Where does the power go?

The physics

Karl-Felix Müller
Georg-August University Göttingen

1 Transistors: general properties

1.1 A brief overview

TRANSISTORS are semiconductor devices used to build circuits on a microscopic level. In fact, transistors are the smallest building units in modern electronics. Their principle of function is to guide the amplitude of electric current flowing though them, allowing to increase or reduce the charge flow. The number of transistors assembled in a single device is a measure for its computing power. The largest transistor count conveys a notion of the order of magnitude these numbers are ranging in. Currently, it is over 30 billion transistors, contained in the Altera Stratix 10 [1].
There are several transistor types. Some of them are not in use anymore, but for understanding the general principle of transistors, it is very useful to start with older models. A transistor can usually be categorized in the following distinction:

- bipolar junction transistor (BJT)
- field-effect transistor (FET)
- junction field effect transistor (JFET)
- metal-oxide-semiconductor field effect transistor (MOSFET)

and rather representing a technology,

- the complementary metal-oxide-semiconductor (field-effect transistor) (CMOS)

1.2 Understanding transistors

A transistor consists of a semiconductor and uses its electronic properties. The atoms of a semiconductor are electrically neutral, unless doped with an element that has one additional electron (n-doped) or one electron less (p-doped) than the semiconductor. When these differently doped materials are brought closely together, as is illustrated in figure 1, a so called *depletion region* emerges, due to the following effect: the deficit of negative charges (electron holes) on one hand's side an the excess of electrons on the other hand's side create a diffusion of negative charges towards the deficit area, and vice versa. This charge diffusion results in an electric field, which counteracts its cause, the diffusion. These two effects are compensating each other in the depletion region, which is therefore electrically neutral, and cannot be passed by any charges.

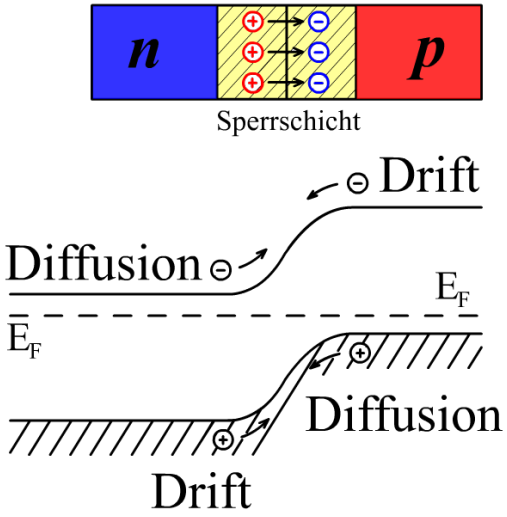

Figure 1: Band diagram of a p-n junction

In a BJT, differently doped areas are assembled the way figure 2 depicts. There are two n-doped electrode, called *emitter* and *collector*, and a very thin, just slightly p-doped electrode dividing them, which is called *basis*. If a small current (I_B) is deployed from emitter to basis, the depletion area between these two electrodes is narrowed down, making it possible for faster electrons to pass into the basis. For the basis is very thin and only slightly doped, the majority (approx. 99%) of these electrons is not recombining with the electron holes in the basis, but is able to overcome the whole basis and transmit into the collector. Once there, yet another small current (I_C) forces the collected electrons to continue into the circuit. Thus, two

small currents I_B and I_C are capable of triggering a large current, and are therefore working as a current amplifier. In case of an n-doped base and p-doped emitter and collector, the transistor would be conducting without a base-emitter current, and would be blocked when applying I_B.

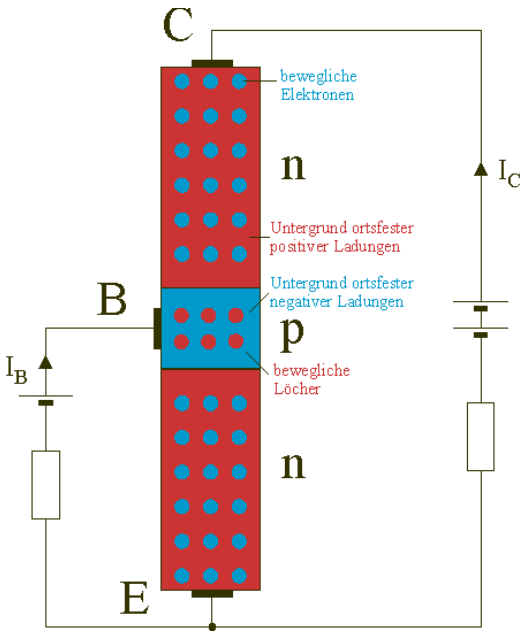

Figure 2: Scheme of a BJT [3]

The function principle of a JFET is similar, but is basd on voltage instead of current: the semiconductor substrate (*n-* or *p-channel*) is embraced by a circular oppositely doped metal oxide electrode (*gate*, in this case i.e. a MOSFET). If a voltage U_D is deployed between the electrodes S (*source*) and D (*drain*), an n-channel JFET is conducting. A second voltage U_{GS} enlarges the depletion region between *gate* and channel. The size of this impassable area is influencing the magnitude of the current flowing through the gate. Thus, the current between source and drain is again controlled (decreased) by a very small voltage. Analoguely to the BJT, the complementary device (a p-channel JFET/MOSFET) works vice versa: it is blocked without a voltage U_{GS} and gets conducting when the latter is applied.

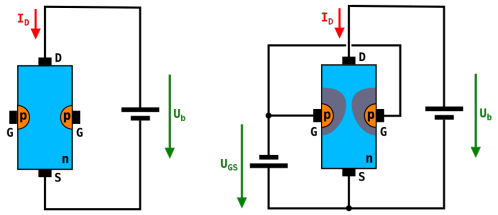

Figure 3: Scheme of a JFET [4]

2 CMOS components

With single nMOS and pMOS components, it is possible to build so called integrated circuits by assembling them in a way such that they can fulfill logical functions. CMOS therefore is rather a technology than a single building block. Circuits that use CMOS design have huge advantages compared to other circuits:

High noise immunity Electromagnetic noise, caused by any EM source including cosmic radiation, can theoretically cause transistors to alter their conducting properties, resulting in unwanted electronic operations. Since CMOS circuits usually use a voltage difference of approx. 5 V between «conducting» and «not conducting» , this huge gap can normally not briged by electromagnetic irritations, making CMOS circuits higly resistible against noise.

low static power consumption Due to the complementary design of CMOS components, one MOSFET is always switched of while the system is operating. This reduces power consumption, as the majority of power is now used for switching inputs of transistors (high voltage to no/low voltage and vice versa).

Low heat production As the amount of power consumed by CMOS components is small, so is the heat output, coming into existence by not perfectly conducting materials (resistance heat production). The switched-off components are not conducting, and therefore not producing heat.

3 Logical units

3.1 Circuit diagram

NOT gate Figure 4 depicts a CMOS inverter, also known as logical NOT gate. Vss is grounded,

while Vdd holds a voltage of about 5 V. The logical unit consists of two MOS components (pMOS at the top, marked by a small circle at the gate, and nMOS at the bottom), the input A and the output Q. Now if the input voltage at A is low, the current capable of flowing through the nMOS is limited. On the opposite, the pMOS gate is at low resistance, allowing current to flow from the top (higher voltage) to the output.
When on the other hand the input is high, the resistance states of the two MOS components are switched, establishing a conductive path from ground to output. This means that the input signal is inverted: low input results into high output, high input results into low output.

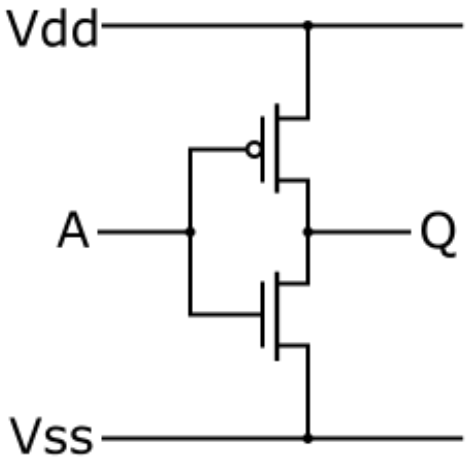

Figure 4: NOT gate in CMOS logic [5]

NAND gate Figure 5 illustrates a CMOS NAND gate, which has two inputs A and B. If both of the A and B inputs are high, then both the nMOS transistors (bottom half of the diagram) will conduct, neither of the pMOS transistors (top half) will conduct, and a conductive path will be established between the output and Vss (ground), bringing the output low. If both of the A and B inputs are low, then neither of the nMOS transistors will conduct, while both of the pMOS transistors will conduct, establishing a conductive path between the output and Vdd (voltage source), bringing the output high. If either of the A or B inputs is low, one of the nMOS transistors will not conduct, one of the pMOS transistors will, and a conductive path will be established between the output and Vdd (voltage source), bringing the output high. As the only configuration of the two inputs that results in a low output is when both are high, this circuit implements a NAND (NOT AND) logic gate.

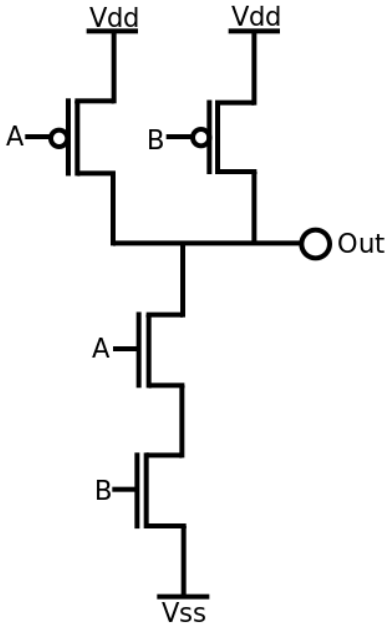

Figure 5: NAND gate in CMOS logic [6]

3.2 Physical layout

The schematic physical layout shown in figure 6 directly corresponds to the circuit diagram in figure 5. The key yields all important information. The smaller doped areas are attached to the main circuit due to practical stability issues.

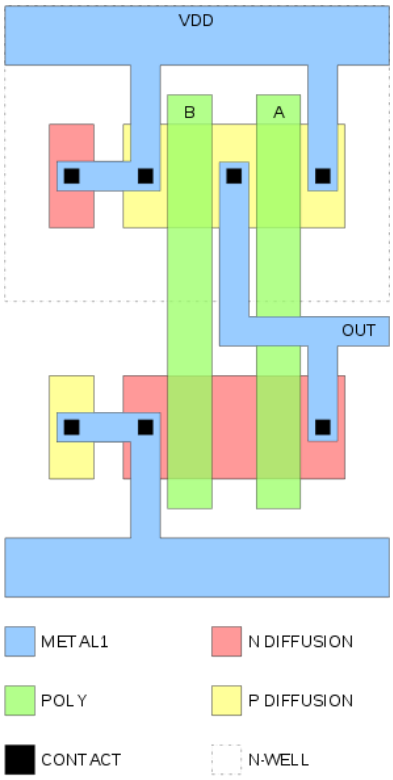

Figure 6: Schematic physical layout of a CMOS NAND gate [7]

4 Energy dissipation

Energy dissipation in transistors, responsible for energy losses and therefore additional cost, have several reasons, and can be categorized in static and dynamic dissipation, respectively.

4.1 Static dissipation

Subthreshold conducting Both nMOS and pMOS transistors have a *threshold voltage* U_{th} between gate and source. If the voltage applied is lower than U_{th}, the current flowing through the channel drops exponentially. However, the transistor is not entirely blocked, and the remaining *subthreshold current* leads to energy losses.

Tunneling current As transistors nowadays have microscopic orders of magnitude, quantum effects have to be taken into account: The tunneling probability of the electrons through the gate electrode drops exponentially with the latters' thickness, but is finite and measurable for electrodes with a thickness in the order of magnitude of an Ångström. The resulting tunneling current is another source of energy dissipation.

Leakage current If a MOSFET is operated in a way that reduces the extent of the depletion area, charge carriers are more likely to overcome this usually impassable region, resulting into a leakage current. Also, if the voltage creating a reverse-biased electrode gets too large, the equilibrium of the depletion area gets overruled and again, a leakage current evokes energy dissipation.

4.2 Dynamic dissipation

Charging and discharging of capacitances During operation, a CMOS circuit charges and discharges load capacitances, whenever the inputs are switched. This switching is determined by the clock cycle f, and accompanied by an *activity factor* α. The dynamic power dissipation can therefore be written as

$$P = \alpha C V f^2, \tag{1}$$

in which C is the particular capacitance, V is the supply voltage, and f is the clock frequency.

Short-circuit power dissipation Both nMOS and pMOS components are not instantaneously switching between conductive and blocking mode. Instead, they have finite switching times. During these peridos, a conductive path is established directly between voltage source and ground, creating a short-circuit. This also dissipates power that remains unused.

5 Mathematical models

Dynamic power models As said before, the most important formula to describe dynamic energy dissipation is

$$P = A \cdot C \cdot V^2 \cdot f. \tag{2}$$

Static losses In recent years, static losses grew more important, since the size of transistors decreased enormously, and quantum effects, especially energy losses due to quantum tunneling became dominant. The other huge power consumer is subthreshold leakage. It occurs when the used voltage is below the threshold voltage for switching the transistor. This leakage energy causes up to 20 - 40% of today's microprocessors. The subthreshold leakage current can be mathematically

modelled as ([Piguet])

$$I_{\text{leak}} = \mu_0 C_{\text{ox}} V_{\text{T}}^2 \frac{W}{L} e^{1.8 \cdot \frac{V_{\text{gs}} - V_{\text{th}}}{n V_{\text{T}}}} \left(1 - e^{-\frac{V_{\text{ds}}}{V_{\text{T}}}}\right). \quad (3)$$

In this formula, mu_0 is the zero bias mobility, C_{ox} is the gate oxide capacitance, $V_{\text{T}} = kT/e$ is the thermal voltage, W and L are the transistor's effective width and length, V_{gs} is the gate-source voltage, V_{th} is the threshold voltage, V_{ds} is the drain-source voltage and n is the subthreshold swing coefcient of the transistor. The crucial knowledge we gain from this model is that the gate-source voltage as well as the drain-source voltage should be as small as possible. High thermal voltage increases the leakage current in the first place, but at very high values, it eventually reduces it. Furthermore, transistor should be built in a compact way such that the quotient of effective width and length equals 1.

6 Architectural solutions

Dynamic dissipation reduction From equation (2), one can easily derive that the two objectives of having a quickly operating computer (high f) and little dynamic energy losses (low f) are interfering, an issue that is critical for further processor design. Decreasing the voltage would strongly reduce energy losses, but would make the processor more susceptible to mistakes caused by noise. Futher possibilities for reducing dynamic losses would be reducing the output capacitance C and reducing the switching activity A.

Static dissipation reduction For reducing dominant static dissipation, the opportunities are more multifarious:

- A dual threshold CMOS yields the advantage, that only in critical paths, a low threshold voltage is deployed. Therefore, the critical switching may happen without too large resistances, while the subthreshold leakage (which is more likely to appear in transistors with low threshold voltage) is prevented due to high threshold voltages in non-critical paths.
- Increasing V_{s} of NMOS transistors causes subthreshold leakage to decrease exponentially, due to negative gate-source voltage. This causality is illustrated in figure (ref).
- Stacking transistors also causes static losses to decrease, as figure (ref) indicates.

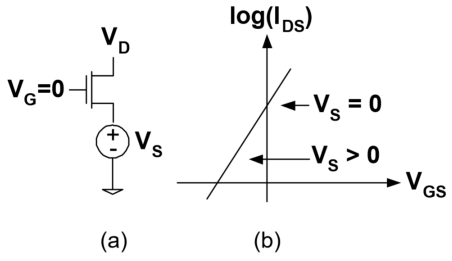

Figure 7: Leakage reduction due to increased V_{s} [Piguet]

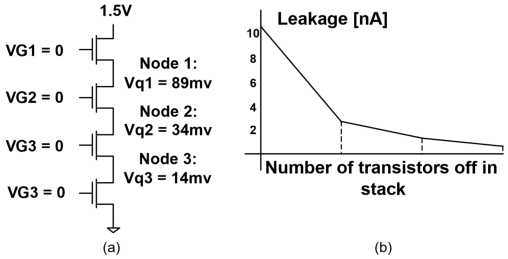

Figure 8: Leakage reduction due to stacked transistors [Piguet]

7 Resources

References

[Piguet] - Piguet, Christian: Low Power CMOS Circuits, Taylor & Francis Group 2006

[1] - `http://www.gazettabyte.com/home/2015/6/28/alteras-30-billion-transistor-fpga.html` - access on 10/18/2016, 7:20 p.m.

[2] - Degreen under Creative Commons Attribution-Share Alike 2.0 Germany license

[3] - `http://www.leifiphysik.de/elektronik/transistor` - access on 10/18/2016, 7:20 p.m.

[4] - Chtaube under Creative Commons Attribution-Share Alike 2.0 Germany license

[5] - Wikimedia Commons, public domain

[6] - JustinForce under Creative Commons Attribution-Share Alike 2.0 Germany license

[7] - Wikimedia Commons, public domain