Metal oxide semiconductor field effect transistors (MOSFETs)

A newer form of transistor which has pretty much replaced BJT technology for all digital applications and much of analog.

Basically, an FET is a resistor whose resistance can be controlled through a third terminal. So the transistor mechanism is much different than that of the BJT. A BJT used *both* electrons and holes injected across the junctions (hence bipolar). An FET uses *either* electrons *or* holes (hence unipolar) flowing by drift current between two contacts (source and drain). The amount of flow is controlled by a voltage applied at the gate.

electron-current device: n-channel MOSFET (NMOS)
hole-current device: p-channel MOSFET (PMOS)

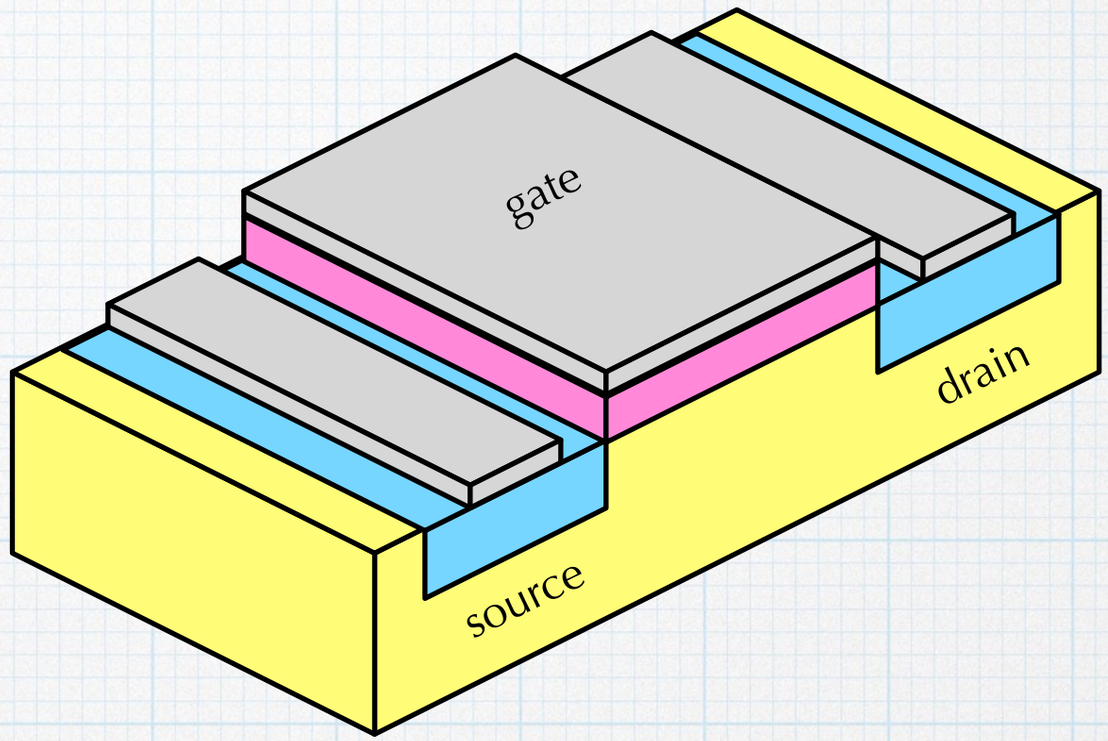
Using NMOS and PMOS together (known as CMOS) offers some significant advantages in circuit design.

Basic NMOS structure

(n-channel device)

Start with a *p*-type substrate.

Form two *n*-type regions (source & drain)



Form a thin layer of silicon dioxide on the *p*-type silicon between source and drain

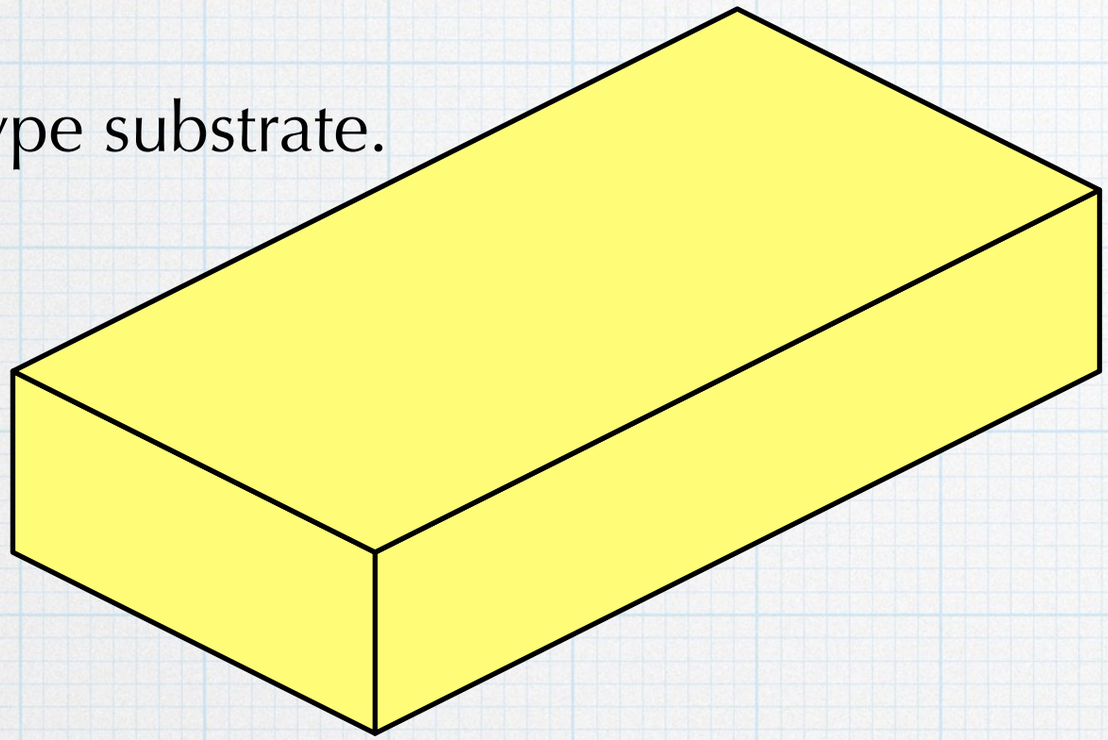
Put metal contacts on the source, drain, and oxide (gate) for electrical connection

PMOS has n- and p-type regions reversed. (Duh!)

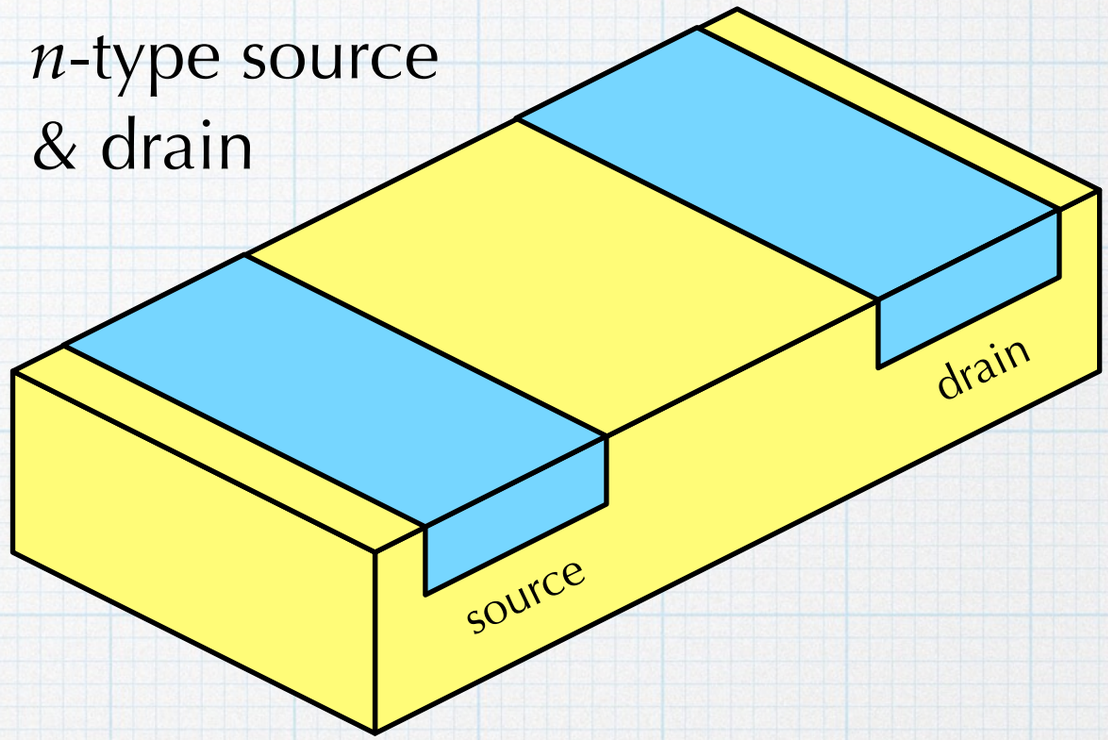
Basic NMOS structure

(n-channel device)

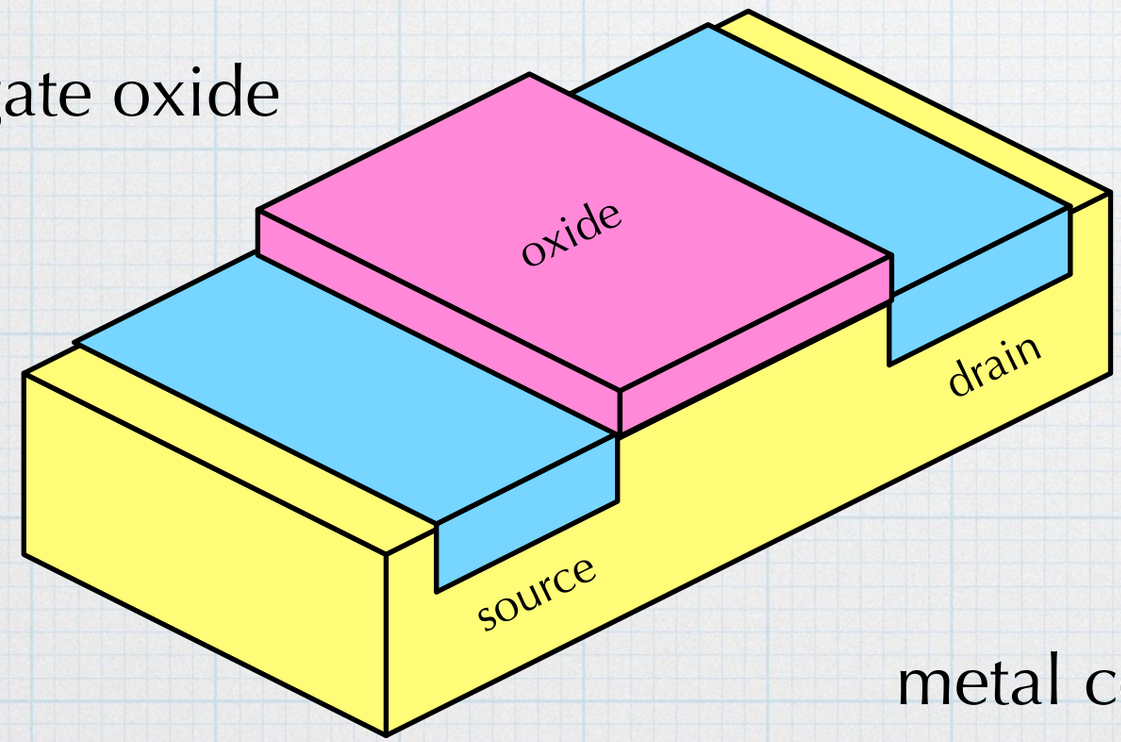
p-type substrate.

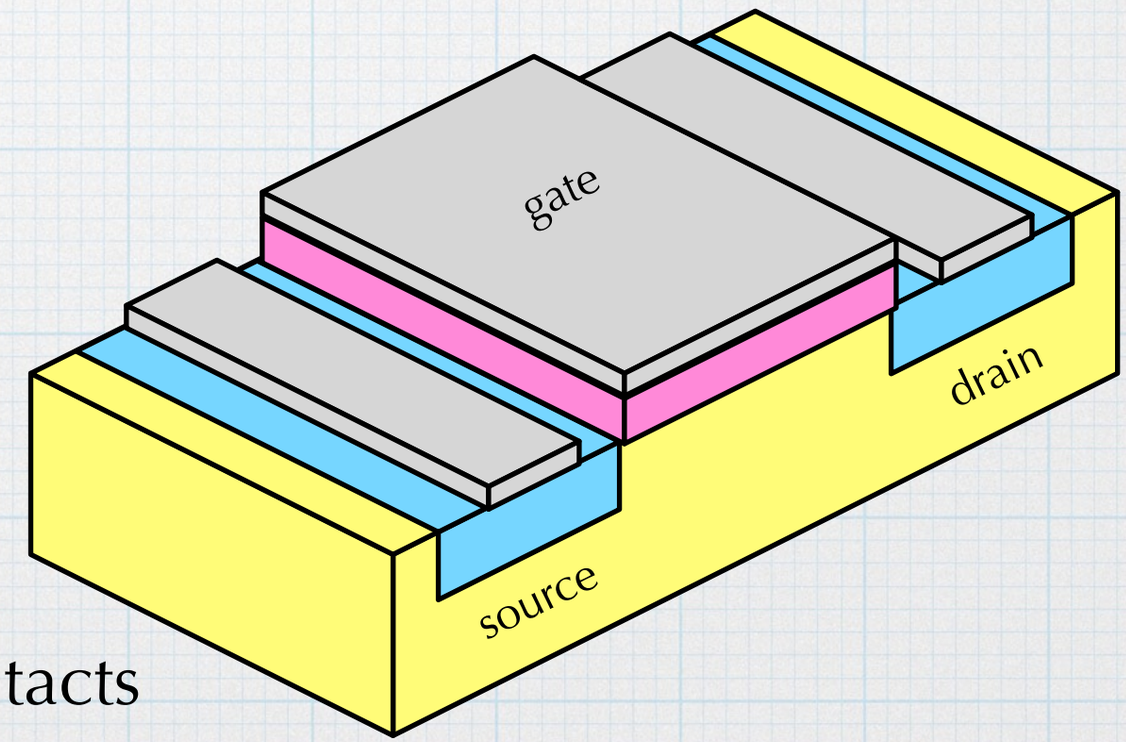


n-type source
& drain



gate oxide

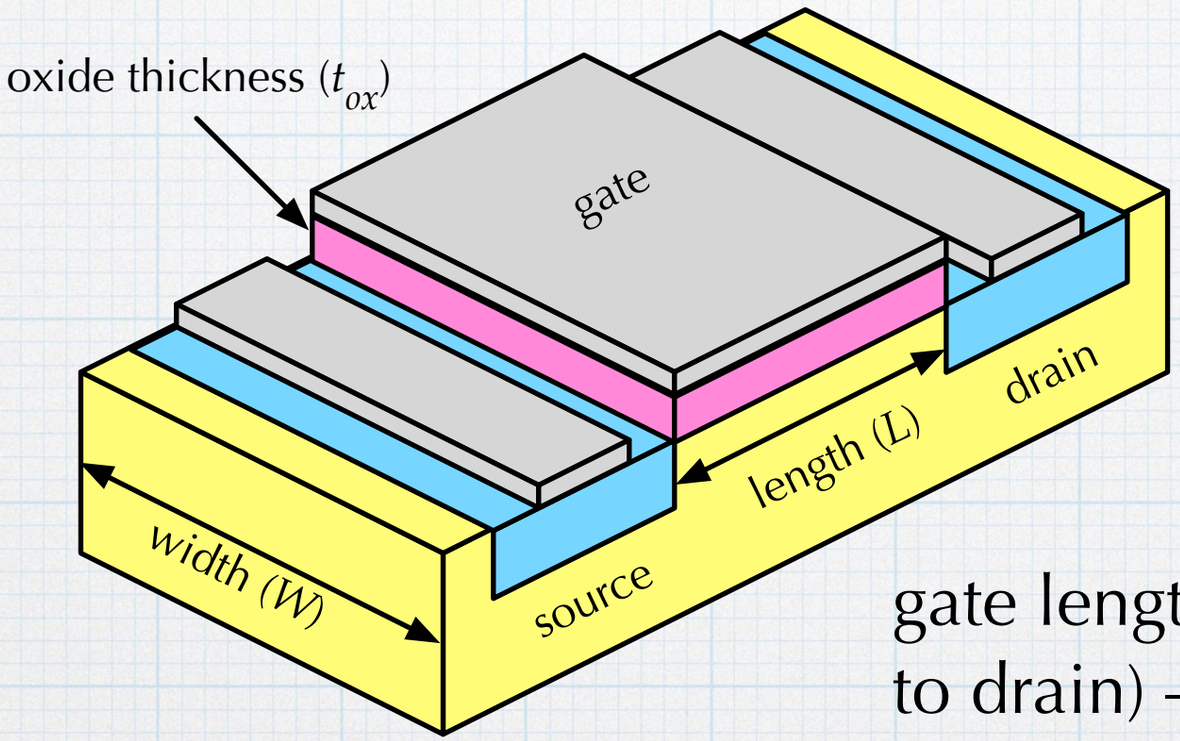




metal contacts

Critical dimensions

oxide thickness: typical 1 - 10 nm.

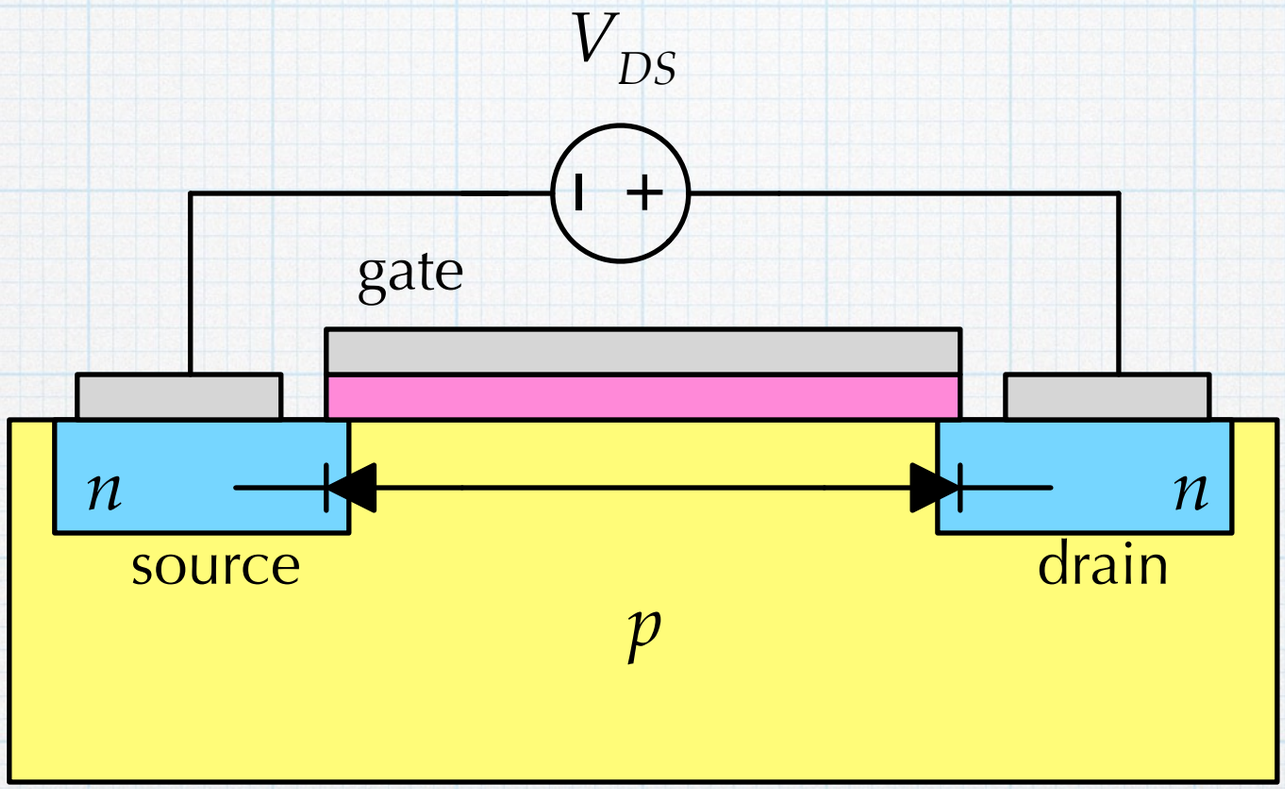


gate length (distance from source to drain) – as small as 20 nm.

width: typical L to 10 L
(W/L ratio is important)

Will current flow?

Apply a voltage between drain and source (V_{DS}).

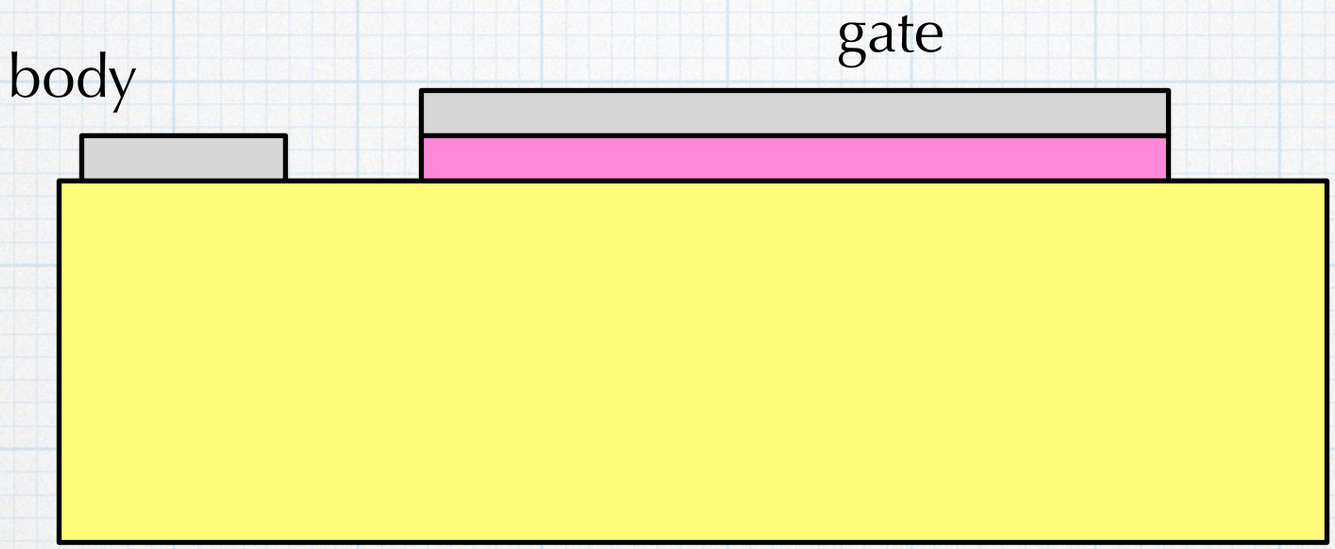


If $V_{DS} > 0$, the diode at the drain end will be reverse-biased, preventing current flow.

Reversing the polarity doesn't help, because then the other junction will be reverse biased.

The MOS capacitor

The secret to MOSFET operation lies in the MOS capacitor (the central part of the FET).



The structure is essentially a little parallel-plate capacitor formed by the gate metal and the semiconductor with the oxide as the dielectric.

$$C_{gate} = \frac{\varepsilon_{ox} A}{t_{ox}}$$

Define oxide capacitance $C_{ox} = \dfrac{\varepsilon_{ox}}{t_{ox}}$

units: $\mathrm{F/m^2}$ (or $\mathrm{F/\mu m^2}$)

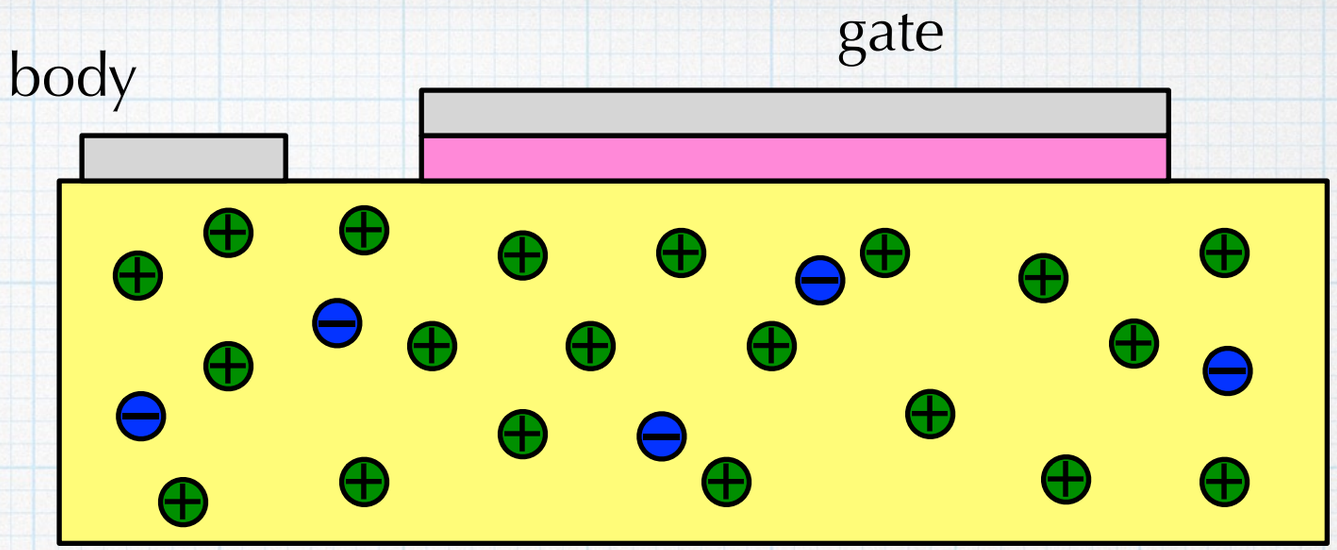
$$\varepsilon_{ox} = 3.9\varepsilon_o = 3.9(8.85\text{x}10^{-12}\ \mathrm{F/m})$$

for $t_{ox} = 10\ \mathrm{nm}\ (10^{-8}\ \mathrm{m})$

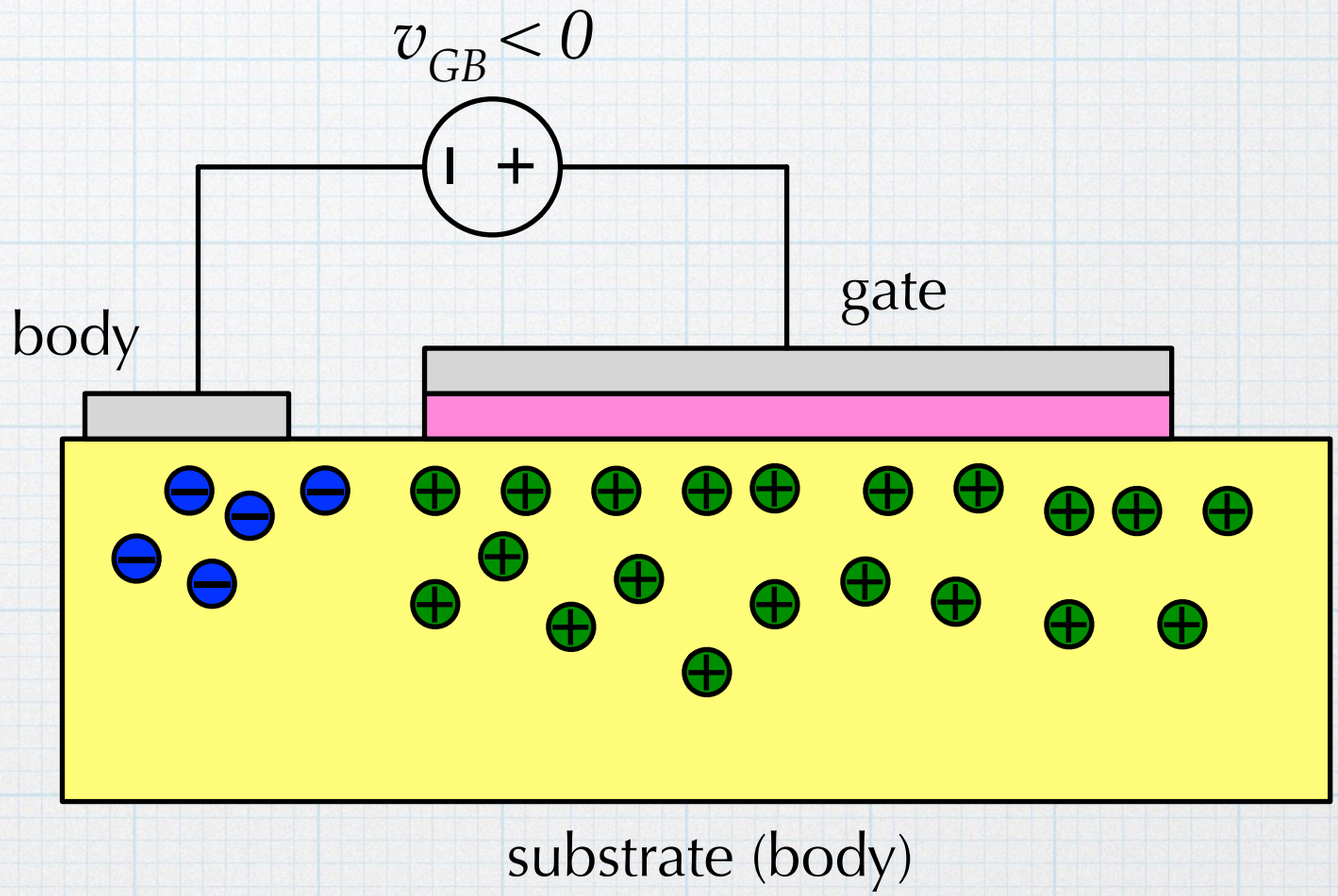
$$C_{ox} = 0.00345\ \mathrm{F/m^2}$$

$$= 3.45\text{x}10^{-15}\ \mathrm{F/\mu m^2}$$

P-type semiconductor – the carriers are mostly holes, but there are few electron lurking around.

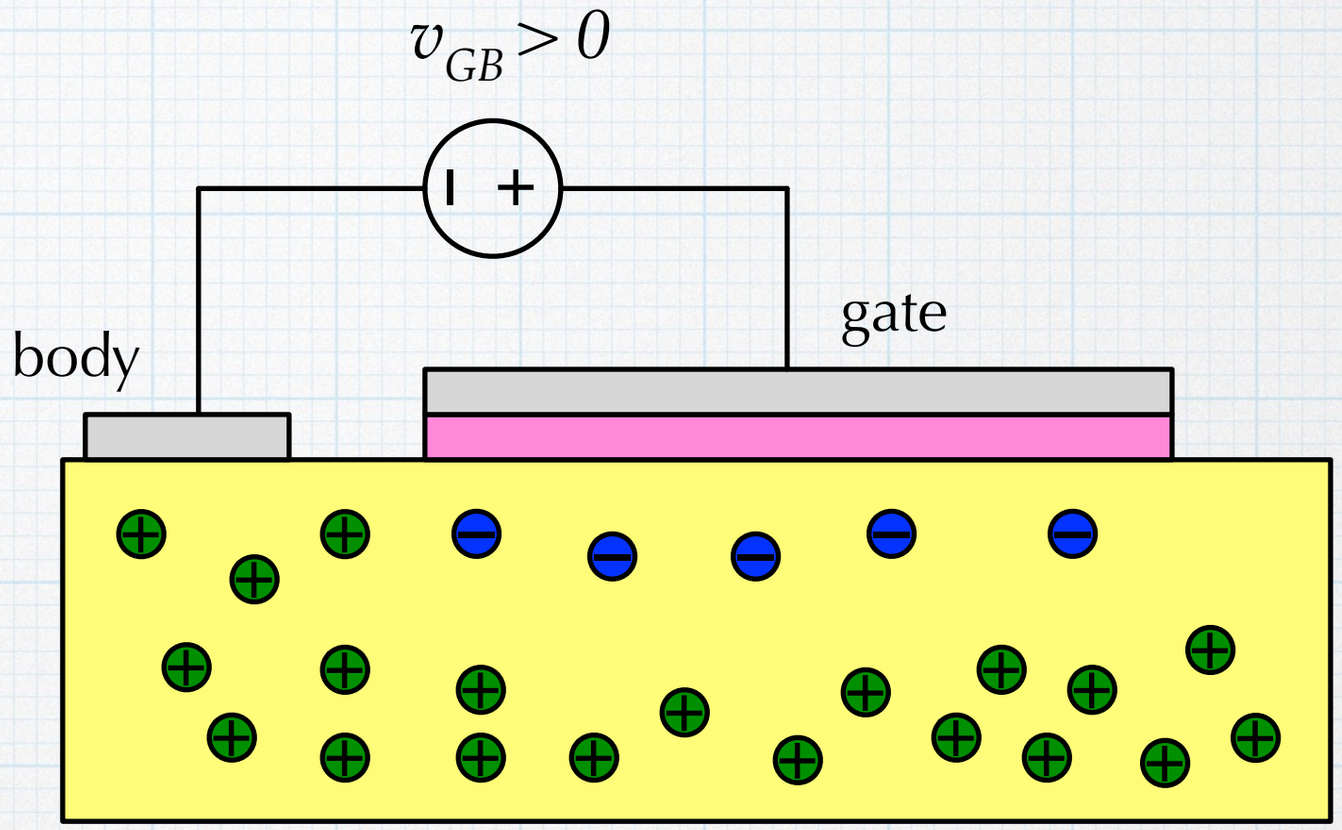


Try applying a negative voltage between the gate and the body: Holes are attracted to the gate region and electrons are pushed away from the negative voltage.



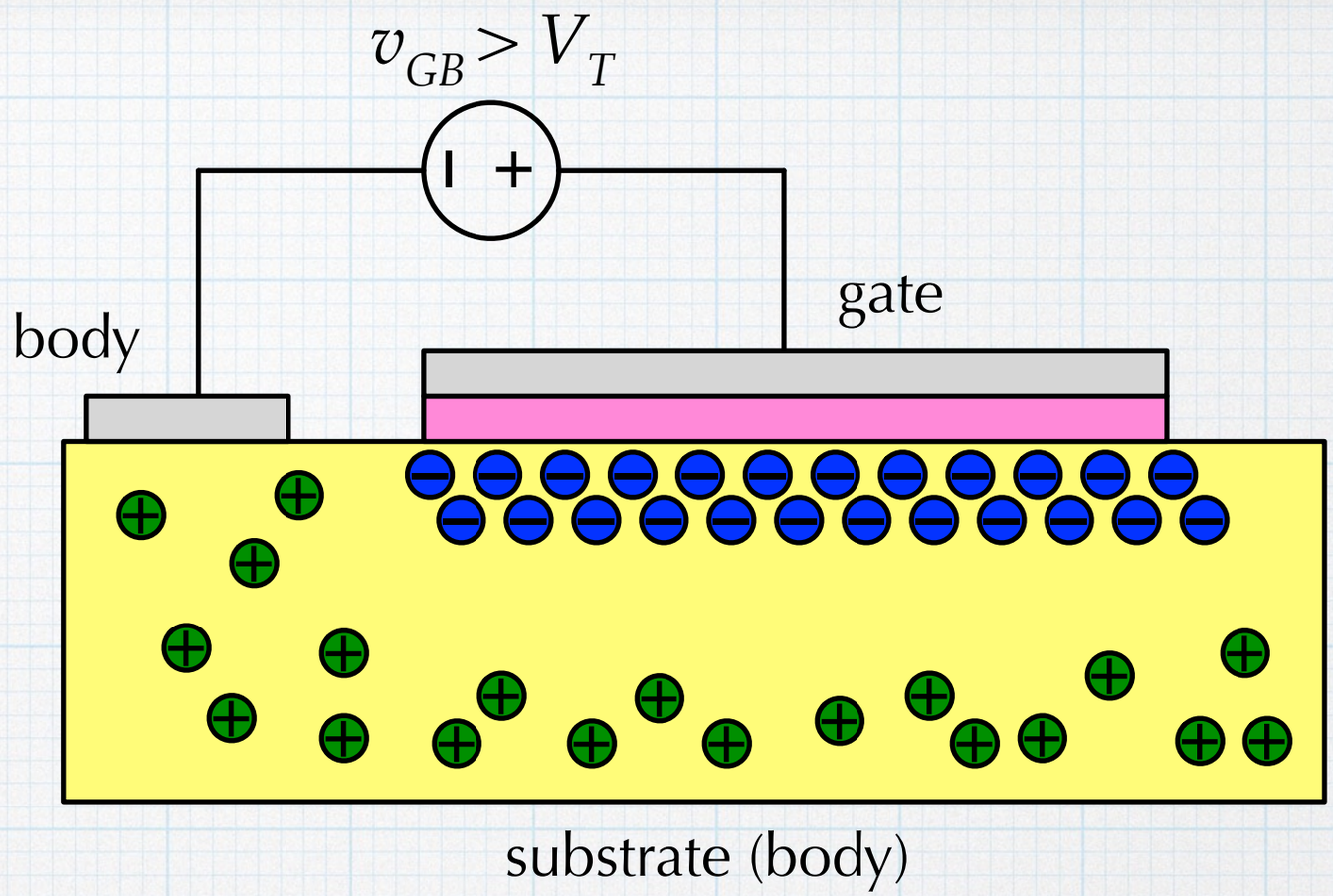
Hole concentration under the gate is enhanced. So negative voltage causes holes to "accumulate". Probably not helpful for our NMOS.

So try applying a positive voltage to the gate. Now the holes are pushed away and the electrons are attracted.



The region under the gate becomes "depleted" of holes. So this is probably not helpful either, but the movement of the electrons towards the gate looks promising.

So try increasing the gate voltage further. Holes are pushed away further and more and more electrons are pulled in.



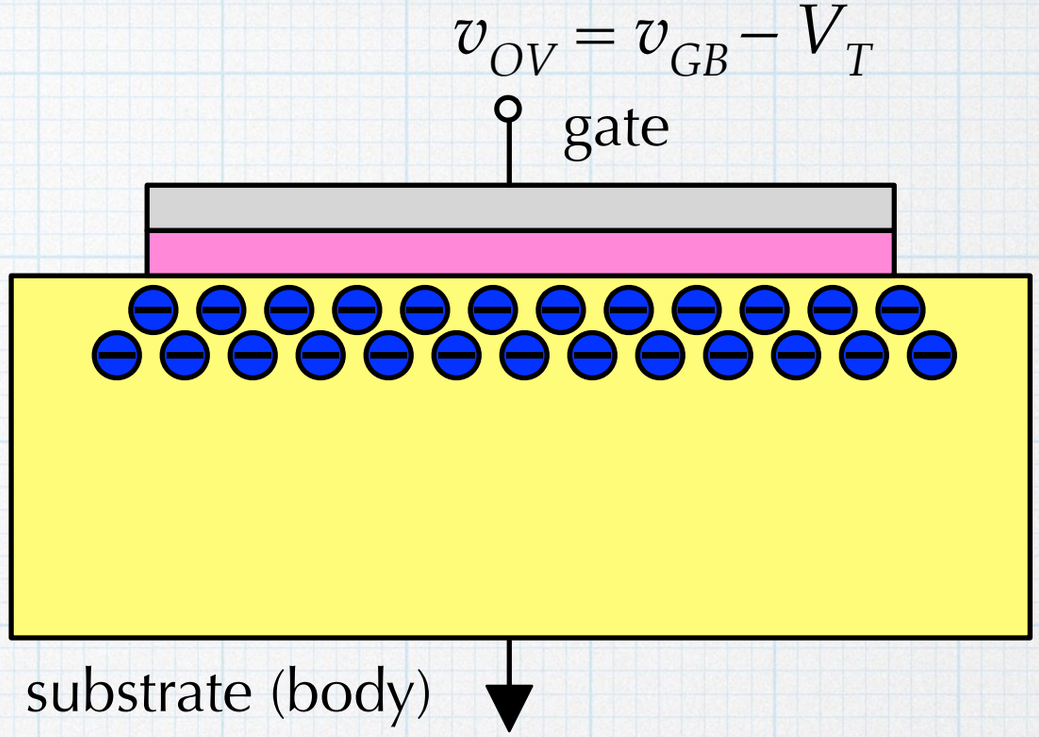
At a sufficiently high voltage, enough electrons will gather together to make the region under the gate "invert" and behave like it was n-type!

The gate voltage required create the electron "inversion layer" is called the threshold voltage, V_T. If the voltage is increase beyond V_T, even more electrons are gathered together under the gate.

The inversion layer

How many electrons are in the inversion layer? The definition of the threshold voltage says that when $v_{GB} = V_T$, the electron concentration in the inversion layer is equal to the hole concentration in the substrate:

$$n(\text{inversion}) = p(\text{substrate}).$$



This gives just enough electrons so that the region under the gate begins to act like it is n-type. A detailed physical description is beyond our capabilities now. (See EE 332 and beyond.) But we can get a good approximation by using the fact that the MOS structure is basically a capacitor. Once the gate voltage is bigger than the threshold, electrons pile up under gate just like charge on a plate of a capacitor.

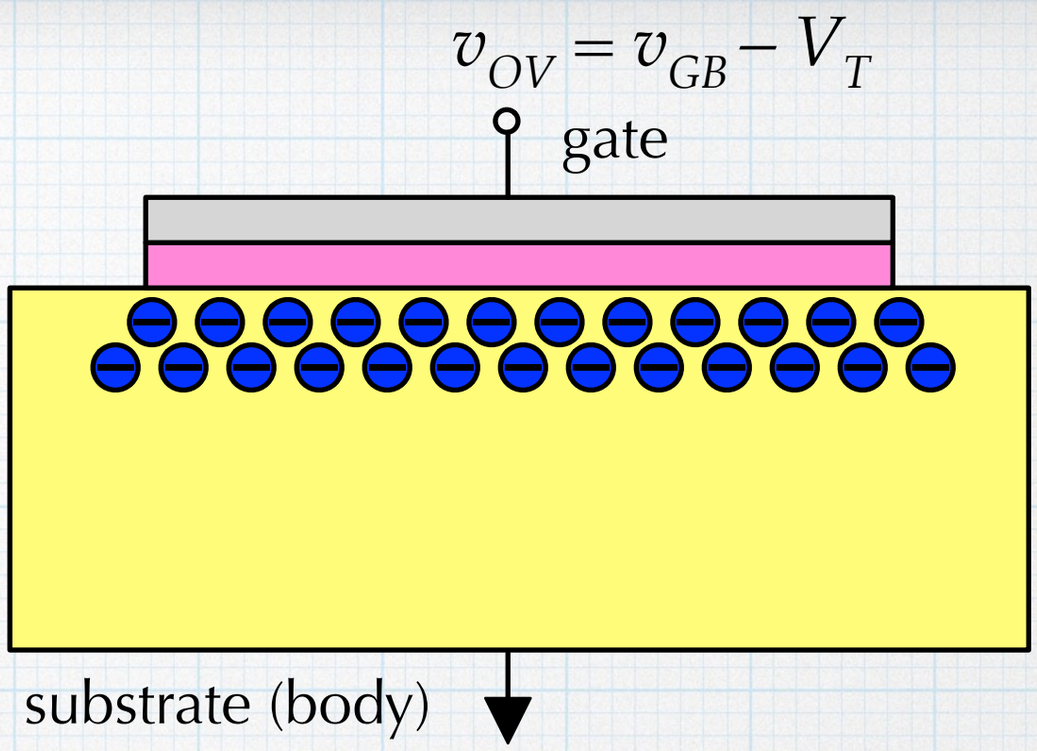
$$Q_{inv} = C_{gate}\left(v_{GB} - V_T\right)$$

$$\frac{Q_{inv}}{A} = C_{ox}\left(v_{BG} - V_T\right) = qn_s$$

where n_s is the *sheet concentration* of electrons in the inversion layer. (electrons per unit area).

The threshold voltage

The threshold voltage for a MOSFET depends primarily on two things. The doping level of the substrate and the thickness of the insulating gate oxide layer.



The definition of threshold is that the electron concentration under the gate is equal to the hole concentration in the rest of the substrate. If the hole concentration in the substrate is higher (due to higher doping) more voltage is required to induce the higher number of electrons needed for inversion. Increased doping means higher threshold.

The electric field in MOS capacitor in the gate that controls what is happening in the inversion layer. Since $\mathcal{E} \propto v_{GB}/t_{ox}$, making the oxide thicker requires more voltage in order to create the same electric field. So a thicker oxide also leads to higher a threshold voltage.

Integrated circuit fabrication engineers work very hard to control the threshold voltages of the MOSFETs. Not easy when there are millions (or billions) of transistors in a chip!

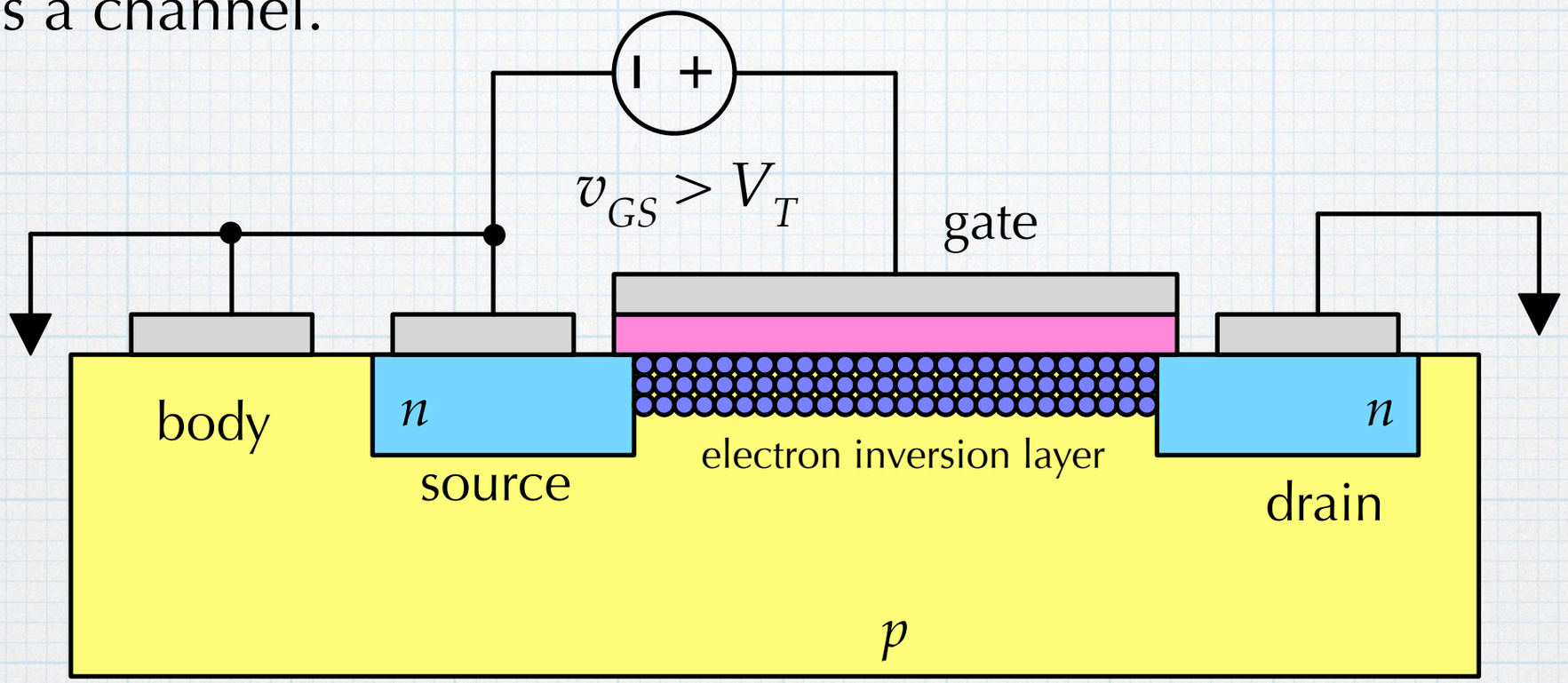
Summary of MOS capacitor operation

So through the application of the gate voltage, we control carriers under the gate.

$v_{GB} < 0$	hole accumulation
$0 < v_{GB} < V_T$	carrier depletion
$v_{GB} > V_T$	inversion – electron layer forms, $qn_s = C_{ox}(v_{GB} - V_T)$

Now we see a mechanism by which we might get current to flow between drain and source.

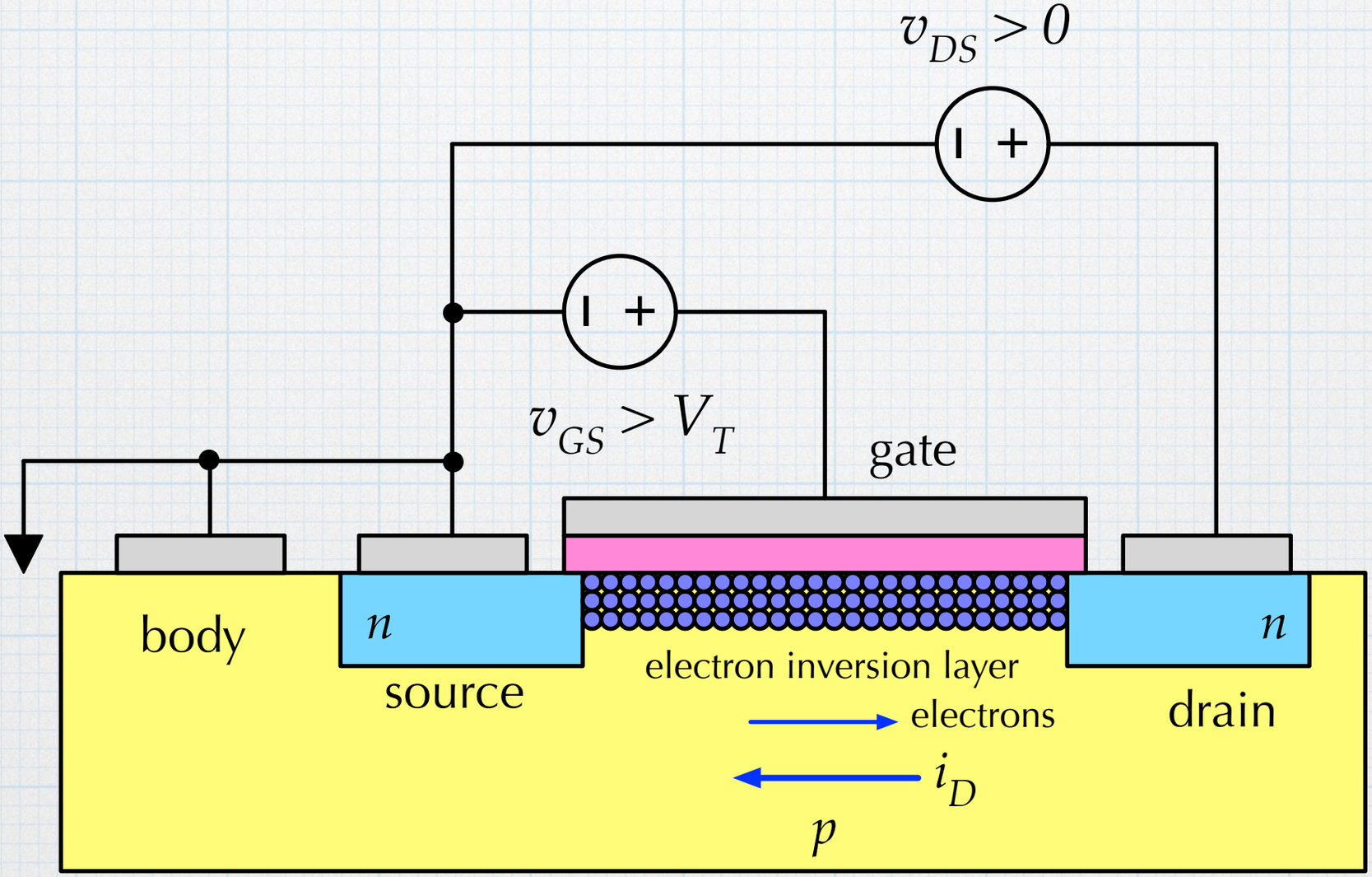
Apply a gate voltage to create an electron inversion layer. Then electrons from the source can flow to the drain using the inversion layer as a channel.



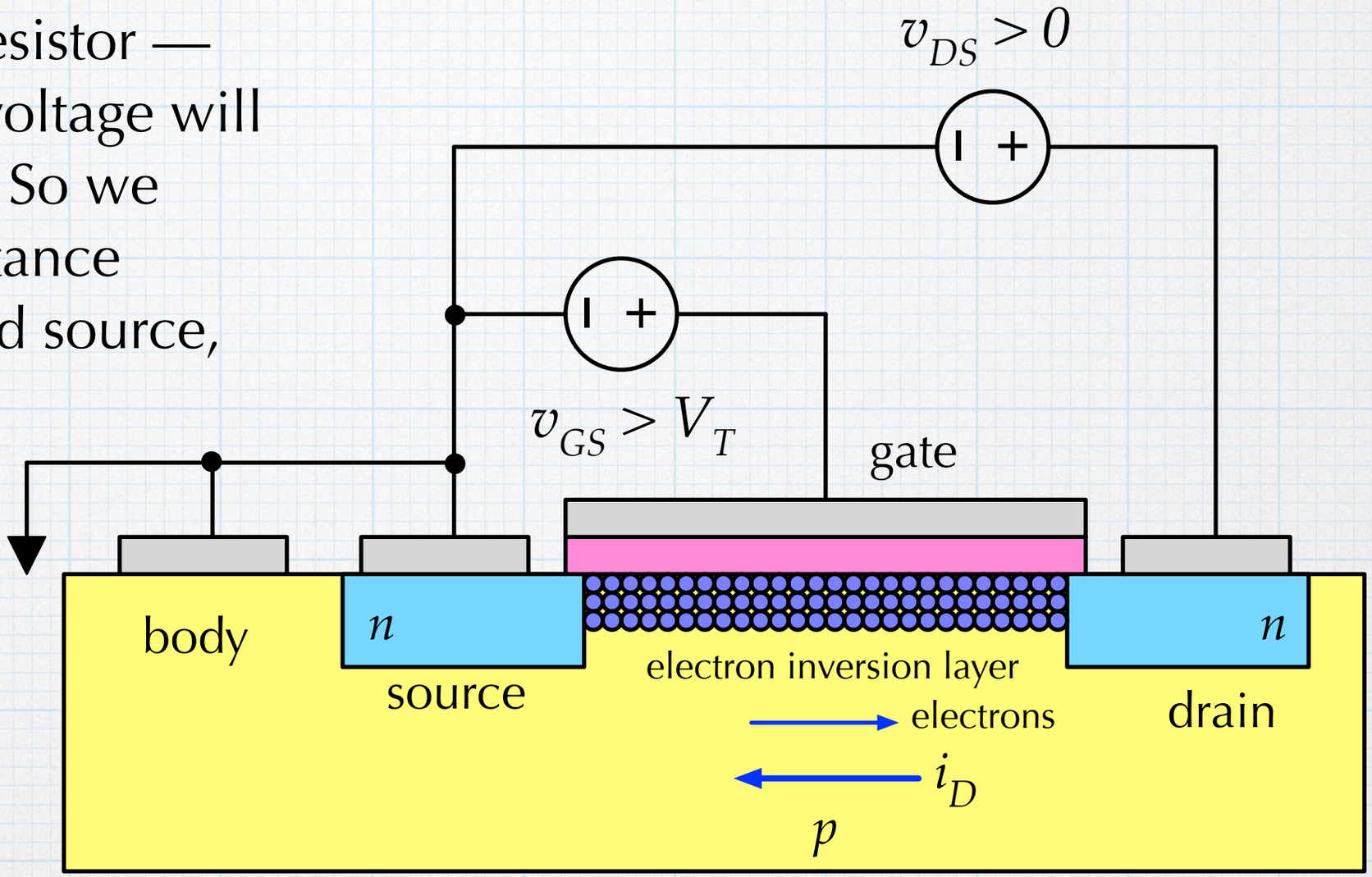
For now, we connect the source to the body (call this ground) and apply the controlling voltage between the gate and the source. This is OK for the time being, but we will have to revisit the issue of the body connection later. With the drain also at ground, the inversion layer (channel) is uniform between source and drain.

Drain resistance

With the inversion layer in place, we can apply a positive voltage between the drain and source. This creates an electric field *along* the channel that will push electrons from the source through the inversion layer to the drain. The moving electrons represent a current in the opposite direction — *from* drain *to* source.



At first inspection, the current is essentially that of a resistor — increasing the drain voltage will increase the current. So we need to find the resistance between the drain and source, R_{DS}.



The total resistance is the sum the resistances of the three regions: $R_{DS} = R_{source} + R_{inversion} + R_{drain}$. Usually, the source and drain are very heavily doped (lots of electrons) and so the resistances in those regions are very small. Most of the resistance is associated with the inversion layer so that $R_{DS} \approx R_{inversion}$.

Using the basic definition for a resistor:

$$R_{DS} = \rho \frac{L}{A} = \left(\frac{\rho}{t_{inv}}\right)\left(\frac{L}{W}\right) \qquad \rho_{inv} = \frac{1}{q\mu_n n_{inv}}$$

where n_{inv} is the electron concentration (m^{-3}) in the inversion layer, t_{inv} is the "thickness" of the inversion layer, μ_n is the electron mobility, q is the charge on one electron and L and W are the gate length and width.

$$i_D = \frac{v_{DS}}{R_{DS}} = \left[\left(q\mu_n n_{inv} t_{inv}\right)\left(\frac{W}{L}\right)\right] v_{DS}$$

The electron concentration and inversion layer thickness are difficult to quantify because of the sheet-like nature of the inversion layer. However, the product $n_{inv} \cdot t_{inv}$ can be re-expressed more easily – it represents the "sheet concentration" (m^{-2}) of electrons in the inversion layer, which we denote as n_s.

We saw earlier that: $qn_s = C_{ox}\left(v_{GS} - V_T\right)$

Using the basic definition for a resistor:

$$R_{DS} = \rho_{inv}\frac{L}{A} = \left(\frac{\rho_{inv}}{t_{inv}}\right)\left(\frac{L}{W}\right) \qquad \rho_{inv} = \frac{1}{q\mu_n n_{inv}}$$

where n_{inv} is the electron concentration (m^{-3}) in the inversion layer, t_{inv} is the "thickness" of the inversion layer, μ_n is the electron mobility, q is the charge on one electron and L and W are the gate length and width. (See slide 4 for a reminder of the 3D geometry.) Putting the equations together:

$$R_{DS} = \left(\frac{1}{q\mu_n n_{inv} t_{inv}}\right)\left(\frac{L}{W}\right)$$

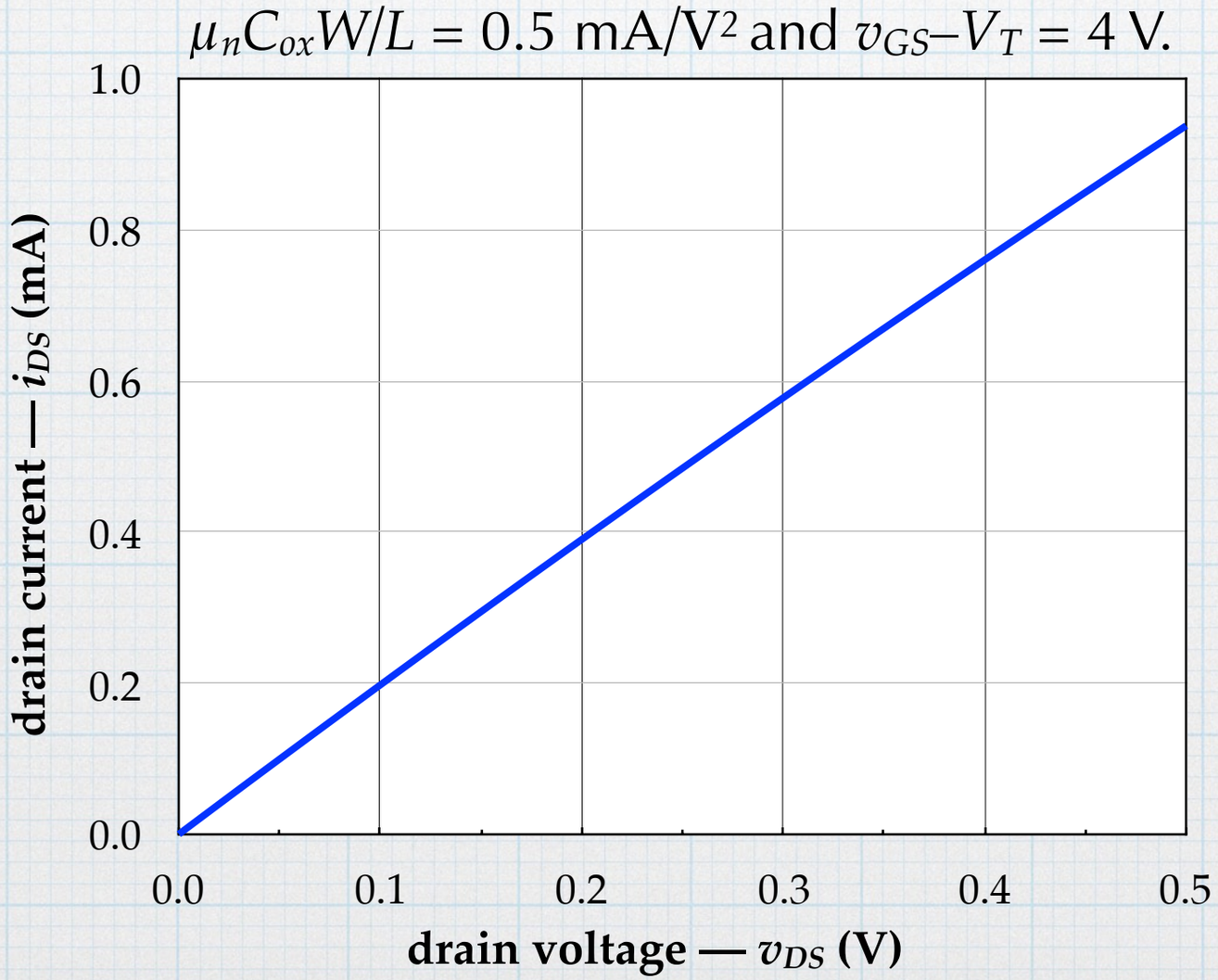
$$i_D = \frac{v_{DS}}{R_{DS}} = \left[\left(q\mu_n n_{inv} t_{inv}\right)\left(\frac{W}{L}\right)\right] v_{DS}$$

The electron concentration and inversion layer thickness are difficult to quantify because of the sheet-like nature of the inversion layer. However, the product $n_{inv} \cdot t_{inv}$ can be re-expressed more easily – it represents the "sheet concentration" (m^{-2}) of electrons in the inversion layer, which we denoted earlier as n_s.

$$i_D = \left[q\mu_n n_s \left(\frac{W}{L} \right) \right] v_{DS}$$

We saw earlier that: $qn_s = C_{ox}\left(v_{GS} - V_T\right)$

$$i_D = \left[\left(\mu_n C_{ox} \frac{W}{L} \right) \left(v_{GS} - V_T\right) \right] v_{DS} \qquad R_{DS} = \frac{1}{\left(\mu_n C_{ox} \frac{W}{L}\right)\left(v_{GS} - V_T\right)}$$

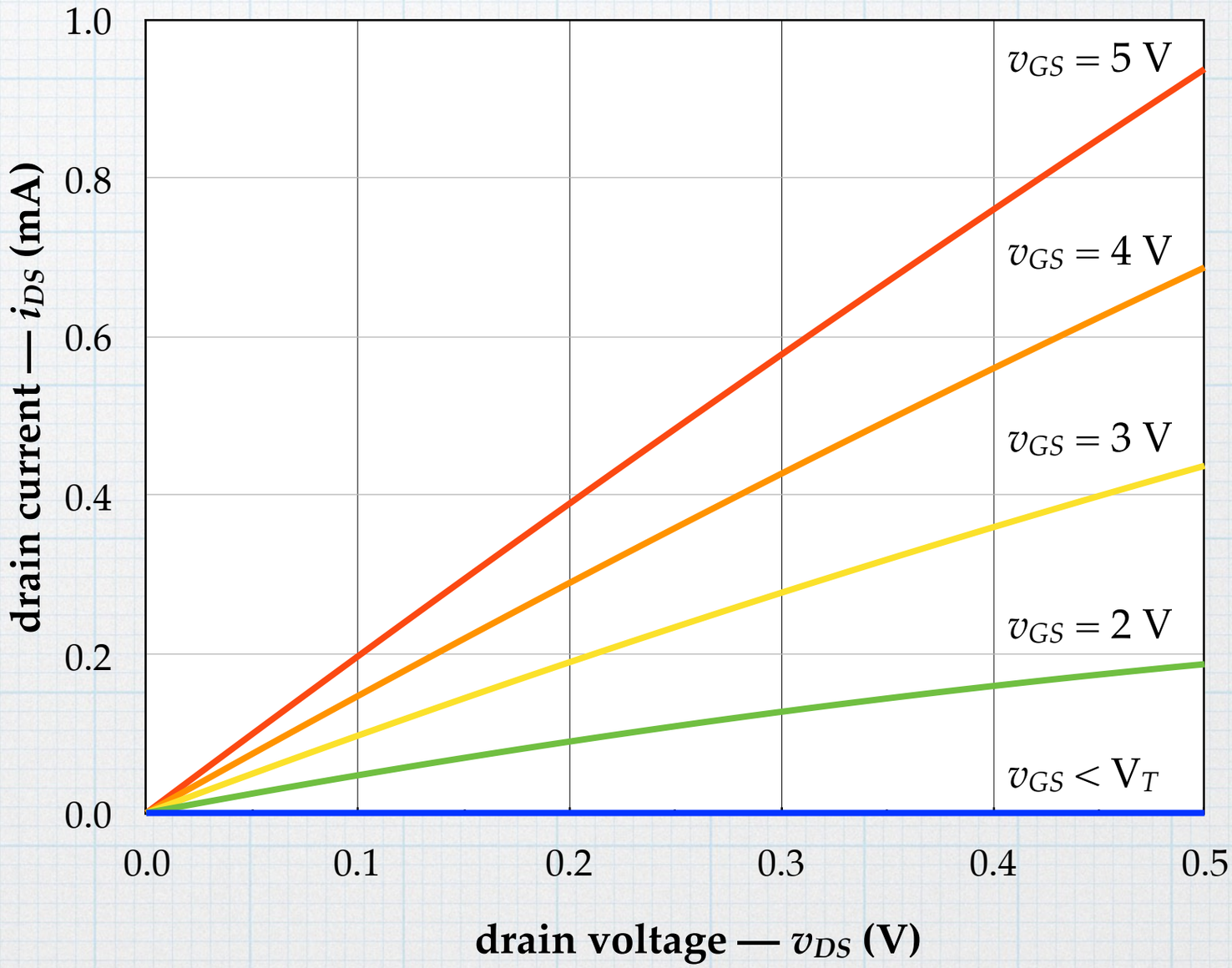


But the most important observation is that the channel resistance depends on the gate voltage! As the gate voltage is increased (which increases the electron sheet concentration in inversion layer), the channel resistance decreases. This a variable resistor. (But not anything like a potentiometer.)

For $v_{GS} < V_T$, there is no inversion layer: $R_{DS} \rightarrow \infty$ and $i_D = 0$ (blue curve).

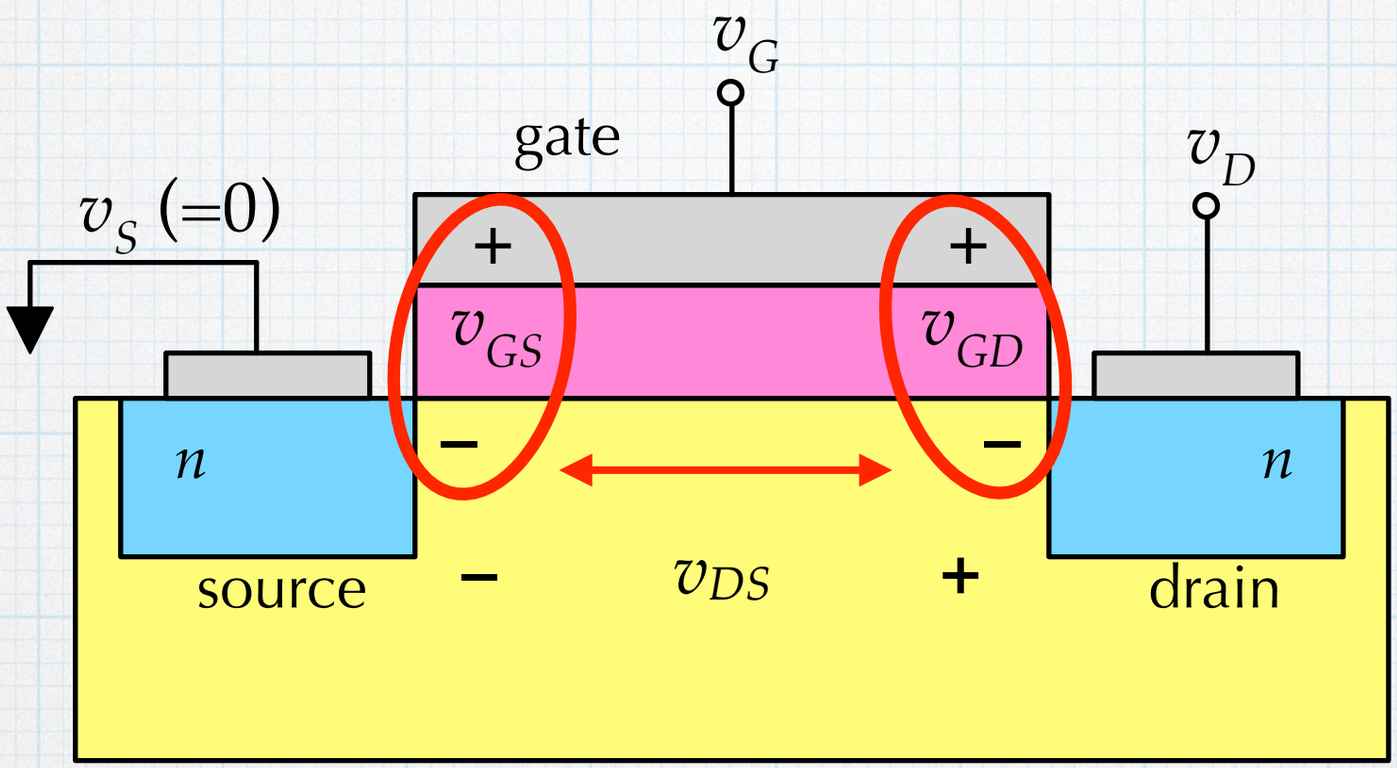
For $v_{GS} \geq V_T$, the resistance decreases (slope increases) as v_{GS} increases.

$\mu_n C_{ox} W/L = 0.5$ mA/V^2 and $V_T = 1$ V.

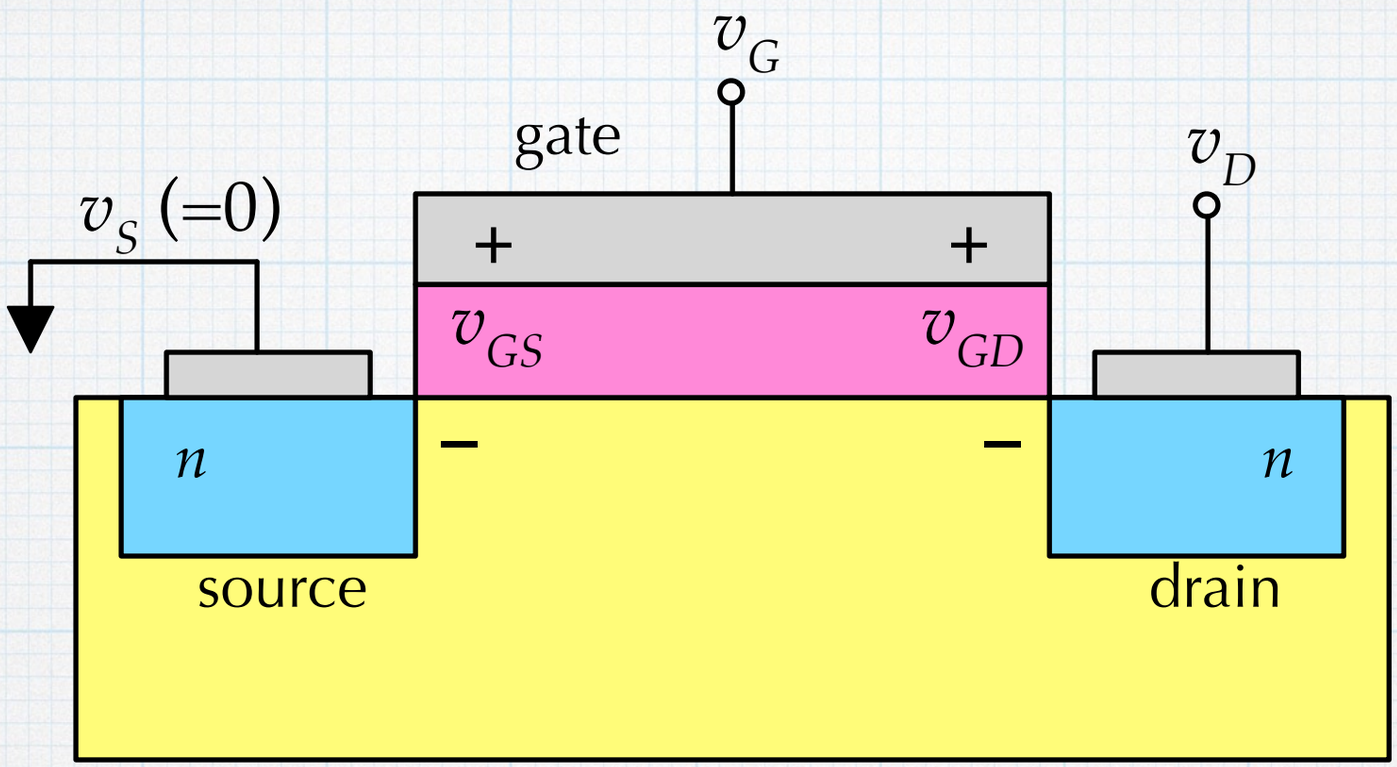


Drain resistance

However, the picture gets more complicated as v_{DS} increases.



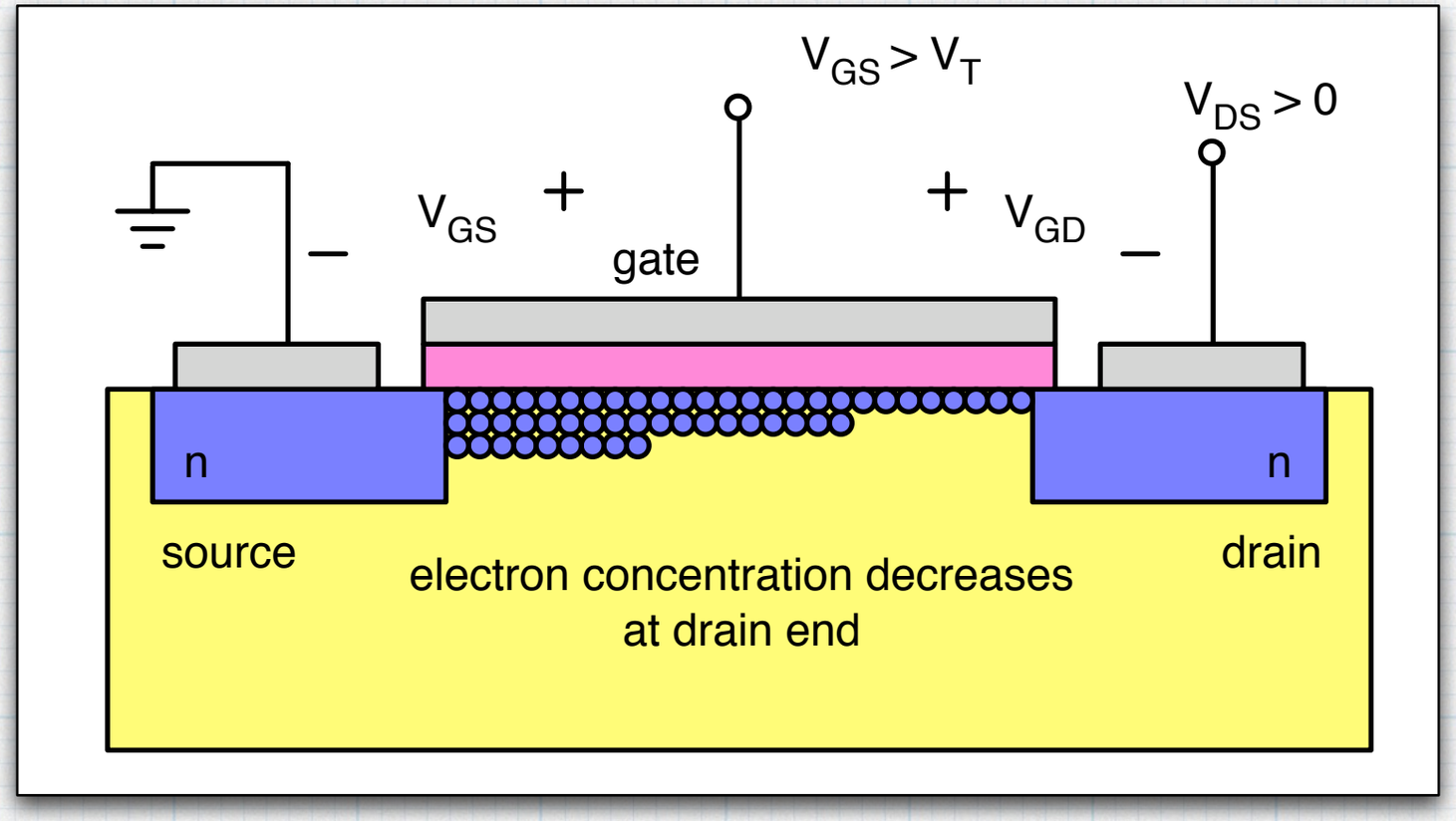
The electron sheet concentration depends on the local voltage of the semiconductor underneath the gate. In the MOS capacitor, the voltage of the semiconductor was the same everywhere, so the electron concentration was uniform. In the case of the MOSFET, the local voltage under the gate can vary from one end to the other. At the source end, the channel voltage is 0, (because we have tied the source to ground). At the drain end, the channel voltage is v_D, because we have applied a voltage there in order to make current flow. In between, the channel voltage ranges between 0 and v_D.



The electron sheet concentration is determined by the difference between the gate voltage and the local channel voltage. At the source end, $qn_s = C_{ox}(v_{GS} - V_T)$. At the drain end, $qn_s = C_{ox}(v_{GD} - V_T)$.

Since $v_{GD} = v_{GS} - v_{DS}$, then $qn_s = C_{ox}(v_{GS} - v_{DS} - V_T)$ at the drain end. Increasing v_{DS} decreases v_{GD} and so decreases the electron concentration at the drain end, *increasing* the incremental resistance at the drain end.

The result of all this is that the inversion layer (channel) resistance is actually non-linear – the resistance changes as the drain voltage changes.



The non-uniform electron concentration along the channel implies that the channel resistance is not uniform — it is greater at the drain end, where there are fewer electrons.

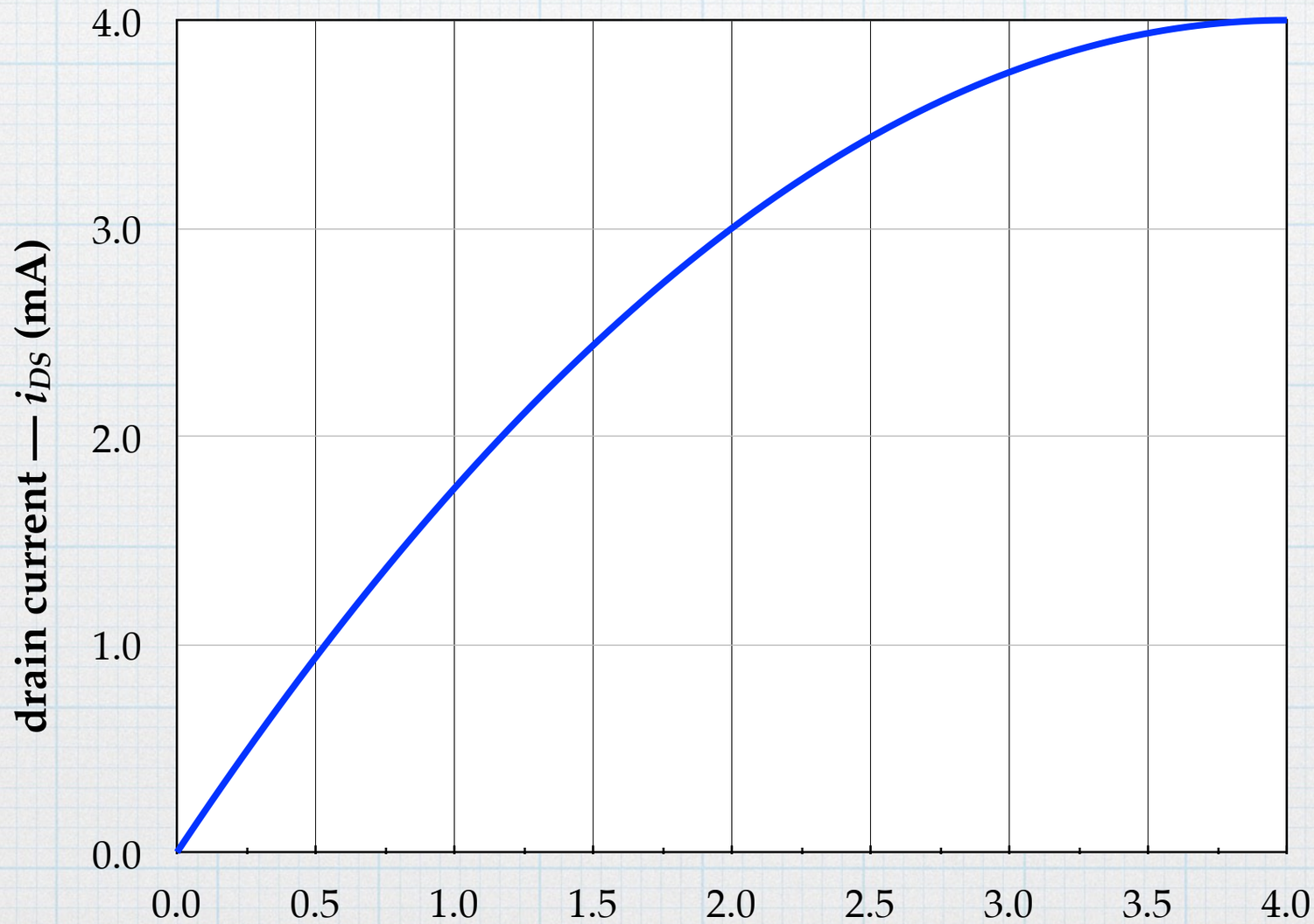
So, the drain voltage increases, the channel resistance increases. The effect is that the I-V curve bends over.

So a more correct analysis of the current flow shows that the i_D-v_{DS} relationship is not linear, but parabolic. (See EE 332.)

$$i_D = \left(\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\right)\left[2\left(v_{GS} - V_T\right)v_{DS} - v_{DS}^2\right]$$

Exercise: Show that this reduces to the linear equation for small values of v_{DS}.

$\mu_n C_{ox} W/L = 0.5$ mA/V² and v_{GS}–$V_T = 4$ V.



This mode of operation is known as "linear" or "ohmic". (Some old codgers also call is "triode" — but don't use that.)

$$i_D = \left(\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\right)\left[2\left(v_{GS}-V_T\right)v_{DS}-v_{DS}^2\right]$$

Notation: To save on ink and time, define the constant in front as:

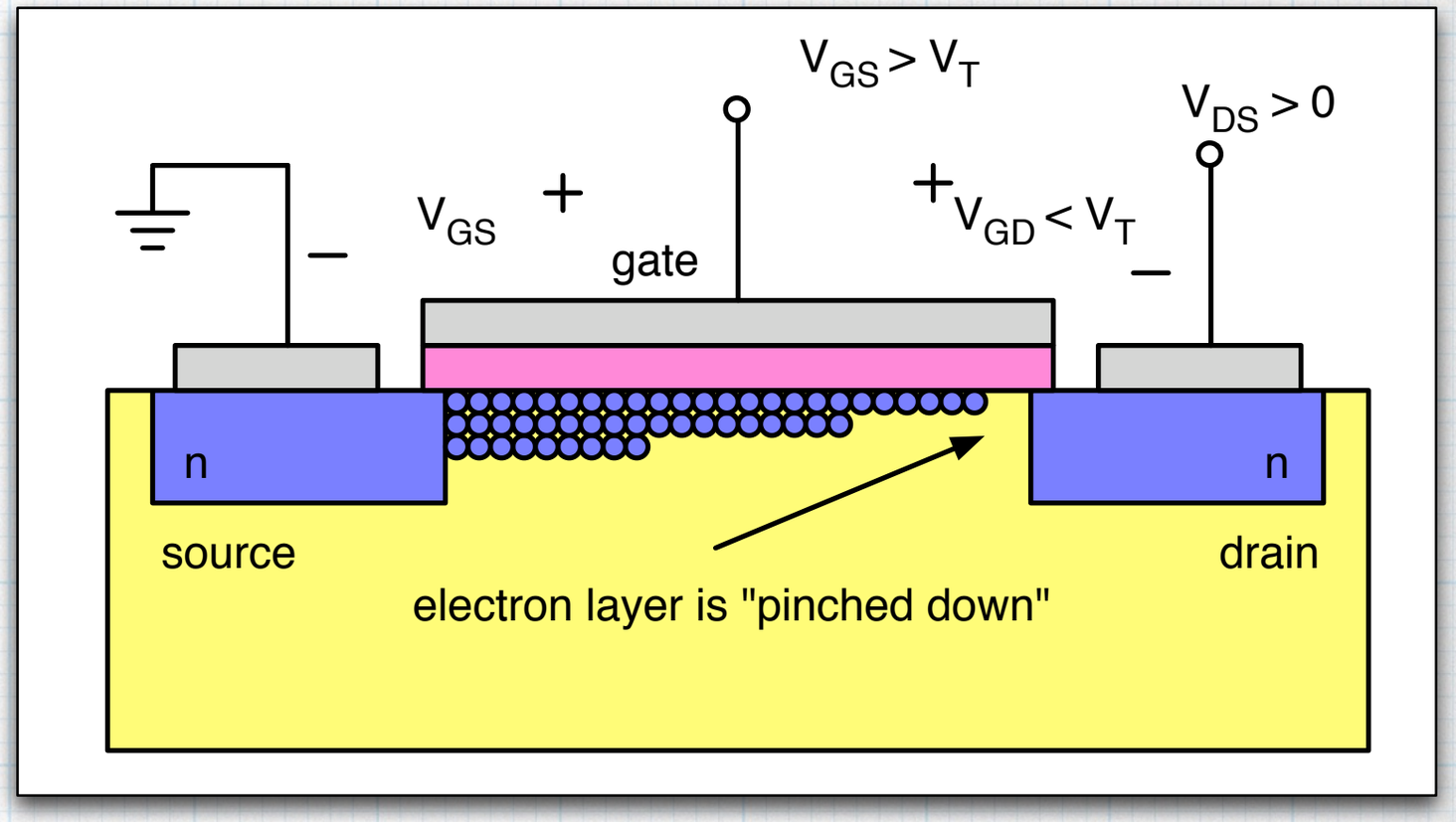
$$K = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}$$

The units of this "MOS current parameter" are $\mathrm{A/V^2}$. Or more typically $\mathrm{mA/V^2}$. (Check it.)

Note: This use of K is not standard. Different textbooks will handle the current parameter differently.

current saturation

Now take the final step: If v_{DS} is increased sufficiently, we reach a point where $v_{GD} < V_T$. The electron concentration should disappear at the drain end!



The usual terminology is to say that the channel is “pinched off” at the drain end, but this is slightly misleading. If the channel were truly pinched off, then the electron concentration would go to zero there and the current would necessarily have to go to zero, also.

Instead, it might be better to say that the channel is “pinched down”. The electron concentration goes to some minimum value, but never really shuts off. In effect, the flowing current works to hold the channel open, even though the drain voltage seems high enough to truly pinch off the channel.

The mathematical analysis of what happens in a FET at pinch-off is a fairly complicated problem in electromagnetics – definitely beyond our current capabilities. (Again, see EE 332.)

However, the end result on the device behavior is easy to grasp – the current saturates at the pinch down value. For higher values of v_{DS}, the current stays constant at a constant value.

The condition for the NMOS going reaching the "pinch-down" condition is $v_{GD} \leq V_T$. This can be re-expressed in terms of v_{DS} and v_{GS}.

$$v_{DS} \geq v_{GS} - V_T$$

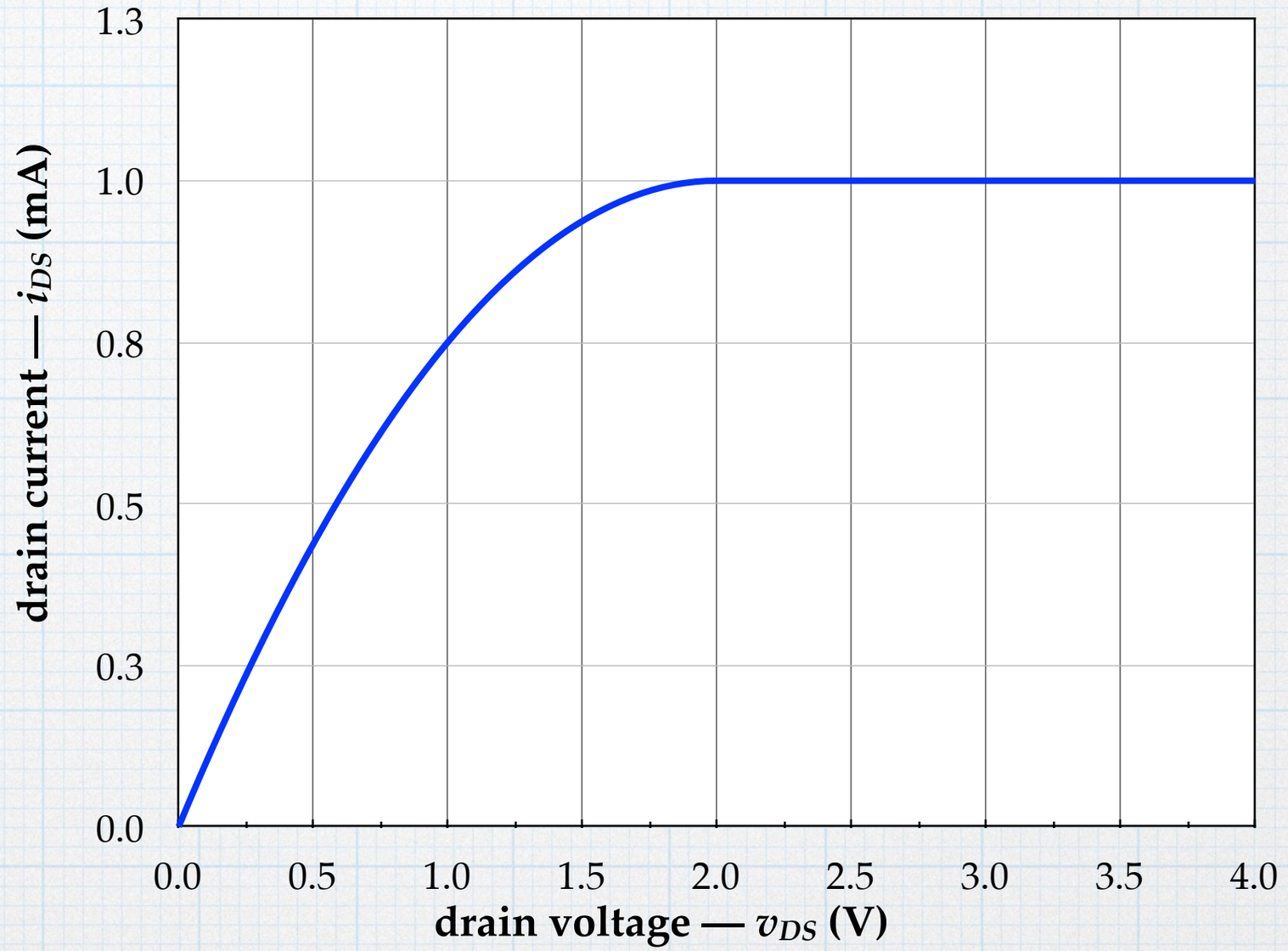
To determine the saturation current, we can insert the pinch-down condition given above into the i_D-v_{DS} equation from the ohmic region of operation. Then the saturated current is

$$i_D\,(sat) = \left(\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\right)\left[v_{GS} - V_T\right]^2 = K\left[v_{GS} - V_T\right]^2$$

(Work out these details for yourself.)

NMOS i_D-v_{DS} curves

Fix v_{GS} and vary v_{DS}.
Plot the drain current.



For small v_{DS} ($v_{DS} < v_{GS}$ - V_T), the NMOS will be in the ohmic region.

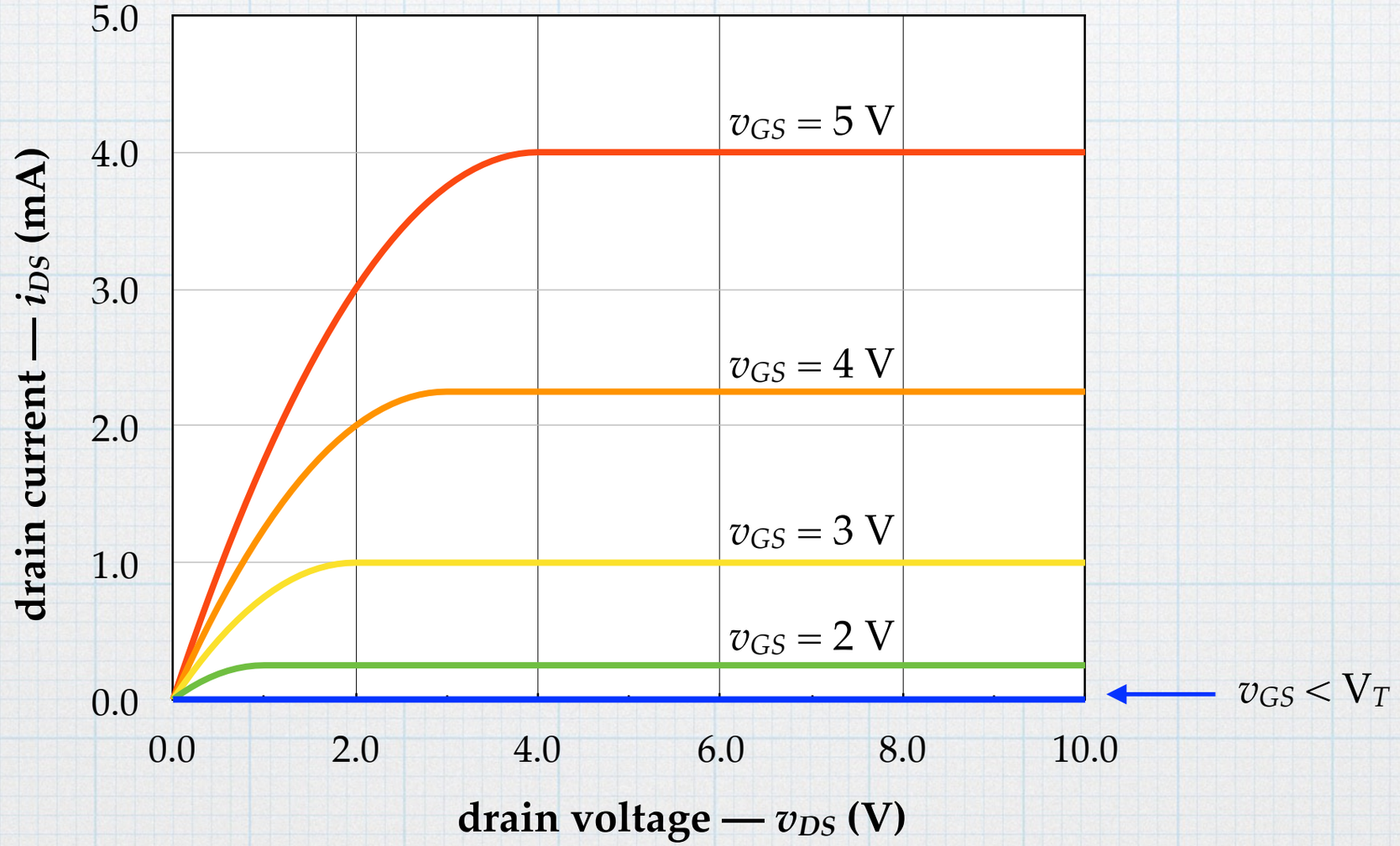
At higher v_{DS} ($v_{DS} \geq v_{GS}$ - V_T), the channel pinches down and the current saturates.

If $v_{GS} < V_T$, there is no inversion layer and the NMOS is off ($i_D = 0$).

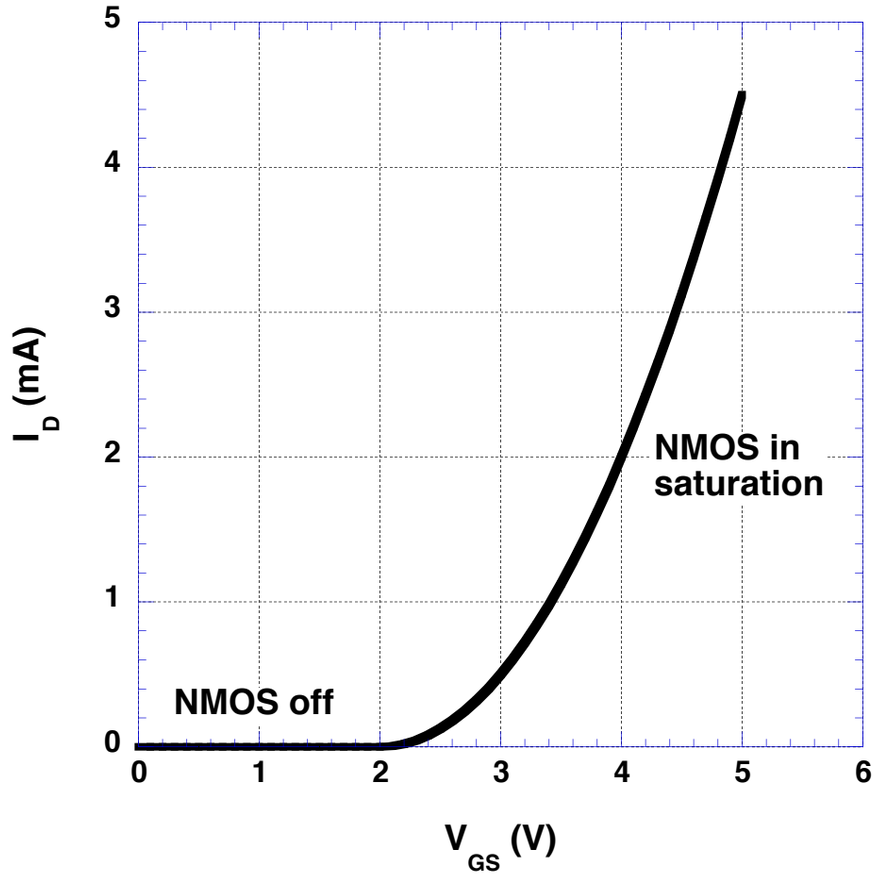
NMOS i_D-v_{DS} curves

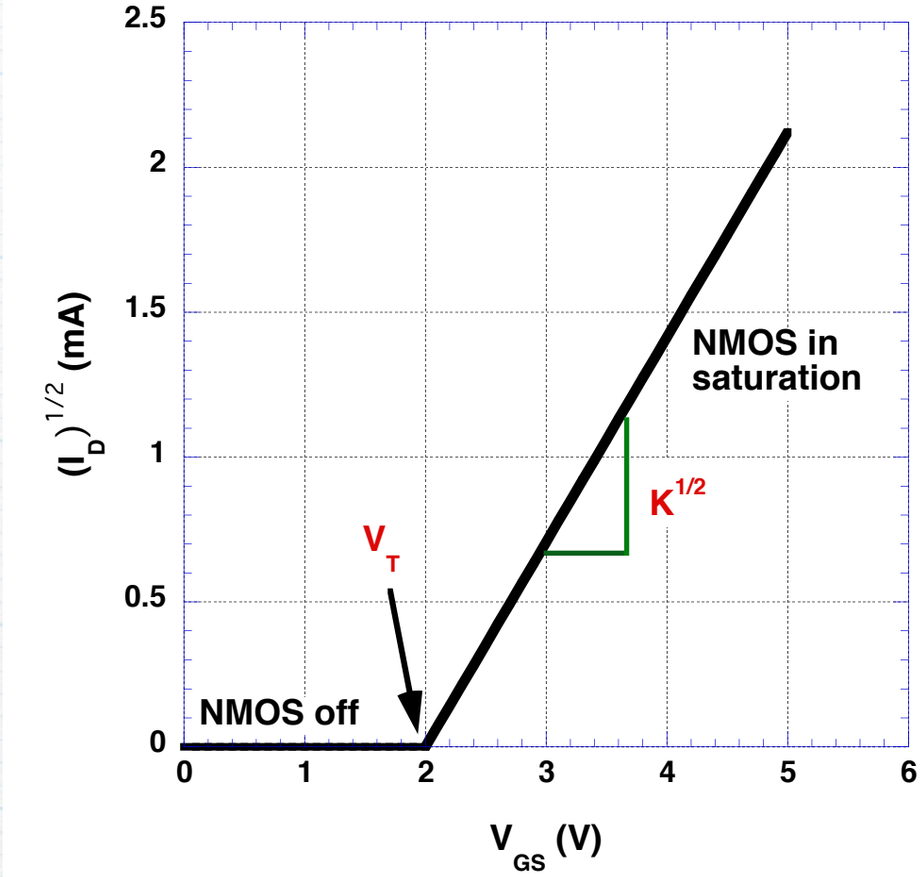
A set of characteristic curves for an NMOS with $V_T = 1$ V and

$$K_n = \frac{1}{2}\mu_n C_{ox}\frac{W}{L} = \frac{1}{2}\left(0.05\frac{\text{mA}}{\text{V}^2}\right)\left(\frac{10\ \mu\text{m}}{1\ \mu\text{m}}\right) = 0.25\frac{\text{mA}}{\text{V}^2}$$



Another useful plot is i_D vs. v_{GS}, with v_{DS} fixed. If v_{DS} is kept large so that the NMOS does not go into the ohmic region, the curve is essentially a plot of the saturation equation.





In a slight modification, we can plot $i_D^{1/2} = K^{1/2}(v_{GS} - V_T)$. From this graph, we can immediately pick out the two important NMOS parameters: V_T and K.

Summary of NMOS equations

$$V_T,\ K = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}$$

$v_{GS} < V_T$ off $i_D = 0$

$v_{GS} \geq V_T$ on $v_{DS} < v_{GS} - V_T$ ohmic or linear

$$i_D = \left(\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\right)\left[2\left(v_{GS} - V_T\right)v_{DS} - v_{DS}^2\right]$$

$v_{DS} \geq v_{GS} - V_T$ saturation

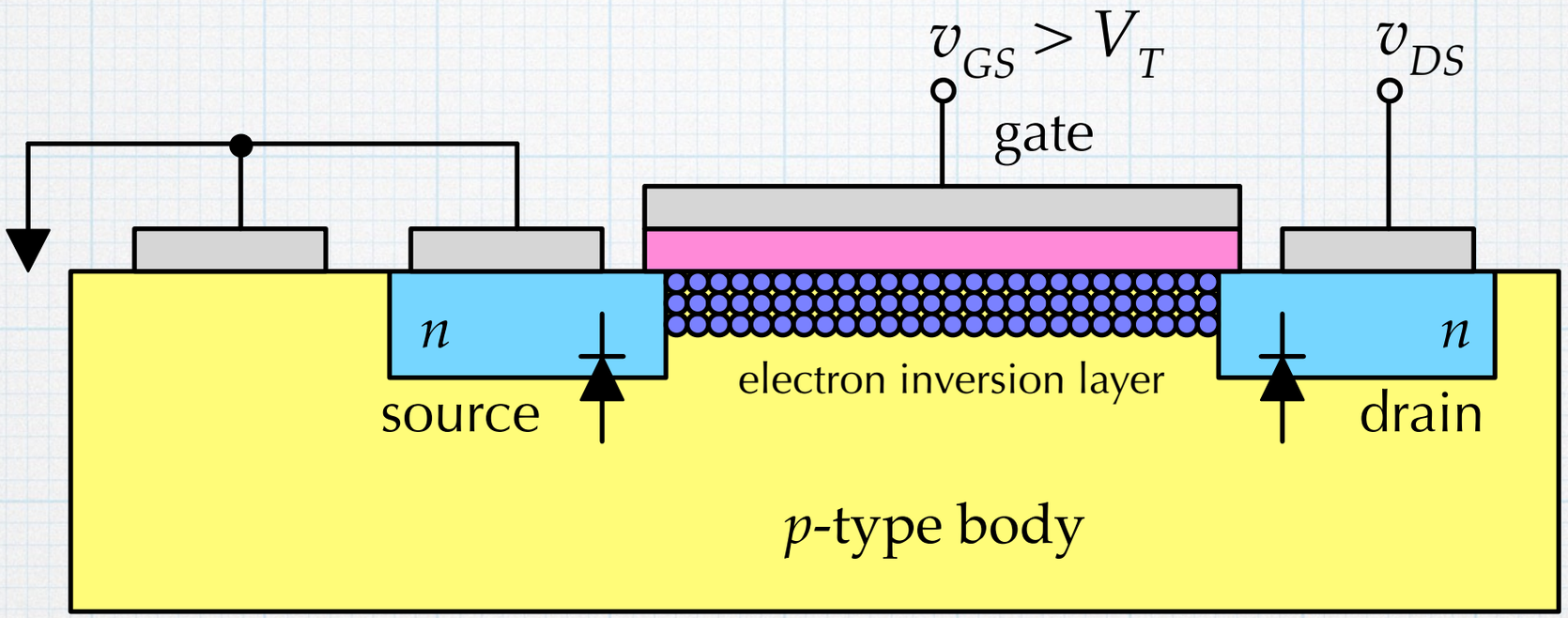
$$i_D = K\left[v_{GS} - V_T\right]^2$$

$i_G = 0$!! (at least at DC)

The equations are exact (within the limits of the simplest MOSFET model). The quadratic dependence of current on voltage means that the circuit analysis will be non-linear and we will have frequent need for the quadratic equation.

The NMOS substrate (body)

In applying the drain voltage to make the drain current flow, is it necessary to use positive value for v_{DS}? It would seem that $v_{DS} < 0$ might be OK. However, it's not.



To see why, consider the p-n junctions formed by the drain and source with the substrate.

If the substrate is at ground potential and we apply a negative voltage to the drain, the p-n junction there would be forward-biased – probably with a very large voltage possibly without a limiting resistor – and a huge forward-bias diode current would flow from the substrate into the drain. The junction would likely be burned out.

The NMOS substrate rule: The NMOS substrate (body) should be connected to the lowest voltage in the circuit – usually the ground. (Although it could be a negative power supply.) Then the source and drain will both always be at the same or higher voltages, and it will be impossible to forward-bias the diodes.