Coupled Storage System for Efficient Management of Self-Describing Data Formats

Michael Kuhn (michael.kuhn@ovgu.de), Kira Duwe (kira.duwe@ovgu.de)
Parallel Computing and I/O • Otto von Guericke University Magdeburg
https://parcio.ovgu.de • https://cosemos.de • https://github.com/julea-io

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Project Description

- CoSEMoS's goal is to rethink the **architecture of storage systems**
  - DFG project to improve performance and data management
  - Built upon JULEA: Modern C11 code, available as open source
  - Currently in its third year with a funding period of 2019–2022
- JULEA provides a **flexible storage framework**
  - Contains necessary building blocks for storage systems
  - Facilitates rapid prototyping and evaluation
- Runs in **user space** and has **few dependencies**
  - Kernel code increases complexity and fragility significantly
  - Possible to use on clusters without root access

Problem Statement

- **Self-describing data formats (SDDFs)** widely used to exchange data
  - Structural information is encoded in the files themselves

1. **Weak treatment of different types of metadata**
   - Strict separation of metadata leads to inefficient file access
   - **File system metadata** is stored on the metadata servers
   - **File metadata** (for example, attributes or additional annotations) is stored within SDDF files on the data servers
2. **Static I/O semantics**
   - Strict consistency and coherence semantics due to POSIX
   - Static approaches are unable to satisfy all requirements
3. **Inefficient data placement**
   - Hierarchical structuring of different hardware is used
   - Data movement across storage tiers is an expensive operation
   - Hardware is available, new approaches need to be developed

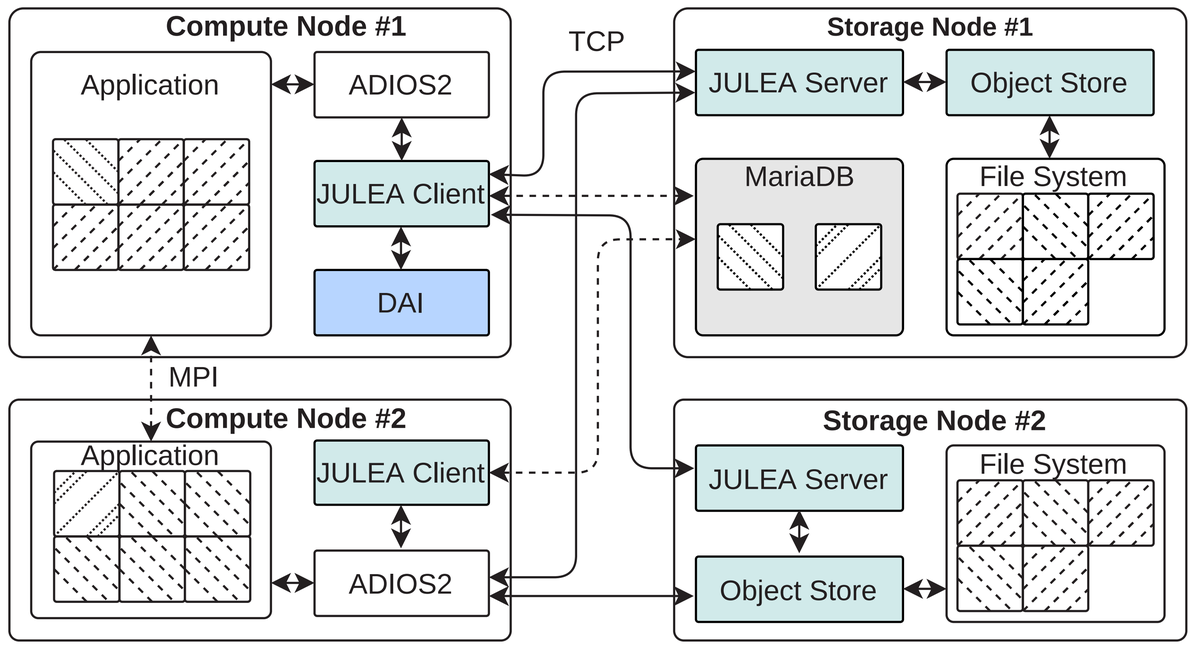
Objectives

1. **Global metadata management**
   - Closely **couple storage system and self-describing data formats**
     - All metadata handled by metadata servers
     - Optimize metadata accesses using database technologies
   - Novel data management approaches via a **data analysis interface**
     - Query file metadata across multiple files in a unified way
2. **Adaptable I/O semantics**
   - Possible to dynamically adapt semantics
     - Provide appropriate interfaces for applications and libraries
     - Specify **requirements regarding atomicity, consistency etc.**
3. **Intelligent storage selection**
   - Use structural information for **informed data placement decisions**
     - Improve performance by optimizing data placement
   - Different parts of self-describing files can be put on different tiers

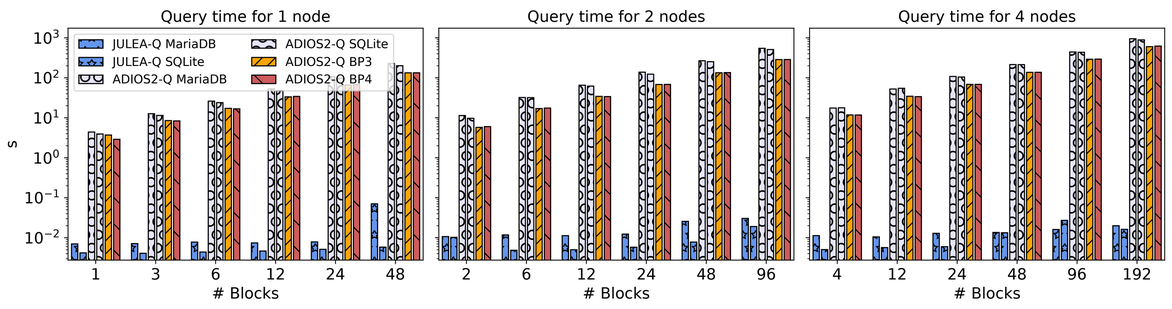
Publications

1. Dissecting Self-Describing Data Formats to Enable Advanced Querying of File Metadata – K. Duwe, M. Kuhn (SYSTOR 2021) https://dl.acm.org/doi/10.1145/3456727.3463778
2. Using Ceph's BlueStore as Object Storage in HPC Storage Framework – K. Duwe, M. Kuhn (CHEOPS 2021 @ EuroSys 2021) https://dl.acm.org/doi/10.1145/3439839.3458734
3. Coupling Storage Systems and Self-Describing Data Formats for Global Metadata Management – M. Kuhn, K. Duwe (CSCI 2020) https://ieeexplore.ieee.org/document/9457959
4. State of the Art and Future Trends in Data Reduction for High-Performance Computing – K. Duwe, J. Lüttgau, G. Mania, J. Squar, A. Fuchs, M. Kuhn, E. Betke, T. Ludwig (Supercomputing Frontiers and Innovations 2020) – https://doi.org/10.14529/jsfi200101

Architecture and Evaluation



- JULEA (**turquoise**), database (**gray**) and data analysis interface (**blue**)
- Small file metadata pieces (**fine-grained pattern**) are stored in a database, data chunks (**coarse-grained pattern**) are stored as objects



- JULEA supports **arbitrary data formats** (currently HDF5 and ADIOS2)
  - DAI offers **direct access** to database and key-value store
  - Previous study with HDF5 shows significant speedups (see website)
- Use case: Query mean values from ADIOS2
  - Mean values are **precomputed** by the ADIOS2-JULEA engine
  - DAI does not need to access the object store containing the data
  - ADIOS2 API has **no way to access the additional metadata**
- Querying 192 blocks: 0.01 s vs. 600+ s (**60,000x speedup**)

Work Packages

- WP1: **Application Interface**
  - T1.1: SDDF Interface ✔
  - T1.2: Application Requirements and Semantics ⌛
  - T1.3: Data Analysis Interface ⚙
- WP2: **Storage Tier Selector and Global Metadata Manager**
  - T2.1: Database Backend ✔
  - T2.2: Database Client ✔
  - T2.3: Metadata Backend Selection ⚙
  - T2.4: Data Storage Tiering ⚙
- WP3: **Evaluation and Dissemination**
  - T3.1: Compatibility Tests ⚙
  - T3.2: Case Study ⚙
  - T3.3: Workshop Organization ✔

Partners

- German Climate Computing Center (Prof. Dr. Thomas Ludwig)
- Intel (Johann Lombardi)
- Max Planck Institute for Meteorology (Uwe Schulzweida)

Acknowledgements and Links

CoSEMoS is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 417705296. All results are being published on the project website and integrated into JULEA.