Techniques for Image Alignment in the Neurosciences using XSEDE

David Lee
Center for Research in Biological Systems (CRBS)
University of California San Diego
9500 Gilman Dr. MC 0608
La Jolla, CA 92093-0608
dlee@ncmir.ucsd.edu

Mark Ellisman
Center for Research in Biological Systems (CRBS)
University of California San Diego
9500 Gilman Dr. MC 0608
La Jolla, CA 92093-0608
mark@ncmir.ucsd.edu

Shava Smallen
San Diego Supercomputer Center (SDSC)
University of California San Diego
9500 Gilman Dr. MC 0505
La Jolla, CA 92093-0505
ssmallen@sdsc.edu

Ilkay Altintas
University of California San Diego
9500 Gilman Dr. MC 0505
La Jolla, CA 92093-0505
altintas@sdsc.edu

ABSTRACT

Data in the sciences are experiencing unprecedented growth in the computational neurosciences where data acquisition rates are growing exponentially. The neurosciences benefit from both the exponential increases in CCD detector resolution and also from advances in tissue sample preparation and image montaging methods. These technologies combined, make possible multi-scale views, enabling the scientist to focus on sub-cellular components in context of entire collections of cells. Multi-scale views into data are key to next-generation scientific questions, but the dramatically increased data sizes (100-1000x) are straining the compute infrastructure and applications designed for much smaller data sizes.

Based on rapid contextual design methods, we analyzed the image processing protocol used by scientists and identified a key bottleneck in the image processing pipeline, specifically image alignment of instrument data in preparation for feature segmentation. Current data analysis protocols are divided across several applications and are commonly run during the span of a week. A typical neuroscience dataset of 32,000 x 24,000 x 1,300 pixel volume of 1.8 TB may take 61 hours to execute on traditional systems. To accelerate this process we applied HPC techniques using XSEDE resources to dramatically increase the throughput of processing and simplify execution. Specifically, we utilized the FLASH storage on the Gordon resource, enhanced legacy applications with a thread-pool wrapper, and encapsulated the data alignment protocol with Kepler workflows.

SDSC's Gordon machine that has a unique 4.4 TB Raid0 SSD scratch filesystem on a subset of nodes in its "bigflash" partition. Following execution profiling of the image alignment algorithms, we paired the application "newstack" with Gordon and its FLASH memory scratch filesystem. Execution of this key step on Gordon yields a 10x increase in throughput while other I/O and CPU bound applications increase throughput by 14x. Our estimates indicate execution on the FLASH storage on Gordon will transform the 61 hour runtime for a 32,000 x 24,000 x 1,300 pixel volume to just 11 hours. In addition to utilizing Gordon's FLASH memory, single threaded legacy applications are run in parallel using a thread-pool model. The thread-pool model allows us to more efficiently use multi-core resources with a near linear increase in throughput, particularly in CPU bound applications such as file format conversions.

Simplifying and automating the execution of previously independent procedures, the Kepler Scientific Workflow System (kepler-project.org) encapsulates the image alignment protocol used by scientists. Kepler workflows are portable and reusable information artifacts that can be shared with other lab members, encapsulate experience and expertise, and can be chained with other analysis protocols.

New methods in the neurosciences are enabling scientific experiments not previously possible. XSEDE resources will dramatically increase throughput and ease of execution, enabling additional experimental conditions, larger experiments, and higher throughput of scientific data analysis.

XSEDE 2013 2013 San Diego, California USA