A Lightweight I/O Scheme to Facilitate Spatial and Temporal Queries of Scientific Data Analytics

Presenter: Yuan Tian

tiany@ornl.gov

University of Tennessee/Auburn University

Zhuo Liu Bin Wang
Weikuan Yu

Auburn University

Tom Clune

NASA Goddard Space Flight Center

Shujia Zhou

Northrop Grumman Corporation

Scott Klasky
Hasan Abbasi

Oak Ridge National Laboratory

Jeremy Logan

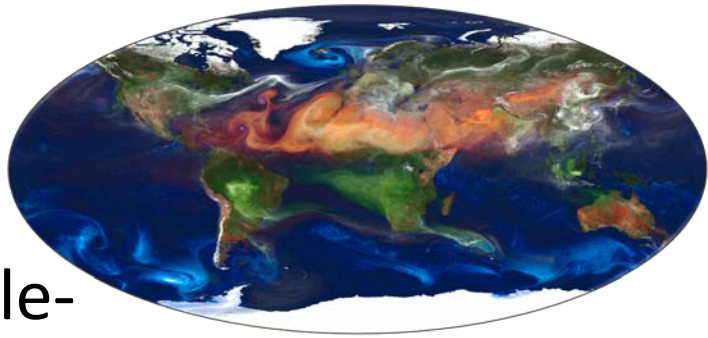
University of Tennessee

Managed by UT-Battelle
for the Department of Energy

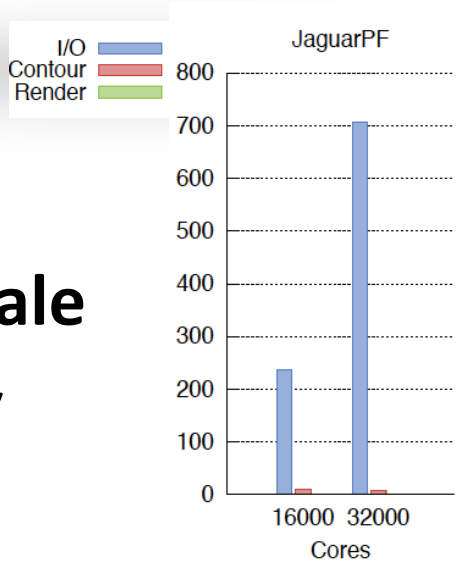
Outline

- Motivation and background
- STAR – Spatial and Temporal Aggregation I/O scheme
- Experimental results
- Conclusion

Motivation and Background



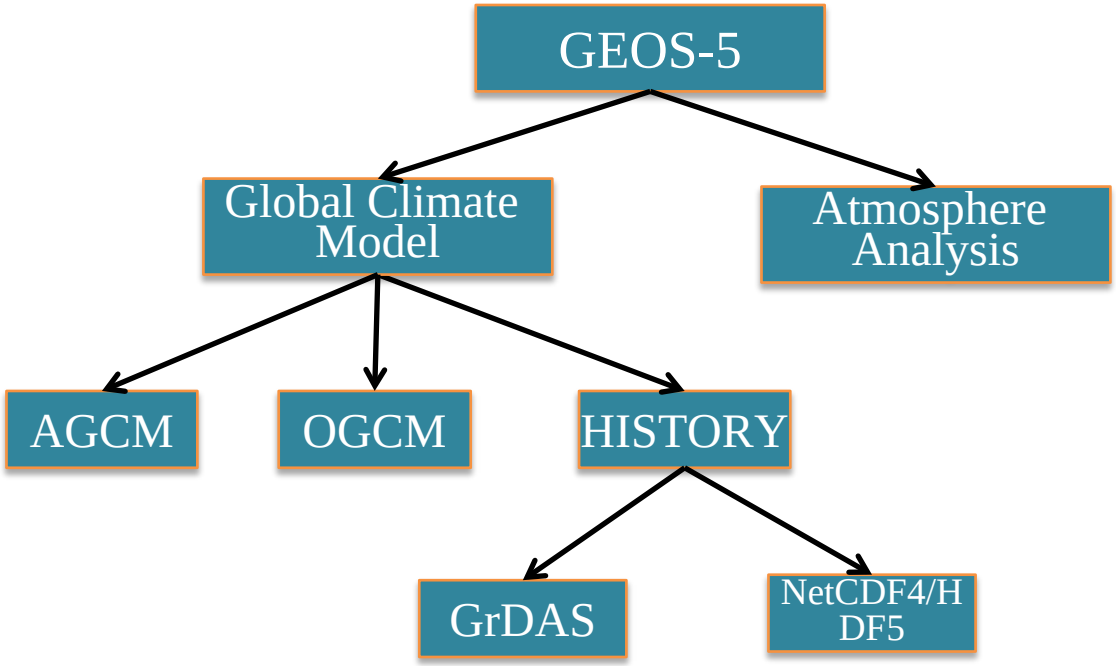
- I/O performance of common scientific data **access patterns** is bottle-necked on HPC system
  - ~90% of execution time spent on I/O in extreme scale visualization[1]
- **Small** variables are difficult to optimize **at scale**
- **Temporal** and **Spatial** queries are commonly performed, yet poorly supported
- Challenges to be addressed:
  - *How to reorganize scientific data storage to enable fast spatial and temporal queries?*
  - *How do we construct such an organization efficiently without degrading the write performance?*



[1] H. Childs, D. Pugmire, S. Ahern, B. Whitlock, M. Howison, Prabhat, G. Weber, and E. Bethel. Extreme scaling of production visualization software on diverse architectures. Computer Graphics and Applications, IEEE, 30(3):22 –31, may-june 2010.

Case Study of GEOS-5

- A system of models integrated using the Earth System Modeling Framework (ESMF) for earth system simulation from NASA
- History component – output diagnosis data
- Two I/O methods: GrDAS and NetCDF/HDF5



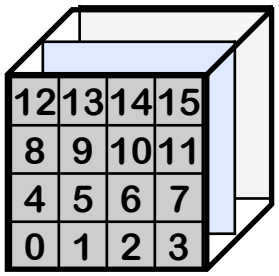
AUBURN UNIVERSITY

RDAV

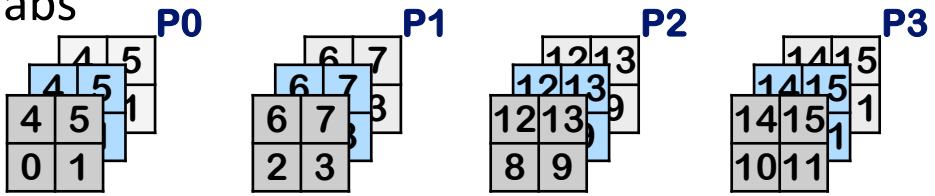
OAK RIDGE National Laboratory

Legacy GEOS-5 I/O Structure

- Data characteristics: many variables, small sizes
  - Variable sizes range from 100KB to 300MB
- Each multidimensional variable is written out in 2D slice to maintain the logically contiguous data layout
  - 3D variable is written out in 2D hyperslabs
  - Loop through all the variables in turn

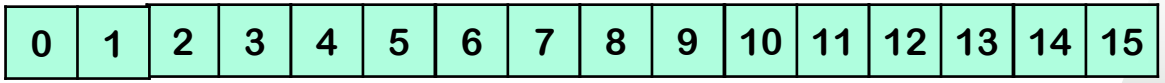


4*4*3 3D variable with 2-D decomposition



1. Send to aggregator process

2. Memory rearrangement at aggregator

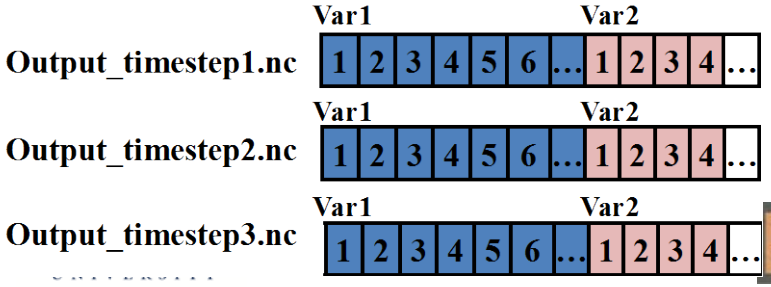


3. Send to root process

P_{root}

4. I/O

Storage



RDAV

OAK RIDGE National Laboratory

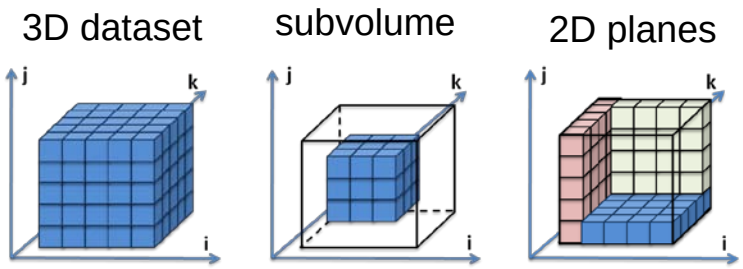
Issues with Legacy I/O Flow

- Requires significant n-m and m-1 communication
- Requires significant memory rearrangement
- Doesn't leverage parallel storage system
- Poor scalability
- Data organization is not optimized for common analysis patterns: spatial and temporal

Such issues are shared with many other applications

Research Targets

- Deficiency of current aggregation techniques
  - Inter-node based: network overhead
  - Intra-node based: limited scope of aggregation
- Deficiency in supporting **spatial** data analytics
  - Common access patterns
    - Read in all of a single variable.
    - Read an arbitrary orthogonal subvolume
    - Read an arbitrary orthogonal full plane
- Deficiency in supporting **temporal** data analytics



Spatial Access Patterns
fastest dimension: **k**; **slowest** dimension: **i**

AUBURN UNIVERSITY

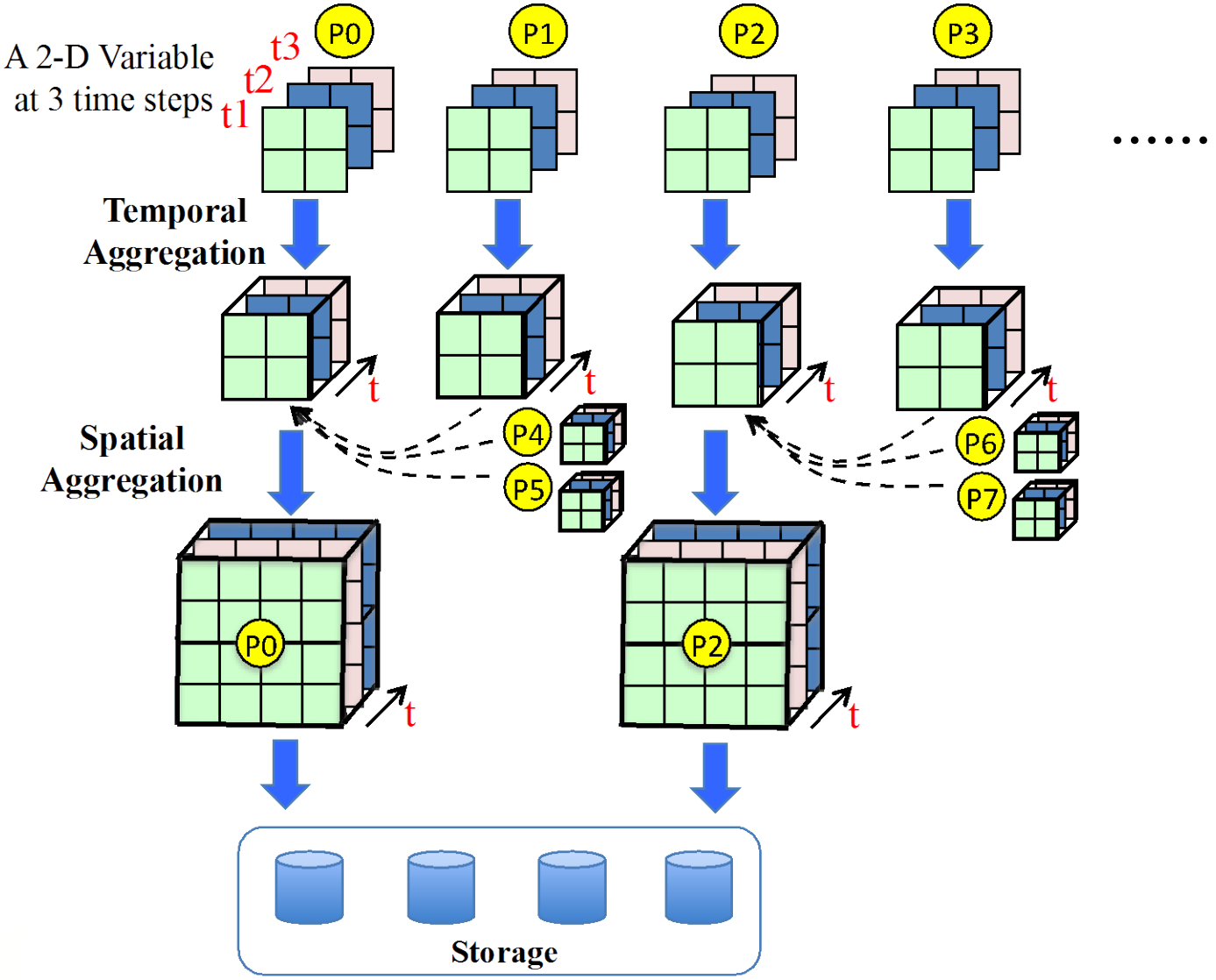
RDAV

OAK RIDGE National Laboratory

STAR - Spatial and Temporal AggRegation

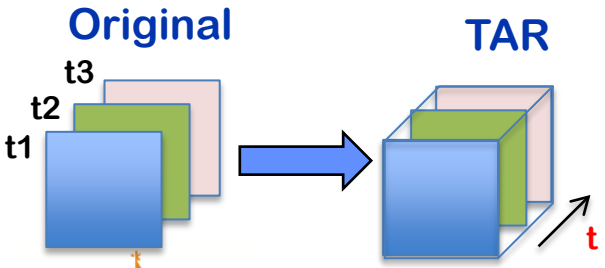
- A **lightweight** I/O layer between application and storage system
- Dynamically construct data across both spatial and temporal dimensions into optimized chunks during output
  - Enable a fully parallelized and high performance I/O flow for scientific applications
  - Exploit the spatial and temporal relationships between variables to further consolidate data
  - Provide a data organization that facilitates the common access patterns of data post-processing
- It consists of **two** key algorithms:
  - Temporal Aggregation (TAR)
  - Spatial Aggregation (SAR)

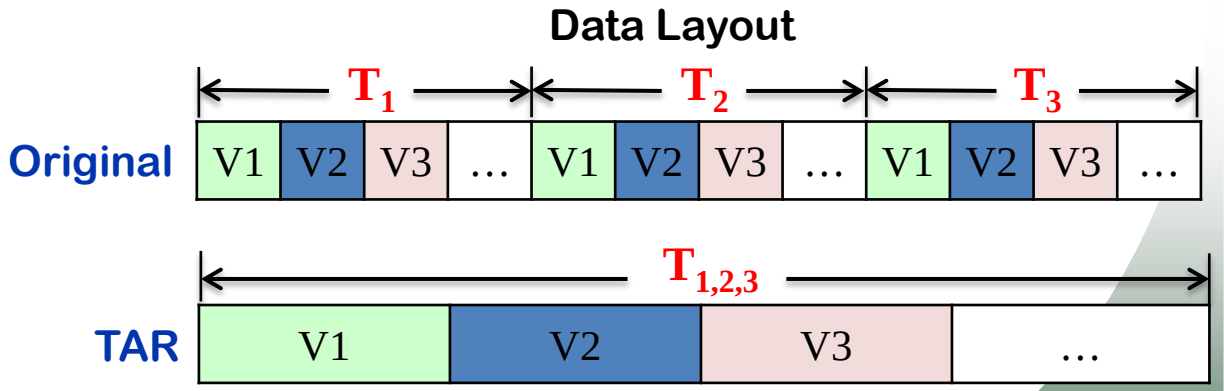
Data Movement of STAR



Temporal Aggregation (TAR)

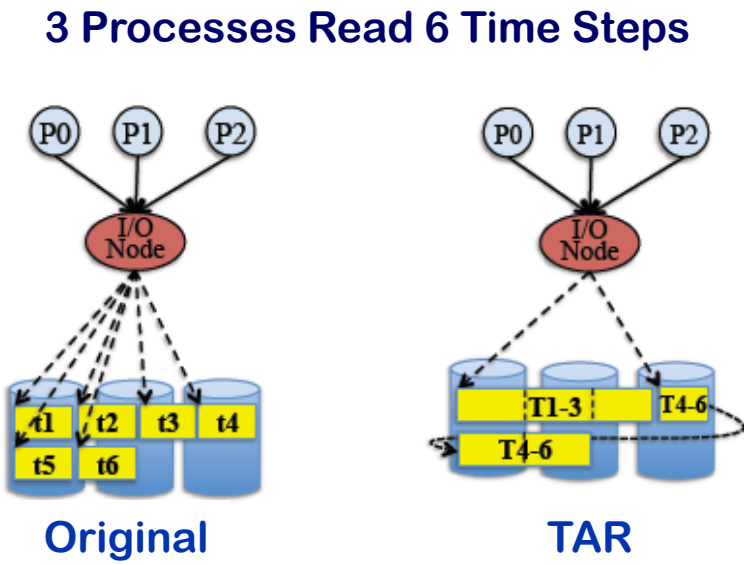
- Temporal aggregation is to open up another horizon to further consolidate data
- Data of multiple time steps are buffered at each process
- Data is written out only at the last time step or reaches the boundary of memory capacity
- Benefits for writing:
  - Number of I/O requests decreases linearly with the degree of aggregation
  - No communication overhead incurred

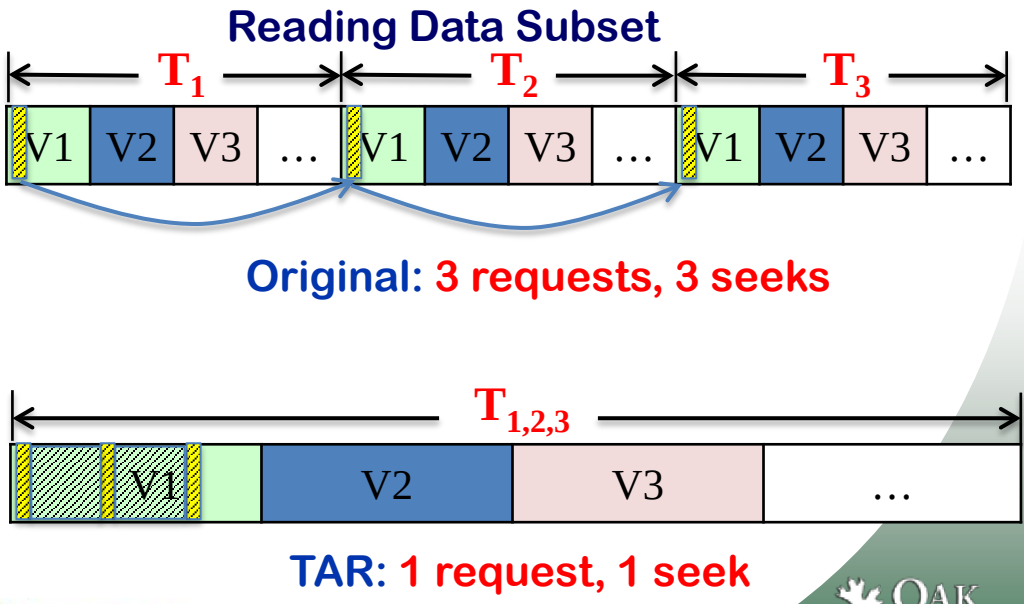




Temporal Aggregation for Reading

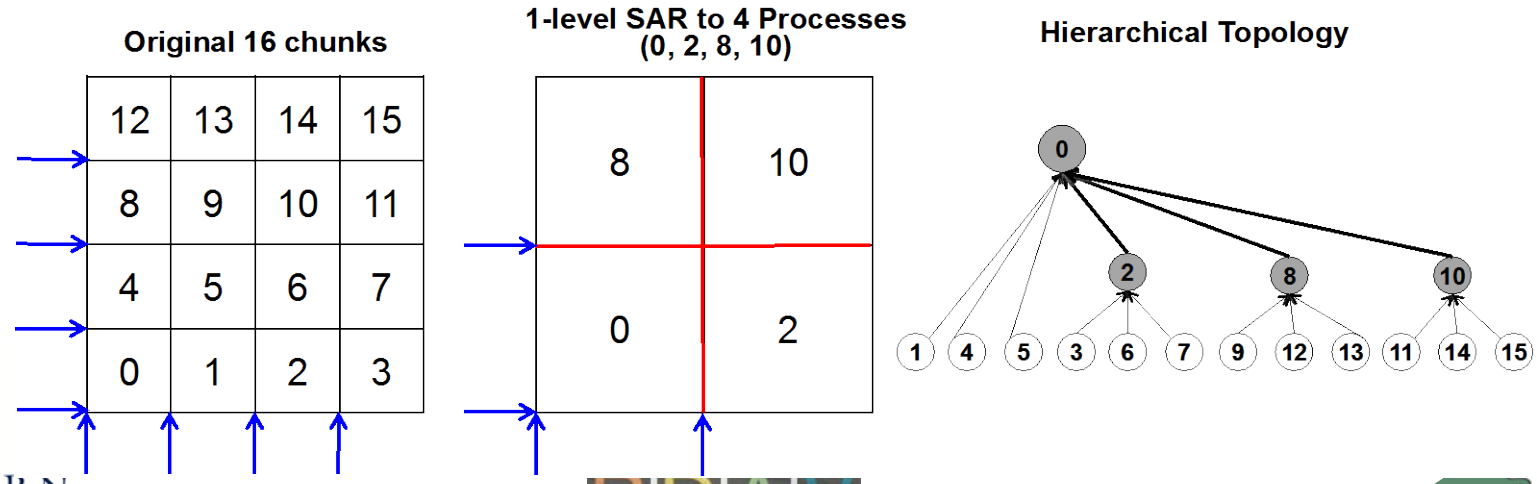
- Number of read request is reduced linearly with the degree of aggregation
- Number of expensive seek operations is reduced for analytics on temporal dimension
- Less contention and interference at storage





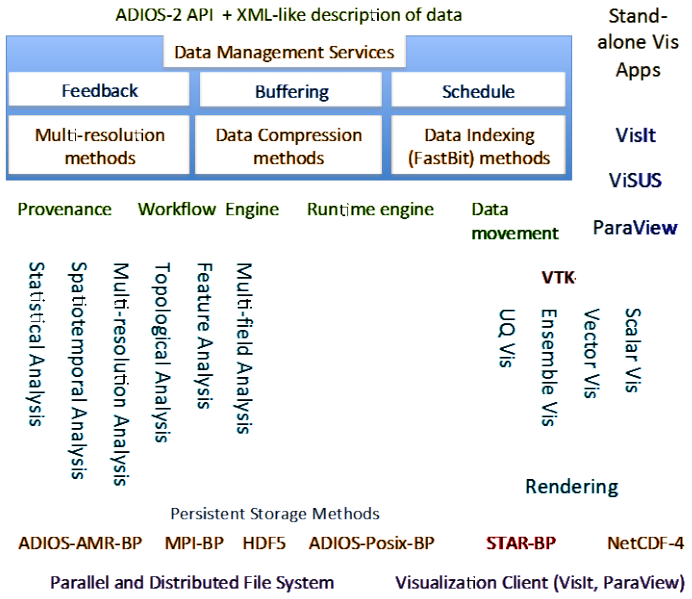
Spatial Aggregation (SAR)

- Chunking data layout leads to large number of small chunks for each variable at scale
  - Lots of seek and read requests are required
- Simply concatenating data chunks doesn't solve above issue
- Spatial Aggregation with hierarchical topology
  - Spatial locality of every data point is reserved
  - Writing: less writers, less contention at storage during output
  - Reading: improve read performance for common spatial access patterns

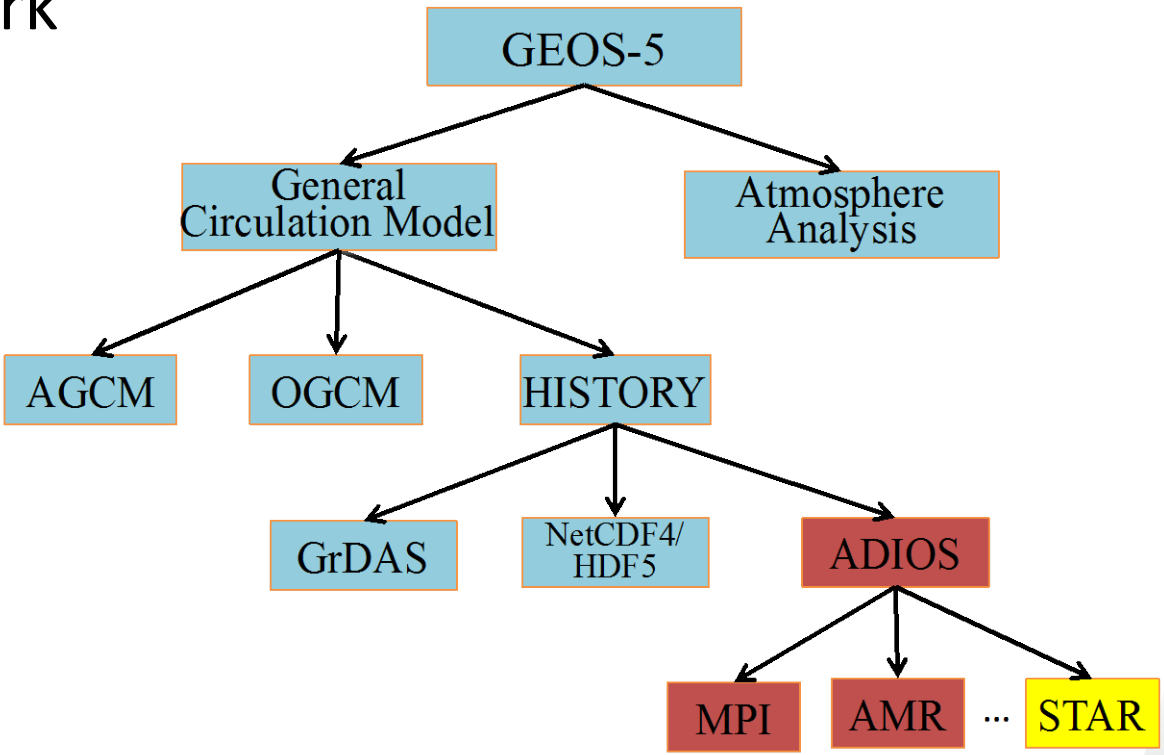


Implementation and Integration with GEOS-5

- STAR is implemented within Adaptable I/O System (ADIOS) I/O framework



ADIOS Framework

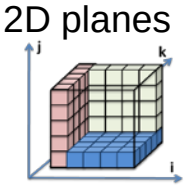


GEOS-5 with ADIOS

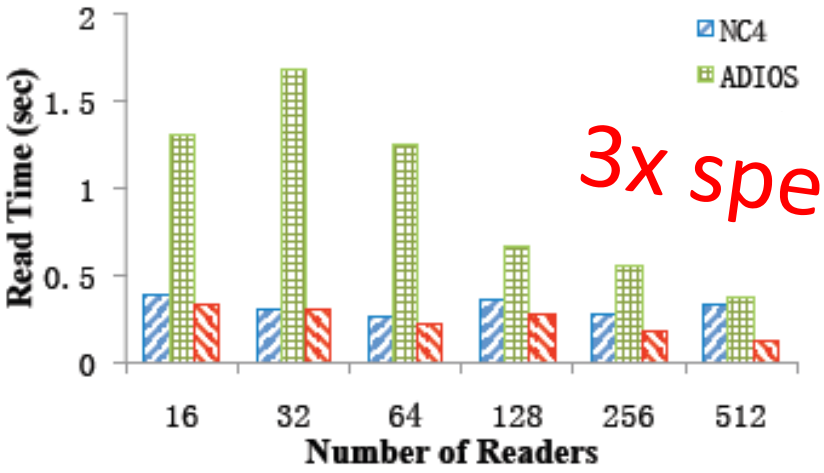
Experiment Platform

- Jaguar Supercomputer at ORNL
  - 18,688 compute nodes
  - Each node contains one 16-core Opteron processor
  - 32GB memory
- Lustre filesystem called Spider
- GEOS-5 employs 2-D domain decomposition
  - 7 bundles consists of 185 2-D variables and 80 3-D variables
  - Output resolution: 576*361*48 (half degree), and 1152*761*48 (quarter degree)
  - Total output size: 3.12GB (half), and 12.4GB (quarter) per time step
  - 30 time steps are generated
- Data organization: NetCDF-4, Original ADIOS and STAR

Planar Read for 1 Time Step

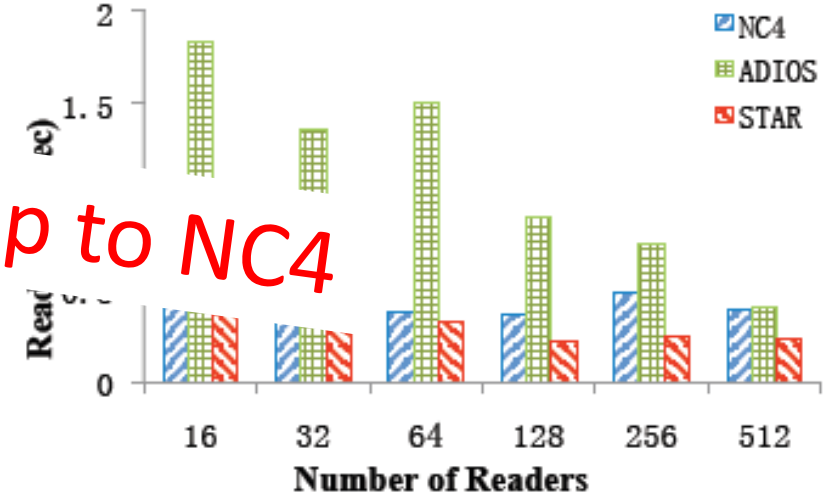


- Variables are generated with 4,096 processes
- 3 2-D slices are read out (k, j), (k, i) and (j, i)
- NC4 suffers from noncontiguous data on the slow dimensions
- Original ADIOS suffers from large amount of read requests



(a) Half Degree 3-D Variable (Dimension: 576x361x48)

3x speedup to NC4



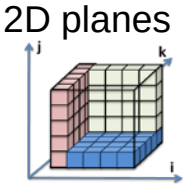
(b) Quarter Degree 3-D Variable (Dimension: 1152x761x48)

AUBURN UNIVERSITY

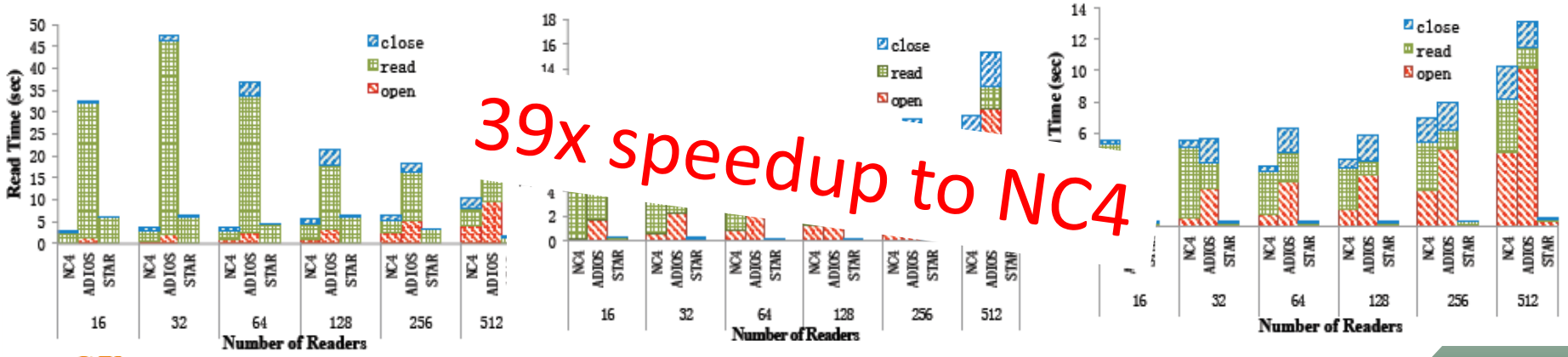
RDAV

OAK RIDGE National Laboratory

Planar Read for 30 Time Steps

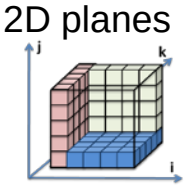


- 30 output files for NC4 and ADIOS, 1 file for STAR
- Variables (half-degree) are generated with 4,096 processes
- 3 2-D slices are read out from 30 time steps
- Both NC4 and ADIOS spends long open and close time
- Chunking-based data layout shows good performance in slow dimensions

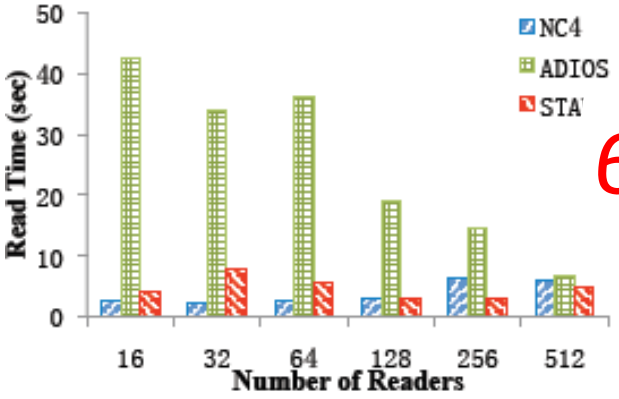


(k, j, t) (k, i, t) (j,i,t)

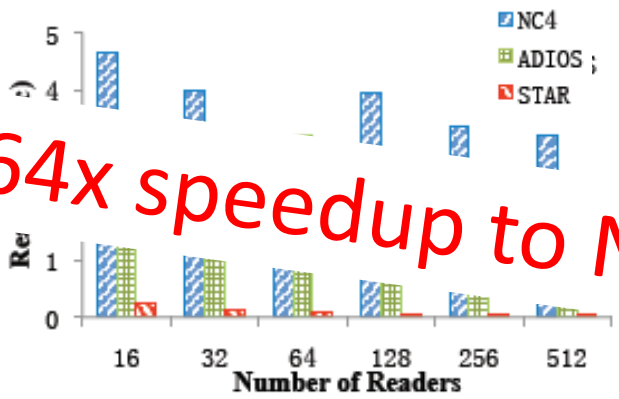
Planar Read for 30 Time Steps



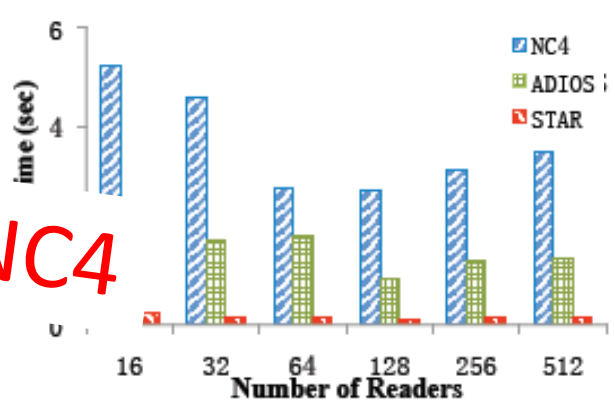
- 1 output file for all test cases, quarter degree simulation
- Only read time is shown
- NC4 suffers from noncontiguous data on the slow dimensions
- ADIOS suffers from small data chunks
- STAR achieves balanced and improved read performance



(k, j, t)



(k, i, t)



(j,i,t)

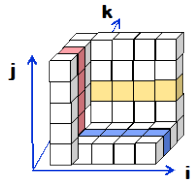
64x speedup to NC4

AUBURN UNIVERSITY

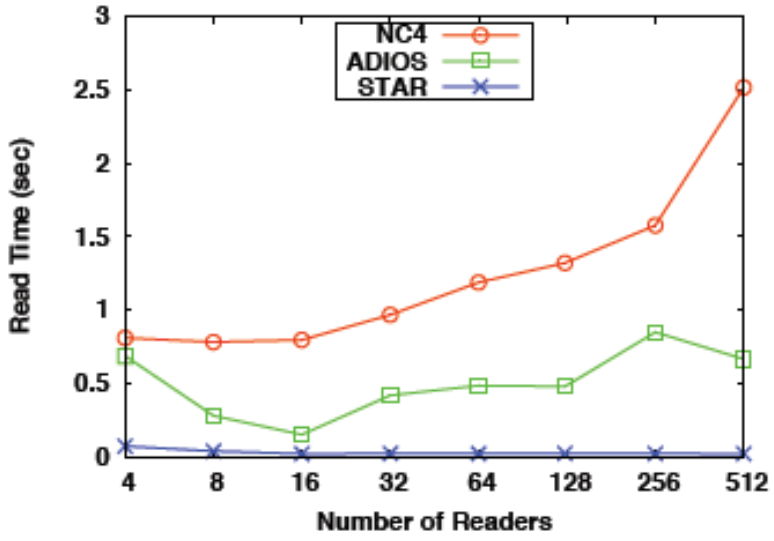
RDAV

OAK RIDGE National Laboratory

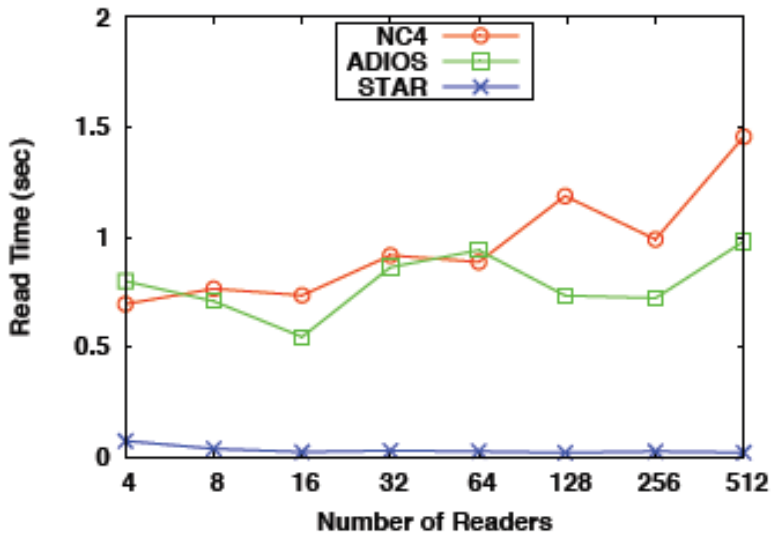
Reading 1-D subset on 30 Time Steps



- A 1-D subset is read out across 30 time steps
- 1 output file for all data layouts
- 74x speedup to NC4 is achieved by STAR



(k, t)



(j, t)

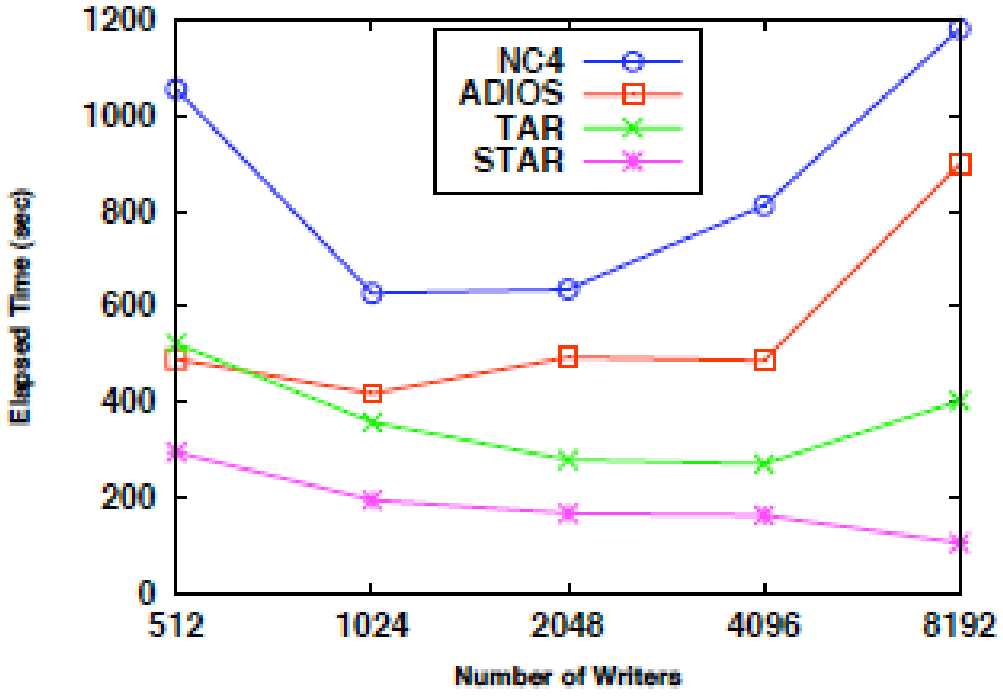
AUBURN UNIVERSITY

RDAV

OAK RIDGE National Laboratory

Write Performance

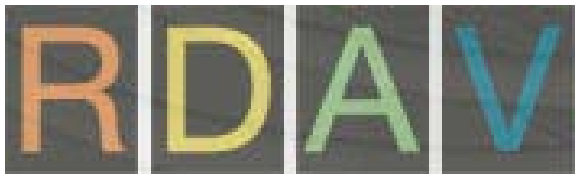
- STAR improves both write performance and scalability
- 11x improvement to NC4 and 4x improvement to TAR



Conclusion

- STAR is able to improve both write and read performance for GEOS-5 through is dynamic data organization strategies
- Temporal Aggregation opens up a new dimension to consolidate data
  - Larger data output is constructed for writing
  - Facilitates the data analytics on time dimension
- Spatial Aggregation further consolidates small data chunks
  - Number of I/O requests are reduced for both writing and reading
- Maximum of 11x speedup is achieved for writing, and 73x speedup is achieved for reading

Sponsors of our research



SciDAC
Scientific Discovery through Advanced Computing

U.S. DEPARTMENT OF ENERGY

AUBURN UNIVERSITY

Questions?