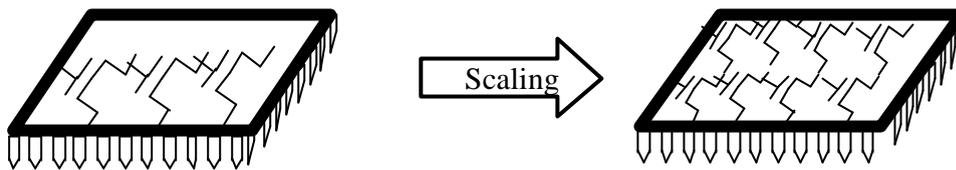


COE360 Course Notes

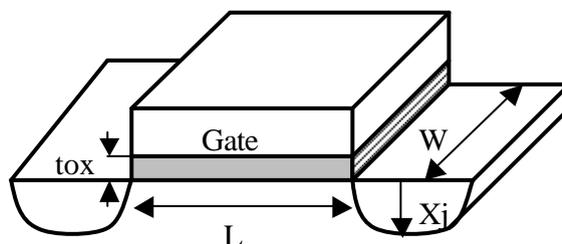
By

Dr. Muhammad Elrabaa

1.1 MOSFET Scaling and Small Geometry Effects



- To increase the number of devices per IC, the device dimensions had to be shrunk from one generation to another (i.e. scaled down)
- In theory, there are two methods of scaling:
 1. Full-Scaling (also called Constant-Field scaling): In this method the device dimensions (both horizontal and vertical) are scaled down by $1/S$, where S is the scaling factor. In order to keep the electric field constant within the device, the voltages have to be scaled also by $1/S$ such that the ratio between voltage and distance (which represents the electric field) remain constant. The threshold voltage is also scaled down by the same factor as the voltage to preserve the functionality of the circuits and the noise margins relative to one another. As a result of this type of scaling the currents will be reduced and hence the total power per transistor ($P=I \times V$) will also be reduced, however the power density will remain constant since the number of transistors per unit area will increase. This means that the total chip power will remain constant if the chip size remains the same (this usually the case).



The table below summarizes how each device parameter scales with S ($S > 1$)

Parameter	Before scaling	After scaling
Channel length	L	L/S
Channel width	W	W/S
Oxide thickness	t_{ox}	t_{ox}/S
S/D junction depth	X_j	X_j/S
Power Supply	VDD	VDD/S
Threshold voltage	V_{TO}	V_{TO}/S
Doping Density	N_A & N_D	$N_A * S$ and $N_D * S$
Oxide Capacitance	C_{ox}	$S * C_{ox}$
Drain Current	I_{DS}	I_{DS}/S
Power/Transistor	P	P/S^2
Power Density/cm ²	p	p

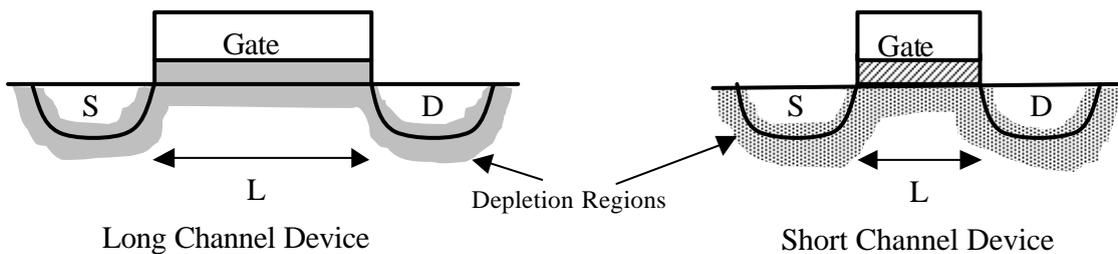
2. Constant-Voltage scaling (CVS): In this method the device dimensions (both horizontal and vertical) are scaled by S, however, the operating voltages remain constant. This means that the electric fields within the device will increase (field = Voltage/distance). The threshold voltages remain constant while the power per transistor will increase by S. The power density per unit area will increase by S^3 ! **This means that for the same chip area, the power chip power will increase by S^3 . This makes constant-voltage-scaling (CVS) very impractical.** Also, the device doping has to be increased more aggressively (by S^2) than the constant-field scaling to prevent channel punch-through. Channel punch-through occurs when the Source and Drain Depletion regions touch one another. By increasing the doping by S^2 , the depletion region thickness is reduced by S (the same ratio as the channel length). **However, there is a limit for how much the doping can be increased (the solid solubility limit of the dopant in Silicon). Again, this makes the CVS impractical in most cases.** The following table summarizes the changes in key device parameters under constant-voltage scaling:

Parameter	Before scaling	After scaling
Channel length	L	L/S
Channel width	W	W/S
Oxide thickness	tox	tox/S
S/D junction depth	Xj	Xj/S
Power Supply	VDD	VDD
Threshold voltage	V _{TO}	V _{TO}
Doping Density	N _A & N _D	N _A * S ² and N _D * S ²
Oxide Capacitance	Cox	S*Cox
Drain Current	I _{DS}	I _{DS} * S
Power/Transistor	P	P*S
Power Density/cn ²	p	p * S ³

In almost all cases, the scaling is a combination of constant-field scaling and constant-voltage scaling, such that the number of devices is increased and the total power/chip does not increase much.

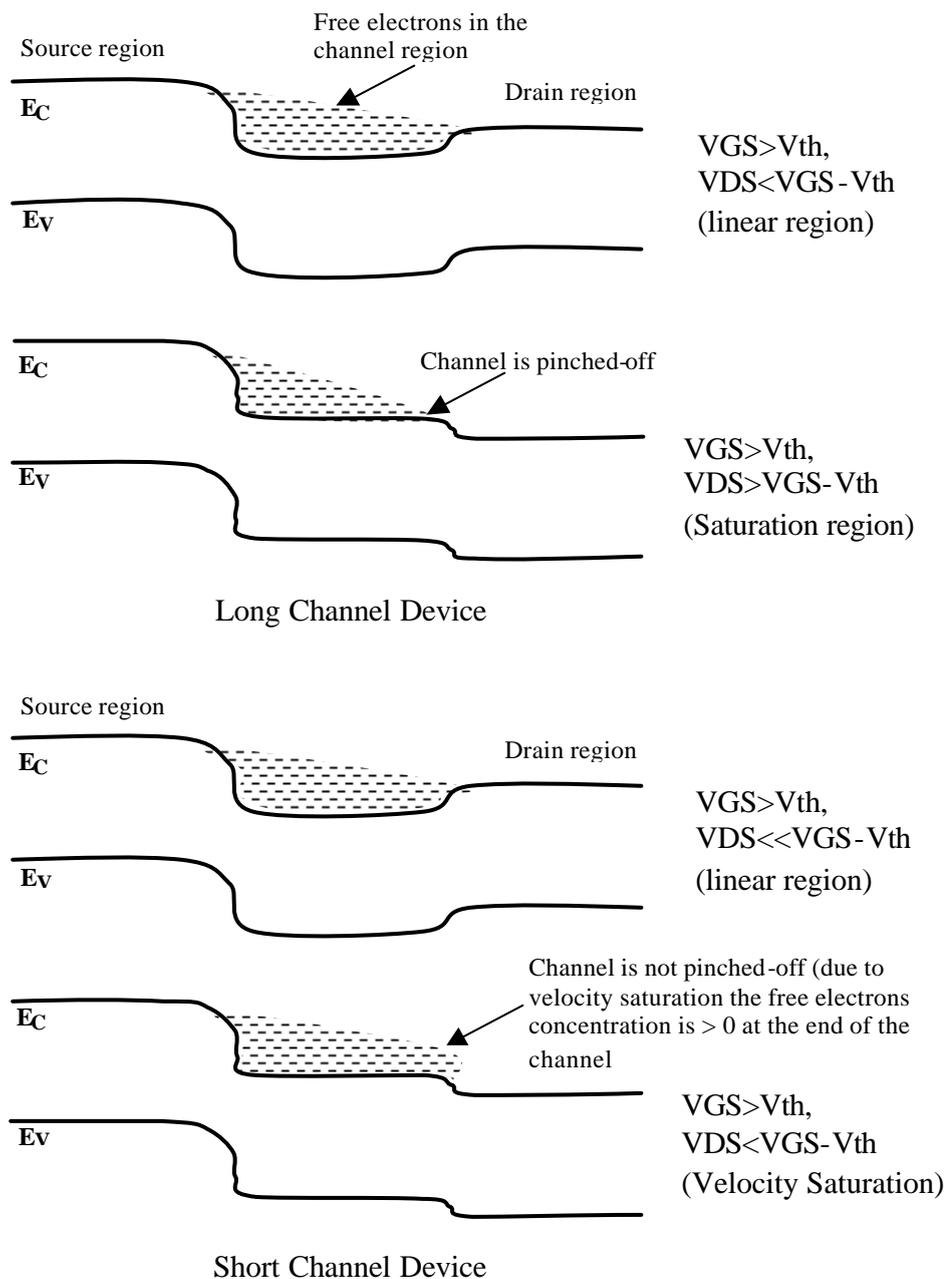
1.2 Small Geometry Effects:

- Short channel effects start to appear as the channel length becomes less than 10 times the depletion region width of the source/drain regions. The figure below demonstrates the difference between long and short channel transistors.



- The major short-channel effects are:

- Carrier velocity saturation:** Recall that the mobility determines the increase in carrier velocity as the electric field increases. But this does not continue forever. As the channel length is reduced, the horizontal electric field between the source and drain increases to a point where the carrier mobility becomes zero, i.e. the carrier velocity won't increase beyond a certain limit (hence the term velocity saturation). Once velocity saturation sets in, the drain current won't increase as V_{DS} increases even if $V_{DS} < V_{GS} - V_{th}$. The figure below shows the effect of velocity saturation using energy band diagram of an NMOS transistor.



In the limit the short-channel device will have a saturation current of:

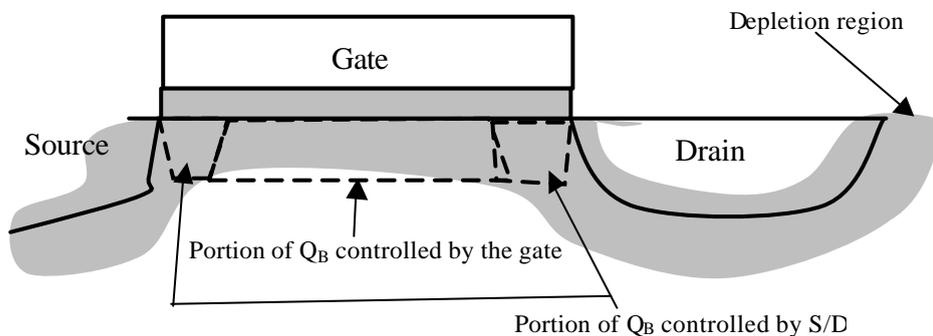
$$I_{DSat} = C_{ox} * W * V_{sat} * V_{DSat}$$

Where V_{sat} is the saturation velocity ($\sim 2E7$ cm/s) and V_{DSat} is the drain to source voltage at the on-set of velocity saturation and depends on L and the substrate doping (semi-empirical formulae are used to determine this voltage). However, for most devices they will have a saturation current characteristic between those of long and short channel devices:

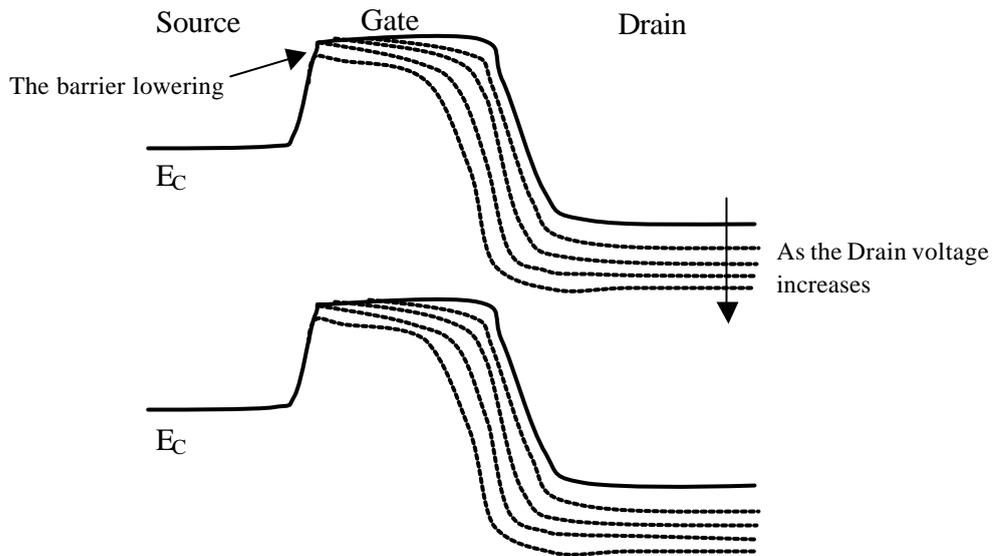
$$I_{DSat} = C_{ox} * W * K (V_{GS} - V_{th})^n$$

Where K and n are constants that are obtained from semi-empirical formulae, $1 < n < 2$ (typically, n is between 1.3 to 1.4). This means that the saturation current won't increase with V_{GS} in a quadratic relation, but rather close to a linear relation, which reduces the speed. Also, once the channel length L goes below the velocity saturation limit, the drain current won't increase with reducing L further, a major set back for scaling efforts (i.e. circuits won't get faster as before with scaling).

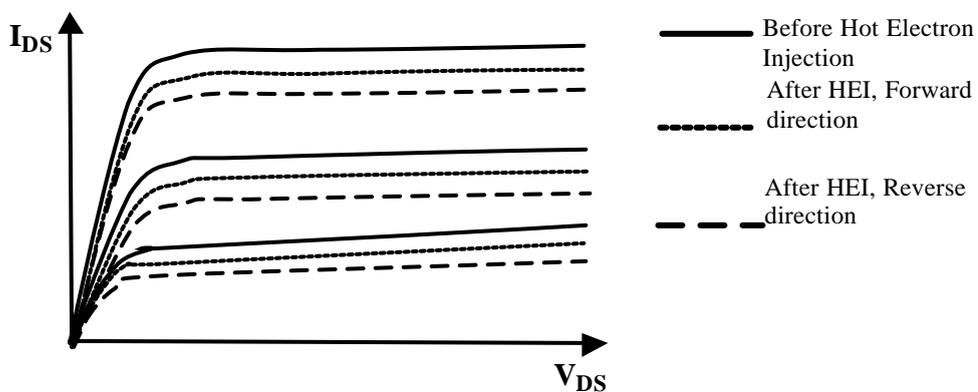
2. **Threshold voltage reduction:** Since for the short-channel devices, a relatively large portion of the channel depletion charge Q_B is controlled by the drain, a smaller amount of gate voltage is required to achieve strong inversion and create the channel, i.e. the threshold voltage will be smaller. The following figure illustrates this.



3. **Drain-Induced-Barrier-Lowering (DIBL):** Because of the short channel length, the drain voltage will reduce the potential barrier between the source and substrate. Hence it become easier for carriers at the source to jump over this barrier and drift to the drain even at the absence of a channel (i.e. when the gate to source voltage is less than the threshold voltage). The DIBL effect causes the MOS devices to “leak” currents when they are turned-off. These currents are called leakage currents or off-currents. If the DIBL effect increases significantly, the MOS switch may become always ON, i.e. it is not controlled by the gate any more. This condition would occur if the source and drain depletion regions reach one another, a condition called **punch-through**. Increasing the substrate doping under the gate/channel reduces the DIBL effect. The figure below demonstrates the DIBL effect using energy band diagram.



4. **Hot Electron Injection into the Gate Oxide:** Due to the high electric fields that result in short-channel devices, electrons can attain enough speed to jump over the energy barrier between the Si channel region and the gate oxide. Once they are injected into the oxide, they become part of its trapped charge. As it is already known, this would alter the device's threshold voltage and may render the device totally useless with time. Also, since the injection occurs near the drain side, where the electrons would have attained the highest speed, the device operation becomes asymmetrical, i.e. the I-V characteristics will exhibit direction dependency. This is because the threshold if the current is flowing from the source to the drain (forward direction) is smaller than the threshold in the reverse direction. The figure below shows the effects of HE injection on the I-V characteristics of an NMOS device.



Effects of Hot Electron Injection (HEI) on NMOS I-V characteristics

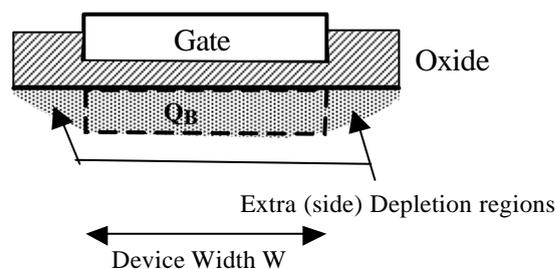
- **Other effects of scaling:**

1. **Formation of gate pinholes:** These are points where the gate electrode is shorted to the substrate. This is due to the scaling down of the gate oxide thickness (t_{ox}), which creates holes in the oxide through which pinholes occur.
2. **Fluctuations in Device parameters:** Due to the smaller channel, fewer dopants get into the channel area. Hence, due to the manufacturing process tolerance, the dopant density will vary from one device to another. Subsequently, the threshold voltage will also vary from one device to another. These fluctuations will increase as the device dimensions shrink (both L and W). Consider as an example a state-of-the-art 0.1 μm technology where we attempt to dope the substrate with a dopant concentration of 10^{16} cm^{-3} . Now if a device has an L of 0.1 μm and a W of 0.5 μm , and assuming the channel thickness to be 50 \AA (Angstroms), then the total number of dopants in the channel area is simply:

$$\# \text{ of dopants} = 0.1\text{E-}4 * 0.5\text{E-}4 * 50\text{E-}8 * 1\text{E}16 = 2.5$$

So even if we assume that there is no process tolerance that means, one device would get 2 dopants, while another would get 3, a difference of 20%.

3. **Narrow Channel Effects:** As the device width gets smaller, the depletion region outside the channel (on the sides) becomes significant relative to the depletion region under the gate. As a result more gate voltage is required to create the required depletion charge under the gate (Q_B) since some voltage is wasted on creating the depletion regions on the sides. Hence, V_{th} will be larger for narrow devices. The following graph illustrates this.



4. **Gate leakage:** Due to the thinning of the gate oxide and the increased vertical fields (between gate and substrate), direct current (DC) start to flow into the gate due to electrons being able to jump over the oxide energy barrier. This causes power dissipation. Also, the MOS now can't be used as a capacitor (which is often done).
5. **Poly Gate depletion:** Due to the high vertical fields, the gate poly starts to form a considerable depletion region, making the effective insulator's thickness larger, thus reducing C_{ox} and consequently the drain current.
6. **Quantum effects:** These are due to the device dimensions approaching the atomic dimensions. At such dimensions, electrons behave like waves rather

than particles, rendering the existing device models useless. The most limiting effect is called band quantization.

7. **Transmission line effects:** As the transistors become faster, the signals rise and fall times become very short, hence long signal wires become like transmission lines. The signal moves as a wave in these lines and is subjected to wave-like behavior (reflections). This causes deterioration in signal quality.

1.3 MOS Capacitances

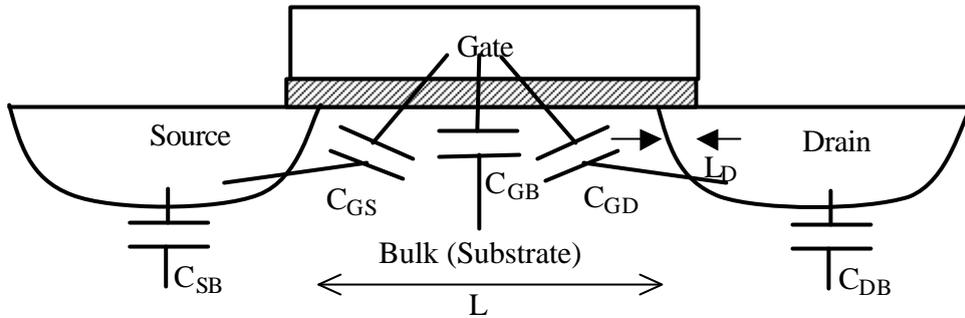
1.3.1 Gate Capacitances

- The build-up and removal of the channel and its associated charge is similar to charging and discharging a capacitor.
- In the case of the channel, this capacitor has an upper plate or electrode, that is the MOS gate, and a lower electrode made of three plates, the source, the bulk (substrate) and the drain. **Hence, charges can enter or leave the upper plate only through the gate terminal. For the lower plate, the charges can enter/leave through any of the three terminals (S, B and D).** Hence the channel charge is lumped (modeled) into three capacitances, as shown in the figure below, Gate-to-Bulk capacitance (C_{GB}), Gate-to-Source capacitance (C_{GS}), and Gate-to-Drain capacitance (C_{GD}). These capacitances are not constant; their values depend on the region of operation. C_{GS} and C_{GD} have two components, called overlap capacitance, that are constant. They basically represent the capacitance between the gate and S/D regions in the overlap area, as shown in the figure below. In the cut-off region, where the channel region is in accumulation (of majority carriers), the gate capacitance is the same as C_{ox} (times $L*W$), and it is all to the bulk (i.e. C_{GS} and $C_{GD} = 0$). When the device is on (i.e. channel is created and surface is in strong-inversion), the channel charge shields the bulk from the gate, i.e. C_{GB} becomes zero and the gate capacitance is distributed between C_{GS} and C_{GD} . In linear region, the gate capacitance is distributed equally between C_{GS} and C_{GD} while in saturation, almost all of the channel charge is controlled by the source, i.e. $C_{GD} = 0$, while $C_{GS} = 2/3 C_{ox} * L * W$. The table below summarizes the values of the gate capacitances for the three different regions of operation as a function of the oxide capacitance C_{ox} , the device length L and width W , and the overlap length L_D between the gate and S/D regions. Also shown below the table, a graph of the gate capacitances versus V_{GS} for the different regions of operation. For this graph, V_{DS} is kept constant. The value of V_{GS} that gives a minimum total gate capacitance is actually the threshold voltage.

1.3.2 S/D Junction Capacitances

- The source and drain forms diodes with the bulk. As seen before these diodes will have junction capacitances that are dependent on the voltage difference between

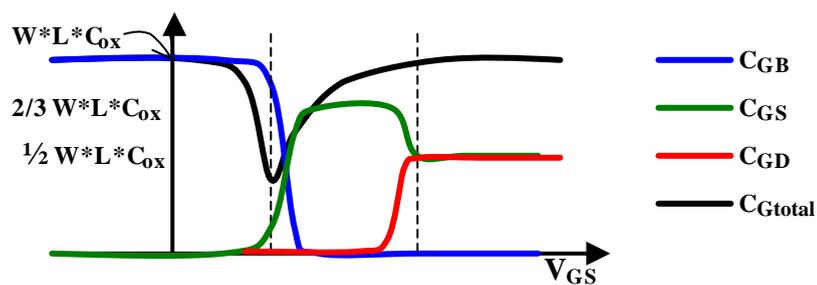
their terminals. Hence, as the source or drain voltages change, these capacitances will be charged or discharged. These capacitances, named C_{SB} and C_{DB} are also shown in the figure below.



The MOS Capacitances

Region \ Capacitance	Cut-Off	Saturation	Linear
C_{GB}	$C_{ox} * W * L$	0	0
C_{GS}	$L_D * W * C_{ox}$	$\frac{2}{3}(C_{ox} * W * L) + L_D * W * C_{ox}$	$\frac{1}{2}(C_{ox} * W * L) + L_D * W * C_{ox}$
C_{GD}	$L_D * W * C_{ox}$	$L_D * W * C_{ox}$	$\frac{1}{2}(C_{ox} * W * L) + L_D * W * C_{ox}$

Values of Gate Capacitances at different regions



The MOS Gate Capacitances versus V_{GS}