Basic MOS Transistor

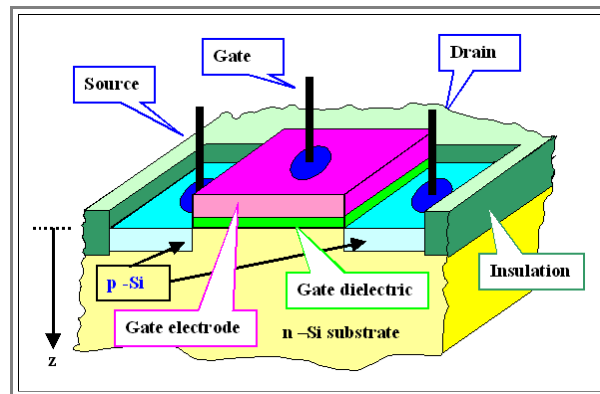### Qualitative Description

▸ The basic concept of a *MOS* **Transistor** transistor is simple and best understood by looking at its structure:

Basics

● It is always an *integrated* structure, there are practically no single individual **MOS** transistors.

▸ A **MOS** transistor is primarily a switch for digital devices. Ideally, it works as follows:

● If the voltage at the **gate electrode** is "*on*" , the transistor is "*on*", too, and current flow between the **source** and **drain** electrodes is possible (almost) without losses.

● If the voltage at the gate electrode is "*off*", the transistor is "*off*", too, and no current flows between the source and drain electrode.

▸ In reality, this only works for a given *polarity* of the gate voltage (in the picture above, e.g., only for negative gate voltages), and if the supply voltage (always called U_{DD}) is not *too small* (it used to be **5 V** in ancient times around **1985**; since then it is declining and will soon hit an ultimate limit around **1 V**).

● Moreover, a **MOS** transistor needs *very thin* **gate dielectrics** (around, or better below **10 nm**), and *extreme control* of materials and technologies if real **MOS** transistors are to behave as they are expected to in "ideal" theory.

▸ What is the working principle of an "ideal" **MOS** transistor?

● In order to understand it, we look at the behavior of carriers in the **Si** under the influence of an external electrical field under the gate region.

▸ Understanding **MOS** transistor *qualitatively* is easy. We look at the example from above and apply some source-drain voltage U_{SD} in either polarity, but *no gate voltage yet*. What we have under these conditions is

● A **n**-type **Si** substrate with a certain equilibrium density of electrons $n^e(U_G = 0)$, or $n^e(0)$ for short. Its value is entirely determined by doping (and the temperature, which we will neglect at the present, however) and is the same everywhere. We also have a much smaller concentration $n^h(0)$ of holes.

● Some **p**-doped regions with an equilibrium concentration of holes. The value of the hole concentration in the source and drain defined in this way also is determined by the doping, but the value is of no particular importance in this simple consideration.

● Two **pn**-junctions, one of which is polarized in forward direction (the one with the positive voltage pole), and the other one in reverse. This is true for any polarity; in particular one junction will *always* be **biased** in reverse. Therefore *no source-drain current I_{SD} will flow* (or only some small reverse current which we will neglect at present).

● There will also be no current in the forwardly biased diode, because the **n-S**i of the substrate in the figure is not electrically connected to anything (in reality, we might simply ground the positive U_{SD} pole and the substrate).

▸ In summary, for a gate voltage $U_G = 0$ **V**, there are no currents and everything is in equilibrium. But now apply a *negative voltage* at the gate and see what happens.

● The electrons in the substrate below the gate will be electrostatically repelled and driven into the substrate. Their concentration directly below the gate will go down, $n^e(U)$ will be a function of the depth coordinate z .
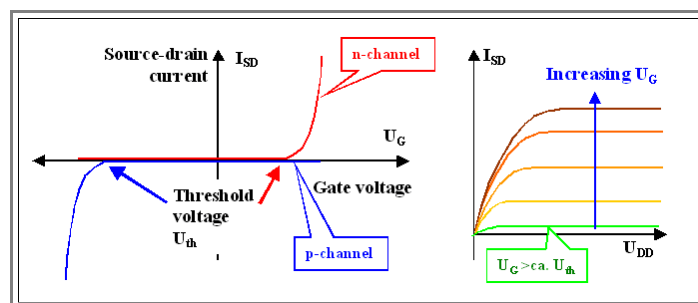
$$n^e \; = \; n^e(z) \; = \; f(n^e(0), \, U)$$

● Since we still have equilibrium, the mass action law for carriers holds anywhere in the Si, i.e. .

$$n^e(z) \cdot n^h(z) \; = \; n_i^2$$

- With n_i = intrinsic carrier density in Si = const.(U,z)
- This gives us

$$n^h(z) = \frac{n_i^2}{n^e(z)}$$

- In other words: If the electron concentration below the gate goes down, the hole concentration goes up.

▶ If we sufficiently decrease the electron concentration under the gate by cranking up the gate voltage, we will eventually achieve the condition $n^h(z=0) = n^e(z=0)$ right under the gate, i.e. at $z=0$
- If we increase the gate voltage even more, we will encounter the condition $n^h(z) > n^e(z)$ for small values of z, ie. for $z_c > z > 0$.

▶ In other words: Right under the gate we now have *more holes than electrons*; this is called a state of **inversion** for obvious reasons. **Si** having more holes than electrons is also called *p-type Si*. What we have now is a **p**-conducting **channel** (with width z_c) connecting the **p**-conducting source and drain.
- There are no more **pn**-junctions preventing current flow under the gate - current can flow freely; only limited by the ohmic resistance of contacts, source/drain and channel.

▶ Obviously, while cranking up the gate voltage *with the right polarity*, sooner or later we will encounter inversion and form a conducting channel between our terminals which becomes more prominent and thus better conducting with increasing gate voltage
- The resistivity of this channel will be determined by the amount of **Si** we have inverted; it will rapidly come down with the voltage as soon as the **threshold voltage** necessary for inversion is reached.

▶ If we reverse the voltage at the gate, we attract electrons and their concentration under the gate increases. This is called a state of **accumulation**. The **pn** junctions at source and drain stay intact, and no source - drain current will flow.
- Obviously, if we want to switch a **MOS** transistor "on" with a *positive* gate voltage, we must now reverse the doping and use a **p**-doped substrates and **n**-doped source/drain regions.
- The two basic types we call "**n-channel MOS**" and "**p-channel MOS**" according to the kind of doping in the channel upon inversion (or the source/drain contacts).
- Looking at the electrical characteristics, we expect curves like this:

- The dependence of the source-drain current I_{SD} on the gate voltage U_G is clear from what was described above, the dependence of I_{SD} on the source-drain voltage U_{SD} with U_G as parameter is maybe not obvious immediately, but if you think about it a minute. you just can't draw currents without some U_{SD} and curves as shown must be expected qualitatively.

▶ What can we say *quantitatively* about the working of an **MOS** transistor?
- What determines the threshold voltage U_{th}, or the precise shape of the $I_{SD}(U_{th})$ curves? Exactly how does the source - drain voltage U_{SD} influence the characteristics? How do the prime quantities depend on material and technology parameters, e.g. the thickness of the gate dielectric and its dielectric constant ϵ_r or the doping level of substrate and source/drain?
- Plenty of questions that are, as a rule, not easily answered. We may, however, go a few steps beyond the qualitative picture given above.

Some Quantitative Considerations

▶ The decisive part is achieving *inversion*. Lets see how that looks like in a band diagram. To make life easier, we make the gate electrode from the same kind of **n-Si** as the substrate, just highly doped so it is as metallic as possible - we have the same kind of band diagram then to the left and right of the gate dielectric
- Lets look schematically what that will give us for some basic cases:
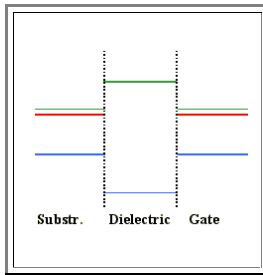
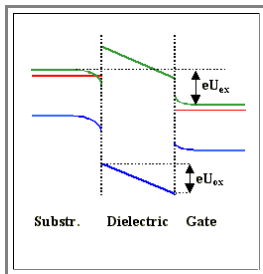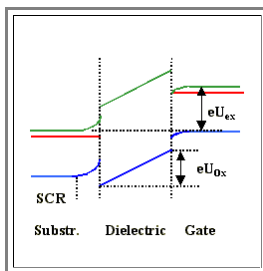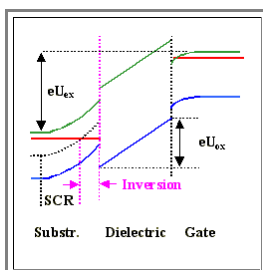| Voltage at the gate	Conditions in the Si	Voltage drop	Charge distribution
Zero gate voltage. "**Flat band**" condition	Nothing happens. The band in the substrate is perfectly flat (and so is the band in the contact electrode, but that is of no interest).	We only would have a voltage (or better potential) drop, if the Fermi energies of substrate and gate electrode were different	There are no *net* charges
Positive gate voltage. **Accumulation**	With a positive voltage at the gate we attract the electrons in the substrate. The bands must bend down somewhat, and we increase the number of electrons in the conduction band accordingly. (There is a bit of a space charge region (**SCR**) in the contact, but that is of no interest).	The voltage drops mostly in the oxide	There is some *positive charge* at the gate electrode interface (with our **Si** electrode from the **SCR**), and *negative charge* from the many electrons in the (thin) accumulation layer on the other side of the gate dielectric.
Small negative gate voltage. **Depletion**	With a (small) negative voltage at the gate, we repel the electrons in the substrate. Their concentration decreases, the hole concentration is still low - we have a layer depleted of mobile carriers and therefore a **SCR**.	The voltage drops mostly in the oxide, but also to some extent in the **SCR**.	There is some *negative charge* at the gate electrode interface (accumulated electrons with our **Si** electrode), and *positive charge* smeared out in the the (extended) **SCR** layer on the other side of the gate dielectric.
Large negagive gate voltage. **Inversion**	With a (large) negative voltage at the gate, we repel the electrons in the substrate very much. The bands bend so much, that the Fermi energy (red line) is in the lower half of the band close to the interface. In this region holes are the majority carriers, we gave *inversion*. We still have a **SCR**, too.	The voltage drops mostly in the oxide, but also to some extent in the **SCR** and the inversion layer.	There is more *negative charge* at the gate electrode interface (accumulated electrons with our **Si** electrode), some *positive charge* smeared out in the the (extended) **SCR** layer on the other side of the gate dielectric, and a lot of *positive charge* from the holes in thin inversion layer.

Qualitatively, this is clear. What happens if we replace the (highly **n**-doped) **Si** of the gate electrode with some metal (or **p**-doped **Si**)?

- Then we have *different Fermi energies* to the left and right of the contact, leading to a *built-in potential* as in a **pn**-junction. We will than have some band bending at zero external voltage, flat band conditions for a non-zero external voltage, and concomitant adjustments in the charges on both sides.

- But while this complicates the situation, as do unavoidable fixed immobile charges in the dielectric or in the **Si**-dielectric interface, nothing new is added.

Now, the decisive part is achieving inversion. It is clear that this needs some minimum threshold voltage U_{th}, and from the pictures above, it is also clear that this request translates into a request for some *minimum charge* on the capacitor formed by the gate electrode, the dielectric and the **Si** substrate.

- What determines the amount of charge we have in this system? Well, since the whole assembly for any distribution of the charge can always be treated as a simple capacitor C_G, we have for the charge of this capacitor.

$$Q_G = C_G \cdot U_G$$

- Since we want U_{th} to be small, we want a *large gate capacitance* for a large charge Q_G, and now we must ask: What determines C_G?

If all charges would be concentrated right at the interfaces, the capacitance *per area unit* would be given simply by the geometry of the resultant plate capacitor to

$$C_G = \frac{\epsilon \epsilon_0}{d_{Ox}}$$

- With $d_{Ox} =$ thickness of the gate dielectric, (so far) always silicon dioxide SiO_2.

Since our charges are somewhat spread out in the substrate (we may neglect this in the gate electrode if we use metals or very highly doped **Si**), we must take this into account.

- In electrical terms, we simply have a second capacitor C_{Si} describing the effects of spread charges in the **Si**, switched in series to the geometric capacitor which we now call **oxide capacitance** C_{Ox}. It will be rather large for concentrated charges, i.e. for accumulation and inversion and small for depletion.

- The total capacitance C_G then is given by

$$\frac{1}{C_G} = \frac{1}{C_{Ox}} + \frac{1}{C_{Si}}$$

For inversion and accumulation, when the most of the charge is close to the interface, the total capacitance will be dominated by C_{Ox}. It is relatively large, because the thickness of the capacitor is small.

- In the depletion range, C_{Si} will be largest and the total capacitance reaches a minimum.

- In total, C_G as a function of the voltage, i.e. $C_G(U)$ runs from a constant value at large positive voltages through a minimum back to about the same constant value at large positive voltages. The resulting curve contains all relevant information about the system. Measuring $C_G(U)$ is thus the first thing you do when working with **MOS** contacts.

- While it is not extremely easy to calculate the capacitance values and everything else that goes with it, it can be done - just solve the Poisson equation for the problem.

All things considered, we want C_{Ox} to be *large*, and that means we want the dielectric to be *thin* and to have a *large* dielectric constant - as stated above without justification.

- We also want the dielectric to have a large breakdown field strength, no fixed charges in the volume, no interface charges, a very small **tg** δ; it also should be very stable, compatible with **Si** technology, and cheap.

- In other words, we wanted SiO_2 - even so its dielectric constant is just a mediocre **3.9** - for all those years of microelectronic wonders. But now (**2001**), we want something better with respect to dielectric constants. Much work is done, investigating, e.g., CeO_2, Gd_2O_3, ZrO_2, Y_2O_3, $BaTiO_3$, **BaO/SrO**, and so on. And nobody knows today (**2002**) which material will make the race!