Dynamic Behavior

The transient behavior of a *pn-junction* was dominated by:

- The movement of **excess minority carrier charge** in the neutral zones.
- The movement of **space charge** in the depletion region.

MOSFETs are *majority carrier* devices.

Their dynamic behavior is determined solely by the time to:

- Charge and discharge the capacitances between the device ports.
- Charge and discharge of the interconnecting lines.

These capacitances originate from three sources:

- The basic MOS structure.
- The channel charge.
- The depletion regions of the reverse-biased pn-junctions of drain and source.

Aside from the MOS structure capacitances, all capacitors are **nonlinear** and **vary with the applied voltage**.

Dynamic Behavior

MOS Structure Capacitances:

The gate of a MOS transistor is isolated from the channel by the gate oxide where:

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

For the I-V equations, it is useful to have C_{ox} as **large** as possible, by keeping the oxide very thin.

This capacitance is called gate capacitance and is given by:

$$C_g = C_{ox} WL$$

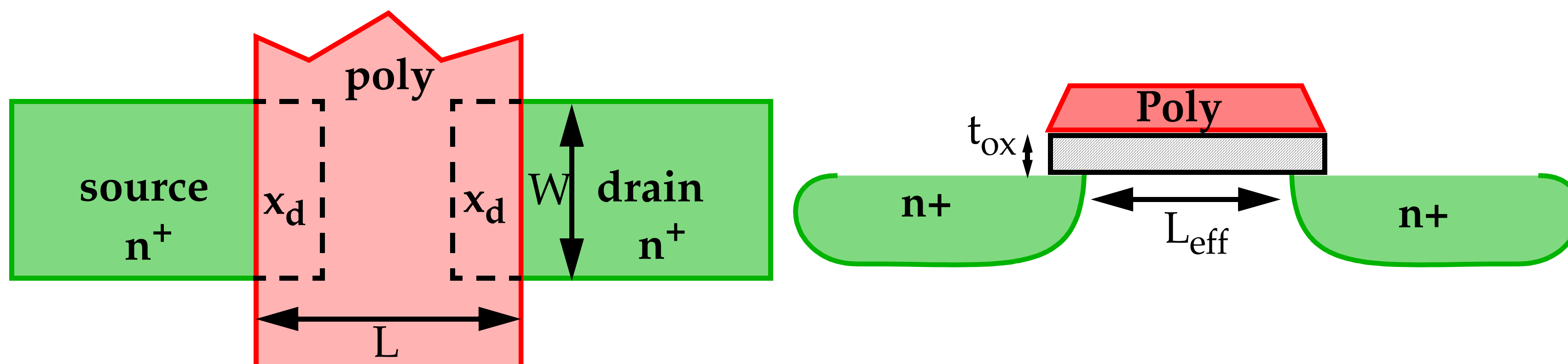
This gate capacitance can be decomposed into several parts:

- One part contributes to the channel charge.
- A second part is due to the topological structure of the transistor.

Let's consider the latter first.

Dynamic Behavior

MOS Structure Capacitances, Overlap:



Lateral diffusion: source and drain diffusion extend under the oxide by an amount x_d.

The effective channel length (L_{eff}) is less than the *drawn length* L by $2*x_d$.

This also gives rise to a linear, fixed capacitance called **overlap capacitance**.

$$C_{gsO} = C_{gdO} = C_{ox}x_dW = C_OW$$

Since x_d is technology dependent, it is usually combined with C_{ox}.

Dynamic Behavior

Channel Charge

The gate-to-channel capacitance is composed of three components, C_{gs}, C_{gd} and C_{gb}. Each of these is non-linear and dependent on the region of operation.

Estimates or average values are often used:

- Triode: C_{gb} ~=0 since the inversion region shields the bulk electrode from the gate.
- Saturation: C_{gb} and C_{gd} are ~= 0 since the channel is pinched off.

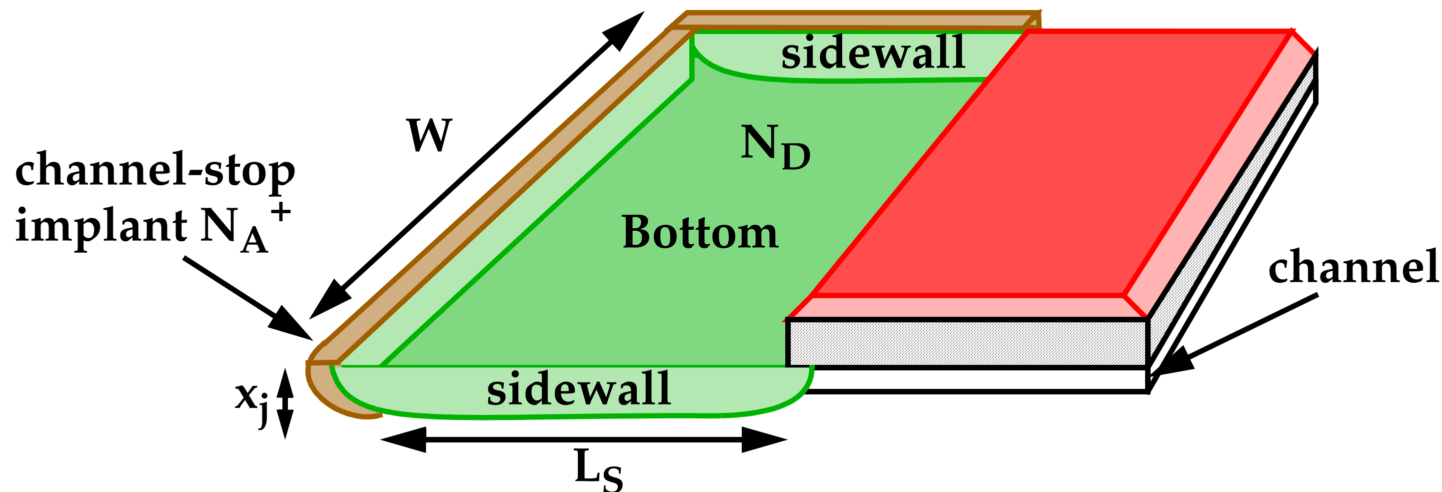
Operation Region	C_{gb}	C_{gs}	C_{gd}
Cutoff	$C_{ox}WL_{eff}$	0	0
Triode	0	$C_{ox}WL_{eff}/2$	$C_{ox}WL_{eff}/2$
Saturation	0	$(2/3)C_{ox}WL_{eff}$	0

Dynamic Behavior

Junction or Diffusion Capacitances

This component is caused by the reverse-biased source-bulk and drain-bulk *pn-junctions*.

We determined that this capacitance is **non-linear** and *decreases* as reverse-bias is *increased*.



- Bottom-plate junction:

  Depletion region capacitance is:

$$C_{bottom} = C_j W L_S$$

with a grading coefficient of m = 0.5 (for an abrupt junction)

Dynamic Behavior

MOS Structure Capacitances, Junction or Diffusion:

- Side-wall junction

Formed by the source region with doping N_D and the p^+channel-stop implant with doping N_A^+.

Since the channel-stop doping is usually higher than the substrate, this results in a higher unit capacitance:

$$C_{sw} = C'_{jsw} x_j (W + 2 \times L_S)$$

with a grading coefficient of m = 1/3.

Note that the channel side is not included in the calculation.

x_j is usually technology dependent and combined with C'_{jsw} as C_{jsw}.

Total junction (small-signal) capacitance is:

$$C_{diff} = C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER$$

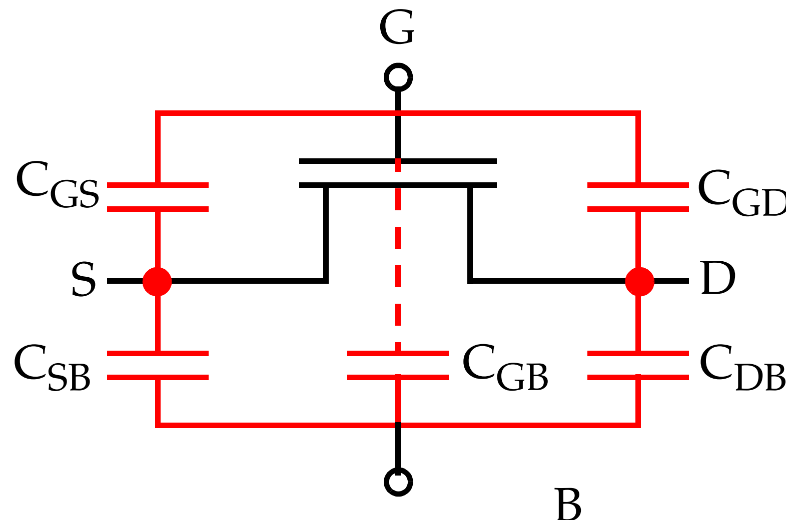
$$C_{diff} = C_j L_S W + C_{jsw}(2L_S + W)$$

As we've done before, we linearize these and use average cap.

Dynamic Behavior

Capacitive Device Model

The previous model can be summarized as:



$$C_{GS} = C_{gs} + C_{gsO}$$
$$C_{GD} = C_{gd} + C_{gdO}$$
$$C_{GB} = C_{gb}$$
$$C_{SB} = C_{Sdiff}$$
$$C_{DB} = C_{Ddiff}$$

The dynamic performance of digital circuits is directly proportional to these capacitances.

Dynamic Behavior

Example:

Given:

$t_{ox} = 6nm$

$L = 0.24um$

$W = 0.36um$

$L_D = L_S = 0.625um$

$C_O = 3 \times 10^{-10} F/m$

$C_{j0} = 2 \times 10^{-3} F/m^2$

$C_{jsw0} = 2.75 \times 10^{-10} F/m$

Determine the zero-bias value of all relevant capacitances.

Gate capacitance, C_{ox}, per unit area is derived as:

$$C_{ox} = \varepsilon_{ox}/t_{ox} = \frac{3.5 \times 10^{-2} fF/um}{6 \times 10^{-3} um} = 5.8 fF/um^2$$

Total gate capacitance C_g is:

$$C_g = WLC_{ox} = 0.36um \times 0.24um \times 5.8fF/um^2 = 0.5fF$$

Overlap capacitance is:

$$C_{GSO} = C_{GDO} = WC_O = 0.108fF$$

Total gate capacitance is:

$$C_{gtot} = C_g + 2 \times C_{GSO} = 0.716fF$$

Dynamic Behavior

Example (cont):

Diffusion capacitance is the sum of bottom:

$$C_{j0}L_DW = 2fF/um^2 \times 0.625um \times 0.36um = 0.45fF$$

Plus side-wall (under zero-bias):

$$C_{jsw0}(2L_D + W) = 2.75\times10^{-1}(2 \times 0.625um + 0.36um) = 0.44fF$$

In this example, diffusion capacitance dominates gate capacitance (0.89 fF vs. 0.716 fF).

Note that this is the worst case condition. Increasing reverse bias reduces diffusion capacitance (by about 50%).

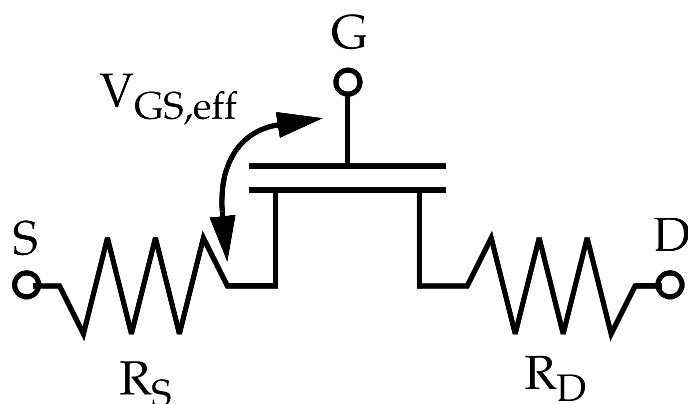
Also note that side-wall dominates diffusion. Advanced processes use SiO_2 to isolate devices (trench isolation) instead of N_A^+ implant.

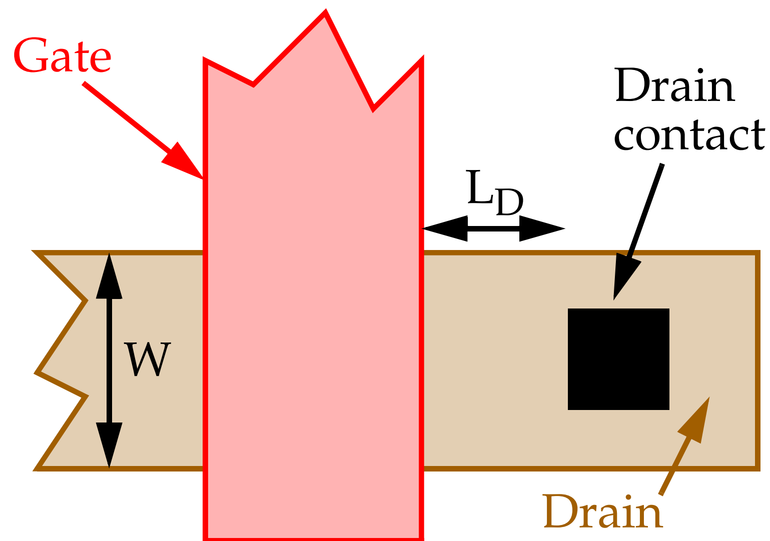
Usually, diffusion is at most equal to gate, very often it is smaller.

Source-Drain Resistance

Scaling causes junctions to be *shallower* and contact openings to be *smaller*.

This increases the parasitic resistance in series with the source and drain.





This resistance can be expressed as:

$$R_{S,D} = \frac{L_{S,D}}{W} R_{\square} + R_C$$

R_C = Contact Resistance

$R_{\square}$ = Sheet resistance(2Ω Ð 100Ω)

$L_{S,D}$ = length of source/drain region.

The series resistance degrades performance by decreasing drain current.

Silicidation used -- low-resistivity material such as *titanium* or *tungsten*.

Secondary Effects

Long-channel devices

One-dimensional model discussed thus far.

Assumed:

- All current flows on the surface of the silicon.
- Electric fields are oriented along that plane.

Appropriate for manual analysis.

Short-channel device

Ideal model does not hold well when device dimensions reach sub-micron range.

The length of the channel becomes comparable to other device parameters such as the depth of the drain and source junctions.

Two-dimensional model is needed.

Computer simulation required.

Secondary Effects

Threshold variations

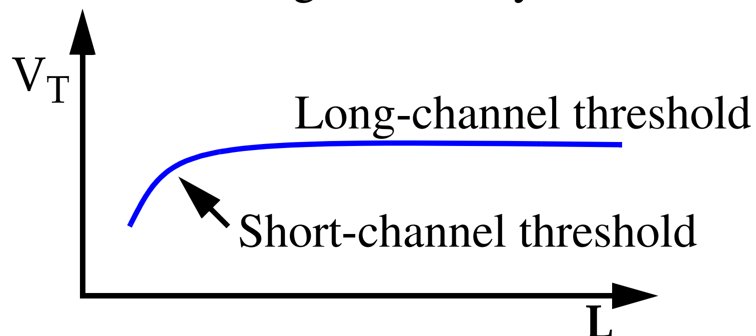
Ideal model assumed *threshold voltage* was only a function of technology parameters and applied body bias, V_{SB}.

With smaller dimensions, the V_{T0} becomes a function of L, W and V_{DS}.

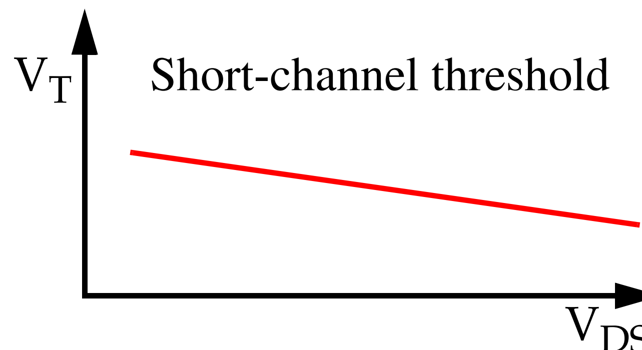
For example, the expression for V_{T0} assumed that all depletion charge beneath the gate originates from the MOS field effects.

We ignored the source and reverse-biased drain depletion regions.

These depletion regions extend under the gate, which in turn **reduces** the threshold voltage necessary to cause strong inversion.



Threshold as a function of length (for low V_{DS})



Drain-induced barrier lowering for small L.

Secondary Effects

Threshold variations

Also, threshold *decreases* with *increasing* V_{DS}.

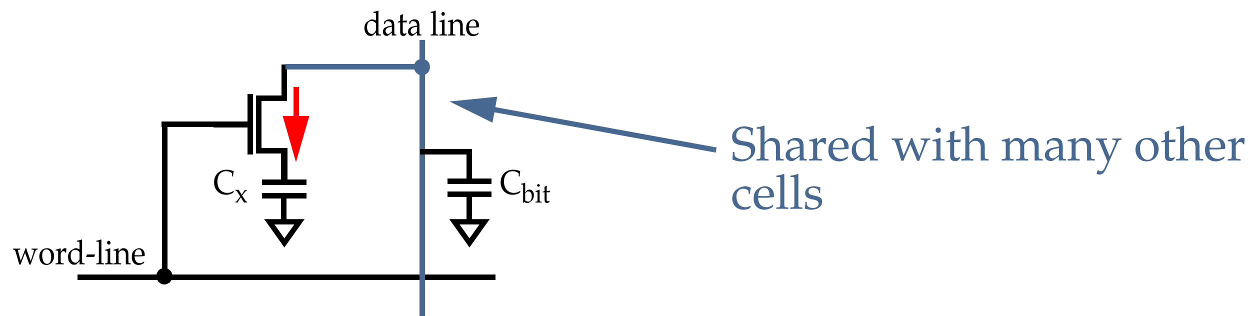
This effect is called *drain-induced barrier lowering* (**DIBL**).

For high values of V_{DS}, the source and drain depletion regions can short together (**punch-through**).

DIBL is a more serious issue than the variation in V_{T0} as a function of length (since most transistors are minimum length transistors).

Particularly for DRAMs.

Leakage current of a cell (e.g. subthreshold current of the access transistor) is a function of voltage on the data line.



Secondary Effects

Threshold variations

Threshold drift also occurs for short-channel devices over time as a result of **hot-carrier effects**.

In the past, *constant voltage scaling* was used which increased the electric field strength and velocity of the electrons.

The electrons can leave the silicon and tunnel into the gate oxide, given enough energy.

Trapped electrons in the oxide *increase* the threshold of NMOS devices and *decrease* the threshold of PMOS devices.

Field strengths of 10^4V/cm are easily reached in submicron devices.

This problem causes **long-term reliability problems**.

Secondary Effects

Variations in the I-V characteristics:

The current-voltage relations deviate significantly from the ideal expressions.

The ideal expressions are:

$$I_D = \frac{1}{2}\mu_n\left(\frac{\varepsilon_{ox}}{t_{ox}}\right)\frac{W}{L}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \qquad \text{(Saturation)}$$

$$I_D = \mu_n\left(\frac{\varepsilon_{ox}}{t_{ox}}\right)\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \qquad \text{(Linear)}$$

The most important reasons for this difference are:

- **Velocity saturation effects**
- **Mobility degradation effects**

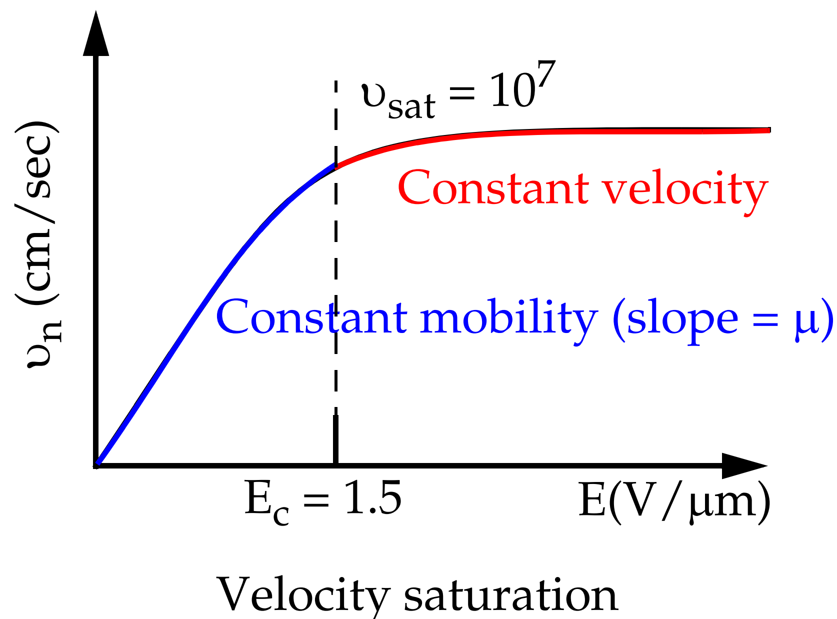
Secondary Effects

Velocity Saturation

We modeled carrier mobility, μ_n, as a constant.

We stated carrier velocity is proportional to the electric field, independent of its value.

This holds up to a critical value of electric field, E_c, after which the velocity of the carriers tends to saturate:



Velocity saturation

Electrons in p-type silicon:

$$E_c = 1 - 5 V/\mu m$$

$$\upsilon_{sat} = 10^7 cm/\sec$$

Therefore, only about 2 volts are needed for NMOS devices with a channel length of 0.25mm.

Holes in n-type silicon:

$$E_c > 10 V/\mu m$$

$$\upsilon_{sat} = 10^7 cm/\sec \quad \text{(same)}$$

Secondary Effects

Velocity Saturation

Revised linear equation:

$$I_D = \frac{\mu_n C_{ox}}{1 + \left(\frac{V_{DS}}{\xi_c L}\right)} \left(\frac{W}{L}\right) \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right]$$

$$I_D = (\mu_n C_{ox}) \left(\frac{W}{L}\right) \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \kappa(V_{DS})$$

with $\kappa(V) = \frac{1}{1 + \left(\frac{V}{\xi_c L}\right)}$

For long-channel devices (L is large) or small values of V_{DS}, κ approaches 1, and the equation simplifies to the traditional equation.

For short channel devices, κ is less than 1 and current is reduced.

Secondary Effects

Velocity Saturation

Revised saturation equation:

$$I_{DSAT} = \upsilon_{sat} C_{ox} W (V_{GS} - V_T - V_{DSAT})$$

$$I_D = \kappa(V_{DSAT}) \mu_n C_{ox} \left(\frac{W}{L}\right) \left[(V_{GS} - V_T) V_{DSAT} - \frac{V_{DSAT}^2}{2} \right]$$

with $$V_{DSAT} = \kappa(V_{GS} - V_T)(V_{GS} - V_T)$$

Further increases in V_{DS} does NOT yield more current and the transistor current saturates at I_{DSAT}.

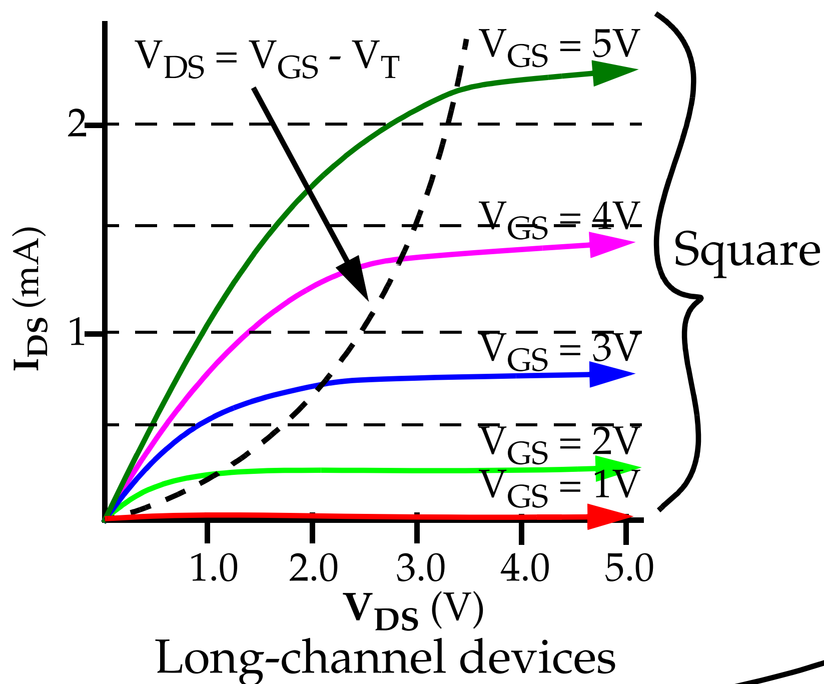
For $V_{DSAT} < V_{GS} - V_T$ (for short channel devices), the device enters saturation before V_{DS} reaches V_{GS}-V_T.

Saturation region is extended.

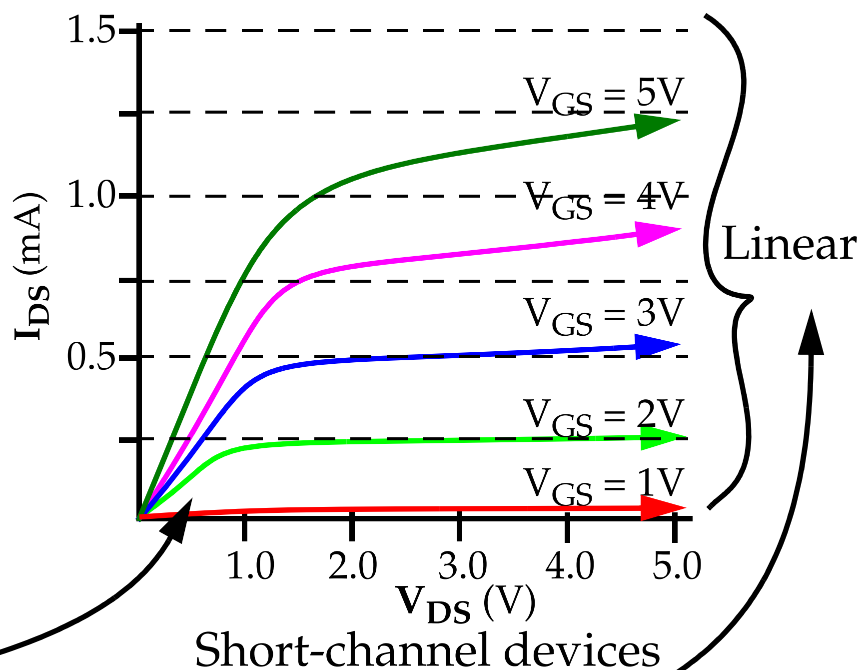
Secondary Effects

Velocity Saturation

This yields a *linear* relationship between the saturation current and the gate-source voltage.



Long-channel devices



Short-channel devices

$W = 100\mu m$
$L = 20\mu m$

Linear relationship with V_{GS}

Extended saturation region (to be discussed).

$W = 4.6\mu m$
$L = 1.2\mu m$

However, reducing the operating voltage does **not** have such a significant effect in submicron devices as it would for long-channel devices.

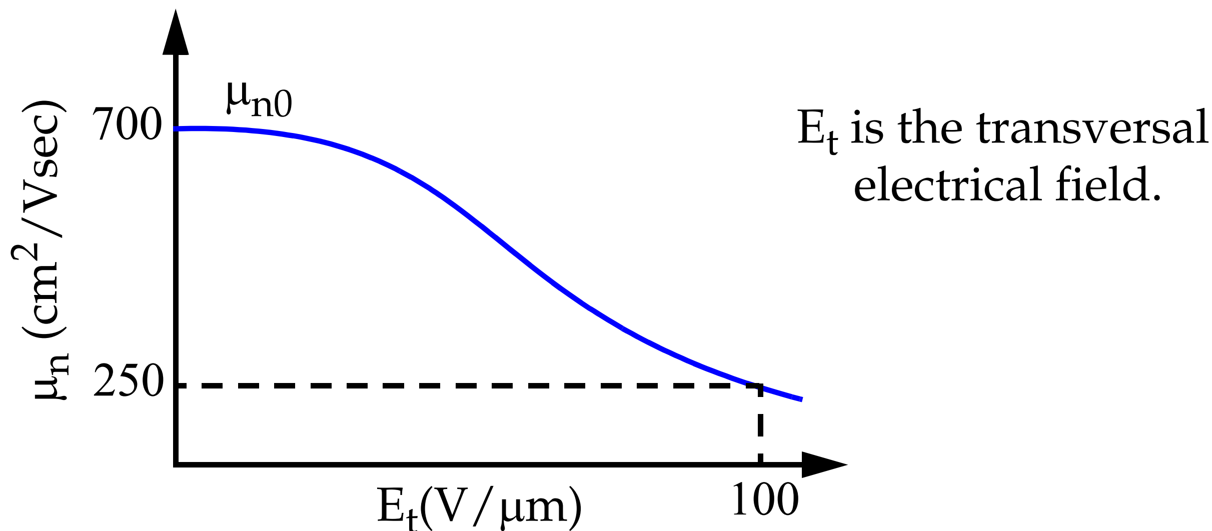
Secondary Effects

Mobility Degradation

Mobility degradation is a second effect of reducing channel-length.

This reduces transistor current even at "normal" electric field levels.

The reduction in the *electron mobility* is caused by the **vertical** component of the electric field (which was ignored before).



E_t is the transversal electrical field.

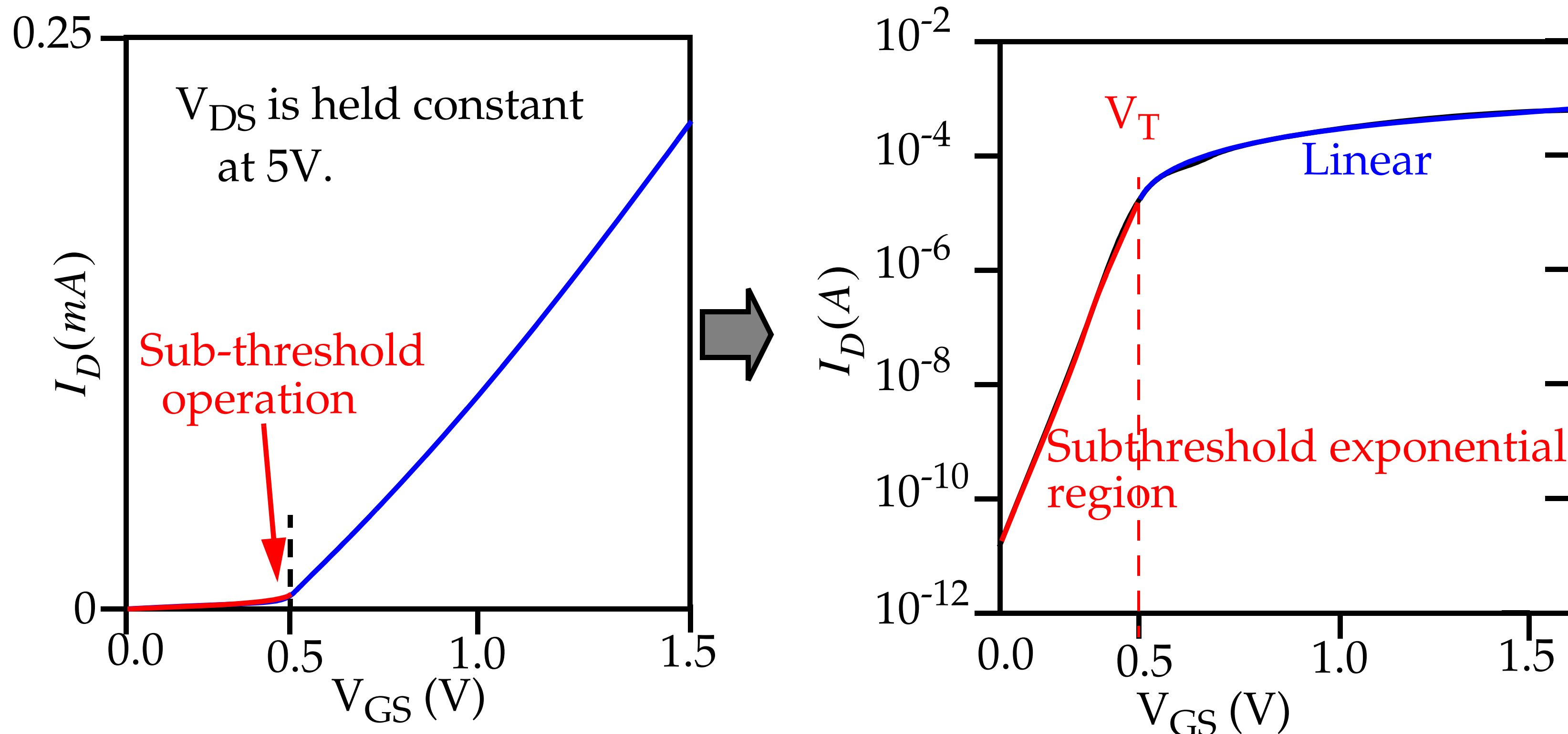
Mobility Degradation

Secondary Effects

Subthreshold Conduction

The transistor is partially conducting for voltages below the threshold voltage.

The region is referred to as *weak-inversion*.



Right logarithmic plot shows current decays in an exponential fashion.

Secondary Effects

Subthreshold Conduction

In the absense of a conducting channel, the n^+ (source) - p (bulk) - n^+ (drain) terminals actually form a *parasitic bipolar transistor*.

The rate of decrease of current is described by:

$$I_D = I_S e^{\frac{V_{GS}}{nkT/q}}\left(1 - e^{\frac{-V_{DS}}{kT/q}}\right)(1 + \lambda V_{DS})$$

where I_S and n are empirical parameters (n ~1.5)

Ideally, I_D should fall to zero very quickly after V_{GS} falls below V_T.

The *inverse* rate of decline of the current w.r.t. V_{GS} below V_T is a quality measure of a device, and can be quantified by the **slope factor** *S*.

$$S = n\left(\frac{kT}{q}\right)\ln(10) \qquad \text{mV/decade}$$

S measures by how much V_{GS} has to be reduced for the drain current to drop by a factor of 10.

Secondary Effects

Subthreshold Conduction

For an ideal transistor, with the sharpest possible roll off, $n = 1$ and $(kT/q)ln(10)$ evaluates to 60 mV/decade at room temperature.

Therefore, subthreshold current drops by a factor of 10 for a reduction in V_{GS} of 60 mV.

Unfortunately, n is greater than 1 for actual devices and current falls at a reduced rate (90 mV/decade for $n = 1.5$).

The current roll-off is *further decreased* by a rise in operating temperature (most chips operate at a temperature considerably above room temp).

Minimizing these leakages is particularly important in *dynamic* circuits, which store logic values as charge on a capacitor.

The value of n is affected by different process technologies, e.g., SOI.

SPICE Models

The complexity of the behavior of the short-channel MOS transistor has resulted in a variety of models of different accuracy and computing efficiency.

The LEVEL parameter in the model statement selects the model:

- **LEVEL 1**:

  Implements the *Shichman-Hodges* model, which is based on the square law long-channel expressions.

  Best used to verify a manual analysis.

- **LEVEL 2**:

  Geometry-based model, which uses **detailed device physics** to define its equations.

  It handles effects such as *velocity saturation*, *mobility degradation* and *DIBL* but is too complex and inaccurate to handle all 3D effects.

SPICE Models

- **LEVEL 3**:

  A semi-empirical model (depends on measured device data to define its parameters).

  Works well for channel lengths down to 1 mm.

- **LEVEL 4 (LEVEL 49):**

  Berkeley Short-Channel IGFET Model (**BSIM**).

  Provides an analytically simple model that is based on a small number of parameters extracted from experimental data.

  It is accurate as well as simple and is the most popular model.

- **LEVEL 5 - n**:

  There are many other models supplied by SPICE vendors and semiconductor manufacturers.

Some of the parameters on the following slides are redundant.

For example, **PHI** can be computed from process model parameters.

User-defined values always override those that can be computed.

SPICE Parameters LEVELs 1-3

Parameter Name	Symbol	SPICE name	Units	Default Value
SPICE Model Index		LEVEL	-	1
Zero-Bias Threshold Voltage	V_{T0}	VTO	V	0
Process Transconductance	k'	KP	A/V^2	1.0E-5
Body-Bias Parameter	γ	GAMMA	$V^{0.5}$	0
Channel Modulation	λ	LAMBDA	1/V	0
Oxide Thickness	t_{ox}	TOX	m	1.0E-7
Lateral Diffusion	x_d	LD	m	0
Metallurgical Junction Depth	x_j	XJ	m	0
Surface Inversion Potential	$2\|\phi_F\|$	PHI	V	0.6
Substrate Doping	N_A, N_D	NSUB	cm^{-3}	0
Surface-State Density	Q_{ss}/q	NSS	cm^{-3}	0
Fast Surface-State Density		NFS	cm^{-3}	0
Total Channel Charge Coef		NEFF	-	1
Type of Gate Material		TPG	-	1
Surface Mobility	μ_0	U0	$cm^2/V\text{-}sec$	600
Maximum Drift Velocity	υ_{max}	VMAX	m/s	0
Mobility Critical Field	E_{crit}	UCRIT	V/cm	1.0E4
Critical Field Exponent in MD		UEXP	-	0

SPICE Parameters LEVELs 1-3

Parameter Name	Symbol	SPICE name	Units	Default Value
Transverse Field Exponent (mobility)		UTRA	-	0
Source Resistance	R_S	RS	Ω	0
Drain Resistance	R_D	RD	Ω	0
Sheet Resistance (Source/Drain)	R/sq	RSH	Ω/sq	0
Zero-Bias Bulk Junction Cap	C_{j0}	CJ	F/m^2	0
Bulk Junction Grading Coeff.	m	MJ	-	0.5
Zero-Bias Side-Wall Junction Cap.	C_{jsw0}	CJSW	F/m	0
Side-Wall Grading Coeff.	m_{sw}	MJSW	-	0.3
Gate-Bulk Overlap Cap.	C_{gbO}	CGBO	F/m	0
Gate-Source Overlap Cap.	C_{gsO}	CGSO	F/m	0
Gate-Drain Overlap Cap.	C_{gdO}	CGDO	F/m	0
Bulk Junction Leakage Current	I_S	IS	A	0
Bulk Junction Leakage Current Density	J_S	JS	A/m^2	1E-8
Bulk Junction Potential	ϕ_0	PB	V	0.8

SPICE Individual Transistor Parameters

The following parameters are specified on the device line, not within the transistor.

Parameter Name	Symbol	SPICE name	Units	Default Value
Drawn Length	L	L	m	-
Effective Width	W	W	m	-
Source Area	AREA	AS	m^2	0
Drain Area	AREA	AD	m^2	0
Source Perimeter	PERIM	PS	m	0
Drain Perimeter	PERIM	PD	m	0
Squares of Source Diffusion		NRS	-	1
Squares of Drain Diffusion		NRD	-	1

Note that zero is assumed for many of these if left unspecified !

NRS and **NRD** multiply the sheet resistance **RSH** specified in the model to give the series source and drain resistance.

```
M1 2 1 0 0 NMOS W=1.8U L=1.2U NRS=0.333 NRD=0.333
   + AD=6.5P PD=9.0U AS=6.5P PS=9.0U
M2 2 1 5 5 PMOS W=5.4U L=1.2U NRS=0.111 NRD=0.111
   + AD=16.2P PD=11.4U AS=16.2P PS=11.4U
```