The Impact of Technology on Information System

W.K. Liebmann
IBM Laboratories, Böblingen, Germany

1. Introduction

An information system receives information in the form of data and commands, it analyzes this information against a preprogrammed set of algorithms, and it issues the results of this analysis - again in the form of data and commands - for further use. This general definition of the tasks of an information system leads automatically to a description of its major hardware components. The system needs input equipment to receive information, it needs a data processing unit to execute the analytical portion of the task, and it needs output equipment to inform the user of the result of the analysis.

To perform the analytical tasks, the data processing unit usually requires fast access to a large collection of background information related to a particular problem. In modern information systems, these background data are stored and organized in data banks. Thus, in addition to input, output and data processing equipment, external storage facilities for data banks become the fourth major hardware component of an information system.

The major hardware components are linked together by equipment designed to facilitate the communication between them. Fig. 1 shows schematically such a basic hardware configuration of an information system.

The base technology for all information system components is quite different. The input/output devices rely heavily on mechanical or display technologies, the external storage devices rely almost exclusively on magnetic media technologies, while the central processing units (CPU) are essentially an embodiment of the silicon large scale integrated circuitry (LSI).

In analyzing the impact of technology on the various components it is, however, quickly noted that also in input/output and in external storage devices, LSI is the major technological driving force, and that in CPU, input/output and external storage alike, information systems progress is most dramatically stimulated through the very rapid evaluation of the silicon integrated circuit technology.

In assembling a report on the impact of technology development on information systems it appears thus most prudent to first evaluate the state of the art and the most likely evolutionary directions of the silicon integrated circuit technology and then - with this background - proceed to an impact assessment of the major information system hardware components.

2. The development of the silicon integrated circuit technology

Silicon - in the form of a great variety of oxides abundent in the earth crust - has a number of physical and chemical properties, which make it uniquely suited for the design and the fabrication of integrated circuits. Silicon is a semiconductor whose band gap of 1.1 eV is such that through addition of appropriate doping elements electronic devices structures, like transistors, diodes resistors, capacitors, can be obtained which operate efficiently at ambient temperatures, for instance in the range between -20 and 100°C. Silicon can easily be (grown into) the form of very pure, defect free single crystals, is machinable with standard tools to obtain very perfect surfaces of a particular crystal orientation and is mechanically stable enough to withstand handling in a modern technology environment. Most important, however, is the fact that, when placed into an oxidizing atmosphere at elevated temperature, a very dense silicon oxide layer grows on the surface of the semiconductor, which adheres well to the base material and which forms an effective diffusion barrier to most of the doping elements which are of technical interest with respect to the generation of the desired electronic properties in the bulk semiconductor. Such elements can only penetrate into the semiconductor surface where the oxide layer has been selectively removed and where the bare semiconductor surface has been exposed to the doping atmosphere. Hydrofluoric acids are efficient etchants to remove the oxide layers with very little attack to the base material. A variety of photoresist materials is available which are resistant to the attack of the hydrofluoric acids. By microphotographic means a pattern can be transposed unto the surface where certain sections of the oxide are shielded from acid attack through the photoresist, while in other sections the oxide has been exposed in the photoresist developing process and can then be removed in the subsequent etch.

The combination of all these properties has rendered to the silicon an exceptional position among all materials which find application in modern technology, and if one looks at its pervasiveness in our high technology world and assesses its impact on our total lives, then it might very well be appropriate to call our time the age of silicon.

Silicon semiconductor components made their entry into commercial application in the fifties; the first integrated circuits, where several components were functionally interconnected on the same silicon chip a few square millimeters in area appeared in the sixties. Today, several thousand circuit elements are integrated into the same area.

The driving force which brought about such rapid technological development was of course entirely economic in nature. The basic manufacturing unit which is processed through the semiconductor lines of today's electronic component manufacturers, is a "wafer" a slice of silicon 75 to 125 mm in diameter (Fig. 2). In a given environment, the cost of processing such a wafer through the various diffusion, oxidation, evaporation and photoresits steps is essentially a constant, independent of the amount of electronic function contained on the wafer. Thus, the more circuit functions can be integrated into a certain silicon area, the cheaper the individual function is going to be.

In driving towards higher and higher integration densities in order to optimize production cost, several very positive side effects appear. Device dimensions and distances between devices become smaller, thus reducing capacitances and carrier travel times, all leading to better circuit performance or power/performance. The number of silicon chip to package connections per functional unit decreases, which in turn leads to better unit reliability, and the number of silicon chips, which must be packaged to obtain a certain function, and consequently their packaging cost decreases.

There are two forces which counteract the rapid progress of LSI: one stems from the fact that in designing so many functions into one individual unit like a silicon chip, the design complexity increases so rapidly that design errors are bound to occur, which will only be detected in a functionality test of the completed hardware, and which lead to expensive and time-consuming design iterations.

The second force stems from the increasing test complexity since a great number of LSI circuits, whose functionality must be individually verified in test, are only accessible through a very small number of input/output contacts. Both sectors, the design and the test, however, are amenable to data processing aids, and with continually improved price/performance of DP equipment, both obstacles to design progress can be overcome efficiently. LSI-designs can be software-simulated and error checked before they are modelled in hardware, which in effect eliminated the need for hardware engineering change recycle. With appropriate combinatorial analysis and iterative test patterns, also the most complex array of internal functions can be tested to a satisfactory level through just a few input/output pads. As an example: At the end of our semiconductor manufacturing line approximately 500 000 test patterns are applied through 23 pads to IBM's 2000 bit random access monolithic memory chip [1]. The test robot requires a few milliseconds to perform this test, which identifies the good and the bad chips on a wafer. Approximately 30 000 test combinations are required to establish functionality a chip with a few hundred circuits and 100 pads; such test patterns can be generated automatically from a particular logic design and can be applied to the chip in a small fraction of a second.

Invention an evolution in three areas stimulated the fast progress of the silicon integrated circuit technology:

1. The improvement of photolithographic dimensions. With today's photolithographic equipment and environmental control, geometrical shapes **with** dimensions approaching the wave lengths of visible light can be transposed into a semiconductor wafer. Electronbeam Lithography and y-ray lithography [2] [3] will extend the dimensional capabilities downward by orders of magnitude. With both technologies the desired device pattern will be directly enscribed into the photoresist-covered silicon wafer (instead of first generating a photomask from which the pattern is transferred to the wafer), and both technologies incorporate the potential to register subsequent exposures to the previous-one automatically which leads to a minimization of the dimensional tolerances that must be provided to account for misregistration of one photomasklayer against the previous one.

2. The integrated advancement of semiconductor processes and semi-

conductor device design. Ten years ago a circuit designer, a semiconductor device designer and a semiconductor process designer were three independent agents who communicated with each other through a set of rules: Today, they work as one team, often this team is integrated by one person with detailed knownledge and experience in all three areas, and synergism takes place to advance the total state of the art faster than the contribution of the individual components would otherwise have permitted. Typical examples of this synergism is the one device FET (field effect transistor) dynamic monolithic memory cell [4] as shown in Fig. 3. This invention, which is the basis for practically all of today's cost-performance FET monolithic memory chips was stimulated by increasing the bit-density on the silicon to increase the bit productivity and decrease the monolithic memory cost. It leads to a bit cell area which is only 8 unit square, where the dimensional unit is the minimum photolythographic line width of the particular semiconductor process. 4 unit square is already required to achieve a pattern of two lines crossing each other. Its practical implementation depends on a semiconductor technology which produces sufficiently low surface and junction leakage currents to retain the capacitive charge at the storage node.

In modern semiconductor processes this control of leakage currents is so good that the time intervals after which the stored charge must be replenished are very long compared to the time required to actually recharge the capacitor, which means that leakage current and recharge considerations have essentially become negligible factors in the operation of one device monolithic memory arrays. Similarly, the desire for smaller and smaller cell areas and consequently less and less stored charge aroused the inventiveness of many circuit engineers to design very efficient FET sense circuits with which the small signals originating from just a 100 000 or so electrons could be reliably sensed [7].

Another example for this synergism is the superintegration of MTL (Merged Transistor Logic) [6]. Fig. 4 shows the semiconductor layout and the circuit schematic of a bipolar MTL - random access static memory cell. This cell requires only 30 unit squares of silicon area, is considerably faster than the FET - one device cell and is static, which means: it maintains its charge as long as the power of the system is turned on and does not require refresh during the operation. The dense layout was made possible with a bipolar semiconductor process which provided inverse n-p-n transistors and lateral p-n-p transistor of sufficiently good characteristics to obtain cell functionality and stability, a circuit design which replaced the large area current limiting resistors with lateral p-n-p transistors and a semiconductor layout highly adaptive to the particular structure of the bipolar semiconductor technology design. Integrated circuits have come a long way from their beginnings where the individual components of a semiconductor circuit were all merely buried side by side into the silicon and then interconnected at the silicon surface.

3. The design of reliable metal/insulator systems for dense, low impedance interconnections on the silicon chip itself and between the chip and its carrier. Particular examples are the use of double layer polysilicon for interconnection and functionality of FET circuits [8], the use aluminum/quartz systems for multilayer chip interconnections [9] and the chrome-copper-gold-lead-tin system for chip to carrier interconnection [10].

The LSI progress has been most dramatic in the area of monolithic memories since the functional and structural regularity of a memory array is most compatible with the capabilities of the semiconductor technology. Monolithic memories entered the data processing market on a large scale with the introduction of the monolithic main memories in the IBM systems 370-135 and 145 [9], which were announced to the market in 1971 and 1970 respectively. The storage unit then was a silicon chip containing 128 bits on about 20 mm². Today, 16K bits integrated on essentially the same area are available in large productions quantities [8], the introduction of 64K bit chips is announced [12], and serial readout monolithic memory chips in CCD (Charge-Coupled-Devices) implementation carrying 128K bits are available. At the high performance end, 1K bit chips with chip access time of 10 Nsec are available.

Performance and density can be traded against each other over a wide performance range as shown in Fig. 5 where the cell area of certain monolithic memories is shown as a function of chip access time.

Projecting into the future the serial access monolithic memories (e.g. CCD`s) will continue their density lead by a factor 2 to 3 over random access memories because they are less leakage current sensitive and allow the design of simpler on-chip signal sensing circuits. In both memory types, random access and serial memories, the limits of the capabilities of the silicon technology are far from exhausted, and the rapid increase in memory chip density will continue, especially since after the introduction of electronbeamlithography no further physical boundaries to further advancement are apparent. The ultimate limits of RAM density is, of course, a subject of professional speculation. Liebmann [a] estimates that this limit in silicon will be approximately 10^7 bits/cm^2, while Mitterer [13] suggests that we can reach 10^9 bits/cm^2. In either case their is plenty of opportunity for further density improvement, which will result in further dramatic memory cost reduction, if the monolithic memory market has the power to absorb all the memory bits. But, right now, there is no saturation in sight. Even with the most dramatic cost reductions, however, it remains questionable whether the monolithic memories will ever be a real cost competitor for the disk storage devices which will probably continue to show better bit costs by an order of magnitude.

The most interesting competing technology to monolithic memories are the magnetic bubble memories [14]. In certain materials with a strong magnetic anisotropy (Orthoferrites, Hexaferrites, Garnets), small cylindrical domains can be formed through the application of an external magnetic field, where the magnetization direction is opposite from the magnetization direction of the bulk material. Fig. 6 shows schematically a magnetic bubble in a slice of anisotropic material. The bubble can be moved by applying externally a magnetic field gradient: it will move into the direction of the lowest external field. The presence of a magnetic bubble at a certain place at a certain time can then be the definition for a digital "one", the absence of the bubble can be digital "zero".

To initiate motion of the bubble along an exactly predetermined path, the gradient fields are structured along the surface by certain patterns shaped from soft magnetic materials, like for instance permalloy. Fig. 7 shows the motion of magnetic bubbles under the influence of an external rotating field in a "T"-bar environment.

The bubble memories are serial access memories by the nature of their storage mechanism. Their performance is gated by the mobility of the bubble through the bulk material. It will always be considerably slower than the performance of semiconductor memories. There is no inherent density - and thus probably cost advantage - of the conventional magnetic bubbles vis-à-vis semiconductor memories; bit-densities of $10^8/cm^2$ are probably achievable. This limit may be extended to considerably higher densities with the "Bubble Lattic File" [16)] where the bubble size can be reduced to the order of magnitude of the crystal lattice of the base material itself. This concept, however, has currently only an advanced technology character.

The main advantage of bubble memories over semiconductor memories is the fact that they are non-volatile read-write memories, which retain their information content also if the power supply current is switched off. They will find initially their major application in those areas where small quantities of non-volatile storage are required, too small to afford the high entry cost of a rotating disk storage device. This could be in point of sale terminals, in small distributed processors which only periodically are linked to a larger central data processing installation, in portable equipment etc. It is very unlikely that the bubbles will displace the monolithic memories from computer main store or control store applications. Even though the bubbles are considerably slower than monolithic memories, they are much faster than rotating disk storage and as such good candidates to take over part of the "fixed head file" market, especially in a hierarchical memory organization (see chapter 4).

Silicon chips for logic applications today carry several hundred to several thousand circuits. In these logic chips depending on the design methodology an interesting trade-off between design effort and manufacturing cost can be made. This trade-off capability is characterized through the three basic design principles by which a certain logic function can be implemented with silicon integrated circuits.

1. A logic design, which represents the desired system function in an optimum manner, is translated circuit by circuit into a semiconductor layout. The result is a semiconductor design of very high circuit density and very good performance. The chip, however, does only represent the particular single function in a "custom design", and is not adaptable to other functions. Any change in the function or any error in circuit or semiconductor design will

result in a full E/C (engineering change) cycle, with associated time delay and cost. The custom design approach relies heavily on density and performance optimization through "hand-honing" and it is thus not very amenable to computer automatization. The design of such a single part number consequently is expensive and only worth while if the total quantity of pieces which are to be manufactured in this custom design is very large.

2. On the opposite side of the semiconductor design spectrum is the "master slice" approach. The attempt is made here to generate an universal semiconductor layout which can implement every desired logic function. This is achieved by generating a standard array of logic circuits of the silicon chip and by providing means to interconnect these circuits differently for each application. The advantage here is - of course - that the semiconductor design only must be done once for a multitude of part-numbers and that in the event of an E/C only the interconnection pattern on the surface is affected. The disadvantage that the universality of the circuit design and the flexibility required to wire widely varying functions results in a loss of circuit density on the chip, and thus in higher manufacturing cost per circuit. The master slice approach is most efficient in the case of many different part-numbers where the quantity of the individual part number is low.

3. The third approach is the implementation of logic functions with the help of a "microprocessor" where the attempt is made to combine the advantages of good manufacturing cost from custom design and high volume for a single part number with the flexibility to implement different logic functions. A microprocessor is a small computer, consisting of a central arithmetic unit, associated control and required input output circuit, scaled down so that everything fits on one silicon chip. In that sense the microprocessor is not a fixed function, but will develop into more powerful computing units as the semiconductor technology progresses, as could be seen from the appearance of originally one bit microprocessors to now 16 bit microprocessors [17)]. The function of the microprocessor is adapted to the particular logic task through an appropriate micro-control program which is stored in an accompanying read-write or read-only store. In manufacturing cost, the microprocessor ranges between the custom design which optimizes circuit utilization, and the masterslice which optimizes flexibi-

lity. The major advantage of the microprocessor is, however, that it clearly separates the function of the design of a silicon integrated circuit chip from the implementation of the function, which the microprocessor later is to perform, and thus permits to take advantage of the functional and cost benefits of the silicon integrated circuit technology without the need to be a semiconductor expert or without disclosing the details of the application to the semiconductor designer or manufacturer. The semiconductor producer, on the other hand, can sell one standard part to a wide variety of customers, with the cost advantages brought through economy of scale and without the need to worry about the application details of his customers. Resulting from this almost ideal situation for all parties concerned is a rapid expansion of microprocessor usage into all technical sectors of our lives.

In silicon, it appears quite feasible to integrate $2x10^7$ logic circuits per cm^2. The circuit count of very large data processing units, like for instance the large IBM /370 CPU`s (Central Processing Units) is in the order of magnitude of 10^5 circuits. Comparison of semiconductor capability and circuits required to implement certain functions show that soon we will have very powerful processing units on one silicon chip. Since there are no homogeneous islands of several thousand circuits in logic designs, these highly dense logic chips will carry a functional mixture of logic circuits interdispersed with read-only and read-write arrays or with circuits which facilitate input/output communication or testing. The master-slice, custom design and microprocessor design approach will thus merge, and on future logic chips we will find master-slice like sections in areas with a high frequency of engineering changes (for instance: control circuits) and there will be custom design for other areas like data flow logic or imbedded arrays. In any event, the "microprocessor" of the late nineteen-eighties might very well represent a data processing power of several MIPS (million instructions per second) on a single chip.

In summary, the technological progress of LSI will continue. There is no saturation of the market or the technological capabilities in sight. Also, there is no real competitive technology to displace the silicon from its predominant role. For logic and memory of very large central electronic complexes the "JOSEPHSON-TUNNEL DIODES" [18)] might eventually qualify. With their limited operating range at superconducting temperatures, it is difficult to imagine that Josephson-devices will ever find the pervasiveness which silicon devices enjoy today.

3. Input/Output

The interface between the user of an information system and the system itself is the input/output equipment. The user could of course be a machine or a robot, but for the purpose of this discussion it will be assumed that the user is a human being. This human user interacts with the information system in a request/ response mode: he addresses a problem statement to the user and expects after a reasonable length of time an answer. In an interactive man-computer environment (and I will limit my comments to this environment), this answer will normally be used to formulate the next problem statement, which receives again an answer, and so on until in moving from general to specific the desired information is obtained. For such interactive operation an interactive input/output terminal is required, with whose help the person can communicate its problem statement to the information system, and which can present the system`s answer in a manner a human being can understand. For practical purposes, the only mechanism at our disposal to communicate to the system are our tactile facilities, the most efficient way to receive the answers from the system is by means of our visual input channel.

There are certainly other means for human-computer interaction. Voice input and audio output are feasible but at present not very efficient. Voice input to the system becomes difficult because of the inherent complexity in the computer analysis of the received message since the human language has a wide spectral distribution and is rich in dialects and synonyms. Voice input nevertheless is potentially useful as long as the application can get along with a relatively limited and simple vocabulary. Portable audio input terminals for stockroom control, where the vocabulary essentially exists only of partnumbers - e.g. the digits zero to nine - and simple commands, might be feasible. Audio output is considerably simpler than voice input, since the analysis of the output message can now be delegated to a very efficient data processing system, namely the human brain, who can easily cope with the problems of a synthetic computer language. But also in audio output the application will be limited to the transmission of relatively simple messages since our interpretive capability of audio signals is far inferior to our visual interpretative capabilities. Special applications of voice output are - however - quite successful, like for instance the use of a synthetic language to communicate between a computer and a blind person [19]. Long range proposals for man-computer interaction include the use of devices for direct electronic coupling to the human brain and transfer directly from the brain waves, with a direct computer interpretation of the ELECTRO-ENCEPHALOGRAM [20]. This input/output channel, though theoretically very broad and fast, can however probably be discounted for practical use in the near future.

The most versatile I/O (input/output) terminal for todays interactive operation is the CRT (Cathode Ray Tube) terminal (Fig. 9). It consists basically of the following components (Fig. 10)

- a keyboard for message input using the human tactile facilities
- processing equipment which interpretes this message and converts it to a bit stream suitable for data processing
- a communications facility to communicate the bit pattern according to a certain protocol to the central processor of the information system, and to receive messages back
- a facility to convert the message into a signal pattern suitable for display on the cathode ray tube,
- the display head itself.

These basic elements are supplemented with devices for control and checking of the terminal and the communications operation.

The CRT-terminals have some unique advantages which make them ideally suited for their task:

- The keyboard essentially taken from a standard typewriter keyboard, is ideally adapted to the tactile capabilities of the human hands.
- Its operation is quiet. Several terminal operators working in close proximity, do not disturb each other excessively.
- The CRT is versatile and can display numerals, alphanumeric characters and even graphics. Size of characters and fonts can easily be changed under program control.
- There are no moving parts, resulting in very good reliability.
- Production costs are low through technology commonality

with the display devices used in modern television.

Technological deficiencies are:

- The image which is generated on the CRT screen is volatile and needs continuous refreshing. The CRT thus requires an image buffer from which the information to be displayed can be retrieved approximately 50 times per second, to guarantee flicker free operation.
- The CRT requires very high operational voltages, and the power consumption is high.
- The weight of the CRT is high.
- The dimensions are bulky due to the distance required between the cathode ray generation and deflection devices and the screen. Bulky dimensions and great weight render the CRT not very suitable for portable devices.
- The maximum display capability is approximately 2000 characters on the screen surface.

Technological progress in the keyboard and the display section of the CRT terminal is largely exhausted. There is room for very solid product engineering and subsequent cost reduction, but it is unlikely that dramatic new technoligical developments will occur. The real technological progress in the CRT-terminals will be stimulated through the progress of the silicon integrated circuit technology. And here it will not so much be a reduction in the price of the CRT unit, but it will be an extension of the terminal functionality which can be obtained for a certain price. The various steps which characterize this expansion of CRT functionality as driven by the silicon integrated circuit technology are clearly discernable:

- First the hard wired logic functions were replaced by monolithic read only memories, the read only memories were personalized by the supplier who now had the flexibility to quickly react to new market requirements just through repersonalization of the ROS (read-only store), but without changing the base design. The character set could be updated easily, the CRT could be quickly adapted to different languages.

- Additional functional flexibility was obtained by executing more and more of the logic and control functions of the system with the help of microprocessors. At very low cost, a standard terminal could be adapted to new emerging functional requirements. The technical absolescence of a certain terminal type was delayed. The user himself could alter the functional characteristics of a terminal by exchanging certain program modules, stored on monolithic read-only chips.

- Next was the advent of the "intelligent" terminal, where enough processing power was added to the CRT so that either different options in existing programs could be exercised, or that the user could develop new programs for specific functions. An example would here be the formatting of the screen under user program control.

- Further advancement of price/performance of semiconductors led then to the completely programmable terminal controller, integrated into the CRT-unit, with its own read-write memory of sufficient size to offload certain processing functions from the central processing unit to the terminal and thus reducing the communications traffic between the host and the terminal.

- Along with expanded functional capabilities went the improvement of circuits for error detection or correction, for data compression or expansion for processor/terminal communication, and for control of the communications traffic and protocol.

All these advancements have made todays CRT-terminals very flexible data input/output station with enough processing power to format the message to the central processing unit, to check it for completeness, to verify its compatibility with the communication protocol, and to detect input errors and in many cases correct, or at least flag them. The format in which the answer from the processing unit is to be displayed can be adjusted to the specific requirements; often repeated program routines can be stored in the terminal under a certain function key, and new routines can be programmed directly at the terminal. The semiconductor integrated circuitry has made the CRT-terminal into a data processing station in its own right.

The most promising competitive technologies are the gas plasma panels [21] light emitting diode arrays [22], and liquid crystal devices [21]. The advantage of these technologies over the CRT-technology is that they do not require refreshing, operate flikker-free, they have a flat screen, are compact built, use low voltage and permit selective erasing of part of the screen. They are, however, not cost competitive with the CRT`s because they cannot share in large consumer market of todays television industry.

In summary then, the CRT-terminal is a uniquely suited interactive input/output device to a communication system. Its major technological progress is characterized by expanded functionality and flexibility with the help of modern integrated semiconductor circuits. No major breakthroughs are to be expected in the CRT base technologies.

External Storage

External storage for communication systems is characterized by the increasing need for larger and larger quantities of on line data. These are data which are always at the disposal of the system and do not require a preparation or set-up time (for instance: fetching a data tape from the library and installing it into a certain tape unit). Such online data are stored on large disk storage devices (DASD) or on online mass storage devices, like for instance the IBM 3850 or the CDC 38500 [24]. I will direct my analysis towards the disk devices, because many of the technological conclusions reached for them do also apply to modern on-line tape devices.
Rotating disk devices are used to store the operating system of the central processing unit, user data and data banks. The largest disk storage devices today have a storage capacity of several hundred megabytes [25]. Parallel to the requirement for more on-line data is the trend towards the "non-removable" disk, where the data storage device cannot be removed from the drive mechanism and transported to another drive spindle. This trend to non-removable disks very well also accomodates the engineering requirements for high precision parts and cleanliness of the atmosphere surrounding the disk, to assure an error free operation.

The key elements of a rotating disk storage device are shown in Fig. 11. Several concentric drives are mounted on a drive spindle and rotate at constant speed. The disks are covered with a thin layer of magnetic material into which or from which the read-write head writes or reads the stored data. To gain maximum efficiency in the read/write process the gap between the read/write head and the magnetic medium must be as small as possible. This is achieved by flying the aerodynamically shaped head on a thin air cushion which is less than one micrometer in thickness. The data are recorded on concentric tracks; there may be a thousand tracks on one disk surface. Under program control the head moves from one track to another to deposit or fetch data. The critical performance parameters of the device is the access time which is the time elapsed between a program command to fetch certain data and their availability at the system channel.

This access is determined first of all by the time it takes to mechanically move the head from one track to the desired next track, and the rotational delay from the time in which the head gets unto the track until the starting address of the desired data set has rotated under the head.
The portion of the access time which is due to the arm moving from track to track can be eliminated by assigning one head to each track. Because of the geometrical restrictions the consequence is a much wider spacing between tracks and consequently a reduced storage capacity of the disk and a higher cost per bit. Such "fixed head files" are only affordable for small sections of storage where a very high access speed is essential.

Analyzing the impact of technology on DASD devices is best done by analyzing the impact on the major functional characteristics of the storage unit. The characteristic parameters are:

- The total storage capacity
- The cost per bit
- The access time
- The rate with which data can be transferred from the disk once the access to the start of data set has been made.

The storage capacity is determined by the number of tracks which can be accomodated per unit diameter of disk, the total diameter of the disk, and the "bit cell length", that is the length along a track which is required to store one single bit. The parameters determining the bit cell length are displayed in Fig. 12. They are the head to disk spacing, the head gap size and the disk coating thickness. The development of these parameters as a function of time, together with the development of the bit cell length is also shown in Fig. 12. The bit cell length changed from 5x10^{-3} cm/bit for IBM`s 1301, which was announced in 1961, to 5x10^{-4} cm/bit for IBM 3350 announced in 1975. These dimensions have all reached the size of the magnetic particles and the limits of machineability and manufacturability. There will certainly be continuous product engineering improvement, but further dramatic progress of these parameters cannot be expected. Here again the LSI circuitry can help to overcome technological deficiencies, especially through better error correction circuitry or better sense amplier circuits integrated into the head, to sense the small magnetic signal.

The number of tracks per unit diameter is determined through the positioning accuracy of the head on the track. The mechanical head positioning devices are optimized to a level where again dramatic improvement is difficult to project. The major improvement here also can come with the help of LSI, to optimize the sense circuitry which signal the correct positioning of the head on the track.

The size of the disk and thus the total number of tracks is limited by the fractural strength of the material (centrifugal forces) and problems in the machineability of the very large disks. This dimension is also unlikely to improve very much.

In summary then, only gradual advancement from todays storage densities of several hundred thousand bits per cm^2 is likely to occur.

The cost per bit is determined by how many bits can be accomodated within the fixed costs of one drive spindle, one power supply and control unit. It is clear from the storage density arguments that low bit costs are most easily achieved with very large storage capacity sizes, while it will remain difficult to produce small capacity devices for small computing devices with still low bit cost.

The bit cost will gradually continue to improve with improving bit density and through additional leverage of cheaper, more reliable LSI circuitry.

The access time can also be improved through better control circuitry, rather than through the improvement of the mechanical elements. In modern disks, for instance, the motion of the head is controlled by microprocessors, who determine the acceleration of the arm, its maximum speed, and they decelerate the arms motion optimally so that the head gets to stop exactly on the desired track, without any overshoot.

Access time improvement is of course possible through hierarchical disk storage structures, through combination of small sections of fixed head file storage, backed-up by large capacity mobile head storage devices. Serial solid state memories (CCD or magnetic bubble devices) may here replace the fixed head files.

The data rate will also show gradual improvements in the same manner as the "bit cell length" decreases since the rotational speed of the disks is close to its technical maximum.

All together, the LSI circuitry will be the major contributor to DASD improvements, through better control and checking, better error correction, through integration of several disk drives into one large storage subsystem, improved electronics for the control of the channel traffic to the central processing unit. These advances of the electronic portion of the rotating disk storage units are one of the prerequisites which make the efficient operation of large data banks possible.

5. The Central Processing Unit

The central processing unit which consists of the central arithmetic unit, the main memory and control store, service and maintenance facilities, channels to external storage and communication facilities to the various input/output devices, has shown the most dramatic price/performance improvement. Since 1960, the price to execute one microinstruction has dropped by more than a factor of 100 [27]. There is no saturation in sight, and driven by the advancement of the silicon integrated circuitry, the price/performance improvement in the CPU will be much more dramatic then either in Input/Output devices or in external storage.

A major impact was the replacement of core-memories in the CPU's mainstore with monolithic memories. The largest monolithic mainmemories on todays CPU's have a capacity up to 16 megabytes of read-write storage. The driving forces which provoked this change - despite the enormous technological success of the magnetic core memory technology - were the monolithic memories better cost, reliability, lower power consumption and volume, and their technological commonality with the logic circuitry surrounding the actual storage array.

In addition to better absolute costs, the monolithic memories also had the advantage that their costs are practically independent of the memory size (core memories always had to carry the fixed overhead of the expensive silicon integrated ciruitry for driving, sensing and decoding, which could best be amortized over very large bit quantities) so that also small, independent memory units could be designed which could be distributed through-out the CPU, embedded directly into the logic functions.

This fact stimulated very much the design of microprogram controlled CPU's, where the microprogram is stored in monolithic read-write control stores. Many engineering changes or new functional requirements of the control functions can now be handled on the microprogram level, that is through software, without the need to redesign the CPU hardware. Monolithic memories can easily be designed with facilities for automatic error correction and detection which effectively hides all technological deficiencies from the user.

The cost/performance trade-off capabilities in monolithic memories, as shown in Fig. 5, has lead to the design of memory hierarchies, where a main memory of several megabytes in size is buffered by a smaller cache memory of much higher performance. The bit capacity of the caches are usually only a few percent of the main memory (example: IBM 370-168: Cache: 32 Kilo-Bytes Main-Memory: 8 Megabytes) which can accordingly be more expensive. The operation of the hierarchical memories then is such that the CPU addresses first the cache, when a certain information is required. When the information is in the cache, it can be presented to the CPU after a very short time (IBM 370-168: Cache cycle time: 80 Nsec Cycle time: the time delay between two subsequent memory addressing operation). Only if the information is not in the cache, then the main memory is addressed and the required block of information - a page - is loaded into the cache. In the 370-168, the corresponding main-memory cycle is 320 Nsec. Since many data processing operations are sequential in nature, the program structure can be organized in such a manner that the cache will contain the required information in the majority of cases, so that the performance of the memory hierarchy is determined by the cache performance, the cost is determined by the main-memory cost. In order to obtain a balanced CPU/memory system the cache cycle should be equal to the CPU cycle, which with the help of monolithic memories is easily achievable.

The advent of inexpensive LSI logic circuits has increased the processing power of the CPU's. Example: IBM 370-168: announcement: 1972 logic technology: 3-4 circuits per chip, 2.3 MIPS, IBM 3033, announcement: 1977 logic technology: 40 circuits per chip, 4.8 MIPS. In addition to just increasing the processor power, the silicon integrated circuitry enabled a considerable CPU task differentiation. Separate processing units for Input/Output processing, the maintenance and service subsystem, or the storage subsystem became affordable, introducing a large amount of parallelism into the CPU operation and thus increasing its throughput. These subsystems carry their own compliment of read-write memory to store the specific segments of the control program, which is necessary to perform their assigned function. Only when their complement of control program is exhausted will they have to go back to the CPU main memory for new control information, and very little interference with the general bus traffic of a CPU will occur.

Some processing tasks may be completely delegated from the CPU to the Input/Output area. Such distributed processing reduces the communication traffic between the CPU and the Input/Output units. The central host will then only be addresses when access to the central data bank is required.

All these facts are elements of the same development: a sharply decreasing cost of data processing. This makes it possible to divert some of the CPU's processing power away from problem solving tasks to assisting the user in his interaction with the data processing system. These "Ease of use" features can be designed to facilitate hardware or software error diagnostics, to aid in application programming, to assist or instruct the user in the operation of the system, or to add security features, which protect the data and programs stored in a computer against the misuse by unauthorized persons.

The trend will be to add so much "ease of use" of the appropriate cost that the power of an efficient information system is available to every authorized person, at all times, at all places. To avail himself of these services, the user does not need to be a computer expert, nor does he have to go through an expensive, time consuming training period. The information system itself will interactively teach the user what he needs to know [26].

6. Summary

- The technological progress of information systems is largely determined by the advancement of the silicon integrated circuitry. The cost/performance progress of this technology shows no signs of saturation in the foreseeable future.

- The silicon integrated circuit technology will continue to dramatically advance the price/performance of the central processing complexes of an information system.

- The advances in Input/Output Units will be more gradual, since much of the enhancement potential of the mechanical technology is exhausted. Here also the LSI technology will be a major contributor to technological progress.

- These different rates of technological progress will cause a shift in cost emphasis away from the CPU to the Input/Output systems and to external storage subsystems.

- The increased CPU processing power can accomodate more powerful operating systems and application programs which will stimulate the use of information systems by a much wider range of non DP professional users.

References

1. R. Remshardt,
   U.G. Baitinger: IEEE J. Solid State Circuits, Vol. Sc 11 (1976), No. 3, Page 352-259

2. F.L. Thompson: Sol. St. Technol. 17, (1974), No. 7, Page 21-30

3. D.L. Spears,
   H.I. Smith: Sol. St. Technol. 15, (1972), No. 7 Page 21-26

4. L.M. Terman: Proc. IEEE 59, (1971), Page 1044-58

5. K.U. Stein,
   H. Friedrich: IEEE J. Sol. St. Circuits, Vol. Sc 8, No. 5 (1973), Page 319-23

6. S.K. Wiedmann: European Solid State Device Conference, Munich 1976

7. K. Horninger: Digital Memory and Storage , W.E. Proebster, ED., Vieweg 1978, Page 121

8. C.N. Ahlquist et al: IEEE J. Solid State Circuits, Vol. Sc-11, No. 5, (1976), Page 570-74

9. P.B. Ghaie, W.R. Gardner, D.L. Crosthwait: IEEE Trans. Reliability, Vol. R2, No.4, (1973), Page 186

10. P. Totia, R. Sopher: IBM J. Res. and Dev., (1969), Page 220

11. W.K. Liebmann: Digital Memory and Storage, W. Proebster, ED., Vieweg 1978, Page 135

12. H. Yoshimura et al: Digest IEEE ISSCC 1978, Page 148-9

13. R. Mitterer: Digital Memory and Storage, W. Proebster, ED., Vieweg 1978, Page 97

14. F.H. De Leeuw: IBID, Page 203

15. A.J. Perneski: IEEE Trans. Magn. Mag-5, 554 (1969)

16. O. Voegeli et al: AIP Conf. Proc. 24, 617 (1975)

17. M. Suzuki et al: Digest IEEE ISSCC 1978, Page 206-207

18. P. Wolf: Digital Memory and Storage, W. Proebster, ED., Vieweg 1978, Page 247

19. J.A. Kutsch, Jr.: Nat. Comp. Conf. Proc., Vol. 46 (1977) Page 357-62

20. C. Fields: IEEE Trans. Prof. Com. VPC-20, No. 1 (1977) Page 2-6

21. G. Chodil: Proc. S.I.D. Vol. 17/1 (1976) Page 14-22

22. B. Kazan: IBID., Page 23-29

23. L.A. Goodman: IBID., Page 30-38

24. E. Lennemann: Digital Memory and Storage, W. Proebster, ED., Vieweg 1978, Page 65

25. P. Wentzel: IBID., Page 33

26. W.K. Liebmann: Proc. 9. Workshop, Institut für Produktionstechnik und Automatisierung, Fraunhofer Gesellschaft, Boeblingen, Nov. 1977.

27. L.M. Terman: Scientific American, Vol. 237, No. 3 (1977), Page 163-177

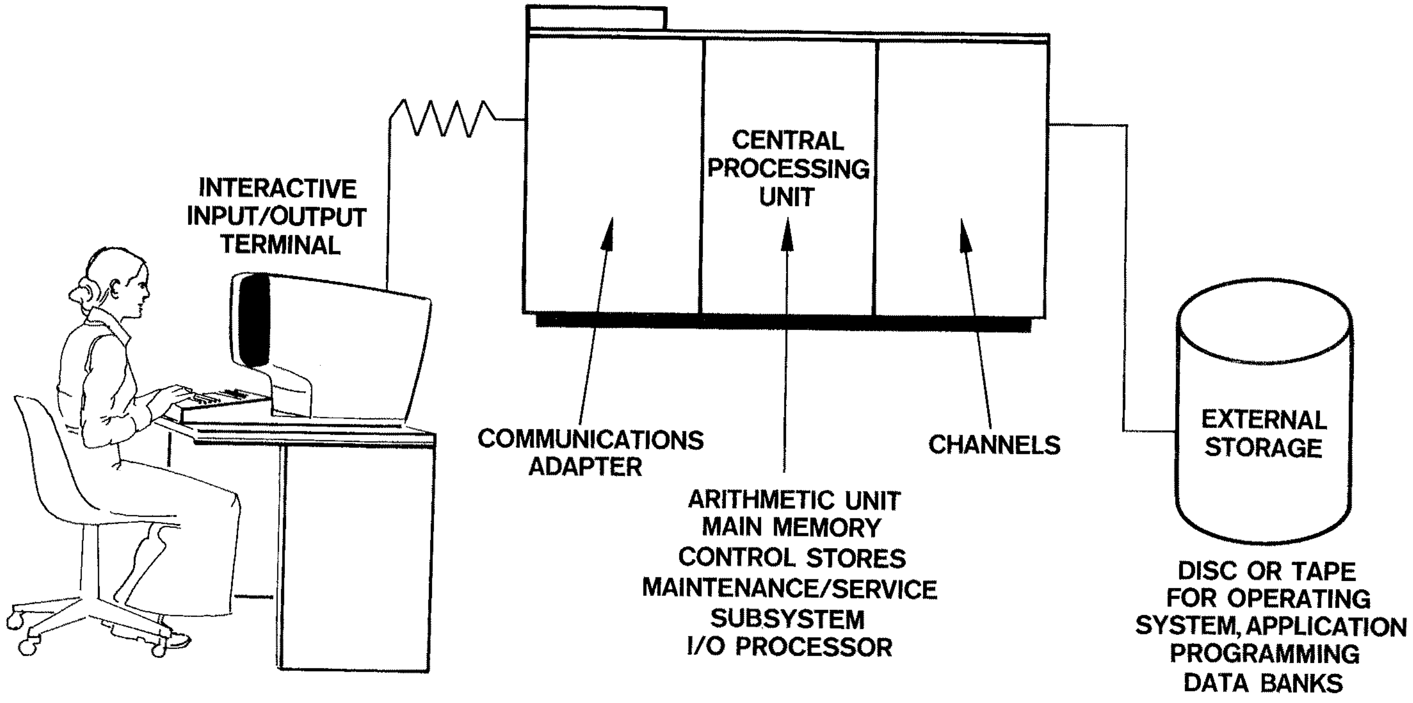


FIG. 1: HARDWARE COMPONENTS OF AN INFORMATION SYSTEM

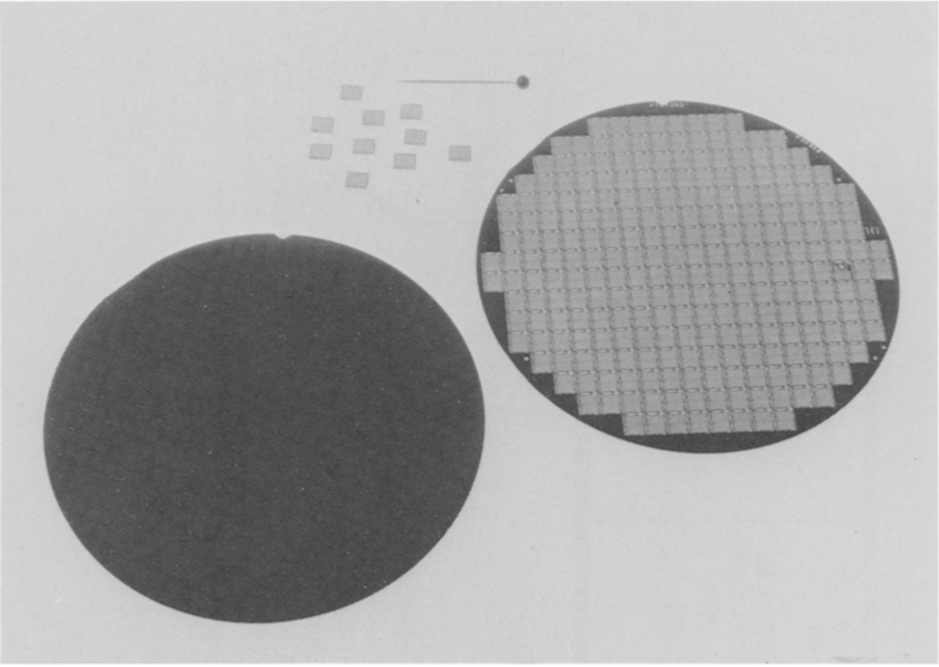


FIG. 2: SILICON WAFER

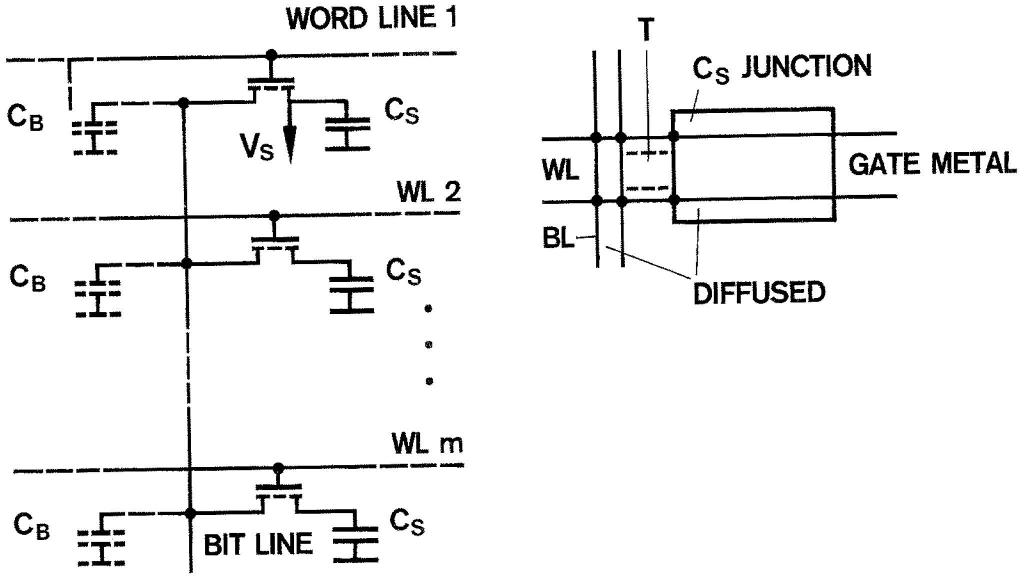
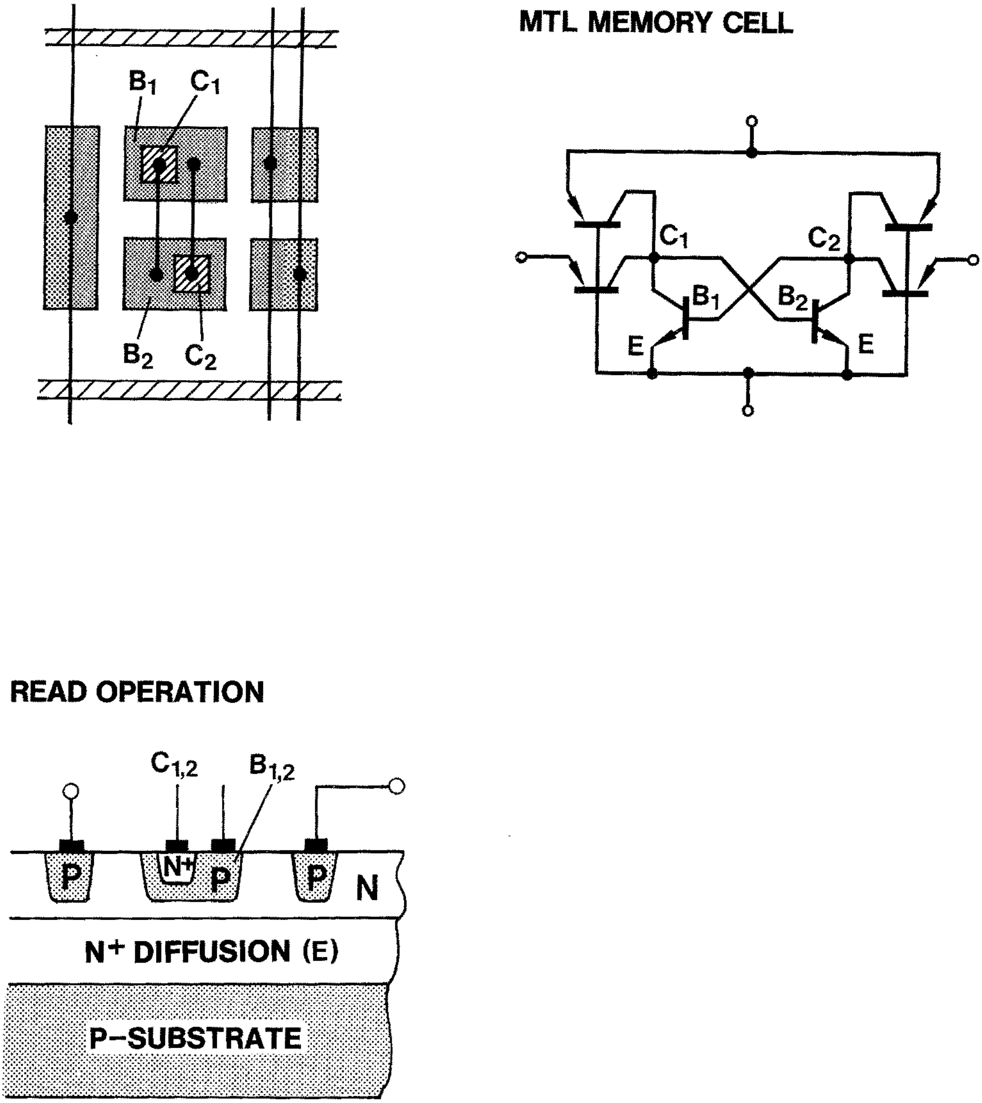


FIG. 3: ONE DEVICE MONOLITHIC MEMORY CELL, CIRCUIT AND LAYOUT EXAMPLE (STEIN A. FRIEDRICH [5.])

MTL MEMORY CELL



MTL MEMORY CELL

READ OPERATION

FIG. 4: MTL BIPOLAR MEMORY CELL (WIEDMANN 6.))

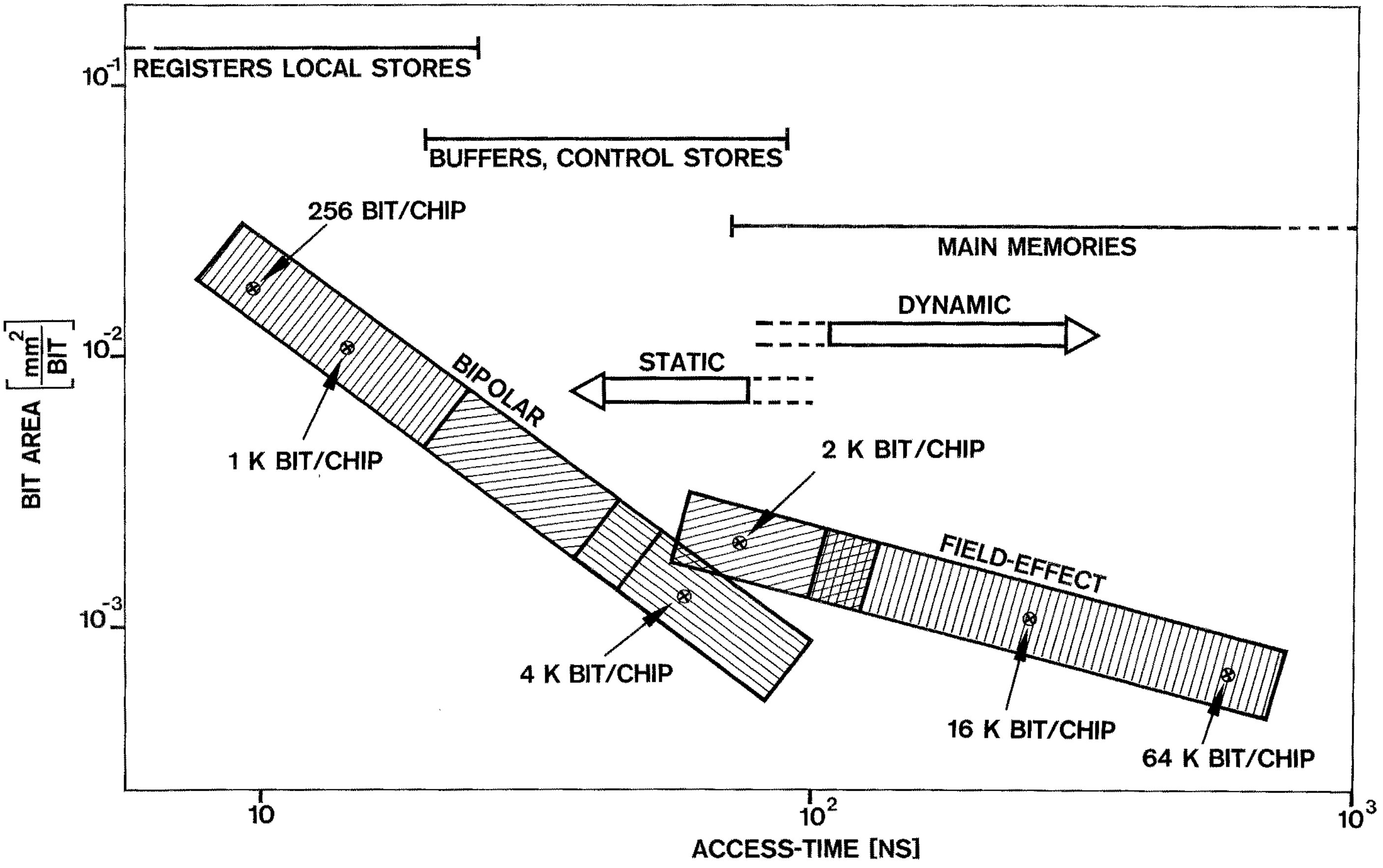


FIG. 5: BIT AREA REQUIREMENT OF MONOLITHIC MEMORIES AS A FUNCTION OF ACCESS TIME

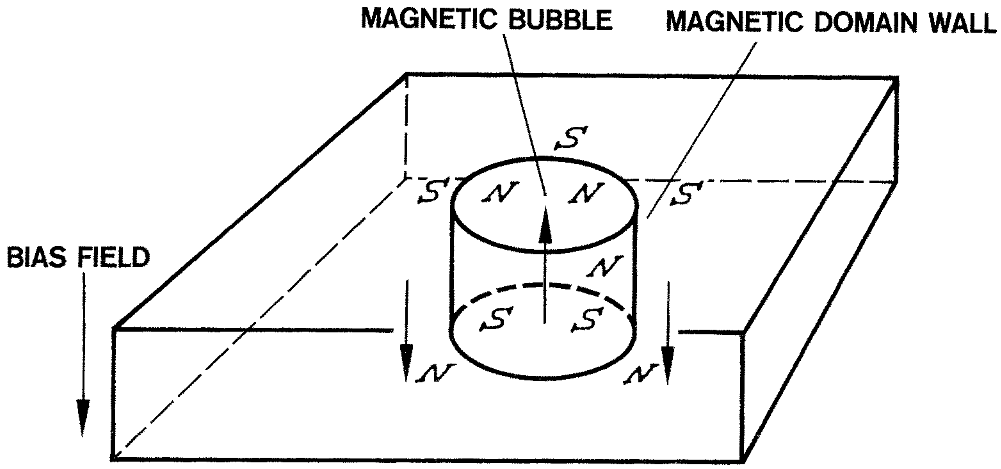


FIG. 6:   MAGNETIC BUBBLE (DE LEEUW 14.))

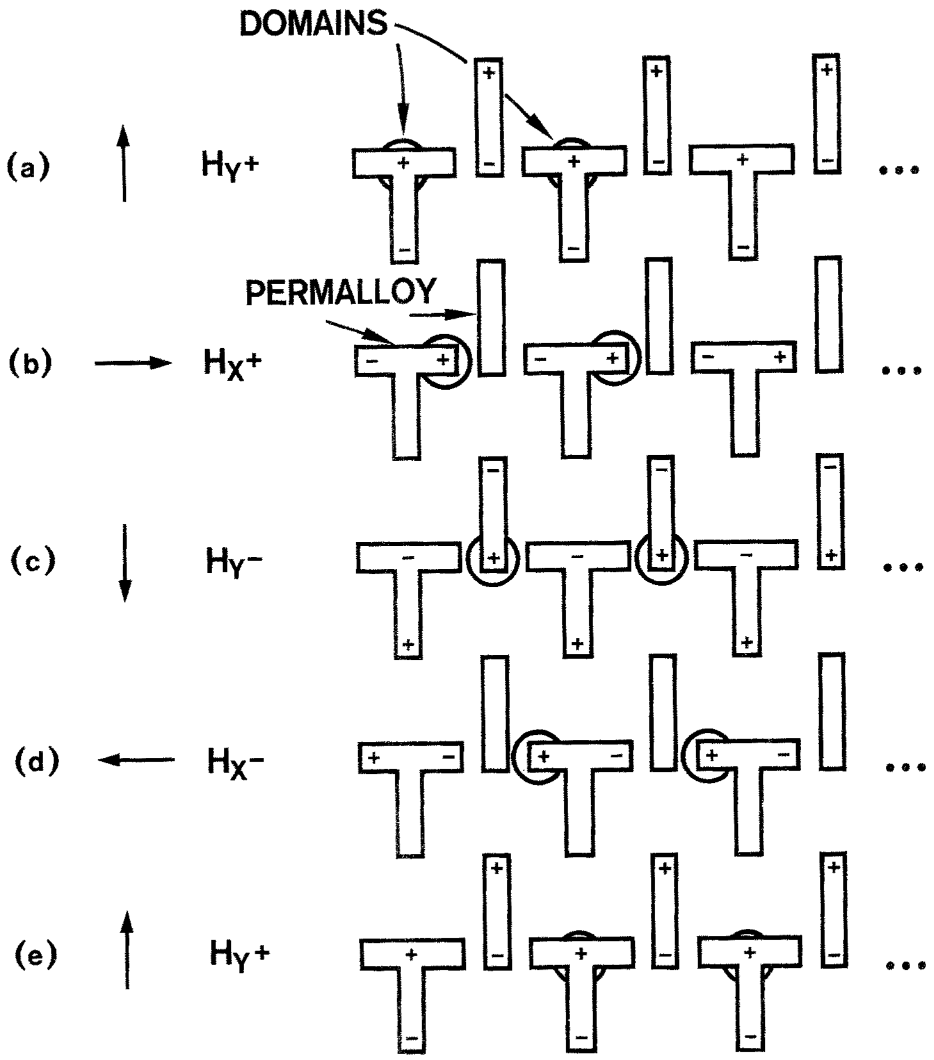


FIG. 7: ROTATING FIELD DRIVE OF MAGNETIC BUBBLES IN T-BARS: ROTATING FIELD: H_{Y+}, H_{X+}, H_{Y-}, H_{X-}. THE EXTERNAL FIELD IS LOWERED AT THE POSITIONS OF THE POSITIVE POLE (PERNESKI [15.])

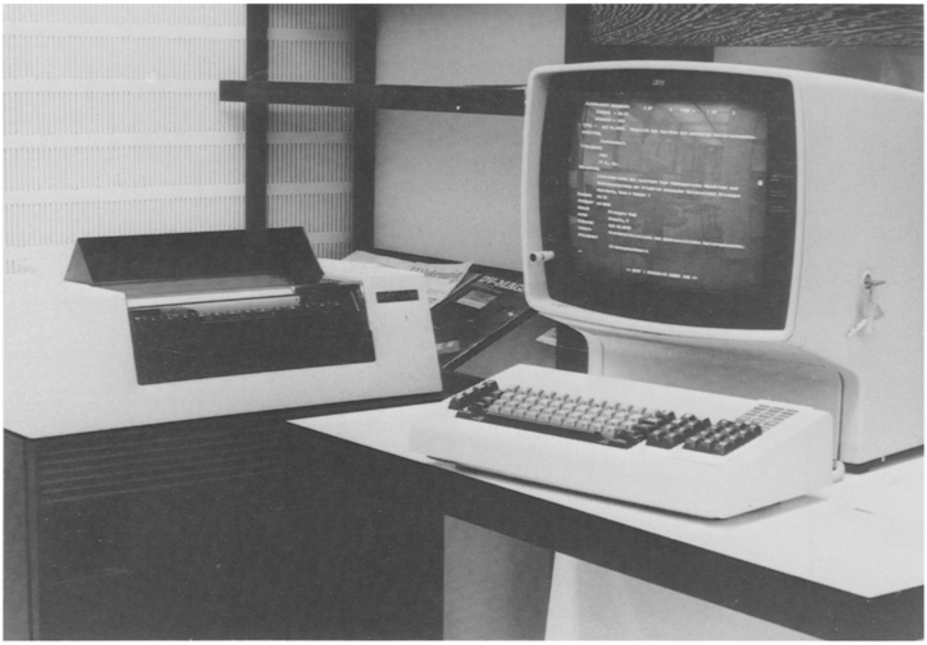


FIG. 9: INTERACTIVE INPUT/OUTPUT TERMINAL: IBM 3270

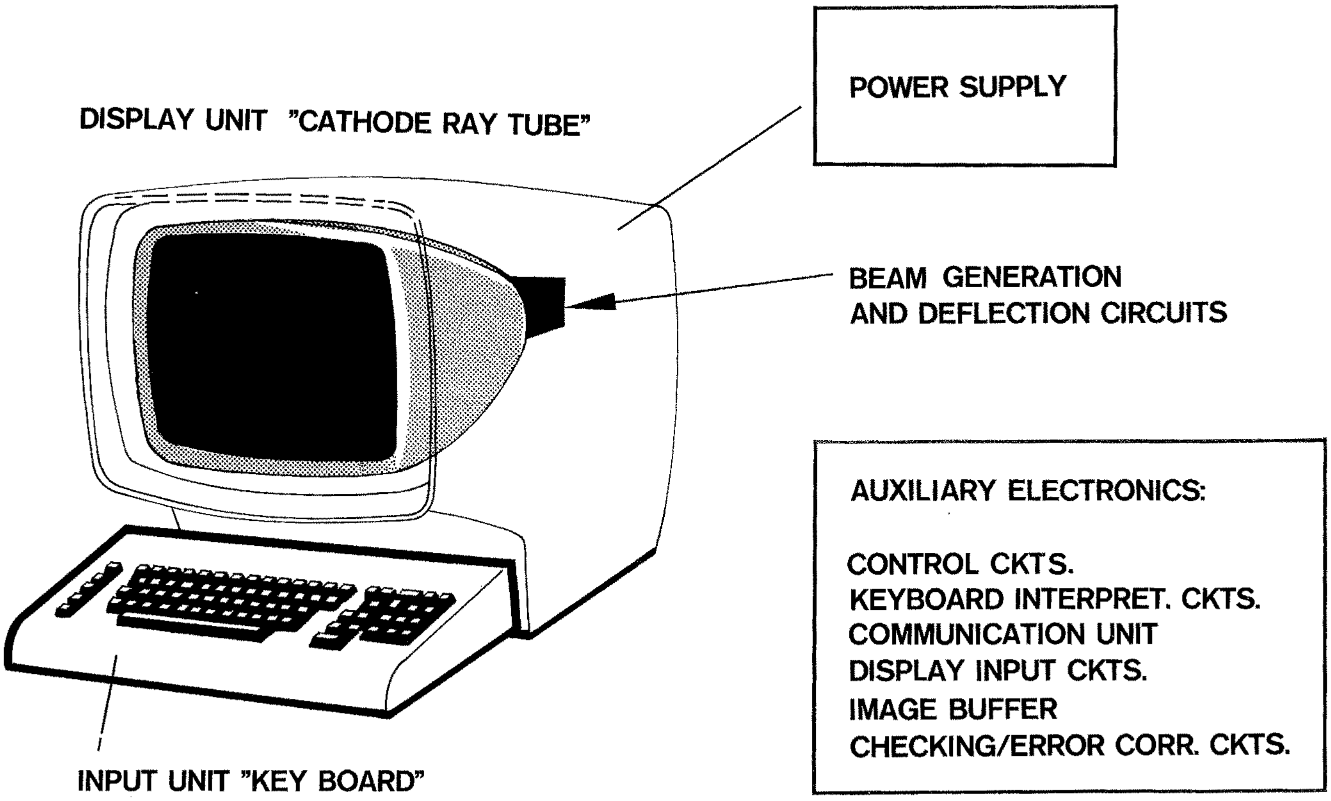


FIG. 10: MAJOR HARDWARE COMPONENTS OF AN INTERACTIVE DISPLAY INPUT/OUTPUT UNIT

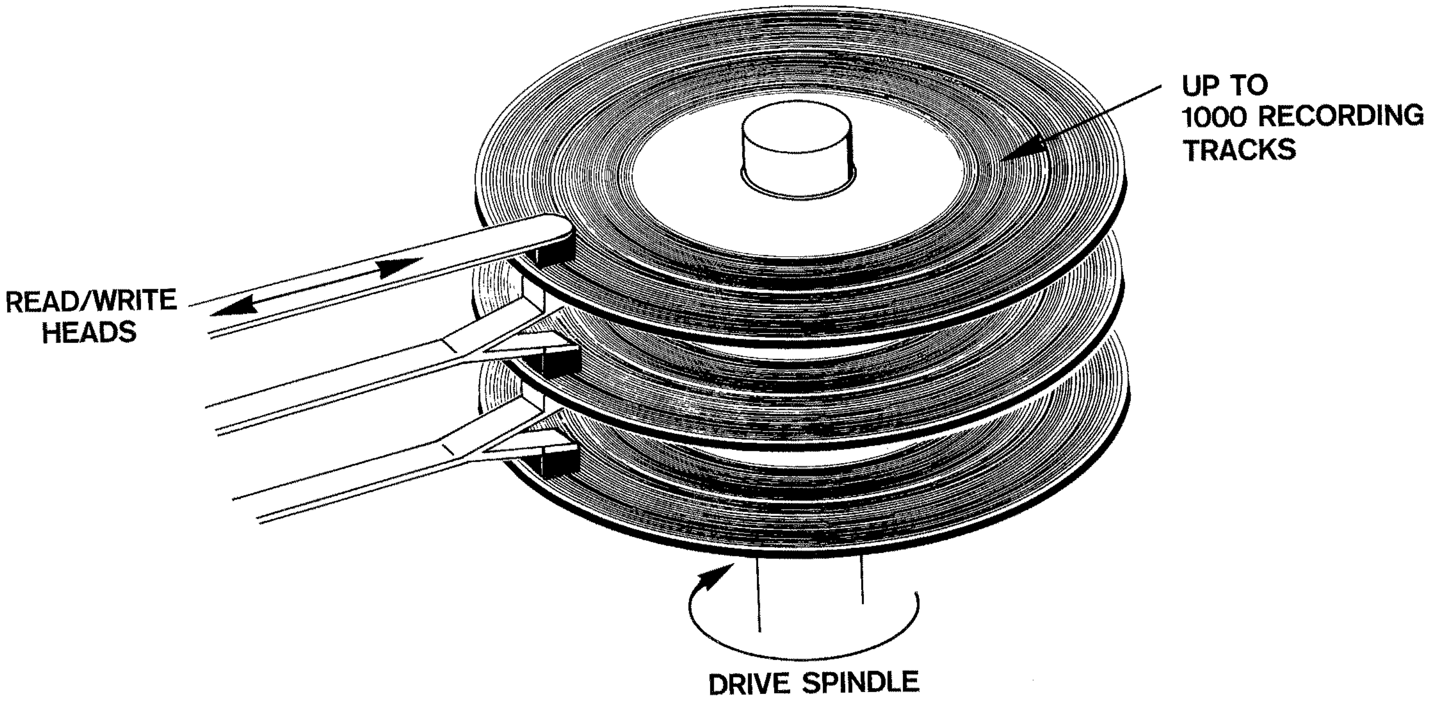


FIG. 11: MAJOR COMPONENTS OF A ROTATING
DISC STORAGE DEVICE

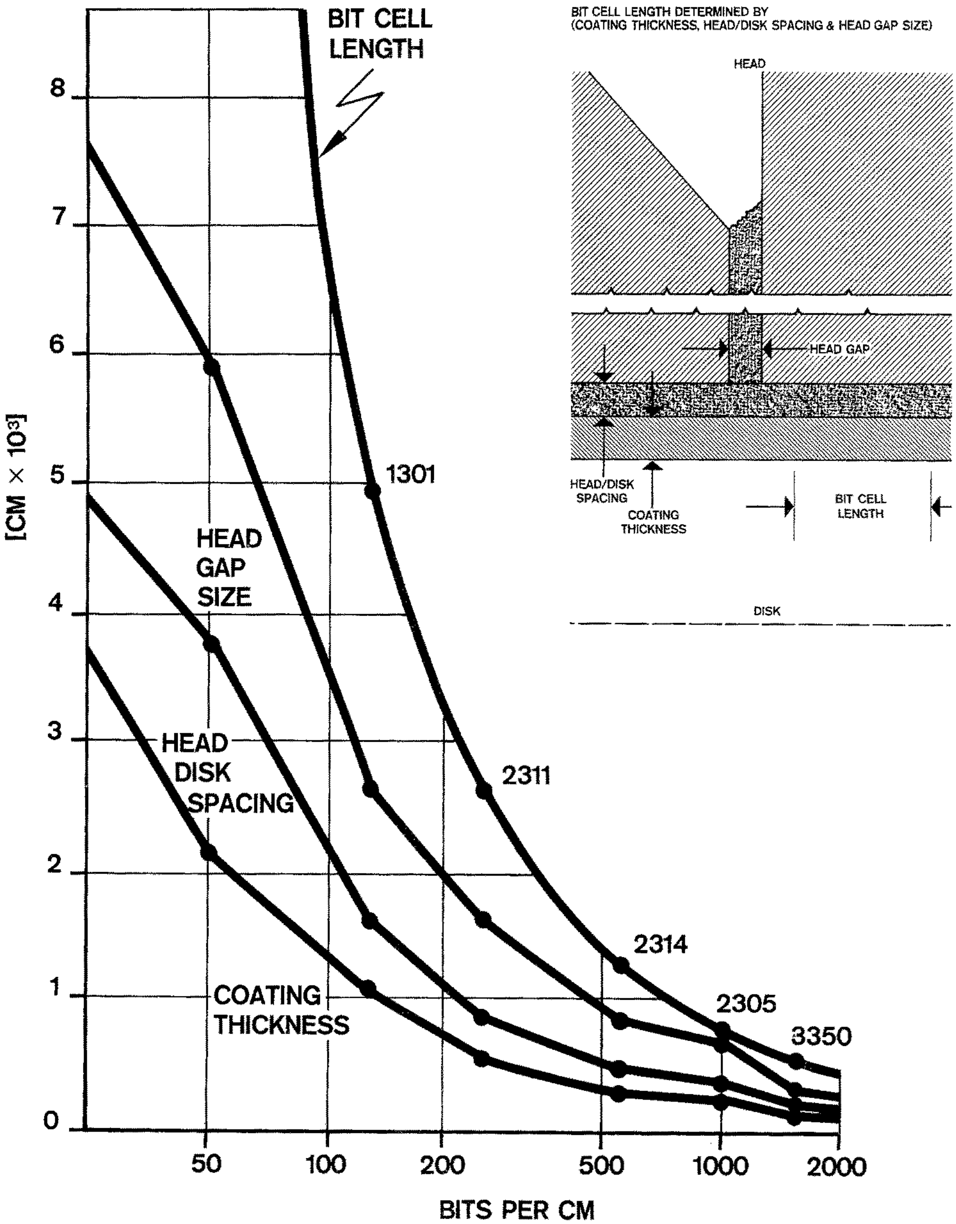


FIG. 12: "BIT CELL LENGTH" FOR VARIOUS DISC PRODUCTS