A System-Aware Optimized Data Organization for Efficient Scientific Analytics

Yuan Tian
Dept. of Computer Science
Auburn University
tianyua@auburn.edu

Scott Klasky
Oak Ridge National Lab
Oak Ridge, 37831
klasky@ornl.gov

Weikuan Yu
Dept. of Computer Science
Auburn University
wkyu@auburn.edu

Hasan Abbasi
Oak Ridge National Lab
Oak Ridge, 37831
habbasi@ornl.gov

Bin Wang
Dept. of Computer Science
Auburn University
bzw0012@auburn.edu

Norbert Podhorszki
Oak Ridge National Lab
Oak Ridge, 37831
pnorbert@ornl.gov

ABSTRACT

Large-scale scientific applications on High End Computing systems produce a large volume of highly complex datasets. Such data imposes a grand challenge to conventional storage systems for the need of efficient I/O solutions during both the simulation runtime and data post-processing phases. With the mounting needs of scientific discovery, the read performance of large-scale simulations has becomes a critical issue for the HPC community. In this study, we propose a system-aware optimized data organization strategy that can organize data blocks of multidimensional scientific data efficiently based on simulation output and the underlying storage systems, thereby enabling efficient scientific analytics. Our experimental results demonstrate a performance speedup up to 72 times for the combustion simulation S3D, compared to the logically contiguous data layout.

Categories and Subject Descriptors

D 4.3 [**Operating System**]: File Systems Management—*Access Method, File organization*

General Terms

Design, Experimentation, Performance

Keywords

I/O, Data Layout

1. INTRODUCTION

The increasing growth of leadership computing capabilities, in terms of both system complexity and computational power, has enabled scientific applications to solve complex scientific problems at large scale. Such phenomenon is accompanied by a gigantic volume of complex scientific data produced, driving the impetus for data intensive computing as a very significant factor in scientific computing.

Many efforts, both past and present, have focused on examinin and improving the I/O performance by studying the output side o the problem [7, 8], despite the importance of read performance i driving the scientific insight through scientific simulation, analysis workflows and visualization. Worse yet, current I/O technique often overlook the need of good read performance and, as a result, have a substantial negative impact on the read performance The main reason is the discrepancy between the physical limitations of linearly ordered magnetic storage and the common acces patterns [5] of the multidimensional scientific data. Significant "dimension dependency" [4] has been observed for range queries, du to the expense of coping with noncontiguity of data points with extra disk seeks or retrieving extra data between needed segments.

We propose a System-aware Optimized Data Organization strategy which uses 1) an Optimized Chunking model to produces th *ideal* sized data chunks based on the simulation output and systen parameters; and 2) a Space Filling Curve reordering to ensure th close-to-optimal data concurrency. The initial experimental result on the Jaguar Cray XT5 [3] supercomputer at Oak Ridge Nationa Laboratories (ORNL) demonstrated that our strategy is able to provide both good balance and high performance for challenging access patterns in scientific applications. Up to 72x speedup to the planar read is achieved for S3D [2] compared to the logically contiguous data layout.

2. OUR APPROACH

Chunking has been commonly recognized as an efficient data layout for multidimensional arrays [4]. For a chunk based data organization, the size of chunks plays a critical role in determining the read performance for query on a data subset.

Intuitively there is a *sweet spot* for the chunk size where the overhead on the seek operation (to traverse through the data chunks) and redundant data retrieval (to read extra data to avoid seeks) achieves the best balance for the slow dimension. Such a sweet spot is where the optimal read performance can be expected. We develop the Optimized Chunking model to pinpoint an *ideal* chunk size for a multidimensional data on a HPC system. We show that significantly varying the chunk size towards either direction results in performance degradation as the I/O becomes burdened by overhead or dominated by disk seeks. Given a 3-D variable, the relationship between the performance for planar read and the chunk size is shown in Figure 1, where N_{ocs} is the sweet spot.

Because dividing large data chunk breaks the contiguity on the

HPDC'12, June 18–22, 2012, Delft, The Netherlands.
ACM 978-1-4503-0805-2/12/06.

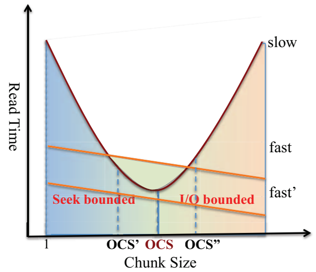
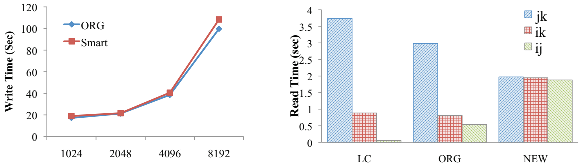


Figure 1: The Read Time vs. the Chunk Size



(a) Data Generation

(b) Peak Planar Read (Jaguar, 4,096 writers, 512 readers, log scale)

Figure 2: I/O Performance

fast dimension, the read performance degrades proportionally with the decreasing chunk size, where the data was originally contiguous. This results in two scenarios as shown in Figure 1, represented by two orange lines. In view of the general performance on all the dimensions, the fastest total read time may not incur at the point *OCS* but still within the *Optimized Region* of *OCS'* and *OCS"*. The performance difference inside such region is within a small margin. As this study is aimed at finding an *optimized* chunk size, we use our solution of *OCS* as the guidance for data organization. Our experimental results in Section 3 demonstrate that this value provides satisfactory performance. After a series mathematical derivation (the detail of the derivation can be found in [6], we can then determine the Optimized Chunk Size: $OCS = BW_{io} \times (CC + T_s)$, where $BW_{i/o}$ is the I/O bandwidth, T_s is the time unit for each seek operation, and CC is the communication cost unit. For a simplified analytical model, the external and internal interferences to the storage system are ignored. Such modeling can help pinpoint a solution that enables near-optimal I/O performance tuning in a timely fashion.

After the *correct* sized data chunks are produced, a Hilbert Space Filling Curve reordering is applied to shuffle the placement of the data chunks on storage targets. This approach is to achieve a close-to-optimal data concurrency for reading based on our earlier study [5].

3. EXPERIMENTAL EVALUATION

We have implemented our new approach within ADaptive I/O System (ADIOS) [1], and evaluated it on the *Jaguar* supercomputer at ORNL. A 3-D variable with a process local dimension of 256 × 256 × 256 is generated by S3D [2]. Based on our equation for optimized chunk size and system parameters, each data chunk is divided into 49 subchunks. Figure 2(a) shows such strategy only introduces a negligible write overhead.

A planar read is then performed on each of three dimensions k, j and i, where k is the fastest dimension. The number of readers varies from 32 to 512. We compare the performance between three different data organizations, namely Logically Contiguous (LC), the original ADIOS (ORG) and our Optimized Data Organization (NEW). The peak performance is shown in Figure 2(b). Our Optimized Data Organization achieves a maximum improvement of 12x compared to ORG and 72x compared to LC, respectively. Moreover, the dimension dependency is significantly alleviated.

4. CONCLUSION

We propose a system-aware optimized data organization to enable the efficient scientific data analytics. Under the governance of Optimized Chunking model and SFC-based chunk reordering, we demonstrate that a balanced and high performance reading for challenging access patterns of scientific data is achieved.

5. 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-4000103043) 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.

6. ADDITIONAL AUTHORS

Additional authors: Ray Grout(National Renewable Energy Laboratories, email: `Ray.Grout@nrel.gov`) and Matt Wolf (Georgia Institute of Technology, email: `mwolf@cc.gatech.edu`).

7. REFERENCES

[1] Adaptable I/O System. http://www.nccs.gov/user-support/center-projects/adios.

[2] J. H. Chen et al. Terascale direct numerical simulations of turbulent combustion using S3D. *Comp. Sci. & Disc.*, 2(1):015001 (31pp), 2009.

[3] NCCS. http://www.nccs.gov/computing-resources/.

[4] T. Shimada, T. Tsuji, and K. Higuchi. A storage scheme for multidimensional data alleviating dimension dependency. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 662 –668, nov. 2008.

[5] Y. Tian, S. Klasky, H. Abbasi, J. Lofstead, N. P. R. Grout, Q. Liu, Y. Wang, and W. Yu. Edo: Improving read performance for scientific applications through elastic data organization. In *CLUSTER '11: Proceedings of the 2011 IEEE International Conference on Cluster Computing*, Washington, DC, USA, 2011. IEEE Computer Society.

[6] Y. Tian and W. Yu. Finding the optimized chunking for multidimensional array on large-scale systems. Technical Report AU-CSSE-PASL/12-TR01, Auburn University, 2012.

[7] W. Yu and J. Vetter. ParColl: Partitioned Collective I/O on the Cray XT. In *International Conference on Parallel Processing (ICPP'08)*, Portland, OR, 2008.

[8] W. Yu, J. Vetter, and H. Oral. Performance characterization and optimization of parallel I/O on the cray XT. *IPDPS*, pages 1–11, April 2008.