ICRAR

International Centre for Radio Astronomy Research

SQUARE KILOMETRE ARRAY

SDP

SCIENCE DATA PROCESSOR

SDP Design for Cloudy Regions

Markus Dolensky

C4SKA, 11/02/2016

Curtin University

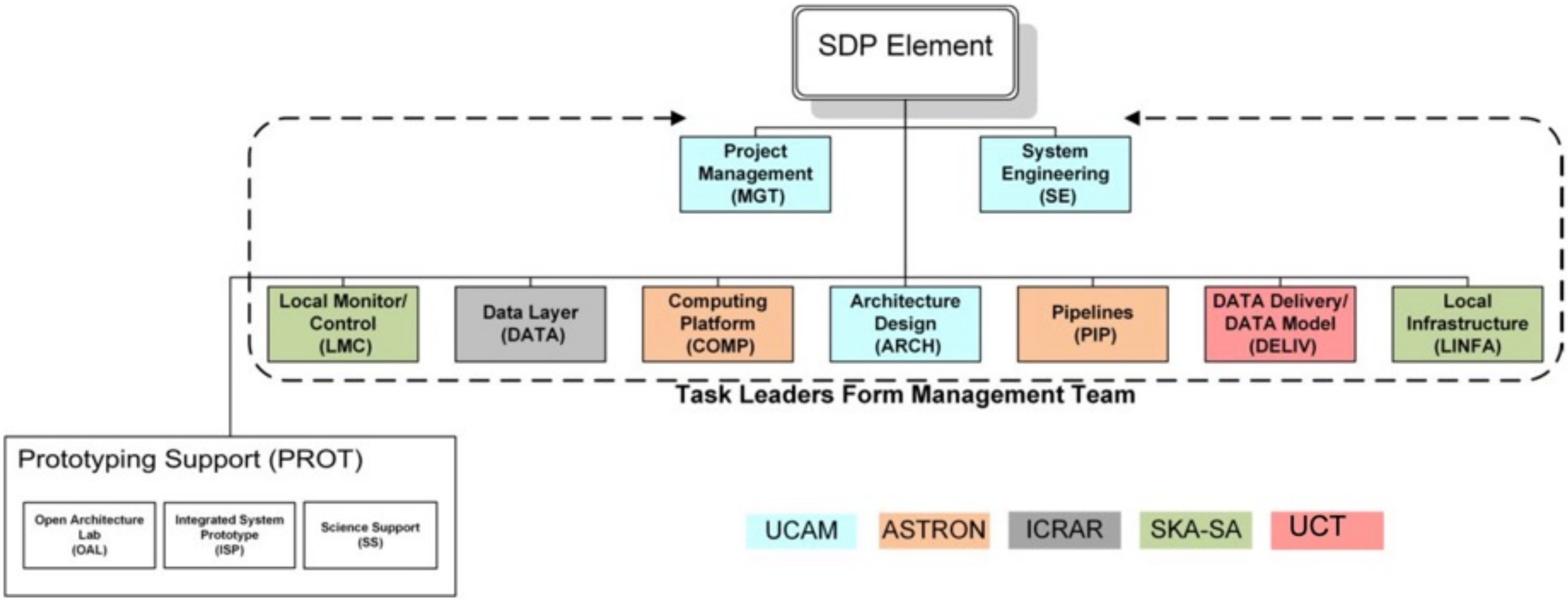
The Government of Western Australia

The University of Western Australia

ICRAR's Data Intensive Astronomy Group

generously borrowed content from above colleagues

SDP Subelements



- Lead: Paul Alexander
- PM: Jeremy Coles
- Deputy PM: Ian Cooper
- PE/Architect: Bojan Nikolic
- SE: Ferdl Graser
- PS: Rosie Bolton

- COMP: Chris Broekema
- PIP: Ronald Nijboer
- DATA: Andreas Wicenec
- DELIV: Rob Simmonds
- LMC: Shagita Gounden
- LINFA: Jasper Horrell

Characteristics after Rebaselining

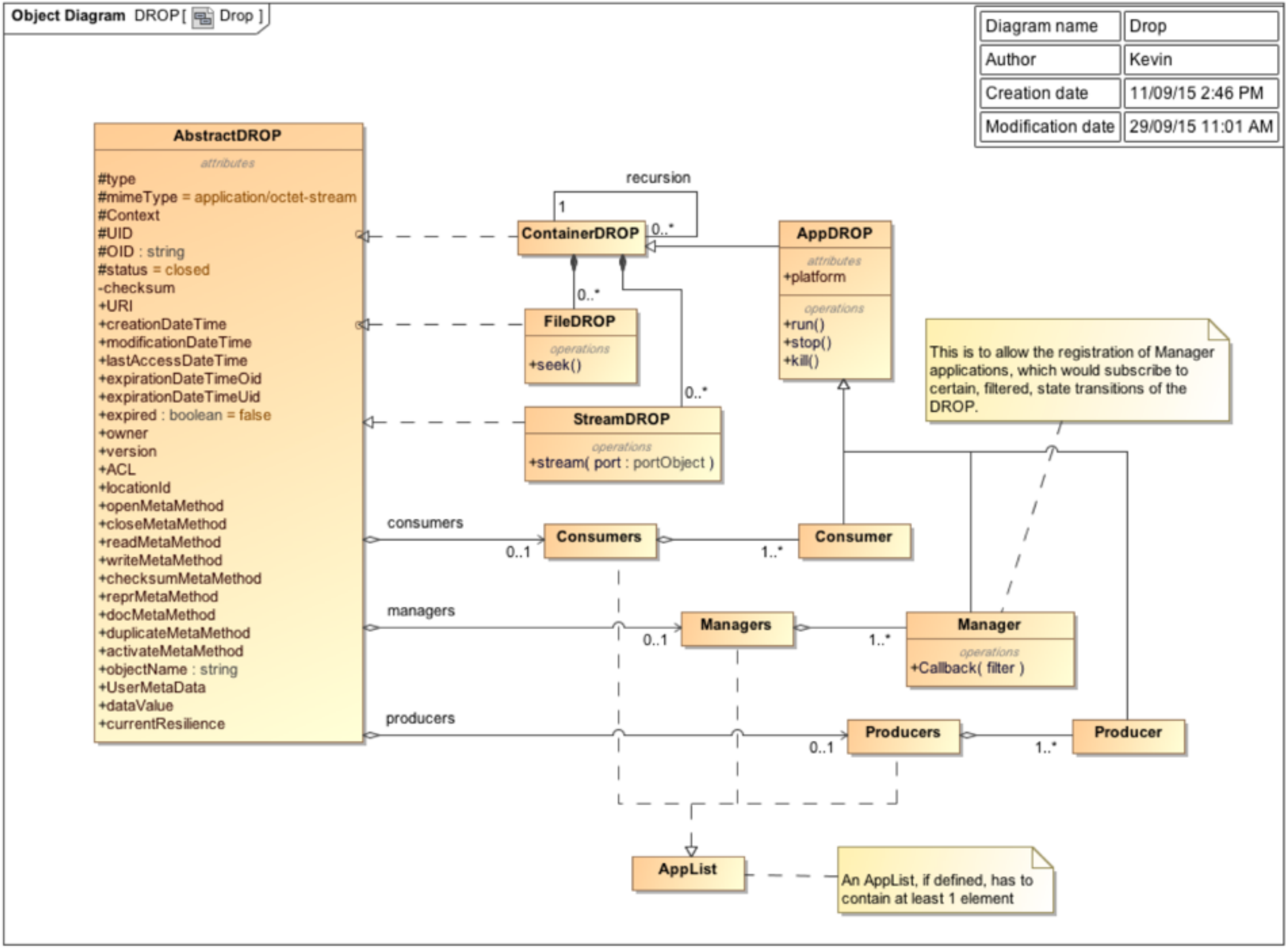
Telescope	SKA1_Low	SKA1_Mid
Antennae / Dishes	130000	200
max. Baseline [km]	65	150
Frequency channels	65,536	65,536
Complex Correlations / s	3.8E+10	6.4E+10
Image side length [pix]	16000	20000

Challenges

- Power Budget: ~5 MW
- Compute Efficiency: target is 25 %
- Sustain Throughput: TByte/s
- Optimize Data Locality
- Error Resilience
- Automated Calibration
- Multiplicity of Input Streams
- Variety of Observing Modes
- ...

Key Architectural Concepts

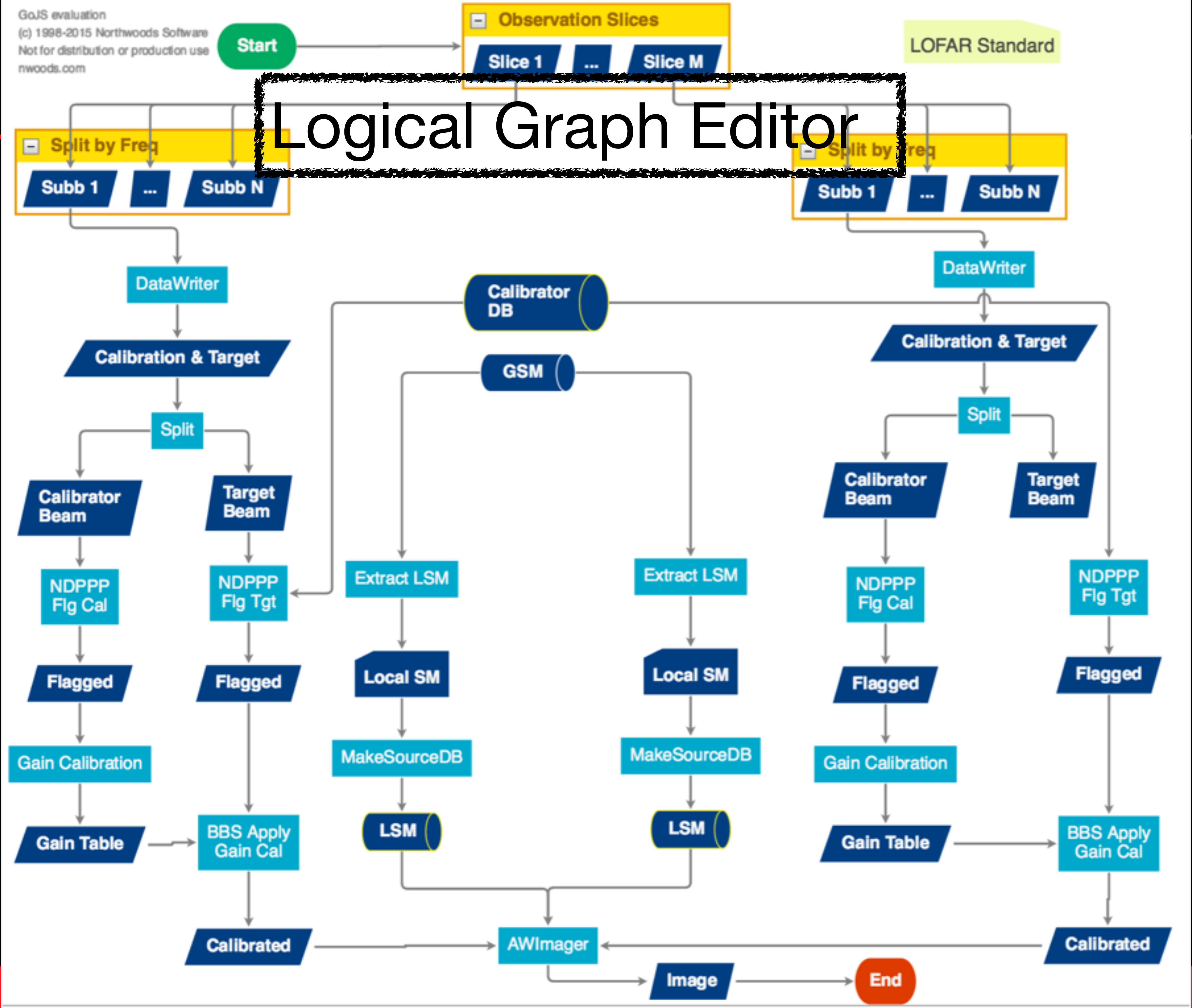
- Data-Driven/Centric
  - data triggers processing
  - focus on data locality
  - exploitation of data parallelism
- Drop
  - an atom of the data flow management systems
  - support of data centric processing


Object Diagram DROP[Drop]
Diagram name
Drop
Author
Kevin
Creation date
11/09/15 2:46 PM
Modification date
29/09/15 11:01 AM
AbstractDROP
attributes
#type
#mimeType = application/octet-stream
#Context
#UID
#OID : string
#status = closed
-checksum
+URI
+creationDateTime
+modificationDateTime
+lastAccessDateTime
+expirationDateTimeOid
+expirationDateTimeUid
+expired : boolean = false
+owner
+version
+ACL
+locationId
+openMetaMethod
+closeMetaMethod
+readMetaMethod
+writeMetaMethod
+checksumMetaMethod
+reprMetaMethod
+docMetaMethod
+duplicateMetaMethod
+activateMetaMethod
+objectName : string
+UserMetaData
+dataValue
+currentResilience
recursion
1
0..*
ContainerDROP
AppDROP
attributes
+platform
operations
+run()
+stop()
+kill()
0..*
FileDROP
operations
+seek()
0..*
StreamDROP
operations
+stream(port : portObject)
This is to allow the registration of Manager applications, which would subscribe to certain, filtered, state transitions of the DROP.
consumers
0..1
Consumers
1..*
Consumer
managers
0..1
Managers
1..*
Manager
operations
+Callback(filter)
producers
0..1
Producers
1..*
Producer
AppList
An AppList, if defined, has to contain at least 1 element

Drops in Pipeline Context

- A pipeline is described in DROP world by a Directed Acyclic Graph (DAG).
- The nodes of such a graph are alternating DataDROPs and ApplicationDROPs. The edges are events.
- We distinguish between *logical* and *physical* graphs.
- Logical graphs contain the pipeline model or template.
- Physical graphs are a mapping of logical graphs onto actually available and suitable hardware.
- The mapping is the real hard bit!!
- The ApplicationDROPs are pipeline components. Essentially wrappers around existing algorithms (e.g. CASA tasks).
- In general these wrappers are implemented as Docker containers.

Logical Graph Editor

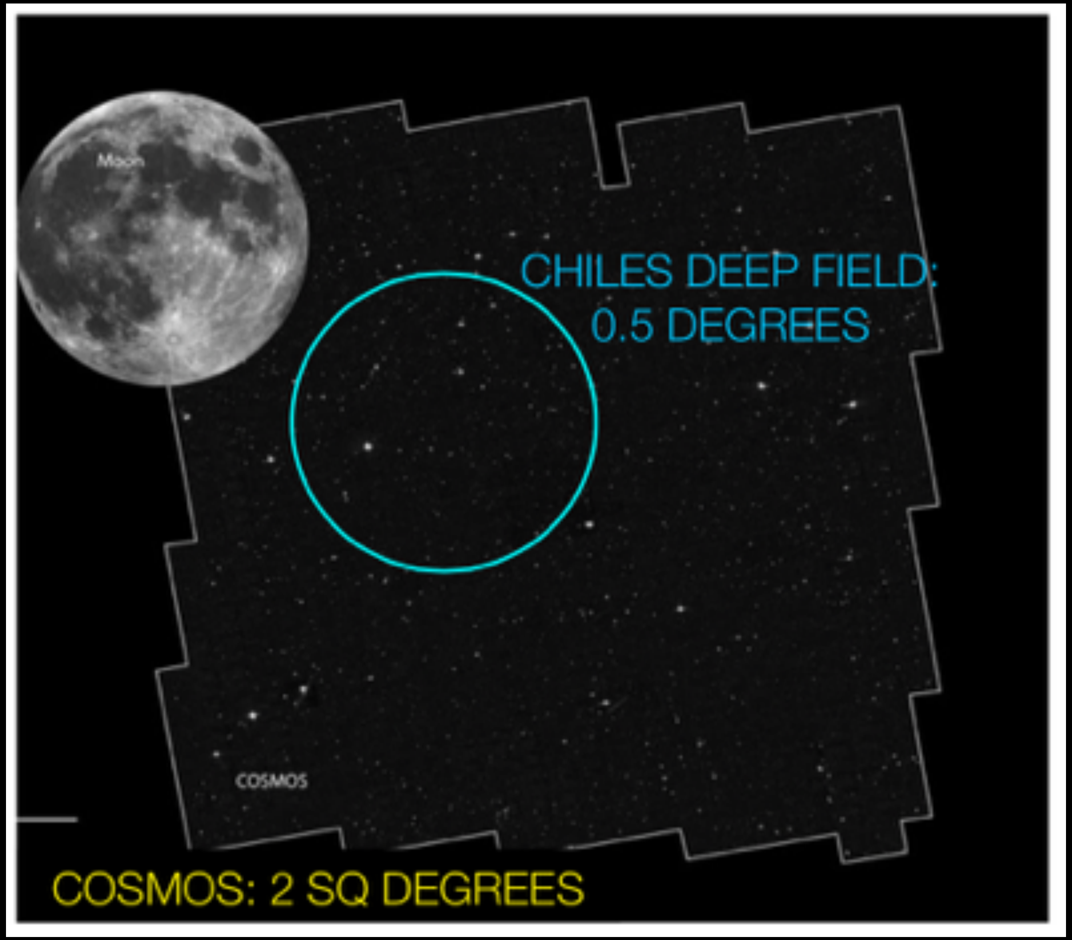


Drop Summary

- Drop concept is a consequence of Data Driven paradigm
- helps pushing S/W architectural design and verification
- current, quite simple prototype allows modelling of real-world pipelines and process JVLA and LOFAR test sets
- prototyping also serves technology evaluation (e.g. Luigi)
- Drop fits well with Object Storage (e.g. Ceph and S3)

CHILES

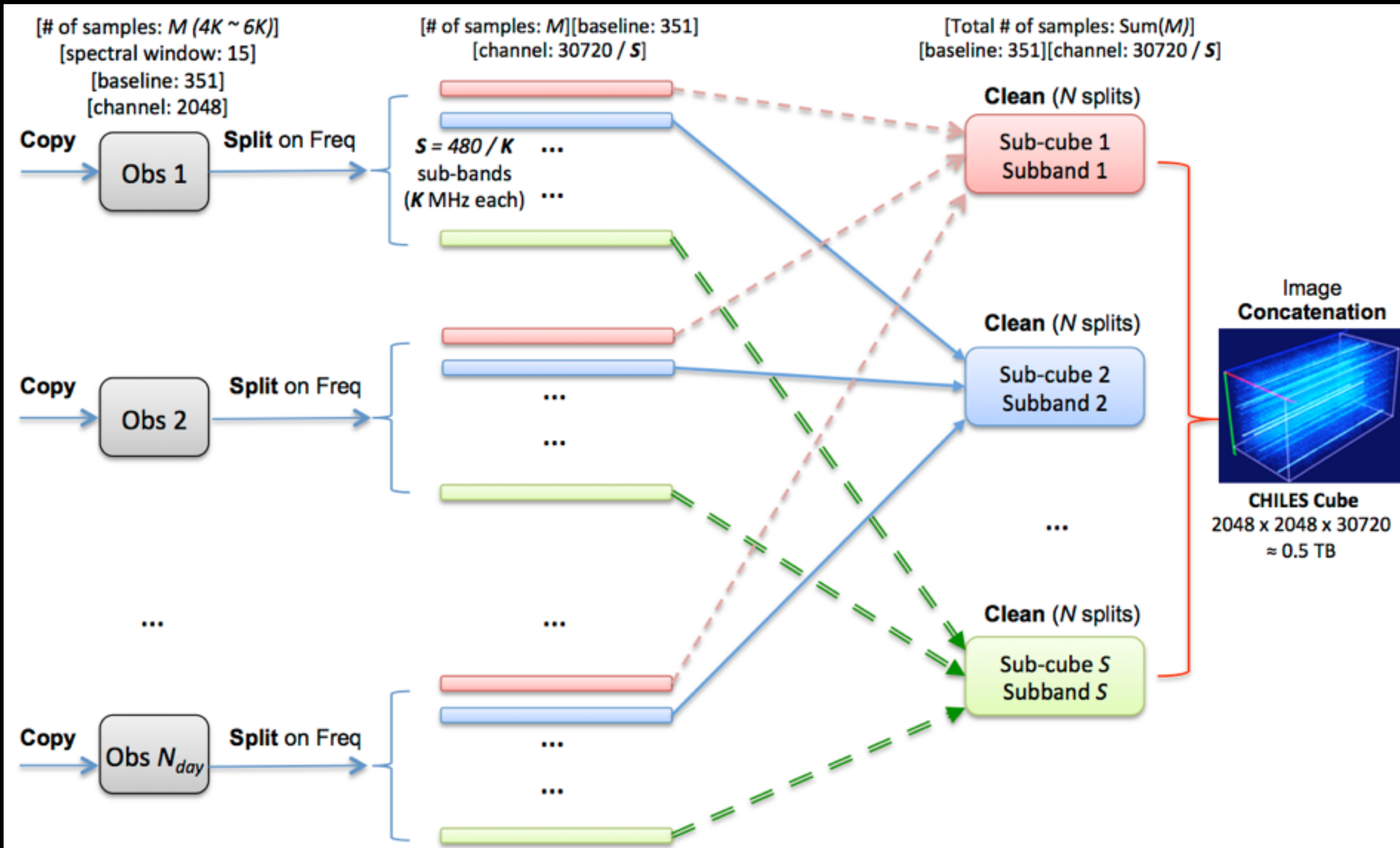
- 1000 hours, single pointing in COSMOS field
- VLA in B-configuration
- freq coverage: ~950 to 1450 MHz (z=0 to z=0.5)
- 30,720 channels (3.5 km/s at z=0)



Computing efforts

Conventional Cluster (pleiades) 5 nodes each node has 2x Intel Xeon X5650 2.66GHz CPUs (6 cores / 12 HTs)	Enough computing power, however it would take weeks
Super computer (MAGNUS) Cray XC40 - 24 cores per node 2.6GHz Intel Xeon E5-2690V3 64GB per Node 35,712 cores available 3PB of storage #58 in the world	
AWS Whatever we wanted r3.xlarge 16 cores 122GB Ram	

Workflow


[# of samples: M (4K ~ 6K)]
[spectral window: 15]
[baseline: 351]
[channel: 2048]
[# of samples: M][baseline: 351]
[channel: 30720 / S]
[Total # of samples: Sum(M)]
[baseline: 351][channel: 30720 / S]
Copy
Obs 1
Split on Freq
S = 480 / K
sub-bands
(K MHz each)
Copy
Obs 2
Split on Freq
Copy
Obs N_day
Split on Freq
Clean (N splits)
Sub-cube 1
Subband 1
Clean (N splits)
Sub-cube 2
Subband 2
Clean (N splits)
Sub-cube S
Subband S
Image
Concatenation
CHILES Cube
2048 x 2048 x 30720
≈ 0.5 TB

Data Reduction

- Exclusively done using CasaPy
- Had to work around limitations of CasaPy
- Trialled 3 ways of splitting out the data
  - SPLIT
  - CVEL
  - MSTRANSFORM
- Tried to keep a common code base in GitHub
  - PBS - Pleiades
  - SLURM - Magnus
  - Python/Boto - AWS

About CasaPy

- CasaPy would not allow parallel access to a large Measurement Set
- CasaPy did not like the Gluster file system on Pleiades. Happy with ext4 or lustre (Magnus).
- With the limited frequency ranges we were using per slice, the noise levels per channel were very sensitive to the weighting scheme

Results - I/O

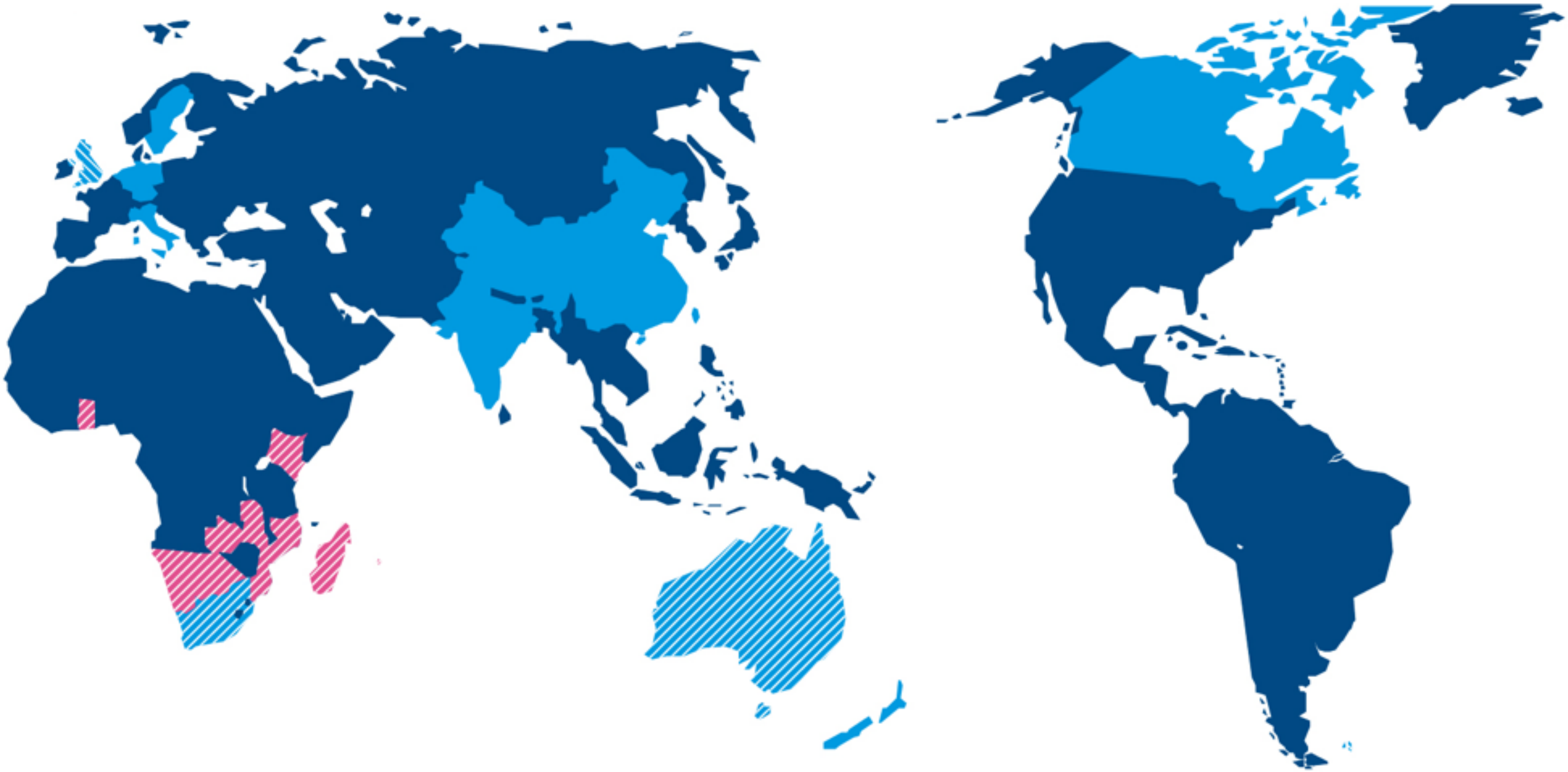
Operation	Platform	Peak Memory	I/O Throughput	CPU Usage	I/O Characteri
SPLIT	AWS (EBS)	420MB	<10MB/s	0.4	Sequential read/write dominate
	Magnus	545MB	40 ~ 100 MB/s	1	
	Pleiades	390MB	60 ~ 100 MB/s	1	
INVERT	AWS (SSD)	60GB	70 ~ 500MB/s	4	Random writes and sequential reads dominate
	Magnus	30GB	50 ~ 400MB/s	1	
	Pleiades	35GB	50 ~ 400MB/s	4	

Results

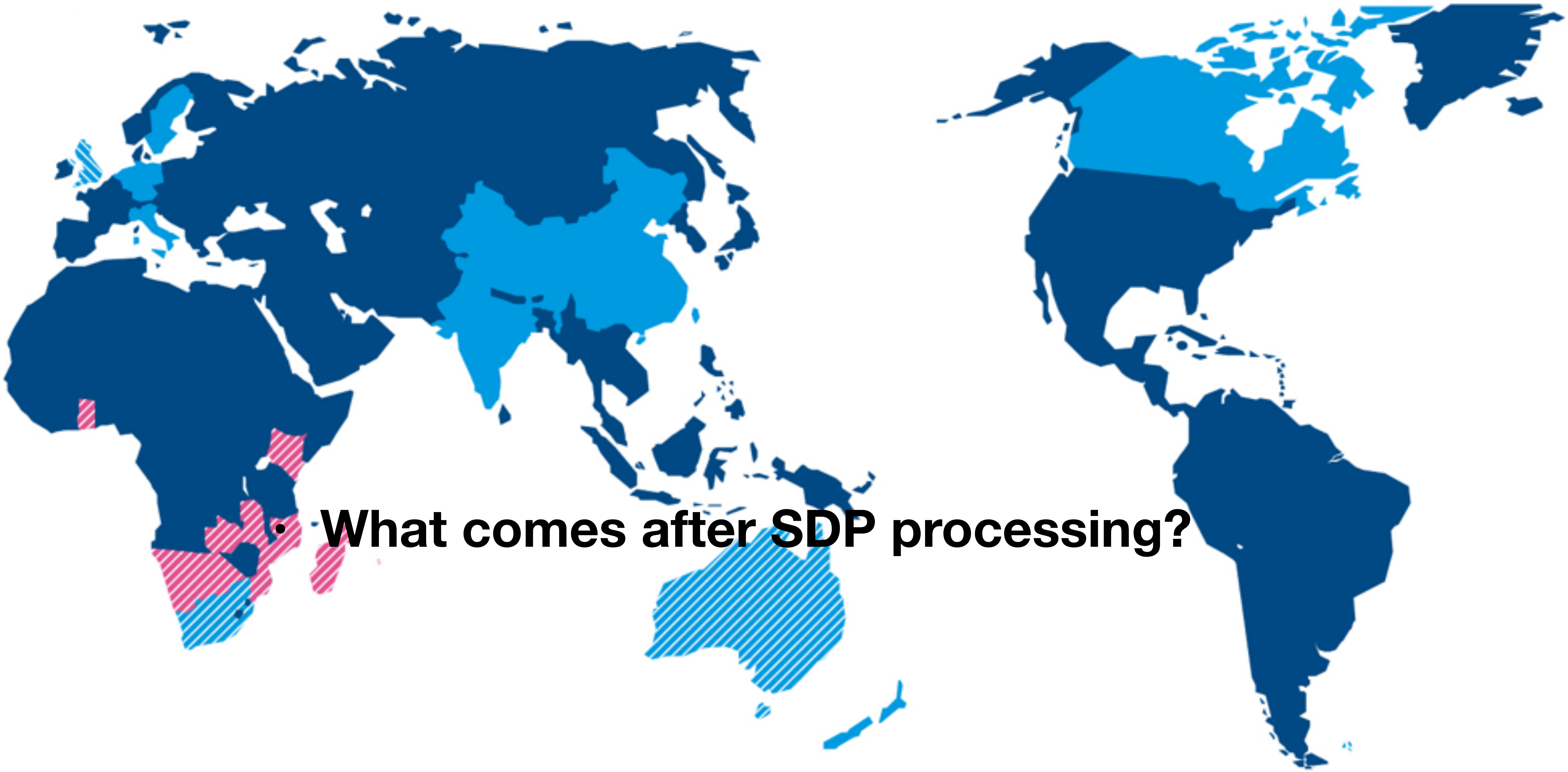
	AWS	Magnus (HPC)	Pleiades
Completion Time	96hr	110hr	1,060hr (est)
Capital Costs	AUD$0	AUD$12,000,000	AUD$50,000
Operational Costs	AUD$2,000	AUD$3,240 (free)	-
Control	Root	Limited	Root
Usability	Complex	Good	Good

Lessons

- In-house Cluster (Pleiades)
  - not very satisfactory
- HPC (Pawsey Centre)
  - very fast
  - no root access
  - additional software is installed by admin
  - in WA it is effectively free
- Cloud (AWS)
  - you can do what you like (a good and a bad thing)
  - EBS volumes are slow
  - directly attached SSDs are fast
  - billing based on usage



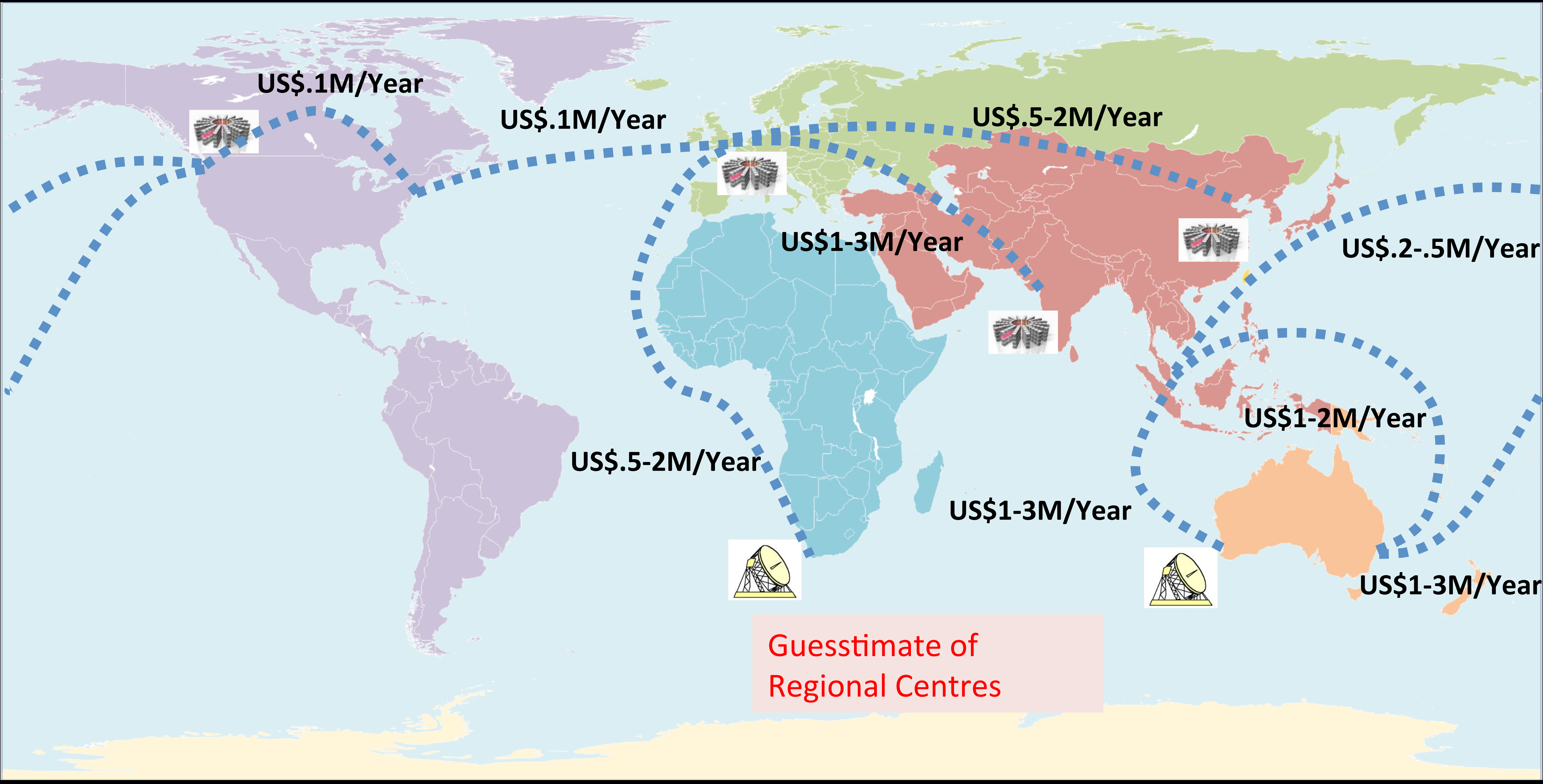
Regional Networking



- **What comes after SDP processing?**
- **How does community get access and maximize scientific return?**

Estimated SDP to World Costs

- 10 year IRU per 100 Gbit circuit 2020-2030
- Guesstimate of Regional Centre locations



Existing Regional Networks

- LHC
  - Tier 0: CERN
  - Tier 1: large computing centres
  - Tier 2: analysis centres
- ALMA
  - regional centres
  - regional centre nodes
- EUMETSAT
  - national meteorological bureaus
  - regional (implementation) centres

Scoping

- technical support for researchers
  retrieval, analysis, visualization
- post & re-processing; software and middleware stacks
- storage/backup: of data products and derived products
- regional outreach

Some Technical Considerations

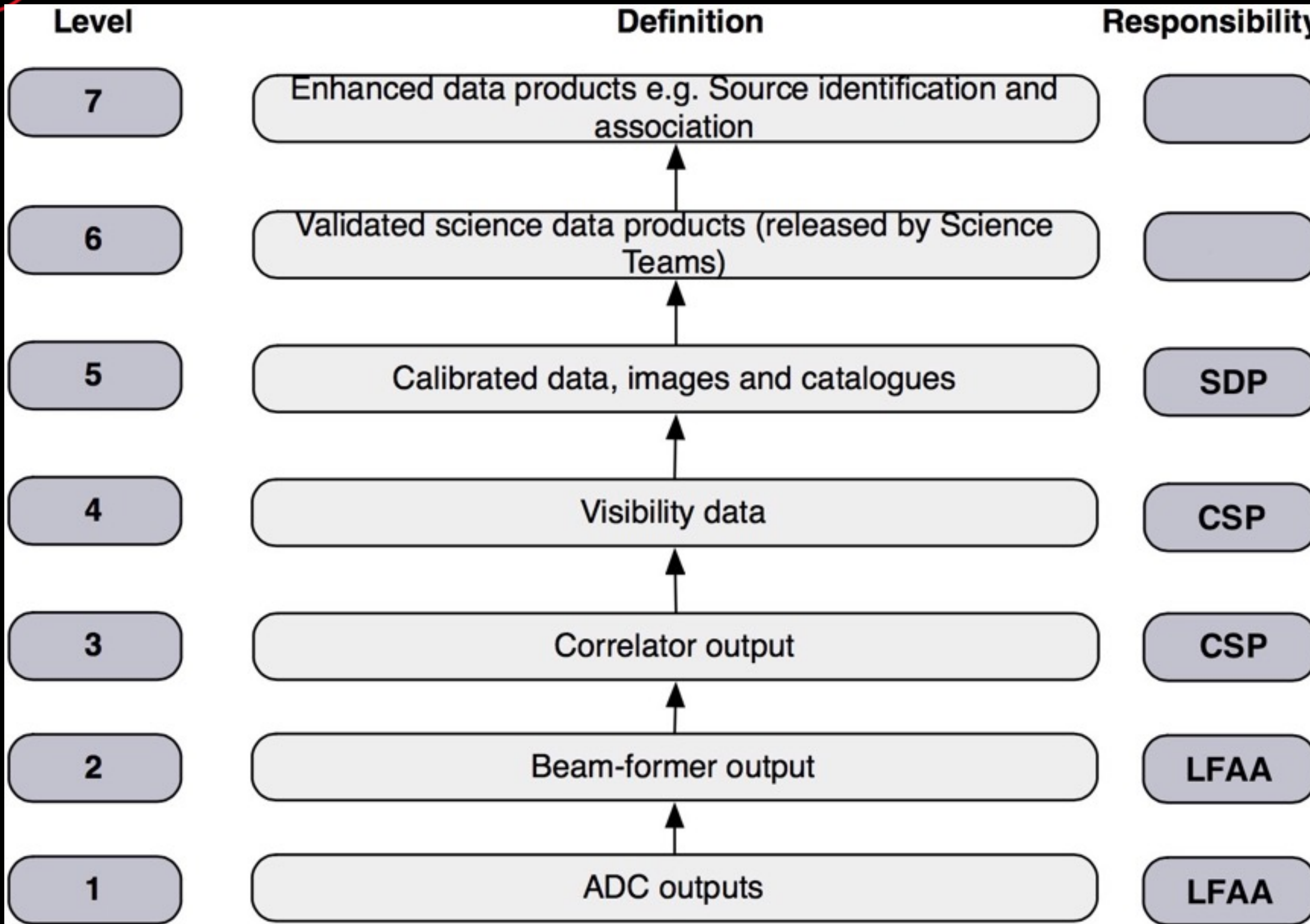
- use of common software tools
- subsetting data
- minimizing data movement
  => requests served by RC instead of SKA site
  => (post)processing moved across regions to the data

Some Technical Considerations

- use of common software tools
- subsetting data
- minimizing data movement
  => requests served by RC instead of SKA site
  => (post)processing moved across regions to the data

=> all of above requires agreements

Data Product Levels



Data Product Levels

Level	Definition	Responsibility
7	Enhanced data products e.g. Source identification and association	
6	Validated science data products (released by Science Teams)	
5	Calibrated data, images and catalogues	SDP
4	Visibility data	CSP
3	Correlator output	CSP
2	Beam-former output	LFAA
1	ADC outputs	LFAA

SDP External Interface

- analyzing requirements/gaps in
  - IVOA support; data discovery, access, preservation, characterization, … VOEvent/Timeseries, SIA, TAP, ObsCore DM, Datalink, …
  - how to enable post processing off-site

- data product types considered (data management perspective)
  - continuum model image
  - spectral line cube image (absorption and emission)
  - sensitivity image
  - representative PSF image
  - moment images for multi-frequency synthesis
  - corresponding residual images (if deconvolved)
  - sensitivity image
  - source catalogue
  - pipeline logs and quality assessment logs

SDP Milestones lying ahead

Milestone	Date
delta-PDR	Q1/2016
Design Maturity Review	Q4/2016
CDR Submission	Q4/2017