Exceptional service in the national interest

Sandia National Laboratories

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

SAND2017-12103 C

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.

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

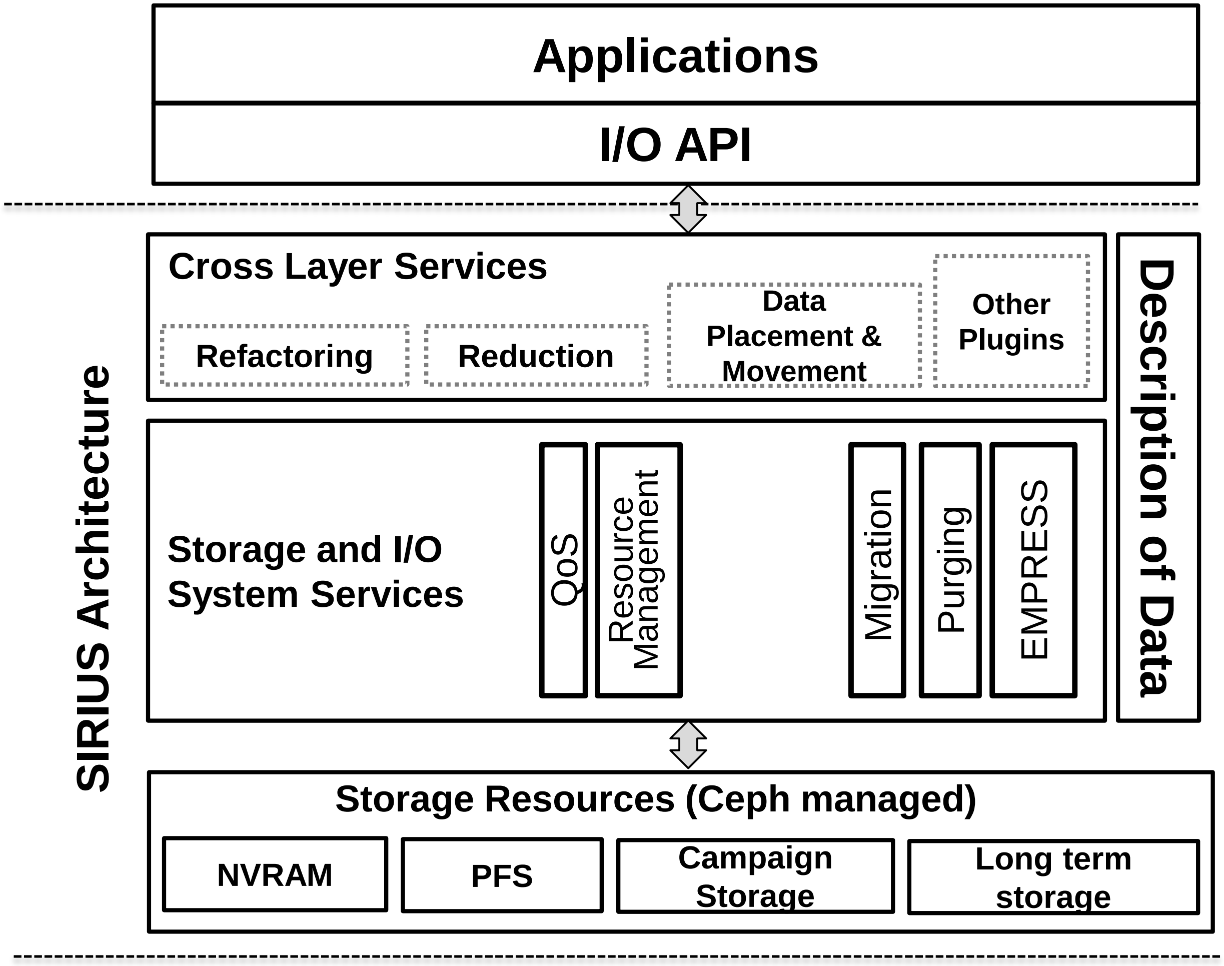
- FastBit – offers data querying based on values, but very limited support for spatial queries and attributes

Why not use a Key-Value Store?

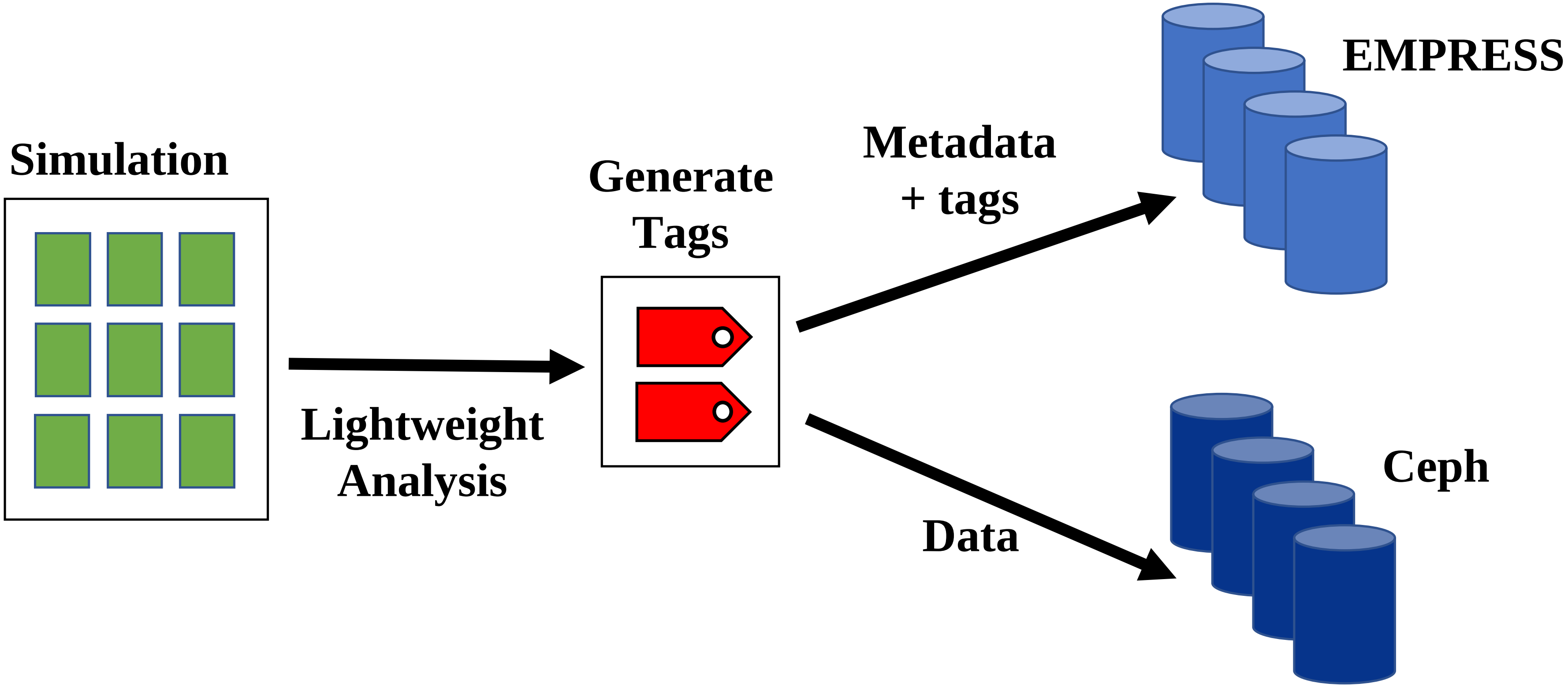
- Custom keys can go a long way, but not far enough
- Two Problems:
  - Inexact matches
  - Custom Metadata

- Relational databases with indices are radically faster at searching like this

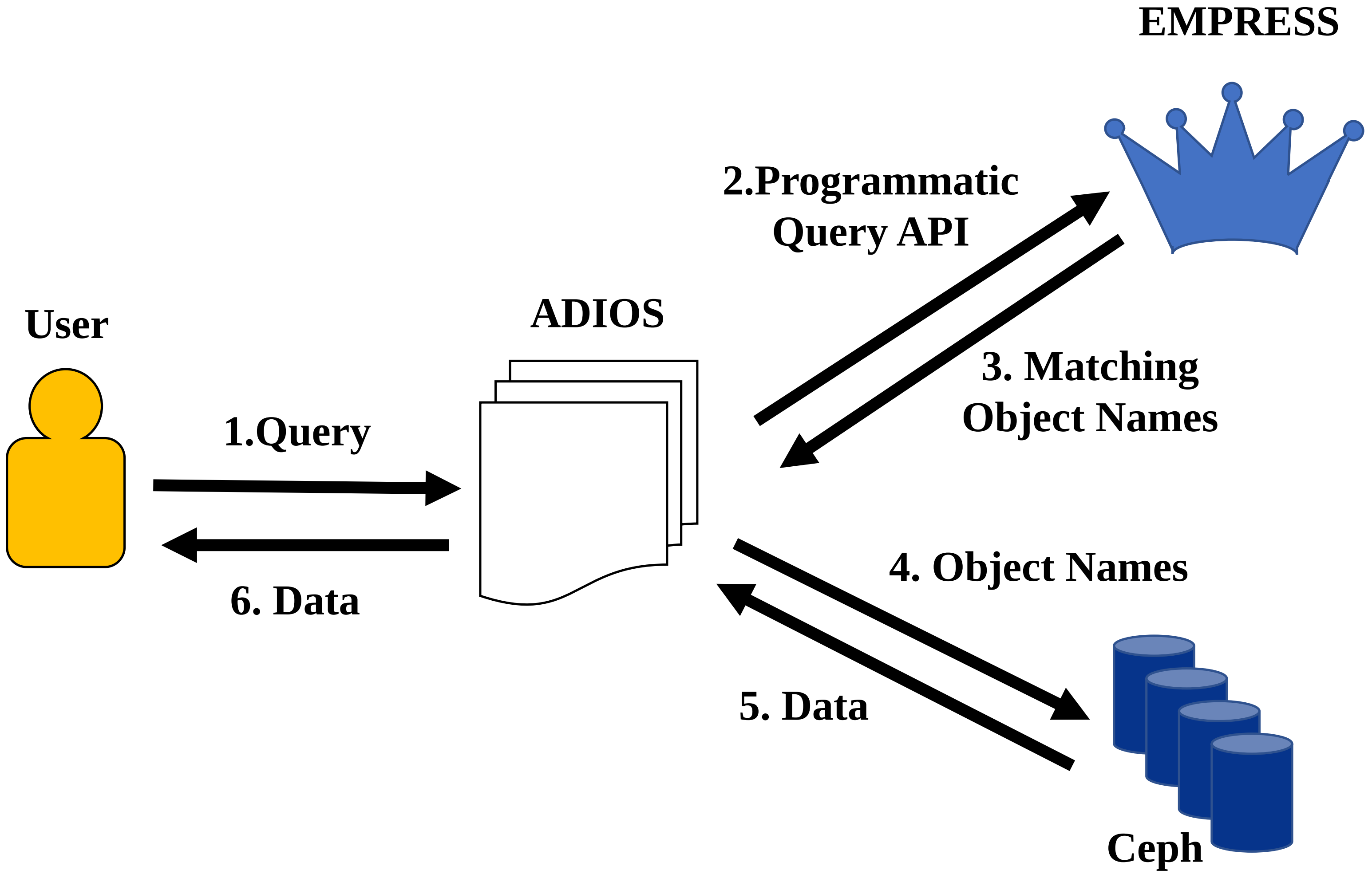
SIRIUS Architecture



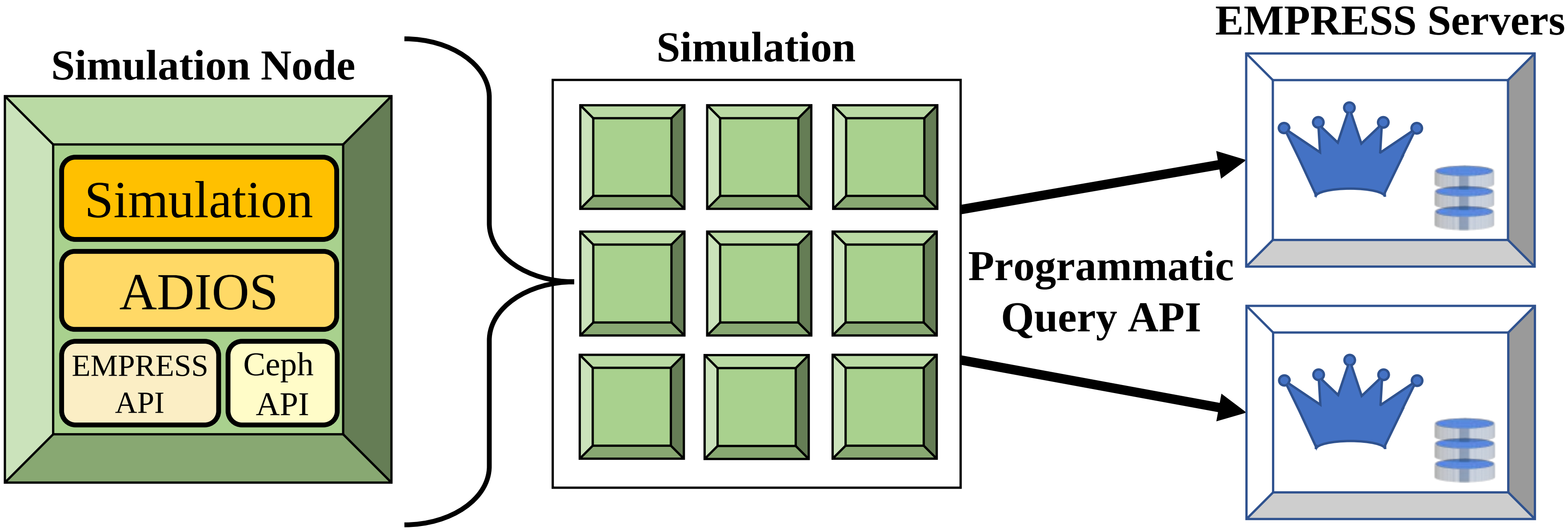
SIRIUS Workflow – Write Process



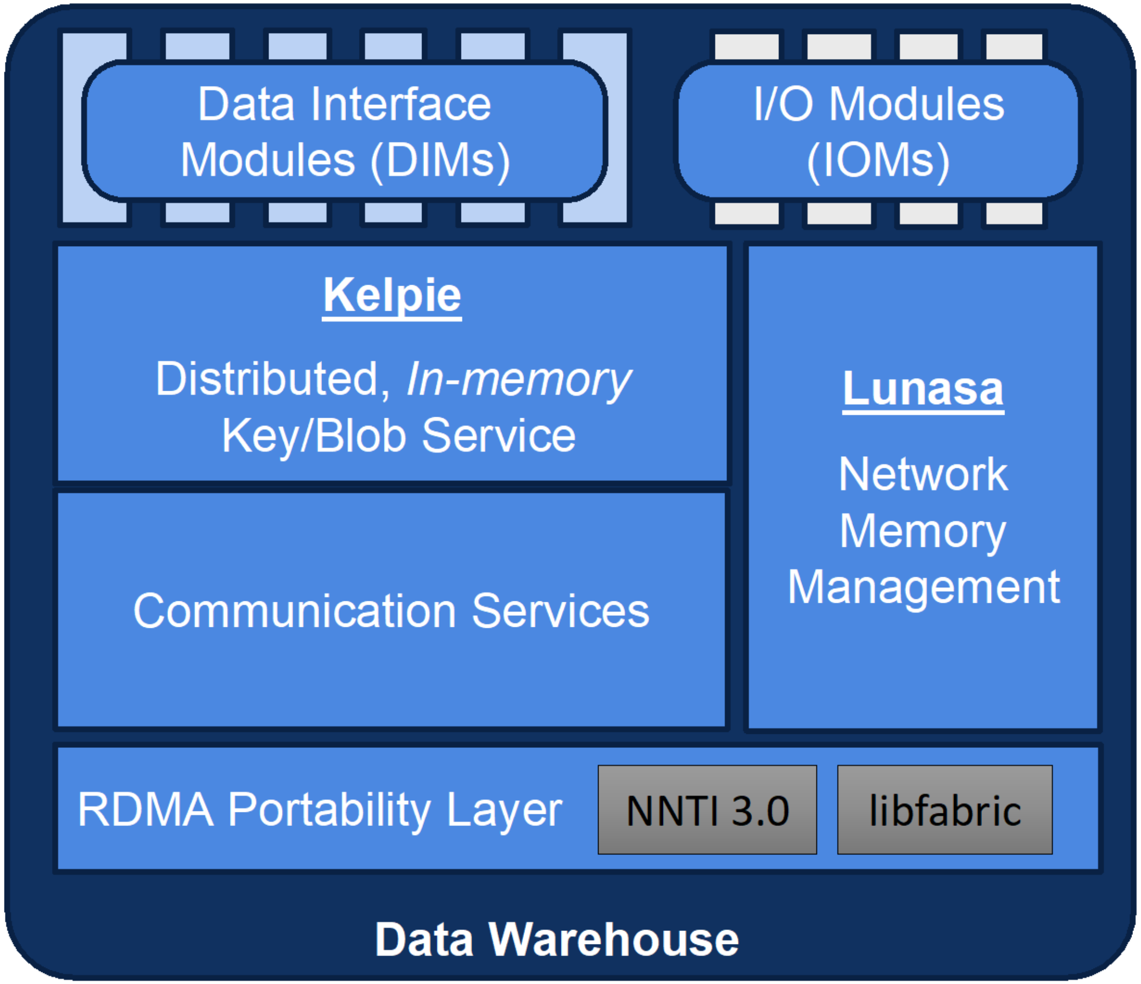
SIRIUS Workflow – Read Process



High Level Design



Faodail



Data Interface Modules (DIMs)
I/O Modules (IOMs)
Kelpie
Distributed, *In-memory* Key/Blob Service
Lunasa
Network Memory Management
Communication Services
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

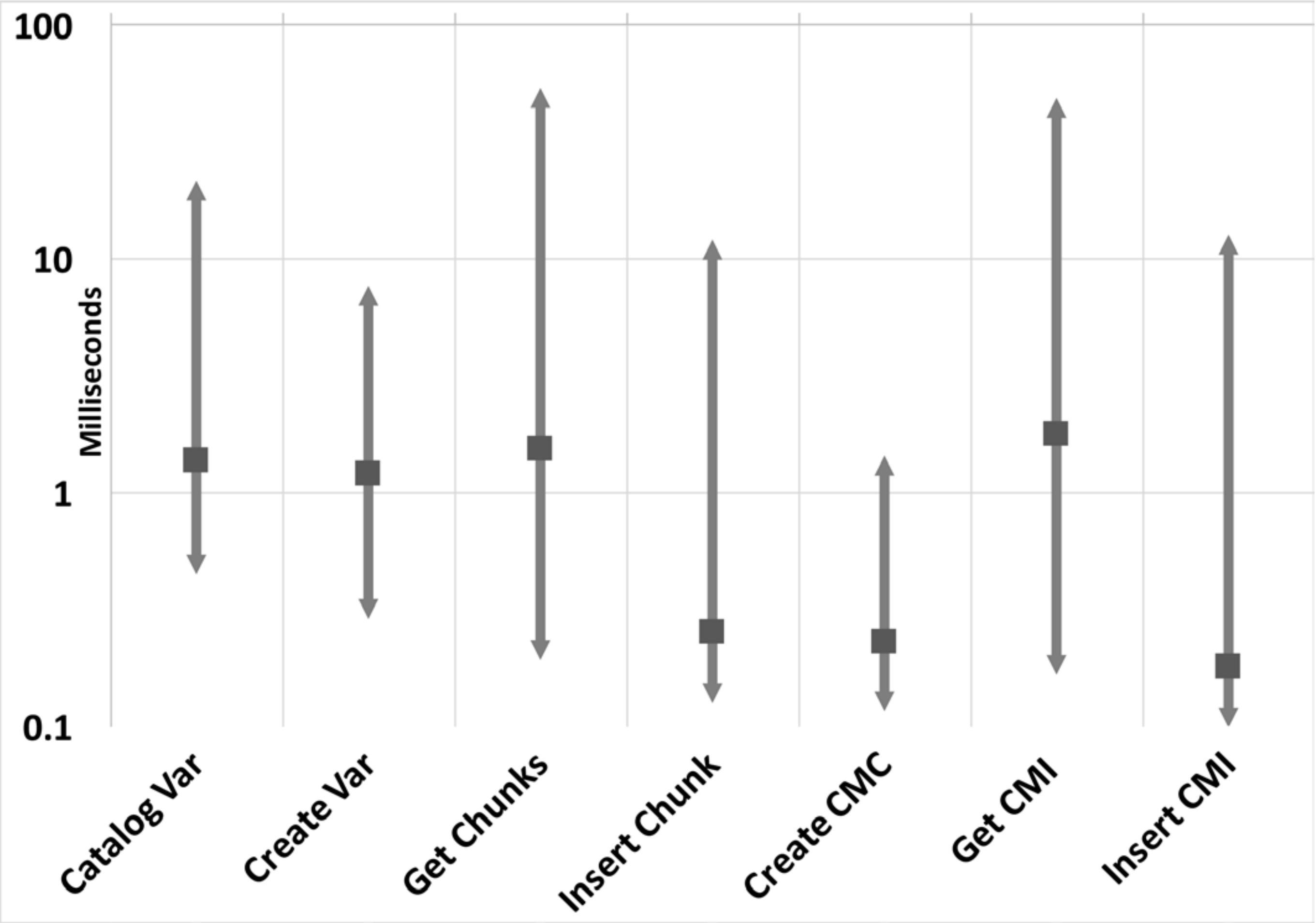
Testing Goals

- 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)

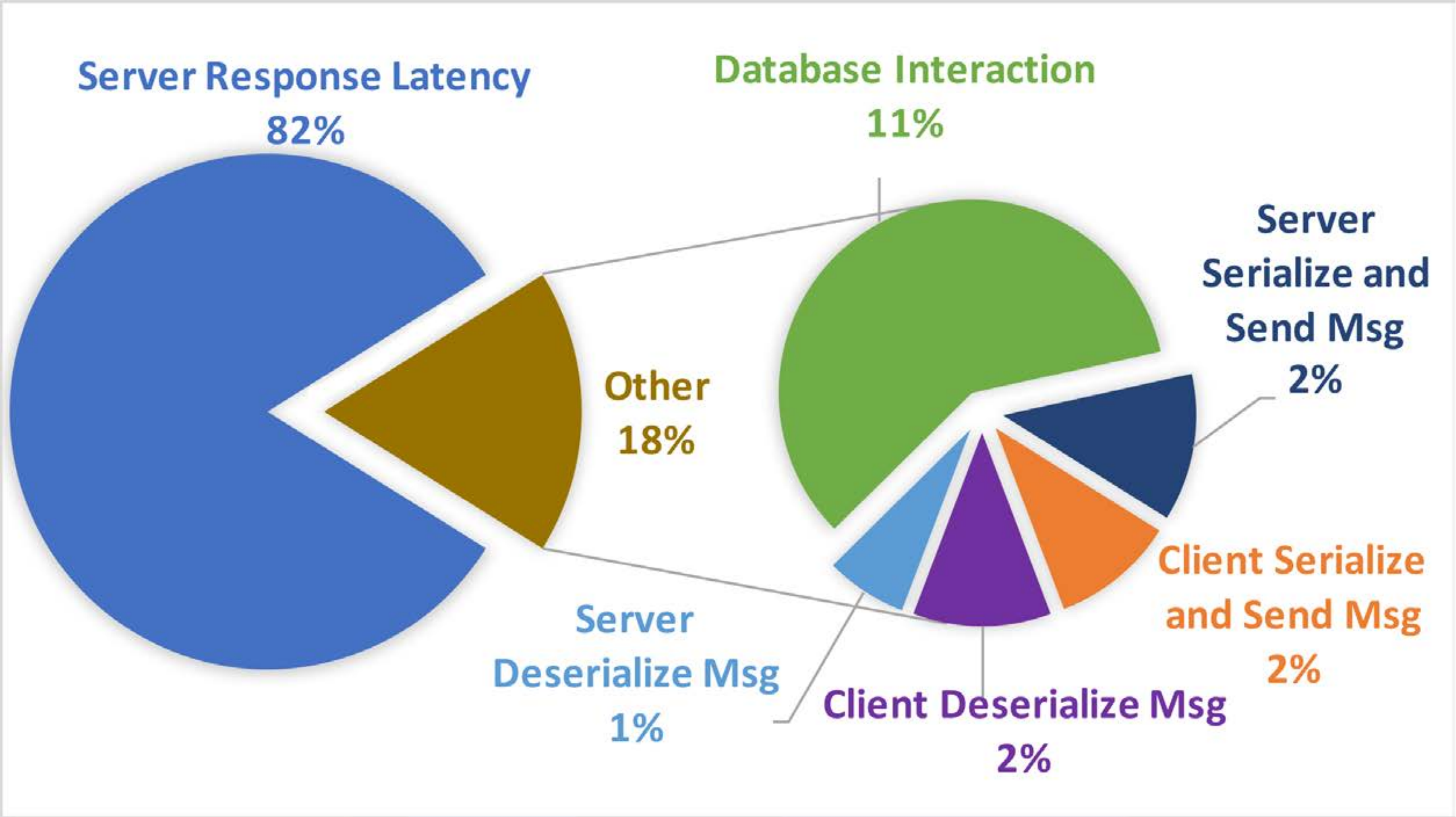
- Proof of concept, can EMPRESS efficiently support:
  - Common writing operations
    - 2 datasets written, each with 10 globally distributed 3-D arrays
  - Common reading operations
    - 6 different read patterns that scientists frequently use (Lofstead, et al. "Six Degrees of Scientific Data")
  - 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)

Testing – Query Times



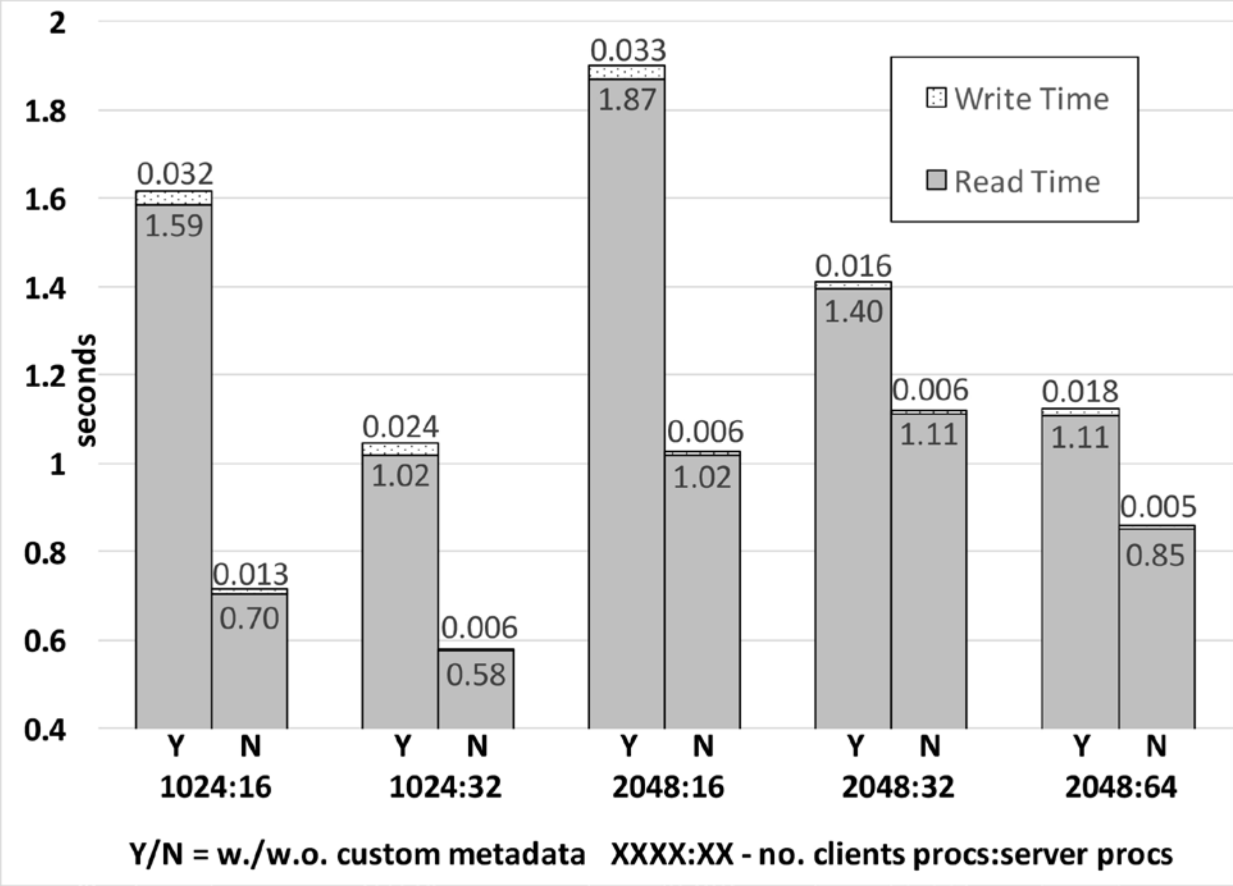
- EMPRESS efficiently supports a wide variety of operations including custom metadata operations

Testing – Chunk Retrieval Time



- Most time is spent waiting for the server to respond
  - Room for improvement in the Faodail infrastructure

Testing – Writing and Reading Time



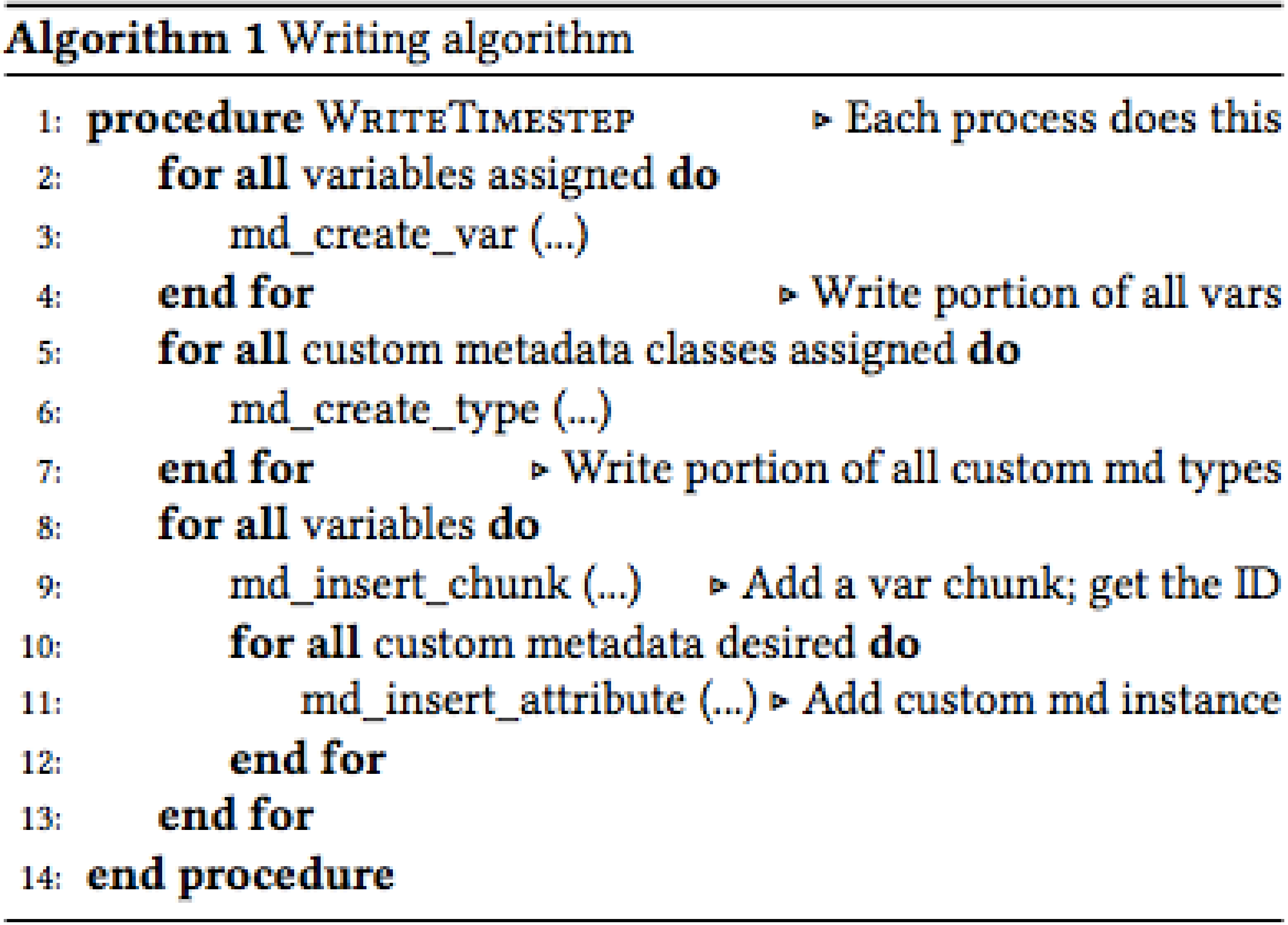
- Good scalability for fixed client-server ratio
- No significant overhead for adding custom metadata
- Client-server ratio greatly affects performance

Future Work

- Increasing EMPRESS’ flexibility, efficiency, and scalability
  - Support more queries
  - Different metadata distribution?

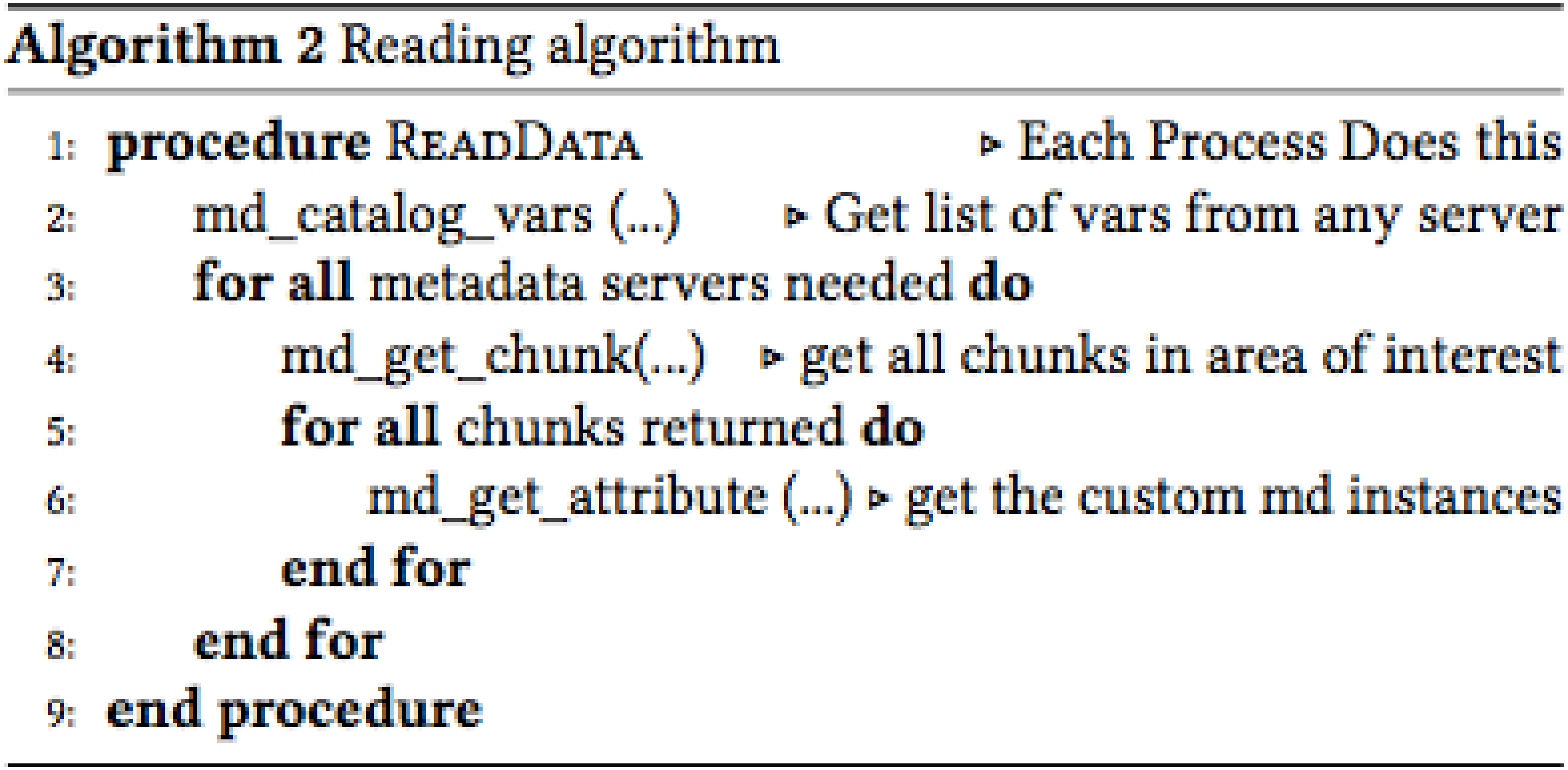
Acknowledgements

- 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

```
procedure WriteTimestep                                  ▷ Each process does this
    for all variables assigned do
        md_create_var (...)
    end for                                              ▷ Write portion of all vars
    for all custom metadata classes assigned do
        md_create_type (...)
    end for                                ▷ Write portion of all custom md types
    for all variables do
        md_insert_chunk (...)                       ▷ Add a var chunk; get the ID
        for all custom metadata desired do
            md_insert_attribute (...)               ▷ Add custom md instance
        end for
    end for
end procedure
```



Algorithm 2 Reading algorithm

```
procedure ReadData                                        ▷ Each Process Does this
    md_catalog_vars (...)                  ▷ Get list of vars from any server
    for all metadata servers needed do
        md_get_chunk(...)             ▷ get all chunks in area of interest
        for all chunks returned do
            md_get_attribute (...) ▷ get the custom md instances
        end for
    end for
end procedure
```