TeraVoxel
2003-2004 Annual Report

Diagram of TeraVoxel's network and hardware setup.
ABSTRACT

1.0 Project Overview
The TeraVoxel project aims to advance the hardware and software technology required to store, process, and visualize large data sets generated both experimentally (laboratory, field, etc.) as well as computationally (numerical simulation of multi-dimensional fields). As such, it needs to provide the capability and support for high-speed/-volume data acquisition; interconnections between data-acquisition and data-storage components; the high-bandwidth/-capacity, real-time data-storage components; and high-volume data post-processing and visualization.

The technology is sufficiently generic to permit it to be shared in the acquisition, analysis, and visualization of laboratory data, such as arise in high-speed, multi-dimensional-data applications, for example (1k^2 image data, at 10^3 frames/sec yield 10^9 measurements/sec), as well as for computational data, such as arise in direct numerical simulations of turbulence, for example (several 512^2·1024 fields that evolve as a function of time).

Technologies brought to bear on the aims of this project are extremely volatile. Significant advances in technology have enabled us to replace most of the specific solutions and approaches outlined in the original proposal with improved off-the-shelf products or more easily developed technologies. This was noted as likely in the proposal and has been used to advantage. The envisaged completed TeraVoxel system is significantly more enhanced in capability as well as portability, as discussed below.

The TeraVoxel project completed its third year at the end of August 2003. The discussion below is a report on progress

2.0 Laboratory front end subsystem

The laboratory font end subsystem under development as part of the TeraVoxel project is mostly hardware development with some additional software/firmware development. This subsystem is under development by Daniel Lang with assistance from Garrett Katzenstein, Brian Kern, Steve Kaye, and supervised by Paul Dimotakis.

1.1 KFS Camera to Datawulf Storage Array

Previously, we decided to use the S-Link (Simple Link) to interface the KFS camera system to a Datawulf storage array as shown in the diagram below. The S-Link is a high-speed fiber-optic link developed by CERN (128 megabytes/second) that connects a data source (the VXIADC2 boards) to a data destination (the Datawulf storage nodes). Eight S-Links gives a maximum aggregate transfer rate of 1 GB/s (gigabyte/second), which is sufficient to allow 800 to 1000 fps (frames/second) with (lossless) data compression enabled.
We have purchased 10 of each of the following S-Link boards (giving us 2 spares of each board):

CT-ODIN-3LSC1 S-Link source board (128MB/s, 1 fiber pair)
CT-ODIN-3LDC1 S-Link destination board (128MB/s, 1 fiber pair),
and,

S32PCI64 S-Link to 64-bit PCI adapter board.

The ODIN S-Link source board plugs into the ADC2SLINK adapter board, which plugs into the auxiliary connector on the VXIADC2 board. The ODIN S-Link destination board plugs into the S32PCI64 S-Link to 64-bit PCI adapter board, which plugs into a PCI slot on the Datawulf storage node.

Jan Lindheim has assembled an 8-node Datawulf storage array for the KFS camera system. Each Datawulf node contains a dual Pentium 4 Xeon system with 2 Raid disk controllers and 14 250-gigabyte disk drives giving 2.9 terabytes/node (12 * 250-gigabytes minus a bit of overhead). The total storage capacity of the 8-node Datawulf storage array is 23 terabytes.
Daniel Lang has designed the ADC2SLINK adapter board and assembled & tested one of them. The image below shows the ODIN S-Link source board (at lower right) installed on the ADC board (the ADC2SLINK board is behind the ODIN S-Link source board).

The VXIADC2 board is communicating with the ADC2SLINK board properly and seems to be driving the ODIN S-Link source board properly. A Datawulf test node running Linux is being used for software development. Currently, the Cern Linux drivers are being adapted to store the incoming data to the disk arrays. The Linux software will take another 4 to 6 weeks of development effort to complete.

1.2 KFS Camera Head Redesign

The KFS camera head was redesigned to place the shutter in front of the optical window and the power board at the rear of the camera, leaving only the CCD board inside the evacuated enclosure. This was done to minimize contamination on the CCD from vacuum grease, heat sink grease, and dust particles from the shutter. The new design also simplifies manufacture of the camera head, reducing the number of hermetic connectors from 34 to 4 (the 32 SMA coaxial connectors are replaced by two 50-pin connectors and 2 video adapter boards fan out the video signals to 32 BNC connectors). The new design also eliminates several custom cable assemblies.

Daniel Lang designed the PC boards (CAMH4KFS KFS CCD board, CAMH4PWR power board, CAMH4TIM timing adapter board, and CAMH4VID video adapter board) and Garrett Katzenstein designed the enclosure for the new camera head. The new camera head has been assembled and tested with the KFS CCD chip and is
working well. The image below shows a side view of the completed KFS camera head. The video signals are brought out through 32 BNC quick disconnect connectors, speeding up assembly/disassembly over the 32 threaded SMA connectors on the old camera head.
The next image shows the inside of the camera head with the CAMH4HIT HIT CCD adapter board installed.

3.0 Networking

The Teravoxel Datawulf nodes are linked internally with a copper Gigabit Ethernet network. For connectivity to graphics rendering clusters, archival storage, and other resources such as the Teragrid, the nodes will be linked via fiber optic Gigabit Ethernet into the CACR Force10 E1200 network switch. This switch is capable of providing full non-blocking connectivity between all ports.

4.0 Storage Systems

The Teravoxel Datawulf is built up from the following components:

  o 1 rack mountable 15" LCD monitor with KVM switch
- 1 Dual 2.4 GHz Intel P4 based control node
- 8 Dual 2.4 GHz Intel P4 based storage nodes, each with 3 TB capacity
- 1 Dell PowerConnect 5012 gigabit switch
- 1 5' rack with wheelbase added

The control node consists of:

- SuperMicro X5DPE-G2 Dual Extended ATX motherboard with 533 MHz front side bus.
- 2 * 2.4 GHz 533 MHz FSB Intel P4 Xeon processors
- 2 GB of ECC registered DDR memory
- 1 Gb/s Syskonnect fiber adapter
- Rackmountpro RM2U2S-P300W Rack Mount Case
- Mitsumi CD-ROM
- Maxtor DiamondMax 120 GB disk system disk
Each storage node is built up from the following components:

- SuperMicro X5DPE-G2 Dual Extended ATX motherboard with 533 MHz front side bus.
- 2 * 2.4 GHz 533 MHz FSB Intel P4 Xeon processors
- 4 GB of ECC registered DDR memory
- 2 * 3Ware Escalade 7500-8 PCI 64-bit/33MHz Ultra ATA RAID5 controller
- 14 * 250 GB 5400 RPM Maxtor disks
- 1 Gb/s Syskonnect fiber adapter
- Mitsumi CD-ROM
- Rackmountpro RM3U3D Rack Mount Case which has 450W redundant power supply and holds 14 hot swap IDE drives

The control node is being used for development and job launching of parallel jobs. It holds a file system /home, which is seen by every node in the cluster. It is also the meta data server for the parallel file system, PVFS. Each storage node runs two io daemons, which provides the data feeds in and out of the two RAIDs it has. With a 1.45 TB capacity partition on each RAID for the Parallel Virtual File System (PVFS), the 8 storage nodes will together make up a 23.2 TB file system, which is available to all storage nodes and the control node.
5.0 Visualization

The laboratory data acquisition system and storage has demonstrated to have different needs throughout the different stages of the pipeline. As such, we have revised our original approach and solution, to a single implementation, with three different applications as appropriate to the different needs of the laboratory researcher. Each of these enabled sub-systems reflect entirely knowledge and expertise acquired throughout the research and work done in the previous years.

All the systems are direct subsets of the SP750 volume rendering cluster demonstrated two years ago. It has been updated to utilize the newest generation of volume rendering hardware, the Volume Pro 1000s. These new cards allow us to store up to 1GigaByte of volume data per card, and is capable of rendering 512x512x512 volumes at interactive rates. On contrast to the work done in the past years, as these systems become online, we have recognized two immediate needs: the ability to quickly explore an acquired data set before committing to store it; and the need to be able to process a complete acquired experiment to produce high quality movie visualizations. All machines have also been expanded with a dual port copper Gig-E network cards and 400GB of disk space (divided into 2 hard drives) per machine.
5.1 Acquired Data Online Quick Exploration System

The first stage furnishes online available rendering capabilities for data on the Datawulf coming directly from the camera. One individual box with two Volume Pro 1000 volume rendering cards will allow the researcher to download a single frame from a recently acquired experiment, and visualize it in realtime in interactive speeds (8-60+fps, depending on rendering quality, volume size, and window size). This system also possesses an ATI Radeon with 2 display capability, and fast memory to graphic card texture uploads.

From this initial analysis the researcher can do a better judgment whether the experiment was successful, and whether it is worth archiving further down the pipeline. The software running on this system is an intuitive interface which allows the researcher to explore the data as needed through simple viewpoint and color transfer function manipulation.

This system, know as VOlumeViewer (VoV), is a platform independent solution which utilizes basic OpenGL, GTK+, and the VolumePro VLI api. VoV was designed and implemented by and for TeraVoxel needs solely, which allows us to be able to customize the interface to the researcher’s new needs in the future.
However, we have also explored and used occasionally KitWare’s VolView, though much of its benefit is on using 3D texture memory on commercial graphic cards, which in our case, having the VolumePro 1000s, becomes redundant and unnecessary. Further visualization capabilities are also offered through the use of NAG’s IrisExplorer and a set of modules used to interface with the VP1000, also implemented as part of this project.

5.2 Batch Mode Rendering System

The second visualization sub-system represents a five-node Volume Rendering cluster also utilizing Compaq SP750 workstations that is used for rendering of full movies in batch mode from a complete acquired experiment. Each systems is furnished with a Volume Pro 1000 volume rendering card, and fast connects (copper GigE) to data storage and the network. The basic software for this system will be the same as the one on the direct data exploratory system (system 1). This means, that basic parameters for creating a full movie can be established at the data exploration head system or at the rendering cluster itself. But complementary, the software allows the moving of the acquired semi-processed data from the Datawulf or HPSS storage onto the local cluster, batch rendering according to the saved parameters, and then collection of all the resulting images.

The head node of this subsystem is furnished with a Matrox Parhelia 256MB card, allowing it to show results into up to 3 tiled screens, for a total maximum combined resolution of 3840x1024 pixels (~4MegaPixels).
An earlier version of this system was already deployed and used in the rendering of a Rayleigh-Taylor simulation performed by Trent Mattner, to create a 500 frame movie at IBM’s T-221 resolution (3840x2400 pixels) as will be shown at SuperComputing Conference 2003 in Phoenix.
5.3 High Resolution Display System

A single node, same as the ones described previously has been equipped with an Nvidea Quadro 2 Pro, and a PCI based Matrox G200 MMS. The latter card allows it up to four displays, with a maximum analog resolution per display of 1920x1200 or digital resolution of 1280x1024. For the purpose of high resolution presentations, we have furnished all three systems (as described so far) with 4 Viewsonic LCD displays VP191b, which offer a resolution of up to 1280x1024, but more importantly have a thin 0.71” bezel, which makes the screens optimal for tiling. The total combined resolution will be of 5.2MegaPixels. When the displays are not being used in conjunction as a 2x2 display screen, the screens are allocated one for the Quick Exploration System, one for the Batch Rendering Farm, and 2 for the High-Resolution Display node.
When not used for high-resolution display presentations, this node will be used as either a development node or an additional node in the batch rendering system.

A previously explored version of this third proposed system, which is still under active is volume rendering environment under high density and high resolution displays. This explorations are taking place at Caltech as part of the NSF Teragrid project as well as DOE’s ASCI/ASAP center at Caltech, and being leveraged by TeraVoxel (without having needed to co-fund any of the work so far). At present, even though the results are highly encouraging, this system is still developed as prototype, making it not stable enough to be used under an ongoing fully functioning and working laboratory. However, as work progresses, one of this systems, may be optimal for visualizing phenomena recorded in the Teravoxel system.

The second-generation Terascale Visualization Workstation (TVW-II) is a re-factoring of the original monolithic TVW vision into a display with no moving parts and two second-generation Datawulf storage nodes. It represents a move from monolithic to distributed and specialized, from air to water cooling, from AC power to central DC
power, from 19 inch racks to 23 inch, and it can operate in either a 1 Gb/s Ethernet or 2 Gb/s Myrinet environment. TVW-II achieves an unprecedented coupling of network bandwidth, processing power and pixel density on a silent desktop. Datawulf-II achieves the highest rack storage density and bandwidth from disk to network known. Together, operating in the Teragrid-attached LAN environment, these devices represent the most capable remote visualization endpoint available today.

6.0 Conclusions

One Datawulf storage array with 23 terabytes storage capacity has been assembled & tested, demonstrating the feasibility of a portable high speed storage array. The Datawulf also can be used for computational processing when it is not being used for recording data. The KFS camera is being interfaced to the Datawulf using the high-speed S-Link fiber-optic link. Finally, a high-speed visualization system is being developed for volume rendering and 3-d display of images.