International Centre for Radio Astronomy Research

Data Intensive Astronomy Primer

Andreas Wicenec

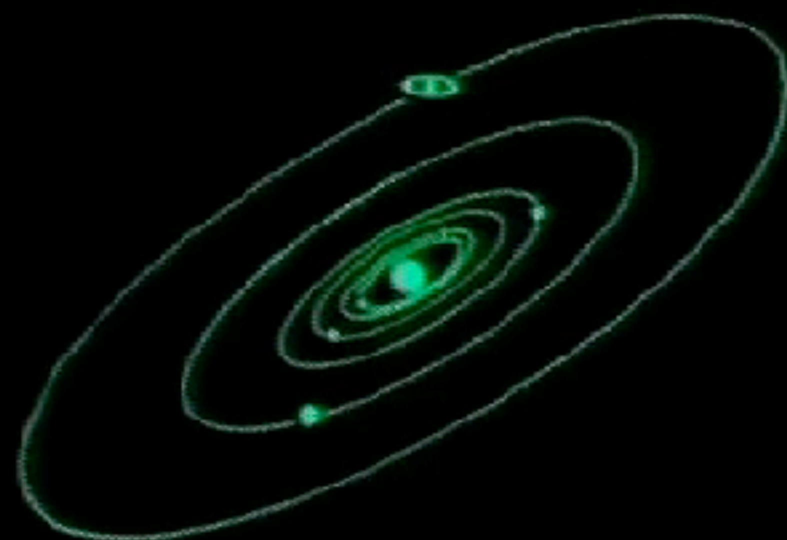
SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

Curtin University

The University of Western Australia
Achieve International Excellence

Why?

INTRODUCTION



SPACE IS BIG. REALLY BIG. YOU JUST WON'T BELIEVE HOW VASTLY, HUGELY, MIND-BOGGLINGLY BIG IT IS. I MEAN,

Video courtesy of BBC

ICRAR International Centre for Radio Astronomy Research

The Hitchhikers Guide to the Galaxy

Douglas Adams was absolutely correct.

Space is really big.

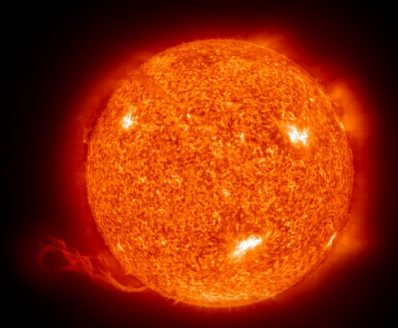
Douglas Adams
1952 - 2001

What’s out there?

Adapted from Quinn

Adapted from Quinn

Radius



Planet	Star	Galaxy
1000 km 10^3 km	10^6 km	10^{17} km
10^{-3}	1	10^{+11}

Distance

10^8 km	10^{13} km	10^{17} km	10^{19} km
10^{+2}	10^{+7}	10^{+11}	10^{+13}

Adapted from Quinn

Are we in a new era of astronomy?

- Amount and complexity of data has to be seen relative to the effort it takes to gather and process it.
- Compared to previous times it is obvious that gathering data nowadays is almost trivial. Too easy??
- Hipparchus and Tycho Brahe spent significant periods of their entire life to observe some 1,000 stars. Their achievements and conclusions are phenomenal! Probably only a few of us today would be able to derive similar knowledge from such data.

Are we in a new era of astronomy?

Tycho Brahe’s Quadrant

The instruments and methods invented and used by ancient and medieval astronomers had always been unique and cutting-edge. Very often stunningly big and complex.

The size of the instruments was always just necessary but never sufficient to derive the knowledge. In addition it always required the brightest minds.

Ancient and Medieval Instruments

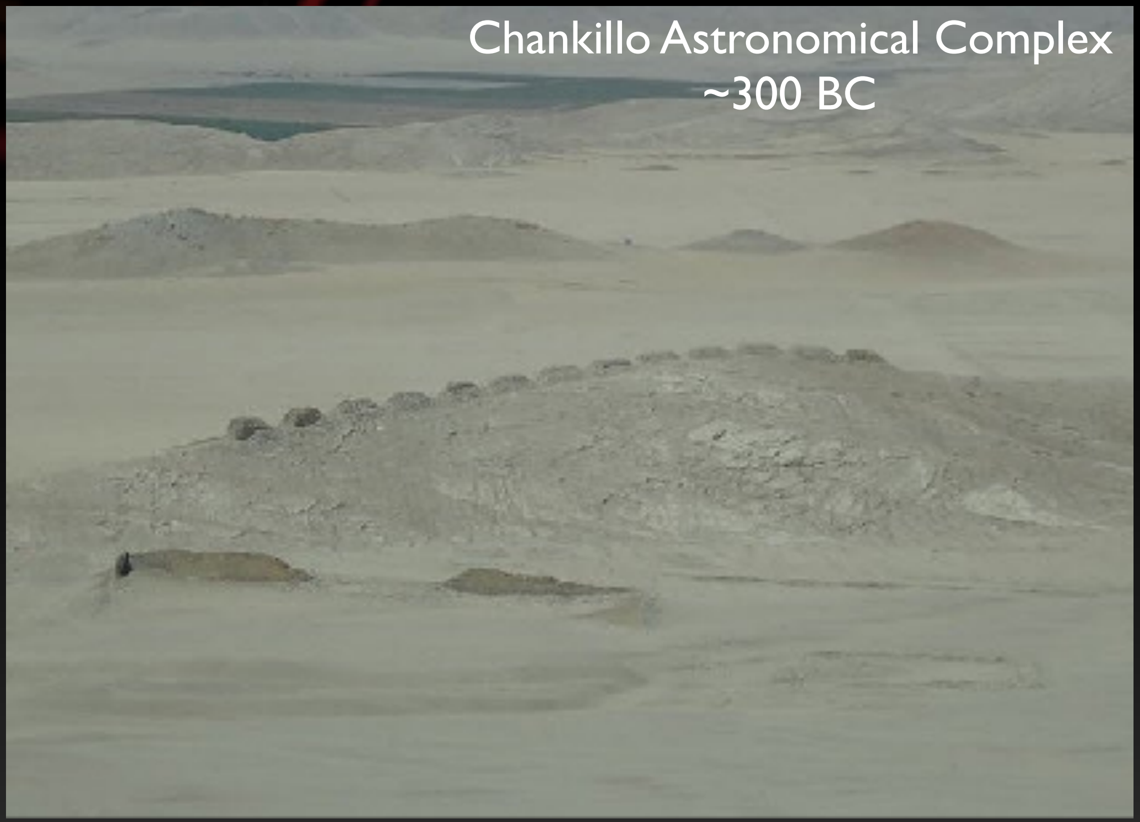


Photo credit: David Edgar

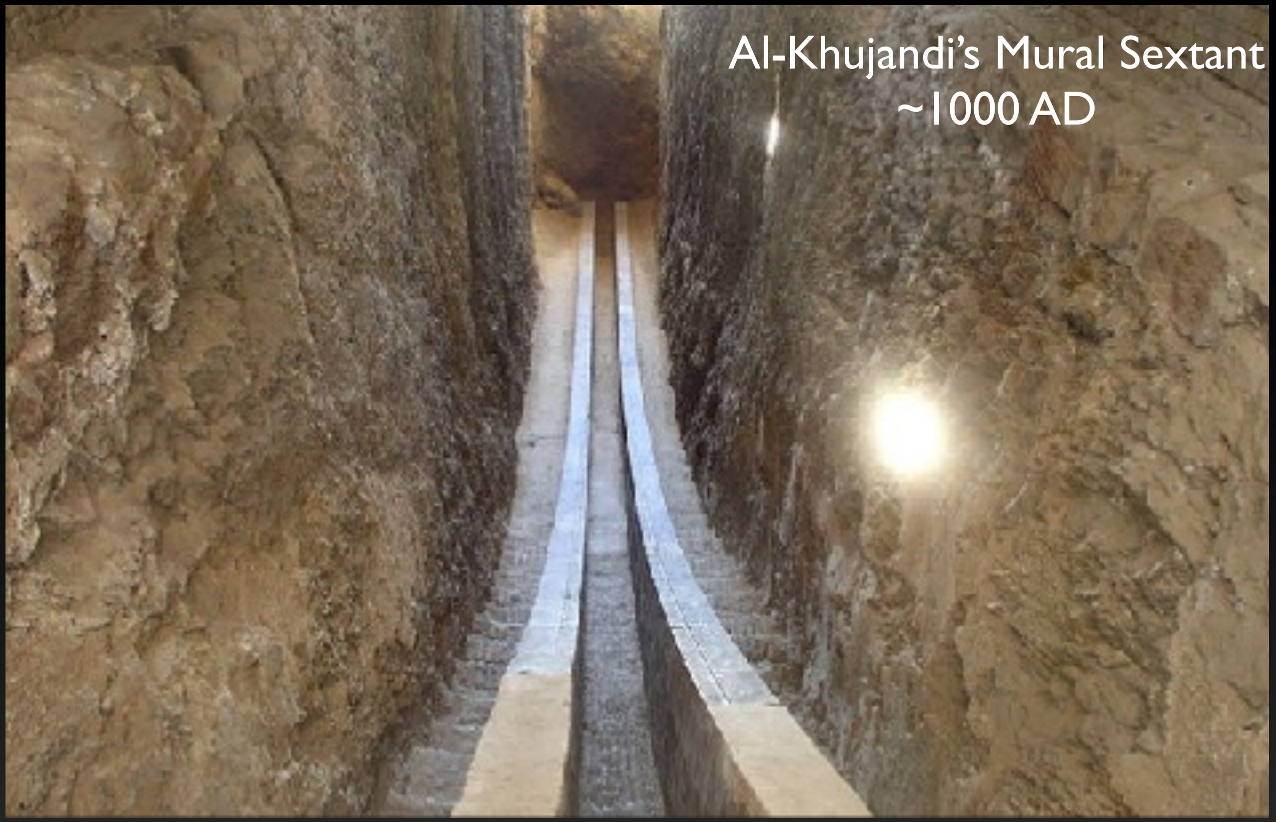
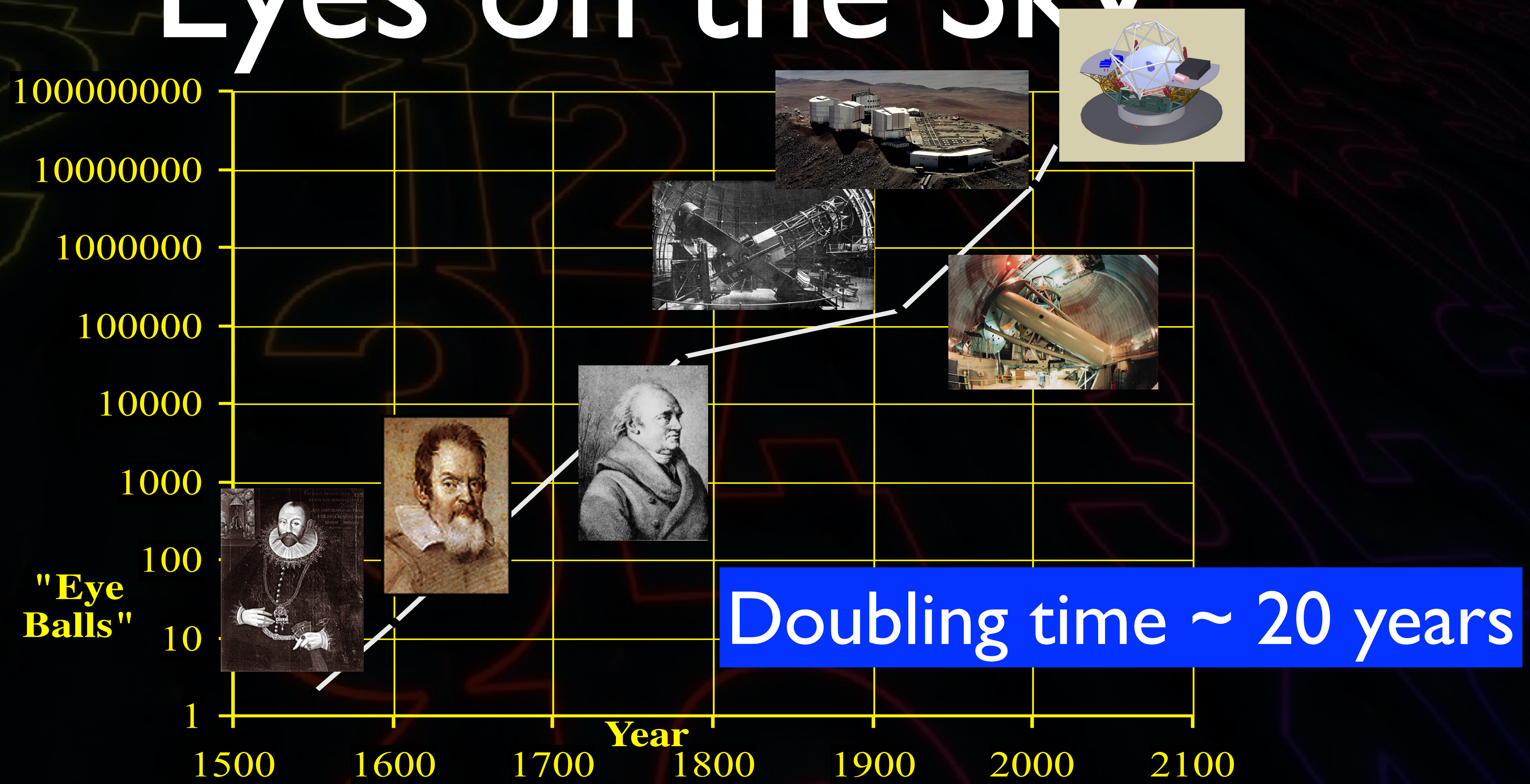


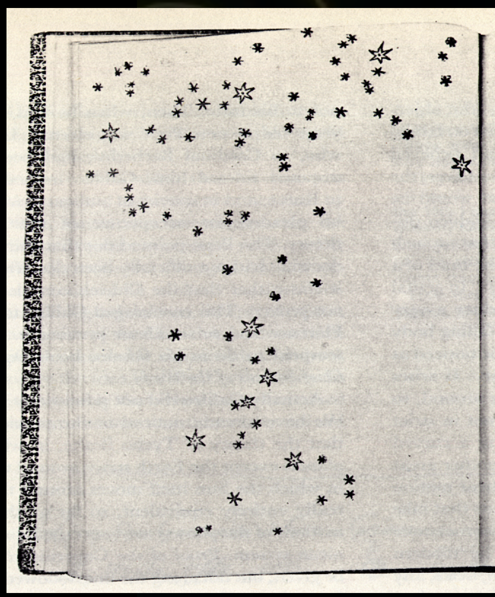
Photo credit: Alaexis

Photo credit: Leoboudv

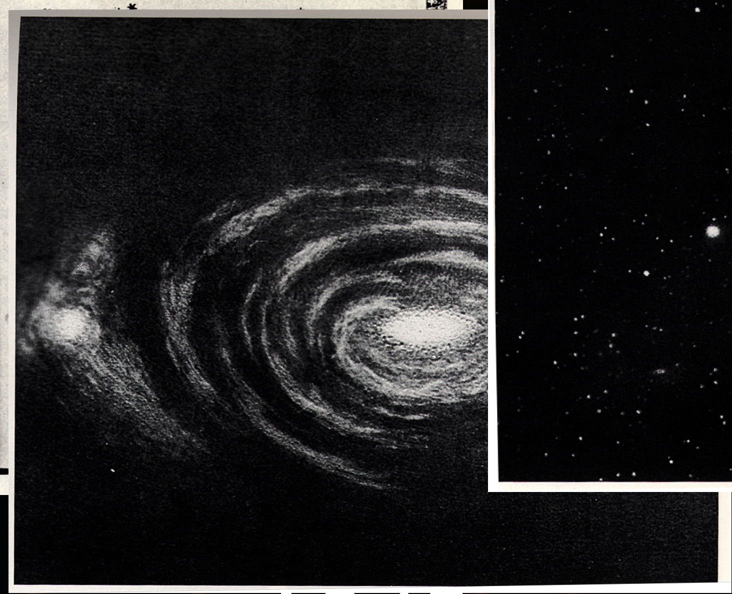
Eyes on the Sky



Gathering information

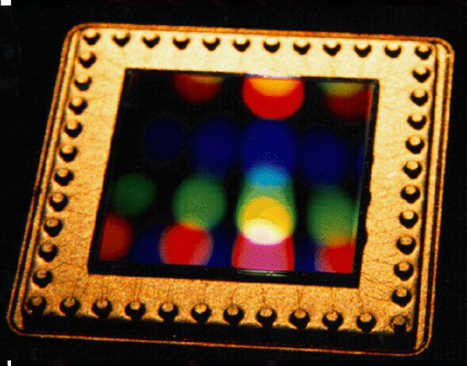


1598
Nearby stars



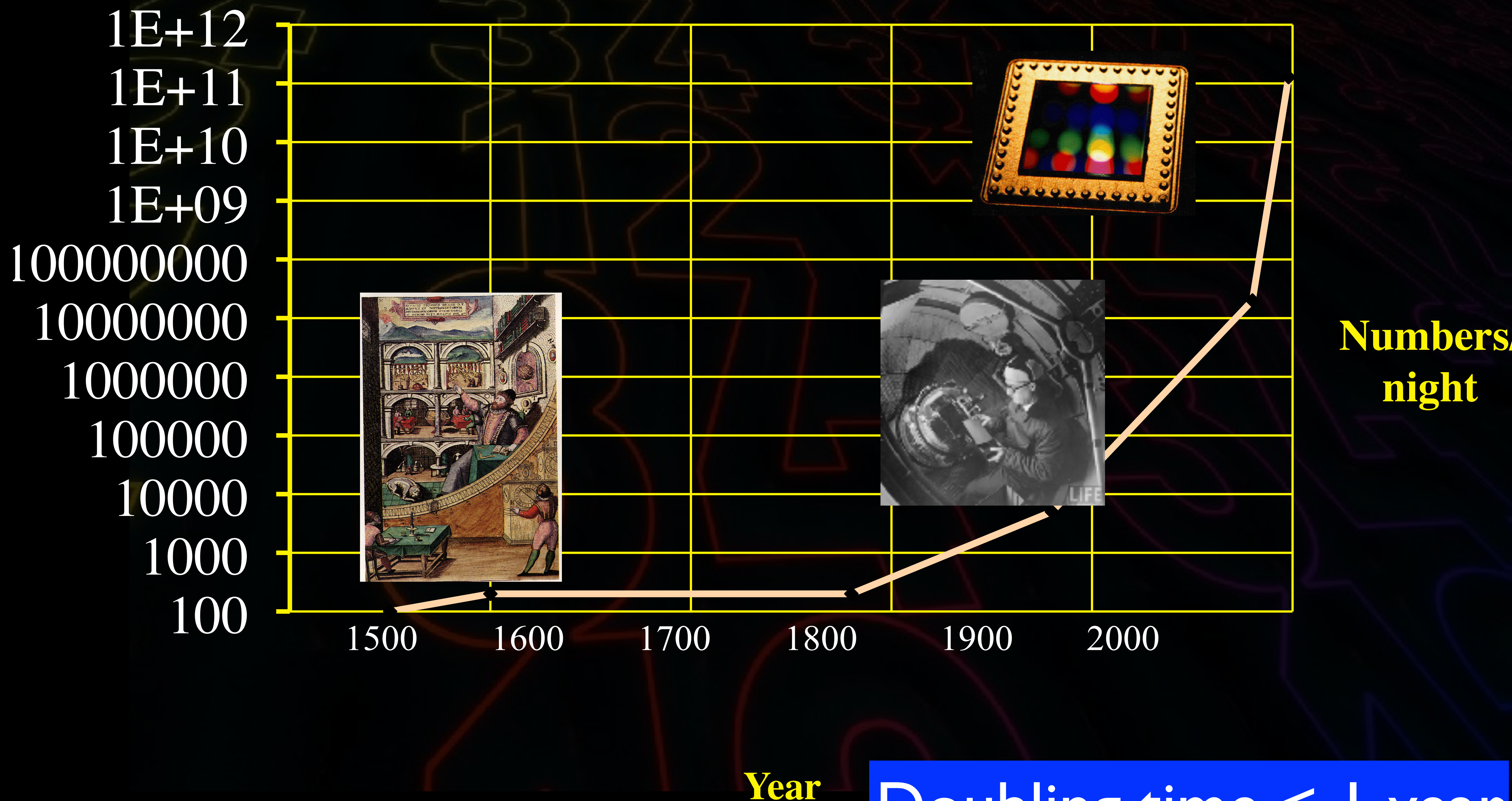
1845
Ink sketch of
nearby galaxy

1880
First photographs



1969
Invention of CCD

Numbers per night

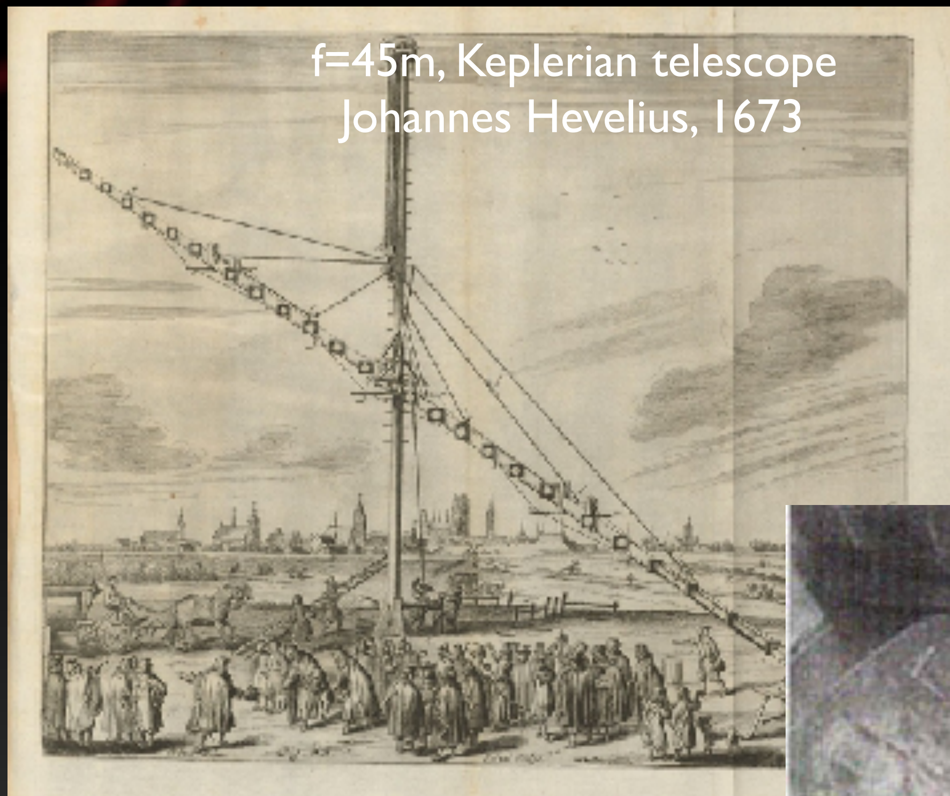


Doubling time < 1 year

The First Revolution: Telescopes and Photographic Plates

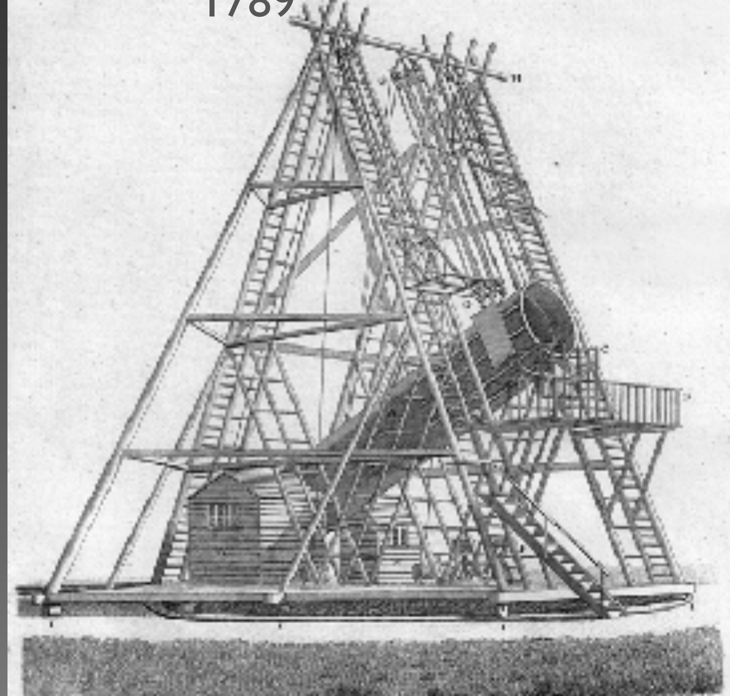
Replica of Newton's second telescope

Photo credit: Andrew Dunn

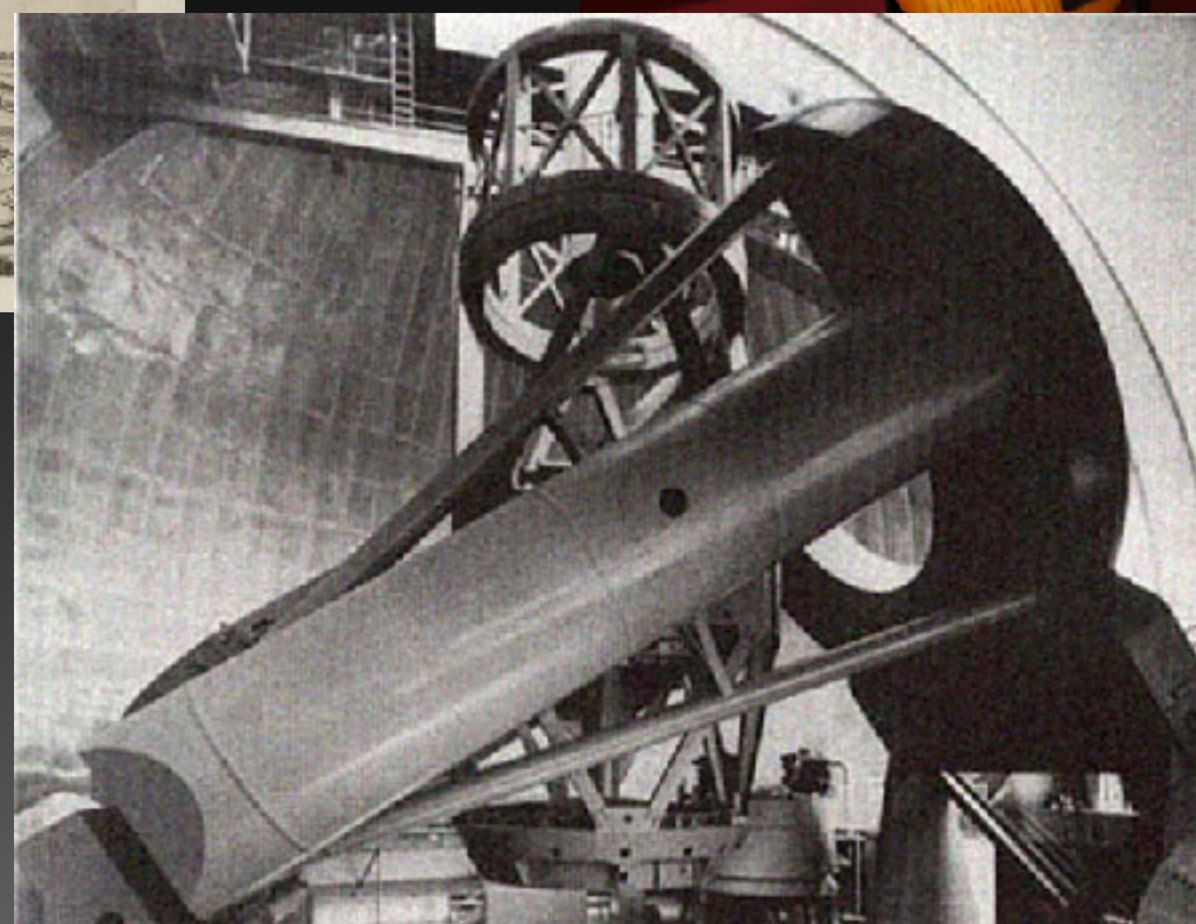


f=45m, Keplerian telescope
Johannes Hevelius, 1673

Houghton Library at Harvard University



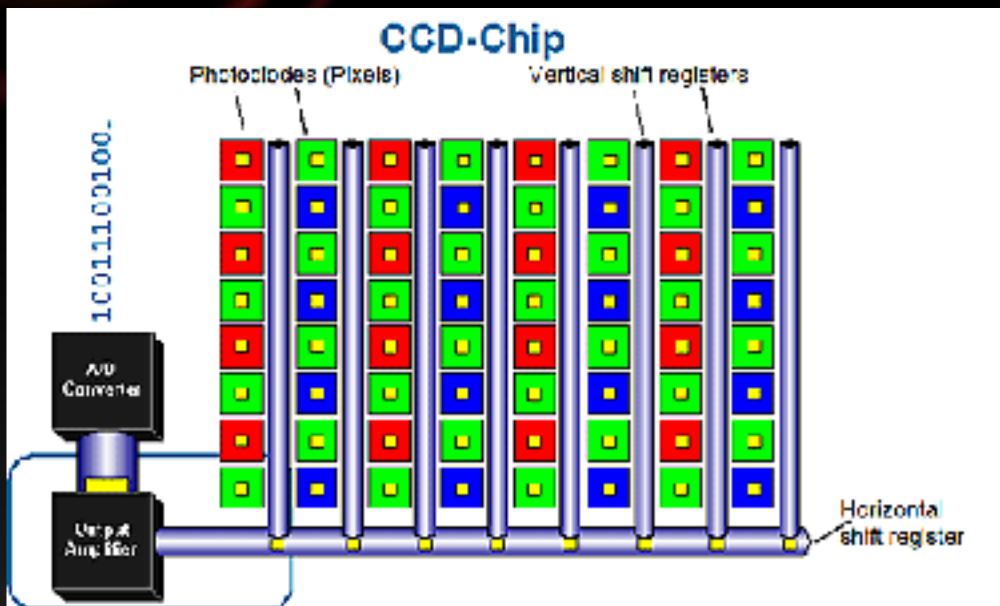
Herschel 49" Telescope
1789



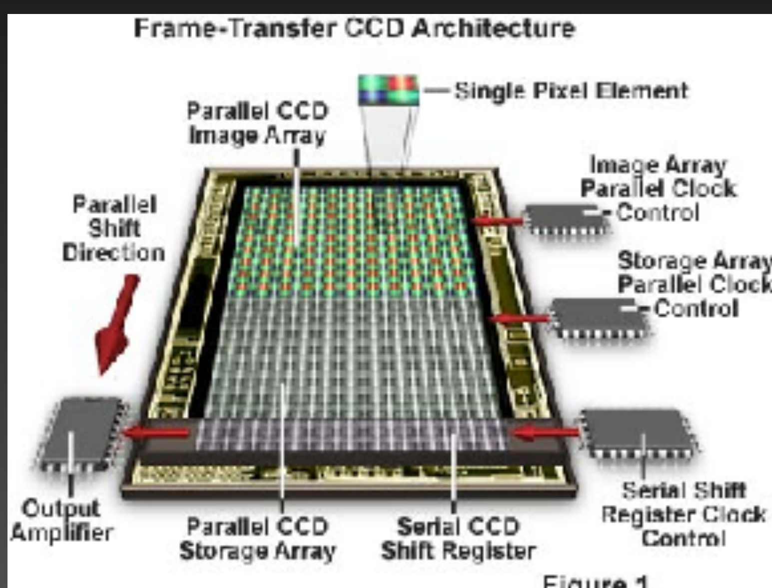
Mt. Palomar's 200-inch Hale Telescope, pointing to the zenith, as seen from the east side.

The Second Revolution: CCDs

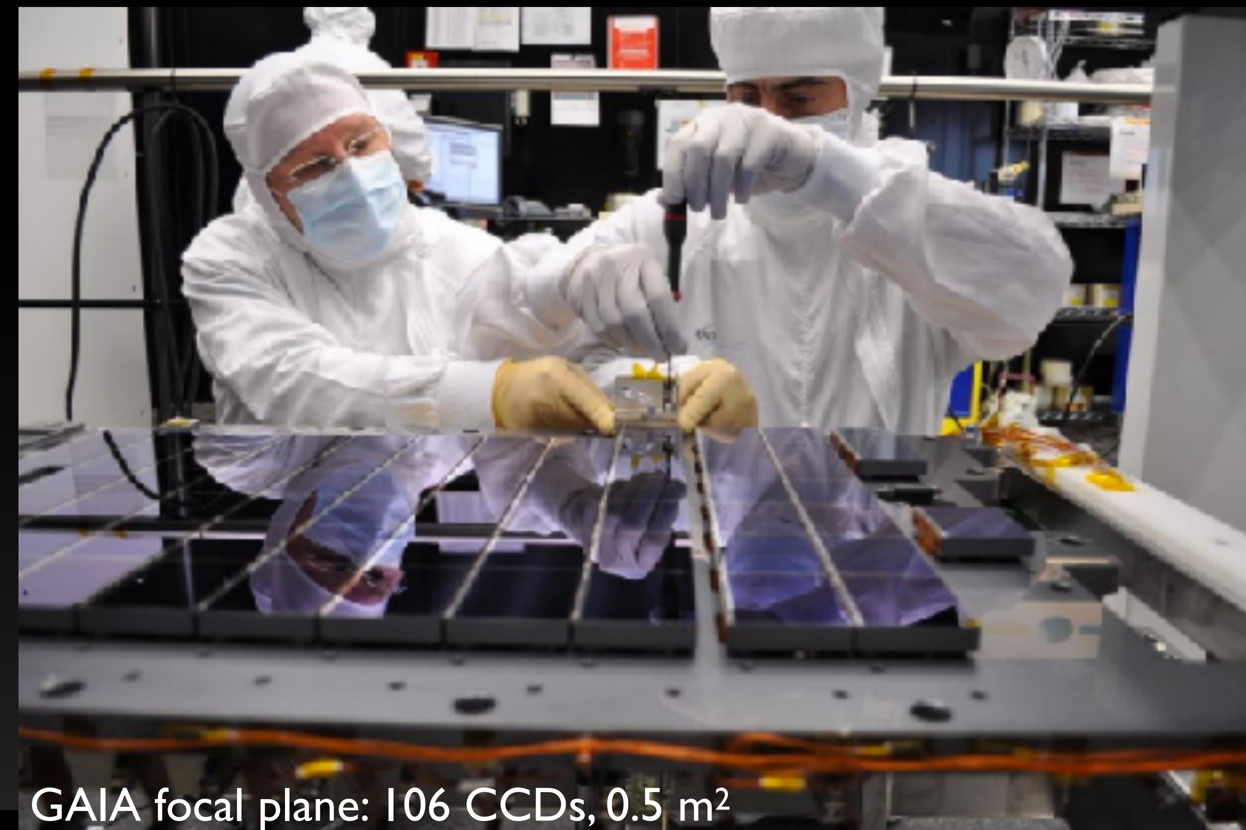
Schematics of a CCD



Graphics credit: nexyad.net



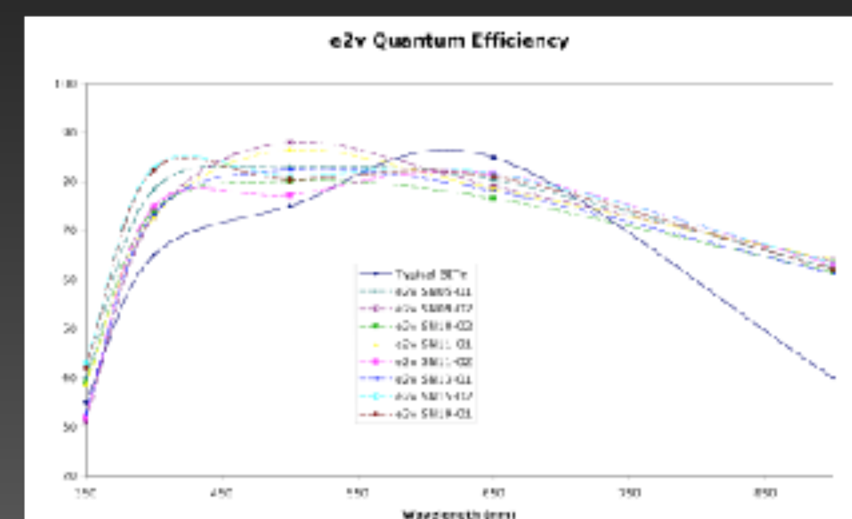
Graphics credit: Olympus Microscopy Research Centre



GAIA focal plane: 106 CCDs, 0.5 m²

Image credit: Astrium

ICRAR Starfox: Wide-field,
low surface brightness imaging
FLI camera: 8173x6132 pix



E2V QE plot

Image credit: NAOA

The Third Revolution: Multi-wavelength and Satellites

Graphics credit: University of Chicago

Astronomical Data Collections in the Past…

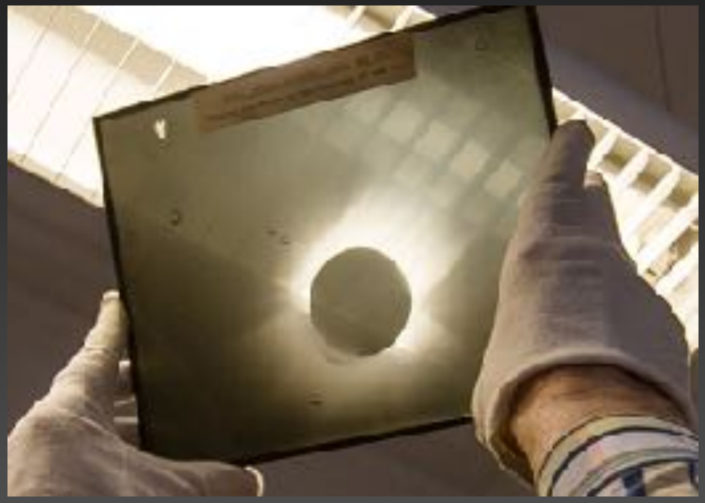
Photographic Plates

Enormous Efforts

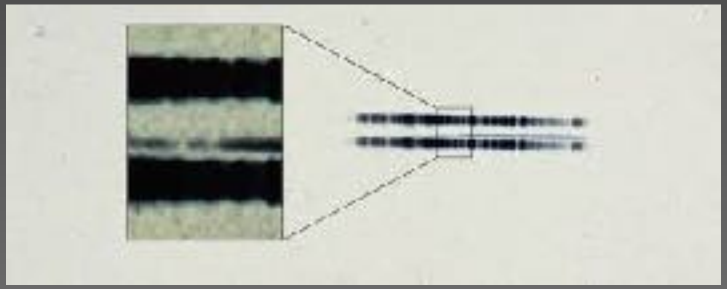
*In July 1850, daguerrotype photographer John Adams Whipple captured the first-ever picture of a star, Vega, using the observatory's mahogany-and-brass Great Refractor. By the **late 1880s**, observatory director Edward Charles Pickering had endeavoured to **photograph the entire sky,** collecting photographs from the northern and southern hemispheres. The college **shipped a 24-inch telescope** to Arequipa, Peru, in 1896, and followed with telescopes in South Africa and other locations. During the next **three decades**, astronomers slipped glass plates into the observing tubes of their telescopes, making exposures of the entire sky, and then gathered them and shipped them back to Cambridge, Mass.*

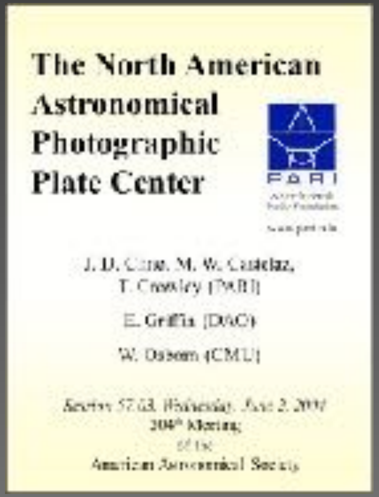
Credit: Rebecca Boyle, Popular Science

https://www.popsci.com/science/article/2011-10/recording-century-night-skies-through-scanner-darkly



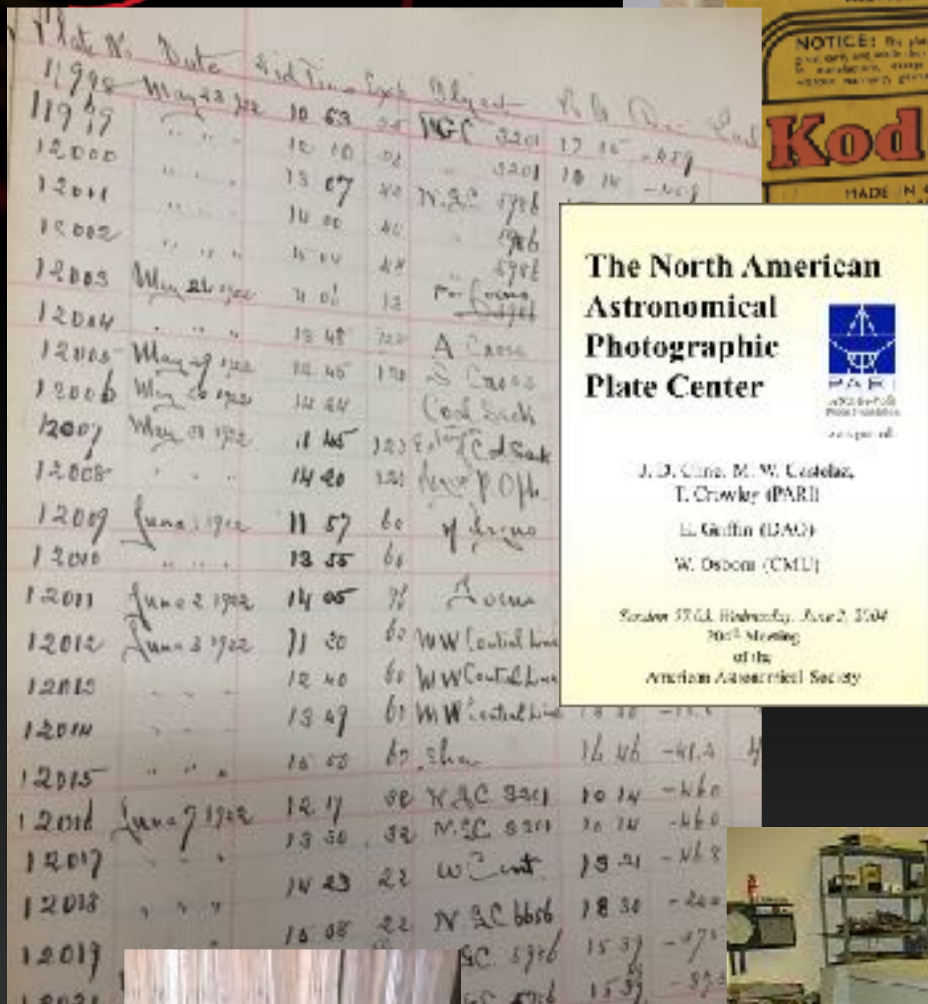
Result: ~525,000 plates!!

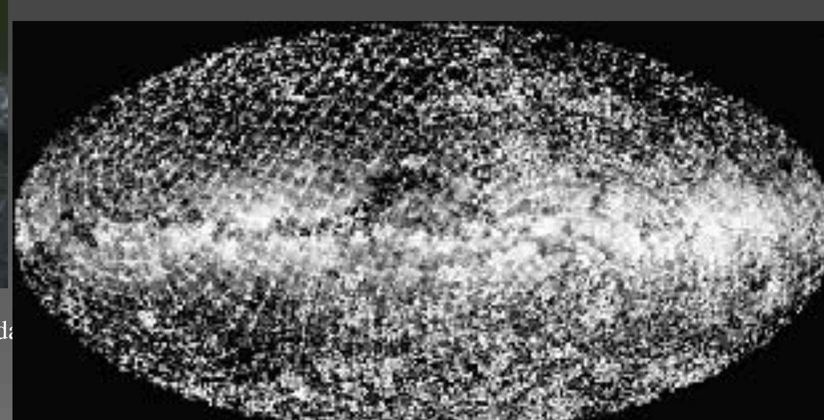



The North American Astronomical Photographic Plate Center

Credit: Hubble/ESA

Enormous Efforts





Credit: Harvard University, PARI, Koda
Mink

Thousands of plates taken with Palomar Schmidt, UK Schmidt and ESO Schmidt telescopes between 1949 and 1990s to cover the whole sky in multiple bands.

Distribution to observatories in glass, film and paper copies.

Scanning started at STScI in 1986, published as DSS in 1994 and then turned into several catalogues as well.

Hubble Space Telescope

More than just data

The Space Telescope European Coordination Facility was born out of a MoU in 1977 between NASA and ESA to provide the FOC, the solar arrays and 15 staff members in exchange of 15% observing time.

Community realised the need for more local support and expertise for observation planning and data reduction as well as a more direct access to HST data.

Credit: Hubble/ESA

More than just data

Credit: Hubble/ESA

ECF was established with ESO in 1983.

Lots of cross-fertilisation of people, software, hardware and archiving technologies between ECF and ESO.

HST data analysis workshops merged with ADASS mid 1990s.

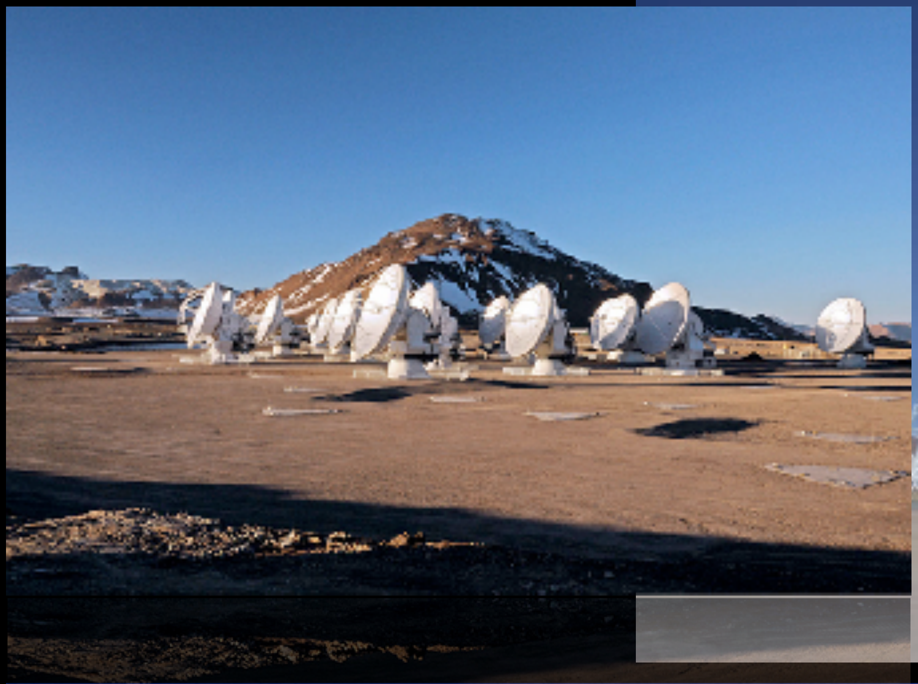
Advanced methods developed by ECF. Truly multi-national collaboration spanning many organisations.

Officially closed end of 2010.

…NOW

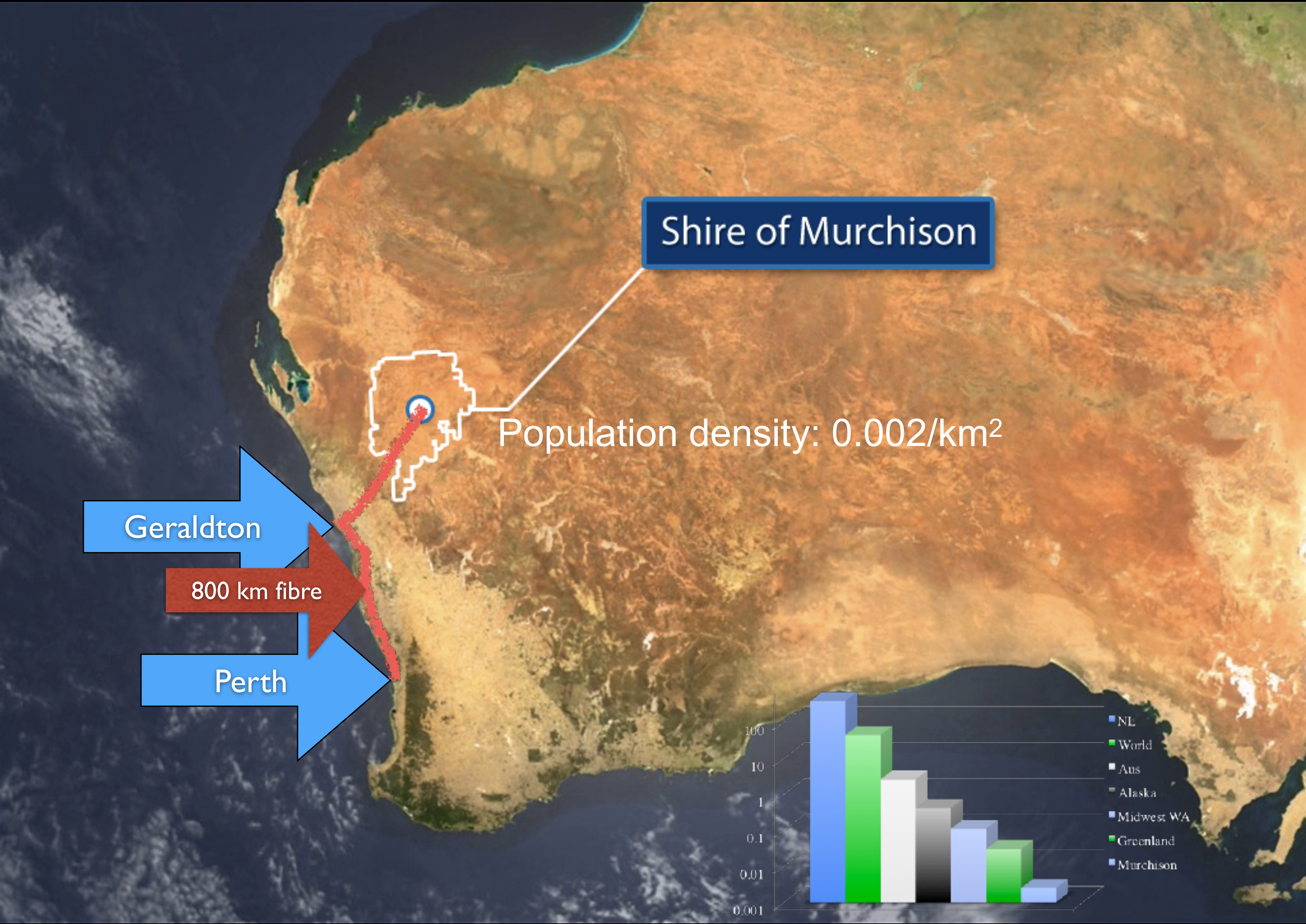
ALMA


Chile
Antofagasta
Cerro Paranal
La Silla
La Serena
Coquimbo
Valparaîso
SANTIAGO

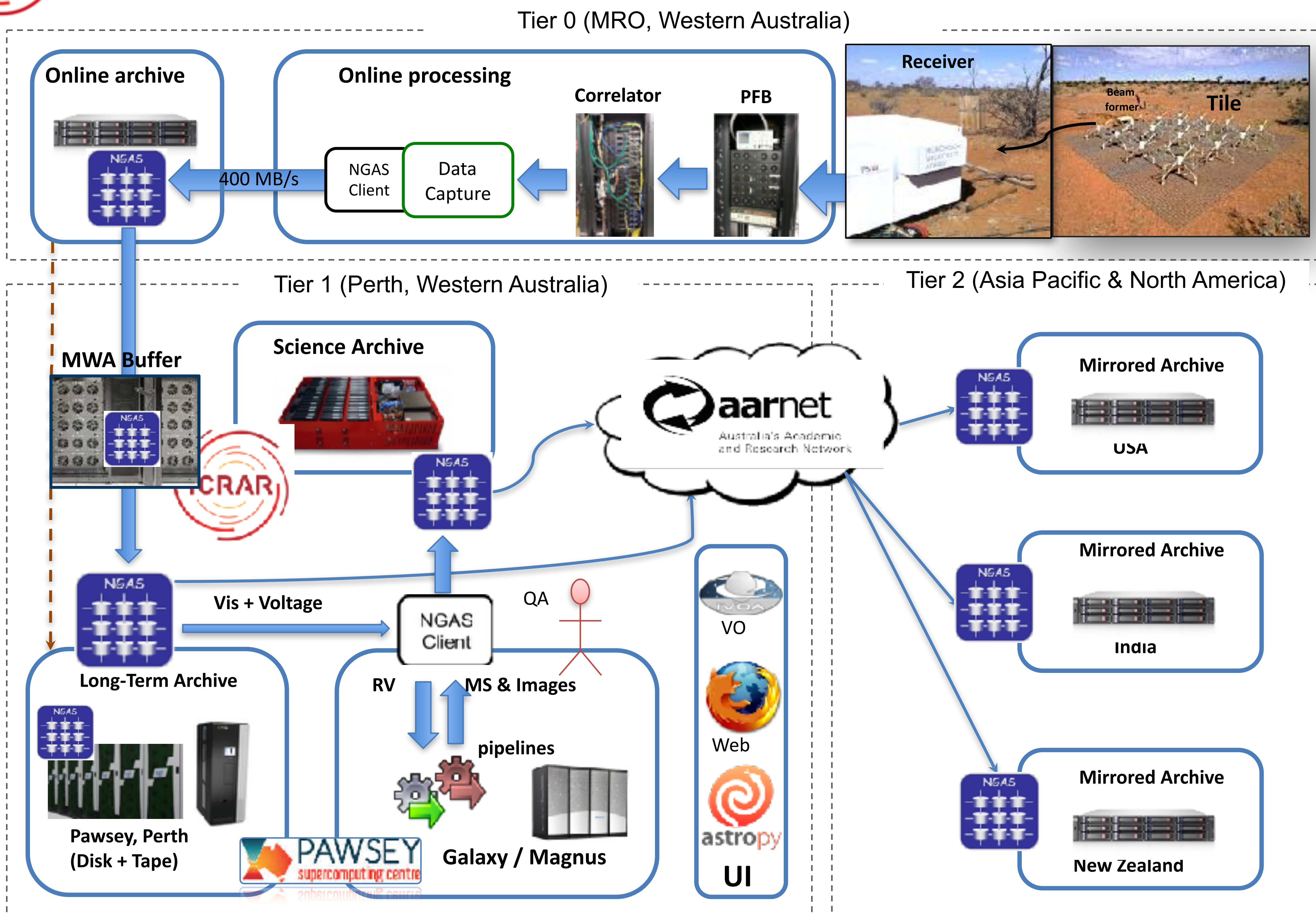


...NOW

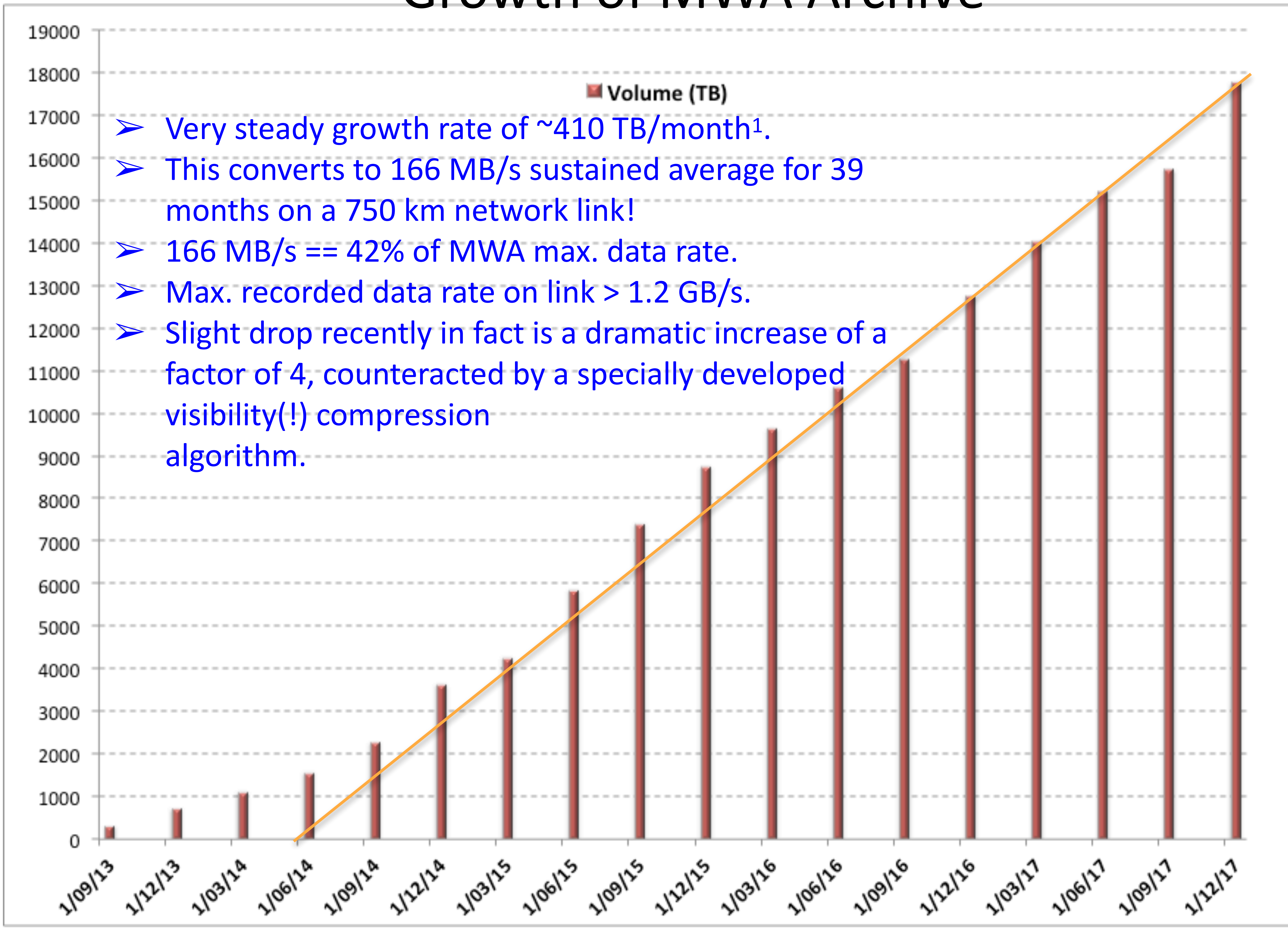
MURCHISON WIDEFIELD ARRAY
MWA


Shire of Murchison
Population density: 0.002/km²
Geraldton
800 km fibre
Perth
100
10
1
0.1
0.01
0.001
NL
World
Aus
Alaska
Midwest WA
Greenland
Murchison

Overall MWA dataflow



Growth of MWA Archive

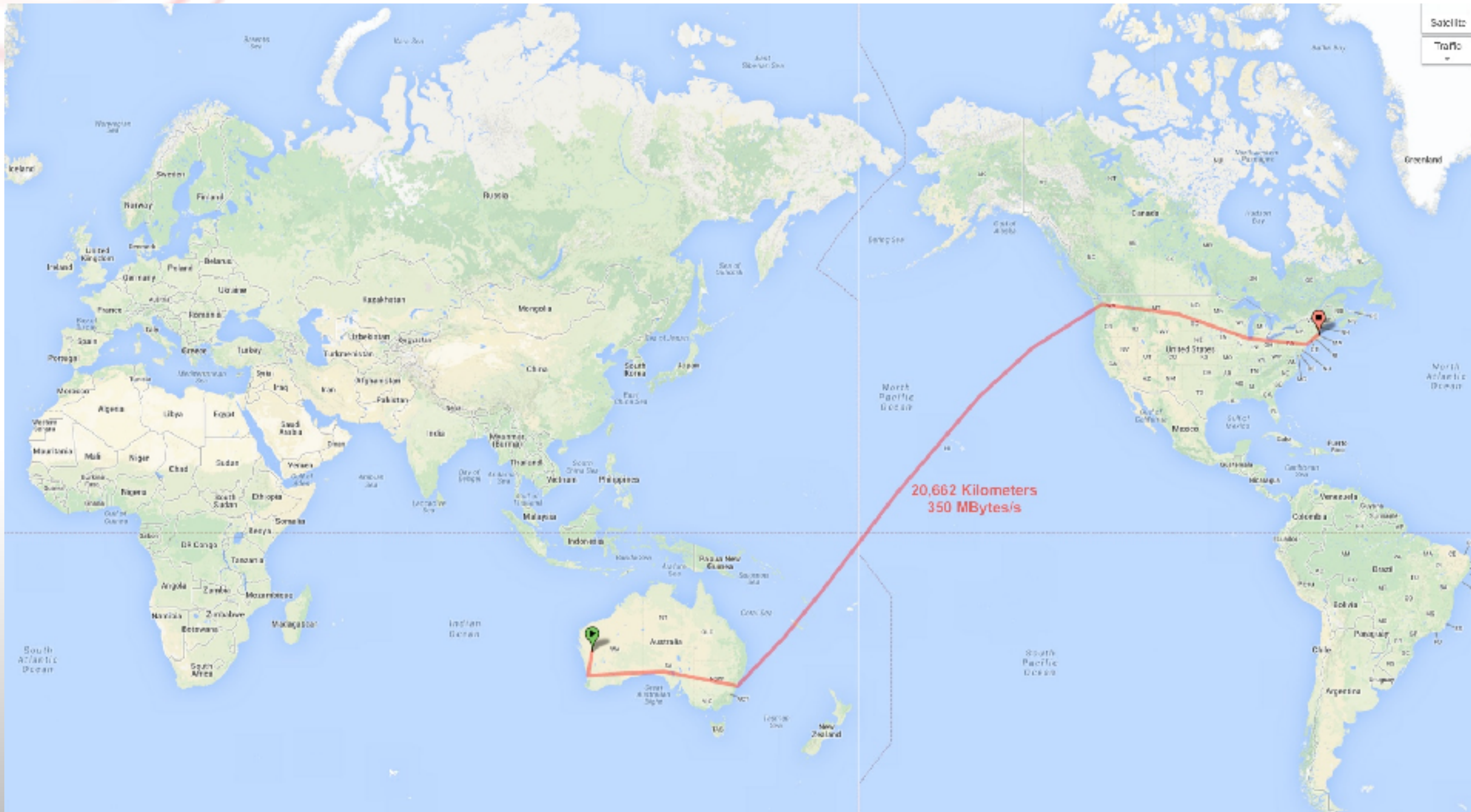


- Very steady growth rate of ~410 TB/month[1].
- This converts to 166 MB/s sustained average for 39 months on a 750 km network link!
- 166 MB/s == 42% of MWA max. data rate.
- Max. recorded data rate on link > 1.2 GB/s.
- Slight drop recently in fact is a dramatic increase of a factor of 4, counteracted by a specially developed visibility(!) compression algorithm.

As of 1 Dec 2017

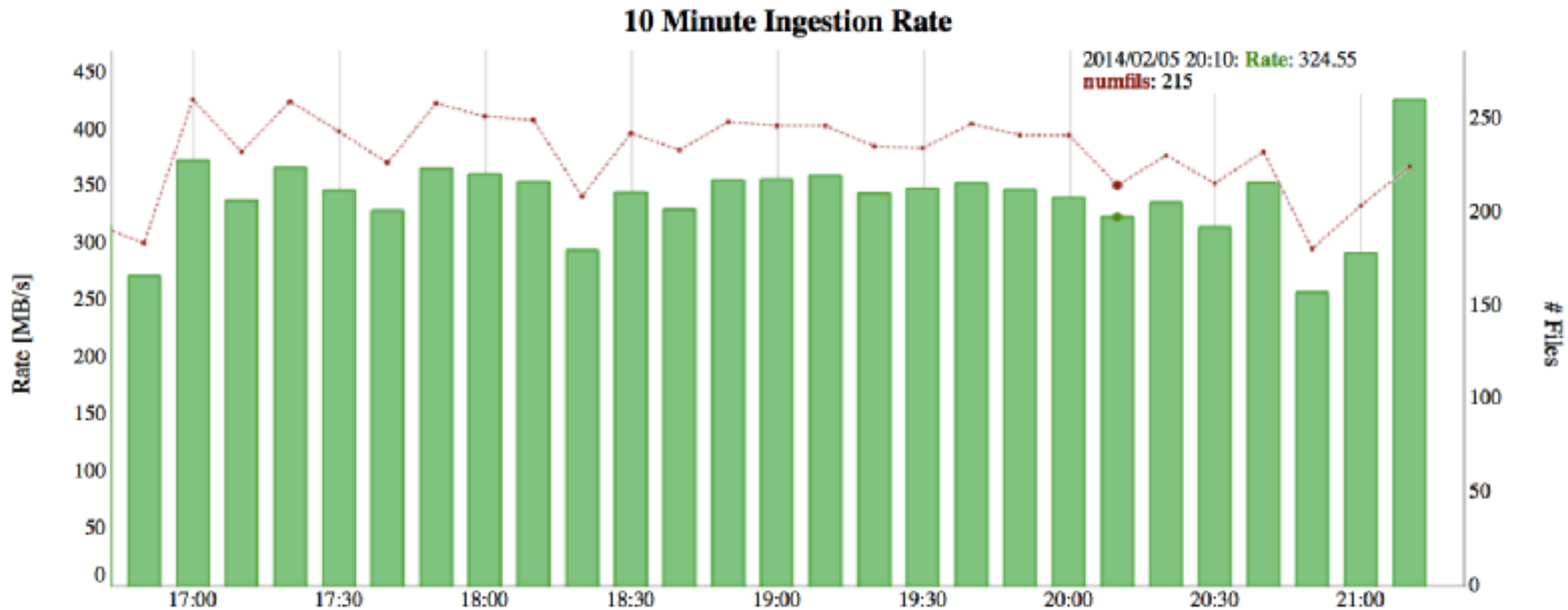
[1] About 2X the **yearly** rate suggested in M. Lacy, D.Halstead, ALMA NAASC memo 110

Long Haul Data Transport



Global Data Transfer

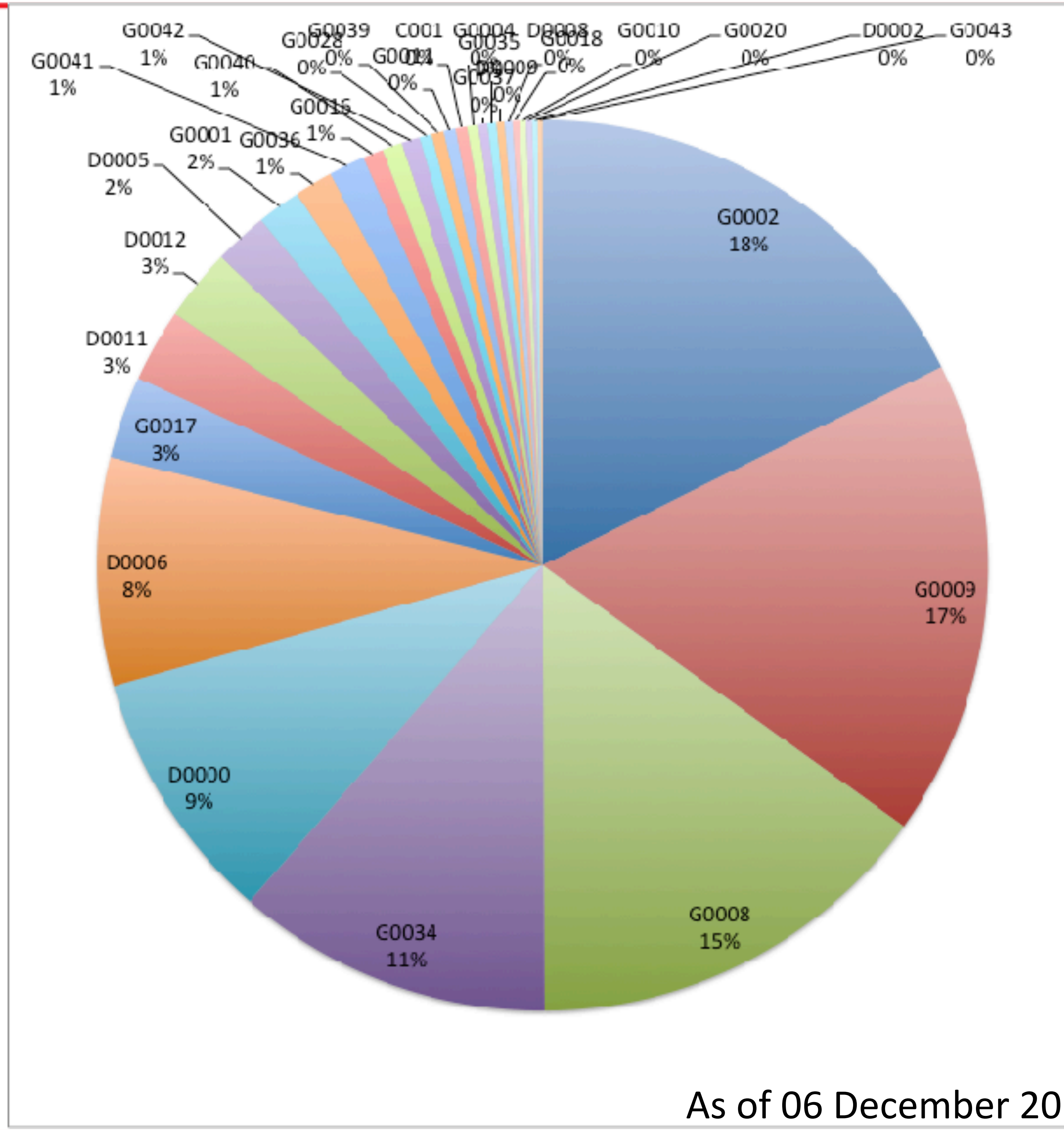
From Perth to Boston (20,000 Kilometers)



- ➢ Almost real-time data transfer of full MWA data rate over 20,000 km!
- ➢ Not using any dedicated link or arrangement.
- ➢ Was one of the highest volume and throughput, long-term sustained, scientific data transfers using just public network.

Usage of MWA Archive

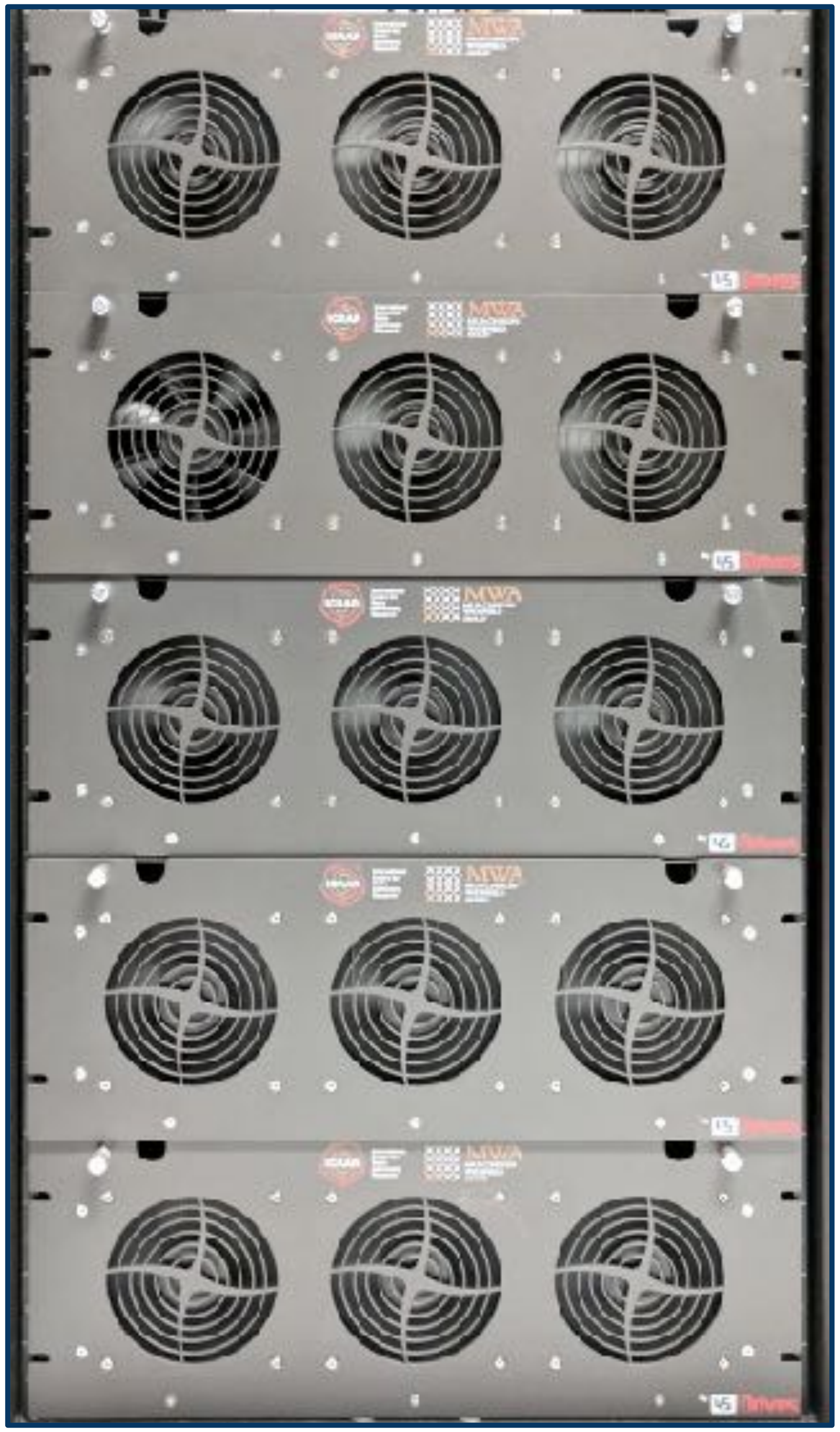
Total: 24197 TB
Not including automatic mirroring of data to partner institutions.



As of 06 December 2017

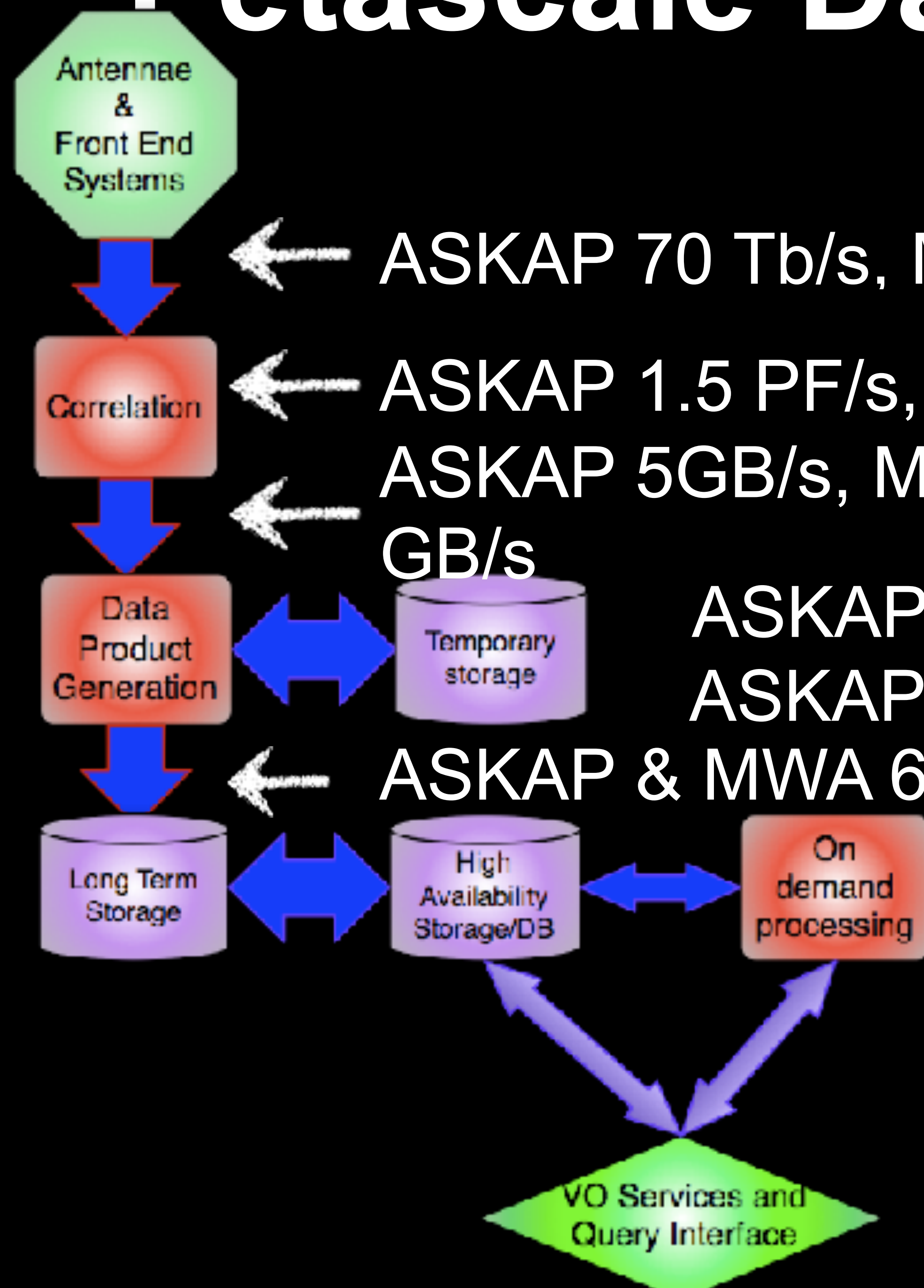
MWA Buffer

- 10 high density storage servers.
- total of ~3 PB usable capacity.
- running 2 NGAS servers each in cache mode.



ICRAR International Centre for Radio Astronomy Research

Petascale Data Flow



ASKAP 70 Tb/s, MWA 320 Gb/s, SKA

ASKAP 1.5 PF/s, SKA >1 EF/s

ASKAP 5GB/s, MWA 8 GB/s, SKA >300 GB/s

ASKAP 13 PB, SKA 800..3000 PB

ASKAP 100 TF/s, SKA 30..100 PF/s

ASKAP & MWA 6 TB/day, SKA

ASKAP 100 TF/s, SKA 30 PF/s

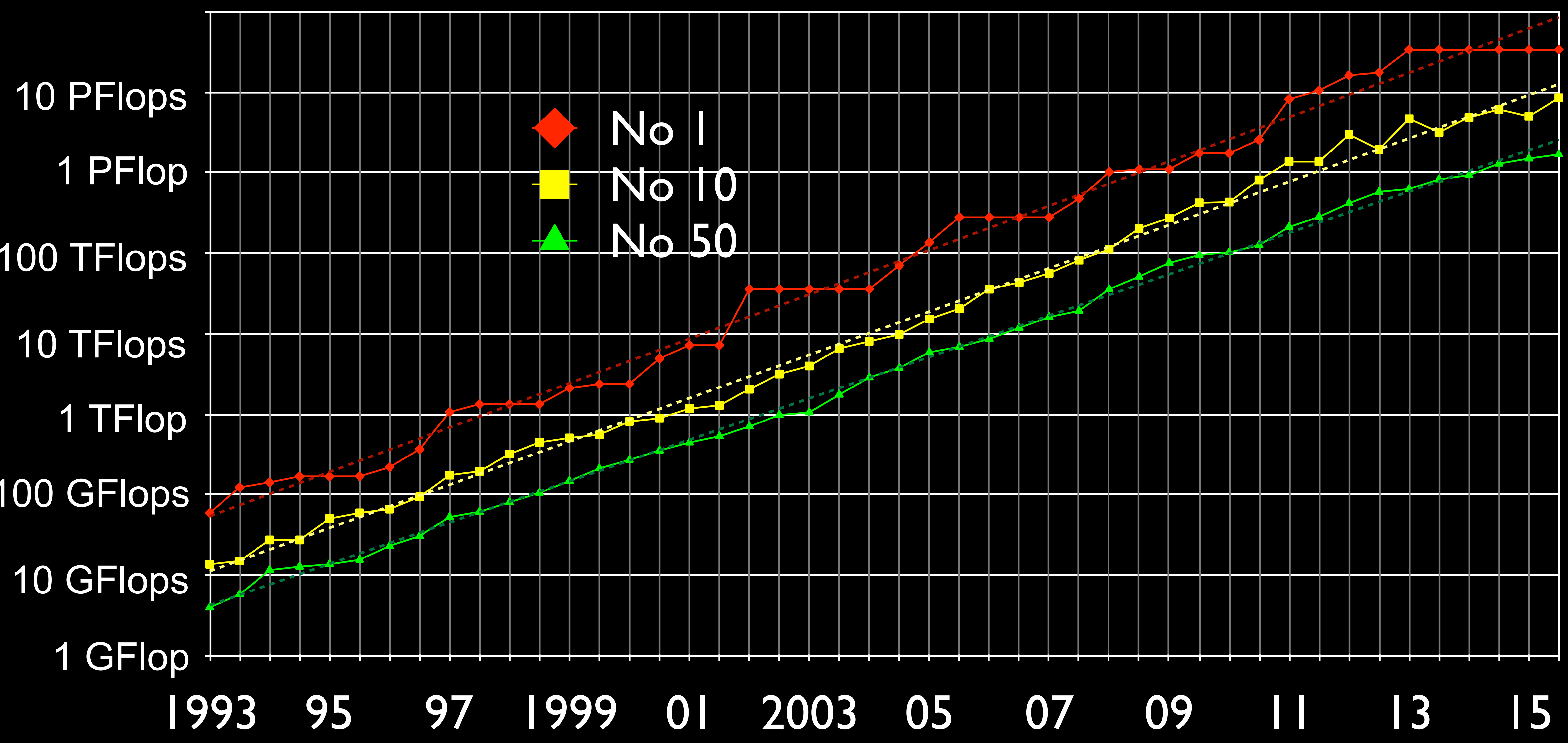
ASKAP 3 PB/year, SKA 18 PB/year

Adapted from Cornwell

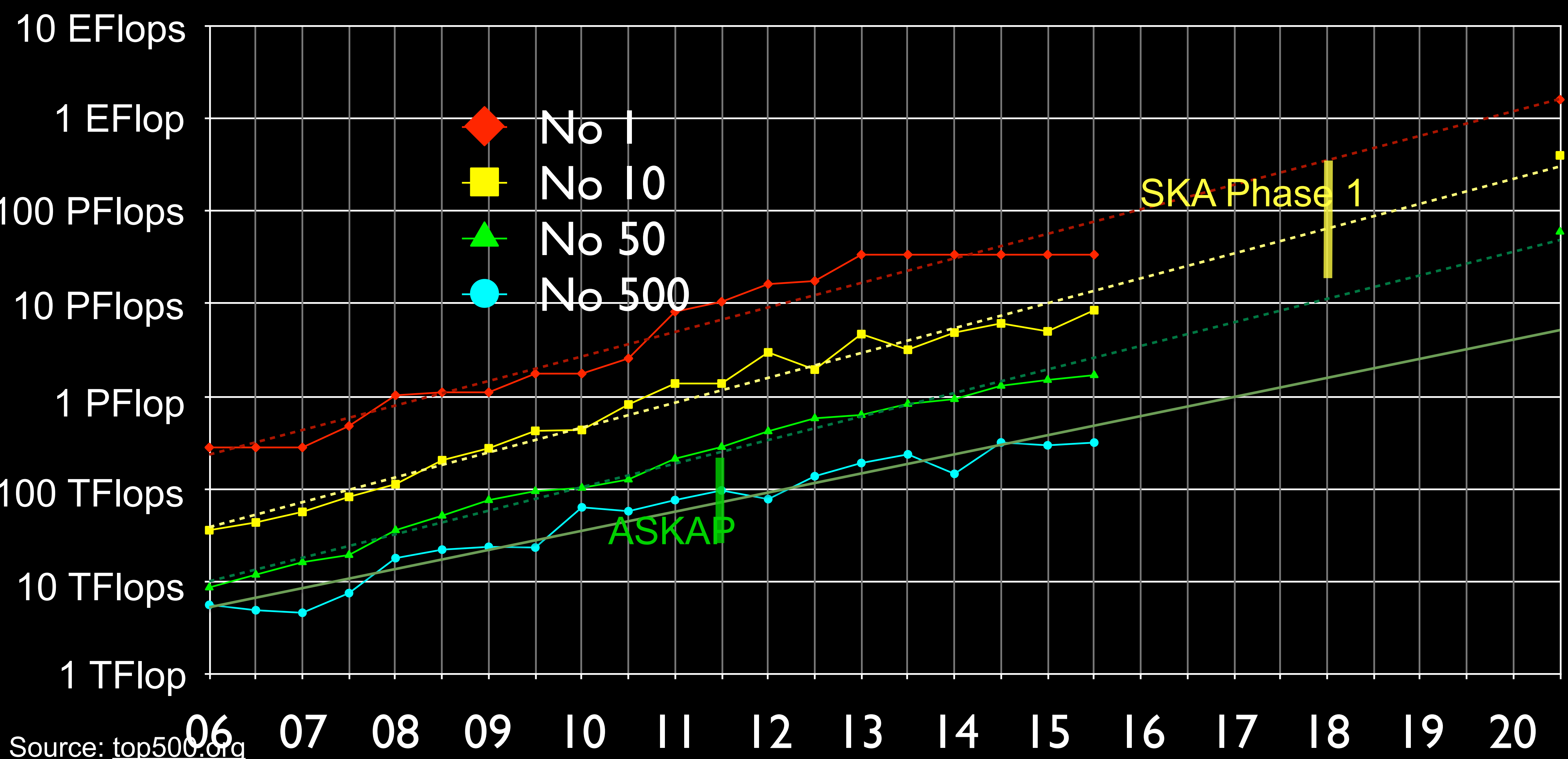
International Centre for Radio Astronomy Research

Why do we think this is possible?

Moore's Law helps



Projected Performance



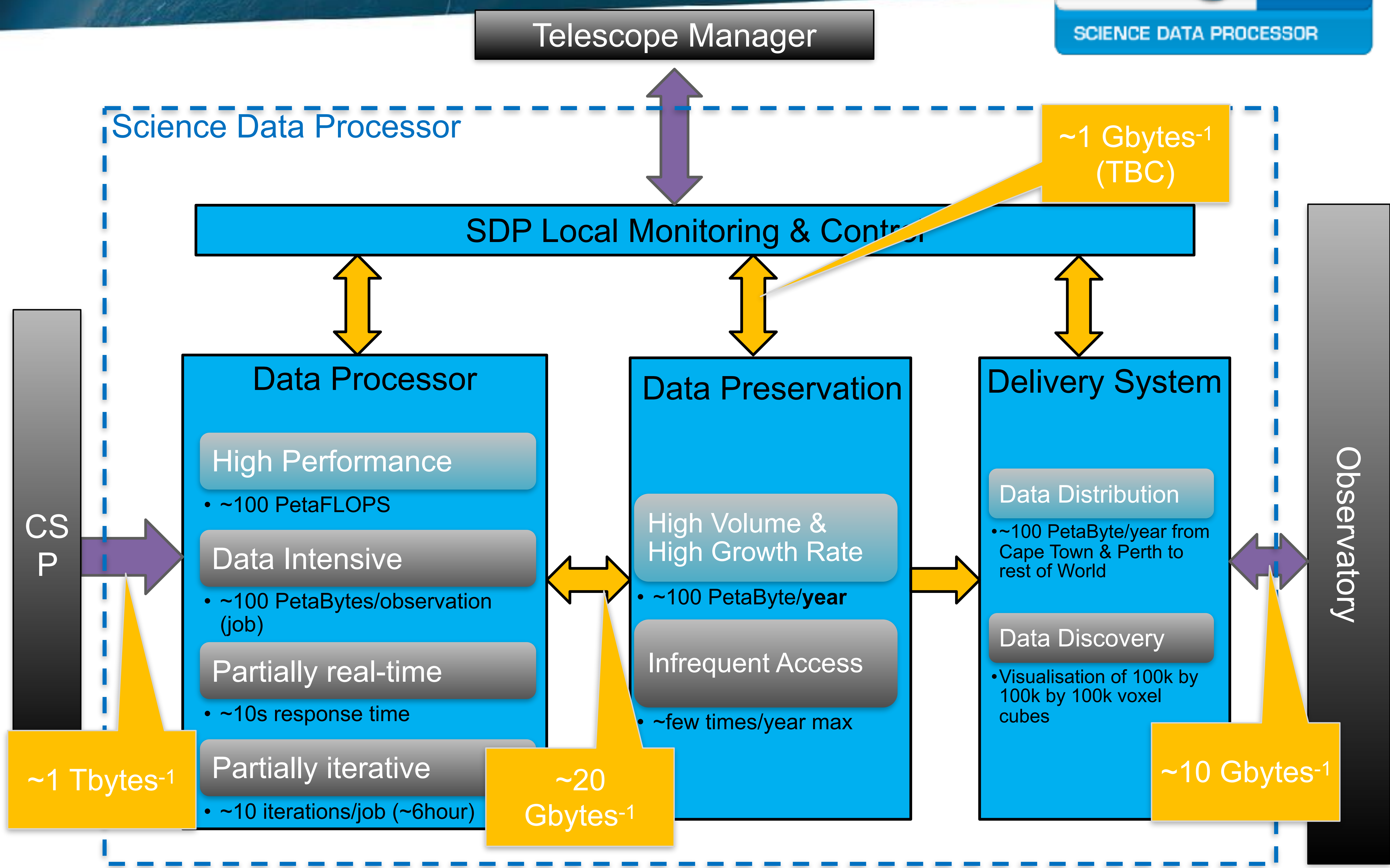
Source: top500.org

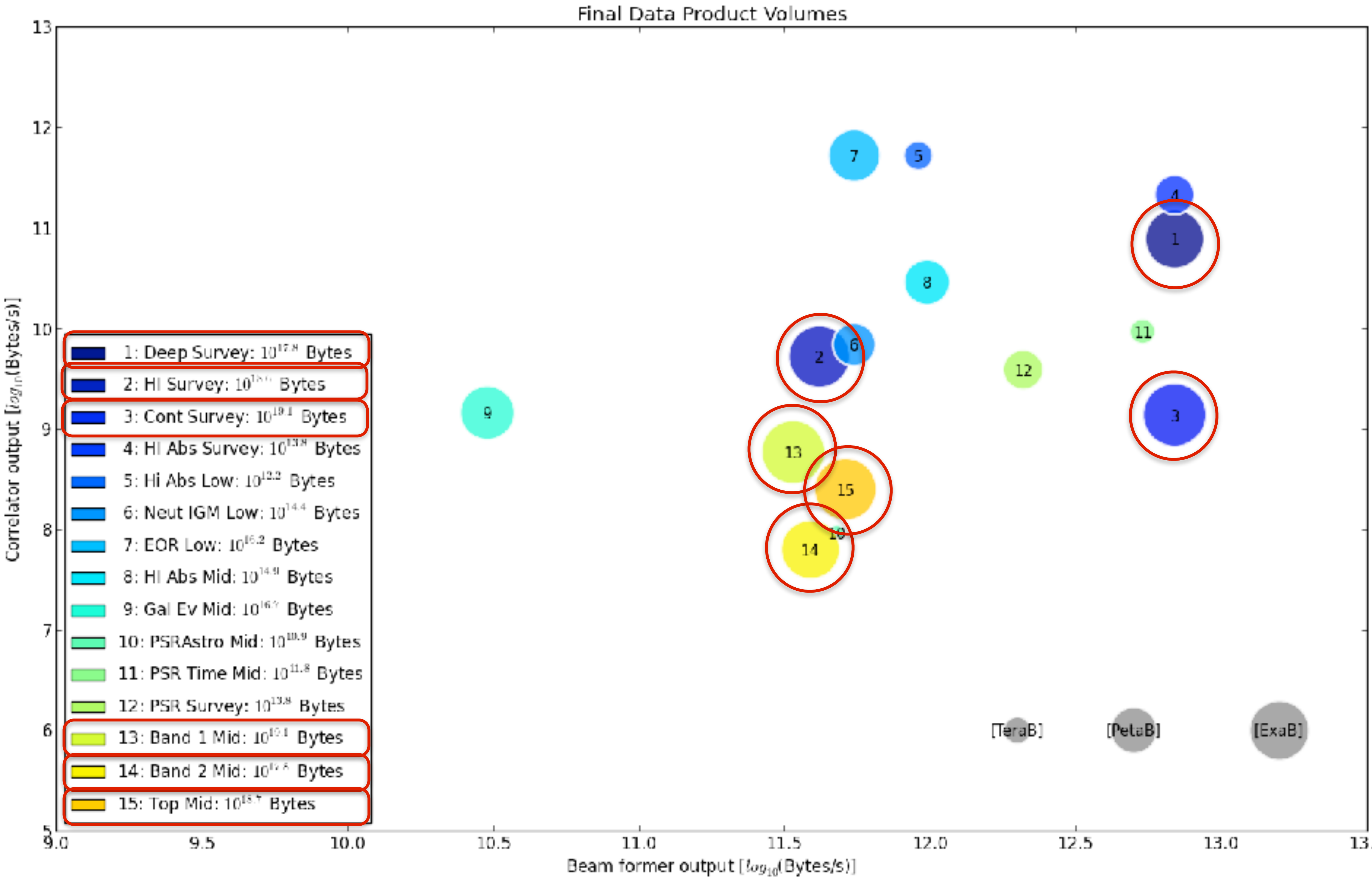
Data Rates and Volume

Telescope	Raw Data Rate	Archive Growth
MWA	1.4 TB/hour	5 PB/year
LSST	1.5 TB/hour	6 PB/year
ASKAP	9 TB/hour	5.5 PB/year
SKA1-LOW	1,400 TB/hour	150 PB/year

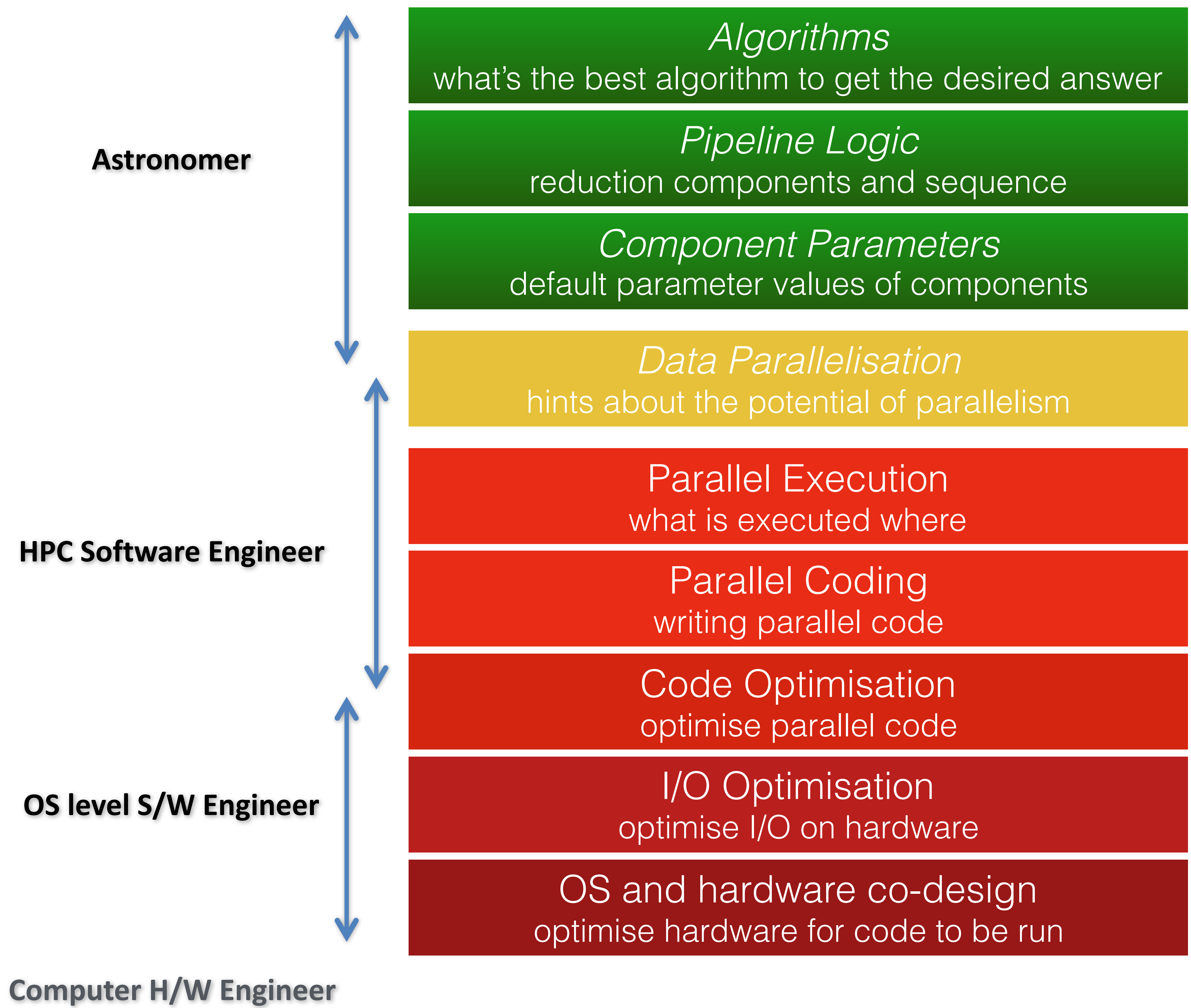
Note: A typical laptop has 1 TB of disk space. You can store all 7 Harry Potter books (~3,500 pages)150,000 times on 1 TB!

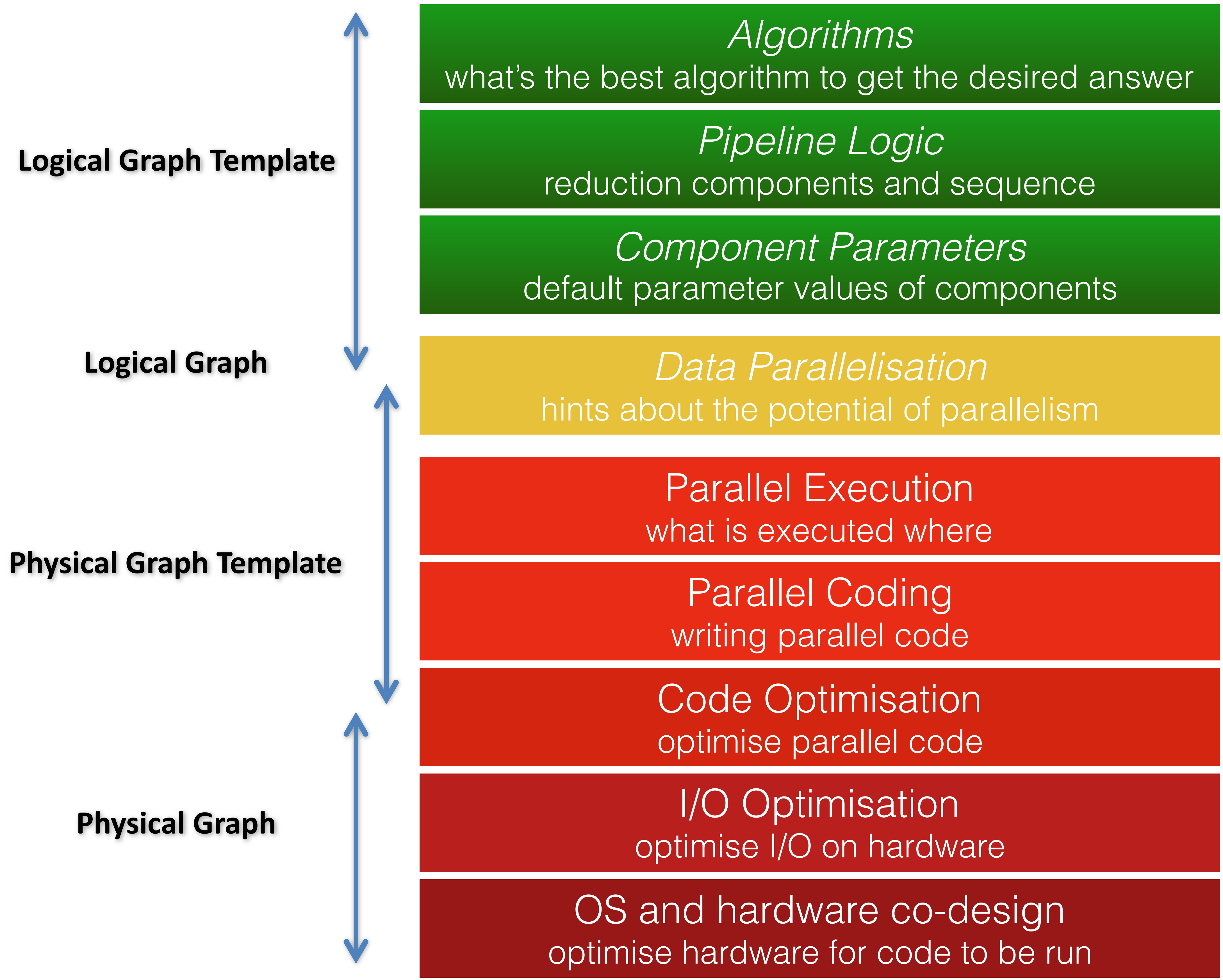
SDP Key Performance Requirements -- **SKA Phase 1**

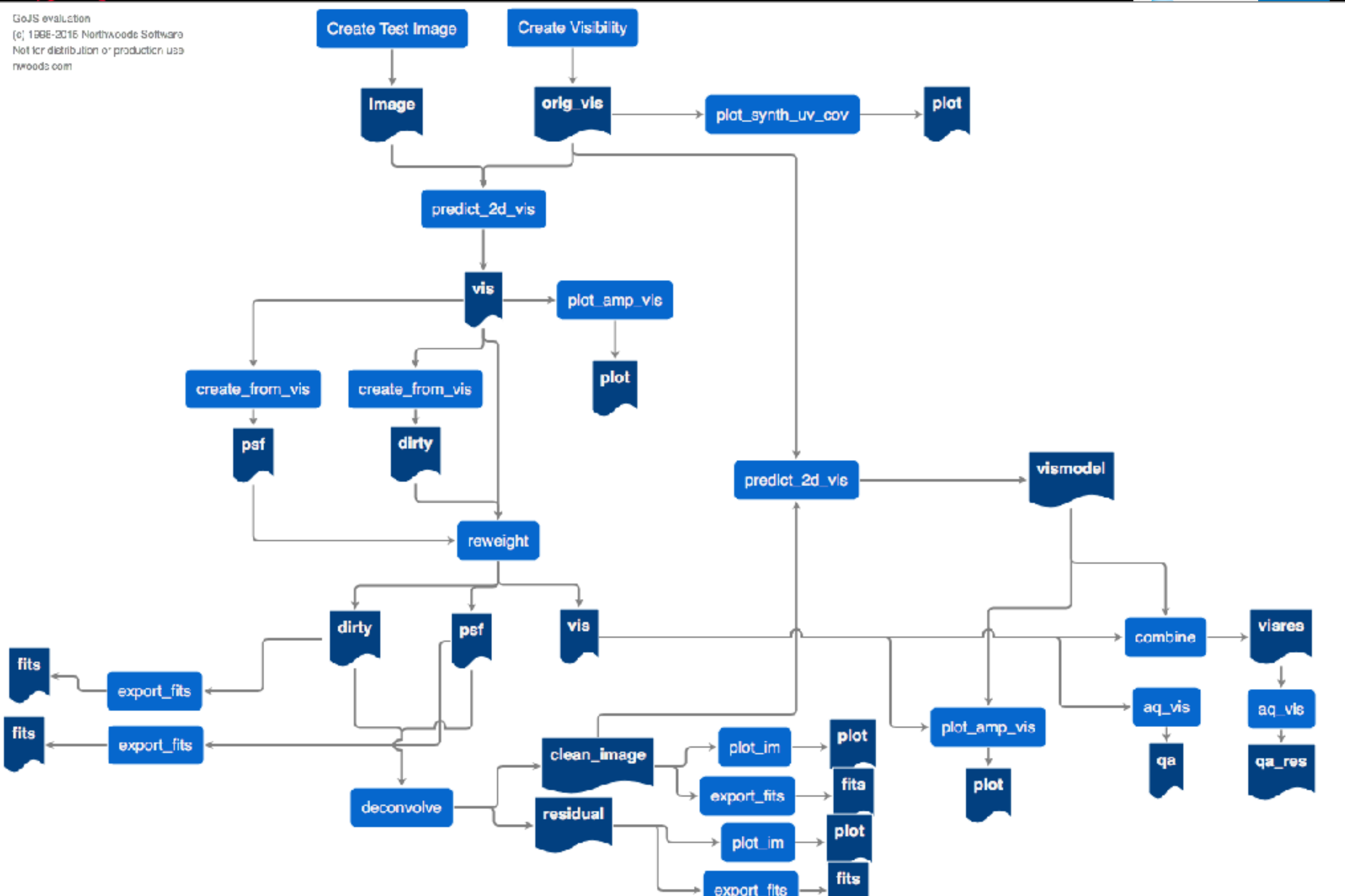




Final Data Product Volumes
Correlator output [log_{10}(Bytes/s)]
Beam former output [log_{10}(Bytes/s)]
1: Deep Survey: $10^{17.8}$ Bytes
2: HI Survey: 10 Bytes
3: Cont Survey: $10^{19.1}$ Bytes
4: HI Abs Survey: $10^{13.8}$ Bytes
5: Hi Abs Low: $10^{12.2}$ Bytes
6: Neut IGM Low: $10^{14.4}$ Bytes
7: EOR Low: $10^{16.2}$ Bytes
8: HI Abs Mid: $10^{14.9}$ Bytes
9: Gal Ev Mid: 10 Bytes
10: PSRAstro Mid: $10^{10.9}$ Bytes
11: PSR Time Mid: $10^{11.8}$ Bytes
12: PSR Survey: $10^{13.8}$ Bytes
13: Band 1 Mid: 10 Bytes
14: Band 2 Mid: 10 Bytes
15: Top Mid: $10^{15.7}$ Bytes
[TeraB]
[PetaB]
[ExaB]


Astronomer
HPC Software Engineer
OS level S/W Engineer
Computer H/W Engineer
Algorithms
what's the best algorithm to get the desired answer
Pipeline Logic
reduction components and sequence
Component Parameters
default parameter values of components
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

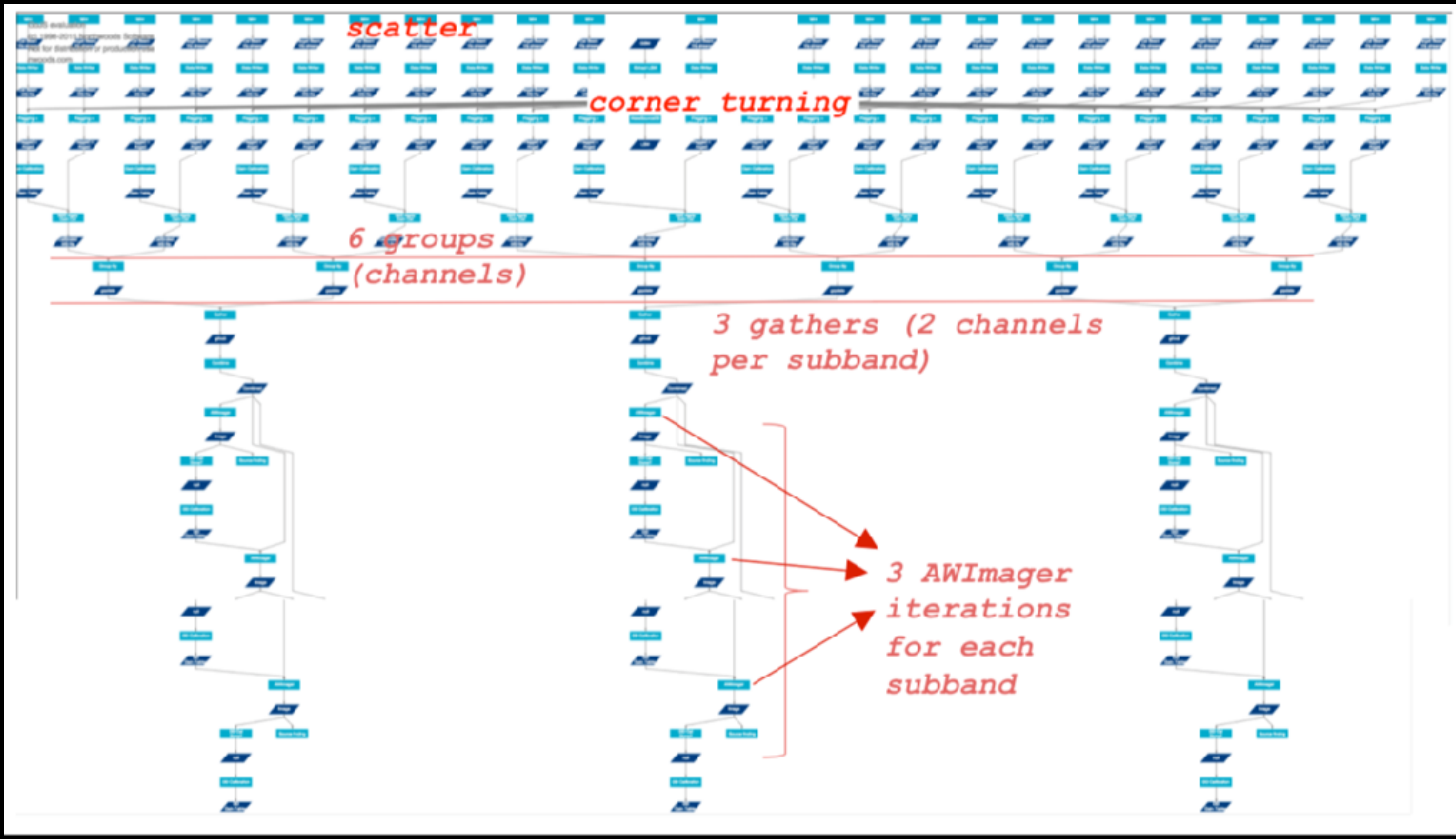

Algorithms
what's the best algorithm to get the desired answer
Pipeline Logic
reduction components and sequence
Component Parameters
default parameter values of components
Data Parallelisation
hints about the potential of parallelism
Parallel Execution
what is executed where
Parallel Coding
writing parallel 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
Logical Graph Template
Logical Graph
Physical Graph Template
Physical Graph


Create Test Image
Create Visibility
Image
orig_vis
plot_synth_uv_cov
plot
predict_2d_vis
vis
plot_amp_vis
plot
create_from_vis
create_from_vis
psf
dirty
predict_2d_vis
vismodel
reweight
dirty
psf
vis
combine
visres
fits
export_fits
fits
export_fits
aq_vis
aq_vis
clean_image
plot_im
plot
plot_amp_vis
qa
qa_res
export_fits
fits
plot
deconvolve
residual
plot_im
plot
export_fits
fits

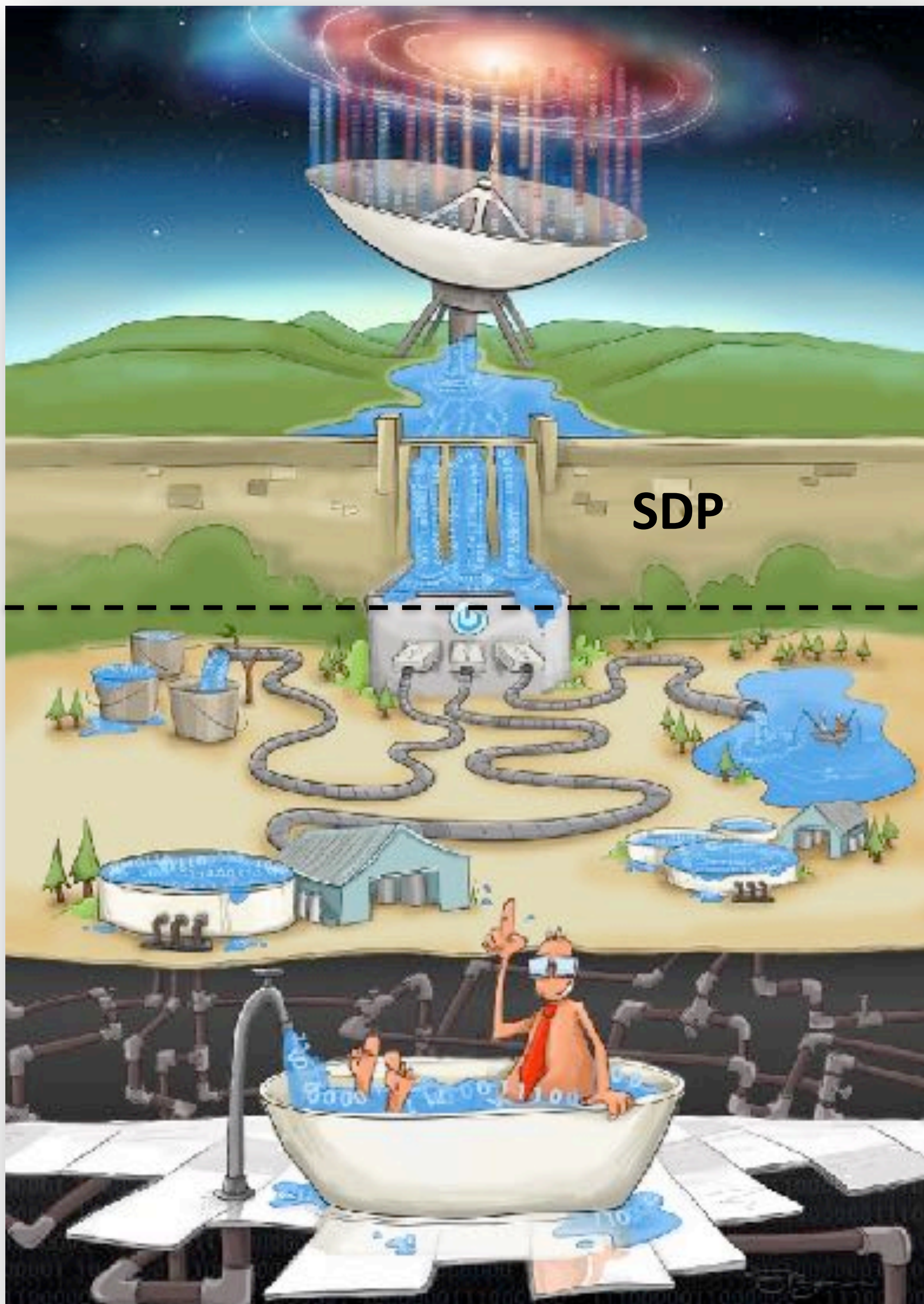
ICRAR International Centre for Radio Astronomy Research

Bird's eye view

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR



SKA = SKAO + SRC



SKA Observatory (SKAO)

Joined SKAO-SRC functions

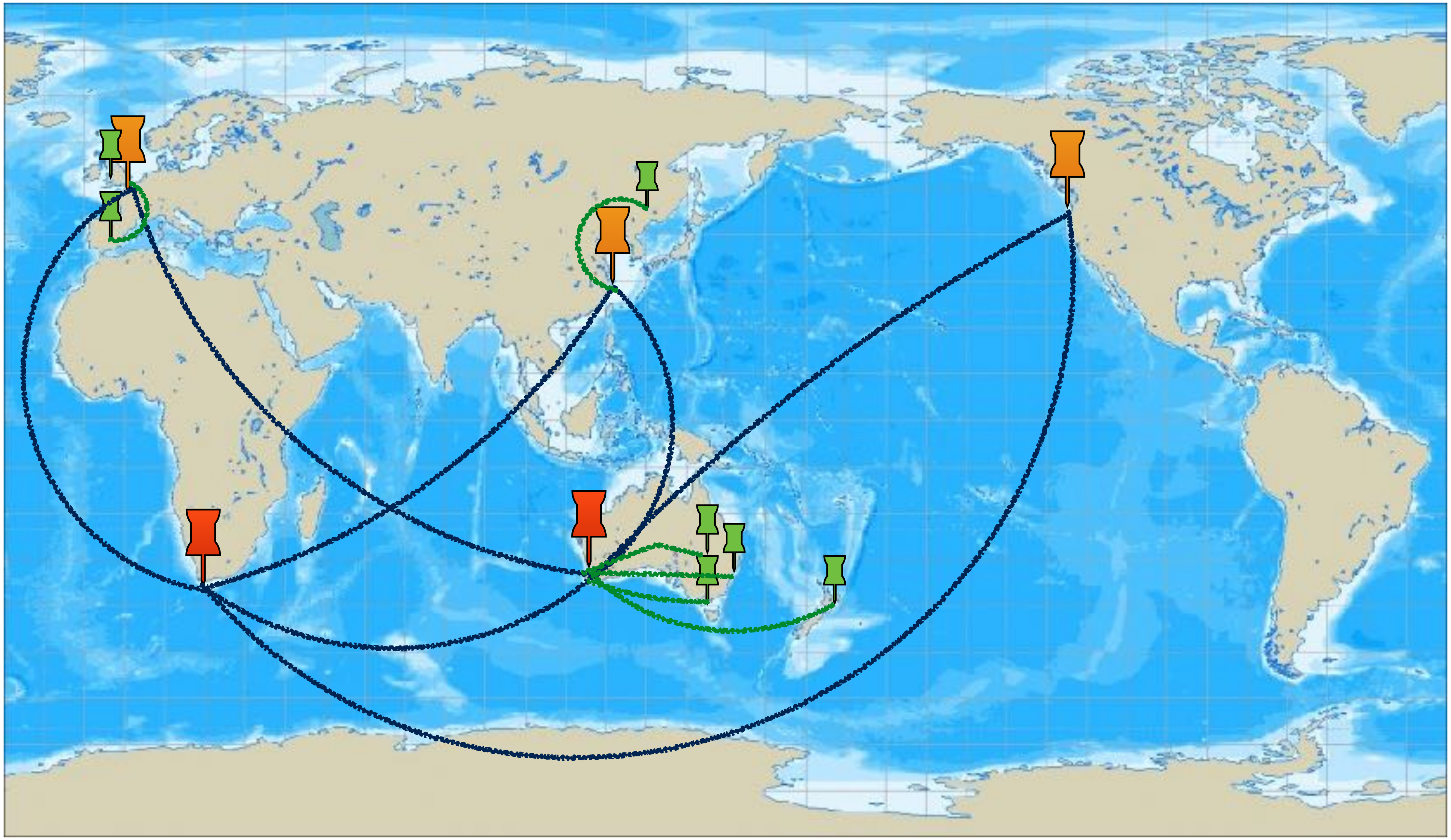
- User support for SKAO data products
- User support for SKAO provided software and tools
- Distribution of SKA data packs to users

SKA Regional Centers (SRC)

Essential SRC functions

- Development and provision of long-term SKA Science Archive
- Provision and management of computational resources for post-processing and analysis
- Provide platforms for continued development of software (pipelines and tools)

Global Network of SRCs



Multiple regional SRCs, locally resourced, heterogeneous in architecture, fully interoperable.

Thank You!