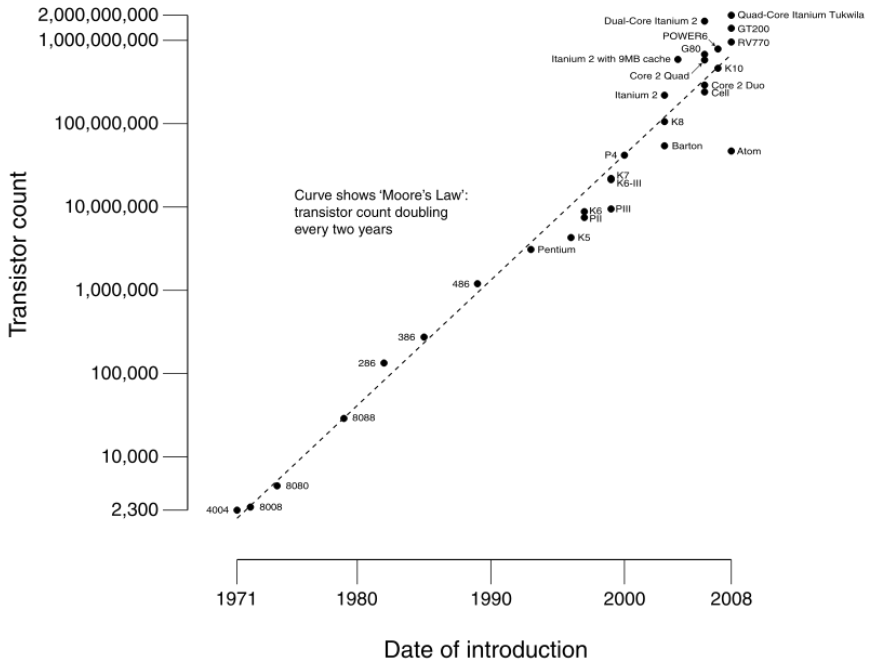
Section V

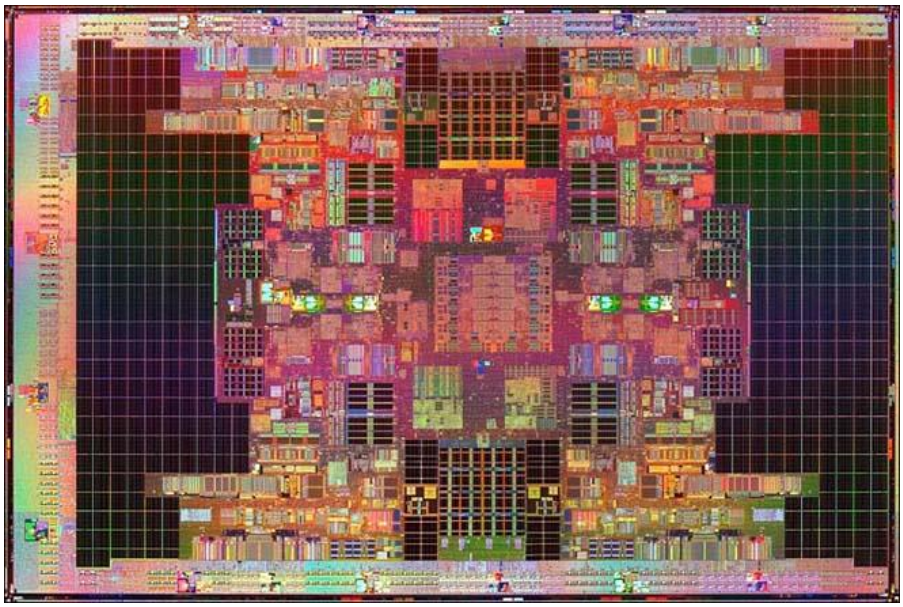
The Transistor

Before 1950 electronic equipment used vacuum tubes, otherwise called valves, which acted as transistors. They usually consumed a few Watts of power each and as a result created a lot of heat. They were also quite bulky as you can see below:

They generally worked by having an incandecent connection (a heated cathode that emits light and electrons) along with a separate anode separated by a grid. Electrons flowed from the incandecent connection to the isolated anode. The grid was connected to a supply voltage which modulated the current flow between the incandescent connection and the isolated anode. In this way signal amplification was possible.
In 1951 William Shockley invented the first transistors based on semiconductors for which he received the Nobel Prize. The advantages of these types of transistors are that they can be made much smaller than a vacum tube and that they consume much less power.
The transistor's impact on electronics has been enormous. Almost all electronic equipment produced today uses semiconductor technology. Before 1951 a computer filled an entire room and cost millions of euro. Today a much better computer sits on a desk with a few billion transistors in its CPU and costs roughly a thousand euro.

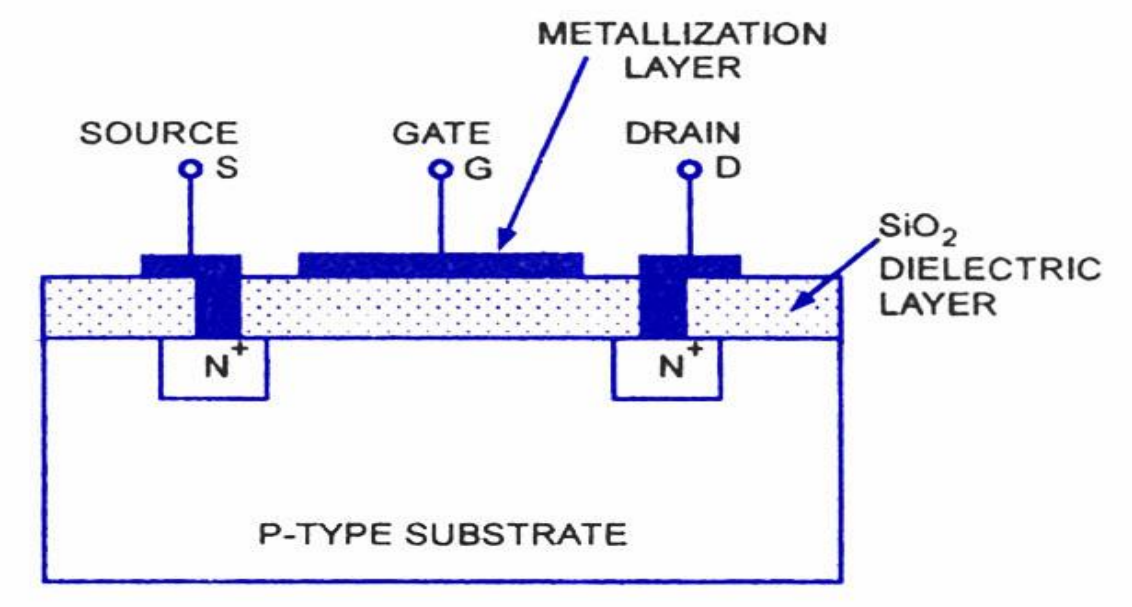
CPU Transistor Counts 1971-2008 & Moore's Law



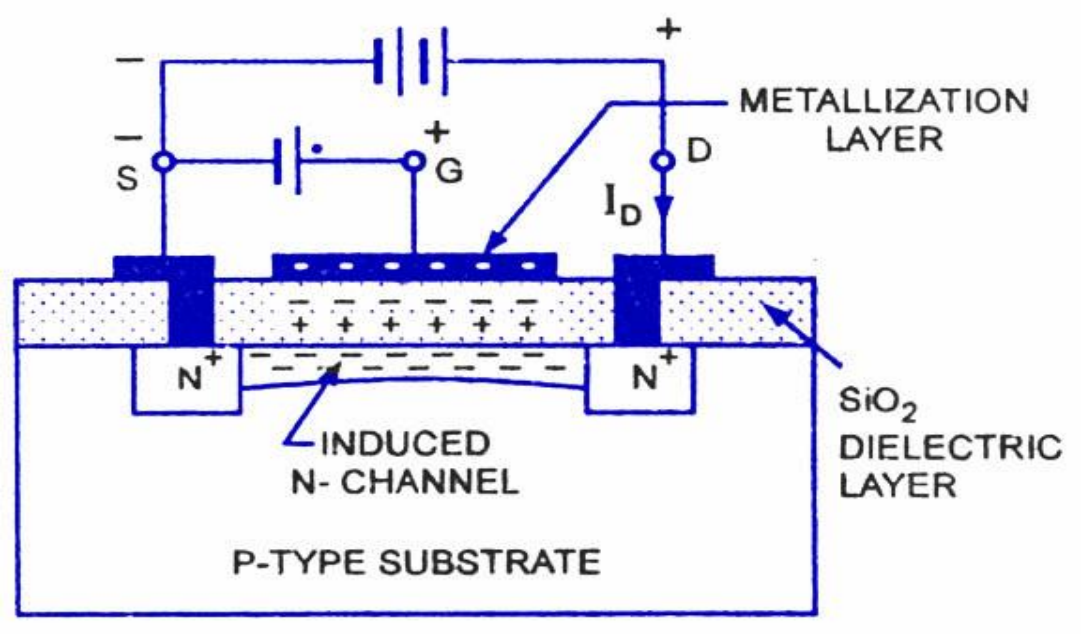


In this course we examine the most widely used transistor, the Enhancement Metal Oxide Field Effect Transistor – the Enhancement MOSFET (E-MOSFET). The advantage of these types of transistors over others is that faster switching speeds are possible which translates into, for example, faster computer processors. Furthermore the E-MOSFET is relatively easy to manufacture and can be used also as a resistor or a capacitor. It does not however achieve the same gain as, for example, the Bipolar Junction Transistor.

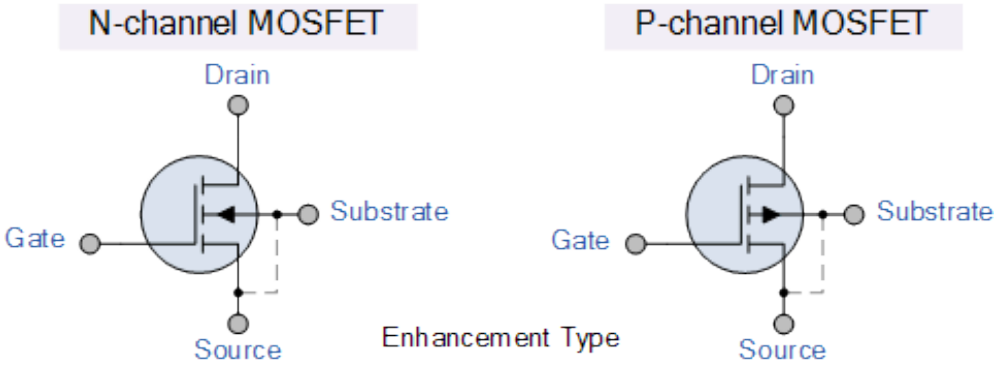
The structure of an n-channel E-MOSFET is as follows:



Between the gate and the substrate there is a layer of silicon dioxide which is a very good insulator. Silicon dioxide is also called silica (glass). It is most commonly found in nature as quartz and is the primary component of sea sand. With a supply voltage connected between the drain and the source (ground) and a zero voltage between the gate and the source no current will flow because of the depletion regions that will exist between the n+ and p-type materials.
As the gate-source voltage is increased electrons are drawn from the source to the gate region forming an n-channel between the drain and the source and current starts to flow from the drain to the source. This n-channel is called an inversion layer.



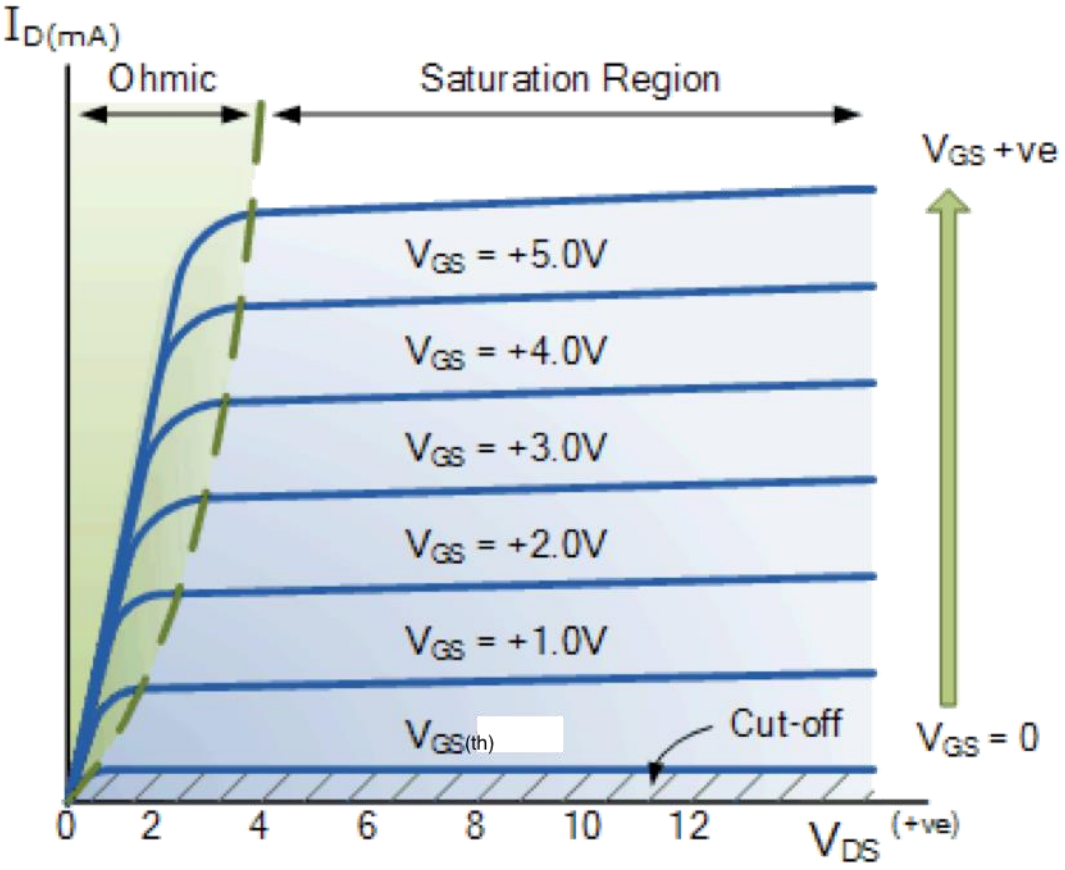
Throughout this section capital letters are used to denote dc voltages/currents which is to say time-independent voltages/currents. Small letters are used to denote time dependent voltages/currents. These may be purely ac or ac with a dc component.
The minimum gate-source voltage that creates an inversion layer and breaks down the depletion regions sufficient for current to flow between the drain and the source, is called the *threshold voltage*, $V_{GS(th)}$. When v_{GS} is less than $V_{GS(th)}$, the drain-source current is zero. When v_{GS} is greater than $V_{GS(th)}$ the n-type inversion layer connects the drain and the source and we get current. $V_{GS(th)}$ can vary from less than 1V to more than 5V depending on the particular device being used. The 3N169 is an example of an enhancement-type MOSFET. It has a threshold voltage of 1.5V.The symbols for enhancement-type MOSFETs are as follows:



The operating principles for p-type MOSFETS are the same as for n-type only that v_{GS} is now negative.
MOSFETs have a very thin layer of silicon dioxide (which is a very good insulator) between the gate and the substrate. This layer is kept as thin as possible to give the gate as much control over the drain current as possible. Because this insulating layer is so thin it is easily destroyed by excessive gate-source voltages. For example, if a MOSFET is supplied with a gate-source voltage greater than its maximum rating, $V_{GS(\max)}$ then the MOSFET can be thrown away since it is most likely destroyed.
A +/-30V rating is typical. However it is easy to destroy a MOSFET in other ways. If you insert the MOSFET into or remove it from a circuit with the power supply on then the inductive back-EMF (from the circuit) may be enough to destroy the transistor. Even picking up a MOSFET may destroy it due to static electricity. This is why MOSFETs are often shipped with a wire ring around the leads. You remove the ring AFTER you have inserted the MOSFET into the circuit. Another way to protect a MOSFET is to build in a Zener diode (a Zener diode is a diode which allows current to flow in the forward direction in the same manner as an ideal diode, but also permits it to flow in the reverse direction when the voltage is above a certain value known as the breakdown voltage) in parallel with the gate-source connection. The idea is that the Zener breaks down before any damage occurs.

Characteristic Curves

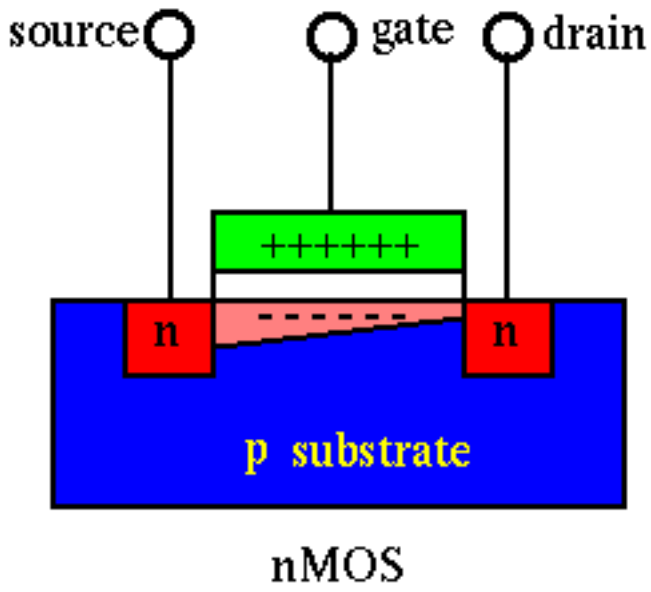
Below is a set of Drain Curves (I_D vs V_{DS} for varying v_{GS}) for an n-channel E-MOSFET. The lowest curve is the $V_{GS(th)}$ curve. When v_{GS} is less than $V_{GS(th)}$ the drain current is extremely small and the transistor is effectively OFF. This region is known as the *Cutoff Region.* When v_{GS} is greater than $V_{GS(th)}$ significant drain current flows, with the amount depending on the value of v_{GS}.



Note how for small V_{DS} there is an approximately linear relationship between I_D and V_{DS}. We therefore refer to this region as the *Linear Region* or *Ohmic Region*.

Each curve then 'flattens out'. With an increase in V_{DS} there is only a small increase in I_D. The reason for this is as follows: An increase in V_{DS} has the effect of forcing greater current flow between the drain and the source. However as V_{DS} is increased the PN Junction at the drain is in

the process of entering into reverse bias (like a diode) which, as the depletion region expands, has the effect of impeding current flow. This phenomenon, known as '*pinchoff'*, results in little increased current flow for increased V_{DS}. The reason it 'pinches off' at the drain and not so much the source is that there is a voltage drop across the n-channel.

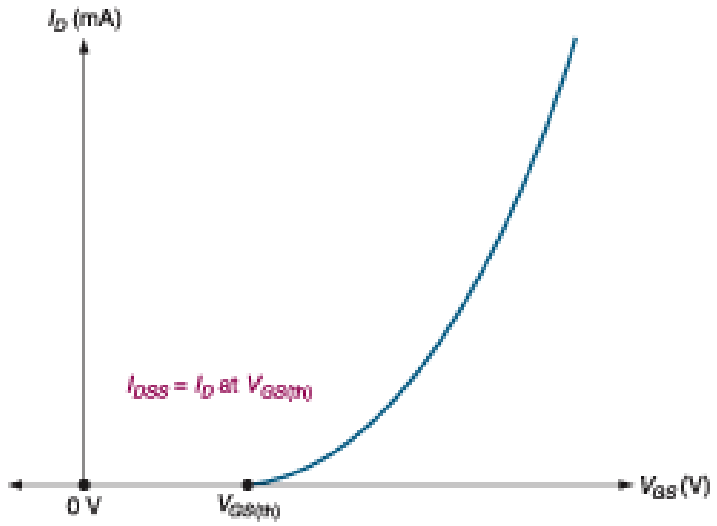


nMOS

Here the transistor is said to be saturated and hence this region is referred to as the *Saturation Region.* It is oftentimes also called the *Active Region.* The reason for this is that, when used as an amplifier, the transistor operates (or is active) in this region. The current in the saturation region is roughly constant for a given V_{GS} and is known as the *Saturation Current* $\left(I_{D(sat)}\right)$ – or sometimes just referred to as the *'On Current'* ($I_{D(on)}$) because the transistor is ON in this region for an amplifier.

Do note however, that for logic gates the transistor is not chosen to be ON in this region. Rather, as we shall see later, it is set to be ON in the Ohmic Region and OFF in the Cutoff Region.

Next is the transconductance curve which is the relationship between the ***Saturation Current*** ($I_{D(sat)}$ or $I_{D(on)}$) and the ***gate-source*** voltage (v_{GS}).



The curve is a parabola with a vertex at $V_{GS(th)}$. The equation for this parabola is:

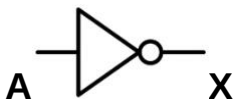
$$I_{D(on)} = K\left(v_{GS} - V_{GS(th)}\right)^2$$

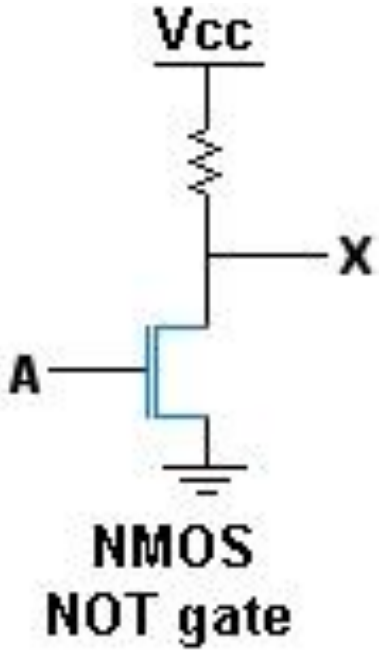
K is a constant that depends on the particular type of MOSFET being used. Data sheets usually give the coordinates for one point on the transconductance curve. This allows you to solve for K using the above equation.

Using transistors to build logic gates:

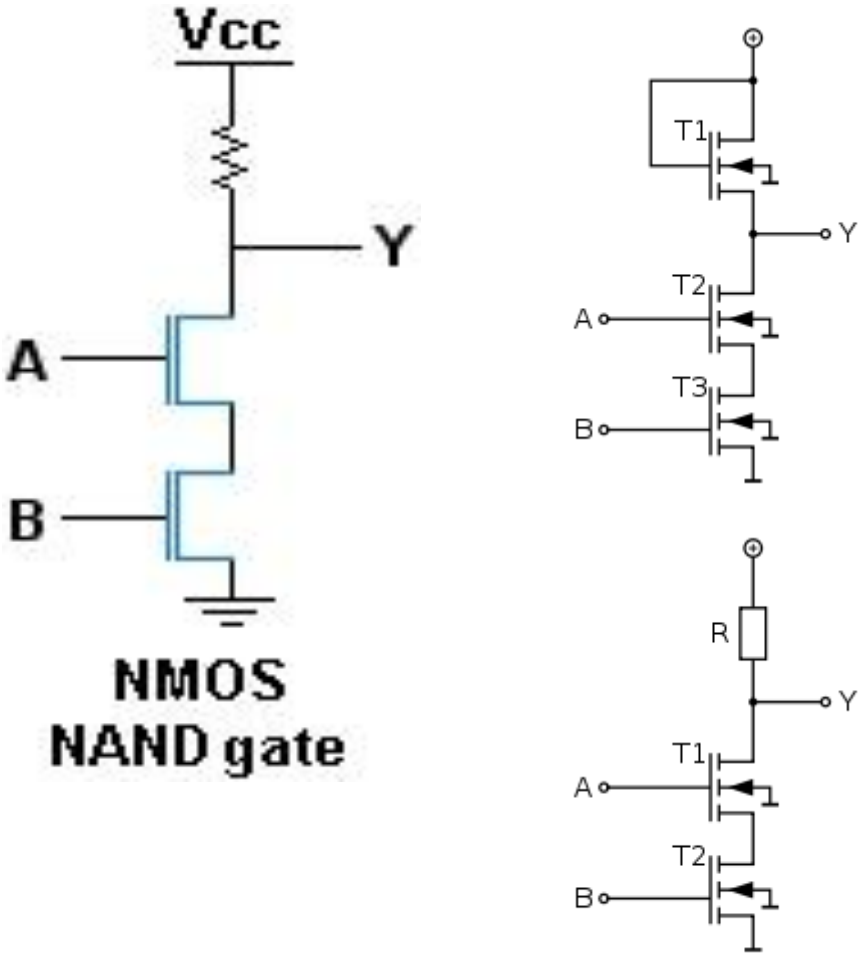
As you can see, from the diagrams below, transistors can be used to build logic gates. The principle of operation is fairly simple. Let's start with the inverter or NOT-gate. If the input is HIGH (i.e. a high voltage) then the transistor is switched ON (current flows from the drain to the source). If the resistor at the drain is chosen to be much greater than the 'ON' resistance of the transistor then the output will be LOW (i.e. a low voltage) – recall the Potential Divider! If the input is LOW then the transistor is OFF. There is no drain current and the output voltage is equal to the supply voltage i.e. HIGH. It is left to the student to figure out how the other logic gates work.

A	X = NOT A
0	1
1	0

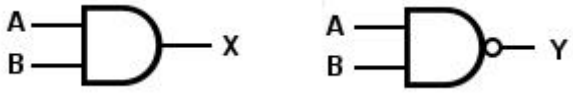


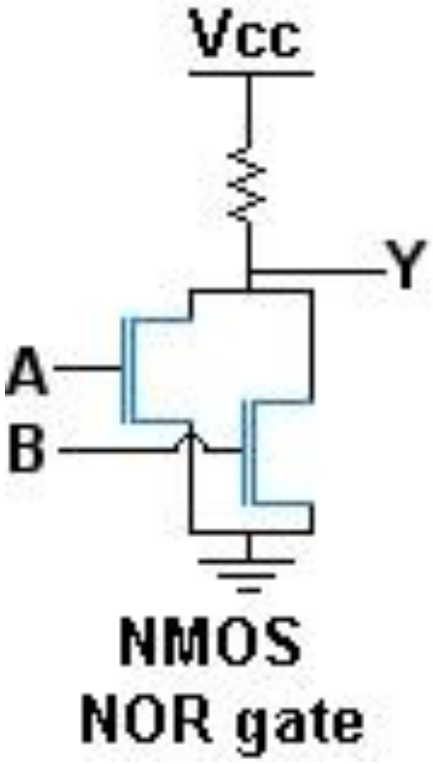


You will note here with the NAND gate that two equivalent circuits are given.
A transistor has a resistance albeit not being an Ohmic device. Since a transistor occupies about 20 times less space on an IC as a resistor it is common to use transistors as ‘resistors’.



A	B	AND: X = A AND B	NAND: Y = NOT(A AND B)
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0





A	B	OR: X = A OR B	NOR: Y = NOT (A OR B)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

Staying with the inverter, one notes that the output voltage when a HIGH voltage is applied at the gate is that of a potential divider.

$$V_{out} = \frac{V_{DD}R_{ON}}{R_D + R_{ON}}$$

- where R_D is the drain resistance and R_{ON} is the 'ON' resistance of the transistor. When we apply a voltage to the gate we want the output to be as close to zero as possible (within certain constraints as will be seen later). This means setting $R_{ON} \ll R_D$. This is achieved by appropriate choice of transistor (we will see later how to set R_D).

Setting $R_{ON} \ll R_D$ also means that the effect of variable resistance of the transistor is minimised giving a better transition curve (from logic 1 to 0 and vice- versa). When a zero voltage exists at the gate the transistor is OFF and the output voltage equals the rail voltage - V_{DD} (FETS) or V_{CC} (Bipolar). In effect what one is doing here is operating the transistor in the Cutoff and Linear (Ohmic) regions for small and large V_{GS} respectively. That is: for a LOW input V_{GS} $(V_{in}) < V_{GS(th)}$ we are operating in the Cutoff Region and for a HIGH input $V_{GS} \approx V_{DD}$ we are operating in the Linear Region. So for logic gates there are two operating points, Linear and Cutoff, corresponding to HIGH and LOW inputs respectively at the gate. For this reason, in Digital Logic, the transistor is said to act as a switch. It is OFF (no current flow) when the input is LOW and ON (current flow) when the input is HIGH.

Biasing the Transistor when used as a Switch

The question now arises as to precisely what voltages (and consequently what resistance values) ought to be applied to the transistor. This process is referred to as *biasing* the transistor. Put another way, *biasing* means applying the correct dc operating voltages for the semiconductor device in question (recall we had Forward and Reverse Bias for diodes) such that it operates as desired.
To this end we employ a very useful concept known as the 'DC Load Line' (there is also an AC Load Line but we won't concern ourselves with that here). The Load Line is simply the relationship between V_{DS} and I_D ***when the transistor is connected into the circuit!*** This relationship is obtained simply from Kirchoff's Voltage Law:

$$V_{DD} = I_D R_D + V_{DS}$$

or

$$I_D = \frac{-V_{DS}}{R_D} + \frac{V_{DD}}{R_D}$$

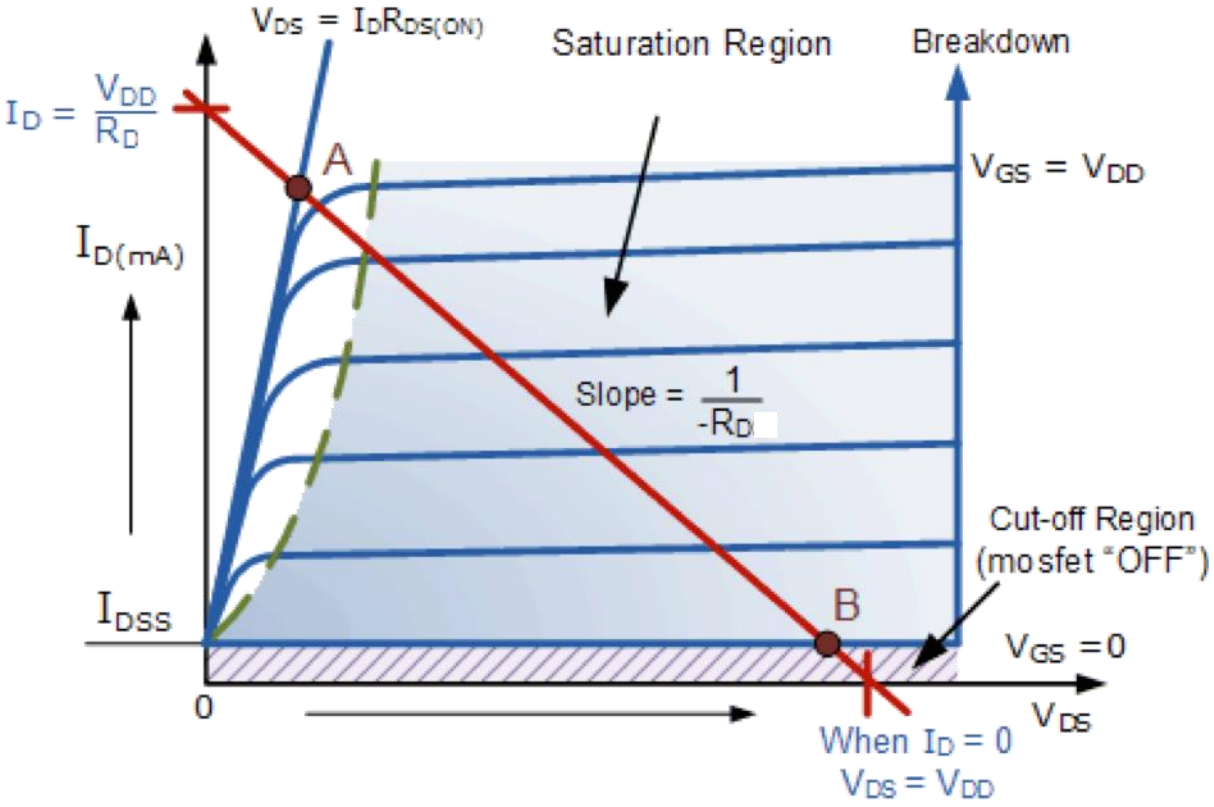
- which is a line with a slope of $-\frac{1}{R_D}$.

Now we have both a relationship between V_{DS} and I_D expressed as a set of drain curves (***regardless of whether the transistor is connected into the circuit or not***) and as a Load Line (***for when the transistor is connected into the circuit***). When the transistor is connected into the circuit both curves must hold true which means that the operating point (V_{DS} and I_D) is somewhere at the intersections of the load line and drain

curves. Which drain curve(s) exactly is (are) determined by one's choice(s) of V_{GS}.

The following is a plot of the Load Line superimposed on the drain curves. You can see that you can control the slope of the Load Line and thus its intersection with the drain curves by varying V_{DD} and R_D.

We will see in the next section how to choose V_{DD}, V_{GS} and R_D.



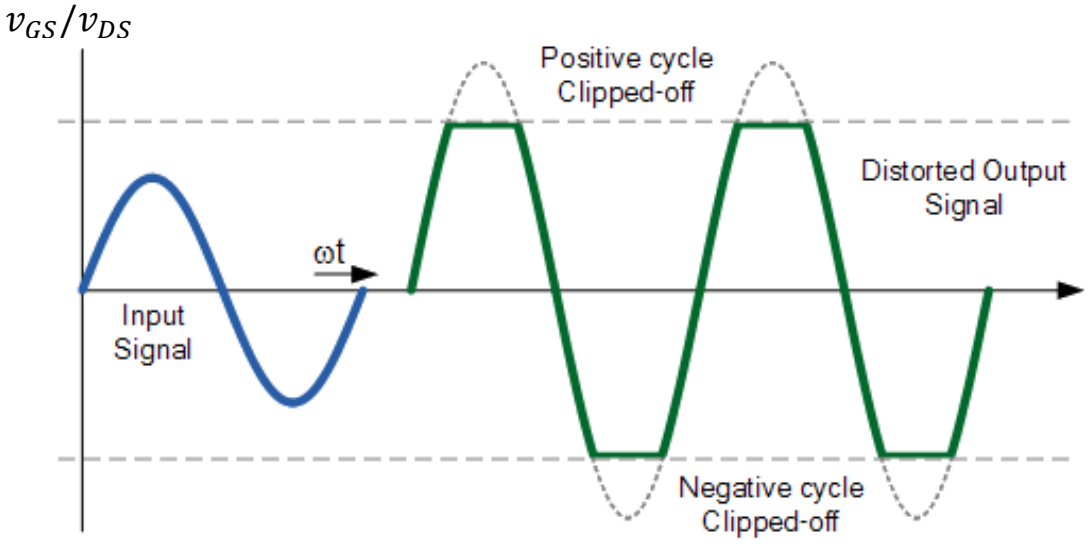
Using transistors to build amplifiers:

It is somewhat more complicated to use a transistor(s) to build amplifiers. Amplifiers are common in electronic systems but probably their most important function in relation to Computer Science is in

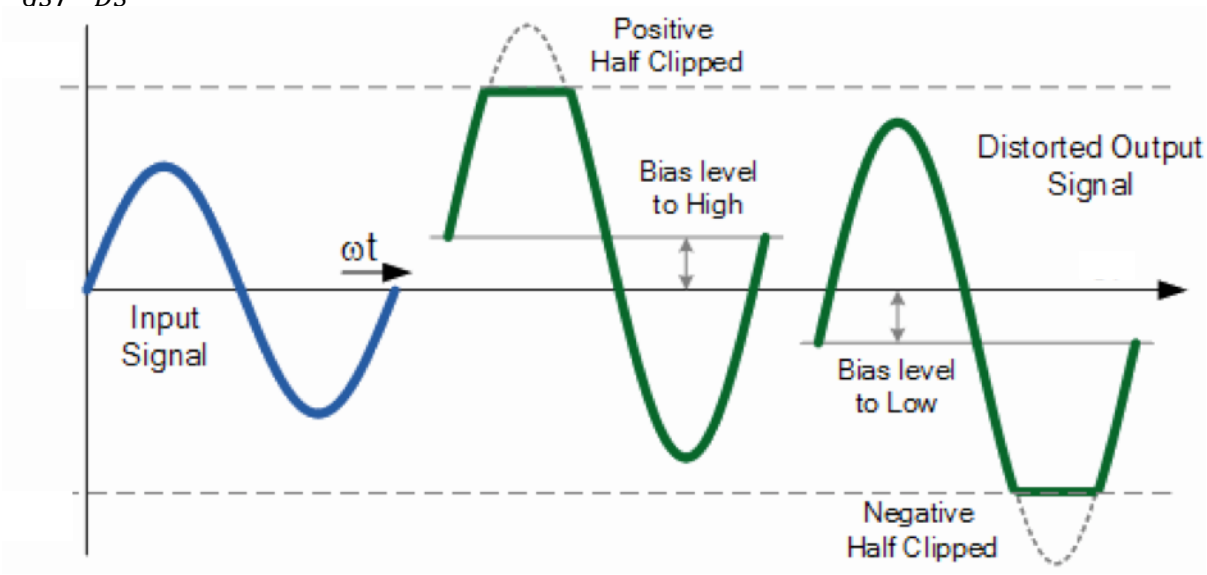
telecommunications where a received wireless signal is amplified. This amplification is necessary to separate out the desired signal from all others and to make it large enough to process (and, for example, to convert to audible sound on phones etc.).

Biasing the E-MOSFETs for Amplifiers:

There are a number of ways by which the enhancement-type MOSFET can be biased when used as an amplifier. We will look at two here, namely, Voltage Divider Bias and Drain-Feedback Bias. As mentioned earlier, the first consideration in biasing an enhancement-type MOSFET is, that when used as an amplifier (as opposed to a switch), we have to ensure that it is on at all times. This means setting V_{GS} to be sufficiently greater than $V_{GS(th)}$ such that variations in the input voltage (v_{GS}) will not cause the MOSFET to switch off while ensuring that V_{GS} is low enough such that operation is not pushed into the Ohmic Region where there will be no further decrease in v_{DS} with increased v_{GS}.
Operation in the Cutoff or Ohmic Regions will result in clipping of the signal. In the Cutoff Region the transistor will simply switch off and the output will be V_{DD}. In the Ohmic Region the output voltage will take on a constant low value.

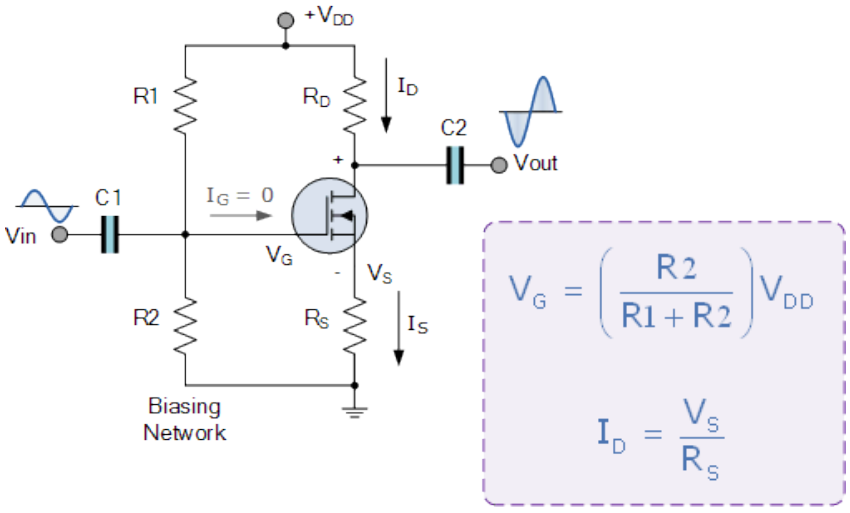


v_{GS}/v_{DS}



Apart from the reasons just given, the operating point for the transistor should be in the Active Region where there is there is the greatest variation in the output voltage v_{DS} versus the input voltage v_{GS}. In other words in this region we get gain. Let us first consider voltage divider bias:

Voltage Divider Bias:



The potential divider formed by R_1 and R_2 provides a dc gate voltage in excess of $V_{GS(th)}$. This ensures that the transistor is on even for negative swings in the input voltage v_{GS}. The input voltage at the gate, v_{GS}, is the sum of the ac input signal and the dc biasing voltage. In other words the signal at the gate is an ac signal with a dc component. If the dc component is high enough the transistor will not switch off in negative swings of the input because the dc component prevents it from doing so. The capacitor C_1 at the gate allows the ac signal through while preventing dc to the source. In other words the capacitor serves to isolate the dc bias of the circuit. Ditto for capacitor C_2. The capacitor values are chosen such that they provide negligible reactance at the frequencies of operation.
The varying voltage at the gate (v_{GS}) causes a varying current flow, i_D, through the transistor. This current also flows through the drain resistor (R_D). Hence a varying voltage appears at the output ($v_{DS} = V_{DD} - i_D R_D$). Hence the output voltage, v_{DS}, can vary from near 0 to V_{DD}. If the supply voltage at the rail (V_{DD}) is greater than the peak voltage at the input then we get same signal at the output as we have at the input albiet with a greater amplitude. ***It is in this way that amplification is achieved with a transistor!***. The gain or amplification is given by:

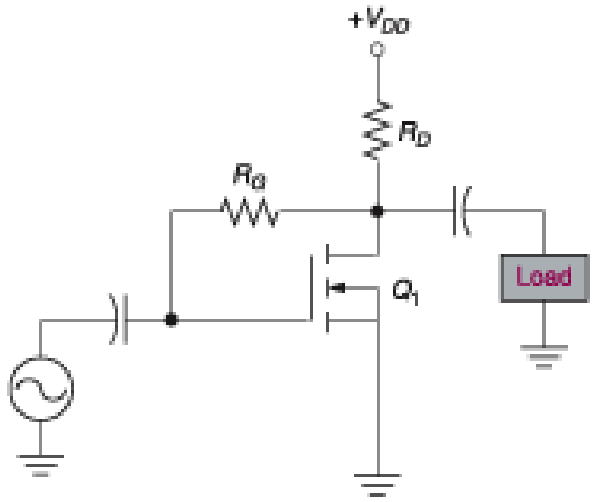
$$G = \frac{v_{out}}{v_{in}}$$

It is also of imortance to note that the output for this amplifier design is inverted wrt the input! - since it is based on an inverter circuit.The solution to this problem is to pass the output through another *amplifier stage.* The source resistance is not absolutely necessary but serves to stabilise the amplifier (keeping its gain constant) in the event of small variations in the rail voltage (V_{DD}). If for some reason V_{DD} increases slightly then i_{DS} *increases*. This means that the voltage across the source resistance increases (since i_{DS} flows through it). The result of this is that the gate-source voltage decreases which has the effect of *reducing* i_{DS} so keeping the gain relatively constant. The same reasoning can be applied for V_{DD} decreasing. Sometimes you will see a capacitor in parallel with the source resistance. The idea is to allow ac to bypass the source resistance which provides stabilisation as described

with its dc voltage while not reducing amplification of the ac signal (the source resistance on its own serves to reduce the ac gain of the amp').

Drain Feedback Bias:

Since resistors take up space on an IC another biasing method known as drain-feedback bias is used which reduces the resistor count.



Here the drain is connected to the gate. There is no resistance at the source. Stabilisation is provided by the gate resistance (figure out how!). Since there is no gate current and no dc flow through the input capacitor, there will be no dc voltage drop across R_G (which has a very high value e.g. $50M\Omega$) and so:

$V_{GS} = V_{DS}$ - for Drain-Feedback Bias

In other words the dc equivalent circuit has the drain and the gate shorted. As before, the ac input signal, v_{GS}, 'sits' on the dc gate-source voltage. That is, the signal at the gate has both an ac and a dc component which can be treated seperately (by Superposition). For the ac signal, since the gate-drain resistance is very high there is practically zero current and so it can be replaced with an open circuit. The drain

feedback provides the bias at the gate to ensure the transistor is always on i.e. $v_{GS} > V_{GS(th)}$ always. Of course this will also depend on the supply voltage V_{DD} and the resistor R_D both of which determine V_{DS} $(= V_{GS})$.

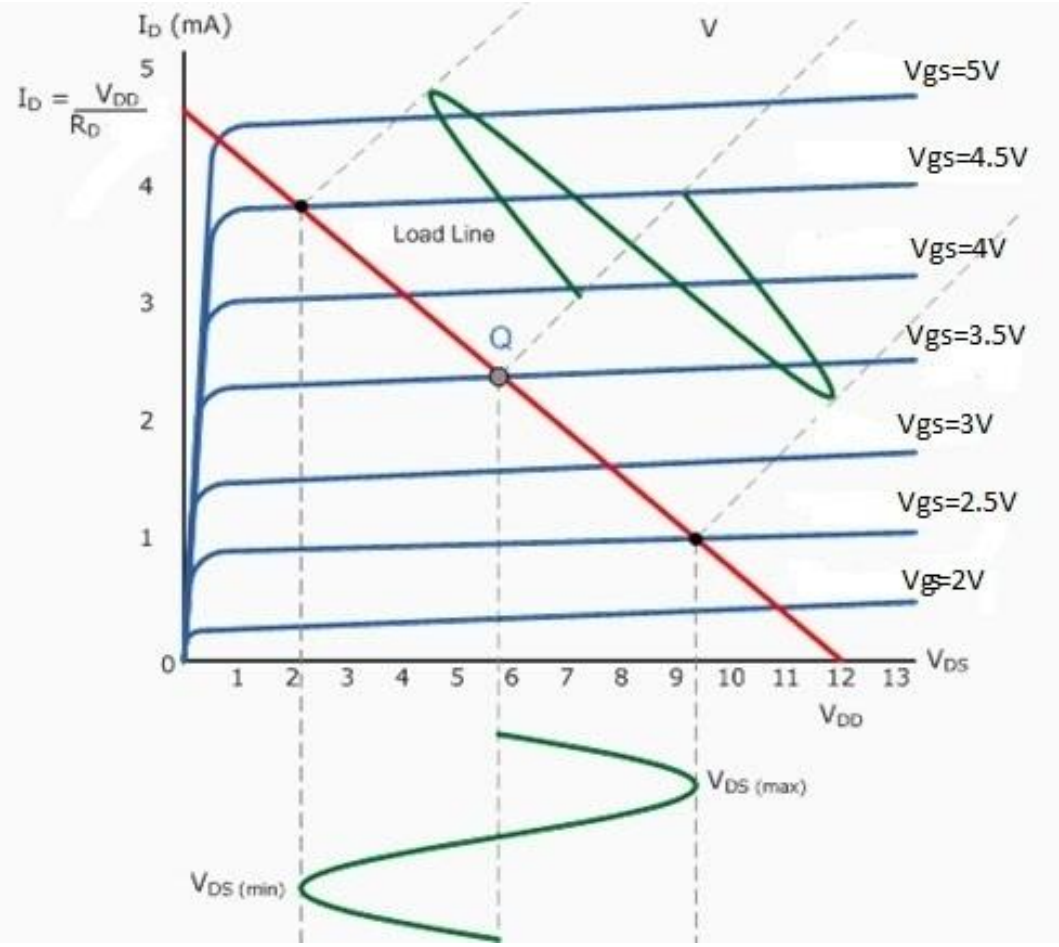
We shall see presently how to choose the bias resistance values for both voltage divider and drain feedback bias.

Choosing the Values of the Biasing Resistors

As before the dc load line is obtained via Ohm's Law and relates i_D and v_{DS}:

$$V_{DD} = v_{DS} + i_D R_D$$

For simplicity we have ignored the source resistance in voltage divider bias (in which case $R_D \rightarrow R_D + R_S$). As before a plot of the load line is then superimposed on the drain curves.



This time the dc operating point, often referred to as the Q-point (quiescent point) of the circuit is chosen to lie in the middle of the active region. There are three reasons for this:

(i) In the Active Region we can get a variation in v_{DS} versus v_{GS}, i.e. we get gain.
(ii) We wish to avoid the transistor going into cutoff where we get clipping at the upper end of the output signal.
(iii) We wish to avoid the transistor going into the linear region of operation where we get clipping at the lower end of the output signal.

You can see from the plot that the operating point moves up and down the Load Line about the Q-Point such that the gate-source voltage can vary without the MOSFET switching off or moving into the Linear Region.

One gets a consequent variation in v_{DS} potentially varying from 0 to V_{DD} - which is an amplification of v_{GS}.

To bias the transistor correctly we do the following:

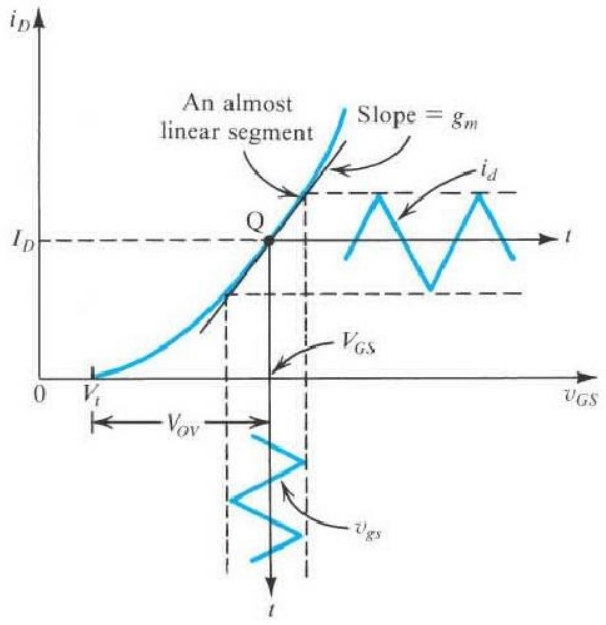
1) Note the range of the input v_{GS} expected.
2) Choose the drain curve corresponding to middle (average) of this v_{GS} range - V_{GSQ}.
3) Choose V_{DD} - usually a given. This gives the point of intersection of the Load Line with the horizontal (v_{DS}) axis.
4) Choose the quiescent point to be at V_{GSQ} roughly midway along the load-line.
5) Choose R_D to be equal to $\frac{V_{DD}-V_{DSQ}}{I_{DQ}}$.
6) In the case of Voltage Divider bias, choose two (fairly large – so as to minimise power dissipation) resistors, R_1 and R_2, to give V_{GSQ} at the gate.
7) In the case of Drain Feedback Bias choose a very large gate resistance (e.g. 50MΩ).

Usually however the datasheet that comes with the transistor will provide biasing information and maximum ratings so that some of this design work is already done for you.

If you want to increase/decrease amplification you must either increase/decrease V_{DD}. Amplification comes at a price however since the power consumption will increase. Typically one tries to keep i_D (and consequently the power dissipation) as low as possible for a given application. However if you want greater amplification you must increase the power.

The operation of a transistor can also be observed on the Transconductance Curve. The output is then, as before,

$$v_{DS} = V_{DD} - i_D R_D.$$



What is notable here is that there is only an *approximately* linear relationship between the output and the input. The greater the amplification the more noticable this non-linearity becomes. Good amplifier design includes good choice of transistor (usually more

expensive) and other components such that this non-linearity is not noticable. However non-linearity cannot be eliminated entirely.

Problem:

For a drain feedback E-MOSFET amplifier the data sheet specifies $I_{D(on)} = 3mA$ for $V_{DS(on)} = 10V$. If $V_{DD} = 25V$ select a value of R_D that allows the MOSFET operate at the specified Q-point.

Solution:

$$v_{DS} = V_{DD} - i_D R_D$$

$$R_D = \frac{25V - 10V}{3mA} = 5k\Omega$$

Problem:

For an n-channel E-MOSFET the manufacturer specifies $V_{GS(th)} = 4V$ and $I_{DS} = 7.2mA$ at $V_{GS} = 10V$. For Drain Feedback Bias with $V_{DD} = 24V$ and $R_G = 100M\Omega$ specify R_D for operation at $V_{DS} = 8V$.

Solution:

Given:

$$i_{D(on)} = K\left(v_{GS} - V_{GS(th)}\right)^2$$

Then:

$$i_{D(on)} = 0.0072 = K\left(v_{GS} - V_{GS(th)}\right)^2 = K(10-4)^2$$

$$\therefore K = \frac{0.0072}{6^2} = 0.0002\ A/V^2$$

Because $i_G = 0$ there is no dc voltage drop across the resistor R_G.

Hence $v_{GS} = v_{DS} = 8V$ and $i_{D(on)} = 0.0002(8-4)^2 = 3.2mA$.

But:

$$v_{DS} = V_{DD} - i_D R_D$$

Solving for R_D gives $R_D = 5k\Omega$

Problem:

For a Voltage Divider Bias configuration using a Depletion MOSFET (which operates for V_{GS} positive and negative) with $V_{DD} = 16V$, $V_D = 12V$ $V_{GSQ} = -2V$, $R_1 = 91K\Omega, R_2 = 47K\Omega, R_D = 1.8K\Omega$ find R_S.

Solution:

$$V_G = \frac{47.16}{47 + 91} = 5.44V$$

$$I_D = \frac{V_{DD} - V_D}{R_D} = 2.22mA$$

Now:

$$V_{GS} = V_G - I_D R_S$$

$$=> R_S = 3.35K\Omega$$

The nearest commercial value is $3.3K\Omega$

Problem:

For a drain feedback biased E-MOSFET we are given the following information: $V_{DS} = \frac{1}{2}V_{DD}$, $I_{D(sat)} = 4mA$ and $V_{GS} = 6V$.
Find V_{DD} and R_D.

Solution:

$$V_{GS} = V_{DS} = \frac{1}{2}V_{DD} = 6V$$

$$=> V_{DD} = 12V$$

$$R_D = \frac{V_{DD} - V_{DS}}{I_D} = 1.5k\Omega$$

Problem:

For an E-MOSFET with Voltage Divider Bias we are given the following information:

R_1, R_2 = $1M\Omega$; $R_D, R_S = 6K\Omega$; $V_{DD} = 10V$; $V_{GS(th)} = 1V$ and

$K = 0.5mA/V^2$.

We may assume that the MOSFET is operating in the Saturation Region as these are recommended values.

Find V_{GS}, V_{DS} and I_D.

Solution:

$$V_G = \frac{R_2}{R_1 + R_2}V_{DD} = \frac{1}{2}V_{DD} = 5V$$

Then:

$$V_{GS} = V_G - V_S = V_G - I_D R_S = 5 - 6I_D$$

The drain current can be computed thus:

$$I_D = K(V_{GS} - V_{GS(th)})^2 = 0.5(5 - 6I_D - 1)^2 = 36{I_D}^2 - 50I_D + 16 = 0$$

This yields two potential solutions for I_D:

$$I_D = 0.89mA$$

and

$$I_D = 0.5mA$$

To determine which of the solutions is correct we compute the gate-source voltage for each:

For $I_D = 0.89mA$, $V_{GS} = 5 - 6I_D = -0.34V$ and for $I_D = 0.5mA, V_{GS} = 5 - 6I_D = 2V$. Since $V_{GS} > V_{GS(th)}$ for operation in the saturation region then:

$$I_D = 0.5mA, \qquad V_{GS} = 2V$$

The corresponding drain voltage is then:

$$V_D = V_{DD} - I_D R_D = 10 - 6I_D = 7V$$

Hence:

$$V_{DS} = V_D - V_S = V_D - I_D R_S = 7 - 3 = 4V$$