2003 Report on Indiana University Accomplishments supported by Shared University Research Grants from IBM, Inc.

Table of Contents

1. Executive Summary
2. Introduction
3. Creation of a distributed high performance computational and storage system
3.1. Distributed SP
3.2. AVIDD - Analysis and Visualization of Instrument-Driven Data
3.3. Massive Data Storage System (MDSS)
3.4. Joint research and development efforts related to high performance and distributed computing
4. Development and implementation of advanced software for the life sciences
4.1. Implementation of new IBM software technology for biomedical data management - the Centralized Life Sciences Data (CLSD) service.
4.2. IBM software
4.3. Creation of new biomedical software tools
4.3.1. Computational phylogenetics
4.3.2. Computational chemistry
4.3.3. Protein data management - the Protein Family Annotator
4.3.4. Radiation oncology - Penelope
4.3.5. Cell modeling
4.3.6. Biomedical text retrieval
5. Computer Science
5.1. The Xport project
5.2. Grid computing with Indiana University and Purdue University
6. Education and Outreach
7. Value to IBM
8. Conclusion
9. References

1. Executive Summary

Indiana University and IBM, Inc. have a very strong history of collaborative research, aided significantly by Shared University Research (SUR) grants from IBM to Indiana University. The purpose of this document is to review progress against recent SUR grants to Indiana University. These grants focus on the joint interests of IBM, Inc. and Indiana University in the areas of deep computing, grid computing, and especially computing for the life sciences. SUR funding and significant funding from other sources, including a $1.8M grant from the NSF and a portion of a $105M grant to Indiana University to create the Indiana Genomics Initiative, have enabled Indiana University to achieve a suite of accomplishments that exceed the ambitious goals set out in these recent SUR grants.

Some of the key accomplishments thus far related to SUR supported activities are the following:

- Creation of a tera-scale distributed computing environment based on IBM hardware and software technology. Thanks in part to SUR support from IBM, Indiana University became the first US university to own a 1 TFLOPS supercomputer system. IU's geographically distributed high performance computing and storage system now includes

a 1 TFLOPS SP, a 1.1 TFLOPS Linux cluster (based on IBM servers and software), and an HPSS installation with 4.5 TBs of disk cache and 500 TBs of tape storage. This provides an excellent testbed and demonstration of grid computing technology. Particularly important is the implementation by IU of novel features that enable IU's tera-scale computing infrastructure to manage and analyze very large data sets. Peter Freeman, Assistant Director of the NSF, has referred to IU's computing environment as an example of the type of cyberinfrastructure that the NSF sees as the future of high performance computing in the US.

- Creation and implementation of new software for the life sciences. IU, in part through efforts supported in part by SUR grants, and in part through Joint Study Agreements with IBM, has created new software for the life sciences and has been an early adopter and tester of new IBM software for the life sciences. Particular activities of note include:
    - Creation of open source software (available as source code and precompiled for AIX and Linux) for computational phylogenetics, simulation of biochemical systems, radiation therapy for cancer, cell modeling, and distributed management of proteomics data.
    - Early implementation of new IBM software. IU has created a centralized life science data retrieval facility for use in biomedical research, based on use of DiscoveryLink and DB2 Information Integrator.
- Indiana University has assiduously and extensively publicized its collaborative work with IBM, Inc. in a number of venues. Indiana University has served as a reference site for IBM technology both as part of press releases and through meetings with other IBM customers. IU has extensively publicized its activities with IBM in the popular and technical press as well. Furthermore, Indiana University has presented the results of collaborative and IBM-supported research in a wide variety of technical venues. As a result, there is broad awareness throughout the high performance and life sciences community about the work IBM and IU have done together.
- Indiana University has also served as a source of recruiting for IBM, Inc. and companies implementing IBM technology. Indiana University's efforts in education about high performance computing, including educational opportunities focused for students from traditionally underrepresented groups, have enhanced the pool of students with recent university degrees that have experience with IBM high performance computing technology.

IBM has delivered significant support to Indiana University through Shared University Research grants. Indiana University has in turn delivered significant value to IBM and to the broader research community, especially in the life sciences. The relationship between IU and IBM continues to grow stronger and more valuable to both parties, and we look forward to the continued strengthening and deepening of this relationship.

2. Introduction

IBM, Inc. and Indiana University have a very strong history of collaborative research aided significantly by Shared University Research (SUR) grants from IBM to Indiana University. Previous grants and accomplishment reports are available online [1]. The purpose of this document is to review progress against the three recent Shared University Research grants to Indiana University:

- "Tightly integrated distributed supercomputing - the Indiana TeraCloud" (2000)
- "Data-intensive and Distributed Computing in the Life Sciences" (2001)

- "Grid and Data Intensive Computing in the Life Sciences" (2002)

All three proposals focus on the commonalities in thinking between IBM and IU: a commitment to new discovery through high performance computing using approaches labeled "deep computing" by IBM; a commitment to advancing the state of the art in distributed and grid computing; and a commitment to advancing the state of the art in life sciences computing (while in the process creating new, open source community codes that enable new life sciences research).

The success of research supported by SUR grants and other funds (most notably the $105M grant from the Lilly Endowment, Inc., to create the Indiana Genomics Initiative [2]) has led to a string of successes and contributions to science that exceed the ambitious goals set out in the grant proposals. These accomplishments are summarized by project below, followed by a brief assessment of the impact of the grant-supported research in the press and on training and recruitment.

3. Creation of a distributed high performance computational and storage system

Through its recent SUR applications, Indiana University has proposed to advance the state of the art in computational grids by creating a distributed high performance computation and storage system at Indiana University. We have indeed done that. In fact, the present computing, storage, and visualization environment at Indiana University matches to a great extent the vision set forth in the recent Cyberinfrastructure report [3], and this is achieved largely through use of IBM technology. This widely acclaimed high performance distributed environment was implemented with significant assistance from and interaction with IBM. The key components of the distributed computing system are as follows:

- An IBM RS/6000 SP with an aggregate computational capacity of 1.005 TFLOPS. This system comprises two geographically distributed components, one located on IU's campus in Indianapolis, and one on IU's campus in Bloomington.
- A distributed Linux cluster with an aggregate computational capacity of 1.1 TFLOPS. This system comprises three geographically distributed components, located at IU's campuses in Indianapolis, Bloomington, and Gary.
- Special purpose computation. Two GRAPE-6 boards (0.5 TFLOPS each) and two MD-GRAPE-2 boards (64 GFLOPS each) provide an aggregate peak theoretical capacity of 1.28 TFLOPS for special-purpose computing [4].
- Storage. A distributed massive data storage system, using HPSS software, with tape silos (500 TB) and disk caches (4.5 TB) located in Indianapolis and Bloomington.
- Networking. The Bloomington and Indianapolis components of the system are connected by an ultra-high speed network running over the I-light network [5]. Using dedicated fiber stretching some fifty miles and high-end Juniper switch technology, the university-owned I-Light network has an aggregate bandwidth of 2 Gigabits per second between the two campuses (this number is slated to go up to 10 Gigabits per second within a year).

3.1. Distributed SP

Indiana University's distributed IBM SP is one of very few SP installations in the world in which two logical SPs are united within a single Loadleveler instance. The upgrade to Indiana University's IBM SP, announced in 2001 and made possible in part by a SUR grant from IBM, created at Indiana University the first 1 TFLOPS supercomputing system owned by a US university. IBM received considerable note in the technical and popular press as a result [6].

3.2. AVIDD - Analysis and Visualization of Instrument-Driven Data

Indiana University recently dedicated the AVIDD (Analysis and Visualization of Instrument-Driven Data) facility [7, 8]. The core of AVIDD is a distributed Linux cluster consisting of:

- Two identical IA32-based clusters, one located at IU Bloomington and one located at IUPUI. Each includes 208 2.4 GHz Prestonia processors.
- One IA64-based cluster, containing 36 1.0 GHz McKinley processors.
- One smaller, IA32-based cluster located at IU Northwest in Gary, containing 18 1.3 GHz PIII processors. This cluster is for instructional use at the IU Northwest campus. (This cluster was funded entirely via IU's 2002 SUR grant. It is playing a particularly important role in the educational aspects of the AVIDD cluster, as it is located on a campus with a high percentage of students from traditionally underrepresented groups.)

The aggregate capacity of the AVIDD system will be 1.1 TFLOPS, 0.5 TB RAM, and 10 TB disk space. There are three aspects of the AVIDD facility that are particularly forward-looking in meeting the coming needs of scientists:

- Managing very large data sets. With a total of 10 TB of spinning disk, and close integration with IU's massive data storage system, scientists will easily be able to manage and analyze multi-TB data sets using the AVIDD facility. Thanks to close integration with IU's Massive Data Storage System, it is possible to move one TB of data between the AVIDD Linux clusters and the MDSS in about 2.5 hours.
- Real-time data analysis. AVIDD provides facilities for analysis of nonscheduled, real-time data streams while also maintaining high overall usage levels of the computational systems. AVIDD has the capability to respond to requests for analysis of incoming data streams in real time by preempting analysis tasks that are not time-sensitive in order. This is accomplished through a mix of kernel-level checkpointing (with LAM/MPI [9]), application level checkpointing, and new features being added to the Maui scheduler [10] at IU's request (IU is a charter member of the Maui Consortium).
- Low-cost, distributed visualization environments. IU has developed three relatively low-cost 3D visualization environments, permitting the installation of immersive 3D visualization devices in several labs and shared research areas. Placing immersive 3D visualization environments in or next to research labs will enable scientists to interact with and analyze their data in ways far more valuable than is typically possible today.

Peter Freeman, Assistant Director of the NSF, has referred to IU's computing environment as an example of the type of cyberinfrastructure that the NSF sees as the future of high performance computing in the US [8].

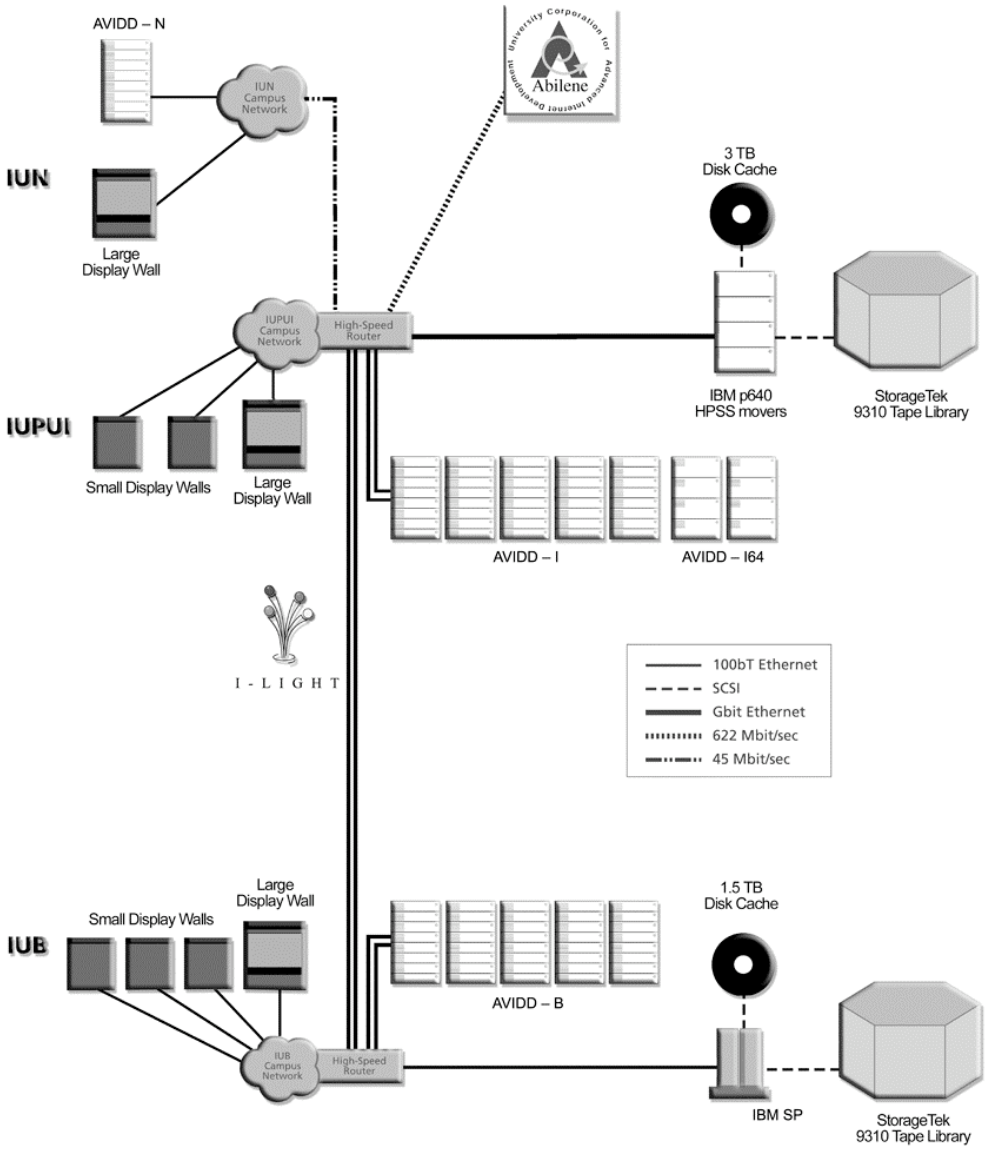

Figure 1. Schematic diagram of IU's high performance computing and massive data storage systems.

3.3. Massive Data Storage System (MDSS)

The IU Massive Data Storage System (MDSS) is based on use of HPSS software, IBM movers and tape drives, and IBM and STK tape silos. The MDSS includes an aggregate of 500 TB of tape storage and 4.5 TB of disk cache. IU currently has a total of 100 TB of data stored, with more than 1 TB of this being biomedical data. IU has more data in storage under control of a massive data storage system than any other university, and we are the only university we are aware of that is currently storing more than 1TB of biomedical data.

IU was the first HPSS installation to implement a distributed HPSS mover, and as a result IU's massive data storage system has become one of the best, and perhaps the best, example of a fault-tolerant distributed massive data storage system [11, 12]. The MDSS includes two tape robots separated by a distance of 50 miles. Each night new data added to the system in one location are copied to the other location, assuring that even in the event of a disaster that might destroy one of

the tape silos, no data would be lost. This is particularly important in the life sciences, as biomedical data is essentially irreplaceable.

HPSS is traditionally used by scientists in physics and astronomy. IU serves not only this traditional, high-end group of scientists, but also offer terabytes of storage to users in diverse and heretofore unrepresented disciplines such as apparel design, business, economics, history, fine arts, music, radiology, recreational sports, theater and drama, and many others. UITS is able to serve this vast array of users because we have made storage easily available, using a secure, Web-based user interface, and by allowing personal computer users to access mass storage directly from an icon on their computer desktops.

IU has invested considerable effort in developing facilities for high-speed transfer between HPSS and GPFS. IU and NERSC put forth a proposal to the HPSS consortium to incorporate a native GPFS interface to HPSS, developing a considerable amount of code in the process. Since the HPSS Consortium has not yet acted upon this request, IU has focused on high-speed transfer via pftp using striped reads and writes, to and from tape and disk. In the process IU has revealed and helped correct significant performance problems in HPSS. As a result of IU's efforts on this matter, it will become possible to deliver sustained 100 Mbit/sec transfer between HPSS and GPFS (a many-fold improvement over the performance of the HPSS code as released).

3.4. Joint research and development efforts related to high performance and distributed computing

IU has worked closely with IBM in the development of a great deal of essential system software. IU participated deeply in the integration of the IBM SP as a single computational resource distributed over a high-speed optical network. IU has contributed substantially to the HPSS code base, particularly in developing the software that enabled the use of geographically distributed HPSS movers. And IU has participated strongly in the creation of system software that has enabled the AVIDD cluster to have many of its cutting edge features. IU's contribution to the development of scheduling and pre-emption capabilities, directly and through the Maui scheduler, is particularly important. IU has also been very active in identifying software problems and working with IBM to correct them. Performance problems in striped reads from HPSS and in GPFS for Linux are just two recent examples of many circumstances in which IU has worked with IBM to enable IBM to correct limitations that we have discovered in IBM-produced software.

4. Development and implementation of advanced software for the life sciences

No area of science has been more radically and rapidly transformed into a massively data-centric science than the life sciences. The human genome has been completely sequenced. The information technology infrastructure and applications that permitted the sequencing of the human genome have received great accolades. However, the most critical problems that lie between our current understanding of genomics and the ability to turn genomic data into understanding and new therapeutic agents are of much greater scale and complexity than sequencing the human genome.

In recent SUR proposals, Indiana University proposed to create applications, frameworks, and solutions for data management, analysis, and visualization for specific biomedical science challenges in the areas of biomedical imaging, cellular modeling, biomedical data retrieval, and computational phylogenetics. We have done that in ways detailed below.

4.1. Implementation of new IBM software technology for biomedical data management - the Centralized Life Sciences Data (CLSD) service.

Indiana University has implemented a distributed database access system called the Centralized Life Sciences Data (CSLD) service, based on use of IBM's DiscoveryLink and DB2 Information Integrator. The CLSD service provides a single, SQL-based interface to a selection of widely used biomedical datasets, including BIND [13], ENZYME [14], LIGAND [15], LocusLink [16], UniGene [17], dbSNP [18], SGD [19], KEGG PATHWAY [20] and a variety of NCBI BLAST databases [21]. CLSD runs on Indiana University's IBM SP supercomputer and Sun E10000 supercomputer, using IBM DB2 database software and IBM's DiscoveryLink software [22] for retrieving data from multiple, heterogeneous data sources.

CLSD is already used in labs within the IU School of Medicine, and the combination of CLSD and DB2 Information Integrator is the core technology IU is using to achieve an elegant yet challenging goal. Specifically, our goal for data accessibility within the IU School of Medicine is that any researcher should be able to query all relevant external data sources and all internal data sources to which the researcher has access permission, from the researcher's desktop in a completely transparent fashion, without needing to know where the data actually reside. The DB2 Information Integrator core of CLSD uses software programs called parsers to transform datasets from their native format into relational databases. Data files are copied across the Internet from their original sources to supercomputers at Indiana University. These files are converted to relational database format using a parser that is specific to each particular data source. The resulting relational database files are then imported into IBM's DB2 relational database management system. This allows, for example, the joining of data from different databases into a single record of information presented to the researcher. CLSD also incorporates BLAST datasets, permitting a researcher to run a BLAST job as if executing a database query.

Clinical and research labs can also make their data available to other IU investigators via CLSD. CLSD and its underlying database products enable clinicians and researchers to provide data both securely and with control over access to the data being provided. Data may be contributed via a wide variety of relational and non-relational formats, including DB2, Oracle, Sybase, SQL Server, Informix, flat files, Excel, and XML. This system allows a quick, simple, and secure means for sharing data while assuring that all researchers with rights to access the data are making use of the most current versions of those data. IU researchers wrote the CLSD system, user interface, and several of the underlying "parser" programs. IU researchers in particular wrote the following parser programs: BIND, ENZYME, KEGG PATHWAY, LIGAND, and UniGENE. Indiana University and IBM are in the process of executing licensing agreements to provide licenses to IBM for the IU-written parsers.

4.2. IBM software

IU runs a variety of IBM software on our tera-scale environments as a critical part of our supercomputing efforts (especially in the life-sciences). A sampling of IBM software currently in use includes Data Explorer, DB2, Teiresias, IBM programming tools, and software environments.

4.3. Creation of new biomedical software tools

IU has been extremely active in creating open source tools for biomedical computing, making these tools available under open source licensing and as precompiled binaries for AIX. A particular feature of our licensing strategy is that, whenever possible, the terms of the "Lesser GNU" license are used [23]. These licensing terms permit the subsequent commercialization of derivative software products, leaving open the possibility of commercialization of these tools by either IBM or IU. A brief description of the software tools made available by IU is provided below; a list and links to download sites is available online [24].

4.3.1. Computational phylogenetics

fastDNAml [25] has become a very popular software tool for inference of evolutionary trees from genetic sequence data. Representatives of IU recently presented a paper about our work with fastDNAml with IU's IBM SP at the SC2001 conference [26]. IU is involved in ongoing improvement of this widely used software [27]. At present IU is testing a portal interface to fastDNAml, as well as a grid implementation of fastDNAml.

4.3.2. Computational chemistry

Dr. Glenn Martyna (formerly of IU and now working in IBM's research labs) and his collaborators have developed a computer package called PINY_MD [28] for the simulation of chemical reactions in the condensed phase for large and complex systems. IU has ported this code to AIX and it is freely available for download [29].

4.3.3. Protein data management - the Protein Family Annotator

IBM and IU have executed a Joint Study Agreement to create a distributed software program for managing information about families of proteins, called the Protein Family Annotator (PFA). Protein Families are groups of related proteins, and a great deal of information can be learned about one protein by accessing information about related proteins. Unfortunately, this information is today scattered and often in disparate formats. The PFA will create a graphical interface to a distributed database system that will permit simple storage and retrieval of data about protein families. The PFA has the particular advantage of letting a community of biomedical scientists contribute data to and retrieve data from a distributed grid of protein data. It also permits scientists to specify which information is to be shared, and which is of proprietary value and not to be shared. Dr. Mehmet Dalkilic of the IU School of Informatics is leading this project. The lead programmer for this project is funded half by IBM Life Sciences and half by IU. A beta version of the PFA will be available for use during the summer of 2003.

4.3.4. Radiation oncology - Penelope

Indiana University has enhanced and optimized the Penelope radiation transport code [30] as part of an effort to improve the efficacy of radiation treatment for cancer using the Gamma Knife [31]. The Gamma Knife is an extremely advanced system for very precise delivery of radiation to treat tumors that cannot be removed surgically. Famed bicycle racer Lance Armstrong was treated for his cancer with the Gamma Knife at the IU School of Medicine. The major challenge in use of the Gamma Knife today is in targeting. A generic model of the human head is used when planning targeting. To the extent that shape or tissue densities of an individual patient vary from the generalized model, there is some chance that inaccuracies in targeting will result. IU's efforts with the Penelope code [32, 33] should make it feasible to plan radiation therapy using the characteristics of an individual patient's head, thus improving the effectiveness of this treatment. These code developments have been done on IU's IBM SP.

4.3.5. Cell modeling

Distinguished Professor Peter Ortoleva's Center for Cell and Virus Theory (CCVT) [34] is developing a cell model called Karyote, which makes use of a mesoscopic biochemistry program called M3. Karyote is a particularly comprehensive computer-based simulator of living cells. It accounts for the compartments in a cell within which specialized metabolic, genetic, and other functions are carried out. Because it includes capabilities to deal explicitly with the internal 3D structure within which intracellular metabolic processes take place, Karyote is uniquely suited to make use of the exponential growth of genomic, proteomic, and cell physiological data. As a result Karyote should be of particular importance in drug discovery and treatment optimization. The mesoscopic metabolic simulation program M3 uses a new space-warping method to compute

the global and other major free-energy minimizing structures of large macromolecules and their aggregates as viruses or cell organelles. M3 can be used to screen large sets of drug candidates. M3 can also be used to understand the functioning of viruses and organelles.

Karyote and M3 are both extremely scalable. At present, researchers throughout the world can request an account through which they may specify model parameters and then run a Karyote simulation on IU's 1 TFLOPS IBM SP [35]. Related to this, the Center for Cell and Virus Theory is planning to create a library of model parameters and model results that will be maintained on IU's IBM SP, and thus form a national and international repository of expert information about cells that will be enhanced through integration with cell modeling.

4.3.6. Biomedical text retrieval

The Bio-SIFTER project (Biomedical Smart Information Filtering Technology for Electronic Resources) is developing systems to retrieve and organize information from massive data stores, based on its relevance to user-specified criteria, using multiple intelligent agents [36]. Bio-SIFTER is being developed specifically to handle bioinformatics data. This development has taken place largely on IU's IBM SP. Both unsupervised and supervised classification methods used in Bio-SIFTER have produced good results in relatively short periods of wall-clock time. A recent publication about Bio-SIFTER appeared in the journal Bioinformatics, describing test results based on classification of tens of thousands of EST sequences from humans [37]. IU's efforts with Bio-SIFTER are taking a somewhat different direction than IBM's own text retrieval efforts. Nonetheless, the techniques developed at IU exploit IBM technology and are advancing the state of the art in text searching and retrieval generally.

5. Computer Science

5.1. The Xport project

The goal of the Xport project is to exploit a combination of advanced networking, middleware services, and remote instrumentation technologies to achieve interactive "better-than-being-there" capabilities for remote experiment planning, instrument operation, and data acquisition, reduction, and analysis [38]. The Xport system is being deployed initially for macromolecular crystallography at several facilities, including the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory and the Advanced Photon Source (APS) at the Argonne National Laboratory.

Determining the structure of large biological molecules by X-ray crystallography is made possible in the US through several synchrotron X-ray sources, such as the ALS and APS. A critical problem is access to and management of beam time at these shared facilities. The Xport project is making it possible for scientists to more effectively use these valuable facilities and to do so remotely rather than by traveling to the actual beam site. The Xport system promises to be highly valuable for access to expensive research instruments that are shared nationally, including specialized telescopes and electron microscopes.

Xport is built on current and emerging middleware standards for grid computing. The user interface is the CCAT Active Notebook [39]. Several plug-in CCAT components provide data management, data reduction, and collaboration tools. These components in turn use the grid computing services provided by Globus and the CoG Toolkit [40] for coordinating compute, storage, and network services when users access the beamline. The below shows the layout of CCAT components on the grid during a typical session. Users can add their own components to the notebook, and a generic component is used to wrap existing codes. The current implementation uses the crystallography code MOSFLM for data analysis and visualization. Another important aspect is the integration of mass storage into the data acquisition phase. This

provides both permanent archiving at the time of the run and universal access by collaborators during and after the experiment. Long-term storage for Xport projects is provided by the Indiana University MDSS system. Computational tasks, primarily image processing and structure determination, are performed on IBM SP and the AVIDD cluster.

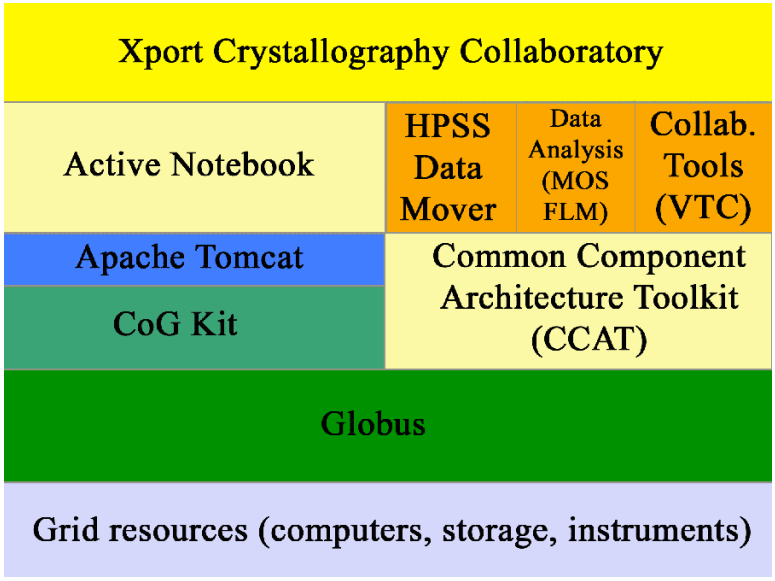

Figure 2. Components of the Xport collaboratory

5.2. Grid computing with Indiana University and Purdue University

Purdue University has recently made significant acquisitions of IBM SP technology. IU and Purdue have leveraged the I-light high speed optical network [5] to create a distributed grid of IBM SPs with an aggregate peak theoretical capacity of nearly 1.5 TFLOPS. This facility has been used in a number of very high-profile experiments and demonstrations. As is the case with IU's distributed SP installation, the IU and Purdue SPs have, for periodic experiments and demonstrations, been seamlessly merged into a single computational resource under a single Loadleveler instance. Using this facility, IU and Purdue have done large scale demonstrations of homeland security and life science applications [41, 6]. Further large-scale integration of Purdue and IU facilities is planned.

6. Education and Outreach

Indiana University has put a great deal of emphasis on educational use of high performance computing in education. Hundreds of graduate students have accounts on IU's IBM SP, thus assuring that many students will have a good deal of exposure to and skill with IBM systems and software technology.

The most recent SUR grant from IBM provided a small (18-processor) Linux cluster for educational use at the IU Northwest (IUN) campus in Gary, Indiana. This campus has the highest proportion of students from traditionally underrepresented groups of any college or university campus in the State of Indiana. IUN has no local equipment to use for classes on parallel, grid, and data-intensive computing. We have already enabled faculty members at IU Northwest to begin making use of this cluster, and have opened up access to this system to undergraduates at IUN. Comments from faculty and students at IUN already indicate that this will be a well used and valued resource. To try to encourage undergraduate use of IU's advanced IT facilities generally, and AVIDD especially, researchers with UITS have created a new class, to be taught

under the auspices of the IU School of Informatics, called "Scientific Informatics." This class will feature hands-on use of AVIDD, the SP, and the MDSS.

Indiana University is already recognized for the breadth of use of its HPC systems [42]. Because of the School of Informatics, Indiana Genomics Initiative, and training within other departments, IU's IBM SP will be used by a large number of undergraduate and graduate students, including the vast majority of MD/PhD students who receive a degree from Indiana University.

IU has made significant progress in outreach activities and events that have drawn attention to the IU/IBM partnership. For example, IBM and Indiana University hosted a reception for IU undergraduate students at the Wrubel Computing Center on October 24, 2001. The purpose of this reception was to give IU students a chance to tour the University Information Technology Services machine room and see IU's recently upgraded IBM SP supercomputer. At just over 1 TFLOPS, this impressive system is the largest university-owned supercomputer in the US. On a very pleasant afternoon over 50 IU students attended the reception, where they also had a chance to hear about career opportunities with IBM. IBM provided documentation about future technology plans and career opportunities, and students had a chance to talk with an IBM recruiter.

Indiana University has a strong commitment to delivering advanced information technology in support of research and education. This event was one of many ways that IU is engaged in increasing awareness and use of high performance computing to advance the state of human knowledge and improve the Indiana economy.

Research and Technical Services Manager Mary Papakhian talks with students about the architecture of the IBM SP frame.

Mary Papakhian leading a tour and explaining theIBM Teraflop SP console area.

UITS staff member Anurag Shankar leading a tour of the Wrubel machine room.

UITS staff member Anurag Shankar discusses IU's distributed massive data storage system

7. Value to IBM

The many accomplishments outlined above detail a variety of technical values to IBM as a result of recent SUR grants:

- The development of significant new software, either released as open source or licensed to IBM under terms of Joint Study Agreements. Thanks in significant part to SUR grants from IBM, IU is now a significant producer and distributor of open source software that is available precompiled for AIX and Linux. Much of this software is related to the life sciences.
- Significant contributions in the development work essential to implement new system software, and in identifying problems and verifying solutions in IBM system software. Some of this has come about through participation in beta tests and some through detection of performance problems in released software.
- Collaborative research and information sharing. Many of the achievements thus far have been accomplished through collaborative work with IBM researchers or product developers. These collaborative relationships have been of great value to both IU and IBM. In addition, the ongoing relationship between IU and IBM has resulted in significant informal sharing of information and expertise.

Indiana University has also served as a reference site and a press contact. Reference site activities have involved significant presentations to other universities and large private companies. Similarly, IU has featured IBM prominently in major announcements. IBM played a major role and received significant press notice as a result of IU's Terascale SP announcement in the fall of 2001, and IU's AVIDD (Analysis and Visualization of Instrument-Driven Data) announcement during the spring of 2003. IU has also consistently made arrangements to feature IBM technology in displays at the annual IEEE/ACM SCxy Supercomputing Conference, and acknowledges IBM's support [43]. A list of press items that describe IU achievements in high performance computing, a very large fraction of which feature IBM technology, is available online at [44].

Indiana University has assiduously acknowledged IBM support through SUR grants and Joint Study Agreements in scholarly publications and conference talks as well. IU researchers have presented talks on matters related to the life sciences at CASCON, SC2001, and most recently at BioITWorld [25, 26, 45]. Papers and posters about IU's massive data storage system have been presented at the Goddard Massive Storage Conference and SC2002 [12, 46]. And Indiana University researchers generally are conscientious in acknowledging IBM support in their scholarly publications. A listing of hundreds of publications that have benefited from IBM SUR grants is available online [47]. A notable recent publication, though not peer reviewed, is a position paper produced by the Coalition for Advanced Scientific Computing [48]. The position paper was prepared at the request of NIH Director Dr. Elias Zerhouni, and addresses the topic of opportunities in the life sciences available through use of high performance computing. The writing of this position paper was led by an IU researcher (Craig Stewart). There is no promotion of IBM technology in the position paper. However, because we (and others in CASC) are in agreement with IBM about many basic principles and strategies regarding high performance computing and the life sciences, many views presented in the position paper are consistent with the views of IBM.

Indiana University has been a recruiting source for IBM and those using IBM technology as well. IBM has recruited from IU at the faculty and student levels. Other companies have recruited from IU in search of expertise with IBM technology as well. Ford has just hired an IU student to participate in their effort to standardize on Linux on IBM servers, as a direct result of this student's experience with Linux and IBM technology at IU.

Indiana University has achieved a leadership position in high performance computing, massive data storage, and networking. IBM support through SUR grants has been an important component and contributor to IU's success in this endeavor. The impact of the SUR grants has been multiplied by IU's success in receiving other funding, thus making the SUR grants a part of a suite of accomplishments that are of much larger scale than the grants themselves. Recent grants of $135M established research centers in computer science (the Pervasive Technology Labs [49]) and the life sciences (the Indiana Genomics Initiative (INGEN) [2]). Most recently Indiana University secured $1.8 million in funding from the NSF to help fund the AVIDD facility. Use of IBM technologies has been a theme running through all of these major activities.

8. Conclusion

IBM has delivered significant support to Indiana University through Shared University Research grants. Indiana University has in turn delivered significant value to IBM and to the broader research community, especially in the life sciences. This value has come through a variety of means. Collaborative research, software development, and involvement in product testing and development have delivered significant value directly to IBM. IU has become a significant producer of open source software, especially in the life sciences. This software is generally made available precompiled for AIX and Linux. IU has involved and featured IBM in a significant array of press events, and has served as a reference site for IBM as well. IU has acknowledged IBM support for IU's activities carefully and prominently. As a result, the impact of IBM SUR grants has been greatly magnified by the consistent involvement of IBM in IU's many activities as a leader in high performance computing, storage, and networking. The relationship between IU and IBM continues to grow stronger and more valuable to both parties, and we look forward to the continued strengthening and deepening of this relationship.

Prepared by:
Craig A. Stewart
Mary Papakhian
David Hart
Anurag Shankar
Andrew Arenson
D.F. (Rick) McMullen
Mathew Palakal
Mehmet Dalkilic
Peter Ortoleva

9. References

[1] Indiana University & IBM Shared University Research Grants. http://www.indiana.edu/~rac/ibm/

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

[3] Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. http://www.cise.nsf.gov/evnt/reports/toc.htm

[4] GRAPE Special-Purpose Compute Systems. http://www.indiana.edu/~rats/research/grapes/grapes.shtml

[5] Indiana's Fiber Optic Initiative. http://www.i-light.org/

[6] Supercomputers at IU, Purdue create new 'tera-scale' grid. http://www.homepages.indiana.edu/062102/text/supercomputer.html

[7] Analysis and Visualization of Instrument-Driven Data. http://www.indiana.edu/~uits/rac/avidd/

[8] Novel computing facility unveiled at Indiana University. http://www.indiana.edu/~uits/cpo/avidd032603/

[9] LAM / MPI. http://www.lam-mpi.org/

[10] Official Maui Homepage. http://supercluster.org/maui/

[11] Anurag Shankar, Gustav Meglicki, Jeff Russ, Haichuan Yang, E. Chris Garrison, Building and supporting a massive data infrastructure for the masses, Proceeding of the 30th annual ACM SIGUCCS User Services conference, Fall 2002. http://www.indiana.edu/~rac/massdatainfra.doc

[12] Shankar, A. 2002. Building a Massive, Distributed Storage Infrastructure at Indiana University. Proceedings of 10th NASA Goddard Conference on Mass Storage Systems and Technologies, Adelphi, MD, April 2002. http://storage.iu.edu/papers/ieee-2002.doc

[13] Samuel Lunenfeld Research Institute University of Toronto. BIND: Biomolecular Interaction Network Database. 2003. http://www.bind.ca/

[14] Swiss Institute of Bioinformatics. ENZYME: Enzyme Nomenclature Database. 2003. http://us.expasy.org/enzyme/

[15] Bioinformatics Center Institute for Chemical Research Kyoto University. LIGAND: Database of Chemical Compounds and Reactions in Biological Pathways. 2003. http://www.genome.ad.jp/ligand/

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

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

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

[19] Department of Genetics at the School of Medicine, Stanford University. SGD: Saccharomyces Genome Database. 2003. http://genome-www.stanford.edu/Saccharomyces/

[20] Bioinformatics Center Institute for Chemical Research Kyoto University. KEGG. 2003. http://www.genome.ad.jp/kegg/kegg1.html

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

[22] IBM, Inc. DiscoveryLink. 2003. http://www-3.ibm.com/solutions/lifesciences/solutions/discoverylink.html

[23] GNU Lesser General Public License. http://www.gnu.org/copyleft/lesser.html

[24] RAC Online Resources. http://www.indiana.edu/~rac/online_resources.html

[25] Stewart, C.A., T.W. Tan, M. Buckhorn, D. Hart, D. K. Berry, L. Zhang, E. Wernert, M. Sakharkar, W. Fischer, and D.F. McMullen. 1999. "Evolutionary biology and high performance computing." Presented at CASCON, Toronto, Canada, Nov. '99.

[26] Stewart, C.A., D. Hart, D. K. Berry, G. J. Olsen, E. Wernert, W. Fischer. 2001. "Parallel implementation and performance of fastDNAml - a program for maximum likelihood phylogenetic inference." Proceedings of SC2001, Denver, CO, November 2001. http://www.sc2001.org/papers/pap.pap191.pdf

[27] Parallel fastDNAml Distribution. http://www.indiana.edu/~rac/hpc/fastDNAml/index.html

[28] The PINY_MD simulation package. http://homepages.nyu.edu/~mt33/PINY_MD/PINY.html

[29] Parallel PINY_MD Distribution. http://www.indiana.edu/~rac/hpc/PINY_MD/

[30] RSICC CODE PACKAGE CCC-713 [PENELOPE-MPI]. http://www-rsicc.ornl.gov/codes/ccc/ccc7/ccc-713.html

[31] Gamma-Knife Center. http://www.clarian.org/clinical/gammaknife/index.jhtml

[32] R.B. Cruise and L.S. Papiez, Integral Equation Formulation of a Mixed Diffusion-Jump Model of Elastic Scattering. Nuclear Mathematical and Computational Science: A Century in Review, a Century Anew. American Nuclear Society, LaGrange Park, Illinois, 2003.

[33] R.B. Cruise, R.W. Shepard and V.P. Moskvin, Parallelization of the Penelope Monte Carlo Particle Transport Simulation Package. Nuclear Mathematical and Computational Science: A Century in Review, a Century Anew. American Nuclear Society, LaGrange Park, Illinois, 2003.

[34] Center for Cell and Virus Theory. http://biodynamics.indiana.edu/

[35] Karyote. http://biodynamics.indiana.edu/cyber_cell/

[36] SIFTER Information Filtering. http://sifter.indiana.edu/

[37] M. Palakal, S. Mukhopadhyay, J. Mostafa, R. Raje, M. N'Cho, and S.K. Mishra, An Intelligent Biological Information Management System, Bioinformatics, 2002.

[38] Extreme! Computing. http://www.extreme.indiana.edu/

[39] The XCAT Project. http://www.extreme.indiana.edu/xcat/index.html

[40] The Globus Project. http://www.globus.org/

[41] New simulation shows 9/11 plane crash with scientific detail. http://news.uns.purdue.edu/UNS/html4ever/020910.Sozen.Pentagon.html

[42] 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 (PDF). Proceedings of SIGUCCS, Portland, OR, October 2001. http://www.indiana.edu/~rac/siguccs_copyright.html

[43] Research in Indiana. http://www.indiana.edu/~rindiana/

[44] RAC in the News. http://www.indiana.edu/~rac/racnews/racnews.html

[45] Stewart, C.A., and R. Repasky. 2003. "High performance computing for university biomedical research: a successful implementation." Proceedings of BioITWorld conference and Expo, Boston, MA, March 2003. http://www.indiana.edu/~rac/papers/IU_bioitworld_2003.pdf

[46] Building a Massive Data Storage Infrastructure for the Masses, Poster presentation at SC2002. http://www.research-indiana.org/techprogram.html

[47] Publications resulting from use of IU High Performance Computing Systems. http://www.indiana.edu/~rac/hpc/papers.html

[48] CASC White Paper on Biomedical Research. http://www.ncsc.org/casc/papers/paper13.pdf

[49] The Pervasive Technology Labs at Indiana University. http://www.pervasivetechnologylabs.iu.edu/