Lecture 2

Summary of the lecture

Production of a MOSFET

Structure of a MOSFET

Working principle

Influence of the vertical E-field component: charge in the channel as function of gate-source voltage

Influence of the horizontal E-field component: drain-source current as function of drain-source voltage

Saturation of I_{ds}

Production of a MOSFET

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

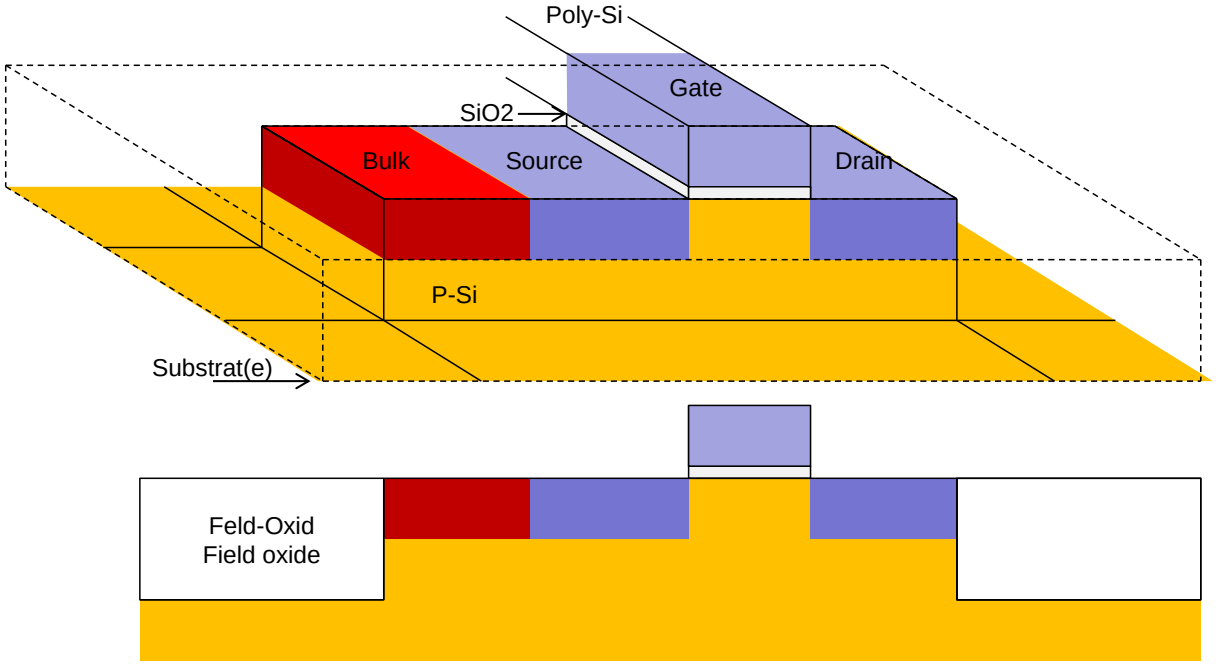


Fig 1: 3D image of an NMOS Transistor

A MOSFET is made of 4 electrodes: source, drain, gate and substrate (called also 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.

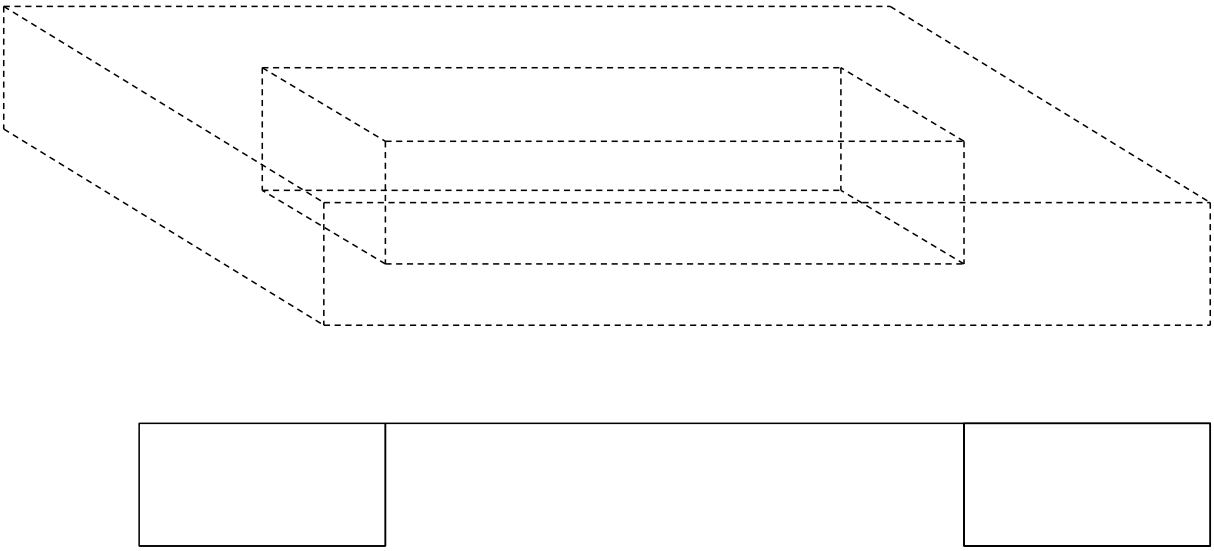


Fig 2: Transistor is surrounded by field oxide

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

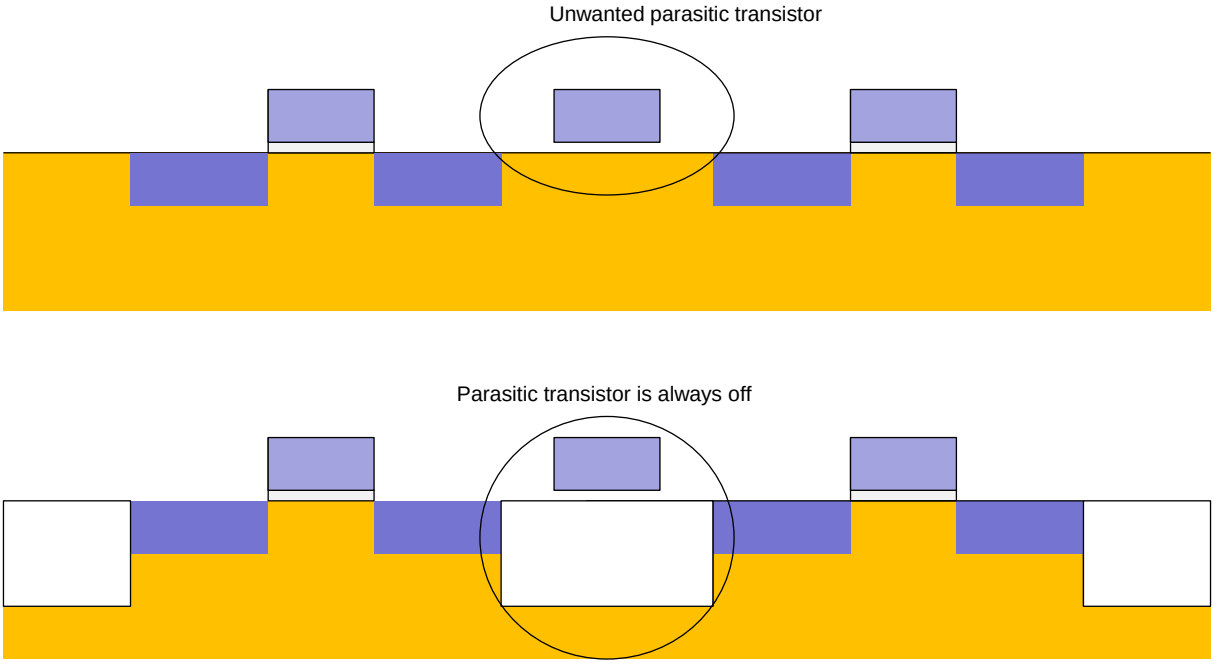


Fig 3: Isolation of the transistors with field oxide

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

The 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).

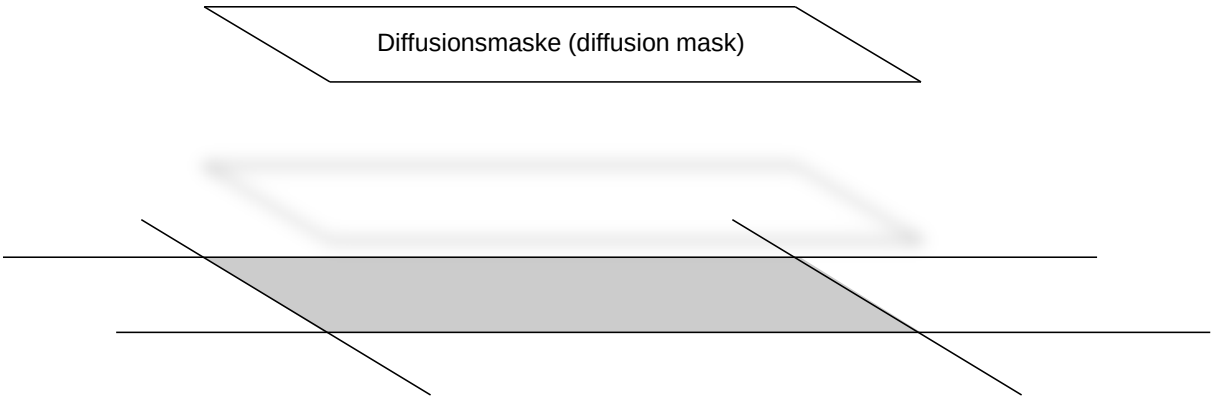


Fig 4: Diffusion mask defines the active region of the substrate where the transistor will be placed

The next step is the production of trenches for the field oxide.

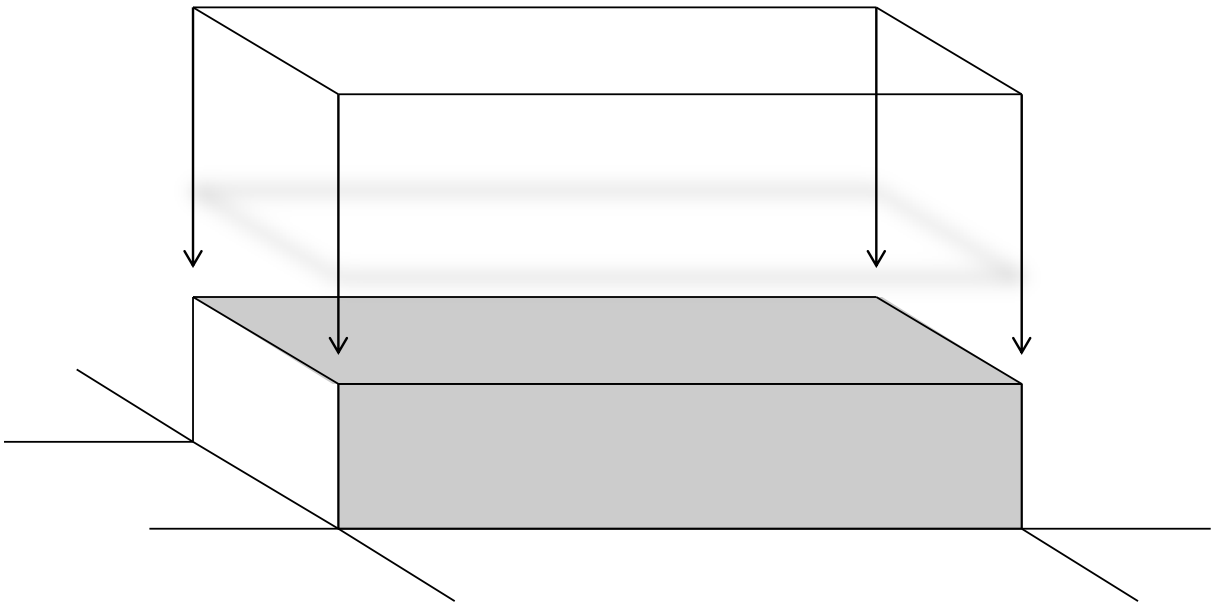


Fig 5: Trench production

The 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

In the following 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.

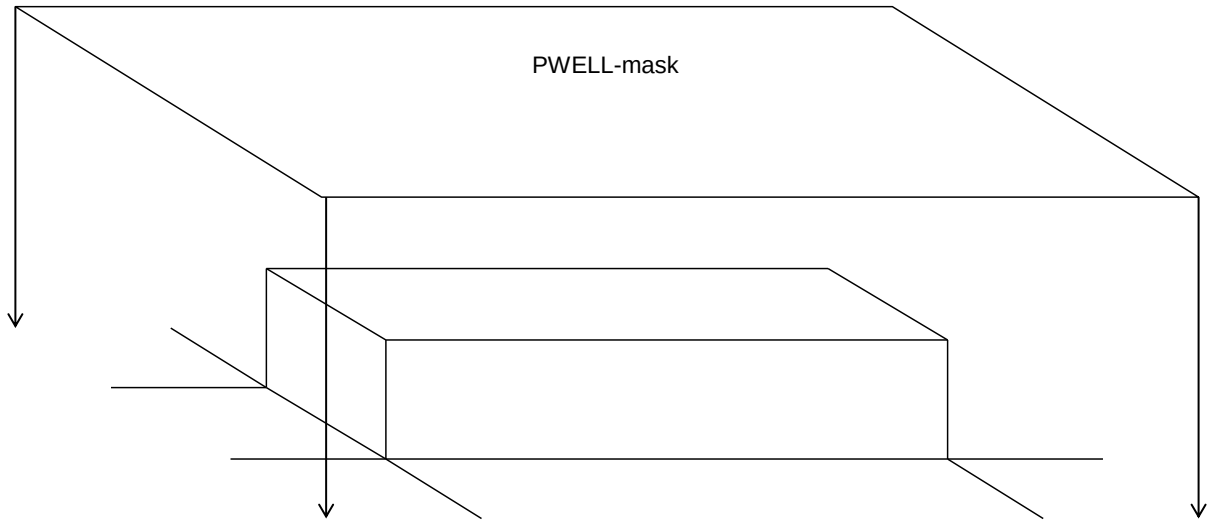


Fig 6: NMOS is placed in a P-substrate

The wells are produced by ion-implantation.

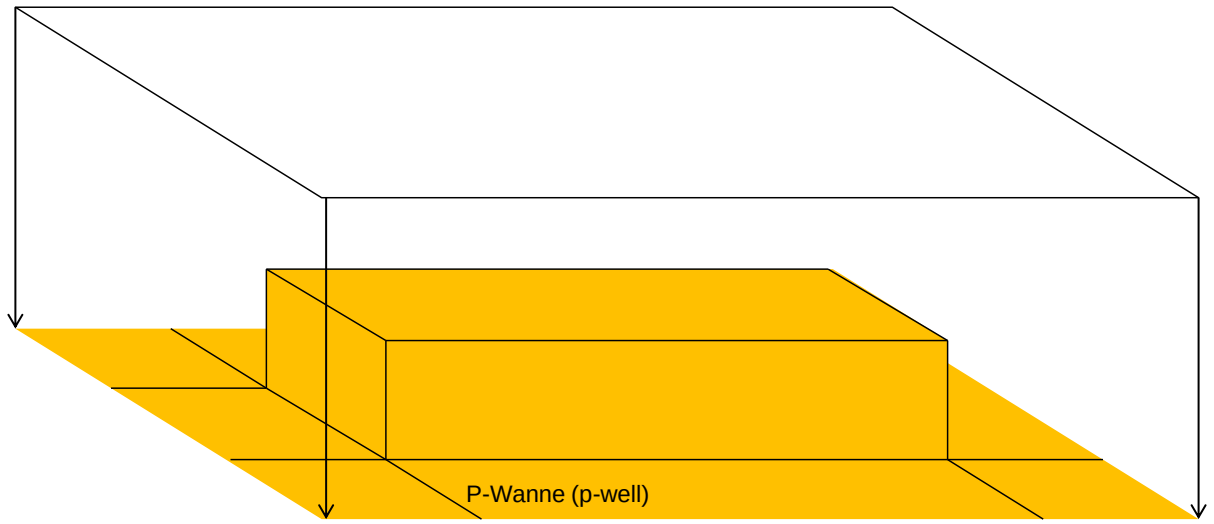


Fig 7: P-well

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 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 capacity is about 13 fF/m^2. As we will see later, oxide capacity 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.

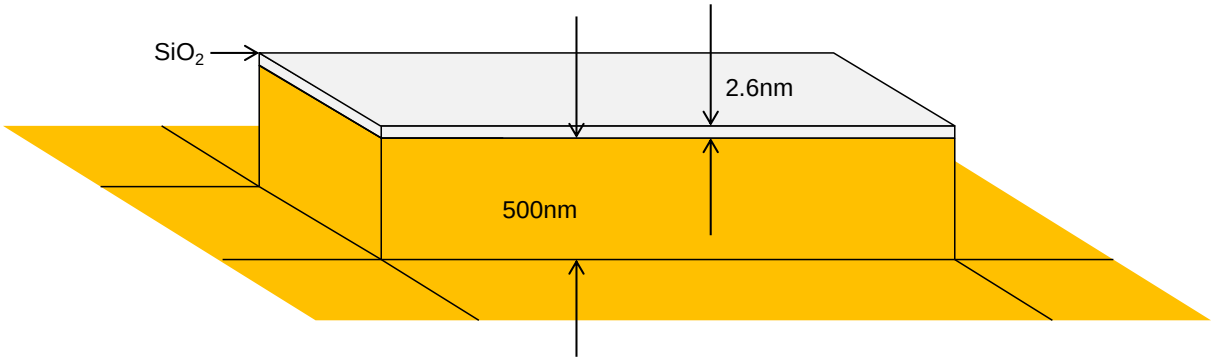


Fig 8: Thermal oxidation

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

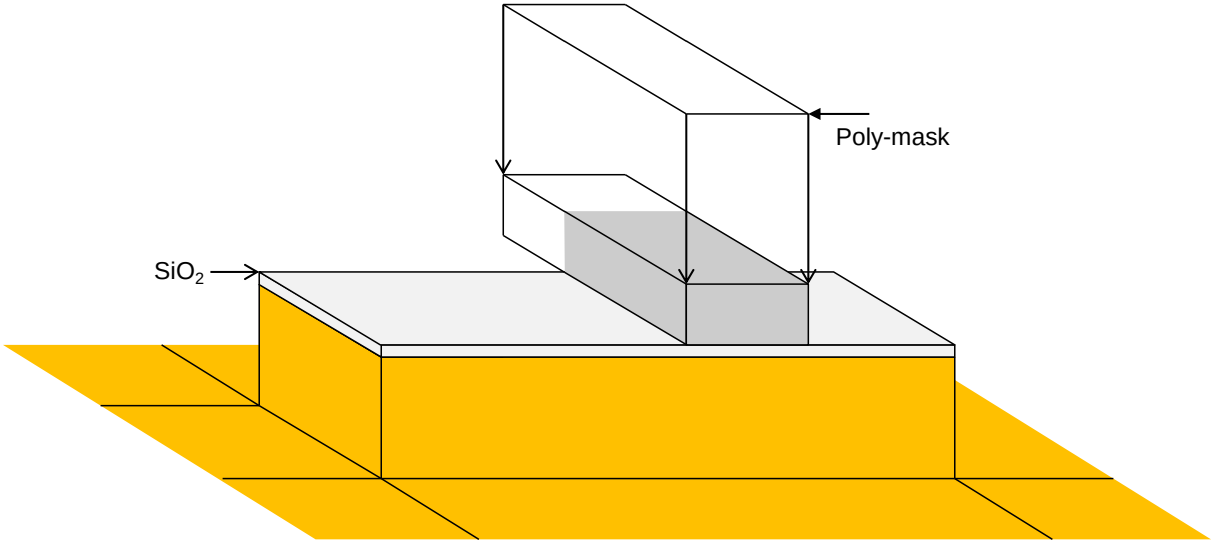


Fig 9: Polysilicon ("poly") mask defines the gate electrode and its connection

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.

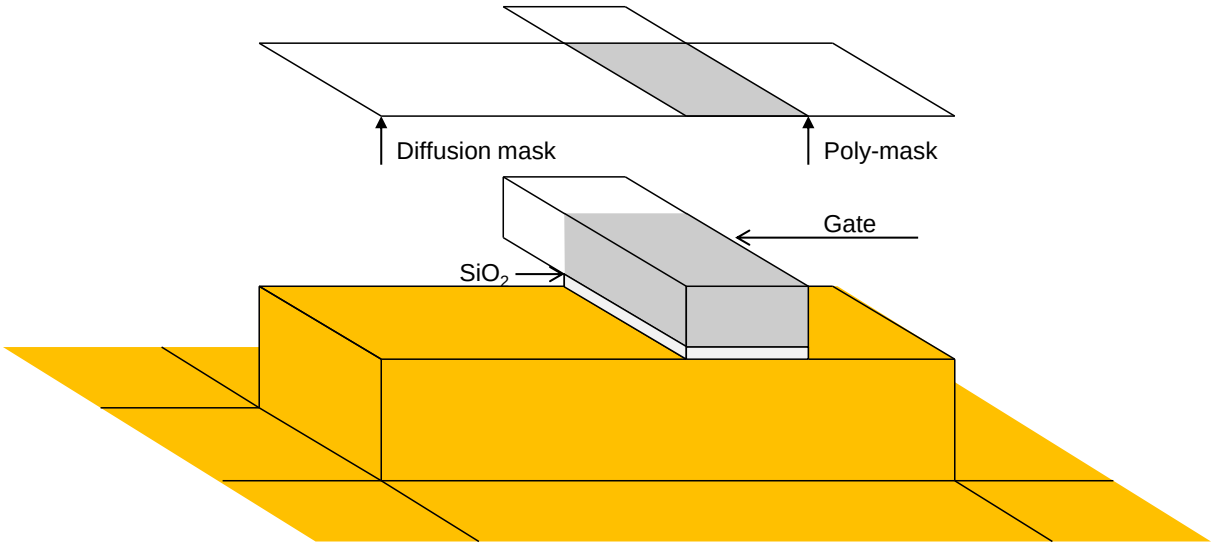


Fig 10: The thin oxide remains in the overlap region of diffusion- and poly-mask

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 the process step is self-aligning.

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 are used, all together, as masks for doping.

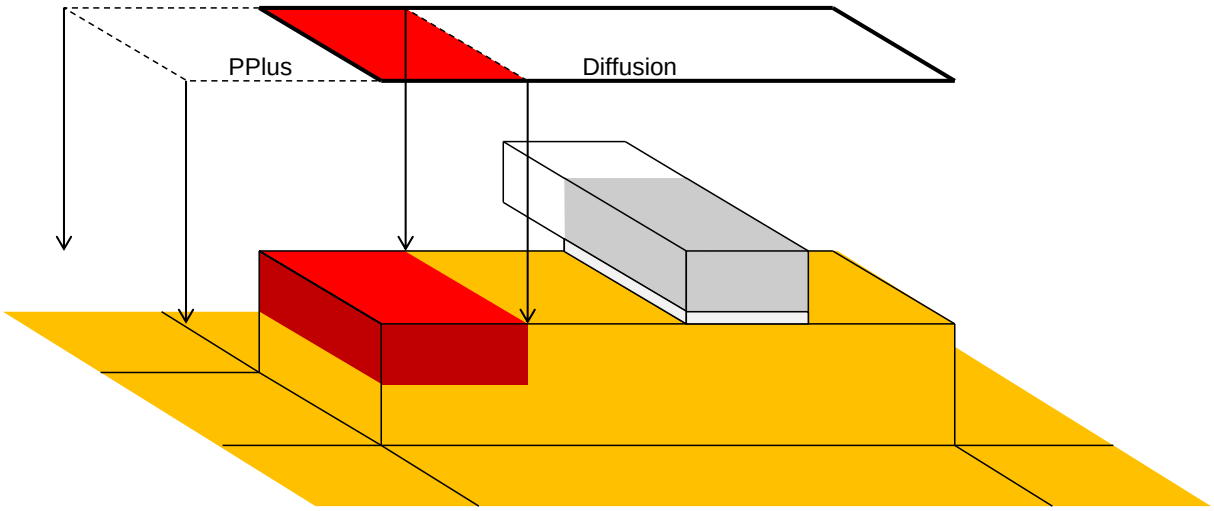


Fig 11: PPlus mask. The overlap region between PPlus- and diffusion masks will be P-doped

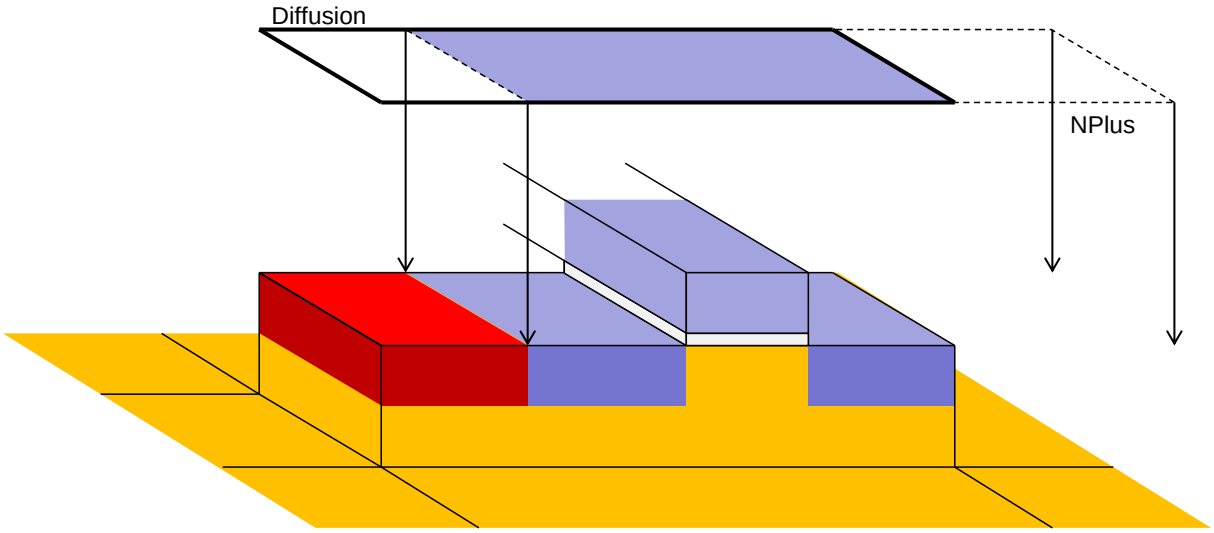


Fig 12: NPlus mask. The overlap region between NPlus- and diffusion masks will be N-doped

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

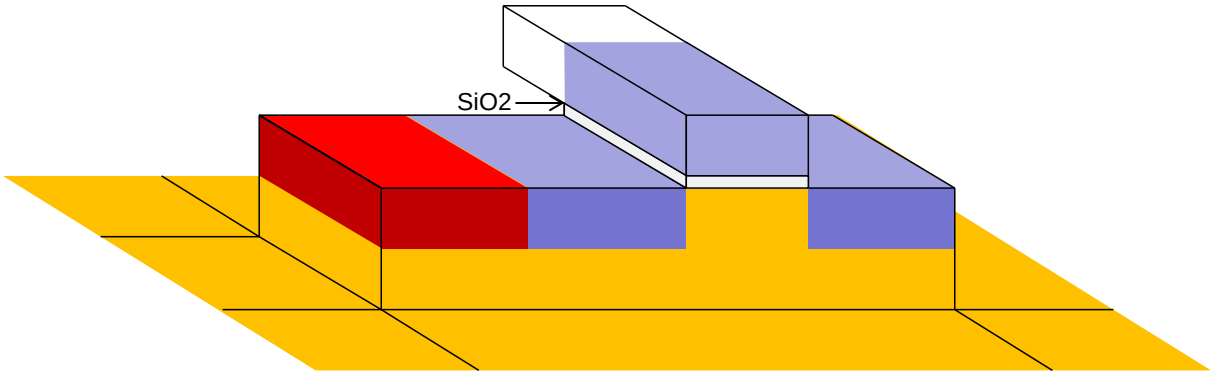


Fig 13: Transistor with all structures

MOSFET

A MOSFET transistor is made up with four electrodes: source, drain, gate and substrate (bulk), as shown in Fig 14.

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.

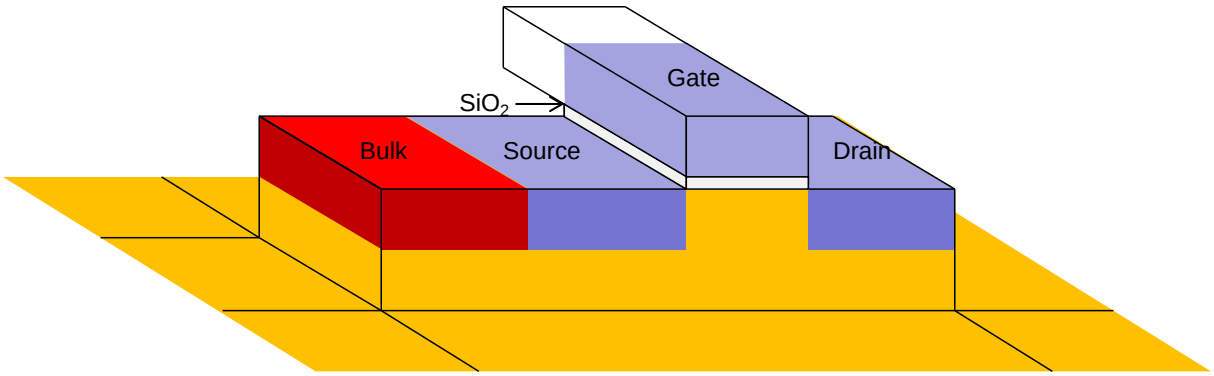


Fig 14: NMOS

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”.

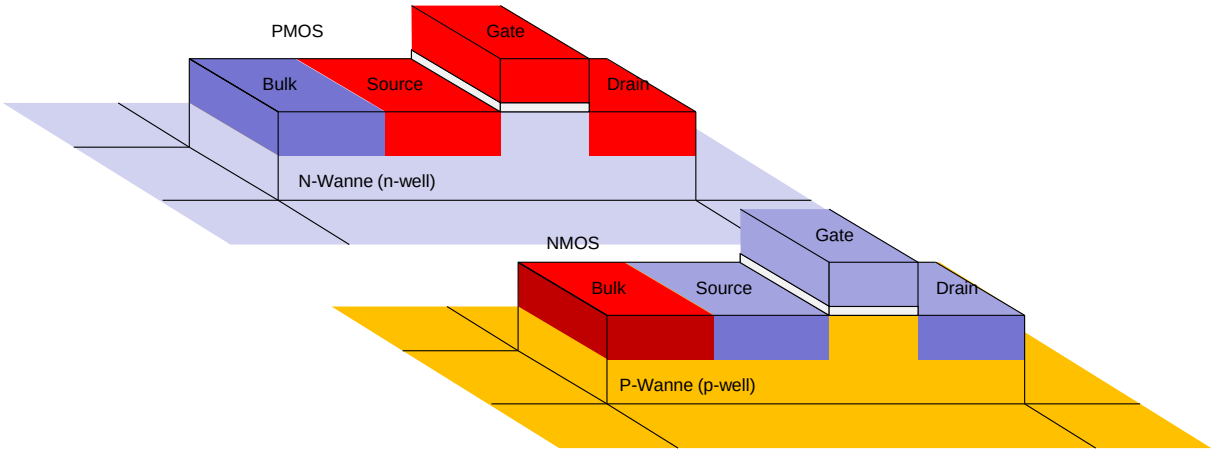


Fig 15: NMOS und PMOS

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.

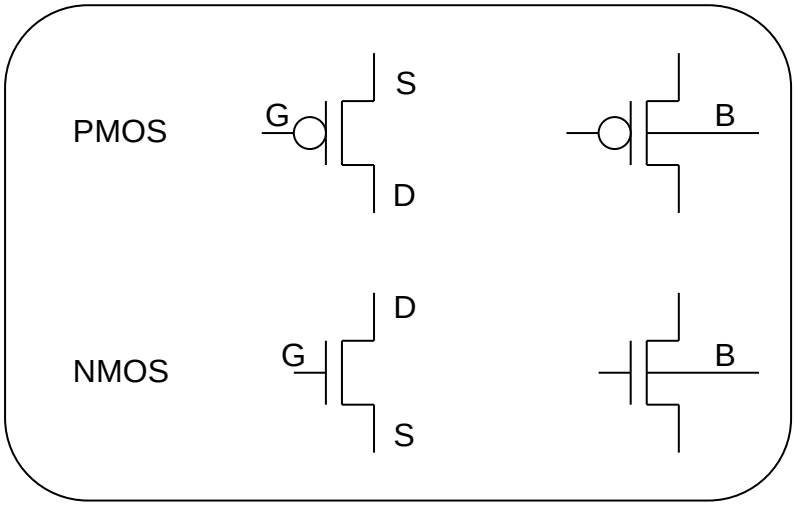


Fig 16: Simplified symbols

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.

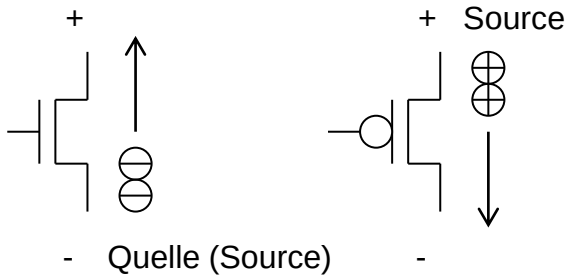


Fig 17: Source

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.

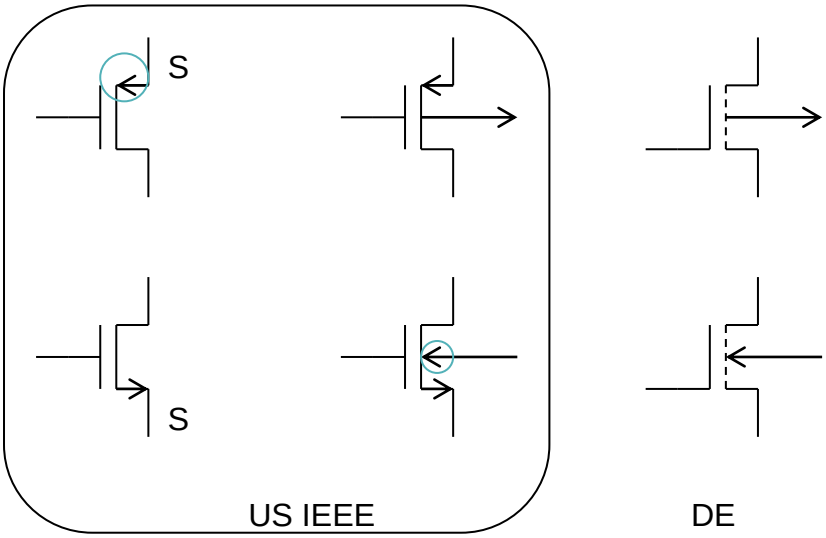


Fig 18: 4-terminal symbols

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.

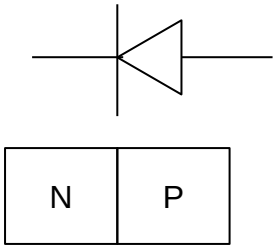


Fig 19: Diode symbol

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. 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.

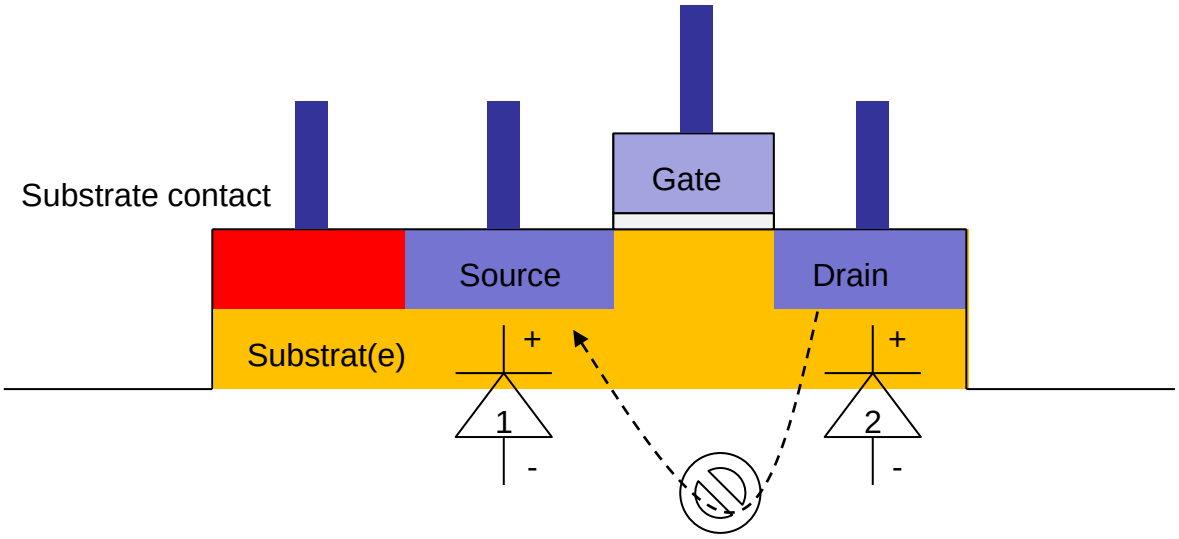


Fig 20: The PN Diodes should be reversely biased

Contact voltages

A contact voltage is induced between P-silicon and N-silicon and between silicon and metal regions, as shown in Fig 21. In order to simplify the analysis, we assume that the metal behaves in a similar way as the N-doped silicon. This is not fully correct, but it does not change significantly the results of further analysis. Under our assumption there is no contact voltage between the N-doped silicon regions and metal. These N-doped silicon regions are the source, the drain and the gate. For simplicity we will also assume that the dopant density is equal in the source, in the drain, in the gate, as well as in the p-substrate (channel region).

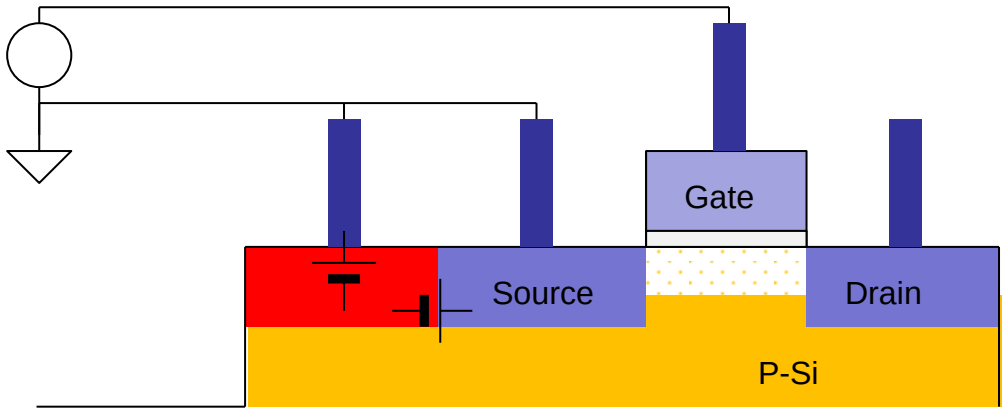
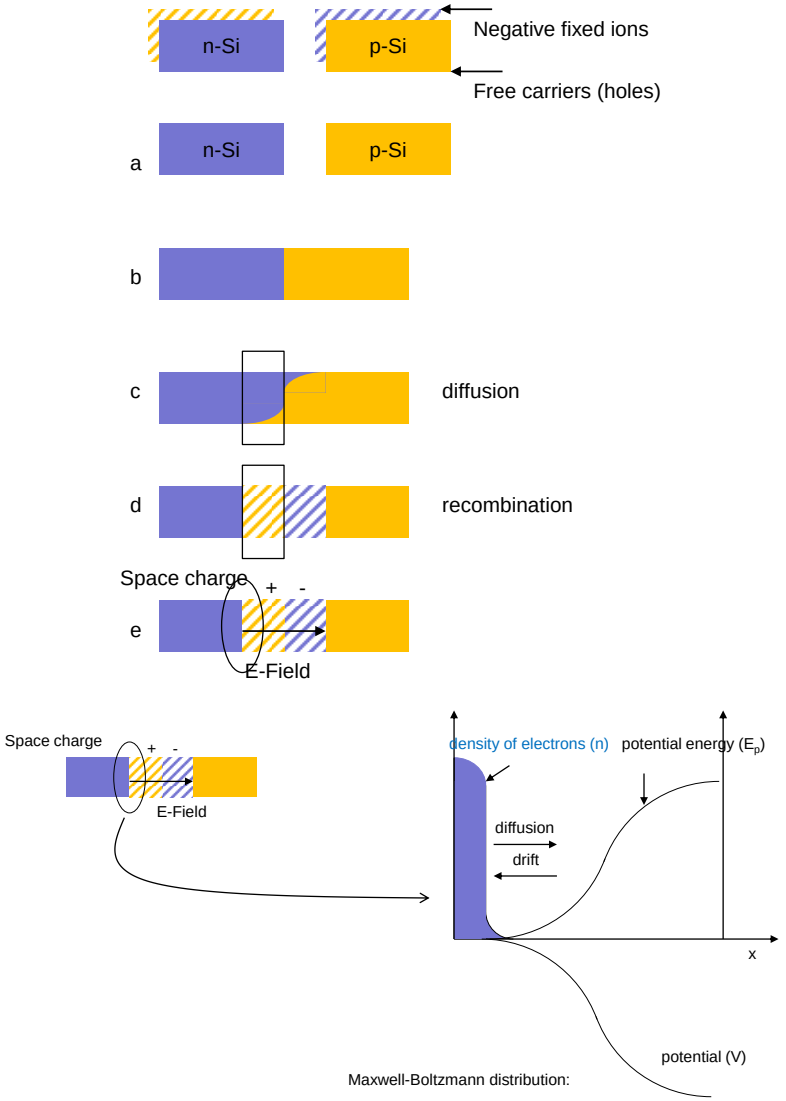


Fig 21: Contact voltages are induced between silicon- and metal regions

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 22: Origin of contact voltage

Simplified explanation (Fig 22): 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 compensated anymore 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 $n_p=n_i/N_a$. 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 22B.

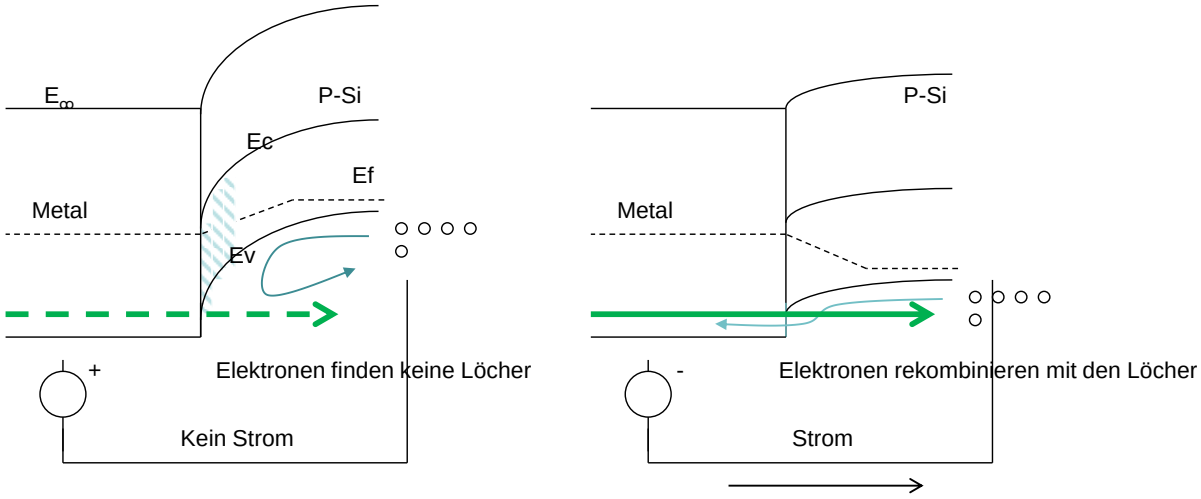


Fig 22B: Schottky diode

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 22C).

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.

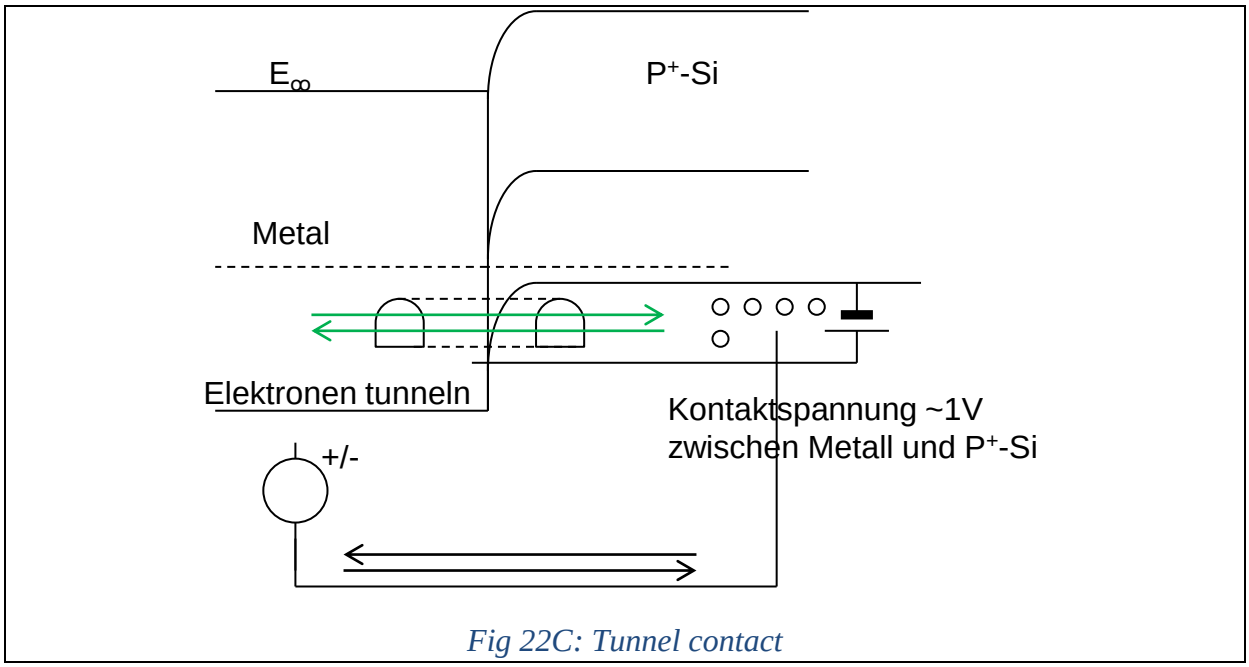


Fig 22C: Tunnel contact

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

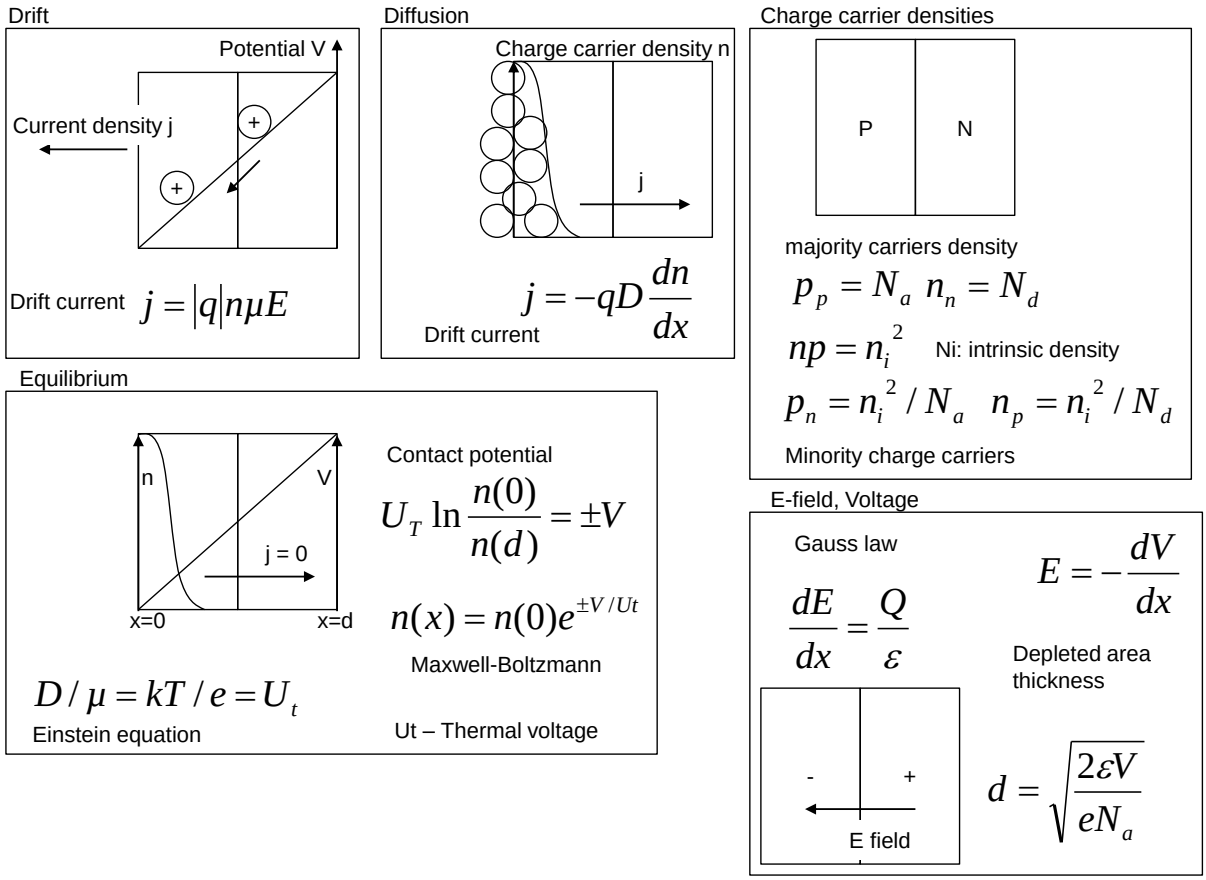


Fig 23: V is potential, E is E-field, n is density of charge carries, q is the charge of the charge carrier (can be negative), Q is the charge density: (Q = q n); e is elementary charge (+1.6 10^{19} C), μ is charge carrier mobility, D is diffusion constant, ε is permittivity (ε = ε_r ε_0), N_a are N_d 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

Fig 24 shows the potentials within the MOSFET when we set source, drain and gate electrodes at 0 V.

We define the potential at the metal contact of source as reference potential of 0 V. Since we made the assumption that there is no contact voltage between the metal and the n-silicon, the potential of the silicon region of source is also 0 V. The drain potential (both silicon and metal) is also 0 V. The bulk contact (metal) is 0 V. However because of the contact voltage, the potential of p-silicon (transistor substrate) is lower. The contact voltage can be calculated using the formula:

$$V_{cont} = U_T \ln\left(\frac{N_a N_d}{n_i^2}\right) \quad (1)$$

Where N_a is the density of the acceptors in the substrate (in the channel region) and N_d the density of the donors in source, n_i is the intrinsic charge carrier density, about 10^{10} cm^{-3} at 300 K. Thermal voltage $U_T = kT/e \sim 26$ mV at 300 K. We assume following values: $N_a \sim 10^{18}$ $cm^{-3} = N_d$. When we substitute these values in (1) we obtain:

$V_{cont} = 0.958$ mV ~ 1V.

The potential oft he p-silicon substrate is therefore -1 V.

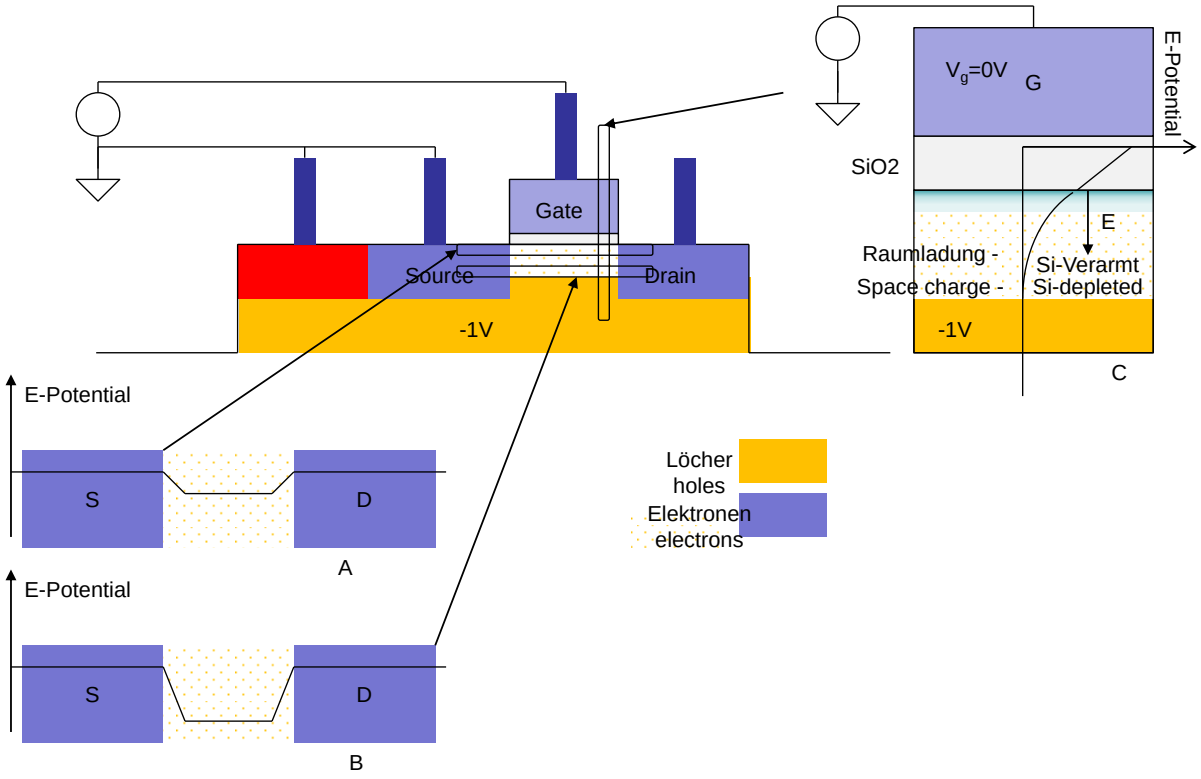


Fig 24: Potentials in different regions of the MOS structure

Fig 24 shows the potential variation in three sections A, B and C (two horizontal sections A and B and one vertical section C).

Generally it holds the following: There are more free electrons in the regions with lower electron potential energy (and higher electrical potential). These regions are source and drain. The electrons drift towards the areas with higher electrical potential. As mentioned, the electron density follows the Maxwell-Boltzmann formula (which is approximation of the Fermi-Dirac distribution).

If we look at the depletion zone in the P-region (Section C), we expect that the electron density is the largest near Si-SiO_2 boundary (also called Si-SiO_2 interface), because the electrical potential is greatest near the boundary.

The potential change in the substrate causes also that the density of holes quickly falls from the value in undepleted substrate N_a to much less then N_a. A depleted region with sharp edge is generated, as shown in Fig 25. Within the depleted region, the negative charge of acceptor ions in not compensated. The total negative charge of the depleted zone must be equal to the positive charge at the gate electrode, otherwise the MOS structure would not be electro-neutral.

If we apply a higher positive voltage to the gate, the holes from the substrate will be further repelled and pushed downwards.

The depletion zone is the expanding. The electron density at the Si-SiO_2 interface is increasing. We call this effect “inversion”.

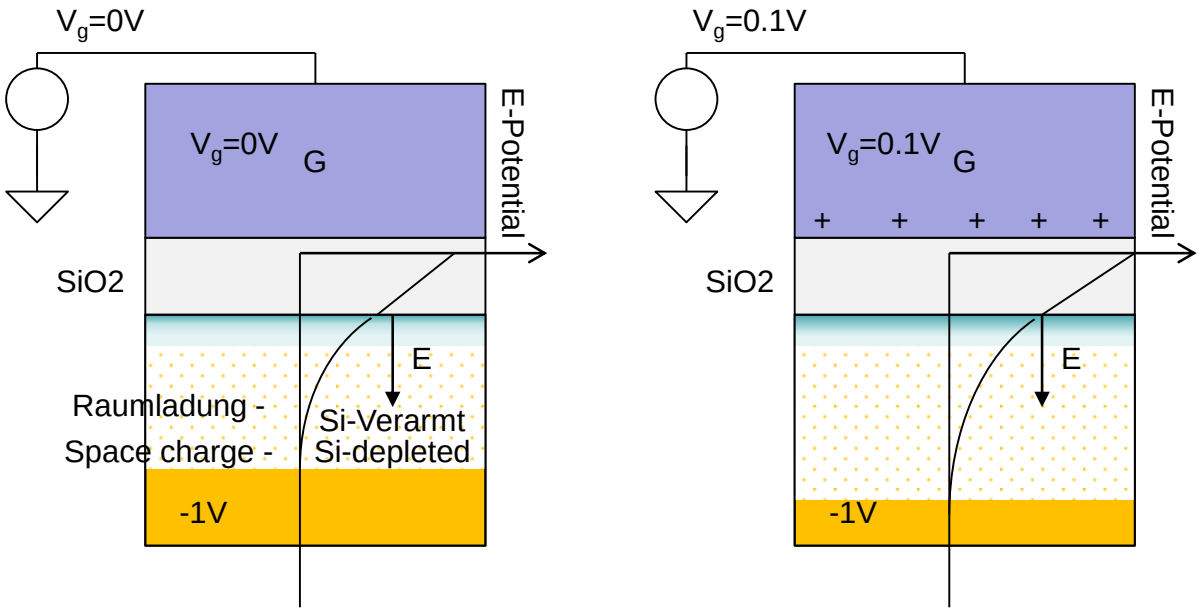


Fig 25: If we apply a positive voltage on the gate, the holes from the substrate will be repelled and pushed downwards

Channel charge

In the following sections we will derive the transistor current as a function of V_g (or V_{gs}). To calculate the current, we should first determine the electron density in the silicon region underneath SiO_2. The electron density depends on the potential at the Si-$Si0_2$ boundary. Therefore let us first calculate the potential at the Si-$Si0_2$ boundary.

This potential is given by the equation:

$$V_{Si,SiO2} = V_{sub} + V_{dep} = V_s - V_{cont} + V_{dep} \approx -1V + V_{dep} \qquad (2)$$

V_{sub} is the potential of the substrate (p-silicon), V_s is the source potential (V_s = 0 V), V_{dep} is the potential change within the depleted zone. In order to calculate V_{dep} as a function of V_g, we will model the MOS structure with an equivalent circuit: When V_{dep} is increased, the amount of negative charge in the depleted zone Q_{dep} is also increased. Therefore the depleted region behaves as a capacitance C_{dep}. Since equal amount of positive charge gets collected at the gate electrode, the capacitance C_{dep} and the gate capacitance C_{ox} form a serial connection. C_{ox} is defined as Q_{gate}/V_{ox} (V_{ox} is the voltage across the gate oxide). It holds:

$$C_{ox} = \varepsilon_0 \varepsilon_{SiO2} \frac{A}{t_{ox}} \sim 8.854\ 10^{-12}\ \frac{As}{Vm} \times 3.9 \times \frac{A}{t_{ox}} \qquad (3)$$

A is the gate area and t_{ox} is the gate thickness, e.g. for a 65 nm technology it is t_{ox} = 2.6 nm.

C_{dep} can be either defined as normal capacitance Q_{dep}/V_{dep} or as dynamic capacitance dQ_{dep}/dV_{dep}. In both cases C_{dep} depends on V_{dep}.

If we define C_{dep} as dynamic capacitance, it holds:

$$C_{dep} \equiv \frac{dQ_{dep}}{dV_{dep}} = \varepsilon_0 \varepsilon_{Si} \frac{A}{t_{dep}} \quad (4)$$

Q_{dep} is the charge in the depleted region, V_{dep} is the potential change in the depleted region and t_{dep} the thickness of the depleted region, $\varepsilon_{Si} \sim 12$. Interestingly, the same formula holds as for normal capacitance.

(For most of calculations it makes sense to define C_{dep} as dynamic capacitance.)

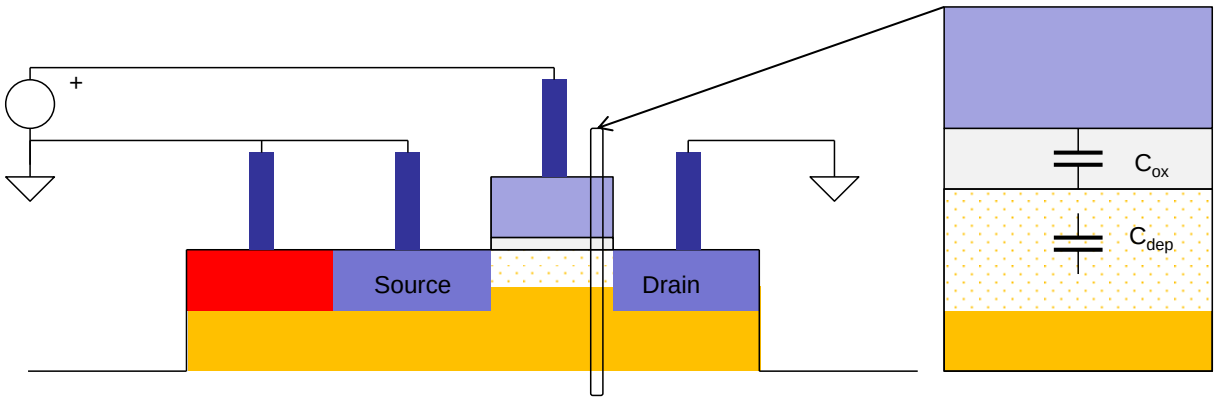
Since t_{dep} increases as function of V_{dep}, C_{dep} is not always the same. It is therefore useful to define the C_{dep} value for the maximum thickness of the depleted region t_{depmax} that we obtain for $V_{dep} = V_{cont}$:

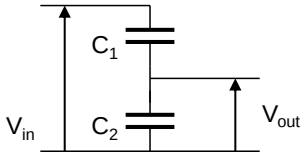
$$C_{dep,min} = \varepsilon_0 \varepsilon_{Si} \frac{A}{t_{dep,max}} \qquad (5)$$

The doping and the oxide thickness are usually chosen in such a way that the value of C_{ox} is approximately $4 \times C_{dep}$.

We also define a factor n called “slope factor” as:

$n = (C_{dep,min} + C_{ox})/C_{ox} = 1.25$ (6)





$$V_{out} = V_{in} \frac{C_1}{C_1 + C_2}$$

Fig 26: Voltage divider

To summarize: We have two capacities 1) the oxide capacity C_{ox}, which is determined by the oxide thickness and 2) the capacity of the depletion zone C_{dep}, which is determined by the depth of the depletion zone. It is useful to define this capacity as dynamic capacity dQ_{dep}/dV_{dep}

Let us now calculate the potential at Si-SiO_2 interface labelled as V_x.

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

$$V_x = V_{sub} + (V_g - V_{sub}) \times \frac{C_{ox}}{C_{ox}+C_{dep}}.$$

For simplicity we will assume that C_{dep} is constant capacitance. A good approximation for calculation of V_x is

$C_{dep} = 2\, C_{dep,min}$ (8)

Because it holds $Q_{dep} \approx 2\, C_{dep,min}\, V_{dep}$.

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:

$$\frac{dE_z}{dz} = -\frac{eN_a}{\varepsilon_0 \varepsilon_{Si}} \Rightarrow E_z = -\frac{eN_a}{\varepsilon_0 \varepsilon_{Si}} z \qquad (9)$$

Potential:

$$-\frac{dV_z}{dz} = E_z \Rightarrow V_z = \frac{eN_a}{\varepsilon_0 \varepsilon_{Si}} \frac{z^2}{2} \quad (10)$$

It follows:

$$t_{dep} = \sqrt{\frac{2\varepsilon_0 \varepsilon_{Si} V_{dep}}{eN_a}} \qquad (11)$$

The charge in the depleted zone is:

$$Q_{dep} = AeN_a t_{dep} = A\sqrt{eN_a 2\varepsilon_0 \varepsilon_{Si} V_{dep}} \qquad (12)$$

The dynamic capacitance of the depleted zone is:

$$C_{dep} \equiv \frac{dQ_{dep}}{dV_{dep}} = A\sqrt{\frac{eN_a \varepsilon_0 \varepsilon_{Si}}{2V_{dep}}} \qquad (13)$$

It holds because of (13) and (11):

$$C_{dep} = A\frac{\varepsilon_0 \varepsilon_{Si}}{t_{dep}} \quad (14)$$

It holds also:

$$Q_{dep} = A\sqrt{eN_a 2\varepsilon_0 \varepsilon_{Si} V_{dep}} = V_{dep} A\sqrt{\frac{2\varepsilon_0 \varepsilon_{Si} eN_a}{V_{dep}}} = 2C_{dep} V_{dep} \qquad (15)$$

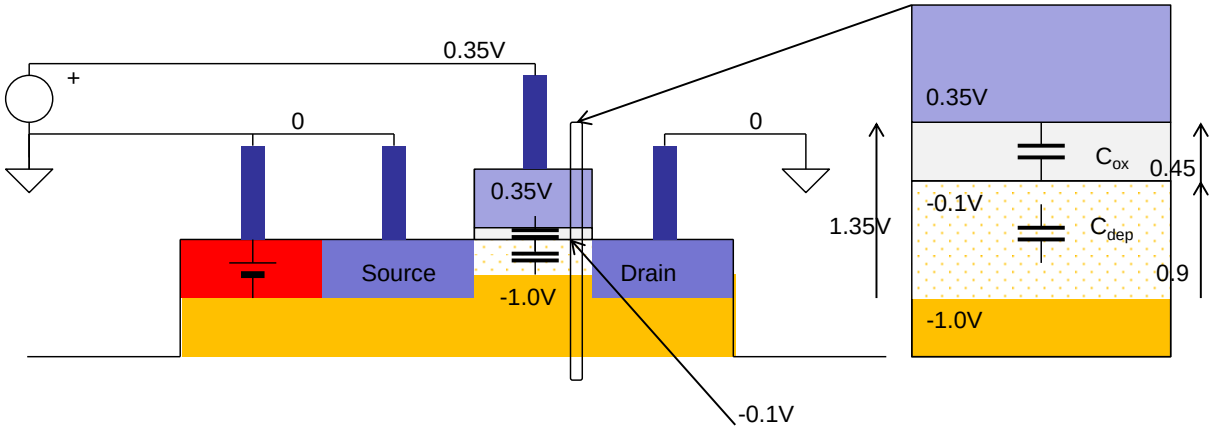


Fig 27: Potential at Si-SiO₂ interface for V_g = 0.35V

Example: We set V_g = 0.35 V. This value is interesting because it is closely below the threshold voltage. In the depletion zone, the potential rises from -1.0V to about -0.1V (by 0.9V). In the oxide the potential increases by about 0.45V. If the gate is at 0.35V, the potential at the silicon-SiO_2 interface is -0.1V.

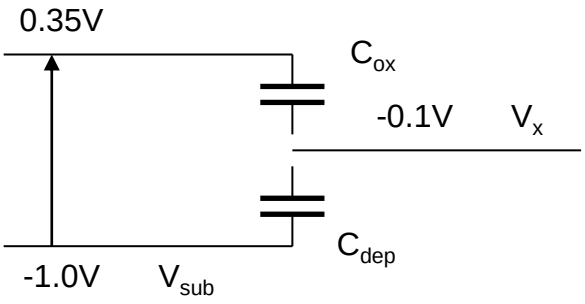


Fig 28: Capacitive voltage divider

A positive potential is a barrier for positive charge (holes) a negative potential is the barrier for electrons. From this follows: an electron can hardly pass the barrier from source to drain, through the deeper P-silicon regions, as 1.0V represents a large potential barrier for electrons, Figure 29.

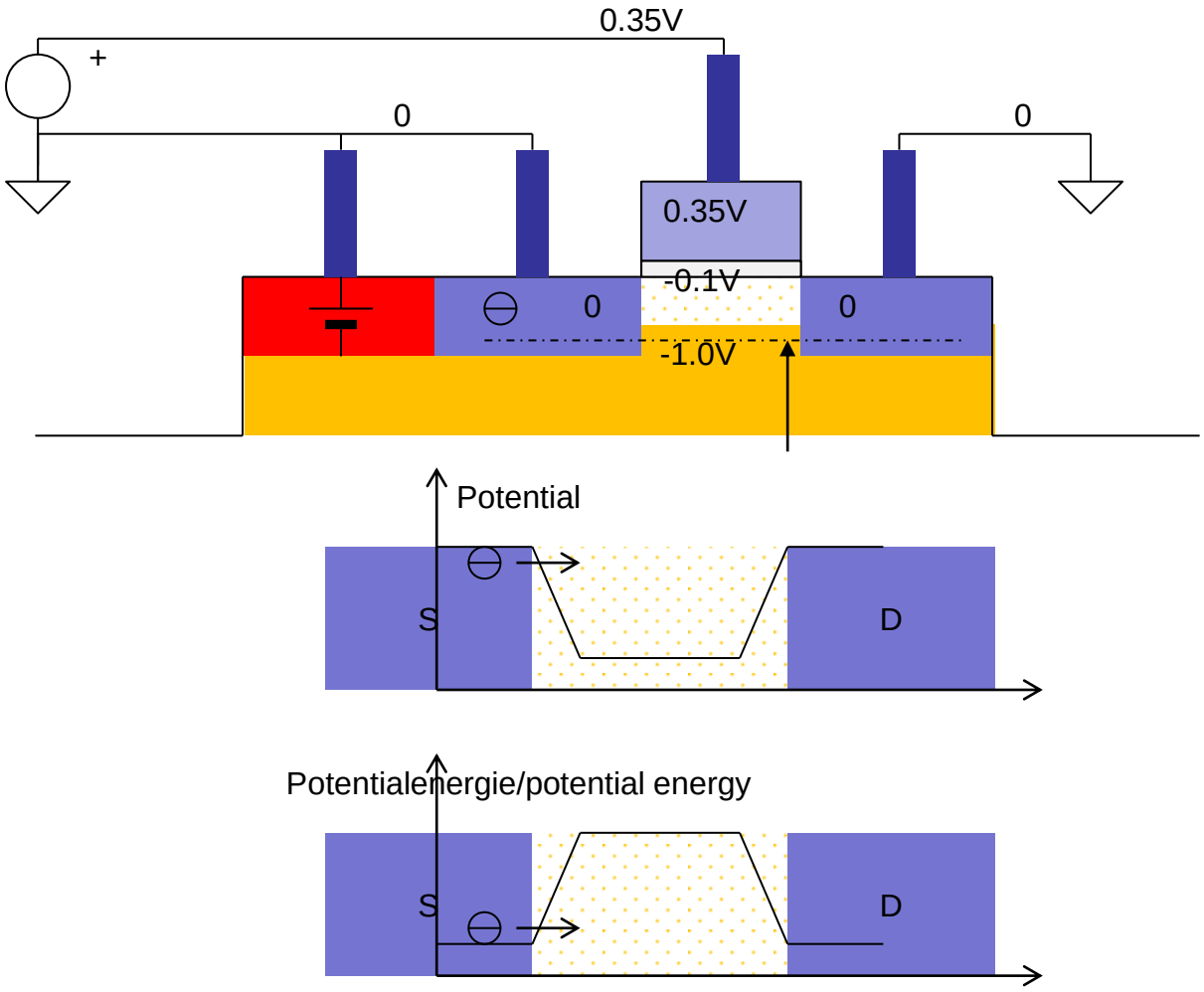


Fig 29: Potential barriers

Unlike the deeper substrate regions, the potential on the substrate surface is only 0.1V lower than in the source and drain. The barrier is smaller and a diffusion current can flow. We will explain this subthreshold current in the next lecture.

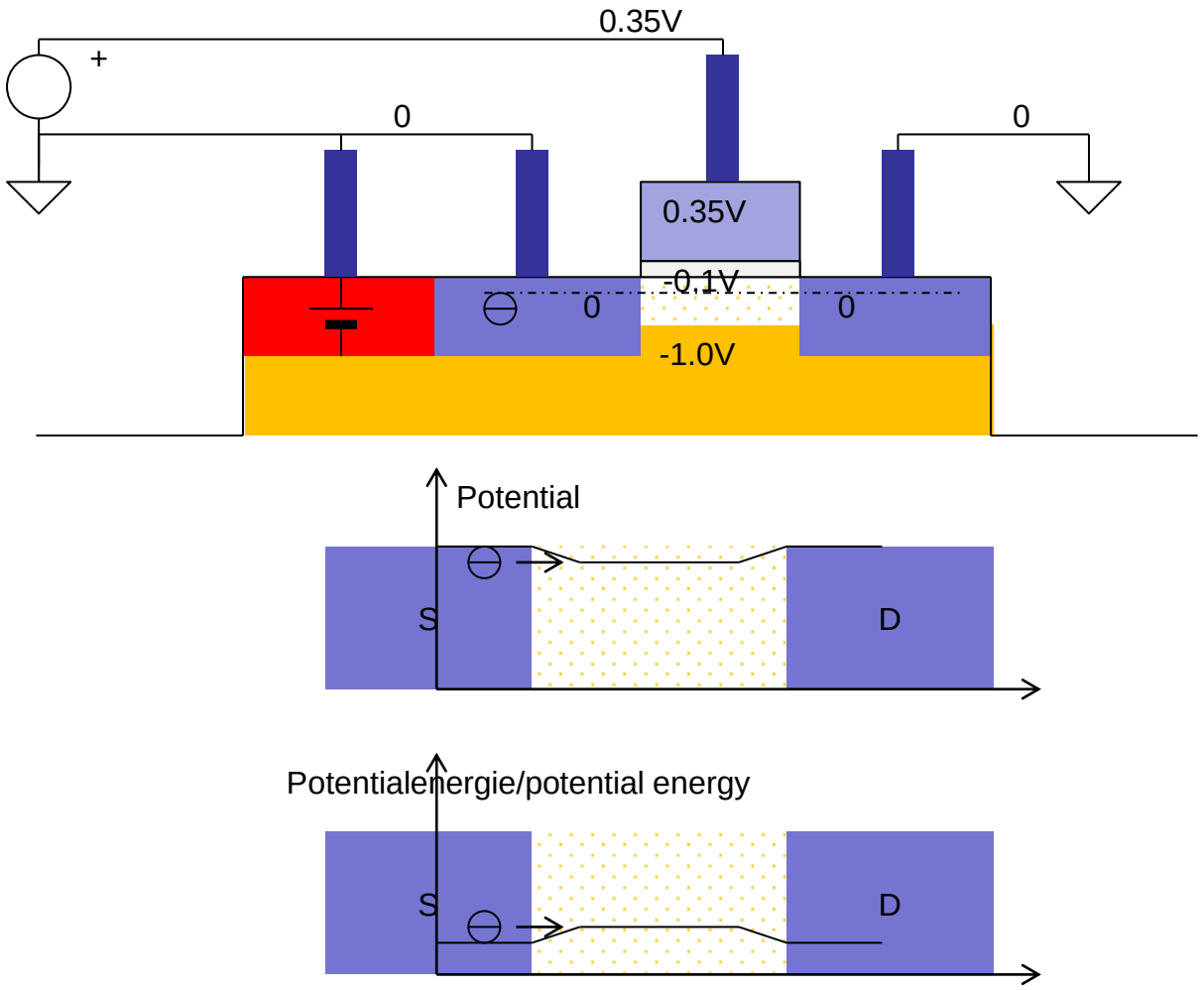


Fig 30: Smaller potential barrier near SiO_2

Let us now increase the gate potential to 0.5 V.

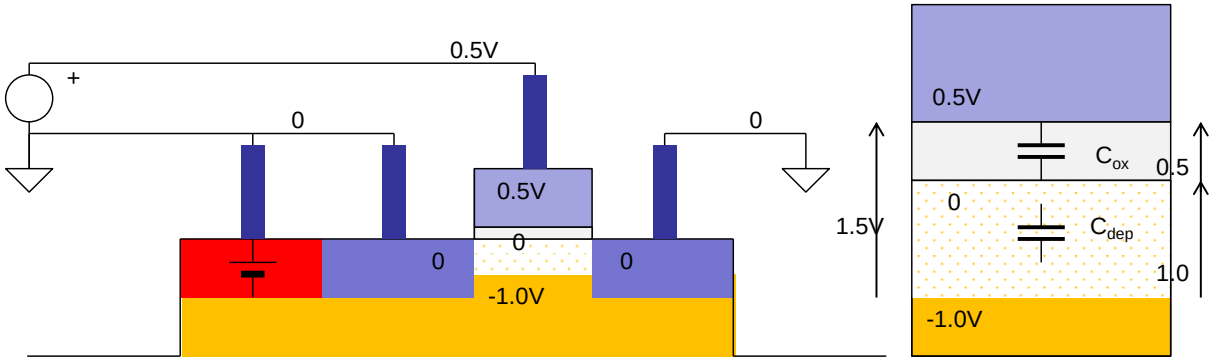


Fig 31: Threshold voltage

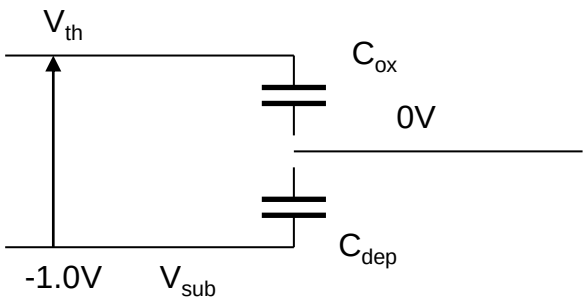


Fig 32: Voltage divider

The potential at the Si-SiO_2 interface (V_x) is 0 V, which is exactly the same as in the source and drain.

We define the threshold voltage V_{th} as the gate source voltage for which following holds $V_x = V_s = V_d$.

The formula for voltage divider leads to the following result:

$$V_{th} = -\frac{C_{dep}}{C_{ox}} \times V_{sub} = \frac{C_{dep}}{C_{ox}} \times V_{cont} = \frac{2C_{dep,min}}{C_{ox}} \times V_{cont} \quad (16)$$

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

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

It holds (Equation 13):

$$C_{dep,min} = A\sqrt{\frac{eN_a\varepsilon_0\varepsilon_{Si}}{2V_{cont}}} \quad (17)$$

Therefore it is:

$$V_{th} = \frac{2C_{dep,min}}{C_{ox}} \times V_{cont} = \frac{A\sqrt{2eN_a\varepsilon_0\varepsilon_{Si}V_{cont}}}{C_{ox}} \quad (18)$$

The contact voltage is given by (1). Since we assume $N_d = N_a$, it holds:

$$V_{cont} = 2U_T \ln\left(\frac{N_a}{n_i}\right) \quad (19)$$

Equation (18) is usually given in the literature.

Smaller thresholds are better if we want to minimize switching currents (AC power) in digital electronics. That is why one tries to make the thickness of the oxide as small as possible.

With temperature rise, the threshold voltage decreases, as the contact potential also decreases. This is the result of the increase of the intrinsic carrier density in silicon and electron density in P-silicon.

If we increase the gate potential (gate source voltage) over about 0.5 V, 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 33).

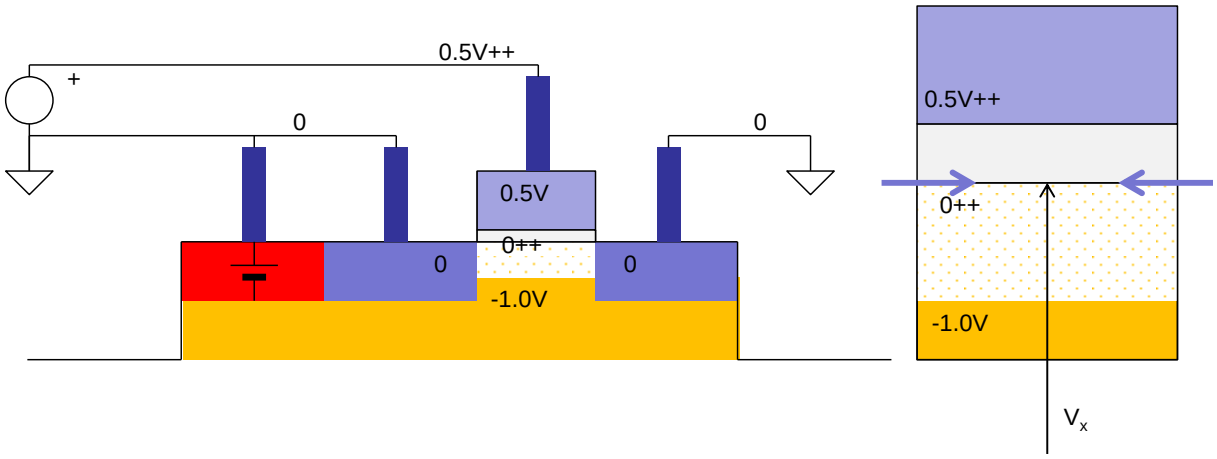


Fig 33: Case $V_g > V_{th}$

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 34. We refer to this operation region strong inversion.

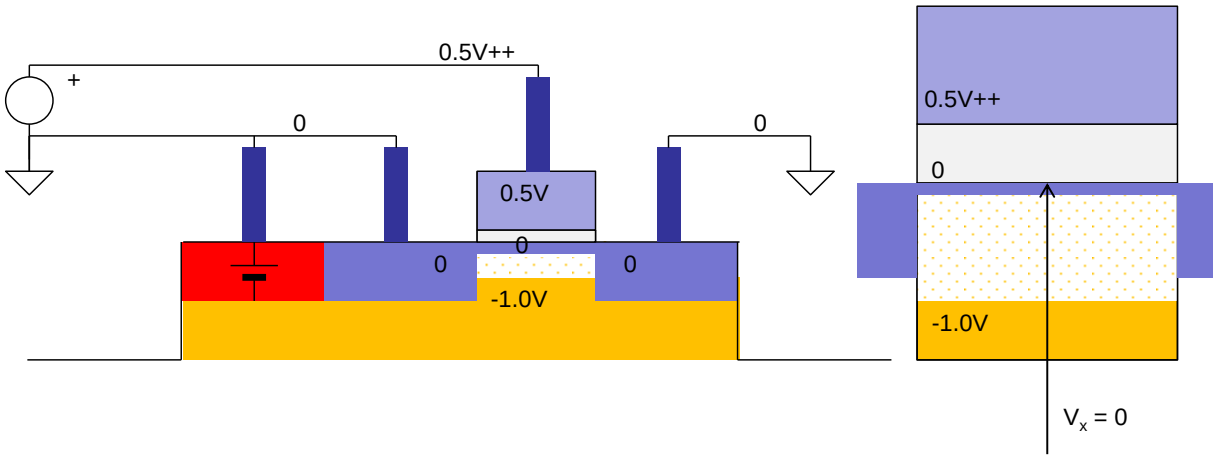


Fig 34: Electrons form the channel. Source and drain are shorted -> $V_x = 0$

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 capacity C_{ox}. When the gate voltage changes by dV_g, the charge amount $C_{ox}\, dV_g$ flows through the voltage source. This charge is formed in the channel.

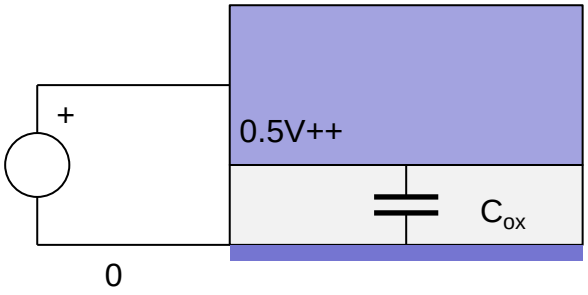


Fig 35: Channel charge

For $V_g = 0.5V$ we have no charge in the channel. For $V_g > V_{th}$ it holds $dQ = C_{ox}\, dV_g$. Therefore it holds

$$Q = C_{ox}(V_g - V_{th})$$

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

$Q = C_{ox}(V_{gs} - V_{th})$ (20)

Summary:

We define the threshold voltage V_{th} ~0.5 V as the gate source voltage for which the potentials in the source and on the substrate surface are approximately equal.

(For $V_{gs} = V_{th}$, the potential barrier between source and drain 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.

Drain-source current

Let us calculate the transistor current for small voltages V_{ds}

Assume that we have a channel of electrons between the source and the drain (Fig. 36).

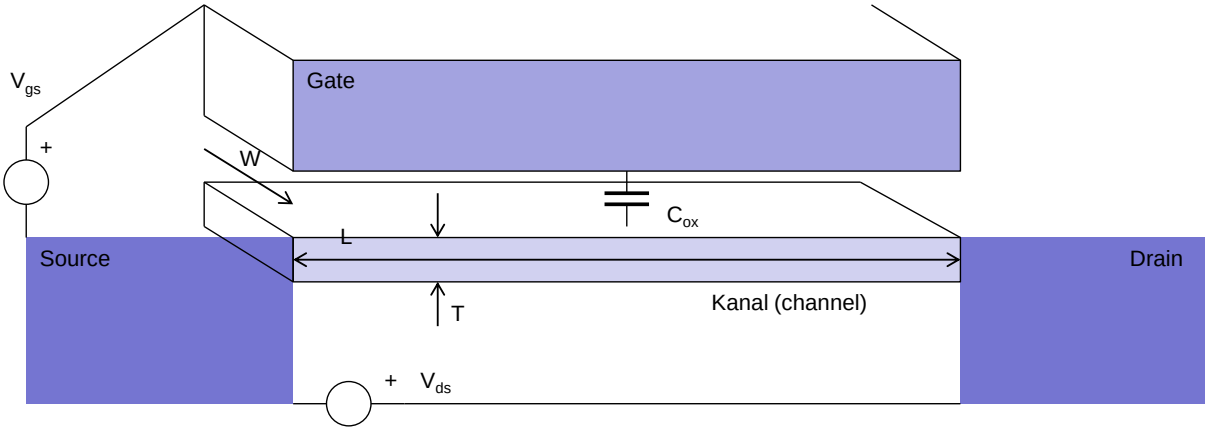


Fig 36: Channel is an ohmic connection between source and drain

If we have a voltage between drain and source (V_{ds}), a current flows from drain to source (I_{ds}).

The channel forms a resistance, the current is given by the following equation (Fig 36):

$I_{ds} = A\, e\, \mu\, n\, E$ (21),

e is the elementary charge, μ is mobility, n is electron density in the channel, E is the horizontal E-field component, A is the channel cross section, which is equal to channel width W multiplied by channel thickness t.

It follows:

$$I_{ds} = W\ t\ e\ \mu\ n\ E$$

It can be derived:

$$n\ t\ e = \frac{Q}{WL} = \frac{C_{ox}\,(V_{gs} - V_{th})}{WL}$$

L is the length of the channel. (We used the result Q = Cox (V_{gs} - V_{th}))

We get:

$$I_{ds} = \mu\ C'_{ox}\ W\ (V_{gs} - V_{th})\ E$$

C'$_{ox}$ is the capacitance per unit area.

E-field is nearly:

E = V_{ds}/L.

Therefore:

$$I_{ds} = \mu C'_{ox} \frac{W}{L} (V_{gs} - V_{th}) V_{ds} \qquad (22).$$

This is the simplest equation for the transistor current.

It is valid for small V_{ds}. The assumption was that the charge is uniformly distributed in the channel.

We see that the current depends on the W/L ratio. This is typical for MOSFETs. In contrast to that, in the case of bipolar transistors, the size does not influence current.

Saturation

How does the current increase when the V_{ds} become larger?

The charge in the channel is expressed with the formula (20)

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

This formula holds when V_{ds} is zero or small.

What happens if we have V_{ds} >> 0?

We can expect following: The channel charge near source is Cox (V_{gs} -V_{th}) and the channel charge near drain is C_{ox} (V_{gd} -V_{th}). This reflects the symmetry of the structure. Formula C_{ox} (V_{gd} -V_{th}) is not fully true because it does not take into account that V_d is different than 0. We will discuss this in the next lecture.

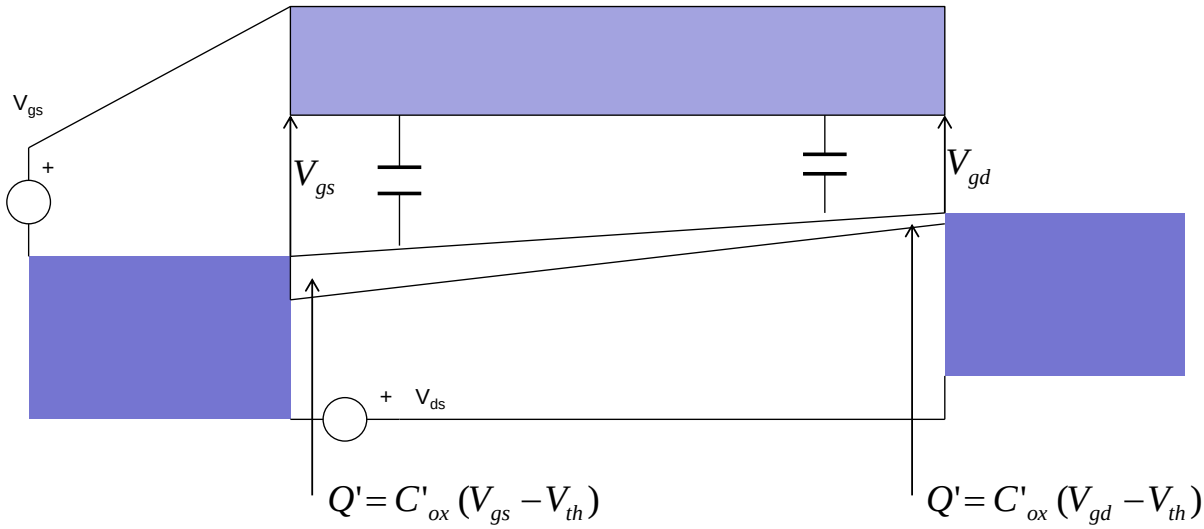


Fig 37: Current as function of V_{ds}. $V_{ds} > 0$

Therefore for a drain voltage $V_d = V_g - V_{th}$ ($V_{gd} - V_{th} = 0$) we have no channel on the drain side. We say that the channel is pinched-off.

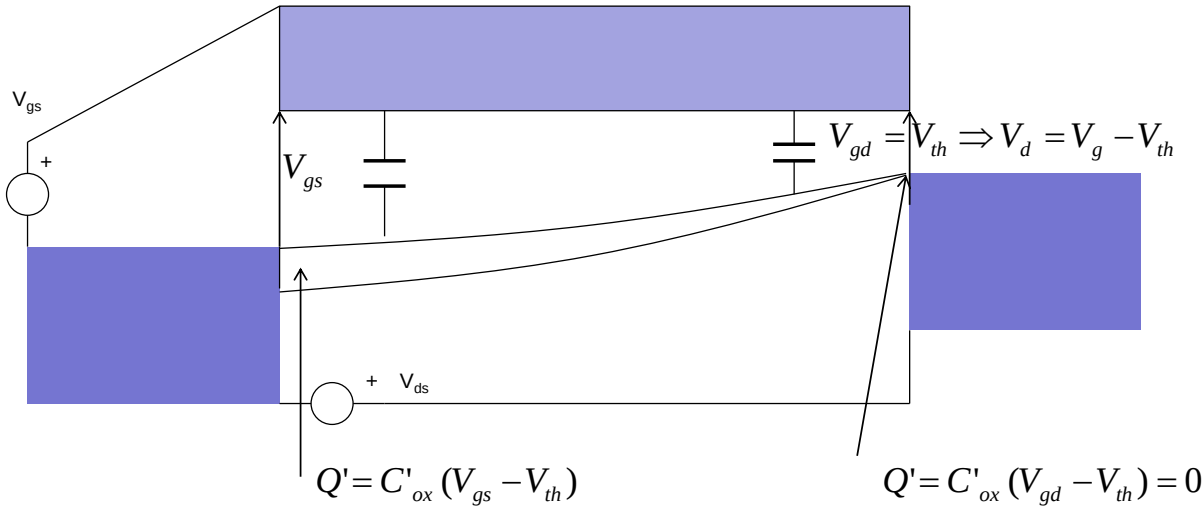


Fig 38: Current as function of V_{ds}. $V_{ds} = V_{gs} - V_{th}$

A further increase in current is inhibited. We have a current saturation. In a first approximation the current will not increase if we continue to increase V_{ds} beyond $V_{gs} - V_{th}$.

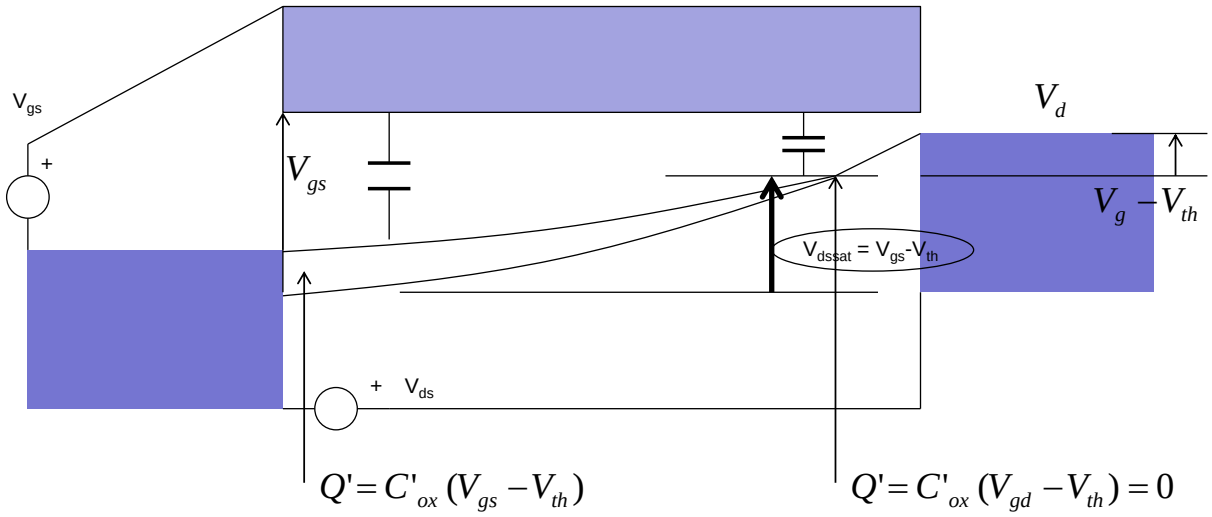


Fig 39: Current as function of V_{ds}. $V_{ds} > V_{gs} - V_{th}$

The condition for the beginning of saturation: $V_{gd} = V_{th}$ can be also expressed as $V_{ds} = V_{gs} - V_{th}$.

We define saturation voltage as

$$V_{dssat} = V_{gs} - V_{th} \qquad (23)$$

For higher V_{ds} the current increases just a little bit because of channel length modulation effect as will be shown in the next lecture.

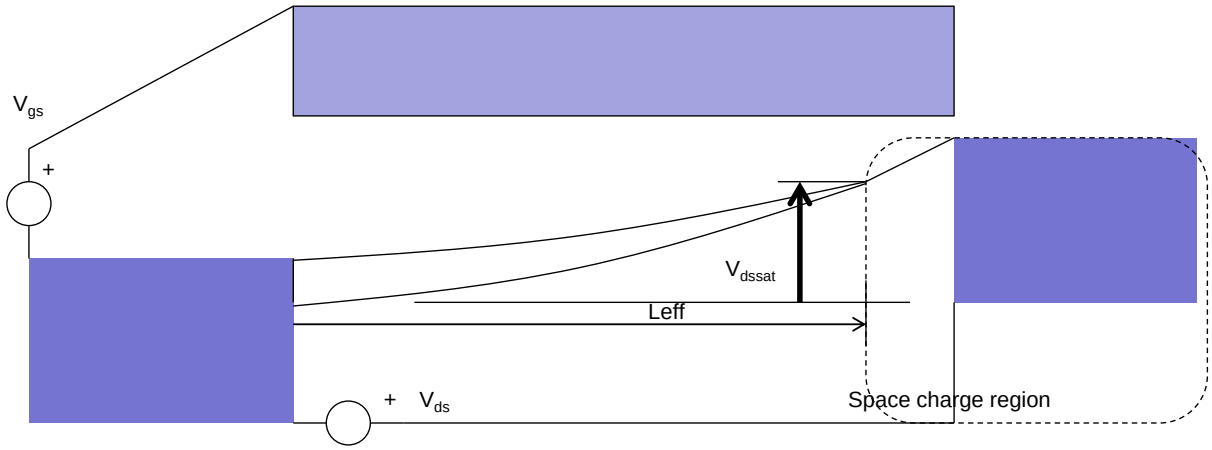


Fig 40: Channel length gets smaller

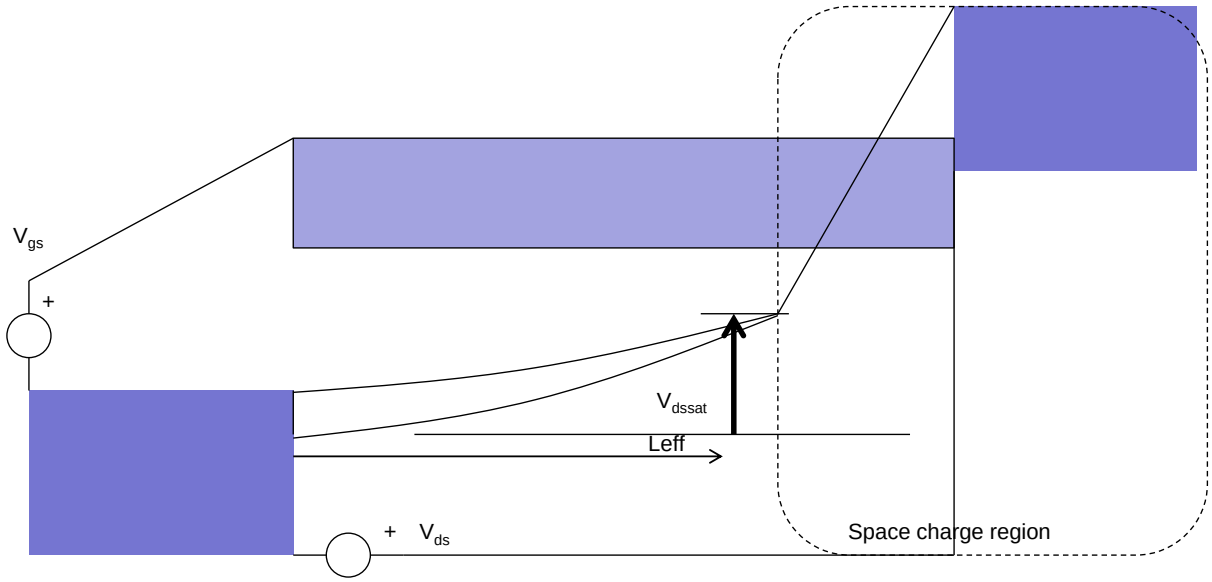


Fig 41: Channel length gets smaller

What is the value of the drain source current for $V_{ds} = V_{dssat}$? (at the beginning of saturation)

Let us make a not so correct assumption (A1): the formula (22), which we have derived for small V_{ds}, applies from $V_{ds} = 0$ to $V_{ds} = V_{dssat}$.

The drain source current for V_{dssat} voltage (saturation current) can be calculated from the equation (22) by using $V_{ds} = V_{gs} - V_{th}$:

$$I_{dssat} = \mu C'_{ox} \frac{W}{L} (V_{gs} - V_{th})^2 \qquad (24)$$

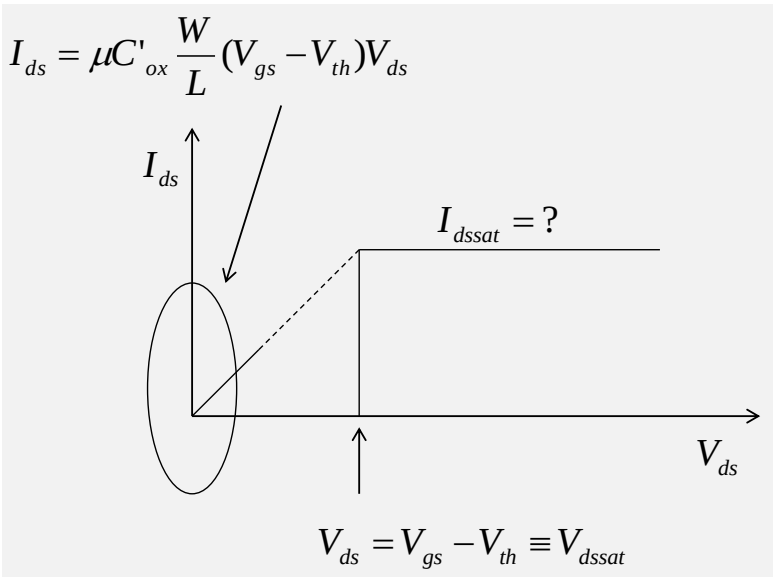


Fig 42: I_{ds} formula – simple derivation

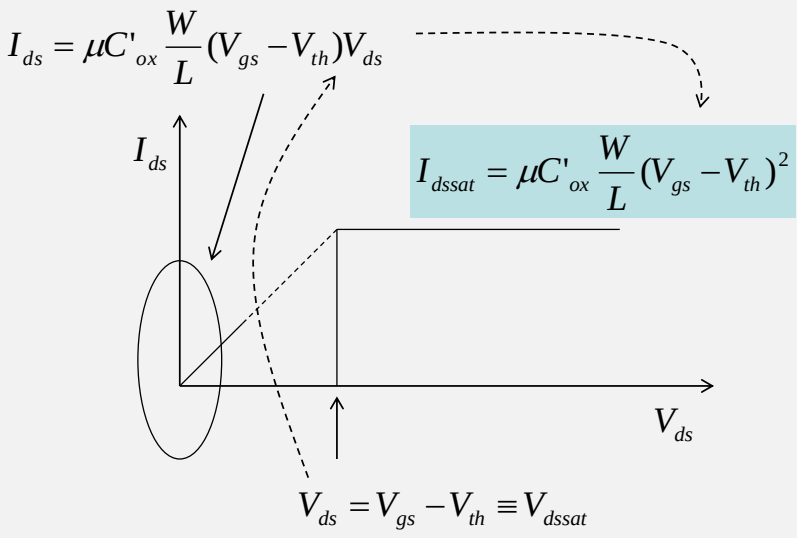


Fig 43: I_{ds} formula – simple derivation

Unfortunately, the assumption (A1) is not quite correct: the current increase is smaller Vds > ~100mV than expected from formula (22).

More precise calculation leads to an additional factor 1/2. The formula for saturation current is as follows:

$$I_{dssat} = \frac{1}{2}\mu C'_{ox}\frac{W}{L}(V_{gs} - V_{th})^2 \quad (25)$$

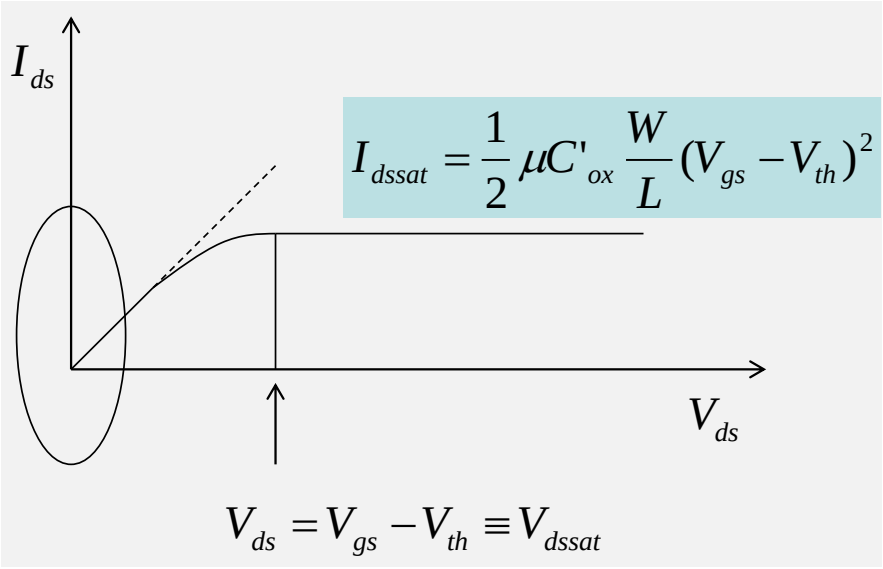


Fig 44*: I_{ds} more precise formula*

This formula is also not very correct. Even more precise models show that the factor ½ has to be replaced by 1/(2nα). For V_{dssat} holds:

$$V_{dssat} = \frac{V_{gs} - V_{th}}{n\alpha} \quad (26)$$

$$\alpha = 1 + \frac{V_{gs}}{nE_{sat}L} \quad (27)$$

Factor α is for long transistors (transistors with L > 1µm) ~1. Factor n is the slope factor

$n = (C_{dep,min} + C_{ox})/C_{ox} \sim 1.25$ (28)

For very short transistors or for large gate source voltages α is significantly larger than 1 and leads to much lower I_{dssat} values than expected. It is an effect of the saturation of mobility. In general, smaller transistors need more complex formulas.

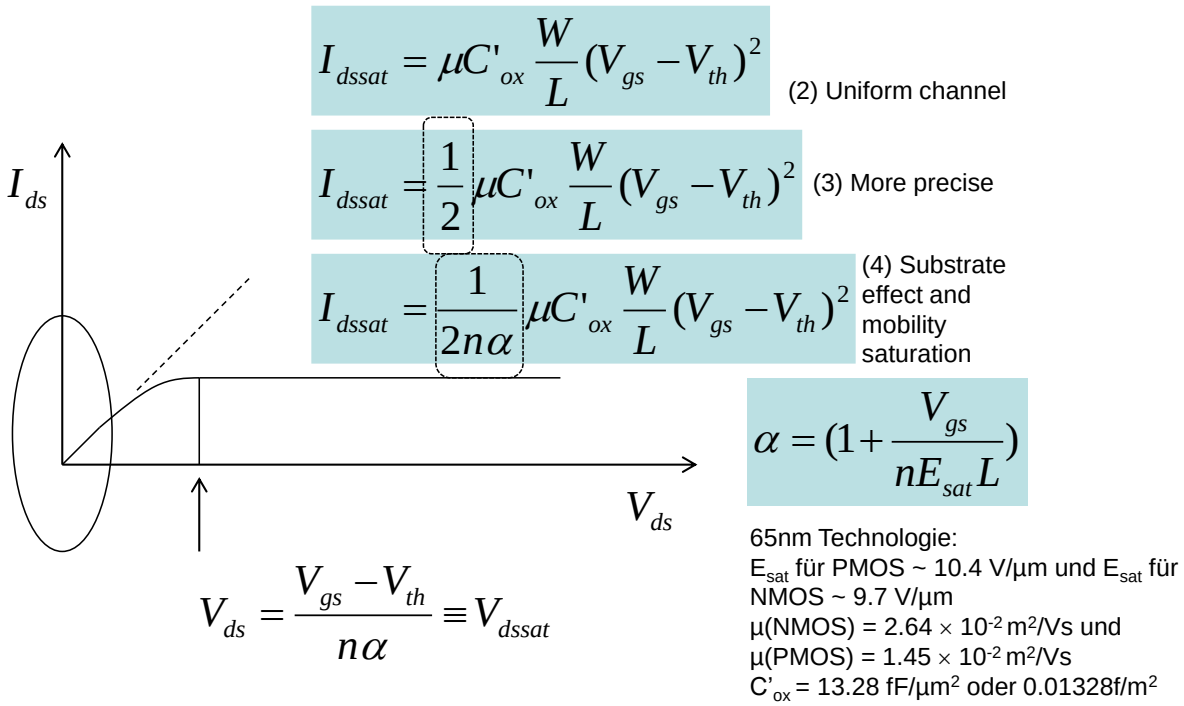


Fig 45: I_{ds} formulas – overview