Cluster Research use cases

To understand how the cluster supports research at Tufts, the following user comments show a wide range of applications. If you wish to contribute a short description of your cluster usage, please contact durwood.marshall@tufts.edu or lionel.zupan@tufts.edu.

Giovanni Widmer

At Tufts Veterinary School of Medicine we are using Illumina technology to sequence PCR amplicons obtained from the bacterial 16S rRNA gene. The analysis of millions of short sequences obtained with this method enables us to assess the taxonomic composition of bacterial populations and the impact of experimental interventions. Some of these analyses are computer-intensive and running them on the cluster saves time. Typically, we use Clustal Omega to align sequences. On the cluster, a samples of a few thousand sequence reads can be aligned in a few minutes. We have also installed mothur on the cluster (mothur.org) and are running sequence analysis programs from this collection. These programs are used to de-noise sequence data and to compute pairwise genetic distance matrices. We visualize the genetic diversity of microbial populations using Principal Coordinate Analysis, which is also computer-intensive. We have adapted this approach to analyze populations of the eukaryotic pathogen Cryptosporidium. Several Cryptosporidium species infect the gastro-intestinal tract of human and animals. Using a similar approach as applied to the analysis of bacterial populations, we assess the diversity of Cryptosporidium parasites infecting a host and monitor the impact of various interventions on the genetic diversity of this parasites.

Daniel Lobo

I am a postdoc in the Biology department and work together with Prof. Michael Levin to create novel artificial intelligence methods for the automated discovery of models of development and regeneration. A major challenge in developmental and regenerative biology is the identification of models that specify the steps sufficient for creating specific complex patterns and shapes. Despite the great number of manipulative and molecular experiments described in the literature, no comprehensive, constructive model exists that explains the remarkable ability of many organisms to restore anatomical polarity and organ morphology after amputation. It is now clear that computational tools must be developed to mine this ever-increasing set of functional data to help derive predictive, mechanistic models that can explain regulation of shape and pattern. We use the Tufts Computer Cluster for running our heuristic searches for the discovery of comprehensive models that can explain the great number of poorly-understood regenerative experiments. Our method requires the simulation of millions of tissue-level experiments, comprising the behavior of thousands of cells and their secreted signaling molecules diffusing according to intensive differential equations. Using the compute cluster, we can massively parallelize the simulation of these experiments and the search for models of regeneration. Indeed, the cluster is an indispensable tool for us to apply cutting-edge artificial intelligence to biological science.

Eric Kernfeld

My work with Prof. Shuchin Aeron (ECE) and Prof. Misha Kilmer (Math) centers around algebraic analysis of image and video data. Even a single video or a small collection of images can require an uncomfortable amount of time to process on a laptop, especially when it comes time to compare algorithms under different regimes. The high-performance computing tools(such as Matlab Distributed Computing Toolbox) at Tufts allowed us to run tests in a manageable time frame and proceed with our projects.

Christopher Burke

Prof. Tim Atherton's group in the physics department makes heavy use of the research cluster. Our focus is soft condensed matter, i.e. complicated solids and fluids such as emulsions, colloids, and liquid crystals. Simulations allow us to understand the behavior of these complex systems, which are often difficult to study analytically. Graduate student Chris Burke is studying how particles can be packed onto curved surfaces. This is in order to understand, for example, how micron-sized polymer beads would arrange themselves on the surface of an oil droplet. He uses the cluster to run large numbers of packing simulations and to analyze the large data sets that result. Post-doc Badel Mbanga and undergraduate Kate Voorhes study the behavior of coalescing droplets coated with liquid crystals. In particular, they are interested in the behavior of defects in the liquid crystal layer as coalescence occurs. They run computationally expensive simulations which would be impractical without the computing power available on the cluster.

Albert Tai

I am the manager and primary Bioinformatican of the TUCF Genomics Core, overseeing the operation of three deep sequencing instruments (Illumina HiSeq 2500, MiSeq and Roche 454 Titanium FLX), and their associated services. As part of these services, I provide primary and secondary data analysis services, and or training associated with these analysis. Deep sequencing generates a large amount of data per run and data analysis requires a significant amount of computing resources, both processing and analytical storage. The high performance research cluster and its associated storage is an essential tool for myself and the users of core facility. The parallel computing capability allow us to analyze large data sets in a timely manner. It also expedites troubleshooting processes, which sometime require us to test multiple analytical parameters on a single data set. As the amount of data generated in biological research increases, high performance computing resources has become an essential resource. I would certainly hope to see the expansion of this crucial computing resource.

Marco Sammon

I recently finished my undergraduate degree in Quantitative Economics, and I am continuing work on my Senior Honors Thesis with Professor Marcelo Bianconi. Two parts of our research in mathematical finance require intense computing power: solving systems of Black-Scholes

equations for implied volatility/implied risk-free rates, and fitting a SUR regression to explain factors that influence the difference between market prices and Black-Scholes prices. Before using the cluster, it took us weeks to process just a few days worth of options data. Now, we are able to work on many days of options data simultaneously, greatly expediting the process. This is important, as it allows us to aggregate a larger time series of data, which allows for much richer analysis.

Hongtao Yu

I am a postdoc in the chemistry department working in Prof. Yu Shan Lin's group. My research involves extensive Molecular Dynamics (MD) simulation of peptides and proteins. We use the MD method to study the folding thermodynamics and kinetics of glycoproteins, stapled peptides and cyclic peptides. The free energy landscape of protein and peptide folding is believed to be rugged. It contains many free energy barriers that are much larger than thermal energies, and the protein might get trapped in many local free energy minima at room temperature. This trapping limits the capacity of effectively sampling protein configuration space. In my research, we use various techniques to overcome the free energy barriers and improve the sampling, for example by using, the Replica-Exchange Molecular Dynamics (REMD) method and the Umbrella Sampling (US) method. In a typical US simulation, the reaction coordinate(s) is broken into small windows, and independent runs have to be done for each window. For example, 36 independent runs have to be performed if we choose a dihedral as the reaction coordinate and use 10 degree window. In a 2D US simulation, the number of independent runs increases to 36x36. Our system usually contains 1 protein molecule and thousands of water molecules; an independent run usually takes about 2.5 hours with 8 CPUs. This means that we have to run 135 days to finish one 2D US simulation on a single 8-core machine! With the large amounts of CPUs provided by the Tufts cluster, we can finish one 2D US simulation within 2 days! The benefit provided by the speed up is that we have the chance to explore more systems and methods.

Rebecca Batorsky

As part of my PhD research in the physics department, I studied various aspects of intra-host virus evolution. I used the cluster in order to run large simulations of evolving virus populations. Our simulations typically ran in Matlab, and we were able to run more than 30 simulations in parallel using multiple compute nodes. This enables faster collection of simulation data and allowed us to study large population sizes than would otherwise have been possible. Furthermore, the ability to access the my files on the cluster and programs like Matlab and Mathematica from any computer was extremely useful.

Scott MacLachlan

My group's research focuses on the development of mathematical and computational tools to enable large-scale computational simulations. We work on a diverse group of problems, including geophysical fluid dynamics, heterogeneous solid mechanics, and particle transport. The Tufts High-Performance Computing Research Cluster supports these activities in many ways. First, by providing significant parallel computing resources, it enables our development of mathematical algorithms and computational codes for challenging problems. For example, we have used the cluster to study parallel scalability of simulation algorithms for the deformation of heterogeneous concretes under load, with higher-resolution models than would otherwise have been possible. Furthermore, by providing access to cutting-edge computing resources, such as the new GPU nodes, we are able to participate in the computing revolution that is currently underway, re-examining the high-performance algorithms that have become the workhorses of the MPI-based parallel paradigm, and developing new scalable techniques that are tuned for these architectures.

Lakshmanan Iyer and Ron Lechan

In collaboration with Dr. Ron Lechan the Chief of Endocrinology at the Tufts Medcal Center, we are applying cutting edge next generation sequencing technology to determine the gene expression profile of Tanycytes,a special cell of glial origin in the brain. While much is known about the anatomy of these cells, their physiologic functions remain speculative and enigmatic. The results of these studies would provide clues towards their function. This analysis requires considerable amount of disk storage and CPU time. Without the Tufts high performance cluster and the associated storage it would be impossible to make sense of this data.

Krzysztof Sliwa, Austin Napier, Anthony Mann and others

Tufts' participants from the Department of Physics use the cluster for analysis of ATLAS data. Compute jobs run continuously in support of this effort. For additional information see **ATLAS** and more recently **Higgs news**.

Joshua Ainsley

Our work at the Laboratory of Leon Reijmers, PhD, Tufts University Neuroscience Department focuses on changes in gene expression that occurs in neurons during learning and memory formation. To examine these events on a genome-wide scale, we use a technique called next generation sequencing which generates millions of "reads" of short nucleotide sequences. By sequencing the RNA that is present before and after a behavioral paradigm designed to induce learning in mice and then comparing the results, we can begin to understand some of the basic steps that occur in a live animal forming a memory. The cluster is essential for our research since figuring out where millions of short DNA sequences map on the mouse genome is a very computationally intensive process. Not only would the results take much longer to obtain on a single desktop, but we would be very limited in our ability to modify parameters of our analysis to see how that affects the results. What would take weeks or months take hours or days thanks to the resources provided by the Tufts cluster.

Chao-Qiang Lai

Our Tufts/HNRC research is focusing on Nutrigenomics to study gene-diet interactions in the area of cardiovascular diseases, utilizing both genetic epidemiology approaches as well as controlled dietary intervention studies. This research involves the investigation of nutrient-gene interactions in large and diverse populations around the world with long-standing collaborations with investigators in Europe, Asia, Australia and the United States. For the current project, I was using the cluster to deal with a large amount of genome data, such as genetic variants in human genomes, which can not be handled with my laptop computer. The cluster is over 50X faster than my laptop. It would not be possible to complete my research project without it!

Anoop Kumar

Professor Lenore Cowen, Matt Menke, Noah Daniels and I used the cluster to hierarchically organize the protein structural domains into clusters based on geometric dissimilarity using the program Matt (http://bcb.cs.tufts.edu/mattweb/). The first step in the experiment was to align all the known protein domains using Matt. To compare all the 10,418 representative domains against each implied running Matt approximately 54 million times. While a single run takes only about 0.1 CPU seconds, running it 54 million times would take approximately 74 days on a single processor. By making use of the ability to run multiple jobs on separate nodes on the cluster we split the job into smaller batches of 0.5 million alignment operations per batch, thus creating 109 jobs that we submitted to the cluster. Each job took approximately 15 hours which is a significant reduction from 74 days. By running the jobs simultaneously on separate nodes of the research cluster we were able to reduce the time taken to perform our analysis from 2.5 months to less than a day. This speed up proved to be an additional benefit when we realized we needed to run an additional experiment using an alternative to Matt, as we were able to run that second experiment without significantly delaying our time to publication. This research has resulted in a paper, "Touring Protein Space with Matt", that has been accepted to the International Symposium on Bioinformatics Research and Applications (ISBRA 2010) and will be presented in May.

Recognizing the value of running large tasks on the research cluster and the future CPU intensive programming requirements of the group, Prof. Cowen has contributed additional nodes to the TTS research cluster. While members of the BCB research group (http://bcb.cs.tufts.edu/) get priority to run programs on those nodes anyone having account on the cluster can run programs on them.

Keith Noto

I'm a postdoc in the Computer Science department, working on anomaly detection in human fetal gene expression data. That is, how does one distinguish "normal" development (meaning: like what we've seen before) from "abnormal" (different from what we've seen before, in the right way) over hundreds of samples with tens of thousands of molecular measurements each, when we don't even really know what we're looking for? I use the Tufts TTS cluster to test our approaches to this problem on dozens of separate data sets. These computational experiments take thousands of CPU hours, so our work cannot be done on just a handful of machines.

Ken Olum, Jose Blanco-Pillado and Ben Shlaer

Ken Olum, Jose Blanco-Pillado and I are using the cluster to attempt to solve an important question in cosmology, namely "How big are cosmic string loops?" Cosmic strings are ultra-thin fast moving filaments hypothesized to be winding throughout the universe, most of it in the form of long loops. There has been much theoretical interest and work in cosmic strings, but before we can connect the theory to future observations, we need to know the typical sizes of the loops the network produces.

It turns out this is an ideal question to solve numerically, since the evolution of each individual string segment is easy to compute, and the tremendous scales over which the network evolves makes analytic work extremely difficult.

What makes this exciting now is that the previous generation of numerical cosmic string simulations disagreed on what the right answer is. We believe that current hardware is sufficient to enable us to answer the question definitively.

Alireza Aghasi

The research that I am doing is very computational and requires a lot of processing and memory. I basically deal with Electrical Resistance Tomography (ERT), for detection of contaminants under the surface of the earth. The problem ends up being a very high dimensional Inverse problem which is intensively ill-posed. Dealing with such a problem without appropriate processing power is impossible. Once I became aware of the cluster I started exploring it and realized that some features of it really help me in the processing speed. The excellent feature which really interested me was the good performance in sparse matrix calculations. Star-P does an excellent job dealing with very large sparse systems compared with other platforms. Personally I experienced some very good results using Star-P.

Umma Rebbapragada

I am a Ph.D. student in computer science, studying machine learning. My research requires me to run experiments in which I test my methods on different data sets. For each data set, I may need to search for or test a particular set of input parameters. For each particular configuration of the experiment, I will need to perform multiple runs in order to ensure my results are statistically significant, or create different samplings of my data. In order to test a wide variety of configurations across multiple data sets, I exploit the cluster's ability to run "embarrassingly parallel" jobs. I have submitted up to 2000 jobs at a time, and have them finish within hours. This has allowed me to test new ideas quickly, and accelerated my overall pace of research. I have different software demands depending on the project I'm working on. These include Java, shell, perl, Matlab, R, C and C++. Fortunately, these are all well-supported on the cluster. I also plan to explore MPI one day and take advantage of products like Star-P, which are available on the cluster.

David Bamman and Greg Crane

We use the cluster now for two main purposes: parallel text alignment (aligning all of the words in a Latin or Greek text like the /Aeneid/ or the /Odyssey/ with all of the words in its English translation) and training probabilistic syntactic parsers on our treebank data. Both of these are computationally expensive processes - even aligning 1M words of Greek and English takes about 8 hours on a single-core desktop, and for my end result, I need to do this 4 separate times. Using a multi-threaded version of the algorithm (to take advantage of each cluster computer's 8 cores) has let me scale up the data to quantities (5M words) that I simply could not have done on our existing desktop computers. Most importantly, though, the cluster environment lets me run multiple instances of these algorithms in parallel, which has greatly helped in testing optimization parameters for both tasks, and for the alignment task in particular lets me run those 4 alignments simultaneously - essentially letting me work not just faster but more accurately as well.

Luis Dorfmann

Evaluating patient-specific Abdominal Aortic Aneurysm wall stress based on flow-induced loading

In this research we develop a physiologic wall stress analysis procedure by incorporating experimentally measured, non-uniform pressure loading in a patient-based finite element simulation. First, the distribution of wall pressure is measured in a patient-based lumen cast at a series of physiologically relevant steady flow rates. Then, using published equi-biaxial stress-deformation data from aneurysmal tissue samples, a nonlinear hyperelastic constitutive equation is used to describe the mechanical behavior of the aneurysm wall. The model accounts of the characteristic exponential stiffening due to the rapid engagement of nearly inextensible collagen fibers and assumes, as a first approximation, an isotropic behavior of the arterial wall. The results show a complex wall stress distribution with a localized maximum principal stress value of 660 kPa on the inner surface of the posterior surface of the aneurysm bulge, a considerably larger value than has generally been reported in calculations of wall stress under the assumption of uniform loading. This is potentially significant since the posterior wall has been suggested as a common site of rupture, and the aneurysmal tensile strength reported by other authors is of the same order of magnitude as the maximum stress value found here. The numerical simulations performed in this research required substantial computational resources and data storage facilities, which were very generously made available by Tufts University. This support is gratefully acknowledged.

Rachel Lomasky and Carla Brodley

We address problems in the two areas of Machine Learning and Classification. A new class of supervised learning processes called Active Class Selection(ACS) addresses the question: if one can collect n additional training instances, how should they be distributed with respect to class? Working with Chemistry's Walt Laboratory at Tufts University we train an artificial nose to discriminate vapors. We use Active Class Selection to choose which training data to generate. And In the area of Active Learning we are interested in the development of tools to determine which Active Learning methods will work best for the problem at hand. We
introduced an entropy-based measure, Average Pool Uncertainty, for assessing the online progress of active learning. The motivating problem of this research is the labeling of the Earth's surface to create a land cover classifier. We would like to determine when labeling more of the map will not contribute to an increase in accuracy. Both Active Class Selection and Active Learning are CPU-intensive. They require working with large datasets. Additionally, experiments are conducted with several methods, each with a large range of parameters. Without the cluster, my research would be so time-consuming to be impractical. For additional details see the Rachel Lomasky pdf attachment.

Eugene Morgan

The Tufts linux cluster allows me to work with large amounts of data within a reasonable time frame. I first used the cluster to interpolate sparse data points over a fairly large 3-dimensional space. The cluster has also dramatically sped up the calculation of semivariance for dozens of sections of seafloor containing vast numbers of data points, quickly performed thousands of Monte Carlo simulations, and computed statistics on one of the largest global wind speed datasets containing ~3.6 billion data points. I have most recently used the cluster find optimal parameters for rock physics equations using a genetic algorithm. Most of these activities have been or will be incorporated in technical publications.

Eric Thompson

We have used the Tufts Linux Cluster to further our understanding of the seismic response of near-surface soils. This behavior, often termed "site response," can often explain why locations heavily damaged by an earthquake are frequently observed adjacent to undamaged locations. Standard modeling procedures often fail to accurately model this behavior. The failure of these models is often attributed to the uncertainty of the soil properties. However, using the Tufts Linux Cluster we have shown that the underlying theoretical assumptions of the standard model (vertically incident plane SH-wave propagation through a laterally constant medium) are responsible for the failure to match the observed site response behavior.

Andrew Margules

The research that I am currently conducting is in the area of Passively Actuated Deformable Airfoils. The largest presence of airfoils today is contained within the aerospace and transportation industries. Like those on commercial and military aircraft, the basic teardrop airfoil shape is augmented with a series external structures which aid in take-off, landing, and cruising flight. While they perform specific and important functions, they add additional weight to a system which is highly immersed in weight management. What my research is looking into, is try find a way to develop an internal structure for an airfoil that would provide similar shape change, without the added external mechanisms. To do this, I am using two different computational software packages. COMSOL Multiphysics allows for the examination of the fluid-structure interaction of the airfoil and moving air. Using different internal rib structures, a goal of finding an appropriate structure is hoped to be achieved. In addition, I am using the computational fluid dynamics package Fluent to help visualize velocity and pressure fields over deformed and undeformed airfoil shapes. If this

software was not available through the academic research cluster, this research would extremely slow process. The governing physics behind these simulations is complex enough that without the computing power of the cluster, I do not believe that we would be able to perform it. In the last twenty or so years, a focus has shifted from passive actuation to active actuation. Hopefully, this research will help to launch a renewed interested in this field.

Ke Betty Li

I am a researcher in the Department of Civil and Environmental Engineering. Our research focuses on the investigation of how various contaminants affect the ground water quality and how we could design remediation systems. An important approach we are using for this type of investigation is modeling contaminant fate and transport in the subsurface on computers. The resources provided by Tufts Cluster Center are very important to us. Our simulations usually take days or even weeks on a single CPU. The clusters can either expedite each simulation if we use simulators that enable parallel computing, or allow us to simulate multiple serial processes simultaneously. The significant improvement in computing efficiency is critical for us to commit work quality to funding sponsors. We expect that our work will improve the cuurent understanding of contamination in the subsurface, provide cutting-edge assessment tools, and stimulate innovative treatment technologies.

Eric Miller

Our work concerns the development of tomographic processing methods for environmental remediation problems. Specifically, we are interested in using electrical resistance tomography (ERT) to estimate the geometry of regions of the subsurface contaminated by chemicals such as TCE or PCE. Though the concept of ERT is not unlike the more familiar computed axial tomography (CAT) used for medical imaging, the physics of ERT are a bit more complicated thereby leading to computationally intensive methods for turning data into pictures. Luckily these computational issues are, at a high level, easily parallelizable. Thus, we have turned to Star-P as the tool of choice for the rapid synthesis of our algorithms.

Michael A. Simon

Nonlinear dynamic modeling of Lepidopteron mechanosensors

The Trimmer Lab is interested in the control of locomotion and other movements in soft bodied animals. I have been analyzing the activity of a specific mechanosensor trying to understand how it influences abdominal movement, a critical question for animals with no rigid components. One particularly powerful analytical tool for analyzing such sensors is nonlinear analysis using Gaussian white noise as a stimulus. One challenge of this technique, however, is that it is computationally complex. Even storing the matrices involved in these computations is beyond the capabilities of the typical personal computer. The Tufts Linux Research Cluster offers me the resources necessary to run these computations and analyze the results without needing to invest in new, complicated, or expensive analytical hardware or software. It also allows me to use software that would have been difficult to acquire for our lab, alone. Without this resource, following this line of inquiry would have proved a costly endeavor, possibly prohibitively so. We hope to apply our results to the development of computer and robotic models, with the eventual goal of designing a soft robot, a groundbreaking engineering application with substantial implications for design in the biomedical engineering arena, as well as in other areas of engineering.

Katherine L. Tucker

Use of the Bioinformatics cluster has been invaluable to our research. We use a genetic analysis software named SOLAR which is Linux/Unix based. This software and the methods used in it are cutting edge. We are able to perform varous genetic computations with ease. In the past some student have had to do these calculations by hand because of a lack of access to such software. However, hand calculations are only possible for small sample sizes and simple genetic analysis. Our current work with Solar includes over 5,000 individuals and we are using some of the most advanced methods available. The cluster allows us to do large computational runs that would not be otherwise possible. Thus, our current work would not have been able without access to SOLAR on the bioinformatics cluster. In addition, this type of analysis is being more common and will be a greater part of our efforts in future years. Use of the bioinformatics cluster helps our research to remain competitive and important in our grant application process. Our lab is the first to use SOLAR on the bioinformatics cluster, however, since we have been using it, many labs have inquired about how to gain access. I sincerely thank you for your work in helping us gain access to the software and the service you have provided through the Bioinformatics cluster.

Jeffery S. Jackson

I am a grad student in Mechanical Engineering and I am conducting research on microfluidic mixers. I use the Cluster01 to create and run fluid flow models on COMSOL Multiphysics. The COMSOL program solves the Navier Stokes equations for transient fluid flow and the convection diffusion equation. For the models that I create to be accurate, though, they require more elements and time steps than my computer, or the computers in the EPDC, can handle. This is where the cluster comes in very handy. I usually have the Cluster run any model that is more complicated than a 2D model with 30,000 elements. The most complicated model I have had the cluster solve consisted of 90,000 elements. This model took 30 hours for the Cluster to solve, which is something that no other computer resource I have access to could do. Another nice benefit of the Cluster is being able to use it from home. I live in Providence, RI and it takes me two hours to get to Tufts by train. So, I only come in when I have to. Having remote access to the Cluster makes this possible. Without the Cluster, or the very helpful people who provide excellent technical support, I would never have been able to do the research I needed to to finish my Master's Thesis.

Erin Munro

I'm studying Computational Neuroscience in the Math department. My research consists of doing MANY simulations. That being said, I would not

be able to do this research without the cluster! I simulate networks of thousands of neurons interacting. While there are some simulations that take a few minutes, the majority of them take 45 minutes to an 1.5 hours on one node. The last time I calculated, I'd like to run over a month's worth of these simulations. On top of this, I've run several very important simulations that take 1.5 days on 16 nodes. I had to run these simulations in order to try to reproduce results from Roger Traub's research. My current project is to try to explain these results. We tried to find a simpler way to explain them without reproducing the full model, but we found that we couldn't do it. With the cluster, I have been able to reproduce the results to the best of my ability. Furthermore, I've been able to dissect the model, and run many more simulations to get a much better understanding of what is going on in his results. I feel like I'm coming close to fully explaining the results, and have just presented a talk at BU explaining my ideas. None of this would have been possible without the cluster.

Casey Foote

My research for my MS in Mechanical Engineering is based on using the software available on the cluster to model a cold forging process. This model, paired with experimental data, will then be used to develop a tool to predict forging work piece cracking. The tool will provide a manufacturer of airfoils for use in the aircraft engine industry a method to rapidly develop new processing while avoiding costly physical trials.

Aurelie Edwards

My graduate student Christopher Mooney performs simulations of unsteady, turbulent fluid flow in a bioreactor with a stir-bar, using Femlab engineering software. Prior to having access to the Tufts cluster, he was experiencing extensive memory usage problems. On a PC with 2GB of RAM using Windows XP, he was only able to
access about 40% of the memory, due to fragmentation issues, and his simulations did not converge. We were both relieved to learn that we could have access to the
Tufts cluster and its Linux platform that offers 4GB+ of memory space. The latter has thankfully allowed us to solve increasingly complex models. For example, using his PC, Chris could solve finite element Navier-Stokes fluid flow problems with an element mesh density that limited the problem to about 100,000 degrees of freedom, beyond which he ran out of memory. He often received "low mesh quality" error messages that hindered the mathematical convergence of the solution. On the cluster, he now has enough memory to refine the mesh and run models with 300,000 degrees of freedom. Chris still runs into "out of memory" problems on the cluster, but much less frequently. The technical staff at Femlab, when told of the kinds of problems we envision solving in the coming years, suggested using a server with 10 to 16GB of memory space to run these models with adequate mesh resolution. In other words, if you were to increase the capacity of the Tufts cluster, we would be takers!

Gabriel Wachman

I use the cluster to conduct experiments relating to my work in machine learning. I am in the computer science department. The experiments I have been running have generally been to aid in the comparison of different learning algorithms. By running many experiments over a range of parameters, I can collect data that helps me to draw conclusions on the behavior of the algorithms. Without the cluster, much of the work I have done would have been impossible or at best severely limited.

Alexandre B. Sousa

I am a grauate student with the High Energy Physics Group and as part of the MINOS experiment collaboration, I have been one of the main people responsible for mass event reconstruction using the Fermilab fixed-target farm. Earlier this year, a Mock Data Challenge was issued to the experiment in order to shake down reconstruction and analysis shortcomings before real data collection starts in January. This effort requested the generation of a rather large MonteCarlo sample, which was subsequently reconstructed at Fermilab. However, the generation of the MC sample was quite hard to setup at Fermilab, where space constraints, e-bureaucracy and competition with other experiments meant we would not be able to do it in a timely manner. That was when I decided to test the Tufts Linux Cluster to perform this task. I was setup with an area on the /cluster/shared space within a day of my original request, and after a few tests, I was able to generate 80% of the total necessary MC sample in less than a week. I was of course lucky to be almost the exclusive user of the cluster for that period, but I really had no problems setting things up and using it in what is seen as a nice success of the Tufts High energy Physics Group. Giving this success we have volunteered to become one of the spearheading institutions taking part on the upcoming MC generation effort which should start later this month, and the gained experience was transformed in a document and relayed to other institutions that are starting to run their own clusters and hope to join this effort. I have used the cluster a second time to do a customized reprocessing data for the CC nue analysis group, which I integrate, which required compilation in the cluster of the MINOS Offline Software, installation of a mysql database and assembling some shell scripts to handle the job output. That went quite well, and the full data sample was processed in 2 hours, with about 1 day of setup. Having worked for 2 years with the Fermilab batch farm, I was mainly impressed by the speed of the network connection of the CPU nodes to the I/O node, almost 20 times the Fermilab data transfer speeds and also by the great flexibility of use given to the users, which implied minimal
back and forth contact with the admins and dramatically improved work efficiency.