### Graduate Mentors in the PURE Program:

<table>
<thead>
<tr>
<th>Mentor</th>
<th>Research / Project Summary</th>
</tr>
</thead>
</table>
| **Abdullah Akce** | **Project-1 Title:** Designing a brain-machine interface for web browsing  
**Project-1 Description:** Brain-computer interfaces enable users to control applications with thoughts instead of with input devices like a mouse or a keyboard. In this project, your task is to design an interface to allow users to control an internet browser with their thoughts. You are also welcome to come up with your own use scenario for brain-computer interfaces. This project requires you to answer the following questions by doing research:  
(1) which paradigm will be used to map brain activity to meaningful commands?  
Several paradigms established in the last decade by researchers, a literature review and discussion with the mentor will be useful.  
(2) how the visual stimuli will be presented to the user?  
There are a few number of interface designs you can find in the literature. You are expected to review the state-of-the-art and to introduce new ideas with the mentor.  
(3) how to implement your design?  
You are free to use any tools, software and programming language of your choice. We don’t require you to work with a neural sensor to measure brain activity, instead you can work in simulation. If students are willing, we can give them access to labs in Beckman institute for doing real experiments. |
| **Abhinav Bhatele** | **Project-2 Title:** Using an indoor-flying helicopter as a personal assistant  
**Project-2 Description:** Mobile phones are the most popular personal assistants. But, they have no motion capability. Imagine having a flying robot equipped with an embedded processor, two cameras, wifi and a GPS at your service. How do you want this robot to assist you? The goal of this project is to design, develop and evaluate the use of a flying robot assistant in a social context. We can provide you an AR-Drone quadrotor ([http://ardrone.parrot.com/parrot-ar-drone/usa](http://ardrone.parrot.com/parrot-ar-drone/usa)). You might want to do one or more of the following:  
1) face recognition with the on-board cameras,  
2) adding on-board computation capability (the device supports control over wifi, but you need iPhone or a laptop to manipulate it),  
3) finding a person in a crowd by doing an efficient search,  
4) tracking a moving person. |

Abhinav works on network topology-aware mapping of parallel objects to processors. This is essentially the same as the graph mapping/embedding problem in mathematics. It involves optimal mapping of one graph on to another while trying to reduce some metric.

The one semester project involves developing a visualization tool to test mapping algorithms. Input to the tool is the decision of the mapping algorithm. The processor topology is generally a 3-dimensional grid. The tool will display which object is mapped on which processor in the 3D grid. It will show which processors are overloaded with objects and which links are overloaded with communication. The tool will be interactive with zooming capabilities to look at one part of the graph and newer features will be added as required.
The DeNovo project is rethinking the design of multicore hardware driven by disciplined parallel programming models. Our work covers the spectrum of novel memory hierarchy design (cache coherence protocols, cache design, synchronization, communication architecture), novel programming language constructs, and new parallel applications and their evaluation. The goal is to provide far more complexity-, power-, and performance-efficient architectures than previously possible.

We work on a framework called K that allows one to specify programming languages formally. Essentially, it is a methodology for specifying languages that is formal (mathematical), so that analysis of programs written in the language, as well as analysis of the language itself, can be performed. Examples of this kind of analysis include all types of model checking, theorem proving, type checking, data-race detection, etc. The framework should also help the designer visualize his or her language features, and potentially to aid writing programs in the language itself. Finally, the framework should also be amenable to translation into other domains. A key feature of our methodology is that the specifications are executable. Thus, one should be able to convert specifications to interpreters or compilers. This is already being done to some extent.

I am open to any project idea in my research area and have three specific project suggestions.

- **Analysis of the temporal evolution of data deduplication systems, and the effect of temporal properties of deduplication relationships on reliability constraint satisfaction.**
  - Deduplication is a new technology which has been deployed extensively in modern storage systems, which involves identifying duplicate data in the data store, and replacing it with references to one or more instances of the data in the system. While this has been very beneficial in terms of storage efficiency, it has reliability consequences which are not fully understood. I have been working with IBM Almaden Research Center to understand these consequences, and while we have submitted our initial study to the File and Storage Technologies conference, we are working to expand our understanding of the reliability consequences of data deduplication through the study of additional systems, and temporal data which will allow us to model how a deduplication system changes over time.

- **Performance analysis of multi-instance data deduplication systems.**
  - See previous project for an overview of deduplication. Our previous work has led us to develop an algorithm for determining a deduplication strategy using multi-instance deduplication to meet certain reliability constraints. We would like to build on this study by modeling the performance consequences of such a deduplication strategy to help further understand the cost imposed by meeting certain reliability constraints in deduplicated file systems.

- **Performance analysis of just in time job re-prediction for priority queuing systems.**
  - An extension of work in the literature to which sorts jobs in a queue to increase throughput based on job size prediction. We propose to expand on this work with a just in time reprediction extension to existing algorithms.

**Desired Skills**

- A strong background in set theory and discrete math.
- Background in C/C++ programming.
There is no pre-requisite for previous experience in Stochastic Modeling, or the specifics for the outlined projects. All projects have direct applications in real systems, and are ideas I would like to pursue for publication with mentees.

Develop 3-D visualization software (based on Paraview, VTK) to analyze and explore oil and gas reservoirs and deep saline formations where carbon dioxide is stored. Software will be developed to provide novel visualizations of several separate fluids (e.g., oil, natural gas, groundwater, CO2), various fluid properties (e.g., viscosity, pressure, temperature), and the various geologic units and their properties. Later phases of project will include development of data mining and autocorrelation techniques for these complex data sets. Project is part of a 3-year US DOE study, in conjunction with the ISGS (Illinois State Geological Survey) and the ISGS stereom-  

SAFECODE is a safety checking compiler built using the LLVM Compiler Infrastructure. SAFECODE enforces memory safety guarantees on code written in C and C++ using a combination of static analysis and program transformation; SAFECODE first tries to prove memory accesses safe using static analysis and then instruments the code as necessary when safety cannot be proven. The memory safety guarantees prevent buffer overflows and uninitialized pointer use and render dangling pointer dereferences harmless. SAFECODE has been used both for user-space applications as well as on kernel code.

Have you ever thought, "Where is that PDF or website I was looking at the other day? I don't remember its name, but I was listening to AC/DC while I was reading it." Currently, you cannot search your computer this way. We have developed new technologies to help allow you to search in that exact way!

In our project, we are looking for a mentee who is interested in helping us implement this type of search tool. Our goal is to distribute this as a free search tool for everyone's computer, so your code may be used by thousands.

My research background is on the design and measurement technologies of computer systems and software. Some research topics I want to investigate with promising undergraduate students are described as follows.

Motivation.

The advances in semiconductor technology brings us both challenges and opportunities. (Challenges) Experts and major semiconductor vendors are predicting the current mainstream memory devices (e.g., DRAM) will not be scalable anymore for example after 5 years (e.g., to below 20nm). Like an
automotive industry (where none competes for the speed of car for sales),
our computing industry will experience a paradigm shift from competing for
high-density chips and high-performance systems to dependable, safe, low-
power, and for example more predictable systems. That would be where we
young researchers need to challenge for high returns and big contributions.
(Opportunity) Material and electrical engineers are actively investigating
new memory and semiconductor technologies. For example, NAND flash
memory is one that has been successfully used in mobile and embedded
systems and is now coming to us (i.e., desktop and server markets) as a
form of solid-state disk (SSD). Also, some memory devices show their
strong potentials as a replacement of current DRAM technology beyond
20nm (e.g., PRAM, MRAM, and FRAM).

Research topics that I can advise include:

- Design, optimization, and evaluation of a machine learning algorithm (i.e.,
  expertise in machine learning is what undergrad student is supposed to
  bring and I'll provide a problem and help them validate/evaluate the models
  and assumptions) for hybrid storage systems (e.g., using HDD and SSD).
  I've proposed and filed a patent on a hybrid storage architecture ~3 years
  ago (on page 255 of http://www.linuxsymposium.org/2008/ols-2008-
  Proceedings-V2.pdf). This architecture is different from the architectures
  presented and led by other companies in the industry (e.g., Hybrid-HDD by
  Microsoft and Turbo Memory by Intel) because this for example manages
  data in a file granularity (than a block granularity). An efficient algorithm
  that can learn patterns in common disk I/O operations and intelligently migrates
  files from one to the other would be an important contribution if designed
  and evaluated. Here a tradeoff is SSD is faster but more expensive than
  HDD. Both SSD and HDD have a fixed capacity, and there is a migration
  overhead.

- Extending the Linux kernel to support hybrid memory architectures (e.g.,
  one with ECC (error correction code) protection, and the other with
  EDC (error detection code) protection). This design is motivated by my
  recent measurement study (http://ieeexplore.ieee.org/xpls/abs_all.jsp?
  arnumber=5544287&tag=1) that shows a large portion of data stored in
  memory is recoverable (if corrupted due to hardware or software fault) by
  simple software techniques. Thus, this type of recoverable data no need to
  use an expensive ECC protection (e.g., ECC protection requires a large
  size of memory space if it's for multi-bit errors, i.e., more common due to
  small transient size).

If you are interested in one of these two topics or some relevant problems
(in dependable computer system design - both HW and SW - and
measurement), please feel free to contact me by email.

Project Title: Fine-Grained Named Entity Recognition

Project Description: One of the key problems in language understanding
and data mining is to identify entities of interest in free-text data. For
example, we might be interested to find all the laptop products in text. We
currently have a solution (please play with our demo: NER Demo) for finding a rather
restricted set of entities: people, locations, and organizations. However, we
are unable to recognize more fine-grained entity types. For example - can
we extract
all the politicians in the text? All the quarterbacks? All the dog breeds? The
project will focus on blending two directions: (1) building context-

independent lists (e.g the list of all politicians) from large datasets (e.g
Wikipedia or the Web) and (2) using these lists in context-dependent way to
solve ambiguity (e.g. decide when “Michael Jordan” refers to the basketball
player, and when it refers to the Berkeley scholar.).

Project Scope: The project goals are ambitious, and will leave plenty of
room for independent research and contributions. At the same time, we
have an attainable and a well-defined goal which we believe is achievable in
one semester, and which will be the core of the project. We will tackle the
ACE 2005 dataset with around 16 entity types, including weapons, vehicles,
countries, cities etc. Extending our current system to achieve new state-of-
the art performance on the ACE 2005 dataset would define a successful project. Bonus points will be given for obtaining this goal with general and scalable techniques which will easily allow addition of other types.

Security requirements change. Many legacy systems fail to cope with the changing requirements because it is infeasible to redesign these systems. Security experts generally say that security "cannot be added on, it must be designed from the beginning." We want to show that it is possible to "add on" security to a system by applying program transformations.

There are many kinds of program transformations. Compilers transfer programs in source form to equivalent programs for a particular machine language. Refactorings are source to source transformations that change the structure of programs but not their behavior. The transformations that we are interested in include both source to source and binary to binary transformations. They improve the security of systems, which means that they do not preserve all types of behavior. They preserve expected behavior, but should change a system's response to security attacks.

We are interested in automated program transformations that change programs to eliminate security threats. They need to preserve as much of the behavior of the system as possible, while changing the behavior that leads to security flaws. We have already worked on some program transformations that increase security; partitioning a monolithic program into multiple processes, running a process in a constrained environment, and checking for buffer bounds for every buffer write operation.

The use of security-oriented program transformations would improve the traditional approach of security engineering. The traditional approach relies on designing security from the ground up and writing patches to fix new vulnerabilities. It is infeasible to redesign software every time a new security threat emerges. On the other hand, a patch fixes the vulnerability at a fixed set of points; it does not globally remove the vulnerability. Our approach to security engineering replaces ad hoc security patches with patches automatically derived from program transformations that globally fix a security vulnerability. This approach not only fixes the vulnerability, but also allows a software developer or a software maintainer to add "security on demand".

Java PathFinder (JPF) is an explicit-state model checker for Java, namely a tool that can be used to find bugs or verify properties of Java programs. JPF is the first open-source tool from NASA. This project will involve extending the functionality and/or improving the performance of JPF, such that it is easier to find bugs, bugs can be found faster, or more bugs can be found. Several previous projects on this topic resulted in undergraduate students publishing research papers and contributing code to JPF.

Project Title: Analyzing Browser Extensions for vulnerabilities

Project Description:

Driving the Internet revolution is the modern web browser, which has evolved from a relatively simple client application designed to display static data into a complex networked operating system tasked with managing
many facets of a user's online experience. To help meet the varied needs of a broad user population, browser extensions expand the functionality of browsers by interposing on and interacting with browser level events and data. Web extensions with security vulnerabilities can have disastrous consequences, as they often run with full browser privileges. Thus an attacker can exploit extension weaknesses to take over the entire browser, stealing cookies, protected passwords, and other confidential information, or even hijack the host system, without revealing its actions to the user.

Based on our studies of security vulnerabilities in browser extensions, we observed that many of the errors translate to certain types of explicit information flows from injectible sources to executable sinks. For extensions written with benign intent, most attacks involve the attacker injecting JavaScript into a data item that is subsequently executed by the extension under full browser privileges. To help understand the behavior of the browser extensions better, we developed a precise information flow analysis tool, VEX. The tool performs static analysis of JavaScript code in order to identify the above information flows, which are likely sources of trouble in Firefox extensions.

For the project, we would like the student to use the tool for finding new vulnerabilities in Firefox extensions. The project would involve understanding browser and browser extension design, understanding known browser extension vulnerabilities and how they can be exploited, understanding static information flow analysis and understanding how the tool works. Once the student gets a fair idea about all the above aspects, she can then use the tool to generate the results for the different Firefox extensions and then analyze the results. This would require a fair amount of programming to improve the current tool.

New mentors can sign up easily by just following the Instructions for Creating a New Mentor Page. If you are a current mentor and would like your profile copied from the previous semester, simply send a blank message to uiucpure [at] gmail dot com with subject "Copy Profile.". Feel free to ask us questions as well.

If you are looking for a Mentor profile from a past semester and do not see it in this listing, check under the PURE Archive.