LA-UR- 96-1245

CONF-9606166-

Title: Distributed-Data Imaging System

Author(s):

D.E. Tolmie
A.G. Dornhoff
A.J. DuBois
S.W. Hodson
F.A. Maestas
K.H. Winkler

Submitted to: 10th Annual International Conference
on High Performance Computers,
June 5-7, 1996, Ottawa, Canada

Los Alamos
NATIONAL LABORATORY

Distributed-Data Imaging System

By: D.E. Tolmie, A.G. Dornhoff, A.J. DuBois, S.W. Hodson, F.A. Maestas, and K.H. Winkler

Los Alamos National Laboratory

Presented at:

10th Annual International Conference on High Performance Computers (HPCS 96)

5 June 1996, Ottawa, Canada

ABSTRACT

A group of eight Digital Equipment Corporation Alpha workstations is interconnected with ATM to form a cluster with supercomputer power. For output, each workstation drives a single "tile" on an 8-tile high-resolution frame buffer. A special purpose adapter is used to convert the workstation's ATM format to the frame buffer's HIPPI format. This paper discusses the rationale behind the workstation farm, and then describes the visualization output path in detail. To provide the system quickly, special emphasis was placed on making the design as simple as possible and using standard software protocols to drive and synchronize the display. The design choices are examined, and the resultant system is described. Previously, a display could connect to a single computer; or a group of computers could drive a fragmented display, e.g., a video wall. Our system is unique in that it provides a high-quality desktop visualization display driven collectively by a group of workstations. A short video will be shown during the presentation to demonstrate the system capabilities.

1 - INTRODUCTION

Throughout the country, supercomputer centers are turning to linked workstations for parallel computing. The reasoning behind this shift is economic. Even large, federally funded organizations are dealing with budget cuts. When money is the issue, a workstation cluster is cheaper than a dedicated, multimillion dollar supercomputer, especially when upgrade costs are considered. (The need to upgrade, which means facing the expensive reality of replacing an entire supercomputer, is often the impetus for an organization to make the change.)

For users needing high-quality visualization, however, the shift has had one major drawback. Until now the data output of linked workstations could not be fed, as a unit, to a display screen without blurriness or visible breaks in the finished image. This is not a small problem. Visualization, one of the key benefits offered by supercomputers, allows researchers to understand computed data without reading tedious lists of numbers. Additionally, visualization has provided a faster route to the intuitive leaps that often underlie scientific and technological advancement. It does this by marrying the computer's "number crunching" power to the human brain's superior pattern recognition ability. Supercomputer visualization systems have been a boon to scientists and engineers. Until workstation clusters can provide the same image quality, those users will find the switch to linked workstations a hard one to make.

The Distributed-Data Imaging System, developed through a cooperative research and development agreement between Los Alamos National Laboratory and Digital Equipment Corporation, allows users to make that switch, and to take advantage of the accompanying economic benefits. By perfectly merging the separate data streams of a workstation cluster, our system produces an animated image of supercomputer quality. That is, it offers the user a seamless display, for a picture without system-related artifacts (no interruption in the flow or color of the pattern); high resolution, for superb detail; and high frame rate, for near-real-time process simulation. It is extremely flexible--the simulation can be looped, slowed, or reversed--and even includes a feature not found in supercomputer systems: variable tile size that lets the user concentrate detail where it is most needed. All of this is done at workstation prices and with desktop technology so high-quality visualization can now be cost-effective.

2 - SYSTEM OVERVIEW

As shown in figure 1, the Distributed-Data Imaging System uses a workstation cluster for parallel computation, transmits the results simultaneously, and merges the separate data streams into one high-quality, unbroken moving picture. To do this we spread a simulation problem across eight Digital Equipment Alpha desktop processors. The processors take on portions of the problem and compute in parallel with each other, sharing information through an asynchronous transfer mode (ATM) switch and eight OC-3c links, each with a data bandwidth of 132 million bits (megabits) per second. (The OC-3c signaling rate is 155 Mbit/s, but overhead reduces the user's data rate to about 132 Mbit/s.) When computation is complete, the workstations send their results over the OC-3c links to a Los Alamos-developed adapter, which merges the eight data streams into a single stream with a combined data bandwidth of 1056 megabits per second. Bandwidth that great requires a double-speed High-Performance Parallel Interface (HIPPI) connection for simultaneous transmission to the frame buffer and, through that, to a high-definition television (HDTV) monitor. The name of our special adapter--the ATM-HIPPI Adapter--reflects its role in our system. (HIPPI is a standard for computer networks that transmit data at 800 megabits per second.) At the HDTV monitor the merged data becomes an animated visualization whose seamless quality belies the distributed computing sources and the composite nature of the finished image.

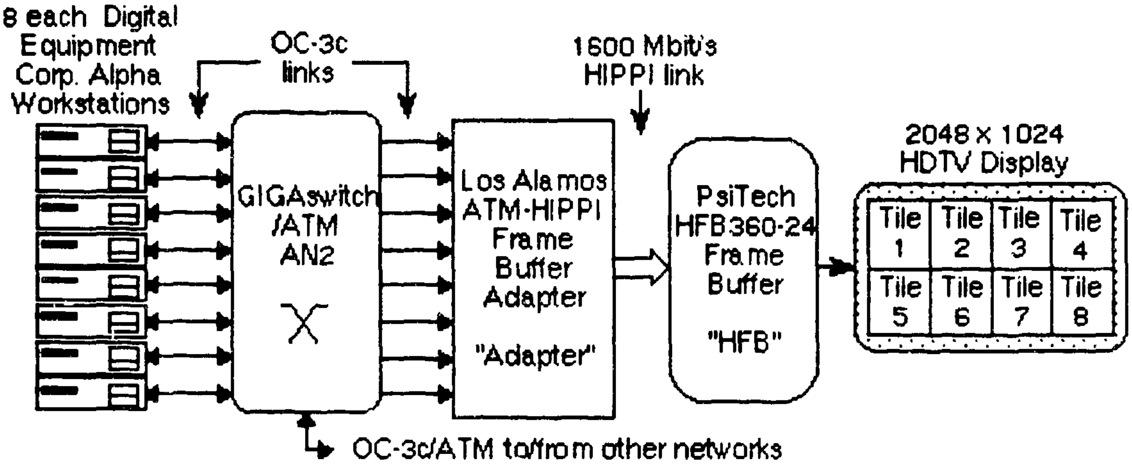

8 each Digital Equipment Corp. Alpha Workstations
OC-3c links
1600 Mbit/s HIPPI link
2048 x 1024 HDTV Display
GIGAswitch /ATM AN2
Los Alamos ATM-HIPPI Frame Buffer Adapter
"Adapter"
PsiTech HFB360-24 Frame Buffer
"HFB"
Tile 1
Tile 2
Tile 3
Tile 4
Tile 5
Tile 6
Tile 7
Tile 8
OC-3c/ATM to/from other networks

Figure 1 - Distributed-Data Imaging System

In reality the HDTV screen is divided into eight pieces, or tiles, each relating to one of the eight workstations. As the ATM-HIPPI Adapter merges the data, it also directs each stream to its designated spot on the viewing screen. Transition from one display tile to another is perfect because the ATM switch allows the workstations to share information about the edge values of adjacent tiles. As a result, the tile boundaries remain invisible unless artificially revealed with a superimposed grid. The visualization is of the highest quality as well because combining the bandwidths of several workstations allows us to exceed the bandwidths of many supercomputers. The moving image has the attributes that allow for the greatest possible insight into the data: high resolution (2048 by 1024 pixels per frame on an HDTV monitor, more on other monitors), excellent color rendition, and rapid frame rate (60 frames per second, twice what is needed for the smooth motion of a near-real-time simulation). To the viewer's eye, frame after frame of the simulation--each frame is a single time-step of the simulation--flows unbroken and highly detailed across the screen, as if generated by a single source. In other words, the Distributed-Data Imaging System coaxes supercomputer visualization from desktop technology.

2 - ADVANTAGES COMPARED TO OTHER SYSTEMS

Our Distributed-Data Imaging System is similar in end product to a high-end supercomputer visualization system--one supercomputer serving a single display monitor--and more distantly to a "video wall"--a tiled display screen with each tile being served by a projection TV. (Video wall are most often used in the entertainment industry but can display computer-generated data from multiple sources, as our system does.)

Our system uses mass-produced commodity parts, reducing costs. Compared to a supercomputer system, the Distributed-Data Imaging System offers lower costs for the computation engine by using commodity workstations, and a modular system that can be easily upgraded in small increments rather than replacing a whole supercomputer. The available bandwidth achieved by aggregating multiple lower-cost commodity workstation interfaces is also greater than the bandwidth provided by many supercomputers.

Compared to a video wall the picture quality is a standout difference, with uniform color and intensity across the whole picture, and no visible tile boundaries to confuse the viewer. Visible tile boundaries result in interrupted or poorly aligned patterns, inhibiting correct interpretation of scientific data. Our desktop form factor is also a plus compared to a room-sized video wall.

With our system, one workstation can drive the entire screen, or tiles can be sized to fit varying amounts of activity. That is, a workstation responsible for a very busy portion of the simulation can drive a smaller tile, providing greater detail exactly where it is needed. Neither a supercomputer system, with its single processor and nontiled display, nor the video wall, with its fixed tile sizes, offers this flexibility.

3 - SYSTEM DETAILS

We used Asynchronous Transfer Mode (ATM) [1,2] to interconnect eight Alpha workstations, and a High Definition television (HDTV) frame buffer display system for visual output. The eight Alpha workstations are model number 3000/600S with a 175 MHz processor. Each has 64 MBytes of memory and 7.8 GBytes of disk. Each workstation is responsible for a single 512 x 512 pixel "tile" of the eight-tile output display, as shown on the right side in figure 1.

The ATM switch is a 20-port Digital Equipment Corporation GIGAswitch/ATM AN2. Each switch port uses SONET OC-3c, 155 Mbit/s, ATM communica tions, over limited distance multimode fiber optic connections. Eight ports go to the Alpha workstations, eight ports to the Adapter, and one port to a Digital Equipment Corporation GIGAswitch for communications with other Los Alamos networks.

The Los Alamos ATM-HIPPI Frame Buffer Adapter, called the "Adapter" for short, converts the ATM-format data received from the Alpha workstations into a High-Performance Parallel Interface (HIPPI) format to drive the frame buffer. HIPPI is an ANSI standard, has been in use since 1988, and is available on many vendor's high-end computing equipment. [3,4] The Adapter is a custom hardware design built specifically for this project.

The PsiTech HFB360-24 HIPPI Frame Buffer, called the "HFB" for short, is a commercially available display system for use with high resolution display screens. A 1600 Mbit/s HIPPI interface is used between the Adapter and the HFB. PsiTech does not presently have an ATM interface for their system, but may offer one in the future.

The 2048 x 1024 HDTV-like display is partitioned into eight tiles, each 512 x 512 pixels. The HDTV standards are not complete at this time, and seem to be going towards a 1920 x 1080 format. A 1920 x 1080 format was awkward with image compression, which operates on multiple 8 x 8 pixel cells. Hence, a 2048 x 1024 format was selected as being close to the HDTV proposal, but easier to implement.

3.1 Minimizing Bandwidth Requirements

A project goal was to display the information with maximum resolution and maximum frame rate, so the user can gain insight into the data being presented --the better the picture the greater the potential insight.

Unfortunately, this is difficult to achieve because of data rate bottlenecks. For example, a 2048 x 1024 display with 24-bit color per pixel requires about 6.3 MBytes of data per frame. At 60 frames per second (fps) this translates into 378 MByte/s, or about 3 Gbit/s--much greater than can be sustained by any component in the system.

The Alpha workstations process the data in parallel, and store the partial results at each workstation. Maximu display performance could only be achieved if the data could also be transmitted and displayed in parallel. It would be a bottleneck if all of the data funneled through a single processor or communications path on its wa to the display. The tack we took was to split the display screen into multiple "tiles", with each workstation responsible for the data for that tile. Separate communications paths were used, coupling each workstation individually to the display system with a commodity OC-3c ATM interface with a user data rate of 132 Mbit/s.

We are limited to a maximum of 20 fps when using 24-bit color. 60 fps is possible with 8-bit color. The choke point is the OC-3c ATM bandwidth. This assumes that all eight workstations are transmitting at the ful OC-3c rate simultaneously.

Joint Photographic Experts Group (JPEG) [5] data compression may be used in the future to effectively increase the data rate by requiring less data to be transferred through the system. Data compression also decreases the storage requirements on the workstations. JPEG compression will be done by software in the workstations, and decompression will be done by hardware in the PsiTech HIPPI Frame Buffer. JPEG operates on 8 x 8 pixel "cells", allowing multiple JPEG chips to be operated in parallel. Early experiments have shown that JPEG compression ratios of up to 7:1 do not cause significant picture degradation of our simulation data images. The compression ratio will be variable to allow experimentation.

Motion Picture Experts Group (MPEG) [6] data compression was also considered since we are in essence transmitting movie data. At the time, the MPEG standards and integrated circuits did not address HDTV size displays. The possibility of using several MPEG chips, one for each tile, was explored, but discarded due to concern about picture artifacts at the tile boundaries. JPEG does not need to pass information between the 8 x 8 pixel cells, but MPEG does. If we increased the bandwidth by using multiple MPEG chips, then we would also need to pass this motion information between the chips, and we did not see a way to do it. Advanced algorithms and chips should allow future HDTV systems to use MPEG-2 with compression ratios better tha that obtainable with JPEG.

3.2 Simplifying the Adapter Design

To provide the system quickly, special emphasis was placed on making the design as simple as possible. Thi led us to put many of the special features in the workstation software, which in turn made the final unit more flexible than if the features were wired into the hardware.

At the start of the project Los Alamos considered building the whole frame buffer, basing it on the design of previous Los Alamos HIPPI-based frame buffer. That design had 1024 x 1024 pixels and 8-bit color. It was decided that in the interest of time and minimal complexity it would be better to purchase a commercial fram buffer rather than design one ourselves. A factor that contributed to this decision was that the resolution of th new system was about 6 times greater, i.e., 1024 x 2048 and 24-bit color. Hence, it requires a higher bandwidth, resulting in a tougher design and layout task. The design of a high-end frame buffer is a complex task even if you discount the bandwidth problems.

The workstation interfaces used ATM communications, and it would have been nice to use ATM throughout Unfortunately, we were not able to find a suitable frame buffer system with an ATM interface. A frame buffer, capable of driving an HDTV display, was obtained from PsiTech Inc. Unfortunately, PsiTech did not offer an ATM interface, only a HIPPI interface. Hence, an "Adapter" to convert the multiple ATM streams from the workstations into a single 1600 Mbit/s HIPPI stream to the PsiTech HFB was designed and built. Los Alamos had considerable experience with HIPPI, and a large set of HIPPI test equipment. Hence, HIPPI in the middle of the system was not viewed as a major detriment.

While the PsiTech HIPPI Frame Buffer (HFB) had some extra features that would not be used in the final system, the extra features did not seem to add significantly to the price or complexity. For example, the HFB provides a path to read the contents of the internal display buffers, and to transmit graphical input from a pointing device like a mouse or trackball. By omitting the possibility of these operations that used a reverse direction data flow, a reverse direction communications channels could also be omitted.

3.2.1 Omitting the Reverse Direction ATM Path

Once the idea of omitting a reverse direction data path was considered, we looked at what other features could not be supported if we did it. Switched virtual circuit (SVC) negotiation, Interim Local Management Interface (ILMI) used for status and control, and Operation Administration and Maintenance (OAM), need a reverse path. While SVCs, ILMI and OAM should probably be part of any commercial production system, it was felt that they were not required in this prototype system. Omitting them resulted in a major software reduction, and allowed a less complex processor to be used in the ATM interfaces, further simplifying the system and shortening the design cycle.

We had originally planned for a feedback mechanism from the Adapter to the workstations to synchronize the workstations and keep the tile data from each workstation in frame-to-frame step. This also would have required a data path from the Adapter to the workstations. The workstations are synchronized by a master workstation sending periodic synchronization messages to the other workstations. By moving the responsibility for synchronization from the Adapter to the workstations, we removed the last barrier to deleting the return ATM path.

Hence, the Adapter's eight ATM interfaces were simplified by implementing the receive side only. Hooks were included for a transmit section if it was needed, but they do not seem necessary now. Figure 1 shows bi-directional ATM paths between the workstations and the ATM switch, allowing the workstations to communicate with each other. A uni-directional path is shown from the ATM switch to the Adapter's ATM interfaces.

3.2.2 Protocols

We selected ATM Adaptation Layer 5, (AAL5) to carry the data from the workstations. Other possible choices were AAL1 or AAL3/4. We chose AAL5 because it seems to be the AAL most commonly used for computer data, and is the easiest to implement. The AAL5 format allows ATM packets up to 64 KBytes in size.

The workstations use User Datagram Protocol (UDP) and Internet Protocol (IP) to transmit the data. Neither protocol is necessary for operation; we use them because they are standards and readily available on the workstations. No return messages were required with UDP, and IP allows the messages to be routed in Local Area Networks (LANs) if necessary. Without UDP/IP, some sort of special driver would have been required to pass the user data to the ATM interface in the workstation.

Our use of the UDP/IP protocols results in an 8-byte UDP header, 20-byte IP header, and 8-byte LLC header, being added to the data packet from the workstation. We use no options, and the headers are always the same size, which simplifies their removal. The ATM-HIPPI Frame Buffer Adapter removes these headers--without examining or using their contents--before sending the data to the HFB.

ATM Virtual Circuits Identifiers (VCIs) are used for routing and addressing; hence, the IP addressing and routing functions can be ignored. Each workstation sends its tile data to the Adapter using VCI = 1024. The XY coordinates in the HIPPI frame buffer command at the beginning of each tile data set determine the particular display tile to which the data are directed. Using a single VCI value allowed all of the Adapter's ATM boards to be hard coded with the same VCI value. Multiple tiles can be sent one at a time over a single OC-3c interface, but multiple tiles cannot be interleaved over a single OC-3c. This limitation made the Adapter design considerably simpler, and was deemed reasonable for this prototype.

We use permanent virtual circuits (PVCs) in the ATM switch. The switch-routing tables are set so that workstation #1's VCI = 1024 is routed to ATM interface #1 on the Adapter. Likewise, workstation #2's VCI = 1024 is routed to ATM interface #2 on the Adapter, and so forth.

3.3.3 Minimizing the AAL5 Processing

AAL5 includes a 32-bit cyclic redundancy check (CRC), and a length field, in the 8-byte trailer of each AAL5 packet. While the workstation hardware will produce the CRC and length, the Adapter's ATM interfaces will discard the CRC without checking it. The length field is only checked to find the last AAL5 packet; the actual number of bytes received is not verified against the length parameter. Ignoring the CRC, and not verifying the AAL5 packet length, allows us to use a complex programmable logic device (CPLD) for the ATM processing. Otherwise available ATM segmentation and reassembly ICs would require additional support hardware to configure and control them. Further minimization includes a fixed format for the AAL5 packets with no pad bytes, as discussed in the section on data flow.

We based our decision to ignore the received CRC on the assumption that the communications circuits would be reliable, negating the need for checking. Random bit errors, resulting in CRC errors, are actually extremely rare due to the short cable distances and benign machine-room environment. Likewise, cells are not lost to congestion in the switch since no other ATM operations are going through at the same time. Also, we use only permanent virtual circuits (PVCs), so the administrator is able to control the interconnections completely. Detected errors cause the associated tile to be discarded, resulting in the in the previous frame's information being redisplayed.

3.3 Data Flow

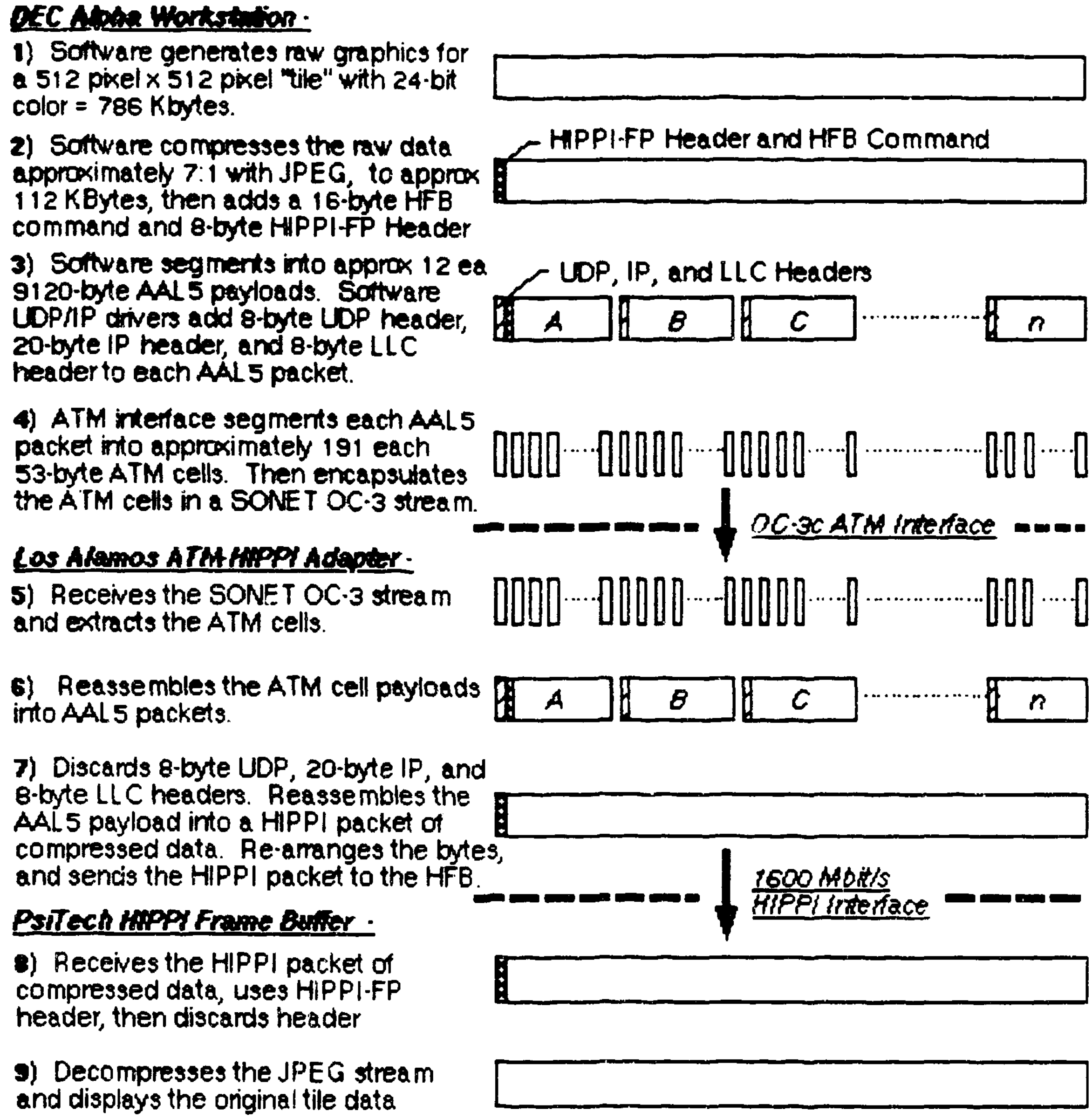
The workstations generate the visualization data one frame at a time, communicating with each other through the ATM switch. Each frame is stored on the disks of the individual workstations. When complete, the data can be recalled from the disks, queued in memory, and the individual frames sent one after the other to the display as a movie.

A fixed AAL5 packet format transmits the data from the workstations. The workstation passes the first 9120 bytes of tile data to the software UDP driver, where an 8-byte UDP header is prepended. The UDP driver in turn passes the data to the IP driver where a 20-byte IP header, and 8-byte LLC header are added. We use a maximum AAL5 packet data size of 9156 bytes to avoid potential IP packet fragmentation. For data sets longer than 9156 bytes the first, and any intermediate AAL5 packets are exactly 9156 bytes long. To differentiate it from the other AAL5 packets, the last AAL5 packet of a data is something other than 9156 bytes long.

When the Adapter receives an AAL5 packet with VCI = 1024, it reassembles the complete data set for the tile before passing it to the HIPPI Frame Buffer. Any AAL5 packets with unknown VCIs are discarded. Reassembly involves discarding the first 28 bytes, that is, the UDP and IP headers, and concatenating the data with the AAL5 packets that follow. As mentioned before, the last AAL5 packet is marked by having other than 9156 bytes.

An AAL5 packet with 9156 bytes, plus the 8-byte AAL5 trailer and 4-byte pad, exactly fills an integral number of 48-byte ATM cell payloads. The 4-byte pad is used to align the data on 8-byte boundaries (the width of the HIPPI interface). Since the pad is always four bytes, except possibly on the last AAL5 packet, it is much easier to concatenate the ATM data to form the HIPPI packet; that is, you do not have to throw away a variable amount of pad bytes. Also, a simple comparison on the length field in the AAL5 trailer tells whether this is the last AAL5 packet in the data set. If it is, the concatenated data set is queued to be sent to the HIPPI Frame Buffer.

Figure 1 gives an overall view of the flow path of data through the system. Figure 2 is a step-by-step summary. Each numbered step from figure 2 is explained here in more detail. Note that compression is not mandatory, it will just reduces the volume of data transferred.



DEC Alpha Workstation -

1) Software generates raw graphics for a 512 pixel x 512 pixel "tile" with 24-bit color = 786 Kbytes.

2) Software compresses the raw data approximately 7:1 with JPEG, to approx 112 KBytes, then adds a 16-byte HFB command and 8-byte HIPPI-FP Header

3) Software segments into approx 12 ea 9120-byte AAL5 payloads. Software UDP/IP drivers add 8-byte UDP header, 20-byte IP header, and 8-byte LLC header to each AAL5 packet.

4) ATM interface segments each AAL5 packet into approximately 191 each 53-byte ATM cells. Then encapsulates the ATM cells in a SONET OC-3 stream.

Los Alamos ATM-HIPPI Adapter -

5) Receives the SONET OC-3 stream and extracts the ATM cells.

6) Reassembles the ATM cell payloads into AAL5 packets.

7) Discards 8-byte UDP, 20-byte IP, and 8-byte LLC headers. Reassembles the AAL5 payload into a HIPPI packet of compressed data. Re-arranges the bytes, and sends the HIPPI packet to the HFB.

PsiTech HIPPI Frame Buffer -

8) Receives the HIPPI packet of compressed data, uses HIPPI-FP header, then discards header

9) Decompresses the JPEG stream and displays the original tile data

Figure 2 - Data segmentation and reassembly

1) The user's software in the workstation does the scientific calculation and generates the raw graphics data for a single tile. In this example, a tile with 512 x 512 pixels and using 24-bit color, will occupy about 786 KBytes. Note that it is possible to use 8-bit color, or other tile sizes, hence with different numbers of bytes.

2) The user's software in the workstation compresses the raw graphics data using the Joint Photographic Experts Group JPEG compression algorithm. Tests have shown that compression ratios of up to 7:1 do not degrade the picture quality significantly. This compression reduces the volume of data that must be stored on the workstation's disk and transmitted through the ATM interface.

After compression, the user's workstation software prepends a PsiTech HIPPI Frame Buffer (HFB) command to the front of the data block. All the HFB commands are 16 bytes in length. For example, a command may give tile coordinates for the graphics data that follows.

The user's workstation software then prepends an 8-byte HIPPI Framing Protocol (HIPPI-FP)[4] header ahead of the HFB command. The size of the data block is contained in the HIPPI-FP header.

3) The user's workstation software then segments the data block into multiple AAL5 payloads. Each payload is exactly 9120 bytes in length, except that the last payload of the data block is shorter.

We chose the value of 9120 bytes so that when an 8-byte UDP header, 20-byte IP header, and 8-byte LLC header, were prepended, the total exactly fills an integral number of ATM cells with a 4-byte AAL5 pad. We also chose the 9120-byte size to ensure that IP packet fragmentation will not occur.

The last AAL5 payload is filled with zeros, if necessary, to exactly fill an integral number of ATM cells.

The user's workstation software then sends each AAL5 payload to the UDP/IP software driver. The resulting AAL5 packets for a particular tile are sent in the proper order, and as a contiguous set.

The UDP/IP drivers will prepend an 8-byte UDP header, 20-byte IP header, and 8-byte LLC header. The contents of these headers are immaterial as they will be discarded by the Adapter. The AAL5 packets, containing the HIPPI-FP, UDP, IP, and LLC headers, will then be forwarded to the ATM interface.

4) The ATM interface transmits the data according to the AAL5 specification, over a SONET OC-3c physical link.

AAL5 specifies that 48 bytes of data be placed in the payload of each 53-byte ATM cell. The last cell of the packet will include an 8-byte AAL5 trailer containing, among other things, a length parameter denoting the number of user bytes in the AAL5 packet, and a CRC-32 checksum. If the number of user bytes, plus the bytes in the AAL5 trailer, is not evenly divisible by 48, then pad bytes are used to fill out the last ATM cell, or possibly the last two cells. Note that by a careful choice of the AAL5 packet size, we are making sure that a consistent 4-byte pad will be used except on the last AAL5 packet of a HIPPI packet.

5) The Adapter, as shown in figure 2, receives the SONET OC-3c stream and extracts the ATM cells from the SONET payload. If a cell does not contain a known VCI, then the cell is discarded.

6) If this is the first cell of an AAL5 packet, then the 8-byte LLC header, 20-byte IP header, and 8-byte UDP header are discarded without their contents being checked.

7) If this is the last cell of an AAL5 packet, then the AAL5 trailer-length parameter is extracted and compared to 9156, that is, (9120 bytes of data) + (8-byte LLC header) + (20-byte IP header) + (8-byte UDP header). If the length parameter = 9156, then this is an intermediate AAL5 packet for the tile and the data is concatenated with other data, if any, for this tile. If the length parameter – 9156, then this is the last AAL5 packet of the tile, and the full tile data will be queued for transmission from the Adapter to the HIPPI Frame Buffer.

8) The HIPPI Frame Buffer uses the HIPPI-FP header values to determine the actual number of bytes in the HIPPI packet. The HIPPI packet header is stripped off in the HIPPI Frame Buffer.

9) The HIPPI Frame Buffer decompresses the data using the JPEG decompression algorithm implemented in hardware. The resultant pixel data are stored in the load buffer.

An Update Image Buffer command is sent from the workstations to transfer the pixel information just sent, and stored in the load buffer to the HIPPI Frame Buffer's display buffer. A single Update Image Data command is sent to update the whole screen, that is, it is independent of the number of tiles used. The Update Image Buffer command may be sent by any of the workstations, but must be timed so that it occurs after all of the tile data has at least started transferring to the Adapter. Timing is the responsibility of the workstations; there is no feedback from the Adapter or HFB to the workstations.

3.4 Synchronization

During the compute phase, the parallel workstation's calculations are synchronized by messages passed over the ATM links. During the visualization phase, one workstation is designated as the master, controlling the timing of the other workstations. For example, the master tells each workstation to send the information for one frame to the Adapter. As each workstation sends their tile for this frame, they report back to the master workstation. When all of the workstations have reported back, the master sends a screen update command to the Adapter to display this completed frame. The master then delays for some period--settable by the user to control the frame rate--before asking the workstations to send the next frame. The messages between the workstations use the UDP/IP protocol.

The user interacts with the master, for example to, speed up or slow down the frame rate, freeze frame, reverse direction, or change the color map.

3.5 Adapter Hardware Implementation

The Adapter is built using a modular design. Each ATM channel is located on a plug-in card with commercial Application Specific Integrated Circuits (ASICs) providing the SONET and ATM layer processing. A specially designed Custom Programmable Logic Device (CPLD) provides the streamlined AAL5 processing. Initialization and monitoring of the channel is exercised by an on-board microprocessor which also provides a serial interface for connecting to a personal computer (PC). The AAL5 packets are processed, that is, UDP/IP headers and AAL5 trailers are removed, as they are received with minimal buffering, and the processed data is placed in a small byte-wide FIFO. Each ATM channel card plugs into a PC-like motherboard containing additional memory buffers, double-wide HIPPI logic, and control logic.

The byte-wide data stream from each of the eight channels is assembled into packets in that channel's separate 4 MByte buffer memory. "Packets" at this level are HIPPI packets that may be either the data set for a display tile or a command for the HIPPI Frame Buffer. When a packet is complete, it is forwarded on to the HIPPI Frame Buffer at double-wide HIPPI speed. With data compression, the packet lengths may vary. A packet sequence of first-started, first-forwarded is followed to keep the packets in order. Other than a few simple length checks to capture gross errors (the packet is assumed to consist of an integral number of 64-bit words), the contents of the packets are ignored by the adapter memory controller. Hence, the workstation software has complete control of the HIPPI Frame Buffer, and allowing flexibility for experimentation at the software level without hardware changes. The 4 MBytes of memory per-ATM-channel allows double buffering of packets at up to twice the expected normal packet maximum.

Two prototype ATM-HIPPI Adapters have been built and checked out. A working system was demonstrated in December, 1995, at Supercomputer '95 in the Digital Equipment Corporation booth. JPEG compression has not been implemented in the prototype. Plans are to use commercial JPEG compression software in the workstations, and to upgrade the PsiTech HIPPI Frame Buffer with JPEG decompression hardware.

4 - SUMMARY

Economics are driving the trend to use workstation clusters to replace supercomputers, but up to now the data output of the clusters could not be fed as a unit to a display screen without blurriness or visible breaks in the finished image. A joint project between the Los Alamos National Laboratory and Digital Equipment Corporation has solved this problem. The Distributed-Data Imaging System uses ATM to interconnect a cluster of workstations. The workstations function as a parallel multiprocessor, generating graphics images for visualizing the results of scientific computations. A special ATM to HIPPI Adapter merges the workstation data streams to drive an high-quality display. The display screen is organized as eight separate tiles, with each workstation responsible for a single tile. JPEG can reduce the amount of data stored and transmitted, allowing the display frame-rate and resolution to be increased to the maximum supportable by the frame buffer. Simplifications involving the protocols, packet formats, and ATM processing, drastically reduced the Adapter's complexity, hardware components, and time-to-implement.

A prototype system was demonstrated in December, 1995, at Supercomputing '95 in San Diego, California. Our system has already elicited the positive response that indicates it is the right technology for the growing trend: parallel computing with workstation clusters.

Acknowledgments

This project was developed through a cooperative research and development agreement between the Los Alamos National Laboratory and Digital Equipment Corporation. Jack DeMember, Dave Loveman, Bob Thomas and Tom Gannon of Digital Equipment Corporation managed the Digital Equipment Corporation end of this project. On the Los Alamos side, Karl-Heinz Winkler was the Principal Investigator--it was his vision and idea that started the whole project, Steve Hodson did the software on the Alphas and tested the compression methods, Frank Maestas programmed the ATM switches, Don Tolmie did much of the Adapter system level design, and Andy DuBois and Gene Dornhoff designed and built the ATM-HIPPI Frame Buffer Adapter. Eileen Peterson of Los Alamos did some excellent editing for submission of this project as an R&D100 entry. Clive Towndrow of PsiTech managed the PsiTech portion of the project.

The Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. This work was performed under auspices of the U.S. Department of Energy.

References

[1] ATM Forum UNI 3.0, ATM User-Network Interface, Version 3.0 Specification.

[2] R. Handel and H. Huber, Integrated Broadband Networks - An Introduction to ATM-Based Networks, Addison-Wesley, Wokingham, England, 1991.

[3] ANSI X3.183-1991, High-Performance Parallel Interface - Mechanical, Electrical, and Signalling Protocol Specification (HIPPI-PH).

[4] ANSI X3.210-1992, High-Performance Parallel Interface - Framing Protocol (HIPPI-FP).

[5] ISO/IEC 10918-1, International Standard, Digital Compression and Coding of Continuous-tone Still Images - Part 1: Requirements and guidelines.

[6] ISO/IEC 11172, Draft International Standard, Coding of Moving Pictures and Associated Audio.

Biographies

Don Tolmie joined the Los Alamos National Laboratory in 1959 as a Technical Staff Member, and has been involved with networking of supercomputers since 1970. His current task is defining the next generation computer network to support higher speeds and visualization, and working with vendors to provide the appropriate products. He has been involved in the computer interface standards activities for over 13 years, and for five years chaired ANSI Task Group X3T9.3 (since renamed to X3T11), responsible for the High-Performance Parallel Interface (HIPPI), Intelligent Peripheral Interface (IPI), and Fibre Channel (FC). Tolmie received a BSEE degree from New Mexico State University in 1959 and an MSEE degree from University of California, Berkeley, in 1961. He can be reached at the Los Alamos National Laboratory, MS-B255, Los Alamos, New Mexico, 87545; e-mail, det@lanl.gov.

Gene Dornhoff has been a Technical Staff Member at Los Alamos National Laboratory since 1975. He was a member of the team that developed the HIPPI ANSI standard, and has been involved in the development of high speed computer networks and related components for more than 20 years. He is currently involved in a project to develop a crosspoint switch for the proposed HIPPI-6400 channel, a proposed standard that will operate at eight times the speed of HIPPI-800. Mr. Dornhoff received a BSEE degree from the University of Nebraska in 1967, and a MSEE degree from the University of New Mexico in 1969. He can be reached at the Los Alamos National Laboratory, MS-B255, Los Alamos, New Mexico, 87545; e-mail, agd@lanl.gov.

Andy DuBois joined the Network Engineering group at Los Alamos National Laboratory in 1991. As a hardware design engineer he has worked on several HIPPI network devices and most recently on the ATM to HIPPI adapter. He received his BSEE and MSEE from the University of New Mexico in 1988 and 1990 respectively. He can be reached at the Los Alamos National Laboratory, MS-B255, Los Alamos, New Mexico, 87545; e-mail, ajd@lanl.gov.

Frank Maestas joined the Los Alamos National Laboratory in 1969 as a computer operator. He has been involved in time sharing systems development since early in his career and his background in the UNIX operating system environment began when he became a Technical Staff Member in 1978. Maestas has been involved since early on in the development of "clustered" environments based on high performance workstations such as the IBM RS/6000 and ALPHA 3000/600. Currently, he is a member of a team using a DEC 8400 with 12, 300Mhz cpus, 4GB of shared memory, and 60GB of disk to parallelize hydrodynamic fortran codes that will generate data to be used in further development of high speed, high bandwidth, networked visualization efforts. Maestas received his BS from the University of New Mexico in 1982, and an MBA from the College of Santa Fe, in 1992.

Dr. Karl-Heinz Winkler joined the Los Alamos National Laboratory as Technical Staff Member in 1984. From 1989 to 1990 he was the Deputy Director for Science, Technology and Education at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, and also a full professor in the Department of Aeronautical and Astronautical Engineering, Mechanical and Industrial Engineering, and Physics at the University of Illinois. At Los Alamos Dr. Winkler is currently the project leader for the Information Architecture Project, Program Manager for Advanced Technology, and is heavily involved in the Accelerated Strategic Computing Initiative. Other positions at Los Alamos included Director of the Numerical Laboratory 1988-1989, and Principal Investigator for the Ultra-Speed Graphics Project 1985-1989. Dr. Winkler received his BS in physics in 1971, and his Ph.D. in Physics Astronomy in 1976, both from Universitaet Goettingen, Germany.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.