William E. Harding

Semiconductor Manufacturing in IBM, 1957 to the Present: A Perspective

Over the past twenty-five years we have witnessed the transition from germanium-based, individual transistors in their hermetically sealed enclosures to VLSI silicon devices interconnected in modular packages containing more than 50 000 logic circuits or as many as 500 000 bits of random-access memory. During this progression, manufacturing facilities producing these modern products have become more complex and technologically more sophisticated than those of any other industry. This review traces these fast-moving changes as they have occurred in IBM, emphasizing the continuous expansion of manufacturing skills and disciplines and how these, in turn, have contributed to the development of today's products and their respective manufacturing systems.

Introduction

Today the semiconductor industry is of worldwide scope, its products pervasive wherever electronics are applied. The remarkable progression, over the past quarter century, from circuits fabricated with discrete germanium transistors to Very-Large-Scale-Integrated (VLSI) silicon products with thousands of interconnected circuits, has been brought about by many contributors. This industry has always been one in which all of its participants continuously add to and share in its progress.

Of the many kinds of semiconductor products, digital applications have experienced the most dramatic improvements in cost, increased function, and overall reliability. Initially paced by their applications in computers, communications, and weapons systems, digital integrated circuits are now the basis for calculators, watches, industrial instrumentation, automation controls, automotive ignition systems, electronic typewriters, and even video games.

The wide acceptance of these products has been heavily motivated by significant reductions in cost; the cost per transistor in a VLSI chip is more than three orders of magnitude less than its original discrete equivalent. More important is the value of the functional product in reducing the cost of electronic packaging and improving reliability. These accelerating improvements are all the more remarkable since semiconductor structure and process have become more complex, and materials, capital, and labor costs have steadily increased.

This paper traces the expansion of semiconductor manufacturing as it occurred in IBM, highlighting its sometimes unique approaches to fulfilling the varied needs of the Corporation's wide range of products. Emphasis is placed on two major themes: first, the evolution of the technology; and second, the contributions of manufacturing (as distinct from product development), outlining its accomplishments, its growth in complexity, capacity, and influence on product design.

Describing the rapid advancements of semiconductors as an evolutionary process is overly simplistic. As well-established techniques are exploited, extended, and optimized, they also approach their basic limitations. These limitations, when coupled with the increased demands of succeeding products, stimulate the need for better methods. The continuous addition of innovative techniques has been the main contributor to maintaining the compound growth rate of semiconductor devices. Since old techniques are frequently used to advantage in the fabrication of new products, this superposition of old upon new, an evolutionary-revolutionary trend, appears to be a better way to explain this accelerated growth.

As these new semiconductor products are introduced, it is the task of manufacturing to assimilate the new technologies and to reduce them to economic practice. But as the manufacturing complex responds in scaling up and controlling new technologies, it assumes an additional responsibility which is equally demanding: New products invariably have characteristics requiring the development of new kinds of test, control, and logistical systems. The execution of this twofold mission has led to a continuous addition of new disciplines and acquisition of new skills by the personnel who create, master, and operate the manufacturing complex.

The paper is organized chronologically, with the flow by product rather than by technology. Following this introduction, the section *The germanium legacy* begins with a description of IBM's fully automatic system for manufacturing germanium alloy transistors. Designed, developed, and implemented by manufacturing, it was the first of its kind in the industry. The next product was the first germanium transistor in the industry to form both junctions with solid state diffusion. It introduced the batch fabrication of transistors in wafer form. This new direction, which all manufacturers were taking, was probably the most significant contributor to the economic advance of semiconductor manufacturing. The limitations of early semiconductor packages are then discussed, illustrating a growing influence of manufacturing on product design.

The next section, *Silicon hybrid integrated circuits*, details the manufacture of Solid Logic Technology (SLT), the first semiconductors uniquely designed for IBM's own computers. Silicon replaced germanium, and a new package interconnected planar semiconductors and passive components to form integrated circuits. Key to these advances was IBM's pioneering of glass passivation to hermetically seal transistors during wafer fabrication. The *Monolithic integrated circuits* section outlines the adaptation and extension of SLT for monolithics. It covers the introduction of the industry's first monolithic memory chips, which led to another major expansion in IBM semiconductor manufacturing to quickly replace a matured base of large-scale magnetic-core manufacturing. A section on *LSI-FET* highlights the development of new manufacturing systems designed to make custom LSI. Silicon planar technology was now so mature that it could be quickly adapted to the integration of Field-Effect Transistors (FETs) at even greater densities than the competing bipolar circuits.

The final section, *The road to VLSI*, brings us to the state of the art as the semiconductor industry faces its next frontier, VLSI. The first 64K-bit RAM memory chips (K = 1024) demonstrate the increased complexity of semiconductor structure and process. To illustrate the reduction of semiconductor development into specific but highly sophisticated disciplines, two aspects of technology are briefly reviewed—lithography and yield management. Finally, in contrast to the isolation of separate technologies, the paper closes with the description of a totally integrated manufacturing system which uniquely combines its own process, tools, computerized measurement, information, and logic systems.

The germanium legacy

By 1957 the decision to design all future IBM circuits with solid state devices had been made. Compared to others in the industry, IBM's production of germanium transistors was modest. The product scope was limited to five switching transistors, two complementary npn and pnp (utilizing alloyed junctions), two more advanced structures (complementary, npn, pnp utilizing alloyed emitter, diffused base), and a medium-power alloyed-diffused transistor. The majority of semiconductor devices in IBM products were procured from the industry, but many were manufactured to our design and process specifications. The alloyed-diffused structures were undergoing rapid improvements that would culminate in the development of the mesa transistor. On the other hand, the less expensive alloy devices were finding large applications where their low performance was adequate; consequently, manufacturing moved to completely mechanize their fabrication.

Completed in 1959, this completely mechanized system [1] assembled single-crystal germanium disks, metal-alloy cylinders (for emitter and collector), a pre-tinned concentric-base electrode, and two contact wires (for emitter and collector) in a multi-piece carbon jig. Each jig, combining just one set of parts, was passed through continuous furnaces to form the junctions and fuse the electrodes. The jigs were disassembled automatically and the extracted transistors were automatically welded to a three-lead stem, chemically etched, cleaned, enclosed with a hermetically sealed cap, and tested. The entire sequence took only three hours. The system could be operated, at variable throughput rates, to a maximum of 3600 units per hour. The fact that the entire procedure can be described in one paragraph testifies to the simplicity of these early transistor processes.

This automatic system was in sharp contrast to the conventional manual assembly line which depended upon the dexterity of experienced operators often working with microscopes. Only three people were now required to monitor the status of the entire system and to direct corrective action as required. Once the system was

installed and debugged, its learning curve was steep. Within three months its yield surpassed that of the batch, manual assembly line and approached the ultimate. In one ninety-day period, operating continuously except for adjustment and maintenance, it yielded an average of 2600 good transistors per hour at final test. The success of this system was not one of economics alone. It convinced us that what was originally perceived as very artful technology could be scientifically understood and controlled. We now saw how essential it was to specify and control each of the fabrication sequences which critically determined the transistor construction.

Given a sound design, the product yield is dependent upon the overall integrity of the manufacturing process. Transistor processes gave almost no allowance for repair or rework, and maintaining process control required dependable machinery as well as techniques which continuously monitored every stage of the fabrication process. A big advantage of the alloy-transistor manufacturing system was the short time it took from start to final test (only three hours). Thus, in-process data could be quickly combined with final test results to diagnose process faults and to take corrective action. Projecting ahead, future products would have increasing numbers of fabrication sequences. Consequently, we would have to find other ways to partition, measure, control, and feed back in each element of the overall process; this would be the guiding principle in yield management.

IBM released what would be the last of its germanium transistors to manufacturing in 1959. This double-diffused device was most important because its technology was a precursor for much of what would become standard practice in succeeding silicon products. Unlike the pnp, alloyed-emitter, diffused-base mesa transistors, this npn device had both the emitter and collector junctions formed by high-temperature diffusions. Even in this early application of the control of donor and acceptor distributions by diffusion it was almost an order of magnitude better than alloying techniques. The unmasked base diffusion was over the entire wafer. But the emitter diffusion was defined by selective masking. The technique (conceived and perfected by manufacturing) consisted of evaporating thick spot-like deposits of sodium chloride through a stencil mask and then vacuum-depositing a blanket film of silicon monoxide. When the salt was subsequently dissolved, the silicon monoxide lid separated, leaving a pattern of circular spots of exposed germanium surrounded by the silicon monoxide which prevented penetration of the diffusing species. This selective masking procedure, crude by today's standards, may have been the first application of what today is commonly called "lift-off" lithography. The procedure made planar emitters. Emitter and collector metal contacts were vacuum-codeposited through masks and were thermally alloyed. Collector junctions were still isolated by chemical machining. In comparison to the typical shape of mesa transistors, the resulting structure was trench-like. We had yet to progress to the all-planar structure.

Transistors were now batch fabricated in two modes: hundreds of devices in one wafer, and many wafers processed simultaneously in separate reactors. The processes of vacuum deposition for metals and insulators, diffusion for defining the localization of doping species, thermal alloying for ohmic contacts, and chemical machining to shape or selectively remove materials, had now become the new techniques for exploitation. Mechanization, so useful for the alloy transistor, would no longer play a significant role in wafer fabrication until its return in the late 1970s. But mechanization was still required to interconnect the transistor to a package and to test it. Semiconductor manufacturing had emerged into two distinct phases: wafer fabrication, and Bond, Assembly, and Test (BAT).

The overall program was successful in bridging manufacturing from alloy to diffused transistors, but it was not without its difficulties. The main problem was that the transistor was becoming too small, and consequently it taxed mechanization to its limits. For example, the automatic system was marginally able to locate and bond wires to electrodes 50 μm wide separated by only 12 μm [2]. The cost of hermetic stems was threatening to exceed that of the transistor itself because of the tight tolerances required for mechanization and the gold plating required to protect the stem metals from corrosive chemical junction stabilization processes, even though gold was then only $35.00/oz. In the wafer processes, the dimensional control of deposited materials by evaporation through stencil masks was also approaching its limits.

As a result of this experience, manufacturing began to pressure development for structures and processes which were "manufacturable." Prior to this time, manufacturing people had little influence on the product design: they accepted the device design and tooled the factory for mass production. But now they wanted participation in the earliest stages of product definition. Encouraged by the success of evaporation and diffusion, they pushed hard to extend these methods, and to find others as well, so that the entire device could be built in the wafer. They also wanted a package whose shape and dimensions would complement automation, eliminate hermetic seals, and minimize the need for chemical processing in the post-wafer stages of assembly. This experience motivated manufacturing to build up its skills by adding the special

disciplines necessary to understand, design, and control these processes. It was now competing with research and development for the same kinds of people.

Silicon hybrid integrated circuits

By the late fifties, markets for computers and peripheral machines expanded, due in part to the success of solid state circuits. It was time for a new componentry which could meet the varied demands of all machines. Toward this goal the development groups switched the bulk of their efforts from germanium to silicon, which promised greater reliability, lower cost, and higher density. Management recognized this increasing dependence on solid state electronic components, and consolidated the related engineering and manufacturing functions. By this time the cooperative efforts of semiconductor device [3], circuit, and packaging development groups were coming to fruition with an overall concept for packaged electronics, tailored for computer systems, called Solid Logic Technology (SLT) [4]. Planar silicon semiconductor devices were combined with passive components on a ceramic substrate to form SLT modules, the first level of the SLT packaging hierarchy [5]. Each module contained a unit building block such as an And-Or-Invert circuit.

The cornerstone of the SLT semiconductor process was optical, chemical lithography. Using optically generated patterns, photoresist was used to define the locations where thin-film masking was to be selectively removed by chemical etching. This method, repeatedly shaping silicon dioxide for diffusion masks or insulation and metals for conductors, introduced a new degree of dimensional and geometrical control [6].

The most significant innovation in SLT was to combine glass passivation [7] with a unique terminal structure [8] to literally build the equivalent of the old three-leaded, glass-metal, hermetic stem into each chip during the final steps of the wafer process. In many ways this was the key to SLT cost and reliability. The resulting chip construction immediately facilitated mechanization of electrical test and bonding to substrates. The package, now unencumbered with the seal function, could be optimized for interconnecting circuits in a near-ideal form factor for interconnecting to printed-circuit cards.

But developing the technology was not easy. The problem was to find a corrosive-resistant, stable glass to match and bond to both silicon dioxide and aluminum. The search was often Edisonian. Having found a chemically resistant glass, we next needed a method of applying the glass which could precisely control thickness and be pinhole-free. It was difficult to chemically machine the vias in the glass in order to contact the aluminum. Finding a metal system which could make contact to the aluminum, seal the glass, and be solderable was an equally challenging problem.

Fortunately, when the glassing process was finally developed [9], it was easily adapted to mass production; but the terminal processes were much more difficult to tool. Preparing the glass and aluminum surfaces required sputter cleaning in a plasma. Within the same reactor, the chromium, copper, and gold had to be evaporated in precise phases to exact thicknesses. Implementing these multiprocess reactors was to become a way of life for semiconductor engineers as future semiconductors kept increasing their use of complex structures, materials, and processes.

To package the semiconductor chips into integrated circuits, paste-like materials for resistors were printed by silk-screen lithography and were fired on precision-made ceramic substrates [10]. With the requirement for hermeticity gone, multiple pins could be economically inserted and staked into the substrate. In one simple "wave-soldering" step the conductivity of screened interconnections was enhanced, the pins were connected, and the lands, where chips would join, were pre-tinned. One of the beauties of this method was the untinnable cermet resistors. Abrasive trimming, under computer control, made precision resistors. Pretested chips were then solder-reflowed to the substrate.

In this process of soldering chips to substrate, the terminal design and glass passivation of the chip were as essential to withstanding the exposure of semiconductors to corrosive flux at high temperatures as they were for hermetically sealing the devices for environmental protection in the ultimate application. Modules were covered by a metal cap, plastically sealed, and tested. Semiconductor products were now manufactured in three phases: semiconductors, substrates, and BAT.

An alternative to these hybrid integrated circuits was monolithic circuits. To be sure, there were opinions within IBM which favored this alternative. To make the decision even more confusing, there were even some who would have persisted on the germanium track. But in late 1961, monolithic designs gave little promise that they could be built economically in the 1962 to 1968 time frame. The really decisive drawback was their limited switching speeds. These early monolithics operated at hundreds of nanoseconds while 70 percent of our applications required switching speeds of less than 30 nanoseconds. The SLT team was further convinced that its technology, at the right time, would readily extend to monolithics.

The program objectives set at the end of 1961 were challenging across the board: module costs were to match those of a discrete transistor; reliability was to be improved by two orders of magnitude; families of modules would cover all performance ranges for a new set of computers to be called System/360; and production volumes, starting at 50 000 for the engineering requirements in 1962, were to be increased by approximately ten times in each of the first four years of production.

These objectives were extremely challenging since none of the existing tools or test systems for germanium products were applicable, the substrate business was new, and the semiconductor tooling industry which exists today was then in its infancy. To obtain the focus we needed, we departed from our traditional organizational structure; we combined the development, product, and manufacturing engineering areas into one dedicated organization. This force was further enhanced with a team from Europe whose first job was to participate in defining the manufacturing system and then to implement a second factory in France. The year 1962 was spent in developing the techniques for each of the production tools and in designing the products. By mid-1963, four circuit families (typically having ten integrated circuits each), spanning performance from 7 to 700 nanoseconds, were in pilot production.

The introduction of multiple products, each with many part numbers, brought a new dimension of complexity to semiconductor production. Added to the old measurements of quality and quantity was a third element called "mix." To meet "mix" required shipments to match the exact distribution of parts in every customer order. New management methods were needed in this changed environment. Previous semiconductor lines characteristically made one product, with relatively fewer processing steps; and tools operating in fixed set-point mode were, by comparison, more predictable and easier to supply. The logistics for SLT modules were significantly more involved with respect to both materials and information moving throughout the system. Subassemblies varied—crystals by type and conductivity, n or p epitaxial wafers, diode or transistor wafers. Each part required a unique set of masks. Routings all varied—fast saturating transistors used gold diffusion for minority carrier lifetime control while other types would skip this step, diodes and transistors each used different passivating glasses which called for different apparatus. Substrate resistors varied in numbers and locations. Specification was required in order to describe the electrical and mechanical sequences of each job and their associated testing.

Superimposed on the normal product flow were all sorts of special jobs, such as priority lots to evaluate engineering changes in product or process, and express lots to make up for yield variations or to correct for changing orders from customers. Information directing the operations flowed to the manufacturing floor, while data on yields, maintenance, work in process, and quality control tests flowed back.

The response to these requirements was the development of a series of computerized information systems for production control and yield management. To facilitate this, the production lines were partitioned into work sections and control gates. The gates were used to regroup batches into production lots classified by the date they entered the system. This discipline forced production to move forward in unison. The result was to minimize the dispersion of production lots so that the results of production could be correlated with the original process and quality control data. These controls were particularly essential when production was building up. During this time, yields, while continuously rising on the average, would vary widely in the short term. Learning to manage these problems proved to be invaluable for handling the highly personalized products of the future.

As production expanded, experience reemphasized the need to maintain a basic effort (not involved with immediate production problems) devoted to understanding the physics and chemistry of our products and process. One such experience was called "the 200-degree disease." Previously, we had correlated an accelerated test (24 hours, 200°C, and bias stress) with long-term life testing at use conditions. We used the test to sample the quality of each day's production. And initially nothing failed. But, as the rate of production grew, we observed an alarming and increasing percentage of product failing. Some good detective work quickly gave the answer. During the reworking of bad patterns at the aluminization step, the phospho-silicate glass (deposited into silicon oxide during emitter diffusion) was inadvertently removed [11]. Without it, devices were demonstrably unstable. The immediate solution was to take the yield loss and not to rework. But the fundamental work which followed to understand and control this reliability mechanism would have great significance in the development of reliable field-effect transistors [12–16].

In 1964, production of the fifth circuit family, called Advanced Solid Logic Technology (ASLT), was introduced. It pressed all aspects of the technology to achieve greater speed and density. Diffusion and lithography improvements reduced the vertical and horizontal dimensions of the transistors. To increase the density of cir-

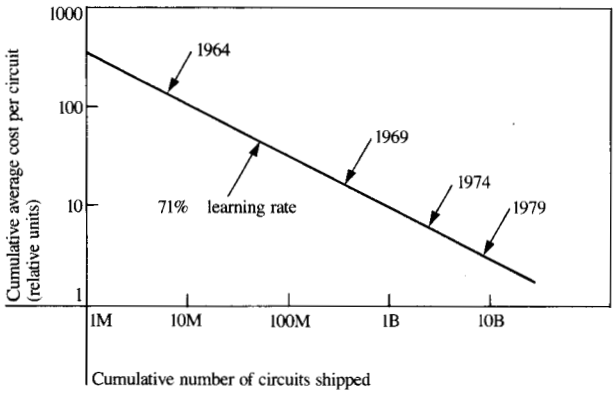


Figure 1 Learning curve for bipolar circuits: relative cumulative average cost per circuit, corrected for inflation, for all bipolar logic and array modules produced at one IBM facility.

cuits, the bottom sides of the substrates were used for added resistors and wiring. Combining multiple transistors per chip and chip capacitors, substrates with increased numbers of pins were piggybacked to make stacked modules each having three and four current-switch circuits. With increased density and with the change from voltage mode to current switching circuits [17], loaded delay was improved to three nanoseconds. Nearing the end of 1964, total production of 50 different module types was approaching one million per month. All of the high-production subsystems had replaced their prototypes and were in their final stages of optimization. Almost every tool in the program was important, but space does not permit reviewing all of these. To give a flavor, we will discuss just one.

Testing 20 000 000 chips per day (required in the years of peak production) necessitated automation of chip testing. Every diode or transistor for a particular module had to be individually predetermined to meet its specification with assurance greater than 99%, so that the fraction of assemblies requiring rework would be acceptably small. These test systems accepted 10 000 chip batches which were poured like sand into vibrating feed bowls, where the unique terminal geometry was used to orient them right-side-up (and correctly polarized) to match the test probes which contacted the chips. Up to 60 separate test formats measured ac and dc parameters and finally sorted them. A combination of two test stations, managed by one computer-controlled test system, could sort for any set of test sequences desired at the rate of 36 000 devices per hour [18].

By 1967, SLT had reached its ultimate efficiency in yield and productivity. Over the preceding three years, the learning curve had been 71%. (This means that the cumulative unit cost of all units produced was reduced by 29% each time the cumulative production was doubled.) By April of 1967, the 250-millionth module was made, and in that month the combined monthly production of modules by three IBM facilities peaked at 50 000 000. The portion of the learning curve of Fig. 1 extending beyond SLT and including all the new bipolar products shows that this cost improvement rate has continued to the present time. To put these achievements into perspective, consider data from the Electronic Industries Association electronic market data book: During the years 1964 through 1966, the entire industry sales for all types of monolithic- and hybrid-integrated circuits was 77 000 000 circuits compared to IBM production of 130 000 000.

The reliability of SLT products surpassed their original projections. Based on more than a billion module hours in the field, the modules produced in 1964 had failure rates of 0.003% per 1000 hours of operation [19]. Two years later, it was three times better, largely the result of automatic manufacturing. This is comparable to a machine with 3000 circuits, such as the CPU of a System/360 Model 30, having only one module fail every five years.

The SLT program had achieved its goals but there was room for vast improvements. SLT chip dimensions had been made purposely large (0.76 mm^2) to allow their terminal spacing to match the capability of chip handling, testing, and screened-thick-film technology. These chips could be made and tested for a few pennies; consequently, the tradeoff of silicon area for low-cost modules made good overall economics. Since the active area of the transistors occupied less than ten percent of the chip, the next challenge was to use the silicon more efficiently by interconnecting more devices per unit area of silicon. This direction promised reduced cost of circuits per module, better performance with increased density, and increased cost effectiveness for the rest of the packaging hierarchy (cards and boards). With fewer interconnecting terminals per circuit, improved reliability was expected.

Monolithic integrated circuits

The invention of the first monolithic integrated circuits is generally attributed to J. Kilby [20] at Texas Instruments and R. Noyce [21] at Fairchild. By 1964, many manufacturers were marketing monolithics. These early products were expensive and relatively slow in performance, but they were clearly the wave of the future.

The first monolithic integrated circuit came to manufacturing in IBM in 1966 [22]. Besides its desirability as a product, this 16-bit random-access memory chip, called SP-95, was an excellent learning vehicle. Its orderly

structure could be densely packed without stressing our lithography. It extended the SLT processes of diffusion and lithography to construct resistors and to effect p-n junction isolation.

This first monolithic memory product pioneered two process innovations. First, the frit seal glass encapsulation was replaced with silicon dioxide deposited by radio frequency sputtering from a quartz target [23]. Sputtering allowed the deposition of high-melting-point materials at low temperatures. Characteristically, the type of glass whose coefficient of thermal expansion matches that of silicon has a melting point greater than that of the aluminum-based alloys used for metal interconnection. The quartz seal made the system mechanically more reliable and had the further advantages of fewer defects and improved thickness control.

Secondly, to meet the needs for more terminals at closer spacings on the substrate, a new procedure was introduced called Controlled Collapse Chip Connections (C-4) [24]. Instead of using copper balls to confine the mating chip and substrate solders, this method substituted a screened and fired glass dam on the substrate which localized a solder volume of sufficient height to eliminate thermal fatigue [8]. Surface tension forces were used to bring chips into exact alignment with substrate pads during solder reflow. In 1981, this process was still in use with products having as many as 289 terminals per chip. In all other respects the substrate-module fabrication was the same as it had been for SLT modules.

During the mid-sixties, we had returned to our traditional organization, which separated the development and manufacturing functions. While development was defining new monolithics for logic and memory, manufacturing, now certain that silicon planar technology was here to stay, began to extend that technology on their own. At the time, three plants were operating and a fourth was being built in Germany. Besides having the general complement of skills necessary for semiconductor manufacturing, each plant also had a special mission for supplying the technology of a particular section of the process hierarchy: One plant would be responsible for diffusion, insulators, and metallurgy; another specialized in chemical vapor deposition, photolithography, or mask technology, etc.

With this deployment of resources, a host of new production subsystems were developed: a computer-controlled, automatic system for crystal growing; a mask-to-wafer alignment system which included mechanized wafer handling; a unique method for flattening wafers [25] to improve the optics; mechanized systems for applying, drying, and developing photoresist; and a new test system for 20-terminal chips.

New logic and memory products moved into manufacturing. While they were developed in one location, remote manufacturing plants were responsible for their production. The logic products, called Monolithic Systems Technology (MST) [26], had two performance categories, six and ten nanoseconds. To minimize logistics, the master slice [27, 28] approach was taken, wherein each product set is made in a common design until the wafers reach the metallization step, at which point each particular circuit has its unique interconnection pattern engraved. We capitalized on our multichip modules to average six circuits per module.

Much more important than MST was the memory program. After the success of a 64-bit chip, the decision was made to build all future memories with semiconductors instead of magnetic cores or thin films. This brought another major buildup in production capacity. Manufacturing started with a 128-bit chip [29, 30]. Four chips in a stacked module made a 512-bit module. Even at the start, it was clear that these memories would require new models almost yearly. As soon as yield could be improved, the optimum level of integration would be raised.

Large-Scale Integration—Field-Effect Transistors

At the start of the seventies, manufacturing added Insulated-Gate Field-Effect Transistor (FET) technology on top of an expanding set of bipolar-based products. Although FET circuits could not operate as fast as bipolars, they could be made with significantly fewer steps at higher circuit densities and at lower costs. The fabrication processes for bipolar and FET were very similar: both used diffused junctions and oxide-insulated metal interconnections, but the basic principles of their operation were entirely different. In the FET, the oxide in the gate structure had to withstand high electric fields. Yield and stable operation, therefore, depended upon close control of the properties of the silicon-oxide interface. Inadvertent introduction of mobile ions, such as sodium, resulted in faulty devices. Consequently, the success of FETs depended heavily on a new degree of process "cleanliness." Both logic and memory products used n-channel devices [31] for better performance but they were highly susceptible to field-induced motion of unwanted contaminants. Over the past decade, basic research throughout the industry and at IBM [32] had developed a sound understanding of the causes and cures for various kinds of FET instabilities. Nevertheless, it was an art to implement, control, and maintain these processes.

Despite the increased sensitivity of FETs to thermal and radiative processes (like radio frequency sputtering the passivation), the FET products were able to utilize the MST passivation and terminal system. Key in this adaptation was the development of new processes for thermal annealing in hydrogen atmospheres [33, 34]. To fabricate these new products, a program with three teams was organized, each at a different site. One team concentrated on perfecting the technology, using a memory product for a test vehicle. They coordinated with the two product program teams, one for main memory and the other for logic, for the extremely cost-sensitive low-performance applications [35].

The logic program, called Emerald, had two additional challenges: one involved the semiconductor fabrication, and the other the first-level package. Emerald was IBM's pioneering effort in making custom semiconductor products tailored to the design and functional requirements specified by each customer. Previous logic products, like MST, consisted of a fixed number of different parts, designed by semiconductor circuit engineers. Each machine designer selected from the set those he required to implement his system. In this new approach, using an automatic design system, each machine group designed a unique set of parts which were optimized for its circuit requirements. Thus, Emerald was referred to as an "open part number set."

Each order of an Emerald product was accompanied with a digital description of its personality and test specification. This information could be delivered in the form of a computer tape or directly over wideband telephone lines. Thus, the blueprint was obsolete. Emerald did not use a master slice; instead, it personalized at all levels. Upon receipt of a Release Interface Tape (RIT) at the factory, computers would verify and convert this information into a set of instructions which computer-controlled machines used to make the mask sets [36].

Manufacturing now had an added measure of their performance: response time to make initial parts and their subsequent engineering changes. The speed at which they could react became a critical factor in the time it took to design and debug an IBM product. In response to this new requirement, the logistics, already compounded with a wide mix of part numbers and several personalization steps, were streamlined by using two wafer fabrication lines. One was built to produce the volume orders. The other was designed to quickly process and test the "engineering" products. These "quick turn-around lines" (one for masks and one for semiconductors) organized their tools to operate on few wafers, with minimum delay between steps. Combined with new production control systems working in real time on tools, wafers, and masks, these shorter turn-around time systems could produce a finished part in 18 days from receipt of a RIT. The conventional line would take two or three times longer.

But these quick lines still used conventional tools. With a need to become even more responsive to the machine design cycle, manufacturing began projects to explore other line configurations, using process tools specifically designed to reduce or eliminate the inhibitors to product flow rates. The result of these early efforts culminated in manufacturing systems like QTAT (Quick Turn-Around Time) [37], which will be discussed in the last section.

The first Emerald products, introduced in late 1971, could only integrate about 300 logic circuits. We were not gated by process yields (since we were already making 18K-bit Read-Only-Memory chips), but by the design automation systems then available, which were limited in their capability to simulate designs and to generate the required test formats. To package these larger chips with more and closer-spaced terminals, manufacturing supplied a new type of substrate. Silk screening gave way to a new process called metallized ceramic (MC) [38], which used an evaporated lamination of chromium, copper, and chromium shaped by photolithography. The area of the substrate was increased fourfold to provide space for up to 100 pins per module. Within a few years, volume production of Emerald circuits began to exceed that of bipolar products. By 1972, the FET memory had taken over most main memory applications and it was already replacing its first 1K-bit model with the 2K-bit chip designed in the Boeblingen, Germany laboratory.

As these FET technologies were being assimilated, new custom-designed bipolar products were added as well. Compared to their FET counterparts, they had higher performance, but the complications of bipolar processes resulted in optimum yields being achieved with only one-third as many circuits. Bipolar memory devices gravitated to performance-oriented applications like cache and high-speed local stores. Figure 2 illustrates the evolution of bipolar and FET memory products at IBM.

Production of semiconductors had reached another plateau. Semiconductor technologists and the machine designers could now work jointly to define semiconductor products. With the added choice of FETs or bipolars, and the ability to design their own functional semiconductors, each machine designer now had a new degree of freedom to optimize the circuits for the unique requirements of his system.

The road to VLSI

Beginning with Medium-Scale Integration (MSI), and rapidly promoting to Large-Scale Integration (LSI), the decade of the seventies witnessed the steady introduction of new semiconductor devices having progressively increased performance and functionality. Today's products mark the era of Very Large-Scale Integration (VLSI). The design, construction, and manufacture of these products are much more involved than simply building larger chips to accommodate more circuits. They have significantly more structural features throughout the silicon and employ more intricate thin-film topography. All dimensions become substantially smaller. As a consequence, these "denser" products require entirely new manufacturing systems.

Two IBM semiconductor products which began manufacture in 1976 illustrate the complexity of modern products. The first of the two is a bipolar, high-performance logic product. It uses a master slice, with 704 NAND gates for logic or receivers and 80 drivers, three levels of interconnection, and 96 input-output terminals [39, 40]. The second, called SAMOS (Silicon and Aluminum Metal Oxide Semiconductor) [41], provides extremely low-cost memory. Four custom designs (18K, 32K, 36K, and 64K bits per chip) cover a wide range of application. Although the levels of integration of these two products differ by more than an order of magnitude, each is an example of VLSI in its respective class of application.

To achieve their increased functionality, today's semiconductors make extensive use of new structural features which exploit a host of new materials and the most advanced process technologies. Some of these technological improvements are 2-μm epitaxy, recessed-oxide isolations (ROI), silicon nitride for self-aligned emitter-base contacts, dual dielectrics of silicon dioxide and nitride, platinum silicide contacts for Schottky barrier diodes and three levels of aluminum-copper alloy insulated with sputtered quartz and polyimide for interconnections, polysilicon for field shields, and ion implantation for more precise deposition of dopants.

These new materials-oriented processes required new tools—three-inch-diameter crystal growers which precisely controlled oxygen content, Chemical Vapor Deposition (CVD) for thin films of silicon nitride, doped oxides and polysilicon, rf sputtering machines for improved "planarized quartz," and ion implantors. There was now a major difference in the construction of a manufacturing facility: A companion industry had been spawned which specialized in supplying the semiconductor houses with many of its capital tools. IBM process strategy was modified to concentrate on developing those tools which were either unique to its process or required special manufacturing attributes. An example of the former is the "planarized quartz" machine [42], and of the latter, the IBM electron-beam direct-writing exposure system [43].

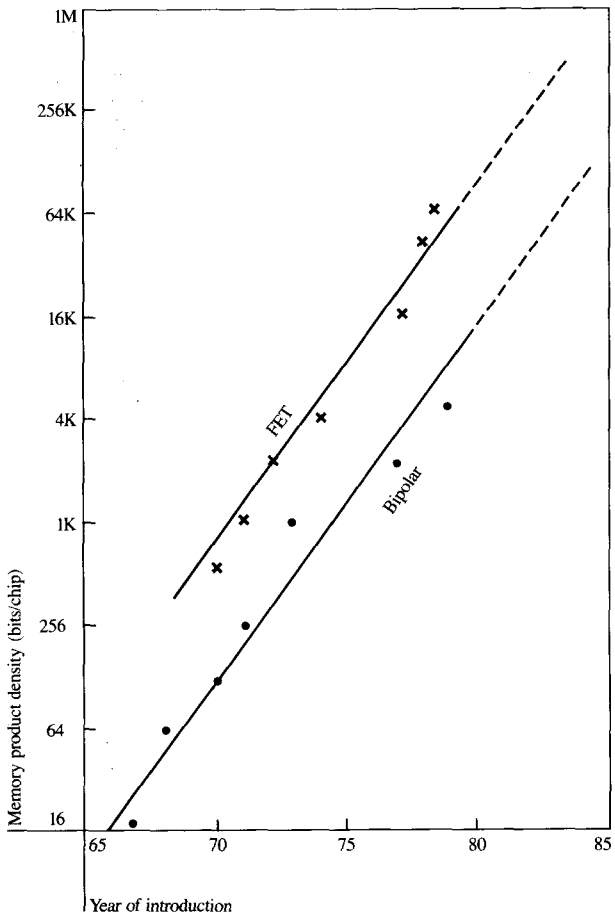


Figure 2 Productivity growth for bipolar and FET memory chips. The dots and crosses mark actual products.

It is hardly possible in this review to describe even the highlights of the manufacturing systems which take advanced logic and memory wafers through more than a hundred steps. But two subsystems, lithography and yield management, which are so integral to the process and its control, warrant discussion.

SLT lithography used contract printing, a single negative resist and developer, and two chemical etchants, and had only four masking steps. By contrast, today's wafers use optical projection and electron-beam printing. Each requires special resists. The optical resist is further complicated by additional layers used in the "lift-off" mode. Plasma and reactive ion etching, instead of wet chemistry, are now used for superior profile control.

Table 1 Complexity and productivity comparisons of two IBM semiconductor logic technologies.

	1964	*1980*
Components per circuit	9	5
Minimum feature size (μm)	12.5	3
Gates per chip	0.25	3234
Wireable circuits per chip	0.25	~1500
Chip size (mm^2)	0.66	22
Circuit density (ckts/mm^2)	0.38	68
Major process steps	16	38
Lithography steps	5	14
Interconnection levels	1	3
Terminals per chip	3	122
Wafer size (mm^2)	800	5280
Circuits per wafer (unyielded)	300	273 000

Modern bipolars have increased to 12 masking steps, and even the FET uses nine. While the smallest feature ever made in SLT was 7.5 μm, new products routinely have 2.5-μm dimensions.

The single most important factor in semiconductor manufacturing is yield. It is the primary determinant of product cost and it also determines the kinds of products that can be made. As the manufacture of one product approaches its ultimate yield, it will become more economical to introduce a new one which has a higher level of integration. Although the yield for the second product will be lower, it will be more efficient in the effective use of silicon, and at the same time will provide more function. In IBM, those activities which control and improve yield are called yield management. It has evolved to a level of sophistication matching that of the other elements of semiconductor technology. Yield management is a combination of disciplines and procedures which guide the daily manufacturing operation in maintaining process control and also sets and monitors the plan to bring each manufacturing system to an ultimate yield objective.

For a modern product, like SAMOS for example, to qualify for manufacturing, a four-year management plan is formulated aimed at achieving an ultimate yield. Each manufacturing engineering department analyzes its particular processes and then places its results into a set of mathematical models to calculate yield. There are models like Design-Limited Yield, Lithography-Limited Yield, and Defect-Limited Yield. The results of each of these separate models are combined in a statistical model which predicts the overall process and final test yield. With these results, a total plan is developed to improve process, tools, and masks to achieve the ultimate yield objective. ("Learning" is planned, so to speak.)

To compare the performance of the manufacturing system to the plan, a set of semiconductor test structures are designed to monitor its daily performance and to provide feedback on proposed design and process modifications. SAMOS uses three kinds of test structures. Process monitors are especially designed to measure the performance of individual process steps. Test sites measure a combination of processes such as the basic FET device. These monitors are placed on one (or between two) of the chip sites of standard production wafers. The third structure estimates the density of defects in each process step. It is a way to simulate the structure within and between levels of personalization and is used to quantify various defect modes. This structure uses its own wafer and is routinely fabricated in the wafer fabrication line. Measurements from these structures are then analyzed and combined to estimate the final process yield. These special test structures are the "hardware" of IBM's yield management system.

QTAT is a good illustration of the kinds of subsystems needed to manufacture today's advanced products. It was designed for two purposes: to reduce the turn-around time on new parts, and to respond to critical "mix" problems. While its logistics do little to change the time to fabricate wafers, the in-process waiting time is an order of magnitude less than the conventional IBM high-volume-oriented line. QTAT uses the principle of "one wafer at a time," each with its own machine-readable serial number. It exploits process automation (as distinct from mechanization) by combining individual wafer logistics with computer-managed process and measurement tools to feed product and test site information forward and back to optimize the processing of each wafer based on its individually measured characteristics.

One technique that is applied is called "End-Point Detect," whereby computerized instruments continually measure an operation as it proceeds and terminate it automatically when the desired results are achieved. As one might expect, operating with this mode has produced yields higher than conventional lines.

The heart of QTAT is IBM's electron-beam direct-writing exposure system which prints the patterns defining the metal and insulator interconnection layers. This system operates with a digital description of each device's topography and thus avoids the time, expense, potential defects, and the added logistics of photo masks. With this machine, up to eight different part numbers can be processed on one wafer. Although the development of

this machine began with the objective of surpassing photo-optics in the control of finer dimensions, its first application was to meet the unique needs of a logistics system.

Conclusion

It is interesting to look back from where we are today. Twenty-five years ago we accelerated our development and manufacture of semiconductors because we knew they would be so important to computers. During the sixties and seventies semiconductor products were key to the economics of computers. Now the computer itself is indispensable to the fabrication of semiconductors. The rates of advancement in both semiconductors and computers have now become mutually dependent.

The Hudson Valley, where much of what we have just reviewed took place, is the home of Rip van Winkle. Had he gone to sleep twenty-five years ago and returned today, he would know there had been a revolution in this industry, and no doubt be astonished by it. From labor- to capital-intensive, the ratio of direct to indirect personnel is completely inverted. Where at one time no computers existed, they now abound. While we once made only a few kinds of transistors in small volume, the Fishkill plant in 1981 alone will make 10 000 different part numbers and release 3000 new ones. At the same time we will still be making significant numbers of all the old products, some of which are twenty years old.

Looking ahead, it seems easy to project the near-in trends. Semiconductors will continue their accelerating advance. The technological alternatives will continue to improve as they try to survive in their battle for a place in the hierarchy. And with these technological advances, new applications will open as density, function, cost, and reliability become better.

But the long-range predictions are tougher to make. One thing this history has taught us: Every time we look back about seven years, we find we have changed in directions which we had not seen originally. If the past is prologue, then while it is unclear what the next ten years may bring, they are certain to bring new and greater challenges.

Acknowledgments

The author wishes to acknowledge J. Ayling, D. Dewitt, W. Harnett, and R. D. Trauben for their help in the preparation of this paper.

References

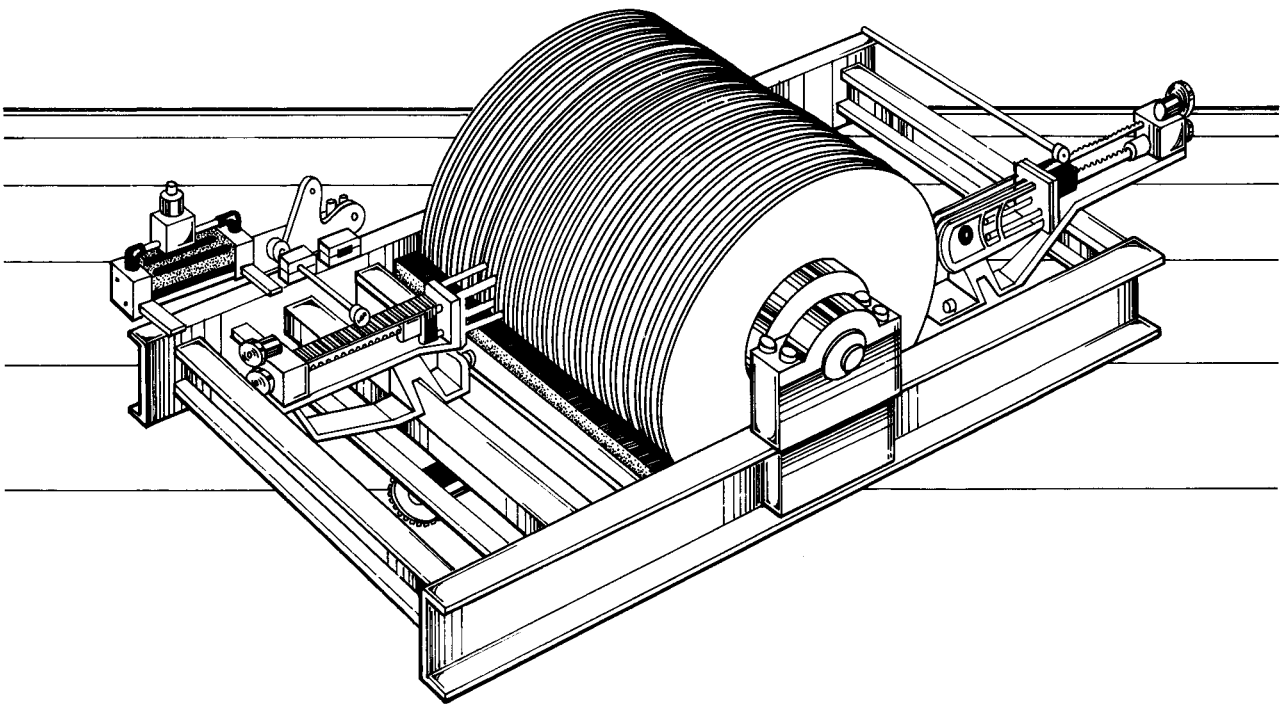
1. T. J. Leach, "Automated Assembly of Alloy-Junction Transistors," *Electronics* **33,** 57 (1960).
2. R. L. Moore, "High-Speed Servo Positioner Bonds Mesa Transistors," *Electronics* **36,** 58 (1963).
3. J. L. Langdon, W. E. Mutter, R. P. Pecoraro, and K. K. Schuegraf, "Hermetically Sealed Silicon Chip Diodes and Transistors," presented at the IEEE Professional Group on Electron Devices Meeting, Washington, DC, 1961.
4. E. M. Davis, W. E. Harding, R. S. Schwartz, and J. J. Corning, "Solid Logic Technology: Versatile, High-Performance Microelectronics," *IBM J. Res. Develop.* **8,** 102 (1964).
5. D. P. Seraphim and I. Feinberg, "Electronic Packaging Evolution in IBM," *IBM J. Res. Develop.* **25,** 617 (1981, this issue).
6. C. J. Frosch and L. Derick, "Surface Protection and Selective Masking during Diffusion in Silicon," *J. Electrochem. Soc.* **104,** 547 (1957).
7. J. A. Perri, H. S. Lehman, W. A. Pliskin, and J. Riseman, "Surface Protection of Silicon Devices with Glass Films," presented as a "Recent News" paper at the Electrochemical Society Meeting, Detroit, MI, October 2–4, 1961.
8. P. A. Totta and R. P. Sopher, "SLT Device Metallurgy and its Monolithic Extension," *IBM J. Res. Develop.* **13,** 226 (1969).
9. W. A. Pliskin and E. E. Conrad, "Techniques for Obtaining Thin Glass Films on Substrates," *Electrochem. Technol.* **2,** 196 (1964).
10. A. H. Mones, J. Boyd, and J. Schottmillen, "Printed Resistors," presented at the Electrochemical Society Symposium, Pittsburgh, PA, April 1963.
11. W. H. Miller and F. Barson, "Semiconductor Devices and Passivation Thereof," U.S. Patent No. 3,343,049, 1967.
12. J. E. Thomas, Jr. and D. R. Young, "Space-Charge Model for Surface Potential Shifts in Silicon Passivated with Thin Insulating Layers," *IBM J. Res. Develop.* **8,** 368 (1964).
13. D. R. Kerr, J. S. Logan, P. J. Burkhardt, and W. A. Pliskin, "Stabilization of SiO_2 Passivation Layers with P_2O_5," *IBM J. Res. Develop.* **8,** 376 (1964).
14. D. R. Kerr, "Effect of Temperature and Bias on Glass-Silicon Interfaces," *IBM J. Res. Develop.* **8,** 385 (1964).
15. P. P. Castrucci and J. S. Logan, "Electrode Control of SiO_2 Passivated Planar Junctions," *IBM J. Res. Develop.* **8,** 394 (1964).
16. D. P. Seraphim, A. E. Brennemann, F. M. d'Heurle, and H. L. Friedman, "Electrochemical Phenomena in Thin Films of Silicon Dioxide on Silicon," *IBM J. Res. Develop.* **8,** 400 (1964).
17. H. S. Yourke, "Millimicrosecond Transfer Current Switching Circuits," *IRE Trans. Circuit Theory* **CT-4,** 236 (1957).
18. W. Schuelke, "Modular Approach to System Design," *Automation* **14,** 77 (1969).
19. E. F. Platz, "Solid Logic Technology Computer Circuits—Billion Hour Reliability Data," *Microelectronics and Reliability*, Vol. 8, Pergamon Press, Inc., Elmsford, NY, 1969, p. 55.
20. J. S. Kilby, "Miniaturized Electronic Circuits," U.S. Patent No. 3,138,743, 1964.
21. R. N. Noyce, "Semiconductor Device and Lead Structure," U.S. Patent No. 2,981,777, 1961.
22. B. Agusta, P. Bardell, and P. Castrucci, "A 16-Bit Monolithic Memory Array Chip," *J. Appl. Phys.* **37,** 574-579 (1966).
23. P. D. Davidse and L. I. Maissel, "Dielectric Thin Films Through RF Sputtering," presented at the Third International Vacuum Congress, Stuttgart, Germany, 1965.
24. L. F. Miller, "Controlled Collapse Reflow Chip Joining," *IBM J. Res. Develop.* **13,** 239 (1969).
25. H. A. Khoury and H. R. Rottman, "Apparatus for Contouring the Surface of Thin Elements," U.S. Patent No. 3,729,966, 1973.
26. O. Bilous, I. Feinberg, and J. L. Langdon, "Design of Monolithic Circuit Chips," *IBM J. Res. Develop.* **10,** 370 (1966).

27. R. C. Piatzek and J. S. Kilby, "Minuteman Integrated Circuits—A Study in Combined Operations," *Proc. IEEE* **52,** 1669 (1964).
28. E. J. Rymaszewski, J. L. Walsh, and G. W. Leehan, "Semiconductor Logic Technology in IBM," *IBM J. Res. Develop.* **25,** 603 (1981, this issue).
29. J. K. Ayling, R. D. Moore, and G. K. Tu, "A High-Performance Monolithic Store," *ISSCC Digest of Technical Papers*, 36 (1969).
30. H. R. Gates, J. D. McKinney, and W. D. North, "Bipolar LSI for Main Memory," *ISSCC Digest of Technical Papers*, 78 (1971).
31. M. J. Kim, "MOSFET Fabrication Problems," *Solid State Electron.* **12,** 557 (1969).
32. E. S. Schlegel, "A Bibliography of Metal-Insulator-Semiconductor Studies," *IEEE Trans. Electron Devices* **ED-14,** 728 (1967).
33. P. Balk, "Low Temperature Annealing in the Aluminum SiO_2 Silicon System," presented at Electrochemical Society Meeting, Buffalo, NY, October 1965.
34. H. S. Lehman, "Method for Controlling the Electrical Characteristics of a Semiconductor Surface and Product Produced Thereby," U.S. Patent No. 3,402,081, 1968.
35. E. W. Pugh, D. L. Critchlow, R. A. Henle, and L. A. Russell, "Solid State Memory Development in IBM," *IBM J. Res. Develop.* **25,** 585 (1981, this issue).
36. R. F. Schauer, "An Integrated Design and Manufacturing Approach to VLSI," presented at the 1980 Conference on Mission Assurance, Los Angeles, CA, April 28–May 3, 1980.
37. N. G. Wu, "Automated Wafer Production," *ISSCC Digest of Technical Papers*, 208 (1980).
38. R. W. Gedney, "Trends in Packaging Technology," presented at the 16th Annual Reliability and Physics Conference, San Diego, CA, April 1980.
39. R. J. Blumberg and S. Brenner, "A 1500-gate Random Logic LSI Masterslice," *ISSCC Digest of Technical Papers*, 60 (1979).
40. T. S. Jen and N. Nan, "Gate Array Experiences in IBM," Paper No. 22-3, *Electro-80 Conference Record*, Boston, MA, May 13–16, 1980; Western Periodicals Co., N. Hollywood, CA.
41. Richard A. Larsen, "A Silicon and Aluminum Dynamic Memory Technology," *IBM J. Res. Develop.* **24,** 268 (1980).
42. J. S. Lechaton, "High Resputtered SiO_2 and Nonoverlap Via Holes," extended abstracts of the Electrochemical Society meeting of October 1979, Abstract No. 585, Vol. 79-2, p. 1466.
43. E. V. Weber and H. S. Yourke, "Scanning Electron-Beam Turns Out IC Wafers Fast," *Electronics* **50,** 96 (1977).

Received October 16, 1980; revised February 26, 1981

The author is located at the IBM General Technology Division laboratory, East Fishkill Facility, Hopewell Junction, New York 12533.

Magnetic Recording Technology



Artist's conception of the finished random access file (1953)

The *Journal* acknowledges the contributions of H. B. Michaelson to the acquisition, review, and editing of the papers in this section.

Editor