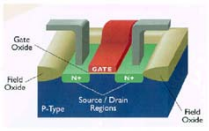
The Devices: MOS Transistor Dynamics

[Adapted from Rabaey's *Digital Integrated Circuits*, ©2002, J. Rabaey et al.]

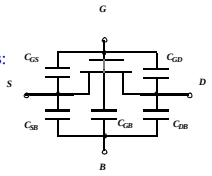
Overview - Transistor Dynamics

- Transistor capacitances
- Sub-Micron MOS Transistor
  - » Threshold Variations
  - » Velocity Saturation
  - » Sub-Threshold Conduction and Leakage
- Latchup
- Process Variations
- Future Perspectives



Dynamic Behavior of MOS Transistor

- MOSFET is a majority carrier device (unlike *pn* junction diode)
- Delays depend on the time to (dis)charge the capacitances between MOS terminals
- Capacitances originate from three sources:
  - basic MOS *structure* (layout)
  - charge present in the *channel*
  - *depletion regions* of the reverse-biased *pn*-junctions of drain and source
- Capacitances are *non-linear* and vary with the applied voltage



MOS Structure Capacitances

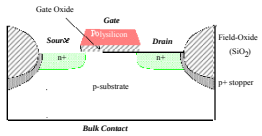
Gate Capacitance

- Gate isolated from channel by gate oxide

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

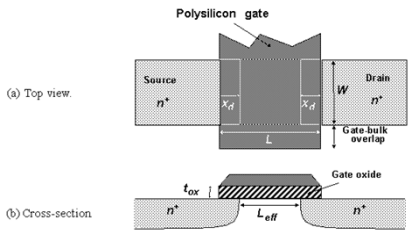
- t_{ox} is very small <10nm
- Results in *gate capacitance* C_g

$$C_g = C_{ox} WL$$



CROSS-SECTION of NMOS Transistor

The Gate Capacitance



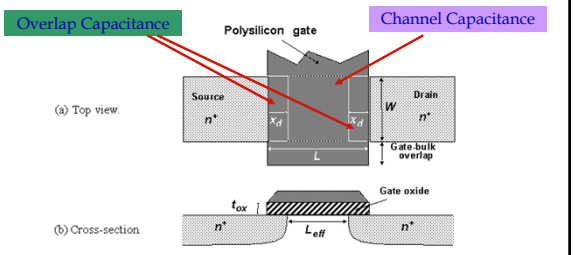
$$C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL$$

The Gate Capacitance

Gate Capacitance depends on
- channel charge (non-linear)
- topology

Capacitance due to topology
- Source and drain extend below the gate oxide by x_d (*lateral diffusion*)
- Effective length of the channel L_{eff} is shorter than the drawn length by factor of $2x_d$
- Cause of parasitic overlap capacitance, C_{gsO}, between gate and source (drain)

The Gate Capacitance



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

The Channel Capacitance

Channel Capacitance has three components
- capacitance between gate and source, C_{gs}
- capacitance between gate and drain, C_{gd}
- capacitance between gate and bulk region, C_{gb}

Channel Capacitance values
- non-linear, depends on operating region
- averaged to simplify analysis

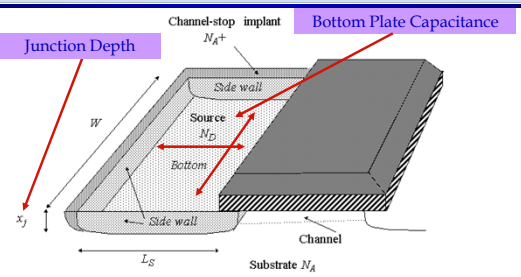
The Channel Capacitance

Different distributions of gate capacitance for varying operating conditions

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

Most important regions in digital design: saturation and cut-off

Diffusion Capacitance



$$C_{diff} = C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER$$
$$= C_j L_S W + C_{jsw}(2L_S + W)$$

Capacitive Device Model

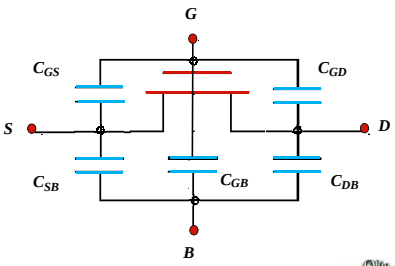
$C_{GS} = C_{gs} + C_{gsO}$

$C_{GD} = C_{gd} + C_{gdO}$

$C_{GB} = C_{gb}$

$C_{SB} = C_{Sdiff}$

$C_{DB} = C_{Ddiff}$



Transistor Capacitance Values for 0.25µ

Example: For an NMOS with L = 0.24 µm, W = 0.36 µm, $L_D = L_S = 0.625$ µm

$C_{GSO} = C_{GDO} = C_{ox} x_d W = C_o W = 0.11$ fF

$C_{GC} = C_{ox} WL = 0.52$ fF

so $C_{gate\_cap} = C_{ox}WL + 2C_oW = 0.74$ fF

Capacitance of both source and drain

$C_{bp} = C_j L_S W = 0.45$ fF

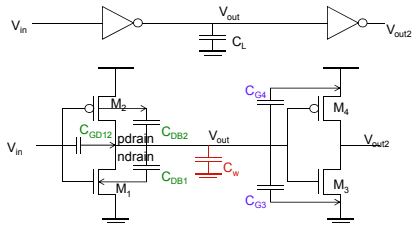
$C_{sw} = C_{jsw}(2L_S + W) = 0.45$ fF

so $C_{diffusion\_cap} = 0.90$ fF

Overlap capacitance

	C_{ox} (fF/µm²)	C_o (fF/µm)	C_j (fF/µm²)	m_j	ϕ_b (V)	C_{jsw} (fF/µm)	m_{jsw}	ϕ_{bsw} (V)
NMOS	6	0.31	2	0.5	0.9	0.28	0.44	0.9
PMOS	6	0.27	1.9	0.48	0.9	0.22	0.32	0.9

Review: Sources of Capacitance



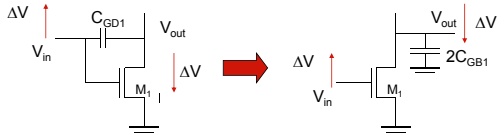
intrinsic MOS transistor capacitances

extrinsic MOS transistor (fanout) capacitances

wiring (interconnect) capacitance

Gate-Drain Capacitance: The Miller Effect

- M1 and M2 are either in cut-off or in saturation.
- The floating gate-drain capacitor is replaced by a capacitance-to-ground (gate-bulk capacitor).



- A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground whose value is two times the original value

Drain-Bulk Capacitance: K_{eq}'s (for 2.5 μm)

- We can simplify the diffusion capacitance calculations *even* further by using a K_{eq} to relate the linearized capacitor to the value of the junction capacitance under zero-bias

$$C_{eq} = K_{eq}\, C_{j0}$$

	high-to-low		low-to-high	
	K_{eqbp}	K_{eqsw}	K_{eqbp}	K_{eqsw}
NMOS	0.57	0.61	0.79	0.81
PMOS	0.79	0.86	0.59	0.7

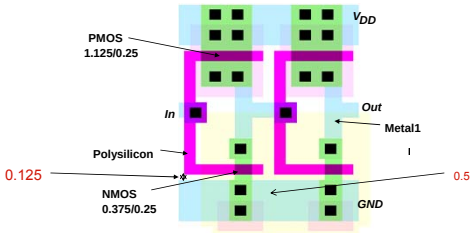
Extrinsic (Fan-Out) Capacitance

- The extrinsic, or fan-out, capacitance is the total gate capacitance of the loading gates M3 and M4.

$C_{fan\text{-}out} = C_{gate}$ (NMOS) + C_{gate} (PMOS)
$= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox})$

- Simplification of the actual situation
  - » Assumes all the components of C_{gate} are between V_{out} and GND (or V_{DD})
  - » Assumes the channel capacitances of the loading gates are constant

Layout of Two Chained Inverters



	W/L	AD (μm²)	PD (μm)	AS (μm²)	PS (μm)
NMOS	0.375/0.25	0.3	1.875	0.3	1.875
PMOS	1.125/0.25	0.7	2.375	0.7	2.375

Components of C_L (0.25 μm)

C Term	Expression	Value (fF) H→L	Value (fF) L→H
C_{GD1}	$2\, C_{on} W_n$	0.23	0.23
C_{GD2}	$2\, C_{op} W_p$	0.61	0.61
C_{DB1}	$K_{eqbpn} AD_n C_j + K_{eqswn} PD_n C_{jsw}$	0.66	0.90
C_{DB2}	$K_{eqbpp} AD_p C_j + K_{eqswp} PD_p C_{jsw}$	1.5	1.15
C_{G3}	$(2\, C_{on}) W_n + C_{ox} W_n L_n$	0.76	0.76
C_{G4}	$(2\, C_{op}) W_p + C_{ox} W_p L_p$	2.28	2.28
C_w	from extraction	0.12	0.12
C_L	Σ	6.1	6.0

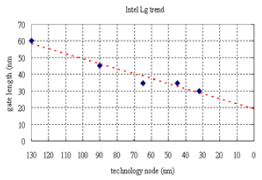
The Sub-Micron MOS Transistor

- Actual transistor deviates substantially from model
- Channel length becomes comparable to other device parameters. Ex: depth of drain and source junctions
- Referred to as a *short-channel* device
- Influenced heavily by secondary effects
- Latchup problems

The Sub-Micron MOS Transistor

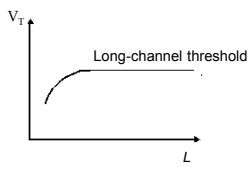
Secondary Effects:

- Threshold Variations
- Parasitic Resistances
- Velocity Saturation
- Mobility Degradation
- Sub-threshold Conduction

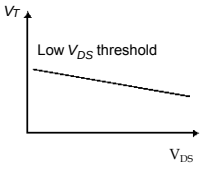


Threshold Variations

- Part of the region below gate is depleted by source and drain fields, which reduce the threshold voltage for short channel.
- Similar effect is caused by increase in V_{DS}, so threshold is smaller with larger V_{DS}



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

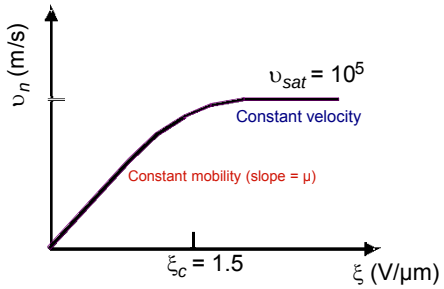


Drain-induced barrier lowering lowers V_T for short channel device

Variations in I-V Characteristics

- The velocity of the carriers is proportional to the electric field up to a point.
- When electric field reaches a critical value, E_{sat}, the velocity *saturates*.
- When the channel length decreases, only a small V_{DS} is needed for saturation
- Causes a *linear dependence* of the saturation current wrt the gate voltage (in contrast to *squared dependence* of long-channel device)
- Current drive cannot be increased by decreasing L

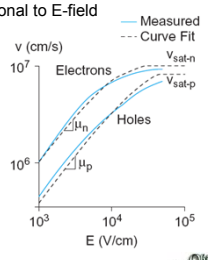
Velocity Saturation



Velocity Saturation

- We assumed carrier velocity is proportional to E-field
  - » $v = \mu E_{lat} = \mu V_{ds}/L$
- At high fields, this ceases to be true
  - » Carriers scatter off atoms
  - » Velocity reaches v_{sat}
    - Electrons: 6-10 x 10^6 cm/s
    - Holes: 4-8 x 10^6 cm/s
  - » Better model

$$v = \frac{\mu E_{lat}}{1 + \frac{E_{lat}}{E_{sat}}} \Rightarrow v_{sat} = \mu E_{sat}$$



Voltage-Current Relation: Velocity Saturation

For short channel devices

- Linear: When $V_{DS} \leq V_{GS} - V_T$

  $I_D = \kappa(V_{DS})\, k'_n\, W/L\, [(V_{GS} - V_T)V_{DS} - V_{DS}^2/2]$

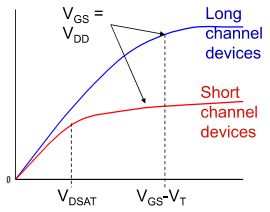
  where

  $\kappa(V) = 1/(1 + (V/(\xi_c L)))$ is a measure of the degree of velocity saturation

- Saturation: When $V_{DS} = V_{DSAT} \geq V_{GS} - V_T$

  $I_{DSat} = \kappa(V_{DSAT})\, k'_n\, W/L\, [(V_{GS} - V_T)V_{DSAT} - V_{DSAT}^2/2]$

Velocity Saturation Effects

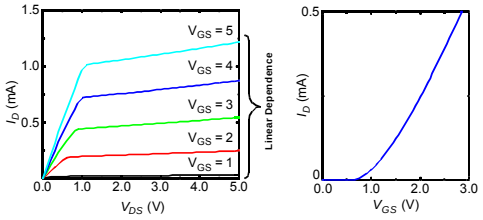


For short channel devices and large enough $V_{GS} - V_T$

- $V_{DSAT} < V_{GS} - V_T$ so the device enters saturation before V_{DS} reaches $V_{GS} - V_T$ and operates more often in saturation

- I_{DSAT} has a linear dependence wrt V_{GS} so a reduced amount of current is delivered for a given control voltage

Velocity Saturation



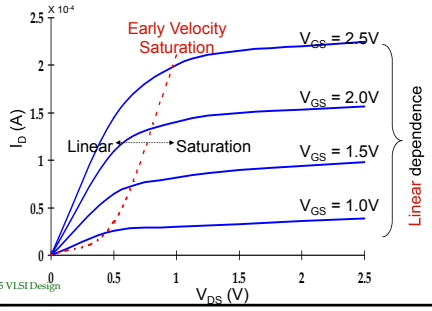
(a) I_D as a function of V_{DS}

(b) I_D as a function of V_{GS} (for V_{DS} = 5 V)

Linear Dependence on V_{GS}

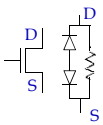
Short Channel I-V Plot (NMOS)

NMOS transistor, 0.25um, L_d = 0.25um, W/L = 1.5, V_{DD} = 2.5V, V_T = 0.4V



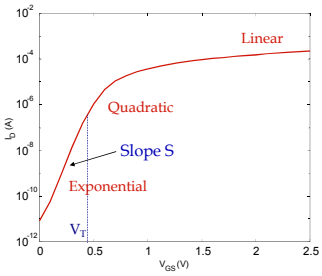
Leakage Sources

- Subthreshold conduction
  - » Transistors can't abruptly turn ON or OFF
- Junction leakage
  - » Reverse-biased PN junction diode current
- Gate leakage
  - » Tunneling through ultrathin gate dielectric



- Subthreshold leakage is the biggest source of DC power dissipation in modern transistors

Sub-Threshold Conduction



The Slope Factor

$$I_D \sim I_0 e^{\frac{qV_{GS}}{nkT}}, \quad n = 1 + \frac{C_D}{C_{ox}}$$

S is ΔV_{GS} for I_{D2}/I_{D1} =10

$$S = n\left(\frac{kT}{q}\right)\ln(10)$$

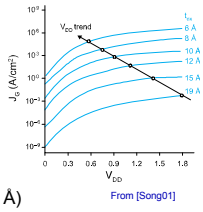
Typical values for S: 60 .. 100 mV/decade

Gate Leakage

- Carriers tunnel thorough very thin gate oxides
- Exponentially sensitive to t_{ox} and V_{DD}

$$I_{gate} = WA\left(\frac{V_{DD}}{t_{ox}}\right)^2 e^{-B\frac{t_{ox}}{V_{DD}}}$$

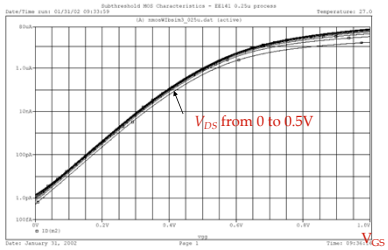
  - » A and B are tech constants
  - » Greater for electrons
    - – So nMOS gates leak more
- Negligible for older processes (t_{ox} > 20 Å)
- Critically important at 65 nm and below (t_{ox} ≈ 10 Å=1nm)



From [Song01]

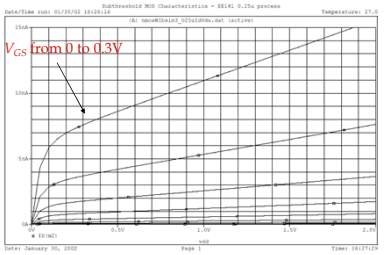
Sub-Threshold I_D vs V_{GS}

$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}}\left(1 - e^{-\frac{qV_{DS}}{kT}}\right)$$

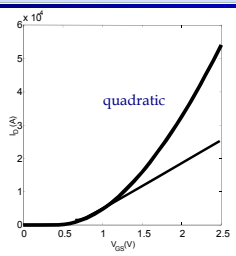


Sub-Threshold I_D vs V_{DS}

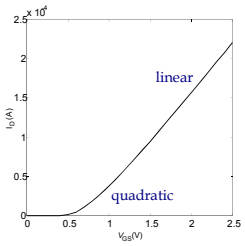
$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}}\left(1 - e^{-\frac{qV_{DS}}{kT}}\right)\left(1 + \lambda \cdot V_{DS}\right)$$



I_D versus V_{GS}

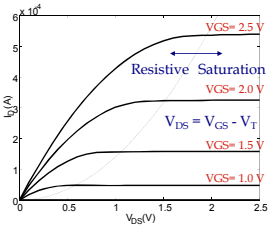


Long Channel

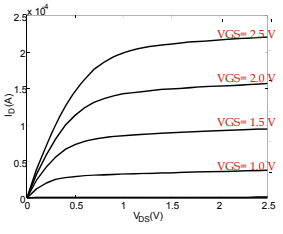


Short Channel

I_D versus V_{DS}

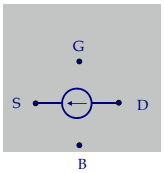


Long Channel



Short Channel

A unified model for manual analysis



$$I_D = 0 \text{ for } V_{GT} \le 0$$

$$I_D = k'\frac{W}{L}\left(V_{GT}V_{min} - \frac{V_{min}^2}{2}\right)(1 + \lambda V_{DS}) \text{ for } V_{GT} \ge 0$$

with $V_{min} = \min(V_{GT}, V_{DS}, V_{DSAT})$,

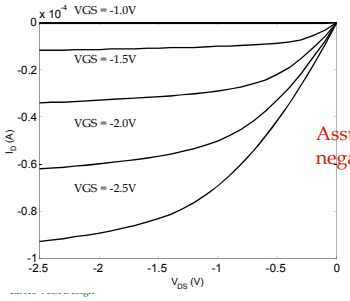
$V_{GT} = V_{GS} - V_T$,

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

	V_{T0}(V)	γ($V^{0.5}$)	V_{DSAT}(V)	k'(A/V^2)	λ(V^{-1})
NMOS	0.43	0.4	0.63	115 x 10^{-6}	0.06
PMOS	-0.4	-0.4	-1	-30 x 10^{-6}	-0.1

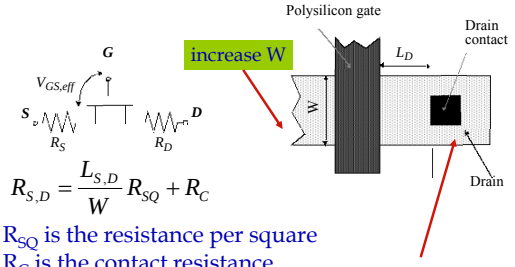
A PMOS Transistor

PMOS transistor, 0.25um, L_d = 0.25um, W/L = 1.5, V_{DD} = 2.5V, V_T = -0.4V



Assume all variables negative!

Parasitic Resistances



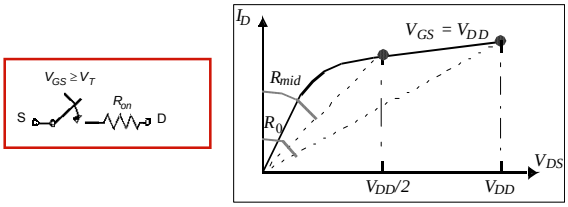
$$R_{S,D} = \frac{L_{S,D}}{W} R_{SQ} + R_C$$

R_{SQ} is the resistance per square

R_C is the contact resistance

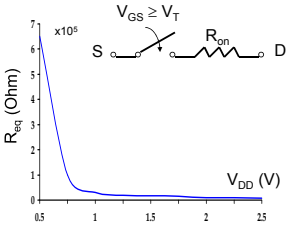
Silicide the bulk region

The Transistor as a Switch



$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}(1+\lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{5}{6}\lambda V_{DD}\right)$$

The Transistor as a Switch



- Resistance inversely proportional to W/L (doubling W halves R_{on})
- For $V_{DD} >> V_T + V_{DSAT}/2$, R_{on} independent of V_{DD}
- Once V_{DD} approaches V_T, R_{on} increases dramatically

V_{DD}(V)	1	1.5	2	2.5
NMOS(kΩ)	35	19	15	13
PMOS (kΩ)	115	55	38	31

(for $V_{GS} = V_{DD}$, $V_{DS} = V_{DD} \rightarrow V_{DD}/2$)

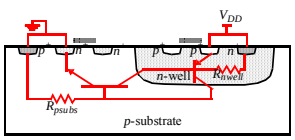
R_{on} (for W/L = 1)

For larger devices divide R_{eq} by W/L

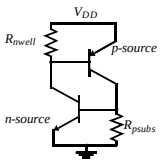
Summary of MOSFET Operating Regions

- Strong Inversion $V_{GS} > V_T$
  - » Linear (Resistive) $V_{DS} < V_{DSAT}$
  - » Saturated (Constant Current) $V_{DS} \geq V_{DSAT}$
- Weak Inversion (Sub-Threshold) $V_{GS} \leq V_T$
  - » Exponential in V_{GS} with linear V_{DS} dependence

Latchup

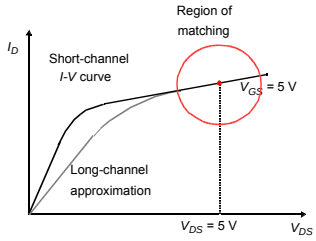


(a) Origin of latchup



(b) Equivalent circuit

Fitting level-1 model to short channel characteristics



Select k' and λ such that best matching is obtained @ $V_{gs} = V_{ds} = V_{DD}$

SPICE MODELS

Level 1: Long Channel Equations - Very Simple

Level 2: Physical Model - Includes Velocity Saturation and Threshold Variations

Level 3: Semi-Emperical - Based on curve fitting to measured devices

Level 4 (BSIM): Emperical - Simple and Popular

Berkeley Short-Channel IGFET Model

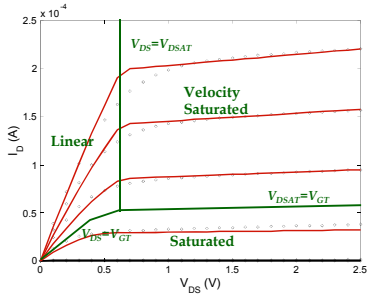
MAIN MOS SPICE PARAMETERS

Parameter Name	Symbol	SPICE Name	Units	Default Value
SPICE Model Index		LEVEL	-	1
Zero-Bias Threshold Voltage	VT0	VT0	V	0
Process Transconductance	k'	KP	A/V2	2.E-5
Body-Bias Parameter	g	GAMMA	V0.5	0
Channel Modulation	l	LAMBDA	1/V	0
Oxide Thickness	tox	TOX	m	1.0E-7
Lateral Diffusion	xd	LD	m	0
Metallurgical Junction Depth	xj	XJ	m	0
Surface Inversion Potential	2|f|	PHI	V	0.6
Substrate Doping	NA,ND	NSUB	cm-3	0
Surface State Density	Qss/q	NSS	cm-3	0
Fast Surface State Density		NFS	cm-3	0
Total Channel Charge Coefficient		NEFF	-	1
Type of Gate Material		TPG	-	1
Surface Mobility	m0	U0	cm2/V-sec	600
Maximum Drift Velocity	umax	VMAX	m/s	0
Mobility Critical Field	xcrit	UCRIT	V/cm	1.0E4
Critical Field Exponent in Mobility Degradation		UEXP	-	0
Transverse Field Exponent (mobility)		UTRA	-	0

SPICE Parameters for Parasitics

Parameter Name	Symbol	SPICE Name	Units	Default Value
Source resistance	R_S	RS	Ω	0
Drain resistance	R_D	RD	Ω	0
Sheet resistance (Source/Drain)	$R_\square$	RSH	Ω/□	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 Capacitance	C_{gb0}	CGBO	F/m	0
Gate-Source Overlap Capacitance	C_{gs0}	CGSO	F/m	0
Gate-Drain Overlap Capacitance	C_{gd0}	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

Simple Model versus SPICE



Technology Evolution

- Semiconductor Industry Association

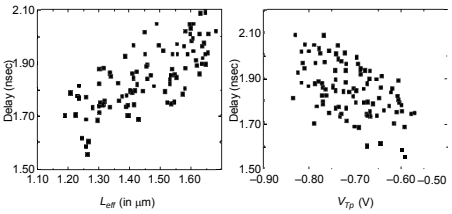
Year	2009	2012	2015	2018	2021
Feature size (nm)	34	24	17	12	8.4
L_{gate} (nm)	20	14	10	7	5
V_{DD} (V)	1.0	0.9	0.8	0.7	0.65
Billions of transistors/die	1.5	3.1	6.2	12.4	24.7
Wiring levels	12	12	13	14	15
Maximum power (W)	198	198	198	198	198
DRAM capacity (Gb)	2	4	8	16	32
Flash capacity (Gb)	16	32	64	128	256

Process Variations

Devices parameters vary between runs and even on the same die!

Variations in the process parameters , such as impurity concentration densities, oxide thicknesses, and diffusion depths. These are caused by non-uniform conditions during the deposition and/or the diffusion of the impurities. Introduces variations in the sheet resistances and transistor parameters such as the threshold voltage.

Variations in the dimensions of the devices, resulting from the limited resolution of the photolithographic process. This causes ***(W/L)*** variations in MOS transistors and mismatches in the emitter areas of bipolar devices.



Delay of Adder circuit as a function of variations in *L* and V_T

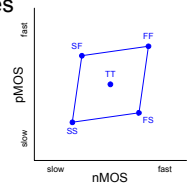
Impact of Device Variations

So What?

- So what if transistors are not ideal?
  - » They still behave like switches.
- But these effects matter for…
  - » Supply voltage choice
  - » Logical effort
  - » Quiescent power consumption
  - » Pass transistors
  - » Temperature of operation

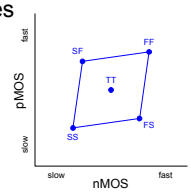
Parameter Variation

- Transistors have uncertainty in parameters
  - » Process: L_{eff}, V_t, t_{ox} of nMOS and pMOS
  - » Vary around typical (T) values
- Fast (F)
  - » L_{eff}: ____
  - » V_t: ____
  - » t_{ox}: ____
- Slow (S): opposite
- Not all parameters are independent for nMOS and pMOS



Parameter Variation

- Transistors have uncertainty in parameters
  - » Process: L_{eff}, V_t, t_{ox} of nMOS and pMOS
  - » Vary around typical (T) values
- Fast (F)
  - » L_{eff}: short
  - » V_t: low
  - » t_{ox}: thin
- Slow (S): opposite
- Not all parameters are independent for nMOS and pMOS



Environmental Variation

- V_{DD} and T also vary in time and space
- Fast:
  - » V_{DD}: ____
  - » T: ____

Corner	Voltage	Temperature
F		
T	1.8	70 C
S		

Environmental Variation

- V_{DD} and T also vary in time and space
- Fast:
  - » V_{DD}: high
  - » T: low

Corner	Voltage	Temperature
F	1.98	0 C
T	1.8	70 C
S	1.62	125 C

Process Corners

- Process corners describe worst case variations
  - » If a design works in all corners, it will probably work for any variation.
- Describe corner with four letters (T, F, S)
  - » nMOS speed
  - » pMOS speed
  - » Voltage
  - » Temperature

Important Corners

- Some critical simulation corners include

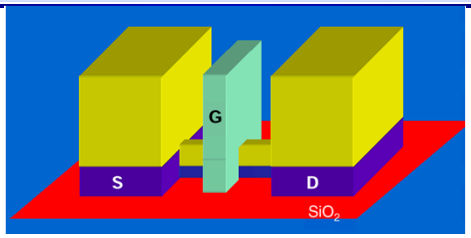
Purpose	nMOS	pMOS	V_{DD}	Temp
Cycle time				
Power				
Subthrehold leakage				

Important Corners

- Some critical simulation corners include

Purpose	nMOS	pMOS	V_{DD}	Temp
Cycle time	S	S	S	S
Power	F	F	F	F
Subthrehold leakage	F	F	F	S

Future Perspectives



25 nm FINFET MOS transistor

Three-Dimensional Integrated Circuits

- Multiple Layers of Active Devices
- Driven by
  - » Limited floorplanning choices
  - » Desire to integrate disparate technologies (GaAs, SOI, SiGe, BiCMOS)
  - » Desire to integrate disparate signals (analog, digital, RF)
  - » Interconnect bottleneck