Improving the I/O Throughput for Data-Intensive Scientific Applications with Efficient Compression Mechanisms

Dongfang Zhao
Department of Computer Science
Illinois Institute of Technology
dzhao8@hawk.iit.edu

Jian Yin
Data Intensive Scientific Computing Group
Pacific Northwest National Labotory
jian.yin@pnnl.gov

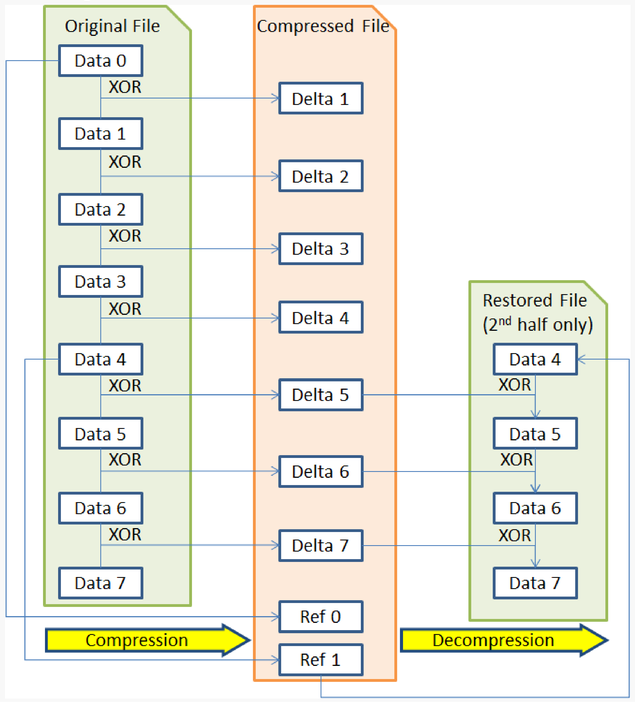
Ioan Raicu
Department of Computer Science
Illinois Institute of Technology
iraicu@cs.iit.edu

Objective

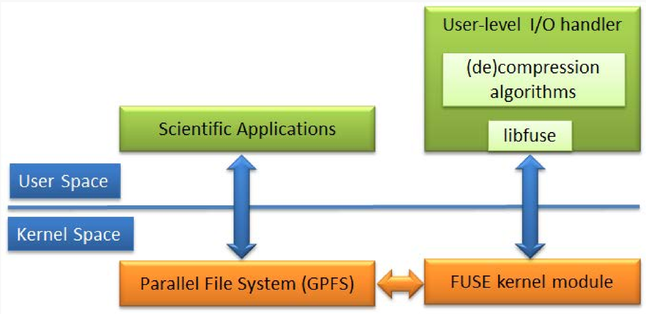
- Design more efficient mechanisms that use compression to reduce I/O overhead in large scale parallel scientific applications
- Requires no modifications to the applications

Method

- Suppose the original data has 8 entries
- The compressed file contains R = 2 reference points and 7 deltas
- The restored 2nd half of the file does not need to decompress the entire compressed-file



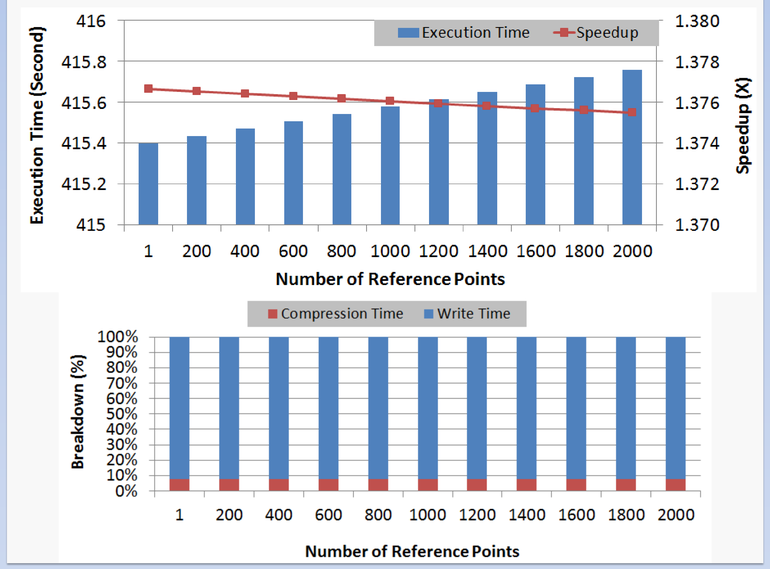
Implementation



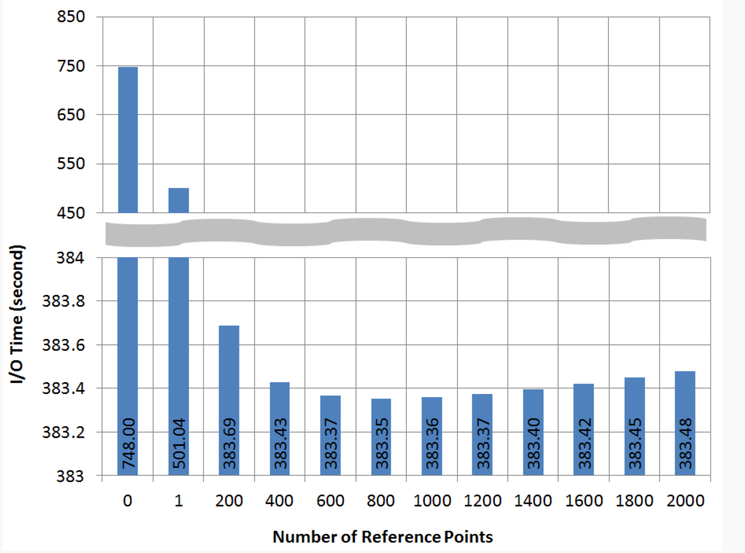
Experiment Setup

- IBM BlueGene/P Supercomputer
- 256 compute nodes / 1024 cores
- 128-storage-node GPFS filesystem
- 244GB Global Cloud-Resolving Model (GCRM) data

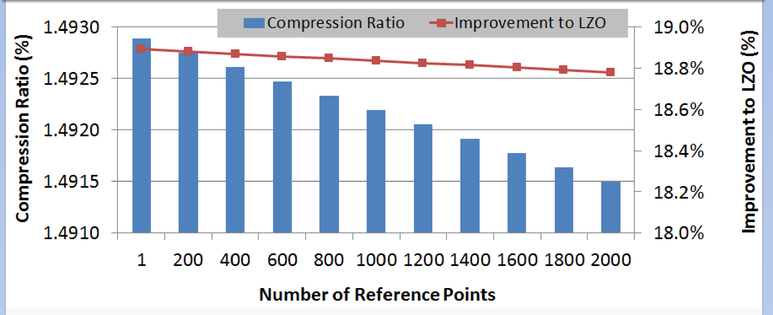
Write All the Data



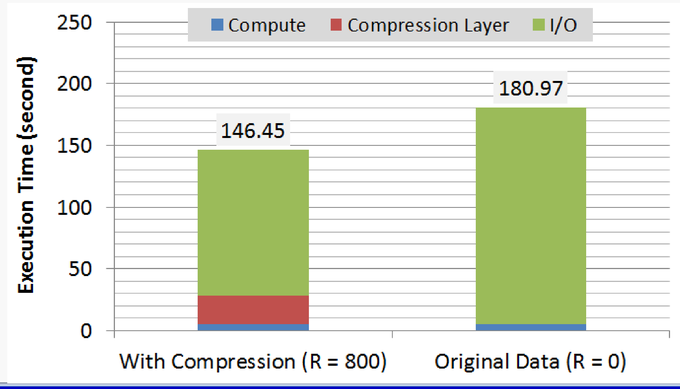
Read the Last Record



Compression Ratio



MMAT Application



Conclusion

- Our mechanism is shown to be highly effective in reducing the I/O overhead of large-scale parallel scientific applications
- The file system with the transparent compression layer delivers a high end-to-end I/O throughput

Future Work

- Explore more elastic mechanisms on determining dynamic reference points
- Leverage GPU to further reduce the computation overhead

Acknowledgement

This work was supported in part by the Office of Biological and Environmental Research, Office of Science, U.S. Department of Energy, under contract DE-ACO2-O6CH11357.