Exploiting Different Storage Types with the Earth-System Data Middleware

Julian Kunkel
University of Reading
j.m.kunkel@reading.ac.uk

Luciana Pedro
University of Reading
l.r.p@reading.ac.uk

Bryan Lawrence
University of Reading
b.n.lawrence@reading.ac.uk

Sandro Fiore
CMCC Foundation
sandro.fiore@cmcc.it

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

***Abstract*—This paper describes early experiences with middleware that provides a high level of abstraction for earth system applications in the presence of storage heterogeneity: the Earth System Data Middleware (ESDM). A preliminary performance assessment of the middleware on different storage technology illustrates the importance of adaptive data placement.**

I. Introduction

A decade ago, the storage landscape of data centres consisted of disk and tape technology providing global shared file systems within hierarchical storage systems. Now, for cost-efficient performance and optimisation across use-cases, the compute and storage systems of HPC environments are increasingly relying on a more heterogeneous architecture. Systems are already adapting to new technology, particularly flash, non-volatile memory, and burst-buffers. This heterogeneity is growing with the increasing prominence of cloud computing technologies and public cloud capability. It is then expected that any storage environment may contain a range of technologies and interfaces, and the environment itself may be distributed.

Organising placement of data on such storage manually by the application or via policy alone is suboptimal. A storage policy may not be aware of subsequent usage, and hence being unable to direct data to locations where it should be optimally available as part of workflows. In an attempt to deal efficiently with the additional complexity that is arising from node-local and rack-local storage, we expect network-attached (non-volatile) memory and solid-state disks to provide the fastest access at the highest cost per unit volume, while tape archives offer capacity for a range of latencies and bandwidths at lower cost/volume ratios.

The Earth System Data Middleware (ESDM) aims to exploit heterogeneous storage landscapes by utilising the available storage concurrently and allowing meaningful subsets of data to be placed on differing storage systems as required. The ESDM targets concurrency from a single parallel application providing dynamic data granularity. The main ideas behind this software-centric co-design effort to address the I/O challenges are: 1. High-level semantics, i.e., awareness of application data structures and scientific metadata and 2. Flexible mapping of data to multiple storage backends. The ESDM is under development and here we introduce our prototype and present some preliminary results.

II. Results

The performance of the ESDM has been investigated on Mistral (DKRZ) for 100, 200, and 500 nodes (see Figure 1) and it has been measured in storing data on different storage systems: a parallel file system, the local SSD, and in-memory storage. A performance of 180 GB/s (write and read) was achieved when using both file systems. In contrast, using only the Lustre02 file system yields less performance. For comparison with the optimal case, the performance of the IOR benchmark has been measured using similar configurations. It is a well-known fact that in-memory and local storage performs better for large systems, and our prototype replicates these expectations. IOR yielded similar results but less performance with one PPN because Lustre benefits from concurrent I/O on a node. The ESDM uses multiple threads automatically and so can improve performance for these cases.

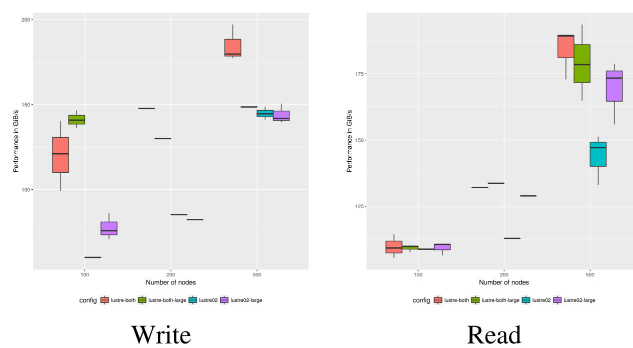

Fig. 1: Performance of ESDM on the Mistral Supercomputer

III. Summary and Status

The Earth-System Data Middleware (ESDM) has been designed to allow interactions with different kinds of backends through data storing and retrieval. It enables an increased usage of the node-local storage and improves model performance within I/O processes. We believe such a high-level of abstraction is the natural next step towards an era of intelligent storage. Please check our git repository[1].

Acknowledgment

This project is funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No. 823988. https://www.esiwace.eu/

[1]https://github.com/ESiWACE