EMPRESS—Extensible Metadata Provider for Extreme-scale Scientific Simulations

Margaret Lawson, Jay Lofstead, Scott Levy, Patrick Widener, Craig Ulmer, Shyamali Mukherjee, Gary Templet, Todd Kordenbrock
Problems Faced

- Simulations with 100s TB per output, run every few minutes
  - Ex. XGC1, Square Kilometer Array Radio Telescope (SKA)

- Storage devices too slow to sift through all output to find “interesting data”

- Scientists have specific data they want to retrieve
  - Ex. “blob” in fusion reactor or a phenomenon in astronomy
Motivating Question

*How can we facilitate scientific discovery from simulations in the exascale age?*
EMPRESS’ Solution

- Allow users to label data and retrieve data based on labels

Features:
- Robust, standard per-process metadata
- User-created metadata that is fully customizable at runtime
- Programmatic query API to retrieve data contents based on metadata
Previous Solutions

- HDF5 and NetCDF – rudimentary attribute capabilities, basic metadata
- ADIOS – per-process metadata

None of these address efficient attribute searching

- FastBit – offers data querying based on values, but very limited support for spatial queries and attributes
Why not use a Key-Value Store?

- Custom keys can go a long way, but not far enough
- Two Problems:
  - Inexact matches
  - Custom Metadata
- Relational databases with indices are radically faster at searching like this
SIRIUS Architecture

Applications

I/O API

Cross Layer Services
- Refactoring
- Reduction
- Data Placement & Movement
- Other Plugins

Storage and I/O System Services
- QoS
- Resource Management
- Migration
- Purging
- EMPRESS

Description of Data

SIRIUS Architecture

Storage Resources (Ceph managed)
- NVRAM
- PFS
- Campaign Storage
- Long term storage
SIRIUS Workflow – Write Process

Simulation

Lightweight Analysis

Generate Tags

Metadata + tags

EMPRESS

Data

Ceph
SIRIUS Workflow – Read Process

1. Query

2. Programmatic Query API

3. Matching Object Names

4. Object Names

5. Data

6. Data

User

ADIOS

EMPRESS

Ceph
High Level Design

Simulation Node

- Simulation
- ADIOS
- EMPRESS API
- Ceph API

Simulation

Programmatic Query API

EMPRESS Servers

Sandia National Laboratories

10
Faodail

Data Interface Modules (DIMs)

Kelpie
Distributed, In-memory
Key/Blob Service

Communication Services

I/O Modules (IOMs)

Lunasa
Network Memory Management

RDMA Portability Layer

NNTI 3.0
libfabric

Data Warehouse
Storage - Tracked Metadata

- Dataset information
  - Application, run, and timestep information
- Variable information
  - Catalogs types of data stored for an output operation
- Variable chunk information
  - Subdivision of simulation space associated with a particular variable
- Custom metadata class
  - Metadata category the user adds for a particular dataset
  - Ex. Max
- Custom metadata instance
  - Ex. Flag for chunk or a bounding box spanning chunks
Testing Goals

- Scalable?
  - Number of client processes: 1024-2048

- Effect of client to server ratio
  - Ratios tested: 32:1 – 128:1

- Overhead of including a large number of custom metadata items
  - Number of custom metadata classes: 0 or 10
  - On average 2.641 custom metadata instances per chunk
Testing Goals (Continued)

- **Proof of concept, can EMPRESS efficiently support:**
  - Common writing operations
    - 2 datasets written, each with 10 globally distributed 3-D arrays
  - Common reading operations
    - 6 different read patterns that scientists frequently use (Lofstead, et al. “Six Degrees of Scientific Data”)
  - A broad range of custom metadata
    - 10 custom metadata classes including max, flag, bounding box (two 3-D points)

- **Scientific validity**
  - A minimum of 5 runs per configuration on 3 computing clusters:
    - Serrano (total nodes: 1122)
    - Skybridge (total nodes: 1848)
    - Chama (total nodes: 1232)
EMPRESS efficiently supports a wide variety of operations including custom metadata operations.
Most time is spent waiting for the server to respond
  - Room for improvement in the Faodail infrastructure
Testing – Writing and Reading Time

- Good scalability for fixed client-server ratio
- No significant overhead for adding custom metadata
- Client-server ratio greatly affects performance
Future Work

- Increasing EMPRESS’ flexibility, efficiency, and scalability
  - Support more queries
  - Different metadata distribution?
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

This work was supported under the U.S. Department of Energy National Nuclear Security Agency ATDM project funding. This work was also supported by the U.S. Department of Energy Office of Science, under the SSIO grant series, SIRIUS project and the Data Management grant series, Decaf project, program manager Lucy Nowell.
Algorithm 1 Writing algorithm

1: procedure WRITE_TIMESTAMP
2:     for all variables assigned do
3:         md_create_var (...)
4:     end for
5:     for all custom metadata classes assigned do
6:         md_create_type (...)
7:     end for
8:     for all variables do
9:         md_insert_chunk (...)  ▶ Add a var chunk; get the ID
10:        for all custom metadata desired do
11:            md_insert_attribute (...) ▶ Add custom md instance
12:        end for
13:     end for
14: end procedure
Algorithm 2 Reading algorithm

1: procedure ReadData \Comment{Each Process Does this}
2: \hspace{0.5em} md_catalog_vars (...) \Comment{Get list of vars from any server}
3: \hspace{0.5em} for all metadata servers needed do
4: \hspace{1em} md_get_chunk(...) \Comment{get all chunks in area of interest}
5: \hspace{1em} \hspace{0.5em} for all chunks returned do
6: \hspace{1.5em} \hspace{0.5em} md_get_attribute (...) \Comment{get the custom md instances}
7: \hspace{1em} \hspace{1em} end for
8: \hspace{1em} end for
9: end procedure