Solutions Summer 2010

Question 1:

a) With the aid of diagrams describe the gate and channel region of a modern CMOS transistor, explain how elements of this device differ from a CMOS device of greater than 1.0μm channel length. Modern devices can be considered to be those in processes from 90nm down to 45nm. Explain the reasons for these process changes.

Older Device
The “stack” largely consists from the bottom up;
Standard <100> silicon lightly doped P type or N type
Gate dielectric invariably pure silicon dioxide
N doped polysilicon as the gate electrode
The S/D on either side may or may not have some drain engineering such as double diffused (DDD), lightly doped drain (LDD)

On a modern device;
The substrate is usually strained silicon or silicon germanium this could also be on and SOI (silicon on insulator substrate.
The gate dielectric down to 90nm is usually nitrided silicon dioxide
In processes below 60nm the gate dielectric is now a high k dielectric such as hafnium dioxide.
The gate electrode down to and including is polysilicon doped N-type over the N-channel devices and P-type over the P-Channel devices. The polysilicon is also silicided, that is the top layer is reacted with some refractory metal such as titanium, nickel or cobalt.
The sides of the gate stack usually have either nitride or oxide spacers, this facilitates the silicidation process as well as the formation of the engineered drains. Also a layer of nitride can be used to induce stress in the substrate to facilitate enhanced mobility in the N-Channel devices

The most recent processes have gone back to metal gates instead of polysilicon/silicide. This is implemented by using poly as a sacrificial layer to form the device structure, then etched out and replaced by a metal.

The description should give a reason for all of the changes mentioned.

b) If a 60nm CMOS process has 1.2nm of silicon dioxide as the gate dielectric, calculate the gate capacitance and also calculate the thickness of dielectric that would be used if the material was changed from silicon dioxide to Hafnia ($Hf0_2$) whilst maintaining the same capacitance per unit area.

Given:
The permittivity of free space is 8.86×10^{-14} F/cm
The dielectric constant of Hafnia ($Hf0_2$) is 25

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

$$C_{ox} = \frac{k_{0x}\varepsilon_0}{t_{ox}}$$

$$C_{ox} = \frac{3.9 \times 8.886 \times 10^{-14} F/cm}{1.2 \times 10^{-7} cm}$$

$$C_{ox} = 2.88 \times 10^{-6} F/cm^2$$

$$t_{Hik} = \frac{k_{Hik}\varepsilon_0}{C_{ox}}$$

$$t_{Hik} = \frac{25 \times 8.886 \times 10^{-14} F/cm}{2.88 \times 10^{-6} F/cm^2}$$

$$t_{Hik} = 7.69 \times 10^{-7} cm$$

$$t_{Hik} = 7.69 nm$$

Question 2:

a) In thermal oxidation of silicon explain why the initial growth rate is faster that the growth rate after a significant oxidation time.

Initially

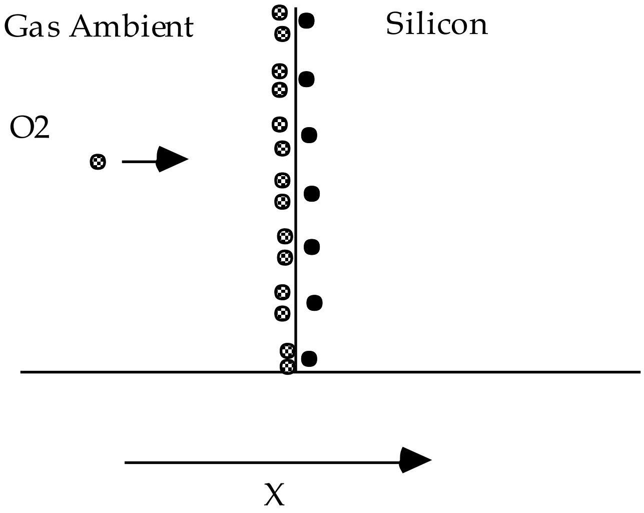


Diagram 1

- Reaction controlled
- How fast the chemical reaction between the oxidizing species and the silicon can take place
- Growth is linear, double the thickness for double the time
- There is no barrier to the oxidation process

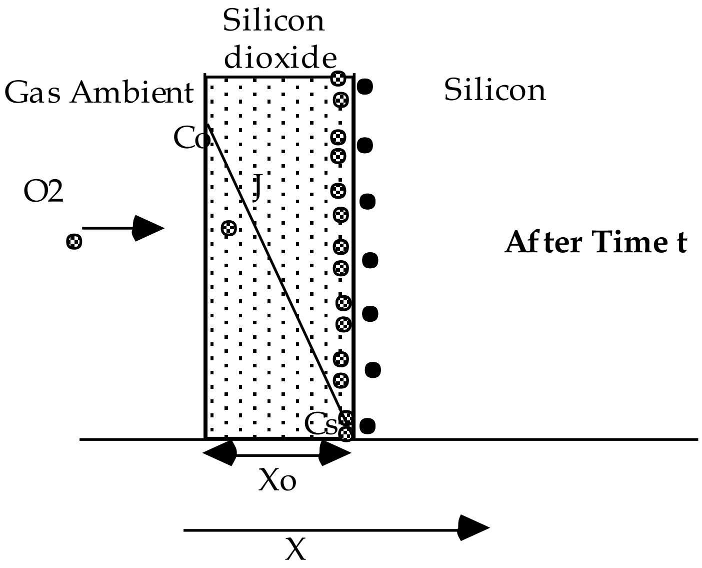


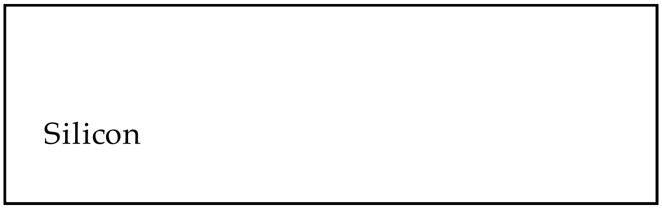
Diagram 2

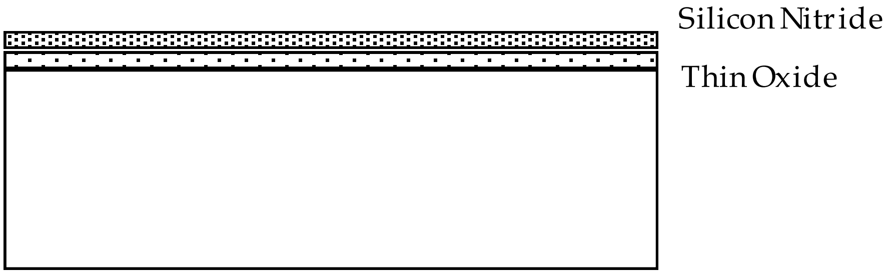
- There is now oxide on the surface of the silicon
- Oxidizing species must diffuse through the oxide to reach the silicon
- As the oxide gets thicker this takes longer
- Growth rate becomes dominated by the diffusion time rather than by the reaction time
- Growth rate parabolic
- As time progresses the growth rate slows down because of the increased diffusion time

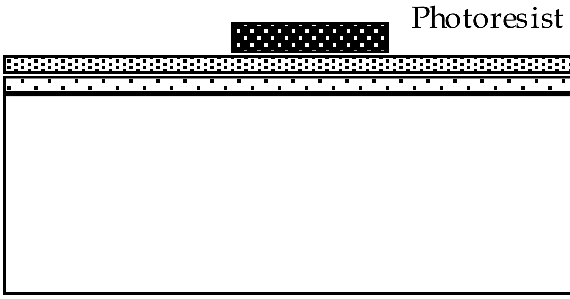
b) Explain the function of the silicon nitride layer in the formation of a LOCOS or Semi-Recessed field oxide.

The area to be designated the active region of a LOCOS processed wafer is defined by patterning a thin layer of silicon nitride over what is to be the active area. The nitride is deposited by LPCVD technique onto a thin layer of thermally grown silicon dioxide. The sharpness or slope of what in the final profile is known as the “birds beak” region is defined by the ratio of the oxide to nitride thickness, typically the nitride is in the region of twice as thick as the oxide.

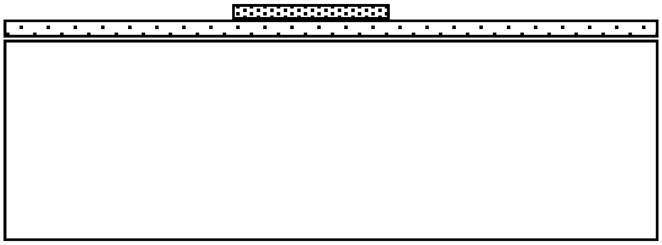
LOCOS or Semi Recessed Oxide



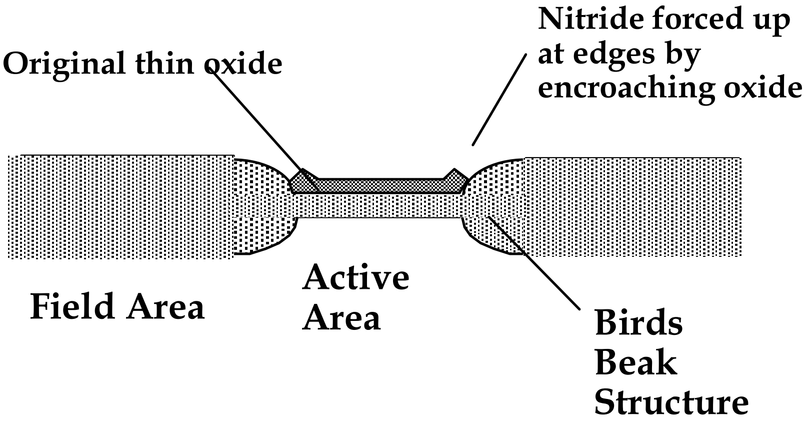




Etch Nitride and
Remove
the Resist



The wafers are then put through a “wet” oxidation process. During this process the oxidising species (water vapour) can reach the silicon surface through the thin oxide to combine with the silicon to form the new thick layer of silicon dioxide. During this oxidation process the silicon surface is being consumed to form the oxide. The silicon nitride protects the active areas, preventing the water vapour from getting through to the silicon surface; this means that there is no consumption of silicon in this area. So the function of the silicon nitride is to prevent the oxidising species getting through to the silicon surface and thus prevent the oxidation of the silicon in these regions. The thick oxide areas then have a partially recessed profile when compared to the active, the overall thickness of the field oxide in the region may be of the order of 1.0um (typical) but the step height from the top of the oxide to the silicon active area will only be about half of that. The reason for this is; because during the oxidation the silicon surface is consumed to form the silicon dioxide.



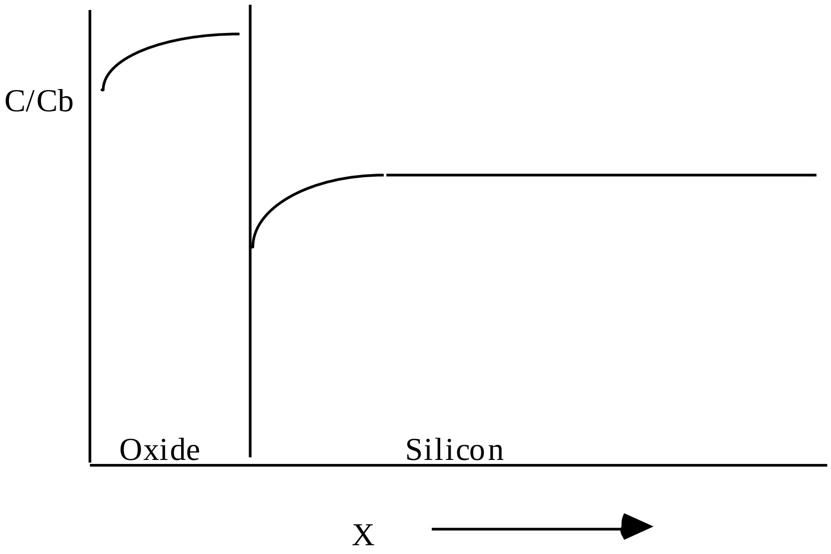
c) Explain why, particularly in P-doped field regions, it is normal to bring up the boron doping concentration under the field oxide by having a field implant prior to oxide growth?

Because most impurities diffuse much slower in silicon dioxide than in silicon we expect that an oxide grown on the surface will effectively seal the impurities in the silicon. The situation however is much more complex!! With any two phases in contact, any impurity will be redistributed between the two phases, until equilibrium is reached. In equilibrium the ratio of the concentrations will be constant.
The ratio of the equilibrium concentrations in the silicon and silicon dioxide is denoted by the term segregation coefficient and is defined by;

$$M= \frac{\text{Equilibrium conc. of impurity in silicon}}{\text{Equilibrium conc. of impurity in silicon dioxide}}$$

When oxide takes up the impurity M<1

Boron has a segregation coefficient of less than 1, in field regions it is important to have the concentration high to prevent the turn on of parasitic MOS transistors (surface inversion). So when a thick field oxide is grown over boron doped regions some of the boron is "sucked out" depleting the doping concentration under the oxide. It is normal to compensate for this loss with a field implant.



d) A silicon <100> wafer, which had been patterned and etched with bare silicon in the patterned windows and 0.4μm in the other areas to protect against the implant, is put through a thermal oxide process at 1000°C in pyrogenic steam for 2 hours 10 minutes, what is the final oxide thickness in:

i) The areas which had no oxide on the surface prior to oxidation?

This is a simple read off from the plots of thickness versus oxidation time for pyrogenic steam:

0.6μm

ii) The areas which had 0.4μm on the surface prior to oxidation?

In the areas where there is already 0.4μm on the surface this thickness must be converted to the time it would take to grow that thickness at 1000C in steam, this is 70 minutes. This is then added to the actual oxidation time of 2 hours 10 minutes or 130 minutes, giving 200 minutes. The resultant thickness in these areas is the equivalent of a 200 minute oxidation or 0.8μm.

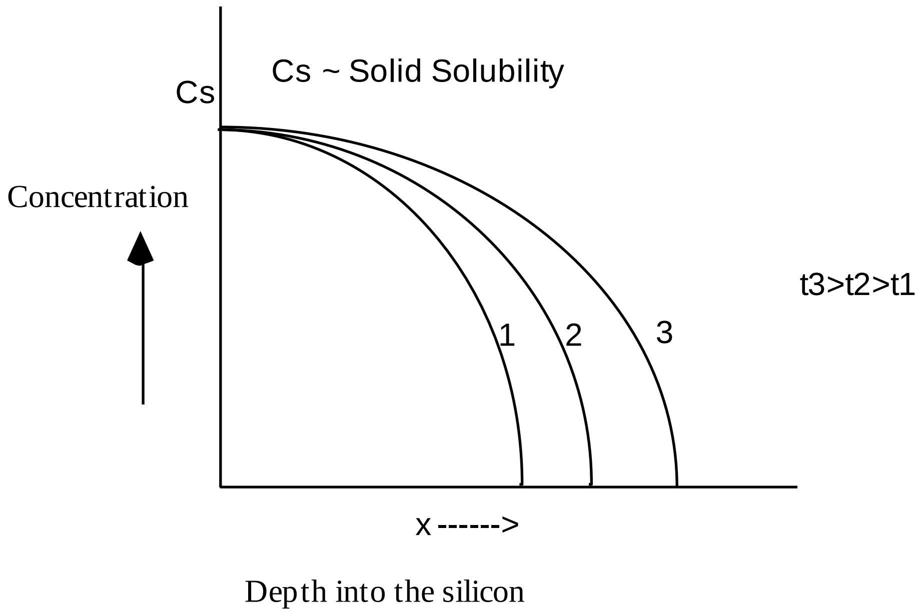
Question 3:

a) Explain why a CMOS process needs ion implant technology and cannot be fabricated using simple furnace doping techniques.

In, for example a CMOS process fabricated on a P-Type substrate it is easy to form the N-Channel transistors. N+ doped regions are diffuse into the substrate to form the Source and drain of the transistor.
On the other hand it is more difficult to form the P-Channel transistors, this would involve P+ regions sitting in an N-"Substrate." To realise this a deep lightly doped N region must be implanted into the P substrate and in this "well" of N- the P-channel transistors can be fabricated. The problem is that the surface doping or the doping in the channel has to be quite light so that the threshold voltage is at a reasonable value, usually the same absolute value as the N Vt but the opposite polarity. Doping regions with this light a concentration can only be formed using implantation. There is not enough control of the doping concentration using thermal doping techniques to enable this lightly doped well. In thermal doping the surface of the silicon goes to the solid solubility level of the dopant in silicon at the process temperature, for most typical predeposition temperatures this is of the order of 10^{20}-10^{21} atoms/cm^3 whereas what is needed for the well is of the order of 10^{15}-10^{16} atoms/cm^3 This can only be achieved with implantation of dopant.

b) Compare the doping concentration versus depth into silicon for pre-deposition and drive-in for different furnace times; explain the reason(s) for the differences.

Pre-Deposition or Constant Source Diffusion :
During constant source diffusion there is a continuous supply of dopant to surface of the silicon. The wafers are in an ambient in which the dopant concentration is kept very high from solid, liquid or gaseous sources. This keeps the surface concentration constant throughout the diffusion, usually at the solid solubility level of the dopant species in silicon at the process temperature. As time progresses the dopant diffuses deeper into the silicon but the surface concentration remains constant.



The boundary conditions for this type of diffusion are:
Initial Condition at : t = 0 → $C_{(x,o)} = 0$
It is assumed that at the start of the process there is no dopant in tha silicon already.
Boundary Conditions are :
1. $C_{(o,t)} = Cs$

The surface concentration almost immediately goes to the solid solubility level of the dopant in the silicon at the process temperature and remains at that level throughout the cycle regardless of time.

2. $C_{(\infty,t)} = 0$
Very deep into the silicon there is no dopant the equations used to describe this are for a process occurring close to the surface.
The equation that satisfies these conditions is

$$C_{(x,t)} = C_s erfc\left[x \big/ 2\sqrt{Dt} \right]$$

Where erfc is the complimentary error function.
This means the total dopant quantity in the silicon is increasing with time.
The total number of dopant atoms per unit area is given by;

$$Q_{(t)} = \int_o^\infty C_{(x,t)} dx$$

$$Q_{(t)} = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt}$$

$$Q_{(t)} = 1.13 C_s \sqrt{Dt}$$

The quantity $Q_{(t)}$ represents the area under one of the diffusion profiles and obviously gets larger as time increases.

Drive-in or Constant Total Dopant Diffusion or Limited Source Diffusion :
A fixed amount of dopant is deposited on the silicon surface during a pre-deposition stage. There is a fixed amount of total dopant available throughout the duration of the drive-in stage. There is no other dopant available during this phase of the process This limited dopant quantity diffuses deeper into the silicon as time increases. As the dopant goes deeper in the silicon the surface concentration reduces while the total dopant quantity remains the same.

The boundary conditions are as follows:
Initial Conditions at t = 0 → $C_{(x,o)} = 0$
This initial condition considers that the dopant has no depth into the silicon at the start of the drive-in. This is often called a delta or spike function

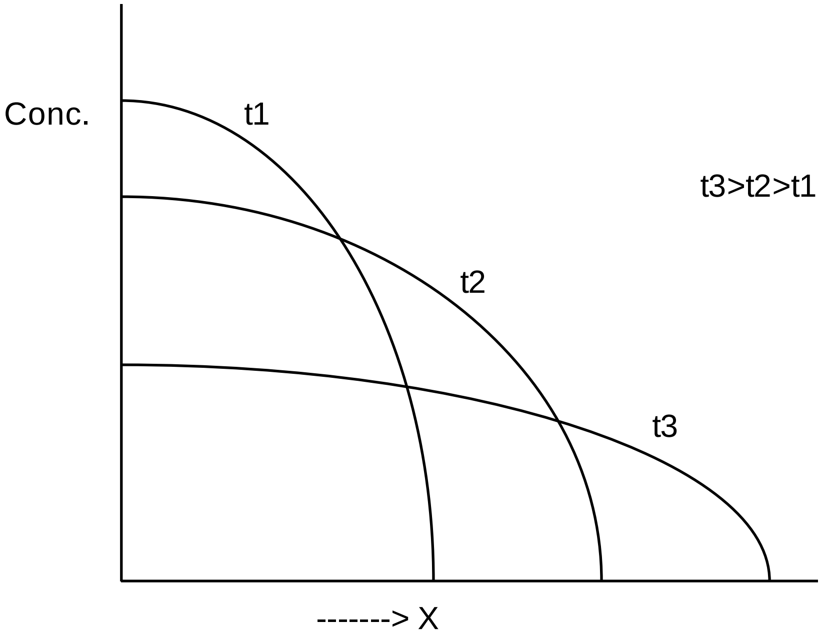
Boundary Conditions:
1. $\int_0^\infty C_{(x,t)} dx = S$
The area under the curve remains the same for different diffusion times. Total dopant quantity is always equal to S, a constant
2. $C_{(x,\infty)} = 0$
After an infinite time the effective concentration is of dopant in the silicon is 0

These conditions are satisfied by the equation below;

$$C_{(x,t)} = \frac{S}{\sqrt{\pi Dt}} \exp\left[\frac{-x^2}{4Dt}\right]$$

by setting x = 0 we can get the surface concentration

$$C_s = C_{(0,t)} = \frac{S}{\sqrt{\pi Dt}}$$



As time progresses the same (constant) quantity of dopant is spread over an increasing depth so as the dopant goes deeper the surface concentration reduces.

c) If Phosphorus at a dose of 1 x 10^{16} ions/cm^2 is implanted into a boron doped wafer with a substrate doping of 1 x 10^{15}atoms/cm^3 through an oxide of thickness 170nm so that the peak of the implant sits at the interface between the silicon and the oxide.

i. Calculate the energy at which the implant is carried out

ii. What will the resulting junction depth be if given a heat treatment that activates the dopant but causes no further dopant diffusion.

iii. How much oxide will be needed to block the implant in the areas that are not to be doped?

Projected Range and Projected Standard Deviation Graphs attached.

The concentration equation $N(x) = Np \exp[-(x-Rp)^2/2\Delta Rp^2]$

at the junction Nx = Nb the bulk or substrate concentration

$$Nb = Np \exp[-(x-Rp)^2/2\Delta Rp^2]$$

$$xj = Rp \pm \Delta Rp\sqrt{2\ln Np/Nb}$$

Find N_p using the following equation $Q = \sqrt{2\pi}Np\Delta Rp$

Where Q is the implant dose

i) If the peak of the implant is at 170nm the implant energy is 150keV

And the ΔR_p for this energy is 0.066μm for silicon dioxide and 0.076 for silicon, either can be used in the calculation, preferably the oxide number, 0.066μm

$$Np = {Q}\Big/{\sqrt{2\pi}\Delta Rp}$$

N_p=6.04 x 10^{20}Atoms/cm^3

Implant energy 60kv straggle 0.063μm
Effectively Rp is 0 as the implant sits at the surface of the silicon and only the positive is therefore valid.

xj = Rp $\pm$ ΔRp $\sqrt{2\ \ln Np\ /\ Nb}$

xj=0+0.066 $\sqrt{2\ln 6.04\times10^{20}/1.0\times10^{15}}$

xj=0.066 x 5.132

xj=0.338μm

xj=0.34μm

To calculate the oxide thickness needed to block the implant assume that any doping “leaking under the oxide should be less than 1/10 of the background doping. And in this case as it is a total thickness the value of Rp in the equation is 200nm or 0.2μm

The equation becomes

$$Xo = Rp + \Delta Rp\sqrt{2\ln 10 Np / Nb}$$

Where in this case Xo is the thickness of the oxide needed to block the implant

Xo=0.17μm+0.066μm $\sqrt{2\ln 10\times 6.04\times10^{20}/1.0\times10^{15}}$

Xo=0.17+0.066(5.56)

Xo=0.17+0.367

Xo=0.537μm

Question 4

a) Describe in simple terms the operation of an enhancement mode NMOS transistor; explain how the gate controls the current flow between source and drain. Use cross-sectional diagrams to illustrate the answer. Indicate what bias is applied at various stages of operation; mention how the biasing of the “back gate” or substrate contact affects the threshold voltage.

Show a diagram of a cross section of a MOS device, showing SD, Gate oxide, Gate electrode, and electrical connections to the S, D, and Gate. Initially show 0V on the gate and a small bias (0.1V) on the Drain, show the Source and substrate grounded.

Description, with 0V on the gate and a small positive voltage on the drain there is no current flowing. Even though there is a potential difference between the source and drain because the positive is applied to the N side of the PN junction diode, this is a reverse biased diode and no current will cross.

If instead of 0V on the gate the voltage applied is increased in the positive direction, no current flows in the gate as the gate electrode is sitting on the gate oxide which is a good insulator and prevents current flow from the gate to the silicon. But if there is a positive voltage with no current flow there will be an associated electric field, this electric field spreads into the channel region of the device. The channel region of the device between the S and D is P type, that is a large number of holes as majority carriers and a smaller number of electrons as minority carriers. By definition these carriers can move under the influence of an electric field, like charges repel each other and unlike charges attract. So under the influence of the gate induced electric field the holes start to move away from the area immediately under the gate oxide, this is known as depletion, the region starts to be come depleted of majority carriers. If the voltage applied to the gate is further increased, the electric field will increase and push the majority carriers further away, at some point not alone are the majority carriers pushed away but minority carriers are pulled into the region. If the voltage on the gate becomes sufficiently high a layer of minority carriers or in this case is formed in the channel region under the gate. This condition is known as inversion, in a region where you expect to find a large number of holes there are a large number of electrons, the type has been inverted.
At this point the conditions for a reverse biased diode are no longer being met, (large no. holes on one side large no. of electrons on the other), now there are electrons for conduction in the N Drain region in the channel under the gate and in the source region, this is now like a resistor and the carriers are free to move between S and D under the influence of the potential difference between the source and drain. The point at which the channel is formed or the surface type is inverted is known as the threshold voltage of the device or the turn point of the device.
If instead of 0V bias (or grounded) substrate, consider a small reverse bias on the substrate wrt the source. This has the affect of narrowing any inversion channel formed in the channel region or making it more difficult with gate bias to form a channel in the first place. This has the overall affect of increasing the threshold voltage.

(b) If the polysilicon gate of an NMOS transistor in a Self-Aligned LOCOS process is over etched during processing by 0.1μm per side, calculate the change in maximum saturated current that would result on a transistor with a design gate length of 1.0μm and width of 10μm. What would the additional affect be if the LOCOS processing meant the “birds beak” (LOCOS edge) also encroached into the active area 0.5μm all round?

Given

V_t = 0.8V and the maximum power supply voltage is 5.0V

$\mu_n = 1450\text{cm}^2/\text{V.s}$

$C_{ox} = 5 \times 10^{-8}\text{F/cm}^2$

Use the saturated current equation

$$I_{DS} = \mu_n C_{ox} {}^{W}\!/_{2L} \left(V_G - V_t\right)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{10}\!/_{(2)1} \times (5.0 - 0.8)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{10}\!/_{(2)1} \times (17.64)$$

$$I_{DS1} = 6.39 \times 10^{-3} A$$

$$I_{DS1} = 6.39mA$$

$$I_{DS} = \mu_n C_{ox} {}^{W}\!/_{2L} \left(V_G - V_t\right)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{10}\!/_{(2)0.8} \times (5.0 - 0.8)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{10}\!/_{1.6} \times (17.64)$$

$$I_{DS1} = 7.99 \times 10^{-3} A$$

$$I_{DS1} = 7.99mA$$

The original design value for the saturated gate current is 6.39mA and the current with the shorter channel length is 7.99mA, the change/difference is 1.6mA

If the LOCOS encroaches by 0.5μm all round this has the affect of reducing the channel W from 10μm to 9.0μm (0.5μm at each end)

$$I_{DS} = \mu_n C_{ox} {}^{W}\!/_{2L} \left(V_G - V_t\right)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{9}\!/_{(2)0.8} \times (5.0 - 0.8)^2$$

$$I_{DS1} = 1450 \times 5 \times 10^{-8} \times {}^{9}\!/_{1.6} \times (17.64)$$

$$I_{DS1} = 7.99 \times 10^{-3} A$$

$$I_{DS1} = 7.19mA$$