A System-Aware Optimized Data Organization for Efficient Scientific Analytics

Yuan Tian¹ Scott Klasky² Weikuan Yu¹ Hasan Abbasi¹ Bin Wang¹ Norbert Podhorszki² Ray Grout¹ Matthew Wolf³

Auburn University¹ Georgia Institute of Technology³
Oak Ridge National Laboratory² National Renewable Energy Laboratory⁴

Problem Statement

• Scientific applications generates massive amount of multi-dimensional arrays
• Read performance is crucial for application execution and data post-processing
• Existing data layouts produces imbalanced read performance for common access patterns of post-processing due to:
  - Inefficiency to alleviate the dimension dependency for common access patterns
  - Poor data concurrency on large-scale storage systems

OUR GOAL - A new data layout provides GOOD and BALANCED read performance for scientific data post-processing.

Performance Evaluation

• Planar Readers - balanced and improved performance, maximum of 66 times speedup
• 4,096 writers, up to 512 readers
• Peak performance comparison among Logically Contiguous (LC), Chunking (ORG), and our new data organization (NEW)

Optimized Chunking

• Goal: mathematically find the Optimized Chunk Size (OCS) that gives the balance between the overhead of seek/read operations and redundant data retrieval.

System-Aware Data Organization

• TWO level of data reorganization:
  - Intra-chunk level: Constructing data chunks into OCS
  - Chunk level: Reorganize data chunks using Space Filling Curve

Acknowledgement

This work is funded in part by a UT-Battelle grant to Auburn University, and in part by National Science Foundation award CNS-1059376. This research is also supported by an UT-Battelle grant (UT-B-400103043) to Auburn University. It used resources of the NCCS at ORNL, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725

Reference: