TeraScope: Distributed Visual Data Mining of Terascale Data Sets over Photonic Networks

Chong (Charles) Zhang, Jason Leigh, Thomas A. DeFanti
Electronic Visualization Laboratory (EVL),
University of Illinois at Chicago
cavern@evl.uic.edu

Marco Mazzucco, Robert Grossman,
Laboratory for Advanced Computing (LAC),
University of Illinois at Chicago

"Where the telescope ends, the microscope begins.
Which of the two has the grander view?" -Victor Hugo

Abstract

TeraScope is a framework and a suite of tools for interactively browsing and visualizing large terascale data sets. Unique to TeraScope is its utilization of the Optiputer paradigm to treat distributed computer clusters as a single giant computer, where the dedicated optical networks that connect the clusters serve as the computer's system bus. TeraScope explores one aspect of the Optiputer architecture by employing a distributed pool of memory, called LambdaRAM, that serves as a massive data cache for supporting parallel data mining and visualization algorithms.

1 Introduction

Areas of research such as Geoscience, Astronomy, and High Energy Physics are routinely producing terabytes, and soon, petabytes of data from direct data gathering, data post processing, and simulations. Algorithmic detection of hidden patterns within these large data sets has been the focus of data mining [6]. Visualization used in this context (often referred to as Visual Data Mining) has been valuable as a way to verify the detected patterns; and in particular, for when algorithmic specifications of the patterns are difficult to derive [2, 3, 8, 9, 10, 15]. In the latter case user-interfaces that allow one to interactively browse, query and visualize enormous data sets need to be developed.

The work described in this paper is motivated by several emerging trends. Firstly scientific databases are becoming highly distributed. Secondly the cost of high speed networking is increasing at a rate far exceeding Moore's Law- network bandwidth is doubling every 8 months whereas processors are doubling in speed every 18 to 24 months. This means that computers, rather than the networks are the bottleneck. Thirdly there is an increasing need and potential, facilitated by these high speed networks, for scientists to publish terabyte data sets on the Web in a manner similar to the way most netizens can create Web pages, so that researchers can make new discoveries by combining data from previously disparate disciplines. For example by correlating data from the World Health Organization with data from the National Center for Atmospheric Research, one could potentially understand how weather patterns influence the spread of diseases.

The Optiputer is a National Science Foundation funded project intended to exploit these trends by interconnecting distributed storage, computation, and visualization resources using extremely high speed photonic networks[13]. The important difference between this and classical Grid computing is that in this new model, the optical networks serve as the system bus for a potentially planetary-scale computer; and compute clusters taken as a whole, serve as the peripherals in the computer. For example, a cluster of computers with high performance graphics cards would be thought of as a single giant graphics card in this

context. In the Optiputer concept, we refer to compute clusters as LambdaNodes to denote the fact that they are connected by multiples of light paths (often referred to as Lambdas) in an optical network. Each computer in a LambdaNode is referred to as a nodule, and collections of LambdaNodes form a LambdaGrid.

TeraScope is an experimental visual data mining toolkit intended to take advantage of the Optiputer paradigm. This paper describes the prototype that was developed and demonstrated at the IGrid 2002 conference in Amsterdam (www.igrid2002.org). Furthermore, this paper describes LambdaRAM, a high performance cache, for the Optiputer.

2 TeraScope

The vision for TeraScope is to provide a way to fluidly work with massive data sets as interactively as one would work with a spreadsheet on a laptop. The goal is not necessarily to massively parallelize visualization algorithms so that a terabyte of points can be plotted. The goal is to use parallel algorithms to process terabyte data sets to produce visual summaries (which we call TeraMaps) to help the user locate regions that are most interesting to them. Once the area of interest has been identified, modest visualization algorithms can be used to depict the derived subset of the data (which we call TeraSnaps). TeraScope has three goals: 1. to develop a software architecture that will allow a variety of data mining algorithms to be easily integrated into the Optiputer framework; 2. to develop ways to create meaningful TeraMaps; 3. to provide browsing interfaces and 2D and 3D visualization tools that are intimately connected to data mining algorithms.

Before we can begin describing the software architecture of TeraScope, we must first explain the Optiputer hardware framework that is currently driving TeraScope.

2.1 The Hardware Architecture

Figure 1A depicts a typical architecture for a modern day PC. Highlighted in yellow are the caches that are a routine part of the components of the architecture. For example, the graphics card has onboard fast graphics RAM, the CPU has L1/L2/L3 caches, and so on. Data from the disk are transferred to the CPU via the PCI bus, whereas data from the CPU is transferred to the graphics card via AGP. Figure 1B shows one possible configuration of the Optiputer mimicking the standard PC architecture except using clusters of computers, optical switches, and multi-gigabit network connections. A similar configuration to this was used for IGrid, although this particular layout is our present configuration. Illustrated are three classes of LambdaNodes. All the LambdaNodes are connected using gigabit network interface adapters. The Lambda**Data**Node is primarily a cluster with large RAIDed disks. The Lambda**Compute**Node is a cluster with large amounts of physical memory and multiple CPUs. The Lambda**Visualization**Node is a cluster with high-end commodity graphics cards (such as the Nvidia Geforce 4 Ti). All network links are presently 1 Gb/s except for the link from Amsterdam to StarLight (at 10Gb/s).

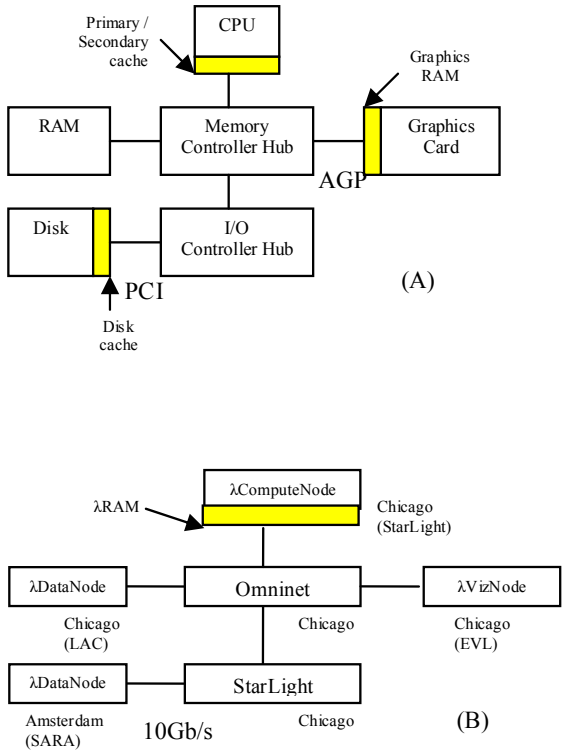


Figure 1. Comparison of a modern CPU architecture with one possible configuration of the Optiputer.

Starlight is a project managed by the University of Illinois at Chicago, to provide an IP-over-Dense Wave Division Multiplexing (DWDM) peering point for national and international optical networks. The goal is to develop a "petri dish" for growing an experimental, photonic Grid whereby clusters of computing resources can directly "dial-up" lambdas between them and use the extreme quantities of bandwidth (on the order of 1-10 Gigabits/s) as a long distance system bus [18]. OMNInet is a project operated by Northwestern University and supported by Nortel Networks, SBC Communications Inc. and Ameritech to assess and validate next-generation photonic technologies, architectures and applications in metropolitan area networks [19]. In our present testbed EVL's clusters connect to OMNInet, which peers with Starlight to reach Amsterdam.

Just as the L1/L2/L3 caches in a CPU are used to overcome the slow data rates and high latencies between RAM and the CPU; in the Optiputer, LambdaRAM performs the same functions for metropolitan and international networks. In the illustrated prototype only a single LambdaRAM cache was implemented as the task (described below) was more computationally intensive than visualization intensive.

The LambdaDataNodes consist of LAC's Project Data Space clusters. Project Data Space's goal is to provide the software infrastructure to allow researchers to publish data on the Web in the same way they would publish documents [5]. Project Data Space's transport protocol (DSTP) is analogous to HTTP for the Web[1]. Data retrieved from DSTP servers are streamed at the maximum capacity of the network using an aggressive data transmission scheme called SABUL (Simple Available Bandwidth Utilization Library)[16], which is based on enhancing UDP with negative acknowledgments to provide reliable data transmission while overcoming the bottlenecks of TCP.

2.2 Driving TeraScope on the Optiputer

The LambdaVisualizationNode at EVL is connected to a tiled display. TeraScope is designed so that the results of a particular query can be displayed on any one of the screens of the tiled display. Tiled displays have been largely used in the past to display a single high resolution image of a data set. TeraScope instead, uses the tiled displays to mosaic visualizations so that the user can view several visualizations simultaneously. Using a Web browser, a user can submit a query to TeraScope. TeraScope decomposes the query and sends it to multiple nodes on the LambdaComputeCluster. Residing on the LambdaComputeCluster is a program to perform data mining calculations. In this particular instance the clusters perform Pearson's correlation calculation over all the attributes of the data to rank the "correlatedness" of pairs of attributes in a multidimensional data set. The correlation is performed by each of the nodes sending parallel queries to remote DSTP servers, which in return, will stream the subset of the query results back. These subsets are stored on LambdaRAM, and the data mining engine works from the local copy on LambdaRAM. The final results of the correlation are tallied up and sent back to the querying interface which loads the reduced data set and visualizes it on one of the tiles on the tiled display. The Web interface also allows multiple visualizations to be produced simultaneously, in which case each of the LambdaVisualizationNodes that drive the tiled display will receive a copy of the data to visualize using a variety of different visualization tools.

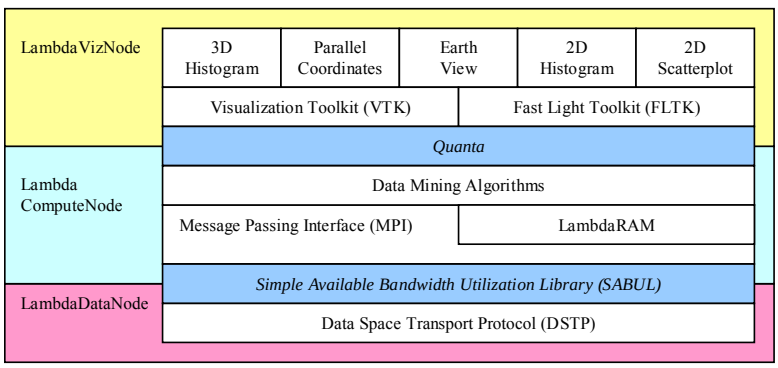


Figure 2. TeraScope software components.

2.3 The TeraScope Software Architecture

Figure 2 shows the overall software modules that are layered to produce TeraScope. At the highest level, TeraScope consists of the visualization tools that reside on the LambdaVisualizationNodes. These tools

were developed with the Visualization Toolkit [11] and the Fast Light Toolkit (FLTK) [21]. Quanta [7], the high performance communication library, is used to communicate between the LambdaVisualizationNode and the LambdaComputeNode to signal the compute nodes to perform the parallel queries of the remote Dataspace data stores. Quanta is also used for fetching the resulting sub-sampled data sets from the LambdaComputeNode to be visualized on the LambdaVisualizationNode. Residing on the LambdaComputeNode is a software framework for executing data mining algorithms such as Pearson's correlation. The framework relies on LambdaRAM and MPI (the Message Passing Interface) [12]. LambdaRAM is also implemented over MPI.

2.3.1 LambdaRAM

LambdaRAM is based on the concept of Network Memory. Prior work in Network Memory (NetRAM) has mainly focused on local area or system area networks because there simply has never been sufficient bandwidth over a wide area network to carry data from memory to memory at rates that are close to memory access rates [4]. The unique difference on the Optiputer is that the high speed optical network that interconnects its components, makes NetRAM over wide areas practical. The concept behind NetRAM is to provide a massive pool of physical memory that is distributed over separate computers. In most current computers, when a program runs out of physical memory it uses its disk drive as virtual memory. The program swaps its data from physical memory to virtual memory as needed. NetRAM changes this paradigm by instead swapping memory to a remote computer rather than to local disk. The advantage of doing this is that it can take significantly less time to swap data to a remote computer using a high speed metropolitan network, than it does to save the data to disk. For example, a SAN such as Myrinet can have as much as a Gigabit of bandwidth with a latency of a few microseconds, our interconnected LambdaNodes will have a bandwidth of 10 Gigabits/s and latencies of approximately 2-5ms; whereas a disk drive only has 300Mbits of bandwidth with a seek time of 10ms.

LambdaRAM's implementation of NetRAM currently provides only Read access. However, even this limited capability has significant utility. Most data mining and visualization algorithms involve the reading of data and the generation of a derived result. Rarely is the original data modified. In order to optimally match the incoming flow of the data with the data access patterns of the data mining and visualization algorithms, an application developer must know when and how much data needs to be prefetched so that the data is available just in time. Since this is a difficult problem to solve, the algorithms are typically modified so that they can perform all the required calculations on a single pass of the data stream. This is clearly not possible for all algorithms- some algorithms simply need to access data values more than once. LambdaRAM is intended to alleviate this problem by allowing the programmer to concentrate on the semantics of the algorithms rather than the optimization of the data fetching.

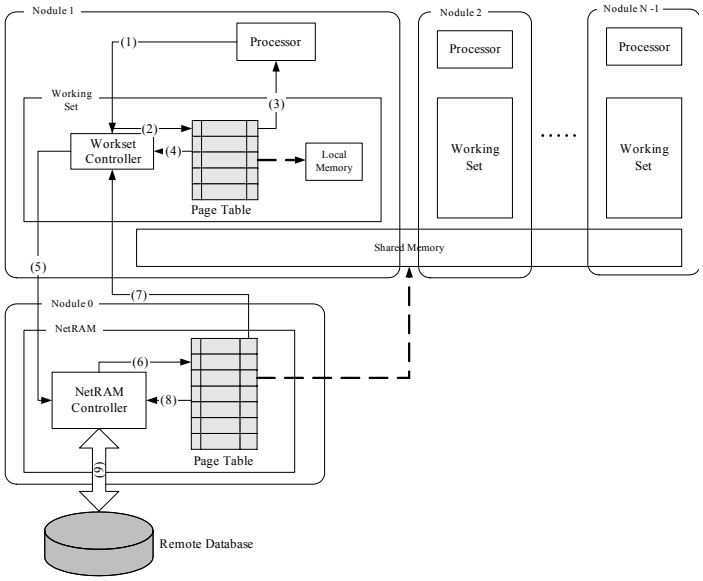


Figure 3. Structure of LambdaRAM

In LambdaRAM, each nodule (a single PC) of the LambdaNode holds a local block of memory, called the NetRAM segment, and a Working Set. Each nodule's Working Set has its own memory controller and page table, whereas there is a single central page table for all the NetRAM segments that are distributed on

the LambdaNode. Hence there are two levels of caches. The Working Set serves as the onboard nodule cache, whereas the NetRAM segment serves as the LambdaNode cache that fetches data from the remote data stores. The structure of LambdaRAM is shown in Figure 3. Its algorithm for memory retrieval is as follows:

(1) When a particular piece of memory is accessed, the request is sent to the Working Set Controller.
(2) The Working Set controller searches for the location of the memory page in the Working Set's page table. If the page is available in the Working Set, Step 3 is performed. Otherwise Step 4 is performed.
(3) The located page of memory is returned as a pointer to the program. Step 10 is then performed.
(4) If the page does not appear in the local Working Set, a page fault occurs in the Working Set Controller.
(5) The page fault causes the Working Set Controller to look for the page on one of the nodules of the LambdaNode. This is achieved by sending a request to the central NetRAM memory controller.
(6) If the NetRAM memory controller is able to identify the nodule on which the sought page resides, Step 7 is performed. Otherwise Step 8 is performed.
(7) The memory page is retrieved from the nodule and stored in the local Working Set. The Working Set Controller may need to decide which page might have to be swapped out. This is decided employing the Second Chance algorithm [20]. Having updated the Working Set's page table, Step 3 is performed.
(8) If the page does not reside on any of the LambdaNode's nodules, a NetRAM fault occurs.
(9) This prompts NetRAM to perform a query to the remote Data Space server to retrieve the needed data. The NetRAM memory controller decides which page is swapped out and updates the page address in NetRAM's central page table. The requested page is moved to the Working Set of the nodule originally requesting the data. Step 3 is then performed.
(10) End of the memory access.

Note that the page size for the Working Set need not be the same as the page size for NetRAM. In fact, to maximize bandwidth utilization, NetRAM's page size is currently set to be 3 times larger than the Working Set's page size. Furthermore, at the present time, NetRAM strictly acts as a virtual memory system rather than a prefetching cache- ie no predictive fetching is performed. Algorithms for prefetching are currently being examined.

2.3.2 TeraScope Visualization Tools

Terascope consists of a variety of information visualization tools which are described below. Recall that the goal is not to draw a terabyte of data, but to produce meaningful visual summaries of the data (called TeraMaps) from which relevant subsets (TeraSnaps) can be visualized on modest systems. The derivation of these visual summaries is what requires the processing power of an Optiputer.

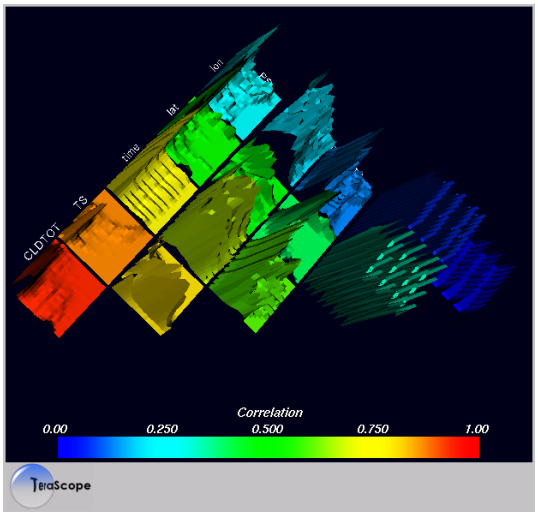


Figure 4. 3D Histogram of atmospheric data from the National Center for Atmospheric Research. Red means there is a high correlation between two attributes along the X and Y axes. The height of the terrain means there are a large number of samples at that particular location in the data set.

3D Histogram – This tool traverses all the requested data points and computes a correlation value between every pair of attributes (Pearson Correlation) in a multidimensional data set. The correlation is then used to color code an overview map of the data set which is a collection of 3D histograms that show the relationship of one attribute to another.

2D Scatterplot and Parallel Coordinates – These tools make use of the Pearson correlation calculation described above to color code the 2D scatterplot. Furthermore the correlation function is used to prioritize the set of attributes that should be placed next to

each other in a parallel coordinate plot. Placing highly correlated attributes next to each other in a parallel coordinate plot helps "untangle" the many lines that cross parallel coordinate diagrams.

3D Scatterplot – Given four attributes (x,y,z, and a scalar attribute), this tool produces both a scatterplot and splat plot highlighting the areas of greatest data concentration.

EarthView – This is analogous to a 2D scatterplot except that the scalar values are plotted on a sphere.

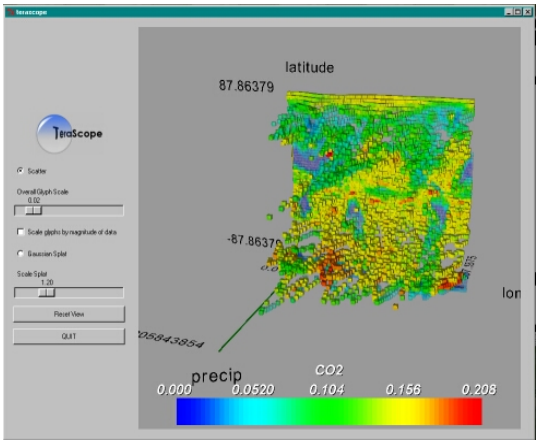


Figure 5. 3D Scatterplot of precipitation and CO_2 levels over the Earth.

3 Results

TeraScope was built in stages. The first prototype included only the visualization tools, and these tools interfaced directly with the data retrieved from the DSTP servers. All correlation calculations in the 3D histogram and 2D scatterplot were performed in the visualization tool. The test case consisted of 100GB of generated data from National Center for Atmospheric Research's (NCAR) Community Climate Model (CCM3). This first prototype was demonstrated at SC2001 in Denver, Colorado.

The second prototype of TeraScope included the LambdaRAM augmentation and was demonstrated at the IGrid 2002 conference in Amsterdam, The Netherlands. At IGrid, TeraScope queried processors at LAC, Amsterdam, Chicago, and Halifax and produced visualizations on the show floor on a tiled display. Unfortunately reliable experimental data could not be gathered during the conference. Work is underway now to benchmark TeraScope to determine how much time it takes for each phase of the process (from data retrieval to visualization,) and to evaluate the effectiveness of LambdaRAM.

4 Discussion and Future Work

This paper has provided an overview of a project recently underway to develop interactive visual data mining tools for exploring massive data sets, using the Optiputer paradigm. At the time of the writing of this paper, a prototype of TeraScope had been demonstrated at IGrid 2002. With the TeraScope framework in place, future work will focus on the development of new tools for creating visual summaries; performance monitoring of the first LambdaRAM prototype; and augmentation of LambdaRAM with adaptive prefetching capabilities.

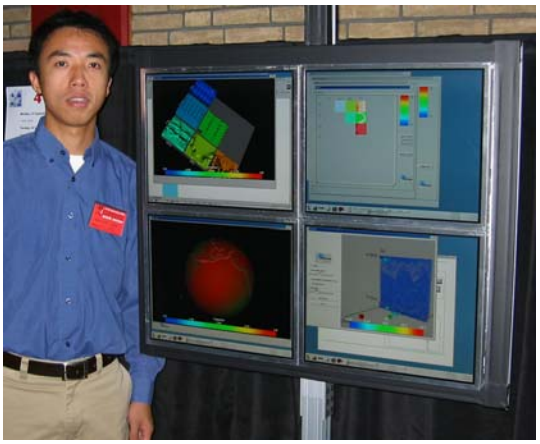


Figure 6. TeraScope tiled display at IGrid 2002. Top left tile shows the 3D histogram; top right shows the 2D scatterplotter and parallel coordinates plotter; bottom left shows EarthView; bottom right shows the 3D scatter and splat plot.

5 Acknowledgments

The virtual reality and advanced networking research, collaborations, and outreach programs at the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago are made possible by major funding from the National Science Foundation (NSF), awards EIA-9802090, EIA-0115809, ANI-9980480, ANI-0229642, ANI-9730202, ANI-0123399, ANI-0129527 and EAR-0218918, as well as the NSF Information Technology Research (ITR) cooperative agreement (ANI-0225642) to the University of California San Diego (UCSD) for "The OptIPuter" and the NSF Partnerships

for Advanced Computational Infrastructure (PACI) cooperative agreement (ACI-9619019) to the National Computational Science Alliance. EVL also receives funding from the US Department of Energy (DOE) ASCI VIEWS program. In addition, EVL receives funding from the State of Illinois, Microsoft Research, General Motors Research, and Pacific Interface on behalf of NTT Optical Network Systems Laboratory in Japan.

StarLight is a service mark of the Board of Trustees of the University of Illinois at Chicago and the Board of Trustees of Northwestern University.

6 References

[1] S. Bailey, R.G., S. Gutti, and H. Sivakumar, A High Performance Implementation of the Data Space Transfer Protocol (DSTP). Proceedings of the KDD 1999 Workshop on High Performance Data Mining, 1999.

[2] M. D. Beynon, C.C., U. Catalyurek, T. Kurc, A. Sussman, H. Andrade, R. Ferreira, and Joel Saltz, Processing large-scale multi-dimensional data in parallel and distributed environments. Parallel Computing, 2002. 28(5): p. 827-859.

[3] Ann Chervenak, I.F., Carl Kesselman, Charles Salisbury and Steven Tuecke, The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Data Sets. Journal of Network and Computer Applications: Special Issue on Network-Based Storage Services, 2000. 23(3): p. 187-200.

[4] K. Li. IVY: A Shared Virtual Memory System for Parallel Computing. In Proceedings of the International Conference on Parallel Processing, 1988

[5] Project DataSpace: www.dataspaceweb.org.

[6] W. J. Frawley, G.P.-S., C. J. Matheus, Knowledge Discovery in Databases: An Overview. Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1991: p. 1-27.

[7] QUANTA: http://www.evl.uic.edu/cavern/quanta

[8] Daniel A. Keim, Information Visualization and Visual Data Mining. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2002. 8(1).

[9] Daniel A. Keim, H.-P.K., Visualization Techniques for Mining Large Databases: A Comparison. IEEE Transactions on Knowledge and Data Engineering, 1996. 8(6).

[10] T. Kurc, U. Catalyurek, C. Chang, A. Sussman, and J. Saltz, Exploration and Visualization of Very Large Datasets with the Active Data Repository. IEEE Computer Graphics and Applications, 2001. 21(4): p. pp. 24-33.

[11] W. Schroeder, B. Lorenson, The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics, Prentice Hall Computer Books, 1996.

[12] The MPI Forum: http://www.mpi_forum.org/docs/docs.html

[13] The OptIPuter: http://www.evl.uic.edu/cavern/optiputer

[14] V. Pemajayantha, Special Canonical Models for Multidimensional Data Analysis for Distributed Computing and Data Mining. http://www.uws.edu.au/qmms/research/reports/, 2002.

[15] N. Sawant, The Tele-Immersive Data Explorer (TIDE): A Distributed Architecture for Tele-immersive Scientific Visualization, Master of Science in Electrical Engineering and Computer Science, Graduate College. 2000, University of Illinois at Chicago.

[16] H. Sivakumar, R.L.G., M. Mazzucco, Y. Pan, Q. Zhang, Simple Available Bandwidth Utilization Library for High speed Wide Area Networks. to be submitted to Journal of Supercomputing, 2002.

[17] P. C. Wong, Guest Editor's Introduction: Visual Data Mining. IEEE Computer Graphics and Applications, 1999. 19(5): p. 20-21.

[18] StarLight: http://www.startap.net/starlight

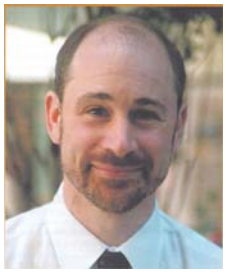
[19] OMNINET: http://www.evl.uic.edu/activity/template_act_project.php3?indi=147

[20] J. Peterson and A. Silberschatz, Operating Systems Concepts, Addison-Wesley, 1986.

[21] Fast Light Toolkit (FLTK): http://www.fltk.org

Chong (Charles) Zhang is an Ph.D. candidate at Department of Computer Science at University of Illinois at Chicago (UIC). Charles is also working as research assistant at Electronic Visualization Laboratory (EVL) at UIC. His current interests include distributed computing, science visualization and data mining.

Jason Leigh is an associate professor of Computer Science at the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago. Leigh is co-chair of the Global Grid Forum's Advanced Collaborative Environments research group; and a co-founder of the GeoWall Consortium. His current research interests include: developing techniques for interactive, remote visual data mining of terascale data sets; application-centric network protocols and data access abstractions; techniques for supporting long-term cooperative work in amplified collaboration environments; and scalable commodity graphics for visualization in immersive environments.



Robert Grossman (grossman@uic.edu) is the Director of the Laboratory for Advanced Computing and National Center for Data Mining at the University of Illinois at Chicago (UIC). He is also the Founder and CEO of the Two Cultures Group.

Marco Mazzucco (marco@dmg.org is a Post-Doctoral fellow at the Univeristy of Wales, Swansea. He is currently working on an ESPRERC funded project in theoretic computer science under the direction of Dr. Martin Otto. He also does research and consulting for the National Center for Data Mining. He received his PhD in Mathematics at UIC in 2000.