PACI REU Proposal Application

Project Title: Programming Gridflows for Scientific Data Management
Relevant URL: http://www.npaci.edu/DICE/SRB/matrix/
No. of students: 2

Desired Budget: $ 9000
Description: Undergraduate students who join our REU programs are usually self-motivated and are looking for avenues that would promote their interest in research and graduate school. They like challenges and want to be treated as a full time professionals who can act on their own while contributing to an overall teamwork.

The SDSC Matrix Project, is a research and development activity within the Data Grid Technologies Group at SDSC focusing on Grid Workflows (Gridflows). The members of this project Matrix Project (except the core architect), have all been students – from UCSD, UCSB and through the REU program before. The project has made a slow but steady progress. The earlier students have participated in the invention (or development) of language called “Data Grid Language” (DGL) as part of the Matrix project. We have reached a stage where in this language the software developed by the students has already used by projects at SDSC. The undergraduate student who participated in this project earlier as a REU is now graduating and will be applying for graduate school in computer science. This time, the plan is to have two students who would continue using the software for two major activities. Each of this activity is challenging, yet accomplishment by undergraduate students. They involve other major academic/research institutions like Fermi Lab in Chicago and Rutherford Appleton Laboratory (RAL) in UK. Even though these activities are of a short span duration, they involve long run projects such Sloan Digital Sky Survey (SDSS) and the Storage Resource Broker Project (SRB). These motivate the students more as they know their work is going to be used by others.

Towards the completion of the project as part of the REU program the student will have added value to themselves in terms:
§ Working in a professional project spanning multiple domains and countries
§ Knowledge about managing very large data sources and techniques to be used in designing gridflow pipelines
§ Programming using the newly invented Data Grid Language
§ Preparing bi-weekly reports of their work and interacting with users (scientists) from other research labs.
§ Promote their interest in Graduate School and Research Work.

The students will be involved in two major activities at SDSC. Both of these activities are very much related and require interaction with the off-site team. The students will participate in our regular meetings using access grid / teleconference / e-mails and develop some customized software on top of the SDSC Matrix software for our users.


Based on our interaction with the above mentioned projects, the astronomy communities in US require a service to co-add of FITS images from the sky. Coadd in this case is a data as well as compute intensive problem. The solution is to have many grid sites participate together in an autonomous way and share the load of the gridflow based on the availability and organization
of the input data. We have planned to integrate the existing SDSC Matrix software and FNAL Coadd daedaleos software. The incoming students will develop some modules that can integrate the software and also help in the deployment of the integrated software in multiple sites across the country.

Activity 2: Data Grid administration tasks. A requirement from UCSD Libraries and other SDSC SRB users is to automate some of the common data grid tasks. Users would like to calculate the md5 checksum of their digital holding to make sure the data is valid and not corrupt. A similar requirement is to start some gridflows to replicate or copy data across sites as expressed by our SRB UK collaborators in Rutherford Appleton Laboratory (RAL) in UK. They would like to have administrative data grid operations start/restart during the weekends only and stop/pause whenever required. The students will add the context of time to the data grid operations in Matrix. This small step would be giant leap in managing the large amount of data in data centers like ours too. Its easy of the students to add few lines of code on top of the existing software – but, it gives a powerful functionality for the data grid operators to schedule replication, archiving time windows.

Team Members, Affiliations: § Jim Annis, Fermi National Laboratory, Astronomy domain Scientist
§ Neha Sharma, Fermi National Laboratory, Software/Computer Science
§ Daniel Moore, University of California, Santa Barbara. Matrix Team

Plan to Support W/M, Persons w/disabilities: The only requirement that we prefer from these young and energetic minds is that these undergraduates have experience in programming - preferably Java. This time we will prioritize on supporting undergraduate women and give them the first preference. We will advertise about this program through UCSD and UCSB departmental mailing lists. In addition information about these will be mentioned in notice boards in multiple departments in UCSD (Computer Science, Physics and Engineering departments). If we are not able to get any undergraduate women our next preference would be for underrepresented minorities. We can even train them with programming with java if required.

Two emails of support from project members: Hi Arun

    I am happy to note that you are applying for this REU. I believe that the proposed project will help undergraduates to get a hands on experience in cutting edge technologies of data grids and web services. I enthusiastically support this work and wish you all the best in this endeavor.

    thanks
    raja

--

=====================================================================
Arcot (raja) Rajasekar Email sekar@sdsc.edu
Director Voice 858-534-8378
Data Grid Technologies Group Secr. 858-534-5121
San Diego Supercomputer Center, MC 0505 Fax 858-534-5077
University of California, San Diego http://www.npaci.edu/dice/srb
9500 Gilman Drive, Bldg 109 __ __
Please Note: A confirmation mail from Reagan Moore (on travel now) would be forwarded to Ange Mason once I receive it. Jim Annis from Fermi Lab has agreed for this over the phone.

Plan to integrate students: The students will participate in our meeting using access grid or teleconferences or regular e-mail discussions. They will play an active role in both these activities as they would be people at SDSC who carry out this project. Arun Jagatheesan will help them in this process and act as an active mentor.

Since both the activities are similar, both the students will work together. One student will focus more on Activity One and collaborate more with the Astronomers from Fermi National Accelerator Laboratory. The other will collaborate with the local UCSD library and also with UK developers who would start looking into this in the January next year. Another student from UCSB (who is already working in this project) will be developing some GUI components along with the two new incoming students.