XEROX
BUSINESS SYSTEMS
Systems Development Department
July 7, 1978

To: Dave Liddle

From: Peter Bishop / SD at Palo Alto

Subject: A new class of I/O device: the OIS archive device.

Copies: W. Lynch, J. Weaver, J. Wick, S. Wallace, R. Sonderegger, R. Metcalfe, W. Kennedy, J. White, V. Schwartz, R. Belleville, T. Townsend, J. Mendelson, R. Wickham, C. Irby, L. Bergsteinsson, W. Bewley, J. Szelong, P. Heinrich, E. Harslem, L. Clark, D. DeSantis, J. Reiley, G. LeCesne, D. Reilly, Dave Thornburg, Brian Rosen, Bill Gunning

Introduction

Developments in LSI technology will have important effects on the cost of memory, and therefore upon the architectures of computers in the 1980s. Many people who try to predict what computer architectures will exist in the 1980s ignore the huge cost of developing software for drastically new architectures. This note assumes that basic computer architectures will not change, i.e. that the D0 will be a state-of-the-art processor (in the low-moderate price range). It is relatively clear that in the 1980s there will be two major factors in the cost of a computer system: peripherals and memory. Therefore the cost of memory can be expected to affect the system architecture. A recent report by Mackintosh Consultants Co. Ltd. [1] has taken a close look at the costs of memory in many different technologies. Using this report, we can estimate what technologies we will be using for memory in the 1980s.

An important factor that cannot be ignored is the effect of price competition from new memory technologies (such as bubble memory and CCDs) on the structure and price of old, established memory technologies (such as disk). One conclusion of the Mackintosh report is that there is no future for fixed head disk technology because by 1985 bubble memories and CCDs will be cheaper while offering a faster access time than fixed head disks. Mackintosh further predicts that beginning in 1980 bubble and CCD memories will be marginally cheaper than RAM, but by 1985 bubble memories will be one order of magnitude cheaper than RAM. The 1985 cost of bubble memories, however, will still be twice as expensive as current costs for non-removable moving head disks. Mackintosh expects the cost of non-removable moving head disks to drop to one-third of their present cost by 1980.

An interesting point made by the Mackintosh report is that it is questionable whether there will be a place for removable disks in the 1980s. The main reason for this is that a removable disk cannot be as tightly sealed as a non-removable disk. A tightly sealed disk can operate with closer tolerances that reduce noise and allow higher densities to be achieved on the disk. Thus although the cost per bit of non-removable disks will fall, it is not clear that the cost per bit of removable disks will also fall. A major advantage of disk memory is that it can be used as a pseudo random access device for the on-line storage of information. Cartridge disks allow the on-line information to be removed, but also to be quickly placed on-line on another system. As the relative cost difference between removable disk and non-removable disk increases, other methods of moving information from system to system and other methods of achieving off-line storage may be found.

We at PARC have already found that high-speed communications lines may be the best way to move information from one computer to another. When we begin to look at technologies that can achieve efficient off-line storage, we notice that magnetic tape achieves an extremely low cost-per-bit for off-line storage. Many people have noticed that magnetic tape is hard to handle, but the tape industry has answered this objection by developing tape cassettes. Tape cassettes are just as easy to handle as cartridge disks, but achieve a much higher surface area. An exciting prospect in the early days of integrated circuits was the possibility of producing three-dimensional circuits. Although this has not been realized in integrated circuit technology, magnetic tape realizes this ideal to a surprising degree.

Use of an Archive Device

There are basically two uses for an archive device: archive and backup. When used for backup, there would be one archive-device volume that contains the last complete dump of the system and all of the changes that have been made to the system since then. If the capacity of the archive volumes is not enough for this, then dumps may be placed on several volumes. When the amount of storage used for incremental backups becomes the same order of magnitude as the storage used for a complete dump, then another complete dump is taken. To restore the system, it is necessary to reload the complete dump and then replay the incremental dumps. The larger the capacity of an archive volume, the less frequently the archive volume will have to be replaced by an operator. If the capacity of an archive volume is 3-10 times larger than the on-line storage of the system, it may not be necessary to ever replace the archive volume, since an old dump can be overwritten with a new dump. In addition, an archive device can be used for archival storage. Reading or writing an off-line volume requires operator intervention to mount the off-line volume. It is acceptable to use a device intended for backup for archival as well since reading or writing archive volumes will only use the device for a short time. The backup volume would be dismounted, the archive volume mounted, the read or write performed, and then the backup volume remounted.

High Performance Requirements

This suggests that there is a place for a special archive device that has an interestingly different behavior from all of the classical digital storage devices. The basic requirements for this device are:

1) low cost for the device itself, since it is a special-purpose device.

2) extremely low cost per bit, especially for off-line storage.

3) high transfer rate - it would be nice to be able to dump a disk onto the archive device and be limited by the disk rather than the archive device. [It would also be interesting to be able to keep a log of X-wire traffic (or an interesting subset of X-wire traffic).] This requires a transfer rate of at least 10^6 bits per second, but transfer rates up to 10^7 bits per second would be nice.

4) very large capacity - we do not want frequent operator intervention in order to make use of

the high transfer rate. One reason for this is so that this device could be used on a system that does not have a full-time operator, but which is using the device for system utility functions, such as backup. Each volume needs a capacity of at least 10^8 bits, but 10^{10} bits would be more desirable.

5) high reliability - we should maintain high reliability even when an off-line volume is kept off-line for a long period of time. There should be a 90-95% probability of recovering all the information on a volume after being kept off-line for 10 or 20 years. In the case of tape, this means that we must be able to splice the tape in several places and still be able to recover all of the original information.

The archive device is intended to be used for backup and archival storage of information. Backup information will probably be read within a month of when it is written. Otherwise there is a low probability of it being read. Archival information will probably not be read for at least a month, at which point the probability of its being read begins to slowly rise until 6 months - 1 year, at which point the probability of accessing the information slowly decreases. Much archival information will never be read, but it has value because it could be read if needed. It is this ability to read that forces us to support reading with high probability. Many current tape systems require that each tape be rewritten once a year. This is acceptable for those who use high-cost archival systems. High cost archival systems are acceptable on high-cost computer systems, but the D0 is aimed at a market that uses much lower cost systems. It is now necessary to have much more efficient archival systems so that the cost of archiving will not be a major component of system cost. In addition, business is accustomed to low-cost archival of information (on paper), so the requirement for a low-cost archival system is even more stringent. The need for low-cost archival precludes the possibility of frequent rewrites of tapes and makes it interesting to develop a system that supports long-term passive storage of information. If the long-term passive storage of information cannot be achieved, then it is one of the more expendable requirements, but achieving this requirement would be a very nice feature on OIS systems.

Low Performance Possibilities

The requirements listed above all specify very high performance, which is in conflict with the first requirement for low cost. In order to achieve low cost, it is necessary to specify a set of performance axes on which extremely low performance is acceptable:

1) It is acceptable to support only rather large blocks of data, possibly as large as 32K bytes. In addition, there can be timing requirements on the computer concerning the amount of real time that can elapse between finishing reading (writing) the last block and beginning reading (writing) the next block. This allows the mechanical transport system to be very inexpensive and to have low performance. In the case of a tape system, there could be very poor start/stop times. This probably needs to be combined with an ability to begin a write pass by first starting the tape, then, only when the tape has come to speed, beginning to erase and rewrite the tape. The software would have to overshoot before beginning a read or a write.

2) The archive device need not be random access, although if it is based on tape, there must be a fast forward, i.e. random access time of less than 5 minutes is necessary. After a random access has been made, a very large amount of information will be transfered starting at that location.

3) There is no need to be able to do computation at the same time as data is being transfered to/from the archive device. Thus 50-75% of the CPU may be used in controlling the transfer of data. There must be enough CPU time left over to initiate disk activity in parallel with the archive device to allow the large buffers to be filled from disk. The ability to use a significant portion of the CPU during a transfer suggests the possibility of making a less expensive controller for the device and placing many of the functions that have classically been placed into the controller into microcode (or into LSI hardware). It also makes it easier to occasionally use software (see point 4) instead of microcode and thereby achieve read algorithms that are two or three orders of magnitude more complex than standard algorithms.

4) Although there is a need for high reliability of storage, it is acceptable for the read rate to drop by one or two orders of magnitude if the data is difficult to read. This means that fairly sophisticated redundancy schemes (performed by software) are interesting on an archive device even though they have played little role in traditional disk or tape systems.

5) A write-once limitation is acceptable. In fact, being unable to modify information that has already been written is an advantage for archival storage. In the case of backup, however, if the capacity of a single volume is 3-10 times larger than the on-line capacity of the rest of the system, then the ability to rewrite would allow a totally automatic backup system to be used. Even restarts could be invoked automatically.

Given a simple mechanical transport system (in the case of tape), the transfer rate of the archive device should be at least 20% slower than the long-term transfer rate of disk so that blocks in high speed memory need not be as large as blocks on the archive device.

Archive Device Technologies

There are two technologies that seem to be able to supply such an archive device: helical scan magnetic tape, and optical disk. The videotape machines that are currently on the mass market meet the cost, capacity, bandwidth, and mechanical specifications of the device. The VHS system (retailing for $900 and including VHF and UHF tuners) comes with 2-hour cassettes. The video system seems to use a bandwidth of about 2 megacycles (analog). This could probably be converted into 1 megabit per second (digital), giving a capacity of 7 x 10^9 bits per cassette. About 5 years ago, International Video Corporation marketed a digital helical scan tape recorder that met the requirements for an archive device and achieved a transfer rate of 8.1 megabits and a capacity of 7.5 x 10^{10} bits on a 7000-foot reel of tape (it used cartridge tape). I think that this product failed because it was not quite time for it, appearing as it did in the heyday of removable disks, and I think it did not take full advantage of the low performance possibilities of an archive device. The helical-scan technology is here, and is mature. Such a device could be produced in a short time. Currently, DEI is marketing rather successfully an inexpensive device (OEM $1500) that begins to take advantage of some of the low performance possibilities of an archive device and begins to supply some of the high performance features. It only has a transfer rate of 200Kbits and a capacity of 10^8 bits per cassette, however, because it does not use helical scan recording.

The other interesting technology for an archive device is optical disk. An optical disk would have much better random access properties that any tape system, but its ability to only be written once makes it less useful for highly variable data. A write-once property is acceptable for an archive device, however, and the high transfer rate (10 megabits) and high capacity (10^{11} bits) meet the requirements for an archive device. In addition, the expected low cost for the optical disk make it possible to use an optical disk as an archive device. An archive device, however, needs to be able to be devoted solely to system utility functions such as backup. The backup volume should be permanently mounted so that incremental dumps can be made automatically by the system. An optical disk may be too valuable a device to use in this way because of its random access abilities. In addition, an optical disk has almost no competitive advantage over helical scan magnetic tape for use as an archive device, since random access is not important. The capacity of an optical volume can be increased significantly by using optical tape, however. Some estimates are that optical tape could store up to 10^{13} bits per volume. Such high density for such low cost might require helical scan optical tape. This seems to be the best device to use for an optical archive device.

Begin Work on Using Archive Devices in OIS

Xerox should make a place for an archive device in OIS architecture and should begin developing prototype archive devices using helical scan magnetic tape. By the time we have been able to iron out exactly what the requirements for an archive device are that enable it to be inexpensive and yet useful, the optical tape technology may have developed far enough to allow an inexpensive optical tape device to be used instead of helical scan magnetic tape. In any case, work should begin now on exactly what redundancy schemes or special hardware is needed to ensure high reliability on off-line tape even when it is stored for 5-20 years. Few computer applications have seriously attempted to provide high reliability at low cost under such conditions.

If archive devices are developed for other systems than Xerox, they are likely to not have a feature that could be taken advantage of by D0-based systems: high memory bandwidth, and microprogram control (or specially-tailored LSI hardware). Digital tape drives will probably place more function in the controller than is needed for a D0-based system. This can have two serious drawbacks: a) it may make it difficult for software to make use of any redundant information placed on the tape by the controller, and b) it may be difficult to add additional redundancy that is only used if the normal error detection finds a hard error. In particular, it may be interesting for software to be able to find which bits the hardware knew had been read incorrectly because the FM rules were violated. This would allow redundant information to be used more efficiently for error correction. Using XOR redundancy techniques it appears to be feasible for 5% redundant storage (in addition to error detection redundancy) to protect against a loss of 2% of the storage medium.

Some Alternatives

There are several alternatives that can be considered for providing backup/archival on OIS systems:

1) use removable disk - this appears to be attractive until better archive devices are developed and until removable disks are shown to be too expensive or in other ways inadequate for this purpose. We must beware of providing a removable disk drive for backup/archival purposes and having it actually used for on-line storage.

2) 9-track tape - this is acceptable for systems on which the cost of the backup/archive system is not important. The difficulty with most 9-track tape drives is that they are very expensive and have low transfer rates and capacities, thereby requiring frequent operator intervention.

3) high density cassette - this device is already included in the OIS Configurations document. It would be interesting to explore redundancy schemes that could be used on this device to provide high reliability over a long time.

4) helical scan magnetic tape - Xerox could produce and sell archive devices for OIS systems based on helical scan magnetic tape. Xerox could contract with a manufacturer of helical scan video tape systems to produce most of the unit, with the electronics and possibly a few mechanical additions being made by Xerox. Before actual manufacturing of such a system, it would probably be beneficial to construct a prototype using an off-the-shelf video tape recorder with mechanical and electronic additions and/or modifications that allow it to be used as an archive device. This prototype might allow work to begin on the software and/or hardware that is needed to allow long-term passive storage and to achieve high reliability.

5) optical tape - The most promising technology for an archive device is optical tape. Xerox should seriously consider becoming a manufacturer of optical tape as well as optical disk. In this case, serious work should begin as soon as possible to solve the problem of long-term passive storage of information on optical tape. This work should not ignore the possibilities of using complex redundancy schemes during reading of the tape to solve some difficult problems that may arise.

References

[1] Mackintosh Consultants Co. Ltd., "Serial Memory Study", a multi-client study, not to be disclosed outside of Xerox, Mountain View, California, 1976, 2 vols.

Appendix

This is an interesting quote from the Mackintosh report on the product prognosis for cassette and cartridge tape drives:

> The attractiveness of tape transports is clearly for applications involving off-line storage and data interchange. In on-line applications moving head disc storage is now cheaper, has two to three orders of magnitude faster access, and higher data rate.
>
> Recent trends to fixed disk moving head storage, both at high and low capacities, are indicative that disc and tape drives will serve complementary rather than competitive functions in future mainframe and minicomputers. Tape provides low cost off-line bulk storage of data and programs in a form convenient for interchange, while disc drives provide facter access for on-line application. This being the case, there is a requirement for tape handlers designed specifically for off-line storage, and rapid transfer from disc to tape. Long blocks of data could be used at high packing and track densities, and, with parallel operation of tracks in tape and serial operation in discs, data rates could be matched. There would be less stringent requirments for start/stop times and at reasonably low cost.
>
> Helical scan rotary head recording, widely used in low cost video recording, could be readily applied to real time transfer from disc to tape since the mode of recording is serial and the data rate could be modified to that of the disc drive.
>
> Half-inch tape systems, e.g. the Emmerson Cartridge-type Tape Pac, described in Section 2.5.2, may well replace open spools in mini-computer applications, providing sealing from contamination and convenient off-line storage.