ADAPT-FS: Active Data Processing and Transformation File System

Arthur W. Wetzel, Jared Yanovich, Robert Budden, Markus Dittrich

Pittsburgh Supercomputing Center, Carnegie Mellon University, Pittsburgh, PA 15213

Introduction

A 2008 JASON Data Analysis Challenges study[1], undertaken for the DoD, projects that multiple areas of experimental science will have exponential growth of data requirements to hundreds of petabytes by 2020. The largest datasets are formed from large image collections that typically represent time and or multi-dimensional series that must be assembled and efficiently processed by multiple layers of computation. Each layer of computation typically receives input files that include the full extent of the dataset and produce new ouput files that are again as large as the original data. This leads to a combinatorial explosion in the number of files and the difficulties of data handling.

This poster outlines an adaptive virtual file concept and its use to simplify the automated assembly, analysis and annotation of petascale multidimensional data. It is being developed for use in high performance and distributed computing environments and will provide an approach to large data intensive computation and data sharing.

Motivation

One area of data intensive work at the PSC that will be a testbed application of our ADAPT-FS software is neural circuit reconstruction. This field, now known as connectomics, begins with large sets of serial section electron microscopy images. In our experience with datasets up to 100 TBytes per specimen the data storage challenge is just as large as the computational challenge. New high-throughput sample handling and imaging techniques will permit the capture of more than one petavoxel from specimens several mm in linear dimension. Many datasets with this combination of resolution and spatial extent will be required for complete reconstructions of insect, and larval fish brains or mammalian cortical circuits containing roughly 100,000 neurons from multiple brain regions and stages of development. New approaches to data handling are required to handle and process these data. A current zebrafish example is shown in Figure 1 below.

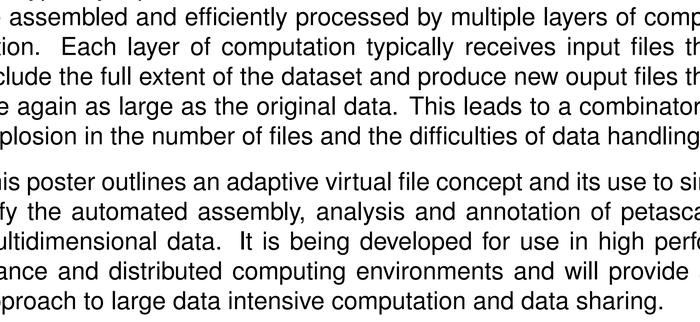

Figure 1: *A greatly reduced cut plane view through a zebrafish connectomics dataset. The full 3D data is 7 TBytes captured in 18,000 serial section electron microscopy images.*

Petascale connectomics datasets will consist of millions of raw EM images that must be accurately registered, normalized, and analyzed during the reconstruction and annotation of anatomical content. A number of interactive and web based tools are being used for the manual and semi-automated analysis of data up to the current Tvoxel level [2,3,4]. It is well known that complete manual annotation and proof reading in these large datasets is very time consuming and will be impossible at the petascale. To take full advantage of these scales the burden of processing has to shift from human to fully automated methods. Conventional processing involves a series of steps to progressively transform raw image data into forms required by 3D analysis tools. Large intermediate files and multiple renditions are often produced during these transformations. File based methods are greatly hampered by the difficulties and inefficiencies of handing and organizing these massive file sets.

We propose an alternative approach to bring many of these stepwise processes into a more practical form using a virtual filesystem model. This concept uses active ”on-the-fly” computation to replace explicit storage of intermediate and finished results. These results can be dynamically presented in multiple and malleable preprocessed formats through a virtual file interface. The faithful implementation of file like semantics will allow a seamless interface to existing file driven analysis programs and to new programs currently in development by the research community.

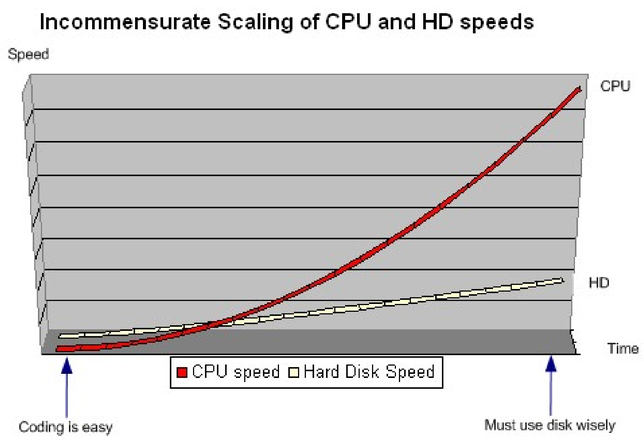

Figure 2: *The relative performance of computation and disk IO has been shifting rapidly. It is increasingly important to optimize data storage factors, particularly data organization and minimization of duplication, to maintain performance and manage storage costs.*

Our approach is driven by trends in the relative cost and performance of computing and storage technologies which show the increasing importance of optimizing IO and reducing explicit storage even at the expense of modest increase in computational load. Figure 2, illustrates the continuing exponential increase of CPU performance while disk performance, in terms of bandwidth and seek rates, has been much less. A similar graph of GPGPU computing performance would show an even sharper increase. The JASON report[1] notes that historical improvements in disk capacity and storage costs are much slower than experienced during the 1990’s and early 2000’s. This growing imbalance of computational and storage access speeds extends beyond disk storage and increasingly affects main memory as well.

Conventional processing compared with the ADAPT-FS

The data flow from capture through a preprocess that produces files is shown in Figure 3. In this conventional mechanism each variation of parameter settings to produce, for example, different geometrical and intensity transforms of raw image data, will produce new sets of intermediate preprocessed files that are approximately the same size as the original raw data. Additional applications that may perform skeleton tracing, segmentation or other analysis operations would then read the appropriate intermediate data files. In practice there are often multiple layers of processing rather than the single stage shown in the diagram. The performance of each processing stage is limited by the time to read its input files and write its output files. Therefore, each processing step is limited to 1/2 of the available IO bandwidth. On high performance striped filesystems this may be a few GB/sec. Unfortunately these filesystems do not improve random access latency or file open/close rates which become limiting factors unless data storage provides optimized localization of access patterns and properly sized records.

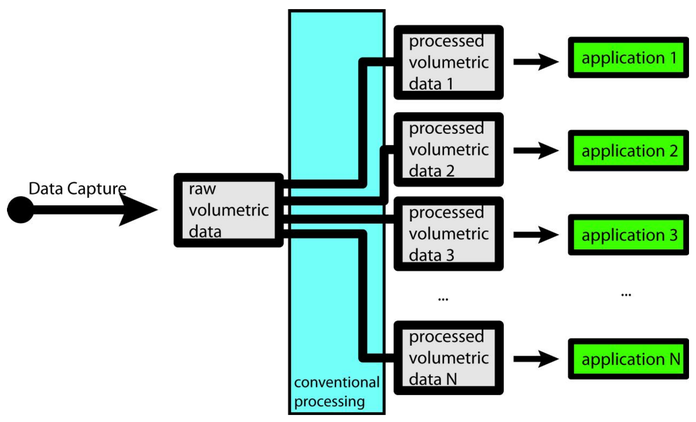

Figure 3: *Conventional file processing typically produces a new set of intermediate files from each variation of preprocess parameters.*

By contrast, the ADAPT-FS approach, shown in Figure 4, provides a framework to incorporate common preprocessing operations into demand driven data flow which operates similar to UNIX pipes. However, the ADAPT-FS interface provides a critical random access capability that is not available with unidirectional pipes. The ability to ”seek” within the virtual files in unanticipated nonlinear order greatly expands the range of applications that can be supported by the ADAPT-FS approach. The essential mechanism to implement the ADAPT-FS is the Filesystem in User Space (FUSE) [5] capability that is now provided on all of the major platforms (MS Windows, Mac and Linux/UNIX). Support for wide area access is provided by PSC’s SLASH2 code base [6]. Each data access by an application program triggers both the retrieval of the relevant raw data and the piecewise transformation of that data by processing within the ADAPT-FS.

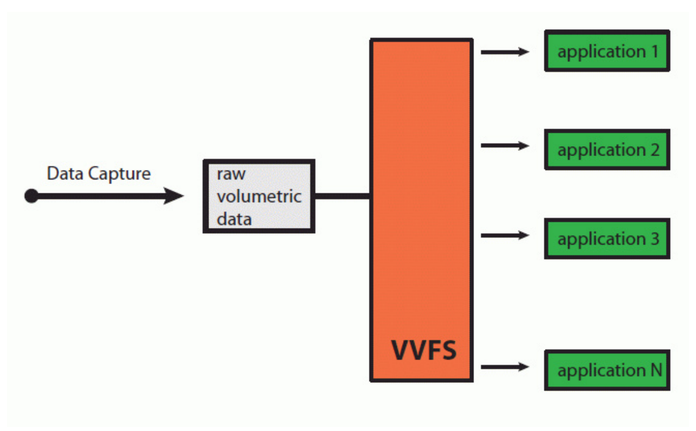

Figure 4: *The ADAPT-FS approach provides a Virtual Volume File System to eliminate explicit intermediate files while providing access by existing file based applications.*

ADAPT-FS will provide efficient processing of multi-dimensional data in high performance computing environments consisting of massively parallel CPU and GPGPU clusters. Therefore, we will implement an extended applications programming interface (API) that supports efficient multi-dimensional data access. The concept will build on frameworks that are familiar from existing filesystems which provide 1-dimensional linear file manipulations. The central element for this capability is a direct and efficient mapping from 3D space onto sets of files and regions within those conventional files as shown in figure 5.

The operations to access arbitrary regions of multi-dimensional data with ADAPT-FS are very similar to the operations required for random access within a normal file system except that retrieved data passes through a specified transformation process before being sent to the application program.

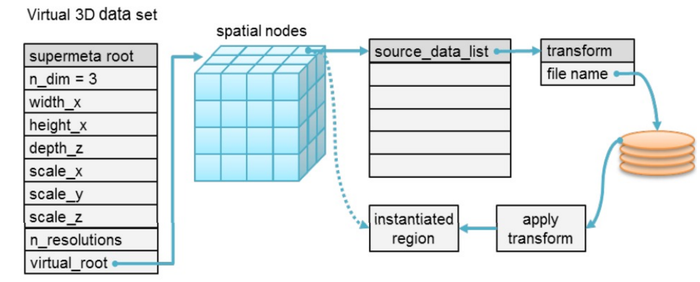

Figure 5: *A simplified data structure, analogous to a Unix/Linux filesystem, as a model for multi-dimensional ADAPT-FS storage organization and on-the-fly processing.*

ADAPT-FS as a CPU/GPU processing pipeline

The contents of the virtual files provided by the ADAPT-FS will be specified by transform descriptions and efficient processing codes to carry out those transformations. The ADAPT-FS will initially include a library of general purpose operations for geometric and other transformations that are similar for connectomics and other scientific applications. It will also provide the capability for users to supply their own codes to provide new application specific libraries that can be made available to other users in an extensible fashion. Examples would include segmentation and feature identification functions specific to EM and neural circuit reconstruction.

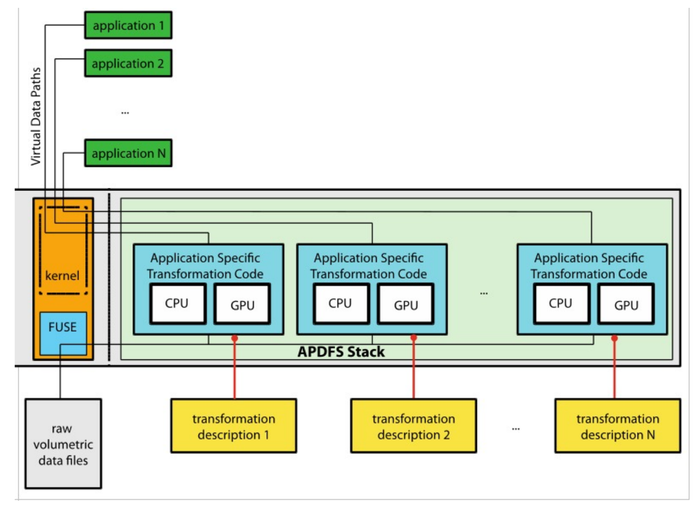

Figure 6: *This diagram shows the flexibility of the ADAPT-FS to operate with many applications at once each of which may require different virtual file transformations, provided via user written application specific transformation codes, while processing raw data that have been captured and stored in real disk files.*

By cascading multiple layers of ADAPT-FS computation complex image registration, normalization, object classification and segmentation operations can be driven from a single well structured version of raw data without file duplication or multiple stages of disk IO.

Summary

We are designing an advanced data processing and management framework to enable efficient, cost-effective, collaborative processing and sharing of scientific datasets at the largest scales with minimal data duplication. The need for such a framework is driven by the staggering pace at which dataset sizes in connectomics and other areas of science are growing. This growth is projected to continue at an accelerated pace, rendering traditional approaches for data transfer, analysis and sharing cumbersome in the near term and impossible in the medium to long term. New and more efficient approaches, as we propose, are required to meet computing challenges which will involve volumetric datasets of unprecedented scale.

Acknowledgements

We appreciate the use of large scale datasets from connectomics collaborators Jeff Lichtman and Florian Engert at Harvard, Davi Bock at Janelia Farm and Clay Reid at the Allen Brain Institute, and support from MMBioS (NIH P41GM103712). This work is funded by NSF SI2-SSE award 1440756.

References

[1] http://www.fas.org/irp/agency/dod/jason/data.pdf

[2] Won-Ki Jeong et. al., SSECRETT and NeuroTrace: Interactive Visualization and Analysis Tools for Large-Scale Neuroscience Datasets *IEEE Computer Graphics and Applications 30*, 2009

[3] http://www.ini.uzh.ch/ acardona/trakem2.html

[4] http://www.catmaid.org/

[5] http://fuse.sourceforge.net/

[6] http://slash2.psc.edu