Middleware for Earth System Data

Julian Kunkel*, Jakob Luettgau*, Bryan N. Lawrence†, Jens Jensen†, Giuseppe Congiu‡, John Readey§
*Deutsches Klimarechenzentrum GmbH, †STFC Rutherford Appleton Laboratory, ‡Seagate Technology LLC, §The HDF Group

***Abstract*—Making the best use of HPC in Earth simulation requires storing and manipulating vast quantities of data. Existing storage environments face usability and performance challenges for both domain scientists and the data centers supporting the scientists. These challenges arise from data discovery/access patterns, and the need to support complex legacy interfaces.**

In the ESiWACE project, we develop a novel I/O middleware targeting, but not limited to, earth system data. Its architecture builds on well established end-user interfaces but utilizes scientific metadata to harness a data structure centric perspective.

I. Challenges

Climate and weather applications are I/O intensive. Historically researchers used to optimize codes for specific supercomputers but with increasingly complex systems this approach is not feasible. As a result, and in an effort to allow for easier exchange and inter-comparison of model and observations, data libraries for standardized data description and optimized I/O such as NetCDF, HDF5 and GRIB were developed. However, many I/O optimizations used in these libraries do not adequately reflect current data intensive system architectures. Additionally, data management needs to scale with future data volumes in a way which provides acceptable data access latencies and data durability, and is cost-effective. Systems now need to support multi-disciplinary research through shared, interoperable interfaces, based on open standards, and deliver environmentally-responsible and flexible hybrid data and compute infrastructures.

II. Related Work

Mehta et al [1] have exploited the Virtual Object Layer (VOL) abstraction provided by the HDF5 library to build a parallel log-structured file system plugin that manages separately scientific data and metadata. Dong et al [2] have used the VOL to build the Scientific Data Service (SDS) framework to adapt stored data to the specific workloads, while the DOE Fast Forwards I/O project [3] integrated the VOL in the new storage system architecture based on Lustre object store extensions.

III. Approach

To meet the aforementioned challenges, we have designed the *Earth System Data (ESD)* middleware, which: 1) understands application data structures and scientific metadata, which lets us expose the same data via different APIs; 2) maps data structures to storage backends with different performance characteristics based on site specific configuration informed by a performance model; 3) yields best write performance via optimized data layout schemes that utilize elements from log-structured file systems; 4) provides relaxed access semantics, tailored to scientific data generation for independent writes, and; 5) includes a FUSE module which will provide backwards compatibility through existing file formats with a configurable namespace based on scientific metadata.

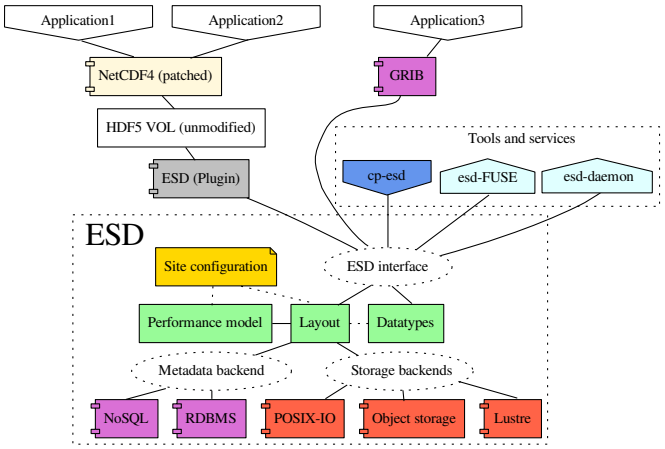

Fig. 1. Architecture overview

Together these allow storing small and frequently accessed data on node-local storage, while serializing multi-dimensional data onto multiple storage backends – providing fault-tolerance and performance benefits for various access patterns at the same time. Compact-on-read instead of garbage collection will additionally optimize and replicate the data layout during reads via a background service. Additional tools allow data import/export for exchange between sites and tape archives.

IV. Summary

The ESD aids the interests of stakeholders: developers have less burden to provide system specific optimizations and can access their data in various ways. Data centers can utilize storage of different characteristics. We expect a working prototype with the core functionality within the next year. Following work will implement and fine-tune the cost model and layout component and provide additional backends.

Acknowledgment

The ESiWACE project received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 675191.

References

[1] K. Mehta, J. Bent, A. Torres, G. Grider, and E. Gabriel, "A Plugin for HDF5 Using PLFS for Improved I/O Performance and Semantic Analysis," in *Proceedings of the 2012 SC Companion: High Performance Computing, Networking Storage and Analysis*, ser. SCC '12', 2012.

[2] B. Dong, S. Byna, and K. Wu, "Expediting scientific data analysis with reorganization of data," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, Sept 2013, pp. 1–8.

[3] DOE Fast Forward I/O Project, Final Report. https://wiki.hpdd.intel.com/download/attachments/12127153/M8.5%20FF-Storage%20Final%20Report%20v3.pdf.