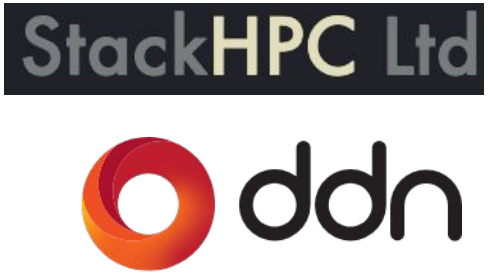
ExCALIData

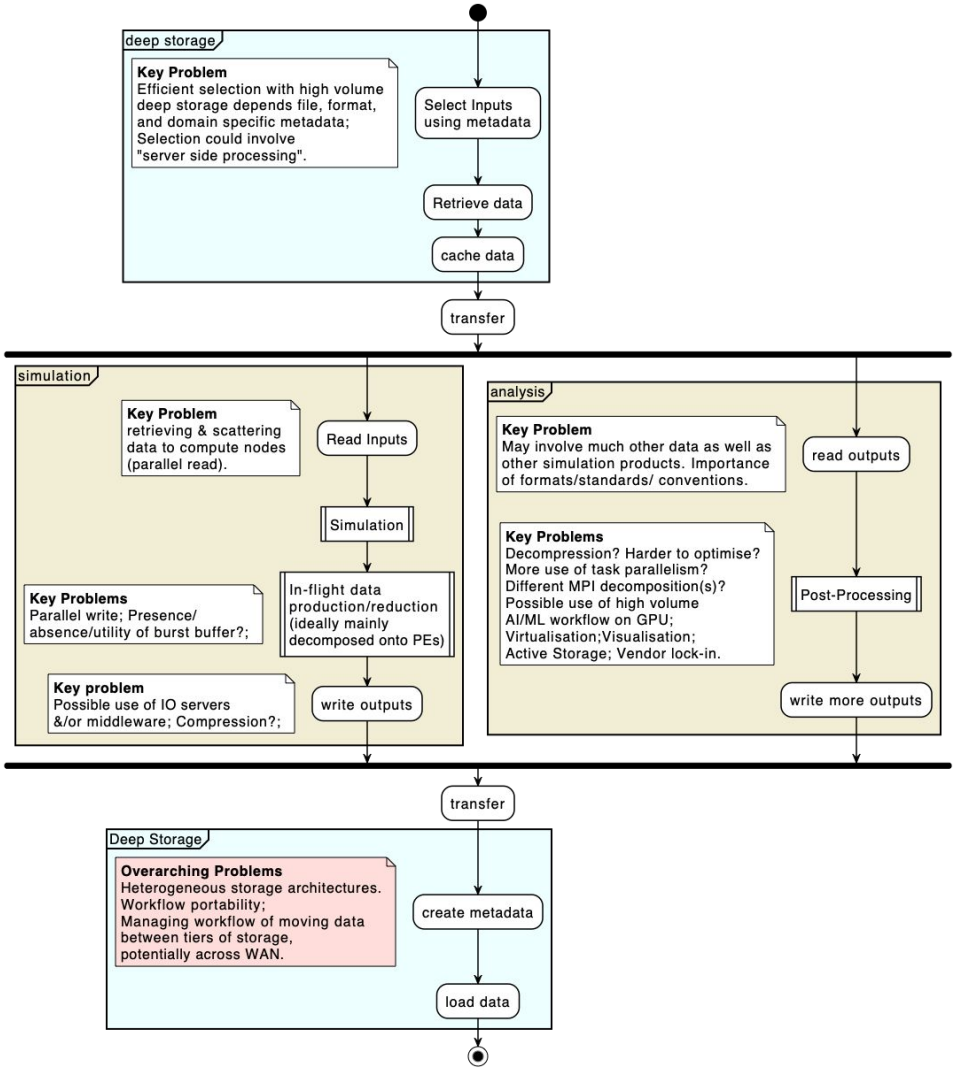
(Two projects:
ExCALIWork and ExCALIStore)



Context

Workflow problem involves

- Selection
- Multiple Tiers of Storage (fast, slow, very slow)
- Multiple Types of Storage (tape, object, posix etc)
- Performance is an issue.
- Compression is needed.
- Lots of technology available, or potentially available, but hard to harness and/or not clearly useful.
- Balance of “difficulty” tilting away from “simulation hard, analysis easy” to “simulation hard, analysis just as hard”.

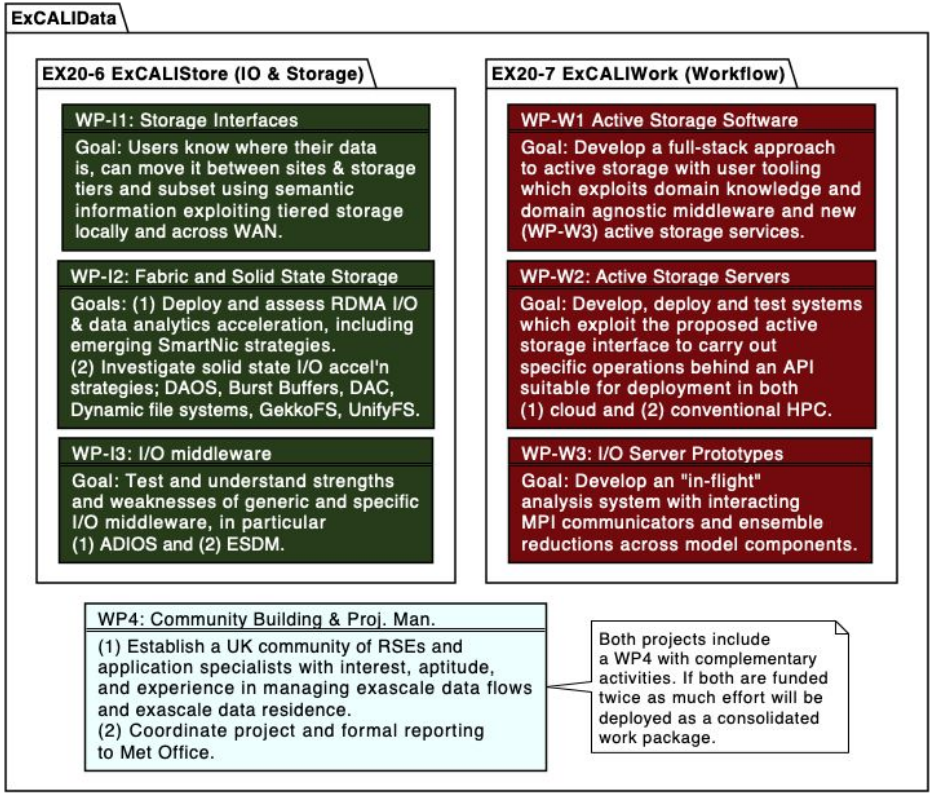


ExCALIdata = ExCALIwork + ExCALIStore

Seven Work Packages; Six distinct activities

1. (WP-I1) Storage Interfaces; Atomic Datasets distributed across storage
2. (WP-I2) Fabric and Solid State Storage; Understanding technology options
3. (WP-I3) Comparing middleware, generic and discipline specific.
4. (WP-W1 and WP-W2) Active Storage in the DASK software stack (W1) and storage (W2)
5. (WP-W3) Extending I/O server functionality
6. (WP4) Community Building

(not including, but not forgetting, project management.)



WP-I1 Storage interfaces and atomic datasets

Problem Statement:

How can an individual user keep track of files held in multiple different storage elements and what is stored in the files, and how could she/he efficiently extract a particular spatio-temporal variable from within those files without having to inspect the contents of those files to find out which files hold which subsets?

Solution:

(involves) an “aggregation abstraction with implementation” which can

(i) apply to any collection of multi-dimensional data which is distributed in hypercube fragments each of which is stored one to a file in such a way that the index coordinate system can be held in NetCDF (or HDF) and

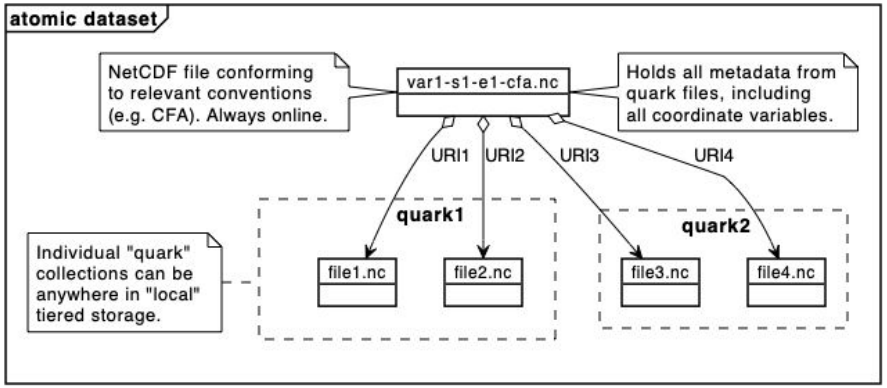
(ii) be used for lazy operations in the full coordinate space without the fragment files being directly available.

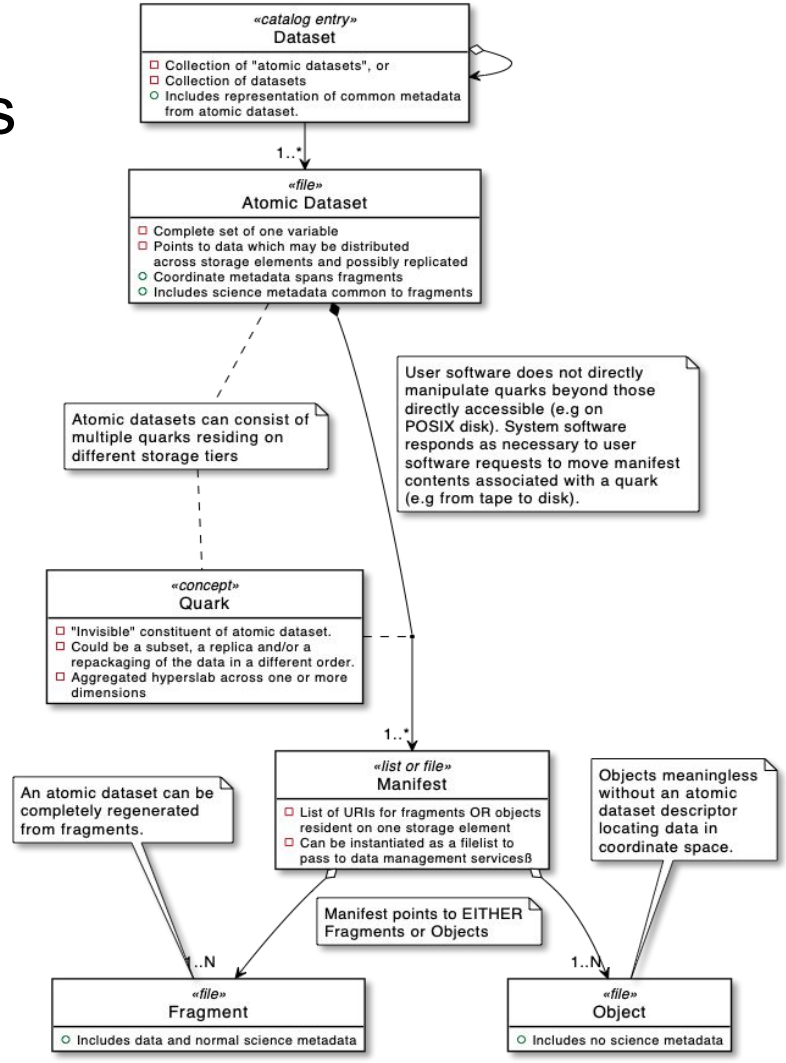
(WP-I1) Persistent Atomic Datasets

We have done a lot of preliminary thinking, and have a clear path as to how to build from an updated set of CF-aggregation rules, to support, fragments which might include:

- NetCDF files,
- Zarr objects,
- GeoTiffs from a SAFE dataset.

We will deliver at least a NetCDF solution here, and the utility of the active storage concept will be demonstrated in W1 (active storage) and a new JASMIN tiered storage system being developed with other funding.





WP-I2 Fabric and Solid State Storage (Cambridge)

WP-I1: Fabric Acceleration

Can we accelerate I/O using Remote Direct Memory Addressing (RDMA) and/or SmartNICs?

Three sub-tasks:

1) Investigating RDMA in synthetic benchmarks
2) Investigating RDMA in real use cases
3) (Investigating real time compression/decompression in SmartNICs.)

WP-I2 Solid State Storage

What are some of the practicalities of ***really*** exploiting solid state storage?

Three sub-tasks: Writing to solid state storage might be easy, but how do we drain them? Are there preferred file system choices?

1) Using remote burst buffers
2) DAOS is supposedly a complete tiered storage solution?
3) Node local ephemeral file systems?

WI2.1-1 and WP I2.1-2 RDMA

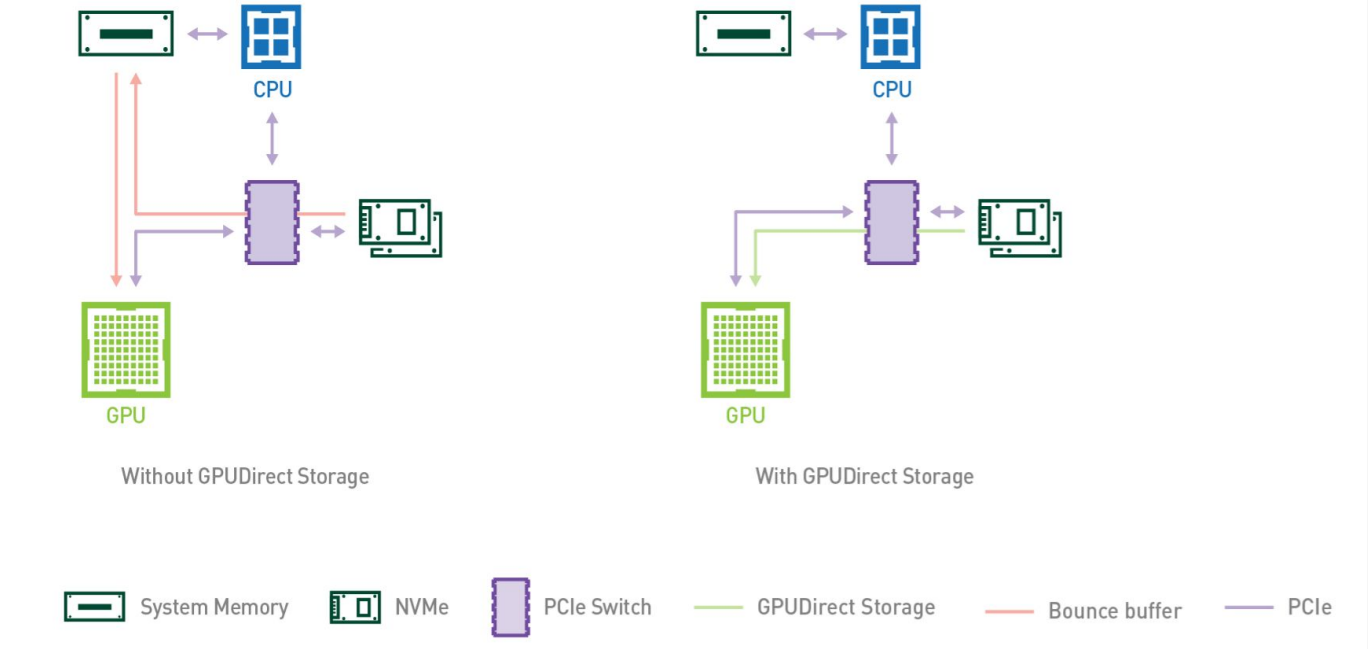
Instead of routing data from storage to system memory and then to the GPU memory (left panel), it is possible to route data direct to the GPU memory with appropriate direct memory access.

Well, yes, but will it make a difference? In

- **Synthetic benchmarks,**

and

- **Real use cases**

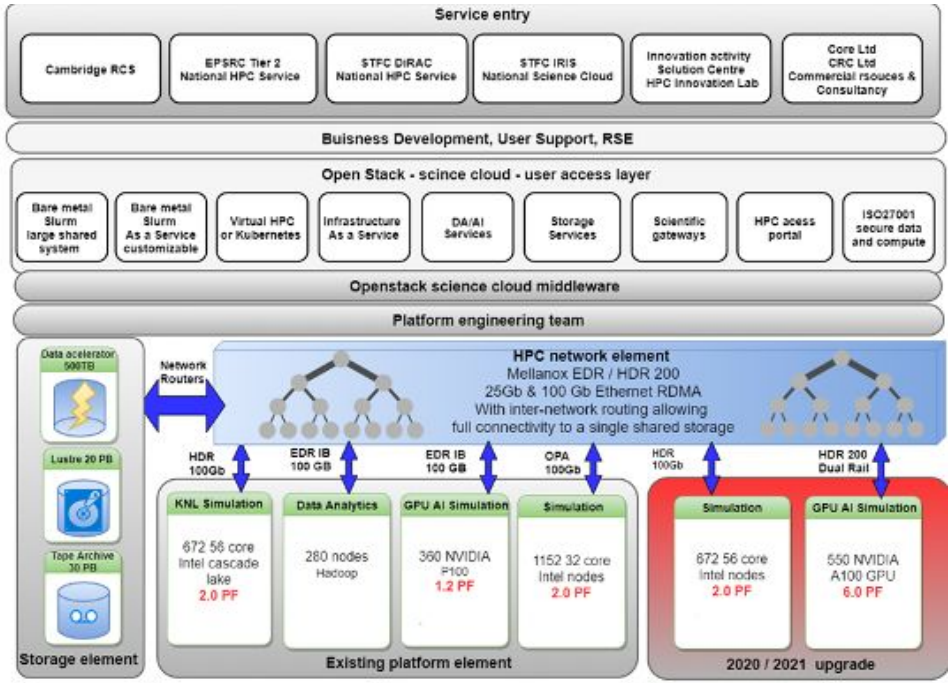


WP-I2 - exploiting CRCS & the Cambridge Solid State Testbed

The CRCS environment is supported by (amongst others) the ExCALIBUR Hardware and Enabling Software project.

CRCS are not being funded to deliver RSE support, but they will also bring significant co-funding that will be directly applied to this project.

We expect WP-I2 to go well beyond what could have been supported without the CRCS testbed and the extra FTE input.



WP-I3: I/O Middleware

ADIOS2 is generic cross-domain I/O middleware being developed in the US exascale programme. The main objective is to facilitate fast I/O for massively parallel jobs in a heterogeneous storage environment. ADIOS also provides some coupling functionality.

ESDM is I/O middleware developed for application in earth system modelling. It is specifically developed for fast I/O in weather and climate applications.

The objective of this work package is to compare and contrast the ease of use of this middleware in weather and climate and fusion use cases. There will be a close relationship between this and WP-W3 where the XIOS sub-system is being used, and we will compare and contrast all three. (XIOS has other capabilities, but we will also look at the ease-of-use and performance in “I/O middleware” mode.)

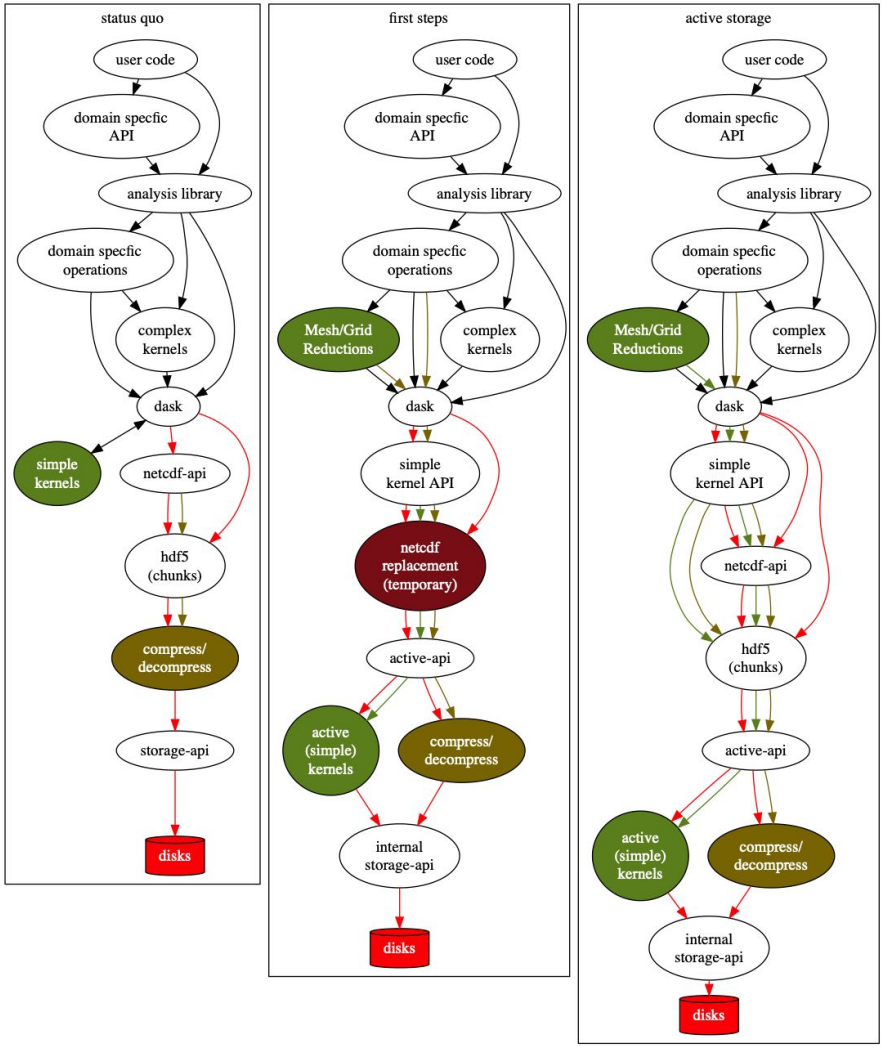
WP-W1;W2: Active Storage

Google says Active Storage is a computer system architecture which utilizes processing power in disk drives to execute application code.

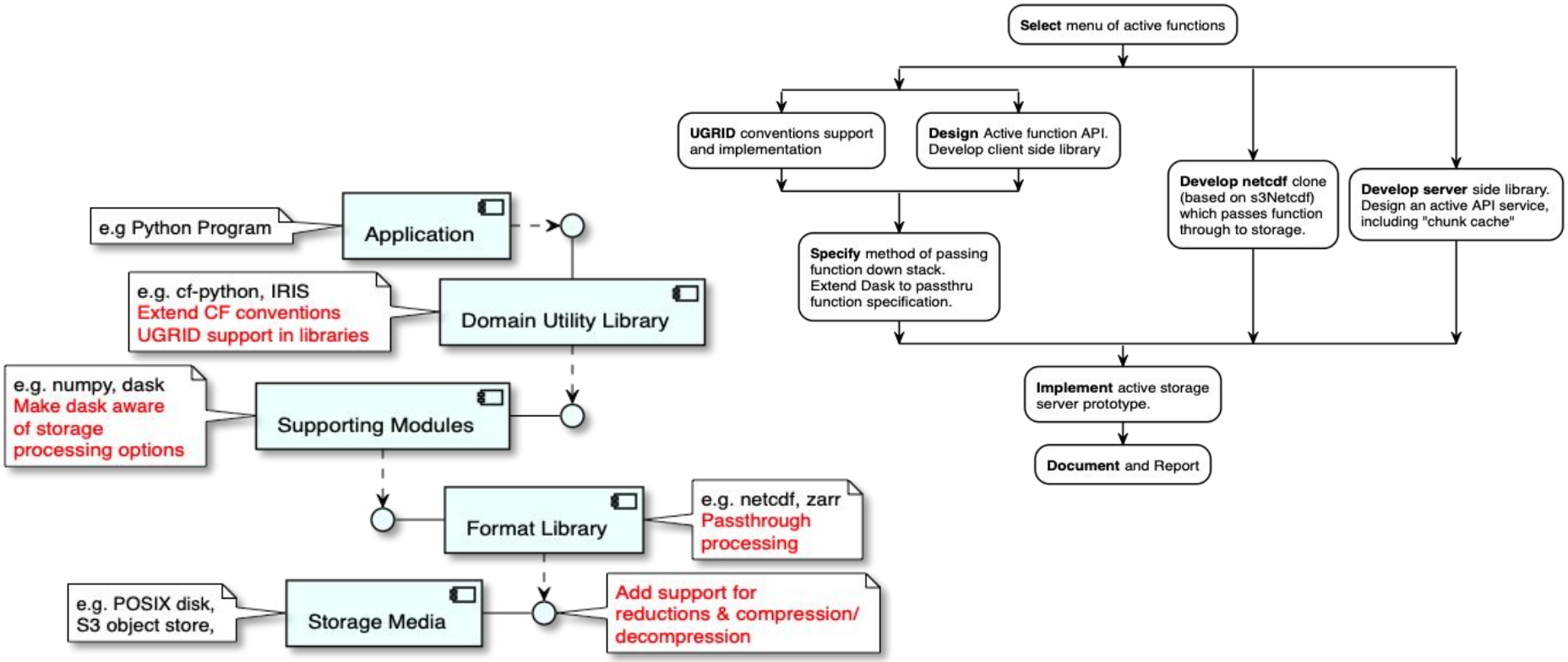
In practice Active Storage needs to consist of a middleware (W1) and server stack (W2) that supports application code in doing data reductions in the storage subsystem.

Why not “arbitrary application code” in the storage subsystem? If for no other reason (e.g. security): most of the algorithmically available parallelism will be in the massively parallel compute system.

In practice we will work using the Dask API, the dask-distributed concept for fragmented arrays and the HDF concept of chunking to push reductions down onto chunks in storage.

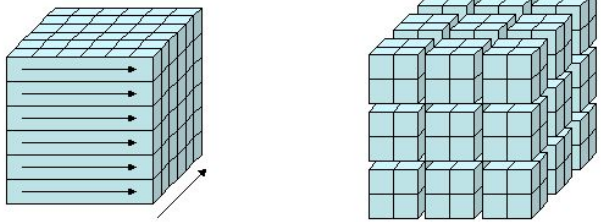


W1 Active Storage Software



W2 Active Storage Servers

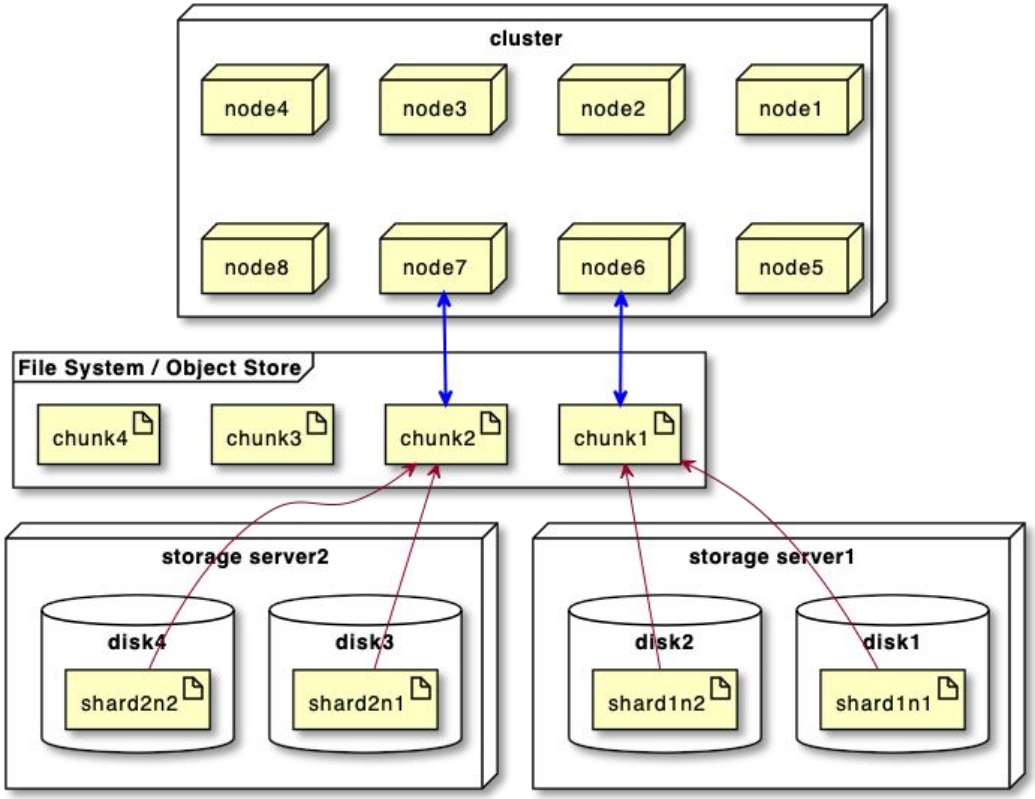
The API talks to “chunks” and needs to work on those



Two implementations:

W2-1 S3 (with StackHPC)

W2-2 IME/RED (POSIX) (with DDN)



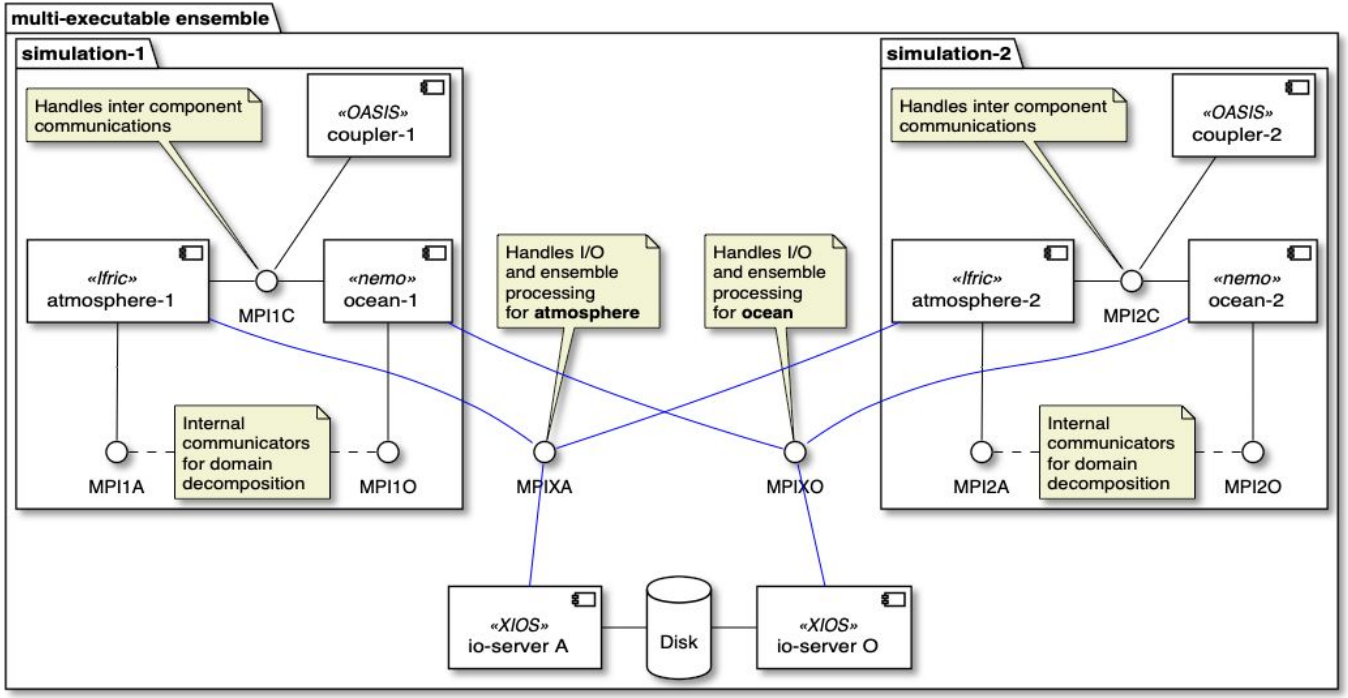
WP-W3: I/O Server Prototypes

We know how to do this for a UM atm-only system.

- But we need to improve the range of diagnostics.

In W3 we will

- Extend to a UM coupled system
- Investigate LFric ensembles.
- Extend to NGMS (LFRic+NEMO)
- (Compare with I3 results for I/O performance.)



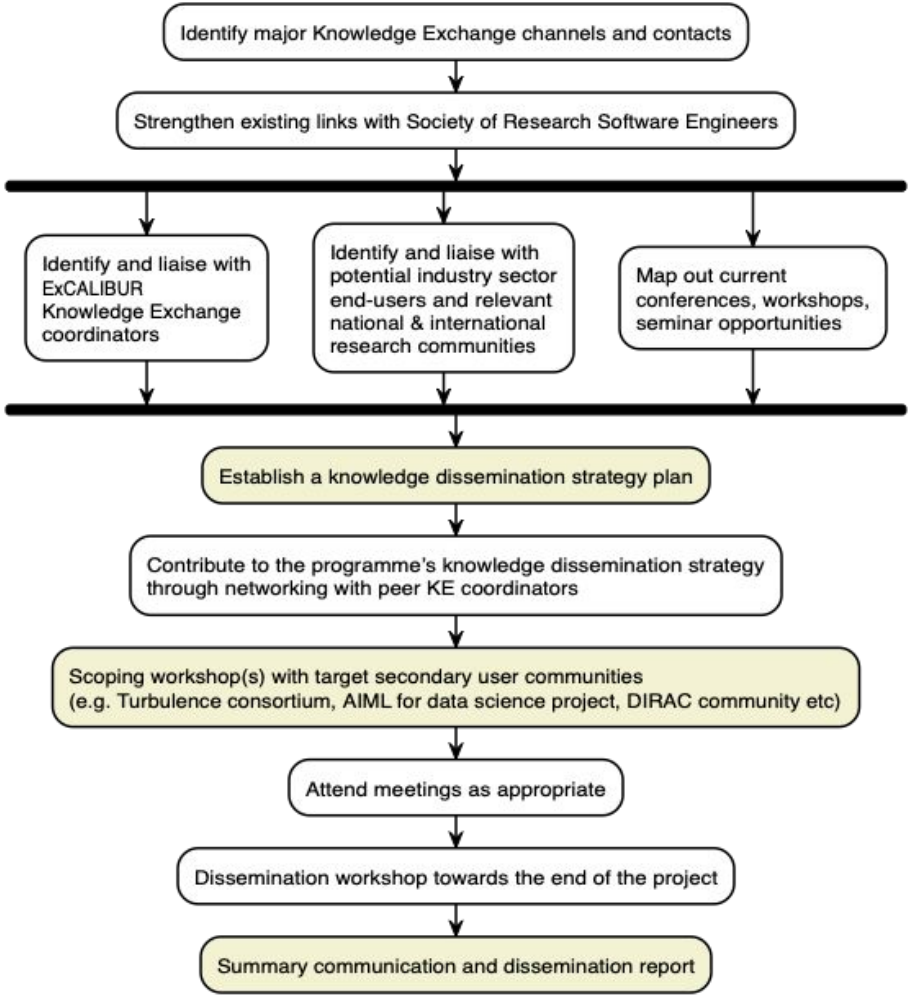
WP4: Knowledge Transfer

First milestone is to develop a plan!

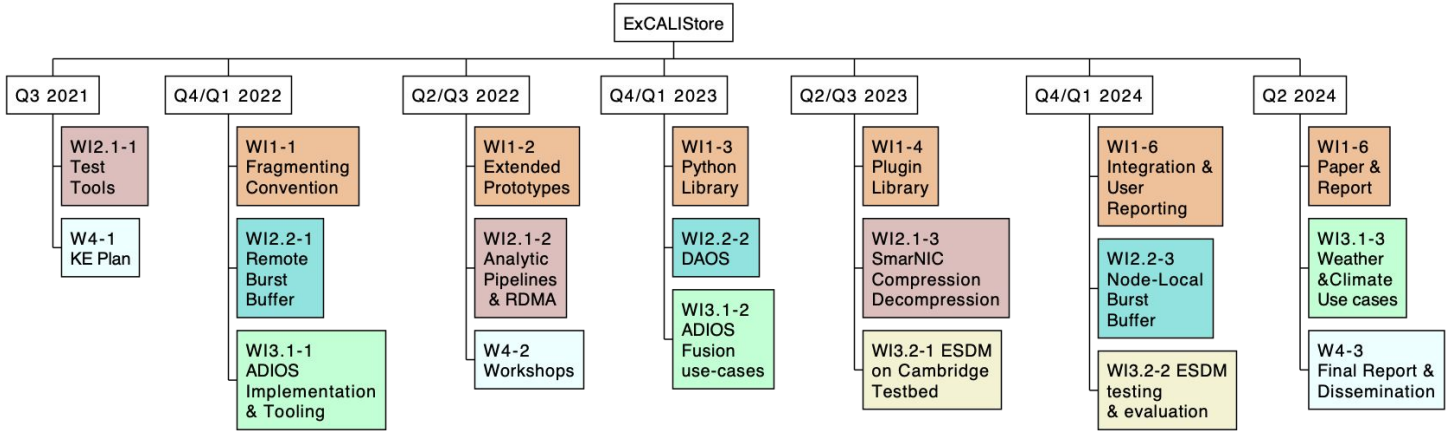
Initial work will be undertaken by Dr Fanny Adloff (ENES Scientific Officer); but we expect to entrain other RSE staff.

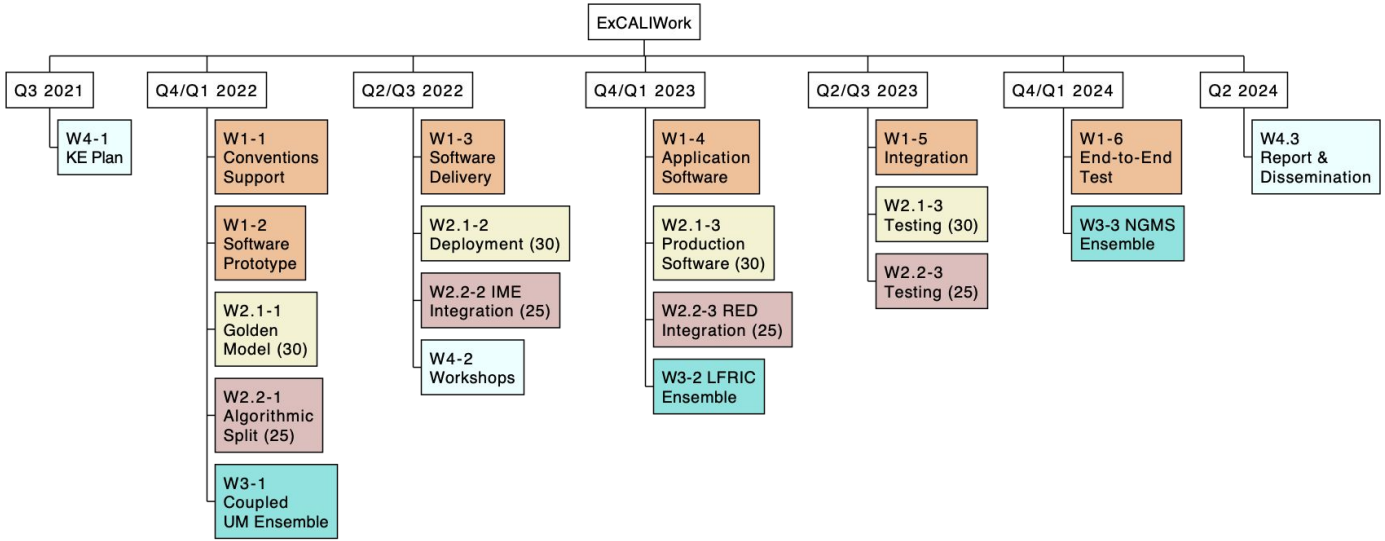
We will also coordinate reaching out to the other use-cases for input to the various WP through this WP.

Fanny and Grenville, with provide initial support to meet Met Office reporting requirements.



Scheduling





Initial Priorities:

- KE Plans
- Sub-Contracts
- Getting staff and tooling sorted in Cambridge
- CF fundamentals for both projects