Tutorial for PADRE-Based Simulation Modules

(PN Junction Lab, MOSCap Lab, BJT Lab, MOSFET Lab, MESFET Lab)

Dragica Vasileska (ASU) and Gerhard Klimeck (Purdue)

July 2009

Work Supported by NSF Under Contract Number:

Table of Contents

1. Why is Simulation Important?

1.1 Silicon-Based Nanoelectronics

Semiconductor device-based electronics industry is the largest industry in the world with global sales of over one trillion dollars since 1998. If current trends continue, the sales volume of the electronics industry will reach three trillion dollars and will constitute about 10% of the gross world product (GWP) by 2010 [1]. The revolution in the semiconductor industry, a subset of the electronics industry, began in 1947 (see Figure 1) with the fabrication of bipolar devices on slabs of polycrystalline germanium (Ge) [2].

Bipolar transistor:	1947
Monocrystal germanium:	1950
First good BJT:	1951
Monocrystal silicon:	1951
Oxide mask, Commercial silicon BJT:	1954
Transistor with diffused base:	1955
Integrated circuit:	1958
Planar transistor:	1959
Planar integrated circuit:	1959
Epitaxial transistor:	1960
MOS FET:	1960
Schottky diode:	1960
Commercial integrated circuit (RTL):	1961
DTL - technology	1962
TTL - technology	1962
ECL - technology	1962
MOS integrated circuit	1962
CMOS	1963
Linear integrated circuit	1964
MSI circuits	1966
MOS memories	1968
LSI circuits	1969
MOS processor	1970
Microprocessor	1971
I^2L	1972
VLSI circuits	1975
Computers using VLSI technology	1977
...	

Figure 1: Some Historic Dates.

Single-crystalline materials were later proposed and introduced, making possible the fabrication of grown junction transistors. Migration to silicon (Si)-based devices was initially hindered by the stability of the Si/SiO$_2$ materials system, necessitating a new generation of crystal pullers with improved environmental controls to prevent SiO$_2$ formation. Later, the stability and low interface-state density of the Si/SiO$_2$ materials system allowed for the passivation of junctions and eventually the migration from bipolar devices to field-effect devices in 1960. By 1968, both complementary metal–oxide–semiconductor devices (CMOS) and polysilicon gate technology, which allowed self-alignment of the gate to the source/drain of the device, had been developed. These innovations permitted a significant reduction in power dissipation and a reduction of the device overlap capacitance, improving frequency performance and resulting in the essential components of the modern CMOS device. Professor Herbert Kroemer’s contributions to heterostructures—from heterostructure bipolar transistors [3] to lasers [4]—culminated in a Nobel Prize in Physics in 2000 and have paved the way for novel heterostructure devices, including those in silicon. The unique properties of the variety of semiconductor materials have enabled the development of a wide variety of ingenious devices that have literally changed our world. To date, there are about 60 major devices, with over 100 device variations related to them.

The metal-oxide-semiconductor-field-effect transistor (MOSFET) and related integrated circuits now constitute about 90% of the semiconductor device market. Combining silicon with the elegance of the field-effect transistor (FET) structure has allowed devices to be simultaneously made smaller, faster, and cheaper—the mantra that has driven the modern semiconductor microelectronics industry. Nowadays, the single factor driving the continuous device improvement is the semiconductor industry's relentless effort to reduce the cost per function on a chip. This is done by putting more devices on a chip while either reducing manufacturing costs or holding them constant. This leads to three methods of reducing the cost per function. The first method is transistor scaling, which involves reducing the transistor size in accordance with some goal (i.e., keeping the electric field constant from one generation to the next). With smaller transistors, more can fit into a given area than in previous generations. The second method is circuit cleverness, which is associated with the physical layout of the transistors with respect to each other. If the transistors can be packed into a tighter space, then more devices can fit into a given area than before. The third method for reducing cost per function is to make the die larger because more devices can be fabricated on a larger die. All the while, the semiconductor industry is constantly looking for technological breakthroughs to decrease the manufacturing cost. All of this effort serves to reduce the cost per function on a chip.

1.2 Need for Simulation

The standard sequence that one follows when modeling device structures of interest involves (1) process simulation step that is followed by a (2) device simulation and is finalized with a (3) circuit simulation step. In this regard, device simulation is the process of using computers to calculate the behavior of electronic devices, i.e. calculating the current-voltage (*IV*) curves of a device in general. The devices are defined mathematically in terms of their dimension, material composition, and other relevant physical information, all of which is obtained from the process simulation step.

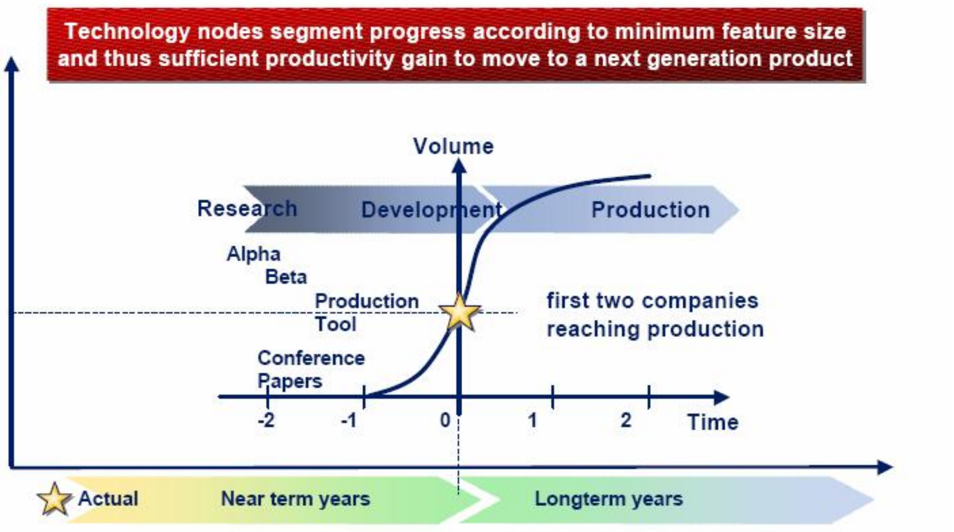

Technology nodes segment progress according to minimum feature size and thus sufficient productivity gain to move to a next generation product
Volume
Research
Development
Production
Alpha
Beta
Production Tool
first two companies reaching production
Conference Papers
-2
-1
0
1
2
Time
Actual
Near term years
Longterm years

Figure 2: ITRS Technology node and related timeline.

There are two issues that make simulation an important step for industry. Consider the product cycle the first issue. Product cycles are getting shorter with each generation, and the demand for production wafers shadows development efforts in the factory. In order for companies to maintain their competitive edge, products have to be taken from design to production in less than 18 months. As a result of this production cycle, the development phase of the cycle is getting shorter. Contrast this requirement with the fact that it takes 2-3 months to run a wafer lot through a factory, depending on the wafers' complexity. The specifications for experiments run through the factory must be near the final phase of the product cycle. While simulations may not be completely predictive, they provide a good initial guess. This process can ultimately reduce the number of iterations during the device development phase. The second issue that reinforces the need for simulation is the production pressures that factories face. In order to meet customer demand, development factories are making way for production space. It is also expensive to run experiments through a production facility because the resources could have otherwise been used to produce sellable product. Again, device simulation can be used to decrease the number of experiments run through a factory. Device simulation can be used as a tool to guide the manufacturing of more robustly designed devices, thereby decreasing the development time and costs (see Figure 2). Besides offering the possibility to test hypothetical devices which have not (or could not) yet been manufactured, device simulation offers unique insight into device behavior by allowing the observation of phenomena that cannot be measured on real devices. It is related to but usually separate from process simulation, which deals with various physical processes such as material growth, oxidation, impurity diffusion, etching, and metal deposition inherent in device fabrication leading to integrated circuits. Device simulation is distinct from another important aspect of computer-aided design (CAD) device modeling, which deals with compact behavioral models for devices and sub-circuits relevant for circuit simulation in commercial packages such as SPICE.

The main components of semiconductor device simulation at any level of approximation are illustrated in Figure 3 [5]. There are two main kernels that must be solved self-consistently with one another, the *transport equations* governing charge flow and the *fields* driving charge flow. Both are coupled strongly to one another and hence must be solved simultaneously. The fields arise from external sources, as well as the charge and current densities which act as sources for the time varying electric and magnetic fields obtained from the solution of Maxwell's equations. Under appropriate conditions, only the quasi-static electric fields arising from the solution of Poisson's equation are necessary. The fields, in turn, are driving forces for charge transport as illustrated in Figure 4 for the various levels of approximation within a hierarchical structure ranging from compact modeling at the top to an exact quantum mechanical description at the bottom.

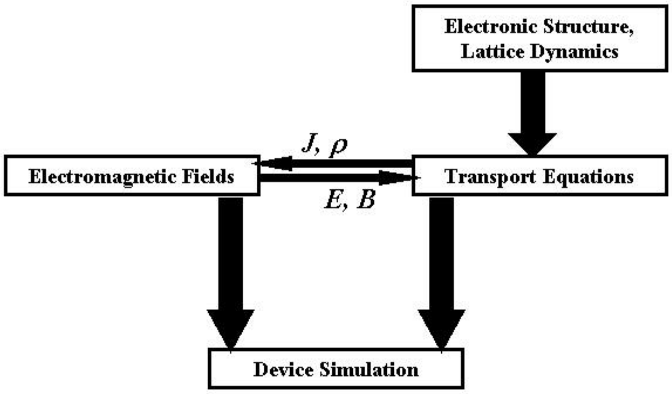
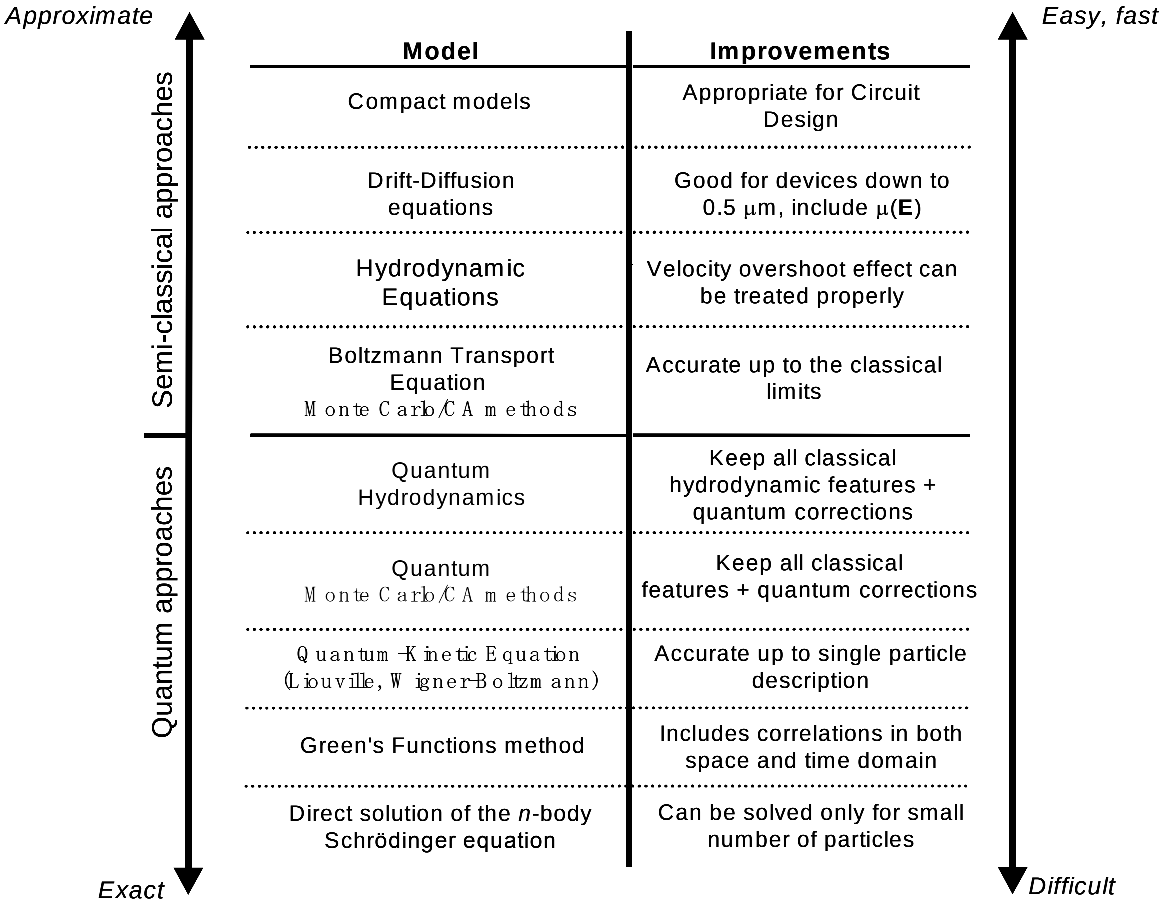


Figure 3: A schematic description of the device simulation sequence.

Note that semiclassical Boltzmann transport has been the mainstay of the semiconductor technology from its early development. Up until now, most device simulations including the full-band Monte Carlo (FBMC) method are based on the solution of the Boltzmann transport equation (BTE) and its simplifications, the hydrodynamic (HD) transport equations and the drift-diffusion (DD) model. But in the last decade, as semiconductor technology has continued to pursue the downscaling of device dimensions into the nanoscale regime, many new and interesting questions have emerged concerning the physics of small devices. Ref. [5] highlights some of the basic physical effects that are viewed as important in nanoelectronics research.

2. PADRE-Supported Simulation Examples

In this section, we want to illustrate some basic PADRE simulation package capabilities. PADRE simulation software developed at Bell Labs has the option to either solve the drift-diffusion model or the hydrodynamic equations. Section 3 clearly illustrates on the example of a simple MOSFET device the limitations of the drift-diffusion model that has been the model of choice for the semiconductor industry until recently. Capabilities of the hydrodynamic model and its limitations are also discussed, and the origin of the deficiencies in the model is clearly explained. For more information on the syntax of PADRE, the user is referred to the PADRE Manual that can be found on the nanoHUB. The manual is especially useful for those users who want to write the input deck themselves and run the PADRE Lab. With the examples that follow, we cover some very important properties of the most commonly used semiconductor devices [6] in real practical applications.



Model	Improvements
Compact models	Appropriate for Circuit Design
Drift-Diffusion equations	Good for devices down to 0.5 μm, include μ(**E**)
Hydrodynamic Equations	Velocity overshoot effect can be treated properly
Boltzmann Transport Equation Monte Carlo/CA methods	Accurate up to the classical limits
Quantum Hydrodynamics	Keep all classical hydrodynamic features + quantum corrections
Quantum Monte Carlo/CA methods	Keep all classical features + quantum corrections
Quantum-Kinetic Equation (Liouville, Wigner-Boltzmann)	Accurate up to single particle description
Green's Functions method	Includes correlations in both space and time domain
Direct solution of the *n*-body Schrödinger equation	Can be solved only for small number of particles

Figure 4: Illustration of the hierarchy of transport models.

2.1 Examples for PN-Junctions

We begin the section with examples that cover basic operation of pn-junctions. We first discuss the ideal diode characteristics and compare them with analytical expressions. With the second example, we illustrate the importance of the generation-recombination mechanisms in the diode operation, high-level injection condition, and the appearance of the series resistance effects.

2.1.1 Ideal diode characteristics at equilibrium and comparison to analytical expressions

In this first example, we calculate the equilibrium diode characteristics and compare them with analytical depletion charge approximation results. The potential and the electric field profiles are shown in Figure 5. From these results we can deduce two things: (1) for low doping concentrations, the analytical and the simulated values agree for both the electrostatic potential and the electric field profile; and (2) the electric field profile is an excellent indicator for the extension of the depletion region and the presence of a net space charge.

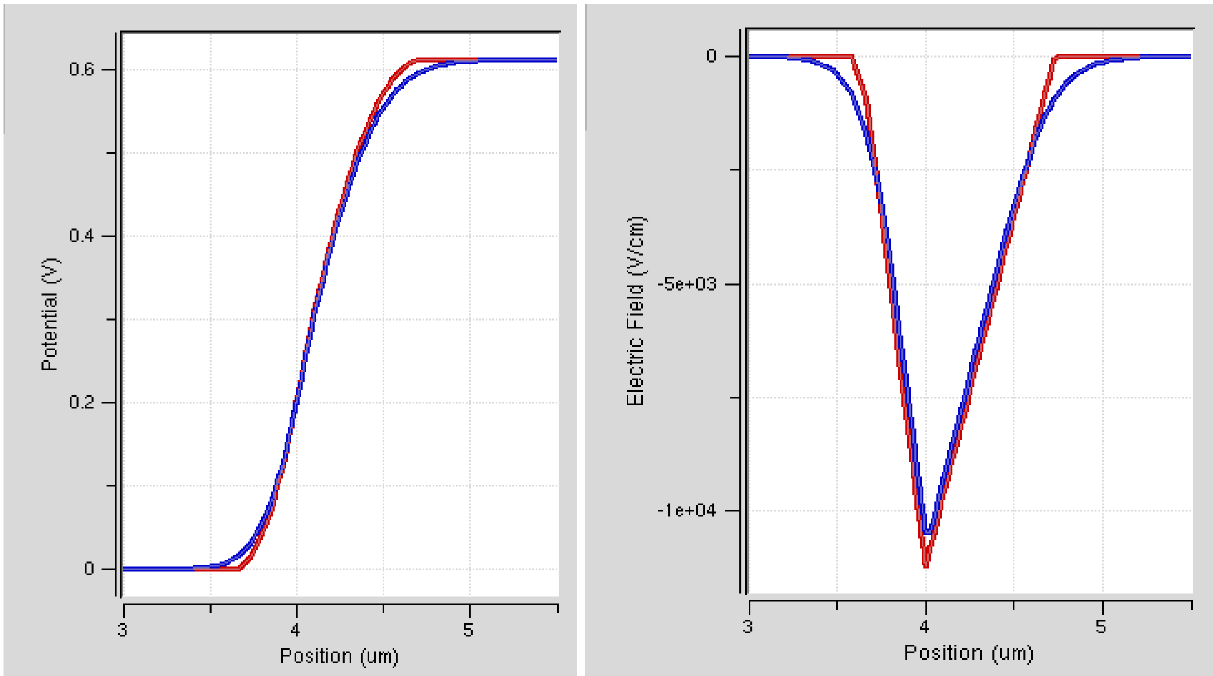


Figure 5: Electrostatic potential (left panel) and electric field profile (right panel) at equilibrium for a diode with $N_A=2\times10^{15}$ cm^{-3} and $N_D=10^{15}$ cm^{-3}.

2.1.2 Non-idealities in pn-diodes

In the derivation of the ideal diode characteristics, it is generally assumed that there is no generation-recombination in the depletion region, and because of that, the minority carrier currents are constant throughout the depletion region, as is illustrated schematically in Figure 6 below.

In reality, this is not the case as the minority carriers have finite lifetimes, and there is always some generation-recombination in the depletion region denoted as space-charge region (SCR) in Figure 6. Shockley-Read-Hall (SRH) recombination process dominates under forward bias conditions, and the SRH generation process dominates under reverse bias conditions. Another diode non-ideality is high levels of injection, which occurs under high forward bias and is a condition in which the excess minority concentration in the quasi-neutral region becomes comparable to the majority carrier concentration. Finally we have the series resistance effect that starts to play role under low doping conditions and/or high biases. It arises because of the finite resistance of the quasi-neutral regions.

To illustrate these effects, let us consider a pn-junction with doping $N_A=N_D=10^{16}$ cm^{-3}. We will apply high bias to the diode and examine the forward bias diode characteristics for three different values of the minority carrier lifetimes: 1 μs, 0.1 μs, and 0.01 μs. In Figure 7, we plot the simulation results for the anode current under forward bias conditions for the case of minority carrier lifetime = 0.01 μs, 0.1 μs, and 1 μs.

What is interesting about the results presented in Figure 7? Let us first focus on the figure on the left, where we have used very small carrier lifetime (0.01 μs). We see sharp increase in current when the voltage turns on, which is followed with a region in which recombination dominates and the current increase with increasing the bias is not as fast as in the normal diode case. This region is characterized with diode ideality factor between 1 and 2, depending upon the importance of the carriers' recombination process. The recombination-dominated region is followed by an ideal diode region because the bias is large enough that the carriers do not spend much time in the depletion region and the recombination process eventually vanishes. High level of injection is next, characterized with ideality factor larger than 1 and the last region is the series resistance dominated region. The interesting question to ask here is: What happens when carrier lifetime is larger? The trend is clearly seen on the figure on the right, where we use carrier lifetime of 1 μs. We see no recombination-dominated region. The figure in the middle that was generated by using carrier lifetime of 0.1 μs is the scenario between very small carrier lifetime and very large carrier lifetime. Thus, from the example presented here, we might conclude that the lifetimes of the carriers play important role in the operation of pn-diodes.

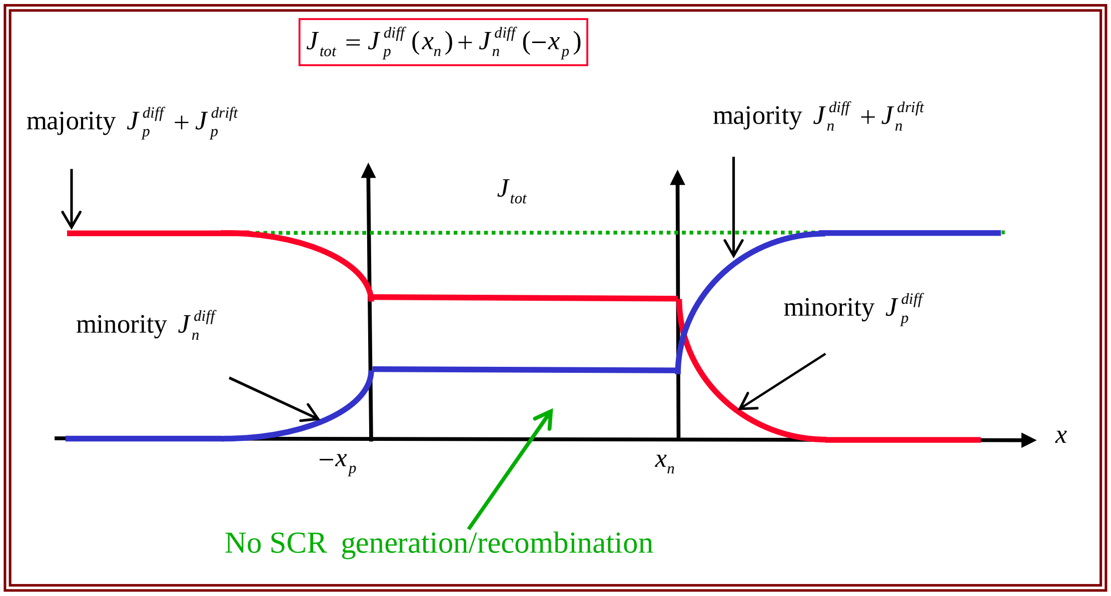

$J_{tot} = J_p^{diff}(x_n) + J_n^{diff}(-x_p)$
majority $J_p^{diff} + J_p^{drift}$
majority $J_n^{diff} + J_n^{drift}$
J_{tot}
minority J_n^{diff}
minority J_p^{diff}
$-x_p$
x_n
x
No SCR generation/recombination

Figure 6: Schematic of the current components in an ideal diode under forward bias conditions.

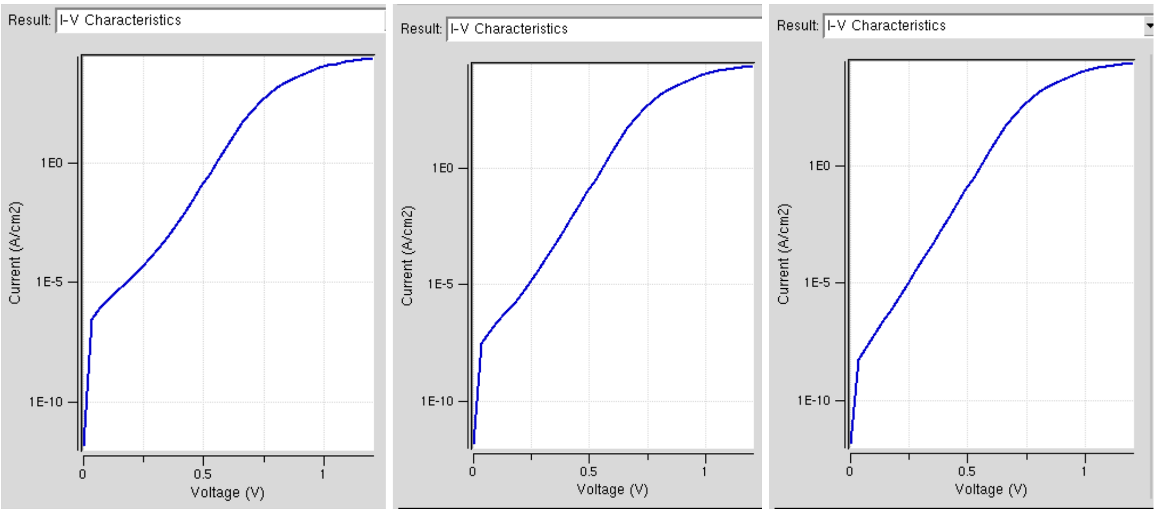
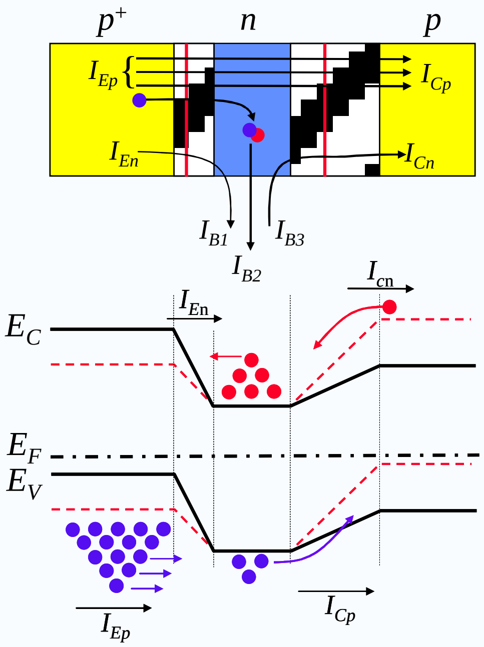


Figure 7: Semi-log plots of the anode current under forward bias conditions for the case of minority carrier lifetime = 0.01 μs, 0.1 μs, and 1 μs (left to right).

2.2 Examples for BJTs

2.2.1 Gummel plot, output characteristics, and current gain

The qualitative description of the transistor operation along with some notation that corresponds to pnp transistor operation is shown in the Figure 8.



- Emitter doping is much larger than base doping
- Base doping larger than collector doping
- Current components:

$$I_E = I_{Ep} + I_{En}$$
$$I_C = I_{Cp} + I_{Cn}$$
$$I_B = I_E - I_C = I_{B1} + I_{B2} - I_{B3}$$

- ***I_{B1}*** = current from electrons being back injected into the forward-biased emiter-base junction
- ***I_{B2}*** = current due to electrons that replace the recombined electrons in the base
- ***I_{B3}*** = collector current due to thermally-generated electrons in the collector that go in the base

Figure 8: Definition of emitter, base, and collector current components and their physical origin.

Regarding circuit definitions, the base transport factor is calculated using

$$\alpha_T = I_{Cp} / I_{Ep} \text{ for } \textit{pnp} \text{ and } \alpha_T = I_{Cn} / I_{En} \text{ for } \textit{npn} \text{ transistor.}$$

If there were no recombination in the base, the base transport factor would be equal to unity. On the other hand, the emitter injection efficiency is calculated according to

$$\gamma = \frac{I_{Ep}}{I_{Ep} + I_{En}} = \frac{I_{Ep}}{I_E} \text{ for } \textit{pnp} \text{ and } \gamma = \frac{I_{En}}{I_{En} + I_{Ep}} = \frac{I_{En}}{I_E} \quad \text{for } \textit{npn} \text{ transistor,}$$

and it approaches unity when the emitter doping is much higher than the base doping. The current amplification under DC operating conditions for a common base configuration is then defined as

$$\alpha_{dc} = \frac{I_C}{I_E} = \frac{I_{Cp} + I_{Cn}}{I_{Ep} + I_{En}} \approx \frac{I_{Cp}}{I_{Ep} + I_{En}} = \alpha_T \gamma \text{ for } \textit{pnp},$$

$$\alpha_{dc} = \frac{I_C}{I_E} = \frac{I_{Cp} + I_{Cn}}{I_{Ep} + I_{En}} \approx \frac{I_{Cn}}{I_{Ep} + I_{En}} = \alpha_T \gamma \text{ for } npn \text{ transistor.}$$

The corresponding current amplification for current emitter configuration is

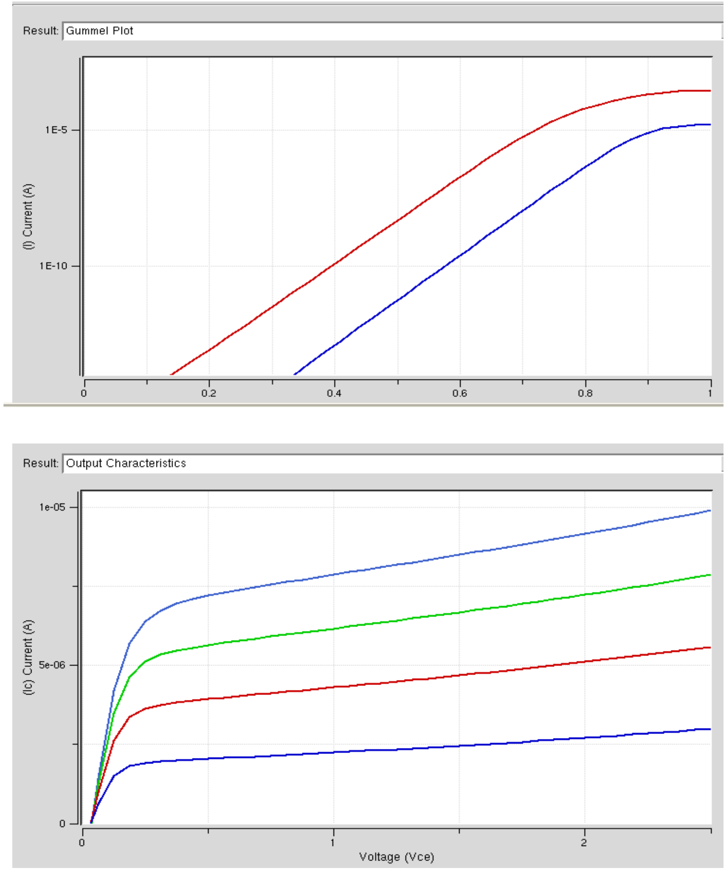
$$\beta_{dc} = \frac{I_C}{I_B} = \frac{I_C}{I_E - I_C} = \frac{\alpha_{dc}}{1-\alpha_{dc}}.$$

What follows is an example of how one can optimize a BJT to get higher and higher current gain in the common emitter configuration. We start with an *npn* transistor with the following parameters: emitter length 0.2 um, base length 0.6 um, base contact length 0.4 um, and collector length 0.8 um. The emitter, base, and collector doping are 10^{18} cm^{-3}, 10^{16} cm^{-3}, and 10^{15} cm^{-3}, respectively. The Gummel plot that shows the dependence of the I_C and I_B current versus the base voltage is given in the Figure 9 top panel. Parameter is V_{CE}=2.5 V. The output characteristics of the BJT are given in the Figure 9 bottom panel. The parameter here is the base current that varies from 1 uA to 4 uA in 1 uA increments. We see that there is a finite output conductance, which means that this is not a well-designed transistor.

From the current density plot of this device, we have: $\alpha_T = I_{Cn} / I_{En}$ =0.889. This value for the base transport factor suggests that there is large recombination effect taking place in the base of this transistor. The emitter injection efficiency is almost unity as there is large difference between the emitter and the base doping (two orders of magnitude difference in doping). Thus $\gamma = I_{En} \big/ I_E$ =0.994 as expected. The $\alpha_{dc} = \alpha_T \gamma$ =0.884 gives current gain in common emitter configuration of $\beta_{dc} = \alpha_{dc} / \left(1-\alpha_{dc}\right)$ =7.6, which is very low.

A strategy that one can follow to improve the current gain in the common base configuration is to first think of reducing the length of the base. Thus, we consider a second *npn* BJT with all parameters the same except the base width is 0.1 um and the base contact width is 0.08 um. With this choice of parameters, we get: $\alpha_T = I_{Cn} / I_{En}$ =0.917, $\gamma = I_{En} \big/ I_E$ =0.997, $\alpha_{dc} = \alpha_T \gamma$ =0.9145, which gives current gain in common emitter configuration of $\beta_{dc} = \alpha_{dc} / \left(1-\alpha_{dc}\right)$ =10.696 . There are two problems with this transistor configuration: (1) the current gain is still low and (2) the output conductance (see output characteristics in Figure 10) is enormous. To improve this situation, we need to increase the base doping which, in turn, requires that we increase the emitter doping to keep the emitter injection efficiency at a value very close to unity.

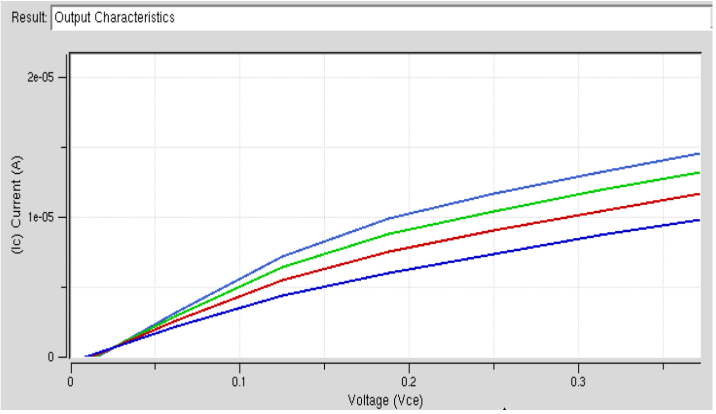
Thus, the third BJT that we consider has the following parameters: emitter length 0.2 um, base length 0.1 um, base contact length 0.08 um, and collector length 0.8 um. The emitter, base, and collector doping are 10^{20} cm^{-3}, 5×10^{17} cm^{-3}, and 10^{15} cm^{-3}, respectively. The Gummel plot that shows the dependence of the I_C and I_B current versus the base voltage is given in the top panel of Figure 11. Parameter is V_{CE}=2.5 V. The output characteristics of the BJT are given in the bottom panel of Figure 11. Parameter here is the base current that varies from 1 uA to 4 uA in 1 uA increments. We see that there is a almost zero output conductance, which means that this is a well-designed transistor. Let us look at the current gain of this transistor: $\alpha_T = I_{Cn} / I_{En}$ =0.96, $\gamma = I_{En} \big/ I_E$ =1, $\alpha_{dc} = \alpha_T \gamma$ =0.96, which gives current gain in common emitter configuration of $\beta_{dc} = \alpha_{dc} / \left(1-\alpha_{dc}\right)$ =24 . This is much better designed device in terms of both output characteristics and the current gain.



Result: Gummel Plot

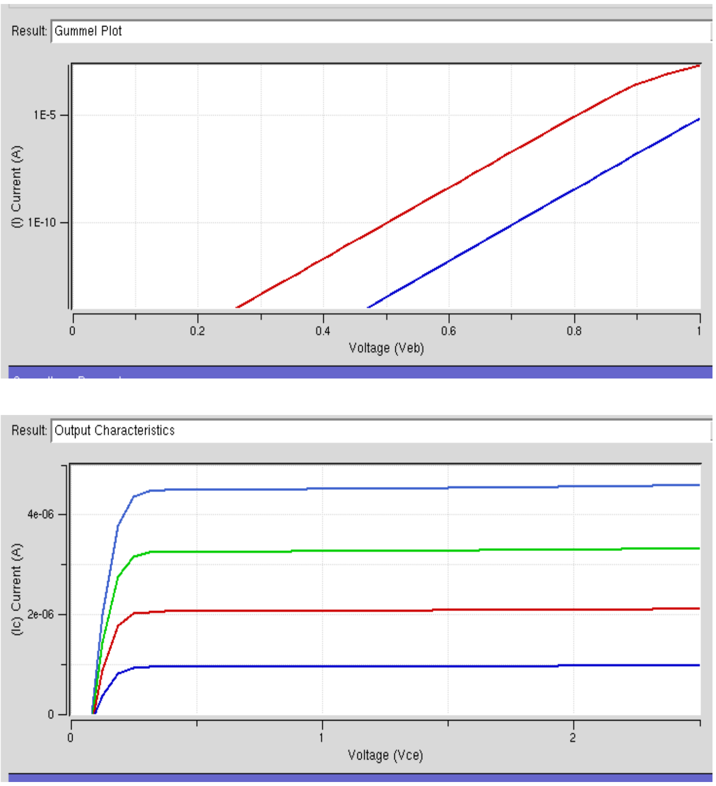
Result: Output Characteristics

Figure 9: Top panel – Gummel plot of transistor #1. Bottom panel – Output characteristics of transistor #1.



Result: Output Characteristics

Figure 10: Output characteristics of transistor #2.


Result: Gummel Plot

Result: Output Characteristics

Figure 11: Top panel – Gummel plot of transistor #3. Bottom panel – Output characteristics of transistor #3.

2.2.2 Early voltage

The Early effect is the variation in the width of the base in a BJT due to a variation in the applied base-to-collector voltage. A greater reverse bias across the collector–base junction, for example, increases the collector–base depletion width, decreasing the width of the charge neutral portion of the base (see Figure 12).

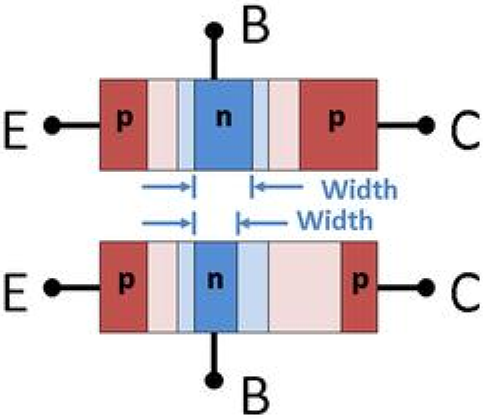


Figure 12: Top: pnp base width for low collector–base reverse bias; bottom: narrower pnp base width for large collector–base reverse bias. Light colors are depleted regions.

In Figure 12, the neutral base width is dark blue, and the depleted base regions are light blue. The neutral emitter and collector regions are dark red, and the depleted regions are pink. Under increased collector–base reverse bias, the lower panel of Figure 12 shows a widening of the depletion region in the base and the associated narrowing of the neutral base region. The collector depletion region also increases under reverse bias more than does that of the base because the collector is less heavily doped. The principle governing these two widths is charge neutrality. The emitter–base junction is unchanged because the emitter–base voltage is the same. The Early effect leads to output characteristics shown in Figure 13.

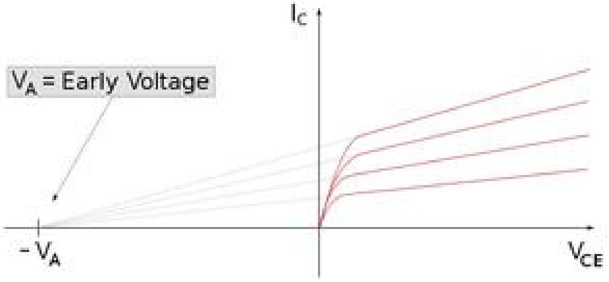


Figure 13: The Early voltage as seen in the output-characteristic plot of a BJT.

Base-narrowing has two consequences that affect the current: (1) There is a reduced chance for recombination within the smaller base region and (2) the charge gradient is increased across the base, and consequently, the current of minority carriers injected across the emitter junction increases.

Both of these factors increase the collector or output current of the transistor with an increase in the collector voltage. This increased current is shown in Figure 13. Tangents to the characteristics at large voltages extrapolate backward to intercept the voltage axis at a voltage called the Early voltage, often denoted by the symbol V_A.

Following the procedure illustrated in Figure 13, we find that for BJT #1 the early voltage is -5 V, for BJT #2 is -0.5 V which is very bad, and for BJT #3 is > 95 V. It is evident that BJT #3 performs the best in terms of the Early voltage and the output characteristics even though its current amplification factor needs further improvement. Typical values for the current amplification in the common-emitter configuration are between 100 and 200.

2.3 Examples for MOS Capacitors

The MOS capacitor consists of a metal-oxide-semiconductor structure as illustrated in Figure 14. The semiconductor substrate with a thin oxide layer and a top metal contact, referred to as the gate, is shown in Figure 14, as well. A second metal layer forms an Ohmic contact to the back of the semiconductor and is called the bulk contact. The structure shown has a p-type substrate. We will refer to this as an n-type MOS or nMOS capacitor since the inversion layer contains electrons.

To understand the different bias modes of n-MOS capacitor, we now consider three different bias voltages: one below the flatband voltage, V_{FB}; a second between the flatband voltage and the threshold voltage, V_T; and finally one larger than the threshold voltage. These bias regimes are called the accumulation, depletion, and inversion mode of operation. These three modes, as well as the charge distributions associated with each of them, are shown in Figure 15.

Accumulation typically occurs for negative voltages where the negative charge on the gate attracts holes from the substrate to the oxide-semiconductor interface. Depletion occurs for positive voltages. The positive charge on the gate pushes the mobile holes into the substrate. Therefore, the semiconductor is depleted of mobile carriers at the interface, and a negative charge resulting from the ionized acceptor ions is left in the space charge region. The voltage separating the accumulation and depletion regime is referred to as the flatband voltage, V_{FB}. Inversion occurs at voltages beyond the threshold voltage. In inversion, there exists a negatively charged inversion layer at the oxide-semiconductor interface in addition to the depletion-layer. This inversion layer is due to the minority carriers that are attracted to the interface by the positive gate voltage.

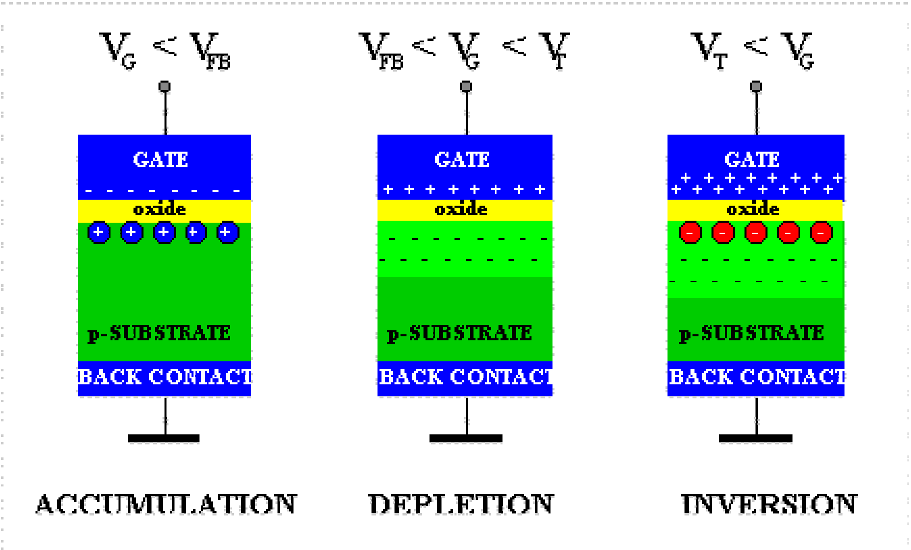
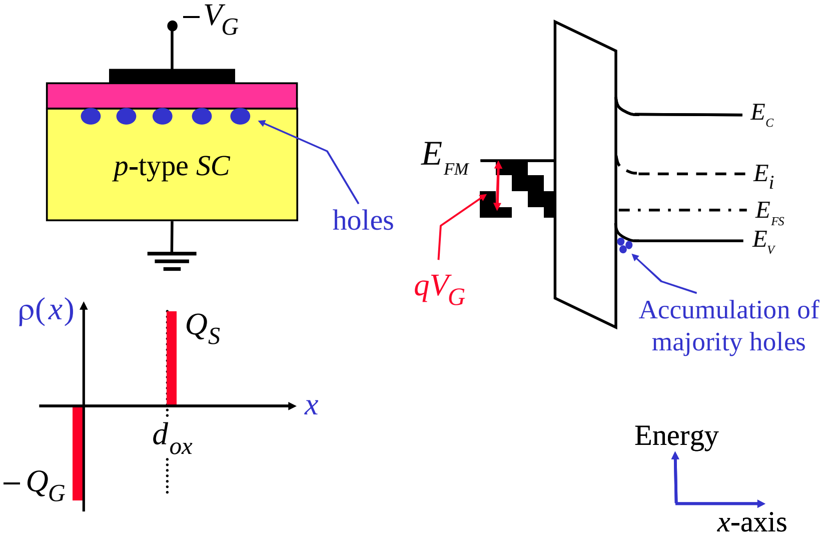
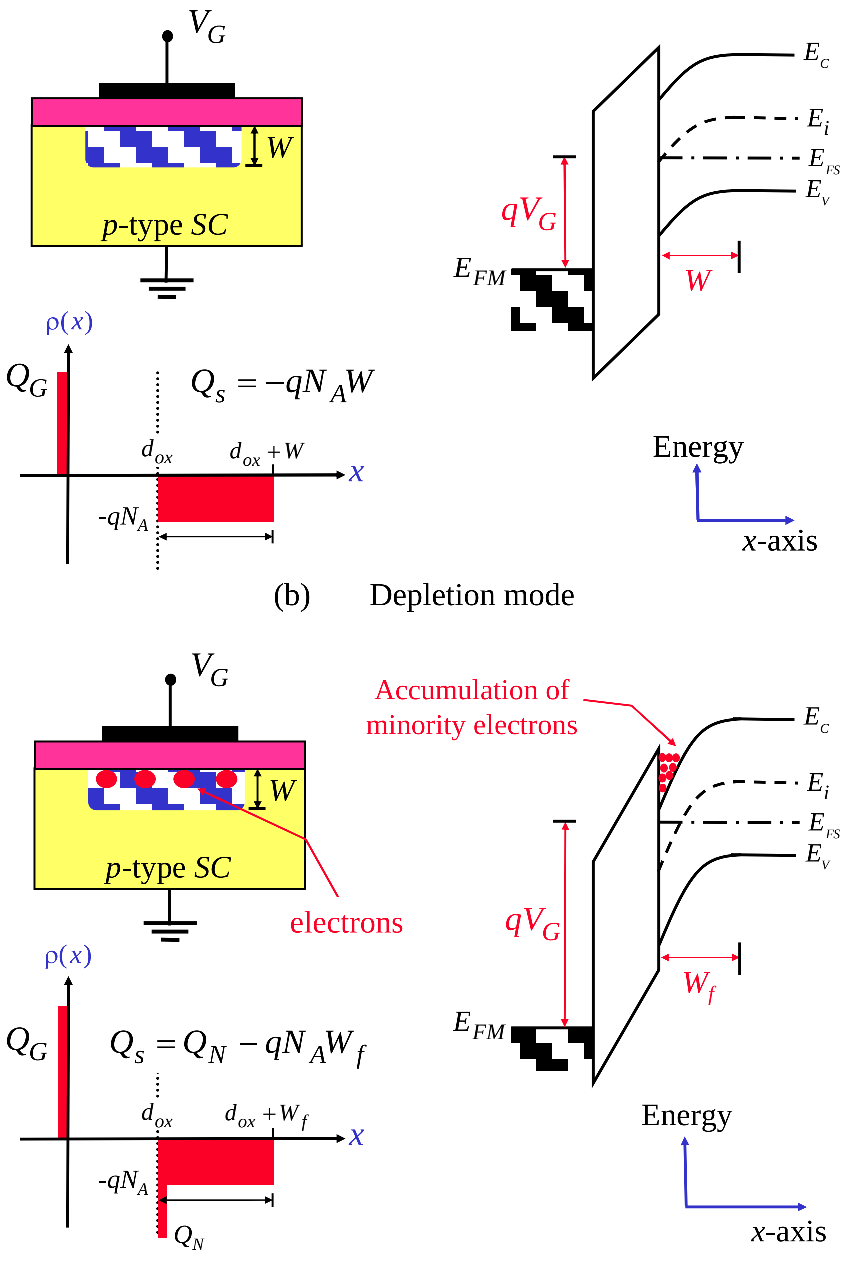


Figure 14: Charges in an n-type metal-oxide-semiconductor structure (p-type substrate) under accumulation, depletion, and inversion conditions.

The energy band diagram of an n-MOS capacitor biased in inversion is shown in Figure 15 (c). The oxide is modeled as a semiconductor with a very large bandgap and blocks any flow of carriers between the semiconductor and the gate metal. The band bending in the semiconductor is consistent with the presence of a depletion layer. At the semiconductor-oxide interface, the Fermi energy is close to the conduction band edge as expected when a high density of electrons is present. The semiconductor remains in thermal equilibrium even while a voltage is applied to the gate. The presence of an electric field does not automatically lead to a non-equilibrium condition, as was also the case for a p-n diode with zero bias.



(a) Accumulation mode



(b) Depletion mode

(c) Inversion mode

Figure 15: Energy band diagram of an MOS structure biased in accumulation (a), depletion (b), and inversion (c).

.

The flatband diagram is by far the easiest energy band diagram. The term flatband refers to fact that the energy band diagram of the semiconductor is flat, which implies that no charge exists in the semiconductor. Note that a voltage, V_{FB}, must be applied to obtain this flat band diagram. The flatband voltage is obtained when the applied gate voltage equals the workfunction difference between the gate metal and the semiconductor. If there is a fixed charge in the oxide and/or at the oxide-silicon interface, the expression for the flatband voltage must be modified accordingly.

Accumulation (Figure 15a) occurs when one applies a voltage less than the flatband voltage. The negative charge on the gate attracts holes from the substrate to the oxide-semiconductor interface. Only a small amount of band bending is needed to build up the accumulation charge so that almost all of the potential variation is within the oxide.

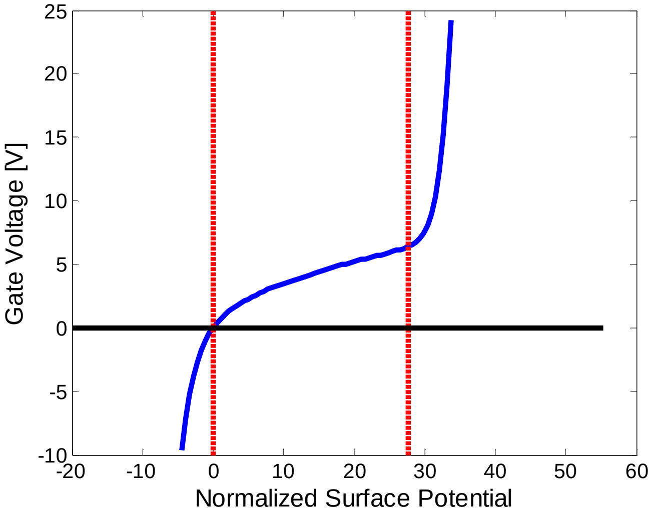
As a more positive voltage than the flatband voltage is applied, a negative charge builds up in the semiconductor. Initially this charge is due to the depletion of the semiconductor starting from the oxide-semiconductor interface (Figure 15b). The depletion layer width further increases with increasing gate voltage.

As the potential across the semiconductor increases beyond twice the bulk potential, another type of negative charge emerges at the oxide-semiconductor interface; this charge is due to minority carriers, which form a so-called inversion layer (Figure 15c). As the gate voltage is increased, the depletion layer width barely increases any further since the charge in the inversion layer increases exponentially with the surface potential.

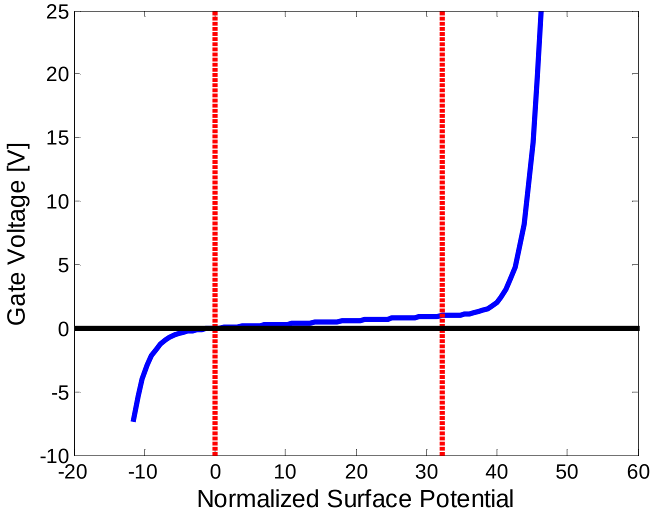
Below we discuss the various modes of operation of a MOS capacitor, the delta-depletion approximation, and an exact analytical model, with emphasis on modeling of MOS capacitors with PADRE simulation software that is used in MOSCap Lab.

2.3.1 Delta-depletion approximation vs. exact analytical model

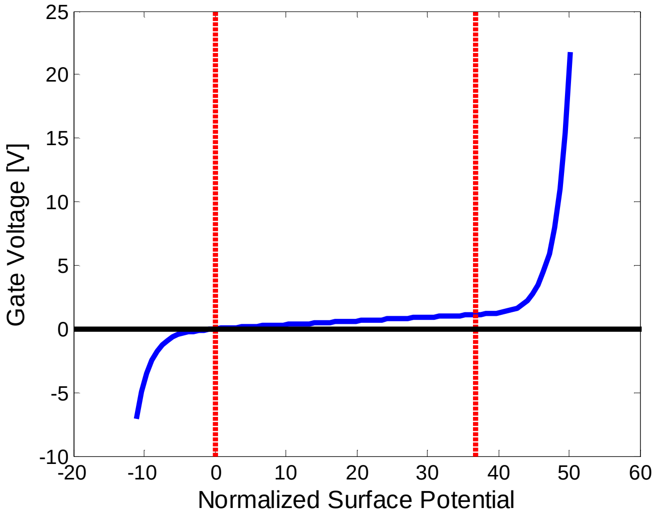
We first want to address when the delta-depletion approximation is valid and when one has to use either the exact analytical or the exact numerical model. In what follows, we will use the exact analytical model results that have been obtained with the script provided at the end of this subsection. To make a comparison between generations of n-channel of devices, we consider two devices at the end of the spectrum and one in the middle. Device #1, which corresponds to old technology has $N_A=10^{16}$ cm^{-3} and T_{ox}=400 um. Device #2 is current technology with $N_A=10^{17}$ cm^{-3} and T_{ox}=2.5 nm. Device #3 is future technology device with $N_A=10^{18}$ cm^{-3} and T_{ox}=1 nm. The results of these simulations are shown in Figure 16. The results presented clearly show that for older technology devices the delta-depletion approximation is rather accurate and the exact analytical model is not necessarily needed. The situation changes for future technology devices for which the differences between the delta-depletion approximation and the exact analytical model are significantly larger in both accumulation and inversion.



(a) Gate voltage vs. surface potential plot for MOS capacitor #1.



(b) Gate voltage vs. surface potential plot for MOS capacitor #2.



(c) Gate voltage vs. surface potential plot for MOS capacitor #3.

Figure 16: Surface potential plots for three different generations of MOS Capacitors.

Listing of the MATLAB code implemented in the MOSCap Lab:

```
T=300
NA=1E24
ND=0
Ni=1E16
dox = 0.001e-6
es = 11.8
eo = 3.9
Vg_min = -10
Vg_max = 30
```

```
kb=1.38E-23
q=1.602e-19
eps0=8.85E-12
eps_sc=es*eps0
eps_ox=eo*eps0
VT=kb*T/q

const = 2*q*Ni*VT/eps_sc
if NA - ND > 0
   fim = -VT*log((NA-ND)/Ni)
else
   fim = VT*log((ND-NA)/Ni)
end

fis1= -4*fim
fis2= 4*fim

dfis= (fis2-fis1)/200
for i=1:200
    fis = fis1 + i*dfis;
    fi(i) = fis;
    term = exp(fim/VT)-exp(fis/VT)+exp(-fim/VT)-exp(-fis/VT)+NA*(fim-fis)/Ni/VT;
    if fis - fim > 0
       field_sc(i) = sqrt(-const*term);
    else
       field_sc(i) = -sqrt(-const*term);
    end
    Vg(i)=eps_sc/eps_ox*dox*field_sc(i) + fis-fim;
end

ii=0
for i=1:200
    if Vg(i) >= Vg_min
        if Vg(i) < Vg_max
            ii=ii+1;
            Vg1(ii)=Vg(i);
            field1(ii)=field_sc(i);
            fi1(ii) = fi(i);
        end
    end
end

figure(1);
plot(fi1/VT-fim/VT,Vg1)
hold on;
surface_pot = -2*fim/VT
a(1) = surface_pot;
b(1) = Vg_min;
a(2) = surface_pot;
b(2) = Vg_max;
c(1) = 0
d(1) = Vg_min
c(2) = 0
d(2) = Vg_max
e(1) = fis1/VT
f(1) = 0
e(2) = fis2/VT
f(2) = 0

plot(a,b)
plot(c,d)
plot(e,f)
```

```
hold off;

clear all;
```

2.3.2 Potential profile, charge distribution, electric-field profile, and low-frequency and high-frequency CV curves of n-channel MOS capacitors

Next, using the MOSCap tool we examine the potential profile, charge distribution, the electric field profile, and the low- and high-frequency CV curves of the MOS capacitor #3 discussed in Section 2.3.1 above. There are several things that we want to point out with the results presented in this section:

1. For the case when there is no charge in the oxide, the potential varies linearly in the oxide region, which means that the electric field, which is a derivative of the potential, is constant (see Figure 17a and b).

2. The electric fields in the oxide versus the semiconductor side of the interface vary as a ratio of the semiconductor versus oxide dielectric constant because if there are no charges at the interface, the perpendicular component of the displacement vector has to be continuous (see Figure 17b).

3. The high-frequency CV curves are obtained when the AC frequency is in the range of MHz. To get the low-frequency CV curves, the AC frequency has to be in few Hz range (see Figure 18).

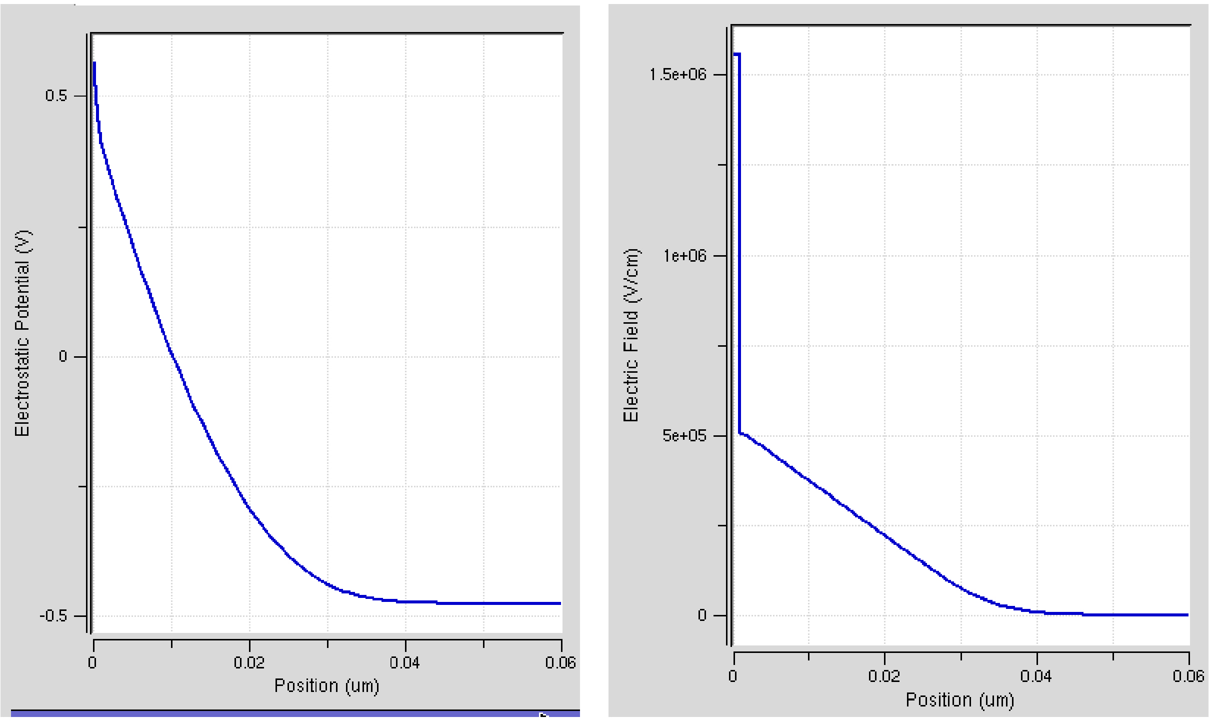
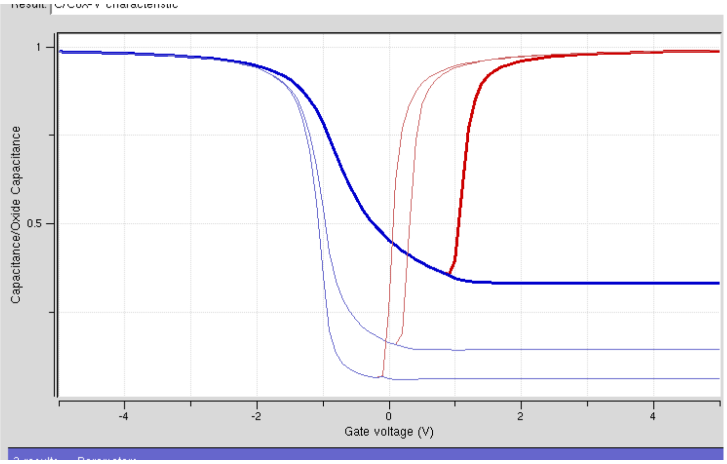
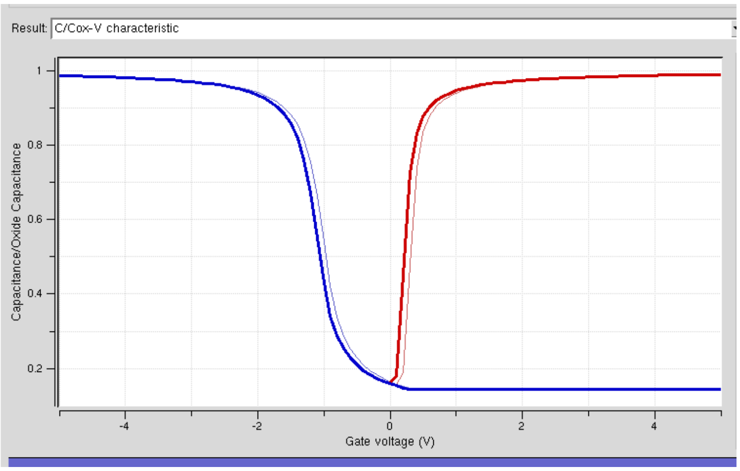


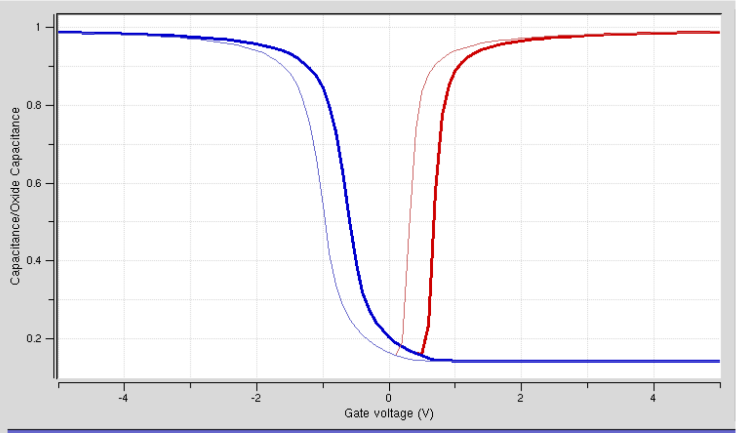
Figure 17: left panel: Electrostatic Potential. b: right panel: Electric field profile.



(a) Low-frequency and high-frequency CV curves for MOS capacitor with T_{ox}=2 nm and N_A=10^{17}, 10^{18}, and 10^{19} cm^{-3}. Note that in obtaining smooth CV-curves, very fine mesh and unrealistically small minority carrier lifetimes are needed. Here, low-frequency is 0.1Hz, and high-frequency is 10 MHz. Notice the increase in the high-frequency curve with the increase in doping.



(b) Low-frequency and high-frequency CV curves for MOS capacitor with T_{ox}=2 nm and N_A=10^{18} cm^{-3}. Notice the small shift in the CV-curves toward negative bias due to the introduction of uniform charge in the oxide with charge density 10^{19} cm^{-3}, which corresponds to sheet charge density of $2\times10^{-7}\times10^{19}$ cm^{-2} = 2×10^{12} cm^{-2}.



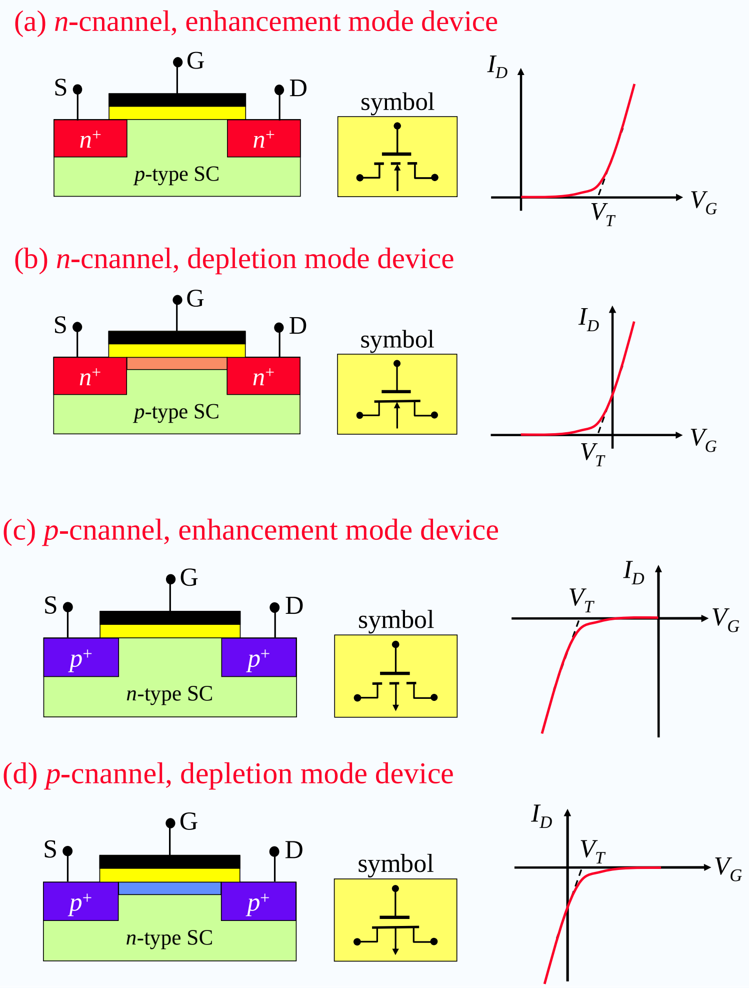
(c) Low-frequency and high-frequency CV curves for MOS capacitor with T_{ox}=2 nm and N_A=10^{18} cm^{-3}. Notice the much larger shift in the CV-curves toward negative bias due to the introduction of charge in the semiconductor/oxide interface with sheet charge density 2×10^{12} cm^{-2}. Notice the much larger shift in the CV curves when same amount of sheet charge density is placed at the semiconductor/oxide interface.

Figure 18: Low frequency and high frequency CV curves of Ideal and Non-Ideal MOS Capacitors.

2.4 Examples for MOSFETs

Another type of transistor, the field effect transistor, operates on the principle that semiconductor conductivity can be increased or decreased by the presence of an electric field. An electric field can increase the number of free electrons and holes in a semiconductor, thereby changing its conductivity. The field may be applied by a reverse-biased pn junction, forming a junction field-effect transistor, or JFET, or by an electrode isolated from the bulk material by an oxide layer, forming a metal-oxide-semiconductor field effect transistor, or MOSFET.

The MOSFET is the most used semiconductor device today. The gate electrode is charged to produce an electric field that controls the conductivity of a channel between two terminals, called the source and drain. Depending on the type of carrier in the channel, the device may be an n-channel (for electrons) or a p-channel (for holes) MOSFET. Although the MOSFET is named in part for its metal gate, polysilicon is typically used in modern devices, instead. The configuration, symbol, and transfer characteristics for n-channel and p-channel enhancement or depletion mode devices are shown in Figure 19.



(a) *n*-cnannel, enhancement mode device

(b) *n*-cnannel, depletion mode device

(c) *p*-cnannel, enhancement mode device

(d) *p*-cnannel, depletion mode device

Figure 19: Configuration, symbols, and transfer characteristics of the four different types of basic MOSFET configurations.

To better understand the operation of a MOSFET device, it is important to understand the role of the gate and the drain electrode individually on the energy band diagram of a MOSFET. We begin with MOSFET in equilibrium, and we want to first understand the influence of the gate electrode (application of positive gate bias) for the case of an n-channel MOSFET. Positive gate voltage on the gate electrode does two things:

(1) Reduces the potential energy barrier seen by the electrons from the source and the drain regions.

(2) Inverts the surface, and increases the conductivity of the channel.

This is schematically shown in Figure 20. Note that we have a flux of carriers: one flux of carriers goes from the source to the drain and is balanced by the flux of carriers that go from the drain to the source, so the net current in equilibrium, as should be expected, is zero.

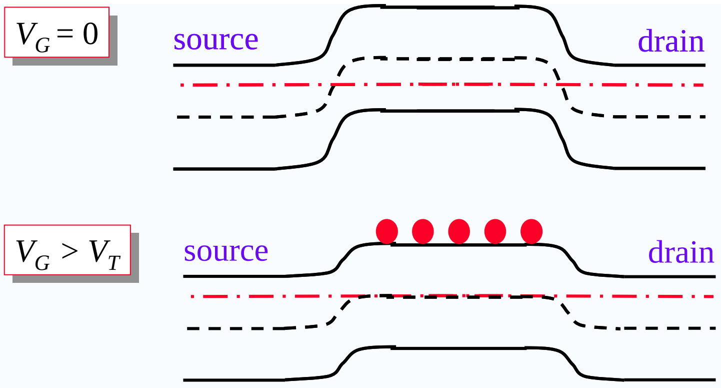


Figure 20: The role of the gate electrode for n-channel MOSFET in equilibrium.

The role of the drain electrode for n-channel MOSFET operation for the two gate bias conditions from Figure 20 is shown in Figure 21.

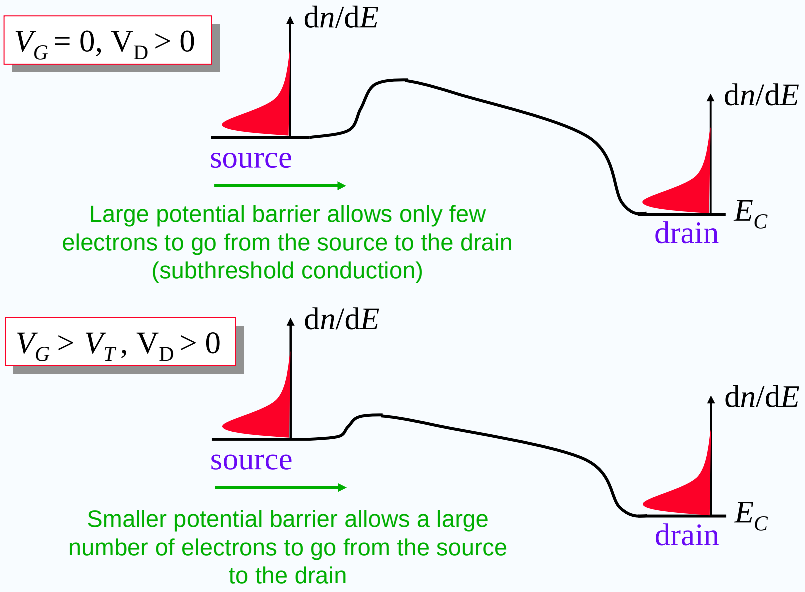
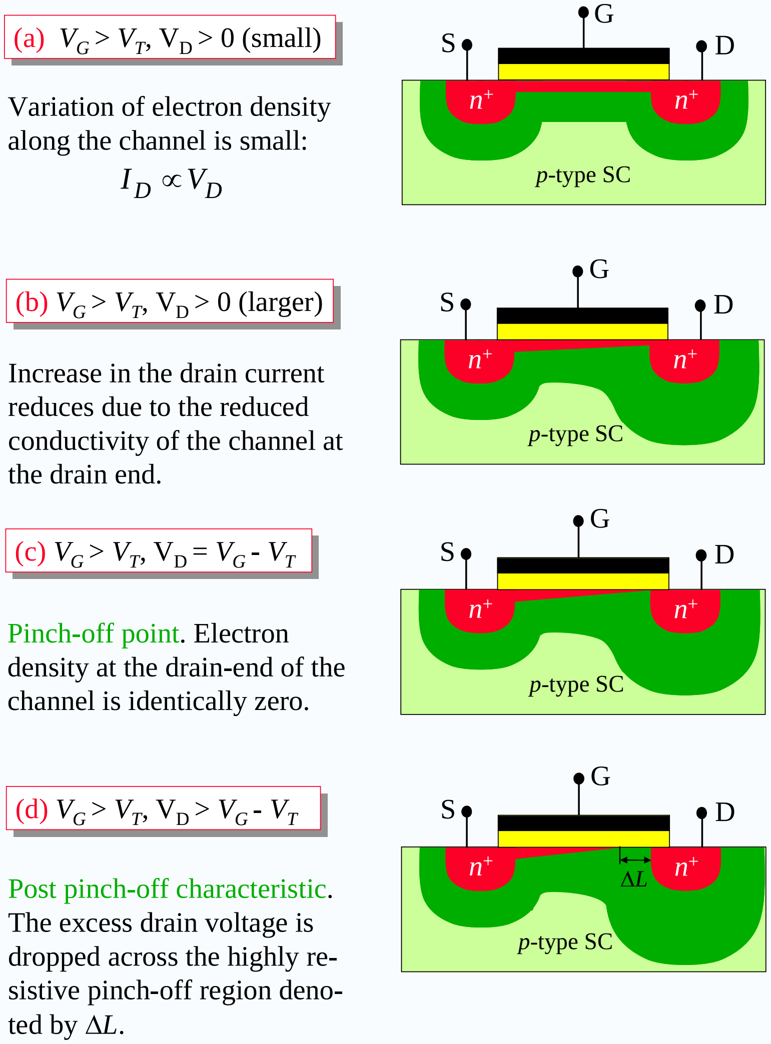

$V_G = 0, V_D > 0$
dn/dE
source
dn/dE
E_C
drain
Large potential barrier allows only few electrons to go from the source to the drain (subthreshold conduction)
$V_G > V_T, V_D > 0$
dn/dE
source
dn/dE
E_C
drain
Smaller potential barrier allows a large number of electrons to go from the source to the drain

Figure 21: The role of the drain electrode.

As can be seen from the schematic diagrams presented in Figure 21, the balance of the source-drain and drain-source fluxes is destroyed with the application of a drain bias. Namely, under positive drain bias, almost all carriers in the drain see a barrier and cannot move toward the source region, i.e. they get reflected back. Thus, the current in this case is approximately equal to the flux of carriers that go from the source to the drain. For V_G=0, we have a small portion of carriers that can overcome the source barrier which, in turn, gives rise to small drain current and subthreshold mode of operation of the MOSFET (top panel in Figure 20). When V_G>V_T, the barrier in the source region is significantly reduced, so much more carriers can go from the source to the drain, and the MOSFET is said to be operating in an on-state (linear or saturation, as we will see later in the text).

Yet another way of describing qualitatively the description of the MOSFET operation is schematically shown in Figure 22. In this set of graphs, we concentrate on the shape of the channel and the depletion regions. We see that for V_G>V_T the channel forms and for small applied drain bias V_D the channel is uniform (top panel in Figure 22). If we increase the drain bias (second panel of Figure 22), two things happen: (1) the drain current increases due to the increase of the lateral electric field that accelerates the carriers from the source to the drain region, and (2) the channel is non-uniform, which means that the larger drain bias leads to smaller gate to channel voltage at the drain side of the channel, which in turn leads to smaller number of inversion electrons. The current is constant along the channel because the velocity of the carriers at the drain end of the channel is much larger when compared to the velocity of the carriers near the source end of the channel; thus the product *nev* is constant as the current is mostly drift current, and the contribution of the diffusion component of the current is insignificant.



(a) $V_G > V_T$, $\mathrm{V_D} > 0$ (small)

Variation of electron density along the channel is small:

$$I_D \propto V_D$$

(b) $V_G > V_T$, $\mathrm{V_D} > 0$ (larger)

Increase in the drain current reduces due to the reduced conductivity of the channel at the drain end.

(c) $V_G > V_T$, $\mathrm{V_D} = V_G - V_T$

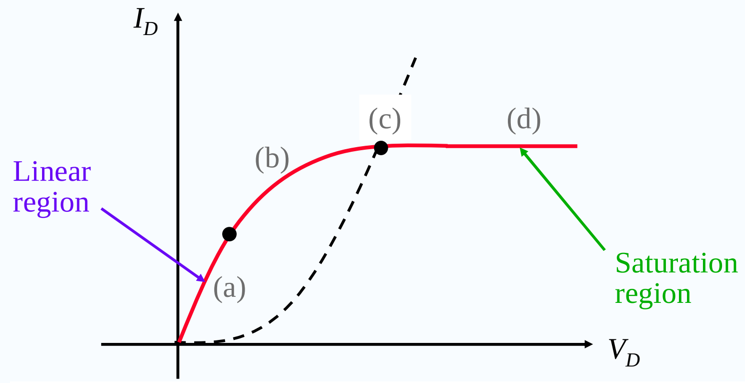
Pinch-off point. Electron density at the drain-end of the channel is identically zero.

(d) $V_G > V_T$, $\mathrm{V_D} > V_G - V_T$

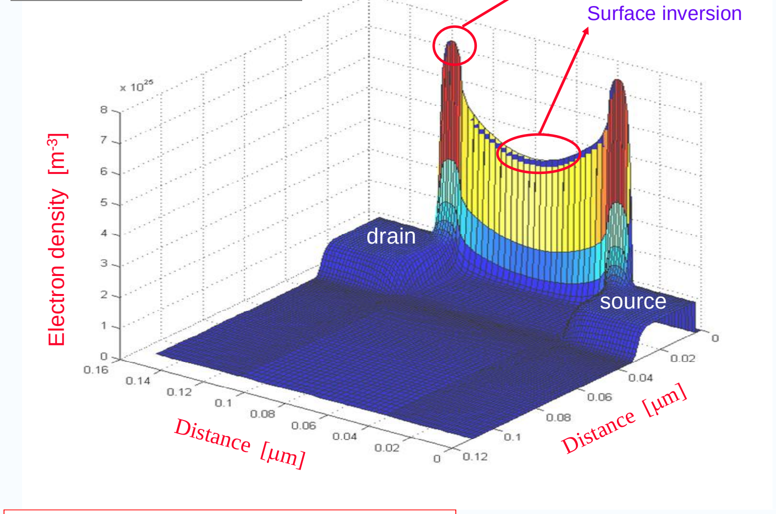
Post pinch-off characteristic. The excess drain voltage is dropped across the highly resistive pinch-off region denoted by ΔL.

Figure 22: Channel and depletion region formation under different biasing conditions ranging from subthreshold to postpinch-off.

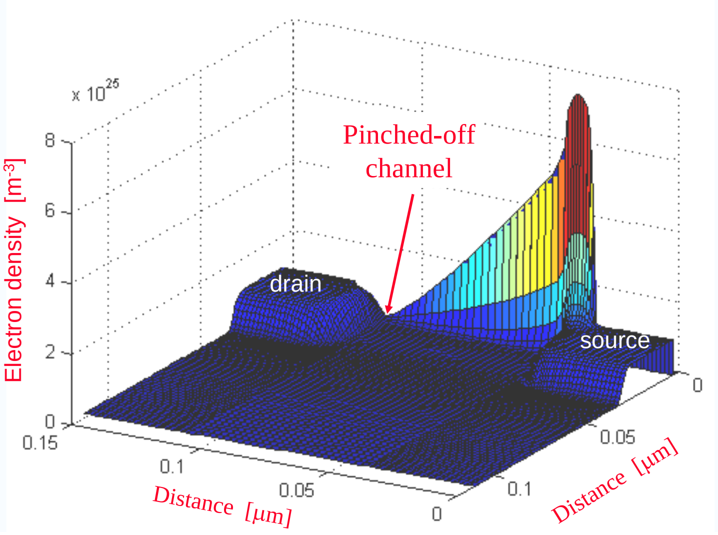
The *IV*-characteristic, shown in Figure 23 explains the various bias conditions given in Figure 22.



$V_G = 0.8\,\mathrm{V},\ V_D = 20\,\mathrm{mV}$



$V_G = 0.8\,\mathrm{V},\ V_D = 0.9\,\mathrm{V},\ V_T = 0.33\,\mathrm{V}$



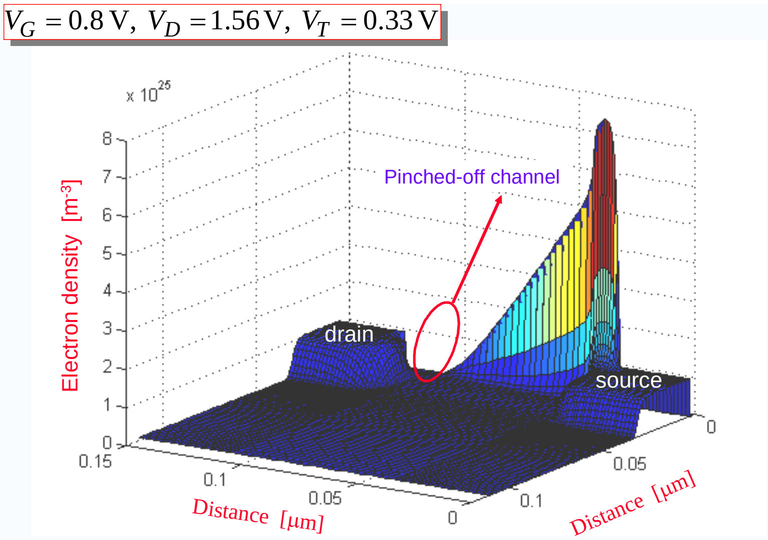


Figure 23: Various modes of operation of a MOSFET device.

Using simple capacitor model, the threshold voltage of a MOSFET device is calculate to be

$$V_T = 2\varphi_F + \frac{1}{C_{ox}}\sqrt{2qN_A k_s \varepsilon_0 \left(2\varphi_F\right)} + V_{FB} \ ,$$

where $\varphi_F = \frac{1}{q}\left|E_F - E_i\right|$, NA is the channel doping, Cox is the oxide capacitance, and the flat-band voltage is given by

$$V_{FB} = \frac{1}{q}\phi_{MS} + \frac{Q_{it}}{C_{ox}} + \frac{Q_f}{C_{ox}} + \gamma_{ot}\frac{Q_{ot}}{C_{ox}} + \gamma_m \frac{Q_m}{C_{ox}} \, ,$$

where ϕ_{MS} is the metal-semiconductor workfunction, Q_{it} is the interface-trap density, Q_f is the fixed oxide charge density, Q_{ox} is the oxide charge, and Q_m is the mobile charge density. In general, the oxide charges are minimized in the device fabrication process.

Analytical or semi-analytical MOSFET models are usually based on the so-called gradual channel approximation (GCA). Contrary to the situation in the ideal two-terminal MOS device, where the charge density profile is determined from a one-dimensional Poisson's equation, the MOSFET generally poses a two-dimensional electrostatic problem. The reason is that the geometric effects and the application of a drain-source bias create a lateral electric field component in the channel, perpendicular to the vertical field associated with the ideal gate structure. The GCA states that, under certain conditions, the electrostatic problem of the gate region can be expressed in terms of two coupled one-dimensional equations: a Poisson's equation for determining the vertical charge density profile under the gate and a charge transport equation for the channel. This allows us to determine self-consistently both the channel potential and the charge profile at any position along the gate. A direct inspection of the two-dimensional Poisson's equation for the channel region shows that the GCA is valid if we can assume that the electric field gradient in the lateral direction of the channel is much less than that in the vertical direction perpendicular to the channel (nanoscale devices). Typically, we find that the GCA is valid for long-channel MOSFETs, where the ratio between the gate length and the vertical distance of the space charge region from the gate electrode, the so-called aspect ratio, is large. However, if the MOSFET is biased in saturation, the GCA always becomes invalid near drain as a result of the large lateral field gradient that develops in this region.

According to the simple charge controlled model, in which the gate charge is completely balanced with the inversion charge, one has the following expressions for the drain current in the linear ($V_D<V_G$-V_T) and saturation region ($V_D>V_G$-V_T) the square law theory gives

$$I_D = \frac{W\mu_{eff}C_{ox}}{L}\left[(V_G - V_T)V_D - \tfrac{1}{2}V_D^2\right], \quad \text{for } V_D \le V_G - V_T$$

$$I_D = \frac{W\mu_{eff}C_{ox}}{2L}(V_G - V_T)^2, \quad \text{for } \; V_D \ge V_G - V_T\,.$$

In the case when the gate charge is balanced with the inversion charge + depletion charge (bulk charge theory), the following results can be derived using SCCM:

$$I_D = \frac{W\mu_{eff}C_{ox}}{L}\left\{\left[(V_G - V_T)V_D - \tfrac{1}{2}V_D^2\right] - \tfrac{4}{3}V_W\varphi_F\left[\left(1+\frac{V_D}{2\varphi_F}\right)^{3/2} - \left(1+\frac{3V_D}{4\varphi_F}\right)\right]\right\}$$

$$V_W = \frac{\sqrt{2k_s\varepsilon_0 qN_A(2\varphi_F)}}{C_{ox}} = \frac{W_T qN_A}{C_{ox}}$$

$$V_{Dsat} = V_G - V_T - V_W\left\{\sqrt{\frac{V_G - V_T}{2\varphi_F} + \left(1+\frac{V_W}{4\varphi_F}\right)^2} - \left(1+\frac{V_W}{4\varphi_F}\right)\right\}.$$

The results of the square law and bulk charge theory are shown schematically in Figure 24. It is evident that the deviation between the square law and bulk charge theory is larger for larger channel doping. The role of the series source and drain resistance is illustrated in Figure 25.

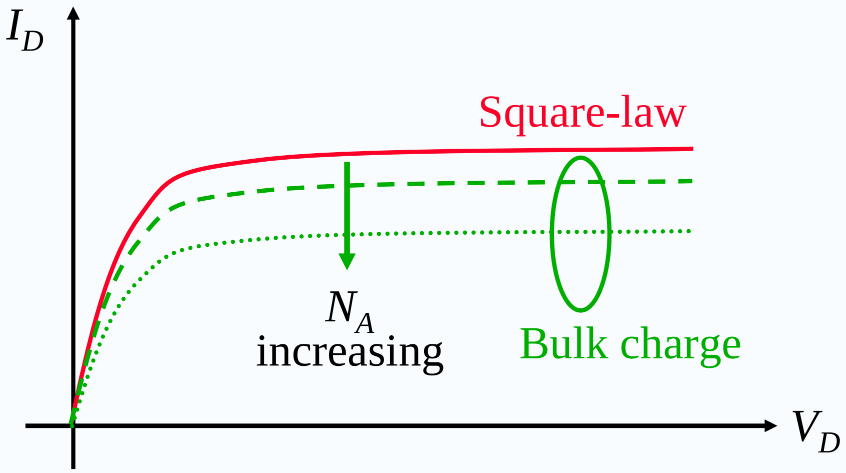
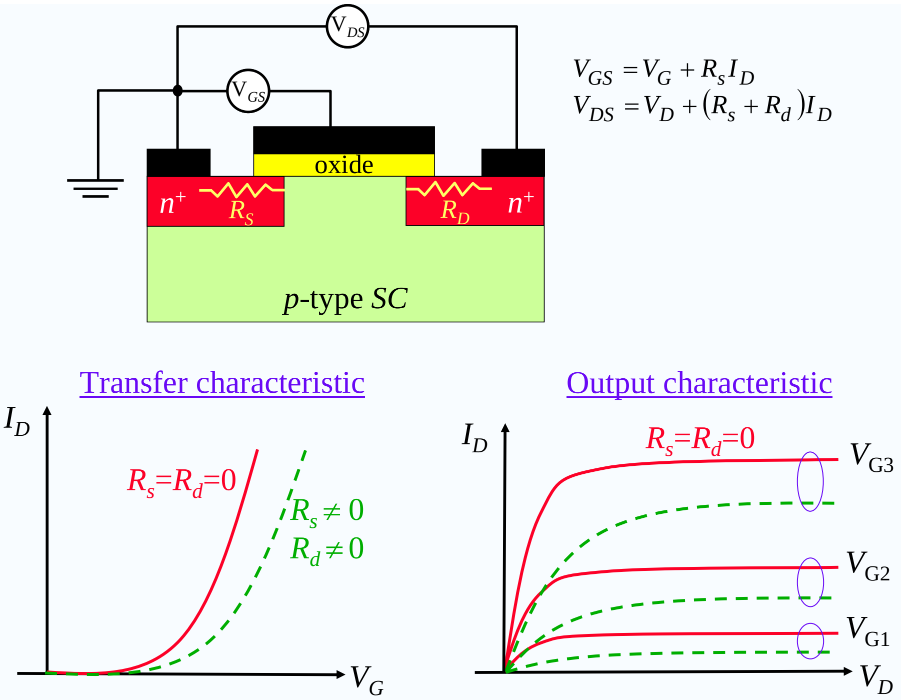


Figure 24: Square-law vs. bulk-charge theory.



$$V_{GS} = V_G + R_s I_D$$
$$V_{DS} = V_D + \left(R_s + R_d\right) I_D$$

Figure 25: Source and drain series resistance effects on transfer and output characteristics.

When the device size is scaled in the nanometer range, first velocity saturation and then velocity overshoot effects start to dominate the device behavior. Velocity saturation arises because the effective carrier mobility is not constant but decreases with increasing the electric field at large in-plane fields. When the device is velocity saturated, then:

$$I_D = -qA_{eff}\, n v_d\,, \quad A_{eff} = Z y_{eff}$$

Device width

Effective thickness of the inversion layer

i.e., the velocity-limited drain current equals to:

$$I_D = -qZy_{eff}\, n v_d = \underbrace{-\,qy_{eff}\, n}_{Q_N} Z v_d = Z v_d C_{ox}\left(V_G - V_T\right)$$

.

2.4.1 Velocity saturation effect for 200 nm channel length technology

For long channel devices, one can apply the gradual channel approximation and consider the MOSFET as a one-dimensional object, but this is not true in nanoscale devices where transport is definitely two-dimensional. In addition to this observation, at small gate length scales, the carrier velocity saturates due to phonon intervalley scattering in Si material system because the gate voltages do not scale proportionally. This behavior is shown in Figure 26.

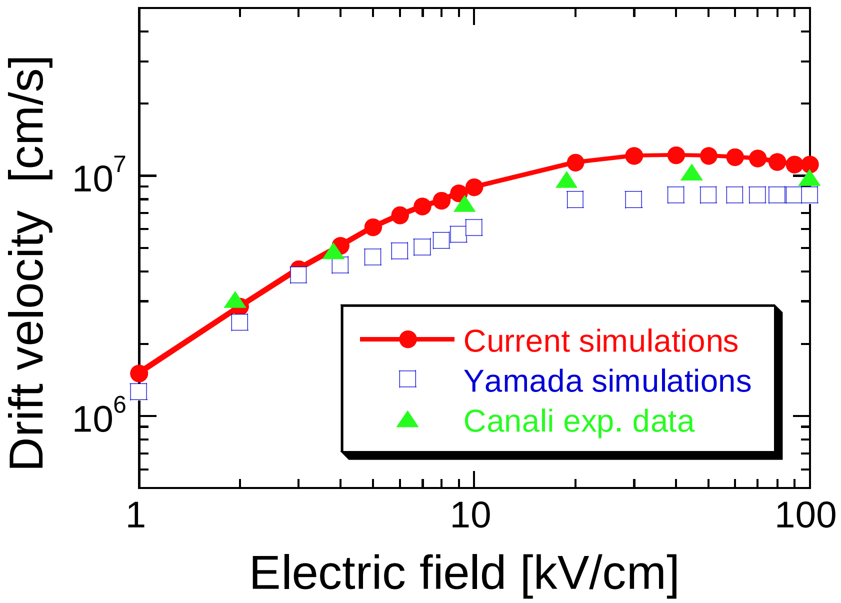


Figure 26: Bulk Monte Carlo simulations that illustrate the velocity saturation effect in silicon. Current simulations refers to the simulations performed by Xiaojiang He at ASU.

Because of the velocity saturation effect, the IV characteristics in saturation regime are no longer proportional to $(V_{GS}\text{-}V_T)^2$, but they become proportional to $(V_{GS}\text{-}V_T)$, which means current no longer increases quadratically with increasing gate voltage but instead has a linear dependence on the gate voltage V_G. The gate-length at which velocity saturation effect starts to become important can be calculated by equating the long-channel and the saturated current characteristics. This is done below, from where it follows that:

$$\begin{aligned} I_{DS} &= \frac{Z\mu_n C_{ox}}{2L}(V_{GS}-V_T)^2 \\ I_{DS} &= Zv_s C_{ox}(V_{GS}-V_T) \end{aligned} \quad \rightarrow L = \frac{\mu_n}{2v_s}(V_{GS}-V_T)\,.$$

Now, if we assume the electron mobility to be 500 cm^2/V-s, V_{GS}-V_T = 5 V, and the saturation velocity to be 10^7 cm/s, we get that the critical length below which velocity saturation effects become important is 1.25 μm. This simplified analysis suggests that if we are going to examine a device with channel length 200 nm, we definitely must include the velocity saturation effect in the model. The simulations for the output characteristics of 200 nm channel length device performed with the MOSFET Lab on the nanoHUB clearly illustrate this behavior (see Figure 27). In this example, we use oxide thickness = 3 nm, channel doping = 10^{17} cm^{-3}, and substrate doping 5×10^{16} cm^{-3}. The gate voltage V_{GS} varies from 1.4 V to 1.8 V in 0.2 V increments. Three curves are sufficient to illustrate the linear dependence of the drain saturation current upon the gate bias.

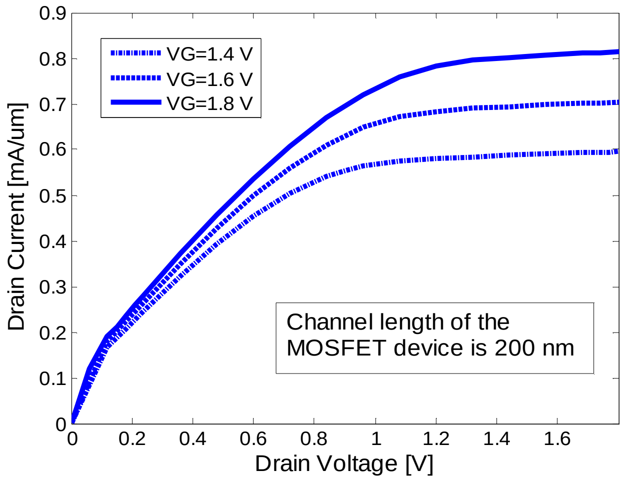


Figure 27: Output characteristics of the 200 nm MOSFET device with parameters described in the text.

2.4.2 Punch-through effect

To have low drain bias and reduce the power-dissipation, the channel doping has to be made as low as possible. This in turn can result in unwanted transistor behavior, namely the transistor might become punch-through. As illustrated schematically in Figure 28, punch-through occurs when the depletion regions of the source-substrate and drain-substrate are connected with each other by high drain voltage.

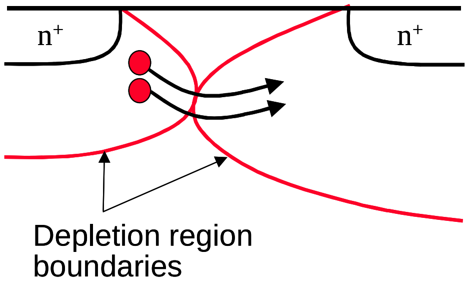


Figure 28: Schematic illustration of the punch-through effect.

As a result of the punch-through effect, the majority electrons in the source are injected into the depletion region, where they are swept by the high electric field. The current flow is deeper in the substrate, and the gate loses the control over the channel. To eliminate this effect, a punch-through stop is used. This is done with deep ion-implantation process, thus leading to very complicated doping profiles in the state-of-the-art devices.

We illustrate the punch-through effect on the example of 200 nm device described in the previous subsection, Velocity saturation effect for 200 nm channel length technology. The only parameter that we vary is the doping of the channel and the doping of the substrate which for simplicity we make them to be equal to each other so that $N_{channel} = N_{substrate} = 5\times10^{14}$ cm^{-3}. The results shown in Figure 29 suggest punch-through starting to manifest itself in the output characteristics for V_G=0.5 V. The results would not even converge for V_G=1 V and V_G=1.5 V.

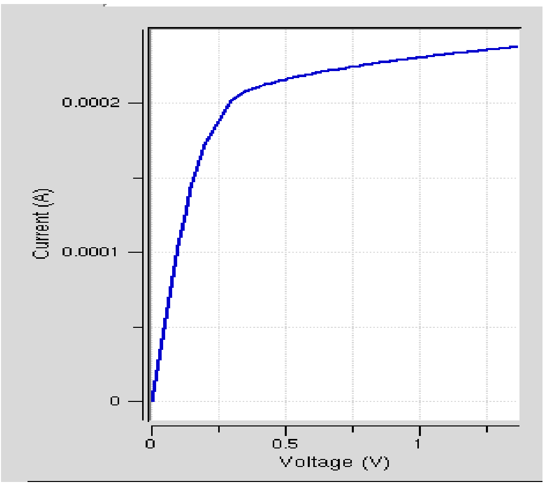


Figure 29: Output characteristics of a MOSFET device that illustrate the punch-through effect.

2.4.3 Drain-induced barrier lowering (DIBL) in nanoscale MOSFETs

The influence of the drain potential on the channel region can have serious impact on the performance of sub-micron MOS transistors. One effect that is very similar to the punch-through effect is drain-induced barrier lowering (DIBL). In the literature, punch-through is sometimes referred to as "subsurface DIBL" in contrast to "surface DIBL," which will be described in this section.

In the weak inversion regime, there is a potential barrier between the source and the channel region. The height of this barrier is a result of the balance between drift and diffusion current between these two regions. If a high drain voltage is applied, the barrier height can decrease (see Figure 30), leading to an increased drain current. Thus the drain current is controlled not only by the gate voltage, but also by the drain voltage. For device modeling purposes, this parasitic effect can be accounted for by a threshold voltage reduction depending on the drain voltage.

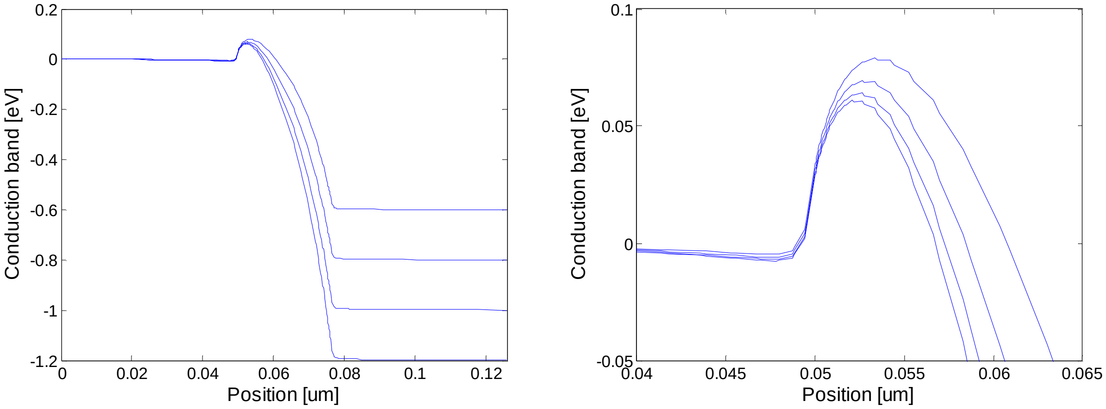
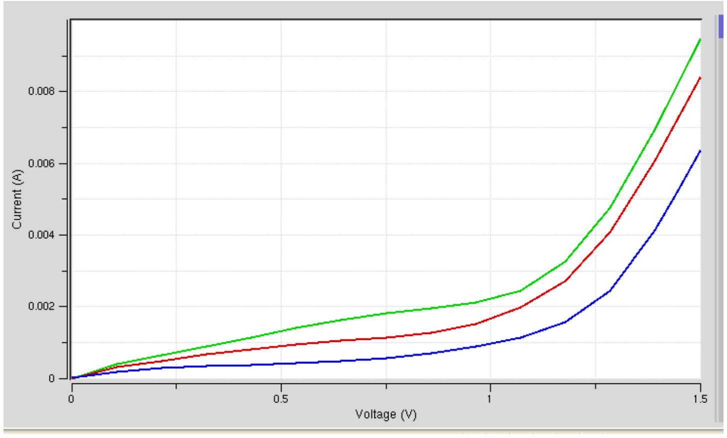


Figure 30: Conduction band profile of a 26 nm channel length device with oxide thickness 1 nm, source-drain doping of 10^{20} cm^{-3}, channel doping of 10^{18} cm^{-3}, and substrate doping of 10^{18} cm^{-3}. The 1D-plots are taken at a depth 10 nm below the semiconductor/oxide interface. Applied gate bias in all these simulations is 1.2 V and the applied drain bias varies between 0.6 V and 1.2 V in 0.2 V increments. The left panel is complete conduction band profile, and the right panel is zoom-in of the barrier at the source end of the channel. These simulations were performed with the MOSFET tool on nanoHUB.org.

SOI Device Structure



2.5 Examples for MESFETs

To understand the operation of a MESFET [7], we consider the section under the gate of Figure 31. The source is grounded, the gate is zero or reverse biased, and the drain is zero or forward biased; that is, $V_G \leq 0$ and $V_D \geq 0$.

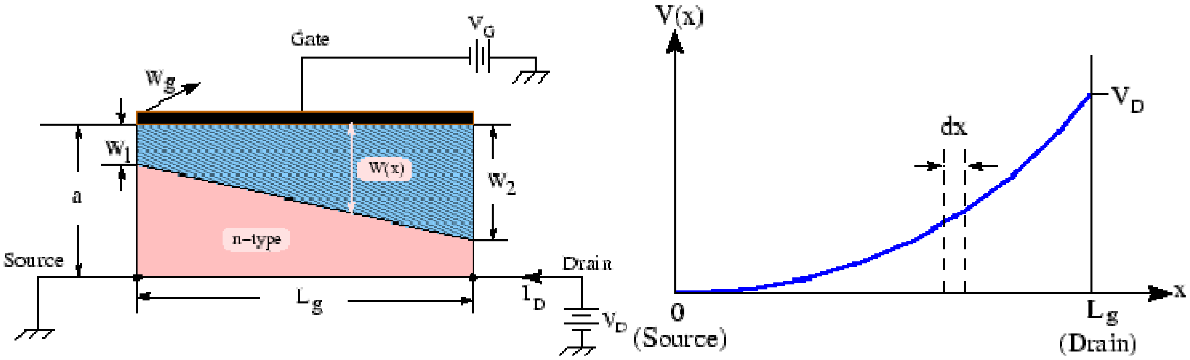


Figure 31: Cross-section of the channel region of a MESFET (left), and drain voltage variation along the channel (right).

The drain voltage along the channel is shown in the right side of Figure 31. The voltage drop across an elemental section *dx* of the channel can be obtained from Figure 31 and is given by

$$dV = \frac{I_D dx}{q\mu_n N_D W_g \left[a - W(x)\right]}$$

and the depletion-layer width at a distance *y* from the source is given by

$$W(x) = \sqrt{\frac{2\varepsilon_s \left[V(x) + V_G + V_{bi} \right]}{qN_D}} \ .$$

Now, the drain current I_D is constant, independent of x. Then one can rewrite

$$I_D dx = q\mu_n N_D W_g \left[a - W(x) \right] dV \ .$$

The differentiation of the drain voltage dV is and integrating from x=0 to x=L_g gives

$$I = I_P \left[\frac{V_D}{V_P} - \frac{2}{3}\left(\frac{V_D + V_G + V_{bi}}{V_P} \right)^{3/2} + \frac{2}{3}\left(\frac{V_G + V_{bi}}{V_P} \right)^{3/2} \right]$$

where

$$I_P = \frac{W_g \mu_n q^2 N_D^2 a^3}{2\varepsilon_s L_g} \text{ and } V_P = \frac{qN_D a^2}{2\varepsilon_s} \quad .$$

The voltage V_P is called the pinch-off voltage, that is the total voltage (V_D+V_G+V_{bi}) at which W_2 = a.

For high-frequency application of MESFETs, an important figure of merit is the cutoff frequency f_T, which is the frequency at which the MESFET can no longer amplify the input signal. It is calculated using

$$f_T = \frac{g_m}{2\pi C_G} < \frac{I_P / V_P}{2\pi W_g L_g \left(\varepsilon_s / \bar{W} \right)} \approx \frac{q\mu_n N_D a^2}{2\pi \varepsilon_s L_g^2} \ .$$

Thus, to improve high-frequency performance, a MESFET with high carrier mobility and short channel length should be used. This is the reason that n-channel SiC MESFET, which has higher electron mobility, is preferred. The derivation of above equation is based on the assumption that the carrier mobility in the channel is a constant, independent of the electric field. However, for very high-frequency operations, the longitudinal field, i.e., the electric field direct from the source to the drain, is sufficiently high that the carriers travel at their saturation velocity. In this case:

$$I_{Dsat} = Aqnv_s = W_g (a - W) qN_D v_s \, .$$

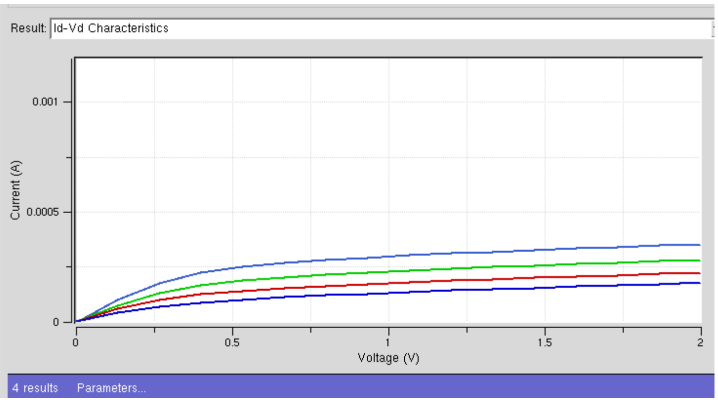
The cutoff frequency under saturation velocity condition is then calculated using

$$f_T = \frac{g_m}{2\pi C_G} = \frac{W_g v_s \varepsilon_s / W}{2\pi W_g L_g \varepsilon_s / W} = \frac{v_s}{2\pi L_g} \, .$$

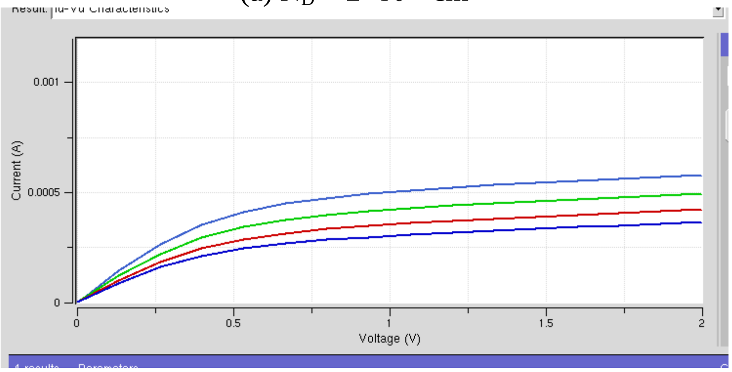
Therefore, to increase f_T, we must reduce the gate length L_g and employ a semiconductor with a high velocity. SiC is superior to other semiconductor materials to operate at higher cutoff frequency due to its higher electron drift velocity.

2.5.1 Output characteristics of a Si MESFET device

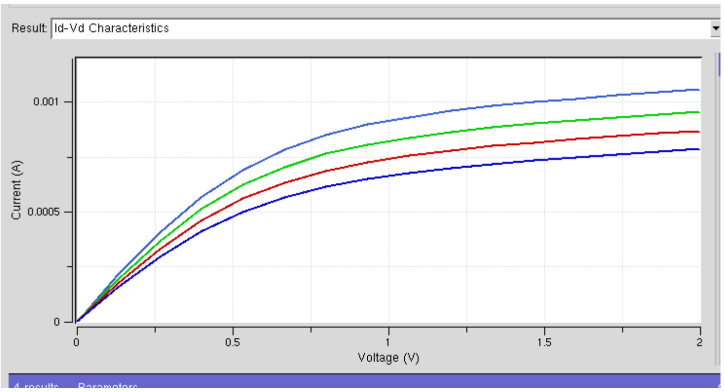
In a typical MOSFET device, there is always a trade-off between the on current and the output conductance and the channel doping. Low channel doping is required to get more inversion layer electrons, but as already noted in Section 2.4.2, that can lead to large output conductance and punch-through effect. To prevent the punch-through and the output conductance effect, larger channel doping densities are typically used. The situation is opposite in MESFET devices. The higher the doping of the channel, the smaller the depletion region width under the Schottky gate and the larger the conductance of the channel, therefore the larger the on current. This behavior is illustrated in Figure 32 where we plot the output characteristics of a 0.3 um channel length MESFET device with source gap of 0.1 um and drain gap of 0.1 um. Parameter in these figures is the substrate doping that equals 2×10^{17} cm^{-3}, 3×10^{17} cm^{-3}, and 5×10^{17} cm^{-3}.



(a) $N_D = 2\times10^{17}$ cm^{-3}



(b) $N_D = 3\times10^{17}$ cm^{-3}



(c) $N_D = 5\times10^{17}$ cm^{-3}

Figure 32: Output characteristics of a MESFET device. Parameter in these figures is the substrate doping.

3. Some Important Things to Remember When Setting Input Parameters and Models

When modeling nanoscale devices, several issues need to be considered. First, the meshing becomes crucial, and this issue is detailed in section 3.1 below. Another point where users of commercial device simulators make errors is the choice of the proper transport model due to the lack of knowledge of physical device behavior. Since this point is very important one, in section 3.2 we first discuss when the drift-diffusion model fails, and in section 3.3, we discuss when the hydrodynamic model fails, in which case one can either use the direct solution of the Boltzmann transport equation if semi-classical transport model is valid. If that is not the case, then tools that implement quantum transport modeling have to be used. The description of the need for quantum transport is discussed in a separate tutorial published on nanoHUB .

3.1 Mesh Size vs. Accuracy

For a stable drift-diffusion/energy balance device simulation, one has to choose the appropriate time step, Δt, and the spatial mesh size (Δx, Δy, and/or Δz). The time step and the mesh size may correlate to each other in connection with the numerical stability. For example, the time step Δt must be related to the plasma frequency. From the viewpoint of the stability criterion, Δt must be much smaller than the inverse plasma frequency. The highest carrier density specified in the device model is used to estimate Δt. The mesh size for the spatial resolution of the potential is dictated by the charge variations. Hence, one has to choose the mesh size to be smaller than the smallest wavelength of the charge variations. The smallest wavelength is approximately equal to the Debye length λ_D. The highest carrier density specified in the model should be used to estimate λ_D from the stability criterion. Critical regions/phenomena where meshing plays crucial role are as follows:

- Proper resolution of the inversion layer charge in a MOSFET or SOI device in either linear or saturation regime of operation.

- Simulation of MOSFET devices with impact ionization model turned on near the breakdown point. Then good meshing is needed near the drain end of the device where the electric fields are the highest.

3.2 When Does the Drift-Diffusion Model Fail?

In principle, the drift-diffusion model is derived under the conditions of low-field transport. The validity and the enormous success that the drift-diffusion model has had in semiconductor device design has been achieved by 'by hands' introducing field-dependent mobilities and diffusion coefficients that allow one to properly capture the velocity saturation effect in devices. This quick fix has worked quite well for many years and has resulted in successful design in new technology nodes, as this model could accurately predict the velocity saturation effect.

In nanoscale devices, the velocity overshoot that results from the non-stationary carrier transport in the device starts to play significant role. Contrary to the velocity saturation effect, this overshoot is a positive effect, as it leads to a situation in which in some or all portions of the channel the velocity of the carriers are larger than the saturation velocity. The so-called velocity overshoot effect can easily be explained by considering a bulk semiconductor to which one suddenly applies a large electric field. This situation is simulated with Arizona State University's bulk Monte Carlo device simulator for the case of both silicon material system and GaAs material system. In silicon material systems, velocity overshoot arises because of the differences in the momentum and the energy relaxation times, whereas in GaAs, it occurs because of the intervalley transfer. In Figure 33 below, we plot the velocity overshoot in silicon calculated using 10,000 particles in the ensemble to reduce the statistical noise.

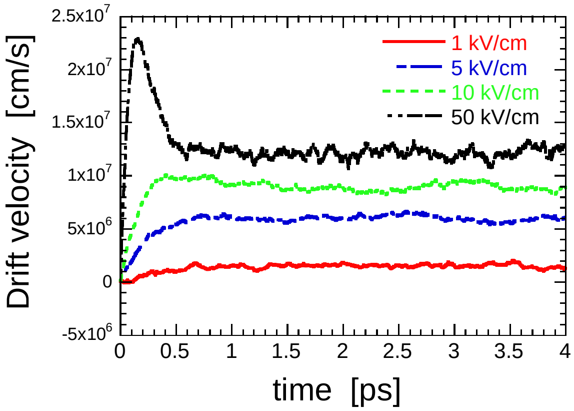


Figure 33: Drift velocity overshoot in silicon.

For the purpose of demonstrating the need for hydrodynamic modeling, three different generations of fully-depleted (FD) silicon on insulator (SOI) devices are being simulated. Since PADRE does not have capability of including the complete hydrodynamic model, the results presented below have been obtained with Silvaco ATLAS simulator discussed in more details in Section 4.1. The characteristic dimensions of the device structures being simulated and schematically shown in Figure 34 are summarized in Table 1. The simulated devices are fully-depleted silicon on insulator (FDSOI) devices.

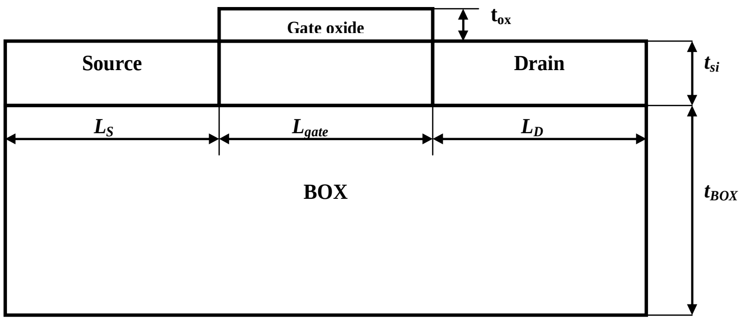


Figure 34: Schematic description of the prototypical FDSOI device structure being simulated.

Table 1 Geometrical dimensions of the simulated fully-depleted SOI nMOSFETs and the applied bias

feature	14 nm	25 nm	90 nm
Tox	1 nm	1.2 nm	1.5 nm
VDD	1V	1.2 V	1.4 V
Overshoot EB/HD	233% / 224%	139% / 126%	31% /21%
Overshoot EB/DD with series resistance	153%/96%	108%/67%	39%/26%

Source/drain doping = 10^{20} cm^{-3} and 10^{19} cm^{-3} (series resistance (SR) case)
Channel doping = 1E18 cm^{-3}
Overshoot= $(ID_{HD}-ID_{DD})/ID_{DD}$ (%) at on-state

We use the Silvaco ATLAS that includes hydrodynamic model with momentum and energy relaxation times of 0.2 ps, Auger generation/recombination, which is important for the proper modeling of the heavily doped source and drain contacts. Schockley-Read-Hall (SRH) generation-recombination mechanism that is not really important for the description of the operation of this device structure, but it is included here for completeness. Impact ionization is not included in these simulations. Since this is a hydrodynamic calculation, it is very important that one uses the NEWTON method for solving the coupled set of equations, otherwise the simulation will not converge. Also note that we consider both the simplified energy balance (EB) model and the complete hydrodynamic model (HD). We present simulation results for the following two cases that we later use to draw conclusions:

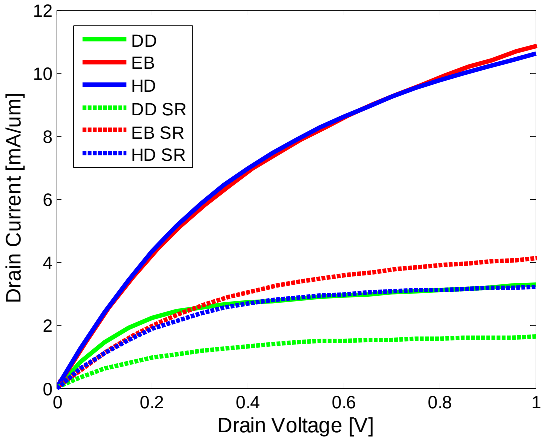
1. Source and drain doping of 10^{20} and 10^{19} cm^{-3} to examine the series resistance effects. This is very important to know as in prototypical Monte Carlo device simulations source and drain regions are usually doped up to 10^{19} cm^{-3} to reduce the computational cost. In these simulations, we assume that the energy relaxation time is 0.2 ps, which is a typical value used for the silicon material system. The results from these simulations are presented in Figure 35 for the 14 nm, 25 nm, and 90 nm channel length device. On the left panel, we show the meshing used in these simulations, and on the right panel, we show the output characteristics for the appropriate on-state gate bias and drift-diffusion and hydrodynamic transport models.

2. In this second case, we perform only hydrodynamic simulations to investigate the sensitivity of the hydrodynamic model to variations in the energy relaxation time which, in principle, is a material and device geometry dependent parameter that makes it almost

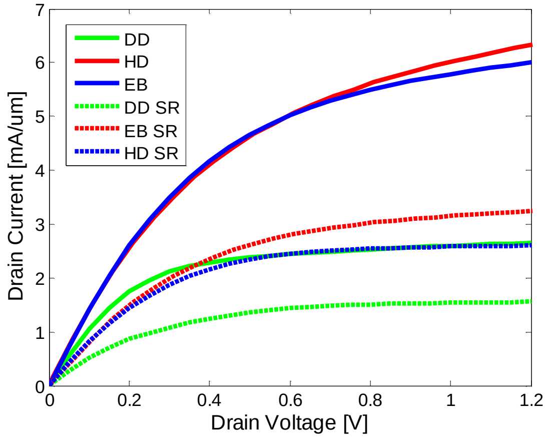
impossible to determine analytically. This variation for the three technology nodes of devices is shown in Figure 36.

From the results presented, it is evident that velocity overshoot plays smaller role in 90 nm gate-length FDSOI devices, whereas the importance of velocity overshoot increases drastically for 14 nm gate length FDSOI device. This result suggests that one must include the energy balance equation if proper modeling of nanoscale devices with gate lengths less than 100 nm is to be achieved.

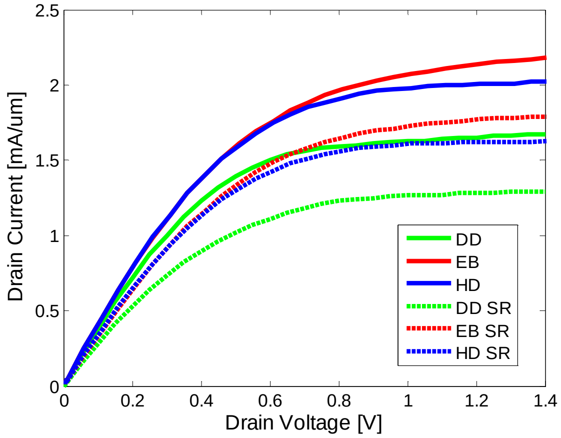
Yet another issue that deserves further attention is the dependence of the simulation results upon the choice of the energy relaxation time. In Figure 36, we plot the output characteristics of 14 nm gate length FDSOI device in which parameter is the energy relaxation time. We see strong dependence of the on-current upon the choice of the energy relaxation time for the smallest structure being simulated, which suggests that proper determination of the energy relaxation time is needed. The energy relaxation time, in turn, is a bias- and geometry-dependent parameter, and its exact determination is impossible. The inability to properly determine the energy relaxation time in hydrodynamic/energy balance models has been the main motivation for the development of particle-based simulators.



(a) channel length = 14 nm. V_G=1 V. SR stands for series resistance.

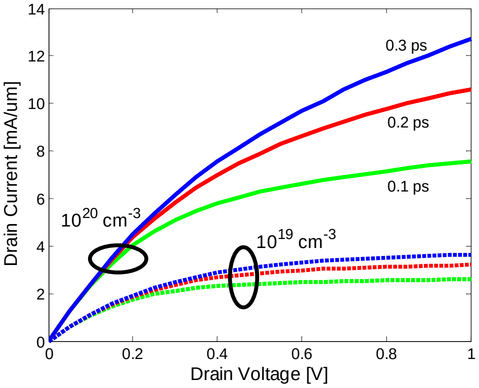


(b) channel length = 25 nm. V_G=1.2 V. SR stands for series resistance.

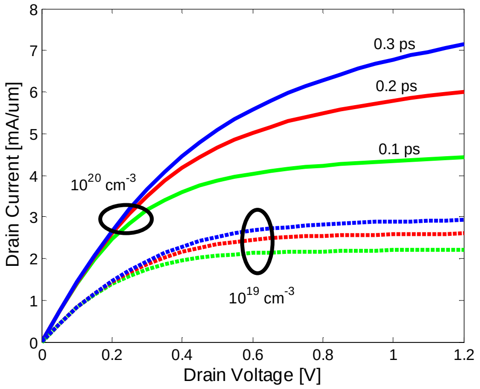


(c) channel length = 90 nm. V_G=1.4 V. SR stands for series resistance.

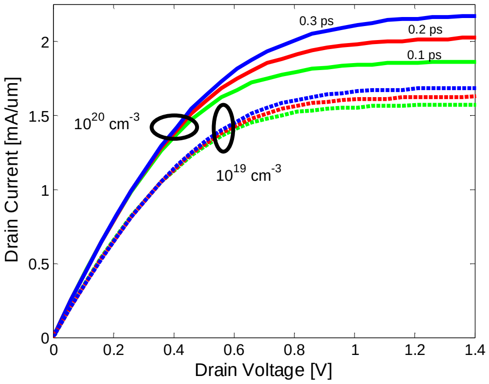
Figure 35: Mesh and output characteristics of 14, 25, and 90 nm channel length FDSOI devices in the on-state when using drift-diffusion and hydrodynamic model.



(a) channel length = 14 nm. V_G=1 V. SR stands for series resistance. Parameters in this figure are the source/drain doping and the energy relaxation time used for the complete hydrodynamic simulation.



(b) channel length = 25 nm. V_G=1.2 V. SR stands for series resistance. Parameters in this figure are the source/drain doping and the energy relaxation time used for the complete hydrodynamic simulation.



(c) channel length = 90 nm. V_G=1.4 V. SR stands for series resistance. Parameters in this figure are the source/drain doping and the energy relaxation time used for the complete hydrodynamic simulation.

Figure 36: Dependence of the on-state current upon the choice of the energy relaxation time for 14 nm channel length FD SOI device.

4. Most Commonly Used Commercial Simulation Tools Capabilities

Silvaco ATLAS

ATLAS is a modular and extensible framework for one-, two-, and three-dimensional semiconductor device simulation [8]. It is implemented using modern engineering practices that promote reliability, maintanability, and extensibility. Products that use the ATLAS framework meet the device simulation needs of all semiconductor applications. ATLAS should only be used with VWF (Virtual Wafer Fab) Interactive Tools. These include DECKBUILD, TONYPLOT, DEVEDIT, MASKVIEWS, and OPTIMIZE. DECKBUILD provides an interactive run-time environment. TONYPLOT supplies scientific visualization capabilities. DEVEDIT is an interactive tool for structure and mesh specification and refinement, and MASKVIEWS is an IC Layout Editor. The OPTIMIZER supports blackbox optimization across multiple simulators. ATLAS is very often used in conjunction with the ATHENA process simulator. ATHENA predicts the physical structures that result from processing steps. The resulting physical structures are used as input by ATLAS, which then predicts the electrical characteristics associated with specified bias conditions. The combination of ATHENA and ATLAS makes it possible to determine the impact of process parameters on device characteristics.

Figure 37 shows the types of information that flow in and out of ATLAS. Most ATLAS simulations use two inputs: a text file that contains commands for ATLAS to execute and a structure file that defines the structure that will be simulated. ATLAS produces three types of output. The run-time output provides a guide to the progress of simulations running and is where error messages and warning messages appear. Log files store all terminal voltages and currents from the device analysis, and solution files store two- and three-dimensional data relating to the values of solution variables within the device for a single bias point.

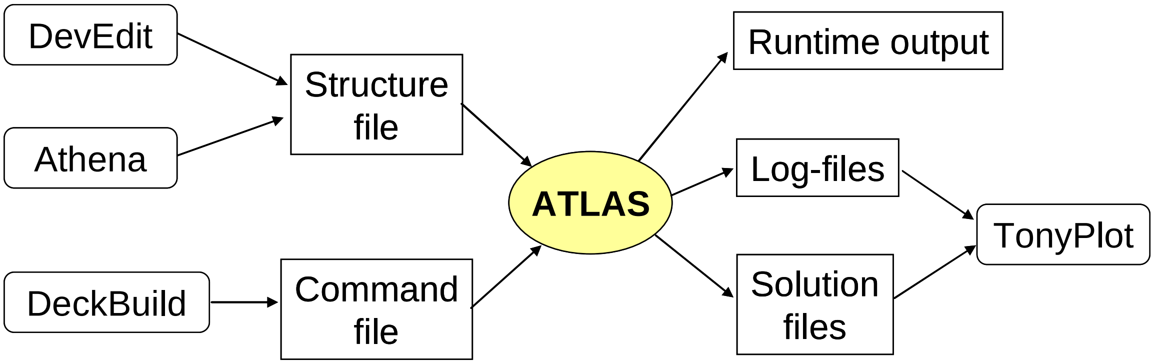


Figure 37: ATLAS Inputs and Outputs.

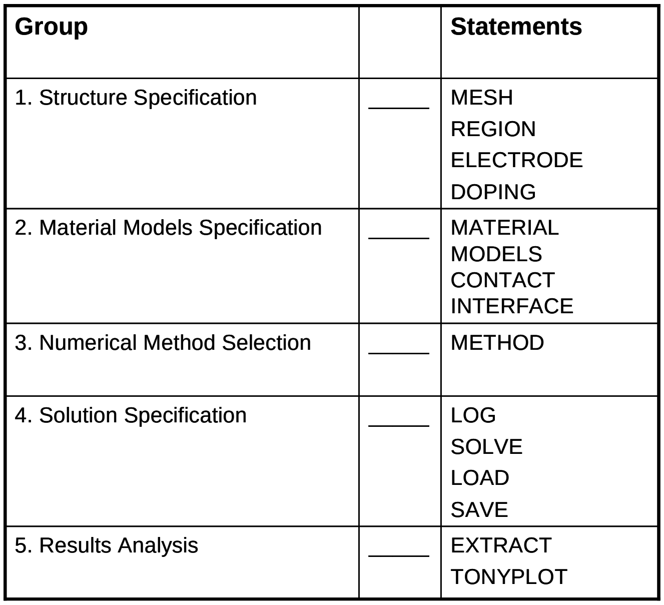
An ATLAS command file is a list of commands for ATLAS to execute. This list is stored as an ASCII text file that can be prepared in DECKBUILD or using any text editor. Preparation of the input file in DECKBUILD is preferred and can be made easier by appropriate use of the DECKBUILD Commands menu. The input file contains a sequence of statements. Each statement consists of a keyword that identifies the statement and a set of parameters. The general format is:

```
<STATEMENT> <PARAMETER>=<VALUE>
```

Some hints on the proper structure of the statements are listed below:

1. The statement keyword must come first, but after this, the order of parameters within a statement is not important.
2. It is only necessary to use enough letters of any parameter to distinguish it from any other parameter on the same statement. Thus `CONCENTRATION` can be shortened to `CONC`. However `REGION` cannot be shortened to `R` since there is also a parameter `RATIO` associated with the `DOPING` statement.
3. Logicals can be explicitly set to false by preceding them with the ^ symbol.
4. Any line beginning with # is ignored. These lines are used as comments.
5. ATLAS can read up to 256 characters on one line; however, it is best to spread long input statements over several lines to make the input file more readable. The character \ at the end of a line indicates continuation.

The order in which statements occur in an ATLAS input file is important. There are five groups of statements, and these must occur in the correct order. These groups are indicated in Figure 38. **Each input file must contain these five groups in order**. Failure to do this will usually cause an error message and termination of the program, but it could lead to incorrect operation of the program. For example, material parameters or models set in the wrong order may not be used in the calculations. The order of statements within the mesh definition, structural definition, and solution groups is also important.



Group		Statements
1. Structure Specification	_____	MESH REGION ELECTRODE DOPING
2. Material Models Specification	_____	MATERIAL MODELS CONTACT INTERFACE
3. Numerical Method Selection	_____	METHOD
4. Solution Specification	_____	LOG SOLVE LOAD SAVE
5. Results Analysis	_____	EXTRACT TONYPLOT

Figure 38: ATLAS Command Groups with the Primary Statements in each group

A device structure can be defined in three different ways in ATLAS:

- An existing structure can be read in from a file. The structure can be created by earlier ATLAS run or by another program such as ATHENA or DEVEDIT. A single statement loads in the mesh, geometry, electrode positions, and `DOPING` of the structure. This statement is `MESH INFILE=<filename>`
- The input structure can be transferred from ATHENA or DEVEDIT through the automatic interface feature of DECKBUILD.
- A structure can be constructed using the ATLAS command language.

The first and second methods are more convenient than the third and are to be preferred whenever possible.

Several different numerical methods can be used for calculating the solutions to semiconductor device problems. Different solution methods are optimum in different situations, so some guidelines will be given here. Different combinations of models will require ATLAS to solve up to six equations. For each of the model types, there are basically three types of solution techniques: (a) de-coupled (`GUMMEL`), (b) fully coupled (`NEWTON`), and (c) `BLOCK`. In simple terms, the de-coupled technique like the Gummel method will solve for each unknown, in turn keeping the other variables constant, repeating the process until a stable solution is achieved. Fully coupled techniques such as the Newton method solve the total system of unknowns together. The combined or block methods will solve some equations fully coupled, while others are de-coupled.

In general, the Gummel method is useful where the system of equations is weakly coupled but has only linear convergence. The Newton method is useful when the system of equations is strongly coupled and has quadratic convergence. The Newton method may however spend extra time solving for quantities which are essentially constant or weakly coupled. Newton also requires a more accurate initial guess to the problem to obtain convergence. Thus, a block method can provide for faster simulations times in these cases over Newton. Gummel can often provide better initial guesses to problems. It can be useful to start a solution with a few Gummel iterations to generate a better guess and then switch to Newton to complete the solution. Specification of the solution method is carried out as follows:

```
METHOD GUMMEL BLOCK NEWTON
```

Sentaurus Software

Sentaurus Device [9] numerically simulates the electrical behavior of a single semiconductor device in isolation or several physical devices combined in a circuit. Terminal currents, voltages, and charges are computed based on a set of physical device equations that describes the carrier distribution and conduction mechanisms. A real semiconductor device, such as a transistor, is represented in the simulator as a virtual device whose physical properties are discretized onto a non-uniform grid (or mesh) of nodes. Therefore, a virtual device is an approximation of a real device. Continuous properties such as doping profiles are represented on a sparse mesh and, therefore, are only defined at a finite number of discrete points in space. The doping at any point between nodes (or any physical quantity calculated by Sentaurus Device) can be obtained by interpolation. Each virtual device structure is described in the Synopsys TCAD tool suite by two files:

- The grid (or geometry) file contains a description of the various regions of the device, that is, boundaries, material types, and the locations of any electrical contacts. This file also contains the grid (the locations of all the discrete nodes and their connectivity).
- The data (or doping) file contains the properties of the device, such as the doping profiles, in the form of data associated with the discrete nodes. By default, a device simulated in 2D is assumed to have a thickness in the third dimension of 1 μm.

Though both numerous and various, the features of Sentaurus Device can be summarized as follows:

- An extensive set of models for device physics and effects in semiconductor devices (drift-diffusion, thermodynamic, and hydrodynamic models).
- General support for different device geometries (1D, 2D, 3D, and 2D cylindrical).
- Mixed-mode support of electrothermal netlists with mesh-based device models and SPICE circuit models.
- Nonvolatile memory simulations are accommodated by robust treatment of floating electrodes in combination with Fowler–Nordheim and direct tunneling, and hot-carrier injection mechanisms.
- Hydrodynamic (energy balance) transport is simulated rigorously to provide a more physically accurate alternative to conventional drift-diffusion formulations of carrier conduction in advanced devices.
- Floating semiconductor regions in devices such as thyristors and silicon-on-insulator (SOI) transistors (floating body) are handled robustly. This allows hydrodynamic breakdown simulations in such devices to be achieved with good convergence.

The mixed device and circuit capabilities give Sentaurus Device the ability to solve three basic types of simulation: single device, single device with a circuit netlist, and multiple devices with a circuit netlist. Multiple-device simulations can combine devices of different mesh dimensionality, and different physical models can be applied in individual devices, providing greater flexibility. In all cases, the circuit netlists can contain an electrical and a thermal section.

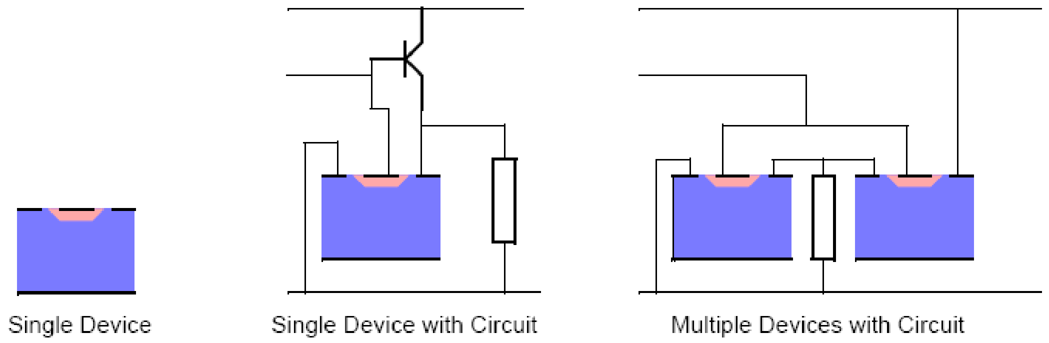


Figure 39: Three types of simulation with Sentaurus.

Device structures can be created in various ways, including 1D, 2D, or 3D process simulation (Sentaurus Process), 3D process emulation (Sentaurus Structure Editor), and 2D (Mdraw and Sentaurus Structure Editor) or 3D (DIP and Sentaurus Structure Editor) structure editors. Regardless of the means used to generate a virtual device structure, it is recommended that the structure be remeshed using Mdraw (2D meshing with an interactive graphical user interface (GUI)) or Mesh (1D, 2D, and 3D meshing without a GUI) to optimize the grid for efficiency and robustness. For maximum efficiency of a simulation, a mesh must be created with a minimum number of vertices to achieve the required level of accuracy. For any given device structure, the optimal mesh varies depending on the type of simulation. It is recommended that to create the most suitable mesh, the mesh must be densest in those regions of the device where the following are expected:

- High current density (MOSFET channels, bipolar base regions);
- High electric fields (MOSFET channels, MOSFET drains, depletion regions in general);
- High charge generation (single event upset (SEU) alpha particle, optical beam).

For example, accurate drain current modeling in a MOSFET requires very fine, vertical mesh spacing in the channel at the oxide interface (of the order 1 Å) when using advanced mobility models. For reliable simulation of breakdown at a drain junction, the mesh must be more concentrated inside the junction depletion region for good resolution of avalanche multiplication. Generally, a total node count of 2000 to 4000 is reasonable for most 2D simulations. Large power devices and 3D structures require a considerably larger number of elements.

A typical device tool flow the creation of a device structure by a process simulation (Sentaurus Process) followed by remeshing using Mdraw (for 2D studies). In this scheme, control of mesh refinement is handled automatically through the file _mdr.cmd (created by Sentaurus Process). Sentaurus Device is used to simulate the electrical characteristics of the device. Finally, Tecplot SV is used to visualize the output from the simulation in 2D and 3D, and Inspect is used to plot the electrical characteristics.

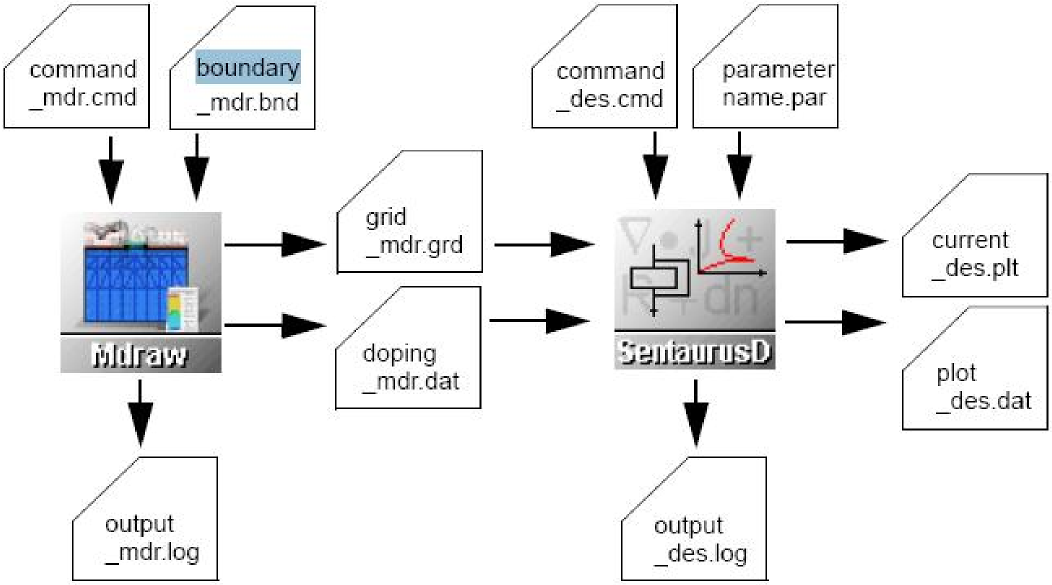


Figure 40: Typical tool flow with device simulation using Sentaurus Device.

5. References

[1] S. M. Sze and G. S. May, *Fundamentals of Semiconductor Fabrication* (John Wiley and Sons Inc., 04 April, 2003).

[2] P. D. Agnello, *IBM J. Res. & Dev.*, Vol. 46, 317 (2002).

[3] B. A. Kramer, R. J. Weber, *Electronics Letters*, Vol. 28, 1106 (1992).

[4] In 1954, Charles Townes and Arthur Schawlow invented the maser. Theodore Maiman invented the ruby laser considered to be the first successful optical or light laser. Many historians claim that Theodore Maiman invented the first optical laser; however, there is some controversy that Gordon Gould was the first.

[5] D. Vasileska and S. M. Goodnick, *Materials Science and Engineering, Reports: A Review Journal* Vol. R38, 181 (2002).

[6] Robert F. Pierret and Gerold W. Neudeck, *Modular Series on Solid State Devices*. Robert F. Pierret, *Semiconductor Device Fundamentals*, Prentice Hall, 1996.

[7] J. C. Bean, "Materials and Technologies," in High-Speed Semiconductor Devices, S. M. Sze, Editor, John Wiley & Sons, New York, 1990.

[8] Silvaco International, Santa Clara, CA, *ATLAS User's Manual*, Ed. 6, 1998 (www.silvaco.com).

[9] www.synopsys.com .