

### Contents

---

1. [Overview](#)
2. [Technology](#)
3. [Resources](#)
4. [Methods](#)
5. [Tools](#)
6. [The Science](#)

### Overview

---

The digital revolution has left us awash in data. Just as Moore's law is fundamentally driven by the resolution of silicon lithography, so too is the abundance, cost and resolution of silicon semiconductor sensor technology. Historically, advances in CPU speed have enabled ever larger and more detailed simulations, resulting in enormous volumes of synthetic data. Today, however, the increasing number and resolution of digital sensors is changing the nature of basic research across disciplines. Addressing these changes will require contributions in the areas of methods, tools, and technology, and this eScience SIG is your place to participate in the revolution.

For now, we're providing this static place holder. In the near future, we'll adopt a more collaborative approach and encourage you to make this page your own by contributing your knowledge to the overall UW eScience effort.

### Technology

---

These are the nuts and bolts usually associated with research computing. Come here to find practical assistance in assembling your own high performance computing (HPC) facility.

#### Storage

eScience is all about the data. Lots and lots of data. This is the place to share techniques for dealing with this problem and experiences with approaches you've tried in the past. Broadly speaking, the storage problem breaks down into three components: performance, capacity, and protection, and the technologies for addressing these three areas often differ substantially.

We'll be providing guidance in the deployment of a variety of storage technologies, from FC SANs, to cluster filesystems, to HSM data protection schemes. We invite you to contribute your experiences, too!

#### Computing

Computing resources are not all created equal. Some problems are lucky enough to be amenable to large-scale distributed approaches, such as [BOINC](#), while others require large shared memory architectures typically only found at the national super computer centers. The vast middle ground of eScience problems is well served by the modern successor to the Beowulf cluster. These come in many flavors, from a few PCs on a shelf connected via 100Mbs ethernet, to large deployments of blades sharing low-latency interconnects such as Infiniband.

In this section we'll share approaches for the deployment of your own HPC clusters, with an emphasis on the hands-on details: How to place an order with a vendor so the nodes automatically PXE boot so you can avoid tedious hands-on configuration of every node; Why choose netbooting? Why choose ROCKS? When to pay the extra cost of Infiniband and when you can avoid it. Look in the tools section for details on schedulers and systems management.

#### Networking

You're producing heaps of data, either from instruments and sensors or from simulations. In any case, you now have to move it from here to there. This section encompasses problems in bandwidth-limited applications (NFS, etc. to hundreds or thousands of nodes), latency-limited applications (fine-grained parallel codes), and surprising combinations (bandwidth delay product in long-haul transfers).

### Resources

---

Where to go for information and help about eScience Resources at UW and beyond. We can't do this alone, and, thankfully, we don't have to. [UW Technologies](#) provides an excellent infrastructure in data-center, networking, and data protection services. And, the new eScience Institute is here to help you navigate the maze of options. Among the resources we'll provide pointers to are:

#### **Proposal Preparation**

The eScience Institute can help you with the preparation of the technical sections of your funding proposals. Everything from helping you evaluate your computational requirements to assisting you in securing budgetary quotes from vendors.

#### **Data-center and Collocation**

Now your proposal is funded and you want to buy that big cluster. Where are you going to put it? We can help you avoid this problem by starting the process of data-center planing while your cluster is still in the planing stages. The more UW Technologies knows about your plans, the better they can accommodate your needs.

#### **UW Networking**

Are you troubled by frustratingly slow data transfers between UW and off-campus sites? Are you planning a multi-site deployment here on the Seattle campus? You're not the first and you're not alone. We can help connect you with the appropriate resources to solve these problems and many others.

#### **UW Storage Services**

Need to back up 100TB of data? Need to preserve your data for years, while only keeping the latest year on local disks? Need to ensure disaster recovery if Seattle suffers an 8.0 quake? Solutions are at hand and we can help you find them.

#### **Remote Compute Resources**

Whether it's a TeraGRID allocation or an introduction to Amazon's EC2 service, we can point you in the right direction.

### Methods

---

The advance of eScience depends on the improvement of existing computational methods as well as the development of new approaches. The following areas have been identified as key to the success of the eScience endeavor.

The Athena project in the institute for nuclear theory, and the physics and astronomy departments has recently hired a scientific programming consultant to assist domain scientists in the development of eScience methods for the Institute of Nuclear Theory and the Astronomy and Physics departments.

#### Sensors & Sensor Networks

Large numbers of tiny but powerful sensors are being deployed to gather data on the sea floor, in the forest canopy, in gene sequencers, in buildings and bridges, in living organisms. These approaches share a common trait: they produce enormous amounts of data that must be captured, transported, stored, organized, accessed, mined, visualized, and interpreted in order to extract knowledge. This “computational knowledge extraction” lies at the heart of 21st century discovery.

#### Machine Learning & Data Mining

Once acquired, knowledge must be extracted from data from simulations and sensor networks. Traditional methods requiring focussed human attention at every step of the process simply do not scale to petabyte data sets. The NSF has recognized the importance of technologies relevant to these problems with their Cyber-Enabled Discovery and Innovation (CDI) program.

#### Visualization

Automated pattern recognition and knowledge extraction from immense data-sets are essential, but enabling the unparalleled human capacity for pattern recognition is equally important. Next-generation visualization methods must be developed which provide productive systems for navigating systems with fantastic resolution and unheard of dynamic range. These methods encompass hardware platforms, software tools, and the interface between them.

#### Multicore and Parallel Techniques

Immense quantities of data require immense computational power for analysis. Parallel data analysis tools are, generally, still in their infancy and must be advanced to address eScience problems. Leveraging recent developments in microprocessor and networking technologies is key to achieving these goals.

## Tools

---

Implementing eScience technologies and methods requires a sophisticated toolkit. The eScience Institute and this SIG are here to help you find the right tool for your job.

### Developer Tools

These include compilers, debuggers, and profilers. We'll provide links to the best-of-breed and advocate for volume pricing or site licenses for commercial products to meet the demand.

### Data Transfer and Management Tools

The seeming simple task of moving your data from a super computer center back to your lab for analysis can be a frustrating experience. We'll provide assistance with Grid tools, such as GridFTP, and hpssh. Aside from the transfer problem, there are huge issues associated with the management of these multi-TB data sets. Look here for help with archiving and data organization tools.

### Systems Management Tools

Deploying and managing parallel compute clusters and large-scale data storage systems can be daunting. Fortunately, many of the thorniest problems have well-established solutions. We can help you navigate the maze of acronyms from ROCKS, through IPMI, through GPFS.

## The Science

---

Below is a *small* sample of the eScience planned or already being conducted at UW. As this SIG grows this will become your place for publicizing your work within the UW eScience community, discovering others engaged in eScience, and building collaborations to leverage our local eScience expertise.

### **Astronomy:**

UW is a participant in the Large Synoptic Survey Telescope (LSST). When this instrument begins science operations in 2015 it will collect 30TB of data every night, resulting in a data set of 150PB.

[http://www.lsst.org/lsst\\_home.shtml](http://www.lsst.org/lsst_home.shtml)

<http://www.astro.washington.edu/becker/LSST/>

The astrophysics theory group is the home of some of the most advanced astrophysics simulation code anywhere. Running on some of the largest supercomputers in the world, it produces detailed pictures of astrophysics phenomena from the scale of asteroids through the large scale structure of the universe

<http://www-hpcc.astro.washington.edu/>

### **Biochemistry:**

The Baker and Schief labs are at the forefront of protein structure prediction. By mid 2008, the combined groups will operate a local compute resource of more than 2,700 CPU cores and more than 400TB of storage. Additionally, the Rosetta@home project currently has roughly 150,000 contributing CPUs. The lab's research spans problems from protein structure prediction to HIV vaccine design.

<http://depts.washington.edu/bakerpg/>

<http://boinc.bakerlab.org/rosetta/>

### **Bioengineering:**

The Daggett Lab is engaged in the realistic simulation of protein dynamics, unfolding/folding, conformational transitions linked to disease, and the design of biomaterials. Locally, the lab operates substantial computational resources, while the National Labs provide tens of millions of hours of additional CPU time annually.

<http://depts.washington.edu/daglab/>

### **JISAO:**

The Joint Institute for the Study of the Atmosphere and Ocean operate at the forefront of climate change research, including studies of the large scale variability and predictability of the coupled atmosphere-ocean system with special emphasis on those processes of most relevance to the Pacific Northwest. As climate models evolve, the volume of data associated with these efforts could exceed 100PB

<http://jisao.washington.edu/>

### **Medicinal Chemistry:**

The Goodlett lab, among other things, is developing computational methods for the analysis of mass spectrometry data. Once developed the technology will be applied to better understand primary brain tumors which are the leading cause of cancer-related death in children.

<http://goodlett.proteomics.washington.edu/>

### **Oceanography:**

Studies in metagenomics and physical oceanography test the limits of computational science at both the observational and simulation boundaries. Analyzing data from a cabled network of underwater sensors, teasing apart the ecology of marine ecosystems from genome data of whole populations of microorganisms, and the ever increasing resolution of simulations in physical oceanography will require unprecedented quantities of computational resources and data storage.

<http://www.ocean.washington.edu/2004/index.html>

<http://armbrustlab.ocean.washington.edu/>