

..and that's a really rough estimate

Separation of Concerns



Pick Algorithms
best algorithm to get the desired answer

Develop Pipeline Logic
reduction components and sequence

Specify Component Parameters
default parameter values of components

Optimise Data Parallelisation
hints about the potential of parallelism

Parallel Execution
what is executed where

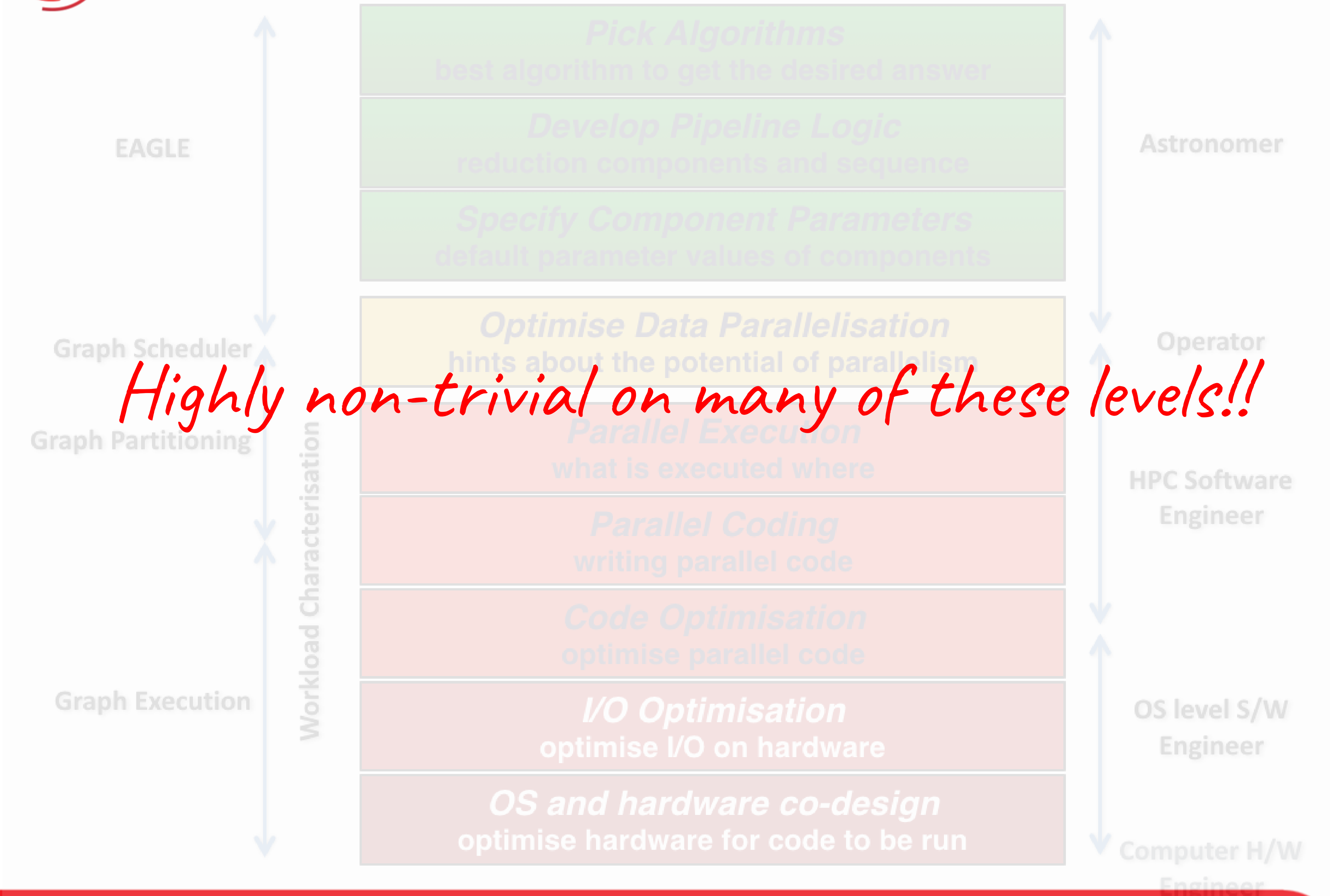
Parallel Coding
writing parallel code

Code Optimisation
optimise parallel code

I/O Optimisation
optimise I/O on hardware

OS and hardware co-design
optimise hardware for code to be run

DALiuGE



EAGLE

Graph Scheduler

Graph Partitioning

Graph Execution

Workload Characterisation

Pick Algorithms
best algorithm to get the desired answer

Develop Pipeline Logic
reduction components and sequence

Specify Component Parameters
default parameter values of components

Optimise Data Parallelisation
hints about the potential of parallelism

Parallel Execution
what is executed where

Parallel Coding
writing parallel code

Code Optimisation
optimise parallel code

I/O Optimisation
optimise I/O on hardware

OS and hardware co-design
optimise hardware for code to be run

Astronomer

Operator

HPC Software Engineer

OS level S/W Engineer

Computer H/W Engineer

Highly non-trivial on many of these levels!!

DALiuGE 流

'Scalabale' for DALiuGE design means thinking about whether a feature could scale and then... asking again whether it really scales.

Result is a share nothing and distributed design and implementation *almost* everywhere!

The Challenges of Simulating the SKA1-LOW:

➢512 stations

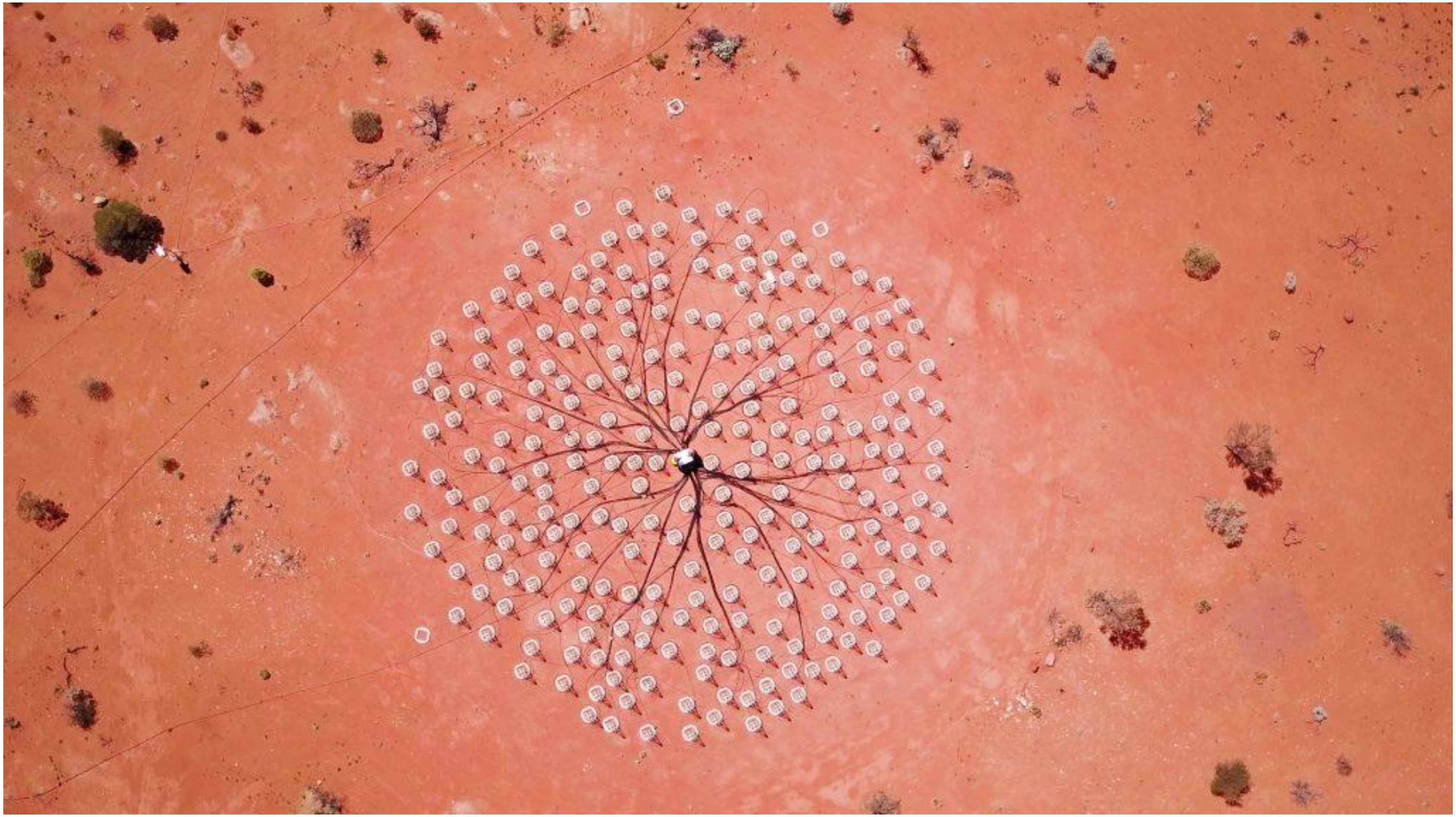
➢131,072 antennas

➢2^{16} channels

➢Complex antenna and station beam patterns

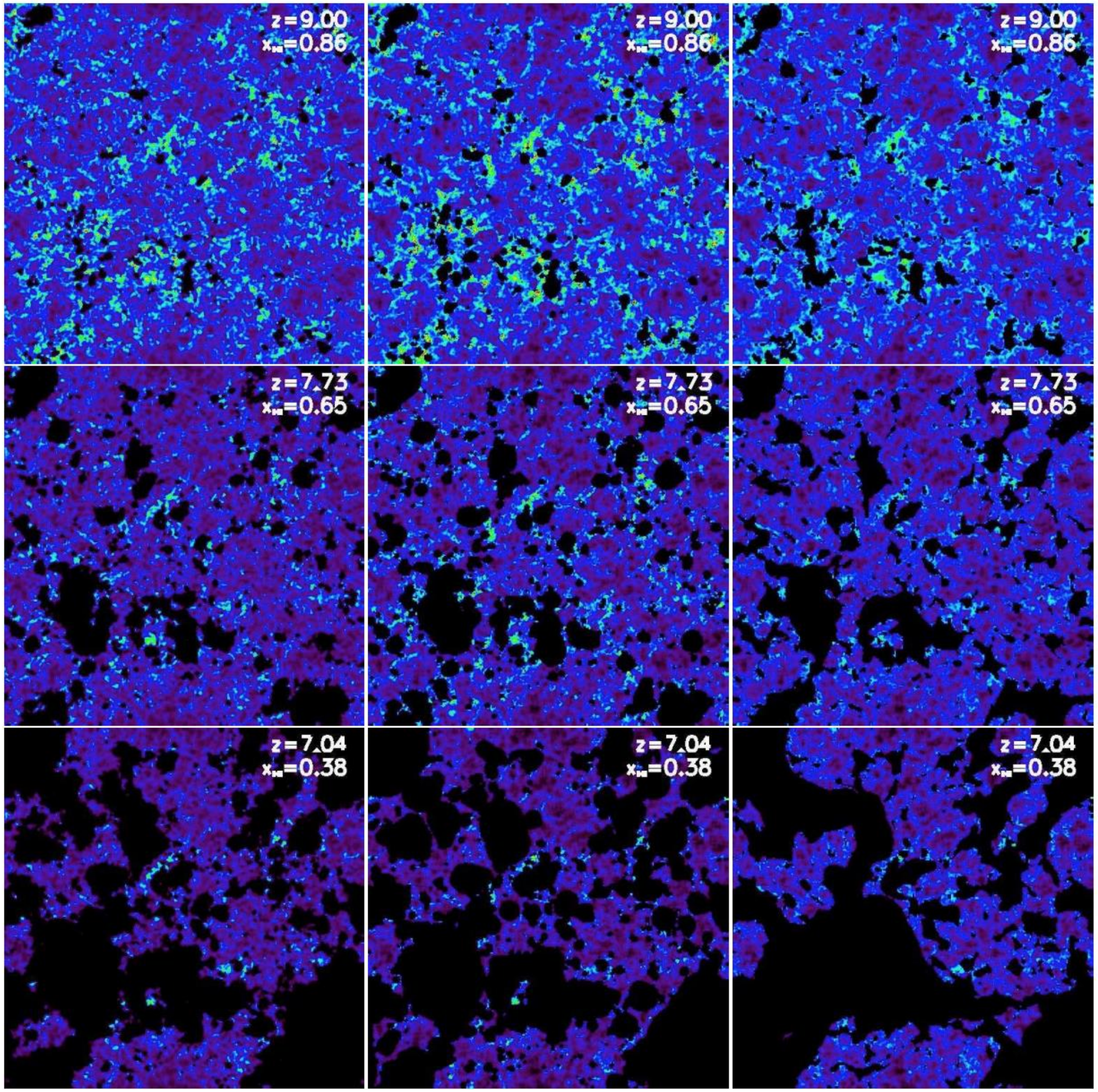
➢Complex sky model

Simulating 131,072 Antennas

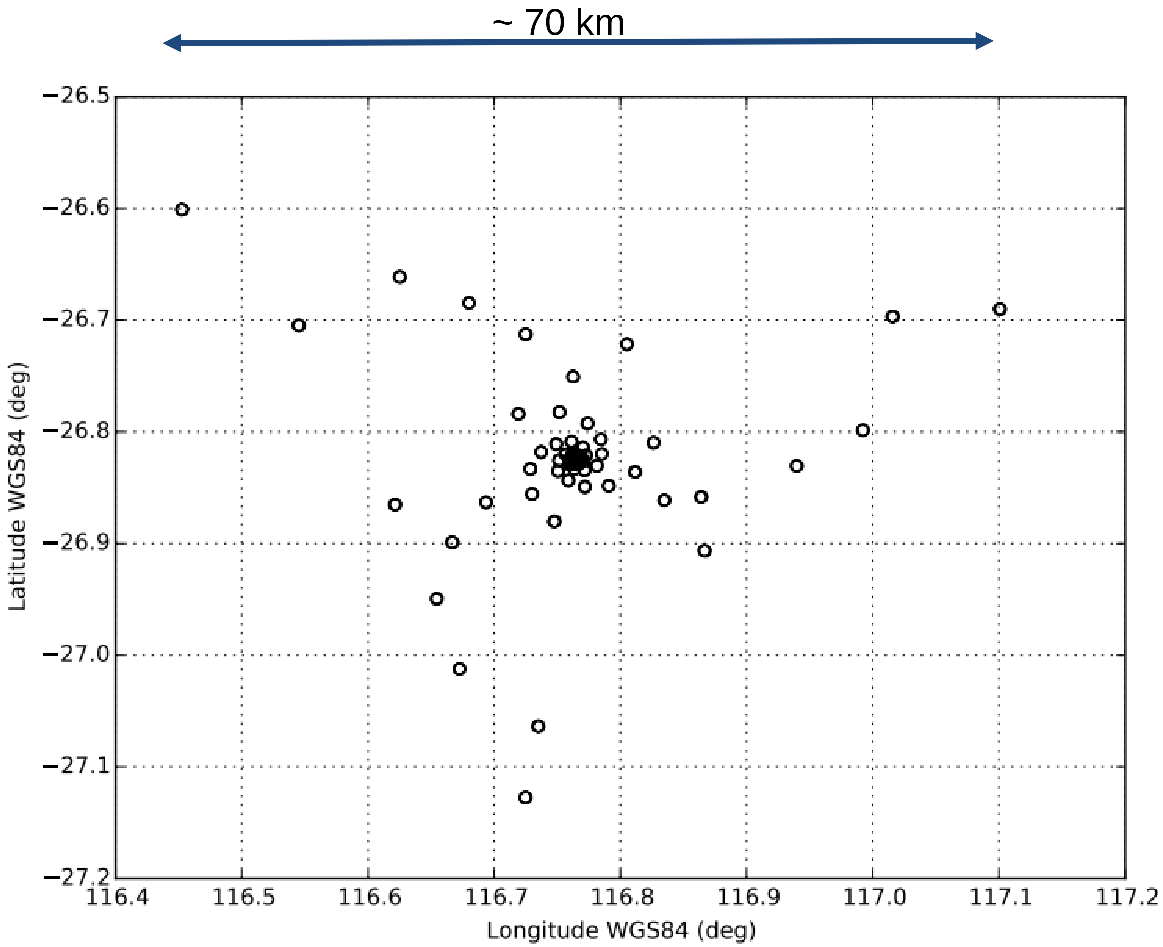


Input: Model of Theoretical Signal

A. Mesinger, S. FurlanettoandR. Cen, Mon. Not. Roy. Astron. Soc. **411**(2011)955*doi* : 10.1111/*j*.1365 – 2966.2010.17731.*x*[*arXiv* : 1003.3878[*astro* – *ph*.*CO*]].

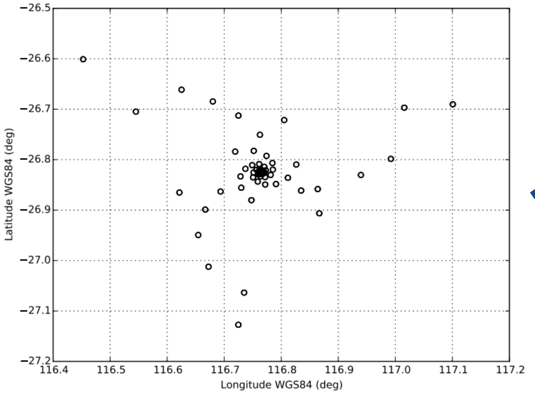


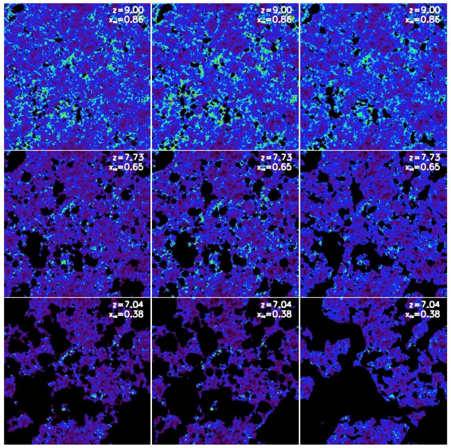
Input: Layout and Response of Antenna Array

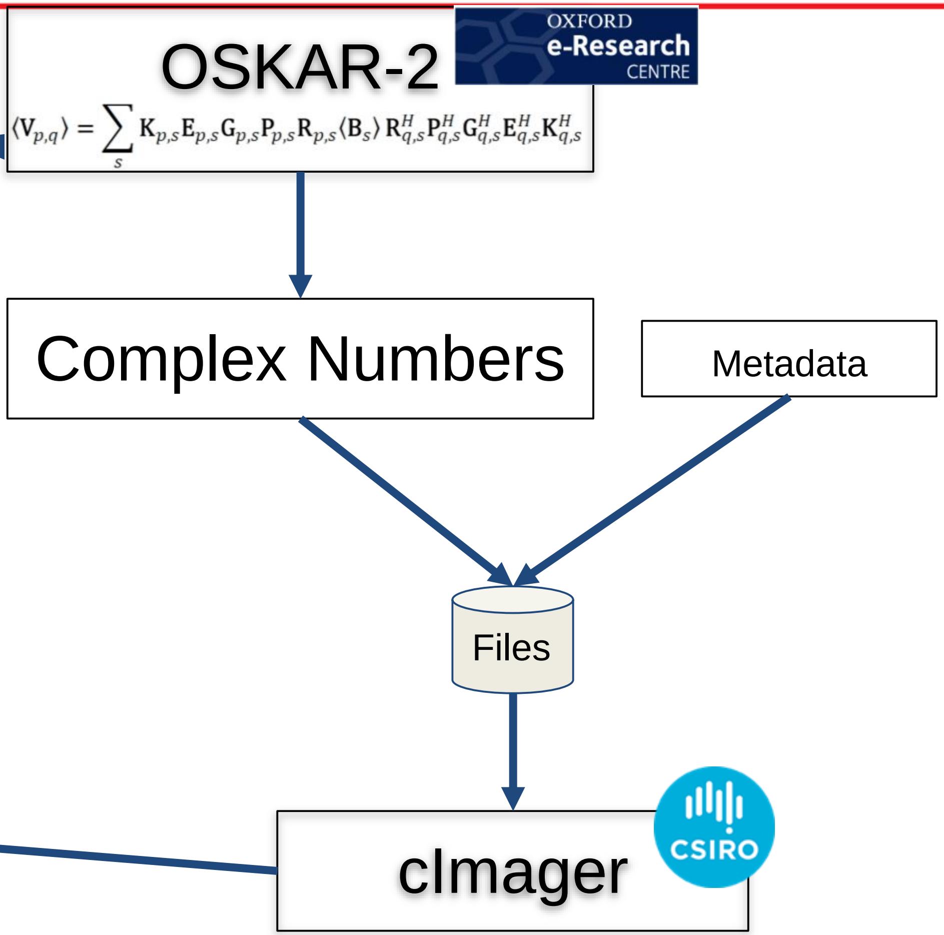


512 antenna stations, 256 antennas/station

Simulating 131,072 Antennas







OSKAR-2

$$\langle V_{p,q}\rangle = \sum_s K_{p,s}E_{p,s}G_{p,s}P_{p,s}R_{p,s}\langle B_s\rangle R^H_{q,s}P^H_{q,s}G^H_{q,s}E^H_{q,s}K^H_{q,s}$$

Complex Numbers

Metadata

Files

cImager

Latest News

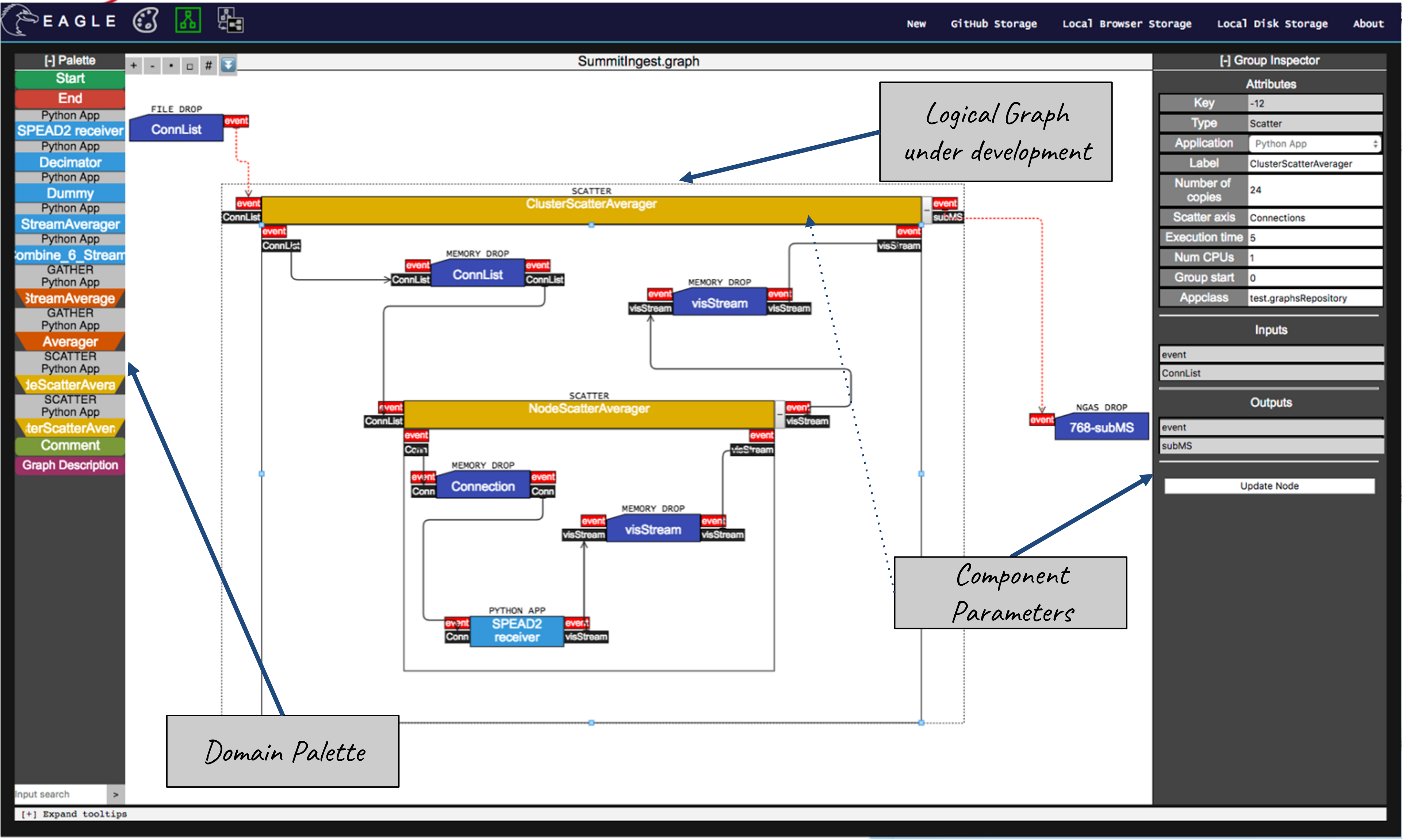
We had been running a SKA1-LOW simulation on SUMMIT using

4,561 nodes
27,360 V100 GPUs

essentially the whole machine available at the time of launching the job...

Workflow

Development



[-] Group Inspector

Attributes

Key	-12
Type	Scatter
Application	Python App
Label	ClusterScatterAverager
Number of copies	24
Scatter axis	Connections
Execution time	5
Num CPUs	1
Group start	0
Appclass	test.graphsRepository

Inputs

event
ConnList

Outputs

event
subMS

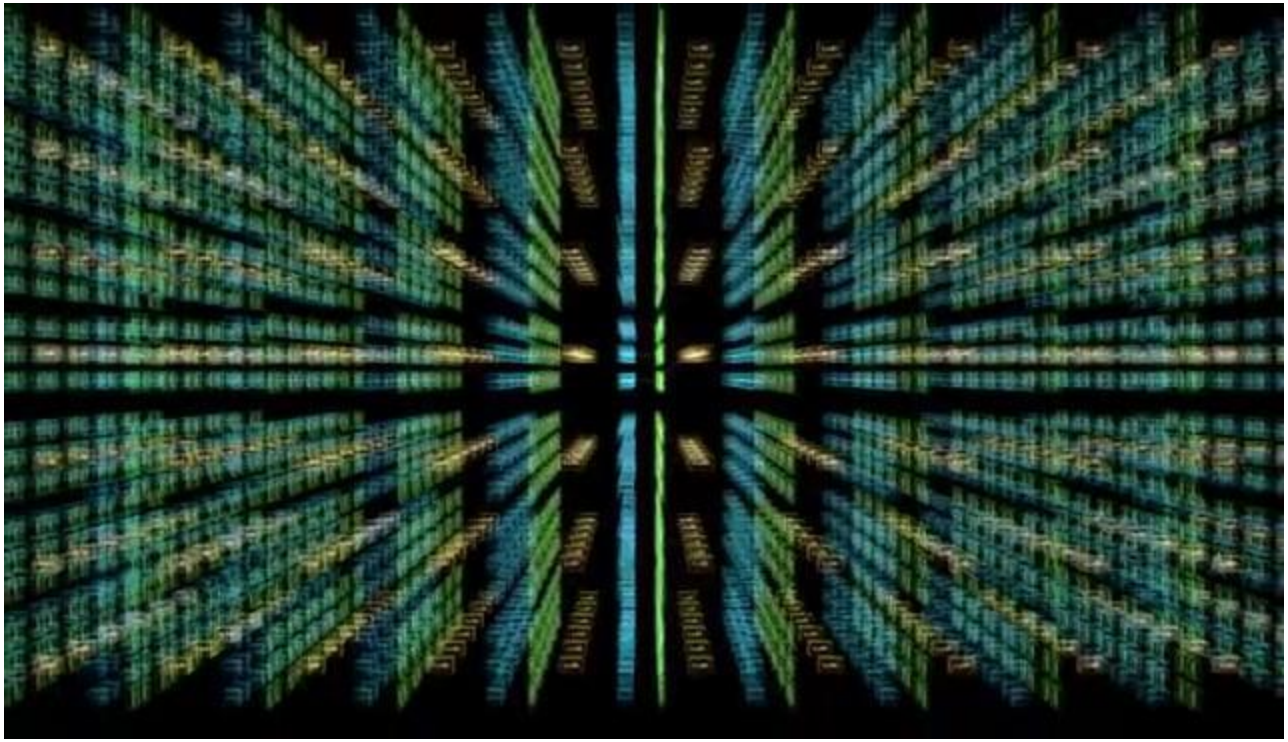
Update Node

Logical Graph under development

Component Parameters

Domain Palette

SKA Simulation Run



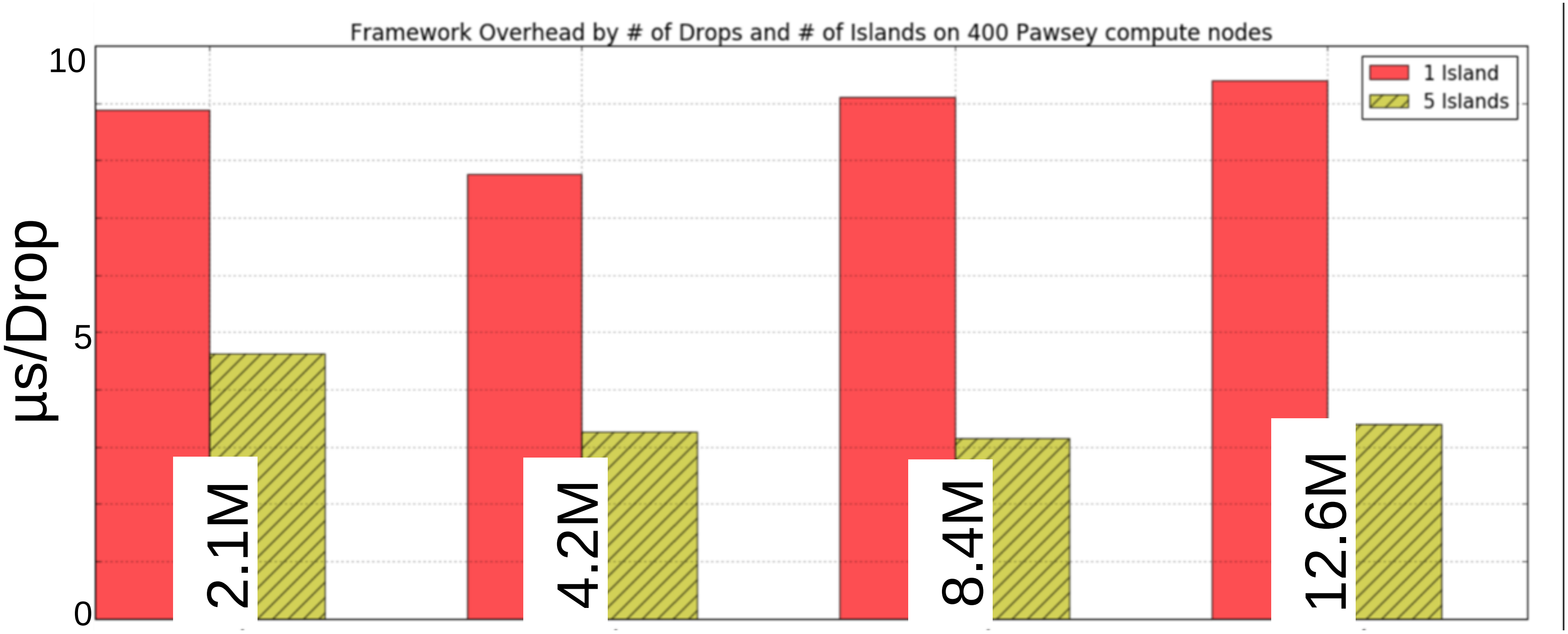
Video credit: Baoqiang Lao, Shanghai Astronomical Observatory

Key Figures:

- ➢3 hours run-time
- ➢2x faster than actual SKA
- ➢27,360 channels
- ➢$3.3*10^{10}$ visibilites/s
- ➢2.6 PB total data flow
- ➢output 112 TB in 760 filesets
- ➢11 GB/s sustained write rate
- ➢$5*10^{14}$ total visibilities written
- ➢Final image cube: ~3.1 GB!!

Does it Scale? 流

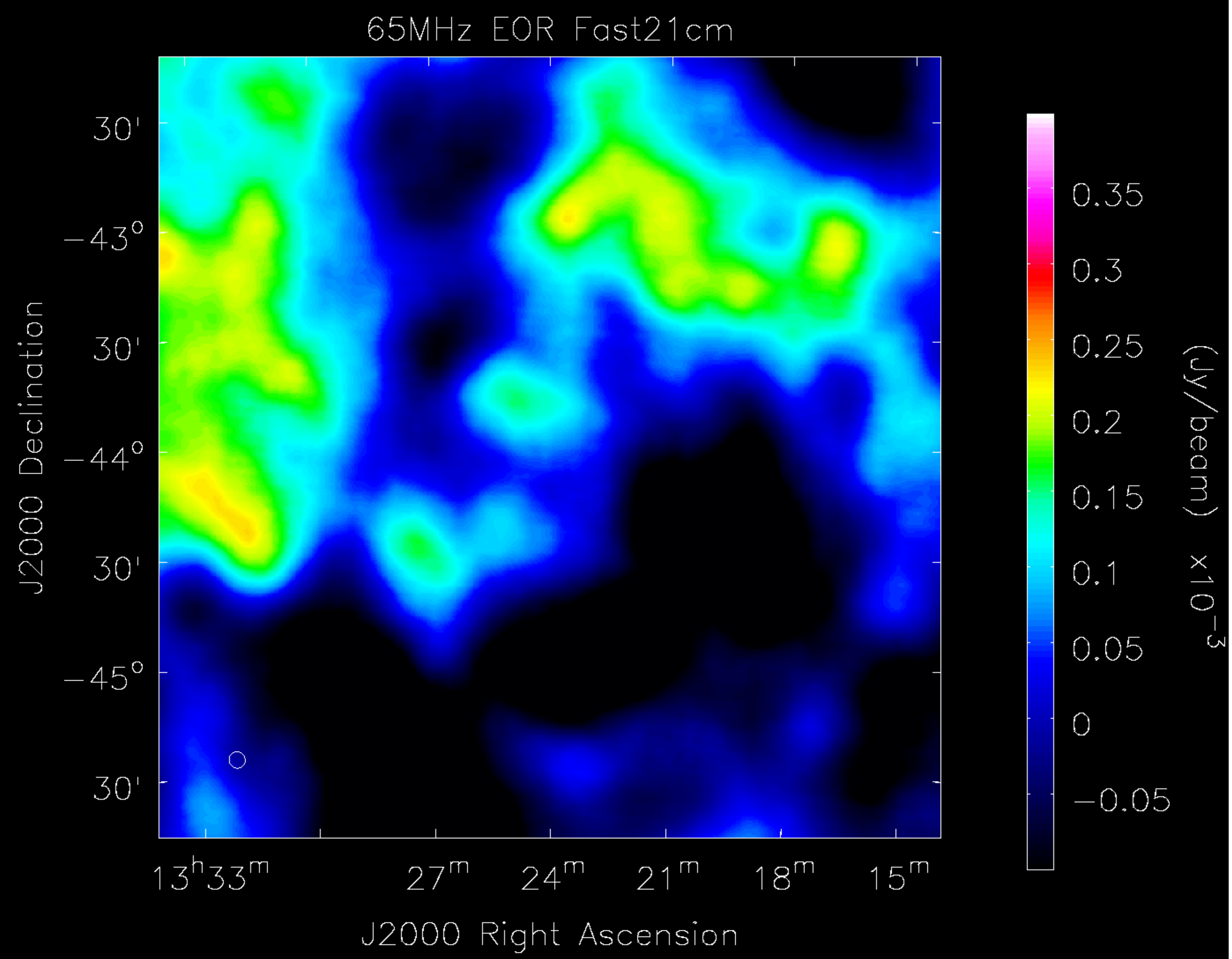
- Complex workflow simulations on Tianhe-2.
- Measured plain overhead imposed by DALiuGE during execution.



Does it Scale? 流

- Yes!

- Test run: 12.6 Million tasks on 400 compute nodes means 31,500 tasks/node.
- Current expectation for SDP 2,500 nodes running 78M tasks. With actual numbers: 31,450 tasks/node.
- Test execution time 420 seconds.
- SDP: several hours.


65MHz EOR Fast21cm
J2000 Declination
30'
−43°
30'
−44°
30'
−45°
30'
J2000 Right Ascension
13h33m
27m
24m
21m
18m
15m
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
(Jy/beam) x10−3

DALiuGE 流

The

Data Activated Flow Graph Engine

Open source, under GitHub and PyPi

https://github.com/ICRAR/daliuge

(https://github.com/ICRAR/EAGLE)