MOSFET devices: physics and technology

Enrico Sangiorgi
University of Bologna and IUNET

ESONN 2019

OUTLINE OF THE COURSE
CLASS 1: Device Modeling

- Introduction to Nanoelectronics

- Modeling electron devices:
  - Electrostatic and electrodynamics in mesoscopic devices
  - The semi-classical approach to model charge transport

OUTLINE OF THE COURSE
CLASS 2: the MOSFET

- MOSFET : basic theory
  - Threshold voltage
  - MOSFET current-voltage characteristics.
  - Short Channel Effects

- MOSFET scaling and scaling strategies
  - Scaled MOSFETS
  - Limits of scaling

OUTLINE OF THE COURSE
CLASS 3: from nano to Tera

- MOSFET Scaling:
  - Recent Scaling Scenario
  - Multigate MOSFET architectures

Suggested Readings

Y. Taur T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, ISBN 0 521 559596

Mark Lundstrom, Fundamentals of Carrier Transport, 2nd Edition, Cambridge University Press

R.S. Muller T.I. Kamins, Device Electronics for Integrated circuits, Wiley

Robert F. Pierret, Semiconductor device Fundamentals, Addison-Wesley, ISBN 0201 54393 1

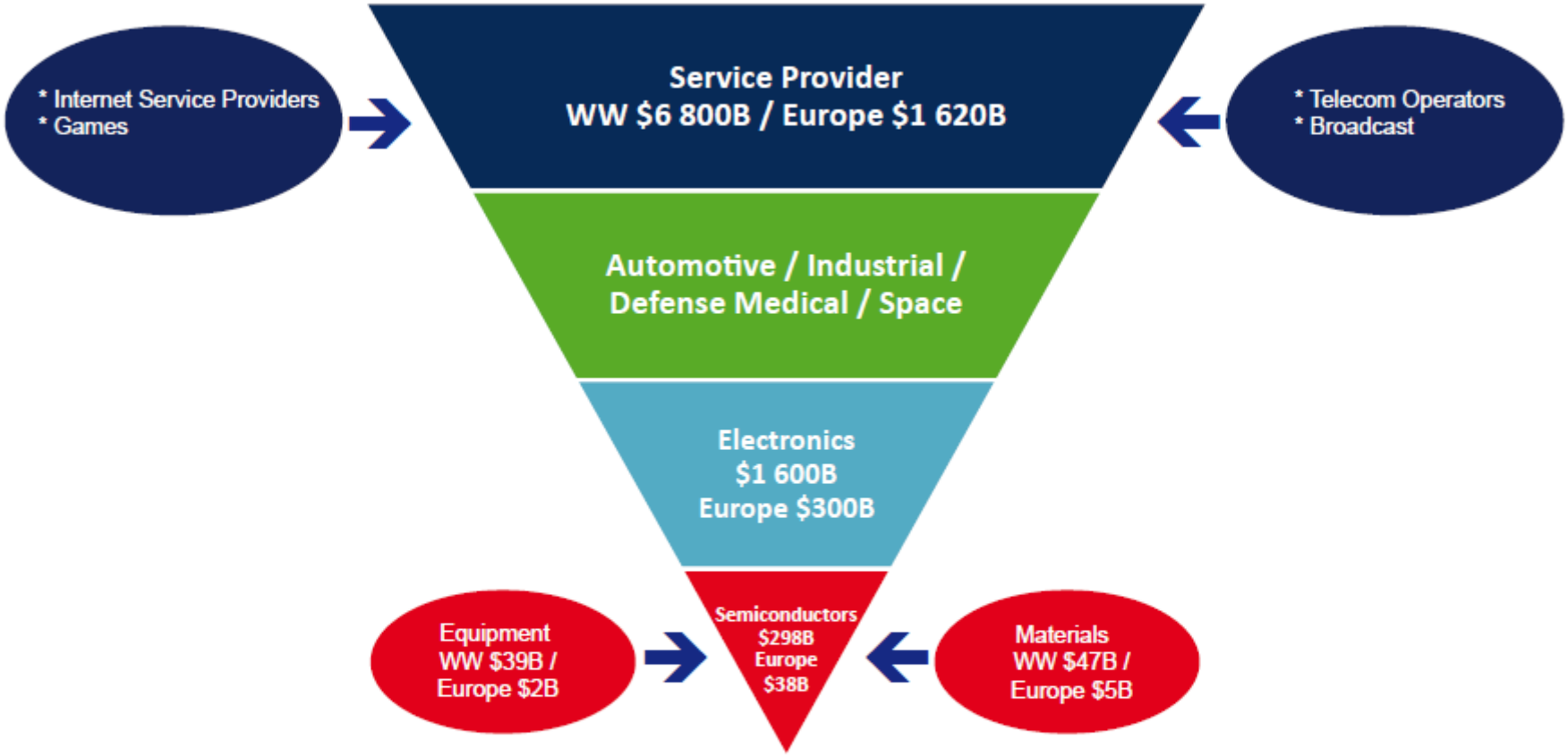
Introduction to Nanoelectronics: The Good, the Bad and the Ugly

The Good, the Bad and the Ugly (Italian: Il Buono, il Brutto, il Cattivo) is a 1966 Italian epic spaghetti western film directed by Sergio Leone, starring Clint Eastwood

Enrico Sangiorgi

The Good of Nanoelectronics: Business and Science.

The pyramid of wealth



Semiconductors provide the knowledge & technologies that generate some 10% of global GDP

2010 World GDP = $73300B

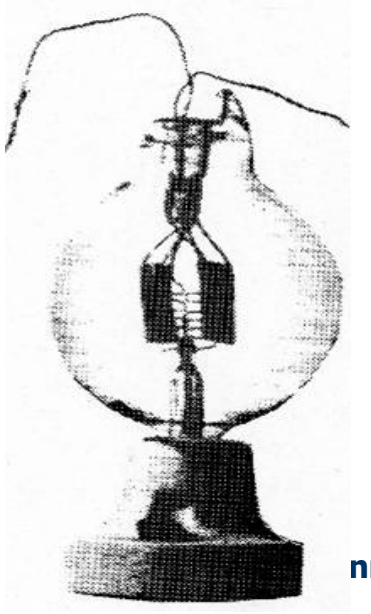
2010 EU GDP = $15040B (ppp based)

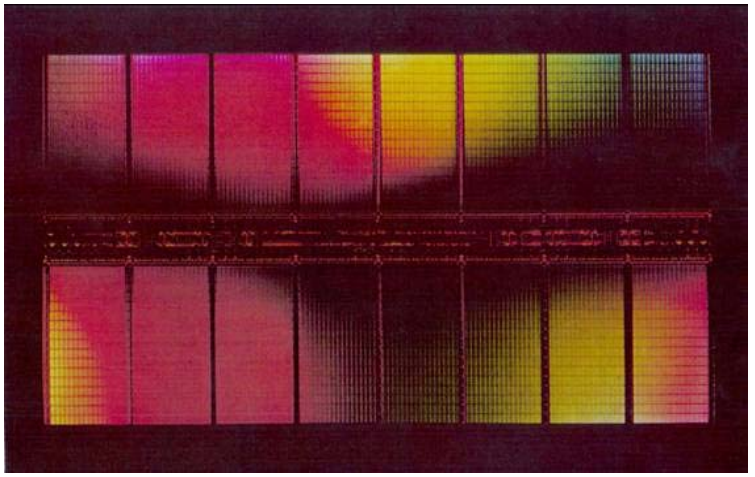
ppp = purchase power parity

Source: CATRENE Program Review 2012 - excerpts

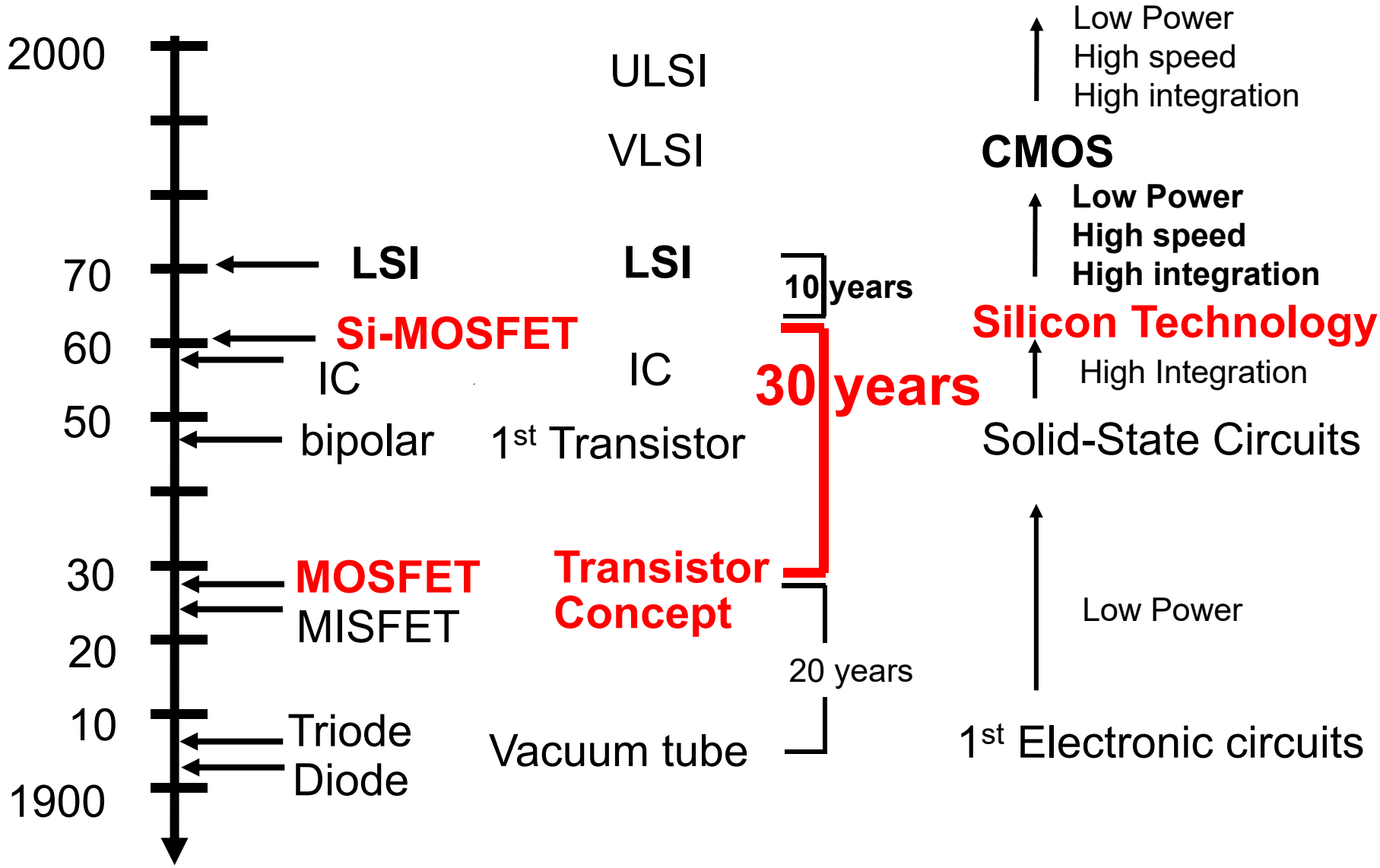
- Electronics is the
  Most relevant invention of the 20th century
- Electronic Circuits in 100 years
  from Vacuum tube to ULSI

J.J. Thomson 1896 electron Lee De Forest 1907 triode





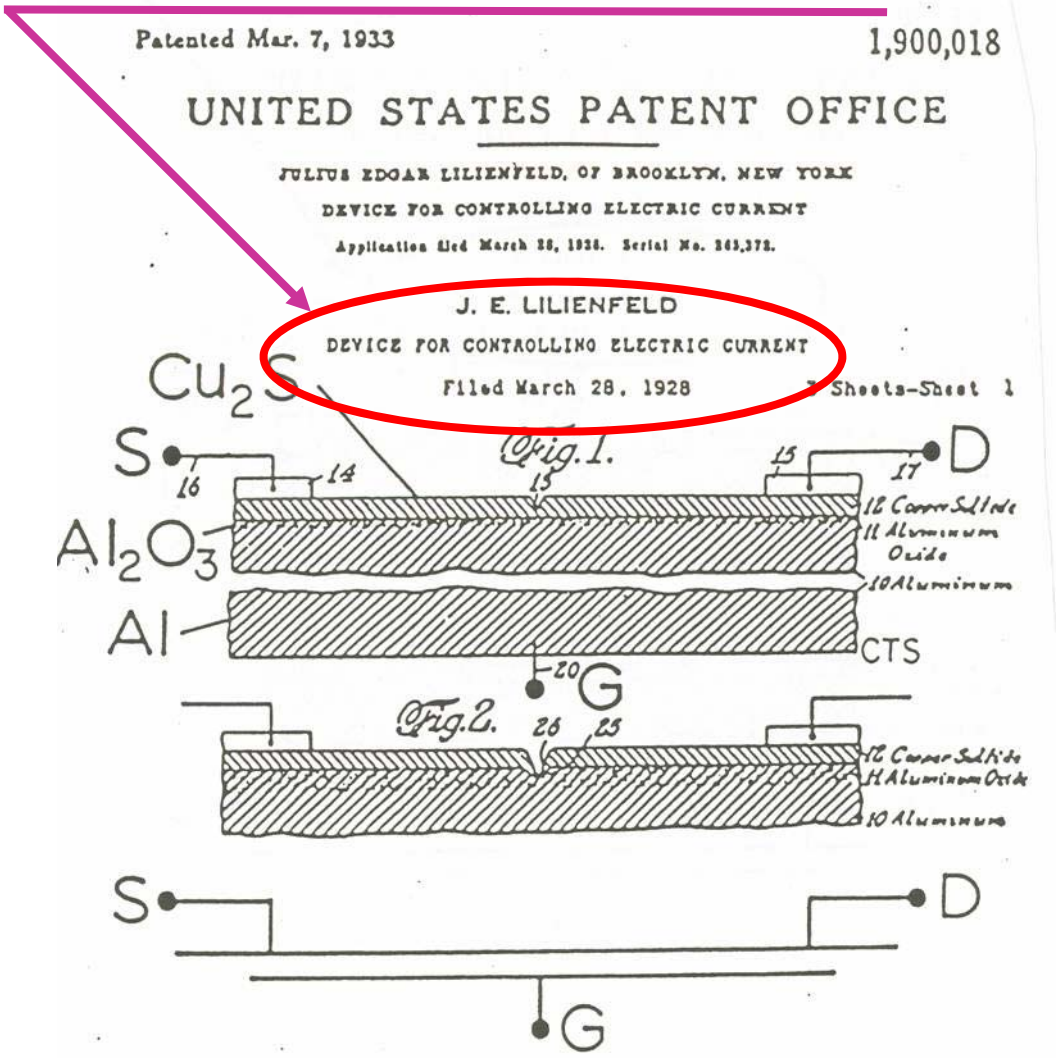
History of Electronic Devices



J. E. LILIENFELD

DEVICES FOR CONTROLLED ELECTRIC CURRENT

Filed March 28, 1928



Patented Mar. 7, 1933 1,900,018

UNITED STATES PATENT OFFICE

JULIUS EDGAR LILIENFELD, OF BROOKLYN, NEW YORK

DEVICE FOR CONTROLLING ELECTRIC CURRENT

Application filed March 28, 1928. Serial No. 263,372.

J. E. LILIENFELD

DEVICE FOR CONTROLLING ELECTRIC CURRENT

Filed March 28, 1928 Sheets-Sheet 1

Cu_2S

S 16 14 15 15 17 D

Al_2O_3

Al

CTS

20 G

Fig. 1.

Fig. 2. 26 25

S D

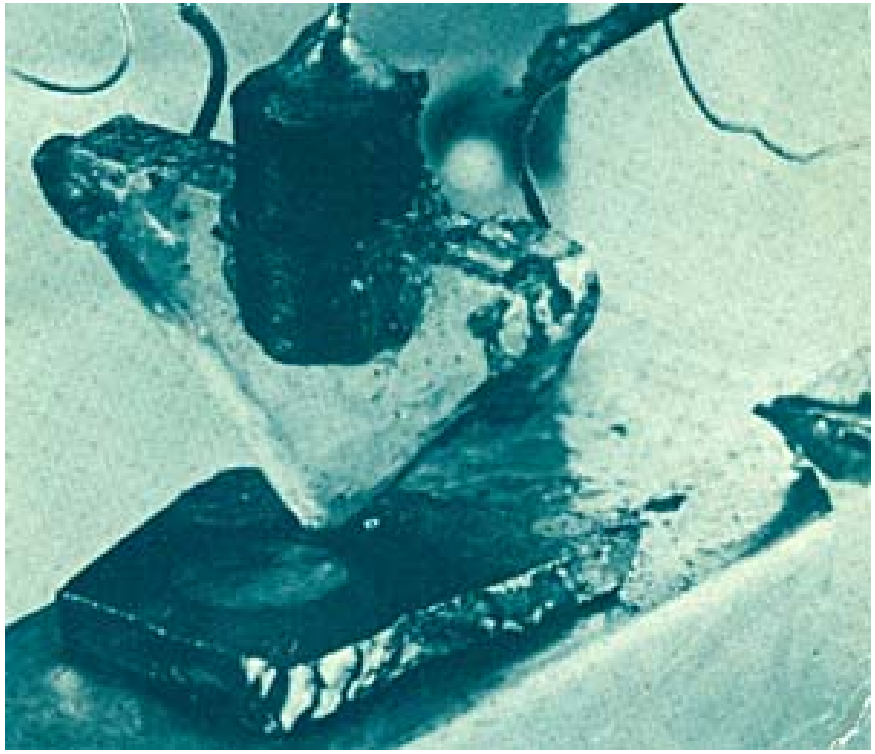
G

J.E.LILIENFELD

1947: 1st transistor

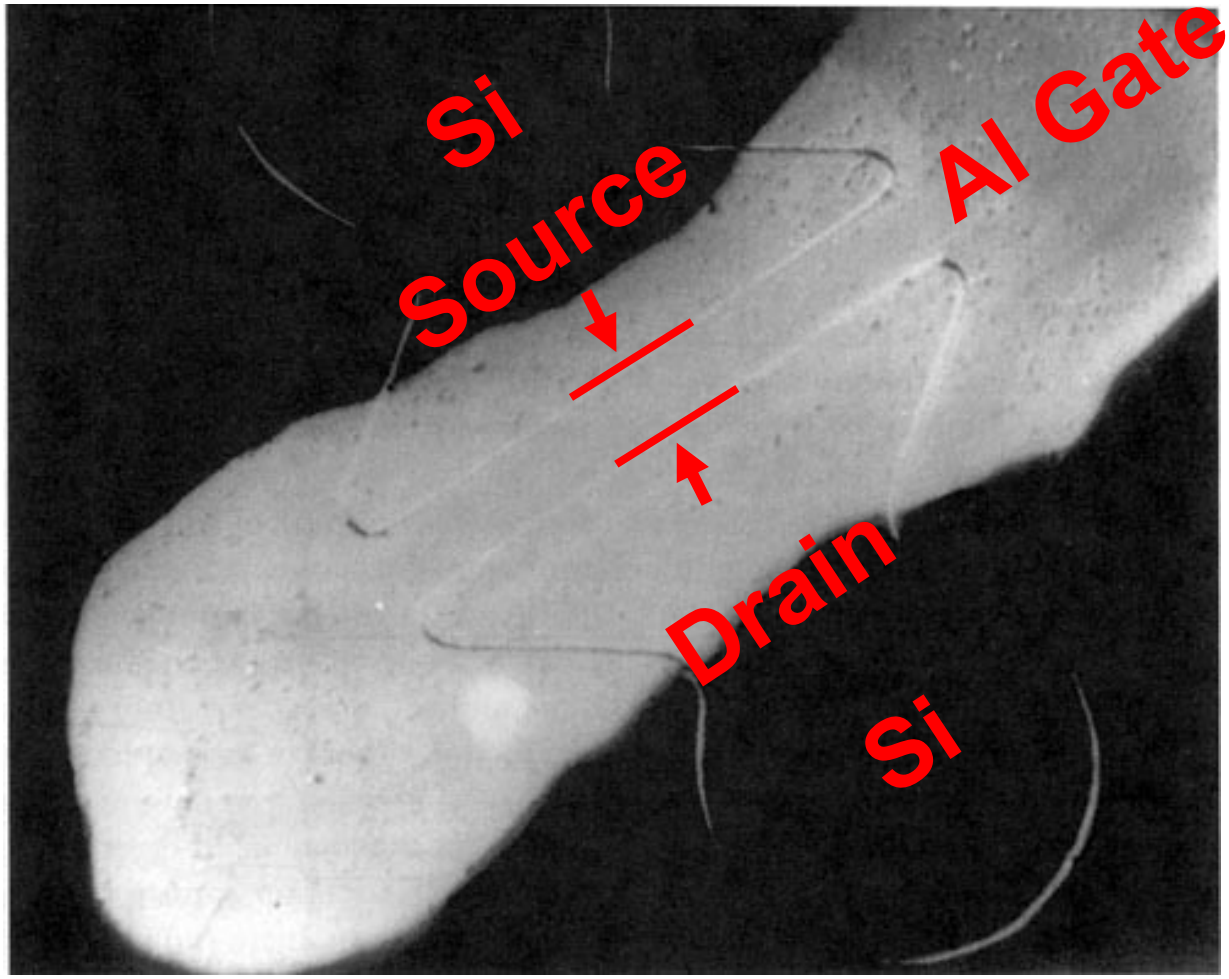
John Bardeen, Walter Brattain, William Shockley

Bipolar using Ge

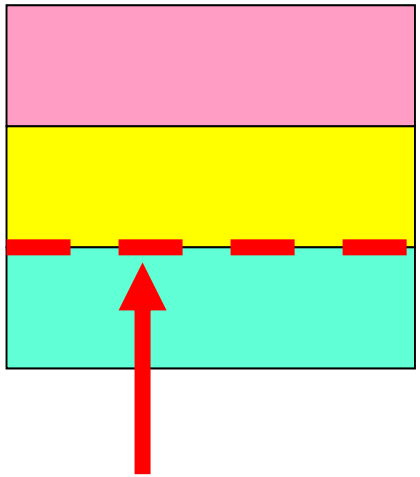


1960: First MOSFET by D. Kahng and M. Atalla

Top View



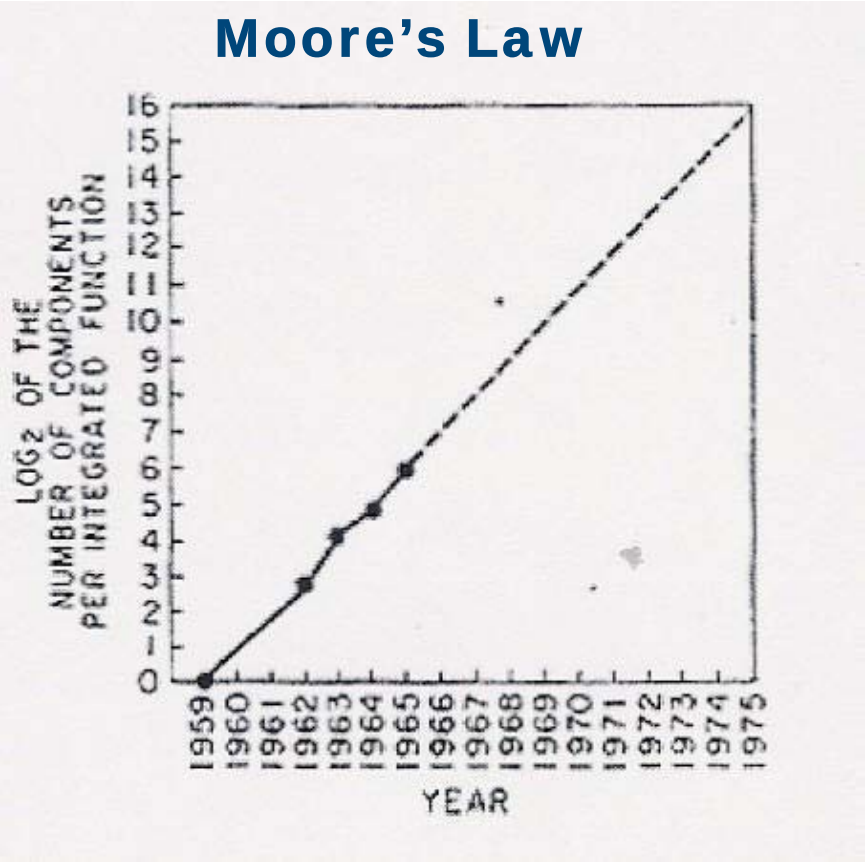
Cross-section



Exceptionally good interface!

Technology for electronic systems: from milli- to micro- to nano-

Moore's Law



Cramming more components onto integrated circuits

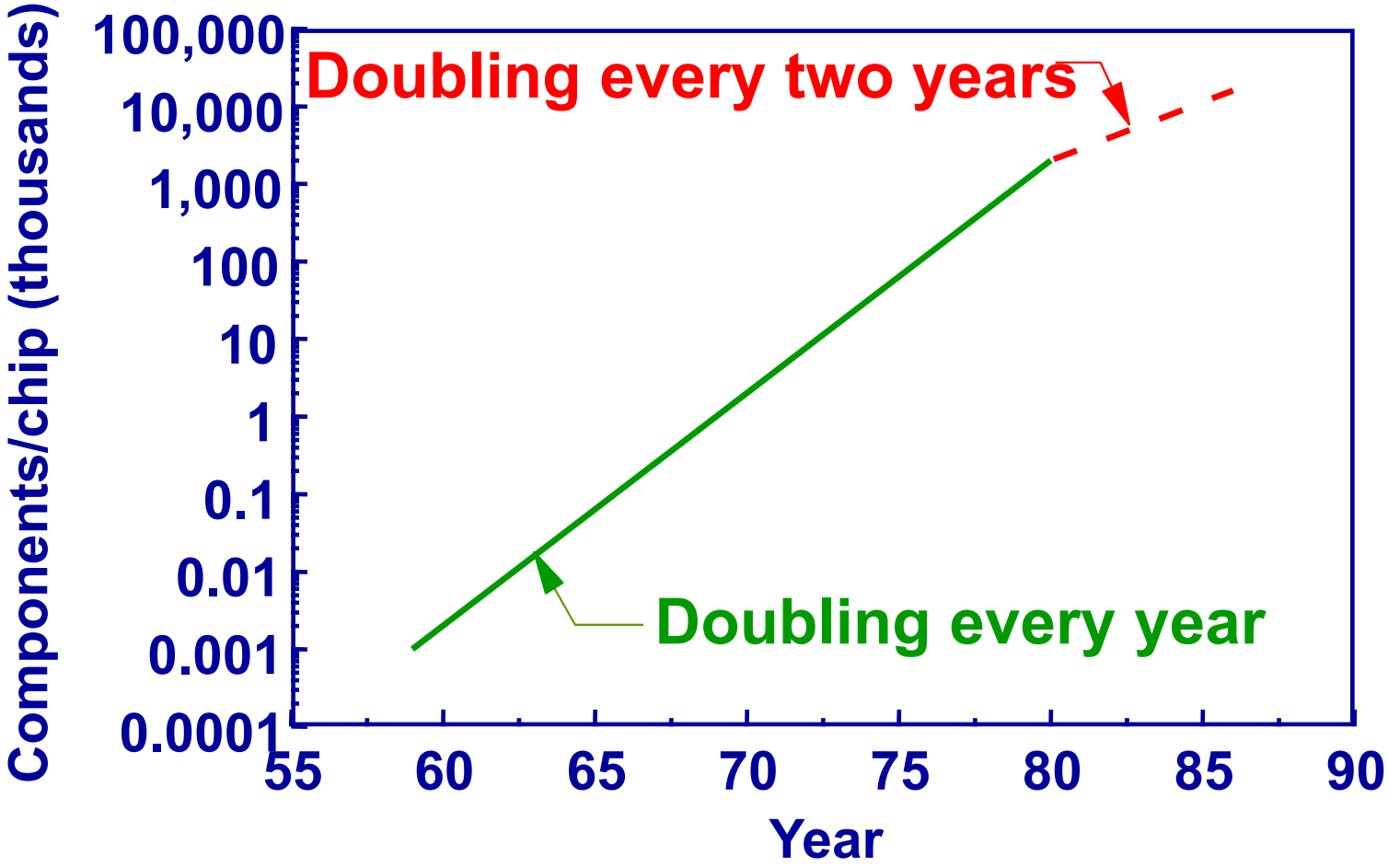
With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

Electronics, Volume 38, Number 8, April 19, 1965

Moore's Law



After G. Moore, 1975 IEDM

Why we scale

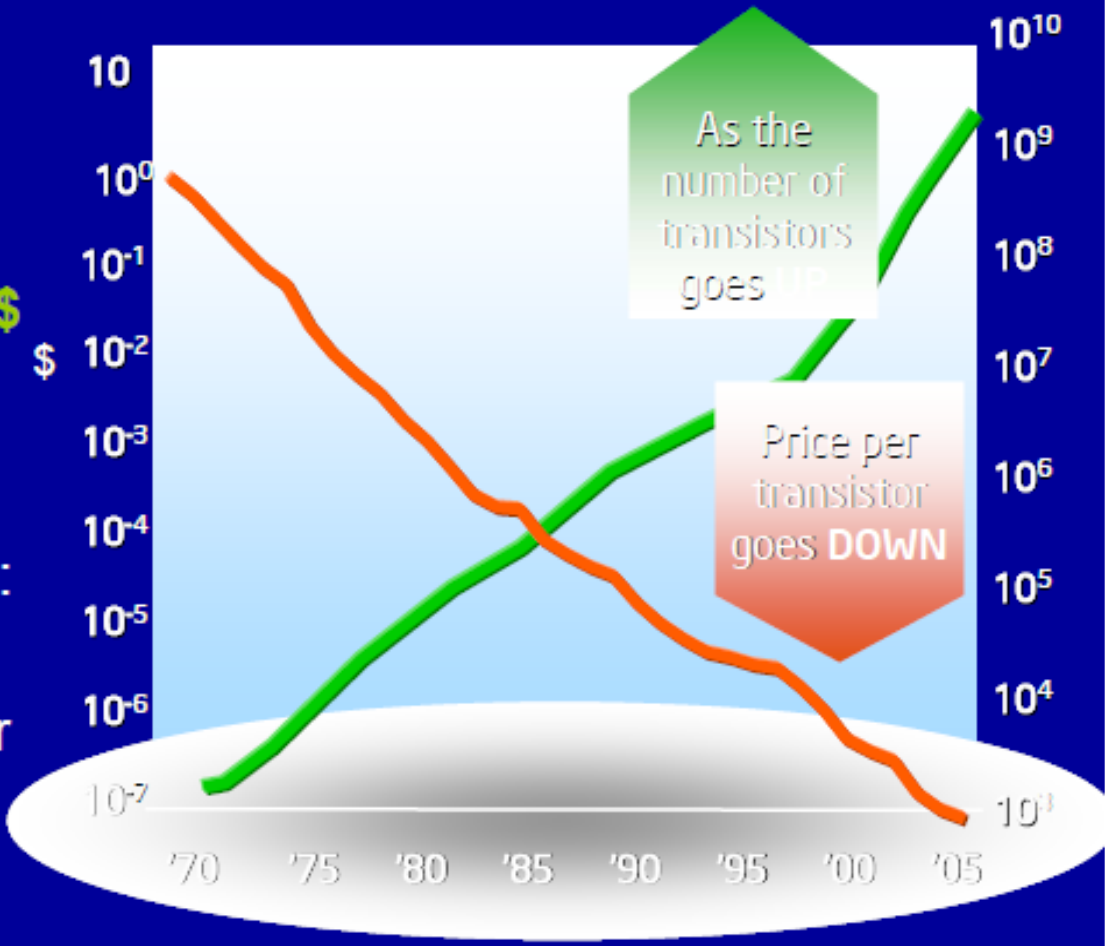
Moore's law

> *"The complexity for minimum component costs has increased at a rate of roughly a factor of two per year"*
>
> **G. Moore, *Electronics* '65**

- **functionality at minimum $**
- **not a direct scaling prediction**

Translation into device Scaling:

- more transistors per area
- more functionality and lower cost per function
- Enables new products

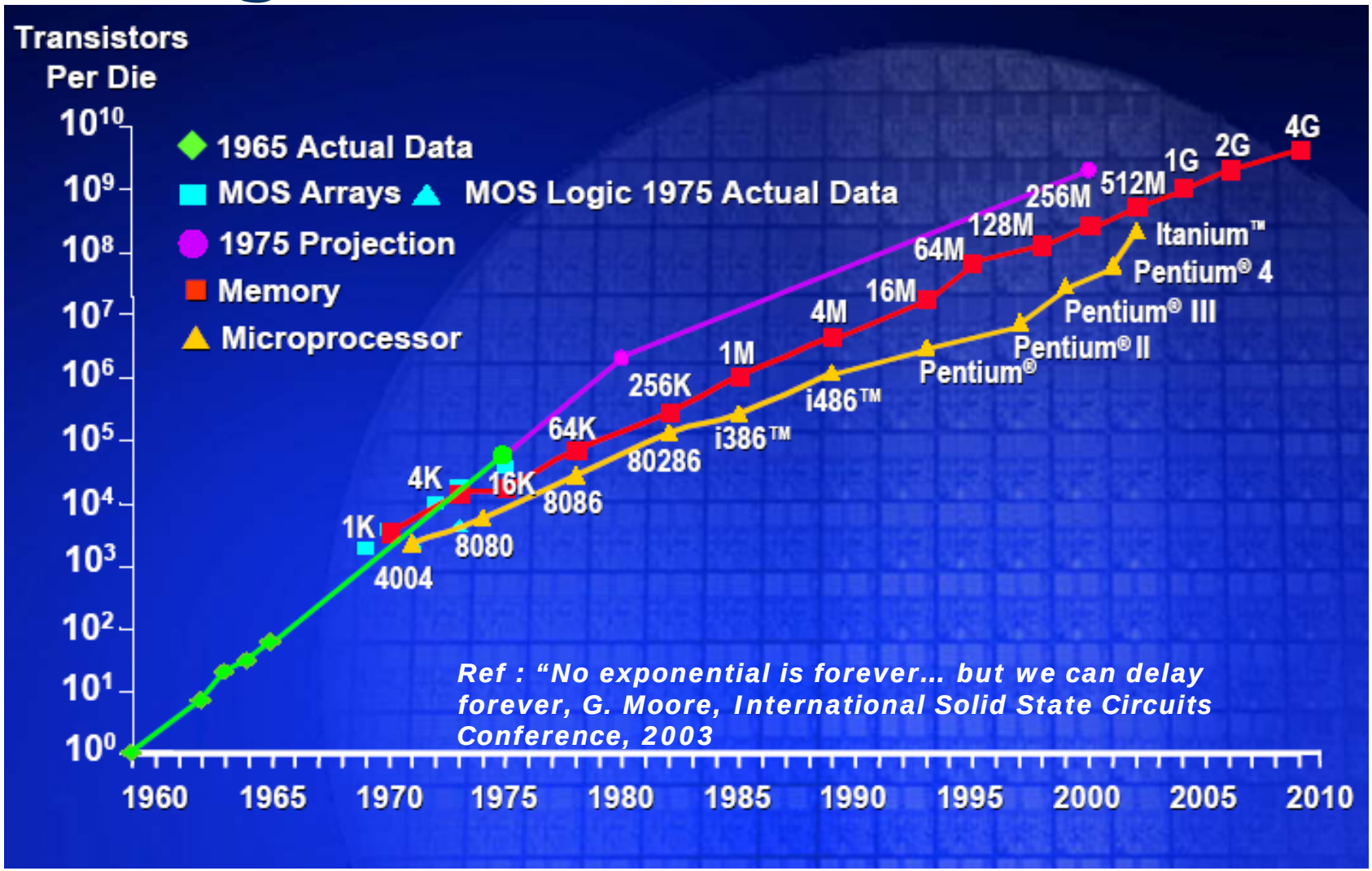


Source: WSTS/Dataquest/Intel

Nanotechnology and Nanoelectronics

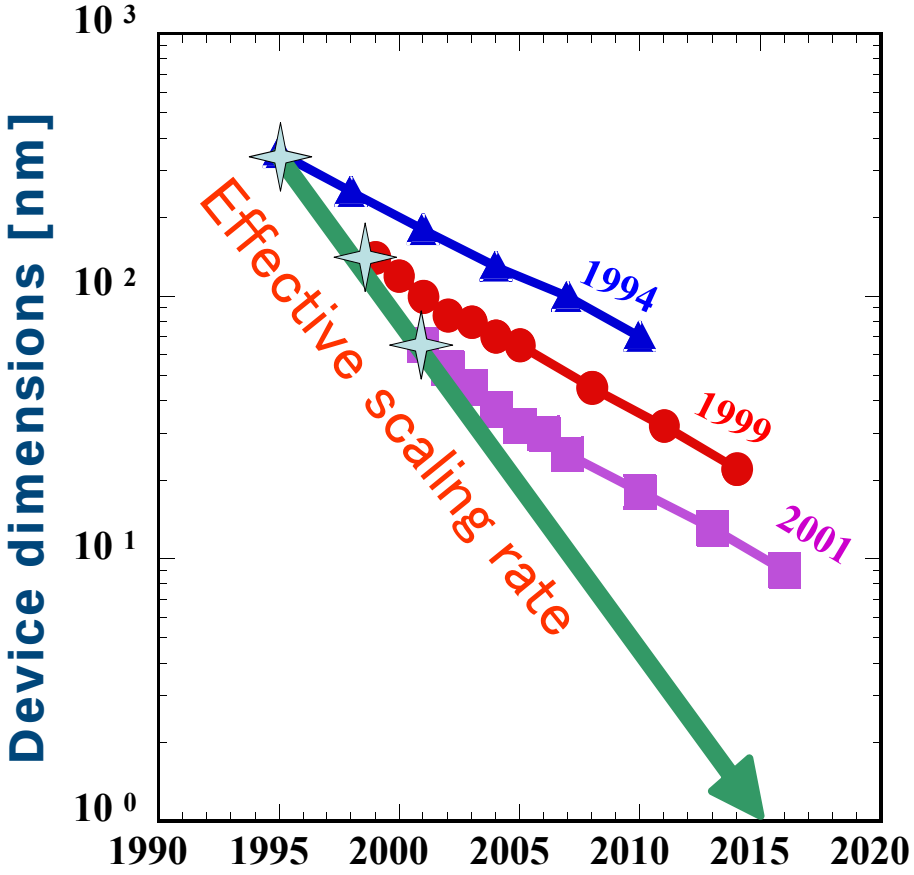
Enrico Sangiorgi

Integration and Moore's Law



The Bad :
complexity, roadblocks and physical limits

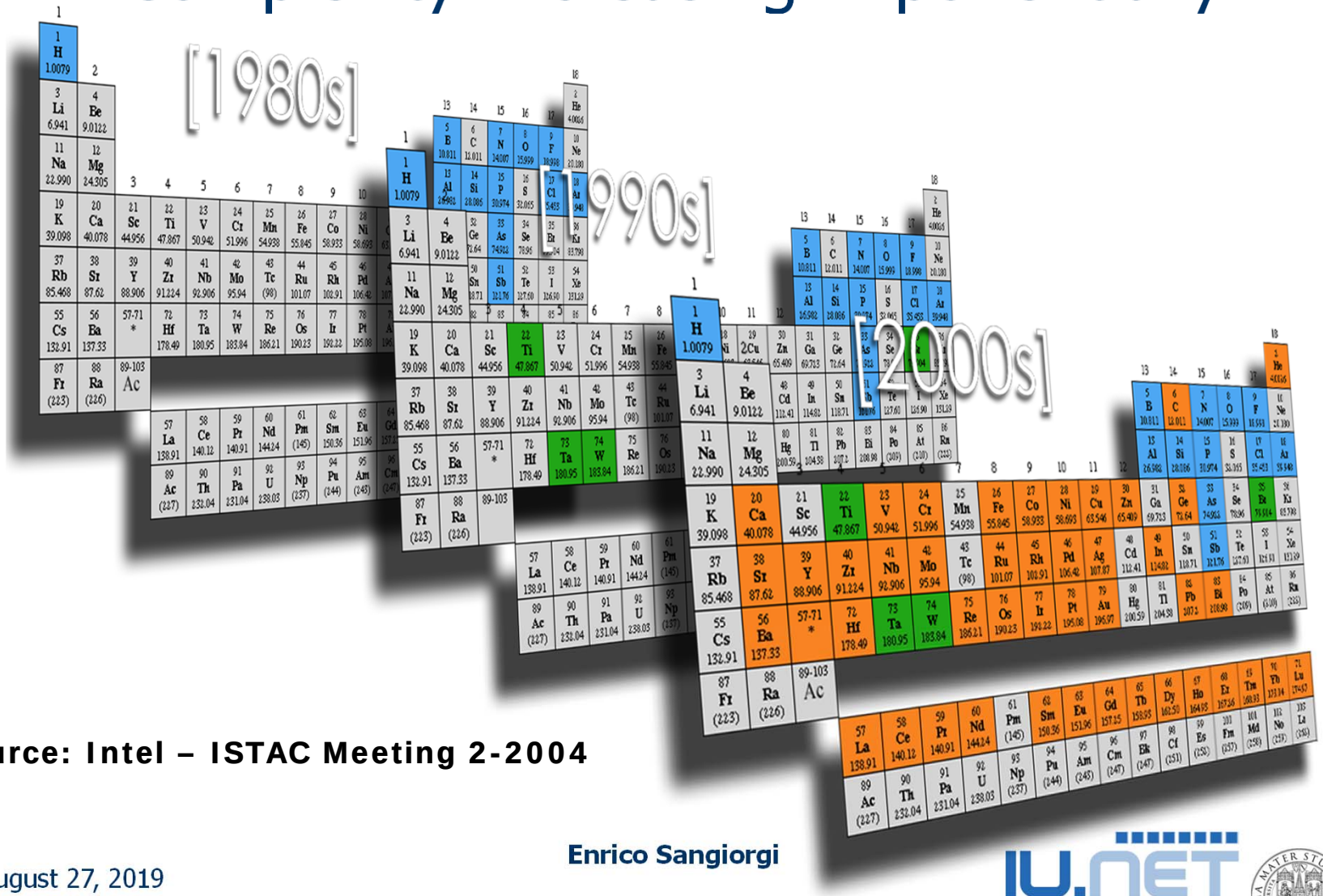
Complexity and acceleration



Silicon-Germanium
Metal gates
High-k dielectrics
Low-k dielectrics
Copper interconnects
Tungsten plugs
Cobalt silicide
Titanium silicide
Nichel silicide
BPSG
Polysilicon
Silicon nitride
Oxinitrides
Indium
Arsenic
Boron
Phosphorous
Aluminum
SiO2
Silicon
2000

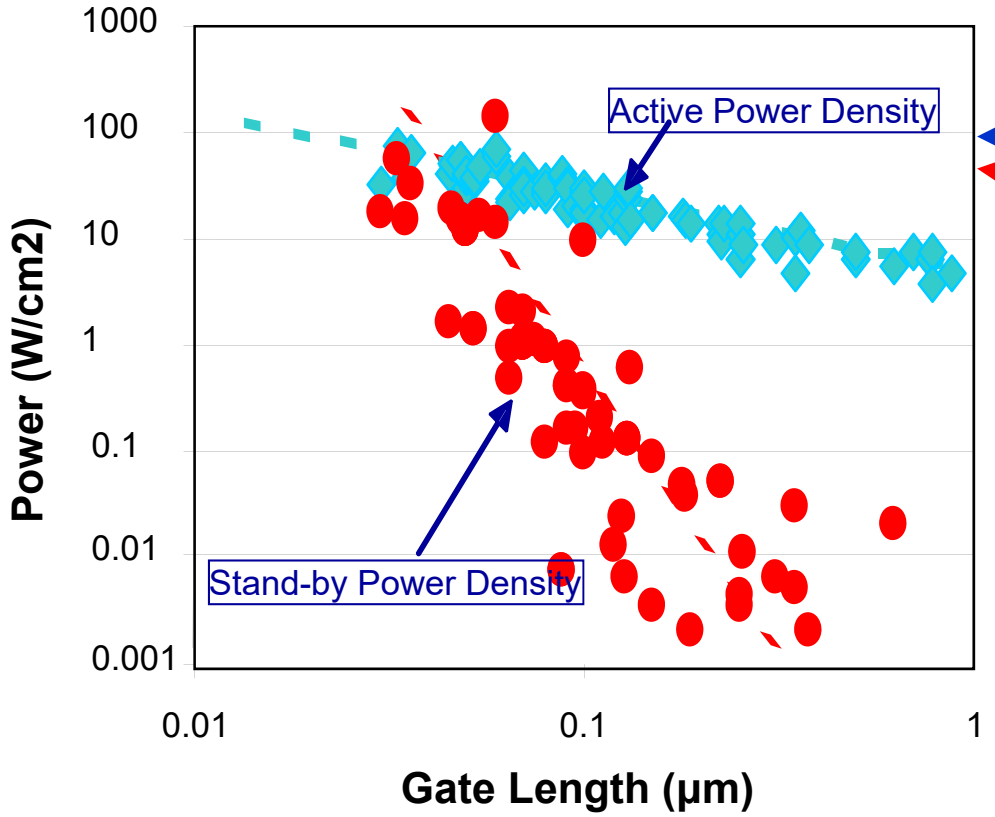
Boron
Phosphorous
Aluminum
SiO2
Silicon
1970

Si Technology: Complexity Increasing Exponentially



Source: Intel – ISTAC Meeting 2-2004

Roadblock: Power Density



Air Cooling Limits (W/cm2)

90 – Unipocessor

60 – Multi Processor

The Blind Race for GHz is Over

Intel Tejas Samples Dissipate 150W of Heat at 2.80GHz

10 Working CPUs Available Now...

by Anton Shilov
01/11/2004 | 12:00 PM

The
in T
mic
arc
Q1

Heat forces Intel to cancel desktop chips, says report

By Peter Clarke Silicon Strategies
05/07/2004, 8:11 AM ET

LONDON -- Intel Corp. plans to announce Friday (May 7) that it has cancelled its next chips for desktop and server computers, codenamed Tejas and Jayhawk, according to a *Reuters* report that cited an un-named source.

The move represents a significant shift in Intel's development plans and a desire to build chips that are computationally powerful without generating excessive amounts of heat, the report said.

The chips to be cancelled include a version of Intel's fourth-generation Pentium 4 chip, code-named Tejas, which was expected to come to market in 2005, and a Xeon processor for low-end computer servers, code-named Jayhawk, the report said referencing its un-named source.

Intel is expected to explain on Friday that it plans to use processors designed for mobile computers for future desktop and server applications, the report said.

Physical limits and switch performance

- Control electrostatics
- Control power dissipation
- Control (Improve) conductivity
- Minimize parasitics

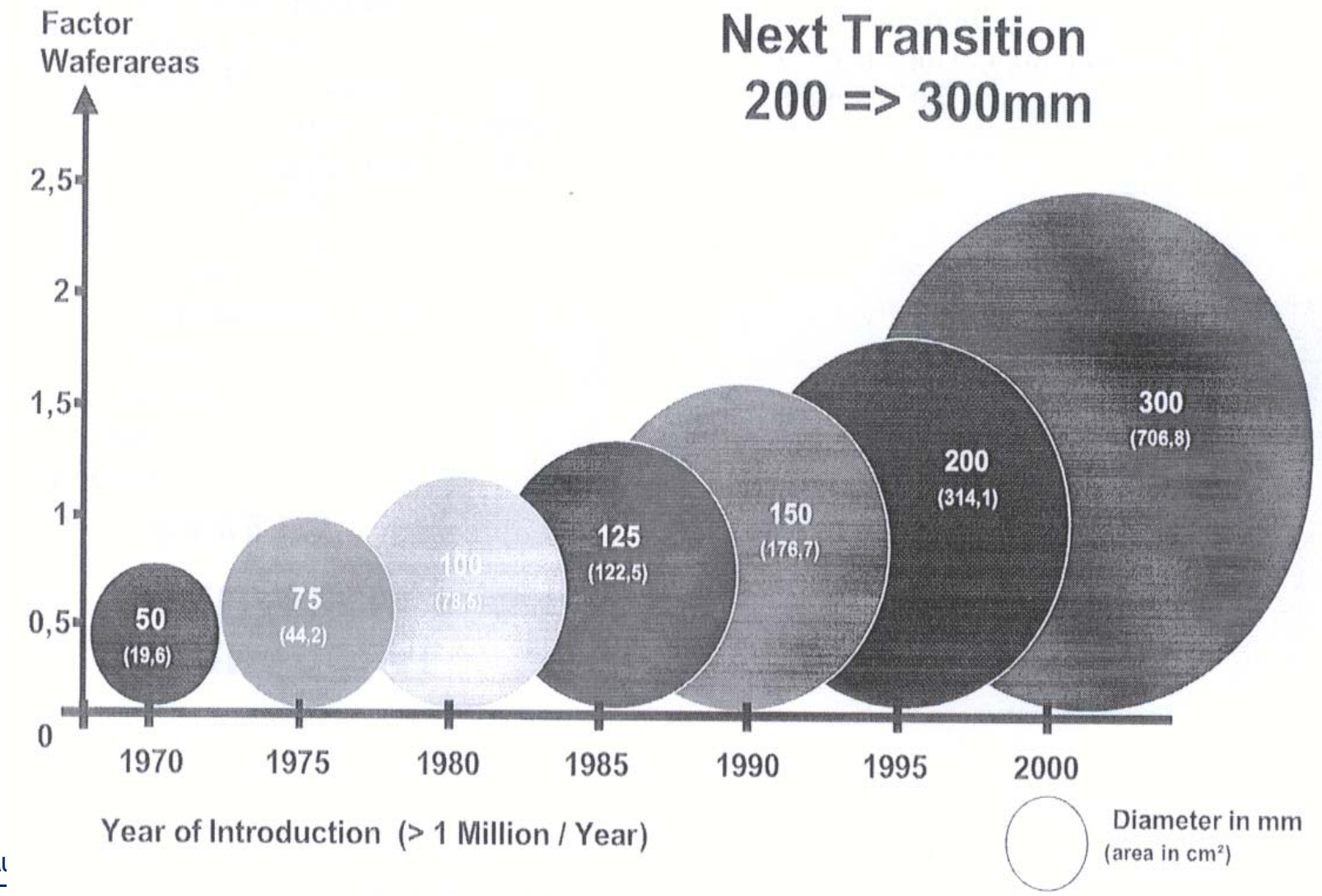
Predictions of Scaling limits since IC started

Period	Expected limit	Cause
Late 1970's	1μm:	SCE
Early 1980's	0.5μm:	S/D resistance
Early 1980's	0.25μm:	Direct-tunneling of gate SiO2
Late 1980's	0.1μm:	Several
Today	50nm:	Several
Today	<10nm:	Fundamental?

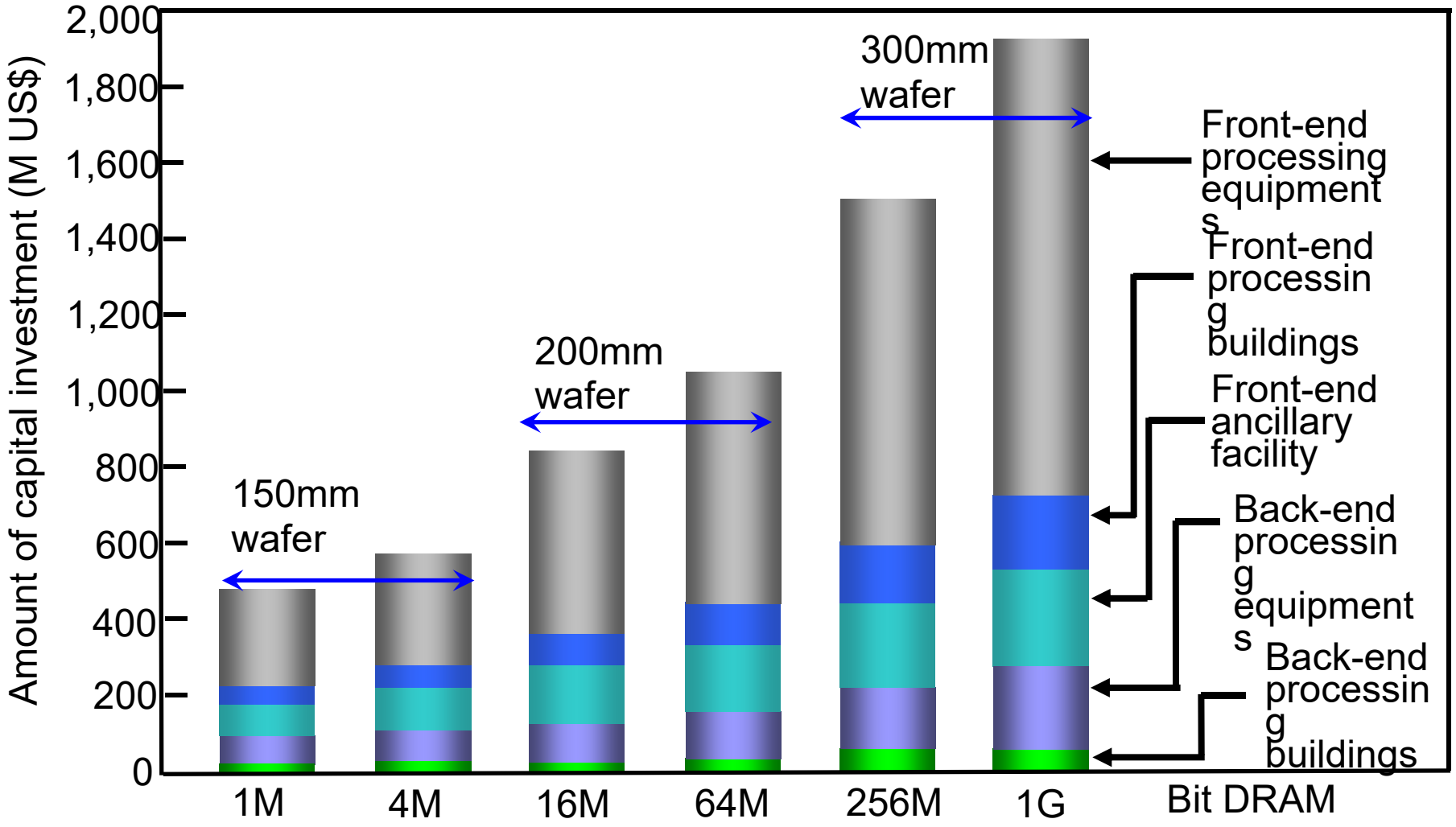
Fundamental: electron wave length (10nm), tunneling distance(3nm), atom distance (0.3nm)

The Ugly : costs

Wafer Dimensions



Amount of capital investment on Each of DRAM generation



Reference : Semiconductor Industry Research Institute Japan

	2012	2013	2014	2015	2016	2017
Semiconductor Capital Spending ($M)	58,742.8	56,704.5	64,745.6	71,305.9	68,790.4	72,399.6
Growth	-11.9%	-3.5%	14.2%	10.1%	-3.5%	5.2%
Capital Equipment ($M)	37,833.2	35,761.6	42,591.0	47,488.8	44,712.0	48,580.9
Growth	-16.1%	-5.5%	19.1%	11.5%	-5.8%	8.7%
Wafer-Level Manufacturing Equipment ($M)	31,445.8	29,900.7	35,293.4	40,400.0	38,867.7	42,179.1
Growth	-17.8%	-4.9%	18.0%	14.5%	-3.8%	8.5%
Wafer Fab Equipment ($M)	29,644.2	27,957.3	32,831.5	37,750.5	36,344.4	39,215.4
Growth	-18.5%	-5.7%	17.4%	15.0%	-3.7%	7.9%
Wafer-Level Packaging and Assembly Equipment ($M)	1,801.6	1,943.4	2,461.9	2,649.5	2,523.3	2,963.7
Growth	-3.1%	7.9%	26.7%	7.6%	-4.8%	17.5%
Die-Level Packaging and Assembly Equipment ($M)	3,867.3	3,503.7	4,258.9	3,922.5	3,232.1	3,548.2
Growth	-10.5%	-9.4%	21.6%	-7.9%	-17.6%	9.8%
Automated Test Equipment ($M)	2,520.0	2,357.2	3,038.7	3,166.3	2,612.2	2,853.5
Growth	0.4%	-6.5%	28.9%	4.2%	-17.5%	9.2%
Other Spending ($M)	20,909.6	20,943.0	22,143.3	23,815.1	24,401.2	24,067.9
Growth	-3.1%	0.2%	5.7%	7.6%	2.5%	-1.4%

Source: Gartner (June 2013)

OUTLINE OF THE COURSE
CLASS 1: Device Simulation

- **Electrostatic and electrodynamics in mesoscopic devices: Schroedinger and Boltzmann equations**
- **The semi-classical approach to model charge transport**
- The method of moments
- The Monte Carlo method

Device simulation and Quantum Mechanics: setting the stage

- At microscopic level (dimensions comparable with atom size) electrons are described by wave functions whose frequency and wave length are related to the particle energy and momentum
- Momentum and position of a particle are known within a given uncertainty (uncertainty principle)

- Electrons are Fermions and thus obey the Pauli exclusion principle: only one electron can be found in a given state (allowed state)
- Energy and momentum of allowed states are solution of the Schrödinger equation

The Schrödinger equation

- The Schrödinger equation in the effective mass approximation and for spherical, parabolic bands

$$j\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m^*}\nabla^2\Psi(\vec{r},t)+\left[E_{C0}(\vec{r},t)+U(\vec{r},t)+U_S(\vec{r})\right]\Psi(\vec{r},t)$$

- **E_{C0}**: conduction band; **U**: potential due to external fields and charge distribution, **Ψ**: envelope wave function; **U_S:** the scattering potential associated to local perturbations due to impurities, lattice vibrations, other particles

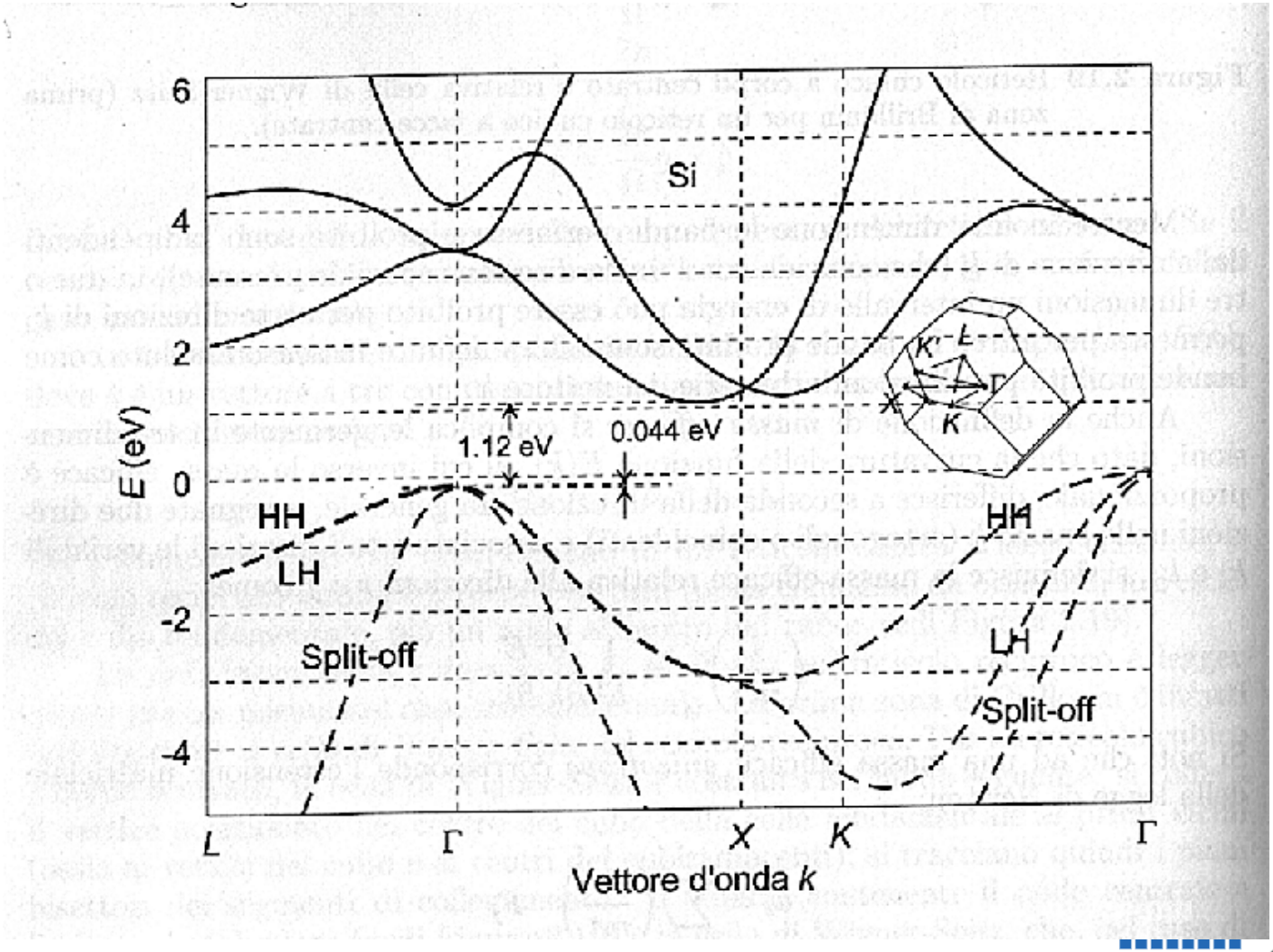
- Electrons in semiconductors are well described by a superimposition of different wave functions, each with defined energy and momentum, and with different weight (wave packet).
- The wave packet energy and momentum are also solution of the Schrödinger equation

- The wave packet is reasonably described by a single value of the crystal momentum **P**=ħ**K** and position **r** corresponding to the centroid of the packet.
- The propagation velocity of the wave packet (group velocity) is given by:

$$\vec{v}_g = \frac{d\vec{r}_{wp}}{dt} = \frac{1}{\hbar}\nabla_{\vec{k}} E(\vec{k})_{\vec{k}=\vec{k}_{wp}}$$

- E(**K**) is called energy-momentum dispersion relation and provides the allowed energy values for each value of the crystal momentum **K,** once the Schrödinger equation is applied to the periodic crystal potential. E(**K**) depends on the material and it is periodic with period **2π/L** (reciprocal reticule) where L is the periodic distance of the crystal in **r** (crystal reticule).

View of the Si bands



- The effects of the periodic crystal potential are included by adopting the E(**k**) dispersion relation.
- Therefore the motion of the particle is described by a single wave packet, with given **k** and **r**, subject to the external forces and to the interaction with the crystal lattice (scattering events).

The semi-classical approach to charge transport

The theoretical frame described above is called "semiclassical approach" to device modeling: each carrier behaves like a classical, electrically charged particle, reacting to the field in accordance to Newton's law, and interacting with lattice vibrations (phonons), defects, or impurities. Quantum mechanics is used only to describe collisions.

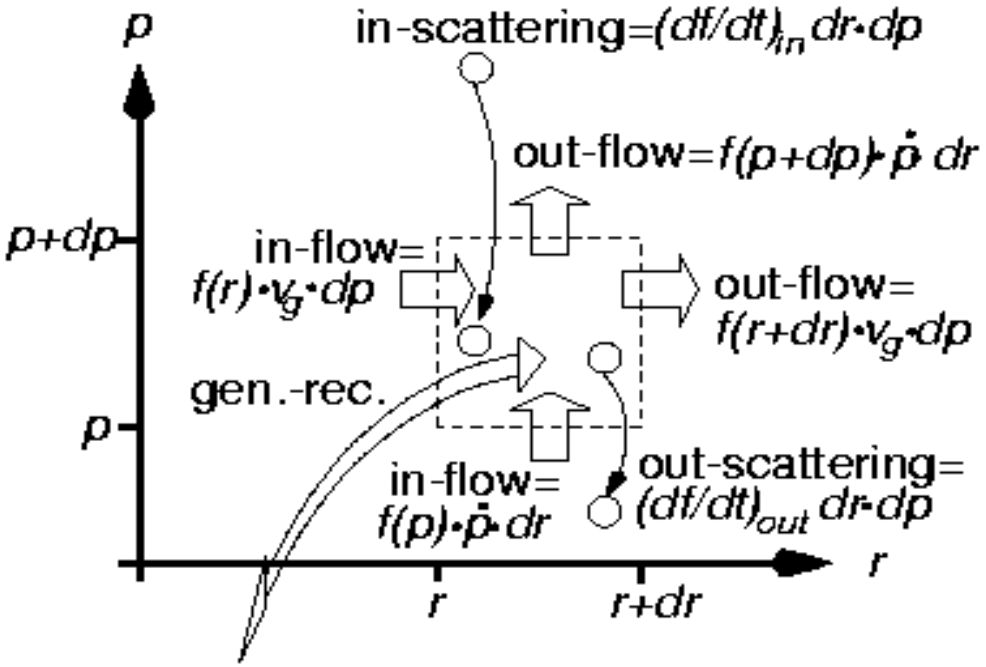
The Botzmann Transport Equation

Carriers are statistically described through a distribution function that gives the probability of occupation of a state with wave-vector *k* and position *r* at time t. The distribution function is the solution of the Boltzmann Transport Equation (BTE) that describes the classical motion of carriers under the action of the electric field *F* and perturbation:

$$\frac{\partial \mathcal{F}}{\partial t} = -\nabla_{\mathbf{r}} \cdot \left(\frac{d\vec{\mathbf{r}}}{dt} \mathcal{F} \right) - \nabla_{\mathbf{p}} \cdot \left(\frac{d\vec{\mathbf{p}}}{dt} \mathcal{F} \right) + \left(\frac{\partial \mathcal{F}}{\partial t} \right)_{C}$$

Where $p=\hbar k$ represents the carrier momentum and F(p,r,t) is the distribution function. The BTE expresses the conservation of particles in position and momentum space as described next.

Boltzmann Transport Equation



$$\frac{\partial \mathcal{F}}{\partial t} = -\nabla_{\mathbf{r}} \cdot \left(\frac{d\vec{r}}{dt} \mathcal{F} \right) - \nabla_{\mathbf{p}} \cdot \left(\frac{d\vec{p}}{dt} \mathcal{F} \right) + \left(\frac{\partial \mathcal{F}}{\partial t} \right)_C$$

Solution of the BTE

- The solution of the BTE for realistic devices is a very difficult task
- Difficult arises from:
  - Non homogeneous device structure (doping and topology
  - Models for collision mechanisms
  - Complexity of the band structure
- Most common solutions: statistical methods (Monte Carlo) or drastic approximations (methods of moments)

Approximations behind the BTE

- Statistical Description using Single Particle Probabilities (Number of Electrons in the active device?)
- Classical Treatment of Motion + QM Scatterings
- Scatterings occur instantly in time and localized in space (interaction with macroscopic fields are neglected during the scattering)
- Fermi Golden Rule for Scattering Rates

OUTLINE OF THE COURSE
CLASS 1

- Electrostatic and electrodynamics in mesoscopic devices: Schroedinger and Boltzmann equations
- The semi-classical approach to model charge transport
- **The method of moments**
- The Monte Carlo method

Moments of the BTE

- Approximation method based on the moments of BTE:
  - Reduction of the number of dimensions obtained substituting the BTE (in $\mathcal{F}$) with a number of equations based on the moments of $\mathcal{F}$ on $\vec{p}$.
  - Approximation of the collision term (Relaxation Time Approximation) into a single parameter τ.

- The first three moments of $\mathcal{F}$ with respect to $\vec{p}$:

$$n(\vec{r},t) = \int_{\vec{p}} \mathcal{F}(\vec{r},\vec{p},t)d\vec{p}$$

$$\vec{P} = \int_{\vec{p}} \vec{p}\mathcal{F}(\vec{r},\vec{p},t)d\vec{p}$$

$$W = \int_{\vec{p}} \frac{p^2}{2m}\mathcal{F}(\vec{r},\vec{p},t)d\vec{p}$$

bear the following physical meaning: n represents the electron concentration, $\vec{P}/n$ the average electron momentum, W/n the average electron energy in the parabolic approximation $p^2/2m = E$.

- The new equations are obtained from BTE by multiplying it by the corresponding power of $\vec{p}$ and then integrating in $\vec{p}$.
- Several physical/empirical approximations are needed in order to obtain a practical system of equations

Drift-diffusion model

- Variables: $n, p, \phi, \vec{J}_n, \vec{J}_p$
- Continuity equations:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n - U$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p - U$$

- Current density equations:

$$\vec{J}_n = -q\mu_n n \nabla \phi + qD_n \nabla n$$

$$\vec{J}_p = -q\mu_p p \nabla \phi - qD_p \nabla p$$

- Poisson's equation:

$$\nabla \cdot (\epsilon \nabla \phi) = -\rho = -q(p - n + N_D^+ - N_A^-)$$

- Parameters: mobility, diffusion coefficient, recombination function ($U = R - G$)

Mobility: the key parameter of the drift-diffusion model

Drift current and mobility (I)

- In the presence of a homogeneous electric field, electrons and holes are subjected to a coulomb force that tends to direct their velocity along the electric filed direction.

$$\vec{J}_{p,drift} = qp\vec{v}_{p,drift}$$

$$\vec{J}_{n,drift} = -qn\vec{v}_{n,drift}$$

- $\vec{J}$: current density vector; $\vec{v}_{drift}$: drift velocity vector;

Drift current and mobility (II)

- by assuming that the average electron at each scattering event loses all the kinetic energy acquired by the field in the time interval between scattering (τ_{sc}):

$$-q\vec{F}\tau_{sc} = m^*\vec{v}_{drift}$$

$$\mu_n = |v_{drift}|/|F| = \frac{q\tau_{sc}}{m^*}$$

- for low electric field values, the drift velocity is proportional to the electric field, and the proportionality constant is called mobility
- $\vec{v}_{p,drift} = \mu_p\vec{F}$, $\vec{v}_{n,drift} = -\mu_n\vec{F}$

$$\vec{J}_{p,drift} = qp\mu_p\vec{F}$$

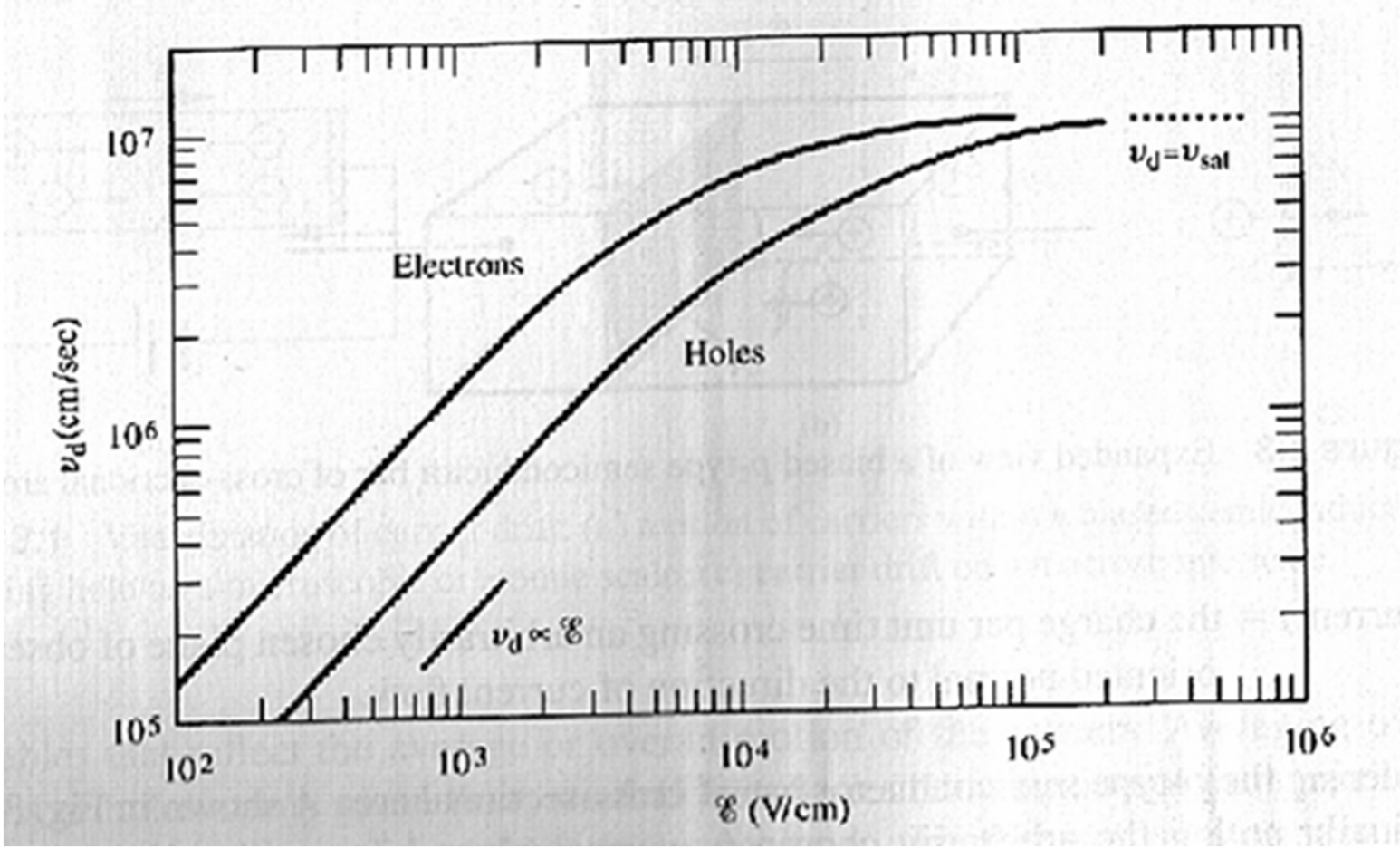
$$\vec{J}_{n,drift} = qn\mu_n\vec{F}$$

Mobility

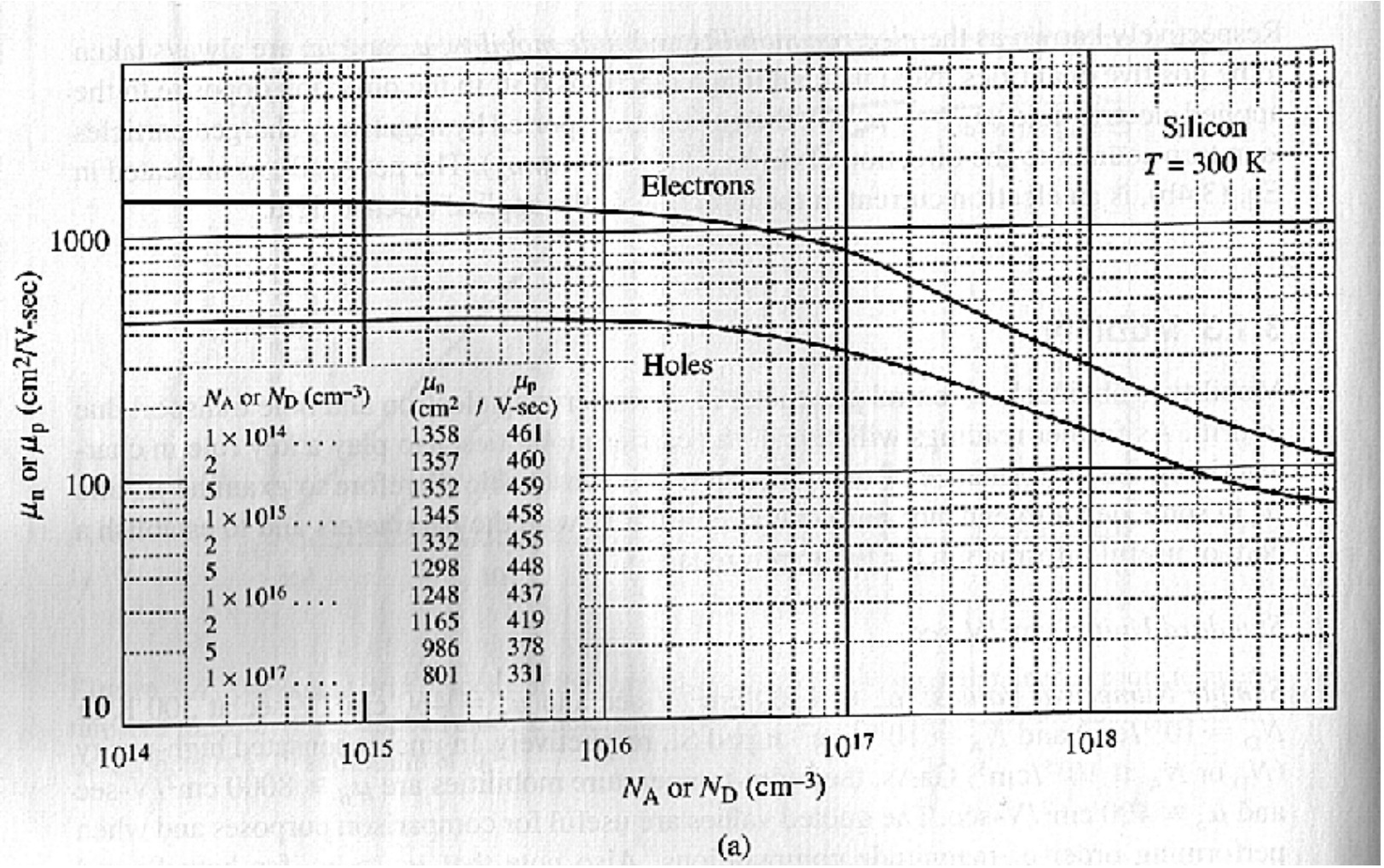
- Main dependencies : material, carrier type, dopant concentration, temperature, field.
- Limited by scattering: lattice phonons (μ decreases when temperature increses); scattering with phonons ccauses exchange of energy with the lattice.
- Impurities (neutral or ionized) perturb the electrostatic potential of the lattice; small change in E, large change in $\vec{p}$.
- A relaxation time τ_{sc} and thus a value of the mobility is associated to each scattering mechanism.
- Concurring scattering events give rise to the mobility composition rule (Mathiessen rule):

$$\frac{1}{\mu} = \sum_i \frac{1}{\mu_i}$$

Drift velocity vs. field

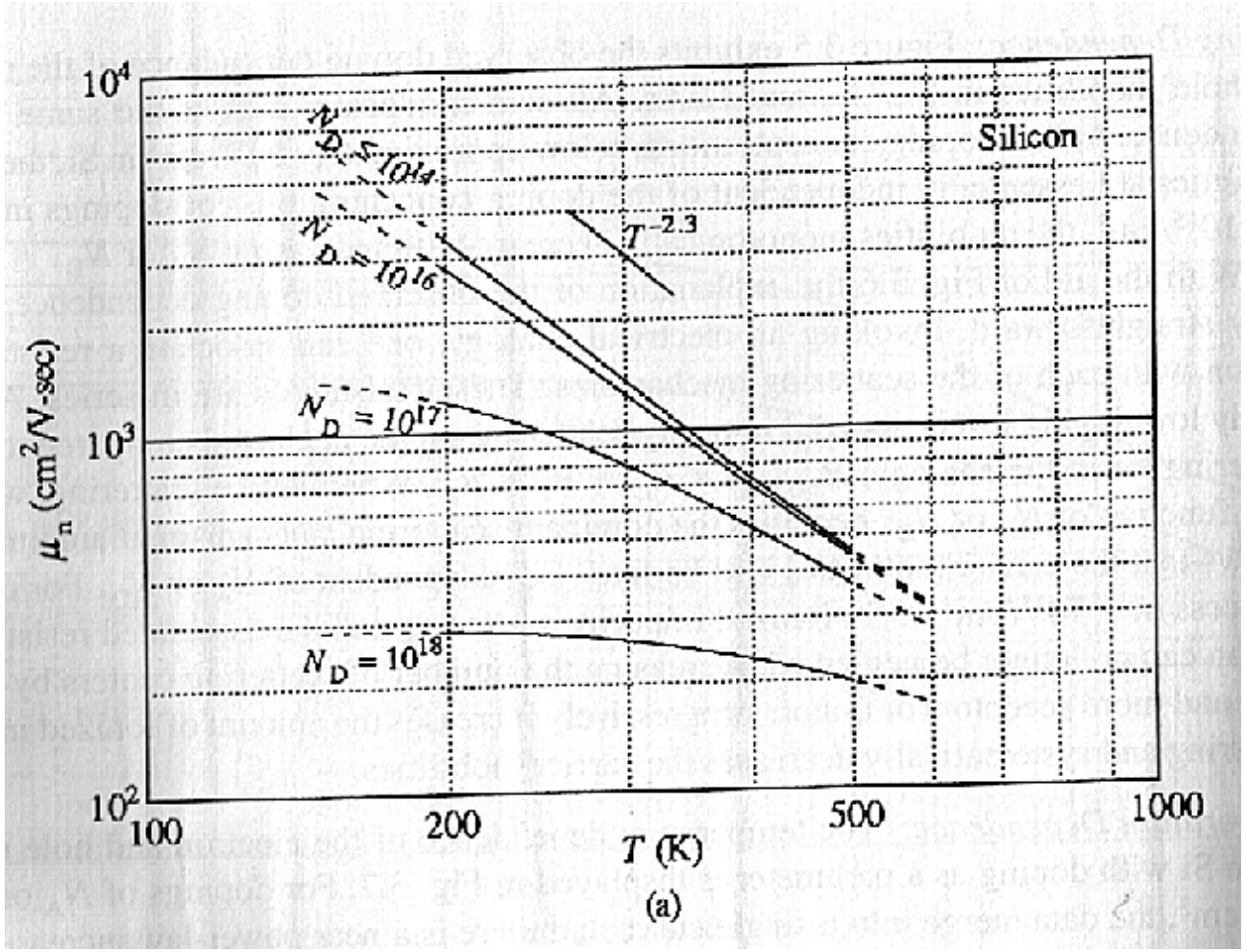


Mobility vs. doping

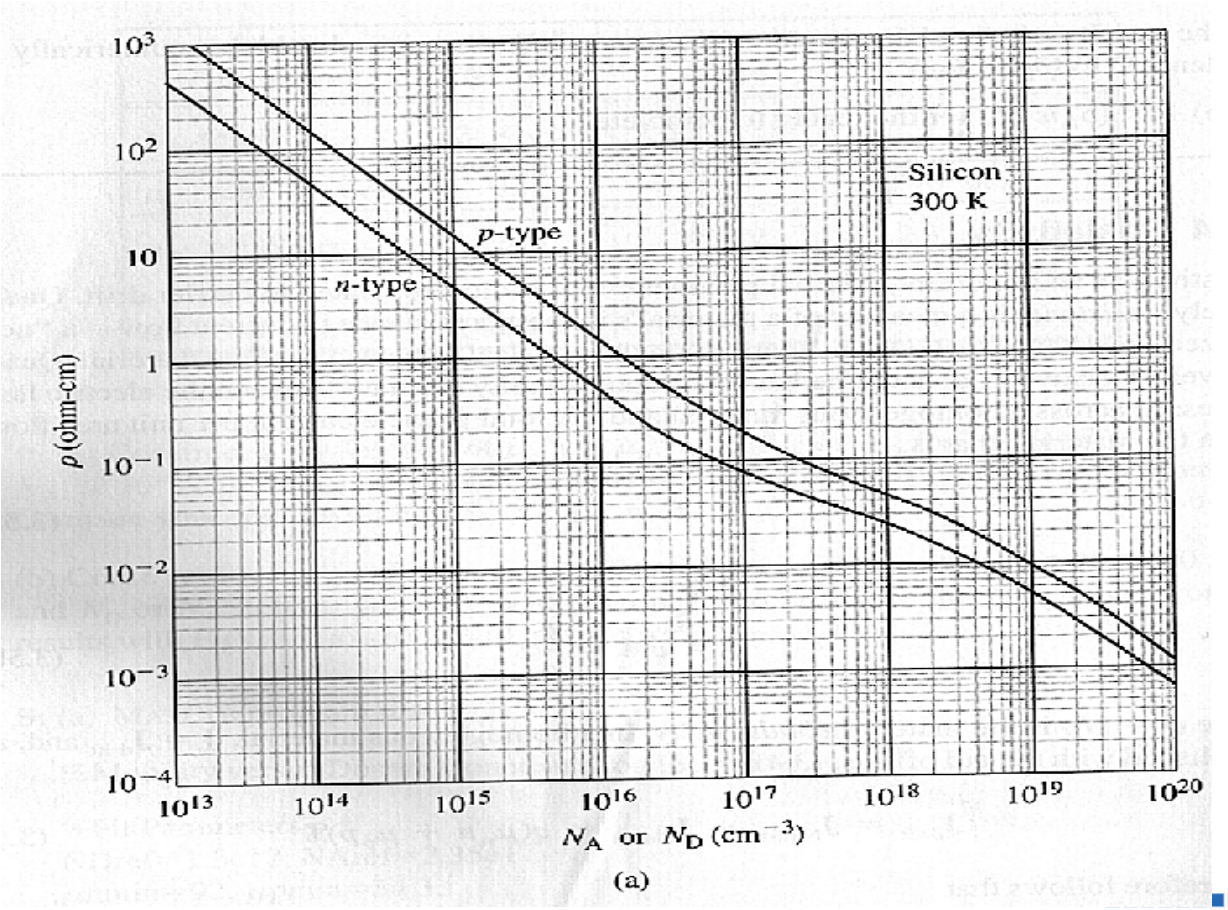


N_A or N_D (cm^{-3})	μ_n (cm^2 / V-sec)	μ_p
1×10^{14}	1358	461
2	1357	460
5	1352	459
1×10^{15}	1345	458
2	1332	455
5	1298	448
1×10^{16}	1248	437
2	1165	419
5	986	378
1×10^{17}	801	331

Mobility vs. Temperature



Resistivity vs. doping



OUTLINE OF THE COURSE
CLASS 1

- Electrostatic and electrodynamics in mesoscopic devices: Schroedinger and Boltzmann equations
- The semi-classical approach to model charge transport
- The method of moments
- **The Monte Carlo method**

Direct solution of the BTE:

The Monte Carlo method (I)

- Il Monte Carlo method is a direct method to solve the Boltzmann transport equation
- The dynamics of each particles is described by the following equation:

$$\frac{d\vec{p}}{dt} = \pm q\vec{F} + R(\vec{r}, \vec{p}, t)$$

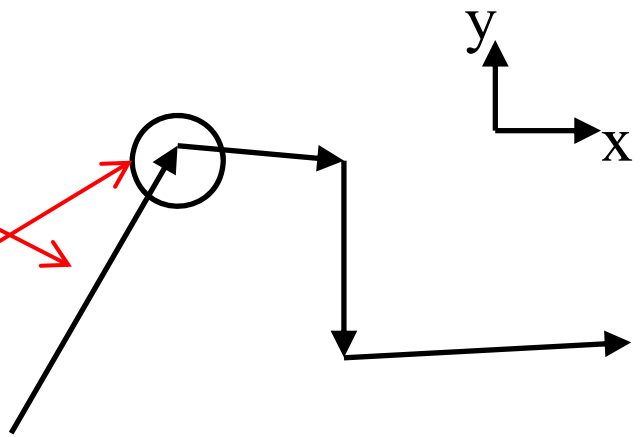
- Where $R(\vec{r}, \vec{p}, t)$ represents the effect *collisions* which modify the carrier dynamics and are due to impurities, lattice vibrations (phonons), etc.

Basic structure of the MC transport

- Motion of sample particles:
  - Free-flights

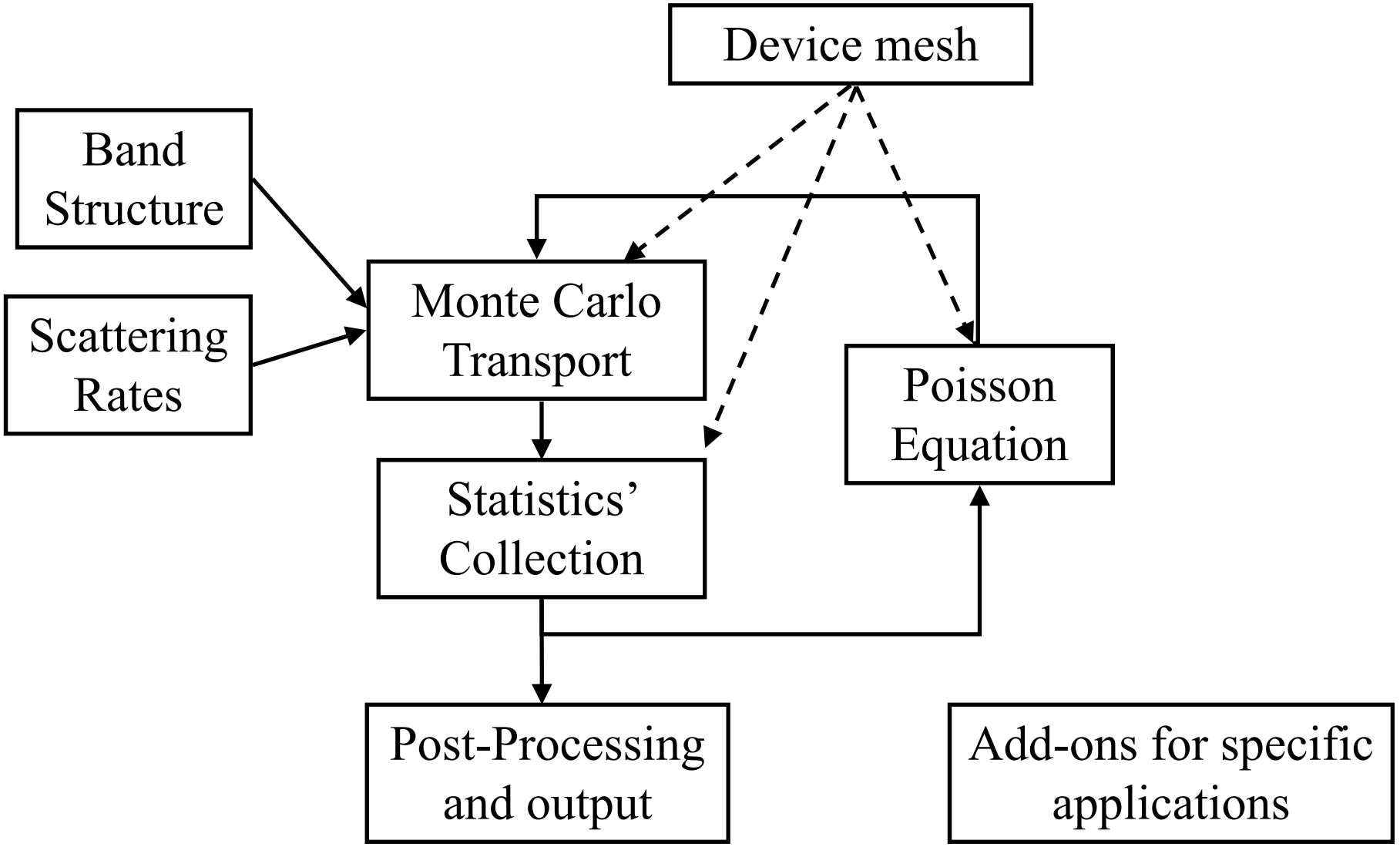
$$\hbar \frac{d\vec{K}}{dt} = e\vec{F}$$

$$\frac{d\vec{r}}{dt} = \nabla_{\vec{K}} E(\vec{K}) / \hbar$$



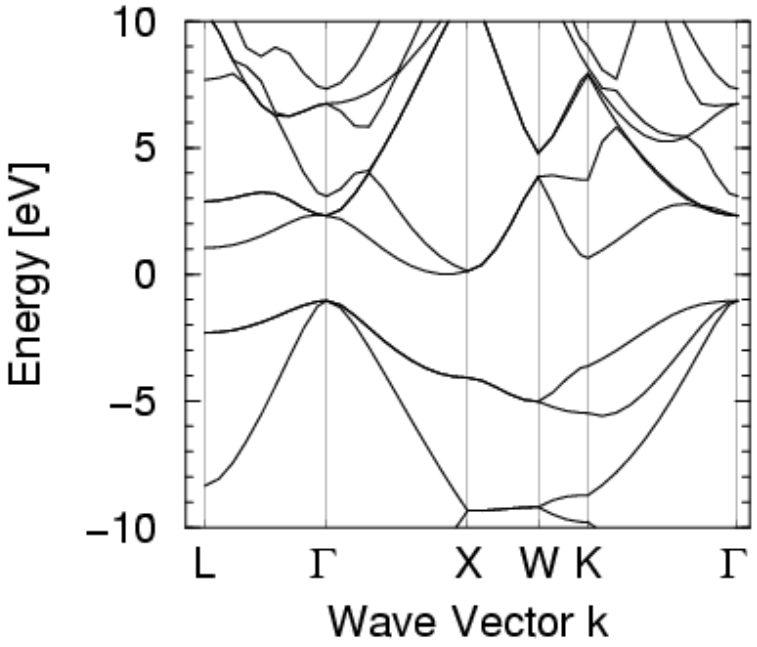
  - Scattering events
    - Determine the free-flight duration
    - “Randomize” K
  - Gathering of the statistics

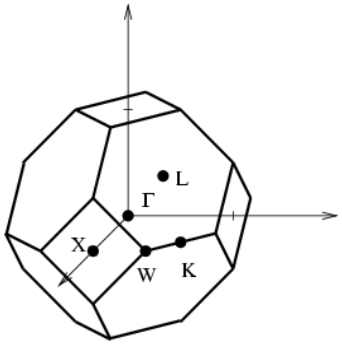
Structure of a Monte Carlo simulator



Full-Band Monte Carlo (1)

The dispersion relationship is parabolic only close to band minima



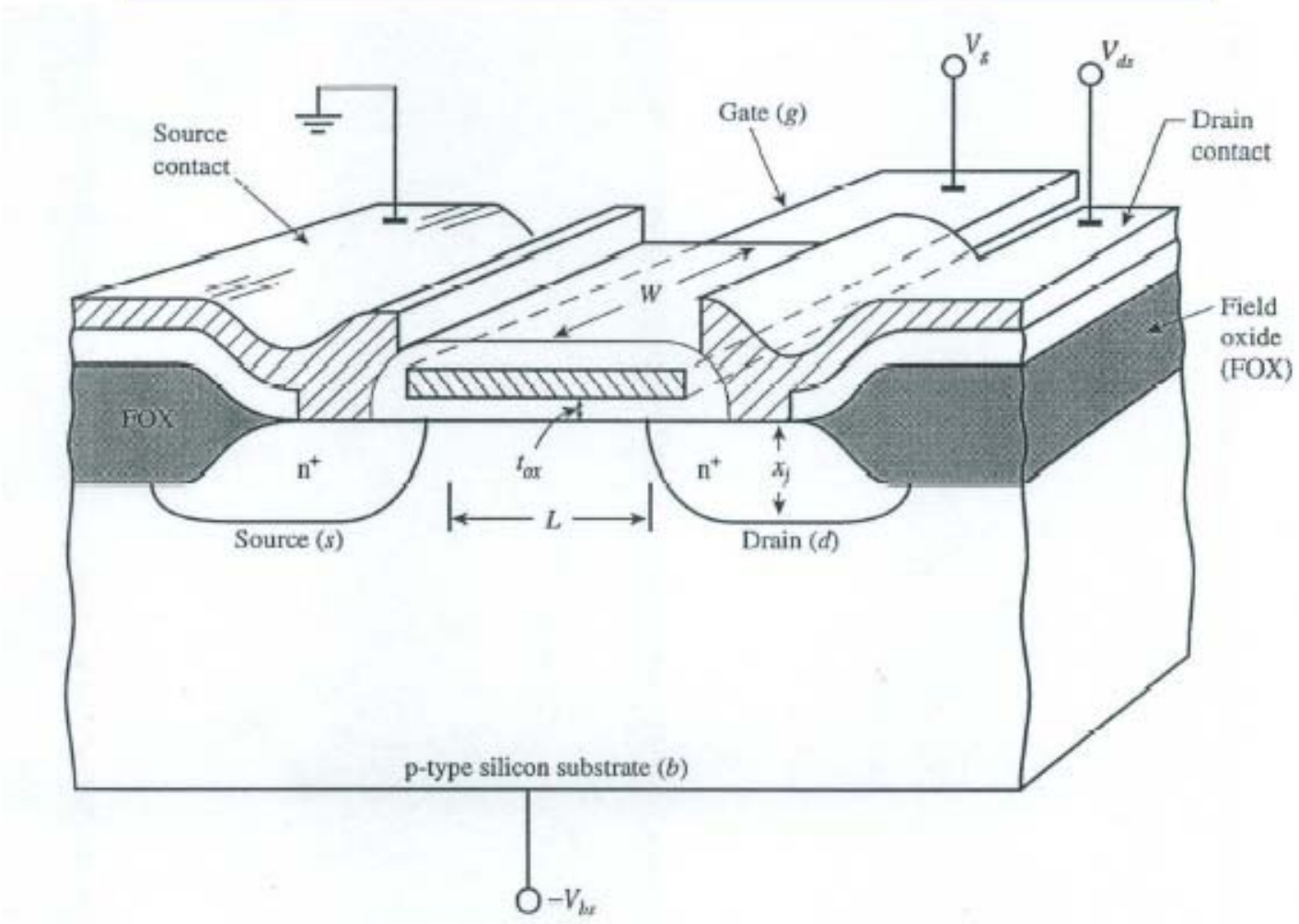


- Full-Band MC are based on the full dispersion relationship (from Non-Local-Pseudopotential method, etc.)
- Needed for high field transport (simulation of hot carriers)
- Integration of the equation of motion is more complicated

OUTLINE OF CLASS 2

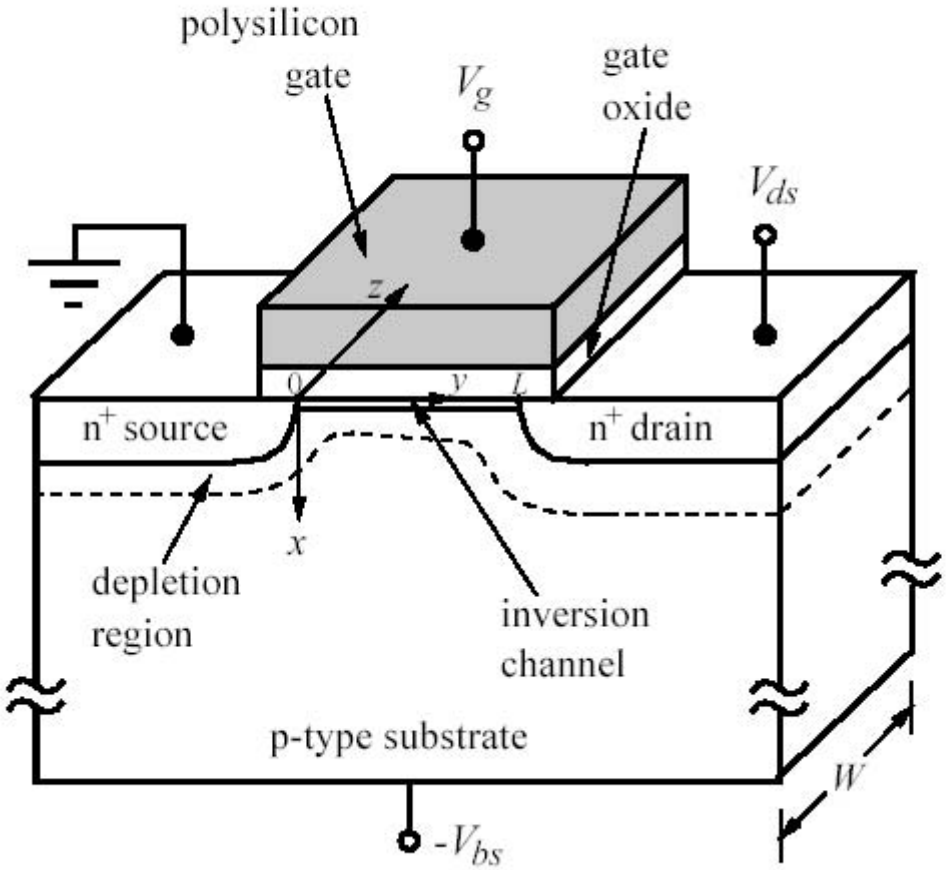
- **MOSFET : basic theory.**
- **Threshold voltage**
- **MOSFET current-voltage characteristics.**
- **Short Channel Effects**
- MOSFET scaling and scaling strategies
- Scaled MOSFETS
- Limits of Scaling

nMOSFET Schematic



- Four structural masks: Field, Gate, Contact, Metal.

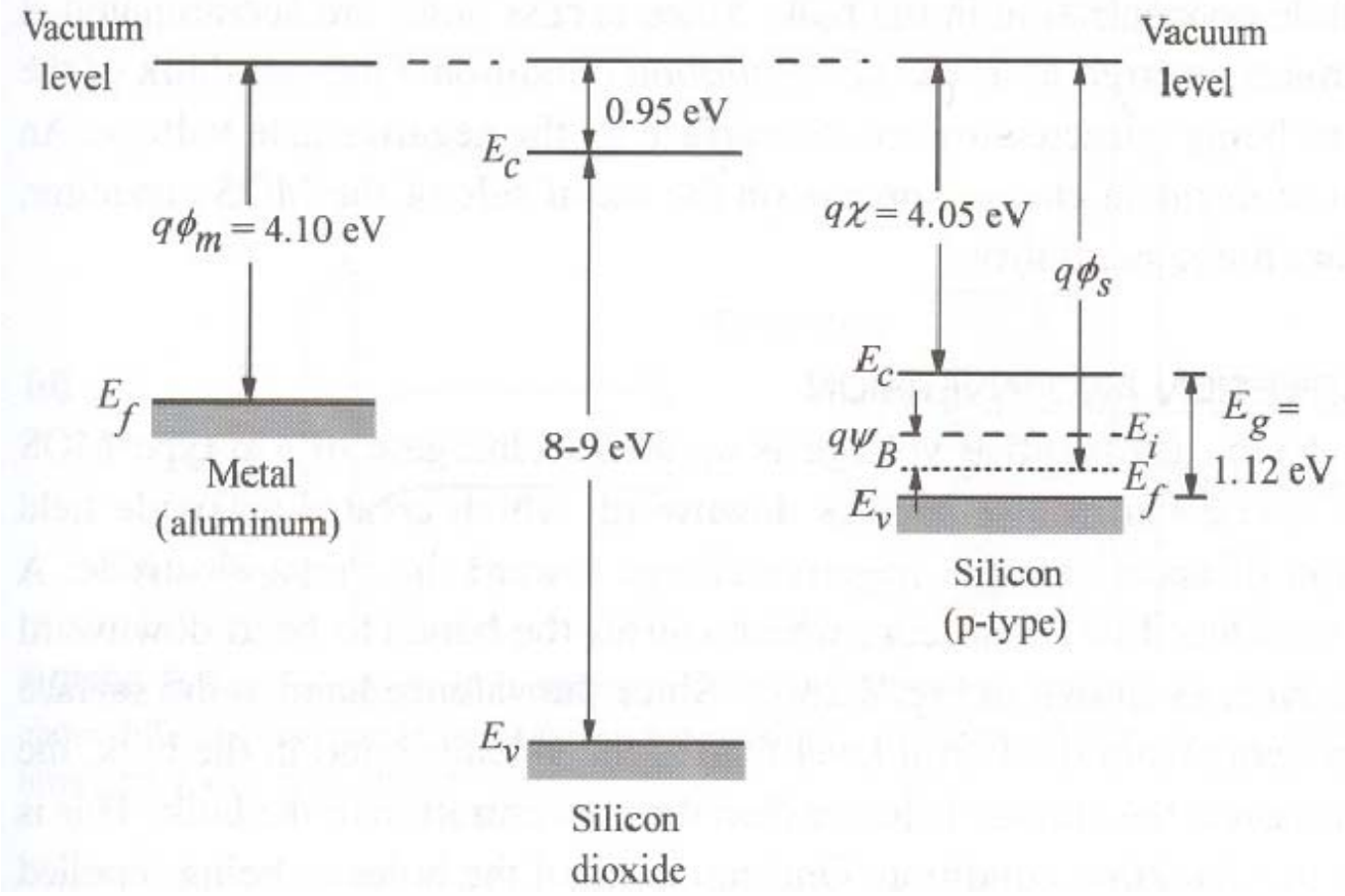
nMOSFET Schematic



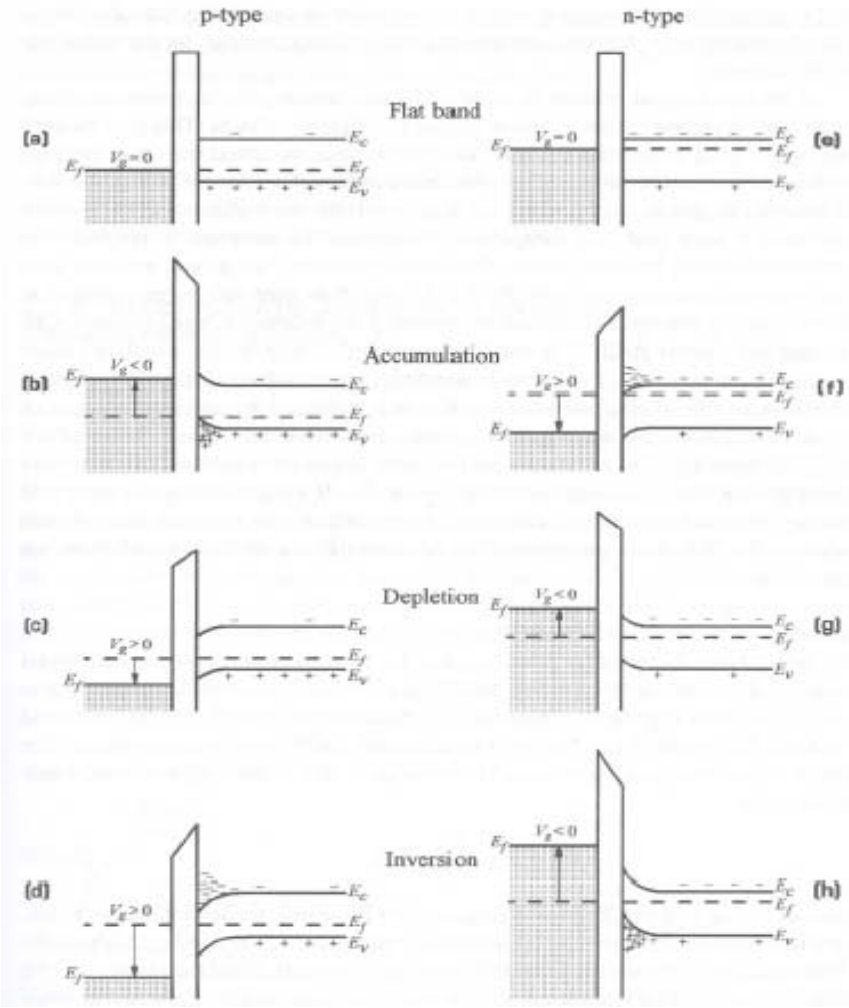
- Source terminal: Ground potential.
- Gate voltage: V_g
- Drain voltage: V_{ds}
- Substrate bias voltage: $-V_{bs}$

- $\psi(x,y)$: Band bending at any point (x,y).
- $V(y)$: Quasi-Fermi potential along the channel.
- $V(y=0) = 0$, $V(y=L) = V_{ds}$.

MOS Energy Band Diagram (I)



Band Diagram (II)



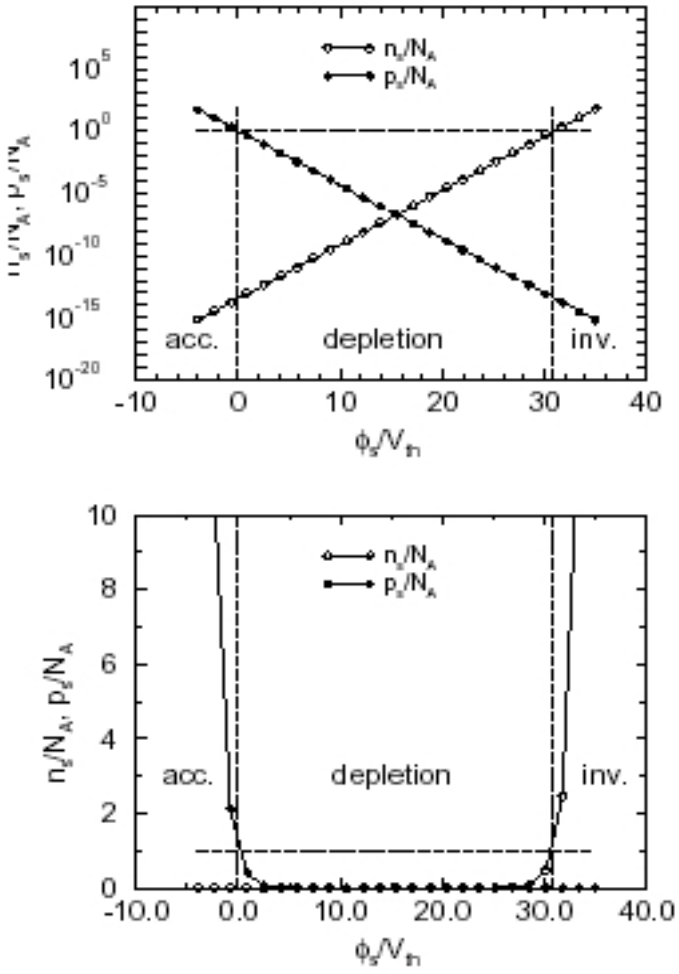
Surface potential and carrier concentration

- If the potential reference in the substrate is chosen to be zero, then electron concentration and potential at the surface and in the substrate can be defined assuming the equilibrium relationships:

$$n_s \;\; ; \;\; \phi_s \;\; ; \qquad n_b = n_i^2/N_A \;\; : \;\; \phi_b = 0$$

$$n_s = \frac{n_i^2}{N_A} \exp\left(\frac{\phi_s}{V_{th}}\right) \qquad p_s = N_A \exp\left(\frac{-\phi_s}{V_{th}}\right)$$

Surface potential and carrier concentration



Depletion and Inversion charges

- For $0 < \phi_s < 2\phi_F$ we can neglect the free carrier charge.
- The bulk charge (Q_s) is solely due to the ionized impurities.

$$\frac{d^2\phi}{dy^2} = -\frac{\rho}{\epsilon_{si}} = \frac{qN_A}{\epsilon_{si}} \quad \Rightarrow \quad y_d = \sqrt{\frac{2\epsilon}{qN_A}\phi_s}$$

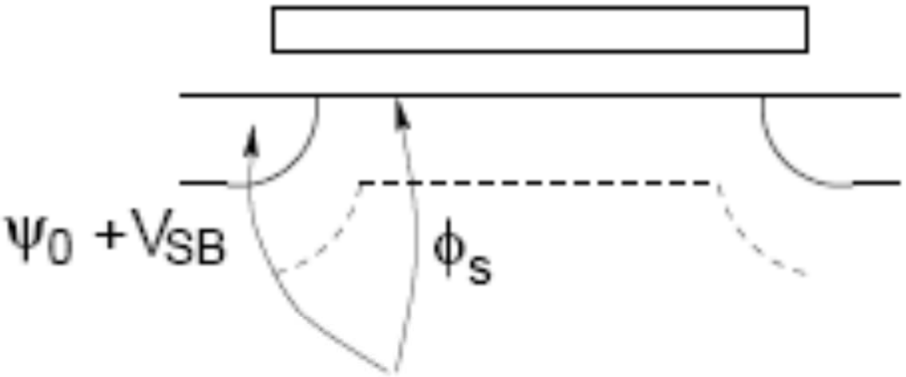
$$Q_s \simeq Q_B(\phi_s) \simeq -qN_A y_d(\phi_s)$$

- For $\phi_s > 2\phi_F$ the inversion charge tends to dominate:

$$Q_n = -\int_0^\infty qn(\phi(y))dy$$

Threshold Voltage

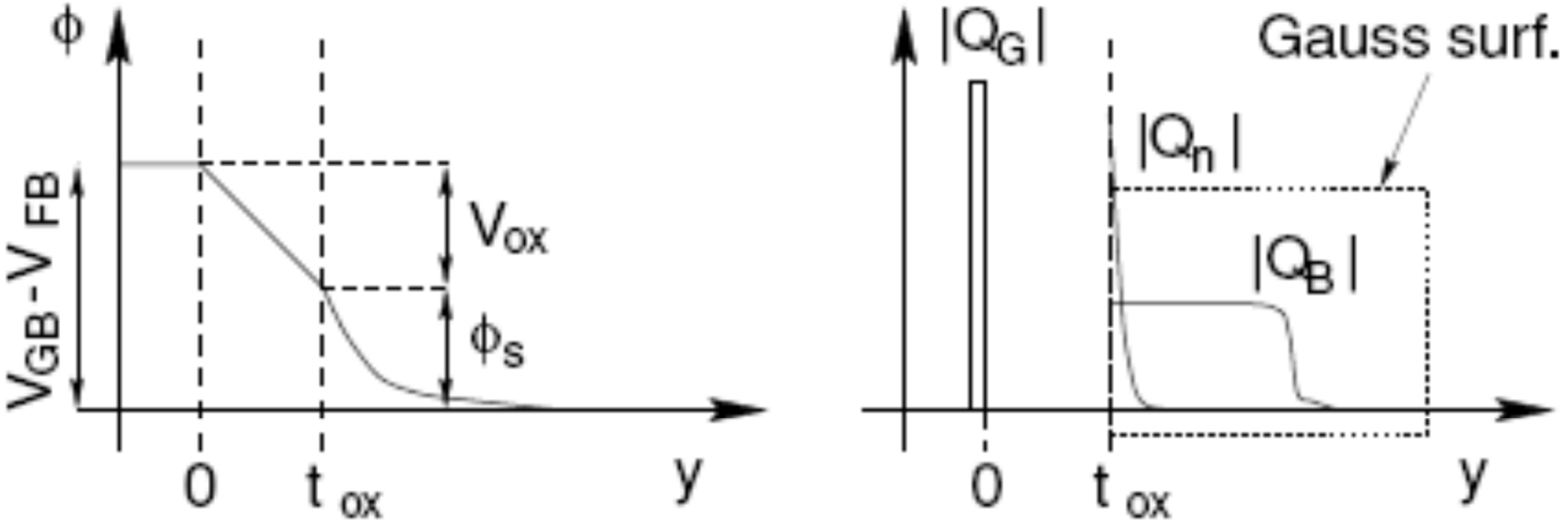
- Defined as the gate voltage that yields $n_s = N_A$
- Find ϕ_s^T such that $n_s = N_A$:



$$n_s = N_D \exp\left(\frac{\phi_s - (V_{SB} - \psi_0)}{V_{th}}\right)$$

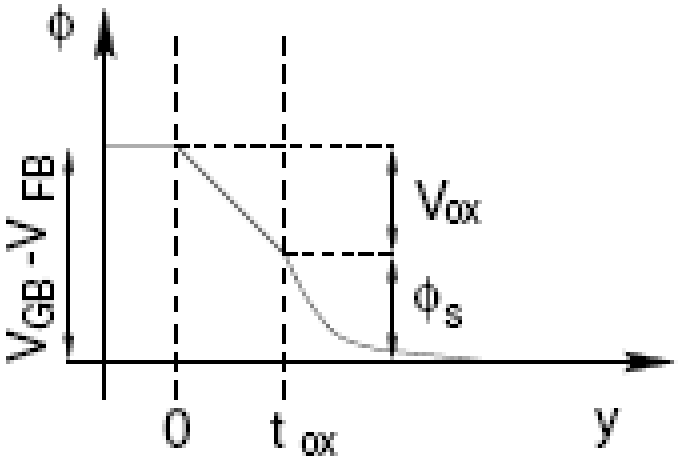
$$\phi_s^T = V_{SB} + V_{th} \ln \frac{N_D N_A}{n_i^2} + V_{th} \ln \frac{N_A}{N_D} = V_{SB} + 2\phi_F$$

- Find V_G such that $\phi_s = \phi_s^T$:



$$V_{OX} = E_{OX} t_{ox} = \frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox} \left(-\frac{Q_n + Q_B}{\epsilon_{si}} \right) = -\frac{Q_n + Q_B}{\epsilon_{ox}}$$

Threshold Voltage



- In general we have:

$$V_{GB} = V_{FB} + \phi_s - \frac{Q_B(\phi_s)}{C_{ox}} - \frac{Q_n(\phi_s)}{C_{ox}}$$

- Below threshold we have $Q_n \simeq 0$. Hence:

$$V_{GB}^T = V_{FB} + \phi_s^T - \frac{Q_B(\phi_s^T)}{C_{ox}} = V_{FB} + \phi_s^T + \frac{qN_A}{C_{ox}} y_d(\phi_s^T)$$

- Using the previous expression for y_d we have:

$$y_d = \sqrt{\frac{2\epsilon_{si}}{qN_A}\phi_s} \quad \Rightarrow \quad y_d^T = \sqrt{\frac{2\epsilon_{si}}{qN_A}(2\phi_F + V_{SB})}$$

$$\begin{aligned} V_{GB}^T &= V_{FB} + 2\phi_F + V_{SB} + \gamma\sqrt{2\phi_F + V_{SB}} \\ V_{GS}^T = V_T &= V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V_{SB}} = \\ &= V_{T0} + \gamma\left[\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right] \\ V_{T0} &= V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F} \end{aligned}$$

- V_{T0} depends on gate material workfunction, oxide capacitance, depletion charge.
- γ (body effect coefficient) depends on oxide capacitance and substrate doping.

Drain Current

- The drift-diffusion electron current density is:

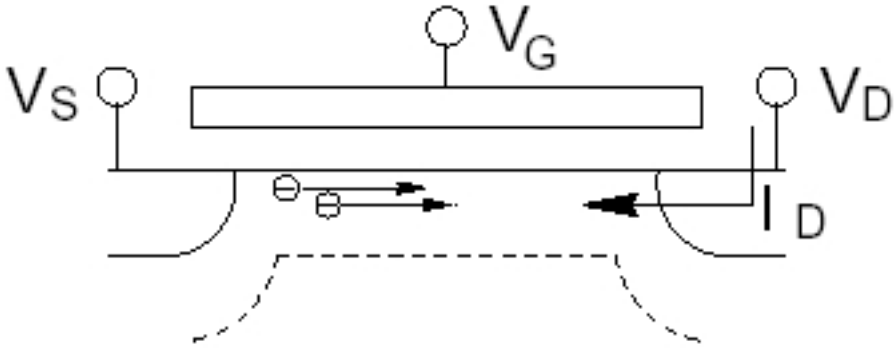
$$J_n = -q\mu_n n\frac{d\phi}{dx} + qD_n\frac{dn}{dx}$$

- Integrating in y and z with constant mobility, assuming a surface sheet of charge and remembering that $D_n = \mu_n V_{th}$:

$$\begin{aligned} I &= -qW\int_0^\infty n\mu_n\frac{d\phi}{dx}dy + qWV_{th}\int_0^\infty \mu_n\frac{dn}{dx}dy \\ &= W\mu_n\frac{d\phi_s}{dx}Q_n(x) - WV_{th}\mu_n\frac{dQ_n(x)}{dx} \end{aligned}$$

- Integrating along x:

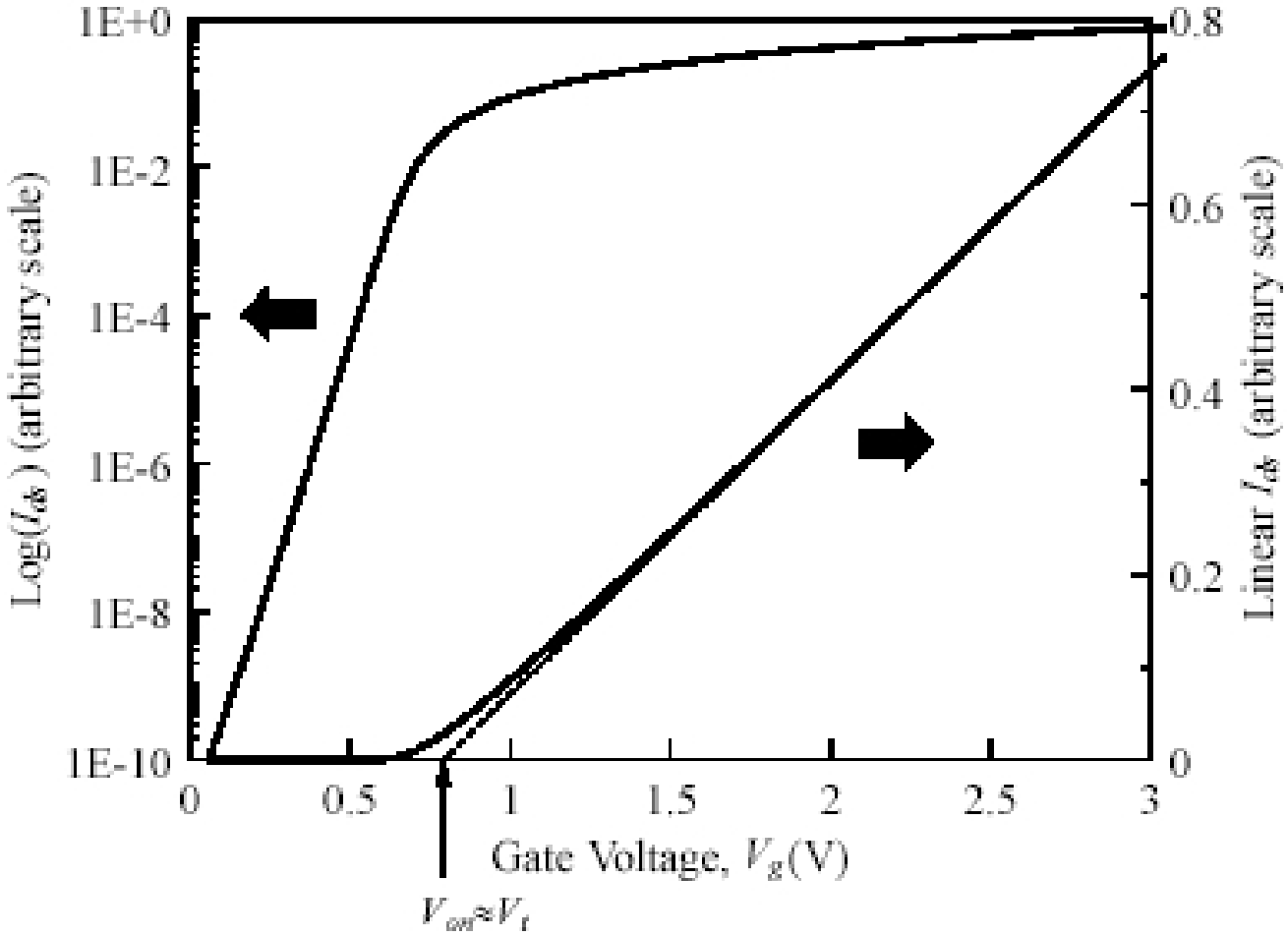
$$\begin{aligned} I &= \frac{W}{L}\mu_n\left[\int_{\phi_{s(0)}}^{\phi_{s(L)}} Q_n(\phi_s)d\phi_s - V_{th}\left(Q_n(\phi_{s(L)}) - Q_n(\phi_{s(0)})\right)\right] \\ &= -I_D \end{aligned}$$



MOSFET: beyond the basics

- Subthreshold characteristics
- Velocity saturation
- Short channel effects

Transfer I-V characteristics



Subthreshold current

$$-Q_s = \varepsilon_{si}\mathcal{E}_s = \sqrt{2\varepsilon_{si}kTN_a}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_a^2}e^{q(\psi_s - V)/kT}\right]^{1/2}$$

Power series expansion: 1st term Q_d, 2nd term Q_i,

$$-Q_i = \sqrt{\frac{\varepsilon_{si}qN_a}{2\psi_s}}\left(\frac{kT}{q}\right)\left(\frac{n_i}{N_a}\right)^2 e^{q(\psi_s - V)/kT}$$

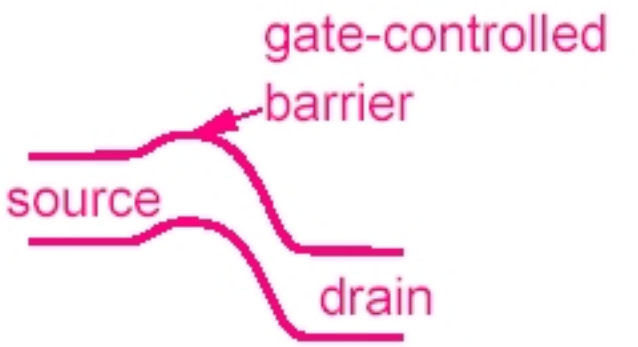
$$\Rightarrow \quad I_{ds} = \mu_{eff}\frac{W}{L}\sqrt{\frac{\varepsilon_{si}qN_a}{2\psi_s}}\left(\frac{kT}{q}\right)^2\left(\frac{n_i}{N_a}\right)^2 e^{q\psi_s/kT}\left(1 - e^{-qV_{ds}/kT}\right)$$

or, $$I_{ds} = \mu_{eff}C_{ox}\frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_g - V_t)/mkT}\left(1 - e^{-qV_{ds}/kT}\right)$$

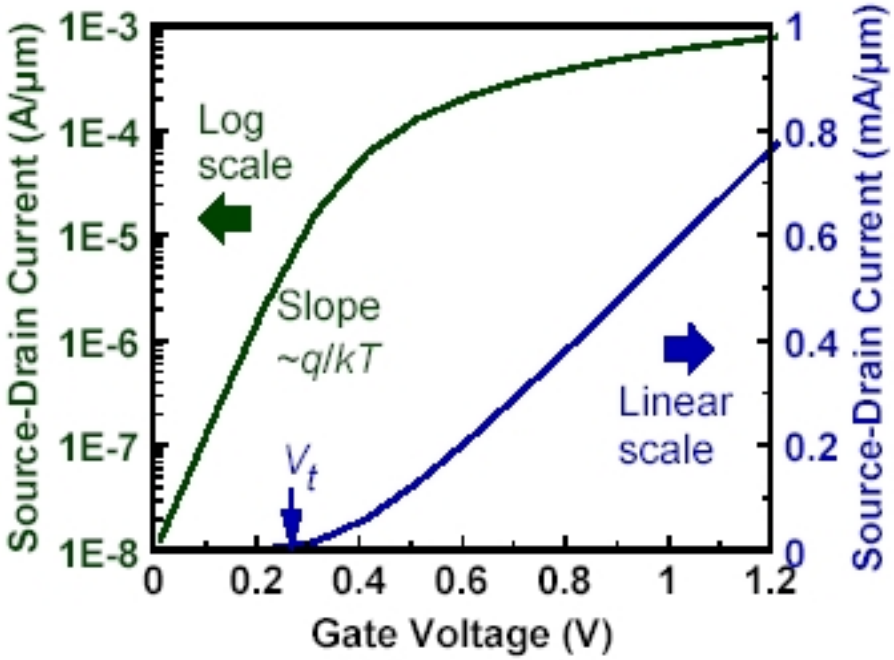
Inverse subthreshold slope:

$$S = \left(\frac{d(\log I_{ds})}{dV_g}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{ox}}\right)$$

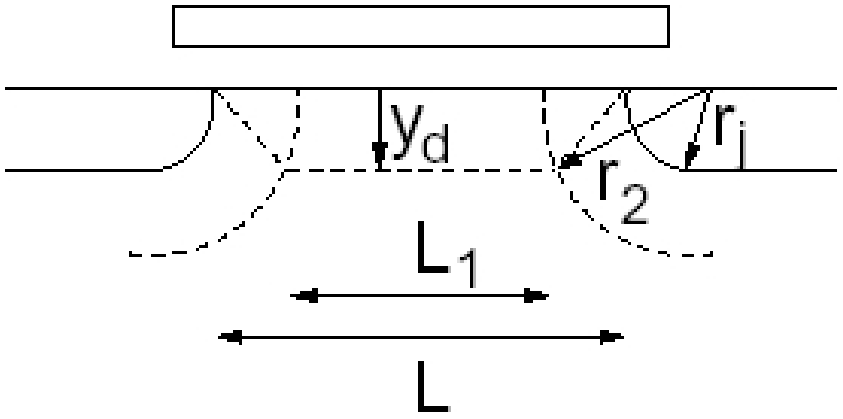
Short Channel Effects (I)



Threshold voltage becomes sensitive to channel length and drain bias.



Short Channel Effects (II)



- Long Channel:

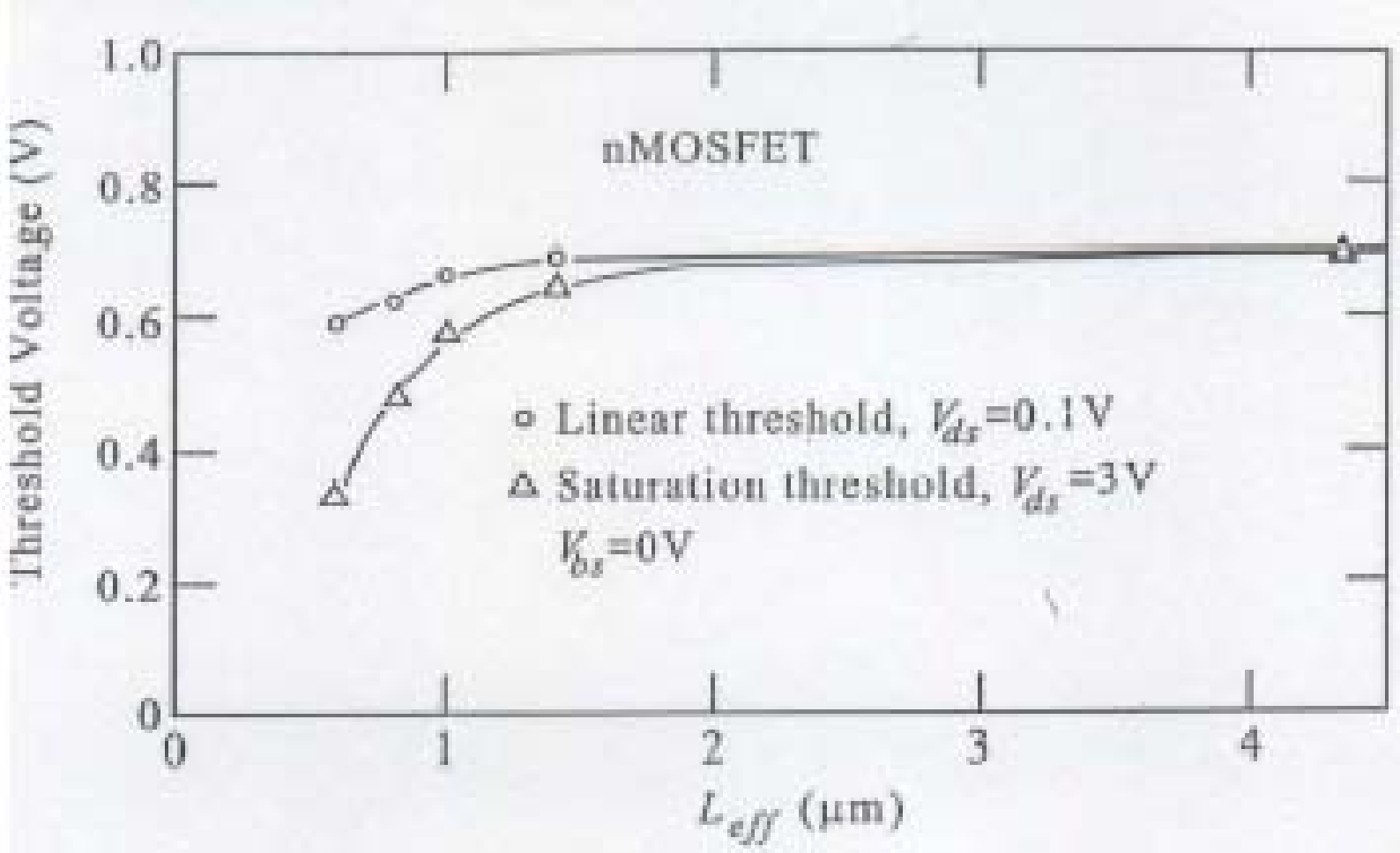
$$Q_B = -qN_A y_d \;\; ; \;\; V_T = V_{FB} + 2\phi_F - \frac{Q_B}{C_{ox}}$$

- Short Channel:

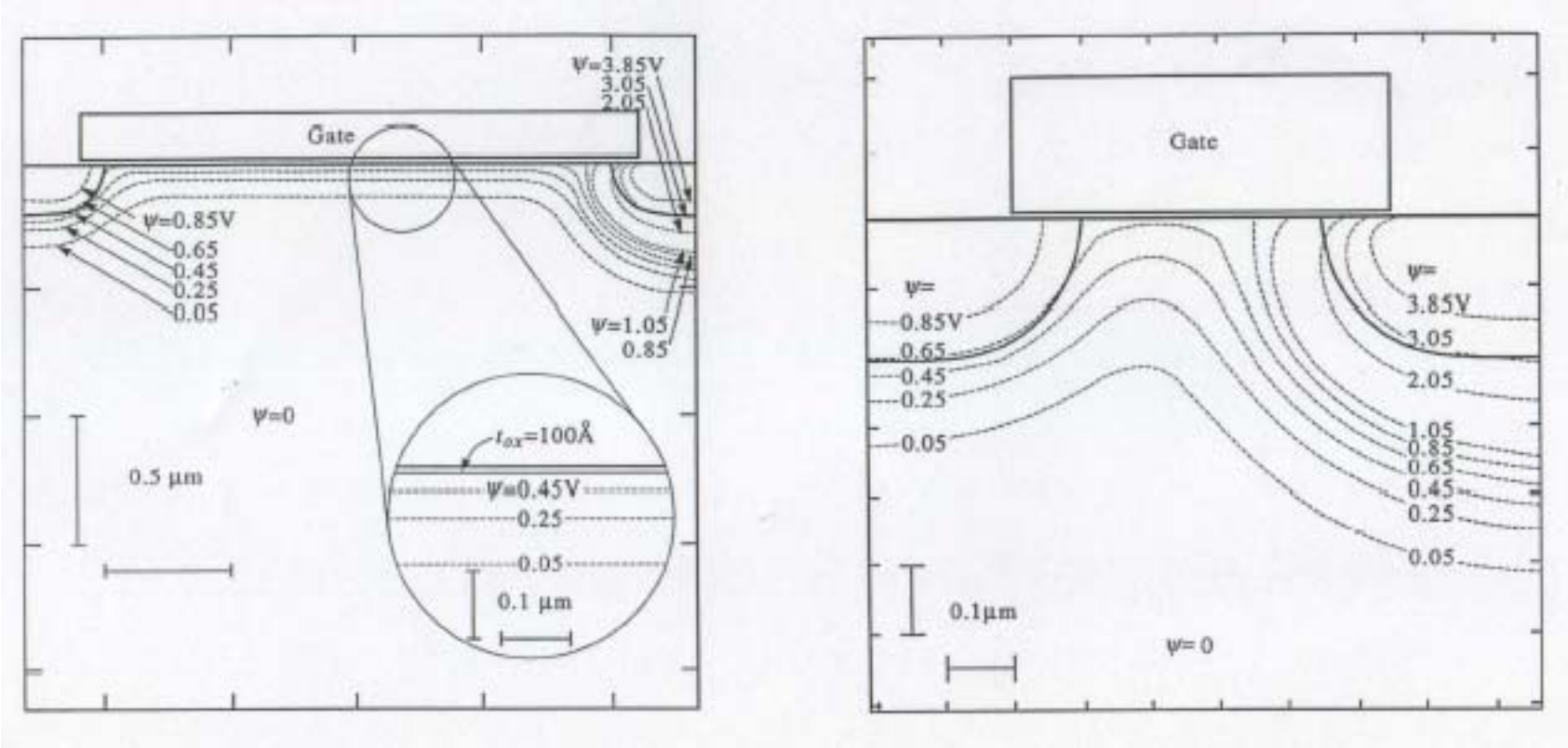
$$Q'_B = -qN_A W y_d \frac{L + L_1}{2}$$

$$V_T = V_{FB} + 2\phi_F - \frac{Q'_B}{C_{ox} W L}$$

Short channel V_T roll-off



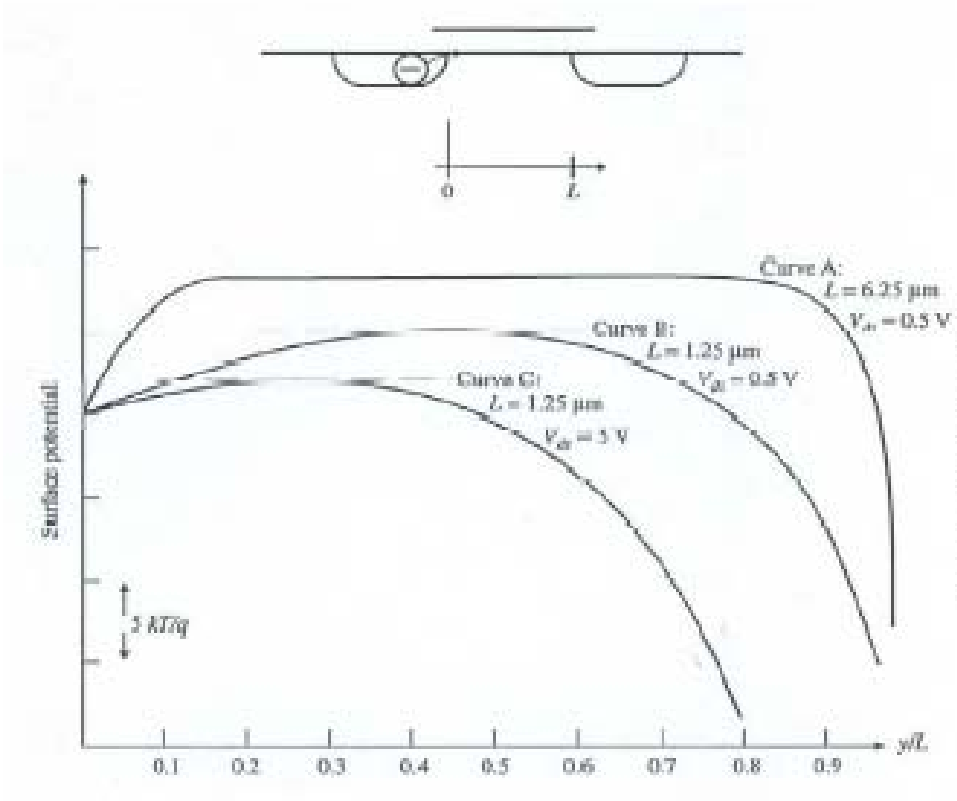
2D simulation of potential

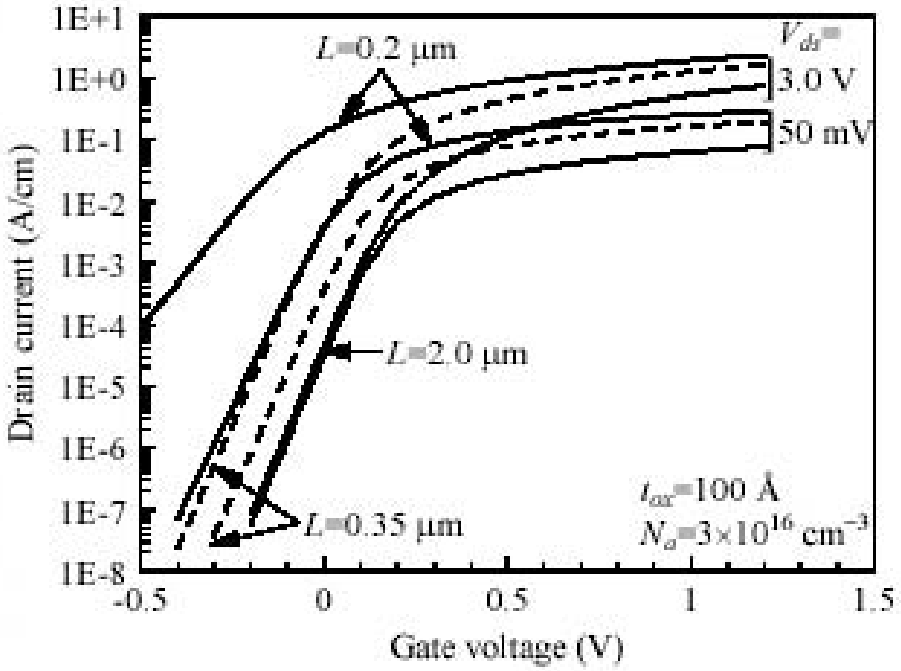


Long channel

Short channel

Drain-induced barrier lowering





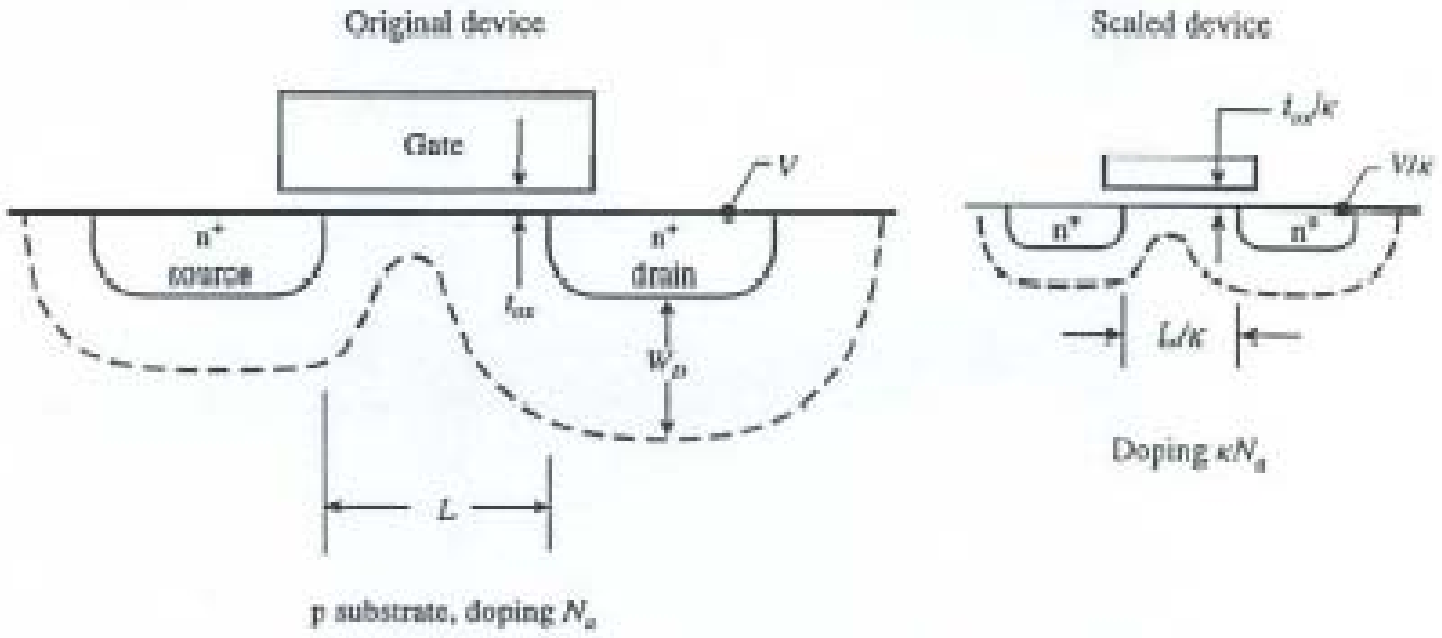
OUTLINE OF CLASS 2

- MOSFET : basic theory.
- Threshold voltage
- MOSFET current-voltage characteristics.
- Short Channel Effects
- **MOSFET scaling and scaling strategies**
- **Scaled MOSFETS**
- **Limits of Scaling**

MOSFET SCALING

Device scaling: simplified design rules for shrinking device dimensions to density and performance gains and power reduction

Principle of constant field scaling (Dennard, 1978)



Scaling

1. **Reduce Capacitance**

   Reduce switching time of MOSFETs

   Reduce power consumption

2. **Increase number of Transistors**

   Increase functionality

   Parallel processing

   Increase circuit operation speed

Predictions of Scaling limits since IC started

Period	Expected limit	Cause
Late 1970's	1μm:	SCE
Early 1980's	0.5μm:	S/D resistance
Early 1980's	0.25μm:	Direct-tunneling of gate SiO2
Late 1980's	0.1μm:	Several
Today	50nm:	Several
Today	<10nm:	Fundamental?

Fundamental: electron wave length (10nm), tunneling distance(3nm), atom distance (0.3nm)

VLSI textbook (Mead and Conway)

Finally, there appears to be a fundamental limit of approximately quarter micron channel length, where certain physical effects such as the tunneling through the gate oxide and fluctuations in the positions of impurities in the depletion layers begin to make the devices of smaller dimension unworkable.

Constant field scaling (I)

The principle of constant field scaling lies in scaling the device voltages and the device dimensions (horizontal and vertical) by the same factor, k>1 , such that the electric field distribution within the device remains unchanged.

Issues: short-channel effects, power density, reliability

Constant field scaling (II)

	MOSFET Device and Circuit Parameters	Multiplicative Factor ($\kappa > 1$)
Scaling assumptions	Device dimensions (t_{ox}, L, W, x_j)	$1/\kappa$
	Doping concentration (N_a, N_d)	κ
	Voltage (V)	$1/\kappa$
Derived scaling behavior of device parameters	Electric field (E)	1
	Carrier velocity (v)	1
	Depletion layer width (W_d)	$1/\kappa$
	Capacitance ($C = \varepsilon A/t$)	$1/\kappa$
	Inversion layer charge density (Q_i)	1
	Current, drift (I)	$1/\kappa$
	Channel resistance (R_{ch})	1
Derived scaling behavior of circuit parameters	Circuit delay time ($\tau \sim CV/I$)	$1/\kappa$
	Power dissipation per circuit ($P \sim VI$)	$1/\kappa^2$
	Power-delay product per circuit ($P \times \tau$)	$1/\kappa^3$
	Circuit density ($\propto 1/A$)	κ^2
	Power density (P/A)	1

Generalized scaling

- It allows the electric field to scale up
- by a factor α, while device dimensions scale down by κ
- Voltages scale by α/κ
- More flexible than constant-field scaling, but has power and reliability concerns

Constant voltage scaling

- Special case of generalized scaling with $\alpha = \kappa$
- Threshold voltage remains unchanged
- Physically incorrect since electric field increases (reliability) and power density increases

Practical scaling

- CMOS technology has gone through a mixture of constant voltage and constant field scaling steps, partially due to system level needs. As a result, field and power density has gone up, but performance gain has been maintained and power per circuit has come down.
- Reliability requirements continue to be met

CMOS Technology generations

Feature Size	Power Supply	Gate Oxide	Oxide Field
2 μm	5 V	350 Å	1.4 MV/cm
1.2 μm	5 V	250 Å	2.0 MV/cm
0.8 μm	5 V	180 Å	2.8 MV/cm
0.5 μm	3.3 V	120 Å	2.8 MV/cm
0.35 μm	3.3 V	100 Å	3.3 MV/cm
0.25 μm	2.5 V	70 Å	3.6 MV/cm

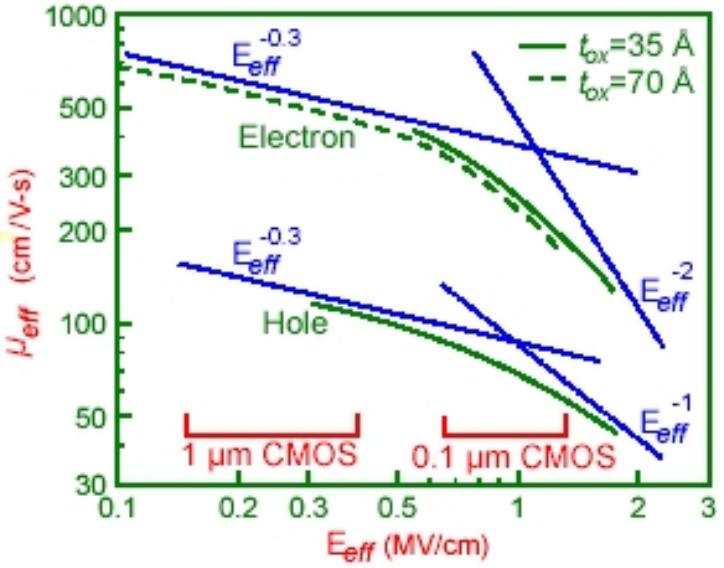
Non-scaling factors

Primary nonscaling factors:

- Built-in potential ψ_{bi} (***Si bandgap***)
- Subthreshold current (***thermal energy kT/q***)

Secondary nonscaling factors (due to higher $\mathcal{E}$):

- ❑ Velocity saturation
- ❑ Decreased mobility at higher fields
- ❑ Oxide reliability (t_{ox} scales less, W_{dm} more)



Scaling of Depletion Width

Maximum drain depletion width: $W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} + V_{dd})}{qN_a}}$

For $N_a \rightarrow \kappa N_a$ and $V_{dd} \rightarrow V_{dd}/\kappa$,
$W_D \rightarrow W_D/\kappa$ if $V_{dd} >> \psi_{bi}$.

However, the source depletion width, $W_S = \sqrt{\frac{2\varepsilon_{si}\psi_{bi}}{qN_a}}$

is indep. of V_{dd} and only scales as $W_S \rightarrow W_S/\sqrt{\kappa}$.

Furthermore, the maximum gate depletion width,

$$W_{dm}{}^{0} = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a / n_i)}{q^2 N_a}}$$

scales even less than $1/\sqrt{\kappa}$.

Other non-scaling factors

- Source and drain resistances
  - Doping level limited by solid solubility
  - Junction abruptness limited by annealing
- Polysilicon gate depletion
- Inversion layer thickness
- Process tolerances:
  - Gate length
  - Dopant number fluctuation

OUTLINE OF THE COURSE
CLASS 3: from nano to Tera

- MOSFET Scaling:
  - Recent Scaling Scenario
  - Multigate MOSFET architectures

Recent Scenario

- CMOS scaling below the 28 - 22 TN is constrained by several parasitic effects, among which are:
  - Lateral SCE
  - Gate leakage due to tunneling
  - Low mobility induced by high doping
- Multi-Gate MOSFETs and high-k dielectrics help to limit SCE allowing for a thicker gate oxide compared to conventional bulk MOSFETs.

Device Scaling

- Channel Width $\propto$ Channel Length
  - To ensure that optimal W/L ratios are available
- Drain-to-Source Voltage $\propto$ Channel Length
  - To maintain a constant electric field from source to drain
- Gate-Oxide Thickness $\propto$ Channel Length
  - To maintain control of the energy barrier in the channel through increasing the capacitive coupling from the gate
- Channel Doping Density $\propto$ 1/(Channel Length)
  - Maintain a reasonable number of dopants in the decreasing channel area

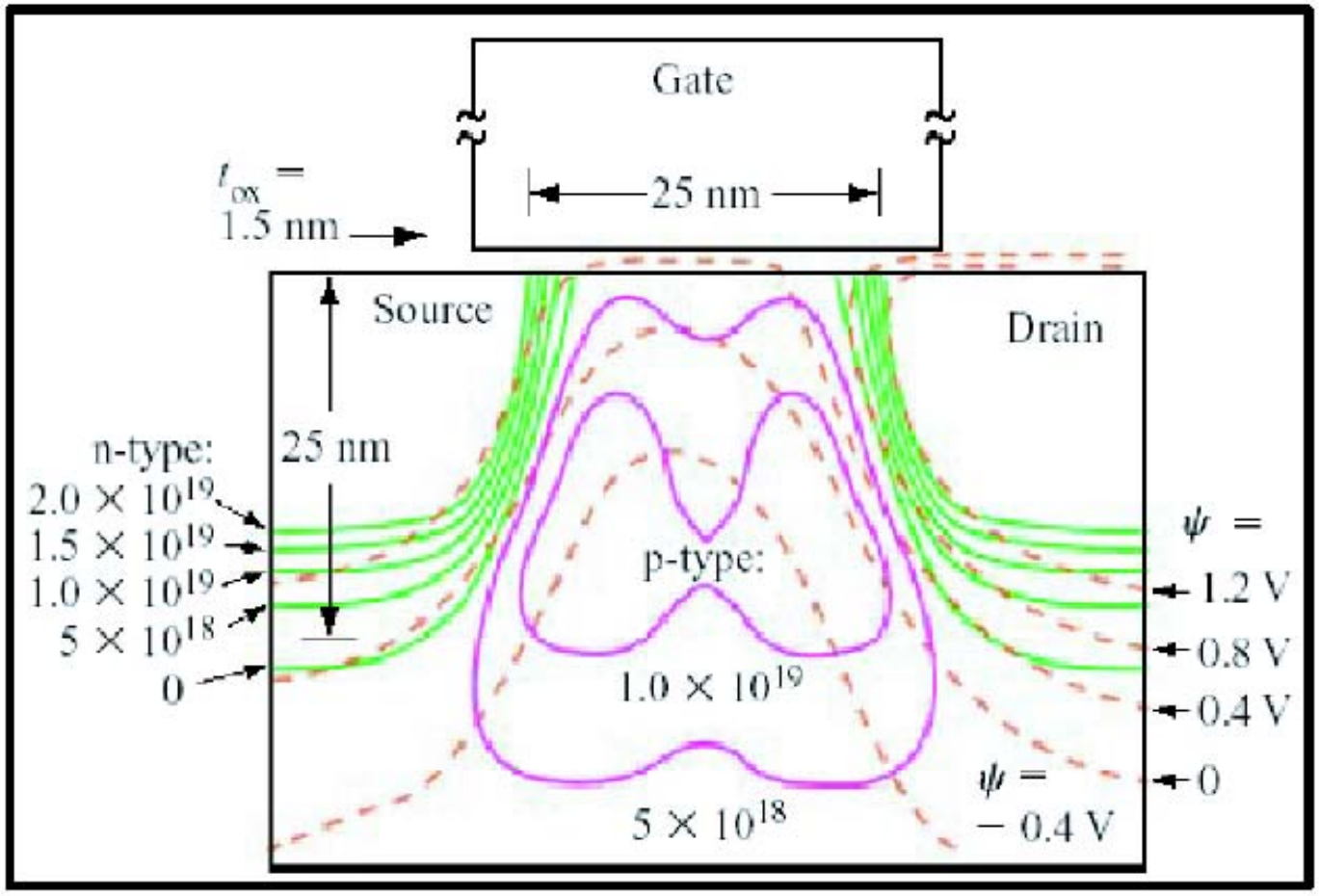
Scaling Issues

- Short-Channel Effects
  - Drain Induced Barrier Lowering (DIBL)

Possible Solution

New channel doping profiles

"Superhalo" doping profile



(Y. Taur IBM J. RES. & DEV. VOL. 46 NO. 2/3 – March/May 2002)

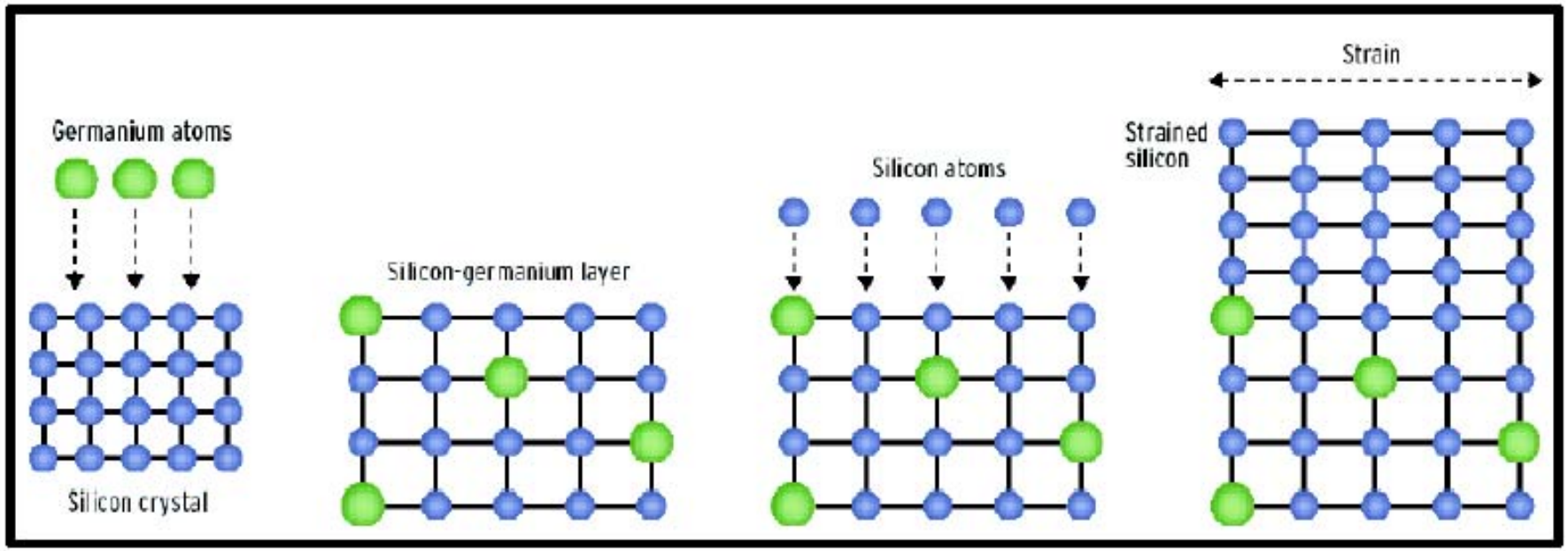
Scaling Issues

- Mobility Degradation
  - Increased Surface Scattering

Possible Solution

"Strained Silicon"

“Strained Silicon”



(L. Geppert – Oct 2002)

Scaling Issues

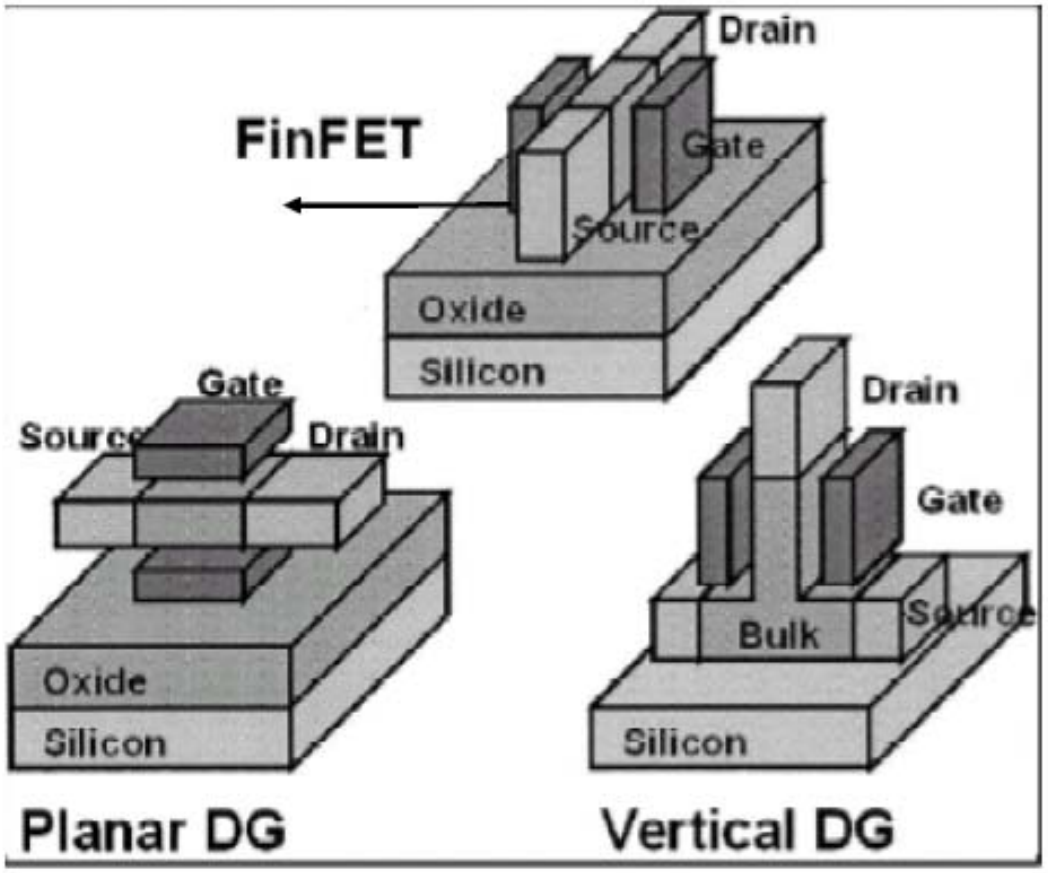
- Doping Density Fluctuations
  - Decreasing channel dimensions make matching channel doping from one device to the next nearly impossible.

<u>Possible Solution</u>

Un-doped Channels

(Not a reasonable solution for the conventional MOSFET)

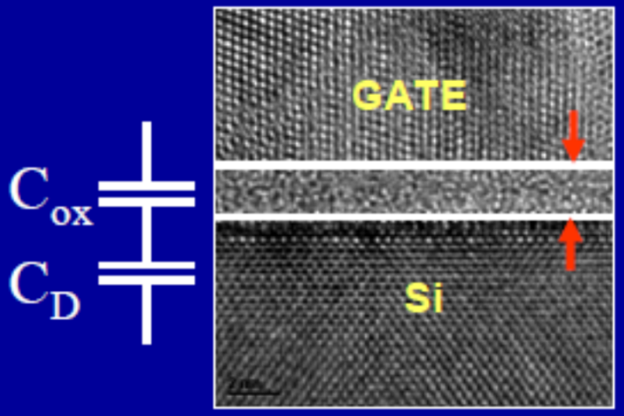
Multigate ultra-thin body architectures allow to keep SCE under control even with almost undoped channels

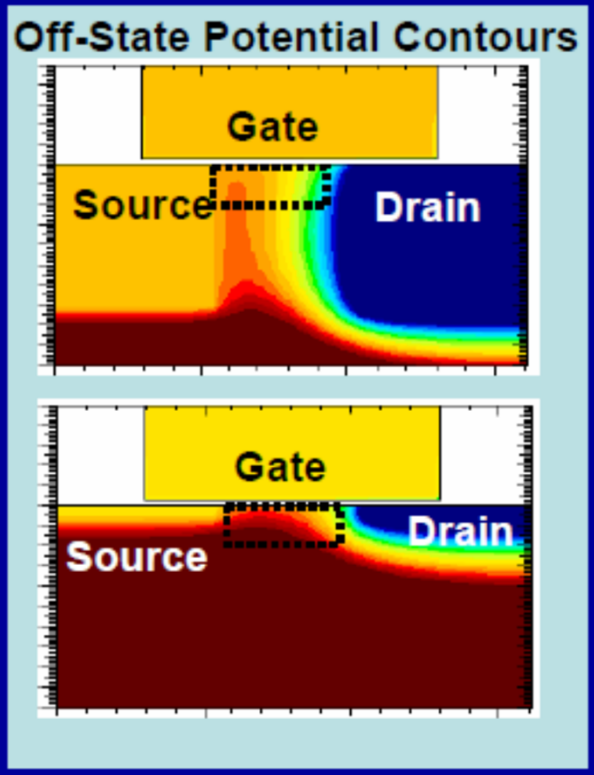


J Kretz, L Dreeskornfeld, J Hartwich, and W Rosner, *Microelectronic Eng.* 67-68, p763-768, 2003

Improving Electrostatics

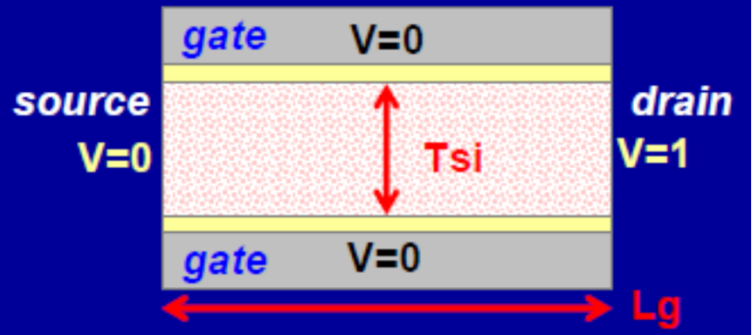
- **Shallow S/D junctions**
- **Well engineering**
  - halo implants
- **Improving electrical Tox**
  - thinning
  - Hi-K
  - Metal gates

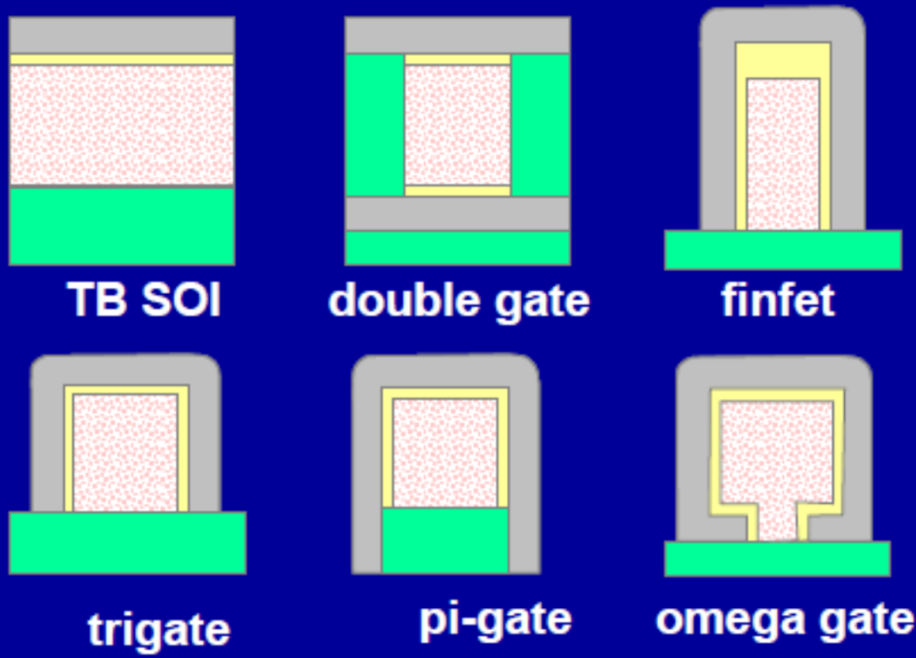


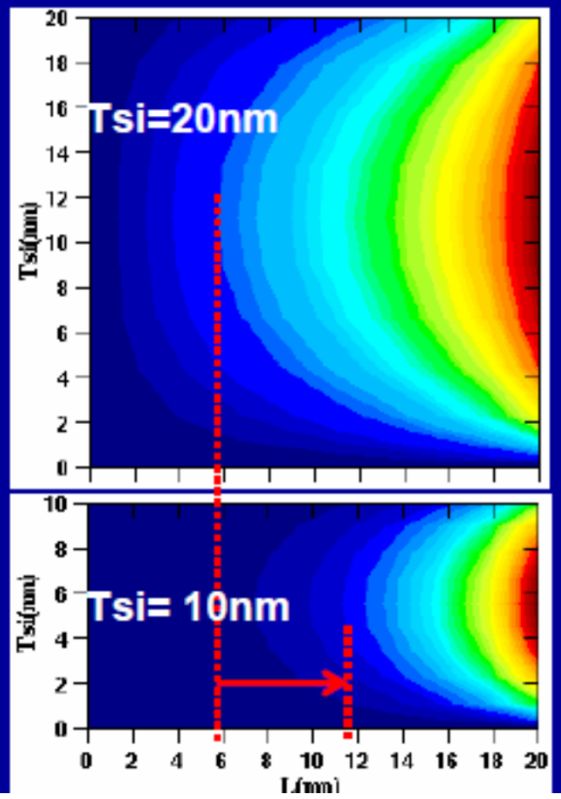


Improving Electrostatics…

- **Gate/channel architecture**
  - Thin body SOI, multigate devices, nanowires

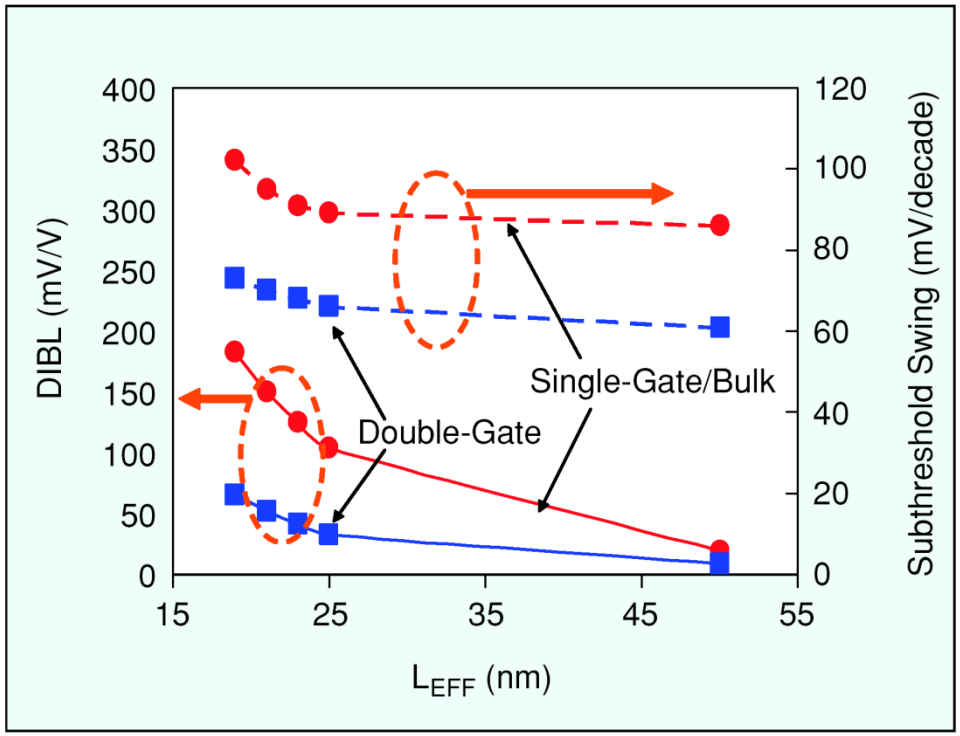






Double-gate FET (DGFET) can reduce Short Channel Effects (SCEs)

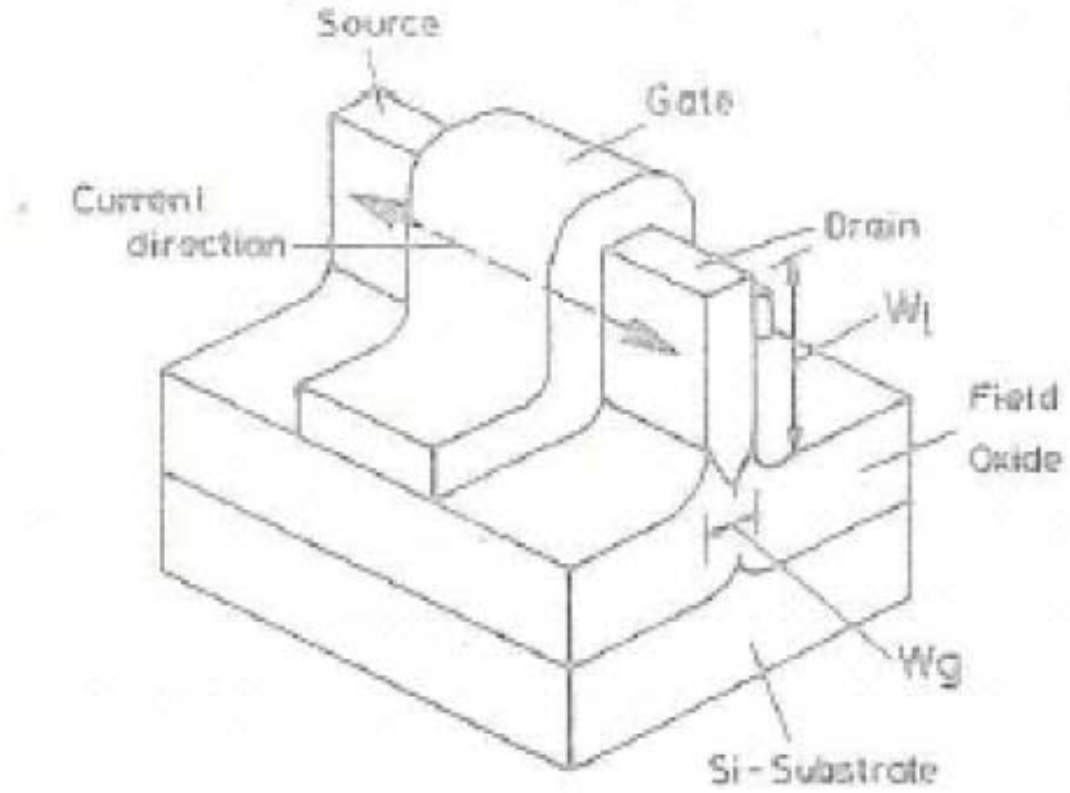
- Reduce Drain-Induced-Barrier-Lowering
- Improve Subthreshold Swing S
- allows for thicker gate oxide compared to conventional bulk MOSFET



•Medici-predicted DIBL and subthreshold swing versus effective channel length for DG and bulk-silicon nFETs

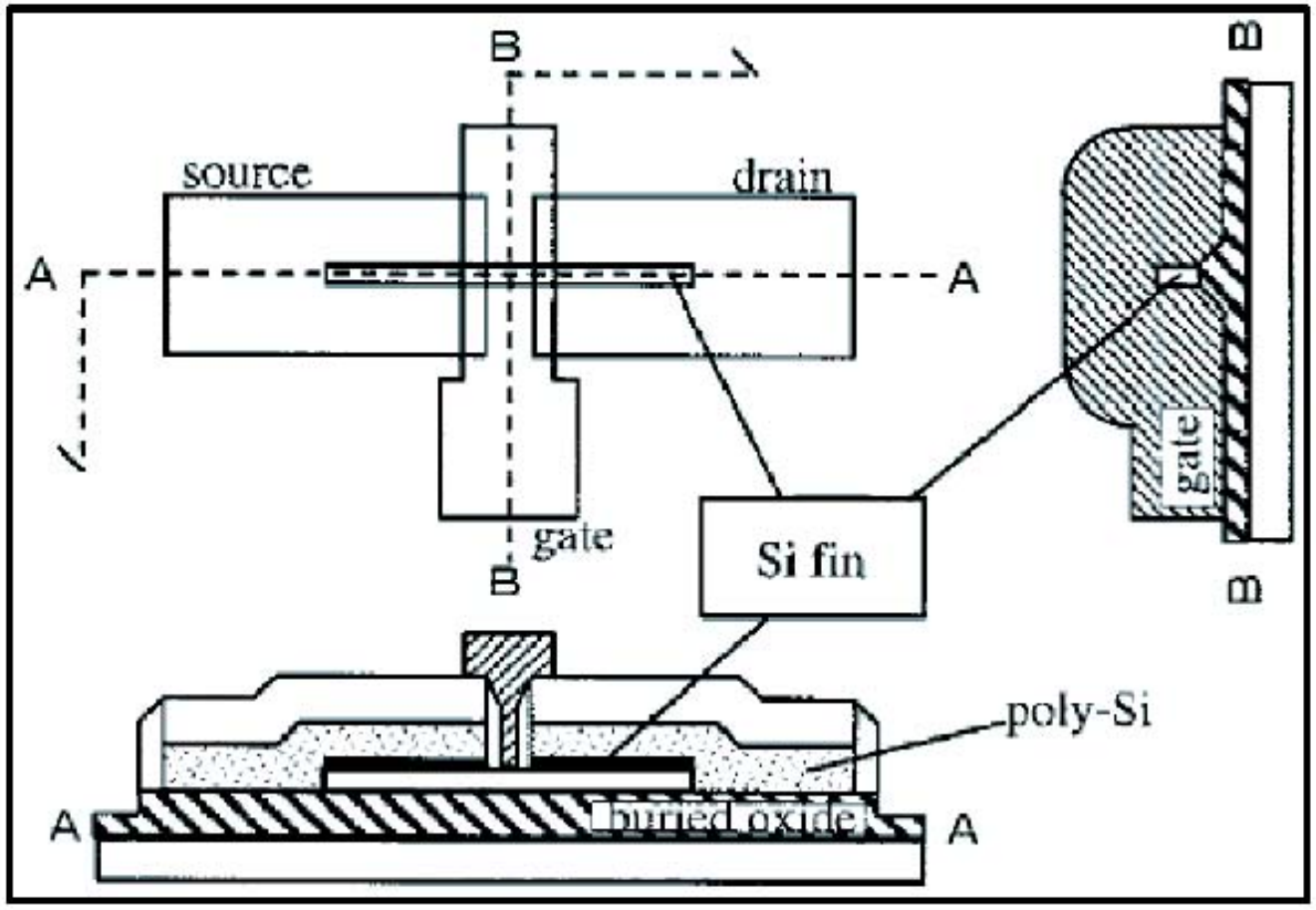
•E J Nowak, I Aller, T Ludwig, K Kim R V Joshi, C-T Chuang, K Bernstein and R Puri, *IEEE Circuits and Dev. Magazine*, p20-31, Jan/Feb 200

First FinFET - DELTA (DEpleted Lean-channel TrAnsistor)



D.Hisamoto, T.Kaga, Y.Kawamoto, and E.Takeda, *IEEE Electron Dev. Lett.*, vol.11, no.1, p36-38, Jan 1990

The FinFET



(E. Anderson, et. al. – Dec 2000)

Design - Geometry

- $H_{fin} >> T_{fin}$
- Top gate oxide thickness >> sidewall oxide thickness

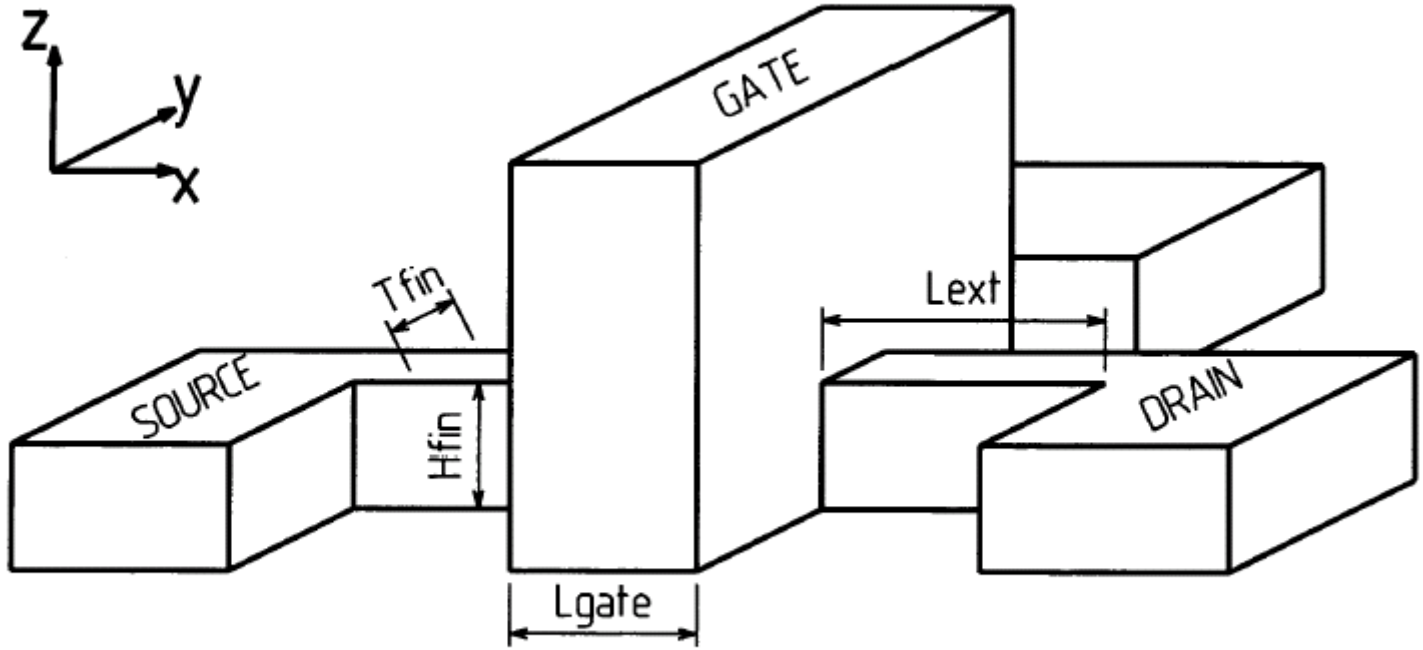


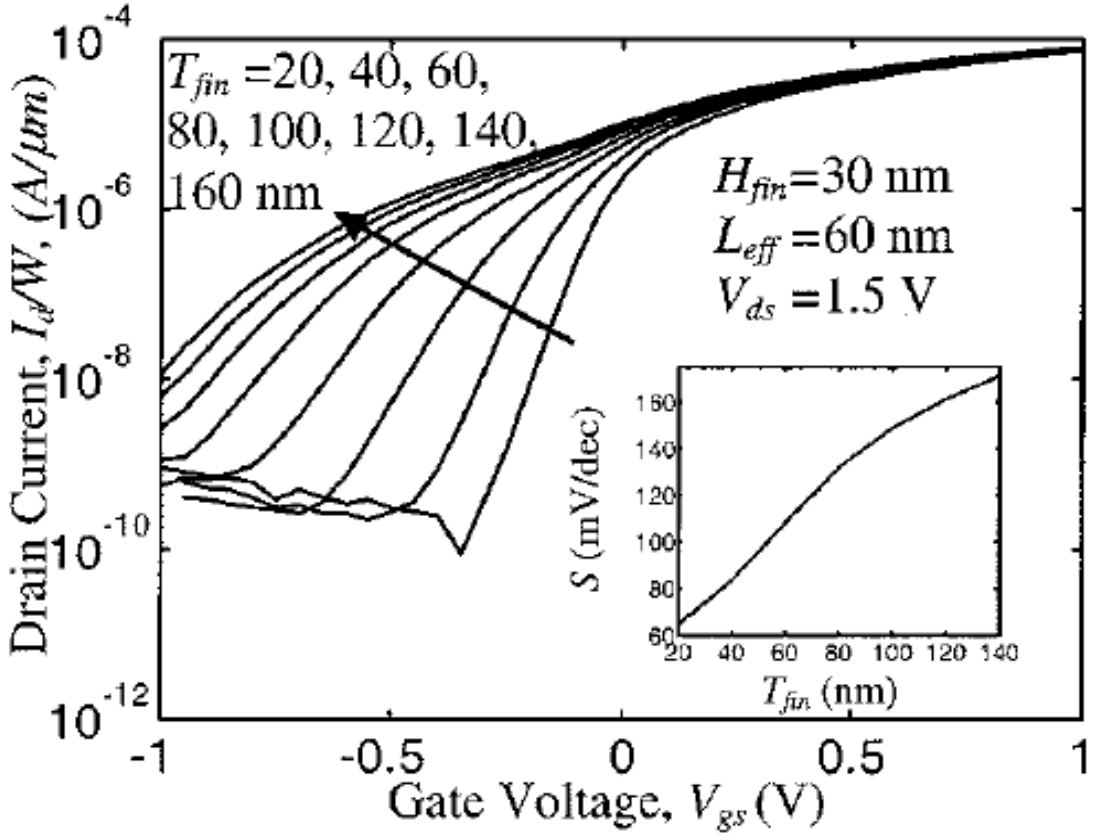
Fig. 1. Schematic of a FinFET structure.

- Effective channel width $W = (T_{fin}) + 2 \times H_{fin}$

•Gen Pei, et al., *IEEE Trans on Electron Dev.*, vol.49, no.8, p1411-1419, 2002

•H -J L Gossmann, et al., *IEEE Trans on Nanotechnology*, vol.2, no.4, p 285-290, 2003

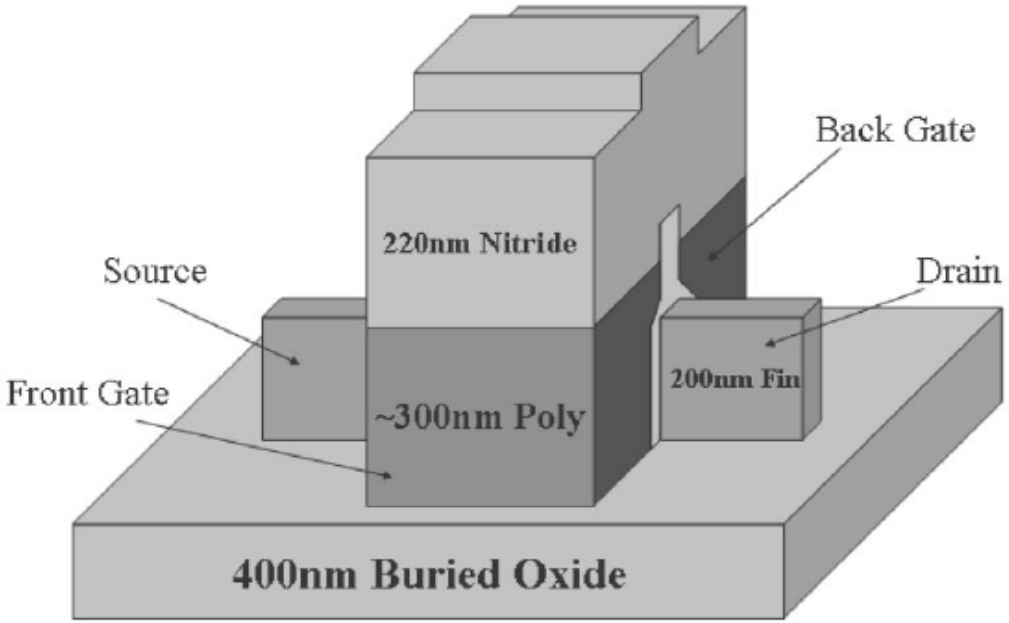
Design - Dependence of sub-threshold I-V characteristic on T_{fin}

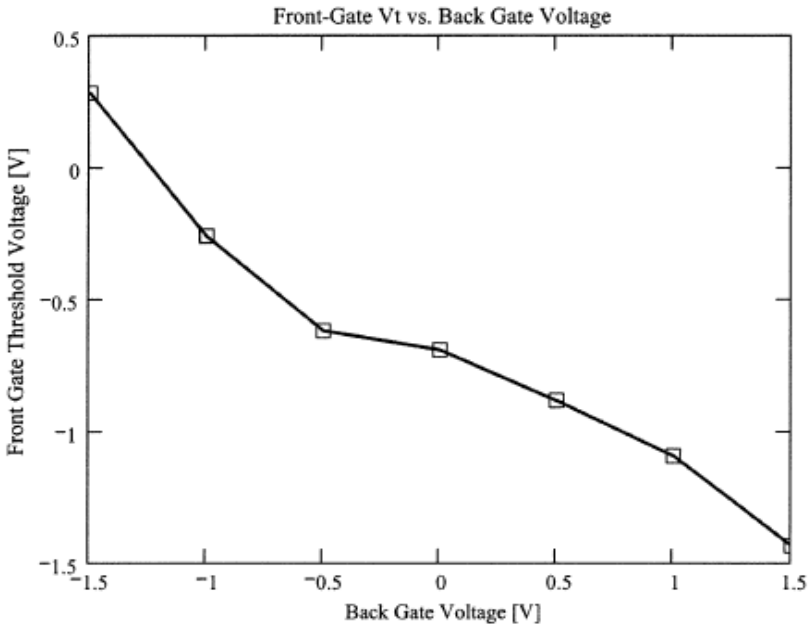


- Very good sub-threshold characteristics for thin silicon layer
- Increasing the thickness of the silicon layer leads to largely degraded short-channel effects

•Gen Pei, et al., *IEEE Trans on Electron Dev.*, vol.49, no.8, p1411-1419, 2002

Independent-Gates FinFET





- The adoption of two separated gate electrodes provides additional control on devices' characteristics (e.g. leakage control)

- M. Fried et al., IEEE TED, Vol. 24, No. 9, 2003

FinFETs - taking advantage of different crystal orientations

- FinFETs can be easily fabricated outside of the traditional (100) plane, allowing to change the crystal orientation to optimize CMOS circuit performance.
- Alternative surface orientations such as (110) and (111) enhance hole mobility while degrading electron mobility, thus allowing for adjustment of the ratio between n- and pMOS transistor drive currents.
- By optimizing the surface orientation, up to a 15% improvement in gate delay can be expected. This value depends upon the type of logic gate, the off-state leakage specification, and technology scaling trends.

Effect of crystal orientation on low-field mobility

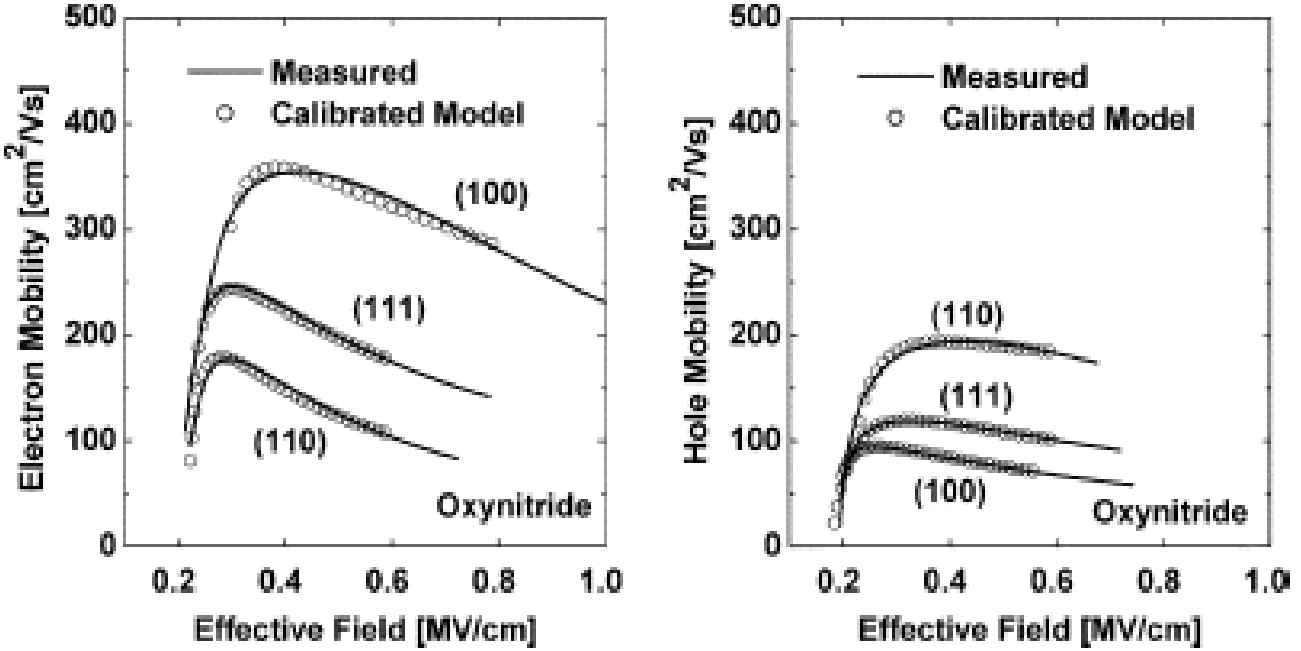


Fig. 1. Electron and hole mobility for (100), (110), and (111) surface orientations. With the use of alternative surface orientations, hole mobility is enhanced while electron mobility is degraded. The direction of current flow is assumed to be perpendicular to the wafer notch. The calibrated mobility models show good agreement with measured data.

•L. Chang et al. IEEE Trans. Elec. Dev. Vol. 51, No. 10, 2004.

Effect of crystal orientation on circuit speed – flexibility offered by FinFETs

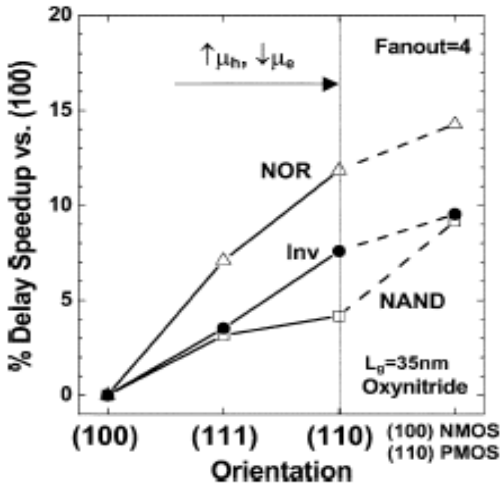


Fig. 3. Gate delay can be improved by moving away from a standard (100) surface orientation due to an enhancement in hole mobility. If a single orientation is used for both nMOS and pMOS devices, degradation in electron mobility reduces the improvement in performance, especially for a NAND gate, which has stacked nMOS devices. However, if different orientations are allowed for the two device types, a significant increase in performance can be expected, especially for a NOR gate, which has stacked pMOS devices.

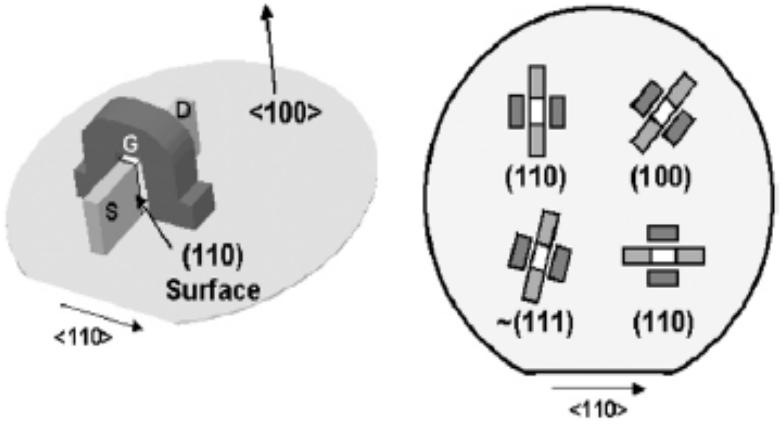


Fig. 10. Because the FinFET device structure is vertical in nature, rotation in the plane of a standard (100) wafer allows for simple modification of the surface orientation. When the device is oriented parallel or perpendicular to the wafer flat, the device lies in the (110) plane. At a 45° angle, the device is in the (100) plane. With an intermediate rotation, the electron and hole mobilities are at intermediate values, which can approximate the (111) plane.

•L. Chang et al. IEEE Trans. Elec. Dev. Vol. 51, No. 10, 2004.

Performance - IV Characteristics

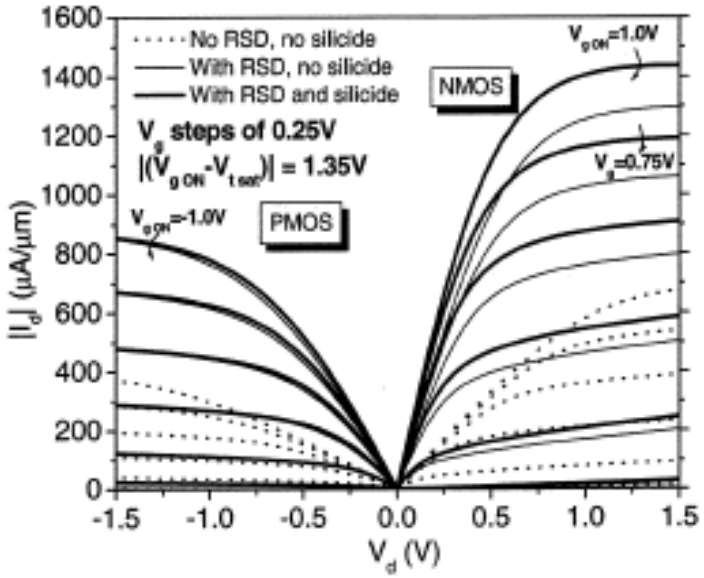


Fig. 13. $I_d - V_d$ plot for short channel, $L_{\text{eff}} = 30$ nm, $T_{\text{si}} = 20$ nm, $\langle 100 \rangle$ directed FiIIFET devices without RSD, with RSD, and with RSD and suicide. NFET V_g varies from -0.5 V to 1.0 V in steps of 0.25 V.

•Optimization of S/D and extensions leads to ON-currents comparable

•to those of bulk MOSFETs with low short-channel effects.

•The performance of the p-FinFET is very good because the hole mobility

•in the (110) channel is enhanced

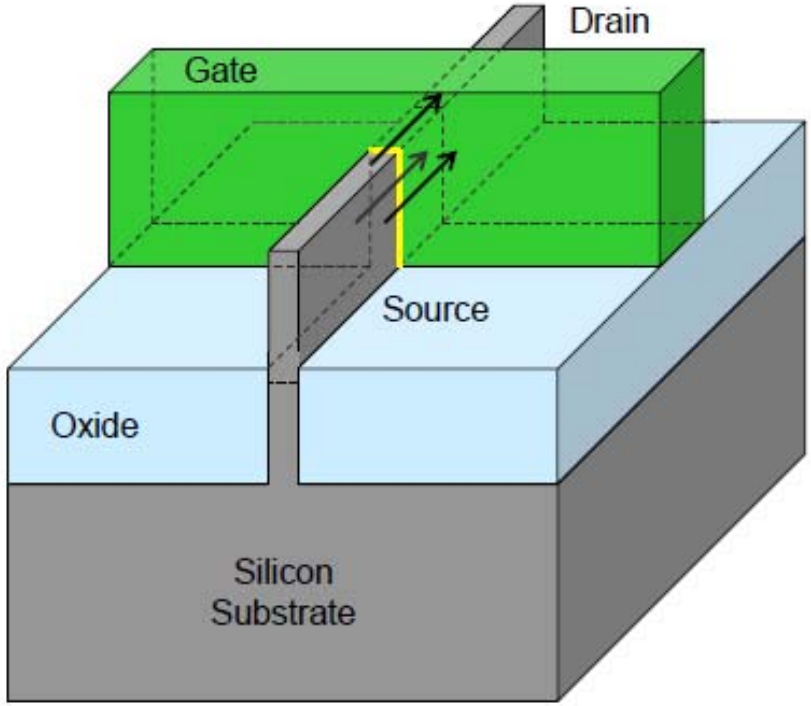
•J. Kedzierski et al. IEEE Trans. Elec. Dev, Vol. 50, No. 4, 2003

Summary

•"Easy in concept----Tough to build"

- Double-gate FET can reduce Short Channel Effects and FinFET is the leading DGFET
- Optimization design includes geometry, S-D fin-extension doping, dielectric thickness scaling, threshold voltage control….
- Fabrication of FinFET is compatible with CMOS process

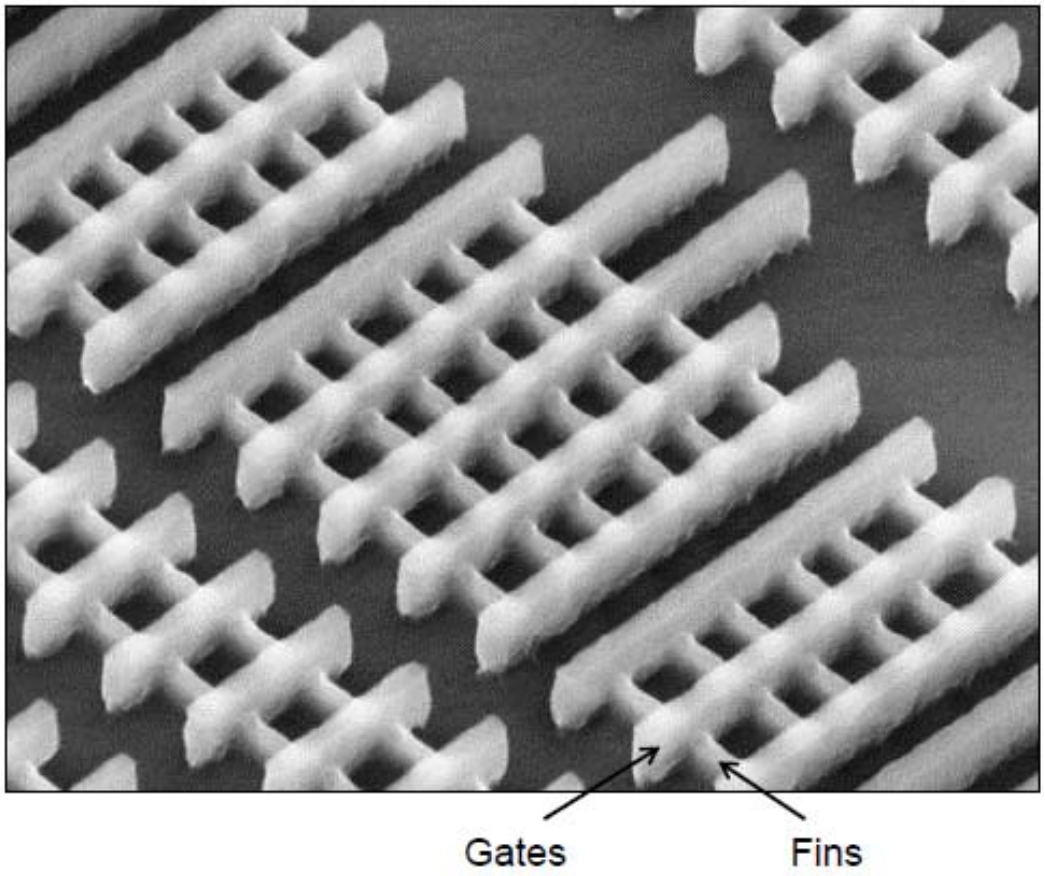
22 nm Tri-Gate Transistor



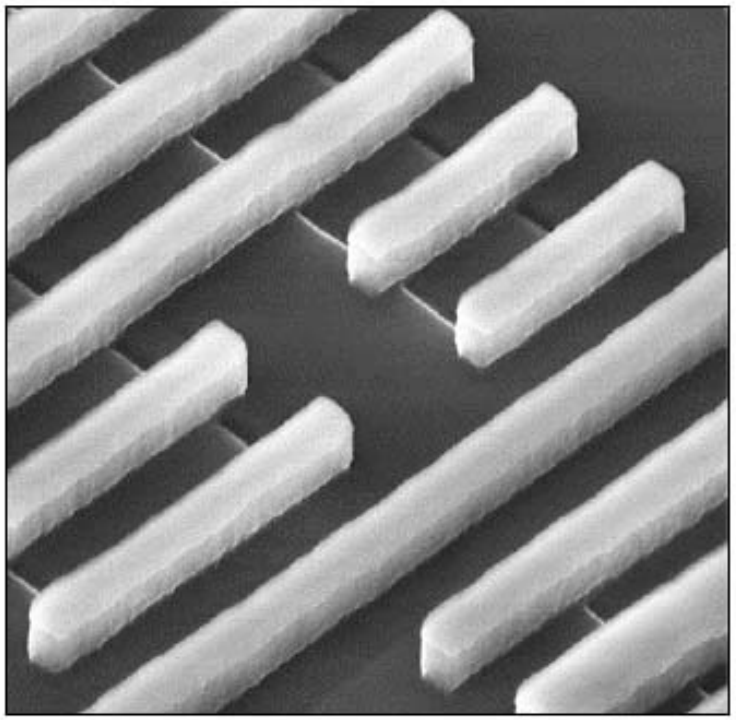
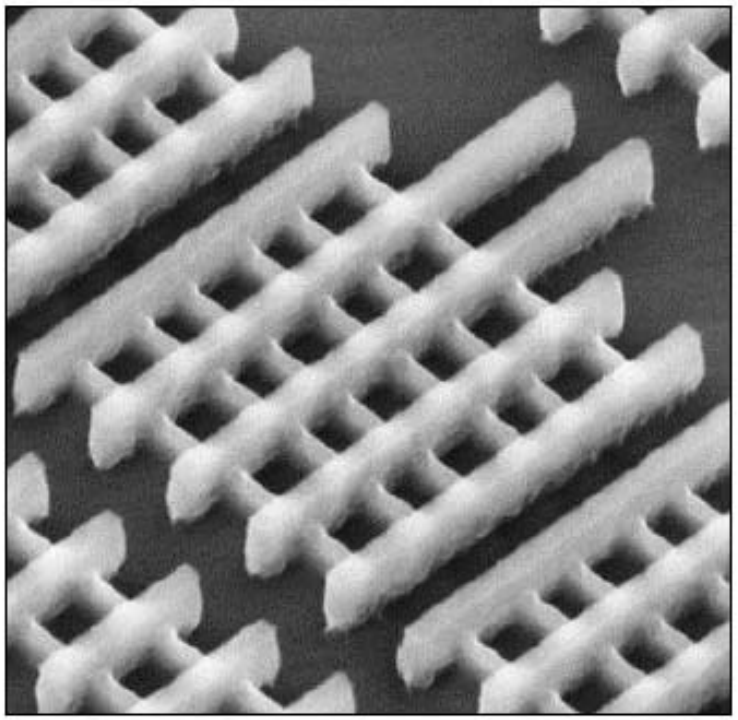
3-D Tri-Gate transistors form conducting channels on three sides of a vertical fin structure, providing "fully depleted" operation

Transistors have now entered the third dimension!

22 nm Tri-Gate Transistor



32 nm Planar Transistors



22 nm Tri-Gate Transistors