

Chapter 2

From Microelectronics to Nanoelectronics

Bernd Hoefflinger

Abstract We highlight key events in over 100 years of electronic amplifiers and their incorporation in computers and communication in order to appreciate the electron as man's most powerful token of information. We recognize that it has taken about 25 years or almost a generation for inventions to make it into new products, and that, within these periods, it still took major campaigns, like the Sputnik effect or what we shall call $10\times$ programs, to achieve major technology steps. From Lilienfeld's invention 1926 of the solid-state field-effect triode to its realization 1959 in Kahng's MOS field-effect transistor, it took 33 years, and this pivotal year also saw the first planar integrated silicon circuit as patented by Noyce. This birth of the integrated microchip launched the unparalleled exponential growth of microelectronics with many great milestones. Among these, we point out the 3D integration of CMOS transistors by Gibbons in 1979 and the related Japanese program on Future Electron Devices (FED). The 3D domain has finally arrived as a broad development since 2005. Consecutively, we mark the neural networks on-chip of 1989 by Mead and others, now, 20 years later, a major project by DARPA. We highlight cooperatives like SRC and SEMATECH, their impact on progress and more recent nanoelectronic milestones until 2010.

2.1 1906: The Vacuum-Tube Amplifier

At the beginning of the twentieth century, the phenomenon of electricity (the charge and force of electrons) had received over 100 years of scientific and practical attention, and signals had been transmitted by electromagnetic waves, but their detection was as yet very limited, because signal levels were small and buried in noise. This changed forever when the vacuum-tube amplifier was invented in 1906

B. Hoefflinger (✉)
Leonberger Strasse 5, 71063 Sindelfingen, Germany
e-mail: bhoefflinger@t-online.de

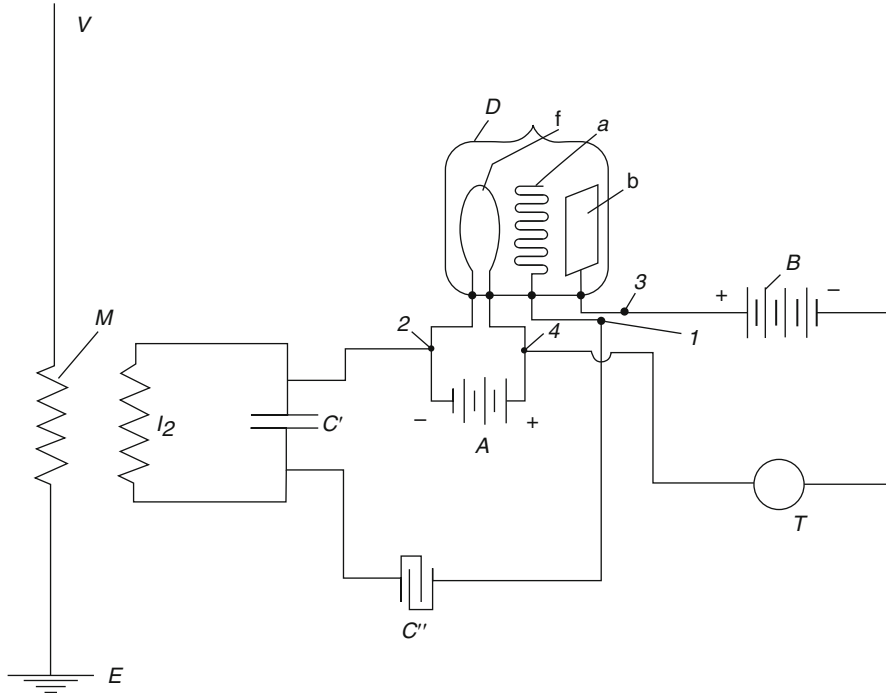
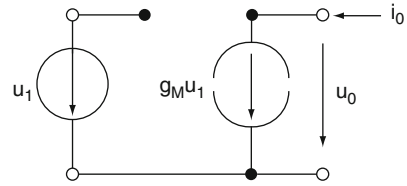


Fig. 2.1 The vacuum-triode amplifier after De Forest 1907. In this field-effect triode, the electrodes are the heated cathode on the left, which we would call the source today, the grid in the center, which would be the gate, and the anode on the right, which we would call the drain (© USPTO)

by Robert von Lieben in Austria [1] and Lee De Forest in the USA [2]. Its predecessor was the vacuum-discharge diode, a two-terminal device consisting of a heated cathode electrode emitting *thermionic electrons*, which are then collected through a high electric field by another electrode, the anode, biased at a high voltage against the cathode. This two-terminal device acts as a rectifier, offering a large conductance in the described case of the anode being at a higher potential than the cathode, and zero conductance in the reverse case of the anode being at a potential lower than the cathode. The invention was the insertion of a potential barrier in the path of the electrons by placing a metal grid inside the tube and biasing it at a low potential with respect to the cathode (Fig. 2.1). The resulting electric field between cathode and anode would literally turn the electrons around. Fewer or no electrons at all would arrive at the anode, and the conductance between cathode and anode would be much smaller. A variation of the grid potential would produce an analogous variation (modulation) of the cathode–anode conductance. This three-terminal device, the vacuum triode, consisting of cathode, anode, and control grid, became the first electronic amplifier: It had a certain voltage gain A_V , because the grid–cathode input control voltage could be made much smaller than the cathode–anode voltage, and it had infinite current gain A_I at low rates of input

Fig. 2.2 Equivalent circuit of an ideal amplifier (voltage-controlled current amplifier)



changes, because there was no current flow in the input between cathode and grid, while large currents and large current changes were effected in the output circuit between cathode and anode. As a consequence, the power gain $A_V A_I$ approaches infinity.

It is worthwhile drawing the abstraction of this amplifier as a circuit diagram (Fig. 2.2), because inventors have by now spent over 100 years improving this amplifier, and they will spend another 100, even if the signal is not electrons. We see that the input port is an open circuit, and the output port is represented by a current source $g_m V_{in}$ in parallel with an output resistance R_{out} .

The inventors of the vacuum-tube amplifiers were tinkerers. They based their patent applications on effects observed with their devices and achieved useful products within just a few years (1912).

10×: Long-range radio (World War I)

These amplifiers launched the radio age, and they triggered the speculative research on building controlled-conductance amplifying devices, which would replace the bulky vacuum tubes with their light-bulb-like lifetime problems.

2.2 1926: The Three-Electrode Semiconductor Amplifier

The best solid-state analogue to the vacuum tube would be a crystal bar whose conductance could be varied over orders of magnitude by a control electrode. This is what the Austro-Hungarian–American physicist Julius Lilienfeld proposed in his 1926 patent application “Method and apparatus for controlling electric currents” [3]. He proposed copper sulfide as the semiconducting material and a capacitive control electrode (Fig. 2.3). This is literally a parallel-plate capacitor, in which the field from the control electrode would have an effect on the conductance along the semiconducting plate.

He did not report any practical results. However, since the discovery of the rectifying characteristics of lead sulfide by K.F. Braun in 1874, semiconductors had received widespread attention. However, it was not until 1938 that Rudolf Hilsch and Richard Pohl published a paper, “Control of electron currents with a three-electrode crystal and a model of a blocking layer” [4], based on results obtained with potassium bromide. Shockley wrote, in his article for the issue of the *IEEE Transactions on Electron Devices* commemorating the bicentennial of the United

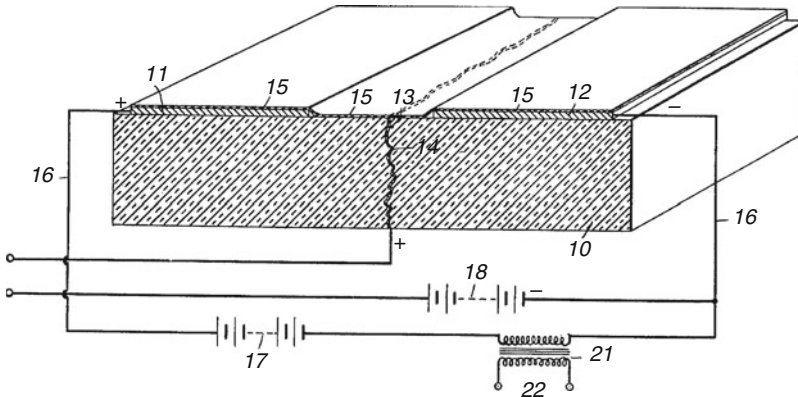


Fig. 2.3 The field-effect triode proposed by Lilienfeld in 1926 (© USPTO)

States in 1976 [5], that he had this idea in December 1939: “It has occurred to me today that an amplifier using semi conductors rather than vacuum is in principle possible”. Research continued during World War II on the semiconductor amplifier [6], but a critical effort began only after the war. As we shall see, it was not until 1959 that the Lilienfeld concept was finally reduced to practice.

2.3 1947: The Transistor

One possible launch date of the Age of Microelectronics is certainly the invention of the transistor in 1947 [5]. Shockley himself described the events leading to the point-contact transistor as the *creative failure mechanism*, because the invention resulted from the failure to achieve the original goal, namely a field-effect transistor (FET) with an insulated gate in the style of the Lilienfeld patent. Nevertheless, this *failure*, implemented as Ge or Si bipolar junction transistors, dominated microelectronics into the 1980s, when it was finally overtaken by integrated circuits based on insulated-gate FETs, the realization of the Lilienfeld concept.

2.4 1959: The MOS Transistor and the Integrated Circuit

Shockley described the first working FET in 1952, which used a reverse-biased pn junction as the control gate, and junction FETs (JFETs) were then used in amplifiers, where a high input impedance was required. In fact, when I was charged in 1967 at Cornell University with converting the junior-year lab from vacuum tubes to transistors, I replaced the vacuum triodes in the General Radio bridges by junction FETs, and one of my students wrote in his lab report: “The field-effect transistor thinks that it is a tube”.

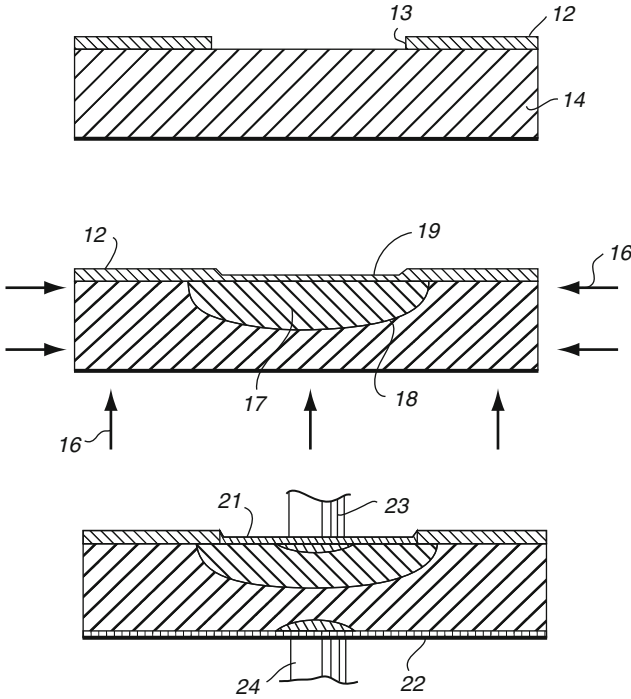


Fig. 2.5 The planar manufacturing process according to Hoerni (© USPTO)

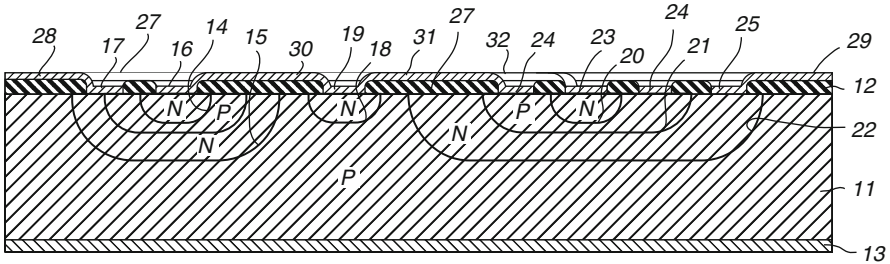


Fig. 2.6 The planar integrated circuit according to Noyce (patent filed 1959) [10] (© USPTO)

1,000 times smaller, the basic method of integrating devices side-by-side in Si is still the same, resulting in a two-dimensional arrangement of transistors and other devices, today at the scale of billions of these on the same chip.

The events of 1956–1959, pivotal for microelectronics, have been covered in great detail, for example in [6 and 11], and we will not elaborate here on the parallel patent by Kilby. It did not have any technical impact because of its Ge mesa technology and soldered flying wires as interconnects.

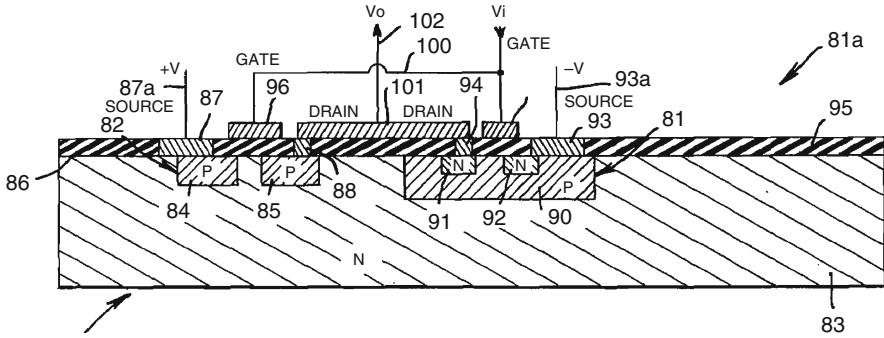
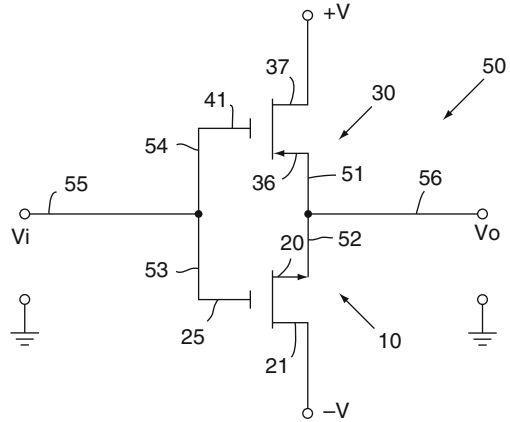


Fig. 2.7 Cross section through the chip surface showing a PMOS and an NMOS transistor side-by-side and isolated from each other, after the Wanlass patent filed in 1963 [12] (© USPTO)

Fig. 2.8 The complementary MOS inverter in the Wanlass patent. Note that the gates are connected and that the PMOS drain is connected to the NMOS drain, making this pair an important fundamental functional unit (© USPTO)



While the Noyce patent showed a bipolar transistor, others in his company concentrated on MOS and came up in 1963 with the ultimate solution for integrated circuits: *complementary MOS (CMOS) integrated circuits*. Frank Wanlass filed this famous patent in 1963 [12], and he presented a paper in the same year with C.T. Sah [13]. Nothing explains the power of this invention better than the figures in the patent (Figs. 2.7 and 2.8). The first one shows a PMOS transistor on the left with p-type doped source and drain, conducting positive carriers (holes) under the control of its gate, and an NMOS transistor on the right with its n-type source and drain conducting electrons under the control of its gate. The transistors are isolated from each other by a reverse-biased p-n junction consisting of a p-well and an n-type substrate. The construction and functionality are listed in Table 2.1.

The complementary transistor pair as connected in Fig. 2.8 is the world's most perfect inverting amplifier for the restoration and propagation of digital signals. It establishes a perfect HI and a perfect LOW at the output V_o (no. 56). There is zero current flow, which means zero power consumption except during a signal transition, when the PMOS would provide current to charge the output to HI and the

Table 2.1 Functionality of complementary MOS transistors

	Charge carriers	Threshold	ON voltage	OFF voltage	Drain–source voltage
NMOS	neg	pos	pos	neg	pos
PMOS	pos	neg	neg	pos	neg

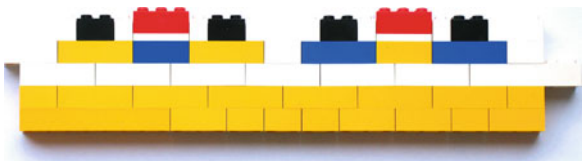


Fig. 2.9 Cross section of a CMOS technology structure on an insulator as the substrate. The Lego blocks indicate materials and aspect ratios. *Yellow*: n-Type Si (doped with donor-type atoms such as As and P), *Blue*: p-Type Si (doped with acceptor-type atoms such as B), *Red*: Metal (or as yet mostly highly As-doped polycrystalline Si), *White*: SiO₂ insulator, *Black*: Interconnect metal

NMOS would discharge the output to LOW. As an amplifier, it has an operating region with infinite voltage gain so that it is the best ever embodiment of the ideal amplifier as shown in Fig. 2.2.

Fairchild did not pursue this ingenious invention, and Wanlass eventually left to join the dedicated MOS companies General Instruments and AMI. His boss Noyce did not embrace CMOS until 1980, almost 20 years later. The real CMOS pushers were at RCA, where a pervasive CMOS culture was established through the first textbook on FETs [14]. They also went one step further towards the ideal CMOS structure by building the transistor layer in a thin Si film on insulator (later named SOI), Fig. 2.9.

This obviously optimizes the vertical isolation, it allows a higher transistor density, and it minimizes parasitic capacitances. In the mid-1960s at RCA, they chose sapphire as the insulating substrate, which was expensive and did not match with the Si lattice so that this approach was only used for space chips because of its immunity to radiation. As of about 2000, 35 years later, the SOI structure finally gained ground, based on massive process development over the last 20 years.

Back at Fairchild, Gordon Moore saw by 1965 the potential for doubling the number of transistors per chip with every new generation of planar IC technology, and in 1965 he had already sufficient data to show that this would happen every 18 months [15], as shown in Fig. 2.10, a reproduction of his 1975 update on this famous curve [16].

Another powerful and lasting driving force for integrating more and more functionality on a single chip is the minimum power and superior reliability achievable in the planar process. We highlight this here with the observation that, if we succeed in doubling the number of transistors or digital gates composed by these on a single piece of silicon, the number of contacts to the outside world will only increase by 40%, or the square root of 2. This is the famous Rent’s rule. Because contacts mean cost and a reduction of reliability (lifetime), this means extra pressure on increasing the chip area beyond the force provided by Moore’s

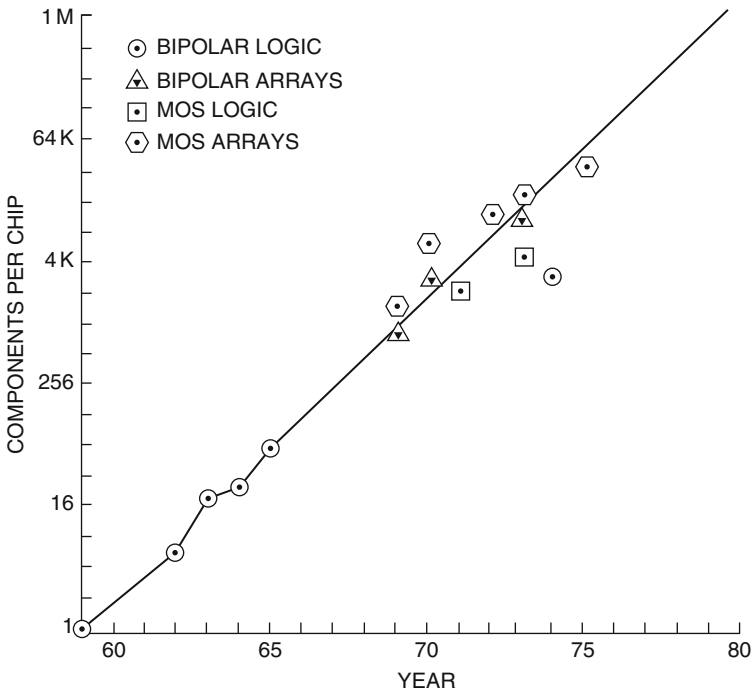


Fig. 2.10 Components per chip vs. year after Gordon Moore, 1965 and 1975 [16] (© 1975 IEEE)

law. These forces led to a new manufacturing paradigm, namely that of producing chips with a manufacturing percentage yield smaller than the high nineties customary in classical manufacturing. This was a fundamental debate in the 1960s between the classical camp of perfect-scale or right-scale integration and the advocates of large-scale integration, who won and went through very-large-scale integration (VLSI) to the giga-scale or giant scale integration (GSI) of today.

As an important historical coincidence, the Sputnik shock provided the perfect scenario for a 10× push.

10×: The computer on a silicon wafer (1966)

This enormous endeavor was launched through the US Air Force Office by giving three contracts to Philco-Microelectronics, RCA, and Texas Instruments. The speaker for the companies was Richard Petritz [17], then with TI and later a co-founder of MOSTEK. This computer never flew, but these projects created design automation, floor planning for yield and testability, automated test and the necessary *discretionary wiring* for each wafer, the birth of *direct electron-beam writing on wafer*, the best technology for rapid prototyping then, today, and probably also in 2020 (see chap. 8).

Another important side-effect of this project was that MOS circuits, which were discredited as being slow and not suitable for computing as compared with bipolar

circuits, gained momentum because of their density, their topological regularity (making them so suitable for design automation), and their much lower power density and overall power. Through this project, the USA advanced toward putting processing and memory functions with higher complexity on a larger scale on a single chip, eventually leading to the first *microprocessor* [6].

When, in the internal Japanese competition for the electronic desktop calculator, Sharp was looking for a competent source for their custom-designed calculator chips, they finally gave a much publicized contract in 1969 to the Autonetics division of Rockwell, because at home in Japan this capability did not exist. This shame became the origin of a big national R&D program funded and coordinated by MITI, the Ministry for International Trade and Industry:

10×: The Joint Very-Large-Scale Integration (VLSI) Laboratory in Japan (1972–1980)

This initiative caught the rest of the world unprepared: the USA was deeply entrenched in Vietnam, and the collapse of the dollar caused a long global recession. The VLSI Lab in Japan had researchers from all the large semiconductor companies united under one roof. The result was the biggest boost, on a relative scale, in microelectronic history for new, large-scale equipment and manufacturing. Memories were identified as lead devices, and by 1980 NEC had become the world's largest semiconductor company and the Japanese dominated the global memory market. Moore's law and the *scaling law* provided simple yardsticks.

The scaling law in its 1974 version became the persistent driving force for microelectronics. It takes us back to the fundamental structure of the FET, as shown schematically in Fig. 2.11.

It had been noted in 1962 (see the history of the 1974 milestone of *The Silicon Engine* [6]), that the FET was uniquely suited to miniaturization, by shrinking its lateral dimensions L and W . To first order, the transistor area would be halved if L and W shrank by a scaling factor of 0.7, providing a simple model for Moore's law. The scaling law declared that, in order to observe the maximum gate-field and source-drain-field limits, the gate and drain–source voltages would have to be

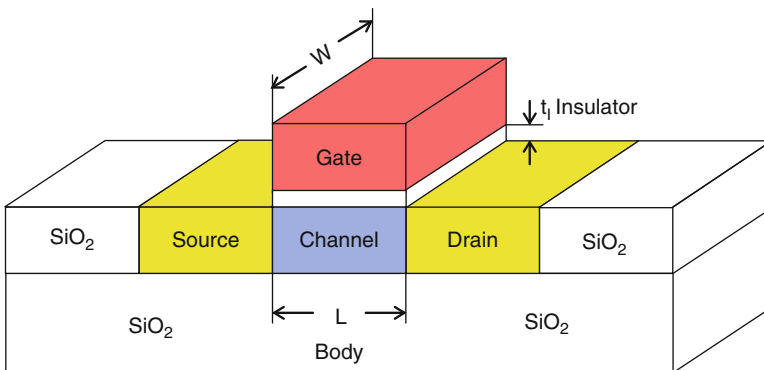


Fig. 2.11 Cross section of the MOS transistor annotated for scaling

reduced by the same amount, as would the insulator thickness t_1 , in order to maintain the aspect ratio of the transistor.

The zero-order model for the drain current in the MOS transistor:

$$\text{current} = \frac{\text{channel charge}}{\text{transit time}}, \quad (2.1)$$

$$\text{channel charge} = \text{capacitance} \cdot \text{voltage} \propto \frac{W \cdot L}{t_1} \cdot \text{field} \cdot L, \quad (2.2)$$

$$\text{transit time} \propto \text{field} \cdot L, \quad (2.3)$$

assumes the simple result in the constant-field case

$$\text{current} \propto \frac{W \cdot L}{t_1}, \quad (2.4)$$

suggesting a constant maximum current per micrometer for a given technology. Many in-depth publications followed, analyzing the potential and limits of scaling-down transistors, and we will discuss some in Sect. 3.2, but the one in 1974 by Dennard et al. [18] triggered a strategy that is very much alive today, more than 30 years later, as we shall see in the International Technology Roadmap for Semiconductors (ITRS) in Chap. 7.

Ever since the 1960s, we have seen an extreme rate of technology improvements marked by the fact that the leaders rapidly increased their R&D budgets disproportionately to well over 20% of their total revenue. We have selected eight technologies to analyze their history and future in Chap. 3. However, for the grand picture, which innovations stand out since 1959 and 1963? The scaling principle took it for granted that devices sat side-by-side in a 2D arrangement and progress would come from making them smaller. Obviously there was the potential of going into the third dimension.

2.5 1979: The Third Dimension

As we saw in Fig. 2.8, the complementary MOS transistor pair shares the input gate contact. This pair was the target of Gibbons and Lee at Stanford, and they produced the *CMOS hamburger* in 1979 [19]. In Fig. 2.12, we see the cross section of their PMOS/NMOS transistor pair.

This can be wired into the basic inverting amplifier of Fig. 2.8 on a footprint of close to one transistor, which would double the circuit density. Their invention is indicative of the process innovation at the end of the 1970s: The upper Si film, in which their NMOS transistor was formed, was deposited as a second polycrystalline-silicon film after the first film for the gate. That upper layer was then recrystallized with the energy from a laser. Recrystallizing Si deposited on oxide was one way to

Fig. 2.12 3D CMOS transistor pair after Gibbons and Lee (1980) [19] (© 1980 IEEE)

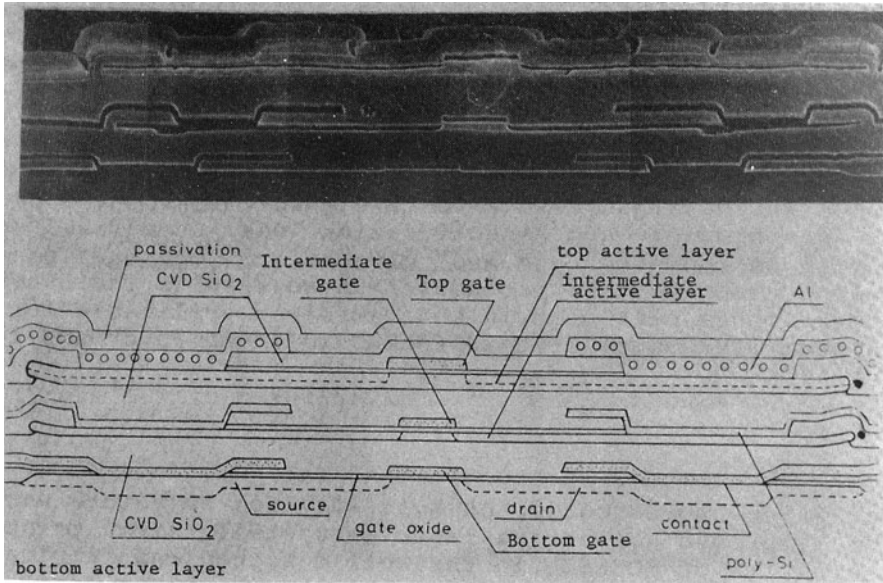
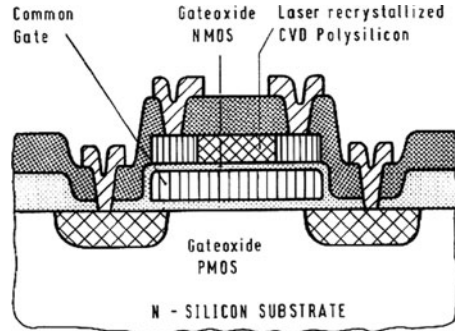


Fig. 2.13 Cross section with three transistor layers, after Kataoka (1986) [20] (© 1986 IEEE)

obtain multiple layers of silicon on top of each other separated by insulating layers of SiO_2 . High-energy implantation of oxygen and proper high-temperature annealing would produce buried layers of SiO_2 , thereby realizing a $\text{Si-SiO}_2\text{-Si}$ sequence of films potentially suitable for multiple device layers. In fact, after the heavily product-oriented VLSI Laboratory, the Japanese launched another, more fundamental effort:

10x: The program on Future Electron Devices, FED (1980)

Here we will only focus on the part *3D integration* and recapitulate some results, because it took until about 2005, roughly 25 years, for this strategic direction to take center-stage. In Fig. 2.13, we see the cross section of a Si chip surface with two crystallized Si films, allowing three transistor layers on top of each other [20]. 3D

integration became part of the research efforts that evolved to be the greatest global renaissance in microelectronics history.

10×: VHSIC, SRC in the USA, ESPRIT in Europe (1981)

At the end of the 1970s, the Vietnam decade and a long global recession, entrepreneurial and intellectual forces stepped forward – and were heard: Simon Ramo, Ray Stata, and Robert Noyce in the USA, François Mitterand and Jean-Jacques Servan-Schreiber in Europe were quoted everywhere in association with new, substantial initiatives. The strategic program on very-high-speed integrated circuits (VHSIC) made possible a set of very advanced, highest quality CMOS manufacturing lines, certified by the RADC (Rome Air Development Center in up-state New York) as qualified manufacturing lines (QMLs), which also provided foundry services to several totally new or refurbished university departments. Hardware, software, and tools for automated chip design were donated on a large scale to US universities. What started as a summer course, given by Carver Mead of Caltech and Lynn Conway of Xerox Palo Alto Research Center (Xerox PARC) at MIT, *VLSI Design*, became a bible for tens of thousands of students and industrial engineers [21]. The Defense Advanced Research Projects Agency (DARPA) launched the foundry service MOSIS (MOS Implementation Service) for universities under the leadership of Paul Losleben and a special university equipment program of hundreds of millions of dollars, incredible dimensions those days. The VHSIC Program produced the first standard hardware-description language VHDL (originally VHSIC Hardware Description Language), which is still the standard today. A uniquely successful university–industry research cooperative, the SRC (Semiconductor Research Cooperative), was founded in 1982 under the chairmanship of Robert Noyce, which quickly grew to a capacity of 500 Ph.D. students working on all aspects of microelectronics together with assignees from industry. A leader from industry, Eric Bloch from IBM, became president of the National Science Foundation (NSF). He created the university Engineering Research Centers (ERCs) on the premise that at least three faculties and more than 15 professors would work together with over 50 Ph.D. students in such a center. The program started with ten centers and now has over 50 in the USA.

The European Commission in Brussels, until then focused on coal, steel, and agriculture, prepared a technology program. As a historical note, as founders and leaders of two of the existing university pilot lines for large-scale integrated circuits in Europe, in Leuven, Belgium, and Dortmund, Germany, Roger Van Overstraeten and I were invited in 1979 to advise Giulio Grata, the responsible person in Brussels, on the elements of a European Microelectronics Research Program, which was launched eventually as ESPRIT (European Strategic Programme for Research in Information Technology).

The 1980s were marked by numerous *megaprojects*, multi-hundred million dollar investments per new facility, many with public money, one typical target being the first 1 Mbit DRAM memory chip. The communist German Democratic Republic went bankrupt on its megaproject, because they had to pay outrageous

amounts of money to acquire US equipment and computers through dark channels in the communist bloc. New regions and states competed for microelectronics industry investments, and today's map of chip manufacturing centers was pretty much established then. That period was also the origin of the global-foundry business model exemplified by TSMC (Taiwan Semiconductor Manufacturing Co.) established by Morris Chang after a distinguished career at Texas Instruments (TI).

One special result of the VHSIC program was high-speed electron-beam direct-write-on-wafer lithography for rapid prototyping and small-volume custom integrated circuits. The variable-shape beam, vector-scan systems Perkin-Elmer AEBLE 150, used by ES2 (European Silicon Structures) in France, and Hitachi HL700, used in Japan, Taiwan, and in my institute in Germany, established electron-beam lithography as a viable technology to help the scaling-driven industry to realize one new technology generation after the other. Line widths of 400 nm or less were considered the limit of optical lithography so that sizeable centers for X-ray lithography were set up at large synchrotrons in the USA at Brookhaven and in Germany at BESSY in Berlin and COSY in Karlsruhe. Wafer shuttles were planned to transport wafers there in volume, plans which never materialized because optical lithography with deep-UV laser sources was able to do the job.

With a capability of millions of transistors on a single chip and supported by a foundry service for up-to-date prototyping, the research community was encouraged to take on big questions, such as information processing not with the von Neumann architecture of processor, program memory, and data memory, but closer to nature, also named biomorphic.

2.6 1989: Neural Networks on Chips

The integration of large numbers of electronic functions on a single microchip was exploited from the early 1960s for array-type tasks, one particular direction being imager arrays made up of optical sensor elements and circuits for reading out the signals from the picture elements (pixels) [14]. The Japanese FED Program produced a 3D technology in which a Si photodiode layer was fabricated on top of two CMOS layers [22], where the CMOS layers would be used for reading and processing the pixel signals (Fig. 2.14).

This can be viewed as an early embodiment of a layered Si retina. A specific retina was then proposed by Mead and Mahowald (Fig. 2.15) in Mead's book *Analog VLSI and Neural Systems* [23]. This book again became a bible and the opener for worldwide activity on building neural networks on Si microchips. Figure 2.15 shows an analog implementation with resistors as synapses. Even in such a rigid setup, powerful functions such as the detection of shapes could be performed. Many other researchers chose general, programmable digital two-layer perceptrons. An example is shown in Fig. 2.16 [24], which can help to explain why these networks have significant potential for intelligent information processing.

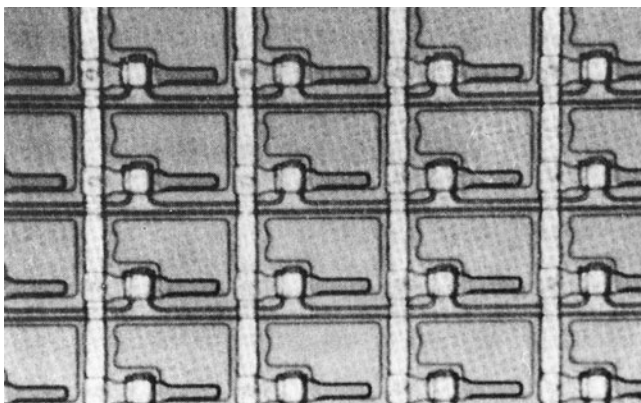


Fig. 2.14 Micrograph of an array of phototransistors with vias to the underlying CMOS layers for read-out electronics (1986) [22] (© 1986 IEEE)

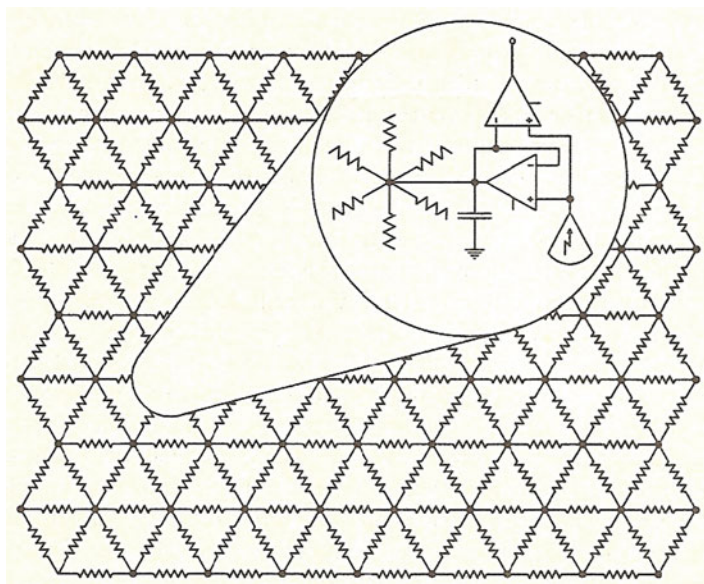


Fig. 2.15 The analog neural network of Mead and Mahowald for vision tasks such as shape detection consisting of active *pixels* and resistors as synapses (1989) [23] (© Springer 1989)

For the neural controller for automatic steering of a vehicle, driving data obtained on a 2 km section of a normal road with a human driver in the car were enough to train the neurocontroller. This is just one example of the prolific research that was launched through the investments in the 1980s.

The 1990s were marked by the worldwide re-engineering of large companies, enhanced by structural consequences after the end of the Cold War. This could have

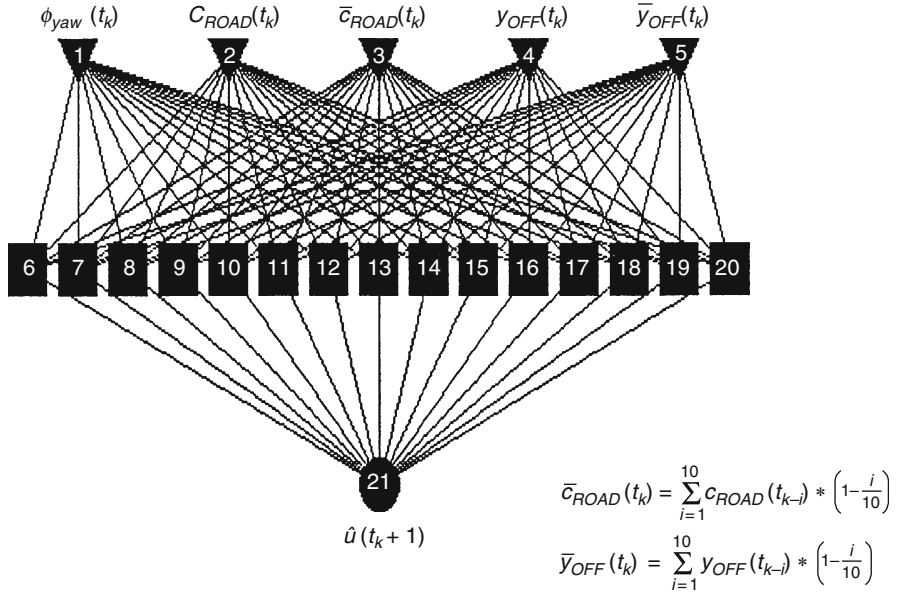


Fig. 2.16 Trainable neural controller with 21 neurons for automatic steering of a vehicle (1993) [24] (© 1993 IEEE)

hit the semiconductor industry much worse if there had not been two large-scale effects offsetting these problems:

- The force of Moore's law and of the scaling law.
- The tremendous push of the Tigers Korea, Taiwan, and Singapore, with just about 80 million people taking on the world of chip manufacturing, and dominating it today.

The effect of both of these factors has been the refocusing of the traditional players and an unparalleled strategic and global alliance of all major players.

10×: International SEMATECH and the Roadmap (1995)

Originally, SEMATECH started in 1988 as a cooperation of the US semiconductor and equipment industry, including government funding, in order to strengthen the US position in the field. In Europe, the JESSI (Joint European Submicron Silicon) project was initiated under the umbrella of EUREKA, the less bureaucratic agency that manages European cooperative strategy, while the funding runs on a per-country basis, which means that national governments fund their constituency. The focus on manufacturing, which relies much on global resources and partnering, soon let non-US companies join so that International SEMATECH was formed in 1995, first concentrating on developing the capability to manufacture on the basis of newly available 300 mm wafers. The International Technology Roadmap for Semiconductors (ITRS) was started with worldwide

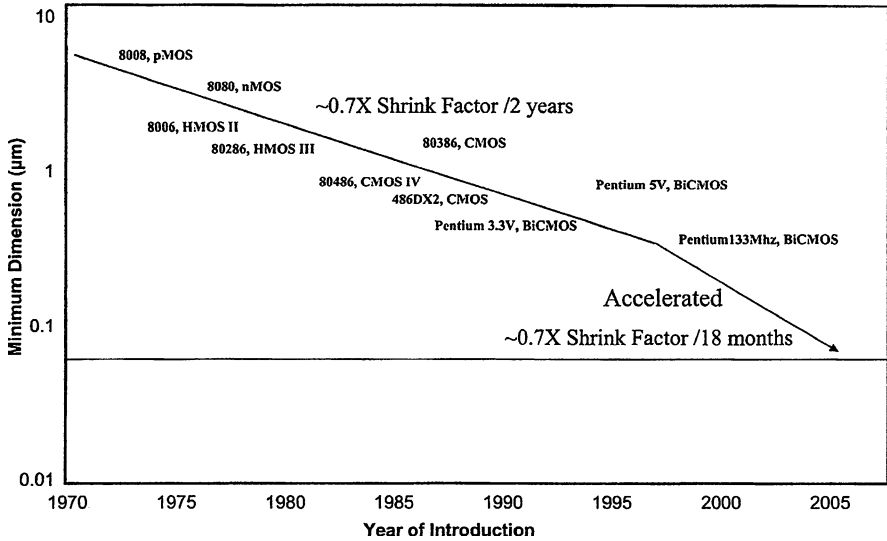
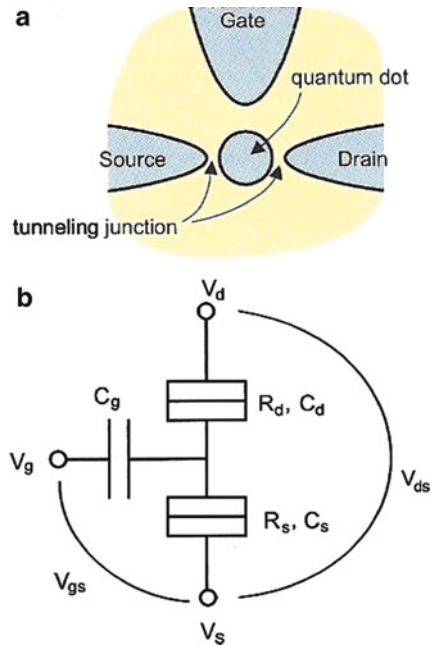


Fig. 2.17 The ITRS, shown here for the minimum dimension: 1999 issue

expert committees, and it became the single most powerful document defining the global strategy of the semiconductor industry. SEMATECH has expanded its charter since 2000 by operating large cooperative R&D facilities in Austin, TX, and Albany, NY, including major funding by the respective states. The Europeans concentrated their most significant efforts in Grenoble, France, and Leuven, Belgium, and IMEC (the Inter-University Microelectronics Center) in Leuven, founded by Roger Van Overstraeten in 1983, is today the world's largest independent R&D center for nanoelectronics. No other industry has developed such a joint global strategy and infrastructure to master the progress manifested by a cumulative R&D budget above 20% of revenue. It is interesting to overlay an early Roadmap for the minimum required lateral dimension on chips with more recent data (Fig. 2.17). It shows that the rate of progress on the nanometer scale has been pushed repeatedly beyond earlier targets [25].

Towards 100 nm, the limits of scaling down MOS transistors were seen as being near, triggering speculative research on that limit and how to go beyond it. A very short channel of well under 100 nm between the metallurgical source and drain junctions was producing a situation in which electrons would tunnel through these junctions and might get trapped in the channel island if the electrostatic “binding” energy exerted by the gate were large compared with the kinetic energy $k_B T$ at the environmental temperature T (k_B is the Boltzmann constant). Obviously, cooling would help to create this situation, as well as a very small channel island, where this electron would be locked up and only released for certain gate voltages and source–drain voltages. This tiny island containing a single electron (or none) was called a *quantum dot*, and it became very attractive as a memory element. A three-terminal *field-effect triode* (Fig. 2.18) consisting of a gate, a quantum dot and two

Fig. 2.18 Concept of a SET
(1987) (© Wiley-VCH)



tunnel junctions was proposed as a single-electron transistor (SET) in 1987 [26]. The first experimental SETs were reported a few years later, and, clearly, one electron per bit of information would be a major quantum step of progress.

For comparison, at the time, a memory transistor stored a ONE as 50,000 electrons on its floating gate, and a ZERO was about 5,000 electrons in a non-volatile memory. The programming pulse had to place these amounts of charge on the floating gate with narrow margins, and the read-out amplifier had to distinguish this difference. Research began on techniques to place three distinguishable amounts of charge on the floating gate, which would establish three threshold voltages and to refine the readout so that it could differentiate between these voltages [27]. Figure 2.19 shows measured threshold-voltage distributions. There was no overlap, and the means were about 1 V apart. Only 10,000 electrons were needed now to distinguish the bits [27]. This capability provided a true quantum jump in bit density. This storage capability development advanced with remarkable speed to about 250 electrons in 2010 as the length of a memory transistor decreased to 34 nm.

2.7 2000: The Age of Nanoelectronics Begins

The Y2K effect in microelectronics was that volume chip production reached the 100 nm lithography level, and compatible overall processing was achieved. Total capital expenditure for a factory based on 300 mm wafers exceeded US\$ 1 billion.

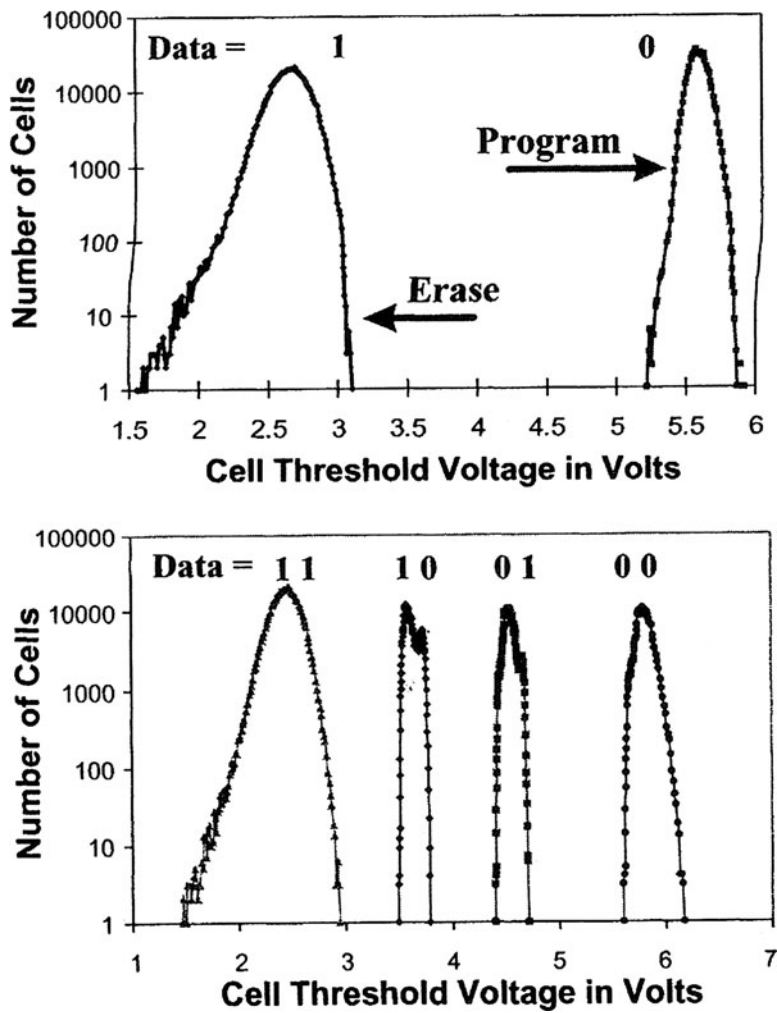


Fig. 2.19 Distribution of multiple threshold voltages achieved in 1995 [27] (© 1995 IEEE Press – Wiley)

10×: Next-Generation Lithography, NGL (2000)

The performance of optical lithography was seen to be at its limits, and this, in fact can be considered as another attribute of the new age of nanoelectronics, namely that, the generation of these sub-100 nm lateral structures and zillions of these on a wafer would require a lithography beyond short-wavelength refractive-optics-based patterning. The largest project in the history of International SEMATECH now became NGL (next-generation lithography), aimed at providing a non-optical alternative for anything smaller than 45 nm to be available for

prototyping in 2005. The contenders were ion-beam lithography (IBL) and extreme ultraviolet (EUV) lithography.

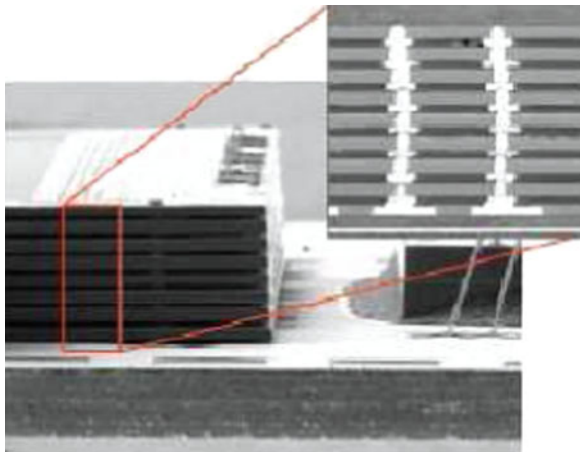
IBL was favored in Europe. A prototype was built in Vienna, Austria, based on a hydrogen-ion beam going through silicon stencil masks produced in Stuttgart, Germany. It was completed and demonstrated 45 nm capability in 2004. However, the international SEMATECH lithography experts group decided to support solely the EUV project, because IBL was assessed to not have enough throughput and less potential for down-scaling. For EUV lithography, a sufficiently effective and powerful source for 13 nm radiation, as well as the reflective optics and masks, had not become available by 2010, and EUV lithography has been rescheduled for introduction in 2012.

On the evolutionary path, optical lithography survived once more as it did against X-rays in the 1980s. At the level of deep UV (DUV), 193 nm, a major breakthrough was achieved by immersing the lens into a liquid on top of the silicon wafer. The liquid has a refractive index several times higher than air or vacuum significantly increasing the effective aperture. This immersion lithography saved the progress on the roadmap for one or two generations out to 22 nm. Thereafter, the future beyond optical is an exciting topic to be covered in Chap. 9.

It is natural that the scaling law cannot be applied linearly. Although lithography provided almost the factor 0.7 per generation, the transistor size could not follow for reasons of manufacturing tolerance, and its switching speed did not follow for physical reasons. Therefore, the industry as of 2000 had to embrace a more diversified and sophisticated strategy to produce miniature, low-power, high-performance chip-size products. It finally embraced the third dimension, 20 years after the FED Program in Japan, which we discussed in Sect. 2.6. Of course, this renaissance now occurred at much reduced dimensions, both lateral and vertical. While in the mid-1980s it was a touchy art to drill a hole (from now on called a *via*) through a 200 μm thick silicon/silicon dioxide substrate, wafers were now thinned to 50 μm or less, and 20 years of additional process development had produced a formidable repertory for filling and etching fine structures with high aspect ratios of height vs. diameter. Test structures on an extremely large scale emerged since 2006 with wafers stacked and fused on top of each other and thin W or Cu vias going through at high density to form highly parallel interconnects between the wafer planes. One early example is shown in Fig. 2.20 [28].

The mobilization in this direction of technology has been remarkable. Stacked wafers with through-silicon vias (TSVs) obviously provided a quantum jump in bit and transistor density per unit area. Enthusiastic announcements were made by industry leaders proclaiming new laws of progress beyond Moore. And it is true that, besides the gain in density, interconnect lengths are much reduced, partly solving the problem of exploding wiring lengths in 2D designs. The added manufacturing cost is accrued at dimensions that are more relaxed, and pre-testing each wafer plane before stacking provides more options for handling the testability of the ever more complex units. Stacking and fusing processor planes and memory planes offers a divide-and-conquer strategy for the diverging roadmaps for

Fig. 2.20 Cross sections through a 3D memory produced by stacking eight wafers with TSVs (2006) [28] (© 2006 IEEE)



processor and memory technologies and for the partitioning of the total system. We will come back to the significance of this type of 3D evolution in Chap. 3.

The first decade of the new millennium has seen further tremendous progress in CMOS technology in nanoelectronics, best described by the recent consensus that NMOS and PMOS transistors have reasonable and *classical or conventional* characteristics down to channel lengths of 5 nm, so that the worldwide design know-how and routines can be leveraged for new and improved products to an extent that is only limited by our creative and engineering capacity to develop these.

For the grand long-term picture, we address two recent achievements that have great potential or that indicate the direction in which we might perform speculative research on how to achieve entirely new levels of electronic functionalities.

2.8 2007: Graphene and the Memristor

The ultimately thin conducting film to which one could apply control, would have a thickness of one atomic layer. It would be a 2D crystal. It is not surprising that, in the context of widespread carbon research, this 2D crystal was eventually realized, observed, and characterized in carbon, where it is called graphene (Fig. 2.21). In 2007, Geim and Novoselov pulled off this single-atom-thick film of carbon from graphite with Scotch tape and transferred it to a SiO_2 layer on top of silicon [29]. The graphene layer fit on the oxide layer so well that the measurements confirmed theories on 2D carbon crystals going back to 1947, and high electron mobilities were observed. The already large carbon research community converged and expanded on graphene. A high-frequency transistor and an inverter with complementary transistors were reported soon after. The film deposition techniques for producing these graphene layers appear to be compatible with large-scale Si manufacturing so that graphene has high potential for future nanoelectronics.

Fig. 2.21 Real graphene single-atom layer with the C atoms clearly visible (From Wikipedia)

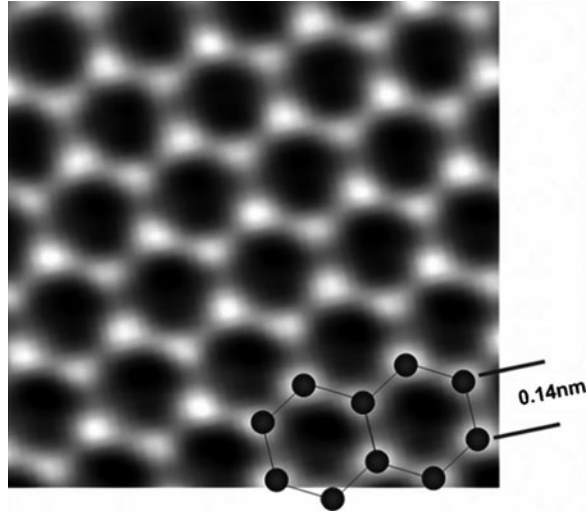
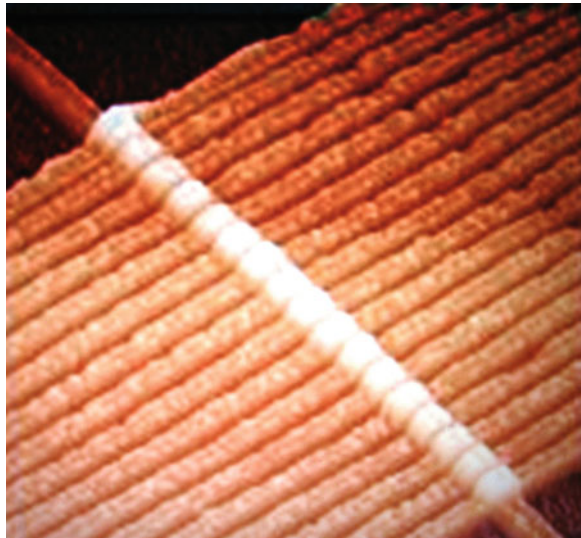


Fig. 2.22 Micrograph of 17 memristors (J.J. Yang, HP Labs)



Another recent achievement resulting from nanometer-scale electronics research is a new two-terminal device, which has an analog memory of its past with high endurance. It was reported in 2007 by Williams and members of his Laboratory for Information and Quantum Systems at Hewlett-Packard [30]. The device consists of two titanium dioxide layers connected to wires (Fig. 2.22). As the researchers characterized their devices, they arrived at a model that corresponded to the *memristor*, a two-terminal device postulated and named by Leon Chua in 1971 [31] on theoretical grounds. The memristor would complement the other three devices resistor, capacitor, and inductor.

A two-terminal device that could be programmed or taught on the go would be very powerful in systems with distributed memory. For example, the resistor synapses in Mead's retina (Fig. 2.15) could be replaced by these *intelligent resistors* to build very powerful neural networks for future silicon brains.

With graphene as a new material and the memristor as a new device, we conclude our grand overview of the technological arsenal that has been developed over more than 60 years and which forms the basis of our 2020 perspective in the following chapter.

References

1. von Lieben R.: Kathodenstrahlenrelais. German Patent No. 179807. Issued 4 March 1906
2. De Forest L.: Device for amplifying feeble electrical currents. US Patent No. 841387, filed 25 Oct 1906. Issued 15 Jan 1907
3. Lilienfeld J.E.: Method and apparatus for controlling electric currents. US Patent No. 1745175, filed 8 Oct 1926. Issued 18 Jan 1930
4. Hilsch R, Pohl R.W.: Steuerung von Elektronenströmen mit einem Dreielektrodenkristall und ein Modell einer Sperrschicht [Control of electron currents with a three-electrode crystal and a model of a blocking layer]. Z. Phys. **111**, 399 (1938)
5. Shockley W.: The path to the conception of the junction transistor. IEEE Trans. Electron Dev. **23**, 597 (1976)
6. See "The Silicon Engine" at www.computerhistory.org/semiconductor/. Accessed Feb 2011
7. Atalla M.M.: Stabilisation of silicon surfaces by thermally grown oxides. Bell Syst. Tech. J. **38**, 749 (1959)
8. Kahng D.: Electric field controlled semiconductor device. US Patent No. 3102230, filed 31 May 1960. Issued 27 Aug 1963
9. Hoerni J.A.: Method of manufacturing semiconductor devices. US Patent No. 3025589, filed 1 May 1959. Issued 20 March 1962
10. Noyce R.N.: Semiconductor device-and-lead structure. US Patent No. 2981877, filed 30 July 1959. Issued 25 April 1961
11. Saxena A.N.: Invention of Integrated Circuits – Untold Important Facts. World Scientific, Singapore (2009)
12. Wanlass F.M.: Low stand-by power complementary field effect circuitry. US Patent No. 3356858, filed 18 June 1963. Issued 5 Dec 1967
13. Wanlass F.M, Sah C.T.: Nanowatt logic using field-effect metal-oxide semiconductor triodes. IEEE ISSCC (International Solid-State Circuits Conference) 1963, Dig. Tech. Papers, pp. 32–33
14. Wallmark J.T, Johnson H (eds.): Field-Effect Transistors. Prentice-Hall, Englewood Cliffs (1966)
15. Moore G.: Cramming more components onto integrated circuits. Electron Mag. **38**(8), 114–117 (1965)
16. Moore G.: Progress in digital integrated electronics. IEEE IEDM (International Electron Devices Meeting) 1975, Tech. Dig., pp. 11–13
17. Petritz R.L.: Current status of large-scale integration technology. In: Proceedings of AFIPS Fall Joint Computer Conference, Vyssotsky, Nov 1967, pp. 65–85
18. Dennard R.H, Gaensslen F.H, Yu H.N, Rideout V.L, Bassous E, LeBlanc A.R.: Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J. Solid-State Circuits **9**, 256 (1974)
19. Gibbons J.F, Lee K.F.: One-gate-wide CMOS inverter on laser-recrystallised polysilicon. IEEE Electron Dev. Lett. **1**, 117 (1980)

20. Kataoka S.: Three-dimensional integrated sensors. IEEE IEDM (International Electron Devices Meeting) 1986, Dig. Tech. Papers, pp. 361–364
21. Mead C, Conway L.: Introduction to VLSI Systems. Addison-Wesley, Reading (1979)
22. Senda K, et al.: Smear-less SOI image sensor. IEEE IEDM (International Electron Devices Meeting) 1986, Dig. Tech. Papers, pp. 369–372
23. Mead C, Ismail M.: Analog VLSI Implementation of Neural Systems, ISBN 978-0-7923-9040-4, Springer (1989).
24. Neusser S, Nijhuis J, Spaanenburg L, Hoefflinger B.: Neurocontrol for lateral vehicle guidance. IEEE Micro. **13**(1), 57 (1993)
25. www.ITRS.net/. Accessed Feb 2011
26. Likharev K.K.: IEEE Trans. Magn. **23**, 1142 (1987)
27. Bauer M, et al.: A multilevel-cell 32 Mb flash memory. IEEE ISSCC (International Solid-State Circuits Conference), Dig. Tech. Papers, 1995, pp. 132–133
28. Lee K et al.: Conference on 3D Architectures for Semiconductor Integration and Packaging, San Francisco, Oct–Nov 2006
29. Geim A.K, Novoselov K.S.: The rise of graphene. Nat. Mater. **6**, 183 (2007)
30. Wang Q, Shang D.S, Wu Z.H, Chen L.D, Li X.M.: “Positive” and “negative” electric-pulse-induced reversible resistance switching effect in $\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$ films. Appl. Phys. A **86**, 357 (2007)
31. Chua L.O.: Memristor – the missing circuit element. IEEE Trans. Circuit Theory **18**, 507 (1971)

Chips 2020

A Guide to the Future of Nanoelectronics

Hoefflinger, B. (Ed.)

2012, XXVIII, 477 p., Hardcover

ISBN: 978-3-642-22399-0