WP4: Data Systems At Scale

Bryan Lawrence and Julian Kunkel

NCAS & University of Reading, UK

Hamburg, 12 March, 2019

WP4 Partners:

CNRS-IPSL, CMCC, DDN, DKRZ, METO, Seagate, STFC, UREAD

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

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

WP4 Philosophy and Methodology

Maximum Impact from a Minimum Change Surface

Solutions (tools addressing the data deluge), need to maximise their impact on data handling by

- Minimising the impact of increasing volumes of data, particularly within large-scale ensembles and/or high resolution runs, while
- minimising interference with existing working practice and codes, and
- minimising requirements of the system environment.

Methodology

- Modify existing tools,
- Develop a minimum of new tools and,
- (where possible) Exploit middleware which can be deployed in userspace, but hide complexity from end-users
- (but at the same time, where we do have access to the system) If appropriate, deploy new services and appliances.

Components/Tasks in WP4

4.1 Leadership and Design: 12 PM

4.2 Ensemble Services (in flight analysis/compression)

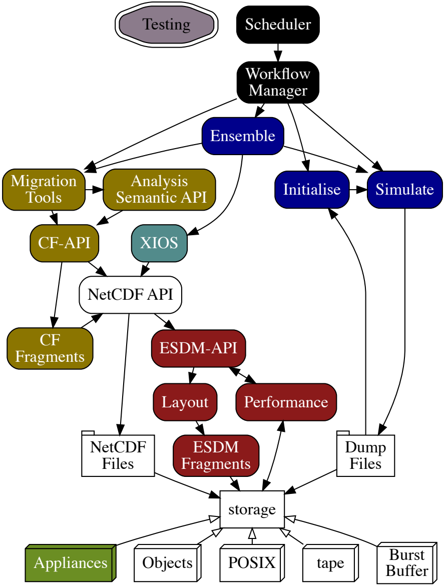
4.3 Earth System Data Middleware (ESDM) - performance in HPC simulation.

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

Core XIOS support for ensemble reductions
CNRS=12
Yann Meurdesoif

IFS support for XIOS ensembles
BSC=12
Contact=tbd

Diagnostics and Control, MPI failure
UREAD=12
Grenville Lister

UM as an Extreme Demonstrator
UKMO=9
Contact=tbd

Compression
ECMWF=1
UREAD=3
tbd

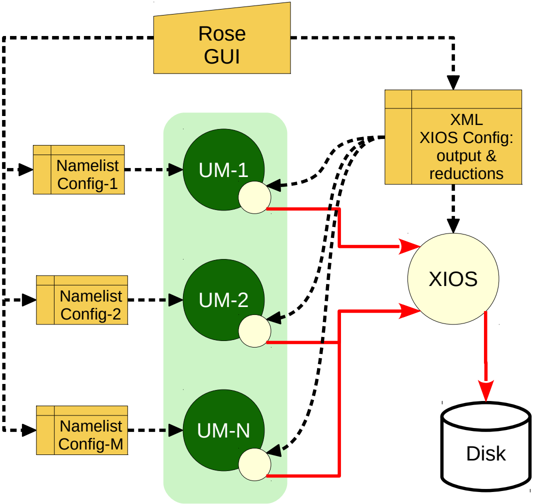
In-Flight Parallel Data Analysis

An ensemble is a set of simulations running different instances of the same numerical experiment. We do this to get information about uncertainty.

Dealing with too much ensemble data

Instead of writing out all ensemble members and doing all the analysis later:

- Calculate ensemble statistics on the fly.
- Only write out some ensemble members.
- (Which ones? A tale for another day, see Daniel Galea’s Ph.D work.)



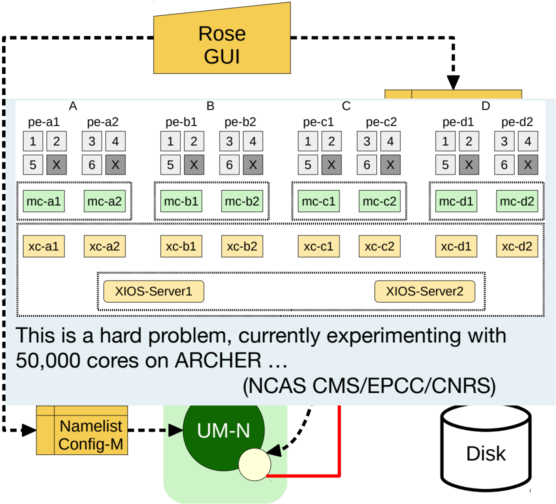
In-Flight Parallel Data Analysis

An ensemble is a set of simulations running different instances of the same numerical experiment. We do this to get information about uncertainty.

Dealing with too much ensemble data

Instead of writing out all ensemble members and doing all the analysis later:

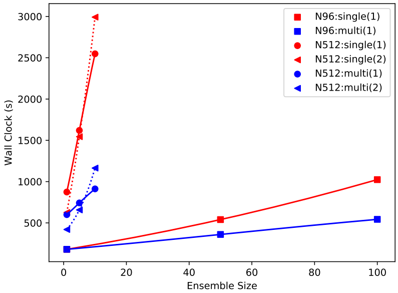
- Calculate ensemble statistics on the fly.
- Only write out some ensemble members.
- (Which ones? A tale for another day, see Daniel Galea's Ph.D work.)

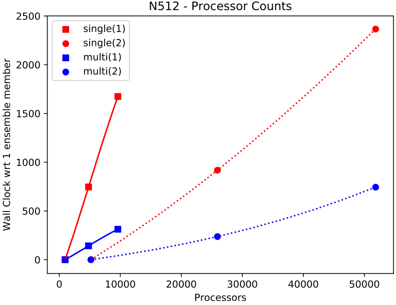


This is a hard problem, currently experimenting with 50,000 cores on ARCHER …

(NCAS CMS/EPCC/CNRS)

Early Results

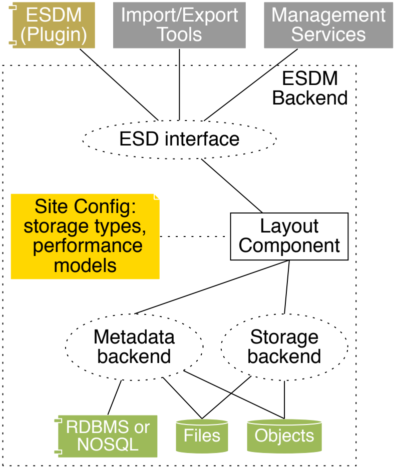




Good results with low resolution, poor with high resolution, but we know where these problems come from. Given the high resolution would output 0.5 PB/model-year, these could well stress any exascale platform.

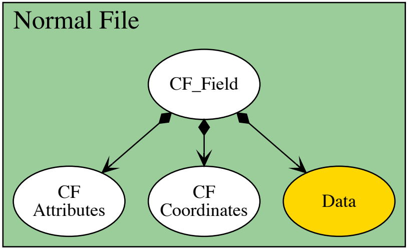
ESD Middleware, 42PM

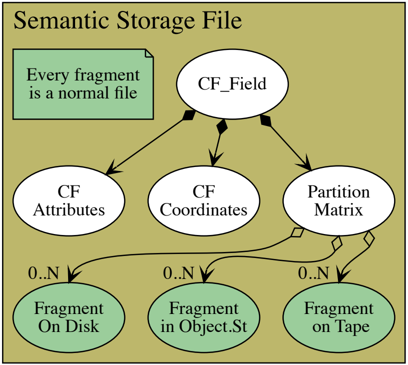
Integration, Hardening, Enhancement



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)

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:

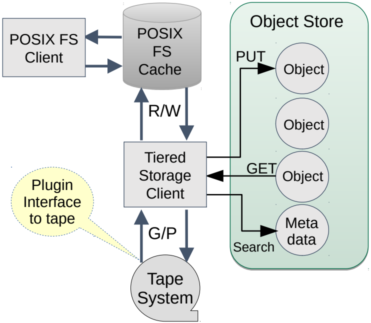
- Tools for manipulating location of fragments (from LAN to WAN)
- Tools for doing science with (local) semantic storage directly

Task 4.4: Semantic Storage Tools

Goal: **Userspace** tools which understand CF semantic storage

Two specific tools

- **Manipulating location of fragments**: evolve the ESiWACE Joint Data Migration App (JDMA) to support ”CF-Interface” and LAN-to-WAN.
- **Science with Semantic Storage**: Middleware library to exploit parallelism and semantics. Support ESDM and evolve ESiWACE S3NetCDF prototype towards Pangeo and zarr?



STFC: 18PM (plus note overlap with ESDM work!)
UKMO: 3 PM (cylc support)

Bring your own software stack: Only delivered by containers?

User needs

1. **Software dependencies** that are numerous, complex, unusual, differently configured, or simply newer or older than what is already provided.
2. **Build-time requirements** unavailable within the centre, such as relatively unfettered internet access.
3. **Validated software stacks** and configuration to meet the standards of a particular field of inquiry.
4. **Portability of environments** between resources, including workstations and other test/development systems not managed by the centre.
5. **Consistent environments** that can be easily, reliably, and verifiably reproduced in the future.
6. **Usability and comprehensibility**.

List from Charliecloud, Priedhorsky and Randles, 2017, https://doi.org/10.1145/3126908.3126925

Workflow Enhancements, Testing

Task 4.5 (19PM)

Goal: Add **explicit** support in SLURM and cylc for scheduling and staging data intensive workloads (with and without ESDM/SemST).

1. Co-Design (DKRZ, 3PM)
2. Enrich Cylc to deal with data dependencies, life-cycle, and support for ESDM/SemST (Backend: MetO, 6PM, Frontend: UREAD 3PM).
3. Enrich SLURM to understand data locality (UREAD, 3PM, DDN, 4PM)

Task 4.6 (3PM + Everyone)

Goal: Testing and continuous integration

1. Continuous Integration Environment (3PM, ICHEC)
2. Component Level QA environment (Everyone)
3. (Eventually) ”Regular” end-to-end testing (Everyone)

Task 4.7: Industry Proof Of Concept

- Uptake of WP4 products will depend on a range of factors, which will include: fitness for purpose, quality, ease of deployment, maintenance and performance.
- For the ESDM in particular, maintenance and performance will be important, which means for sustainability, some sort of vendor engagement will be crucial *in some environments*. Which in turn means they need something to sell which depends on the ESDM.
- **Seagate and DDN** will be developing appliances, that is customised solutions which integrate customised ESDM layout and performance algoirithms and implementations with their storage hardware and software.

Deliverables, Milestones, Interactions

Deliverables

4.1 Month 24: Advanced Software Stack, including ESDM and ensemble code.

4.2 Month 42: Report on appliances (availability, configuration, performance).

4.3 Month 48: Documentation and Roadmap beyond ESiwACE2.

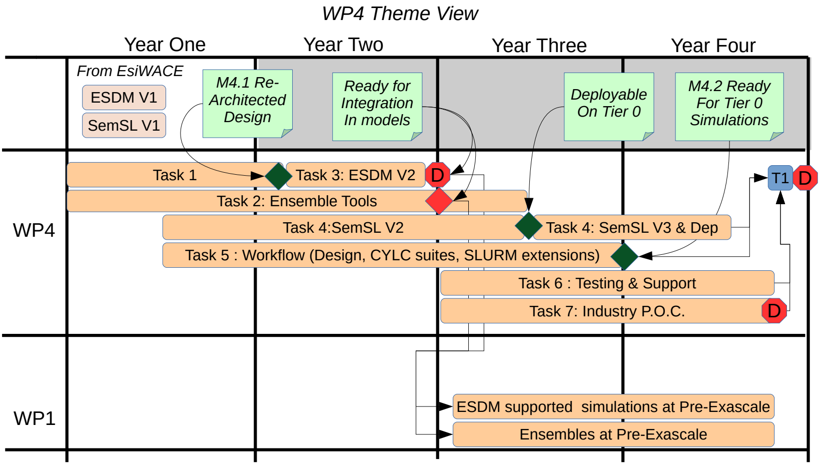
Milestones

1. Month 15: M4.1 Architecture Update
2. Month 36: M5.2 Workflow Extensions

Key Interactions (not all)

- With the XIOS team in this project and IS-ENES3.
- With the WP1 team
  - ESDM support
  - High Resolution Ensembles with the UM, and hopefully, EC-Earth
- With the WP2 Containers Team
- With WP5, ESDM into post-processing, analytics and visualisation
- WP6 Training.

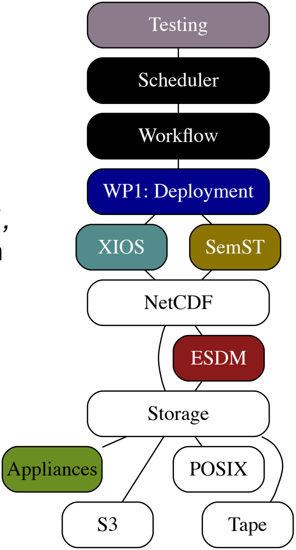
Timeline


WP4 Theme View
Year One
Year Two
Year Three
Year Four
From EsiWACE
ESDM V1
SemSL V1
M4.1 Re-Architected Design
Ready for Integration In models
Deployable On Tier 0
M4.2 Ready For Tier 0 Simulations
WP4
Task 1
Task 3: ESDM V2
D
Task 2: Ensemble Tools
Task 4:SemSL V2
Task 4: SemSL V3 & Dep
T1
D
Task 5 : Workflow (Design, CYLC suites, SLURM extensions)
Task 6 : Testing & Support
Task 7: Industry P.O.C.
D
WP1
ESDM supported simulations at Pre-Exascale
Ensembles at Pre-Exascale

(and with WP5 throughout ...)

Summary

- WP4 is about scale - performance and volume.
- An ambitious programme, but we have learnt some lessons about that from ESiWACE(1).
- Aiming for deployment, in WP5 and *in WP1, from month 24*, will need engagement from WP1 task leaders.
- Several activities at various places in the stack.
- Aiming for portability (some elements will be containerised) and sustainability with *key role for industrial partners* - both in terms of the appliances and helping us deliver on portability.



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