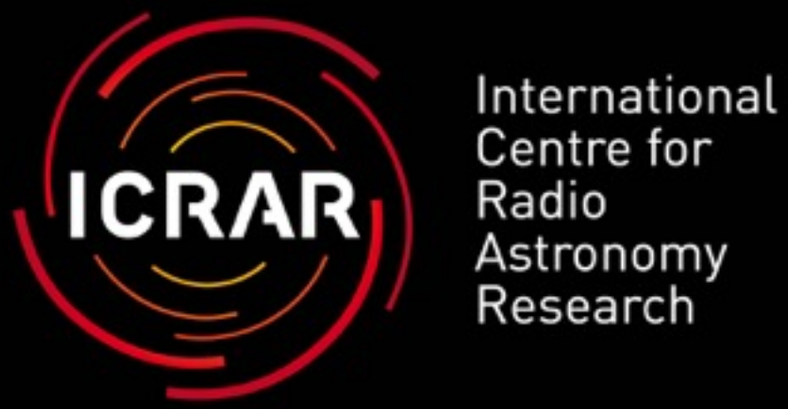
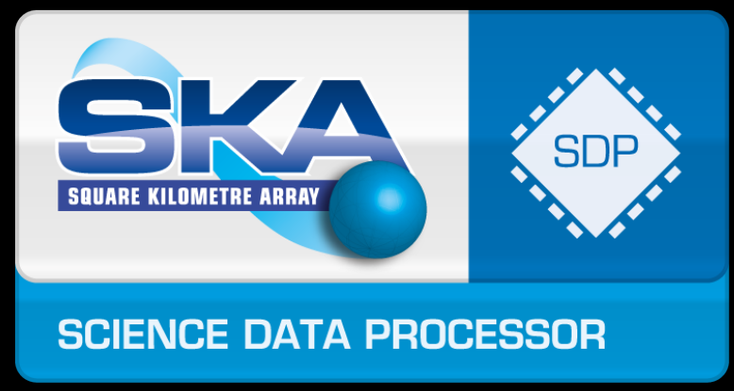


International
Centre for
Radio
Astronomy
Research



SDP Data Management Challenges

Andreas Wicenec

on behalf of

the Data Intensive Astronomy team at ICRAR

&

others

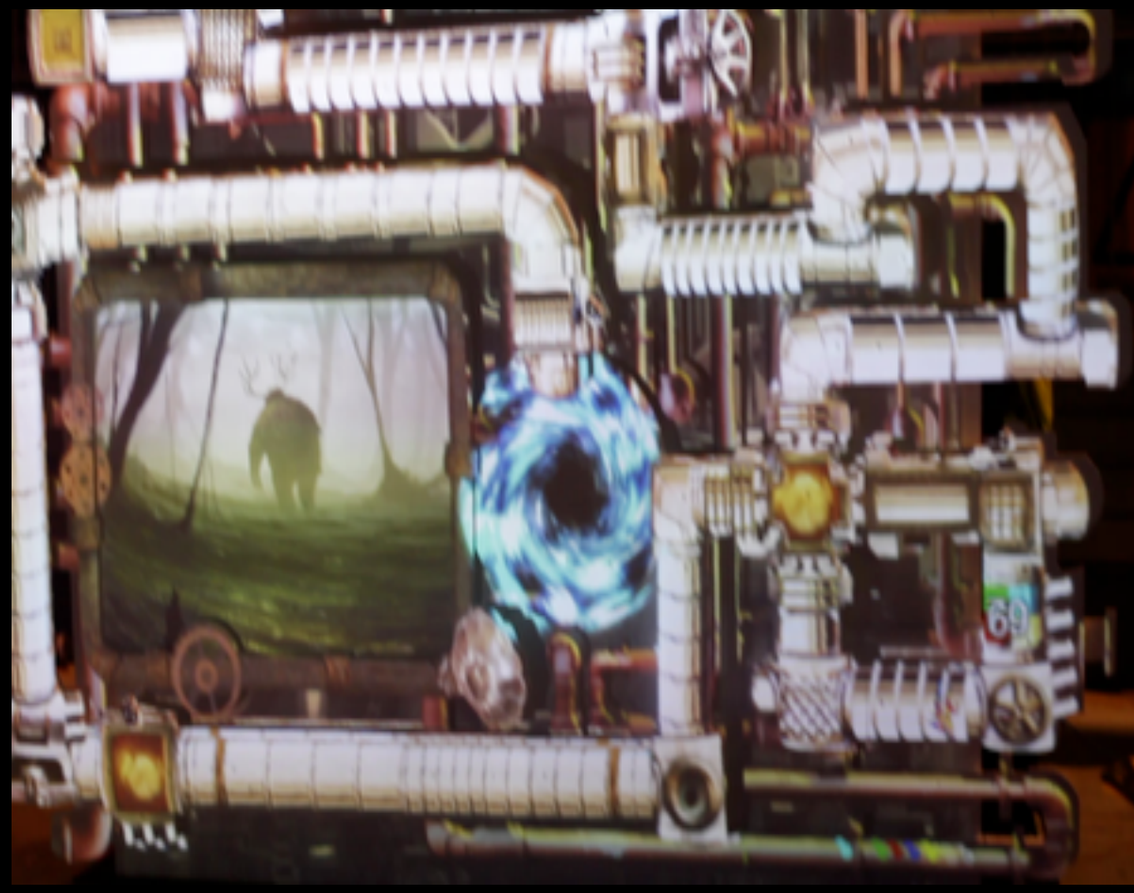
Context

Boundary Conditions

- Ingest rates ~ 400 GB/s per site
- Multiple concurrent observing projects
- Data sharing between projects
- Capital and operational budget limited
  - Power, Cooling
  - Acquisition, maintenance & software development costs
- Data parallelism: Millions of related tasks on thousands of nodes.
- Throughput: produce ~0.2-10 Tera Voxels/second
- Automatic 23/7 type of operation

Pipeline processing

- New ones are ‘invented’ for almost every single project.
- Very often based on scripting together modules from CASA, Miriad and AIPS as well as homegrown modules into an unmaintainable monster, that only a few people understand.
- Very dynamic and project specific. One of those monsters does not work for the next project.



We are at the limit!

★ SKA, ASKAP, MWA, LOFAR, MeerKAT

- are producing very high data volumes at very high rates
- are a challenge for currently available compute infrastructures (at least at affordable costs)

★that means

- just throwing more hardware at the problem won’t do the trick anymore.
- we need to use existing hardware more efficiently.
- at least the SKA requires significant innovation in order to approach the science potential of the arrays.

Proposal & Prototype:

… a tiny bit of innovation

ICRAR International Centre for Radio Astronomy Research

SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR

DALiuGE!

Data Activated Graph flow Engine

...think about deluge!

Separation of Concerns

Separation of concerns

★ Let software engineers think about and write software:

- Optimised code using the most appropriate language
- Novel ways of using latest hardware
- Using modern I/O techniques.
- Using advanced DB technologies
- Parallel code (even only a few software engineers can do this well!)
- HPC coding and optimisation (even less people can do this well!)
- ….

Separation of concerns

★ Let astronomers think about and do astronomy:

- Astronomical algorithms
- Pipeline logic
- Novel ways of extracting science
- New science
- Interpretation of extracted information
- Training of AI methods
- ….

Astronomer

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

HPC Software Engineer

Parallel Execution
what is executed where

Parallel Coding
writing parallel code

Code Optimisation
optimise parallel code

OS level S/W Engineer

I/O Optimisation
optimise I/O on hardware

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

Computer H/W Engineer

Logical Graph Template

Algorithms
what's the best algorithm to get the desired answer

Pipeline Logic
reduction components and sequence

Component Parameters
default parameter values of components

Logical Graph

Data Parallelisation
hints about the potential of parallelism

Physical Graph Template

Parallel Execution
what is executed where

Parallel Coding
writing parallel code

Physical Graph

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

Algorithms
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

Astronomer

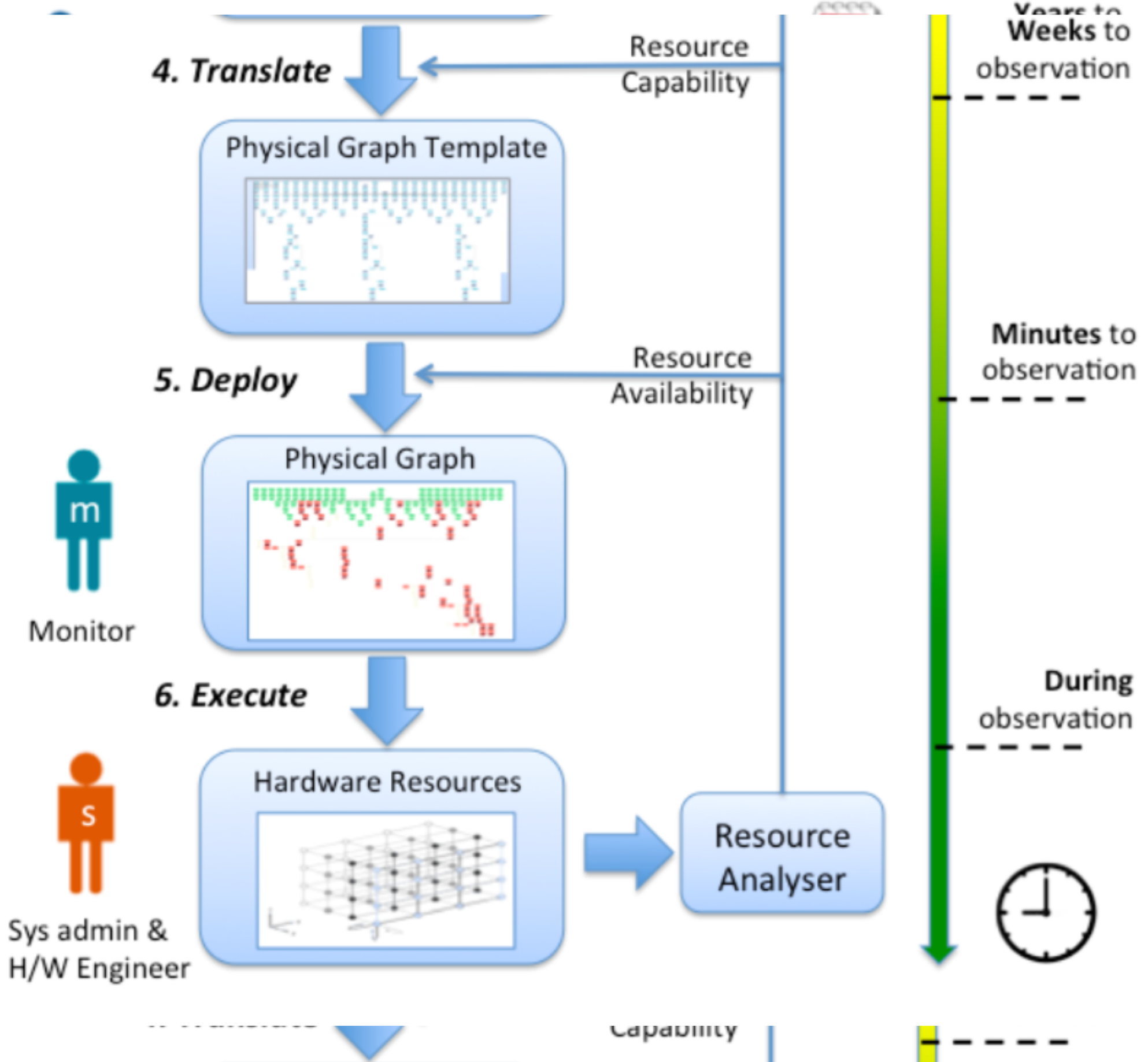
Operator

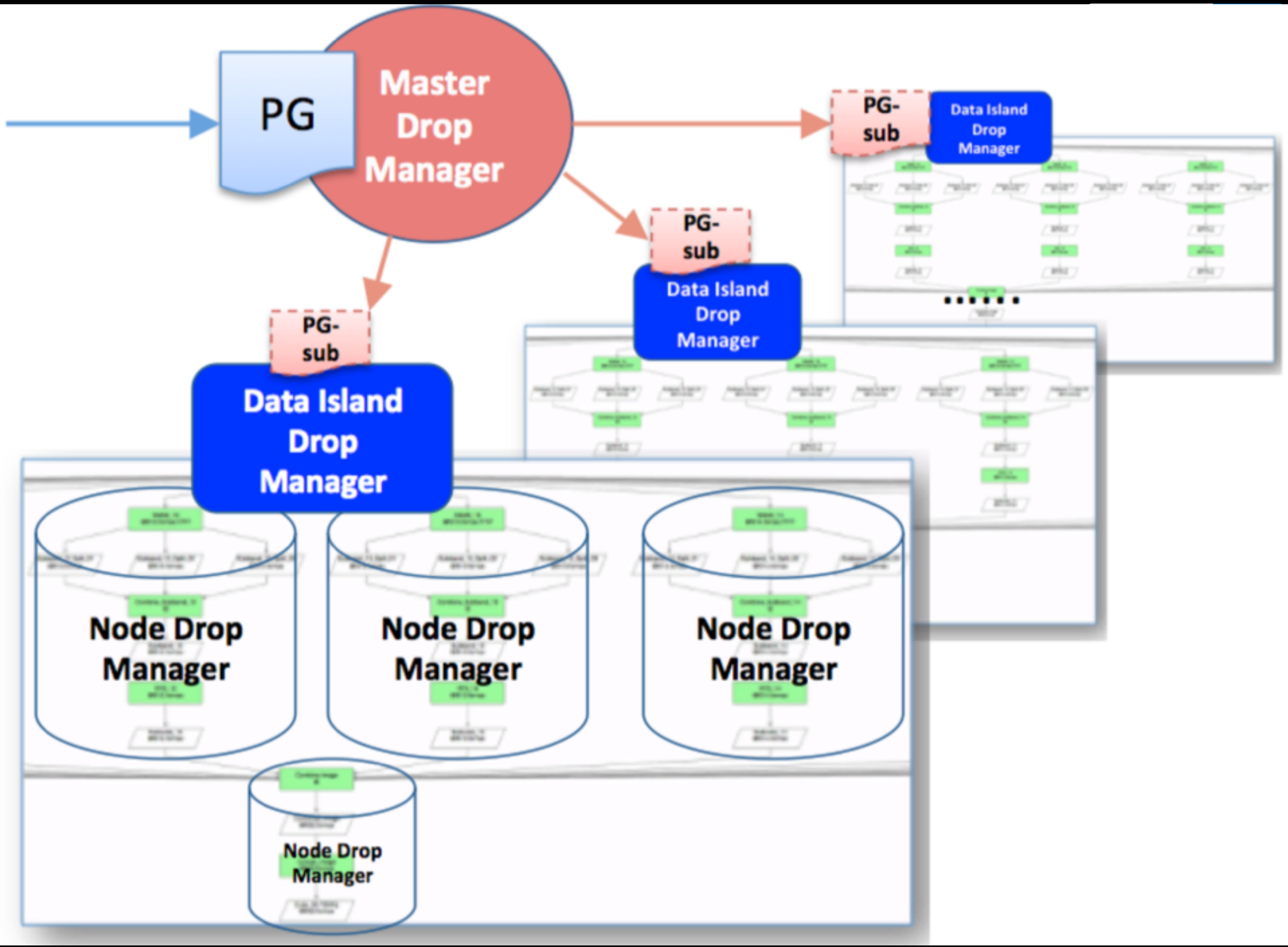
HPC Software Engineer

OS level S/W Engineer

Computer H/W Engineer

Development, Deployment & Asynchronous Execution


4. Translate
Resource Capability
Physical Graph Template
5. Deploy
Resource Availability
Physical Graph
m
Monitor
6. Execute
Hardware Resources
s
Sys admin & H/W Engineer
Resource Analyser
Weeks to observation
Minutes to observation
During observation


PG
Master Drop Manager
PG-sub
Data Island Drop Manager
PG-sub
Data Island Drop Manager
PG-sub
Data Island Drop Manager
Node Drop Manager
Node Drop Manager
Node Drop Manager
Node Drop Manager

Execution (1)

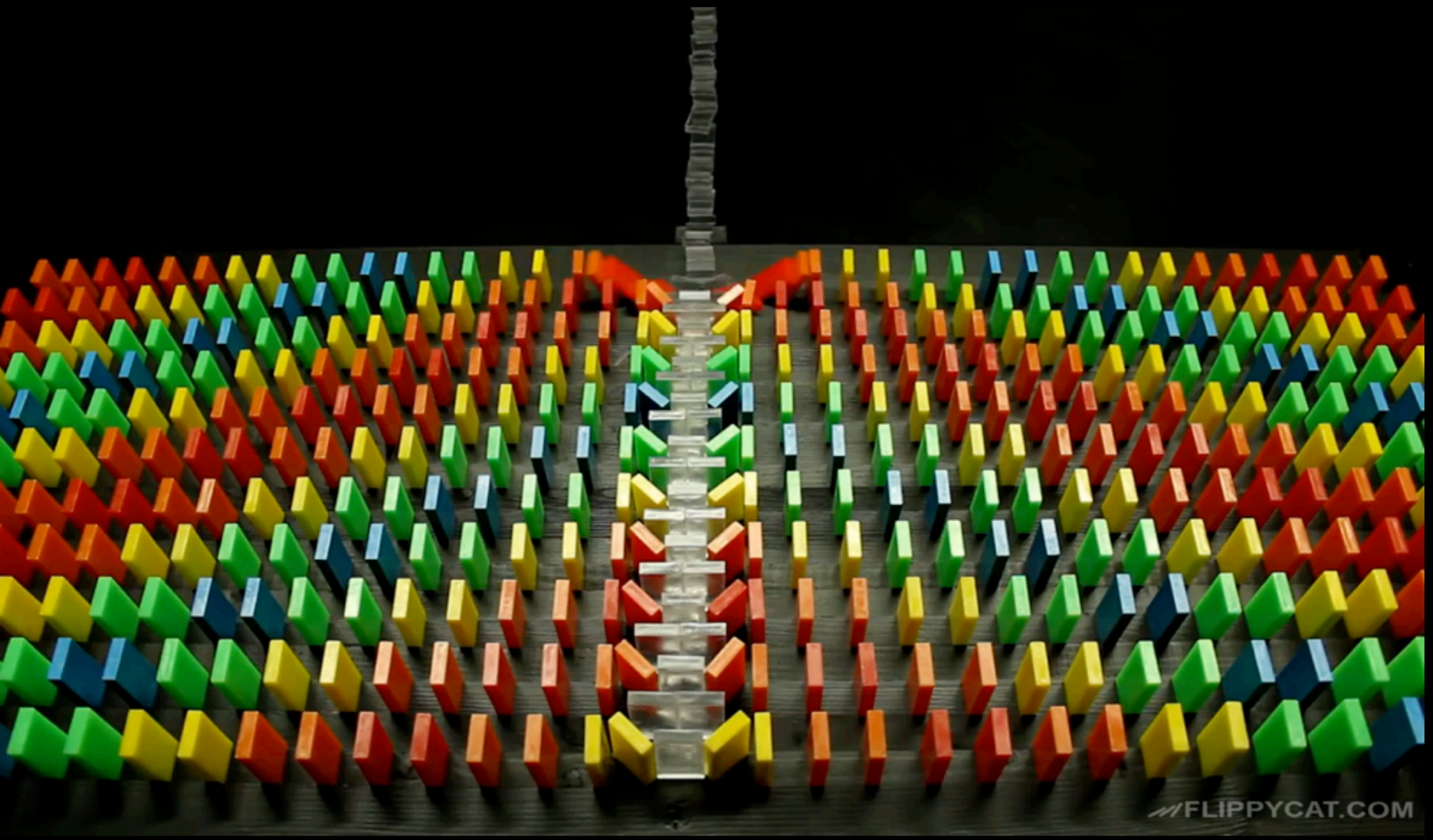
- DALiuGE's graph implementation translates Data constructs into *DataDrop* objects.
- A Drop is an active software object wrapper referring to a payload.
- Typically the data is represented as a URL reference in a Drop object.
- A Drop payload can be anything, actual data but also executable binaries (*ApplicationDrop* objects)
- Enables the implementation of a Drop life-cycle management system for data and applications.
- Enables the deployment of Data

Execution (2)

- Drop objects also define event channels, which are implemented as ZeroMQ PUB/SUB sockets.
- Drop objects implement a state machine.
- Combine these:
  - When a DataDrop object transitions to the ‘completed’ state it fires an event to its subscribers.
  - This triggers one or more consumer ApplicationDrops to act on the data payload.
  - That means a single root DataDrop on a PG can trigger the execution of the whole rest of the PG.
  - Very similar to a complex domino challenge.

ICRAR International Centre for Radio Astronomy Research

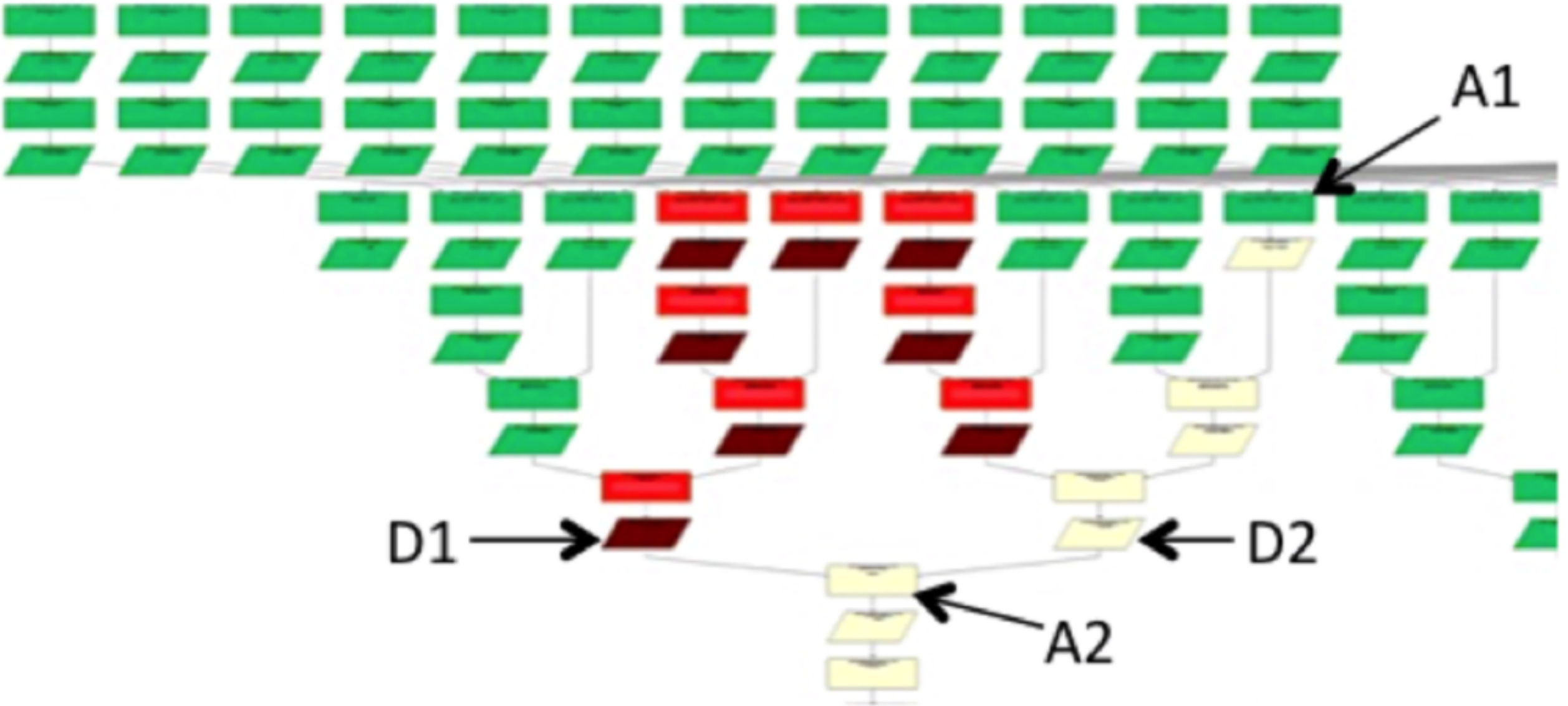
SKA SQUARE KILOMETRE ARRAY SDP SCIENCE DATA PROCESSOR



Error Handling

Errors

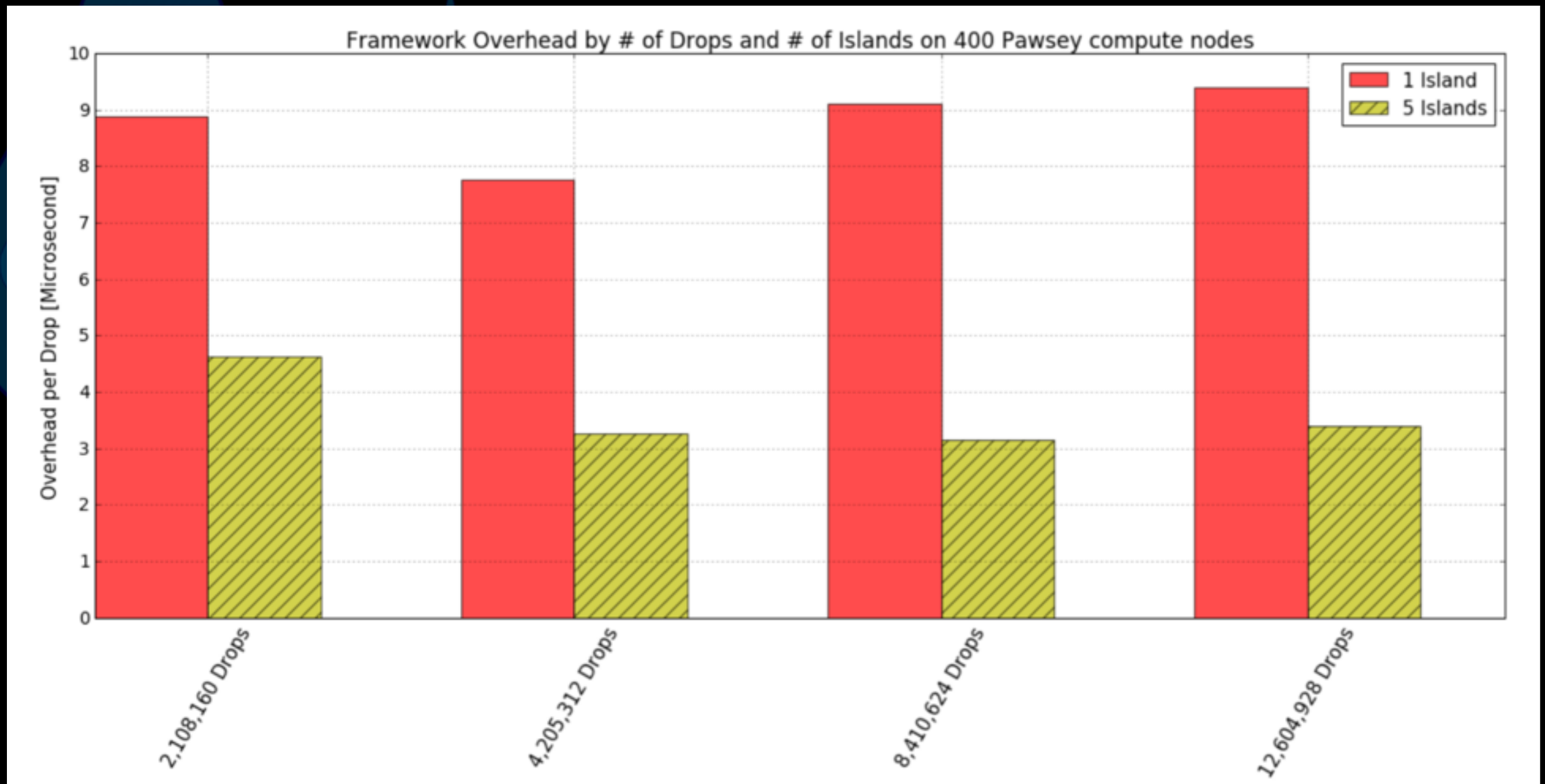
- Like with dominos, some of the chains WILL fail.
- We have constructs with thresholds to deal with that….



Scalability

How many Drops?

- We have executed a whole series of scalability tests on various platforms, including Tianhe-2 in China.
- Measured plain overhead during execution time.



Real World Examples

CHILES

- DALiuGE has been verified using CHILES data on AWS, in-house cluster, Magnus and Galaxy.
- The current version of the code creates 40+ Node managers all running on separate heterogeneous AWS instances; with a single Data Island Manager controlling them.
  - These graphs contain 7,000 ~ 88,000 Drops
  - The graph generator knows the AWS instance types and can deploy more CPU/IO intensive tasks to more powerful nodes.
- The CasaPy tasks are all run from within Docker containers controlled by the DALiuGE

Viewer Display Panel (DJ) (on epeius.icrar.org)

Data Display Panel Tools View Help

Display

clean_1408~1412.image-raster

2117.1 km/s

J2000 Declination

20"
26'00"
40"
20"
25'00"
40"
20"
2°24'00"
40"

10h01m59s 56s 54s

J2000 Right

Animators

Channels

Rate: 5 Jump 152 256

150 200

Images

Cursors

clean_1408~1412.image-raster

+3.19447e-05 Jy/beam Pixel: 642 1253 0 153
10:01:55.850 +02.25.46.455 I 2113.8 km/s (lsrk/radio velocity)

Spectral Profile - clean_1408~1412.image-raster (on epeius.icrar.org)

Rectangle Region Profile

(mJy/beam)

0.4
0.3
0.2
0.1
0
-0.1
-0.2

1408 1408.5 1409 1409.5 1410 1410.5 1411

frequency [MHz]

…and more

- we have ported the basic version of a MWA GLEAM pipeline to Daliuge.
- ASTRON is working to port and run the LOFAR pipeline.
- Kunming University ported their existing MUSER pipeline.
- we are also integrating OSKAR2 to simulate and reduce SKA-scale data sets on Tianhe-2.
- code is available on SKA SDP github.
- documented and fully tested code (continuous integration with loads of test code)
- Graph translation and optimised scheduling is a really hard problem…

Daliuge run on Tianhe2
66473 Drops

Step 1/9