A Distributed Digital Library: Planning, Building, and Using

William J. Adams
Department of EE & CS
United States Military Academy
West Point, NY 10996
Voice: (914) 938-5575
FAX: (914) 938-5956
adams@exmail.usma.edu

Bernard J. Jansen
Department of EE & CS
United States Military Academy
West Point, NY 10996
Voice: (914) 938-3233
FAX: (914) 938-5956
jjansen@acm.org

Todd L. Smith
Department of EE & CS
United States Military Academy
West Point, NY 10996
Voice: (914) 938-2200
FAX: (914) 938-5956
smith@exmail.usma.edu

Abstract

Cost cutting and personnel restructuring are forcing organizations to make difficult decisions on where to spend money. Among the areas hit hard by budget cuts are education and training. Academic, industrial, and governmental institutions are all seeking means to leverage technology to improve the timeliness, efficiency, and standardization of their required training. One way to extend budgets while continuing to deliver training is by constructing distributed digital libraries. A distributed digital library consists of material on separate machines connected via a network. The challenge of managing this information is deciding how to store the information and how users will connect, search, and retrieve the material. One method, the monolithic library, forces all user interactions through a single, controlling node of the library network. Another is called the distributed library, which hides the actual server architecture by allowing the user to interact with whichever library node is nearest to him. Using the model of the U.S. Army's Army Training Digital Library as an example, this paper will discuss challenges and solutions to indexing, searching, and retrieving material from globally distributed digital libraries. In particular, this paper will compare the costs and benefits of using a monolithic library structure with that of a distributed digital library. We also present lessons learned from the project so far, specific in the areas of classroom development and video streaming.

Please Cite*:* Adams, W. J., Jansen, B.J., & Smith, T. L. 1998. A Distributed Digital Library: Planning, Building, and Using. *IEEE Conference on Research Issues in Data Engineering.* Orlando, Florida.

See Other Publications

Background

Schools, businesses, and governmental organizations are turning to Distance Learning to bolster enrollments, share expertise, extend the geographical extent of training programs, and broaden their customer base [1],[3]. Within an educational environment, Distance Learning (DL) provides a means to keep faculty employed and low enrollment courses viable through video conferencing and digital libraries [4],[7]. The mechanisms required to deliver an instructor’s video taped lecture or printed material to students at various locations and times are fairly well documented and routine. The challenge lies in formulating a method to distribute interactive, multimedia educational resources on large scale, in a timely and cost-effective manner.

Research is continuing on the most effective utilization of networks for DL [8]. Current DL programs are using a mixture of 3.5" floppy disks and CD-ROMs to distribute multimedia files to students [5], [6]. The 1.44 MB size restrictions of a floppy disk make it impractical for anything but text files or application files to be distributed. Likewise, CDs present two challenges to instructional developers. First is the 650 MB capacity limit of the compact disk. With a typical AVI file averaging 12 MB per minute of video, this storage limitation is quickly reached. Second, because of the "write-once, read-many" nature of compact disks, changes to any file on the disk, no matter how slight, require a new master and all of its

accompanying charges, effectively erasing most of the savings realized through the use of CDs in the first place.

To provide learning materials quickly and efficiently to students, many schools have turned to digital libraries. Some institutions like the University of Minnesota, utilize a type of monolithic storage arrangement to provide learning material to a continuing education student population that is scattered over several hundred miles of the northern United States [2]. Others, like Virginia Tech use digital libraries as a reference store for local students [4]. While the exact methods of storage and retrieval may differ, both examples allow students access to repositories of multimedia information through some type of network connection. These server-based edifices allow any student with network access, either through a local area network or dial-up access, to access or browse learning material of any type. Because of the immediacy of server access by instructors and developers, the material is the most current it can be, without the lag time of disk mastering or disk distribution. As much, digital libraries become a natural extension of any distance learning plan.

Problem

There are some issues with operating and maintaining digital libraries, however. These problems originate with the diverse nature of the user population. The variables in the retrieval equation are the users' connection method, connection speed, knowledge of search queries, and preference of file formats. The enormity of this problem becomes more evident when one looks at the global nature of the users of Army training material.

Digital libraries must be accessible if they are to be successful. Therefore, where the information is stored and how these storage sites are interconnected is of concern. Where do we place the servers and how do we connect them to each other? Of equal importance is that material in the digital library must be arranged and organized so that users do not spend an inordinate amount of time searching for the material they need.

Solution

The best manner to accomplish these tasks is still a subject for research in the networking, information retrieval, and digital library fields. However, some trends are emerging. There are two ways to ensure accessibility. The first method is to consolidate all the material at a single site. This method is suitable for small schools or single colleges, institutions with a narrow range of topics or a small staff of developers. Larger, more diverse organizations need a more distributed solution. The organization and retrieval of networked information is also an on-going research area.

The exponential growth of the World-Wide Web (Web) can provide much information about the characteristics one can expect from a digital library's typical users. Although this is valuable information, the Web may not be the best model for organizing information in a digital library. The closed nature of digital libraries provides organizations with indexing and storage opportunities not available to the Web. Another way to view this is to think of DL as the Web with some forced structure and chosen proponents for a given subject area.

While the investment in technology has been extensive, one of the remaining challenges is deciding on the best design to facilitate access and retrieval of information from the schools' digital libraries. The remainder of this paper focuses on the discussion and the decision-making rationale so far for a major distance learning project for the US Army. The goal is to share our experiences, successes, and lessons learned in order to aid other institutions in the development of their DL programs and digital libraries.

Background

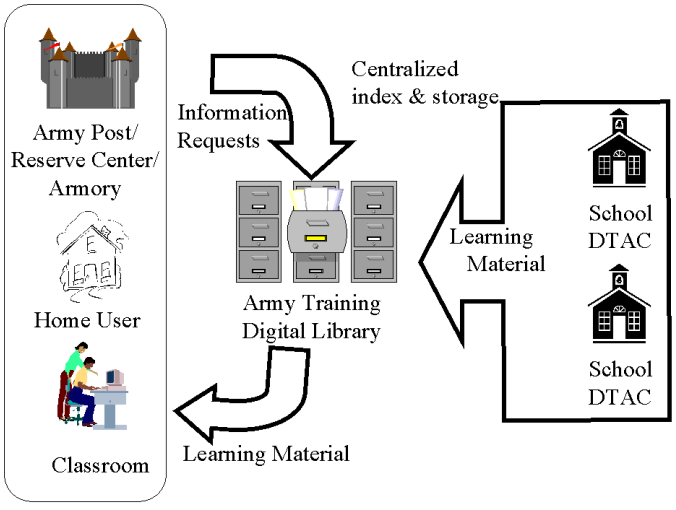
The United States Army has made a significant investment into distance learning capabilities. The goal is for soldiers to be able to access training material from any location. In 1995, the U.S. Army's Training and Doctrine Command (TRADOC) embarked upon a plan that is intended on "leveraging technology to improve training" at the 21 Army schools [9]. TRADOC is similar to a University and the 21 TRADOC schools are similar to a Department within the University. A significant difference though is geographical dispersion.

The 21 TRADOC schools are in various locations across the United States. TRADOC is responsible for developing the training doctrine, the training material, training standards for both military tasks and doctrine. Previously, TRADOC distributed this material via hard copy. The organization then moved to disk distribution; however, this soon proved unworkable on a global, Army-wide scale. TRADOC began exploring other options for distribution.

The long-term intent of the Army Distance Learning Plan (ADPL) is to provide high-quality, standardized training material to soldiers around the world. By providing access to this material, soldiers can receive training or review material in any combination of locations. First, attendees of resident courses will access training material in the classroom through local area networks. Second, upon returning to their workplace the soldiers can review the material from their local education centers, armories, Reserve Centers, or their offices by way of the Internet. In the near future, soldiers will use the Internet from their home, using their personal computer.

There are many ways to quantify the efficiency and cost-effectiveness of this vision. The most immediate means are strictly monetary. The cost of connecting users is undeniably less than the amount currently spent on: mailing and updating course books for correspondence and nonresident courses; travel and per diem costs for resident training; and the lost productivity of students that have to travel. The greater, but harder to quantify, benefit is the guarantee of standardized, on-demand training anywhere in the world. This benefit is especially important for the military, with a large, mobile, and geographically diverse population.

The key component to this architecture is the backbone connectivity that will be used to transfer the training material between school and student, provided in this case via the Internet. The Internet does not provide instant access to the DL; but it does provide a global path. Each DL on the network would contain a cached copy of material from other DLs. If the cached copy were too old or too big to store, then the local DL would request an update from the source DL.



Within the ADLP, the training material that is transported over the network is a combination of multimedia files. Video, audio, and text are the three largest components of learning resources. Once these materials

are retrieved from their source, they are temporarily stored at the network access point's Digital Training Access Center (DTAC.) From the DTAC, users can view and replay the material at their convenience, for their use or for a class.

The planned training environment redefines the concept of the training site. The training site could be a school classroom, a learning lab at a military post anywhere in the world, or a National Guard Armory. Regardless of its location, the site is equipped with a set of hardware that enables a specified set of functions. These functions include: Internet access, World Wide Web browsing, multimedia capability, and local area network connectivity.

The distance learning retrieval process works as follows:

1. A soldier arrives for training. He accesses the local DTAC and requests training material.
2. The local DTAC searches its index of material that is currently stored there. If it finds what the user needs, the DTAC notifies the user that requested material is available and returns a list of Uniform Resource Locators (URLs) of the appropriate files. The process now jumps to step 5.
3. If the requested material is not on the local DTAC, the request is forwarded to the ATDL to find the course material he needs. This material supports correspondence courses, individual professional development, on-the-job training, or review of material learned at a previous resident course.
4. The ATDL searches its database to find course material related to the soldier's request. It returns a list of material in the form of URLs of the appropriate files. The DTAC displays the list of material with estimated download times, calculated from the current network load. The user can select any or all of the material, which is then downloaded to the local DTAC.
5. Once this download is complete, the DTAC notifies the user and prepares for delivery.

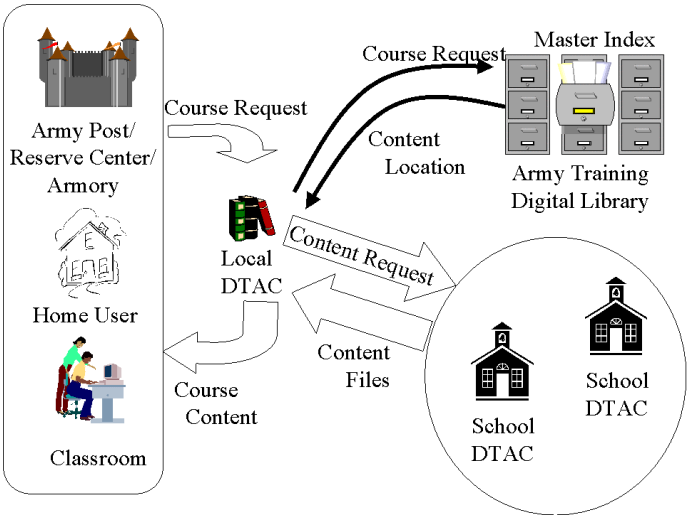
Delivery is dependent on both media and the user. Text, either in the form of HTML or Adobe Acrobat files, is downloaded to the user via a Web browser. Audio and video files represent more challenges however, although streaming products have simplified their delivery.

Discussion of access

We explored two options concerning storage and assess to information: a monolithic and a distributed solution. It is the management of the DTAC and its material that is at the heart of the challenge of a distributed digital library. The challenge is how to manage the storage, access, and retrieval of material from the DTACs.

The Monolithic Solution

Consolidating all an institution's academic material on a single server has several seductive benefits. Consolidated digital libraries can be easier to index, and users only have a single location to access. Figure 1 illustrates this concept. Unfortunately, after analyzing this option, the problems quickly outweighed the benefits. First, the challenge of providing users with a satisfactory response time is very involved. This is comprised of getting enough bandwidth to the site, spreading the traffic load among several servers, and providing security against the eventual hardware failure or network interruption.



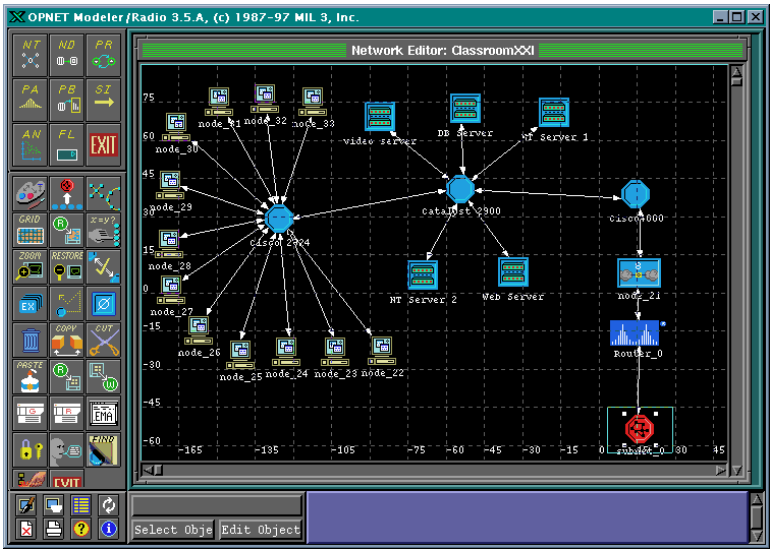
Other, more subtle problems are also present. The first is the fact that a centralized library often has the unintentional effect of absolving the authors of ownership of their material. Each school is the proponent for a specific area of doctrine and training. For example, the Signal School at Ft. Gordon, Georgia is responsible for writing and maintaining all the Army's communications doctrine. This problem exhibits itself in difficulties in getting material updated, defeating the primary purpose of making material available electronically in the first place. Second, access and download times are proportional to the user's distance from the library. This could have the effect of undermining user confidence in the library's services. This is especially important concerning the TRADOC mission of providing training materials and courses to soldiers world-wide.

The schools (which own the DTAC machines) plan to use the DTAC as a place to store developmental materials. This would complicate indexing and retrieval because of the additional layer of marking to prevent users from accessing material that is not ready for distribution. It would also complicate the updating and transferring of information from the schools generating the training material.

The Distributed Solution

The distributed solution has some apparent problems. Namely, the information is not centrally located; therefore, it can be more difficult to manage. This is true in terms of hardware, networking, and information retrieval. Figure 2 illustrates this distributed concept.

The distributed model has several advantages, however. For example, the distributed model provides a location and space for schools to develop and distribute material. Approved material is placed in a specific directory structure where a master index would know to look for it. A distributed model also has the advantage of eliminating a single point of failure. A mirror plan can be implemented where is one server is down, another server with a copy of the information can still provide access to users. With information stored locally, the authors retain ownership of the documents and multimedia stored on their server. The distributed model also can reduce access time for certain users, especially if highly utilized information



is mirrored to different geographical locations. Several researchers have studies the workloads of servers and proxies [13][14]. The distributed model would provide space for development and distribution needs. Approved material would be placed in a specific directory structure where the master index would search, catalog, and retrieve it.

The Distant End

Many discussions of digital libraries and distance learning architecture focus almost exclusively on the network and the servers. This is understandable considering that this aspect of the architecture is many times the most exciting. However, the distance end, i.e., the classroom itself is to the user the most important component of the system. For it is here that the student and professor will interact. It is here that the system must perform. Figure 3 is a computer model of our existing classroom, modeled in OPNET, a military networking modeling and simulation application.

Since the development of the classroom has received so little attention, many may be unaware of the expensive involved in setting up a classroom capable of supporting the wide range of digital library media and mediums. The media may include text, images, video, etc. The medium may be a intranet in addition to a connection to the Web capable of supporting multiple and simultaneous accesses. Plus, there is all the normal support required for a professor in a classroom. Additionally, we have discovered that to take full advantage of the new technology and to emphasis small group exercises, some modifications to the standard lecture hall environment are desirable.

These modifications include, among others, flooring to accommodate the network cabling, projectors for viewing of instructor or student computer screens, and even special sound proofing for the classroom walls. As an example, in our prototype classroom the expenses total over $$270,000. Table 1 presents an itemized expense list for a classroom capable of supporting multimedia instruction. As one can see, a fully equipped, multimedia classroom can be rather expensive.

Item or Component	Cost in US dollars	Comments
Instructor Station		
• Projector	$7,378.00	
• *Video Visualizer*	$4,305.00	

• *Computer*	$3,500.00	
Classroom Modifications		
• *Flooring*	$47,425.00	
• *Lighting*	$5,000.00	
• *Wall Covering*	$30,600.00	
• *Furniture*	$15,300.00	
• *Chairs*	$16,305.50	
• *Tables and Desk*	$5,600.00	
Classroom Computers		
• *Compaq Laptop*	$2,423.00	
• *Laptops*	$57,760.00	20 total
• *Scanner and Printer*	$2,500.00	
Network	$7,000.00	Supplemented by existing university equipment (i.e., cable, router, switch, connectors)
Server Upgrades	$6,200.00	Plus an additional server from the university
Video Server and Software	$59,000.00	Due to the nature of the project, purchased this particular server for evaluation.
Grand Total	$270,296	

Table 1: Items and Cost of Multimedia Capable Classroom.

There are three general groups of equipment purchased for Classroom XXI, User Classroom Computers, Networking, and Servers. For the classroom:, we purchased twenty (20) laptops (P166, 80 MB RAM, 2.1 GB Hard Drive, NT 4.0 Workstation). Each laptop has a 100 Mbps 3Com FastEtherLink PCMCIA card.

For the Networking component, there are 36 CAT5 UTP connections installed in the classroom. Eventually these will be complemented with 36 multi-mode fiber drops as well. All network connectivity is through boxes recessed into the floor of the classroom. The UTP drops are connected to two 24 port Cisco Catalyst switches. This enables us to have a star network configuration connecting users to the servers, which are in an adjoining room.

In order to operate as a stand-alone classroom outside of the university network, we've procured and installed a range of servers for use by the students in the classroom. The TRADOC's current DTAC architecture calls for a web server (Nina), a file/database server (Pinta), and a hypercube video server

(Santa Maria). These servers store training material in an Oracle database. Users make requests through the web server interface.

Nina and Pinta are dual processor Pentium Pro 200's with 128 MB Ram and two hard drives for 4 GB hard drive storage. They use the onboard 10/100 Etherlink cards for network connectivity and operate under NT 4.0 Server. Santa Maria is a dedicated video server with 4 processors arranged in a hypercube. Each processor runs under a proprietary version of Unix called Transit and shares 512 MB RAM. Santa Maria is administered through the web server.

In addition to the three DTAC machines, we have procured and installed a primary domain controller (Isabella) and a backup domain controller/DNS server (Ferdinand). Isabella is a 486/66 with 64 MB Ram and a 2.1 GB hard drive. Ferdinand is a P200 with 32 MB Ram and 2.1 GB hard drive.

Video from Source to Classroom

One of the primary goals of this project is to delivery multi-media files to the user. Multi-media applications tax almost every component of a digital library, the network, data storage, the operating system, servers, and the processors on the user machines. Despite the increased research concerning digital libraries, there are still many unresolved issues, especially in the area of delivery of multi-media files. For example, many some digital library projects desire to use network-delivered multimedia files. Rather than downloading video files, students would use streamed videos that have been indexed and linked to concepts, which the video supports. Streamed video has a lot of advantages in the teaching environment.

What is not usually mentioned, however, is that the cost of providing enough bandwidth to each student workstation to enable television-quality video can be prohibitive. A small (120 x 160 pixel) window can require a significant investment of networking resources per workstation. Additionally, the benefits of the networking can be realized only if the workstation hardware can support video decompression without degradation of quality. We wanted to test if our infrastructure could support the goal of streamed video.

We were interested in whether or not our system (i.e., the network, servers, end user computer systems) could support streamed video. Unfortunately, we could locate no published data on this subject, much less for a given set of hardware. So, we decided to determine this for ourselves.

For our experiment, we wanted to maximize control of the bandwidth while at the same time providing significant traffic. The USMA network's average bandwidth utilization rate is 9% on a FDDI backbone, and the maximum-recorded utilization rate has been 30%. We wanted to approximate this range of traffic in a controlled setting.

We chose a USMA football game that the campus television studio was broadcasting tape-delayed at 9:00 p.m., on a Saturday. The video signal was broadcast over the campus television system, and we captured it using a Dual Processor, Osprey Video Real Encoder (200 MHz Pentium). We configured the software to encode at the maximum quality, which required a constant stream of 500 Kbps from a RealVideo Server. This is a commercial video server capable of delivering 15 frames per second.

The digital stream was then forwarded to a second machine along a shared 10 Mbps channel. This machine was a Dual Processor Server with two Pentium 166s. It contained the Real Video Server, which distributed the signal to the Campus-Area network. Figure 3 illustrates the server configuration.

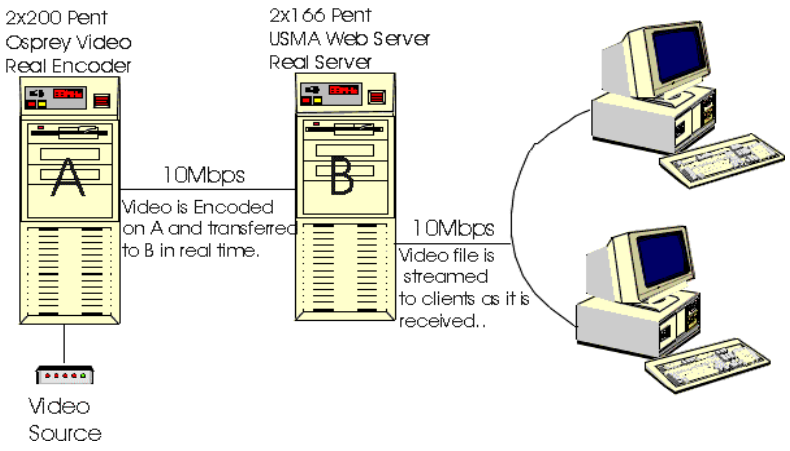


Figure 3: Server Configuration.

Video requires bandwidth that can exceed the capabilities of even the most robust networks. Equation 1 shows that a "postage-stamp" size window requires almost 100 Mbps without dropping any frames. Equation 2 shows the same for a "half-screen" window except that the requirement quadruples. Note that this does not include the audio requirements, which can exceed 1 Mbps.

$$\frac{176\,rows}{frame} * \frac{144\,pixels}{row} * \frac{256\,bits}{pixel} * \frac{15\,frames}{\mathrm{sec}} = 97\,Mbps$$

$$\frac{352\,rows}{frame} * \frac{288\,pixels}{row} * \frac{256\,bits}{pixel} * \frac{15\,frames}{\mathrm{sec}} = 389\,Mbps$$

On a 10 Mbps or even 100 Mbps network, uncompressed video is not possible; however, most vendors provide compression in the range of 1:1 to 200:1. The resulting bit streams can range from 8 Kbps to 600 Kbps depending on the quality of the stream. We set the Real Video Server at the maximum quality setting which provided 15 frames per second at 500 Kbps per stream. Television quality is 30 frames per second. For the experiment, we had 19 workstations in a classroom requiring 19 streams or 9.5 Mbps.

The initial results were disappointing. We were able to receive a maximum of 5 frames per second regardless of the number of workstations receiving the live stream. Surprisingly, the network measured 5% utilization, and the average classroom processor utilization was 60%. We determined that the bottleneck was the Video Encoder coupled with the 10 Mbps segment between the Video Encoder and the Video Server. Figure 4 shows the processor performance of a typical machine when decoding the live stream. The variation is caused by the normal and 2X frame size provided by Real Video. Doubling the frame size increased the processor utilization by an average of 15%.

Next, we streamed a previously encoded file. This test more closely simulated our intended purpose, which was to stream video lessons or assignments from a server. In the worst case, all nineteen machines would be viewing video simultaneously, and we were able to accomplish this task. Although this test better suited our purpose, it did not provide the operational test that we had originally desired.

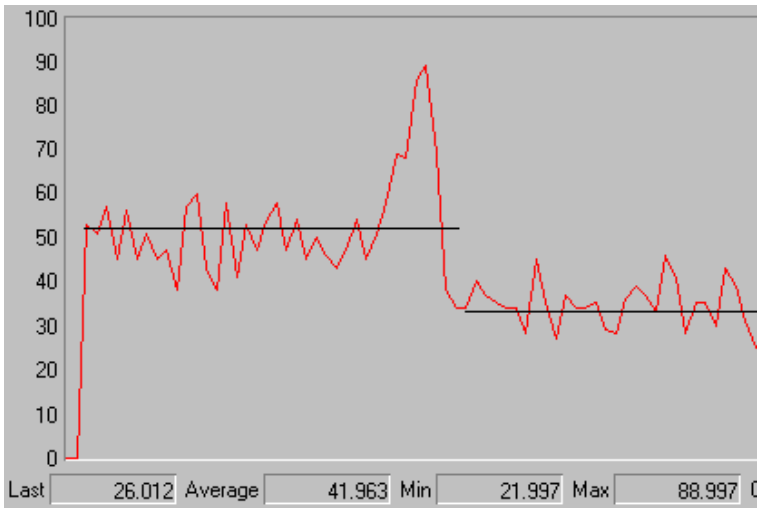


Figure 4: Processor Utilization.

Our bottleneck did shift from the Video Decoder to the workstation processor, however. We were now able to achieve 15 frames per second on each workstation. The total bandwidth requirement was 9.5 Mbps, which was easily supported on our network. However, the Pentium 90 processors dropped several frames, as processor utilization reached 100%.

Although fifteen frames per second and a frame size of 352 X 288 was certainly adequate for comprehension, it was not enjoyable. Images experienced blurring and pixelization, and the video experienced momentary pauses. However, during a separate classroom trial, the majority of students indicated that the quality was adequate. One thing this experiment demonstrates is how far we have to go to address all of the practical, implementation issues of multi-media delivery from source to destination.

Summary and Conclusion

Digital libraries, whether distributed across the globe or across a small campus, provide a quick and cost effective means to distribute learning resources to students, employees, or soldiers. Two methods to manage a digital library were discussed: monolithic and distributed. The monolithic system used a single point to store and deliver all the course material for the organization. While this simplifies the maintenance and indexing of the digital library, user access and information maintenance can present a problem. The monolithic method may be best suited for organizations with small, highly specialized information bases.

The distributed system requires material that is approved for distribution to be placed in a specific directory structure on the dispersed machines for a single master index mechanism to find. Users contact the system through the closest machine, which uses a distributed index to find the material requested by the user. This method is suited for large or widely dispersed organizations, since it optimizes user access and dispersed storage. As such, the distributed method is best suited for the military and other dispersed organizations. The distributed system does, however, increase the maintenance complexity since index updates and material transfers can significantly impact system performance. Choosing the best system requires an analysis of the organization, its location, its user population, and the type of material being provided to the users.

There must be a realization that there is a cost in establishing a multimedia infrastructure. These costs include those of the classroom and the computing and networking infrastructure to support multimedia delivery.

References

1. Dance, Muriel. *The Promise of Distance Learning*. http://Weber.u.washington.edu/~jamesher/mdance.htm. Accessed November 1997.
2. Duin, A. Hill, and E. A. Nater. *Designing and Managing Virtual learning Environments for Secondary, Post-Secondary, Graduate, and Continuing Education: A Land Grant Perspective*. Proceedings of the World Conference on Educational Multimedia and Hypermedia 1996. Pp. 202 - 207.
3. Etter, D.M.; Orsak, G.C.; Johnson, D.H. *A distance learning laboratory design experiment in undergraduate digital signal processing*. 1995 International Conference on Acoustics, Speech, and Signal Processing. Conference Proceedings. p. 2885-7 vol. 5.
4. Fox, Edward. *Digital Libraries, WWW, and Educational Technology: Lessons Learned*. Proceedings of the World Conference on Educational Multimedia and Hypermedia 1996. P. 246 - 251.
5. Harris, J.A.; Murden, C.; Webster, L.L. *The potential of interactive multimedia on CD-ROM to enhance laboratory work in physical science and engineering*. 1994 IEEE First International Conference on Multi-Media Engineering Education Proceedings. p. 296-301.
6. Lollar, R.B. *Distance learning for non-traditional students to study, near home, toward a UNC Charlotte BSET degree*. Proceedings IEEE Southeastcon '95 Visualize the Future. p. 366-7.
7. Palounek, Andrea P. T., et.al. *Distributed Computing Network for Science and Math Education in Rural New Mexico*. Proceedings of the World Conference on Educational Multimedia and Hypermedia 1996. P 557-562
8. Stanford. *An On-Line Distance Learning System Using Digital Video and Multimedia Networking Technologies*. http://minas.stanford.edu/project/project.html. Accessed November 1997.
9. U.S. Army Training and Doctrine Command. *The Army Distance Learning Plan.* Available on-line from http://www-dcst.monroe.army.mil/adlp/adlp.htm. Accessed 1 December 1997.