Exemplar Tiger Team Projects

Jason Lawrence: Image Super-Resolution by Example

Dr. Jason Lawrence is an Assistant Professor in the Department of Computer Science. He is a lead researcher on an NSF-funded project that is investigating a radically different approach to image processing tasks. While many important image processing tasks involve solving some type of inverse problem (e.g., eliminating noise, enhancing a black-and-white image with plausible colors, increasing the resolution of an image), Dr. Lawrence’s team poses the question of whether an image can be improved by evaluating trillions of image patches constructed from millions of on-line images and using the set of relevant patches to improve the quality of the original.

The team is working with approximately 10 million high-resolution digital images scraped from the popular photo sharing website Flickr, which it is using to develop a system that can recover plausible details (e.g., a detailed bark texture consistent with the pixilated input and that a human would accept as real). The key barrier in deploying this type of example-based image enhancement system is constructing methods for efficiently searching this large amount of image data to identify similar patches. Fortunately, these types of nearest neighbors searches are known to be embarrassingly parallel and so the task of searching a collection of images can be easily distributed over multiple machines.

Dr. Lawrence is a shining Tiger Team success story: before bringing his project and needs to our attention, he was struggling to find resources adequate for his research. UVACSE helped the group identify computing resources available through the TeraGrid infrastructure that are appropriate for a project of this scope. Immediately, UVACSE helped the team successfully apply for a startup allocation on the Ranger computer at the Texas Advanced Computing Center (TACC). The PI became so enthusiastic about this resource (200,000 cpu hours, or 25 CPU years) that he taught himself MPI in one weekend in order to allow his code to run on Ranger.

Dr. Lawrence quickly exceeded his initial allocation. Our Campus Champions have been working with him to get him more CPU hours quickly so he can continue his research. UVACSE supplied expertise in technology that allows Dr. Lawrence to achieve his research goals.
According to Dr. Lawrence:
UVACSE has been instrumental in helping my group identify computing resources available through the TeraGrid infrastructure that are appropriate for a project of our scope. In particular, with the help of UVACSE staff we were successfully granted a startup allocation on the Ranger computer at the Texas Advanced Computing Center (TACC). This resource has proved to be an enabling technology that has carried our work from the proverbial drawing board to reality. At the current date, we have successfully performed nearest neighbors queries over $10^{12}$ training patches in ~80 minutes using 4,096 processing cores. The same feat would require roughly ~228 days of processing on a single computer!

Simonetta Liuti: Self-Organizing Maps for Quarks and Gluons at the LHC

Dr. Liuti is an Associate Professor in the Department of Physics. Dr. Liuti is researching the extension of computational methodologies based on neural networks to theoretical analyses of nuclear and particle physics. She has requested help from UVACSE in parallelizing her code. We are currently nearing the completion of a new version of her code that should better suit her needs.

Partons are particles consisting of quarks and gluons (primarily the proton and neutron) and their distribution functions are not particularly well known, especially in certain energy ranges. The goal of the project is to determine better approximations of these distribution functions through the application of self-organizing maps. From this, we can estimate the error of the functions. In addition, we have applied a genetic-algorithm technique to minimize the deviation from experimental data, so that the final distribution functions should be an improved estimate of the behavior.
UVACSE staff has created a code that is able to mix a variety of theoretical constructions of the parton distribution functions in order to reduce the bias toward a particular function. The code development that was performed enabled the PI to make a number of changes that she wants to do on her own. This project also involves a Masters student and led to his thesis and one or more publications.

UVACSE has been and continues to be a resource for expertise and collaboration in Dr. Liuti’s research.

According to Dr. Liuti:

The tiger team program has been vital for launching my research project in the field of computational particle and nuclear physics. In my case, in particular, some preliminary work had already been accomplished previously and it had provided a successful initial test. It had however come to a stall due to the difficulties in interpreting and connecting the different parts of the code. This was initially written, for reasons given in detail below, mostly in Python, a language most physicists are not familiar with. In addition the code needed to be optimized and parallelized. Thanks to the tiger teams we have now a completely reorganized code using Fortran 95 that is flexible and well documented. Our interaction could not have been more successful.

The reason why we contacted the tiger teams concerned the rewriting of our code used for our initial work in a more coherent and unifying way.

... our research progress using advanced computational techniques were highly facilitated by the tiger teams. The interaction with Dr. Katherine Holcomb was ideal especially because she could suggest a number of useful improvements in the coding, besides undertaking the major task of translating it into Fortran 95 and making it run in parallel. Both her high expertise and her unique ability to follow research issues in particle physics have played an important role. This very important contribution of hers was completed in a timely manner.

I lend my highest support to the tiger teams at the University of Virginia. Our case is indeed a success story. All of the goals in assisting my research project by addressing optimization, parallelization, workflow, and data management issues were accomplished ideally.

Douglas Taylor: Annotation of High-Throughput Transcriptome Data

Dr. Taylor is a Professor and Chair of the Department of Biology. The PI has recently generated a large quantity (~400 million base pairs) of DNA sequence data, representing the expressed genes of a single plant species (*Silene vulgaris*). Assembly of these data produces tens of thousands of gene sequences, and he would like to develop/implement an approach to systematically annotate these genes based on their similarity to other known genes. Such an approach will depend on performing BLAST searches against large databases of previous sequences and analyzed genes and parsing/storing the resulting output.

UVACSE staff met with the PI and other participants and discussed breaking up the problem into two major phases: 1) doing BLAST comparisons between their database and reference databases and 2) gathering annotation information from NCBI for interesting hits in comparisons.
UVACSE staff designed a directory structure to organize the data for the project: the *Silene vulgaris* DNA sequence database, the community reference genome databases, results for the BLAST comparisons and annotations, and status information for progress for each chunk of work.

Since the meeting, the PI’s student Dan Sloan provided: 1) more detailed information about the likely annotation scripts that will be used; 2) the reference databases to be compared against; 3) a description of how each reference database will be processed (the process is slightly different for each reference database); and 4) Three sample chunks of the *Silene vulgaris* DNA sequence database of various sizes (10, 100, 1000 sequences).

Using this information, UVACSE staff downloaded the relevant reference databases from links provided by Dan, formatted them appropriately where necessary, and produced the necessary job submission files to execute the BLAST runs appropriately for each reference database. UVACSE also created a test directory populated with the real reference databases, job submission files, and the sample chunks of the *Silene vulgaris* DNA sequence database. UVACSE also ran a test script that ran the full BLAST and annotation pipeline for all 3 chunks of the *Silene vulgaris* data against all of the reference databases.

The test ran successfully using the grid as the mechanism to run the BLAST comparisons, while annotations were done on a single machine (the machine where the driver script was executed). The test helped to debug the driver script and provided statistics for the runtime behavior of the various processing steps.

UVACSE provided the support that has made Dr. Taylor’s research a reality. In turn, this research has become a resource for other scholars.

According to Dr. Taylor:

*The Tiger team played an essential role in our analysis of plant genomes and gene expression. One of the most exciting developments in the field of biology in recent decades has been the rapid advance of DNA sequencing technology, but these advances have come with increased computational requirements that often cannot be met with local computing resources. John Karpovich and the Tiger Team successfully implemented a strategy to deploy our DNA sequence analysis pipeline on the university’s distributed computing resources (Cross Campus Grid). They were valuable partners in this project, taking the time to meet with us and develop the tools for our custom needs. Simply put, we could not have completed this analysis without them.*

*The dataset generated by this project is now available in an online database ([http://faculty.virginia.edu/silenetranscriptome/](http://faculty.virginia.edu/silenetranscriptome/)) that is being accessed by labs all over the world. We are also planning on publishing a paper that describes this genomic resource. The dataset also represents an important tool being utilized in an ongoing project that was recently funded by the NSF (MCB-1022128; $638,952) to study the interaction between multiple genomes within cells. Finally, the genomic resources generated by this project will constitute a component of the dissertation research of at least 2 Ph.D. students in my lab (Dan Sloan and Andrea Berardi).*
Federico Ciliberto: Inference on Market Power in Markets with Multiple Equilibria

Dr. Ciliberto is an Assistant Professor in the Department of Economics. He is investigating how the measurement of market power is biased when the endogeneity of market structure (i.e., which firms serve the market) is not controlled for.

Dr. Ciliberto wanted to parallelize an integration over a sample distribution by having different sections of the integral calculated in parallel. This is trivial using Matlab’s parallel “for” loop. UVACSE staff has parallelized Federico’s code and have set up his account so the PI can start doing production runs of his code on the Linux clusters. Our testing indicates that using 16 processors optimizes the tradeoff between communication and computation for his code.

UVACSE’s Ed Hall worked with Dr. Ciliberto to parallelize his Matlab code so he could run sufficient simulations for a conference paper. Mr. Hall’s expertise and support were essential to the project’s progress and success.

According to Dr. Ciliberto:

*Clearly, the second stage estimator is computationally very intensive since we need to compute the distribution of the residuals and of the unobservables for many parameter values and at many points of the distribution. This would not be possible without parallelizing the computation. To this end, Ed’s contribution has been fundamentally crucial for this project. I can state here that without Ed’s contribution this paper would have never been completed.*