Evolution of the MOS Transistor—From Conception to VLSI

CHIH-TANG SAH, FELLOW, IEEE

Invited Paper

Historical developments of the metal-oxide-semiconductor field-effect-transistor (MOSFET) during the last sixty years are reviewed, from the 1928 patent disclosures of the field-effect conductivity modulation concept and the semiconductor triodes structures proposed by Lilienfeld to the 1947 Shockley-originated efforts which led to the laboratory demonstration of the modern silicon MOSFET thirty years later in 1960. A survey is then made of the milestones of the past thirty years leading to the latest submicron silicon logic ***CMOS*** *(Complementary MOS) and* ***BICMOS*** *(Bipolar-Junction-Transistor CMOS combined) arrays and the three-dimensional and ferroelectric extensions of Dennard's one-transistor dynamic random access memory (****DRAM****) cell. Status of the submicron lithographic technologies (deep ultra-violet light, X-ray, electron-beam) are summarized. Future trends of memory cell density and logic gate speed are projected. Comparisons of the switching speed of the silicon MOSFET with that of silicon bipolar and GaAs field-effect transistors are reviewed. Use of high-temperature superconducting wires and GaAs-on-Si monolithic semiconductor optical clocks to break the interconnect-wiring delay barrier is discussed. Further needs in basic research and mathematical modeling on the failure mechanisms in submicron silicon transistors at high electric fields (hot electron effects) and in interconnection conductors at high current densities and low as well as high electric fields (electromigration) are indicated.*

I. Introduction

Some of the geometrics and basic concepts underlying the **MOS** transistor or the metal-oxide-semiconductor field-effect transistor (**MOSFET**) were described sixty years ago by Julius Edgar Lilienfeld of Brooklyn and Cedarhurst, NY, in the first two of his three patents [1]–[3]. (A list of acronyms and abbreviations follows.) Applications of Lilienfeld's three patents were filed with the United States Patent Office on October 8, 1926 [1], March 28, 1928 [2], and December 8, 1928 [3]. They were granted in 1930, 1933 and 1932, respectively. Lilienfeld was born in Poland in 1882 and was a professor of physics at the University of Leipzig from 1910 to 1926. He immigrated to the United States in 1926 and was the director of research of the Ergon Research Laboratory at Malden, MA. He became a naturalized U.S. citizen in 1935, retired to the Virgin Islands soon after with his American-born wife, the former Beatrice Ginsberg, and died on August 28, 1963, at the age of 81 [4].

In order to see how Lilienfeld arrived at his MOS transistor concepts and structures, we first give a summary of the analyses of the patents, using modern semiconductor terminology and the semiconductor device theory of Shockley. My description is helped by the modern knowledge of basic transistor theory developments since 1947. As we shall see, field-effect transistor structures and several other transistors were described by Lilienfeld's first two patents and a later (1935) patent by Heil [5], except the p-n junction field-effect and minority-carrier injection bipolar transistors which were invented and experimentally demonstrated by Bardeen, Brattain, Shockley and their colleagues at the Bell Telephone Laboratories during 1947–1952 [6]. My analyses show that there are five transistor structures in the three Lilienfeld patents. We have no evidence, however, if any of the transistors proposed by Lilienfeld were built and worked.

The first Lilienfeld patent [1] gives the **MESFET** (Metal/Semiconductor FET).

The second Lilienfeld patent gives two transistor structures [2]. They are derived from the MESFET of the first patent by inserting an insulator layer between the metal gate and the semiconductor film resulting in the depletion-mode MOSFET and the **GMSR** (Gated Metal/Semiconductor Rectifier Transistor or **GMSRT**). I call the GMSR a transistor instead of just a gated diode, a term I used in 1962 (Surface Controlled Diode or Surface Controlled Tetrode) since it is a three-terminal device although it may not give power amplification except under certain conditions as demonstrated by the first confirmed amplification in a solid state device in Brattain's December 4, 1947 experiment. (See Fig. 13 on p. 609 and text on pp. 609 and 611 of Shockley's review [6].) I use the term 'transistor' in this paper as a generic name

Manuscript received August 1, 1986; revised July 14, 1988. The research was supported in part by two sequential grants from the U.S. National Science Foundation, three sequential and ongoing contracts with the Semiconductor Research Corporation concerning silicon MOS aging and failure mechanisms as well as reliability physics, chemistry, and modeling, and an unrestricted Intel grant for the author's stipend of the summer of 1988.

The author was with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL. He is now with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611, USA.

IEEE Log Number 8823837.

0018-9219/88/1000-1280$01.00 © 1988 IEEE

for solid-state electron devices with three or more terminals. The technical definition used by Bell Lab researchers and IEEE is that it must have a power gain greater than unity. Brattain's experiment was on p-type silicon with a thin surface p-n junction from an n-type surface inversion layer, suggested by Bardeen and Shockley in the first transistor patent filed by Bell Labs in 1948 [7], described in Bardeen's Nobel lecture, and reviewed in 1976 by Shockley on the invention of the bipolar junction transistor [6]. The gated p-n junction diode structure I used in 1962 was given as Fig. 1 of Shockley's June 26, 1948 bipolar transistor patent [8], and also Brattain's December 4, 1947 notebook figure, part II, given in Fig. 13 of Schockley's review [6], unknown to me in 1962. I will discuss the recent results later.

The third Lilienfeld patent [3] gives two more transistor structures, the Metal Base Transistor or the Semiconductor/Metal/Semiconductor Transistor (**SMST**), and the Schottky-Barrier-Emitter/Schottky-Barrier-Collector Transistor or the Metal/Semiconductor/Metal Transistor (**MSMT**). These two transistors do not operate on the conductivity modulation or field-effect principle as do the three transistors of the first two patents.

In addition, several integrated amplifier structures consisting of different intergrated configurations of these five basic transistor structures were also given by Lilienfeld. For examples, the second patent gives two integrated structures, the first containing three transistors (MESFET, MOSFET, GMSR), and the second containing a GMSR transistor and a metal-semiconductor rectifier (**MSR**) diode with a distributed and voltage dependent resistor, and the third patent gives a five-layer MPMPM transistor where *M* is Metal and *P* (p-type semiconductor) is Cu_2S.

Circuit diagrams with dc bias voltages are also given by Lilienfeld with the correct polarities for proper operation as an amplifier although he was limited to batteries with negative common. He even gives a complete receiver circuit in the first patent using four MESFETs as RF and AF amplifiers and a MSR as the detector.

The patents describe in considerable detail methods for making the thin-film transistor structures and Lilienfeld's experiences on how the transistor characteristics are affected by varying the fabrication conditions, as well as such characteristics as the magnitude of the breakdown voltage, the reproducibility of the Schottky barrier or metal-semiconductor rectifier diode and the goodness of the ohmic contacts. Copper sulfide is mentioned as the 'compound' or semiconductor film in the first (1926) patent, which also gives two methods each on deposition of the copper film (by spattering or thermal evaporation in vacuum and by deposition from a colloidal suspension) and sulphuration of the film (in sulfur vapor or in carbon bisulphide liquid). In the first patent, which gives the MESFET, the 0.0001" or 2.5-μm-long aluminum gate is very ingeniously made by squeezing together an aluminum foil between the two edges of a broken glass slide on which the copper film is then deposited and sulphurized. Copper sulfide, copper oxide, and lead oxide are specifically mentioned in the second patent as possible candidates for the semiconductor compound film by sulfuration or oxidation. The second patent also gives the chemical formulae explicitly, Cu_2S, Cu_2O and PbO_2 and describes Cu_2S as copper disulphide while the modern names are cuprous sulfide and cuprous oxide. They are known today to be p-type semiconductors due to the copper vacancy produced during sulfuration and oxidation and have a rather low hole mobility, about 1 cm^2/V-s in Cu_2S.

My foregoing description of Lilienfeld's transistors relies on an analysis of his transistor structures using the concepts of hole, p-type semiconductor and energy bands. These concepts were not used by Lilienfeld, which prevented him from developing the basic physics of these transistors. They were not used probably because they were just developed by A. H. Wilson in 1931 [9], [10] after Lilienfeld's patent applications were filed in 1926 and 1928. However, Lilienfeld understands well a key concept, conductivity modulation by a transverse electric field, which he repeatedly uses to describe signal amplification of the transistors in the patents. He uses this concept to explain the operation of both the MESFET in the first patent and MOSFET in the second patent. In addition, he knows well where to place the rectifying junctions and the ohmic junctions for contacts to the external circuits. The proliferation of transistor structures in Lilienfeld's patents further indicates his grasp of the conductivity modulation principle and his ability to apply it in combination with ohmic and nonrectifying contacts. However, we still do not know and have no evidence if Lilienfeld made a working, amplifying transistor.

In February 1964, after Lilienfeld had died the previous year, Bottom [11] published a suggestion that the transistor in Lilienfeld's first patent is the bipolar (minority carrier injection) junction transistor. Bottom's erroneous conclusion was made under the false assumption that the compound semiconductor film, Cu_2S, is n-type and that aluminum had diffused through the thin film to form a p-type region during the pulsed 'forming' operation while making the transistor. This was corrected by Johnson [12], the discoverer of the Johnson Noise. Johnson pointed out that Cu_2S is p-type and that the Al/Cu_2S Schottky diode was known in the 1930s as a rectifier with high resistance when Al is biased positively relative to Cu_2S. Johnson also reported that he had tried to duplicate the structure but could not get amplification or even modulation. He attributed the failure to the low mobility of holes (1 cm^2/V-s) and the trapping of holes at the surface states and further stated that Lilienfeld's receiver could not have worked due to the transistor's frequency limitation. But, Lilienfeld stated a gate length of 0.0001" or 2.5 μm which would have given a hole transit time of 62 ns or a cutoff frequency of 2.5 MHz at a drain-to-source voltage of 1 V for a mobility of 1 cm^2/V-s. Even high trap density would not have increased the transit time since the traps are already mostly filled by the trapped holes. Thus, Johnson proved it was not a bipolar transistor. It is uncertain if the MESFET nature of Lilienfeld's first transistor would have made Johnson's experiments successful since the impurity ion density in the Cu_2S film and the surface state density at the glass-substrate|Cu_2S-film interface could also be too high.

A patent application, filed by Oskar Heil, a native of Germany, with the British Patent Office on March 4, 1935 and granted on December 6, 1935 [5] provides the first description of the MOSFET operation using modern semiconductor concepts of electrons and holes. Both n-type and p-type semiconductor films were specified with Te, I, Cu_2O and V_2O_5 as explicit examples. The MOSFET structure described by Heil is a long thin semiconductor film whose two ends

are covered by a metal stripe, to serve as the drain and the source contacts. It also has a gate metal electrode overlapping but insulated from the semiconductor film and the drain and source metal stripes to modulate the conductance of the semiconductor film. Both the single gate over one surface of the semiconductor film and two gates over both surfaces of the semiconductor film are given. Heil's Metal/n-film/Metal (M/n/M) structure using an n-type semiconductor film is the modern depletion mode MOSFET just like Lilienfeld's Au/p/Au which uses ohmic metal (Au) contact to the p-Cu_2S. Heils' M/p/M structure, using a p-type semiconductor film, would be the modern inversion-channel or enhancement-mode MOSFET since its souce and drain contacts are metal/p-semiconductor rectifying junctions. The enhancement-mode MOSFET is not contained in Lilienfeld's three patents whose transistor structures have at most one rectifying junction. However, Heil did not explain how his M/p/M inversion-channel transistor would work. Thus, I conclude that he did not recognize the necessity of having a gate-voltage induced surface inversion channel (n-channel for the M/p/M transistor) in order to have amplification. The inversion channel idea was first recognized by Bardeen in 1947 [6], [7]. It led to the invention of several inversion channel field-effect transistors, including a predecessor of the bipolar point-contact transistor by Bardeen and Brattain in 1948 [13] and the modern MOSFET with diffused source and drain junctions, all of which we will discuss later.

One is curious about what motivated Lilienfeld and provided him the precedents that led him to so many transistor structures. It seems that Lilienfeld's earlier research on electron emission in vacuum, published in a 1920 Leipzig university journal [12], and his experimental expertise in vacuum tube electronics must have led him to invent the solid state devices. More recent examples of transistor inventions [7], [8] based on previous or related knowledge in vacuum tubes were those of Bardeen, Brattain, and Shockley, as described by Bardeens's Nobel lecture [7] and by Shockley, whom I quote: "accidents (inventions) favor the prepared mind" [6].

To visualize Lilienfeld's MOSFET inventions, I make use of several figures from his first two patents in the following descriptions.

The first figure of Lilienfeld's first (1926) patent [1] is reproduced in Fig. 1(a) with modern symbols added, D = Drain, S = Source, and G = Gate. An enlarged view of the Al/p-Cu_2S contact region is given in Fig. 1(b) which helps to ascertain that it is a MESFET or Schottky-gate FET and that Lilienfeld has the proper reverse bias (positive voltage) applied to the gate and (negative voltage) to the drain. One can readily see that the amplifier circuit is a source follower although the placements of the power supplies are somewhat unconventional, probably dictated by having only positive power supplies. Lilienfeld's understanding of the MESFET operation principle is clearly indicated by his patent descriptions on lines 84–86 and 92–96 on page 1 [1]. Some explanations are given in brackets, { }, in the following quote: "the metal foil electrode {14 in Fig. 1(a) where 10 is a broken glass slide} and the compound {p-Cu_2S film labeled 15} form an element of unidirectional conductivity {Schottky barrier or metal-semiconductor rectifier} and the thickness of the film is minute {he stated elsewhere in the patent it is 0.0001 inch or 2.5 μm} and that the electrical conductivity

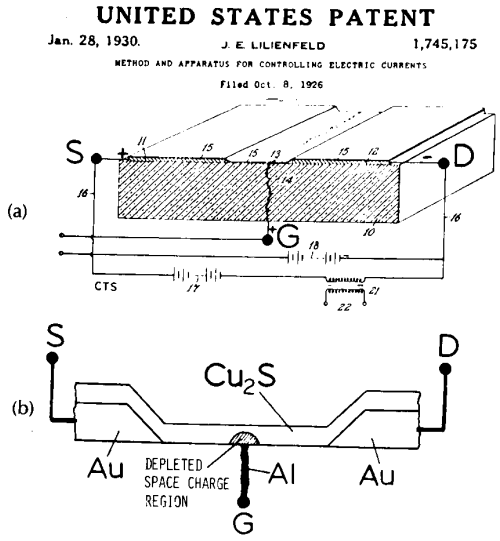


Fig. 1. (a) Amplifier circuit and dc bias polarity of Lilienfeld's 1926 patent [1] which apparently described the Schottky-barrier gate or metal-semiconductor gate field-effect transistor (MESFET) structure. (b) The enlarged view of the active gate and channel regions of transistor in (a).

is influenced by the electrostatic force applied across the metal foil compound film contact."

The first two figures of Lilienfeld's second patent [2], the MOSFET patent, are reproduced in Fig. 2(a) and (b). The modern circuit symbol is given in Fig. 2(c). Lilienfeld has all the ingredients of the modern depletion-mode MOSFET. He has an aluminum substrate as the gate electrode, the aluminum oxide as the gate insulator and p-type cuprous sulfide (Cu_2S) as the semiconductor. His two structures in Fig. 2 are in fact the thin film version of the MOSFET. It is

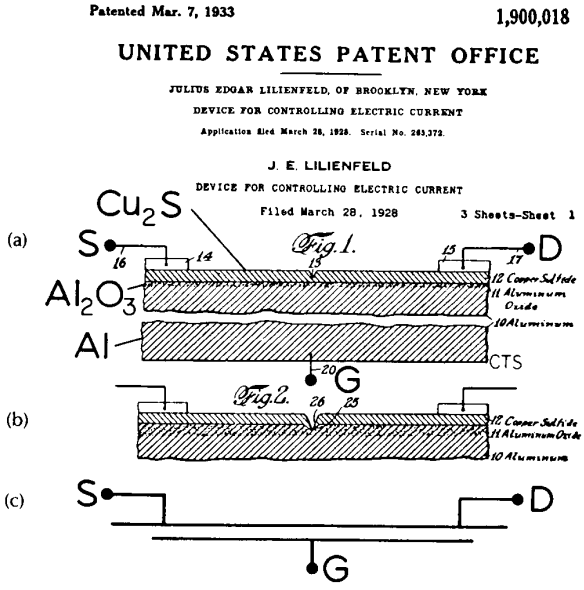
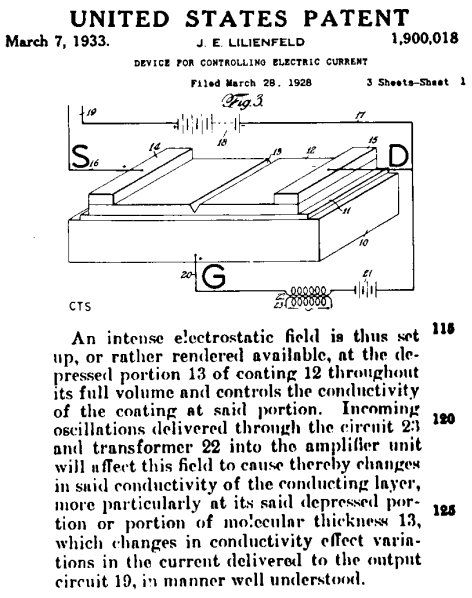


Fig. 2. (a), (b) The first two figures in Lilienfeld's 1933 patent [2] giving two structures of the metal-oxide-semiconductor field-effect transistor (MOSFET) he proposed. (c) The modern circuit symbol of the MOSFET.

also a facsimile of the modern depletion mode MOSFET which is now made with a doped conduction channel produced by diffusion, epitaxial growth or ion implantation on the surface of a silicon substrate. It finds applications as an amplifier as well as a "pinched" resistor load in modern silicon integrated circuits. In the thin film version, the conducting film is made by recrystallization of a deposited polycrystalline silicon thin film. This depletion MOSFET has no rectifying junction for the drain and the source in contrast to the enhancement or inversion channel MOSFET. Stability and reproducibility problems have prevented experimental CdS and silicon thin-film transistors from reaching the marketplace so far. However, significant recent development efforts have continued on growing low-defect silicon film by recrystallization of polycrystalline Si films deposited on oxide covered silicon substrate, known today as **SOI** (Silicon On Insulator). These efforts are motivated by the potential of denser integrated circuits using the third and vertical dimension.

Lilienfeld's understanding of the MOS field-effect transistor operation principle is further demonstrated by the descriptions in his second patent. He states that a high gate electric field is required to operate the transistor, about 10^7 V/cm computed from the 100 V and 10^{-4} mm numbers given in the patent. However, the transistor operation voltages he states are about 10 V so that the field is about 10^6 V/cm which falls right in the range of modern silicon MOSFET operation. The circuit given in Fig. 3 again shows that Lilienfeld gives the correct bias polarities to operate the transistor as a source-follower amplifier.


UNITED STATES PATENT

March 7, 1933. J. E. LILIENFELD 1,900,018

DEVICE FOR CONTROLLING ELECTRIC CURRENT

Filed March 28, 1928 3 Sheets-Sheet 1

Fig. 3

CTS

An intense electrostatic field is thus set up, or rather rendered available, at the depressed portion 13 of coating 12 throughout its full volume and controls the conductivity of the coating at said portion. Incoming oscillations delivered through the circuit 23 and transformer 22 into the amplifier unit will affect this field to cause thereby changes in said conductivity of the conducting layer, more particularly at its said depressed portion or portion of molecular thickness 13, which changes in conductivity effect variations in the current delivered to the output circuit 19, in manner well understood.

Fig. 3. Amplifier circuit, dc bias polarity, and description of the device operation physics given in Lilienfeld's 1933 patent [2] of the metal-oxide-semiconductor field-effect transistor.

A most important feature of Lilienfeld's invention was that he understood the conductivity modulation principle of the transistor, based on electrostatics, at a time when quantum mechanics was in its infancy and semiconducting hole concepts had been just developed by A. H. Wilson in 1931 [9], [10], which Lilienfeld was probably not aware of or did not use. His description in the patent is reproduced in the lower part of Fig. 3, which explicitly states that the conductivity of the thin semiconductor channel region (portion 13) is modulated by the input signal applied to the gate (labeled 20) through the input transformer.

The foregoing analyses of Lilienfeld's three patents and Heil's one patent led me to conclude that Lilienfeld proposed five transistor structures: the MESFET in 1926; and the depletion-mode MOSFET, the gated metal-semiconductor transistor, the metal-base transistor, and the metal-semiconductor-metal transistor, all in 1928. And Heil proposed three transistor structures: the single-gate inversion-channel or enhancement-mode MOSFET, the dual-gate enhancement-mode MOSFET, and the dual-gate depletion-mode MOSFET. We must also concede that Lilienfeld understood the conductivity modulation principle and how the depletion-mode MESFET and MOSFET work as is evident by the proliferation of integrated device structures given by him in the patents. However, it is clear that neither Lilienfeld nor Heil recognized the necessity of the surface inversion channel for the operation of the enhancement-mode MOSFET proposed by Heil or the gated metal-semiconductor rectifier transistor GMSRT proposed by Lilienfeld. It is also clear that the MESFET and MOSFETs proposed by these two inventors are field-effect devices depending on conductivity modulation, and not on minority-carrier injection. The minority-carrier injection phenomena underlying the basis of the bipolar junction transistor was proposed by Shockley on January 23, 1948 [6], [8], demonstrated by Shive experimentally on February 18, 1948 [6], and suggested in the point-contact transistor patent of Bardeen and Brattain [13] filed on June 17, 1948. The minority carrier injection principle was not known or recognized by Lilienfeld and Heil.

II. Organization of This Review

In the remaining review, I would like to focus on the highlights of the evolution of the **modern** silicon MOS transistor by pointing out some of the important milestones in the development and manufacturing technology of the transistor as well as its integrated circuits, by describing some recent advances of the silicon MOS integrated circuit technology, by reviewing some comparative studies on the performance of the silicon MOSFET relative to silicon bipolar junction transistor (**BJT**) and the compound semiconductor (GaAs) field-effect transistors (FETs), and finally, by indicating some of the current material and technology problems which must be solved to realize the full potential of the Si MOSFET.

The chronological advances of the MOSFET can be divided into three phases: a discovery phase during the first thirty years from 1928 to 1958, a technology development and new device structure invention phase during the next ten years from 1959 to 1968, and the MOS transistor integration and integrated circuit manufacturing phase from 1968 to today. This choice is based on the two singularly most important contributions which affected and started the feverish development efforts in the second and third phases. The choice is clearly unequivocal. Phase I might be thought to have started when Lilienfeld applied for his patents during 1926–1928. The first phase really began around

1947 at Bell Laboratories by Bardeen, Brattain, and Shockley when the basic transistor physics was clearly understood and applied to the new transistor structures. Atalla, Tannenbaum, and Scheibner started the second phase by successfully completing a Bell Laboratory project in 1959 which had the specific objective of finding a passivation method to stabilize the silicon surface using the thermally grown silicon dioxide. And finally, Noyce and Moore started the third phase in 1968, by first inventing the monolithic integrated circuit concept in 1960 and then starting a silicon MOS integrated circuit manufacturing company (Intel) in 1968, resulting in volume delivery in 1970.

These three phases will be discussed in the next three Sections (III, IV, and V). Some anticipations of future advances are then made in the next three Sections (VI, VII, and VIII). Performance trends of the silicon MOSFET are described in Section VI. Comparisons of silicon with gallium arsenide transistors are summarized in Section VII. Research needs to achieve higher speed and density are discussed in Section VIII. A summary is given in Section IX.

In this review of history, I do not explicitly distinguish the invention of a transistor structure or technology from the discovery of a fundamental transistor principle since this distinction is hopefully self-evident in my presentation. However, I want to reiterate two commonly held convictions: i) Recent inventions and new ideas have mostly come directly or indirectly from expanding and improving old ideas and previous knowledge. Even Lilienfeld's 1926–1928 transistor inventions were probably influenced by his earlier (1920) and others' experiences in vacuum tube research. The use of previous knowledge by a prepared mind was amply described and illustrated by Shockley in his descriptions of the invention of the bipolar junction transistor during 1947 [6]. ii) It was not enough just to have invented a new device or apparatus. It takes the ingenuity, persistency, and basic hard work of many dedicated engineers and technicians to put the inventors' ideas into practice in order to produce inexpensive, high performance and reliable integrated circuit chips, systems and consumer products. In the limited space available, I cannot include the many reported and untold dedications, efforts, foresights, and interesting stories, but their importance in propelling the electronics industry to the present state of the art cannot be more recognized and appreciated.

III. Discovery Phase (1928-1957)

A milestone chart of the first phase is given in Table 1. There are about fifty important contributions and events during the first phase of thirty years on transistor inventions, new technology demonstrations, discovery of new device principles as well as the founding of two companies, the Shockley Semiconductor Laboratory—Shockley Transistor Corporation in 1955 and the Fairchild Semiconductor Laboratory in 1957 which seeded and initiated the thriving growth of the integrated circuit companies around Palo Alto, CA.

Included also in Table 1 are historical events on the discoveries and developments of the metal-semiconductor rectifier [14]–[16], known today as the Schottky barrier diode, which led to Lilienfeld's MESFET [1] and MOSFET [2]. It includes Shockley's MESFET [6], as well as Bardeen–Brattain's [13] and Shockley's junction-gate FET (**JGFET**) [6]. It also includes the Shockley invention of the minority carrier injection concept, the bipolar junction transistor [8], and Bardeen–Brattain's suggestion [13] of the minority-carrier-injection point-contact transistor. These ideas and developments have significantly influenced and impacted the advances of the silicon MOSFET. References are given in the right column of this and later tables. This review will highlight selected events.

A. Early Transistor Inventions and Working Transistors

The next wave of transistor inventions after Lilienfeld's three patents of 1926–1928 and Heil's patent in 1935 started about 1939 when Shockley, after three years at Bell Telephone Laboratories in search of a solid-state amplifier, was asked to get involved with Brattain's copper-oxide rectifier research [6]. This was interrupted by World War II but Shockley came up with two transistor proposals in 1939 before leaving for the war. The first was the analog transistor made of a grid of oxidized copper wire-screen imbedded into a semiconductor [6] instead of alkali halide crystal, which had been proposed and built a year earlier by Hilsch and Pohl [15] and had a cut-off frequency of about 1 Hz owing to its size of about 1 cm. The second idea from Shockley was the MESFET [6] which was not influenced by Lilienfeld's 1926 patent [1] since it had not been recognized that one of Lilienfeld's 1926 patent disclosures is indeed the MESFET until my analysis made during the preparation of this review.

The transistor innovations started anew after the war when Shockley returned to **BTL** in mid-1945 [6] and Bardeen joined BTL later that year and began a basic study of surface states [17] after Shockley showed him the experimental results by Pearson of very low conductivity modulation of silicon and germanium films in their field-effect experiments [18]. In the following year, 1946, Bardeen completed the basic theory of surface states which accounted for this lack of conductivity modulation.

The thin-film field-effect transistor experiment was reported in detail by Shockley and Pearson in 1948 [18] after initial trials in 1945 [6]. They observed conductivity modulation on p-type germanium, p-type cuprous oxide (Cu_2O), and n-type silicon. However, the conductivity modulation they measured was a thousand times smaller than expected. They attributed this deficiency to trapping of the modulated charges by surface states according to Bardeen's theory of surface states [17] since low mobility of the deposited film cannot account for such a small conductivity modulation. Their tabulated data gave a surface state density of 5×10^{13} states/cm^2-eV. This is almost one surface state or dangling silicon bond per ten silicon surface atoms. It is 10000 times higher than that on the oxidized silicon of today's MOS VLSI circuit chips.

At the same time, experiments on Bardeen's gated surface inversion layer [7] was carried out by Brattain with suggestions from Bardeen and Shockley [6]. An electrode insulated by a drop of electrolyte or an insulating film was placed around or next to a point contact over the n-type surface layer of p-type silicon in the first experiment and the p-type surface layer of a n-type Ge block in the second experiment. This structure is similar to the **GMST** structure described in Lilienfeld's second patent [2] which has a metal/p-Cu_2S rectifying junction and a Al_2O_3 dielectric film in place of the

Table 1 Milestones During the Discovery Phase of MOSFET (1926–1957)

Disclosure Dates[a] Subm.	Pub.	Authors-Inventors Development Team	Institutions or Locations[b]	Device and Technology	Reduction to Practice	Ref.
1926	1930	Lilienfeld	NYC	MESFET Al/Cu^2S	?	[1]
1928	1933	Lilienfeld	NYC	MOSFET Al/Al_2O_3/Cu_2S	?	[2]
1928	1933	Lilienfeld	NYC	n+-n & p+-p Ohmic Contacts	?	[2]
1928	1933	Lilienfeld	NYC	GMST Gated MS Junction Diode	?	[2]
1928	1932	Lilienfeld	NYC	SMST, MSMT, MpMpM	?	[3]
	1931	Wilson	GB	Holes and Semiconductor Theory		[9], [10]
1934	1935	Heil	GER	MOSFET Inversion Channel	?	[5]
	1938	Davydov	USSR	P-N Junction Minority Carrier		[15]
	1938	Hilsch, Pohl	GER	Analog Transistor Ionic		[15]
	1939	Davydov	USSR	M/S Rectifier Diffusion Theory		[15]
	1939	Schottky	GER	M/S Rectifier Diffusion Theory		[15], [16]
	1939	Mott	GB	M/S Rectifier Diffusion Theory		[15], [16]
1939	1976	Shockley	BTL	Analog Transistor Semiconductor		[6]
1939	1976	Shockley	BTL	MESFET Proposal		[6]
1940	1942	Scaff, Theuerer, Ohl	BTL	p-n Junction Fabricated		[15]
	1942	Bethe	MIT	M/S Rectifier Thermionic Theory		[14]
	1942	Bethe	MIT	Diffused M/S Rectifier Proposal		[14]
	1946	Serin, Stephens	MIT	Diffused M/S Rectifier	Lab	[14]
19MAR	1946	Bardeen	BTL	Surface State Theory		[6], [17]
24APR	1947	Shockley	BTL	Minority Carrier Injection Theory		[6]
04DEC	1947	Bardeen, Brattain	BTL	Sourceless MOSFET	Lab	[7]
1947	1948	Bardeen, Brattain	BTL	Point Contact Transistor	Lab	[13], [19]
18FEB	1948	Shive	BTL	BJT First Demonstrated	Lab	[21]
	1948	Shockley, Pearson	BTL	Conductivity Modulation	Lab	[18]
1948	1951	Shockley	BTL	BJT Diffused Base		[8]
1948	1951	Shockley	BTL	BJT Grown Junction		[8]
1948	1951	Shockley	BTL	BJT Integrated pnpnp		[8]
1948	1951	Shockley	BTL	High/Low n+-n & p+-p Contact		[8]
1948	1951	Shockley	BTL	Negative Resistance Diode		[8]
	1949	Shockley	BTL	BJT and Junction Theory		[22]
	1950	Teal, Little	BTL	Grown Ge Crystal	Lab	[15]
	1950	Hall, Dunlap	GE	Alloy-Diffusion p-n Junction	Lab	[35]
	1951	Scaff, Theuerer	BTL	Gas-Diffusion p-n Junction	Lab	[36]
	1951	Teal, Spark, Buehlor	BTL	BJT Grown GE Junction	Lab	[33]
	1951	Shockley	BTL	Hot Electron Theory		[56]
	1952	Teal, Buehler	BTL	Grown Si Crystal	Lab	[15]
	1952	Shockley, Read, Hall	BTL-GE	Recombination Kinetics Theory		[51], [52]
	1952	Theuerer, Pfann	BTL	Zone Refining of Material	Lab	[42]
	1952	Shockley	BTL	Junction FET Theory		[30]
1952	1953	Brattain, Bardeen, Shockley	BTL	Surface State Concepts	Lab	[23]
1953	1953	Brown	BTL	Surface Inversion Channel	Lab	[31]
	1953	Decay, Ross	BTL	Junction FET Built	Lab	[34]
1954	1958	Shockley	BTL	Ion Implanted Junction	Lab	[47]
	1954	Fuller, Ditzenberger	BTL	Gas-Diffusion p-n Junction	Lab	[37]
	1954	Kleinknecht, Seiler	Siemens	Generation Current in Si	Lab	[53]
	1955	Pell, Roe	GE	Generation Current in Ge	Lab	[54]
	1955	Shockley	SSL	Shockley Semiconductor Founded		
1954	1957	Frosch, Derrick	BTL	Oxide Diffusion Mask	Lab	[39]
1955	1957	Ross	BTL	MOSFET Ferroelectric Gate	Lab	[32]
	1956	Tanenbaum, Thomas	BTL	Diffused Diodes & BJT	Lab	[38]
1956	1957	Sah, Noyce, Shockley	STC	Recombination Current in Si	Lab	[55]
	1957	Noyce, Moore + six	FSC	Fairchild Semiconductor Founded		

[a]Subm. = Submission or first disclosure date. Pub. = Publication date.
[b]NYC = New York City; GER = Germany; BTL = Bell Telephone Laboratories; GE = General Electrical Research and Development Laboratory; STC = Shockley Transistor Corp.; SSL = Shockley Semiconductor Laboratory; FSC = Fairchild Semiconductor Lab.; MIT = M.I.T. Radiation Lab.; USSR = Soviet Union; GB = England.

electrolyte. A voltage amplification of 2 and power amplification of 330 was observed on December 8, 1947 [6] and current amplification of 1000 was stated in Shockley's transistor patent [8]. Shockley stated (see p. 613, second paragraph in the left column of [6]) that this was probably the only transistor that had amplified when his bipolar junction transistor patent was filed on June 26, 1948 [8], one week before the public announcement of the point-contact transistor [19], [20]. The point contact transistor has a different structure than the above Bardeen–Brattain surface field-effect transistor [13]. (See also Bardeen's Nobel lecture [7].) It consists of two closely-spaced point-contact electrodes made of phosphorus bronze wires or metal stripes from an evaporated gold film. These electrodes contact a p-type surface inversion layer on an n-type germanium, through a thin oxide film.

Earlier (around November 17, 1947), Brattain was also doing other surface state and conductivity modulation experiments with Bardeen using the two point contacts on an n-type Ge [6]. These experiments led to the invention and experimental proof of the bipolar junction transistor in the following three events: i) minority carrier injection and junction transistor structure proposed by Shockley on January 23, 1948 [6], [8]; ii) experimental demonstration of the bipolar transistor by Shive on February 18, 1948 [21], [6] using a double-surface Ge point-contact transistor which had the emitter and collector contacts on the opposite surfaces of a thin n-Ge slice to eliminate the surface inversion path of the hole current; and iii) Bardeen and Brattain's explanation of the bipolar point-contact transistor action in the patent they filed on June 17, 1948 [13], [19], [20].

The key features of the bipolar transistor invention were: i) exponential dependence of the injected minority carrier current upon emitter junction voltage, and ii) the high impedance of the collector junction [6].

B. Surface Inversion Layer and Channel

Modulation of the conductance of a surface inversion layer was suggested by Bardeen in his first transistor patent [7]. He stated that the modulation of the surface conductance can be obtained by reverse biasing either an insulated or a Schottky barrier gate over the surface of the surface inversion layer. In fact, Bardeen–Brattain's experiments [7] on the point-contact transistor [13] also gave the first working junction-gate field-effect transistor based on the modulation of the conductance of a surface inversion channel. As pointed out by Bardeen and Brattain in their 1948 analysis of the current-voltage characteristics of the point-contact transistor [19], [20], there are two 'minority' carrier (hole) current components in the emitter and collect leads of their point-contact transistor through two paths: i) the hole current flowing along the p-type surface inversion channel to the collector which is really a majority carrier current, and ii) the current from holes injected into the n-type Ge base (and hence a minority carrier current) by the forward biased p-n junction under the emitter point contact and these holes transverse laterally through the n-type base between the two point contacts and reach the collector. They stated that in their first experiments the majority carrier (hole) current in the p-type surface channel on the n-type Ge, i), probably dominated the current-voltage characteristics of the point-contact transistor, and not the minority carrier injection current along the longer path through the base, ii). The cross-sectional view of a surface channel junction-gate field-effect transistor appeared in Figures 10 and 11 of the Bardeen–Brattain 1948 point-contact transistor patent [13] and these two figures also gave the biasing circuits as an amplifier with power gain using the base as the input.

C. Surface States (Interface States) Fundamentals

A fundamental concept of surface states was introduced by Bardeen in 1947 [17] to explain the lack of metal-work-function-dependence of the metal/semiconductor or Schottky barrier diode as well as the lack of conductivity modulation of the field-effect experiments performed by Pearson and Shockley at BTL [17], [18]. In a later experimental paper by Brattain and Bardeen on germanium surfaces [23] they and Shockley introduced three important fundamental concepts of surface states which have been the foundation of all subsequent research on the surface properties of semiconductors and surface effects on transistor characteristics. These are: i) the surface recombination velocity; ii) a two-surface-state model having a donor level near the conduction band and an acceptor level near the valence band; and iii) the equivalent circuit model of resistors to represent the surface recombination losses at these surface states.

It is now known that the two-level surface state model, ii), corresponds to **extrinsic** traps due to impurities at the semiconductor surface. For the **intrinsic** traps due to the dangling bonds on the host atoms (Ge in the Brattain and Bardeen experiment) the sign of the charge of the two levels is just the opposite to that of the impurity surface states: the positively charged or neutral donor level is near the valence band edge and the negatively charged or neutral acceptor level is near the conduction band edge.

The third concept, iii), was extended by Shockley in 1958 who developed a comprehensive equilibrium equivalent circuit model for electron-hole recombination and generation at bulk and surface traps which included also the diffusion and drift currents [24]. Shockley's extension and influence-at-a-distance led me to complete the circuit model development to include the arbitrary nonequilibrium situation in 1961 [25]–[27]. The equivalent circuit model has not only provided a most familiar tool to teach the physics of electrons and holes in transistors to undergraduate electrical engineering students who thrive on circuits but has also given a systematic means to solve the Shockley semiconductor equations numerically to obtain accurate solutions of the highly nonlinear characteristics of the transistor [28].

D. Surface State (Interface State) Terminology

The terminology used in semiconductor surface research has been confusing probably because surface effects on transistor characteristics are complex, irreproducible, and unstable. Thus, the catch-all term 'surface states' has been used for many years to denote all the traps on and near the surface of a semicondutor covered by a fabricated (thermally grown or deposited) insulator or native oxide. A recent survey made by Bruce E. Deal led to a consensus on the notation and nomenclature to replace the catch-all term, surface states. It is now accepted and adhered to by most transistor researchers and integrated circuit engineers to call the surface states "interface traps" in transistor device and integrated circuit work. The term, interface traps, is most appropriate since no practical surfaces are bare. They are interfaces resulting from the coverage of the surface by a layer of native or fabricated oxide or other materials, such as an evaporated metal layer or an epitaxially grown semiconductor layer. The term "trap," as used by Bardeen, Brattain, and Shockley, is also the most appropriate to indicate that the electronic states under consideration are electronic bound states whose bound electron wave functions are localized in all three space dimensions at these trapping centers. Modifications, extensions, and concise definitions of these surface state phenomena occuring at today's oxidized VLSI silicon surface have been given by this author in a recent British IEE data review handbook, *Properties of Silicon* [29]. For example, the Deal committee also intro-

duced the term 'oxide fixed charge' as the second component of the old 'surface states' observed on oxidized silicon. This term has been discarded and replaced by 'oxide traps' and 'oxide charges' without the word 'fixed' since it has been well known that charges in the oxide are not fixed but can change by diffusion, drift, capture or emission of an electron or a hole, and by creation via bond breaking or destruction via recombination with hydrogen or with adjacent dangling bonds. In fact, the change of the density of charges trapped on the oxide traps is the very cause that degrades the performance and reduces the operating life of the transistors and hence the integrated circuit chip.

E. Shockley's Minority Carrier Injection and Field-Effect Transistors in the Bulk of a Semiconductor

Surface states, or the interface and oxide traps, in the point contact experiments and the unsuccessful conductivity modulation experiments hindered the reproducibility. Shockley invented the bipolar junction transistor (BJT) theory in 1949 which depends solely on minority carrier injection [22] and the junction-gate field-effect transistor (JGFET) and published the theory in 1952 [30]. Both of these transistors in Shockley's device structures have their electrically active regions away from the surface in the bulk of the semiconductor. Surface effects should not be important, which has largely been realized in modern silicon transistors. Residual deleterious surface effects still require careful transistor design and clean fabrication processing in today's VLSI circuit chips to minimize and eliminate surface related electrical instabilities. The two basic surface phenomena first described by Bardeen, Brattain, and Shockley during 1947 to 1951—the surface inversion channel and surface recombination—are the basic causes of degradation of the transistor characteristics. The temporal changes of the oxide charge density will change the size of the surface channel because it is induced or modulated by the oxide charge. The temporal changes of the oxide charge density will also change the rate of recombination of electrons and holes in the surface states because of the change of the induced electron and hole densities at the interface by the oxide charge. These temporal changes give rise to electrical instabilities and aging of the transistor characteristics and cause integrated circuits to fail.

The key concept for the bipolar junction transistor is **minority carrier injection over a potential barrier,** such as the potential barrier of the p-n junction or the surface inversion layer which gives the exponential dependence of the injected minority carrier current upon the emitter junction voltage. The second key feature is the high impedance of the collector junction. The mathematical theory was first developed by Shockley on April 24, 1947 for the reverse biased p-n junction [6]. It was then extended to forward bias when he developed the p-n junction transistor theory [22] on January 23, 1948 [6]. Experimental proof of minority carrier injection was first given by Shive's double-surface transistor [21]. Detailed studies of minority carrier injection were carried out by Haynes, Ryder, and Shockley [6] during 1948.

F. Tracing the Origin of the Modern Surface-Inversion Insulated-Gate Field-Effect Transistors

The n-type surface inversion layer on p-type Si and the p-type surface inversion layer on an n-type Ge were the principal reason for the successful operation of the field-effect point contact transistors invented by Bardeen and Brattain in 1947 [7]. The strong influence of the surface inversion channel on the bipolar transistor current-voltage characteristics was also noted by Bardeen and Brattain in 1947 [13]. This influence was further demonstrated by Walter L. Brown [31] in 1953 who gave experimental proofs of conduction in an n-type surface inversion channel across the surface of the p-type base layer of a Ge n-p-n bipolar transistor. Brown's conclusion is unequivocal in spite of the highly unstable surface conditions in his experiments. The instability was because there was no gate electrode and the surface channel was induced by surface charge from the ambient gas. Ian M. Ross was the first to describe the modern p-n junction or enhancement MOSFET structure in a 1957 patent disclosure using Brown's observations [32]. Instead of depending on the ambient to control the surface channel conductance as in Brown's experiment, Ross placed a ferroelectric crystal over the surface of the p-type base layer of the n-p-n transistor and covered the ferroelectric with a silver paste as the gate electrode. He then applied a voltage to the silver gate to control the polarization of the ferroelectric and the conductance of the n-type surface channel over the p-type base. The n-type channel over the p-type base extends from the n-type emitter to the n-type collector. Ross's intention was to use the ferroelectric polarization as a memory mechanism. But his transistor structure is precisely also the modern inversion-channel MOSFET, that is, it has the mandatory ingredient: two rectifying junctions one at each of the two ends of the surface inversion channel, which is also a structure given by Heil in 1935 [5] without knowing the surface inversion principle. The modern version of the MOSFET uses a thermally grown oxide film in place of a ferroelectric film for the gate insulator.

The evolution of the inversion channel MOSFET is illustrated in Fig. 4 using figures taken from the various patents and articles just cited. Fig. 4(a) is the one-rectifying-junction (sourceless or drainless) MOSFET proposed in Lilienfeld's second patent in 1928 [2]. Fig. 4(b) is the two-rectifying-junction surface inversion MOSFET described in Heil's 1935 patent [6]. Fig. 4(c) is Bardeen's single-rectifying-junction electrolyte-insulated-gate FET which gave the first recorded power gain in a solid-sate amplifier [6], [7]. Fig. 4(d) is Shockley's p-n junction diode with an insulated gate and it is the inversion channel MOSFET without the source (or drain) junction. Fig. 4 (e) is Brown's two junctions structure without a gate, used in his surface channel experiments [31]. Fig. 4(f) is Ross's inversion channel MOSFET using a ferroelectric gate insulator as memory site [32]. Fig. 4(g) is the modern inversion-channel MOSFET. It is evident that the chronological sequence depicted in this figure illustrates that new ideas and innovations can evolve directly or indirectly from previous knowledge. I was told that the transistor inventions at Bell Labs during 1947 and 1948 were not influenced by the patents of Lilienfeld and Heil. (See also Shockley's 1976 review [6] and Bardeen's Nobel lecture text and 1987 review [7].)

G. Transistor Technology

There were a number of laboratory demonstrations of new methods of transistor fabrication during 1950 to 1956. They contributed directly or indirectly but all very signif-

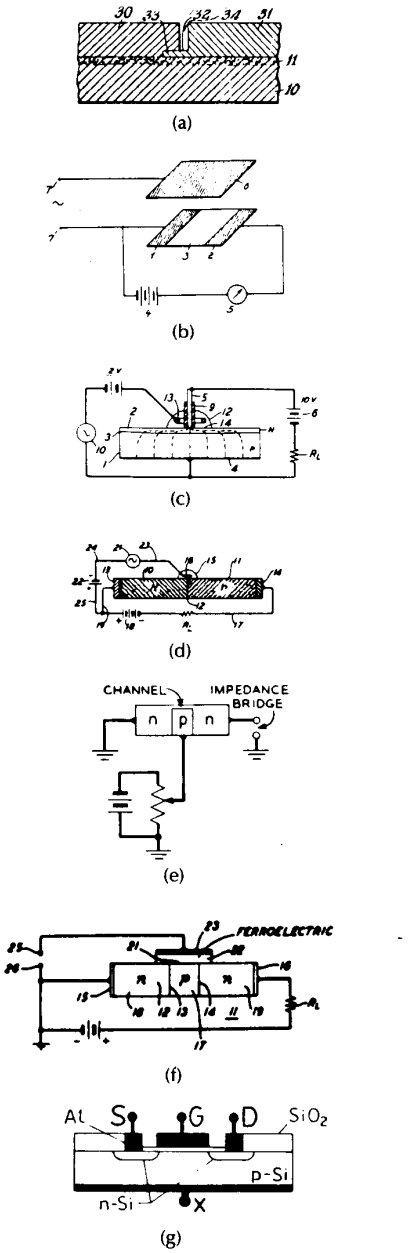


Fig. 4. An evolutionary chronology of the inversion-channel MOSFET. (a) Fig. 4 of Lilienfeld's second (1928) patent [2] showing a gated metal/p-Cu_2S junction. (b) Fig. 4 of Heil's 1935 patent [5] showing the inversion n-channel MOSFET which has the two **mandatory** rectifying junctions between the p-type semiconductor film (3) and the metal source and drain (1 and 2). (c) Fig. 3 of the first modern transistor patent (Bardeen's 1948 [7]) showing an insulated gate modulating an n-type surface layer in series with a reverse-biased bulk n-p junction. It is a sourceless MOSFET. (d) Fig. 1 of Shockley's 1948 transistor patent [8] showing a gated p-n junction which is a sourceless modern inversion-channel MOSFET. (e) Fig. 2(c) of Walter Brown's 1953 surface inversion channel experiment [31]. (f) Fig. 1 of Ian Ross's 1955 ferroelectric MOSFET [32] showing the **two mandatory rectifying junctions**. This must be credited as the first written disclosure of the modern surface-inversion-channel MOSFET. (g) The cross-sectional view of a modern Si surface-inversion-channel n-channel MOSFET.

icantly to the eventual successes in the realization of the silicon MOS transistor and integrated circuits. They are also listed in Table 1 and briefly discussed next.

The first BJT was built by Teal, Spark, and Buehlor in 1951 [33] from alternatively doping the germanium melt with n-type and p-type dopant impurities during the growth of a germanium single crystal using the single-crystal growth technique developed by Teal and Little [15]. The first JGFET in the semiconductor bulk was built by Dacey and Ross in 1953 [34]. Solid state diffusion to improve diode characteristics was suggested by Bethe in 1942 [14] and reported by Serin and Stephens in 1946 [14] in order to improve the uniformity of silicon crystal rectifiers for radar and microwave detectors. A very thin diffused Ge p-n junction layer was introduced by Hall and Dunlap in 1950 [35] using impurities from an alloy source. The gaseous diffusion technology used in the production of silicon integrated circuit today was invented by Scaff and Theuerer in 1951 [36] and further developed by Fuller and Ditzenberger in 1954 [37] which led to the fabrication of the first high frequency diffused bipolar junction transistors by Tannenbaum and Thomas in 1956 [38] and also the Ge p-n-p transistor by C. A. Lee *et al.* The technique of using an oxide mask against impurity penetration during impurity diffusion into the silicon wafer was introduced by Frosch and Derrick in 1957 [39], [40]. The first quantitative kinetic and process modeling of the oxide masking technique of Frosch and Derrick was given by Sah, Sello, and Tremere in 1958 [41] for oxide masking against phosphorus donor impurity diffusing into silicon. The masking data (oxide thickness required to mask or prevent phosphorous from penetrating into the silicon surface versus the temperature and time of exposure to the phosphorus vapor) obtained by Sah, Sello, and Tremere was used by Hoerni to produce the oxide stabilized planar silicon transistor (personal communication from Jean Hoerni, 1959) and these data and their extensions are used in today's VLSI production.

Other transistor fabrication technologies introduced or proposed during this period included the principle of zone refining (float-zone had been used to give high purity silicon) of semiconductor single crystal developed by Theuerer and Pfann in 1952 [42]–[46], the ion implantation technique proposed by Shockley in 1954 [47], [48], the first epitaxial transistor demonstrated by Theurerer, Kleimack, Loar, and Christenson in 1959 [49] which was directed by Ross [50], and simultaneous epitaxial Ge transistor work by Tannenbaum *et al.* These technologies took another decade to develop into effective production processes.

H. Transistor Material and Device Physics

Research on the effects of electronic traps or recombination centers on diode and transistor characteristics began during the 1950s. Electronic traps arise from atomic imperfections in the crystalline lattice of the semiconductor. Electrons and holes can recombine and be generated at high rates at the electronic traps. The studies have been based on the electron-hole recombination rate theory developed by Shockley and Read [51] and Hall [52] in 1952, known today as the Shockley-Read-Hall (**SRH**) or Hall-Shockley-Read (**HSR**) recombination statistics. ('Recombination **kinetics**' is the correct term.) Generation of electrons and holes by thermally exciting them out of the electronic traps in the

space charge layer of a p-n junction in the bulk of a semiconductor was shown to be an important mechanism of the reverse leakage current in silicon diodes by Kleinknecht and Seiler in 1954 [53] and in germanium by Pell and Roe in 1955 [54]. Recombination of electrons and holes in these traps in a forward biased silicon junction space charge layer was shown by Sah, Noyce, and Shockley [55] to control the silicon bipolar transistor gain at low currents and the switching voltage and holding current of the silicon p-n-p-n four-layer diodes and silicon controlled rectifiers. The generation-recombination current gradually discharges the stored charge on the capacitance and is the reason that the MOSFET dynamic random access memory (DRAM) cell has to be refreshed every millisecond.

Another transistor related basic concept of singular importance today is the theory of hot electrons advanced by Shockley in 1951 [56], [57]. It predicts the observed reduction of electron mobility at increasing electric field and the observed saturation of the electron drift velocity to about 10^7 cm/s at very high electric field. These effects are due to electron energy loss during collision with the vibrating silicon atoms resulting in the generation of optical phonons or silicon vibrations at infrared frequencies. These hot electron effects pose a fundamental limit on the speed of today's silicon as well as GaAs field-effect transistors. All the fundamental ideas and zeroth order theory of the hot electron effects are contained in this Shockley paper [56] but they have only been gradually rediscovered by the current generation of semiconductor researchers and engineers who are investigating transistor reliability and oxide failure in high electric fields such as the maximum dielectric breakdown field of the gate oxide of MOSFET which limits its ultimate performance. This will be discussed later.

I. Birth and Growth of the Silicon Valley

Toward the end of this first phase (1957), accelerated development and manufacturing of silicon transistors and integrated circuits began in the Silicon Valley which has grown to cover an area of about fifty miles in length extending from San Francisco south to Stanford and then to San Jose, now heavily populated by electronics and biotechnology related companies. The predecessor of the Silicon Valley was started by F. E. Terman, engineering dean at Stanford, when his students, Hewlett, Packard, and the Varian brothers formed the companies bearing their names around World War II. The rapid growth of the Silicon Valley began when Shockley left Bell Laboratories to start his own **silicon** transistor manufacturing company in 1955 and obtained the financial backing of fellow Caltech alumnus, Arnold Beckman. Shockley Semiconductor Laboratory began operation in the fall of 1955 as a division of the Beckman Instruments, Inc. in an army barracks (some people call it army hut) at 391 South San Antonio Road in Palo Alto. Twelve technical persons were personally interviewed and hired by Shockley: Horseley, Noyce, Moore, Grinich, Roberts, Hoerni, Last, Jones, Kleiner, Blank, Knapic, and Sah. It was incorporated in 1957 as the Shockley Transistor Corporation.

Eight of these, Noyce, Moore, Grinich, Roberts, Hoerni, Last, Kleiner, and Blank were then recruited by Fairchild Camera and Instruments Corp. to start the Fairchild Semiconductor Corporation with $2M when negotiation of the eight with A. O. Beckman to set up a separate $1M silicon transistor production facility faltered. The Fairchild laboratory and production line began operation at 844 Charleston Road in Palo Alto, about ten blocks east of the Shockley Laboratory, with the focused mission to mass-produce high performance silicon (bipolar) transistors. Among the many alternatives considered in a detailed 1957 plan of Noyce, the double diffused Si n-p-n mesa transistor was selected as the first product. It required only two photoresist steps to define the emitter and metal contacts.

Most of the early integrated circuit companies in the Silicon Valley were started or jointly started by Fairchild engineers, including four of the original eight, Noyce, Moore, Last, and Hoerni. One can say Shockley seeded the Silicon Valley and Fairchild was the first born that started its growth into the multi-billion dollar integrated circuit farm of today. A genealogy map has been prepared by the Semiconductor Equipment and Materials Institute in Mountain View, CA, and the first companies started by the Shockley line are listed in the map as follows. Last and Hoerni started Amelco in 1961, which was later renamed Teledyne Semiconductor. Hoerni went on to start Union Carbide Electronics in 1964, Intersil in 1967, and then moved to Europe to start several companies. Noyce and Moore started Intel in 1968. Other earlier and later Fairchild employees, who were not the original eight from Shockley Semiconductor, started more than twelve companies. These have since multiplied into slightly more than 100 companies by 1983, averaging about four startups per year during the twenty-five years since 1957 when Fairchild was founded. The entrepreneurship continues and fuels more startups. Most of the startups are not large companies, which reflects the opportunities created by the very many niche markets in integrated circuit applications based on the unique ideas of the founders of the startups. Most of the recent startups use silicon processing foundries to produce specialty integrated circuit chips. They have no in-house silicon wafer processing facility due to the very high cost of the processing equipment. The large number of so many new startups in the Silicon Valley also reflects the proximity of Stanford and (University of California at) Berkeley and their faculty's active involvement.

IV. Technology Development and Device Innovations (1958–1968)

Development of new silicon transistor fabrication technologies and invention of new transistor and circuit structures are highlights of the MOSFET evolution during the second phase which spans the decade from 1958 to 1968. These advances are tabulated in Table 2. It began when successful efforts to stabilize the surface of silicon p-n junction rectifiers and transistors were reported.

Stabilization of the MOSFET was the principle concern of researchers during 1958 to 1962. Headlines were made by a Berkeley professor during the 1961 **WESCON** (IRE-IEEE Western Convention and Show in August) and quoted by a San Francisco Chronical reporter who insisted that the MOSFET will never be practical due to electrical instabilities caused by surface states (interface and oxide traps). The bipolar junction transistor was predicted to continue to dominate. However, intensive research on stabilizing the surface of silicon rectifiers and transistors was going on at several laboratories, including Bell Telephone, IBM, and

Table 2 Basic Technology and Transistor Innovation Phase of MOSFET Evolution (1958–1968)

Disclosure Dates[a] Subm.	Pub.	Authors-Inventors Development Team	Institutions or Locations[b]	Device and/or Technology	Reduction to Practice	Ref.[c]
1958	1959	Atalla, Tannenbaum, Scheibner	BTL	Oxide Surface Passivation	Lab	[58]
1958	1959	Sah, Sello, Tremere	STC-FSC	Oxide Diffusion Mask, PSG	Lab	[41]
1959	1959	Moll	SEL	MOS Capacitor (MOSC Coined)	Lab	[59]
1960	1960	Hoerni	FSC	Si Planner BJT	Prod	[63]
1960	1960	Noyce	FSC	Si Monolithic IC	Prod	[65]
1960	1960	Kahng, Atalla	BTL	Si MOSFET	Lab	[70]
	1960	Ross, Team	BTL	Si Epitaxial Transistor	Lab	[49], [50]
1960	1961	Terman, Moll	SEL	MOS CV Technique	Lab	[60]
1961	1961	Sah	FSC	BIMOS	Lab	[79], [80]
1961	1961	Sah	FSC	Floating Gate Charge Storage	Lab	[85]
1961	1962	Sah	FSC	Surface Channel and Recombination	Lab	[81]
1962	1963	Wanlass, Sah, Moore	FSC	CMOS	Lab	[90], [92]
	1962	Engineering	FSC	FI-100 PMOS (MOSFET Coined)	Prod	FIG8
	1962	Engineering	RCA	3N98 NMOS (Depletion Mode)	Prod	FIG9
	1962	Engineering	FSC	5-Micron Channel	Lab	[PR]
1962	1963	Deal, Grove, Snow, Wanlass	FSC	Expanding MOS Research		[PR]
	1964	Young, Seraphim, Team	IBM	Expanding MOS Research		[88]
1964	1964	Kerr, Young	IBM	PSG Stabilization	Lab	[100]
1964	1965	Snow, Deal, Grove, Sah	FSC	Sodium Ion Drift Reduced	Lab	[99]
1963	1966	Heiman	RCA	Body Effect	Lab	[117]
	1965	Balk	IBM	H_2 Anneal of Interface Traps	Lab	[103], [104]
	1965	Kooi	Philips	H_2 Anneal of Interface Traps	Lab	[105]
	1965	Balk, Burkhardt, Gregor	IBM	Oxide Charge 100<100<111	Lab	[106]
	1965	Delord, Hoffman, Stringer	Reed	Oxide Charge 100<111	Lab	[107]
	1965	Miura	NEC	Oxide Charge 100<110<211<111	Lab	[108]
	1965	Snow, Deal	FSC	PSG Polarization	Lab	[102]
	1966	Mead	Caltech	MESFET Demonstrated	Lab	[98]
1964	1967	Miller, Barson	IBM	PSG Bulk Impurity Getter	Lab	[140]
1967	1967	Kahng	BTL	Floating Gate MOSFETs	Lab	[121]
1967	1967	Kahng, Sze	BTL	Floating Gate Demonstrated	Lab	[120]
1963	1967	Kerwin, Klein, Sarace	BTL	Poly Si Gate Self-Align	Lab	[111]
	1968	Sarace, Kerwin, Klein, Edwards	BTL	Silicon Nitride Oxide Mask	Lab	[112]
	1968	Bower, Dill	Hughs	Ion-Implant Self-Align	Lab	[113]
1967	1968	Dennard	IBM	One-Transistor DRAM Cell	Lab	[118]
	1969	Chou, Eldridge	IBM	PSG Without Polarization	Lab	[140], [141]
	1969	Hatzakis	IBM	Liftoff Metallization	Lab	[140], [141]
	1968	Noyce, Moore	INTEL	Intel Corp. Founded		

[a]Subm. = Submission date. Pub. = Publication date.
[b]BTL = Bell Telephone Labs. or Bell Labs. or AT&T Bell Labs.; SEL = Stanford Electronics Labs.; FSC = Fairchild Semiconductor Corporation; NEC = Nippon Electric Corp.; Reed = Reed College.
[c][PR] = Personal recollection or record.

Fairchild Semiconductor Laboratories. These projects were aimed at discovering the causes of the electrical instabilities and inventing practical ways to stabilize the silicon surface. The effort was motivated not only by the fabrication simplicity of MOSFETs that could result in volume production of inexpensive complex MOS integrated circuits with low power dissipation, but also by the need to stabilize the electrical characteristics of silicon bipolar transistors for operation at low currents or high voltages where performance deterioration (current gain and leakage current) due to surface instabilities had been severe.

A. Surface Stabilization by Thermally Grown Oxide Film on Silicon

The first successful attempt to stabilize the silicon surface was reported by a group led by Atalla (Atalla, Tannenbaum, Scheibner) at Bell Telephone Laboratories during 1958 [58], while early attempts were made by Frosch also at Bell Telephone Laboratories around 1954 (private communication from J. Bardeen, Oct. 1, 1988). On the surface where the silicon p-n junction intercepts, they grew a thin layer of silicon dioxide of 150 to 300 A in dry oxygen at 920°C for 10 to 30 minutes. They reported a ten to one hundred times reduction of the reverse leakage current of silicon n+-p diode rectifiers and a stabilization of leakage current in p-n-p-n diode switches. Low frequency $1/f$ noise, due to random trapping of electrons and holes at the interface traps, was also reduced tremendously in the oxide passivated diode rectifiers as indicated by a reduction of the noise corner frequency from 10 000 Hz to less than 10 Hz. Those of us active in silicon material and device research during 1956–1960 considered this successful effort by the Bell Labs group led by Atalla to stabilize the silicon surface the most important and significant technology advance, which blazed the trail that led to silicon integrated circuit technology developments in the second phase and volume production in the third phase.

B. Origin of the Acronym MOS

The acronym MOS (Metal-Oxide-Semiconductor or Metal-Oxide-Silicon) was coined by John Moll in 1959 [59] who moved to Stanford from Bell Telephone Laboratories

a year earlier. He gave a paper at WESCON on using the MOS capacitor (**MOSC**) as a voltage controlled variable capacitor. Fig. 5 shows Moll's experimental capacitance versus dc voltage (***CV***) curve (curve C) and his theoretical curve (curve *B*) based on a simple theory in which the electrons and holes are neglected in the surface space-charge layer, known as the carrier (or electron-hole) depletion theory.

C. The MOS Capacitance Voltage Diagnostic Technique for Surface States

The use of the MOS capacitance-voltage (**MOSCV**) characteristics to monitor and measure the density of surface states (interface plus oxide traps) was suggested by Moll [59] and demonstrated by Lewis M. Terman [60] during 1959–1960 while doing his doctoral research under Moll at Stanford. Terman measured the frequency dependences of the MOSCV curves from 100 Hz to 5 MHz [60], similar to the experimental curve taken at one frequency given by Moll (see Fig. 5). He then extrapolated the experimental *CV* curves

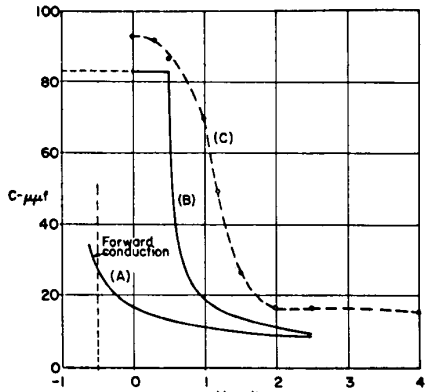


Fig. 5. The first experimental metal-oxide-silicon capacitance-voltage (dc) (MOSCV) characteristic given by John Moll and comparison with theory [59].

to zero frequency. From the difference between the high frequency and the low frequency *CV* curves, he computed the surface state density, which contained only the density of the interface traps and not the oxide traps. It also gave the variation of the density of the interface traps as a function of electron energy in the silicon energy gap over the whole range of frequencies or time constants. The extrapolated zero-frequency *CV* curve gives the density of surface states (interface traps) of all time constants up to 0.2 microseconds or 5 MHz. An extension of Terman's method has been widely used not only to monitor silicon VLSI production processes but also in research on the origin and properties of the surface states. This extension uses only the easy-to-measure high-frequency *CV* curve and compares it with the theory to give the total interface state density versus trapped electron energy. It is the most powerful electrical diagnostic method to monitor manufacturing reproducibility of silicon MOS integrated circuits. It has been known as the Terman Method (or Terman-Moll method) and the **HFCV** (High Frequency Capacitance Voltage) method. Other methods of measuring surface states have been investigated by later researchers but none can match the speed, accuracy, and range of the Terman-Moll method. Over a billion bytes of data on the charging-generation-annealing kinetic rates of surface states (oxide and interface traps) have been accumulated by Sah and his graduate students using the Terman-Moll method on state-of-the-art oxidized silicon fabrication under controlled and designed fabrication conditions [61], [62]. Additional data are still needed to form a database of wide range for the prediction of the reliability and operating life or endurance of silicon MOS and bipolar transistors and integrated circuit chips.

D. Invention and Production of the Oxide Passivated Silicon Transistors

The production implementation of the Atalla-Tannenbaum-Scheibner silicon transistor passivation technique by thermal oxide was first reported in 1960 by Jean Hoerni at the Fairchild Semiconductor Corporation [63]. Hoerni's oxide-protected silicon n+-p-n planar transistor was fabricated in the following way: He first diffused the entire surface of an n-type silcon wafer with boron at a high temperature (about 1200°C) using the gaseous diffusion method invented by Scaff and Theuerer in 1951 [36] and developed by Fuller and Ditzenberger in 1954 [37] to form the p-n collector junction. He then grew an oxide and etched an array of 15-mil diameter holes spaced 50-mils apart in the oxide (1 mil = 1/1000 inch = 25.4 micrometers). Using the data of oxide mask against the phosphorus diffusion developed by Sah, Sello and Tremere in 1959 [41], Hoerni diffused phosphorous into the oxide holes in order to obtain an array of diffused n^+ emitters of 15-mil diameter. (The + sign means heavily doped or high concentration of phosphorous donor impurity.)

The first silicon bipolar transistors in volume production by solid-state diffusion were the so-called mesa transistors, the n-p-n 2N696 [64], whose surface did not have a passivating thermal oxide. The term 'mesa transistor' was coined by J. M. Early (the discoverer of the Early Effect) of Bell Labs and was adopted since the finished transistor surface topology resembles a flat-top mesa found in the southwestern United States. The 2N696 was known as the double-diffused silicon mesa transistor since both the collector and emitter junctions were obtained by high-temperature solid-state diffusion of impurities, boron and phosphorus respectively, into silicon using Hoerni's procedure just described. The emitter junction was a 15-mil diameter circular dot defined by a hole chemically etched into the oxide film using photoresist masking. The oxide film over the entire wafer surface was chemically etched off after the emitter diffusion. Thus, the emitter junction intercepted a nearly bare silicon top surface covered only by a few monolayers of residual oxide left by the chemical etch. The area of the collector junction was defined by chemically etching a 30-mil diameter silicon mesa centered at the emitter dot. The collector junction then intercepted the bare side-wall of the etched silicon mesa. The bare surface of the perimeters of the emitter and collector junctions was the very cause of the high collector junction leakage current, low and unstable current gain at low emitter current, and unstable

breakdown voltages at the emitter and collector junctions in the 2N696 transistor. Some stabilization was achieved in Fairchild production by a short (about an hour) high temperature (300C) bake-out, which resulted in a very thin protective oxide layer covering the junction perimeters on the surface.

Using additional photolithography steps, Hoerni [63] replaced the mesa etched collector by a 30-mil diffused tub. The diffused tub was first used at Bell Labs. Furthermore, he left a thick (about 1 micrometer) high-temperature (1000°C) thermally grown oxide on the silicon surface over the perimeter of both the emitter and collector tubs where the emitter and collector junctions intercept the surface. These oxide covered surface-intercepts of the emitter and collector surface junctions greatly improve the transistor current gain at low currents as well as the leakage current and breakdown voltage of the emitter and collector junctions. These improvements were due to the reduction and stabilization of the surface states (interface and oxide traps) at the surface perimeter of the emitter and collector junctions by the thermally grown oxide.

This transistor passivation and stabilization technology was termed by Hoerni as the **planar** process [63]. The two planar silicon BJTs 2N613 (n-p-n) and 2N869 (p-n-p) were produced and delivered in volume a year later in 1961 [64].

E. Invention of the Monolithic Integrated Circuit Technology

During the time Hoerni was perfecting the planar oxide-passivation techniques, Noyce invented the monolithic integrated circuit concept at Fairchild [65] and used the planar technique to fabricate the first monolithic silicon integrated circuits. In the monolithic integrated circuit, planar diffused silicon bipolar transistors and resistors are interconnected by thin and narrow aluminum lines over the passivation oxide. The aluminum interconnect lines are fabricated by etching an evaporated aluminum layer over the entire oxide surface using the photolithographic technique. This has been known as the monolithic integrated circuit technology.

In contrast, Kilby at Texas Instruments made a flip-flop from a single chip of Ge. Gold wires were used for intraconnections . This is known today as the hybrid integrated circuit technology. According to a personal account given by Kilby in a review [66], a U.S. Appeal Court had ruled Noyce the inventor of the monolithic technique using adherent oxide, junction isolation and deposited and photographically etched aluminum film interconnection lines adherent to the oxide. But it is obvious that Kilby's flip-flop is a predecessor which could have led to the monolithic integrated circuit. It took the insight of Noyce to conceive the monolithic integrated circuit which, like many other inventions and discoveries, is directly or indirectly built on previous knowledge.

A family of monolithic transistor-transistor (direct coupled transistor logic or **DCTL**) logic elements containing four or more silicon BJTs on one silicon chip was introduced by Fairchild Semiconductor in February 1960 [67], [68]. It was called the "Micrologics."

The planar and the monolithic technologies of Hoerni and Noyce laid the foundation for the development of integrated circuits, using BJTs initially, MOSFETs during the last two decades (1965–1985), and more recently a mixture of both, known as the **BIMOS** and BICMOS (to be discussed in a later section). The short time lag (a few months) during 1959–1960 between the conception and the volume production of silicon transistors and integrated circuits using the planar and monolithic technologies was an example of the dexterity of its inventors and developers. I still vividly recall watching Jean Hoerni in 1959 pushing many 1-inch and 1.5-inch diameter silicon wafers into the oxidation and diffusion furnaces himself to find the optimum boron and phosphorus diffusion depths and oxide masking conditions, Bob Noyce coating and exposing photoresist on many oxidized and aluminized silicon wafers to find the optimum oxide and aluminum etching conditions in the dark room on evenings and weekends, Gordon Moore personally managing the first planar silicon transistor and integrated circuit manufacturing line of thirty or so women, Vic (Victor H.) Grinich connecting and running the instruments to characterize the transistors and integrated circuits, all next to my laboratory workbench where I was measuring the surface-controlled or gated diodes and transistors, at Fairchild's first home, 844 Charleston Road in Palo Alto. That was the birth of volume produced, oxide protected and stabilized silicon transistors and integrated circuits; when there was no precedence or existence proof, the best and fastest route to reduction to practice was the do-it-yourself approach taken by these four pioneers—Grinich, Hoerni, Moore, and Noyce.

F. A 1960 Supercomputer Design and a Missile Using Silicon Planar Transistors

Two of the first silicon bipolar n-p-n transistor products should go into the historical record book, not only for the enormous profits they generated which enabled Fairchild Semiconductor Laboratory to greatly increase its research and development efforts that led to the rapid introduction of whole families of volume produced silicon transistors and integrated circuits, but also for setting the pace on computer system designs based on the availability of certain superior transistor performance characteristics, such as speed and especially reliability.

The origin of the first product was the gold-doped high-speed (16 ns) switching n-p-n transistor, 2N706. It was a smaller mesa (three-times smaller diameter at 5-mil or an area of 1.2×10^{-4} cm^2) and higher speed version of the 2N696 bipolar silicon n-p-n discussed in Section IV-D which had been marketed by Fairchild in 1960. Gold is a highly efficient recombination center for electrons and holes. In order to increase the switching speed, gold was diffused into the transistor to reduce the minority carrier lifetime and thus the charge storage time in the base and collector layers of the 2N706. Based on this existence proof, Control Data Corporation awarded Fairchild Semiconductor Laboratory a $500 000 development contract to produce a still higher speed silicon transistor switch to meet the first requirement—the high switching speed (less than three nanoseconds) of the 10-MHz (3**MIPS**) CDC-6600 scientific computer [69]. The second requirement was reliability since there were 600 000 transistors in the **CPU**. That contract was followed up by a $5M production contract for 10 million units of high speed, gold-diffused, transistors and 2.5 million units of high speed, gold-diffused, diodes in September 1964. In fact, the transistor specifications of 3-ns and

high reliability were arrived at by the CDC computer designers based on the required speed and reliability to complete a numerical solution of a scientific problem without interruption from a computer hardware failure [69]. In order to achieve several thousand hours of CPU run-time without failure, high reliability from the individual silicon planar transistors was the most critical consideration owing to the large number of transistors (600 000) used in the CPU of the CDC-6600. Noyce's monolithic technology has greatly improved the numerics of reliability today. For example, the 600 000 transistors in CDC-6600 is only about one-half of the number of transistors contained in a 1-mega-bit (Mbit) DRAM chip which has a projected chip operating life of 10 years and as many as nine or more 1-Mbit chips can be used in a single personal computer today which rarely experiences MOS memory failures and whose failures are usually due to the crash of the mechanical magnetic disk drive. To meet both the 3-ns and high-reliability specifications, Fairchild engineers shrunk the circular 16-ns mesa 2N706 transistor down to a three-finger stripe geometry and used oxide passivation for stabilization. They also improved the yield by using an epitaxial layer to control the resistivity. The result was the 2N709 which met the 3-ns switching time and high reliability requirements. It gave a 2000 CPU-hour operating time before a transistor fails. This was a very large development and production contract for the design and delivery of only one transistor type—by comparison, it took only about $250 000 to start a silicon transistor manufacturing company in 1960. High speed and high reliability of the 2N709 met the critical requirements that made the first scientific computer possible.

The second volume-produced silicon transistor of historical significance was the planarized (oxide passivated and stabilized) version of the 2N696 [64]. Its high temperature operation capability (compared with germanium) and high reliability met the principal requirements of the Minuteman missile electronics program. This high reliability part was sampled at $250 each and delivered in volume at $100 per transistor in 1961 at a manufacturing cost of less than 50 cents. The profit helped to fund the expanding technology development efforts at Fairchild during the early 1960s which rapidly advanced the first generation silicon integrated circuit technology.

G. The First Laboratory Demonstration of the Modern Silicon MOSFET

A major development in the evolution of the silicon MOSFET was reported in June 1960 by Kahng and Atalla at the **SSDRC** (the Solid State Device Research Conference sponsored by the IEEE Electron Device Society) [70]. They described the first successful operation and the pentode-like characteristics of the inversion-channel or enhancement-mode silicon field-effect transistor using a thermally grown oxide for the gate insulator over the n-type surface inversion channel between two n+-p junctions on a p-type silicon substrate [70]–[75].

The historical precedents to the Kahng-Atalla MOSFET structure are traced in Fig. 4: Lilienfeld's gated junction of 1928 [2]; Heil's inversion channel MOSFET between two Schottky or metal/semiconductor junctions of 1935 [5] (which one must concede as the very first public disclosure of the inversion channel MOSFET structure although Heil did not recognize the surface inversion channel nor stated this concept in his patent); Bardeen's [7] and Shockley's [6] 1947 gated p-n junction, with the former giving the first historically recorded power gain [6]; Bardeen and Brattain's 1947 realization that their point-contact transistor would also work as a FET in the presence of a surface inversion channel (p-channel on a piece of n-Ge) [19], [20] and Shockley's further account of the FET current component in the point-contact transistor in 1976 [6]; Shockley's 1952 theory of current saturation owing to a pinched-off channel in the junction-gate field-effect transistor [30] which led Walter Brown, who was working under Shockley in 1953, to experimentally demonstrate the presence of a surface inversion channel over the surface of the p-type base of a Ge n-p-n transistor which provided the surface channel conduction path connecting the n-type emitter and n-type collector regions [31]; Ian Rose's 1955 surface-inversion MOSFET patent [32] which employed a ferroelectric gate dielectric over the surface of the base layer of a n-p-n or p-n-p transistor and which must be recognized as the invention of the (modern) inversion-channel MOSFET with **two rectifying junctions** that electrically isolate the source and drain regions; Hoerni's planar structure of an oxide over the surface intercept of a diffused silicon p-n junction; and Noyce's aluminum over the oxide covered p-n junction intercept. With all of these precedents, Kahng and Atalla's patents were restricted to other new modes of operation and did not contain claims of the pentode-like transistor operation, namely, the channel current saturation predicted by Shockley's channel pinch-off theory [30] which was employed by Brown [31] to describe the experimental current when one of two rectifying junctions is reverse biased. It seems that Lilienfeld's, Heil's, and the other precedents had so completely described both the transistor structure and its operating principle that the United States Patent Office frequently states: "all the subsequent works are 'obvious to one familiar with the art.'" This hindrance to patent claims is further reconfirmed recently by Bardeen's recollection [76]. Nonetheless, to those of us in the field, Kahng and Atalla are the recognized inventors of the modern MOSFET.

H. Some Basic Experiments on Surface Effects

During this time in 1960, I made my first contact with MOS transistors. Having written the article on the recombination current in the bulk junction-space-charge-layer with Noyce and Shockley in 1957 [55], which had provoked a question on surface currents from a Bell Lab colleague during my talk at the March Meeting of the American Physical Society in 1956 at Monterey, CA, I embarked on a series of experiments to delineate and understand the effects of surface states, surface recombination and surface channel on the current-voltage characteristics of silicon p-n junction diode and bipolar junction transistors. This pursuit was greatly aided by Hoerni's planar technology which enabled me to build a gated p-n junction using thermally grown oxide to control the surface effects. An aluminum gate electrode was photo-lithographically etched onto the oxide that covers a p-n junction diode or the emitter junction of a bipolar transistor [77]–[81]. With the help of my long-time assistant, Douglas Tremere, two bipolar transistor structures were fabricated in 1961. One has a MOS gate on the emitter junction (Fig. 6(a)) [77], [78]. The other has an additional diffused

(a)

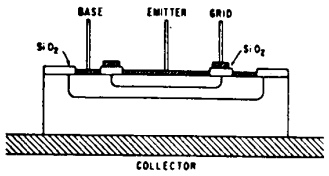
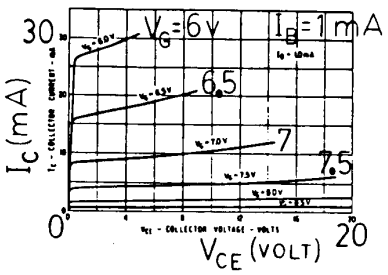


Figure 1—Cross-sectional view of tetrode. The grid electrode is shown on top of the SiO_2 which covers the emitter base junction.

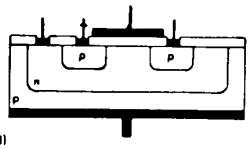
(Below)

Figure 2—Collector family for the *SCT*.

(b)



(c)



(d)

Fig. 6. (a) The cross-sectional view [77] and (b) the output characteristics [82] of the silicon surface potential controlled tetrode (SCT). (c) The first BIMOS [79] built and used to give the high gain reported in (b) and shown in [77] and [82].

drain junction covered by the gate (Fig. 6(c)) [79], [80]. I named them the Surface Controlled Tetrode (**SCT**) [77] and made extensive current-voltage measurements [81] to demonstrate and simulate the importance of surface recombination and surface channels in reducing and changing the leakage current and the current gain of bipolar transistors. We, at Fairchild, also thought that this might be a very useful transistor because of its high-gain and high-input impedance at the gate electrode [82] and received many inquiries, some about product delivery from potential startups who wanted to replace the high input impedance electrometer tube by the surface controlled tetrodes. But lack of stability of the oxide (known today as ion migration due to the sodium ion drift in the oxide; see a later part of this review) and the high gate voltage required due to the thick gate oxide (2000 A) used kept it in the laboratory.

Three new device ideas evolved from this gated BJT or SCT. The first is the inclusion of a second diffused p-type region in the surface of the n-type base under the gate to serve as the drain so that the output (between the source and drain) of a MOSFET is connected across the emitter-base junction. This was not disclosed in the 1961 journal article on the SCT transistor [77] due to impending patent applications [78], [80] but a facsimile of its cross-sectional view was revealed in the *1963 McGraw-Hill Yearbook of Science and Technology* (reproduced in Fig. 6(c)) [79] and also in my second SCT transistor patent [80]. This is the basic unit of today's BIMOS and BICMOS (bipolar-MOS) integrated circuits with cells having a vertical p-n-p BJT integrated with a p-channel MOSFET and a vertical n-p-n BJT integrated with an n-channel MOSFET. The second idea is the addition of another diffused bulk junction to give an MOS-gate-controlled p-n-p-n switch which has a very high input impedance MOS gate (see U.S. Patents listed in [78], [80]). This is known today as the MOS-**SCR** (Silicon Controlled Rectifier) and it has been extensively developed by General Electric engineers [83], [84] for household appliance (such as light dimmer), system, and power control applications. A third idea concerned storing charge on the gate (or grid as it was called then) when the metal gate on the oxide over the emitter junction is disconnected or floating [77], [85]. In my experiment [85], the underlying MOSFET current-voltage characteristics were controlled by the amount of charge stored on the floating gate and the characteristics or channel conductance were calibrated so that the voltage on the floating-gate can be calculated from the shift of the transistor characteristics or channel conductance as a function time. Knowing the gate capacitance, the measured time constant then gives the leakage resistance through the gate oxide [85]. This observation and measurement were the precedents of the subsequent development of the floating gate nonvolatile MOS memory (more in Sections IV-J and IV-P).

The gated BJT transistor structures of Fig. 6(a) and (c) had given a wealth of information about the surface effects on diode and transistor characteristics. The vast amount of new data and the insights derived had overwhelmed us in 1961 that little attention was paid to the results of my floating gate experiments. Instead, the time dependent characteristics due to charging the floating gate and its surrounding oxide was looked upon as a cause of the undesirable electric instability. Although surface effects are minimized, stabilized, and under exceptional control in today's silicon VLSI factories, the atomic origin and chemical structures of the surface states (or oxide and interface traps as they are called today) are still elusive and are occupying an increasing number of the best scientific research minds over the world (in the hundreds as estimated from the attendee counts of the IEEE Interface Specialists Conference and the Gordon Research Conferences). The importance of surface effects has begun to re-emerge again in recent years as the transistor size shrinks to below one micron, causing the electric field in the oxide to increase and forcing energetic or hot electrons to be injected into the oxide. (See [86] and also the comprehensive reviews in twenty sections by this author with his graduate students in a data-review handbook titled *Properties of Silicon* which was published by the British IEE-INSPEC and IEEE press in May 1988 [87].)

I. Two Management Decisions on MOS Research and Development

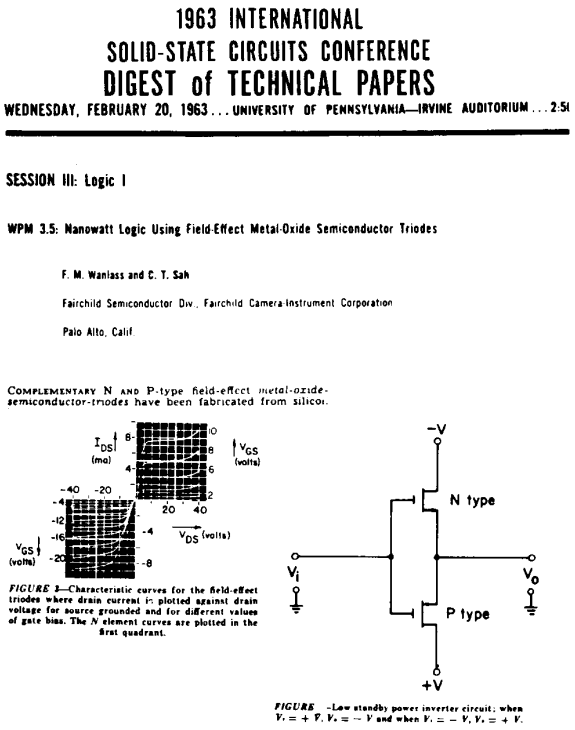
Two major management decisions occurred during 1961 and 1962 which shaped the future of the silicon MOS transistor and silicon MOS integrated circuits. At the Fairchild Semiconductor Research and Development Laboratory, a management decision was made in 1961 to greatly expand its silicon material and device research efforts. I was asked by Gordon Moore, Bob Noyce, and Vic Grinich to take charge of this effort while I was taking a one-semester leave at the University of Illinois in Urbana to teach Bardeen's semiconductor physics course, which had given me my first

exposure to transistor physics eight years earlier in the Spring of 1953. I commuted from Urbana to Palo Alto monthly on a DC-3 to interview applicants. But with Fairchild's fabulous successes on the silicon planar transistors during those days, it was easy for us to attract and hire Frank Wanlass, Andy Grove, and Ed Snow (in that order), all fresh out of graduate school. Wanlass and Snow were both solid state physicists from Utah while Grove was a chemical engineer from Berkeley. In order to complement these bright and young doctorate upstarts, we were also able to attract a veteran chemist, Bruce Deal. Two MOS projects were started, one with Deal, Grove, and Snow on material and device physics and the other with Wanlass on circuit applications (such as the CMOS). Some device phenomenon observations, device structure innovations, device principle discoveries, process chemistry delineations, and novel circuit designs made by these four are still used in today's silicon VLSI technology and even dominate the current technology. Some of these will be summarized later.

Another major management decision occurred 3000 miles away in New York at about the same time (1963). A plan was implemented by the IBM management to develop the n-channel FET for use as the main (CPU) memory in their next generation mainframe computers. Its success led to the delivery of the mainframe computer, IBM-370/158, in 1973 using MOS memory. According to Donald R. Young and Donald P. Seraphim in their introductory remarks of a special issue of the *IBM Journal of Research and Development* published in September 1964 [88], IBM's initial MOS efforts were stimulated by my detailed 1961 studies of the effects of surface recombination and channel on the Si transistor characteristics [77] and earlier work by Shockley and Pearson [18], Bardeen [17], Brown [31], and McWhorter, Kingston, Cutler, Bath, Garrett, and Brattain [88]. IBM's efforts involved about fifteen veteran engineers and material scientists managed by Young at the IBM Components Division at Poughkeepsie and about the same number managed by Seraphim at the IBM Research Center at Yorktown Heights. (Personal communication from D. R. Young, May 3, 1987.) Many of today's basic MOS technologies and circuit structures were developed by these and other IBM groups. Most of these MOS technologies are also used in BJT and bipolar **IC** productions. They will be described in the following sections. At IBM, the adopted acronym for the MOS field-effect transistor was just FET without the MOS. The 'camouflage' confused me when I heard it later in my first visit to IBM-East-Fishkill as a consultant to D. R. Young during 1968–1973. It was not camouflage—IBM always had its own acronyms and symbols. (Personal communication from Lewis M. Terman, June 1, 1988.) The FET shorthand often required clarification since there are many types of FETs, such as the junction gate FET or JGFET, Schottky barrier or metal gate FET or MESFET, the thin-film FET or **TFT** or **TFFET**, and others, although none has made an impact in today's integrated circuits compared with the MOSFET.

J. The Conception of the CMOS Circuit and the Floating Gate Memory Idea

Wanlass was a most inventive young circuit designer although his doctoral research was in theoretical solid state physics [89] under the famous quantum chemist, Henry Eyring. His experience came from owning a circuit engineering company in the late 1950s while still a graduate student at Utah. Within three months after arriving at Fairchild in the summer of 1962, he conceived two major circuit and application ideas. One is the CMOS (Complementary MOS) inverter circuit first reported at the 1963 **ISSCC** (International Solid State Circuit Conference) [90]–[93]. Two figures from the extended abstract of the conference proceeding [90] are combined and shown in Fig. 7. (See also a popular account on Wanlass [94].)


1963 INTERNATIONAL
SOLID-STATE CIRCUITS CONFERENCE
DIGEST of TECHNICAL PAPERS
WEDNESDAY, FEBRUARY 20, 1963 ... UNIVERSITY OF PENNSYLVANIA—IRVINE AUDITORIUM ... 2:5

SESSION III: Logic I

WPM 3.5: Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes

F. M. Wanlass and C. T. Sah

Fairchild Semiconductor Div., Fairchild Camera-Instrument Corporation

Palo Alto, Calif

COMPLEMENTARY N AND P-type field-effect metal-oxide-semiconductor-triodes have been fabricated from silicon.

FIGURE 3—Characteristic curves for the field-effect triodes where drain current is plotted against drain voltage for source grounded and for different values of gate bias. The *N* element curves are plotted in the first quadrant.

FIGURE —Low standby power inverter circuit; when $V_i = +V$, $V_o = -V$ and when $V_i = -V$, $V_o = +V$.

Fig. 7. Characteristics and circuit of the complementary metal-oxide-silicon (CMOS) transistor circuit from the first CMOS paper by Wanlass and Sah [90].

The other idea was on the use of the floating-gate MOSFET as a memory device. The memory idea was not generally known outside of the Fairchild laboratory but recorded in Fairchild's monthly reports. The idea came from the observation reported by Sah a year earlier that a charge can be stored for a long time (several days) on the gate electrode of the MOS-gated BJT or the SCT transistor [85] and that charges can also be injected onto the floating gate by a high voltage applied to the underlying source or drain junction. This stored charge on the floating gate was used to measure the leakage resistance of the gate oxide of the transistors shown in Fig. 6 [85]. The stored charge shifted the current-voltage characteristics in Fig. 6. Victor H. Grinich, the associate research director at the Fairchild Research and Development Laboratory, and then Wanlass recognized immediately its potential as a floating gate memory device. Electrical instability in MOS transistors so occupied the Fairchild managers and engineers at the time that this idea was not pursued until almost ten years later when Frohman-Bentchkowsky from Fairchild joined Intel and invented a most ingenious charge injection method and a transistor

structure to make the floating gate transistor a successful commercial product. (More in Section IV-P and also in IV-Q, the latter describes a new competitor, the nonvolatile ferroelectric MOS DRAM and **SRAM** cells, which can potentially replace the volatile **NMOS** or NMOS-CMOS DRAMs as well as the CMOS SRAMs to incite another revolution in computer technology.)

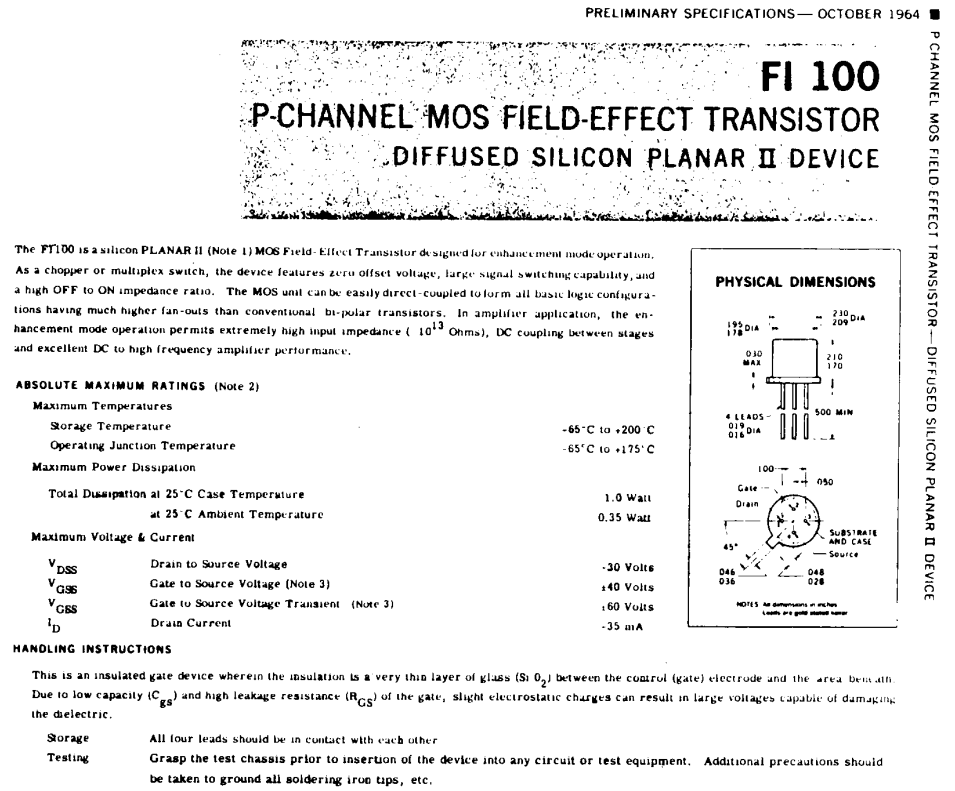
K. First Commercial MOSFETs

The first two commercial MOSFETs were announced in late 1964, one by Fairchild and a second by RCA. Fairchild announced the production of an enhancement-mode p-channel silicon MOSFET, FI-100, whose data sheet is given in Fig. 8. It employed the diffused planar-II process. It was designed for switching and logic applications. One month later, RCA announced a depletion-mode n-channel MOSFET for small-signal applications shown in Fig. 9. RCA then followed up with a tetrode version which has a second gate electrode placed between the first gate electrode and the drain electrode. This dual-gate MOS tetrode has been the popular front-end amplifying transistor for stereo high-fidelity **FM** tuners first introduced by Scott, a leading and domestic FM stereo receiver and high-fidelity equipment manufacturer of that time. The selection of the dual-gate MOS tetrode was due to its immunity to overload by strong signals and its high stable gain from the large reduction of the Miller or feedback capacitance by RF grounding the second gate near the drain. This also influenced my choice of the first stereo FM receiver I owned. The dual-gate MOSFET tetrode is another idea borrowed from precedents, the vacuum tube tetrode and pentode, which have one and two screen grids to capacitively shield the input grid from the output plate.

L. The Acronym Race

There were many names and acronyms proposed and used for the metal-oxide-silicon field-effect transistors. Eight are known to this author. They give an amusing glimpse of the intensity of competition not only at the marketplace but also in engineering and scientific disclosures, within one company and among companies.

Fairchild's October 1964 data sheet (Fig. 8) called it the MOS Field-Effect Transistor. Fairchild's November 1964 application bulletin (Fig. 10) called it MOS FET with a space, sometimes half-space, between the two groups of letters. Fairchild's October 1964 **AM** radio circuit application paper (Fig. 11) called it **MOST**, an acronym coined by Sah and appeared in the draft version of the Wanlass-Sah ISSCC CMOS paper issued as a Fairchild technical report [91]. It seemed that Fairchild engineers could not make up their minds. Wanlass and I knew that MOST is the "most," but our marketing department was not convinced and preferred other names or acronyms. When our boss Gordon Moore heard about it, he exclaimed, "MOST—what a name!," which probably eliminated its appearance in the title of our CMOS paper [95] in the 1963 ISSCC Digest of Technical Papers [90]. However, Sah was persistent and used it in his next one-author MOSFET paper in 1964 [96].



PRELIMINARY SPECIFICATIONS— OCTOBER 1964

P CHANNEL MOS FIELD EFFECT TRANSISTOR— DIFFUSED SILICON PLANAR II DEVICE

FI 100

P-CHANNEL MOS FIELD-EFFECT TRANSISTOR

DIFFUSED SILICON PLANAR II DEVICE

The FT100 is a silicon PLANAR II (Note 1) MOS Field-Effect Transistor designed for enhancement mode operation. As a chopper or multiplex switch, the device features zero offset voltage, large signal switching capability, and a high OFF to ON impedance ratio. The MOS unit can be easily direct-coupled to form all basic logic configurations having much higher fan-outs than conventional bi-polar transistors. In amplifier application, the enhancement mode operation permits extremely high input impedance (10^{13} Ohms), DC coupling between stages and excellent DC to high frequency amplifier performance.

PHYSICAL DIMENSIONS

ABSOLUTE MAXIMUM RATINGS (Note 2)

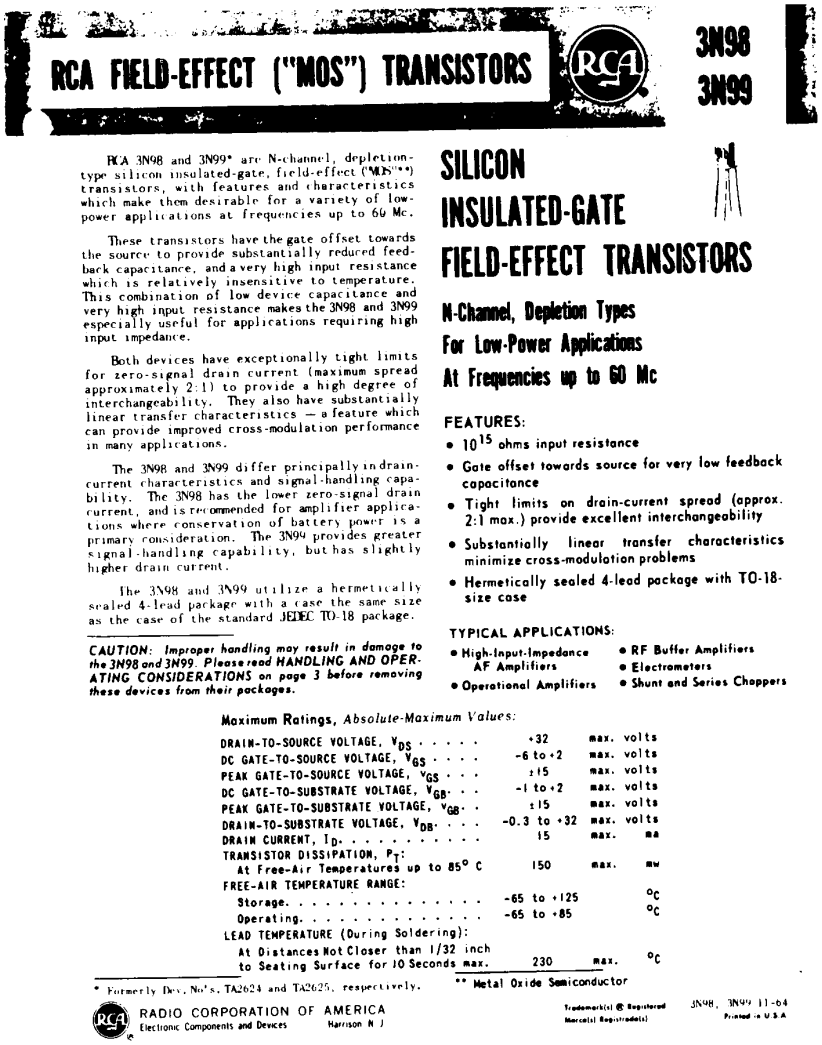
Maximum Temperatures		
Storage Temperature		-65°C to +200°C
Operating Junction Temperature		-65°C to +175°C
Maximum Power Dissipation		
Total Dissipation at 25°C Case Temperature		1.0 Watt
at 25°C Ambient Temperature		0.35 Watt
Maximum Voltage & Current		
V_{DSS}	Drain to Source Voltage	-30 Volts
V_{GSS}	Gate to Source Voltage (Note 3)	±40 Volts
V_{GSS}	Gate to Source Voltage Transient (Note 3)	±60 Volts
I_D	Drain Current	-35 mA

HANDLING INSTRUCTIONS

This is an insulated gate device wherein the insulation is a very thin layer of glass ($Si\,0_2$) between the control (gate) electrode and the area beneath. Due to low capacity (C_{gs}) and high leakage resistance (R_{GS}) of the gate, slight electrostatic charges can result in large voltages capable of damaging the dielectric.

Storage	All four leads should be in contact with each other
Testing	Grasp the test chassis prior to insertion of the device into any circuit or test equipment. Additional precautions should be taken to ground all soldering iron tips, etc.

Fig. 8. The Fairchild product data sheet of the first volume-produced silicon MOSFET, a p-channel enhancement-mode transistor.



RCA FIELD-EFFECT ("MOS") TRANSISTORS

3N98
3N99

RCA 3N98 and 3N99* are N-channel, depletion-type silicon insulated-gate, field-effect ("MOS"**) transistors, with features and characteristics which make them desirable for a variety of low-power applications at frequencies up to 60 Mc.

These transistors have the gate offset towards the source to provide substantially reduced feedback capacitance, and a very high input resistance which is relatively insensitive to temperature. This combination of low device capacitance and very high input resistance makes the 3N98 and 3N99 especially useful for applications requiring high input impedance.

Both devices have exceptionally tight limits for zero-signal drain current (maximum spread approximately 2:1) to provide a high degree of interchangeability. They also have substantially linear transfer characteristics — a feature which can provide improved cross-modulation performance in many applications.

The 3N98 and 3N99 differ principally in drain-current characteristics and signal-handling capability. The 3N98 has the lower zero-signal drain current, and is recommended for amplifier applications where conservation of battery power is a primary consideration. The 3N99 provides greater signal-handling capability, but has slightly higher drain current.

The 3N98 and 3N99 utilize a hermetically sealed 4-lead package with a case the same size as the case of the standard JEDEC TO-18 package.

CAUTION: Improper handling may result in damage to the 3N98 and 3N99. Please read HANDLING AND OPERATING CONSIDERATIONS on page 3 before removing these devices from their packages.

SILICON
INSULATED-GATE
FIELD-EFFECT TRANSISTORS

N-Channel, Depletion Types
For Low-Power Applications
At Frequencies up to 60 Mc

FEATURES:

- 10^{15} ohms input resistance
- Gate offset towards source for very low feedback capacitance
- Tight limits on drain-current spread (approx. 2:1 max.) provide excellent interchangeability
- Substantially linear transfer characteristics minimize cross-modulation problems
- Hermetically sealed 4-lead package with TO-18-size case

TYPICAL APPLICATIONS:

- High-Input-Impedance AF Amplifiers
- Operational Amplifiers
- RF Buffer Amplifiers
- Electrometers
- Shunt and Series Choppers

Maximum Ratings, *Absolute-Maximum Values:*

DRAIN-TO-SOURCE VOLTAGE, V_{DS}	+32	max.	volts
DC GATE-TO-SOURCE VOLTAGE, V_{GS}	-6 to +2	max.	volts
PEAK GATE-TO-SOURCE VOLTAGE, v_{GS}	±15	max.	volts
DC GATE-TO-SUBSTRATE VOLTAGE, V_{GB}	-1 to +2	max.	volts
PEAK GATE-TO-SUBSTRATE VOLTAGE, v_{GB}	±15	max.	volts
DRAIN-TO-SUBSTRATE VOLTAGE, V_{DB}	-0.3 to +32	max.	volts
DRAIN CURRENT, I_D	15	max.	ma
TRANSISTOR DISSIPATION, P_T:			
At Free-Air Temperatures up to 85° C	150	max.	mw
FREE-AIR TEMPERATURE RANGE:			
Storage	-65 to +125		°C
Operating	-65 to +85		°C
LEAD TEMPERATURE (During Soldering):			
At Distances Not Closer than 1/32 inch to Seating Surface for 10 Seconds max.	230	max.	°C

* Formerly Dev. No's. TA2624 and TA2625, respectively.

** Metal Oxide Semiconductor

RADIO CORPORATION OF AMERICA
Electronic Components and Devices Harrison N J

Trademark(s) ® Registered
Marca(s) Registrada(s)

3N98, 3N99 11-64
Printed in U.S.A.

(a)

RCA-3N98, 3N99

SILICON INSULATED-GATE

FIELD-EFFECT ("MOS") TRANSISTORS

(b)

Fig. 9. (a) The front and (b) the back covers of the RCA product data sheet of the second volume-produced silicon MOSFET, an n-channel depletion-mode transistor.

RCA could not make up its mind either. Their data sheet gave two names on the front cover (Fig. 9(a)) and a third in the back (Fig. 9(b)). On the front cover, they called it the **FIELD-EFFECT ("MOS") TRANSISTOR.** Note the quotation around MOS which is doubly emphasized with a parenthesis. It must be a foreign object. Then to make it more domestic, they also called it the **SILICON INSULATED-GATE FIELD-EFFECT TRANSISTOR** since RCA was ready to market a CdS insulated-gate field-effect transistor at the time (it failed to reach the marketplace). Just to be sure not to miss the boat on the name-calling contest, RCA's marketing manager put down all the possible descriptives on the back of their data sheet (Fig. 9(b)). However, he gave only one permutation of the nine words: **SILICON INSULATED-GATE FIELD-EFFECT ("MOS") TRANSISTOR.** Note **MOS** is still foreign, quoted, and bracketed. Will the marketing genius from RCA please stand up?

There were two other acronyms of lesser virtue than these first six. These were: the **INSULATED GATE FIELD EFFECT TRANSISTOR** or **IGFET** probably from RCA. This was also used by Simon Sze in the first (1969) edition of his popular device physics book [98]. But, it did not stick. It was too hard to pronounce **IGFET**, (I-G-FET?), and Sze conceded to the easier name, MOSFET in his second (1981) edition [98].

APPLICATIONS OF THE SILICON PLANAR II MOS FET

By JOHN S. MacDOUGALL

E. R. deAtley, Technical Writer and Editor

I. INTRODUCTION

The FI100 Metal Oxide Silicon Field-Effect Transistor (MOSFET) is the first of a new line of products manufactured by a refined planar process known as PLANAR II. In earlier planar MOSFET's the dielectric was subject to charge migration due to heavy electric fields (up to 1 million volts per centimeter) often met in operation. This charge migration caused gross changes in device operating parameters and deterioration of P-N junctions. The PLANAR II MOSFET, however, can withstand electric fields of the order of 2 million volts per centimeter without dielectric charge migration and therefore is an inherently stable device.

This note describes the physical characteristics of the PLANAR II MOSFET and gives a number of applications made possible by the special properties of this device.

II. PHYSICAL CHARACTERISTICS

Figure 1 is a cross-sectional representation of the MOSFET showing DC biasing polarities. Two P islands are diffused close together into a high-resistivity N substrate and covered by a thin insulating layer of silicon dioxide (stipled area in the diagram). Metal electrodes called the "source" and the "drain" are evaporated in the oxide layer in contact with the P islands. A third metal electrode called the "gate" is evaporated on the oxide layer insulated from the source and drain, and a fourth—the "base substrate"—is connected at the bottom of the N-type material.

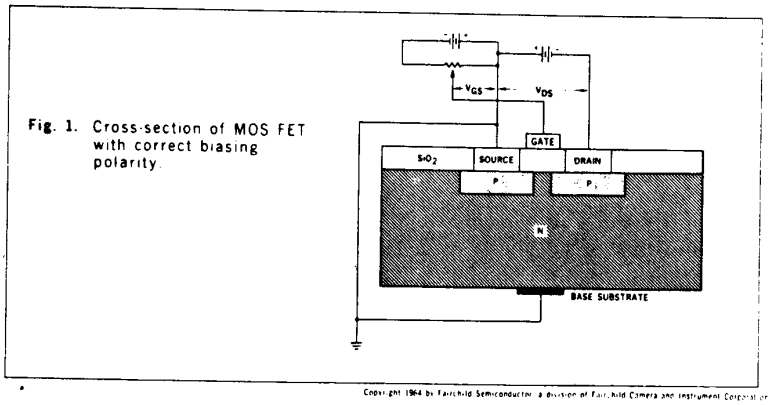


Fig. 1. Cross-section of MOS FET with correct biasing polarity.

Copyright 1964 by Fairchild Semiconductor, a division of Fairchild Camera and Instrument Corporation

FAIRCHILD SEMICONDUCTOR

313 FAIRCHILD DRIVE, MOUNTAIN VIEW, CALIFORNIA, (415) 962 5011, TWX 910 379-6435

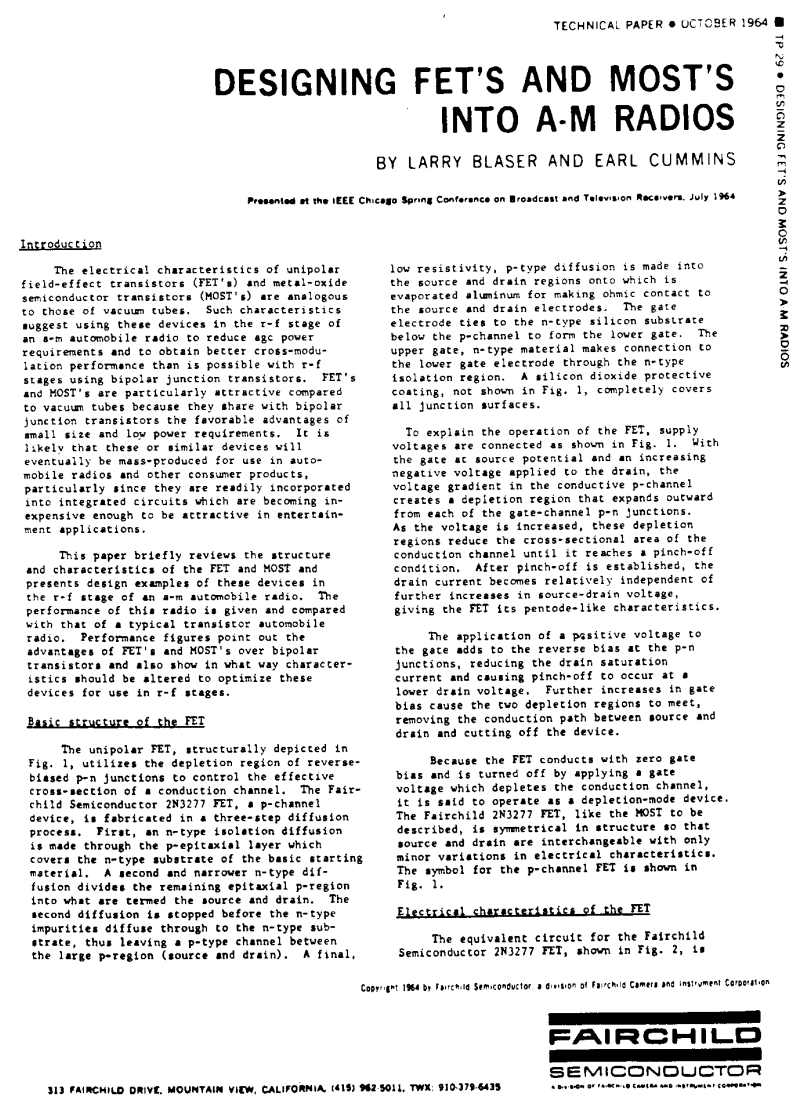
Fig. 10. Cover page of a Fairchild application bulletin on the silicon MOSFET.

An eighth name was also proposed by someone who called it the **Metal Insulator Semiconductor Field Effect Transistor.** This was also not popular although it is a more accurate description of the structure since many FETs use several gate insulators other than just oxide alone, such as the nitride/oxide dual dielectric gate. It is more generic and gives no indication of what the insulator is. It is still used by some unaware engineers as it has a rather notorious sounding acronym, **MISFET.** With one typing error or a bit error from a failed MOSFET in one's word processor, our MOST (MISFET) could indeed become a 'misfit'.

MOSFET, originated from the Fairchild marketing department in October 1962, has been the general consensus for the past decade, including IBM who had used FET and IGFET for a decade. But, younger engineers and writers from the orient have frequently used the first and original acronym, MOST, which Frank Wanlass and I felt most appropriate in the summer of 1962 when we were writing the CMOS manuscript for the 1963-February ISSCC proceeding. As a compromise, I am using the easy-to-say name **MOS TRANSISTOR** in the title of this review, following the good example of John Moll who coined the term MOS CAPACITOR. The most generic acronym to describe the basic principle of operation is probably the FET for all field-effect transistors and **MIT** or **M-CIT** for the minority-carrier injection transistor. BJT instead of MIT has been used for the bipolar junction transistor which is an unfortunate choice since the word 'junction' could mean a metal/semiconductor, liquid/solid or even gas/solid interface or junction as well as a p-n junction.

M. New Technology Developments

The first five years of the 1959–1969 decade saw a number of important innovations and developments on the MOSFET we have just discussed, such as the demonstration of the use of thermally grown oxide for transistor stabilization



TECHNICAL PAPER • OCTOBER 1964 ■ TP 29 • DESIGNING FET'S AND MOST'S INTO A-M RADIOS

DESIGNING FET'S AND MOST'S INTO A-M RADIOS

BY LARRY BLASER AND EARL CUMMINS

Presented at the IEEE Chicago Spring Conference on Broadcast and Television Receivers, July 1964

Introduction

The electrical characteristics of unipolar field-effect transistors (FET's) and metal-oxide semiconductor transistors (MOST's) are analogous to those of vacuum tubes. Such characteristics suggest using these devices in the r-f stage of an a-m automobile radio to reduce agc power requirements and to obtain better cross-modulation performance than is possible with r-f stages using bipolar junction transistors. FET's and MOST's are particularly attractive compared to vacuum tubes because they share with bipolar junction transistors the favorable advantages of small size and low power requirements. It is likely that these or similar devices will eventually be mass-produced for use in automobile radios and other consumer products, particularly since they are readily incorporated into integrated circuits which are becoming inexpensive enough to be attractive in entertainment applications.

This paper briefly reviews the structure and characteristics of the FET and MOST and presents design examples of these devices in the r-f stage of an a-m automobile radio. The performance of this radio is given and compared with that of a typical transistor automobile radio. Performance figures point out the advantages of FET's and MOST's over bipolar transistors and also show in what way characteristics should be altered to optimize these devices for use in r-f stages.

Basic structure of the FET

The unipolar FET, structurally depicted in Fig. 1, utilizes the depletion region of reverse-biased p-n junctions to control the effective cross-section of a conduction channel. The Fairchild Semiconductor 2N3277 FET, a p-channel device, is fabricated in a three-step diffusion process. First, an n-type isolation diffusion is made through the p-epitaxial layer which covers the n-type substrate of the basic starting material. A second and narrower n-type diffusion divides the remaining epitaxial p-region into what are termed the source and drain. The second diffusion is stopped before the n-type impurities diffuse through to the n-type substrate, thus leaving a p-type channel between the large p-region (source and drain). A final, low resistivity, p-type diffusion is made into the source and drain regions onto which is evaporated aluminum for making ohmic contact to the source and drain electrodes. The gate electrode ties to the n-type silicon substrate below the p-channel to form the lower gate. The upper gate, n-type material makes connection to the lower gate electrode through the n-type isolation region. A silicon dioxide protective coating, not shown in Fig. 1, completely covers all junction surfaces.

To explain the operation of the FET, supply voltages are connected as shown in Fig. 1. With the gate at source potential and an increasing negative voltage applied to the drain, the voltage gradient in the conductive p-channel creates a depletion region that expands outward from each of the gate-channel p-n junctions. As the voltage is increased, these depletion regions reduce the cross-sectional area of the conduction channel until it reaches a pinch-off condition. After pinch-off is established, the drain current becomes relatively independent of further increases in source-drain voltage, giving the FET its pentode-like characteristics.

The application of a positive voltage to the gate adds to the reverse bias at the p-n junctions, reducing the drain saturation current and causing pinch-off to occur at a lower drain voltage. Further increases in gate bias cause the two depletion regions to meet, removing the conduction path between source and drain and cutting off the device.

Because the FET conducts with zero gate bias and is turned off by applying a gate voltage which depletes the conduction channel, it is said to operate as a depletion-mode device. The Fairchild 2N3277 FET, like the MOST to be described, is symmetrical in structure so that source and drain are interchangeable with only minor variations in electrical characteristics. The symbol for the p-channel FET is shown in Fig. 1.

Electrical characteristics of the FET

The equivalent circuit for the Fairchild Semiconductor 2N3277 FET, shown in Fig. 2, is

Copyright 1964 by Fairchild Semiconductor, a division of Fairchild Camera and Instrument Corporation

FAIRCHILD SEMICONDUCTOR

313 FAIRCHILD DRIVE, MOUNTAIN VIEW, CALIFORNIA, (415) 962-5011, TWX: 910-379-6435

Fig. 11. Cover page of a Fairchild technical paper on designing the silicon MOSFET into AM radios.

by Atalla, Tannenbaum, and Scheibner; the laboratory operation of the modern silicon MOSFET using the thermally grown oxide for the gate insulator by Kahng and Atalla; the planar process by Hoerni; the monolithic technique by Noyce; the CMOS circuit invented by Wanlass; the major management decisions to develop the silicon NMOS FET integrated circuits for the IBM mainframe computers; and the delivery of the first commercial silicon MOSFET, a p-channel triode by Fairchild and an n-channel tetrode (two gates) by RCA.

The second five years of this period, to be discussed in this section, saw major advances in the understanding of oxide stabilization mechanisms, key developments of the fabrication technology and the invention of two novel integrated MOSFET device structures: i) the four-terminal MOSFET, with a bias voltage applied to the substrate gate, by IBM and Heiman of RCA which made it possible to use the higher performance n-channel MOSFET as the main memory in the IBM-370/158 mainframe, and ii) the one-transistor dynamic random access memory (DRAM) cell invented by Dennard of IBM. We shall first summarize the four oxide technology advances listed in Table 2.

In 1964, Snow and co-researchers [99] identified sodium ion drift in thermally grown oxide as the principal cause of threshold voltage instability in the electrical characteristics of silicon MOSFETs. Also in 1964, Kerr and Young [100], [101] at IBM East Fishkill Laboratory discovered that the silicon dioxide film can be electrically stabilized to eliminate the sodium ion drift by growing a phosphorus silicate glass (known as **PSG**) layer [41] onto the surface of the oxide film to getter the sodium. Deal and Snow showed that the threshold voltage shifts of MOSFETs, induced by polarization in the PSG layer on the oxide surface, can be controlled [102].

In 1965, Pieter Balk at IBM Yorktown Heights Research Laboratory [103], [104] suggested that hydrogen can anneal

out the surface states (interface traps at the oxide/silicon interface) by tying up the dangling silicon and oxygen bonds. In the same year, Kooi of the Philips Research Laboratory [105] reported that his extensive measurements confirmed Balk's idea and experiments.

Since the abstract volumes [103], [104] of Balk's two papers are difficult to get and Balk's concluding statements are so scientifically precise and have withstood the test of time—in fact, they have been proven correct by all the measurements reported in the last twenty-three years—we quote the entire concluding paragraph and summary made by Balk from his abstracts to remind the young silicon technologists who do not know why, and the old who have forgotten from wearout of memory cells. In the 1965 San Francisco paper [103] Balk stated (My comments are added in curly brackets):

> "The main effect of the H_2 treatment appears to be the annihilation of fast states {another name for interface states}. If these states are related to vacancies, accompanied by chemically unsaturated bonds {has been known as dangling bonds} and unpaired electrons near the interface, then the H_2 {not hydrogen ion or proton as some think} annealing may be in effect the chemical saturation {now known as hydrogenation} with H atoms of these bonds at the vacancies. The low state density obtained upon steam oxidation is probably caused by hydrogen, evolved during oxidation, and retained in the oxide. The similarity in action between H_2 and Al remains as yet unexplained."

This unexplained issue on the effect of aluminum gate electrode was then resolved by Balk's second paper presented in October of the same year at the Buffalo meeting of the Electrochemical Society [104] in which he stated:

> "The similarity of the annealing behavior of the electrical interface properties of Si-SiO_2 in H_2 and Al-SiO_2-Si in N_2 around 300°C suggests that the same mechanism is operative in both cases. Hydrogen released in a reaction between Al and hydroxyl groups in the oxide is proposed as the active agent in the Al-SiO_2-Si case. This model is supported by the absence of any annealing effects on 'ultra-dry' oxide."

These conclusions of Balk have had profound effects on the progress made towards the stabilization of the silicon VLSI integrated circuit and increasing manufacturing yields. In many cases, the young process engineers and developers have followed the ancient recipe of annealing in forming gas, unknowing of its basic reasons nor its origin from these two 1965 conclusions of Balk. Balk's hydrogen bond model of passivating and deactivating the interface traps has also been the chemical-atomic base for the characterization of the generation, annealing and charging kinetics of the interface and oxide traps due to silicon and oxygen dangling bonds [61], [62], [87].

Also in 1965, Balk, Burkhardt and Gregor at IBM [106], DeLord, Hoffman, and Stringer at Reed College in Oregon [107], and Miura [108] at NEC-Japan independently discovered that the surface or interface state density is lower on the oxidized (100) silicon surface than the (110) and (111) surfaces. This is consistent with the Bardeen-Shockley-Handler silicon dangling bond model [109], [110]. Recent reliability experiments have shown just the opposite but **expected** trend based on the hydrogen bond breaking model—interface traps are generated faster on (100) than (111) surface during high electric field stress although the initial pre-stress densities are smaller on the (100) than the (111) surface [61].

These four discoveries still form a part of the basis in today's silicon VLSI fabrication technology. Sodium ion contamination is minimized and almost completely eliminated by adhering to clean transistor processing procedures such as minimizing contact of processed silicon VLSI wafers with salty air and contaminated handling utensils. High temperature sodium gettering steps have also been employed to further reduce sodium concentration in the oxide, such as using a PSG layer formed at around 800 or 900°C over the oxide surface. More recently, chlorine from HCl, **TCE** (trichloroethylene), or **TCA** (trichloroethane) is used to leech out the sodium in the fused-quartz furnace tube at high temperature (1000–1100°C). A small amount of chlorine (3 percent) from one of these sources is also added to the oxygen during oxidation so that chlorine is incorporated into the oxide which apparently helps to trap and immobilize the residual sodium in the oxide film at room temperatures in addition to leeching them out during high temperature oxidation and diffusion. Starting silicon wafers with (100) surface orientation have been almost exclusively used by all VLSI manufacturers due to the lower interface trap density on the oxidized (100) silicon surface. Further reduction of surface or interface traps by hydrogen has been implemented also by most VLSI manufacturers following Balk's original idea of passivation of the silicon and oxygen dangling bonds by hydrogen. During the final high-temperature processing step, the VLSI chip is soldered (known as die-attach, which this author terms chip-bond in analogy to wire-bond) to the chip carrier or integrated circuit carrier at 400–450°C in a forming gas atmosphere of about 10-percent hydrogen and 90-percent nitrogen. As a result of these modern processing steps, the surface states density (interface plus oxide traps) of 10^{13} states/cm^2 observed by Shockley and Pearson in the first silicon field-effect experiments in 1948 [18] has now been reduced by about ten thousand times, to less than 10^9 traps/cm^2 at the oxide/silicon interface.

There were two other technological innovations which have found continued widespread use. One was the silicon gate process reported by Kerwin, Klein, and Sarace of Bell Labs in 1963 [111]. The other was the silicon nitride mask, reported by Sarace, Kerwin, Klein, and Edwards also of Bell Labs in 1968 [112]. Nitride mask was used at Bell Labs at Allentown in late 1965 or early 1966 (private communication from J. M. Early, Oct. 1, 1988). Both are used as masks against silicon oxidation and impurity diffusion. These two technologies have been developed further [113]–[116] and were used in the production of the first silicon MOS integrated circuit by Intel [116]. The silicon gate process provides self-alignment of the gate over the drain and source junctions. It is an indispensable technology today to shrink the geometry of silicon MOSFETs and BJTs to micron and submicron dimensions in order to put several hundred thousands to millions of MOS and bipolar transistors on a single VLSI silicon chip of a quarter to a half [97] of an inch on the sides. The nitride mask against diffusion and oxidation is not used today as much as it was used in producing the second and third generation silicon chips five to ten years ago. How-

ever, it is still widely used as the dielectric for the charge storage capacitor of the dynamic MOSFET memory cell since Si_3N_4 has a dielectric constant of 7.5/3.9 = 1.9 or about twice that of SiO_2 so that higher capacitance or charge storage can be attained in a smaller area which helps to increase the cell density on a silicon chip. The nitride mask is also still universally used in the **LOCOS** (local oxidation of silicon) process to grow thick field oxides adjacent to the thin gate oxide in order to reduce the fringe electric field [97] and junction capacitance.

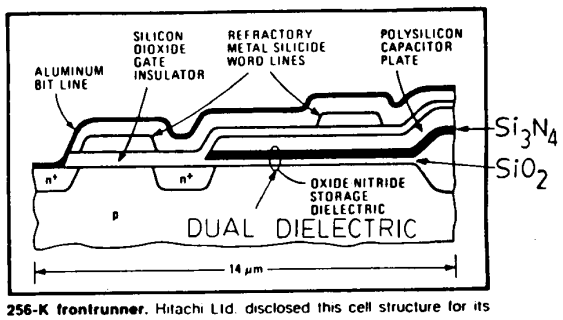
N. The Body Effect

During this period, there were also two major device innovations among the many technology and device inventions and developments just discussed. The first was invented to control the threshold voltage of the n-channel MOSFET by means of the body effect. In 1965, Heiman [117] at RCA and a group at IBM showed that a dc voltage applied to the silicon substrate can control the threshold gate voltage of the MOSFET. This was the very technique that allowed IBM to select the performance-competitive (against magnetic core memory's access time of 1 μs) higher-speed n-channel Si MOSFET (NMOS), instead of the slower p-channel Si MOSFET (**PMOS**), for the mainframe memory of the IBM-370/158 computer that was delivered in 1973. The threshold voltage adjustment by the substrate (or body) dc voltage was necessary to cut off the n-type surface inversion channel on the p-type silicon substrate (body) of an NMOS since an n-type channel was induced by the presence of unavoidable residual positive charges in the oxide during this earlier technology era. Today, substrate bias is no longer necessary since controlled amounts of boron, phosphorus, arsenic or antimony can be ion implanted into the silicon surface layer under the gate oxide to adjust the threshold voltage of both the N-MOSFET and P-MOSFET in order to compensate any residual oxide charges that might still be present in a particular production run. However, substrate bias also reduces the junction capacitances of the source, drain and channel, giving an important performance advantage [97].

O. Invention of the One Transistor DRAM Cell

The second device invention in MOSFET during the second half (1963–1968) of the second phase has had a major and long-lasting impact on the electronic industry today which will extend into the distant future. Its influence is comparable to the invention of the transistor itself, a worldwide consensus. This was the invention of the one transistor (**1-T**) dynamic memory cell for use in the random access memory (DRAM) by Dennard [118], [119] at the IBM Yorktown Heights Research Laboratory in 1967. For this 1967 invention and the development of the scaling law to design smaller MOSFET (see Section V-A), Dennard received the 1982 Cledo Brunetti Award of the IEEE and was elected to the National Academy of Engineering. The cell contains one MOSFET switch and one charge storage capacitor integrated together. The MOSFET serves as the switch to charge (write) and discharge (read) the capacitor. The 1-T DRAM cell is now the most abundant man-made object on this planet earth. The annual production was projected to reach about 100 trillion (1-T) DRAM cells in 1985. If the rate continues to increase as in the past, 10^{20} DRAM cells could be produced in the beginning of the twenty-first century. This would equal the number of neurons in one billion brains (if each brain contains 100 billion neurons or neural cells). This no doubt will help to hasten the development of artificial intelligence and neural computers which need a large amount of fast memory bits.

The cross-sectional view of the 1-T DRAM cell in one of the first volume-produced 256-kbit DRAM chips is shown in Fig. 12. The charge storage capacitance has a two-layer



256-K frontrunner. Hitachi Ltd. disclosed this cell structure for its 256-K dynamic RAM, to be available in 1983. The cell measures 14 by 7 μm, using 2-μm minimum mask geometries, and gets a 50-femtofarad storage capacitance with a two-layer dielectric

Fig. 12. The cross-sectional view of the first publicly announced 256-kbit/chip silicon NMOS DRAM using the one-transistor Dennard cell and a dual-dielectric storage capacitor at the 1982-IEDM, from Electronics [154].

or dual dielectric insulator consisting of a nitride on a thin thermally grown oxide. The silicon nitride has a higher dielectric constant (7.5) than the silicon oxide (3.9), resulting in a higher capacitance per unit area, a larger amount of charge stored in a smaller area, and a higher bit density. The nitride film also seals pinholes in the thin oxide which improves the yield.

The DRAM cell will be the device (or circuit) unit used to trace the evolution of the integrated circuits technologies in the reviews to follow.

P. Invention and Development of the Floating Gate Transistor Cell for Electrically Erasable Nonvolatile Memory

The one-transistor dynamic random-access memory or DRAM cell is volatile, that is, the stored bit (or charge) is lost when the power supply is turned off. In contrast, the bit or magnetic polarization stays in the magnetic core, disk, and tape memories. The idea of using charge stored on the floating gate of a MOSFET to nearly permanently (10 years or longer) store a bit of information, so that it is not lost when the power supply is turned off, was well known to Wanlass, Grinich, Tremere, and other Fairchild engineers and managers since 1961 when it was observed by Sah in the spring of 1961 at Fairchild that the potential on the floating gate (unconnected to external leads) of a MOSFET will not decay overnight after the power supply was turned off. (See Sections IV-H and IV-J.) But they were engrossed in the very profitable bipolar transistor and integrated circuit development and manufacturing activities as well as research and development into the very exciting PMOS and micropower CMOS (complementary MOS, see Section IV-J). They did not pursue the nonvolatile memory idea further at that

time. The first study of the floating gate was reported by Kahng and Sze [120] on an $Al/ZrO_2/Zr/SiO_2(50A)/n$-Si MOSFET. It was based on a patent filed by Kahng in 1967 [121], which followed the 1955 invention of Ross [32] on a MOSFET memory using the ferroelectric gate, which was an extension of Brown's 1953 surface channel structure [31]. Moreover, the surface channel was already known to Bardeen in the first modern patent on solid-state amplifiers [7], to Bardeen and Brattain in their point-contact transistor experiments [19] and to Shockley in his patented gated p-n junction amplifier [8]. Furthermore, the surface inversion channel was implicitly assumed by Heil in his 1935 MOSFET patent [5]. The foregoing traces the historical connections and precedents. Kahng's extension used the silicon MOSFET and a floating gate conductor buried in the SiO_2 gate dielectric instead of the ferroelectric gate dielectric over the p-base layer of a Ge n-p-n BJT of Ross. Fowler–Nordheim electron tunneling through the 50-A SiO_2 between the floating Zr gate and the silicon substrate was used to charge and discharge the floating gate by Kahng. It was ten years later when an engineering effort was made by Dov Frohman-Bentchkowsky at Intel [122]–[125] with new innovations that resulted in a practical memory transistor cell which has been successfully volume produced and initially marketed as a 2 kbit nonvolatile erasable-programmable-read-only-memory (EPROM). For this 1971 effort, Frohman-Bentchkowsky received the 1982 Jack A. Morton Award of the IEEE. In this EPROM chip, ultraviolet light was employed to erase the charge on the floating gates. Avalanche multiplication of electrons and holes from the silicon surface beneath the floating gate was utilized to inject charges onto the floating gates, and it requires several milliseconds to complete charging (or erasing) the floating gates on the 2-kbit chip. This cell was known initially as the **FAMOS** (Floating-gate Avalanche-injection MOS) cell and is currently known as the **UV-EPROM** (Ultra Violet light Erasable Programmable Read Only Memory). Intel engineers later developed a faster and nonoptical scheme to write and erase the charge by the use of electron tunneling through a thin oxide to charge and discharge the floating gate which was first proposed in Kahng's 1967 patent [121]. This is generically known as the **EEPROM** (Electrical Erasable Programmable Read Only Memory) and specifically **FLOTOX** (Floating-gate Tunnel Oxide). This was the second time that the quantum mechanical tunneling phenomenon is used successfully in a volume-produced product. (The first time was the low-voltage p-n junction voltage reference diode where tunneling dominates below ten volts. The tunnel or Esaki diode was another quantum mechanical tunneling device but it failed in the marketplace.) The write speed has been improved to 1 ms from a previous value of more than 10 ms in the so-called flash **EPROM**. Three comprehensive earlier reviews have been written on nonvolatile semiconductor memories [126]–[128].

Q. A Nonvolatile Silicon DRAM Cell

An exciting new nonvolatile memory cell in a 512-bit test chip was reported in the last **IEDM** (December 1987) by Kinney, Shepherd, Miller, Evans, and Womack of Krysalis Microelectronics Corporation [129], [130]. They replaced the dielectric capacitor of Dennard's DRAM cell by a ferroelectric capacitor. The idea of using a ferroelectric capacitor for information storage was first proposed in 1955 by I. M. Ross of Bell Telephone Laboratories [32]. Ross's device uses the ferroelectric as the gate dielectric of the MOSFET (actually over the base of a Ge n-p-n-BJT) instead of a separate memory capacitor as in the Krysalis cell. Many recent attempts have been made to improve the wearout rate and write speed of ferroelectric gate without success [131]. The new DRAM cell disclosed by the Krysalis group [129] is novel. It combines two old ideas, Dennard's dielectric capacitor memory and Ross's ferroelectric polarization to replace the dielectric capacitor. The separation of the gate dielectric (still a thermally grown oxide) of the MOSFET switch from the ferroelectric film in the memory capacitor provides process and design flexibility and cell reliability. High write speed (<100 ns) and wearout tolerance (10^{15} write cycles) have been projected from laboratory experiments [132]. The very excellent performance comes from the physical separation of the ferroelectric polarization storage site from the MOSFET gate in this new cell. It is an experience taught in the design of the DRAM and EEPROM. This new success of combining the Si MOSFET and a ferroelectric capacitor brings renewed interest that a practical nonvolatile memory with high write speed may soon become available. (More in Section V-I.) The development of this ferroelectric DRAM cell has also been undertaken by a second company, Ramtron Corporation [133].

V. MOS Integrated Circuits (1969 to 1988 and Beyond)

The third phase of the MOSFET evolution begins when Noyce and Moore, two of the Fairchild Semiconductor founders, finally decided to start their own company with one objective in mind: to manufacture silicon MOS integrated circuits. They could not attain this goal at Fairchild for several reasons, even though they were in charge of the whole Fairchild operation: i) Fairchild's main products and bread and butter winner were the bipolar silicon integrated circuits which they originated; ii) the engineers assigned to transfer the first MOS integrated circuit product from the Fairchild R&D laboratory to the manufacturing plant two miles away always leave to start their own company (to my recollection, this occurred thrice in three successive years, 1961–1963); and iii) financing from the dwindling bipolar profit and Fairchild's parent firm was inadequate to start a separate MOS manufacturing plant.

We shall next describe some of the historical highlights and firsts in the development and volume production of the silicon MOSFET integrated circuits during the third phase, from **SSI** (Small Scale Integration, about 100 MOSFETs on a 100×200 square mil silicon chip), to **MSI** (Medium Scale Integration, 1000 MOSFET/chip), to **LSI** (Large Scale Integration, 10 000 MOSFET/chip) to **VLSI** (Very Large Scale Integration, more than 100 000 MOSFET/chip) and to **ULSI** (Ultra Large Scale Integration, several million or more MOSFET/chip). The chip area has increased gradually from SSI to the current VLSI, by a factor of about four while the cell area and MOSFET line widths both have decreased about twenty-five times, the cell area from greater than 75 to 3 square micrometers, and the line width from 25 to less than 1 micrometer. The invention, laboratory demonstration and first production of both the DRAM and some SRAM (Static RAM) MOS chips using different MOSFET circuits [134] for the memory cell during 1969 to 1972 are summarized in

Table 3 Evolution of the Silicon MOS Random Access Memory (RAM) (1969–1988)

Disclosure Dates[a] Subm.	Pub.	Authors-Inventors Development Team	Institutions or Locations[b]	Memory Density (bit/chip)	Device and Technology[c]	Reduction to Practice[d]	Ref.
1965	1965	Schmid	FSC	64	SRAM PMOS 6-T Al-gate	Eng	[135]
	1968	Noyce, Moore	INTEL		Intel Corp. Founded		
1968	1969	1101	INTEL	256	SRAM PMOS Si-Gate	Lab	[137]
1968		Engineering	IBM	512	SRAM NMOS Al-Gate	Lab	[140]
	1970	Spampinato, Terman	IBM		DRAM 4T	Lab	[140]
	1970	1103 Engineering Team	INTEL	1k	DRAM PMOS 3T Si-Gate	Prod	[137]
	1971	Engineering	INTEL	1k	SRAM NMOS 6T Si-Gate	Prod	[137]
	1971	Frohman-Bentchkowski	INTEL	2k	EPROM PMOS 2T Al-Gate	Lab	[123]
	1971	Sonoda	IBM		Voltage Multiplier	Lab	[140]
	1972	Various Manufacturers			Logic CMOS Watch Chip	Prod	[143]
	1972	2104 Engineering	INTEL	4k	DRAM NMOS 1T Si-Gate	Prod	[137]
	1973	IBM 370/158, 168	IBM	1→2k	SRAM NMOS 256kB/card	Prod	[140]
	1974	21MX 16bit Minicomputer	HP	4k	DRAM NMOS 4kB/card	Prod	[154]
	1975	Yu, Dennard *et al.*	IBM	8k	SRAM NMOS 6T 1μm EB	Lab	[140]
	1976	2116 Engineering	INTEL	16k	DRAM NMOS 1TM 2-Poly	Prod	[152]
1975	1979	SAMOS	IBM	64k	DRAM NMOS 1TM Si-Gate	Prod	[140]
	1979	IBM 4331 SAMOS	IBM	64k	DRAM NMOS 1MB/card,SiAl	Prod	[140]
	1982	9000 32bit Minicomputer	HP	128k	SRAM NMOS	Prod	[154]
	1982	Sunami	Hitachi		3D Capacitance	Lab	[164]
1982	1983	Electronic Week	Hitachi	256k	DRAM NMOS RM,Al,Nitride	Lab	[154]
	1984	Various		256k	DRAM NMOS	Prod	[154]
	1984	IBM Essex Junction		1M	DRAM NMOS 1T2d SAMOS	Lab	[154]
	1984	Hitachi, NEC, NTT, Toshiba, Ti		1M	DRAM NMOS 1T3d trench-C	Lab	[170]
	1985	IBM Essex Junction		1M	DRAM NMOS 1T2d SAMOS	Prod	[154]
	1985	ATT, Fujitsu, Hitachi, Toshiba		1M	DRAM NMOS 1T2d	Eng	[154]
	1985	TI		1M	DRAM NMOS 1T3d trench-C	Eng	[154]
	1985	IBM E. Fiskill, Yorktown Ht		64k-4M	DRAM PMOS 1T3d trench-C	Lab	[169]
	1985	Hitachi, Toshiba, NEC		4M	DRAM NMOS 1t3d trench-C	Lab	[170]
	1985	Chatterjee *et al.*	TI	4M	DRAM NMOS 1T3d >1μm TCT	Lab	[173]
	1985	IBM Research	IBM	16M	DRAM NMOS 1T2d 0.5μm EB	Est	[198]
	1986	IBM-3090	IBM	1M	DRAM NMOS 1T2d SAMOS	Prod	[154]
	1986	MicroVAX-2 Toshiba/Chrislin		1M	DRAM NMOS 16MB/card	Prod	[154]
	1988	Matsushita, Toshiba, Hitachi		16M	DRAM CMOS 1T3d trench-C	Eng	[177]–[179]
	1995	SRC University Research		64M	DRAM CMOS 1T3d 0.25μm	Est	[130]
	1987	Miller *et al.*	KRYSALIS	512	DRAM FMOS 1T2D FDRAM	Lab	[129]
	1988	Sheffield *et al.*	RAMTRON	512	DRAM FMOS 1T2d FDRAM	Lab	[133]

[a]Subm. = Submission date. Pub. = Publication date.
[b]FSC = Fairchild Semiconductor Corporation; HP = Hewlett-Packard; TI = Texas Instruments.
[c]T = MOSFET; TM = MOSFET-Merged; k = 1024; M = 1024k; RM = Refractory Metal; EB = Electron Beam Lithography; trench-C = trench-capacitor; TCT = trench capacitor-trench transistor; nd = n-dimensional; FMOS = Ferroelectric MOS cell.
[d]Lab = Laboratory; Eng = Engineering sampling; Prod = Production; Est = Estimation.

Table 3. Mass production and volume applications began about 1972, when Intel marketed the first 4-kbit DRAM, and these events are also listed in Table 3.

A. From SSI to LSI (1969 to 1972)

The first laboratory MOS **RAM** chip was reported in 1965 by Schmidt of the Fairchild Semiconductor Laboratory [135], [136] who built a 16 × 4 (4 bits per word) or 64-bit SRAM (Static Random Access Memory) using the 6-transistor PMOS static memory cell or flip-flop [134]. The first product was an INTEL 256 bit RAM chip using the Si-gate PMOS technology in 1969 which was quickly succeeded by chips with higher bit count within a year [137]–[142]. IBM had a 512-bit aluminum gate NMOS memory built for laboratory testing purpose one year earlier [140], [141]. The first mass produced and volume delivered part was the INTEL 3-transistor, Si-gate PMOS, 1-kbit DRAM (part number, 1103) announced in 1970 [137] although it was quickly replaced by the 4-kbit DRAM described in the next subsection.

A most important MOSFET design technique was advanced by Dennard and co-researchers at IBM around 1973 [143]. They showed that the transistor dimensions can be reduced without compromising its current-voltage characteristics. This has been known as the scaling law.

Probably the earliest volume commercial application of the MOSFET was the use of the CMOS logic inverter circuit [90] in the watch chip as the frequency divider [261]. This began around 1970. Since CMOS uses very little power, inexpensive wrist watches with battery life of a few months became a popular commodity. The rather short battery life of the Texas Instruments' digital watch was due to the heavy current drain of the time display using the light emitter diode (**LED**) which killed the LED watch. The battery life was increased to one year or more when the low power **LCD** (Liquid Crystal Display) replaced the brighter but power-thirsty LED. The larger scale integration using CMOS in recent years has made it possible to produce long-battery-life watches with many more timing functions and even a calculator on the wrist watch. The electronic watch, known today as the digital watch, was reviewed by Eleccion in the April 1973 issue of the *IEEE Spectrum* [261].

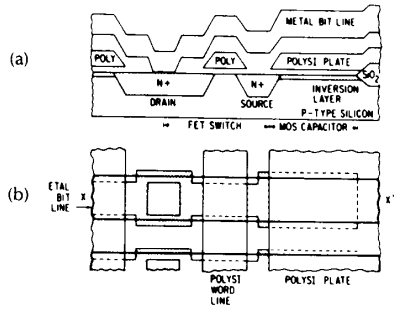
B. DRAM Technology Advances from 4 kbits to 64 kbits (1972 to 1979)

The success in the volume production of the silicon MOS integrated circuit appears to have been controlled and driven by the introduction of new technology and new production equipment [144] until about 1982 when the manufacturing technology and production planning became mature from past experiences. Since 1983, market demand and recovery of research, development, and especially production equipment and clean room costs appear to have delayed the introduction of higher (finer line) technology into MOS integrated circuit manufacturing [145]-[149]. Recent delays have been caused by the production shakedown of the submicron lithography tools.

Recovery of equipment and development costs of 256-kbit and 1-4-Mbit DRAMs has become a major factor that has dictated the three-year product introduction and delivery cycle in order to make business sense. For example, during a typical recent 3-year cycle, less than 1 million chips would be shipped for sampling during the introduction year, 5 million chips for workstations in the first production year, 50 million chips for mainframes in the second production year, and over 500 million chips for personal computers and consumer products in the third and peak production year. This trend is best illustrated using the one-transistor DRAM cell for two reasons. First, it is the largest volume product. Second, the repetitive memory structure makes it a good test vehicle to advance the silicon integrated circuit process technology since it requires the least engineering man-year to design a full memory array on a chip (with several hundred thousands to sixteen million transistors on the chip today) in order to run a full-scale test of a new technology in the factory environment. However, in the last few years high density MOS logic arrays have been increasingly used as the test vehicle by some American manufacturers [150]. The turn to logic chips as test vehicle was due in part to the lower production cost of DRAM by Japanese, Korean, and Taiwan manufacturers which had forced most of the American manufacturers to abandon the DRAM market after the price erosion of the 64-kbit DRAM chip began around 1983. The American preference of the logic test chip, however, is also due to the engineering and manufacturing expertise accumulated from producing the dense logic chips such as the popular (and extremely profitable) Intel 8080, 8086, 80186, 80286 and 80386 **MPU**s (Micro Processor Units) and the Motorola 68000 MPU families. Today, there is an increasing trend in using CMOS-SRAMs as the test vehicle for putting new technologies into production since the CMOS logic and SRAM circuits are easier to design than DRAM which requires a refresh clock [97]. The latest trend has been in the direction of BICMOS, the combination of bipolar junction transistor and the CMOS devices to gain further speed (8-12 ns, now 5 ns) while keeping down the power consumption of the chip.

The lower two-thirds of the milestone chart given in Table 3 provides a summary of the progresses made since 1972. The first widely used silicon MOS memory chip, especially in computers, was the 4-kbit DRAM chip (part number 2104), mass produced first by INTEL in 1972. Intel employed the Si-gate NMOS technology, the one-transistor Dennard DRAM Cell, and a source and drain region formed by diffusion. (In the merged transistor cell, the source was eliminated—see the description on the 16-kbit DRAM chip given later.) Intel was so successful that AT&T closed its 4-kbit DRAM line in Allentown, PA, after an unsuccessful start in 1973 and placed a $2M, DRAM order with INTEL, for prototype equipment development. AT&T experienced a repeat ten years later in 1983 and at the 256-kbit and 1-Mbit DRAM levels when it abandoned the completion of a 1-million (or multi-hundred-thousand) square-foot class-1 factory in Kansas City that was designed for the 1-Mbit DRAM production. Similar fate was experienced by several other American and foreign manufacturers during 1986-1987, and new mega square-foot, mega-bit DRAM factory buildings are still standing, equipmentless, empty shells today. The intricacy of the new production technology, the very expensive production equipment (more than $1M for a production plasma etcher or a production optical lithography wafer stepper), and the intense market competitions including accused pervasive dumping below cost by foreign producers to gain market share, have been the causes of these business failures. To assure supply, a U.S. DRAM Consortia has recently been proposed [262].

There were two popular cell geometries used in the 4-kbit DRAM memory chips in 1972. One was the INTEL geometry using a polysilicon word line and aluminum metal bit line. Its cross-sectional view and top view are shown in Fig. 13.



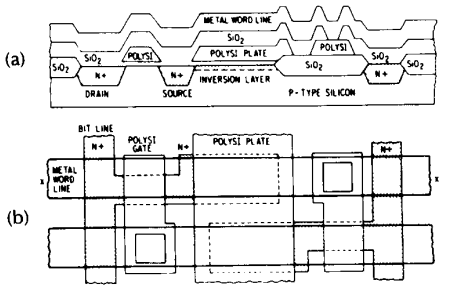
Single-polysilicon cell with metal bit line. The cross section shown in (a) is taken along line XX' of (b).

Single-Polysilicon Cell, Metal Bit Line
4K DRAM CELL
Minimum Feature L = 8 um
Cell Area $20L^2$ = 1280 um^2
Storage Capacitance = 120 fF

Fig. 13. The cross-sectional view of the single-poly-world-line/single-metal-bit-line 4-kbit DRAM cell. Metal is aluminum [144], [152].

The other 4-kbit cell, invented by MOSTEK (a recent casualty of the fierce competition which caused MOSTEK to be sold twice, first to United Technology and then to SGS-Thompson Semiconductor), used an aluminum metal word line and drain diffusion for the bit line. This is shown in Fig. 14. The diffused bit-line makes the cell smaller. At the 4-kbit chip level, the large series resistance of the diffused bit line does not degrade the performance or the switching speed significantly and the chip is easier to make [144], [152], [153].

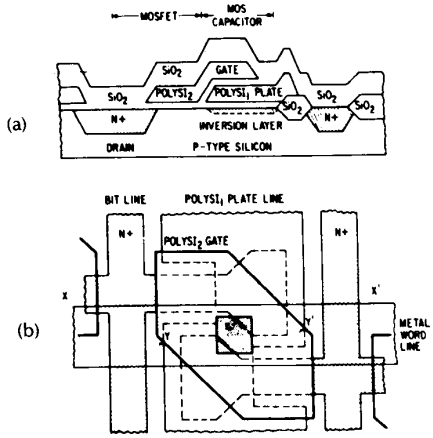
The next increase of memory bit density, from 4 to 16 kbits per chip, was announced in 1976. Its commercial success came about via three factors shown on Fig. 15: i) from a reduction of the 8-micron design rule for 4 kbits to the 5-



Single-polysilicon cell with diffused bit line. The cross section shown in (a) is taken along the line XX' in (b).

Single-Polysilicon Cell, Diffused Bit Line	
4K DRAM CELL (MOSTEK MK 4027)	
Minimum Feature	L = 8 um
Cell Area	$20L^2$ = 1280 um^2
Storage Capacitance	= 120 fF

Fig. 14. The cross-sectional view of the single-poly-word-line/diffused-bit-line 4-kbit DRAM cell [144], [152].



Double-polysilicon cell with diffused bit line. The cross section shown in (a) is taken along line XX' of (b). Positions Y and Y' indicate a potential metal step coverage problem.

Double-Polysilicon Cell, Diffused Bit Line				Intel-III
16K and 64K DRAM		16K	64K	64K
Minimum Feature	L =	5 um	3 um	2 um
Cell Area	$20L^2$ =	500 um^2	180 um^2	138 um^2
Storage Capacitance	=	60 fF	40 fF	145 fF

Fig. 15. The double-poly diffused bit line merged or sourceless 16- and 64-kbit DRAM cell [144], [152].

micron design rule for 16 kbits, ii) the source diffusion is removed which became known as the merged one-transistor dynamic RAM cell (**MDRAM** where the symbol 'M' is understood today), and iii) overlapping double polysilicon gates, one for the sourceless transistor and one for the charge storage capacitor, which is necessary to make the merged cell geometry. The density increase from 4 to 16 kbits was achieved by using only these innovative transistor device designs. No new process technology nor new equipment were required [144].

The road going from 16- to 64-kbit DRAM was a very rocky one which shook the financial foundation of many American manufacturers who eventually dropped out of the DRAM manufacturing business. It was a difficult transition since very expensive new production equipment was needed for four new technologies to be described in order to maintain yield, cost, and performance so that the 64-kbit DRAM product could be made price-performance competitive to the 16-kbit DRAM. There was no need for new cell designs since the merged one-transistor dynamic memory cell had nearly attained the most compact geometry in the two-dimensional layout. These four new technologies and associated production equipment are given in Table 4, which also lists earlier production technologies and design rules. They are (see the 1979 listings in Table 4): i) parallel plate (instead of surface inversion) to give higher charge storage capacitance from the small area of the 1-T DRAM cell, since higher capacitance (about 32 fF to store 1 million electrons at 5 V) is needed to reduce the soft errors due to noise electrons generated by alpha particles from the package materials of the chip, cosmic rays, and other noise sources; ii) dual dielectrics for the charge storage capacitor, using a higher dielectric constant nitride layer on a thin thermally grown oxide layer to provide higher charge storage capacitance, but also to reduce pinholes in the thinner oxide; iii) plasma dry (or gaseous) etching technology to produce steeper walls or trenches in order to reduce tapers in oxide, metal, and silicon steps which take up silicon real estate (chip area); and iv) optical wafer stepper to shrink the lithographic line width or feature size from 3 to below 2 micrometers.

C. MOS Advances Driven by Production Technology and Equipment Availability

The influence of the new process technology and production equipment on the commercial success of the 64-kbit DRAM around 1984 are illustrated using geometrical size given in Fig. 16 [154]. There are three size effects. The silicon wafer area increased from three to four inch diameter in 1983 (and is now six inch in the production of 1-Mbit DRAM chip or the 80386 MPU and IBM has just moved its volume production of the 1-Mbit DRAM to 8-inch wafers [155]) in order to get more dice (or chips) per wafer to lower or maintain the cost per chip. The chip area has increased only slightly (about 1.2×) so that the DRAM can still fit into a standard 300-mil dual-in-line package. The cell area and transistor dimensions have decreased to increase the bit density and thus the number of bits on each chip or die.

The chip area given on the left vertical axis of Fig. 16 is an indirect measure of the performance since at a smaller area for a given bit per chip (the x-axis), the transistor size is smaller which means higher speed. Higher productivity is also obtained owing to a larger number of chips from a larger silicon wafer and smaller chip area. The four generation lines show the decrease of the chip area at each bit density (x axis). The decrease reflects the use of the new technology and equipment at each bit density. The first two generations of the 64-kbit DRAM chips were manufactured by pushing the older technology and equipment of the 16-kbit DRAM chip to their limits in order to get to the market first but with a lower-performance, more-costly (lower-yield) and larger-size 64-kbit DRAM chip. The third and fourth generations corresponded to the use of the four new technologies and equipment to produce the smaller linewidth and chip-size and higher-performance 64-kbit DRAM chip, resulting in higher yields, lower cost, and presumably higher profit before price erosion set in. The consequence

Table 4 Production Process and Equipment Trends (1962–1995)

Prod Year	Gate & Process	Ram Product (bit)	Design Rules Line (μm)	Design Rules Oxide (A)	Cell Type	New Equipment and Technology
1962	Al PMOS		25			
1965	Al PMOS	64	15	1500	6T-SRAM	
1969	Si PMOS	256	10			CVD Poly CVD Oxide
1970	Si PMOS	1k	10		3T-DRAM	
1971	Si NMOS	1k	10		6T-SRAM	
1974	Si NMOS	4k	8		1T-DRAM	CVD Nitride Ion Implant
1976	Si NMOS	16k	5	700	1T-M-DRAM	Double Poly for Merged DRAM
1977	Si HMOS-1	4k	4	700	6T-SRAM	
1979	Si HMOS-2	64k	3	400	1T-M-DRAM	Low Yield
1979	Si NMOS	64k	2	200	1T-M-DRAM	Parallel Plate Dual Dielectrics Plasma Etcher Wafer Stepper
1983	RSi NMOS	256k	2	200	1T-M-DRAM	Silicide/Nitride
1985	RSi NMOS	256k	1.5	200	1T-M-DRAM	Triple Poly/Aluminum
1986	RMS NMOS	1M	1.5		1T-M-DRAM	2-D Planar (Toshiba)
1987	RMS NMOS	1M	1.5		1T-M-DRAM	3-D T-C Epitaxy
1987	MSi CMOS	4M	0.8	150	1T-M-DRAM	3-D T-C Epitaxy (MA,TO)
1987	MSi CMOS	4M	0.8	150	1T-M-DRAM	3-D S-C Epitaxy (FI,HI,OK)
1987	MSi CMOS	4M	1.25	150	1T-M-DRAM	3-D T-T-C Epitaxy (TI,NEC)
1988	MSi CMOS	16M	0.5		1T-M-DRAM	3-D T-C (Matsushita) 50fF
1988	MSi CMOS	16M	0.6		1T-M-DRAM	3-D S-T-C (TO,HI) 30,33fF
1995	MSi CMOS	64M	0.25	110	1T-M-DRAM	3-D X-ray (Deep UV?)

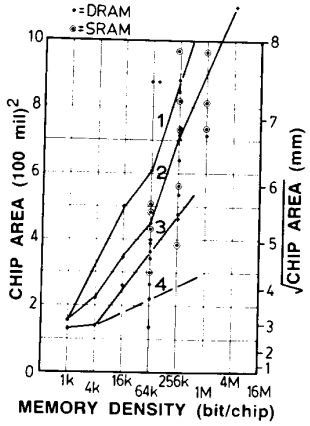


Fig. 16. Chip area versus bit-per-chip for 1- to 256-kbit DRAM cells with four generation lines related to production equipment availability or line width [154].

of the premature transition from the first to the second generation is well known. The many aggressively designed 64-kbit DRAM chips by American engineers, using smaller charge storage capacitors that stored less than 500 thousand electrons (80 fC or 16 fF at 5 V) and narrower linewidths to put more bit on one chip, could not be produced at high yields with the older 16-kbit technologies and equipment. The conservative Japanese designs, which maintained a larger capacitor, wider linewidth and larger cell, gave higher yield using old equipment and thus met the volume demands (mainly from IBM's need who lacked an in-house manufactured 16-kbit chip [97]). This gave the Japanese producers the opportunity to penetrate and eventually dominate the DRAM market worldwide. After belatedly putting the equipment of the new technologies in place to produce the smaller chips shown in the third and fourth generation lines indicated in Fig. 16, price erosion, due to reduced manufacturing cost in foreign factories mainly from learning curve and some automation, gave the fatal blow to many American manufacturers trying to re-enter the 64-kbit DRAM market. Many had to drop out of the unprofitable DRAM business altogether, giving the Japanese producers an 80-percent share of the 1985–1986–1987 market. There was an indication of trend reversal in 1987 when the federal import price control went into effect at the beginning of the year which limited further price erosion and moved the 256-kbit price up from less than $2 to about $3, which has since risen to $6–$9 in the spot market in July 1988. Several American companies have planned to return to DRAM manufacturing. Micron Technology, who sweated it out has turned profitable ($100M profit from $300M gross in 1988) [263].

The decrease in chip area observed in the four 64-kbit DRAM generations just discussed has manifested itself in the three-year product cycle of sampling-to-volume delivery of the denser DRAMs, such as the 256-kbit (1982–1985) and 1-Mbit (1985–1988) DRAMS. This is also evident in the manufacturing of the 64-kbit to 1-Mbit SRAMs indicated by the circled dots in Fig. 16 which has four to six MOSFETs per bit or cell.

For the developmental 256-kbit DRAM, refractory metal silicides and aluminum metal on double and triple polysilicon technologies have been used. For the 1-Mbit DRAM chip, the first generation samples were just a shrunk ver-

sion of the 2-dimensional (2-D) planar 256-kbit designs. In the second generation 1-Mbit DRAM, whose production volume has greatly increased in 1988 as predicted (dictated) by **MITI** of Japan, two 3-dimensional (3-D) charge storage capacitor geometries of great promise to further reduce the cell dimension are to be employed in the production cell design. In the stack capacitor design, multilayers of conductors (poli-Si or Al) and insulators (polySi-oxide or Si-nitride) are stacked on top of the MOSFET switch. In the trench capacitor design, a hole or trench is etched into the silicon and a MOS capacitor is fabricated inside the hole or trench while the MOSFET switch is still on the planar surface. As of mid-1988, these cell designs still need a significant amount of technology development to reach volume production and their implementation may be postponed to the 4-Mbit DRAM production around 1990. These are further discussed in the next two subsections.

D. Interconnect Wiring Delay and Silicide/Metal Technology

At 256 kbits, the dimension or linewidth was reduced to 2 micrometers. Smaller linewidths are used for the 1-Mbit chips (1–1.2 micrometer), 4-Mbit chips (0.7–0.8 micrometer), and 16-Mbit chips (0.5–0.6 micrometer). At these dimensions, the silicon MOSFET transistor itself does not limit the logic switching or memory access speeds but its ability to drive the capacitive load does. For example, the intrinsic speed or signal delay (known as the intrinsic gate delay) due to electrons drifting through a 2-micron length channel at their phonon-scattering limited velocity of 10^7 cm/s [56], [57] is $2 \times 10^{-4}/10^7 = 20$ ps while the cycle and access times, which I shall call the chip (or die) delay time, of a 64- to 256-kbit chip are around 100 ns which is 5000 times larger than the intrinsic gate delay. Thus, the chip delay of a high density VLSI memory or logic chip cannot be limited by the transistor's gate delay itself unless 5000 of the 20-ps gates are all in series which is not likely to occur even in a large logic chip but certainly does not in a memory chip. One possible source of a long delay in a high-bit-count (256-kbits or larger) memory chip is the capacitive loading of the resistance of the word and bit lines, especially the diffused-silicon and the poly-silicon lines if only one of the two lines is strapped by the highly conductive aluminum metal. This results in a large resistance-capacitance or ***RC*** delay of the interconnection lines (known also as the wiring delay). For example, the worst-case signal delay from one corner to the diagonal corner of a 7×7 mm^2 chip can be $(7 + 7)^2 \times 7000$ ps $= 1.4$ μs using the 20 ohms per square sheet resistance value for a polysilicon line given in Table 5. Chip designers have been able to reduce this delay by segmenting the memory cell arrays. The chip delay can be further reduced to $[2(7 + 7)]^2 \times 90$ ps $= 4.2$ ns if both lines were covered with Al metal (using the values for the Al line again given in Table 5). However, the double metal technology was not developed until the last three years for the 1- and 4-Mbit chips.

Lower resistivity refractory metal silicides are also being developed to reduce the interconnect resistance by covering the poly Si lines with refractory metals and then heating to form a silicide/poly-Si double layer. The double layers are known as polycide or salicide (Self-Aligned Silicides) in which the lower poly-Si layer is used for masking against source and drain diffusion so that the poly-Si gate is automatically self-aligned with the edge of the source and drain without having to use lithography for a precise alignment). Table 5 gives the *RC* delay time of poly-Si, silicides, refractory metals, and aluminum metal. The delay time computed is the time constant of a simple parallel plate transmission line. Thus it is the best-case estimate since the additional edge capacitances due to fringing field of a narrow conductor line are not included. In the examples, the line consists of a 1-mm-long and 1000-A-thick conductor (poly-Si, silicide, refractory metal or aluminum-metal) over a 1000-A-thick silicon dioxide film and the resistivity data of the conductor films were taken from the literature [154], [156]–[162]. In a typical VLSI chip of 200×200 square mil or 5×5 mm^2, the longest signal path may be somewhat over 1 cm and the oxide thickness is unlikely to be much more than 0.5 micron, thus increasing the delay to twenty times the values given in the right column of Table 5, about 140 ns for the poly-Si line and 1.8 ns for the Al line. It is evident from this table that to maintain or improve the chip speed as the transistor density is increased on VLSI chips, the lower resistivity silicides, refractory metals, and eventually the aluminum metal must be used in place of the doped polysilicon. Table 5 shows an improvement in chip delay of a factor of 12 from poly-silicon to $TiSi_2$, and another factor of 3 to tungsten metal or 6 to aluminum metal, giving an 80-fold improvement from poly-Si to Al-metal.

Although the single-layer aluminum conductor technology, known as 'metal line' to VLSI engineers, has been used for one of the two lines (bit and word lines) in the first 1-kbit DRAM cell in 1970, process incompatibility has prevented its use in a double-metal technology for both word and bit lines in memory chips or in multi-interconnect-layer logic chips until the last few years. This is due to the high chemical reactivity of aluminum at high temperatures (above about 500°C) which had prevented its use because of the lack of a reliable and controllable low-temperature thin-insulator technology to separate the two aluminum layers. The insulator thickness should be less than the linewidth to avoid etching deep vertical insulator walls. Double

Table 5 Interconnection Delay in Silicon VLSI Chip

Conductor Materials	Resistivity (μohm-cm)	Thickness (A)	Sheet Resist (ohm/square)	Delay[a] (ps/mm^2)
POLY-Si	1000	5000	20	7000
TaSi	46	1000	4.6	1587
Ti	42.7	1000	4.3	1484
Pd Si	32	1000	3.2	1104
MoSi	22	1000	2.2	759
TiSi	17	1000	1.7	586
Ta	13.1	1000	1.3	448
WSi	12.5	1000	1.3	448
Pd	10.5	1000	1.1	380
W	5.3	1000	0.53	183
Mo	5.3	1000	0.53	183
Al	2.6	1000	0.26	90
SiO_2	[b]	—	—	6.58 ps/mm
Vacuum	[b]	—	—	3.33 ps/mm

[a]The interconnect *RC* delay time constant is computed using a parallel plate transmission line on a zero resistance ground plane. Fringing capacitance is ignored. The oxide dielectric thickness is taken as 1000 A while normal thickness is 5–10K A. Delay/$L^2 = RC/L^2 = (\epsilon_{ox}/t_{ox})R_s = 345(R_s/t_{ox})$ (ps/mm^2) where $\epsilon_{ox} = 3.9 \times 8.85 \times 10^{-14}$ F/cm and an oxide thickness of $t_{ox} = 1000$ A are used.

[b]The delay of light in the lossless SiO_2 and in vacuum are given by delay $= \sqrt{LC}/c = 3.33\sqrt{\epsilon_{ox}}$ (ps/mm).

metal technology is required to connect the memory cells in dense chips of over 1 Mbit/chip and picosecond logic gates in order to maintain or reduce the interconnection delays. Multi metal layer processes have been used for bipolar integrated circuits for a decade. It was expensive to adapt to MOS.

The above analysis was made in 1983 for the first presentation of this review. Interconnection technology has leap-frogged the silicides. Double and even triple layer metal (Al) lines have now indeed been used in the 1-, 4- and 16-Mbit DRAM designs in which the poly-Si lines are strapped by Al layers. However, as the dimensions shrink to below one micron, the source and drain junctions of the MOSFET and the emitter junction of the BJT must be made shallower, less than 0.5 micron, to maintain the transistor's electrical performance (mainly to avoid the reduction of the threshold voltage owing to the short-channel effect). The risk of aluminum penetration short-circuiting the shallow silicon p-n junctions increases during the post-metallization anneal and die attach or chip bonding operation around 400°C. In addition, the pure aluminum line has a lower onset current density of metal electromigration, at about 100 kA/cm^2. Aluminum doped with a few percent of silicon has been employed to reduce junction shorts at the Al/Si contact and to increase the electromigration current density, but further improvements are required for the next generations of still smaller and denser MOSFETs and BJTs in ULSI chips. Renewed development efforts are being made to find an optimum silicide or refractory metal (such as tungsten [159], [161]) to overcome these two limitations [161], [162].

Many refractory metal silicides and refractory metals have been used as the interconnection bit-line and word-line during the 256-kbit DRAM development several years ago to reduce and avert the reactivity problem of aluminum. The cross-sectional views of some of the chip designs from a number of American and Japanese companies are shown in Figs. 17 and 18 [154], the latter also includes one of the first designs using double metal (Al) over double poly layers. In response to a question from a DEC engineer during the first presentation of this review on March 1, 1983, the author replied that tungsten and tungsten silicides should be the preferred silicide and refractory metal due to their chemical stability at high temperatures to give process compatibility, and their low reactivity at room and operating temperatures to give operating reliability. Indeed, tungsten seems to be the choice five years later today, as indicated by recent trade reports on the many efforts of production process development and production equipment delivery for laying down tungsten film by the chemical vapor deposition (**CVD**) process [154], [159], [161]. Reports on the characterization of tungsten silicide and its oxidation have also appeared recently for use in VLSI-ULSI logic chips where both line and contact resistances must be minimized since speed is the cardinal objective [162].

E. Superconductor Transmission Line for Interconnect Wiring

Recent discoveries of high temperature superconducting ceramics have raised the question of whether the speed limit due to *RC* delay along the interconnect lines can be improved or removed. The data given in Table 5 provide an answer since the signal delay of a superconducting metal

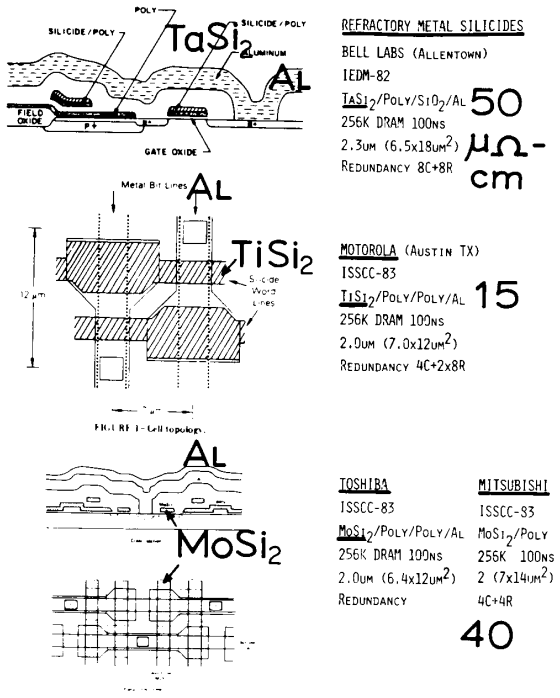


Fig. 17. Cross-sectional views of three 256-kbit DRAM cells using silicide for one interconnect line and aluminum metal for the second interconnect line [154].

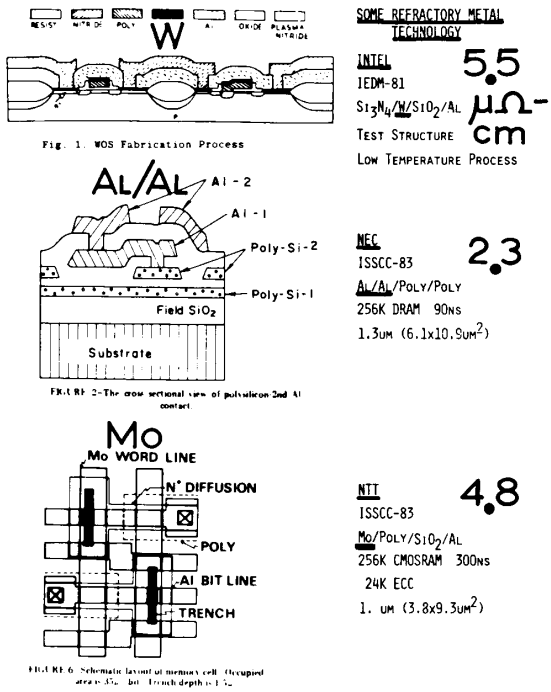


Fig. 18. Cross-sectional views of three 256-kbit DRAM cells using refractory metal and aluminum metal for the two interconnect lines [154].

line over an SiO_2 insulator would be just the ***LC*** delay or the delay of the electromagentic wave propagating through a lossless dielectric media. We again use the ideal parallel plate geometry assumed in Table 5 which disregards the edge capacitances of a narrow line and we obtain the computed delays for one transistor and four lines given in Table 6. It lists the delay of a 1-micron Si MOSFET operated at 300

Table 6 Delay of Superconducting Wires

Wire Length (mm) =			1.0	1.13	7.0
Wire Type	Delay	Unit	Total Delay (ps)		
Al/SiO_2	88.7	ps/mm^2	90	113	435
Al/SiO_2[a]	5.9	ps/mm^2	6	7.5	290
MOSFET (1 μm)	80	ps	80	80	80
Super-C/SiO_2	6.58	ps/mm	6.6	7.4	46
Super-C/Air	3.33	ps/mm	3.3	3.8	23

[a]$(T_{Fox}) \times (T_{al}) = 5000\ A \times 3000\ A$

K, two *RC* metal interconnect lines whose Al/SiO_2 layer thicknesses are (1000 A/1000 A) and (3000 A/5000 A), and two superconductor lines of Superconductor/SiO_2 and Superconductor/Air. Delays along three line lengths, 1.0, 1.13, and 7.0 mm are given where 1.13 mm is a cross-over length below which a superconductor does not provide a significant speed advantage while 7 mm is roughly the size of a current VLSI chip.

The table shows that at a wire length of 1.13 mm the *RC* delay in the thick Al/SiO_2 (3000A/5000A) line is about equal to the *LC* delay of a superconductor line on the SiO_2 dielectric, 7.5 ps. For this hypothetical example, the *RC* delay may become dominant along longer lines on larger chips as indicated by the 7-mm examples in the right column of Table 6, then, superconductor could give some improvements. This cross-over length comes about since the *RC* delay is proportional to the square of the line length (both *R* and C are proportional to length) while the *LC* delay is linearly proportional to line length since the velocity of signal propagation (or light) is a constant. Thus, resistive metal lines can give very high speeds if the signal path length is limited which means a low level of integration in small chips (below about $1 \times 1\ mm^2$ in this example). For such a small chip and low level integration, the interconnect delays between the small chips would also have to be taken into account. (See Section VIII on clock skew.)

The limited improvement from high temperature superconductors has been further elaborated recently by Gordon E. Moore of Intel and Ralph E. Gomory, senior vice president for science and technology at IBM, and their reservations were quoted in a technology preview of the year 2000 [163].

F. Three-Dimensional DRAM Cells for 1-Mbit, 4-Mbit and Beyond

A major advance in the design of the DRAM cell was made by Sunami and co-workers at Hitachi in 1982 [164] who introduced the 3-dimensional capacitance design. This was followed up by a number of Japanese manufacturers [165]–[167] as well as TI [168] and IBM [169]. The 3-dimensional design was further extended in 1985 by a group managed by Chatterjee at Texas Instrument. The new TI cell is known as the trench transistor cross-point cell. This began when various attempts were made at American and Japanese VLSI laboratories to extend Dennard's 1-T DRAM cell from the two-dimensional (**2-D**) plane geometry to the third or depth and height dimension. These earlier efforts were directed at reducing the area of the charge storage capacitor while keeping the capacitance value constant at more than 32 fF in order to hold more than 10^6 electrons at 5 V to limit soft errors. (The generally quoted critical charge or noise is 35 fC which is about 200 000 electrons or twenty percent of noise in a signal of one million electrons.) Two capacitor geometries were employed, the stacked capacitor and the trench capacitor. In the former, multiple layers of conductors and insulators are stacked on top of the gate of the switching MOSFET. In the latter, the capacitor is built inside a deep (10 micron) and small (6–9 $micron^2$) hole or trench etched into the silicon surface next to the MOSFET gate. The trench and stack capacitance approaches have used the one-micron or submicron lithographic technologies to go beyond 1 Mbit per chip since the cell area in a 4-Mbit chip must be less than about 10 square microns to keep the chip area and cost down. The submicron technology is about to be put into production for the 4-Mbit chip in late 1988 or 1989. Sunami gave a 1985 review on the trench capacitor DRAM cell developments at various laboratories [170] and Lu gave a 1987 review [171].

A novel approach was then reported by Chatterjee [172] and Texas Instruments engineers Richardson *et al.* [173] and Banerjee *et al.* [174], first at the December 1985 IEDM [173]. Their new **3-D** cell puts not only the storage capacitor but also the MOSFET switch into the trench as indicated in Fig. 19. This allows them to build a cell in an area of only 9 square micrometers using the then available 1.25 micron technology (they later called it a one-micron technology). This cell also simplifies the *x-y* interconnect topology, which further contributed to the reduction of the cell size. The TI group suggested that the 4-Mbit DRAM chips can be built

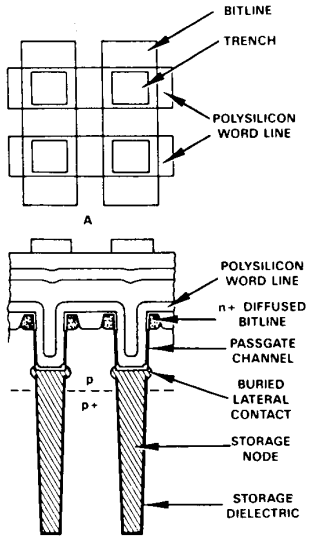


Fig. 19. The Texas Instrument three-dimensional DRAM cell known as the trench-transistor cross-point [173], which has both the MOSFET switch and the charge storage capacitor inside the trench to shrink the cell area down to 9 square microns for 4 Mbit density without using a submicron technology (1.25 micron).

with this cell design using 1 micron optical lithography. If successful, then the dependence of the 4-Mbit chip on the availability of submicron manufacturing equipment and process technology, especially lithography, could be averted for the near term. It was predicted [173]-[175] that a 4-Mbit DRAM chip using optical lithograph could reach the market in one or two years (1988), which would be much sooner than previously predicted when it was assumed that submicron lithography linewidth was necessary for a 4-Mbit DRAM chip. However, delivery of the 4-Mbit TI cross-point DRAM chip will probably not begin at the end of this year (1988) according to Chatterjee [97] and a recent survey [176]. Instead, 4-Mbit chips using submicron optical lithography and trenched or stacked capacitor cells will reach the volume market first [176].

G. Lithography Technologies for Submicron MOSFETs

The TI trench-transistor cross-point DRAM cell is very near the optimum three-dimensional cell area of four squares for maximum bit density at a given lithographic width or feature size. Further increase in density to 16 Mbit, 64 Mbit, 256 Mbit, and 1 Gbit would rely on the reduction of feature size or linewidth to below the present production optical lithography limit of about 0.7 micrometer, although the 16-Mbit DRAM chip may still be volume manufacturable using the 0.5 micrometer I-line, deep-UV optical lithography as reported at the February 1988 ISSCC [177]-[179]. Three lithographic techniques have been proposed and reviewed [180], [181]. In historical sequence, these are electron-beam [180]-[185], X-ray [180], [181], [186]-[192], and deep ultraviolet light from excimer laser [180], [181], [193]-[195].

Fabrication of chips with more than 8-kbit MOSFETs was demonstrated using electron beam by IBM in 1975 [185] and a 1-μm 4-kbit NMOS 6T-SRAM using X-ray by Bell Labs in 1983 [186]. More recently, it was estimated that deep UV from excimer laser can reach down to 0.25 to 0.3 micron linewidth but at only 25-30 4-inch wafers per hour throughput which is less than half of the 60-wafer per hour throughput of optical lithography used in current production [180], [181], [193], [194]. However, the sub-0.5-micrometer deep-UV optical lithography has the advantage over electron-beam or X-ray of using the relatively familiar current optical technologies as a base [195]. Experimental deep-UV lithography machines are becoming available [193], [194], and extensive refinements of the technology are still needed.

Below about 0.5 micron linewidth, X-ray offers probably the best potential for volume production with high throughput (sixty 4-inch or 6-inch wafers per hour) but four production technologies must be developed: i) intense X-ray source (from a synchrotron storage ring), ii) X-ray mask, iii) X-ray resist, and iv) alignment method [180], [181], [187]-[192]. Developments of X-ray technology are underway; for example, a compact 15 × 6 foot race-track-shaped synchrotron has been ordered by the IBM Fishkill development laboratory from Oxford Instruments, England, in mid-1987 at a reported cost of $16.3M to support ten to as many as thirty step-and-repeat stations, for submicron logic and memory development work [196]. This huge cost further reinforces the Gordon Moore criteria: recovery of equipment and development costs is the prime consideration for the manufacturing and volume delivery of future generation of submicron VLSI and ULSI chips. X-ray lithography could produce chips down to 250-A linewidth; however, the first volume product at 2500 A or shorter channel length probably will not reach the marketplace until 1995 or later due to the many necessary technology developments even if it is not further delayed by future 5-year cycles of market and economic conditions. In June 1988, IBM Federal System Division proposed a multi-year effort to the Naval Research Laboratory to develop a 0.25 micrometer linewidth **ASIC** (Application Specific Integrated Circuit) gate array technology using X-ray lithography [197].

The anticipated long delay of the X-ray lithographic technology has prompted several manufacturers to turn to the direct-write electron-beam lithographic technique for the interim fabrication of sub-0.5-micron chips since electron-beam technology requires only the development of the electron-beam resist instead of all four technologies required by X-ray and a good electron-beam resist has been available. However, wafer throughput is only about 12 per hour for 4-inch wafers with low-complexity circuits [180]-[182] which could still be economical for producing the higher priced high speed logic arrays of ASIC of 2500 A or smaller feature sizes for mainframe and supercomputer CPUs as well as real-time graphic controllers. For very complex circuits, throughput can drop to less than one wafer per hour. In April 1985, IBM discussed a 0.5 micron Al-line direct-write electron beam NMOS technology for a 16-Mbit DRAM chip with 8.5 μm^2 cell area and for a 100 kbit-gate logic chip at one-volt power supply voltage [198]. In December 1986, it was reported that IBM had contracted Grumman Corporation to build 40 to 50 direct-write electron-beam lithography systems [199] based on the IBM in-house design, EL3 [182]. And in late 1987, a successful fabrication of a 1000A channel N-MOSFET operated at 77 K was reported by IBM [200]-[202] that had a 10-ps gate delay. Using the sub-0.5-micron lithography and plasma etching technologies to fabricate the trench capacitor and trench transistor cell, 64-Mbit DRAM chips could be possible but it cannot be made economically with electron-beam lithography due to its very low throughput, only with X-ray or possibly excimer laser optical lithography.

H. A Recapitulation of the Berglund-Moore-Intel Scenarios

During 1981 to 1982 when Neil Berglund gave his talks under Intel's College Seminar Program [144], he showed that the MOS VLSI advances had been driven by the availability of manufacturing equipment and technology. Two years later in 1983, excess production capacity from both domestic and Asian producers had turned it into a market driven business, as pointed out by Gordon Moore [145]. Although technology development and production equipment have continued to advance rapidly toward submicron production of VLSI chips using light, electron-beam, X-ray and deep ultra-violet (excimer laser light) lithographies and plasma dry etch, the market scenario of supply and demand of Gordon Moore continues to be the driving force today. Profitability or the recovery of equipment investment has continued to determine the volume production and delivery dates [176], although yield was the cause of the recent delay of the 1-Mbit DRAM chip. Volume delivery of DRAM has followed a three year cycle. For example, the laboratory 1-Mbit DRAM chip was first described at IEDM and ISSCC

in late 1984 and early 1985, but volume delivery for personal computer use had not begun in high gear at the end of 1987 although manufacturing capacities were already in place. The delay was designed to allow the recovery of the manufacturing capital equipment costs of the previous generation (256 kbits). The delay further helps the recovery of the equipment costs to produce the current generation (1-Mbit chip) by premium initial pricing to give high profit margin before substantial price drop when volume delivery begins.

I. The Nonvolatile Ferroelectric MOS (FMOS) RAM—A Replacement for the Magnetic Disk and Main Memory?

A most exciting development was reported at the last IEDM in December 1987 on a new nonvolatile MOSFET DRAM memory cell by Kinney, Shepherd, Miller, Evans, and Womack [129], and at the February 1988 ISSCC by Sheffield-Eaton, Bulter, Parris, Wilson, and McNeillie [133]. They succeeded in using the electrical polarization of a ferroelectric capacitor to store the information bit semi-permanently or nonvolatilly, getting experimental write speeds of 200 ns [132] and 60 ns [133] which are nearly 100 times faster than a EEPROM (1 ms) or UV-EPROM (10 ms) and without fatigue after 10^{12} write cycles [129], [130]. The projected operating life is 10^{15} read/write cycles which is 75 years at a cycle time of 100 ns [132]. (See also Section IV-P.)

This FMOS cell is a potential replacement of the DRAM cell. It could also provide some simplification of the system design, since ferroelectric polarization retention is nearly perpetual just like the magnetic core memory so that refresh is not needed.

The author has played the acronym game of successful simplification with the following: FECDRAM, FCDRAM, FERRAM, FERAM, FDRAM, FRAM, FECHMOSFET, FERFET, FEFET, FFET, FECMOS, FEMOS, and FMOS. FRAM and FERAM or any of its derivatives are appropriate only for the memory chip or array containing many cells. This can also be applied to the two possible versions, the FDRAM and FSRAM, but they are not appropriate for the individual cell or memory circuit. By limiting an acronym to four characters, **FMOS** was picked as the winner for the individual memory cell since the cell contains a ferroelectric capacitor (for F) as the load or charge storage element and a MOS transistor gate to sense and write. **FRAM** is picked as the acronym for the memory chip containing very many FMOS cells. It can be extended to distinguish the ferroelectric version of one-transistor cell of Dennard, FDMOS cell, and the six-transistor static flip-flop, FSMOS cell. The FMOS choice is analogous to NMOS, PMOS or CMOS which all designate a circuit type, for example, N stands for an N-channel MOSFET as the load of the gate which is also an N-channel MOSFET while C stands for the Complementary or P-channel MOSFET as the load of the N-MOSFET gate. The original ferroelectric MOSFET of Ian Ross [32] using a ferroelectric material as the gate would be a MFSFET.

Looking ahead, the FRAM (**FDRAM** and **FSRAM**) memory arrays using the FMOS cell have the potential not only to replace the hard and floppy magnetic disks but also to eliminate one level of memory hierarchy from today's computer systems if high yields can be attained at the **WSI** (Wafer Scale Integration) level to produce a four- or six-inch silicon disk whose FMOS cells are connected and routed to the contact pads at the silicon disk's perimeter. Defect elimination of a WSI FRAM is not so critical and designed-in redundancy is not required since 'bad tracks' can be deleted by the first-time user or by distributor before shipment following the current practice of formating the hard (and floppy) disk prior to first use. Such a **FMOS RAM Disk** or **FRAM Disk,** hard or soft (silicon thin film on flexible substrate?) and removable or fixed, not only could provide increased reliability compared with the present magnetic disk drive since the silicon FRAM disk has no moving parts, but also substantially higher read/write/access speed, approaching the main memory speed (50 to more than 100 times faster than the current magnetic disks). The availability of such a FMOS disk near the main memory speed and at the density of a secondary (disk) memory should significantly if not drastically influence the design philosophy and architecture of future generations of computers.

It seems that the development and volume production of the FRAM disk or some other form of high speed and high density nonvolatile main memory is inevitable for the stake is too high. The FMOS and FRAM have the potential for success since the present silicon MOS VLSI technologies are all applicable to the FMOS, and the costs of manufacturing the silicon part of it, the DRAMs, are dropping continually. Only the fabrication technology of the ferroelectric thin film on silicon and silicon dioxide substrates require further development and engineering. But this should not be very different from nor more difficult than the well-established magnetic thin film disk technology. The combination of the ferroelectric technology with the Si MOS technology should also not be difficult since much experiences have been accumulated on combining the refractory metal and silicide technology with the silicon MOS technology.

Based on the price per bit cross-over point, the demise of the Winchester magnetic hard disk drives owing to the lowering Si DRAM price has already been overly optimistically extrapolated to occur in 1990 for the 14-inch disk and 1993 for the 5.25-inch disk [203]. The timing may be a decade or more off [97] since magnetic disk technology is also advancing and new operating system software requires long development time. The nonvolatile FMOS and the FMOS disk will exceed not only the speed advantage of the DRAM over the magnetic disk, but also has the nonvolatility of the magnetic disks not provided by the DRAM. Such a revolutionary advance in main memory technology has already occurred once before when the volatile semiconductor (Si-DRAM) memory [140], [141] replaced the nonvolatile magnetic core [204] in the IBM 370/158. This second revolution will reverse the volatility characteristics again but will have an even greater impact owing to the possibility of replacing and even eliminating the secondary mass storage memory, the magnetic disk, and also having a nonvolatile main memory with higher speed and much larger capacity and density than the magnetic core. The elimination of one layer of computer memory hierarchy could alter the design philosophy of computers significantly as von Neumann [205] once stated while searching for a memory (in 1946), "Ideally one would desire an indefinitely large memory capacity such that any word would be immediately available." Ideally maybe, but in practice there will probably still be a memory hierarchies to improve access time [97].

After I completed writing the above, I received on January 28, 1988 a copy of a forecast made by a member of my 1962

Fairchild team on MOSFET development, Howard Z. Bogert [82], who has recently turned his talent to reporting and forecasting on the semiconductor industry. His enthusiasm [206] on the FMOS is no less than mine and his independent assessment is similar to mine but short of my bold prediction that the FMOS can appear in the form of a FRAM disk and the FRAM disk will replace both the hard disk and the main memory to incite another revolution of computer design.

VI. Performance Trends of Silicon MOSFET

There are three factors which must be considered when comparing silicon MOS transistors and integrated circuits with other transistors such as the silicon bipolar transistor and the compound semiconductor transistors. One of these is the ultimate speed of the transistor in an integrated circuit configuration with load and parasitic capacitances, not just the intrinsic transistor itself, since transistors must be interconnected to perform information processing functions. The signal delay from the parasitic resistances and capacitances of the interconnection lines must also be included since it will contribute to the chip delay of future chips as we have already discussed in Section V-D. A highly optimistic best-case minimum delay can be estimated by assuming a single load, i.e., one transistor driving an adjacent transistor on the chip such as in a ring oscillator, and disregarding the inter-transistor *RC* line delay. The second factor is the reliability, operating life, or wearout rate of the transistor and the interconnection wire. The third is the manufacturing yield of the many-transistor chip which partially determines the cost. In certain applications, such as military electronics, the cost may not be a primary concern but reliability is, which is related to yield. However, all three factors are important in most consumer, commercial, industrial, and other nonmilitary applications, while cost at acceptable reliability is the most important in consumer applications.

We shall focus our comparison only on the first factor, which is determined by the fundamental material properties and device physics. The second factor, reliability or operating life, is also determined by fundamental device and material physics, but the failure rates have been analyzed only empirically using statistical failure data on production chips, and a fundamental base is incomplete which is necessary to review the reliability based on material and device physics. Some empirical engineering black-box approaches have been used for quick fixes of reliability problems, for example, on the drift of the Si MOSFET characteristics due to the hot or energetic electrons accelerated by the high electric field which is present in micron and submicron MOSFETS. Basic research on the failure mechanisms as well as failure modeling and prediction has just begun.

The chronological advances of the Si MOSFET performance are illustrated in Figs. 20 and 21. The two declining straight lines in Fig. 20 give the change of the lithographic linewidth or feature size of the transistor with time, decreasing at the rate of about one decade in twenty years. The upper straight line is drawn through the solid dots each representing an announced production DRAM chip and all have used optically lithography. The lower straight line is drawn through circles each representing a laboratory or

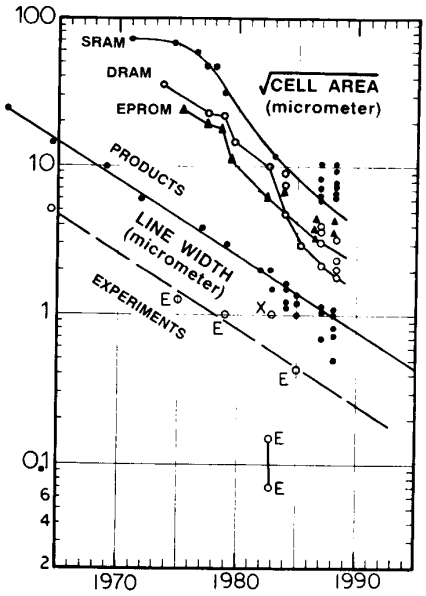


Fig. 20. Trends of the laboratory and production lithographic line width (two lower straight lines) and of the cell size (square root of cell area) of production SRAM, DRAM, and EPROM cells [154].

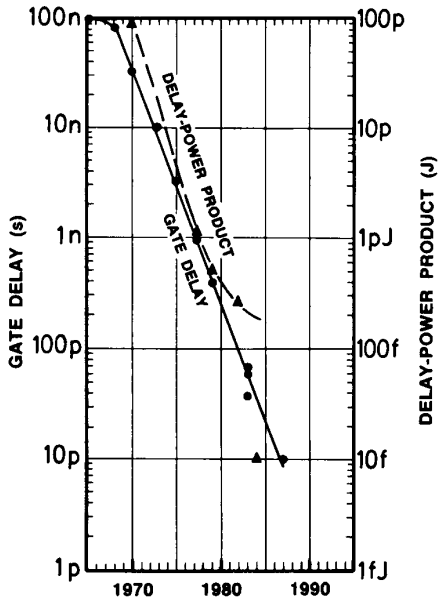


Fig. 21. Trends of gate delay and delay-power product of Si MOSFETs [154].

experimental test transistor or a chip containing several transistors. The test transistor and chip have employed one of the three lithographic techniques, optical, electron-beam (labeled *E*), and X-ray (labeled *X*). Notice that there is almost a ten-year delay from the first experimental demonstration of a technology to the engineering sampling of a product using that technology as indicated by the ten-year horizontal separation of the two declining straight lines. For

example, a 1.2-micron 8-kbit RAM was fabricated by IBM in 1975 using the direct-write electron beam (**EB**) method [185] but Si chips with 1.2 micron size transistors did not reach volume production until 1985 and not by the EB method but still by optical lithography. A similar ten-year development and engineering effort may be anticipated for the X-ray technique to produce sub-0.5-micron chips. The first X-ray chip was a 4-kbit 5-ns SRAM reported by O'Connor and Kushner of Bell Labs in 1983 [186]. At the present stage of development with most of the four X-ray technologies still to be developed for the factory, volume production using X-ray lithography probably will not occur before 1993 according to the history of the ten-year laboratory-to-production delay.

The three upper curves in Fig. 20 give the size of the square root of the area of the three types of memory cells. The top upper curve is for the static RAM or SRAM cell using either the four-transistor-plus-two resistor (polycrystalline silicon resistors) or the six-transistor CMOS flip-flop which is becoming dominant. The middle curve is for the DRAM chip using the one-transistor (N-MOSFET)/one-capacitor Dennard cell, first in the planar 2-D and now in the 3-D configuration with stacked, trenched or stacked-and-trenched capacitor and trenched transistor. The lower curve is for the UV-EPROM cell using the floating gate buried in the gate oxide. The relative level of these three curves reflect the complexity of the memory cells. The EPROM cell has dropped to 9 square micrometers in a 1-Mbit chip in 1987 from the conception size of 600 square micrometers when Frohman-Bentchkowsky first built a 2-kbit chip in 1976. The size of the DRAM cell has similarly decreased from 1000 square micrometers in the Intel 4-kbit chip in 1973 to 4 square micrometers in the 16-Mbit test cells reported at the IEDM-87 in December.

The TI 3-D DRAM cross-point cell with both the capacitor and the transistor in the trench first described in 1985 [173] had produced a sharp drop of cell dimension to 9 square micrometers using the 1 micron optical technology. Submicron technology applied to the TI trench transistor cell could steepen the drop of the cell dimension.

The second indication on the performance trend of the Si MOSFET is given in Fig. 21 which shows that the gate delay and the power-delay product also decrease when cell size and transistor dimension are reduced, the latter as given in Fig. 20. Some of the experimental data points on the continuously decreasing straight line are the "gate" (i.e., ring oscillator) delay calculated from the maximum frequency of oscillation of small test circuits with tens of MOSFET connected as a ring oscillator. This gate delay appears to decrease at about 2 to 2.5 decade per 10 years, nearing 10 ps in a 1000-A n-channel Si MOSFET at 77 K reported by IBM in mid-1987 [200]–[202]. The other slightly curved line in Fig. 21 is the figure of merit, power-delay product (power-dissipation times gate-delay). Assuming that a gate in a CPU or logic array is switched only about 30 percent of the time [92], [207], then, a factor of 3 improvement could be expected of CMOS over NMOS since CMOS draws essentially no standby power when not switching while NMOS draws about 10 percent of the maximum current when not switching and about half of the NMOS gates are in the on or maximum current state. Higher speed and lower power-delay product can be achieved with smaller feature size due to lower capacitance and current. Additional improvement of the power-delay product can be realized by reducing the dc supply voltage at the expense of a smaller noise margin. The smaller feature size is illustrated by the 10-fJ power-delay product of the recently reported submicron CMOS and the 1000-A–10-ps NMOS shown in Fig. 21. The data given in the figure show that the improvement in gate delay tracks the reduction in feature size and it should continue as electron-beam and X-ray technologies reach production for sub-1000-A transistors.

VII. Comparing Silicon with Gallium-Arsenide

There have been many review articles as well as spirited discussions and headline news in electronic trade journals on the merit of silicon versus gallium arsenide and on the silicon bipolar versus silicon MOS. We summarize and analyze some of the key points.

A. Silicon MOSFET versus Silicon BJT

The principal early argument in favor of silicon MOS over BJT is its ease of larger scale or higher level of integration due to the fewer lithographic alignment and high temperature steps. These are the results of the topology of the MOS transistor itself: It is on the surface, i.e., planar and lateral, and its source and drain junctions are self-isolated. These features ease the interconnection of many transistors. The fewer processing steps is illustrated by Figs. 22 and 23. The cross-sectional views in Fig. 22 give a comparison of the

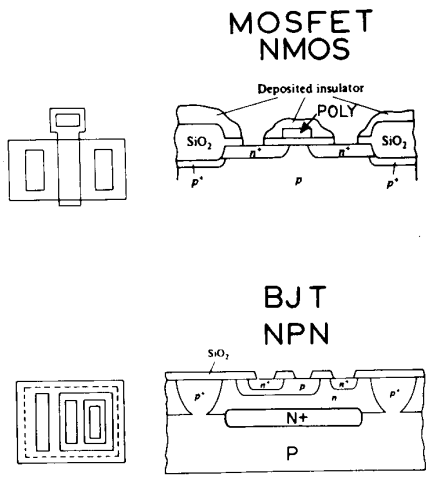


Fig. 22. Comparison of the structural complexity of MOSFET with BJT. (From D. A Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Fig. 1.14(f) on p. 15 for BJT and Fig. 1.15(b) on p. 17 for MOSFET, McGraw-Hill Book Co., New York, 1983.)

structural complexity of a MOSFET with a BJT. It is evident that the MOSFET requires fewer steps to fabricate. A second comparison is given in Fig. 23 for a CMOS with a textbook **C-BJT** (complementary BJT gate). The early processing simplicity advantages of MOS over BJT are diminished to some extent in the recent choice of CMOS as the technology leader since the processing steps of fabricating one of the two MOS transistors in the CMOS are similar to that of a bipolar n-p-n or p-n-p transistor. For examples, the n-chan-

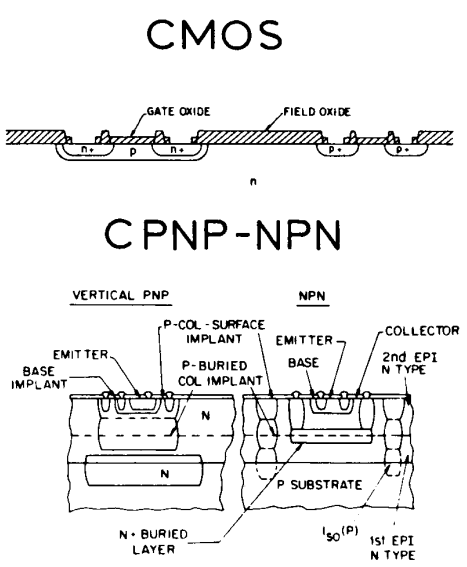


Fig. 23. Comparison of the structural complexity of CMOS with CBJT. (From A. B. Glasser and G. E. Subak-Sharpe, *Integrated Circuit Engineering*, Fig. 6.15 on p. 271 for CBJT and Fig. 6.21(g) on p. 282 for CMOS.)

nel MOSFET on the left in Fig. 23 in a p-tub requires oxide masking and impurity diffusion steps similar to the n-p-n transistor, and so is the complement, a p-channel MOSFET in an n-tube shown in Fig. 6(c). However, the increased complexity of the current CMOS technology has approached if not exceeded that of the bipolar technology.

The dominance of MOSFET over BJT in cost and density is also evident in production statistics. After the initial lag behind the BJT in the early 1960s due to surface problems, the silicon MOSFET has dominated the integrated circuit marketplace since 1970 and has been the overwhelming leader in both unit volume of production and annual as well as cumulative dollar of sale. The latest trend towards the use of CMOS instead of NMOS, especially for the peripheral drivers on the previously all-NMOS DRAM chip [208]-[214], has further lowered the standby power dissipation in comparison with BJT memory chips. Developments were recently directed to combine the BJT and NMOS known as the BIMOS (Bipolar-MOS) [215] and now the efforts have turned to combining BJT and CMOS or the BICMOS [151], [216]-[220]. For 1-ns small area gates, the BIMOS gate achieves speed advantage at 2 fan-outs while for large area gates this occurs at 6 fan-outs. In the BIMOS and BICMOS RAM chips, the BJT is used as the peripheral driver owing to its higher transconductance which can drive larger capacity loads while it is used as inter-gate driver in logic chips and more recent RAM designs. The combination of BJT and MOS has nearly eliminated the advantage of the larger output driving capability of an all-bipolar circuit over a MOS circuit. However, all-bipolar **ECL** (Emitter-Coupled Logic) gate array and cache (or buffer) memory chips have continued to be the choice in very high speed memory and logic applications, such as in the CPU of mainframes and supercomputers since ECL gates still lead in speed over NMOS, CMOS, BIMOS, and BICMOS gates. Cooling is the main concern in ECL chips due to the higher power density of the all-bipolar ECL circuits.

A comparison of the speed of two silicon and two compound semiconductor transistors was given by Solomon of IBM [221], who selected four transistors as the basis for comparison. These are the silicon MOSFET, the silicon BJT, the GaAs MESFET (Metal Semiconductor Field-Effect Transistor), and the GaAs **HEMT** (High-Electron Mobility Transistor also known as the **MODFET** or Modulation Doped Field-Effect Transistor). Two linewidths were considered, 1 and 0.1 micrometer. The latter was near the theoretical limit to give the highest speed. The calculated speed was based on the best assumed transistor geometry which minimizes the loading delay from the junction capacitances of the transistor and the following transistor load and from the parasitic overlap capacitances. The gate insulator and base thicknesses as well as the applied voltage were also optimized. The intrinsic (t_i) and total (t_t) delay times, unloaded and loaded capacitances, as well as the small-signal and large-signal transconductances (g_m and G_m) are summarized in Table 7. The intrinsic delay is the signal delay along the FET channel (1.0 or 0.1 micrometer length) or through the BJT base layer. It is about equal to the reciprocal cutoff frequency of the transconductance loaded by the active capacitances. The total delay is the minimum practical gate delay, including capacitance loading from the preceding (or drive) and the following (or load) transistors and parasitics.

These results are revealing. Compare first the Si BJT and Si MOSFET at 1.0 micrometer linewidth. The loaded or total delay in the MOSFET (81.4 ps) is 4.5 times longer than its intrinsic delay (18 ps). In Si, BJT, it is only 2.3 (23/10 = 2.3) times longer. This difference is mainly due to the smaller transconductance of the MOSFET (G_m = 0.16 mS), or for that matter any FET, compared with the BJT (G_m = 2.5 mS). However, this MOSFET or FET deficiency can be largely overcome by the BIMOS or BICMOS [151], [216]-[220] which uses the high transconductance BJT to drive the load capacitance from the next stage or gate, provided the intra MOSFET-BJT coupling has minimum capacitance loading delay.

Near 0.1 micrometer, there was essentially no difference in the total delay between the Si BJT (6.0 ps) and the Si MOSFET (6.1 ps). This is due to the smaller input or oxide gate capacitance of the MOSFET (0.85 fF) compared with the large input capacitance (15 fF) of the forward biased emitter junction of the BJT which offsets the smaller MOSFET transconductance.

B. Si versus GaAs Transistors

Comparing silicon with the two GaAs transistors in Table 7, we see that at 1 micron and room temperature, Si BJT (23 ps) and GaAs MESFET (22 ps) have comparable total gate delays while Si MOSFET is slower (82 ps). Operated at 77 K, the 1-micron HEMT-MODFET is about a factor of two faster (11 ps) than the room-temperature operation of the Si BJT or GaAs MESFET if we assume that velocity saturation overshoot occurs in the HEMT at 77 K. If it does not, then the gate delay is twice larger (22 ps) for the HEMT-MODFET at 77 K which is equal to the Si BJT (23 ps) and GaAs MESFET (22 ps). There was early reservation on such a comparison at two different temperatures since reducing the operating temperature of the Si BJT and the GaAs MESFET could significantly increase the speed or decrease the gate delay due to the higher electron mobilities and saturation velocities at lower temperatures. However, recent theoretical and

Table 7 Comparisons of Semiconductor Devices for High-Speed Logic

Channel Length Base Thickness[a]	Operation Paramater[b]	Silicon BJT	Silicon MOSFET	GaAS MESFET	MODFET or HEMT (77K)
1.0 μm	V (V)	0.4	2.5	2.0	0.5
1.0 μm	I (mA)	0.5	0.2	0.4	0.26
1.0 μm	P (mW)	0.2	0.5	0.8	0.13
1.0 μm	g_m (mS)	19.2	0.4	0.7	1.44
1.0 μm	G_m (mS)	2.5	0.16	0.35	1.03
1.0 μm	C_i (fF)	192	7.2	7	7.5
1.0 μm	C_t (fF)	57.3	13	7.6	11
1.0 μm	t_i (ps)	10	18	10	5.2
1.0 μm	t_t (ps)	23	81.4	21.6	10.7
1.0 μm	$P*t_t$ (fJ)	4.6	41	17	1.4
0.1 μm	V (V)	0.4	0.6	0.5	0.3
0.1 μm	I (mA)	0.5	0.04	0.08	0.06
0.1 μm	P (mW)	0.2	0.024	0.04	0.02
0.1 μm	g_m (mS)	7.6	0.24	0.69	0.88
0.1 μm	G_m (mS)	2.5	0.14	0.22	0.48
0.1 μm	C_i (fF)	19	0.33	0.29	0.18
0.1 μm	C_t (fF)	15	0.85	0.58	0.53
0.1 μm	t_i (ps)	2.5	1.4	0.42	0.20
0.1 μm	t_t (ps)	6.0	6.1	2.6 (3.6)	1.1(2.2)[c]
0.1 μm	$P*t_t$ (fJ)	1.2	0.15	0.10(0.14)	0.02(0.04)[c]

[a]BJT base thickness is 0.1 micron plus 0.2 micron of collector depletion layer instead of 1.0 micron; and 0.023 micron plus 0.05 micron instead of 0.1 micron.
[b]V = voltage swing; I = average; $P = IV$; $t_i = C_i/g_m$ = intrinsic delay; $t_t = C_t/G_m$ = gate delay with one load.
[c]The larger figure is obtained if saturation velocity overshoot is omitted.

experimental work [222] on Si BJT at 77 K have shown a decreasing in speed as the temperature is lowered due to the deionization of the dopant impurity and lowered mobility from impurity scattering of the electrons.

Low temperature operation to demonstrate a significant increase of speed of MOS transistors was known for twenty years [223] and a system demonstration was reported in December 1985 by DEC engineers [224] who dipped a PDP-11 CPU board into liquid nitrogen (77 K) and observed an increase of the maximum clock frequency from 20 to 40 MHz. The cooled silicon chip in this case was a 3-micron Si CMOS 32-bit CPU. The increase in speed of the CMOS chip is about a factor of two when lowered to 77 from 300 K which is extremely valuable in computer design. This comes mainly from an increase of the mobility and a smaller increase of the saturation velocity of the electrons in the n-type channel. At 3-micron linewidth, the increase of speed from the lower resistance of the interconnect lines is small. Other advantages of operating at 77 K are the increased reliability and better heat dissipation. There have been several studies on the performance of silicon integrated circuit at 77 K since this recent publicity including the 10-ps 1000-A NMOS reported by IBM in mid-1987 [202] and a number of reports at the 1987-IEDM last December for both Si CMOS and BJT [222].

A theoretical limit to maintain the transistor characterisitics and to reach the best performance was estimated to be about 0.1 micron by Solomon [221] and is given in Table 7. The total gate delay of the GaAs MESFET at 300 K, 3.6 ps, (we omit the electron velocity overshoot in GaAs) and the 0.1 μm limit is 40 percent smaller than the gate delay of the Si MOSFET, 6.1 ps. This is mostly due to the higher electron mobility and saturation velocity (from lower electron effective mass) in GaAs resulting in higher transconductance to drive the capacitive load. The 2.2 ps delay of the HEMT (High Electron Mobility Transistor) at 77 K should again be compared with the speed of the other transistors (MOST only since BJT does not work well at 77 K) at the same temperature, 77 K instead of 300 K and the differential would diminish.

In this comparison, different technologies are compared instead of comparing just purely materials and transistor types (MOS versus BJT). Thus, higher mobility should in principle be achievable in BJT (as well as MOST) if highly doped impurity superlattices are grown on the thin silicon layers by the same molecular beam epitaxy technique that has been used to fabricate the high-electron-mobility GaAs transistors, HEMT. Electron mobility in silicon would be higher in an impurity superlattice since random impurity scattering would be eliminated. Recently, Morkoc and Solomon [225]–[227] have tabulated the gate delay data of the recent (up to 1983) laboratory MODFET-HEMTs. The fastest at 300 K was reported by Rockwell at 12 ps. They also projected that gate delays will drop to less than 10 ps in MODFETs within a few years. However, a 10-ps gate delay has already been demonstrated in a 1000-A (0.1-micrometer) Si NMOS at 77 K [200]–[202], with a technology which has demonstrated high-yield capabilities at larger (1-micrometer) linewidth.

C. Synopsis of Si versus GaAs

The main message of the above comparisons is that there is not that much speed advantage of GaAs over silicon even when the state-of-the-art expensive GaAs technology of

seemingly perpetual minuscule yield is used to compare with the matured, high-yield and low-cost silicon production technology. Comparing isolated transistors at 0.1 micrometer with no wiring delay, the GaAs MESFET is slightly faster than the Si BJT, by about a factor of two. bncreasing the integration scale to greater than 1000 gates rapidly diminishes this GaAs FET speed advantage due to the interconnect wiring delays [218] and the low transconductance and low drive capability of the FET. (See [264] for a recent analysis of Si versus GaAs.)

It should also be noted that the 23-ps Si BJT and 82-ps MOSFET speeds at 1 micron and 300 K have not been attained by the clock frequencies of the logic gates and memory access times of today's production Si VLSI chips, which are in the 10- to 100-ns range or about 1000 times larger. The principal limiting factor is not the transistor itself but the signal delay due to loading the MOS transistor by the capacitive load and the interconnect lines or wires which interconnect the transistors, factors not taken into account by the computed delay times given in Table 7. Probably most of the future application needs can be satisfied by the Si MOSFET and Si BJT and the BICMOS when the interconnect and loading delays are reduced and the computed individual transistor gate delay listed in Table 7 and the experimental state-of-the-art data shown in Fig. 21 are approached in volume-produced full-size integrated circuit chips.

Finally, Tables 5 and 6 also gave the limiting propagation delay of light which is 3.33 ps/mm in vacuum and 6.58 ps/mm in a lossless or superconducting silicon dioxide light guide. So just to increase the transistor speed alone (6 ps or less) would not increase the chip speed. Optimum routing of the interconnection wires, innovative or novel interconnection schemes, new logic and chip organization, and even new computer architectures must also be considered in the design of the next generation higher density and larger area ULSI chips with many million gates, memory cells, and transistors.

VIII. Research Needs

One of the most pressing fundamental device and circuit problems in ULSI (1M logic gates or 1 Giga memory bits) and larger chips with more than 1 square centimeter chip area is the random signal delay known as the timing or clock skew [228]. This noise comes from the very nature of the computer signals which are random in time and follow different and random paths along the interconnecting conductor networks over different parts of a large chip. Computer-aided design has been used extensively to optimize routing and critical paths at the current technology level with subnanosecond delay per gate. Various novel methods have been suggested to improve or provide better clock synchronization on all parts of a large-area ULSI chip. The most intriguing is the use of an optical clock distribution system [228]. The most attractive, based on our experience in monolithic integrated circuit technology, is the use of semiconductor-insulator optical waveguides on top of a silicon integrated circuit chip as reviewed by Goodman [228]. As one can see from Table 5, the delay of the light signal (3.3 ps/mm) is about 30 times smaller than the *RC* delay of aluminum metal wire (90 ps/mm^2). A possible very-high-speed ULSI chip could be one using **MBE** (Molecular Beam Epitaxy) [229]–[232] to fabricate strategically located GaAs light emitters and optical interconnecting waveguides over a very high-speed silicon integrated circuit chip. The GaAs light emitters over the entire surface of the large silicon chip can then be synchronized to provide picosecond timing in order to overcome timing or clock skew. GaAs on Si MBE technology has been under intensive investigation in many laboratories for light emitters and junction lasers [230]–[232], MODFETs [233], [234] MESFETs [235], [236], and **HBJTs** (Heterojunction Bipolar Junction Transistors) [237], [238]. The principal problem has been the growth of a defect-free GaAs layer on a silicon substrate due to lattice mismatch. Growth of GaAs on a tilted silicon surface to provide lattice match and lower interfacial disorder and defects have shown considerable promise recently [237], [238]. A review of the status of GaAs on Si was given by A. Y. Cho at the IEDM-87 in December [239] and some recent results [240], [241] were given by TI engineers Richard J. Matyi and Hisashi Shichigo [242].

Aside from comparing the transistor performance in different device types, circuits, materials, and process technologies discussed above, two other factors, reliability and manufacturing yield, must be considered in a realistic assessment of the application merits of future transistors. The reliability issue is probably the more important of the two but one would expect the manufacturing yield to be closely related to reliability since a higher yield would imply fewer weak transistors about to fail and hence a longer and more reliable operating life. Production yields of Si VLSI logic gate chips containing a few hundred thousand MOSFETs, such as the 256-kbit DRAM, have routinely exceeded 50 percent approaching 90 percent in some unconfirmed private claims. However, the densest GaAs integrated circuit chip on an all-GaAs substrate, containing about 1000 gates or a few thousand GaAs MESFETs, had a yield of less than 1 percent in 1984 [243] and the yield has only marginally improved in the last two years [244]–[246] mainly from lower crystal defects, with the latest report of 1.3-percent yield with a total of only seven fully functional 16-kbit GaAs SRAMs [247]. Notwithstanding, a next generation supercomputer planned to be delivered during 1989–1990, the CRAY-3, has committed to use depletion mode GaAs MESFET chips of 300–500 gates and 1-kbit SRAMs [248]–[251], but delay due to GaAs chips has occurred [265]. There was a recent report that an IBM mainframe may also use GaAs in 1995 [266].

There is a steady growth of the application of GaAs light emitting and laser diodes as well as single high-speed and microwave transistors in hybrid integrated circuits [252], [253] with worldwide 1987 sales of $2.1B, of which $1.6B were light emitting diodes and almost none ($17M) in GaAs digital integrated circuits. A growth in high speed and military applications of GaAs digital integrated circuits should be expected in view of the existence of more than one dozen GaAs manufacturers [252], [253] and the soon to be available thin-film GaAs on Si-substrate wafers [251].

Even though silicon integrated circuits have attained high yield and reliability, engineering design rules on yield and reliability have been highly empirical and statistical. As the transistor dimension decreases below one micrometer and gate or bit density increases further, intrinsic failures due to fundamental electronic conduction phenomena become increasingly important. Examples are the creating and charging of electronic [87], [254]–[256] and protonic [257],

[258] traps (due to moisture and processing chemicals [259]) in the presence of the high electric fields in submicron transistors, known as the hot electron and hot hole effects. Reliability can no longer be accurately and broadly characterized by the traditional empirical and statistical approach. Fundamental understanding of the failure mechanisms is necessary to provide physics-based design rules for reliable operation of the VLSI-USLI chips over wide ranges of transistor sizes and operating conditions (voltage and temperature). One of the major emphases in the current basic research in silicon VLSI integrated circuits is on the atomic or chemical origin of transistor failure and on the modeling of the failure mechanisms in order to provide computer simulators for predicting aging and failure of the integrated chips and the systems using the chips. The basic research on the aging and failure mechanisms has been sponsored by the American silicon integrated circuit manufacturers through the Semiconductor Research Corporation [260]. A main objective is to bring the art of reliability prediction from random statistical analysis to a quantitative engineering science.

IX. Concluding Remarks

Tremendous advances have been made in silicon integrated circuit technology during the last 20 years, beginning about 1968, which have now given us the ability to produce electrical circuits approaching biological dimensions: the size of a large molecule of a few tens to a hundred atoms. This review has focused on a historical survey of one particular solid-state device, the silicon MOSFET, from its invention 60 years ago by Lilienfeld in 1928 to the latest development on the 16-Mbit submicron DRAM memory chips. Today's production lines have given very low cost, affordable integrated circuits, at about one millicent per bit and still decreasing: −$0.50 for a 64-kbit memory chip and $2 for a 256-kbit memory chip before the **FMV** (Foreign Market Value) restriction imposed by the U.S.-Japan Semiconductor Trade Agreement took effect in January 1986. However, the full capabilities of the present silicon integrated circuit technology have not been completely developed and exploited as yet and can give still higher speed, put more functions on a chip and reduce the cost further. Optical lithography will face the fundamental wavelength barrier at linewidths of 0.3 to 0.5 micrometer. The X-ray lithography, which can engrave lines narrower than 0.3 micrometer, is being developed with possible volume production of chips around 1995. Meanwhile, electron-beam (25 keV) lithography is being used to fabricate 0.1–0.3 micrometer size test transistors and very-high-speed small-scale integrated circuits (logic gate arrays) for mainframe and supercomputer CPUs owing to its flexibility in personalization of chips.

Significant payoffs in other areas have also been made from using the micron and submicron lithographic patterning techniques developed for silicon integrated circuit manufacturing. These include the smaller, submicron thin-film magnetic heads for higher density magnetic disks and tapes as well as high-density optical memories, the finer pole pieces for multi-pole subminiature stepping motors, the macromolecular-size electrode arrays for biological sensors, the subminiature electronically-controlled liquid sensors and valves, and a host of other applications where micron and submicron lines and patterns are needed.

The applications in electronics have grown exponentially with time in a short span of less than two decades due to the availability of inexpensive and high performance silicon integrated circuits, especially the MOS type in the last ten years. Their volume application in the marketplace has not only proceeded at a high rate, but the market demands have also provided the positive feedback loop to further accelerate the technology advances in order to give still higher performance and larger varieties of applications. New applications using higher density, lower power and higher speed silicon MOSFETs include: personal workstations with high resolution real-time graphics and voice-input capabilities, automotive and aircraft electronics, and computerized offices, laboratories, and households with immense volume demands.

An equally if not more important application area is in the life sciences, including further and rapid advances in the understanding of the basic biological mechanisms using the fine-line lithography technology to fabricate biomolecular sensors, and in applications for medical instrumentations and diagnostics as well as patient monitoring.

One of the most important current problems to be solved to further advance the silicon integrated circuit technology concerns the mechanisms and modeling of the electrical failures of the smaller and higher speed MOS and BJT transistors and circuits owing to the presence of very high electric fields (several million volts per centimeter) in small transistors, and very high current densities (several hundred kiloamperes per square centimeters) in the thin and narrow metal lines which interconnect hundreds of thousands to many million transistors on a chip. Fundamental studies are necessary to obtain a detailed understanding of the atomic and chemical models and chemical kinetics of the aging and failure processes during the operation of the transistor at high electric fields and during the operation of the interconnect metal line at high current densities. Mathematical models must then be developed to accurately describe the kinetics of these failure processes in order that they can predict, using computer simulators, the failure rates of transistors and interconnect lines under a given operating condition for a given transistor structure and wiring design. Using these transistor and conductor failure models to stimulate the intergrated circuit operation on a computer, more reliable and higher performance circuits can then be designed and manufactured at high yields and lower cost. Another critical need is the development of more accurate models for process simulation and design at lower processing temperatures. Lower processing temperatures are necessary to control the fabrication of submicron transistors.

Acronyms, Abbreviations, and Symbols

Acronyms Abbreviations Symbols	Section First Appeared	Description (and commonly used words that follow)
1-T	IV-O	One Transistor (dynamic random access memory cell)
2-D	V-F	Two Dimensional
3-D	V-F	Three Dimensional
AM	IV-L	Amplitude Modulation
ASIC	V-G	Application Specific Integrated Circuit

Acronyms Abbreviations Symbols	Section First Appeared	Description (and commonly used words that follow)
BICMOS	Abstract	Bipolar and Complementary MOS (circuit)
BIMOS	IV-E	Bipolar and MOS (circuit)
BJT	II	Bipolar Junction Transistor
BTL	III-A	Bell Telephone Laboratories
C-BJT, CBJT	VII-A	Complementary Bipolar Junction Transistor (circuit)
CMOS	Abstract	Complementary Metal Oxide Semiconductor (circuit)
CPU	IV-F	Central Processing Unit
CV	IV-B	Capacitance Voltage (characteristics)
CVD	V-D	Chemical Vapor Deposition
DCTL	IV-E	Direct Coupled Transistor Logic (circuit)
DRAM	Abstract	Dynamic Random Access Memory (cell, chip or array)
EB	VI	Electron Beam (lithography, lithographic machine)
ECL	VII-A	Emitter Coupled Logic (circuit)
EEPROM	IV-P	Electrically Erasable Programmable Read Only Memory
EPROM	IV-P	Erasable Programmable Read Only Memory (cell, chip)
FAMOS	IV-P	Floating-gate Avalanche-injection MOS (memory cell)
FDRAM	V-I	Ferroelectric Dynamic Random Access Memory (cell, chip)
FET	I	Field Effect Transistor
FLOTOX	IV-P	Floating-gate Tunnel Oxide (programmable memory cell)
FM	IV-K	Frequency Modulation
FMOS	V-I	Ferroelectric MOS (memory cell)
FMV	IX	Foreign Market Value
FRAM	V-I	Ferroelectric Random Access Memory (cell, chip)
FSRAM	V-I	Ferroelectric Static Random Access Memory (cell, chip)
GMSRT	I	Gated Metal Semiconductor Rectifier Transistor
GMST	III-A	Gated Metal Semiconductor Transistor
HBJT	VIII	Heterojunction Bipolar Junction Transistor
HEMT	VII-A	High Electron Mobility Transistor
HFCV	IV-C	High Frequency Capacitance Voltage (characteristic)
IC	IV-I	Integrated Circuit
ICCD	Acknowledgment	International Conference on Computer Design
IEDM	IV-Q	International Electron Device Meeting

Acronyms, Abbreviations, and Symbols

Acronyms Abbreviations Symbols	Section First Appeared	Description (and commonly used words that follow)
IGFET	IV-L	Insulated Gate Field Effect Transistor
ISSCC	IV-J	International Solid State Circuits Conference
JGFET	III	Junction Gate Field Effect Transistor
LC	V-E	Inductance Capacitance (delay, time constant)
LCD	V-A	Liquid Crystal Display
LED	V-A	Light Emitting Diode
LOCOS	IV-M	Local Oxidation of Silicon
LSI	V	Large Scale Integration or Integrated (circuit)
MACSTA	Acknowledgment	Mid American Chinese Science & Technology Association
MBE	VIII	Molecular Beam Epitaxy
MDRAM	V-B	Merged 1-T DRAM
MESFET	I	Metal Semiconductor Field Effect Transistor
MIPS	IV-F	Million Instruction Per Second
MISFET	IV-L	Metal Insulator Semiconductor Field Effect Transistor
MIT	IV-L	Minority-carrier Injection Transistor (or M-CIT)
MITI	V-C	Ministry of Information Technology Industry (of Japan)
MODFET	VII-A	Modulation Doped Field Effect Transistor
MOS	I	Metal Oxide Semiconductor
MOSC	IV-B	Metal Oxide Semiconductor Capacitor
MOSCV	IV-C	Metal Oxide Semiconductor Capacitance Voltage (curve)
MOSFET	I	Metal Oxide Semiconductor Field Effect Transistor
MOST	IV-L	Metal Oxide Semiconductor Transistor
MPU	V-B	Microprocessing Unit
MSI	V	Medium Scale Integration
MSMT	I	Metal Semiconductor Metal Transistor
MSR	I	Metal Semiconductor Rectifier
NMOS	IV-J	N-channel MOS (circuit)
PMOS	IV-N	P-channel MOS (circuit)
PSG	IV-M	Phosphorus Silicate Glass (layer)
RAM	V-A	Random Access Memory (cell)
RC	V-D	Resistance Capacitance (delay, time constant)
SCR	IV-H	Silicon Controlled Rectifier
SCT	IV-H	Surface Controlled Tetrode
SMST	I	Semiconductor Metal Semiconductor Transistor
SOI	I	Silicon (or Semiconductor) on Insulator
SRAM	IV-J	Static Random Access Memory (cell or chip or array)

Acronyms, Abbreviations, and Symbols

Acronyms Abbreviations Symbols	Section First Appeared	Description (and commonly used words that follow)
SRH	III-H	Shockley Read Hall (recombination kinetics)
SSDRC	IV-G	Solid State Device Research Conference
SSI	V	Small Scale Integration or Integrated (circuit)
TCA	IV-M	Trichloroethane
TCE	IV-M	Trichloroethylene
TFFET	IV-I	Thin Film Field Effect Transistor
TFT	IV-I	Thin Film Transistor
ULSI	V	Ultra Large Scale Integration or Integrated (circuit)
UV	V-G	Ultra Violet (light)
UV-EPROM	IV-P	Ultra-Violet Erasable Programmable Read Only Memory
VLSI	Title	Very Large Scale Integration or Integrated (circuit)
WESCON	IV	Western Electronic Convention and Show
WSI	V-I	Wafer Scale Integration or Integrated (circuit)

Acknowledgment

This paper is dedicated to Prof. John Bardeen and Prof. William Shockley, who taught the author transistor physics, and to the late Dean William L. Everitt and Dean Frederick E. Terman who influenced the author's education and career path as teachers and family friends. The author was most fortunate to have received his first exposure to transistor physics and electronics from Bardeen in the Spring of 1953 when, as an undergraduate, he was permitted by Bardeen to audit Bardeen's second offer of a graduate semiconductor physics course (EE/Physics 435) at the University of Illinois in Urbana, and to have associated with Bardeen in teaching the subject for a quarter of a century (1963–1988). The author was doubly fortunate to be able to work also for Shockley at the beginning of the author's professional career (1956–1959) when the author finished his doctoral thesis on traveling wave tubes at Stanford and Shockley started his transistor company in Palo Alto. Those were the most productive three years of learning transistor physics under Shockley's tutelage.

This review is based on an invited talk of the same title by the author first given on March 1, 1983, as a Semiconductor Engineering and Research Seminar at the Digital Equipment Corporation. Eight subsequent lectures were at: the University of California-Davis; Los Alamos National Laboratory; the 1984 Caltech-MIT VLSI Research Symposium at MIT; the Chinese Science and Technology University in Beijing and the Fudan University in Shanghai in April, 1984; the **MACSTA** (Mid-American Chinese Science and Technology Association) Spring Symposium in 1985; the Northwestern University; and the 1985 International Conference on Computer Design (**ICCD**). Two additional presentations were made to about 60 students in two second-year graduate courses on transistor physics in Urbana which provided further inputs on the amount of details to include in this written version. With the cooperation of the PROCEEDINGS OF THE IEEE Technical Editor, Hans P. Leander, I was able to add updates and details in the revision in order to include recent developments up to the fourth quarter of 1988 and to include suggestions made by the personal reviewers I recruited. The preparation of this review was started when a former graduate student, Dr. Chang Su Kim (previously with the Digital Equipment Corporation (DEC) and now the senior managing director of the semiconductor division of Goldstar Co., Ltd. of Korea) asked me in December 1982 to present a research seminar at the Semiconductor Engineering Laboratory of DEC and suggested this subject when I asked for a topic. Dr. Donald E. Nelsen of DEC initiated the invited presentations at the 1984 Caltech-MIT VLSI Research Symposium in Cambridge and at the 1985-ICCD in December in New York. Dr. Robert B. Hammond, formerly of Los Alamos National Laboratory, and Prof. Fredrik A. Lindholm of the University of Florida suggested to the PROCEEDINGS OF THE IEEE to invite me to write this review. Dr. Neil Berglund, former general manager of technology at INTEL, introduced me to the model of manufacturing equipment availability on the growth of DRAM production and provided me some of the state-of-the-art information including the viewgraphs of the graduate seminar he presented in Urbana on April 19, 1982. It is Berglund's introduction of this subject to me and his subseqent technical supports, via a former graduate student Dr. Leopoldo D. Yau of Intel as the research project mentor of my reliability physics contract from the Semiconductor Research Corporation, that greatly influenced and in fact started my second-round of research on silicon transistor manufacturing problems, after a nearly twenty-year diversion. Dr. Gordon E. Moore of Intel generously provided me inputs of his market-oriented model of Si MOS VLSI production trend.

Many colleagues who have been participants in the earlier and current phases of the MOS evolution have advised me on the history presented here. In particular, I thank Prof. Bardeen whose suggestions have rectified several historical inaccuracies in the second draft and the galley proof. I also wish to especially thank Prof. Shockley whose comments and suggestions have helped to improve the presentation and eliminate ambiguous and misleading statements. I also thank Lewis Terman for his detailed reading and remarks of the second draft and the galley proof, which have helped to improve the technical and historical accuracies as well as presentation. I would like also to thank Dr. Donald Young for correcting a historical inaccuracy in the second draft on the size of the initial IBM effort on MOS technology development during 1963–1964. I also thank Dr. James Early for clear comments and concise suggestions on the final draft, and very detailed written recommendations on the galley proof which have helped to improve the presentation and accuracy. In addition, Fred Tsang, formerly of Intel, pointed out to me that the practical CBJT is less complex than the textbook picture of Fig. 23. Dr. William D. Miller sent me information on Krysalis' ferroelectric MOS memory. Dr. Toshi Nishida commented from a graduate student's perspective, obtained some references, and also helped in entering some of the revisions in the second draft and suggested some editorial and semantic changes. To all of these people, I am deeply grateful.

I have endeavored to make this review historically and technically accurate with added personal hands-on expe-

riences and viewpoints. About twice as many references could not be listed nor discussed due to limited space. Many of the cited references will lead to additional and earlier references of historical interest. For any remaining oversights and historical inaccuracies, I apologize and hope that a future letter to the editor or addendum based on reader's comments will provide clarifications and corrections.

The drafts were composed and edited by the author personally on a personal MicroVAX-2 using the MASS-11 scientific word processor and on a VAX-11/750 donated by the Digital Equipment Corporation in 1983 with the support and at the insistence of Chang Su Kim and Rich Hollingsworth, both of DEC, as a result of the first lecture on this subject given by the author at DEC on March 1, 1983.

References

[1] J. E. Lilienfeld, "Method and apparatus for controlling electric currents," U.S. Patent 1 745 175. Application filed Oct. 8, 1926, granted Jan. 18, 1930.

[2] ——, "Device for controlling electric current," U.S. Patent 1 900 018. Application filed Mar. 28, 1928, granted Mar. 7, 1933.

[3] ——, "Amplifier for electric currents," U.S. Patent 1 877 140. Application filed Dec. 8, 1928, granted Sept. 13, 1932.

[4] Obituaries, *Physics Today*, vol. 16, no. 11, p. 104, Nov. 1963. See also [76] on a recent announcement about Lilienfeld.

[5] O. Heil, "Improvements in or relating to electrical amplifiers and other control arrangements and devices," British Patent 439 457, application filed Mar. 4, 1935, granted Dec. 6, 1935. Germany Convention Date, Mar. 2, 1934.

[6] W. Shockley, "The path to the conception of the junction transistor," *IEEE Trans. Electron Devices*, vol. ED-23, no. 7, pp. 597–620, July 1976. Reprinted, vol. ED-31, no. 11, pp. 1523–1546, Nov. 1984.

[7] J. Bardeen, "Three-electrode circuit element utilizing semiconductive materials," U.S. Patent 2 524 033. Application filed Feb. 26, 1948, granted Oct. 3, 1950. See also, "Semiconductor research leading to the point contact transistor," Nobel Lecture, Dec. 11, 1956; and "Solid state physics—1947," *Solid State Technol.*, pp. 68–71, Dec. 1987.

[8] W. B. Shockley, "Circuit element utilizing semiconductive material," U.S. Patent 2 569 347. Application filed June 26, 1948, granted Sept. 25, 1951.

[9] A. H. Wilson, "The theory of electronic semiconductors, I," *Proc. Roy. Soc.*, vol. A133, pp. 458–468, 1931.

[10] ——, "The theory of electronic semiconductors, II," *Proc. Roy. Soc.*, vol. A134, pp. 277–287, 1931.

[11] V. E. Bottom, "Invention of the solid-state amplifier," *Physics Today*, vol. 17, no. 2, pp. 24–26, Feb. 1964.

[12] J. B. Johnson, "More on the solid-state amplifier and Dr. Lilienfeld," *Physics Today*, vol. 17, no. 5, pp. 60–62, May 1964.

[13] J. Bardeen and W. H. Brattain, "Three-electrode circuit element utilizing semiconductive materials," U.S. Patent 2 524 035. Application filed June 17, 1948, granted Oct. 3, 1950.

[14] H. C. Torrey and C. A. Whitmer, *Crystal Rectifiers*. New York, NY: McGraw-Hill, 1948; reprinted by Boston Technical Publishers, Inc., 1964.

[15] G. L. Pearson and W. H. Brattain, "History of semiconductor research," *Proc. IRE*, vol. 43, no. 12, pp. 1794–1806, Dec. 1955.

[16] E. H. Rhoderick, *Metal-Semiconductor Contacts*. New York, NY: Oxford University Press, 1978.

[17] J. Bardeen, "Surface states and rectification at a metal semiconductor contact," *Phys. Rev.*, vol. 71, no. 10, pp. 717–727, May 15, 1947. For a one-dimensional theory of surface states, see W. Shockley, "On the surface states associated with a periodic potential," *Phys. Rev.*, vol. 56, pp. 317–323, Aug. 15, 1939.

[18] W. Shockley and G. L. Pearson, "Modulation of conductance of thin films of semi-conductors by surface charges," *Phys. Rev.*, vol. 74, No. 2, pp. 232–233, July 15, 1948.

[19] J. Bardeen and W. H. Brattain, "The transistor, a semi-conductor triode," *Phys. Rev.*, vol. 74, no. 2, 230–231, July 15, 1948.

[20] W. H. Brattain and J. Bardeen, "Nature of the forward current in Germanium Point Contacts," *Phys. Rev.*, vol. 74, no. 2, 231–232, July 15, 1948; vol. 75, no. 5, p. 1208, 1949.

[21] J. N. Shive, "The double surface transistor," *Phys. Rev.*, vol. 75, pp. 689–690, 1949. First reported in a closed door Bell Lab project review meeting on Feb. 18, 1948 [6].

[22] W. Shockley, "Theory of p-n junctions in semiconductors and p-n junction transistors," *Bell Syst. Techn. J.*, vol. 28, no. 7, pp. 436–489, July 1949.

[23] W. H. Brattain and J. Bardeen, "Surface properties of germanium," *Bell Syst. Tech. J.*, vol. 32, no. 1, pp. 1–41, Jan. 1953.

[24] W. Shockley, "Electrons, holes and traps," *Proc. IRE*, vol. 46, no. 6, pp. 973–990, June 1958.

[25] C. T. Sah, "The equivalent circuit model in solid-state electronics—Part I: The single energy level defect centers," *Proc. IEEE*, vol. 55, no. 5, pp. 654–771, May 1967; "Part II: The multiple energy level impurity centers," pp. 672–684.

[26] ——, "The equivalent circuit model in solid-state electronics—III (Conduction and displacement currents)," *Solid-State Electron.*, vol. 13, no. 12, pp. 1547–1575, Dec. 1970.

[27] ——, "Equivalent circuit models in semiconductor transport for thermal, opitcal, auger-impact and tunnelling recombination-generation-trapping processes," *Phys. Status Solids*, vol. (a)7, pp. 541–559, Oct. 16, 1971.

[28] F. A. Lindholm and C. T. Sah, "Circuit technique for semiconductor-device analysis with junction diode open circuit voltage decay example," *Solid-State Electron.*, vol. 31, no. 2, pp. 197–204, Feb. 1988. For a latest extension, see C. T. Sah, "New integral representation of circuit models and elements for the circuit technique for semiconductor device analysis," *Solid-State Electron.*, vol. 30, no. 12, pp. 1277–1281, Dec. 1987.

[29] C. T. Sah, "Interface traps on Si surface," in *Properties of SILICON*. London, England: INSPEC, The Institution of Electrical Engineers, section 17.1, pp. 499–507, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

[30] W. Shockley, "A unipolar field effect transistor," *Proc. IRE*, vol. 40, no. 11, pp. 1365–1376, Nov. 1952.

[31] W. L. Brown, "n-type surface conductivity on p-type germanium," *Phys. Rev.*, vol. 91, no. 5, pp. 518–527, Aug. 1, 1953.

[32] I. M. Ross, "Semiconductive translating device," U.S. Patent 2 791 760. Application field Feb. 18, 1955, granted May 7, 1957.

[33] G. K. Teal, M. Sparks, and E. Buehler, "Growth of germanium single crystals containing p-n junctions," *Phys. Rev.*, vol. 81, no. 2, p. 637, Feb. 1951.

[34] G. C. Dacey and I. M. Ross, "Unipolar field-effect transistor," *Proc. IRE*, vol. 41, no. 8, pp. 970–979, Aug. 1953.

[35] R. N. Hall and W. C. Dunlap, "p-n junctions prepared by impurity diffusion," *Phys. Rev.*, vol. 80, no. 3, pp. 467–468, Nov. 1, 1950.

[36] J. H. Scaff and H. C. Theuerer, "Semiconductor comprising silicon and method of making it," U.S. Patent 2 567 970. Application filed Dec. 24, 1947, granted Sept. 18, 1951.

[37] C. S. Fuller and J. A. Ditzenberger, "The diffusion of boron and phosphorus into silicon," *J. Appl. Phys.*, vol. 25, no. 11, pp. 1439–1440, Nov. 1954.

[38] M. Tannenbaum and D. E. Thomas, "Diffused emitter and base silicon transistors," *Bell Syst. Tech. J.* vol. 35, no. 1, pp. 1–15, Jan. 1956.

[39] C. J. Frosch and L. Derrick, "Surface protection and selective masking during diffusion in silicon," *J. Electrochem. Soc.*, vol. 104, no. 5, pp. 547–552, May 1957.

[40] L. Derrick and C. J. Frosch, "Manufacture of silicon devices," U.S. Patent 2 804 405. Application filed Dec. 24, 1954, granted Apr. 27, 1957.

[41] C. T. Sah, H. Sello, and D. A. Tremere, "Diffusion of phosphorus in silicon dioxide film," *J. Phys. Chem. Solids*, vol. 11, no. 3/4, pp. 288–298, Mar. 1959.

[42] H. C. Theuerer, "Method of processing semiconductive materials," U.S. Patent 3 060 123. Application filed Dec. 17, 1952, granted Oct. 23, 1962.

[43] ——, "Removal of boron from silicon by hydrogen water vapor treatment," *Trans. AIME*, vol. 206, pp. 1316–1319, Oct. 1956.

[44] For a review, see W. G. Pfann, "Principles of zone-refining," *Trans. AIME*, vol. 194, pp. 747–799, Mar. 1952.

[45] See also, W. G. Pfann, "Techniques of zone melting and crystal growth," *Solid State Phys.*, vol. 1, pp. 423–451, 1957, F. Seitz and D. Turnbull, Eds. New York, NY: Academic Press; and [46].
[46] W. G. Pfann, "The semiconductor revolution," *J. Electrochem. Soc.*, vol. 121, no. 1, pp. 9C–15C, Jan. 1974.
[47] W. Shockley, "Forming semiconductive devices by ionic bombardment," U.S. Patent 2 787 564. Application filed Oct. 28, 1954, granted Apr. 12, 1958.
[48] For a review see J. F. Gibbons, "Ion implantation in semiconductors—part I. Range distribution theory and experiments," *Proc. IEEE*, vol. 56, no. 3, pp. 295–319, Mar. 1968.
[49] H. C. Theuerer, J. J. Kleimack, H. H. Loar, and H. Christenson, "Epitaxial diffused transistors," *Proc. IRE*, vol. 48, no. 9, pp. 1642–1643, Sept. 1960.
[50] I. M. Ross, 1963 Morris N. Leibmann Award of I.R.E. "For contribution to the development of the epitaxial transistor and other semiconductor devices."
[51] W. Shockley and W. T. Read, Jr., "Statistics of recombination of holes and electrons," *Phys. Rev.*, vol. 87, no. 9, pp. 835–842, Sept. 1952.
[52] R. N. Hall, "Germanium rectifier characteristics," *Phys. Rev.*, vol. 83, p. 228, 1951.
[53] H. Kleinknecht and K. Seiler, "Einkristalle und pn schichtkristalle aus silizium," *Z. Phys.*, vol. 139, no. 12, pp. 599–618, Dec. 1954.
[54] E. M. Pell and G. M. Roe, "Reverse current and carrier lifetime as a function of temperature in germanium junction diodes," *J. Appl. Phys.*, vol. 26, no. 6, pp. 658–665, June 1955.
[55] C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier generation and recombination in p-n junction and p-n junction characteristics," *Proc. IRE*, vol. 45, no. 9, pp. 1228–1243, Sept. 1957.
[56] W. Shockley, "Hot electrons in germanium and Ohm's law," *Bell Syst. Tech. J.*, vol. 30, no. 10, pp. 990–1034, Oct. 1951.
[57] For a followup, see W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electron.*, vol. 2, no. 1, pp. 35–67, Jan. 1967.
[58] M. M. Atalla, M. Tannenbaum, and E. J. Scheibner, "Stabilization of Silicon Surface by Thermally Grown Oxides," *Bell Syst. Tech. J.*, vol. 38, no. 3, p. 123, May 1959. For a recent review and analysis of the interface state properties of oxidized silicon surfaces, see [29].
[59] J. Moll, "Variable capacitance with large capacity change," in *1959 WESCON (Western Convention and Show) Record*, part 3, pp. 32–36, IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855–1331.
[60] L. M. Terman, "An investigation of surface states at a silicon/silicon diode interface metal-oxide-silicon diodes," *Solid-State Electron.*, vol. 5, pp. 285–299, *Sept.-Oct.* 1962.
[61] C. T. Sah, M. S.-C. Luo, C. C.-H. Hsu, T. Nishida, and A. J. Chen, "Interface traps on oxidized Si from two-terminal, dark capacitance-voltage measurements on MOS capacitors," in *Properties of SILICON*. London, England: INSPEC, The Institution of Electrical Engineers, section 17.4, pp. 521–531, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855–1331.
[62] C. T. Sah and C. C.-H. Hsu, "Oxide traps on oxidized Si," in *Properties of SILICON*. London, England: INSPEC, The Institution of Electrical Engineers, section 17.5, pp. 532–547, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855–1331.
[63] J. A. Hoerni, "Planar silicon transistors and diodes," presented at the 1960 IRE International Electron Device Meeting, Oct. 27–29, 1960. Technical Article and Paper Series, No. TP-14, 9 pp., 1961. Fairchild Semiconductor Corporation, 645 Whisman Road, Mountain View, CA.
[64] These were recorded in product data sheets, titled "Fairchild Silicon Transistors," during 1959 to 1961. The transistor number, data sheet number, date of issue and transistor type are: 2N696: SL-4/1(npn,mesa), SL-5/4(9-61,npn, planar); 2N706: SL-13/3(6-61,npn,mesa,gold-doped,switch); 2N709: SL-52/3(2-62,npn,planar, gold-doped,switch); 2N869: SL-42/3(6-61,pnp,planar); 2N914: SL-36/3(6-61,npn,planar, epitaxial, gold-doped,switch); 2N1131: SL-6/4(6-61,pnp, mesa); 2N1613: SL-17/4(9-61,npn, planar).
[65] R. N. Noyce, "Semiconductor device-and-lead structure," U.S. Patent 2 981 877. Application filed July 30, 1959, granted Apr. 25, 1961.
[66] For a personal account from the co-inventor of the integrated circuit, see Jack S. Kilby, "Invention of the integrated circuit," *IEEE Trans. Electron Devices*, vol. ED-23, no. 7, pp. 648–654, July 1976.
[67] R. H. Norman, J. T. Last, and I. Hass, "Solid state micrologic elements," in *IRE-IEEE Solid State Circuits Conference*, Feb. 1960, and full article published in the Fairchild Technical Articles and Papers series, no. TP-7, 8 pp. See also [68].
[68] R. H. Norman, "Status report on micrologic elements, "presented at the 51st Bumblebee Guidance Panel, June 22, 1960, and full article published in the Fairchild Technical Articles and Papers series, no. TP-10, 8 pp.
[69] J. E. Thornton, "Design of a computer, the Control Data 6600," pp. 5–6 and pp. 19–32, Glenview, ILL: Scott, Foresman and Company, 1970. The performance figures quoted in the text come from the Boston Computer Museum exhibit and are different from Thorton's book which gives 20 MHz and 400 000 Si transistors. See also, "CDC 6600 in stretch class," Datamation, p. 13, May 1961. The dollar and unit of the production contract was reported by Norman H. Bowan, "Business As Usual, Good News—But Good," Palo Alto Times, Sept. 1, 1964. A particularly timely statement read "But this order of Fairchild's is not for military uses. It's a commercial order."
[70] D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced surface devices," presented at the IRE-AIEE Solid-State Device Research Conference at Carnegie Institute of Technology, Pittsburgh, PA, 1960. See also a personal account given by D. Kahng [71] and a summary of our analysis [72].
[71] D. Kahng, "A historical perspective on the development of MOS transistors and related devices," *IEEE Trans. Electron Devices*, vol. ED-23, no. 7, pp. 655–657, July 1976.
[72] One patent was awarded to each of these two inventors. Atalla's patent [73], filed second but granted first claims an analog transistor device operation in the punch-through or space charge limited mode. Kahng's patent [74], filed first but granted second, claims a circuit arrangement or circuit configuration to operate the device. Kahng's patent describes a MOSFET with two diffused p-type region on an n-type silicon and Atalla's patent describes one with two diffused n-type region on an intrinsic or high resistivity p-type or pi-type silicon. In the personal account by Kahng given in [71], he claims that all the possible current-voltage characteristics were presented at the Solid State Device Research conference and predicted by his unpublished analyses made in Jan. 1961 [75]. This claim is supported by Kahng's drain current equation (22)' $\{I_p = [s\mu\epsilon_i/2\delta]V_p(V_i - V_p/2)$ where s is the channel width divided by the channel length, μ is the hole mobility in the surface channel, ϵ_i is the dielectric constant of the gate oxide, V_p is the drain-to-source voltage, and V_i is the gate (or field plate) voltage.$\}$ and the computed current voltage characteristics given in Kahng's figure 6 for the experimental p-channel MOSFET with boron diffused source and drain regions in an n-type silicon.
[73] M. M. Atalla, "Semiconductor triode," U.S. Patent 3 056 888. Application filed Aug. 17, 1960, granted Oct. 2, 1962.
[74] D. Kahng, "Electric field controlled semiconductor device," U.S. Patent 3 102 230. Application filed May 31, 1960, granted Aug. 27, 1963.
[75] ——, "Silicon-silicon dioxide surface device," Memorandum for file, MH-2821-DK-pg, 23 pp. and 10 figures, Bell Telephone Laboratories, Jan. 16, 1961. A copy can be obtained from the author, P.O. Box 550, Martinsville, NJ 08836.
[76] W. Sweet, "American Physical Society establishes major prize in memory of Lilienfeld," *Physics Today*, vol. 41, no. 5, pp. 87–89, May 1988. Received by this author on May 23, 1988. A copy of the initial versions of the last section of this article, captioned "Bardeen's evaluation," were given to this author by Bardeen on May 5, 1988, during a three-hour discussion of the second (Jan. 25, 1988) version of the manuscript of this evolution paper.
[77] C. T. Sah, "A new semiconductor tetrode, the surface-potential controlled transistor," *Proc. IRE*, vol. 49, no. 11, pp. 1623–1634, Nov. 1961.

[78] ——, "Surface-potential controlled semiconductor device," U.S. Patent 3 204 160. Application filed Apr. 12, 1961, granted Aug. 31, 1965.
[79] ——, "Transistors," in *1963 McGraw-Hill Yearbook of Science and Technology*. New York, NY: McGraw-Hill, pp. 560–562. In particular, Fig. 2(d) shows the cross-sectional view of a BIMOS with a p/n/p bipolar junction transistor integrated with an enhancement-mode p-channel MOSFET.
[80] ——, "Surface-potential controlled semiconductor device," U.S. Patent 3 243 699. Application filed June 11, 1962, granted Mar. 29, 1966.
[81] ——, "Effect of surface recombination and channel on p-n junction and transistor characteristics," *IRE Trans. Electron Devices*, vol. ED-9, no. 1, pp. 94–108, Jan. 1962.
[82] Applications of the tetrode was described by H. Z. Bogert, C. T. Sah, and D. A. Tremere, "Applications of the surface potential controlled transistor tetrodes," in *Proceedings of the IEEE 1962 Int. Solid State Circuits Conf.*, pp. 34–35, Feb. 1962.
[83] See references 68, 69, and 70 in M. S Adler *et al.*, "The evolution of power device technology," *IEEE Trans. Electron Devices*, vol. ED-31, no. 11, pp. 1570–1591, Nov. 1984. See also a popular account by A. Pshaenich [84].
[84] A. Pshaenich, "MOS thyristor improves power-switching circuits," *Electron. Des.*, pp. 165–170, May 12, 1983.
[85] For a detailed description of the floating gate charge storage experiment, see section D, titled 'Grid (Gate) input impedance,' on p. 1625 in [77].
[86] T. H. Ning *et al.*, "1 um MOSFET VLSI technology: Part IV. Hot electron design constraints," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 346–353, Apr. 1979.
[87] See the comprehensive review of the research literature on the properties of the silicon-dioxide/silicon interface given by Sah and his graduate students in the datareview handbook *Properties of Silicon*. London, England: British IEE, May 1988. The reviews are in sections 17.1–17.5, 17.7–17.21, pp. 497–639. This chapter (Chapter 17) also contains authoritative reviews on the properties of the oxide film on silicon by other active researchers. The book can be acquired from the INSPEC Dept. IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.
[88] D. R. Young and D. P. Seraphim, "Surface effects on silicon: introduction," *IBM J. Res. Develop.*, vol. 8, no. 4, pp. 366–367, Sept. 1964. This whole issue is devoted to the research results obtained by this group during 1963–1964.
[89] F. M. Wanlass, "Gas-solid interactions," Ph.D. Thesis, Department of Physics, University of Utah, printed June 1962, 123 pp.; approved by his thesis committee, May 1960.
[90] F. M. Wanlass and C. T. Sah, "Nanowatt logic using field-effect metal-oxide semiconductor triodes," in *Technical Digest of the IEEE 1963 int. Solid-State Circuit Conf.*, pp. 32–33, February 20, 1963. See also [91, 92, 93].
[91] ——, "Nanowatt logic using field-effect metal-oxide semiconductor triodes (MOSTs)," Fairchild Tech. Rep. 98, Jan. 23, 1963.
[92] G. E. Moore, C. T. Sah, and F. Wanlass, "Metal-oxide-semiconductor field-effect devices for micropower logic circuitry," in *Micropower Electronics*, Edward Keonjian, Ed. New York, NY: Pergamon Press, 1964, pp.41–55.
[93] F. M. Wanlass, "Low stand-by power complementary field effect circuitry," U.S. Patent 3 356 858, filed June 18, 1963, issued Dec. 5, 1967.
[94] For a recent report on Wanlass, see C. Barney, "He started MOS from scratch," *Electronics Week*, p. 64, Oct. 8, 1984.
[95] The acroynm MOST given in the title of the Fairchild technical memorandum [91] did not appear in the article published in the 1963-ISSCC proceeding [90]. After twenty-six years, my recollection on the reason for this omission is hazy and I seem to recall my anxiety after F. Wanlass told me during Christmas 1962 about G. E. Moore's remark discussed in the text which probably resulted in its omission. However, it was used in the first article I wrote on the characteristics of the MOS transistor [96].
[96] C. T. Sah, "Characteristics of the metal-oxide-semiconductor transistors," *IEEE Trans. Electron Devices*, vol. ED-11, no. 7, pp. 324–345, July 1964.
[97] L. M. Terman, private communications, June 1, 1988.
[98] S. M. Sze, *Physics of Semiconductor Devices*. New York, NY: Wiley, Ch. 10, first edition 1969, ch. 8, second edition, 1981.
[99] E. H. Snow, B. E. Deal, A. S. Grove, and C. T. Sah, "Ion transport phenomena in insulating films using the MOS structure," *J. Appl. Phys.*, vol. 36, no. 5, pp. 1664–1673, May 1965.
[100] D. R. Kerr and D. R. Young, "Method of improving electrical characteristics of semiconductor devices and products so produced," U.S. Patent 3 303 059, filed June 29, 1964, issued Feb. 7, 1967.
[101] D. R. Kerr, J. S. Logan, P. J. Burkhardt, and W. A. Pliskin, "Stablization of SiO2 passivation layers with P_2O_5," *IBM J. Res. Develop.*, vol. 8, no. 4, pp. 376–384, Sept. 1964.
[102] E. H. Snow and B. E. Deal, "Polarization phenomena and other properties of phosphosilicate glass films on silicon," *J. Electrochem. Soc.*, vol. 113, no. 3, pp. 263–269, Mar. 1966.
[103] P. Balk, "Effects of hydrogen annealing on silicon surfaces," presented at the Electrochemical Society Spring Meeting, San Francisco, CA, May 9–13, 1965. Extended Abstracts of Electronics Division, vol. 14, no. 1, abstract no. 109, pp. 237–240, May 1965.
[104] ——, "Low temperature annealing in the Al-SiO_2-Si system." presented at the Electrochemical Society Meeting, Buffalo, NY, Oct. 10–15, 1965. Extended Abstracts of Electronics Division, vol. 14, no. 2, abstract no. 111, pp. 29–32, Oct. 1965. Also abstracted in *J. Electrochem. Soc.*, vol. 112, no. 8, abstract no. 111, p. 185C, Aug. 1965.
[105] E. Kooi, "Effects of low temperature heat treatments on the surface properties of oxidized silicon," *Philips Res. Rep.*, vol. 20, pp. 578–594, Oct. 1965 who stated, ". . . Balk has proposed that the effect of hydrogen during annealing (heat treatment of oxidized silicon in 300–500°C range) is due to a chemical saturation with H atoms of certain unsaturated bonds (dangling bonds) at the oxide-silicon interface. The experiments to be described in this paper suggest that the main effect of water and an aluminum electrode on the oxide during low-temperature heat treatments has to be explained in a similar way, although charge redistribution (due to sodium ion contamination of the early oxidation technology) in the oxide-silicon system cannot always be neglected."
[106] P. Balk, P. G. Burkhardt, and L. V. Gregor, "Orientation dependence of built-in surface charge on thermally oxidized silicon," *Proc. IEEE*, vol. 53, no. 12, pp. 2133–2134, Dec. 1965.
[107] J. F. Delord, D. G. Hoffman, and G. Stringer, "Use of MOS capacitors in determining the properties of surface states at the Si-SiO2 interface," *Bull. Amer. Phys. Soc.*, serial II, vol. 10, no. 4, abstract KF9, 546, Apr. 26, 1965.
[108] Y. Miura, "Effect of orientation on surface charge density at silicon-silicon dioxide interface," *Japan. J. Appl. Phys.*, vol. 4, no. 12, pp. 958–961, Dec. 1965.
[109] P. Handler, "Electrical properties of a clean germanium surface," in *Semiconductor Surface Physics*, R. H. Kingston, Ed. Philadelphia, PA: University of Pennsylvania Press, 1957, pp. 23–51. See also a detailed recent review and a new and corrected tabulation of the orientational dependences of the dangling bond density by Sah [29], [110].
[110] C. T. Sah, "Interface traps on oxidized Si from X-ray photoemission spectroscopy, MOS diode admittance, MOS transistor and photogeneration measurements," in *Properties of SILICON*, London, England: INSPEC, The Institution of Electrial Engineers, section 17.3, pp. 512–520, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. See also section 17.1 cited in [29].
[111] R. E. Kerwin, D. L. Klein, and J. C. Sarace (Bell Telephone Laboratories), "Method for making MIS structures," U.S. Patent 3 475 234, filed Mar. 27, 1967, issued Oct. 28, 1969.
[112] J. C. Sarace, R. E. Kerwin, D. L. Klein, and R. Edwards, "Metal-nitride-oxide-silicon field-effect transistors with self-aligned gates," *Solid-State Electron.* vol. 11, no. 7, pp. 653–660, July 1968.
[113] H. G. Dill (Hughes Aircraft Company), "Insulated gate field-effect transistor (IGFET) with semiconductor gate electrode," U.S. Patent 3 544 399, filed Oct. 26, 1966, issued Dec. 1, 1970.
[114] R. W. Bower and H. G. Dill, 1966 IEDM paper 16.6 which cov-

ered the first ion implanted MOSFET using self-aligned aluminum metal gate.
[115] R. W. Bower, H. G. Dill, K. G. Aubuchon, and S. A. Thompson, "MOS field-effect transistors formed by gate masked ion implantation," *IEEE Trans. Electron Devices*, vol. ED-15, no. 10, pp. 757-761, Oct. 1968.
[116] L. L. Vadasz, A. S. Grove, T. A. Rowe, and G. E. Moore, "Silicon gate technology," *IEEE Spectrum*, vol. 6, no. 10, pp. 28-35, Oct. 1969.
[117] F. P. Heiman, "Integrated insulation-gate field-effect transistor circuit on a single substrate employing substrate-electrode bias," U.S. Patent 3 233 123, application filed Feb. 14, 1963, granted Feb. 1, 1966.
[118] R. H. Dennard, "Field-effect transistor memory," U.S. Patent 3 387 286, application filed July 14, 1967, granted June 4, 1968. See also a recent personal account by the inventor [119].
[119] —, "Evolution of the MOSFET dynamic RAM—a personal view," *IEEE Trans. Electron Devices*, vol. ED-31, no. 11, pp. 1549-1555, Nov. 1984.
[120] D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," *Bell Syst. Tech. J.* (Brief), vol. 46, no. 4, pp. 1288-1295, July-Aug. 1967.
[121] D. Kahng, "Field effect semiconductor apparatus with memory involving entrapment of charge carriers," U.S. Patent 3 500 142, filed June 5, 1967, issued Mar. 10, 1970.
[122] D. Frohman-Bentchkowsky, "Floating gate transistor and method for charging and discharging same," U.S. Patent 3 660 819, filed June 15, 1970, issued May 2, 1972.
[123] See also D. Frohman-Bentchkowsky, "A fully decoded 2048-bit electrically-programmable MOS-ROM," in *ISSCC Digest of Tech. Papers*, pp. 80-81, Feb. 1971, and a full size paper with the same title in *IEEE J. Solid-State Circuits*, vol. SC-6, no. 10, pp. 301-306, Oct. 1971.
[124] D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.*, vol. 18, pp. 332-334, Apr. 15, 1971.
[125] —. "FAMOS—A new semiconductor charge storage device," *Solid-State Electron.*, vol. 17, pp. 517-529, 1974.
[126] J. F. Verwey, "Nonvolatile semiconductor memories," *Advances in Electronics and Electron Physics*, vol. 41, pp. 249-309, 1976.
[127] J. J. Chang, "Nonvolatile semiconductor memory devices," *Proc. IEEE*, vol. 64, no. 7, pp. 1039-1059, July 1976.
[128] Y. Nishi and H. Iizuka, "Nonvolatile memories," in *Silicon Integrated Circuits*, Dawon Kahng, Ed. New York, NY: Academic Press, 1981, pp. 121-251.
[129] W. I. Kinney, W. Shepherd, W. Miller, J. Evans, and R. Womack (Krysalis Microelectronic), "A non-volatile memory cell based on ferroelectric storage capacitors," in *Technical Digest of the 1987 IEDM*, paper 3.9, pp. 850-851, Dec. 1987.
[130] See also "Ferroelectrics: the latest in RAD-hard, nonvolatile memory," *Electronics*, p. 162, Jan. 7, 1988.
[131] For a detailed discussion of some speculated advantages of the nonvolatile and high-speed ferroelectric MOS memory, see Section IV, titled FEFET, on page 1053 in J. J. Chang's review cited in [127] which also listed twenty-eight references of research reports on using ferroelectric for memory during 1967 to 1974.
[132] R. Womack, W. Kinney, B. Shepherd, B. Miller, and J. Evans, "An experimental 512-bit nonvolatile memory with ferroelectric storage cell," submitted to the IEEE Int. Solid State Circuit Conf., San Francisco, CA, Feb. 12, 1988. Manuscript available from Bill Miller. See also "Krysalis Co-Founders leave in realignment," *Electronic News*, p. 24, July 4, 1988; "Krysalis to file Chapter 11," *Electronic News*, p. 22, Sept. 19, 1988.
[133] S. Sheffield-Eaton, D. B. Bulter, M. Parris, D. Wilson, and H. McNeillie, "A ferroelectric nonvolatile memory,"; in *IEEE Int. Solid State Circuits Conf.* (ISSCC), Digest of Technical Papers, XXXI, pp. 130-131, Feb. 1988. IEEE Cat. No. 88CH2562-7. Publisher: Lewis Winner, Coral Gables, FL 33134.
[134] For an early review of the MOS memory cell designs, see L. M. Terman, "MOSFET memory circuits," *Proc. IEEE*, vol. 59, no. 7, pp. 1044-1058, July 1971.
[135] J. D. Schmidt, "Integrated MOS transistor random access memory," *Solid State Design*, pp. 21-25, Jan. 1965. A 256 word (72 bit/word) silicon BJT memory system using the six-transistor cell and 36-bit per chip (60 x 80 square mil area) was described by H. A. Perkins and J. D. Schmidt [136].
[136] H. A. Perkins and J. D. Schmidt, "An integrated semiconductor memory system," in *Proceedings of IEEE Fall Joint Computer Conf.*, pp. 1053-1064, 1965, which used the CTμL (Complementary Transistor Micrologic) elements.
[137] For a comprehensive compilation of silicon memory developments and products prior to 1972, see the collected reprint volume by D. A. Hodges, *Semiconductor Memories*. New York, NY: IEEE Press, 1972. See also reviews given in [138]-[142].
[138] Special Issue on Semiconductor Memory, *IEEE Trans. Electron Devices*, vol. ED-26, no. 6, pp. 825-918, June 1979.
[139] Special Issue on Logic and Memory, *IEEE J. Solid-State Circuits*, vol. SC-17, no. 5, pp. 790-963, Oct. 1982.
[140] For a more recent survey of the silicon memory technology, see E. W. Pugh, D. L. Critchlow, R. A. Henie, and L. A. Russell, "Solid state memory development at IBM," *IBM J. Res. Develop.*, vol. 25, no. 5, pp. 585-602, Sept. 1981 and [141].
[141] W. E. Harding, "Semiconductor manufacturing in IBM, 1957 to the present: A perspective," *IBM J. Res. Develop.*, vol. 25, no. 5, pp. 647-658, Sept. 1981.
[142] S. Asai (Hitachi Central Research Laboratory), "Semiconductor memory trends," *Proc. IEEE*, vol. 74, no. 12, pp. 1623-1635, Dec. 1986.
[143] R. H. Dennard *et al.*, "Design of ion implanted MOSFET's with very small physical dimensions," *J. Solid-State Circuits*, vol. SC-9, pp. 256-268, May 1973.
[144] C. N. Berglund, "The evolution of MOS process technology," presented at a graduate seminar of the Coordinated Science Laboratory of the University of Illinois on Apr. 19, 1982, under the Intel College Seminar Program. Dr. Berglund started a laser-based semiconductor manufacturing system company (Ateq Corp. in Beaverton, OR) in 1984 see "Intel Mgr. leaves to start semicon gear Mfg. venture," *Electronic News*, p. 1,22, Jan. 9, 1984, and M. Mehler, "TAGs (Technical Advisory Groups): A form of organized bean spilling," *Electronic Business*, pp. 42-44, Aug. 1, 1986. He recently left and became an independent expert consultant.
[145] G. E. Moore presented a market-driven scenario on the silicon MOS integrated circuit advances to the American Electronics Association San Francisco Council, Feb. 1984. I am indebted to him for a private communication to provide the information. For other recent surveys on the price trend of the DRAM chip, see [146]-[149].
[146] M. P. Lepselter and S. M. Sze, "DRAM pricing trends—the pi rule," IEEE *Circuits and Devices Mag.*, vol. 1, pp. 53-54, Jan. 1985.
[147] R. Bambrick, "Premature erosion of DRAM prices changing economics of memory market," *Electronic News*, vol. 31, no. 1543, p. 1, Apr. 1, 1985.
[148] "DRAMs near millicent per bit, speed up megabit race," *Electronic News*, vol. 31, no. 1556, p. 1, July 1, 1985.
[149] M. R. Leibowitz, "4Mb DRAMs: Tougher to make; tougher to sell?" *Electronic Business*, vol. 14, no. 1, pp. 110-114, Jan. 1, 1988.
[150] J. R. Lineback, "TI turns to EPROMs as technology driver," *Electronics Magazine*, pp. 16-17, Sept. 30, 1985, and "DRAM loses ground as process driver," *Electronics Magazine*, pp. 24-25, Dec. 2, 1985.
[151] J. Thompson, "SRAMs lead DRAM in BICMOS process push," *Electronic News*, p. 23, July 4, 1988.
[152] V. L. Rideout, "One-device cells for dynamic random-access memories: A tutorial," *IEEE Trans. Electron Devices*, vol. ED-26, no. 6, pp. 839-852, June 1979.
[153] P. K. Chatterjee, G. W. Taylor, R. L. Easley, H.-S. Fu, and A. F. Tasch, Jr., "A survey of high-density dynamic RAM cell concepts," *IEEE Trans. Electron Devices*, vol. ED-26, no. 6, pp. 827-839, June 1979.
[154] The data plotted in the figures and discussed in the text were gathered by the author from the following sources up to the latest issues. Digests of Technical Papers of the following annual conferences: ISSCC, IEDM, Symposium on VLSI Technology (U.S.-Japan), all sponsored by the IEEE. *IEEE Trans. Electron Devices* (monthly). *IEEE J. Solid-State Circuits* (bimonthly). *IEEE Spectrum* (monthly). *Solid State Technol.* (monthly). *Electronic News* (weekly newspaper). *Electronics*

(biweekly). *Electronic Business* (biweekly). *Electronic Design* (biweekly). *Computer World* (weekly). *Computer Design* (monthly). *Physics Today* (monthly). *Wall Street Journal* (daily). *Fortune* (monthly). *Sci. Am.* (monthly). The author has added some new data points to the figures since the completion of the first draft on July 4, 1986, up to the date of page setting by the publisher around July 4, 1988.

[155] "A Matter of Substance: . . .IBM has moved volume production of 1M DRAMs to 8-inch silicon wafers (photograph . . . and IBM claims its new 8-inch wafers can yield 450 1M DRAMS compared with 150 on a 5-inch wafers. The (increase) reportedly multiplied IBM's DRAM production in Burlington, VT., so much the company could set aside plans to put new memory capacity in its Manassas, VA, facilities." *Electronic News*, photograph on p. 17, June 6, 1988.

[156] S. P. Murarka, "Refractory silicides for integrated circuits," *J. Vac. Sci. Technol.*, vol. 17, no. 4, pp. 775–792, July/Aug. 1980.

[157] A. K. Sinha, "Refractory metal silicides for VLSI applications," *J. Vac. Sci. Technol.*, vol. 19, no. 3, pp. 778–785, Sept./Oct. 1981.

[158] S. P. Murarka, *Silicides for VLSI Applications*. New York, NY: Academic Press, 1983.

[159] P. A. Gargini, "Interconnect and contact technologies for VLSI applications," in *Solid State Devices 1983*, Conf. Series 69, E. H. Rhoderick, Ed. London, England: The Institute of Physics, 1983, pp. 141–159.

[160] F. M. d'Heurle and P. Gas, "Kinetics of formation of silicides: A review," *J. Material Res.*, vol. 1, no. 1, pp. 205–221, Jan./Feb. 1986.

[161] R. S. Blewer, "Progress in LPCVD tungsten for advance microelectronics applications," *Solid State Technol.*, pp. 117–126, Nov. 1986.

[162] A whole session of five papers are devoted to the tungsten interconnect technology in the 1987-IEDM last December. See papers 9.1 to 9.5, pp. 200–220, in *Technical Digest, 1987 International Electron Device Meeting*, IEEE Catalog No. 87CH2515-5, IEEE Inc., 445 Hoes Lane, Piscataway, NJ.

[163] G. Bulinsky, "Technology in the year 2000," *Fortune*, pp. 92–98, July 18, 1988.

[164] H. Sunami *et al.*, "A corrugated capacitor cell (CCC) for megabit dynamic MOS memories," in *Technical Digest, IEEE Int. Electron Device Meeting*, pp. 806–808, Dec. 1982. Also published as "A corrugated capacitor cell (CCC)," *IEEE Electron Device Lett.*, vol. EDL-4, pp. 90–91, 1983, and *IEEE Trans. Electron Devices*, vol. ED-31, pp. 746–753, 1984.

[165] K. Nakamura, M. Yanagisawa, Y. Nio, K. Okamura, and M. Kikuchi, "Buried isolation capacitor (BIC) cell for megabit MOS dynamic RAM," in *Technical Digest of the IEEE Int. Electron Devices Meeting*, pp. 236–239, Dec. 9–12, 1984. IEEE Catalog No. 84CH2099-0.

[166] S. Nakajima, K. Miura, K. Minegishi, and T. Morie, "An isolation merged vertical capacitor cell for large capacity DRAM," in *Technical Digest of the IEEE Int. Electron Devices Meeting*, pp. 240–243, Dec. 9–12, 1984. IEEE Catalog No. 84CH2099-0.

[167] M. Wade, K. Hieda, and S. Watanabe, "A folded capacitor cell (F.C.C.) for future megabit DRAMS," in *Technical Digest of the Int. Electron Devices Meeting*, pp. 244–247, Dec. 9–12, 1984. IEEE Catalog No. 84CH2099-0.

[168] M. Elahy *et al.*, "Trench capacitor leakage in Mbit DRAM," in *Technical Digest of the Int. Electron Devices Meeting*, pp. 248–251, Dec. 9–12, 1984.

[169] N. Lu *et al.*, "The SPT cell—a new substrate-plate trench cell for DRAMs," in *Technical Digest of Int. Electron Device Meeting*, pp. 771–772, Dec. 1–4, 1985. IEEE Publication Catalog No. 85CH2252-5.

[170] H. Sunami, "Cell structures for future DRAMs," in *Technical Digest of the 1985 Int. Electron Device Meeting*, pp. 694–697, Dec. 1–4, 1985. IEEE Publication Catalog No. 85CH2252-5.

[171] N. Chau-chun Lu, "Advanced cell structures for dynamic RAMs," in *Proc. 1987 Symp. on VLSI-TSA*, pp. 163–168, May 13–15, 1987.

[172] P. K. Chatterjee was a former graduate student in the semiconductor physics course, EE/Physics 435, I taught during spring semester of 1974 at the University of Illinois in Urbana with the assistance of M. McNutt, S. Pantelides, and T. H. Ning as teaching assistants during the early 1970 offerings. Pallab was one of the best students I have taught in this advanced semiconductor physics course over the years. Chatterjee won the J. J. Ebers Award of the IEEE Electron Device Society in 1986 for leading this and other TI's DRAM development effort.

[173] W. F. Richardson *et al.* and P. K. Chatterjee, "A trench transistor cross-point DRAM cell," in *Technical Digest, Int. Electron Device Meeting*, pp. 714–717, Dec. 1–4, 1985. IEEE Catalog No. 85CH2252-5.

[174] S. K. Banerjee *et al.* and P. K. Chatterjee, "Characterization of trench transistors for 3-D memories," in *Digest of Technical Papers, 1986 Symp. on VLSI Technology*, pp. 79–80. IEEE Catalog No. 86CH2318-4.

[175] For a pedestrian account, R. Ristelhueber, "TI Redesigning 4M DRAM to fit 300-mil package by '89," *Electronic News*, vol. 32, no. 1612, p. 32, July 28, 1986.

[176] R. Bambrick, "Shift to CMOS FMVs (Foreign Market Value) may hike DRAM tags," *Electronic News*, p. 1, 22–23, July 4, 1988.

[177] M. Inoue *et al.* (Matsushita Semiconductor Research Center), "A 16Mb DRAM with an open bit-line architecture," in *Technical Digest of Papers, ISSCC XXXI*, pp. 246–247, Feb. 1988.

[178] S. Watanabe *et al.* (Toshiba VLSI Research Center), "An experimental 16Mb CMOS DRAM chip with a 100MHz serial read/write mode," in *Technical Digest of Papers, ISSCC XXXI*, pp. 248–249, Feb. 1988.

[179] M. Aoki *et al.* (Hitachi Central Research Laboratory), "An experimental 16Mb DRAM with transposed data-line structure," in *Technical Digest of Papers, ISSCC XXXI*, pp. 250–251, Feb. 1988. For a pedestrain account of these three papers [177]–[179], see B. C. Cole, "The next wave: 16-MBIT DRAMS FROM JAPAN," *Electronics*, vol. 61, no. 4, pp. 68–69, Feb. 18, 1988.

[180] A. S. Oberai, "Lithography—challenges of the future," *Solid State Technol.*, vol. 30, no. 9, pp. 123–128, Sept. 1, 1987. For an earlier review, see R. K. Watts and J. H. Bruning, "A review of fine-line lithographic techniques: present and future," *Solid State Technol.*, vol. 24, no. 5, pp. 99–105, May 1981.

[181] L. Waller, "Fine tuning optical steppers," *Electronics*, pp. 134–135, Oct. 15, 1987.

[182] R. D. Moore, "EL systems: high throughput electron beam lithography tools," *Solid State Technol.*, vol. 26, no. 9, pp. 127–132, Sept. 1, 1983.

[183] H. C. Pfefer, "Recent advances in electron-beam lithography for the high volume production of VLSI devices," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 663–674, Apr. 1979.

[184] W. D. Grobman *et al.*, "1 μm VLSI technology: Part IV—electron-beam lithography," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 360–368, Apr. 1979.

[185] H. N. Yu, *et al.*, "Fabrication of a miniature 8k-bit memory chip using electron-beam exposure," *J. Vac. Sci. Technol.*, vol. 112, no. 6, pp. 1297–1300, Nov./Dec. 1975.

[186] K. J. O'Connor and R. A. Kushner (Bell Labs.), "A 5ns 4K $\times$ 1 NMOS static RAM," in *Technical Digest 1983 IEEE Int. Solid State Circuits Conf.*, pp. 104–105, 201, 1983.

[187] G. A. Garrettson and A. P. Neukermans, "X-ray lithography," *Hewlett-Packard J.*, pp. 14–17, Aug. 1982.

[188] "The x-ray stepper heads for the VLSI production line," *Electronics*, pp. 41–44, Mar. 10, 1986.

[189] J. Lyman, "It's a three-way race in x-ray lithography," *Electronics*, pp. 46–49, Mar. 17, 1986.

[190] C. L. Cohen, "Japan kicks off x-ray fabrication project," *Electronics*, vol. 44, p. 44, Aug. 7, 1986.

[191] M. Miyake *et al.* (NTT), "Subhalf-micrometer p-channel MOSFET's with 3.5-nm gate oxide fabricated using x-ray lithography," *IEEE Electron Device Lett.*, vol. EDL-8, no. 6, pp. 266–268, June 1987.

[192] C. L. Cohen, "NEC draws 0.2-micron lines with x-rays," *Electronics*, p. 34, Oct. 29, 1987.

[193] G. E. Fuller, "Optical lithography status," *Solid State Technol.*, vol. 30, no. 9, pp. 113–118, Sept. 1, 1987.

[194] C. Stedman, "Excimer-laser stepper, Report IBM, AT$T order GCA prototype," *Electronic News*, p. 45, 47, Monday, June 22, 1987. See also B. Santo, "SRC deep-UV litho. effort hit by funding problems," *Electronic News*, p. 45, 47, June 22, 1987.

[195] For a recent development on photoresist and earlier technical references, see K. J. Orvek, W. C. Cunningham, Jr., and J. C. McFarland, "An organosilicon photoresist for use in

excimer laser lithography," in *Technical Digest of 1987 IEDM*, paper 32.5, pp. 929–932, Dec. 1987.
[196] J. Fallon and B. Santo, "IBM orders compact synchrotron," *Electronic News*, p. 45, June 22, 1987.
[197] J. Robertson, "IBM to develop 0.25-micron ASIC gate array," *Electronic News*, p. 25, June 13, 1988. For a recent development, see "IBM claims X-ray litho advance, fully-scaled '0.25 μm' NMOS ICs at for exposure levels reported," *Electronic News*, p. 38, Sept. 12, 1988.
[198] See *Electronics Week*, p. 18, Apr. 15, 1985; "IBM facility's chip research progressing," *Computer World*, p. 88, Apr. 22, 1985; "IBM works toward 16M-Bit Ram," *Mini-Micro Systems*, July 1985. All of these three news releases stated that IBM Research employed a 0.5 micron aluminum metal NMOS electric beam technology to produce chips with 16M-bit RAM with 8.5 square micron cell area and 100 000 logic elements with 1700 MOSFETs per 100 × 100 square micron area.
[199] B. Santo, "Say IBM signs Grumman for E-beam mfg.," *Electronic News*, p. 38, Dec. 1, 1986. (This reported the IBM order of 40–50 direct-write electron beam systems.)
[200] "IBM researchers build ULSI devices," *Electronics*, vol. 60, no. 17, p. 110, Aug. 20, 1987.
[201] "Getting tight in Yorktown Heights," *Electronic News*, p. 29, Aug. 17, 1987. For a later account and photograph of the electron beam lithography machine, see J. A. Armstrong, "Solid state technology and the computer: 40 years later, small is still beautiful," *Solid State Technol.*, pp. 81–83, Dec. 1987.
[202] G. A. Sai-Halasz *et al.*, "Experimental technology and characterization of self-aligned 0.1μm-gate-length low-temperature operation NMOS devices," in *Technical Digest of IEDM87*, paper 16.5, pp. 397–400, Dec. 1987.
[203] M. Murphy and L. Morgenthaler, "Why chips will soon kill off disk drives," *Electronic Business*, pp. 74–78, Nov. 1, 1987.
[204] E. W. Pugh, *Memories That Shaped an Industry—Decisions Leading to IBM System/360.* Cambridge, MA: The MIT Press, 1984.
[205] For the 1946 von Neumann prophecy, see p. 34 of G. Bell and A. Newell, *Computer Structures: Readings and Examples.* New York, NY: McGraw-Hill, 1971.
[206] H. Z. Bogert, "FERRAM: the memory the market always wanted," *Dataquest Newsletter*, Dataquest, Inc., Jan. 1988.
[207] This 1-percent assumption strongly depends on application and CPU architecture. There are many detailed studies in recent years [97].
[208] For CMOS development history, see R. D. Davies, "The case for CMOS," *IEEE Spectrum*, vol. 20, no. 10, pp. 26–32, Oct. 1983 and the following articles [209]–[214].
[209] B. C. Cole, "CMOS memories replacing n-MOS in megabit storage chips," *Electronics Week*, pp. 53–61, Nov. 26, 1984.
[210] ——, "CAD, CMOS, and VLSI are changing analog world," *Electronics*, pp. 35–39, Dec. 23, 1985.
[211] Y. Nishi, "Direction of VLSI CMOS technology," *Hewlett-Packard J.*, pp. 24–25, June 1987.
[212] R. Wilson, "CMOS VLSI sets the direction for new ICs," *Computer Design*, pp. 79–91, Dec. 1987.
[213] For a latest report on optical submicron technology for fabricating CMOS, see "0.25 μm CMOS technology using p+ polysilicon gate PMOSFET," in *Technical Digest of 1987 IEDM*, paper 15.5, pp. 367–370, Dec. 1987, and other articles in this digest. The readers should note that the so-called CMOS DRAM chip employs the one-transistor one-capacitor Dennard NMOS DRAM cell for the memory bits and the CMOS inverter gate is used only for peripheral drivers while a SRAM chip uses CMOS for both the six-transistor bistable flip-flop memory bit [134] as well as the peripheral drivers. The commonly used description 'CMOS DRAM,' in trade releases and engineering articles could be misleading since it could be interpreted as to mean that the memory cell is made of CMOS which is not.
[214] "CMOS overtakes NMOS: ICE report," *Electronic News*, p. 28, Mar. 14, 1988. See also the Feb. 4, 1988, issue of the *Electronics* magazine, vol. 61, no. 3, pp. 55–69, which discussed the question of BICMOS as the next technology driver and the production efforts of several American companies (TI, National, LSI Logic, AMCC-American Micro Circuit Corp., and Saratoga).
[215] The integrated bipolar-mosfet (BIMOS) cell was first used by C. T. Sah in his 1961 SCT (Surface-controlled Tetrode) experiments [77], [81] which were patented in Sah's two patents [78], [80]. The cross-sectional view of the structure used by Sah is given in the second patent of Sah [80] and also the 1963 McGraw-Hill Yearbook [79] shown in Fig. 6(c).
[216] For additional accounts of recent developments on BIMOS and BICMOS, see C. L. Cohen, "Cells combine CMOS, bipolar transistors," *Electronics Week*, p. 17–18, Nov. 5, 1984, and the following references [217]–[220].
[217] L. Waller, "Behind Motorolas silence: ambitious product plans, company unwraps a slew of products, commits to BIMOS," *Electronics*, pp. 18–19, Nov. 25, 1985.
[218] E. Connolly, "BiMOS devices give designers the best of two worlds," and "How BiMOS reduces delays," *Computer Design*, pp. 22–25, June 1, 1987.
[219] C. L. Cohen, "Here comes a 256-k SRAM (1-μm BiCMOS) and it's from Fairchild," *Electronics*, pp. 34–35, June 11, 1987.
[220] S. Weber, "TI to roll out its first family of products made with BiCMOS," *Electronics*, pp. 81–82, Oct. 1, 1987.
[221] P. H. Solomon, "A comparison of semiconductor devices for high-speed logic," *Proc. IEEE*, vol. 70, no. 5, pp. 489–509, May 1982.
[222] F. Gaensslen and D. E. Nelsen, "Solid-state devices—low temperature device operation," in *Technical Digest of IEDM-87*, pp. 379–408, IEEE Publication Catalog No. 87CH2515-5.
[223] L. M. Terman, 1967 (unpublished) [97].
[224] A system level experiment was performed by Digital Equipment Corporation researchers in which a 3-micron 2-chip 32-bit CMOS based CPU (DEC part DCJ11) was cooled from 300 to 77K. A factor of two improvement in the maximum clock frequency was observed, from 20 to 40 MHz. See G. Gildenblat, L. Colonna-Romano, D. Lau, and D. E. Nelsen, "Investigation of cryogenic CMOS performance," in *Technical Digest of Int. Electron Device Meeting*, pp. 268–271, Dec. 1-4, 1985. IEEE Publication Catalog No. 85CH2252-5.
[225] H. Morkoc and P. M. Solomon, "The HEMT: a superfast transistor," *IEEE Spectrum*, pp. 28–35, Feb. 1984.
[226] P. M. Solomon and H. Morkoc, "Modulation doped GaAs/AlGaAs heterojunction field effect transistors (MODFETs): Ultra high speed devices for super computer," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 1015–1028, 1984.
[227] For a recent account on GaAs devices in Japan, see T. E. Bell, "Japan reaches beyond silicon," *IEEE Spectrum*, pp. 46–52, Oct. 1985.
[228] J. W. Goodman, F. I. Leonberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proc. IEEE*, vol. 72, no. 7, pp. 850–866, July 1984. See also the whole issue which is devoted to optical electronics and integrated optics.
[229] B. Y. Tsaur *et al.*, "Low dislocation density GaAs epilayers grown on Ge-coated Si substrates by means of lateral epitaxial growth," *Appl. Phy. Lett.*, vol. 41, no. 4, pp. 347–349, Apr. 1982.
[230] R. M. Fletcher, D. K. Wagner, and J. M. Ballantyne, "GaAs light-emitting diodes fabricated on Ge-coated Si substrates," *Appl. Phys. Lett.*, vol. 44, pp. 967–969, 1984.
[231] T. H. Windhorn, G. M. Metze, B.-Y. Tsaur, and J. C. C. Fan, "AlGaAs double-heterostructure diode lasers fabricated on a monolithic GaAs/Si substrate," *Appl. Phys. Lett.*, vol. 45, pp. 309–311, 1984.
[232] T. Windhorn and G. M. Metz, "Room temperature operation of GaAs/AlGaAs diode lasers fabricated on a monolithic GaAs/Si substrate," *Appl. Fhys. Lett.*, vol. 47, pp. 1031–1033, 1985.
[233] R. Fischer *et al.*, "GaAs/AlGaAs MOSFETs grown directly on (100) silicon," *Electronics Letters*, vol. 20, pp. 945–947, 1984.
[234] R. Fischer *et al.*, "Monolithic integration of GaAs/AlGaAs modulation doped field effect transistors and N-metal oxide semiconductor silicon circuits," *Appl. Phys. Lett*, vol. 47, pp. 983–985, 1985.
[235] H. K. Choi, B.-Y. Tsaur, G. M. Metze, G. W. Turner, and J. C. C. Fan, "GaAs MESFETs fabricated on monolithic GaAs/Si substrates," *IEEE Electron Devices Lett.*, vol. EDL-5, pp. 207–208, 1984.
[236] J. Nonake, M. Akiyama, Y. Kawarada, and K. Kaminiski, "Fabrication of GaAs MESFET ring oscillator on MOCVD grown on GaAs/Si(100) substrate," *Japan J. Appl. Phys.*, vol. 23, pp. L919–L921, 1984.
[237] R. Fischer *et al.*, "GaAs/AlGaAs heterojunction bipolar transistors on Si substrates," in *Digest of Int. Electron Devices*

Meeting, pp. 332–335, Dec. 2–4, 1985. IEEE Catalog No. 85CH2252-5.
[238] R. Fischer *et al.*, "GaAs bipolar transistors grown on (100) Si substrate by molecular beam epitaxy," *Appl. Phys. Lett.*, vol. 47, pp. 397–399, 1985.
[239] A. Y. Cho, "Recent advances in GaAs on Si," in *Technical Digest, IEDM-87*, pp. 901–904, Dec. 1987. IEEE Publication No. 87-CH2515-5.
[240] H. Shichijo and R. J. Matyi, "GaAs-on-Silicon integrated circuits: savior for GaAs? Or for Si?," in *Technical Digest, IEDM-87*, pp. 88–91, Dec. 1987. IEEE Publication No. 87-CH2515-5.
[241] "TI claims Si, GaAs linked on chip," *Electronic News*, p. 20, July 4, 1988. A cross-sectional view is given of a GaAs MESFET in a GaAs epitaxy layer grown on a 3°−off>100> Si substrate and integrated with a twin-well Si CMOS by R. J. Matyi and H. Shichijo.
[242] H. Shichijo is another former graduate student at the University of Illinois who took my semiconductor physics course, EE/Physics 435, in the spring semester of 1977. He was one of the best students of that class.
[243] Data released by DARPA (Defense Advanced Research Project Agency), see R. Beresford, "Military VLSI technology," *VLSI Design*, pp. 46–54, Apr. 1984.
[244] Progresses are being made in the last two years due to a $60M U.S. government funding, for example, see H. Bierman, "Move over silicon! Here comes GaAs making complex devices starts to get easier," *Electronics*, pp. 39–44, Dec. 2, 1985, and the following references [245], [246].
[245] T. Naegele, "U.S. GaAs makers move to catch up," *Electronics*, p. 17, Mar. 17, 1986. See however, [246].
[246] A. Wing, "GaAs startups struggle with high costs, low densities and sparse yield," *Electronic News*, vol. 32, no. 1549, p. 1, Mar. 24, 1986.
[247] "Rockwell breaks 1% yield barrier for GaAs 16kbit RAM," *Electronics*, p. 162, Jan. 7, 1988.
[248] L. Waller, "Commercial quantities of LSI GaAs are finally here," *Electronics*, pp. 48–49, Sept. 17, 1987.
[249] W. R. Iversen, "Cray is still using GaAs—but not its own," *Electronics*, p. 49, Sept. 17, 1987. See also [265].
[250] A. Pollack, "Gallium arsenide chips making research gains," *The New York Times*, Sunday, Oct. 25, 1987.
[251] "Kopin offers GaAs-on-Silicon wafers for research," *Electronic News*, p. 39, Oct. 19, 1987.
[252] V. Rice, "The making of a military GaAs market," *Electronic Business*, pp. 46–50, July 15, 1987.
[253] J. Lineback, "Hybrid (GaAs-Si) amp(lifier) has double the AGC of silicon only," *Electronics*, pp. 170–172, Jan. 7, 1987.
[254] C. T. Sah, "Models and experiments on degradation of oxidized silicon," in *Proceeding of IFOS-87*, also in *Applied Surface Science*, vol. 30, pp. 311–12, 1987.
[255] ——, "VLSI device reliability modeling," in *Proceedings of 1987 Int. Symp. on VLSI Technology, Systems and Applications*, pp. 153–162, May 13–15, 1987.
[256] For a review of the fundamental physics and analysis of the published data related to failure of silicon dioxide film on silicon, see [87].
[257] C. T. Sah, "Hydrogenation and dehydrogenation of shallow acceptors and donors in Si: fundamental phenomena and survey of the literature," in *Properties of SILICON*. London, England: INSPEC, The Institution of Electrical Engineers, section 17.16, pp. 584–604, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.
[258] C. T. Sah *et al.*, "Hydrogenation and dehydrogenation of shallow acceptors and donors in Si: kinetic rate data," in *Properties of SILICON*. London, England: INSPEC, The Institution of Electrical Engineers, section 17.17, pp. 604–613, May 1988. Available from INSPEC Dept., IEEE Service Center, 455 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.
[259] For a pedestrain account, see W. R. Iversen, "Model may help solve chip-reliability problem," *Electronics*, vol. 59, no. 12, pp. 21–24, Mar. 24, 1986.
[260] Semiconductor Research Corporation, "Industrial Mentor Handbook," ch. III, pp. 9–17, SRC Publication S87011, July 1987.
[261] M. Eleccion, "The electronic watch," *IEEE Spectrum*, vol. 10, no. 4, pp. 24–32, Apr. 1973.
[262] J. Robertson, "SIA: DRAM Consortium are eyed," *Electronic News*, p. 6, Sept. 19, 1988; also J. Titta, "Drumming up DRAM support," *Computerworld*, p. 8, Oct. 31, 1988.
[263] *Electronic News*, p. 36, Oct. 3, 1988; also A. M. Hayashi, "How Micron is marching through the DRAM minefield," *Electronic Business*, pp. 42–46, June 15, 1988.
[264] "ECL vs. GaAs chips: Let the clock-speed games begin," *Electronic Business*, pp. 103–104, June 15, 1988.
[265] For recent Cray-3 problem due to GaAs chips, see J. Bailey and R. Gibson, "Cray stumbles with new supercomputers," *Wall Street J.*, p. 1, Oct. 26, 1988; J. S. Bozman, "Cray revenue slips, Y-MP, Cray-3 woes hurt supercomputer firm," *Computerworld*, p. 6, Oct. 31, 1988; "Snags in GaAs circuits hit Cray 3; Charge: $10M," *Electronic News*, p. 14, Oct. 31, 1988.
[266] "IBM may go to GaAs for '90s CPUs: Study," *Electronic News*, p. 46, Sept. 12, 1988.

Chih-Tang Sah (Fellow, IEEE) was born in Beijing, China, in November 1932, and came to the United States in September 1949.

He began his transistor career in February 1953 at the University of Illinois in Urbana when he audited Prof. John Bardeen's second giving of the graduate course titled, "Conduction of Electricity in Semiconductors." From 1953 to 1956, he was diverted to a doctoral thesis research project on an artificial periodic structure, the traveling wave tube, at Stanford University under the tutelage of Prof. Karl Spangenberg. From 1956 to 1959, he apprenticed with William Shockley at the Shockley Transistor Laboratory where he wrote the p-n junction recombination-current paper in the *Proceedings of the IRE* with Robert N. Noyce and William Shockley. In 1959 he joined the Fairchild Semiconductor Corporation and later became the Manager and Head of the Physics Department of the Fairchild Semiconductor Laboratory where he led a team which grew to sixty-five members and developed many of the first generation silicon integrated circuit technology during 1959–1964, including the stable MOS technology with Bruce Deal, Andy Grove, and Ed Snow for which they received the Franklin Institute Prize in 1975 and the CMOS circuit with Frank Wanlass, Gordon Moore, and Vic Grinich in 1962. A paper on the effects of surface recombination and channel on diode and transistor characteristics in the 1962 *IRE Transactions on Electron Devices* won him the Browder J. Thompson Prize of the Institute of Radio Engineers for the best paper of an author under thirty. In 1962, he was recruited by Profs. John Bardeen and Edward C. Jordan, Head of the Electrical Engineering Department of the University of Illinois, and joined the faculty in 1963 as a Professor of Electrical Engineering and a Professor of Physics. At Illinois, he started the undergraduate teaching laboratory on transistor fabrication in 1964 under an NSF undergraduate instructional equipment grant which has provided hands-on experiences to 4000 undergraduate junior and senior electrical engineering students. He directed 40 doctoral theses in electrical engineering and in physics and wrote about 250 journal articles with his graduate students and colleagues. He was listed as one of the world's 1000 most cited scientists during 1965–1978 in a survey by the Institute of Scientific Information of Philadelphia and has given more than 100 invited lectures on transistor physics and technology. His current research interest concerns the aging and failure mechanisms of silicon transistors and integrated circuits under the sponsorship of the Semiconductor Research Corporation.

Dr. Sah was awarded the Doctor Honoris Causa by the University of Leuven, Belgium, in 1975, the J. J. Ebers Award by the IEEE Electron Device Society in 1981, and the first achievement award by the Asian American Manufacturing Association for contributions in transistor physics and technology. He is a Fellow of the American Physical Society and a member of the U.S. National Academy of Engineering. On August 21, 1988, he was appointed the first Eminent Scholar Chair in the College of Engineering of the University of Florida at Gainesville, the Robert C. Pittman Chair, where he was also appointed a Graduate Research Professor and a Professor of Electrical Engineering.