

Narrative for Extreme Scale Biology Workshop: Gabrielle Allen

Extreme scale investigations in simulation-based biology face many of the challenges seen in other disciplines: coupling of multi-scale and multi-physics science, for example in integrating molecular dynamics with fluid flow; algorithms able to scale numerical methods using complex data structure to petascale and beyond compute resources; complexity in the software that needs to be continuously developed and extended particularly in regard to debugging, validation, profiling; handling of large data sets including archiving of simulation data, production and indexing of metadata; interactive visualization of multiple large data sets; integration with cyberinfrastructure services including workflow, computational steering and others.

My research is focused around the development of the Cactus Framework as an enabling technology for large scale scientific computing. Cactus is a general component framework designed for large scale parallel simulation code development for applications in science and engineering. The largest user base for Cactus is the numerical relativity community, where over 15 groups around the world use Cactus as the underlying framework for their codes to model black holes, neutron stars and other astrophysical objects. In this field, state of the art Cactus codes for black hole modeling include some 60 Cactus components (thorns) for computational infrastructure (coordinates, I/O, parallel driver, checkpointing, etc) and science (initial data, boundary conditions, evolvers, analysis). Parallelization is handled by a Cactus thorn (Carpet) that implements adaptive mesh refinement, with production simulations deployed on some 2000 cores of TeraGrid resources, and simulations taking some weeks to complete. An NSF project (XiRel) is working to improve the scaling of Carpet and Cactus for numerical relativity, where a mixed MPI-OpenMP approach currently scales to around 16,000 cores for the high order methods used in the black hole evolver. Other disciplines using Cactus with a need for large scale computing include coastal modeling, computational fluid dynamics, quantum gravity and petroleum engineering. Ongoing funded projects related to Cactus include developing application or framework level tools for debugging and profiling, interfaces for coupling multi-physics components, generation of scientific software from high level specifications of partial differential equations to reduce complexity and the introduction of errors.

Visualization and interaction are also of great importance to Cactus users. Cactus has a well defined steering interface that can be invoked, e.g. through a web interface, to dynamically steer parameters at run time. Previous work on visualization has integrated steering and data streaming into visualization applications such as OpenDX and Amira to provide interactive, real time visualization. A current project is developing a system for distributed visualization over high speed networks. As shown in Figure 1, the visualization pipeline can be distributed over remote resources to provide advanced capabilities over what can be achieved on a single system. Data servers distributed across compute clusters (or integrated into a running application) serve data across high speed optical networks to GPU clusters for parallel rendering, with resulting images streamed to large display devices. Integrated tangible devices control the system, which supports collaborative visualization. Such a system can not only provide improved response times and framerates com-

pared to a local system, but can scale to large data sets which cannot fit into the memory of a single system.

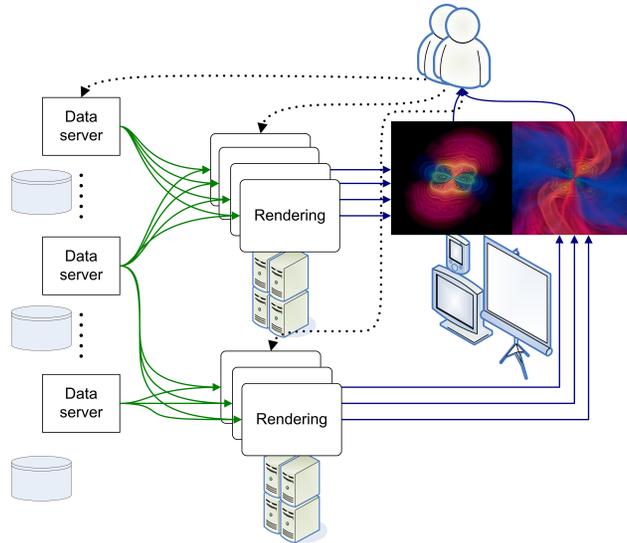


Figure 1: Distributed system for interactive, collaborative visualization.

Current efforts that address future challenges include:

- Development of application-level tools for debugging and profiling complex, multicomponent codes. The use of a framework allows us to provide general high level capabilities. This work includes using real time visualization (currently via ViSit) for debugging purposes. A web interface to Cactus is also being expanded to provide debugging capabilities for any simulation code using Cactus, for example providing the ability to step through Cactus methods and monitor the code.
- Provision and storage of metadata for the simulation to accompanying large scale data archiving. The metadata will then be used both for provenance and also for example to be able to automatically create high level reports on simulations. A domain specific language for relativity will be used to better describe science components for Cactus.
- Strategies for visualization of large data sets produced by simulations. Visualization needs to be interactive (steering), responsive (high frame rate, small delay) and collaborative.
- Integration of simulation codes with Web 2.0. We have investigating using e.g. Twitter and Flickr directly from a running simulation to announce monitoring information and display images, this has the benefit of reliable servers and flexible, easy to use interfaces, where the collaborative capabilities are all handled by the server and integrate well with user's usual environment.

A.R. Lawrence

High Resolution and Wide Field Electron Microscope Tomography

We discuss the software package “Transform Based Tracking Bundle Adjustment and Reconstruction” (TxBR) in the context of exascale biological research and beyond.

- **Electron Microscope Tomography (EMT) Serves a Wide Community of Researchers in the Biosciences**

Electron tomography is an emerging technology for the three dimensional imaging of cellular ultrastructure. In combination with other techniques this technology can provide three dimensional reconstructions of protein assemblies, correlate 3D structures with functional investigations at the light microscope level and provide structural information which extends the findings of genomics and molecular biology. Because of rapid advances in instrumentation and computer-based reconstruction, the routine imaging of molecular structure in context appears to be likely within the next few years.

- **Electron Microscope Tomography Provides an Essential Input to Physiological Modeling at Multiple Spatial Scales**

Physiological models are only as good as the structural models on which they are based. Dynamical models of molecular interactions, ion fluxes, transport and signaling depend upon the geometric details from the molecular and subcellular level of membranes, microfilament networks and various intra- and extra-cellular channels. Realistic physical details are essential for the petascale task of modeling over many spatial scales.

- **Wide-Field EMT Serves as a Critical Bridge Between Details at the Molecular Level and Structure Recoverable by Light Microscopy**

EM images span spatial scales ranging from a fraction of a nanometer to about 50 micrometers, and typical 3D reconstructions may cover $1/10^{15}$ of the volume of a typical optical microscope reconstruction. The limit of resolution of light microscopy is on the order of a hundred nanometers, and even though super resolution techniques applied in “pointilliste optical tomography” may give much better resolution, but rare events and contextual information are missed.

- **Continuing Ultrastructural Investigations Across the Range of Relevant Spatial Scales is an Exascale Computational Problem**

In order to bridge the gap, the spatial range of the electron microscope has been expanded by various techniques. Large sensor arrays and wide-field camera assemblies have increased the field dimensions by a factor of ten over the past decade, with a further 10X expansion possible and new techniques for serial tomography (z-axis) and montaging (x and y axes) make possible the assembly of tens of thousands of three dimensional reconstructions. Even though we are far from closing the spatial scale gap with a single experimental preparation, the amount of data generated by this enterprise threatens to overwhelm computational resources. A single data set for tomographic processing may have as many as 360 images, each 8K by 8K. Simple backprojection is order ≈ 3 so a data set of this size requires about 10^{13} or more coordinate evaluations. Because coordinate evaluations involve nonlinear functions in our case, the number of elementary operations is hundreds of times more, and bit error rates require double precision. The number of reconstruction volumes necessary to image an average cell down to the protein assembly level is of the order 10^4 , so counting exponents, we are into the exascale range. Data sets for this are well into the petascale range.

- **New Software for EMT Provides High-Quality Reconstructions to Support this Effort**

Tomographic reconstruction from large format electron microscope images requires special procedures to handle geometric distortions arising from electron optics as opposed to light-ray optics. In particular, electrons travel in curvilinear paths through the sample, and defocus and other aberrations can be more severe. We have developed a software package, TxBR, which is based on a generalization of the inverse problem associated with the ray transform. This software compensates for instrument optics, sample degradation, and other deleterious effects. In addition we have developed techniques to handle a wide range of data acquisition modes.

- **Further Development of Tomographic Software Requires Both Computational and Human Resources**

The present state of the art can exist only because of the scientific culture which has provided knowledge of mathematics and algorithms in addition to the scientific instrumentation and computer hardware which permit the experimenter to obtain the data and run the reconstructions. We appeal to a particular example which makes this point. Our tomography software runs in 4 days instead of 4 years on the particular data set mentioned above because a mathematician noticed that evaluation of high degree polynomials on a regular grid can be reduced to a simple recursion consisting of additions, and that the number of additions is linear (rather than exponential) in the degree of the polynomial. This, on a workstation. Furthermore, the recursion

can be made parallel to a very high degree, and the algorithm can be mapped to the simplified processors comprising, for example, a graphics processor unit. Programming this on a GPU board reduces the 4 days to less than 4 hours.

- **Progress in Electron Microscope Tomography is Dependent on Fundamental Research**

Tomography, or more properly from a mathematician's point of view, integral geometry is a core area in mathematics. One may find applications in many areas of mathematics, physics, and engineering. The mathematics of inverse transport, for example, is not far from inverse Radon transforms, and inverse transport is fundamental enough to provide some top ten problems. We get into problems of inverse transport when we consider beam-sample interactions and theorists are beginning investigate higher-order scattering phenomena in the context of novel modes of tomography.

- **Research in Electron Microscope Tomography will Stimulate Progress in Other Vital Areas of the Imaging Sciences**

We should also consider the entire workflow of EM tomography. Preparation of the data set requires tracking of features, and ultimate publication of the observations requires segmentation of structures in the reconstruction. Manpower limitations make the automation of these processes highly desirable, if not absolutely essential. Automated tracking and segmentation are aspects of automated pattern recognition which make a number of top ten lists in the imaging sciences. Containing the combinatorial explosions associated with pattern recognition may be accomplished by the application of high power parallel computers, but the research is just beginning in this area. It should go without saying that progress in this area will increase, rather than decrease the demand for computer resources. Recent work in tropical semirings also deserves a brief mention in this context. These objects appear in studies of the theory of computation, automata theory, graph theory, image analysis, neural circuitry, and surprisingly, have also become a fashionable area of research in algebraic geometry. Doubtless, there are yet to be discovered synergies between pure mathematics and high power computation in biology.

- **Building Research Communities Should be Given High Priority**

We should not neglect the potential contribution of the mathematical sciences and the mathematical community to the solution of some very hard problems associated with our research efforts. A little bit of math at the beginning can go a very long way, and we are not anywhere near to the point of diminishing returns. Some consideration should be given to the recruitment of mathematical talent, and the maintenance of communities with common interests

in the biological sciences. Virtual organizations and the internet may offer some possibilities, but this is the starting point of a much longer discussion.

For further information and references:

https://www.nbcv.net/pub/wiki/index.php?title=Tomography_Day_2008

The National Center for X-ray Tomography (NCXT) is a new DOE/NIH advanced imaging facility. The center is a resource for Soft X-ray Tomography, live cell light microscopy and super-resolution high aperture light microscopy. In addition to the new techniques discussed below the NCXT also performs a variety of closely related quantitative cell measurement and selection methods such as FACs and optical laser tweezers. The center includes sophisticated cell culture capabilities for a wide variety of organisms, and a rich spectrum of in-house whole cell analysis capabilities for tackling modern systems biology in a reproducible and meaningful way.

Soft X-ray tomography is a powerful new method for imaging whole cells and tissue samples. One of the most exciting research directions enabled by this method is the volumetric imaging of native cell structure at better than 50nm resolution. In addition, the measurement is highly quantitative and produces artifact free isotropic resolution tomographic reconstructions. The technique is high throughput, and can easily generate many hundreds of whole cell tomograms per day. Perhaps most important is the application of soft x-ray tomography to rapidly vitrified cells; in fact, simple freezing is the preferred method for preparing samples for any type of high-resolution imaging modality. This ensures that the observed cell structure is truly representative of the state of the specimen.

The other new technique invented and pioneered at the NCXT is high numerical aperture cryogenic light microscopy. In this method a specialized low temperature light microscope has been developed to view the same vitrified sample that is then viewed by x-ray tomography. These two methods can be intimately coupled and used to produce an information rich correlated data set, i.e a sub 50nm isotropic x-ray tomogram over-laid with a high-resolution multicolor quantitative fluorescence image. The new cryo-light microscope can be used to perform almost any light microscope technique, including super resolution fluorescence imaging. At liquid nitrogen temperatures the fluorescent signal is many times more stable and quantitative measurements can be performed.

Automated data collection robots are being developed to enable statistically significant throughput of specimens. The NCXT provides unique capabilities which are producing a completely new type of supremely high fidelity data for comparison with mathematical models of cellular function.

Due to the throughput and multimodal nature of the data, we have an urgent need for sophisticated computational development at the highest level. At the data collection level we need to develop online "real time" data alignment and reconstruction. We want microscope users to be able to view reconstructed tomograms within seconds of data collection. This requires invention of fast model based alignment and reconstruction algorithms. Analysis of the data generated by users of the NCXT is similarly in great need of sophisticated computational approaches. For example, analysis of a soft x-ray tomogram requires new algorithms for segmentation and shape analysis. The isotropic resolution and natural state character of the data means that for the first time it is sensible to develop software based tomographic analysis suitable for automatically processing thousands of data sets.

Possibly the highest level computational challenge generated by the capabilities of the NCXT is the modeling and analysis of the multimodal image data generated by the combination of selected live cell microscopy followed by high resolution cryo-fluorescence and then high resolution soft x-ray tomography. This challenge involves both more technique oriented problems such as automatic multimodal data set alignment, and problems of deep scientific significance such as how to understand and make useful models of cell behavior.

The new techniques developed at the NCXT are amongst the most precise and reproducible imaging methods available, but due to the complexity and throughput requirements necessary for meaningful measurements on biological systems it is clear that parallel advancements in computing must be made to fully capitalize on these new methods. The combination of new techniques developed at the NCXT and high performance computing will make a significant contribution to our ability to engineer organisms for global needs, such as fuel, food and medicine production.

T. Tasdizen

The National Academy of Engineering has selected *reverse engineering the brain* as one of their grand challenges with the motivation that part of the problem with state-of-the-art thinking machines is that they have been designed without much attention to real ones¹. Furthermore, the impact of figuring out how the brain works goes beyond building smarter computers, it has implications in understanding neuro-degenerative diseases and building neural implants such as artificial retinas to cure blindness. We believe that a crucial component of reverse engineering the brain is deciphering the precise circuit wiring maps of its neural systems.

Neural circuit reconstruction (NCR) is the mapping of all individual neurons and their synaptic contacts in a specimen to create its circuit map, also known as the *connectome* [1,2]. Models of neural circuits are essential to the study of the nervous system, but state-of-the-art models are largely not based on ground truth connectivity. NCR can provide this ground truth. Electron microscopy (EM) is an unique imaging modality for NCR because it has a resolution that is high enough to identify synaptic contacts and gap junctions. However, deciphering the wiring maps of neural systems is a long-standing problem in neuroscience that has been hindered due to the impracticalities in acquisition and analysis of large scale EM images. Biologists have been limited to studying the connectivities of a very small number of neurons with the only exception being the decade-long effort to reconstruct the circuit map of the *C. elegans* which still has just over 300 neurons. The last five years have seen a revitalization of this field with the invent of automated EM acquisition strategies and the involvement of researchers from the computational sciences. For instance, with automated image acquisition, we can now capture approximately 4000 transmission electron microscopy (TEM) images in 24 hrs. Neuroscientists are now interested in the circuit wiring maps for larger neural systems such as the fly brain and the mammalian retina [3]. The bottleneck for generating the next round of scientific results in this field is the analysis, not the acquisition, of EM images.

There are two major algorithmic barriers to large-scale reconstruction of neural circuitry from serial-section TEM: volume assembly and process tracking/synapse detection. In [3] we demonstrate completely automatic and robust approaches for mosaicking of TEM sections from thousands of tiles and for three-dimensional volume assembly by section-to-section registration. Figure 1 shows a mosaic of a single-section from transgenic rabbit retina, approximately 60 Gigabytes, acquired and assembled with these tools. Since tracking neural processes in 3D, even through small distances, requires hundreds of such 2D sections, NCR projects have significant data storage requirements.

Automated tracking of neuronal process and synapse detection presents an even harder algorithmic and computational challenge. Solutions based on supervised machine learning have been demonstrated to be successful in different types of EM images for the cell membrane vs. non-membrane pixel classification task [4,5]. Once membranes in a section are detected, individual cells can be segmented and tracked across the volume [6]. Jain et al. propose to use convolutional artificial neural networks on serial block-face scanning EM images [4] whereas Jurrus et al. develop a series artificial neural network strategy for serial-section TEM [5]. A significant shortcoming of both approaches is the computational cost involved in the training of classifiers on the billions of pixels that compose a typical EM training dataset. Hence, researchers are forced to work on images that are down-sampled by factors as high as 16^2 or to use small portions of the available training image. In our experiments, we have found that even at a down-sampling factor of 16^2 , training the classifier on a section such as the one shown in Figure 1 takes approximately 40 hrs using 5 processors. These practical solutions compromise the accuracy of the algorithms. Furthermore, certain EM modalities such as TEM contain rich textural detail which can be used to go beyond the cell membrane detection task and to identify different types of cells and synapses which is a crucial component of NCR. The classifier architectures proposed in [4,5] can be extended to this multiple label case, a problem known as *scene parsing* in computer vision; however, this extension is not practical as it further increases the computational cost of the algorithms. Fortunately, the

¹ <http://www.engineeringchallenges.org/cms/8996/9109.aspx>

backpropagation algorithm used in training these classifiers can be parallelized using batch learning [7]. Therefore, the algorithms can easily be extended to take advantage of better computational resources.

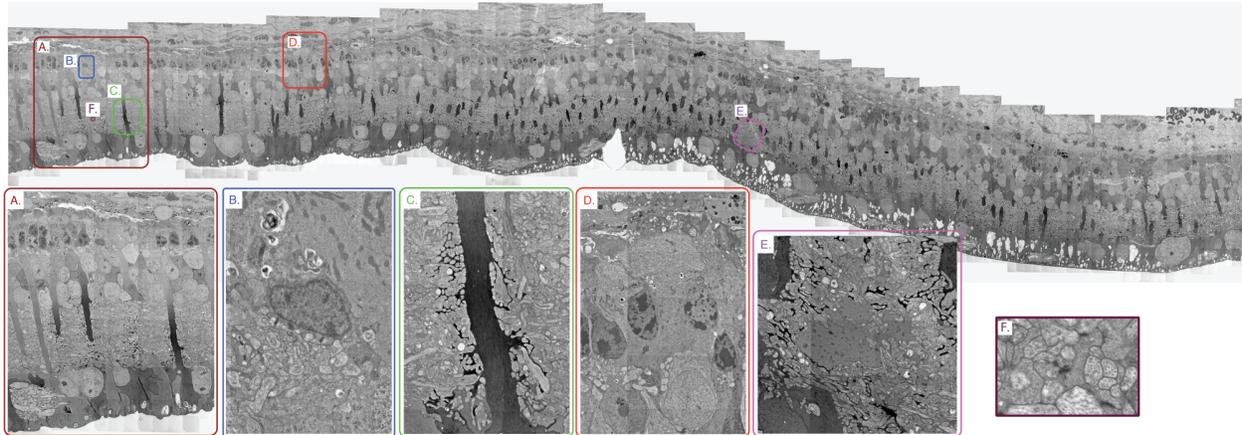


Figure 1: Image of a transgenic rabbit retina is a mosaic of over 2200 separate TEM assembled in a completely automated fashion through solutions derived from the previous funding phase of this project. Each tile is an image with approximately 4000 x 4000 pixels. Insets show several areas at varying levels of zoom to demonstrate the amount of information available in the mosaic.

There are currently efforts such as the Blue Brain Project² which attempt to reverse engineer the mammalian brain by simulating highly complex models of a neocortical column involving tens of thousands of neurons. The Blue Brain Project team estimates that real-time simulation of a single neocortical column with 10,000 neurons and approximately 10^8 synapses will require 10,000 processors; in other words, one processor per neuron. While the model used in the Blue Brain Project is based on anatomical, genetic and electrical data, the amount of neural connectivity data available to constrain the model is very limited. Therefore, a solution to the NCR problem will also aid such efforts by providing ground truth for connectivity and allowing more realistic models to be built.

References

- [1] O. Sporns, G. Tononi, and R. Kotter, "The human connectome: A structural description of the human brain," *PLoS Comput. Biol.*, vol. 1, pp. e42, Sep 2005.
- [2] K. L. Briggman and W. Denk, "Towards neural circuit reconstruction with volume electron microscopy techniques," *Current Opinion in Neurobiology*, vol. 16, no. 5, pp. 562–570, October 2006.
- [3] J.R. Anderson, B.W. Jones, J.-H. Yang, M.V. Shaw, C.B. Watt, P. Koshevoy, J. Spaltenstein, E. Jurrus, Kannan U.V., R.T. Whitaker, D. Mastronarde, T. Tasdizen, and R.E. Marc, "A computational framework for ultrastructural mapping of neural circuitry," *PLoS Biology*, vol. 7, no. 3, pp. e74, 2009.
- [4] V. Jain, J.F. Murray, F. Roth, S. Turaga, V. Zhigulin, K.L. Briggman, M.N. Helmstaedter, W. Denk and H.S. Seung, Supervised learning of image restoration with convolutional networks. *IEEE Int Conf Computer Vision*, pp 1–8, 2007.
- [5] E. Jurrus, A. R. Paiva, S. Watanabe, R. Whitaker, E. M. Jorgensen, and T. Tasdizen, "Serial neural network classifier for membrane detection using a filter bank," Tech. Rep. UUSCI-2009-006, University of Utah, 2009 (submitted to *Medical Image Analysis*)
- [6] E. Jurrus, R.T. Whitaker, B. W. Jones, R. E. Marc and T. Tasdizen, An optimal-path approach for neural circuit reconstruction. *Proc. IEEE Int Symposium Biomedical Imaging*, pp. 1609-1612, 2008.
- [7] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

² <http://bluebrain.epfl.ch/page17871.html>