

# TeraVoxel 2001-02 Annual Report

## ABSTRACT

The TeraVoxel project will develop a data-acquisition, computational/data-processing, visualization, and archiving infrastructure to service large data sets arising from both laboratory and numerical-simulation sources. Progress in the TeraVoxel project during the second (2001-02) year permitted the final design and implementation of the various components to be undertaken. At this writing, these are being implemented with an aim of completing the project by the end of August 2003.

## 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  $512^2 \times 102^4$  fields that evolve as a function of time).

Technologies brought to bear on the aims of this project are extremely volatile. Significantly, almost all the specific solutions and approaches outlined in the original proposal have been superceded by improved off-the-shelf, or within-development-reach capabilities. 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 second (or three) years at the end of August 2002. The discussion below is a report on progress.

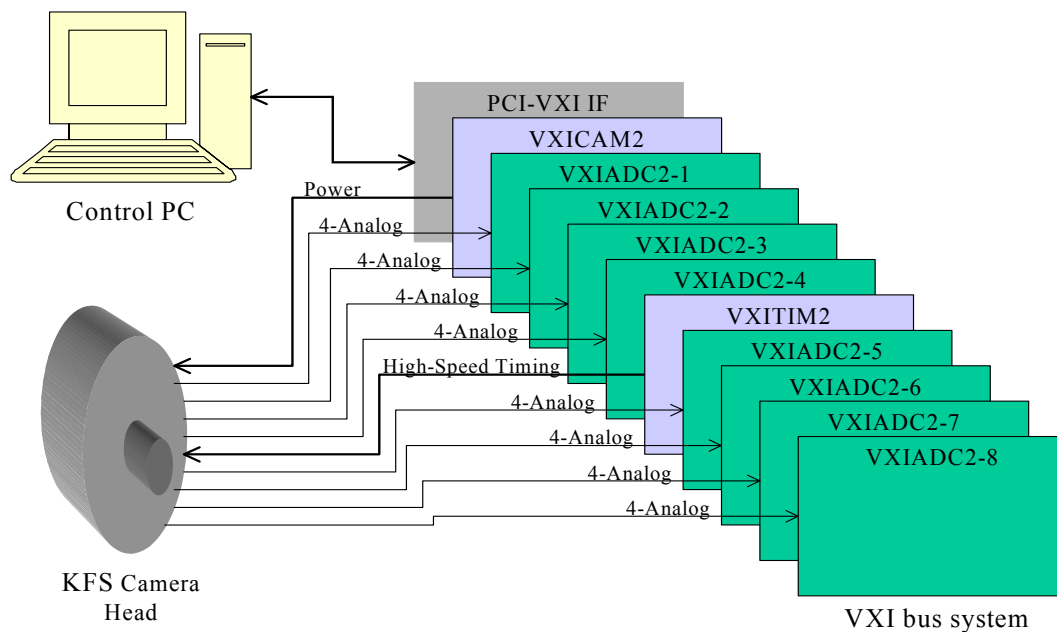
## 2.0 Laboratory front end subsystem

The laboratory front end subsystem under development as part of the TeraVoxel project is mostly hardware development with some additional software/firmware development.

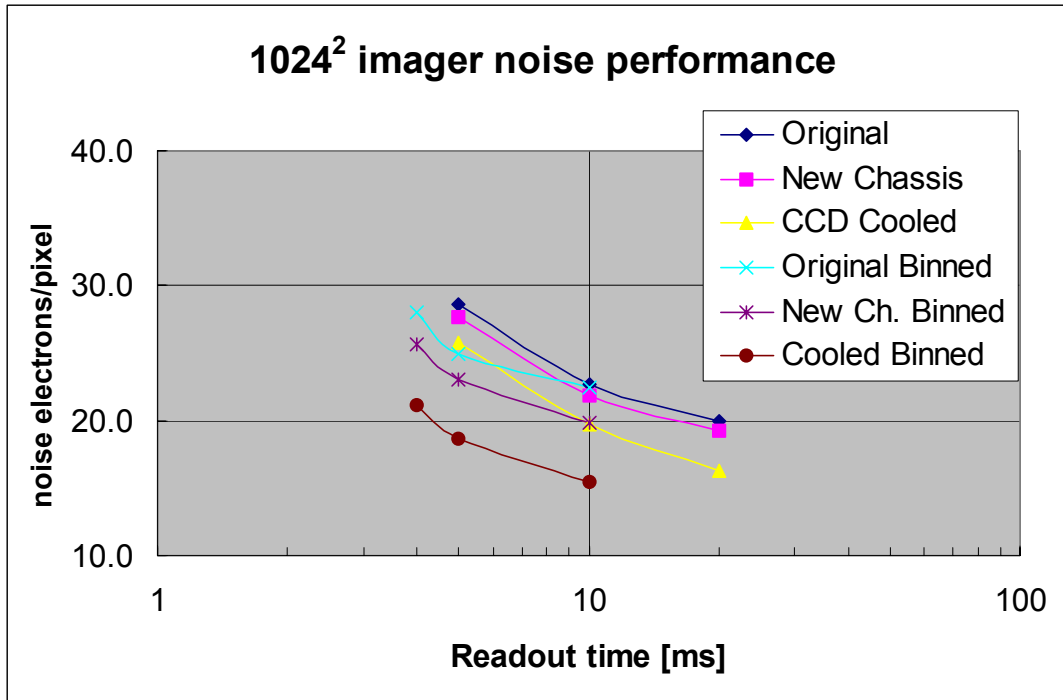
This subsystem is under development by Daniel Lang with assistance from Garrett Katzenstein, Brian Kern, Steve Kaye, and supervised by Paul Dimotakis.

## 2.1 Improved KFS Noise Figures

The KFS camera system is a camera system designed for high-speed low-noise image acquisition. The KFS camera CCD image sensor was designed by Mark Wadsworth of JPL. It has a resolution of 1024 by 1024 pixels and has 32 output channels to acquire up to 1000 images per second. The KFS camera head electronics were designed by Daniel Lang with assistance from Steve Kaye. The camera head housing was designed by Pavel Svitek. The KFS camera system includes a VXI bus enclosure with 8 4-channel A/D converter boards (VXIADC2) for a total of 32 channels, a VXI timing board (VXITIM2), a VXI power board (VXICAM2), and a control PC (See image below).



The KFS image-acquisition system has been completed and is working well. Low-light level applications, however, as encountered in its use at the Palomar observatory last July 2002, can benefit from improved noise figures. The plot below shows the original noise figures (1024×1024) and original binned noise figures (binned to 512×512) in electrons/pixel.



A new VXI enclosure was installed with a separate low-noise power supply for the analog circuits. Also, the analog inputs on the VXI A/D converter boards were adjusted to maximize the Common Mode Rejection Ratio (CMRR). The improved noise figures are shown on the *New Chassis* and *New Ch. Binned* graphs on the plot.

Finally, the CCD camera was connected to a liquid nitrogen dewar and the assembly was evacuated. The liquid nitrogen cools the CCD using a cold finger designed by Brian Kern for operational temperature down to  $-38^{\circ}\text{C}$ . The final improved noise figures are shown on the *CCD Cooled* and *Cooled Binned* graphs on the plot.

## 2.2 KFS Camera Head Redesign

A problem encountered with evacuating the camera and cooling the CCD is that vacuum grease and heat sink grease evaporate from the warmer surfaces and condense on the coldest item, the CCD chip itself. Another problem is that the shutter deposits small dust particles on the surface of the CCD. This has led to a redesign of the camera head to minimize these problems and simplify manufacture of the camera head. The new design places 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. The new design also reduces the number of hermetic connectors from 34 to 4 (the 32 coaxial video connectors are replaced by two 50-pin connectors), reducing the chances of leaks and simplifying assembly. Also, a custom power cable is replaced with a standard 25-pin cable. Currently, Daniel Lang is designing the PC boards and Garrett Katzenstein is designing the enclosure for the new camera head.

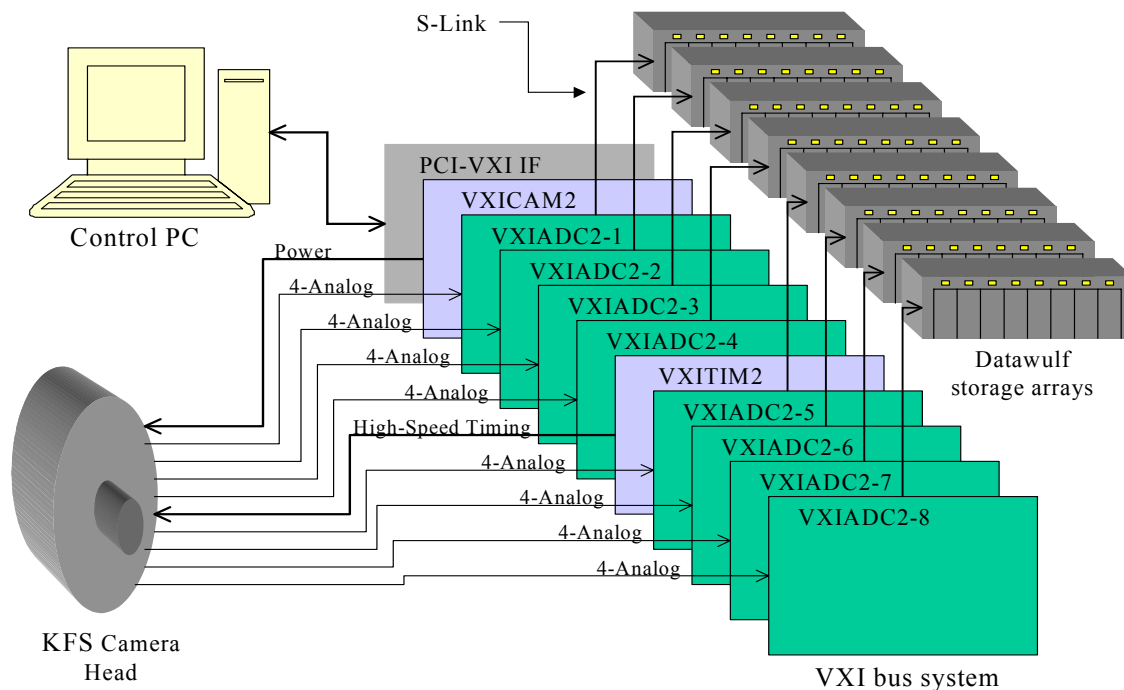
### 2.3 KFS Camera to Datawulf Storage Array

The original proposal called for a Fibre-Channel link from each VXI A/D converter board (VXIADC2) to a Fibre-Channel disk array. Since that time, developments in the PC world have decreased the cost effectiveness of Fibre-Channel. High capacity Ultra-ATA disk drives are now available for a much lower cost:

- Maxtor DiamondMax D540X, 160 GB, Ultra-ATA-133, 5400RPM      \$235
- Seagate Cheetah ST3146807FC 146GB, Fibre-Channel, 10000RPM      \$1105
- Seagate Barracuda ST1181677FCV 181GB, Fibre-Channel, 7200RPM      \$1406

Jan Lindheim has developed the Datawulf storage system for the central data/file storage portion of the TeraVoxel project. The Datawulf is based on a high-end PC with 64-bit PCI slots, 2 Raid disk controllers, 14 low-cost Ultra-ATA disk drives and a gigabit Ethernet interface. The Datawulf storage systems will be used for local data acquisition systems as well, further improving economies of scale.

The basic configuration of the Datawulf storage system has one spare 64-bit PCI slot that can be used for a fiber-optic adapter to receive data from the A/D converter boards. Since the Datawulf storage systems can accommodate any PCI adapters that have Linux drivers, we are no longer tied to the Fibre-Channel protocol. Since the Fibre-Channel protocol is very complex, other fiber-optic links were considered. Most of the other fiber-optic links also have complex protocols that would add substantially to development time. It was decided to rely on the S-Link (Simple Link) interface developed by CERN. The S-Link acts as a simple high-speed data pipe from the A/D converter boards to the Datawulf storage arrays (see image below).



The following S-Link boards have been ordered (10 each):

- CT-ODIN-3LSC1 S-Link Source Card (128MB/s, 1 fiber pair)
- CT-ODIN-3LDC1 S-Link Destination Card (128MB/s, 1 fiber pair),

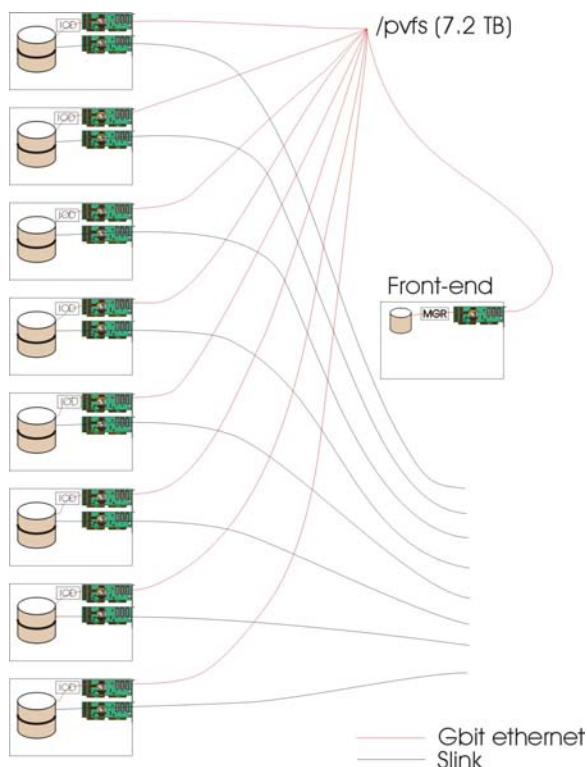
and,

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

S-Link source cards will be mounted on the A/D converter boards and an adapter board will be designed to interface between the A/D converter board auxiliary interface and the S-Link source card. The S-Link destination cards plug into the S32PCI64 S-Link to 64-bit PCI adapter cards, which then plug into the Datawulf storage systems. 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.

The adapter boards between the VXIADC2 A/D converter boards and the S-Link source cards will take approximately 3 weeks to design and 2 weeks to fabricate. This means that we should be ready to start testing in about 10 to 12 weeks (it will take about 5 to 6 weeks to complete the new camera head design). The time-pacing item is the S-Link to 64-bit 66-MHz PCI bus card, which has a 15-week delivery time (orders have been placed with delivery expected by the end of December).

### 3.0 Local data storage and processing subsystem



A storage cluster providing about 15 TB of RAID storage will record the 8 streams of data, coming from the front-end subsystem over S-Link connections, as described above. The cluster will be using gigabit Ethernet between all nodes and will have enough power in itself to do post processing of the data. We also envision this storage cluster being part of a larger cluster with more computational power.

Each storage node is built up from the following components:

- SuperMicro P4DL6 Dual Extended ATX motherboard (\$595)
  - Dual Intel Xeon Processor (up to 2.4 GHz)
  - DDR ECC registered memory support
  - ServerWorks GC-LE chipset

- 1 64-bit 133 MHz PCI-X slot
- 4 64-bit 100 MHz PCI-X slots
- 1 64-bit 66 MHz PCI slot
- 1 Intel 10/100 Ethernet
- 1 Broadcom Gigabit Ethernet
- ATI Rage XL graphics,
- 2 2.2 GHz Intel P4 Xeon (\$245)
- 2 GB of ECC registered DDR memory (\$307)
- 2 × 3Ware Escalade 7500-8 PCI 64-bit/33MHz Ultra ATA RAID5 controller (\$365) supports 8 disks, on-board processor
- 14 × 160 GB 5400 RPM Maxtor (\$235)
- 1 Gb/s Syskonnect fiber adapter (\$564)
- Rackmountpro RM3U3D Rack Mount Case (\$1390) 450W redundant power supply, holds 14 hot swap IDE drives

A front-end node will consist of:

- Supermicro P4DL6 Motherboard (\$595)
- 2 2.2 GHz Intel P4 Xeon (\$245)
- 2 GB of ECC registered DDR memory (\$307)
- 1 Gb/s Syskonnect fiber adapter (\$564)
- Rackmountpro RM2U2S-P300W Rack Mount Case (\$179)
- Maxtor DiamondMax 120 GB disk (\$124)
- Mitsumi CD-ROM (\$13)

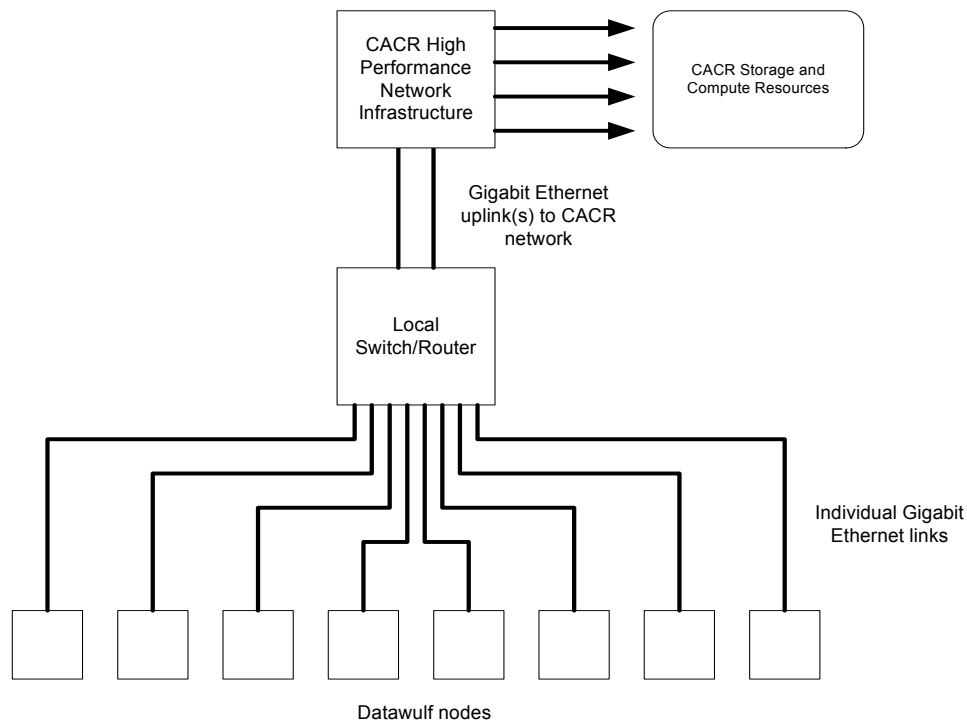
The RAID5s in each Datawulf will be split in two partitions, one for local scratch space and one for Parallel Virtual File System (PVFS).

Because of the long delivery time on the S-Link adapters to be used in the storage nodes, the parts for the nodes will be ordered 5 weeks ahead of the delivery time of the S-Link adapters. This way, we will get better prices on the hardware. Delivery time for the parts is approximately two weeks and assembly will take about two weeks.

Disk prices have been at a peak the last couple of weeks, but are assumed to come down again in a few weeks when also higher-density drives become available.

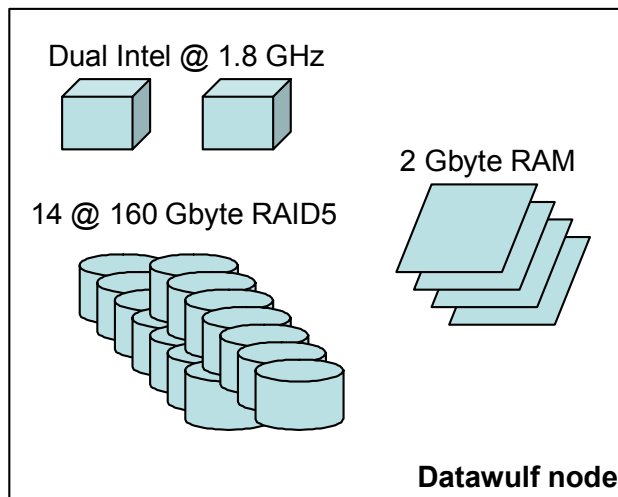
## 4.0 Network subsystem

Interconnectivity will rely on a gigabit ethernet switch that will be installed into the Datawulf rack and each Datawulf node will have a gigabit ethernet link into the switch. The switch will then be linked into the CACR high performance network, providing access to the CACR compute and storage resources. A sufficient number of strands of fiber will be installed between the instrument in the Guggenheim/Karman complex for laboratory data and CACR. This link will consist of one or more gigabit ethernet links, as necessary. The actual count will be determined based on cost and performance requirements.



## 5.0 Central data/file storage and processing subsystem

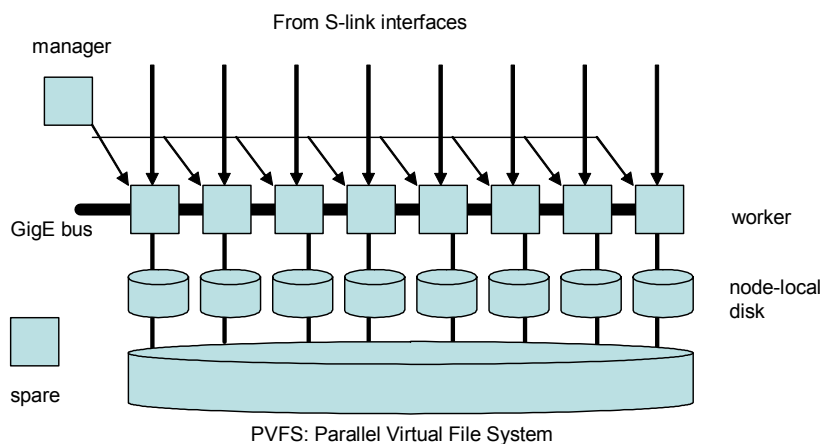
For the data and processing system, we have chosen to use the Beowulf concept for maximum price-performance; however, a new name was used earlier, *Datawulf*. This is because the implication is that data-storage is the primary function rather than computation.



The datawulf system consists of a number of powerful data-computing nodes, as illustrated in this figure.

Schematic Datawulf node (left), with 1.8 TByte reliable storage, with two powerful processors and ample memory. This node can be constructed for less than \$8,000. Such nodes will be connected by Gigabit Ethernet into a parallel storage and computational system using the Beowulf idea. The overall architecture is illustrated below.

Of the 14.4 TBytes capacity available on Datawulf, half is configured as part of node-local devices, i.e., any given node can access only its own part of the storage, and part is configured with the Parallel Virtual File System. This is open-source software from



Clemson University and converts many disk volumes on different machines to a single, large virtual system. A schematic of the configuration of the Datawulf machine as 8 worker nodes, plus a manager and a spare, is on the left. The disk space is partitioned into

local and parallel parts.

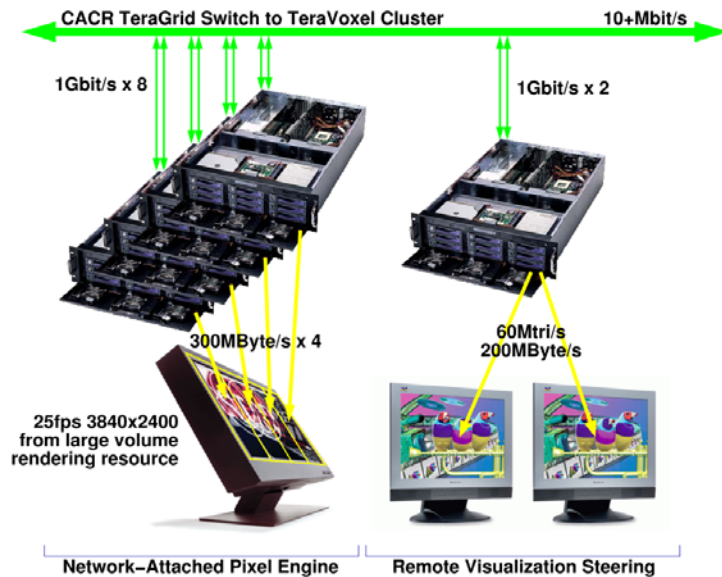
The node-local disk will be used to obtain maximum writing efficiency during data taking, storing compressed data there. After the experiment, data analysis will begin by decompressing the local data into the global PVFS, followed by calibration and visualization steps.

## 6.0 Visualization

The visualization subsystem under development as part of the TeraVoxel project can be mostly classified as software/hardware visualization development. This subsystem is under development by John McCorquodale and Santiago V. Lombeyda.

## 6.1 Hardware Infrastructure

The ability to interactively visualize large-scale volume datasets is central to the TeraVoxel effort. We have constructed a single-user cluster called the *Terascale Visualization Workstation* (TVW), which achieves the scalability and resolution of a multiprojector display wall in a desktide form factor.



Unlike traditional desktop rendering, the TVW is intended to interactively display results of a large-scale parallel rendering computation on a centralized computational resource. This computational resource may or may not have 3D graphics hardware in each node. The TVW is intended to solve bandwidth problems of delivering pixel data to the eye in a scalable and affordable way, enabling research into the parallel rendering application structures necessary to the

goals of the Teravoxel project.

The cluster consists of five nodes, each of which is a 2-processor AMD AthlonMP 2100+ on an AMD 760MPX system chipset. Each node has 4 GByte of RAM and 1.3 TByte of disk. The cluster occupies a small rack, which is the same height as the average desk and has been equipped with low-noise fans for a stealthy lab presence.

Four of the nodes drive four sections of an IBM T221 ("Bertha") display, which provides 3840x2400 resolution at over 200dpi. Two gigabit ethernet pipes and a high-performance AGP graphics coprocessor deliver an observed data bandwidth from network to screen of 950 MByte/s, corresponding to 25 frames per second at 4 bytes per pixel. The disks on these nodes are organized as a hardware-accelerated RAID-0 optimized for recording interactive rendering sessions for later reuse. These four nodes' collective 5.2 TByte disk array can record up to four hours of rendering session.

The remaining node is a traditional high-end graphics PC driving two high-contrast 1600x1200 fast LCD displays and providing the user's keyboard and mouse. Traditional GUI interaction with computation and rendering happens here, with results displayed on bertha. Additionally, this machine has a high-end 3D graphics coprocessor for traditional non-parallel visualization tasks.

Bertha delivers pixel density at the limits of visual acuity. The twin steering displays provide high contrast and outstanding response time (25ms) for crisp motion. Sufficient

real bandwidth is available from network to pixels to reduce large-scale parallel rendering to a problem of pure computation. These resources create a compelling and capable workstation for meeting Teravoxel visualization challenges.

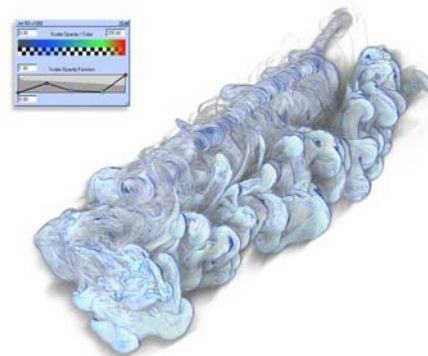
## 6.2 Software/Hardware Visualization



On November 2001, at the SuperComputing conference in Dallas (TX) we demonstrated interactive volume rendering of a  $512 \times 512 \times 512$  12-bit volume at rates of 24+ frames per second. This was done based on the implementation of an 8-node parallel volume rendering cluster equipped with specialized volume rendering hardware (VolumePro 500 by RTViz/TeraCon), as well as specialized dynamic image blending bus across the nodes; as presented at the Parallel Volume Graphics symposium in San

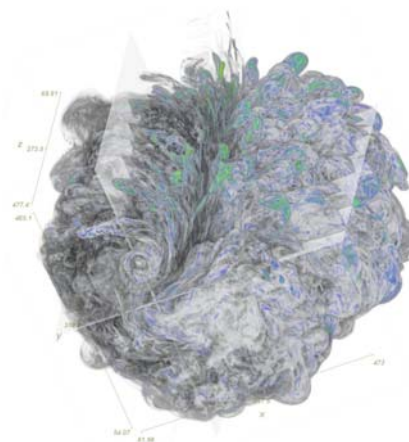
Diego.

Our next iteration of this same cluster replaces the original volume rendering hardware with faster, more flexible, and more capable hardware (VolumePro 1000 by TeraCon). With this new implementation we should be able to render a  $1024 \times 512 \times 512$  volume at interactive framerates, or do multiple-field volume rendering of a  $512 \times 512 \times 512$  volume.



Similarly, in collaboration with Compaq-HP/Tandem, we are investigating the use of commercial off-the-shelf graphic accelerator cards (e.g., Radeon 9700) to implement texture based volume rendering replacing the specialized hardware, resulting in images which can then directly be transferred from the graphics card to the pixel bus via DVI interconnect.

Once the cluster is stable under one or several configurations, the efforts will shift towards creating a remote interface, which will allow users to log in remotely, transfer in data (parallel delivery as sub-volumes to each of the nodes), and the use the available visualization resources from their desktop. Similar delivery of images will also be implement in the context of the transfer and distribution of high-resolution images ( $3840 \times 2400$  pixels) to a parallel display (T-221, as explained above).



A parallel effort at Caltech focuses on the integration of a parallel systems framework, from a visual programming paradigm. As this endeavor takes shape, it will be possible to encompass many of these operations from acquisition all the way to

visualization from a dynamic human-understandable visual paradigm.

## REFERENCES

“Scalable Interactive Volume Rendering Using Off-the-Shelf Components” at PVG 2001  
<http://www.cacr.caltech.edu/projects/ldviz/results/pvr/>

VolumePro by TeraRecon: [http://www.terarecon.com/products/vp\\_1000\\_1\\_prod.html](http://www.terarecon.com/products/vp_1000_1_prod.html)

T221 Display by IBM: <http://researchweb.watson.ibm.com/deepview/>

Parallel Virtual File System: <http://parlweb.parl.clemson.edu/pvfs/>

Sepia by Compaq-HP/Tandem Labs: <http://www.research.compaq.com/SRC/pamette/>