Adaptive Tier Selection for NetCDF/HDF5

Jakob Luettgau*, Eugen Betke*, Olga Perevalova‡, Julian Kunkel*, Michael Kuhn‡
*German Climate Copmuter Center (DKRZ), ‡University of Hamburg (UHH)

Abstract

Scientific applications on supercomputers tend to be I/O intensive. To achieve portability and performance data description libraries such as HDF5 and NetCDF are commonly used. Unfortunetly, the libraries often default to suboptimal access patterns for reading/writing data to multi-tier distributed storage. This work explores the feasibility of adpatively selecting tiers depending ona applications I/O behavior.

Overview

The contributions presented in this work are:

- A proof of concept prototype implementation demonstrating the benefit of adaptive tier selection on a real system
- An architecture for I/O middleware beyond adaptive tier selection for more intelligent data placement from user space

Opportunities using HDF5 Virtual Object Layer

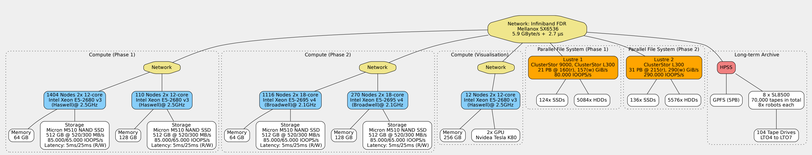
Hierarchical Data Format (HDF5): HDF5 is an open source, hierarchical and self-describing format that combines data and metadata. Advantages of this format make it widely used by scientific applications.

Virtual Object Layer (VOL): The VOL is an abstraction layer in the HDF5 library with the purpose to expose the HDF5 API to applications while allowing to use different storage mechanism. The VOL intercepts all API calls and forwards those calls to plugin object drivers. The VOL exports an interface that allows third party plugin development.

Plugin for Seperate Metadata Handling: A VOL plugin was developed to handle data and metadata seperately. For adaptive tier selection this is necessary to keep track of alternating data sources but it also offers additional opportunties. Generated simulation data is routinely published, but at the moment not automatically catalogued. With the VOL plugin it would be possible to extracted a dataset description to make them available for search in a catalogue as the dataset is written.

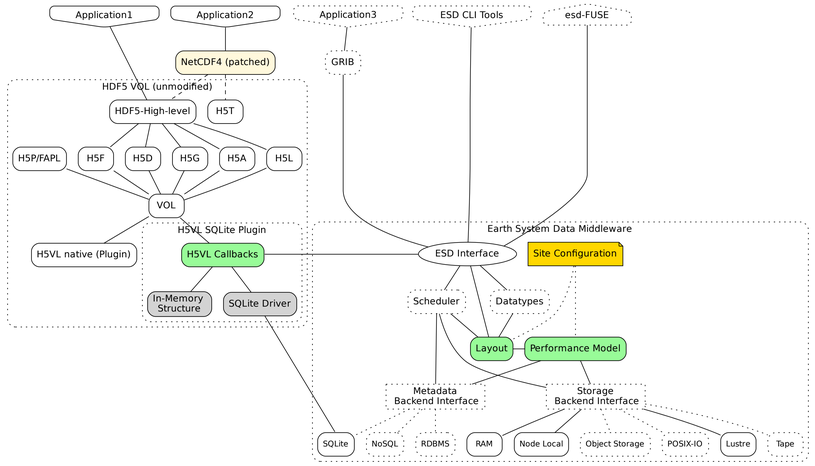
Mistral Supercomputer

Performance evaluation was carried out on the Mistral supercomputer. The computer, listed #38 in the Top500 (June'17), is located at the German Climate Compute Center (DKRZ) in Hamburg and exclusively serves the climate research community. The current site configuration is as follows:



Architecture for Adaptive Tier Selection

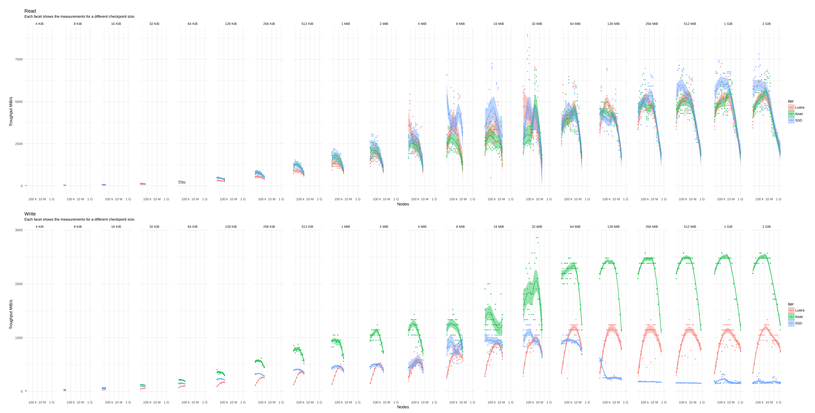
Adaptive tier selection is realized as depicted in the architecture illustration below. The prove of concept decision component accounts for two runtime information made available by MPI and HDF5: 1) domain decomposition and 2) the domain description of a dataset. The tier selection policies are based on benchmark measurements obtained at an earlier time.



Performance Evaluation

The following plots show throughput of each tier for `READ` and `WRITE` in comparison to the performance when a VOL plugin that adaptively selects the most appropriate tier. The following tiers are considered:

- Shared Memory: Small random I/O and in expectance of burst buffers.
- Local SSDs: For medium random I/O not shared with other nodes.
- Parallel File System: When performing large sequential file I/O.



Each configuration consists of a checkpoint size `{xdim,ydim,zdim}` which is read and written `t` times, a domain decomposition `{nodes+ppn}` and the I/O path `{SHM,SSD,Lustre,adaptive}`.

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

A main goal of the Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE) is to improve efficiency and productivity of numerical weather and climate simulation on high-performance computing platforms by supporting the end-to-end workflow of global Earth system modeling in HPC environment. Part of the project is the development of a middleware for earth system data featuring:

- Access to shared data with different APIs
  - NetCDF4, HDF5 or GRIB
- Data-center optimized layouts
  - Advanced data placement optimizing for cost and performance
  - Support for different backends: object storage, file systems

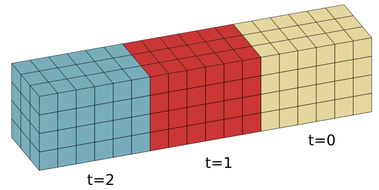
Summary and Future Work

Adaptive tier-selection promises to be a viable approach for performance optimization of I/O performance. As storage systems become more heterogeneous in the wake of burst buffers and non-volatile memory I/O middleware can help to avoid exposing unnecessary complexity to users. In future work the decision component should automatically extract tier-selection rules from benchmark measurements and observed access patterns. The integration of various storage tiers is continued as part of the ESiWACE project. In particular the following backends deserve further exploration:

- Object storage mappings for short term storage of working sets
- Tape and other nearline storage for affordable long-term archival

NetCDF Benchmark

NetCDF Performance Benchmark Tool (NetCDF-Bench) was developed to measure NetCDF performance on devices ranging from notebooks to large HPC systems. It mimics the typical I/O behavior of scientific climate applications and captures the performance on each node/process. Details: `https://github.com/joobog/netcdf-bench`



Acknowledgments

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