Exceptional service in the national interest

Sandia National Laboratories

Using a Robust Metadata Management System to Accelerate Scientific Discovery at Extreme Scales

Margaret Lawson, Jay Lofstead

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

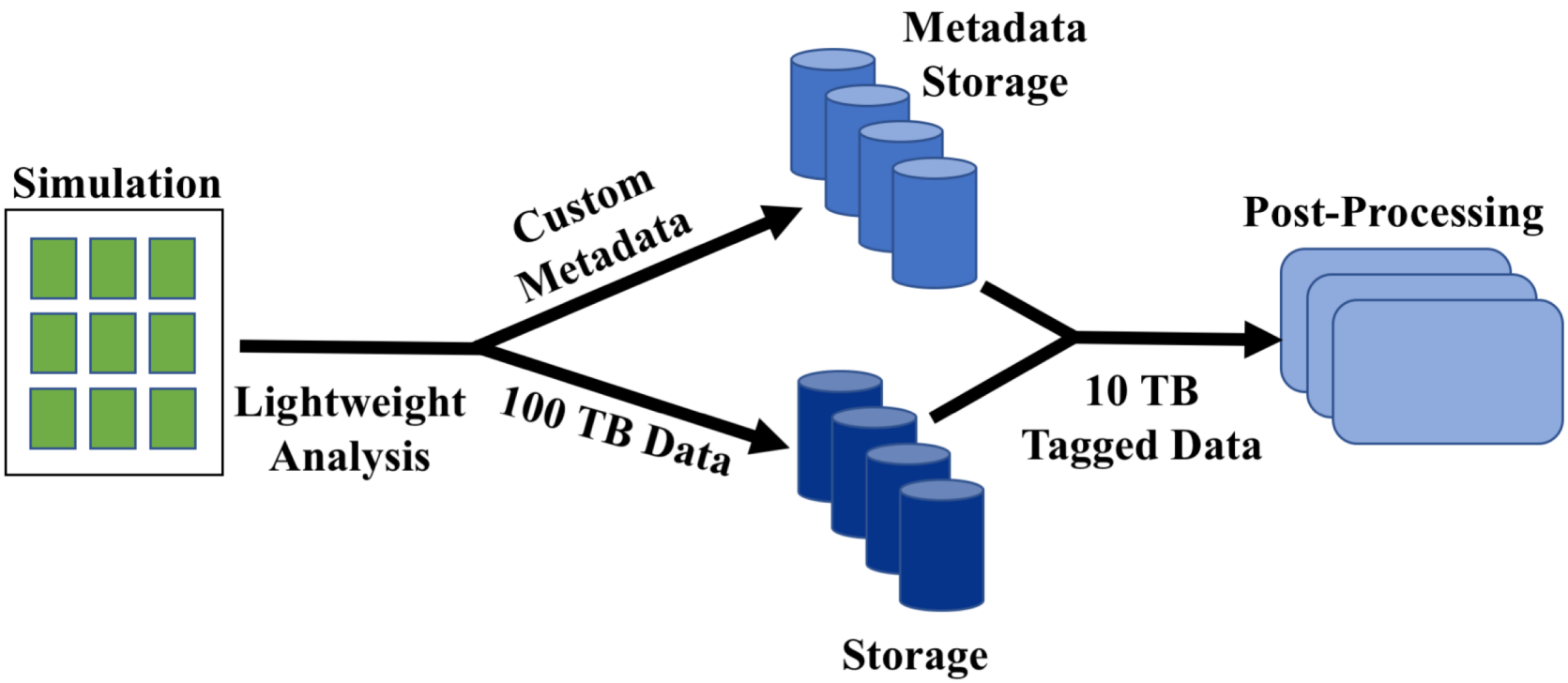
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Problems Faced

Sandia National Laboratories

- A single output can produce a dataset in the terabyte to petabyte range
- Large datasets are very slow to move and search
- Scientists have limited allocations of computational resources

Custom Metadata Solution



Previous Work - EMPRESS 1.0

Sandia National Laboratories

- Proof of concept
  - Rich, custom metadata management can be supported with reasonable efficiency and scalability
- Next steps
  - Improving the efficiency, scalability, and functionality to create a viable production system

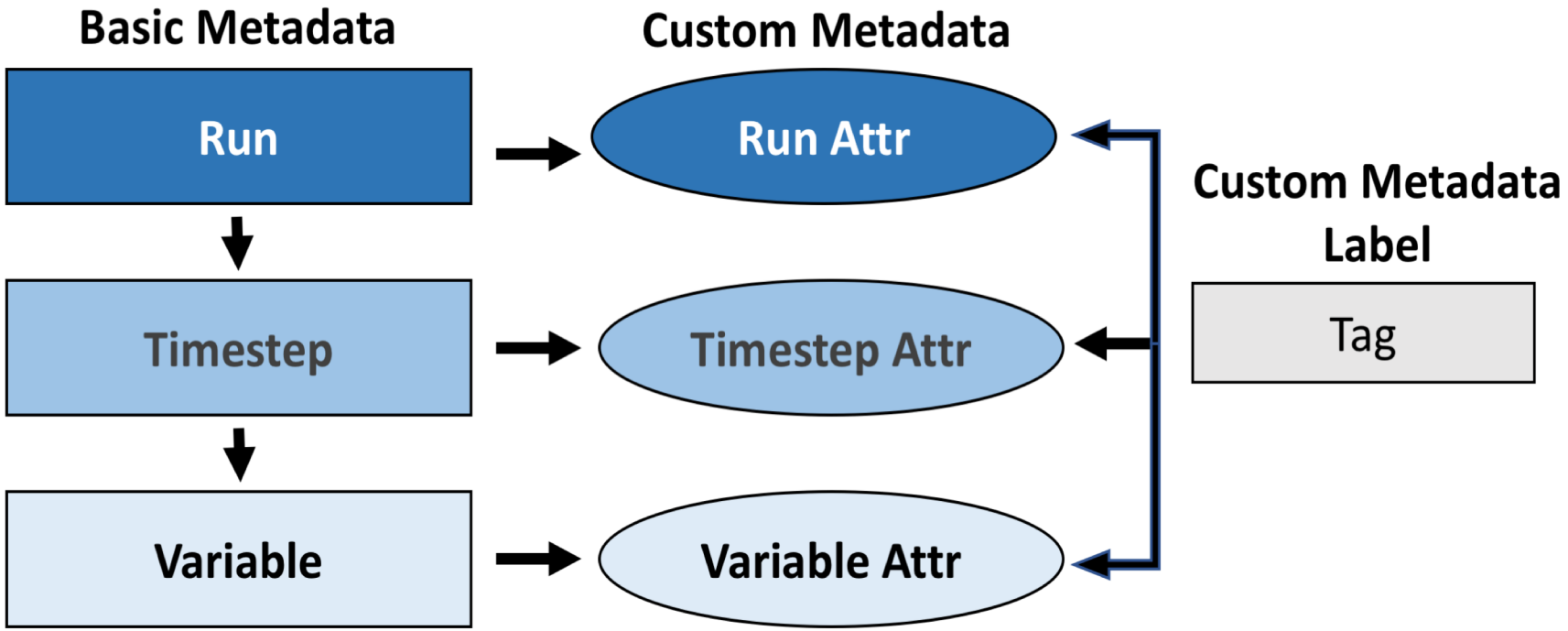
Paper Contributions

Sandia National Laboratories

- EMPRESS 2.0
  - Queries
  - Atomic operations
  - Fault tolerance
  - Portability

- RDBMS is a viable HPC technology for data-oriented metadata

Metadata Model



Custom Metadata Queries

Sandia National Laboratories

- Supports a wide variety of queries including global, spatial, temporal and multivariate
  - E.g., list all runs or timesteps that contain a “blob” near the reactor edge

Atomic Operations

Sandia National Laboratories

- Low overhead transactions
  - Transactions are atomic (committed in their entirety or aborted)
  - Metadata is given a transaction id that determines its external visibility
    - Eliminates the need for locks or blocking of service
  - The implementation is largely based on the D^2T system[1]

Fault Tolerance

Sandia National Laboratories

- Users can choose how to recover from failures occurring at the function, transaction, and hardware levels

- Basic metadata may be redundant, preventing data loss
  - E.g., if used with an I/O system

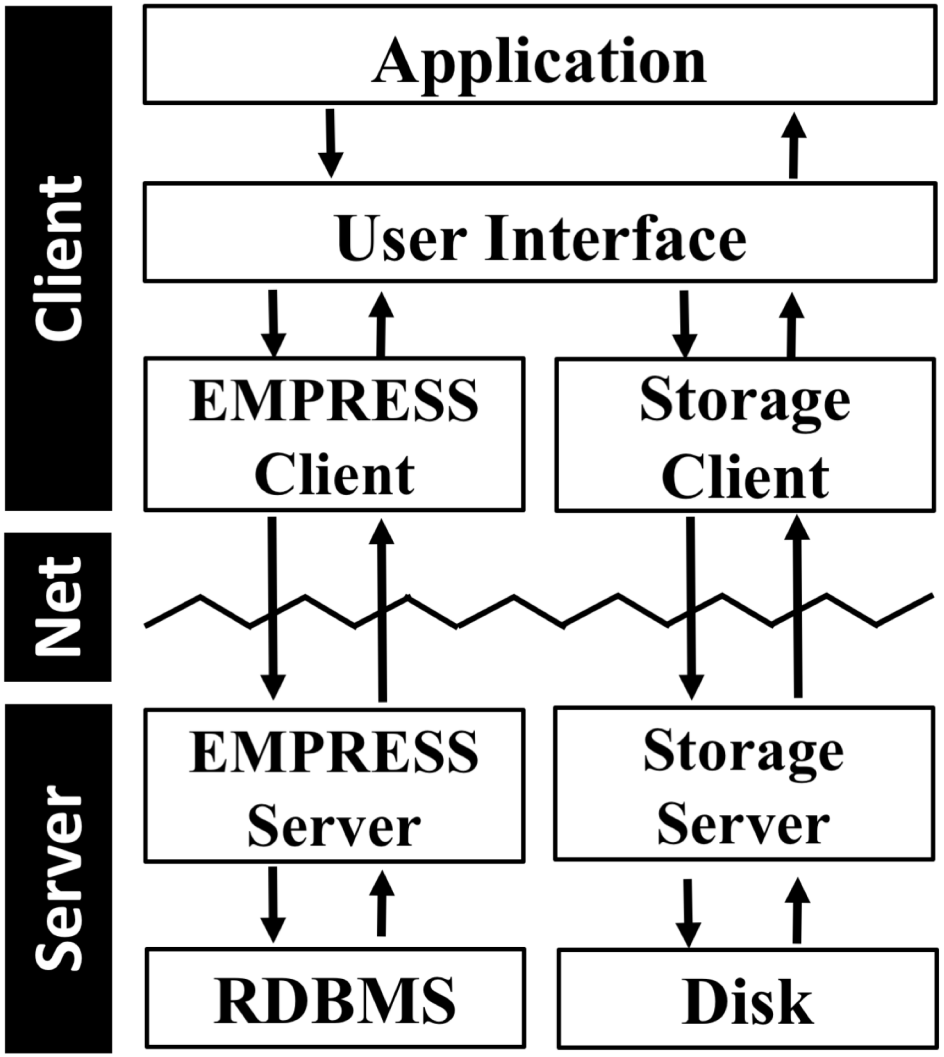
Portability

Sandia National Laboratories

- Directly storing the names of associated data objects limits portability and scalability

- EMPRESS 2.0 does not store the names, it uses a function to generate them
  - All EMPRESS metadata is portable

Implementation



Evaluation - Experiment Types

Sandia National Laboratories

Test Type	# Write Procs	# Read Procs	# Metadata Servers
EMPRESS 2.0 + HDF5	1000	100	1
EMPRESS 2.0 + HDF5	2000	200	2
EMPRESS 2.0 + HDF5	4000	400	4
HDF5	1000	100	N/A
HDF5	2000	200	N/A
HDF5	4000	400	N/A

Evaluation – Write Process

Sandia National Laboratories

- Run structure:
  - One application run, three timesteps, ten 3-D variables
- Data
  - Each process writes 0.4GB of data (10% of RAM) per timestep
- Custom metadata:
  - 10 different tags of varying frequency
  - On average, each process writes 26 attributes per timestep (2.6 per variable)

Evaluation – Read Process

1. 6 common read patterns[2] are performed including
   1. An entire variable
   2. A plane and partial plane in each dimension
   3. A 3-D subspace

2. Custom metadata is used to identify potential features of interest and the associated data is read in

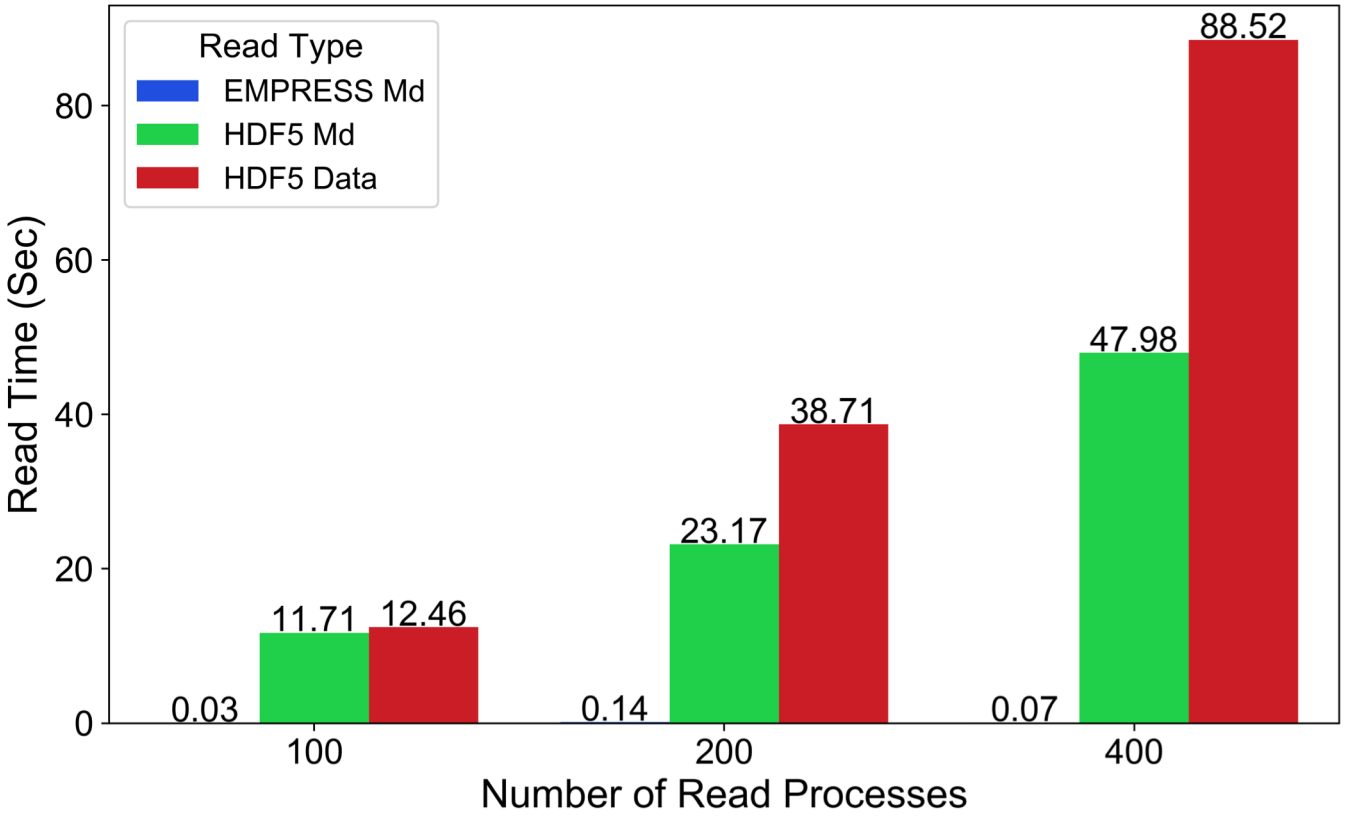
Evaluation – Writing

Sandia National Laboratories

# Write Procs	Data Write	EMPR Md Write	EMPR Md Overhead	HDF5 Md Write	HDF5 Md Overhead
1000	1753s	1.53s	0.09%	0.66s	0.04%
2000	3852s	1.65s	0.04%	1.80s	0.05%
4000	7406s	1.56s	0.02%	3.22s	0.04%

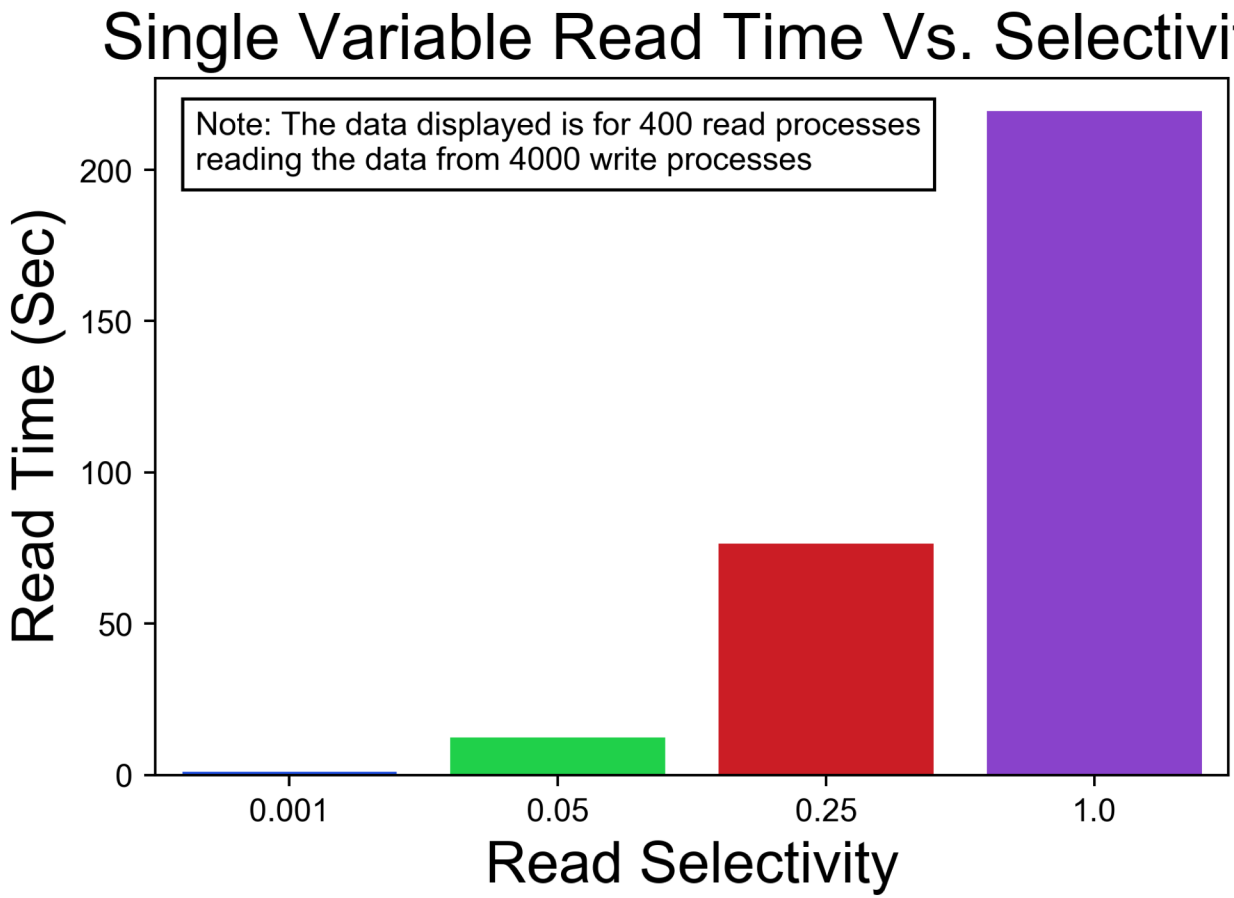
- Both can do efficient metadata writes at the evaluated scales
  - But EMPRESS can scale out to achieve constant performance

Evaluation – Metadata Read



- HDF5 takes almost as long to do the metadata query as it does to read the data

Evaluation – Accelerating Data Reads



- EMPRESS can significantly accelerate data reads by limiting the scope to data of interest

Future Work - EMPRESS

Sandia National Laboratories

- Evaluation
  - Potential bottlenecks & solutions
  - Comparison to more alternatives
  - NoSQL vs RDBMS
- Functionality
  - Expanding the application classes that EMPRESS can support

Conclusions

Sandia National Laboratories

- Custom metadata is an important tool for accelerating post-processing
- Current I/O tools cannot efficiently support custom metadata services
- EMPRESS 2.0 offers insights on the functionalities needed for a production system & how to implement them scalably

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.

Citations

- [1] J. Lofstead, J. Dayal, K. Schwan, and R. Oldfield, “D2t: Doubly dis- tributed transactions for high performance and distributed computing,” in *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*. IEEE, 2012, pp. 90–98.
- [2] J. Lofstead, M. Polte, G. Gibson, S. Klasky, K. Schwan, R. Oldfield, M. Wolf, and Q. Liu, “Six degrees of scientific data: reading patterns for extreme scale science IO,” in *Proceedings of the 20th international symposium on High performance distributed computing*, ser. HPDC ’11. ACM, 2011, pp. 49–60. [Online]. Available: http://doi.acm.org/10.1145/1996130.1996139