



**National Centre for
Atmospheric Science**
NATURAL ENVIRONMENT RESEARCH COUNCIL



**University of
Reading**

ExCALIData: Exascale I/O & Storage and Workflow

[Bryan Lawrence + many]

<https://excalibur.ac.uk/projects/excalidata/>



StackHPC Ltd



**UNIVERSITY OF
CAMBRIDGE**
Research Computing Services



Overview

Exascale Computing ALgorithms & Infrastructures Benefiting UK Research (ExCALIBUR)

ExCALIBUR is a UK research programme that aims to deliver the next generation of high-performance simulation software for the highest-priority fields in UK research. It started in October 2019 and will run through until March 2025, redesigning high priority computer codes and algorithms to meet the demands of both advancing technology and UK research.

ExCALIBUR is built around four pillars – **separation of concerns, co-design, data science, investing in people**

<https://excalibur.ac.uk/>

ExCALIBUR Themes

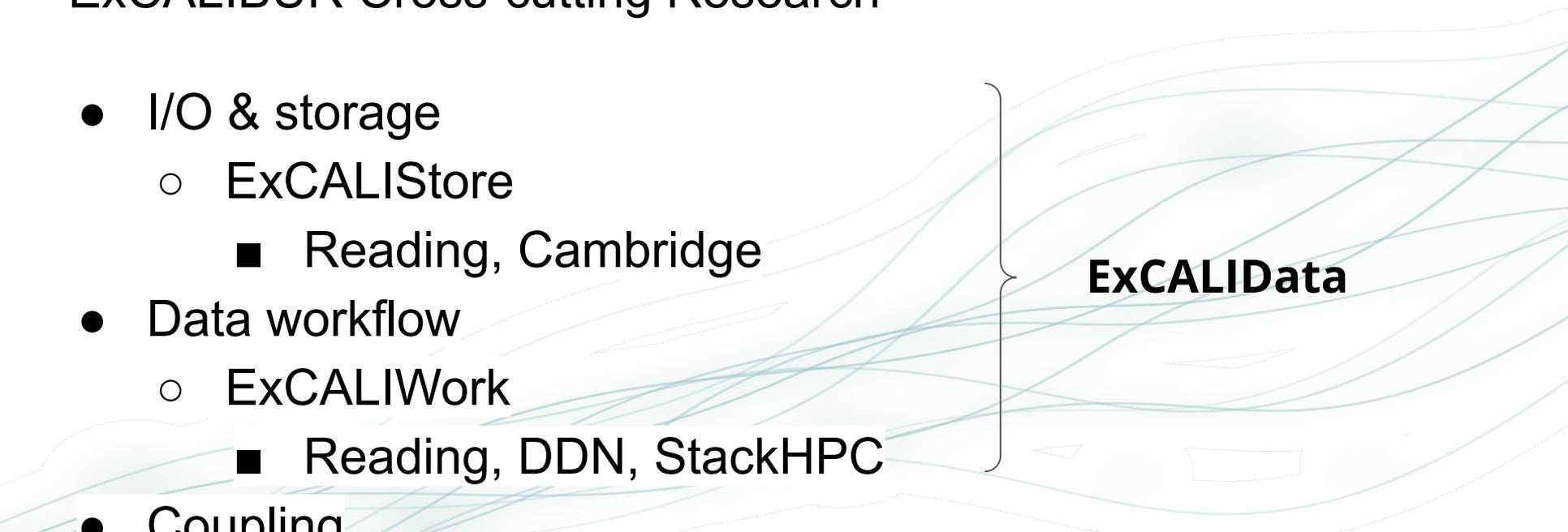
- High Priority use cases
 - Weather and Climate, Fusion
- Emerging Requirements for High Performance Algorithms
 - Social Sciences, Humanities, Biomedicine...
- Hardware and Enabling Software
 - Testbeds, early access to novel hardware and software
- Cross-cutting Research
- RSE Knowledge Integration
 - Grow interdisciplinary RSE community, fill skills gap, training...

ExCALIBUR Cross-cutting Research

- I/O & storage
- Data workflow
- Coupling
- Domain Specific Languages

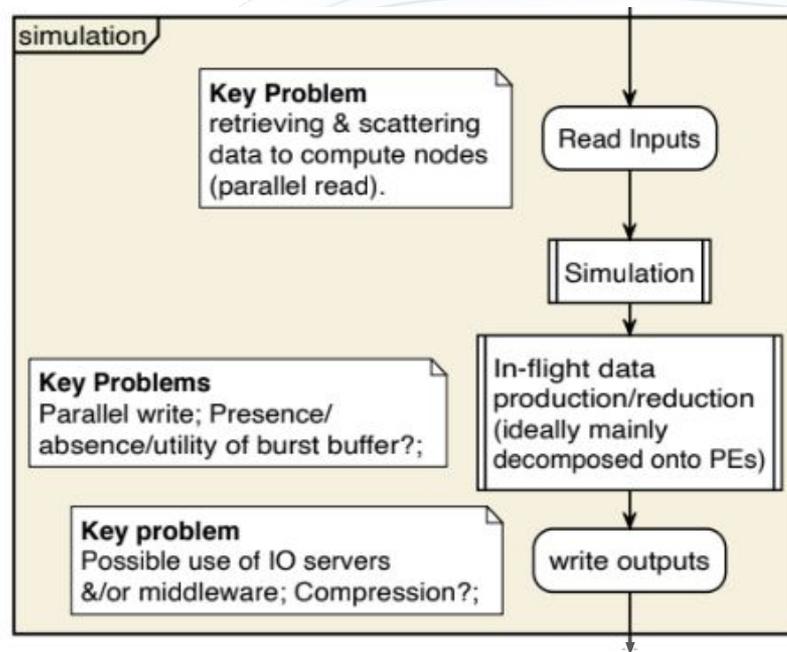
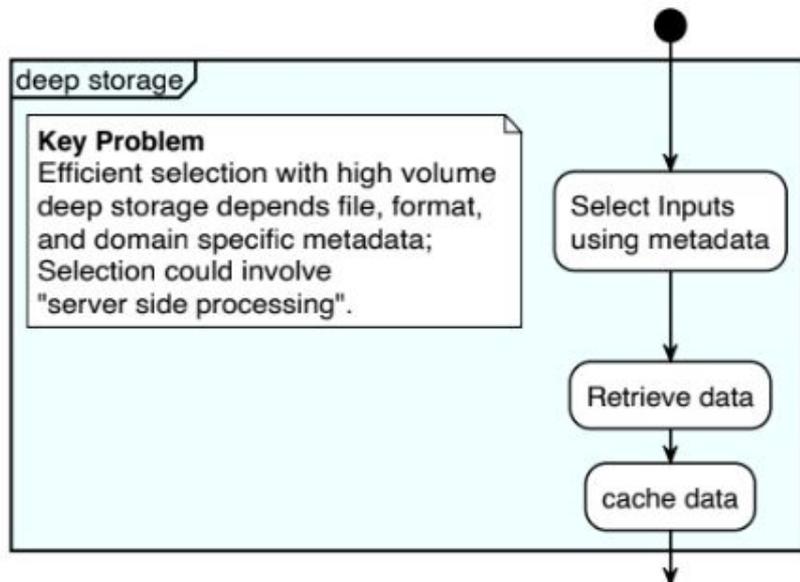
ExCALIBUR Cross-cutting Research

- I/O & storage
 - ExCALIStore
 - Reading, Cambridge
- Data workflow
 - ExCALIWork
 - Reading, DDN, StackHPC
- Coupling
- Domain Specific Languages

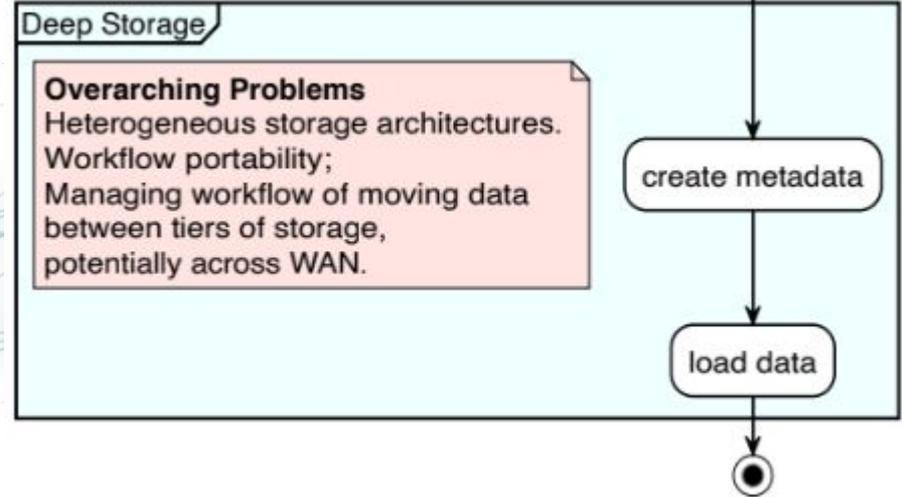
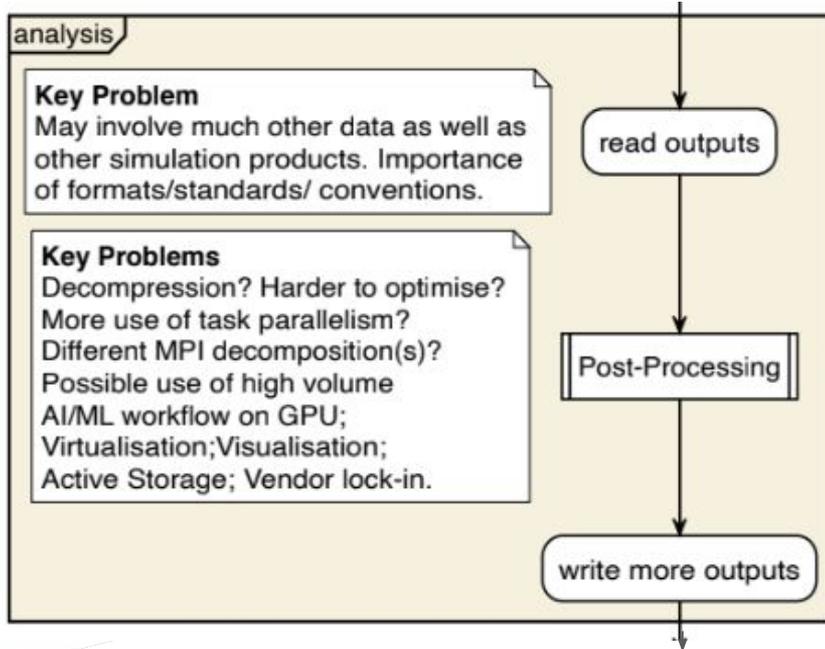


ExCALIData

ExCALIData



ExCALIData



ExCALIStore

Development and demonstration of novel approaches to optimising aspects of the data flow to improve I/O for large-scale applications.

- Storage interfaces
- Data management across storage tiers and between institutions
- Accelerating IO
 - Network fabric
 - Remote Direct Memory Addressing (RDMA) and/or SmartNICs
 - Advanced burst buffers
 - IO middleware
 - ADIOS
 - ESDM
- Synthetic tests and application to Weather & Climate and Fusion cases

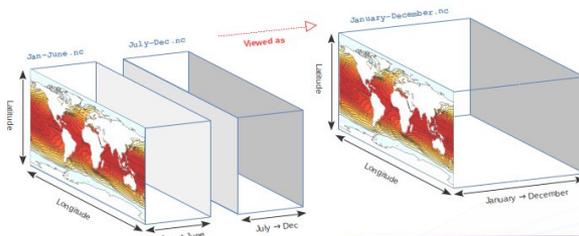
ExCALIStore

extract a particular spatio-temporal variable from a set of files

cf aggregation - underlying data remains unaltered, the aggregation presents the user with a single file

keep track of files held in multiple different storage elements

cf store - a method of cataloguing the list of atomic datasets available to user and tools for manipulating the atomic datasets (such as, listing all atomic datasets in a particular storage element, identifying duplicates, and moving subsets between storage elements).



```
netcdf January-December.nc
dimensions:
  // Aggregated dimensions
  > time = 12;
  > latitude = 73;
  > longitude = 144;
  // Fragment dimensions
  f_time = 2;
  f_latitude = 1;
  f_longitude = 1;
  1 = 3; // i = number of aggregated dimensions
  j = 2; // j = maximum of fragment dimension sizes

Variables:
double temp; // Aggregation variable, encoded as a scalar
temp:standard_name = "sea_surface_temperature";
temp:units = "K";
temp:cell_methods = "time: mean";
temp:aggregated_dimensions = "time latitude longitude";
temp:aggregation_location = "location: aggregation_location";
temp:aggregation_file = "aggregation_file";
temp:aggregation_format = "aggregation_format";
temp:aggregation_address = "aggregation_address";

float time(time);
time:units = "days since 2022-01-01";
float latitude(latitude);
latitude:units = "degrees_north";
float longitude(longitude);
longitude:units = "degrees_east";
// Aggregation instruction variables
string aggregation_addresses(f_time, f_latitude, f_longitude);
string aggregation_file(f_time, f_latitude, f_longitude);
int aggregation_location(i, j);

// Global attributes
:Conventions = "CF-1.9 CFA-0.6.1"; // CF and CFA conventions

data:
time = 0, 31, 59, 90, 120, 151, 181, 212, 243, 273, 304, 334;
temp = ;
aggregation_location = 6, 6, // Each fragment spans half the time range
73, 1; // All fragments span the whole latitude range
aggregation_file = "file:///data1/Jan-June.nc", "file:///data2/July-Dec.nc";
aggregation_format = "nc";
aggregation_address = "tos", "tos";
```

aggregated_dimensions: Dimensions of the aggregated data

aggregated_data: Instructions for aggregating the fragments

location: Locations of fragments in the aggregated data

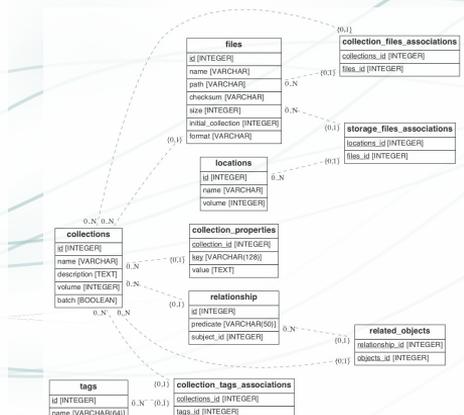
file: URIs of fragments

Format: Format of fragment files

Address: Addresses of data in fragment file

Dimensions indexing the number of fragments along each aggregated dimension

Dimensions indexing the fragment data sizes along each aggregated dimension (padded with missing values as required)



cf store

cf aggregation

David Hassell, Sadie Bartholomew, George O'Brien

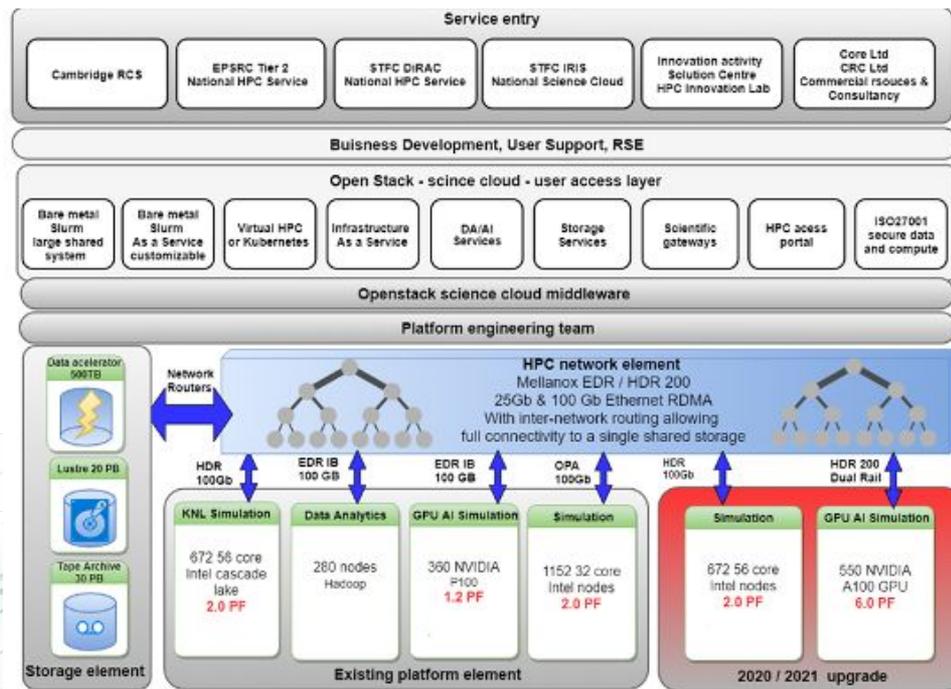


- <https://ncas-cms.github.io/cf-python>
- https://ncas-cms.github.io/cf-python/aggregation_rules.html
- <https://github.com/NCAS-CMS/cfstore>

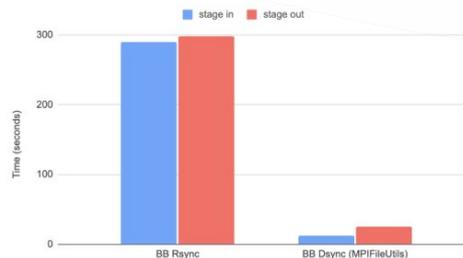
ExCALIStore

- CRCS testbed
- Cambridge Data Accelerator (DAC)
 - Burst Buffer
 - Improved NVMe file system life cycle
 - deploying a per job ephemeral directories
 - mpifileutils dsync as opposed to simple rsync for improved stage in and stage out performance
 - SLURM Burst buffer Lua functionality
 - Build a DAOS prototype system, integrate into the burst buffer framework

Paul Calleja, Wojciech Turek, Chris Edsall



Drain-back 100GB Dataset



ExCALIWork

Development and demonstration of novel approaches to taking certain computations to the data to reduce the need for data movement to improve I/O in large-scale applications

- Support for moving computation away from the traditional analysis phase
 - Increase concurrency/Reduce data movement
 - Move data reductions into the storage layer
 - Active Storage
 - User facing
 - Storage compute
 - Move ensemble manipulations into the simulation phase
 - Leverage XIOS capability
 - Current model (UM)
 - LFRic
- Synthetic tests and application to Weather & Climate and Fusion cases

ExCALIWork

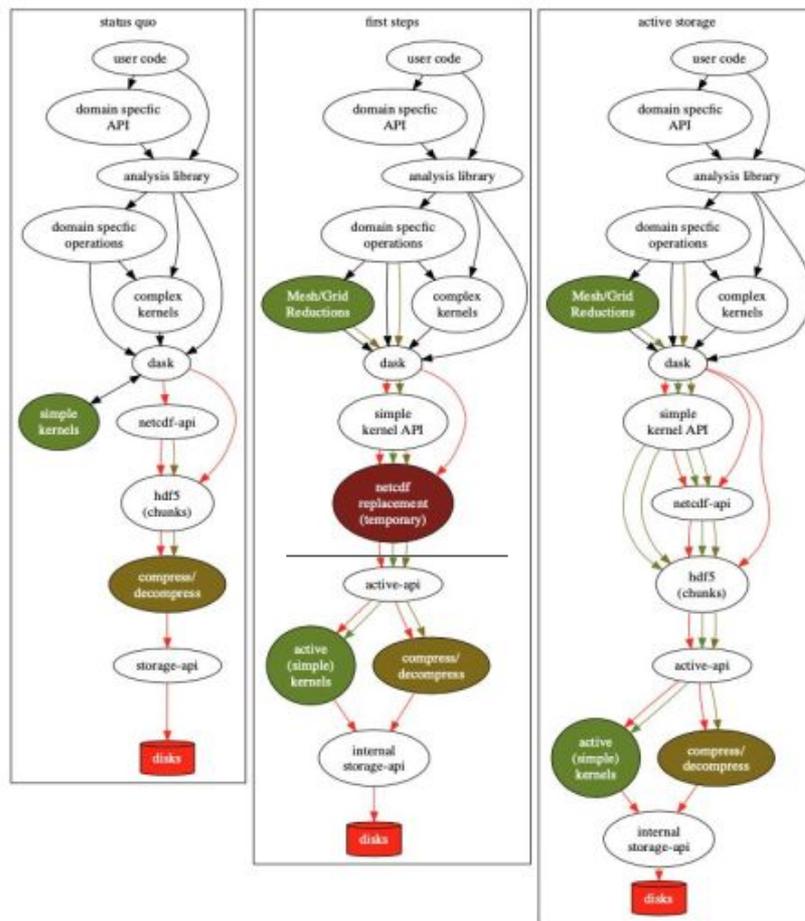
Active Storage -

a computer system architecture which utilizes processing power in disk drives to execute application code.

- Interface design
- Simple reduction(s)
- Storage compute
- Client compute (DASK)

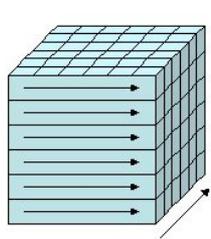
David Hassell, Valeriu Predoi, Stig Telfer, J-T Acquaviva

<https://github.com/NCAS-CMS/ActiveStorage>

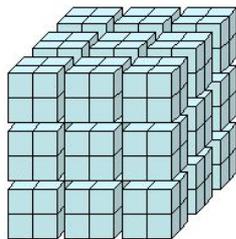


Active Storage Servers

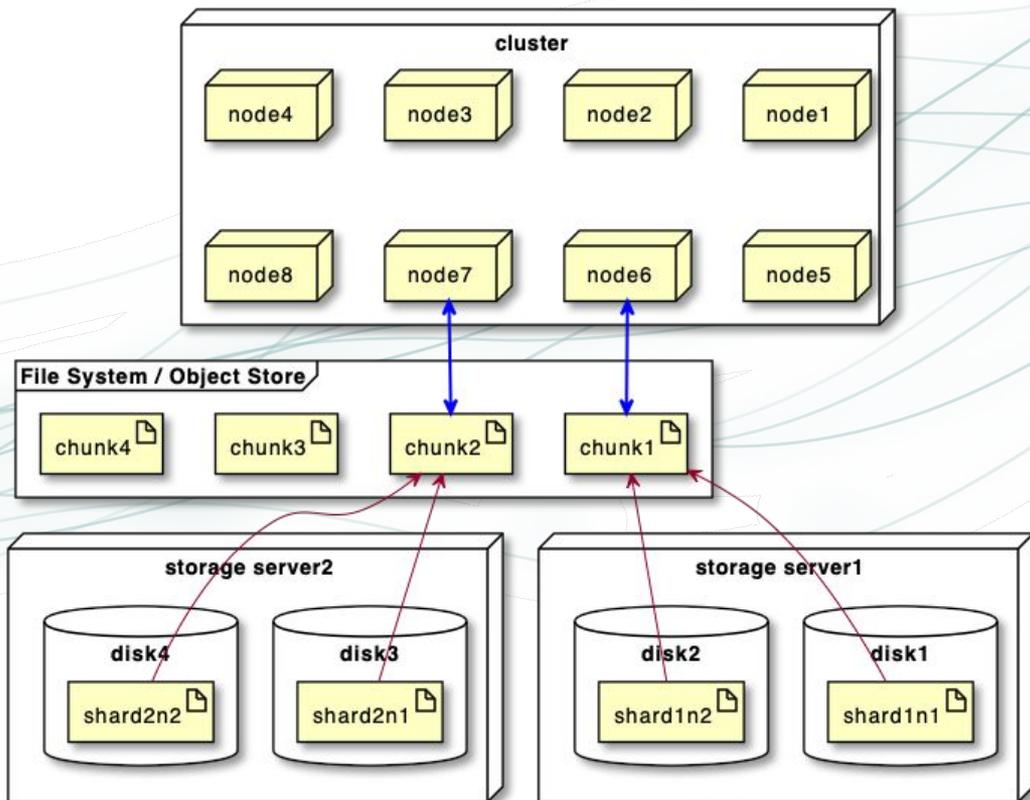
The API talks to “chunks” and needs to work on those



index order



chunked



Two implementations:

S3 (with StackHPC)

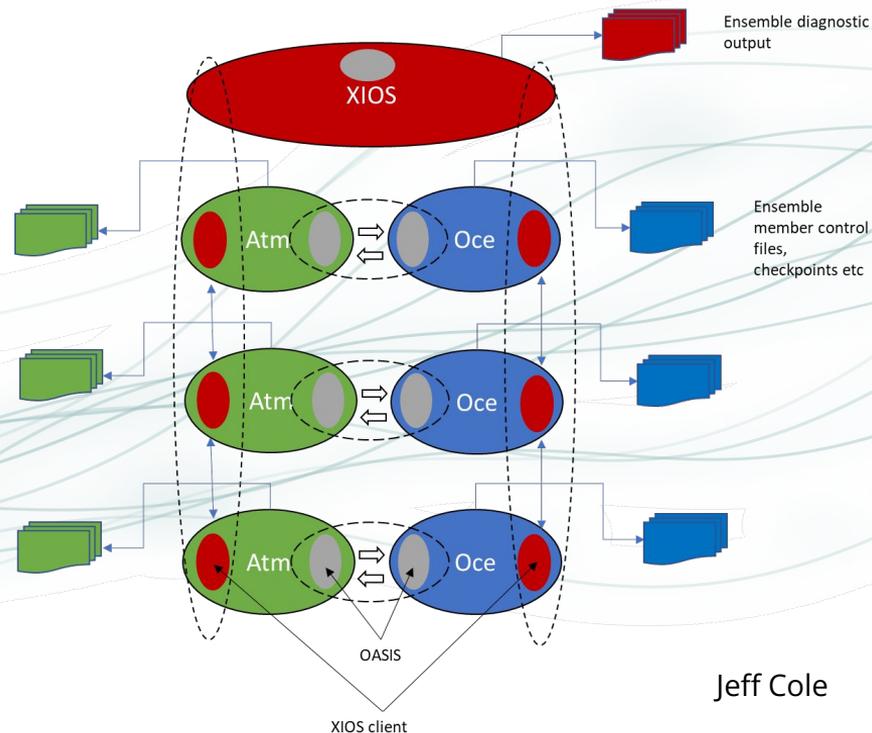
IME/RED (POSIX) (with DDN)

ExCALIWork

Coupled UM Ensemble -

Exploit XIOS capability to deliver *in-flight* ensemble diagnostics

```
<grid id="um-atmos_grid_uv_pfl35225">  
  <domain domain_ref="um-atmos_grid_uv" />  
  <axis axis_ref="um-atmos_pfl35225" />  
  <axis axis_ref="ensemble" />  
</grid>  
<grid id="um-atmos_grid_uv_pfl35225_ensmean">  
  <domain domain_ref="um-atmos_grid_uv" />  
  <axis axis_ref="um-atmos_pfl35225" />  
  <scalar id="um-atmos_grid_uv_pfl35225_ensmean">  
    <reduce_axis operation="average" />  
  </scalar>  
</grid>
```



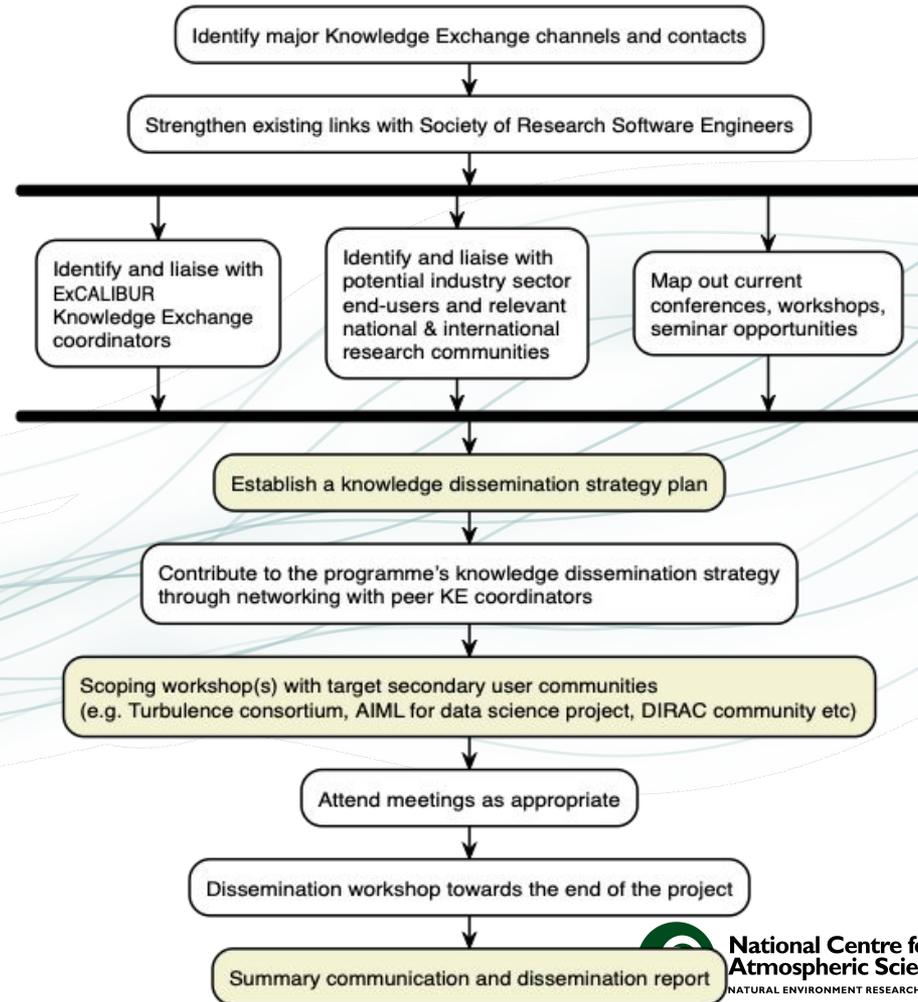
Jeff Cole

Extend to LFRic - incorporate XIOS ensemble developments

ExCALIData

Knowledge Exchange

- Position the next generation of software engineers at the cutting edge of scientific computing
- Ensure integration across the programme activities
- Establish connections with potential beneficiaries in academia, Public Sector Research Establishments and industry



ExCALIData

Watch this space - thanks