EELE 414 – Introduction to VLSI Design

Module #2 – MOSFET Operation

- **Agenda**

  1. MOSFET Operation

     - Device Physics
     - MOSFET Structure
     - IV Characteristics
     - Scaling
     - Small Geometry Effects
     - Capacitance

- **Announcements**

  1. Read Chapter 3

MOSFET Operation

- **MOSFET**

  - Metal Oxide Semiconductor Field Effect Transistor

  - we need to understand the detailed operation of the MOSFET in order to use it to build larger blocks such as Inverters, NAND gates, adders, etc…

  - we will cover the theory of the device physics, energy bands, and circuit operation

  - we will do homework to analyze the behavior by hand

  - in the real world, we typically use SPICE simulations to quickly analyze the MOSFET behavior

  - but we need to understand what SPICE is calculating or:

    1) we won't be able to understand performance problems
    2) we won't be able to troubleshoot (is it the tool, is it the circuit, is it the process?)

Semiconductors

- **Semiconductors**

  - a semiconductor is a solid material which acts as an insulator at absolute zero. As the temperature increases, a semiconductor begins to conduct

  - a single element can be a semiconductor:

    Carbon (C), Silicon (Si)

  - a compound material can also form a semiconductors (i.e., two or more materials chemically bonded)

    Gallium Arsenide (GaAs), Indium Phosphide (InP)

  - an alloy material can also form semiconductors (i.e., a mixture of elements of which one is a metal):

    Silicon Germanium (SiGe), Aluminum Gallium Arsenide (AlGaAs)

  - Silicon is the most widely used semiconductors for VLSI circuits due to:

    - it is the 2nd most abundant element (25.7%) of the earth's crust (after oxygen)
    - it remains a semiconductor at a higher temperature
    - it can be oxidized very easily

Semiconductors

- **Charge Carriers**

  - since we want to use Si to form electronics, we are interested in its ability to conduct current. A good conductor has a high concentration of charge carriers.

  - an electron can be a charge carrier.

  - a hole (the absence of an electron) can be a charge carrier.

  - "Intrinsic" Silicon means silicon that is pure or it has no impurities. We sometimes called this i-typed Silicon

  - Since there are no impurities, the number of charge carriers is determined by the properties of the Silicon itself.

  - We can define the **Mobile Carrier Concentrations** as:

    n = the concentration of conducting electrons
    p = the concentration of conducting holes

  - these are defined per unit volume (1/cm^3)

Semiconductors

- **Charge Carriers**

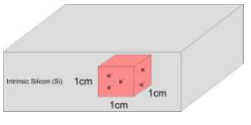
  - Intrinsic Silicon has a carrier concentration of :

    n_i = 1.45 x 10^{10} cm-3

  - notice the units are "carriers per cubic centimeter"

  - notice that we give the subscript "i" to indicate "intrinsic"

  - this value is dependant on temperature and is defined above at T=300 K (i.e., room temperature)



  - there are about 5x10^{22} Atoms of Silicon per cubic centimeter in a perfect intrinsic lattice

Semiconductors

- **Charge Carriers**

  - The equilibrium of the carriers in a semiconductor always follows the **Mass Action Law**

$$n \cdot p = n_i^2$$

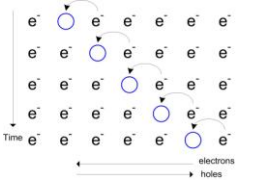
  - this means there is an equal number of p and n charge carriers in intrinsic Silicon

- **Electrons vs. Holes**

  - electrons have a charge of q=-1.6x10^{-19} Coulomb (C)

  - holes are the "absence" of electrons in an orbital of an atom. When an electron moves out of an orbital, it leaves a void (or hole). This hole can "accept" another electron

  - as electrons move from atom to atom, the holes effectively move in the opposite direction and give the impression of a positive charge moving



Energy Bands

- **Energy Bands**
  - the mobility of a semiconductor increases as its temperature increase.
  - Increasing the mobility of a semiconductor eventually turns the material into a conductor.
  - this is of interest to electronics because we can control the flow of current
  - we can also cause conduction using an applied voltage to provide the energy
  - we are interested in how much energy it takes to alter the behavior of the material
  - Energy Band Diagrams are a graphical way to describe the energy needed to change the behavior of a material.

Energy Bands

- **Energy Bands**
  - Quantum Mechanics created the concept of bands to represent the levels of energy that are present at each "state" of an atom.
  - the electrons on an atom occupy these energy states
  - For a given number of electrons in an atom, we begin filling in the energy bands from lowest to highest energy until all of the electrons have been used.
  - electrons only exist in the bands. By convention, electrons are forbidden from existing in between bands
  - there is a finite amount of energy that exists to move an electron from one band to another
  - if given enough energy (via heat or E-fields), electrons can receive enough energy to jump to a higher energy band.

Energy Bands

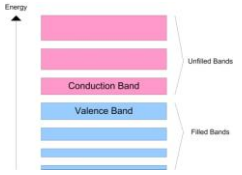
- **Energy Bands**

Valence Band : the highest range of electron energies where electrons are **normally** present at absolute zero.

: this is the highest "filled" band

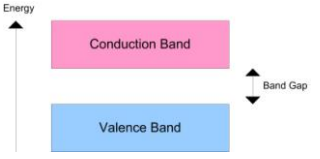
Conduction Band : the range of electron energy sufficient to make the electrons free to accelerate under the influence of an applied electric field (i.e., current).

: this is the lowest "unfilled" band



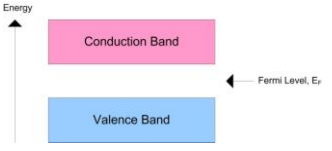
Energy Bands

- **Band Gap**
  - the band gap energy is the energy between the lowest level of the "conduction band" and the top of the "valence band"
  - this can be thought of as the amount of energy needed to release an electron for use as current at absolute zero.



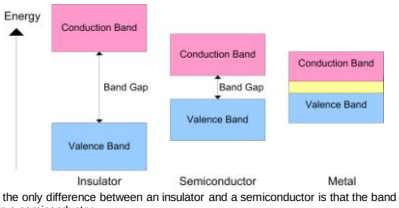
Energy Bands

- **Fermi Level**
  - the Fermi Level (or energy) represents an energy level that at absolute zero:
    - all bands below this level are filled
    - all bands above this level are unfilled
  - the Fermi Level at room temperatures is the energy at which the probability of a state being occupied has fallen to 0.5
  - at higher temperatures, in order for an electron to be used as current, it needs to have an energy level close to the Fermi Level
  - this can also be thought of as the equilibrium point of the material



Energy Bands

- **Band Gap Comparisons**
  - the following shows the relationship of Band Gap energies between insulators, semiconductors, and metals



  - notice that the only difference between an insulator and a semiconductor is that the band gap is smaller in a semiconductor.
  - notice that there is an overlap between the conduction and valence bands in metals. This means that metals are always capable of conducting current.

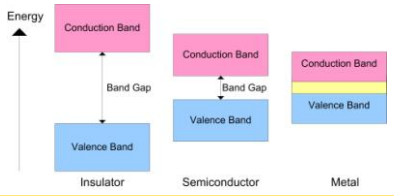
Energy Bands

- **Band Gap Comparisons**

Insulator Band Gap : it is large enough so that at ordinary temperatures, no electrons reach the conduction band

Semiconductor Band Gap : it is small enough so that at ordinary temperatures, thermal energy can give an electron enough energy to jump to the conduction band

: we can also change the semiconductor into a conductor by introducing impurities



Energy Bands

- **Band Gap Comparisons**

- we typically describe the amount of energy to jump a band in terms of "Electron Volts" (eV)

- 1 eV is the amount of energy gained by an unbound electron when passed through an electrostatic potential of 1 volt

- it is equal to (1 volt) x (unsigned charge of single electron)

- 1 Volt = (Joule / Coulomb)

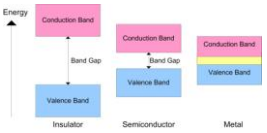
- (V x C) = (J/C) x (C) = units of Joules

- 1eV = 1.6×10^{-19} Joules

- we call materials with a band gap of ~ 1eV a "semiconductor"

- we call materials with a band gap of much greater than 1eV an "insulator"

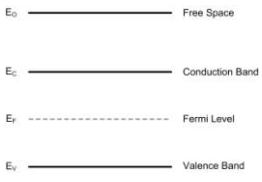
- and if there isn't a band gap, it is a "metal"



Energy Bands

- **Band Diagram**

- in a band diagram, we tabulate the relative locations of important energy levels



- Note that E_O is where the electron has enough energy to leave the material all together (an example would be a CRT monitor)

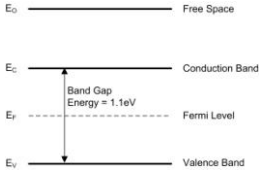
- as electrons get enough energy to reach near the Fermi level, conduction begins to occur

Energy Bands

- **Band Diagram of Intrinsic Silicon**

- Intrinsic Silicon has a band gap energy of 1.1 eV

  - @ 0 K, Eg=1.17 eV
  - @ 300 K, Eg=1.14 eV

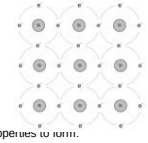


Doping

- **Doping**

- the most exploitable characteristic of a semiconductor is that impurities can be introduced to alter its conduction ability

- Silicon has a valence of 4 which allows it to form a perfect lattice structure. This lattice can be broken in order to accommodate impurities



- VLSI electronics use Silicon as the base material and then alter its properties to form:

1) n-type Silicon : material whose majority carriers are electrons
: introducing a valence-of-5 material increases the # of free negative charge carriers
: Phosphorus (P) or Arsenic (As) are typically used (group V elements)

2) p-type Silicon : material whose majority carriers are holes
: introducing a valence-of-3 material increases the # of free positive charge carriers
: Boron (B) is typically used (a group III element)

- when Silicon is doped, it is called **"Extrinsic Silicon"** due to the presence of impurities

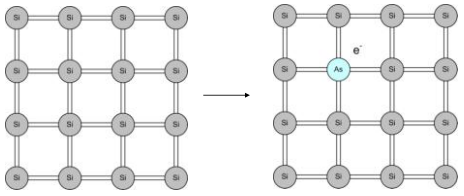
N-type Doping

- **N-type Doping**

- a perfect Silicon lattice forms covalent bonds with neighbors on each side

- there is an equal number of p and n charge carriers ($n \cdot p = n_i^2$)

- inserting an element into the lattice with a valence of 5 will form 4 covalent bonds PLUS have an extra electron

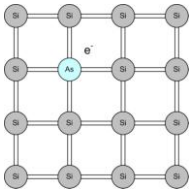


N-type Doping

- **N-type Doping**
  - this extra electron increases the n-type charge carriers
  - we call the additional element that provides the extra electron a **Donor**
  - the concentration of donor charge carriers is now denoted as N_D
  - we call N_D the doping concentration of an n-type material
  - we can use the Mass Action Law to say:

$$N_D \cdot p_{n-type} = n_i^2$$

$$N_D = \frac{n_i^2}{p_{n-type}}$$

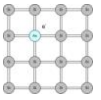


N-type Doping

- **N-type Doping**
  - Doping Silicon can achieve a Donor Carrier Concentration between 10^{13} cm^{-3} to 10^{18} cm^{-3}
  - Doping above 10^{18} cm^{-3} is considered degenerate (i.e., it starts to reduce the desired effect)
  - We give postscripts to denote the levels of doping (normal, light, or heavy)
  - Remember that Silicon has a density of ~10^{21} atoms per cm

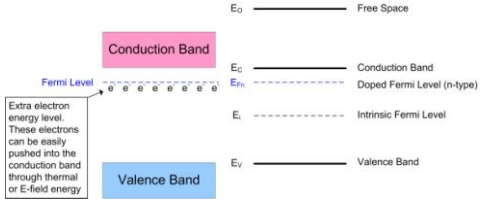
Example:

n-	: light doping	: $N_D = 10^{13}$ cm^{-3}	: 1 in 100,000,000 atoms
n	: normal doping	: $N_D = 10^{15}$ cm^{-3}	: 1 in 1,000,000 atoms
n+	: heavy doping	: $N_D \geq 10^{17}$ cm^{-3}	: 1 in 10,000 atoms



N-type Doping

- **Effect on the Band Structure**
  - by adding more electron charge carriers to a material, we create new energy states
  - by adding more electrons to Silicon, we decrease the energy that it takes for an electron to reach the conduction band
  - this moves the Fermi Level (the highest filled energy state at equilibrium) closer to the conduction band



N-type Doping

- **Effect on the Band Structure**
  - We can define the **Fermi Potential** (ϕ_F) as the difference between the intrinsic Fermi Level (E_i) and the new doped Fermi Level (E_{Fn})

$$\phi_{F_n} = \frac{E_{F_n} - E_i}{q}$$

  - note: that ϕ_{Fn} has units of volts and is positive since $E_{Fn} > E_i$,
  - note: that E_i and E_{Fn} have units of eV, which we convert to volts by dividing by q
  - note: we use q=1.6x10^{-19}C, which is a positive quantity
  - the **Boltzmann approximation** gives a relationship between the Fermi Level and the charge carrier concentration of a material (a.k.a, the Quasi Fermi Energy).
  - This expression relates the change in the Fermi Level (from intrinsic) to the additional charge carriers due to n-type doping.

$$n = n_i \cdot e^{\frac{(E_{F_n} - E_i)}{k_B \cdot T}}$$

where, k_B = the Boltzmann Constant = 8.62x10^{-5} (eV/K) or = 1.38x10^{-23} (J/K)

notice that the $(E_{Fn} - E_i)$ term in the exponent represents a positive voltage since $E_{Fn} > E_i$

N-type Doping

- **Effect on the Band Structure**
  - if we rearrange terms and substitute n=N_D…

$$N_D = n_i \cdot e^{\frac{(E_{F_n} - E_i)}{k_B \cdot T}}$$

$$\frac{N_D}{n_i} = e^{\frac{(E_{F_n} - E_i)}{k_B \cdot T}}$$

$$\ln\left(\frac{N_D}{n_i}\right) = \left(\frac{E_{F_n} - E_i}{k_B \cdot T}\right)$$

$$k_B \cdot T \cdot \ln\left(\frac{N_D}{n_i}\right) = \left(E_{F_n} - E_i\right)$$

Then plug into the Fermi potential

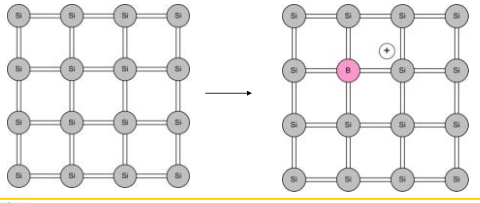
$$\phi_{F_n} = \frac{E_{F_n} - E_i}{q}$$

$$\phi_{F_n} = \frac{k_B \cdot T}{q} \cdot \ln\left(\frac{N_D}{n_i}\right)$$

- since $N_D > n_i$ the natural log is taken on a quantity that is greater than one
- this makes ϕ_{Fn} POSITIVE

P-type Doping

- **P-type Doping**
  - inserting an element into the silicon lattice with a valence of 3 will form 3 covalent bonds but leave one orbital empty
  - this is called a **hole** and since it "attracts an electron", it can be considered a positive charge with a value of +1.6e^{-19} C



P-type Doping

- **P-type Doping**

  - this extra electron increases the p-type charge carriers

  - we call this type of charge carrier an **Acceptor** since it provides a location for an electron to go

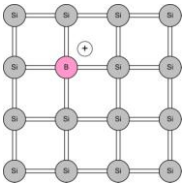
  - the concentration of acceptor charge carriers is now denoted as N_A

  - we call N_A the doping concentration of a p-type material

  - we can use the Mass Action Law to say:

$$n_{p-type} \cdot N_A = n_i^2$$

$$N_A = \frac{n_i^2}{n_{p-type}}$$



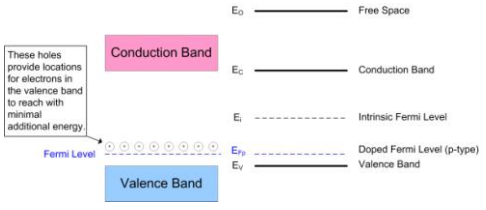
P-type Doping

- **Effect on the Band Structure**

  - by adding more hole charge carriers to a material, we also create new energy states

  - holes create new "unfilled" energy states

  - this moves the Fermi level down closer to the Valence band



P-type Doping

- **Effect on the Band Structure**

  - We again define the **Fermi Potential** (ϕ_F) as the difference between the intrinsic Fermi Level (E_i) and the new doped Fermi Level (E_{Fp})

$$\phi_{F_p} = \frac{E_{F_p} - E_i}{q}$$

  - we again use the **Boltzmann approximation**, which gives a relationship between the Fermi Level and the electron concentration of a material.

  - notice that the $(E_{Fp} - E_i)$ term yields a negative potential since $E_{Fp} < E_i$

  - note that for the P-type doping the Fermi level moves down below the original Intrinsic level. This original expression stated the ***increase*** in electron energy achieved by the doping.

  So we need to swap the p and n_i terms to use this equation.

$$p = n_i \cdot e^{\frac{E_i - E_{F_p}}{k_B \cdot T}} \Leftrightarrow n_i = p \cdot e^{-\frac{E_i - E_{F_p}}{k_B \cdot T}}$$

  notice that the $(E_i - E_{Fp})$ term in the exponent represents a positive voltage since $E_i > E_{Fp}$

P-type Doping

- **Effect on the Band Structure**

  - if we rearrange terms and substitute p=N_A…

$$n_i = N_A \cdot e^{-\frac{E_i - E_{F_p}}{k_B \cdot T}}$$

$$\frac{n_i}{N_A} = e^{-\frac{E_i - E_{F_p}}{k_B \cdot T}}$$

$$\ln\left(\frac{n_i}{N_A}\right) = -\left(\frac{E_i - E_{F_p}}{k_B \cdot T}\right)$$

$$k_B \cdot T \cdot \ln\left(\frac{n_i}{N_A}\right) = -\left(E_i - E_{Fp}\right)$$

$$k_B \cdot T \cdot \ln\left(\frac{n_i}{N_A}\right) = E_{F_p} - E_i$$

Then plug into the Fermi potential

$$\phi_{F_p} = \frac{E_{F_p} - E_i}{q}$$

$$\phi_{F_p} = \frac{k_B \cdot T}{q} \cdot \ln\left(\frac{n_i}{N_A}\right)$$

  - since $N_A > n_i$ the natural log is taken on a quantity that is between 0 and 1

  - this makes ϕ_{Fp} NEGATIVE

Work Function

- **Electron Affinity & Work Function**

  - another metric of a material is the amount of energy it takes to move an electron into Free Space (E_o)

Electron Affinity : the amount of energy to move an electron from the conduction band into Free Space.

$$q\chi = E_O - E_C$$

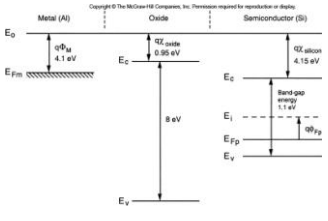
Work Function : the amount of energy to move an electron from the Fermi Level into Free Space.

$$q\Phi_S = q\chi + \left(E_C - E_F\right)$$

Work Function

- **Work Function of Different Materials**

  - When materials are separate, we can compare their band energies by lining up their Free Space energies

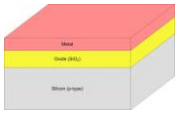


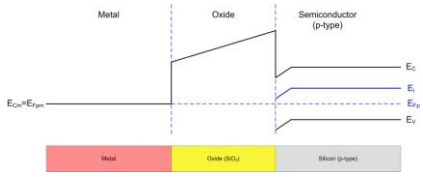
MOS Structure

- **MOS Structure**

  - When materials are bonded together, their Fermi Levels in the band diagrams line up to reflect the thermodynamic equilibrium.

  - of special interest to VLSI is the combination of a Metal Oxide Semiconductor (p-type) structure



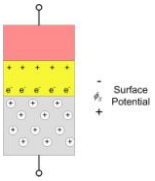


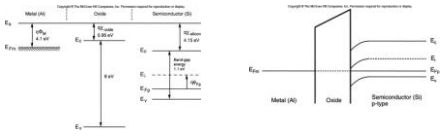
MOS Structure

- **Built-In Potential**

  - there is a built in potential due to the mismatches in work functions that causes the bands to bend down at the oxide-semiconductor junction

  - this is due to the PN junction that forms due to the p-type Si and the oxide. The oxide polarizes slightly at the surface

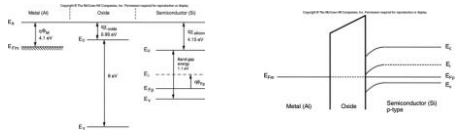




MOS Structure

- **Built-In Potential Example**

  - Example 3.1 in text. Given $q\phi_{Fp}$=0.2eV, what is the built in potential in the following MOS structure?



Solution: We need to find the difference in work functions between the Silicon substrate and the metal gate. We are given the metal gate work function (4.1eV) so we need to find the Silicon work function:

$$q\Phi_S = 4.15eV + \left(\frac{1.1eV}{2} + 0.2eV\right) = 4.9eV$$

Now we just subtract the Silicon work function from the Metal Gate work function:

$$q\Phi_M - q\Phi_S = 4.1eV - 4.9eV = -0.8eV$$

MOS Under Bias

- **MOS Accumulation**

  - If we apply an external bias voltage to the MOS, we can monitor how the charge carriers are affected

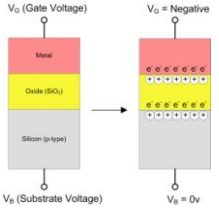
  - assume a "body" voltage of 0v (V_B=0)

  1) let's first apply a negative voltage to the "gate" (V_G=negative)

    - the holes of the p-type semiconductor are attracted to the Oxide surface

    - this causes the concentration of charge carriers at the surface to be greater than that of the normal concentration (N_A)

    - this is called the **Accumulation** of charge carriers in the semiconductor



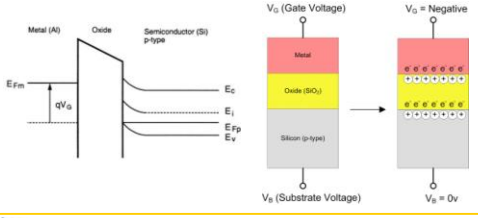
MOS Under Bias

- **MOS Accumulation**

  - applying a negative voltage to the metal raises its highest electron energy state by $q \cdot V_G$

  - the surface accumulation of energy can be reflected in the energy bands "bending up" near the Oxide-Semiconductor surface

  - note also that the minority carriers (electrons) in the p-type semiconductor are pushed away from the oxide surface (not shown)



MOS Under Bias

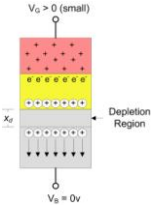
- **MOS Depletion**

  2) now let's apply a *small* positive voltage to the gate

    - the holes of the p-type semiconductor are repelled back away from the oxide surface

    - as V_G increases, it will approach a level where there are no mobile carriers near the Oxide-Semiconductor junction

    - the region without mobile carriers is called the **Depletion** Region

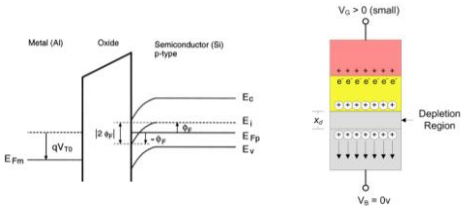


MOS Under Bias

- **MOS Depletion**

  - the positive voltage that develops at the Oxide-Semiconductor surface bends the energy bands downward to reflect the decrease in electron energy in this region.

  - the thickness of the depletion region is denoted as x_d



MOS Under Bias

- **MOS Depletion**

  - the depletion depth x_d is a function of the surface potential ϕ_S

  - if we model the holes as a sheet of charge parallel to the oxide surface, then the surface potential (ϕ_S) to move the charge sheet a distance x_d away can be solved using the Poisson equation.

  - the solutions of interest are:

    1) the depth of the depletion region:

$$x_d = \sqrt{\frac{2 \cdot \varepsilon_{Si} \cdot |\phi_S - \phi_F|}{q \cdot N_A}}$$

    2) the depletion region charge density:

$$Q = -q \cdot N_A \cdot x_d = -\sqrt{2 \cdot q \cdot N_A \cdot \varepsilon_{Si} \cdot |\phi_S - \phi_F|}$$

MOS Under Bias

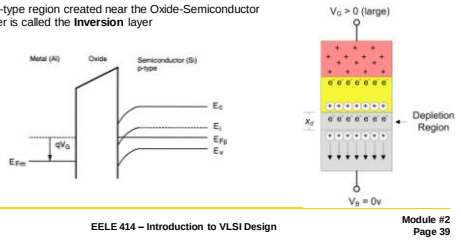
- **MOS Inversion**

  3) now let's apply a *larger* positive voltage to the gate

  - the positive surface charge in the Oxide is strong enough to pull the minority carrier electrons to the surface.

  - this can be seen in the band diagrams by "bending" the mid-gap (or E_i) energy at the surface of the Oxide and semiconductor until it falls below the Fermi Level (E_{Fp})

  - the n-type region created near the Oxide-Semiconductor barrier is called the **Inversion** layer



MOS Under Bias

- **MOS Inversion**

  - this region has a higher density of minority carriers than majority carriers during inversion

  - by definition, the region is said to be "inverted" when the density of mobile electrons is equal to the density of mobile holes

  - this requires that the surface potential has the same magnitude as the bulk Fermi potential, ϕ_F

  - as we increase the Gate voltage beyond inversion, more minority carriers (electrons) will be pulled to the surface and increase the carrier concentration

  - however, the inversion depth does not increase past its depth at the onset of inversion: $\phi_s = -\phi_F$

  - this means that the ***maximum depletion depth*** (x_{dm}) that can be achieved is given by:

$$x_{dm} = \sqrt{\frac{2 \cdot \varepsilon_{Si} \cdot |2\phi_F|}{q \cdot N_A}}$$

  - once an inversion layer is created, the electrons in the layer can be moved using an external E-field

MOSFET Operation

- **MOSFET Operation**

  - we saw last time that if we have a MOS structure, we can use V_G to alter the charge concentration at the oxide-semiconductor surface:

    1) Accumulation : $V_G < 0$
    : when the majority carriers of the semiconductor are pulled toward the oxide-Si junction

    2) Depletion : $V_G > 0$ (small)
    when the majority carriers of the Si are pushed away from the oxide-Si junction until there is a region with no mobile charge carriers

    3) Inversion : $V_G > 0$ (large)
    : when V_G is large enough to attract the minority carriers to the oxide-Si junction forming an *inversion layer*

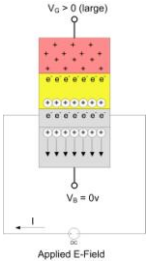
MOSFET Operation

- **MOSFET Operation (p-type substrate)**

  - Inversion is of special interest because we have created a *controllable* n-type channel that can be used to conduct current.

  - these electrons have enough energy that they can be moved by an electric field

  - if we applied an E-field at both ends of this channel, the electrons would move



  NOTE: In a p-type material, the holes are also charge carriers. But since they exist in all parts of the Si, we can't control where the current goes.

  We use the minority charge carriers in *inversion* because we can induce a channel using the MOS structure.

MOSFET Operation

- **MOSFET Operation (p-type substrate)**

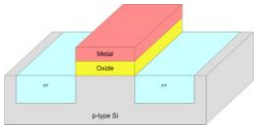
  - in order to access the channel created by inversion, we add two doped regions at either end of the MOS structure

  - these doped regions are of the minority carrier type (i.e., n-type)

  - current *can* flow between these terminals if an inversion is created in the p-type silicon by V_G

  - since we are controlling the flow of current with a 3rd terminal, this becomes a "transistor"

  - since we use an *E-field* to control the flow, this becomes the **MOS Field Effect Transistor**



MOSFET Operation

- **Terminal Definition**

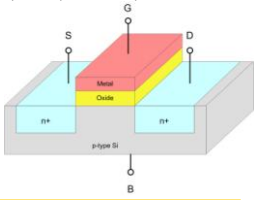
  Gate : The terminal attached to the metal of the MOS structure.

  Source : One of the doped regions on either side of the MOS structure. Defined as the terminal at the lower potential (vs. the Drain)

  Drain : One of the doped regions on either side of the MOS structure. Defined as the terminal at the higher potential (vs. the Source)

  Body : The substrate

  NOTE: we often don't show the Body connection



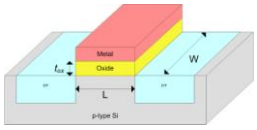
MOSFET Operation

- **MOSFET Dimensions**

  Length : the length of the channel. This is defined as the distance between the Source and Drain diffusion regions

  Width : the width of the channel. Notice that the metal, oxide, source, and drain each run this distance

  t_{ox} : the thickness of the oxide between the metal and semiconductor



MOSFET Operation

- **MOSFET Materials**

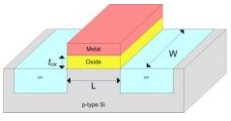
  Metal : Polysilicon. This is a silicon that has a heavy concentration of charge carriers. This is put on using Chemical Vapor Deposition (CVD). It is naturally conductive so it acts like a metal.

  Oxide : Silicon-Oxide (SiO_2). This is an oxide that is grown by exposing the Silicon to oxygen and then adding heat. The oxide will grow upwards on the Silicon surface

  Semiconductor : Silicon is the most widely used semiconductor.

  P-type Silicon : Silicon doped with Boron

  N-type Silicon : Silicon doped with either Phosphorus or Arsenic



MOSFET Operation

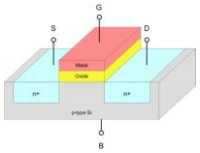
- **MOSFET Type**

  - we can create a MOSFET using either a p-type or n-type substrate. We then can move current between the source and drain using the minority carriers in inversion to form the conduction channel

  - we describe the type of MOSFET by describing what material is used to form the channel

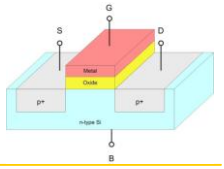
N-Channel MOSFET

- p-type Substrate
- n-type Source/Drain
- current carried in n-type channel



P-Channel MOSFET

- n-type Substrate
- p-type Source/Drain
- current carried in p-type channel



MOSFET Operation

- **Enhancement vs. Depletion MOSFETS**

  Enhancement Type : when a MOSFET has no conduction channel at V_G=0v
  : also called *enhancement-mode*
  : we apply a voltage at the gate to turn **ON** the channel
  : this is used most frequently and what we will use to learn VLSI

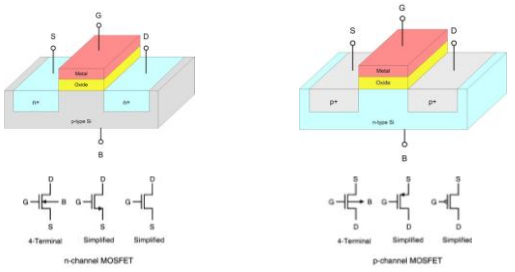
  Depletion Type : when a MOSFET **does have** a conducting channel at V_G=0v
  : also called *depletion-mode*
  : we apply a voltage at the gate to turn **OFF** the channel
  : we won't use this type of transistor for now

  Note: We will learn VLSI circuits using enhancement-type, n-channel MOSFETS. All of the principles apply directly to Depletion-type MOSFETs as well as p-channel MOSFETs.

MOSFET Operation

- **MOSFET Symbols**

  - there are multiple symbols for enhancement-type MOSFETs that can be used



MOSFET Operation

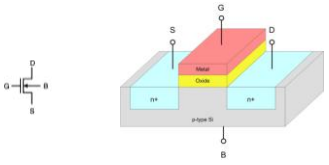
- **Terminal Voltages**

  - all voltages in a MOSFET are defined relative to the Source terminal

V_{GS} : Gate to Source Voltage

V_{DS} : Drain to Source Voltage

V_{BS} : Body to Source Voltage



MOSFET Operation Under Bias

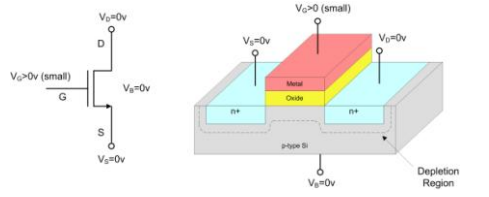
- **MOSFET under Bias (Depletion)**

  - let's begin with an n-channel, enhancement-type MOSFET

  - we bias the Source, Drain, and Body to 0v

  - we apply a *small* positive voltage to the gate, $V_{GS} > 0$ (small)

  - this creates a *depletion region* beneath the Gate, Source, and Drain that is void of all charge carriers



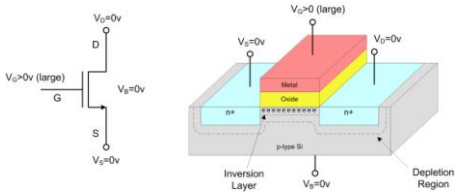
MOSFET Operation Under Bias

- **MOSFET under Bias (Inversion)**

  - as V_{GS} gets larger, it will form an *inversion layer* beneath the Gate oxide by attracting the minority carriers in the substrate to the oxide-Si surface.

  - when the surface potential of the gate reaches the bulk Fermi potential, the surface inversion will be established and an n-channel will form $\phi_s = -\phi_F$

  - this channel forms a path between the Source and Drain



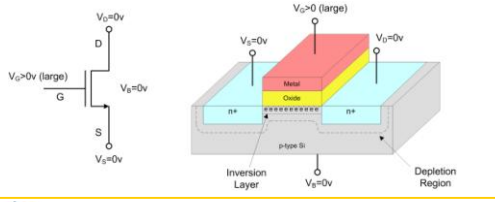
Threshold Voltage

- **MOSFET under Bias (Inversion)**

  - as V_{GS} gets larger, it will form an *inversion layer* beneath the Gate oxide by attracting the minority carriers in the substrate to the oxide-Si surface.

  - when the surface potential of the gate reaches the bulk Fermi potential, the surface inversion will be established and an n-channel will form $\phi_s = -\phi_F$

  - this channel forms a path between the Source and Drain



Threshold Voltage

- **MOSFET under Bias (Inversion)**

  - we are very interested when an inversion channel forms because it represents when the transistor is ON

  - we define the Gate-Source voltage (V_{GS}) necessary to cause inversion the ***Threshold Voltage*** (V_{T0})

when $V_{GS} < V_{T0}$ there is no channel so no current can flow between the Source and Drain terminals

when $V_{GS} \geq V_{T0}$ an inversion channel is formed so current can flow between the Source and Drain terminals

NOTE: We are only establishing the *channel* for current to flow between the Drain and Source. We still have not provided the necessary V_{DS} voltage in order to induce the current.

- just as in the MOS inversion, increasing V_{GS} beyond V_{T0} does not increase the surface potential or depletion region depth beyond their values at the onset of inversion.

It does however increase the concentration of charge carriers in the inversion channel.

Threshold Voltage

- **Threshold Voltage**
  - the threshold voltage depends on the following:
    1) the work function difference between the Gate and the Channel Φ_{GC}
    2) the gate voltage necessary to change the surface potential $2\cdot\phi_F$
    3) the gate voltage component to offset the depletion region charge $\frac{Q_{B0}}{C_{ox}}$
    4) the gate voltage necessary to offset the fixed charges in the Gate-Oxide and Si-Oxide junction $\frac{Q_{ox}}{C_{ox}}$
  - putting this all together gives us the expression for the threshold voltage at ***Zero Substrate Voltage***

$$V_{T0} = \Phi_{GC} - 2\cdot\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$$

Threshold Voltage

- **Threshold Voltage with Non-Zero Substrate Bias**
  - sometimes we can't guarantee that the substrate will be zero at all points of the IC:
  - when a potential develops in the substrate, it pushes the Source terminal of the MOSFET to a higher potential. We typically describe this as V_{SB} (instead of V_{BS})
  - to predict the effect of a substrate bias voltage (V_{SB}), we must alter the expression for the depletion charge density term:

$$\frac{Q_{B0}}{C_{ox}} \rightarrow \frac{Q_B}{C_{ox}}$$

  - this changes the expression for the Threshold Voltage to:

$$V_T = \Phi_{GC} - 2\cdot\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$$

Threshold Voltage

- **Threshold Voltage with Non-Zero Substrate Bias cont…**
  - **V_{T0}** is hard to predict due to uncertainties in the doping concentrations during fabrication. As a result, V_{T0} is measured instead of calculated.
  - this means for a typical transistor, it is a given quantity
  - however, the non-zero Substrate Bias is a quantity that still must be considered.
  - we want to get an expression for V_T that includes V_{T0} (a given)

$$V_T = \Phi_{GC} - 2\cdot\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_B - Q_{B0}}{C_{ox}} = V_{T0} - \frac{Q_B - Q_{B0}}{C_{ox}}$$

  - the depletion charge density is a function of the material and the substrate bias:

$$\frac{Q_B - Q_{B0}}{C_{ox}} = -\frac{\sqrt{2q\cdot N_A\cdot\varepsilon_{Si}}}{C_{ox}}\cdot\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$$

Threshold Voltage

- **Threshold Voltage with Non-Zero Substrate Bias cont…**
  - we can separate the material dependant term into its own parameter separate from V_{SB}

$$\gamma = \frac{\sqrt{2q\cdot N_A\cdot\varepsilon_{Si}}}{C_{ox}}$$

  where γ is called the ***substrate-bias*** or ***body-effect*** coefficient
  - this leaves our complete expression for threshold voltage as:

$$V_T = V_{T0} + \gamma\cdot\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$$

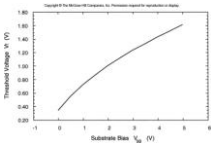
  - a few notes on this expression:

	ϕ_F	γ	V_{SB}
1) in an n-channel, the following signs apply:	–	+	+
2) in a p-channel, the following signs apply	+	–	–

Threshold Voltage

- **Threshold Voltage with Non-Zero Substrate Bias cont…**
  - the following plot shows an example of threshold dependence on substrate bias for an enhancement-type, n-channel MOSFET

$$V_T = V_{T0} + \gamma\cdot\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$$



  - the threshold voltage increases with Substrate bias. This means as noise gets on the substrate, it takes more energy to create the channel in the MOSFET. This is a BAD thing…

MOSFET I-V Characteristics

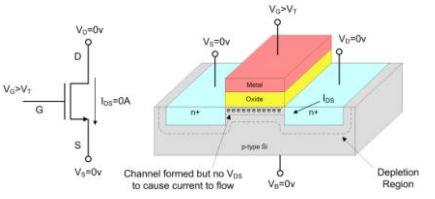
- **MOSFET I-V Characteristics**
  - we have seen how the Gate-to-Source voltage (V_{GS}) induces a channel between the Source and Drain for current to flow through
  - this current is denoted **I_{DS}**
  - remember that this current doesn't flow unless a potential exists between V_D and V_S
  - the voltage that controls the current flow is denoted as **V_{DS}**
  - once again, we start by applying a small voltage and watching how I_{DS} responds
  - notice that now we actually have two control variables that effect the current flow, **V_{GS}** and **V_{DS}**
  - this is typical operating behavior for a 3-terminal device or *transistor*
  - we can use an enhancement n-channel MOSFET to understand the IV characteristics and then directly apply them to p-channel and depletion-type devices

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Cutoff Region**
  - when $V_{GS} < V_T$, there is no channel formed between the Drain and Source and hence I_{DS}=0 A
  - this region is called the *Cutoff Region*
  - this region of operation is when the Transistor is OFF

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**
  - When $V_{GS} > V_T$, a channel is formed. I_{DS} is dependant on the V_{DS} voltage
  - When V_{DS} = 0v, no current flows



MOSFET I-V Characteristics

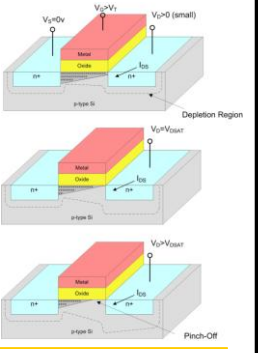
- **MOSFET I-V Characteristics : Linear Region**
  - If $V_{GS} > V_T$ and $V_{DS} > 0$, then a current will flow from the Drain to Source (I_{DS})
  - the MOSFET operates like a voltage controlled resistor which yields a *linear* relationship between the applied voltage (V_{DS}) and the resulting current (I_{DS})
  - for this reason, this mode of operation is called the ***Linear Region***
  - this region is also sometimes called the *triode region* (we'll use the term "linear")
  - V_{DS} can increase up to a point where the current ceases to increase linearly (saturation)
  - we denote the highest voltage that V_{DS} can reach and still yield a linear increase in current as the *saturation voltage* or V_{DSAT}

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**
  - when a voltage is applied at V_D, its positive charge pushes the majority charge carriers (holes) that exist at the edge of the depletion region further from the Drain.
  - as the depletion region increases, it becomes more difficult for the Gate voltage to induce an inversion layer. This results in the inversion layer depth decreasing near the drain.
  - as V_D increases further, it eventually causes the inversion layer to be *pinched-off* and prevents the current flow to increase any further.
  - this point is defined as the *saturation voltage* (V_{DSAT})
  - from this, we can define the *linear region* as:

    $V_{GS} > V_T$

    $0 < V_{DS} < V_{DSAT}$



MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**
  - the Drain to Source current (I_{DS}) is given by the expression:

$$I_{DS_{linear}} = \frac{u_n \cdot C_{ox}}{2} \cdot \frac{W}{L} \cdot \left[2 \cdot (V_{GS} - V_{T0}) \cdot V_{DS} - V_{DS}^2\right]$$

  - where:

u_n	= electron surface mobility (units in cm²/V·s)
C_{ox}	= Unit Oxide Capacitance (units in F/cm²)
W	= width of the gate
L	= length of the gate

  - remember this expression is only valid when :

    $V_{GS} > V_T$

    $0 < V_{DS} < V_{DSAT}$

A note on electron mobility (u_n):

u_n relates the drift velocity to the applied E-field

Drift velocity is the average velocity that an electron can attain due to an E-field.

We are interested in Drift Velocity because it tells us how fast the electron can get from the Source to the Drain.

Since current is defined as I=ΔQ/ Δt, u_n relates how much charge can move in a given area per-time and per E-field

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**
  - what is linear about this equation?

$$I_{DS_{linear}} = \frac{u_n \cdot C_{ox}}{2} \cdot \frac{W}{L} \cdot \left[2 \cdot (V_{GS} - V_{T0}) \cdot V_{DS} - V_{DS}^2\right]$$

  - most of the parameters are constants during evaluation. They are sometimes lumped into single parameters

$$k' = u_n \cdot C_{ox} \qquad I_{DS_{linear}} = \frac{k'}{2} \cdot \frac{W}{L} \cdot \left[2 \cdot (V_{GS} - V_{T0}) \cdot V_{DS} - V_{DS}^2\right]$$

or

$$k = u_n \cdot C_{ox} \cdot \frac{W}{L} \qquad I_{DS_{linear}} = \frac{k}{2} \cdot \left[2 \cdot (V_{GS} - V_{T0}) \cdot V_{DS} - V_{DS}^2\right]$$

  - Notice that W and L are parameters that the designers have control over. Most of the other parameters are defined by the fabrication process and are out of the control of the IC designer.

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**

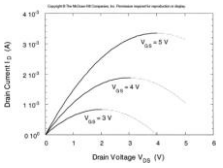
  - what is linear about this equation?

$$I_{DS_{linear}} = \frac{k}{2}\cdot\left[2\cdot(V_{GS}-V_{T0})\cdot V_{DS}-V_{DS}^2\right]$$

For a fixed V_{GS}, then I_{DS} depends on V_{DS}

V_{DS}^2 has a smaller effect on I_{DS} at low values of V_{DS} since it is not multiplied by anything

  - the $-V_{DS}^2$ term alters the function shape in the linear region. As it becomes large enough to significantly *decrease* I_{DS} in this function, the transistor enters *saturation* and this expression is no longer valid.

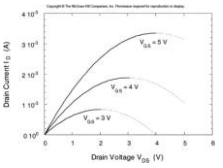


MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Linear Region**

  - since we know what the current will not decrease as V_{DS} increases past V_{DSAT}, we can use this expression to define V_{DSAT}:

$$I_{DS_{linear}} = \frac{k}{2}\cdot\left[2\cdot(V_{GS}-V_{T0})\cdot V_{DS}-V_{DS}^2\right]$$



  - when $V_{DS}>(V_{GS}-V_T)$, then I_{DS} in this expression begins to decrease

  - we can then define $V_{DSAT} = (V_{GS}-V_T)$

  - so now we have the formal limits on the linear region and the validity of this expression:

Linear Region : $V_{GS}>V_T$

$0 < V_{DS} < (V_{GS}-V_T)$

MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : Saturation Region**

  - a MOSFET is defined as being in saturation when:

Saturation Region : $V_{GS} \geq V_T$

$V_{DS} \geq (V_{GS}-V_T)$

  - an increase in V_{DS} does not increase I_{DS} because the channel is *pinched-off*

  - However, an increase in V_{GS} **DOES** increase I_{DS} by increasing the channel depth and hence the amount of current that can be conducted.

  - measurements on MOSFETS have shown that the dependence of I_{DS} on V_{GS} tends to remain approximately constant around the peak value reached for $V_{DS}=V_{DSAT}$

  - a substitution of $V_{DS}=(V_{GS}-V_{T0})$ yields:

$$I_{DS_{sat}} = \frac{k}{2}\cdot\left[2\cdot(V_{GS}-V_{T0})\cdot(V_{GS}-V_{T0})-(V_{GS}-V_{T0})^2\right]$$

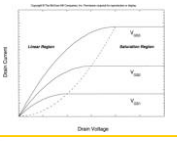
$$I_{DS_{sat}} = \frac{k}{2}\cdot(V_{GS}-V_{T0})^2$$

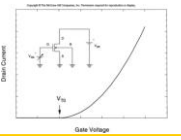
MOSFET I-V Characteristics

- **MOSFET I-V Characteristics : IV Curves**

  - now we have 1st order expressions for all three regions of operation for the MOSFET

Region	Conditions	I_{DS}
Cutoff	$V_{GS} < V_T$	$I_{DS_{cutoff}} = 0$
Linear	$V_{GS} \geq V_T$ $V_{DS} < (V_{GS}-V_T)$	$I_{DS_{linear}} = \frac{k}{2}\cdot\left[2\cdot(V_{GS}-V_{T0})\cdot V_{DS}-V_{DS}^2\right]$
Saturation	$V_{GS} \geq V_T$ $V_{DS} \geq (V_{GS}-V_T)$	$I_{DS_{sat}} = \frac{k}{2}\cdot(V_{GS}-V_{T0})^2$





MOSFET I-V 2nd Order Effects

- **Channel Length Modulation**

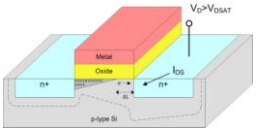
  - the 1st order IV equations derived earlier are not 100% accurate. They are sufficient for 1st order (gut-feel) hand calculations

  - we can modify these IV equations to include other effects that alter the IV characteristics of a MOSFET

  - Channel Length Modulation refers to additional I_{DS} current that exists in the *saturation* mode that is not modeled by the 1st order IV equations

  - when the channel is pinched off in saturation by a distance ΔL, a depletion region is created next to the Drain that is ΔL wide

  - given enough energy, electrons in the inversion layer can move through this depletion region and into the Drain thus adding additional current to I_{DS}



MOSFET I-V 2nd Order Effects

- **Channel Length Modulation**

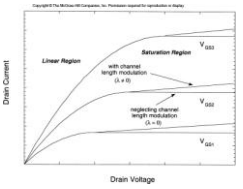
  - we can model this additional saturation current by multiplying the I_{DS} expression by:

$$(1+\lambda\cdot V_{DS})$$

  - λ is called the ***channel length modulation coefficient*** and is determined via empirical methods

  - this term alters the $I_{DS_{sat}}$ expression to be:

$$I_{DS_{sat}} = \frac{k}{2}\cdot(V_{GS}-V_{T0})^2\cdot(1+\lambda\cdot V_{DS})$$



MOSFET I-V 2nd Order Effects

- **Substrate Bias Effect**
  - another effect that the 1st order IV equations don't model is substrate bias
  - we have assumed that the Silicon substrate is at the same potential as the Source of the MOSFET
  - if this is not the case, then the Threshold Voltage may increase and take more energy to induce a channel
  - we've already seen how we can model the change in threshold voltage due to substrate bias:

$$V_T = V_{T0} + \gamma \cdot \left(\sqrt{\left| -2\phi_F + V_{SB} \right|} - \sqrt{\left| -2\phi_F \right|} \right)$$

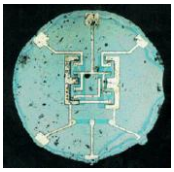
  - for the IV equations to accurately model the substrate bias effect, we must use V_T instead of V_{T0}

$$I_{DS_{linear}} = \frac{k}{2} \cdot \left[2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2 \right]$$

$$I_{DS_{sat}} = \frac{k}{2} \cdot (V_{GS} - V_T)^2 \cdot (1 + \lambda \cdot V_{DS})$$

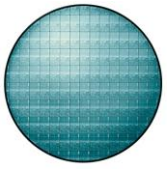
Scaling Theory

- **What is Scaling?**
  - Moving VLSI designs to new fabrication processes
  - Shrinking the size of the circuitry



1961
First Planar Integrated Circuit
Two Transistors

2001
Pentium 4 Processor
42 Million Transistors



2006
Itanium 2 Dual Processor
1.7 Billion Transistors

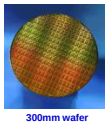
Scaling Theory

- **Why do we Scale?**

1) Improve Performance
   - More complex systems

2) Increase Transistor Density
   - Reduce cost per transistor & size of system

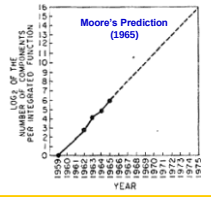
3) Reduce Power
   - Smaller transistors require less supply voltage



300mm wafer

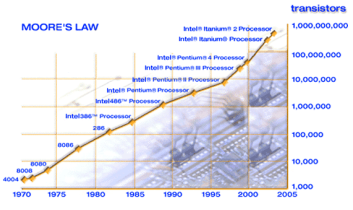
Scaling Theory

- **Scaling Predictions**
  - In 1965, Gordon Moore of Intel predicted the exponential growth of the number of transistors on an IC.
  - Transistor count will doubled every 2-3 years
  - Predicting >65,000 transistors in 1975



Scaling Theory

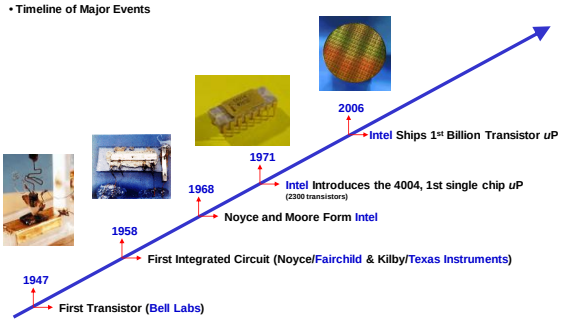
- **More than just a prediction**
  - Transistor count has doubled every 26 months for the past 30 years



  - Today this trend is used to target future process performance and prepare necessary infrastructure (Design Tools, Test, Manufacturing, Engineering Skills, etc…)

Scaling Theory

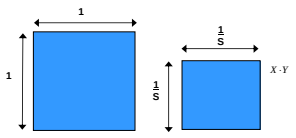
- **Timeline of Major Events**



Scaling Theory

- **How much can we shrink?**

  - Chip Area (A)



$$A \propto$$

$$A \propto \left(\frac{1}{S}\right)\cdot\left(\frac{1}{S}\right)$$

Chip Area for a Circuit *(A)* scales following : $\frac{1}{S^2}$

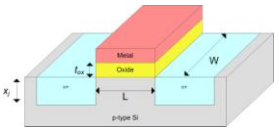
Note: In addition, the die sizes have increased steadily, allowing more transistors per die

Full Scaling

- **Full Scaling (Constant-Field)**

  - Reduce physical size of structures by 30% in the subsequent process

W	= Width of Gate
L	= Length of Gate
t_{ox}	= thickness of Oxide
x_j	= depth of doping



  - Reduce power supplies and thresholds by 30%

  - we define: S ≡ *Scaling Factor* > 1

  - Historically, S has come in between 1.2 and 1.5 for the past 30 years

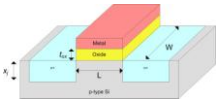
  - sometimes we use √2 ≡ 1.4 for easy math

Full Scaling

- **Full Scaling (Constant Field)**

  - The following quantities are altered during fabrication

  - we use a prime (') to denote the new scaled quantity



Quantity	Before Scaling	After Scaling
Channel Length	*L*	*L' = L/S*
Channel Width	*W*	*W' = W/S*
Gate Oxide Thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power Supply Voltage	V_{DD}	$V_{DD}' = V_{DD}/S$
Threshold Voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping Densities	N_A	$N_A' = N_A \cdot S$

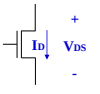
  - Note that the doping concentration has to be increased to keep achieve the desired Fermi level movement due to doping since the overall size of the junction is reduced

Full Scaling

- **Scaling Effect on Device Characteristics : Linear Region**

  - by scaling t_{ox}, we effect C_{ox} :

$$C_{ox}' = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{\varepsilon_{ox}}{t_{ox}/S} = S\cdot\frac{\varepsilon_{ox}}{t_{ox}} = S\cdot C_{ox}$$



  - since $k = u_n \cdot C_{ox}\cdot\frac{W}{L}$ then $k' = u_n \cdot C_{ox}'\cdot\frac{W}{L} = S\cdot k$

  - The voltages V_{GS}, V_{TO}, and V_{DS} also scale down by S, which creates a1/S^2 in this expression:

$$I_{DS_{linear}} = \frac{k}{2}\cdot\left[2\cdot(V_{GS}-V_{T0})\cdot V_{DS}-V_{DS}^2\right] \Rightarrow I_{DS_{linear}}' = \frac{S\cdot k}{2}\cdot\frac{1}{S^2}\left[2\cdot(V_{GS}-V_{T0})\cdot V_{DS}-V_{DS}^2\right]$$

  - which results in:

$$I_{DS_{linear}}' = \frac{I_{DS_{linear}}}{S}$$

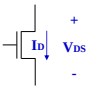
$I_{DS_{lin}}$ scales down by S, this is what we wanted!!!

Full Scaling

- **Scaling Effect on Device Characteristics : Saturation Region**

  - again, k effects I_{DS}

$$I_{DS_{SAT}} = \frac{k}{2}\cdot(V_{GS}-V_{T0})^2 \Rightarrow I_{DS_{SAT}}' = \frac{S\cdot k}{2}\cdot\frac{1}{S^2}\cdot(V_{GS}-V_{T0})^2$$



  - which gives

$$I_{DS_{SAT}}' = \frac{I_{DS_{SAT}}}{S}$$

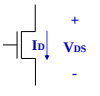
$I_{DS_{sat}}$ scales down by S, this is what we wanted!!!

Full Scaling

- **Scaling Effect on Device Characteristics : Power**

  - Static Power in the MOSFET can be described as:

$$P = I_{DS}\cdot V_{DS}$$



  - both quantities scale by 1/S

$$P' = \frac{I_{DS}}{S}\cdot\frac{V_{DS}}{S} = \frac{P}{S^2}$$

Power scales down by S^2, this is great!!!

Full Scaling

- **Scaling Effect on Device Characteristics : Power Density**

  - Power Density is defined as the power consumed per area

  - this is an important quantity because it shows how much heat is generated in a small area, which can cause reliability problems

$$P_{Density} = \frac{P}{Area}$$

  - Power scales by $1/S^2$

  - Area scales by $1/S^2$ (because W and L both scale by S and Area=W·L)

  - this means that the scaling cancels out and the Power Density remains constant

This is OK, but can lead to problems when IC's get larger in size and the net power consumption increase

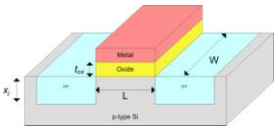
Constant-Voltage Scaling

- **Constant-Voltage Scaling**

  - sometimes it is impractical to scale the voltages

  - this can be due to:

    1) existing I/O interface levels
    2) existing platform power supplies
    3) complexity of integrating multiple power supplies on chip



  - Constant-Voltage Scaling refers to scaling the physical quantities (W, L, t_{ox}, x_j, N_A) but leaving the voltages un-scaled (V_{T0}, V_{GS}, V_{DS})

  - while this has some system advantages, it can lead to some unwanted increases in MOSFET characteristics

Constant-Voltage Scaling

- **Scaling Effect on Device Characteristics : Linear Region**

  - we've seen that scaling t_{ox}, W, and L causes:

$$k' = S \cdot k$$

  - if the voltages (V_{GS}, V_{T0}, and V_{DS}) aren't scaled, then the I_{DS} expression in the linear region becomes:

$$I'_{DS_{linear}} \frac{S \cdot k}{2} \cdot \left[2 \cdot (V_{GS} - V_{T0}) \cdot V_{DS} - V_{DS}^2\right]$$

  - which results in:

$$I'_{DS_{linear}} = S \cdot I_{DS_{linear}}$$

$I_{DS_{lin}}$ actually increases by S when we get smaller, this is NOT what we wanted!!!

Constant-Voltage Scaling

- **Scaling Effect on Device Characteristics : Saturation Region**

  - this is also true in the saturation region:

$$I'_{DS_{SAT}} = \frac{S \cdot k}{2} \cdot (V_{GS} - V_{T0})^2$$

  - which results in:

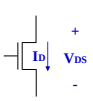
$$I'_{DS_{SAT}} = S \cdot I_{DS_{SAT}}$$

$I_{DS_{SAT}}$ also increases by S when we get smaller, this is NOT what we wanted!!!

Constant-Voltage Scaling

- **Scaling Effect on Device Characteristics : Power**

  - Instantaneous Power in the MOSFET can be described as:



$$P = I_{DS} \cdot V_{DS}$$

  - but in Constant-Voltage Scaling, I_{DS} increases by S and V_{DS} remains constant

$$P' = S \cdot I_{DS} \cdot V_{DS} = S \cdot P$$

Power increases by S as we get smaller, this is not what we wanted!!!

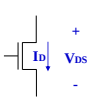
Constant-Voltage Scaling

- **Scaling Effect on Device Characteristics : Power Density**

  - Power Density is defined as the power consumed per area

  - we've seen that Power increases by S in Constant-Voltage Scaling

  - but area is still scaling by $1/S^2$



$$P'_{Density} = \frac{S \cdot P}{\left(\frac{Area}{S_2}\right)} = S^3 \cdot P$$

  - This is a very bad thing because a lot of heat is being generated in a small area

Scaling Choices

- **So Which One Do We Choose?**
  - Full Scaling is great, but sometimes impractical.
  - Constant Voltage can actually be worse from a performance standpoint

Quantity	Full Scaling	Constant-V Scaling
C_{ox}'	S	S
I_{DS}'	$1/S$	S
Power'	$1/S^2$	S
Power Density'	1	S^3

  - We actually see a hybrid approach. Dimensions tend to shrink each new generation. Then the voltages steadily creep in subsequent designs until they are in balance. Then the dimensions will shrink again.
  - Why scale if it is such a pain?
    - the increase in complexity per area is too irresistible.
    - it also creates a lot of fun and high paying jobs.

Scaling Trends

- **How Does Scaling Effect AC Performance?**
  - Assume Full Scaling
  - Resistance (R)

$$R \propto$$

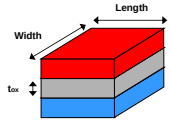
$$\frac{V_{DD}}{I_{ds}} \quad R \propto \frac{(1/S)}{(1/S)}$$

Device Resistance *(R)* remains constant : 1

OK

Scaling Trends

- **How Does Scaling Effect AC Performance?**
  - Total Gate Capacitance (C)



$$C \propto$$

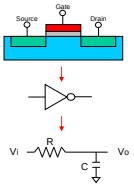
$$\frac{W \cdot L}{t_{ox}} \quad C \propto \frac{(1/S)\cdot(1/S)}{(1/S)}$$

Gate Capacitance *(C)* scales following : $\frac{1}{S}$

Good!

Scaling Trends

- **How Does Scaling Effect AC Performance?**
  - Gate Delay (τ)



$$\tau \propto$$

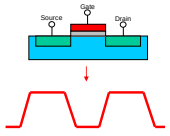
$$R \cdot C \quad \tau \propto 1 \cdot (1/S)$$

Gate Delay *(τ)* scales following : $\frac{1}{S}$

Good!

Scaling Trends

- **How Does Scaling Effect AC Performance?**
  - Clock Frequency (τ)



$$f \propto$$

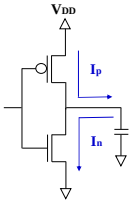
$$\frac{1}{\tau} \quad f \propto \frac{1}{(1/S)}$$

Clock Frequency *(f)* scales following : S

Good!

Scaling Trends

- **How Does Scaling Effect AC Performance?**
  - Dynamic Power Consumption (P)



$$P \propto$$

$$C \cdot V^2 \cdot f \quad P \propto (1/S)\cdot(1/S)^2 \cdot (S)$$

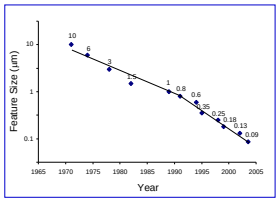
Dynamic Power *(P)* scales following : $\frac{1}{S^2}$

Great!!!

Does Scaling Work?

- **How accurate are the predictions?**

  - For three decades, the scaling predictions have tracked well

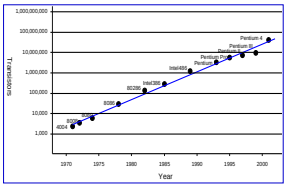


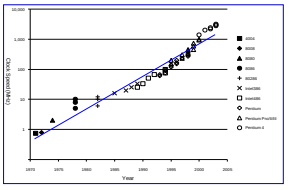
Feature Sizes have been reduced by >30%

Does Scaling Work?

- **How accurate are the predictions?**

Transistor Count has increased exponentially





Clock Rates have improved >43%

Can We Keep Scaling?

- **Why not just keep scaling?**

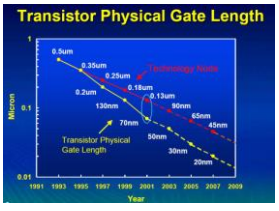
  - If it was easy, we wouldn't have jobs!

  - each time we get smaller, a couple new major problems arise.

  - Over the years, we have a list of issues that we call "Small Geometry Effects" that have posed a barrier to future scaling.

  - But until now, all of the problems have been solved with creative engineering and we continue on to the next process.

  - You will need to solve the problems in the next generation of process sizes. Good luck!



Small Geometry Challenges

Short Channel Effects

- a MOSFET is called a *short channel device* when the channel length is close to the same size as the depletion region thickness ($L \approx x_{dm}$). It can also be defined as when the effective channel length L_{eff} is close to the same as the diffusion depth ($L_{eff} \approx x_j$)

Velocity Saturation

- as the device gets smaller, the relative E-field energy tends to increase and the carriers in the channel can reach higher and higher speeds.

- the long channel equations (i.e., the 1st order IV expressions) shows a linear relationship between the E-field and the velocity of the carrier.

- however, at a point, the carriers will reach a maximum speed due to collisions with other electrons and other particles in the Silicon

- At this point, there is no longer a linear relationship between the applied E-field (V_{DS}) and the carrier velocity, which ultimately limits the increase in I_{DS}

- we can model this effect by altering the electron mobility term, u_n

$$u_n(eff) = \frac{u_{n_0}}{1 + \eta \cdot (V_{GS} - V_T)}$$

where η is an empirical coefficient

Small Geometry Challenges

Subthreshold Leakage

- we've stated that when $V_{GS} < V_T$, there is no inversion in the channel and hence, no charge carriers to carry current from the Drain to Source

- this transition from no-inversion to inversion doesn't happen instantaneously

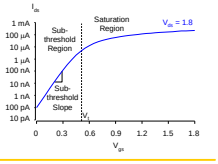
- there is a small amount of current that does flow when $V_{GS} < V_T$.

- we call this current *Subthreshold leakage current*.

- as devices get smaller, this current has become a non-negligible quantity.

- current in this region follows the relationship:

$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_t}{nv_T}} \left(1 - e^{\frac{-V_{ds}}{v_T}}\right)$$



- lowering the V_{T0} makes this problem worse

Small Geometry Challenges

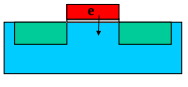
Oxide Breakdown

- the Oxide in a MOSFET serves as an insulator between the Gate electrode and the induced channel in the semiconductor

- as t_{ox} gets thinner and thinner, it becomes difficult to grow a planar surface. The *thin* parts of the non-planar oxide can be so thin that they will short out to the semiconductor

- another problem is that electrons can be excited enough in the Gate to have the energy to jump through the oxide.

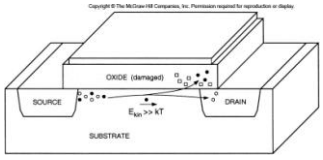
- this effectively shorts the Gate to the Source/Drain



Small Geometry Challenges

Hot Carrier Injection

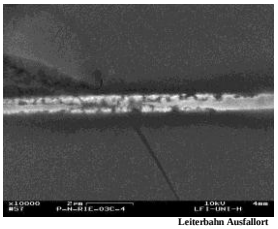
- as geometries shrink and doping densities increase, electrons can be accelerated fast enough to actually inject themselves into the oxide layer.
- this creates permanent damage to the oxide (effectively doping it to become a conductor)



Small Geometry Challenges

Electromigration

- when the metal interconnect gets smaller, its current density increases.
- the ions in the conductor will actually *move* due to the momentum of conducting electrons and diffusion metal atoms
- this can leave holes in the metals which lead to opens
- this can also build up regions of unwanted metal that may short to an adjacent trace

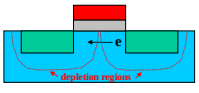


Leiterbahn Ausfallort

Small Geometry Challenges

Punch Through

- when the depletion regions around the Source and Drain get large enough to actually touch
- this is an extreme case of channel modulation
- this leads to a very large diffusion layer and causes a rapid increase in I_{DS} versus V_{DS}
- this limits the maximum operating voltage of the device in order to prevent damage due to the high electron acceleration



Small Geometry Challenges

Drain Induced Barrier Lowering (DIBL)

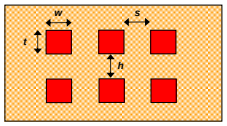
- if the gate length is scaled without properly scaling the Source/Drain regions, the Drain voltage will cause an un-proportionally large inversion layer
- this inversion interferes with the desired inversion layer being created by the Gate voltage
- this effectively lowers the Threshold voltage because it takes less energy to create inversion since the Drain is providing some inversion itself.

Small Geometry : Interconnect

- **Interconnect**

  - Quantities altered during fabrication

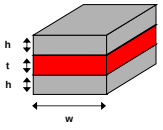
Quantity	Before Scaling	After Scaling
Width	*w*	*w' = w/S*
Spacing	*s*	*s' = s/S*
Thickness	*t*	*t' = t/S*
Interlayer oxide height	*h*	*h' = h/S*



Small Geometry : Interconnect

- **Scaling Effect on Interconnect**

  - Resistance, Capacitance, & Delay



$R \propto \frac{1}{w \cdot t}$ — **Resistance scales following : S^2** — **Horrible!!!**

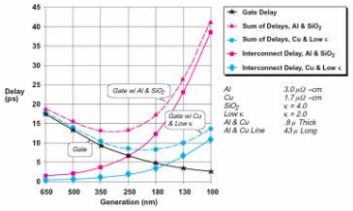
$C \propto \frac{w}{h}$ — **Capacitance scales following : 1** — **OK**

$\tau_{int} \propto \frac{1}{h \cdot t}$ — **Delay scales following : S^2** — **Horrible!!!**

Small Geometry : Interconnect

- **Interconnect Delay**

  - Device delay scales following 1/S
  - Interconnect delay scales following S^2



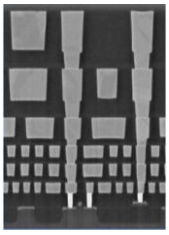
Interconnect Delay Dominates below 0.25um

Small Geometry : Interconnect

- **Interconnect Delay**

  - DSM Interconnect doesn't full scale due to resistance:
  - Interconnect structures are becoming "tall"
  - Moving up the Z-axis decreases density

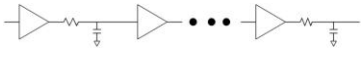
$$R \propto \frac{1}{w \cdot t}$$



Small Geometry : Interconnect

- **Interconnect Delay**

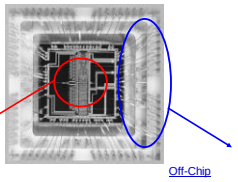
  - Repeaters are used to make the "delay vs. length" linear
  - Repeaters take power
  - Repeaters require diffusion layer access



Small Geometry : Interconnect

- **On-Chip vs. Off-Chip Performance Mismatch**

  - On-Chip and Off-Chip Features are not scaling at the same rate



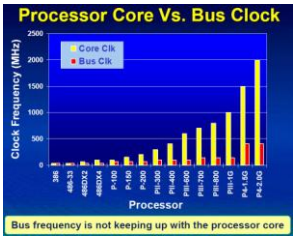
On-Chip
- f > 4GHz
- signal count scales exponentially
- Cheap

Off-Chip
- f < 2GHz
- signal count scales linearly (if that!)
- Expensive

Small Geometry : Interconnect

- **On-Chip vs. Off-Chip Performance Mismatch**

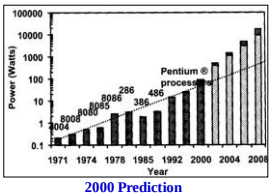
  - Getting data off-chip is the system bottleneck



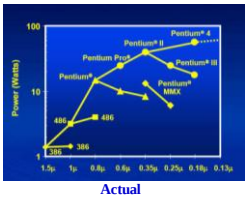
Small Geometry : Power

- **Power Consumption**

  - Dynamic Power scales at $1/S^2$ under "Full Scaling"
    but…
  - Full Scaling is impractical so we don't get full $1/S^2$ scaling
  - Die sizes are increasing ~25% per generation



2000 Prediction

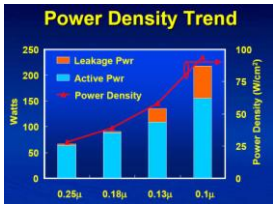


Actual

Small Geometry : Power

- **Power Consumption**

  - Lowering V_{TO} and V_{DD} reduces dynamic power
    but…
  - Leakage current increases exponentially

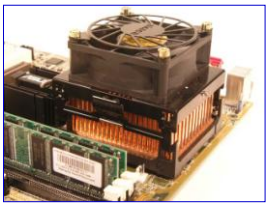


Approaching 50% in DSM

Small Geometry : Power

- **Power Consumption**

  - We're now breaking the 100W mark @ 40W/cm^2
  - Distribution and Cooling become very difficult (if not impractical)



MOSFET Capacitance

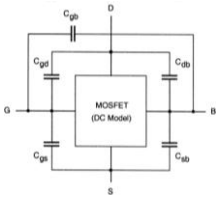
- **MOSFET Capacitance**

  - We have looked at device physics of the MOS structure
  - We have also looked at the DC I-V characteristics of the MOS Transistors
  - We have not looked at AC performance
  - Capacitance is the dominating imaginary component on-chip (i.e., we don't really have inductance)
  - the Capacitances of a MOSFET are considered *parasitic*
  - "parasitic" means *unwanted* or *unintentional*. They are unavoidable and a result of fabricating the devices using physical materials.
  - we can use the capacitances of the MOSFET to estimate factors such as rise time, delay, fan-out, and propagation delay

MOSFET Capacitance

- **MOSFET Capacitance**

  - Capacitance = Charge / Volt = (C/V)
  - as we've seen, the charge in a semiconductor is a complex, 3-dimensional, distribution due to the materials, doping, and applied E-field
  - we develop simple approximations for the MOSFET capacitances for use in hand calculations
  - we define each of the following lumped capacitance for an AC model of the transistors
  - each capacitance will have multiple contributions and different values depending on the state of the transistor (i.e., cutoff, linear, saturation)



MOSFET Capacitance

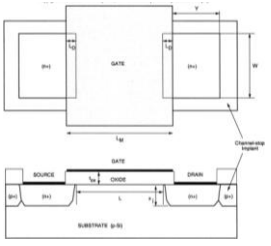
- **MOSFET Dimensions**

  - We need to define the geometric parameters present in the MOSFET structure

  Mask Length

  - we draw a gate length during fabrication
  - we call this the *Drawn Length*, L_M
  - in reality, the diffusion regions extend slightly under the gate by a distance, L_D
  - this is called *overlap*
  - the actual gate length (L) is given by:

$$L = L_M - 2 \cdot L_D$$



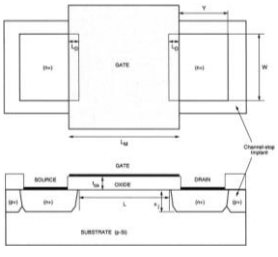
MOSFET Capacitance

- **MOSFET Dimensions**

  W = Channel Width
  t_{ox} = Oxide thickness
  x_j = diffusion region depth
  Y = diffusion region length

- **Channel-Stop Implants**

  - in order to prevent the n+ diffusion regions from adjacent MOSFETS from influencing each other, we use "channel-stop implants"
  - this is a heavily doped region of opposite typed material (i.e., p+ for an n-type)
  - these electrically isolate each transistor from each other



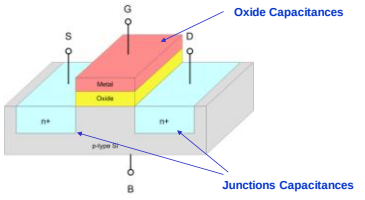
MOSFET Capacitance

- **MOSFET Capacitance**

  - We group the various capacitances into two groups

    1) Oxide Capacitances - capacitance due to the Gate oxide

    2) Junction Capacitances - capacitance due to the Source/Drain diffusion regions



Oxide-Related Capacitance

- **Oxide-Capacitance**

  - Oxide Capacitance refers to capacitance which uses the gate oxide as the insulator between the parallel plates of the capacitor

  - as a result, these capacitances always use the Gate as one of the terminals of the capacitor

  - we are concerned with the following capacitances:

    C_{gb} = Gate to Body capacitance
    C_{gd} = Gate to Drain capacitance
    C_{gs} = Gate to Source capacitance

  - again, each of these values will differ depending on the mode of operation of the MOSFET

Oxide-Related Capacitance

- **Overlap Capacitance**

  - capacitance from the Gate to the Source/Drain due to the overlap region (L_D)

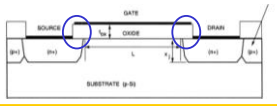
  - this creates:

$$C_{gs}(overlap) = C_{ox} \cdot W \cdot L_D$$
$$C_{gd}(overlap) = C_{ox} \cdot W \cdot L_D$$

  - where C_{ox} is the unit-area capacitance (i.e., multiply by area to find total capacitance, F/m^2 or F/um^2)

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

  - NOTE : this capacitance does NOT depend on the external bias of the MOSFET since the Gate and the Source/Drain do not have their carrier density altered during bias.

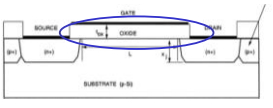


Oxide-Related Capacitance

- **Gate-to-Channel Capacitance**

  - the gate-to-channel configuration results in 3 capacitances (C_{gb}, C_{gs}, C_{gb})

  - these capacitances change as a result of external bias since in effect, the "bottom plate" of the capacitor is being *moved* around during depletion/inversion

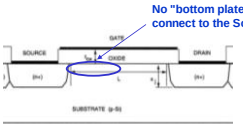


Oxide-Related Capacitance (Cut-Off)

- **Gate to Source Capacitance (C_{gs}) : Cut-Off**

  - During Cut-off, there is no channel beneath the Gate.

  - since there is no channel that links the Gate to the Source (i.e., no ΔQ), there is no Gate-to-Channel capacitance.



  - this leaves the overlap capacitance as the only component to C_{gs} in cut-off:

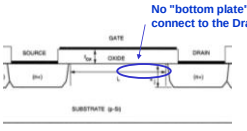
$$C_{gs_{(cut-off)}} = C_{ox} \cdot W \cdot L_D$$

Oxide-Related Capacitance (Cut-Off)

- **Gate to Drain Capacitance (C_{gd}) : Cut-Off**

  - Just as C_{gs}, during Cut-off, there is no channel beneath the Gate.

  - since there is no channel that links the Gate to the Drain (i.e., no ΔQ), there is no Gate-to-Channel capacitance.

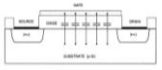


  - this leaves the overlap capacitance as the only component to C_{gd} in cut-off:

$$C_{gd_{(cut-off)}} = C_{ox} \cdot W \cdot L_D$$

Oxide-Related Capacitance (Cut-Off)

- **Gate to Body Capacitance (C_{gb}) : Cut-Off**
  - There is a capacitor between the Gate and Body
  - The bottom plate is the conductor formed by the p-type silicon since it has majority charge carriers and acts as a conductor



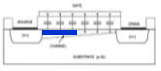
  - we can describe the Gate-to-Body Capacitance as:

$$C_{gb_{(cut-off)}} = C_{ox} \cdot W \cdot L$$

  - remember that L = (L_M-2·L_D)

Oxide-Related Capacitance (Linear)

- **Gate to Source Capacitance (C_{gs}) : Linear Region**
  - When operating in the linear region, a channel is present in the substrate.
  - this can be thought of as a conductor (or metal plate) that contacts the Source and Drain
  - this results in a capacitance between the Gate and the Source/Drain
  - we split this capacitance between the Source and Drain for simplicity



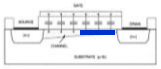
  - the Gate-to-Channel contribution to C_{GS} is (1/2)C_{ox}WL
  - the total C_{GS} capacitance in the linear region includes the overlap capacitance:

$$C_{gs_{(linear)}} = \frac{1}{2} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$$

Oxide-Related Capacitance (Linear)

- **Gate to Drain Capacitance (C_{gd}) : Linear Region**
  - The Gate-to-Drain Capacitance is identical to the Gate-to-Source Capacitance in the Linear region:

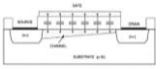
$$C_{gd_{(linear)}} = \frac{1}{2} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$$



Oxide-Related Capacitance (Linear)

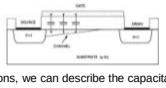
- **Gate to Body Capacitance (C_{gb}) : Linear Region**
  - Since the channel (inversion layer) looks like a metal plate to the gate, the Gate can't actually *see* the substrate anymore
  - this means that the capacitance between the Gate and Body is zero when a channel is present

$$C_{gb_{(linear)}} = 0$$



Oxide-Related Capacitance (Saturation)

- **Gate to Source Capacitance (C_{gs}, C_{gd}, C_{gb}) : Saturation Region**
  - When operating in the saturation region, the channel is pinched off
  - We can make the assumptions that:
    1) There is no longer a link between the Gate and Drain
    2) Roughly 2/3 of the channel is still present linking the Gate to the Source (2L/3)
    3) The pinched off channel still effectively shields the gate from the body



  - from these approximations, we can describe the capacitances in the saturation region

$$C_{gs_{(sat)}} = \frac{2}{3} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$$

$$C_{gd_{(sat)}} = C_{ox} \cdot W \cdot L_D$$

$$C_{gb_{(sat)}} = 0$$

Oxide-Related Capacitance (Summary)

- **Summary of Oxide-Related Capacitance**

Cut-off	Linear	Saturation
$C_{gs_{(cut-off)}} = C_{ox} \cdot W \cdot L_D$	$C_{gs_{(linear)}} = \frac{1}{2} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$	$C_{gs_{(sat)}} = \frac{2}{3} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$
$C_{gd_{(cut-off)}} = C_{ox} \cdot W \cdot L_D$	$C_{gd_{(linear)}} = \frac{1}{2} \cdot C_{ox} \cdot W \cdot L + C_{ox} \cdot W \cdot L_D$	$C_{gd_{(sat)}} = C_{ox} \cdot W \cdot L_D$
$C_{gb_{(cut-off)}} = C_{ox} \cdot W \cdot L$	$C_{gb_{(linear)}} = 0$	$C_{gb_{(sat)}} = 0$

Oxide-Related Capacitance (Total)

- **Total of Oxide-Related Capacitance**

  - if we assume that these three capacitances are in parallel, then their total values add:

$$C_{oxide} = C_{gs} + C_{gd} + C_{gb}$$

  - the **lowest** oxide-related capacitance that is present is in the *saturation* region:

$$C_{oxide_{(max)}} = \frac{2}{3} \cdot C_{ox} \cdot W \cdot L + 2 \cdot C_{ox} \cdot W \cdot L_D = 0.66 \cdot C_{ox} \cdot W \cdot (L + 3 \cdot L_D)$$

  - the **largest** oxide-related capacitance that is present is in the *cut-off* & the *linear* regions:

$$C_{oxide_{(max)}} = C_{ox} \cdot W \cdot (L + 2 \cdot L_D)$$

  - for quick hand-calculations, we can use the largest oxide capacitance to find a worst-case value

Junction Capacitance

- **Junction Capacitance**

  - Junction Capacitance refers to capacitance between the diffusion regions of the Source & Drain to the doped substrate surrounding them.

  - they are called "junction" because these capacitances are due to the PN junctions that are formed between the two materials

  - we are concerned with the following junction capacitances:

    C_{sb} = Source to Body capacitance
    C_{db} = Drain to Body capacitance

  - these capacitances are highly dependant on the bias voltages since the effective distance between *plates* is the depth of the built in depletion region that forms at the PN junction

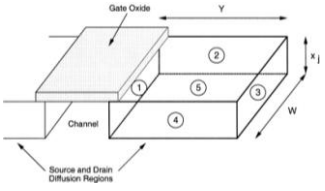
Junction Capacitance

- **Junction Capacitance**

  - the Source and Drain regions will have similar geometries so we will start by describing the PN junctions for only one region

  - Consider the numbers in the following figures illustrating the PN junctions that exist

  - Let's start with an N-type MOSFET and identify all of PN junctions



Junction Capacitance

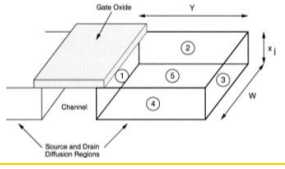
- **Junction Capacitance**

  - remember that the MOSFET is surrounded by a channel-stop implant to prevent the diffusion regions from coupling to other MOSFETs.

  - This implant is heavily doped (p+), usually $10 \cdot N_A$

  - These areas are also called *sidewalls*

  - remember that the diffusion regions are heavily doped (n+)



Junction Capacitance

- **Junction Capacitance**

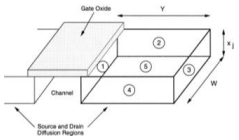
  1) n+ / p junction = diffusion region to substrate beneath gate

  2) n+ / p+ junction = diffusion region to channel-stop implant in back (sidewall)

  3) n+ / p+ junction = diffusion region to channel-stop implant on side (sidewall)

  4) n+ / p+ junction = diffusion region to channel-stop implant in front (sidewall)

  5) n+ / p junction = diffusion region to substrate underneath



Junction Capacitance

- **Junction Capacitance**

  - the capacitance will be proportional to the area of the junction

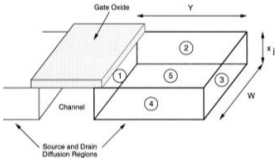
  1) Area = $W \cdot x_j$

  2) Area = $Y \cdot x_j$

  3) Area = $W \cdot x_j$

  4) Area = $Y \cdot x_j$

  5) Area = $W \cdot Y$



Junction Capacitance

- **Junction Capacitance**

  - first we need to express the capacitance for an abrupt, PN junction under reverse-bias

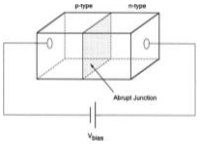
  - we begin with finding the depletion region thickness

  - this is similar to the expression for depletion thickness of a MOS structure, except that the region will protrude into both materials instead of just the semiconductor as before.

  - as a result, the carrier concentration of both materials is now described: $\frac{N_A \cdot N_D}{N_A + N_D}$

  - the depletion thickness is given by:

$$x_{d_{pn}} = \sqrt{\frac{2 \cdot \varepsilon_{Si}}{q} \cdot \frac{N_A + N_D}{N_A \cdot N_D} \cdot (\phi_0 - V)}$$



  - where the built in junction potential is given by:

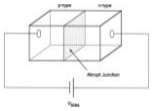
$$\phi_0 = \frac{k_B \cdot T}{q} \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right)$$

Junction Capacitance

- **Junction Capacitance**

  - the depletion-region charge (Q_j) can be written as:

$$Q_j = A \cdot q \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot x_{d_{pn}}$$



  - substituting $x_{d_{pn}}$ and rearranging terms, we get:

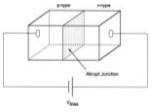
$$Q_j = A \cdot \sqrt{2 \cdot \varepsilon_{Si} \cdot q \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot (\phi_0 - V)}$$

Junction Capacitance

- **Junction Capacitance**

  - the capacitance of the junction is defined as:

$$C_j = \left|\frac{dQ_j}{dV}\right|$$



  - we can differentiate our expression for junction charge with respect to voltage to get the capacitance as a function of junction voltage:

$$C_j = \left|\frac{dQ_j}{dV}\right| = \frac{d}{dV}\left(A \cdot \sqrt{2 \cdot \varepsilon_{Si} \cdot q \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot (\phi_0 - V)}\right)$$

$$C_j(V) = A \cdot \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot (\phi_0 - V)}$$

Junction Capacitance

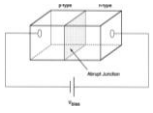
- **Junction Capacitance**

  - from this expression, we can define the *zero-bias junction capacitance* per unit area:

$$C_{j0} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot \frac{1}{\phi_0}}$$

  - putting this back into a more generic expression for $C_j(V)$, we get:

$$C_j(V) = \frac{A \cdot C_{j0}}{\sqrt{1 - \frac{V}{\phi_0}}}$$



  - remember that we are assuming an "abrupt" PN junction

Junction Capacitance

- **Junction Capacitance**

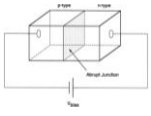
  - since the total capacitance depends on the external bias voltage, it can be a complicated to find the equivalent capacitance when the bias voltage is a transient.

  - we need to make an assumption to simplify the expression.

  - let's assume that the voltage change across the junction is linear. Then we can find the equivalent or average capacitance using:

$$C_{eq} = \frac{\Delta Q}{\Delta V} = \frac{Q_j(V_2) - Q_j(V_1)}{V_2 - V_1}$$

$$C_{eq} = \frac{1}{V_2 - V_1} \cdot \int_{V_1}^{V_2} C_j(V) \cdot dV$$



Junction Capacitance

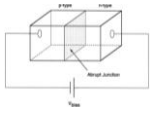
- **Junction Capacitance**

  - substituting in our expression for $C_j(V)$, we get:

$$C_{eq} = -\frac{2 \cdot A \cdot C_{j0} \cdot \phi_0}{(V_2 - V_1)} \cdot \left[\sqrt{1 - \frac{V_2}{\phi_0}} - \sqrt{1 - \frac{V_1}{\phi_0}}\right]$$

  - which we can simplify even further by defining a dimensionless coefficient K_{eq}

$$C_{eq} = A \cdot C_{j0} \cdot K_{eq}$$



  - where K_{eq} is the *voltage equivalence factor* ($0<K_{eq}<1$):

$$K_{eq} = -\frac{2 \cdot \sqrt{\phi_0}}{(V_2 - V_1)} \cdot \left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1}\right)$$

Junction Capacitance

- **Junction Capacitance**

  - How do we use this?

    - for a given diffusion region, we calculate C_{eq} for each of the 5 PN junctions

    - note that there really are only two different regions (n+ / p and n+ / p+)

    - the sidewalls (2,3,4) will have their own zero-bias junction capacitance since they have a unique carrier concentration (i.e., n+ / p+).

    - the inner and bottom junctions (1,5) will have their own zero-bias junction capacitance since they have a unique carrier concentration (i.e., n+ / p).

    - when solving for the sidewall contribution, you can add the areas for 2,3,4 and solve once

    - when solving for the inner and bottom junction contributions, you can add the areas for 1 and 5 and solve once

    - since this is a reverse biased PN junction the voltages for V1 and V2 are actually negative when plugged into the K_{eq} expression

Junction Capacitance

- **Junction Capacitance**

  - What's the difference between the Source and Drain?

    - If the Source is grounded, then there is no voltage change across it. This means its capacitance is simply the *zero-bias capacitance*

    - you will still need to calculate the sidewall and inner/bottom C_{j0} capacitances separately

    - the drain typically sees a voltage change (V_{DS}). However, one good thing is that typically the source voltage is 0v, so the expression simplifies somewhat

Junction Capacitance

- **Junction Capacitance**

  - Doesn't this take a lot of time?

    - Yes! And remember that we have made a lot of assumptions along the way

    - For this reason, we typically rely on computer models of the capacitance

    - We do the hand calculations to get a *gut feel* for what factors affect capacitance

    - *Gut Feel* makes for good designers because design is about balancing trade-offs

    - If you don't have *Gut Feel* and rely totally on simulators, you will struggle when asked to innovate and trouble-shoot.