# Exploiting Different Storage Types with the Earth-System Data Middleware

Julian Kunkel (University of Reading), Luciana Pedro (University of Reading), Bryan Lawrence (University of Reading), Sandro Fiore (CMCC), Huang Hua (Seagate)

Department of Computer Science, University of Reading

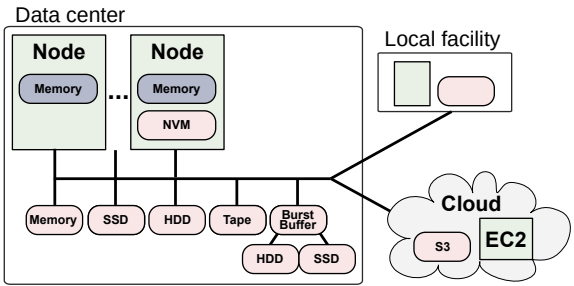
18 November 2019

# Outline



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

# The Coexistence of Storage

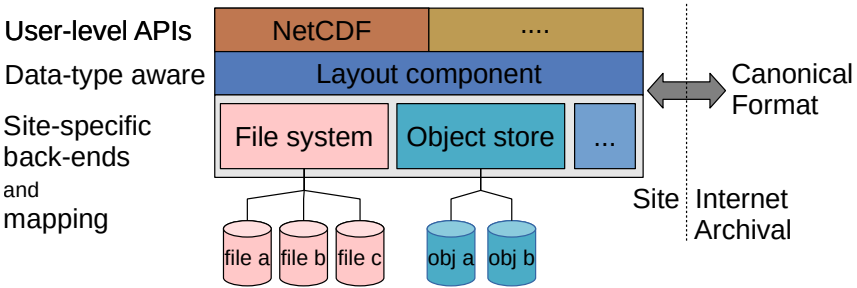



- Goal: We shall be able to exploit all storage technologies concurrently
  - Without explicit migration, put data where it fits
  - Administrators just add new technology (e.g., SSD pool) and users benefit from it
- May utilize local storage, SSDs, NVMe
  - Even without communication used in workflows

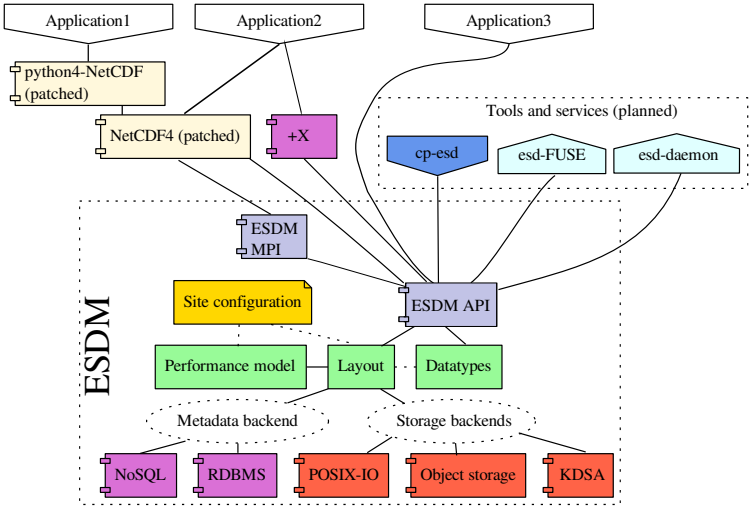
# Architecture of the Earth-System Data Middleware

## Key concepts

- Middleware utilizes layout component to make placement decisions
- Applications work through existing API
- Data is then written/read efficiently; potential for optimization inside library



# Architecture: Detailed View of the Software Landscape



# Backends

## Storage backends

- POSIX: Backwards compatible for any shared storage
- CLOVIS: Seagate-specific interface, will be open sourced soon
- WOS: DDN-specific interface for object storage
- KDSA: Specific interface for the Kove cluster-wide memory
- PMEM: Non-volatile storage interface (`http://pmem.io`)

## Metadata backends

- POSIX: Backwards compatible for any shared storage
- Investigated performance of ElasticSearch, MongoDB as potential NoSQL solutions

# Evaluation

## System

- Test system: DKRZ Mistral supercomputer
- Nodes: 100, 200, 500

## Benchmark

- Uses ESDM interface directly; metadata on Lustre
- Write/read a timeseries of a 2D variable; 3x repeated
- Grid size: 200k $\times$ 200k $\times$ 8 Bytes $\times$ 10 iterations
- Data volume: size = 2980 GiB; compared to IOR performance (partially shown)

## ESDM configurations

- Splitting data into fragments of 100 MiB or 500 MiB
- Use one Lustre, two Lustre fs, TMPFS or Local SSD

# Performance Growth of ESDM on Lustre (PPN = 1)

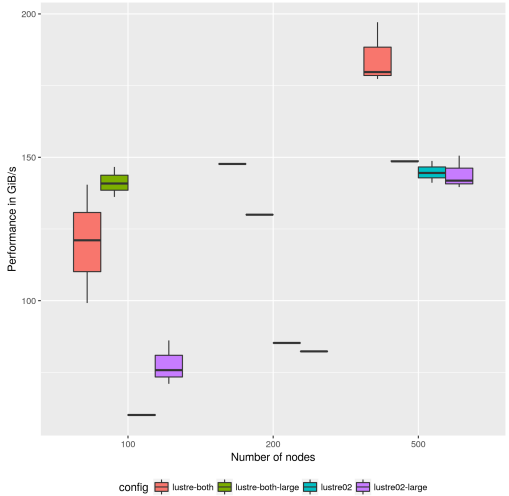



Figure: Write

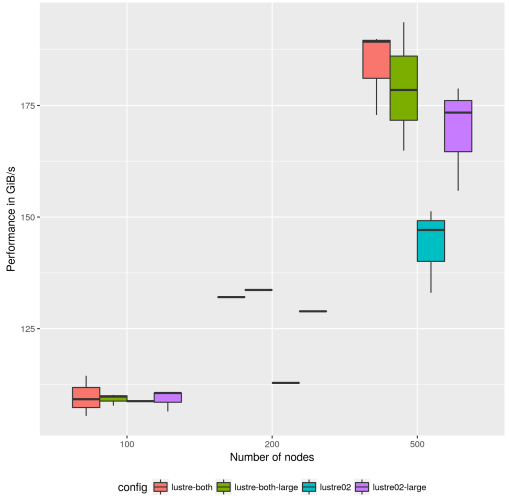



Figure: Read

# Performance on TMPFS vs. IOR (nodes = 500, varied PPN)

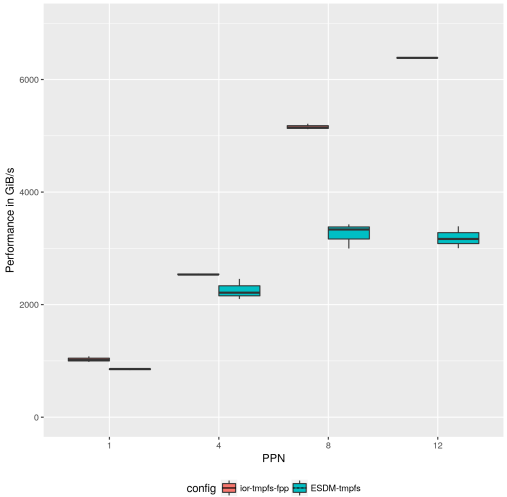



Figure: Write

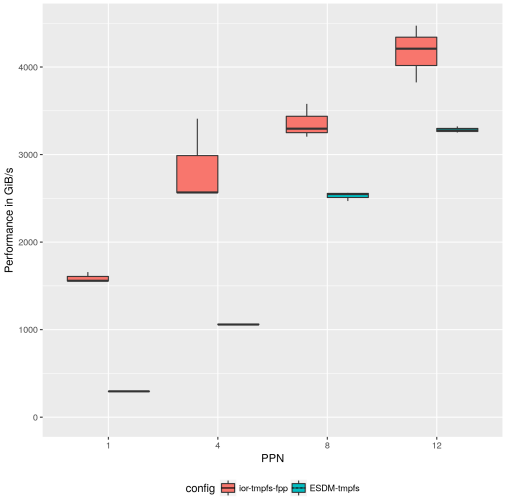



Figure: Read

# Performance on NVDIMMs

- ESDM on the NextGenIO Prototype with a first naive approach (with PMEM)
- Test run on four dual-socket nodes with 80 GByte of data
- Theoretic HW performance per node (12 NVDIMMs) W: 96 GB/s, R: 36 GB/s
- Max test: explore best case performance (single file)

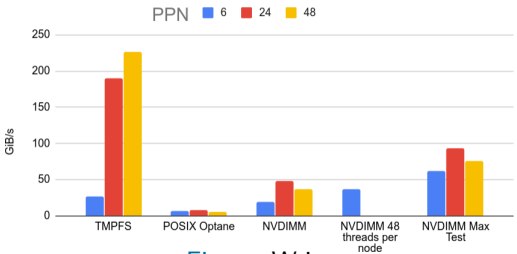



Figure: Write

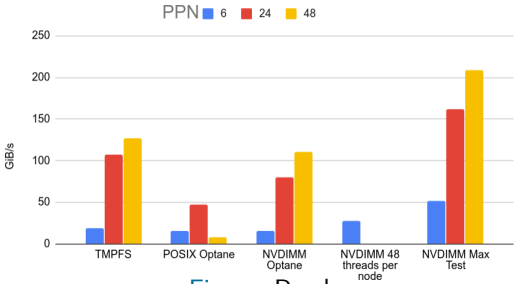



Figure: Read

# Status

- NetCDF: Done, trivial issues to fix, use tests for checking compatibility
  - netcdf4-python: Available, derived tests with supported features
- First tools implemented (esdm-mkfs, esdm-rm)
- Deployed daily regression testing using Jenkins (Webpage will go public soon)
- FUSE prototype to dynamically build a hierarchical namespace on semantics
  - E.g., `<model>/<date>/<variable>`

## ESiWACE2 Plans

- Hardening and optimisation of ESDM
- Integrate an improved performance model
- Industry proof of concepts for EDSM, i.e., shipping of HW with software
- Workflow support and active storage

# ESiWACE: http://esiwace.eu

## The Centre of Excellence in Simulation of Weather and Climate in Europe

- Prepare the European weather and climate community
  - Make use of future exascale systems
- Goals in respect to HPC environments
  - Improve efficiency and productivity
  - Supporting the end-to-end workflow of global Earth system modelling
  - Establish demonstrator simulations that run at the highest affordable resolution
- Funding via the European Union's Horizon 2020 program (ESiWACE2 2019-2022)

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

# Earth-System Data Middleware

## A transitional approach towards a vision for I/O addressing

- Scalable data management practice
- The inhomogeneous storage stack
- Suboptimal performance and performance portability
- Data conversion/merging

## Design goals of the Earth-System Data Middleware

1. Relaxed access semantics, tailored to scientific data generation
2. Site-specific (optimized) data layout schemes
3. Ease of use and deploy a particular configuration
4. Enable a configurable namespace based on scientific metadata

# Performance Discussion

- Benefit when accessing multiple global file systems
- Write performance benefits from using both file systems
  - Most benefit when using 200 nodes (2x)
  - 500 nodes: 180 GiB/s vs. 140 GiB/s (single fs)
- Read performance shows some benefit for larger configurations
- ESDM achieves similar performance regardless of PPN (not shown)
- What is the performance when we use node-local storage?

# Discussion

- Node-local storage is much faster than global storage
  - TMP achieves 750-1,000 GB/s for write (500 SSDs, some caching)
  - TMP reads are actually cached (6 GB data per node)
  - TMPFS achieves up to 3,000 GB/s
- TMP write is invariant to PPN
  - ESDM configured to use at least four threads per node
- TMPFS write depends on PPN
  - ESDM configured to not use threads, could use them to improve performance!
- IOR is faster; potential to improve ESDM path further
  - Localization of fragments using r-tree

The ESiWACE1/2 projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No **675191** and No **823988**