**Some Cyberinfrastructure Undergraduate Projects in Community Grids laboratory**

**Visualization and Analysis of Chemical Compound and Biology Data**

Science discovery is fascinating. The broad availability of scientific data compels us to rethink new approaches to store, retrieve, process, and analyze this abundance information. Our group studies computer system architecture and novel software technologies. However we stress study of applications so we ensure our work is relevant. In biomedical informatics, we are looking in four areas:

a) **EST** (Expressed Sequence Tag) sequence assembly program using DNA sequence assembly program software CAP3. This uses MapReduce technologies.

b) **Pairwise Alu sequence alignment** using Smith Waterman dissimilarity computations followed by MPI applications for Clustering and MDS (Multi Dimensional Scaling)

c) **Correlating Childhood obesity with environmental factors** by combining medical records with Geographical Information data with over 100 attributes and performing correlation computation, MDS and genetic algorithms for choosing optimal environmental factors. We also integrate the statistics package R into analysis system.

d) **Mapping over 20 million entries in PubChem into two or three dimensions** to aid selection of related chemicals for drug discovery with convenient Google Earth like Browser. This uses either hierarchical MDS (which cannot be applied directly as it is too time consuming) or GTM (Generative Topographic Mapping).

Each of these areas is research in collaboration with application scientists at IUB or IUPUI. There are opportunities for students in all areas. See <http://grids.ucs.indiana.edu/ptliupages/publications/CetraroWriteupJune11-09.pdf>, or <http://grids.ucs.indiana.edu/ptliupages/publications/cloud_handbook_final-with-diagrams.pdf>.

**New paradigm of computing - MapReduce and Cloud Technologies**

Today’s supercomputers can handle tens of trillions of computations per second. Imagine one can remotely run applications accessing the power of supercomputers from a desktop machine. Cloud computing provides the type of “on-demand” services of computation and storage infrastructure. There have been several important commercial developments of computing technologies that have important implications for scientific computing. Cloud computing is best known for the systems like Amazon EC2, Eucalyptus and Azure which use virtual machines to provide flexible, dynamic, easy to use computing on demand. Another important development is MapReduce systems that were developed to support the huge information retrieval industry. This is perhaps the largest data analysis problem and so it is particularly interesting to examine for scientific data processing which is of growing importance as the data deluge continues. We have opportunities in both virtual machine and MapReduce areas. These technologies are applied to applications in several areas including Bioinformatics, Particle Physics and Polar Science. See <http://grids.ucs.indiana.edu/ptliupages/publications/MTAGS09-23.pdf>.

**Datamining and Visualization of Web Data**

The amount of data in the Internet powered by people’s intelligence, especially including rating and social bookmarking—is steadily growing. The analysis of such data can lead us to discover hidden knowledge but the huge size of web data remains challenging in many machine learning algorithms. We have explored the possible algorithms for data analysis and its computational efficiencies by using multicore and cluster technologies. Especially, we demonstrate the analysis of social bookmarking data and Netflix movie rating data by using the parallelized deterministic annealing clustering algorithm. We have also developed new robust algorithms (Deterministically Annealed Generative Topographic Mapping) for mapping high dimensional data to lower dimensions. See <http://grids.ucs.indiana.edu/ptliupages/publications/presentations/CCT.pdf>.

**FutureGrid**

Learning to use “distributed super computer” may be challenging. But the opportunity is great. FutureGrid (<http://uitspress.iu.edu/news/page/normal/11841.html>) is a major new project led by Indiana University to develop a distributed testbed where new approaches to scientific computing can be developed that exploit clouds, grids and parallel computing with large numbers of distributed multicore nodes. Cloud technologies—such as Amazon Web Services and the open-source Eucalyptus system—are increasingly used to support online resources used by researchers and the public, and have the potential to make a significant impact on the 21st century economy. The US federal government is also exploring the use of cloud technologies to better serve the public—including the proposed development of a federal computing cloud—and government officials are working with industry partners to establish standards for cloud computing. Partners in the FutureGrid project include: Purdue University, San Diego Supercomputer Center at University of California San Diego, University of Chicago/Argonne National Labs, University of Florida, University of Southern California Information Sciences Institute, University of Tennessee Knoxville, University of Texas at Austin/Texas Advanced Computing Center, University of Virginia, and the Center for Information Services and GWT-TUD from Technische Universtität Dresden. There are opportunities for work in both the computer science of new software models for distributed and parallel computing and for developing new applications using the technologies of the future.