## Lecture 2

The following topics will be covered in this lecture

Fabrication of the MOSFET

Working principle of the MOSFET

Channel charge as a function of gate voltage

## Production of a MOSFET

Fig 1 shows the 3D view of an N-channel Metal Oxide Semiconductor (MOS) Field Effect Transistor (FET) – shortly NMOS.

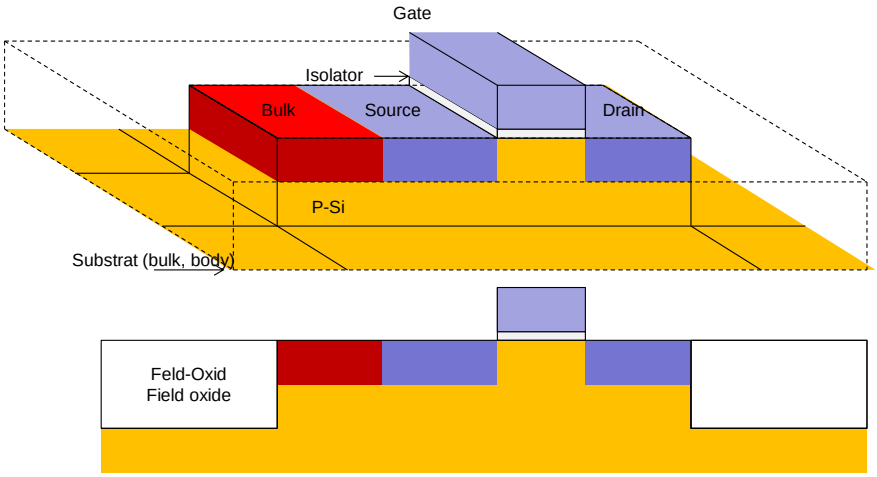



*Fig 1: 3D image of an NMOS Transistor*

A MOSFET is made of 4 electrodes: source, drain, gate and substrate (also called bulk). The source electrode is the source of the charge carriers (NMOS: electrons, PMOS: holes). The drain collects the charge carriers. The gate is the electrode that controls the transistor current. The source and drain are implanted in the substrate. In the case of an NMOS, the substrate is N-doped and the source, the drain and the gate are P-doped.

*In this lecture we consider the planar transistors. The latest semiconductor technologies use different transistor types, such as fin-FETs.*

The transistor is surrounded by the insulator-region that we call field oxide (made with $SiO_2$). The field oxide is used to isolate adjacent transistors. This is shown in Fig 2.

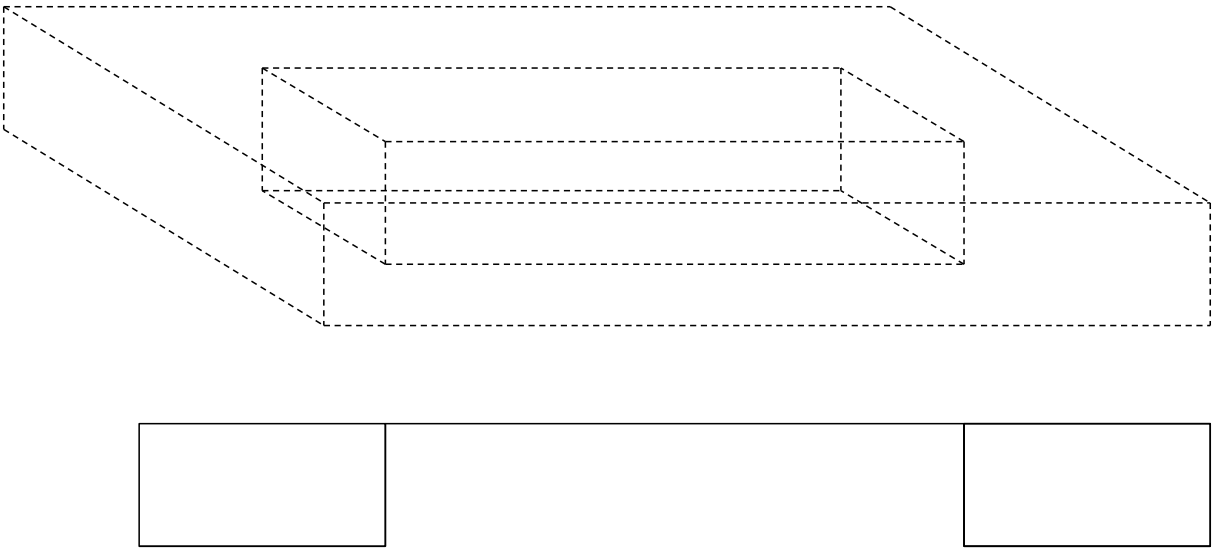


*Fig 2: Transistor is surrounded by field oxide*

The field oxide isolates the neighbouring transistors from each other. It prevents the formation of parasitic transistor structures between the regular transistors (Fig 3).

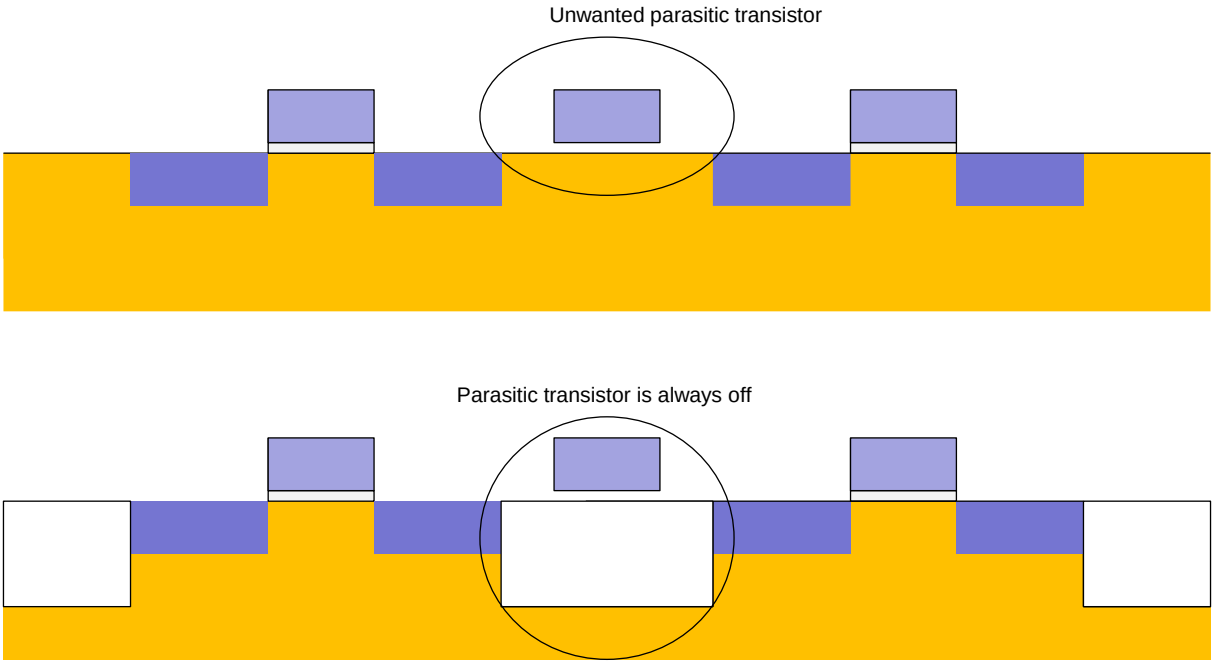



*Fig 3: Isolation of the transistors with field oxide*

The following figures describe the series of steps in the MOSFET fabrication.

In the first step, the local substrates (called wells) are generated. For NMOS we need a P-well (sometimes called P-tub) and for PMOS (P-channel MOSFET) an N-well.

A mask layer (PWELL) defines the p-well. The structures in the photomask (made of quartz glass and metal) are transferred using a photoresist to silicon by exposing the photoresist with UV light through the mask. Photoresist serves as a mask for the layers underneath (silicon dioxide and silicon nitride). Photoresist is used to protect the areas that should not be etched. The silicon dioxide and silicon nitride structures are used as masks for doping. Diffusion or ion implantation are used as doping techniques.

N-regions can be doped with phosphorus or arsenic. P-areas are doped with boron. Ion implantation works in the following way: The accelerated dopant ions penetrate into silicon substrate. They have relatively constant range, which depends on its kinetic energy. In this way, the dopant density has a defined maximum. During this process, radiation damage occurs in the crystal grid of the semiconductor. Therefore, the substrate must be annealed (cured) after an implantation step by warming it up to a high temperature.

https://de.wikipedia.org/wiki/Ionenimplantation

The next mask called diffusion-mask defines the active silicon area where the transistor will be placed. The field oxide is made outside this active region (Fig 4).

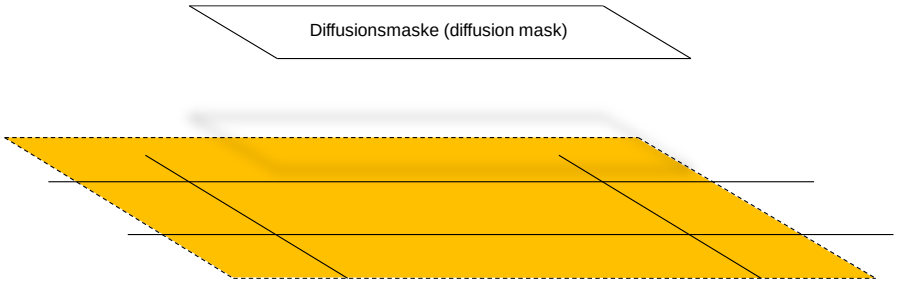



*Fig 4: Diffusion mask defines the active region of the substrate where the transistor will be placed*

Trenches for the field oxide are made by etching.

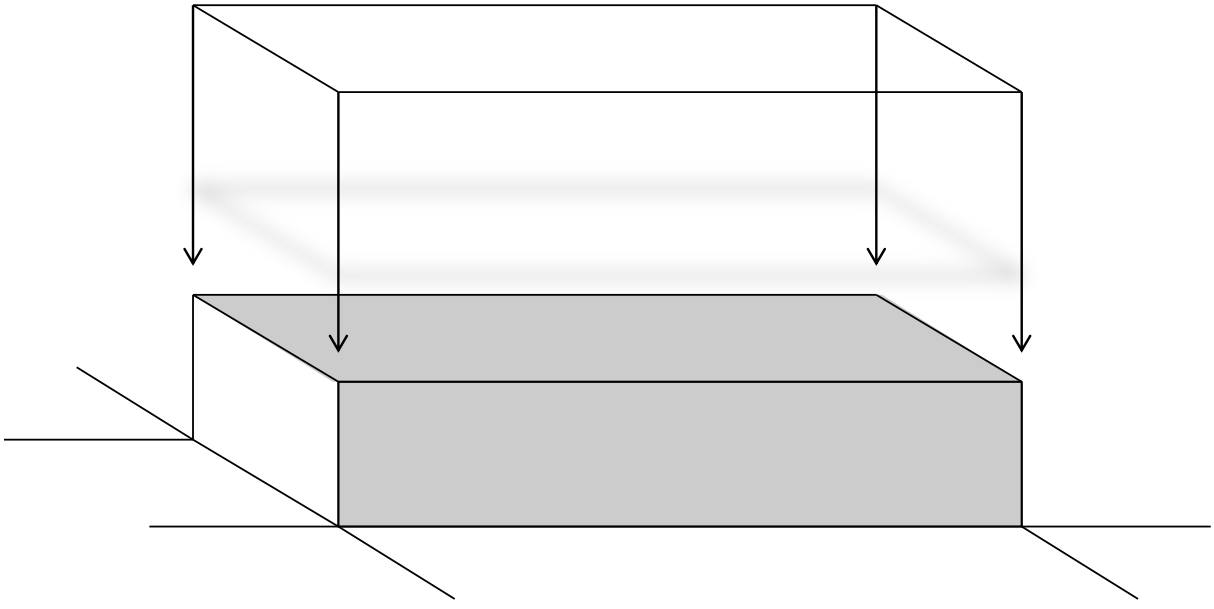


*Fig 5: Trench production*

Anisotropic etching process (e.g. reactive ion etching) is used. The trenches are filled with silicon dioxide. For this purpose, CVD procedure (chemical vapour deposition) is used. A more detailed description can be found here: https://de.wikipedia.org/wiki/Grabenisolation

The next step is production of the gate oxide. Thin gate oxide is grown by thermal oxidation. (We consider a semiconductor technology which uses silicon dioxide as gate oxide. The new technologies use other materials with higher relative permittivity such as $HfO_2$.) A thin oxidation layer is crucial for good electrical properties of transistors. The oxide capacitance is about 13 fF/µm$^2$. As we will see later, oxide capacitance determines the threshold voltage and the transconductance of the transistor. Typical thickness of gate oxide is 2.6 nm (in 65 nm technology). This corresponds only to about 5 atom layers, since the grid constant of $SiO_2$ is about 0.5 nm.

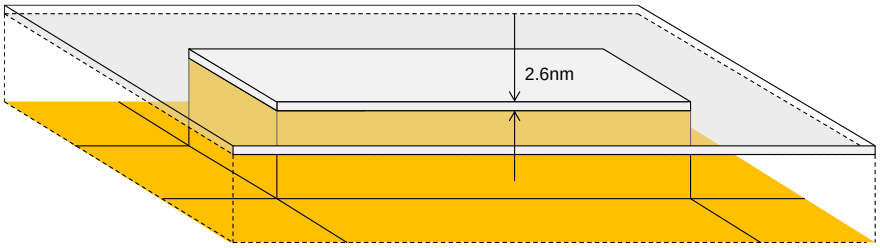



*Fig 6: Thermal oxidation*

In a further step, the gate electrode is made. Gate is made of polycrystalline silicon produced by LPCVD (low pressure chemical vapour deposition).

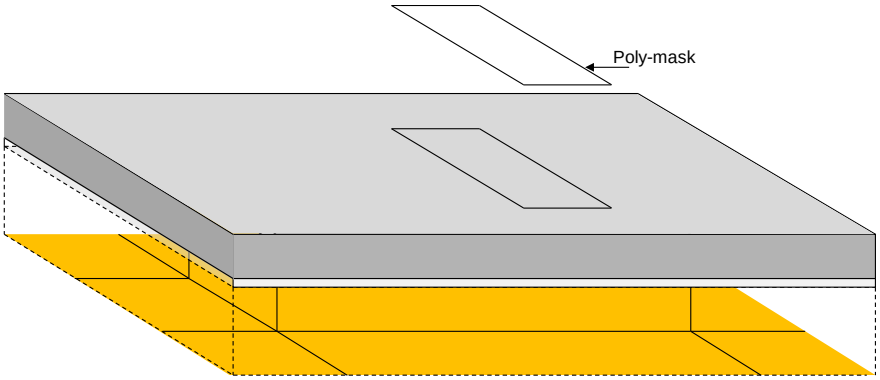



*Fig 7: Polysilicon ("poly") mask defines the gate electrode and its connection*

The position of the gate is defined by the overlap between the diffusion-mask and the polysilicon-mask. The thin oxide remains in the overlap region.

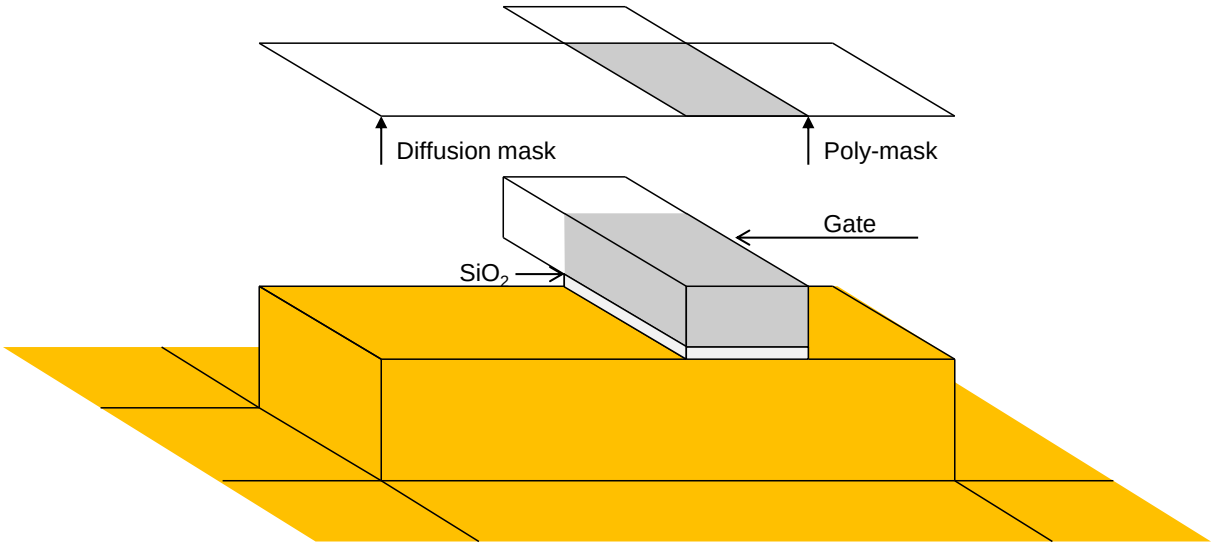



*Fig 8: The thin oxide remains in the overlap region of diffusion- and poly-mask*

The first FET transistors used metal gate electrode. The name MOSFET originates from Metal-Oxide-Semiconductor. Why is polysilicon used for the gate electrode and not metal?

There are three reasons for this:

1. Gate electrode from polysilicon can be doped, which leads to a lower threshold voltage, as will be explained later.

2. Polysilicon gate can be used as a mask for subsequent doping of source and drain. We say that this process step is self-aligned.

3. Polysilicon has a significantly higher melting temperature than aluminium, and the following process steps can be performed at higher temperatures. For instance, the doping of source and drain by diffusion and annealing.

After the production of the gate, the next step is doping of the source, the drain and the substrate contact. Two methods can be used: diffusion and ion implantation. The masks called: P/NPuls, the gate electrode and the field oxide define doped areas.

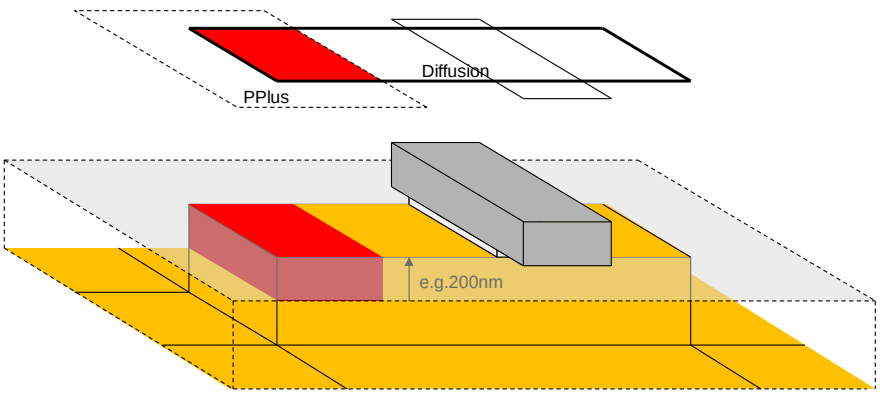



*Fig 9: PPlus mask. The overlap region between PPlus- and diffusion masks will be P-doped*

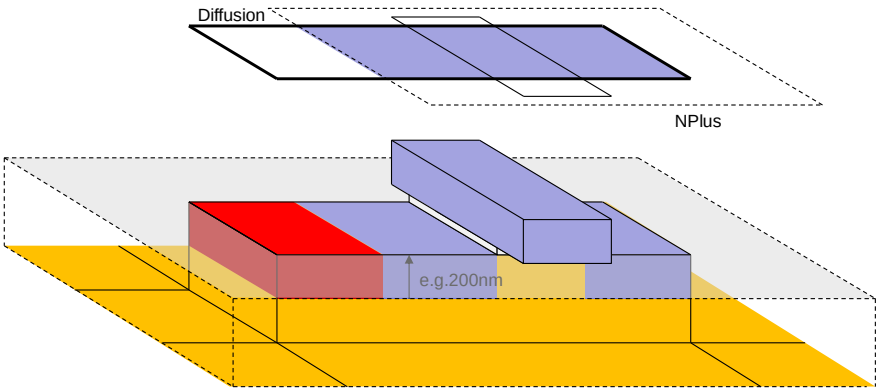



*Fig 10: NPlus mask. The overlap region between NPlus- and diffusion masks will be N-doped*

With this step, the production of the main transistor structures is accomplished.

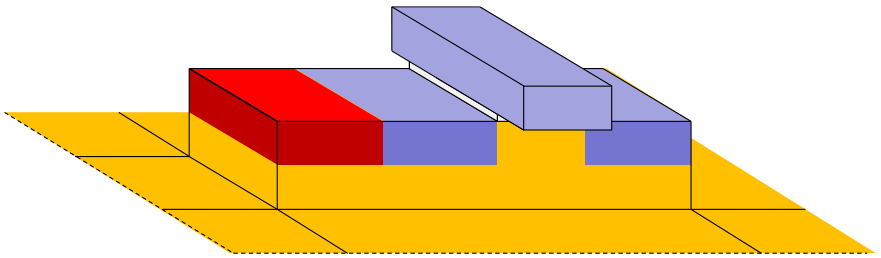


*Fig 11: Transistor with all structures*

## MOSFET

A MOSFET transistor is made up with four electrodes: source, drain, gate and substrate (bulk), as shown in Fig 12.

The source is the source for the free charge carriers (NMOS: electrons, PMOS: holes) and the drain collects them. The gate serves for control of the drain-source current. Source and drain are placed in the substrate. The substrate has its own contact, in the figure called bulk contact.

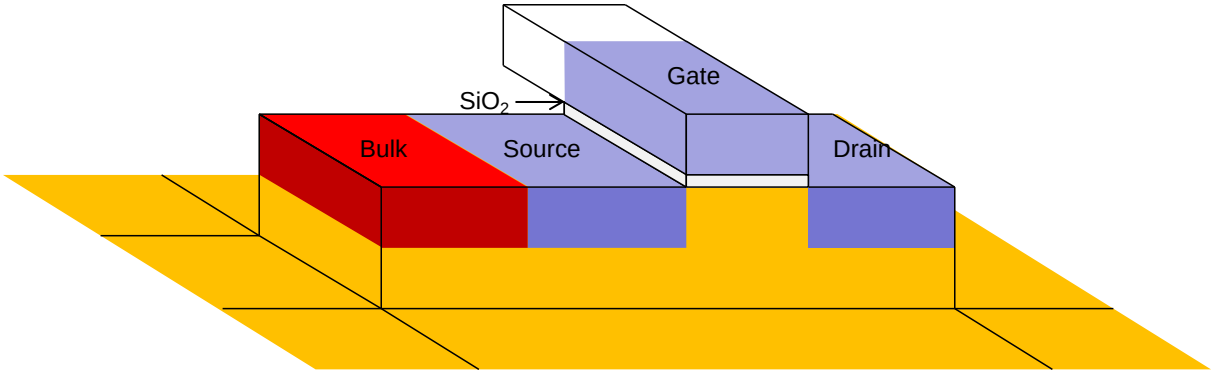



*Fig 12: NMOS*

A PMOS can be obtained by exchanging all the dopants (N->P, P->N). A PMOS is located in an N-type substrate. Since PMOS and NMOS are placed on the same wafer (silicon substrate) the transistors are usually located in the local substrates called "wells" or “tubs”.

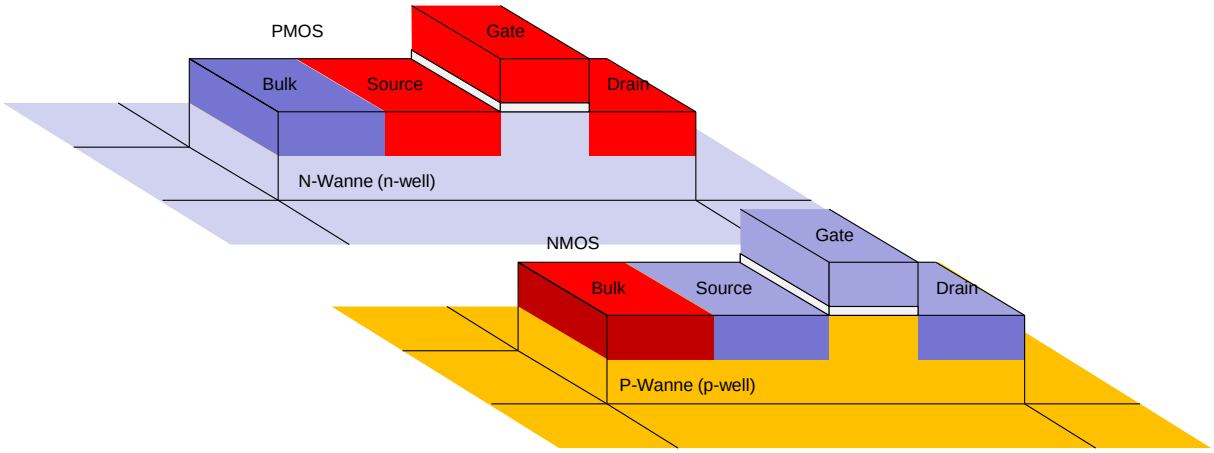



*Fig 13: NMOS und PMOS*

In this course we will use the switch-like transistor symbols, with or without substrate electrodes. These symbols are symmetrical, like the transistor structure itself.

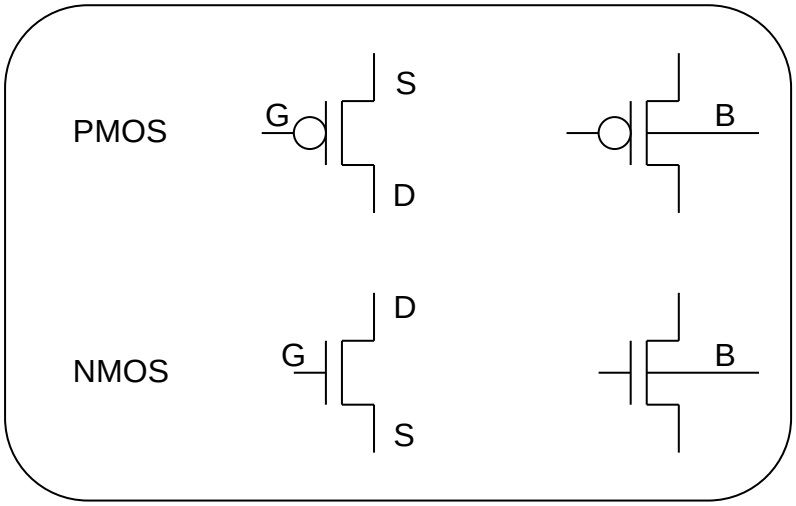



*Fig 14: Simplified symbols*

How do we recognize source and drain?

In the case of NMOS, the source is the electrode that has lower potential. In the case of PMOS, the source has higher potential.

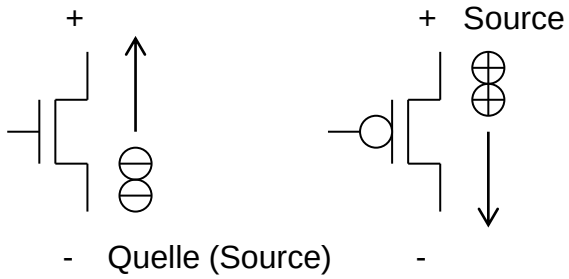



*Fig 15: Source*

Alternatively, we can use the asymmetric symbols with arrows. If the substrate electrode is missing in the symbol, it is connected either to the source or to a fixed voltage.

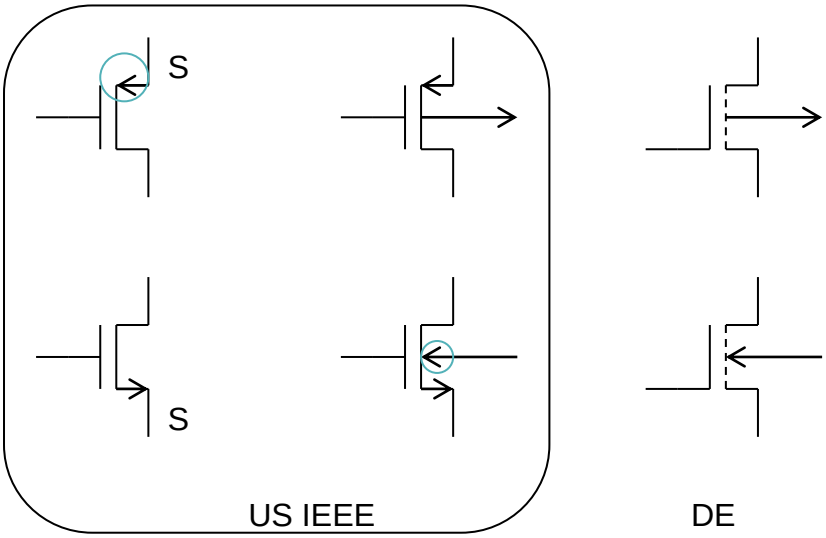



*Fig 16: 4-terminal symbols*

How to remember the substrate-arrow direction? The arrow shows the direction from p to N-region. (NMOS: from P-substrate to N-channel) It is similar as the symbol of a PN diode. Its shape shows the current direction when it is directly biased. The current flows then from P- to N-region.

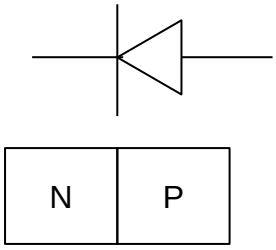



*Fig 17: Diode symbol*

**Operation of the MOSFET and derivation of the current equation**

The MOSFET functionality was explained in the lecture "Electronic circuits". Here we will summarize the most important facts and describe some special properties of small MOSFETs.

Let us consider an NMOS. The structure contains two PN diodes: source/substrate and drain/substrate (Fig 18). The substrate potential must be chosen in such a way that both diodes are reversely biased. Otherwise, a MOSFET does not work properly. Therefore, in the case of NMOS, the substrate must have a lower potential than the source and drain. In such a state, no current flows between the drain and the source if the gate source voltage is zero.

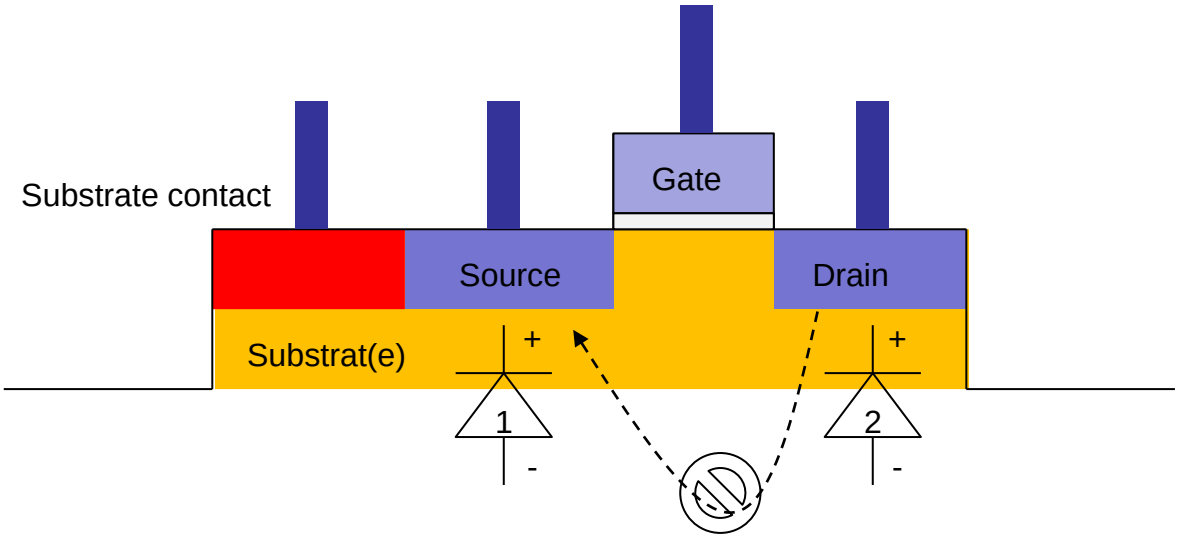



*Fig 18: The PN Diodes should be reversely biased*

**Contact voltages**

Silicon and metal form either ohmic contacts or Schottky diodes. A contact that conducts only in one direction wouldn't be suitable for the MOSFET structure.

Contacts between metal and *high-doped* silicon conduct current in both directions, even if they are Schottky diodes. Why? The potential barrier (Schottky barrier) is then very thin. The electrons tunnel through the barrier and current can also flow in the direction of the barrier. To make use of this effect, the silicon part of the substrate contact is highly doped (see Fig 18).

Contact voltages are generated between silicon and metal electrodes.

The contact voltages make the potentials of the silicon electrodes ($V_g$*, $V_s$*, $V_d$* and $V_{b*}$) different than the potentials on the metal electrodes $V_g$, $V_s$, $V_d$ and $V_b$ (see Fig 19).

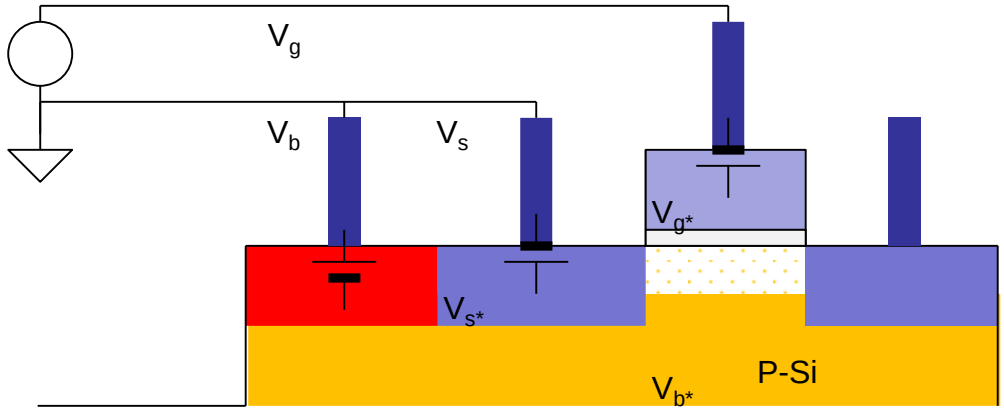
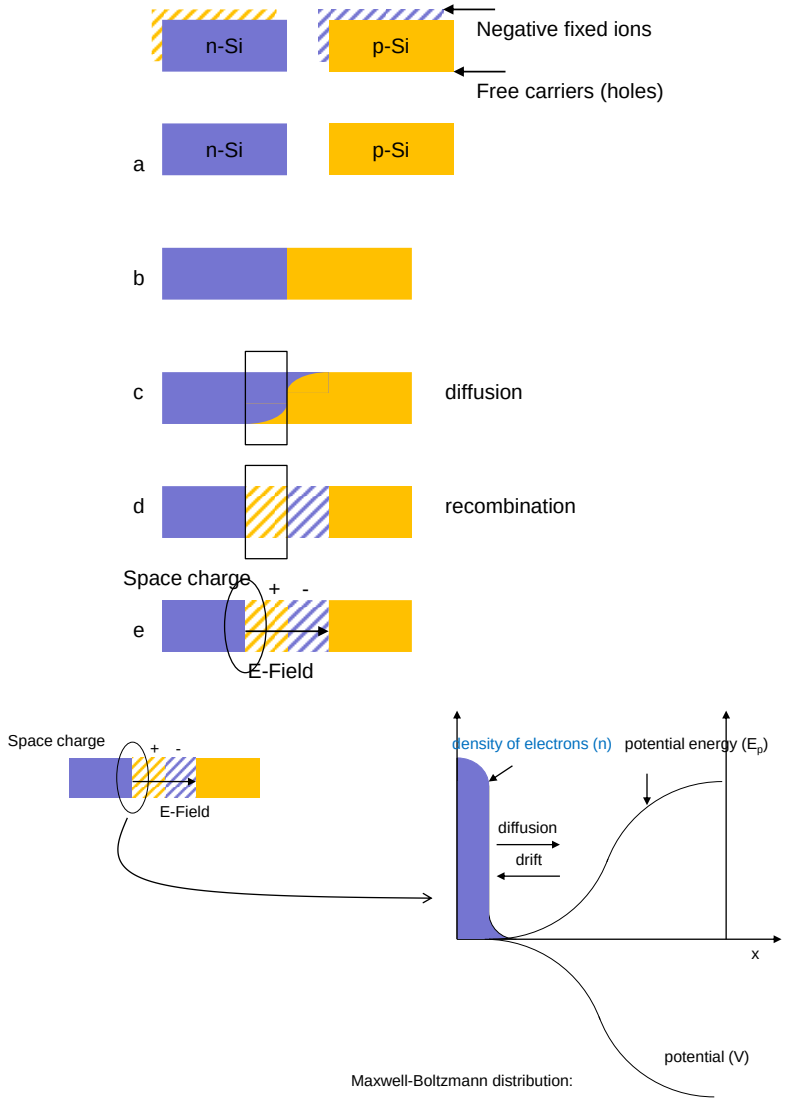



Fig 19: Contact voltages are induced between silicon- and metal regions

What is the origin of the contact voltage?




$$n(x) = n(0)e^{-E_p/kT}$$

$$n(x) = n(0)e^{eV/kT} = n(0)e^{V/U_T}; U_T = kT/e$$

*Fig 20: Origin of contact voltage*

Simplified explanation (Fig 20): Let us consider the electrons first. In N-silicon and metal, the density of electrons is higher than in P-silicon. When we "connect" N-silicon or metal with P-silicon (b), a diffusion current of electrons towards P-silicon starts (c). The electrons and holes recombine and the negative charge of acceptor ions is not anymore compensated by holes. A negatively charged space charge region is formed in P-silicon (d). From the same reasons, a positively charged region is formed in N-silicon. In this way, E-field and contact voltage are generated (e). E-field yields to a drift current that compensates for the diffusion. An equilibrium state is achieved. The density of electrons and holes is described by Maxwell-Boltzmann distribution. More precisely, electrons follow Fermi-Dirac distribution, but it can be approximated by Maxwell-Boltzmann formula within their energy bands.

The electron density in N-silicon $n_n$ is nearly equal to the donor density $N_d$. The electron density in P-silicon $n_p$ can be expressed with the following equation $n_p=n_i^2/N_a$. Intrinsic charge carrier

density is: $n_i = 10^{10}/cm^3$, density of silicon atoms is: $n_{si} = 5 \times 10^{22}/cm^3$. Using Maxwell Boltzmann formula, we can calculate the contact voltage as function of $n_n$ und $n_p$:

$$V = U_T \ln\left(\frac{n_n}{n_p}\right) = U_T \ln\left(\frac{N_a N_d}{n_i^2}\right); n_i \sim e^{\frac{-E_g}{kT}}$$

Since instrinsic denisty $n_i$ increases with temperature, the contact voltage decresases with temperature. We model the contact voltages with constant voltage sources.

**Tunnel effect contact**

The contact between metal and silicon is normally a Schottky diode. The current can flow only in one direction, when the external voltage lowers the potential barrier, as shown in Fig 20B.

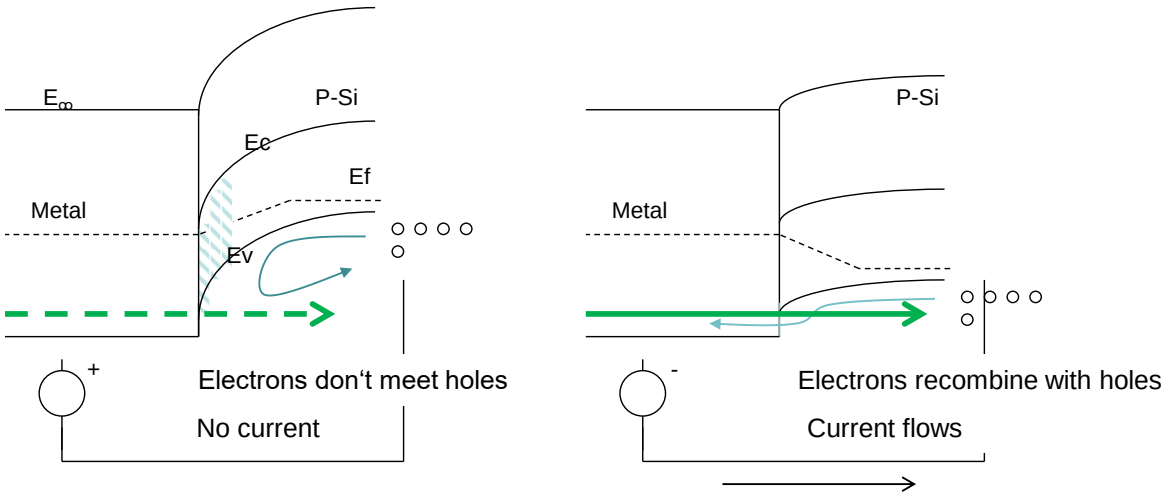



*Fig 20B: Schottky diode*

We use a trick to establish a normal contact in both directions. When silicon is highly doped, the potential barrier in silicon is very narrow and the charge carriers can tunnel (quantum mechanical tunnel effect) through the barrier in both directions. The contact between silicon and metal conducts then in both directions, it is an "ohmic contact" (Fig 20C).

The substrate contact of the MOSFET is implemented as tunnel contact. For this reason, the silicon end of the bulk contact is additionally p+ doped. Also, a tunnel contact has a contact voltage.

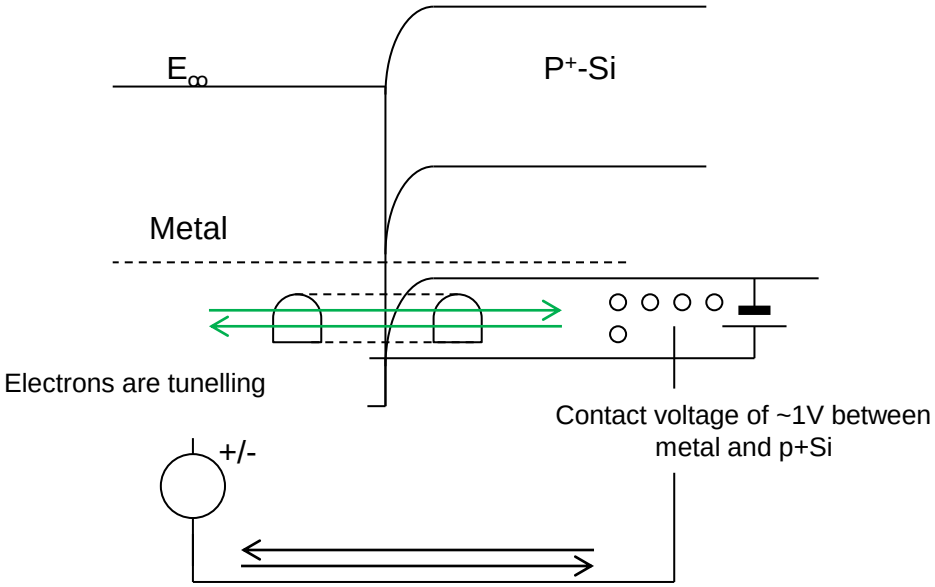



*Fig 20C: Tunnel contact*

The following figure summaries the most important formulas for understanding of semiconductors.

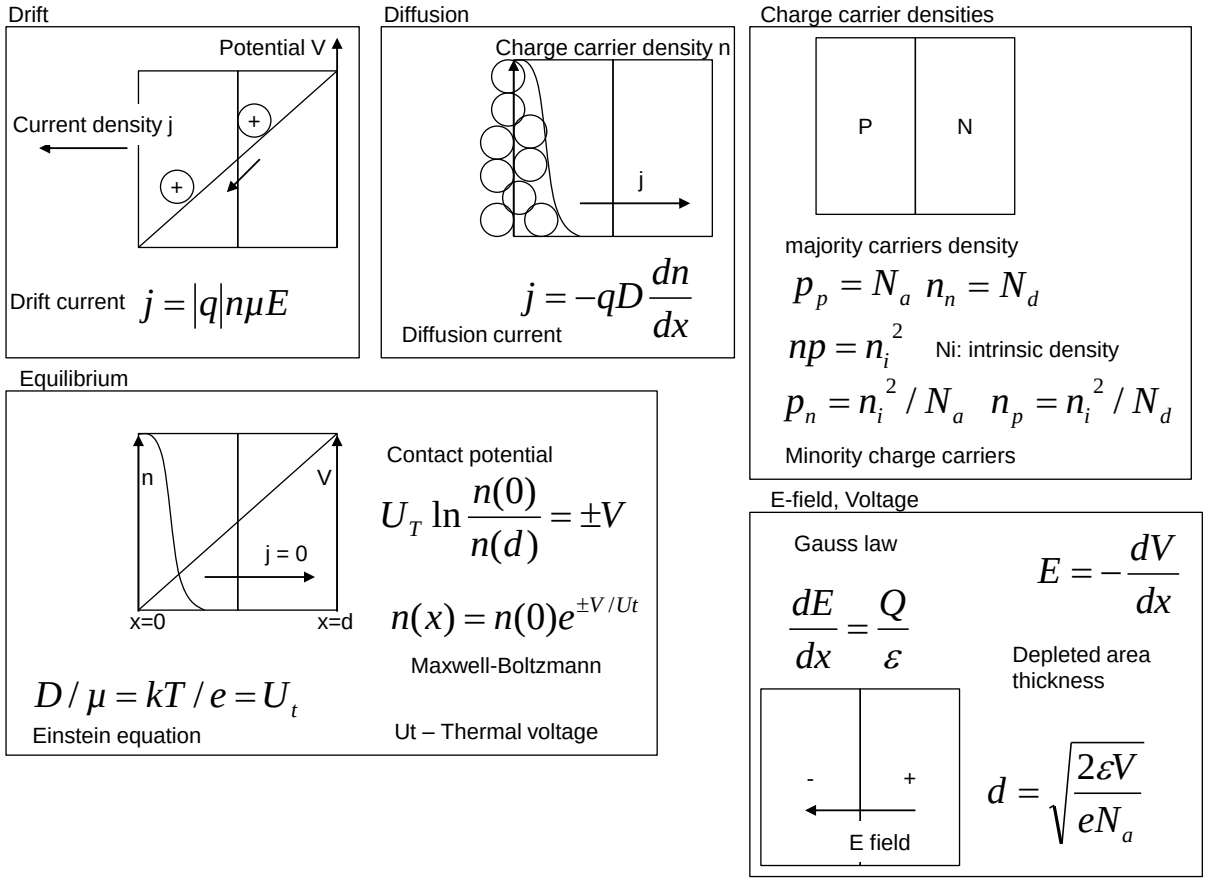



*Fig 21: V is potential, E is E-field, n is density of charge carries, q is the charge of the charge carrier (can be negative), Q is the charge density: (Q = q n); e is elementary charge (+1.6 $10^{19}$ C), μ is charge carrier mobility, D is diffusion constant, ε is permittivity (ε = $\varepsilon_r$ $\varepsilon_0$), $N_a$ are $N_d$ are densities of acceptors and donors.*

E-field generates drift current (Fig. drift).

If the charge density is inhomogeneous, a diffusion flow exists. (Fig. diffusion)

If the drift and diffusion currents compensate each other, we have a balanced state - equilibrium. The charge density is then described by the Maxwell-Boltzmann formula. Actually, the electrons are distributed according to Fermi-Dirac distribution. However, this distribution can be approximated by Maxwell-Boltzmann formula within one energy gap.

Gauss law describes how E-field is generated by charge. E-field is defined as the potential gradient.

**Potential within the MOS structure**

To simplify the analysis, we will assume that the metal behaves in the same way as the N-doped silicon in the source/drain/gate. (The metal/P-substrate junction has, in our approximation, exactly the same contact voltage as the N-source/P-substrate junction). We will also assume that the N-Source, N-Drain and N-Gate have the same doping density as the P-substrate (channel region). This is not entirely true in reality; however, it does not influence the results of the analysis. In our case, the contact potentials cancel out each other.

Under this assumption, there is no contact voltage between the N-doped regions – the source, the drain and the gate and the metal. (Gate, like source and drain, is n-doped.)

Therefore

$V_g^* = V_g$

$V_s^* = V_s$

$V_{gs}^* = V_{gs}$ (2)

The contact voltage between the substrate-contact and the silicon-substrate is given by the following formula:

$$V_b^* = V_b - V_{cont,np};\ V_{cont,np} = U_T \ln\left(\frac{N_a N_d}{n_i^2}\right) \quad (3)$$

$N_a$ and $N_d$ are the densities of acceptors in the substrate (in the channel area) and the donors in the source/gate/drain, $n_i$ is the intrinsic charge carrier density in silicon, about $10^{10}\,cm^{-3}$ at 300K. Thermal voltage $U_T = kT/e \sim 26$ mV at 300 K. We will assume the following values:

$N_a \sim 10^{18}\,cm^{-3} = N_d$.

When we substitute these values, we get:

$V_{cont,np} = 0.958$ mV ~ 1V

Figure 22 shows the potentials within the transistor.

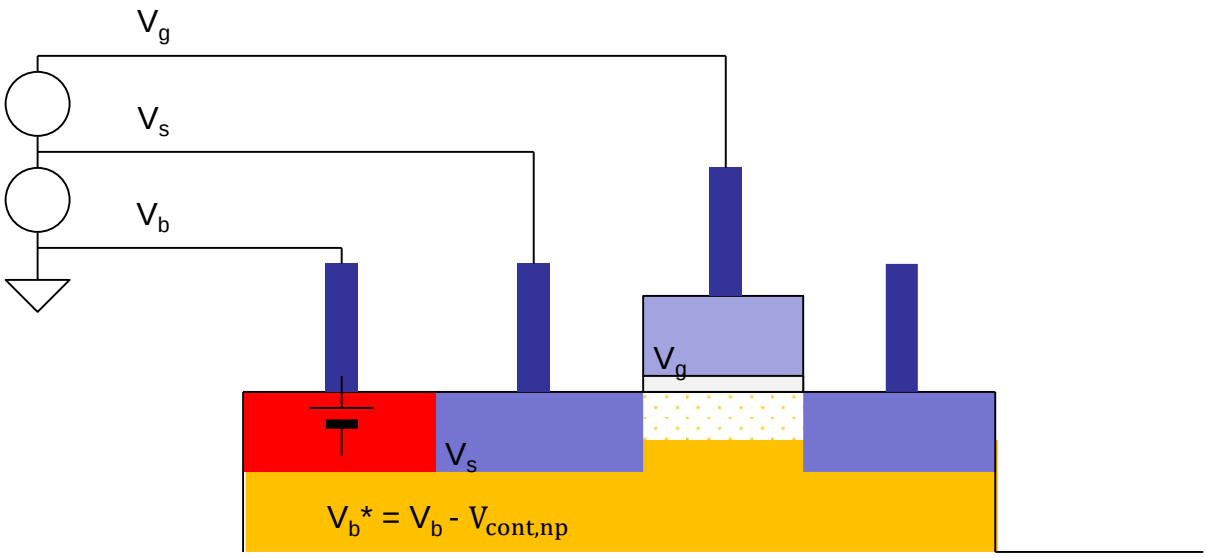



*Figure 22: Potentials in different areas of the MOS structure*

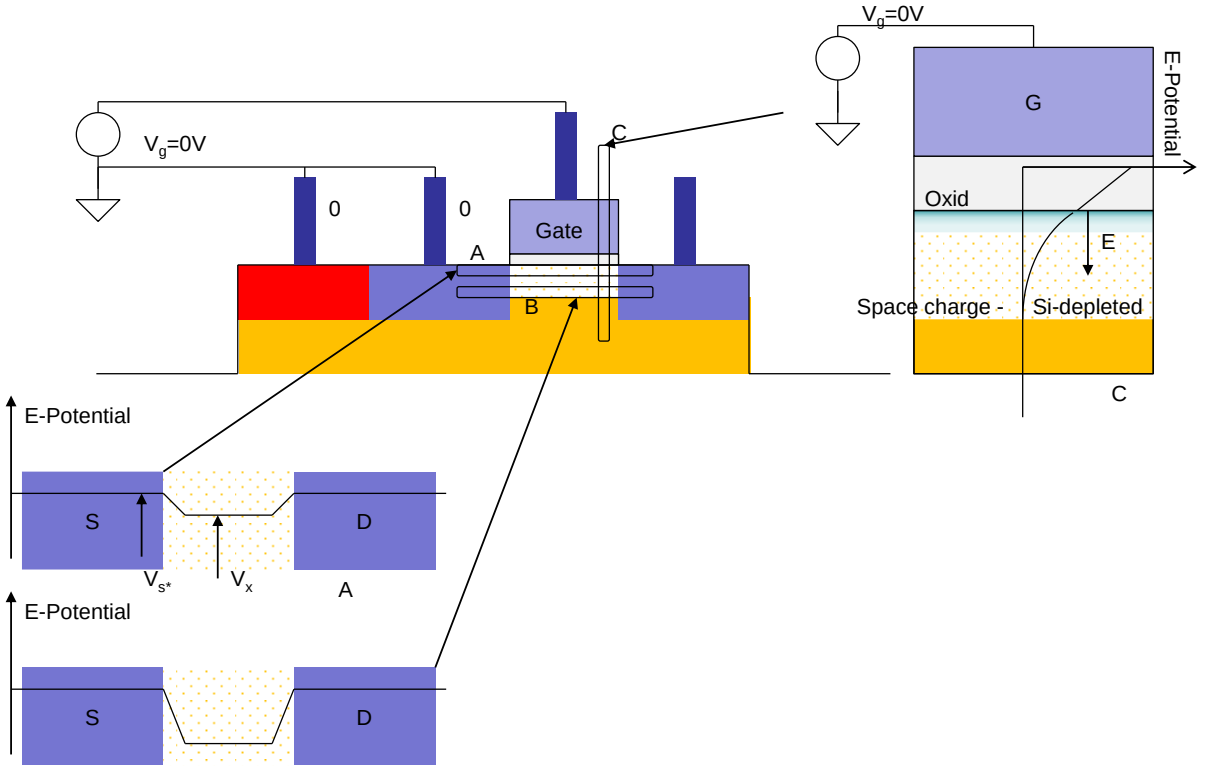



*Fig 23: Potentials in different regions of the MOS structure*

Let us now take a look at the structure in Fig 23. There are many electrons in the areas with higher electrical potential since electrons drift there. The holes are missing in these areas. The electron density (and hole density) can be approximated in thermodynamic equilibrium by the Maxwell-Boltzmann formula.

Let us define the regions A, B, C (two horizontal regions A and B and one vertical C) in the transistor structure (Fig 23).

**Vertical potential change**

Let us examine the space charge region in the P-substrate (region C in Fig 23). The potential change in the substrate causes the hole density to drop very quickly from $N_a$ to << $N_a$. (Holes are pushed downwards.) This causes a depletion zone (space charge zone) in the substrate, where the negative charge of the acceptor ions is not compensated (Fig 24). The total negative charge of the space charge region must be equal to the positive charge in the gate electrode (electro-neutrality).

If we look at the space charge zone in the P-substrate (region C), we expect that the electron density is largest close to the silicon-insulator interface, since this is where the electric potential is highest. If we apply a positive voltage to the gate (Fig 24, bottom), the holes will be pushed from the substrate even further downwards. The depletion zone gets larger. The electron density at the silicon-insulator boundary increases.

We call the effect when negative charge carriers accumulate in a P-silicon region "inversion". A p-region becomes an n-like.

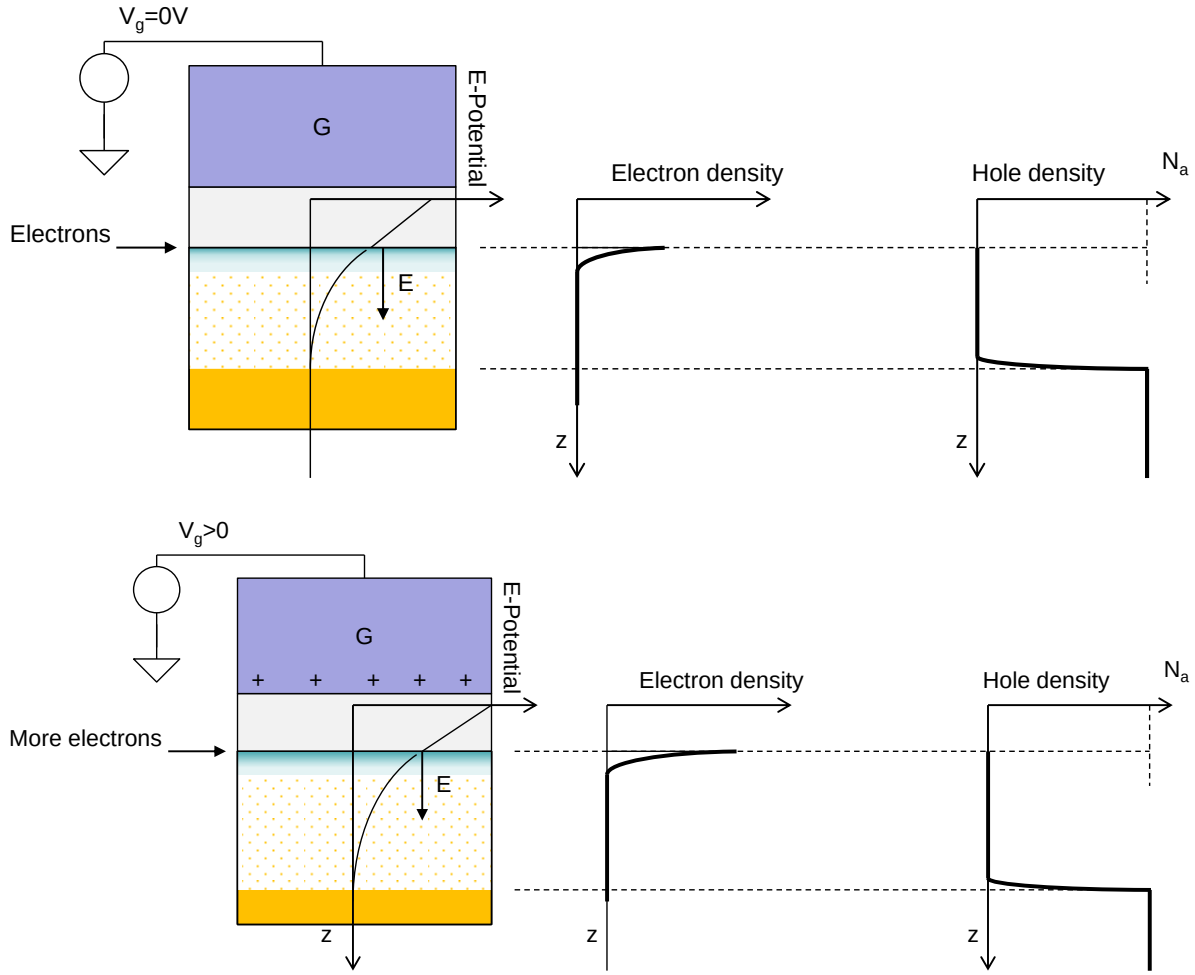



*Fig 24: If we apply a positive voltage on the gate, the holes from the substrate will be repelled and pushed downwards*

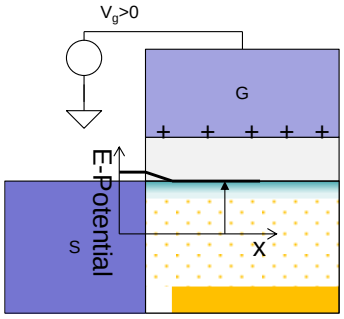
**Horizontal potential change**

Let us take look at the sections A and B in Fig 23. Positive potential change is a barrier for the positive charge (holes) and negative potential change is a barrier for electrons. The space charge zone under the gate represents a potential barrier for the electrons. It is difficult for an electron to pass from source to drain through the deeper P-silicon layers, where the potential barrier is large. When the electron moves just below the insulator layer, it sees the smallest potential barrier. The source-drain current flows in this narrow area, called the channel.

The height of the potential barrier between the source and the channel region is:

$U_B = | V_x - V_{s*}|$

$V_x$ is the potential of the silicon below the insulator.




One can show that not only the gate potential affects the density of electrons in the channel, but also the source potential. Electron density in the channel depends on the height of the barrier $U_B$. This may look strange. We have said the electron density can be approximated by Maxwell-Boltzmann formula. The potential $V_x$ does not depend on $V_s$. However, the density of electrons at the bottom of the depletion zone in the substrate depends on $V_s$. Therefore, the electron density below the insulator is also influenced by $V_s$.

The charge carrier density at the silicon-insulator boundary is described for $U_B < 0$ by the following formula:

$$n = n_0 e^{-U_B/U_T}$$

$n_0$ is the electron density in source, which is equal to the density of the donor ions: $n_0 = N_d$.

### Potential barrier as a function of $V_{dep}$

In the following text, we will derive the transistor current as a function of $V_{gs}$.

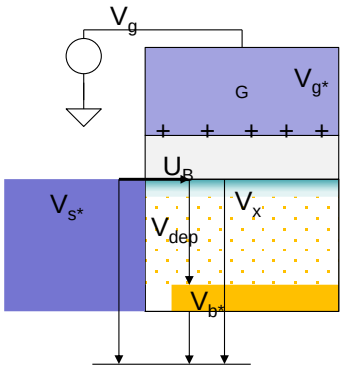



*Fig 25*

To calculate the current, we will first calculate the electron density in the channel. As mentioned, the electron density depends on the height of the potential barrier $U_B$, which depends on $V_{gs}$, and $V_{sb}$. We will derive this result in several steps:

$$-U_B = V_x - V_{s*} \quad (4)$$

The following holds (see Fig 25):

$$V_x = V_{b*} + V_{dep} \quad (5)$$

$V_{b*}$ is the potential in the substrate (p-silicon), $V_{dep}$ is the potential change within the depletion zone.

### Capacitances $C_{ox}$ and $C_{dep}$

To calculate $V_{dep}$ as a function of $V_g$, we will model the MOS structure with the series of two capacitances.

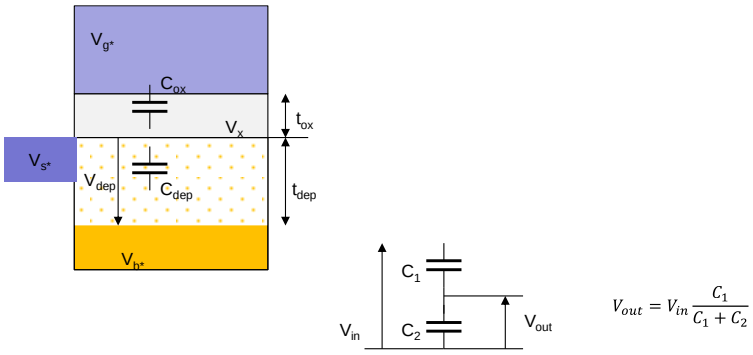



Figure 26: The capacitances in the MOS structure form a voltage divider. There are two capacitances: 1) the oxide capacitance: $C_{ox}$ – It depends on the oxide thickness and 2) the capacitance of the depletion zone: $C_{dep}$. $C_{ox}$ and $C_{dep}$ form a capacitive voltage divider.

*Why?*

When $V_{dep}$ increases, the negative charge in the depletion zone $Q_{dep}$ also increases and the depletion zone behaves like a capacitance $C_{dep}$. Since equal positive amount of charge $Q_{dep}$ accumulates in the gate electrode, the capacitance of the depletion zone $C_{dep}$ and the gate capacitance $C_{ox}$ make a series circuit, Figure 26. $C_{ox}$ is defined as $Q_{gate}/V_{ox}$ ($V_{ox}$ is the voltage in the gate oxide). The following applies:

$$C_{ox} = \varepsilon_0 \varepsilon_{SiO2} \frac{A}{t_{ox}} \sim 8.854\ 10^{-12}\ \frac{As}{Vm} \times 3.9 \times \frac{A}{t_{ox}} \qquad (6)$$

A is the gate area and $t_{ox}$ is the gate thickness. For a 65 nm technology $t_{ox} \sim 2.6$ nm.

$C_{dep}$ can be defined either as the normal DC-capacitance, $C_{dep,dc} = Q_{dep}/V_{dep}$, or as the dynamic AC capacitance $C_{dep,ac} = dQ_{dep}/dV_{dep}$. In both cases, $C_{dep}$ depends on $V_{dep}$.

The AC capacitance can be described with the following formula (proof in the box):

$$C_{dep,ac} \equiv \frac{dQ_{dep}}{dV_{dep}} = \varepsilon_0 \varepsilon_{Si} \frac{A}{t_{dep}} \quad (7)$$

$Q_{dep}$ is the charge in the depletion zone, $V_{dep}$ is the potential change within the depletion zone, and $t_{dep}$ is the thickness of the depleted zone, $\varepsilon_{si} \sim 12$.

It can be shown (see proof below) that following formula holds:

$C_{dep,dc} = Q_{dep}/V_{dep} = 2\ C_{dep,ac}$ (8)

Proof: (optionally)

We can calculate the size of the depleted region in the following way.

Let us calculate the E-field in z-direction. Z-coordinate is defined to be 0 at the bottom edge of the depleted region and it shows upwards. Gauss's law:

$$\frac{dE_z}{dz} = -\frac{eN_a}{\varepsilon_0 \varepsilon_{Si}} \Rightarrow E_z = -\frac{eN_a}{\varepsilon_0 \varepsilon_{Si}} z \qquad (B1)$$

Potential:

$$-\frac{dV_z}{dz} = E_z \Rightarrow V_z = \frac{eN_a}{\varepsilon_0\varepsilon_{Si}}\frac{z^2}{2} \quad (B2)$$

It follows:

$$t_{dep} = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}V_{dep}}{eN_a}} \quad (B3)$$

The charge in the depleted zone is:

$$Q_{dep} = AeN_a t_{dep} = A\sqrt{eN_a 2\varepsilon_0\varepsilon_{Si}V_{dep}} \quad (B4)$$

The dynamic capacitance of the depleted zone is:

$$C_{dep,ac} \equiv \frac{dQ_{dep}}{dV_{dep}} = A\sqrt{\frac{eN_a\varepsilon_0\varepsilon_{Si}}{2V_{dep}}} \quad (B5)$$

It holds because of (B5) and (B3):

$$C_{dep,ac} = A\frac{\varepsilon_0\varepsilon_{Si}}{t_{dep}} \quad (B6)$$

It holds also:

$$Q_{dep} = A\sqrt{eN_a 2\varepsilon_0\varepsilon_{Si}V_{dep}} = V_{dep}A\sqrt{\frac{2\varepsilon_0\varepsilon_{Si}eN_a}{V_{dep}}} = 2C_{dep,ac}V_{dep} \quad (B7)$$

and

$$C_{dep,dc} = \frac{Q_{dep}}{V_{dep}} = 2C_{dep,ac} \quad (B8)$$

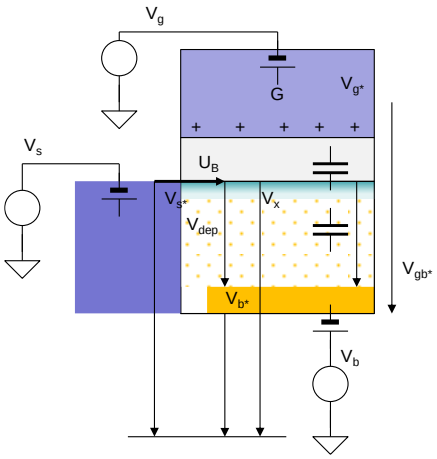
The doping of silicon and the oxide thickness are usually chosen so that $C_{ox}$ is about 4 times greater than $C_{dep,ac}$.

It holds: $C_{ox} = 4 \times C_{dep,ac}$ (5) and $C_{ox} = 2 \times C_{dep,dc}$

**We define a factor (called a "slope factor") as:**

**$n = (C_{dep,ac} + C_{ox})/C_{ox} \sim 1.25$ (9)**

**Potential barrier as a function of $V_{gs}$**




Let us calculate the potential barrier: $-U_B = V_x - V_{s*}$ as a function of $V_{gs}$.

It holds:

$$-U_B = V_x - V_{s*} = V_{b*} + V_{dep} - V_{s*}$$

We can use the formula for voltage divider. It follows:

$$V_{dep} = V_{gb*}\frac{C_{ox}}{C_{dep} + C_{ox}}$$

and:

$$-U_B = V_{b*} + V_{gb*}\frac{C_{ox}}{C_{ox} + C_{dep}} - V_{s*} =$$

$$V_{g*}\frac{C_{ox}}{C_{ox} + C_{dep}} + V_{b*}\frac{C_{dep}}{C_{ox} + C_{dep}} - V_{s*} =$$

$$V_{gs*}\frac{C_{ox}}{C_{ox} + C_{dep}} + V_{bs*}\frac{C_{dep}}{C_{ox} + C_{dep}} =$$

$$\frac{C_{ox}}{C_{ox} + C_{dep}}\left(V_{gs*} - \frac{C_{dep}}{C_{ox}}V_{sb*}\right)$$

Because (2) and (3), we get:

$$-U_B = \frac{C_{ox}}{C_{ox}+C_{dep}}\left(V_{gs} - \frac{C_{dep}}{C_{ox}}(V_{cont,np} + V_{sb})\right) \quad (10)$$

**Threshold voltage**

We define the threshold voltage as $V_{gs}$ at which $U_B = 0$. Let us start with (10):

$$-U_B = \frac{C_{ox}}{C_{ox} + C_{dep}}\left(V_{gs} - \frac{C_{dep}}{C_{ox}}(V_{cont,np} + V_{sb})\right)$$

It follows:

$$V_{gs} = \frac{C_{dep}}{C_{ox}}(V_{cont,np} + V_{sb}) \Rightarrow U_B = 0$$

We refer to the threshold for $V_{sb}$ unequal 0 as $V_{thsb}$.

$$V_{thsb} \equiv \frac{C_{dep}}{C_{ox}}(V_{cont,np} + V_{sb}); C_{dep} = C_{dep,dc} \text{ für } V_{dep} = V_{cont,np} + V_{sb}$$

The threshold at $V_{sb} = 0$ can be understood as the "base line value" of the threshold - $V_{th}$ and the following approximation can be written for $V_{thsb}$:

$$V_{thsb} \sim \frac{C_{dep,dc}}{C_{ox}}V_{cont,np} + \frac{C_{dep,ac}}{C_{ox}}V_{sb}; C_{dep,ac}C_{dep,dc} \text{ for } V_{dep} = V_{cont,np}$$

or

$$V_{thsb} \sim V_{th} + (n-1)V_{sb}; V_{th} \equiv \frac{C_{dep,dc}}{C_{ox}}V_{cont,np} \quad (11)$$

It holds:

$$-U_B = V_x - V_{s*} = \frac{C_{ox}}{C_{ox}+C_{dep}}\left(V_{gs} - V_{thsb}\right) \quad (12)$$

Let us calculate $V_{th}$ as function of $V_{cont}$. (optionally)

It holds (Equation B5):

$$C_{dep,ac} = A\sqrt{\frac{eN_a\varepsilon_0\varepsilon_{Si}}{2V_{cont}}} \quad (13)$$

Therefore, it is:

$$V_{th} = \frac{C_{dep,dc}}{C_{ox}} \times V_{cont} = \frac{2C_{dep,ac}}{C_{ox}} \times V_{cont} = \frac{A\sqrt{2eN_a\varepsilon_0\varepsilon_{Si}V_{cont}}}{C_{ox}} \quad (14)$$

The contact voltage is given by (3). Since we assume $N_d = N_a$, it holds:

$$V_{cont} = 2U_T \ln\left(\frac{N_a}{n_i}\right) \quad (15)$$

Equation (14) is usually given in the literature.

Note that the threshold voltage decreases when $C_{ox}$ is increased.

Smaller threshold is better when the supply voltage is low. A low supply voltage results in lower power consumption. Therefore, one tries to maximize $C_{ox}$ or to make the thickness of the oxide as small as possible.

As the temperature rises, the threshold voltage decreases (14) because the contact potential decreases (15). It is the consequence of the increase in intrinsic carried density in silicon $n_i$.

## Strong Inversion

If we increase the gate potential (gate source voltage) over the threshold voltage, the potential at the substrate surface should increase over 0 V. However, this would cause that electrons from the source and drain flow into the regions below the oxide, since a potential minimum for them is formed (Fig 27).

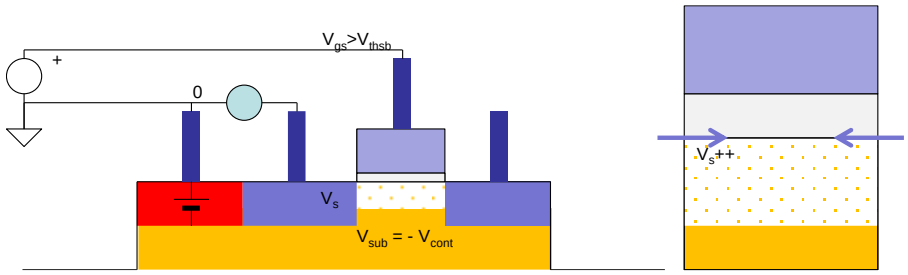



*Fig 27*: Case $V_{gs} > V_{thsb}$

The electron density would be higher than in source and drain. In reality, the electrons get collected and form a conductive channel. The electrons in the channel short the source, drain and the channel region together and thus keep, through their own charge, the channel potential at the level of source and drain. The channel and the source/drain are therefore short-circuited Fig 28. We refer to this operation region strong inversion.

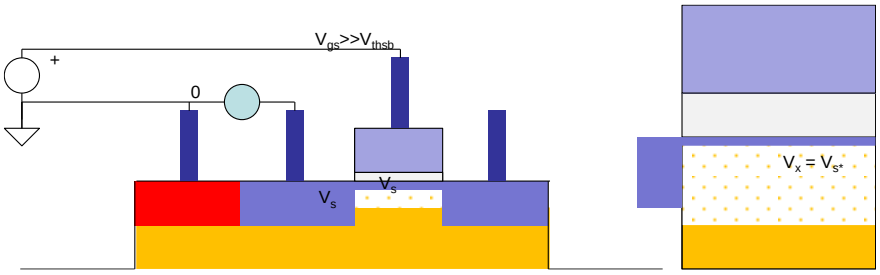



*Fig 28*: Electrons form the channel. Source and drain are shorted -> $V_x = 0$

## Channel Charge

Let us calculate the charge in the channel:

The bottom electrode of $C_{ox}$ is at a fixed potential. The voltage at $C_{dep}$ is constant. The voltage source at the gate therefore “sees” only the input capacitance $C_{ox}$. When the gate voltage changes by $dV_{gs}$, the charge amount $C_{ox}\, dV_{gs}$ flows through the voltage source. This charge is formed in the channel.

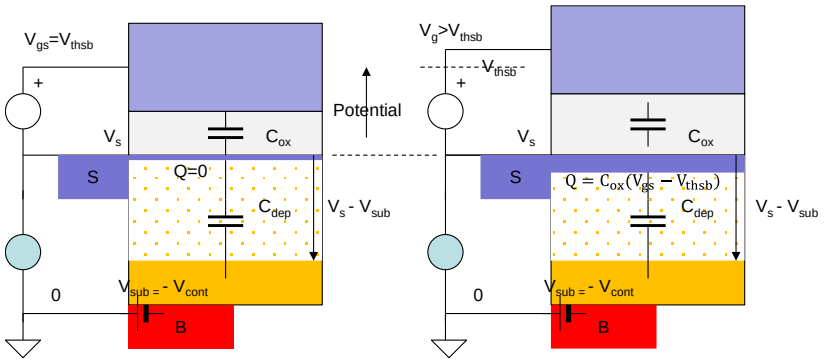



*Fig 29*: Channel charge

We make the following assumption:

For $V_{gs} = V_{thsb}$ there is no charge in the channel.

This assumption is not correct for every application (see "weak inversion")

For $V_g > V_{thsb}$ it holds $dQ = C_{ox}\, dV_{gs}$. Therefore, it holds

$$Q = C_{ox}(V_{gs} - V_{thsb})$$

Since we have $V_s = 0$, we can also write the following formula:

$$Q = C_{ox}(V_{gs} - V_{thsb}) \quad (20)$$

Summary:

We define the threshold voltage at drain side $V_{thsb}$ as the gate source voltage for which the potentials in the source and on the substrate-surface are approximately equal.

(For $V_{gs} = V_{thsb}$, the potential barrier between source and channel is zero.)

When the gate source voltage rises above the threshold, the electrons collect in the channel. We have a strong inversion.

For gate voltages below the threshold, the potential at the substrate surface is not sufficient for channel formation.