Towards Structure-Aware Earth System Data Management

Jakob Luettgau
German Climate Compute Center
luettgau@dkrz.de

Julian Kunkel
German Climate Compute Center
kunkel@dkrz.de

Bryan N. Lawrence
University of Reading
bryan.lawrence@ncas.ac.uk

Sandro Fiore
CMCC Foundation
sandro.fiore@cmcc.it

Huang Hua
Seagate Technology LLC
hua.huang@seagate.com

ABSTRACT

Current storage environments confront domain scientist and data center operators with usability and performance challenges. To achieve performance portability data description libraries such as HDF5 and NetCDF are widely adopted. At the moment, these libraries struggle to adequately account for access patterns when reading and writing data to multi-tier distributed storage systems. As part of the ESiWACE[1] project, we develop a novel I/O middleware targeting, but not limited to, earth system data. The architecture builds on top of well established end-user interfaces but utilizes scientific metadata to harness a data structure centric perspective.

1 INTRODUCTION

As scientists are adapting their codes to take advantage of the next-generation exascale systems, the I/O bottleneck becomes a major challenge[1–3] because storage systems struggle to absorb data at the same pace as it is generated. Especially, simulation codes such as climate and numerical weather prediction periodically experience bursty I/O, as they are writing so called checkpoints to achieve fault tolerance and for data analysis. Technological and budgetary constraints have lead to complex storage hierarchies.

2 APPROACH

The overall architecture of the *Earth System Data (ESD)* middleware is depicted in Figure 1. It is designed to address multiple I/O challenges, in particular this includes:

(1) awareness of application data structures and scientific metadata, which lets us expose the same data via different APIs;-
(2) map data structures to storage backends based on performance characteristics of storage configuration of site
(3) optimize for write performance by combining data fragmentation and elements from log-structured file systems;
(4) provides relaxed access semantics, tailored to scientific data generation for independent writes, and;
(5) backwards compatability to existing file formats is provided by a FUSE module which exposed a virtual configurable namespace based on scientific metadata

As result small and frequently accessed data is kept 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. The middleware also uses structural information to optimize workflow performance.

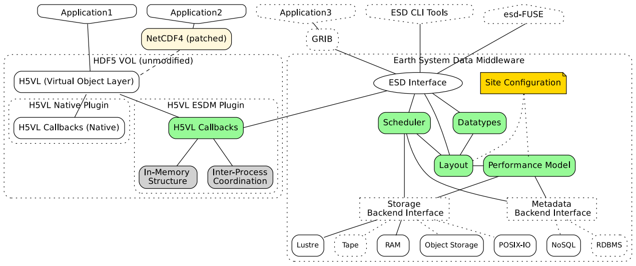

Figure 1: Overview of the architecture, which allows the middleware optimize for site specific data services without requiring changes to applications.

3 ACTIVITIES AND STATUS

A first prototype was developed demonstrating the viability of adaptively choosing tiers based on policies to achieve performance gains. By using information exposed by HDF5, MPI and SLURM it was possible to account for checkpoint size and domain decomposition `{nodes+ppn}`. Without changing application code interfacing with NetCDF/HDF5, it was possible to choose different I/O paths (e.g. SHM, SSD, Lustre) for different requests. In addition, the design of the proposed architecture of the middleware thoroughly documented covering access semantics, (meta)data backends and processing throughout the I/O stack. The full report is available on the ESiWACE website [1]. Currently, the preliminary interfaces for the data and metadata backends are being developed.

4 SUMMARY

The ESDM addresses the challenges of multiple 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 coming year. Following work will implement and fine-tune the cost model and layout component and provide additional backends. Besides storage backends, an integration of scientific workflows with workload manager requires investigation as it offers new opportunities to automatically reduce data movements by exploiting spacial and temporal data locality.

REFERENCES

[1] [n. d.]. ESiWACE | Centre of Excellence in Simulation of Weather and Climate in Europe. ([n. d.]). https://www.esiwace.eu/
[2] [n. d.]. NEXTGenIO | Next Generation I/O for the Exascale. ([n. d.]). http://www.nextgenio.eu/
[3] Intel, The HDF Group, EMC, and Cray. 2014. Fast Forward Storage and I/O - Final Report. (June 2014).