SKA Data Archive Requirements

Peter Hague

University of Cambridge

prh44@cam.ac.uk

SKA Data Archives

- Data products produced by SDP are to be archived on site, and then mirrored in part or in whole (*TBD*) at regional centres

- Limit on size of archive is primarily rate at which it fills rather than cost of storage

- Support multi-wavelength science at regional centres

SKA Data Archives

- MID is the South Africa site
- LOW is the Australia site
- Total archive requirements for both sites expected to be ~555PB over 5 years
- This is a minimum estimate - ideally would like to make full use of 100Gbit/s connection to archives

Table 10: Archive size estimates, power requirements and cost

	A Data rate in to Archive (Gbits / s)	B Growth rate of Archive (TBytes / day)	C Growth rate of Archive (PBytes / year)	D Total archive after 5 years (PBytes)	Actual Power* associated with total size in column D	Storage CAPEX associated with total size in column D
MID (HPSOs)	9	90	34	170	99 KW peak 2 KW ave	3.6 M€
LOW (HPSOs) **Excluding EoR** uv archive	3	30	11	55	32 KW peak 0.7 KW ave	1.2 M€
LOW uv archive	22	220	77	385	225 KW peak 4.7 KW ave	6.6 M€

*Note: Power estimate excludes PUE and power factor.

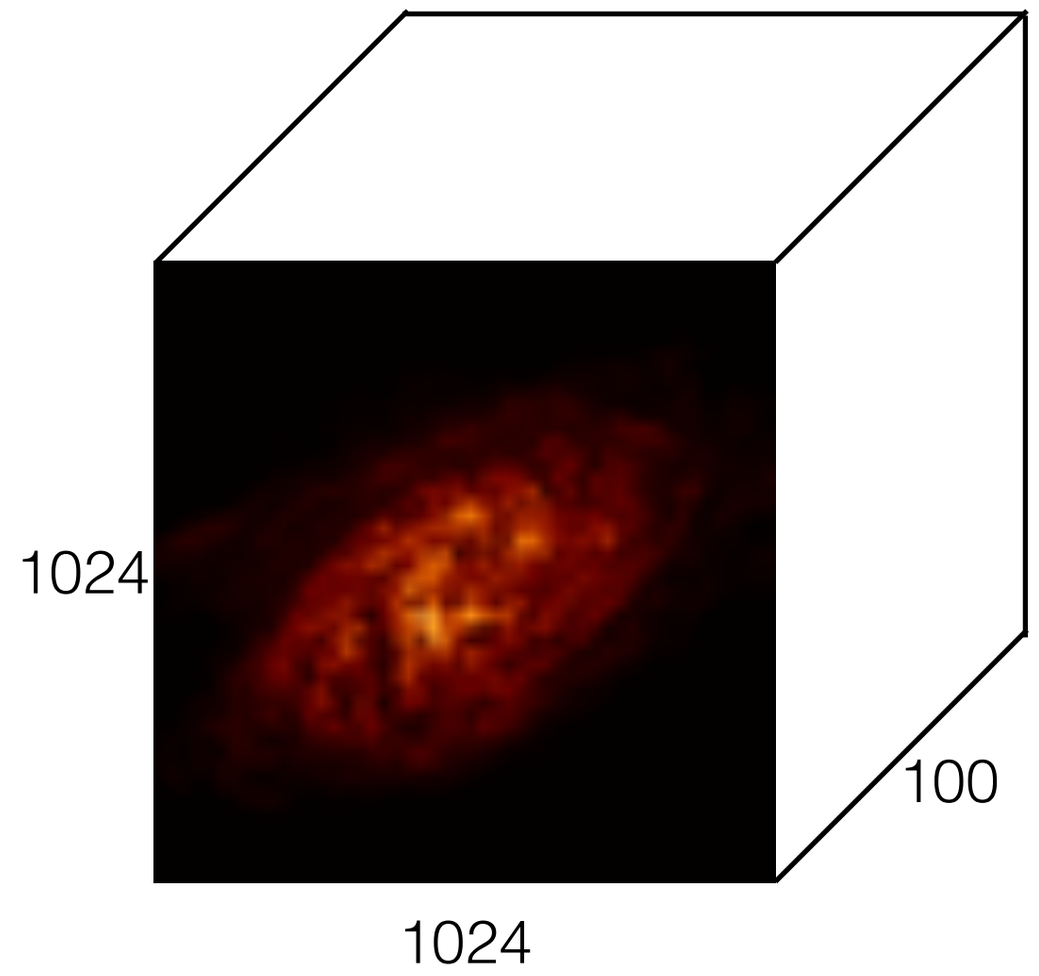
From SKA SDP (Science Data Processor) document number SKA-TEL-SDP-0000038

Worst case scenario

- 50,000x50,000 image with 65,536 channels ~1 petabyte
- 1 such image every 6 hours ~2 exabytes per year
- Most projects will discard either spatial or frequency resolution to some extent

Use case - THINGS

- The HI Nearby Galaxy Survey, using VLA L-band
- 1024x1024 images with 100 frequency channels
- HI flux and velocity maps of ~30 nearby galaxies
- Only interested in 21cm line; such a survey with SKA wouldn't use maximum number of channels



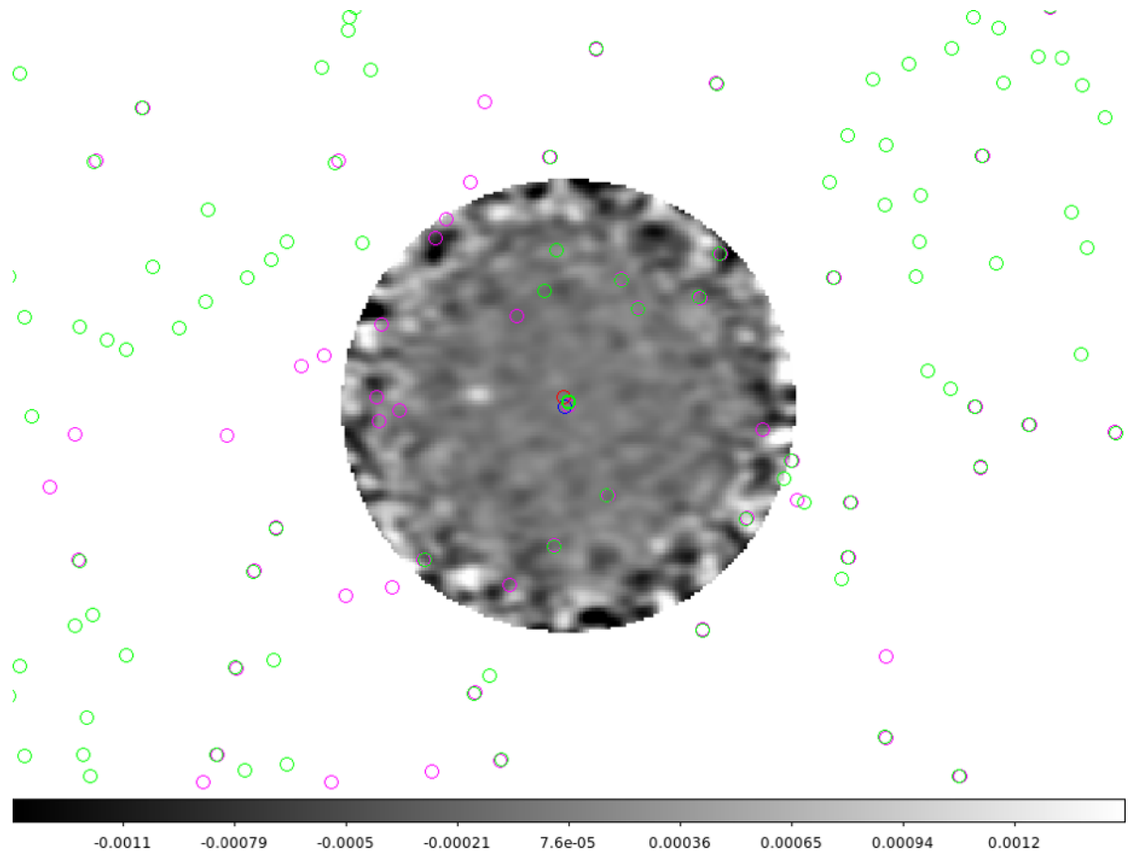
NGC 2403

Current project (ALMA)

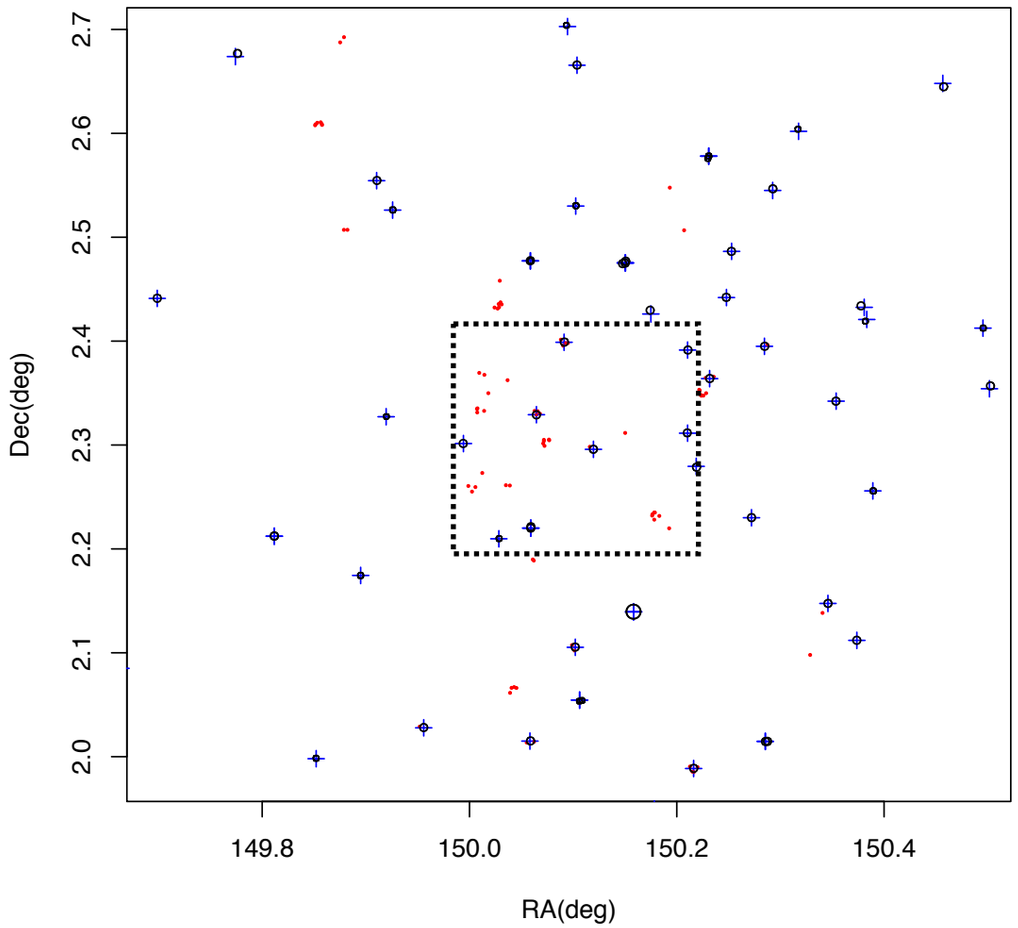
- Source matching between an existing quasar catalogue (Veron 2010) and the ALMA archive

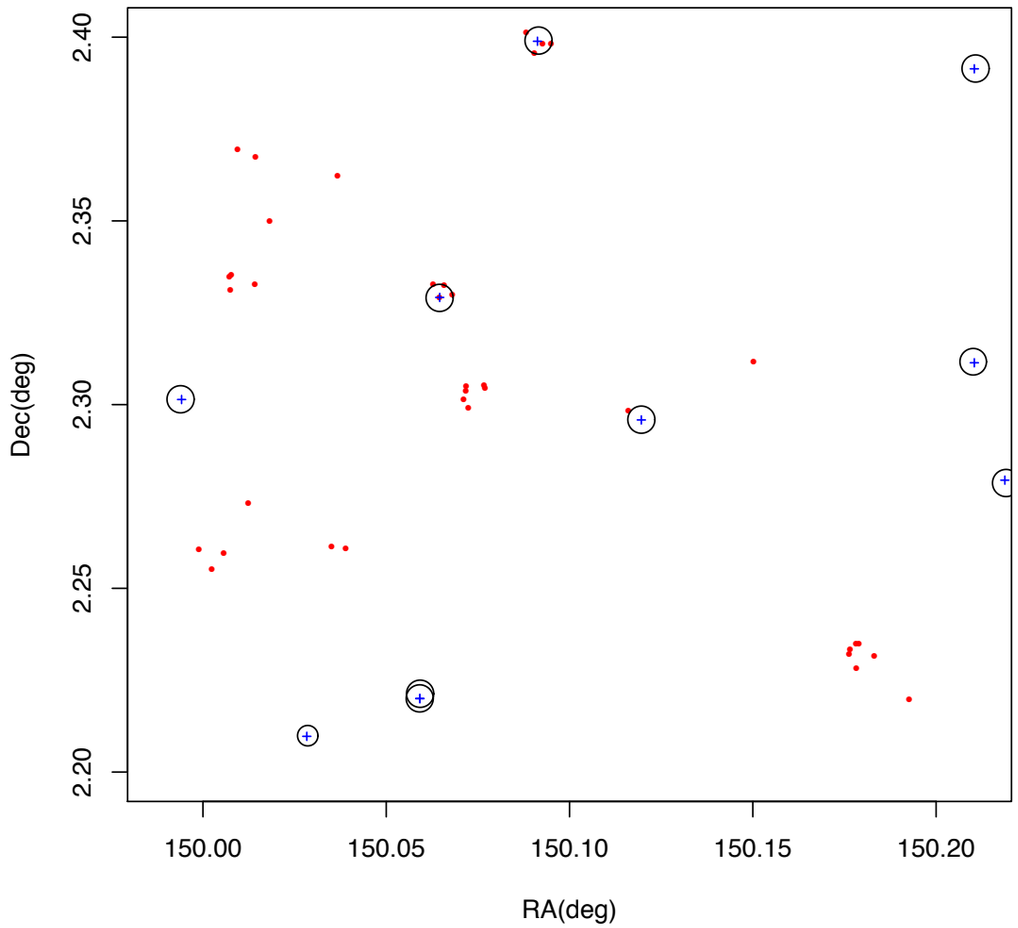
- Application of SExtractor to ~35 terabytes of data products

- Typical image 300x300, currently only investigating single channels



Current project (ALMA)





Red dots - ALMA sources
Blue crosses - Veron (2010) quasars
Circles - HWHM of observational hits

Scaling to SKA

- Takes ~10s per 300x300 channel on a typical machine

- So naive scaling to images that can be produced in SKA would take ~77 hours **per channel**

- Each full image with 2^{16} channels would thus take ~577 years using SExtractor in this manner

Summary

- Not enough to just store the archive - it must be usable
- Geared towards multi-wavelength science
- Size presents more challenges for processing than for storage
- **3.4 D-ANA plans** - Bayesian source extraction