# WP4: Data Systems At Scale (Review Update)

Bryan Lawrence, Julian Kunkel *et.al.*
NCAS & University of Reading, UK

21 October, 2020

WP4 Partners:
CNRS-IPSL, CMCC, DDN, DKRZ, METO, Seagate, STFC, UREAD

# WP4: Data Systems at Scale

## Objectives

*to mitigate the effects of the data deluge from high-resolution simulations (project objective-d)*

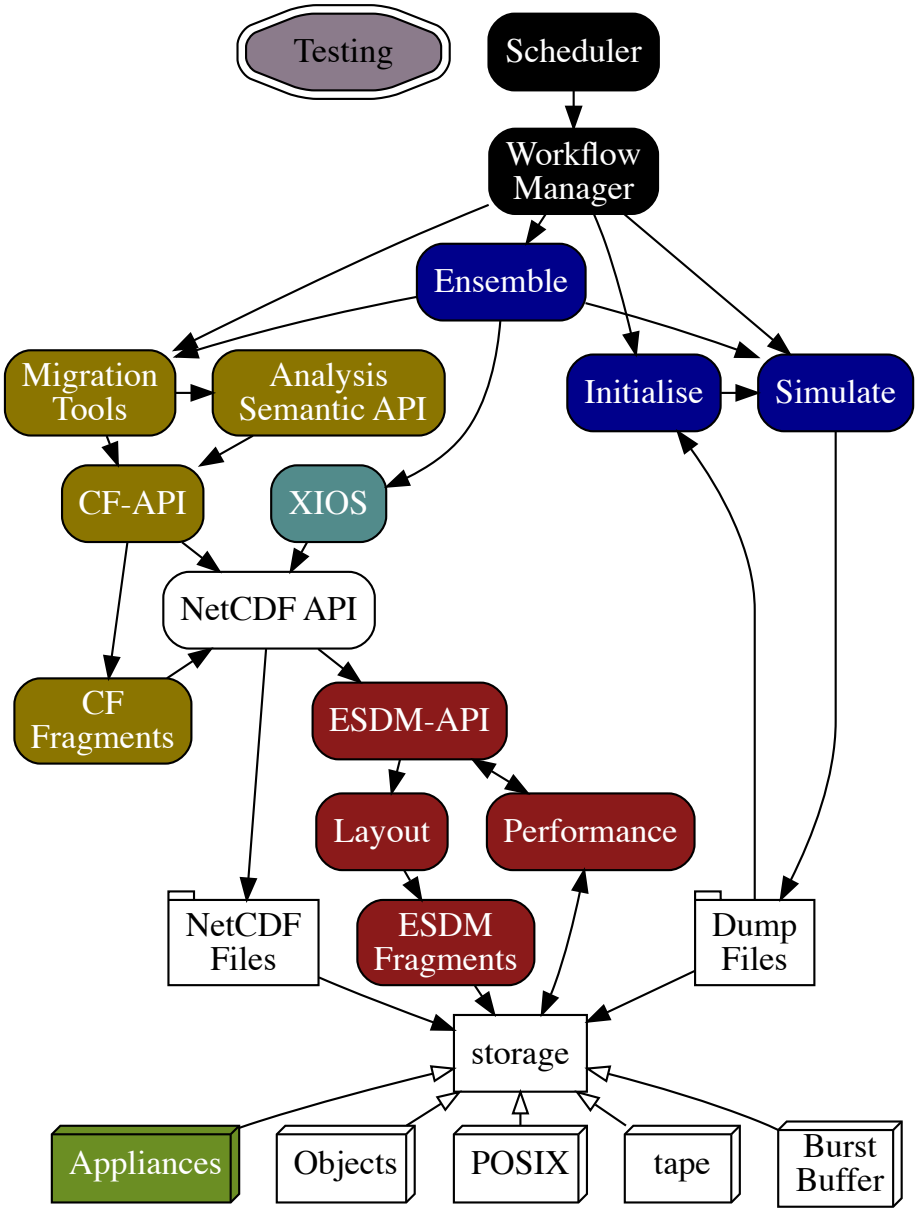
by

1. Supporting data reduction in ensembles by providing tools to carry out ensemble statistics “in-flight” and compress ensemble members on the way to storage, and
2. Providing tools to:
   1. transparently hide complexity of multiple-storage tiers (middleware between NetCDF and storage) with industrial prototype backends, and
   2. deliver portable workflow support for manual migration of semantically important content between storage on disk, tape, and object stores.

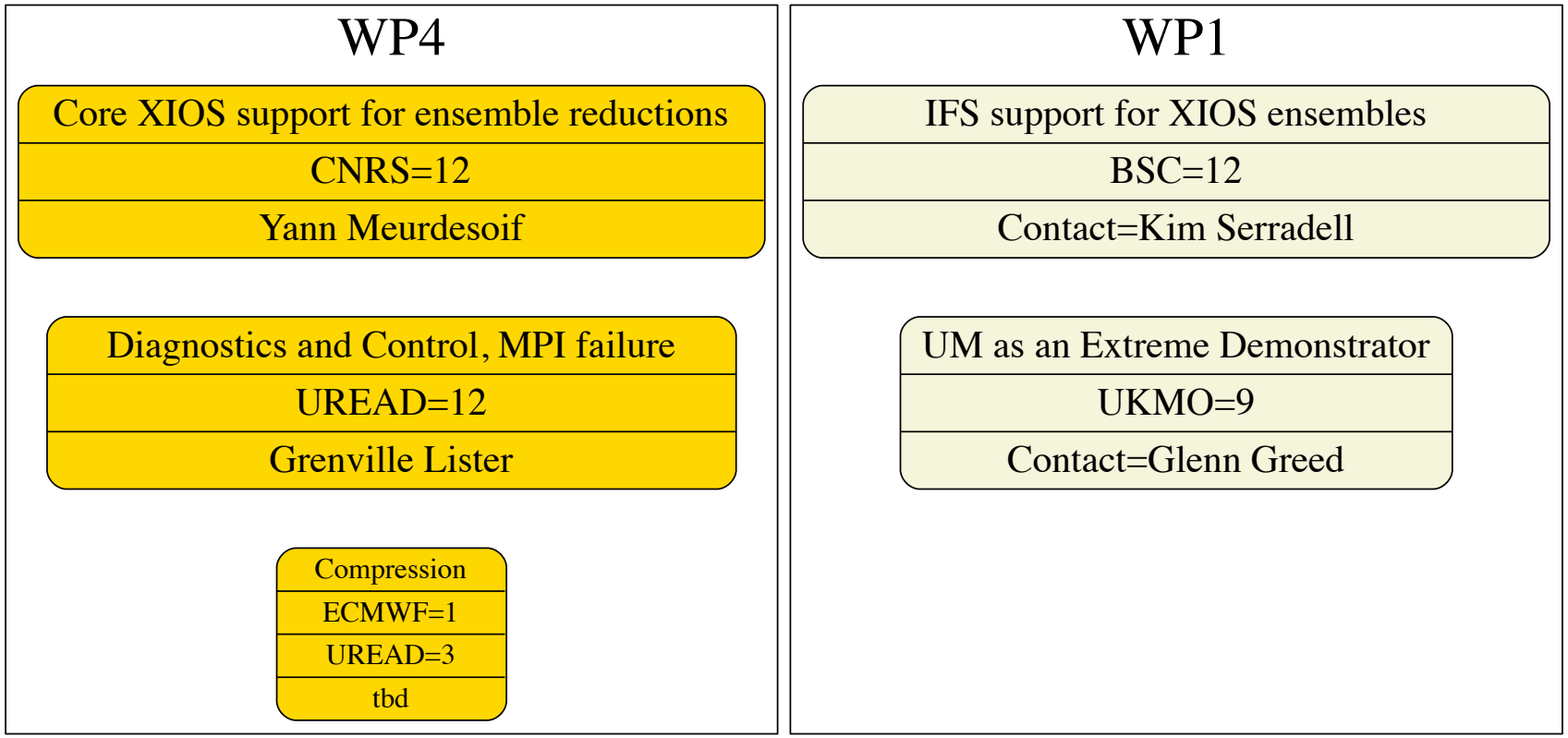
*ensemble tools, storage middleware, storage workflow*

# Components/Tasks in WP4

- 4.1 Leadership and Design: 12 PM
- 4.2 Ensemble Services (in flight analysis/compression)
- 4.3 Earth System Data Middleware (ESDM) - performance in HPC simulation.
- 4.4 Semantic Storage Tools (SemST) - userspace tools for handing volume.
- 4.5 Workflow Support (enhancements to SLURM/cylc)
- 4.6 Component and End-to-End Testing
- 4.7 Industrial Proof-of-Concept Appliances.

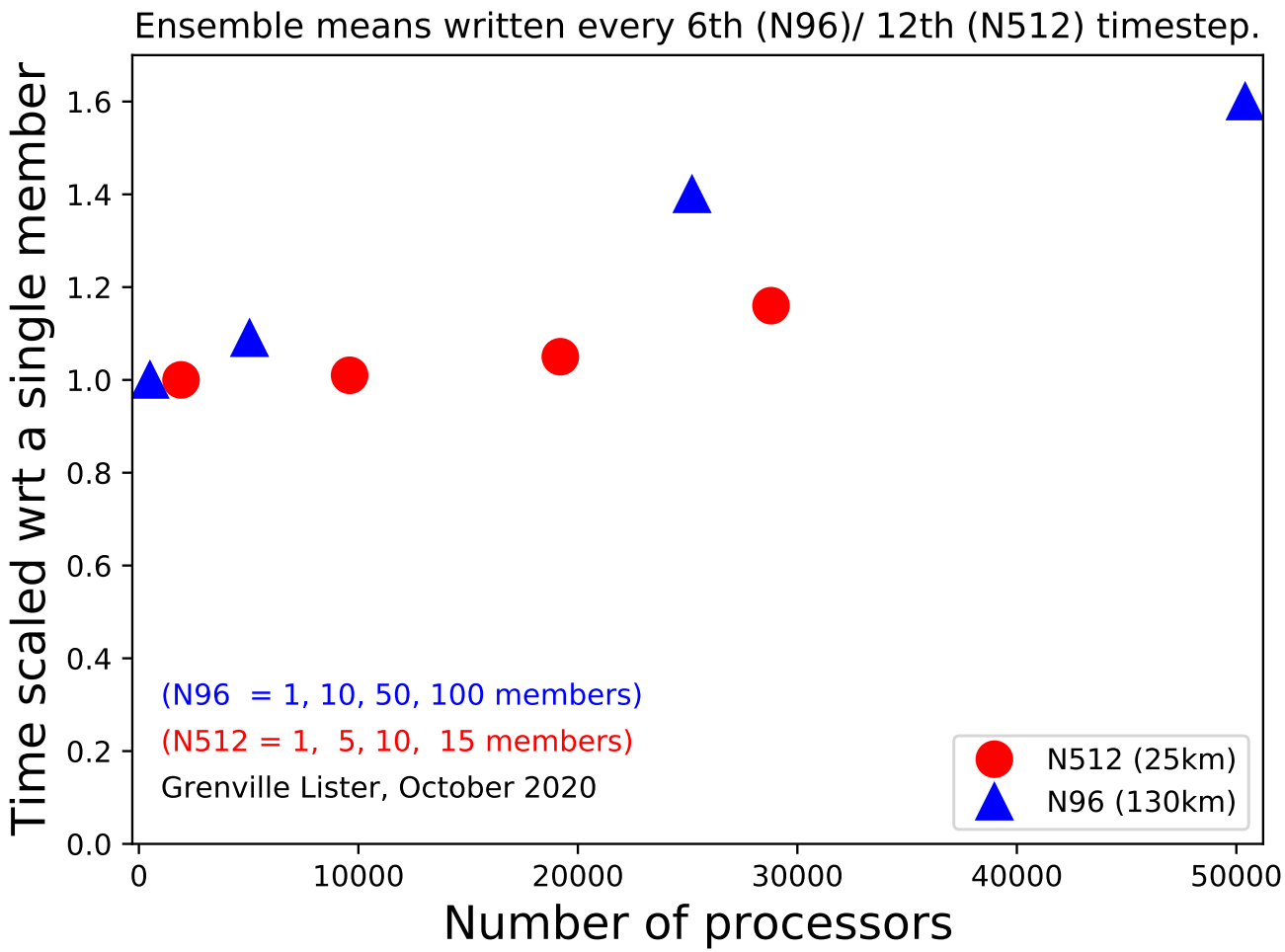


# Task 4.2: Ensemble Services




XIOS and Ensembles

# Task 4.2: Ensemble Services - Performance (Lister, Oct 2020)




- Ensemble processing with XIOS scales very well!
- Real potential for reducing need for all data from all ensemble members to be stored!

# Task 4.2: Ensemble Services - XIOS integration (Coles, Oct 2020)

## Integration Services

Thus far:

- Developed a standalone XIOS testbed.
- Developed an XIOS-ESDM test suite (running on JASMIN)
- Integrating XIOS with *Coupled* UM system.
- Integrate support for output on a reduced Gaussian grid.

Next Steps:

- Integrate XIOS CMIP6 configuration with Rose-CYLC.
- Extend XIOS ensemble configuration to a coupled model system.

# Task 4.2: Ensemble Services - Error Handling (Wilson, Oct 2020)

## Ensemble Failure - Working Example

- Deferred handling within MPI, moved to managing within CYLC.
- System shares MPI communicator. Trap and signal exceptions from running (UM) model to CYLC.
- Stop ensemble. Reconfigure ensemble averaging.
- Restart from last checkpoint without bad ensemble member.

## Next Steps

- Currently handles only one failure. Extend to many.
- Instead of removing member, allow restart reconfiguration
- At end of each cycle, run analysis code via workflow and pro-actively trigger member removal via preconfigured criteria.
- (With WP1) Investigate IFS error trapping and restart options.

# XIOS Futures (Meurdesoif, Oct 2020)

## IS-ENES3 Underpinning

- …2018: XIOS very successful in support of CMIP6, but …
- Very large unwieldy code base, so …
- 2019-2020: IS-ENES3 investment in robustness and reliabilty. Code overhaul. Reducing memory footprint. New architecture. Test suite, and more.
- Rewrite of transfer, and transformation filters.
- Stable version expected mid 2021.

## ESiWACE2 Futures

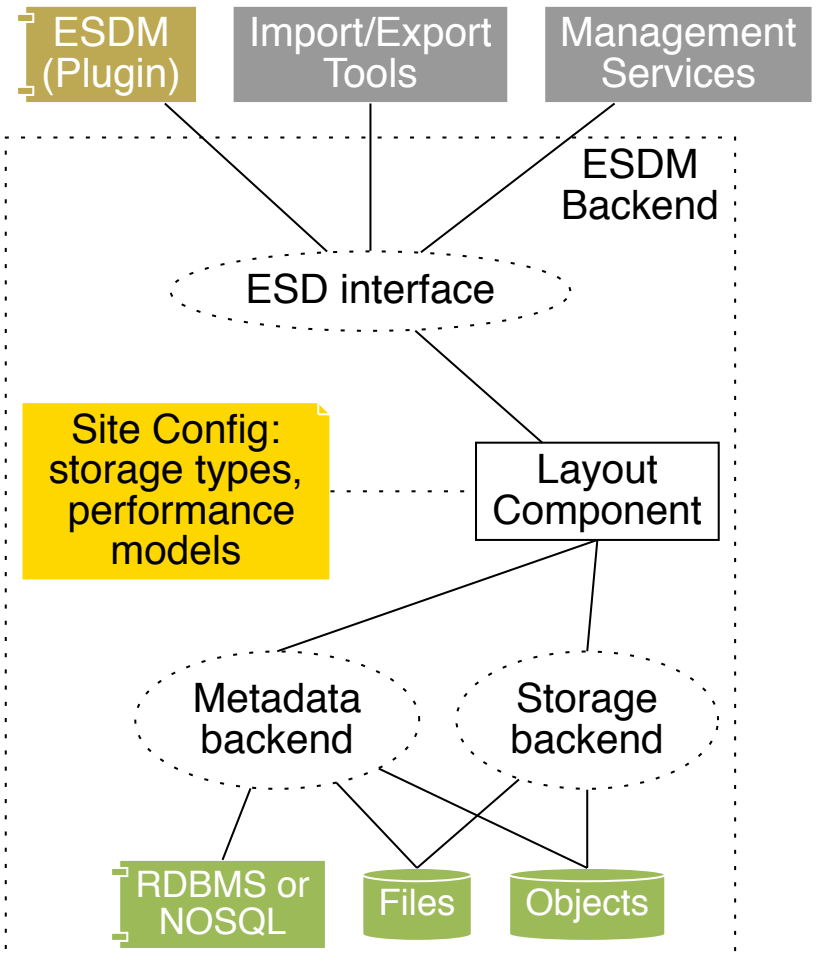
2021+

- Support for XIOS restartability.
- Move to ensemble support with model members on their own communicator.
- Dynamic connection from enesmble emmbers to an ensemble service.
- Ensemble reductions in server.
- ...

all complementing and extending the work done with previous versions.

# ESD Middleware, 42PM

## Integration, Hardening, Enhancement




1. Native NetCDF (bypass HDF, UREAD 9PM)
2. Harden, Optimise (DKRZ, 6PM)
3. Improve Performance Model Component (DKRZ, 6PM)
4. Compression Enhancements (UREAD, 6PM)
5. Backends:
   - DDN (3PM)
   - Seagate (3PM)
   - Ophidia (CMCC, 3PM)
   - S3 (STFC, 6PM)

# ESDM Interfaces

## NetCDF (UoR)

(Requirement: provide a NetCDF interface to ESDM.)

- **esdm-netcdf** library released in April 2020 (on github).
- Main NetCDF functionality supported, except for groups and compression.
- Includes test suite which covers expected functionality.

(J. Kunkel and L. Pedro)

## Ophidia (CMCC)

(Ophidia Requirement: deliver transparent data load/store over heterogeneous storage devices and libraries.)

- Ophidia can now import/export via the **e**sdm-netcdf library.
- Direct ESDM import/export planned over next few months.

(D. Elia)

# ESDM Interfaces (cont)

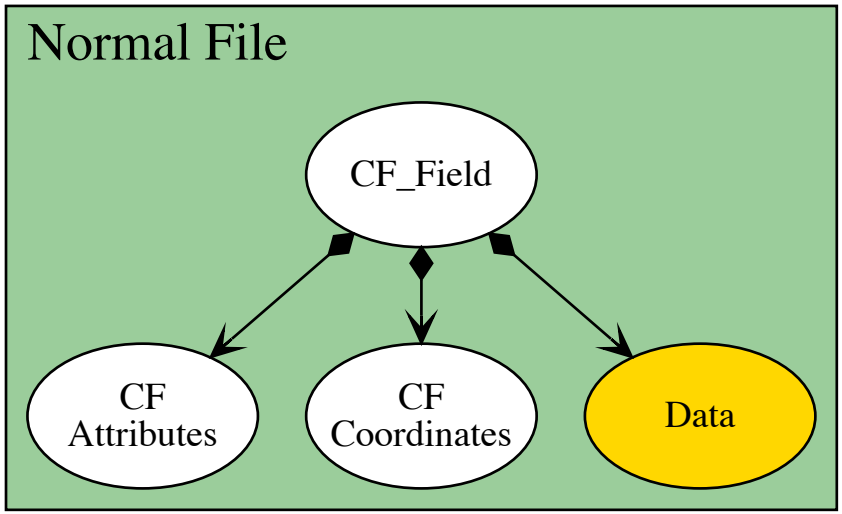
## Seagate: Clovis

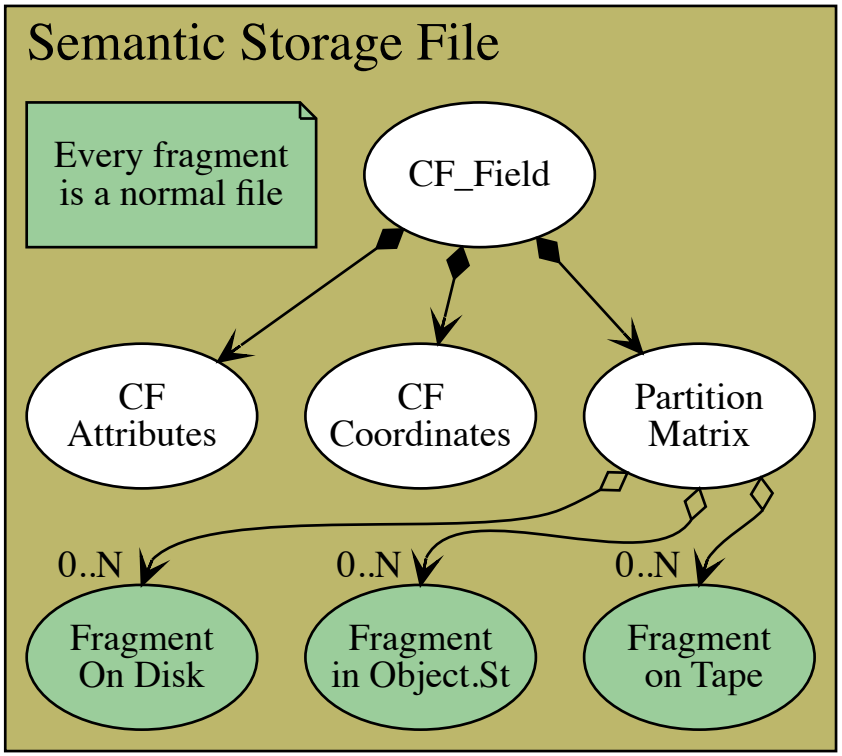
- Clovis backend updated to latest ESDM interfaces, performance optimisation underway. Multiple internal adaptations (metadata improvements, storing segments in single Clovis object).
- Investigating adaptive parameters for specific I/O requirements and patterns.

## Other

- Experiments with generic object stores an technology (e.g. zarr) also underway.
- (ESDM/XIOS test suite already mentioned).

# Semantic Storage (for weather and climate)




Build on CF Data Model
& CF Aggregation Framework
https://doi.org/10.5194/gmd-10-4619-2017
& http://www.met.reading.ac.uk/~david/cfa/0.4/

## Using semantic storage files will depend upon:

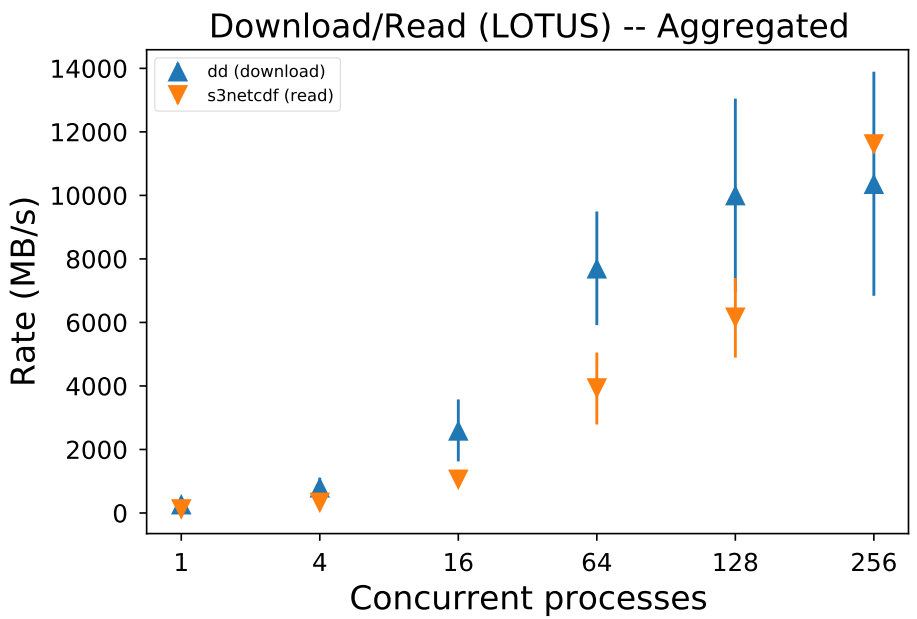
- Tools for manipulating location of fragments (from LAN to WAN)
- Tools for doing science with (local) semantic storage directly

# October 2020 Update

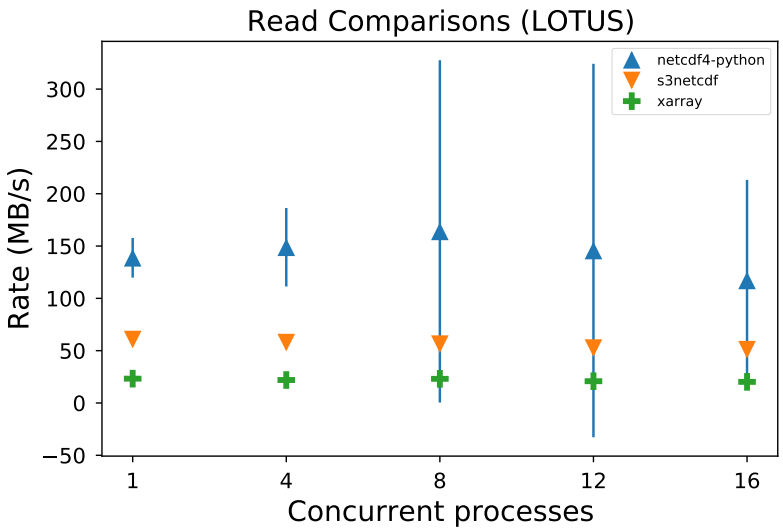
## Progress:

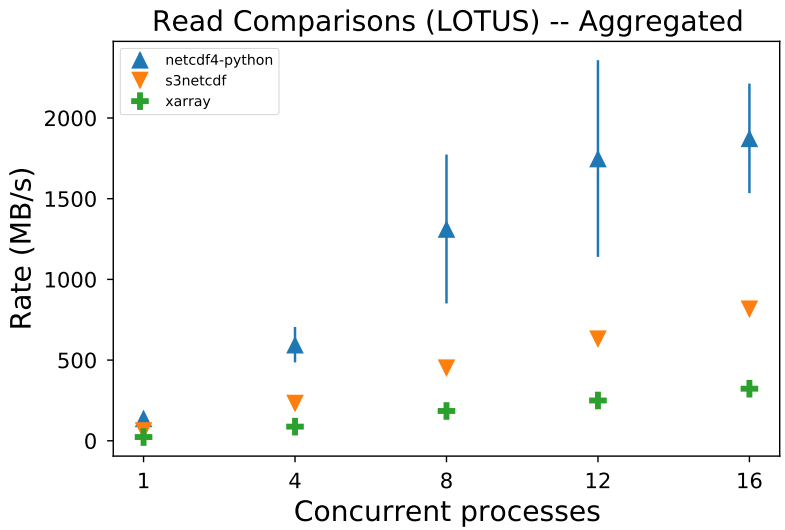
- S3NetCDF4 completely refactored, V2 released 21/08/2020, already up to 2.0.3.
- Performance being evaluated a) in ideal test cases, and b) in CMIP6 evaluation exercise (aiming for results for AGU 2020).
- Next steps; considering dask integration, additional zarr backend, integration with JDMA, interaction of chunk size with fragment size.

- Excellent aggregate performance reading data out to 250+ processes.
- (Massey, Jones, Lawrence)



# S3NetCDF comparative performance




- Comparison of netcdf4-python, s3netcdf, and xarray: all three relatively flat in performance for parallel reads out to 16 processes, with good aggregate outcomes.
- netcdf4-python is reading from POSIX files on Quobyte storage. S3netcdf is reading from S3netcdf files fragmented into Caringo high performance object storage. xarray is reading from zarr objects in Caringo storage.
- It is not worth over construing the absolute numbers (nc4 benefiting from unrealistic caching, object size can influence the others), but the relationships seem robust.

# Task 4.7: Industry Proof Of Concept (Oct 2020)

## DDN POC

Concentrating on function shipping (compute into active storage)

- Step One: client only support: client takes a function as input, requires byte offset in a file. Client performs the read request, fetches data from the server, applies function and returns result. (This step will be implemented and available by end of Autumn 2020.)
- Step Two: server support: client sends function to server, server performs read, applies function and returns results to client. (Schedule TBD).

## Seagate POC

Previously concentrating on ESDM backend

- Mero is now fully open-sourced ( allowing POC to be deployed on bare-metal hardware (servers and drives).
- Plans underway for deployment closer to application.
- Initial discussions underway for methodology of handling function shipping.

Mero: https://github.com/Seagate/cortx

# Summary

- WP4 is about scale - performance and volume.
- Several activities at various places in the stack.
- Significant progress on all fronts:
  - T4.2 ensemble handling,
  - T4.3 earth system data middleware (ESDM),
  - T4.4 semantic middleware — S3NetCDF and JDMA (JDMA not discussed today), and
  - T4.7 industry proof of concepts.
- More activities still to spin up during the next reporting period.

The projects ESiWACE and ESiWACE2 have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements numbers **675191** and **823988**.

esiwace2

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

*Disclaimer: This material reflects only the view of the author(s) and the EU-Commission is not responsible for any use that may be made of the information it contains*