Fabrizio Luccio

The transistor turns 75
April 2023

The first transistor patented in 1948
The transistor contributed to changing human activities as very few other inventions in history. Its implementation in integrated circuits has marked a spectacular reduction in size and energy consumption over time, with a consequent increase in the number of transistors per chip and a huge variety of its applications. We follow the transistor along its seventy-five years of life up to today’s state of the art, indicate expectations for the near future, and discuss the physical limits that will be imposed on them.
Main references:

Journals and conference proceedings, in particular:

IEEE Transactions on Components and Packaging
Communications of the ACM
IEEE Spectrum

Reports from manufacturers and evaluation agencies

IRDS (commission sponsored by IEEE)

C. Miller. Chip war: the fight for the world’s most critical technology. 2022 (politics)

Wikipedia (images)
2018
a year of innovation

ASML NXE:3300B
Conducting materials (copper, aluminum, tungsten), semiconducting materials (silicon, germanium, gallium arsenide), and insulating materials (silicon bioxide) will come into play.
$10^{22}$ atoms/cm$^3$

Van der Waals radius
210 pm = 0.21 nm
n-silicon doped with phosphorous

impurity \(10^{13}\) atoms/cm\(^3\)
doping \(10^{15}\) atoms/cm\(^3\)

p-silicon doped with boron

The phosphorus atom donates its fifth valence electron. It acts as a free charge carrier.

The free place on the boron atom is filled with an electron. Therefore a new hole ("defect electron") is generated. This holes move in the opposite direction to the electrons.
Two adjoining portions of n-silicon and p-silicon form a junction, a key electronic structure in all electronic devices.
junction pn
electron and hole diffusion
region depleted of charge carriers

In the *depletion region* an electric field $E$ is formed.
Forward voltage overcomes the field $E$ and causes current to flow.

Inverse voltage adds to the field $E$ and does not cause current to flow.
This is how a semiconductor diode is born. The diode has been a fundamental component of the first computing circuits, but is insufficient to build all the possible Boolean functions . . . .

. . . hence to implement any possible algorithm by means of a circuit
Transistor: the basic component

Invented: 1947
Patented: 1948

Brattain, Bardeen, Shockley
Nobel Prize: 1956
all transistors were born in the Bell Labs

Point-contact transistor, Bardeen and Brittain 1947
Bipolar-junction transistor (BJT), Shokley 1948
MOSFET (NMOS) transistor, Kahng and Atalla 1959
offshoot to Lilienfeld’s FET design 1925
MOSFET stands for Metal Oxide Semiconductor Field Effect Transistor

NMOS transistor
INVERTER
NAND and NOR are universal operators, i.e. any circuit implementing an arbitrary Boolean function can be implemented using NANDs or NORs only.
A flip-flop is made with two coupled inverters

RAM cell
DRAM
In 1963 Frank Wanlass and Chi-Tang Sah invented the CMOS (Complementary MOS) technology: a substantial energy saving is obtained by eliminating resistors. CMOS has become the standard since 1968.

The CMOS transistor actually consists of a series of one PMOS (called pull-up) and one NMOS (called pull-down). Only minor parasite heat dissipations occur.
CMOS

Vdd
A
Vss
Q
p-n-p
n-p-n
From PLANAR to FINFET

In 2000 Chenming Hu proposed a new MOS structure called FINFET growing in **three dimensions**. FINFET came into use in 2011 and is now the standard.
FGMOS used in flash memories

Below the control gate, an insulated floating gate is separated from the channel by an extra-thin insulator layer.
Writing is based on transit of electrons between base and floating gate by hot injection.

Reading is based on the electric field generated by the electrons possibly present in the floating gate. This field contrasts the gate tension, blocking the current between source and drain.

Cancelation by tunnel effect.
**Integrated circuits (IC’s)**

Independently proposed in 1959 by Jack Kilby of Texas Instruments and Robert Noyce of Fairchild:

_A new miniaturized electronic circuit built on a semiconducting body where all components are completely integrated._

statement in Kilby’s patent application

After a legal debate the two companies made an agreement and a massive production started in 1966.
1970
The first microprocessor

INTEL 4004
2300 transistors
10μm minimal feature

Federico Faggin, Ted Hoff Jr., Stanley Mazor
**ASIC** (Application Specific Integrated Circuit)

Standard Cells are pre-characterized cells implemented in ASICs

**Gate Array** (fixed structure)

Three functions of eight variables obtained through overlapping connections

the component * is useless
PLA Programmable Logic Array
FPGA (Field Programmable Gate Array)

extension of the PLA model, internally divided in different connected modules.
## Standard Cell ASIC Vs Gate Array Vs FPGA

<table>
<thead>
<tr>
<th></th>
<th>Standard Cell ASIC</th>
<th>Gate Array</th>
<th>FPGA</th>
</tr>
</thead>
</table>
| **Positives**    | • Highest performance: ~1 Billion transistors at multi-GHz rates. Often only way to meet a spec.  
• Lowest high volume cost (→ $ per die) | • Fairly Low design, CAD and up-front costs  
• Time from design ready to first part 1-2 weeks  
• Lowest mid-volume price | • Low design, CAD and up-front costs  
• Time from design ready to first part almost zero |
| **Negatives**    | • High design, CAD and wafer costs  
• Long time to first-product to market (long design time + >4 weeks for fab) | • Performance not much more than FPGA | • Low performance (Millions of implemented logic gates @ 10s to 100s of MHz)  
• High unit cost ($1,000’s) |
| **Comments**     | • Often used for FPGA shrink | | • Especially useful in markets that change fast or have low volumes |
1965  Gordon Moore

Fairchchild,
then founder of Intel

the circuits of computers, mobile phones, and other control systems will be regulated by a yearly doubling in the number of components that can be economically packed in an integrated circuit

Economics was at the core of Moore’s 1965 paper.
For any particular generation of manufacturing technology, there is a cost curve. The cost of making a component declines the more you pack onto an integrated circuit, but past a certain point, yields decline and costs rise. The sweet spot, where the cost per component is at a minimum, moves to more and more complex integrated circuits over time.
Chip fabrication starts with a silicon ingot

Diameter up to 30 cm

Mono-crystalline
Wafer thickness 0.5 mm

Poly-crystalline
Mainly used for MOS gate electrodes
Lithography

1. Grow field oxide

2. Etch oxide for pMOSFET

3. Diffuse n-well

4. Etch oxide for nMOSFET

5. Grow gate oxide

6. Deposit polysilicon

7. Etch polysilicon and oxide

8. Implant sources and drains

9. Grow nitride

10. Etch nitride

11. Deposit metal

12. Etch metal
Masks

individual chip

wafer
Wavelengths

visible light: down to 390 nm
ultraviolet (UV): 380 to 200 nm
extreme ultraviolet (EUV): 200 to 10 nm
X rays (soft): 10 to 0.1 nm
X rays (hard): below 0.1 nm
In principle with a radiation of wavelength $\lambda$ we cannot “draw” details smaller than $\lambda$. The problem is overcome by using multiple patterning, i.e. making shifts of the masks of a few nanometers to draw each area.

Up to 2018 the minimum wavelength used was 193 nm, allowing a minimal feature down to $\sim 30$ nm.

Then the whole picture changed using EUV radiation at 13.5 nm.
EUV scanner ASML NXE:3300B  > 100 M$  1.5 MWatt

The first 13.5 nm machine, 2018
Pulses of laser light are sent into a vessel where they collide with tin droplets

Assembling the scanner
Tin droplets are flattened by one laser pulse and then converted to EUV-emitting plasma by a second pulse.

Wavelength 13.5 nm
Generating EUV and focusing through a mirror (Zeiss) made of 40 pairs of alternating silicon and molybdenum layers
Array of FINFET
light 193 nm       EUV 23.5 nm
The problem of connections:

“Wiring is the major bottleneck for semiconductors”
Common opinion in all conferences on IC

Material used:
from Aluminum (originally) to

Copper has the lowest resistivity but it is prone to electromigration (atom dislodge by hitting accelerated electrons) for wires starting at 14 nm
Thin copper interconnection broken by electromigration
To protect thin copper connections from electromigration, they are lined with a 1-2 nm cobalt or graphene border.
Several tens of layers containing metallic connections may be piled up.

Wires in adjoining layers are connected through vias, i.e. holes in the insulator filled with copper or tungsten.

Over ten kilometers of wiring are contained today in a chip area of 1 cm$^2$. 
Parameters controlling functioning

The main electric function in transistor commutation is loading and unloading the capacitance $C$ arising between gate and substrate, through the connection resistance $R$. Power dissipation occurs through resistances, mainly in the source-drain path. Many parasitic capacitances and resistances also exist.

\[ v_g(t) = V_g(1-e^{-t/RC}) \]

where $RC$ is the circuit *time constant*

\[ w = \Sigma Ri^2 \quad e = wt \]

The smaller the circuit parameters, the higher the speed and the smaller the energy consumption

\[ i = C \frac{dv}{dt} \]

The deeper the voltage front, the higher the current and the source power
Number of transistors per chip

2022    Apple M1 Ultra
114 billion transistors in an area of ~ 860 mm$^2$
TSMC 5nm technology

Designed in Cupertino, CA, USA
Fabricated in Hsinchu, Taiwan
Assembled in various locations in China
TSMC (Taiwan Semiconductor Manufacturing Company) already produces microprocessors with 3nm technology and will get into 2nm by the end of this year.

Intel, Samsung, IBM, and others are going the same way.

Since the “diameter” of a silicon atom is ~ 0.4 nm, the chip minimal feature regards a few atoms.

Moore’s law will face limits very soon. Progress will be partly attained through system advances. As far as IC design, major attention is now directed to logic in memory and three-dimensional structures.
Energy and speed of computing vs data transfer

**Adding** two 32 bit numbers may require:

- energy: 20 fJ (20 femto Joule = $2 \times 10^{-14}$ J)
- time: 150 ps (150 pico seconds = $1.5 \times 10^{-10}$ s)

**Moving** two di 32 bit words by 1 mm may require:

- energy: 2 pJ (2 pico Joule = $2 \times 10^{-12}$ J)
- time: 15 ns (15 nano seconds = $1.5 \times 10^{-8}$ s)

AMD Zen3
Produced by TSMC
3D chips

Much shorter connections on the average reduce delays
3D chips: Samsung is testing CMOS construction putting pull-up and pull-down transistors one on top of the other
R&D expenses

Major investors are Intel, Samsung, and TSMC
The evolving foundry market
According to the marketing agency IC Insight about $2 \cdot 10^{21}$ transistors have been produced in 2022: about 250 billions for each inhabitant of the world.