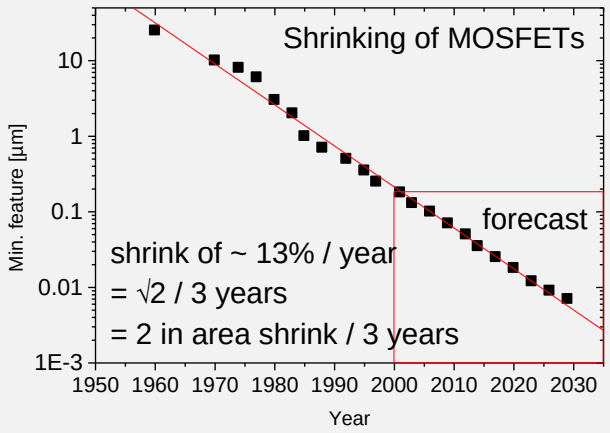
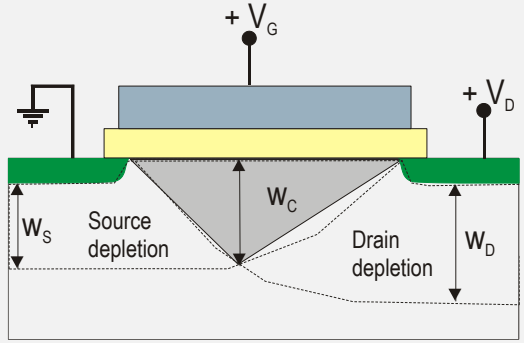
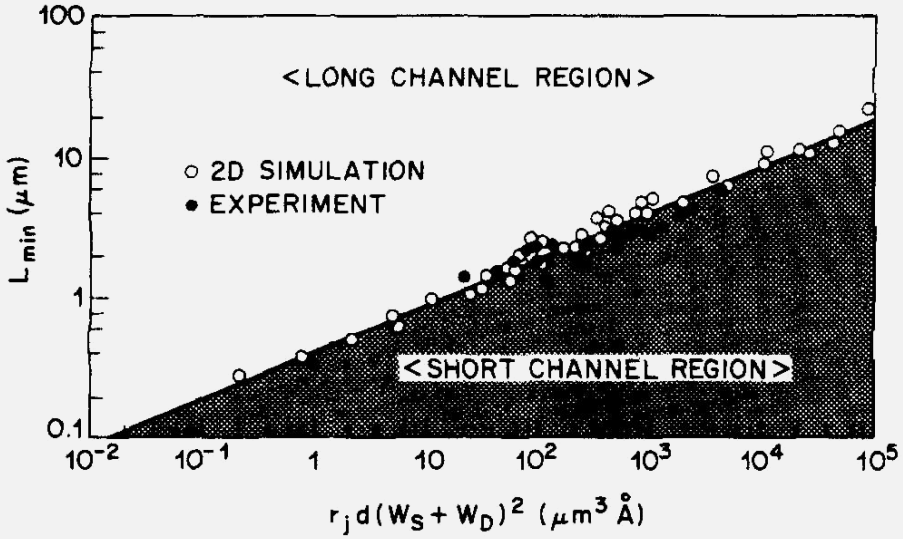
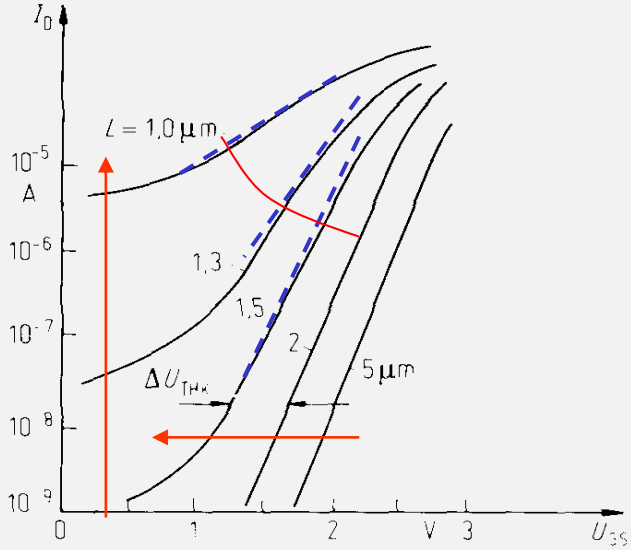

Shrinking of MOSFETs
forecast
shrink of ~ 13% / year
= √2 / 3 years
= 2 in area shrink / 3 years
Min. feature [µm]
Year
10
1
0.1
0.01
1E-3
1950 1960 1970 1980 1990 2000 2010 2020 2030

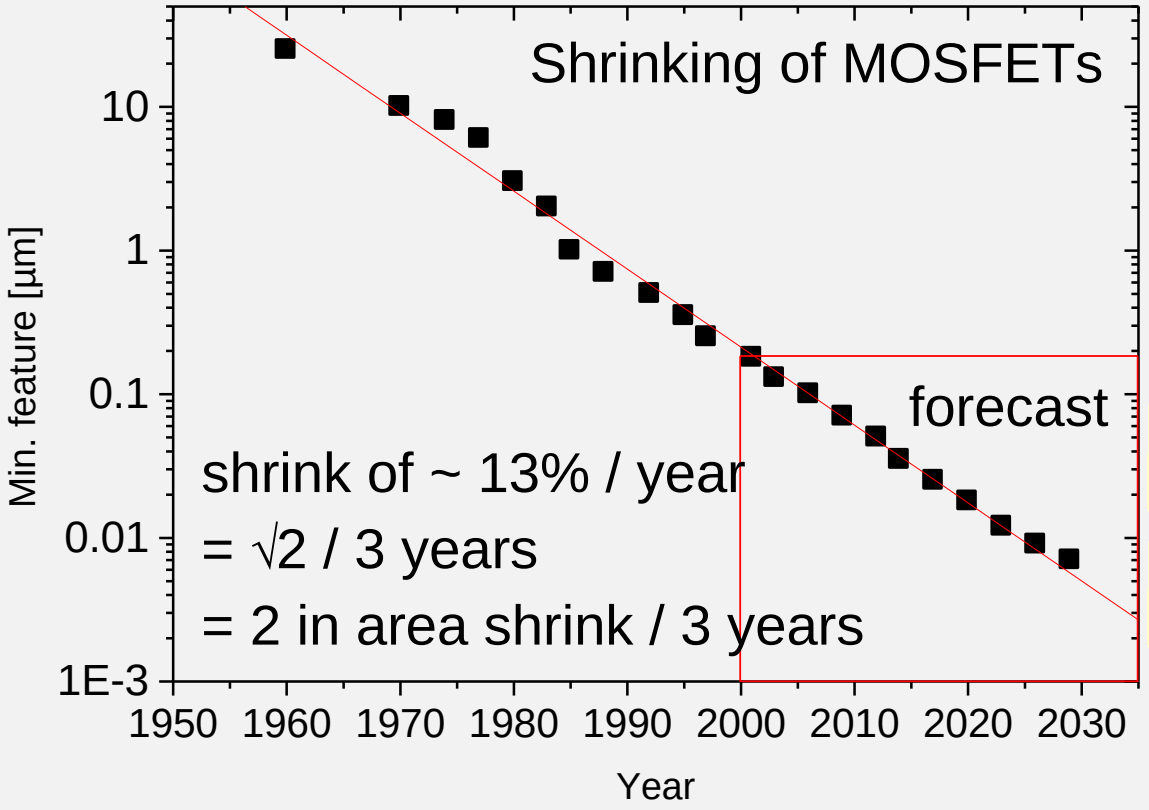

+ V_G
+ V_D
W_C
W_S
Source depletion
Drain depletion
W_D


<LONG CHANNEL REGION>
○ 2D SIMULATION
● EXPERIMENT
<SHORT CHANNEL REGION>
L_min (µm)
100
10
1
0.1
10^-2 10^-1 1 10 10^2 10^3 10^4 10^5
r_j d(W_S + W_D)^2 (µm^3 Å)


I_D
L = 1,0 µm
10^-5 A
10^-6
10^-7
10^-8
10^-9
1,3
1,5
2
5 µm
ΔU_TH
0 1 2 V 3 U_GS

Historical Trend

History:



1960: 25 µm

2000: 180 nm
2005: 90 nm
2008: 65 nm
2012: 22 nm
2010: 32 nm
2014: 14 nm
2016: 10 nm
2018: 7 nm
2020: 5 nm ?

Since the first MOSFET (1960) the feature size is shrinked about 13 % per year for 50 years now

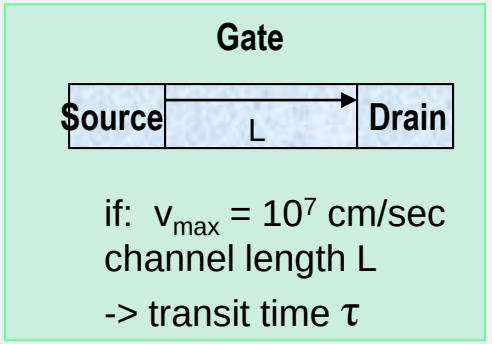
Why ?

Economics: A shrink of 15 % approximately reduces the fabrication costs by 50% / device

(see chap.2 and exercise 1)

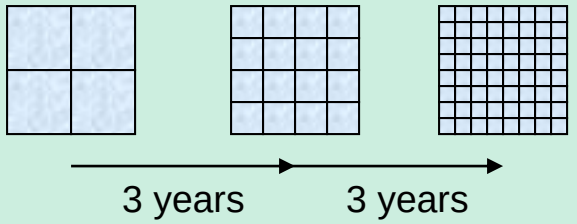
Shrinking - The Advantage

Discretes



if: $v_{max} = 10^7$ cm/sec
channel length L
-> transit time τ

Integration



every 3 years the number of devices on the same area increases by 4

Shrinking dimensions:

increase speed

$$\tau = \frac{L}{v} \longrightarrow \tau' = \frac{L/S}{v} \longrightarrow \tau' = \frac{\tau}{S}$$

reduce device area

$$A = w \cdot L \longrightarrow A' = w/S \cdot L/S \longrightarrow A' = \frac{A}{S^2}$$

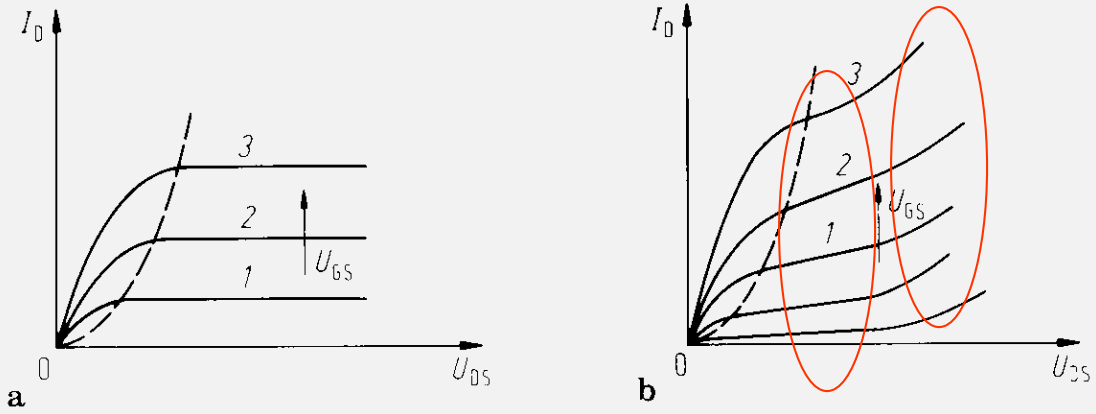
reduce energy consumption (power-delay)

$$P = U \cdot I \cdot \tau \longrightarrow P' = U/S \cdot I/S \cdot \tau/S \longrightarrow P' = \frac{P}{S^3}$$

Shrinking dimensions induces:

+ better dynamic behavior (higher speed, lower power)
+ better economics (more devices per wafer)

Shrinking - The Disadvantage

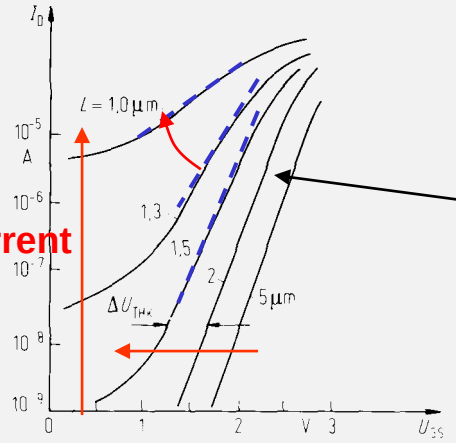


Output chracteristics:

* **no saturation** -> bad for analog application

* early, weak **break-through**

Sub-threshold behavior:



* increase in **subthreshold slope S**
(from ideal value 60mV/dec -> higher values)

-> higher voltage swing needed to turn device on/off

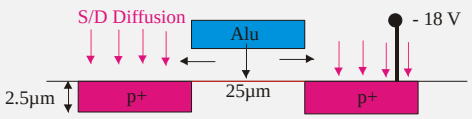
* increase of **leakage current**

Drain Induced Barrier Lowering (DIBL)

* **dynamic lowering of threshold voltage**

Shrinking dimensions induces: worse static behavior (higher leakage, lower reliability)

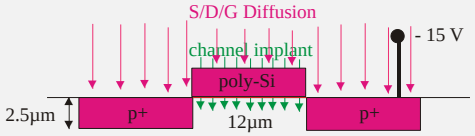
The Discovery of Short-Channel Effects



1960

1) hot processes first (Diffusion 1000°C)
2) cold processes (Alu 450°C)

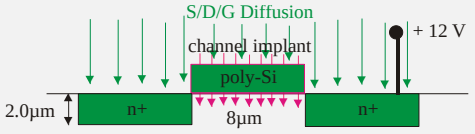
-> gate has to be adjusted -> large channels
-> diffusion -> large depth



1966

1) hot poly-gate first (Diffusion 1000°C)
2) channel implant (new!)

-> S/D self-adjusted -> short channels
-> diffusion -> large depth

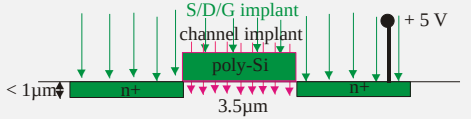


1974

4k-DRAM, CMOS-logic

1) short channel
2) large diffusion depth
3) high voltages

-> serious **„short-channel“ effects**



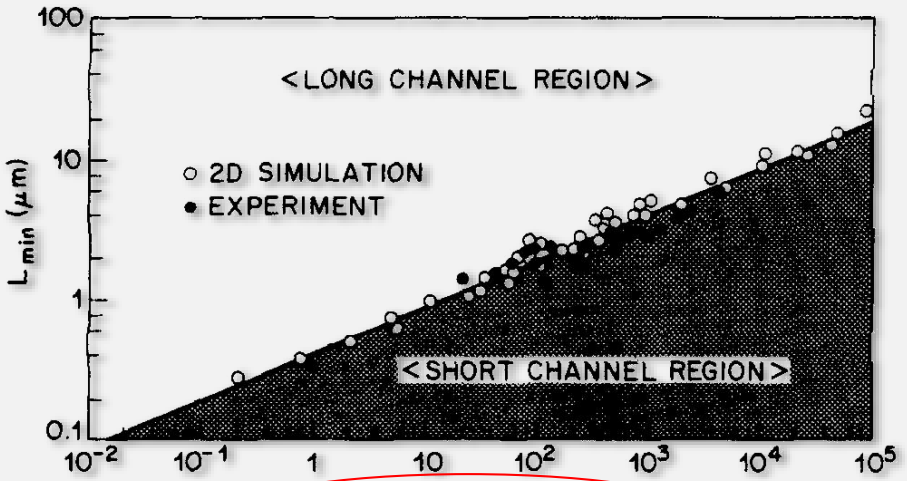
1978

INTEL HMOS I

1) S/D implant -> shallow junctions
2) low voltage -> reduce electric fields
3) e-beam-lithography for experimental short channel studies

discovery

study

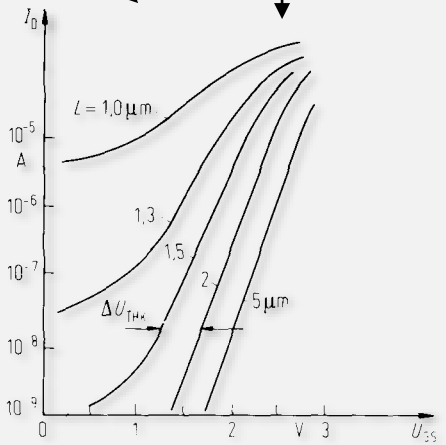


Brews, Fichtner, Nicollian, Sze: "Generalized Guide for MOSFET Miniaturization", IEEE Electron Device Letters, 1(1980)2-4

First **scaling** rules 1974

1980

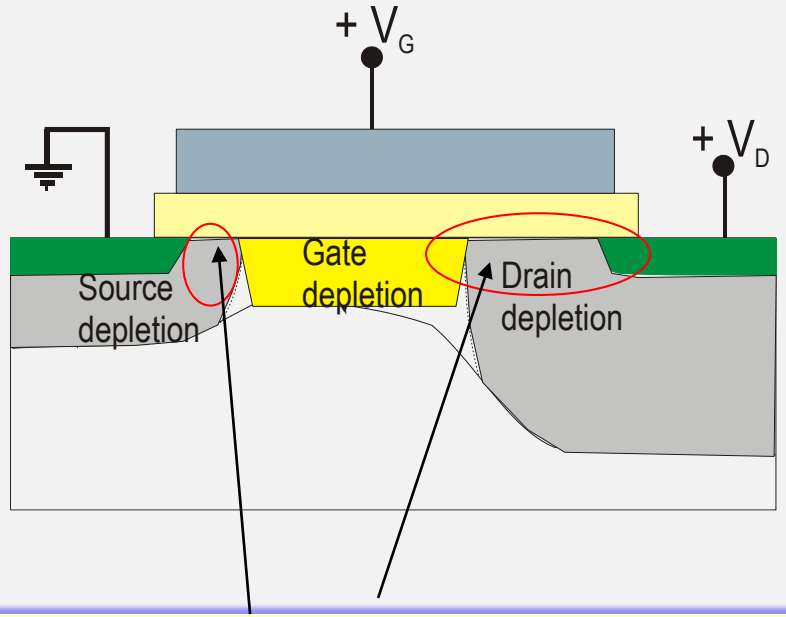
Empirical relationship



geometry- and voltage dependant parameters must be designed together !

Qualitative Look Inside the MOSFET

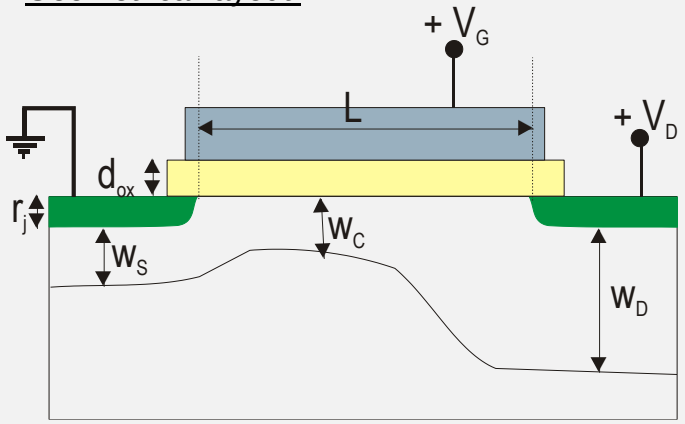
Result: **As early as in 1973 it was recognized by experimental devices, that:**



- **Short-channel effects appear, if the length of the Source/Drain depletion zone is no longer small compared to the channel length**
- **These, not gate-controlled, depletion charges change the gate-defined threshold voltage**
- **In addition, 2-dim or 3-dim field distributions must be handled**

Quantitative Look inside the MOSFET

Geometrical layout:



Geometrical parameters:

- L : **channel length** → limited by technology
- r_j : **junction depth of S/D doping** → defined by technology
- d_{ox} : **oxide thickness** → uncritical by technology

Voltage dependent parameters:

width of depletion layer
(one-sided abrupt junction)

$$w = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot(V_{bi}\pm V_{ext})}{qN_{dop}}}$$

built-in voltage:

$$V_{bi} = \frac{kT}{q}\cdot\ln\left[\frac{N_A\cdot N_D}{n_i^2}\right]$$

dependent on doping

Some calculations for historical MOSFETs: see page 25:

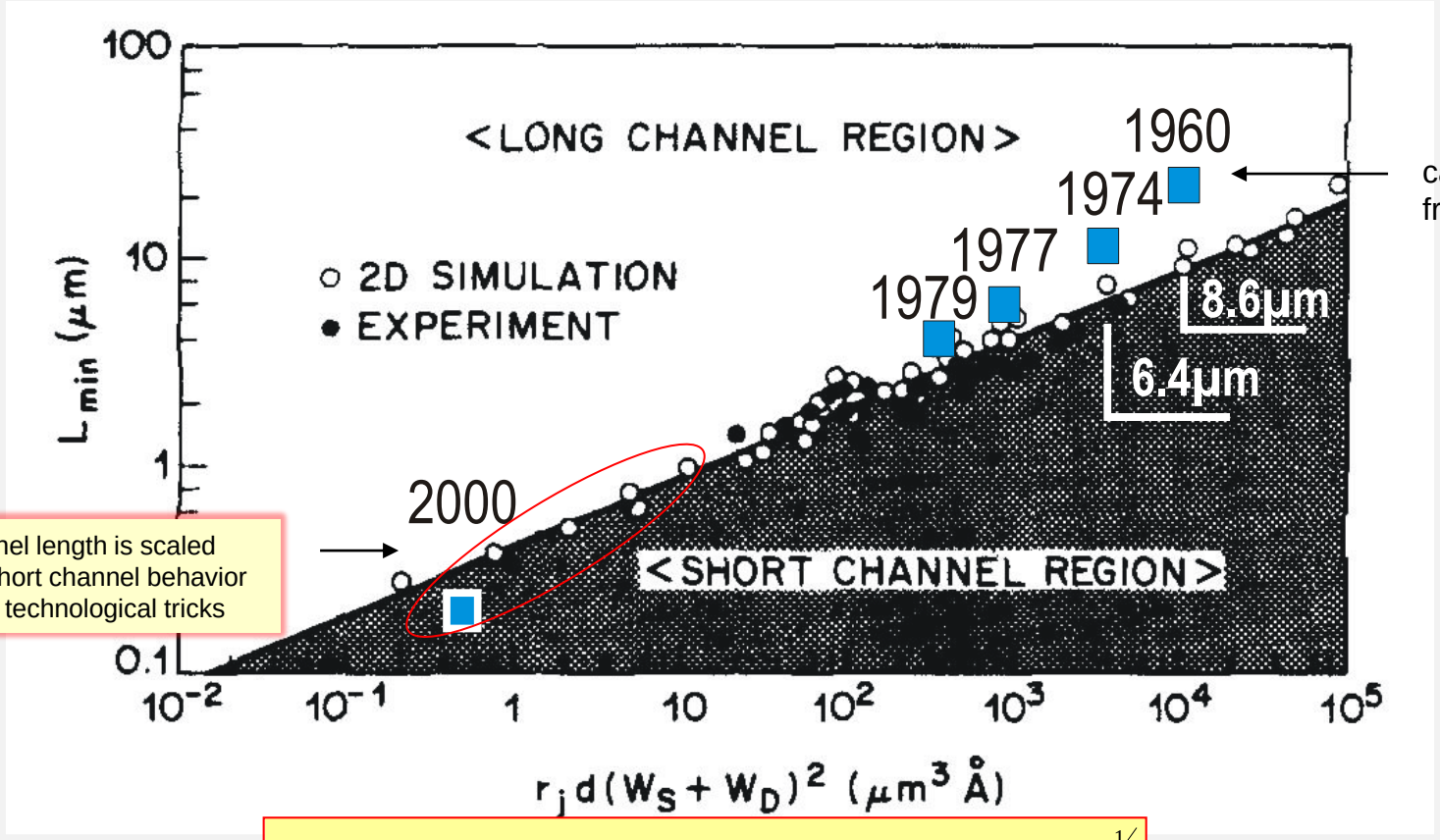
year	L µm	r_j µm	d_{ox} nm	V_{dd} V	w_S µm	w_D µm	$r_j\cdot d_{ox}\cdot(w_S+w_D)^2$	L_{min} µm
1960	25	2.5	200	18	0.26	1.1	9248	8.6
1974	8	2	120	12	0.26	1.0	3810	6.4
1977	6	1	100	5	0.26	0.6	740	3.7
1979	3.5	0.7	70	5	0.26	0.6	362	2.9
2000	0.15	0.04	4	1.5	0.26	0.3	0.6	0.34

w_s, w_D are calculated using: $N_A = 10^{16}$ cm^{-3}, $N_D = 10^{20}$ cm^{-3}

Empirical formula:

$$L_{min}[\mu m] = 0.41\cdot\left[r_j[\mu m]\cdot d_{ox}[\mathring{A}]\cdot(w_S[\mu m]+w_D[\mu m])^2\right]^{1/3}$$

Quantitative Look inside the MOSFET



calculated values from table

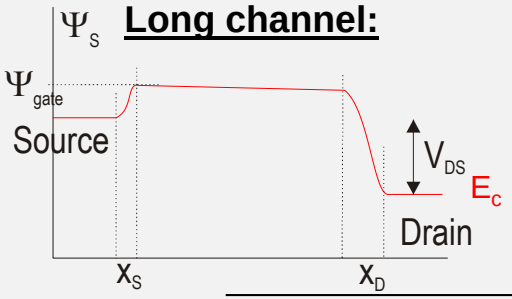
Since 1985 the channel length is scaled more aggressively, short channel behavior partly suppressed by technological tricks

$$L_{min}[\mu m] = 0.41 \cdot \left[r_j[\mu m] \cdot d_{ox}[\mathring{A}] \cdot (w_S[\mu m] + w_D[\mu m])^2 \right]^{1/3}$$

There exists no (electrical) lower limit for a short-channel device, if junction depth r_j, oxide thickness d_{ox} and doping is properly scaled down

Overview of Basic Short Channel Models

Long channel:



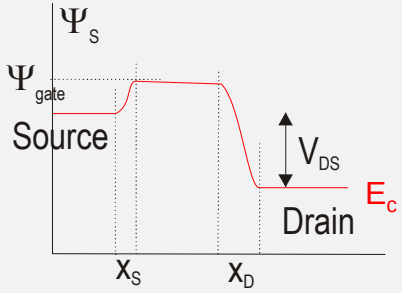
* the surface potential Ψ_S, determined by the gate voltage, is constant all over the channel

* the S/D depletion zones are negligible compared to the gate length

Gradual Channel Approximation

Short channel:

1 Charge sharing model:

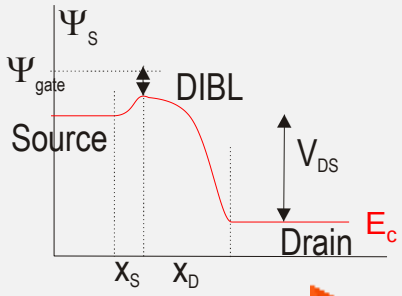


Assumption: the surface potential Ψ_S is still constant, but the depletion charge is shared by the Source/Drain depletion zones and the gate depletion

→ the threshold voltage is reduced due to already depleted Source/Drain depletion

This model is a good description, as long as a remarkable part of the channel is not occupied by the S/D

2 Drain-induced Barrier-Lowering model (DIBL):



Assumption: the surface potential Ψ_S varies along the channel due to the S/D potentials

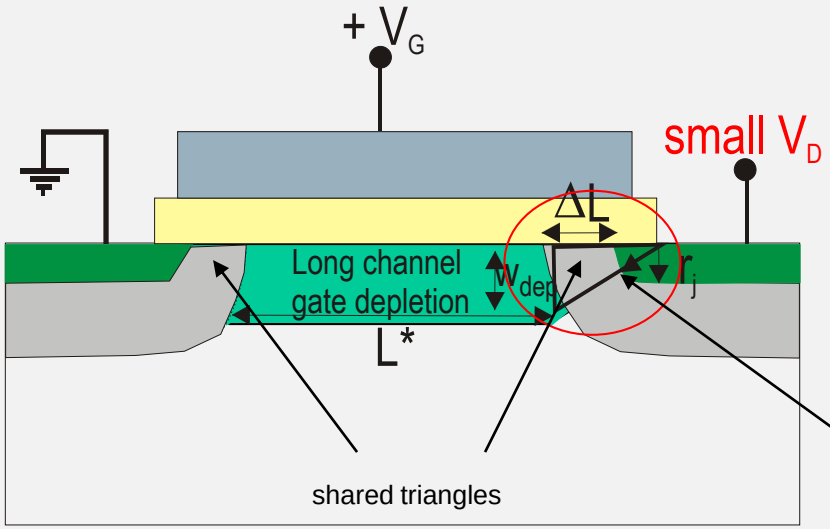
→ the threshold voltage is reduced due to a Source/Drain induced barrier lowering

This model is a good description, if the S/D depletion zones occupy a remarkable part of the channel

These analytical models result in equations, which may be adjusted for practical use

In contrast, with increasing influence of 2-dim or 3-dim effects, analytical models must be replaced by numerical simulations

The Charge Sharing Model (Yau-Model)



Yau: "A simple theory to predict the threshold voltage of short-channel IGFETs", Solid-State Electronics, 17(1974)1059-1063

Calculations for n-channel:

1) the charges in the gate capacitor equal to zero: $Q_{gate} + Q_{ox} + Q_{inv} + Q_{dep} = 0$

2) in voltages: $V_{gate} = V_{FB} + \Psi_S + \frac{Q_{dep}}{C''_{ox} \cdot A}$ with: $Q_{dep} = qN_A w_{dep} A$

3) with threshold voltage: $V_{th} = V_{FB} + 2\Psi_{bulk} + \frac{Q_{dep}}{C''_{ox} \cdot A}$ with: $\Psi_S = 2\Psi_{bulk}$

1

Short channel, but the contribution of S/D depletion is small:
(channel not too short and/or V_{DS} = zero/ small)

1) the gate controlled charge is approximated by a trapezoid:

Approach: $Q^*_{dep} = qN_A w_{dep} \cdot w_{gate} \cdot \left(\frac{L + L^*}{2}\right)$ w_{gate} = width of gate

2) the shared (S/D symmetrical) triangles can be calculated by trigometric: $w_{dep}^2 + (\Delta L + r_j)^2 = (w_{dep} + r_j)^2$ $\Delta L = \frac{L - L^*}{2}$

$$\Rightarrow \quad \frac{L + L^*}{2} = L - r_j \cdot \left(\sqrt{1 + \frac{2w_{dep}}{r_j}} - 1\right)$$

$$w_{dep} = \sqrt{\frac{2\varepsilon_0 \varepsilon_{Si} \cdot 2\Psi_{bulk}}{qN_A}}$$

3) the threshold voltage shift is then:

+ : p-channel
- : n-channel

$$\Delta V_{th} = V_{th}^{new} - V_{th}^{old} = V_{th}^{SC} - V_{th}^{LC} = \frac{qN_A w_{dep} \cdot \left(w_{gate} \cdot \frac{L + L^*}{2}\right)}{C_{gate}} - \frac{qN_A w_{dep} \cdot (w_{gate} \cdot L)}{C_{gate}} = \frac{qN_A w_{dep}}{C''_{gate}} \cdot \left(\frac{L + L^*}{2L} - 1\right) = \pm \frac{|q| N_A w_{dep} \cdot r_j}{LC''_{gate}} \cdot \left(\sqrt{1 + \frac{2w_{dep}}{r_j}} - 1\right)$$

the long channel threshold voltage decreases with shrinking channel length L, independent of V_{DS} at low V_{DS}

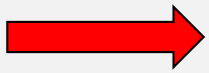
The Charge Sharing Model (Yau-Model)

Let us have a look to the empirical scaling law:

$$L_{\min} = 0.41 \cdot \left[r_j[\mu m] \cdot d_{ox}[\overset{o}{A}] \cdot \left(w_S[\mu m] + w_D[\mu m] \right)^2 \right]^{1/3}$$

From analytics of Charge-Sharing model:

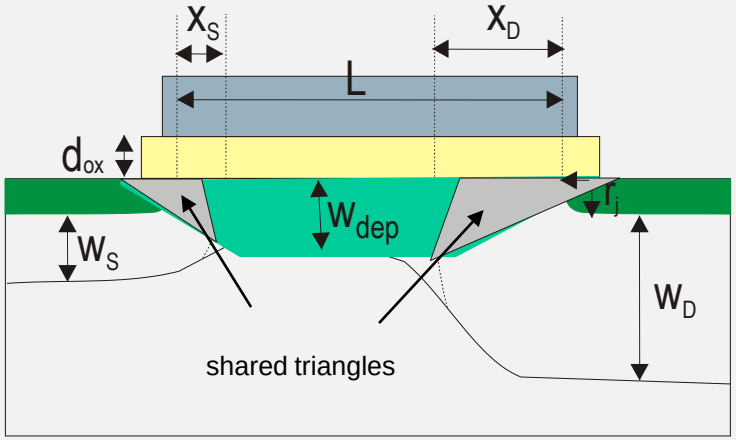
$$\Delta V_{th} \approx \pm \frac{|q| N_A w_{dep} \cdot r_j}{LC''_{gate}} \approx const. \cdot \frac{w_{dep} \cdot r_j \cdot t_{ox}}{L}$$

If we ask for: $\Delta V_{th} \leq 10\% \leq const. \cdot \frac{w_{dep} \cdot r_j \cdot t_{ox}}{L}$ 

$$L \geq const. \cdot w_{dep} \cdot r_j \cdot t_{ox}$$

which corresponds to the old scaling law

The Charge Sharing Model (Yau-Model)



2

Short channel, but the contribution of S/D depletion is high:
(very short channel and/or V_{DS} high)

1) with high V_{DS} the shared tringles increase to unsymmetrical triangles:

$$x_S = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot(V_{bi}-\Psi_S)}{qN_A}} \approx \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot(V_{bi}-2\Psi_{bulk})}{qN_A}}$$

at Source using: $\Psi_S = 2\,\Psi_{bulk}$

$$x_D = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot(V_{bi}-\Psi_S+V_{DS})}{qN_A}} \approx \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot(V_{bi}-2\Psi_{bulk}+V_{DS})}{qN_A}}$$

at Drain (reverse biased)

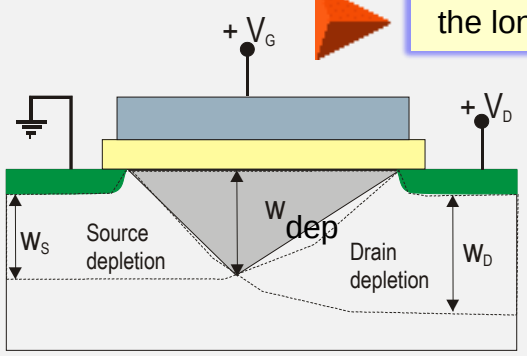
keeping $\Psi_S = 2\,\Psi_{bulk}$ approximately constant -> w_{dep}~const $\rightarrow$ $w_{dep} = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si}\cdot 2\Psi_{bulk}}{qN_A}}$

+ : p-channel
- : n-channel

2) the threshold voltage shift is then:

$$\Delta V_{th} = V_{th}^{new} - V_{th}^{old} = V_{th}^{SC} - V_{th}^{LC} = \pm\frac{|q|N_A w_{dep}\cdot r_j}{2LC''_{gate}}\cdot\left[\left(\sqrt{1+\frac{2x_S}{r_j}}-1\right)+\left(\sqrt{1+\frac{2x_D}{r_j}}-1\right)\right]$$

the long channel threshold voltage now decreases with shrinking channel length L **and** V_{DS} (via Ψ_S or x_D)



3

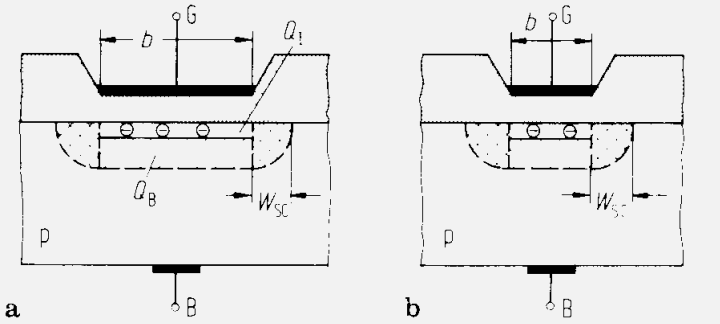
Short channel, the contribution of S/D depletion dominates:
(very short channel and/or V_{DS} high)

1) the gate controlled charge is assumed to reduce to a triangle:

$$w_{dep} = \frac{L}{2r_j}\cdot\left(r_j+\frac{L}{4}\right) \rightarrow$$ independent of V_{DS}

The Charge Sharing Model (Yau-Model)

Narrow channel



1) In reality the gate voltage also depletes the regions on both sides of the channel, which means an unintentionally increase in channel width

The gate charge, which in the long channel case, is enough to induce inversion, is no longer sufficient, because a part of this gate charge depletes the channel sides.

As a result, a higher gate charge is needed for inversion, this corresponds to higher threshold voltage

2) Assuming, that the additional depletion is approximately cylindrical, the total depletion charge is:

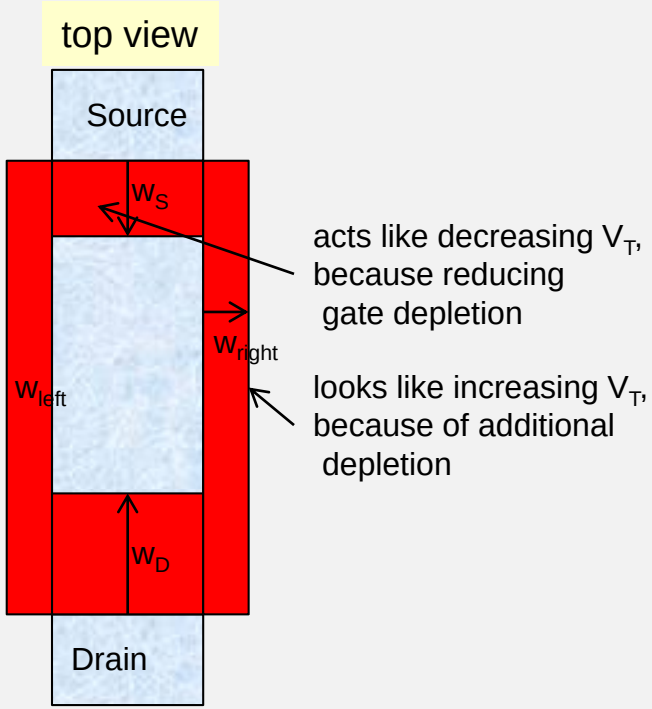
$$Q^*_{dep} = qN_A w_{dep} \cdot w_{gate} \cdot \left(L + \frac{\pi}{2} \cdot \frac{w_{dep}}{w_{gate}} \right)$$

3) In consequence, the threshold voltage is increased to:

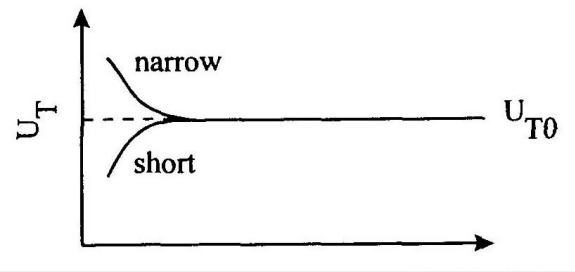
$$V_{th} = V_{FB} + 2\Psi_{bulk} + \frac{\sqrt{2\varepsilon_0 \varepsilon_{Si} q N_{dop} \cdot 2\Psi_{bulk}}}{C''_{ox}} \cdot \left(1 + \frac{\pi}{2} \cdot \frac{w_{dep}}{w_{gate}} \right)$$

4) In reality, the metal gate overlaps the sloped regions:

-> analytical corrections must be done
-> or 2-dim/ 3-dim numerical solutions of the Poisson equation (simulations)

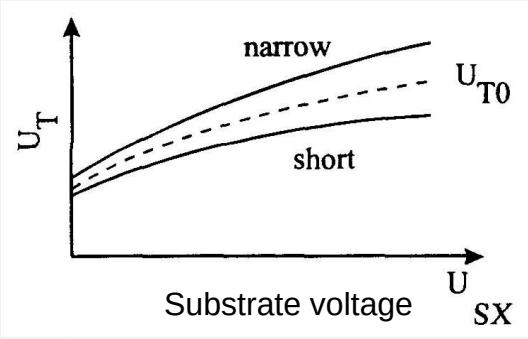


Effect of Short-Channel on Threshold and Current



geometry: L or w

Threshold Voltage



Substrate voltage

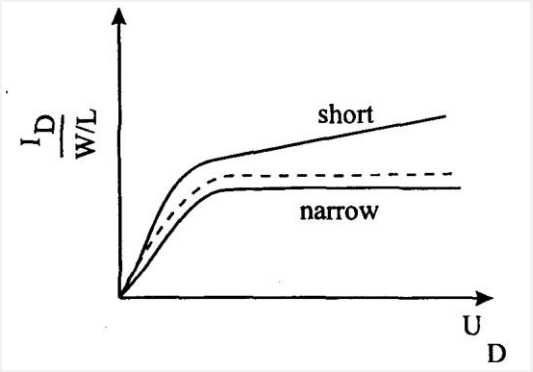
- the opposite behavior of threshold voltage shift can not be used for compensation, since both effects are nonlinear connected
- in addition, the channel length is fixed and defined by technology, the gate width is needed to be flexibel in circuit design

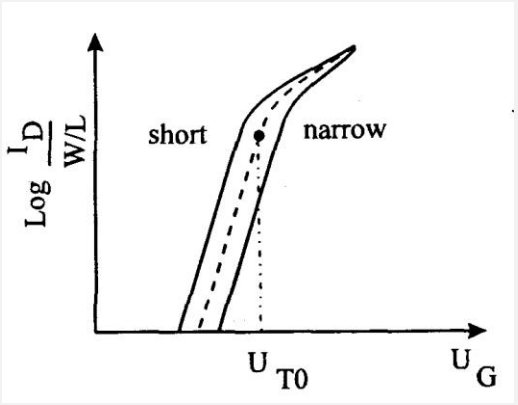
I-V characteristics:

+ : p-channel
- : n-channel

- : p-channel
+ : n-channel

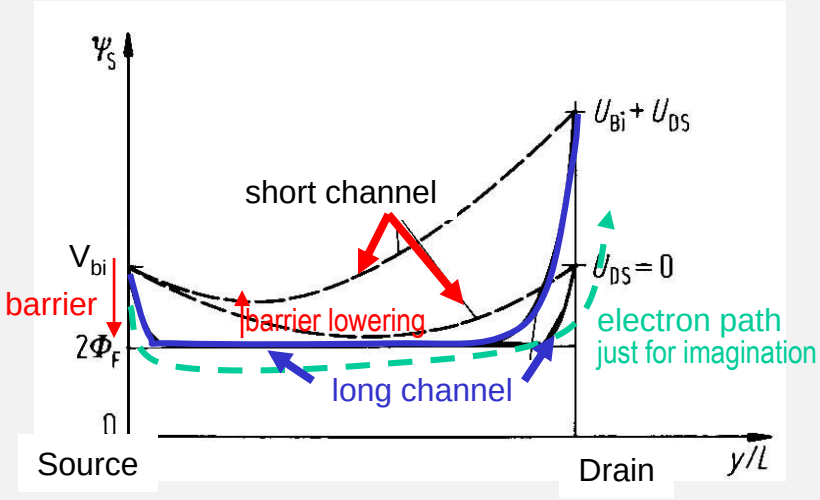
Since short channel and narrow channel effect can not be separated, the simple approach: $V_{th}^* = V_{thLC} \pm |\Delta V_{thSC}| \pm |\Delta V_{thNC}|$ can not be used in the IV-characteristics:





- in practice, complicated analytical models must be assumed to calculate $V_{th}^* = f(N, L, w, r_j, V_{DS})$ or 3-dim simulations

The Drain-Induced Barrier-Lowering (DIBL) Model



Long channel:

Channel formation (strong inversion) starts when $\Psi_S = 2\,\Psi_{bulk}$

- Ψ_S is constant all over the channel
- Ψ_S determines the value of threshold voltage V_{th}

$$V_{th} = V_{FB} + 2\Psi_{bulk} + \frac{\sqrt{2\varepsilon_0\varepsilon_{Si} q N_{dop} \cdot 2\Psi_{bulk}}}{C_{ox}''}$$

Short channel:

- a symmetrical barrier lowering exists
- with additional V_{DS} the barrier maximum is shifted to Source and in addition further lowered.

Due to empirical observations the lowering is approximately linear in V_{DS}:

Approach: $$V_{th,SC} = V_{th,LC} - A_{pene} \cdot V_{DS} - B_{neigh}$$

where A_{pene} is a so-called **penetration constant**, how effective the V_{SD} penetrates into the channel
and B_{neigh} a so-called **neighbouring-effect** B = f(L, x_j, d_{ox}, N_{dop}, ...)

For A_{pene} several approaches exist:

$$A_{pene} = \frac{6 d_{ox}}{w_{dep}} \cdot \exp\left[-\frac{n \cdot L}{4 w_{dep}}\right]$$

$A_{pene} = 1/L$

$A_{pene} = L^{-3}$ as it is used in SPICE

the long channel threshold voltage decreases with shrinking channel length L and linear decrease in V_{DS}

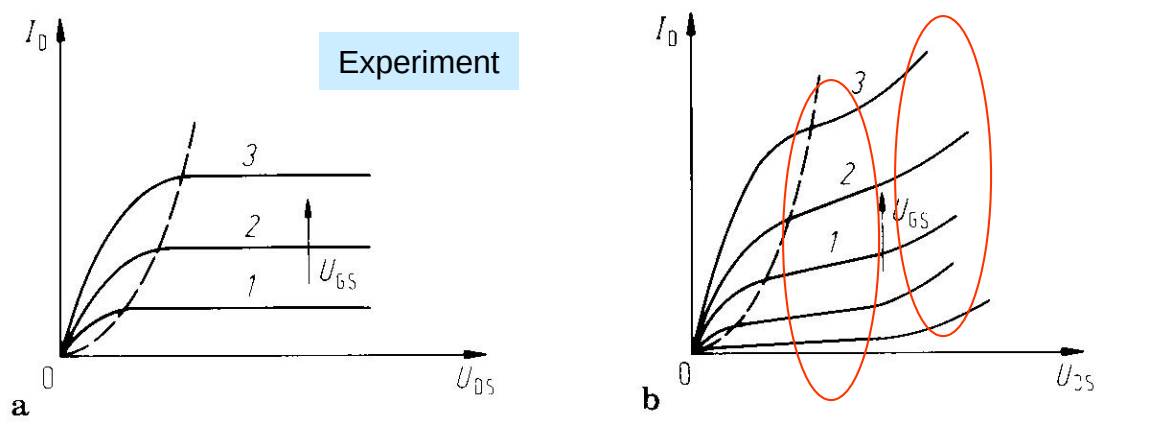
Characterization of the DIBL-model:

disadvantage: the physical description is nearly lost, mainly empirical trends are used, which are fitted to (a special ?) reality

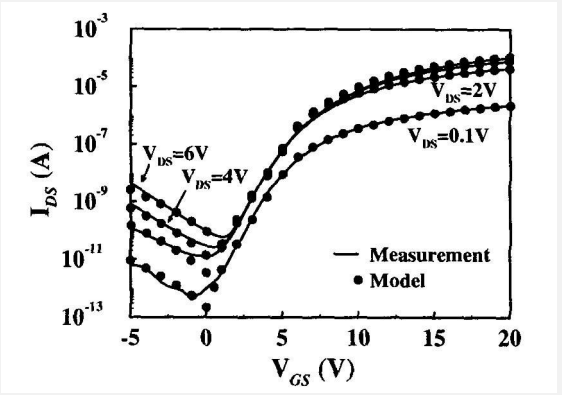
advantage: In principle the 2/3-dim Poisson-equation must be solved. Because solving Poisson´s equation is done in any numerical simulation the DIBL-model is very powerful when strong inhomogenities (e.g. channel doping) exists

Comparison of Models vs Experiment

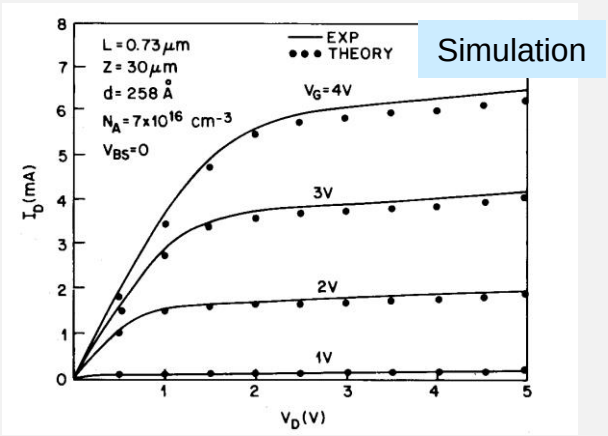
Output characteristics:

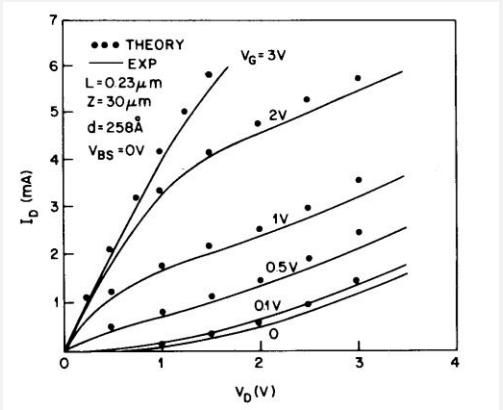


Transfer characteristics:

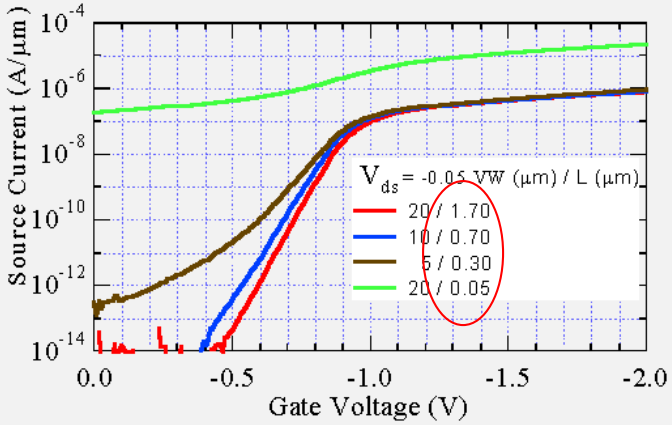


a) Long-channel, L = 0.73µm → b) Short-channel, L = 0.23µm



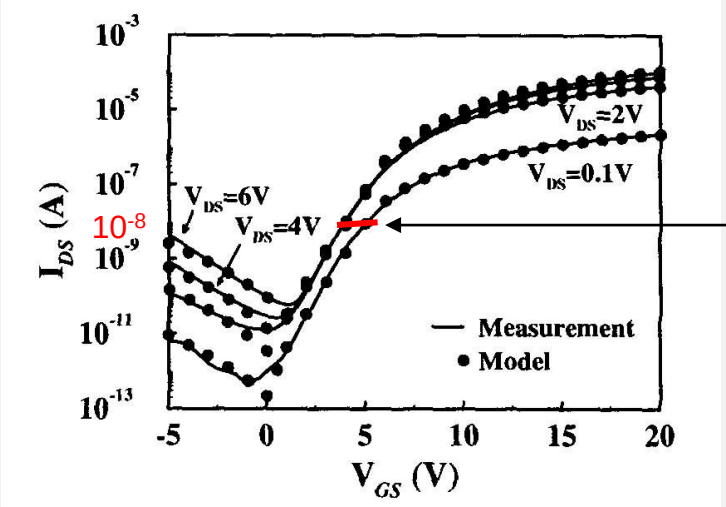


weak inversion = subthreshold



Comparison of experimental MOSFETs and numerical simulations show good agreement

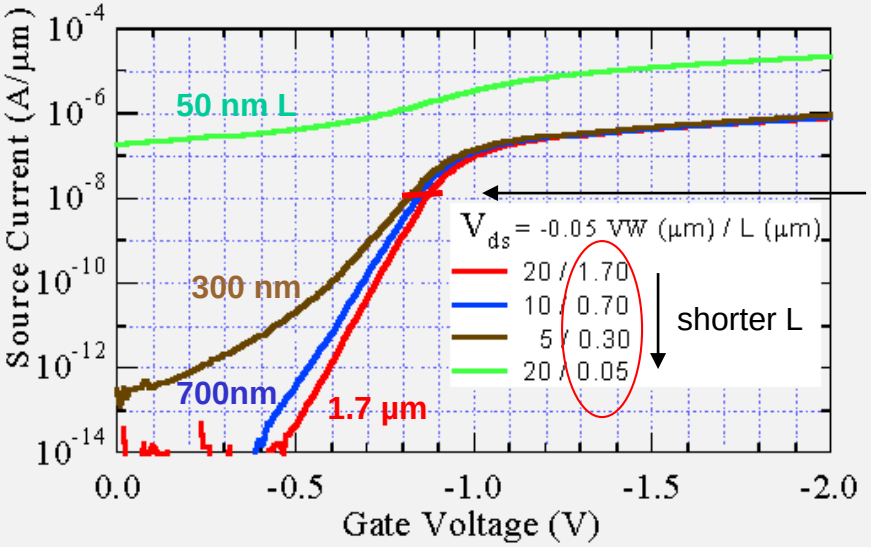
The DIBL: - Measurements



The DIBL is characterized as a threshold voltage shift

1) In a single device commonly the gate voltage shift at a given drain current I_D (mostly at ~10^{-8} A/µm, which was old industry way for fast determing V_T) is used between the characteristics of 0.1V (low field) and the supply voltage (or sometimes V_{DS}=1V)

This effect is called: **DIBL**

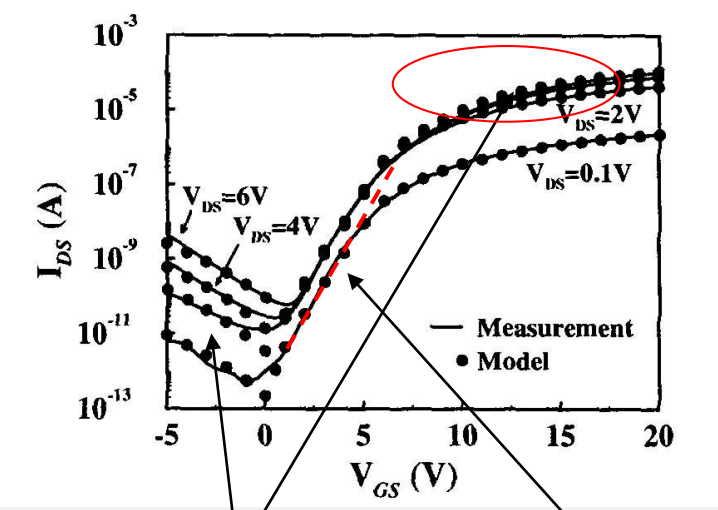


2) The design of processes and geometries for shrinking MOSFETs consumes a lot of money. Industry prefers devices, where not much has to be changed with shrinking.
As a test of robustness of technology for future shrinking, MOSFETs are fabricated with constant technology, but shorter channels.
Now, the gate voltage shift at a given drain current I_D [~10^{-8} A/µm] for various channel length´s is used.

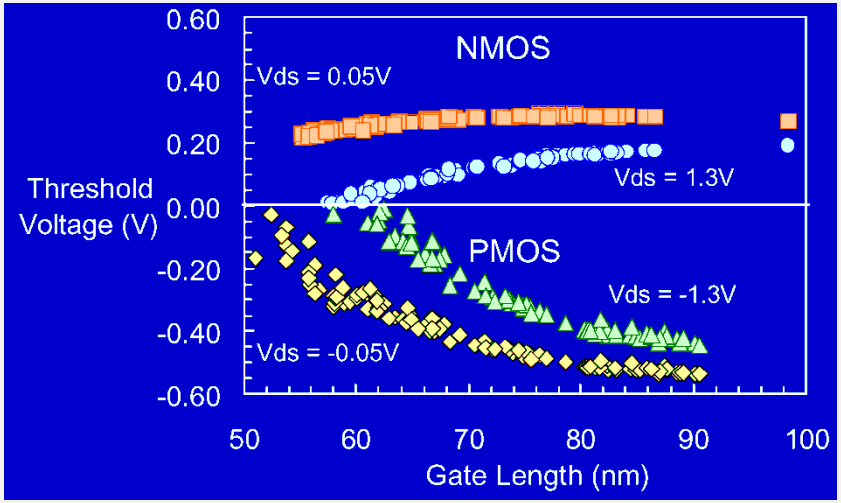
This effect is called: **V_T roll-off**

The DIBL: - Measurements

Example of DIBL
(measured in one single device at different voltages)



Example of V_T Roll-off
(measured in different devices at same voltages)



Same device, same technology, but just shorten the channel

1) the I_{on} - current remains the same, because the internal barrier vanishes at high gate voltage anyhow (if there is V_{DS} barrier lowering (DIBL) or not)

2) but if DIBL exist, the threshold voltage decreases and therefore leakage current I_{off} increases.
3) if I_{off} increases and I_{on} remains the same the subthreshold swing S must increase in addition

Summary of High-Field Effects

In addition to the small geometry effects all **high-field effects** have a serious impact to short channel MOSFETs:

High Electric Fields induce:

- Modification in carrier transport
  - Mobility degradation due to Gate field
  - Velocity saturation due to Source-drain field
- 2-dim field effects
  - Channel length modulation in saturation
- Break-through effects
  - 1 Hot electrons -> Impact ionization
    - Avalanche multiplication (Drain break-through)
    - Parasitic Bipolar Transistor
    - Gate injection
  - 2 Gate break-through
  - 3 Punch-through
- Parasitics
  - Source-Drain resistors

High electric fields modify (degrade) the MOSFET performance

Scaling

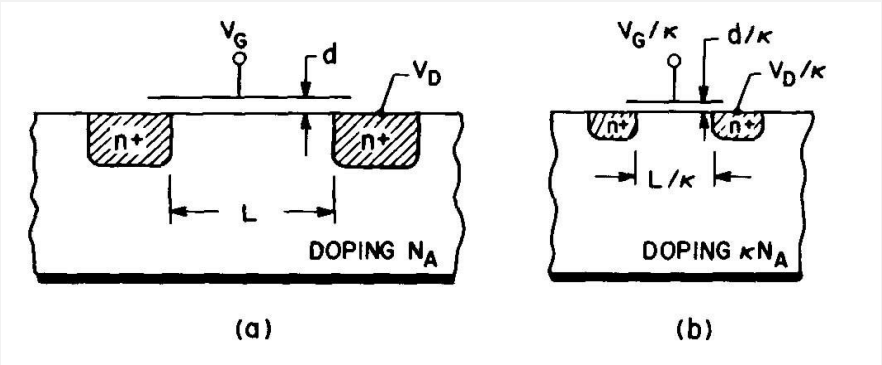
To avoid short-channel effects (2/3-dim geometry, hot electrons (high-field) und transport degradations) **scaling laws** are investigated

The aim of scaling is to reduce the economically needed device size
and
keep the desired electrical advantages (saturation, power reduction, speed increase)
without the electrical disadvantages (degradation, hot carrier effects, reliability reduction)

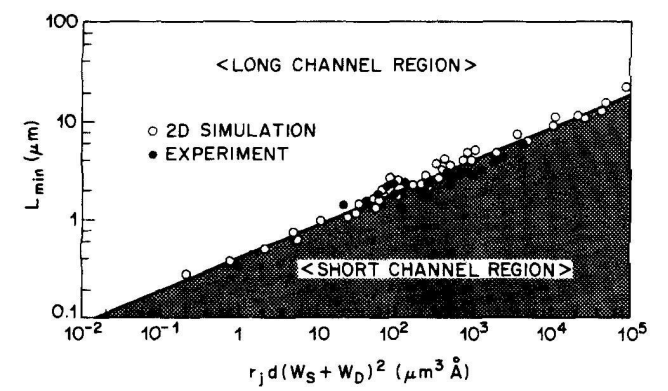
Two basic scaling approaches exist

Electrostatic scaling
(based on physics, advantage for analog application, but practical limits)

- ideal (constant field)
- constant voltage
- quasi-constant voltage
- free scaling



Subthreshold scaling
(empirical, focus for digital ULSI-circuits (leakage, switching))



Electrostatic Scaling

1 Constant-Field Scaling

	Scaling	expression	limited by
scaling			
lateral dimensions channel length L	1 / S		lithography
vertical dimensions Oxide thickness, junction depth	1 / S		breakdown, tunneling
voltage V	1 / S		
Doping N_A	* S	to achieve reduction of depletion regions * compromising $1/\sqrt{S}$ without voltage * compromising $1/S$ with high voltage	
consequences of scaling			
area	$1/S^2$		
area gate capacity C´´	*S	$C'' = \frac{C''}{d/S}$	
drain current	1 / S	$I_{DS} \propto S/C'' \cdot V_G/S \cdot V_D/S$	
switching time	1 / S		
power loss	$1 / S^2$		
power-delay-product	$1 / S^3$		
current density	*S		
interconnection resistance	*S	$R \propto \frac{l}{d \cdot b}$	increases !
voltage drop	not scaling	$\Delta V = I \cdot R$	
RC-time	not scaling	$\tau = RC$	
contact resistance	*S	$\Delta V_{cont} = \frac{\rho}{A} \cdot I$	increases !
not influenced by scaling:			
work function difference Φ_{MS}	thermodynamic	used in Vth	
subthreshold swing S			same voltage swing for ion/ioff
not linear scaling:			
bulk Fermi-level, inversion-onset		$\Psi_{bulk} = f(\ln \frac{N_{dop}}{ni})$	
width of depletion zone w	ideal should be: $1/\sqrt{S}$	$w_{dep} = \sqrt{\frac{2\varepsilon_0\varepsilon_{Si} \cdot \left(V_{bi}\left(\ln N_{dop}\right) \pm V_{ext}/S\right)}{qSN_A}}$	
threshold voltage		$V_{th} = \Phi_{MS} - \frac{Q_{ox}}{C_{ox}} + 2\Psi_b(\ln(NA)) + \frac{\sqrt{2\varepsilon_0\varepsilon_{Si}qSN_A \cdot 2\Psi_b(\ln(NA))}}{S \cdot C_{ox}}$	
subthreshold current		$I_{D,sub} \approx I_0\left(\sqrt{N_A}, \ln(N_A)\right) \cdot \exp\left[V_{GS}/S\right]$	

* shrinking all geometrical dimensions and voltages by the same factor S keeps the electrical fields in the MOSFET ~ constant

* because depletion width is scaling with $\sqrt{V'/N'_{dop}} -> \sqrt{\frac{V/S}{N*S}} = \frac{1}{S} \cdot \sqrt{V/N_{dop}}$

the doping N is scaled by S*N to reduce depletion width by w/S

Advantages in dynamic behavior
increasing switching speed
consuming less power

But increasing parasitic effects

Limits by not scaling parameters

Electrostatic Scaling

Due to several restrictions in the past (and future), e.g. keeping voltage levels constant or technological capabilities the „ideal“ Constant-Field Scaling was modified by keeping Poisson´s equation constant

With the **Poisson equation** the electrical potentials in a semiconductor device are calculated:

$$\frac{d^2\varphi}{dx^2} = -\frac{\rho(x)}{\varepsilon_0\varepsilon_{Si}} \approx -\frac{qN_{dop}}{\varepsilon_0\varepsilon_{Si}}$$

Poisson equation is invariant when transforming: Geometries: x*= x/a with a factor a

potentials: φ*= φ /b with a factor b

doping: N*= N(a²/b) with a factor a²/b

$$\longrightarrow \quad \frac{d^2\varphi^*}{dx^{*2}} = \frac{d^2\varphi/b}{d\left(x/a\right)^2} = \frac{a^2}{b}\cdot\frac{d^2\varphi}{dx^2} = -\frac{a^2}{b}\cdot\frac{qN_{dop}}{\varepsilon_0\varepsilon_{Si}} = -\frac{qN^*_{dop}}{\varepsilon_0\varepsilon_{Si}}$$

Choosing special values of a, b results in special scaling models:

a = b =: S, -> a²/b = S -> E=const, **Constant-Field Scaling**
a = S, b = const, -> **Constant-Voltage Scaling**
a > b, **Quasi-Constant-Voltage Scaling**
a, b **General Scaling**

Electrostatic Scaling

	Constant-Field Scaling: a = b = S	Constant-Voltage Scaling: 1 < b < a but V = const	Quasi-Constant Voltage Scaling: 1 < b < a	free Scaling 1 < b < a
scaling				
lateral dimensions channel length L	1 / S	1/a	1/a	1/a
vertical dimensions Oxide thickness, junction depth	1 / S	1/b	1/a	1/a
voltage V	1 / S	const=1	1/b	1/b
Doping NA	* S	* a	*a	a^2/b
consequences of scaling				
area	$1/S^2$	a^2		
drain current	1 / S	b	a/b^2	a/b^2
switching time	1 / S	$1/a^2$	a/b^2	a/b^2
power loss	$1 / S^2$	b	a/b^3	a/b^3
power-delay-product	$1 / S^3$	a/b^2	a^2/b^5	a^2/b^5
interconnection resistance	*S	*a	*a	*a
voltage drop	not scaling	not scaling	not scaling	not scaling
RC-time	not scaling	not scaling	not scaling	not scaling
contact resistance	$*S^2$	a^2	a^2	a^2
not influenced by scaling:				
work function difference Φ_{MS}	thermodynamic			
subthreshold swing S				same ion/ioff-voltage
not linear scaling:				
bulk Fermi-level, inversion-onset				
width of depletion zone w				
threshold voltage				
subthreshold current				

2 Constant-Voltage Scaling

* was used to keep the voltage level (e.g. TTL) constant for circuits
* to avoid high gate fields the vertical dimensions are less reduced (b < a)

3 Quasi-Constant-Voltage Scaling

* the voltage is also reduced by b
  -> depletion width scales less
  -> electric fields increase less

4 Free Scaling

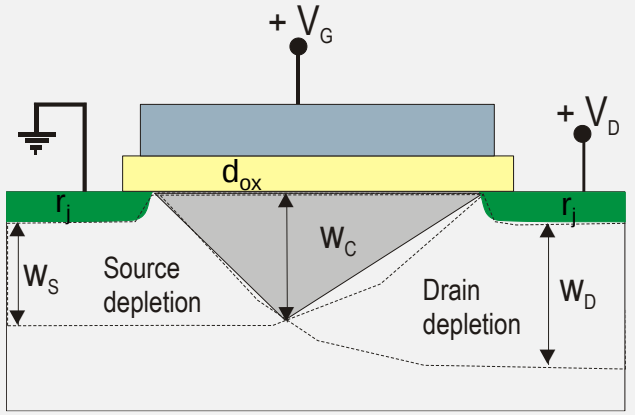
* full use of Poisson-scaling
  -> if special parameters are limited (e.g. oxide thickness) scaling can be done nevertheless

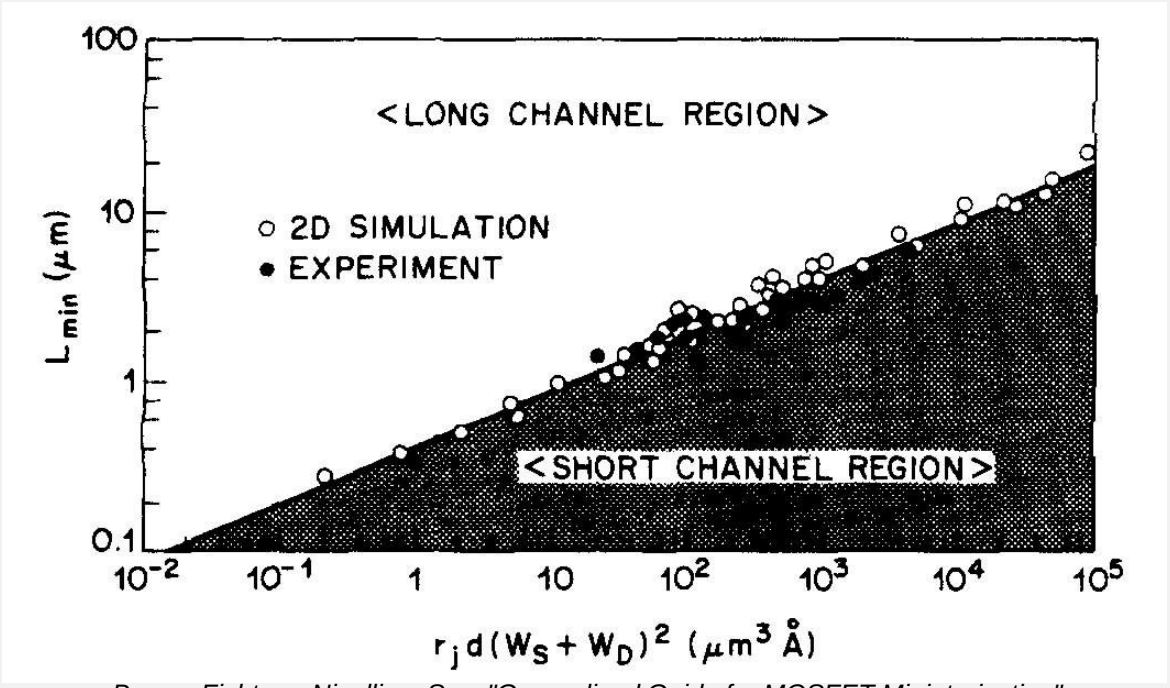
Subthreshold Scaling

Subthreshold scaling:

It was found emirically (1980), that the short-channel behavior can be avoided and the subthreshold characteristics is not degraded, if the following relationship is fulfilled:

$$L_{\min} = 0.41 \cdot \left[r_j[\mu m] \cdot d_{ox}[\overset{o}{A}] \cdot \left(w_S[\mu m] + w_D[\mu m] \right)^2 \right]^{1/3}$$





Brews, Fichtner, Nicollian, Sze: "Generalized Guide for MOSFET Miniaturization", IEEE Electron Device Letters, 1(1980)2-4

Most flexible scaling, because only the product of parameters must be scaled

MOSFET Development - Overview

Scaling relevant values of parameters in MOSFET generations:

DRAM-complexity	1k	4k	16k	64k	256k	1M	4M	16M	64M	256M	1G	1G *	4G	16G	64G	256G	Scaling S
Year	1970/71	1973	1976/77	1979/80	1982	1985	1988	1993	1995	1997	1999	2001	2003	2006	2009	2012	All 3 years
Process	p-chan Si-gate	n-chan Si-gate	n-chan 2-poly	n-chanl 2-poly						forecast: ITRS-roadmap 1997							
DRAM cell structure	3 T	1 T	1 T														
Min.Feature [µm] (channel length)	8 - 10	7 - 8	6-7 3-4	3.5 2-2.5	2.0 1.5	1.0 0.8	0.7 0.5	0.5 0.3	0.35	0.25	0.18	150	130	100	70	50	S=1/√2=0.7
junction depth [nm]	2500	2000	1000	700	450	250	150	150		100-50	72-36	60-30	52-26	40-20	30-15	20-10	
gate oxide thick [nm]	120	120	100	70, 70	45	30-20	15-10	15-10		5-4	4-3	3-2	3-2	2-1.5	< 1.5	< 1.0	
gate width µm			5-7	3-4	2	1.2	0.5	0.5									
interconnect width µm			5	4	3	2	1	1									
interconnect thick µm			1	1	0.9	0.75	0.6	0.6									
cell area [µm²]	3700	900	450	170		30	10	3.9	1.5	0.56	0.22	0.14	0.09	0.036	0.014	0.006	S²/1.3
Chip size [mm²]	13	19	18	21	38	55	80	145	190	280	400	445	560	790	1120	1580	1 / 1.5
Udd [V]	18/12	12	12,5	5, 5	5	5	5-3.3	3.3		2.5- 1.8	1.8-1.5	1.5	1.2	1.2-0.9	0.9-0.6	0.6-0.5	
Power dissipation [W]				0.2						70	90	110	130	160	170	175	
comments			HMOS	HMOS	HMOS II												

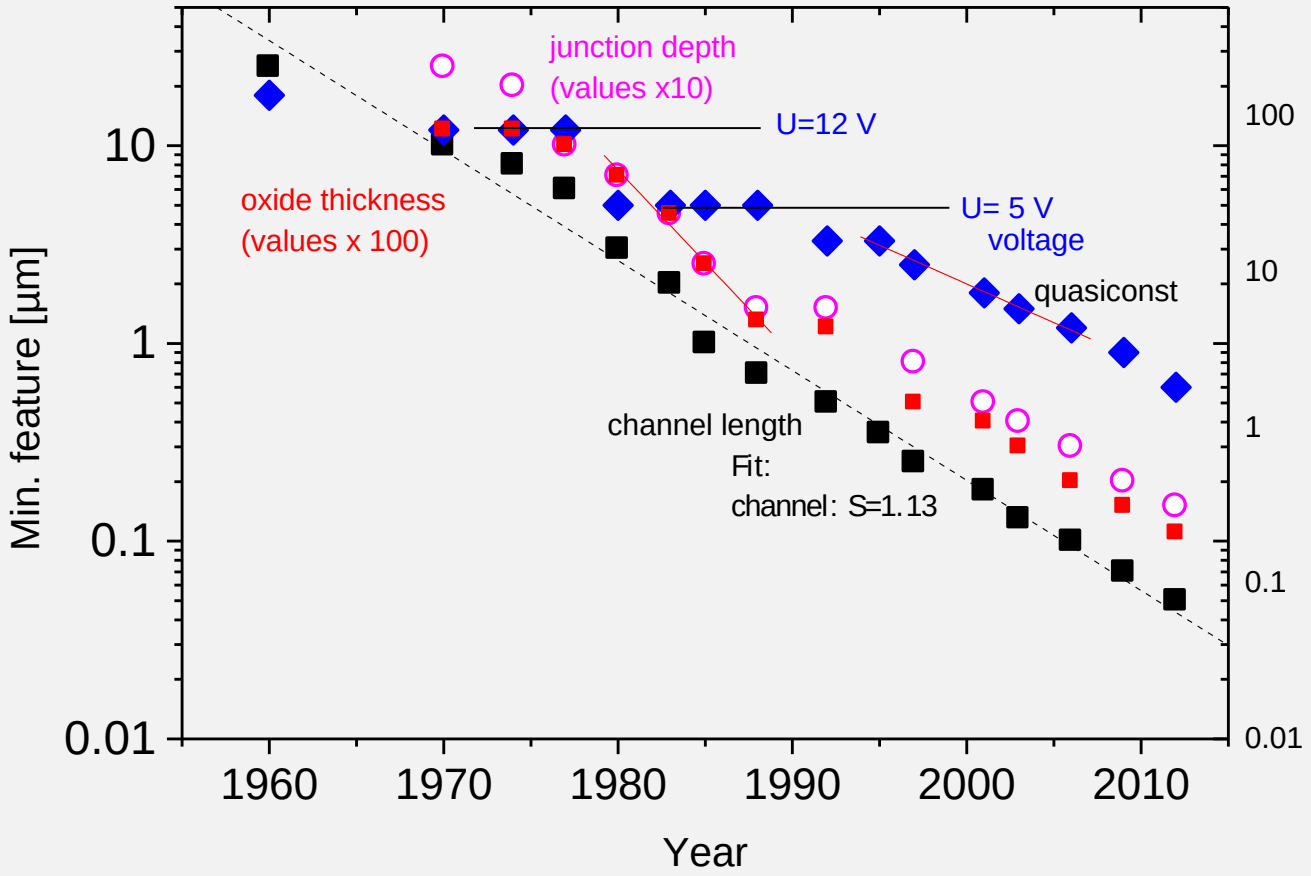
Note:
Usually between experimental prototypes -> industrial prototype -> high volume fabrication (~ 1mill. chips/year) several years are passing.
Dependent on source the data correlation year – properties may be shifted for 1-3 years

Note:
A factor 4 increase in chip complexity is achieved by:

- decreasing size by 13% per year -> √2 in 3 years -> factor 2 in area
- reducing functionality area by 1.3 per 3 years
- increasing chip size by 1.5 per 3 years

Historical Scaling

Shrinking of MOSFETs



1970-1976:
* shrinking of channel length
* no shrink of voltage, oxide thickness
-> discovery of short channel effects
-> voltage reduction, increased oxide scaling

1978-1987:
* const. Voltage 5V TTL-level
* more vertical shrink to come back to GCA
! Not const.voltage scaling !

1995 - 2000:
* variable voltage
* ~ quasi-const. scaling

> 2000 forecast:
* ~ subthreshold scaling

The Law: From economical pressure a lateral shrink (channel length) of 13% / year must be achieved -> area shrink of factor 2 / 3 years

Due to the technological capabilities of various companies short-time deviations from long term S=1.13 are the result

Beside the general Const.-Field scaling the scaling models may change from generation to generation

Historical Scaling

We have seen:

"Scaling" is the theoretical instruction how to avoid SCE (Short-Channel Effects) in the I-V characteristics

Reduce minimum feature size by S
-> L/S
-> t_{ox}/S
-> r_j/S
-> other geometries (e.g. interconnects)/S
-> N*S

but following these instructions new problems arise:

t_{ox} < 3nm -> through-diffusion of Boron from poly-Gate in channel causes V_T-shift
-> quantum-mechanical tunnel leakage current from channel through gate oxide
possible solutions: -> high-k gate dielectrics
-> and/or metal gate (due to gate depletion in poly-Silicon gate)

r_j thinner -> increasing resistance of S/D junctions
possible solutions: -> raised S/D by selective epitaxial growth-> thicker S/D -> lower resistance
-> insert of Ge in S/D results in higher solubility of doping atoms in S/D -> lower resistance
-> realization of vertical S/D

shrinking dimensons of interconnects -> inceasing RC-delay in interconnects
possible solutions: -> Cu and low-k intermetalic dielectrics / or air-gap dielectrics

Limits of Gate Dielectrics Scaling

By using thermal oxidation of silicon $Si + O_2 \xrightarrow{\sim 1000°C} SiO_2$ with SiO_2 the first working MOSFETs was possible (1960)

Some properties: high temperature stable
good adhesion to silicon and metals
selectively etching to silicon and metals

very good isolator: bulk resitivity 10^{15} Ωcm
very high breakthrough field: >10 MV/cm
dielectric constant: 3.9
best interface to silicon: $D_{it} < 10^8$ $(eVcm)^{-1}$ possible

From scaling theory a continous shrinking of oxide thickness is necessary to avoid short-channel effects (SCE)

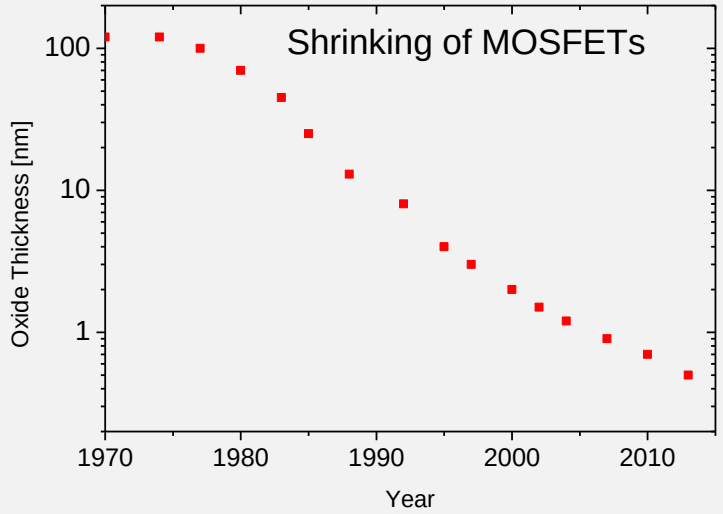


Table 1. Roadmap (ITRS 2001)

Year	'02	'04	'06	'07	'10	'13	'16
Node(nm)	130	90	70	65	45	32	22
Lg(MPU)(nm)	53	37	28	25	18	13	9
EOT (nm) Min	1.2	1.4	0.7	0.6	0.5	0.4	0.4
Max	1.5	0.9	1.2	1.1	0.8	0.6	0.5
Supply voltage	1.1	1.0	0.9	0.7	0.6	0.5	0.4

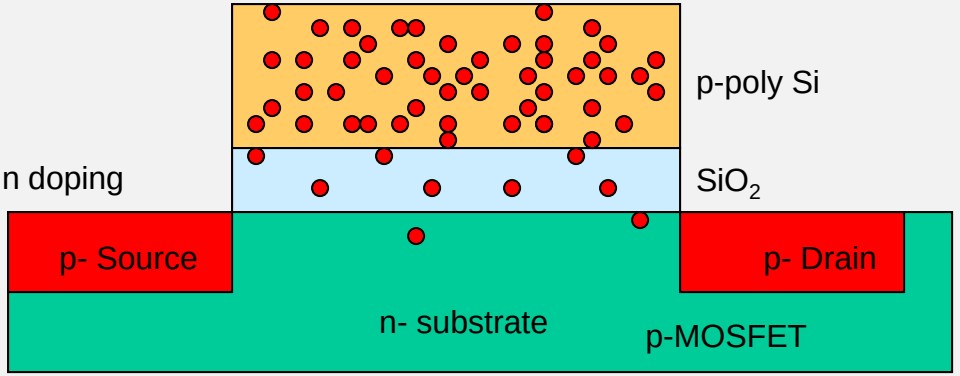
But when gate oxide thickness was reduced to 4 nm and above two serious problems arose:
- boron diffusion
- gate leakage current

Limits of Gate Dielectrics Scaling

Boron diffusion

A CMOS-inverter consists of an p-MOSFET and an n-MOSFET

The gate metal is poly-Silicon, highly-doped together with Source/Drain doping

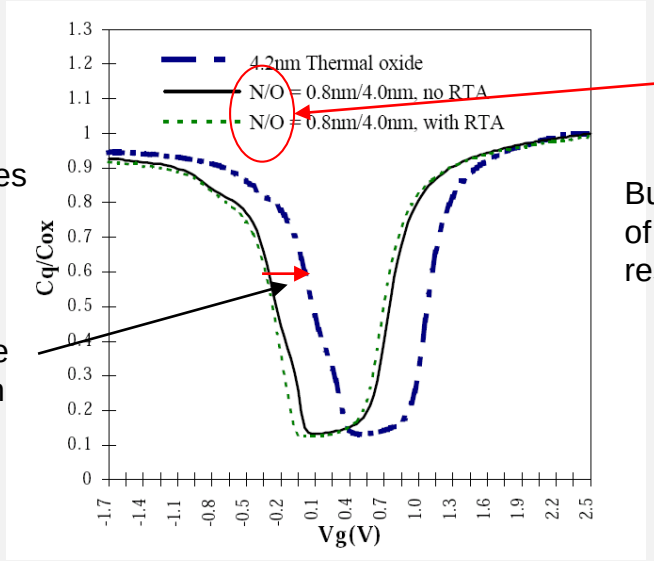


Boron (due to its high solubility in silicon) is used for the p-MOSFET, implanted as B^+ or BF_2^+

But Boron has a higher solubility in SiO_2 -> boron from gate poly-Si diffuse into SiO_2 and also in the silicon channel
These positive charges within the SiO_2 and also within the channel will shift the threshold voltage of the MOSFET and increase scattering, thus reducing mobility and on-current.

This effect, called **boron penetration**, is observed, when gate oxide thickness was around 4 nm and below (beginning 1990)

C-V characteristics of MOS-Gates



Shift of threshold voltage due to boron penetration

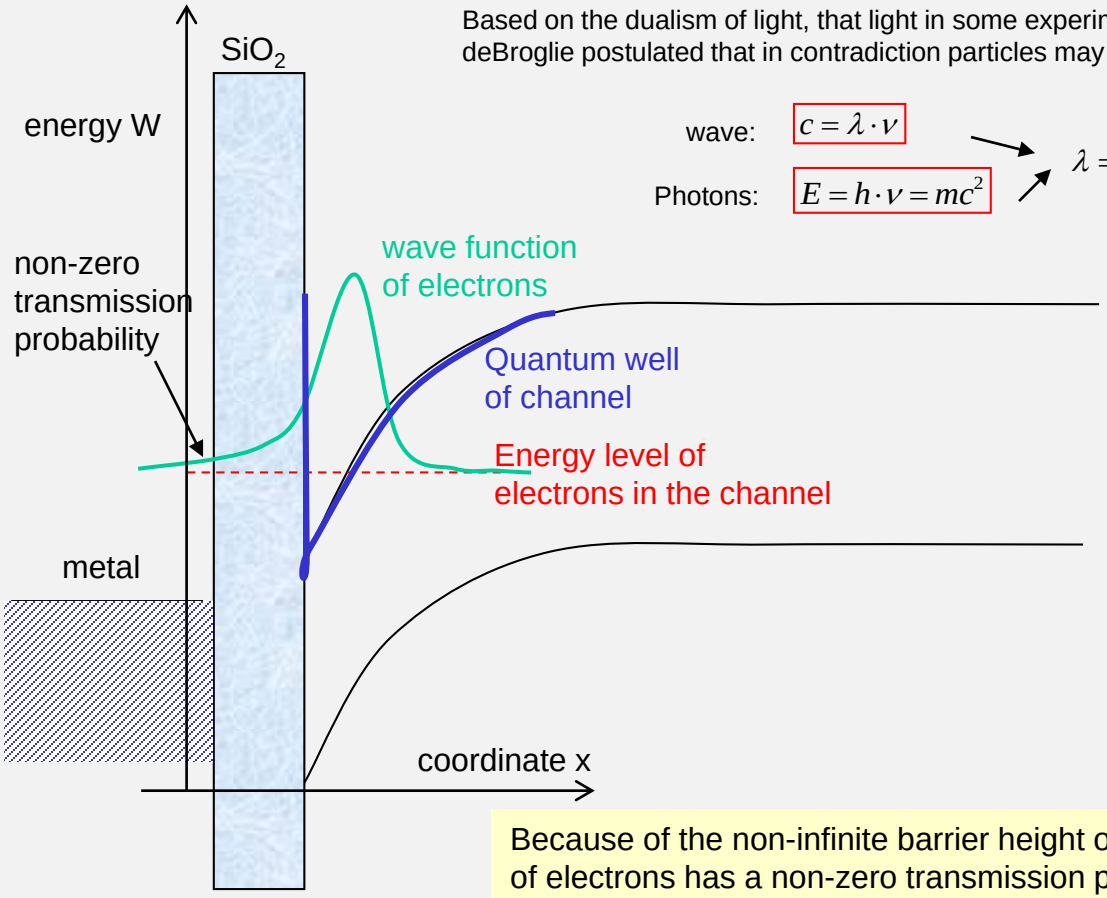
As a solution the incorporation of nitrogen within the oxide was/is used.

But nitrogen was found to diffuse to the interface of SiO_2 and Si and there increasing surface "roughness", resulting in lowering of mobility and therefore I_{on}

from: Wu, Lucovski, Int.Conf. on Characterization and Metrology for ULSI-Technology, 1998

Limits of Gate Dielectrics Scaling

Gate leakage current



de Broglie 1923:

Based on the dualism of light, that light in some experiments behave like an electromagnetic wave in other experiments like a particle deBroglie postulated that in contradiction particles may be described as wave with a wavelength depending on velocity (momentum):

wave: $c = \lambda \cdot \nu$

Photons: $E = h \cdot \nu = mc^2$

$$\lambda = \frac{h}{mc}$$

$$\lambda = \frac{h}{mv} = \frac{h}{p}$$

Dualism of wave and particle

for electrons

$$\lambda = \frac{1.23nm}{\sqrt{E_{kin}[eV]}}$$

for thermal electrons with E=0.026 eV the wave length is: λ ~ 7.6 nm

Because of the non-infinite barrier height of SiO_2 (3.1eV) the quantum-mechanical wave function of electrons has a non-zero transmission probability through the gate oxide.

This results in leakage current and long-term degradation of the gate oxide

Limits of Gate Dielectrics Scaling

From material properties of SiO_2 we calculate: very high breakthrough field: >10 MV/cm = 1V/nm

With reduced 0.5 V /nm the tunneling current due to Fowler-Nordheim is ~ 10^{-9} A /cm²

The increase of tunneling current is ~ 1 order/ 0.25 nm (1 monolayer)

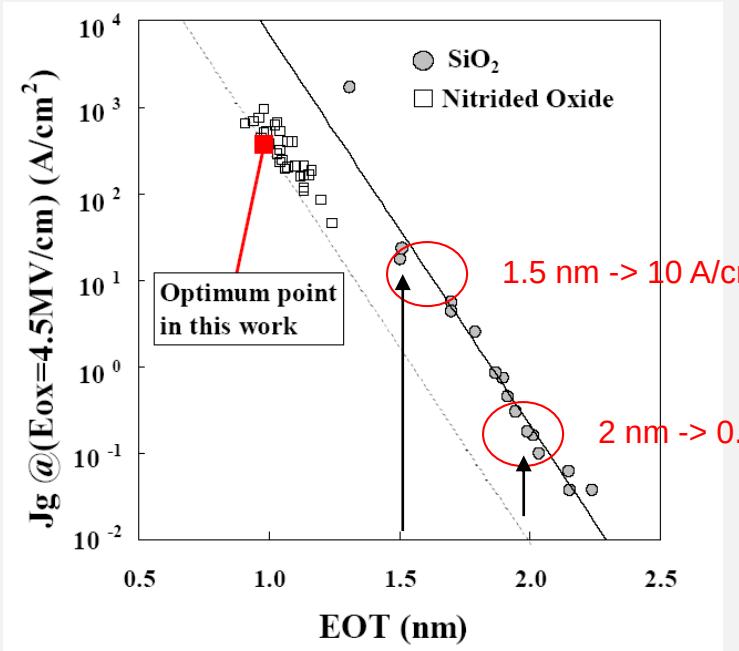
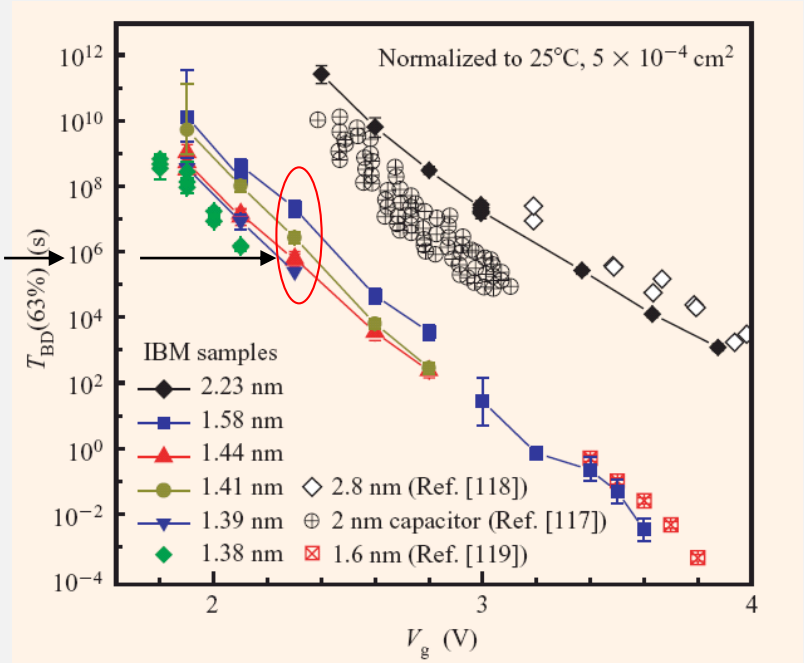


Fig.5 Dependence of Gate leakage current(Jg) on EOT.

from: Yanagiya et al. (Toshiba), "A 65nm CMOS Technology ...", IEDM 2002

12 days →



from: Stathis (IBM),IBM J. Res. & Dev., 46 (2002)

This current damages the gate oxide and reduces the lifetime of the gate oxide

Limits of Gate Dielectrics Scaling

But from roadmap to avoid SCE
we need oxide thickness < 1nm

Table 1. Roadmap (ITRS 2001)

Year	'02	'04	'06	'07	'10	'13	'16
Node(nm)	130	90	70	65	45	32	22
Lg(MPU)(nm)	53	37	28	25	18	13	9
EOT (nm) Min	1.2	1.4	0.7	0.6	0.5	0.4	0.4
Max	1.5	0.9	1.2	1.1	0.8	0.6	0.5
Supply voltage	1.1	1.0	0.9	0.7	0.6	0.5	0.4

Solution: We are searching for new gate dielectrics with k> 3.9 (SiO_2) -> high-k dielectrics

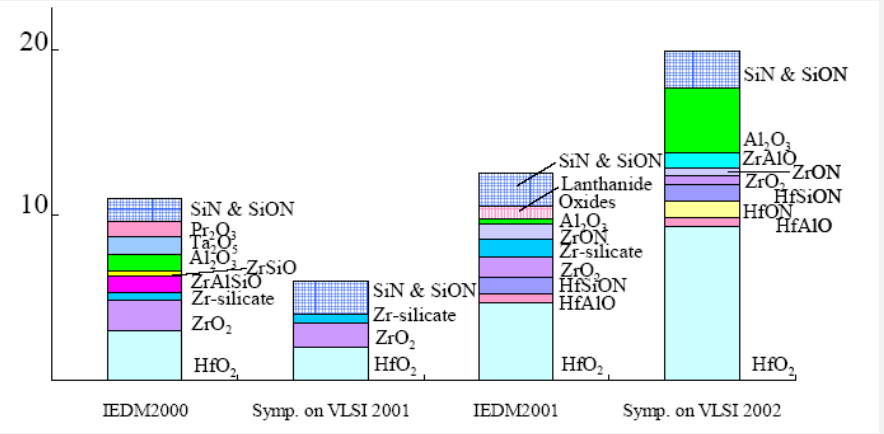
Because: $C = \frac{\varepsilon_0 \varepsilon_{ox} \cdot A}{t_{ox}}$ we achieve the same high capacitance C with higher ε_{ox} and thicker t_{ox}

Example: $\varepsilon_{SiO2} = 3.9$, if we use $\varepsilon_{ox} = 20$, we can use $t_{ox} = t_{SiO2} * \varepsilon_{ox} / \varepsilon_{SiO2} = 5 * t_{ox}$
-> if $t_{SiO2} < 1nm$ is needed from roadmap, we can use $t_{ox} < 5nm$, which is too thick for tunneling

Possible candidates:

Material	k		
NO stack	5-6	$HfAl_xO_y$	10-15
Al_2O_3	8-9	$HfSi_xO_yN_z$	10-15
$HfSi_xO_y$	10-15	ZrO_2, HfO_2	20-30
		Lanthanide Oxides	15-30

most promising candidates



Limits of Gate Dielectrics Scaling: New high-k materials

A new gate dielectric material must fulfill many requirements (as it is done by SiO_2):

1) $3.9 < k < \sim 20$

2) isolating material -> bandgap > 3 eV

3) band position compared to Si > high barrier for electrons in conduction band + high barrier for holes in valence band

4) density of interface states < 10^{10} $eV^{-1}cm^{-2}$

5) amorphous or single-crystalline lattice with no phase changes between room temperature and 1050°C, 30 sec
(if Gate is fabricated first, then S/D. But fabrication of dummy-gate, then S/D (high-temperature), then Gate may be possible)

6) No reaction with surrounding material, e.g. Si and gate metal
(if Gate is fabricated first, then S/D. But fabrication of dummy-gate, then S/D (high-temperature), then Gate may be possible)

7) compatible elements to CMOS-technology (for example no iron-oxides)

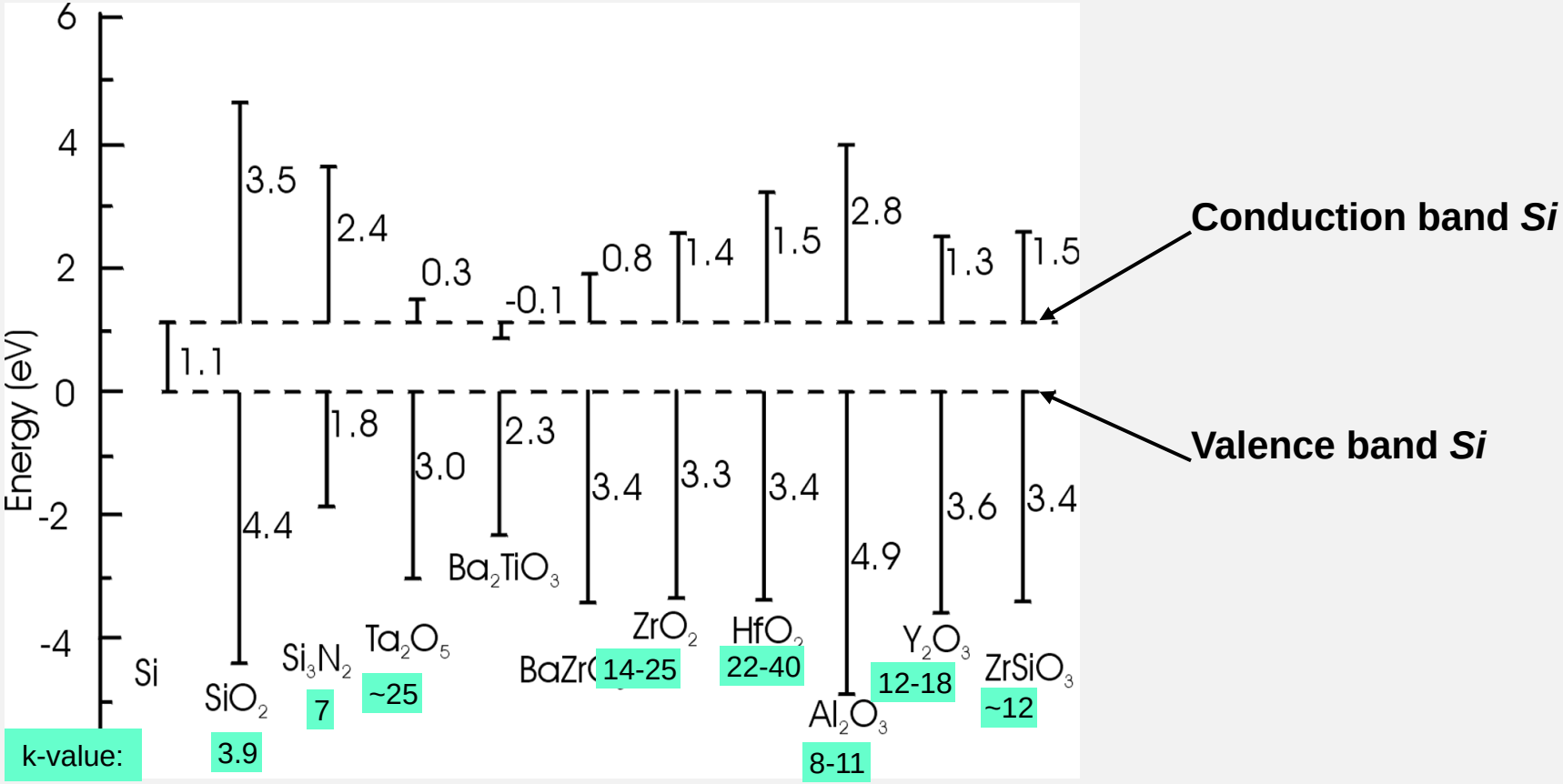
8) good adhesion to Si, SiO_2, nitride, ...

9) dry etch possible (-> all elements should build volatible compounds)

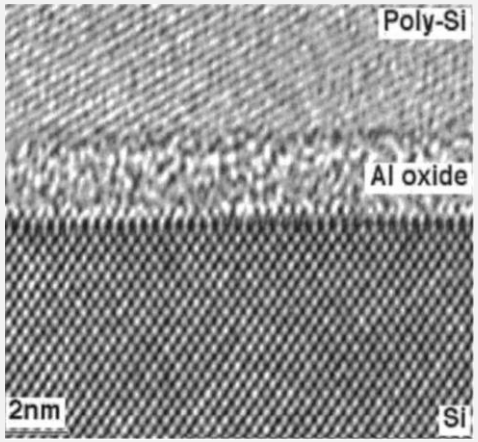
10) selectively etchable to Si, SiO_2, nitride, ...

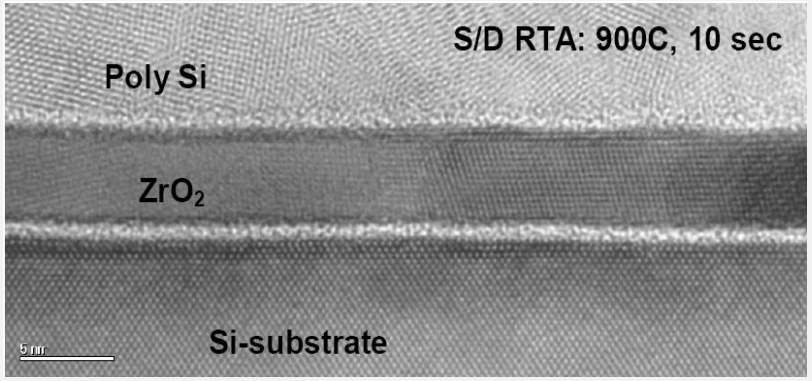
Limits of Gate Dielectrics Scaling: New high-k materials

Selection parameters: k, Egap, band position

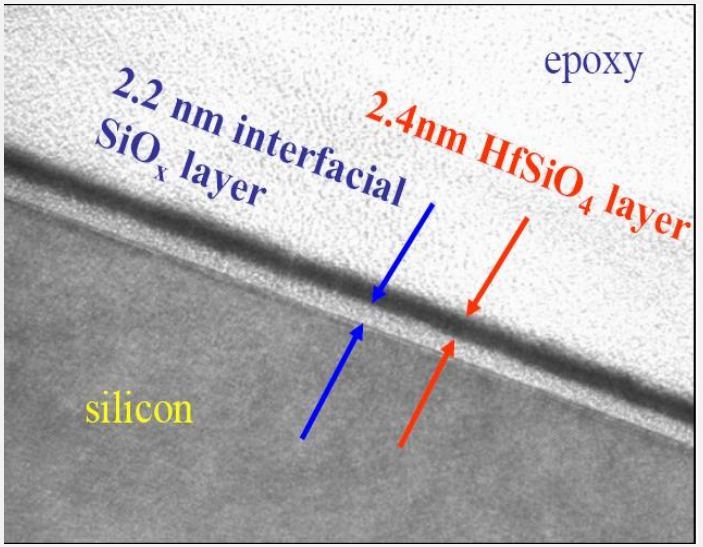


Limits of Gate Dielectrics Scaling: New high-k materials





ZrO_2, k> 15
after temperature treatment 900°C, 10 sec
interfacial layers of SiOx are created to Si and poly-Si



In most cases the high-k dielectrics is not stable
in the gate stack concerning high temperature

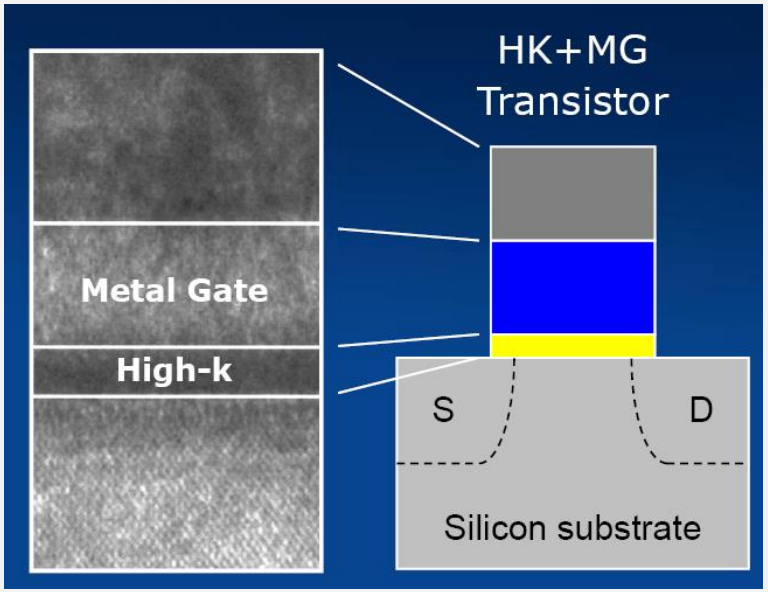
Possible solutions:

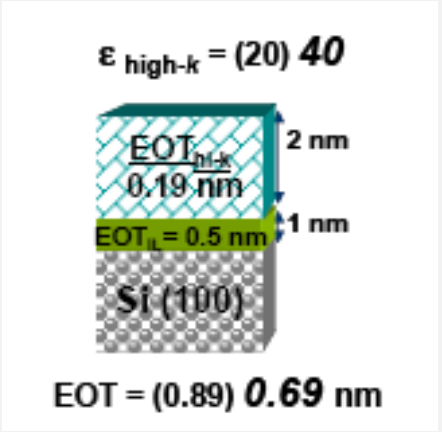
* further search and development of materials
* replacement gate (also good for metal gate)

Leading candidate materials in 2005: HfO_2 (K_{eff}~15 - 30); $HfSiO_x$ (K_{eff}~12 - 16)
Materials, process, integration issues to solve

and then in 2007
INTEL presented HfOSiN

Limits of Gate Dielectrics Scaling: New high-k materials





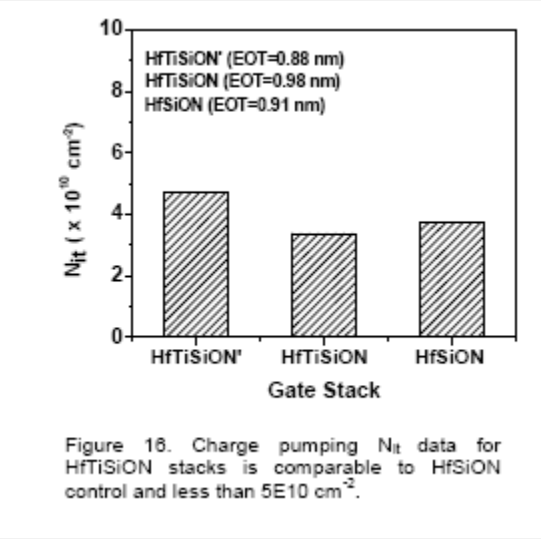


Figure 16. Charge pumping N_{it} data for HfTiSiON stacks is comparable to HfSiON control and less than 5E10 cm^{-2}.

Interface states < $5*10^{10}$ $cm^{-2}eV^{-1}$

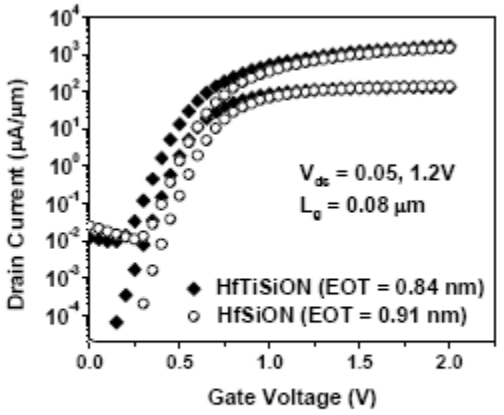


Figure 13. I_d-V_g for 10 x 0.08 μm devices indicates comparable subthreshold slope for HfTiSiON vs. HfSiON control.

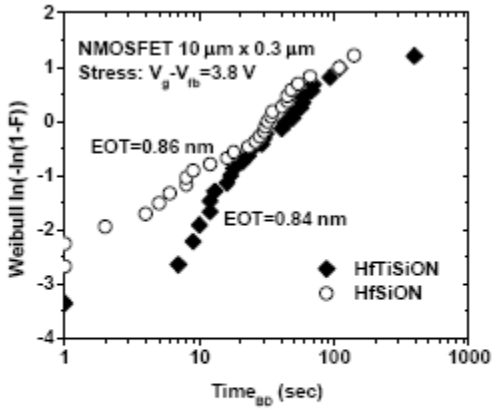
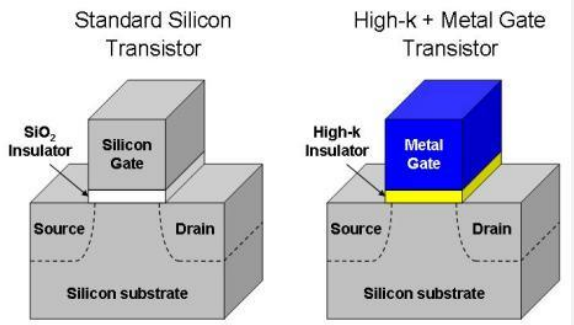


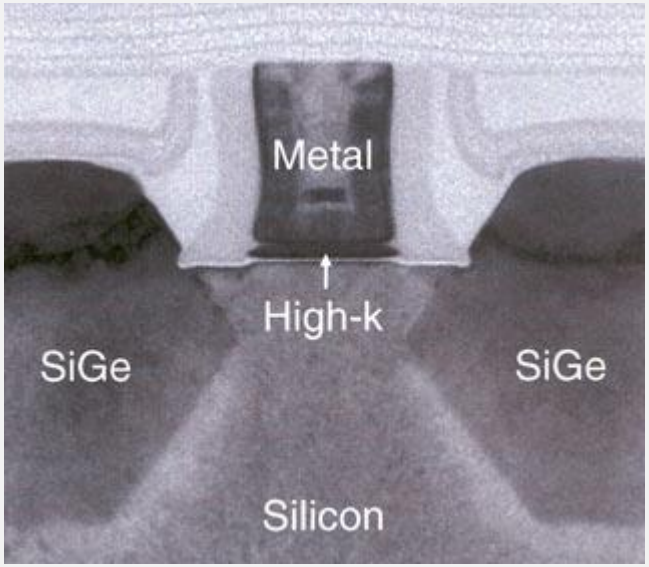
Figure 18. Time dependent dielectric breakdown for HfTiSiON is similar to HfSiON of similar EOT. Fewer early fails for HfTiSiON may be due to physically thicker dielectric stack.

INTEL, IEDM 2007

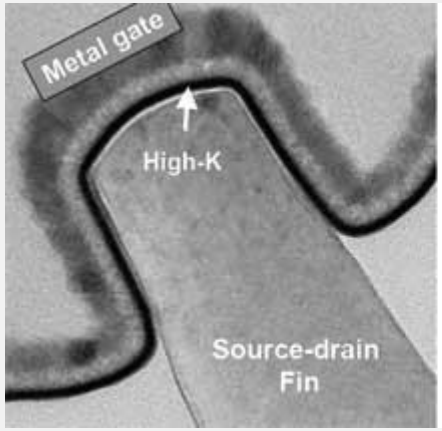


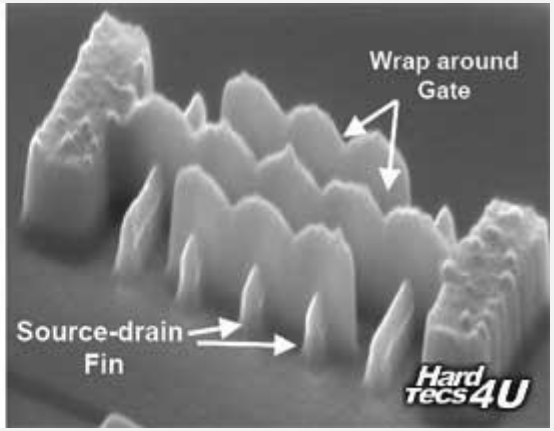
Limits of Gate Dielectrics Scaling: New high-k materials

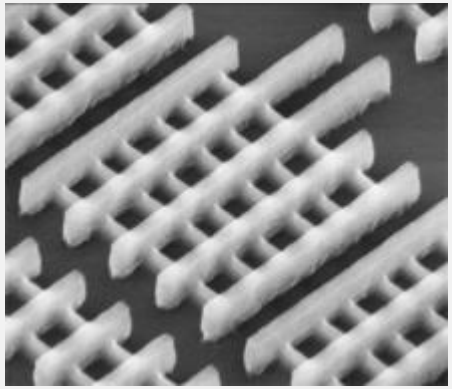
INTEL, IEDM 2007



INTEL, 2012







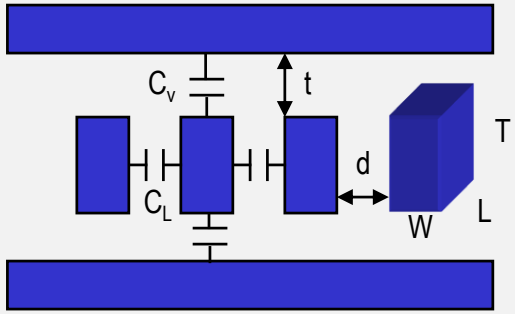
Limits of interconnect- Scaling: RC-Delay

Due to shrinking the resistance of interconnects itself and by narrowing the parasitic capacitances are increasing. Both increases switching times, although the single transistor is getting faster and faster.

The interconnect resistivity:

$$R = \rho \cdot \frac{L}{T \cdot W}$$ (usually an aspect ratio of T : W = 2 :1 can be filled by technology without problems)

$$R = \rho \cdot \frac{L}{2W \cdot W}$$ (statistically most interconnects exhibit a length of about 100µm, but longer interconnects are existing, see next page)



Example: 100nm Generation

$$R = \rho \cdot \frac{L}{2 \cdot W^2} = 3.0 * 10^{-6}\,\Omega cm \cdot \frac{100\mu m}{2 \cdot (100nm)^2} = 150\Omega$$

The Parasitic Capacity:

$$C = 2 \cdot C_{lateral} + 2 \cdot C_{vertikal} = 2\varepsilon_0\varepsilon_{SiO2} \cdot \frac{L \cdot T}{d} + 2\varepsilon_0\varepsilon_{SiO2} \cdot \frac{L \cdot W}{t}$$

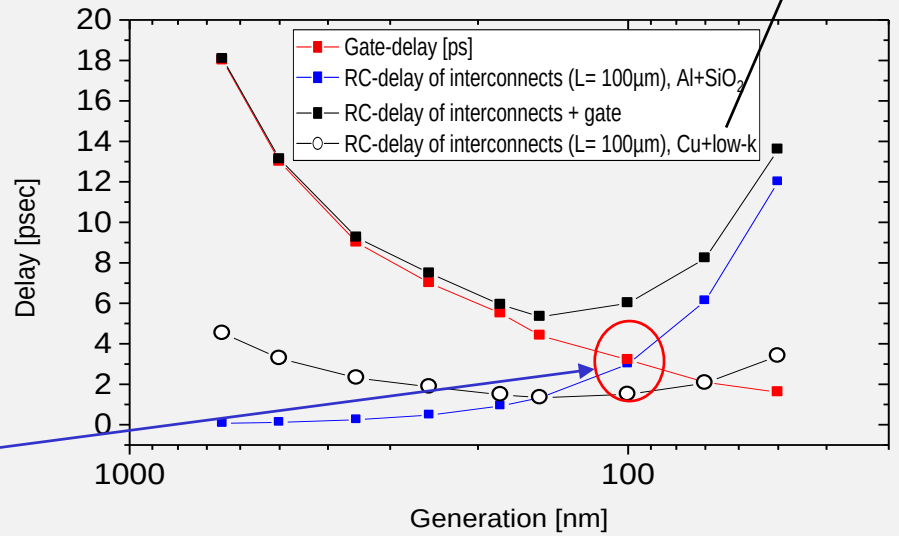
$$C = 2\varepsilon_0\varepsilon_{SiO2} \cdot L \cdot \left[\frac{T}{d} + \frac{W}{t}\right] \approx 2\varepsilon_0\varepsilon_{SiO2} \cdot L \cdot \left[\frac{2W}{W} + \frac{W}{W}\right] = 6\varepsilon_0\varepsilon_{SiO2} \cdot L$$

(d,t ~W, see next page)

Example: 100nm Generation

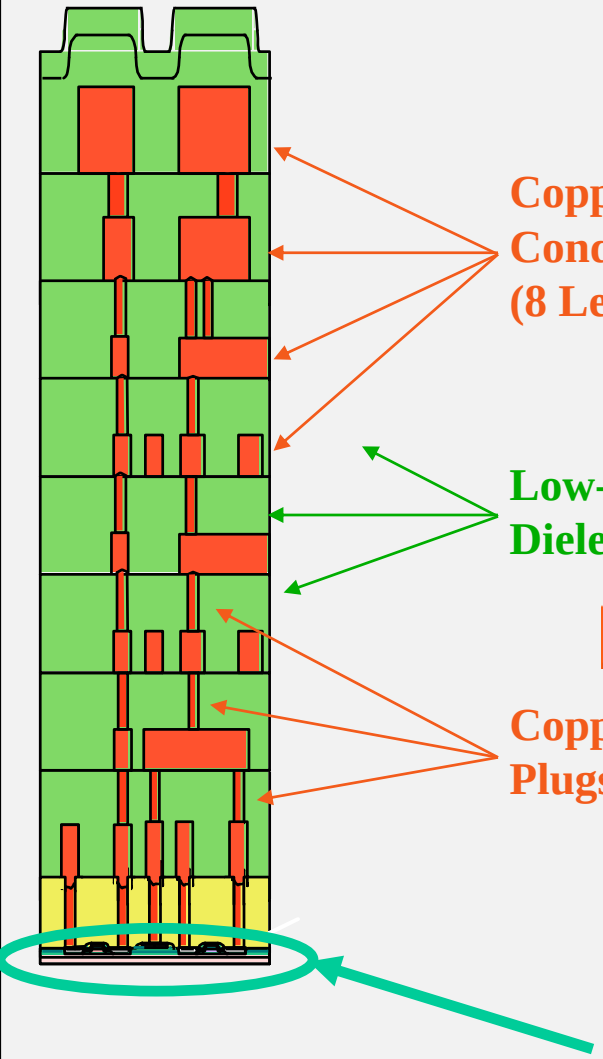
$$C = 6\varepsilon_0\varepsilon_{SiO2} \cdot L = 6 \cdot 8.85 * 10^{-14}\,F/cm \cdot 3.9 \cdot 100\mu m \approx 2 * 10^{-14}\,F$$

All together: RC = 150Ω * 2*10^{-14} F = 3 psec



ε_{SiO2} = 3.9
low-k: ε < 3.9

Limits of Interconnect Scaling



Year / Parameter	2003	2005	2008	2011	2014
Technology(nm)	120	100	70	50	35
# of Transistors	95.2M	190M	539M	1523M	4308M
Clock Frequency	1724 MHz	2000 MHz	2500 MHz	3000 MHz	3600 MHz
Chip Area (mm^2)	372	408	468	536	615
Wiring Levels	8	8-9	9	9-10	10
Pitch(L/I/G)(nm)	330/420/690	265/340/560	185/240/390	130/165/275	95/115/190
A/R (L/I/G)	1.6/2.2/2.8	1.7/2.4/2.8	1.9/2.5/2.9	2.1/2.7/3.0	2.3/2.9/3.1
Dielectric Const.	2.2-2.7	1.6-2.2	1.5	<1.5	<1.5

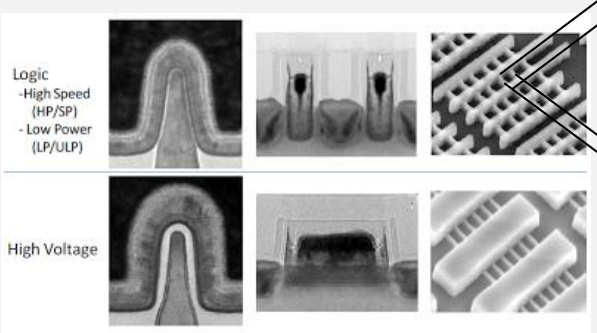
At sub-0.25um feature sizes, 90% of the total capacitance is dominated by line-to-line capacitance.

Increasing clock speeds and reduced sizes require new dielectrics with k values of less than 3.

Limits of interconnect: Metal levels and pitch

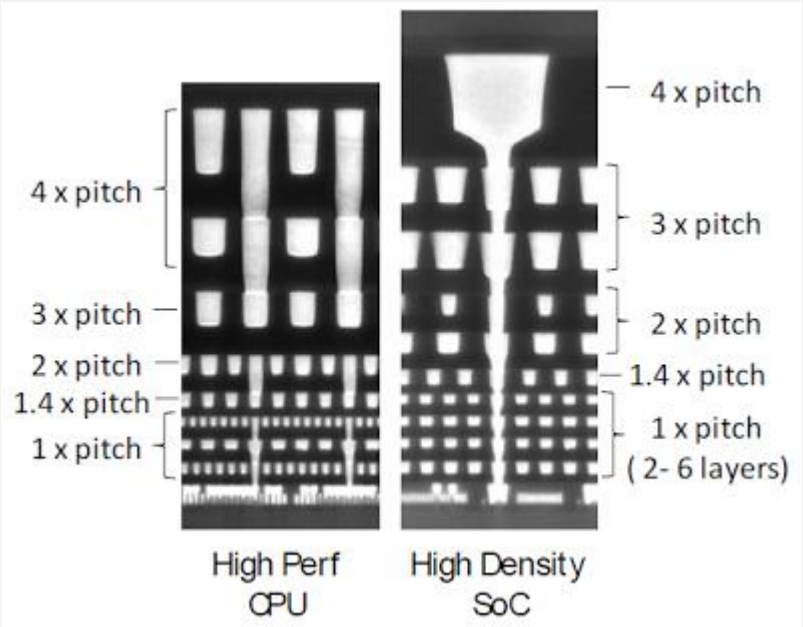
INTEL, Dec.2012, IEDM conference

22nm FINFETs



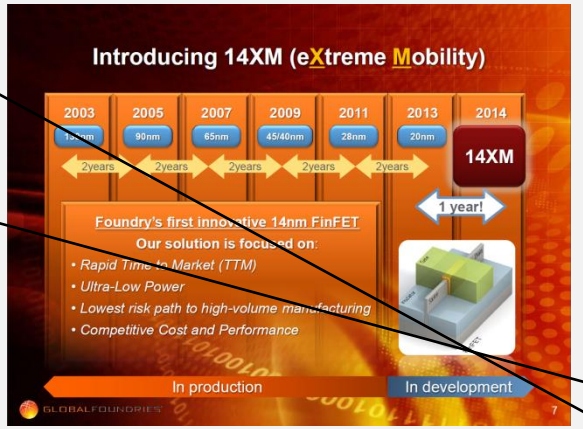
Gate pitch: 90nm

Fin pitch: 60nm



up to 12 metallization levels

GlobalFoundries, 2012 Common Platform Tech Forum (CPFT)



20nm node
- replacement-gate
- metal1 pitch: 64nm

14nm node
Fin pitch: 48nm
Contacted gate pitch: 64nm

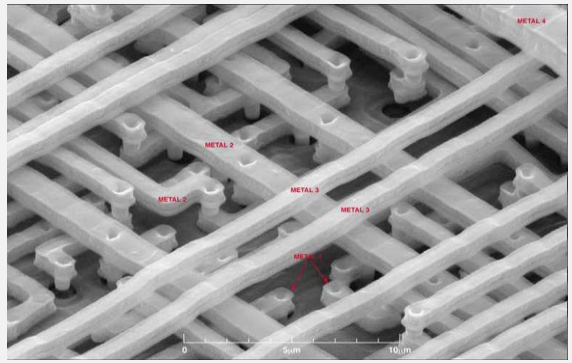
Limits of interconnect: Low-k materials

•Material	•ε_{di}
•Free space	•1
•Acrogels	•1.5
•Teflon AF	•2.1
•Aromatic thermosets (SiLK)	•2.6 – 2.8
•Polyimides (organic)	•3.1 – 3.4
•Silicon dioxide	**•3.9 – 4.5**
•Glass epoxy (PCBs)	•5
•Silicon nitride	•7.5
•Alumina (package)	•9.5
•Silicon	•11.7

Low-k Materials:

1) Vacuum/air exhibit the smallest possible dielectric constant, this is k=1

Free, non-supported interconnects are not mechanically stable,
CMP is not possible (may be with etching-out isolation at last step)



2) Try to introduce air in isolation -> Nanoporos

Since (2005) the most promising candidates are polymers, which create nanopores when annealed

Principles for the Reduction of the k-Value

Dielectric Constant Reduction Methods

Reduce **polarization strength** and **density**

basic concept is to replace strong Si-O bond by weaker bond types as provided by C, F

basic concept is to create polymers, because during formation a lot of free space (ε =1) is created

more based on inorganic SiO_2 → hybrides (Si,O,C,F,H) ← more based on organics C, O, H, ..

Table of polarization strength with covalent C- bonds

from: Miller et al., Macromolecules 23(1990)3865 (polarizability)
Pine, Organic chemistry 5th ed., 1987 (bond strength)

Bond	C≡N	C≡C	C=C	C=O	O-H	C-H	C-O	C-F	C-C
Polarizability [A]	2.24	2.04	1.64	1.02	0.71	0.65	0.58	0.56	0.53
bond strength [kcal/mole]	213	200	146	176	102	99	84	116	83

avoid 3-fold bonds → avoid 2-fold bonds → low polarizability + high bond strength

additional porosity by space creators (porogenes) → air gaps ← avoid 3-fold bonds

Market Overview

About 10 manufacturer are developing low-k materials, the market value is about 300 Mill. € (2005)

status 2007

Equipment manufacturer try to force the development of **CVD**-materials

Advantage: for SiO_2-based CVD-layers established tools and integration experience exists

Disadvantage: due to temperature limitation T< 400°C in the metal layers there is not much freedom for material development

Black Diamond von Applied Materials
(based on trimethylsilane, fabricated by Dow Corning, distributed by Air Products)

CORAL, TOMCATS (porous SiO:C) from Novellus
(based on tetramethyl silane, tetramethylcyclotetrasiloxane, distributed by Air Products)

ORION (porous SiO:C, k>2.2) from Trikon Technologies

Most suppliers of CVD-tools offer processes for deposition of SiO_2-based layers with k between 2.7 to 2.4, layers with k~2.0 were demonstrated

Manufacturers of chemicals offer **Spin-on** materials

Advantage: due material composition additional porosity can be created -> Ultra-low k (ULK)
Disadvantage: problems with integration

organics: (based on C,O,H,F) (polymeric)

SiLK (porous polymer, k>2.2) from Dow Chemical

GX-3 from Honeywell Electronic Materials (Sunnyvale, Calif.)

inorganics: (mainly based on Si,O,H,F)
silicon-based systems with carbon (or methyl) groups and/or hydrogen attached (hydrogen silsesquioxane)
* advantage: better temperature stability, lower density compared to organics

XLK from Dow Corning (porous version of FOx)

Nanoglass E (porous SiO_2) from Honeywell Electronic Materials

Air Products (world´s largest producer of SiF_4) porous silica material called **MesoELK**

LKDxx (porous MSQ, k>1.9) from JSR (Tokyo)

ZIRKON (porous MSQ, k>1.8) from Rohm and Haas (Shipley)

Example of Spin-on Low-k Material

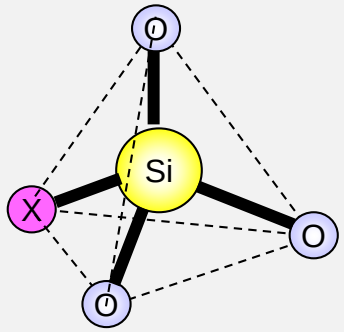
Basic Process: spin-on

typical: Silsesquioxane (abreviated SQ or SSQ)

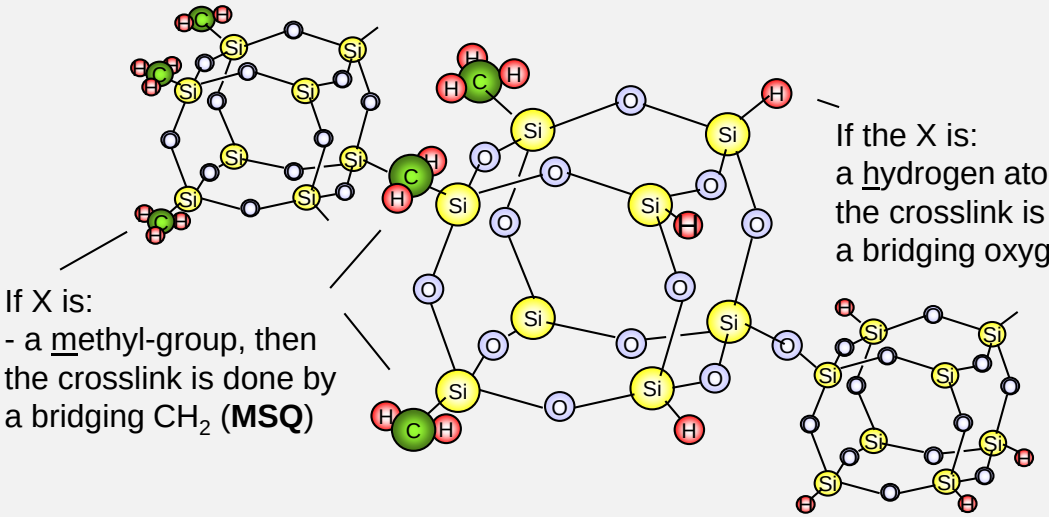
Silsesquioxane forms precise 3-dim structures (like bucky balls) and with radicals X properties like polymerisation, solubilization, polaribility, stability,.. can be tailored.

some of the polar Si-O bonds are replaced by less polar Si-X bonds

Formation of cross-linked Oligomeres (here octamere) (creating free space by cage-like arrangement)



Si-O-Si bond is named siloxane



If X is:
- a methyl-group, then the crosslink is done by a bridging CH_2 (**MSQ**)

If the X is:
a hydrogen atom then the crosslink is done by a bridging oxygen (**HSQ**)

Properties:

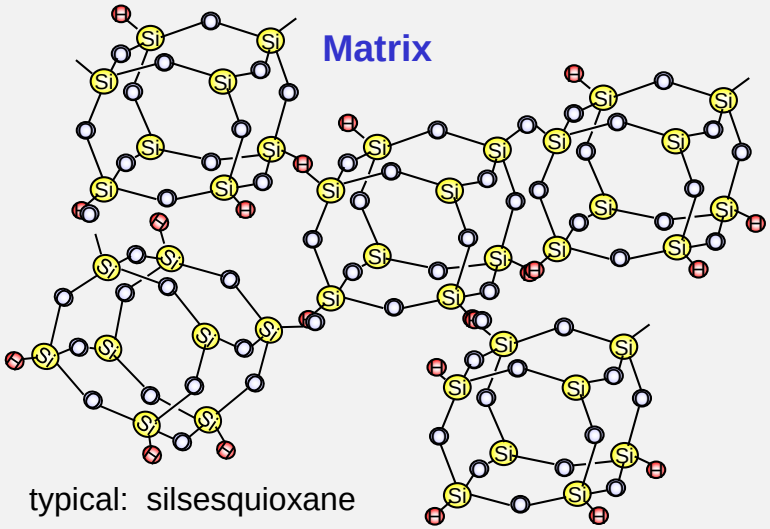
+: $k_{min} > 2.6$

-: weak mechanical strength
-: low thermal stability, T ~ 500°C
-: bad adhesion -> improved by promoters

HSQ: k> 2.9
MSQ: k> 2.7

Example of Spin-on Ultra Low-k Material

Typically in the low-k polymer (the matrix) an additional thermal instable polymer (called porogen, ~ 25 vol %) is added.
At formation temperature of the low-k network the instable polymer decomposes and evaporates, leaving nanopores

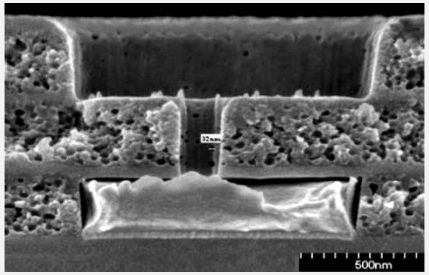


typical: silsesquioxane

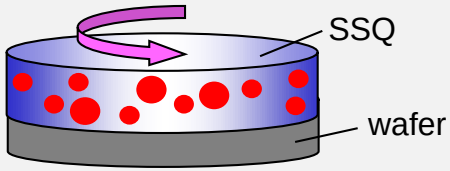
Porogen

typical: co-polymeres

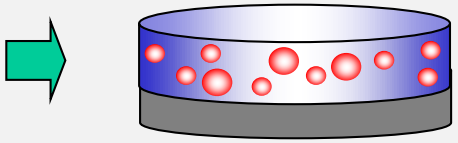
- poly(methyl methacrylate-*co*-dimethylaminoethyl methacrylate)
= P(MMA-*co*-DMAEMA)



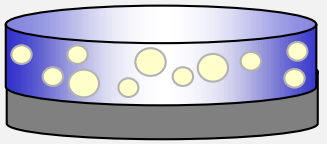
Typical Process:



spin-coat mixture of matrix
and porogen



bake-out at 430°C, 5min
-> matrix crosslinking at ~200°C
-> decomposition of porogen at 400°C

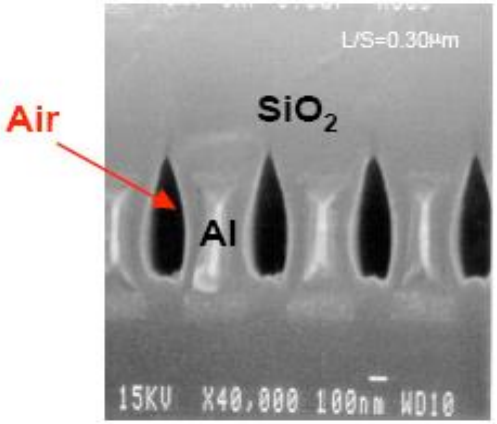


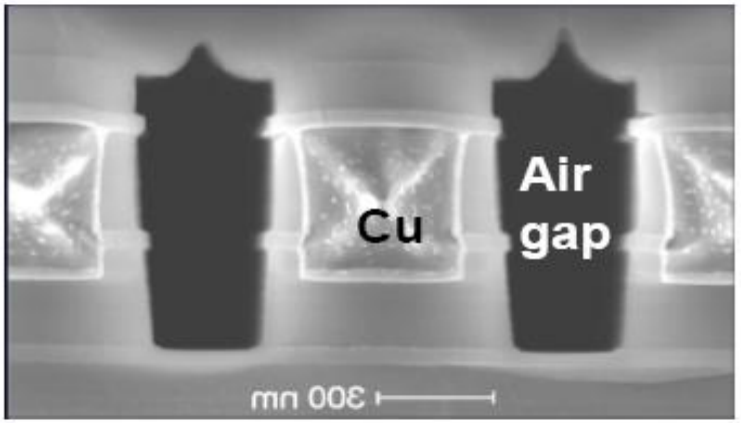
at room temperature
porous are remaining
(up to 90% of porogen content)

Air-Gaps as Low-k

Air-gap as Low-k Dielectrics

Air-Gap Interconnect Structure





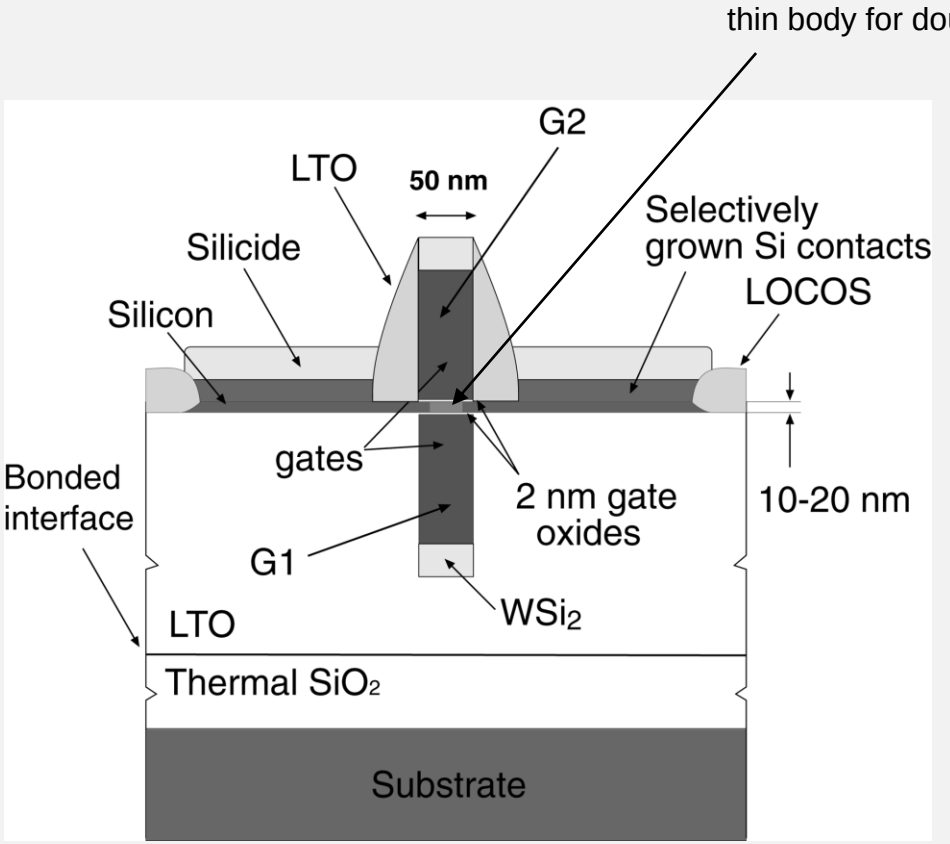
Ref: Shieh, Saraswat & McVittie. IEEE Electron Dev. Lett., January 1998

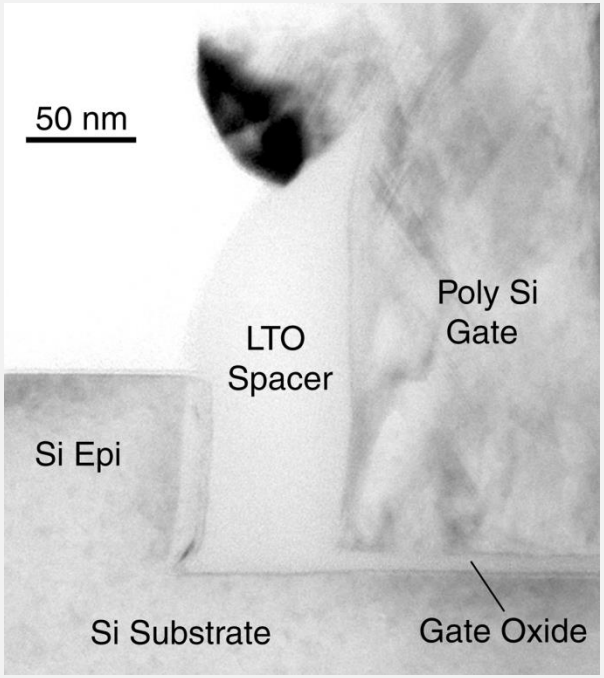
Source: Werner Pamler, Infinion

Old dielectric SiO_2 K = 4

Ultimate limit is air with K = 1

Limits of Junction Depth Scaling

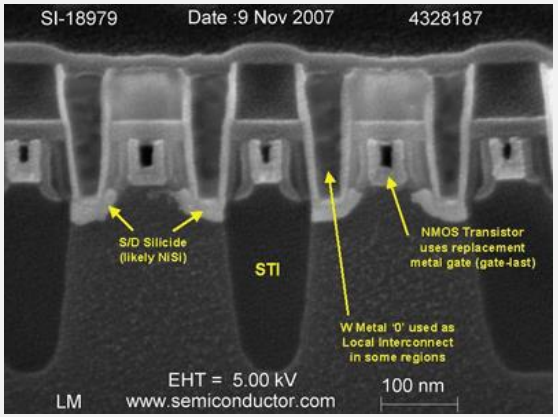




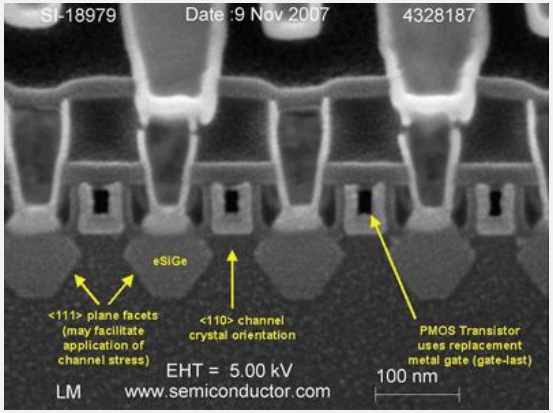
from: Nanostructures Laboraty at MIT
http://nano-web.mit.edu/annual-report00/13.html

To achieve low contact resistance and shallow junctions selective epitaxial growth could be used

Limits of Junction Depth Scaling



INTEL, 45nm n-MOS



INTEL, 45nm p-MOS

Saving area but keeping resistance low
-> try vertical S/D
-> but: sophisticated design for SCE

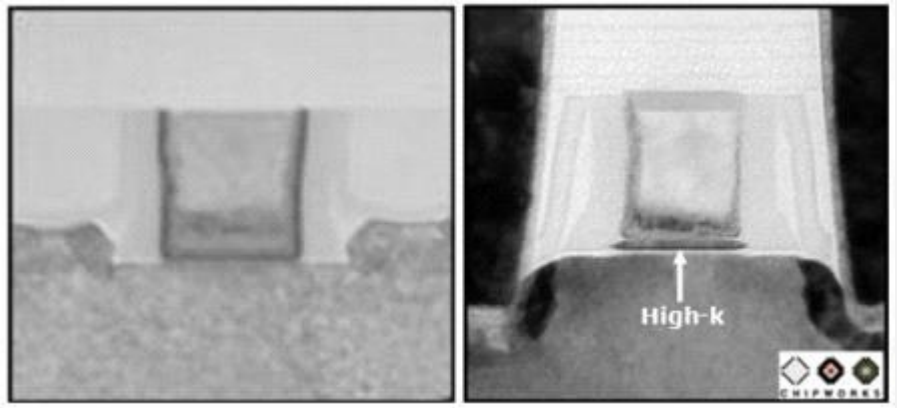


Fig. 2 Intel IDF image of 32-nm NMOS gate (left) and Chipworks image of 45-nm NMOS gate

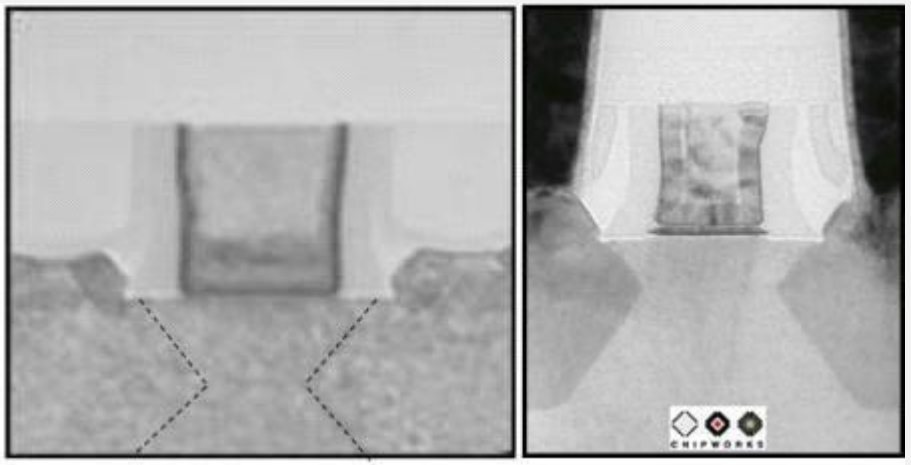


Fig. 3 Comparative cross-sections of Intel 32-nm NMOS (left) and 45-nm PMOS transistors

4.1 Shrinking

- Economical pressure requires a lateral shrink (channel length) of ~ 13% / year -> area shrink of factor 2 every 3 years
- Shrinked devices have the **advantage of better dynamic performance**, but the **disadvantage of worse static behavior**

4.2 Short Channel Effects

- The disadvantage of worse static behavior appears, if the electrical fields are changed
- The classical short channel effect appears, if the Source/Drain depletion zones occupy most of the channel length
  - This results in a **lowering of the threshold voltage** and in an **IV-characteristics dependent on S/D voltage**
    - This results in higher leakage current, lower S-factor
- The classical short channel effect can be analyzed by two similar models:
  - The Charge-sharing model explains the lowering of threshold voltage by shared depletion zones of Gate and S/D
  - The DIBL model explains extreme lowering of threshold voltage by channel barrier lowering due to connected S/D depletion zones

4.2 Short Channel Effects (cont.)

- In addition to the classical short channel effect shrinked devices usually show **higher and 2/3-dim electric fields**
  - This results in many effects like modifications in carrier transport and break-through phenomena

4.3 Scaling

- Scaling describes the attempt to keep the electric fields the same in a shrinked device
  - The starting point of scaling is the simple, classical long-channel MOSFET without any modifications
- Due to external requirements (as given by technological capability and circuit requirements) several scaling models exist:
  - Electrostatic scaling, with special conditions like Constant-Field Scaling, (Quasi)-Constant-Voltage Scaling, General Scaling
  - Subthreshold scaling
- Scaling rules are overruled by addition of doping structures in the MOSFET structure
- Todays and future MOSFET design is high sophisticated task to balance all parameters to a **good compromise**

4.1 Shrinking
- Advantages
- Disadvantages

4.2 Short Channel Effects
- Discovery and first analysis
- Quantitative analysis
  - Charge Sharing Model
  - Drain-Induced Barrier-Lowering Model

4.3 Scaling Models
- Electrostatic Scaling
- Subthreshold Scaling

4.4 Limits of Scaling
- Gate dielectrics, high-k
- Interconnect Materials: Cu and low-k

End of Chapter 4