Recent Innovations in Data Handling for CANARI

Bryan Lawrence and a cast of thousands (as you will see)

NCAS and University of Reading

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

8th ENES HPC Workshop on HPC for high-resolution climate and weather modelling
Lecce, 2024-05-24

University of Reading

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

1. CANARI
   - Context and Requirements, from ARCHER2 to tape!



2. Atomic Datasets and Aggregation
   - How do we think about products and how (will) we manipulate them?
   - (re) introduction to CFA

3. Using CFA for data management as well as analysis.
   - A new interface and mechanism for managing atomic datasets and their physical location.

4. Futures, including Active Storage
   - A new technique for managing remote data manipulation.

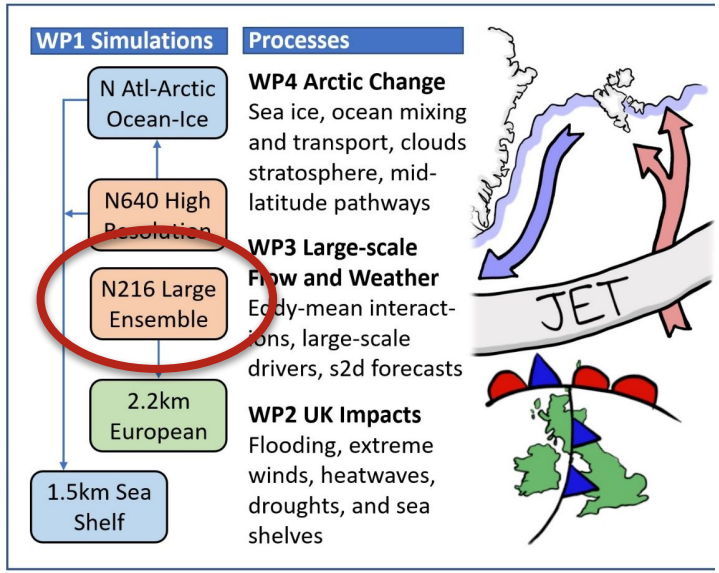




National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Climate Change in the Arctic-North Atlantic Region and Impacts on the UK

Advance understanding of the impacts on the UK arising from climate variability and change in the Arctic-North Atlantic region, with a focus on extreme weather and the potential for rapid, disruptive change





Len Shaffrey

WP5 Knowledge Integration
Process Understanding
combined with
User Knowledge

National Oceanography Centre

UK Centre for Ecology & Hydrology

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

BGS British Geological Survey

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

cpom Centre for Polar Observation and Modelling
Natural Environment Research Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Met Office

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI Large Ensemble



Reinhard Schiemann

Model

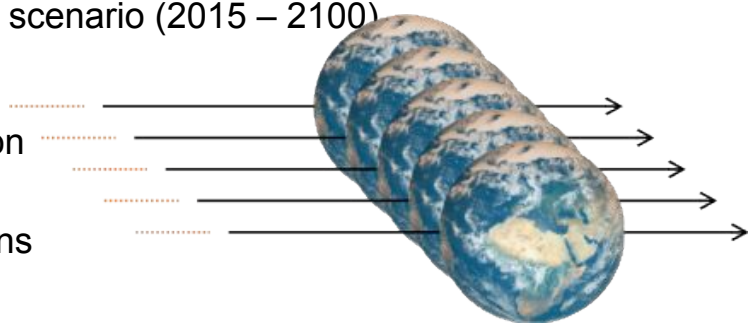
- HadGEM3-GC3.1 (the UK Met Office CMIP6 model)
- N216-ORCA0.25 (~60 km atmosphere, ¼° ocean)

Experiments

- CMIP6 historical (1950 – 2014)
- CMIP6 SSP3-7.0 scenario (2015 – 2100)

40 members

- Hybrid initialisation
- 8 macro ICs x 5 micro perturbations



Output

- 6000 years - “CMIP-like”.
- UM boundary conditions (“Frames”) for identified European storms.
- RCM boundary conditions for all CORDEX regions
- UM boundary conditions (“Frames”) for NZ and East-Asia

Data

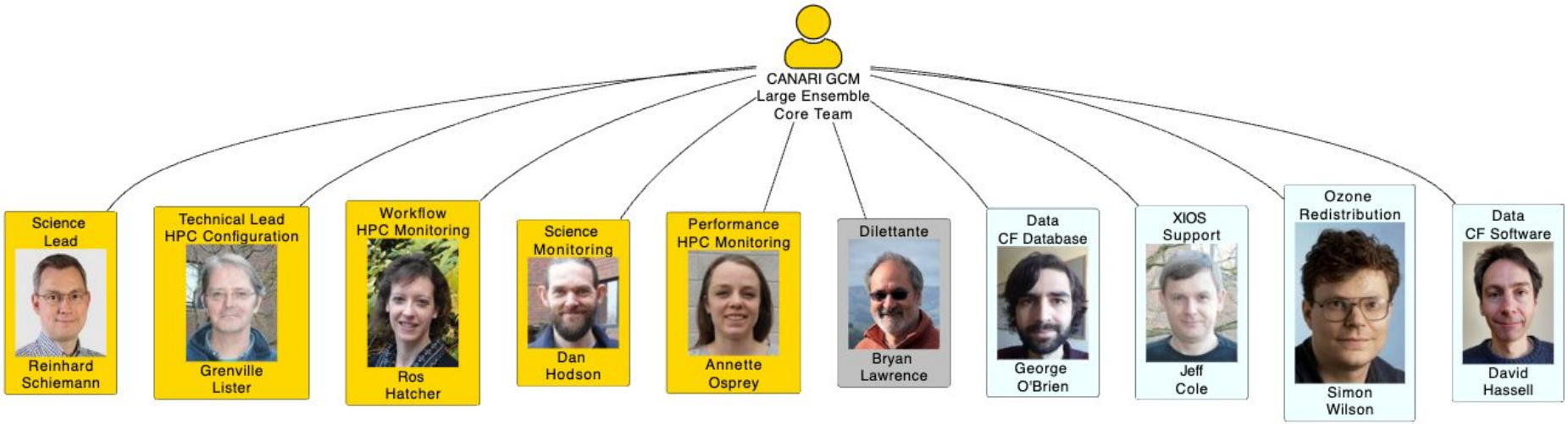
Volumes (compressed): ~5.5 PB

- 2 PB atmos fields
- 0.5 PB atmos restarts
- 0.4 PB ocean fields
- 0.1 PB ocean restarts
- 0.1PB ice fields and restarts
- 2.1 PB of frames etc

4.5 Million Files!

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

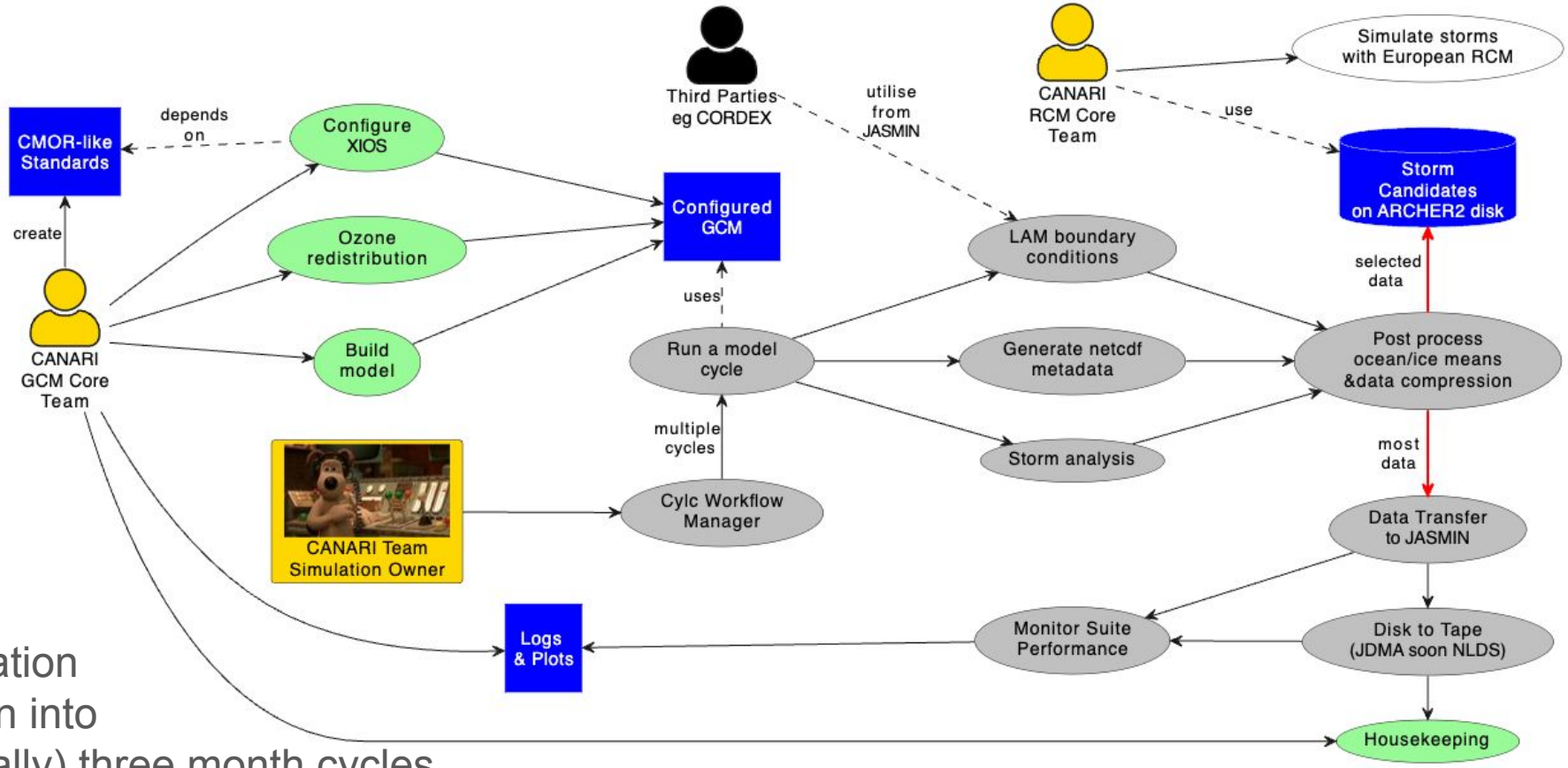
CANARI: Large team put the model and workflow into place



Every one of whom had/has a role to play in setting up and/or sustaining the large-ensemble.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI Activities

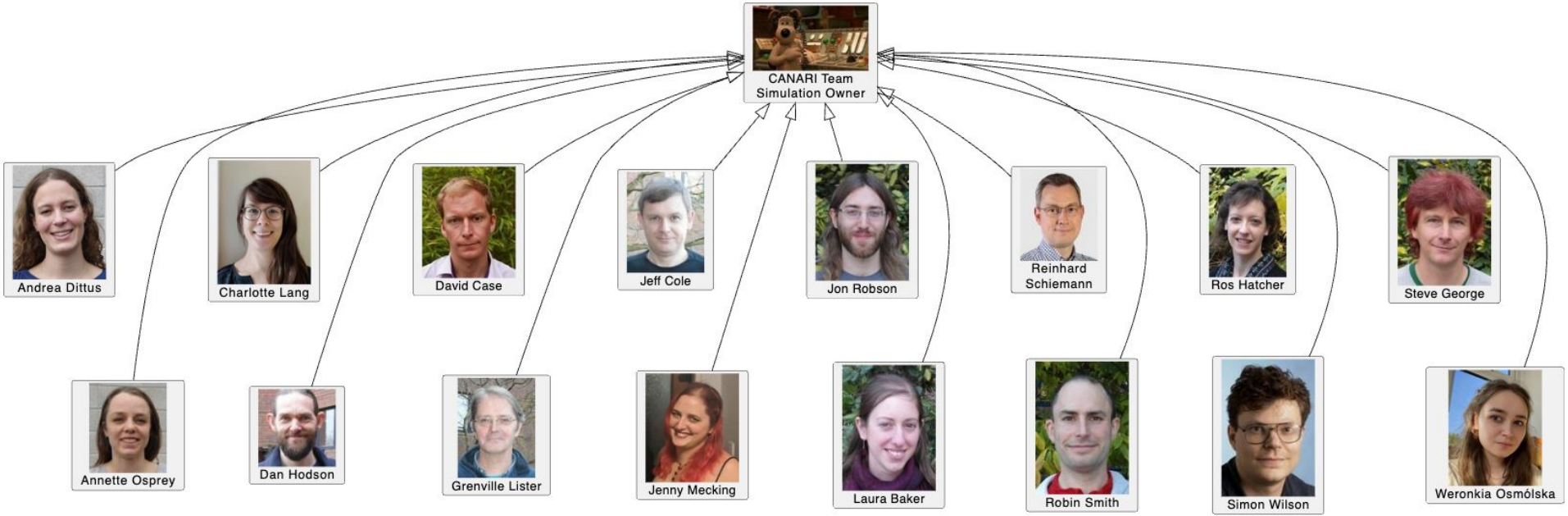


Each simulation broken into (typically) three month cycles.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI: Large team involved in running the model

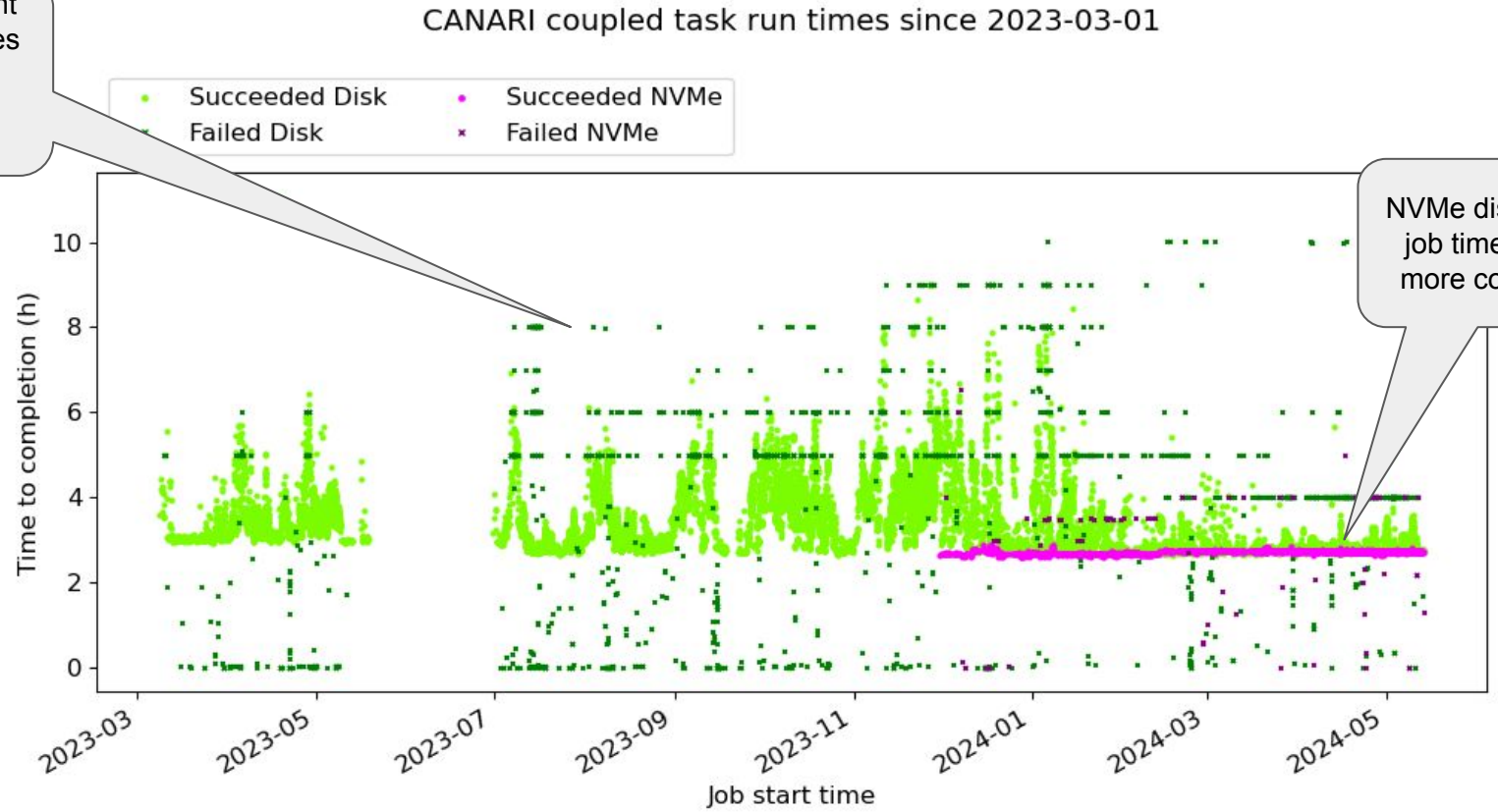
Many individuals, each taking responsibility for cycling one or more simulations through the machine



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Lots of failed jobs, lots of variability

Mostly silent node failures and/or file system issues



NVMe disk makes job times much more consistent

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

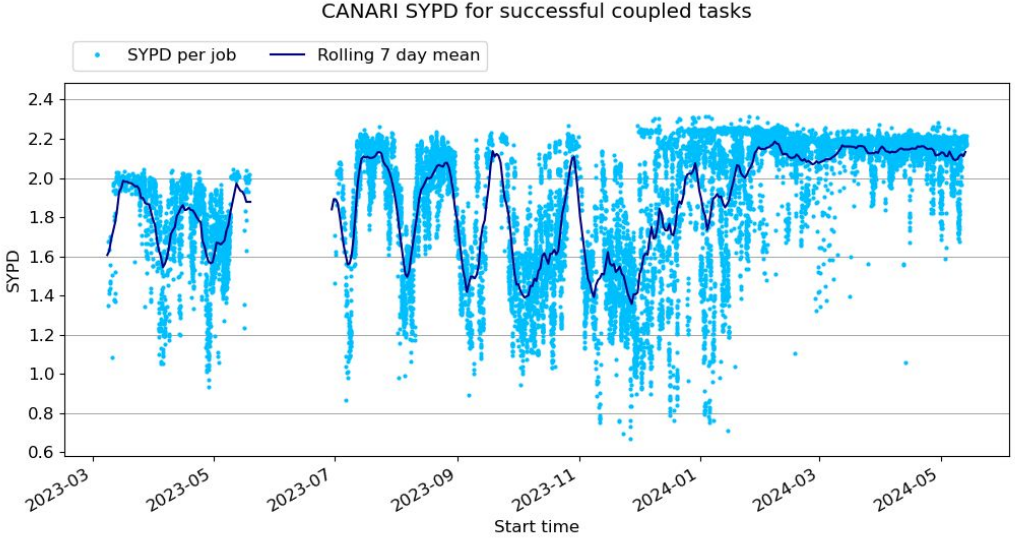
Reported Average Speeds are Misleading


Grenville Lister

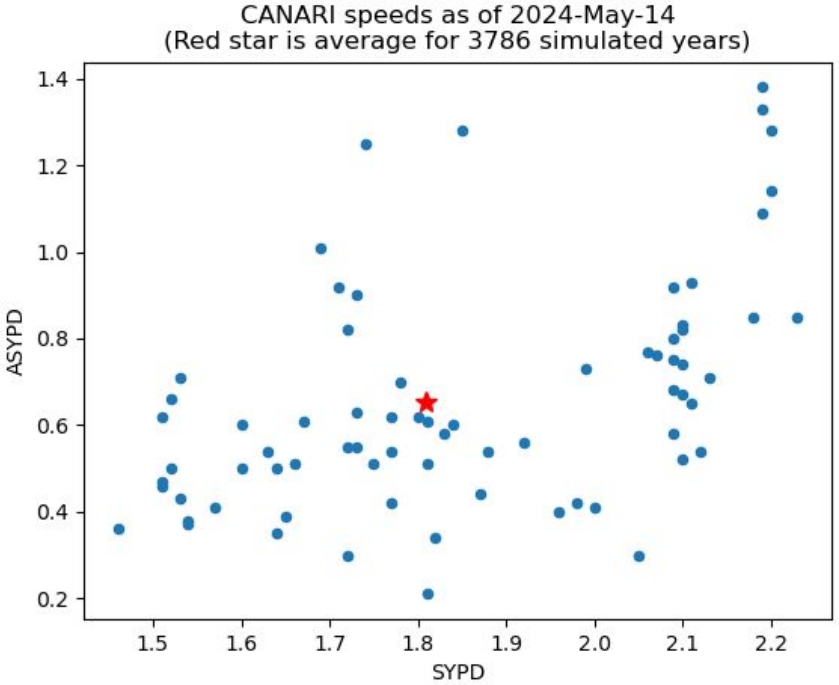

Ros Hatcher


Annette Osprey

Each dot is one cycle completed.



Each dot is one simulation (completed or in progress).



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

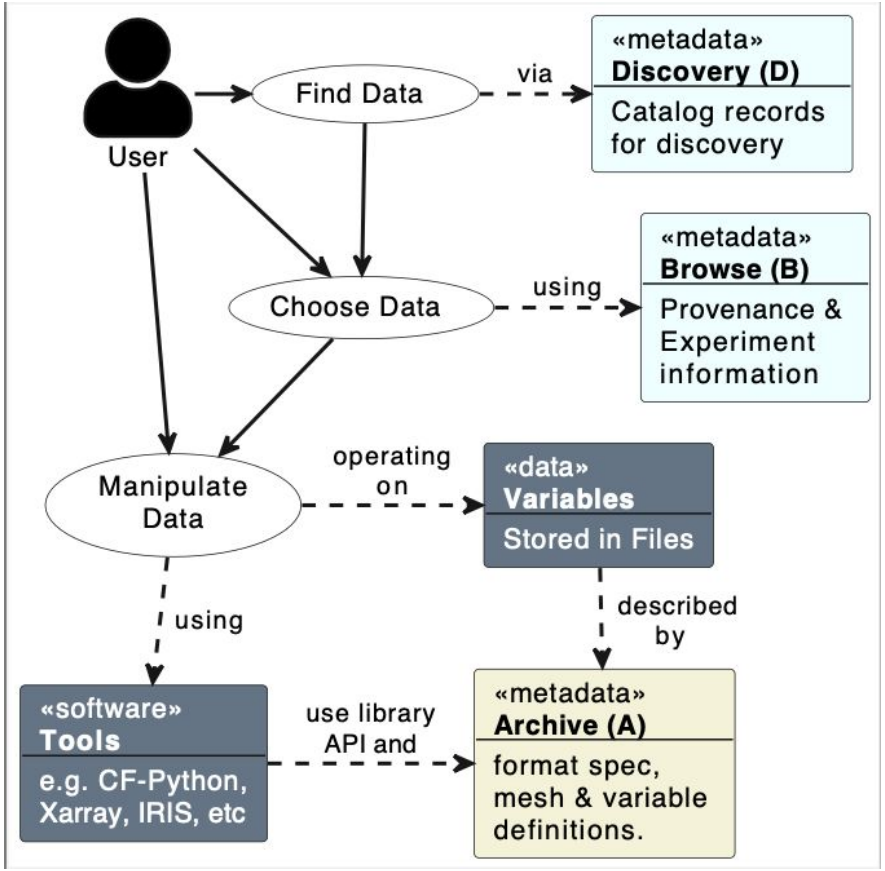
TOO MANY FILES

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

VOCAB: Atomic Datasets and Aggregation

Reminder:

- An atomic dataset is the entire sequence of data from one simulation for one field variable with a common sampling frequency and cell method.
- Typically consists of part or all of multiple files - whatever the format used.
- Discovery (D) metadata often defines it, but up to tools to deliver an aggregated view of contents (including B metadata).



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Climate Forecast Aggregation (CFA) Conventions

A normal dataset

```
dimensions:
  time = 12 ;
  latitude = 73 ;
  longitude = 144 ;
variables:
  double temp(time, latitude, longitude);
    temp.standard_name = "air_temperature" ;
    temp.units = "K" ;
    temp:cell_methods = "time: mean" ;
```

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Climate Forecast Aggregation (CFA) Conventions

A normal dataset

```
dimensions:
  time = 12 ;
  latitude = 73 ;
  longitude = 144 ;
variables:
  double temp(time, latitude, longitude);
    temp.standard_name = "air_temperature" ;
    temp.units = "K" ;
    temp:cell_methods = "time: mean" ;
```

A CFA dataset

As CF-like as possible: all CF metadata is unchanged, just the specification of the data array differs

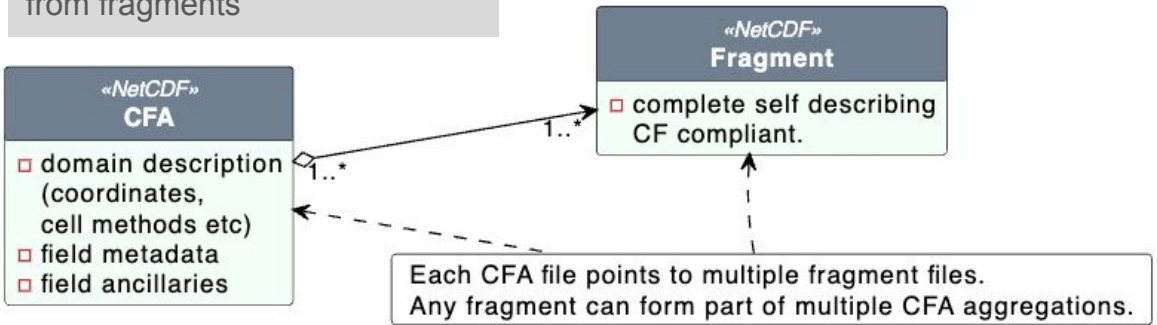
```
dimensions:
  time = 12 ;
  latitude = 73 ;
  longitude = 144 ;
variables:
  double temp ;
    temp.standard_name = "air_temperature" ;
    temp.units = "K" ;
    temp:cell_methods = "time: mean" ;
    temp:aggregated_dimensions = "time latitude longitude" ;
    temp:aggregated_data = "location: aggregation_location
                            file: aggregation_file
                            format: aggregation_format
                            address: aggregation_address" ;
```

scalar

actual dimensions

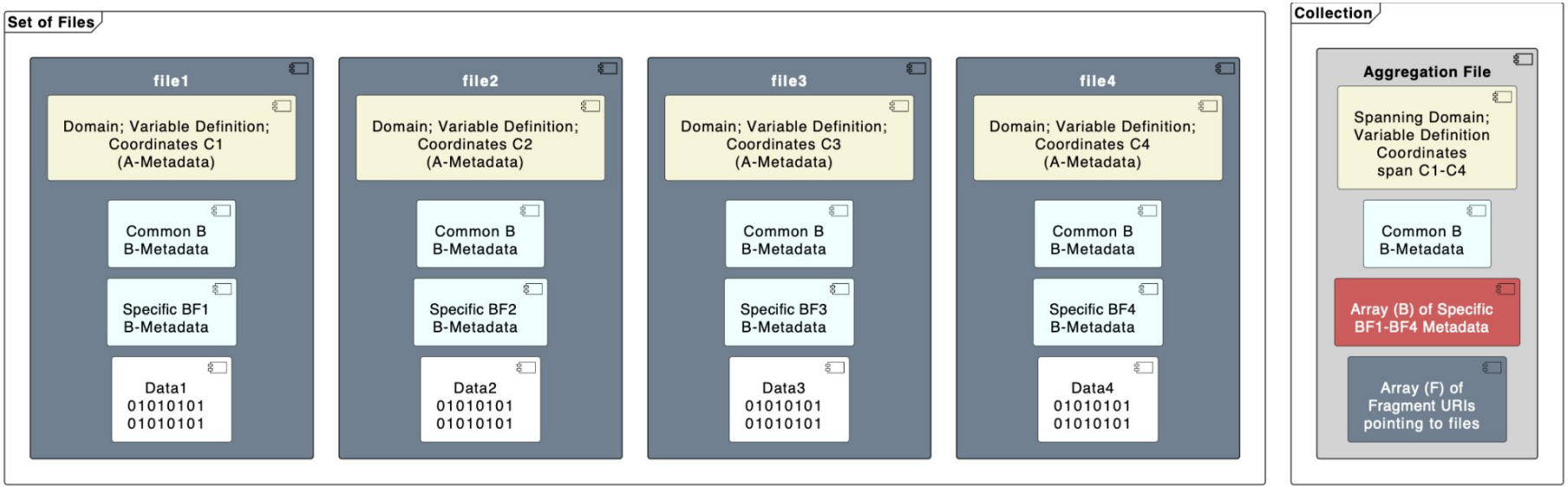
Instructions on how to create the data

Complete A and B metadata from fragments



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA - All the benefits of CF with aggregation!



Fragments can be duplicated in multiple locations

- Push the atomic analogy too far: quarks, being subsets in different locations!
- We can record the location of quarks and move them around!

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA Example: Two fragment Files

cfdump is a tool of **cfdm** (which is used by **cf-python**).
cfdm is the "formal implementation of the CF data model".
cf-python is a tool which (transparently) uses **cfdm.**

```
% cfdump temperature1.nc
Field: air_temperature (ncvar%tas)
----------------------------------
Data            : air_temperature(time(3), latitude(324), longitude(432)) K
Cell methods    : area: time(3): mean
Dimension coords: time(3) = [2015-01-01 12:00:00, 2015-01-02 12:00:00, 2015-01-03 12:00:00]
360_day
                : latitude(324) = [-89.72222137451172, ..., 89.72223663330078] degrees_north
                : longitude(432) = [0.4166666567325592, ..., 359.58331298828125] degrees_east
                : height(1) = [1.5] m
Cell measures   : measure:area (external variable: ncvar%areacella)
% cfdump temperature2.nc
Field: air_temperature (ncvar%tas)
----------------------------------
Data            : air_temperature(time(9), latitude(324), longitude(432)) K
Cell methods    : area: time(9): mean
Dimension coords: time(9) = [2015-01-04 12:00:00, ..., 2015-01-12 12:00:00] 360_day
                : latitude(324) = [-89.72222137451172, ..., 89.72223663330078] degrees_north
                : longitude(432) = [0.4166666567325592, ..., 359.58331298828125] degrees_east
                : height(1) = [1.5] m
Cell measures   : measure:area (external variable: ncvar%areacella)
```

```
!du -sh temperature*.nc

60K     temperature.nc
1.7M    temperature1.nc
4.9M    temperature2.nc
```

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA Aggregated View

Currently a different tool to look at CFA files; when CFA formally in CF, cfdump will do the same thing!

```
cfa -vm temperature.nc
Field: air_temperature (ncvar%tas)
----------------------------------
Data            : air_temperature(time(12), latitude(324), longitude(432)) K
Cell methods    : area: time(12): mean
Field ancils    : id%tracking_id(time(12), latitude(324), longitude(432)) =
[[[0551c766-f4e2-4dc5-a47d-234a40919bf9, ..., 77ff6beb-d7b9-475a-83fe-b3a1a789cabb]]]
Dimension coords: time(12) = [2015-01-01 12:00:00, ..., 2015-01-12 12:00:00] 360_day
                : latitude(324) = [-89.72222137451172, ..., 89.72223663330078] degrees_north
                : longitude(432) = [0.4166666567325592, ..., 359.58331298828125] degrees_east
                : height(1) = [1.5] m
Cell measures   : measure:area (external variable: ncvar%areacella)
```

Field ancillaries are non-domain metadata that vary over the domain

```
!du -sh temperature*.nc
60K     temperature.nc
1.7M    temperature1.nc
4.9M    temperature2.nc
```

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA - Transparent to Application

Note apparent dimension of tracking-id (but it does only take up N numbers in memory) where N is the number of fragments.

In [6]:
```
import cf
import numpy as np
f = cf.read('temperature.nc')[0]
```

In [7]:
```
print(f)
```

```
Field: air_temperature (ncvar%tas)
----------------------------------
Data            : air_temperature(time(12), latitude(324), longitude(432)) K
Cell methods    : area: time(12): mean
Field ancils    : id%tracking_id(time(12), latitude(324), longitude(432)) = [[[0551c766-f4e2-4dc5-a47d-234a40919bf
9, ..., 77ff6beb-d7b9-475a-83fe-b3a1a789cabb]]]
Dimension coords: time(12) = [2015-01-01 12:00:00, ..., 2015-01-12 12:00:00] 360_day
                : latitude(324) = [-89.72222137451172, ..., 89.72223663330078] degrees_north
                : longitude(432) = [0.4166666567325592, ..., 359.58331298828125] degrees_east
                : height(1) = [1.5] m
Cell measures   : measure:area (external variable: ncvar%areacella)
```

In [8]:
```
g=f[2:4]
g
```

Out[8]:
```
<CF Field: air_temperature(time(2), latitude(324), longitude(432)) K>
```

In [9]:
```
tid = g.field_ancillary('id%tracking_id')
tid=tid.data
print(tid.shape)
print(np.unique(tid))
```

```
(2, 324, 432)
['0551c766-f4e2-4dc5-a47d-234a40919bf9'
 '77ff6beb-d7b9-475a-83fe-b3a1a789cabb']
```

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA - Use logicals to help with locations

This particular example was written by David Hassell, and one of the file attributes was:

```
cfa_file:substitutions = "${location}: file:// /home/david/cf-python/"
```

So that the fragment files are recorded in the file as :

```
cfa_location =
  "${location}temperature1.nc",
  "${location}temperature2.nc" ;
```

When I have the same data in my own directory, I can read the data with:

```
g = cf.read('temperature_sub.nc',
    cfa={'substitutions': {'location': ' /home/bryan/data}})
```

This same mechanism can be used to copy a CFA file to another location and/or copy the data files to another location, and make things work provided there is a suitable data service available and understood by the tool reading the file.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA Timeline

- **CFA 0.4** - used by s3netCDF v1 (ESIWACE1).....
- **CFA 0.5** - Neil's somewhat hacked version used by s3netCDF v2
- **CFA 0.6** - Aug 2021 - Improved version of 0.5.
- **CFA 0.6.1** - Dec 2021 - reference for the library implementations
- **CFA-C** - Apr 2022 - first implementation of the reference
- **CFA-Python** - Jul 2022 - Python wrapper to CFA-C
- **CFA-CPP** - Jul 2022 - C++ wrapper to CFA-C
- **CF-Python Support** - Apr 23
- **Agreement by CF community** to include in CF conventions - Sep 23

- *Formal proposal to CF*
  - June 24
- *Xarray support*
  - *Summer 24*
- *Finalise CFA in CF*
  - *Late Summer 24*


Bryan Lawrence


David Hassell

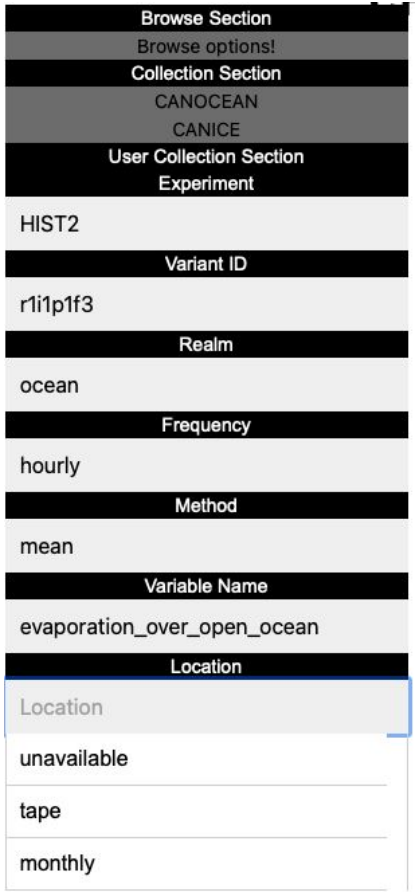

Jonathan Gregory


Neil Massey


Sadie Bartholomew

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI and Atomic Datasets



- The complete set of atomic datasets for CANARI is known a priori.
- We can (and have) built a website for browsing the CANARI contents (a complete SQL representation of CF metadata).
- We keep CFA files as the ground truth, and use them both to track locations of files in several states:
  - Unavailable (not yet simulated and/or not yet at JASMIN)
  - Tape (on tape)
  - On disk (or where-ever, we can make use of the location substitutions).

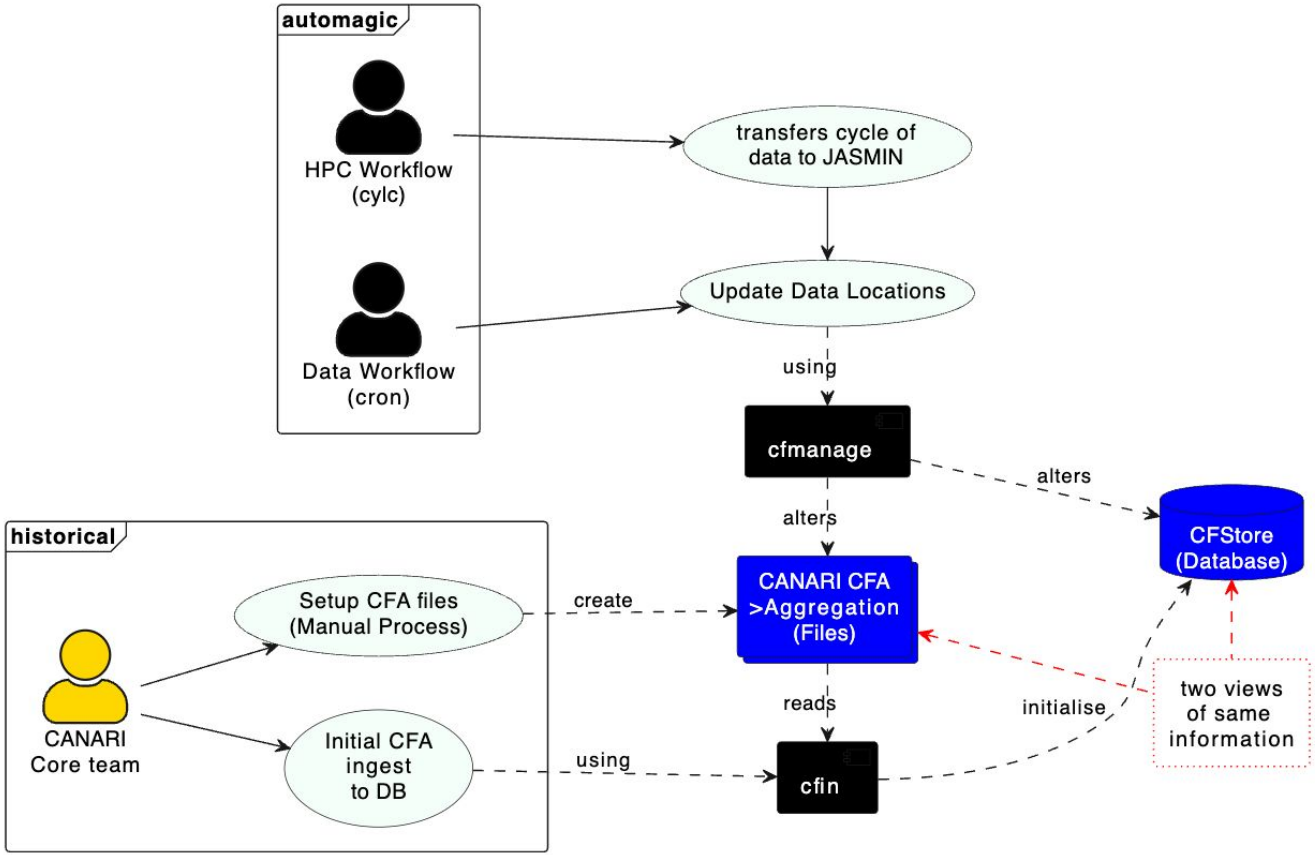

Bryan Lawrence


David Hassell


George O'Brien

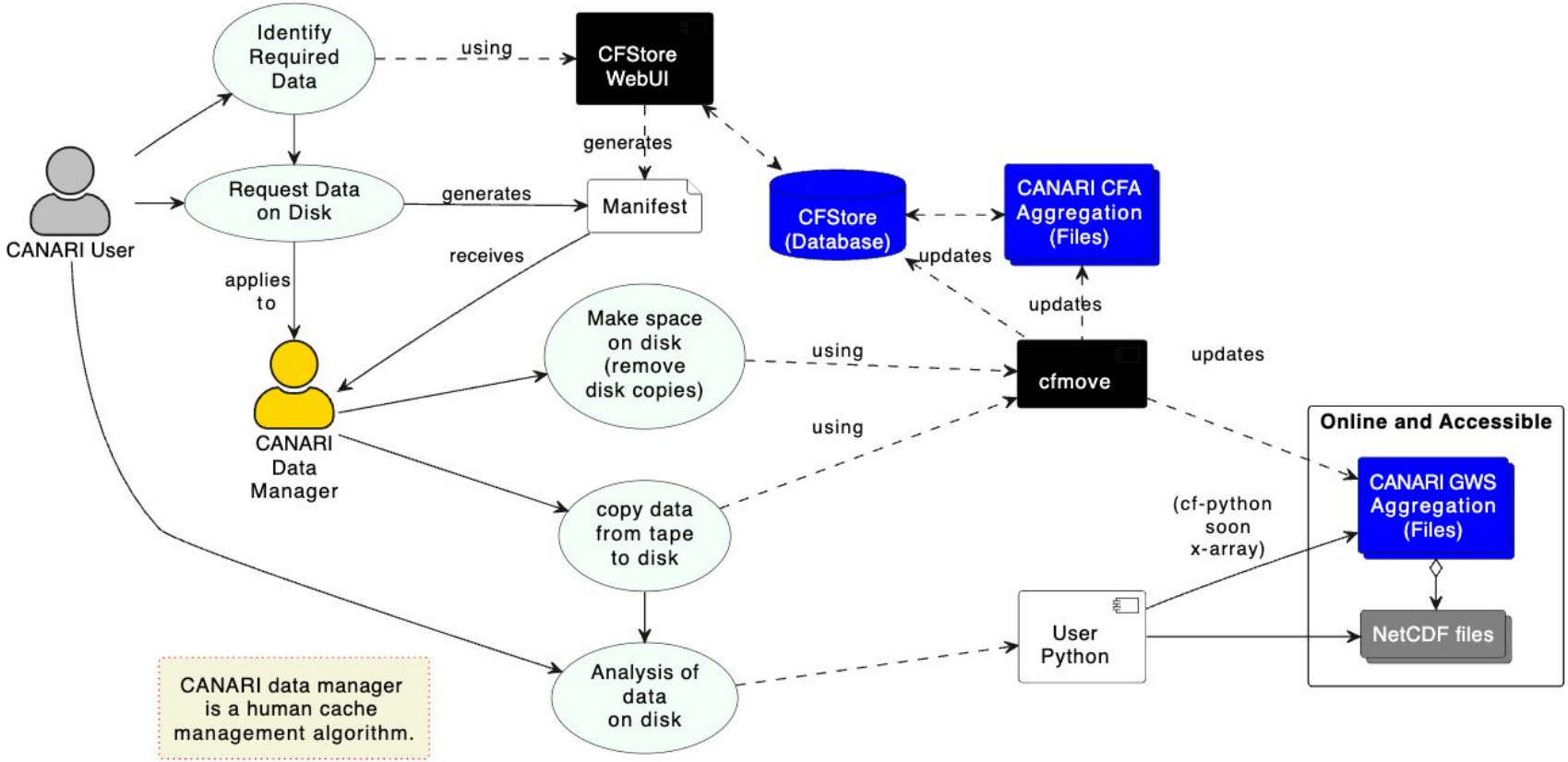
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI: Managing the CFA Files



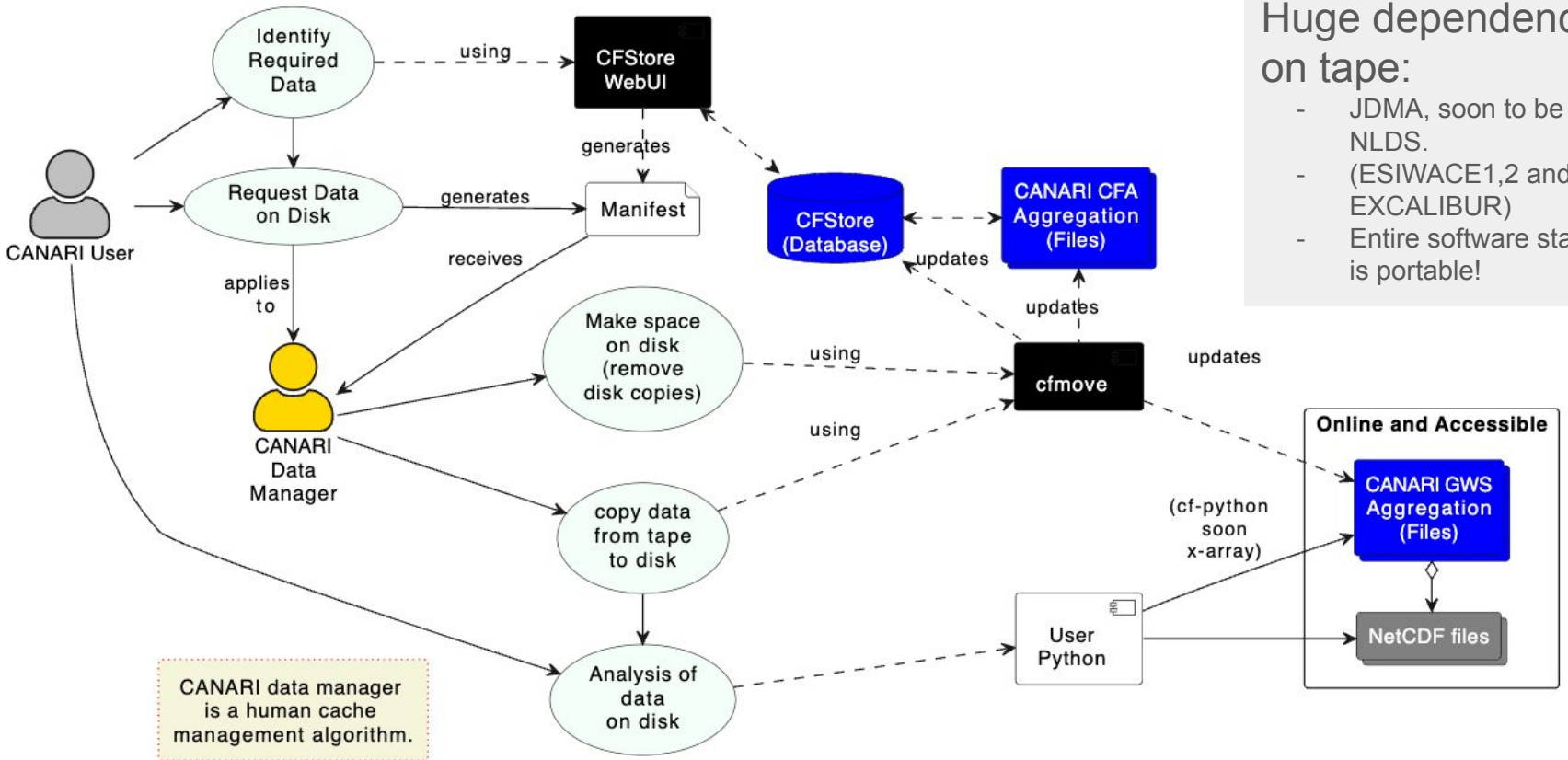
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA - Using for data management



National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CFA - Using for data management



Huge dependency on tape:

- JDMA, soon to be NLDS.
- (ESIWACE1,2 and EXCALIBUR)
- Entire software stack is portable!

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI: One CFAmaster per simulation per realm

```
In [6]: cd atmos
/mnt/canari/cfa/generatedncfiles/atmos
In [7]: ls
CANARI_1_cs125_atmos.nc   CANARI_26_db301_atmos.nc
 …40 files!

In [8]: import cf
In [9]: fx = cf.read('CANARI_1_cs125_atmos.nc')
```

- 822 atmosphere atomic datasets for one simulation
- 32,800 (in 40 files) for CANARI
- correspond to > 2 million actual NetCDF files.

```
In [36]: len(fx)
Out[36]: 822
In [37]: f = fx[0]
In [38]: f
Out[38]: <CF Field: eastward_wind(time(93600),
atmosphere_hybrid_height_coordinate(85), latitude(324), longitude(432)) m s-1>
```

```
In [39]: x=f.subspace(T=cf.wi(cf.dt('1990-01-01'),
                               cf.dt('1991-01-01')))
In [40]: x.data.get_filenames()
Out[40]:
{'unavailable:cs125_1_6hr_u_pt_cordex__198912-198912.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199001-199001.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199002-199002.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199003-199003.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199004-199004.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199005-199005.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199006-199006.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199007-199007.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199008-199008.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199009-199009.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199010-199010.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199011-199011.nc',
 'unavailable:cs125_1_6hr_u_pt_cordex__199012-199012.nc'}
```

These data not yet simulated!
Could be tape:... or disk:

All CF metadata on disk, even if fragments only on tape.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Futures - Higher Resolution and Distribution

- For some CANARI users, downloading subsets to remote sites is possible, but for most, it will not be.
- We expect future (maybe) large ensembles at even higher resolution.
- We expect to intercompare these data with high-resolution and/or large ensembles elsewhere.
  - Moving the data is not an option, we must move the computation.
  - But what if the data is distributed?
- Lots of plans to support distributed data analysis
  - Web Processing Service (WPS for pre-canned operations)
  - Active Storage
  - Dask Multicloud

EXPECT

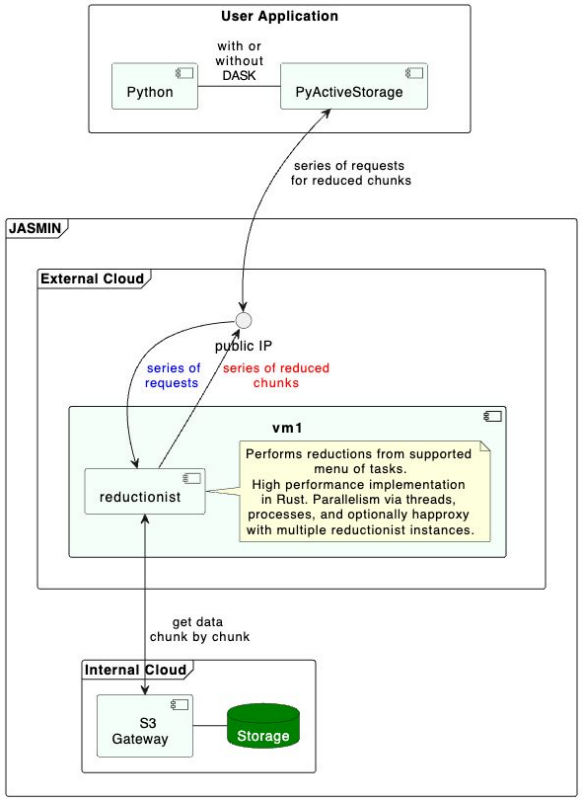
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

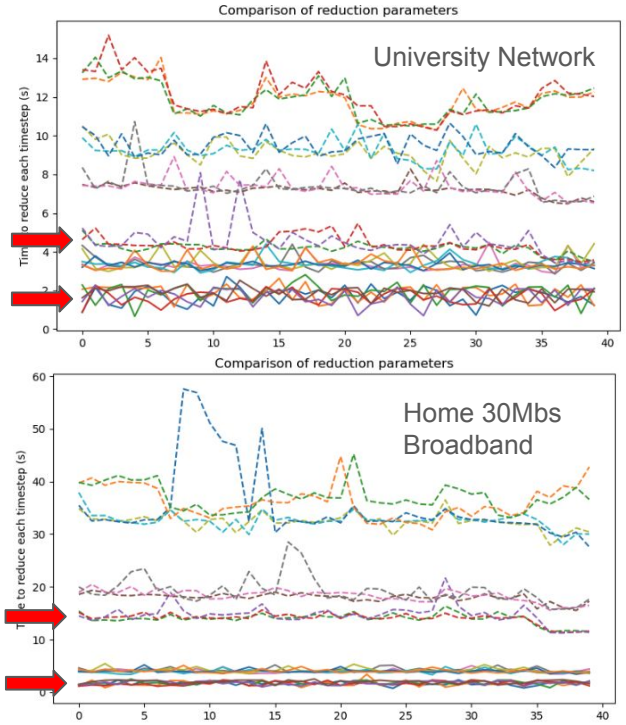
Active Storage: Reductions in storage (a la MPI reductions)

Python support for remote reductions in (or near) storage. S3 version is production ready, POSIX version in development.

Very small code changes (keywords and one or two lines at most) to get full support in Python.

Example on the right is doing hemispheric means of 40 days of 10km atmospheric data. Benefits much greater with higher resolution data.





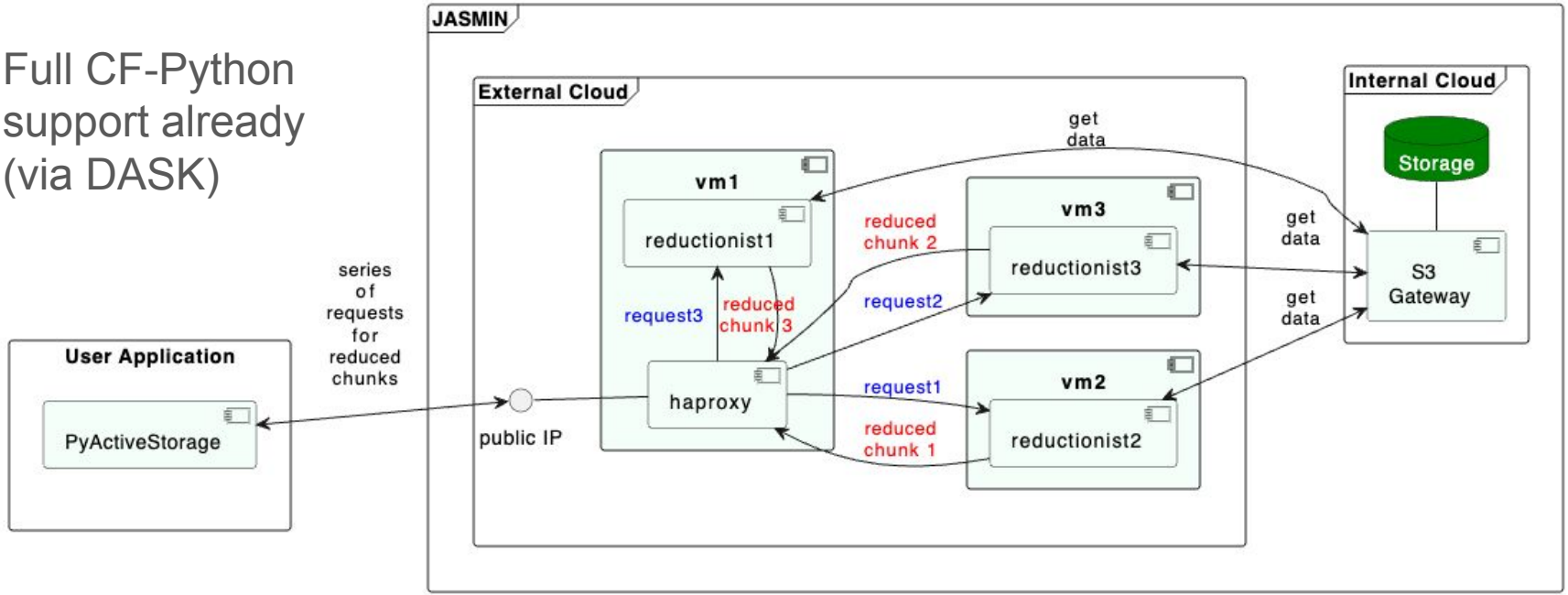
Testing with multiple threads and S3 parameters
Laptop on the university network, connected at 1 Gbit/s all the way to the S3 server, the best config of standard reductions take around 5s. Active is 3 times faster. On home broadband, the fastest option is 8 times faster and is comparable with the university network. The use of active storage has taken the local network out of the game.

ExCALIBUR 10 · University of Reading · National Centre for Atmospheric Science · Stack HPC · ddn

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Active Storage - Next Steps

Full CF-Python support already (via DASK)




Bryan Lawrence


David Hassell


V. Predoi

- Testing high availability, high performance S3.
- DDN working on POSIX version.
- Will deploy at selected European sites during:

EXPECT

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

On formats and future

- We will provide full support for HealPix in CF NetCDF via EXPECT - soon!
- We plan to provide CFA support in Xarray
- People keep asking me about Zarr: but it is not a priority for us *in workflow tooling* , because

    a) we don't (yet) see benefits from Zarr that we don't already get from:

        (1) Properly chunking NetCDF files

        (2) Properly organising the b-tree (HDF File chunk index) so it is itself not chunked and

        (3) Using pure python HDF5 readers so we can deliver thread safety when reading data.

    b) we do not want to handle that many files in a our workflow. It's bad for file systems, it's difficult to be sure of data integrity (less of a problem on AWS for sure)

- We will probably support zarr at some point in CF-Python, just as we aim to eventually support a restricted subset of GRIB as well; but we won't produce any of either if we can avoid it.

2 PB of CANARI atmos data.

- At recommended zarr chunk sizes of 10 MB = 200 million files. Who wants to keep that checksum manifest?

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Summary

- The CANARI large ensemble is a mammoth logistical effort.
- ARCHER2 is a difficult machine in which to manage a large project,
  - but the use of NVMe disk improves things a lot
- We are exploiting a lot of ESIWACE1, ESIWACE2, EXCALIBUR intellectual property to manage the workflow.
  - We continue to need sustained investment in data management tools as well as everything else in HPC.
- CFA is an aggregation format which provides a lot of benefits
  - Keeping all the metadata online, in NetCDF, regardless of what data is on tape, has many benefits!
- Looking forward, we are continuing to try and find ways of minimising data movement at the same time as minimising the consequential intrusion into code.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CANARI is a “Long-Term-Science-Multiple-Centre” (LTSM) project funded by NERC

UKRI Natural Environment Research Council

The large ensemble would not have been possible without the HPC platforms and all these projects!

archer2 JASMIN

The ESiWACE project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 675191

The project Centre of Excellence in Simulation of Weather and Climate in Europe Phase 2 (ESiWACE2) has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No 823988.

Apologies to the other ExcaliData participants whose names and photos are not included because I didn’t have time to sort the GDPR issues.

StackHPC ddn

The EXCALIBUR work described here was done in partnership with StackHPC and DDN. EXCALIBUR is delivered by UKRI and the Met Office

Met Office UKRI UK Research and Innovation

EXPECT is funded by the European Union's Horizon Europe research and innovation programme under grant agreement no.101137656 with additional support from UKRI.

UKRI UK Research and Innovation