Advanced Information Technology Support for Life Sciences Research

Craig A. Stewart[1] stewart@indiana.edu
David Hart[2] dhart@indiana.edu
Richard Repasky[2] rrepasky@Indiana.edu
Mary Papakhian[2] mpapakhi@indiana.edu
Anurag Shankar[2] ashankar@indiana.edu
Eric Wernert[2] ewernert@indiana.edu
Andrew D. Arenson[3] aarenson@iupui.edu
Gerry Bernbom[2] bernbom@indiana.edu

ABSTRACT
The revolution in life sciences research brought about by the sequencing of the human genome creates new challenges for scientists and new opportunities for computing support organizations. This may involve significant shifts in computing support strategies, particularly as regards interacting with life sciences researchers who maintain a medical practice. This paper describes Indiana University's experience in a large-scale initiative in supporting life sciences research, as well as several strategies and suggestions relevant to colleges and universities of any size. Computing organizations and support professionals have many opportunities to facilitate and accelerate life sciences research.

Categories and Subject Descriptors
J.3 Life and Medical Sciences
K.6 Management of Computing and Information Systems

General Terms
Documentation, human factors, management

Keywords
Life sciences, computational biology, bioinformatics

1. INTRODUCTION
Indiana University announced in 2000 the Indiana Genomics Initiative [1], a major effort to enable new treatments for human disease and to improve human health. This initiative involved a significant information technology component. The purposes of this paper are: to describe the steps taken by Indiana University's University Information Technology Services (UITS) in support of life sciences research; to identify key lessons learned; and to suggest strategies that colleges and universities of any size may employ to enhance life sciences research.

This paper first describes some of the hardware facilities implemented by Indiana University (IU) in support of life sciences research. Systems without adequate support are doomed to be used poorly [2]; therefore, this paper also describes strategies and some new applications developed to support life sciences research.

As a result of these efforts, IU medical researchers' use of advanced IT facilities has increased one hundredfold during the past two years. During May 2003, researchers from the IUPUI campus (Indiana University – Purdue University Indianapolis, which includes the IU Medical School) for the first time used more CPU hours than researchers from IU's Bloomington campus (IUB). This is noteworthy because most of the researchers in fields that traditionally make heavy use of supercomputers (*e.g.* physics, astronomy, and chemistry) are located at IUB. Results of the annual UITS survey of computer user satisfaction [3] also indicate progress. One question on this survey addresses overall satisfaction with UITS's research computing services. Respondents from the IU School of Medicine provided an average rating of 3.9 on a 1-5 Likert scale (with 5 = excellent); 97.5% rated services as at least satisfactory (3 or higher).

Copyright is held by the author/owner(s).
SIGUCCS'03, September 21–24, 2003, San Antonio, Texas, USA.
ACM 1-58113-665-X/03/0009.

2. DISCUSSION
2.1 Advanced IT facilities for life sciences
Many life sciences computing challenges require massive computational power and storage facilities. Indiana University upgraded its IBM SP supercomputer to a peak theoretical processing capability of 1 TFLOPS (one trillion mathematical operations per second) in large part to support life sciences research, and IU maintains other large HPC systems as well [4]. IU's computational systems are, overall, fairly centralized, and in particular there has not been a great proliferation of small Linux clusters on campus. This has led to greater value being received from equipment purchased by the University than might otherwise have been the case. In particular, computers in individual labs may sometimes be idle. Centralization of computing systems provides economies of scale and opportunities to increase levels of use for any university or college. Equipment can be centralized in a fashion that provides at-will access to equipment purchased by a particular research group, while making such resources available for general use when idle [5].

Many computational resources are also available via the World Wide Web. Many Web sites dedicated to the delivery of biomedical data provide significant computational capacity available without cost [*e.g.* 6]. Such facilities may provide significant help for smaller institutions in meeting the computational needs of biomedical researchers. For those researchers who require large-scale computing resources, allocations on national supercomputer centers are available without cost through a peer-review application process [7]. Furthermore, much of the software widely used in academic research is available as open source, and scripts that facilitate installation of several such packages are available [8]. As a result, given some investment in knowledgeable support staff, it is possible to deliver useful services in support of life sciences research at many levels of investment in computing systems.

Life sciences research involves significant data storage challenges. IU manages a massive data storage system containing more than one terabyte of life sciences data [9]. An important aspect of IU's massive data storage system is that it is geographically distributed, including one robotic tape silo in Indianapolis and another in Bloomington – roughly 50 miles apart. Data is generally copied to both locations

simultaneously, ensuring that data will be preserved even in the event of a catastrophe destroying one of the machine rooms. This is particularly important in the case of life sciences data, which is essentially irreplaceable. It is possible to collect multiple data sets that are similar, but each is as unique as the individual creatures involved; a data set once lost is lost forever. Reasonably low-cost solutions exist for moderate-scale data storage problems, including off-site vaulting options that provide secure (albeit slow) data storage and recovery.

To support visualization of biomedical data, IU has installed several immersive, 3D, virtual reality visualization environments. Relatively low-cost 3D display systems [10] can be installed in life sciences research labs, providing an opportunity for impromptu analyses of life sciences data at the convenience of the researcher.

There are, then, a wide variety of options for meeting the computing, storage, and data visualization needs of biomedical researchers, spanning a range of scales and costs. However, all of these options require staff dedicated to supporting their use in order to successfully fill these needs.

2.2 Support for life sciences researchers

The same general strategies previously described [2] for supporting high performance computing in academic research also work for life scientists, particularly provision of easy access to computing systems and consulting and programming support. One of the lessons learned early on was to adapt these support strategies to match the needs and work practices of life sciences researchers, especially medical researchers, who often stay active in clinical practice. This imposes stress and time constraints on their schedules rarely faced by researchers in other disciplines. Consequently, extra efforts are required to make advanced IT infrastructure accessible and valuable to these scientists.

There is a "chicken and egg" problem regarding advanced IT facilities for life sciences. The life sciences do not have a long tradition of use of large-scale computing systems, as exists, for example, in physics and astronomy. A life scientist persuaded to try using advanced (and sometimes hard-to-use) information technology facilities will only continue to use these resources if they promise to increase his productivity. From the standpoint of the academic computing center, it may be difficult to justify the staff effort required to install and maintain a suite of software and data resources while waiting for life sciences researchers to begin making heavy use of such resources. It seems clear, however, that progress will be most rapid if the computing center takes the first step.

One critical step in improving the perceived accessibility of Indiana University's computing systems was the creation of a suite of batch scripts for the software packages most often used by life scientists [11]. These scripts are designed to do what the life sciences researcher is most likely to want done and report to the user exactly what was done as part of the analysis. Determining the most appropriate behavior was the largest part of the work in creating these scripts. They are available for download and only minor modifications should be required to adapt them for use at other institutions.

A second critical step in improving the accessibility of these facilities was the creation of a hardcopy document describing their use and how they could enhance the research programs of life sciences researchers [12]. This document was widely distributed within the IU School of Medicine, and has proven to be very useful – largely because the hardcopy format made it easy for researchers to carry it around to read during spare moments, or refer to as they worked at their terminals. Also, a Web page specifically designed to help life sciences researchers navigate easily to the most critical Web-based information resources was created [13].

Finally, an aggressive outreach campaign followed, focused on visits to researchers' labs and offices to ask them to describe their research activities, so that ways in which the university's advanced IT facilities could be put to use to the benefit of their research programs might become apparent.

2.3 Example software implementations

Indiana University's software support strategies have extended beyond offering assistance with existing software to include optimization of software and creation of new software applications. Some software implementations have required significant effort, while others have been very straightforward. Some examples follow.

Supercomputers enable many processors to work together to solve very large computational problems. They also enable many independent problems of more modest scale to be solved simultaneously. Programming support, involving the tuning and/or parallelization of existing software or the creation of entirely new software, has in some cases dramatically reduced the amount of time required for researchers to perform important analyses – reducing processing time from days to hours or from hours to minutes. The capabilities of one researcher were enhanced substantially by the addition of a particular piece of software to the IBM SP supercomputer which enabled processing large numbers of data analyses concurrently. The researcher had previously been limited to processing just a few data sets at a time, one per PC in the researcher's lab. Using a supercomputer to process larger numbers of data sets simultaneously was simple to implement and very valuable to the researcher. Use of supercomputers both accelerates research and helps facilitate the transformation of experimental analyses into clinical practice.

Considerations related to technology transfer may be complicated when dealing with life scientists. New and improved treatments are often put into use through commercialization, and this requires clarity in ownership of intellectual property. UITS employs written agreements defining the terms of extended consulting projects. When a UITS programmer creates a fundamentally new piece of software, UITS files an invention disclosure with Indiana University's technology transfer office. A CVS repository is maintained, in part because this is good programming practice and in part to help clarify ownership of intellectual property. No software has at this time been commercialized, but several life sciences applications have been released under the terms of LGPL or BSD licenses [14]. These licenses permit software to be exchanged as open source and still permit commercialization – important because economic development is among the objectives of the Indiana Genomics Initiative [1].

Considerable effort has also been expended to address data management and visualization issues, which are particularly important given the burgeoning data stores now being made available to life sciences researchers. Indiana University has created a Centralized Life Sciences Data (CLSD) service, which provides a single, SQL-based interface for querying a variety of public biomedical databases and databases created within Indiana University [15]. UITS has also created several visualization applications for biomedical research [16-18].

There are a number of roles that computing organizations and professionals are uniquely positioned to fulfill that will lead to the enhancement of research in the life sciences and, in turn, better health for humans overall. As the life sciences become more and more dependent upon advanced information technology, proactive assistance from computing support organizations and professionals can dramatically accelerate life sciences research. Indiana University has used the strategies described here with significant success; the senior author (CAS) is currently visiting a supercomputing center in Germany [19] to assist with implementation of several of these strategies there.

3. CONCLUSIONS

The support strategies identified here, particularly as regards interactions with life scientists, may be used successfully across a wide range of universities and colleges. There are likewise options available for meeting computational and storage needs that range in cost from free to modest. At the same time, some of the services Indiana University is able to provide are in part a result of significant investment and grant support. The possibility of providing such services is good grounds for seeking additional external or internal funding. There is at present an opportunity for computing support organizations and professionals to greatly accelerate life sciences research. Such progress will be most rapid if computing organizations and support professionals take the first steps in establishing service suites, and reach out actively searching for collaborations with life scientists.

4. ACKNOWLEGEMENTS

This paper is dedicated to Dr. Christopher S. Peebles on the occasion of his retirement as Associate Vice President for Research and Academic Computing. Funding support has been provided via the Indiana Genomics Initiative, which is funded in part by the Lilly Endowment, Inc. This material is based upon work supported by the National Science Foundation under Grant No. 0116050 and grant CDA-9601632. IU acknowledges multiple Shared University Research grants awarded by IBM, Inc., and IU's relationship with IBM as an IBM Life Sciences Institute of Innovation. IU acknowledges its relationship with Sun Microsystems, Inc., as a Center of Excellence.

5. REFERENCES

[1] The Indiana Genomics Initiative. http://www.ingen.iu.edu/

[2] Stewart, C.A., C.S. Peebles, M. Papakhian, J. Samuel, D. Hart, Stephen Simms. 2001. High Performance Computing: Delivering Valuable and Valued Services at Colleges and Universities. Proceedings of SIGUCCS, Portland, OR. http://www.indiana.edu/~rac/stewart_siguccs.pdf

[3] UITS User Survey. http://www.indiana.edu/~uitssur/

[4] Research and Technical Services http://www.indiana.edu/~rats/

[5] Agnihorti, P., V.K. Agarwala, J.J. Nucciarone, K.M. Morooney, C. Das. 1998. The Penn State computing condominium scheduling system. Proc. ACM/IEEE conference on Supercomputing, San Jose, CA.

[6] National Center for Biotechnology Information. 2003. http://www.ncbi.nlm.nih.gov/

[7] National Computational Science Alliance Allocations. 2003. http://www.ncsa.uiuc.edu/UserInfo/Allocations/

[8] Biolinux. http://www.biolinux.org/soft.html

[9] Shankar, A. Research Data Storage Services at IU, in Proceedings of the I-Light Applications Workshop 2002, http://www.iupui.edu/~ilight/proceedings/presentations/ilight_wrkshp02_panel2_shankar.pdf

[10] Technology - Affordable, commodity-based systems for virtual reality & visualization, http://www.avl.iu.edu/technology/affordable/

[11] UITS Bioinformatics Support, http://www.indiana.edu/~rac/bioinformatics/

[12] UITS, 2002. INGEN's advanced IT facilities: the least you need to know. Indiana University, Bloomington, IN. 52 pp.

[13] A Field Guide to UITS Computing Services for the IU School of Medicine. http://www.indiana.edu/~rac/INGEN/fieldguide.html

[14] Open Source Initiative OSI – Licensing, http://www.opensource.org/licenses/

[15] Guide to CLSD, http://www.indiana.edu/~rac/clsd/

[16] AVL Projects - 3DIVE, http://www.avl.iu.edu/projects/3DIVE/index.shtml

[17] AVL Projects - ToothPics, http://www.avl.iu.edu/projects/ToothPics/

[18] AVL Projects - FAS/3D Scanner, http://www.avl.iu.edu/projects/FAS/

[19] High Performance Computing Center Stuttgart. http://www.hlrs.de/

6. ADDRESS INFORMATION

[1] Indiana University, Bloomington, IN & Höchstleistungsrechenzentrum Universität Stuttgart

[2] Indiana University, Bloomington, IN

[3] Indiana University – Purdue University Indianapolis, Indianapolis, IN