Creation of the AVIDD Data Facility: A Distributed Facility for Managing, Analyzing and Visualizing Instrument-Driven Data (AVIDD)

Final Report

NSF Award ID: EIA-0116050
Project Dates: 9/1/2001 to 8/31/2003
Principal Investigator: McRobbie, Michael
Co-Investigators: Bramley, Randall, Huffman, John C., Stewart, Craig A.
Organization: Indiana University

Table of Contents

Executive Summary…….….. 1

Activities and Findings
The AVIDD Facility
The Computational Component…….…....4
The Data Management and Storage Component……..……..6
The Visualization Component……8
The Integrated AVIDD System…..…….....….10

Research Enabled by AVIDD
Computer Science……..……...12
Life Sciences (Biology, Chemistry, Medicine)……14
Physics and Astrophysics…….....17
Geology…….21
Geography, Atmospheric Sciences, and GIS……....21
Artistic ……22

Economic Outcomes…….23

Human Resources Outcomes……25

Publications……..26

Outreach Activities……..29

Implementation of the AVIDD system (Analysis and Visualization of Instrument-Driven Data)

1) Executive Summary

By early 2001, researchers and information technologists at Indiana University recognized a pressing need to provide better facilities and support for data-intensive science. In response, a group of IU scientists led by Vice President for Information Technology Michael A. McRobbie, submitted a proposal to the National Science Foundation's Major Research Inititiative Program, entitled "Creation of the AVIDD Data Facility: A distributed facility for managing, Analyzing and Visualizing Instrument-Driven Data". The proposed solution called for geographically distributed Linux clusters built on commodity components with high-speed networks with a large amount of high-performance disk to be integrated into IU's existing massive data storage facility, as well as state-of-the-art visualization and data presentation environments. The proposal was successful, with a NSF grant award of $1.8 million in October, 2001. Request for proposals (RFPs) were sent out to vendors; Intel 32-bit and 64-bit platforms from IBM with Myricom interconnects were selected in July, 2002. Hardware began arriving in September 2002, with the first cluster PentiumIII cluster installed on the IUN campus in November 2002. Two PentiumIV clusters were installed on the IUB and IUPUI campuses in January 2003, and a fourth Itanium2 cluster was installed at IUPUI in May 2003. By September 2003 installation was also completed for the AVIDD visualization components at IUN, IUB and IUPUI.

Initial use of the AVIDD clusters, with the exception of the instructional component at IUN, was limited to the research teams of the principal investigators named on the NSF grant proposal. By July, 2003, access to AVIDD was opened up to all IU faculty and their sponsored graduate students and staff. Since that time the user base has increased to ~300 users from a variety of disciplines, with high utilization by researchers in Chemistry, Physics, Informatics, Computer Science, and Astronomy. The ability of AVIDD to handle the storage, computational and visualization needs of applications with very large datasets generated from scientific instruments, such as the iVDGL and GlueX physics projects, was indeed gratifying to AVIDD's initial users, who until then had been largely limited to analyzing subsets of the data read in from tape on far less capable departmental servers. But AVIDD also proved very popular with data-mining applications which relied on processing online databases or data culled from the WWW by multiple simultaneous processes. New research was empowered by AVIDD's ability to pull in data quickly, either from AVIDD's high-performance disk, from online sources, or from IU's mass store tape silos, process the data with multiple, very fast processors, and then write results back to AVIDD's high-performance disk via its parallel filesystems. Similarly, deployment of AVIDD's powerful visualization components (tiled displays, display walls and smaller 3D stereo display devices) at various locations at each of the three campuses allowed enhanced 3D display of researchers' data within convenient access.

Beyond serving the expanded data handling needs for which AVIDD was intended, the AVIDD facility has also attracted a number of IU researchers already familiar with Linux workstations and clusters. The ease of porting their Linux applications and the large compute and data capacity available within a fairshare allotment of resources has made AVIDD particularly attractive to users who had not previously made much use of IU's high performance computing systems. And because AVIDD is architected on commodity components and an open source operating system, expanding its resources to meet anticipated future demand should be relatively inexpensive, relatively easy to implement, and non-disruptive to users.

2) Introduction

For much of its history the primary focus of the computational sciences has been on the speed of numerical computation and the hardware, algorithms, and software required to maximize this speed. The needs of data-intensive science received secondary consideration. But the steady proliferation of scientific instruments that generate vast amounts of data is demanding a new generation of facilities constructed specifically to support the needs of modern instrument-driven data-intensive science. Such facilities must address the full data life cycle: data capture and onsite data reduction; high-speed data transfer; real time data analysis and processing; data storage; data retrieval; data analysis and post-processing; data visualization; and the use of remote data stores.

Aided by a major grant from the National Science Foundation, Indiana University has created a distributed data management facility for the **A**nalysis and **V**isualizing **I**nstrument-**D**riven **D**ata flows (AVIDD). It is distributed among three of Indiana University's campuses, and integrated with very high bandwidth using the high-speed, university-owned I-light Network [1]. The AVIDD system has already resulted in dozens of publications (see References), presentations (see Appendix 1), and press announcements (see Appendix 2). This report summarizes the architecture of the AVIDD system, the information technology advances created by Indiana University that made this system possible, some of the research breakthroughs that have been enabled by this system, plus other benefits accrued from AVIDD.

Figure 1. Photos from the AVIDD Kickoff Event at Wrubel Computing Center, March 26, 2003.

3) The AVIDD Facility

AVIDD consists of the following three main components:

- Computational component - A distributed quartet of Linux clusters with a combined total of 2.2 teraflops of processing capacity and 10 terabytes of distributed storage.
- Massive data storage facilities for storage of hundreds of terabytes of data.
- A network of 3D visualization systems installed in laboratories throughout the university.

The technical details of these components are described below.

Computational component of AVIDD

The computational component of AVIDD consists of a quartet of IBM IA32 and IA64 Linux clusters. Ten terabytes of disk space are available, providing a 5:1 ratio of storage to processing capability. This ratio is unusually high compared to other existing high performance computing facilities and is one aspect of AVIDD that makes it ideally suited for data-intensive computing. A total of 470 processors provide an aggregate processing capacity of 2.2 teraflops:

- Two clusters, each with 208 Prestonia 2.4-GHz processors, provide the bulk of the processing power.
- A third cluster provides 36 64-bit McKinley processors for computer science research.
- The fourth cluster, located at Indiana University Northwest, has 18 1.3-GHz PIII processors. This cluster was funded by a Shared University Research grant from IBM and, along with advanced visualization equipment located at IUN, plays a particularly important role in the educational aspects of the AVIDD facility.

Figure 2. Photo of the IUB PentiumIV AVIDD cluster

Some of the initial research carried out with AVIDD facilities focused on the interaction between cluster resource managers and the Linux kernel's real-time features [23]. Resource management tools evaluated included OpenPBS, PBSPro, and Condor. The production AVIDD clusters initially used PBSPro to take advantage of its preemption capabilities. Later, PBSPro was replaced with Torque (an Open-PBS based resource manager) to benefit from its extended scalability and additional functionality. Most significant is its integration with the Maui Scheduler, which provides much-needed capabilities for batch job scheduling on AVIDD, including fairshare, advanced reservation, multidimensional job throttling policies and preemption support

Indiana University is one of five charter members of the Maui Consortium (http://www.supercluster.org/mauicon), which aims to advance HPC cluster and grid scheduling tools, and manage further development of the Maui Scheduler. Several enhancements have been implemented in the Maui scheduler specifically for the AVIDD

clusters, and these enhancements will be of general use to the cluster and grid communities generally. Specifically, the Maui scheduler has been improved to include:

- Fault tolerance during resource manager failures
- Support for the preemption facilities in OpenPBS
- Tracking of resource manager specific names for referencing job objects
- Multidimensional and QOS-based job throttling policies
- Support for network adapters as consumable resources

In addition, the following enhancements to the Maui scheduler are currently being implemented:

- Generic resource definition and consumption tracking
- Floating consumable resource support
- Multiple resource manager integration, including license managers and global file systems

Also, as a member of Maui Consortium, Indiana University has been able to influence the development of code enhancements for a highly scalable open source cluster resource manager based on OpenPBS, formerly referred to as ScalablePBS but now known as Torque (Tera-scale Open Resource and Queue manager) [2]. Indiana University's experience with SPBS/Torque on AVIDD has resulted in several bug fixes, plus efficiency and functionality enhancements beyond what is offered in OpenPBS today.

Data management and storage component of AVIDD

The archival data storage component of AVIDD was implemented as an expansion of IU's High Performance Storage System (HPSS) infrastructure. To deliver high-speed archival data storage to the AVIDD Linux clusters in a dedicated fashion, two IBM p640 servers, four IBM 2104 RAID disk arrays (providing 1.7 TB of HPSS disk cache), and six StorageTek T9940A drives were installed at IUPUI. Along with two existing IBM p640 servers, four data movers (with both directly attached disk and tape drives) were configured in HPSS to construct four-way, striped disk and tape storage for AVIDD. Parallel data movement between four AVIDD Linux cluster nodes configured with HPSS clients and the four HPSS data movers occurred via a Cisco 4108 gigabit Ethernet switch.

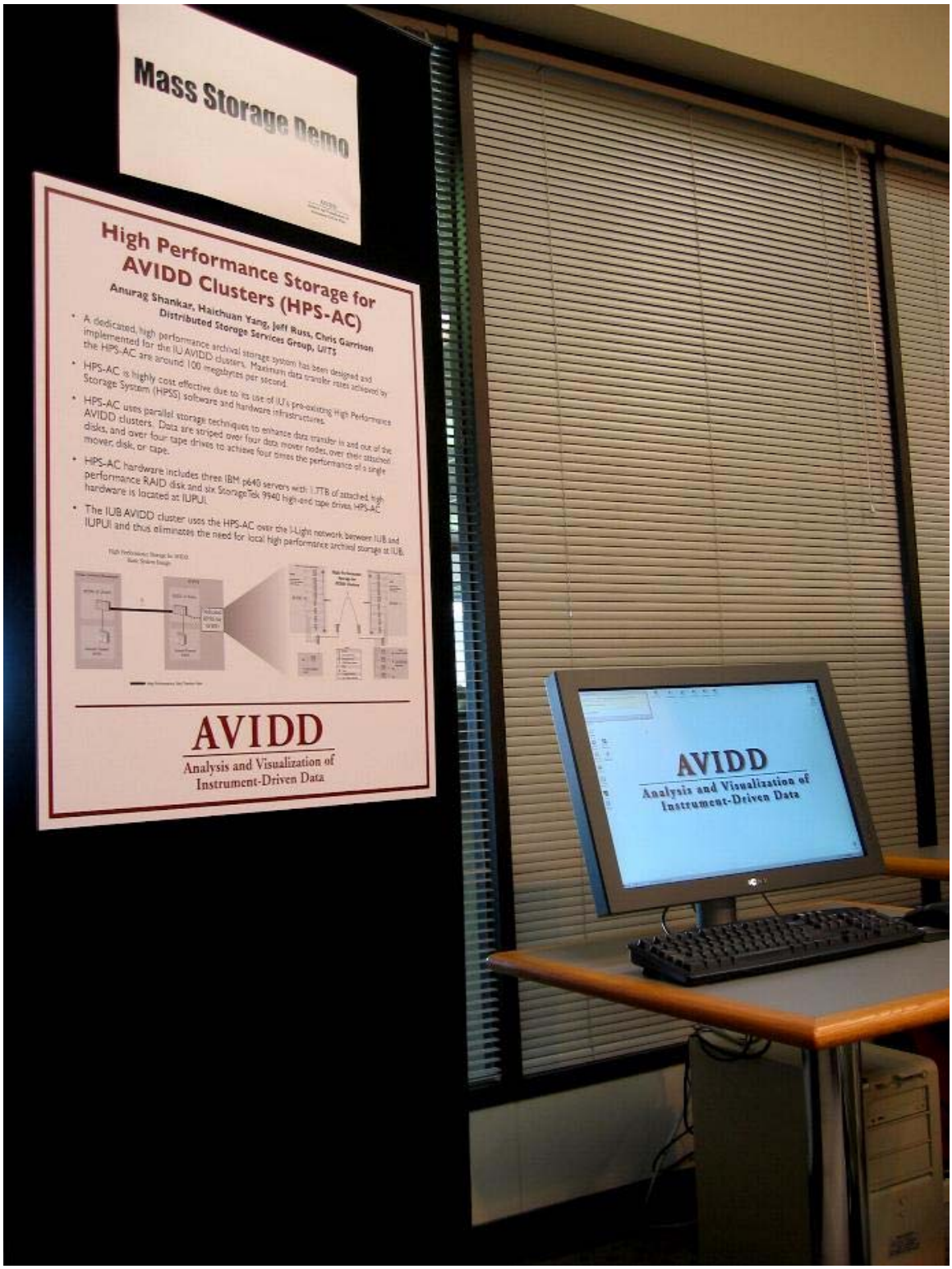


Figure 3. Exhibit displaying AVIDD's high-performance data storage configuration at the AVIDD Kickoff Event, March 26, 2003.

The high performance (4-way, striped, parallel) disk cache comprising 1.7 TB allow users on the AVIDD Linux cluster at IUPUI to read data from HPSS at a peak speed of 200MB/s. Writes to the disk cache occur at a peak speed of 100MB/s. At these speeds, a 1 TB file can be copied from the disk cache to the AVIDD cluster in roughly 1.4 hours. The reverse (1TB file copied from the AVIDD cluster to the HPSS disk cache) takes around 2.8 hours. The six tape drives permit the loading of a 1 TB file directly from four-way, striped tape in approximately 5 hours. This allows users to move large data files between the AVIDD Linux clusters and the disk cache very rapidly, while at the same time dramatically improving the speed with which data is staged from tape to disk and back.

The institution of four-way, parallel HPSS data transfers for AVIDD required significant collaboration between IBM Global Services in Houston (which provides HPSS software support) and IU. The AVIDD high-performance HPSS represents the first production instance of parallel HPSS data movers anywhere. In fact, the AVIDD project allowed the HPSS collaboration an opportunity to test the HPSS multinoded parallel FTP system. Numerous bugs and performance problems in the software were uncovered and fixed during the AVIDD project.

An example of a project at IU that currently benefits from high performance storage on AVIDD is the high energy physics Hall-D project. Physicists at IU engaged in designing the data analysis system for this next-generation, high energy physics experiment store large files (tens of GB to a TB) on 4-way disk cache and then to tape, and then stage these files back to AVIDD's high-performance filesystem disk to perform data analysis. Without the AVIDD high performance, archival data storage design, this kind of data analysis would not be possible.

Visualization component of AVIDD

IU has developed and tested three separate types of display systems to be deployed as part of the AVIDD visualization component:

- Small, portable systems with 3' x 4' passive stereo displays called John-e-Boxes.
- Wall displays, very similar in technology to the portable displays, but with ceiling mounted projectors. (These may be either rear or front projected.)
- Larger, tiled display walls.

.

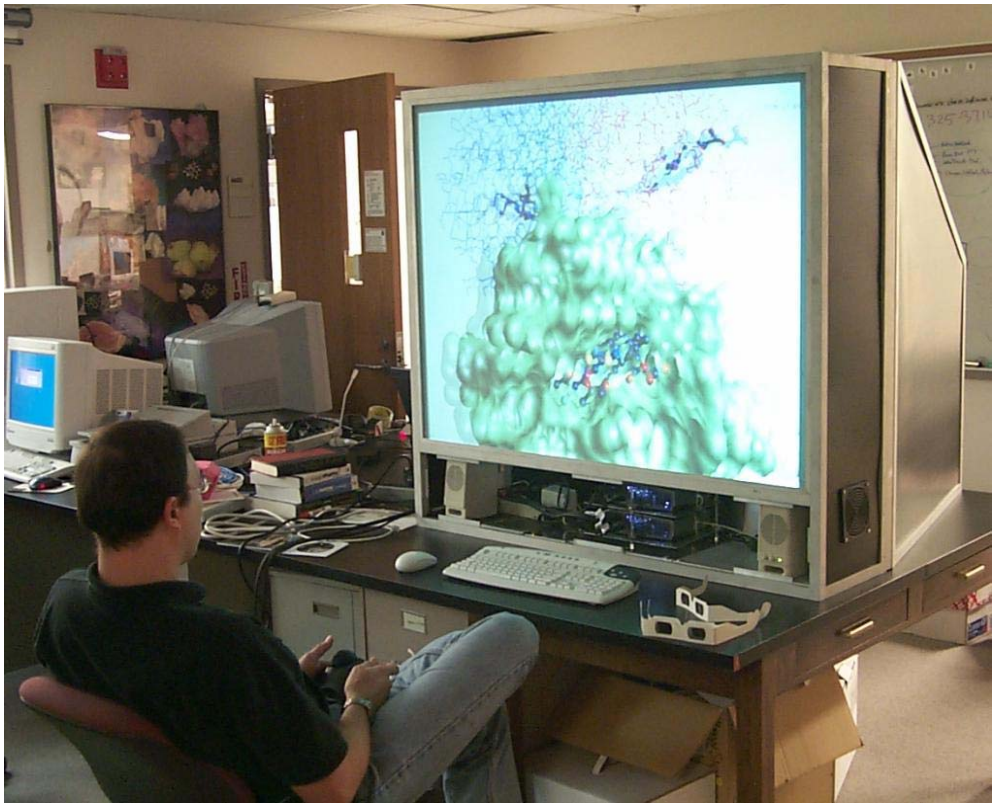


Figure 4. Prototype of the John-e-Box in use for examining molecular structures

The John-e-Box (see Figure 1) is a small-scale (50" cube) device that provides very good quality 3D visualization, and which is now available commercially. This is a passive 3D display based on commodity components, leveraging design principles from other PACI visualization projects but with highly novel design and packaging [3].

Deployment of John-e-Boxes as part of the AVIDD project includes the following locations:

- School of Informatics (IU Bloomington)
- Department of Chemistry (IUB)
- University Information Technology Services (IU B)
- University Information Technology Services (Indiana University Purdue University Indianapolis – IUPUI)
- IU Northwest
- Department of Biology (IUB)
- Department of Geology and Indiana Geological Sciences (IUB)
- Department of Physics (IUB)
- Department of Computer and Information Science (IUPUI)
- IU East

As a result o the AVIDD project, then, half of IU's eight campuses are equipped with 3D visualization and collaboration environments.

These visualization systems are being used for a variety of purposes, including molecular visualization, GIS analysis, volumetric rendering, and biomedical research visualization.

The John-e-Box systems have also been instrumental in a number of educational and outreach programs, including: Supercomputing 2002 and 2003, I-Light 2002 and 2004, the Indiana Statewide GIS Conference 2004, NSF ESTME event 2004, and IU's annual IT awareness campaign "Making IT Happen" at IUB, IUPUI, IUN, and IUK campuses.

While the John-e-Boxes are the core of the AVIDD visualization facilities, other types of visualization facilities have been installed based on needs of the application scientists who use AVIDD. A non-stereo, high-resolution, high-brightness/contrast system has been deployed at the IU School of Medicine (Department of Nephrology) for use with biomedical microscopy visualization. A larger 2x2 tiled display system has been deployed at the School of Informatics at IUB. This fixed display wall has been configured as an Access Grid Node and has been used to host meetings with collaborators accross the US, as well as others in Germany, the UK, and Australia. This Access Grid Node has been instrumental in facilitating a class on Virtual Environments that is being jointly taught between IU Bloomington and Purdue University. The success of this class has led to plans to expand the course to include other institutions and other IU campuses.

With supplemental funding from other sources, UITS is creating two smaller, mobile versions of the Access Grid at IUB and IUPUI.

The AVIDD grant has been leveraged in several ways, including enhancement of the visualization component beyond that initially envisioned in the AVIDD proposal. With supplemental funding from another source, UITS is deploying an 8-node, dual-Opteron cluster with large memory and Infiniband interconnect. This system will be used to test a number of advanced rendering topologies and network-based visualization techniques, including: high-resolution, rendering for tiled displays; synchronous, real-time rendering for multi-walled VR environments; massive data rendering using data distribution and image compositing methods; and remote rendering and visualization techniques, including visualization portals. This system will be deployed on the IUPUI campus by the end of June 2004.

The integrated AVIDD system

Indiana University has indeed done what it set out to do with the AVIDD project. We have created an integrated system of computational, storage, and visualization resources that spans the entire range of the data life cycle from collection, through analysis, visualization, and storage, and does so in a geographically distributed fashion that meets the needs of application scientists while also providing a valuable resource for computer science research. A schematic of the entire AVIDD system is shown in Figure 5.

2.

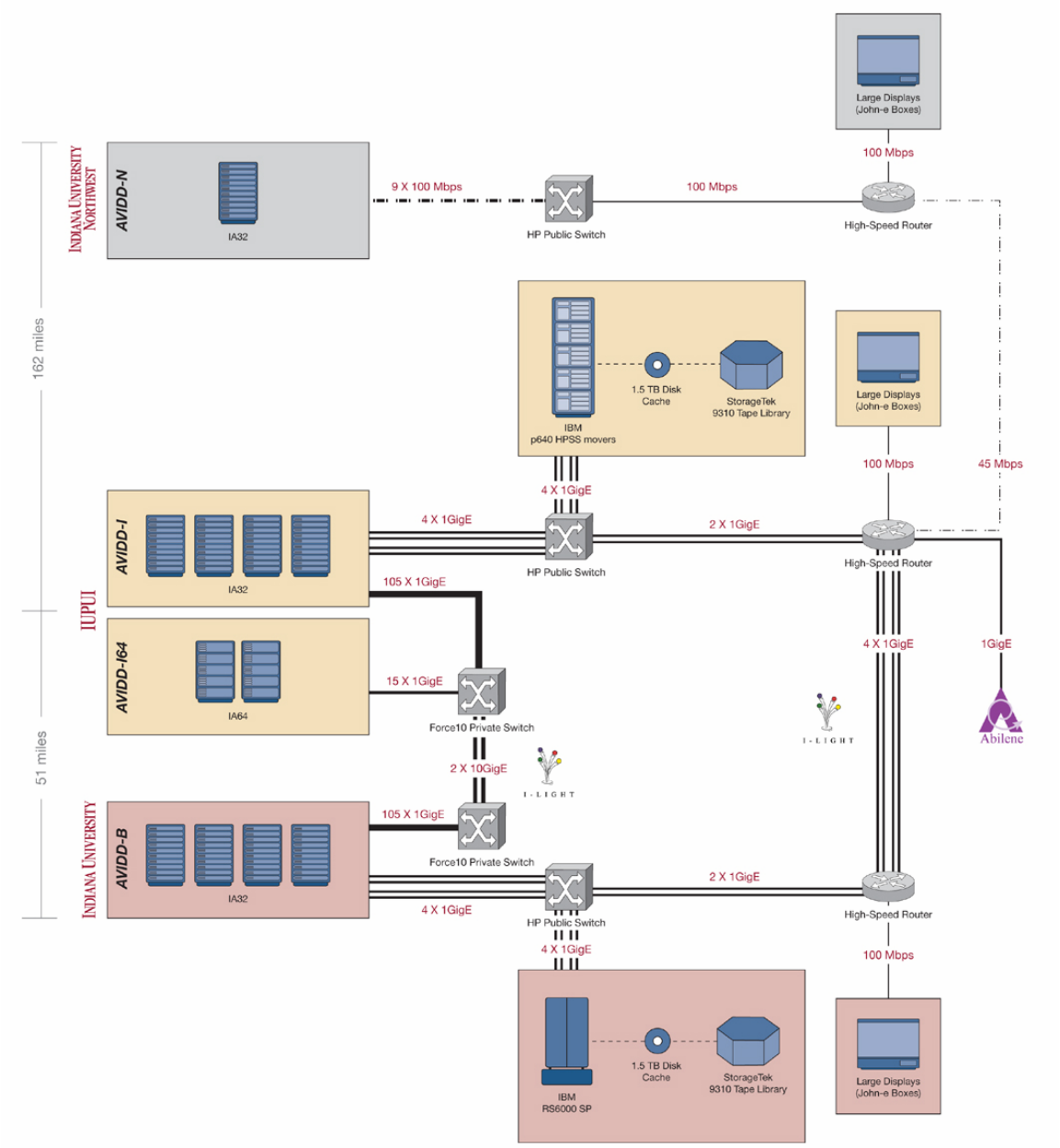


Figure 5. Schematic diagram of the entire AVIDD system.

4) The Scientific Outcomes: Research Enabled by AVIDD

Our key project finding has been that it is possible to enable major, new, scientific breakthroughs as the result of implementing an integrated information technology system that addresses the full data life cycle, from data collection, storage and reduction, analysis and simulation, to visualization. Implementation of a system for the purpose of turning massive streams of digital data into scientific insight calls for a fundamentally different

configuration, implementation, and management plan than the common approach of using Linux clusters to attain highly cost-effective systems for parallel applications with very high computation to communication ratios. Nonetheless, the AVIDD distributed Linux cluster has provided very high-end computational capabilities. AVIDD was listed in 50th place on the June 2003 Top500 list – the highest ranked distributed Linux cluster on that list, and the first distributed Linux cluster to document the achievement of more than 1 TFLOPS on the Linpack benchmark [41].

Computer Science

Grid Computing (Dennis Gannon, Department of Computer Science, IU Bloomington)

AVIDD is a grid architecture that is similar to the NSF Teragrid in terms of its software stack. Part of the work in computer science that uses AVIDD has been looking at the design of software infrastructure for Grid systems. This research involves a software component architecture (XCAT) based on the CCA standard and high performance communication protocols between distributed components (the Proteus Multiprotocol Library) and web service frameworks for distributed scientific applications [20]. AVIDD is an essential research testbed for this work.

Scientific and Grid Computing (Randall Bramley, Department of Computer Science, IU Bloomington)

A major computer science research project that relies on AVIDD is our work on the "MxN problem" [13]. This is a scenario of increasing importance and frequency in scientific computing and occurs when codes that simulate different physics and time/length scales are connected to create a larger, interdisciplinary simulation. This is the case in climate modeling where ocean, atmosphere, land ice, sea ice, groundwater transport, and chemistry are connected to create a higher fidelity simulation. It is has been identified as one of the most pressing issues in fusion energy simulations, where codes for transport, MHD, radio frequency heating, materials properties, and other effects need to be combined to achieve fully predictive fusion burns in tokamaks. On a smaller scale, it is happening at many universities, as scientists have begun to develop multidisciplinary models from their existing discipline-specific codes.

The MxN problem occurs when those codes are parallel, in which case a code running on M processors needs to directly communicate in parallel with a code running N processors. At IU we have been investigating this problem, in activities ranging from the semantics of parallel remote method invocation to actual implementations. Two subprojects in particular are using AVIDD. One uses an MPI I/O type data to a logical file. However, the underlying I/O device is a network transfer mechanism that automatically sets up the necessary connections and transfers the data between two live programs. This allows users to read and write a logically serialized data object, but in practice the communications are completely in parallel. An advantage of this MPI I/O solution to the MxN problem is that users are provided with a familiar API, and dynamically at runtime it can be determined if the data will go to a file or through parallel

network streams to another code. For scalability and testing, however, it requires a large number of I/O nodes, geographically distributed sets of nodes, and fast network transfers between the sets of nodes - which are precisely the features that AVIDD was designed to provide .

A second subproject also addresses the MxN problem, but in the context of the Common Component Architecture (CCA), an effort by the national laboratories and several universities to develop standards for high-performance computing software components [14,17]. The CCA requires a "uses/provides" design pattern, and this places strictures on the MxN transfer. In "DCA: A Distributed CCA Framework Based on MPI" we have identified those constraints and designed the first CCA framework which handles components that are both parallel and distributed. Like the MPI I/O approach, this work presents to an application user an interface that uses familiar MPI terminology and leverages the underlying MPI run-time system as much as possible, while handling the parallel data transfers automatically. The DCA has been used in a prototype demonstration, showing that a time-optimal solution to a signal processing problem does require differing numbers of processors for each component, and can be done in fully scalable fashion.

Development of LAM/MPI (Andrew Lumsdaine, Department of Computer Science & Pervasive Technology Labs IU-Blooomington)

As a part of the development of LAM/MPI, the LAM Development Team uses AVIDD-B for testing and performance evaluation purposes. LAM is an open source, production-quality implementation of MPI [11,28-34].

During development, individual modules are developed and tested on AVIDD for functionality, performance measurements, and stress testing. With every release of LAM/MPI, AVIDD is utilized for extensive testing, including platform compatibility and scalability testing. AVIDD's SMP capabilities are also a critical component in LAM's development, testing, and release procedures in order to ensure correct functionality in both pure shared memory and mixed share memory/network parallelism models.

Since LAM now natively supports MPI jobs in a PBS batch scheduler environment, the AVIDD administrators can exert positive control on parallel jobs submitted on the cluster as well as generate accurate usage statistics on a per-user basis.

Network Flow Analysis for Computer Security (David Monnier, University Information Technology Services, IU Bloomington)

This project focused on real-time analysis and visualization of network flow data from Indiana University's backbone and campus routers. Using AVIDD we are able to detect network anomalies and, by use of the high speed inter-node network, visualize network utilization in near real-time. The scale of the network data is beyond that of standard computational capabilities. Current data set size is 35,000 samples per second and will soon be scaled up to 1,250,000 samples per second.

The ability to detect network anomalies (attacks, scans, etc) in real-time will lead to the development of self-reacting network security measures. ITSO (IT Security Office) and the NOC (Network Operations Center) hope to develop and deploy this capability in the near future.

Visualization of network data allows for the detection of network problems in an intuitive fashion. Allowing engineers to work with visualized data all the way down to the protocol level. Where traditionally router load would be the only data visualized.

Life Sciences (Biology, Biochemistry, Chemistry, Medicine)

BioMap (Mathew Palakal, Department of Computer Science, Indiana University Purdue University Indianapolis)

The biological literature databases continue to grow rapidly with vital information that is important for conducting sound biomedical research. The purpose of the BioMap project is the development of a scalable knowledge base (BioMap) of biological relationships from vast amount of literature data. The results of this may ignificantly enhance the ability of biological researchers with diverse objectives to efficiently utilize on-line resources, generate methods for analysis of biological data such as identifying biological pathways, and provide computerized support for disease target and new drug discovery. BioMap will be a new type of secondary knowledge resource derived from primary resources such as MEDLINE. It will be the "window" to every biomedical researcher who will be seeking knowledge from the literature databases, however, without being overwhelmed by its large volume [21].

In this work some preliminary results were generated as we progress toward with the development of the BioMap system. A key component of the system, the Knowledge Bases for multiple organisms, has been fully developed. As an initial step toward discovering knowledge embedded as object-object relationships, the object identification problem was investigated using multiple dictionaries. Preliminary results based on 30,000 MEDLINE abstracts showed that existing dictionaries can only resolve less than 50% of biological objects names. These results indicate the need for developing intelligent algorithms for resolving biological object names. The final goal of the project is to build BioMap, a knowledge base of biomedical literature. The very first step of this project is to collect data from all kinds of biological literature databases. So far, we downloaded MEDLINE documents, which contain about 12 million abstracts.Using Brill tagging tool, we have analyzed the MEDLINE documents, and extracted the necessary information which has been loaded into an Oracle database.

Optimization of Integration of X-ray Diffraction Intensities (John C. Huffman, Department of Chemistry, IU Bloomington)

A critical phase in the elucidation of molecular structures by X-ray diffraction techniques is the integration of the individual intensities in the "CCD frame data" collected either at local sources or national synchrotron beamlines. One of the problems is that there are a large number of independent parameters that can be adjusted to obtain optimal integration. We are investigating the use of the AVIDD system to allow the user to input a range of parameters so that significantly improved data can be obtained in a realistic time frame.

With the knowledge of this study we will be able to develop a software/graphical interface to run on the AVIDD system to facilitate the parameter adjustments and analysis of the results of the integrated intensities more rapidly.

To test the procedure, X-ray diffraction intensities of a crystalline sample of a small organic compound at low temperature (-136 C) were recorded. The data consisted of 3600 frames from a 4K x 4K CCD Bruker SMART 6000. The recorded intensities were integrated using the SAINT software for Linux platform (NDA from Bruker AXS).

A total of 9 integration parameters were varied and hence 9 separate SAINT integration processes (~ 3 min per process) were running on on the AVIDD cluster simultaneously. The result of each integration process was examined and the preliminary result indicated that the default parameters in the SAINT software do not result in the best integration profiles for the above mentioned data set .

*Protein Structur*e (Thomas Hurley, Department of Biochemistry & Molecular Biology, IUPUI)

The Center of Structure Biology at IUPUI campus focuses on solving the structures of proteins and protein-ligand complexes using X-ray crystallography methods. SHELX97 software package is especially useful in solving protein structures at ultra-high resolution. SHELX97 has been optimized to run more than three times faster and ported to the AVIDD system. Typical run times have been reduced from more than three days per computation to a single day.

In addition, the use of AVIDD-64 facilitates the solution of very large protein structures without the need to subdivide the data into smaller subsets. The ALS (Advanced Light Source) [7] isn't exactly running yet, but the code is already on AVIDD-64 waiting for them.

Cell Growth and Protein Content and Synthesis (James Reilly, Department of Chemistry, IU Bloomington)

The Reilly Group analyzes variations in protein content and synthesis as a function of cell cycle stage and growth conditions, using MALDI-TOF mass spectrometry. The laboratory will eventually generate over 1 gigabyte of data per hour, requiring automated analysis. The software for this analysis has been prototyped and tested, ported to AVIDD, and is currently being parallelized.

Understanding and Improving Transition Metal-based Anticancer Drugs (Mu-Hyun Baik, School of Informatics, IU Bloomington)

Cisplatin, a simple Pt(II)-complex (cis-[Pt(NH3)2Cl2]), is one of the most widely used anticancer drugs today. It is particularly successful against ovarian, testicular, head, neck and small cell lung cancer. Over the last few decades, much research effort has been dedicated to understanding how cisplatin reacts with its primary cellular target, genomic DNA. As a result, there is now general agreement on the overall mechanism of antitumor activity. However, many crucial details of the mechanism are not well understood. We are interested in identifying the electronic details of how cisplatin interacts with DNA and rationally deriving general drug design strategies for obtaining cisplatin analogues that display higher antitumor activity, are more specific for tumor cells or allow for overcoming cisplatin resistance. This work involves both studying possible drug candidates specifically and understanding in a broader sense how transition metals interact with nucleobases [10].

Reactivity of Titanium and Vanadium complexes of Beta-Diketiminate Ligands (Mu-Hyun Baik, School of Informatics, IU Bloomington)

In collaboration with an experimental chemist in the chemistry department of IUB, we are exploring the reactivity of Titanium and Vanadium Complexes of beta-diketiminate ligands. These molecules display unusual structural features that make them potentially interesting as new catalysts for a number of industrial processes. Catalysts promote otherwise difficult chemical transformations and allow for carrying out chemical reactions that usually require high temperatures and high pressures at much milder conditions. Catalysts are key to energy conservation and industrial production of materials that would not be available otherwise [8,12].

Computational Studies of Metalloproteins (Mu-Hyun Baik, School of Informatics, IU Bloomington)

Transition metal centers frequently play a crucial role in biologically important processes. Mostly embedded into a protein scaffold, they often catalyze demanding reactions at ambient conditions serving as the reactive centers in metalloenzymes. We are interested in studying a number of metalloproteins for two purposes. First, we are curious about how the protein environment impacts the electronic structure of the metal center. Being inorganic chemists, we like to think of the protein as a gigantic ligand that can perform a number of difficult tasks, such as isolating reactive intermediates from solvents and other reactants, controlling the hydrophobicity of the local surrounding or enforcing otherwise impossible structural distortions that promote a certain reaction. We would like to understand in detail, how nature has integrated these features into a biological machinery. Second, we are interested in designing a minimalist model of the natural system that can mimic the enzymatic behavior and carry out the same or similar reactions. The goal is to design biomimetic catalysts that can be used in a technical setting for carrying out reactions that require high temperatures and/or high pressures [9].

Phylogenetic Inference: Global Analysis of Arthropod Evolution, (Craig Stewart, David Hart, Richard Repasky, University Information Technology Services, IU Bloomington)

This project was an entry in the HPC Challenge competition at the SC2003 meeting in Phoenix, Arizona, November 2003. The leading institutions were University Information Technology Services at Indiana University, the Center for Genomics and Bioinformatics at Indiana University, and the High Performance Computing Center at the University of Stuttgart; many other institutions throughout the world contributed their resources. Computers around the world were used to investigate two questions regarding the evolution of arthropods: are hexapods paraphyletic? are ecdysozoans paraphyletic?

These issues are currently being intensively debated. Its analyses are computationally intensive and the availability of computing resources limits the number of organisms or the number of genes analyzed; more data would produce a more reliable prediction. We assembled molecular data from 12 mitochondrial genes in 67 species.

The project won the prize for "most distributed application" in the contest. Grid computing tools were used to distribute, launch and control geographically distributed jobs and to visualize results. Tools were built using the PACX-MPI parallel programming interface and the COVISE collaborative visualization and simulation environment. It showed that PACX-MPI and COVISE can be successfully used to construct large-scale grid applications. The analysis was performed using 23 computer systems on 5 continents (AVIDD prominent among them) [18,19].

Physics & Astrophysics

International Virtual Data Grid Laboratory (iVDGL) and ATLAS High Energy Physics Experiment (Frederick Luehring, Department of Physics, IU Bloomington)

Starting in April of 2003, the IU ATLAS Experiment prototype Tier 2 center was relocated from older hardware to the AVIDD cluster. A Tier 2 center is a regional computing center for running scientific computations requiring large amounts of computing and storage using grid-enabled applications. ATLAS is one of two large general-purpose particle physics experiments that will take data using particle beams produced by the Large Hadronic Collider (LHC) at CERN, the European Laboratory for Particle Physics located near Geneva Switzerland. The work on the Tier 2 center is part of the research conducted by the International Virtual Data Grid Laboratory that is NSF-funded. Approximately $41,000 of iVDGL funding was spent in adding about 1.5 TB of high-speed fiber channel disks to the AVIDD cluster for use in ATLAS and iVDGL research. The Tier 2 center is assigned 64 CPUs of the AVIDD-B cluster and has a complete suite of Globus-based grid middleware installed on it. The Tier 2 center was

fully operational in June and the previous Tier 2 hardware was permanently shutdown in August.

The Tier 2 center fully participated in the Grid 2003 project and will participate in the follow-up projects (Grid3+ etc.) as well as ATLAS data challenge productions. Grid 2003 was a project sponsored by iVDGL to demonstrate operating a very large grid at the 2003 Supercomputing Conference in Phoenix, Arizona. The Grid 2003 project assembled a grid of approximately 2500 CPUs from about 25 computer centers throughout the US and Korea. The IU Tier 2 center made one of the larger contributions of CPU cycles in the demonstration (392,978 cpu-minutes). Following the GRID 2003 demonstration, the grid assembled has remained active and continues to be used by a number of scientific collaborations for research. The users of AVIDD during Grid 2003 included high energy physics experiments (ATLAS, CMS, BTeV), astronomers (SDSS), and gravitational wave experiments (LIGO). Also using GRID 2003 were biologists and computer scientists [22].

Exotic Mesons Confinement Mechanisms (Alex Dzierba, Department of Physics, IU Bloomington)

Physicists associated with the light-quark physics effort at Indiana University include experimentalists and theorists. The physics goal of this effort is an understanding of the mechanism – the confinement mechanism – in quantum chromodynamics (QCD) responsible for holding together the constituents (quarks and gluons) of a large class of elementary particles called hadrons. So strong is this binding that free quarks and gluons have never been observed. Understanding the confinement mechanism is recognized as being one of the outstanding fundamental problems in physics.

Hadrons number in the hundreds and include baryons (of which protons and neutrons are examples) and mesons (the π and K mesons are examples). Baryons consist of three quarks and mesons of a quark and anti-quark. If our current notions of how confinement arises are correct, a new class of mesons, called exotic mesons, should exist whose quantum numbers cannot be explained by a simple quark-antiquark configuration. These new mesons can be produced in collisions and observation of their decays leads to a determination of their mass, lifetime (decay width) and other properties including spin and parity. To identify the exotic mesons among the myriad of normal mesons that are produced requires a complex quantum-mechanical amplitude analysis to describe the angular decay characteristics. Large statistical samples (tens of millions of events) are required to make the identification. Each event includes from four to eight or more decay products (charged particles and/or photons) each of which in turn is described by three components of momentum and one of energy. These components are reconstructed from detector signals. This reconstruction is followed by a sophisticated description of the decay in terms of quantum-mechanical amplitudes. The amplitudes are determined using maximum likelihood techniques. Along with all this is the necessity to generate Monte Carlo events with even larger statistics (by an order of magnitude) to understand the corrections needed to take into account imperfect detector response. Multiple passes

through the data are necessary to study systematic effects due to detector response, reconstruction criteria and model assumptions in the final analysis.

The Indiana group led an experiment at Brookhaven National Lab in 1990-1995 that resulted in a large data set including decays into three π mesons. The statistical sample size collected at that time could not be fully analyzed using computational power available at the time but a small subset was analyzed. A tantalizing signal for an exotic meson was obtained. Our group also leads a new experimental program called GlueX that will use beams of photons to produce exotic mesons at Jefferson Lab. This project, and the associated energy upgrade of the Jefferson Lab accelerator needed to accomplish this physics, were recently identified by the Secretary of the US Department of Energy as one of several near-term high priority projects as part of the new 20-year plan of facilities. The analysis of the Brookhaven data is a natural step in the direction of analysis of data from GlueX that will exceed the Brookhaven experiment by several orders of magnitude.

Complex Amplitude Analysis of Large Data Sets (Alex Dzierba, Department of Physics, IU Bloomington)

Indiana physicists are also working with particle physicists from Cornell University and computer scientists from the local Community Grids Lab of Indiana's Pervasive Technology Lab, to establish a Physics Analysis Center to develop efficient techniques and develop the phenomenology to perform complex amplitude analysis of large data sets.

Since AVIDD became operational the group has used approximately 10,000 cpu hours to start the analysis of the 3π sample. The first reconstruction pass took about one month compared to the 10 months required for a similar previous reconstruction using all available resources at the time. Monte Carlo modeling has consumed about 30,000 cpu hours. In parallel, members of the local analysis team have been making improvements to the algorithms that have yield a factor of 500 in efficiency compared to older algorithms that depended on storing intermediate results on disk (when computation speeds were the limiting factor). This is truly textbook quality physics of unprecedented precision and indeed we plan to publish new results on well-known mesons soon as we march along in extracting information about possible exotics. All this now possible with the use of the AVIDD facility.

MINOS Experiment (Brian Rebel and Stuart Mufson, Department of Physics, IU Bloomington)

The High Energy Astrophysics Group has used the AVIDD cluster extensively as part of its effort on the MINOS experiment, which is based at Fermilab. The cluster has proved to be a valuable tool for generating monte carlo events, testing event reconstruction software, and processing monte carlo events and data from the experiment. The AVIDD cluster has allowed the processing of several months worth of data in the span of about a

week, allowing the group to test many variations on event reconstruction and data processing.

MINOS is an experiment designed to search for neutrino oscillations in the region of parameter space that best accounts for the atmospheric neutrino results obtained by the SuperKamiokande experiment in Japan. MINOS will study muon neutrinos directed to northern Minnesota from Fermilab. The experiment will explore the neutrino mass region below 1 eV2. The collaboration appointed me, Stuart Mufson, as one of six data coordinators for the MINOS experiment. My data analysis responsibilities include underground muon physics with the MINOS detector. In addition, Brian Rebel and I have been responsible for the atmospheric neutrino analysis that searches for evidence for neutrino oscillations in cosmic rays.

Chromoelectric Flux Tubes (Patrick Bowman and Adam P. Szczepaniak, Department of Physics, IU Bloomington)

The interactions between quarks, anti-quarks and gluons, the fundamental constituents of protons and neutrons are described by a part of the Standard Model known as Quantum Chromodynamics or QCD. Unlike other known forces these interactions are strong and therefore not amendable to techniques which otherwise have proven powerful in understanding fundamental interactions, such as interaction of electromagnetic forces with matter or weak radioactivity. The most reliable, first principles analysis of strong interaction phenomena require large computer simulations known as QCD lattice simulations. Alternatively approximate modes of quark-gluon systems can be devised and tested against the more powerful albeit expensive simulations or the experimental data. One such test involves studies of patterns of energy distribution near static sources, for example a quark and anti-quark pair. This is an analog to the electric dipole in electrostatic in QCD, however, one expects the field lines to be collimated into thin flux-tubes instead of being dispersed as it is the case of the electric dipole. The figure below shows such flux-tubes emerging in a quark-anti-quarks system. This is not a standard QCD lattice computation; the theory has been approximated to take advantage of special properties of the Coulomb gauge quantization. The computations have been performed on the AVIDD cluster using 100 nodes running in parallel to independently compute filed density at a single pint in the plane (x-z) containing the quark-antiquark pair [15].

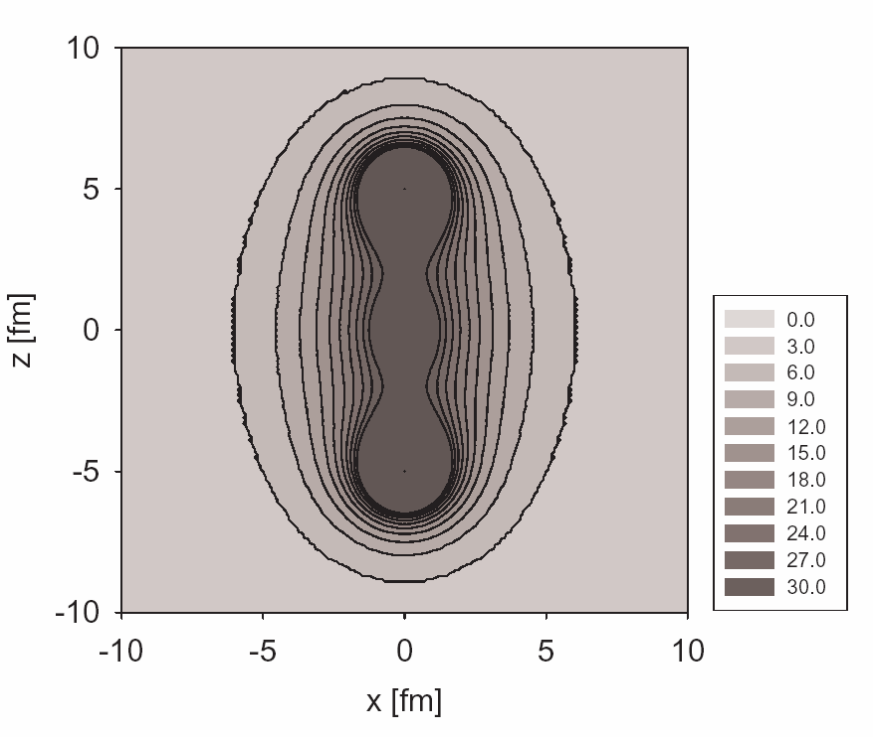


Figure 6. Coulomb energy density profile (in arbitrary units) for a quark-antiqurk system located along the z-axis 10^{-14} m apart along the z-axis.

Geology

Seismic Research (Gary Pavlis, Department of Geology, IU Bloomington)

Many seismic analysis applications are data driven. As the physical data ranges from crystals to continents, datasets range to gigabytes. Pavlis' dbpmel and pmelgrid applications perform simultaneous estimation of improved earthquake locations and ensembles of path corrections. Other applications such as time-series processing or geophysical data mapping, involve the same basic steps, reading data of a certain type, processing it, then writing something out.

We have created a generic mechanism to support the data-driven applications. The library is object oriented and is implemented in C++. To use the library in his application, a user only needs to implement certain abstract data types following C++'s data type inheritance rules, based on the intended behavior of their own applications. We were able to implement and test the library on AVIDD using pmelgrid, with satisfactory results.

Geography, Atmospheric Sciences, and Geographical Information systems

Sara Pryor (Department of Geography, Atmospheric Science Program, IU-Bloomington)

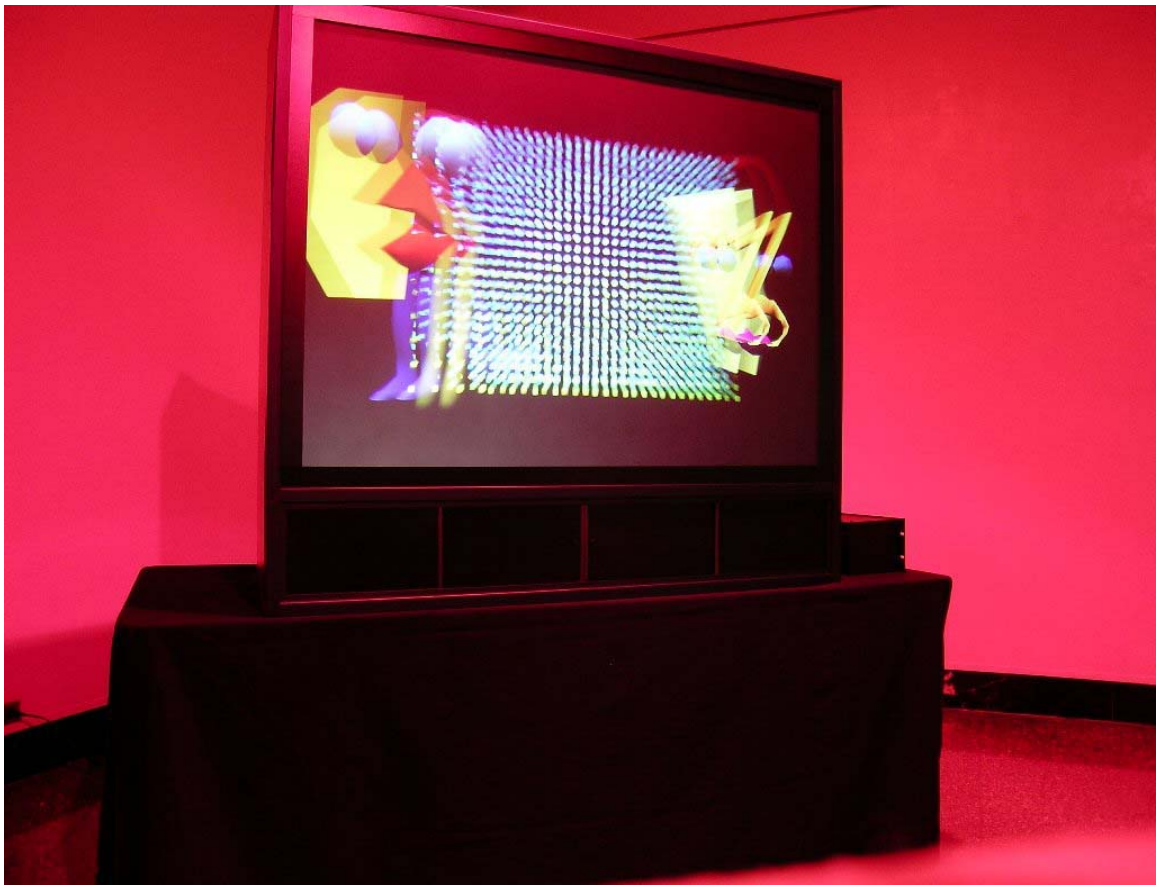
The AVIDD system was used primarily to conduct research into development of global climate simulations with a focus on prediction of near-surface wind speeds. Preliminary research conducted in part using this computational resource is forming the basis of a joint proposal with Iowa State University to NSF Climate Dynamics division to undertake development of wind speed projections for the twenty-first century for the contiguous USA [24,25].

GIS Datasets (Anna Radue, University Information Technology Services, IU Bloomington)

The Data Management Support Group (DMS) of UITS has installed Lizardtech's Geoexpress 3.1 on the AVIDD system. The National Agriculture Imagery Program (NAIP) will soon release one meter color aerial photographs for the entire state of Indiana. To facilitate the network transfer of these 135 MB Geotiff images, DMS plans to convert the files to MrSID generation 3 compressed format using Geoexpress. The AVIDD system's large disk space is needed for processing this 400 GB dataset. Because Geoexpress has both an application license and a data cartridge license, the installation on AVIDD this past July was a pilot project to confirm that DMS could install the software on the system. At this time DMS is waiting delivery of the NAIP imagery from the Natural Resource Conservation Service office in Fort Worth, Texas.

---

5) Beyond Science: Artistic Outcomes



A John-e-Box has been provided (through university funds independent of the AVIDD project) to Assistant Professor of Fine Arts Margaret Dolinsky (http://dolinsky.fa.indiana.edu/). Her CAVE artwork has been successfully demonstrated on the new VR display. Together with AVL staff, she is currently investigating new ways for the audience to interact in the virtual experience, as with sound-activated 3D scenes directly controlled by the participants' voices. The John-e-Box's portable technology has enabled installations of her artwork in the IU Art Museum and in the School of Fine Arts Gallery. The flexibility and multi-disciplinary utility of the John-e-Box has already been demonstrated through its use as a platform for artistic works based on virtual reality technology (see Figure 4). In addition, University Information Technology Services is discussing the possible use of this device at the Indianapolis Museum of Art [5].

Figure 7. John-e-Box installed at the IU Art Museum

6) Economic Outcomes

Additional Grants

The AVIDD facility has been leveraged in grant applications to the NSF and other funding agencies. Successful grants that have leveraged the AVIDD facility include the following:

- IP-Grid. M.A. McRobbie et al. 2003. Indiana Universitys successful NSF proposal to become one of the new participants in the NSF-funded Teragrid.
- Center of excellence in Homeland Security. A. Churvedi et al. 2003. Indiana University is a partner in this successful Purdue-lead proposal to the State of Indiana 21st Century fund.
- Grid and Data Intensive Computing in the Life Sciences . C.A. Stewart et al. 2002. Thus IBM SUR grant provided the linux cluster installed at IU Northwest (Gary) as part of the AVIDD project.
- Creation of the Curation and Alignment Tool for Protein Analysis (CATPA): a system for organizing and sharing information about protein families. M. Dalkilic, P. Cherbas, C.A. Stewart. Joint development contract with IBM, Inc.
- Informatics Core for the Collaborative Initiative on Fetal Alcohol Spectrum Disorders (FASD). C.A. Stewart et al. NIH U24 core grant.

The AVIDD facility is mentioned as a critical piece of infrastructure in several other pending grant applications. Given computing resources that will greatly expand Indiana University□s ability to conduct computationally intensive research, IU has been very aggressive and successful in pursuing funding for research that will use the AVIDD facility as a computational and data analysis resource.

Technology Transfer

The development of the John-e-Box and its successful commercialization has been a particular highlight of the AVIDD project. At the time the proposal to create AVIDD was written, one prototype John-e-Box existed (see Figure 1). While it was determined by ARTI (the Advanced Research and Technology Institute – responsible for IU's technology transfer processes) that there was no patentable technology in this design, the design has been successfully licensed to the private sector. John-e-Boxes are now part of the product offerings of CAE-Net (http://www.cae-net.com/), a central Indiana engineering and technology company. Transferring this technology was definitely facilitated by the AVIDD project, and by the same token the licensing of this design aided the AVIDD project. It is both cheaper and quicker to purchase the commercially produced John-e-Boxes than to make them in-house.The John-e-Box is now commercially available as a regular product offering of CAE-Net. CAE-Net plans to sell this device to research and educational institutions, and believes that it is likely to be very popular for use in museums [4].

Open Source Software

Several of the investigators' codes required substantial porting and optimization in the course of the AVIDD project: those of Hurley, Gardner, Reilly, Palikal, Pryor and Pavlis deserve particular mention. In addition, several other research groups' codes have been

ported to AVIDD. Not all of these codes are expected to become open-source, and not all which are, have matured to the point of general release.

Codes which have been significantly developed in the course of the AVIDD project, and which have been publicly released as open-source, include the following:

- **Nbpack** implements the Barnes-Hut hierarchical tree method for classical N-body simulations. NBpack includes functions for computing potential energies and forces for gravitational, coulomb (electrostatic) and Lennard-Jones interactions; intended application areas include gravitational problems in astrophysics, molecular dynamics problems in physics, chemistry and biology, plasma physics and nuclear collision theory.
- **GeneIndex** is a very efficient program for finding in a DNA sequence the frequencies and positions of all words of a given length.
- **fastDNAml** computes the likelihood of various phylogenetic trees, starting with aligned DNA sequences from a number of species. First written many years ago, important new modeling capabilities have been added to fastDNAml. This application, to run in parallel on heterogenous and widely distributed systems, was used in the demonstration which won the HPC Challenge at SC2003 [18], [19].
- PENELOPE is a highly accurate radiation transport algorithm, available from ORNL. Used by the IU Medical School's Department of Radiation Oncology in investigating accurate treatment planning algorithms, PENELOPE was parallelized and returned to ORNL for redistribution.

7) Human Resources Outcomes

This project has provided research and training in the area of high performance computing, particularly as regards Linux clusters, for roughly one dozen professional staff, and more than two dozen faculty and graduate student researchers.

The highlight of the training and development effort related to AVIDD was the creation of a new course to the curriculum of the IU School of Informatics entitled P573 – Scientific Computing. This course was offered as a first-year graduate class, but was also open to upper-level undergraduates. This class was developed and taught by Dr. Zdzislaw Meglicki, PhD. It was offered to graduate and undergraduate students on the Bloomington, Indianapolis and Gary campuses during Indiana Universitys Fall 2003 semester via multicast. The intent of the course was to acquaint students with techniques for managing very large data sets within distributed-memory cluster systems. The course covered the basic mechanics of parallel computing with special emphasis on parallel IO (MPI-IO and HDF5). Parallel databases and data mining techniques were also covered. The AVIDD clusters in Bloomington, Indianapolis and Gary were used for class assignments. More than 20 students completed this class, and additional offerings are

planned for the future. The offering of this class on the Gary campus is particularly important, as the IU Northwest campus in Gary serves a larger percentage of students from traditionally underserved groups than any other campus of Indiana University.

Indiana University's execution of the AVIDD project has had important effects in helping create the workforce of tomorrow – and attract that workforce from the full richness of America's diverse population.

Of particular note is University Information Technology Services' Minority Intern program. Two undergraduate interns in this program have had extensive involvement in use of the AVIDD facility.

This project has provided research and training in the area of high performance computing, particularly as regards Linux clusters, for roughly one dozen professional staff, and more than two dozen faculty and graduate student researchers.

References

[1] I-Light: Indiana's Optical Fiber Initiative, http://www.i-light.org.

[2] Torque Resource Manager, http://www.supercluster.org/torque.

[3] The John-e-Box: Realizing the promise of affordable and accessible visualization and virtual reality systems, http://avl.indiana.edu/.

[4] John-e-Box Visualization, http://www.cae-net.com/John-E-Box_Brochure.pdf.

[5] The Indianapolis Museum of Art, //www.ima-art.org/.

[6] The Top500 Supercomputer Sites, http://www.top500.org/.

[7] The Advanced Light Source, http://www-als.lbl.gov/.

[8] Baik, Mu-Hyun, Richard A. Friesner and Stephen J. Lippard, "cis-{Pt(NH3)2(L)}2+/+ (L = Cl, H2O, NH3) Binding to Purines and CO: Does π-Backdonation Play a Role?" Inorg. Chem., 42, 8615-8617 (2003).

[9] Baik, Mu-Hyun, http://mypage.iu.edu/~mbaik/research/metalloproteins.htm.

[10] Baik, Mu-Hyun, http://mypage.iu.edu/~mbaik/research/cisplatin.htm.

[11] Barrett, Brian, Jeff Squyres, and Andrew Lumsdaine, Integration of the LAM/MPI environment and the PBS scheduling system, in Proceedings, 17th Annual International Symposium on High Performance Computing Systems and Applications, Quebec, Canada, May 2003.

[12] Basuli, Falguni, Brad C. Bailey, John C. Huffman, Mu-Hyun Baik and Daniel J. Mindiola, “Terminal and Four-Coordinate Vanadium(IV) Phosphinidene Complexes. A Pseudo Jahn-Teller Effect of Second Order Stabilizing the V-P Multiple Bond", Journal of the American Chemical Society 126 (7), 2004.

[13] Bertrand, Felipe, Kenneth Chiu, Yongquan Yuan, and Randall Bramley, “An Approach to Parallel MxN Communication”. Proceedings of the Los Alamos Computer Science Institute (LACSI) Symposium, Sante Fe, NM, October 2003.

[14] Bertrand, Felipe, Randall Bramley , DCA: A Distributed CCA Framework Based on MPI, accepted for 9th International Workshop on High-Level Parallel Programming Models and Supportive Environments, Sante Fe, NM, April 26-30, 2004.

[15] Bowman, Patrick , and Adam Szczepaniak, "Chromoelectric Flux Tubes", e-Print Archive: hep-ph/0403075, March 2004.

[16] Cruise, Robert, et al., □INGEN□s advanced IT facilities: The least you need to know□, UITS pamphlet, January 2002.

[17] Govindaraju, Madhusudhan, Sriram Krishnan, Kenneth Chiu, Aleksander Slominski, Dennis Gannon, Randall Bramley , Merging the CCA Component Model with the OGSI Framework, Proceedings of CCGrid-2003, Tokyo, Japan, March 2003.

[18] Keller, Rainer, et al. , Global analysis of Arthropod evolution., SC2003 Technical Program, Phoenix, AZ, November 2003.

[19] Keller, Rainer, et al., http://www.hlrs.de/news-events/2003/sc2003/HPC-CHALLENGE/, November 2003.

[20] Krishnan, Sriram, Randall Bramley, Dennis Gannon, Rachana Ananthakrishnan, Madhusudhan Govindaraju, Aleksander Slominski, Yogesh Simmhan, Jay Alameda, Richard Alkire, Timothy Drews, and Eric Webb, The XCAT Science Portal, Journal of Scientific Programming, Volume 10, Number 4, pp. 303--317, 2002.

[21] Kumar, K., M. Palakal, S. Mukhopadhyay, M. Stevens, and H. Li, BioMap: Toward the Development of a Knowledge Base of Biomedical Literature, Proceedings of the ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.

[22] Luehring, Frederick, http://www.ivdgl.org/grid2003/index.php.

[23] Plale, Beth et. al, □Real Time Response to Streaming Data on Linux Clusters□, presented at the Linux HPC Revolution conference in St. Petersburg, Florida , October , 2002.(http://www.linuxclustersinstitute.org/Linux-HPC-Revolution/Archive/PDF02/25-Plale_B.pdf).

[24] Pryor S.C. and Barthelmie R.J., "Long term variability of flow over the Baltic", International Journal of Climatology 23, 2003.

[25] Pryor S.C. and Barthelmie R.J., Past and future wind climates: Flow regimes in a non-stationary climate. The science of making torque from wind - EWEA special topic conference, Delft, April 2004.

[26] Samuel, J.V., Peebles, C.S., Noguchi, T., Stewart, C.A., "Gauging IT Support Strategies: User needs then and now." (PDF) Proceedings of SIGUCCS 2002 Conference, Providence, RI, November 20-23, 2002.

[27] Samuel, J.V., Wilhite, K.J., Stewart, C.A.. "Getting More for Less: A Software Distribution Model." (PDF) Proceedings of Educause, Atlanta, GA, October 1-4, 2002.

[28] Sankaran, Sriram, Jeffrey M. Squyres, Brian Barrett, and Andrew Lumsdaine, Parallel Checkpoint/Restart for MPI Applications, Open Systems Laboratory, Indiana University, in LACSI Symposium, Lawrence Berkeley National Laboratory, October 2003.

[29] Sankaran, Sriram, Jeffrey M. Squyres , Brian Barrett, and Andrew Lumsdaine, Checkpoint/Restart System Services Interface (SSI) Modules for LAM/MPI API Version 1.1.0 / SSI Version 1.0.0, http://www.lam-mpi.org/ , Open Systems Laboratory Pervasive Technologies Labs, Indiana University, CS TR578.

[30] Squyres, Jeffrey M., Brian Barrett, and Andrew Lumsdaine, The System Services Interface (SSI) to LAM/MPISSI Version 1.0.0, http://www.lam-mpi.org/, Open Systems Laboratory Pervasive Technologies Labs, Indiana University, CS TR575.

[31] Squyres, Jeffrey M., Brian Barrett, and Andrew Lumsdaine, Boot System Services Interface (SSI) Modules for LAM/MPI API Version 1.0.1 / SSI Version 1.0.0 , http://www.lam-mpi.org/, Open Systems Laboratory Pervasive Technologies Labs, Indiana University, CS TR576.

[32] Squyres, Jeffrey M., Brian Barrett, and Andrew Lumsdaine, MPI Collective Operations System Services Interface (SSI) Modules for LAM/MPI API Version 1.1.0 / SSI Version 1.0.0. http://www.lam-mpi.org/ Open Systems Laboratory Pervasive Technologies Labs, Indiana University, CS TR577.

[33] Squyres, Jeffrey M., Brian Barrett, and Andrew Lumsdaine, Request Progression Interface (RPI) System Services Interface (SSI) Modules for LAM/MPI API Version

1.1.0 / SSI Version 1.0.0, http://www.lam-mpi.org/,Open Systems Laboratory Pervasive Technologies Labs, Indiana University, CS TR579.

[34] Squyres,Jeffrey M, and Andrew Lumsdaine, A Component Architecture for LAM/ MPI, Open Systems Lab, Indiana University, in Proceedings, Euro PVM/MPI, October 2003.

[35] Stewart, C.A., D. Hart, R.W. Sheppard, H. Li, R. Cruise, V. Moskvin, L. Papiez.,. Parallel computing in biomedical research and the search for peta-scale applications. Proceedings of Parco2003, Sept. 2-5, 2003, Dresden, Germany, September 2-5, 2003.

[36] Stewart, C.A., Repasky, R., Hart, D., Papakhian, M., Shankar, A., Wernert,E., Arenson, A., and G. Bernbom, "Advanced Information Technology Support For Life Sciences Research," (PDF) Proceedings of SIGUCCS 2003, San Antonio, TX, Sept. 21-24, 2003.

[37] Stewart, C.A., and R. Repasky, "High performance computing for university biomedical research: a successful implementation," (PDF) Proceedings of BioITWorld Conference and Expo, Boston, MA, March 2003.

[38] Stewart, C.A., Roskies, R.Z., Subramaniam, S., "Opportunities for Biomedical Research and the NIH through High Performance Computing and Data Management," (PDF) CASC White Paper, January 2003.

[39] Stewart, C.A., and D.P. Moffett (eds), "Proceedings of the I-Light Applications Workshop,"http://www.i-light.iupui.edu/proceedings.html ", Indianapolis, IN, December 4, 2002.

[40] Stewart, C.A., "Implementing Advanced IT Facilities for the Indiana Genomics Initiative." (PDF) Proceedings of HPC@IDC meeting, Taos, NM, April 2002.

[41] Wang, Peng, et al., "1 TFLOPS Achieved with a Geographically Distributed Linux Cluster", In : Laurence T. Yang, Guest Editor. High performance computing: Paradigm and Infrastructure", John Wiley & Sons, Inc., 2004.

Appendix 1. Presentations related to AVIDD

Daniel Lauer and George Turner, "The AVIDD Linux Clusters", booth at annual IU LinuxFest event, IUB, March 30, 2004.

AVIDD, and in particular the John-e-Box, was featured in IU's display at the NSF ESTME event, March 15-19, 2004.

Hasan Akay, “Advances in Parallel Metacomputing of Solid-Fluid Interaction Problems”, poster session/demo featuring visualizations on the John-e-Box, I-Light Symposium 2004, IUPUI, March 9, 2004.

The AVIDD project was featured in the IU-lead 'Research in Indiana' display at SC2003, Phoenix, AZ, November 17-20, 2003.

George Turner, Peng Wang et al, “1 TeraFLOPS achieved with distributed Linux cluster”, poster session at SC2003, Phoenix, AZ, November 18, 2003.

Craig Stewart, "Global analysis of arthropod evolution", presented at SC2003, Phoenix, AZ, November 18, 2003.

Craig Stewart, “Computational Biology” tutorial presented at SC2003, Phoenix, AZ, November 16, 2003.

Mary Papakhian, Daniel Lauer, Stephen Simms and George Turner,“Linux in High Performance Computing at IU”, keynote talk featuring AVIDD at annual IUPUI LinuxFest event, April 24, 2003.

Mary Papakhian, Matthew Allen, Daniel Lauer, Stephen Simms and George Turner,“Linux in High Performance Computing at IU”, keynote talk featuring AVIDD at annual Bloomington LinuxFest event, April 16, 2003.

Display featuring AVIDD at IU's annual IT awareness campaign "Making IT Happen" at IUB, IUPUI, IUN, and IUK campuses, March, 2003.

Daniel Lauer and George Turner, “The AVIDD Linux Clusters”, booth at annual IU LinuxFest event, March 30, 2004.

Anurag Shankar, "Handling the Data Deluge at Indiana University", CIC TechForum 2003, Madison, WI, October 30, 2003.

Andrew Arenson, "The Centralized Life Sciences Data (CLSD) Service", UITS IUB Infoshare , October 20, 2003.

Andrew Arenson, "The Centralized Life Sciences Data (CLSD) Service", UITS IUPUI Infoshare , October 22, 2003.

Andrew Arenson, "Resources for Bioinformatics at Indiana University", Presented to School of Informatics class, IUPUI, October 7, 2003.

Anurag Shankar, "High Performance Storage at Indiana University", UITS Infoshares , IUB, September 30, 2003.

Anurag Shankar, "High Performance Storage at Indiana University", UITS Infoshares, IUPUI, October 1, 2003.

Andrew Arenson, "The Centralized Life Sciences Data Service (CLSD) and supported applications at Indiana University", Presented at the IBM Institute of Innovation inaugural event, IUB, September 25, 2003.

David Hancock, "High Performance Computing Resources at IU", UITS IUPUI Infoshare, July 29, 2003.

Stephen Simms, "High Performance Computing Resources at IU", UITS IUB Infoshare, July 30, 2003.

Anurag Shankar, "HPSS at Indiana University: A Site Report", Presented at the 2003 HPSS User Forum, Asheville, NC, June 19, 2003.

Andrew Arenson, "The Centralized Life Sciences Data (CLSD) Service", UITS Infoshare, IUPUI, June 3, 2003.

Andrew Arenson, "The Centralized Life Sciences Data (CLSD) Service", Center for Genomics and Bioinformatics Roundtable, IUB, May 29, 2003.

Anurag Shankar, "High Performance Storage at Indiana University", UITS Infoshare, IUPUI, May 28, 2003.

Anurag Shankar, "High Performance Storage at Indiana University", UITS Infoshare, IUB, May 27, 2003.

Eric Wernert, "An Overview of Visualization Resources at IU - presentation to the Center for Genomics and Bioinformatics." May, 2003.

Greg Cook, "A Survey of Video Compression Technologies" - a series of presentations given to staff of the AVL and the Indiana Center for Biomedical Microscopy, May 2003.

Andrew Arenson, "Resources for Bioinformatics at Indiana University", Computational Biology class, IUPUI, April 28, 2003.

E. Chris Garrison, "High Performance Storage at Indiana University", UITS Infoshare, IUPUI, April 16, 2003.

Anurag Shankar,. "High Performance Storage at Indiana University", UITS Infoshare, IUB, April 15, 2003.

Stephen Simms, "High Performance Computing Resources at IU", UITS Infoshare, February 19, 2003..

Craig Stewart, "INGEN's advanced IT facilities." Presented to Dept. of Psychiatry, IU School of Medicine. February, 2003.

Craig Stewart, "INGEN's advanced IT facilities." Presented to Dept. of Physiology, IU School of Medicine. February, 2003.

Craig Stewart and Andrew Arenson, "Discovery Link at IU: The Centralized Life Science Data (CLSD) Service", IBM/Lilly/IU Data Integration Conference, IUPUI, January 17, 2003

Anurag Shankar, "Research Data Storage Services at IU", I-Light Workshop, IUPUI, December 4, 2002.

Dave Hart and Mary Papakhian, "High Performance Computing Resources at Indiana University", presentation and panel discussion, I-Light Workshop, IUPUI, December 4, 2002.

Andrew Arenson, Mary Papakhian, Richard Repasky, and Eric Wernert, "Biomedical Applications over I-Light", poster session at I-Light Workshop, December 4, 2002.

Anurag Shankar, "Building and Supporting a Massive Data Infrastructure for the Masses", Presented at the SIGUCCS 2002 Conference, Providence, RI, November 21, 2002.

Andrew Arenson, Mary Papakhian, Richard Repasky, and Anurag Shankar, "Proteomics Software and Services at University Information Technology Services", poster session at Indiana Proteomics Symposium, November 15, 2002.

Craig Stewart, "INGEN's advanced IT facilities." Presented to Division of of Nephrology, IU School of Medicine. November, 2002.

Computational Biology. panel moderated by Craig Stewart at SC2002 conference, Baltimore, MD,. November, 2002.

Mary Papakhian, "High Performance Computing Resources at IU", UITS IUB Infoshare, October 17, 2002.

Dave Hancock, "High Performance Computing Resources at IU", UITS IUPUI Infoshare, October 19, 2002.

Craig Stewart, "INGEN's advanced IT facilities." Presented to Indiana Health Industry Forum meeting, October 2002.

Richard Repasky, "UITS Bioinformatics & Supercomputing Facilities." Biostatistics, Indianapolis, September 2002.

Richard Repasky, "How IU supercomputers can help your research." Bioinformatics Roundtable, Bloomington, Center for Genomics and Bioinformatics, September 2002.

Robert Cruise, "Adaptive Monte Carlo Methods," Department of Radiation Oncology Colloquium, IU School of Medicine, June 14, 2002.

Craig Stewart, "INGEN's advanced IT facilities." Presented to Wells Center, IU School of Medicine. June, 2002.

Richard Repasky, "High-volume BLAST searching on IU supercomputers." Bioinformatics Roundtable, Bloomington, Center for Genomics and Bioinformatics. May 2002

Michael Boyles and Mary Papakhian, "High-Performance Computing Resources at IU" - IUPUI Summerfare course, June 8, 2002.

Craig Stewart, "Implementing Advanced IT facilities for Indiana Genomics Initiative." Presentation to HPC@IDC User Forum, April, 2002.

Eric Werner, "Virtual Reality Methods for Medical Education." Presented at the Learning Resource Center conference hosted by the School of Nursing; supplemented by AVL lab demonstrations. April, 2002.

Craig Stewart. "INGEN's advanced IT facilities." Presented to Department of Endocrinology, IU School of Medicine. March, 2002.

AVL Open House - events at IUPUI and IUB showcasing new hardware and software technologies, including John-e-Box, March, 2002.

Eric Wernert, "Overview of Visualization and Tele- collaboration Technologies." Presented to staff of the Brown Cancer Center at the University of Louisville. March, 2002.

Craig Stewart, "INGEN's advanced IT facilities." Presented to Department of Medical Genetics, IU School of Medicine. January, 2002.

Craig Stewart, University Information Technology Services/INGEN IT Core display booth at Iconomy conference, Indiana Marriot, January, 2002.

Appendix 2. Press Announcements related to AVIDD

Indiana University news article on the development, licensing and deployment of the John-e-Box, http://newsinfo.iu.edu/news/page/normal/1175.html, November 4, 2003.

Indiana University press release describing the benchmark results on combined AVIDD-B and AVIDD-I clusters, http://www.indiana.edu/~uits/cpo/avidd062403, June 24, 2003.

Indiana University press release announcing the unveiling of a novel computing facility for the Analysis and Visualization of Intrument-Driven Data, http://www.indiana.edu/~uits/cpo/avidd032603, March 26, 2003.

HPCWire article 72683 describing IU's installation and deployment plans for the AVIDD facility, November 29, 2001.

HPCWire article 101487 describing IU's AVIDD NSF MRI award-winning proposal to create the AVIDD facility, November 2, 2001.

Press conference announcing NSF award, focused particularly on its impact on diversity in the HPC community, http://www.indiana.edu/~uits/cpo/mri/index.html, October 2001.