DOCUMENT RESUME

ED 344 610 IR 054 022

AUTHOR Thornton, Roberta
TITLE The Potential Use of Electronic File Transfer in the National Archives.
SPONS AGENCY National Archives and Records Administration, Washington, DC.
PUB DATE 90
NOTE 27p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Electronic Equipment; Feasibility Studies; Federal Government; *Information Dissemination; *Information Networks; Information Systems; Information Technology; *Information Transfer; Modems; Public Agencies; *Telecommunications
IDENTIFIERS *National Archives DC

ABSTRACT

This paper reviews the incompatibilities among federal government electronic records and explores the potential use of electronic file transfer in the National Archives. It begins by explaining the procedures of the current Center for Electronic Records (NNX) for dealing with accessioning, preservation, and reference tapes. The advantages and disadvantages of existing methods for electronic file transfer are then described, including the use of modems, dedicated lines, the Integrated Services Digital Network (ISDN), local area networks (LANs), and internetworking. The paper concludes with several recommendations: (1) that study of the methods and developments in file transfer technology be continued; (2) that current accessioning and reference procedures be improved with the help of the Data General minicomputer; (3) that the title list and other general finding aids be placed on an electronic bulletin board or on diskette to enable researchers to log on for general information; and (4) that the system for the destruction or return of agency tapes be improved. Three figures illustrating modem operation, star topology, and bridge and gateway linkups are attached. (MAB)

THE POTENTIAL USE OF ELECTRONIC FILE TRANSFER IN THE NATIONAL ARCHIVES

by
Roberta Thornton
Archivist, Technical Services Branch
Center for Electronic Records

June 29, 1990

PART I - INTRODUCTION

The technology of electronic file transfer is one which is constantly under development, and has become common to both the individual and professional user. The result has been a plethora of services and equipment which operate at varying levels of cost and efficiency. Many file transfer services are not compatible with more than a handful of other services. Therefore, while universities, businesses, Federal agencies, and individuals may choose from a variety of file transfer hardware and software, more likely than not, they are limited to communicating with only those who possess identical equipment.

Federal agencies are free to purchase file transfer equipment which best suits their needs. Businesses, universities, and individuals expect to be able to access Federal electronic records with their own file transfer equipment. The resulting confusion confronts The Center for Electronic Records (NNX), which accessions electronic records, and answers to both Federal agencies and private citizens.

Due to the overwhelming variety of hardware and software available on the market, there are many problems with electronic records in the Federal government. Different agencies of the Federal government require specialized computer equipment and facilities, and are entitled to make decisions regarding their agencies' computer requirements. For example, the scientific

data generated from one satellite alone may constitute one terabyte of data per day. The type of equipment necessary to process such massive amounts of data is inappropriate for an agency which uses PCs primarily for word processing and electronic mail, with no requirements for complex computing utilities.

As the National Archives and Records Administration (NARA) and NNX move into an era of automation and improved methods of data retrieval, improved methods of accessioning and reference service of electronic records must be developed. Better and faster data transfer is necessary if NNX is to exist as a useful archives of electronic records.

NNX is responsible for maintaining electronic records of the Federal government which are determined to be of permanent archival value. Certain datasets should be made available to researchers quickly, as some categories of data have high current interest as well as enduring value. NNX staff must possess the ability to accession, describe, preserve, and make available electronic records from all agencies of the Federal government in a manner which is consistent with NARA's mission. Certainly this is no easy task, and rapidly changing technology makes it more difficult to handle a wide variety of hardware and software.

Current procedures for the accessioning, processing and reference service of NNX's holdings are slow and outdated in comparison to the current level of data transfer and retrieval

technology used in the government and private sector. This is damaging to the image of NNX and NARA.

There is equipment on the market which may improve NNX's ability to provide electronic file transfer, some of which is reasonably priced. For example, the surplus Data General mini-computer which NNX acquired from the Small Business Administration will be able to provide faster service for preservation copying than that which is currently possible. Before reviewing such options, an explanation of current NNX procedures is necessary to provide the reader with a basis from which to judge the electronic file transfer options indicated in this paper.

PART II - CURRENT PROCEDURES

A. Accessioning

The Archival Services Branch (NNXA) of NNX is responsible for the appraisal, accessioning, validation and description of datasets as they arrive in NNX. Current procedures for doing this are slow and tedious. The tape arrives, usually with an enclosed SF 258 (Transfer of Federal Records). In addition to the tape is documentation which includes code books to interpret the data, and record layouts.

The accessioning archivist must submit the dataset(s) to the Technical Services Branch (NNXR) for a Tapemap [1] and Tape Dump [2]. When output is received, the archivist validates the records by hand with the tape dump, or by use of Statistical Analysis System (SAS) software. Validation consists of verifying that the record layout provides correct information about each field within a record, that there are no erroneous fields, and that the records match those described in the documentation. Once validation is completed the archivist fills out a P-1 form requesting preservation work. The P-1 is recorded in the Preservation Log and the work is assigned to an NNXR programmer.

---

[1] If the tape is labelled, this utility provides information about each dataset stored on the tape such as logical record length, blocksize, blocking factor, number of blocks, and coding format. In addition, this utility will determine if the tape has data checks.

[2] Prints a block of records.

Once the P-1 is assigned, the programmer pulls the input (agency) tape(s) from the temporary storage vault, pulls two or more blank preservation tapes from the blank tape storage area,[3] logs them out from NNXR, and sends them via courier to the National Institutes of Health (NIH) computer center. It takes approximately one-half day for the tapes to arrive at NIH, and they must be logged into the system before they can be accessed by the NNXR staff. Generally, it takes one day to produce a tapemap and dump, and two days to receive printouts from tapemaps and dumps.

B. Preservation

After NNXA validates the agency tapes, the preservation copy work is performed by first creating a master tape, and then a backup. Skeleton programs, written by NNXR archivists and stored on-line on WYLBUR (the NIH computer's text editor), assist in preservation copying. NNXR programmers modify these programs, which are written in Job Control Language (JCL) and Common Office Business Oriented Language (COBOL), to copy each accession. A Compare [4], Tapemap, and Tape Dump are performed after copying is complete. The tapes are returned to NNXR and are checked by the programmer responsible for making the copies.

---

[3] Blank tapes must pass a strict evaluation test on ComputerLink tape cleaner/evaluator equipment before use in preservation work.

[4] Compares the data stored on the input tape to what has been copied onto the output tape. This is a byte to byte comparison.

It generally takes one week for the work to be completed and for the tapes to be returned to NNX.

Copying multiple datasets onto individual reels is time consuming because each dataset must be copied and compared individually, rather than making use of the tape-copy utility (the tape-copy utility copies all datasets residing on a reel sequentially, but does not compare the output reel to the input reel). Currently, the only use of the tape-copy utility by NNX is for reference work with multiple dataset reels.

Individual copying and comparison is slow and tedious, but must be done in order to verify that the output datasets exactly match the input datasets. Otherwise, the preservation work does not fulfill the archival function. Datasets are often copied in groups of 10 to 20. It is efficient to copy as many datasets as can fit onto the output reel to preserve space. This procedure uses fewer reels and provides less expensive output to researchers (researchers are charged by how many reels have to be mounted in the copy job).

After the copy work is completed, the output tapes are logged into the FilePro TAPES database and the input tape is returned to the originating agency or destroyed. The FilePro TAPES database is a system which contains information about each dataset in the custody of NNX. This information is used for many purposes, including: ten-year re-copying, reference work, the annual sample project, and holdings maintenance. The copies,

printouts, and completed P-1 forms are then checked by another staff member before the work is considered complete.

An additional activity for the staff involves the return of agency original tapes. A backlog of work exists from past years when NNX lacked the staff to work on this project, and there were no formal procedures for the return of agency tapes. Electronic file transfer would reduce this task, because agency tapes would never leave the agency of origin.

After the tapes are copied, the accessioning archivist prepares an NNXA documentation package which describes the dataset(s) on the accessioned tape(s).

Although the procedures described above have worked in the past, there is a large backlog of accessioning and preservation work extending back several years. In addition, the NNX staff faces an enormous increase in the number of accessioned datasets in both the short and long-term, and there is every reason to believe that these numbers will continue to increase, especially as NNX aggressively pursues the appraisal and accessioning of electronic records. When one considers the increase in accessioned datasets in addition to the number of tapes which must be recopied every ten years, the amount of preservation work required becomes overwhelming.

C. Reference

The NNXR staff is responsible for creating output tapes for researchers. Reference tapes are provided on a cost recovery basis. Although they are copied onto a lower quality

storage media,[5] they require time commitment and programming utilities comparable to those described in the preservation process (with the exception that multi-dataset reels are copied via the tape-copy utility rather than by single dataset copy and compare).

All of the procedures described above do not take into consideration what occurs when documentation does not match agency records, or when input tapes have data checks or other problems which prevent the preservation process from operating smoothly. There are many such datasets residing in "archival limbo" because the problems with these tapes are either beyond the current capabilities of the NNXR staff or the time commitment required to solve the problem is too great.[6] The procedures described also do not take into account the possibility that researchers may want something other than complete copies of datasets. For example: many researchers request extracts from large bodies of data.

---

[5] These blank tapes are also evaluated on the ComputerLink equipment, but at a lower standard than that required for preservation tape.

[6] This constitutes another backlog. For example: some earlier accessioned tapes were written in NIPS character code rather than the standard ASCII or EBCDIC format. These tapes were in the process of being converted to EBCDIC via a contract with CorDatum, but a moratorium was issued against CorDatum and the work has been halted.

PART III - ELECTRONIC FILE TRANSFER

Accessioning data electronically via rapid telecommunication lines might eliminate some problems. Agencies would be able to keep their tapes, so NNX would not have to return them. If errors occur during transmission, or data checks appear on output tapes, agencies can re-transmit data. In this way, it is possible that agencies would be more inclined to give NNX copies of datasets, as they would never lose custody, even temporarily, of original tapes. One issue to consider, however, is how to ensure agencies transmit only scheduled records.

The next five sections will consider five options for transferring electronic records via telecommunications.

A. Modems

To understand the potential for electronic file transfer, one must first become familiar with the principles of telecommunications.[7] The most common telecommunications device, the telephone, accesses the public network. The public network (the telephone) is the system which allows each of us to communicate with others from a distance. The simplest device for connecting a computer to the public network is the modem. A modem is a device which converts a digital data bit stream by

---

[7] "...the exchange of information, usually over a significant distance and using electronic equipment for transmission." William J. Beyda, Basic Data Communications (Prentice Hall: Englewood Cliffs, NJ), 1989: p.5.

modulation into an analog transmission, which can be sent via phone lines (see Figure 1). A modem at the receiving end demodulates the analog signal back into digital pulses. Most modems operate at speeds of up to 9,600 baud (bits per second). Many individuals who own PCs also own 9,600 baud modems.

There are two limiting factors which inhibit the use of modems for electronic file transfer. Analog transmission is designed for transmitting the human voice. Analog amplifiers increase the voice signal if it weakens. These amplifiers also increase analog distortion on the circuit. This system is not conducive to transmitting digital data, because there are no digital repeaters tc maintain the integrity of the data. In addition, signal distortion cccurs as a current encounters resistance while traveling through a circuit. As a result, the demodulated signal is not identical to the modulated signal, and data may contain errors or erroneous characters.

Modem speed is the second obstacle to using modems for file transfer. With a speed of 9,600 baud, it would take 10 hours to send one file containing 40 megabytes of data. One reel of data at 6250 cpi may contain up to 180 megabytes of data, which would require 45 hours of transmission time. Many reels in the custody of NNX contain large amounts of data. It is less expensive to send a copy of a large file through the mail or courier than via modem. Shipping reels eliminates the need to re-transmit data when circuit problems occur.

Modems are useful for low volume transfers. Providing the NNX title list or list of new accessions by modem would be an appropriate use of this utility.

Advances in modems in recent years make them more useful and worth watching as they continue to develop. Modems can now have multiport capabilities; making them similar to multiplexers.[8] Multiport modems are able to transmit more than one synchronous data stream over a single transmission line.

Unfortunately, even if some type of standard modem transmission were to be adopted by the Federal government, this would not affect telecommunications with the general public. Therefore, it is not feasible to engage in electronic file transfer via modem, except in the case of small files.

B. Dedicated Lines

Other methods of file transfer involve the use of special lines: leased, switched, and T1. Such lines are available from carriers such as AT&T, Bell Atlantic, etc. They are lines to which the customer must lease or subscribe. They are very expensive, and special wiring is usually required. These lines allow file transfer to occur at speeds much faster than a modem.

Leased and switched lines are more economical than the faster T1 lines, but transfer too slowly for NNX's needs (8 and 22 MB/hour respectively). A T1 transfers at 120 MB/hour, but is

---

[8] A multiplexer is a device which combines multiple data streams into one higher speed data stream; resulting in economical transmission. Multiplexers allow multiple devices to share the same circuit and are usually used in pairs, with one multiplexer at the end of each circuit.

enormously expensive. NNX would have to transfer incredible amounts of data to justify the cost, and the fee for researchers would be exorbitant. Listed below is a chart comparing methods of file transfer and time factors.

		Time to Process (hours)		
Method	Speed	3 MB	40 MB	500 MB
Modem	4MB/hour	.75	10	125
Leased	8MB/hour	.37	5	62.5
Switched	22MB/hour	.13	1.8	22.7
T1	120MB/hour	.02	.33	4.1

C. ISDN

The Integrated Services Digital Network (ISDN) is an end-to-end digital network which will solve some of the transmission problems associated with digital-to-analog-to-digital transmission. As explained earlier in this paper, the public network transmits in analog signal, and modems (or similar equipment) convert digital signal to analog signal prior to transmission, while the end user re-converts the analog signal back to digital. ISDN transmits signal in digital format from end-to-end with no conversion to analog signal.

Aside from data transmission, plans for ISDN include voice and image transmission. Transmission will occur over high-speed circuit and packet switching networks. The Basic Rate Interface (BRI) will utilize clear channel signal consisting of two separate channels: a bearer channel which carries user information, and a separate channel with signal information for

circuit switching. In the United States, the Primary Rate Interface (PRI) will be the equivalent to T-1 carrier service. This, however, will be available only to large organizations.

Because ISDN is in the development stage, and because standards have not yet been defined, this method of file transfer is not available for use, but bears further consideration.

D. Local Area Networks

Local Area Networks (LANs), are physical connections which allow data transfer at high speeds over short distances. There are several types of LANS which are commonly used; these networks transmit data at speeds ranging from 4 to 16 megabits per second. A LAN is independent of the computers attached to it. LANs are highly reliable and have low error rates due to the short distances involved and the control mechanisms built into LAN operating systems. There are three basic requirements for a LAN: there is no intervention by processors; all data enters the network in a standard format, so that all computers equipped with the proper interface can communicate on the network; and the allocation of data transmission capacity is on demand.

There have been many developments in recent years which have improved the reliability of LANs. Different topologies and connections have allowed more rapid data transfer and direct destination data broadcast (see Figure 2).

LANs are privately owned systems; the owner of the network is also the owner of the wiring. The speed of the LAN is limited

by the transmission media which the owner selects and by the protocol of the LAN. Use of optical-fiber rather than copper-conductor cable provides immunity to external noise through use of the short wavelength light signal. Because transmission speed is often high (millions of bits per second), users are able to exchange files. It would appear that the LAN is an attractive choice for electronic file transfer. However, the high speed available is a result of the short distance traveled (a few thousand meters), and is limited to nodes physically residing within the LAN. Therefore, it is not relevant to file transfer between agencies and NNX because it is impossible to link all agencies with NNX via a LAN.

E. Internetworking

Internetworking may be defined as:

> ...a set of connected networks that act as a coordinated whole. The chief advantage of an internet is that it provides universal interconnection while allowing individual groups to use whatever network hardware is best suited to their needs. [9]

Internetworking connects different networks together. Internetworks are also known as wide area networks or data transport networks. There are many wide area networks in use by government, business, and academia. These networks perform only the low level function of the connection, or data transport. Internetworking ignores the upper level functions of the individual networks. By providing no functions other than data

---

[9] Douglas E. Comer, Internetworking with TCP/IP: Principles, Protocols, and Architecture (Prentice Hall, Englewood Cliffs, NJ), 1988: 9.

transportation, internetworking is possible, because data transport networks are not physical networks, but provide links between or among physical networks. The physical networks must be compatible at higher levels, because the internetwork does not intervene or verify compatibility. The result is that separate networks which operate by different protocols [10] may communicate with each other, but the user is responsible for verifying that the individual systems are compatible.

Internetworking has grown as a means of connecting networks. Separate networks, using different communications methods, may be connected by a utility called a gateway. Networks using the same communications methods may be connected by a simpler utility called a bridge. For example: two packet switching networks can be connected by a bridge, while a packet switching network can be connected to a token ring network by a gateway (see Figure 3). This means that NNX could transmit and receive data through data transport networks, but would have to be able to communicate with any number of networks at higher level OSI protocols.

Speed is a major problem with internetworking because high-speed networks cover short distances while slower networks cover great distances.

Internet protocols allow the transfer of large files. Again, however, the transfer of large datasets over Internet

---

[10] Protocols provide formulas for communications which are independent of vendor standards.

would be expensive and time consuming. Providing such transfer on a cost recovery basis would be enormously expensive to the researcher.

Some networks are specifically designed to limit file size in transport. BITNET (Because It's Time) is one such network. On BITNET, it is only possible to transmit approximately 150 pages at a time. This is not feasible for datasets containing thousands of records. If NNX were to attempt to transmit a large dataset over BITNET, the dataset would have to be broken and transmitted in bursts. Hence, the problem with time and transmission error comes into play.

DARPA [11] is an agency which provides specifications for interconnecting networks and routing traffic. It specifies Transmission Control Protocol/Internet Protocol (TCP/IP) [12] for the DARPA network (Darpanet). TCP/IP technology is a fundamental protocol used to connect many institutions. Without basic protocols such as TCP/IP, basic physical transportation would not be possible.

Standardization is one way of eliminating incompatibilities among physical networks linked by internetworking. TCP/IP is a de facto standard. The International Standards Organization has

---

[11] Internet research funded by the Defense Advanced Projects Research Agency.

[12] TCP - Transmission Control Protocol - allows a process on one machine to send a stream of data to a process on another. IP - Internet Protocol - provides the basis for connectionless packet delivery. TCP/IP are fundamental protocols.

formally endorsed another set of standards, called the Open Systems Interconnection (OSI). The OSI protocols cover a broader range of communications functions. Unfortunately, the range is so broad that it is possible that two networks which both conform to OSI cannot communicate with each other. The National Institute of Standards and Technology has established a subset of the OSI protocols, called the Government OSI Protocol (GOSIP) as standard for the Federal government. Within a few years, all agencies will have to conform to GOSIP when procuring ADP equipment. Unfortunately, many non-standard networks already in Federal agencies will remain in use for years. Furthermore, GOSIP will not increase internetworking speeds, or reduce incompatibility with the private sector.

PART IV - CONCLUSION

The possibilities for electronic file transfer are exciting and can offer many benefits to NNX in the areas of both accessioning and reference. However, even with the availability of internet, the plethora of available file transfer options makes it virtually impossible for NNX to choose a method which will connect all Federal agencies, institutions, and researchers. File transfer speed and cost eliminate the use of modems, which are certainly common, from file transfer beyond limited file size.

It is simply not possible to choose one network or one piece of equipment which will meet the needs of all PCs and computer facilities with which NNX may communicate. Even if the possibilities for reference are ignored, NNX does not have the funds to afford a T1 line, and does not accession enough datasets from agencies with rapid speed file transfer capabilities to justify the expense. It is less expensive to ship datasets, obtain replacement copies when there are data checks, and return or destroy agency original tapes, than to engage in electronic file transfer.

Rather than delve into the multitude of equipment and technology which could upgrade NNX's telecommunications capabilities, a better use of time and energy would be spent examining the problems existing in NNX, and revising procedures

to help eliminate them. For example, using the Data General to copy reference output reels will shorten the turn-around time in processing reference requests. This will help researchers who are usually computer literate and expect only a short wait for datasets, because one of the fundamental purposes of electronic records is quick access.

If NNX is to use the Data General mini-computer for copying tapes, a second 6250 cpi tape drive or a cartridge drive must be purchased for tape-to-tape copying, otherwise NNXR will have to copy tape-to-disk and then disk-to-tape for every dataset. Such a procedure would be a waste of hard disk space and would be slow.

The advantage of using the Data General is that many tapes could be copied on-site, rather than at NIH. This will eliminate time wasted while tapes are in transit or waiting to be logged-in to the NIH tape library. It could also be of great benefit when rush orders occur or when there are problems with the courier. Use of the Data General, however, will in no way replace the use of off-site computer facilities, because it lacks some capabilities which are necessary for certain types of preservation work.

It is perfectly appropriate to have aggressive accession policies, but only in tandem with equally aggressive preservation and reference service policies. Otherwise, as is the case, accessioned datasets reside in a backlog because NNX lacks the staff, equipment, and time to complete the preservation process.

My recommendations are: continue to study methods and developments in file transfer technology, improve current accessioning and reference procedures with the help of the Data General mini-computer, consider placing the title list and other general finding aids on an electronic bulletin board or on diskette, so researchers can log on for general information, and improve the system for the destruction or return of agency tapes.

Figure 1

MODEM OPERATION

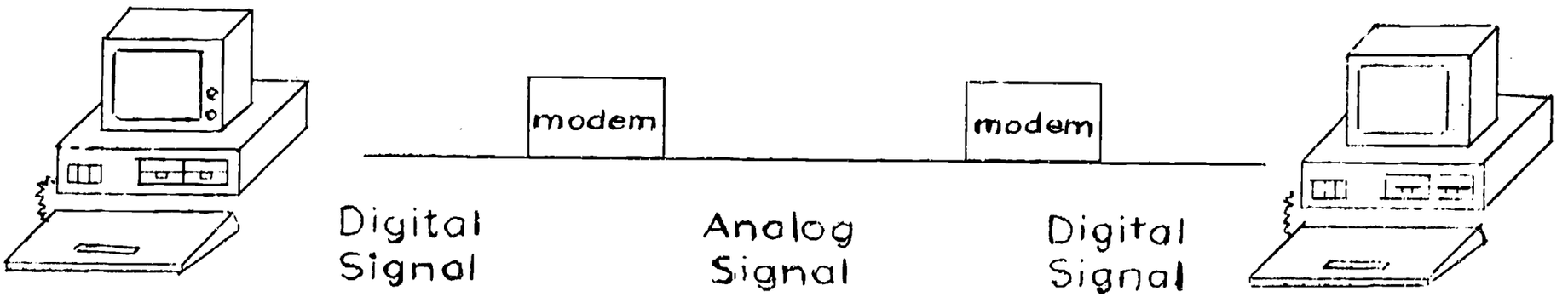

Figure 2

STAR TOPOLOGY

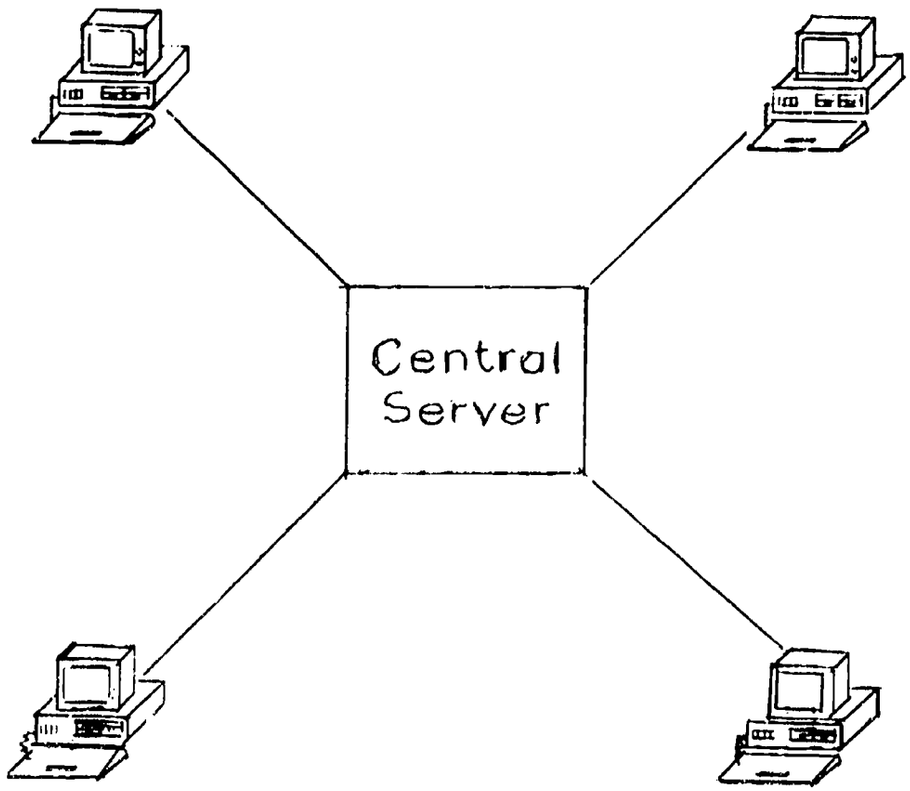

Figure 3

GATEWAY & BRIDGE