



EE-0715652

Solid state devices

Prepared by Prof. M. El-Banna

Outline

Part I: Semiconductor Physics

Chapter 01: Physics and Properties of Semiconductors – a Review

Part II: Device Building Blocks

Chapter 02: p-n Junctions

Chapter 03: Metal-Semiconductor Contacts

Chapter 04: Metal-Insulator-Semiconductor Capacitors

Part III: Transistors

Chapter 06: MOSFETs

Chapter 1: Physics and Properties of Semiconductors

- 1.1 Introduction**
- 1.2 Crystal Structure**
- 1.3 Energy Bands and Energy Gap**
- 1.4 Carrier Concentration at Thermal Equilibrium**
- 1.5 Carrier-Transport Phenomena**
- 1.6 Phonon, optical, and thermal properties**
- 1.7 Heterojunctions and nanostructures**

1.1 Introduction

- This chapter focuses on the two most-important semiconductors: silicon (Si) and gallium arsenide (GaAs).
- The silicon is the basic material in all commercial electronics products.
- The GaAs
 - Direct band gap for photonic applications (GaAs).
 - Generating microwaves because of intervalley-carrier transport and higher mobility .

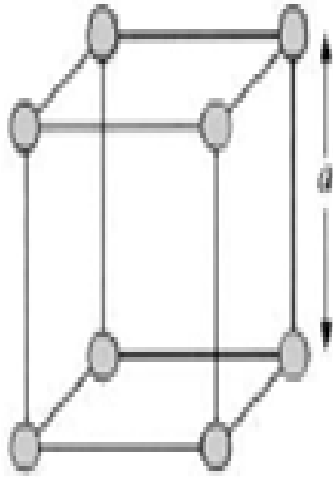
1.2 Crystal Structure

- ▶ 1.2.1 Primitive Cell and Crystal Planes:
 - ▶ A crystal is characterized by having a well-structured periodic placement of atoms.
 - ▶ The smallest assembly of atoms that can be repeated to form the entire crystal is called a primitive cell.
 - ▶ The distance between atoms is “a” which is defined as lattice constant.
 - ▶ The bond between two nearest neighbors is formed by two electrons with opposite spins

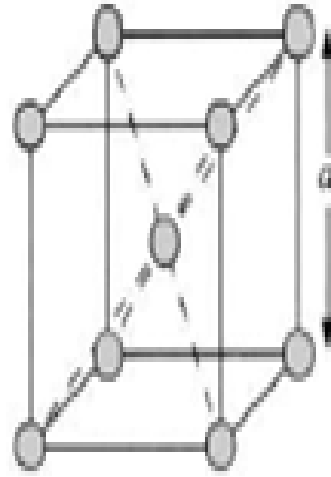
1.2 Crystal Structure

Unit cells

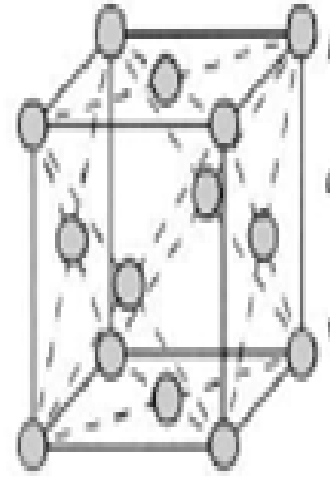
Basic primitive cells of some solid state materials



Simple cubic
(Po)
(a)



Body-centered cubic
(Na, W, etc.)
(b)

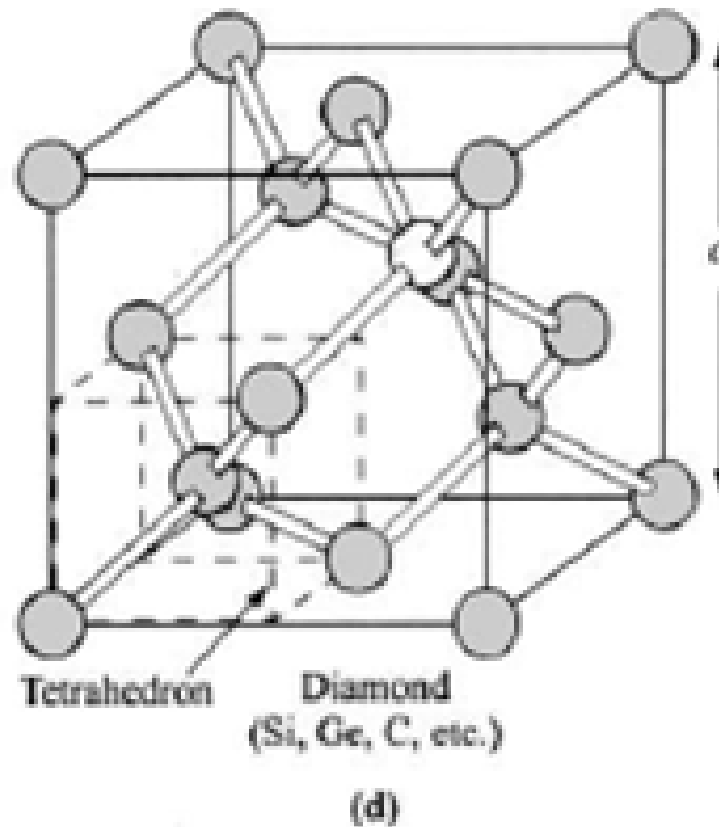


Face-centered cubic
(Al, Au, etc.)
(c)

1.2 Crystal Structure

The diamond structure is considered as two interpenetrating face-centered cubic (fcc) lattices.

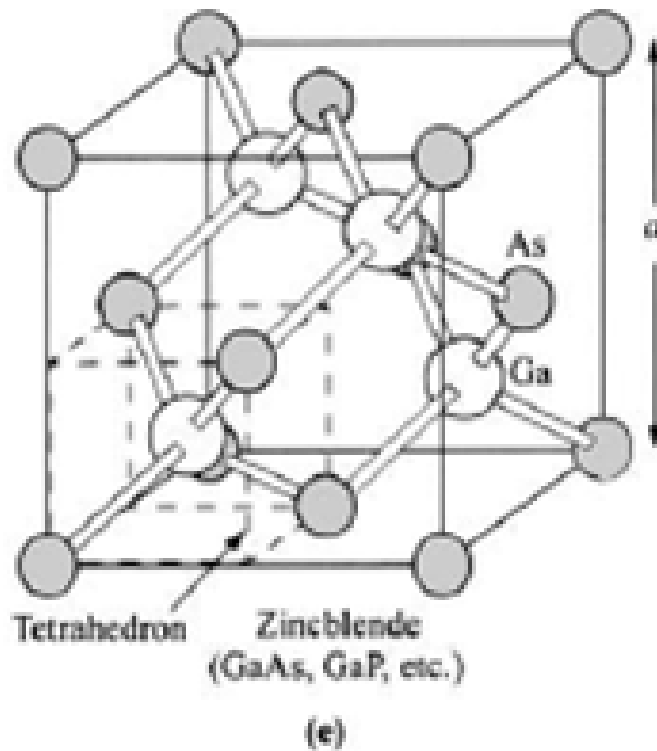
All atoms are alike (Si)



1.2 Crystal Structure

The zincblende structure is considered as two interpenetrating face-centered cubic (fcc) lattices.

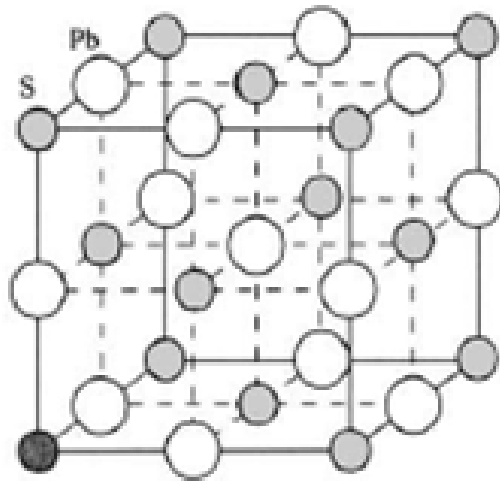
One atom is Ga and the other is As (GaAs)



1.2 Crystal Structure

The rock-salt lattice, which again can be considered as two interpenetrating face-centered cubic lattices.

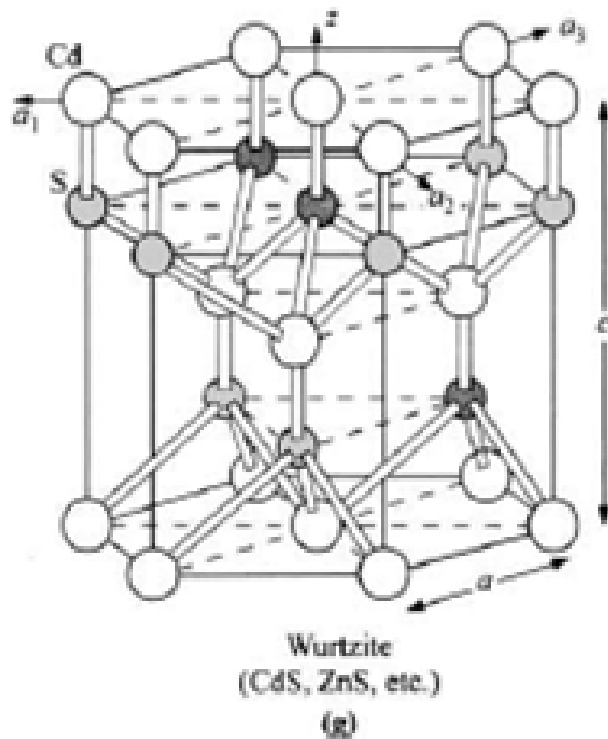
Each atom has six nearest neighbors



Rock-salt
(PbS, PbTe, etc.)
(f)

1.2 Crystal Structure

The wurtzite lattice, which can be considered as two interpenetrating hexagonal close-packed lattices.



1.2 Crystal Structure

Miller indices

- ▶ The orientations and properties of the surface crystal planes are important.
- ▶ They are determined by :
 - ▶ Finding the intercepts of the plane with the three basis axes in terms of the lattice constants.
 - ▶ Taking the reciprocals of primitive axis and reducing them to the smallest three integers having the same ratio.
 - ▶ The result is enclosed in form (hkl) called the Miller indices
 - ▶ Three primitive basis vectors, a , b , and c of a primitive cell.

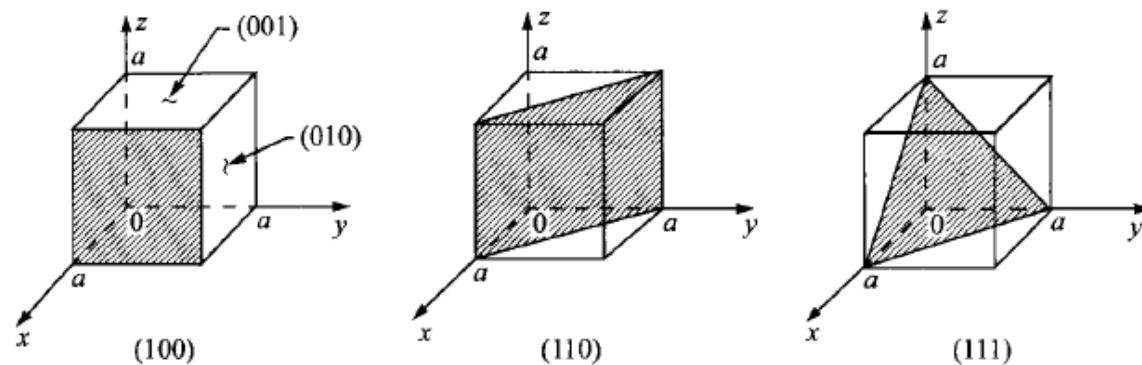


Fig. 2 Miller indices of some important planes in a cubic crystal.

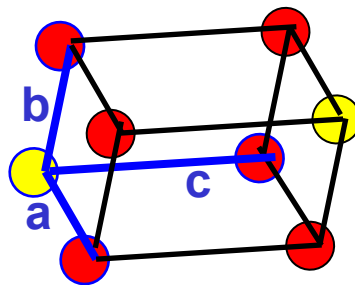
1.2 Crystal Structure

Three primitive basis vectors, **a**, **b**, and **c** of a primitive cell, describe a crystalline solid.

The crystal structure remains invariant under translation through any vector of the form:

$$\mathbf{R} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c}$$

where **m**, **n**, and **p** are integers.



1.2.2 Reciprocal Lattice

- For a given set of the direct basis vectors, a set of reciprocal lattice basis vectors $\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*$ can be defined as:

$$\mathbf{a}^* \equiv 2\pi \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}},$$

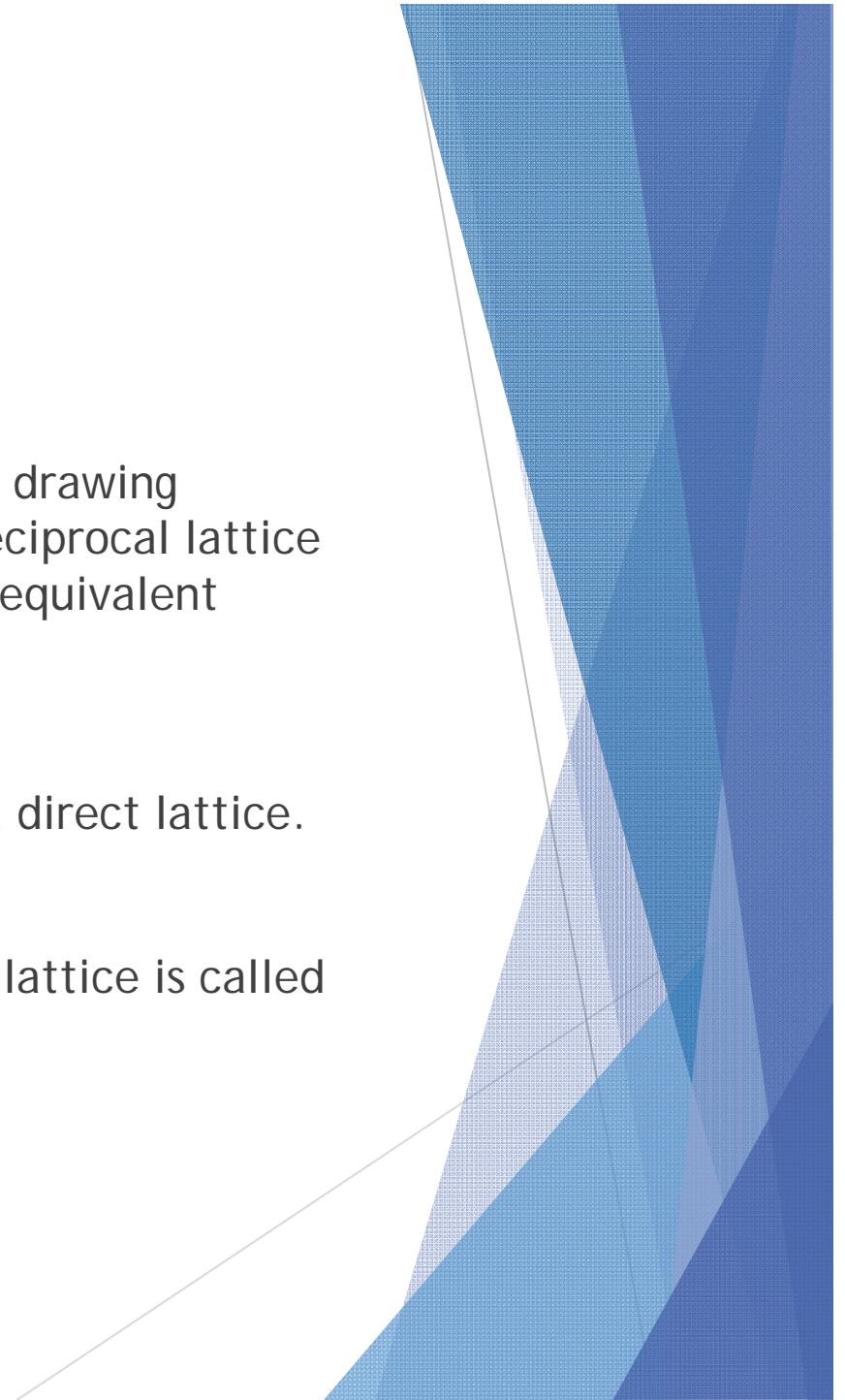
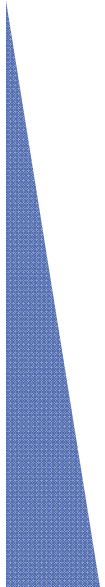
$$\mathbf{b}^* \equiv 2\pi \frac{\mathbf{c} \times \mathbf{a}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}},$$

$$\mathbf{c}^* \equiv 2\pi \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}$$

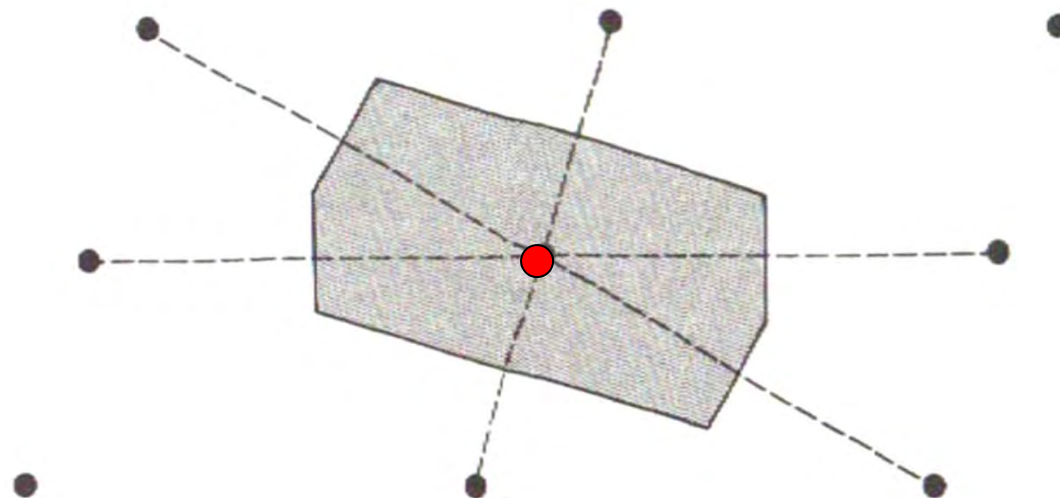
- $\mathbf{a} \cdot \mathbf{a}^* = 2\pi, \mathbf{a} \cdot \mathbf{b}^* = 0,$
- The denominators are identical due to the equality that
- $\mathbf{a} \cdot \mathbf{b} \times \mathbf{c} = \mathbf{b} \cdot \mathbf{c} \times \mathbf{a} = \mathbf{c} \cdot \mathbf{a} \times \mathbf{b}$
- The general reciprocal lattice vector is given by
- $\mathbf{G} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ where hkl are integers (miller).
- $\mathbf{G} \cdot \mathbf{R} = 2\pi \times \text{Integer}$

First Brillouin zone

- ▶ The Wigner-Seitz cell is constructed by drawing perpendicular bisector planes in the Reciprocal lattice from the chosen center to the nearest equivalent reciprocal lattice sites.
- ▶ This technique can also be applied to a direct lattice.
- ▶ The Wigner-Seitz cell in the reciprocal lattice is called the first Brillouin zone.

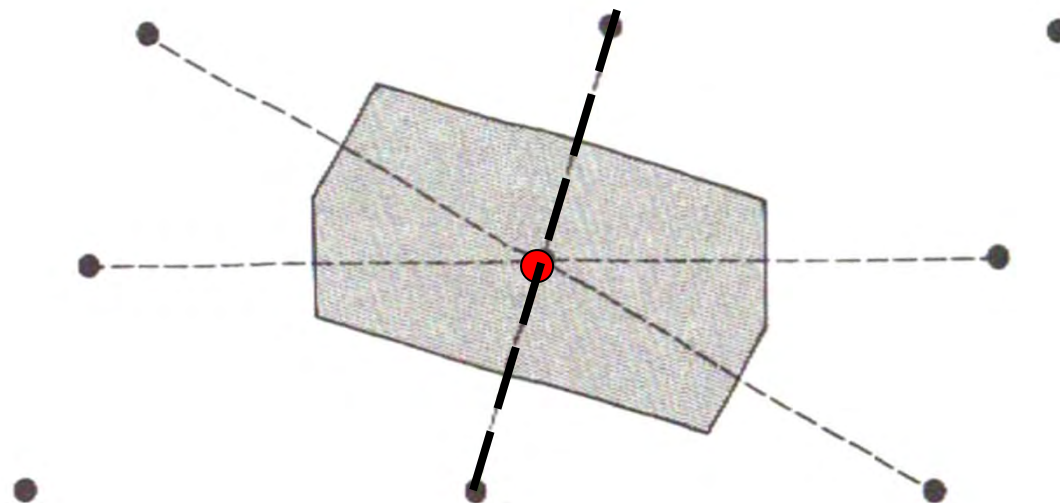


2D example of how to find a Wigner Seitz cell:



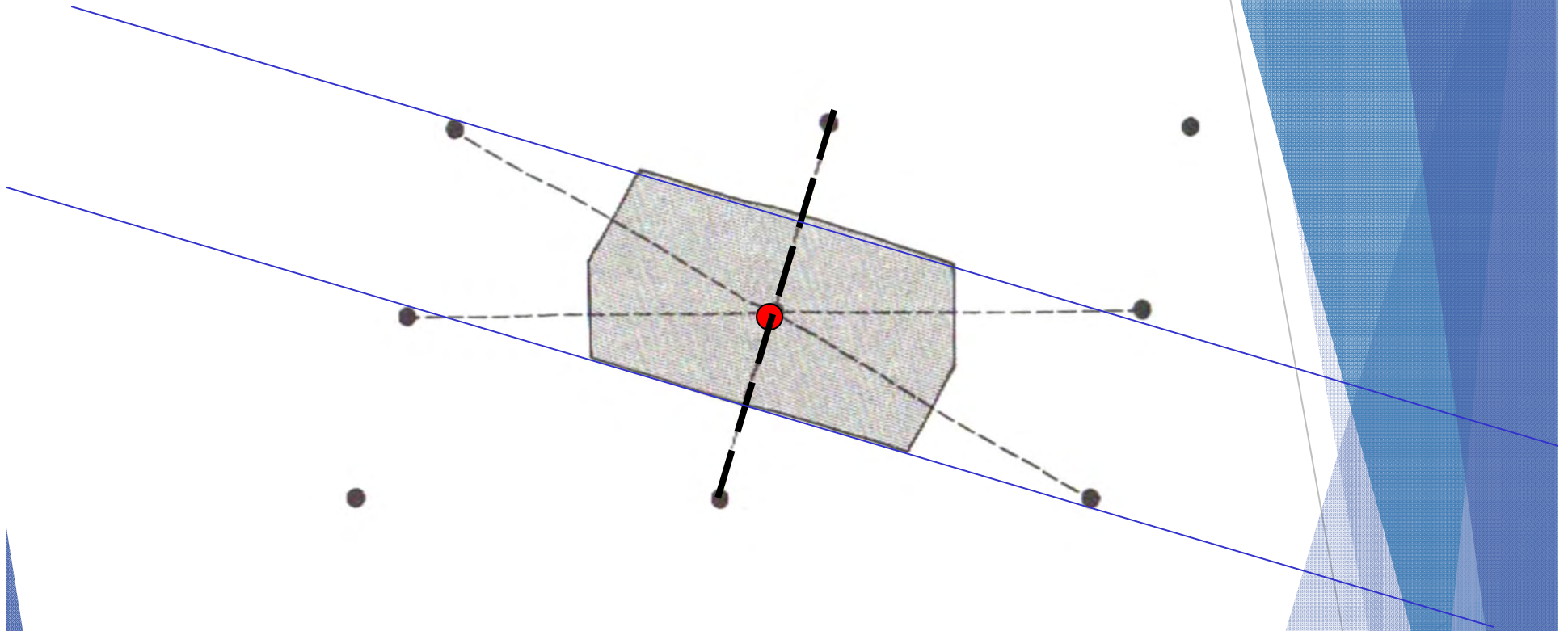
Pick center

2D example of how to find a Wigner Seitz cell:



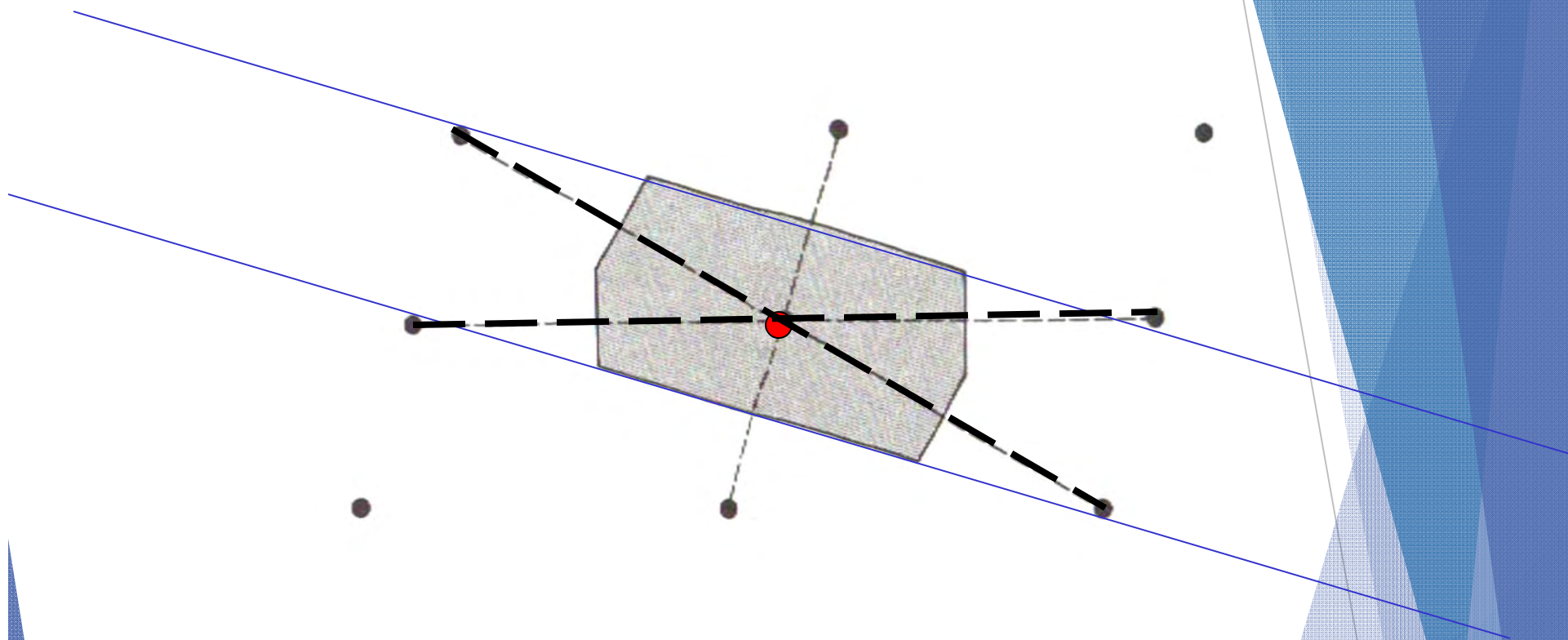
Nearest neighbors

2D example of how to find a Wigner Seitz cell:



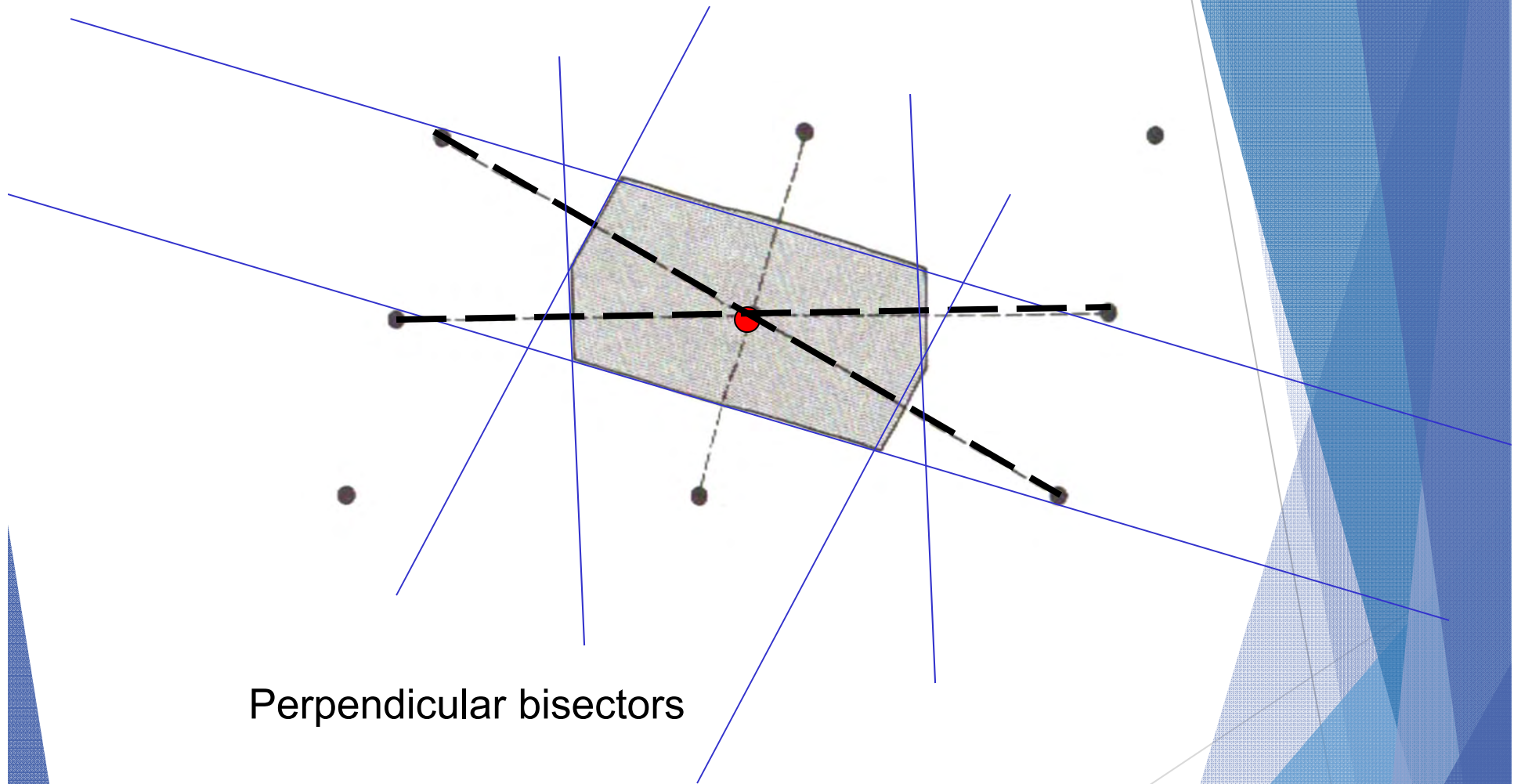
Perpendicular bisectors (represents a plane)

2D example of how to find a Wigner Seitz cell:



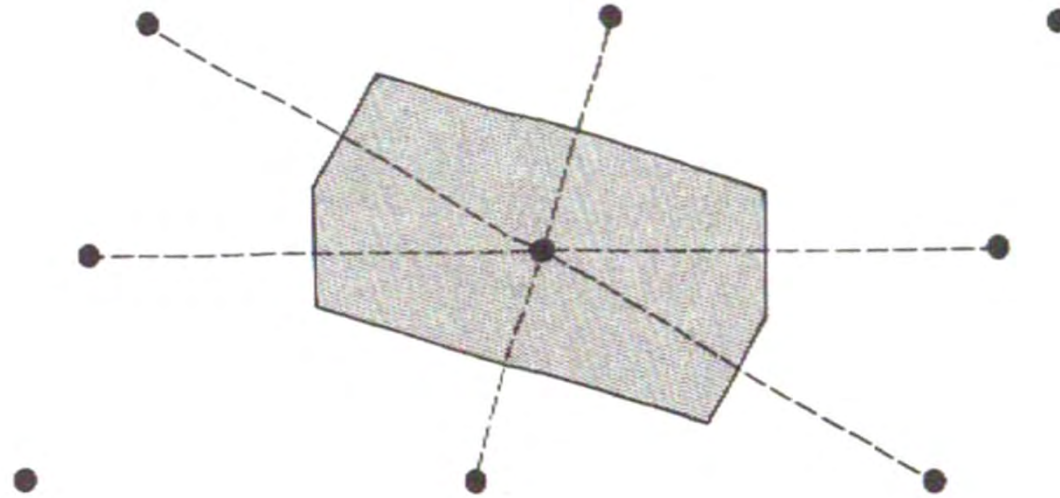
Next nearest neighbors

2D example of how to find a Wigner Seitz cell:



Perpendicular bisectors

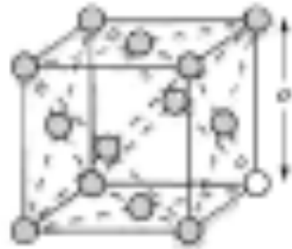
2D example of how to find a Wigner Seitz cell:



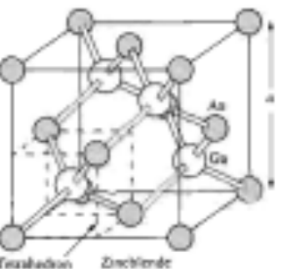
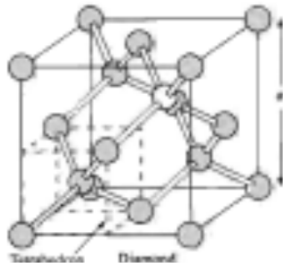
Wigner Seitz cell is the shaded area (in 2D)
Can do this in direct space or reciprocal space

Direct space (lattice)

Conventional cubic Unit cell

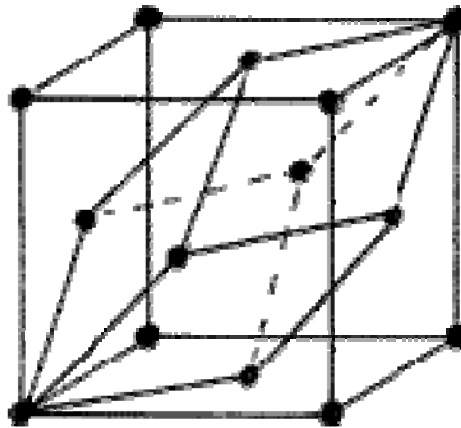


Face-centered cubic



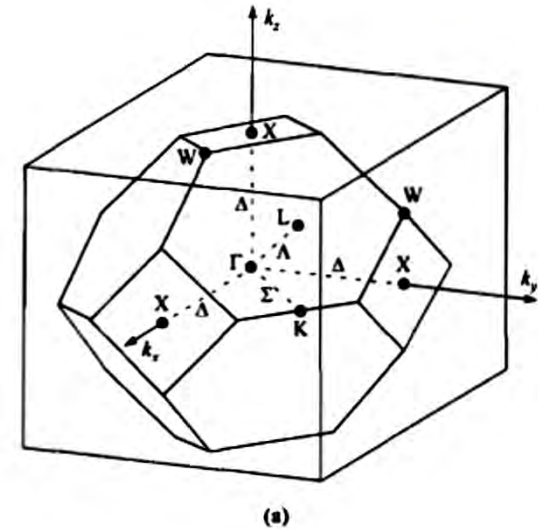
Direct space (lattice)

Primitive cell for:
fcc, diamond, zinc-blende,
and rock salt



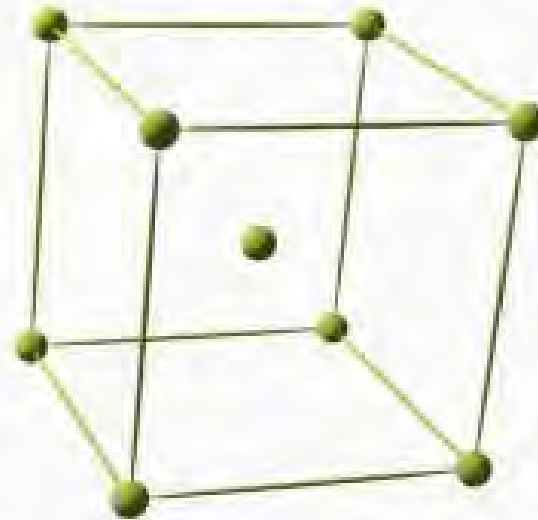
Reciprocal space (lattice)

Reciprocal space = first
Brillouin zone for:
fcc, diamond, zinc-blende,
and rock salt



The reciprocal lattice is useful to visualize the E - k relationship when the coordinates of the wave vectors k ($|\mathbf{k}| = k = 2\pi/\lambda$) are mapped into the coordinates of the reciprocal lattice to draw the bandgap.

The bcc structure



conventional unit cell

1.3 Energy Bands and Energy Gap

- ▶ The band structure of a solid is the energy-momentum ($E-k$) relationship for carriers in a lattice and is used to determine whether the material emits a photon or phonon.
- ▶ It also characterizes the effective mass and the group velocity.
- ▶ By solving the Schrodinger equation of an approximate one-electron problem
- ▶ The Bloch theorem, one of the most-important theorems basic to band structure.
- ▶ It states that if a potential energy $V(\mathbf{r})$ is periodic in the direct lattice space, then the solutions for the wave function $\Psi(\mathbf{r},\mathbf{k})$ of the Schrodinger

$$\left[-\frac{\hbar^2}{2m^*} \nabla^2 + V(\mathbf{r}) \right] \psi(\mathbf{r}, \mathbf{k}) = E(\mathbf{k}) \psi(\mathbf{r}, \mathbf{k})$$

1.3 Energy Bands and Energy Gap

are of the form of a Bloch function

$$\psi(\mathbf{r}, \mathbf{k}) = \exp(j\mathbf{k} \cdot \mathbf{r})U_b(\mathbf{r}, \mathbf{k})$$

- ▶ Here b is the band index, $\Psi(\mathbf{r}, \mathbf{k})$ and $U_b(\mathbf{r}, \mathbf{k})$ are periodic in \mathbf{R} of the direct lattice.

- ▶ Since
$$\begin{aligned}\psi(\mathbf{r} + \mathbf{R}, \mathbf{k}) &= \exp[j\mathbf{k} \cdot (\mathbf{r} + \mathbf{R})]U_b(\mathbf{r} + \mathbf{R}, \mathbf{k}) \\ &= \exp(j\mathbf{k} \cdot \mathbf{r})\exp(j\mathbf{k} \cdot \mathbf{R})U_b(\mathbf{r}, \mathbf{k}),\end{aligned}$$

$$\mathbf{G} \cdot \mathbf{R} = 2\pi \times \text{Integer},$$

- ▶ Note that: $\mathbf{k} \cdot \mathbf{R} = 2n\pi$
- ▶ \mathbf{G} is replaced with \mathbf{k} for visualizing the E - \mathbf{k} relationship
- ▶ Energy $E(\mathbf{k})$ is periodic in the reciprocal lattice, that is, $E(\mathbf{k}) = E(\mathbf{k} + \mathbf{G})$ so that by solving $E(\mathbf{k})$ bottom of E_c and top of E_v the E - \mathbf{k} relationship can be approximated by a quadratic equation:

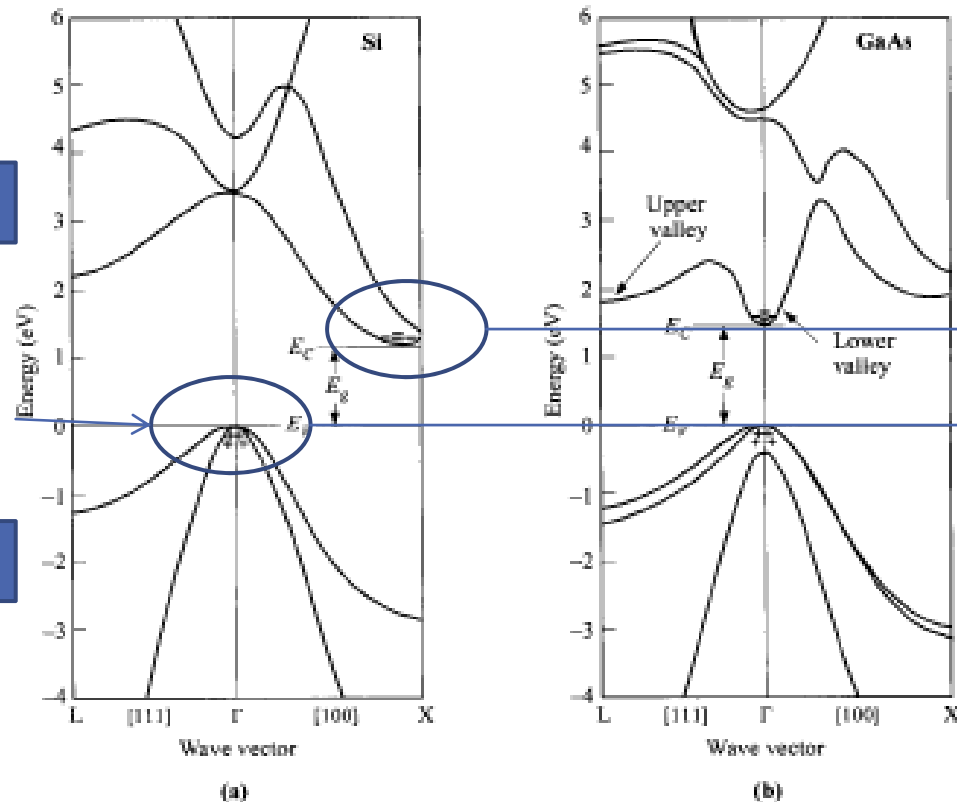
$$E(\mathbf{k}) = \frac{\hbar^2 \mathbf{k}^2}{2m^*},$$

- ▶ where m^* is the associated effective mass

Conduction band

Two parabolic bands

Valance band



Heavy electron band

Heavy hole band

Fig. 4 Energy-band structures of (a) Si and (b) GaAs, where E_g is the energy bandgap. Plus signs (+) indicate holes in the valence bands and minus signs (-) indicate electrons in the conduction bands. (After Ref. 20.)

- ▶ The effective mass in general is tensorial with components defined as: $\frac{1}{m_{ij}^*} \equiv \frac{1}{\hbar^2} \frac{\partial^2 E(k)}{\partial k_i \partial k_j}$
- ▶ Carriers in motion are also characterized by a group velocity:

$$v_g = \frac{1}{\hbar} \frac{dE}{dk}$$

1.3 Energy Bands and Energy Gap

- ▶ Considering that the valence-band maximum (E_v) occurs at Γ .
- ▶ The conduction-band minimum can be aligned or misaligned in k -space in determining the bandgap.
- ▶ If its aligned then it is a direct band gap as in (GaAs) if not then it is an indirect band gap (Si).
- ▶ This bears significant consequences when carriers transfer between this minimum gap in that momentum (or k) is conserved for direct band gap but changed for indirect bandgap

1.3 Energy Bands and Energy Gap

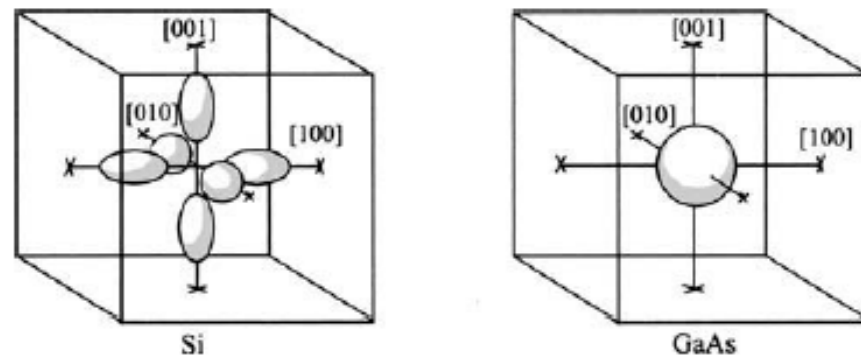


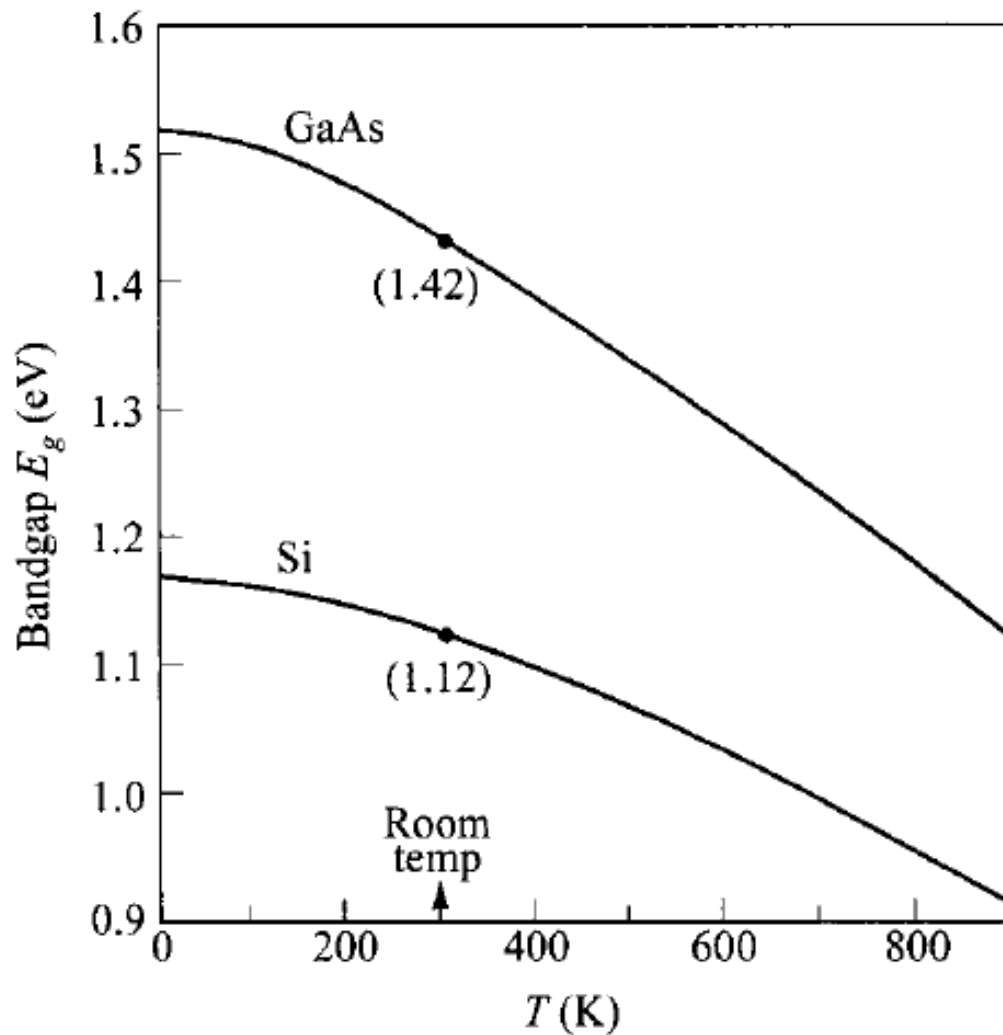
Fig. 5 Shapes of constant-energy surfaces for electrons in Si and GaAs. For Si there are six ellipsoids along the $\langle 100 \rangle$ -axes with the centers of the ellipsoids located at about three-fourths of the distance from the Brillouin zone center. For GaAs the constant-energy surface is a sphere at zone center. (After Ref. 21.)

- ▶ This results in obtaining the electron effective masses; one for GaAs and two for Si, m^* along the symmetry axes and m^* transverse to the symmetry axes.

1.3 Energy Bands and Energy Gap

$$E_{gap}(T) \approx E_{gap}(0K) - \frac{\alpha T^2}{T + \beta}$$

- At room temperature, the values of the bandgap are 1.12 eV for Si and 1.42 eV for GaAs due to impurity, the temperature coefficient dE/dT is negative for both semiconductors
- It is an approximate solution in case of Si and GaAs by supposing that it has high purity, the energy gap will decrease



	$E_g(0)$ (eV)	α (eV/K)	β (K)
GaAs	1.519	5.4×10^{-4}	204
Si	1.169	4.9×10^{-4}	655

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T + \beta}$$

Fig. 6 Energy bandgaps of Si and GaAs as a function of temperature. (After Refs. 22–23.)

Appendix F Properties of Important Semiconductors

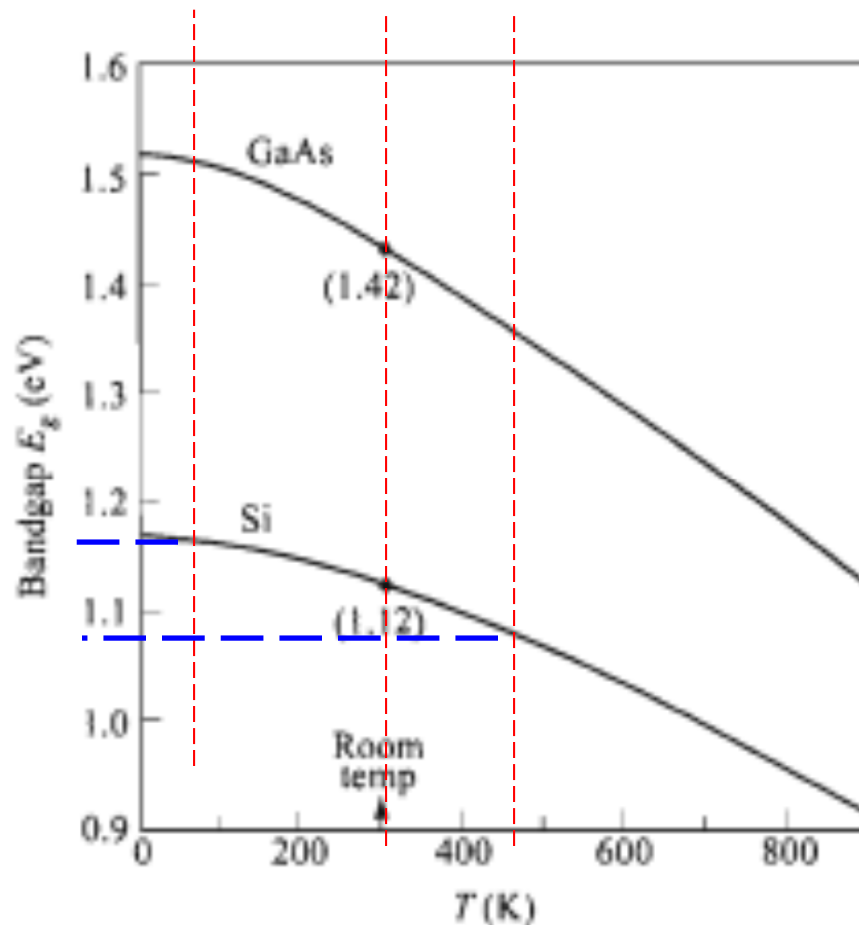
Semiconductor		Crystal Struct.	Lattice Const. at 300 K (Å)	Bandgap (eV)		Band	Mobility at 300 K (cm ² /V-s)		Effective Mass		ϵ_s/ϵ_0
				300 K	0 K		μ_n	μ_p	m_n^*/m_0	m_p^*/m_0	
C	Carbon (diamond)	D	3.56683	5.47	5.48	I	1,800	1,200	0.2	0.25	5.7
Ge	Germanium	D	5.64613	0.66	0.74	I	3,900	1,900	1.64 ^l , 0.082 ^t	0.04 ^{lh} , 0.28 ^{hh}	16.0
Si	Silicon	D	5.43102	1.12	1.17	I	1,450	500	0.98 ^l , 0.19 ^t	0.16 ^{lh} , 0.49 ^{hh}	11.9
IV-IV	SiC Silicon carbide	W	$a=3.086, c=15.117$	2.996	3.03	I	400	50	0.60	1.00	9.66
III-V	AlAs Aluminum arsenide	Z	5.6605	2.36	2.23	I	180		0.11	0.22	10.1
	AlP Aluminum phosphide	Z	5.4635	2.42	2.51	I	60	450	0.212	0.145	9.8
	AlSb Aluminum antimonide	Z	6.1355	1.58	1.68	I	200	420	0.12	0.98	14.4
	BN Boron nitride	Z	3.6157	6.4		I	200	500	0.26	0.36	7.1
	" "	W	$a=2.55, c=4.17$	5.8		D			0.24	0.88	6.85
	BP Boron phosphide	Z	4.5383	2.0		I	40	500	0.67	0.042	11
	GaAs Gallium arsenide	Z	5.6533	1.42	1.52	D	8,000	400	0.063	0.076 ^{lh} , 0.5 ^{hh}	12.9
	GaN Gallium nitride	W	$a=3.189, c=5.182$	3.44	3.50	D	400	10	0.27	0.8	10.4
	GaP Gallium phosphide	Z	5.4512	2.26	2.34	I	110	75	0.82	0.60	11.1
	GaSb Gallium antimonide	Z	6.0959	0.72	0.81	D	5,000	850	0.042	0.40	15.7
	InAs Indium arsenide	Z	6.0584	0.36	0.42	D	33,000	460	0.023	0.40	15.1
	InP Indium phosphide	Z	5.8686	1.35	1.42	D	4,600	150	0.077	0.64	12.6
	InSb Indium antimonide	Z	6.4794	0.17	0.23	D	80,000	1,250	0.0145	0.40	16.8
	II-VI	CdS Cadmium sulfide	Z	5.825	2.5		D			0.14	0.51
" "		W	$a=4.136, c=6.714$	2.49		D	350	40	0.20	0.7	9.1
CdSe Cadmium selenide		Z	6.050	1.70	1.85	D	800		0.13	0.45	10.0
CdTe Cadmium telluride		Z	6.482	1.56		D	1,050	100			10.2
ZnO Zinc oxide		R	4.580	3.35	3.42	D	200	180	0.27		9.0
ZnS Zinc sulfide		Z	5.410	3.66	3.84	D	600		0.39	0.23	8.4
" "	W	$a=3.822, c=6.26$	3.78		D	280	800	0.287	0.49	9.6	
IV-VI	PbS Lead sulfide	R	5.9362	0.41	0.286	I	600	700	0.25	0.25	17.0
	PbTe Lead telluride	R	6.4620	0.31	0.19	I	6,000	4,000	0.17	0.20	30.0

D = Diamond, W = Wurtzite, Z = Zincblende, R = Rock salt. I, D = Indirect, direct bandgap. *l, t, lh, hh* = Longitudinal, transverse, light-hole, heavy-hole effective mass.

Physics of Semiconductor Devices, 3rd Edition
 by S. M. Sze and Kwok K. Ng
 Copyright © 2007 John Wiley & Sons, Inc.

Example problem:

A satellite in low earth orbit has a temperature swing of +200°C sun side to - 200°C dark side over 24 h. Its electronics are Si-based. Find the range of E_{gap} and compare it to operation on earth.



$$T = 200^{\circ}\text{C} = 473 \text{ K}$$

$$T = -200^{\circ}\text{C} = 73 \text{ K}$$

$$E_{\text{gap}} (73 \text{ K}): \text{graph estimate: } 1.16 \text{ eV}$$

$$E_{\text{gap}} (300 \text{ K}): \text{on graph: } 1.12 \text{ eV}$$

$$E_{\text{gap}} (473 \text{ K}): \text{graph estimate: } 1.08 \text{ eV}$$

1.4 Carrier Concentration at Thermal Equilibrium

- ▶ Doping with different types and concentrations of impurities vary the semiconductor's resistivity.
- ▶ Intrinsic :very pure and contains a negligibly small amount of impurities
- ▶ n-type Si with donor (phosphorus).
- ▶ p-type Si with acceptor (boron).
- ▶ silicon atom shares its four valence electrons with the four neighbors .

1.4 Carrier Concentration at Thermal Equilibrium

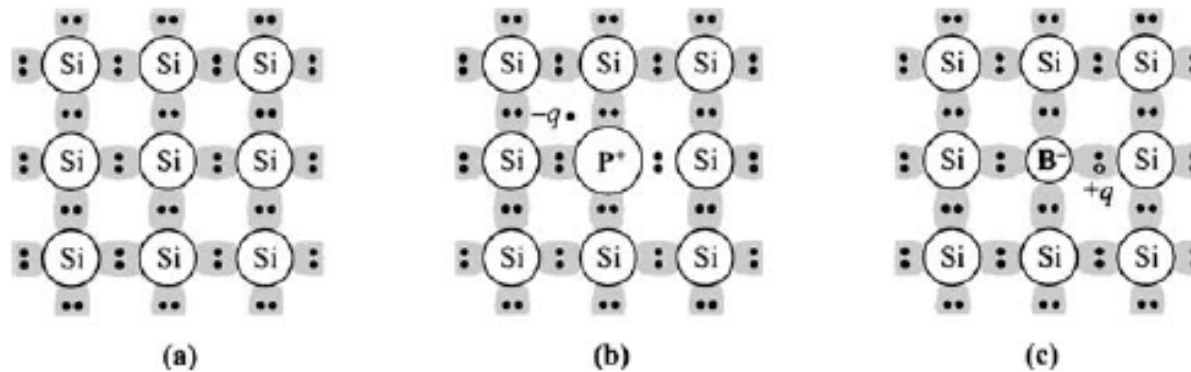



Fig. 7 Three basic bond pictures of a semiconductor. (a) Intrinsic Si with no impurity. (b) *n*-type Si with donor (phosphorus). (c) *p*-type Si with acceptor (boron).

1.4.1 Carrier Concentration and Fermi Level

- ▶ In intrinsic case (without impurities added)
- ▶ The number of electrons in conduction-band levels is given by the total number of energy states $N(E)$ multiplied by the probability of occupancy $F(E)$, integrated over the conduction band:

$$n = \int_{E_C}^{\infty} N(E) F(E) dE.$$



$$N(E) = M_C \frac{\sqrt{2} m_{de}^{3/2} (E - E_C)^{1/2}}{\pi^2 \hbar^3}.$$

- ▶ Where M_C is the number of equivalent minima in the conduction band
- ▶ m_{de} is the density - of- state effective mass for electrons:

$$m_{de} = (m_1^* m_2^* m_3^*)^{1/3}$$

- ▶ Si $m_{de} = (m_t^* m_t^{*2})^{1/3}$

1.4 Carrier Concentration at Thermal Equilibrium

- ▶ Fermi-Dirac distribution function (F(E)):

$$F(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

- ▶
$$n = \int_{E_C}^{\infty} \frac{(E - E_C)^{\frac{1}{2}}}{(1 + e^{\frac{E - E_F}{kT}})} \cdot (N_C) dE$$

- ▶ where N_C is the effective density of states in the conduction band and is given by:

$$N_C \equiv 2 \left(\frac{2 \pi m_{de} kT}{h^2} \right)^{3/2} M_C.$$

1.4 Carrier Concentration at Thermal Equilibrium

$$n = \int_{E_C}^{\infty} \frac{[(E - E_C)/kT]^{1/2} dE}{1 + \exp[(E - E_F)/kT]}$$
$$n = \int_0^{\infty} \frac{\eta^{1/2} d\eta}{1 + \exp(\eta - \eta_F)}$$

- By changing variables at which $\eta = (E - E_C)/kT$, and $\eta_f = (E_f - E_C)/kT$

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) \equiv F_{1/2}(\eta_F) = \int_{E_C}^{\infty} \frac{[(E - E_C)/kT]^{1/2} dE}{1 + \exp[(E - E_F)/kT]}$$

Nondegenerate Semiconductors:

- ▶ By definition, the doping concentrations are smaller than N_c and the Fermi levels are more than several kT below E_c , (negative η_F),
- ▶ Boltzmann's approximation.

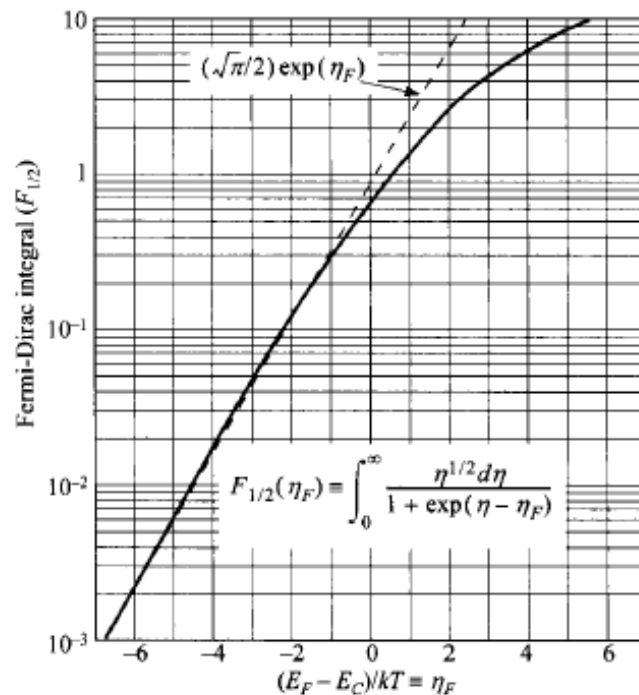


Fig. 8 Fermi-Dirac integral $F_{1/2}$ as a function of Fermi energy. (After Ref. 27.) Dashed line is approximation of Boltzmann statistics.

Nondegenerate Semiconductors:

$$F_{1/2}\left(\frac{E_F - E_C}{kT}\right) = \frac{\sqrt{\pi}}{2} \exp\left(-\frac{E_C - E_F}{kT}\right)$$

n-type semiconductors



$$n = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad \text{or} \quad E_C - E_F = kT \ln\left(\frac{N_C}{n}\right).$$

p-type semiconductors



$$p = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \quad \text{or} \quad E_F - E_V = kT \ln\left(\frac{N_V}{p}\right),$$

Degenerate Semiconductors

- ▶ This case represents energy levels where n or p are near or beyond the effective density of states (N_C or N_V) and the approximation cannot be used.
- ▶ For $\eta_f > -1$ we cannot use Boltzmann's approximation.
- ▶ The integral has weaker dependence on the carrier concentration.
- ▶ Also the Fermi levels are outside the energy gap.
- ▶ Fermi level as a function of carrier concentration is given by,
- ▶ for n-type semiconductor:

$$E_F - E_C \approx kT \left[\ln \left(\frac{n}{N_C} \right) + 2^{-3/2} \left(\frac{n}{N_C} \right) \right],$$

- ▶ for P-type semiconductor:

$$E_V - E_F \approx kT \left[\ln \left(\frac{p}{N_V} \right) + 2^{-3/2} \left(\frac{p}{N_V} \right) \right].$$

Intrinsic Concentration

- ▶ For intrinsic semiconductors at finite temperatures, thermal agitation occurs causing continuous excitation of electrons from the valence band to the conduction band so an equal holes is left in the valence band
- ▶ This process is balanced by recombination of the electrons in the conduction band with holes in the valence band.
- ▶ Net of carriers $n = p = n_i$, where n_i is the intrinsic carrier density.
- ▶ To define fermi energy level:

$$E_F = E_i = \frac{E_C + E_V}{2} + \frac{kT}{2} \ln\left(\frac{N_V}{N_C}\right) \rightarrow \text{To compensate for thermal agitation}$$
$$= \frac{E_C + E_V}{2} + \frac{3kT}{4} \ln\left(\frac{m_{dh}}{m_{de} M_C^{2/3}}\right).$$

Intrinsic Concentration

- ▶ Fermi level $E_f = E_i$ of an intrinsic semiconductor generally lies very close to $E_g/2$, but not exactly at, the middle of the bandgap.
- ▶ The intrinsic carrier density n_i can be obtained by:

$$n_i = N_C \exp\left(-\frac{E_C - E_i}{kT}\right) = N_V \exp\left(-\frac{E_i - E_V}{kT}\right) = \sqrt{N_C N_V} \exp\left(-\frac{E_g}{2kT}\right)$$

- ▶ the mass-action law:

$$\begin{aligned} pn &= N_C N_V \exp\left(-\frac{E_g}{kT}\right) \\ &= n_i^2, \end{aligned}$$

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad \text{or} \quad E_F - E_i = kT \ln\left(\frac{n}{n_i}\right),$$

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad \text{or} \quad E_i - E_F = kT \ln\left(\frac{p}{n_i}\right)$$

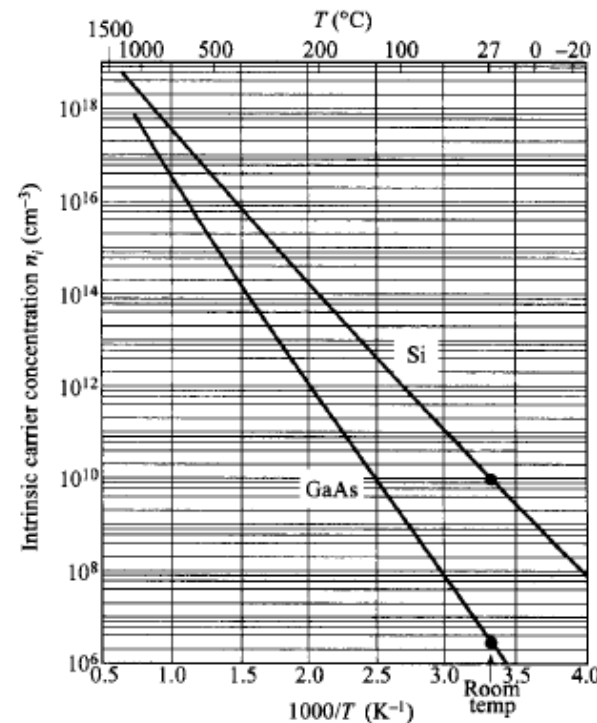
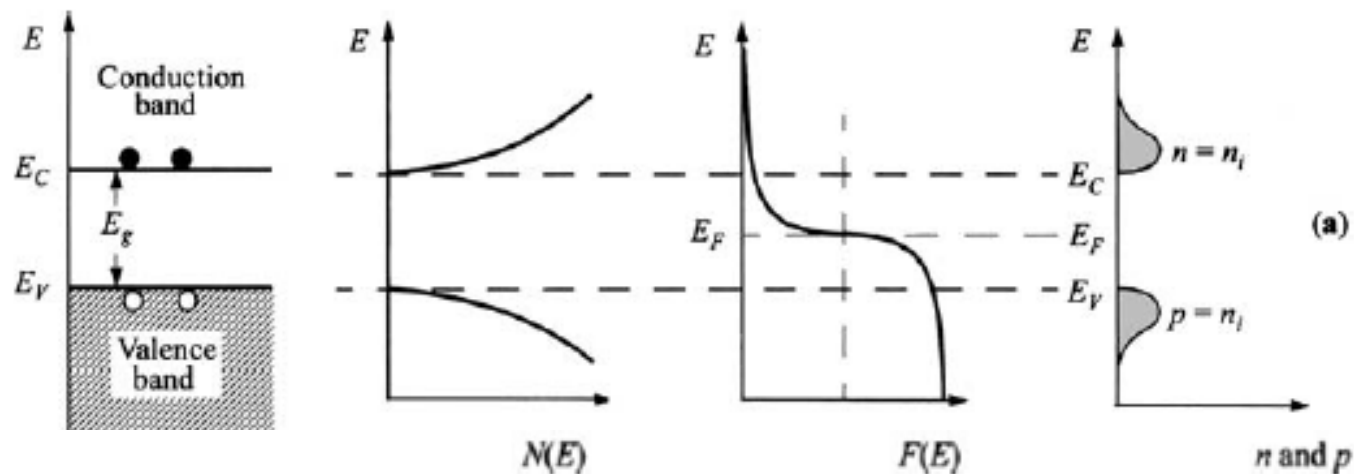


Fig. 9 Intrinsic carrier concentrations of Si and GaAs as a function of reciprocal temperature. (After Refs. 22 and 29.)

1.4.3 Calculation of Fermi Level

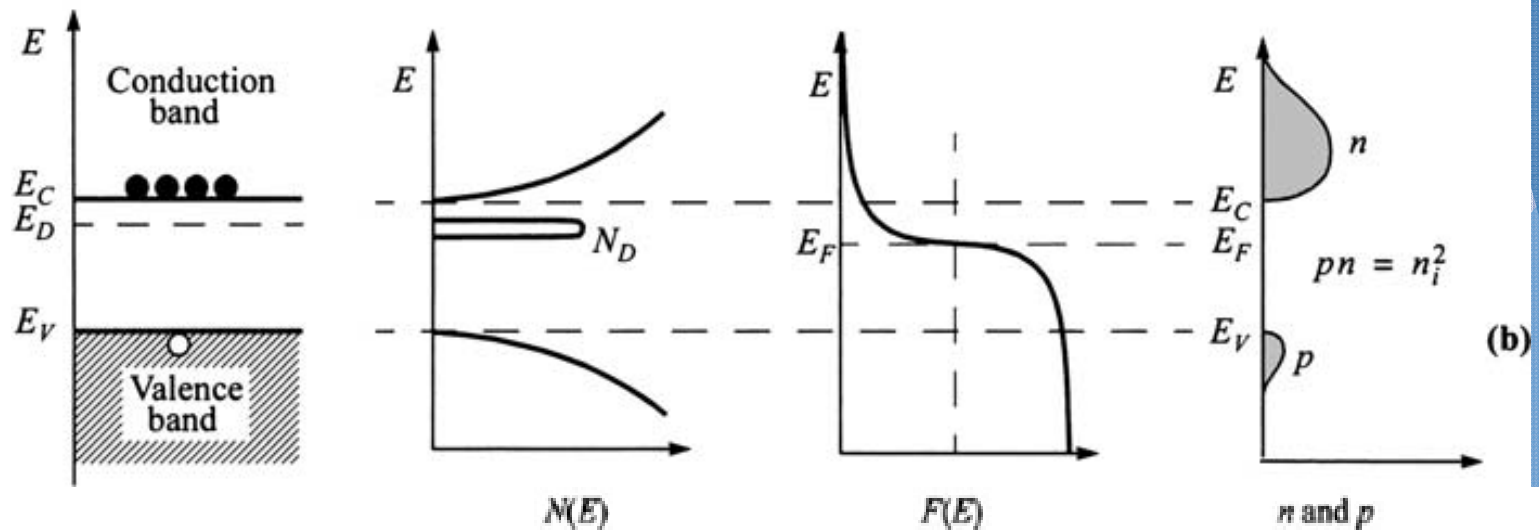
- ▶ The Fermi level for the intrinsic semiconductor lies very close to the middle of the bandgap. $E_f = E_g/2$



- ▶ When impurities are introduced to the semiconductor crystals, The ionized concentration for donors:

$$N_D^+ = \frac{N_D}{1 + g_D \exp[(E_F - E_D)/kT]}$$

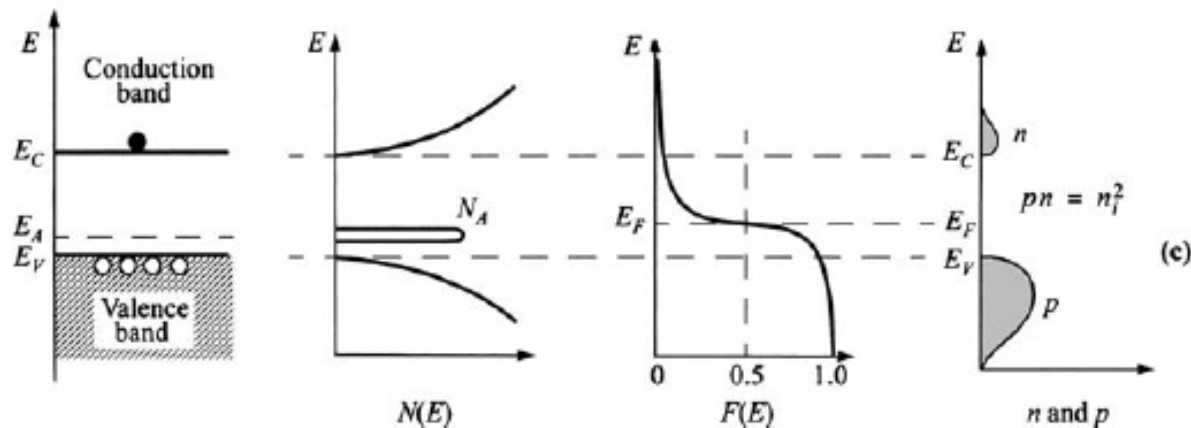
- ▶ where g_D , is the ground-state degeneracy of the donor impurity level and equal to 2 because a donor level can accept one electron with either spin or can have no electron.



- ▶ When acceptor impurities of concentration N_A are added to a semiconductor crystal, a similar expression can be written for the ionized acceptors

$$N_A^- = \frac{N_A}{1 + g_A \exp[(E_A - E_F)/kT]}$$

- ▶ where the ground-state degeneracy factor g_A is 4 because in most semiconductors each acceptor impurity level can accept one hole of either spin and the impurity level is doubly degenerate as a result of the two degenerate valence bands at $k = 0$.



- ▶ When impurity atoms are introduced, the total negative charges (electrons and ionized acceptors) must equal the total positive charges (holes and ionized donors), represented by the charge neutrality.

$$n + N_A^- = p + N_D^+.$$

- ▶ pn product is always independent of the added impurities.
- ▶ where donor impurities with the concentration added to the crystal. The charge neutrality condition becomes, neglecting the concentration of holes:

$$\begin{aligned} n &= N_D^+ + p \\ &\approx N_D^+ \end{aligned}.$$

- ▶ So by substituting in the n and N_D equations

$$N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \approx \frac{N_D}{1 + 2 \exp[(E_F - E_D)/kT]}.$$

- ▶ it can be shown that for $N_D \gg 0.5 \times N_C \exp[-(E_C - E_D)/kT] \gg N_A$, the electron concentration can be approximated by:

$$n \approx \sqrt{\frac{N_D N_C}{2}} \exp\left[-\frac{(E_C - E_D)}{2kT}\right].$$

$$N_A \gg \frac{1}{2} N_C \exp[-(E_C - E_D)/kT].$$

- ▶ the approximate expression for the electron density is then:

$$n \approx \left(\frac{N_D - N_A}{2N_A}\right) N_C \exp\left[-\frac{(E_C - E_D)}{kT}\right].$$

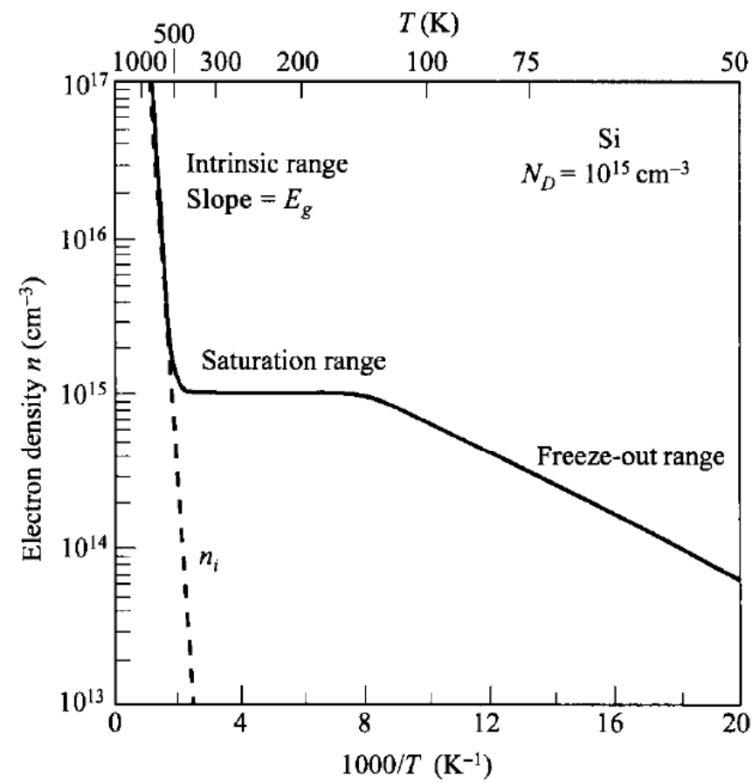


Fig. 13 Electron density as a function of temperature for a Si sample with donor impurity concentration of 10^{15} cm^{-3} . (After Ref. 5.)

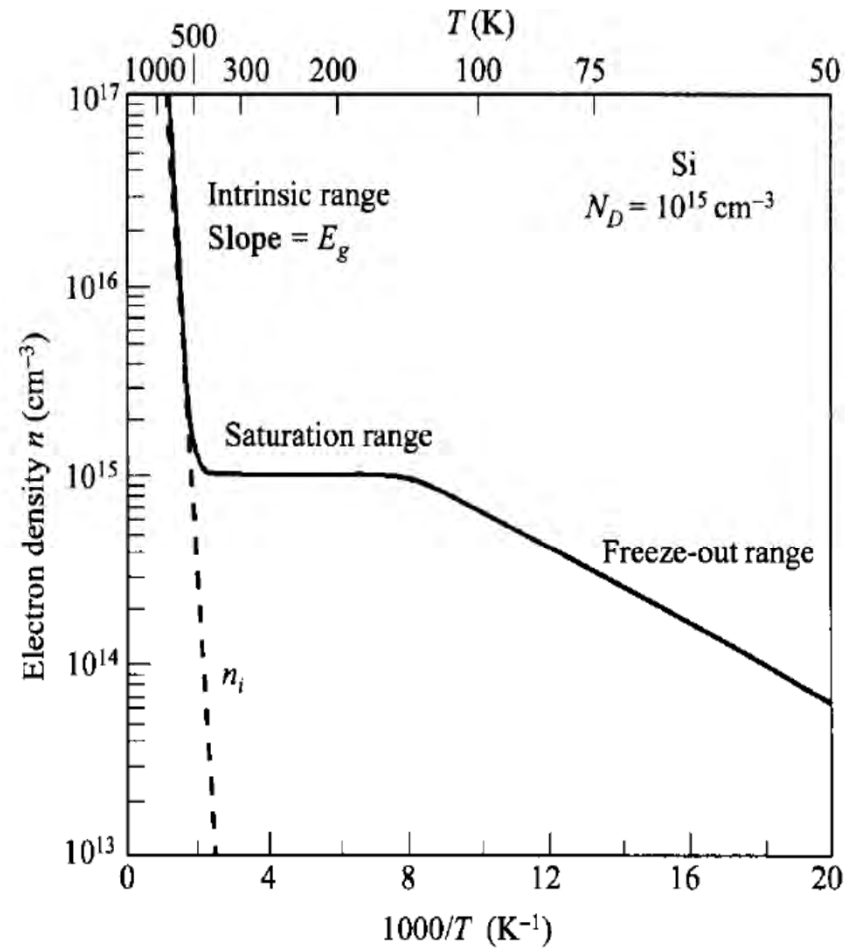


Fig. 13 Electron density as a function of temperature for a Si sample with donor impurity concentration of 10^{15} cm^{-3} . (After Ref. 5.)

- At high temperatures we have the intrinsic range since

$$n \approx p \approx n_i \gg N_D.$$

- At relatively high temperatures, most donors and acceptors are ionized, so the neutrality condition can be approximated by :

$$n + N_A = p + N_D.$$

- In an n-type semiconductor where $N_D > N_A$

$$n_{no} = \frac{1}{2}[(N_D - N_A) + \sqrt{(N_D - N_A)^2 + 4n_i^2}]$$

$\approx N_D$ if $|N_D - N_A| \gg n_i$ or $N_D \gg N_A$,

$$p_{no} = \frac{n_i^2}{n_{no}} \approx \frac{n_i^2}{N_D}.$$

- ▶ The Fermi level can be obtained from:

$$n_{no} = N_D = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) = n_i \exp\left(\frac{E_F - E_i}{kT}\right).$$

- ▶ For p-type:

$$p_{po} = \frac{1}{2}[(N_A - N_D) + \sqrt{(N_A - N_D)^2 + 4n_i^2}]$$

if $|N_A - N_D| \gg n_i$ or $N_A \gg N_D$,

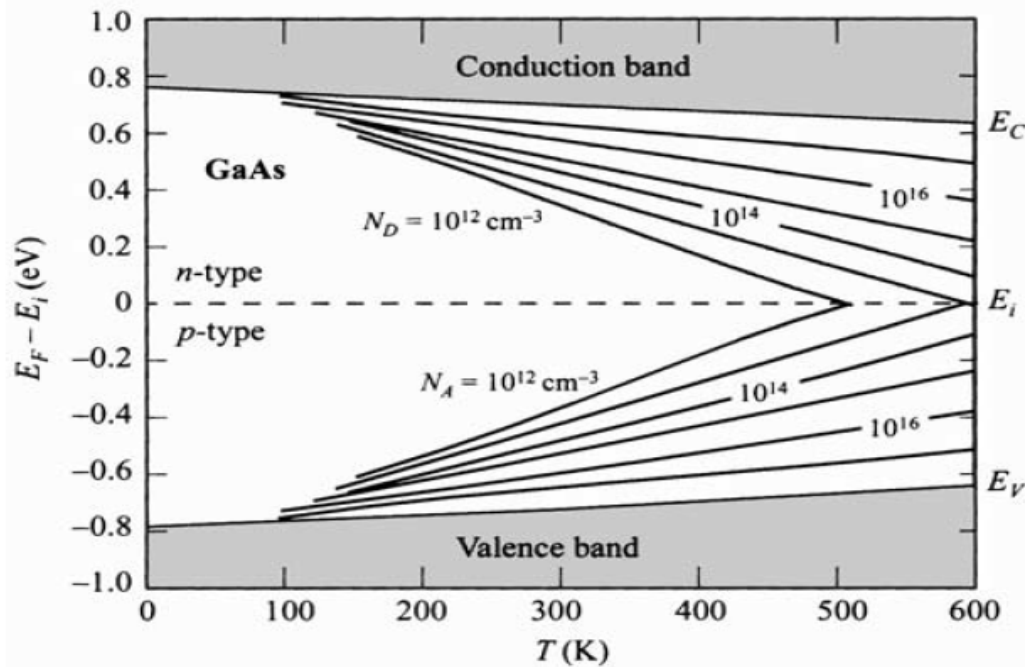
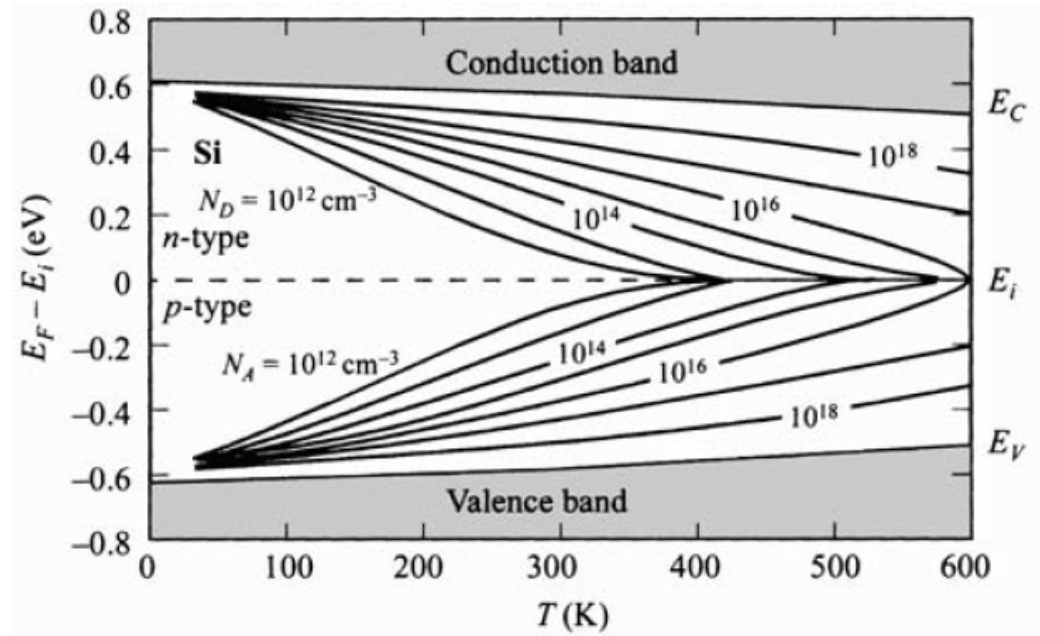
$$\approx N_A$$

$$n_{po} = \frac{n_i^2}{p_{po}} \approx \frac{n_i^2}{N_A},$$

$$p_{po} = N_A = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) = n_i \exp\left(\frac{E_i - E_F}{kT}\right).$$

- ▶ subscript "0" refers to the thermal equilibrium condition.
- ▶ For n-type semiconductors the electron is referred to as the majority carrier and the hole as the minority carrier.

► Which material can be used in high temperature Application?



1.5 Carrier-Transport Phenomena

▶ 1.5.1 Drift and Mobility:

- ▶ At low electric fields, the drift velocity (\mathcal{V}_d) is proportional to the electric field strength (\mathcal{E})

$$v_d = \mu \mathcal{E}.$$

- ▶ The mobility is the proportionality constant (μ)
- ▶ Intervalley scattering in which an electron is scattered from the vicinity of one minimum to another minimum and an energetic phonon (optical phonon) is involved.
- ▶ Qualitatively, since mobility is controlled by scattering, it can also be related to the mean free time τ_m or mean free path λ_m :

$$\mu = \frac{q \tau_m}{m^*} = \frac{q \lambda_m}{\sqrt{3kTm^*}}.$$

$$\lambda_m = v_{th} \tau_m$$

- ▶ where V_{th} is the thermal velocity given by $v_{th} = \sqrt{\frac{3kT}{m^*}}$.

- ▶ Actually the mobility inside material depends on the μ_i ionized impurities and μ_l of material.

$$\mu_l = \frac{\sqrt{8\pi} q \hbar^4 C_l}{3 E_{ds}^2 m_c^{*5/2} (kT)^{3/2}} \propto \frac{1}{m_c^{*5/2} T^{3/2}} \quad m_c^* = \text{Conduction effective mass}$$

- ▶ As the impurity concentration increases (at room temperature most shallow impurities are ionized) the mobility decreases.
- ▶ The mobility from ionized impurities μ_i can be described by:

$$\mu_i = \frac{64 \sqrt{\pi} \epsilon_s^2 (2kT)^{3/2}}{N_I q^3 m^{*1/2}} \left\{ \ln \left[1 + \left(\frac{12 \pi \epsilon_s kT}{q^2 N_I^{1/3}} \right)^2 \right] \right\}^{-1} \propto \frac{T^{3/2}}{N_I m^{*1/2}}$$

- ▶ where N_I is the ionized impurity density.
- ▶ The combined mobility, which includes the two μ_l and μ_i is given by the Matthiessen rule:

$$\mu = \left(\frac{1}{\mu_l} + \frac{1}{\mu_i} \right)^{-1}.$$

The relation between mobility and impurity concentration

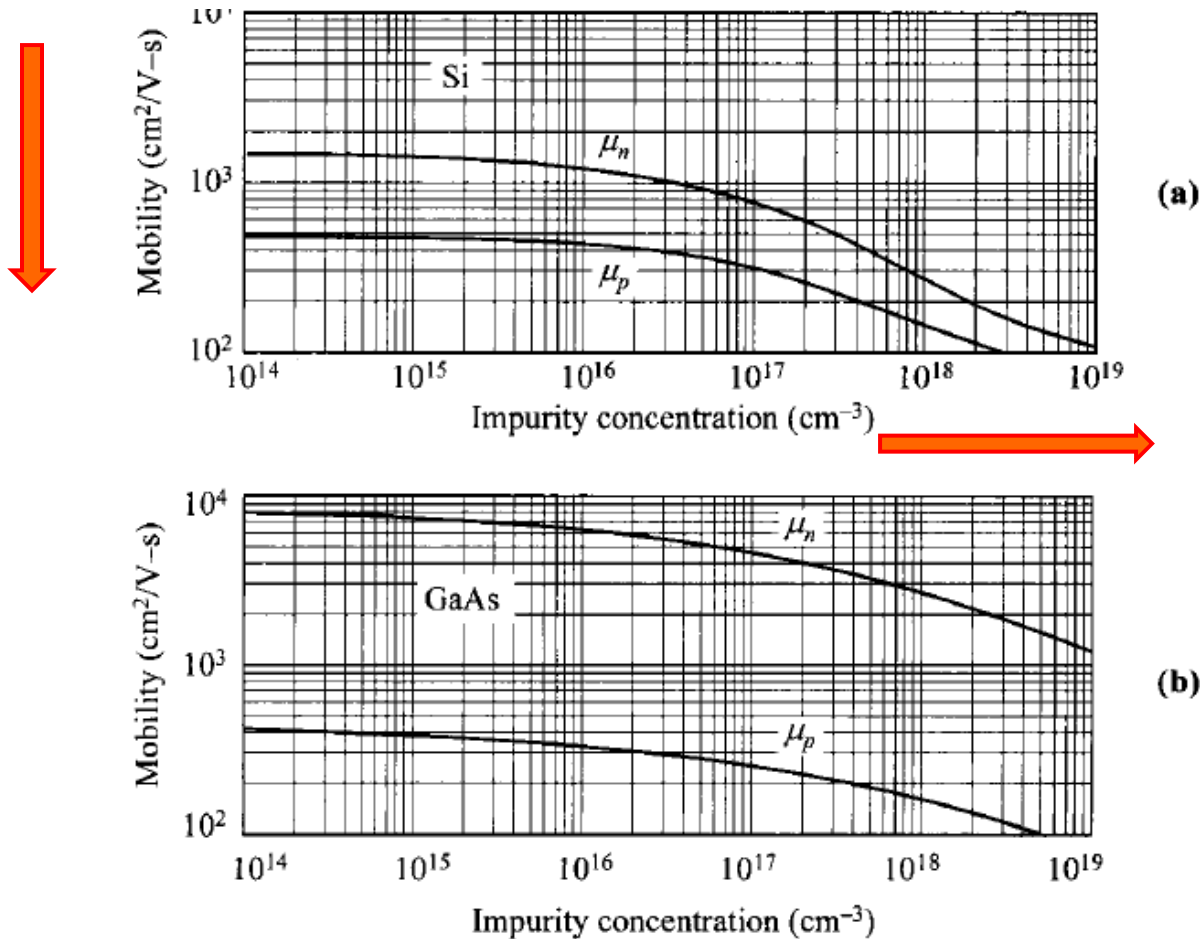


Fig. 15 Drift mobility of (a) Si (After Ref. 40.) and (b) GaAs at 300 K vs. impurity concentration (after Ref. 11).

Mobility of electrons and holes in Si as a function of temperature

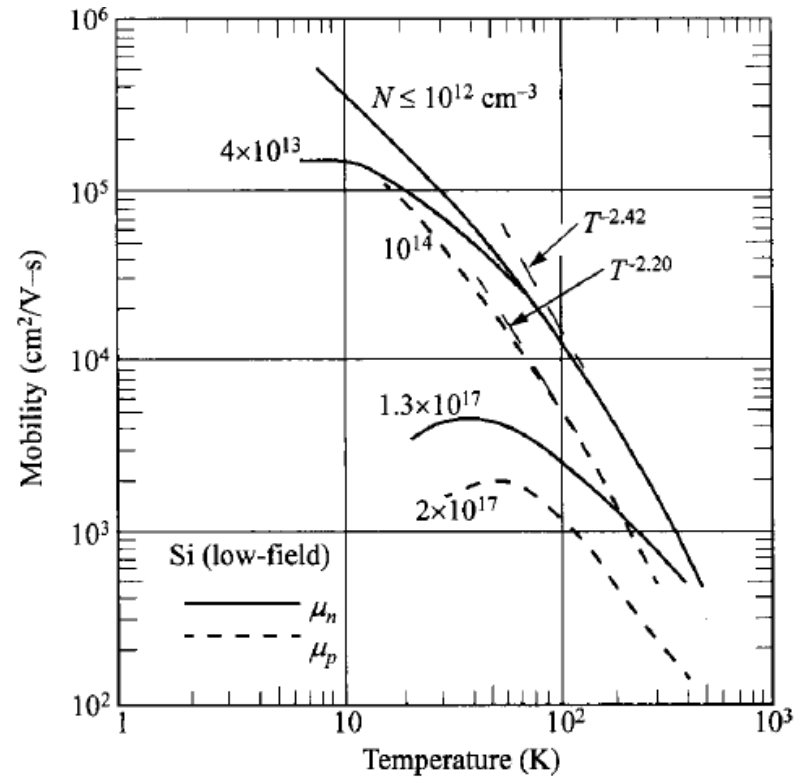


Fig. 16 Mobility of electrons and holes in Si as a function of temperature. (After Ref. 41.)

1.5.2 Resistivity and Hall Effect

- ▶ For semiconductors with both electrons and holes as carriers, the drift current under an applied field :

$$\begin{aligned} J &= \sigma \mathcal{E} \\ &= q(\mu_n n + \mu_p p) \mathcal{E} \end{aligned}$$

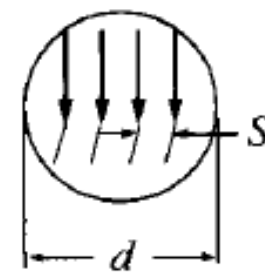
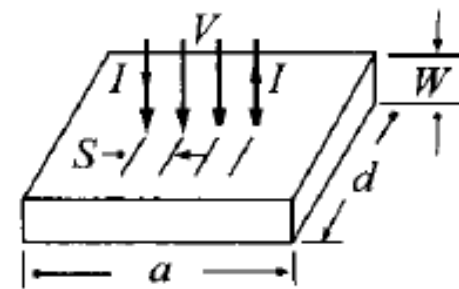
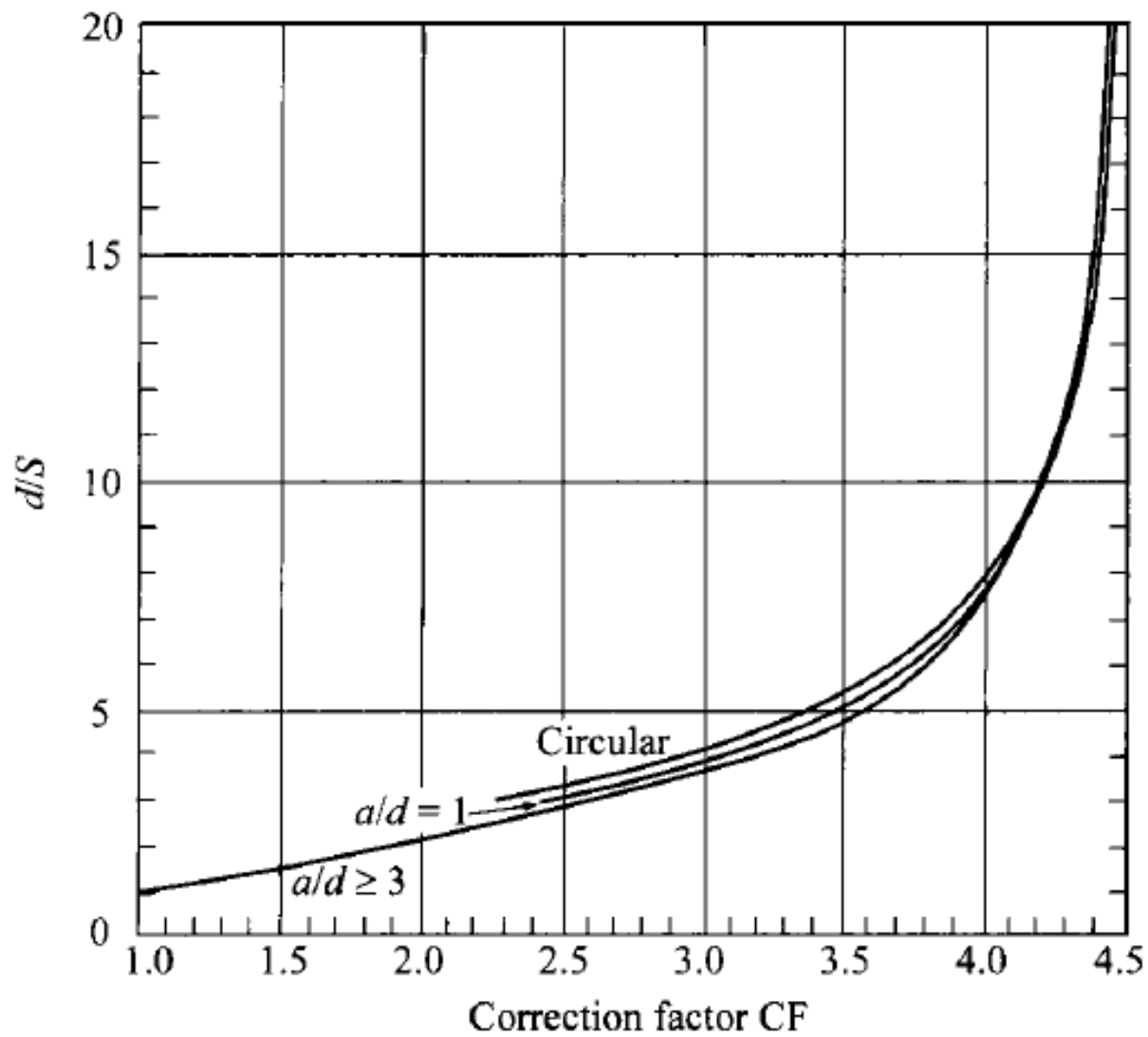
- ▶ where σ is the conductivity:

$$\sigma = \frac{1}{\rho} = q(\mu_n n + \mu_p p)$$

- ▶ ρ is the resistivity. If $n \gg p$, as in n-type semiconductors

$$\begin{aligned} \sigma &= q\mu_n n. \\ \rho &= \frac{1}{q\mu_n n} \end{aligned}$$

- ▶ The most-common method for measuring resistivity is the four-point probe method



$$R_s = \frac{V}{I} CF \quad (\Omega/\square)$$

$$\rho = R_s W \quad (\Omega\text{-cm})$$

- ▶ A small constant current is passed through the outer two probes and the voltage is measured between the inner two probes.
- ▶ For a thin wafer with thickness W much smaller than either a or d , the sheet resistance R_{\square} :

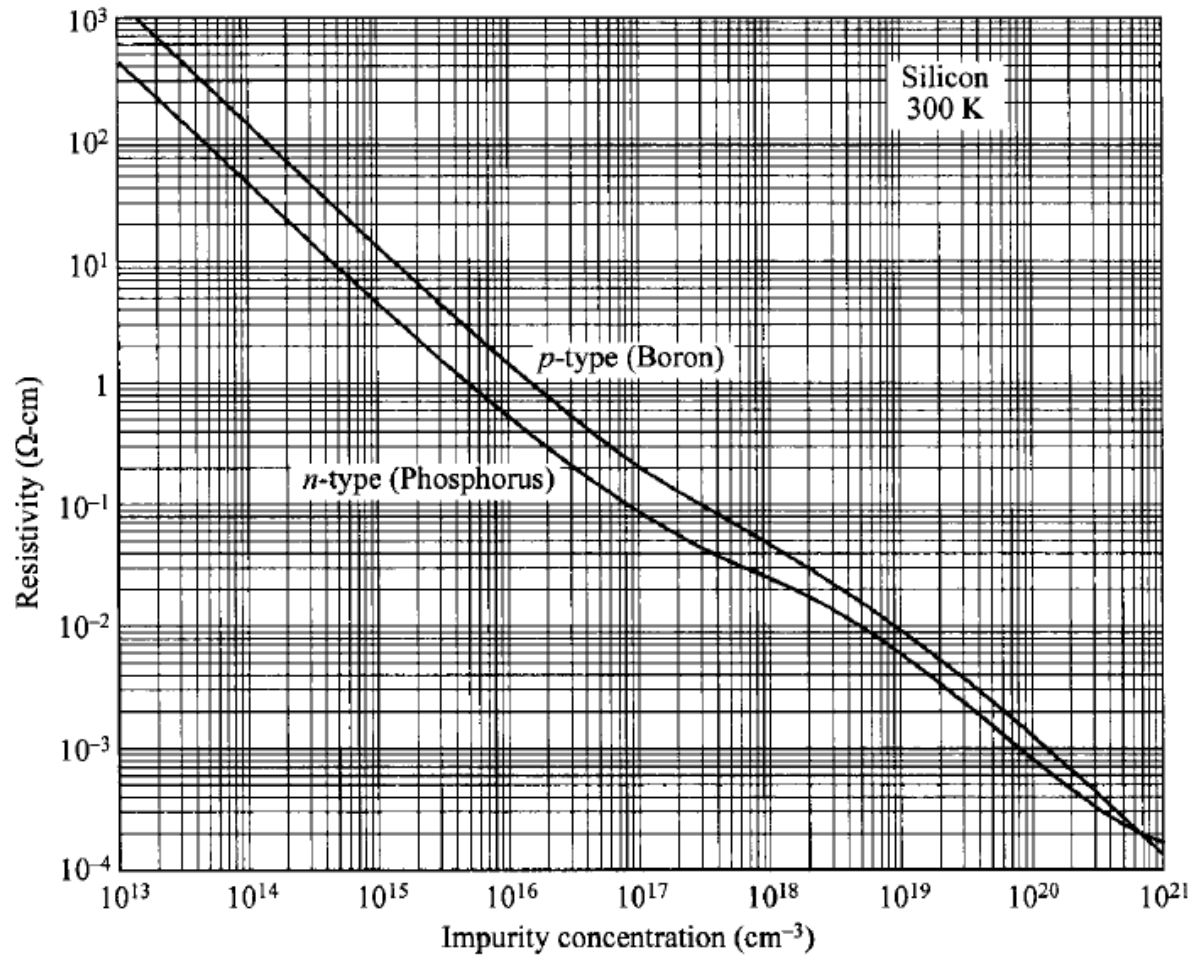
$$R_{\square} = \frac{V}{I} \cdot \text{C.F.} \quad \Omega/\square$$

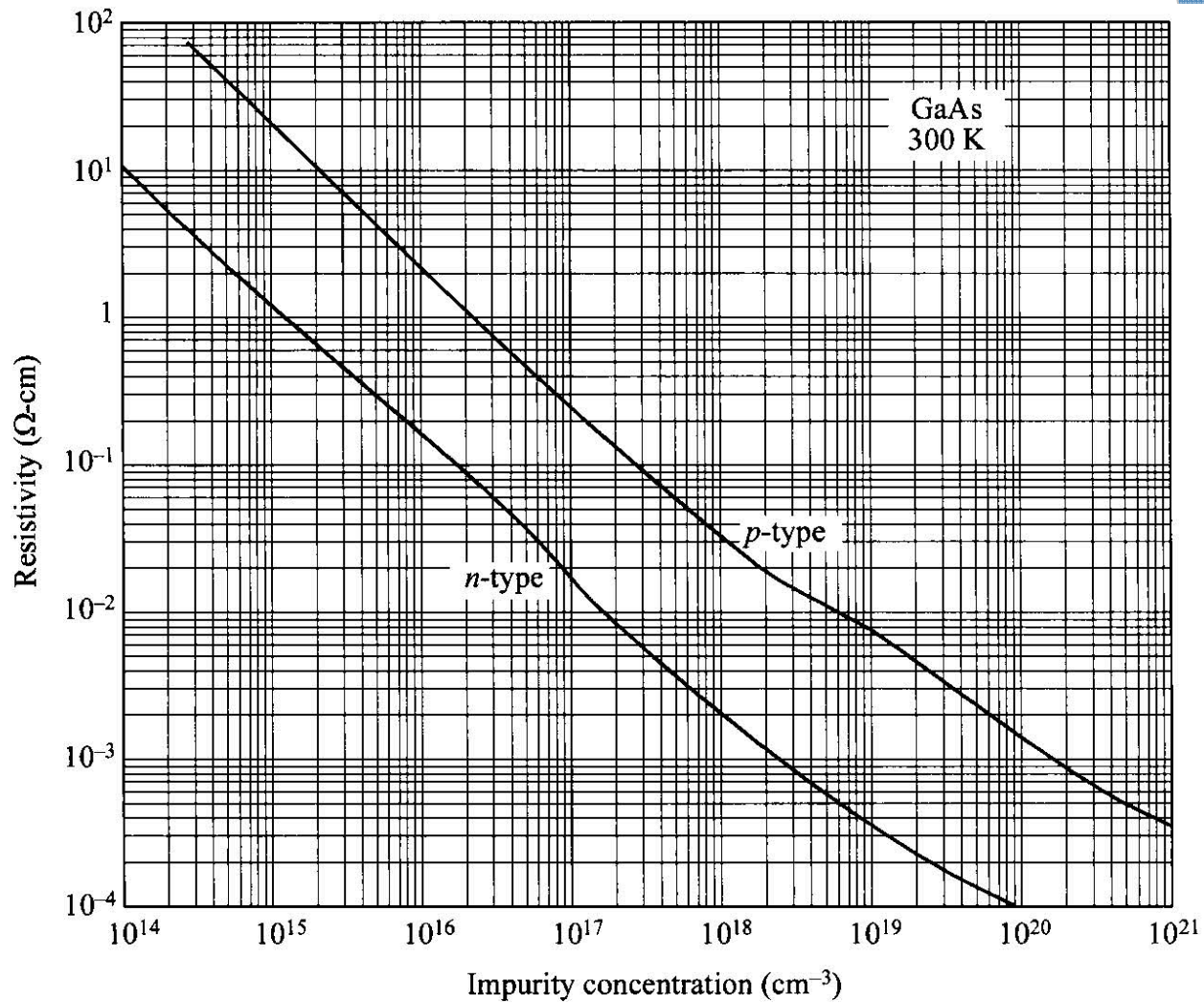
- ▶ Where C.F is the correction factor
- ▶ Resistivity is function of sheet resistance \times thickness

$$\rho = R_{\square} W \quad \Omega\text{-cm.}$$

- ▶ *Under a condition that $d \gg S$, where S is the probe spacing, the correction factor becomes $\frac{\pi}{\ln 2} = 4.54$.*

Relation between resistivity and doping concentration





(b)

Fig. 18 Resistivity vs. impurity concentration at 300 K for (a) silicon (after Ref. 40) and (b) GaAs (after Ref. 35).

Hall Effect

- ▶ Measurement of the resistivity only gives the product of the mobility and carrier concentration.
- ▶ Hall effect is to measure the mobility and carrier concentration parameters directly.
- ▶ The electric field is applied along the x-axis and a magnetic field is applied along the z-axis
- ▶ Consider a p-type sample
- ▶ The Lorentz force exerts an average downward force on the holes

$$\text{Lorentz force} = qv_x \times \mathcal{B}_z,$$

- ▶ The downward-directed current causes a piling up of holes at the bottom side of the sample, which in turn gives rise to an electric field E_y .
- ▶ Originally there is no net current along the y-direction in the steady state, the electric field along the y-axis (Hall field)

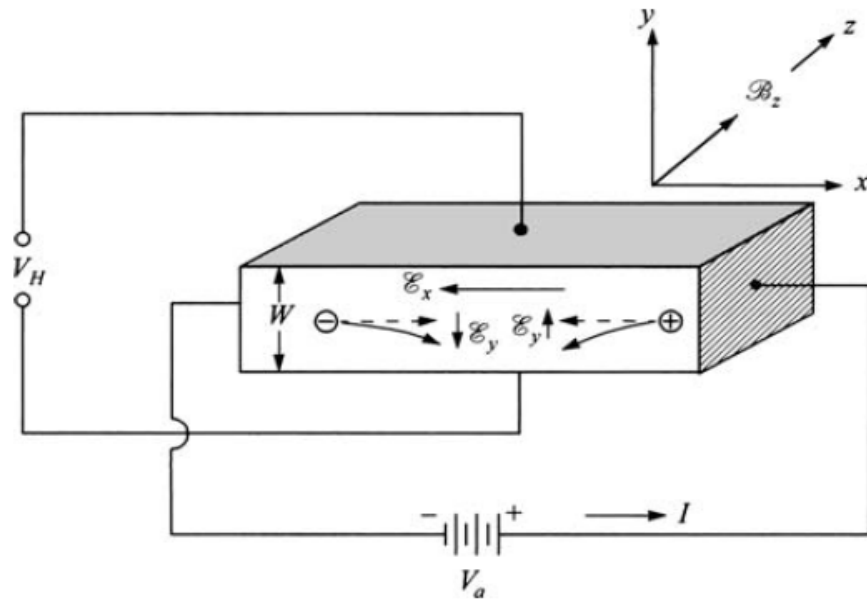


Fig. 19 Basic setup to measure carrier concentration using the Hall effect.

- ▶ the carriers travel in a path parallel to the applied field E_x .
- ▶ The carrier velocity v is related to the current density by:

$$J_x = qv_x p.$$

- ▶ Since for each carrier the Lorentz force must be equal to the force exerted by the Hall field

- ▶ Hall voltage: $V_H = \mathcal{E}_y W = \frac{J_x \mathcal{B}_z W}{qp}$ $q\mathcal{E}_y = qv_x \mathcal{B}_z,$

- ▶ R_H is Hall coefficient $V_H = R_H J_x \mathcal{B}_z W$

- ▶ The carrier concentration and carrier type (electrons or holes from the polarity of the Hall voltage) can be obtained directly from the Hall measurement, provided that one type of carrier dominates
- ▶ the sign of R_H , and thus V_H , reveals the majority type of the semiconductor sample.

$$R_H = \frac{r_H}{qp} \quad p \gg n,$$

$$R_H = -\frac{r_H}{qn} \quad n \gg p,$$

$$R_H = \frac{r_H}{q} \frac{\mu_p^2 p - \mu_n^2 n}{(\mu_p p + \mu_n n)^2}.$$

$$r_H \equiv \frac{\langle \tau_m^2 \rangle}{\langle \tau_m \rangle^2}.$$

- The parameter τ_m for the Hall factor is the mean free time between carrier collisions, which depends on the carrier energy.

$$\tau_m = C_1 E^{-s},$$

- ▶ The Hall mobility μ_H is defined as the product of the Hall coefficient and conductivity:

$$\mu_H = |R_H| \sigma.$$

- ▶ The Hall mobility, the Hall factor r_H and the carriers mobility are related by:

$$\mu_H = r_H \mu.$$

1.5.3 High-Field Properties

- ▶ When the fields are sufficiently large, however, nonlinearities in mobility and, in some cases, saturation of drift velocity are observed.
- ▶ At still larger fields, impact ionization occurs. First, we consider the nonlinear mobility.
- ▶ At thermal equilibrium the carriers both emit and absorb phonons and the net rate of exchange of energy is zero.
- ▶ In the presence of an electric field the carriers acquire energy from the field and lose it to phonons by emitting more phonons than are absorbed.
- ▶ As the field increases, the average energy of the carriers also increases and they acquire an effective temperature T_e , that is higher than the lattice temperature T

- ▶ Balancing the rate at which energy is transferred from the field to the carriers by an equal rate of energy loss to the lattice.

- ▶ To deduce the rate (Ge and Si):

$$\frac{T_e}{T} = \frac{1}{2} \left[1 + \sqrt{1 + \frac{3\pi(\mu_0 \mathcal{E})^2}{8c_s^2}} \right] \quad v_d = \mu_0 \mathcal{E} \sqrt{\frac{T}{T_e}}$$

- ▶ where μ_0 is the low-field mobility, and c_s the velocity of sound.
- ▶ For moderate field strength when $\mu_0 \mathbf{E}$ is comparable to c_s , the carrier velocity v_d starts to deviate from being linearly dependent on the applied field, by a factor of $\sqrt{T/T_e}$.
- ▶ The carriers start to interact with optical phonons and T/T_e no longer accurate.

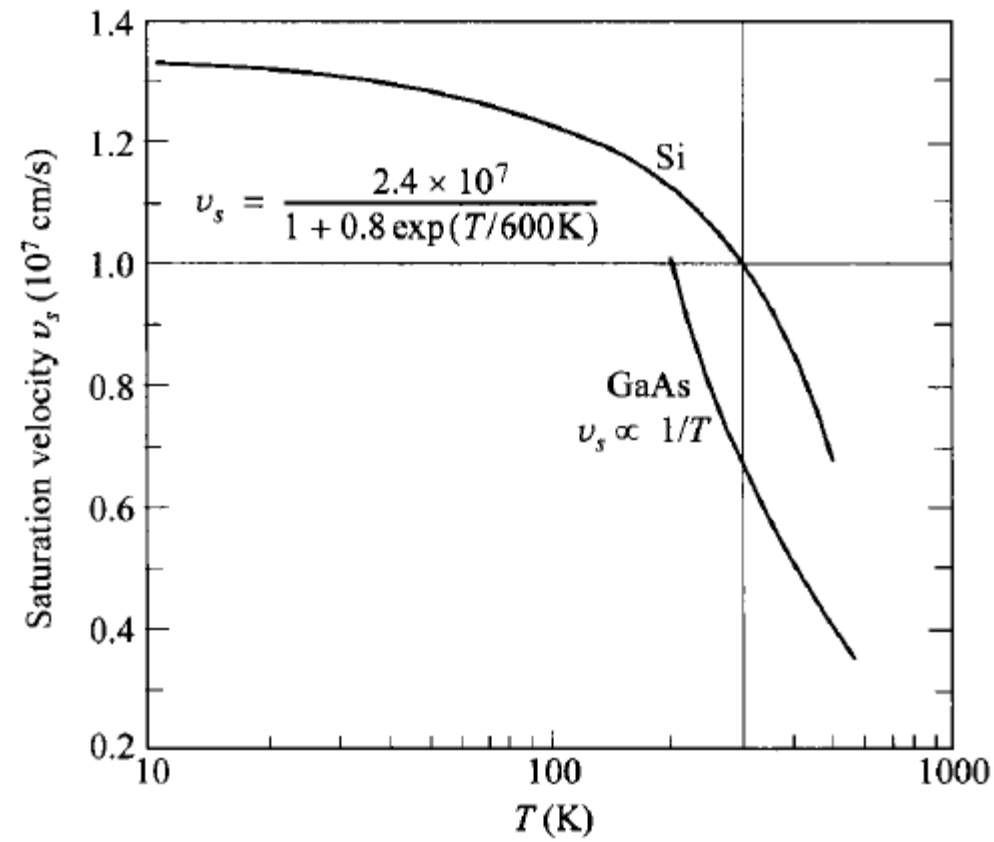
$$v_s = \sqrt{\frac{8E_p}{3\pi m_0}}$$

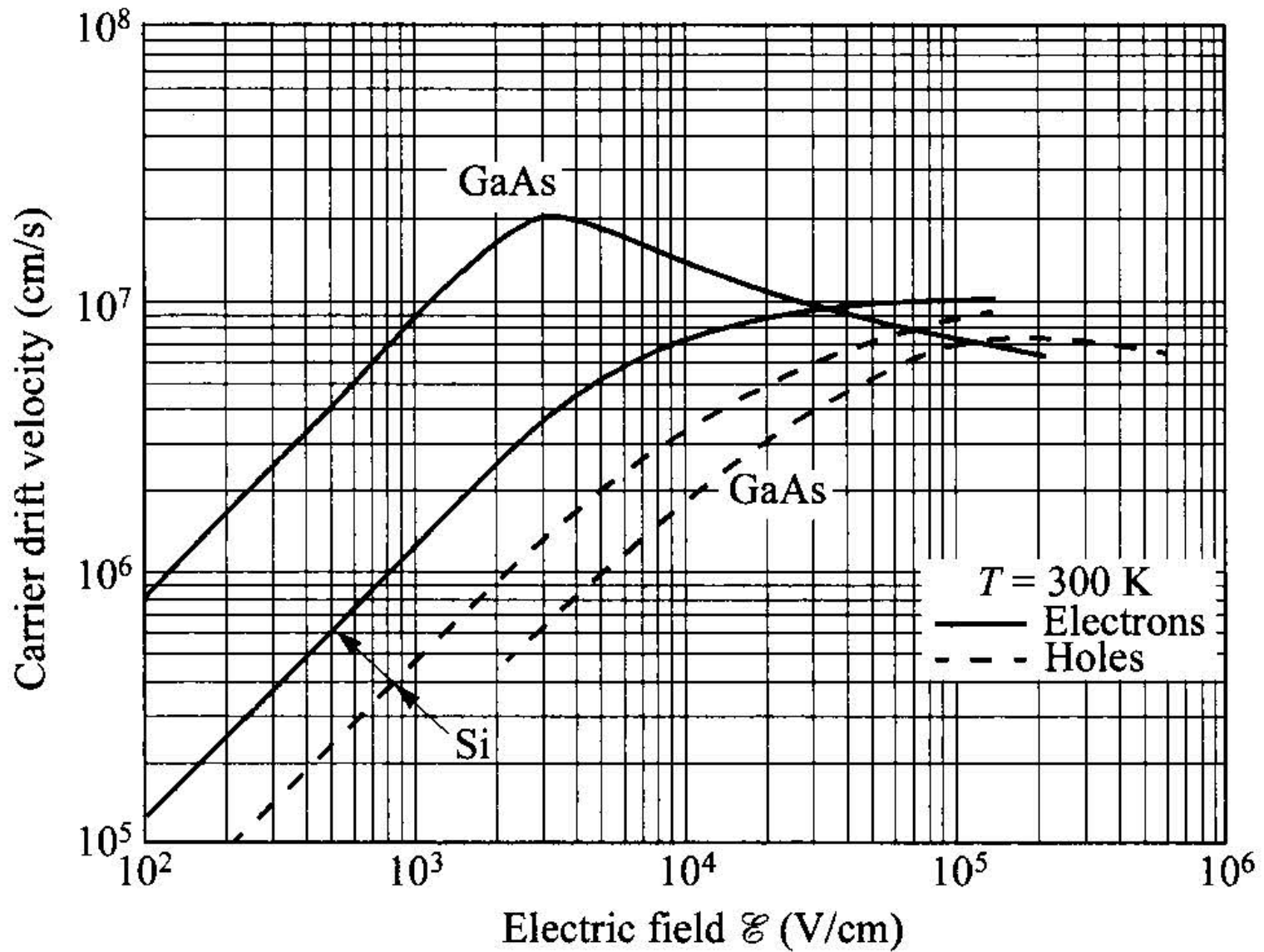
- ▶ where E_p is the optical-phonon energy “ depends on the material “

- ▶ Then the drift velocity describing the whole range, from low-field drift velocity to velocity saturation is expressed as:

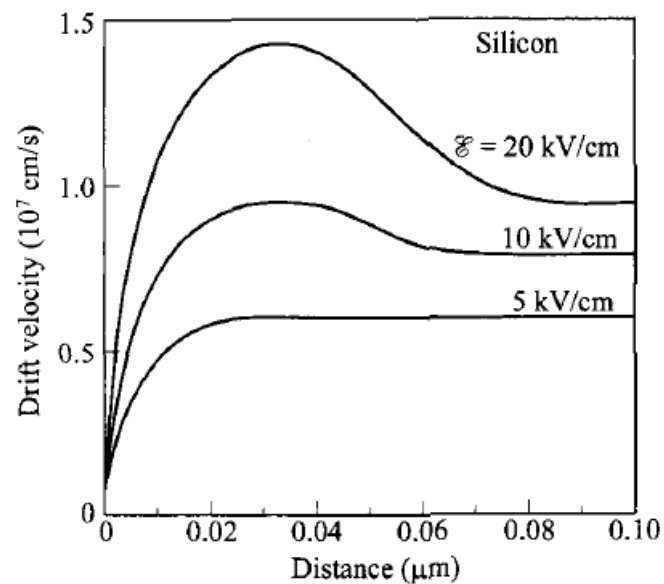
$$v_d = \frac{\mu_0 \mathcal{E}}{[1 + (\mu_0 \mathcal{E}/v_s)^{C_2}]^{1/C_2}}$$

- ▶ C_2 is depending on the electrons (2) or holes (1) and it is a function of temperature.





- ▶ At critical dimension where carriers transit across becomes smaller than the mean free path, ballistic transport is said to occur before carriers start to be scattered.
- ▶ At high fields, drift velocity can attain a higher value momentarily than that at steady state, within a short space (of the order of mean free path) and time (of the order of mean free time).
- ▶ This phenomenon is called velocity overshoot.



- ▶ Considering both electrons and holes, the generation rate:

$$\begin{aligned}\frac{dn}{dt} = \frac{dp}{dt} &= \alpha_n n v_n + \alpha_p p v_p \\ &= \frac{\alpha_n J_n}{q} + \frac{\alpha_p J_p}{q} .\end{aligned}$$

- ▶ Conversely, at any given time, the carrier density or current varies with distance and can be shown to be:

$$\frac{dJ_n}{dx} = \alpha_n J_n + \alpha_p J_p,$$

$$\frac{dJ_p}{dx} = -\alpha_n J_n - \alpha_p J_p.$$

- ▶ The total current ($J_n + J_p$) remains constant over distance but ionization rate “ α ” is depending on electric field

$$\alpha(\mathcal{E}) = \frac{q\mathcal{E}}{E_I} \exp\left\{-\frac{\mathcal{E}_I}{\mathcal{E}[1 + (\mathcal{E}/\mathcal{E}_p)] + \mathcal{E}_T}\right\}$$

- ▶ where E_I is the high-field effective ionization threshold energy, and E_T, E_p are threshold fields for carriers to overcome the decelerating effects of thermal, optical phonon, and ionization scattering.
- ▶ For Si, the value of E_I is found to be 3.6 eV for electrons and 5.0 eV for holes.

$$\alpha(\mathcal{E}) = \frac{q\mathcal{E}}{E_I} \exp\left(-\frac{\mathcal{E}_I}{\mathcal{E}}\right), \quad \text{if } \mathcal{E}_p > \mathcal{E} > \mathcal{E}_T,$$

$$\alpha(\mathcal{E}) = \frac{q\mathcal{E}}{E_I} \exp\left(-\frac{\mathcal{E}_I \mathcal{E}_p}{\mathcal{E}^2}\right), \quad \text{if } \mathcal{E} > \mathcal{E}_p \text{ and } \mathcal{E} > \sqrt{\mathcal{E}_p \mathcal{E}_T}.$$

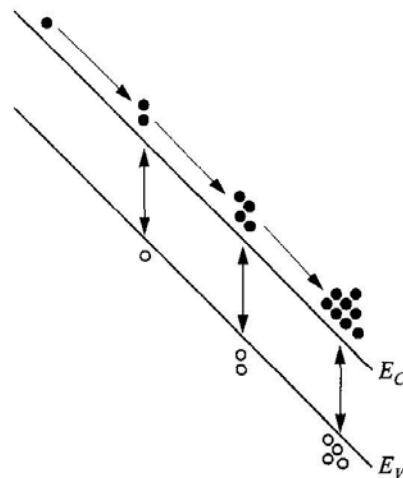


Fig. 22 Multiplication of electrons and holes from impact ionization, due to electrons (α_n) in this example ($\alpha_p = 0$).

1.5.4 Recombination, Generation, and Carrier Lifetimes

- ▶ Whenever the thermal-equilibrium condition of a semiconductor system is disturbed $pn \neq n_i^2$, processes exist to restore the system to equilibrium.
- ▶ These processes are recombination when $pn > n_i^2$ and thermal generation when $pn < n_i^2$
- ▶ The energy of an electron in transition from the conduction band to the valence band is conserved by emission of a photon (radiative process) or by transfer of the energy to another free electron or hole (Auger process)
- ▶ The radiative process is the inverse of direct optical absorption, and the Auger process is the inverse of impact ionization.

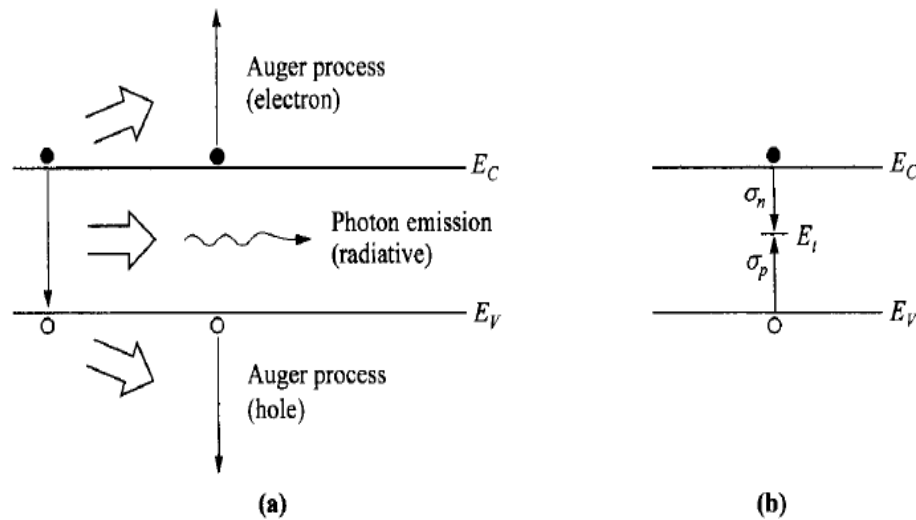
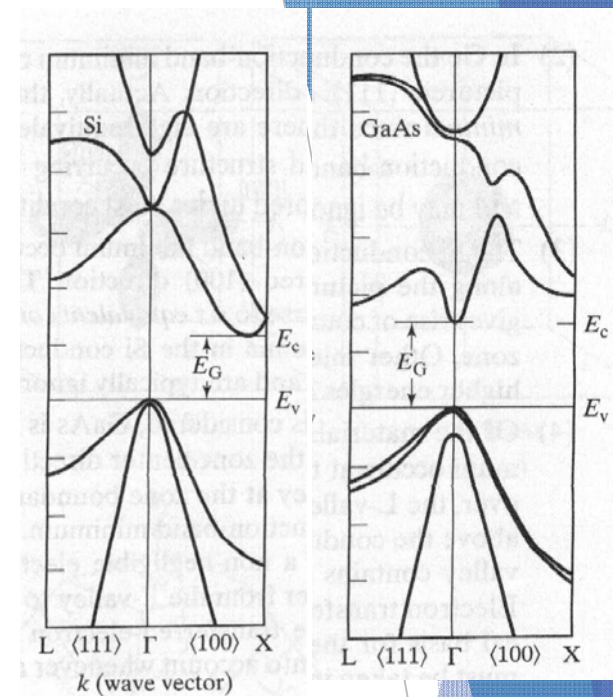


Fig. 25 Recombination processes (the reverse are generation processes). (a) Band-to-band recombination. Energy is exchanged to a radiative or Auger process. (b) Recombination through single-level traps (nonradiative).



- ▶ Band-to-band transitions are more probable for direct-bandgap semiconductors.

Note: Recombining e- must have a momentum value that matches the crystal momentum of the hole it is dropping into.

Direct bandgap = OK all the way to the valence band

- ▶ For this type of transition, the recombination rate is proportional to the product of electron and hole concentrations, given by:

$$R_e = R_{ec}pn.$$

- ▶ The term R_{ec} called the recombination coefficient, is related to the thermal generation rate G_{th} by

$$R_{ec} = \frac{G_{th}}{n_i^2}.$$

- ▶ R_{ec} is a function of temperature and is also dependent on the band structure of the semiconductor.
- ▶ In thermal equilibrium, since $pn = n_i^2$, $R_e = G_{th}$ and the net transition rate $U = R_e - G_{th}$ equals zero.

***n*-type material $p_n = p_{no} + \Delta p$ and $n_n \approx N_D$, the net transition rate is given by**

$$\begin{aligned} U &= R_e - G_{th} = R_{ec}(pn - n_i^2) \\ &\approx R_{ec}\Delta p N_D \equiv \frac{\Delta p}{\tau_p} \end{aligned}$$

- ▶ where the carrier lifetime for holes:

$$\tau_p = \frac{1}{R_{ec} N_D},$$

in p-type material $\tau_n = \frac{1}{R_{ec} N_A}.$

- ▶ In indirect-band gap semiconductors such as Si and Ge, the dominant transitions are indirect recombination and generation via bulk traps, of trap density N_t and energy E_{trap} present within the band gap.
- ▶ The single-level recombination can be described by two processes—electron capture and hole capture.
- ▶ The net transition rate can be described by the Shockley-Read-Hall statistics.

$$U = \frac{\sigma_n \sigma_p v_{th} N_t (pn - n_i^2)}{\sigma_n \left[n + n_i \exp\left(\frac{E_t - E_i}{kT}\right) \right] + \sigma_p \left[p + n_i \exp\left(\frac{E_i - E_t}{kT}\right) \right]}$$

- ▶ where σ_n and σ_p are the electron and hole capture cross sections, respectively.

- ▶ U is maximized when $E_t = E_i$, indicating for an energy spectrum of bulk traps, only those near the mid-gap are effective recombination and generation centers

$$U = \frac{\sigma_n \sigma_p v_{th} N_t (pn - n_i^2)}{\sigma_n (n + n_i) + \sigma_p (p + n_i)}$$

- ▶ for low-level injection in n-type semiconductors, the net recombination rate becomes:

$$U = \frac{\sigma_n \sigma_p v_{th} N_t [(p_{no} + \Delta p)n - n_i^2]}{\sigma_n n}$$

$$\approx \sigma_p v_{th} N_t \Delta p \equiv \frac{\Delta p}{\tau_p} \quad \tau_p = \frac{1}{\sigma_p v_{th} N_t}$$

- ▶ for p-type semiconductor, the electron lifetime is given by:

$$\tau_n = \frac{1}{\sigma_n v_{th} N_t}$$

- ▶ the lifetime arising from indirect transitions is inversely proportional to the trap density N_{trap}

1.5.5 Diffusion

- ▶ Here excess carriers are introduced locally, causing a condition of non-uniform carriers.
- ▶ Examples are local injection of carriers from a junction, and non-uniform illumination.
- ▶ The diffusion occurs by which the carriers migrate from the region of high concentration toward the region of low concentration.
- ▶ To drive the system toward a state of uniformity.
- ▶ This flow or flux of carriers (Fick's law):

$$\left. \frac{d\Delta n}{dt} \right|_x = -D_n \frac{d\Delta n}{dx},$$

- ▶ The proportionality constant is called the diffusion coefficient or diffusivity D_n . This flux of carriers constitutes a diffusion current, given by:

$$J_n = qD_n \frac{d\Delta n}{dx}, \quad J_p = -qD_p \frac{d\Delta p}{dx}.$$

- ▶ Diffusion is due to random thermal motion of carriers as well as scattering.

$$D = v_{th} \tau_m.$$

- ▶ The diffusion current is depending on mobility.
- ▶ we consider an n-type semiconductor with nonuniform doping concentration but without an external applied field.
- ▶ The zero net current necessitates that the drift current exactly balances the diffusion current.

$$qn\mu_n \mathcal{E} = -qD_n \frac{dn}{dx}.$$

- ▶ In this case, the electric field is created by the non-uniform doping

$$\mathcal{E} = dE_c / qdx.$$

$$qn\mu_n\mathcal{E} = -qD_n\frac{dn}{dx}.$$

$$\begin{aligned}\frac{dn}{dx} &= \frac{-q\mathcal{E}}{kT}N_C\exp\left(-\frac{E_C-E_F}{kT}\right) \\ &= \frac{-q\mathcal{E}}{kT}n.\end{aligned}$$

$$D_n = \left(\frac{kT}{q}\right)\mu_n.$$

$$D_p = \left(\frac{kT}{q}\right)\mu_p.$$

diffusion length:

$$L_d = \sqrt{D\tau}$$

The diffusion length can also be viewed as the distance carriers can diffuse in a carrier life time before they are wipe out

1.5.6 Thermionic Emission

- ▶ Another current conduction mechanism is thermionic emission.
- ▶ It is a majority carrier current and is always associated with a potential barrier.
- ▶ Note that the critical parameter is the barrier height, not the shape of the barrier.
- ▶ For thermionic emission to be the controlling mechanism, the criterion is that collision or the drift-diffusion process within the barrier layer is to be negligible.
- ▶ The barrier width has to be narrower than the mean free path. (classical mechanism of tunneling)

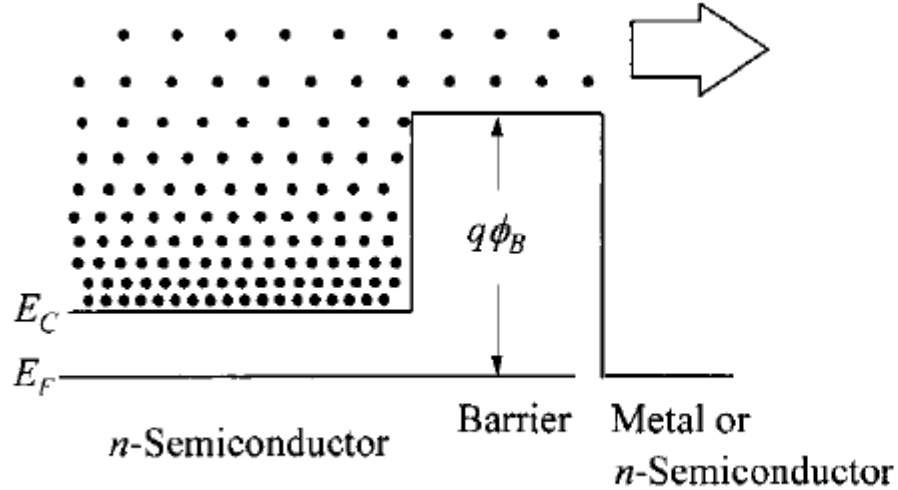


Fig. 26 Energy-band diagram showing thermionic emission of electrons over the barrier. Note that the shape of the barrier (shown as rectangular) does not matter.

- ▶ The region behind the barrier must be another n-type semiconductor or a metal layer to neglect the effect of diffusion current in that region.
- ▶ The density of electrons (for n-type substrate) decreases exponentially as a function of their energy above the conduction band edge.

- ▶ The target here is to define the integrated number of carriers above the barrier height.
- ▶ These carriers which are thermally generated carriers are no longer confined by the barrier so they contribute to the thermionic-emission current.
- ▶ The total electron current over the barrier is given by:

$$J = A^* T^2 \exp\left(-\frac{q\phi_B}{kT}\right).$$

- ▶ where ϕ_B is the barrier height, and

$$A^* \equiv \frac{4\pi q m^* k^2}{h^3}$$

- ▶ A^* is called the effective Richardson constant and is a function of the effective mass.

1.5.7 Tunneling

- ▶ Tunneling is a quantum-mechanical phenomenon
- ▶ In classical mechanics:
 - ▶ The carriers are completely confined by the potential walls
 - ▶ Only those carriers with excess energy higher than the barriers can escape.
 - ▶ As in the case of thermionic emission
- ▶ In quantum mechanics:
 - ▶ an electron can be represented by its wavefunction
 - ▶ The wave function does not terminate abruptly on a wall of finite potential height.
 - ▶ it can penetrate into and through the barrier
 - ▶ The probability of electron tunneling through a barrier of finite height and width is thus not zero.

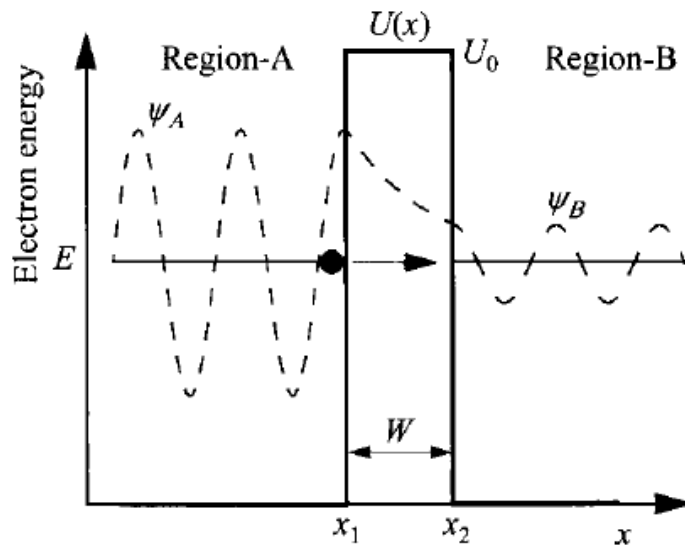


Fig. 27 Wavefunctions showing electron tunneling through a rectangular barrier.

- ▶ tunneling probability from the Schrodinger equation:

$$\frac{d^2 \psi}{dx^2} + \frac{2m^*}{\hbar^2} [E - U(x)] \psi = 0.$$

- ▶ In the case of a simple rectangular barrier of height U_0 and width W , ψ has a general form of $e^{\pm ikx}$

$$\text{where } k = \sqrt{2m^*(E - U_0)}/\hbar$$

- ▶ Note that for tunneling, the energy E is below the barrier U_0 , so that the term within the square root is negative and k is imaginary.

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} = \left[1 + \frac{U_0^2 \sinh^2(|k|W)}{4E(U_0 - E)} \right]^{-1}$$

$$\approx \frac{16E(U_0 - E)}{U_0^2} \exp\left(-2 \sqrt{\frac{2m^*(U_0 - E)}{\hbar^2}} W\right).$$

- ▶ For more complicated barrier shapes, simplification of the Schrodinger equation is made by the WKB (Wentzel-Kramers-Brillouin) approximation if the potential $U(x)$ does not vary rapidly.

$$T_t = \frac{|\psi_B|^2}{|\psi_A|^2} \approx \exp\left\{-2 \int_{x_1}^{x_2} |k(x)| dx\right\}$$

$$\approx \exp\left\{-2 \int_{x_1}^{x_2} \sqrt{\frac{2m^*}{\hbar^2} [U(x) - E]} dx\right\}.$$

- ▶ The tunneling current J_t can be calculated from the product of the number of available carriers in the originating Region-A and the number of empty states in the destination Region-B.

$$J_t = \frac{qm^*}{2\pi^2\hbar^3} \int F_A N_A T_t (1 - F_B) N_B dE$$

- ▶ where F_A , F_B , N_A , and N_B represent the Fermi-Dirac distributions and densities of states in the corresponding regions.

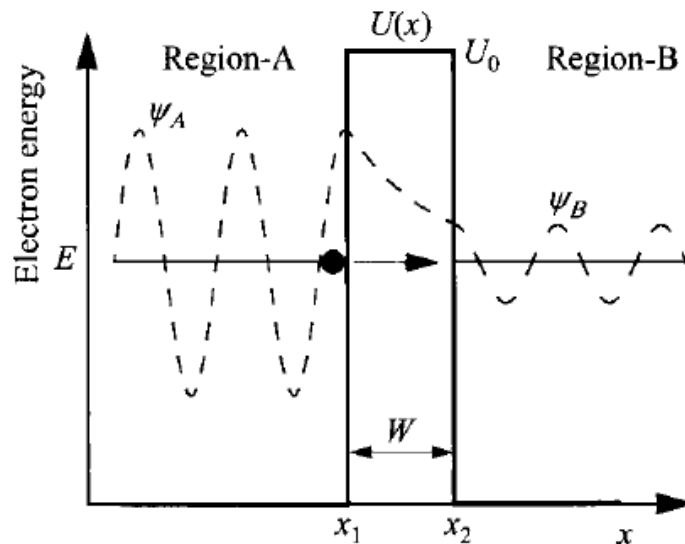


Fig. 27 Wavefunctions showing electron tunneling through a rectangular barrier.

1.5.8 Space-Charge Effect

- ▶ The space charge in a semiconductor is determined by both the doping concentration and the free-carrier concentration.

$$\rho = (p - n + N_D - N_A)q$$

- ▶ In the neutral region of a semiconductor, $n = N_D$ and $p = N_A$, so that the space-charge density is zero.
- ▶ The vicinity of a junction formed by different materials, dopant types, or doping concentrations, n and p could be smaller or larger than N_D and N_A , respectively.
- ▶ Under bias, the carrier concentrations n and p can be increased beyond their values in equilibrium.
- ▶ When the injected n or p is larger than its equilibrium value as well as the doping concentration, the space-charge effect is present.

- ▶ The injected carriers thus control the space charge and the electric-field profile.
- ▶ This results in a feedback mechanism where the field drives the current, which in turn sets up the field.
- ▶ The space-charge effect is more common in lightly doped materials, and it can occur outside the depletion region.
- ▶ space-charge-limited current : In the presence of a space-charge effect, the current is dominated by the drift component of the injected carriers.

$$J = qnv.$$

- ▶ The space charge again is determined by the injected carriers giving rise to the Poisson equation:

$$\frac{d^2 \psi_i}{dx^2} = \frac{qn}{\epsilon_s}$$

- ▶ The carrier velocity v is related to the electric field by different functions, depending on the field strength. In the low-field mobility regime

$$v = \mu \mathcal{E}$$

- ▶ In the velocity-saturation regime, velocity v_s is independent of the field. In the limit of ultra-short sample or time scale, we have the ballistic regime where there is no scattering.

$$v = \sqrt{\frac{2qV}{m^*}}$$

- ▶ the space-charge-limited current in the mobility regime (the Mott-Gurney law) can be solved to be:

$$J = \frac{4\epsilon_s}{9L^2} \left(\frac{2q}{m^*}\right)^{1/2} V^{3/2}.$$

- ▶ in the velocity-saturation regime

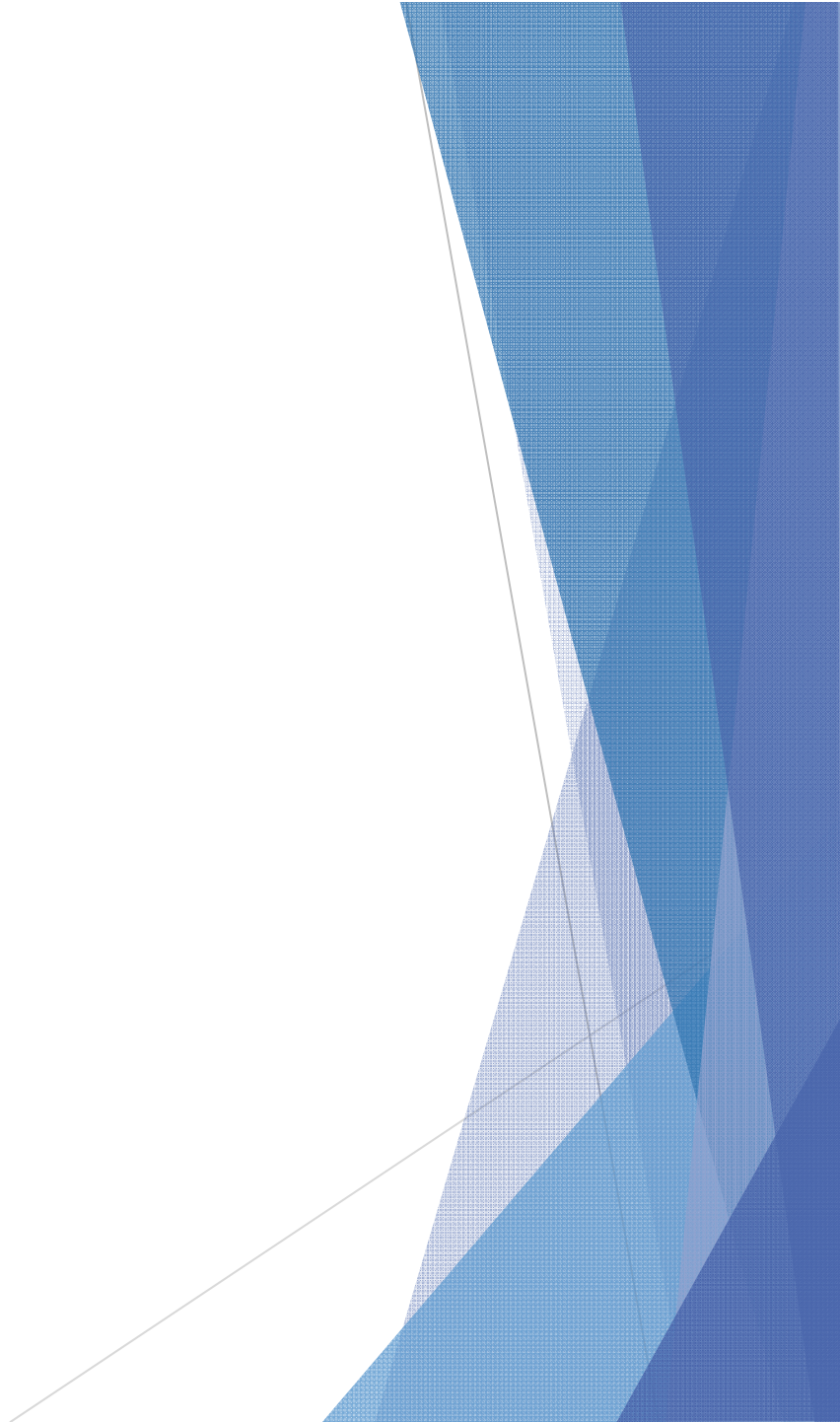
$$J = \frac{2\epsilon_s v_s V}{L^2},$$

$$J = \frac{9\epsilon_s \mu V^2}{8L^3}$$

- ▶ Here L is the length of the sample in the direction of the current flow.

this is the ballistic regime

The end





Chapter 2:p-n Junctions

Chapter 2:p-n Junctions

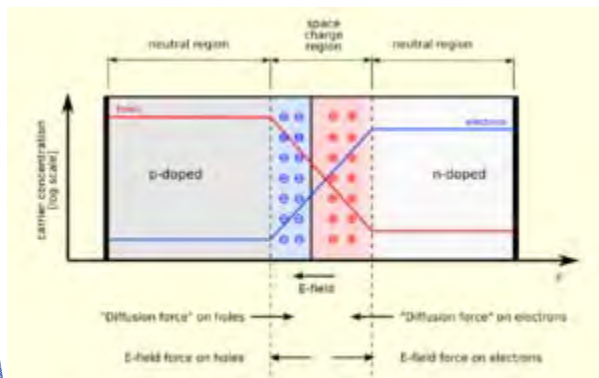
- ▶ 2.1 INTRODUCTION
- ▶ 2.2 DEPLETION REGION
- ▶ 2.3 CURRENT-VOLTAGE CHARACTERISTICS
- ▶ 2.4 JUNCTION BREAKDOWN
- ▶ 2.5 TRANSIENT BEHAVIOR AND NOISE
- ▶ 2.6 TERMINAL FUNCTIONS
- ▶ 2.7 HETEROJUNCTIONS

2.1 INTRODUCTION

- ▶ The p-n junction theory serves as the foundation of the physics of semiconductor devices.
- ▶ The basic theory of current voltage characteristics of p-n junctions was established by Shockley.
- ▶ The basic equations presented in Chapter 1 are used to develop the ideal static and dynamic characteristics of p-n junctions.

Junction

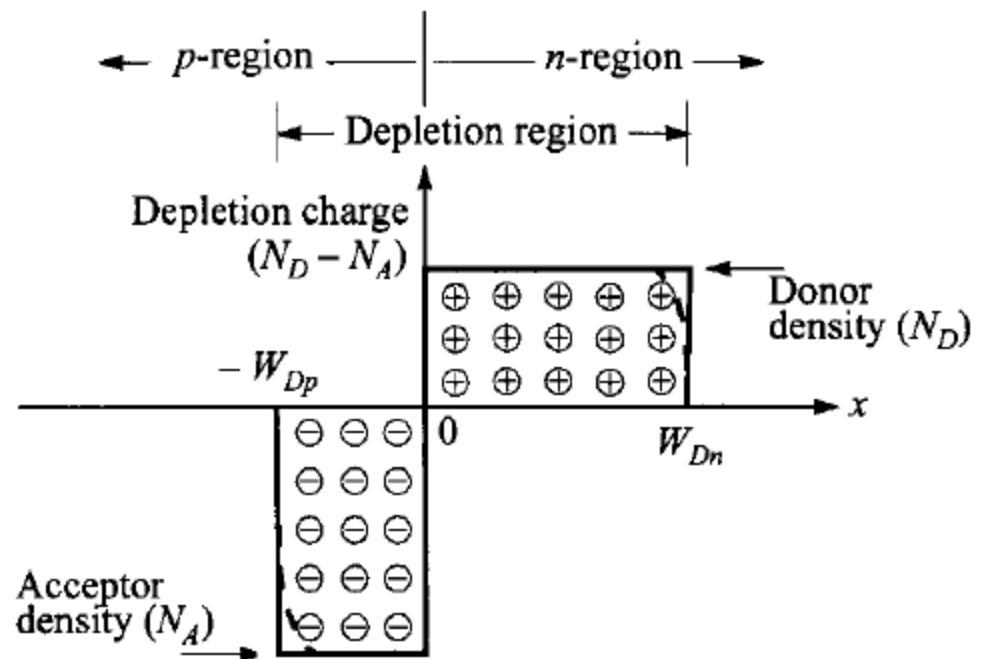
- ▶ A **p-n junction** is a boundary or interface between two types of semiconductor material.
- ▶ p-n junctions are elementary "building blocks" of most semiconductor electronic devices such as diodes, transistors, solar cells, LEDs, and integrated circuits.



JUNCTION DIODE

DEPLETION REGION

- ▶ 2.2.1 Abrupt Junction:
- ▶ When the impurity concentration in a semiconductor changes abruptly from acceptor impurities N_A to donor impurities N_D , obtains a one-sided abrupt $p^+ - n$ or $n^+ - p$ junction.



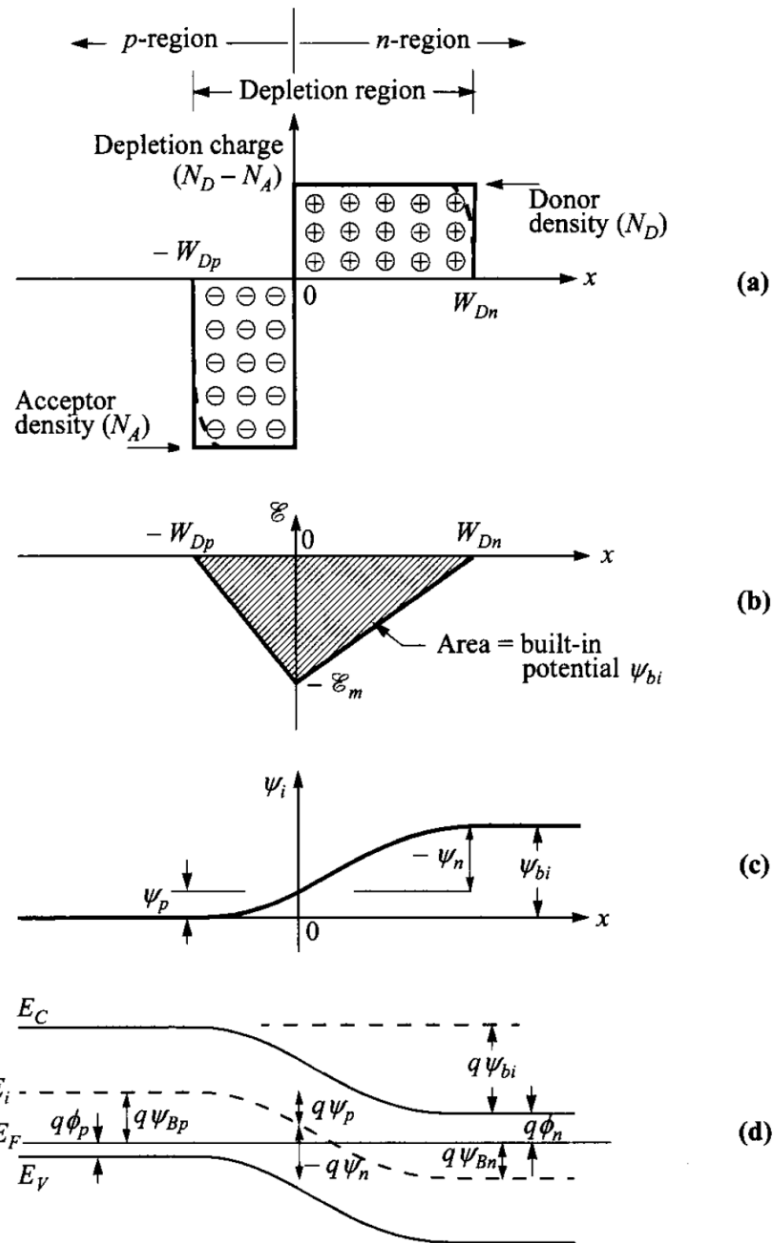


Fig. 1 Abrupt p - n junction in thermal equilibrium. (a) Space-charge distribution. Dashed lines indicate corrections to depletion approximation. (b) Electric-field distribution. (c) Potential distribution where ψ_{bi} is the built-in potential. (d) Energy-band diagram.

Built-in Potential and Depletion-Layer Width

- ▶ We first consider the thermal equilibrium condition, that is, one without applied voltage and current flow.

- ▶ From the current equation of drift and diffusion:

$$J_n = 0 = q\mu_n \left(n\mathcal{E} + \frac{kT}{q} \frac{dn}{dx} \right) = \mu_n n \frac{dE_F}{dx}$$

- ▶ Similarly

$$J_p = 0 = \mu_p p \frac{dE_F}{dx}$$

- ▶ The condition of zero net electron and hole currents requires that the Fermi level must be constant throughout the sample.
- ▶ The built-in potential ψ_{bi} , or diffusion potential

$$q\psi_{bi} = E_g - (q\phi_n + q\phi_p) = q\psi_{Bn} + q\psi_{Bp}$$

- ▶ For non-degenerate semiconductors:

$$\begin{aligned}\psi_{bi} &= \frac{kT}{q} \ln\left(\frac{n_{no}}{n_i}\right) + \frac{kT}{q} \ln\left(\frac{p_{po}}{n_i}\right) \\ &\approx \frac{kT}{q} \ln\left(\frac{N_D N_A}{n_i^2}\right) .\end{aligned}$$

Since at equilibrium $n_{no} p_{no} = n_{po} p_{po} = n_i^2$

$$\psi_{bi} = \frac{kT}{q} \ln\left(\frac{p_{po}}{p_{no}}\right) = \frac{kT}{q} \ln\left(\frac{n_{no}}{n_{po}}\right) .$$

This gives the relationship between carrier densities on either side of the junction.

- ▶ If one or both sides of the junction are degenerate, care has to be taken in calculating the Fermi-levels and built-in potential.
- ▶ Incomplete ionization has to be considered

$$n_{no} \neq N_D \text{ and/or } p_{po} \neq N_A$$

- ▶ we proceed to calculate the field and potential distribution inside the depletion region.
- ▶ To simplify the analysis, the depletion approximation is used which assumes that the depleted charge has a box profile.
- ▶ Since in thermal equilibrium the electric field in the neutral regions of the semiconductor must be zero.
- ▶ the total negative charge per unit area in the p-side must be equal to the total positive charge per unit area in the n-side
- ▶ Solving by using poison equation

$$N_A W_{Dp} = N_D W_{Dn}.$$

$$-\frac{d^2 \psi_i}{dx^2} = \frac{d\mathcal{E}}{dx} = \frac{\rho(x)}{\epsilon_s} = \frac{q}{\epsilon_s} [N_D^+(x) - n(x) - N_A^-(x) + p(x)].$$

- ▶ Inside the depletion region $n(x) = p(x) = 0$, and assuming complete ionization.

$$\frac{d^2 \psi_i}{dx^2} \approx \frac{qN_A}{\epsilon_s} \quad \text{for } -W_{Dp} \leq x \leq 0 ,$$

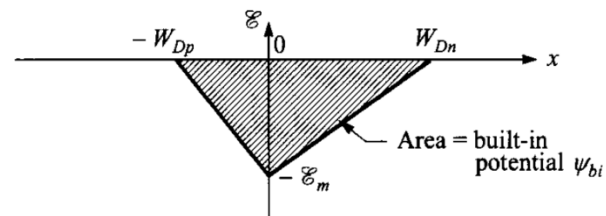
$$-\frac{d^2 \psi_i}{dx^2} \approx \frac{qN_D}{\epsilon_s} \quad \text{for } 0 \leq x \leq W_{Dn} .$$

The electric field is then obtained by integrating the above equations:

$$\mathcal{E}(x) = -\frac{qN_A(x + W_{Dp})}{\epsilon_s} \quad \text{for } -W_{Dp} \leq x \leq 0 ,$$

$$= -\frac{qN_D}{\epsilon_s}(W_{Dn} - x) \quad \text{for } 0 \leq x \leq W_{Dn}$$

$$\mathcal{E}(x) = -\mathcal{E}_m + \frac{qN_D x}{\epsilon_s}$$



(b)

- ▶ where E_m the maximum field that exists at $x = 0$ and is given by:

$$|E_m| = \frac{qN_D W_{Dn}}{\epsilon_s} = \frac{qN_A W_{Dp}}{\epsilon_s}.$$

- ▶ To get the built in potential distribution we integrate the equation of the electric field given above.

$$\psi_i(x) = \frac{qN_A}{2\epsilon_s}(x + W_{Dp})^2 \quad \text{for } -W_{Dp} \leq x \leq 0,$$

$$\psi_i(x) = \psi_i(0) + \frac{qN_D}{\epsilon_s}\left(W_{Dn} - \frac{x}{2}\right)x \quad \text{for } 0 \leq x \leq W_{Dn}$$

- ▶ With these, the potentials across different regions can be found as:

$$\psi_p = \frac{qN_A W_{Dp}^2}{2\epsilon_s},$$

$$|\psi_n| = \frac{qN_D W_{Dn}^2}{2\epsilon_s}, \quad \psi_{bi} = \psi_p + |\psi_n| = \psi_i(W_{Dn}) = \frac{|E_m|}{2}(W_{Dp} + W_{Dn})$$

- ▶ where E_m can also be expressed as:

$$|E_m| = \sqrt{\frac{2qN_A \psi_p}{\epsilon_s}} = \sqrt{\frac{2qN_D |\psi_n|}{\epsilon_s}}.$$

- ▶ the depletion widths are calculated to be:

$$W_{Dp} = \sqrt{\frac{2\epsilon_s \psi_{bi}}{q} \frac{N_D}{N_A(N_A + N_D)}},$$

$$W_{Dn} = \sqrt{\frac{2\epsilon_s \psi_{bi}}{q} \frac{N_A}{N_D(N_A + N_D)}},$$

$$W_{Dp} + W_{Dn} = \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) \psi_{bi}}.$$

- ▶ The following relationships can be deduced:

$$\frac{|\psi_n|}{\psi_{bi}} = \frac{W_{Dn}}{W_{Dp} + W_{Dn}} = \frac{N_A}{N_A + N_D},$$

$$\frac{\psi_p}{\psi_{bi}} = \frac{W_{Dp}}{W_{Dp} + W_{Dn}} = \frac{N_D}{N_A + N_D}.$$

- ▶ For a one-sided abrupt junction (*p⁺-n or n⁺-p*)

- ▶ In this case, the majority of the potential variation and depletion region will be inside the lightly doped side.

$$W_D = \sqrt{\frac{2\epsilon_s \psi_{bi}}{qN}}$$

- ▶ where *N* is *N_D* or *N_A* depending on whether *N_A* >> *N_D* or vice versa:

$$\psi_i(x) = |\mathcal{E}_m| \left(x - \frac{x^2}{2W_D} \right).$$

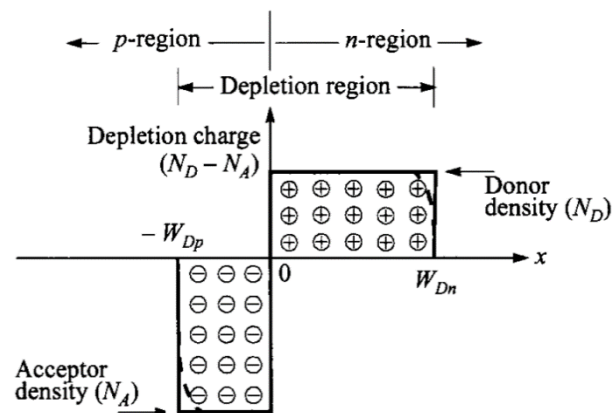
- ▶ The depletion-layer properties can be obtained by considering the majority-carrier contribution in addition to the impurity concentration.

- ▶ $n_{\text{side}} \quad \rho \approx q[N_D - n(x)] \quad p_{\text{side}} \quad \rho \approx -q[N_A - p(x)]$

- ▶ There is a correction factor of kT/q should be subtracted from ψ_{bi} to get the depletion width because of the two majority-carrier distribution tails near the edges of the depletion region.

- ▶ Each contributes a correction factor kT/q . The depletion-layer width at thermal equilibrium for a one-sided abrupt junction becomes:

$$W_D = \sqrt{\frac{2\epsilon_s}{qN} \left(\psi_{bi} - \frac{2kT}{q} \right)}$$



(a)

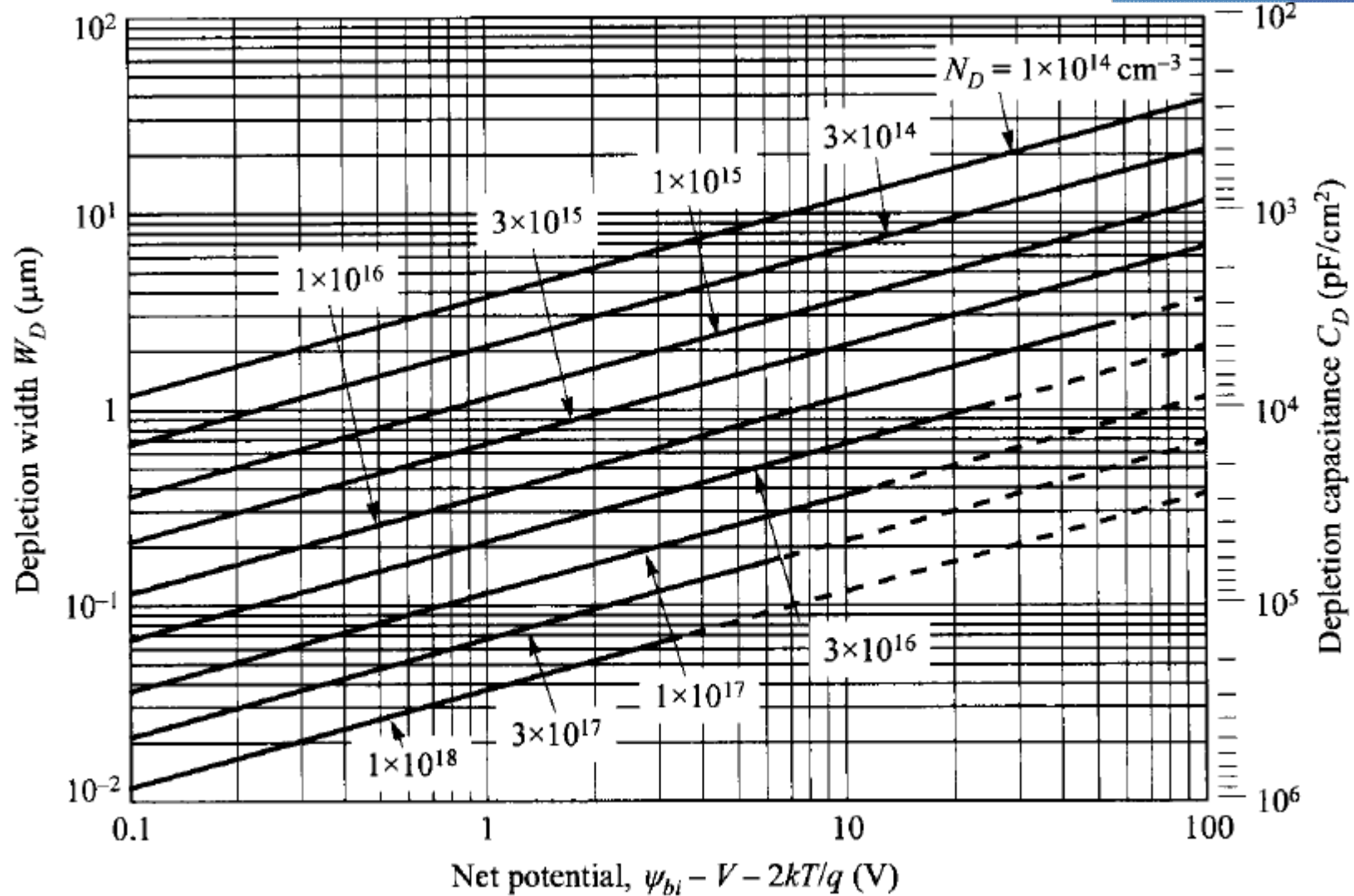
- ▶ V is applied to the junction, the total electrostatic potential variation across the junction

$$(\psi_{bi} - V)$$

- ▶ V is positive for forward bias (positive voltage on p-region with respect to n-region)
- ▶ V is negative for reverse bias (positive voltage on n-region with respect to p-region)
- ▶ Then the voltage will affect the depletion width

$$\psi_{bi} - V - \frac{2kT}{q}$$

- ▶
$$W_D = \sqrt{\frac{2\epsilon_s}{qN} \left(\psi_{bi} - V - \frac{2kT}{q} \right)}$$



To obtain the depletion-layer width for other semiconductors such as Ge, one must multiply the results of Si by the factor

$$\sqrt{\epsilon_s(\text{Ge})/\epsilon_s(\text{Si})}$$

Depletion-Layer Capacitance.

- ▶ The depletion-layer capacitance per unit area is defined as

$$C_D = dQ_D/dV = \epsilon_s/W_D.$$

- ▶ where dQ_D is the incremental depletion charge on each side of the junction

$$C_D = \frac{\epsilon_s}{W_D} = \sqrt{\frac{q\epsilon_s N}{2}} \left(\psi_{bi} - V - \frac{2kT}{q} \right)^{-1/2}$$

- ▶ where V is positive for reverse bias and negative for forward reverse bias.
- ▶ Rearrange the capacitance equation leads to:

$$\frac{1}{C_D^2} = \frac{2}{q\epsilon_s N} \left(\psi_{bi} - V - \frac{2kT}{q} \right),$$

$$\frac{d(1/C_D^2)}{dV} = -\frac{2}{q\epsilon_s N}.$$

- ▶ The slope gives the impurity concentration of the substrate (N), and the extrapolation to $1/C^2 = 0$ gives $(\psi_{bi} - \frac{2kT}{q})$

$$\frac{1}{C_D^2} = \frac{2}{q\epsilon_s N} \left(\psi_{bi} - V - \frac{2kT}{q} \right),$$

$$\frac{d(1/C_D^2)}{dV} = -\frac{2}{q\epsilon_s N}.$$

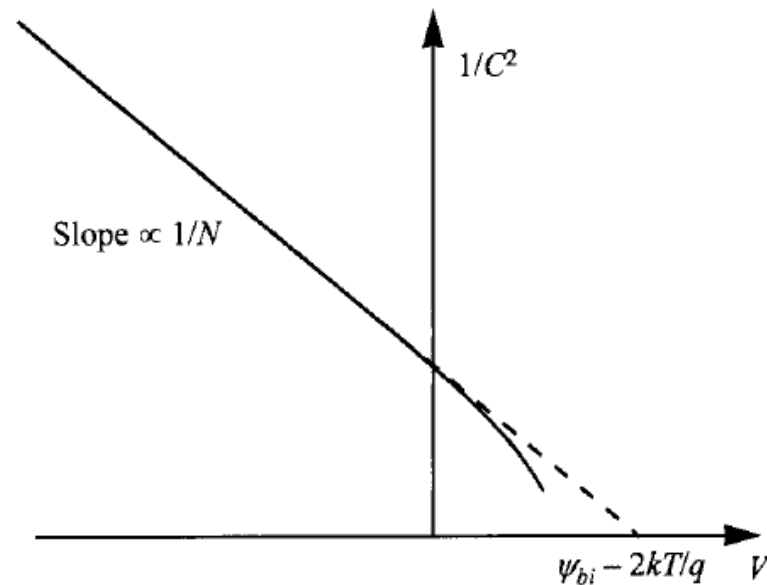


Fig. 3 A $1/C^2$ - V plot can yield the built-in potential and doping density N .

- ▶ The semiconductor potential and the capacitance-voltage data are insensitive to changes in the doping profiles

$$\frac{d(1/C_D^2)}{dV} = -\frac{2}{q\epsilon_s N}$$

- ▶ The Debye length L_D is a characteristic length for semiconductors and is defined as:

$$L_D \equiv \sqrt{\frac{\epsilon_s kT}{q^2 N}}$$

- ▶ This Debye length gives an idea of the limit of the potential change in response to an abrupt change in the doping profile.
- ▶ Consider a case where the doping has a small increase of ΔN , in the background of N_D , the change of potential $\Delta\psi_i(x)$

$$n = N_D \exp\left(\frac{\Delta\psi_i q}{kT}\right),$$

$$\frac{d^2 \Delta \psi_i}{dx^2} = -\frac{q}{\epsilon_s} (N_D + \Delta N_D - n) = -\frac{q N_D}{\epsilon_s} \left[1 + \frac{\Delta N_D}{N_D} - \exp\left(\frac{\Delta \psi_i q}{kT}\right) \right]$$

$$\approx -\frac{q N_D}{\epsilon_s} \left[1 + \frac{\Delta N_D}{N_D} - \left(1 + \frac{\Delta \psi_i q}{kT} \right) \right] \approx \frac{q^2 N_D}{\epsilon_s kT} \Delta \psi_i$$

$$L_D \equiv \sqrt{\frac{\epsilon_s kT}{q^2 N}}$$

- ▶ “This implies that if the doping profile changes abruptly in a scale less than the Debye length, this variation has no effect and cannot be resolved.”
- ▶ if the depletion width is smaller than the Debye length, the analysis using the Poisson equation is no longer valid.
- ▶ At thermal equilibrium the depletion-layer widths of abrupt junctions are about $8L_D$ for Si and $10L_D$ for GaAs.
- ▶ The Debye length as a function of doping density and temperature.

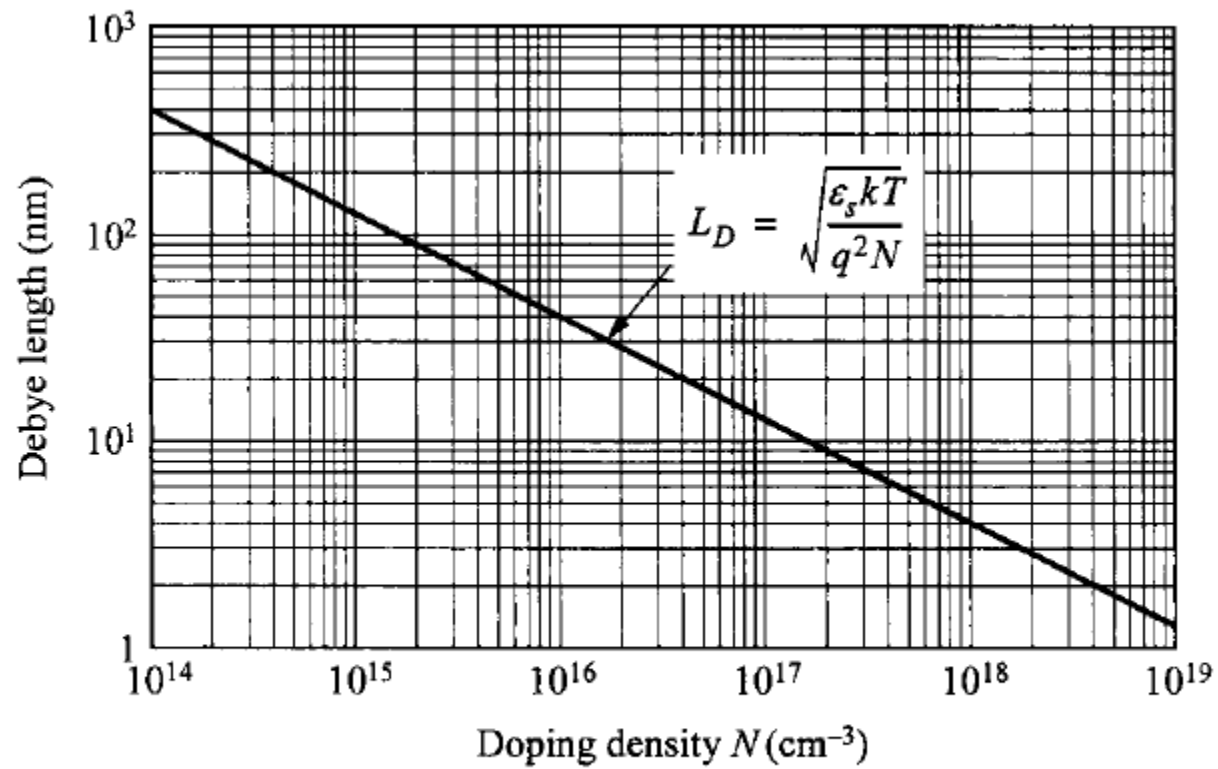
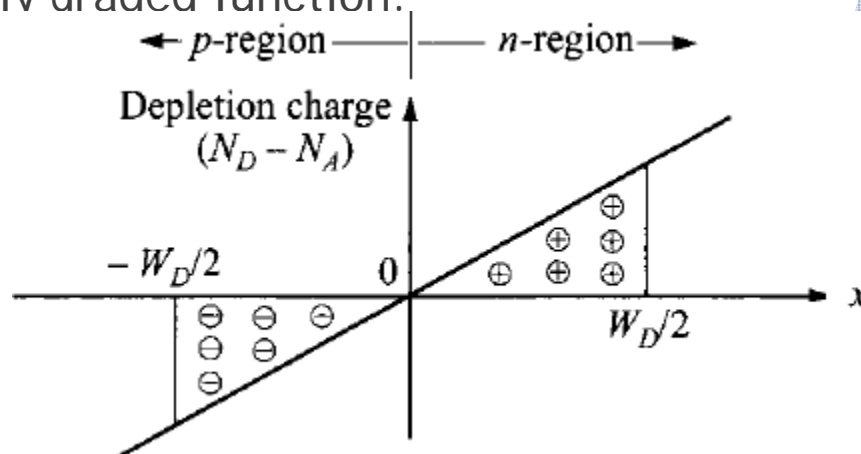


Fig. 4 Debye length in Si at room temperature as a function of doping density N .

2.2.2 Linearly Graded Junction

- ▶ In practical devices, the doping profiles are not abrupt.
- ▶ where the two types meet and they compensate each other.
- ▶ The depletion widths terminate within this transition region, the doping profile can be approximated by a linear function.
- ▶ At thermal-equilibrium case first The impurity distribution for a linearly graded junction.



- ▶ The Poisson equation for this case is:

$$-\frac{d^2 \psi_i}{dx^2} = \frac{d\mathcal{E}}{dx} = \frac{\rho(x)}{\epsilon_s} = \frac{q}{\epsilon_s}(p - n + ax)$$

$$\approx \frac{qax}{\epsilon_s}$$

$$-\frac{W_D}{2} \leq x \leq \frac{W_D}{2}$$

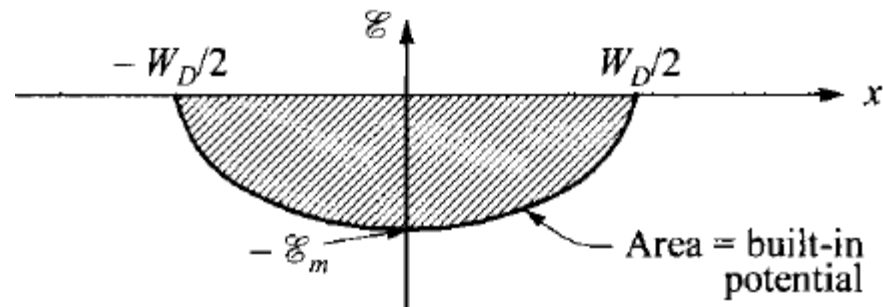
- ▶ where a is the doping gradient in cm^{-4} .

$$\frac{d\mathcal{E}}{dx} = \frac{\rho(x)}{\epsilon_s} = \frac{q}{\epsilon_s}(p - n + ax)$$

$$\mathcal{E}(x) = -\frac{qa}{2\epsilon_s} \left[\left(\frac{W_D}{2} \right)^2 - x^2 \right] \quad -\frac{W_D}{2} \leq x \leq \frac{W_D}{2}$$

- ▶ with the maximum field E_m at $x = 0$,

$$|\mathcal{E}_m| = \frac{qaW_D^2}{8\epsilon_s}$$



- ▶ Integrating $E(x)$ it gives the potential distribution

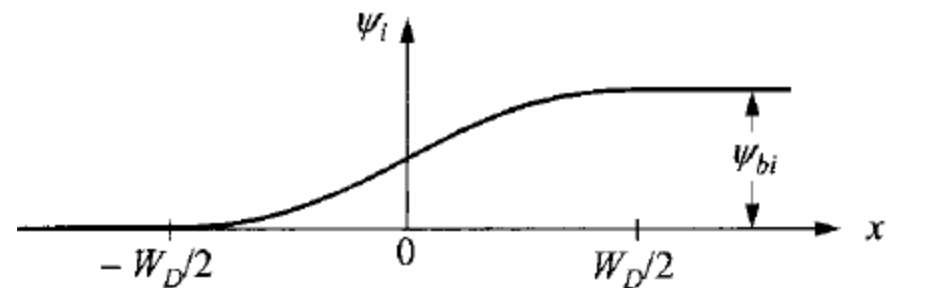
$$\psi_i(x) = \frac{qa}{6\epsilon_s} \left[2\left(\frac{W_D}{2}\right)^3 + 3\left(\frac{W_D}{2}\right)^2 x - x^3 \right] \quad -\frac{W_D}{2} \leq x \leq \frac{W_D}{2}$$

- ▶ from which the built-in potential can be related to the depletion width

$$\psi_{bi} = \frac{qaW_D^3}{12\epsilon_s} \quad W_D = \left(\frac{12\epsilon_s\psi_{bi}}{qa} \right)^{1/3}$$

- ▶ Since the values of the impurity concentrations at the edges of the depletion region ($-W_d/2$ and $W_d/2$) are the same and equal to $aW_d/2$,

$$\begin{aligned} \psi_{bi} &\approx \frac{kT}{q} \ln \left[\frac{(aW_D/2)(aW_D/2)}{n_i^2} \right] \\ &\approx \frac{2kT}{q} \ln \left(\frac{aW_D}{2n_i} \right) \end{aligned}$$



- ▶ the built-in potential can be calculated explicitly by an expression as a gradient voltage V_g :

$$V_g = \frac{2kT}{3q} \ln\left(\frac{a^2 \epsilon_s kT}{8n_i^3 q^2}\right)$$

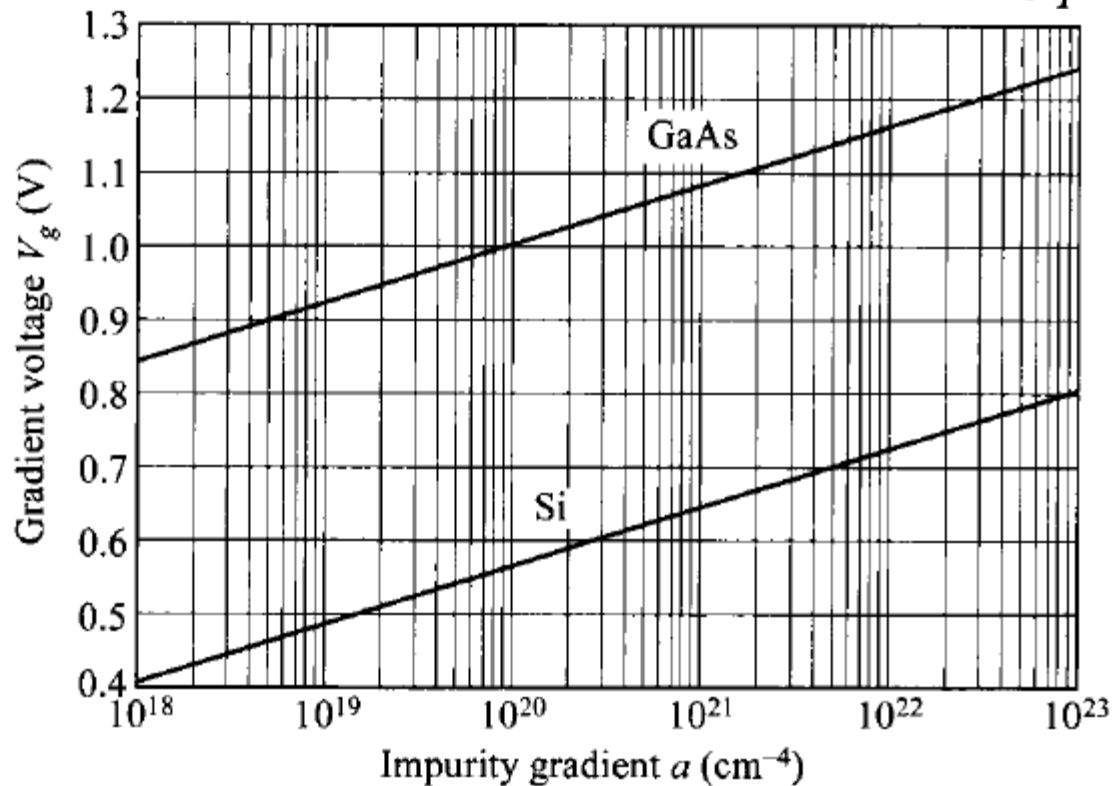


Fig. 6 Gradient voltages for linearly graded junctions in Si and GaAs.

- ▶ The depletion-layer capacitance for a linearly graded junction is given by

$$C_D = \frac{\epsilon_s}{W_D} = \left[\frac{qa\epsilon_s^2}{12(\psi_{bi} - V)} \right]^{1/3}$$

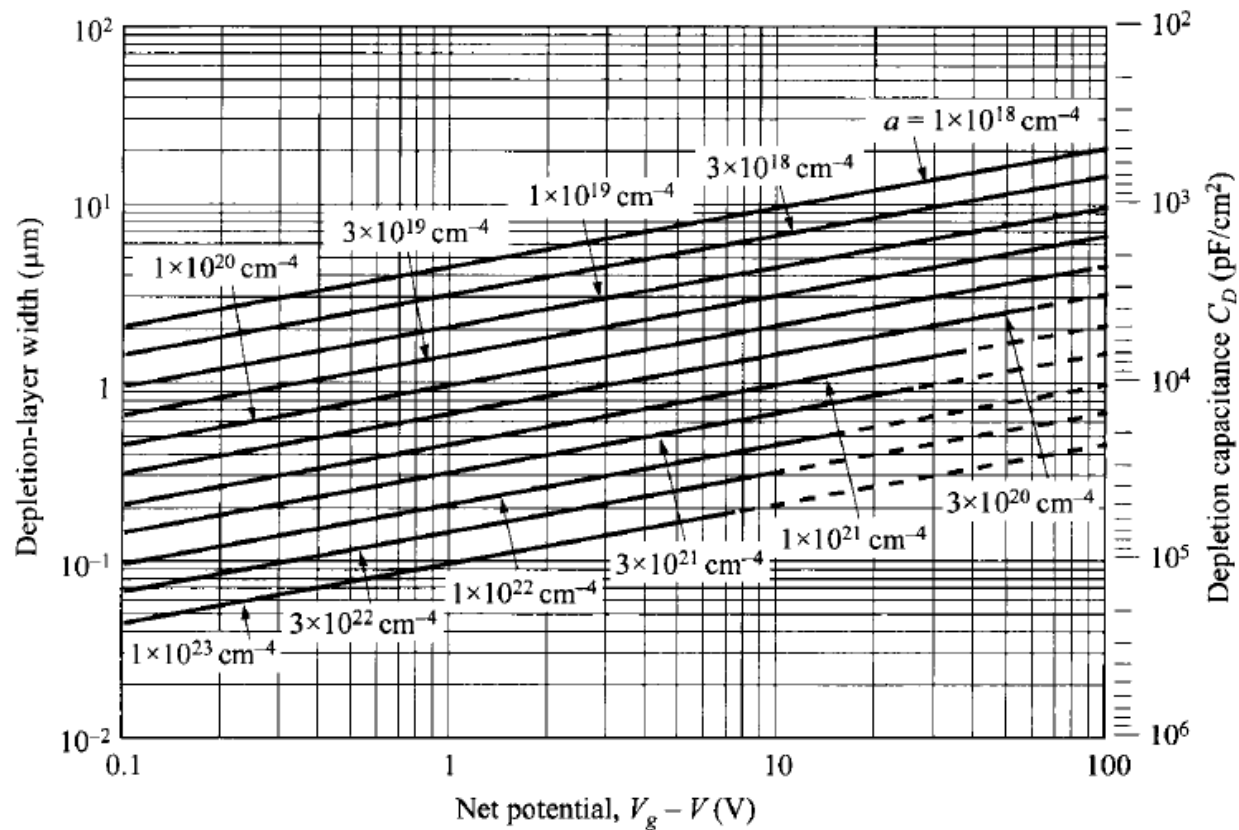


Fig. 7 Depletion-layer width and depletion-layer capacitance per unit area as a function of net potential ($V_g - V$) for different impurity gradients in linearly graded junctions in Si. Dashed lines represent breakdown conditions.

2.2.3 Arbitrary Doping Profile

- ▶ The doping near the junction to be of any arbitrary shape.
- ▶ the net potential change at the junction is given by integrating the total field across the depletion region.

$$\psi_n = \psi_{n0} - V = -\int_0^{W_D} \mathcal{E}(x) dx = -x\mathcal{E}(x) \Big|_0^{W_D} + \int_{\mathcal{E}(0)}^{\mathcal{E}(W_D)} x d\mathcal{E},$$

where ψ_{n0} is ψ_n at zero bias.

$$\psi_n = \int_{\mathcal{E}(0)}^{\mathcal{E}(W_D)} x \frac{d\mathcal{E}}{dx} dx = \frac{q}{\epsilon_s} \int_0^{W_D} x N_D(x) dx.$$

- ▶ total depletion-layer charge is given by
- ▶ Differentiating the quantities with respect to the depletion width gives:

$$Q_D = q \int_0^{W_D} N_D(x) dx.$$

$$\frac{dV}{dW_D} = - \frac{d\psi_n}{dW_D} = - \frac{q N_D(W_D) W_D}{\epsilon_s},$$

$$\frac{dQ_D}{dW_D} = q N_D(W_D).$$

$$C_D = \left| \frac{dQ_D}{dV} \right| = \left| \frac{dQ_D}{dW_D} \times \frac{dW_D}{dV} \right| = \frac{\epsilon_s}{W_D}.$$

$$\begin{aligned} \frac{d(1/C_D^2)}{dV} &= \frac{d(1/C_D^2)}{dW_D} \frac{dW_D}{dV} = \frac{2W_D dW_D}{\epsilon_s^2 dV} \\ &= - \frac{2}{q \epsilon_s N_D(W_D)}. \end{aligned}$$

2.3 CURRENT-VOLTAGE CHARACTERISTICS

▶ 2.3.1 Ideal Case-Shockley:

▶ The ideal current-voltage characteristics are based on the following four assumptions:

1. the abrupt depletion-layer approximation; that is, the built-in potential and applied voltages are supported by a dipole layer with abrupt boundaries, and outside the boundaries the semiconductor is assumed to be neutral.
2. the Boltzmann approximation

$$n = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \quad \text{or} \quad E_C - E_F = kT \ln\left(\frac{N_C}{n}\right).$$

3. the low-injection assumption; that is, the injected minority carrier densities are small compared with the majority-carrier densities.
4. no generation-recombination current exists inside the depletion layer, and the electron and hole currents are constant throughout the depletion layer.

- ▶ first consider the Boltzmann relation. At thermal equilibrium this relation is given by:

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right),$$

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right).$$

- ▶ at thermal equilibrium, the pn product from the above equations is equal to n_i^2 .
- ▶ By applying V , the minority-carrier densities on both sides of the junction are changed, and the pn product is no longer equal to n_i^2 .
- ▶ the quasi-Fermi (imref) levels is equal:

$$n \equiv n_i \exp\left(\frac{E_{Fn} - E_i}{kT}\right),$$

$$p \equiv n_i \exp\left(\frac{E_i - E_{Fp}}{kT}\right),$$

- ▶ where E_{Fn} and E_{Fp} are the quasi-Fermi levels for electrons and holes, respectively.

$$E_{Fn} \equiv E_i + kT \ln\left(\frac{n}{n_i}\right),$$

$$E_{Fp} \equiv E_i - kT \ln\left(\frac{p}{n_i}\right).$$

- ▶ The pn product becomes:

$$pn = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fp}}{kT}\right).$$

- ▶ For a forward bias, $(E_{Fn} - E_{Fp}) > 0$ and $pn > n_i^2$,
- ▶ For a reversed bias, $(E_{Fn} - E_{Fp}) < 0$ and $pn < n_i^2$.
- ▶ the electron and hole current densities are proportional to the gradients of the electron and hole quasi-Fermi levels, respectively.

$$\begin{aligned} \mathcal{E} &\equiv \nabla E_i / q & J_n &= q\mu_n \left(n\mathcal{E} + \frac{kT}{q} \nabla n \right) = \mu_n n \nabla E_i + \mu_n kT \left[\frac{n}{kT} (\nabla E_{Fn} - \nabla E_i) \right] \\ & & &= \mu_n n \nabla E_{Fn} . \end{aligned}$$

- ▶ If $E_{Fn} = E_{Fp} = \text{constant}$ (at thermal equilibrium), then $J_n = J_p = 0$.
- ▶ The variations of E_{Fn} and E_{Fp} with distance are related to the carrier concentrations as given in:

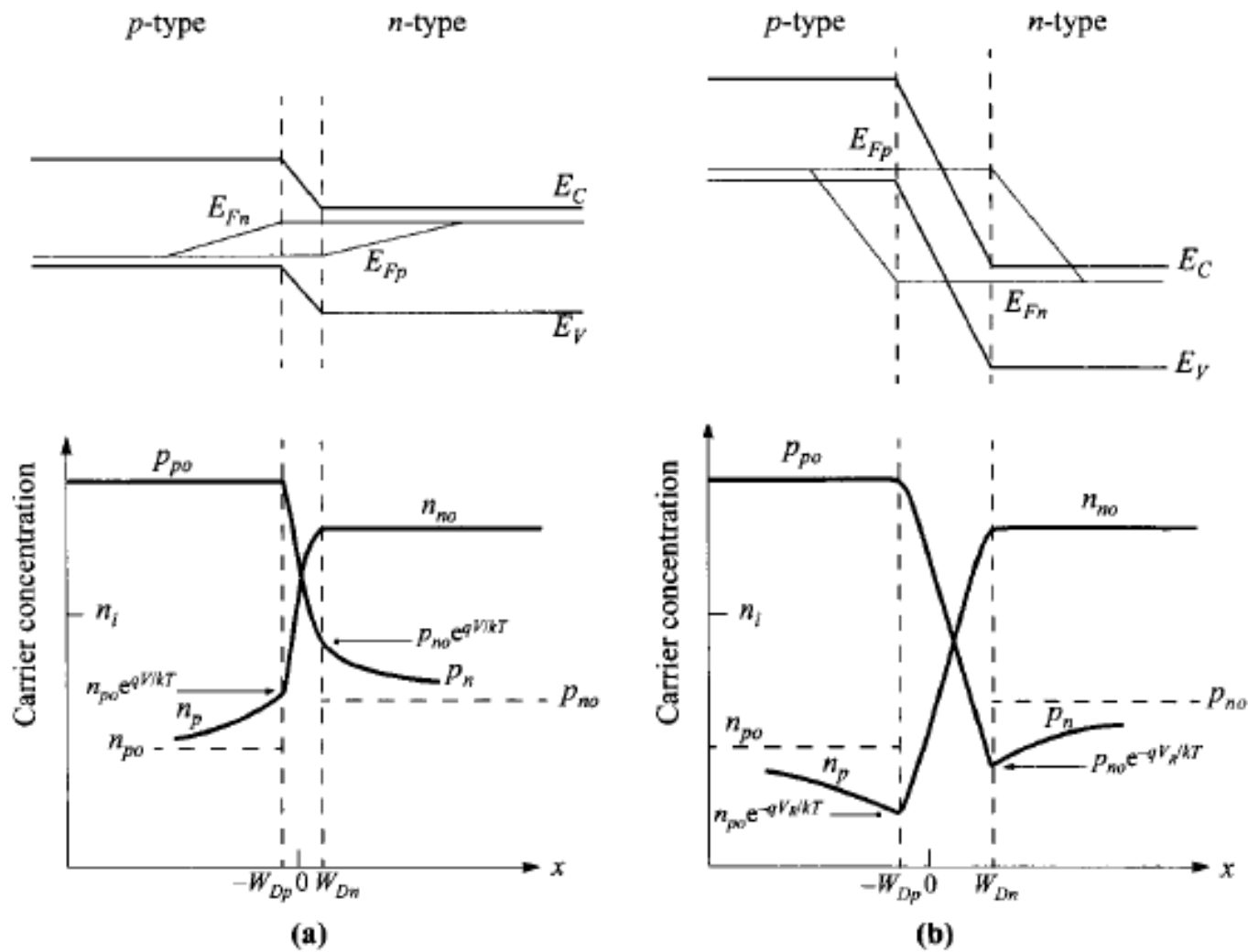


Fig. 8 Energy-band diagram, with quasi-Fermi levels for electrons and holes, and carrier distributions under (a) forward bias and (b) reverse bias.

- ▶ Inside the depletion region E_{Fn} and E_{Fp} remain relatively constant.
- ▶ This comes about because the carrier concentrations are relatively much higher inside the depletion region.
- ▶ The currents remain fairly constant, the gradients of the quasi-Fermi levels have to be small.
- ▶ The depletion width is typically much shorter than the diffusion length.
- ▶ The total drop of quasi-Fermi levels inside the depletion width is not significant.

$$qV = E_{Fn} - E_{Fp}.$$

- ▶ the electron density at the boundary of the depletion-layer region on the p-side ($x = -W_{Dp}$):

$$n_p(-W_{Dp}) = \frac{n_i^2}{p_p} \exp\left(\frac{qV}{kT}\right) \approx n_{po} \exp\left(\frac{qV}{kT}\right)$$

- ▶ where $p_p = p_{po}$ for low-level injection, and n_{po} is the equilibrium electron density on the p-side.

$$p_n(W_{Dn}) = p_{no} \exp\left(\frac{qV}{kT}\right)$$

- ▶ From the continuity equations we obtain for the steady-state condition in the n-side of the junction:

$$-U + \mu_n \mathcal{E} \frac{dn_n}{dx} + \mu_n n_n \frac{d\mathcal{E}}{dx} + D_n \frac{d^2 n_n}{dx^2} = 0,$$

$$-U - \mu_p \mathcal{E} \frac{dp_n}{dx} - \mu_p p_n \frac{d\mathcal{E}}{dx} + D_p \frac{d^2 p_n}{dx^2} = 0.$$

- ▶ U is the net recombination rate.
- ▶ majority carriers need to adjust their concentrations

$$(n_n - n_{no}) = (p_n - p_{no})$$

The substituting in previous equation to have :

$$-\frac{p_n - p_{no}}{\tau_p} - \frac{n_n - p_n}{(n_n/\mu_p) + (p_n/\mu_n)} \mathcal{E} \frac{dp_n}{dx} + D_a \frac{d^2 p_n}{dx^2} = 0$$

Where D_a is the ambipolar diffusion coefficient,

Einstein relation $D = (kT/q)\mu$,

$$\tau_p \equiv \frac{p_n - p_{no}}{U}.$$

$$D_a = \frac{n_n + p_n}{n_n/D_p + p_n/D_n}$$

- ▶ the low-injection assumption [e.g., $p_n \ll (n_n = n_{no})$] in the n-type semiconductor so neglecting P_n

$$-\frac{p_n - p_{no}}{\tau_p} - \mu_p \mathcal{E} \frac{dp_n}{dx} + D_p \frac{d^2 p_n}{dx^2} = 0$$

- ▶ In the neutral region where there is no electric field:

$$\frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{D_p \tau_p} = 0.$$

$$p_n(x) - p_{no} = p_{no} \left[\exp\left(\frac{qV}{kT}\right) - 1 \right] \exp\left(-\frac{x - W_{Dn}}{L_p}\right)$$

$$L_p \equiv \sqrt{D_p \tau_p}.$$

- ▶ At $x = W_{Dn}$ the hole diffusion current is

$$J_p = -qD_p \frac{dp_n}{dx} \Big|_{W_{Dn}} = \frac{qD_p p_{no}}{L_p} \left[\exp\left(\frac{qV}{kT}\right) - 1 \right].$$

- ▶ we obtain the electron diffusion current in the p-side:

$$J_n = qD_n \left. \frac{dn_p}{dx} \right|_{-W_{Dp}} = \frac{qD_n n_{po}}{L_n} \left[\exp\left(\frac{qV}{kT}\right) - 1 \right].$$

- ▶ It is interesting to note that the hole current is due to injection of holes from the p-side to the n-side.
- ▶ the magnitude is determined by the properties in the n-side only.
- ▶ The total current is given by the sum J_p and J_n :

$$J = J_p + J_n = J_0 \left[\exp\left(\frac{qV}{kT}\right) - 1 \right],$$
$$J_0 \equiv \frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n} \equiv \frac{qD_p n_i^2}{L_p N_D} + \frac{qD_n n_i^2}{L_n N_A}.$$

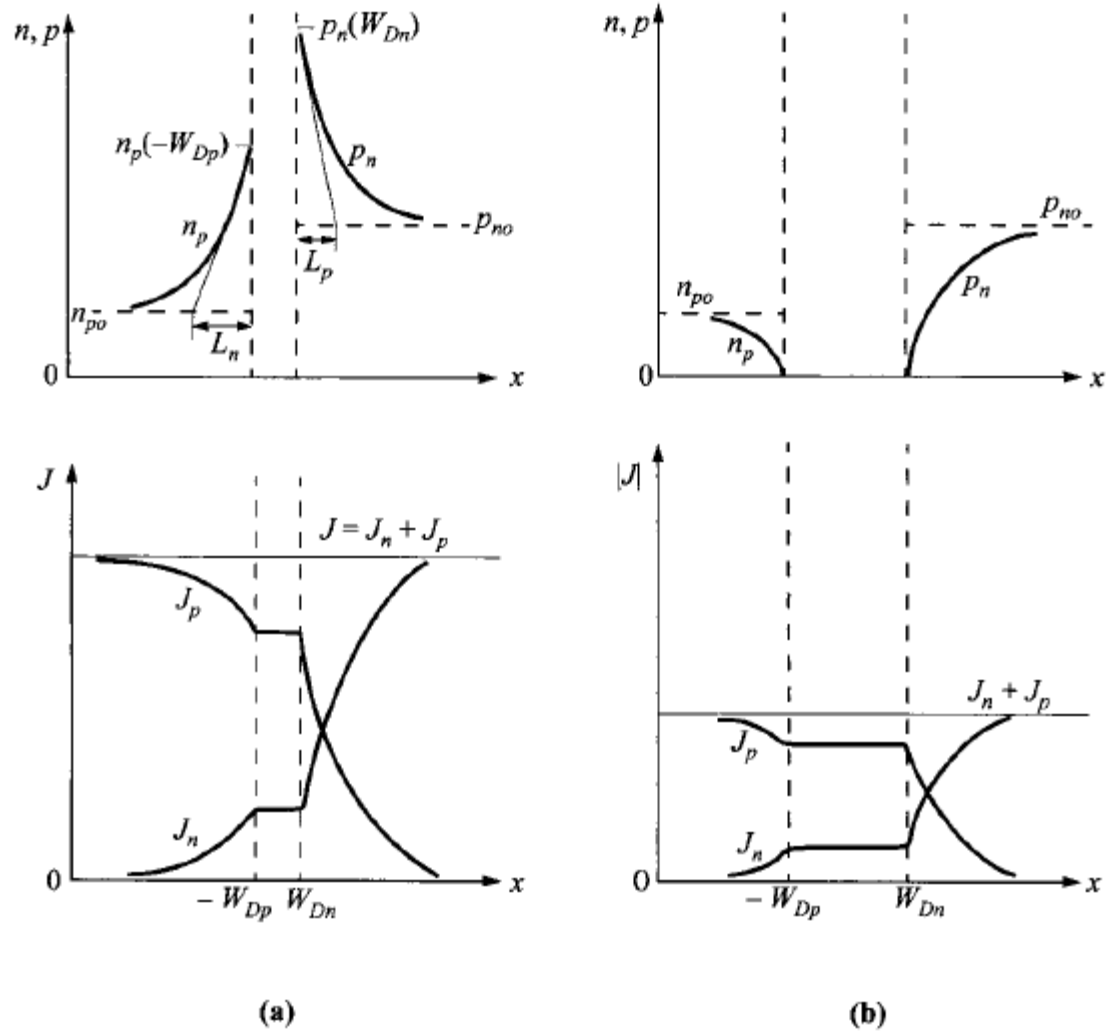


Fig. 9 Carrier distributions and current densities (both linear plots) for (a) forward-biased conditions and (b) reverse-biased conditions.

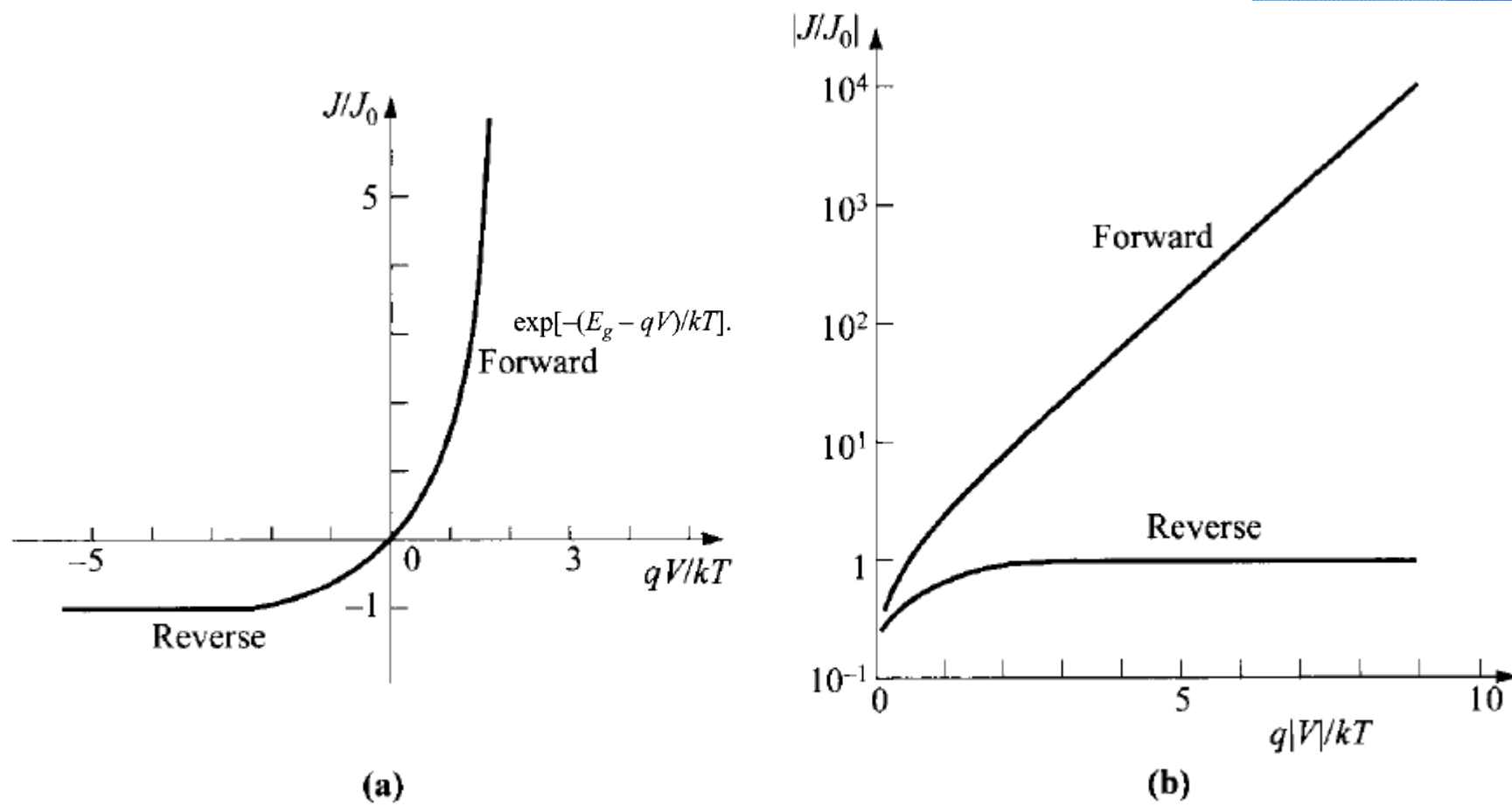


Fig. 10 Ideal current-voltage characteristics. (a) Linear plot. (b) Semilog plot.

$$\begin{aligned}
 J_0 &\approx \frac{qD_p p_{no}}{L_p} \approx q \sqrt{\frac{D_p}{\tau_p}} \frac{n_i^2}{N_D} \propto T^{\gamma/2} \left[T^3 \exp\left(-\frac{E_g}{kT}\right) \right] \\
 &\propto T^{(3 + \gamma/2)} \exp\left(-\frac{E_g}{kT}\right) .
 \end{aligned}$$

where γ is a constant

- ▶ The temperature dependence of the term $T^{(3+\gamma/2)}$ is not important compared with the exponential term.
- ▶ The slope of a plot J_0 versus $1/T$ is determined mainly by the energy gap E_g .
- ▶ It is expected that in the reverse direction, where $|J_R| \approx J_0$.
- ▶ The current will increase approximately as $e^{-\frac{E_g}{kT}}$ with temperature.
- ▶ The forward bias current density is function of the applied voltage
- ▶ The current will increase approximately

$$J_F \approx J_0 \exp(qV/kT),$$

- ▶ The Shockley equation adequately predicts the current-voltage characteristics of germanium p-n junctions at low current densities.
- ▶ For Si and GaAs p-n junctions however, the ideal equation can only give qualitative agreement

- The ideal relation of the Shockley equation is not suitable due to:
- (1) the generation and recombination of carriers in the depletion layer
 - (2) the high-injection condition that may occur even at relatively small forward bias,
 - (3) the parasitic IR drop due to series resistance,
 - (4) the tunneling of carriers between states in the bandgap.
 - (5) the surface effects.
 - (6) Under sufficiently larger field in the reverse direction, the junction will breakdown.

- The surface effects on p-n junctions are primarily due to ionic charges on or outside the semiconductor surface that induce image charges in the semiconductor
- This cause the formation of the so-called surface channels or surface depletion-layer regions.
- The channel is formed, it modifies the junction depletion region and gives rise to surface leakage current.
- For Si planar p-n junctions, the **surface leakage current** is generally much smaller than the generation-recombination current in the depletion region.

2.3.2 Generation-Recombination Process

- ▶ First the generation current under the reverse-bias condition.
- ▶ The reduction in carrier concentration under reverse bias
- ▶ The rate of generation of electron-hole pairs under Condition $p \ll n_i$ and $n \ll n_i$:

$$(pn \ll n_i^2)$$

$$U = - \left\{ \frac{\sigma_p \sigma_n v_{th} N_t}{\sigma_n \exp[(E_t - E_i)/kT] + \sigma_p \exp[(E_i - E_t)/kT]} \right\} n_i \equiv - \frac{n_i}{\tau_g}$$

- ▶ where τ_g is the generation lifetime

- ▶ The current due to generation in the depletion region is thus given by:

$$J_{ge} = \int_0^{W_D} q|U|dx \approx q|U|W_D \approx \frac{qn_iW_D}{\tau_g}$$

- ▶ where W_D is the depletion-layer width.
- ▶ If the generation lifetime (τ_g) is varying slowly as function of temperature, the generation current will then have the same temperature dependence as n_i .
- ▶ At a given temperature, J is proportional to the depletion-layer width, which in turn is dependent on the applied reverse bias.

- ▶ For abrupt junctions $J_{ge} \propto (\psi_{bi} + V)^{1/2}$

- ▶ For linearly graded junctions

$$J_{ge} \propto (\psi_{bi} + V)^{1/3}$$

- ▶ The total reverse current can be approximated by the sum of the diffusion component in the neutral region and the generation current in the depletion region

(for $p_{no} \gg n_{po}$ and $|V| > 3kT/q$)

$$J_R = q \sqrt{\frac{D_p n_i^2}{\tau_p N_D}} + \frac{q n_i W_D}{\tau_g}$$

- ▶ For semiconductors with large values of n_i (Ge)



the diffusion component will dominate at room temperature and the reverse current will follow the Shockley equation

- ▶ if n_i is small as (Si), the generation current may dominate.
- ▶ At sufficiently high temperatures, however, the diffusion current will dominate.

- ▶ At forward bias, where the major recombination-generation processes in the depletion region are the capture processes.
- ▶ a recombination current J_{re} in addition to the diffusion current

$$U = \frac{\sigma_p \sigma_n v_{th} N_t n_i^2 [\exp(qV/kT) - 1]}{\sigma_n \{n + n_i \exp[(E_t - E_i)/kT]\} + \sigma_p \{p + n_i \exp[(E_i - E_t)/kT]\}}$$

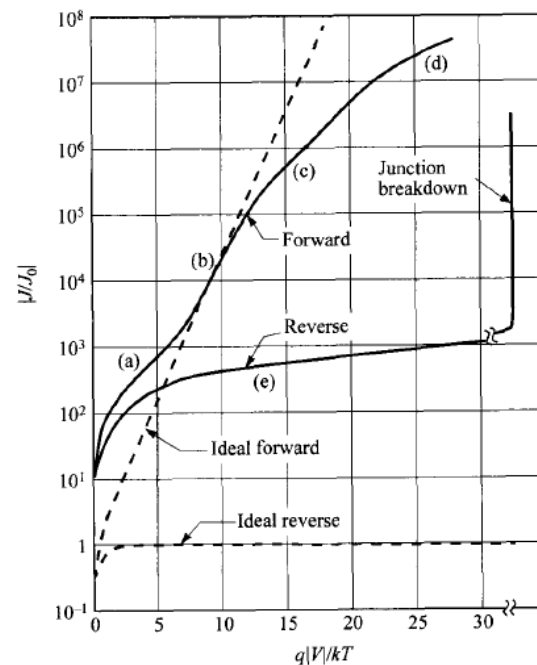


Fig. 11 Current-voltage characteristics of a practical Si diode. (a) Generation-recombination current region. (b) Diffusion-current region. (c) High-injection region. (d) Series-resistance effect. (e) Reverse leakage current due to generation-recombination and surface effects.

- ▶ Under the assumptions that $E_t = E_i$ and $\sigma_n = \sigma_p = \sigma_s$ so :

$$U = \frac{\sigma v_{th} N_t n_i^2 [\exp(qV/kT) - 1]}{n + p + 2n_i}$$

$$= \frac{\sigma v_{th} N_t n_i^2 [\exp(qV/kT) - 1]}{n_i \{ \exp[(E_{Fn} - E_i)/kT] + \exp[(E_i - E_{Fp})/kT] + 2 \}}$$

- ▶ The maximum value of U exists in the depletion region where E_i is between E_{Fn} and E_{Fp} ,

$$U \approx \frac{1}{2} \sigma v_{th} N_t n_i \exp\left(\frac{qV}{2kT}\right)$$

$$J_{re} = \int_0^{W_D} qU dx \approx \frac{qW_D}{2} \sigma v_{th} N_t n_i \exp\left(\frac{qV}{2kT}\right) \approx \frac{qW_D n_i}{2\tau} \exp\left(\frac{qV}{2kT}\right).$$

- ▶ It is assumed that most part of the depletion layer has this maximum recombination rate.

$$J_{re} = \int_0^{W_D} qU dx = \sqrt{\frac{\pi}{2}} \frac{kT n_i}{\tau \mathcal{E}_0} \exp\left(\frac{qV}{2kT}\right)$$

- ▶ the electric field at the location of maximum recombination

$$\mathcal{E}_0 = \sqrt{\frac{qN(2\psi_B - V)}{\epsilon_s}}$$

- ▶ The recombination current in forward bias is also proportional to n_i .
- ▶ The total forward current can be approximated by

$$J_F = q \sqrt{\frac{D_p}{\tau_p} \frac{n_i^2}{N_D}} \exp\left(\frac{qV}{kT}\right) + \sqrt{\frac{\pi}{2} \frac{kT n_i}{\tau_p \epsilon_0}} \exp\left(\frac{qV}{2kT}\right).$$

$$J_F \propto \exp\left(\frac{qV}{\eta kT}\right)$$

- ▶ where the ideality factor η equals 2 when the recombination current dominates.
- ▶ η equals 1 when the diffusion current dominates
- ▶ When both currents are comparable, η has a value between 1 and 2.

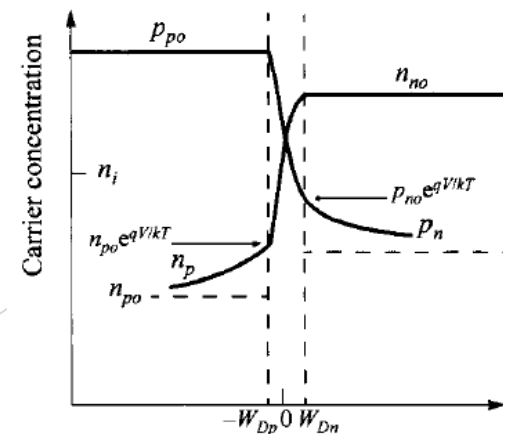
2.3.3 High-Injection Condition

- ▶ At high current densities under the forward-bias condition
- ▶ The injected minority-carrier density is comparable to the majority concentration
- ▶ The drift and diffusion current components must be considered
- ▶ The individual conduction current densities can:

$$\begin{aligned} \mathbf{J}_n &= q\mu_n \left(n\mathcal{E} + \frac{kT}{q} \nabla n \right) = \mu_n n \nabla E_i + \mu_n kT \left[\frac{n}{kT} (\nabla E_{Fn} - \nabla E_i) \right] \\ &= \mu_n n \nabla E_{Fn} . \end{aligned}$$

$$\mathbf{J}_p = \mu_p p \nabla E_{Fp} .$$

- The quasi-Fermi level for holes E_{Fp} increases in right .
- The quasi-Fermi level for electrons E_{Fn} decreases to the left.
- the separation of the two quasi-Fermi levels must be equal to or less than the applied voltage.



$$pn \leq n_i^2 \exp\left(\frac{qV}{kT}\right)$$

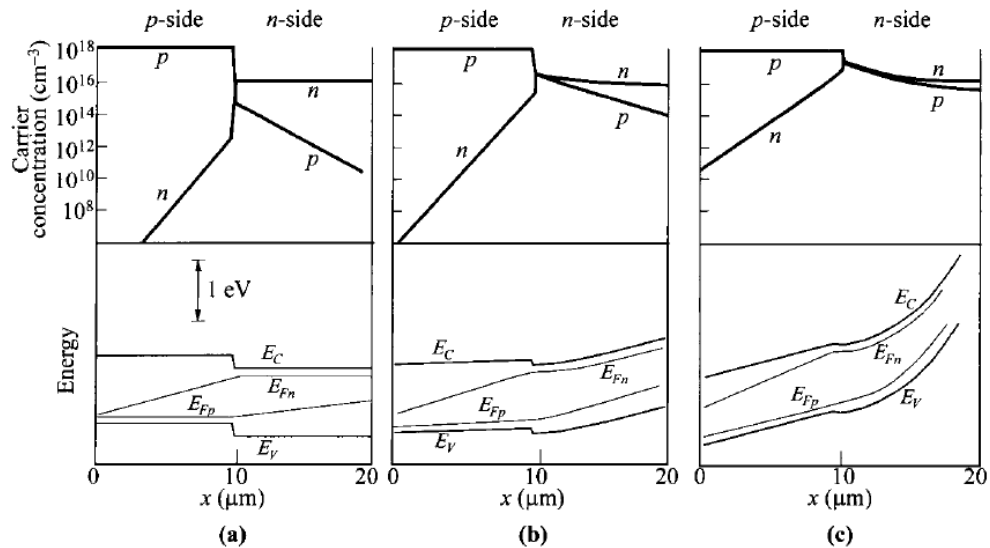


Fig. 12 Carrier concentrations and energy-band diagrams for a Si p^+n junction operated at different current densities. (a) 10 A/cm^2 . (b) 10^3 A/cm^2 . (c) 10^4 A/cm^2 . Device parameters: $N_A = 10^{18} \text{ cm}^{-3}$, $N_D = 10^{16} \text{ cm}^{-3}$, $\tau_n = 3 \times 10^{-10} \text{ s}$, and $\tau_p = 8.4 \times 10^{-10} \text{ s}$. (After Ref. 10.)

- The previous figure are numerical simulation results for carrier concentrations and energy-band diagram with quasi-Fermi levels for a silicon p^+n step junction.
- The current densities in Figs. a, b, and c are 10 , 10^3 and 10^4 A/cm^2 , respectively
- At 10 A/cm^2 the diode is in the low-injection regime
 - The potential drop occurs across the junction.
 - The hole concentration in the n-side is small compared to the electron concentration.
- At 10^3 A/cm^2 the electron concentration near the junction exceeds the donor concentration appreciably (from charge neutrality, injected carriers $\Delta p = \Delta n$)
- An ohmic potential drop appears on the n-side.
- At 10^4 A/cm^2 we have very high injection; the potential drop across the junction is insignificant compared to ohmic drops on both sides of the neutral regions.

- ▶ It is shown that the separation of the quasi-Fermi levels is equal to or less than the applied voltage (qV).
- ▶ we obtain $p_n(x = W_{Dn}) = n_i e^{\frac{qV}{2KT}}$
- ▶ The current then becomes roughly proportional to $e^{\frac{qV}{2KT}}$
- ▶ At high-current levels we should consider another effect associated with the finite resistivity in the quasi-neutral regions.
- ▶ This resistance absorbs an appreciable amount of the applied voltage (the slope starts to change)

One can estimate the series resistance from comparing the experimental curve to the ideal curve ($\Delta V = IR$).

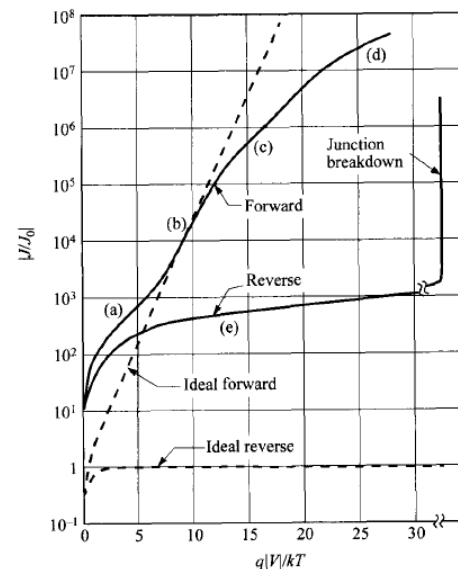


Fig. 11 Current-voltage characteristics of a practical Si diode. (a) Generation-recombination current region. (b) Diffusion-current region. (c) High-injection region. (d) Series-resistance effect. (e) Reverse leakage current due to generation-recombination and surface effects.

2.3.4 Diffusion Capacitance

- ▶ The depletion-layer capacitance considered previously accounts for most of the junction capacitance when the junction is reverse-biased.
- ▶ The diffusion capacitance is in forward-biased, there is a additional contribution to junction capacitance from the rearrangement of minority carrier density.

- ▶ When a small ac signal is applied to a junction that is forward-biased at a dc voltage V_0 and current density J_0 , the total voltage and current are defined by:

$$V(t) = V_0 + V_1 \exp(j\omega t),$$

$$J(t) = J_0 + J_1 \exp(j\omega t)$$

- ▶ where V_1 small-signal voltage and J_1 is small-signal current density,

- ▶ The admittance J_1/V_1 will give the diffusion conductance and diffusion capacitance.

$$Y \equiv \frac{J_1}{V_1} \equiv G_d + j\omega C_d .$$

- ▶ The electron and hole densities at the depletion region boundaries can be obtained by replacing V with $[V_0 + V_1 \exp(j\omega t)]$

$$V_1 \ll V_0,$$

$$n_p(-W_{Dp}) = \frac{n_i^2}{p_p} \exp\left(\frac{qV}{kT}\right) \approx n_{po} \exp\left(\frac{qV}{kT}\right)$$

$$p_n(W_{Dn}) = p_{no} \exp\left(\frac{qV}{kT}\right)$$

$$p_n(W_{Dn}) = p_{no} \exp\left\{ \frac{q[V_0 + V_1 \exp(j\omega t)]}{kT} \right\}$$

$$\approx p_{no} \exp\left(\frac{qV_0}{kT}\right) + \frac{p_{no}qV_1}{kT} \exp\left(\frac{qV_0}{kT}\right) \exp(j\omega t) \approx p_{no} \exp\left(\frac{qV_0}{kT}\right) + \tilde{p}_n(t)$$

dc component, small-signal ac compone

$$G_p = \mathcal{E} = d\mathcal{E}/dx = 0; \quad j\omega\tilde{p}_n = -\frac{\tilde{p}_n}{\tau_p} + D_p \frac{d^2\tilde{p}_n}{dx^2} \quad \frac{d^2\tilde{p}_n}{dx^2} - \frac{\tilde{p}_n}{D_p\tau_p/(1+j\omega\tau_p)} = 0.$$

- ▶ if the carrier lifetime is expressed as:

$$\tau_p^* = \frac{\tau_p}{1+j\omega\tau_p}$$

$$J = J_p + J_n = J_0 \left[\exp\left(\frac{qV}{kT}\right) - 1 \right],$$

$$J = \left(qp_{no} \sqrt{\frac{D_p}{\tau_p^*}} + qn_{po} \sqrt{\frac{D_n}{\tau_n^*}} \right) \exp\left\{ \frac{q[V_0 + V_1 \exp(j\omega t)]}{kT} \right\}$$

$$\approx \left(qp_{no} \sqrt{\frac{D_p}{\tau_p^*}} + qn_{po} \sqrt{\frac{D_n}{\tau_n^*}} \right) \left[\exp\left(\frac{qV_0}{kT}\right) \right] \left[1 + \frac{qV_1}{kT} \exp(j\omega t) \right],$$

- ▶ with the ac component being:

$$J_1 = \left(\frac{qD_p p_{no} \sqrt{1+j\omega\tau_p}}{L_p} + \frac{qD_n n_{po} \sqrt{1+j\omega\tau_n}}{L_n} \right) \left[\exp\left(\frac{qV_0}{kT}\right) \right] \frac{qV_1}{kT}.$$

- ▶ From J_1/V_1 , both G_d and C_d can be found and they are frequency dependent.

- ▶ For relatively low frequencies ($\omega\tau_p, \omega\tau_n \ll 1$), the diffusion conductance

$$G_{d0} = \frac{q}{kT} \left(\frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n} \right) \exp\left(\frac{qV_0}{kT}\right) \quad \text{mho/cm}^2$$

- ▶ The low-frequency diffusion capacitance C_{d0}

$$\sqrt{1 + j\omega\tau} \approx (1 + 0.5j\omega\tau)$$

$$C_{d0} = \frac{q^2}{2kT} (L_p p_{no} + L_n n_{po}) \exp\left(\frac{qV_0}{kT}\right) \quad \text{F/cm}^2.$$

- ▶ This diffusion capacitance is proportional to the forward current. For an $n^+ - p$ one-sided junction, it can be shown that

$$C_{d0} = \frac{qL_n^2}{2kTD_n} J_F.$$

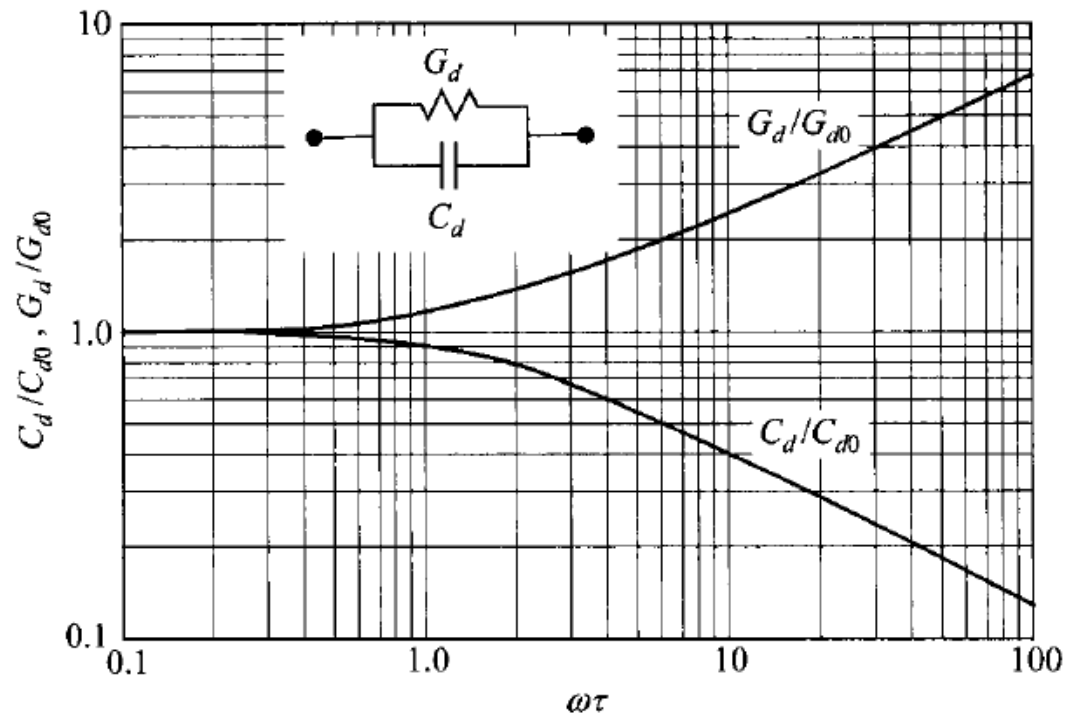


Fig. 13 Normalized diffusion conductance and diffusion capacitance versus $\omega\tau$. Inset shows the equivalent circuit of a $p-n$ junction under forward bias.

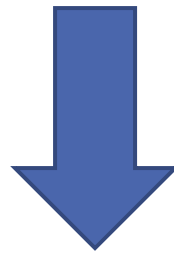
- ▶ the diffusion capacitance decreases with increasing frequency.
- ▶ For high frequencies, C_d is approximately proportional to $\omega^{-1/2}$.
- ▶ The diffusion capacitance is also proportional to the dc current
- ▶ The C_d is especially important at low frequencies and under forward-bias conditions

2.4 JUNCTION BREAKDOWN

- ▶ When a sufficiently high field is applied to a p-n junction.
- ▶ the junction *break down* and conducts a very large current.
- ▶ Breakdown occurs only in the reverse-bias regime because high voltage can be applied resulting in high field.
- ▶ There are basically three breakdown mechanisms:
 - ▶ (1) thermal instability
 - ▶ (2) tunneling
 - ▶ (3) avalanche multiplication.

2.4.1 Thermal Instability

- ▶ Breakdown due to thermal instability is responsible for the maximum dielectric strength in most insulators at room temperature.
- ▶ a major effect in semiconductors with relatively small bandgaps.
- ▶ the heat dissipation caused by the reverse current at high reverse voltage, the junction temperature increases.
- ▶ The temperature increase, in turn, increases the reverse current in comparison with its value at lower voltages.



This positive feedback is responsible for breakdown.

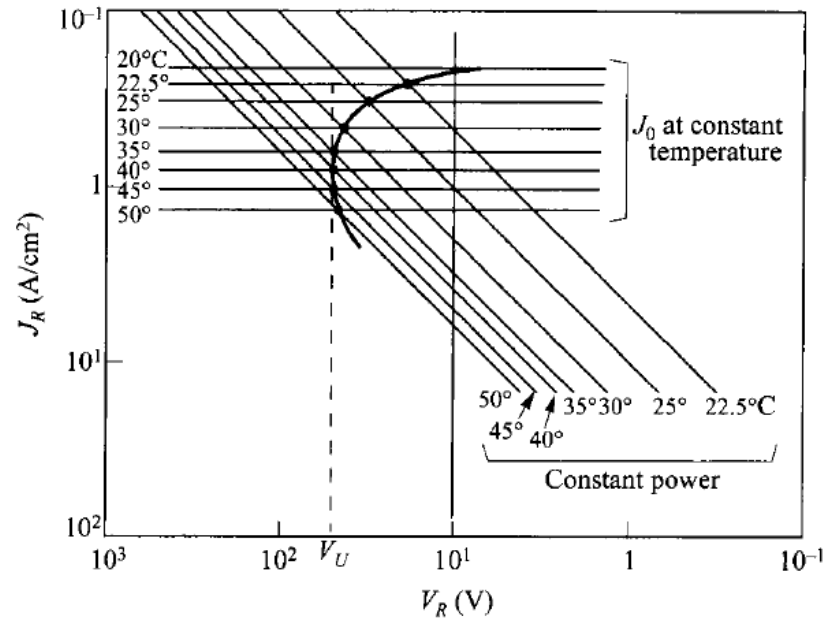


Fig. 14 Reverse current-voltage characteristics of thermal breakdown, where V_U is the turn-over voltage. (Note decreasing values of coordinates.) (After Ref. 12.)

- Each line represents the current at a constant junction temperature.
- The current with slop $T^{3+\frac{\gamma}{2}} e^{-\frac{E_g}{KT}}$
- The heat dissipation hyperbolas which are proportional to the power.
- These lines also have to satisfy the curves of constant junction temperature.
- The reverse current-voltage characteristics are obtained by the intersection points of these two sets of curves.
- The heat dissipation at high reverse voltage, the characteristics show a negative differential resistance.
- Because of the heat dissipation the diode will be destroyed unless some special measure such as a large series-limiting resistor is used.

- ▶ This effect is called thermal instability or thermal runaway.
- ▶ The voltage V_U is called the turnover voltage.
- ▶ The thermal instability is important at room temperature.
- ▶ At very low temperatures it becomes less important compared with other mechanisms.

2.4.2 Tunneling

- ▶ The junction is under a large reverse bias.
- ▶ The carriers can tunnel through a potential barrier if this barrier is sufficiently thin.
- ▶ In this particular case, the barrier has a triangular shape with the maximum height given by the energy gap.
- ▶ The tunneling current of a p-n junction:

$$J_t = \frac{\sqrt{2m^*} q^3 \mathcal{E} V_R}{4\pi^2 \hbar^2 \sqrt{E_g}} \exp\left(-\frac{4\sqrt{2m^*} E_g^{3/2}}{3q\mathcal{E}\hbar}\right).$$

- ▶ \mathcal{E} is some average field inside the junction.

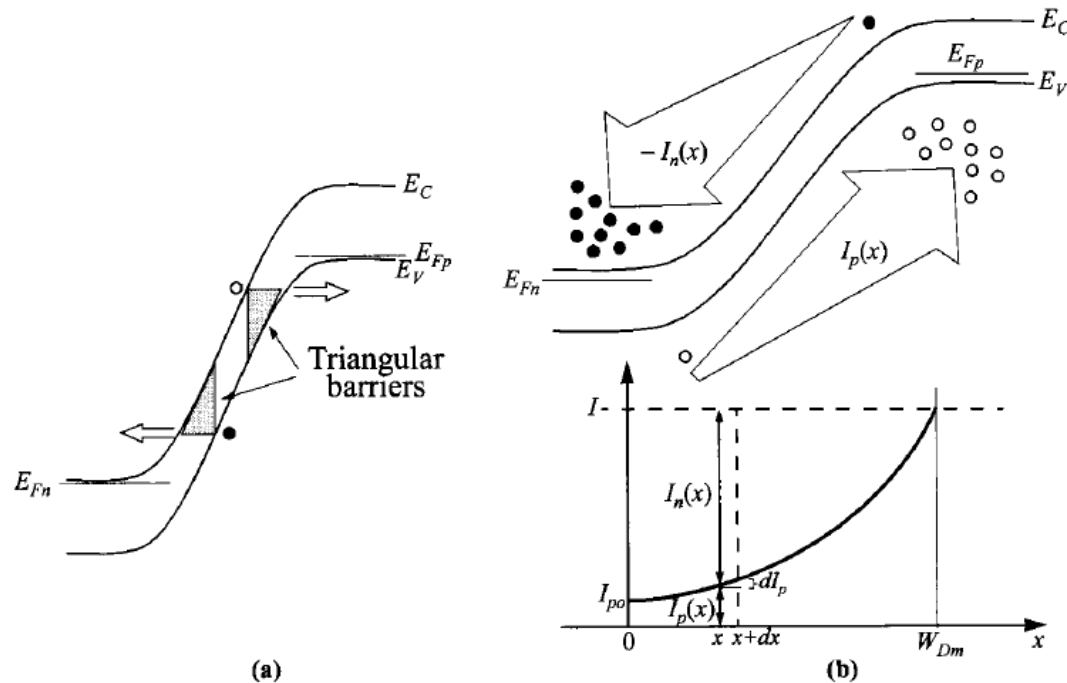


Fig. 15 Energy band diagrams showing breakdown mechanisms of (a) tunneling and (b) avalanche multiplication (example initiated by hole current I_{po}).

- When the field approaches 10^6 V/cm in Si, significant current begins to flow by means of this band-to-band tunneling process.
- To obtain such a high field, the junction must have relatively high impurity concentrations on both the p-side and n-side.
- The mechanism of breakdown for p-n junctions with breakdown voltages less than about $4E_g/q$ is due to the tunneling effect.
- For junctions with breakdown voltages in excess of $6E_g/q$, the mechanism is caused by avalanche multiplication.
- At voltages between 4 and 6 E_g/q , the breakdown is due to a mixture of both avalanche and tunneling.

- ▶ the breakdown voltage in these semiconductors due to the tunneling effect has a negative temperature coefficient.
- ▶ the breakdown voltage decreases with increasing temperature.
- ▶ This is because a given breakdown current J_t can be reached at smaller reverse voltages (or fields) at higher temperatures.
- ▶ This temperature effect is generally used to distinguish the tunneling mechanism from the avalanche mechanism.
- ▶ which has a positive temperature coefficient.
- ▶ The breakdown voltage increases with increasing temperature.

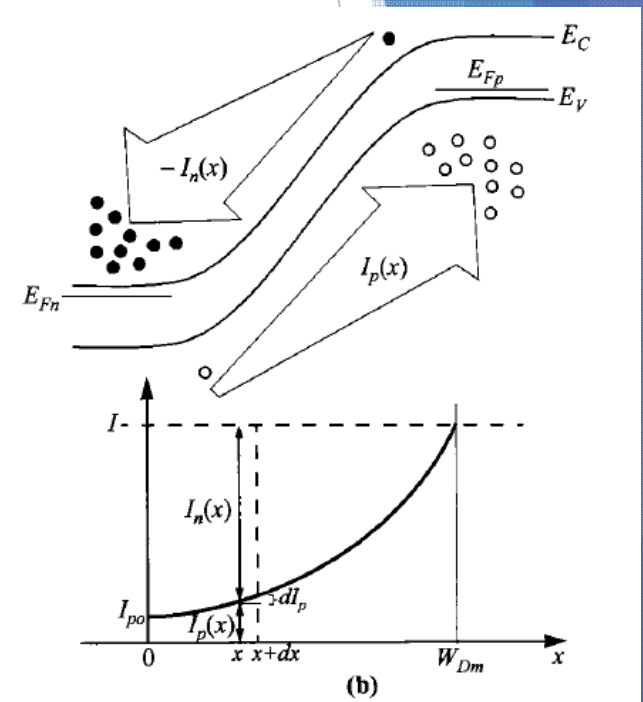
2.4.3 Avalanche Multiplication

- ▶ Avalanche multiplication, or impact ionization, is the most-important mechanism in junction breakdown.
- ▶ The avalanche breakdown voltage imposes an upper limit on the reverse bias
 - ▶ most diodes
 - ▶ on the collector voltage of bipolar transistors
 - ▶ on the drain voltages of MESFETs and MOSFETs.
- ▶ The impact ionization mechanism can be used to generate microwave power, as in IMPATT devices,
- ▶ Amplify optical signals, as in avalanche photodetectors.

Derive the basic ionization integral which determines the breakdown condition.

- ▶ Assume that a current I_{po} is incident at the left-hand side of the depletion region with width W_D
- ▶ The electric field in the depletion region is high enough that electron-hole pairs are generated by the impact ionization process.
- ▶ The hole current I_p will increase with distance through the depletion region and reach a value

$$M_p I_{po} \text{ at } x = W_{Dm}$$



- ▶ The electron current I_n will increase from $I_n(W_{Dm}) = 0$ to $I_n(0) = I - I_{p0}$
- ▶ where the total current $I (= I_p + I_n)$ is constant at steady state.

- ▶ The incremental hole current is equal to the number of electron-hole pairs generated per second in the distance dx .

$$dI_p = I_p \alpha_p dx + I_n \alpha_n dx \quad \frac{dI_p}{dx} - (\alpha_p - \alpha_n)I_p = \alpha_n I.$$

$$I_p(x) = I \left\{ \int_0^x \alpha_n \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] dx + \frac{1}{M_p} \right\} / \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right]$$

$I = I_p(W_{Dm}) = M_p I_{p0}$ where M_n is the multiplication factor of holes and is defined as

$$M_p \equiv \frac{I_p(W_{Dm})}{I_p(0)} \equiv \frac{I}{I_{p0}}.$$

$$\int_0^{W_{Dm}} (\alpha_p - \alpha_n) \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] dx = - \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] \Big|_0^{W_{Dm}} = - \exp \left(\left[- \int_0^{W_{Dm}} (\alpha_p - \alpha_n) dx' \right] + 1 \right),$$

evaluated at $x = W_{Dm}$ and be rewritten as

$$1 - \frac{1}{M_p} = \int_0^{W_{Dm}} \alpha_p \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] dx.$$

- ▶ M_p is a function of α_n in addition to α_p . The avalanche breakdown voltage is defined as the voltage where M_p approaches infinity.
- ▶ the breakdown condition is given by the ionization integral.

$$\int_0^{W_{Dm}} \alpha_p \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] dx = 1.$$

- ▶ If the avalanche process is initiated by electrons instead of holes, the ionization integral is given by:

$$\int_0^{W_{Dm}} \alpha_n \exp \left[- \int_x^{W_{Dm}} (\alpha_n - \alpha_p) dx' \right] dx = 1.$$

- ▶ The breakdown condition depends only on the behavior within the depletion region and not on the carriers

▶ The situation does not change when a mixed primary current initiates the breakdown.

▶ For semiconductors with equal ionization rates $(\alpha_n = \alpha_p = \alpha)$

$$\int_0^{W_{Dm}} \alpha dx = 1.$$

▶ From the breakdown conditions described above and the field dependence of the ionization rates, the breakdown voltage, maximum electric field, and depletion-layer width can be calculated.

▶ The electric field and potential in the depletion layer are determined from the solutions of the Poisson equation.

▶ With known boundaries we obtain the breakdown voltage

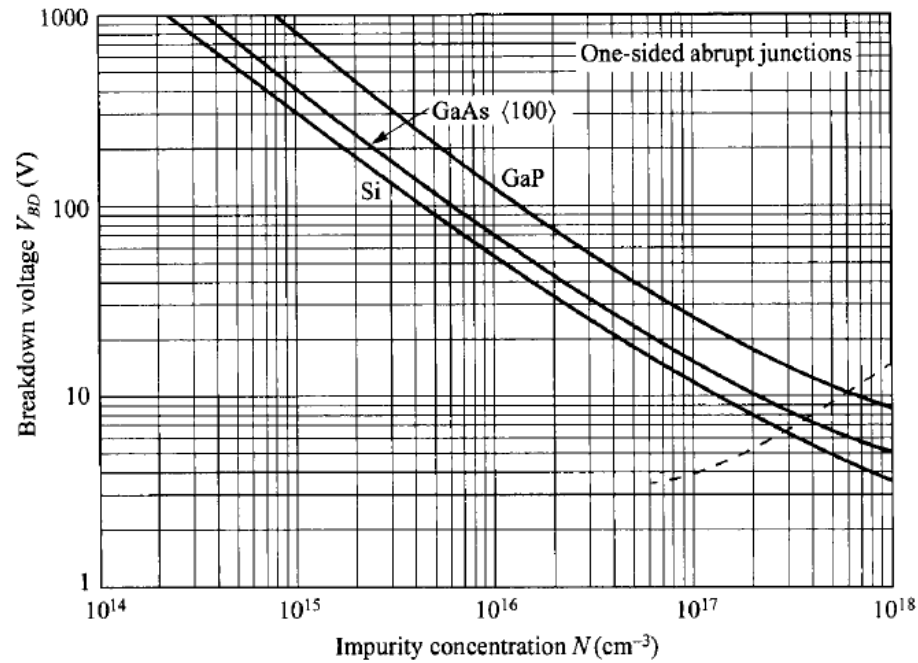
$$V_{BD} = \frac{\mathcal{E}_m W_{Dm}}{2} = \frac{\epsilon_s \mathcal{E}_m^2}{2qN}$$

for one-sided abrupt junctions

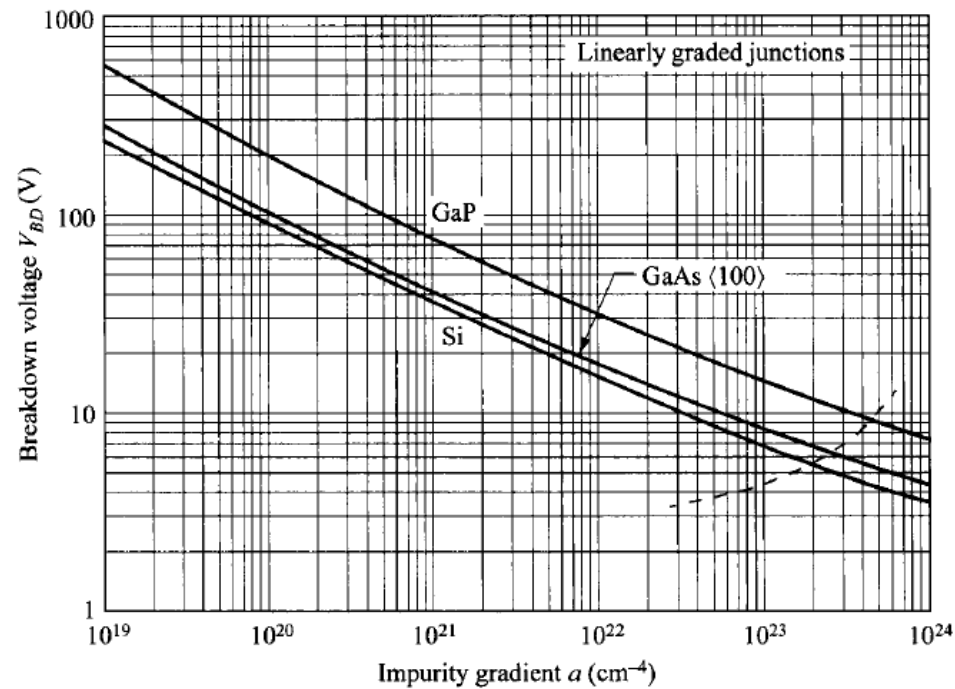
$$V_{BD} = \frac{2\mathcal{E}_m W_{Dm}}{3} = \frac{4\mathcal{E}_m^{3/2}}{3} \left(\frac{2\epsilon_s}{qa} \right)^{1/2}$$

linearly graded junctions

- N is the ionized background impurity concentration of the lightly doped side
- a the impurity gradient
- \mathcal{E}_m the maximum field



- ▶ the calculated breakdown voltage as a function of N for abrupt junctions in Si, (100)-oriented GaAs, and GaP.
- ▶ The dashed line in the figure indicates the upper limit of N for which the avalanche breakdown calculation is valid.
- ▶ This limitation is based on the criterion of $6E_g/q$
- ▶ Above the corresponding values of N , the tunneling mechanism will contribute to the breakdown process and eventually dominates.



- The calculated breakdown voltage versus the impurity gradient for linearly graded junctions.
- The dashed line indicates the upper limit of a for which the avalanche breakdown calculation is valid.
- The calculated values of the maximum field \mathcal{E}_m
- For the Si abrupt junctions, the maximum field at breakdown can be expressed as

$$\mathcal{E}_m = \frac{4 \times 10^5}{1 - (1/3) \log_{10}(N/10^{16} \text{ cm}^{-3})} \quad \text{V/cm}$$

- ▶ The maximum field at breakdown, sometimes called the *critical field*.
- ▶ The *critical field* varies very slowly with either N or a.
- ▶ we can assume that for a given semiconductor that has a fixed value.
- ▶ For abrupt junctions: $V_{BD} \propto N^{-1.0}$
- ▶ For linearly graded junctions: $V_{BD} \propto a^{-0.5}$
- ▶ for a given N or a, the breakdown voltage increases with the energy bandgap of the material.
- ▶ It should be mentioned that the critical field is only a guide line for material but not a fundamental material property
- ▶ For large distance the critical field is considered uniform.
- ▶ the total voltage (field times distance) needs to be larger than the bandgap for band-to-band carrier multiplication.
- ▶ An example is the high field but small voltage drop in an accumulation layer.

\mathcal{E}_m

- ▶ For abrupt junctions where E_g is the room-temperature bandgap in eV and N is the background doping in cm^{-3} .

$$V_{BD} \approx 60 \left(\frac{E_g}{1.1 \text{ eV}} \right)^{3/2} \left(\frac{N}{10^{16} \text{ cm}^{-3}} \right)^{-3/4} \quad \text{V}$$

- ▶ For linearly graded junctions where a is the impurity gradient in cm^{-4} .

- ▶ For diffused constant doping on one side $V_{BD} \approx 60 \left(\frac{E_g}{1.1 \text{ eV}} \right)^{6/5} \left(\frac{a}{3 \times 10^{20} \text{ cm}^{-4}} \right)^{-2/5} \quad \text{V}$:ion and a

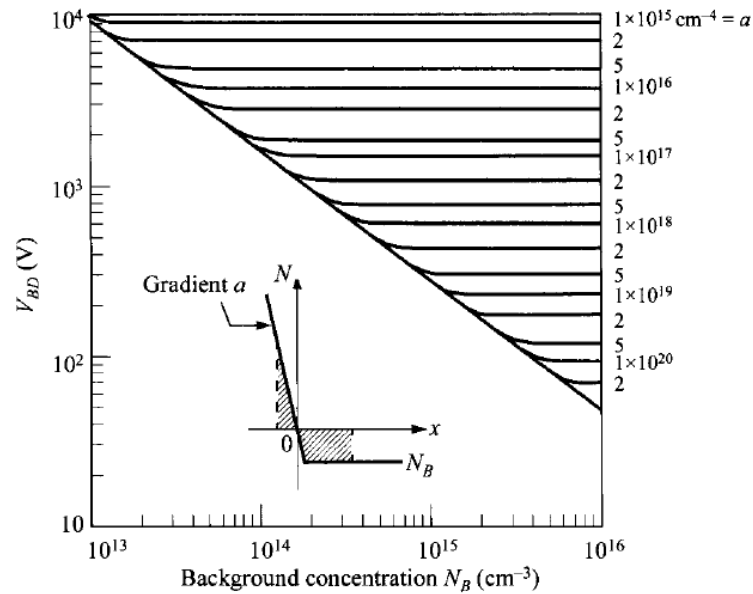


Fig. 18 Breakdown voltage for Si diffused junctions at 300 K. The inset shows the space-charge distribution. (After Ref. 18.)

- ▶ For large a , the breakdown voltage V_{BD} is determined by the abrupt junction results (bottom line)
- ▶ For small a V_{BD} will be given by the linearly graded junction results (parallel lines) and is independent of N_B .

- ▶ it is assumed that the semiconductor layer is thick enough to support the maximum depletion-layer width W_{DM} at breakdown.

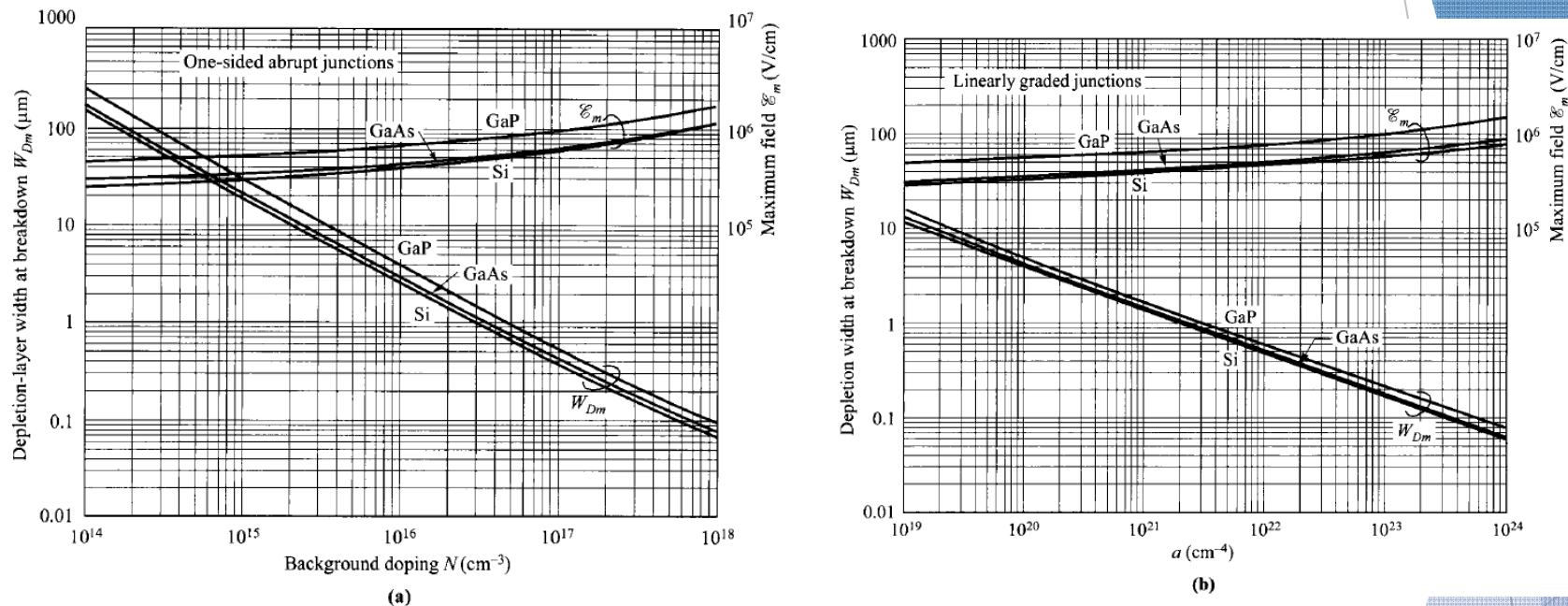


Fig. 17 Depletion-layer width and maximum field at breakdown in Si, $\langle 100 \rangle$ -oriented GaAs, and GaP for (a) one-sided abrupt junctions and (b) linearly graded junctions. (After Ref. 14.)

- ▶ The semiconductor layer W is smaller than W_{DM} the device will be punched through (i.e the depletion layer reaches the $n+$ substrate) prior to breakdown.
- ▶ As the reverse bias increases further, the depletion width cannot continue to expand and the device will break down prematurely.

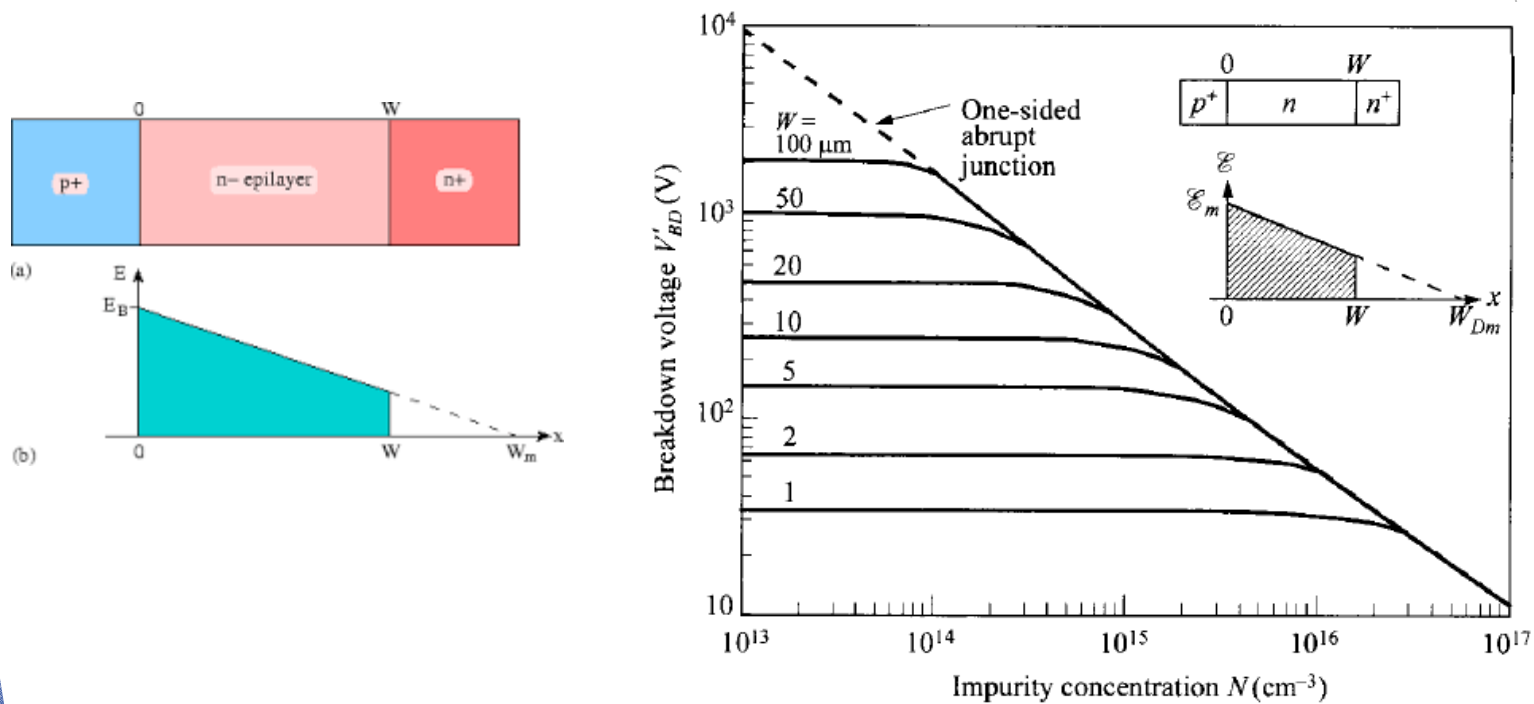


Fig. 19 Breakdown voltage for Si $p^+-\pi-n^+$ and $p^+-\nu-n^+$ junctions, where π stands for lightly doped p -type and ν for lightly doped n -type. W is the thickness of the π - or ν -region.

- ▶ For a given thickness, the breakdown voltage approaches a constant value as the doping decreases, corresponding to the punch-through of the epitaxial layer.
- ▶ The results shown so far are for avalanche breakdowns at room temperature.
- ▶ At higher temperatures the breakdown voltage increases.
- ▶ A qualitative explanation of this increase is that hot carriers passing through the depletion layer under a high field
- ▶ The carriers lose a part of their energy to optical phonons via scattering, resulting in a smaller ionization rate.
- ▶ The carriers lose more energy to the crystal lattice along a given distance at a constant field.
- ▶ The carriers must pass through a greater potential difference (or higher voltage) before they can acquire sufficient energy to generate an electron-hole pair.

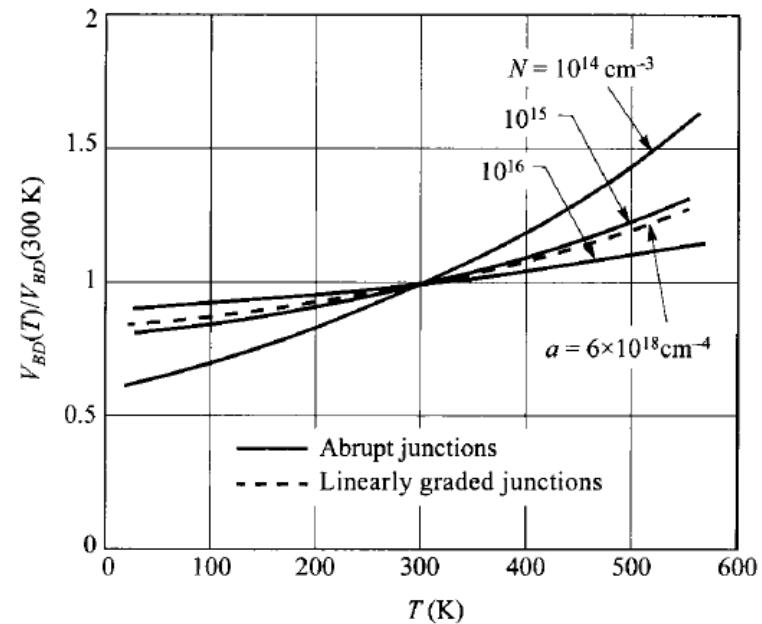


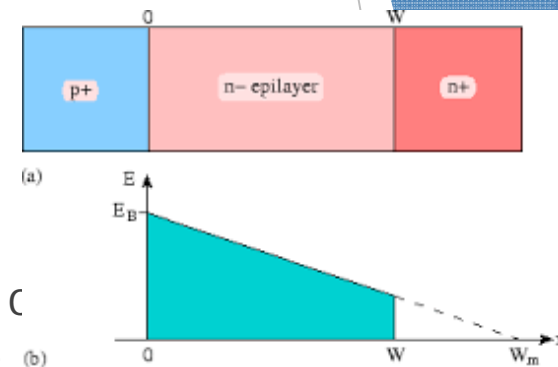
Fig. 20 Normalized avalanche breakdown voltage versus lattice temperature, in silicon. The breakdown voltage generally increases with temperature. (After Ref. 19.)

- ▶ There are substantial increases of the breakdown voltage, especially for lower doping (or small gradient) at higher temperatures

- ▶ Punched-through diode is the regular device but with extended region for distance W to reduced V_{BD} to be V_{BD}' .
- ▶ The reduced breakdown voltage V_{BD}' for the punched-through diode, compared to a regular device with V_{BD} for the same doping, can be given by:

$$\frac{V_{BD}'}{V_{BD}} = \frac{\text{Shaded area in figure insert}}{(\mathcal{E}_m W_{Dm})/2} = \left(\frac{W}{W_{Dm}}\right) \left(2 - \frac{W}{W_{Dm}}\right)$$

- ▶ Punch-through sufficiently low as in a $p^+-\pi-n^+$ or p^+-v-n^+ diode
- ▶ where π stands for a lightly doped p-type and v for a lightly doped n-type semiconductor.
- ▶ In Fig. 19 as a function of the background doping for Si one-sided abrupt junction formed on epitaxial substrates



- ▶ For a given thickness, the breakdown voltage approaches a constant value as the doping decreases, corresponding to the punch-through of the epitaxial layer.
- ▶ The results shown so far are for avalanche breakdowns at room temperature.
- ▶ At higher temperatures the breakdown voltage increases.
- ▶ A qualitative explanation of this increase is that hot carriers passing through the depletion layer under a high field
- ▶ The carriers lose a part of their energy to optical phonons via scattering, resulting in a smaller ionization rate.
- ▶ The carriers lose more energy to the crystal lattice along a given distance at a constant field.
- ▶ The carriers must pass through a greater potential difference (or higher voltage) before they can acquire sufficient energy to generate an electron-hole pair.

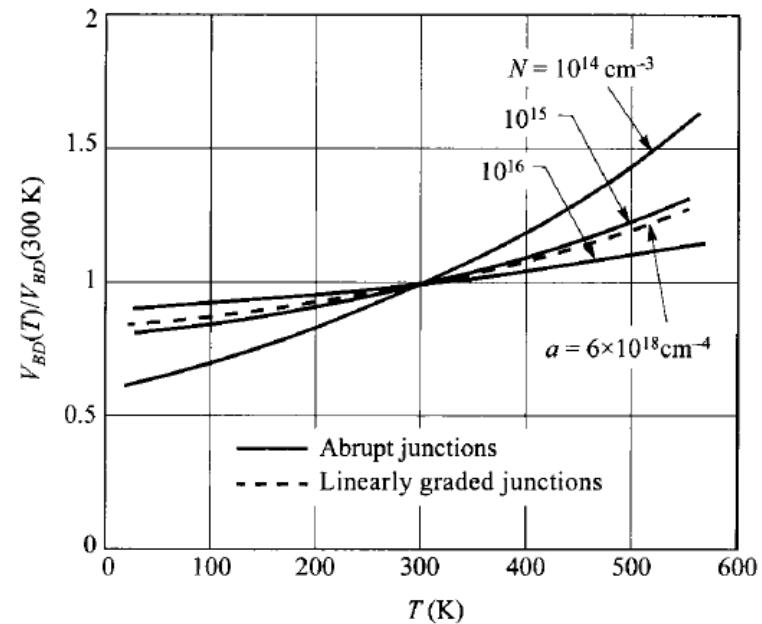


Fig. 20 Normalized avalanche breakdown voltage versus lattice temperature, in silicon. The breakdown voltage generally increases with temperature. (After Ref. 19.)

- ▶ There are substantial increases of the breakdown voltage, especially for lower doping (or small gradient) at higher temperatures

Edge Effects.

- ▶ For junctions formed by a planar process, a very important junction curvature effect at the perimeter should be considered.
- ▶ at the perimeter, the depletion region is narrower and the field is higher.

- Since the cylindrical and/or spherical regions of the junction have a higher field intensity.
- the avalanche breakdown voltage is determined by these regions.
- The potential $\Psi(r)$ and the electric field $\mathcal{E}(r)$ in a cylindrical or spherical p - n junction can be calculated from Poisson equation:

$$\frac{1}{r^n} \frac{d}{dr} [r^n \mathcal{E}(r)] = \frac{\rho(r)}{\epsilon_s}$$

- where n equals 1 for the cylindrical junction, and 2 for the spherical junction.

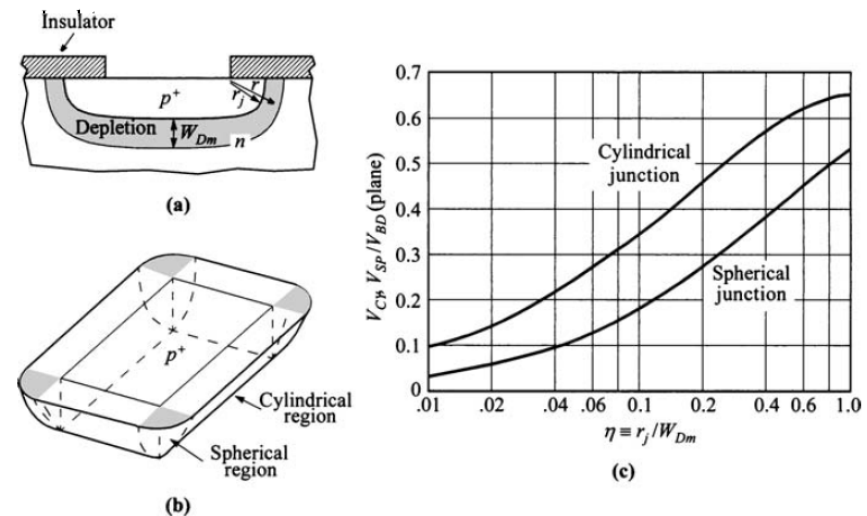


Fig. 21 (a) A planar diffusion or implantation process forms a junction curvature near the edges of the mask with r_j the radius of curvature. (b) Three-dimensional view of the junction curvature showing the spherical region at the corners. (c) Normalized breakdown voltage of cylindrical and spherical junctions as a function of the normalized radius of curvature. (After

- ▶ The solution for $\mathcal{E}(r)$ be obtained from this equation and is given by:

$$\mathcal{E}(r) = \frac{1}{\epsilon_s r^n} \int_{r_j}^r r^n \rho(r) dr + \frac{C_1}{r^n}$$

- ▶ where r_j is the radius of the junction, and the constant C_1 , must be adjusted so that the integration of the field is equal to the built-in potential.
- ▶ The calculated results for Si one-sided abrupt junctions at 300 K can be expressed by a simple equation:
- ▶ for cylindrical junctions:

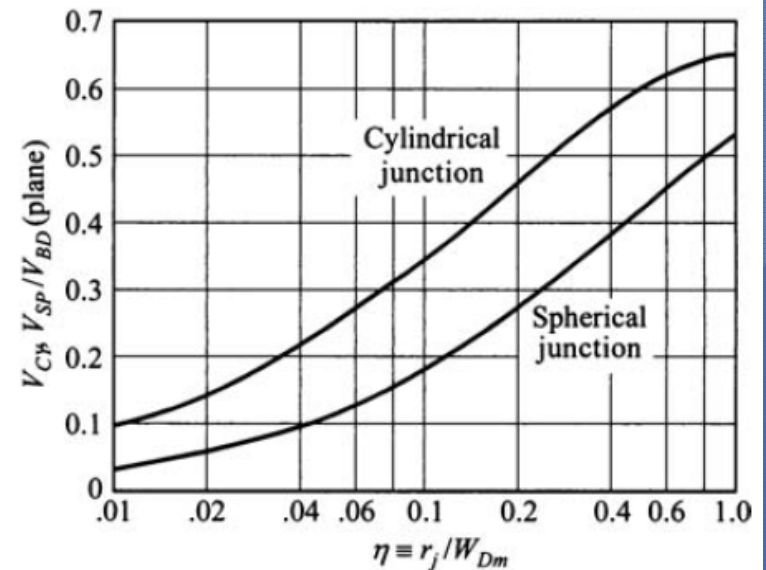
- ▶ for spherical junctions:

$$\frac{V_{CY}}{V_{BD}} = \left[\frac{1}{2}(\eta^2 + 2\eta^{6/7}) \ln(1 + 2\eta^{-8/7}) - \eta^{6/7} \right]$$

$$\frac{V_{SP}}{V_{BD}} = \left[\eta^2 + 2.14\eta^{6/7} - (\eta^3 + 3\eta^{13/7})^{2/3} \right]$$

- where V_{CY} and V_{SP} are the breakdown voltages of cylindrical and spherical junctions respectively.
- V_{BD} and $W_{D,}$ are the breakdown voltage and maximum depletion width of a plane junction having the same background doping,
- $\eta = r_j / W_{Dm}$

- ▶ illustrates the numerical results as a function of η .
- ▶ As the radius of curvature becomes smaller, so does the breakdown voltage.
- ▶ for linearly graded cylindrical or spherical junctions, the calculated results show that the breakdown voltage is relatively independent of its radius of curvature.
- ▶ Another edge effect that causes premature breakdown is due to an MOS (metal oxide semiconductor) structure over the junction at the surface.
- ▶ This configuration is often called a gated diode.



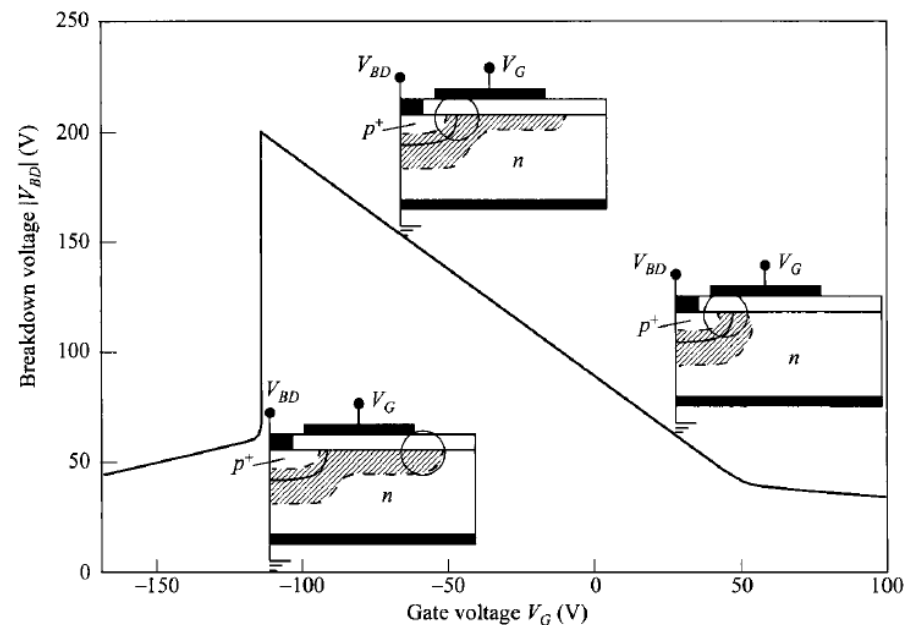


Fig. 22 Gate-voltage dependence of breakdown in a gated diode. The location of high-field breakdown shifts with gate bias. (After Ref. 22.)

- At certain gate biases, the field near the gate edge is higher than in the planar portion of the junction and breakdown changes location from the surface area of the metallurgical junction to the edge of the gate.
- This gate voltage dependence of breakdown voltage.
- At high positive gate bias on a $p^+ - n$ junction, the p^+ -surface is depleted while the n -surface is accumulated.
- Breakdown occurs near the metallurgical junction at the surface.
- As the gate bias is change to have higher negative value the location of breakdown moves toward the n -side (to the right).
- The breakdown voltage has a linear dependence on the gate bias. $V_{BD} = mV_G + \text{constant}$

- ▶ *Where $m < 1$.* At some high negative gate bias, the field directly under the gate edge is high enough to cause breakdown, and the breakdown voltage collapses.
- ▶ This gated diode breakdown phenomenon is reversible and the measurement can be repeated.
- ▶ To minimize this edge effect, the oxide thickness should be above a critical value.
- ▶ This mechanism is also responsible for the gate-induced drain leakage (GIDL) of the MOSFET.

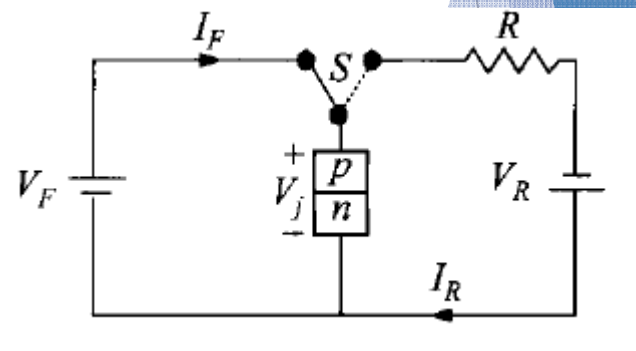
2.5 TRANSIENT BEHAVIOR AND NOISE

2.5.1 Transient Behavior:

- ▶ For switching applications the transitions from forward bias to reverse bias or vice versa must be nearly abrupt and the transient time short.
- ▶ For a p-n junction the response from forward to reverse is limited by minority carrier charge storage.

- ▶ Example:

- a simple circuit in which a forward current I_F flows in p-n junction
- at time $t = 0$, the switch S is suddenly inverted to reverse bias
- an initial reverse current of $I_R = (V_R - V_F)/R$ flows.



- ▶ The transient time is defined as the time in which the current drops to 10% of the initial reverse current I_R and is equal to the sum of t_1 and t_2

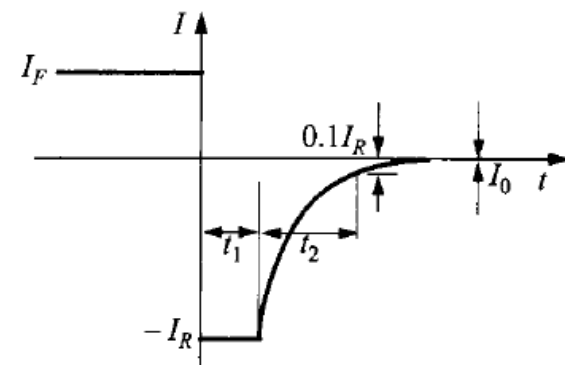
- ▶ where t_1 and t_2 are the time intervals for the constant-current phase and the decay phase.

- ▶ Consider the constant-current phase (also called storage phase) first.

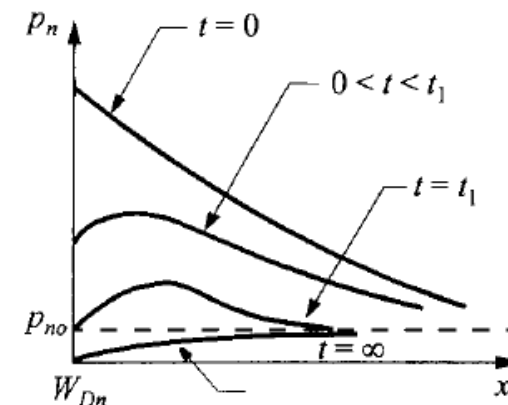
- ▶ The continuity equation has

$$\frac{\partial p_n(x, t)}{\partial t} = D_p \frac{\partial^2 p_n(x, t)}{\partial x^2} - \frac{p_n(x, t) - p_{no}}{\tau_p}$$

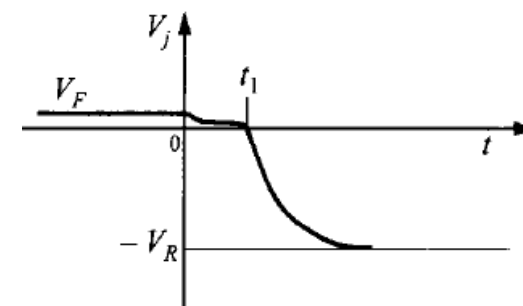
- Transient current response.
- Minority-carrier distribution outside depletion edge for various time intervals.
- Transient junction-voltage response.



(b)



(c)



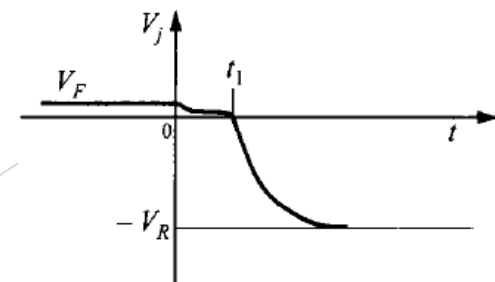
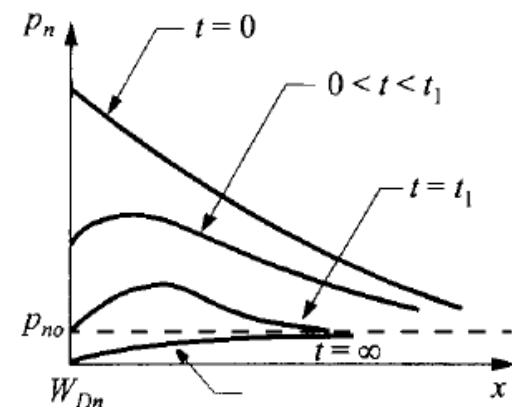
(d)

- ▶ The boundary conditions are that at $t = 0$ the initial distribution of holes is a steady state solution to the diffusion equation.
- ▶ under forward bias the voltage across the junction is given:
- ▶ The distribution of the minority carrier density p_n with time

$$V_j(t) = \frac{kT}{q} \ln \left[\frac{p_n(0, t)}{p_{no}} \right].$$

- As long as $p_n(0, t)$ is greater than p_{no} (in the interval $0 < t < t_1$), the junction voltage V_j remains of the order of kT/q
- In this time interval the reverse current is approximately constant and we have the constant-current phase.
- The solution of the time-dependent continuity equation gives t_1 by the transcendental equations

$$\operatorname{erf} \sqrt{\frac{t_1}{\tau_p}} = \frac{1}{1 + (I_R/I_F)}.$$



- ▶ The stored minority carrier charge in the lightly doped side is given by the integral:

$$Q_s = qA \int \Delta p_n dx.$$

- ▶ Integration of the continuity equation, after the current is switched to the reversed mode.

$$-I_R = \frac{dQ_s}{dt} + \frac{Q_s}{\tau_p}.$$

- ▶ With the initial condition given by the for

$$Q_s(0) = I_F \tau_p$$

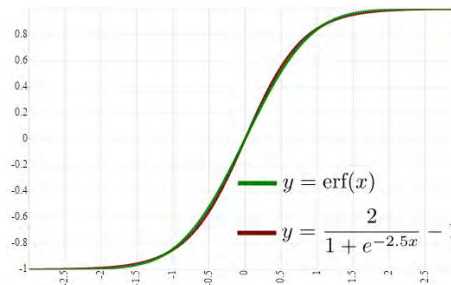
- ▶ By setting $Q_s(t) = \tau_p \left[-I_R + (I_F + I_R) \exp\left(\frac{-t}{\tau_p}\right) \right]$.

- ▶ $\frac{I_F}{I_R} = 0.1, \quad 2$

$$t_1 = \tau_p \ln\left(1 + \frac{I_F}{I_R}\right).$$

- ▶ $\frac{I_F}{I_R} = 10, \quad \text{erf} \sqrt{\frac{t_1}{\tau_p}} =$

$$\text{erf} \sqrt{\frac{t_1}{\tau_p}} =$$



$$\text{erf} \sqrt{\frac{t_1}{\tau_p}} = \frac{1}{1 + (I_R/I_F)}.$$

- ▶ After t_1 the hole density starts to decrease below its equilibrium value p_{no}
- ▶ The junction voltage tends to reach $-V_R$ and a new boundary condition now holds.
- ▶ This phase is the decay phase with the initial boundary condition $p_n(0, t_1) = p_{no}$
- ▶ The solution for t_2 is given by another transcendental equation:

- ▶ The plane junction with the length of the n-type material W much greater than the diff

- ▶ For a large I $\operatorname{erf} \sqrt{\frac{t_2}{\tau_p}} + \frac{\exp(-t_2/\tau_p)}{\sqrt{\pi t_2/\tau_p}} = 1 + 0.1 \left(\frac{I_R}{I_F} \right)$. approximated by:

- ▶ for $W \gg L_p$

$$t_1 + t_2 \approx \frac{\tau_p}{2} \left(\frac{I_R}{I_F} \right)^{-2}$$

$$t_1 + t_2 \approx \frac{W^2}{2D_p} \left(\frac{I_R}{I_F} \right)^{-2}$$

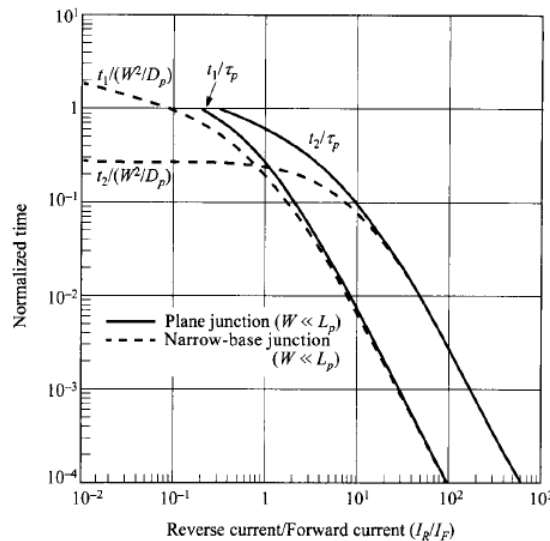


Fig. 24 Normalized time versus the ratio of reverse current to forward current. W is width of the n -region in a p^+n junction. (After Ref. 24.)

- The solid lines are for the plane junction with the length of the n -type material W much greater than the diffusion length ($W \gg L_p$).
- The dashed lines are for the narrow-base junction with ($W \ll L_p$.)

For example:

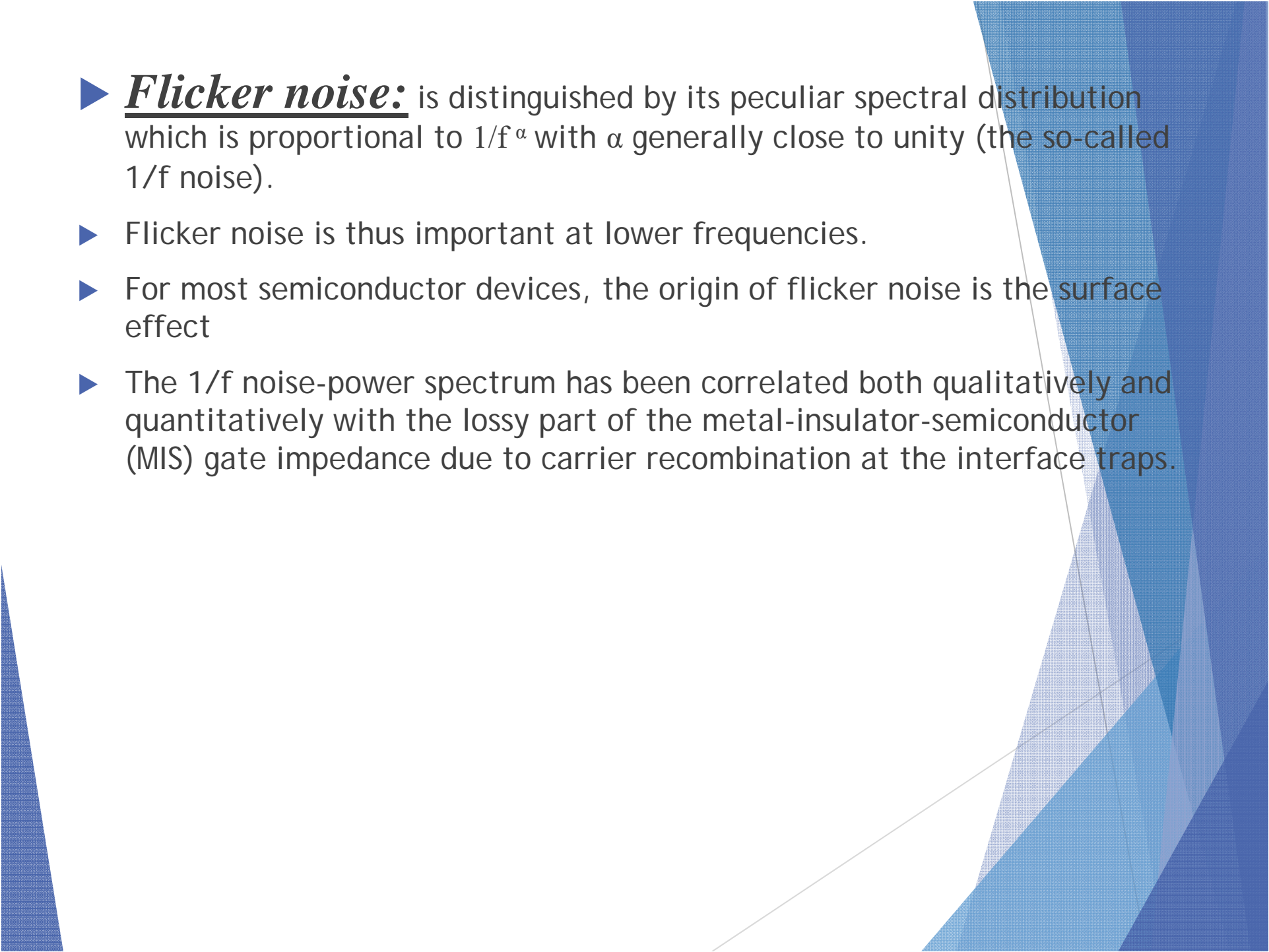
- If one switches a junction (of $W \gg L_p$) from forward 10 mA to reverse 10 mA ($I_R/I_F = 1$), the time for the constant-current phase is $0.3 \tau_p$, and that for the decay phase is about $0.6 \tau_p$
- Total transient time is then $0.9 \tau_p$.
- A fast switch requires that τ_p be small for all cases.
- The lifetime τ_p can be substantially reduced by introducing impurities with deep levels in the forbidden gap, such as gold in silicon.

2.5.2 Noise

- ▶ The term “noise” refers to spontaneous fluctuations in the current passing through, or the voltage developed across, semiconductor bulk materials or devices.
- ▶ The semiconductor devices are mainly used to amplify small signals or to measure small physical quantities
- ▶ The spontaneous fluctuations in current or voltage set a lower limit to these signals.
- ▶ The noise is generally classified into
 - 1) Thermal noise or Johnson noise,
 - 2) Flicker noise
 - 3) Shot noise.

- ▶ **Thermal noise**: occurs in any conductor or semiconductor device and is caused by the random thermal motion of the current carriers.
- ▶ It is also called white noise because its level is the same at all frequencies.
- ▶ The open circuit mean-square voltage of thermal noise is given by:
- ▶ where B is the bandwidth in Hz, and R the real part of the dynamic impedance (dV/dQ) between terminals.
- ▶ At room temperature, for a semiconductor device with 1 KΩ resistance.
- ▶ The root-mean-square voltage measured with a 1-Hz bandwidth is only about 4 nV.

$$\langle V_n^2 \rangle = 4kTBR$$

- 
- ▶ **Flicker noise:** is distinguished by its peculiar spectral distribution which is proportional to $1/f^\alpha$ with α generally close to unity (the so-called $1/f$ noise).
 - ▶ Flicker noise is thus important at lower frequencies.
 - ▶ For most semiconductor devices, the origin of flicker noise is the surface effect
 - ▶ The $1/f$ noise-power spectrum has been correlated both qualitatively and quantitatively with the lossy part of the metal-insulator-semiconductor (MIS) gate impedance due to carrier recombination at the interface traps.

- ▶ **Shot noise** is due to the discreteness of charge carriers that contribute to current flow
- ▶ it constitutes the major noise in most semiconductor devices. It is independent of frequency (white spectrum) at low and intermediate frequencies.
- ▶ At higher frequencies the shot-noise spectrum also becomes frequency-dependent.
- ▶ The mean square noise current of shot noise for a p-n junction is given by:
- ▶ where I can be forward or reverse current. For low injection the total mean-square noise current ($\langle i_n^2 \rangle = 2qB|I|$ /f noise) is the sum
- ▶ the Shockley equation we obtain:

$$\langle i_n^2 \rangle = \frac{4kTB}{R} + 2qB|I|.$$

$$\frac{1}{R} = \frac{dI}{dV} = \frac{d}{dV} \left\{ I_0 \left[\exp\left(\frac{qV}{kT}\right) - 1 \right] \right\} = \frac{qI_0}{kT} \exp\left(\frac{qV}{kT}\right).$$

- ▶ the forward-bias condition:

$$\langle i_n^2 \rangle = 4qI_0B \exp\left(\frac{qV_F}{kT}\right) + 2qI_0B \left[\exp\left(\frac{qV_F}{kT}\right) - 1 \right]$$
$$\approx 6qI_0B \exp\left(\frac{qV_F}{kT}\right) .$$

- ▶ Experimental measurements indeed confirm that the mean-square noise current is proportional to the saturation current I_0 which can be increased by irradiation

2.6 TERMINAL FUNCTIONS

- ▶ A p-n junction is a two-terminal device that can perform various terminal functions, depending upon its biasing condition as well as its doping profile and device geometry.

- ▶ **2.6.1 Rectifier:**

- ▶ A rectifier is a two-terminal device that gives a very low resistance to current flow in one direction
- ▶ In the other direction it has a very high resistance.
- ▶ The forward and reverse resistances of a rectifier can be derived from the current-voltage relationship of diode:

- ▶ where I_0 is the saturation current and generally has a value between 1 (for 2 (for recombination current)).
$$I = I_0 \left[\exp\left(\frac{qV}{\eta kT}\right) - 1 \right]$$

- ▶ The forward dc (or static) resistance R_F and small-signal (or dynamic) resistance r_F are:

$$R_F \equiv \frac{V_F}{I_F} \approx \frac{V_F}{I_0} \exp\left(\frac{-qV_F}{\eta kT}\right),$$

$$r_F \equiv \frac{dV_F}{dI_F} \approx \frac{\eta kT}{qI_F}.$$

- ▶ The reverse dc resistance R_R and small-signal resistance r_R are given by:

$$R_R \equiv \frac{V_R}{I_R} \approx \frac{V_R}{I_0},$$

- ▶ the dc rectification the ac rectification ratio r_R/r_F varies with $\frac{dV_R}{dI_R} = \frac{\eta kT}{qI_0} \exp\left(\frac{q|V_R|}{\eta kT}\right)$ $(V_R/V_F) e^{\frac{qV_F}{\eta kT}}$ while

$$e^{\frac{q|V_R|}{\eta kT}}$$

- ▶ p-n junction rectifiers generally have slow switching speeds.
- ▶ a significant time delay is necessary to obtain high impedance after switching from the forward conduction state to the reverse-blocking state.
- ▶ This time delay proportional to the minority-carrier lifetime is of little consequence in rectifying 60-Hz currents.
- ▶ For high-frequency applications, the lifetime should be sufficiently reduced to maintain rectification efficiency.
- ▶ The majority of rectifiers have
 - ▶ Power dissipation capabilities from 0.1 to 10 W
 - ▶ reverse breakdown voltages from 50 to 2500 V
 - ▶ switching times from 50 ns for low-power diodes to about 500 ns for high-power diodes.
- ▶ Circuit application:
 - ▶ It is used to transform ac signals into different special waveforms
 - ▶ clipper and clamper circuits, peak detector (demodulator).
 - ▶ a ESD (electrostatic discharge) protection device.

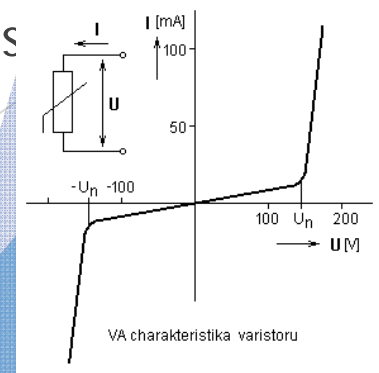
2.6.2 Zener Diode

- ▶ A Zener diode (also called voltage regulator) has a well-controlled breakdown voltage, called the Zener voltage.
- ▶ sharp breakdown characteristics in the reverse bias region
- ▶ the diode has a very high resistance; after breakdown the diode has a very small dynamic resistance.
- ▶ The terminal voltage is thus limited (or regulated) by the breakdown voltage.
- ▶ This is used to establish a fixed reference voltage.
- ▶ Most Zener diodes are made of Si, because of the low saturation current in Si diodes and the advanced Si technology.
- ▶ They are special p-n junctions with higher doping concentrations on both sides.
- ▶ for breakdown voltage V_{BD} larger than $6Eg/q$ ($= 7\text{ V}$ for Si) is the breakdown mechanism is mainly avalanche multiplication.

- ▶ the temperature coefficient of V_{BD} is positive
- ▶ for breakdown voltage V_{BD} smaller than $4E_g/q$ ($= 5\text{ V}$ for Si) the breakdown mechanism is band-to-band tunneling.
- ▶ the temperature coefficient of V_{BD} is negative.
- ▶ For $4E_g/q < V_{BD} < 6E_g/q$, the breakdown is due to a combination of these two mechanisms.
- ▶ a negative-temperature-coefficient diode in series with a positive-temperature-coefficient diode to produce a temperature-independent regulator

2.6.3 Varistor

- ▶ A varistor (variable resistor) is a two-terminal device that shows nonohmic behavior
- ▶ *An* interesting application of varistors is their use as a symmetrical fractional-voltage ($\approx 0.5 V$) limiter by connecting two diodes in parallel.
- ▶ The two-diode unit will exhibit the forward I-V characteristics in either direction.
- ▶ A varistor, being a nonlinear device, is also useful in microwave modulation, mixing, and detection (demodulation).
- ▶ Varistors based on metal-semiconductor contacts are more common due to their higher speed from the absence of minority-charge s



2.6.4 Varactor

- ▶ The term *varactor* comes from *variable reactor* and means a device whose reactance (or capacitance) can be varied in a controlled manner with a dc bias voltage.
- ▶ Varactor diodes are widely used in parametric amplification, harmonic generation, mixing, detection, and voltage-variable tuning.
- ▶ For this application, the forward bias is to be avoided because of excessive current which is undesirable for any capacitor.
- ▶ The basic capacitance-voltage relationships of abrupt and linearly graded doping distributions to a more general case. The one-dimensional Poisson equation is given as:

$$\frac{d^2 \psi_i}{dx^2} = - \frac{qN}{\epsilon_s}$$

- ▶ where N is the generalized doping distribution (negative sign for donors)

$$\frac{d^2 \psi_i}{dx^2} = -\frac{qN}{\epsilon_s} \quad N = Bx^m \quad \text{for } x \geq 0.$$

- ▶ For $m = 0$ we have $N = B$, corresponding to the uniformly doped (or one-sided abrupt junction) case.
- ▶ For $m = 1$, the doping profile corresponds to a one-sided linearly graded case.
- ▶ For $m < 0$, the device is called a "hyper-abrupt" junction.
- ▶ The hyper abrupt doping profile can be achieved by an epitaxial process or by ion implantation.

$$W_D = \left[\frac{\epsilon_s(m+2)(V_R + \psi_{bi})}{qB} \right]^{1/(m+2)},$$

$$C_D \equiv \frac{\epsilon_s}{W_D} = \left[\frac{qB\epsilon_s^{m+1}}{(m+2)(V_R + \psi_{bi})} \right]^{1/(m+2)} \propto (V_R + \psi_{bi})^{-s},$$

$$s \equiv \frac{1}{m+2}.$$

- ▶ V_R is the applied reverse voltage and ψ_{bi} is the built-in potential. Integrating the Poisson equation the boundary conditions,

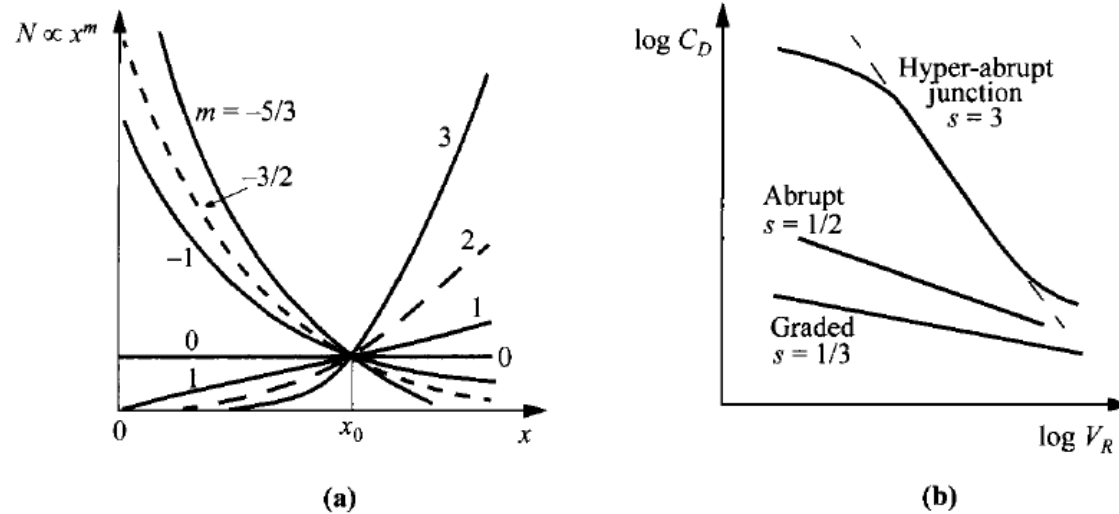


Fig. 25 (a) Various impurity distributions (normalized at x_0) for varactors. (b) Log-log plot of depletion-layer capacitance versus reverse bias. (After Refs. 29 and 30.)

$$-\frac{dC_D}{C_D} \frac{V_R}{dV_R} = -\frac{d(\log C_D)}{d(\log V_R)} = \frac{1}{m+2} = s.$$

- The larger the s , the larger will be the capacitance variation with biasing voltage.
- For linearly graded junctions, $m = 1$ and $s = 1/3$
- for abrupt junctions, $m = 0$ and $s = 1/2$;
- for hyper-abrupt junction with $m = -1, -3/2$ and $-5/3$. the value of s is 1, 2 and 3.
- The hyper-abrupt junction, as expected, has the highest sensitivity and gives rise to the largest capacitance variation.

2.6.5 Fast-Recovery Diode

- ▶ Fast-recovery diodes are designed to give ultrahigh switching speed.
- ▶ The devices can be classified into two types:
 - ▶ p-n junction diodes
 - ▶ metal-semiconductor diodes.
- ▶ The total recovery time ($t_1 + t_2$) for a p-n junction diode can be substantially reduced by introducing recombination centers.
- ▶ Although the recovery time is directly proportional to the lifetime τ
- ▶ to reduce the recovery time indefinitely by introducing an extremely large number of recombination centers N_t
- ▶ For direct bandgap semiconductors, such as GaAs, the minority-carrier lifetimes are generally much smaller than that of Si.
- ▶ This results in ultra-high-speed GaAs p-n junction diodes with recovery times of the order of 0.1 ns

- ▶ For Si the practical recovery time is in the range of 1 to 5 ns.
- ▶ The metal-semiconductor diodes (Schottky diodes) fundamentally exhibit ultrahigh- speed characteristics,
- ▶ they are majority-carrier devices and the minority-carrier storage effect is negligible.

2.6.6 Charge-Storage Diode

- ▶ In contrast to fast-recovery diodes, a charge-storage diode is designed to store a charge while conducting in the forward direction.
- ▶ switching to the reverse direction, to conduct a reverse current for a short period.
- ▶ A particularly interesting charge-storage diode is the step-recovery diode (also called the snapback diode) that conducts in the reverse direction for a short period.
- ▶ it is desirable here to reduce the decay phase or t_2 without shortening the storage phase or t_1 .
- ▶ Most charge storage diodes are made from Si with relatively long minority-carrier lifetimes ranging from 0.5 to 5 ps.
- ▶ the lifetimes are about 1000 times longer than for fast-recovery diodes.

2.7 HETEROJUNCTIONS

- ▶ The two semiconductors have the same type of conductivity, the junction is called an *isotype* heterojunction.
- ▶ The conductivity types differ, the junction is called an *anisotype* heterojunction which is a much more useful and common structure than its counterpart.
- ▶ Shockley proposed the abrupt heterojunction to be used as an efficient emitter-base injector in a bipolar transistor
- ▶ The heterojunctions have been extensively studied, and many important applications have been made,
 - ▶ the room-temperature injection laser,
 - ▶ light-emitting diode (LED),
 - ▶ photodetector
 - ▶ solar cell

2.7.1 Anisotype Heterojunction

- ▶ The slight modification of the model is needed to account for non-ideal cases such as interface traps.
- ▶ the energy-band diagrams of two isolated semiconductors of opposite types.
- ▶ The two semiconductors are assumed to have different bandgaps E_g , different permittivity ϵ_s , different work functions ϕ_m and different electron affinities χ .
- ▶ Work function and electron affinity are defined as the energy required to remove an electron from the Fermi level E_F and from the bottom of the conduction band E_C respectively.
- ▶ The difference in energy of the conduction-band edges in the two semiconductors is represented by ΔE_c , and that in the valence-band edges by ΔE_v ,

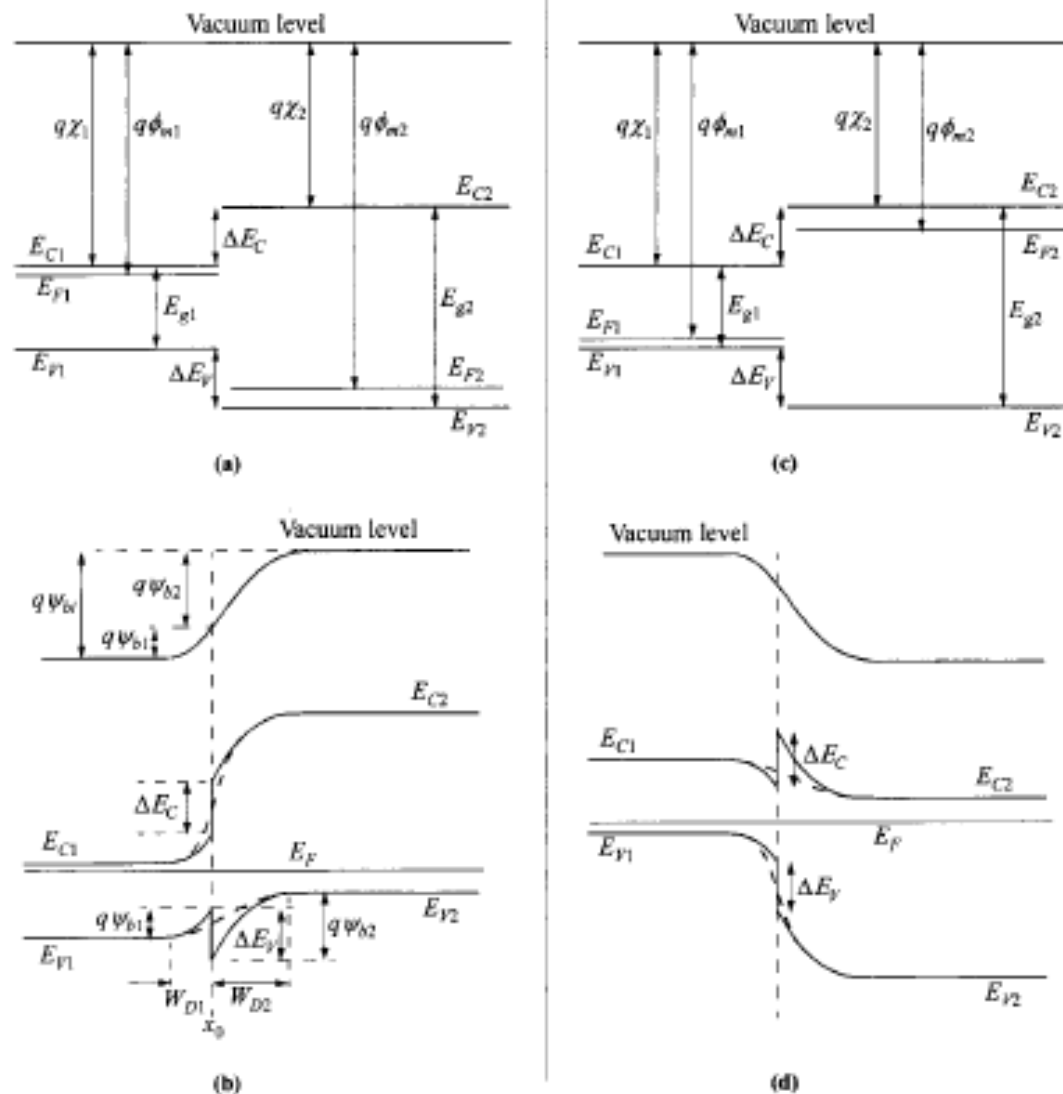


Fig. 27 Energy-band diagrams for (a) two isolated semiconductors of opposite types and different E_g (of which the smaller bandgap is *n*-type) and (b) their idealized anisotype heterojunction at thermal equilibrium. In (c) and (d), the smaller bandgap is *p*-type. In (b) and (d), the dashed lines across the junctions represent graded composition. (After Ref. 40.)

- ▶ The electron affinity rule ($\Delta E_c = q\Delta\chi$)
- ▶ When a junction is formed between these semiconductors, the energy-band profile at equilibrium for an n-p anisotype heterojunction.
- ▶ Since the Fermi level must coincide on both sides in equilibrium and the vacuum level is everywhere parallel to the band edges and is continuous.
- ▶ the discontinuity in the conduction-band edges (ΔE_c) and valence-band edges (ΔE_v) is invariant with doping in those cases where E_g and χ are not functions of doping
- ▶ The total built-in potential Ψ_{bi} is equal to the sum of the partial built-in voltages
- ▶ where Ψ_{b1} and Ψ_{b2} are the electrostatic potentials supported at equilibrium by semiconductors 1 and 2.
- ▶ it is apparent that since at $(\psi_{b1} + \psi_{b2})$
- ▶ equilibrium, $E_{F1} = E_{F2}$, the total built-in potential is given by

$$\psi_{bi} = |\phi_{m1} - \phi_{m2}|.$$

- ▶ The depletion widths and capacitance can be obtained by solving the Poisson equation for the step junction on either side of the interface.
- ▶ One boundary condition is the continuity of electric displacement, at the interface. We obtain,

$$\mathcal{D}_1 = \mathcal{D}_2 = \epsilon_{s1} \mathcal{E}_1 = \epsilon_{s2} \mathcal{E}_2$$

$$W_{D1} = \left[\frac{2N_{A2}\epsilon_{s1}\epsilon_{s2}(\psi_{bi} - V)}{qN_{D1}(\epsilon_{s1}N_{D1} + \epsilon_{s2}N_{A2})} \right]^{1/2},$$

$$W_{D2} = \left[\frac{2N_{D1}\epsilon_{s1}\epsilon_{s2}(\psi_{bi} - V)}{qN_{A2}(\epsilon_{s1}N_{D1} + \epsilon_{s2}N_{A2})} \right]^{1/2},$$

- ▶ The relative voltage supported in each semiconductor is

- ▶ where the applied voltage $V = V_1 + V_2$ is supported in the two regions

$$V = V_1 + V_2$$

$$\frac{\psi_{b1} - V_1}{\psi_{b2} - V_2} = \frac{N_{A2}\epsilon_{s2}}{N_{D1}\epsilon_{s1}}$$

- ▶ It is apparent that the foregoing expressions will reduce to the expression for the p-n junction.
- ▶ the conduction band edge E_c increases monotonically while the valence-band edge E_v goes through some peak near the junction.
- ▶ The hole current could become complicated due to the added barrier which might present a bottle-neck in thermionic emission in series with diffusion.
- ▶ The analysis can be greatly simplified by assuming a graded junction where ΔE_c and ΔE_v become smooth transitions inside the depletion region.

- ▶ The electron and hole diffusion currents are:

$$J_n = \frac{qD_{n2}n_{i2}^2}{L_{n2}N_{A2}} \left[\exp\left(\frac{qV}{kT}\right) - 1 \right],$$

$$J_p = \frac{qD_{p1}n_{i1}^2}{L_{p1}N_{D1}} \left[\exp\left(\frac{qV}{kT}\right) - 1 \right].$$

- ▶ the band offsets ΔE_c and ΔE_v are not in these equations, and also that each diffusion current component depends on the properties of the receiving side only.

$$J = J_n + J_p = \left(\frac{qD_{n2}n_{i2}^2}{L_{n2}N_{A2}} + \frac{qD_{p1}n_{i1}^2}{L_{p1}N_{D1}} \right) \left[\exp\left(\frac{qV}{kT}\right) - 1 \right].$$

- ▶ Of particular

$$\begin{aligned} \frac{J_n}{J_p} &= \frac{L_{p1}D_{n2}N_{D1}n_{i2}^2}{L_{n2}D_{p1}N_{A2}n_{i1}^2} = \frac{L_{p1}D_{n2}N_{D1}N_{C2}N_{V2} \exp(-E_{g2}/kT)}{L_{n2}D_{p1}N_{A2}N_{C1}N_{V1} \exp(-E_{g1}/kT)} \\ &\approx \frac{N_{D1}}{N_{A2}} \exp\left(\frac{-\Delta E_g}{kT}\right). \end{aligned}$$

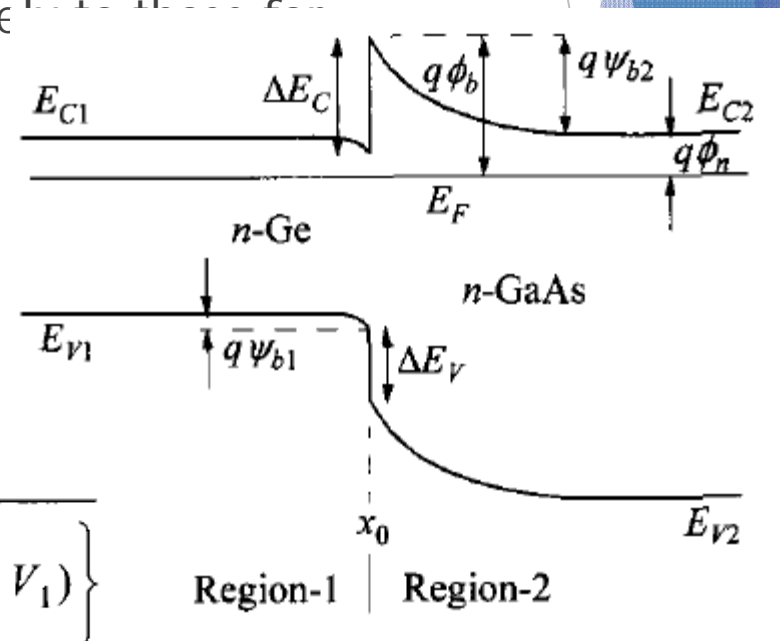
- ▶ The injection ratio depends exponentially on the bandgap difference, in addition to their doping ratio.
- ▶ This is critical in designing a bipolar transistor where the injection ratio is directly related to the current gain.
- ▶ The heterojunction bipolar transistor (HBT) uses a wide-bandgap emitter to suppress the base current.

2.7.2 Isotype Heterojunction

- ▶ The case of an isotype heterojunction is somewhat different.
- ▶ In an n-n heterojunction, since the work function of the wide-bandgap semiconductor is smaller.
- ▶ The energy bands will be bent opposite to the n-p case

- The relation between $(\Psi_{b1} - V_1)$ and $(\Psi_{b2} - V_2)$ can be found from the boundary condition of continuity of electric displacement ($D = \epsilon_s E$) at the interface.
- The electric field in region 1

$$E_1(x_0) = \sqrt{\frac{2qN_{D1}}{\epsilon_{s1}} \left\{ \frac{kT}{q} \left[\exp \frac{q(\Psi_{b1} - V_1)}{kT} - 1 \right] - (\Psi_{b1} - V_1) \right\}}$$



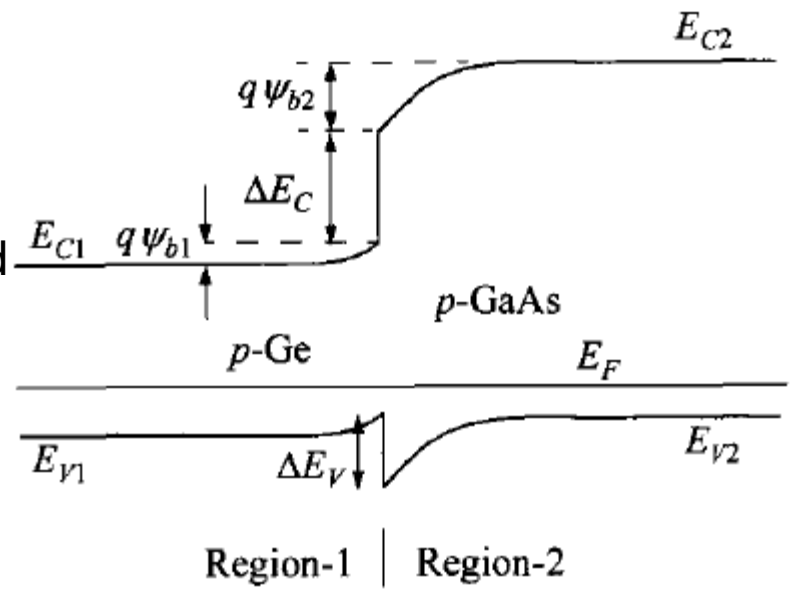
- ▶ The electric field at the interface for a depletion in Region-2 is given by:

- ▶ if the ratio is 0 $\mathcal{E}_2(x_0) = \sqrt{\frac{2qN_{D2}(\psi_{b2} - V_2)}{\epsilon_{s2}}}$.

- ▶ $\epsilon_{s1}N_{D1}/\epsilon_{s2}N_{D2}$ $\psi_{bi}(\equiv \psi_{b1} + \psi_{b2}) \gg kT/q$

- ▶ where V is the total $\exp\left[\frac{q(\psi_{b1} - V_1)}{kT}\right] \approx \frac{q}{kT}(\psi_{bi} - V)$ $(V_1 + V_2)$

- The idealized equilibrium energy-band diagram for p-p heterojunctions.
- The conduction mechanism is governed by thermionic emission of majority carriers, electrons in this case



$$J = qN_{D2} \sqrt{\frac{kT}{2\pi m_2^*}} \exp\left(\frac{-q\psi_{b2}}{kT}\right) \left[\exp\left(\frac{qV_2}{kT}\right) - \exp\left(\frac{-qV_1}{kT}\right) \right]$$

- ▶ The current-voltage relationship:

$$J = \frac{q^2 N_{D2} \psi_{bi}}{\sqrt{2\pi m_2^* kT}} \exp\left(\frac{-q \psi_{bi}}{kT}\right) \left(1 - \frac{V}{\psi_{bi}}\right) \left[\exp\left(\frac{qV}{kT}\right) - 1\right].$$

- ▶ Since the current is thermionic emission as in a metal-semiconductor contact, the pre exponential factor is often expressed in terms of the effective Richardson constant A^* and the barrier height ϕ_b .
- ▶ The appropriate expression for N_{D2} , the current equation above becomes:

$$J = \frac{q \psi_{bi} A^* T}{k} \left(1 - \frac{V}{\psi_{bi}}\right) \exp\left(\frac{-q \psi_{bi}}{kT}\right) \exp\left(\frac{-q \phi_b}{kT}\right) \left[\exp\left(\frac{qV}{kT}\right) - 1\right] \quad \text{or}$$

- ▶ This expression for a metal-semiconductor contact

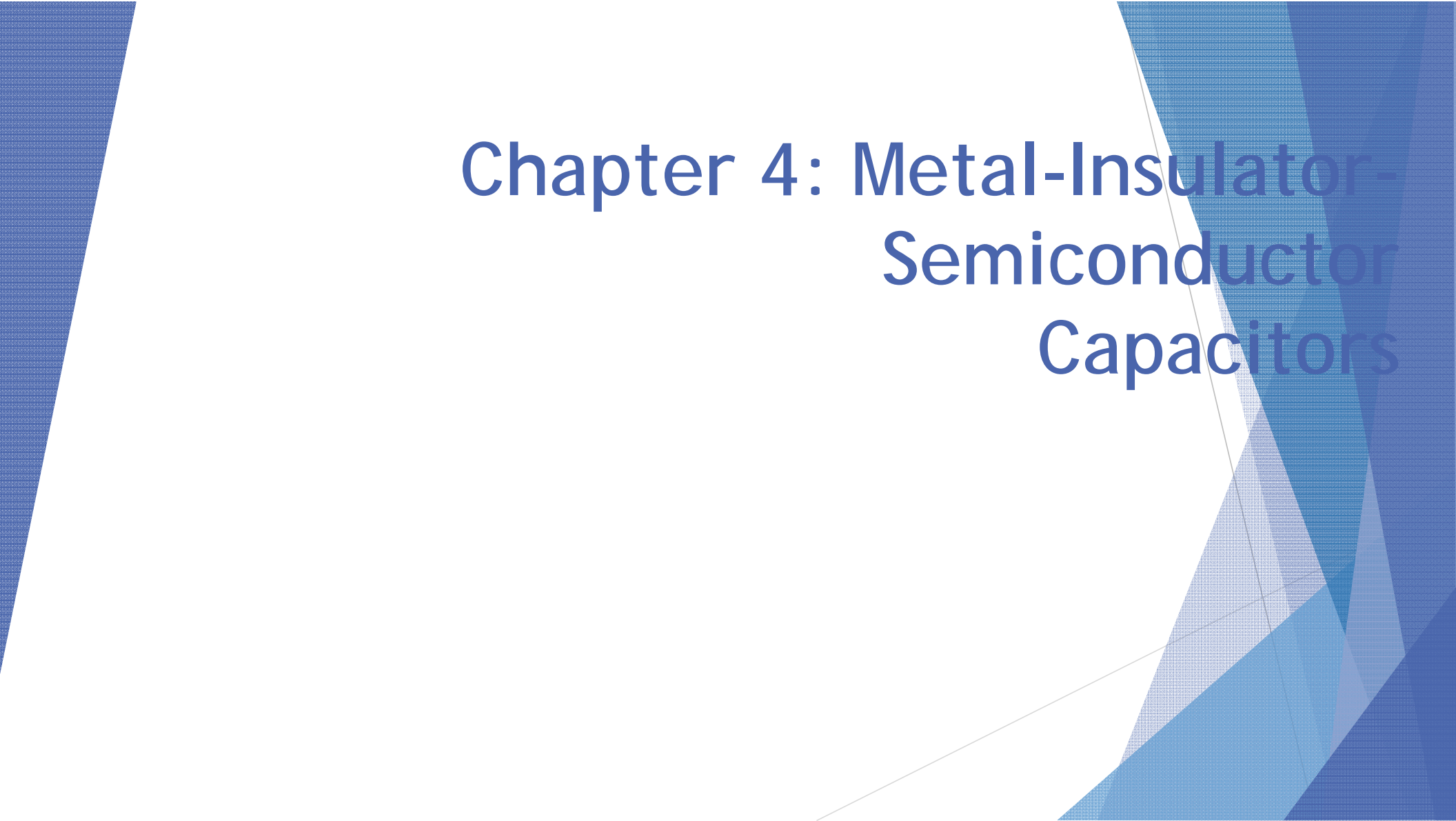
$$= J_0 \left[\exp\left(\frac{qV}{kT}\right) - 1\right].$$

- ▶ The value of J_0 is different from $A^* T^2 \exp(-q \phi_B / kT)$ is temperature dependence.

from $A^* T^2 \exp(-q \phi_B / kT)$

- ▶ The reverse current never saturates but increases linearly with voltage at large - V
- ▶ In the forward direction, the dependence of J on V can be approximated by an exponential function

$$J \propto \exp(qV/\eta kT).$$



Chapter 4: Metal-Insulator- Semiconductor Capacitors

Metal-Insulator-Semiconductor Capacitors

- ▶ 4.1 INTRODUCTION
- ▶ 4.2 IDEAL MIS CAPACITOR
- ▶ 4.3 SILICON MOS CAPACITOR

4.1 INTRODUCTION

- ▶ The metal-insulator-semiconductor (MIS) capacitor is the most useful device in the study of semiconductor surfaces.
- ▶ In this chapter we are concerned primarily with the metal-oxide-silicon (MOS) system.
- ▶ This system has been extensively studied because it is directly related to most silicon planar devices and integrated circuits.
- ▶ The MIS structure was first proposed as a voltage-controlled varistor (variable capacitor) in 1959 by Moll' and by Pfann and Garrett
- ▶ Its characteristics were then analyzed by Frank P and Lindner
- ▶ The first successful MIS structure was made of SiO_2 grown thermally on silicon surface by Ligenza and Spitzer in 1960.

4.2 IDEAL MIS CAPACITOR

- ▶ The metal-insulator-semiconductor (MIS) structure is shown, where d is the thickness of the insulator and V is the applied voltage.

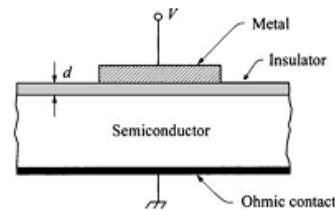


Fig. 1 Metal-insulator-semiconductor (MIS) capacitor, in its simplest form.

- ▶ The energy-band diagram of an ideal MIS structure without bias is shown, for both n-type and p-type semiconductors.

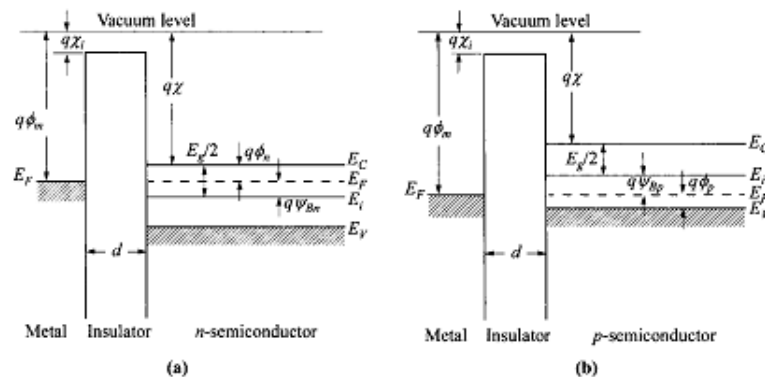


Fig. 2 Energy-band diagrams of ideal MIS capacitors at equilibrium ($V = 0$). (a) n -type semiconductor. (b) p -type semiconductor.

► An ideal MIS capacitor is defined as follows:

- (1) The only charges that can exist in the structure under any biasing conditions are those in the semiconductor and those, with an equal but opposite sign, on the metal surface adjacent to the insulator that there is no interface trap nor any kind of oxide charge;
- (2) There is no carrier transport through the insulator under dc biasing conditions or the resistivity of the insulator is infinite.

► for simplicity we assume the metal is chosen such that the difference between the metal work function φ_m and the semiconductor work function is zero, or $\varphi_{ms} = 0$.

$$\varphi_{ms} = \varphi_m - \left(\chi + \frac{E_g}{2q} - \psi_{Bn} \right) = \phi_m - (\chi + \phi_n) = 0 \text{ for } n\text{-type}$$

$$\varphi_{ms} = \varphi_m - \left(\chi + \frac{E_g}{2q} - \psi_{Bp} \right) = \phi_m - \left(\chi + \frac{E_g}{q} - \phi_p \right) = 0 \text{ for } p\text{-type}$$

- ▶ The band is flat (flat-band condition) when there is no applied voltage.
- ▶ When an ideal MIS capacitor is biased with positive or negative voltages, basically three cases may exist at the semiconductor surface.

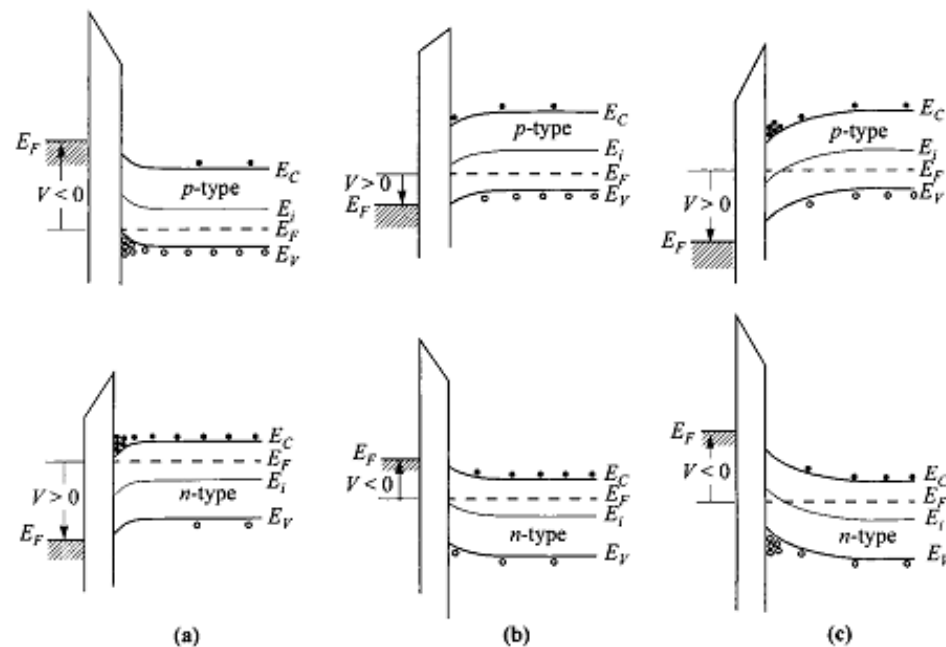
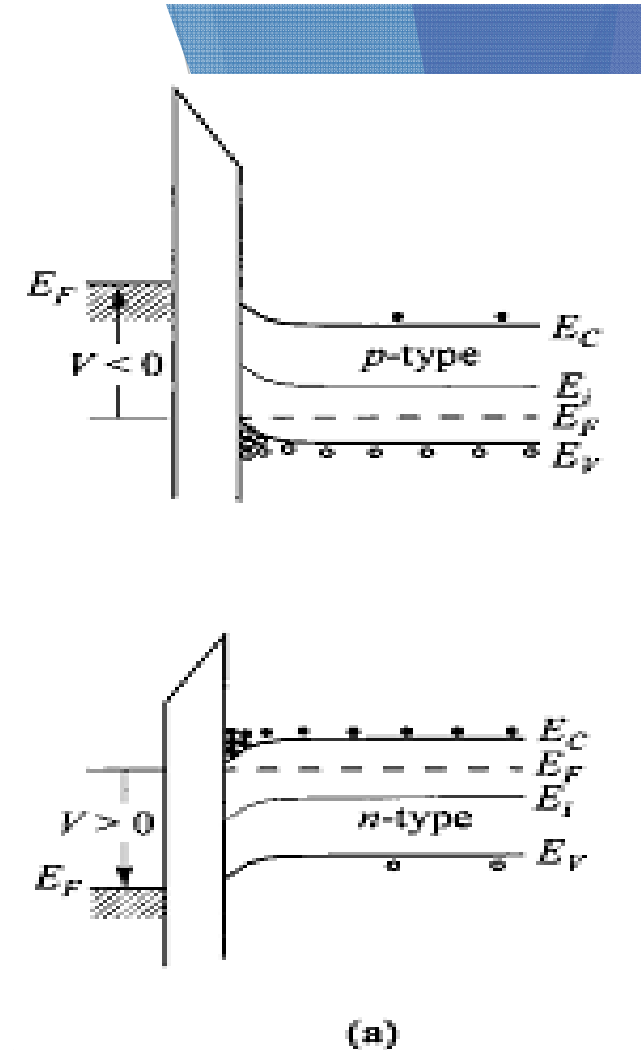


Fig. 3 Energy-band diagrams for ideal MIS capacitors under different bias, for the conditions of: (a) accumulation, (b) depletion, and (c) inversion. Top/bottom figures are for *p*-type/*n*-type semiconductor substrates.

- ▶ The ideal MIS capacitor theory to be considered in this section serves as a foundation for understanding practical MIS structures and to exploring the physics of semiconductor surfaces.
- ▶ When a negative voltage ($V < 0$, p-type) is applied to the metal plate.
- ▶ The valence-band edge E_v bends upward near the surface and is closer to the Fermi level.
- ▶ For an ideal MIS capacitor, no current flows in the structure ($dE_F/dx = 0$)
- ▶ The Fermi level remains flat in the semiconductor.
- ▶ The accumulation is for that the carrier density depends exponentially on the energy difference ($E_F - E_v$), this band bending causes an accumulation of majority carriers (holes) near the semiconductor surface.

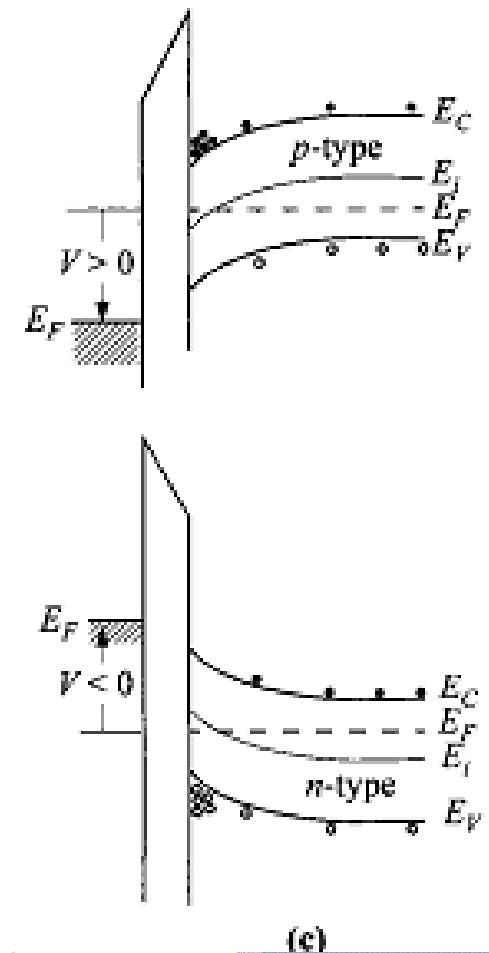


▶ The depletion case.

- ▶ When a small positive voltage ($V > 0$, p-type) is applied, the bands bend downward, and the majority carriers are depleted.

▶ The inversion case:

- ▶ When a larger positive voltage is applied, the bands bend even more downward so that the intrinsic level E_i at the surface crosses over the Fermi level E_F
- ▶ At this point the number of electrons (minority carriers) at the surface is larger than that of the holes.
- ▶ the surface is thus inverted .



4.2.1 Surface Space-Charge Region

- ▶ we derive the relations between the surface potential, space charge, and electric field.
- ▶ The potential $\psi(x)$ is defined as the potential $E_i(x)/q$ with respect to the bulk of the semiconductor.

$$\psi_p(x) \equiv -\frac{[E_i(x) - E_i(\infty)]}{q}$$

- ▶ At the semiconductor surface $\psi(0) = \psi_s$ is called the surface potential.

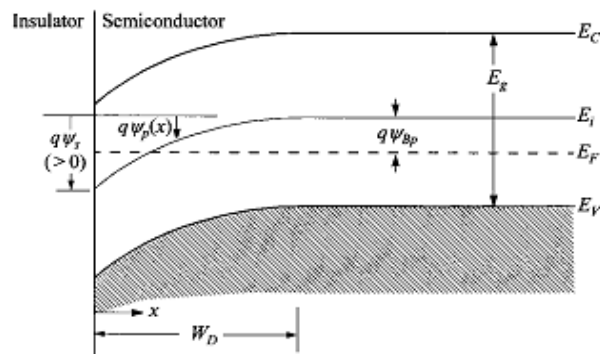


Fig. 4 Energy-band diagram at the surface of a *p*-type semiconductor. The potential energy $q\psi_p$ is measured with respect to the intrinsic Fermi level E_i in the bulk. The surface potential ψ_s is positive as shown. Accumulation occurs when $\psi_s < 0$. Depletion occurs when $\psi_{Bp} > \psi_s > 0$. Inversion occurs when $\psi_s > \psi_{Bp}$.

- ▶ The electron and hole concentrations as a function of ψ_s are given by the following relations:

$$n_p(x) = n_{p0} e^{\left(\frac{q\psi_p}{KT}\right)} = n_{p0} e^{(\beta\psi_p)}$$

$$\beta \equiv \frac{q}{KT}$$

$$p_p(x) = p_{p0} e^{\left(\frac{-q\psi_p}{KT}\right)} = p_{p0} e^{(-\beta\psi_p)}$$

where ψ_p is positive when the band is bent downward n_{p0} and p_{p0} are the equilibrium densities of electrons and holes, respectively.

At the surface the densities are:

$$n_p(0) = n_{p0} e^{(\beta\psi_s)}$$

$$p_p(0) = p_{p0} e^{(-\beta\psi_s)}$$

$\psi_s < 0$	Accumulation of holes (bands bending upward).
$\psi_s = 0$	Flat-band condition.
$\psi_{Bp} > \psi_s > 0$	Depletion of holes (bands bending downward).
$\psi_s = \psi_{Bp}$	Fermi-level at midgap, $E_F = E_i(0)$, $n_p(0) = p_p(0) = n_i$.
$2\psi_{Bp} > \psi_s > \psi_{Bp}$	Weak inversion [electron enhancement, $n_p(0) > p_p(0)$].
$\psi_s > 2\psi_{Bp}$	Strong inversion [$n_p(0) > p_{p0}$ or N_A].

- ▶ The potential $\psi_p(x)$ as a function of distance can be obtained by using the one dimensional Poisson equation:

$$\frac{d^2\psi_p}{dx^2} = -\frac{\rho(x)}{\epsilon_s}$$

$$\begin{aligned}\frac{d^2\psi_p}{dx^2} &= -\frac{q}{\epsilon_s}(n_{p0} - p_{p0} + p_p - n_p) \\ &= -\frac{q}{\epsilon_s}\{p_{p0}[\exp(-\beta\psi_p) - 1] - n_{p0}[\exp(\beta\psi_p) - 1]\}.\end{aligned}$$

where $\rho(x)$ is the total space-charge density given by:

$$\rho(x) = q(N_D^+ - N_A^- + p_p - n_p)$$

In the bulk of the semiconductor, far from the surface, charge neutrality must exist.

Therefore at $\psi_p(\infty) = 0$, we have $\rho(x) = 0$.

$$(N_D^+ - N_A^- = -p_{p0} + n_{p0})$$

The resultant Poisson equation to be solved within the depletion region is therefore

$$\frac{d^2\psi_p}{dx^2} = -\frac{q(-p_{p0} + n_{p0} + p_p - n_p)}{\epsilon_s} = -\frac{q(p_{p0}[e^{(-\beta\psi_p)} - 1] - n_{p0}[e^{(\beta\psi_p)} - 1])}{\epsilon_s}$$

▶
$$\int_0^{\psi_p} \frac{d\psi_p}{dx} d\left(\frac{d\psi_p}{dx}\right) = \int_0^{\psi_p} -\frac{q(p_{p0}[e^{(-\beta\psi_p)}-1]-n_{p0}[e^{(\beta\psi_p)}-1])}{\epsilon_s} d\psi_p$$

▶ gives the relation between the electric field ($E \equiv -d\psi_p/dx$) and the potential ψ_p :

$$E^2 = \left(\frac{2KT}{q}\right)^2 \left(\frac{qp_{p0}\beta}{2\epsilon_s}\right) [e^{(-\beta\psi_p)} + \beta\psi_p - 1] + \frac{n_{p0}}{p_{p0}} [e^{(-\beta\psi_p)} - \beta\psi_p - 1]$$

We shall use the following abbreviations:

$$L_D \equiv \sqrt{\frac{KT\epsilon_s}{p_{p0}q^2}} \equiv \sqrt{\frac{\epsilon_s}{qp_{p0}\beta}}$$

$$F\left(\beta\psi_p, \frac{n_{p0}}{p_{p0}}\right) \equiv \sqrt{[e^{(-\beta\psi_p)} + \beta\psi_p - 1] + \frac{n_{p0}}{p_{p0}} [e^{(-\beta\psi_p)} - \beta\psi_p - 1]}$$

where L_D is the extrinsic Debye length for holes.

- ▶ The electric field is given by:

$$E(x) = \pm \frac{\sqrt{2KT}}{qL_D} F\left(\beta\psi_p, \frac{n_{po}}{p_{po}}\right)$$

To determine the electric field at the surface E_s we let $\psi_p = \psi_s$:

$$E_s = \pm \frac{\sqrt{2KT}}{qL_D} F\left(\beta\psi_s, \frac{n_{po}}{p_{po}}\right)$$

we can deduce the total space charge per unit area by applying Gauss' law:

$$Q_s = -\epsilon E_s = \pm \frac{\sqrt{2KT}}{qL_D} F\left(\beta\psi_s, \frac{n_{po}}{p_{po}}\right)$$

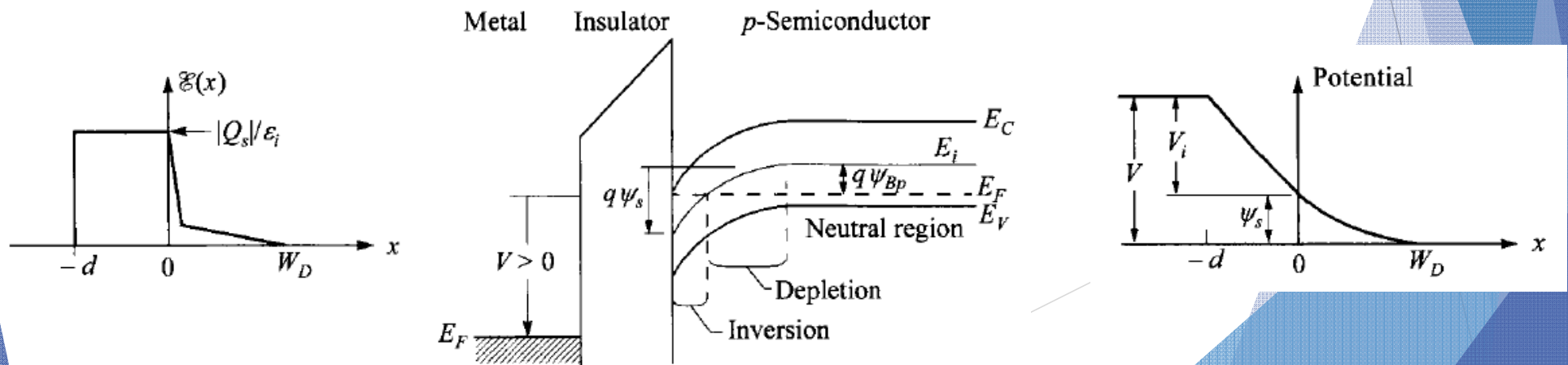
$$\psi_s(\text{strong inversion}) \approx 2\psi_{Bp} \approx \frac{2kT}{q} \ln\left(\frac{N_A}{n_i}\right).$$

4.2.2 Ideal MIS Capacitance Curves

- ▶ For charge neutrality of the system, it is required that the charge on metal surface is equal to Q_s

$$Q_M = -(Q_n + qN_A W_D) = -Q_s$$

- ▶ where Q_M is charges per unit area on the metal, Q_n is the electrons per unit area near the surface
- ▶ the inversion region, $qN_A W_D$ is the ionized acceptors per unit area in the space-charge region with depletion width
- ▶ Q_s is the total charges per unit area in the semiconductor.



- ▶ In the absence of any work-function difference, the applied voltage will partly appear across the insulator and partly across the semiconductor.

$$V = V_i + \psi_s$$

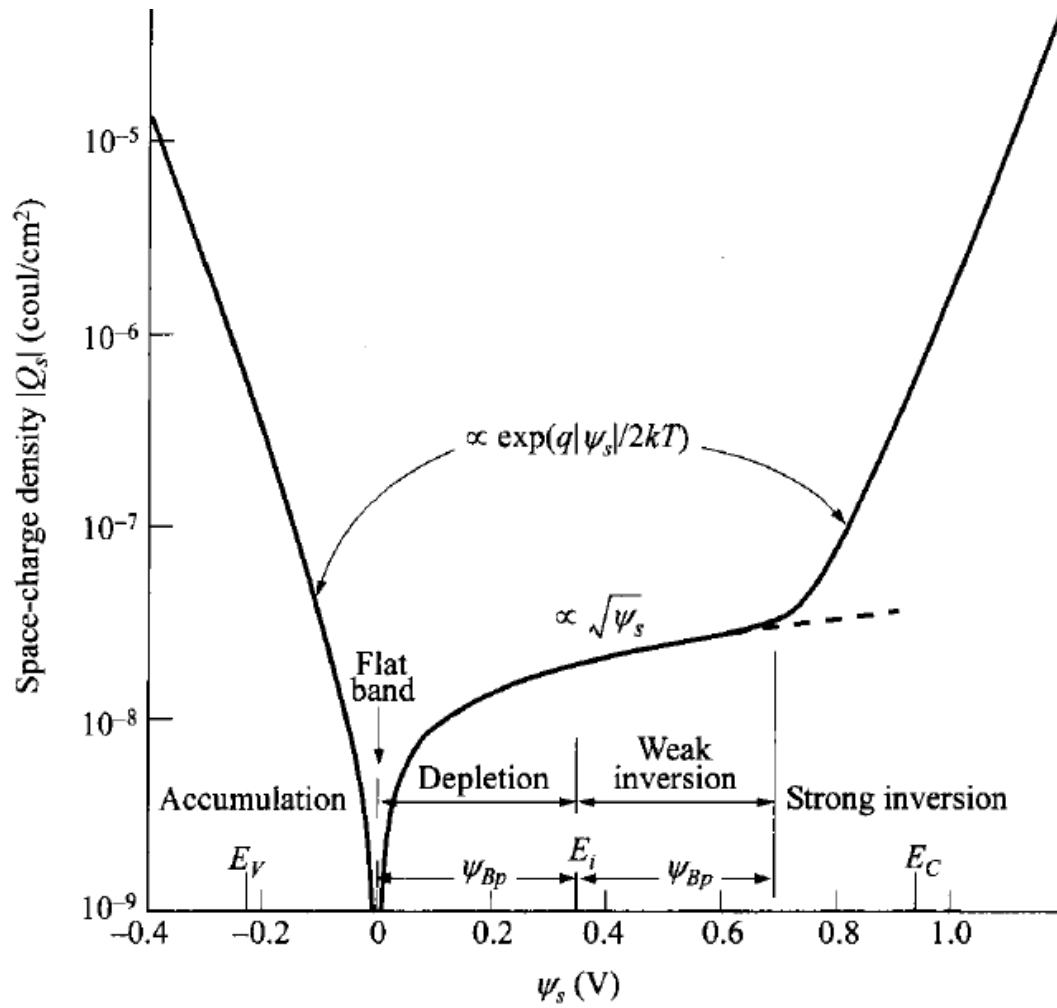
- ▶ where V_i is the potential across the insulator

$$V_i = E_i d = \frac{|Q_s|d}{\epsilon_i} = \frac{|Q_s|}{C_i}$$

- ▶ The total capacitance C of the system is a series combination of the insulator capacitance and the semiconductor depletion-layer capacitance C_D .

- ▶
$$C = \frac{C_i C_D}{C_i + C_D}$$

- ▶ For a given insulator thickness d the value of C_i is constant and corresponds to the maximum capacitance of the system.
- ▶ The semiconductor capacitance C_D not only depends on the bias (ψ_s), it is also a function of the measurement frequency.



Capacitance and frequency

- ▶ **Low-Frequency Capacitance:** The capacitance of the semiconductor depletion layer is obtained by differentiating the total static charge in the semiconductor side with respect to the semiconductor surface potential

$$C_D \equiv \frac{dQ_s}{d\psi_s} = \frac{\epsilon_s}{\sqrt{2LD}} \frac{1 - e^{(-\beta\psi_s)} + \left(\frac{n_{p0}}{p_{p0}}\right)[e^{(\beta\psi_s)} - 1]}{F(\beta\psi_s, \frac{n_{p0}}{p_{p0}})}$$

- ▶ In describing this low-frequency curve we begin at the left side (negative voltage)

- ▶ where we have an accumulation of holes and therefore a high differential capacitance of the semiconductor.
- ▶ As a result the total capacitance is close to the insulator capacitance.
- ▶ As the negative voltage is reduced to zero, we have the flat-band condition.
- ▶ Since the function F approaches zero, C_D has to be obtained:

$$C_{D(\text{flat} - \text{band})} = \frac{\epsilon_s}{L_D}$$

- ▶ The total capacitance at flat-band condition:

$$C_{FB}(\psi_s = 0) = \frac{\epsilon_i \epsilon_s}{\epsilon_s d + \epsilon_i L_D} = \frac{\epsilon_i \epsilon_s}{\epsilon_s d + \epsilon_i \sqrt{KT \epsilon_s / N_A q^2}}$$

- ▶ where ϵ_i and ϵ_s are the permittivities of the insulator and the semiconductor respectively and L_D is the extrinsic Debye length.

- ▶ It can be shown that under depletion and weak inversion conditions, i.e. $2\psi_{Bp} > \psi_s > kT/q$, the function $F = \sqrt{\beta\psi_s}$
- ▶ the space-charge density can be reduced to:

$$Q_s = \sqrt{2\epsilon_s q p_{p0} \psi_s} = qWDNA$$

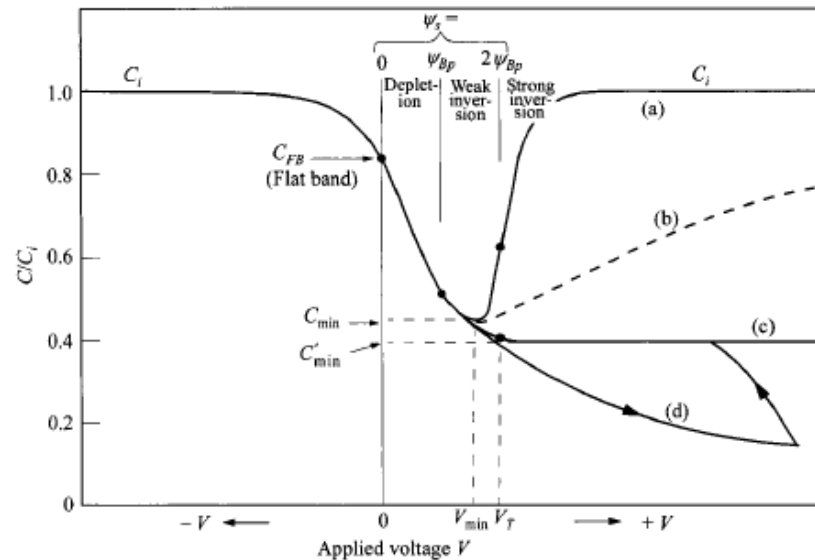


Fig. 7 MIS C - V curves. Voltage is applied to the metal relative to the p -semiconductor. (a) Low frequency. (b) Intermediate frequency. (c) High frequency. (d) High frequency with fast sweep (deep depletion). Flat-band voltage of $V = 0$ is assumed.

- ▶ we can express the depletion width as a function of the terminal voltages. The quadratic equation gives a solution of:

$$W_D = \sqrt{\frac{\epsilon_s^2}{C_{ox}^2} + \frac{2\epsilon_s V}{qND}} - \frac{\epsilon_s}{C_{ox}}$$

- ▶ Once W_D is known, C_D and ψ_s are deduced. The depletion capacitance be estimated by:

$$C_D = \sqrt{\frac{\epsilon_s q p_{p0}}{2\psi_s}} = \frac{\epsilon_s}{W_D}$$

- ▶ With further increase in positive voltage, the depletion region widens which acts as a dielectric at the semiconductor surface in series with the insulator
- ▶ The total capacitance continues to decrease.

- ▶ The capacitance goes through a minimum and then increases again as the inversion layer of electrons forms at the surface.
- ▶ The minimum capacitance and the corresponding minimum voltage are designated C_{\min} and V_{\min} .
- ▶ Since C_i is fixed, C_{\min} can be found by the minimum value of C_D
- ▶ Experimentally, it is found that for the metal-SO₂-Si system the range in which the capacitance is most frequency-dependent is between 5 Hz and 1 kHz
- ▶ This is related to the carrier lifetime and thermal generation rate in the silicon substrate.

High-Frequency Capacitance.

- ▶ The high-frequency curve can be obtained using an approach analogous to a one-sided abrupt *p-n* junction.
- ▶ When the semiconductor surface is depleted, the ionized acceptors in the depletion region are given by- $qN_A W_D$
- ▶ Integrating the Poisson equation yields the potential distribution in the depletion region:

$$\psi_p(x) = \psi_s \left(1 - \frac{x}{W_D}\right)^2$$

$$\psi_s = \frac{qN_A W_D^2}{2\epsilon_s}$$

- ▶ When the applied voltage increases ψ_s and W_D increase.
- ▶ Once strong inversion occurs, the depletion-layer width reaches a maximum.
- ▶ When the bands are bent down far enough that $\psi_s = 2\psi_{Bp}$, the semiconductor is effectively shielded from further penetration of the electric field by the inversion layer very large increase in the charge density within the inversion layer.

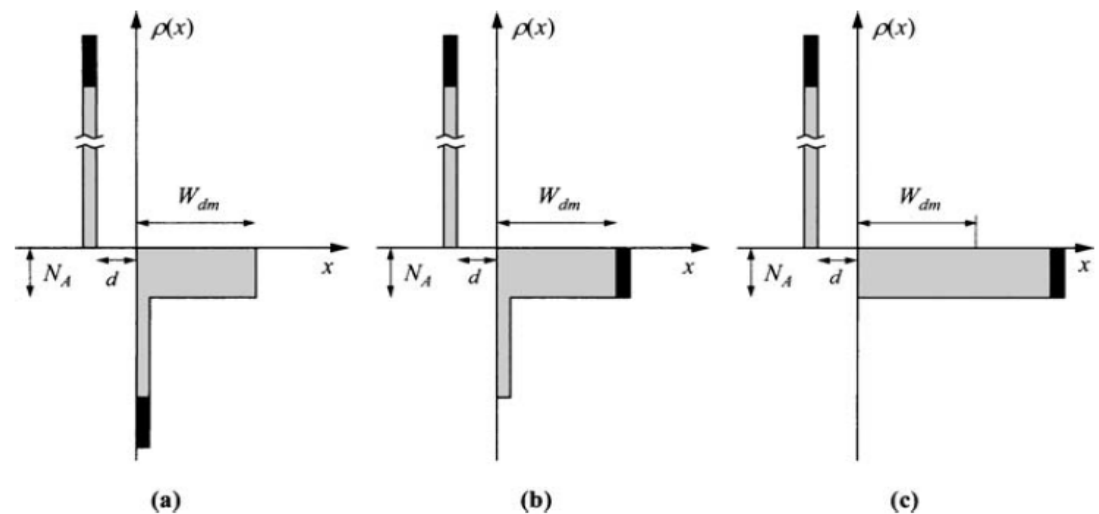


Fig. 8 In strong inversion, capacitance is a function of the small-signal frequency and the quiescent sweep rate. The incremental displacement charge (black area) is shown in cases of (a) low frequency,

- ▶ The maximum width W_{Dm} of the depletion region under steady state condition:

$$W_{Dm} = \sqrt{\frac{2\varepsilon_s\psi_s(\text{strong inv})}{qN_A}} = \sqrt{\frac{4\varepsilon_sKT\ln\left(\frac{N_A}{ni}\right)}{q^2N_A}}$$

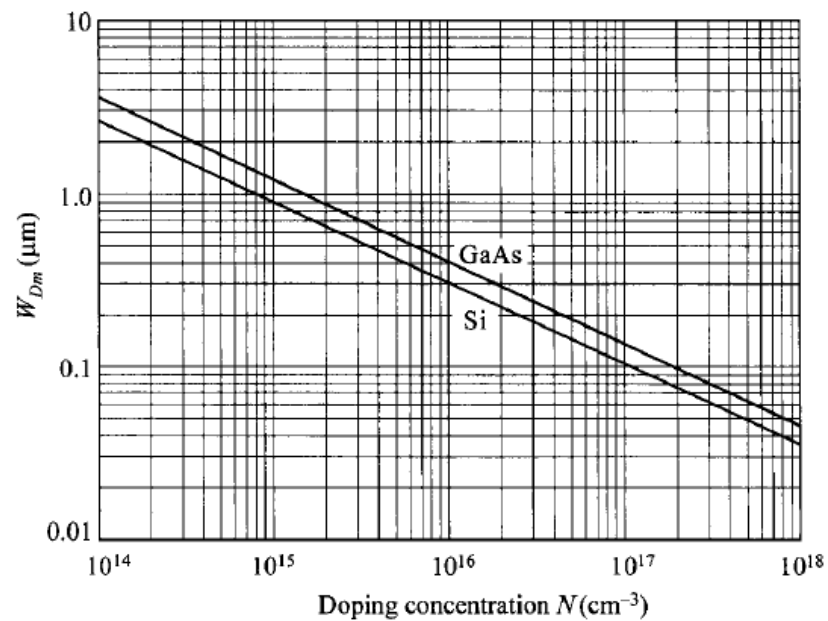


Fig. 9 Maximum depletion-layer width versus impurity concentration of semiconductors Si and GaAs under a heavy-inversion condition.

Threshold voltage

- ▶ The turn-on voltage or threshold voltage, V_T at which strong inversion occurs.

$$V_T = \left| \frac{Q_s}{C_i} \right| + 2\psi_{Bp} = \sqrt{\frac{2\varepsilon_s q N_A (2\psi_{Bp})}{C_i}} + 2\psi_{Bp}$$

- ▶ Note that even though the slow-varying quiescent voltage puts the additional charge at the surface inversion layer, the high-frequency small signal is too fast for the minority carriers

- ▶ The depletion capacitance is simply given by ϵ_s/W_D , with a minimum value corresponding to the maximum depletion width W_{DM}

$$C_{min} = \frac{\epsilon_i \epsilon_s}{\epsilon_s d + \epsilon_i W_{Dm}}$$

- ▶ Complete ideal C- V curves of the metal- SiO_2 -Si system have been computed for various oxide thicknesses and semiconductor doping densities.
- ▶ The conversion to n-type silicon is achieved simply by changing the sign of the voltage axes. Converting to other insulators requires scaling the oxide thickness with the ratio of the permittivity of SiO_2 and the other insulator.

$$d_c = d_i \frac{\epsilon_i(SiO_2)}{\epsilon_i(insulator)}$$

- ▶ where d_c is the equivalent SiO_2 thickness to be used in these curves, d_i and ϵ_i are the thickness and permittivity of the new insulator.

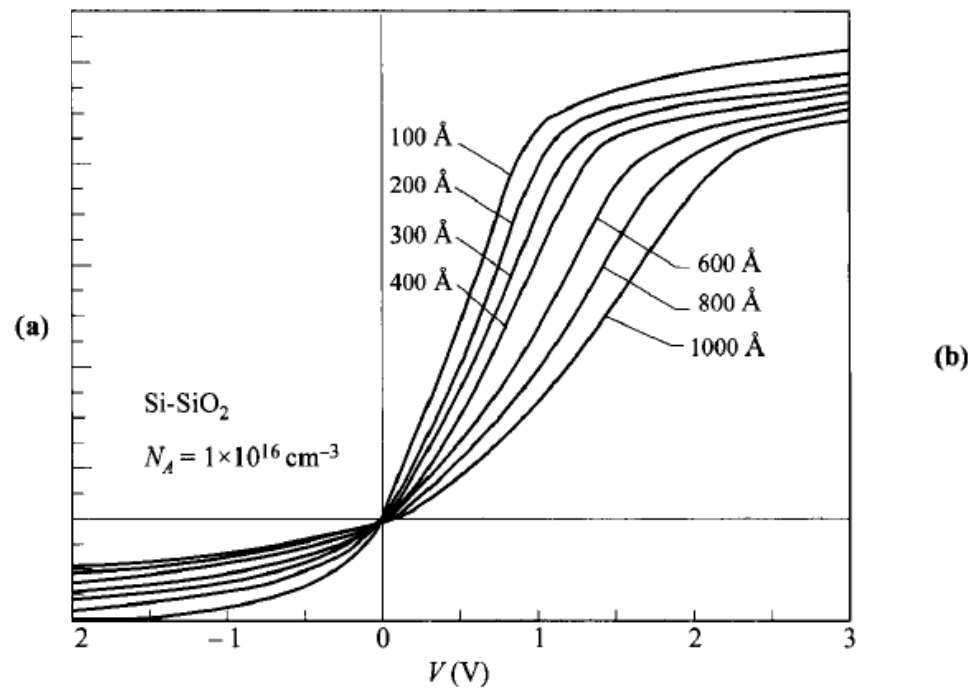
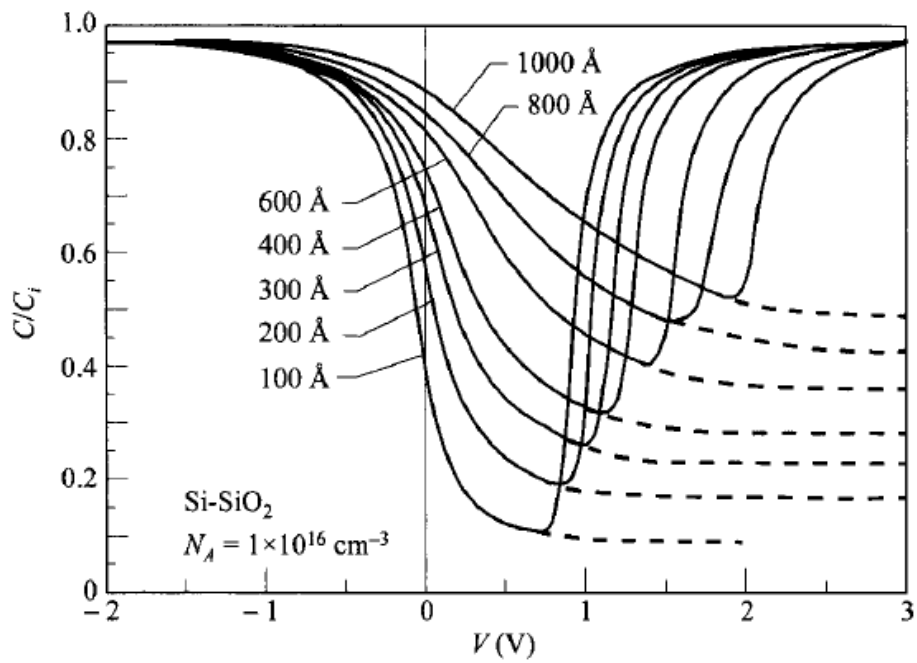


Fig. 10 (a) Ideal MOS C - V curves for various oxide thickness. Solid lines for low frequencies. Dashed lines for high frequencies. (b) Surface potential ψ_s vs. applied voltage. (After Ref. 15.)

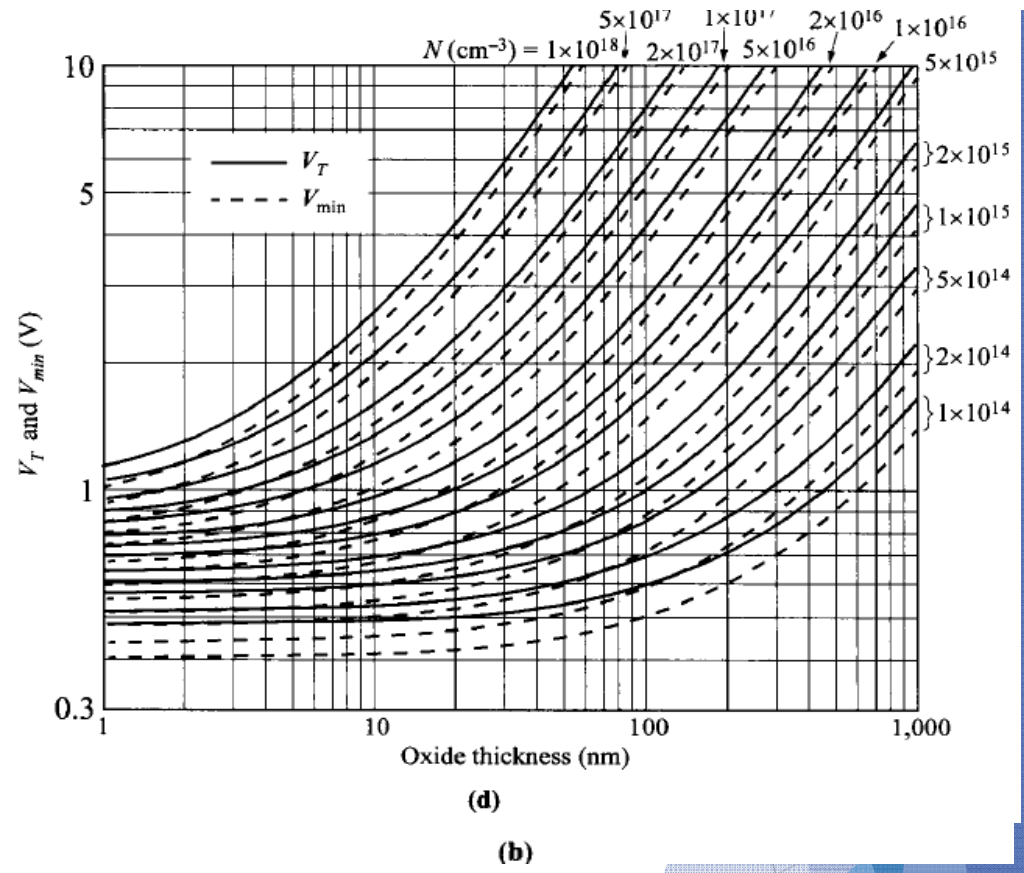
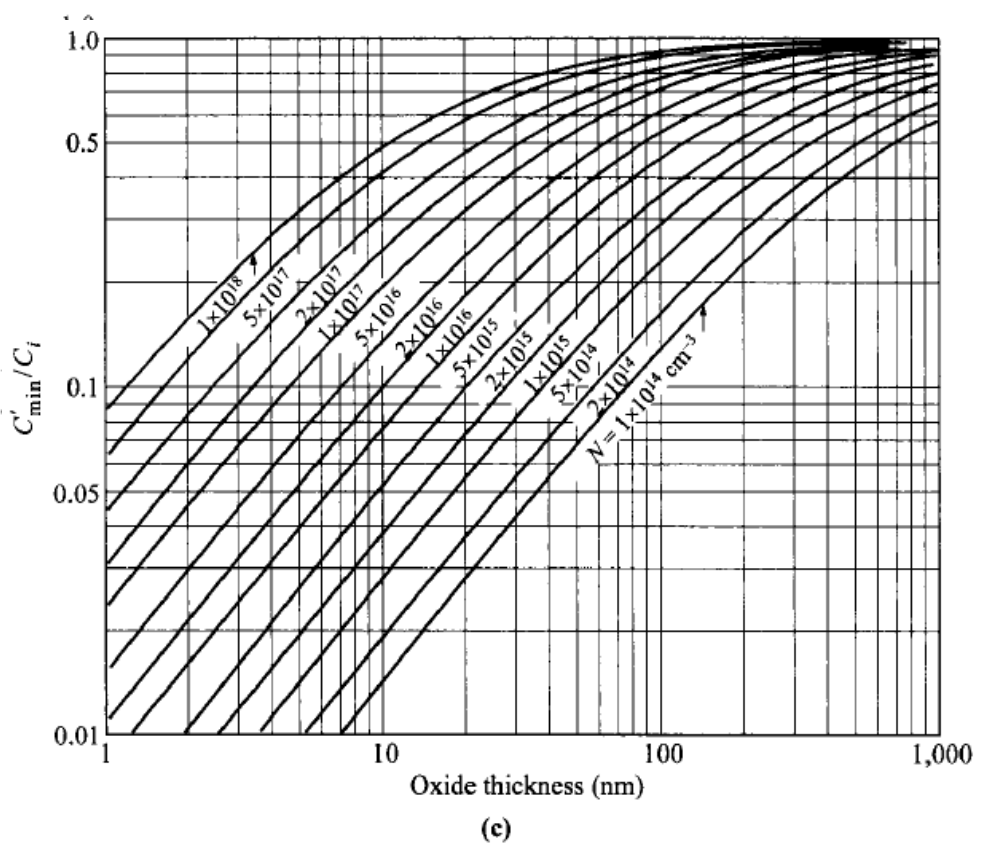


Fig. 11 Critical parameters of ideal SiO_2 -Si MOS capacitors as a function of doping level and oxide thickness. (a) Flat-band capacitance (normalized). (b) Low-frequency C_{min} (normalized) (c) High-frequency C'_{min} (normalized). (d) V_T and low-frequency V_{min} .

4.3 SILICON MOS CAPACITOR

- ▶ It consists of a single-crystal silicon followed by a monolayer of SiO_x that is, incompletely oxidized silicon, then a thin strained region of SiO_2 and the remainder stoichiometric, strain-free, amorphous SiO_2 .
- ▶ The compound SiO_x is stoichiometric when $x = 2$ and nonstoichiometric when $2 > x > 1$
- ▶ Practically
- ▶ In MOS capacitor, interface traps and oxide charges exist that will, in one way or another, affect the ideal MOS characteristics.

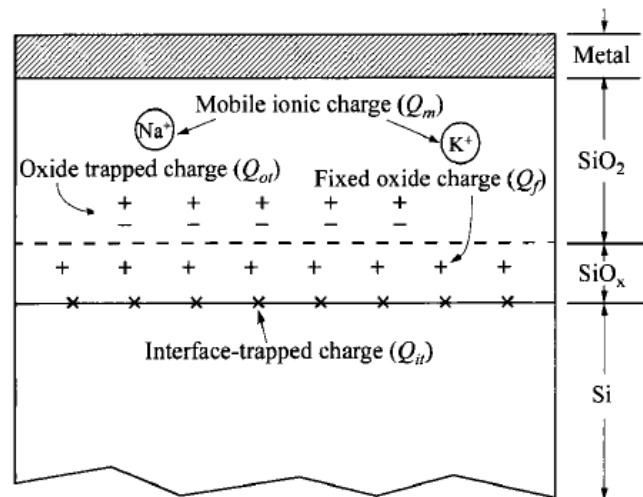


Fig. 12 Terminology for charges associated with thermally oxidized silicon. (After Ref. 16.)

4.3.1 Interface Traps

- ▶ the charge Q_{it} interface traps (historically also called interface states, fast states, or surface states)
- ▶ The Q_{it} exists within the forbidden gap due to the interruption of the periodic lattice structure at the surface of a crystal.
- ▶ Shockley and Pearson experimentally found the existence of Q_{it} in their surface conductance measurement.
- ▶ Measurements on clean surfaces in an ultra-high-vacuum system confirm that Q_{it} can be very high in the order of the density of surface atoms (= 10^{15} atoms/cm²)

- ▶ For the present MOS capacitors having thermally grown SiO₂ on Si, most of the interface-trapped charge can be neutralized by low-temperature (450°C) hydrogen annealing.
- ▶ The total surface traps can be as low as 10¹⁰ cm⁻²,
- ▶ An acceptor interface trap is neutral and becomes negatively charged by accepting an electron.
- ▶ The distribution functions (occupancy) for the interface traps are similar to those for the bulk impurity levels.

$$F_{SD}(E_t) = \left[1 - \frac{1}{1 + \frac{1}{g_D} e^{(E_F - E_t)/kT}} \right]$$

$$F_{SA}(E_t) = \frac{1}{1 + g_A \exp[(E_t - E_F)/kT]}$$

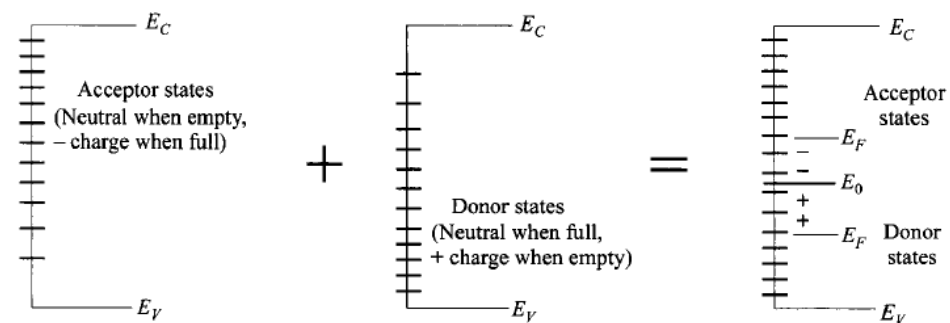


Fig. 13 Any interface-trap system consisting of both acceptor states and donor states can be interpreted by an equivalent distribution with a neutral level E_0 above which the states are of acceptor type and below which of donor type. When E_F is above (below) E_0 , net charge is - (+).

- ▶ A convenient notation is to interpret the sum of these by an equivalent D_{it} distribution, with an energy level called neutral level E_0 above which the states are of acceptor type, and below which are of donor type.
- ▶ The occupancy takes on the value of 0 and 1 above and below E_F .
- ▶ With these assumptions, the interface-trapped charge can now be easily calculated by:

$$\begin{aligned} Q_{it} &= -q \int_{E_0}^{E_f} D_{it} dE \\ &= +q \int_{E_f}^{E_0} D_{it} dE \end{aligned}$$

- ▶ interface-trap density distribution: $D_{it} = \frac{1}{q} \frac{dQ_{it}}{dE}$
- ▶ This is the concept used to determine D_{it} experimentally-from the change of Q_{it} in response to the change of E_F or surface potential ψ_s

- ▶ When a voltage is applied, the Fermi level moves up or down with respect to the interface-trap levels and a change of charge in the interface traps occurs.
- ▶ This change of charge affects the MIS capacitance and alters the ideal MIS curve.

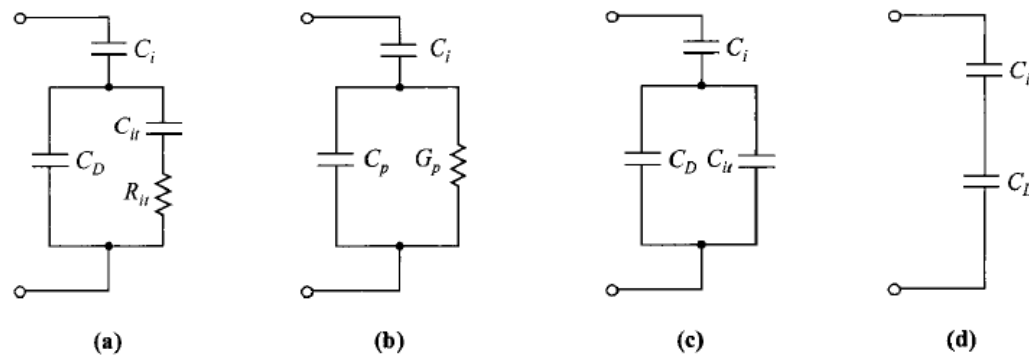


Fig. 14 (a)–(b) Equivalent circuits including interface-trap effects, C_{it} and R_{it} . (After Ref. 21.)
 (c) Low-frequency limit. (d) High-frequency limit.

- ▶ Where C_i and C_D are the insulator capacitance and the semiconductor depletion-layer capacitance.
- ▶ C_{it} and R_{it} are the capacitance and resistance associated with the interface traps and also functions of energy.
- ▶ The product C_{it} and R_{it} is defined as the interface-trap lifetime τ_{it} related to the frequency behavior of the interface traps.

- ▶ The parallel branch of the equivalent circuit can be converted into a frequency-dependent capacitance C_p in parallel with a frequency dependent conductance G_p

$$C_p = C_D + \frac{C_{it}}{1 + \omega^2 \tau_{it}^2}$$

And

$$\frac{G_p}{\omega} = \frac{C_{it} \omega \tau_{it}}{1 + \omega^2 \tau_{it}^2}$$

- ▶ Physically it means that the traps are not fast enough to respond to the fast signal.
- ▶ The total terminal capacitance for these two cases (low-frequency C_{LF} and high-frequency C_{HF})

- ▶
$$C_{LF} = \frac{C_i(C_D + C_{it})}{C_i + C_D + C_{it}}$$

- ▶
$$C_{HF} = \frac{C_i C_D}{C_i + C_D}$$

- ▶ Oxide charges, other than that of the interface traps, include the fixed oxide charge Q_f , the mobile ionic charge Q_m and the oxide trapped charge Q_{ot} .
- ▶ In general, unlike interface-trapped charges, these oxide charges are independent of bias, so they cause a parallel shift in the gate-bias direction.
- ▶ The flat-band voltage shift due to any oxide charge is given by Gauss' law:

$$\Delta V = \frac{1}{C_i} \left[\frac{1}{d} \int_0^d x \rho(x) dx \right]$$

- ▶ where $\rho(x)$ is the charge density per unit volume.
- ▶ The effect on the voltage shift is weighted according to the location of the charge.
- ▶ The closer to the oxide-semiconductor interface, the more shift it will cause.
- ▶ Positive charge is equivalent to an added positive gate bias for the semiconductor so it requires a more negative gate bias to achieve the same original semiconductor band bending.

- ▶ Q_f can be regarded as a charge sheet located at the Si-SiO₂ interface:

$$\Delta V_f = -\frac{Q_f}{C_i}$$

- ▶ Mobile ionic charges can move back and forth through the oxide layer, depending on biasing conditions, and thus give rise to voltage shifts.

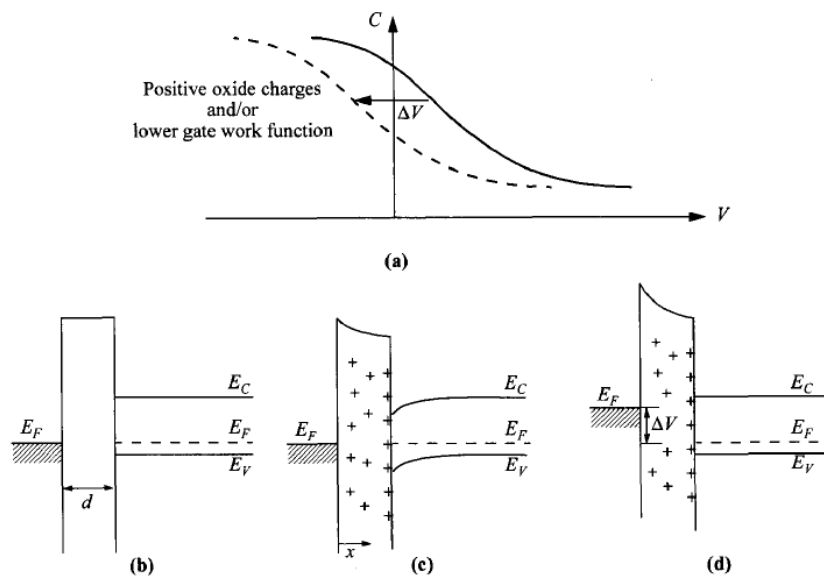


Fig. 19 (a) High-frequency C - V curve (on p -semiconductor), shifted along the voltage axis due to positive oxide charges. (b) Band diagram at flat band, original. (c) With positive oxide charges and (d) new flat-band bias.

- ▶ Reliability problems in semiconductor devices operated at high temperatures and voltages may be related to trace contamination by alkali metal ions.
- ▶ The voltage shift is given by :

$$\Delta V_m = -\frac{Q_m}{C_i}$$

- ▶ where Q_m is the effective net charge of mobile ions per unit area at the Si-SiO₂ interface and the actual mobile ions $\rho(x)$ is used.
- ▶ To prevent mobile ionic charge contamination of the oxide during device life, one can protect it with a film impervious to mobile ions such as amorphous or small-crystallite silicon nitride.
- ▶ For amorphous SiN_x, there is very little sodium penetration. Other sodium barrier layers include AlO and phosphosilicate glass.

- ▶ The oxide traps are usually initially neutral and are charged by introducing electrons and holes into the oxide layer.
- ▶ This can occur from any current passing through the oxide layer, hot-carrier injection, or by photon excitation. The shift due to the oxide trapped charge:

$$\Delta V_{ot} = -\frac{Q_{ot}}{C_i}$$

- ▶ where Q_{ot} is the effective net charge per unit area at the Si-SiO₂ interface.
- ▶ The total voltage shift due to all the oxide charges is the sum:

$$\Delta V = \Delta V_f + \Delta V_m + \Delta V_{ot} = \frac{-(Q_f + Q_m + Q_{ot})}{C_i}$$

Work-Function Difference.

- ▶ For the preceding discussions on ideal MIS capacitor it has been assumed that the work-function difference for a p-type semiconductor.

$$\phi_{ms} \equiv \phi_m - \left(\chi + \frac{E_g}{2q} + \psi_{Bp} \right)$$

- ▶ If the value of ϕ_{ms} is not zero, the experimental C - V curve will be shifted from the theoretical curve by the same amount in gate bias.
- ▶ This shift is in addition to the oxide charges, so the net flat-band voltage becomes:

$$V_{FB} \equiv \phi_{ms} - \frac{(Q_f + Q_m + Q_{ot})}{C_i}$$

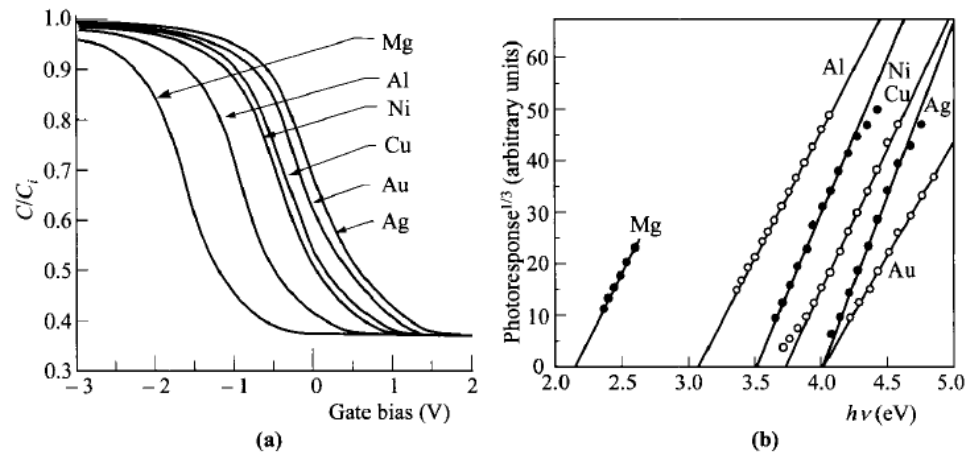


Fig. 21 Correlation of (a) flat-band voltage from capacitance measurement and (b) barrier height from photoresponse. (After Ref. 29.)

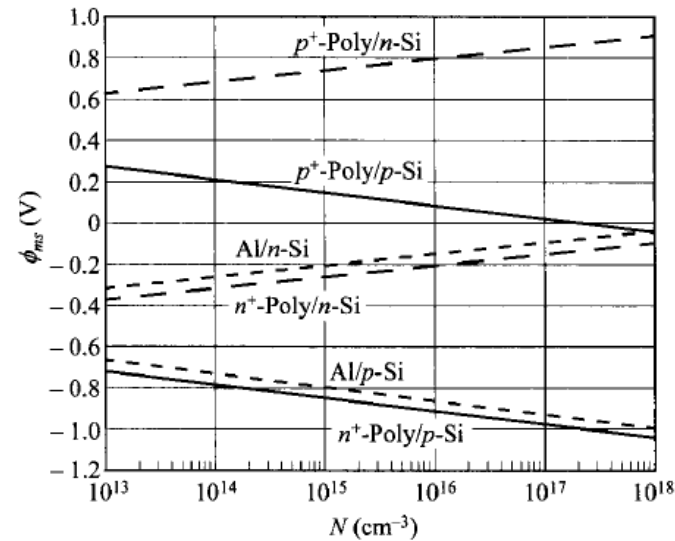


Fig. 22 Work-function difference ϕ_{ms} vs. doping, for gate electrodes of degenerate polysilicon and Al on p - and n -Si.

4.3.4 Carrier Transport

- ▶ Real insulators, however, show some degree of carrier conduction when the electric field or temperature is sufficiently high.
- ▶ To estimate the electric field in an insulator under biasing conditions:

$$E_i = E_s \left(\frac{\epsilon_s}{\epsilon_i} \right) \approx \frac{V}{d}$$

- ▶ where E_i and E_s are the electric fields in the insulator and the semiconductor respectively. ϵ_i and ϵ_s are the corresponding permittivity's

Table 2 Basic Conduction Processes in Insulators

Process	Expression	Voltage & temperature dependence
Tunneling	$J \propto \mathcal{E}_i^2 \exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar\mathcal{E}_i}\right]$	$\propto V^2 \exp\left(\frac{-b}{V}\right)$
Thermionic emission	$J = A^{**}T^2 \exp\left[\frac{-q(\phi_B - \sqrt{q\mathcal{E}_i/4\pi\epsilon_i})}{kT}\right]$	$\propto T^2 \exp\left[\frac{q}{kT}(a\sqrt{V} - \phi_B)\right]$
Frenkel-Poole emission	$J \propto \mathcal{E}_i \exp\left[\frac{-q(\phi_B - \sqrt{q\mathcal{E}_i/\pi\epsilon_i})}{kT}\right]$	$\propto V \exp\left[\frac{q}{kT}(2a\sqrt{V} - \phi_B)\right]$
Ohmic	$J \propto \mathcal{E}_i \exp\left(\frac{-\Delta E_{ac}}{kT}\right)$	$\propto V \exp\left(\frac{-c}{T}\right)$
Ionic conduction	$J \propto \frac{\mathcal{E}_i}{T} \exp\left(\frac{-\Delta E_{ai}}{kT}\right)$	$\propto \frac{V}{T} \exp\left(\frac{-d'}{T}\right)$
Space-charge-limited	$J = \frac{9\epsilon_i\mu V^2}{8d^3}$	$\propto V^2$

A^{**} = effective Richardson constant. ϕ_B = barrier height. \mathcal{E}_i = electric field in insulator. ϵ_i = insulator permittivity. m^* = effective mass. d = insulator thickness. ΔE_{ac} = activation energy of electrons. ΔE_{ai} = activation energy of ions. $V \approx \mathcal{E}_i d$. $a \equiv \sqrt{q/4\pi\epsilon_i}$. b , c , and d' are constants.

- ▶ At low voltage and high temperature, current is carried by thermally excited electrons hopping from one isolated state to the next.
- ▶ This mechanism yields an ohmic characteristic exponentially dependent on temperature.
- ▶ The ionic conduction is similar to a diffusion process. Generally, the dc ionic conductivity decreases during the time the electric field is applied because ions cannot be readily injected into or extracted from the insulator.
- ▶ After an initial current flow, positive and negative space charges will build up near the metal-insulator and the semiconductor-insulator interfaces, causing a distortion of the potential distribution.
- ▶ The Frenkel-Poole emission, shown in below is due to emission of trapped electrons into the conduction band.
- ▶ The supply of electrons from the traps is through thermal excitation. For trap states with Coulomb potentials, the expression is similar to that of the Schottky emission.
- ▶ The barrier height, however, is the depth of the trap potential well. The barrier reduction is larger than in the case of Schottky emission by a factor of 2, since the barrier lowering is twice as large due to the immobility of the positive charge.

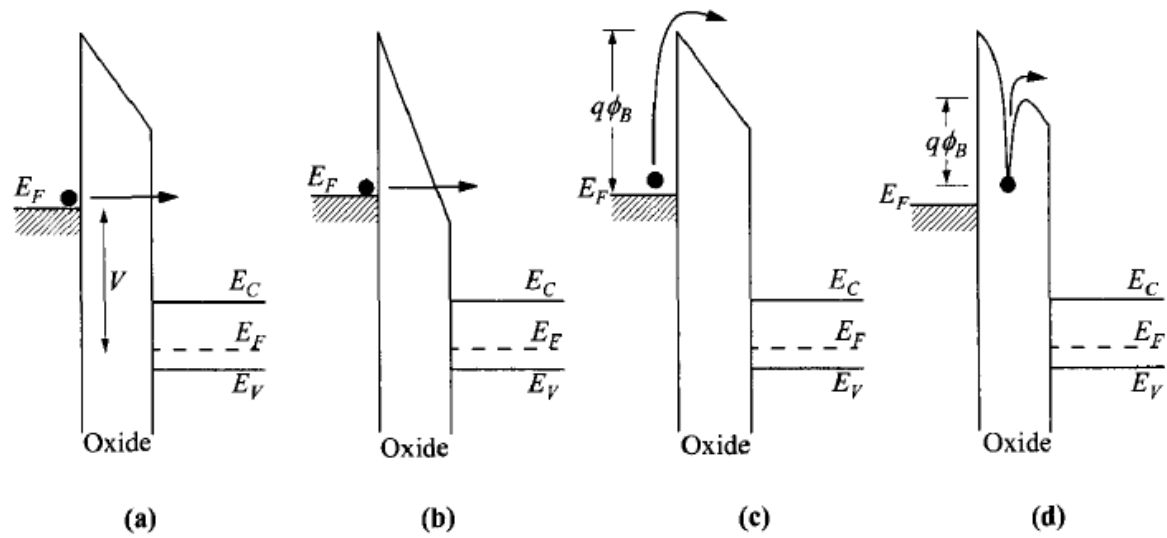


Fig. 23 Energy-band diagrams showing conduction mechanisms of (a) direct tunneling, (b) Fowler-Nordheim tunneling, (c) thermionic emission, and (d) Frenkel-Poole emission.

4.3.6 Accumulation- and Inversion-Layer Thickness

- ▶ For an MIS capacitor, the maximum capacitance is equal to ϵ_i / d which implies that charges on both sides of the electrodes cling to the two interfaces of the insulator.
- ▶ While such an assumption is valid on the metal-insulator interface, detailed examination on the insulator-semiconductor interface reveals that it can lead to considerable error, especially for thin oxides.
- ▶ This is due to charges on the semiconductor side, either accumulation or strong-inversion charges, have a distribution as a function of distance from the interface.
- ▶ Effectively this would reduce the maximum capacitance given by ϵ_i / d .

► Classical Model.

The charge distribution is controlled by the Poisson equation. Using Boltzmann statistics,

$$p(x) = N_A \exp\left(\frac{-q\psi_p}{KT}\right)$$

$$\frac{d^2\psi}{dx^2} = -\frac{\rho}{\epsilon_s} \approx -\frac{qN_A e^{-\frac{q\psi_p}{KT}}}{\epsilon_s}$$

$$\psi_p(x) = -\frac{kT}{q} \ln\left(\sec^2\left\{\cos^{-1}\left[\exp\left(\frac{q\psi_p}{2kT}\right)\right] - \frac{x}{\sqrt{2}L_D}\right\}\right).$$

- ▶ The total accumulation layer thickness where ψ_p approaches zero is equal to $\pi L_D/2$

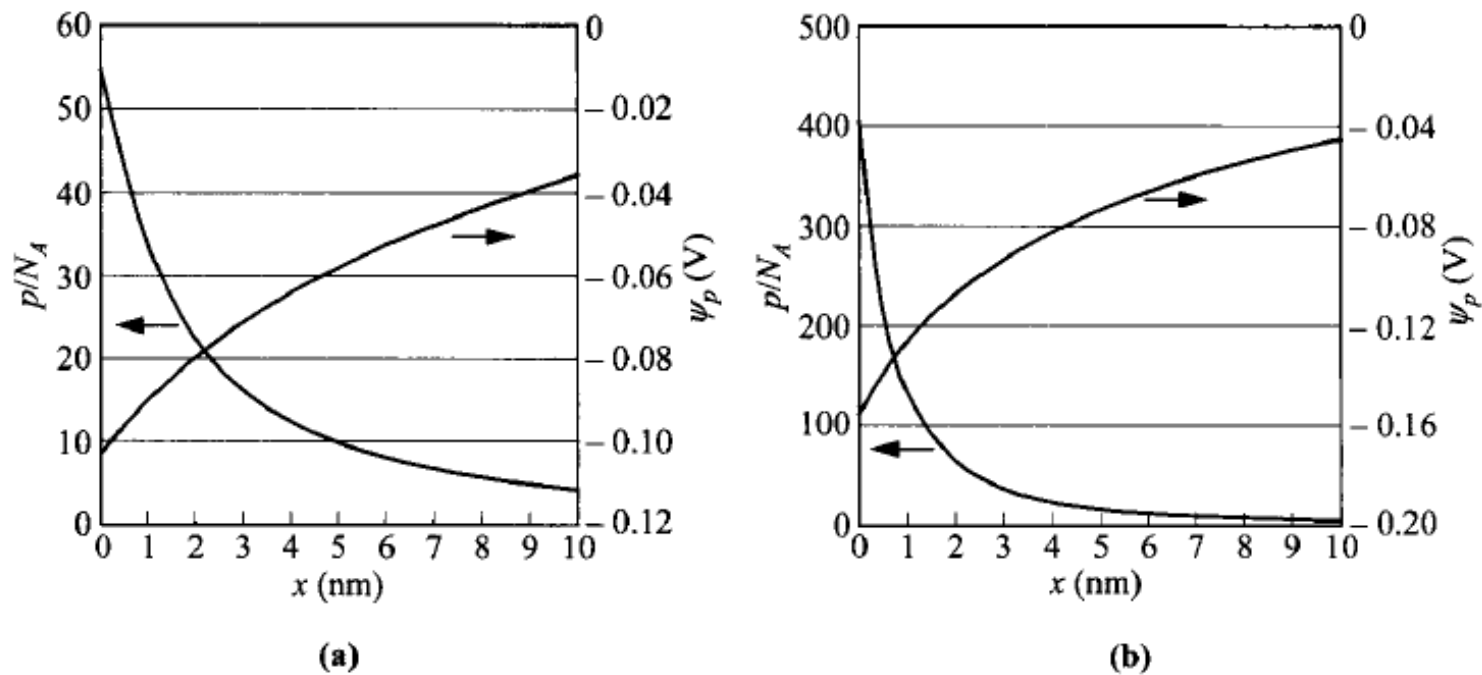


Fig. 28 Classical calculation of potential and carrier profiles, with a surface potential ψ_s of (a) $4kT/q$ and (b) $6kT/q$.

Quantum-Mechanical Model

- ▶ In quantum mechanics, the wave function associated with the carriers is near zero at the insulator-semiconductor interface because of the high barrier of the insulator.
- ▶ the carrier concentration peaks at some finite distance from the interface.
- ▶ This distance is approximately 10 \AA

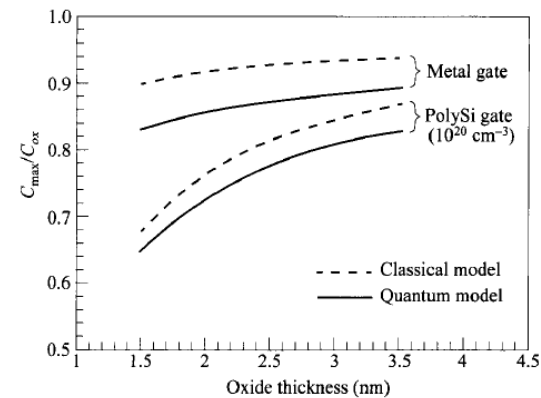


Fig. 29 Quantum-mechanical calculation of capacitance reduction. Also shown are results from classical model and those including depletion effect from polysilicon gate. (After Ref. 43.)



MOSFETs

MOSFETs

- ▶ 6.1 INTRODUCTION
- ▶ 6.2 BASIC DEVICE CHARACTERISTICS
- ▶ 6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE
- ▶ 6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS
- ▶ 6.5 MOSFET STRUCTURES
- ▶ 6.6 CIRCUIT APPLICATIONS
- ▶ 6.7 NONVOLATILE MEMORY DEVICES
- ▶ 6.8 SINGLE-ELECTRON TRANSISTOR

6.1 INTRODUCTION

- ▶ The metal-oxide-semiconductor field-effect transistor (MOSFET) is the most-important device for forefront high-density integrated circuits such as microprocessors and semiconductor memories.
- ▶ The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld and Heil.
- ▶ It was subsequently studied by Shockley and Pearson in the late 1940s. In 1960, Ligenza and Spitzer produced the first device-quality Si-SiO₂ MOS system using thermal oxidation.
- ▶ The basic device characteristics have been initially studied by Ihantola and Moll, Sah and Hofstein and Heiman.
- ▶ The reduction of the gate-length dimension in production ICs since 1970.
- ▶ This dimension has been decreasing at a steady pace and will continue to *shrink* in the foreseeable future.

- ▶ The number of components per integrated circuit chip has grown exponentially. The rate of growth is expected to slow down because of increasing technological challenge and fabrication cost.

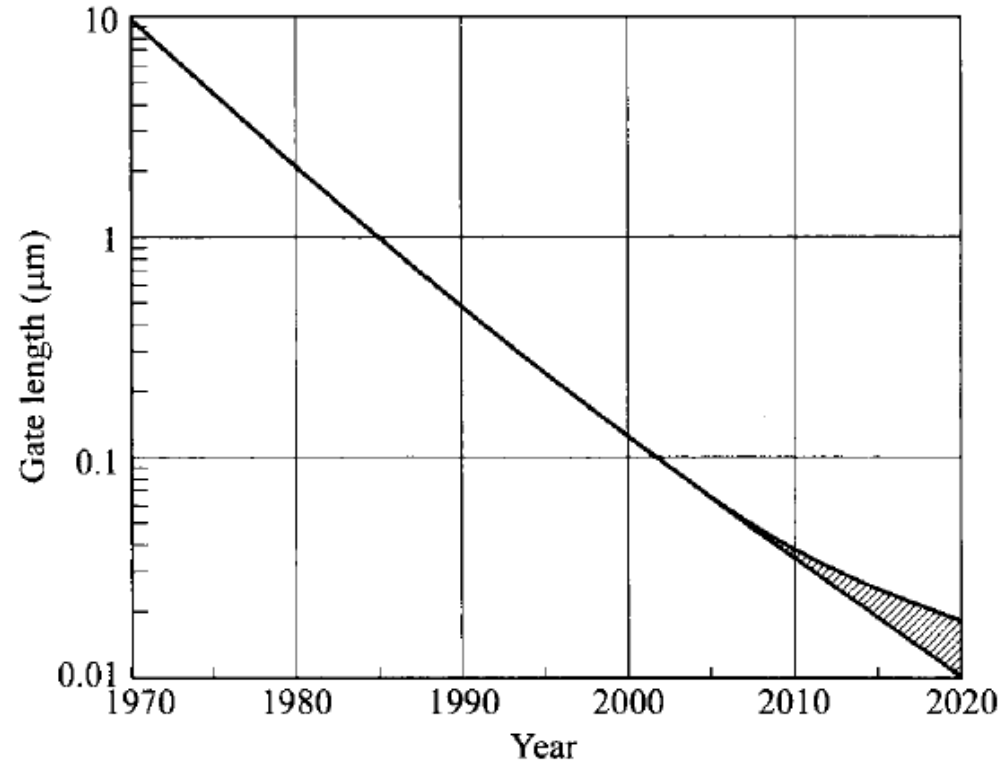


Fig. 1 Minimum gate dimension in commercial integrated circuit as a function of the year of production.

6.1.1 Field-Effect Transistors: Family Tree

- ▶ The MOSFET is the main member of the family of field-effect transistors.
- ▶ A transistor in general is a three-terminal device where the channel resistance between two of the contacts is controlled by the third.
- ▶ The difference between the FET and the PET is the way the control is coupled to the channel.
- ▶ In an FET, the channel is controlled capacitively by an electric field (hence the name *field effect*).

- ▶ In a name

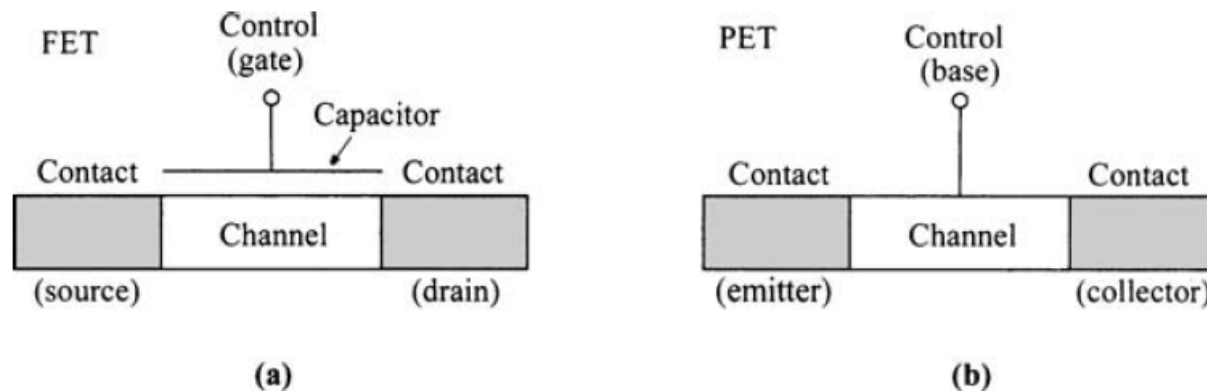


Fig. 2 Distinction between (a) field-effect transistor (FET) and (b) potential-effect transistor (PET).

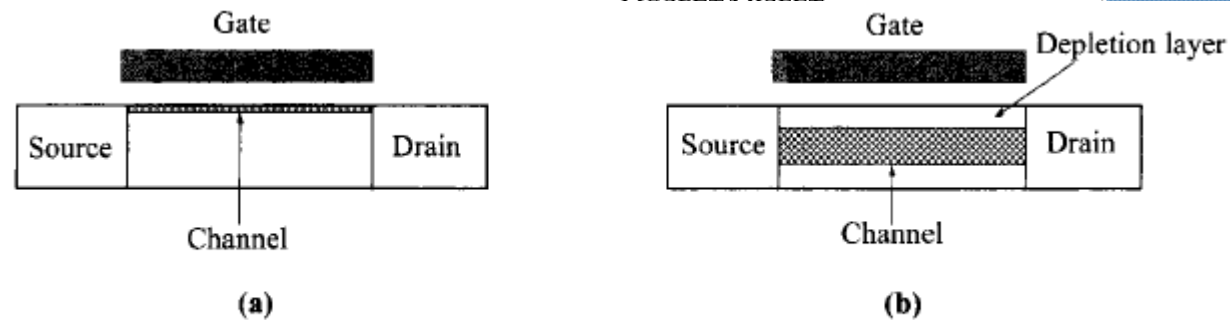


Fig. 5 FET channels: (a) surface inversion channel and (b) buried channel.

- ▶ The FETs have considerably higher input impedance than bipolar transistors, which allows the input of a FET to be more readily matched to the standard microwave system.
- ▶ The FET has a negative temperature coefficient at high current levels; that is, the current decreases as temperature increases.
- ▶ This characteristic leads to a more uniform temperature distribution over the device area and prevents the FET from thermal runaway or second breakdown.
- ▶ In addition, the devices are basically square-law or linear devices; intermodulation and cross-modulation products are smaller than those of bipolar transistors.

6.2 BASIC DEVICE CHARACTERISTICS

- ▶ Throughout this chapter we assume the channel carriers are electrons-an n-channel device.
- ▶ All discussion and equations will be applicable to the counterpart p-channel devices with appropriate substitution of parameters and the reversal of polarity of the applied voltages.
- ▶ A common MOSFET is a four-terminal device that consists of a p-type semiconductor substrate into which two n⁺-regions, the source and drain, are formed, usually by ion implantation.
- ▶ The SiO₂ gate dielectric is formed by thermal oxidation of Si for a high quality SiO₂-Si interface.
- ▶ The metal contact on the insulator is called the gate.

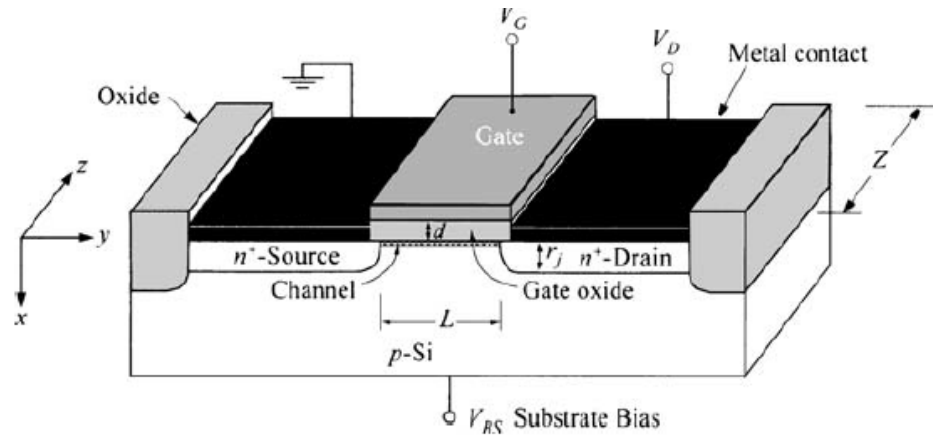


Fig. 6 Schematic diagram of a MOSFET.

- The gate electrode use doped polysilicon or a combination of silicide and polysilicon.
- The basic device parameters are the channel length L , which is the distance between the two metallurgical n+-p junctions.
 - Channel width Z .
 - Insulator thickness d .
 - Junction depth rj .
 - Substrate doping N_A .
- When ground or a low voltage is applied to the gate, the main channel is shut off, and the source-to-drain electrodes correspond to two p-n junctions connected back to back.
- When a sufficiently large positive bias is applied to the gate so that a surface inversion layer (or channel) is formed between the two n+-regions.
- The source and the drain are then connected by a conducting surface n-channel through which a large current can flow

6.2.1 Inversion Charge in Channel

- ▶ When a voltage is applied across the source-drain contacts.
- ▶ the MOS structure is in a non-equilibrium condition
- ▶ the minority carrier is in a quasi-Fermi equilibrium
- ▶ The two-
 $V_D = V_{Bs} =$

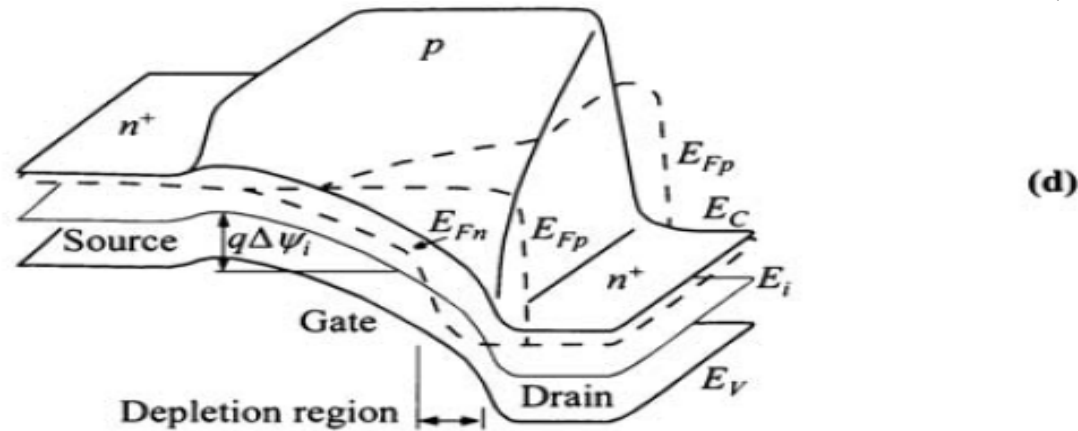


Fig. 7 Two-dimensional band diagram of an *n*-channel MOSFET. (a) Device configuration. (b) Flat-band zero-bias equilibrium condition. (c) Equilibrium condition ($V_D = 0$) under a positive gate bias. (d) Nonequilibrium condition under both gate and drain biases. (After Ref. 20.)

- ▶ The gate voltage required for inversion at the drain is larger than the equilibrium case in which $\psi_s(\text{inv}) = 2\psi_{Bn}$
- ▶ In other words, the inversion-layer charge at the drain end is lowered by the drain bias.
- ▶ This is because the applied drain bias lowers the E_{Fn} and an inversion layer can be formed only when the surface potential meets the criteria of $[E_{Fn} - E_i(0)] > q \psi_{Bn}$, where $E_i(0)$ is the intrinsic Fermi level at $x = 0$.

- ▶ For the non equilibrium case, the depletion-layer width is deeper than W_{Dm} and is a function of the drain bias V_D .
- ▶ The surface potential $\psi_s(y)$ at the drain at the onset of strong inversion is:

$$\psi_s(inv) \approx VD + 2\psi_B$$

- ▶ The characteristics of the surface space charge under the non-equilibrium condition are derived under two assumptions:
 1. the majority-carrier quasi-Fermi level E_{Fp} is the same as that of the substrate and it does not vary with distance from the bulk to the surface (constant with x).
 2. the minority-carrier quasi-Fermi level E_{Fn} is lowered by the drain bias by an amount dependent on the y -position.
- ▶ The first assumption introduces little error when the surface is inverted, because majority carriers are only a negligible part of the surface space charge.

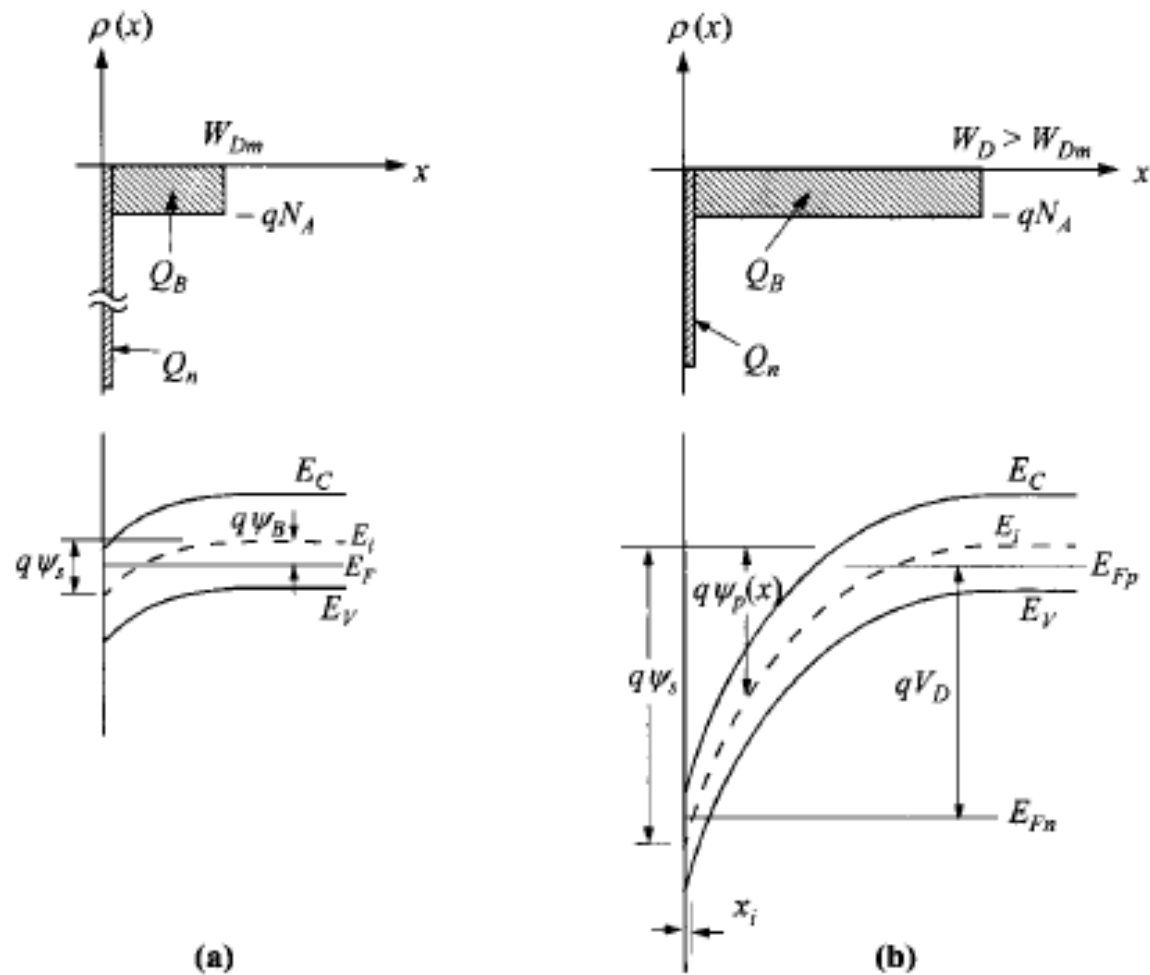


Fig. 8 Comparison of charge distribution and energy-band variation of an inverted p -region in (a) equilibrium and (b) nonequilibrium at the drain end. (After Ref. 21.)

$$\frac{d^2y}{dx^2} = \frac{q}{\epsilon_s} (NA - p + n)$$

► Where

$$p_{po} = NA = \frac{ni^2}{n_{po}}$$

$$p = NAe^{-\beta\psi_p}$$

$$n = n_{po}e^{\beta\psi_p - \beta VD}$$

► Conceptually the charge due to minority carriers within the inversion layer, is given by:

$$|Q_n| = q \int_0^x n(x) dx = q \int_{\psi_s}^{\psi_B} \frac{n(\psi_p)}{d\psi_p/dx} d\psi_p$$

$$= q \int_{\psi_s}^{\psi_B} \frac{n_{po} e^{\beta\psi_p - \beta VD} d\psi_p}{\left(\sqrt{2} \frac{KT}{qL_D}\right) F(\beta\psi_p, VD)}$$

- ▶ where x_i denotes the point at which $q\psi_p(x) = E_{Fn} - E_i(x) = q\psi_B$ and the function F

$$F\left(\beta\psi_p, V_D, \frac{n_{po}}{p_{po}}\right) \equiv \sqrt{\left[e^{(-\beta\psi_p)} + \beta\psi_p - 1\right] + \frac{n_{po}}{p_{po}} \left[e^{(-\beta V_D)} [e^{(\beta\psi_p)} - \beta\psi_p e^{(\beta V_D)}] - 1\right]}$$

- ▶ For the practical doping ranges in silicon, the value of x_i is quite small, of the order of 3 to 30 nm.
- ▶ The surface electric field in the x-direction at the drain end is given by:

$$E_s = -\left.\frac{d\psi_p}{dx}\right|_{x=0} = \pm \left(\sqrt{2} \frac{KT}{qL_D}\right) F\left(\beta\psi_s, VD, \frac{n_{po}}{p_{po}}\right)$$

- ▶ The total semiconductor surface charge is then obtained from Gauss' law

$$Q_s = -\varepsilon_s E_s = \mp \left(\sqrt{2} \frac{KT}{qL_D}\right) F\left(\beta\psi_s, VD, \frac{n_{po}}{p_{po}}\right)$$

- ▶ where the Debye length is:

$$L_D \equiv \sqrt{\frac{KT\varepsilon_s}{N_A q^2}}$$

- ▶ The inversion charge per unit area Q_n after strong inversion is given by:

$$Q_n = Q_s - QB$$

- ▶ where the depletion bulk charge is:

$$Q_B = -qNAWD = -\sqrt{2qNA\varepsilon_s(VD + 2\psi_B)}$$

- ▶ the inversion charge Q_n at the drain end can be simplified to:

$$|Q_n| \approx \sqrt{2}qN_A L_D \left[\sqrt{\beta\psi_s + \left(\frac{n_{po}}{p_{po}}\right)e^{(\beta\psi_s - \beta VD)}} - \sqrt{\beta\psi_s} \right]$$

- ▶ This solution is still difficult to use because at strong inversion, Q_n is very sensitive to the surface potential ψ_s
- ▶ the relationship to the terminal bias, that is V_G is still missing

Charge-Sheet Model

- ▶ In the charge-sheet under strong-inversion conditions, the inversion layer is treated as a charge sheet with zero thickness ($x_i = 0$).

- ▶ this assumption implies that the potential drop across this charge sheet is also zero.

$$\Delta\psi_i(y) \equiv \frac{E_i(x=0,y=0) - E_i(x=0,y)}{q},$$

- ▶ From Gauss' law, the boundary conditions on both sides of the charge sheet are:

$$E_{ox}\epsilon_{ox} = Es\epsilon_s - Qn$$

- ▶ In order to express $Q_n(y)$ throughout the channel, the surface potential is generalized:

$$\psi_s(y) = \Delta\psi_i(y) + 2\psi_B$$

- ▶ where $\Delta\psi_i(y)$ is the channel potential with respect to the source end.

$$\Delta\psi_i(y) \equiv \frac{E_i(x = 0, y = 0) - E_i(x = 0, y)}{q}$$

- ▶ The electric field of oxide:

$$E_{ox} = \frac{V_G - \psi_s}{d} = \frac{V_G - (\Delta\psi_i + 2\psi_B)}{d}$$

$$E_s = \sqrt{\frac{2qNA(\Delta\psi_i + 2\psi_B)}{\epsilon_s}}$$

- ▶ An ideal MOS system with zero work-function difference is assumed.

$$C_{ox} = \frac{\epsilon_{ox}}{d}$$

$$|Qn(y)| = [V_G - (\Delta\psi_i + 2\psi_B)]C_{ox} - \sqrt{2qNA\epsilon_s(\Delta\psi_i + 2\psi_B)}$$

- ▶ This final form will be used as the *channel charge responsible for the current conduction*.

6.2.2 Current-Voltage Characteristics

- ▶ We shall now derive the basic MOSFET characteristics under the following idealized conditions:
 1. The gate structure corresponds to an ideal MOS capacitor that is there are no interface traps nor mobile oxide charge.
 2. Only drift current will be considered.
 3. doping in the channel is uniform.
 4. Reverse leakage current is negligible.
 5. the transverse field (E_x in the x-direction) in the channel is much larger than the longitudinal field (E_y the y-direction).
- ▶ This last condition corresponds to the so-called gradual-channel approximation.

- ▶ the requirements of zero fixed oxide charge and work-function difference are removed.
- ▶ their effects are included in a flat-band voltage V_{FB} required by the gate to produce the flat-band condition.
- ▶ The V_G is replaced by $V_G - V_{FB}$ for the inversion charge.

$$|Qn(y)| = [V_G - V_{FB} - (\Delta\psi_i + 2\psi_B)]C_{ox} - \sqrt{2qNA\epsilon_s(\Delta\psi_i + 2\psi_B)}$$

- ▶ Under such idealized conditions, the channel current at any y -position is given by:

$$I_D(y) = Z|Qn(y)|v(y)$$

where $v(y)$ is the average carrier velocity

- ▶ the current has to be continuous and constant throughout the channel LENGTH :

- ▶ Type equation here.

$$I_D(y) = \frac{Z}{L} \int_0^L |Qn(y)|v(y)dy$$

- ▶ The carrier velocity $v(y)$ is a function of the y -position since the longitudinal field $E(y)$ is a variable.
- ▶ For shorter channel lengths, higher field causes velocity saturation and ultimately ballistic transport.

CONSTANT MOBILITY

$$\begin{aligned} I_D(y) &= \frac{Z\mu_n}{L} \int_0^L |Qn(y)|E(y)dy = \frac{Z\mu_n}{L} \int_0^L |Qn(y)| \frac{d\Delta\psi_i}{dy} dy = \frac{Z\mu_n}{L} \int_0^{V_D} |Qn(\Delta\psi_i)| \frac{d\Delta\psi_i}{d\Delta\psi_i} d\Delta\psi_i \\ &= \frac{Z\mu_n}{L} \text{cox} \left\{ \left(V_G - V_{FB} - 2\psi_B - \frac{V_D}{2} \right) V_D - \frac{2\sqrt{2qNA\epsilon_s}}{C_{ox}} [(V_D + 2\psi_B)^{3/2}] - 2\psi_B^{3/2} \right\} \end{aligned}$$

- ▶ A given V_G the drain current first increases linearly with drain voltage (the linear region).
- ▶ A gradually levels off (the nonlinear region).
- ▶ finally approaching a saturated value (the saturation region).

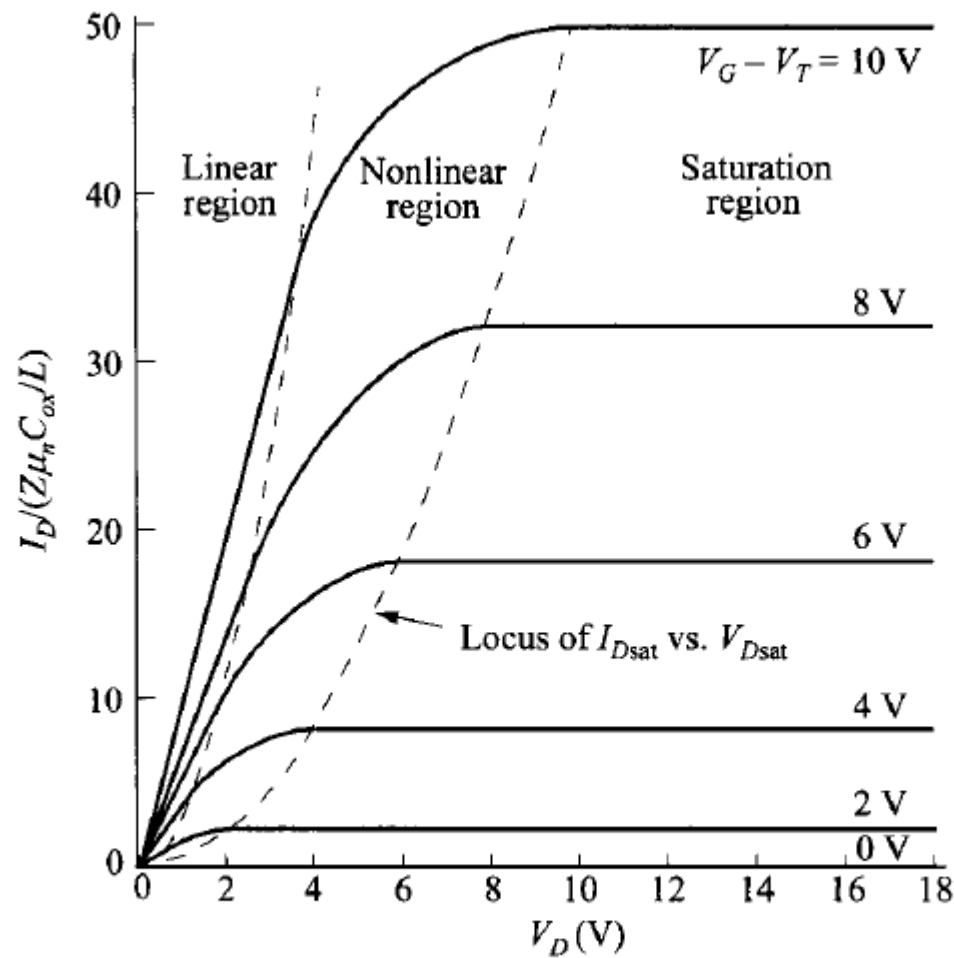
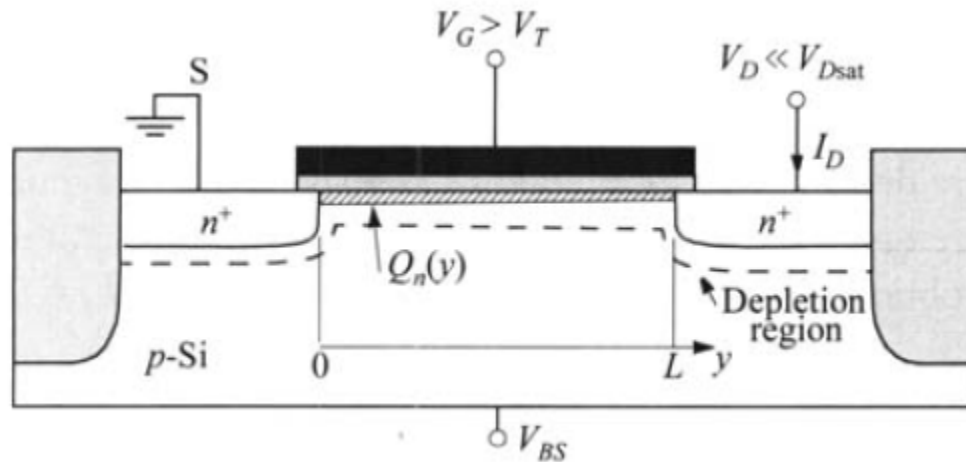
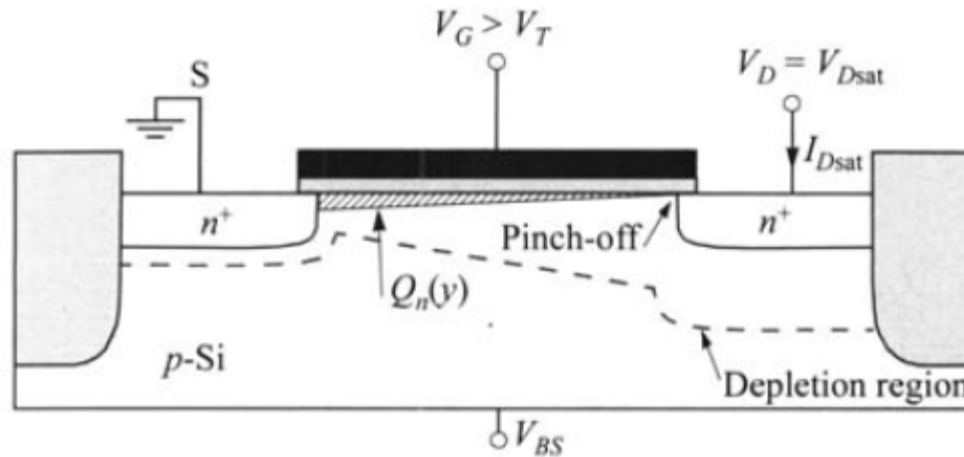


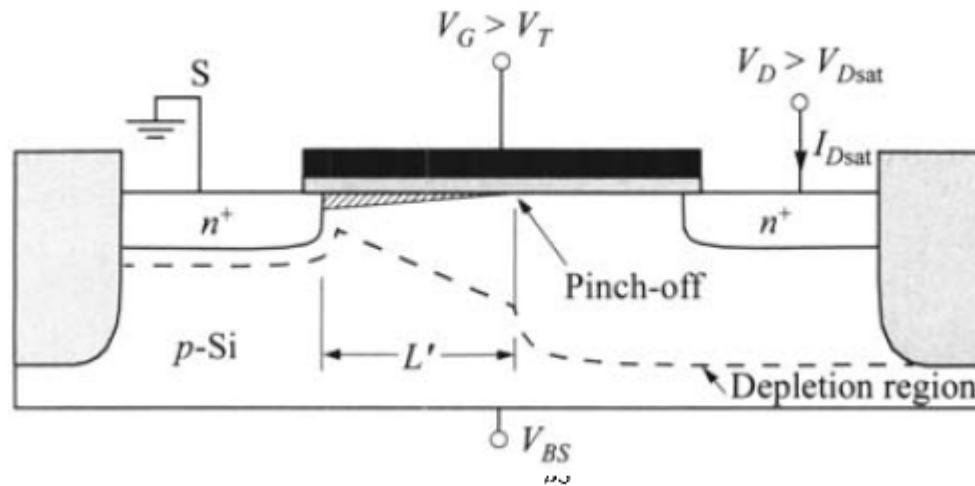
Fig. 9 Idealized drain characteristics (I_D vs. V_D) of a MOSFET. The dashed lines separate the linear, nonlinear, and saturation regions.



- ▶ Let us consider that a positive voltage is applied to the gate, large enough to cause an inversion at the semiconductor surface.
- ▶ If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel.
- ▶ The channel acts as a resistor, and the drain current I , is proportional to the drain voltage V_D .
- ▶ This is the linear region.



- ▶ As the drain voltage increases, the current deviates from the linear relationship .
- ▶ the charge near the drain end is reduced by the channel potential $\Delta\psi_i$.
- ▶ It eventually reaches a point at which the inversion charge at the drain end $Q_n(L)$ is reduced to nearly zero.
- ▶ This location of $Q_n = 0$ is called the pinch-off point.
- ▶ "In reality $Q_n(L)$ is not zero for current continuity, but small because of its high field and high carrier velocity." "



(c)

- ▶ Beyond this drain bias, the drain
- ▶ current remains essentially the same, because for $V_D > V_{Dsat}$, the pinch-off point starts to move toward the source but the voltage at this pinch-off point remains the same (V_{Dsat})
- ▶ the number of carriers arriving at the pinch-off point from the source and hence the current, remains essentially the same, apart from a decrease in L to the value L' .
- ▶ This change of effective channel length will increase the drain current only when the shortened amount is a substantial fraction of the channel length.
- ▶ This pinch-off point occurs because the relative voltage between the gate and the semiconductor is reduced.

- ▶ To deduce the current terms for the operation regions:

$$I_D = \frac{Z\mu_n}{L} C_{ox} \left\{ \left(V_G - V_{FB} - 2\psi_B - \frac{V_D}{2} \right) V_D - \frac{2}{3} \frac{\sqrt{2qNA\epsilon_s}}{C_{ox}} \left[3 \sqrt{\frac{\psi_B}{2}} V_D \right] \right\} = \frac{Z\mu_n}{L} C_{ox} \left(V_G - V_T - \frac{V_D}{2} \right) V_D$$

- ▶ where V_T is the threshold voltage, one of the most-important parameters, given by:

$$V_T = V_{FB} + 2\psi_B + \sqrt{\frac{2\epsilon_s q N_A (2\psi_B)}{C_{ox}}}$$

- ▶ The drain voltage and the drain current at this point are designated as V_{DSAT} and I_{DSAT} .
- ▶ Beyond the pinch-off point the current remains independent of V_D and enters to saturation regions
- ▶ The value of V_{dsat} under the condition $Q_n(L) = 0$

$$V_{Dsat} = \Delta\psi_i(L) = VG - VFB - 2\psi_B + K^2 \left[1 - \sqrt{1 + \frac{2(V_G - VFB)}{K^2}} \right]$$

- ▶ Alternatively, the same solution can be obtained by setting $dI_D/dV_D = 0$.
- ▶ The saturation current I_{DSat}

$$I_{Dsat} = \frac{Z}{2ML} \mu_n C_{ox} (V_G - VT)^2$$

M is a function of doping concentration and oxide thickness

$$M = 1 + \frac{K}{2\sqrt{\psi_B}}$$

- ▶ It has a value slightly larger than unity and it approaches unity with thinner oxide and lower doping.

$$V_{DSAT} = \frac{V_G - VT}{M}$$

- ▶ The transconductance in the saturation region:

$$g_m = \frac{dI_D}{dV_G} = \frac{Z}{ML} \mu_n C_{ox} (V_G - VT)$$

- ▶ Finally, the nonlinear region in between these two extreme cases can be described well by:

$$ID = \frac{Z}{L} \mu_n C_{ox} \left(V_G - VT - \frac{MV_D}{2} \right) V_D$$

- ▶ An approximation of the following form, taking advantage of the definition of threshold voltage, can be made:

$$|Q_n(y)| = C_{ox} [V_G - VT - M\Delta\psi_i(y)]$$

Velocity-Field Relationship

- ▶ As technology advances and pushes for device performance and density, the channel length gets shorter and shorter
- ▶ The internal longitudinal field $E(y)$ in the channel also increases as a result.
- ▶ For low fields, the mobility is constant.

- ▶ This low-field
$$v(\mathcal{E}) = \frac{\mu_n \mathcal{E}}{[1 + (\mu_n \mathcal{E}/v_s)^n]^{1/n}} = \frac{\mu_n \mathcal{E}}{[1 + (\mathcal{E}/\mathcal{E}_c)^n]^{1/n}}$$
- ▶ In the extreme saturation velocity v_s .

- ▶ In between the constant-mobility regime and the saturation-velocity regime, the carrier velocity can be described by:

$$v(E) = \frac{\mu_n E}{\left[1 + \left(\frac{\mu_n E}{v_s}\right)^n\right]^{1/n}} = \frac{\mu_n E}{\left[1 + \left(\frac{E}{Ec}\right)^n\right]^{1/n}}$$

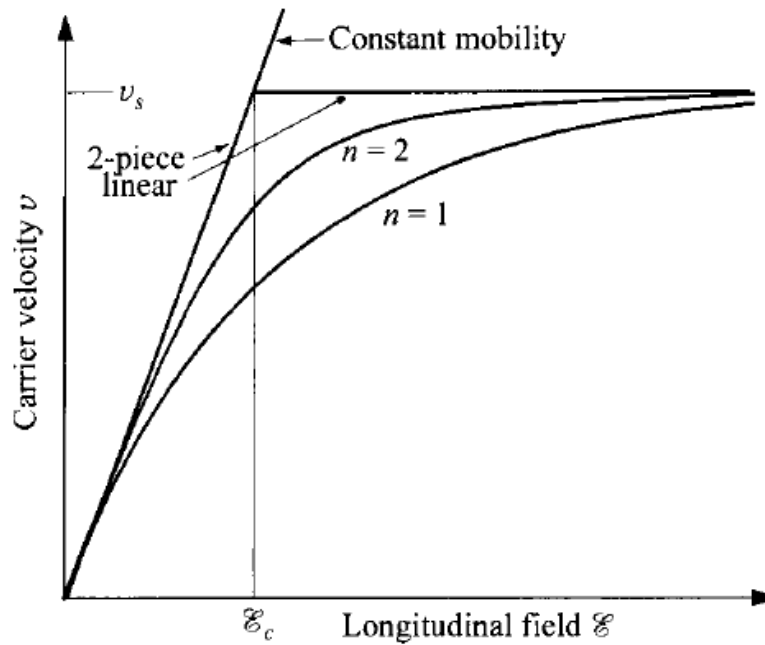


Fig. 11 v - \mathcal{E} relationship (Eq. 33) for $n = 1$ and 2 , and two-piece linear approximation. The critical field $\mathcal{E}_c \equiv v_s / \mu$, where μ is low-field mobility, is also indicated.

- ▶ where μ_n is the low-field mobility.
- ▶ The value of n changes the shape of the curve, but μ_n, v_s and the critical field $E_c (= v_s / \mu_n)$ remain the same.
- ▶ It has been observed that in silicon for electrons $n = 2$ and for holes $n = 1$ have the best fit.
- ▶ As the terminal voltage V , is increased from zero, current is increased because of higher field and higher velocity.
- ▶ the velocity reaches the maximum value of v_s , and the current also saturates to a constant value.
- ▶ it is due to velocity saturation of carriers, before the pinch-off condition can occur.

Field-Dependent Mobility: Two-Piece Linear Approximation

- ▶ In the two-piece linear approximation, the constant-mobility model is valid up to the point when the maximum field near the drain exceeds E_c .

$$I_D(y) = ZCox\mu_n E(V_G - VT - M\Delta\psi_i).$$

- ▶ the drain bias is increased to a value where $E(L) = E_c$.

$$I_D(y) = ZCox\mu_n E_c(V_G - VT - MV_{DSat}).$$

$$I_D(y) = \frac{ZCox\mu_n}{L} \left(V_G - VT - \frac{MV_{DSat}}{2} \right) V_{DSat}$$

$$V_{DSat} = LE_c + \frac{V_G - VT}{M} - \sqrt{(LE_c)^2 + \left(\frac{V_G - VT}{M} \right)^2}$$

- ▶ Since V_{Dsat} here is always smaller than $\frac{V_G - VT}{M}$ the field-dependent mobility always gives a lower I_{Dsat} .

$$I_D \left(E_C + \frac{d\Delta\psi_i}{dy} \right) = ZC_{ox}\mu_n E_C (V_G - VT - M\Delta\psi_i) \frac{d\Delta\psi_i}{dy}$$

$$I_D = \frac{ZC_{ox}\mu_n E_C}{LE_C + VD} \left(V_G - VT - \left(\frac{MV_D}{2} \right) V_D \right)$$

replace L with $L + \frac{V_D}{E_C}$

$$V_{Dsat} = LE_C \left[\sqrt{1 + \frac{2(V_G - VT)}{MLE_C}} - 1 \right]$$

Velocity Saturation.

- ▶ it is interesting and insightful to look at the extreme case of short-channel devices where velocity saturation completely limits the current flow.
- ▶ In such case we set $v = v_s$, consequently
- ▶ Q_n has to be fixed for current continuity, is approximated to be $(V_G - V_T)C_{ox}$.
- ▶ $I_{Dsat} = \frac{Z}{L} \int_0^L |Q_n(y)| v(y) dy = \frac{Z}{L} |Q_n| v_s L = Z(V_G - VT)C_{ox}v_s$
- ▶ The transconductance becomes:

$$g_m \equiv \frac{dI_{DSat}}{dV_G} \equiv ZCOXv_s$$

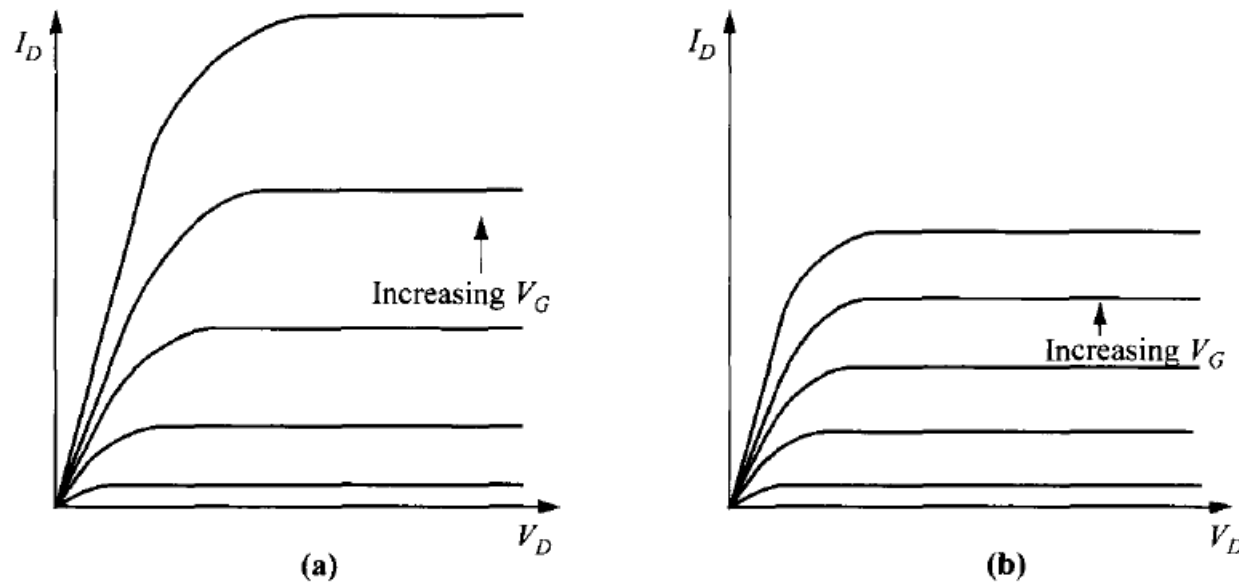
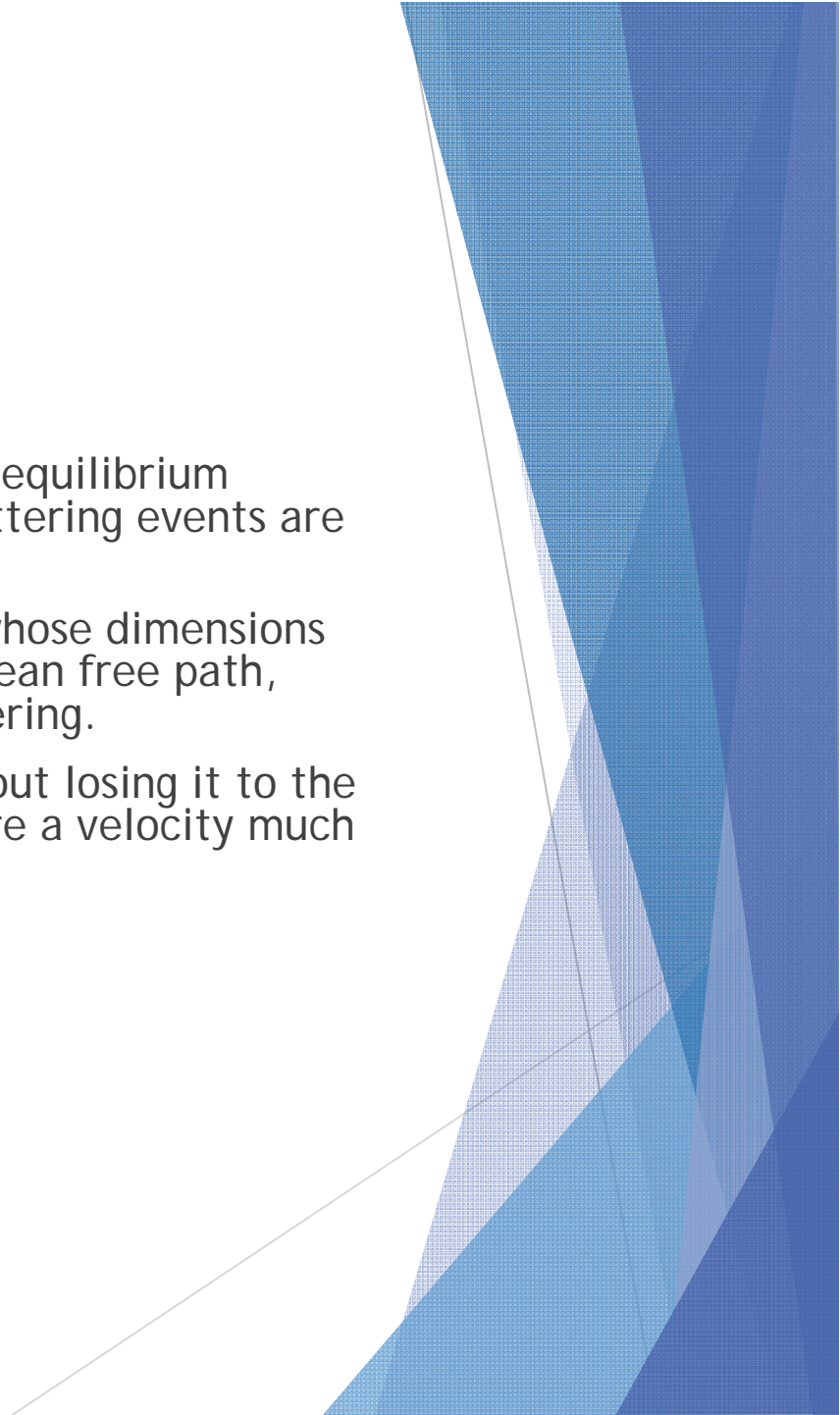
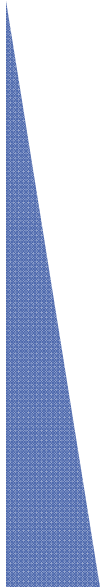


Fig. 12 Comparison of I - V characteristics for (a) constant mobility and (b) velocity saturation. All other parameters are the same.

- ▶ First I_{Dsat} and V_{Dsat} are both lowered by velocity saturation, while the linear regions remain similar.
- ▶ The g_m (which is the current difference between V_G steps) also becomes a constant, independent of V_G
- ▶ shows an interesting phenomena that the saturation current no longer depends on the channel length.

Ballistic Transport.

- ▶ The velocity saturation is a steady-state, equilibrium phenomena at high field, when many scattering events are allowed to happen.
- ▶ However, in ultra-short channel lengths whose dimensions are on the order of or shorter than the mean free path, channel carriers do not suffer from scattering.
- ▶ They can gain energy from the field without losing it to the lattice through scattering, and can acquire a velocity much higher than the saturation velocity.



- ▶ This effect is called ballistic transport
- ▶ The ballistic transport is important since it points out that the current and transconductance can be higher than that of saturation velocity, giving an additional incentive for shrinking the channel length.
- ▶ At positions closer to the source, the velocity decreases.

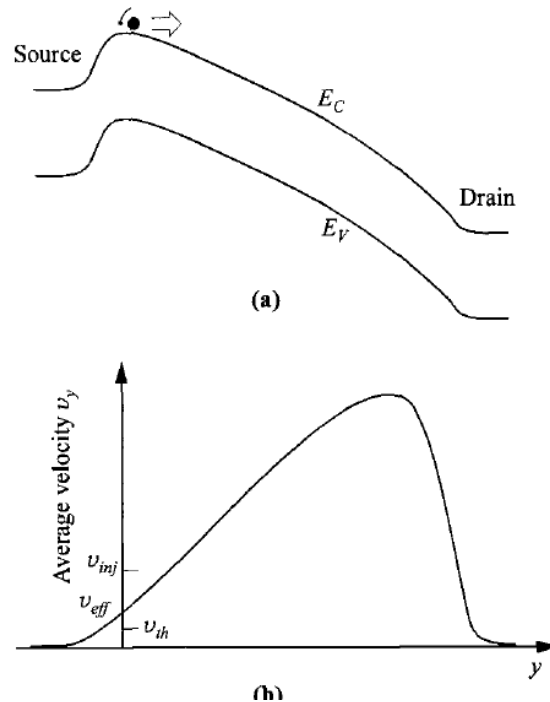


Fig. 13 (a) Under a drain bias, the potential maximum is the bottleneck for the current flow and is used to calculate the current. (b) Average carrier velocity (y -component) as a function of the channel position. Note that v_{eff} is between the values of v_{inj} and v_{th} , and that the maximum velocity near the drain can be higher than v_{inj} .

- ▶ In order to have current continuity, the channel potential and inversion charge must adjust themselves such that the product of velocity and charge would remain constant throughout the channel.
- ▶ The bottleneck for the current flow, at the extreme of ultra-short channel length, would be at the position of maximum charge and minimum field.
- ▶ which means the potential maximum near the source end.

$$I_{Dsat} = Z|Q_n|v_{eff}$$

- ▶ where $|Q_n|$ has the maximum value at the source as $C_{ox}(V_G - V_T)$, and v_{eff} is an effective average carrier velocity that should match the final experimental saturation current.
- ▶ The only critical parameter is v_{eff} .
- ▶ The maximum value of v_{eff} according to classical thermal equilibrium, is simply the thermal velocity $v_{eff} [= (2kT/\pi m^*)^{1/2}]$
- ▶ This is a quantum-mechanical effect, called carrier degeneration where the mean carrier energy is pushed to a higher state than the thermal energy.
- ▶ This higher value is called the injection velocity v_{inj} .
- ▶ the Fermi energy with respect to the quantized energy E_n inside the potential well where carriers reside:

$$v_{inj} = \sqrt{\frac{2KT}{\pi m^*} \frac{F_{1/2}\left[\frac{E_F - E_n}{KT}\right]}{\ln(1 + e^{(E_F - E_n)/KT})}}$$

- ▶ where $F_{1/2}$ is the Fermi-Dirac integral
- ▶ Small inversion charge ($E_F - E_n$) reduced to $\sqrt{\frac{2KT}{\pi m^*}}$ and $v_{inj} = v_{th}$

- ▶ If the inversion charge is high

$$v_{inj} = \frac{8h}{3m^*} \sqrt{\frac{|Q_n|}{2\pi q}} = \frac{8h}{3m^*} \sqrt{\frac{C_{ox}(VG - VT)}{2\pi q}}$$

- ▶ Which is a function of the inversion charge or gate overdrive.
- ▶ The maximum current, which is a product of $Q_n v_{inj}$:

$$ID_{sat} = r_n Z |Q_n| v_{inj} = \frac{8r_n Z h}{3m^*} \left[\frac{C_{ox}(VG - VT)}{\sqrt{2\pi q}} \right]^{\frac{3}{2}}$$

- ▶ where r_n is the *index of ballisticity* ($= v_{eff} / v_{inj}$). In the extreme of ballistic transport, $r_n = 1$, and it sets the ultimate current drive for L $\rightarrow 0$.



- ▶ The trans-conductance is given by:

$$g_m = \frac{4 r n h}{2 \pi m^*} \sqrt{\frac{C_{ox}(V_G - V_T)}{2 \pi q}}$$

- ▶ It is seen here that both I_{Dsat} and g_m *are* independent of channel length L.
- ▶ The index of ballisticity is also interpreted by back scattering R of channel carriers at the drain back to the source.

$$r_n = \left(\frac{v_{eff}}{v_{inj}} \right) = \frac{1 - R}{1 + R} = \left[\frac{1}{v_{inj}} + \left(\frac{1}{\mu_n E(0)} \right) \right]^{-1}$$

- ▶ where $E(0)$ is the field at a potential kT down from the maximum toward the drain.

6.2.3 Threshold Voltage

- ▶ To account for the threshold shift from nonzero flat-band voltage whose main cause comes from fixed oxide charges Q_f and the work-function difference ϕ_{ms} .

$$V_T = V_{FB} + 2\psi_B + \sqrt{\frac{2\varepsilon_s q N_A (2\psi_B)}{C_{ox}}} = \left(\phi_{ms} - \frac{Q_f}{C_{ox}} \right) + 2\psi_B + \sqrt{\frac{4\varepsilon_s q N_A (\psi_B)}{C_{ox}}}$$

- ▶ V_T is the gate bias beyond flat-band just starting to induce an inversion charge sheet
- ▶ the sum of voltages across the semiconductor $2\psi_B$ and oxide layer (Q_f) charge.
- ▶ The square-root term is the total depletion-layer charge.

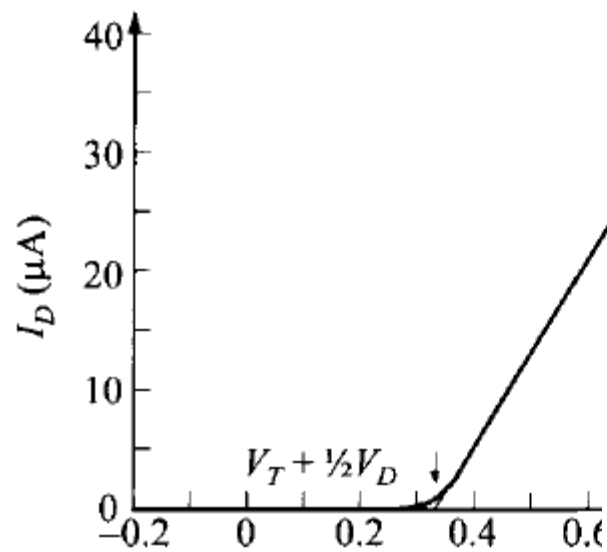
- ▶ When a substrate bias is applied (negative for n-channel or p-substrate), the threshold voltage becomes:

$$V_T = V_{FB} + 2\psi_B + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B - V_{BS})}}{C_{ox}}$$

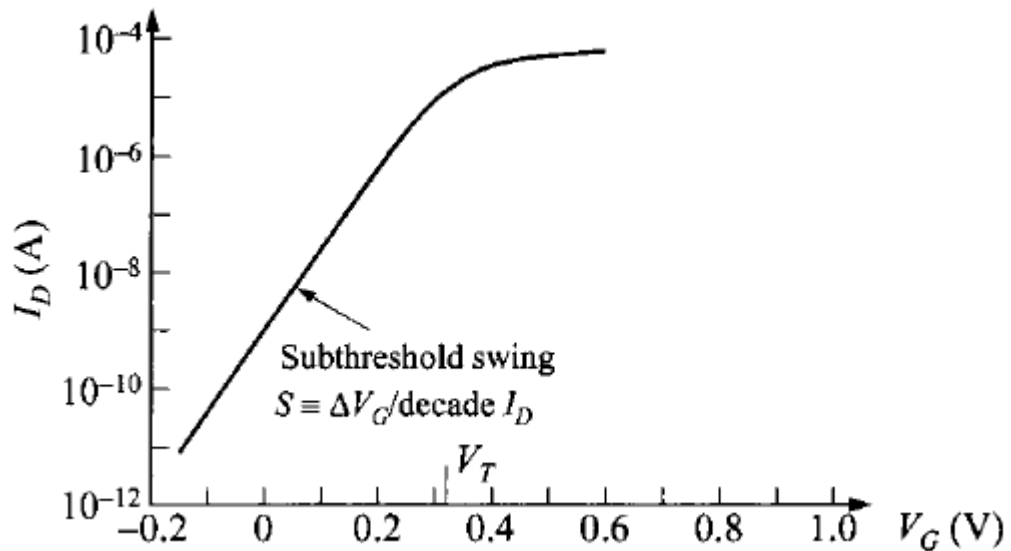
- ▶ it is shifted by the amount of:

$$\Delta V_T = V_T(V_{BS}) - V_T(V_{BS} = 0) = \frac{\sqrt{2\varepsilon_s q N_A}}{C_{ox}} (\sqrt{2\psi_B - V_{BS}} - \sqrt{2\psi_B})$$

- ▶ In practice it is often necessary to minimize this threshold-voltage shift due to substrate bias.
- ▶ low substrate doping and thin oxide thickness are preferred.



(a)



(b)

Fig. 15 Transfer characteristics (I_D vs. V_G) in the linear region ($V_D \ll V_G$). (a) I_D in linear scale to deduce V_T . Deviation from linearity at higher V_G is due to lower mobility. (b) I_D in logarithmic scale to show subthreshold swing.

6.2.4 Subthreshold Region

- ▶ When the gate bias is below the threshold and the semiconductor surface is in weak inversion or depletion.
- ▶ The corresponding drain current is called the subthreshold current.
- ▶ The subthreshold region tells how sharply the current drops with gate bias
- ▶ the MOSFET is used as a switch in digital logic and memory applications.
- ▶ In weak inversion and depletion, the electron charge is small.
- ▶ the drift current is low.
- ▶ The drain current is dominated by diffusion and is derived in the same way as the collector current in a bipolar transistor.
- ▶ Considering the electron-density gradient in the channel, the diffusion current is given by

$$I_D = -ZqDn \frac{dN'(y)}{dy} \approx ZqDn \frac{N'(0) - N'(L)}{L}$$

where N' is the electron density per unit area, integrated over the depletion width.

$$N'(0) \approx \left(\frac{1}{\beta}\right) \sqrt{\frac{\epsilon_s}{2q\psi_s N_A}} n_{p0} e^{\beta\psi_s}$$

- ▶ Similar result can be obtained by assuming an effective thickness (x_i) of the surface charge layer.
- ▶ the exponential dependence of electron density on the potential ψ_p , x_i corresponds to the distance in which ψ_p decreases by KT/q
- ▶ The electron density at the drain end is lowered exponentially by the drain bias.

$$N'(L) = N'(0)e^{-\beta V_D}$$

$$I_D = \frac{Z\mu_n}{L\beta^2} \sqrt{\frac{q\epsilon_s N_A}{2\psi_s}} \left(\frac{n_i}{N_A}\right)^2 e^{\beta\psi_s} [1 - e^{-\beta V_D}] \cong \frac{Z\mu_n}{L\beta^2} \sqrt{\frac{q\epsilon_s N_A}{2\psi_s}} \left(\frac{n_i}{N_A}\right)^2 e^{\beta\psi_s}$$

- ▶ The subthreshold region the drain current varies exponentially with ψ_s .
- ▶ for drain voltage V_D larger than $= 3KT/q$, the current becomes independent of V_D .

$$V_G - V_{FB} = \psi_s + \frac{\sqrt{2\epsilon_s \psi_s q N_A}}{C_{ox}}$$

This quadratic equation will not give a simple expression of ψ_s as a function of

- ▶ The parameter to quantify how sharply the transistor is turned off by the gate voltage is called the subthreshold swing S (inverse of subthreshold slope)
- ▶ the relative change of V_G and ψ_s :

$$\frac{dV_G}{d\psi_s} = 1 + \frac{1}{C_{ox}} \sqrt{\frac{\epsilon_s q N_A}{2\psi_s}} = \frac{C_{ox} + CD}{C_{ox}}$$

- ▶ By definition, the subthreshold swing can now be calculated:

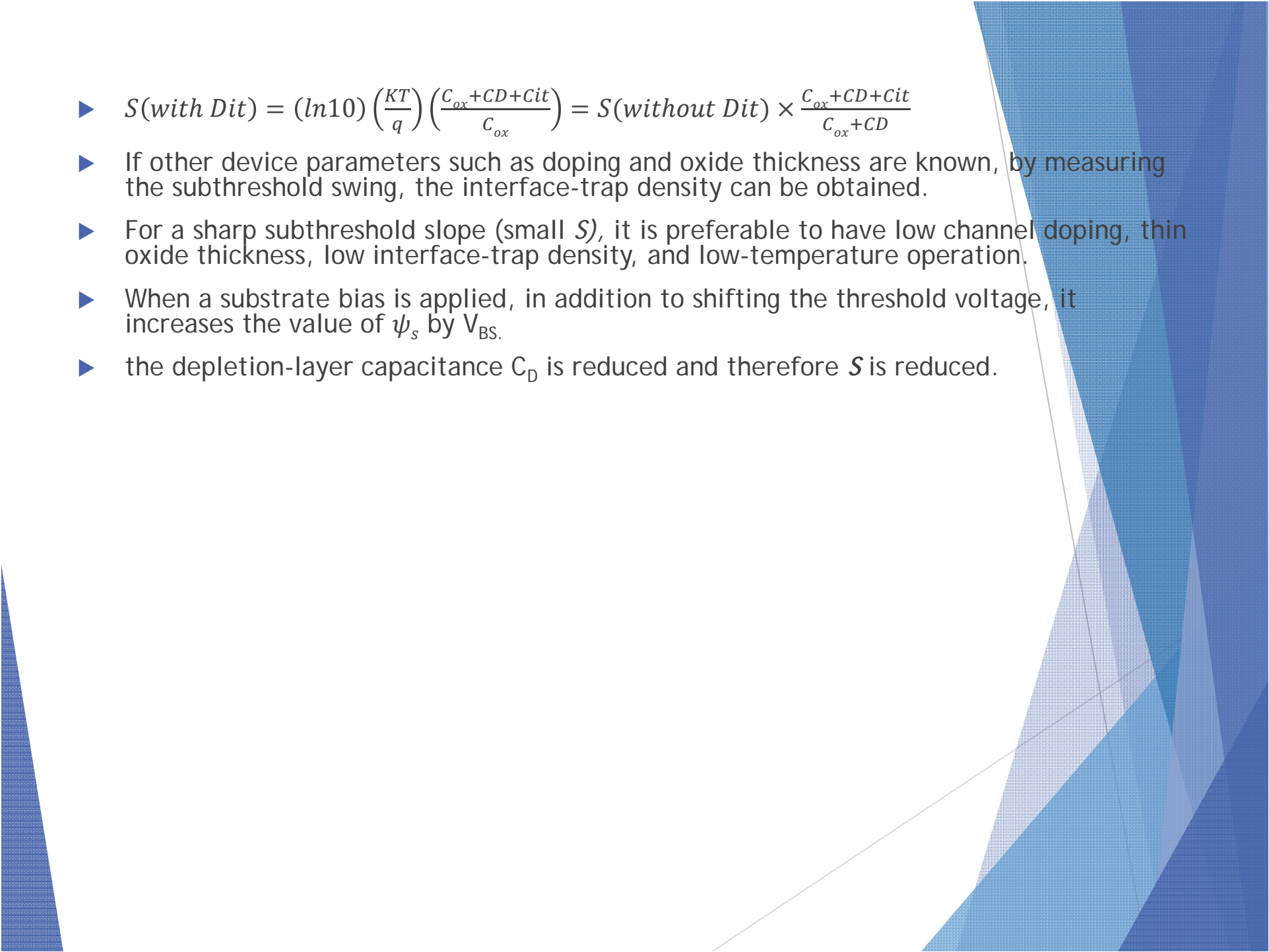
$$S \equiv (\ln 10) \frac{dV_G}{d(\ln I_D)} = (\ln 10) \frac{dV_G}{d(\beta\psi_s)} = (\ln 10) \left(\frac{KT}{q} \right) \frac{C_{ox} + CD}{C_{ox}}$$

- ▶ In the extreme of zero oxide thickness, the exponential characteristics are identical to the familiar case of the diffusion current in a p-n junction.

- ▶ For nonzero oxide thickness, the swing is just degraded by a factor which is a voltage divider of two capacitors in series, whose ratio is

$$\frac{C_{ox} + C_D}{C_{ox}}$$

- ▶ One also notices that since the depletion width (and C_D) varies with ψ_s the subthreshold swing is a weak function, but not exactly constant with V_G .
- ▶ In the presence of a significant interface-trap density D_{it} .
- ▶ C_{it} ($= q^2 D_{it}$) is in parallel with the depletion-layer capacitance C_D .

- 
- ▶ $S(\text{with } Dit) = (\ln 10) \left(\frac{KT}{q} \right) \left(\frac{C_{ox} + CD + Cit}{C_{ox}} \right) = S(\text{without } Dit) \times \frac{C_{ox} + CD + Cit}{C_{ox} + CD}$
 - ▶ If other device parameters such as doping and oxide thickness are known, by measuring the subthreshold swing, the interface-trap density can be obtained.
 - ▶ For a sharp subthreshold slope (small S), it is preferable to have low channel doping, thin oxide thickness, low interface-trap density, and low-temperature operation.
 - ▶ When a substrate bias is applied, in addition to shifting the threshold voltage, it increases the value of ψ_s by V_{BS} .
 - ▶ the depletion-layer capacitance C_D is reduced and therefore S is reduced.

- ▶ Near the threshold voltage, the drain current does not turn off as sharply.
- ▶ This is due to diffusion current which is the dominant current near and below threshold.
- ▶ The total drain current density including both drift and diffusion components is given by:

$$J_D(x, y) = q\mu_n n E_y + qDn \frac{dn}{dy} = Dn n(x, y) \frac{dE_{Fn}}{dy}$$

- ▶ The drain current based on the gradual-channel approximation is:

$$\begin{aligned} ID &= Z \int_0^{x_l} JD(x, y) dx = \frac{ZDn}{L} \int_0^L \frac{dE_{Fn}}{dy} \int_0^{x_l} n(x, y) dx dy \\ &= \frac{Z\epsilon_s \mu_n}{LL_D} \int_0^{V_D} \int_{\psi_B}^{\psi_s} \frac{e^{(\beta\psi_p - \beta\Delta\psi_i)}}{F(\beta\psi_p, \Delta\psi_i, \frac{n_{po}}{p_{po}})} d\psi_p d\Delta\psi_i \end{aligned}$$

- ▶ The gate voltage V_G is related to the surface potential ψ_s by:

$$V_G - V_{FB} = -\frac{Q_s}{C_{ox}} + \psi_s = \frac{2\varepsilon_s K T}{C_{ox} q L_D} F\left(\beta\psi_s, \Delta\psi_i, \frac{n_{po}}{p_{po}}\right) + \psi_s$$

- ▶ The V_D , V_G and I_D calculated numerically to give accurate results from the linear region to the saturation region.

6.2.5 Mobility Behavior

- ▶ The channel carriers are confined to a thin inversion layer.
- ▶ The drift velocity v and mobility μ are expected to be influenced by the thickness of this inversion layer.
- ▶ When a small longitudinal field E_y is applied (parallel to the semiconductor surface).
- ▶ The drift velocity varies linearly with E_y and the proportionality constant is the low field-mobility.
- ▶ The Si inversion layers show that this low field mobility, while independent of E_y is a unique function of the transverse field E_x that is perpendicular to the current.
- ▶ This dependence is not directly on the oxide thickness or doping density, but through their impact of E_x in the inversion layer.

- ▶ When many devices with different oxide thicknesses and doping levels are measured, the mobility is found to correlate well with a single parameter that is related to E_x .
- ▶ The mobility decreases with an increasing *effective* transverse field.
- ▶ The field averaged over the electron distribution in the inversion layer:

$$(E_x)_{eff} = \frac{1}{\epsilon_s} \left(Q_B + \frac{1}{2} Q_n \right)$$

Depend on half of the inversion-layer charge Q_n and depletion-layer charge Q_B .

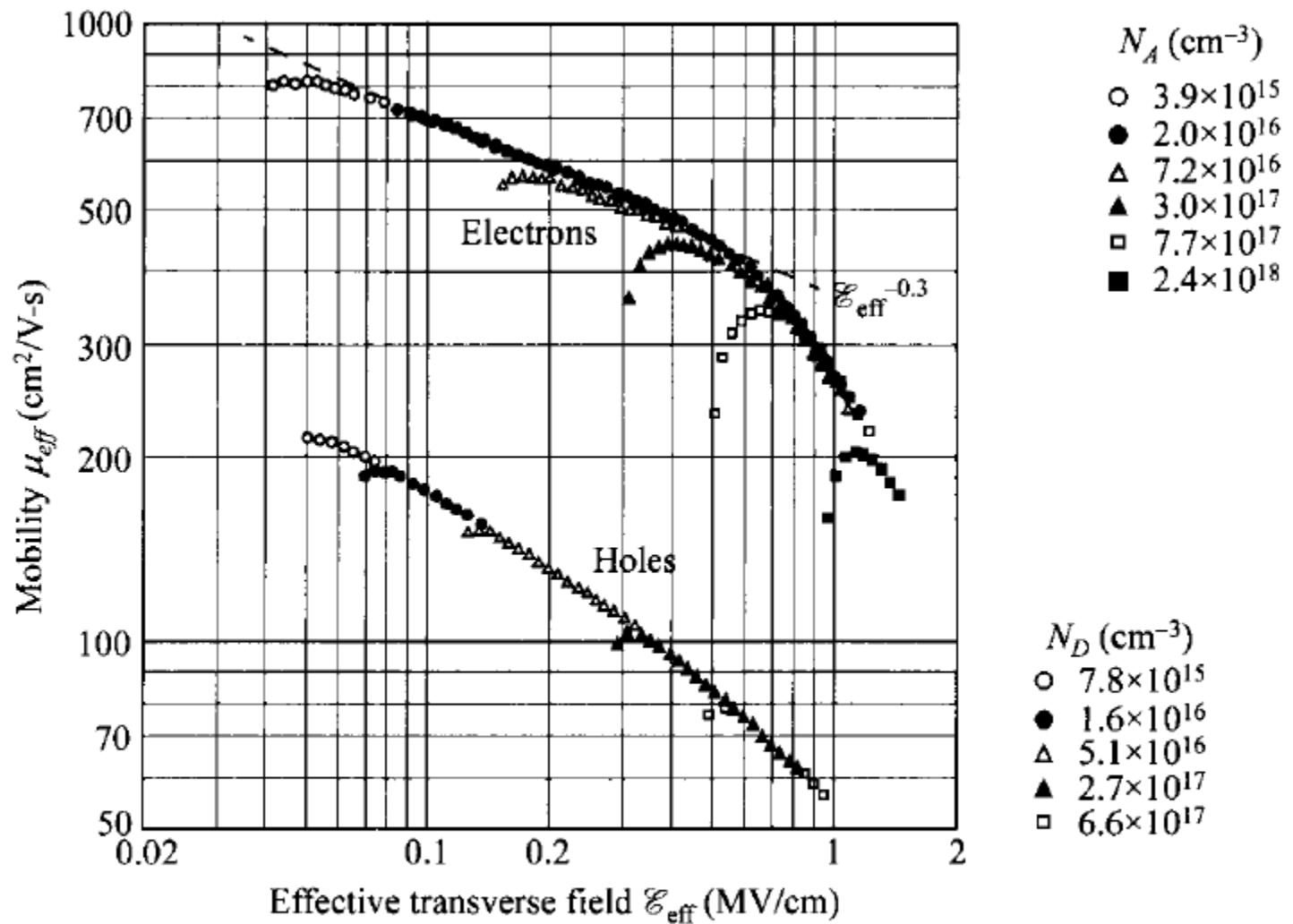


Fig. 16 Electron and hole inversion-layer mobilities vs. effective transverse field, at room temperature on Si (100) surface. (After Ref. 33.)

- ▶ When the longitudinal field increases, the $E-v$ relationship starts to deviate from linearity.
- ▶ The mobility at any field is defined as the ratio of v/E_y .
- ▶ It can be seen here also that the saturation velocity v_s is independent of the low-field mobility

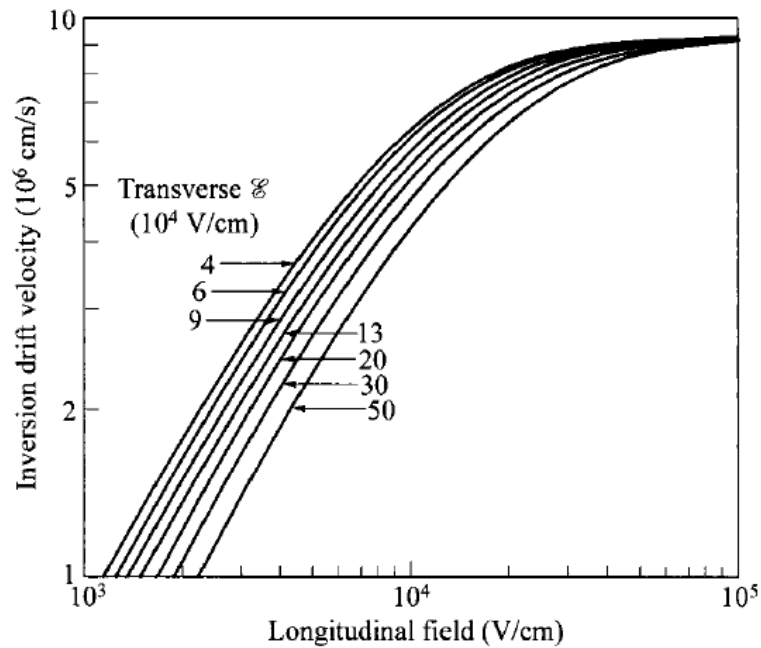


Fig. 17 Electron surface drift velocity vs. longitudinal field for various transverse fields. The slope at low longitudinal field is mobility. (After Ref. 34.)

6.2.6 Temperature Dependence

- ▶ Temperature affects device parameters and performance, in particular mobility, threshold voltage, and subthreshold characteristics.
- ▶ The effective mobility in inversion layer has a T^{-2} dependence on temperatures around 300 K at gate biases corresponding to strong inversion.

This gives rise to high V_T vs V_T :

$$V_T = \phi_{ms} - \frac{Q_f}{C_{ox}} + 2\psi_B + \frac{\sqrt{4\epsilon_s q N_A \psi_B}}{C_{ox}} \quad \text{re.}$$

- ▶ Because the work-function difference ϕ_{ms} and the fixed oxide charges are essentially independent of temperature.

$$\frac{dV_T}{dT} = \frac{d\psi_B}{dT} \left(2 + \frac{1}{C_{ox}} \sqrt{\frac{\epsilon_s q N_A}{\psi_B}} \right)$$

From the basic equations of:

$$\psi_B = \frac{KT}{q} \ln\left(\frac{N_A}{n_i}\right)$$

$$n_i^2 \propto T^3 e^{-\frac{E_{g0}}{KT}}$$

where E_{g0} is the energy gap at $T = 0$.

$$\frac{d\psi_B}{dT} \approx \frac{1}{T} \left(\psi_B - \frac{E_{g0}}{2q} \right)$$

the quantity $|dV/dT|$ can increase or increase with the substrate doping.

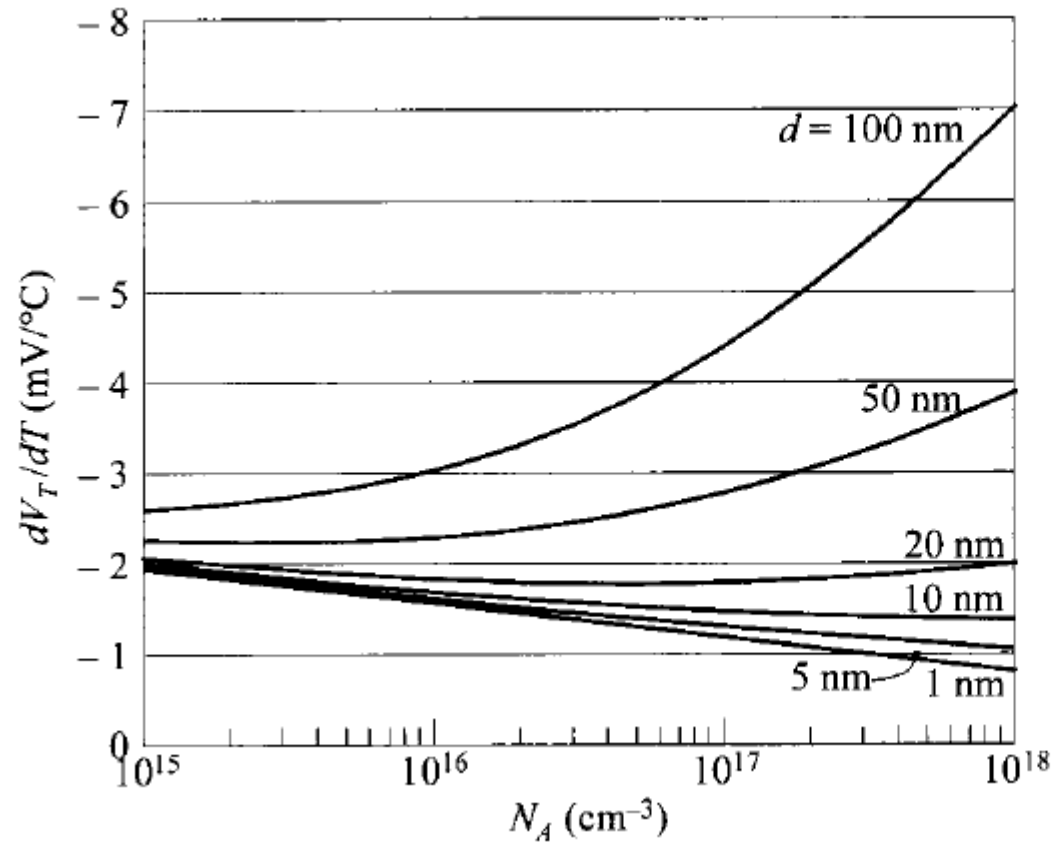


Fig. 18 Threshold-voltage shift (dV_T/dT) of a Si-SiO₂ system at room temperature vs. substrate doping, with oxide thickness d as a parameter.

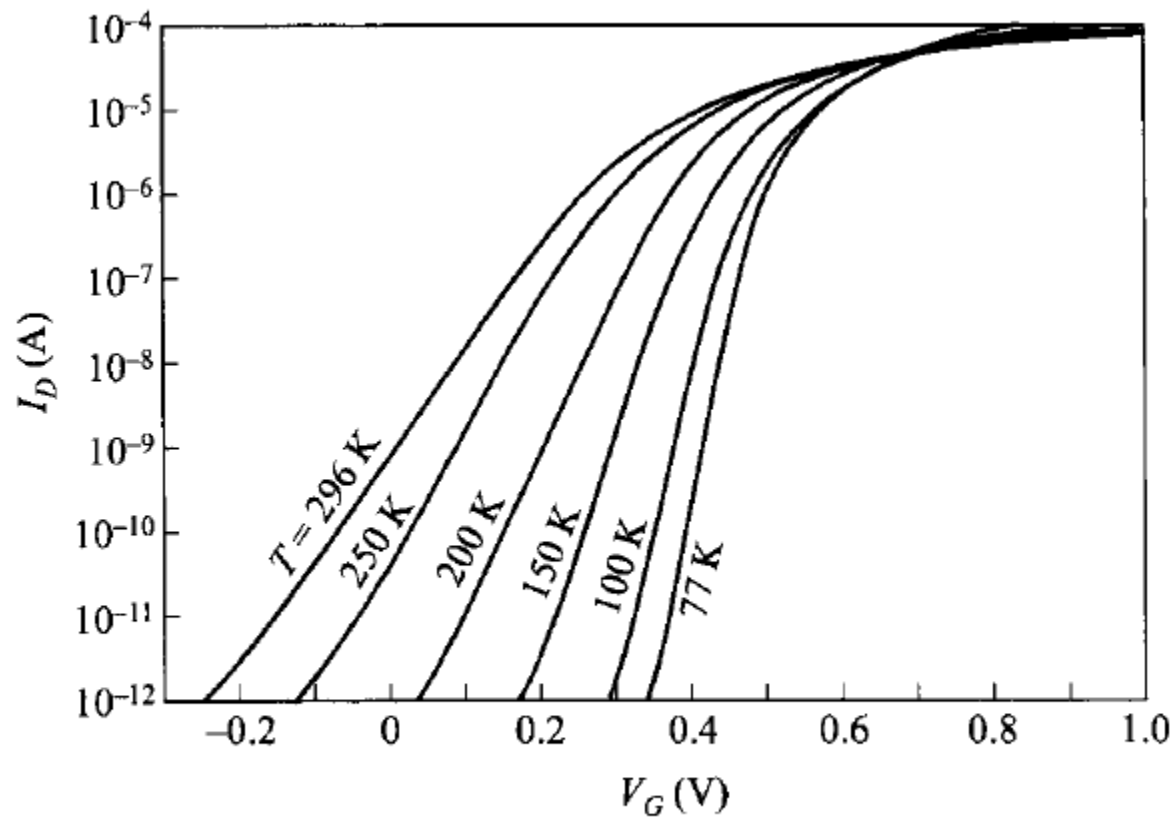


Fig. 19 Subthreshold characteristics for a long-channel MOSFET ($L = 9 \mu\text{m}$) with temperature as a parameter. (After Ref. 36.)

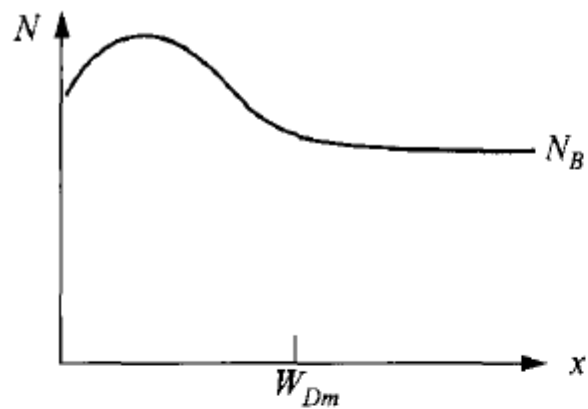
6.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICE

- ▶ We consider next the effect of nonuniform channel doping on device characteristics, especially on threshold voltage and depletion width which in turn affects subthreshold swing and the substrate-bias effect.
- ▶ Note that what is most important for determining V_T is the doping profile within the depletion region.
- ▶ The profile outside the depletion is important for considerations of capacitance and substrate sensitivity.

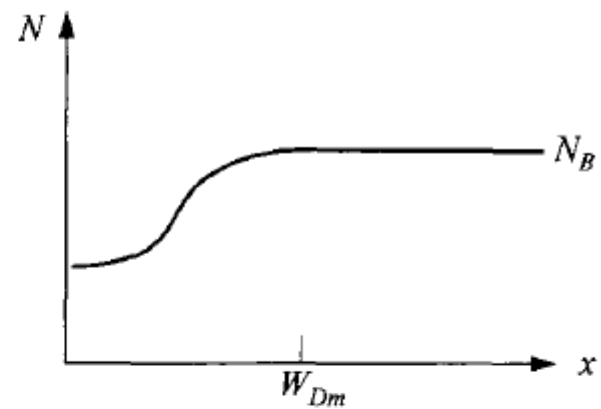
$$\begin{aligned} V_T &= V_{FB} + \psi_s + \frac{Q_B}{C_{ox}} \\ &= V_{FB} + 2\psi_B + \frac{q}{C_{ox}} \int_0^{W_{DM}} N(x) dx \end{aligned}$$

where Q_B is the depletion-layer charge.

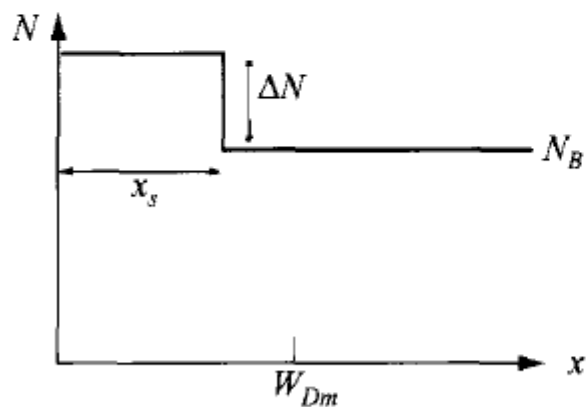
The limit for integration, that is the maximum depletion width W_{DM}



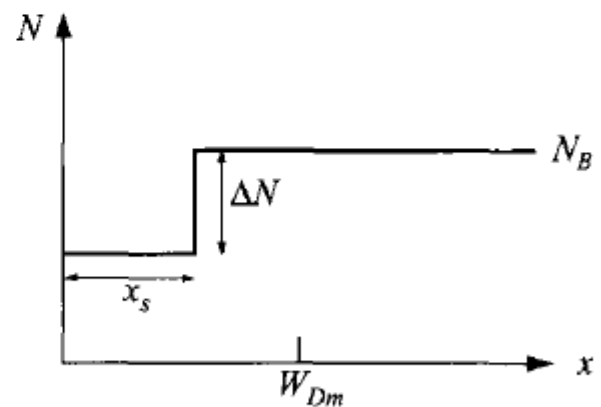
(a)



(b)



(c)



(d)

Fig. 20 Nonuniform channel doping profiles. (a) High-low profile. (b) Low-high (retrograde) profile. (c) – (d) Their approximations using step profiles.

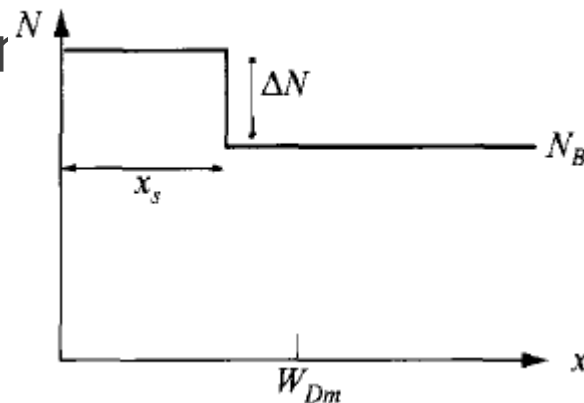
- ▶ the Poisson equation with the onset of strong inversion being the boundary condition:

$$\psi_s = 2\psi_B = \frac{q}{\epsilon_s} \int_0^{W_{DM}} xN(x)dx$$

- ▶ Note that for a nonuniform profile, the definitions of ψ_B and V_{FB} become nontrivial and complicated.
- ▶ the background doping of N_B for these values is found to be sufficiently accurate.

6.3.1 High-Low Profile

- ▶ To derive the threshold voltage shift due to ion implantation.
- ▶ we shall consider an idealized step profile
- ▶ The implant profile, after thermal anneal, is approximated by the step function with step depth x_s .
- ▶ The equal to the sum of the projected range and the standard deviation of the original implant.
- ▶ For a wider x_s the depletion-layer width W_{DM} under N_B is within x_s .



- ▶ The surface region can be considered a uniformly doped region with a higher concentration.
- ▶ If $W_{DM} > x_s$ the threshold voltage

$$\begin{aligned}
 V_T &= V_{FB} + \psi_s + \frac{Q_B}{C_{ox}} \\
 &= V_{FB} + 2\psi_B + \frac{q}{C_{ox}} \int_{x_s}^{W_{DM}} N(x) dx \\
 &= V_{FB} + 2\psi_B + \frac{qN_B W_{DM} + q\Delta N x_s}{C_{ox}} \\
 &= V_{FB} + 2\psi_B + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_B \left(2\psi_B - \left(\frac{q\Delta N x_s^2}{2\epsilon_s}\right)\right)} + \frac{q\Delta N x_s}{C_{ox}}
 \end{aligned}$$

$\psi_s = 2\psi_B$ for strong inversion:

$$W_{DM} = \sqrt{\frac{2\epsilon_s}{qN_B} \left(2\psi_B - \frac{q\Delta N x_s^2}{2\epsilon_s}\right)}$$

- ▶ The V_T shift is largest with the added doping closest to the surface.
- ▶ For the limiting case of a delta function of dose localized at the Si-SiO₂ interface ($x_s = 0$), the threshold shift is simply

$$\Delta V_T \approx \frac{qD_I}{C_{ox}}$$

- ▶ where D_I is the total dose $\Delta N x_s$

Such approach is called threshold adjust, which has the same effect as changing the work-function difference $q\phi_{ms}$ or changing the total fixed oxide charge.

- ▶ The step-profile approach described above can give first-order results for the threshold voltage.
- ▶ To obtain a more accurate V_T we have to consider the actual doping profile, because the step width x_s is not well defined for nonuniform doping.

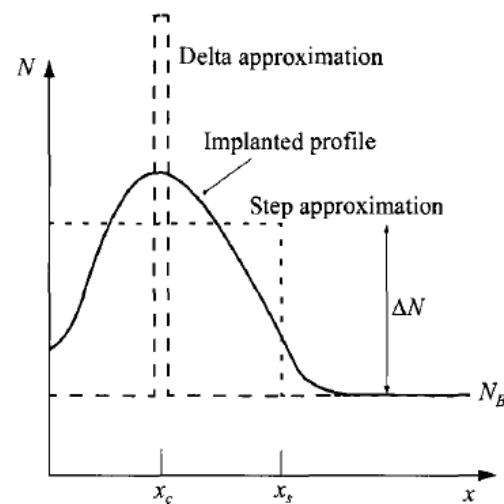


Fig. 21 Approximation of an actual implanted profile by step and delta profiles.

- ▶ For a typical case, the threshold voltage depends on the implanted dose D_I and the centroid of the dose x_c .

$$D_I = \int_0^{W_{DM}} \Delta N(x) dx$$

$$x_c = \frac{1}{D_I} \int_0^{W_{DM}} x \Delta N(x) dx$$

$$V_T = V_{FB} + 2\psi_B + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_B \left(2\psi_B - \frac{qx_c D_I}{\epsilon_s} \right)} + \frac{qD_I}{C_{ox}}$$

$$W_{DM} = \sqrt{\frac{2\epsilon_s}{qN_B} \left(2\psi_B - \frac{qD_I x_c}{\epsilon_s} \right)}$$

- ▶ As x_c increases, the dose becomes less effective in changing V , and the depletion width W_{DM} decreases also at the same time.
- ▶ The condition for which x_c starts to be equal to W_{DM}
- ▶ Where W_{DM0} is the original W_{DM} with background doping N_B .

$$D_I(x_c = W_{DM}) = \frac{N_B(W_{DM0}^2 - x_c^2)}{2x_c},$$

- ▶ Eventually, as x_c moves beyond the W_{Dm0} it is no longer has any effect on threshold voltage and depletion width.
- ▶ we have interpreted the subthreshold swing by comparing the gate-oxide capacitance C_{ox} to depletion capacitance C_D .
- ▶ once the depletion width is known, the subthreshold swing can be calculated.
- ▶ For the high-low profile, the added doping decreases W_{DM} increases C_D
- ▶ The substrate sensitivity can be calculated also by substituting $2\psi_B$ with $2\psi_B + V_{BS}$ in calculating V_T

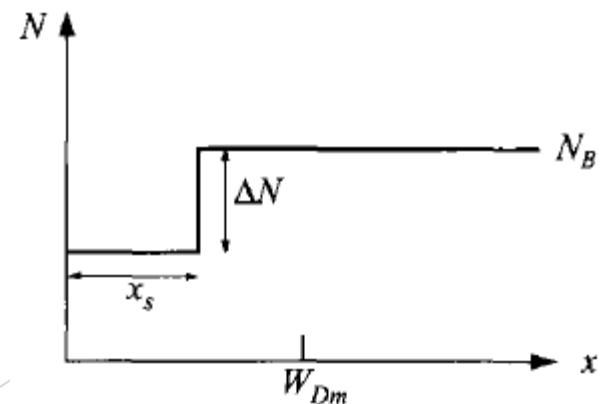
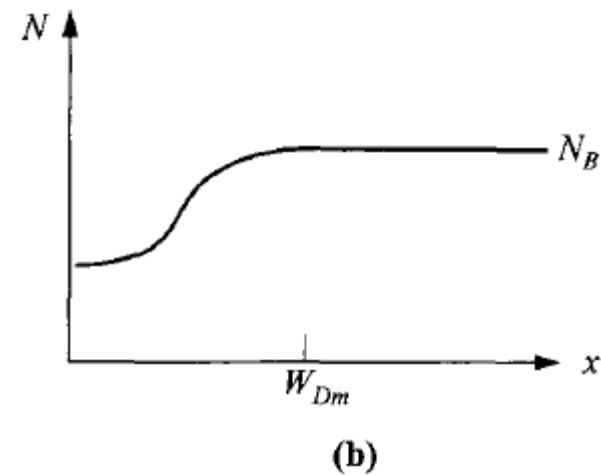
6.3.2 LOW-HIGH PROFILE

- ▶ Analysis of the low-high profile also called the retrograde profile is similar to the high-low case with a ΔN being subtracted from the background doping.

$$V_T = V_{FB} + 2\psi_B + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_B \left(2\psi_B + \frac{q\Delta N x_s^2}{2\epsilon_s} \right) - \frac{q\Delta N x_s}{C_{ox}}}$$

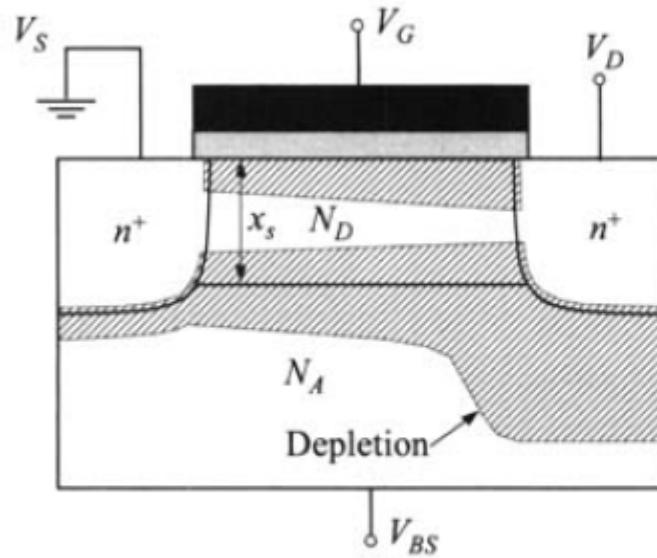
$$W_{DM} = \sqrt{\frac{2\epsilon_s}{qN_B} \left(2\psi_B + \frac{q\Delta N x_s^2}{2\epsilon_s} \right)}$$

The threshold voltage is, thus, decreased and the depletion width is increased by a dip at the surface doping.

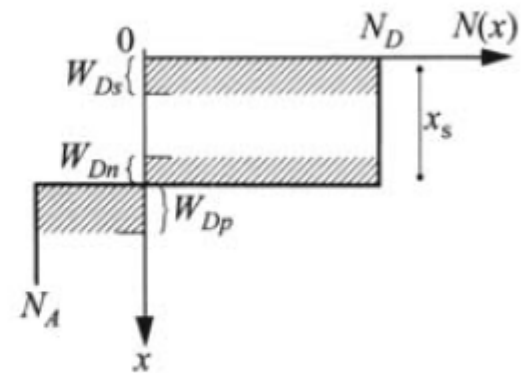


6.3.3 Buried-Channel Device

- ▶ the surface doping can be of the opposite type of the substrate. When this happens, and if part of the surface doped layer is not fully depleted
- ▶ there exists some neutral region, current can conduct through this buried layer.
- ▶ The gate voltage can change the surface depletion layer, thus controlling the net opening of the channel thickness and controlling the current flow.
- ▶ With a large positive gate bias, the channel is fully open, and an addition surface inversion layer can be induced at the surface, similar to a regular surface channel, resulting in two channels in parallel.



(a)



(b)

Fig. 22 (a) Schematic of a buried-channel MOSFET under bias. (b) Its doping profile and depletion regions.

- ▶ The net channel thickness is reduced from x_s by the amounts of surface depletion W_{DS} and the bottom p-n junction depletion W_{Dn}

$$W_{DS} = \sqrt{\frac{2\varepsilon_s}{qN_D} (V_{FB}^* - VG) + \left(\frac{\varepsilon_s}{C_{ox}}\right)^2} - \frac{\varepsilon_s}{C_{ox}}$$

$$V_{FB}^* = V_{FB} + \psi_{bi}$$

- ▶ where V_{FB} keeps the p-substrate as reference. The bottom depletion width is from the p-n junction theory, given by:

$$W_{Dn} = \sqrt{\frac{2\varepsilon_s\psi_{bi}}{qN_D} \left(\frac{N_A}{N_D + N_A}\right)}$$

- ▶ Of special interest is the threshold voltage V_T at which gate bias the channel width is totally consumed by both depletion regions.
- ▶ Setting the condition of: $x_s = W_{DS} + W_{Dn}$

- ▶ The threshold voltage is obtained :

$$V_T$$

$$= V_{FB}^* - qNDx_s \left(\frac{x_s}{2\varepsilon_s} + \frac{1}{C_{ox}} \right) + \left(\frac{x_s}{\varepsilon_s} + \frac{1}{C_{ox}} \right) \sqrt{\frac{2q\varepsilon_s N_D N_A \psi_{bi}}{N_D + N_A}} - \frac{N_A \psi_{bi}}{N_D + N_A}$$

- ▶ Once the channel dimensions are known, the channel charge can be calculated easily.
- ▶ Depending on the gate-bias range, we can have different amounts of bulk charge Q_B and surface inversion charge Q_I .

$$Q = Q_B = (x_s - W_D s - W_D n) N_D \quad V_T < V_G < V_{FB}^*$$

$$Q = Q_B + Q_I = (x_s - W_D n) N_D + C_{ox} (V_G - V_{FB}^*) \quad V_{FB}^* < V_G$$

- ▶ The drain current can be deduced from charge But

Table 1 Current Equations for Buried-Channel MOSFETs, Based on Long-Channel Constant Mobility (After Refs. 15 and 37)

$$V_T \leq V_G \leq V_{FB}^*$$

$$I_D = \frac{W\mu_B C_{ox}}{L(1+\sigma)} \left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 \right], \quad V_D \leq V_{Dsat}$$

$$= \frac{W\mu_B C_{ox}(V_G - V_T)^2}{L(1+\sigma)2\alpha}, \quad V_D \geq V_{Dsat}$$

$$V_G \geq V_{FB}^*$$

$$I_D = \frac{W\mu_B C_{ox}}{L(1+\sigma)} \left\{ (V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + (r-1) \left[(V_G - V_{FB}^*)V_D - \frac{1}{2}V_D^2 \right] \right\}, \quad V_D < V_G - V_{FB}^*$$

$$= \frac{W\mu_B C_{ox}}{L(1+\sigma)} \left[(V_G - V_T)V_D - \frac{1}{2}\alpha V_D^2 + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2 \right], \quad V_G - V_{FB}^* \leq V_D < V_{Dsat}$$

$$= \frac{W\mu_B C_{ox}}{L(1+\sigma)} \left[\frac{(V_G - V_T)^2}{2\alpha} + \frac{1}{2}(r-1)(V_G - V_{FB}^*)^2 \right], \quad V_D \geq V_{Dsat}$$

where

$$V_{Dsat} = (V_G - V_T)/\alpha \quad \sigma = \frac{C_{ox}x_s}{\epsilon_s} \left(\frac{C_{ox}x_s}{2\epsilon_s} + 1 \right) \quad \alpha = 1 + (1 + \sigma) \frac{\gamma}{4\sqrt{\psi_{bi}}}$$

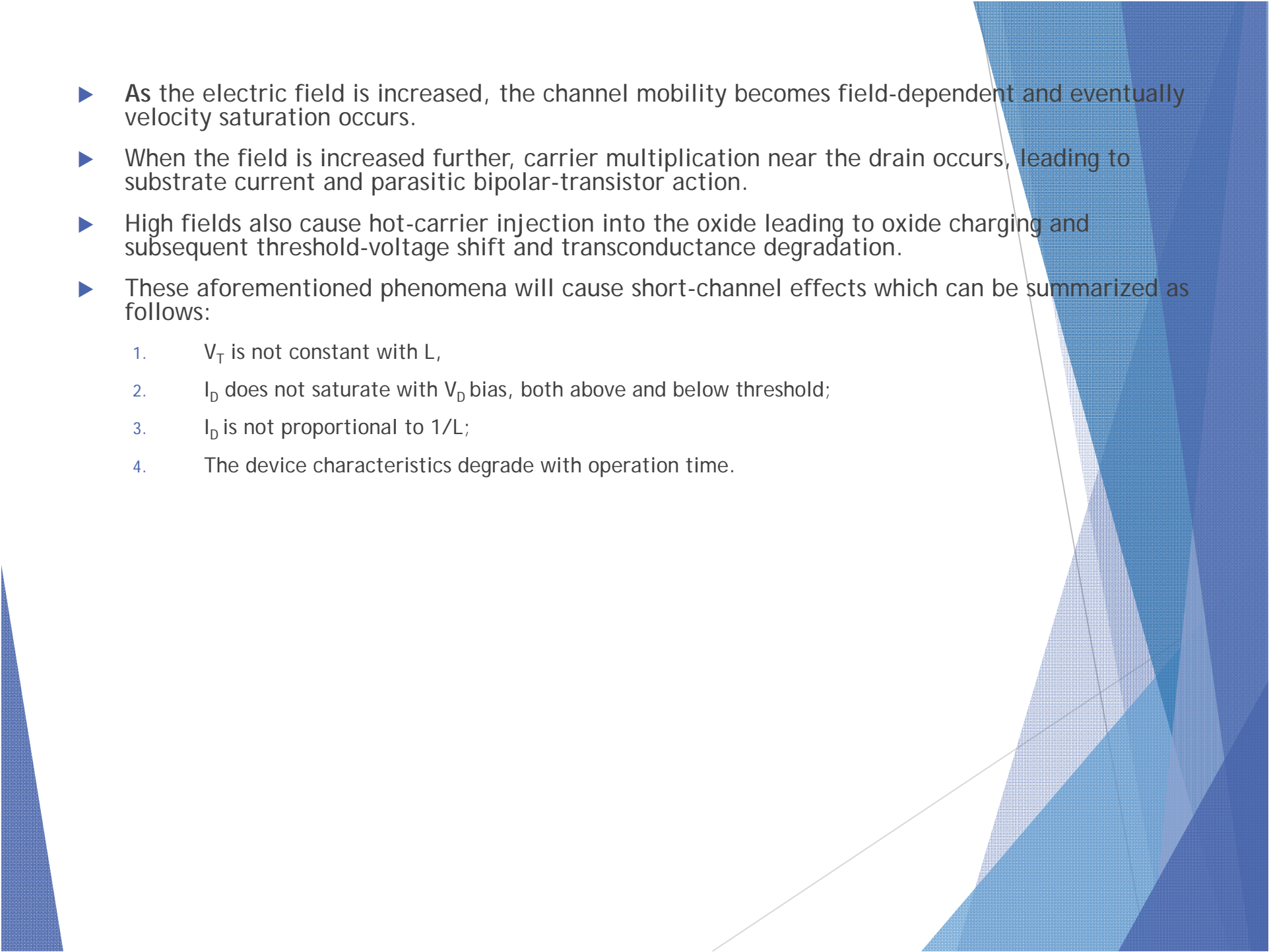
$$\mu_B = \text{bulk mobility} \quad r = (1 + \sigma) \frac{\mu_s}{\mu_B} \quad \gamma = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}}$$

$$\mu_s = \text{surface mobility}$$

$$S = (\ln 10) \frac{kT}{q} \left[1 + \frac{\epsilon_{ox} W_{Ds} + \epsilon_s d}{\epsilon_{ox}(W_{Dn} + W_{Dp})} \right],$$

6.4 DEVICE SCALING AND SHORT-CHANNEL EFFECTS

- ▶ As the MOSFET dimensions shrink, they need to be designed properly to preserve the long-channel behavior as much as possible.
- ▶ As the channel length decreases, the depletion widths of the source and drain become comparable to the channel length and punch-through between the drain and source will eventually occur.
- ▶ A higher channel doping will increase the threshold voltage, and in order to control a reasonable threshold voltage, a thinner oxide is necessary.
- ▶ Even with the best scaling rules, as the channel length is reduced, departures from long-channel behavior are inevitable.
- ▶ The potential distribution in the channel now

- 
- ▶ As the electric field is increased, the channel mobility becomes field-dependent and eventually velocity saturation occurs.
 - ▶ When the field is increased further, carrier multiplication near the drain occurs, leading to substrate current and parasitic bipolar-transistor action.
 - ▶ High fields also cause hot-carrier injection into the oxide leading to oxide charging and subsequent threshold-voltage shift and transconductance degradation.
 - ▶ These aforementioned phenomena will cause short-channel effects which can be summarized as follows:
 1. V_T is not constant with L ,
 2. I_D does not saturate with V_D bias, both above and below threshold;
 3. I_D is not proportional to $1/L$;
 4. The device characteristics degrade with operation time.

0.4.1 DEVICE SCALING

- ▶ The most-ideal scaling rule to avoid short-channel effects is simply to scale down all dimensions and voltages of a long-channel MOSFET.
- ▶ the internal electric fields are kept the same.
- ▶ The shrunk by the same scaling factor K .

Table 2 MOSFET Scaling

Parameter	Scaling factor: Constant- \mathcal{E}	Scaling factor: Actual	Limitation
L	$1/\kappa$	$/$	$/$
\mathcal{E}	1	> 1	$/$
d	$1/\kappa$	$> 1/\kappa$	Tunneling, defects
r_j	$1/\kappa$	$> 1/\kappa$	Resistance
V_T	$1/\kappa$	$\gg 1/\kappa$	Off current
V_D	$1/\kappa$	$\gg 1/\kappa$	System, V_T
N_A	κ	$< \kappa$	Junction breakdown

In ideal constant-field scaling parameters are scaled by the same factor.
In reality the scaling factors are limited by other reasons and skewed.

- ▶ The doping level is increased by K
- ▶ All voltages are reduced by K
- ▶ The junction depletion width is reduced by K
- ▶ the subthreshold swing S remains essentially the same
- ▶ S is proportional to $1 + C_D/C_{ox}$ and both capacitances are scaled up by the same factor K .
- ▶ Unfortunately such an ideal scaling rule is hindered by other factors that are fundamentally not scalable.

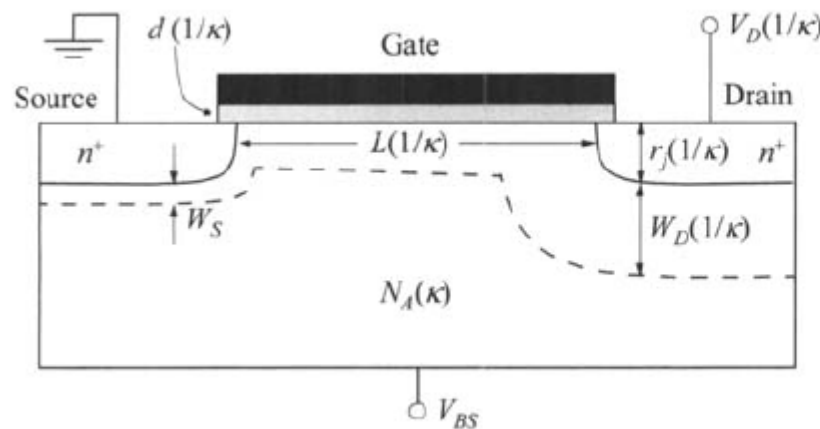


Fig. 25 Physical parameters for MOSFET scaling. Scaling factors for constant-field are indicated.

- ▶ First, the junction built-in voltage and the surface potential for the onset of weak inversion do not scale (only 10% change for 10 times increase in dopings).
- ▶ The range of gate voltage between depletion and strong inversion is approximately 0.5 V.
- ▶ These limitations from the energy gap and thermal energy kT remain constant
- ▶ The gate oxide thickness has the technological difficulty of defects as it approaches the low-nm scale.
- ▶ Tunneling through the oxide is another fundamental limitation
- ▶ The source and drain series resistance increases when r_j is decreased.
- ▶ The threshold voltage cannot be scaled due to the off-current consideration, even with a fixed subthreshold swing.

- ▶ The expression for minimum channel length for which long-channel behavior can be observed is found to follow a simple empirical relation:

$$L \geq \sqrt[3]{C_1[r_j d (W_s + W_D)^2]}$$

- ▶ where C_1 is a constant, and $W_s + W_D$ is the sum of the source and drain depletion widths in a one-dimensional abrupt junction formulation

$$W_D = \sqrt{\frac{2\epsilon_s}{qN_A} (VD + \psi_{bi} - VBS)}$$

- ▶ Such a high-K gate dielectric can relax the physical thickness, improving the defect density and reducing the field for tunnelling.

6.4.2 Charge Sharing from Source/Drain

- ▶ the inversion charge and depletion charge is completely balanced by the charge on the gate.
- ▶ 2-dimensional examination at the ends of the channel reveals that some of the depletion charge is balanced by the n⁺ source and drain.
- ▶ the charge conservation principle to the region bounded by the gate, the channel, and the Source/drain:

$$Q_M' + Q_n' + Q_B' = 0$$

- ▶ where Q_M is the total charge on the gate, Q_n is the total inversion-layer charge, and Q_B is the total ionized impurity in the depletion region.

- ▶ The threshold voltage, which can be viewed as voltage required to deplete the total bulk charge Q_B in the maximum depletion width, is given by:

$$V_T = V_{FB} + 2\psi_B + \frac{Q_B}{C_{ox}A}$$

- ▶ where A is the gate area $Z \times L$. For long-channel devices, $Q_B = qAN_A W_{DM}$ where W_{DM} is the maximum depletion-layer width

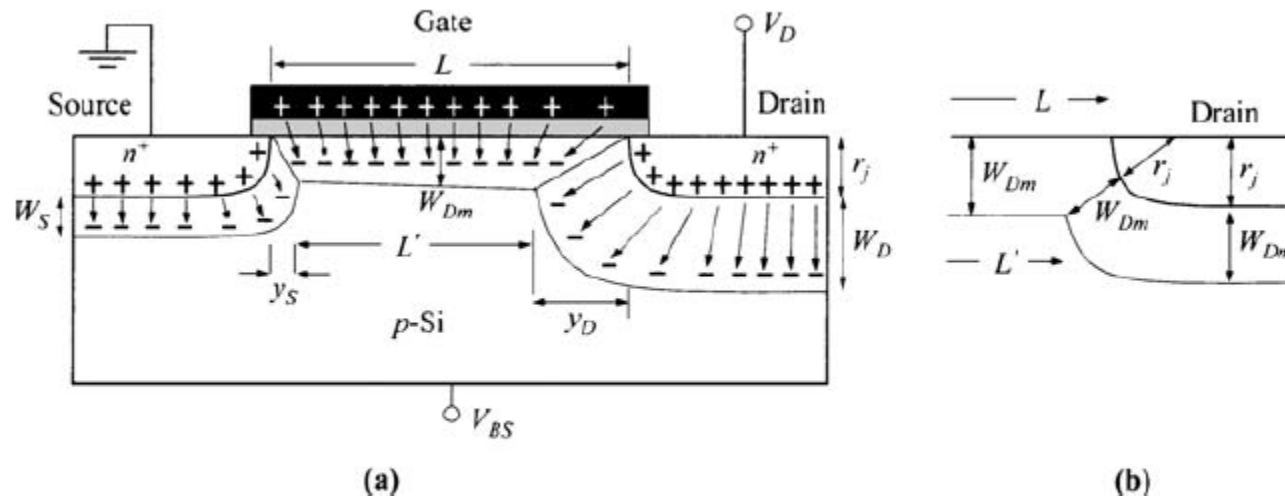


Fig. 26 Charge-conservation model for (a) $V_D > 0$, and (b) $V_D = 0$ where $W_D \approx W_S \approx W_{Dm}$. (After Ref. 47.)

$$W_{DM} = \sqrt{\frac{2\epsilon_s(2\psi_B - VBS)}{qN_A}}$$

- ▶ For short-channel devices, the full effect of Q_B on the threshold voltage is reduced, because near the source and drain ends of the channel.
- ▶ First-order estimation of the threshold voltage can be made by considering the charge partition.

- ▶ The total bulk depletion charge can be estimated by the trapezoid

$$Q'_B = \frac{ZqN_A W_{DM}(L + L')}{2}$$

- ▶ For small drain bias, we can assume that $W_n, W_c = W_{nm}$ and by straightforward trigonometric analysis

$$L' = L - 2(\sqrt{r_j^2 + 2W_{DM}r_j} - r_j).$$

- ▶ The threshold-voltage shift from long-channel behavior is then given by:

$$\begin{aligned}\Delta V_T &= \frac{1}{C_{ox}} \left(\frac{Q'_B}{ZL} - qN_A W_{DM} \right) = - \frac{qN_A W_{DM}}{C_{ox}} \left(1 - L + \frac{L'}{2L} \right) \\ &= - \frac{qN_A W_{DM} r_j}{C_{ox} L} \left(\sqrt{1 + \frac{2W_{DM}}{r_j}} - 1 \right)\end{aligned}$$

The negative sign means V_T is lowered and the transistor is easier to turn on.

- ▶ To take into account the effect of the drain voltage and the substrate bias.

$$\Delta V_T = -\frac{qN_A W_{DM} r_j}{2C_{ox} L} \left[\left(\sqrt{1 + \frac{2y_s}{r_j}} - 1 \right) + \left(\sqrt{1 + \frac{2y_D}{r_j}} - 1 \right) \right]$$

► where y_s and y_D are given as:

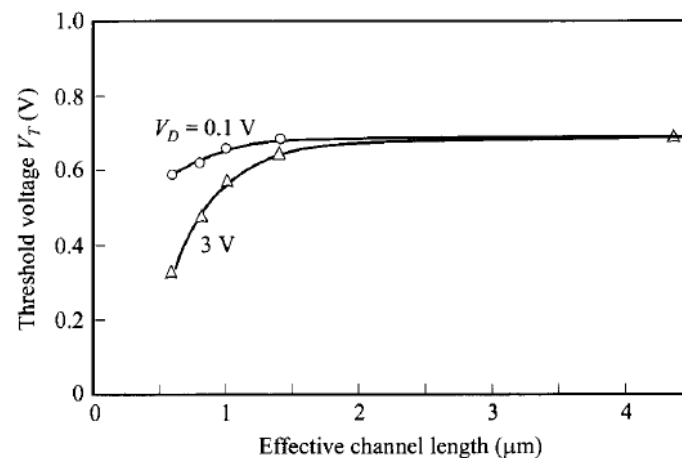
$$y_s \approx \sqrt{\frac{2\varepsilon\psi}{qN_A} (\psi_{bi} - \psi_s - V_{BS})}$$

$$y_D \approx \sqrt{\frac{2\varepsilon\psi}{qN_A} (\psi_{bi} + V_D - \psi_s - V_{BS})}$$

Note that the threshold voltage becomes a function of both L and V_D .

6.4.3 Channel-Length Modulation

- ▶ The y_D is a high-field region where carriers are swept out efficiently.
- ▶ The y_s is a transitional region where the electron concentration is higher than that in the main channel.
- ▶ for consideration of the channel drift region, the *effective* channel length:



- ▶ This factor of the effective channel length accounts for the channel length modulation effect.
- ▶ The change in V_T is only linearly dependent on the channel length.

6.4.4 Drain-Induced Barrier Lowering (DIBL)

- ▶ We have pointed out that when the source and drain depletion regions are a substantial fraction of the channel length, short-channel effects start to occur.
- ▶ In extreme cases when the sum of these depletion widths approaches the channel length ($y_s + y_D = L$)
- ▶ This condition is commonly called punch-through.
- ▶ The net result is a large leakage current between the source and drain, and that this current is a strong function of the drain bias.

- ▶ The origin of punch-through is the lowering of the barrier near the source, commonly referred to as DIBL (drain-induced barrier lowering).
- ▶ When the drain is close to the source, the drain bias can influence the barrier at the source end, such that the channel carrier concentration at that location is no

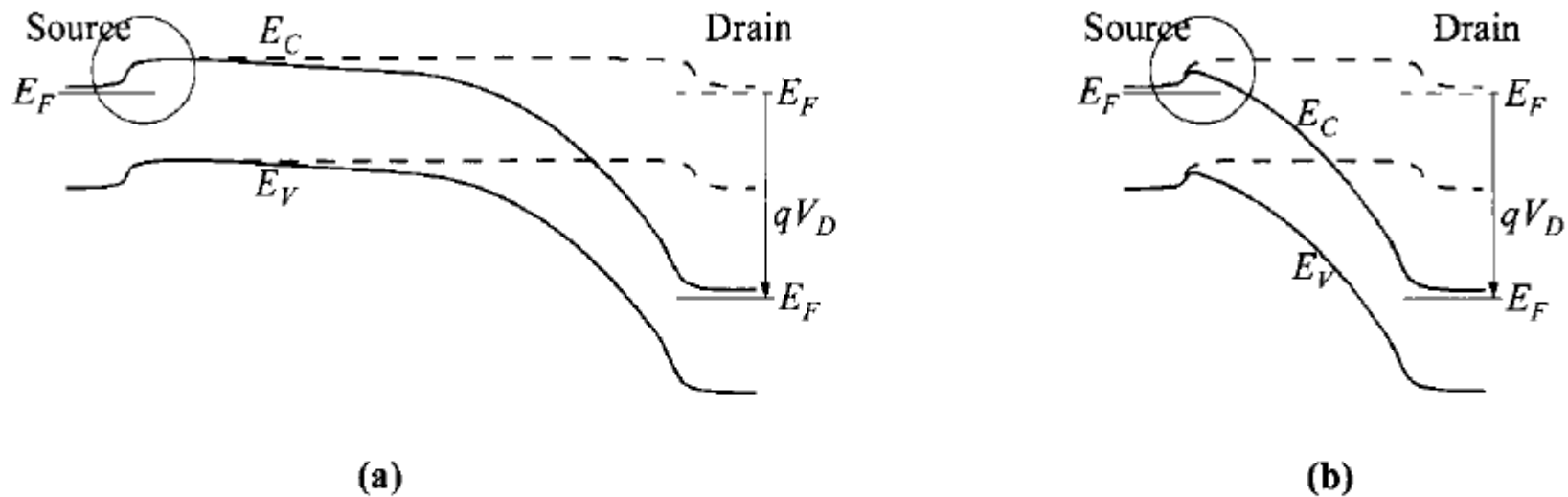
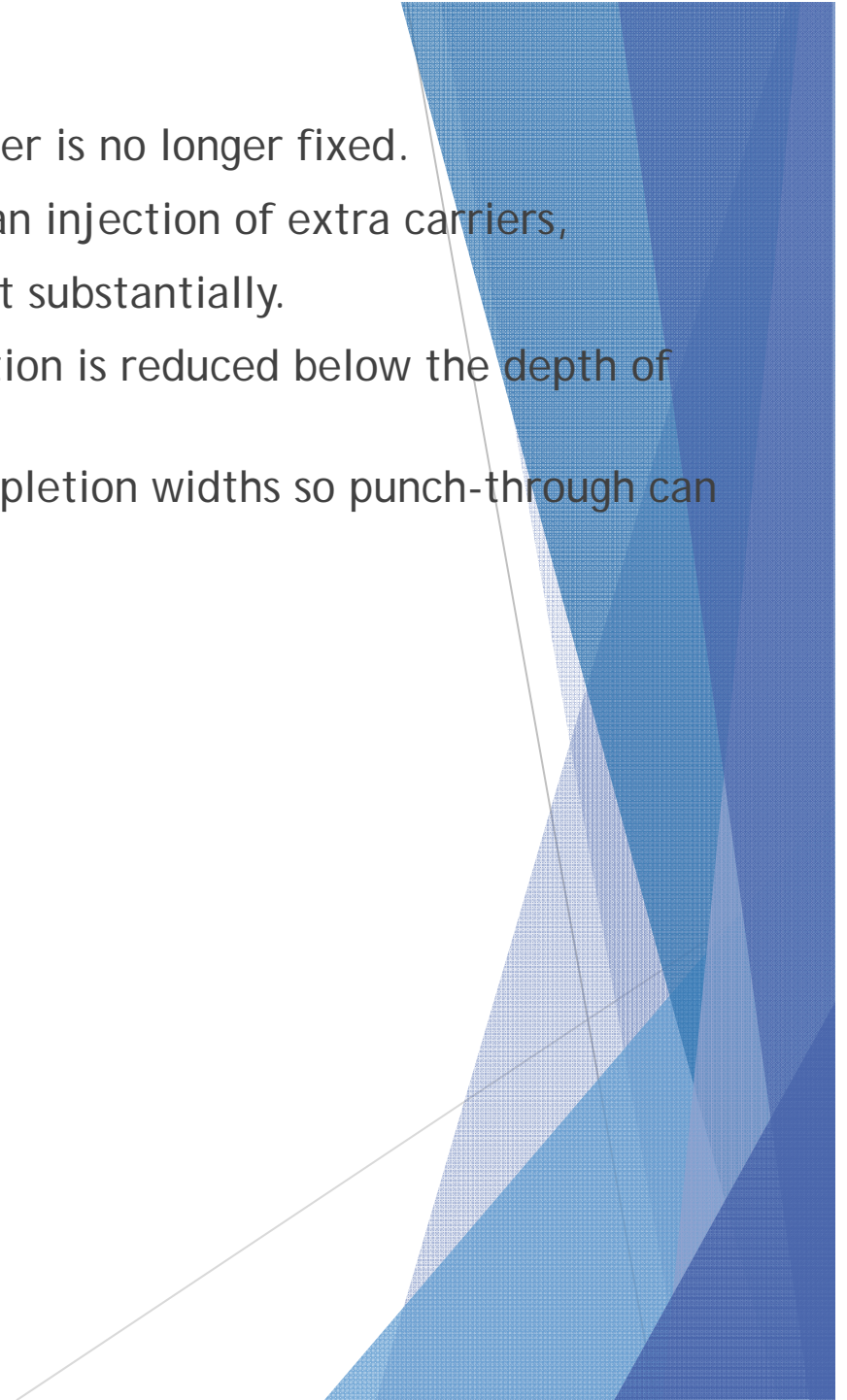
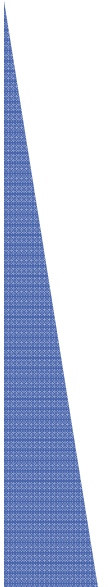


Fig. 28 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines $V_D = 0$. Solid lines $V_D > 0$.

- ▶ For a short-channel device, this same barrier is no longer fixed.
- ▶ The lowering of the source barrier causes an injection of extra carriers,
- ▶ These extra carriers increasing the current substantially.
- ▶ it is common that the substrate concentration is reduced below the depth of the source/drain junction r_j .
- ▶ A reduced substrate doping widens the depletion widths so punch-through can also happen via a path in the bulk.



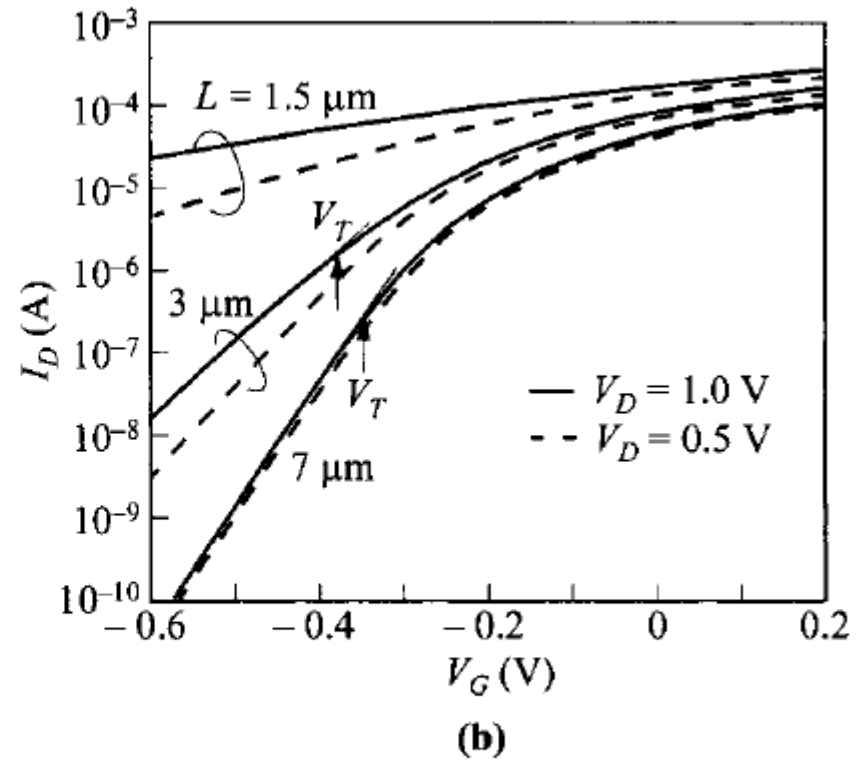
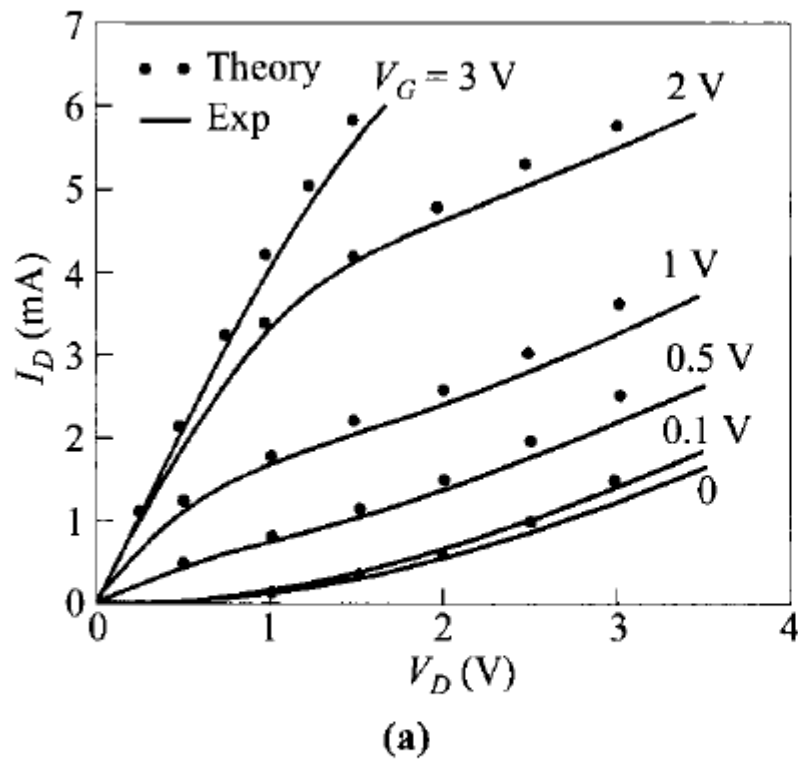


Fig. 29 Drain characteristics of MOSFETs showing DIBL effect. (a) Above threshold. $L = 0.23 \mu\text{m}$. $d = 25.8 \text{ nm}$. $N_A = 7 \times 10^{16} \text{ cm}^{-3}$. (b) Below threshold. $d = 13 \text{ nm}$. $N_A = 10^{14} \text{ cm}^{-3}$. (After Ref. 50.)

- ▶ The punch-through drain voltage can be estimated by the depletion approximation to be:

$$V_{pt} = \frac{qN_A(L - ys)^2}{2\epsilon_s} - \psi_{bi}$$

Drain current will be dominated by the space-charge-limited current:

$$I_D \approx \frac{9\epsilon_s\mu_n AV_D^2}{8L^3}$$

- ▶ where A is the cross-sectional area of the punch-through path.
- ▶ The space-charge limited current increases with V_D^2 and is parallel to the inversion-layer current.

6.5 MOSFET STRUCTURES

6.5.1 Channel Doping Profile

- ▶ Typical high-performance MOSFET structure based on planar technology.
- ▶ The channel doping profile has a peak level slightly below the semiconductor surface.
- ▶ Such a retrograde profile is achieved with ion implantation, often of multiple doses and energies.
- ▶ The low concentration at the surface has the advantages of higher mobility due to reduced normal field and low threshold voltage

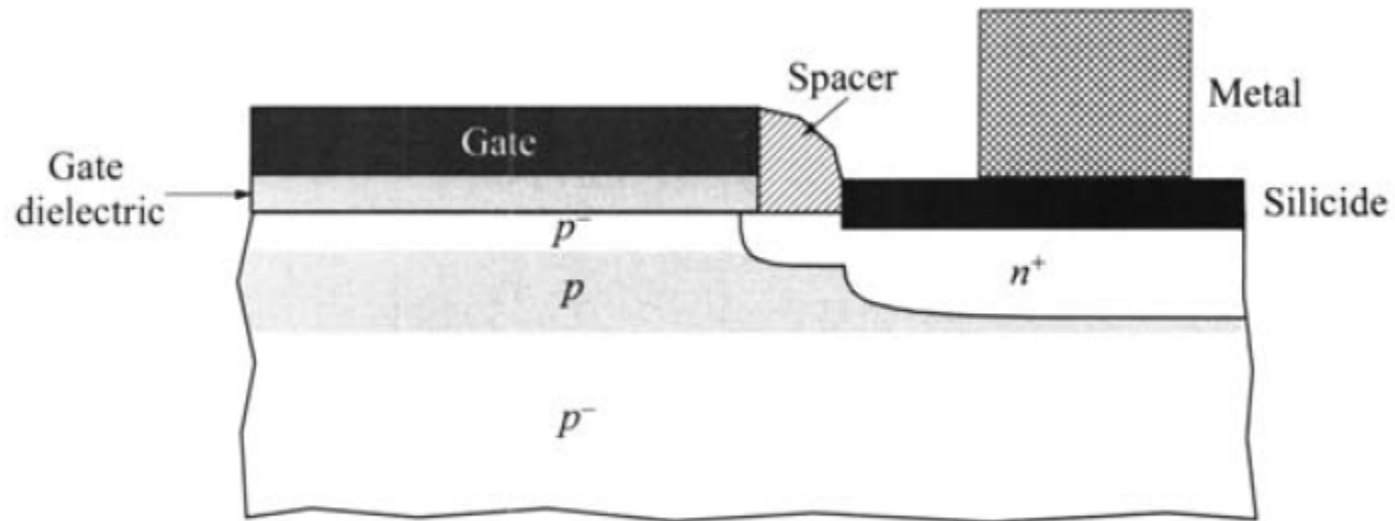


Fig. 32 High-performance MOSFET planar structure with a retrograde channel doping profile, two-step source/drain junction, and self-aligned silicide source/drain contact.

6.5.2 Gate Stack

- ▶ The gate stack consists of the gate dielectric and the gate contact material.
- ▶ The gate dielectric has been exclusively SiO_2 right from the birth of MOSFET.
- ▶ In fact, the ideal Si- SiO_2 interface is the main factor responsible for the success of MOSFET.
- ▶ As the oxide thickness is scaled into the range below = 2 nm.
- ▶ fundamental problem of tunneling and technological difficulty of defects start to demand alternatives
- ▶ The solution that is actively sought after is a material with high dielectric constant, called high-K dielectric.

- ▶ K dielectric can have a thicker physical thickness for the same capacitance, thus reducing its electric field and technological problem related to defects.
- ▶ equivalent oxide thickness EOT = thickness x $K(\text{SiO}_2)/K$.
- ▶ Material options being examined are Al_2O_3 , HfO_2 , ZrO_2 , Y_2O_3 , La_2O_3 and TiO_2 .
- ▶ The dielectric constants for these materials range from 9 to 30, except for TiO_2 which is larger than 80.

- ▶ The gate material has been polysilicon for a long time.
- ▶ The advantages of a poly-Si gate:
 - ▶ its compatibility with the silicon processing
 - ▶ its ability to withstand high-temperature anneal that is required after self-aligned source/drain implantation.
 - ▶ Another important factor is that the work function can be varied by doping it into n-type and p-type.
- ▶ One limitation of the poly-Si gate is its relatively high resistance.
- ▶ This does not result in penalty of dc characteristics since the gate terminates on the gate insulator
- ▶ The penalty shows up in high-frequency parameters such as noise

6.5.3 Source/Drain Design

- ▶ Typically the junction has two sections. The extension near the channel has shallower junction depth to minimize short-channel effects.
- ▶ Sometimes it is doped less heavily to reduce the lateral field for consideration of hot-carrier aging.
- ▶ For this purpose it is called a lightly doped drain (LDD).
- ▶ The deeper junction depth away from the channel helps to minimize the series resistance.

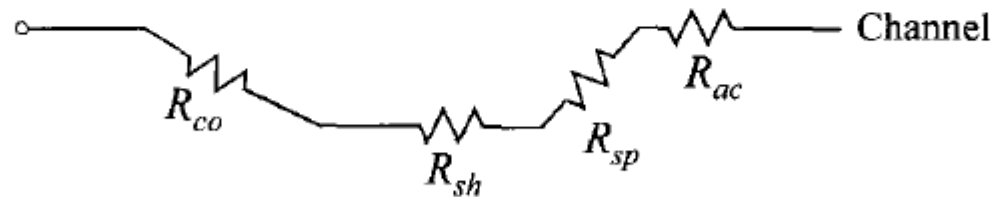
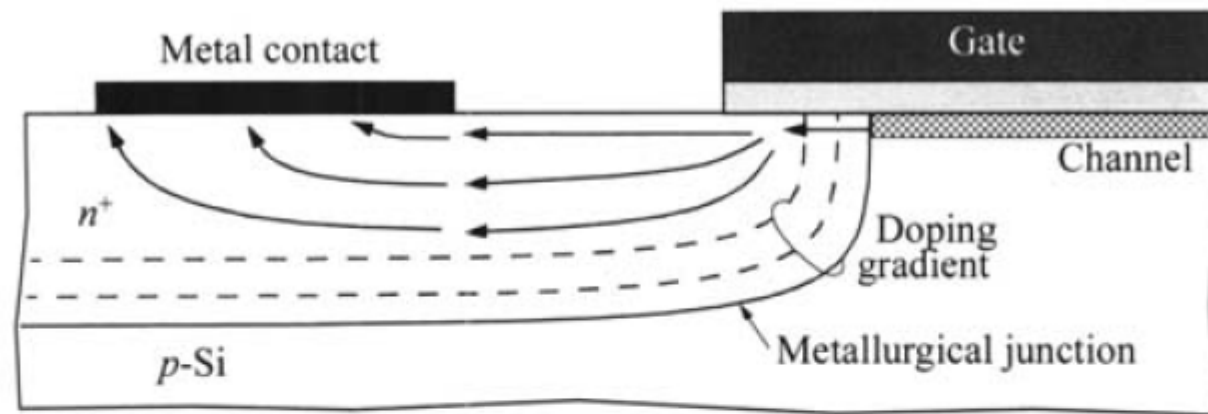


Fig. 33 Detailed analysis of different components of parasitic source/drain series resistance. R_{ac} is accumulation-layer resistance due to doping gradient. R_{sp} = spreading resistance. R_{sh} = sheet resistance. R_{co} = contact resistance. (After Ref. 60.)

- ▶ In practice the profile is never perfectly abrupt, and there exists a region of accumulation layer (of n-type) before current spreads into the bulk.
- ▶ This accumulation-layer resistance R_{ac} is related to the transition distance before the doping reaches a critical level.
- ▶ Unlike the metal contact, the silicide can be made self-aligned to the gate.
- ▶ thus minimizing the sheet-resistance component (R_{sh}) between the contact and the channel.
- ▶ In this way the silicide has become the metal contact because contact resistance between metal and silicide is very small.
- ▶ This self-aligned silicide process has been coined salicide.

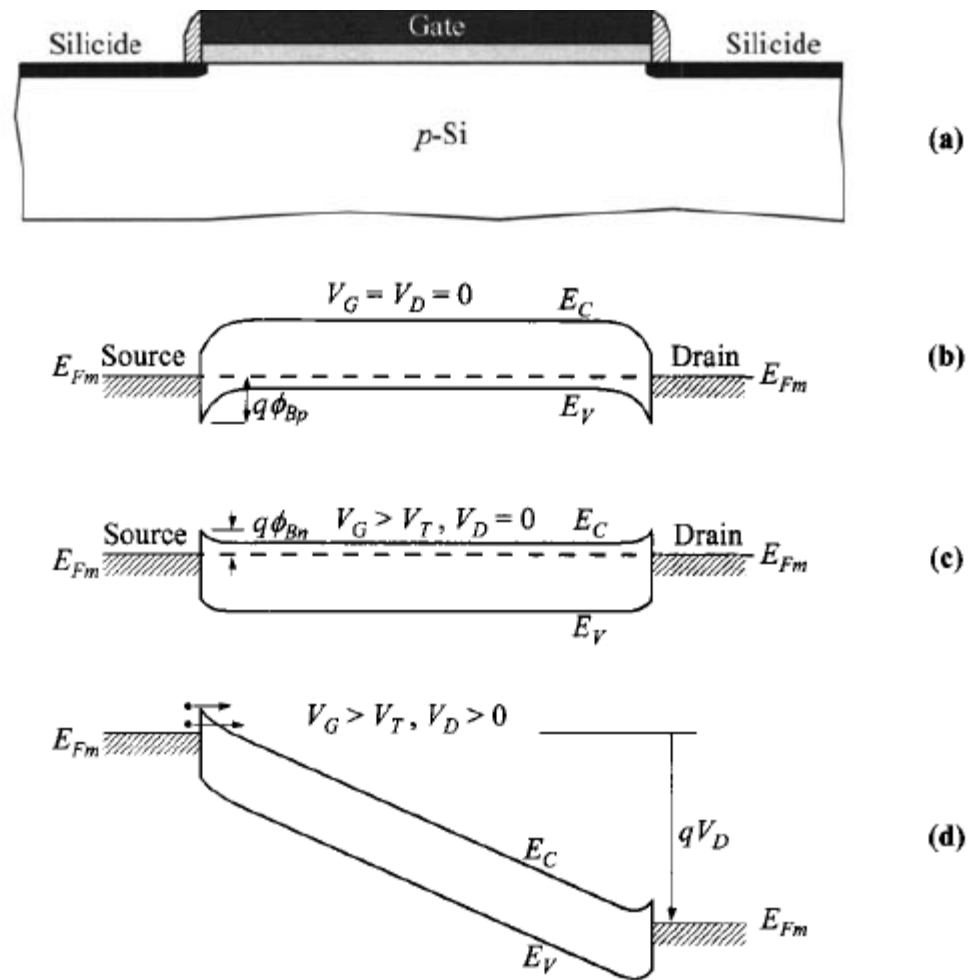


Fig. 34 MOSFET with Schottky-barrier source and drain. (a) Cross-sectional view of the device. (b)–(d) Band diagrams along semiconductor surface under various biases.

- ▶ The silicide process is described as follows.
- ▶ After the gate definition, an insulator spacer is formed on the sides of the gate.
- ▶ A metal layer for silicidation is deposited uniformly, which at this stage is shorting the gate and the source/drain.
- ▶ After a thermal reaction at low temperature ($= 450^{\circ}\text{C}$), the metal reacts with silicon to form silicide on the source/drain region.
- ▶ Silicide formation on the gate is optional depending on whether the gate is capped with an insulation layer as part of the gate stack.
- ▶ Metal over the spacer region and the field region (between transistors, not shown) remains metal since there is no exposed silicon for reaction.
- ▶ The metal is then removed with a selective chemical that etches metal only without etching silicide, thereby removing the shorting paths.
- ▶ the consumption of silicon during silicide formation. Examples for silicides are CoSi_2 , NiSi_2 , TiSi_2 and PtSi .

Schottky-Barrier Source/Drain.

- ▶ Instead of $p-n$ junction, use of Schottky-barrier contacts for the source and drain of a MOSFET can result in some advantages in fabrication and performance.
- ▶ For a Schottky contact, the junction depth can be effectively be made zero to minimize the short-channel effects.
- ▶ The $n-p-n$ bipolar-transistor action is also absent for undesirable effects such as bipolar breakdown and latch-up phenomenon in CMOS circuits.
- ▶ Eliminating high-temperature implant anneal can promote better quality in the oxides and better control of geometry.

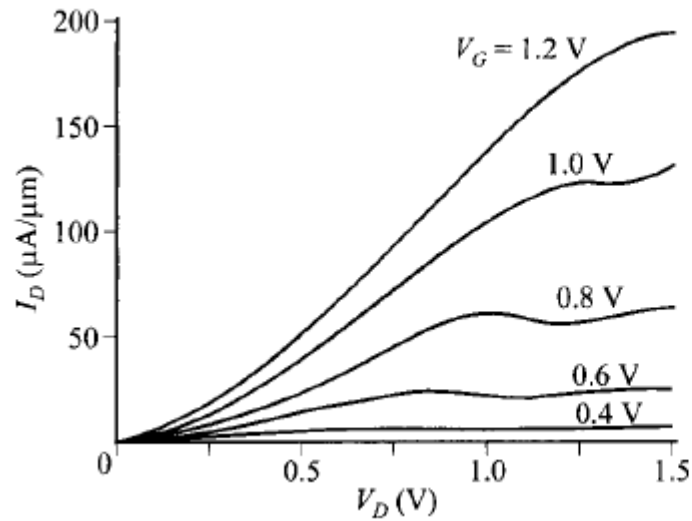


Fig. 35 I - V characteristics of n -channel MOSFET with Schottky source/drain. (After Ref. 63.)

- ▶ The disadvantages of the Schottky source/drain are high series resistance due to the finite barrier height, and higher drain leakage current.
- ▶ Typical I - V curves show that current is started at low-drain bias
- ▶ The metal or silicide contact has to extend underneath the gate for continuity.
- ▶ This process is much more demanding than a junction source/drain which is done by self aligned implantation and diffusion.

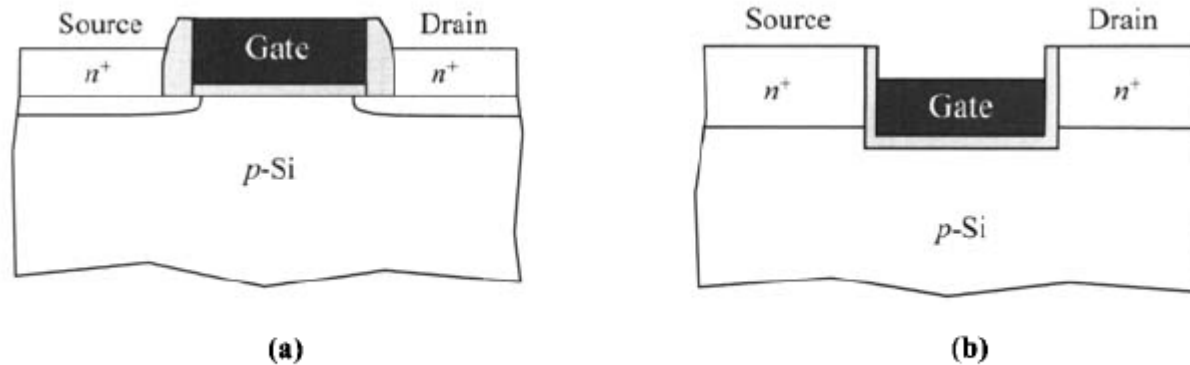


Fig. 36 Means to reduce source/drain junction depth and series resistance. (a) Raised source/drain. (b) Recessed channel.

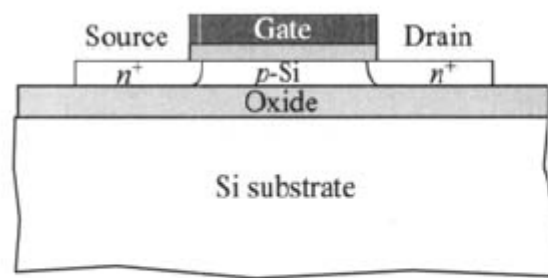
► **Raised Source/Drain:**

- An advanced design is the raised source/drain where a heavily doped epitaxial layer is grown over the source/drain regions
- The purpose is to minimize junction depth to control short-channel effects.
- An alternative is the recessed-channel MOSFET where the junction depth r_j is zero or negative
- The drawback of the recessed-channel structure, especially for submicron is the difficulty in controlling the contour and the oxide thickness at the corners where the threshold voltage is determined.
- The oxide charging may be worsened because more hot-carrier injection will occur.

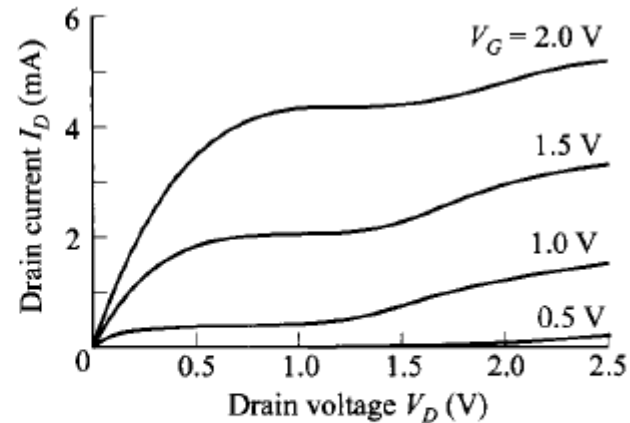
6.5.4 SOI and Thin-Film Transistor (TFT)

- ▶ SOI:
 - ▶ the top silicon layer of an SOI (silicon-on-insulator) wafer is high-quality single-crystalline material that is suitable for high-performance
 - ▶ Many forms of SOI structures have been demonstrated with different insulator materials and holding substrates.
 - ▶ These include silicon-on-oxide, silicon-on-sapphire (SOS), silicon-on-zirconia (SOZ), and silicon on- nothing (air gap).
 - ▶ In SOS and SOZ technologies, a single-crystalline silicon film is epitaxially grown on a crystalline insulating substrate.
 - ▶ In these cases, the insulators are the substrates themselves Al₂O₃ in SOS and ZrO₂ in SOZ.
 - ▶ The difficulties in these techniques are the material quality when the film gets thinner.

- ▶ using oxide as an insulator and another Si wafer as the holding substrate, is by far the most popular.
- ▶ Among them SIMOX (separation by implantation of oxygen) where high-dose oxygen is implanted onto a silicon wafer followed by high-temperature anneal to form the buried SiO₂ layer
- ▶ Another uses laser recrystallization, transforming amorphous silicon deposited onto the oxide layer into single-crystalline material or into poly-crystalline form with

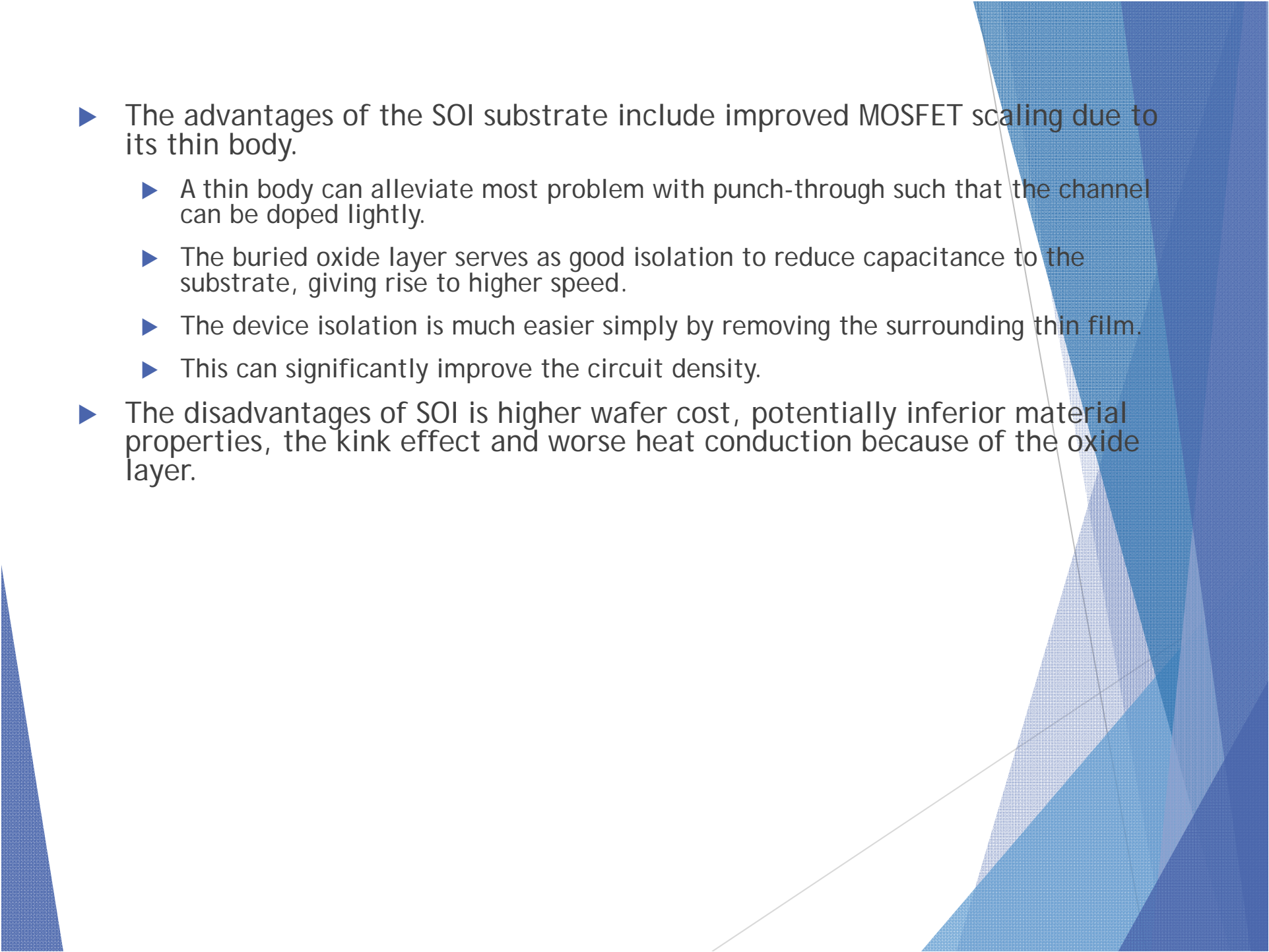


(a)



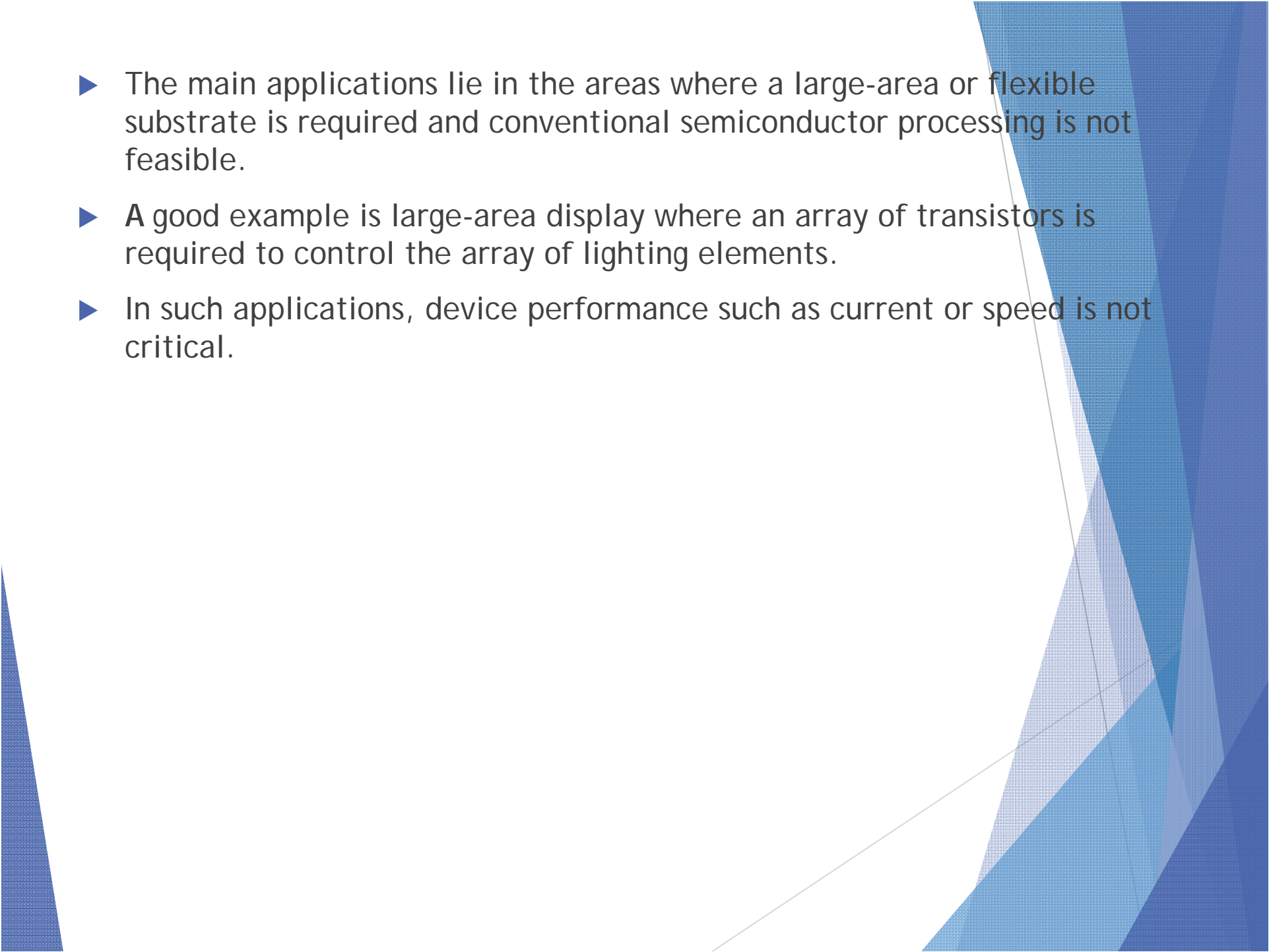
(b)

Fig. 37 (a) Typical structure of MOSFET on SOI wafer, and (b) its drain characteristics. (After Ref. 66.)

- 
- ▶ The advantages of the SOI substrate include improved MOSFET scaling due to its thin body.
 - ▶ A thin body can alleviate most problem with punch-through such that the channel can be doped lightly.
 - ▶ The buried oxide layer serves as good isolation to reduce capacitance to the substrate, giving rise to higher speed.
 - ▶ The device isolation is much easier simply by removing the surrounding thin film.
 - ▶ This can significantly improve the circuit density.
 - ▶ The disadvantages of SOI is higher wafer cost, potentially inferior material properties, the kink effect and worse heat conduction because of the oxide layer.

Thin-Film Transistor (TFT):

- ▶ The thin-film transistor usually refers to MOSFET as opposed to other kinds of transistors.
- ▶ The structure is similar to MOSFET built on SOI with the exception that the active film is a deposited thin film and that the substrate can be of any form
- ▶ Because the semiconductor layer is formed by deposition the amorphous material has more defects and imperfections than in single-crystalline semiconductors, resulting in more complicated transport processes in the TFT.
- ▶ To improve device performance, reproducibility, and reliability, the bulk and interfacetrapped densities must be reduced to reasonable levels.

- 
- ▶ The main applications lie in the areas where a large-area or flexible substrate is required and conventional semiconductor processing is not feasible.
 - ▶ A good example is large-area display where an array of transistors is required to control the array of lighting elements.
 - ▶ In such applications, device performance such as current or speed is not critical.

6.5.5 Three-Dimensional Structures

- ▶ In device scaling, the optimum design is with MOSFET built on a body of ultra-thin layer such that the body is fully depleted under the whole bias range.
- ▶ A design to achieve this more efficiently is to have a surround gate structure that encloses the body layer from at least two sides.
- ▶ They can be classified according to their current-flow pattern; the horizontal transistor and vertical transistor
- ▶ There challenging from a fabrication point of view, the horizontal scheme is more compatible with SOI technology and more data are reported

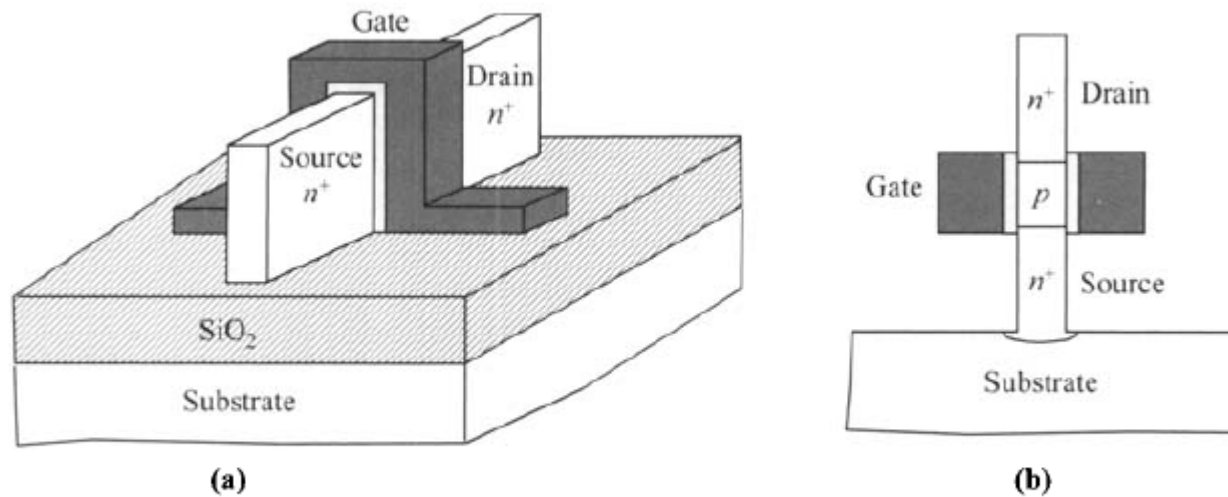


Fig. 38 Schematic 3-dimensional MOSFETs. (a) Horizontal structure. (b) Vertical structure.
Note commonality of surround gate and thin body.

- ▶ This fact presents great challenges in achieving a smooth channel surface from etching and growth or deposition of gate dielectrics on these surfaces.
- ▶ Formation of the source/drain junction is no longer trivial by means of ion implantation.
- ▶ Salicide formation will also be much more difficult.
- ▶ Whether one of these turns out to be the device choice.

6.5.6 POWER MOSFET IS

- ▶ In general, power MOSFETs employ thicker oxides, deeper junctions, and have longer channel lengths. These generally post a penalty on device performance such as transconductance (g_m) and speed f_T .
- ▶ **DMOS:** As the name implies, in the DMOS (double-diffused MOS)
 - ▶ the channel length is determined by the higher diffusion rate of the p-dopant (e.g., boron) compared to the n⁺-dopant (e.g., phosphorus) of the source.
 - ▶ This technique can yield very short channels without depending on a lithographic mask.
 - ▶ The channel is followed by a lightly doped n--drift region.
 - ▶ This drift region is long compared to the channel, and it minimizes the peak electric field in this region by maintaining a uniform field.
 - ▶ Usually the drain is located at the substrate contact.
 - ▶ The field near the drain is the same as in the drift region, so avalanche breakdown, multiplication, and oxide charging are lessened compared to conventional MOSFETs.

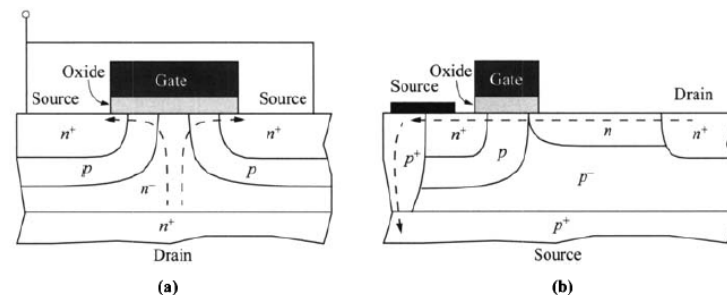


Fig. 39 (a) Vertical DMOS transistor and (b) LDMOS transistor. Current path is indicated by dashed line. In LDMOS transistors, it is common to connect the source to the substrate to reduce inductance of the bonding wire.

6.7 NONVOLATILE MEMORY DEVICES

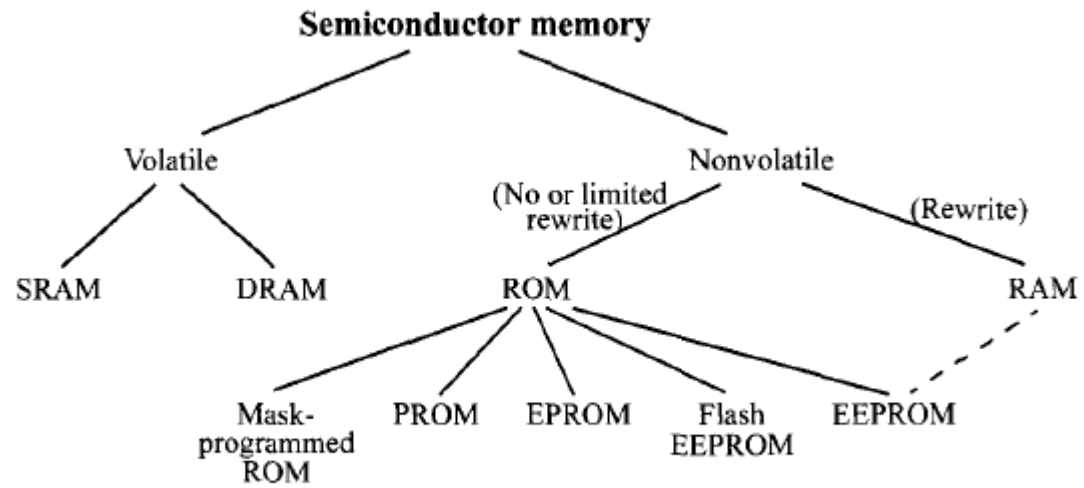


Fig. 43 Classification of semiconductor memories.

▶ **Mask-programmed ROM:**

The memory content is fixed by the manufacturer and is not programmable once it is fabricated. Sometimes mask-programmed ROM is simply referred to as ROM

▶ **PROM:**

Programmable ROM is sometimes called field-programmable ROM or fusible-link ROM. The connectivity of the array is custom programmed using the technique of fusing or antifusing. After programming, the memory works as a ROM.

▶ **EPROM:**

In an electrically programmable ROM, programming is performed by hot-electron injection or tunneling to the floating gate, and it requires biases on both the drain and the control gate. Global erase is by exposure to a UV light or x-ray. Selective erase is not possible.

▶ **Flash EEPROM:**

A flash EEPROM, as opposed to a full-feature EEPROM below, can be erased electrically but only by a large block of cells simultaneously. It loses byte selectivity but maintains a one-transistor cell. It is, thus, a compromise between an EPROM and a full-feature EEPROM.

- ▶ **EEPROM:** In an electrically erasable/programmable ROM, not only can it be erased electrically, but also selectively by byte address. To erase selectively, a select transistor is needed for each cell, leading to a two-transistor cell. This makes it less popular than a flash EEPROM.
- ▶ **Nonvolatile RAM:** This memory can be viewed as a nonvolatile SRAM, or EEPROM with short programming time as well as high endurance. If technology allows the aforementioned features, this would be the ideal memory.
- ▶ The two groups of nonvolatile memory devices are the floating-gate devices and the charge-trapping devices
- ▶ In both types of devices, charges are injected from the silicon substrate across the first insulator and stored in the floating gate or at the nitride-oxide
- ▶ The stored charge gives rise to a threshold-voltage shift, and the device is at a *high-threshold state* (programmed).
- ▶ For a well-designed memory device, the charge retention time can be over 100 years

0.7.1 Floating-Gate Devices

- ▶ In a floating-gate memory device, charge is injected to the floating gate to change the threshold voltage.
- ▶ The two modes of programming are hot-carrier injection and Fowler-Nordheim tunneling.

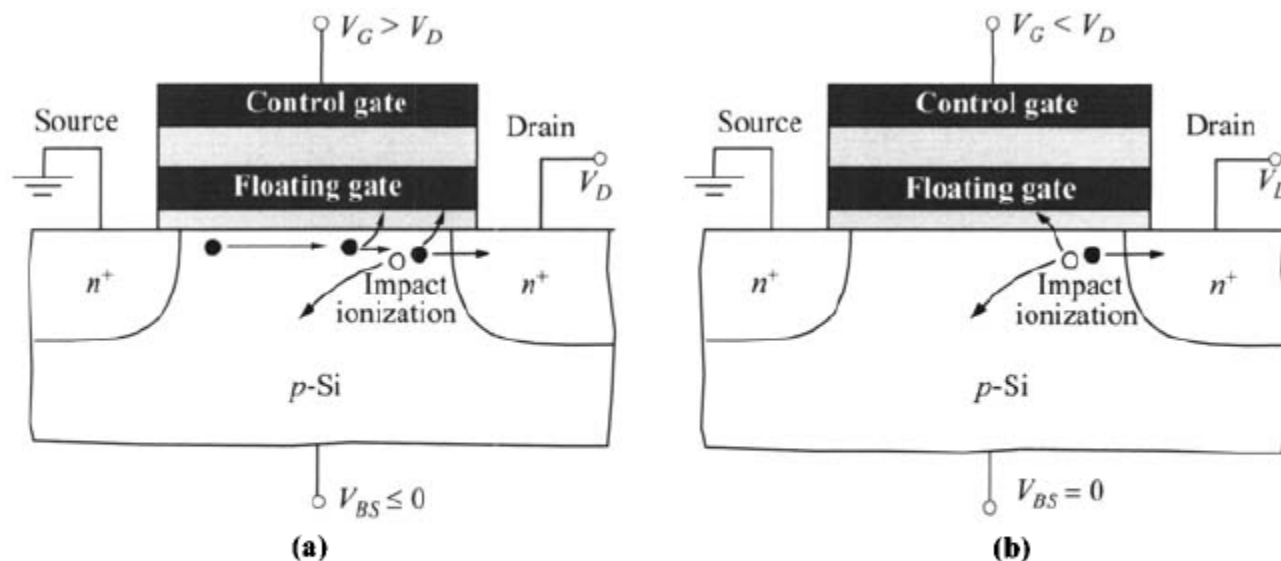
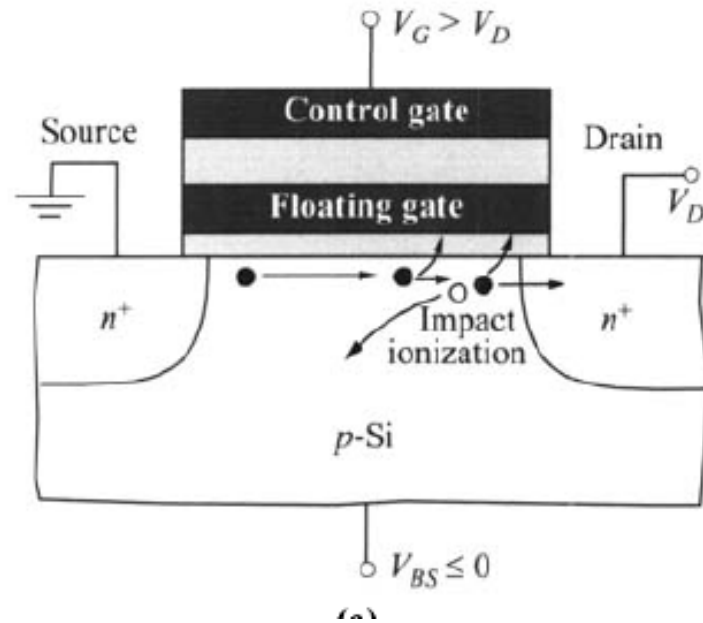
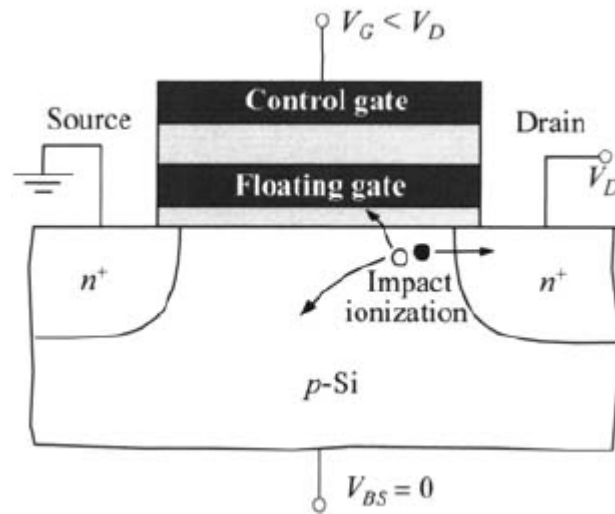


Fig. 45 Charging of the floating gate by hot carriers. (a) Hot electrons from channel and impact ionization. (b) Hot holes from drain avalanche. Note difference in gate bias between the two figures.



- ▶ The channel carriers (electrons) acquire energy from the field and become hot carriers. When their energy is higher than the barrier of the Si/SiO₂ interface, they can be injected to the floating gate.
- ▶ At the same time, the high field also induces impact ionization. These generated
- ▶ secondary hot electrons can also be injected to the floating gate. The hot-carrier injection currents give rise to the equivalence of gate current in a regular MOSFET.
- ▶ This gate current peaks at $V_{FG} = V_D$ where V_{FG} is the potential of the floating gate



- ▶ the original method of hot-carrier injection using drain-substrate avalanche.
- ▶ the floating-gate potential is more negative such that hot holes are injected instead.
- ▶ This injection scheme is found to be less efficient and is no longer used in practice.
- ▶ Besides hot-carrier injection, electrons can be injected by tunneling.
- ▶ In this programming mode, the electric field across the bottom oxide layer is most critical.
- ▶ On application of a positive voltage V_G to the control gate, an electric field is established in each of the two insulators

$$\begin{aligned}\varepsilon E_1 &= \varepsilon E_2 + Q \\ V_G &= V_1 + V_2 = d_1 E_1 + d_2 E_2 \\ E_1 &= \frac{V_G}{d_1 + d_2 \left(\frac{\varepsilon_1}{\varepsilon_2}\right)} + \frac{Q}{\varepsilon_1 + \varepsilon_2 \left(\frac{d_1}{d_2}\right)}\end{aligned}$$

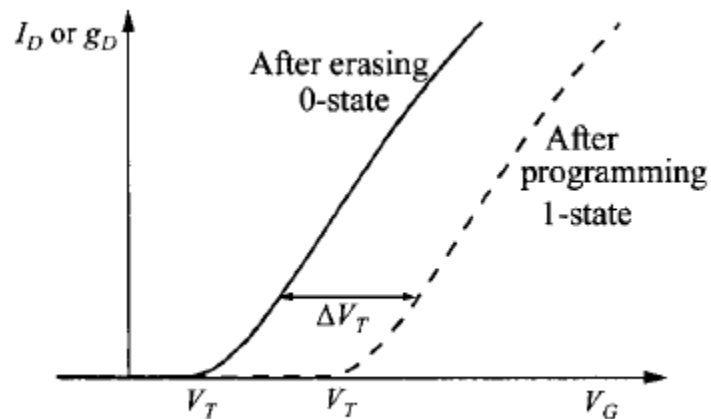
- ▶ The current transport in insulators is generally a strong function of the electric field. When the transport is Fowler-Nordheim tunneling, the current density has the form:

$$J = C_4 E^2 e^{-\left(\frac{E_0}{E_1}\right)}$$

- ▶ where C_4 and E_0 are constants in terms of effective mass and barrier height
- ▶ the total stored charge Q is equal to the integrated injection current since the gate is floating.

$$J = C_4 \mathcal{E}_1^2 \exp\left(\frac{-\mathcal{E}_0}{\mathcal{E}_1}\right)$$

- ▶ This causes a shift of the threshold voltage by the amount: $\Delta V_T = \frac{d_2 Q}{\epsilon_2}$
- ▶ This threshold-voltage shift can be directly measured as shown in the I_D - V_G plot
- ▶ the threshold-voltage shift can be measured from the drain conductance.
- ▶ The change in V_T results in a change in the channel conductance g_D , of the MOSFET.
- ▶ For small drain voltage, the drain current of an n-channel MOSFET is $I_D = \frac{Z}{L} \mu C_{ox} (V_G - V_T) V_D$, $V_G > V_T$.



$$g_D = \frac{I_D}{V_D} = \frac{Z}{L} \mu C_{ox} (V_G - V_T) \quad V_G > V_T$$

- ▶ To erase the stored charge, a negative bias is put on the control gate or a positive bias on the source/drain.

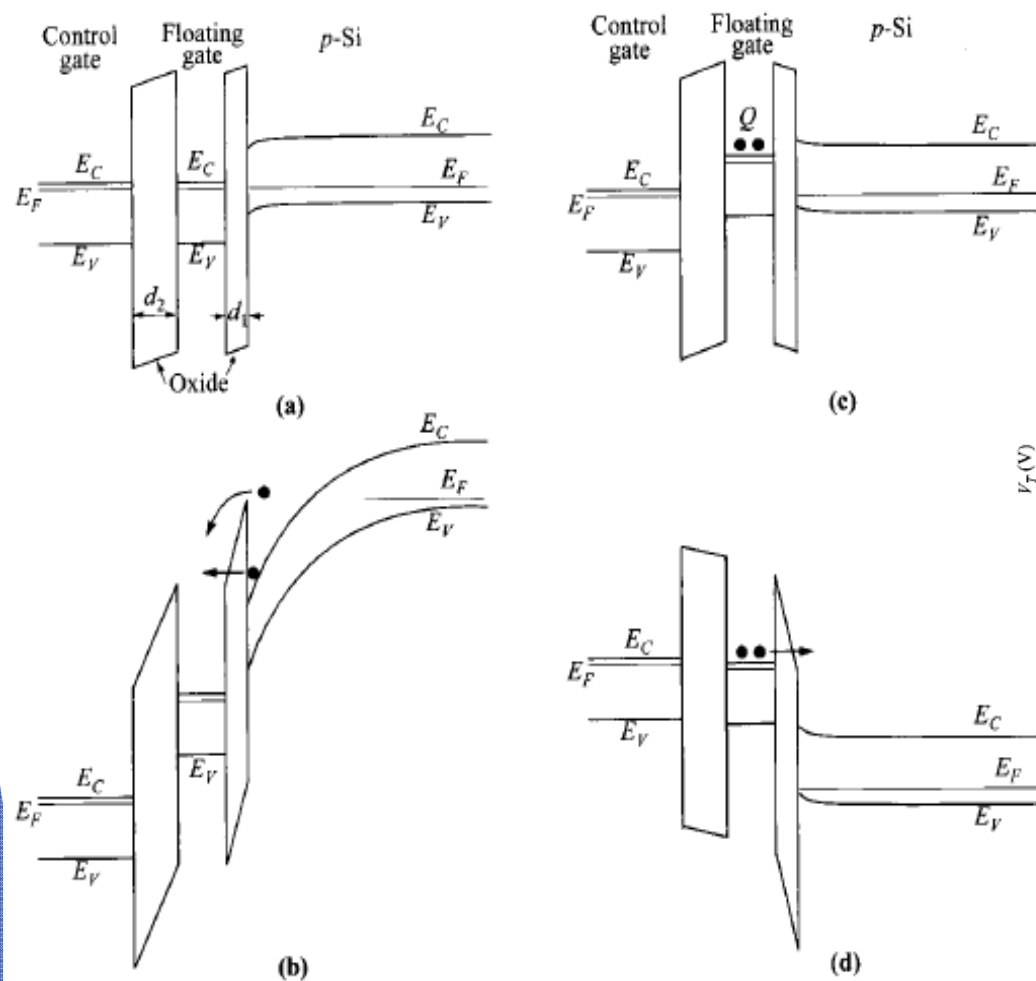


Fig. 47 Energy-band diagrams for a stacked-gate memory transistor at different stages of operation. (a) Initial stage. (b) Charging by hot electrons or electron tunneling. (c) After charging, the floating-gate having charge Q (negative) is at higher potential and V_T is increased. (d) Erasing by electron tunneling.

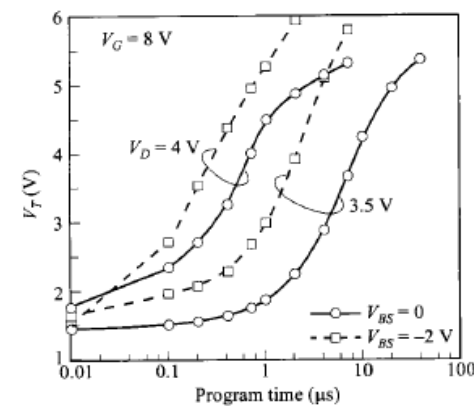
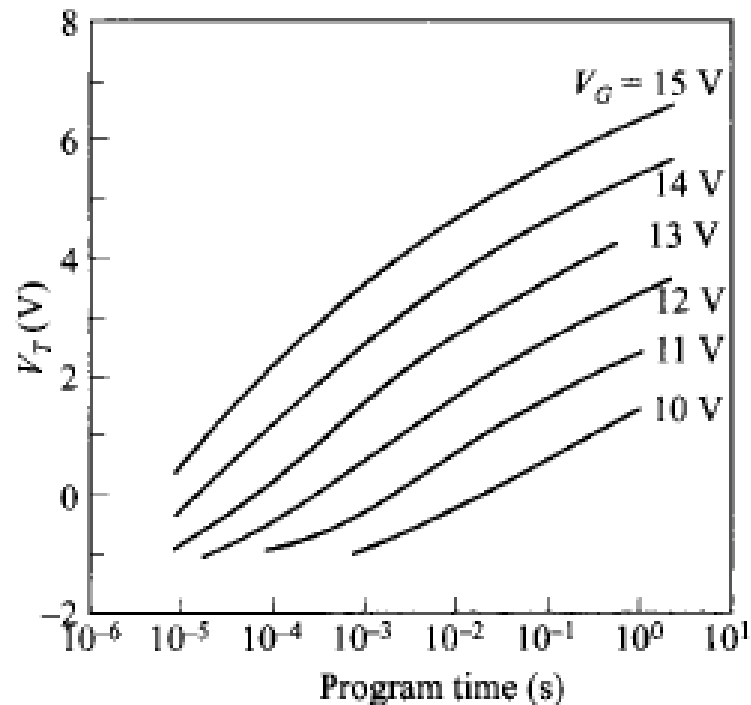


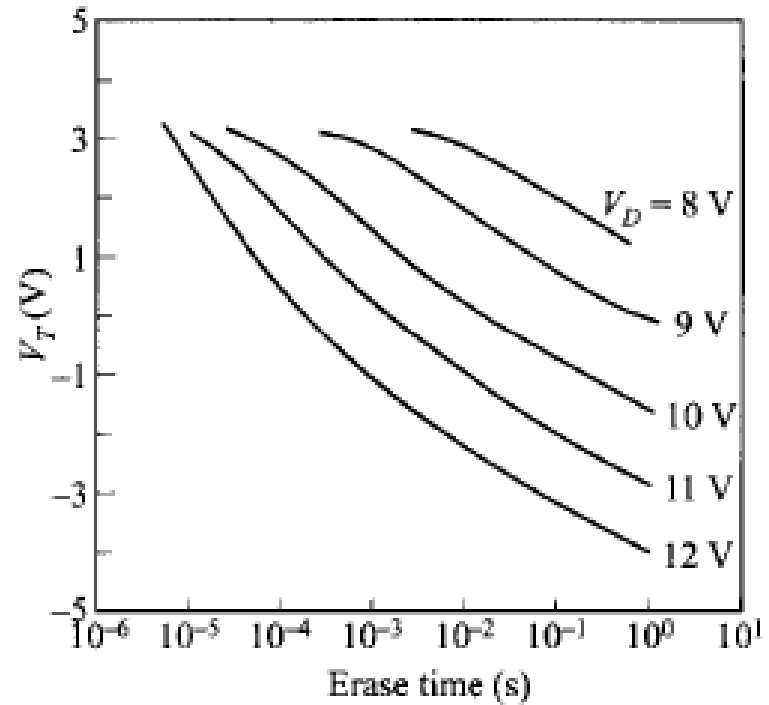
Fig. 48 Programming of floating-gate memory using hot-electron injection. (After Ref. 80.)

The floating-gate potential is given by

$$V_{FG} = R_{CG} V_G.$$



(a)



(b)

Fig. 50 Typical programming and erasing times for FLOTOX memory device. (After Ref. 76.)

$$t_R = \frac{\ln(2)}{v \exp(q\phi_B/kT)} \quad (122)$$

6.8 SINGLE-ELECTRON TRANSISTOR

- ▶ With the continuing advancement of technology to nano-scale device geometry, there are new experimental observations that have not been possible before.
- ▶ One of them is the charge-quantization effect in
- ▶ The structure of the single-electron transistor is shown in the schematic diagram below.

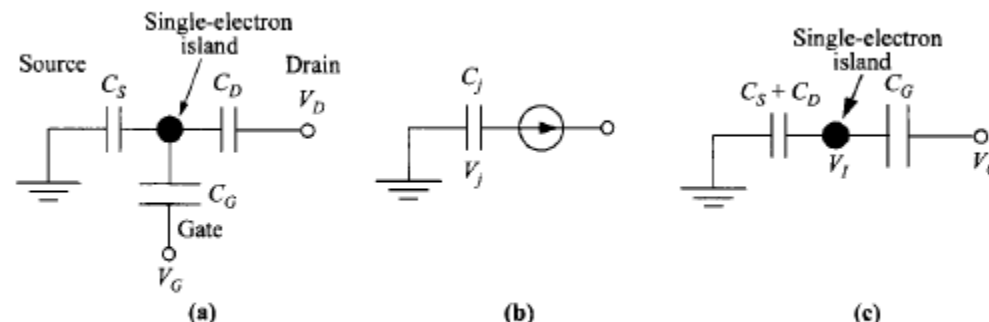


Fig. 53 Circuit representations of (a) single-electron transistor, (b) charging a tunneling capacitor, and (c) single-electron box.

- ▶ It has a central single-electron island that has to be extremely small
- ▶ The island is connected between the source and drain via capacitors through which tunneling occurs to conduct current.
- ▶ The third terminal is the insulated gate and its purpose is to control the current between the source and drain, similar to the case of an FET.
- ▶ The opportunity to observe quantization of charge comes directly from the small dimension of the single-electron island.
- ▶ The minimum energy needed to transport a single electron charge to and from the island is $q^2/2C_{\Sigma}$
- ▶ where C_{Σ} is its total capacitance

$$C_{\Sigma} = C_S + C_D + C_G$$

- ▶ This energy also must be much larger than the thermal energy for experimental observation, requiring that:

$$\frac{q^2}{2C_{\Sigma}} > 1000KT$$

- ▶ The capacitors between the island and the source or drain are characterized by the tunneling resistances R_{Ts} and R_{TD} .

$$R_{Ts} \approx RTD > \frac{h}{q^2}$$

- ▶ ($h/q^2 = 25.8 \text{ k Ohm}$) and they should be above

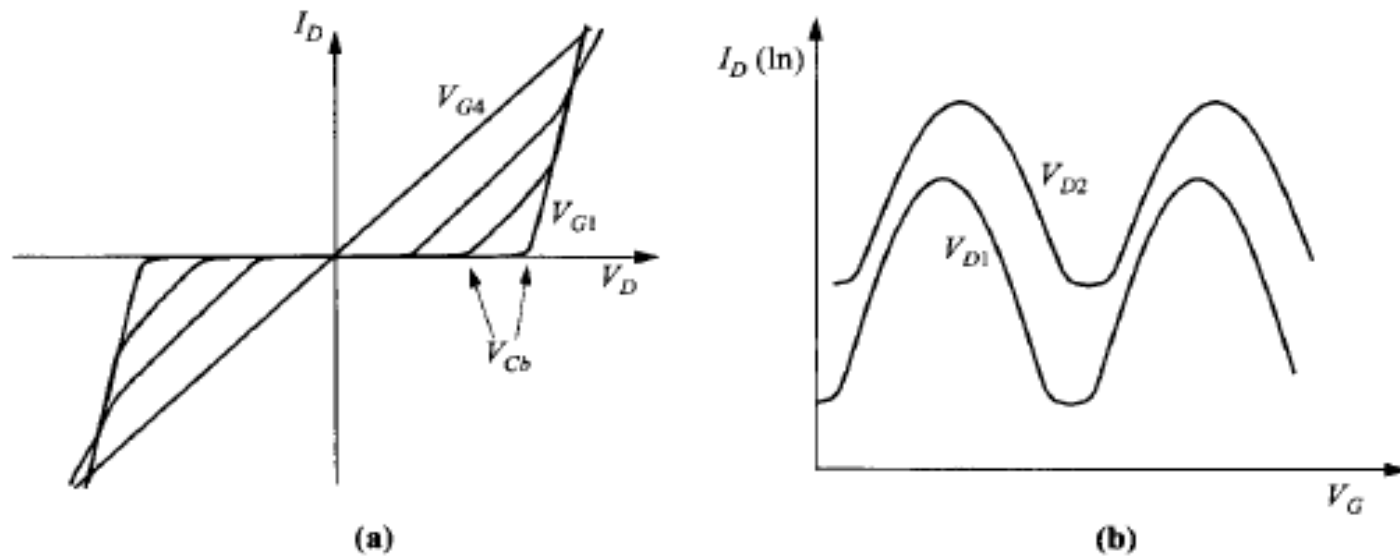


Fig. 54 (a) I - V characteristics of SET for various V_G . The Coulomb-blockade voltage can be varied by V_G . (b) Drain current (logarithmic scale) as a function of V_G for various V_D . Note that V_G is shifted by V_D .

- ▶ the capacitor is charged by a small current source, so the junction voltage V_j will increase until an electron can tunnel.
- ▶ The basis for the Coulomb blockade is that it requires a certain minimum V_j before there is enough energy for an electron to tunnel.
- ▶ The minimum energy needed is $q^2/2C_j$ which will be the change of energy of the capacitor when an electron tunnels.
- ▶ This is also the same as the energy gained by the electron when tunneling across the capacitor of voltage V_j

$$\frac{q^2}{2C_j} = qV_j$$

$$\frac{q^2}{2C_j} = qV_j$$

- ▶ Next, we consider a single-electron box where an island is placed between two capacitors, the same as the situation when the source and drain of an SET is tied together
- ▶ As the gate voltage is increased, the island voltage (V_I) is also increased accordingly, although scaled down by a factor of C_G/C_Σ .
- ▶ It can be seen that the gate voltage at which multiple values of N_i can coexist is a N

$$V_G = \frac{q}{C_G} \left(Ni + \frac{1}{2} \right)$$

- ▶ This condition implies degeneracy: multiple Ni can exist without a change of energy and one electron can tunnel freely to and from the island.

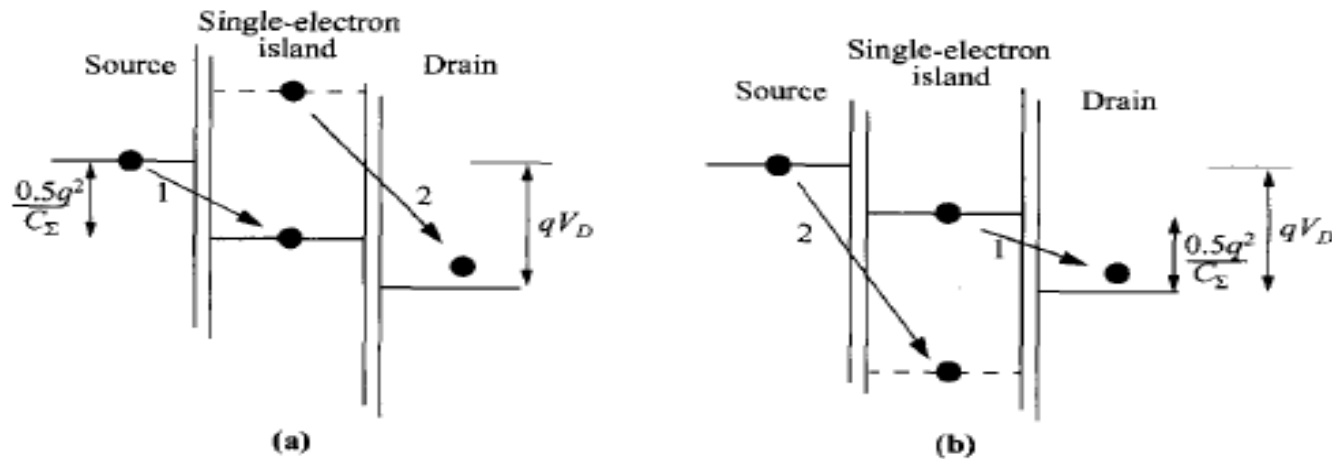


Fig. 57 Energy-band diagrams showing the sequence of tunneling events when the triggering process occurs at the junction (a) between the single-electron island and the source, and (b) between the island and the drain. Event-1 occurs before event-2. Note that the island potential changes by q/C_Σ after each tunneling event.

- ▶ the charging energy of a single-electron box:

$$E_{ch} = \frac{Ni^2q^2}{2C_\Sigma} - \frac{N_iV_G C_G}{C_\Sigma}$$

- ▶ We can now return to the SET and explain the two most-important phenomena:
 - ▶ the Coulomb blockade
 - ▶ Coulomb-blockade oscillations.

