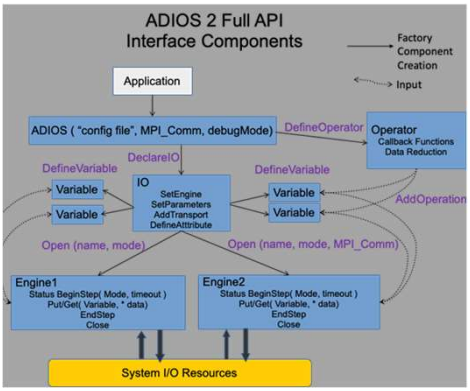
Optimizing Data Management in HPC with ADIOS2: A Comparative Study

Raven Campbell

Ana Gainaru

Define The Problem

Traditional I/O methods like CSV and POSIX falter in I/O intensive projects due to limited scalability and speed. ADIOS2, developed at Oak Ridge National Laboratory, transforms high-performance computing with fast I/O, self-describing data formats, and advanced compression. Its modular design allows for optimized performance in demanding applications such as climate modeling, significantly outperforming traditional methods.



Performance

To demonstrate the performance gap, we benchmarked a relatively modest dataset of 96,453 records of hourly meteorological data (temperature, humidity, and wind speed). This dataset, though relatively small, serves to showcase the shortfalls of conventional CSV approaches as well as the potential efficiency improvements achievable through the application of ADIOS2.

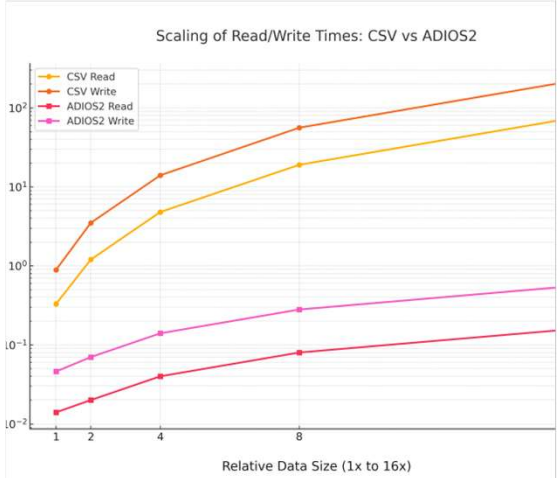
Read Time:
It took just 0.0142 seconds to read ADIOS2, while reading CSV took 0.3277 seconds — a 95% improvement.

Write Time:
ADIOS2 wrote the dataset in 0.0464 seconds, while CSV wrote it in 0.8866 seconds, thus making ADIOS2 nearly 20 times faster.

The benefits of ADIOS2 are evident even when compared in this limited dataset: higher data transfer rates, reduced I/O overhead, and improved scalability. As datasets become larger, these benefits become more and more important, and thus ADIOS2 is an especially well-suited candidate for high-performance computing and data-intensive computing.

Implementation and Case Studies

ADIOS2 has shown impressive performance gains over traditional CSV and text-based formats, especially in dealing with large-scale scientific data sets. For example, for surface water quality data with over 460,000 observations, CSV read and write times tend to scale linearly or worse because of text parsing overhead, leading to multi-second to minute-order latency for I/O. In contrast, ADIOS2 read and write operations consistently achieve sub-second performance even when data sets are doubled or quadrupled.



Why This Dataset is a Classic HPC Problem:

- It covers several decades, necessitating historical examination.
- It contains millions of readings if it is scaled or mixed with related environmental data.
- Analysis typically requires time-windowed or region-based subsetting (querying).
- Future scaling will probably be more than millions of records since sensors are still gathering data.

In related applications, ADIOS2 may be utilized in:

- Climate Simulation (E3SM, CESM)
- Nuclear Reactor Simulations (NEAMS)
- Environmental Monitoring (Flood, Air, or Water)
- Scalable Machine Learning Pipelines on High-Performance Computing

Future Directions

ADIOS2 is under development to enhance data compression, real-time processing, and usability for industrial and scientific applications. Planned developments involve campaign management to facilitate the grouping of related datasets together under one metadata file for querying efficiently, and derived variables that are computed from the primary data. Users will also benefit from remote data access, local visualization of distant datasets, and adaptive data retrieval, which prioritizes approximate data based on accuracy needs to enhance performance. All these will ensure ADIOS2 is more scalable, more efficient, and indispensable for high-performance, data-heavy workflows.

Conclusion

ADIOS2 has been a game-changer for data management in the HPC environment, setting new standards for performance, flexibility, and scalability. Its novel architecture, supporting fast I/O operations, complex data querying, and high-performance data compression, has been essential in enabling the massive data requirements of modern computational research. ADIOS2 greatly enhances the effectiveness and productivity of research activities by dramatically reducing read and write times, as well as enabling high-throughput data processing. This breakthrough enables scientists and engineers to achieve accelerated breakthroughs in their respective fields.

OAK RIDGE National Laboratory
THE UNIVERSITY OF TEXAS UTSA AT SAN ANTONIO
U.S. DEPARTMENT OF ENERGY
PELLISSIPPI STATE COMMUNITY COLLEGE
THE UNIVERSITY OF TENNESSEE KNOXVILLE