N71-32437

NASA CR-111873

AN EVALUATION OF SOLID STATE PHENOMENA AND DEVICES FOR SENSOR APPLICATIONS

CASE FILE COPY

Final Report on

NASA Contract NASI-9316

January 1971

L. K. Monteith
M. A. Littlejohn
J. R. Hauser

Department of Electrical Engineering
North Carolina State University
Raleigh, North Carolina
27607

NASA CR-111873

AN EVALUATION OF SOLID STATE PHENOMENA AND DEVICES FOR SENSOR APPLICATIONS

Final Report on

NASA Contract NASI-9316

January 1971

L. K. Monteith
M. A. Littlejohn
J. R. Hauser

Department of Electrical Engineering
North Carolina State University
Raleigh, North Carolina
27607

ABSTRACT

This report documents the results of a survey of solid state phenomena which can be employed in sensors and transducers. The phenomena are divided for study into bulk phenomena and semiconductor junction phenomena. The third major section is an assessment of state-of-the-art sensors with emphasis on their limitations. As part of the assignment, the authors make the following observations and recommendations based on the results of the survey.

(1) Many bulk effects and devices have been reported and suggested for sensors. Most of these ideas have not been implemented or discarded with cause.

(2) The character of the semiconductor industry is likely to become more consumer oriented in the next decade. The R and D effort available for sensors will decrease and research will swing toward high volume electronics. Consequently, NASA will find industry increasingly unresponsive to its sensor needs. For this reason, it is recommended that NASA establish a strong in-house sensor development program based on contemporary solid state technology.

(3) The following is the recommended priority for the sensor R and D effort based on the probability of quick return.

(a) the improvement of presently available sensors

(b) the exploitation of previously identified effects

(c) the study of new materials

(d) the search for new phenomena

(4) One of the potentially fruitful research areas is in the magnetic properties of solids. There appears to be no reason why the magnetic properties of a large class of materials (ferrimagnetic for example) could

not and should not be used as sensors.

(5) Thin film technology has been advanced to the point where it could be used effectively for producing bulk sensors.

(6) The electrical properties of the transition metal compounds should be investigated for possible application to stress and temperature sensors.

(7) Ultraviolet sensors using bulk or junction effects appear to be feasible.

TABLE OF CONTENTS

Section Page

1.0 THE BULK PROPERTIES OF SOLID STATE MATERIALS AND THEIR APPLICATION AS SOLID STATE SENSORS..... 2

1.1 INTRODUCTION..... 3

1.2 ELECTRICAL PROPERTIES IN SOLIDS - THE BOLTZMANN - RELAXATION TIME APPROACH..... 5

1.2.1 Electrical Conductivity..... 7
1.2.2 Scattering Mechanisms Which Influence Conductivity..... 13
1.2.3 Impurity Effects on Conductivity..... 20
1.2.4 The Transition Metal Compounds..... 31

1.3 DIELECTRIC PROPERTIES OF SOLIDS..... 36

1.3.1 Polarization of Dielectrics..... 38
1.3.2 Sources of Polarization in Crystals..... 40

1.4 FERROELECTRIC MATERIALS..... 45

1.4.1 The Occurrence of Ferroelectricity..... 47
1.4.2 Other Aspects of Ferroelectrics..... 52

1.5 THE INFLUENCE OF EXTERNAL STIMULI ON THE ELECTRICAL PROPERTIES OF SOLIDS..... 54

1.5.1 Piezoresistance - The Effect of Stress and Strain on the Electrical Conductivity of Solids..... 54
1.5.2 The Concept of Strain Gauge Factor..... 58
1.5.3 Piezoresistance in Metals..... 60
1.5.4 Piezoresistance in Semiconductors..... 62
1.5.4.1 The Pinch Effect..... 73
1.5.5 The Piezoelectric Effect..... 74
1.5.6 Modes of Operation of Piezoelectric Materials..... 84
1.5.7 The Pyroelectric Effect..... 93
1.5.8 Magnetoresistance and The Hall Effect..... 98
1.5.9 Photoconductivity in Solids.....113
1.5.9.1 Processes Involved in Photoconductivity.....120
1.5.9.2 Other Aspects of Photoconductivity.....122
1.5.10 The Photodielectric Effect.....139
1.5.11 The Thermoelectric Effect.....142
1.5.11.1 The Thompson Effect.....151
1.5.11.2 The Seebeck Effect.....153
1.5.11.3 The Peltier Effect.....160
1.5.11.4 Thermoelectric Power Generation.....163
1.5.12 Theromagnetic Effects.....171

1.6 MAGNETIC PROPERTIES OF SOLIDS.....175

1.6.1 The Physical Origins of the Magnetic Properties of Solids.....176

1.6.2 The Classification of Magnetic Materials..... 179
1.6.3 Ferromagnetic Materials..... 189
1.6.4 Antiferromagnetic Materials..... 197
1.6.5 Ferrimagnetic Materials..... 200
1.6.6 Magnetostriction..... 205
1.6.7 Pyromagnetic Effect..... 210

2.0 DEVICE PROPERTIES AND JUNCTION SENSORS.....216

2.1 SEMICONDUCTOR P-N JUNCTIONS.....218

2.1.1 Basic Theory.....219
2.1.1.1 Basic Properties.....219
2.1.1.2 Junction Space Charge Region Properties.....221
2.1.1.3 Current Densities at p-n Junctions.....223
2.1.1.4 Avalanche Multiplication at Large Reverse Bias...230
2.1.1.5 Tunnel Diode Properties.....232
2.1.1.6 p-i-n Diode Properties.....236
2.1.1.7 Noise.....238
2.1.1.7.1 Fabrication Techniques for p-n Junctions.240
2.1.1.8 Dynamic Properties of Devices.....241

2.1.2 Optical Sensors.....242
2.1.2.1 Introduction.....242
2.1.2.2 Basic Optical Properties of p-n Junction.....243
2.1.2.3 p-i-n Photodiode.....262
2.1.2.4 Avalanche Photodiode.....266
2.1.2.5 Frequency Response of Photodiodes.....269
2.1.2.6 Noise Properties.....274

2.1.3 Particle Sensors.....277

2.1.4 Temperature Sensors.....280

2.1.5 Stress Sensors.....287

2.1.6 Magnetic Field Sensors.....310

2.2 SEMICONDUCTOR HETEROJUNCTIONS.....321

2.2.1 Basic Theory.....322

2.2.2 Optical Sensors.....338

2.2.3 Stress Sensors.....347

2.2.4 Other Heterojunction Sensors.....360

2.3 OTHER SEMICONDUCTOR p-n JUNCTION DEVICES.....361

2.3.1 Bipolar Transistors.....361

2.3.2 P-N-P-N Devices....369

2.3.3 Junction Field Effect Transistors....374

2.4 METAL - SEMICONDUCTOR DEVICES....381

2.4.1 Basic Theory....381
2.4.2 Optical Sensors....395
2.4.3 Stress Sensors....411
2.4.4 Other Sensors....415

2.5 METAL INSULATOR SEMICONDUCTOR (MIS) DEVICES....417

2.5.1 Basic Theory....418
2.5.2 Optical Sensors....431
2.5.3 Stress Sensor....434
2.5.4 Magnetic Field Sensors....444
2.5.5 Other Sensors....452

2.6 SENSOR ARRAYS....457

2.6.1 Optical Sensors....458
2.6.2 Other Sensors....466

3.0 STATE-OF-THE-ART FOR SOLID-STATE SENSORS....474

3.1 TRANSDUCTION PRINCIPLES FOR MECHANICAL TRANSDUCERS....476

3.1.1 Strain-Gauges....476
3.1.1.1 Metal Wire Strain Gauges....477
3.1.1.2 Semiconductor Strain Gauges....491

3.1.2 Junction Strain Gauge....510
3.1.2.1 P-n Junction (Piezojunction)....511
3.1.2.2 Tunnel Diode....517
3.1.2.3 Other Junction Devices....518

3.1.3 Thin Film Strain Gauges....520

3.1.4 Piezoelectric Transduction....522

3.1.5 Magnetic Transduction....529

3.1.6 Summary for Mechanical Transducers....533

3.2 OPTICAL TRANSDUCTION....538

3.2.1 Photoemission....540
3.2.1.1 Ultraviolet Photoemission (2000 A°- 3500 A°)....543
3.2.1.2 Ultraviolet Photoemission (1050 A°- 2000 A°)....546
3.2.1.3 Visible Photoemission (3500 A°- 7000 A°)....547
3.2.1.4 Infrared Photoemission....553

3.2.2 Photoconductivity..... 554
3.2.2.1 Ultraviolet Photoconductors..... 556
3.2.2.2 Visible Photoconductors..... 559
3.2.2.3 Infrared Photoconductors..... 573

3.2.3 Photo-Junction Transduction..... 587
3.2.3.1 Ultraviolet Photojunctions..... 589
3.2.3.2 Visible Photojunctions..... 593
3.2.3.3 Infrared Photojunctions..... 612
3.2.3.4 Summary of Optical Transduction 620

4.0 CONCLUSIONS AND RECOMMENDATIONS..... 632

4.1 STRESS SENSORS - BULK EFFECTS..... 633

4.2 STRESS SENSORS - JUNCTION DEVICES..... 635

4.3 OPTICAL SENSORS - BULK EFFECTS..... 636

4.4 OPTICAL SENSORS - JUNCTION DEVICES..... 637

4.5 TEMPERATURE SENSORS - BULK EFFECTS..... 638

4.6 TEMPERATURE SENSORS - JUNCTION DEVICES..... 639

4.7 MAGNETIC SENSORS - BULK EFFECTS..... 640

4.8 MAGNETIC FIELD SENSORS - JUNCTION DEVICES..... 641

4.9 CHEMICAL SENSORS..... 642

4.10 INTEGRATED SENSORS..... 643

4.11 SENSOR RESEARCH AND DEVELOPMENT..... 645

LIST OF FIGURES

Figure — Page

1.1 Experimental Hall, resistivity, and mobility data for a sample of p-type germanium.[from D. M. Brown and R. Bray, Phys. Rev. 127: 1597 (1962).] 14

1.2 A plot of the electrical resistivity of several metals illustrating the proportionality to absolute temperature 18

1.3 Mobility-temperature curves for pure lattice scattering, neglecting scattering by optical modes. The quantities C_1 and C measure the strength of the coupling of the carriers to intra - and intervalley modes, respectively; ω^1 is the frequency of the intervalley mode..19

1.4 The resistance of three samples of sodium.....21

1.5 The resistivity of copper-nickel alloys showing the existence of the same temperature-dependent resistivity superimposed on varying amounts of residual resistance.....22

1.6 The resistivity of copper-gold alloys. The full line refers to annealed samples and the dashed line to quenched samples.....24

1.7 Intrinsic carrier densities n_i in silicon and germanium, obtained as best fits to experimental data.....27

1.8 Electron concentration in n-type silicon as a function of temperature 29

1.9 The Fermi level in silicon as a function of temperature, for various impurity concentrations.....31

1.10 Conductivities of oxides (Morin[36]).....32

1.11 Resistivity vs reciprocal temperature at different pressures for $(V_{0.96}Cr_{0.04})_2O_3$. Inset shows I-M transition at room temperature as a function of pressure. Below 10 kbar the AF → I transition occurs on warming. Above 10 kbar the AF → M transition occurs. Above 40 kbar the sample is metallic from 4.2°K to 298°K..... 35

1.12 The dependence upon applied frequency of the real and imaginary parts of the complex dielectric constant in an ionic solid containing permanent dipoles. The imaginary part of the dielectric constant (or dielectric loss factor) is large whenever the frequency reaches the upper limit for a dielectric response mechanism. This happens for dipole orientation in the microwave range, for ionic polarizability in the infrared, and for bound electron displacements in the ultraviolet.....42

1.13 (a) Ferroelectric hysteresis loop, the coercive force is E_c (b) Circuit for display of loop. The voltage across the ferroelectric crystal C_x is applied to the horizontal plates of the oscilloscope. The linear capacitor C_o is in series with C_x. The voltage across C_o is proportional to the polarization of C_x and is applied to the vertical plates..... 46

1.14 Calculated values (a) of the spontaneous polarization and (b) of the static dielectric constant as a function of temperature, with parameters as for barium titanate..... 51

1.15 Spontaneous polarization versus temperature, for a second-order phase transition. (The curve is not realistic at low temperatures: the third law of thermo-dynamics requires that $dP_s/dT \to 0$. 52

1.16 Change of resistance of the alkali metals at room temperature with change of pressure 62

1.17 (a) E(k) curves in <111> and <100> directions in Si and Ge. The important energy band separation values at 0°K, determined mainly from optical measurements at liquid helium temperature, are indicated. The levels at the principal symmetry points are labeled in single group notation, and the corresponding double group labels are given in parentheses..... 66

1.17 (b) E(k) curves in <111> and <100> directions in GaAs. The important energy differences determined experimentally at o K are indicated. Notations are as in Figure 1.17(a) 67

1.18 The model used by Lord Kelvin to explain the piezoelectric effect in quartz..... 76

1.19 Deformation Modes of Single Piezoelectrical Elements 86

1.20 A thin piezoelectric crystal under an alternating electric field..... 87

1.21 The equivalent circuit of a piezoelectric crystal under the applied voltage condition of Figure 1.20 and also in the longitudinal mode of vibration..... 89

1.22 Piezoelectric transformers..... 91

1.23 Relation between voltage amplifications and load resistance for transverse ceramic transformer with constant voltage drive =0.09 volt..... 93

1.24 Relative response as a function of frequency deviation for a ring-type ceramic filter with various loads..... 94

1.25 Temperature dependence of the spontaneous polarization of triglycine sulphate as deduced from pyroelectric data (circles) and from hysteresis data (triangles)..... 96

1.26 The transverse magnetoresistivity of 22 metals plotted on a reduced Kohler diagram..... 105

1.27 Magnetoresistive effect in Ge plotted as a function of temperature at B = 0.4 weber/m^2. The Ge sample is n-type and has a resistivity of 11.5 Ω-cm..... 106

1.28 Variations in $\Delta\rho/\rho H^2$ as H is rotated with respect to I in n-type silicon. Certain cardinal directions of H are as indicated. The fixed directions of I are as given..... 107

1.29 Relative resistance R_B/R_o of field-plates as a function of the magnetic induction B. D, intrinsic InSb-NiSb. L, M, doped InSb-NiSb. P, T, raster plates with doped InSb..... 112

1.30 Temperature coefficient $\Delta R/R_B$. ΔT of field-plates at room temperature as a function of the magnetic induction B. For symbols identifying curves see Figure 1.29..... 113

1.31 Basic arrangement for observation of photoconduction..... 115

1.32 Variation of photo-current with wavelength for large and small values of the surface recombination velocity s..... 119

1.33 Spectral response of typical "intrinsic" photoconductors..... 120

1.34 Typical electronic transitions in an n-type photoconductor..... 122

1.35 Factor of merit, D* (or N.E.P.)$^{-1}$ in unit band-width (for detector of unit area) for ideal photon detector assuming that incident radiation is just below cut-off wavelength λ_c. The detector and surroundings are at a temperature of 77° K(curve (a)), 195° K(b) 290°K (c), and 500° K(d).....128

1.36 Factor of merit, D*, for various detectors operated at room temperature: (a) thallium-sulphide photoconductive cell; (b) PbS photoconductive cell; (c) PbSe photoconductive cell; (d) InSb photoconductive or PEM cell; (e) Hilger-Schwartz thermocouple; (f) Golay cell; (g) thermistor bolometer; (h) represents the thermal noise limit for a non-selective detector operated at represents the thermal noise limit for a non-selective detector operated at 300°K..... 130

1.37 Factor of merit, D*, for various photoconductive detectors operated at 77°K; (a) PbS; (b) InSb; (c) PbSe; (d) Au-doped Ge, p-type: (e) PbTe; (f) represents the thermal noise limit if the surroundings of the detector are at 300°K..... 131

1.38 Relationship between peak detectivity and temperature-long wavelength limit product..... 132

1.39 The spectral detectivity of a number of low temperature infrared detectors. Typical frequency of the noise measurement was 1000 Hz..... 133

1.40 Representative spectral relative response curves for various compositions (percent). The curves are not normalized to each other..... 134

1.41 The curve of the specific heat of a solid (per mole) as a function of temperature, according to the Debye model in three dimensions. The experimental points have been plotted here with an abscissa scale appropriate for θ_D of 200°K. (An optimum fit for the lowest temperature points would require a slightly higher Debye temperature). Note that the experimental points are above the Debye curve at high temperatures because C_p is the quantity actually measured, not C_v..... 149

1.42 Schematic illustration of the Seebeck Effect..... 155

1.43 The thermoelectric powers of three useful low-temperature thermocouples compared with each other. (a) Au - 2.1 at .-% Co against Cu,[357] (b) Cu against constantan,[357] and (c) Au-0.03 at. -% Fe against chromel[74]..... 157

1.44 Thermoelectric power of a semiconductor. At low temperatures the simple formulas fail because of failure of the relaxation-time approximation. In an extrinsic material, S increases slightly with temperature in the range where minority carriers are negligible because of the temperature dependence of N_c or N_v..... 161

1.45 Thermolectric power of p-type Ge (1.5×10^{14} acceptors per cubic centimeter). Dashed curve represents the calculated value. The difference between the curves is the result of phonon drag..... 162

1.46 Schematic illustration of the Peltier Effect..... 163

1.47 A simple semiconductor thermoelectric device, which may be used as either a generator (as shown here) or a refrigerator..... 164

1.48 Electrical and thermal conductivities and thermoelectric coefficient as a function of carrier concentration. (It is implied that band gap decreases toward the right.)..... 167

1.49 Maximum coefficient of performance Φ_{max} as a function of figure of merit Z for different hot junction temperatures T_H(T_C = 300°K)...168

1.50 (a) The temperature dependence of the figure of merit, Z for p-type telluride semiconductors and their alloys.....169

1.50 (b) The temperature dependence of the figure of merit, Z for n-type telluride semiconductors and their alloys.....170

1.51 Plot of susceptibility per gm versus reciprocal temperature for powdered $CuSO_4.K_2SO_4.6H_2O$, showing the Curie Law temperature dependence.....182

1.52 Characteristic magnetic susceptibiliteis of diamagnetic and paramagnetic substances..... 184

1.53 Hysteresis loop of ferromagnet..... 186

1.54 The low temperature ordering (if any) of neighboring dipoles, and the consequent behavior of spontaneous magnetization and/or susceptibility, for A, paramagnetism, B, antiferromagnetism, C, ferromagnetism, and D, ferrimagnetism..... 188

1.55 Diagram illustrating how Eq. (1.163) can be solved graphically for $H_{ext}=0$. A solution for the value of spontaneous magnetization exists only when the straight line whose slope is a function of temperature intersects the curve..... 191

1.56 Relative magnetization versus T for the three ferromagnetic elements of the iron group..... 192

1.58 Enhancement of magnetic induction by cylindrical rods as a function of the ratio p=length of rod/rod diameter, with the permeability μ as parameter..... 195

1.59 The susceptibility of an antiferromagnetic material as a function of temperature in reduced units..... 200

1.60 Comparison of the behaviors of paramagnetic susceptibility as a function of temperature in antiferromagnetic, ideal paramagnetic, and ferromagnetic materials..... 201

1.61 Reciprocal susceptibility of magnetite, $FeO.Fe_2O_3$, above the Curie temperature 203

1.62 Stress-strain relation for ferromagnetic material compared with that for a nonmagnetic material of equal stiffness..... 207

1.63 Magnetostriction of some common materials, showing expansion or contraction.....208

1.64 The influence of stress on the magnetization of Permalloy..... 209

2.1 Basic p-n Junction.....,220

2.2 Typical theoretical characteristic of p-n junction diode [7].....228

2.3 Experimental I-V characteristics of Ge, Si, and GaAs diodes [4]..229

2.4 Typical dependence of reverse current on voltage near breakdown..232

2.5 Energy bands for tunnel diode.....234

2.6 Typical current-voltage characteristics of Ge, GaSb, and GaAs tunnel diodes at 300°K. The ratio I_p/I_v is 8/1 for Ge and 12/1 for GaSb and GaAs tunnel diodes.....235

2.7 P-I-N Junction diode. (a) Structure (b) Impurity distribution...237

2.8 P-N Junction used as optical sensor.....245

2.9 Reflectivity of Si for photon energies above the band gap energy..246

2.10 Measured absorption coefficients near and above the fundamental absorption edge of pure Ge, Si, and GaAs [9].....247

2.11 Model for p-n junction optical detector.....248

2.12 Sketch of terminal I-V characteristic of photodiode.....255

2.13 Effective quantum efficiency η as a function of wavelength for a silicon and a germanium photodiode. (Laser lines within this wavelength range are indicated) [16].....257

2.14 Internal efficiency or response of GaAs p-n junctions with (a) illuminated p-surface and (b) illuminated n-surface. (1) 370°K (2) 293°K (3) 90°K, after Gutkin et. al. [17].....258

2.15 Measured and calculated values of diode sensitivity. Both the measured quantum efficiency η and the calculated internal quantum η_i efficiency are shown. After Morse, et.al. [19].....260

2.16 Open circuit voltage vs. illumination intensity [20].....262

2.17 Semiconductor p-i-n photodiode. (a) Basic structure (b) Light intensity versus distance (w is depletion layer width).....264

2.18 Experimental results on change in relative response (R) with bias (V_B) for n+-n-p+ type GaAs detector [22].....265

2.19 Cross-sectional view of planar photodiode (Ref. [23][24]).....267

2.20 Static current-voltage characteristics of a Ge photodiode for various primary currents [16].....268

2.21 Typical multiplication vs. bias for avalanche photodiode. The different curves correspond to different dc photocurrents measured at low reverse voltage [25].....270

2.22 Radiation wavelength versus cutoff frequency operating region for Ge and Si photodiodes ([9] page 668).....273

2.23 (a) Photodiode equivalent circuit. (b) Noise equivalent circuit.....275

2.24 Signal and Noise power output of a Ge avalanche photodiode. Both are measured with respect to 1 mW power level [9].....276

2.25 Illustration of p-i-n particle detector.....279

2.26 Variation of p-n junction voltages with temperature at a constant forward current....283

2.27 Variation of forward voltage for a GaAs diode with temperature [30]....285

2.28 Typical zener diode temperature coefficient vs. breakdown voltage [31]....287

2.29 Reverse emitter-base breakdown voltage vs. temperature for integrated small geometry transistor [29]....288

2.30 Forward-biased I-V characteristic of a mesa diode subjected to stress (the applied force is shown) [35]....289

2.31 Typical changes in minority carrier lifetime with stress for silicon (Stress in <100> direction) [37]....292

2.32 Effect of a general stress on energy bands of silicon....294

2.33 Theoretical ratio of stressed to unstressed minority carrier density for a Hydrostatic, [100], [011], and [111] uniaxial, compressive stress for silicon diodes [43]....299

2.34 Forward I-V characteristics, limited to the first positive resistance region of an n-type GaSb tunnel diode as a function of hydrostatic pressure....309

2.35 Current-voltage characteristics at 77°K for the InSb tunnel diode in different magnetic fields perpendicular to the current flow across the p-n junction [49]....312

2.36 p-n Junction showing current flow and magnetic field....313

2.37 Calculated current distribution for a p-n junction in a magnetic field (series resistance neglected) [51]....314

2.38 General view of a magnetodiode. a. Ideal magnetodiode. 1) metal; 2) p-type semiconductor; 3) n-type semiconductor; 4) p-n junction. b. Test specimen. 1) indium; 2) tin; 3) n-type Ge; 4) p-n junction....317

2.39 Forward branches of I-V characteristics of magnetodiode in different magnetic fields. 1) B=0, 2) B=2.5 kgauss, 3) B=5.3 kgauss, 4) B=7.8 kgauss, 5) B=10.1 kgauss, 6) B=12.6 kgauss, 7) B=19.5 kgauss, 8) B=23 kgauss....320

2.40 Energy band diagram of n-p abrupt heterojunction with a positive reverse barrier [57][58]....323

2.41 Energy band diagrams of n-p heterojunctions with (a) negative reverse barrier and (b) no reverse barrier....325

2.42 Illustration of energy band diagram of n-n heterojunction.....327

2.43 Expected energy band diagram for a graded band gap p-n heterojunction.....328

2.44 Graphical solution of voltage equations for heterojunction.....332

2.45 Illustration of three types of electron barriers in a heterojunction.....333

2.46 Heterojunction optical sensor with semiconductors of two different band gaps.....339

2.47 Illustration of heterojunction response to different wavelengths for $E_{gp} > E_{gn}$.....342

2.48 Short-circuit photocurrent of a GaAs-Ge p-n heterojunction per incident photon vs. photon energy [60].....344

2.49 Relative photoresponse current spectrum for device 42 GaAs-GaP....345

2.50 Illustration of the effect of a piezoelectric semiconductor on the space charge regions of a heterojunction, n-type material assumed to be piezoelectric.....352

2.51 Graphical solution of heterojunction voltage equations for a net piezoelectric effect (Q_I=0 for this case).....353

2.52 Effect of strain on forward I-V characteristics of CdSe -Se heterojunction [67].....357

2.53 Voltage-strain characteristic of CdSe-Se heterojunction [67].....358

2.54 Cross sectional view of bipolar transistor.....362

2.55 Operation of phototransistor.....363

2.56 Typical collector current vs. collector voltage characteristics of a phototransistor [31].....365

2.57 Illustration of transistor used as stress sensor.....366

2.58 Ratio of stressed to unstressed collector current of a 2N1958 n-p-n transistor with constant base current and constant emitter current [40].....367

2.59 P-N-P-N device and equivalent n-p-n, p-n-p transistor model.....370

2.60 Sketch of p-n-p-n device characteristics.....371

2.61 Simple oscillator circuit using p-n-p-n device.....372

2.62 Oscillator amplitude and frequency change (Δf) as a function of force for a p-n-p-n oscillator under stress [70].....373

2.63 Sketch of junction field effect transistor..... 375

2.64 Terminal characteristics of junction field effect transistor..... 377

2.65 Equivalent circuit of junction FET employed as a photodetector.... 380

2.66 Energy band diagram for metal semiconductor device..... 383

2.67 Energy band diagram near the interface showing Schottky barrier lowering..... 386

2.68 Energy band diagrams for metal semiconductor barrier under forward and reverse bias..... 388

2.69 Illustration of thermal emission and tunneling currents for metal semiconductor barrier..... 389

2.70 Illustration of two modes of incident photon energy..... 396

2.71 Excitation of carriers within metal for metal-semiconductor junction detector..... 397

2.72 $\sqrt{\eta h\nu}$ vs $h\nu$ for 60 A° and 2000 A° Au-p-type Si diodes at 77°K... 401

2.73 $\sqrt{\eta h\nu}$ vs $h\nu$ for 40 A° and 2000 A° Au-p-type Ge diodes at 5°K.....403

2.74 Efficiency of a Cr-GaAs metal semiconductor diode as a function of photon wavelength [78].....409

2.75 Forward-bias voltage-current characteristic of 0.25 mm radius Au-$KTaO_2$ diode for various forces. Steel stylus radius, 0.25 mm [79].....413

2.76 Standard MIS device structure.....419

2.77 One possible variation of MIS geometry for thin film device.....420

2.78 Types of surface space charge regions for an MIS device with a p-type semiconductor.....421

2.79 Ideal MIS capacitance -voltate curves. (a) Low frequency, (b) High frequency, (c) Nonequilibrium or transient case.....425

2.80 Sketch of MIS device with an applied drain-to-source voltage.....428

2.81 Theoretical characteristics of ideal MIS field effect transistor...430

2.82 Effect of illumination on C-V characteristics of silicon MIS device [82].....432

2.83 Equivalent circuit of MIS transistor used as a photodetector.....434

2.84 Stress dependent surface charge density of piezoelectric MIS device.....441

2.85 Equivalence circuit of MIS piezoelectric transistor used as a stress sensor..... 444

2.86 Structure of MIS Hall effect device..... 445

2.87 Dependence of Hall voltage on position along channel..... 449

2.88 Operation of MIS Hall device with Hall electrodes at saturation for increased sensitivity [87].....450

2.89 Pyroelectric MIS temperature sensor..... 453

2.90 MIS (or MOS) capacitor type meteroroid detector..... 455

2.91 Electron beam scanning of diode optical sensor array..... 460

2.92 XY scanned phototransistor array [91]..... 463

2.93 Output response of XY scanned phototransistor in the change storage mode [91]..... 465

3.1 Temperature effects on gauge factor [4]..... 485

3.2 Nonlinearity of two gauges at room temperature [6]..... 486

3.3 Resistance bridge..... 487

3.4 Gauge-factor K of p and n silicon for different crystal orientations as function of resistivity ρ [111]..... 492

3.5 Gauge factor as a function of temperature [10] 493

3.6 Apparent strain $\Delta\varepsilon$ of a gauge bonded to steel as function of the temperature T. [11]..... 496

3.7 Resistance as a function of temperature [10]..... 497

3.8 Relative resistance change as a function of strain of lightly doped p-type silicon at room temperature [2]..... 498

3.9 Typical values of F_1 and F_2 for unbonded p-silicon gauges [7]..... 499

3.10 Relative resistance change as a function of strain of heavily doped p- and n-type silicon at room temperature. [2]..... 500

3.11 Gauge factor vs. temperature for n- and p-types (unbounded) tested in tension. [15]..... 504

3.12 $\Delta R/R$ vs. temperature for selected resistivities and types of semiconductor elements. Note: Gauges are unbounded [15]..... 505

3.13 $\Delta R/R$ vs. strain for n- and p-types. [15]..... 506

3.14 Compensated bridge using p-type and n-type gauge elements..... 507

3.15 Pressure transducer configuration utilizing the needle sensor. [20].....516

3.16 Simple equivalent circuit for piezoelectric transduction.....525

3.17 Variation of the d_{11} coefficient of quartz with temperature, referred to d_{11} at 20°C [37].....527

3.18 Variation of dielectric constant of quartz at higher temperature [37].....528

3.19 Variation of resistivity of quartz at higher temperature. [37]..528

3.20 Illustration of Hall voltage.....532

3.21 Relative spectral response characteristic for S-13 photocathode [53].....544

3.22 Simple model for analyzing pulsed response of photocathode.....546

3.23 Spectral response curves of alkali diodes.[54].....548

3.24 Relative spectral response characteristic for S-20 photocathode [53].....552

3.25 Relative spectral response characteristic for S-1 photocathode [53].....554

3.26 Spectral response curve of Ag-O-Cs cathode showing long wavelength threshold.[56].....555

3.27 Spectral response of photoconductivity for a photosensitive CdS crystal with a trace of iodine [69].....560

3.28 Detectivity of ideal detectors. The Planck photon distribution for 290°K black-body radiation and the detectivities for ideal thermal and photoconductive detectors exposed to this background are shown. The dot on the curve for photoconductive detectors indicates the detectivity of a detector of quantum efficiency (or spectral response) given by the lowest curve. [79].....578

3.29 The spectral detectivity of a number of low temeprature infrared detectors. Typical frequency of the noise measurement was 1000 Hz [82]..... 583

3.30 Absorption coefficient of silicon as a function of wavelength [83]..... 590

3.31 Normal incidence reflectivity as a function of photon energy for single crystal, etched silicon at 300°K. [86]..... 591

3.32 Spectral responsivity of typical silicon junction devices [32]... 592

3.33 Conversion efficiency as a function of energy gap for ideal current voltage characteristics. [90]..... 598

3.34 Spectral response of Pt and $Cu_{1.8}$ Se barrier cells [88] 599

3.35 Spectral responses of various types of Si cells [89]..... 600

3.36 Pass band response of heterojunction..... 604

3.37 Silicon solar cell [96]..... 606

3.38 Effect of illumination on junction's. The characteristic of the dark junction is referred to the solid axes. The characteristic of the illuminated junction is referred to the dotted axes..... 607

3.39 Quantum efficiency, Q, of silicon solar cells [98]..... 609

3.40 (a) Spectral response for p +/n cells No. 903, No. 71 and n+/p cell No. 276. Symbols for identification. Measured points each 0.02 μ (b) Variation of spectral response of p+/n cells as junction depth is decreased. Symbols for identification. Measured points each 0.02 μ [97].....611

3.41 V-I-Φ characteristics of a silicon photodiode [99].....612

3.42 Spectral response for InAs diode [101].....618

3.43 D* versus wavelength for InAs diode [101].....619

LIST OF TABLES

Section 1

Table		Page
1	Ideal Resistivities at 273°K of a Number of Monovalent, Divalent, Trivalent, and Transition Metals	12
2	Electrical Properties and Band Parameters of Some Common Semiconductors	23
3	Energy Band Parameters of Semiconductors and Closely related Semimetals	25
4	Ionization Energies of Donors and Acceptors in Si and Ge,ev	28
5	The Range of Resistivity observed at Room Temperature in Different Materials	30
6	Experimental Values of High and Low Frequency Dielectric Constants for Ionic Crystals	43
7	Relative Permittivities and Losses in Dielectrics at Frequencies of 10^6 to 10^9 Hz	44
8	Ferroelectric Crystal Data	48
9	Elastic Constants in Units of 10^{11} Dynes/cm^2 for Some Cubic Crystals at Room Temperature	57
10	Elastoresistance Coefficients of Some Semiconductors	58
11	Gage Factors of Wires of Different Compositions	63
12	Comparison of Several Properties of a Semiconductor Strain Gage to like Properties of a Strain Gage Wire	64
13	Deformation Potential Constants for Silicon, Germanium, and Gallium Arsenide in Units of eV per Unit Strain	70
14	Characteristics of Some Piezoelectric Transducers and Materials	79
15	Characteristics of Semiconductor Hall Generators	108
16	Hall Effect Properties of Several Semiconductors	110
17	Properties of Various Cadmium Sulfide Photoconducting Materials	126
18	Calculated values of response time for 10^{10} traps per cm^3 per KT	136
19	Experimental values of $K/\sigma T$	147

20 Thompson Coefficients μ in microvolts/degree of the alkalis just above and just below the melting-point..... 154

21 Parameters for some Monovalent Metals at 0°C Theoretical value for $10^8 \rho_o k_T/T$ is 2.45 [joules/(coul) (°K)]^2..... 158

22 The spontaneous magnetization, Curie temperature, and effective number of Bohr magnetons in a selected number of magnetic materials..... 193

23 Data for some Antiferromagnetic Compounds 199

Section 2

1 Measured Schottky Barrier Heights for Selected Semiconductors [9]..... 384

2 Piezoresistance Coefficients for Silicon p-type inversion layers (10^{-12} cm^2/dyne)..... 437

Section 3

1 Properties of Wire Materials for Use in Strain Gauges [2]..... 478

2 Thermal Coefficient of Expansion for Some Common Materials [2]..... 480

3 Properties of Silicon Strain Gauge and Wire Resistance Strain Gauge [11]..... 494

4 Apparent Strain for 0.018 ohm-cm p-type Silicon Strain Gauge Bonded to the Listed Materials [8]..... 501

5 Comparison of Deposited Semiconductors, Resistance Wire and Bulk Semiconductor Strain Gauges [34]..... 522

6 Spectral Response and Related Photocathode Characteristics [53]..... 550

7 Properties of Various Cadmium Sulfide Photoconducting Materials [69]..... 564

8 Examples of Photoconductive Detectors [79]..... 584

9 Efficiencies of Some Solar Cell Material [93]..... 602

10 Minimum Film Thickness Required for Various Materials [93].... 603

An Evaluation of Solid State Phenomena and Devices for Senor Applications

1.0 THE BULK PROPERTIES OF SOLID STATE MATERIALS AND THEIR APPLICATION AS SOLID STATE SENSORS

The object of this section is to discuss the physical mechanisms which control the bulk electrical and magnetic properties of solids. These properties will generally be considered to be independent of the surfaces which are present in any real solid.

There is probably no fundamental electrical property of solid state materials which has not been used in some manner as a sensor. Thus, much of this section is devoted to the consideration of the electrical properties. In contrast, there has been very little exploitation of the magnetic properties of materials for sensor applications. It appears that there are some promising sensor concepts which could be developed using the magnetic properties of materials.

This section has been organized so that the important material phenomena are first discussed. This is followed by a presentation of the uses of the particular phenomena as sensors of the various types, such as optical sensors, mechanical sensors, magnetic field sensors, and temperature sensors. The performance of a specific type of sensor is then compared to those utilizing a different material phenomena. Charts and tables are given to allow rapid comparison of the properties of various materials.

1.1 INTRODUCTION

Any complete review of the properties of solid state materials would fill hundreds of volumes, and before the task could be completed new information would be obtained which would make such a review out-of-date. Fortunately, the scope of this review is limited by its confinement to the application of material properties to bulk effect sensors. Another restriction will also be imposed on this review. In order to keep the length to a reasonable number of pages very little mathematical rigor will be followed. Sufficient references will be given so that the reader can carry out detailed study if desired. Finally, not all bulk effects will be covered, since the intent of this section is to examine those effects which have been, and are hoped to be, prominent in sensor development, both past and future.

It is obvious that their are probably at least six major areas which can be classified as distinct fields dealing with the various properties of solid-state materials. These include the electrical properties, magnetic properties, mechanical properties, optical properties, thermal properties, and chemical properties. In this report all of these areas cannot be covered. Since one is concerned with those properties which relate directly to solid-state sensors, the major emphasis in this section will be on the electrical and magnetic properties. The reasoning here is that the majority of solid-state sensors have as a requirement the fact that the output of the sensor which is used as a measure of the external stimuli invariably is an electrical signal. While this is not always true, it does seem to be a desirable feature.

Using this reasoning, this section on bulk properties can be outlined as follows. First there are two major sections which deal with the electrical and magnetic properties of solid-state materials. In each of these major sections the influence of external stimuli on these properties are discussed. The external stimuli are those quantities, such as light, temperature, force, and magnetic field intensity, which one desires to sense. The electrical and magnetic properteis provide the sensing mechanism for the external stimuli and an output mechanism for measurement. Thus the input-output aspect of solid-state sensors is the major theme of this section.

1.2 ELECTRICAL PROPERTIES IN SOLIDS - THE BOLTZMANN - RELAXATION TIME APPROACH

The electrical properties of solids is an area dealing with the mechanisms by which charge is transported within these materials while under the influence of electric fields. This general area has proved to be one of the more fruitful in terms of its application to sensors. The existing transport theory involves obtaining solutions to the Boltzmann transport equation for the electron distribution function, which may be interpreted as the probabiltiy that an allowed state in phase space is occupied by an electron. From the distribution funtion and a knowledge of the allowed state distribution (the energy bands) electrical current density can be normally calculated as a function of the electric field. If this relation is linear, an electrical conductivity can be defined.

The Boltzmann transport equation can be written as

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial t}\Big)_{coll} + \frac{\partial f}{\partial t}\Big)_{fields}$$

or

$$\frac{\partial f(\bar{k}, \bar{r}, t)}{\partial t} = \frac{\partial f}{\partial t}\Big)_{coll} - \frac{1}{\hbar}\ \bar{F}.\nabla_k f - \bar{v}.\nabla_r f \qquad (1.1)$$

Here f is the probability that an allowed state in phase space (the six-dimensional space composed of three k and three r co-ordinates) is occupied by an electron, k is the magnitude of the electron wave vector, r is the magnitude of the position vector of the electron, and $\bar{F}$ is the force acting on the electron. Ther term $\partial f/\partial t)_{coll}$ represents the influence of the scattering mechanisms on the distribution function

and is of basic importance. Generally, for calculational purposes this term is taken as

$$\left(\frac{\partial f}{\partial t}\right)_{coll} = -\frac{f - f_o}{\tau} \qquad (1.2)$$

where f_o is the equilibrium Fermi-Dirac distribution function and τ is the relaxation time. Generally, τ is temperature and energy dependent. Equation (1.2) is called the relaxation time approximation. Its usefulness and correctness have quite often been debated [1]; but when a relaxation time can be shown to exist the approximation is extremely useful in solving Equation (1.1).

Once Equation (1.1) is solved for f, the electric current density can be calculated from

$$\bar{J}(\bar{r}, t) = \frac{-q}{4\Pi^3} \int_{\text{all } \bar{v}} f(\bar{r}, \bar{k}, t)\, \bar{v}\,(\bar{k})\, d\bar{k} \qquad (1.3)$$

where the integral is taken over all allowed electronic states in the first Brillouin zone. Here q is the magnitude of the electron charge. The most fundamental factors which determine $\bar{J}$ are the energy band structure which relates the allowed states for a given material, and the scattering mechanisms which determine τ. This can be seen by recalling that, according to the Bloch theory for an electron in a band, the velocity of an allowed state is given by [1]

$$v(k) = \hbar^{-1} \nabla_k E(\bar{k}) \qquad (1.4)$$

where $E(\bar{k})$ is the energy versus wave vector, or energy band, of the allowed states. Thus the allowed velocities are determined by the allowed energy states.

1.2.1 Electrical Conductivity

For the case of steady state with $\bar{F} = q\,\bar{E}$, where $\bar{E}$ is the electric field intensity, a formal solution of Equation (1.1) and (1.3) can be obtained by assuming the existence of a relaxation time and by neglecting spatial variations in f. This solution is

$$J_i = \sigma_{ij}\, E_j \quad , \; ij = x,y,z$$

where

$$\sigma_{ij} = -\frac{q^2}{4\Pi^3} \int_{\text{all } \bar{v}} \tau\, v_i\, v_j \frac{\partial f_o}{\partial E}\, d\bar{k} \tag{1.5}$$

In this equation summation oven repeated indices is implied. Equation (1.5) illustrates the basic anisotropic nature of electrical conductivity, and also illustrates the importance of the energy band structure in determining this anisotropy. Since one can write [1]

$$\bar{dk} = \frac{ds\, dE}{|\nabla_k E(k)|}$$

then (1.6)

$$\sigma_{ij} = \frac{-q^2}{4\pi^3} \int_{\text{all } E} \frac{\partial f_o}{\partial E}\, dE \int_{s(E)} \frac{\tau\, v_i\, v_j\, ds}{|\nabla_k E(k)|}$$

Here ds is the differential element of surface area on a constant energy surface in k-space, and the first integral is taken over all allowed energies in a given band. Thus for an isotropic band structure the term $|\nabla_k E(k)|$ will be isotropic and the integral over constant energy surfaces will be zero except when i = j.

The theory for conductivity presented so far is valid as long as there are negligible interactions between bands, such as transitions due to external forces or scattering, and as long as the relaxation time approximation is valid [2]. This theory is also invalid for the case where two separate bands are degenerate at one or more points in k-space ([2] , page 170). Interband transitions do occur in some metals and semiconductors. Also, large electric fields can cause interband transitions in semiconductors, and many semiconductors, such as silicon, have degenerate valence bands. Thus, in general one needs information about the band structure of a material in order to interpret conductivity experiments as well as to determine whether a given model will apply to the results. Care must be taken before the results of the Boltzmann theory is applied to a given material. The references listed so far discuss these added effects and they will not be covered here.

The electrical conductivity of most materials is isotropic so that the tensor Equation (1.6) reduces to a scalar quantity. In fact, for any material whose crystalline structure contains at least one axis of symmetry, there are at most only two components of the conductivity tensor [3]. Many metals and semiconductors possess cubic symmetry and thus have an isotropic conductivity. The

tensor aspect of electrical conductivity has been brought up here because of the fact that upon application of certain external stimuli to an isotropic material, the conductivity can become isotropic. For example, this is the case when a uniaxial stress is applied to silicon.

When a solid is isotropic the conductivity can be expressed in the well-known form of

$$\sigma = \frac{q^2 n \bar{\tau}}{m^*} \tag{1.7}$$

where n is the electron concentration, $\bar{\tau}$ is an effective relaxation time, and m^* is an effective mass. For metals the effective relaxation time is ([2] p. 184)

$$\bar{\tau} = \tau(E_F) \tag{1.8}$$

where E_F is the Fermi energy. For non-degenerate semiconductors $\bar{\tau}$ is more complicated and can be expressed as [1]

$$\bar{\tau} = \frac{\int_0^\infty \tau(E)\, E^{3/2} \exp(-E/kT)\, dE}{\int_0^\infty E^{3/2} \exp(-E/kT)\, dE} \tag{1.9}$$

where k is Boltzmann's constant. In a first order theory, this is a basic difference between the conductivities of metals and semiconductors, aside from the fact the n can be several orders of magnitude larger for metals than for semiconductors due to differences in the band structure.

The mobility can be defined as

$$\mu = \frac{q\,\bar{\tau}}{m^*} \tag{1.10}$$

which gives the well-known equation

$$\sigma = q\mu n$$

In any solid at T=0°K which has an empty energy band seperated by an energy gap from a completely filled band one must usually consider another contribution to the electrical conductivity. This is due to the fact that the completely filled band (valence band) will not be completely filled for T>0°K, and thus conduction can occur through this band. This effect is most conveniently formulated in terms of holes which, is simple terms, represent the motion of empty states in the valence band. For this case the electrical conductivity can be expressed as

$$\sigma=q(\mu_n\, n + \mu_p\, p) \tag{1.11}$$

where μ_n, μ_p, n and p are the mobilities and free carrier concentrations of electrons and holes, respectively.

Another case which is important is that of a material with more than one equivalent conduction band minima in momentum space. Such a material can be called a many-valley material. The semiconductor silicon and germanium are classic examples. For these materials the conductivity can be expressed as [4]

$$\sigma=q^2 \sum_{i=1}^{M} n_i \cdot \frac{\bar{\tau}_i}{m_i^*} \tag{1.12}$$

where m_i^* and $\bar{\tau}_i$ are the effective mass and relaxation time of the ith conduction band, respectively, and M is the number of conduction bands.

As has been shown, the conductivity of most materials is determined by the energy level structure and by the scattering mechanisms in the solid. Hence the great interest in these two area of solid state physics.

There has been two essentially equivalent developments of techniques to describe the energy level structure of materials, which are sometimes called the zone theory and the band theory [5]. The band theory evolves from a consideration of the interactions of the electrons with the lattice. The zone theory is most useful for discussing properties of metals while the band theory is used to describe semiconductors and insulators. This is primarily true because the zone theory can only predict whether or not an energy band gap exists, but cannot predict its width. The difference in the energy band structure of metals and semiconductors is due primarily to the existence of an energy gap between a completely filled band and a completely empty band at T=0°K. For metals, a fundamental criterion is that at absolute zero, the electrons reside in unfilled bands. It is thus apparent that as temperature increase the density of conduction electrons is considerably greater for metals than for semiconductors. For metals, the free electron density should be on the order of the density of valence electrons, while for semiconductors this density will be determined at a finite temperature, by those electrons which can be thermally excited across the energy gap to the empty conduction band states. The theory for conductivity must take these basic facts into account in evaluating the free electron density [3].

Table 1 shows a comparison of experimental and theoretical resistivities for several metals at 273°K. The agreement is not bad for most monovalent metals, but for the transistion metals there is almost as much as two orders of magnitude difference between theory and experiment. The source of this disagreement is largely unknown. It has been attributed to such things as incomplete knowledge of the band structure and to inaccurate knowledge of the Debye characteristics temperature ([1] p. 195) which greatly influences the mobility in metals.

Table 1 Ideal Resistivites at 273°K of a Number of Monovalent, Divalent, Trivalent, and Transition Metals

Element	Valence	Resistivity at 273°K, μohm-cm	ρ_{exp}/ρ_{theor}
Li	1	8.5	1.40
Na	1	4.27	0.77
K	1	6.3	0.80
Rb	1	11.6	1.31
Cs	1	19.0	1.50
Cu	1	1.55	1.87
Ag	1	1.50	1.16
Au	1	2.04	3.26
Be	2	2.78	7.33
Cd	2	6.73	8.05
Mg	2	3.94	3.23
Zn	2	5.45	5.75
Al	3	2.50	3.62
In	3	8.2	8.2
Co	Trans	5.2	14.7
Cr	Trans	15.0	34.2
Fe	Trans	8.71	26.0

Table 1, Continued

Mo	Trans	5.03	25.2
Ni	Trans	6.58	9.9
Pd	Trans	9.77	13.0
Pt	Trans	9.81	17.2
Ta	Trans	12.4	36.9
Ti	Trans	42	81.0
V	Trans	18.2	48.6
W	Trans	4.89	35.2

In semiconductors, the theory for electrical conductivity is in a much more satisfactory state. For many of these materials detailed band calculation have been made, and scattering theory is in a well-developed state. Independent calculations for both n and μ can be made and the theoretical values agree well with experiment. For example, Figure 1.1 shows resistivity, mobility, and Hall co-efficient (assumed proportional to p^{-1}) for p-type germanium. Theories exist to explain these data and they will be discussed in later sections.

1.2.2 Scattering Mechanisms Which Influence Conductivity

There are several mechanisms which result in the scattering of an electron in a crystal, and which limit its mean free path to finite values. Generally, the scattering mechanisms fall into the category of crystal imperfections, and can be classified as stationary imperfections (defects) and lattice vibrations. Lattice

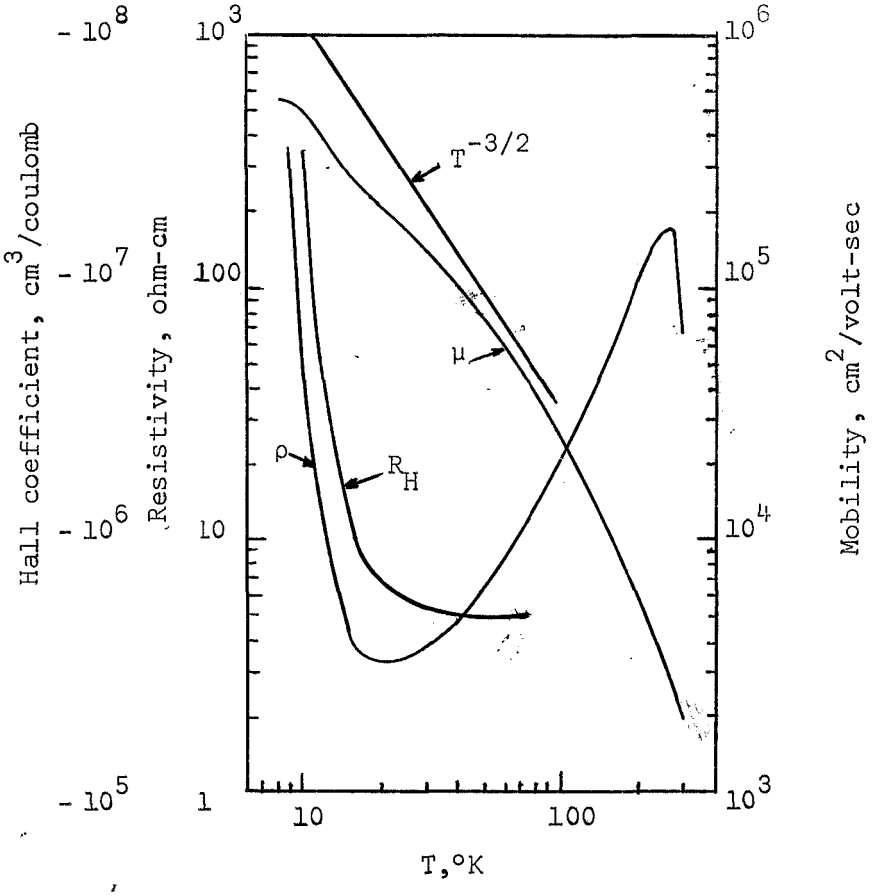


Figure 1.1. Experimental Hall, resistivity, and mobility data for a sample of p-type germanium. [from D. M. Brown and R. Bray, Phys. Rev., 127: 1597 (1962).]

vibrations are a result of the pertubation of the perfect lattice arising from the thermal motion of the lattice atoms. Imperfections in the lattice such as immobile impurities and other point defects, as well as line defects such as dislocations, can also be important factors in determining the relaxation time. The scattering of charge carriers in metals and semiconductors is basically the same process. The difference between these materials in terms of their electronic properties is largely due to the drastic difference in their carrier concentrations. This difference does result in a distinct behavior for electrons in these two types of solids in terms of their

reaction to external influences. Also, a basic calculational difference results in the manner in which formal integration of the transport Equation (1.3) is carried out. This is due to the fact that for metals the Fermi energy lies in the conduction band states, while for semiconductors it lies in the energy gap several kT in energy away from the conduction band edge. This results in different approximations in evaluating the transport integrals ([2], p. 180).

The scattering of conduction electrons by lattice vibrations (electron-phonon scattering) in metals may be of two types, which are known as normal processes or Umklapp processes. These processes are distinguishable through the momentum conservation law [6]

$$\bar{k} + \bar{q} + \bar{g} = \bar{k}' \qquad (1.13)$$

where k is the initial electron wave vector, $\bar{k}'$ is the final electron wave vector, $\bar{q}$ is the phonon wave vector, and $\bar{g}$ is a reciprocal lattice vector. For normal processes $\bar{g} = 0$ while it is non-zero for Umklapp processes. The normal processes lead to small angle scattering events while Umklapp processes lead to large angle scattering events, and this second type of scattering event is important in metals at all temperatures above about 50°K [2]. The Umklapp processes make optical phonon and short-wavelength acoustic phonon scattering most important in metals.

In semiconductors and insulators, Umklapp process are generally not important. However, optical phonon scattering can be important

because of the fact that many of these materials have two or more atoms per unit cell, and thus the phonon spectrum contains at least three optical branches. Another complication arises because of the many-valley nature of some semiconductors which lead to another form of lattice scattering called inter-valley scattering. This type of scattering is very important in determining the mobilities of electrons in silicon and germanium. In fact, the many-valley scattering mechanism is the major factor in the large piezoresistance effects which occur in silicon and germanium. This has led to increased usage of silicon as a strain sensor in many areas of stress analysis and pressure measurements. This topic will be discussed in a latter section.

It is obvious that lattice scattering is of basic importance in determining the electrical properties of materials and is thus of basic importance in any electrical phenomena which might have sensor application. However, a theoretical treatment is beyond the scope of this review, and such treatments are available to the reader [1] - [8]. The results will only be stated here. It should be noted that these results can be assumed characteristic of lattice scattering in the particular materials for which calculations are quoted, and any deviation from these results can be assumed to arise from other effects which will be quoted later.

For metals, the theory of resistivity due to lattice scattering is based on the Bloch-Gruneisen integral [6], which has found application over a suprisingly wide temperature range. This equation yields the results

$$\rho(T) = \begin{cases} AT & T >> \Theta_R \\ BT^5 & T << \Theta_R \end{cases} \qquad (1.14)$$

where Θ_R is a characteristic temperature for resistivity which is quite similar to the Debye characteristic temperature in a solids' lattice specific heat [6]. The Bloch-Gruneisen equation is a semi-emperical equation which is a valuable tool for analyzing experimental data. Its success is very suprising since, among other things, it neglects completely Umklapp scattering. Much of the success has been due to the adjustable parameter Θ_R and the comparison between theory and experiment is basically a curve fitting scheme. Another factor is that the agreement is superficial at high temperatures since most metals do vary linearly with temperature in this range. Only at low temperatures, were significant deviations from the T^5 behavior are generally found, is the Bloch-Gruneisen theory put to the test. The characteristic temperature, Θ_R, is generally not a very low temperature, varying from about 37°K for cesium to as high as 1240°K for Be [6]. For most metals it lies in the range of about 200-400°K. Figure 1.2 illustrates the resistivity versus temperature for several metals over a wide temperature range. The discontinuites in some of the curves occur at the melting point of the metal.

For semiconductors with spherical energy surfaces the mobility due to lattice scattering varies with temperature as

$$\mu_L = CT^{-3/2} \qquad (1.15)$$

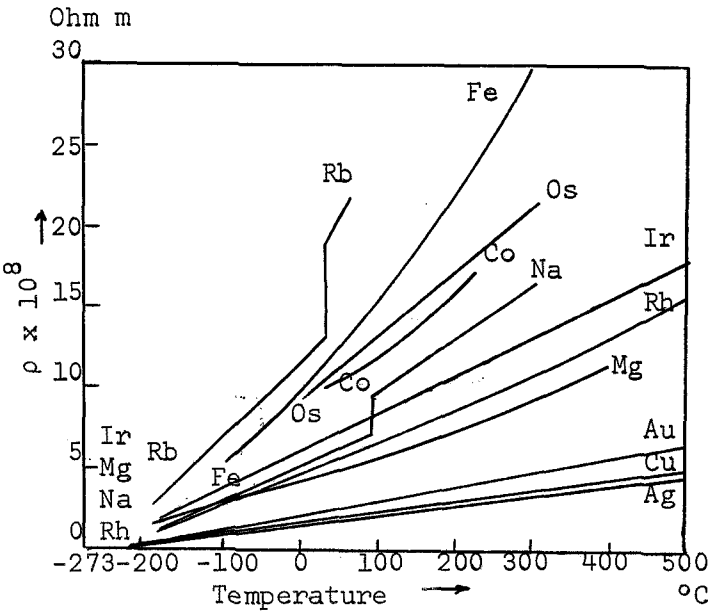


Figure 1.2. A plot of the electrical resistivity of several metals illustrating the proprotionality to absolute temperature.

For semiconductors with many-valley energy band structures the mobility generally varies with temperature to a power less than -1.5 due to intervalley scattering [1]. Figure 1.3 shows a plot of mobility versus temperature for a many-valley semiconductor. The deviation form a $T^{-3/2}$ law can be clearly seen.

It should be mentioned that the temperature dependence of resistivity has been given here for metals while only mobility versus temperature has been given for semiconductors. This is because of the fact that in semiconductors the free carrier concentration is almost solely determined by the impurity content of the material. This aspect will be discussed in the next section.

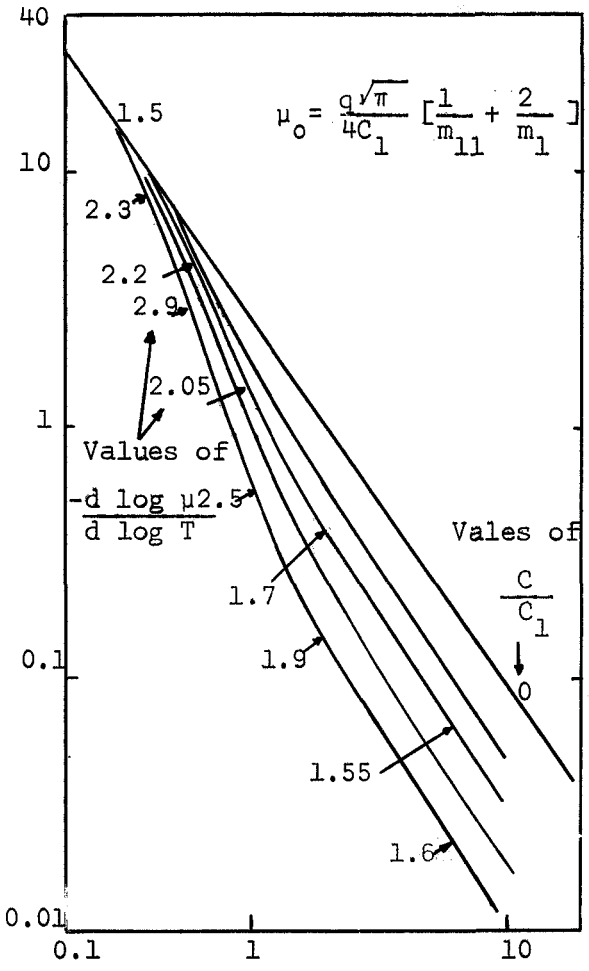


Figure 1.3. Mobility-temperature curves for pure lattice scattering, neglecting scattering by optical modes. The quantities C_l and C measure the strength of the coupling of the carriers to intra- and intervalley modes, respectively; ω^l is the frequency of the intervalley mode.

1.2.3 Impurity Effects on Conductivity

Impurities can have a large influence on the electrical properties of solids and on the behavior of bulk solid-state sensors as well. This is especially true for semiconductors. These impurities are generally in the form of chemical defects such as vacancies, interstitials, and foreign atoms.

For metals, chemical impurities contribute to the conductivity through what is called the residual resistance. This term results from the experimental observation that as temperature is lowered to near 0°K, the resistivity of metals does not approach zero according to Equation (2.14) but instead approaches some constant value as shown in Figure 1.4 for sodium. This constant value is called the residual resistivity, and its presence leads to what is known as Matthiesson's rule for metals. This rule states that

$$\rho_T = \rho_o + \rho_L \tag{1.16}$$

where ρ_T is the total resistivity, ρ_L is the lattice resistivity, and ρ_o the residual resistivity. That ρ_o is in fact due to impurities can be seen from Figure 1.5 which shows resistivity versus temperature for copper containing varying small atomic per centages of nickel. The number on the curves indicate the atomic per cent of nickel. This illustrates that the residual resistance is an impurity effect in metals.

Another important part of the behavior of the conductivity of metals is the influence of alloying. If an impurity is added in sufficient quantities, then alloys can be formed between the impurity

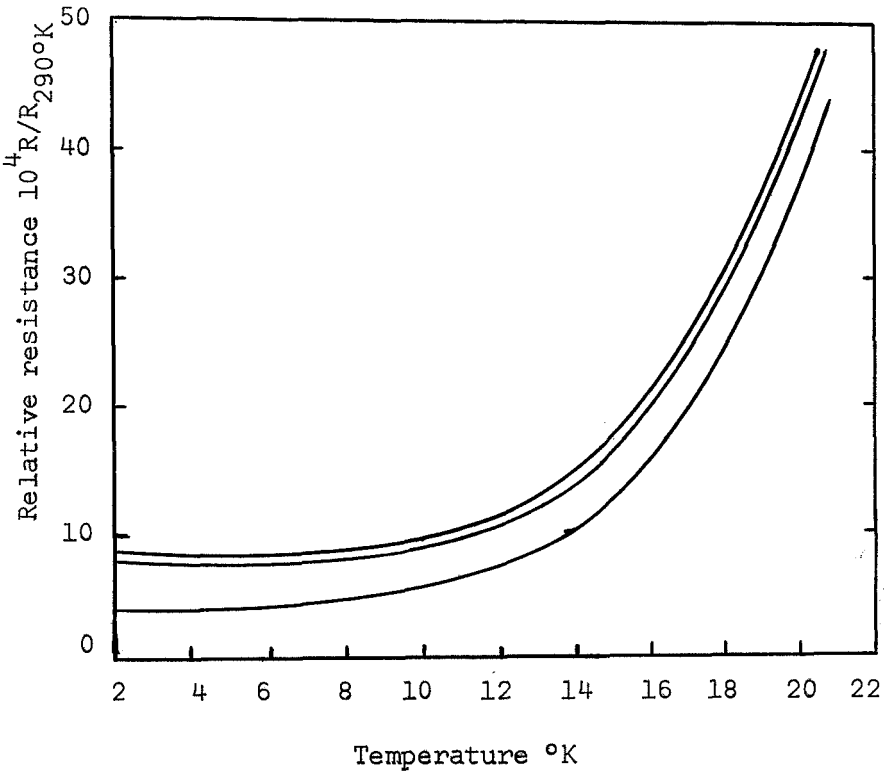


Figure 1.4. The resistance of three samples of sodium.

and host crystal atoms which can result in new crystal structures. Figure 1.6 illustrates this effect for copper-gold alloys. As the amount of gold in the copper is changed the value of the resistivity depends on whether a new crystal structure is formed. In this figure the dips in resistivity are due to the formation of Cu_3 Au and Cu Au alloys formed by proper high-temperature annealing of gold-doped copper.

In semiconductors very minute amounts of impurities on the order of a few parts per million can completely determine the free carrier concentration. In a pure semiconductor with a band gap greater than a few tenths of an electron volt the free carrier concentration due to thermal excitation (the intrinsic carrier concentration) is usually several orders of magnitude less than one free carrier for each host atom, so that electrically active impurities with concentrations

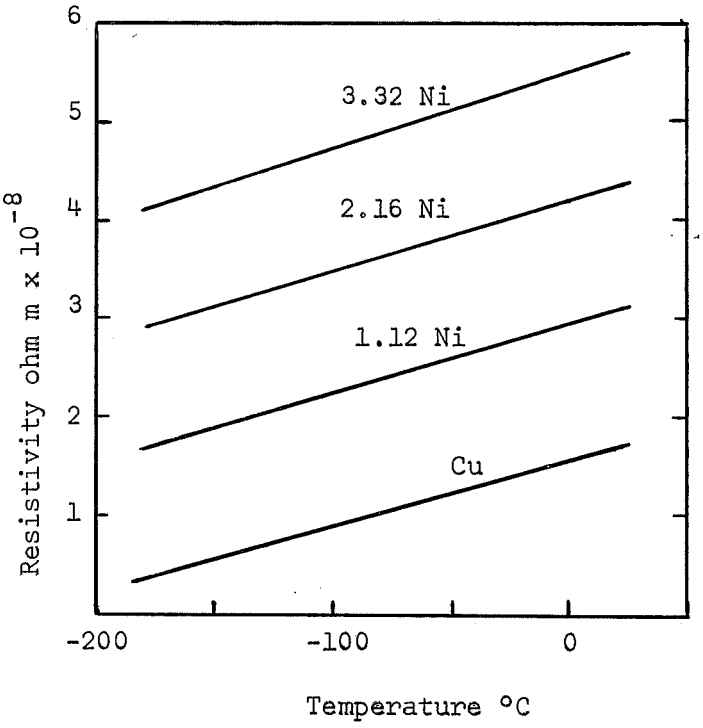


Figure 1.5. The resistivity of copper-nickel alloys showing the existence of the same temperature-dependent resistivity superimposed on varying amounts of residual resistance.

around a few parts per million will be the determining factor for the carrier concentration the material is called extrinsic. The intrinsic carrier concentration varies as [7]

$$n_i = \frac{1}{\sqrt{2}} \left(\frac{kT}{\pi h^2}\right)^{3/2} (m_c\, m_v)^{3/4} \exp\left(\frac{-E_g}{2kT}\right) \qquad (1.17)$$

where m_c and m_v are the effective masses in the conduction and valence bands, respectively, h is Planck's constant, and E_g is the energy of the forbidden gap. Table 2 lists some of the important parameters which determine n_i, for several semiconductors. Also shown are experimental values for the electron and hole mobilities as a function of temperature, and the deviation from a $T^{-3/2}$

Table 2 Electrical Properties and Band Parameters of Some Common Semiconductors.

Semiconductor	Energy gap, ev		Mobility, cm^2/volt-sec ($T = 300°K$)		Mobility variation $\mu \propto T^{-n}$		Effective mass m/m_o		Dielectric constant ϵ_r
	300°K	0°K	Electrons	Holes	Electrons	Holes	Electrons	Holes	
Ge	0.67	0.75	3,950 (3,900)	3,400 (1,900)	1.66	2.33	m = 1.6 m = 0.082	0.3,0.04	16
Si	1.106	1.153	1,900 (1,350)	425 (480)	2.5	2.7	m = 0.97 m = 0.19	0.5,0.16	12
InSb	0.16	0.26	78,000	750	1.6	2.1	0.013	0.6	17
InAs	0.33	0.46	33,000	460	1.2	2.3	0.02	0.41	14.5
InP	1.29	1.34	4,600	150	2.0	2.4	0.07	0.4	14
GaSb	0.67	0.80	4,000	1,400	2.0	0.9	0.047	0.5	15
GaAs	1.39	1.58	8,500	400	1.0	2.1	0.072	0.5	12.5
GaP	2.24	2.40	110	75	1.5	1.5	.....	.....	10

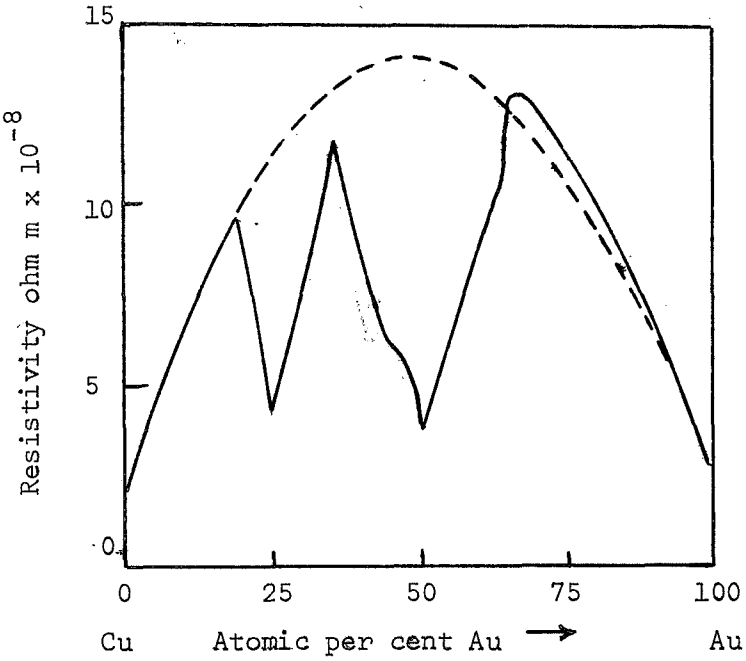


Figure 1.6. The resistivity of copper-gold alloys. The full line refers to annealed samples and the dashed line to quenched samples.

variation should be noted. Table 3 gives a more complete description of the band structure, band paramters, and bad gap energy of many elemental and compound semiconductors [9]. Finally, Figure 1.7 shows the intrinsic carrier concentration for silicon and germanium as a function of temperature. It is quite obvious that the intrinsic carrier concentration will control the temperature behaviro of the resisitivity of an intrinsic semiconductor, and that the resistivity will be a very strong function of temperature.

The expected temperature dependence of the carrier concentration of a doped semiconductor is depicted in Figure 1.8, and is compared with the intrinsic carrier concnetration. It can be seen that at high temperatures the material approaches intrinsic conduction. The most striking feature is that over a large range of temperatures the carrier concentration is constant. This is called the extrinsic or exhaustion region,

Table 3 Energy Band Parameters of Semiconductors and closely related Semimetals

		Valence band maximum			Conduction band minimum			
Material	Crystal structure	State[a]	Energy surfaces	Effective mass parameters[b]	State	Energy surfaces	Effective mass parameters[b]	Energy gap at ∿ 0°K (in eV)
Si	Cubic (Diamond)	Γ_8 + 2	warped spheres	0.50 }hh 0.15 }hh	Δ_1	6 spheroids	0.9163 (0.9716)} 0.1905 (0.1945)}	1.17
Ge	Cubic (Diamond)	Γ_8 + 2	warped spheres	0.32 }hh 0.043}hh	L_1	4 spheroids	1.588 (1.349) } 0.08152 (0.0791)}	0.744
α-Sn	Cubic (Diamond)	Γ_8 + Single	warped spheres	———	Γ_8 +	Warped sphere[c]	0.02	0 (semimetal)
InSb	Cubic (Zinc blende)	Γ_8	2 warped spheres	0.4 0.016 }hh (0.016)}lh	Γ_6	Single spherical	0.0155 (0.0149)	0.235
InAs	Cubic (Zinc blende)	Γ_8	2 warped spheres	0.41 }hh 0.026 }lh (0.029)	Γ_6	Single spherical	0.024 (0.025)	0.42
GaSb	Cubic (Zinc blende)	Γ_8	2 warped spheres	0.052 }lh (0.053)	Γ_6	Single spherical	0.042 (0.046)	0.81
InP	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	Γ_6	Single spherical	0.077 (0.075)	1.42
GaAs	Cubic (Zinc blende)	Γ_8	2 warped spheres	0.68 }hh 0.12 }lh (0.10)	Γ_6	Single spherical	0.07(77°K) (0.083)	1.52
AlSb	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	X_1 or Δ_1	3 or 6 spheroids	0.39	1.60
GaP	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	X_1 or Δ_1	3 or 6 spheroids	0.35	2.34

(continued)

Table 3, Continued

CdTe	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	Γ_6	Single spherical	0.0096 (0.11)	1.605
ZnTe	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	Γ_6	Single spherical	———	2.39
ZnSe	Cubic (Zinc blende)	Γ_8	2 warped spheres	———	Γ_6	Single spherical	0.17 (77°K) (0.21)	2.80 (80°K)
HgTe	Cubic (Zinc blende)	Γ_8	Single warped sphere	———	Γ_8	Warped sphere[c]	0.029 (0.025)	0 (semimetal)
CdS	Hexagonal (Wurtzite)	Γ_9	Single spheroidal	———	Γ_7	Single spheroidal	0.2 (0.18)	2.58
CdSe	Hexagonal (Wurtzite)	Γ_9	Single spheroidal	———	Γ_7	Single spheroidal	0.13 (0.13)	1.84
PbS	Cubic (Rock salt)	L_6 +	4 spheroids	0.105} 0.075}	L_6-	4 spheroidal	0.105} 0.080}	0.29
PbSe	Cubic (Rock salt)	L_6 +	4 spheroids	0.068} 0.034}	L_6-	4 spheroids	0.07 } 0.040}	0.15
PbTe	Cubic (Rock salt)	L_6 +	4 spheroids	0.031} 0.022}	L_6-	4 spheroids	0.24 } 0.024}	0.19
SnTe	Cubic (Rock salt)	L_6 -	4 spheroids	———	L_6+	4 spheroids	———	0.3

[a] Double group notation. Detailed structure near band edge of crystals not having inversion symmetry is ignored here in specifying band extrema.

[b] Experimental values are given first, and k.p derived theoretical values are given in parentheses immediately below for some of the materials. All values are in units of the free electron mass, and are for ∿0°K unless indicated otherwise. Spherical average values of anisotropic masses are given unless indicated otherwise. The hh and lh represent heavy and light hole masses, respectively, and the and symbols represent effective mass values parallel and perpendicular, respectively, to the symmetry axis of a constant energy spheroid.

[c] The conduction band in these materials is generically related to the normal light hole valence band, so that the surfaces of constant energy should be at least slightly warped; they appear spherical, however, from results obtained so far.

and for this range of temperatures the thermal variation in resistivity of a semiconductor is determined by its mobility. The extrinsic region is determined by the density of impurity atoms and their ionization energy. Generally, a semiconductor will remain extrinsic over a range of temperature for which the Fermi energy is at least 3kT from the impurity energy level in the forbidden gap. Table 4 shows the ionization energies for several defects in the forbidden gap of silicon and germanium and Figure 1.9 shows the position of the Fermi energy with temperature for

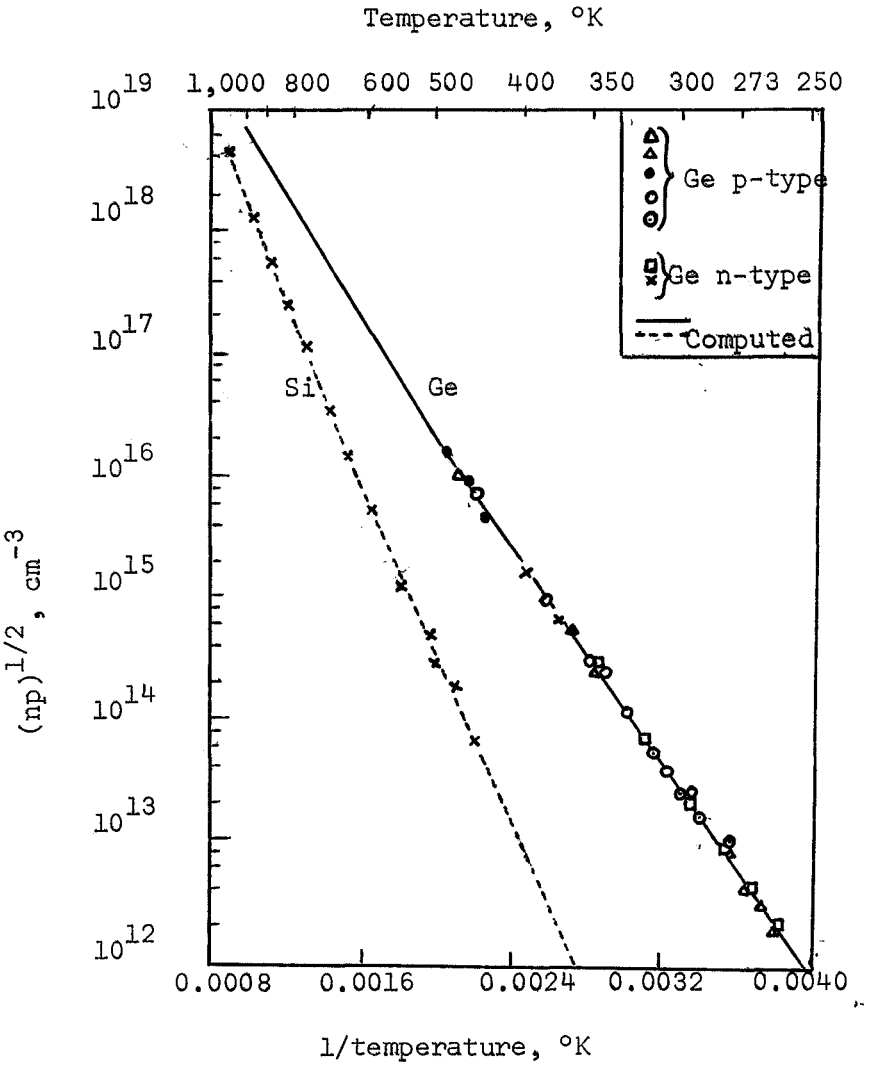


Figure 1.7. Intrinsic carrier densities n_i in silicon and germanium, obtained as best fits to experimental data.

Table 4 Ionization Energies of Donors and Acceptors in Si and Ge,ev

Impurity	Germanium	Silicon
Donors		
Li		0.033
P	0.0120	0.044
As	0.0127	0.049
Sb	0.096	0.039
Bi		0.069
Acceptors		
B	0.0104	0.045
Al	0.0102	0.057
Ga	0.0108	0.065
In	0.0112	0.16

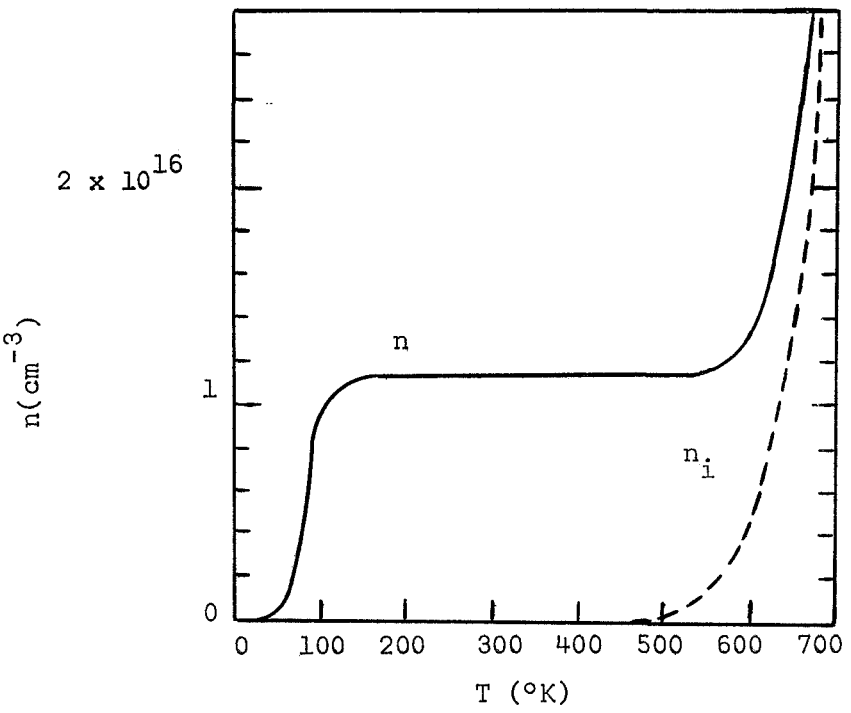
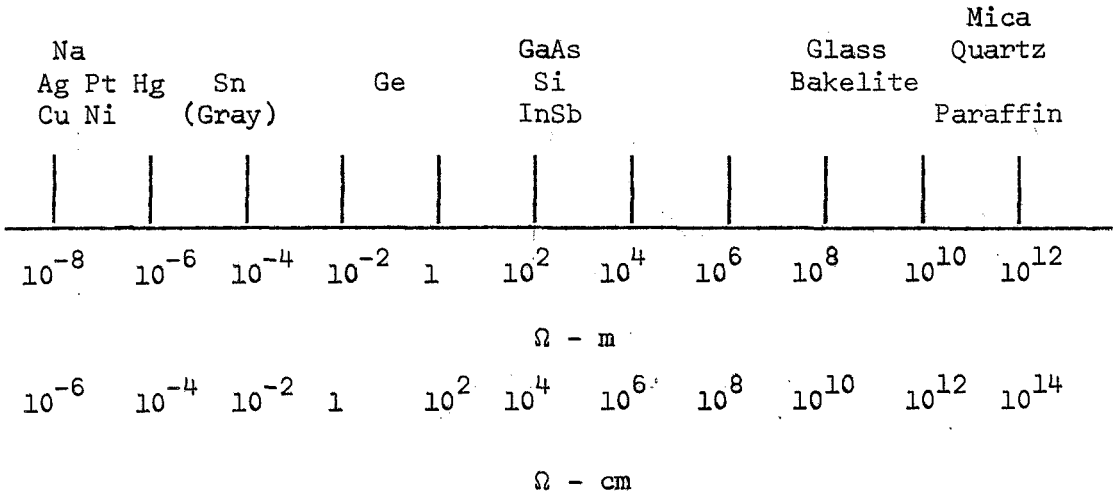


Figure 1.8. Electron concentration in n-type silicon as a function of temperature .

various impurity concentrations in silicon. It can be seen that silicon is extrinsic over a quite wide temperature range, except for near degenerate doping levels.

In summary, the conductivity is probably the most important material property used in sensor development. Its magnitude and temperature dependence and the components of which it is composed are of basic importance. The range of resistivities which are attainable and these values cover the range from about 10^{-6} to 10^{4} ohm cm as depicted in Table 5. The twenty orders of magnitude range of resisitivty values illustrates a basic fact of material properties: the choice of values of any material parameter is usually limited only by the number and types of materials which are available, and which have properties which are desirable to accomplish a

Table 5 The range of resistivity observed at room temperature in different materials.



given task.

1.2.4. The Transition Metal Compounds

A particular class of materials which have very interesting electrical properties,and which have received a great deal of attention in the literatue during he past several years, are some of the transition metal oxides and sulfides exhibit what is commonly called a semiconductor-to-metal transition at some finite temperature. For example, Figure 1.10 shows conductivity versus reciprocal temperature for a few of the transition metal oxides. With the exception of TiO, these materials have a semiconducting type of behavior at lower temperatures, and then undergo an

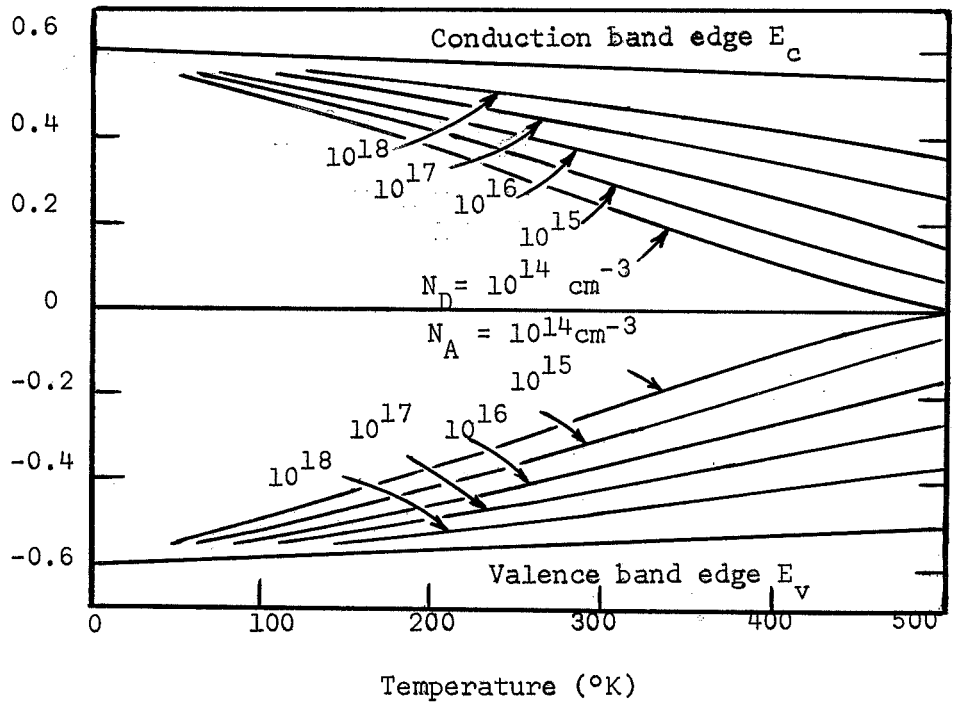


Figure 1.9. The Fermi level in silicon as a function of temperature, for various impurity concentrations.

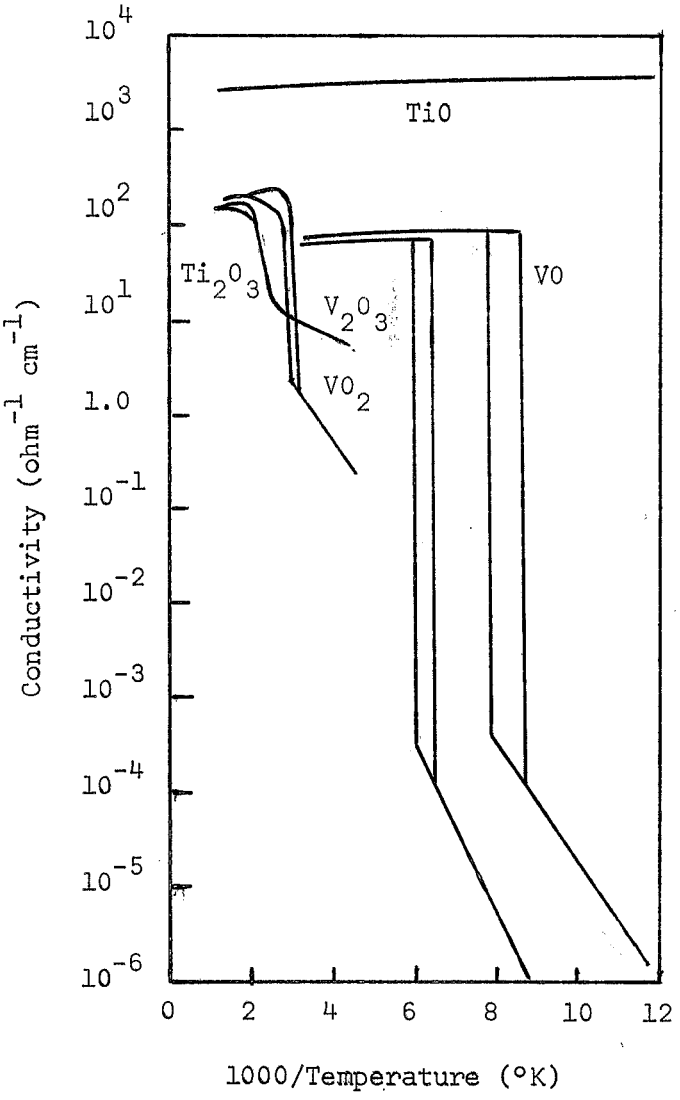


Figure 1.10. Conductivities of oxides (Morin [36]).

abrupt transition to a metallic form of conduction at a given temperature, which is called the transition temperature. In fact, at the lower temperatures some of these materials behave as very good insulators. The conductivity change at the transition temperature is quite large, being as much as five orders of magnitude for the two oxides of vanadium, VO and V_2O_3. Transition metal compounds which have shown the semiconductor-to-metal transition include V_2O_3, VO_2, VO, V_6O_{13}, Ti_2O_3,

NbO_2, Fe_2O_4, V_3O_5, V_4O_7, Ti_3O_5, NiS, CrS, and FeS. Recently, mixed oxides of vanadium and chromium, $(V_{1-x}Cr_x)_2O_3$, have shown the transition for values of $x \geq 0.009$ [10]. Figure 1.11 shows resistivity versus reciprocal temperature for a mixed oxide with x-0.04. Also shown in this figure is the pressure dependence of resistivity at room temperature, as well as the effect of pressure on the semiconductor-to-metal transition. It appears that pressure drastically affects the transition temperature as well as the temperature dependence of resistivity of these oxides. It has also been observed that the resisitivity of V_2O_3 in the metallic state is very pressure dependent, and that the transition can be induced by applying pressure to this material [10]. Most of the efforts of researchers interested in materials which exhibit the semiconductor-to-metal transition have been directed toward understanding the physics of the transition. However, it does appear that this class of materials could have definite sensor applications, the most obvious being in the area of temperature and mechanical sensors.

There have been several attempts at explaining the semiconductor-to-metal transition. The accepted model for the conductivity change in most of the transition metal compounds involves crystalline symmetry changes and/or antiferromagnetism [11]. This effect will be discussed in a later section. In this model, the semiconductor energy gap at low temperatures results from the splitting of the first Brillouin zone by either an antiferromagnetic exchange interaction or by a crystalline-structure distortion to a state of lower symmetry. This model predicts a carrier concentration dependent energy gap of the form

$$E_g = E_{go} - \beta n \qquad (1.18)$$

where β is a constant, E_g is the energy gap, and E_{go} the energy gap for $n \to 0$. At low temperature n is small and the material behaves as a semiconductor with an energy gap of E_{go}. As temperature increases, n increases and E_g decreases. Since, for a intrinsic semiconductor, n is exponentially related to E_g, then at some temperature there is a very rapid decrease in E_g with a corresponding increase in n. The energy gap approaches zero and an metallic conduction occurs. This model has been found to account for the major features of the semiconductor-to-metal transition.

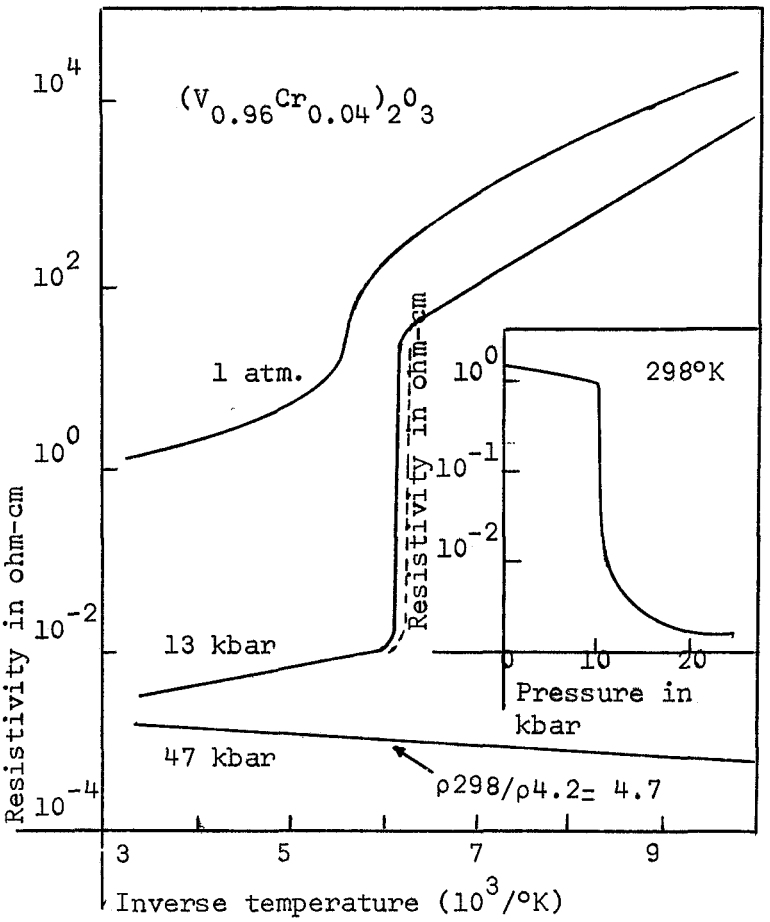


Figure 1.11. Resistivity vs reciprocal temperature at different pressures for $(V_{0.96}Cr_{0.04})_2O_3$. Inset shows I-M transition at room temperature as a function of pressure. Below 10 kbar the AF → I transition occurs on warming. Above 10 kbar the AF → M transition occurs. Above 40 kbar the sample is metallic from 4.2°K to 298°K.

1.3 DIELECTRIC PROPERTIES OF SOLIDS

The dielectric properties of solids is an area of physics which deals with the influence of external electric fields on the internal microscopic charge distributions present in a material media. This generally does not include processes which produce a net charge transport, although transport of charge can influence the dielectric properties of solids.

The charge distributions in a material media are a result of its atomic structure; yet any observation of the influence of an externally applied electric field on these distributions must be made macroscopically. As in the theory of conductivity the problem here is to relate the microscopic to the macroscopic properties. Mathematically this is done through Maxwell's equations of electromagnetics which relate electromagnetic fields to their sources. Since in Maxwell's equations charge is the source of an electric field, and since Maxwell's equations are macroscopic in nature, this formulation provides the appropriate link between macroscopic observables and microscopic sources.

The most common macroscopic manifestation of the effect of an electric field upon a material media is through the capacitance of an ideal parallel plate capacitor. For this type of geometry, the capacitance is proportional to the dielectric constant of the medium from which the capacitor is made. The dielectric constant is defined by the relation

$$\varepsilon_r = \varepsilon/\varepsilon_0 \tag{1.19}$$

where ε is the permittivity of the medium, ε_o is the permittivity of free space, and ε_r is the dielectric constant of the medium. Thus the presence of a material medium in a parallel plate capacitor manifests itself by changing the value of the capacitance, which is a measurable macroscopic quantity.

The dielectric constant is different from unity in a material medium due to the phenomenon of polarization. This phenomena is the induction of electric dipole moments, or the orientation of existing dipoles, by the applied field. The dipole moment is the result of a separation of positive and negative charges in the material by the applied field. Some materials have permanent dipole moments which exist in the absence of an applied field. These materials, which are called ferroelectrics, will be treated in a later section. Quite often materials which do contain permanent dipole moments are called polar materials, while those that are polarizable only in an electric field are called non-polar.

The result of a polarization in a material leads to the following equation relating electric flux density (displacement vector) to the total electric field intensity and the net polarization per unit volume

$$\overline{D} = \varepsilon_o \overline{E} + \overline{P} \tag{1.20}$$

where $\overline{D}$ is the electric flux density and $\overline{P}$ is the polarization per unit volume. Since $\overline{P}$ is produced by $\overline{E}$ it must be mathematically related to $\overline{E}$. As an example, consider the case where $\overline{P}$ is proportional to $\overline{E}$, and in the same direction as $\overline{E}$. Such a material is said to be linear and isotropic, and is often called a class A dielectric. If

one writes

$$\overline{P} = \varepsilon_o \chi \overline{E}$$

then

$$\overline{D} = \varepsilon_o (1 + \chi) \overline{E} \qquad (1.21)$$

$$= \varepsilon_o \varepsilon_r \overline{E}$$

$$= \varepsilon \overline{E}$$

For this case, χ is called the electric susceptibility and is dimensionalless. The quantity $(1 + \chi)$ is the relative dielectric constant, ε_r, and the field Equation (1.20) becomes the same as that for free space, except for the fact that the permittivity of free space is replaced by the permittivity of the dielectric medium.

Unfortunately, not all dielectrics are isotropic and the relationship between $\overline{D}$ and $\overline{E}$ is a tensor relation [12]. In tensor notation this relation is

$$D_i = \varepsilon_{ij} E_j \qquad i=x,y,z \qquad (1.22)$$

where summation over repeated indices is implied. For the majority of crystals the tensor is diagonal so that the relation is simplified somewhat, and for crystals with cubic symmetry a linear relation holds, and such materials are isotropic.

1.3.1 Polarization of Dielectrics

An applied electric field polarizes a dielectric. The dipole

moments so produced have an electric field associated with them. In solids and liquids, the electric field intensity due to any dipole moments must be included, along with the applied electric field, in a calculation of the total electric field intensity at a given point. This is not true for gases due to the much larger atomic spacings as compared to those for solids and liquids. In solids and liquids the total electric field acting on a dipole is not simply the applied field, a complication which makes the theory much more difficult. The total field at a given point is called the local field and the polarizability of a dipole is given by

$$\alpha = P/E_{local} \tag{1.23}$$

where α is the polarizability ($f.m^2$), E_{local} is the magnitude of the local field, and p is the magnitude of the dipole moment given by p= P.V., where V is the volume of the crystal. The manner in which the local field is calculated and related to the average or macroscopically measured field is of fundamental importance. A good general treatment of these calculations and a discussion of the assumptions necessary to make these calculations is given in Kittel [13].

The local field at an atomic position in the interior of a dielectric with a general geometry can be expressed as

$$\overline{E}_{local} = \overline{E}_o + \overline{E}_{Dipole} \tag{1.24}$$

where $\overline{E}_o$ is the external field and $\overline{E}_{Dipole}$ is the field due to the dipole moment of the other atoms in the crystal. In general

$$\overline{E}_{Dipole} = \frac{1}{4\pi\epsilon_o} \sum_i \frac{3(\overline{p}_i \cdot \overline{r}_i) - r_i^2 \overline{p}_i}{r_i^5} \tag{1.25}$$

where the sum is over all atoms [13]. The evaluation of this term appears to be rather formidable task. However, it has been found expedient to decompose the dipole field into three terms which allows the summation

to be replaced by integrations over a surface enclosing a spherical cavity. The process is readily carried out for crystals having cubic symmetry, as well as for isotropic crystals, and the result is [13]

$$\overline{E}_{local} = \overline{E} + \frac{\overline{P}}{3\varepsilon_o} \tag{1.26}$$

where now $\overline{E}$ is the macroscopic field used in Maxwell's equations. Now, the polarization per unit volume can be expresses as

$$\overline{P} = \sum_i N_i \alpha_i \overline{E}_{local} \tag{1.27}$$

where N_i is the concentration of atoms of polarizability α_i, and the sum allows the inclusion of different mechanisms for polarization. Using Equations (1.26) and (1.21) gives

$$3\varepsilon_o \cdot \frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \sum_i N_i \alpha_i \tag{1.28}$$

This equation is known as the Clausius-Mossotti relation which relates the macroscopic parameters ε_o and ε_r to the microscopic polarizability of a crystal.

1.3.2 Sources of Polarization in Crystals

There are three possible distinct causes of polarization in solids. These can be simply states as [13]:

a) Electronic polarizability - Due to the fact that atoms consist of positively charged nuclei surrounded by negatively charged electrons, an applied electric field will tend to change the spherical symmetry of such an arrangement producing an electric dipole moment. Every atom in a crystal should have an electronic polarizability.

b) Ionic polarizability - If the atomic bonding in a crystal is of an ionic nature, then an applied electric field can polarize the ions by displacing them from their equilibrium positions.

c) Orientational polarizability - Certain molecules have a permanent dipole moment due to their internal structure. In the solid-state the orientation of these dipoles is random which, in the absence of an applied field, results is no net dipole moment for the solid. An external field tends to orient these dipoles however, giving rise to an effective polarizability.

In a solid all three of these mechanisms can contribute to the polarizability, and thus to the dielectric constant for fields with low frequencies. Up to this point, nothing has been said about the time dependence of the electric field, but it is recognized that alternating as well as steady fields can induce polarization. As the frequency of the applied field is increased only those polarization mechanisms which can respond to the time behavior of the field can contribute to the dielectric constant. Thus, at high frequencies, the dielectric constant is expected to be complex. The real and imaginary parts of the dielectric constant are connected by the Kramers - Kronig dispersion relations [12]. The imaginary part of the dielectric constant represents the energy loss mechanisms which are involved with the polarization process. Figure 1.12 illustrates the expected behavior of the real and imaginary parts of the dielectric constant as a function of frequency for an ionic crystal containing permanent dipoles. The variation with frequency for the different polarization processes occurs in specific frequency and physical models have been developed to explain this behavior [12] , [13] , [14]. Generally, the frequency dependence of the dielectric constant due to orientational polarizability occurs in the microwave region ($\sim 10^{10}$ Hz), that due to ionic polarization occurs in the infrared region ($\sim 10^{13}$ Hz), and that due to electronic polarization occurs in the ultraviolet region ($\sim 10^{15}$ Hz). For frequencies much larger than 10^{15} Hz the dielectric constant is very close to unity. Table 6 gives values of the dielectric constant of several solids for low frequencies (ε_r) and for frequencies in the visible region (ε_∞). Table 7 shows a comparison of the low frequency dielectric constant

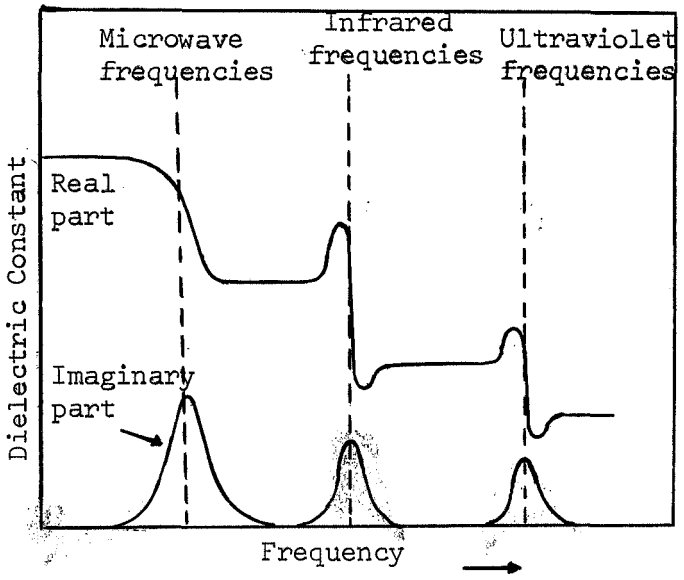


Figure 1.12. The dependence upon applied frequency of the real and imaginary parts of the complex dielectric constant in an ionic solid containing permanent dipoles. The imaginary part of the dielectric constant (or dielectric loss factor) is large whenever the frequency reaches the upper limit for a dielectric response mechanism. This happens for dipole orientation in the microwave range, for ionic polarizability in the infrared, and for bound electron displacements in the ultraviolet.

and loss tangent for several different types of solids which have practical applications. The loss targent is defined as

$$\text{Tan } \delta = \frac{\sigma}{\omega \varepsilon} \qquad (1.29)$$

where σ is the conductivity of the dielectric, which for an ideal dielectric should be zero. Note that the class of ferroelectric materials seem to have usually large dielectric contstant. These materials will be discussed in a later section, and it can be noted that these materials do have useful sensor applications. In fact, dielectric properties have been discussed in order that ferroelectric behavior can be introduced at an appropriate level.

Table 6 Experimental values of high and low frequency dielectric constants for ionic crystals.

Material	Lattice type	ε_∞	ε_r	Material	Lattice type	ε_∞	ε_r
LiF	NaCl	1.92	9.27	MgO	NaCl	2.95	9.8
LiCl	"	2.75	11.05	CaO	"	3.28	11.8
LiBr	"	3.16	12.1	SrO	"	3.31	13.3
LiI	"	3.80	11.03	BaO	"		34.
NaF	"	1.74	6.0	CsCl	CsCl	2.60	7.20
NaCl	"	2.25	5.62	CsBr	"	2.78	6.51
NaBr	"	2.62	5.99	CsI	"	3.03	5.65
NaI	"	2.91	6.60	TICI	"	5.10	31.9
KF	"	1.85	6.05	TIBr	"	5.41	29.8
KCI	"	2.13	4.68	CuCl	ZnS	3.57	10.0
KBr	"	2.33	4.78	CuBr	"	4.08	8.0
KI	"	2.69	4.94	ZnS	"	5.07	8.3
RbF	"	1.93	5.91	BeO	ZnO	2.95	7.35
RbCl	"	2.19	5.0	CaF_2	CaF_2	1.99	8.43
RbBr	"	2.33	5.0	SrF_2	"	2.08	7.69
RbI	"	2.63	5.0	BaF_2	"	?.09	7.33
AgCI	"	4.01	12.3	TiO_2	TiO_2	____	80.0
AgBr	"	4.62	13.1				

As far as dielectric materials applied to sensors are concerned, the ferroelectric materials have the greatest application.

Another important class of dielectric materials which are useful in the sensor area are those materials which exhibit the piezoelectric effect. These materials will also be discussed in a later section.

Finally, for the most complete tabulation of the properties of solid-state dielectrics, the reader is referred to the book by Von Hippel [15]. While this book is quite old, it contains comprehensive data on dielectric materials as well as a good summary of some of the theory and applications of these solids.

Table 7 Relative Permittivities and Losses in Dielectrics at frequencies of 10^6 to 10^9 Hz

	ε_r	tan δ
Ceramics		
Electrical porcelain	5.7	0.007 to 0.02
Steatites	9	0.001
Forsterites	6 to 6.5	0.0005
Glasses		
Fused silica	3.8	0.0002
Soda glass	7.6	0.009
Lead glass	7.5	0.009
Borasilicate	4.5	0.01
Micas		
Mica (muscovite)	6.5 to 8.7	6 to 10 x 10^{-4}
Mica (phlogopite)	5 to 6	1.5 to 8 x 10^{-4}
Ferroelectrics		
Rutile titanium oxide	80	0.0005
Barium titanate	1200	0.0014
Barium/strontium titanate	3000	0.005
Plastics		
Polystyrene	2.55	0.0012
Lucite	2.57	0.0032

1.4 FERROELECTRIC MATERIALS

Ferroelectric crystals are materials which exhibit a spontaneous electric polarization. In the limit as the applied electric field goes to zero, the polarization of the material does not approach zero as it does for other dielectrics. Also the direction of this spontaneous polarization can be reversed by applying an electric field of proper magnitude and direction. This process is called switching and is accompanied by an hysteresis effect. Hysteresis leads to a highly non-linear relationship between polarization and electric field intensity, such as shown in Figure 1.13. This so-called hysteresis loop is a distinquishing characteristic of ferroelectric materials. The phenomenon of hysteresis is explainable by the use of thermodynamics in the study of phase changes in solids. A discussion of this field is beyond the scope of this review; however, the reader is referred to the book by Burfoot [16] for a complete discussion of hysteresis.

Since the hysteresis loop is a feature common to all ferroelectrics, it should contain information descriptive of the material from which it is obtained. The parameters of a hysteresis loop which do characterize the given material are:

a) The spontaneous polarization, P_s - this is the value of polarization obtained by extrapolating the linear portion of the hysteresis loop to zero applied field.

b) The coercive field strength, E_c - the value of electric field intensity required to reduce the polarization to zero. E_c is not a well-defined quantity since it depends on the measuring frequency as well as the wave-form of the applied voltage.

c) The slope of the hysteresis loop connecting P and E in the limit as E approaches zero. This slope yields

the electric susceptibility of the ferroelectric material.

Most ferroelectrics cease to be ferroelectric above a temperature known as the transition temperature. Above this transition temperature the material is said to be paraelectric, and it behaves much like an ordinary dielectric. At the transition temperature is is observed that the permittivity rises to a very high value, while the electric polarization begins to decrease to small values and eventually goes to zero. The temperature at which the permittivity becomes large and that at which the polarization goes to zero do not necessarily coincide. Again, this is due to the type of phase transition which occurs in the material. The temperature at which the polarization goes to zero is called the Curie temperature. In this review, the transistion temperature will be denoted by T_o and the Curie temperature by T_c. One practically

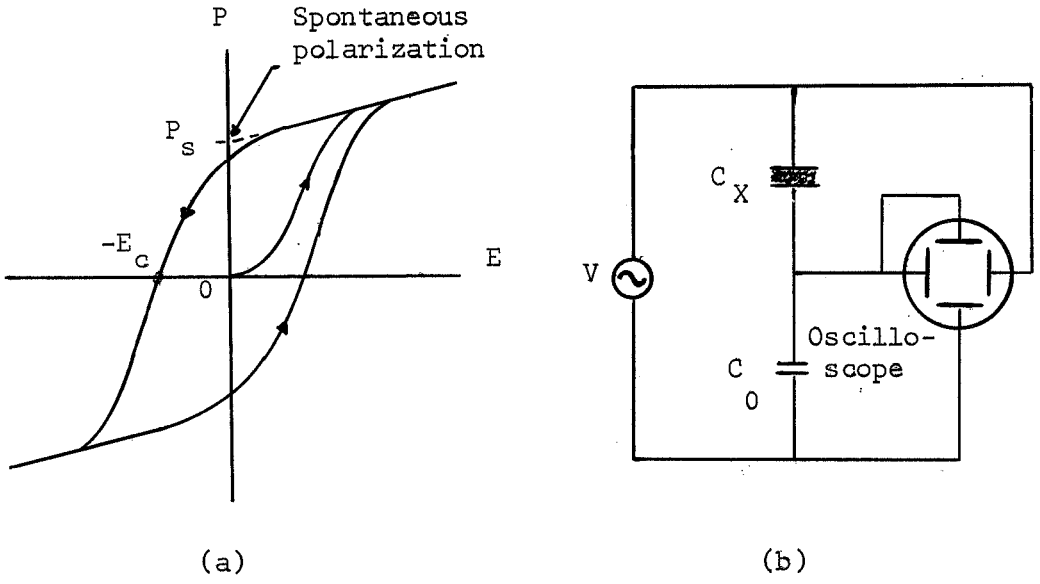


Figure 1.13. (a) Ferroelectric hysteresis loop. the coercive force is E_c (b) Circuit for display of loop. The voltage across the ferroelectric crystal C_x is applied to the horizontal plates of the oscilloscope. The linear capacitor C_0 is in series with C_x. The voltage across C_0 is proportional to the polarization of C_x and is applied to the vertical plates.

important point is that a wide range of Curie temperatures are available for ferroelectric materials. Table 8 illustrates the wide range of Curie temperatures and spontanteous polarization values which are available in ferroelectrics. In this table the spontaneous polarization is given in the esu or cgs system of units. To obtain the mks units these values should be multiplied by 2.655×10^{-7}. Also the Curie temperature can be changed by many external variables, such as pressure and electric field. It can also be modified by impurities in the ferroelectric material as well as by composition changes. This fact is an attractive one in the consideration of the use of ferroelectrics as sensors.

1.4.1 - The Occurrence of Ferroelectricity

There are basically two effects which have led to the understanding of ferroelectricity, and these two effects will be summarized now.

1) The Polarization Catastrophe

The Claussius-Mossotti relation of Equation (1.28) can be re-written as

$$\varepsilon_r = \frac{1 + \frac{2}{3\varepsilon_o} \sum_i N_i \alpha_i}{1 - \frac{1}{3\varepsilon_o} \sum_i N_i \alpha_i} \tag{1.30}$$

Thus the dielectric constant becomes infinite, when

$$\sum_i N_i \alpha_i = 3\varepsilon_o \tag{1.31}$$

From Equations 1.26 and 1.27, this is a necessary, but not sufficient, condition for the polarization to be finite for zero applied field.

Table 8 Ferroelectric crystal data

		T_c, °K	P_s(esu), at T°K	
Rochelle salt group	$NaK(C_4H_4O_6).4H_2O$	297(upper) 255(lower)	800	[278°K]
	$NaK(C_4H_2D_2O_6).4D_2O$	308(upper) 251(lower)	1100	[279]
	$LiNH_4(C_4H_4O_6).H_2O$	106	660	[95]
KDP group	KH_2PO_4	123	16,000	[96]
	KD_2PO_4	213	27.000	-
	RbH_2PO_4	147	16,800	[90]
	KH_2AsO_4	111	--	-
	KD_2AsO_4	96	15,000	[80]
	KD_2AsO_4	162	--	-
	CsH_2AsO_4	143	--	-
	CsD_2AsO_4	212	--	-
Perovskites	$BaTiO_3$	393	78,000	[296]
	$SrTiO_3$	32	(9000)	[4]
	WO_3	223	--	-
	$KNbO_3$	712	90,000	[523]
	$PbTiO_3$	763	150,000	[300]
Ilmenite (approx.)	$LiTaO_3$	-	70,000	[720]
TGS group	Tri-glycine sulfate	322	8400	[293]
	Tri-glycine selenate	295	9600	[273]

Equation 1.31 is the condition for a polarization catastrophe. The theory for the polarization catastrophe is based on thermodynamics, and on the nature of the phase transitions which occur in the material. In this theory the Gibbs' free energy function is related to polarization through [13]

$$G(P,T) = g_o + (1/2)g_2 P^2 + (1/4)g_4 P^4 + (1/6)g_6 P^6 + \ldots \quad (1.32)$$

The importance of the Gibbs' free energy is that the electric field intensity is given by

$$E = \frac{\partial G}{\partial P})_T \quad (1.33)$$

The ferroelectric state is characterized by a value of g_2 which varies with temperature according to

$$g_2 = \gamma (T - T_c) \quad (1.34)$$

where T_c is called the Curie temperature. This variation of g_2 with temperature can be accounted for through lattice vibrations [13].

There are two types of phase transitions which occur in ferroelectrics. These are first order and second order transistions. Thermodynamically, these transitions in ferroelectric materials are characterized by the sign of g_4. If:

$g_4 > 0$, second order phase transition

$g_4 < 0$, first order phase transition

The consideration of these two distinct thermodynamic cases leads to the following observations [13],[16]:

a) second order phase transition, $T_o = T_c$

b) first order phase transition, $T_o > T_c$

This theory also predicts that

$$\varepsilon_r = 1 + \frac{1}{\gamma \varepsilon_o} \cdot \frac{1}{T - T_c} \quad , T > T_c \tag{1.35}$$

which is valid for both first and second order phase transitions above the Curie temperature. The temperature dependent term in Equation (1.34) is often called the Curie-Weiss law. Thermodynamics also yields a good description of the dielectric constant below the Curie temperature and Figure 1.14 shows the dielectric constant versus temperature above and below the Curie temperature for barium titanate, one of the more useful ferroelectric materials. Also, Figure 1.15 shows curves of spontaneous polarization versus temperature for first- and second-order phase transitions. These curves illustrate a basic distinquishing feature of these two types of phase transitons.

2) Low Frequency Optical Phonons

It has been noted in Section 1.3.2 that the permittivity of a dielectric is a function of frequency. An important relationship which relates the low and high frequency dielectric constants in a crystal having some ionic character is the Lyddane-Sachs-Teller relation which is

$$\frac{\varepsilon(o)}{\varepsilon(\infty)} = \left(\frac{\omega_L}{\omega_T}\right)^2 \tag{1.36}$$

Here ω_T is the phonon frequency corresponding to the long wavelength limit of the optical phonon spectra, and ω_L is a frequency

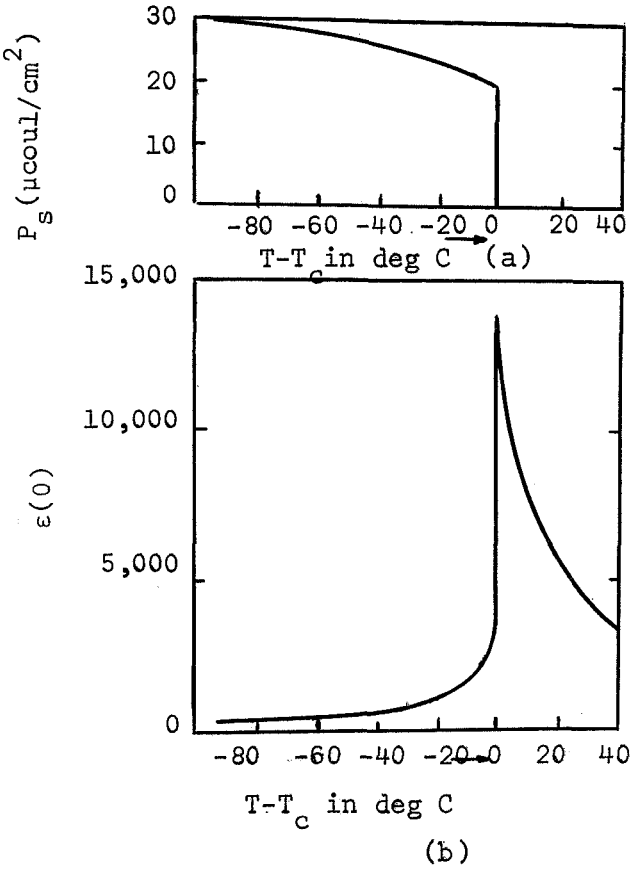


Figure 1.14. Calculated values (a) of the spontaneous polarization and (b) of the static dielectric constant as a function of temperature, with parameters as for barium titanate.

related to ω_T by

$$\omega_L^2 = \omega_T^2 + S/\varepsilon(\infty) \qquad (1.37)$$

where S is a constant. Also, $\varepsilon(o)$ is the low frequency permittivity and $\varepsilon(\infty)$ is the high frequency permittivity. Now, the range of values of $\varepsilon(\infty)$ not vary widely for most crystals. Thus, whenever very large values of $\varepsilon(o)$ are observed one would expect very low values of ω_T. For example ω_T for $BaTiO_3$ is about $.3 \times 10^{12}$ Hz at room temperature which is an extremely low value for an optical phonon frequency. The

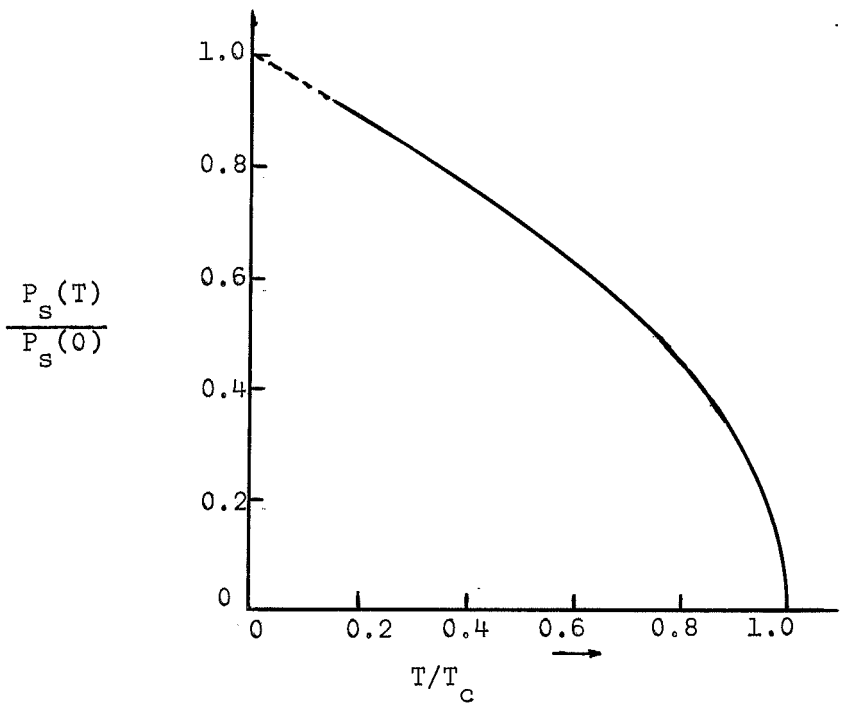


Figure 1.15. Spontaneous polarization versus temperature, for a second-order phase transition. (The curve is not realistic at low temperatures: the third law of thermo-dynamics requires that $dP_s/dT \to 0$ as $T \to 0$.

association of a large dielectric constant with a low frequency optical mode has also been supported very well by experiments on $SrTiO_3$ [13]. The lattice dynamics theory for ferroelectricity is considered by some to be a major breakthrough, and a complete discussion is contained in the review of Fatuzzo and Merz [17] .

1.4.2 Other Aspects of Ferroelectrics

Another important feature of the ferroelectric materials is that such crystals consist of regions called domains in which the polarization is in the same direction, but in which the polarization in adjacent domains are in different directions. The demarcation between two domains

is called a domain wall, and the wall is usually thought of as being so thin that its volume is negligible compared to the volume of the domains. For example in Ba Ti 0_3 the domain widths are between 10^{-4} and 10^{-2} cm, while the wall width may be of the order of one lattice constant. Each domain is an ordered region in that its dipoles have the same orientation. Quite often, adjacent domains are ordered to the same extent so that the magnitude of the polarization in these domains are the same, even though the directions are not. However, the magnitudes are usually different due to stresses and fields set up by neighboring domains.

A single crystal specimen can possess domains as long as it does not have more than one crystallite. This implies that a single crystal can have domain walls, but no grain boundaries.

The importance of domains in ferroelectrics is considerable. Not only do they affect the switching of the spontaneous polarization by an applied field, as well as to introduce additional hysteresis in the thermal phase transistions; they can also be moved through the crystal by applied fields and stresses. This phenomena can influence the properties of ferroelectrics. For a discussion of the influence of the motion of domain walls the reader is referred to Chapter 19 of Reference 15.

1.5 THE INFLUENCE OF EXTERNAL STIMULI ON THE ELECTRICAL PROPERTIES OF SOLIDS

The previous discussion in this section had dealt with the basic electrical properties of solid-state materials. Also contained in this previous discussion has been the influence of the external stimuli temperature on some of these material properties, particularly the conductivity. Due to the fact that the theory for the electrical properties is based to some extent on the statistically probable behavior of systems with approximately 10^{23} atoms/cm^3, temperature is an inherent parameter in the theory. In fact temperature is introduced into the solid state only through the concept of a thermodynamic distribution function. For this reason temperature generally affects all solid-state material properties to some extent. The temperature variations in some problems may cancel one another, but temperature will always be present in most fundamental material properties.

The purpose of the following sections is to introduce the influence of other external stimuli on the electrical properties of solids. Usually these factors do not appear in the fundamental material properties and must be considered as perturbations on the system.

1.5.1 Piezoresistance - The effect of Stress and Strain on the Electrical Conductivity of Solids

One of the most basic sensor applications is contained in the phenomena of piezoresistance. This phenomena has been used in strain and pressure measurements for many years, and these measurements have been refined to a point where extremely accurate data can be obtained for a wide variety of experimental conditions. Section 3 of this review

contains a state-of-the art summary of these measurements, and this section will be concerned primarily with the formal theory of piezoresistance itself.

First, of all, when external forces are applied to a solid, it undergoes a state of deformation, and its volume changes. If, after the deformation is removed, the solid goes back to its original shape and volume, it is said to be elastic. This review will solely be concerned with elastic solids. The deformation of an elastic solid leads to a state of stress and strain in the material. In a one-suffix tensor notation, the stresses and strains are related by

$$\sigma_i = c_{ij}\,\varepsilon_j$$

$$\varepsilon_i = s_{ij}\,\sigma_j \qquad (1.38)$$

Here σ_i are the stress components and ε_i are the strain components where i, j = 1, 2, ..., 6. The c_{ij} are called the compliance coefficients and the s_{ij} are called the stiffness coefficients. The components for i = 1, 2, 3 are called principle stresses or strains, and the components for i = 4, 5, 6 are called shear stresses or strains. The terminology here will be to refer to the effect of a stress or strain on the resistivity as piezoresistance; quite often the literature uses the term piezoresistance to refer to the influence of stress on resistivity, while the term elastoresistance is used for the effect of strain on resistivity. The compliance and stiffness coefficients reflect the symmetry of the crystal and many of ther terms vanish as the symmetry

of a crystal increases. For a good summary of the effects of crystal symmetry on the tensor properties of solids the reader is referred to Reference 18. For the cubic system the compliance co-efficient tensor is

$$[c_{ij}] = \begin{bmatrix} C_{11} & C_{12} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{12} & C_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{44} \end{bmatrix} \tag{1.39}$$

The stiffness co-efficient tensor is identical to Equation (1.39). Table 9 shows some typical values of compliance co-efficients of cubic crystals at room temperature. The parameter s in this table is defined as

$$s = (1/2)\,\frac{(C_{11} - C_{12})}{C_{44}} \tag{1.40}$$

and is called the degree of isotropy. It should be equal to unity for a perfectly isotropic solid [12].

Piezoresistance can be formulated in a formal manner due to the nature of Equation (1.5). Since any stress or strain other than hydrostatic removes the isotropy of a crystal, then the stressed conductivity of this crystal will be anisotropic. The change in the conductivity due to a general strain will be

$$\Delta\sigma_i = M_{ij}\,\varepsilon_j \tag{1.41}$$

Table 9 Elastic Constants in Units of 10^{11} Dynes/cm^2 for Some Cubic Crystals at Room Temperature.

Crystal	Structure	c_{11}	c_{12}	c_{44}	s
Cu	fcc	16.84	12.14	7.54	0.31
Ag	fcc	12.40	9.34	4.61	0.33
Al	fcc	11.29	6.65	2.78	0.84
C	diamond	95	39	43	0.65
Ge	diamond	12.88	4.83	6.71	0.60
Si	diamond	16.57	6.39	7.95	0.64
W	bcc	50.1	19.8	15.14	1.0
K	bcc	0.459	0.372	0.263	0.17
KCl	NaCl	3.98	0.62	0.625	2.69
NaCl	NaCl	4.86	1.28	1.27	1.41
AgBr	NaCl	5.63	3.3	0.720	1.62
AgCl	NaCl	6.01	3.62	0.625	1.91

where the change in conductivity has been written in a single suffix notation. The tensor $[M_{ij}]$ has the same symmetry as the crystal and for cubic crystals is of the same form as Equation (1.39). A more meaningful form of this equation is

$$\frac{\Delta\sigma_i}{\sigma(o)} = m_{ij}\,\varepsilon_j \qquad (1.42)$$

where the $m_{ij} = \frac{M_{ij}}{\sigma(o)}$ are called the elastoresistance co-efficients, and $\sigma(o)$ is the unstrained conductivity. In terms of the resistivity this equation is often replaced by

$$\frac{\Delta\rho_i}{\rho(o)} = \Pi_{ij}\,\sigma_j \qquad (1.43)$$

where now σ_j is the stress component and the Π_{ij} are denoted as the

Table 10 Elastoresistance Coefficients of Some Semiconductors

	Temperature, °K	No. of carriers, cm^{-3}	Shear coefficients		Volume dilation coefficient
			$(m_{11} - m_{12})/2$	m_{44}	$(m_{11} + 2m_{12})/3$
n-InSb	77	4.6×10^{18}	- 1.1	--1.3	-16.0
n-GaSb	300			-38.6	
n-GaAs	300		+ 0.5	- 1.4	- 7.5
n-Si	300		-79.5	-10.8	+ 5.0
n-Ge	300	$\sim 5 \times 10^{18}$	+ 0.1	-93.4	-13.6
p-InSb	77		+21.0	+101.0	- 4.4
p-Si	300		+ 3.9	+110.0	+ 6.0
p-Ge	300		+66.0	- 6.0	+ 1.0

piezoresistance co-efficients. Table 10 list the elastoresistance co-efficients of several semiconductors with cubic structure. Elasto-resistance measurements and the tabulation of such data as appears in Table 10 have been extremely valuable in studying the band structure of semiconductors.

1.5.2 The Concept of Strain Gauge Factor

The last section discussed the formal concept of piezoresistance. Practical devices are used for stress and strain measurements, and such devices are called strain gauges. The piezoresistance co-efficient indicate which materials are more sensitive, but a more practical measure of the sensitivity is important to strain gauge users. One such measure of sensitivity is the so-called gauge factor. The gauge factor for a resistance strain gauge is

defined as

$$G_F \equiv \frac{\Delta R}{R_o \quad \varepsilon_1} \tag{1.44}$$

where R_o is the unstrained resistance of the gauge, ΔR is the change in resistance due to strain, and ε_1 is the strain along the gauge axis. This can be expressed as

$$G_F \equiv 2 + \frac{\Delta \mu}{\mu \varepsilon_1} \tag{1.45}$$

where now the first term represents changes due to strain of the gauge dimensions and the second term represents changes in mobility due to strain. This definition of gauge factor is only meaningful for small strains where the resistance contains at most only terms of second order in the strain, and where the change is resistance with strain is small. If the resistance is highly non linear with strain a more meaningful definition is

$$G_F \equiv \frac{1}{R} \frac{dR}{d\varepsilon} \tag{1.46}$$

which is a "small signal" gauge factor. One can view this as the sensitivity of a resistance strain gage to small strain measurements about a given bias strain, and generally Equation (1.46) predicts a smaller sensitivity than Equation (1.44) since resistance versus strain tends to saturate. If, however, a material could be found whose resistance was super liner with strain, more sensitivity could be

achieved for small strain measurements by biasing the gage at a large strain with respect to that being measured.

For a strain gauge to be useful in strain measurements, there are three major inherent factors (factors which depend only on the gauge material before the gage is actually placed in the measurement environment) which are important to the user. These are the temperature co-efficient of resistance (TCR), the temperature co-efficient of gauge factor (TCGF),and the magnitude of the gage factor. Generally, speaking one can state that a metal will have much a lower TCR and TCGF than a semiconductor, while at the same time the metal will have a much lower gauge factor. The difference in TCR's for metals and semiconductors is apparent from Equation (1.14) and (1.15) and gauge factors will be discussed in the next two sections. The state-of-the art of semiconductor and metal strain gauge is discussed in Section 3 of this review.

1.5.3 Piezoresistance in Metals

The theory of piezoresistance in metals is in a much less satisfactory state than is the practical application of metals as strain sensors. In fact, one would expect that if the theory of resistivity is not in a satisfactory state, then neither would be the theory of piezoresistance. Some progress in this regard has been made in the last few years [19]. However , a Ziman points out the calculations are complicated, and the mathematical formulation can only be unraveled with aid of presently lacking experimental observations.

Most measurements of piezoresistance in metals are made on non-

crystalline samples, either in the form of extruded wire or rolled foils. Thus the anisotropic nature of piezoresistance is not important and is seldom mentioned in the literature. The majority of metals have a resistivity which decreases with increasing pressure, and with an upward curvature. In some metals, the resistivity eventually passes through a minimum and begins to increase with further increasing pressure. Figure 1.16 illustrates this point for several of the akali metals. Such structure in resistivity versus pressure curves are due to many effects such as scattering processes, and even phase transitions [6]. A good summary of the present state of pressure effects in metals is contained in Reference 6, Chapter 6. Table 11 illustrates typical values of gage factor for some of the more commonly used strain gage metals, along with a value for nickel which is considered to be anomalous. Then Table 12 gives values of gage factor, TCR, and TCGF for a semiconductor and a metal strain gage. These numbers are quite typical for strain gages of these two types of materials. Also included in this table are values of Seebeck co-efficient and linear thermal expansion co-efficient for these two materials. The Seebeck co-efficient is important in determining thermocouple effects when leads are attached to the gage, and the linear expansion co-efficient is important in considering bonding agents for attaching strain gauges to test vehicles. The Seebeck effect is discussed in a later section.

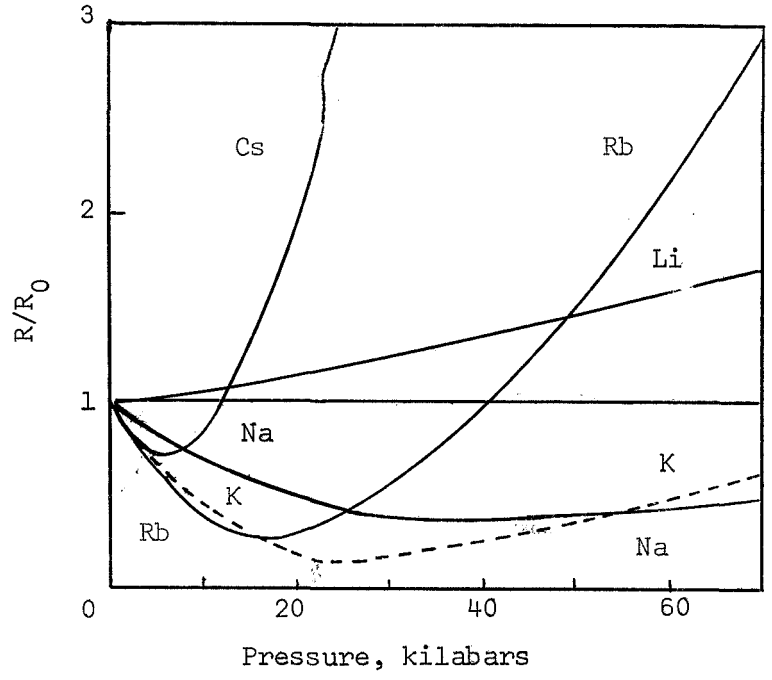


Figure 1.16. Change of resistance of the alkali metals at room temperature with change of pressure.

1.5.4 Piezoresistance in Semiconductors

The theory of piezoresistance in semiconductors is in a reasonably good state of affairs for the elemental semiconductors silicon and germanium. It has been an important phenomenon in the study of the band structure of other semiconductors and has often provided the experimental information necessary to carry out theoretical calculations in this area. However, the information is not by any means complete. For example, it is only quantitatively understood why p-type silicon and germanium have large piezoresistance co-efficients. By and large, the piezoresistance effect in semiconductors is usually determined by the band structure of these materials, and one can obtain gauge factors two orders of magnitude larger than dimensional changes alone would yield. This fact can generally be attributed

Table 11 Gauge Factors of Wires of Different Compositions

Material	Approximate Composition	Average G_F
Advance	55 Cu, 45 Ni	2.1
Nichrome	75 Ni, 12 Fe, 11 Cr, 2 Mn	2.5
Constantan	60 Cu, 40 Ni	1.8
Isoelastic	36 Ni, 8 Cr 50 Fe, Mn, Si, Mo	3.54
Manganin	9-18 Mn, 1 1/2 -4 Ni, Remainder Cu	0.5
Nickel		-12

to complex energy band structures of the many-valley or degenerate type [20]. Such effects must be highly orientation dependent, and this is observed for piezoresistance in semiconductors. For example, the gauge factor in one ohm-cm n-type silicon is about 125 for the (100) crystal orientation and it is about 2.5 for gauge cut along the (111) direction. Obviously, this introduces another consideration for semiconductor strain gauge users. Other factors aside from the band structure can contribute to the piezoresistance of semiconductors. Impurities can have a pronounced effect on not only the gauge factor, but on the TCR and TCGF as well. These are areas which are less understood than band structure effects, and it will be pointed out later that these appear to offer promising areas for strain gauge research.

A state of deformation in a crystal changes the allowed energy states for electrons since the relative atomic spacings in the lattice are changed. The change in the conduction bands and the valence bands with strain determine changes in carrier concnetration as well as

Table 12 Comparison of several properties of a semiconductor strain gauge to like properties of a strain gauge wire.

Property	Material: 0.1 ohm-cm p-type Si, 111	Material: Karma strain gauge wire	Ratio
Strain gauge factor	125 (nom)	2.0	62.5
Thermal coeff. res., $\frac{dR}{R_0} \cdot \frac{1}{\Delta T} \left(\frac{\%}{^\circ C} \right)$	12×10^{-4}	0.2×10^{-4}	60
Thermal coeff. G.F., $\frac{dG}{G_0} \cdot \frac{1}{\Delta T} \left(\frac{\%}{^\circ C} \right)$	16×10^{-4}	5×10^{-4}	3
Seebeck coeff., alpha (uv/°C)	600	40	15
Linear expansion coeff., K $\left(\frac{\text{inch/inch}}{^\circ C} \right)$	4×10^{-6}	10×10^{-6}	0.4

influence the mobility of these carriers, and such band changes must be calculated if a theory for piezoresistance in semiconductors is to be formulated. Unfortunately, there is no unified theory for all semiconductors which can accomplish this task, since such a theory depends on the un-deformed band structure of the semiconductor in question. A good summary of the state of knowledge of the detailed nature of the band structure of twenty elemental, III-V, and II-VI semiconductors is contained in Appendix C of Reference 20. Since no unified theory exists for the effects of strain on the energy bands of a given semiconductor, the following treatment will summarize the results for the three most widely studied and understood semiconductors, which are silicon, germanium, and gallium arsenide. This treatment serves to introduce the concept of the deformation potential, as well as to illustrate a wide variety of conditions which might be encountered in the study of any semiconductor.

Silicon and germanium, most of the III-V compound semiconductors, and several of the II-VI compound semiconductors have valence bands which are of the complex degenerate type. These consist of two bands whose constant energy surfaces are warped spheres which are degenerate at k=0, and a single spin-orbit split-off band. The two degenerate bands are known as the heavy-and light-hole bands, because of the large difference in effective mass of the two. The conduction bands of silicon and germanium are of the simple many-valley type [20], while the majority of the III-V and II-VI semiconductors have a single conduction band at k=0. Such compound semiconductos are called direct band gap semiconductors because their conduction and valence bands extrema both

lie at k=0. Figure 1.17 illustrates the band structure of silicon, germanium, and gallium arsenide.

The effects of stress on the band structure of these three materials has been tabulated by Wortman and Wooten [21]. The theoretical procedure for the valence bands requires a knowledge of the effects of strain on the Hamiltonian for holes in these bands. If this is known, the technique is to diagonalize this Hamiltonian in matrix form to obtain the energy eigenvalues, which will be a function of strain. For

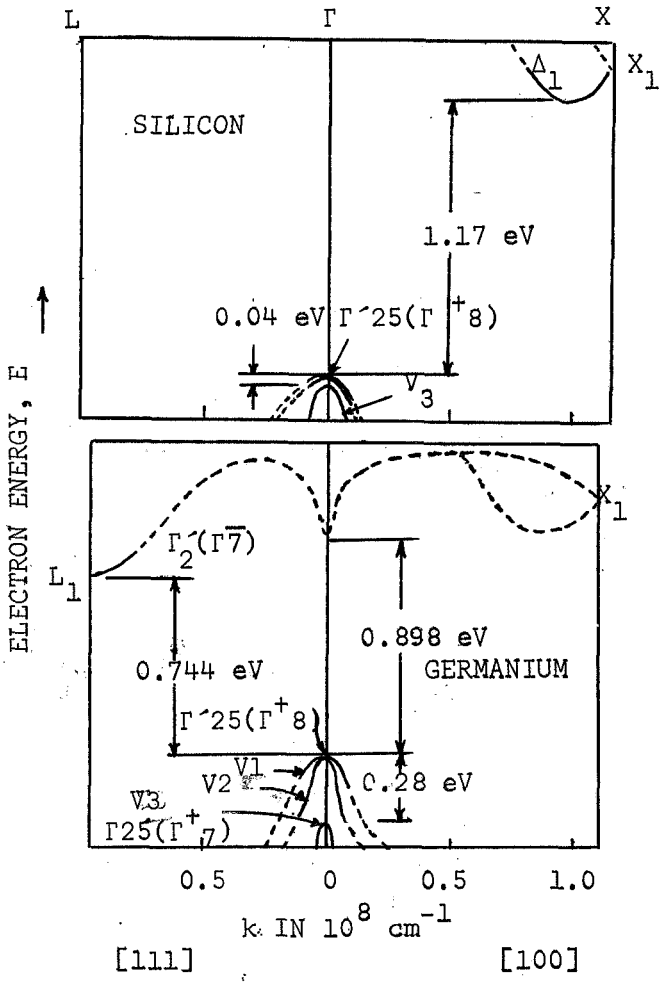


Figure 1.17a. E(k) curves in <111> and <100> directions in Si and Ge. The important energy band separation values at 0°K, determined mainly from optical measurements at liquid helium temperature, are indicated. The levels at the principal symmetry points are labeled in single group notation, and the corresponding double group labels are given in parentheses.

silicon and germanium this procedure yields

$$\Delta E_{V1} = D_d\varepsilon + \{(2/3\ D_u)^2\ (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 - \varepsilon_1\ \varepsilon_2 - \varepsilon_1\ \varepsilon_3 - \varepsilon_2\ \varepsilon_3)$$

$$+ 1/3(D_u^1)^2(\varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2)\ \}^{1/2} \qquad 1.47$$

$$\Delta E_{V2} = D_d\varepsilon - \{(2/3\ D_u)^2(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 - \varepsilon_1\ \varepsilon_2 - \varepsilon_1\ \varepsilon_3 - \varepsilon_2\ \varepsilon_3)$$

$$+(1/2)\ (D_u^1)^2(\varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2)\}^{1/2}$$

where $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$. Here ΔE_{V1} is the change in the heavy hole band and ΔE_{V2} is the change in the light hole band. The parameters D_d, D_u, and D_u^1 are deformation potential co-efficients, and they describe the behavior of the band edge points with strain. It can be seen that

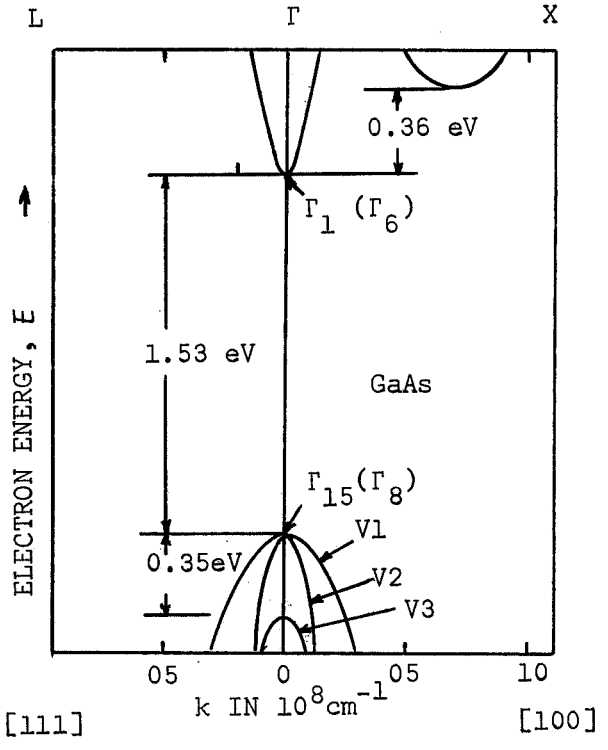


Figure 1.17b. $E(k)$ curves in <111> and <100> directions in GaAs. The important energy differences determined experimentally at 0° K are indicated. Notations are as in Fig. 1.17a.

the band edges are shifted in energy as well as split in energy; that is, the degeneracy at k=0 is removed. Here, the split-off band has been completely neglected, which should be a reasonably good approximation for temperatures less than the split-off energy divided by Boltzmann's constant, which is generally greater than room temperature.

The conduction bands of both silicon and germanium can be described by a phenomenogical deformation potential approach for strains less than 1%. For larger strains in silicon the description is complicated by the fact that the band edge points in k-space shift with strain. The phenomena will be neglected here. The deformation potential approach states that

$$\Delta E_c^{(i)} = \sum_{j=i}^{6} \Xi_j^{(i)} \varepsilon_j \tag{1.48}$$

where $\Delta E_c^{(i)}$ is the change in energy of the ith conduction band edge. The $\Xi_j^{(i)}$ are again conduction band deformation potential co-efficients. By the use of symmetry considerations, it can be shown that two deformation potential constants are sufficient to describe the conduction band shifts of both silicon and germanium [22]. For silicon, this treatment yields:

$$\begin{aligned} \Delta E_{c1} &= \Xi_D \varepsilon + \Xi_u \varepsilon_1 \\ \Delta E_{c2} &= \Xi_D \varepsilon + \Xi_u \varepsilon_2 \\ \Delta E_{c3} &= \Xi_D \varepsilon + \Xi_u \varepsilon_3 \end{aligned} \tag{1.49}$$

where the subscripts 1, 2, and 3 refer to the valleys along the (100), (010), and (001) direction and their inverses, respectively. For germanium, the results are

$$\Delta E_{c1} = (\Xi_D + 1/3\ \Xi_u)\ \varepsilon + 1/3\ \Xi_u\ (\varepsilon_4 + \varepsilon_5 + \varepsilon_6)$$

$$\Delta E_{c2} = (\Xi_D + 1/3\ \Xi_u)\ \varepsilon + 1/3\ \Xi_u\ (\varepsilon_4 - \varepsilon_5 - \varepsilon_6)$$

$$\Delta E_{c3} = (\Xi_D + 1/3\ \Xi_u)\ \varepsilon + 1/3\ \Xi_u\ (-\varepsilon_4 + \varepsilon_5 - \varepsilon_6) \qquad (1.50)$$

$$\Delta E_{c4} = (\Xi_D + 1/3\ \Xi_u)\ \varepsilon + 1/3\ \Xi_u\ (-\varepsilon_4 - \varepsilon_5 + \varepsilon_6)$$

Here the subscripts 1, 2, 3 and 4 refer to the valleys along the (111), (111), $(1\bar{1}\bar{1})$, and $(\bar{1}1\bar{1})$ directions and their inverse, respectively.

For gallium arsenide, the conduction band minimum at k=0 can be described by a single deformation potential, that is

$$\Delta E_c = D_d^c\ \varepsilon \qquad (1.51)$$

The valence bands are assumed to have the same dependence on strain as the bands of germanium and silicon, as given by Equation 1.47. For gallium arsenide, the deformation potential D_d in this equation is usually denoted by D^V_d. For most III-V and II-VI compounds the deformation potential constants are not known, although the dependence of the band gap on hydrostatic pressure has been experimentally determined. The following table contains the information which is known about the deformation potential constants for silicon, germanium, and gallium arsenide.

Table 13 - Deformation Potential Constants for Silicon, Germanium, and Gallium Arsenide in Units of eV per Unit Strain

Silicon	Germanium	Gallium Arsenide
$D_u = 2.04$	$D_u = 3.15$	$D_d^c - D_d^V = 8.7$
$D_u^1 = 2.68$	$D_u^1 = 6.06$	$D_u = 3.15$
$\Xi_u = 11$	$\Xi_u = 19.2$	$D_u^1 = 5.63$
$D_d - (\Xi_D + 1/3\ \Xi_u) = -1.44$	$D_d - (\Xi_D + 1/3\ \Xi_u) = 4.82$	

An examination of Table 13 leads one to believe that the effects of stress on the three listed semiconductors would be about the same. However, while band edge changes are on the same order for most semiconductors, the piezoresistance effect is not. The reason for this is that an additional mechanism not contained in band edge changes by themselves has a drastic effect on piezoresistance. This effect is called the population transfer effect. It is due to the fact that in many-valley semiconductor the various conduction (or valence bands) change energy with respect to each other and thus become unequally populated. Since the mobilities of carriers in the different valleys are generally unequal (the conductivity mobility is still anisotropic), carriers can be transferred from high to low mobility bands, and vice versa, which is a piezoresistance effect. For a many valley n-type semiconductor, the conductivity can be written as

$$\sigma = q \sum_{i=1}^{m} \mu_i n_i = q \mu_{eff} n$$

and (1.52)

$$\mu_{eff} = \sum_{i=1}^{m} \mu_i \left(\frac{n_{ci}}{n}\right)$$

Here m is the number of valleys and n is the total electron concentration. Since changes in n due to strain are on the order of dimensional changes the gage factor for small strains is approximately

$$G_F = \frac{1}{\mu_{eff}} \frac{\partial \mu_{eff}}{\partial \varepsilon} \tag{1.53}$$

For silicon in the (100) direction the gage factor becomes

$$G_F = -\frac{2}{3} \frac{K_u - 1}{2K_u + 1} \frac{\Xi_u}{k_B T} \tag{1.54}$$

where $K_u = \frac{\mu_T}{\mu_L}$, the ratio of transverse to longitudinal valley mobility, is about 4.7 for silicon. This gives a gage factor of about 105 at 300°K which is close to the measured value. An important part of Equation 1.54 is that it illustrates the tremendous variation of gauge factor with temperature of a silicon strain gauge. It predicts a temperature coefficient of 33×10^{-4} at room temperature. This can be compared with the values listed in Table 12.

One other point should be mentioned about the piezoresistance of semiconductors. It is possible to cause the total electron concentration to change significantly larger with strain than dimensional changes would

allow by using impurities to compensate the unstrained resistivity. If an impurity is electrically active and has an energy near the Fermi energy, then by straining the crystal the ionization state of the defect can be changed. This will change the total electron concentration. For this case the gage factor becomes

$$G_F = \frac{1}{\mu_{eff}} \frac{\partial \mu_{eff}}{\partial \varepsilon} + \frac{1}{n} \frac{\partial n}{\partial \varepsilon}$$

or (1.55)

$$G_F = G_F(0) + \frac{1}{n} \frac{\partial n}{\partial \varepsilon}$$

where $G_F^{(0)}$ is the gage factor due to the population transfer effect as given by Equation (1.54). This additional term can change both the magnitude of gage factor as well as the TCGF and the TCR. For example, the use of radiation damage caused by electrons has been shown to reduce the gage factor in n-type silicon to about 50-100 while simultaneously reducing both the TCR and TCGF to values which approach those by metals [23]. This is an area which offers promise in improving the performance of semiconductor gages since impurities can be introduced chemically which should cause the same effect as electron radiation.

In conclusion, it is apparent that while semiconductor (specifically silicon) strain gauges have larger sensitivities, they also have large thermal variations which are difficult to overcome. Thus the tendency in the strain measurements seems to have been to avoid semiconductor strain gages unless temperature variations are not important. However,

in certain cases the thermal variations of silicon strain gages has not detracted from their use over a wide range of temperatures [24]. In fact, in this work, miniature pressure transducers have been constructed which have zero shifts and sensitivity shifts (see section 3 for definitions of these factors) of less than ±3% throughout the temperature range of -450°F to + 250°F. This type of performance has been attributed to the optimization of gauge characteristics through choice of doping level, as well as optimization of excitation source impedance, bridge circuit parameters, and the housing of the sensor. The point to be made is that in some instances there are many more factors than simply those inherent to the strain gauge material which are important to strain sensor design. Often these parameters as well as those of the gauge can be varied to obtain optimum sensor design. These factors will be discussed in section 3 of this report.

1.5.4.1 The Pinch Effect

Large stress effects have been reported in the Russian literature on intrinsic germanium, and these results have been recently summarized by Pittelli and Rindner [25]. This effect has been given the name "pinch effect" because the carrier concentration and current flow can be confined to a thin region near one of the sample surfaces [26], when the sample is under the influence of an applied stress. This effect leads to a non-linear voltage current characteristic under stress, which has been attributed to the occurrence of transverse (i.e. perpendicular to the direction of the applied field) carrier concentration gradients which are present due to the anisotropic nature of the stressed semi-

conductors. From the reported experimental results, the pinch effect appears to give current changes more than an order of magnitude larger than those observed in optimally oriented extrinsic semiconductor strain gages. The mechanism is extremely sensitive to the minority carrier concentration, and in extrinsic materials, the pinch effect is reduced by the ratio of the minority to the majority carrier concentrations. Thus it would not be of any significance in only slightly extrinsic material, such as n-type silicon or germanium where $n_o \geq 10\ n_i$. This is severe restriction on its application; however, many semiconductors can be produced in intrinsic form, and the pinch effect might provide new materials from which stress sensors could be made.

The mathematical treatment of the pinch effect is extremely complicated, and will not be treated here. The reader is referred to the treatments in References 25 and 26. The usefulness of the pinch effect as a stress sensor can be summarized by stating that it seems to offer very high sensitivities and allows the use of some materials (such as InSb) which have not found much application as mechanical sensors. These advantages are offset by the stringent requirement of crystallographic orientation control, by the stringent requirement placed on the control of surface recombination velocity, a high frequency response limit, and large non linearities in the stress dependence of the effect [25].

1.5.5 The Piezoelectric Effect

The direct piezoelectric effect in a crystal results when an electric polarization is produced by the application of a stress. The converse piezoelectric effect refers to a mechanical deformation of the crystal

when it is subjected to an external electric field. There is a one-to-one correspondence between the direct and converse effects since a reversal of the direction of the electric field reverses the sign of the field- induced strain. This is an important fact since it distinguishes the phenomena of piezoelectricity from that of electrostriction. Electrostriction is the straining of a material from its being placed in an electric field. However, electrostriction is a second-order effect in that the strain is an even function of the applied field and the strain will not change sign by changing the direction of the applied field. Furthermore, there is no inverse effect in that no polarization will result when the crystal is strained due to electrostriction alone. Usually, electrostrictive effects are much smaller than piezoelectric effects except in certain ferroelectric crystals [27]. Thus electrostriction will be neglected in the following discussion.

An excellent example of the mechanisms for piezoelectricity in quartz can be seen in the following discussion. Quartz is probably the most well-known piezoelectric crystal. Figure 1.18a shows a pictorial model of the quartz structue with each silicon atom assumed to be positively charged and each double oxygen atom to be negatively charged. For a compressional stress, the angle Θ changes to Θ', the distance O-O increases due to a Coulomb repulsion, while the distance A-B stays the same since there is no net force on the atoms at A & B. Thus, in Figure 1.18b it can be seen that there now results a net dipole moment in the direction of B-A due to the displacement of positive sites C and E to the left and negative sites D and F to the left. Figure 1.18c illustrates a similiar effect for a "shear" stress perpendicular to the compressional

stress. This particular model is only descriptive and can be carried further only with great care. More complete analysis requires use of the dielectric properties of soldis discussed in Section 1.3 and of the mechanical relations discussed Section 1.5.1.

Some general observations can be made about the types of crystals which will exhibit piezoelectricity. A detailed discussion is beyond the scope of this review; however, the interested reader is referred to Reference 18 for such a discussion. These observations can be summarized simply by stating that any crystal prossessing a center of symmetry (also called a center of inversion) cannot exhibit the piezoelectric effect.

Piezoelectricity is a tensor property of a crystal and is an anisotropic effect. In terms of the polarization per unit volume due tc an applied stress the piezoelectric co-efficients are defined by the

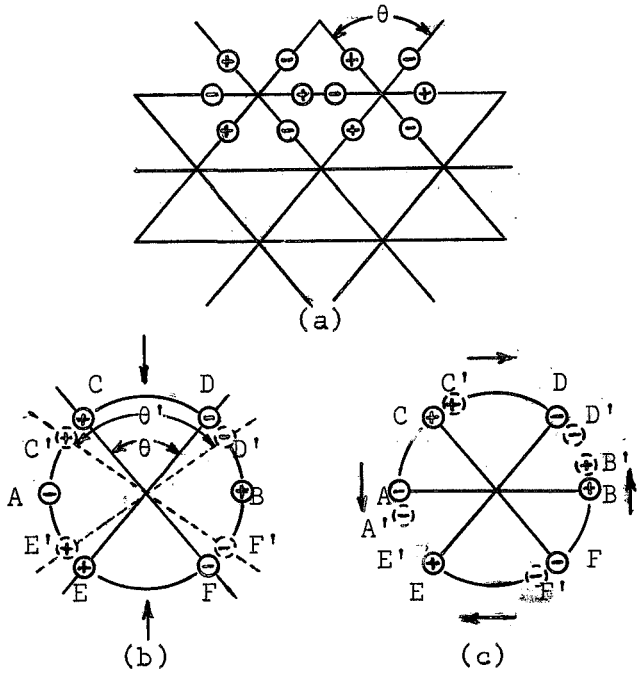


Figure 1.18. The model used by Lord Kelvin to explain the piezoelectric effect in quartz.

relation

$$P_i = d_{ij}\,\sigma_j \tag{1.56}$$

In a similar manner the inverse piezoelectric effect is defined by the field-induced strain by the equation

$$\varepsilon_i = d_{ij}\,E_j \tag{1.57}$$

Thus for piezoelectric materials, Equations (1.21) and (1.37) must be replaced by

$$\begin{aligned} D_i &= \varepsilon_o\,\varepsilon_{r_{ij}}\,E_j + d_{ij}\,\sigma_j \\ \varepsilon_i &= d_{ij}\,E_j + S_{ij}\,\sigma_j \end{aligned} \tag{1.58}$$

To avoid confusion the tensor permittivity components have been written as $\varepsilon_o\,\varepsilon_{r_{ij}}$, while the strain components have been denoted as ε_i. As has been noted, crystal symmetry has a definite effect on the tensor properties of a crystal, and a listing of the piezoelectric co-efficients for the twenty crystal classes showing piezoelectric properties is contained in Reference 18.

There is a physical interpretation one can place on the d- co-efficients of a piezoelectric material, as well as other co-efficients one can identify for these materials which themselves have physical significance. First of all, the d - co-efficients can be interpreted as the amount of charge per unit area output for a unit stress (force)

input. Similiarly, it can be interpreted as the amount of strain output obtained for a unit applied voltage input with no stress applied. These two conditions follow from Equation (1.58) by setting E=0 and $\sigma_j = 0$, respectively. Ohter co-efficients which are useful for piezoelectric materials are:

1) The g-co-efficient is obtained by dividing the d-co-efficient by the permittivity of the material, which gives the generated electric field output per unit stress (force) input under open circuit conditions. This can be seen from Equation (1.58) when $D_i = 0$.

2) The h-co-efficient is obtained by multiplying the g-co-efficient by Young's Modulus for the appropriate crystal orientation of the material. This then gives the generated electric field output per unit strain input under open circuit conditions, since Young's modulus is defined as

$$Y_i = \frac{\sigma_i}{\varepsilon_i} \tag{1.59}$$

3) The coupling co-efficient, or the k-co-efficient, can be computed from the square root of the product of the d- and h- co-efficients. This represents the square root of the ratio of mechanical energy to electrical energy and can be considered a measure of the efficiency of the crystal as an energy converter. It should be noted, that since the k-co-efficient uses the definition of the d- and h co-efficients stated above, it refers to an unloaded output condition without mechanical restraint. This is not the normal way in which piezoelectric transducers are operated. The k- co-efficient gives an ideal efficiency.

A more elaborate discussion of the material parameters and co-efficients which characterize piezoelectric materials can be found in the book by Mason [28]. A tabulation of several piezoelectric materials, along with values of the d-, g-, and k-coefficients and other useful parameters, is given in Table 14. Some of the materials listed in this table are denoted by their commercial names, and references to these names are given at the end of the table.

Table 14
Characteristics of some piezoelectric transducer materials

Group Number		Cut	Basic-plate action	Free dielectric constant	d-coefficient	g-coefficient
	Dimensions	..	..	..	$\frac{\text{coulb}^{1}}{\text{newton}}$	$\frac{\text{V/m}}{\text{N/m}^2}$
	Multiply figures by	..	..	1	10^{-12}	10^{-2}
1	Quartz	0°X	TE	4.5	2.3	58
		0°X	LE	4.5	2.3	58
2	Tourmaline . . .	0°Z	TE	6.6	1.9	33
		0°Z	VE	6.6	2.4	41
3	Rochelle salt at 30°C .	0°X	FS	350^{2}	550	180
		45°X	LE	350^{2}	275	90
		0°Y	FS	9.4	54	664
		45°Y	LE	9.4	27	332
4	Ammonium dihydrogen phosphate, ADP . . .	0°Z	FS	15.3	48	354
		45°Z	LE	15.3	24	177
5	Lithium sulphate,LH .	0°Y	TE	10.3	16	175
		0°Y	VE	10.3	13.5	148
6	Barium titanate compounds: 'Ceramic' A^{4} .	..	TE	1700	190	12.6
		..	LE	1700	78	5.2
		..	VE	1700	34	2.2
	'Ceramic' B^{4} .	..	TE	1200	140	13.0
		..	LE	1200	56	5.3
		..	VE	1200	28	2.4
	Unipiezo H 254^{5} .	..	TE	1000	80	9.1
		..	LE	1000	30	3.4
		..	VE	1000	20	2.3
	TCL 1^{6} . . .	..	TE	1400	150	12.2
		..	LE	1400	60	4.9
		..	VE	1400	30	2.4

Table 14 (cont'd)

TCL 2^6	..	TE	1100	100	10.3
	..	LE	1100	40	4.1
	..	VE	1100	20	2.1
TCL 3^6	..	TE	1050	83	9
	..	LE	1050	35	3.8
	..	VE	1050	13	1.4

Table 14 (cont'd)

Group number	Material		Out	Basic-plate action	Free dielectric constant	d-coefficient	g-coefficient	
	Dimensions		..	..	...	$\frac{\text{coulb}^1}{\text{newton}}$	$\frac{\text{V/m}}{\text{N/m}^9}$	
	Multiply figures by		..	..	...	10^{-12}	10^{-4}	
7	Lead-zirconate-titanate compounds	TCL 4^6 . . .	..	TE	630	90	16-1	
			..	LE	630	37	6.7	
			..	VE	630	10	2.7	
		TCL 6^4 . . .	..	TE	1700	100	6.7	
			..	LE	1700	40	2.7	
			..	VE	1700	20	1.3	
		TCL 7^6 . . .	..	TE	1650	105	7.3	
			..	LE	1650	42	2.9	
			..	VE	1650	21	1.4	
		LZ $4a^7$ . . .	..	TE	1200	..	..	
			..	LE	1200	131	12.5	
			..	VE	1200	..	..	
		LZ 5 a^7 . . .	..	TE	1500	320	23	
			..	LE	1500	140	10	
			..	VE	1500	40	3	
8	Lead niobate compounds	$Pb(NbO_3)_2{}^3$ . .	..	TE	270	90	37	
			..	LE	270	33	14	
			..	VE	270	24	9	
		'Ceramic' 260^2.	..	TE	1500	200	15	
			..	LE	1500	80	6	
			..	VE	1500	40	3	
		" " 330^9.	..	TE	900	80	11	
			..	LE	900	32	4	
			..	VE	900	10	3	

Group Number	Coupling Coefficient k	Volume Resistivity	Density	Young's Modulus	Maximum safe stress	Maximum safe temp.	Safe Humidity Range	Remarks
	%	ohm m	$kg\ m^{-3}$	$newton\ m^{-3}$	$newton\ m^{-2}$	C	%	
	1	10^9	10^3	10^9	10^6	1	1	
1	11	>1000	2.65	80	98	550	0-100	$Q=10^5$
	10	>1000	2.65	80	98	550	0-100	
2	10	> 100	3.10	160	..	...	0-100	
	..	> 100	3.10	...	..	...	0-100	
3	76	> 10	1.77	...	14.7	45	40-70	Twister bimorph
	73	> 10	1.77	19.3	14.7	45	40-70	Bender bimorph
	32	> 10	1.77	...	14.7	45	40-70	Twister bimorph
	29	> 10	1.77	10.7	14.7	45	40-70	Bender bimorph
4	32	$>0.1^3$	1.80	...	20.6	125	0-94	Twister bimorph
	28	$>0.1^3$	1.80	19.3	20.6	125	0-94	Bender bimorph
5	38	> 10	2.06	46	..	75	0-95	
	..	> 10	2.06	...	..	75	0-95	
6	50	> 100	5.7	110	80	70	0-100	
	21	> 100	5.7	110	80	70	0-100	Q=400
	..	> 100	5.7	...	..	70	0-100	
	46	> 100	5.5	110	80	70	0-100	
	18.5	> 100	5.5	110	80	70	0-100	
	..	> 100	5.5	...	..	70	0-100	
	30	> 10	5.5	100	..	70	0-100	
	17	> 10	5.5	100	..	70	0-100	
	..	> 10	5.5	100	..	70	0-100	
	40	> 30	5	91	..	70	0-100	
	15	> 30	5	91	..	70	0-100	
	..	> 30	5	..	..	70	0-100	
	29	>0.1	5.3	83	..	70	0-100	
	11	>0.1	5.3	83	..	70	0-100	
	..	>0.1	5.3	..	..	70	0-100	
	28	>100	5.7	106	..	70	0-100	
	12	>100	5.7	106	..	70	0-100	
	..	>100	5.7	..	..	70	0-100	

Table 14 (cont'd)

Group Number	Coupling coefficient k	Volume resistivity	Density	Young's modulus	Maximum safe stress	Maximum safe temperature	Safe humidity range	Remarks
	%	ohm m	$\frac{kg}{m^3}$	$\frac{newton}{m^2}$	$\frac{newton}{m^2}$	°C	%	
	1	10^9	10^3	10^9	10^6	1	1	
7	9.3	> 5	7.2	6	95	250	0-100	
	4	> 5	7.2	6	95	250	0-100	
	..	> 5	7.2	..	..	250	0-100	
	19.5	> 5	7.5	55	..	250	0-100	
	7.4	> 5	7.5	55	..	250	0-100	
	..	> 5	7.5	..	..	250	0-100	
	20	> 30	7.5	55	..	250	0-100	
	8.2	> 30	7.5	55	..	250	0-100	
	..	> 30	7.5	..	..	250	0-100	
	..	$>10^4$	..	81.5	..	250	0-100	
	35.5	$>10^4$	..	81.5	..	250	0-100	Q = 500
	..	$>10^4$	..	81.5	..	250	0-100	
	67.5	$>10^4$	..	67.5	..	250	0-100	
	31.8	$>10^4$	..	67.5	..	250	0-100	Q = 75
	..	$>10^4$	..	..	..	250	0-100	
8	31	..	..	29	..	250	0-100	
	11.5	..	..	29	..	250	0-100	
	..	..	..	..	..	250	0-100	
	51	>100	5.9	88	20	270	0-100	
	21	>100	5.9	88	20	270	0-100	
	..	>100	5.9	..	..	270	0-100	
	27	>100	6	92	20	270	0-100	
	11	>100	6	92	20	270	0-100	
	..	>100	6	..	..	270	0-100	

1.5.6 Modes of Operation of Piezoelectric Materials

The most common mode of operation for a piezoelectric crystal is when an electric field is used to excite an elastic resonance in a sample cut in a specific lattice orientation. Another major field of use is in the detection and generation of acoustic power, where the piezoelectric ceramic materials have found great application. While such applications are different in the sense that they require different sample configurations as well as different input conditions, their basic operation can be described in terms of several similar modes in which a piezoelectric crystal can be excited. These so-called natural modes of vibration are: [28].

a) Longitudinal expander (LE) - electric field perpendicular to length

b) Longitudinal expander (LE) - electric field parallel to length

c) Thickness expander (TE) - electric field parallel to thickness

d) Thickness shear (TS) - electric field parallel to thickness

e) Face shear (FS) - electric field normal to larger faces

f) Volume expander (VE) - sample subjected to hydrostatic pressure

g) Planor mode of a disk

h) Hoop made of a ring (axial electric field)

i) Hoop made of a ring (radical electric field)

j) Breathing mode of a spherical shell

Modes a-f are the most common modes of operation and can generally be excited in all piezoelectric crystals. Modes g-h are less common and

often require the use of certain properties of piezoelectric crystals such as the permanent polarization of the ferroelectric ceramics in order to be excited [28]. Figure 1.19 illustrates the deformation modes a-f for single piezoelectric elements.

The explanation of these basic modes of operation for piezoelectric elements is carried out using a continuum model to arrive at an electrical equivalent circuit for both the electrical and mechanical portions of the system. As an example, consider a thin dielectric slab as shown in Figure 1.20. Here an alternating electric field is applied along the z-axis, and the thickness b and width a are assumed to be much smaller than the length ℓ. This corresponds to the longitudinal expander (LE) mode, and the surfaces normal to the y and z axes can be considered free of any stress. Thus in the previous notation

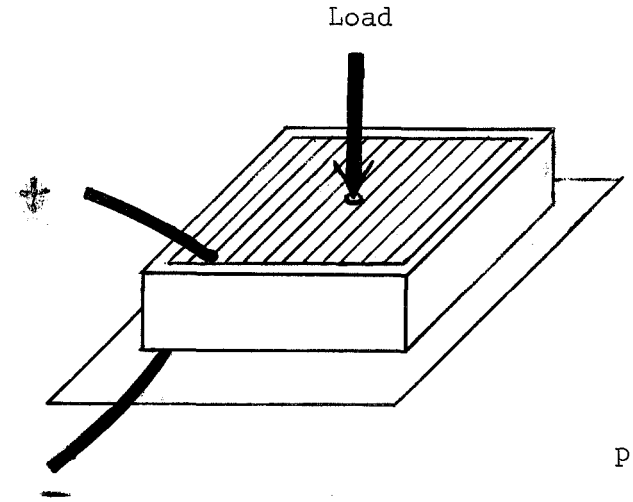
$$\varepsilon_1 = S_{11}\,\sigma_1 + d_{13}\,E_3 \tag{1.60}$$

is x- component of strain. Now, from Newton's Laws of motion, the displacement of the sample in the x direction, ξ, obeys the following differential equation

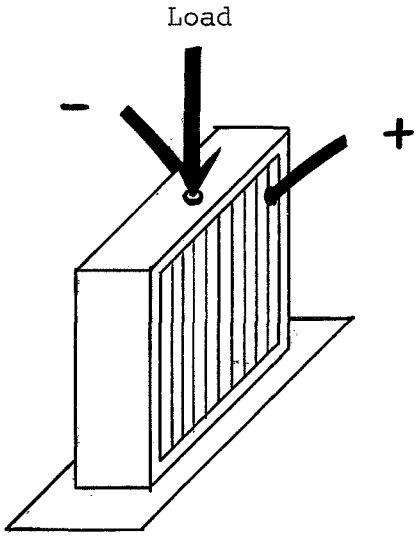
$$\rho_m \frac{\partial^2 \xi}{\partial t^2} = \frac{\partial \sigma_1}{\partial x} \tag{1.61}$$

where ρ_m is the mass density of the sample. Since

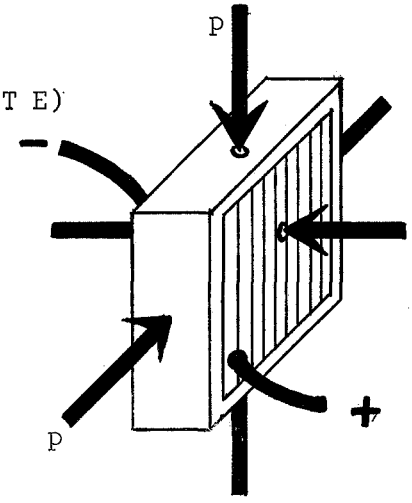
$$\varepsilon_1 = \frac{\partial \xi}{\partial x} \tag{1.62}$$



(a) Thickness expander (T E)

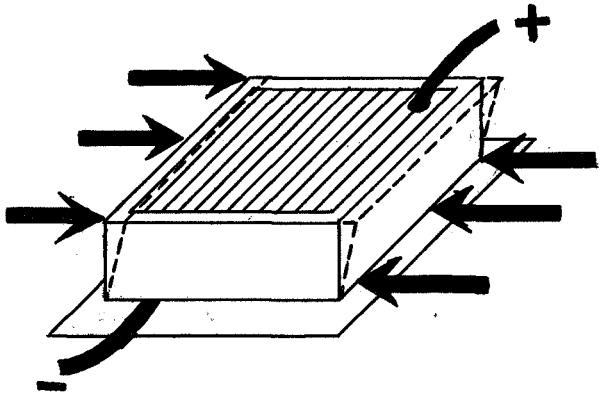


(b) Length expander (L E)

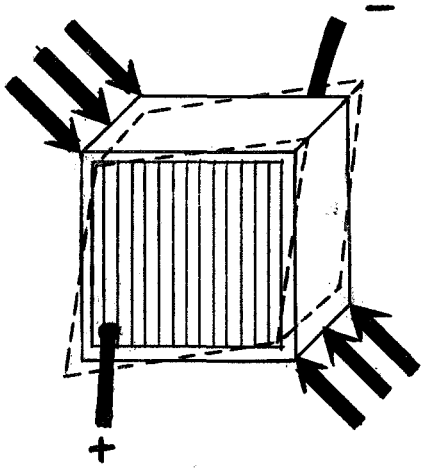


(c) Volume expander (V E)

(Hydraulic pressure p)



(d) Thickness shear (TS)



(e) Face shear (F S)

Figure 1.19. Deformation Modes of Single Piezoelectric Elements

then the equation of motion (a wave equation) becomes

$$\rho_m \frac{\partial^2 \xi}{\partial t^2} = \frac{1}{S_{11}} \frac{\partial^2 \xi}{\partial x^2} - \frac{d_{13}}{S_{11}} \frac{\partial E_3}{\partial x} \qquad (1.63)$$

This is the type of equation which almost invariably arises the considering the vibrational modes of a piezoelectric crystal. The approach here is to solve Equation (1.63) in the sinusoidal steady state to obtain ε_1. The displacement vector D_3 can then be found from Equation (1.58) since the electric field intensity is known. Next, the current in the material can be found from

$$I_3 = j\omega \int_0^{\ell} D_3 \, a \, dx \qquad (1.64)$$

where $j = \sqrt{-1}$ and ω is the radian frequency of E_3. This assumes that the current is totally a displacement current. However, losses can be included in the model if desired [27]. For the example problem

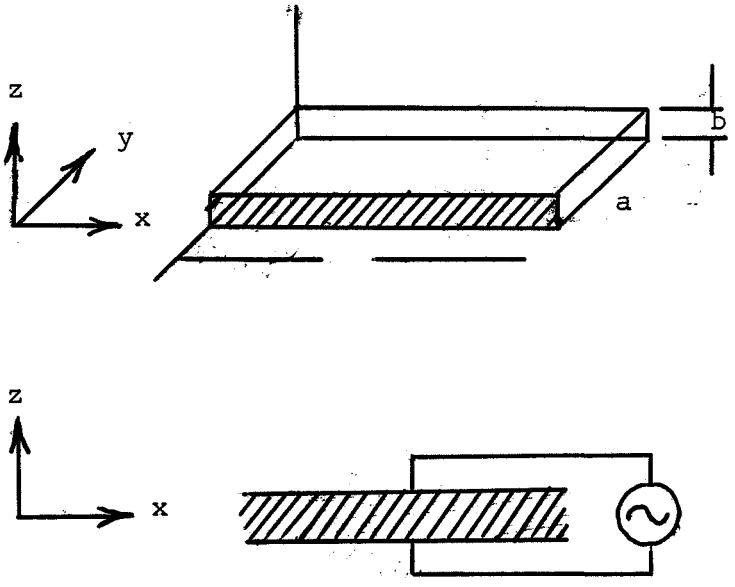


Figure 1.20 A thin piezoelectric crystal under an alternating electric field.

stated here, the result is

$$I_3 = j\omega[(\varepsilon_{P_{33}} - \frac{d_{13}^2}{s_{11}})\, a\, l + \frac{d_{13}^2}{s_{11}} \frac{2a}{k} \tan(\frac{k\ell}{2})\,]E_3 \qquad (1.65)$$

Here $k = \omega/v$ where $v = (\sqrt{\rho_m\, S_{11}})^{-1}$ is the velocity of sound propogation in the x direction of the crystal. From this equation an admittance function can be found, and an electrical equivalent circuit along with frequency variation of admittance. There is one pole (resonant frequency) and one zero (anti-resonant frequency) of the admittance as seen from Equation (1.65). These frequencies satisfy the following equation

$$f_r = \frac{1}{2\ell\sqrt{\rho_m\, S_{11}}} \qquad (1.66)$$

and

$$\frac{1 - K^2}{K^2} = \frac{v}{\pi\, f_a\, \ell} \tan(\frac{\pi\, f_a\, \ell}{v})$$

where $K = \dfrac{d_{13}}{\sqrt{\varepsilon_{P_{33}}\, S_{11}}}$ is quite analogous to the k-co-efficient.

These resonant frequencies are determine solely by the material parameters of the piezoelectric crystal. Figure 1.21 shows the equivalent circuit which would result if losses due to internal mechanisms were included. The loss mechanism allows one to define a quality factor for the circuit at anti-resonance as

$$Q = \frac{\omega L_1}{R_1} \qquad (1.67)$$

For piezoelectric crystals Q's as high as 10^5 are common. An excellent detailed discussion of equivalent circuits for other modes of operation of piezoelectric crystals is contained in Reference 28. This reference discusses mechanical as well as electrical equivalent circuit. The equivalent circuit just discussed is an unloaded case where the applied stresses are zero. Mason [24] discusses both unloaded and loaded equivalent circuits.

Probably the most obvious application of the piezoelectric effect to sensors is in the area of strain and pressure measurements. Indeed quartz pressure sensors have proven to be versatile instruments, featuring high linearity and repeatability over a wide range of pressures from cryogenic temperatures to temperatures near 500°K [29]. The high natural frequencies of piezoelectric materials allow measurement of high frequency pressure variations and pulsed pressures in explosions and in internal combustion engines. Piezoelectric pressure sensors also can be used in a differential mode to measure very low pressures.

Generally, piezoelectric strain and pressure sensors are restricted to dynamic measurements due to finite leakage resistance which causes the strain-induced charge to decay with time. However, through the use of electrostatic charge amplifiers, whose time constants can be as long as several days, near static measurements can be made [29].

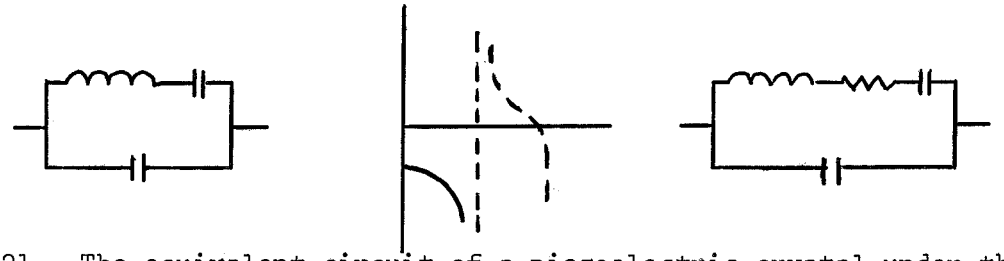


Figure 1.21. The equivalent circuit of a piezoelectric crystal under the applied voltage condition of Figure 1.20 and also in the longitudinal mode of vibration.

The sensitivity of piezoelectric pressure and strain sensors can be arrived at by considering the open circuit output voltage once a given geometry and mode of operation is chosen [30]. For example, for a piezoelectric hollow cylinder used as a pressure transducer with no internal pressure, the sensitivity is usually given in terms of $V/p_e d_o$, where V is the output voltage, p_e is the external pressure, and d_o is the outer diameter of the cylinder. This sensitivity is a function of the piezoelectric g-coefficients and the thickness-to-diameter ration [30]. A typical value for a barrium titanate sensor with thickness-to-diameter ratio of 0.1 is about 0.35 volts/(p.s.i.inch). This is a very large voltage sensitivity for reasonable pressures and dimensions. Another type of pressure sensor used as a blast pressure gauge is simply a piezoelectric cylinder mounted in a well-designed mechanical housing. The particular type of sensor has a charge sensitivity of around 100 picocoulombs/p.s.i, and a voltage sensitivity of about .4 volts/p.s.i. Piezoelectric materials can be used as strain gauges in which a thin slab operated in the longitudinal expander mode is bonded to a test vehicle, and when the gauge is strained a charge is generated across the output terminals. The output voltage is determined primarily by the h-coefficients and this type of gauge gives extremely large output voltages. For example, for a strain of 10^{-6} inch/inch, a quartz strain would give an output voltage of several volts while barium titanate gives a voltage of few tenths of a volt. These outputs are several hundred times larger than those from a semiconductor gauge at the same strain level. While basically only suitable for dynamic measurements, the use of piezoelectric strain gauges with charge sensitive amplifiers should allow almost static measurements. This type of arrangement offers much more sensitivity for low level strain measurements than other types of strain gauges.

Another application of the piezoelectric effect in piezoelectric ceramic materials is their use as ceramic transformers and filters. Figure 1.22

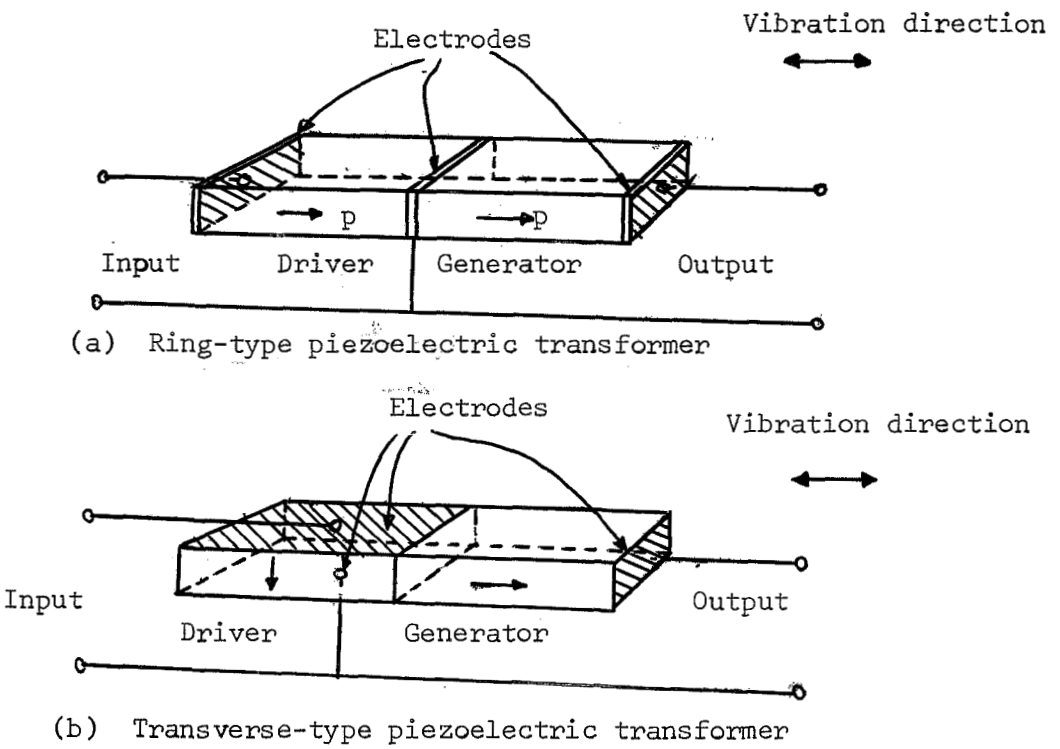


Figure 1.22 Piezoelectric Transformers.

shows how a piezoelectric crystal can be used as a transformer or a filter. An electric field applied to the left-half of the bar (the driver section) causes a mechanical vibraiton of the complete bar through the converse piezoelectric effect. As has been seen the strain distributions are standing waves resonant modes which occur at specific frequencies of the driving voltage source. These strain distributions in the right half of the bar (the generator section) produce an electric field at the output through the direct piezoelectric effect. A voltage step-up (or a voltage amplification) can be obtained in this device and the gain can be quite large. For example,Figure 1.23 shows a comparison between the experimental and theoretical voltage gain versus load resistance for a barium titanate ceramic transformer. This particular device had a Q > 400, and a maximum

efficiency of about 80% with a load of 0.5 Mohm.

The ceramic transformer can obtain large gains and efficiencies only near mechanical resonance. This restriction of energy transfer from input to output terminals to narrow frequency ranges make the ceramic transformer appear to have application as an electrical filter. Figure 1.24 shows the frequency response of a ceramic filter operating near resonance. The ceramic transformer is best suited for high-voltage low-current operation, and has found application for use with cathode ray tubes, Geiger-Muller tubes and photomultipliers. For a complete discussion of ceramic transformers as well as another good review of piezoelectric materials in general, the reader is referred to the book by Katz [31].

1.5.7 The Pyroelectric Effect

The pyroelectric effect is due to an electric polarization of a crystal when the temperature of the crystal is changed uniformly. Pyroelectric materials have a finite value of polarization, even when there are no applied electric fields or external stresses, so that the pyroelectric polarization is a spontaneous polarization in the sense of the definition given in Section 1.4. In this previous section it was illustrated that the spontaneous polarization of ferroelectric materials was highly temperature dependent, so that ferroelectric materials are expected to be pyroelectric. It is generally assumed that if a spontaneous polarization exists in a material, then that material which is not pyroelectric must have zero polarization for no applied electric field and stress. Thus all ferroelectric materials are pyroelectric and all pyroelectric materials are piezoelectric, but the converse of the statements is not necessarily true.

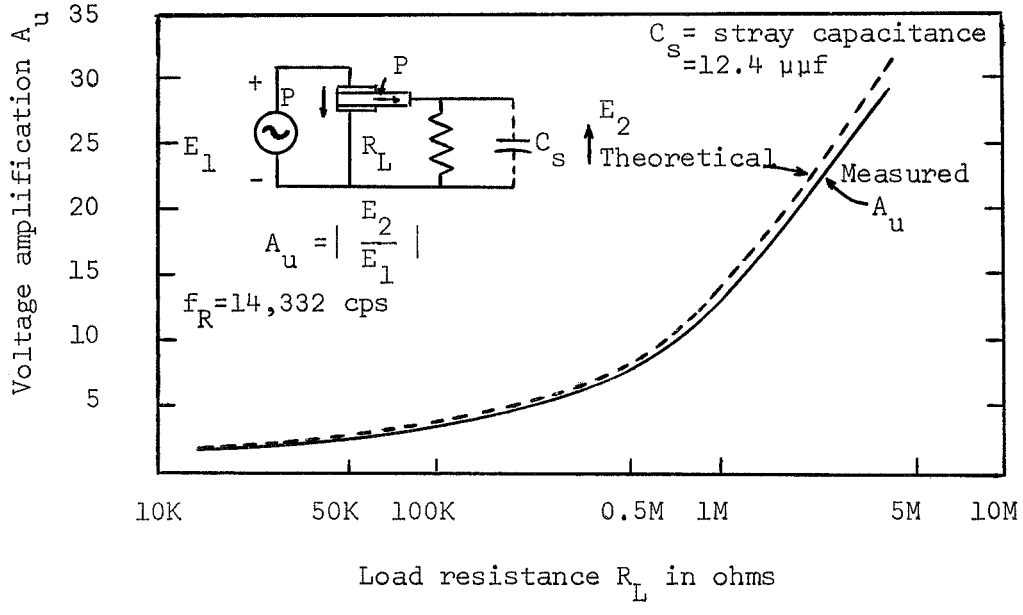


Figure 1.23. Relation between voltage amplifications and load resistance for transverse ceramic transformer with constant voltage drive=0.09 volt.

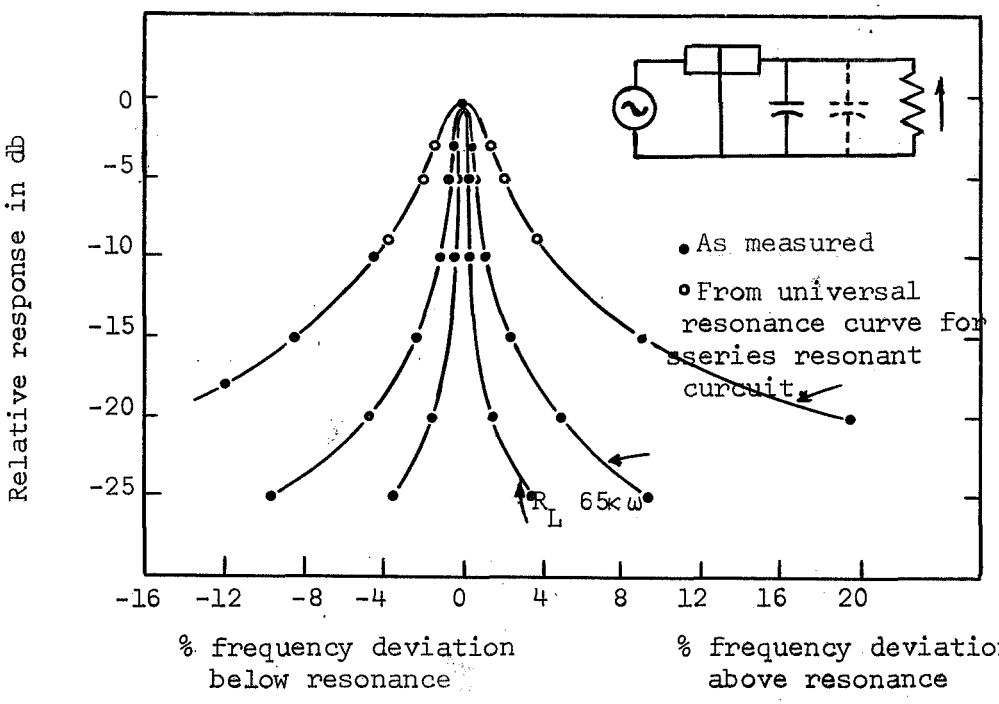


Figure 1.24. Relative response as a function of frequency diviation for a ring-type ceramic filter with various loads.

The pyroelectric effect has been useful in studying spontaneous polarization in ferroelectric materials, and the techniques used in these experiments offer the possibility of application to temperature and optical sensors. In these experiments a sample is heated by exposure to a strong light pulse, and a pyroelectric current results whose value is given by

$$i(t) = \frac{dP_s}{dt} = \frac{dP_s}{dT}\,\frac{dT}{dt} \tag{1.68}$$

or

$$i(t) = Q\,\frac{dT}{dt}$$

Here Q is called the pyroelectric coefficient and, in order to calculate Q from a measurement of i, it is necessary to know the heating rate, dT/dt. This is difficult to determine experimentally and for this reason the absolute value of Q is seldom given in the published literature. An idea of the magnitude of Q for triglycine sulfate (TGS) is shown in Figure 1.25 . This curve shows values of spontaneous polarization determined from hysteresis data and pyroelectric data, and the agreement is excellent. Values of Q for TGS range from about 4×10^{-9} coul/cm^2°C at -100°C to 2×10^{-7} coul/cm^2°C at + 50°C. The pyroelectric current in TGS would be very sensitive to temperature changes in the -50°C to +50°C temperature range. However, the current levels would be extremely low unless large heating rates could be used. Also, the pulsed light source necessary to induce sample heating would be a nuisance, and a temperature sensor using this type of pyroelectric mode of operation seems impractical. Another type of temperature sensor could be made using a pyroelectric material as a capacitor, and biasing it with a constant a-c current. The voltage across the capacitor would then be a function of temperature. The voltage sensitivity would be

$$\frac{dV}{dT} = \frac{I}{\omega C_o} \frac{d}{dT}\left(\frac{1}{\varepsilon_r}\right) \tag{1.69}$$

where V is the r.m.s value of the capacitor voltage, I the r.m.s. value of the current, ω the radian frequency, and ε_r the relative dielectric constant. Using typical values for TGS [17] yields voltage sensitivities of about 200 microvolt/deg, which is possibly sensitive enough to warrant its use as a temperature sensor. However the current

levels would be very small due to the large capacitive reactances which result from using ferroelectric materials, and this would not be advantageous.

While the pyroelectric effect has properties which are available for use as temperature sensors, these properties do not seem to be overly attractive and do not appear to offer any advantages over other devices which are available as temperature sensors (see section 3 of this report). However, the pyroelectric effect can be used to detect optical radiation, especially in the infrared portion of the spectrum [34]. Presently, there are commercial infrared photon detectors made from TGS which cover the region of wavelengths from 2 microns

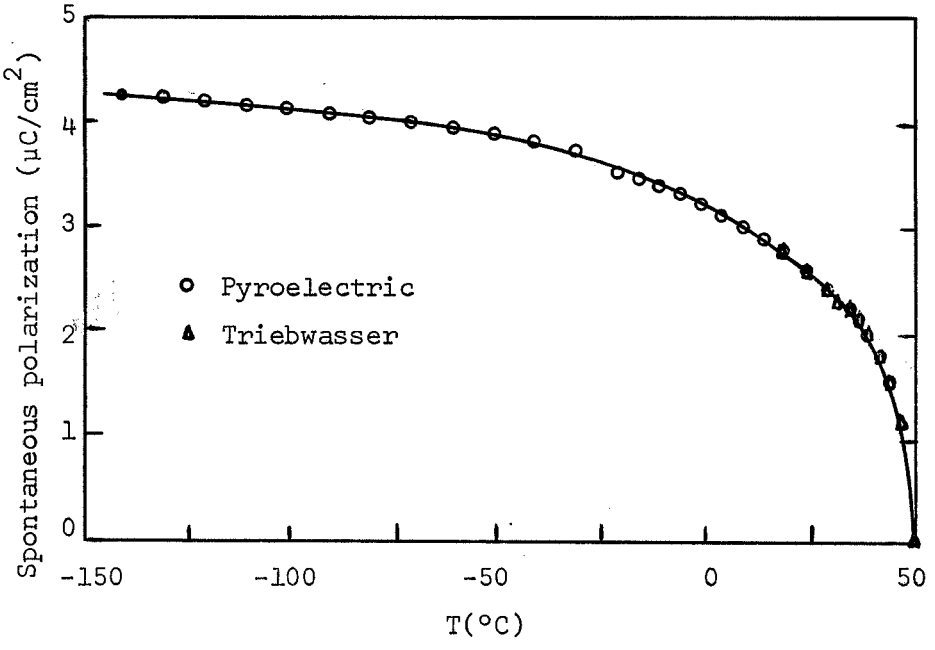


Figure 1.25. Temperature dependence of the spontaneous polarization of triglycine sulphate as deduced from pyroelectric data (circles) and from hysteresis data (triangles).

to 20 microns, with a response independent of wavelength. This wavelength range is limited strictly by the window material, and the TGS pyroelectric material is inherently uniformly sensitive to radiation from the ultraviolet ($\sim$. 1 microns) to beyond 300 microns. The pyroelectric detectors are about a factor of two more sensitive than the thermocouple type of detector (see section 1.5.9).

1.5.8 Magnetoresistance and The Hall Effect

If an external magnetic field is applied to a material which has a finite conductivity, it is observed that the resistance of the sample can change. For example, suppose an electric field is applied to a sample in the z direction. Then the resistance of the sample is the ratio of the voltage and current in the z direction. If a magnetic field is also applied in the z direction it is then found that the resistance of some metals and semiconductors are almost always higher than that obtained in the absence of the field. The fractional change in the resistance is called the longitudinal magnetoresistance. However, if the magnetic field is applied in the y direction the resistance will also change, but the fractional change will be different than for the magnetic field in the z direciton. This fractional change is called the transverse magnetoresistance. In addition to the transverse magnetoresistance, an electric field which is perpendicular to both the applied electric and magnetic fields will be generated in this case. This phenomena is called the Hall effect. For small magnetic fields the Hall effect is generally first order in the magnetic field while magnetoresistance is of second order. Both these phenomena can be generally classified under the name of galvanomagnetic effects.

The theory for magnetoresistance is based on the Boltzmann transport equation described in section 1.2. However, first a distinction should be made here. The results to be presented in this section assume the magnetic field to be a small perturbation on the zero magnetic field condition. This requirement can be shown to be applicable as long as the cyclotron frequency is much less than the reciprocal of the collision

time, or that the mean free path of the electron be much less than the extent of the cyclotron orbit [1]. This condition can be expressed as

$$\omega_c = q\ B/m^* << \frac{1}{\tau}$$

or

$$\frac{q\tau}{m^*}\ B << 1 \qquad (1.70)$$

or

$$\mu\ B << 1$$

where ω_c is the cyclotron frequency, B is the magnetic field intensity, m^* is the effective mass, τ the relaxation time and $\mu = q\tau/m^*$ is the mobility. While some simple cases can be treated ignoring the size of the field, these cases are not generally experimentally justified and likewise the first and second order nature of the Hall and magneto-resistance effect, respectfully, no longer hold.

If the force on an electron in an electric and magnetic field is given by the Lorentz equation

$$\overline{F} = g\ (\overline{E} + \overline{v} \times \overline{B}) \qquad (1.71)$$

where $\overline{v}$ is the electron velocity, the distribution function can be expressed as [27]

$$f = f_o + q\tau \frac{\partial f_o}{\partial \varepsilon} \cdot \frac{\bar{E}.\bar{v} + \frac{q\tau}{m^*} \bar{E}.(\bar{v} \times \bar{B})}{1 + (\frac{q\tau}{m^*} B)^2} \qquad (1.72)$$

If $\bar{B}$ is parallel to $\bar{E}$ then $\bar{E}.(\bar{v} \times \bar{B}) = 0$. Thus equation(1.72) predicts no longitudinal magnetoresistance. However, it has already been noted that the longitudinal magnetoresistance effect has been observed in some materials. This apparent discrepancy will be discussed later in this section, and it will be seen that, like the large piezoresistance effects observed in a many-valley semiconductor, the existence of a longitudinal magnetoresistance will depend on the band structure of the material.

Now for the specific case of $\bar{B}$ in the z direction and $\bar{E}$ taken to lie in the xy plane, the x and y components of the current can be expressed as [1]

$$J_x = K_1 E_x + K_1' E_y$$

$$J_y = K_1 E_y - K_1' E_x \qquad (1.73)$$

The only difference between metals and semiconductors in this equation is the value of the co-efficients K_1 and K_1'. Generally these co-efficients can be written as

$$K_1 = \frac{q^2 n_o}{m^*} \frac{\tau}{1 + (\omega_c \tau)^2}$$

$$K_1' = \frac{q^2 n_o}{m^*} \frac{\omega_c \tau^2}{1 + (\omega_c \tau)^2} \qquad (1.74)$$

Here n_o is the equilibrium electron concentration. The brackets imply an average over constant energy surfaces as defined in Equation (1.3). For metals the brackets are eliminated by simply evaluating the relaxation time at the Fermi energy. Now, to determine the transverse magnetoresistance and the Hall effect, the boundary condition $J_y = 0$ is used. This condition is easily realized experimentally. This condition gives

$$\frac{E_y}{E_x} = \frac{K_1'}{K_1} \tag{1.75}$$

The transverse magnetoresistance coefficient, R_m, and the Hall coefficient, R_H, are defined by

$$R_m = \frac{\Delta\rho}{\rho_o B_z^2}$$

$$R_H = \frac{E_y}{J_x B_z} \tag{1.76}$$

where $\Delta\rho/\rho_o$ is the fractional change of the resistivity. The theory presented so far predicts that for metals

$$R_m = 0$$

$$R_H = \frac{1}{q\, n_o} . \tag{1.77}$$

This is quite suprising since the theory at this point predicts that no magnetoresistance effect can be observed for any material (metal or semiconductor) where the relaxation time for all carriers is the same, i.e. for the case where the carriers do not have a velocity distribution [27]. For the case where there is a distribution of carrier velocities, the magnetoresistance and Hall coefficients become for $\omega_c \tau << 1$

$$R_m = \left[\frac{q \langle\tau\rangle}{m^*} \right] \frac{\langle\tau^3\rangle \langle\tau\rangle - \langle\tau^2\rangle^2}{\langle\tau\rangle^4}$$

$$R_H = \frac{1}{q\, n_o} \quad \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} \tag{1.78}$$

Thus from Equations (1.76) and (1.78) it can be seen that for the transverse magnetoresistance effect and the Hall effect the resistivity decreases as the square of the magnetic field intensity and the Hall field is directly proportional to the magnetic field intensity. Both these effects have been used as sensors of magnetic fields. Note that the variation of R_m with temperature should follow the variation of the square of the mobility with temperature, while the Hall coefficient should follow the variation of the inverse carrier concentration with temperature. It is obvious that these are important experimental tools as well as mechanisms with sensor applications.

To further explain the magnetoresistance effect different models are needed. First of all a two-band model will yield a finite transverse magnetoresistance, even if the relaxation times are constant, as long as the relaxation times in the bands are different from each

other [7]. For example, for a semiconductor will both holes and electrons, the transverse magnetoresistance coefficient becomes

$$R_m = \frac{n_o^2}{(n_o\, b + p_o)^2}\ \mu_p^2\, b(1 + b)^2 \tag{1.79}$$

which for an n-type semiconductor such that $n_o >> p_o$ gives

$$R_m = \frac{n_i^2}{n_o^2}\ \frac{(1 + b)^2}{b}\ \mu_p^2 \tag{1.80}$$

This magnetoresistance has a much stronger temlerature dependence than the squared mobility alone would have, and also is generally smaller than Equation (1.78) predicts.

To explain the more complicated magnetoresistance effects such as the longitudinal magnetoresistance and other effects which can occur in many-valley semiconductor, a much more complicated procedure is required. This procedure is too complicated to allow detailed discussion here, and the reader is referred to Reference 1, Chapters 5 and 8 for a complete discussion. The basic approach of this technique is to formally expand the distribution function in powers of the magnetic field and then use this expansion to evaluate the current density components. This procedure yields the following expression to second order in the magnetic field components for the current density components.

$$J_i = \sigma_{ij}\, E_j + \sigma_{ijk}\, E_j\, B_k + \sigma_{ijkl}\, E_j\, B_k\, B_l + \ldots \tag{1.81}$$

Here σ_{ij} is the zero-magnetic field conductivity tensor and σ_{ijk} and σ_{ikjl} are often called the magnetoconductivity tensor coefficients. The form of these co-efficients are quite similar to the conductivity tensor of Equation (1.5), except for the fact that they are much more complicated. This type of approach does give both a transverse and logitudinal magnetoresistance. It also can be applied to the many-valley semiconductors, and yields results which are comparable to theory. It also helps to explain the band structure of this type of semiconductor, and actually compliments piezoresistance measurements in this regard [20].

Another aspect of the theory of magnetoresistance which has found application to metals is Kohler's rule [6]. This rule states that the fractional change in resistivity due to a magnetic field is given by

$$\frac{\Delta\rho}{\rho_o} = f\left(\frac{B}{\rho_o}\right) \tag{1.82}$$

where f(x) is a function which depends on the orientation of B and E as well as on the orientation of the crystal axes, if the sample is a single crystal. The important point is that f is independent of temperature and impurity content for small impurity concentrations [1].

Figure 1.26 shows a plot of the transverse magnetoresistance of twenty two metals on a reduced Kohler plot, where the zero field resistivity is normalized to the resisitvity at the Debye temperature

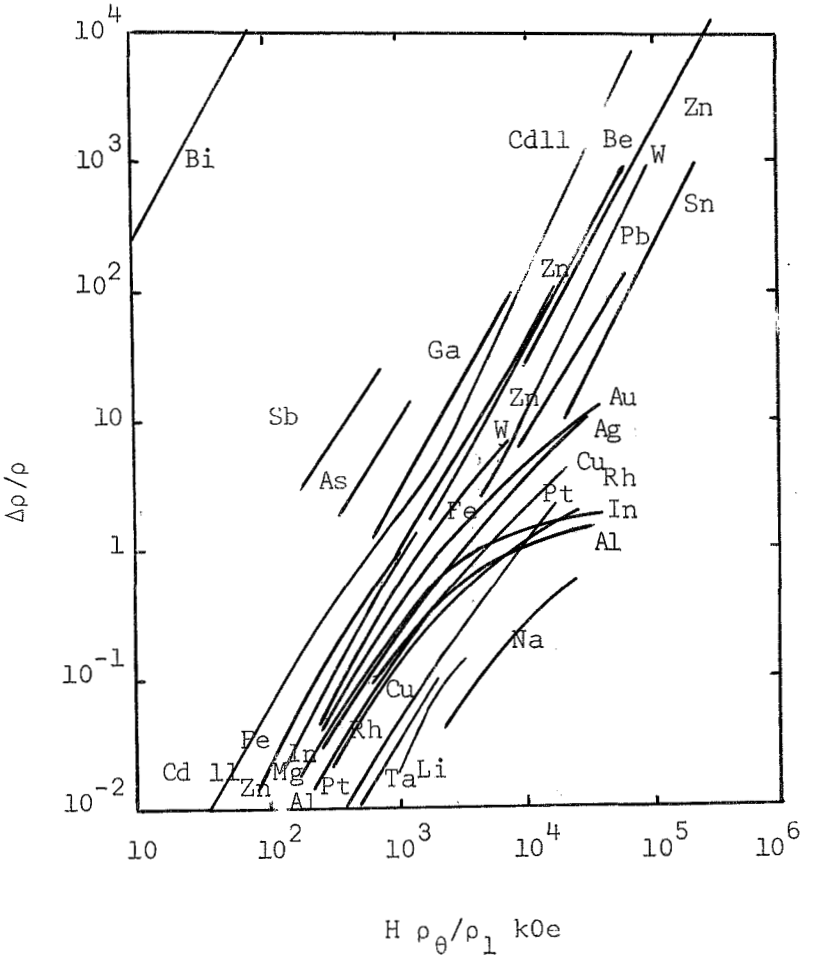


Figure 1.26. The transverse magnetoresistivity of 22 metals plotted on a reduced Kohler diagram

of the material. Notice that bismuth has an extremely large magnetoresistance. This material is very sensitive and has been used as a sensor of magnetic fields. Generally, magnetoresistance is smallest for metals of the simple alkali group and largest for metals of the bismuth group. Figure 1.27 shows both the longitudinal and transverse magnetoresistance as function of temperature for the n-type (11.5 ohm cm) semiconductor germanium. This figure illustrates that

above about 100°K lattice scattering is the dominant scattering mechanism. Since magnetoresistance varies as μ^2 and for lattice scattering μ varies as $T^{-3/2}$. Figure (1.28) shows the oscillatory behavior of magnetoresistance for n-type silicon as the angle between current and magnetic field is varied. It can be seen that for some orientations the transverse magnetoresistance is larger than the longitudinal magnetoresistance, while the opposite is true for other orientations.

In summary, the magnetoresistance and Hall effects are two important phenomena which can give electrical outputs which are proportional to magnetic field, and thus are useful as magnetic field sensors. Generally,

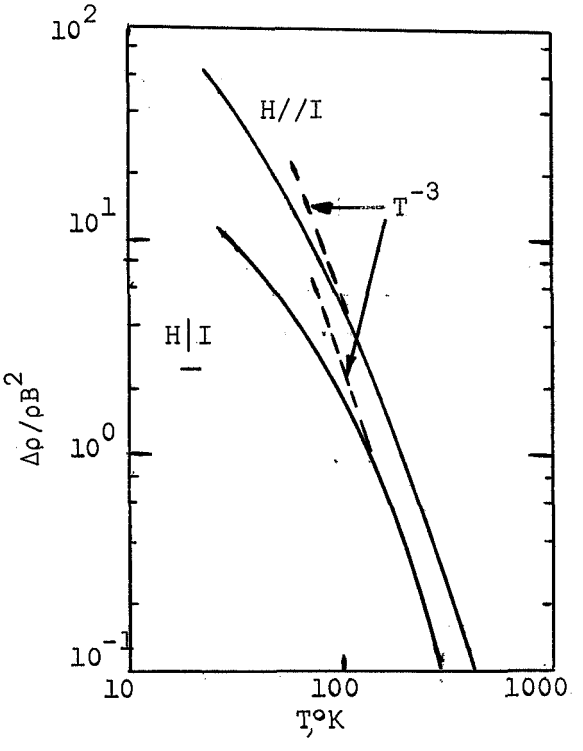


Figure 1.27. Magnetoresistive effect in Ge plotted as a function of temperature at B = 0.4 weber/m^2. The Ge sample is n-type and has a resistivity of 11.5 Ω-cm.

the outputs can be described by

$$E_H = E_a \ (\mu B)$$

$$\frac{\Delta\rho}{\rho_o} = C \ (\mu B)^2 \tag{1.83}$$

where E_H is the Hall field, and E_a is the applied field for the Hall effect. $\Delta\rho/\rho_o$ is the fractional change in resistance of a magnetoresistor and C is a constant which depends on the material used. Here $\mu B << 1$ for these equations to apply. For maximum sensitivities both the Hall effect and magnetoresistance require materials with large mobilities, which implies a small effective mass in the band structure. This likewise implies that semiconductors would be the most sensitive materials to use for such devices, and indeed this is the

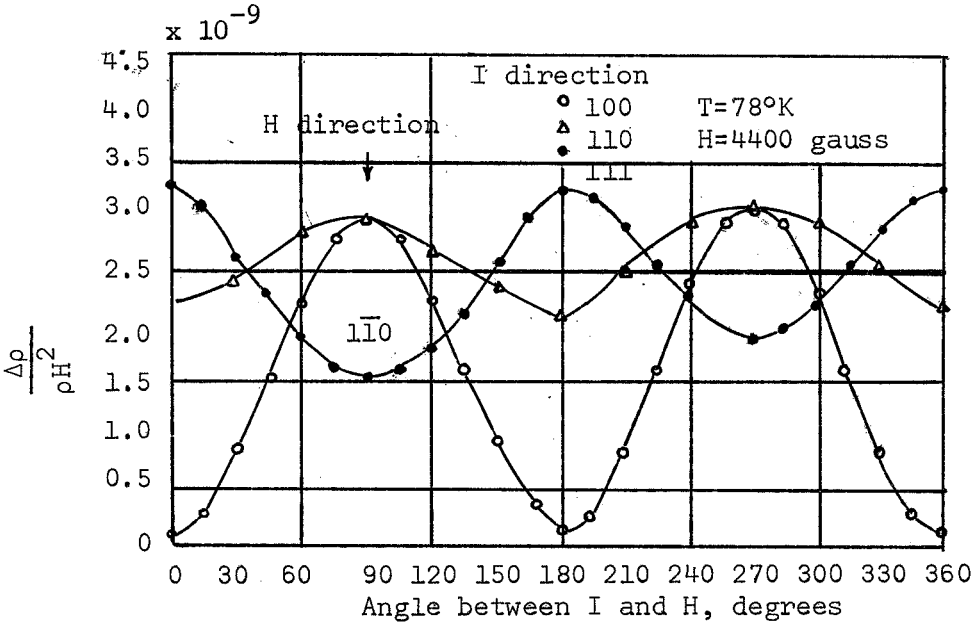


Figure 1.28. Variations in $\Delta\rho/\rho H^2$ as H is rotated with respect to I in n-type silicon. Certain cardinal directions of H are as indicated. The fixed directions of I are as given.

case. The two most commonly used materials for these devices are InSb and InAs.

Table 15 shows a comparison of the characteristics of two typical Hall generators constructed from InSb and InAs. While the areas of these two devices are not the same, and thus their sensitivities cannot be directly compared, one can still see that these materials are very sensitive to magnetic field. For the particular InAs device, a one gauss magnetic field will give an output of 50μv.

Table 15 Characteristics of Semiconductor Hall Generators

Material	Surface (mm^2)	Nominal drive current i_1 (mA)	Sensitivity	Residual Hall voltage (mV)	Hall voltage (%/°C)	Residual Hall voltage (μV/°C)
InSb	1	60	40 mV/Maxwell	20	-2	200
InAs	80	400	50 mV/kG	1.5	-0.07	1

Measurements of magnetic fields whose flux density is as low as 2×10^{-3} gauss are possible with Hall effect devices, and the theoretical limit is even lower. This theoretical limit is almost achievable over short periods of time if special precautions and techniques are utilized [32] and section 1. 6. 3. for further discussion). In table 15

information is given on some of the thermal variations in Hall effect devices. It is desirable that the Hall voltage be independent of temperature, and this is probably one reason InAs seems to be preferred over InSb. From table 15 it is seen that the InAs has a much better temperature coefficient of Hall voltage than InSb. In table 15 numbers are listed for the temperature dependence of the residual Hall voltage. This term is due to the fact that a voltage can be observed across the Hall voltage terminals, even when there is no applied magnetic field, if a longitudinal current is flowing. This residual Hall voltage is due to Hall probe misalignment as well as a self-induced magnetic field due to the current flow in the longitudinal direction. It introduces non-linear terms into the Hall effect. For example,it can cause the Hall voltage versus magnetic flux density to deviate from a linear relation. The fact that the residual Hall voltage is a function of temperature further complicates this problem, so that the temperature coefficient of the residual Hall voltage is of interest. Again InAs performs much better than InSb in this respect. Table 16 shows the temperature dependence of the Hall coefficient, R_H, and the input resistance of the Hall probes, R_{11}, for several different materials, including an alloy on InAsP. Also shown is data taken on thin films for comparison purposes.

Simple methods are available for the compensation of the temperature variations in Hall generators. For example, one simple technique uses a nickel resistor with a positive temperature coefficient in parallel with the Hall generator, which offsets the negative temperature

Table 16 Hall Effect Properties of Several Semiconductors.

Material	R_H (cm^3/A-sec)	$\frac{dR_H}{dT} \times \frac{1}{R_H}$ (%)	$\frac{dR_{11}}{dT} \times \frac{1}{R_{11}}$ (%)	Temp. range (°C)	Method of Preparation
InSb	-400	-2.2	-1.8	0-50	crystals
InAs	-100	-0.07 to -0.01	+0.2	0-100	
$InAs_{0.8}P_{0.2}$	-200	-0.04	+0.2	0-100	grown from the melt
InSb		-1.5	-1.5	0-50	evaporated
InAs		-0.1	+0.1	0-100	films

coefficient of the device. Other techniques are discussed in Reference 45. Also, Hall generators do have a very slight aging characteristic, unless the devices are properly packaged. With suitable packaging, such as epoxy potting, the output characteristics of Hall generators vary with time by no more than experimental measurement error, and the residual Hall voltage changes less than the Hall voltage induced by a magnetic field of 0.1 gauss over a period of eight months [32].

The magnetoresistance effect is used in a practical device called a field-plate to measure magnetic fields. As has been mentioned, the metal bismuth was used in early practical applications to measure strong magnetic fields. As far as the measurement of magnetic fields are concerned, the magnetoresistance effect is not as attractive as the

Hall effect because of its second order dependence of magnetic field. However, the magnetoresistance effect does have some advantage over the Hall effect. Probably the most significant is that a field plate is an two-termanal device. Thus, it may be substitued for conventional resistors in electronic circuits. This is an important practical advantage. Recently, magneto-resistance field plates have been suggested for use as potentiometers and volume controls [33]. It is proposed that these devices are constructed through application of integrated circuits technology, and thus should be compatible with integrated circuits to some extent. Another advantage of field plates, which is related to the previously mentioned advantage, is that their zero magnetic field resistance can be chosen arbitrarily over a range of several orders of magnitude. This also offers advantages for circuit applications.

Semiconductor field plates have a sensitivity which is greater than that of bismuth devices by as much as two orders of magnitude [32]. Figure 1.29 shows a plot of the ratio of the resistances with and without magnetic fields, R_B/R_o, as a function of magnetic flux density for several important magnetoresistance materials. The notation on the curves is explained in the figure caption. The curves are normalized to each other at zero magnetic field. The raster plate referred to in Figure 1.28 is an arrangement which is used to increase the total resistance of a magnetoresistance device for use of materials which have a low resistivity Reference 32. Figure 1.30 shows the temperature coefficient of resistance at room temperature as a function of magnetic flux density for the same materials listed in Figure 1.29 . Generally

the temperature coefficient of magnetoresistance increases at lower temperatures due to the strong dependence on electron mobility.

Finally, the choice of field plates over Hall effect generators is usually made through economic considerations and through simplicity of circuit applications. While this section has discussed some of the major considerations of these two devices, a much more elaborate discussion of problems and applications of magnetoresistance devices is contained in the recent and excellent book by Weiss [32].

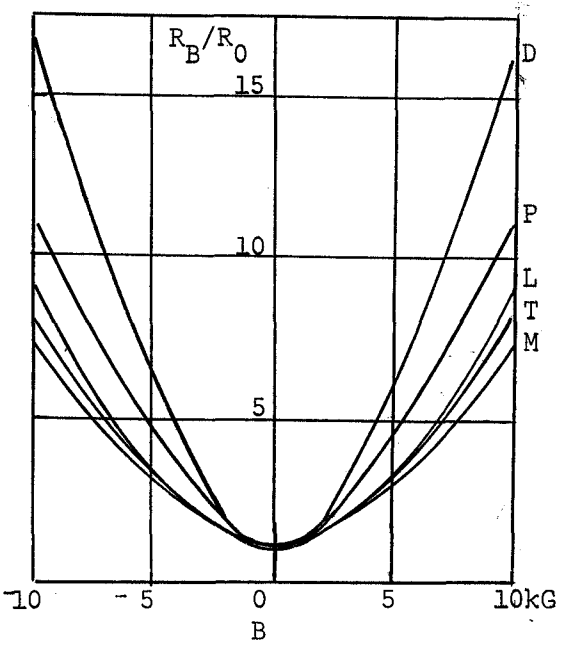


Figure 1.29. Relative resistance $R_B R_0$ of field-plates as a function of the magnetic induction B. D, intrinsic InSb-NiSb. L, M, doped InSb-NiSb. P, T, raster plates with doped InSb.

1.5.9 Photoconductivity in Solids

When light is incident on a solid, the electrical conductivity of a sample can be increased if the light is absorbed. The process is called photoconductivity and it has many practical applications in such areas as television cameras, light meters, and in the photographic process. However, its most important application is probably in the sensor area where bulk solid state materials can be used as detectors of light in specific regions of the optical spectrum.

The largest photoconductive effects occur in insulators and semiconductors, since for these materials the background (dark) electron and hole concentrations can be made much smaller than the photogenerated concentrations. Photoconductive effects are not important in metals and will not be discussed here.

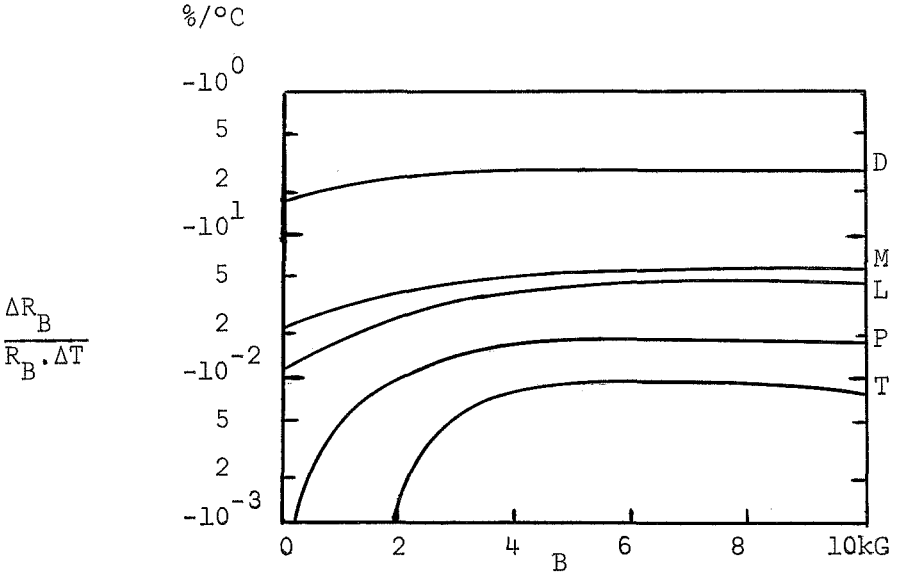


Figure 1.30. Temperature coefficient $\Delta R_B/R_B.\Delta T$ of field-plates at room temperature as a function of the magnetic induction B. For symbols identifying curves see Figure 1.29.

The fundamental absorption of light (as distinguished from absorption of light due to free carriers) is a process which generates hole-electron pairs which in turn increase the photoconductivity. Because of this fact one would expect the photoconductive threshold (the wavelength at which photoconductivity occurs) to be the same as the "fundamental absorption edge" (the wavelength at which a solid begins to strongly absorb light). This is generally true, although excitons (bound hole-electron pairs) can influence the absorption edge at low temperatures [1]. The absorption edge occurs at wavelengths which correspond approximately to the energy band gap of the solid. The energy band gap of most semiconductors fall in the range of about .16 eV to 3.0 eV which corresponds to wavelengths between about .42 microns to 7.8 microns. These values cover the range from below the visible to well into the infrared region of the spectrum. Other means are available which extend the response of photoconductors to higher wavelengths, so that a wide range of wavelengths can be detected with bulk photodetectors.

Generally, bulk photodetectors can be classified as either instrinsic or extrinsic detectors. For an intrinsic detector the photoconductivity is due to electronic transitions between the conduction and valence bands, while for an extrinsic photodetector the photoconductivity results an excitation of either holes or electrons from energy levels in the forbidden gap which are due to impurities. Since the impurity levels lie in the forbidden gap the excitation energy of such an impurity is less than the band gap. Thus the extrinsic photoconductivity in

a given material will occur at longer wavelengths than the intrinsic photoconductivity. In this way the response of bulk semiconductor photoconductors can be extended into the far infrared. It will be seen later that efficient photoconductivity has been achieved in semiconductors at wavelengths longer than 40 microns.

Now, the following simple model will be used to illustrate the salient features of photoconductivity. While it will not apply in general to extrinsic photoconductors this model will point out several features which are important in determining the response of both extrinsic and intrinsic photoconductors. Figure 1.31 shows an arrangement which could be used to observe photoconductivity. A sample of length ℓ, width W, and thickness d is illuminated over its cross-sectional area $W\ell$. Electrodes are attached to the surface ℓ d and the current through the sample is monitored. The current which results due to the finite resistance of the sample is called the dark current and the current due to the light is called the photocurrent. A general expression for the fractional change of the conductivity due to light can

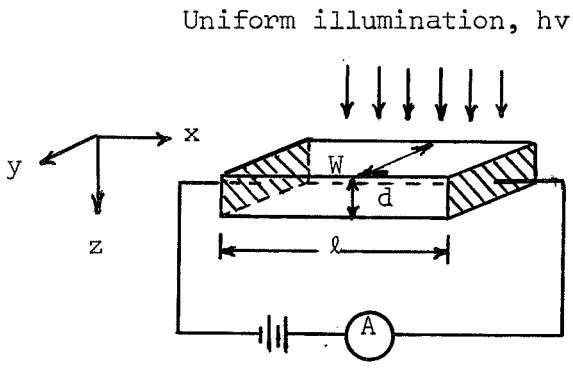


Figure 1.31. Basic arrangement for observation of photoconduction.

be written as

$$\frac{\Delta\sigma}{\sigma_o} = \frac{\mu_n \Delta n + \mu_p \Delta_p}{\mu_n n_o + \mu_p p_o} \simeq \frac{(1 + b)}{p_o + b n_o} \Delta n \qquad (1.84)$$

Here $b = \mu_n/\mu_p$ and Δn and Δ_p are the changes in electron and hole concentrations due to light. For convenience it has been assumed that $\Delta n = \Delta p$ which neglects any trapping effects. Trapping is usually of importance in photoconductivity and this point will be discussed later.

Now, the number of photons per unit are a which penetrate the sample can be written as

$$N(z) = N_o (1 - R) e^{-\alpha z} \qquad (1.85)$$

where N_o is the density of incident photons, R is the surface reflection coefficient, and α is the absorption coefficient usually given in units of cm^{-1}. The energy per unit area, quite often called the incident intensity, is thus given by

$$W_i = \frac{N(z)}{h} = \frac{N_o}{h\nu} (1 - R) e^{-\alpha z} \qquad (1.86)$$

where $h\nu$ is the energy per incident photon. The number of photons per unit time which are absorbed in an incremental thickness dz is then given by

$$\#\text{photons/time} = (W\ell)\,[N(z) - N(z + dz)]$$

$$\simeq \alpha W\ell\, N(z)\, dz \qquad (1.87)$$

Thus the hole-electron pair generation rate (the number of hole-electron pairs generated per unit volume per unit time) becomes

$$G = \alpha N(z) \eta \tag{1.88}$$

where η is called the internal quantum efficiency, that is the number of hole electron pairs generated per incident absorbed photon. Generally, for wavelengths near the fundamental absorption edge η can be taken as unity. Since the generation rate is a function of position the hole and electron concentration will also depend on position and the continuity equations govern this spatial dependence. Thus

$$D_p \frac{d^2(\Delta p)}{dz^2} - \frac{\Delta p}{\tau_p} = - G(z) \tag{1.89}$$

Here D_p is the diffusion constant for holes and τ_p is the hole lifetime. This is a simple differential equation which can be solved for $\Delta p(z)$, with appropriate boundary conditions. The photocurrent can then be found from the relation

$$I_p = q W (1 + b) \mu_p E \int_0^d \Delta p (z) dz \tag{1.90}$$

Here E is the electric field and is simply $E = V_a/\ell$, where V_a is the applied voltage. For convenience, this integral will be evaluated by letting $\ell \to \infty$. This corresponds to the case where $\alpha\ell >> 1$; the case for $\alpha\ell << 1$ can be treated by simply noting for $\alpha\ell << 1$, $\Delta p(z)$ is approximately constant. For this case

$$\Delta p \simeq G \tau_p = \alpha\eta N_o (1-R) \tau_p$$

and $\alpha\ell \ll 1$ (1.91)

$$I_p = q\, W(1 + b)\, \mu_p\, \alpha\eta\, N_o\, (1 - R)\, \tau_p\, E$$

Now, Equation (1.90) can be evaluated as $\ell \to \infty$. This gives for the photocurrent, I_p,

$$I_p = \frac{q\eta W\ell(1 + b)\mu_p \tau_p\, (1 - R)E}{(\ell + S\, \tau_p)} \cdot \left[1 + \frac{S\tau_p}{\ell(\, 1 + \alpha\ell)}\right] \tag{1.92}$$

Here S is a parameter called the surface recombination velocity and is defined by the boundary condition

$$D_p \left.\frac{d\ (\Delta p)}{dz}\right|_{z=0} = S\ \Delta p(\ z=0) \tag{1.93}$$

This parameter reduces the lifetime of carriers near the surface [7] and is a physically meaningful parameter. The larger the surface recombination velocity, the smaller the photo-generated carrier concentrations will be at the surface. Figure 1.32 shows a plot of photocurrent versus wavelength from Equation (1.92). For S=0 the photocurrent monotonically increases with decreasing wavelength. For larger values of S there will be a maximum photocurrent which occurs approximately at the wavelength where $\alpha\ell \simeq 1$. In Figure 1.32 , the photocurrent is shown for a linear monotonically decreasing absorption coefficient versus wavelength for comparison purposes. Figure 1.33 illustrates typical experimental photoconductivity spectra for ten semiconductor materials. This figure vividly depicts the wide wavelength range for which semiconductors are photosensitive.

In this figure all of the response curves have been normalized to the same maximum value. In reality the maximum photocurrents for these materials will vary considerably.

This simple model presented in Equation (1.92) predicts the basic behavior of an ideal photoconductor in terms of its photo-current response to light of varying wavelength. As mentioned before, this treatment has neglected the presence of trapping levels which can greatly influence the photoconductivity of all materials. Generally, a trap is a center which has a large capture cross section for a carrier of one type,and a correspondingly low cross section for carriers of other types. This type of center tends to increase the lifetime of the carrier for which the capture cross

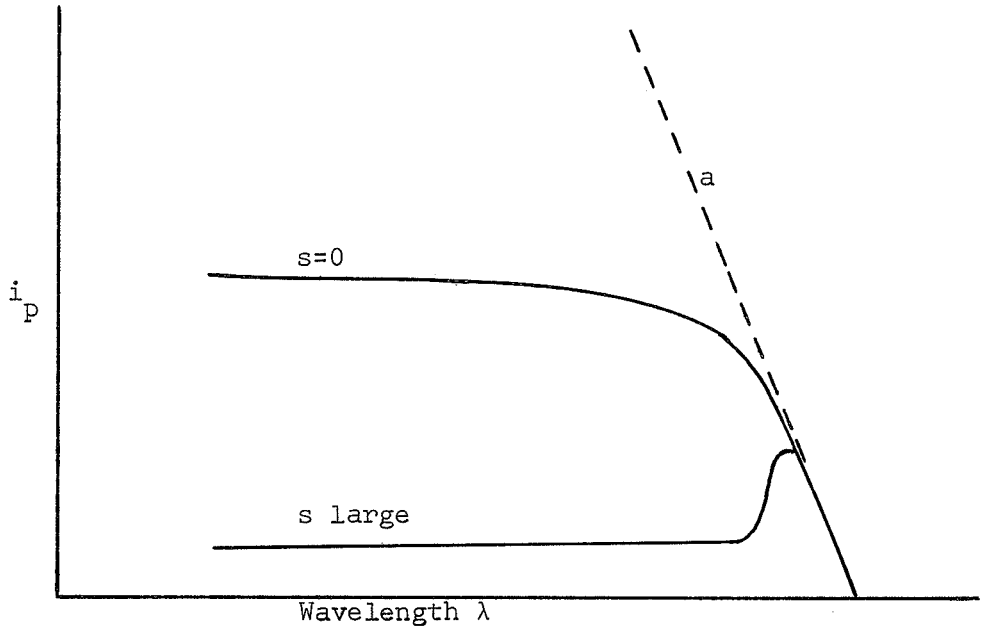


Figure 1.32. Variation of photo-current with wavelength for large and small values of the surface recombination velocity s.

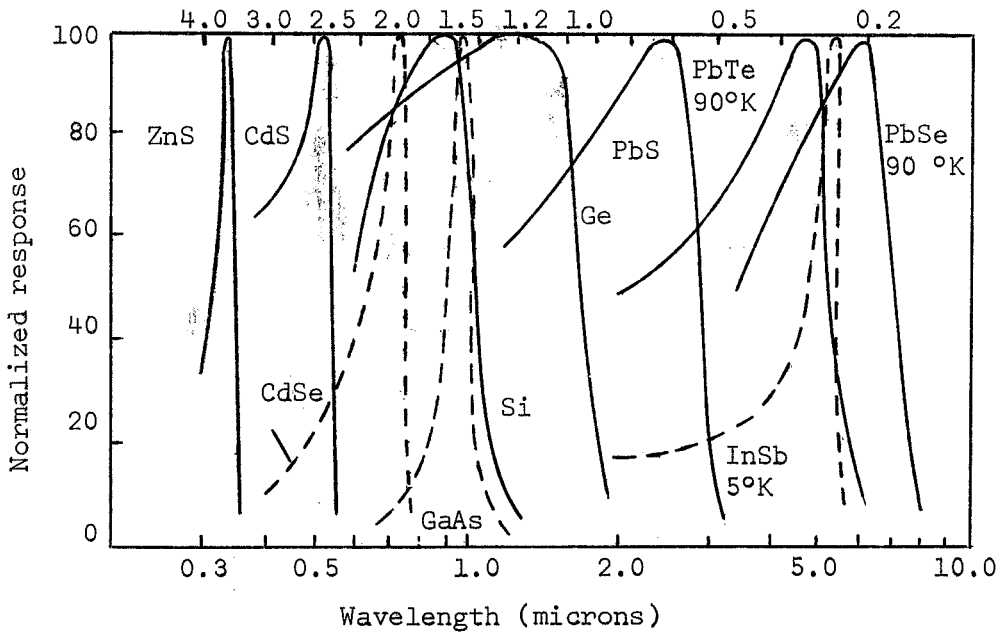


Figure 1.33. Spectral response of typical "intrinsic" photoconductors.

section is small. Thus such traps are expected to increase the photocurrent. However, this increase in photocurrent is obtained only with the sacrifice of response time of the photoconductor to abrupt changes in light intensity or wavelength. Trapping effects will be discussed in more detail in a later section.

1.5.9.1 Processes Involved in Photoconductivity

As has been mentioned, the basic electronic transitions taking place in the photoconductive process involve both band-to-band transitions as well as transitions from defect levels in the forbidden energy gap to either the conduction or valence band. All of the important aspects of photoconductivity are determined to a large extent by these transitions. These aspects, among which are sensitivity, spectral response, speed of response, and dark

conductivity will be discussed in the next sections. This section will be concerned with the electronic transitions which are important in photoconductors. A good basic discussion has been given in Reference 34, and will be followed here.

Figure 1.34 shows the basic transitions which can occur in a photoconductor. These processes are:

1) Creation of hole-electron pairs by the absorption of light greater than or equal to bandgap.

2) Excitation of an electron which is bound to a defect whose binding energy lies in the forbidden gap. Generally the binding energy must lie below the electron quasi-Fermi level for this process to be important. Usually, this implies a deep-lying defect level.

3) The capture of a photoexcited hole by a deep-lying defect. The defect should be occupied by an electron for this process to be important.

4) The capture of a free electron by a deep center which has previously captured a free hole. This process is called recombination and largely controls the lifetime of the photoexcited carriers.

5) The capture of a photoexcited electron by a shallow defect, which is usually called an electron trap. This center usually has a very low probability for capturing a free hole.

6) The thermal excitation of an electron trapped in a shallow defect.

7) The optical excitation of an electron trapped by a shallow level.

8) The optical excitation of a trapped hole.

9) The thermal excitation of a trapped hole.

Transitions 1 and 2 determine the variation of the photocurrent versus the wavelength of excitation, i.e. the spectral response. Transitions 3 and 4 together specify the electron and hole lifetimes, which in

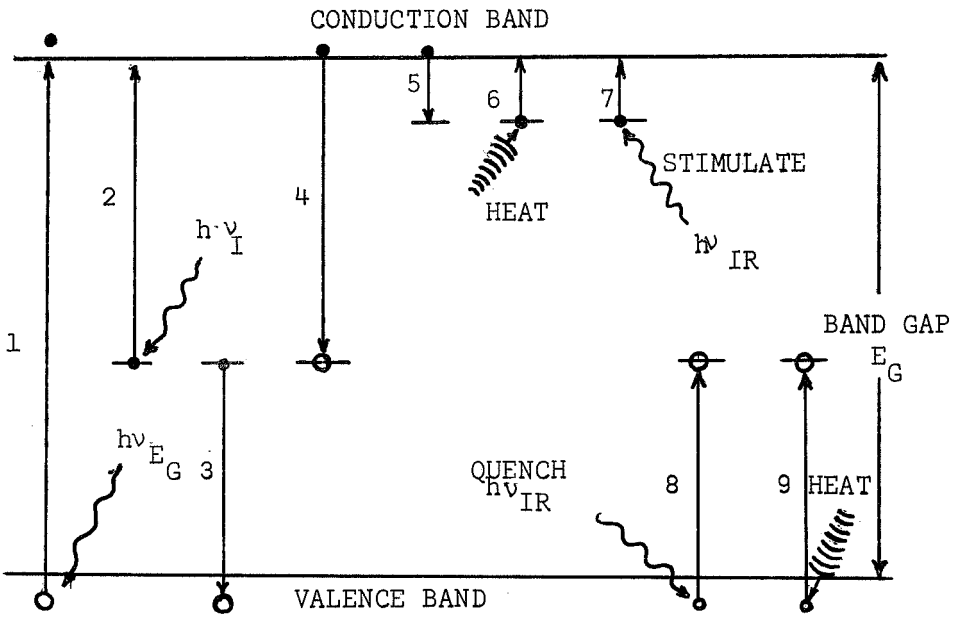


Figure 1.34. Typical electronic transitions in an n-type photoconductor.

turn determine the photosensitivity of the material. Transistions 5 and 6 determines the speed of the photoconductor to changes in light intensity. Transition 7 can lead to increased photoconductivity and along with transition 2, determine what is known as extrinsic or impurity photoconductivity. These processes increase the wavelength range to which a given material is sensitive. Finally, transitions 8 and 9 lead to what is known as optical and thermal quenching of photoconductivity, when the centers involved have a small probability for capturing an electron. These processes will not be discussed here and the reader is referred to the review by Bube [35]. For a good discussion of the mathematics of the transition processes, the reader is referred to [42].

1.5.9.2 Other Aspects of Photoconductivity

a) The importance of Ohmic Contacts - The types of electrodes

used to apply the electric field to a photoconductor is of utmost importance to its performance. Generally it is required that the contacts be "ohmic", so that space charge neutrality is maintained throughout the crystal while in the presence of illumination. This insures that the photocurrent is dominated by recombination mechanisms in the bulk of the material and not by electrode injection effects, nor by the formation of depletion regions near the contacts, which reduces the carrier concentrations and changes the generation and recombination rates in these areas.

Generally, ohmic contacts can be formed by diffusing a metal into the semiconductor where the contacts are desired. For n-type material a metal which acts as a donor offers the greatest opportunity for success, while in a p-type material a metal which acts as an acceptor should be employed. These diffusions form what are known as high-low junctions near the contact areas. As long as the applied voltage is low this type of junction will exhibit ohmic behavior. In a photoconductivity experiment one generally desires to eliminate the possibility of contact effects by not illuminating the contact area.

b) Photoconductivity Gain (Sensitivity) - For a given generation rate the photoconductivity of a material is determined largely by the mobility-lifetime product, as can be seen in Equation (1.91). Another useful parameter which also characterizes photoconductors is the photoconductivity gain, g. This is defined as the ratio of the number of charge carriers which pass between the electrodes per second to the number of photons absorbed by the material per second [32]. Thus, the relation for g is

$$g = \frac{I_p/q}{G} \tag{1.94}$$

The gain can also be related to the lifetime and transit times of the carriers by

$$g = \frac{\tau_n}{t_n} + \frac{\tau_p}{t_p} \tag{1.95}$$

where τ_n and τ_p are the electron and hole lifetimes and t_n and t_p are the respective transit times. The transit times are related to the length of the sample and the carrier velocities. That is

$$t_n = \ell/\mu_n E$$
$$t_p = \ell/\mu_p E \tag{1.96}$$

Now the photoconductivity gain can be written as

$$g = (\tau_n \mu_n + \tau_p \mu_p) \cdot V/\ell^2 \tag{1.97}$$

where V is the applied voltage. This is just another example which points out the importance of the lifetime - mobility product. Typical values of photoconductivity gain are in the $10^1 - 10^3$ range, with values as high as $10^4 - 10^5$ not uncommon [35].

The concept of photoconductive gain is one of several attempts at defining a sensitivity for photoconductors. Another useful definition of sensitivity is the ratio of light to dark conductivity [34]. In infrared photoconductors it is often useful to

define sensitivity in terms of the amount of light power required to excite a photocurrent which is just equal to the noise. The topic of noise will be discussed in a later section. Another type of sensitivity is called the specific sensitivity and equals the product of the conductance and the square of the electrode spacing divided by the absorbed light power. Thus the specific sensitivity, is, is

$$s = \frac{\left(\frac{I_p}{V}\right) \ell^2}{P} = \frac{q\mu_n (1 + b) \Delta n}{p} \qquad (1.98)$$

where P is the total light power absorbed, and p is the total light power per unit volume absorbed. Table 17 gives some typical values of specific sensitivity for several different types of cadmium sulfide, probably the most widely studied photoconductor. The wide range of values depicted in this table illustrate the two important points that 1) a wide range of sensitivities are available in a given material, and 2) impurities have a very drastic effect on the photosensitivity of a given material. The topic of response time shown in this table will be discussed in a later section.

c) The Detectivity of Infrared Photoconductors - As stated in the last section the sensitivity of infrared photoconductors is usually quoted in a different manner than for other wavelength-sensitive materials. All the definitions of sensitivity previously mentioned are valid independent of the spectral response of a material. However, in the infrared region of the spectra one of the most important limiting factors in the sensitivity of a given photoconductor is the thermal fluctuations present in the background radiation incident

Table 17 Properties of various Cadmium Sulfide Photoconducting Materials

Material	Specific Sensitivity, cm^2/ohm watt	Electron Lifetime, sec.[a]	Response Time at 10^{-4} ft-c, sec
Impure CdS (Contains recombination centers such as Ni	10^{-9}	10^{-10}	---
Pure CdS (No spectrographically detectable impurities)	10^{-5}	10^{-6}	---
Sensitive CdS (Contains deliberately incorporated impurities such as Cl and Cu)	10^{-1}	10^{-2}	10
Special Sensitive CdS (Contains only traces of iodine impurity)	10^{-1}	10^{-2}	10^{-1}

[a]Also the response time for very high light level excitation.

on the crystal as well as other sources of noise present in the photoconductor itself. Since this factor is of a random nature (noise), it seems appropriate to use a sensitivity factor which involves the noise in a non-deal photoconductor, and to compare the sensitivity of such a detector with a truly background-limited,or ideal detector.

Background noise is primarily due to random fluctuations in the temperature of the detector. The temperature of a material changes when it absorbs radiation, and the random changes of temperature

set a lower limit on the amount of radiation which can be detected. The response of an ideal background limited photoconductor can be derived using Boltzmann statistics, in a manner similar to the Nyquist formula for electrical noise [36] . This response is stated in terms of the noise equivalent power (N.E.P.), which is the amount of radiation in watts which must be incident on the detector to give an electrical signal output equal to the r.m.s. value of the total noise (which for an ideal detector is thermal noise). The result is [36]

$$(N.E.P.)^2 = 4\pi\left[\frac{h^2\nu_o^2}{c^2}\int_0^\infty \frac{\eta\,\nu^2\exp(h\nu/k_BT)}{[\exp(h\nu/k_BT)-1]^2}\,d\nu\right] A(\Delta f) \qquad (1.99)$$

In this equation ν is the frequency of the radiation, and c is the speed of light. T is the absolute temperature, A is the area of the detector, and Δf is the bandwidth of the measured electrical signal. Likewise ν_o is the frequency which is being detected and η is the quantum efficiency. This equation assumes that the detector and its surroundings are at the same temperature. If the detector is at a temperature much less than the surroundings Equation (1.99) should be divided by 2. For an ideal photodetector the quantum efficiency is taken to be

$$\eta(\lambda) = \begin{cases} 1, & \lambda < \lambda_c \\ 0, & \lambda > \lambda_c \end{cases} \qquad (1.100)$$

where $\lambda = c/\nu$ is the wavelength and λ_c is called the cutoff wavelength. Figure 1.35 shows a plot of detectivity, D*, versus wavelength

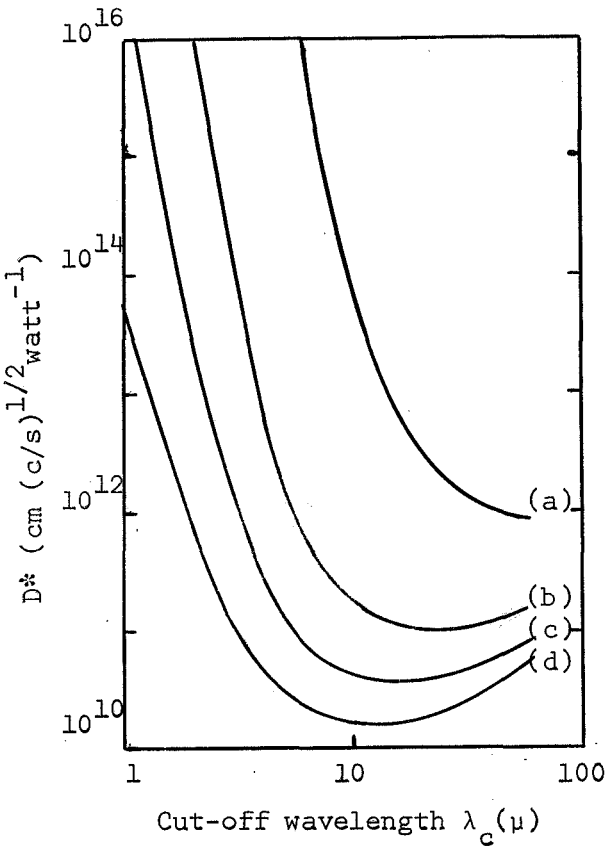


Figure 1.35. Factor of merit, D* (or N.E.P.)$^{-1}$ in unit band-width (for detector of unit area) for ideal photon detector assuming that incident radiation is just below cut-off wavelength λ_c. The detector and surroundings are at a temperature of 77° K(curve (a)), 195° K (b), 290° K (c), and 500° K (d).

assuming $\lambda_o = \lambda_c$ for an ideal photodectector. Here detectivity is defined as

$$D^* = (\text{N.E.P.})^{-1} \qquad (1.101)$$

Here is vivid evidence of the importance of temperature on the detectivity of a photodetector, since decreasing temperature increases D*, and thus reduces the N.E.P. Generally, to reduce the N.E.P. one must cool the surroundings of the detector as well as the detector itself [36], so that most infrared detectors

are used in either liquid nitrogen or liquid helium cryostats.

A practical infrared detector has sources of noise other than thermal noise which causes its detectivity to be less than that shown in Figure 1.35. Other sources of noise are:

1) Generation-Recombination Noise
2) Contact and Surface Noise
3) Circuit and Amplifier Noise

The first type of noise is due to random fluctuations in the carrier concentrations which influence the generation-recombination processes. These factors are discussed in Reference 36. For a non-ideal photoconductor the detectivity is thus defined as

$$D^* = \frac{S_N \sqrt{A\Delta f}}{P} \tag{1.102}$$

where S_N is the total signal-to-noise ratio. Quite often values for D* are quoted for unit area and unit bandwidth. Figure 1.36 shows values of D* versus wavelength defined this way for several types of detectors. Some of these detectors have not been discussed and here the reader is referred to Reference 36. Figure 1.37 shows detectivities versus wavelength for several semiconducting photodetectors which have proven quite useful. In Figures 1.36 and 1.37, D* is shown for an ideal non-selective detector, which does not take into account the blackbody nature of such a device, as has been done in Equation (1.99).

It is apparent from Figure 1.36 and 1.37 that the detectivity always has a maximum value for some specific wavelength in a given

material. When the photoconductor response is limited by generation-recombination noise, there exists a unique relationship between this maximum detectivity, D^*_{max}, and the wavelength, λ_m, at which this maximum value occurs, which is valid only if the material is an extrinsic photoconductor. This expression is given by [37].

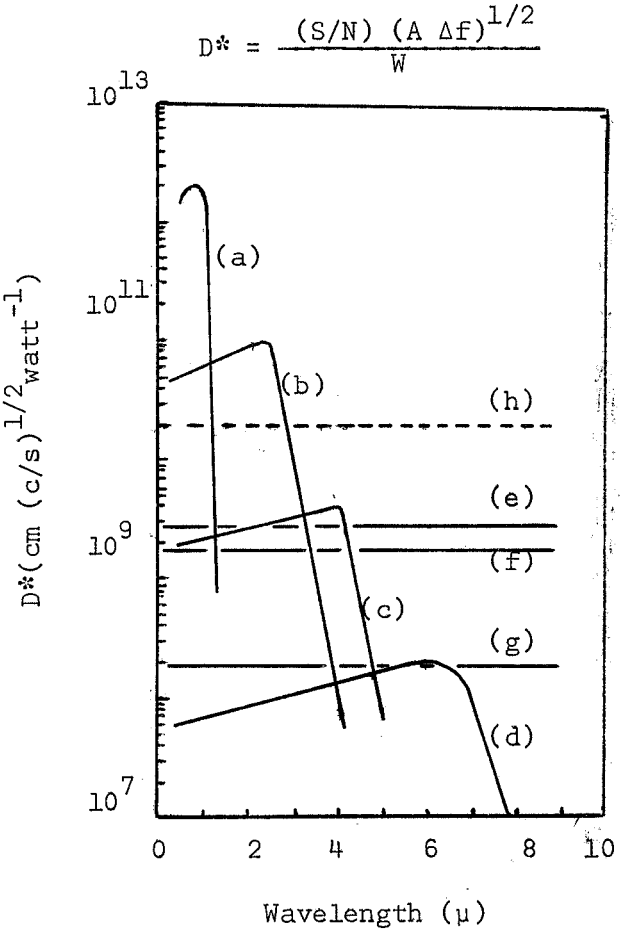


Figure 1.36. Factor of merit, D*, for various detectors operated at room temperature : (a) thallium-sulphide photoconductive cell; (b) PbS photoconductive cell; (c) PbSe photoconductive cell; (d) InSb photoconductive or PEM cell; (e) Hilger-Schwartz thermocouple; (f) Golay cell; (g) thermistor bolometer; (h) represents the thermal noise limit for a non-selective detector operated at represents the thermal noise limit for a non-selective detector operated at 300° K.

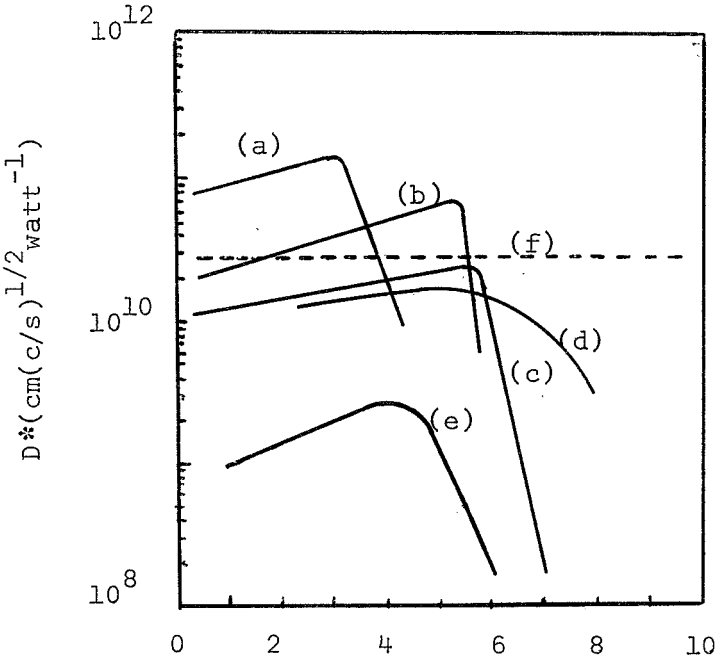


Figure 1.37 Factor of merit, D*, for various photoconductive detectors operated at 77°K: (a) PbS; (b) InSb; (c) PbSe; (d) Au-doped Ge, p-type (e) PbTe; (f) represents the thermal noise limit if the surroundings of the detector are at 300°K.

$$\ell n(D^*_{max}) = \ell n\ (A) + 0.434\ \left(\frac{hc}{2k_B}\right)\ \left(\frac{1}{\lambda_m T}\right) \qquad (1.103)$$

where A is a material constant only slightly dependent on temperature. Figure 1.38 shows experimental values of D^*_{max} as a function of $(\lambda_m T)^{-1}$ for several intrinsic as well as extrinsic photoconductors. The extrinsic photoconductors follow Equation (1.103) to a reasonable agreement. The lead telluride and lead sulfide detectors quoted in this figure are constructed as barrier-type devices and these devices are discussed in section 2. The dotted line in this figure is the best fit to Equation (1.103) for the extrinsic photoconductors, while the solid line is drawn for arbitrary value of the constant A.

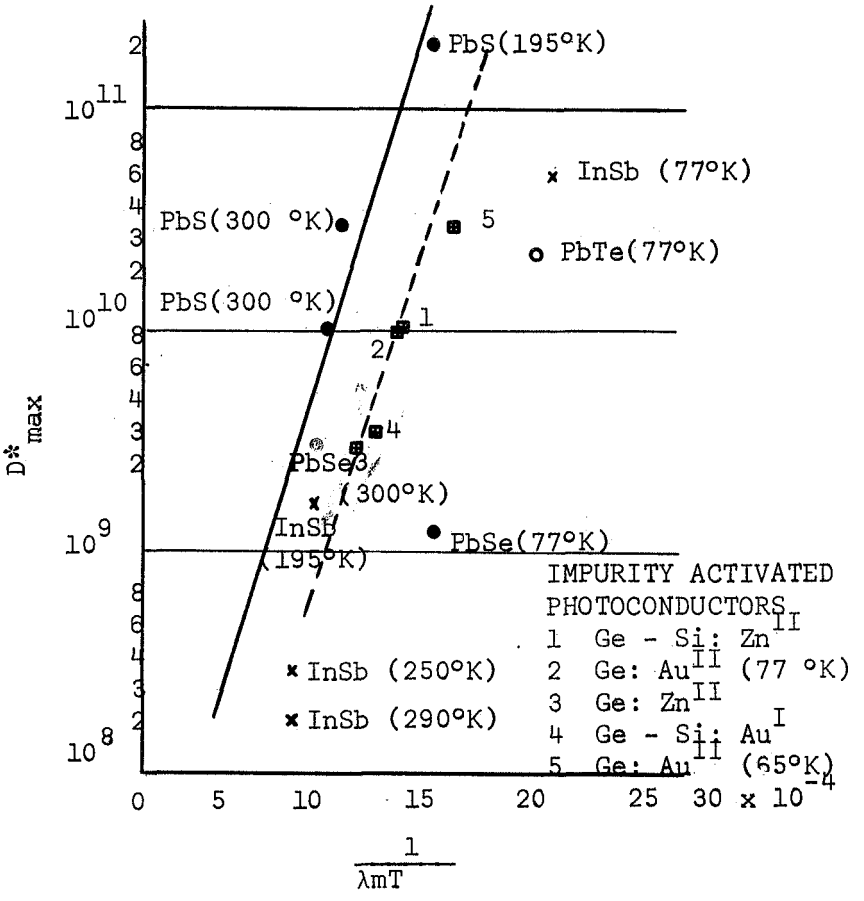


Figure 1.38. Relationship between peak detectivity and temperature long wavelength limit product.

d) Infrared Photoconductivity of Alloys - While the elemental, group III-V, and group II-VI semiconductors offer a wide range of values for λ_m in Equation (1.103), there are still gaps in the spectrum which are not efficiently covered by these materials. For example, Figure 1.39 shows detectivity versus wavelength over the spectral region of 1-40 microns for several extrinsic and intrinsic photoconductors. The use of alloys of the III-V and II-VI semiconductors, as well as alloys of the elemental semiconductors silicon and germanium offer the unique possibilities of tayloring the spectral response of

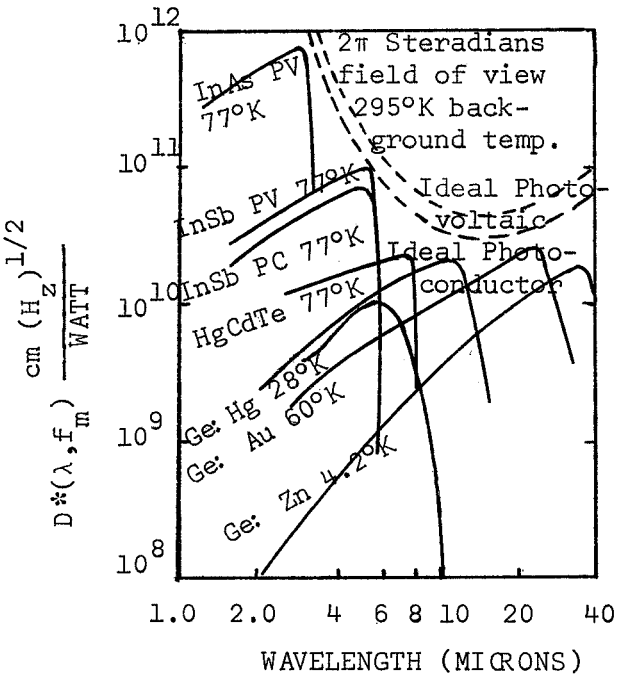


Figure 1.39. The spectral detectivity of a number of low temperature infrared detectors. Typical frequency of the noise measurement was 1000 Hz.

a photoconductor to any part of the spectrum for which a detector is desired. A specific instance which vividly illustrates this point is shown in Figure 1.40 . Here is shown normalized spectral response for alloys of the semiconductor HgCdTe, for varying compositional percentages of the element Cd in HgTe. It can be seen that by varying the percentage of Cd in HgTe form about 17% to 61% maximum responses can be obtained over the wavelength range from 1.5 microns to 20 microns. This type of material is considered to be an extrinsic photoconductor, and the behavior shown in Figure 1.40 is due to the compositional dependence of the bandgap of HgCdTe on the atomic percent of Cd in this alloy [38]. This feature of alloys is extremely important in photoconductor development, and there seems to

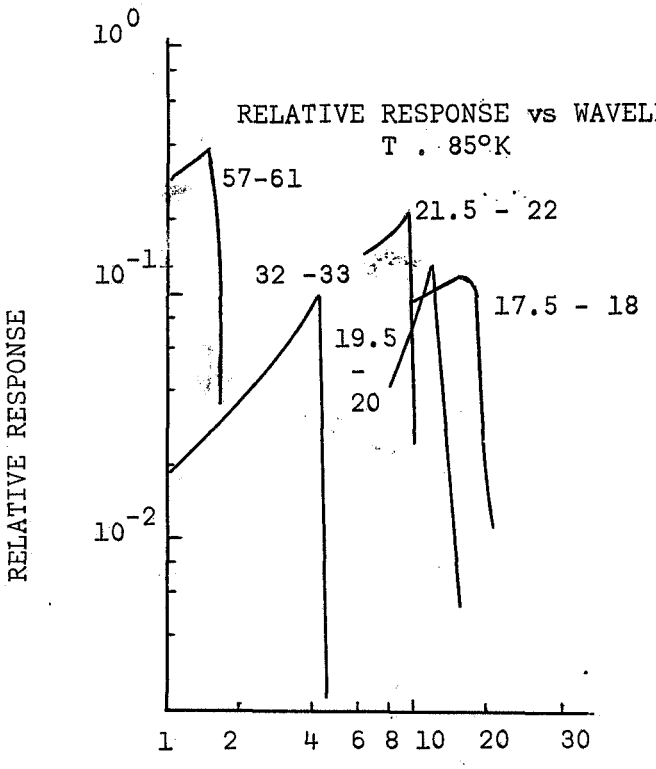


Figure 1.40. Representative spectral relative responce curves for various compositions (percent). The curves are not normalized to each other.

be much work needed here. The advances being made in materials preparation technology have opened up this new area of research which seems to be a very fruitful one from the standpoint of the development of new and improved optical sensors.

e) Speed of Response - While the sensitivity of a photoconductor to light of various wavelengths is extremely important, it is not the only consideration. Another equally important parameter is the response time, or speed of response. This is the time required for conductivity to respond to changes in the light intensity. If there are no trapping centers, then the speed of response is determined solely by the free carrier lifetime [39]. However, for most

photoconductors the response time is generally several orders of magnitude larger than the lifetime of free carriers, and this can be attributed to the trapping of free carriers. For example, if the light intensity is increased then the free electron concentration will increase. Since, by definition, a trap is in quasi-equilibrium with the conduction band, then the density of trapped electrons will also increase. The process of filling the traps is basically a slow process, and if the density of traps is large, the time required to fill the traps to the level required for quasi-equilibrium will be very long. In general, the response time can be expressed as [40].

$$\tau_o = (1 + \frac{n_t}{n}) \tau_n \simeq \frac{n_t}{n} \tau_n \qquad (1.104)$$

where n_t is the density of trapped electrons and n is the free electrons concentration, and it is assumed $n_t >> n$. As shown by Bube [39] an approximate expression for n_t as a function of τ_o is

$$n_t = \frac{4 \times 10^{12}}{\ell} L \tau_o \qquad (1.105)$$

where L is the light intensity in foot candles, and ℓ is the distance in the crystal over which the light is absorbed. Table 18 illustrates some typical response times calculated from this equation assuming a trap density of 10^{10} cm^{-3}, which is probably a lower limit to a practically realizable trap density. The important point is that response time becomes smaller the larger the light intensity, so that a photoconductor can be extremely slow to small changes in light intensity.

Table 18 Calculated values of response time for 10^{10} traps per cm^3 per kT

Light Intensity		Response Times	
Ft-c	Number Equivalent of 5000 A Photons, $cm^{-2}sec^{-1}$	$\ell = 10^{-2}cm$ [a]	$\ell = 10^{-5}cm$ [a]
1	4×10^{12}	25 μsec [b]	0.025 μsec [b]
10^{-3}	4×10^{9}	25 msec	25 μsec [b]
10^{-6}	4×10^{6}	25 sec	25 msec
10^{-9}	4×10^{3}	7 hr	25 sec

[a] ℓ is the thickness of photoconductor over which the light is assumed to be absorbed.

[b] Since a value of response time less than 10 msec implies that the photoconductor involved has proportionately less sensitivity than standard cadmium-sulfide-type photo cells, it would be more realistic to insert 10 msec in these places in the table in order to maintain the same high photosensitivity throughout.

f) Dark Conductivity - It has been mentioned that all semiconductors and insulators will exhibit photoconductivity to some extent. One criterion which determines how sensitive a material will be as a photoconductor is its dark or unilluminated conductivity. Depending on the magnitude of the band gap energy and the types of impurities present in the crystal, the dark conductivity can vary by many orders of magnitude. If the dark resistivity of a crystal at room temperature is greater than about 10^6 ohm-cm, then the photo excited carrier concentration is usually greater than the dark free carrier concentration. Such materials are usually classified as insulators. If the

dark resistivity is between about 10^2 to 10^6 ohm-cm then the dark free carrier concentration is greater than the photoexcited carrier concentration, and the material is not very suitable as a photoconductor [41].

g) Sensitization of Photoconductors by Imperfections - It has been observed for many single crystals of the II-VI compound semiconductors that perfect (as pure a state as technology will allow) crystals are much less photosensitive than those for which suitable imperfections have been incorporated. For example, in pure CdS the hole and electron lifetimes are comparable, each being on the order of a microsecond. However, in photosensitive materials the electron lifetime is on the order of a few milliseconds, while the hole lifetime is decreased by as much as three or more orders of magnitude. It can be concluded that the defects which control the lifetimes, and thus the sensitivity, have a small probability for capturing an electron (a small capture cross section) and large probability for capturing a hole (a large capture cross section). A compensated acceptor type imperfection with a net negative charge in thermal equilibrium could lead to such recombination properties [35]. For example,a Cu center substituting for a Cd atom, or a S vacancy could be postulated as causing this sensitizing process. However, the physical and chemical identification of those defects which lead to sensitization of photoconductivity have not been accomplished at the present time. There are indications which do lead to the conclusion that the sensitizing process is due to intrinsic defects in the crystal. Defects such as the sulfur

vacancy or vacancy-impurity complexes are the most likely candidates [35]. Table 17 contains a summary of the information which has been accumulated in CdS, and which has led to the previous conclusion. Evidence is also available to indicate that these are the primary source of sensitizing centers in other II-VI compounds such as CdTe, ZnS, and ZnTe [35].

1.5.10 The Photodielectric Effect

If an insulating photoconductor is used as the dielectric of a capacitor, and if light is incident on the capacitor when it is under the influence of an applied alternating electric field, the dielectric constant appears to change if the light is absorbed by the photoconductor. This phenomena has been given the name of the photodielectric effect. This effect has been observed primarily in powders of ZnS, ZnO, and CdS, as well as in single crystals of these materials [34]. The change in the dielectric constant has been reported to be about a factor of two to three in ZnS, while it is between about seven to one hundred in ZnO. Also changes as large as 10^6 (i.e. capacitance of 10^6 times the dark capacitance) have also been reported in single crystals of CdS with non ohmic contacts [41]. It is felt that these large changes are simply due to the effective reduction of the thickness of the capacitor because the bulk of the photoconductor is no longer an insulator when illuminated, and the capacitance is controlled by the space charge regions near the contacts.

According to Bube [32] there are three possible explanations for the photodielectric effect. The first is that the capacitance changes are simply due to the normal photoconductive effects in the material. If the photoconductor was non-homogeneous, then the photoconductive effect would decrease the effective distance between the capacitor plates causing an apparent increase in the dielectric constant as manifested by the capacitance. Supporting evidence for this hypothesis

is contained in the correlation between the infrared stimulation of normal photocurrents and capacitance changes. The second hypothesis assumes that there are real changes in the dielectric constant caused by the presence of polarization due to a separation between electrons bound to positively charge defects levels. The proponents of this argument support their views by emperiments which examine the frequency dependence of the dielectric constant change and the dielectric loss (loss tangent) change. Generally, it was found that these factors followed the relations

$$\frac{\Delta\varepsilon_o - \Delta\varepsilon_\infty}{\Delta\varepsilon - \Delta\varepsilon_\infty} \sim \omega^2$$

and (1.106)

$$\frac{1}{\Delta\gamma} \sim \omega^{-2}$$

Here $\Delta\varepsilon_o$ is the low frequency dielectric constant change , $\Delta\varepsilon_\infty$ is the high frequency dielectric constant change, and $\Delta\varepsilon$ is the dielectric constant change at frequency ω, all due to illumination. Also $\Delta\gamma$ is the change in the dielectric loss tangent. There is much more substantiating evidence to show that changes in dielectric constant due to illumination is a real effect (for a review the reader is referred to Reference 34). Finally, a third explanation for the photodielectric effect assumes that a real change in ε_r occurs due to space charge regions at grain boundaries, which

remain insulating under illumination while the remaining regions of the material are photoconducting. This explanation of the photodielectric effect normally is limited to noncrystalline solids.

As pointed out by Bube, all three explanations for the photodielectric effect are plausible, and all three could possibly attribute to observations in any given sample. While it is an interesting phenomena, it does not seem to offer any advantages over the normal intrinsic or extrinsic photoconductivity effects in terms of its application to optical sensors. However, if the effects observed in CdS with non-ohmic contacts can be reproduced and controlled, then it would seem that the phenomena could offer advantages in this area. To this writer's knowledge little information is known about this specific effect, and it has not been applied as an optical sensor.

1.5.11 The Thermoelectric Effects

It has been seen that current flow in solids arises from the transport of charge carriers through the material when an external electric field is applied. It is apparent that at the same time energy is transported through the crystal as the charge carriers are transported. It is to be anticipated that there is an intimate relationship between charge and energy transport. It can be recalled that in the derivations of conductivity in section 1.2, it was assumed that the distribution function was independent of spatial position in the crystal. Now, if this condition is relaxed and the sample is allowed to possess thermal gradients then an additional "force" is present in the crystal which influences both the charge and energy transport, and the electrical and thermal forces will interact. Also, in the case where the electric fields are zero, any thermal gradients in the sample will lead to both an electric current and a thermal current, i.e. a transport of charge and energy. The electric current is called diffusion current, and the thermal current leads to the identification of thermal conductivity. In terms of sensors the most important effects results from the various conditions which can occur when both electric field and thermal gradients are present in a given material. These effects are called thermoelectric effects.

In order to lay the theoretical foundation for the consideration of the thermoelectric effects, consider a one-dimensional solid with

an electric field, E_x, and a thermal gradient, $\partial T/\partial x$, in the x-direction. Using the relaxation time approximation the distribution function can be expressed as [7]

$$f = f_o - \tau v_x \left[k T \frac{\partial}{\partial T} \left(\frac{E - E_F}{k T} \right) \frac{\partial T}{\partial x} - qE_x \right] \frac{\partial f_o}{\partial E} \tag{1.107}$$

where v_x is the x directed velocity, E is the energy and E_F the Fermi energy of the particle distribution. Now, the electric and heat (or thermal) current densities in the x-direction can be expressed as

$$J_x = - q \int v_x f \, d\bar{p}$$

$$Q_x = \frac{1}{2m} \int p^2 v_x f \, d\bar{p} \tag{1.108}$$

Here p is the magnitude of the momentum (i.e. $p = \hbar k$) and $d\bar{p} = dp_x \, dp_y \, dp_z$ in a rectangular co-ordinate system. Also m is the mass of the electron and Q_x represents the energy transported per unit area per unit time by electrons traveling in the x-direction.

By substituting Equation (1.107) into Equation (1.108), the following general equations can be obtained for J_x and Q_x

$$J_x = \frac{qn}{m} \left[q \langle \tau \rangle E_x + \langle \tau E \rangle \frac{1}{T} \frac{\partial T}{\partial x} + \langle \tau \rangle k T \frac{\partial}{\partial T} \left(\frac{E_F}{kT} \right) \right.$$

$$\left. \frac{\partial T}{\partial x} \right]$$

$$Q_x = - \frac{n}{m} [q \langle \tau E \rangle E_x + \langle \tau E^2 \rangle \frac{1}{T} \frac{\partial T}{\partial x} + \langle \tau E \rangle kT \frac{\partial}{\partial T} (\frac{E_F}{kT}) \frac{\partial T}{\partial x}] \qquad (1.109)$$

where the brackets denote the average value as defined in Equation (1.9). In these equations is contained a reasonable theoretical basis for all the thermoelectric effects. Before examining the thermoelectric effects, several interesting features of Equation (1.109) can be noted. First of all, the last term in the expression for J_x gives rise to the diffusion current which is due to a particle density gradient. This is true since this term is more fundamentally written as

$$J_{x_{Diff}} = \frac{qn}{m} \langle \tau \rangle kT \frac{\partial}{\partial x} (\frac{E_F}{kT}) \qquad (1.110)$$

That is, even in the absence of an applied thermal gradient, if there exists a particle gradient the Fermi energy must change with position. For the case of a non-degenerate semiconductor such that

$$n = N_c \exp[- \frac{E_c - E_F}{kT}] \qquad (1.111)$$

Equation (1.102) becomes

$$J_{x_{Diff}} = - \frac{q^2 \langle \tau \rangle}{m} \cdot \frac{kT}{q} \frac{\partial n}{\partial x} = - q D_n \frac{\partial n}{\partial x} \qquad (1.112)$$

where $D_n = \frac{kT}{q} \mu_n$ defines the electron diffusion co-efficient from the well-known Einstein relation [7]. Equation (1.109) also

contain information about the thermal conductivity due to electron energy transport in a thermal gradient, where there is no net flow of electron current. By setting $J_x=0$ Equation (1.109) can be solved for E_x as a function of the thermal gradient. Here E_x is a"built-in field" which is necessary to cancel the diffusion current established because of the thermal gradient. By substituting for E_x in the heat current density equation one obtains

$$Q_x = \frac{n}{mT}\left[\frac{\langle\tau E\rangle^2 - \langle\tau\rangle\langle\tau E^2\rangle}{\langle\tau\rangle}\right]\frac{\partial T}{\partial x} \qquad (1.113)$$

The definition of electronic thermal conductivity, K_{Te}(in watts/cm °K), is

$$K_{Te} = -\frac{Q_x'}{\frac{\partial T}{\partial x}}$$

or (1.114)

$$K_{Te} = \frac{n}{mT}\left[\frac{\langle\tau\rangle\langle\tau E^2\rangle - \langle\tau E\rangle^2}{\langle\tau\rangle}\right]$$

Now, since the electrical conductivity is given by $\sigma = \frac{q^2\langle\tau\rangle}{m} n$ one has that

$$\frac{K_{Te}}{\sigma} = \left[\frac{\langle\tau\rangle\langle\tau E^2\rangle - \langle\tau E\rangle^2}{k^2 T^2 \langle\tau\rangle^2}\right] \cdot \frac{k^2}{q^2} T \qquad (1.115)$$

For a completely degenerate semiconductor (or a metal), this ratio is

$$\frac{K_{Te}}{\sigma} = \frac{\pi^2}{3}\left(\frac{k^2}{q^2}\right) T \simeq 2.5 \times 10^{-8} T \qquad (1.116)$$

which is called the Wiedemann-Franz law. Table 19 shows a table of $K_{Te}/\sigma T$ in mks units for several metals. As can be seen the agreement of these experimental values with Equation (1.116) is reasonable. For a semiconductor with a relaxation time of the form $\tau(\varepsilon) = a\,\varepsilon^{-s}$, Equation (1.115) can be shown to become [7]

$$\frac{K_{Te}}{\sigma T} = \left(\frac{5}{2} - s\right)\left(\frac{k^2}{q^2}\right) \qquad (1.117)$$

For a semiconductor with lattice scattering only, s = 1/2, and the Wiedemann-Franz ratio for a non-degenerate semiconductor should differ by no more than 33% from that of a metal or a degenerate semiconductor. Generally, at high temperatures measurements confirm these preditions [43] . This is an important fact relating the thermal and electrical properties of a wide variety of materials.

The discussion up to this point has, however, neglected a very important point. It has not considered that energy can be transported in a solid by means other than electron energy transport. For example, a most important mode of energy transport in solids is manifested by longitudinal and transverse elastic waves which can propogate through the crystal. Such waves are called phonons, and for insulators and high resistivity semiconductors

Table 19 Experimental values of $K/\sigma T$

Metal	273°K	373°K
Al	2.19×10^{-8}	2.22×10^{-8}
Cu	2.23	2.34
Zn	2.22	2.34
Rh	2.56	2.54
Ag	2.31	2.37
Sn	2.52	2.49
Bi	3.31	2.89

the thermal conductivity can be almost completely determined by phonons. Heat conduction in these materials can be thought of as being brought about by the diffusion of high energy phonons along the temperature gradient. If energy transport by both phonons and electrons are considered, then to a first approximation the total thermal conductivity can be expressed as

$$K_T = K_{Te} + K_{TP} \qquad (1.118)$$

where K_{Te} is the electron contribution to K_T, and is given by Equation (1.114) to (1.117) and K_{TP} is the phonon contribution to

K_T. Some typical examples of the order of K_{Te} and K_T have been given by Smith [7]. For one ohm - cm germanium $K_T \simeq 60$ watt/cm°C, and $K_{Te} \simeq 4.4 \times 10^{-4}$ watt/ cm °C for lattice scattering, which gives $K_{Te}/K_T \simeq 10^{-5}$. However, for 10^{-3} ohm cm $Bi_2 Te_3$, $K_T \simeq 2.25$ watt/cm °C. Thus $K_{Te}/K_T \simeq 0.2$. A mathematical expression is available for the phonon thermal conductivity, according to the Debye theory for the specific heat of solids. This equation is

$$K_{TP} = \frac{1}{3} C_v T \ell_m \qquad (1.119)$$

where C_v is the lattice specific heat at constant volume, v_T is the mean thermal speed of a phonon, and ℓ_m is the mean free path between phonon scattering events. Generally v_T is taken to be the speed of sound in the medium, and for materials where ℓ_m does not vary drastically with temperature K_{TP} should vary approximately with temperature as the specific heat. In highly discordered materials, such as quartz glass where ℓ_m is approximately equal to the mean distance between molecules, this fact has been born out [43] Figure 1.41 illustrates the lattice specific heat as a function of temperature normalized to the Debye characteristic temperature. While the data points on this curve are for yttrium, the curve is basically a universal curve for any material with the Debye temperature as an adjustable parameter. For temperatures below the Debye temperature, the specific heat, and thus the lattice thermal conductivity, varies rapidly with temperature. Above θ_D the specific heat and thermal conductivity should change only very slightly with temperature.

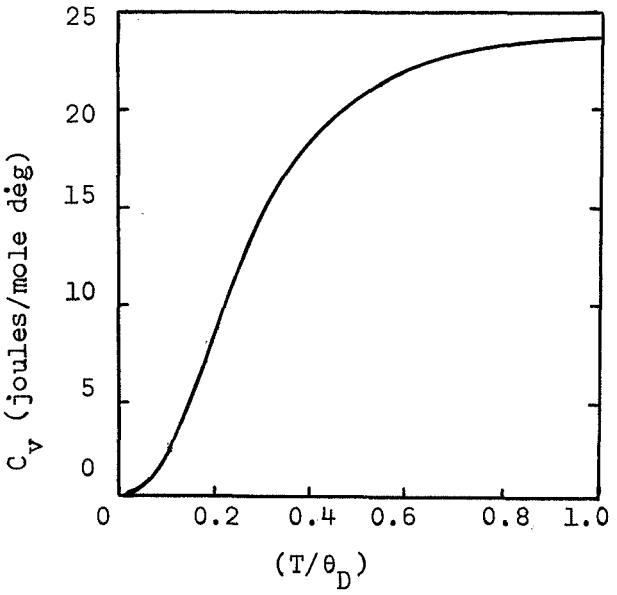


Figure 1.41. The curve of the specific heat of a solid (per mole) as a function of temperature, according to the Debye model in three dimensions. The experimental points have been plotted here with an abscissa scale appropriate for θ_D of 200°K. (An optimum fit for the lowest temperature points would require a slightly higher Debye temperature.) Note that the experimental points are above the Debye curve at high temperatures because C_p is the quantity actually measured, not C_v.

The thermal conductivity is an important parameter in determining the thermoelectric behavior of materials, and thus is of importance in the sensor area. It is also an important parameter to be considered in the design of sensor housings and device encapsulations. While only a few sensors directly use the thermal conductivity as the basic means to sense external stimuli, it is a sufficiently important material property to warrant the discussion presented here.

The significance of the thermoelectric effect, which is simply the interation between the electrical and thermal conduction mechanisms, can very easily be seen from Equation (1.109) and the use of a converva-tion theorem from thermodynamics. This relation states that

$$\frac{dH}{dt} = \bar{J} \cdot \bar{E} - \nabla \cdot Q$$

or in one dimension (1.120)

$$\frac{dH}{dt} = J_x E_x - \frac{\partial Q_x}{\partial x}$$

Here H is the rate at which heat is generated per unit volume, $\bar{J} \cdot \bar{E}$ is the rate at which electrical heat energy is generated per unit volume and $\nabla \cdot \bar{Q}$ is the rate at which heat is absorbed per unit volume. Using Equation (1.109) and substituting into Equation (1.120) the following relation is obtained,

$$\frac{dH}{dt} = \frac{J_x^2}{\sigma} + \frac{\partial}{\partial x}\left(K_T \frac{\partial T}{\partial x}\right) + J_x\left(\frac{kT}{q}\right) \frac{\partial}{\partial T}\left[\frac{\langle \tau E \rangle}{\langle \tau \rangle} \frac{1}{kT}\right] \frac{\partial T}{\partial x} \quad (1.121)$$

which can be expressed as

$$\frac{dH}{dt} = \frac{J_x^2}{\sigma} + \frac{\partial}{\partial x}\left(K_T \frac{\partial T}{\partial x}\right) + \frac{dQ'}{dt} \quad (1.122)$$

The first and second terms in this equation represent the heat generated by electrical conduction only and heat generated by thermal conduction only, respectively. dQ'/dt is responsible for the thermoelectric effects, and is present due to an interaction between electrical current flow and thermal current flow, as is evident from the simultaneous dependence on J_x and $\partial T/\partial x$.

There are three thermoelectric effects which are manifested through Eqution (1.122). These are:

a) The Thompson effect- the generation or absorption of heat within the volume of a material due to the simultaneous presence of a thermal gradient and an electrical current flow

b) The Seebeck effect - the generation of thermoelectric voltage in a circuit consisting of two joined materials, the two junctions of which are maintained at different temperatures

c) The Peltier effect - the generation or absorption of heat at the junction between two materials as on electrical current is passed through the junction.

These three thermoelectric effects will now be discussed seperately.

1.5.11.1 The Thompson Effect

From the discussion leading up to Equation (1.113) it is obvious that, when a thermal gradient exists along a material, there is a generated electric field which is proportional to the thermal gradient for the condition of zero electrical current density. It has also been pointed out that the heat flow in the presence of both a thermal and electrical current is different than when either of these two currents are zero. This additional heat flow is created reversibly in that a reversal of the direction of either the thermal current or the electrical current causes heat to be absorbed from the surroundings of the material instead of generating heat. This is seen by examining the relation for dQ'/dt in Equation (1.122). This reversible heat generated per unit volume per unit time is called the Thompson heat or Thompson effect, The electric field

generated due to the thermal gradient when the electric current is zero will be seen to be related to the Thompson effect.

The Thompson coefficient is formally defined from Equation (1.122) as the factor which relates the heat absorbed per unit volume per unit time to the electrical current density and the thermal gradient [45]. Thus

$$\sigma_T = - \frac{dQ'/dt}{(J_x)\ (\partial T/\partial x)} \tag{1.123}$$

and is taken to be positive if heat is generated by an electric current passing from regions of high to low temperature. From Equation (1.121) it can be seen that

$$\sigma_T = - \frac{kT}{q} \frac{d}{dT} \left[\frac{\langle \tau E \rangle - \langle \tau E_F \rangle}{kT \quad \langle \tau \rangle} \right] \tag{1.124}$$

Also, from Equation (1.109), the electric field due to the thermal gradient with no net electric current flow is given by

$$E_x = \frac{-1}{qT} \left[\frac{\langle \tau E \rangle - \langle \tau E_F \rangle}{\langle \tau \rangle} \right] \frac{\partial T}{\partial x} \tag{1.125}$$

By comparing these equations, the relationship between the Thompson coefficient and the electric field is clearly seen. This relationship will become more important in the discussion of the Seebeck effect in the next section. Generally, it can be shown that the magnitude of the Thompson coefficient is on the order of [7].

$$\sigma_T \sim \begin{cases} -\frac{k}{q}\left(\frac{kT}{E_F}\right), \text{ metals} \\ -\frac{k}{q}\left(\frac{E_F}{kT}\right), \text{ semiconductors} \end{cases} \qquad (1.126)$$

Now k/q equals 86.2 microvolts per degree. Since E_F for metals is on the order of several electron volts, while it is on the order of a few tenths of an electron volt for a semiconductor, it can be seen that σ_T is orders of magnitude larger for semiconductors than it is for metals. Table 20 shows a list of the Thompson co-efficients for several of the alkali metals for temperatures slightly above and slightly below the melting point, along with theoretical values calculated from Equation (1.124). With the exception of lithium the agreement is quite reasonable.

1.5.10.2 The Seebeck Effect

The Seebeck effect is the most well-known of the thermoelectric effects since it is the principle on which the thermocouple temperature sensor is based. This effect utilizes two materials connected together in a manner similar to Figure(1.42) In this circuit the junctions between the two materials A and B are maintained at the temperatures T_1 and T_2, respectively. If the circuit is operated so that no electrical current can flow, then from the discussion of the Thompson effect, it should be expected that a voltage will be developed across the open terminals of the circuit.

The absolute thermoelectric power, S, of a material is defined by [1]

$$E_x = S \frac{\partial T}{\partial x} \qquad (1.127)$$

Table 20 Thompson coefficients μ in microvolts/degree of the alkalis just above and just below the melting-point

		Li	Na	K	Rb	Cs
μ/T:	Solid	0.022	(0.005)	-0.036	-0.036	(0.008)
	Liquid	0.030	-0.048	-0.043	-0.085	-0.076
	Calculated	-0.016	-0.023	-0.036	-0.041	-0.048

where, from Equation (1.125), S is given by

$$S = -\frac{1}{qT}\left[\frac{\langle \tau E\rangle - \langle \tau E_F\rangle}{\langle \tau\rangle}\right] \tag{1.128}$$

and obviously

$$\sigma_T = T\frac{dS}{dT} \tag{1.129}$$

Equation (1.129) is known as one of Kevin's relations, and can be proven from thermodynamic arguments. The open circuit voltage which is developed in the circuit of Figure 1.42 can be obtained from Equation (1.127) as

$$V_o = -\int_o^L E_x\, d_x = \int_{T_1}^{T_2} (S_A - S_B)\, dT \tag{1.130}$$

where S_A and S_B are the absolute thermoelectric powers of materials A and B, respectively. The quantity $S_{AB} = S_A - S_B$ is referred to in the literature as the thermoelectric power. The distinction between absolute thermoelectric power and thermoelectric power is made, due to the fact that the absolute thermoelectric power is difficult to

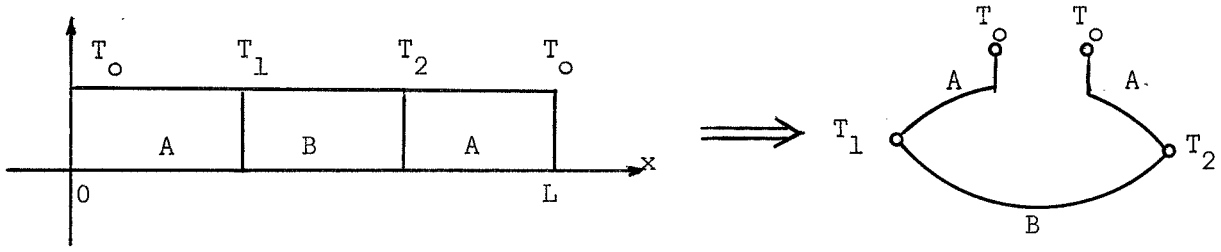


Figure 1.42. Schematic illustration of the Seebeck Effect

measure experimentally. The Seebeck effect measures differences between the properties of two materials, rather than the properties themselves. In order to measure the absolute thermoelectric power of a given material the Seebeck effect experiment must be performed using a material whose thermoelectric power is known as a reference. A common material for this purpose is lead, which has become a secondary standard [1].

The Seebeck effect in metals has been utilized as a temperature sensor, commonly known as a thermocouple. The principle of operation is based on Equation (1.130). Metals are useful materials for thermocouples because they can be made in the form of wire and junctions can be formed quite easily. Thermocouples give reasonable accuracies with reasonable output voltages which can be detected with standard d-c measurement techniques. They have the advantages of very small size, negligible thermal capacity and heat generation, and are reliable, reproducible and inexpensive.

The most common thermocouple material is probably a junction between copper and constantan. The thermoelectric power of such a thermocouple varies from about 6 μV/°K at 20°K to about 40 μV/°K at room temperature. Reliable tables are available for this thermocouple with respect to 0°C down to 10°K. However, one should always make certain that the composition of the materials being used for a thermocouple is the same as that for which the tables were compiled. This can be a problem in using published thermocouple voltage tables. Figure 1.43 shows the thermoelectric power versus temperature for several thermocouple materials. The numbers on the figure indicate the atomic percent of the alloy constituents.

The theory outline up to this point predicts that for an "ideal metal", the absolute thermoelectric powe varies with temperature as [1].

$$S = \begin{cases} -\frac{\pi^2 k}{q}\left(\frac{kT}{E_F}\right), & T >> \Theta_R \\ -\frac{\pi^2 k}{3q}\left(\frac{kT}{E_F}\right), & T << \Theta_R \end{cases} \qquad (1.131)$$

These equations predict negative values of S which are on the order of a few microvolts per degree. Generally this theory is completely inadequate for most metals, and the observed temperatures dependence of S is drastically different from a linear function, especially at low temperatures below about 80°K. Also, some metals have positive values of S, a condition which is quite opposite from Equation (1.132). A more general expression for S can be derived which takes into

account the shape of the Fermi surface Reference 2. This result predicts that non-spherical Fermi surfaces can lead to a positive value for the absolute thermoelectric power, although the results cannot be verified exclusively because of an uncertainty in the influence of the relaxation time on the results. It is generally concluded that positive absolute thermoelectric powers result from a combination of both a complicated relaxation time and a complicated Fermi surface. Table 21 summarizes the properties of several metals,

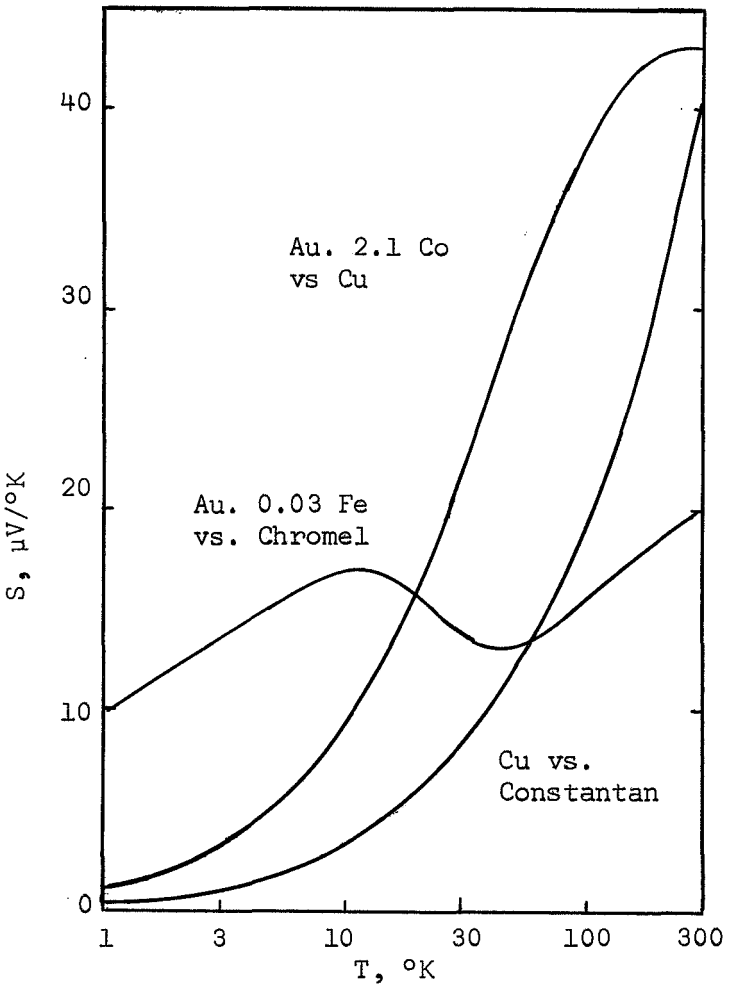


Figure 1.43. The thermoelectric powers of three useful low-temperature thermocouples compared with each other. (a) Au-2.1 at.-% Co against Cu, (b) Cu against constantan, and (c) Au-0.03 at. -% Fe against chromel.

Table 21 Parameters for some monovalent metals at 0°C theoretical value for $10^{8}\rho_{o}K_{T/T}$ is 2.45 [joules/(coul) (°K)]^2

Metal	Density g/cm^3	10^{-22}n, cm^3	Debye temperature θ, °K	$10^{8}\rho_{o}$, ohm- m	α_{o}, μv/°K	K_T, watts/(m)(°K)	10^{-16}T(μ)/m*(coul)(°K)] sec/kg	$10^{8}\rho_{o}K_{T/T}$ [joules/(coul)(°K)]^2
Li	0.53	4.6	363	8.55	+11.6	71	1.0	2.22
Na	0.97	2.5	202	4.2	- 8.3	135	3.7	2.18
K	0.86	1.3	163	6.15	-15.6	100	4.9	2.25
Rb	1.53	1.1	85	11.3	-10.2	...	3.1	
Cs	1.89	0.85	54	18.8	- 0.1	...	2.4	
Cu	8.92	8.5	333	1.60	+ 1.73	390	2.9	2.28
Ag	10.5	5.8	223	1.47	+ 1.35	420	4.6	2.26
Au	19.3	5.9	175	2.19	+ 1.85	310	3.0	2.48

including values of S(in this table α_o = S), and gives experimental values of the variable 10^8 $[\rho_o K_T/T]$ in [joules/coul °K]^2. The theoretical value is 2.45. Another mechanism which has been found to greatly influence the thermoelectric power at low temperatures is called phonon-drag. This mechanism is determined by the interaction of electrons and phonons, and generally the Umklapp processes mentioned in Section 1.2.5 are very important. This mechanism will not be discussed here and the reader is referred to References 1 and 2.

The absolute thermoelectric power for semiconductors is in about the same situation as that for metals, although the high temperature results are in reasonable agreement with theory. Based on Equation (1.128) and assuming a relaxation time of the form of $\tau(E) = a\, E^{-s}$ gives for the absolute thermoelectric power of an n-type non-degenerate semiconductor an equation of the form

$$S = -\frac{k}{q}\left[\frac{5}{2} - s + \ln\left\{\frac{2(2\pi m^* kT/h^2)^{3/2}}{N_D}\right\}\right] \qquad (1.132)$$

in the extrinsic temperature range where the electron concentration is independent of temperature and equals the doping level N_D. For a semiconductor which has both hole and electron conduction, or for a many-valley semiconductor, the relation for S can be written as [1]

$$S = \frac{\sum_i (\sigma_i S_i)}{\sum_i \sigma_i} \qquad (1.133)$$

where σ_i is the conductivity for the ith conduction or valence band valley, or where it is the conductivity for either holes or electrons, and where S_i is the absolute thermoelectric power under the same conditions. The value of S for holes is positive and of about the same magnitude as that for electrons, so that S for an intrinsic semiconductor is determined primarily by the carriers having the highest mobility. Figure (1.44) shows a plot of S versus temperature which should be expected for a semiconductor on the basis of Equation (1.128). Figure 1.45 shows experimental results for p-type germanium and compares experiment with theory. At low temperatures the agreement is atrocious. The primary reason for disagreement is the phenomenon of phonon drag and the neglect of complicated scattering mechanisms in Equation (1.133). A treatment of this effect in semiconductors can be found in Reference 1.

1.5.10.3 The Peltier Effect

The Peltier effect is the phenomena whereby heat is generated or absorbed at the junction of two dissimiliar materials which are maintained at constant temperature and through which is passed a constant electrical current. Consider the simple one dimensional situation depicted in Figure 1.46. The total heat generated in the volume $A(x_1 + x_2)$ per unit time is given by

$$Q_v = A \int_{x_1}^{x_2} \frac{dH}{dt} \, dx \qquad (1.134)$$

Using Equation (1.121) and integrating with the limit $x_1 \to x_2 \to 0$ gives

$$Q_v = AJ_x \frac{kT}{q} \left[\left(\frac{\langle \tau E \rangle - \langle \tau E_F \rangle}{\langle \tau \rangle} \right) - \left(\frac{\langle \tau E \rangle - \langle \tau E_F \rangle}{\langle \tau \rangle} \right) \right] \tag{1.135}$$

The Peltier coefficient is defined by

$$\Pi = \frac{Q_v}{J_x A}$$

which is (1.136)

$$\Pi = T(S_2 - S_1)$$

where S_2 and S_1 are the absolute thermoelectric powers of materials 2 and 1, respectively. This is another of the Kevin relations. From

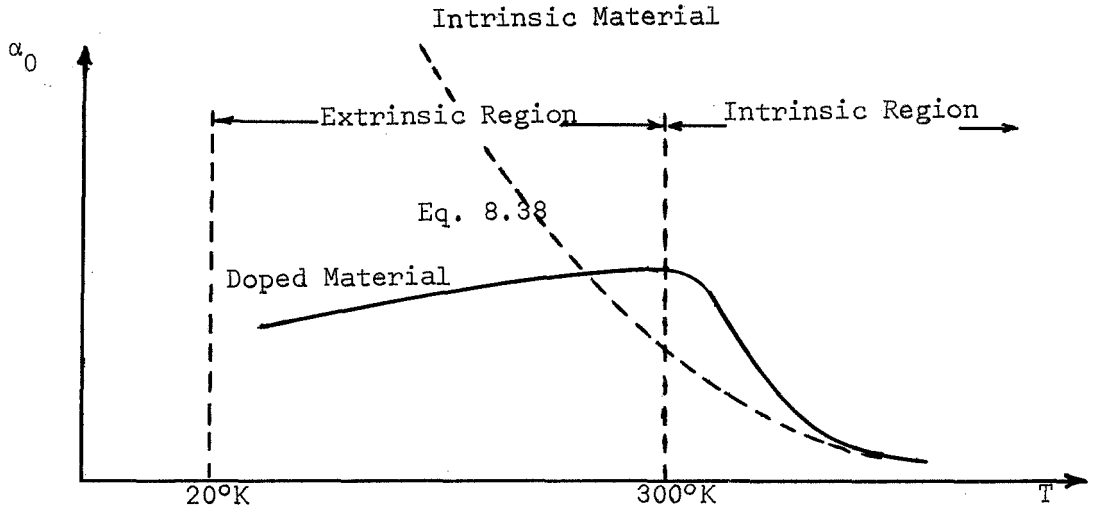


Figure 1.44. Thermoelectric power of a semiconductor. At low temperatures the simple formulas fail because of failure of the relaxation-time approximation. In an extrinsic material, α_0 increases slightly with temperature in the range where minority carriers are negligible because of the temperature dependence of N_c or N_v.

Equation (1.136) and (1.129) it can be seen that the thermoelectric effects are all intimately related to the Seebeck effect through the absolute thermoelectric power of a given material. The absolute thermoelectric power should be considered a basic material parameter and the experimental measurement of S can be very complimentary to the Hall effect in determining the properties of materials.

The Peltier effect is a most useful thermoelectric effect in terms of its application to the conversion of electrical energy to cause a cooling or refrigeration effect. More will be said about this effect in the next section.

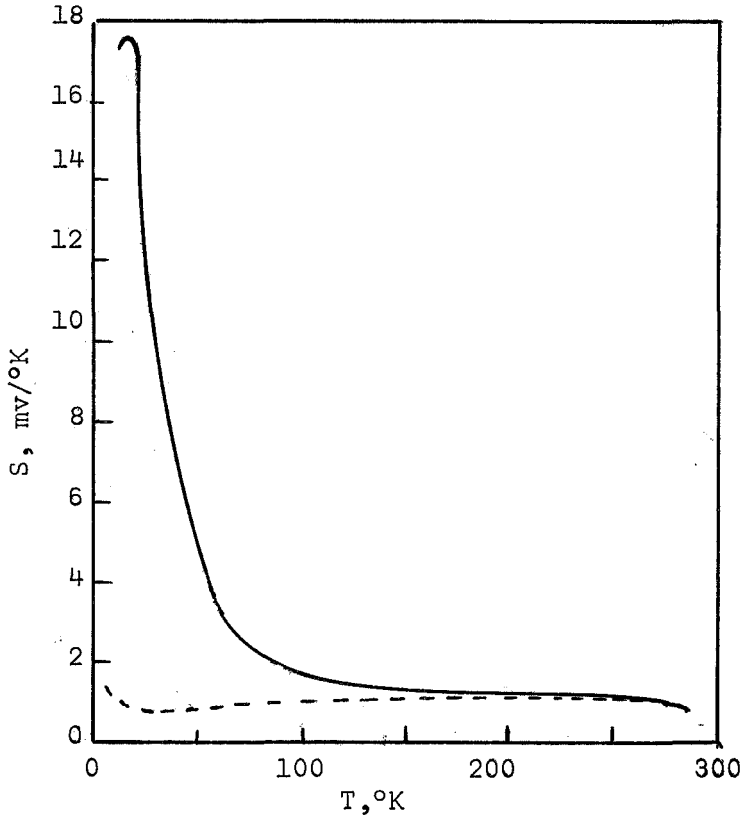


Figure 1.45. Thermolectric power of p-type Ge (1.5×10^{14} acceptors per cubic centimeter). Dashed curve represents the calculated value . The difference between the curves is the result of phonon drag.

1.5.10.4 Thermoelectric Power Generation

The thermoelectric effects obviously have application as temperature sensors, and thermocouples utilizing the Seebeck effect are probably the simplest and least expensive means by which temperature is presently measured. The Seebeck and Peltier effects have further application since they involve the generation or absorption of heat energy in the manner in which they manifest themselves. Thus the thermoelectric effects should be applicable as sensors of heat flux or thermal gradients in such areas as aerodynamic research and testing. Sjpecifically, the measurement of these properties in wind tunnel testing of flight vehicle models is an area application for the thermoelectric effects.

A simple thermoelectric cell is shown in Figure (1.147). In this figure a generator is illustrated since heat at the hot junction causes electrical power to be delivered to the load resistance. A refrigerator can be made by replacing the load resistance by a source of electrical energy and through the Peltier effect heat from the hot junction results in a cooling effect. For the time being the discussion will be directed toward the generator, which is basically a

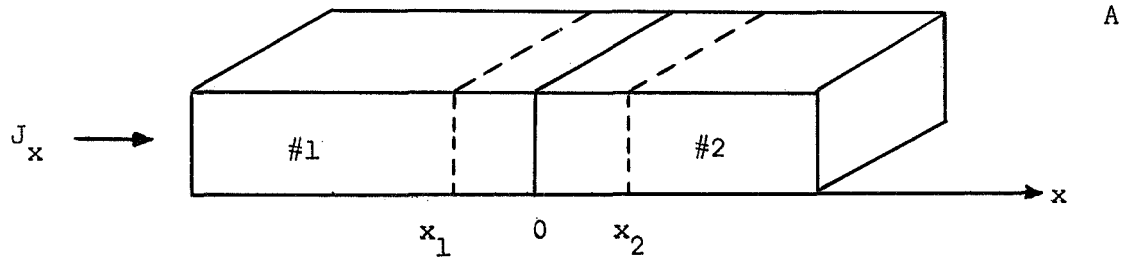


Figure 1.46. Schematic Illustration of the Peltier Effect

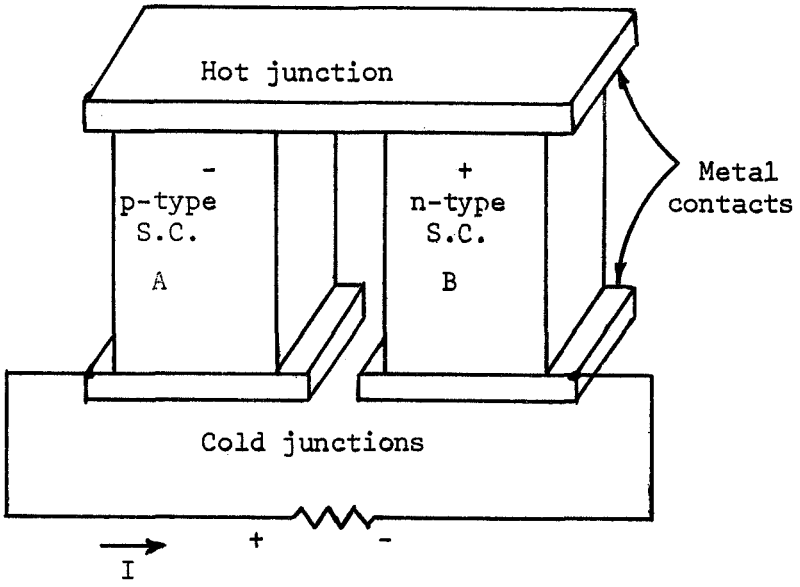


Figure 1.47. A simple semiconductor thermoelectric device, which may be used as either a generator (as shown here) or a refrigerator.

heat flux sensor. Here heat is supplied to the hot junction having a temperature T_H. The other junction is maintained at a temperature T_c. Due to the temperature difference a current will flow through the circuit because of the Seebeck effect, and the heat can be converted into electrical energy. The efficiency of the thermoelectric generator is defined as

$$\eta = \frac{\text{power supplied to load}}{\text{heat absorbed at hot junction}} \qquad (1.137)$$

Now, the power supplied to the load is simply I^2R_L. The heat absorbed at the hot junction consists of the Peltier heat abosrbed at the hot junction plus the heat which is withdrawn from the junction and dissipated in the form of electrical energy in the semiconductor elements and in the form of thermal conduction down the semiconductor elements.

There is also heat withdrawn from the junction in the form of Thomson heat in the semiconductor elements and this will be assumed to be neglectable [45]. Thus the efficiency can be written as

$$\eta = \frac{I^2 R_L}{S_{AB} I T_H + C_K (T_H - T_c) - 1/2\ I^2 R} \qquad (1.138)$$

Here $S_{AB} = S_A - S_B$ is the thermoelectric power of the n- and p-type semiconductor thermocouple element, C_K is the thermal conductance of the two branches in parallel, and R is the sum of the resistances of the semiconductor elements. For maximum power transfer to the load for a given heat absorption, the derivative of η with respect to R_L/R should be set equal to zero. This yields an expression for the maximum efficiency of

$$\eta_{max} = \frac{T_H - T_c}{T_H} \cdot \frac{\sqrt{1 + Z\bar{T}} - 1}{\sqrt{1 + Z\bar{T}} + (T_c/T_H)} \qquad (1.139)$$

In this Equation $\bar{T} = 1/2\ (T_H + T_c)$ is the mean temperature and the factor Z is given by

$$Z = \frac{S^2_{AB}}{RC_K} \qquad (1.140)$$

The quantity Z is called the thermoelectric figure of merit. Equation (1.130) can be seen to contain the factor $(1 - T_H/T_c)$, which is the thermodynamic efficiency of a reversible Carnot cycle, multiplied by

a factor less than unity which reduces the maximum efficiency because of heat conduction and Joule heating in the semiconductor elements. The heat absorption in the branches can be minimized by choosing

$$\frac{R_A}{R_B} = \frac{\rho_A K_{TA}}{\rho_B K_{TB}} \tag{1.141}$$

where ρ_A, ρ_B and K_{TA},K_{TB} are the electrical resistivities and thermal conductivities of the n- and p- type materials, respectively. Under this condition, the figure of merit becomes

$$Z = \frac{(S_A - S_B)^2}{[\sqrt{P_A K_{TA}} + \sqrt{\rho_B K_{TB}}]^2} \tag{1.142}$$

and for the case where $S_A = - S_B$ and $\rho_A K_{TA} = \rho_B K_{TB}$ Z becomes

$$Z = \frac{S_A^2}{\rho_A K_{TB}} \tag{1.143}$$

A quantitative reason why semiconductors make the best thermoelectric generators can be seem by considering Figure 1.48. This figure shows representative plots of thermoelectric power, electrical conductivity , and thermal conductivity as a function of carrier concentration which spans the range from insulator to semiconductor to metal. The variation of Z with carrier concentration is also shown and it is seen that Z has a maximum for semiconducting material whose carrier concentration is around 10^{19} cm^{-3}. Also, from Equation (1.139), it can be seen that

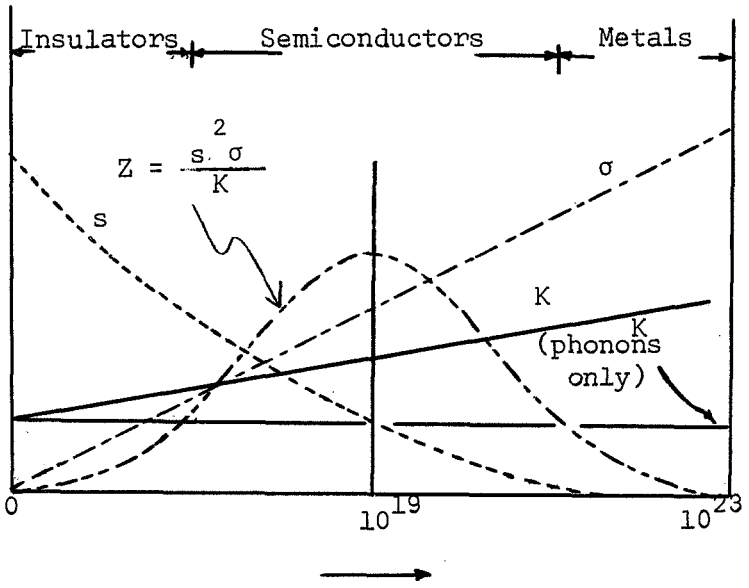


Figure 1.48. Electrical and thermal conductivities and thermoelectric coefficient as a function of carrier concentration. (It is implied that band gap decreases toward the right.)

to obtain large values of η_{max} one desires to obtain large values of the product $Z\overline{T}$. This fact is illustrated in Figure 1.49 where η_{max} is plotted as a function of Z for various values of T_H and for T_c = 300°K. Figure 1.50 shows typical values of Z versus temperature for several semiconductors and semiconductor alloys which have been used as thermoelectric materials. The materials cover the temperature range of 0-750°C quite adequately. Other materials research has lead to the development and study of Ge-Si alloys as thermoelectric elements, and these compounds have a maximum absolute thermoelectric power of almost 1.1 mV/deg for an alloy of 15% Si and a carrier concentration of 1.3×10^{15} cm^{-3} [46]. The usefulness of these materials at high temperatures (900°K - 1300°K) is also promising. Other alloys and III-V compound semiconductors have shown promise as thermoelectric

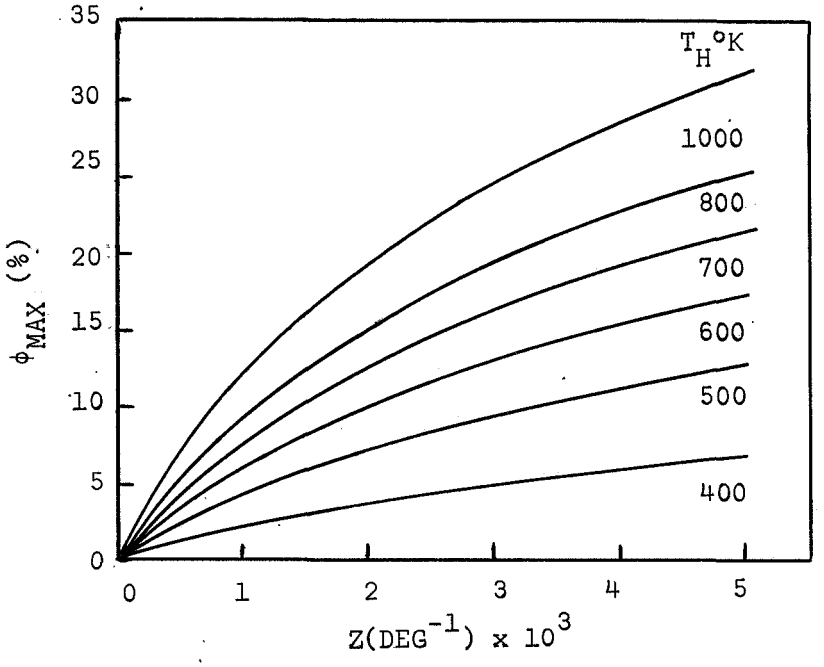


Figure 1.49. Maximum coefficient of performance Φ_{max} as a function of figure of merit Z for different hot junction temperatures T_H(T_C = 300°K).

materials and a great deal of research is going on in this area.

Finally, the arrangement shown in Figure 1.47 can be used as a heat flux sensor by measuring the open circuit voltage across the thermoelectric elements. This voltage will be directly proportional to the heat flux incident on the hot junction since

$$V_o = S_{AB} (T_H - T_c)$$

$$Q_h = C_K (T_H - T_c) \qquad (1.144)$$

Thus $V_o = \frac{S_{AB}}{C_K} Q_h$

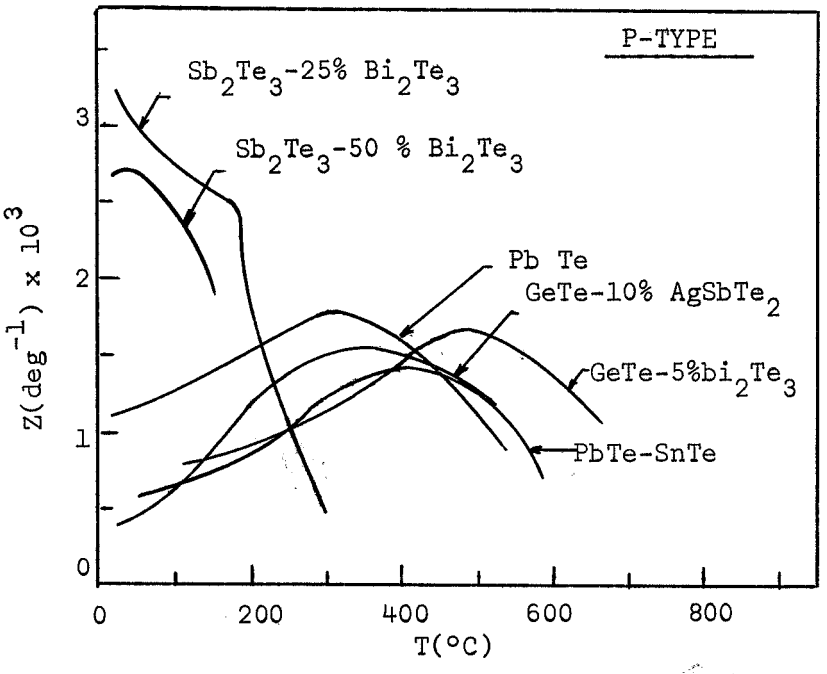


Figure 1.50a. The temperature dependence of the figure of merit, Z for p-type telluride semiconductors and their alloys.

where V_o is the output and Q_h is the incident heat flux. The other quantities have been defined previously. Using typical values of S_{AB} = 100 μ V/°C, K_T = .05 watt/cm °C, A = 10^{-4} cm^2, and ℓ = 1 cm gives a sensitivity of

$$\frac{V_o}{Q_h} = 20 \ \frac{\mu v}{\mu watt}$$

which is a very good value for sensitivity for the detection of small heat fluxes.

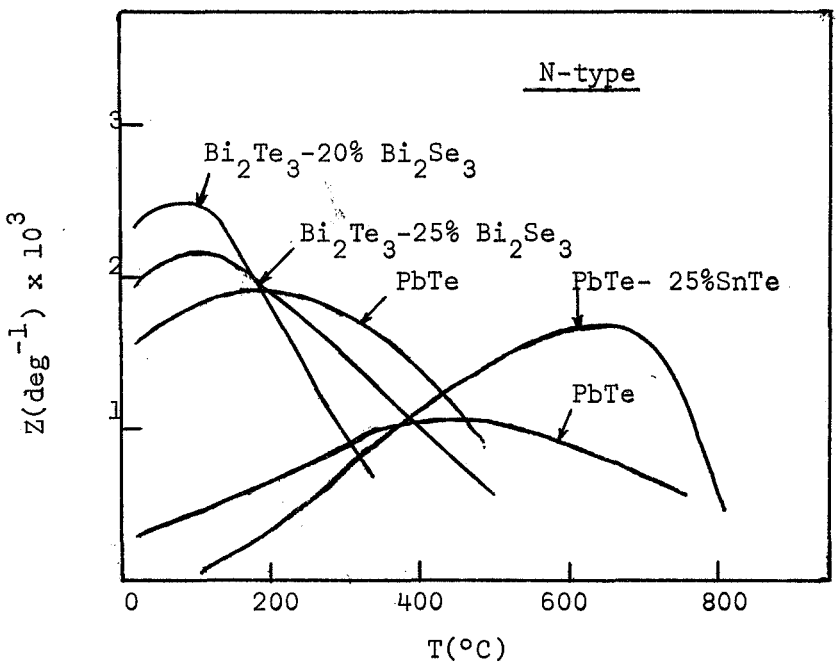


Figure 1.50b. The temperature dependence of the figure of merit, Z for n-type telluride semiconductors and their alloys.

1.5.11 Theromagnetic Effects

Whenever a thermal current exists in a solid, the application of a magnetic field gives rise to several effects which are analogous to the galvonomagnetic effects, and which are a result of the coupling between the thermal and electrical currents which flow. Basically, these effects result from the fact that the magnetic field deflects the charge carriers from the direction which they would normally follow, as evidenced for example, by the Hall effect. This means that energy will also be transported, which will generally result in establishment of thermal gradients in directions for which they would normally not exist in the absence of an applied magnetic field. These effects do not seem to offer any unique application to sensors. However, they are effects which can lead to errors in the use of the Hall effect and magnetoresistance effects as sensors and thus appear to be worth mentioning. These effects will only be listed and basically defined here. For a more detailed discussion the reader is referred to Reference 47. These effects are: a) The Ettinghausen effect is the development of a thermal gradient which is mutually perpendicular to the electrical current and applied magnetic field in a Hall effect type of experiment. The Ettinghausen coefficient, A_E, is defined by the relation

$$\frac{\partial T}{\partial y} = - A_E B_z J_x \qquad (1.145)$$

In metals, the Ettinghausen effect is small, leading to thermal gradients on the order of 10^{-2}°K/cm. In semiconductors the effect can be larger, and is often large enough to cause error in Hall effect measurements. The Ettinghausen effect leads to an effective Hall co-efficient of

$$R_{eff} = R \pm A_E S \tag{1.146}$$

where R is the normal Hall coefficient and S is the absolute thermoelectric power. Generally, by proper control of the experiment, $A_E S$ can be made less than the experimental error in R_{eff}, and is unimportant.

b) The Nernst effect is the appearance of a thermal gradient along the direct of current flow when a transverse magnetic field is applied. The Nernst coefficient, A_N, is defined by

$$\frac{\partial T}{\partial x} = A_N J_x \tag{1.147}$$

Little information is available on experimental values of A_N. However, it is known to be a function of magnetic field.

c) The Righi-Leduc effect is due to a thermal gradient which appears mutually perpendicular to the direction of a longitudinal heat current and the applied magnetic field in a Hall effect experiment. Obviously this effect is analogous to the Ettinghausen effect. The Righi-Leduc coefficient is defined by

$$\frac{\partial T}{\partial y} = - A_{RL} B_z Q_x \tag{1.148}$$

Often, an alternate expression for the Righi-Leduc coefficient, A'_{RL}, is defined by

$$\frac{\partial T}{\partial y} = - A'_{RL} B_z \frac{\partial T}{\partial x} \tag{1.149}$$

and this is used to relate the transverse thermal gradient to the longitudinal thermal gradient when one exists. Two relations which describe the magnitude of the Righi-Leduc coefficients are

$$\frac{A'_{RL}}{A_{RL}} = K_T$$

$$A'_{RL} = R_H/\rho \tag{1.150}$$

d) The Ettinghausen-Nernst effect is due to an electric field which is induced mutually perpendicular to the direction of a longitudinal heat flow and a transverse magnetic field. The Ettinghausen-Nernst coefficient A_{EN} is defined by the relation

$$E_y = - A_{EN} B_z Q_x \tag{1.151}$$

This effect can also lead to errors in Hall effect experiments since the induced field is in the same direction as the Hall field.

A mathematical treatment of these effects would follow along the same lines as the development of the thermoelectric effects, with the inclusion of a magnetic force in the expressions for the electrical and heat current densities. A good treatment is contained in Reference 7, and these effects are also summarized in Reference 44.

1.6 Magnetic Properties of Solids

The magnetic properties of solids is an extremely important area of solid-state physics, with many various practical applications. Magnetic materials have been used as inductors and as switching elements in linear and digital electronic equipment. They have applications in the field of magnetic recording and in microwave equipment. These are just a few of the important applications of magnetic materials. However, the direct use of the magnetic properties of solids in the sensor area has been generally limited. This can be attributed to the simple fact that when one desires to sense a given physical variable, he generally desires that the sensing be accomplished through an electrical means. Thus an electrical signal is desired which is proportional in some manner to the given stimuli. Bulk devices which could employ magnetic materials and which will give an electrical output are resistors, inductors, and capacitors. Of these bulk devices, only an inductor utilizes any unique features of magnetic materials, and generally there would be no advantages as far as sensors are concerned in using magnetic materials in a resistive or capacitive function. The inductive property requires a-c excitation which can detract from its use in a sensor application, although this is not always true. By and large there do not seem to be many unique advantages for using magnetic materials for this particular physical property alone as sensors. For this reason this section will be brief and descriptive in nature.

1.6.1 The Physical Origins of the Magnetic Properties of Solids

As in the study of the dielectric properties of solids the problem is to relate the microscopic sources of the magnetic properties of a solid to the macroscopic measureables which can be observed. The three physical quantities which describe the magnetic behavior of a solid, and which are macroscopic parameters, are the magnetic field intensity, $\overline{H}$, the magnetic flux density, $\overline{B}$, and the magnetization, $\overline{M}$. The magnetization is due to the motion of charged particles in a solid. This is because the motion of charge leads to an electrical current, and current is the source of magnetic field intensity in Maxwell's equations. Generally, the current flow due to individual charges in a solid is thought of as occuring in the form of a current loop. It can be demonstrated that these current loops correspond to a magnetic dipole which can be though of as two "magnetic charges" separated in space in the exact manner in which an electric dipole exist [27]. This non-real magnetic dipole gives rise to the same net magnetization as does the current carrying loop. It also allows many concepts analogous to electrostatics to be applicable in magnetostatics.

Now, the macroscopic magnetization, M, is defined as the average density per unit volume of the magnetic moment in a given direction. The magnetic dipole moment, m, is defined by

$$m = IA \tag{1.152}$$

where I is the current in the loop and A is the area of the loop. Thus

$$M = m/V \tag{1.153}$$

where V is the "volume" of the loop. The total magnetic flux density produced in a magnetic material is the sum of the flux produced by the internal magnetic dipoles plus the external applied magnetic field. This fact then yields the auxiliary Maxwell's equation [27]

$$\overline{B} = \mu_o (\overline{H} + \overline{M}) \tag{1.154}$$

or

$$\overline{B} = \mu_o (1 + \chi_m) \overline{H}$$

where μ_o is the permeability of free space. Here χ_m is called the magnetic susceptibility and Equation (1.144) assumes a linear relation between $\overline{H}$ and $\overline{M}$. The quantity $(1 + \chi_m)$ is often called the relative permeability, μ_r.

In many cases, magnetization is not simply linearly related to magnetic field intensity. Often, this relation is in a tensor form

$$\mathbf{M}_i = \chi_{m_{ij}} H_j \tag{1.155}$$

The form of the magnetic suceptibility tensor has the same form as the permittivity tensor given in Equation (1.22), so that for most materials the tensor aspect of the magnetic properties are not important.

From a quantum mechanical point of view, the magnetic properties of a solid are determined by the state of the angular momentum which the electrons of the constituent atoms possess. This includes both the orbital angular momentum and the spin angular momentum. A substance can possess a magnetic moment in the absence of any applied magnetic fields if the angular momentum of all the electrons does not sum up vertorically to be zero. In this case the material is said to possess a permanent magnetic dipole moment, or a permanent magnetization. Another source of magnetization in a solid comes from the fact that an applied magnetic field can induce magnetic dipole moments, since the field can cause changes in the orbital motion of an electron. The induced orbital motion of an electron in an applied magnetic field leads to what is known as the Larmor precession, and the induced magnetic moment is directly proportional to the magnetic flux density.

An estimate of the size of both the induced magnetic moment and the permanent magnetic moment can be made based on simple calculations [27]. For a single electron the permanent dipole moment is equal to 9.27×10^{-24} amp-m^2. This is called a Bohr magneton. The induced magnetic moment of the same electron in the first Bohr orbit and with an applied magnetic flux density of 1 weber/m^2 (10^4 gauss) gives a magnetic moment of 2×10^{-29} amp-m^2. Thus, for small fields the permanent moment will be much greater than the induced moment. Also, if each atom in a solid possess one Bohr magneton, and if all these magnetic moments are in the same direction, one would expect a magnetization of about 9×10^5 amp/m. This corresponds to an internal flux density of about 1 weber/m^2. These simple ideas give some idea

about the size of magnetization which can result in a solid. However, there are many different varieties of magnetic materials whose magnetic properties vary greatly. The differences in these properties are what distinguish the magnetic materials and are the factors which determine their applicability to sensors.

1.6.2 The Classification of Magnetic Materials

Depending on whether or not a material has permanent magnetic dipole moments, and on the orientation of these dipoles with respect to one another, magnetic materials can be placed into five different categories. These categories are:

1) Diamagnetic materials
2) Paramagnetic materials
3) Ferromagnetic materials
4) Antiferromagnetic materials
5) Ferrimagnetic materials

If a material has no permanent magnetic dipoles, then it is said to be diamagnetic. In this case the diamagnetic susceptibility is very small (its magnitude is much less than unity) and it is negative. This is because of the fact that from Lenz's law the induced current due to a magnetic field produces another field which opposes the time rate of change of the applied field. Classical and quantum mechancial treatments both predict a diamagnetic magnetic susceptibility of

$$\chi_{mo} = -\frac{q^2 Z N \mu_o}{6 m_o} \langle r^2 \rangle \qquad (1.155)$$

where N is the number of atoms per unit volume, m_o is the electron mass, Z is the atomic number of the material, and $\langle r^2 \rangle$ is the mean squared (expectation value) distance of the electrons from the nucleus. Experimentally, the diamagnetic susceptibility is approximately $10^{-6}Z$, which gives $\sqrt{\langle r^2 \rangle} \sim 10^{-8}$ cm. [48] . Also, χ_{mo} should be almost independent of temperature because the only factor that is a function of temperature is $\langle r^2 \rangle$. Due to the small value of χ_{mo}, diamagnetic materials do not have a magnetic behavior which differs much from that of free space.

Paramagnetic substances have permanent magnetic dipoles due to a nonvanishing angular momentum. In the absence of an applied field, the orientation of these dipoles are random so that the net magnetic moment is zero. The dipole orientations at zero applied field intensity tends to be randomized because of thermal fluctuations. If a field is applied which is strong enough to overcome the thermal energy of the system then a net magnetic moment results. For a given field strength one then expects this effect to be larger at lower temperatures since the thermal energy is smaller. The energy density of a dipole in a magnetic field is on the order of the magnetic flux density times the internal magnetic field intensity per dipole. Thus the energy per dipole is on the order of Bm. Thermal energy at room temperature is on the order of kT which is 4.3×10^{-21} joule. To align the random dipoles in the direction of B requires a value of B greater than $(4.3 \times 10^{-21}/\text{m})$ which is about 5×10^2 webers/m^2 or five million gauss. This is a much larger flux density than is normally available in the laboratory. Thus about all the applied field can do is to slightly

reorient the dipoles in the direction of the field. From this discussion one expects the paramagnetic susceptibility to be positive, so that the magnetization is in the direction of the applied field.

There are three important contributions to the theory of paramagnetism in solids. The first of these is called Langevin paramagnetism in which the solid is treated as containing n atoms per unit volume with each atom having a magnetic moment m. A treatment using Boltzmann statistics for the distribution of magnetic moments yields for the magnetization the relation see [13]

$$M = nm\, L\,(x) \tag{1.157}$$

where $x = m\mu_o H/kT$ and H is the applied magnetic field intensity. Here L(x) is called the Langevin function and is given by

$$L(x) = \coth\,(x) - \frac{1}{x} \tag{1.158}$$

As shown in the above discussion x is generally much less than unity so that the magnetization reduces to

$$M \simeq \frac{nm^2\mu_o H}{3kT} \tag{1.159}$$

and

$$\chi_m = M/H = \frac{nm^2\mu_o}{3kT} = \frac{C}{T}$$

The parameter C is called the Curie constant, and Equation (1.159) is called the Curie law. A quantum mechanical treatment can be carried out which includes both spin and orbital angular momentum,

and which for $x << 1$ yields the same result as Equation (1.159). This type of law can be verified for many solids, and is illustrated in Figure 1.51. An approximate value of the paramagnetic susceptibility can be made assuming the magnetic moment is equal to the Bohr magneton and assuming n to be the density of atoms in a solid state

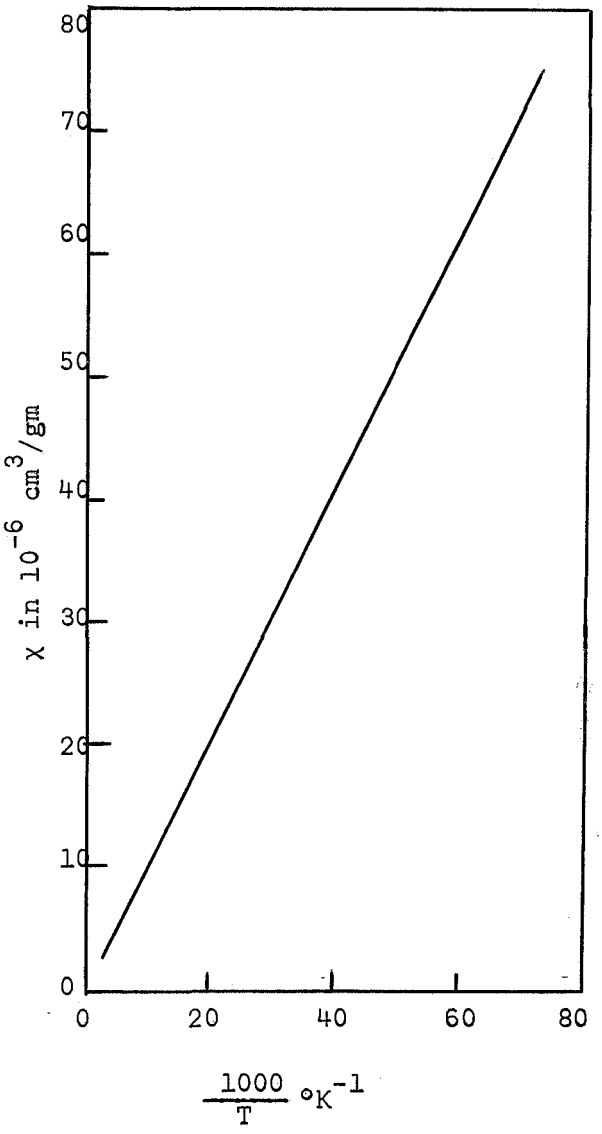


Figure 1.51. Plot of susceptibility per gm versus reciprocal temperature for powdered $CuSO_4.K_2SO_4.6H_2O$, showing the Curie law temperature dependence.

substance. This yields a value of χ_m of about 10^{-3}.

Another type of susceptibility is due to Pauli, and is often called the Pauli spin susceptibility [49]. This type of susceptibility is due to conduction electrons. Here classical results predict a susceptibility which is too large and which has too a strong a temperature variation. When Fermi-Dirac statistics are used, these errors are corrected and the susceptibility becomes

$$\chi_{m_n} \simeq \frac{nm^2\mu_0}{kT_F} \tag{1.160}$$

where $k\,T_F = E_F$, the Fermi energy. Since $kT << E_F$ over a large temperature range, Equation (1.159) does predict a much reduced value for χ_m, and it also predicts only a very minute temperature variation.

Another contribution to the theory of paramagnetism has been given by Van Vleck [50]. This theory is purely quantum mechanical and has no classical analogure. It takes into account the fact that excited states of atoms can have non-zero orbital magnetic moments, and thus should contribute to the susceptibility. Generally, one expects the excited states to be only very sparsely populated, and for this case the Van Vleck theory predicts a susceptibility of [13]

$$X = K\left(\frac{n}{\Delta}\right) \tag{1.161}$$

where Δ is the energy difference between the excited and ground state. K is determined by the matrix element between the ground

state and the excited state. This result was determined from second order perturbation theory, and can be very difficult to apply to actual solids in practice.

Figure 1.52 illustrates the characteristic magnetic susceptibilities of both diamagnetic and paramagnetic materials as a function of temperature. From this figure and from the approximate calculations which have been made the obvious conclusion to be drawn is that for ordinary diamagnetic and paramagnetic substances the susceptibility is much less than unity. Thus the permeability of these materials can be taken to be the permeability of free space for most application. Thus usually neglected when materials for magnetic applications are

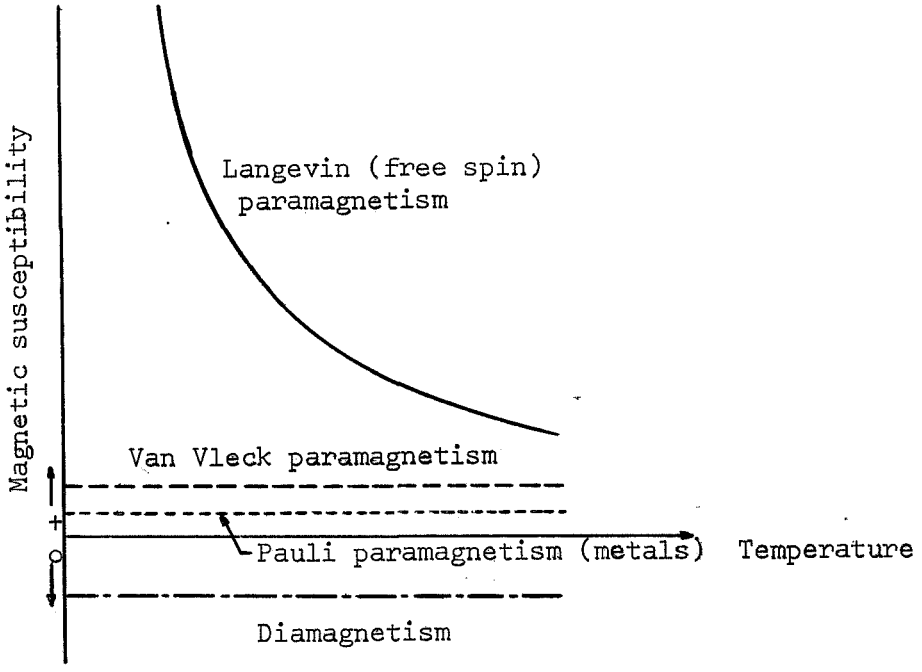


Figure 1.52. Characteristic magnetic susceptibilities of diamagnetic and paramagnetic substances.

sought, and generally there seems to be no unique sensor applications for this type of solid. However, this should not imply that these materials are of academic interest only. There are instances where the minute value of magnetic susceptibility of these materials can perform marvels [27]. For example, the development of the solid-state maser has depended on the use of a single crystal of a paramagnetic substance, such as the hydrated sulfates, oxides, halide salts, and nonmagnetic ruby doped with magnetic impurities. The maser is used as a very low-noise microwave amplifier, or as an oscillator. Another application of paramagnetic materials is in the cryogenic area. Here the process of cooling by the adiabatic demagnetization of a paramagnetic salt has become a universal method for achieving temperatures near absolute zero. This technique has yielded temperatures as low as 10^{-3} °K [31].

The last three classifications of magnetic materials, ferromagnetic, antiferromagnetic, and ferrimagnetic, are the most interesting as far as sensor applications are concerned. In a ferromagnetic substance, the permanent magnetic dipoles are strongly aligned, and a spontaneous magnetization exist in the absence of any applied magnetic fields, in a manner analogous to ferroelectric materials. Thus, one expects that for these materials the relation of applied magnetic flux density to induced magnetic field intensity should be a hysterisis loop. A typical B-H curve is shown in Figure 1.53 . This curve is characterized by the parameters noted on the graph. These parameters are the saturation flux density, B_s, the remanent flux density, B_r, and the coercive field intensity, H_c. The remanent flux density and the coercive field

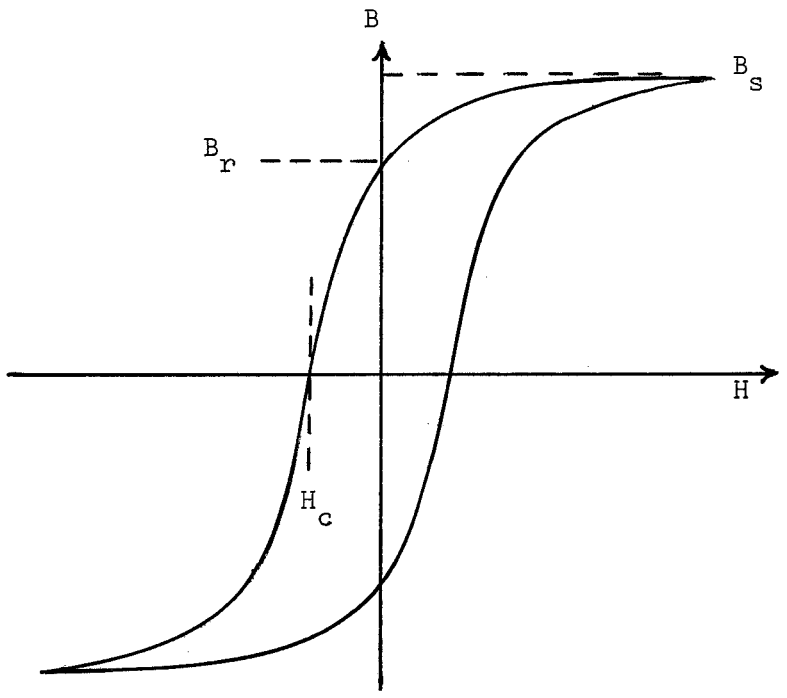


Figure 1.53. Hysteresis loop of ferromagnet.

intensity are structure sensitive quantities which depend quite strongly on the presence of voids, inclusions, internal strains, and other types of materials imperfections. The saturation flux density is structure insensitive and depends on the chemical composition of the material, and the temperature [31].

A number of magnetic materials contain more than one species of magnetic ions or atoms in their crystalline composition. The ions located at different crystal sites can affect the net magnetization differently, and for this reason it is often convenient to divide the magnetic ions into subgroups according to their positions in the lattice. These subgroups are often called sublattices. The magnetic ions in the different sublattices can interact with each other, and the magnetic

moments of the different sublattices can be aligned either parallel or anti-parallel to each other. In ferromagnetic substances the preferred arrangements are parallel, while in antiferromagnetic and ferrimagnetic materials the preferred alignment is anti-parallel. The reason for this will be discussed in a later section. In an antiferromagnetic material the net magnetization is zero in the absence of an applied magnetic field due to an exact cancellation of the magnetic moment of one sublattice by the magnetic moment of another sublattice. In a ferrimagnetic material, the moments on adjacent sublattices do not exactly cancel, yielding a net spontaneous magnetization. Both antiferromagnetic and ferrimagnetic materials have a critical temperature, called the Neel temperature, above which the orderly arrangement of magnetic moments on adjacent sublattices cease to exist. Above this temperature the moments are randomly oriented and the materials behave as paramagnets. Below the Neel temperature the material is said to be in an ordered state, while above this temperature the material is said to exist in a disordered state, as far as the magnetic properties are concerned.

Figure 1.54 illustrates the expected temperature dependence of the magnetic susceptibility and the spontaneous magnetization for paramagnetic, antiferromagnetic, ferromagnetic, and ferrimagnetic materials. In these figures the arrows represent the magnetic moments of the materials, and the critical temperatures of interest are shown. The remaining sections will be devoted to discussions of these curves for ferromagnetic, antiferromagnetic, and ferrimagnetic materials. Examples of sensor applications will be given.

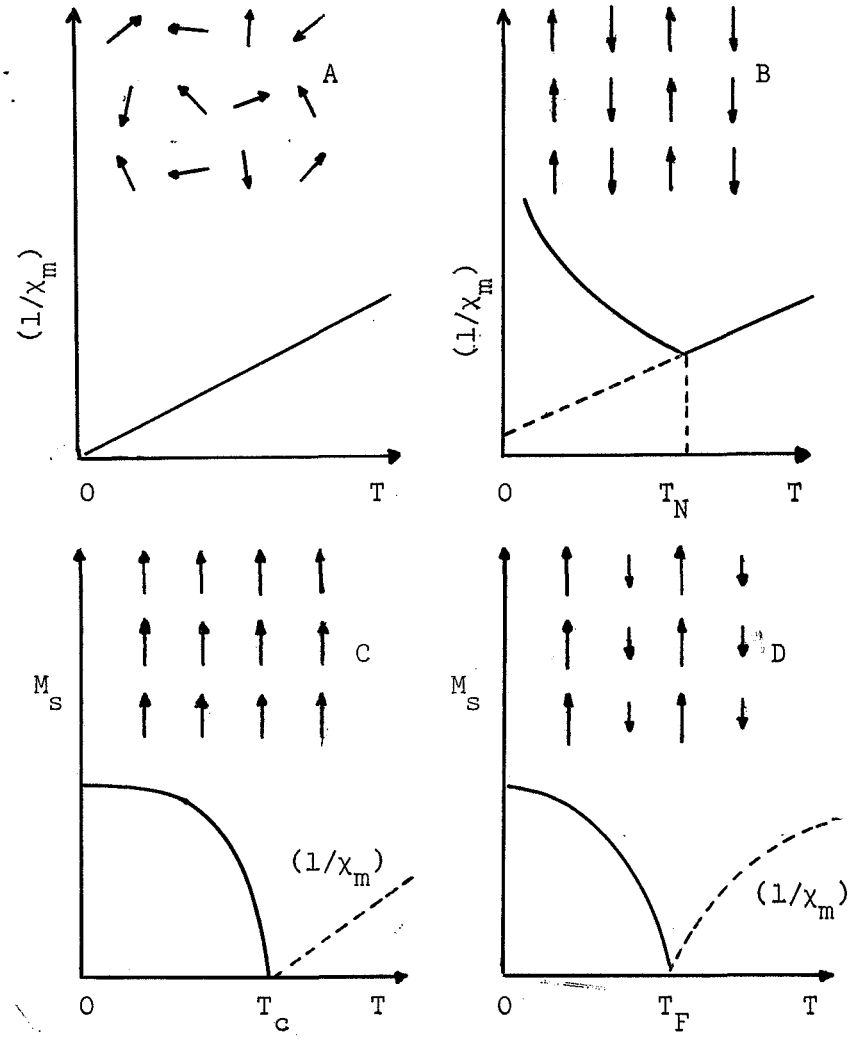


Figure 1.54. The low temperature ordering (if any) of neighboring dipoles, and the consequent behavior of spontaneous magnetization and/or susceptibility, for A, paramagnetism, B, antiferromagnetism, C, ferromagnetism, and D, ferrimagnetism.

1.6.3 Ferromagnetic Materials.

It has been shown quantitatively that the alignment of magnetic dipole moments in a solid requires extremely large externally applied fields. Since ferromagnetic substances possess a spontaneous magnetization in the absence of any applied magnetic fields, the permanent dipole magnetic moments must be aligned to a large extent. The question then arises as to where the energy comes from which is required to align these dipoles. Weiss originally postulated that this energy comes from a very large internal field which has come to be known as the Weiss molecular field [49] . This field represented a magnetic interaction between neighboring magnetic ions. Weiss postulated that the field in a material consisted of the external applied field and the internal molecular field (as in Equation (1.155)) which is of the form

$$H_T = H_{ext} + \alpha M \tag{1.162}$$

where α is a very large constant, independent of H_{ext}. Here H_T is the total magnetic field intensity, and H_{ext} is the applied magnetic field intensity. Then Weiss assumed that the magnetization is still given by the Langevin relation of Equation (1.156). Thus

$$M = nm\ L\ (x_T) \tag{1.163}$$

where now $x_T = m\mu_o H_T/kT$. Obviously for a fixed H_{ext}, Equation (1.163) is a transcendental equation for M, which can possess solutions even when H_{ext} is zero. Also, for temperatures greater than some critical

temperature solutions of Equation (1.163) do not exist. This can be seen from Figure 1.55 which shows a plot of M versus x_T for $H_{ext} = 0$, along with a plot of the straight line $M = (kT/\alpha m\mu_o)\, x_T$ for different values of the temperature. At low temperatures, the slope of this straight line is small enough so that it intersects with the Langevin curve to give a solution of Equation (1.163). At some critical temperature the slope of the straight line is equal to the slope of the Langevin function at $x_T = 0$, and for temperatures greater than this, the only solution of Equation (1.163) is at $x_T = 0$ which requires M=0. The critical temperature is called the Curie temperature and is defined by

$$nm \frac{dL(x_T)}{dx_T} \Big| x_T = 0 \quad = \frac{kT_c}{\alpha m \mu_o} \tag{1.164}$$

which gives a Curie temperature of

$$T_c = \frac{n\, m^2 \mu_o \alpha}{3k} \tag{1.165}$$

It is obvious that the solutions of (1.163) for $H_{ext} = 0$ gives the spontaneous magnetization, and that the spontaneous magnetization is a function of temperature which approaches a constant value at low temperatures. Below the Curie temperature a universal magnetization curve can be obtained by the above procedure and is shown in Figure 1.56 , along with experimental data for iron, nickel, and cobalt. As can be seen the agreement is excellent. Above the Curie temperature the magnetization is not truly zero and the ferromagnetic

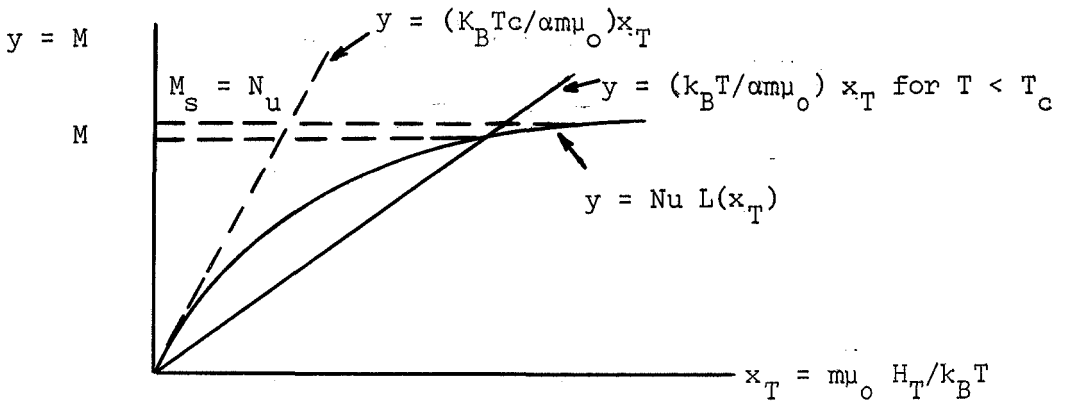


Figure 1.55. Diagram illustrating how Eq. (1.163) can be solved graphically for $H_{ext} = 0$. A solution for the value of spontaneous magnetization exists only when the straight line whose slope is a function of temperature intersects the curve.

material is paramagnetic with a small magnetic susceptibility. Expanding the Langevin function for $x_T << 1$ gives

$$\frac{M}{H_{ext}} = \chi_m = \frac{C}{T - T_c}, \quad T > T_c \tag{1.166}$$

where T_c is the Curie temperature and $C = T_c/\alpha$ is called the Curie constant. Equation (1.166) is called the Curie-Weiss law, and predicts paramagnetic behavior above the Curie temperature.

The phenomenogical Curie-Weiss law predicts most of the essential features of paramagnetic materials. It does not explain, however, the origin of the molecular field. The Weiss field, αM, is of electrostatic origin and can only be explained by quantum mechanics. This will not be done here and the reader is referred to Reference 27 for a discussion.

Typical values of α are between 10^3 - 10^4, so that the internal field is indeed large enough to orient the dipoles at most temperatures. The discussion of ferromagnetic materials is concluded by the data presented in Table 22. This table gives some typical values of the properties for several ferromagnetic materials.

Ferromagnetic materials have had some application in the sensor area. One such application has been in the measurement of pressure in shock waves [51]. Here one side of a magnetic core is hit by a shock wave. A coil is wound around the core, and due to the fact that the remanent magnetic flux density, B_r, is a function of pressure, a voltage is induced in the coil because of the changing flux in the core. The magnetic materials used were 50% nickel-iron alloys, 49% - 49% - 2% iron-cobalt-vanadium alloys.

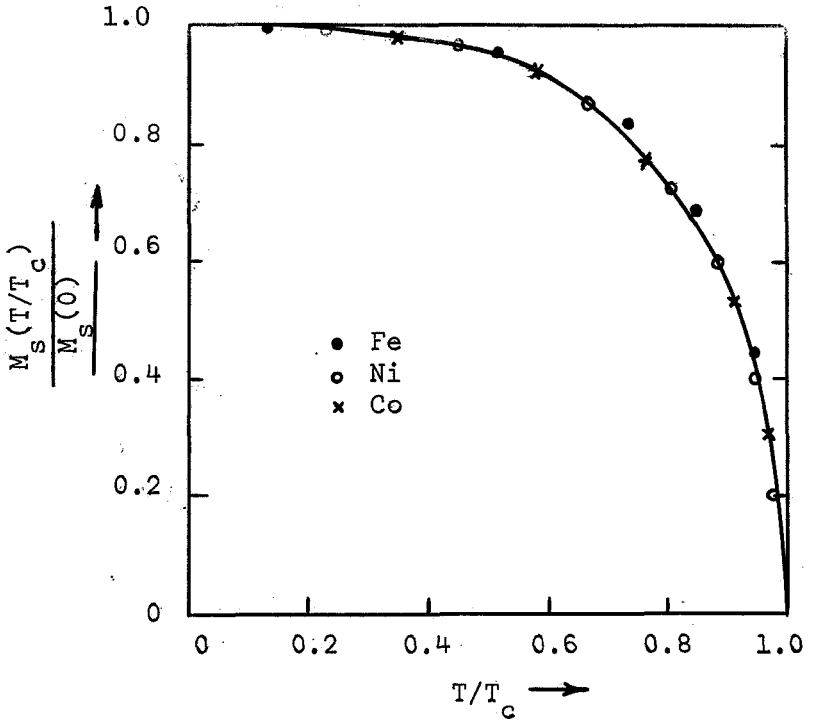


Figure 1.56. Relative magnetization versus T for the three ferromagnetic elements of the iron group.

Table 22 The spontaneous magnetization, Curie temperature, and effective number of Bohr magnetons in a selected number of magnetic materials

Substance	Spontaneous magnetization (10^3 amp/m)		Curie temperature, T_c °K	Effective No. of Bohr magnetons, n_{eff}(0°K)
	300°K	0°K		
Fe	1707	1752	1043	2.22
Co	1400	1446	1400	1.72
Ni	485	510	631	0.606
Gd	1090	1980	289	7.10
Dy	....	1830 (80°K)	105	
$NiOFe_2O_3$	240	.....	863	2.3
Cu_2MnAl	430	580	603	(4.0)
MnAs	670	870	318	3.4

The magnetic core operates at extremely high voltages and currents, of up to 300kv and 100 amp. Needless to say the times of interest are very short. These experiments defined a "transducer effect" which is the ratio of output energy to the energy stored in the magnetic core. This is certainly a unique application of ferromagnetic materials as a sensor.

Another very interesting application of ferromagnetic materials as sensors has been in the area of magnetic field measurements. Cylindrical ferromagnetic rods have been used to increase the sensitivity of Hall effect semiconductor probes in the measurement of the earth's magnetic field. Figure 1.57 illustrates the arrangement used for this purpose. From Equation (1.162) it is obvious that the magnetic

field intensity in the ferromagnetic rods is greater than that external to the rods. If a Hall effect generator is placed in a small air gap between the two rods then a considerably higher magnetic field intensity is incident on the Hall probe and an amplification effect can be achieved. In this case an increase in the effective sensitivity of the magnetic field measurement is achieved. A theoretical calculation of the amplification factor , ν, versus the ratio of rod length to rod diameter, p, with rod permeability as a parameter is shown in Figure (1.58). It is clear that large gains can be achieved. These curves were calculated for a continuous rod, and the fact that the rod has a small air gap in it tends to reduce the gain achievable in practice. Large gains are still obtained in these devices, and magnetic field sensors have been constructed in this manner which have a gain of larger than 400 [32]. The device has an open circuit sensitivity of 40 mV/gauss with an internal resistance of 2 ohms. The temperature coefficient of the open circuit voltage was 0.2%/°C. The ferromagnetic material used is called Mu-metal, a trade name for a particular type of material. Probes have been made in this manner which are capable of measuring the earth's magnetic field [32].

The Faraday effect in ferromagnetic materials can also be used to measure magnetic field intensity. This effect occurs due to the rotation of the plan of polarization of polarized light incident on a crystal when it is subjected to an applied magnetic field. The angle of rotation is proportional to the magnetic field intensity, and the use of a ferromagnetic material greatly enhances the effect due to the intrinsic magnetization in the same manner as was done with

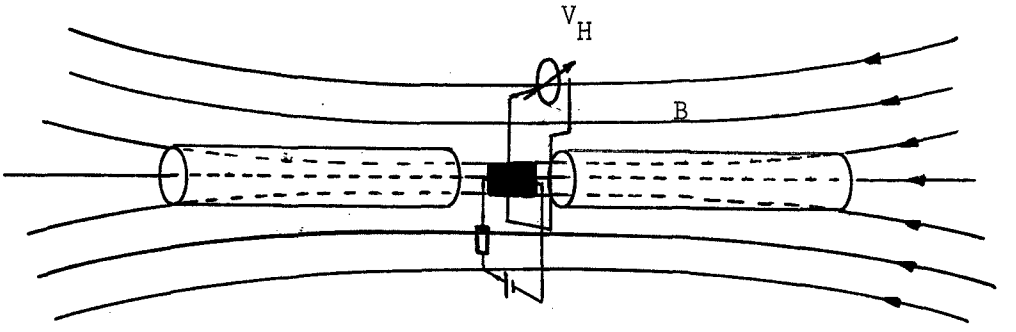


Figure 1.57. Principle of measurement of smallest magnetic fields with a Hall generator and ferromagnetic rods.

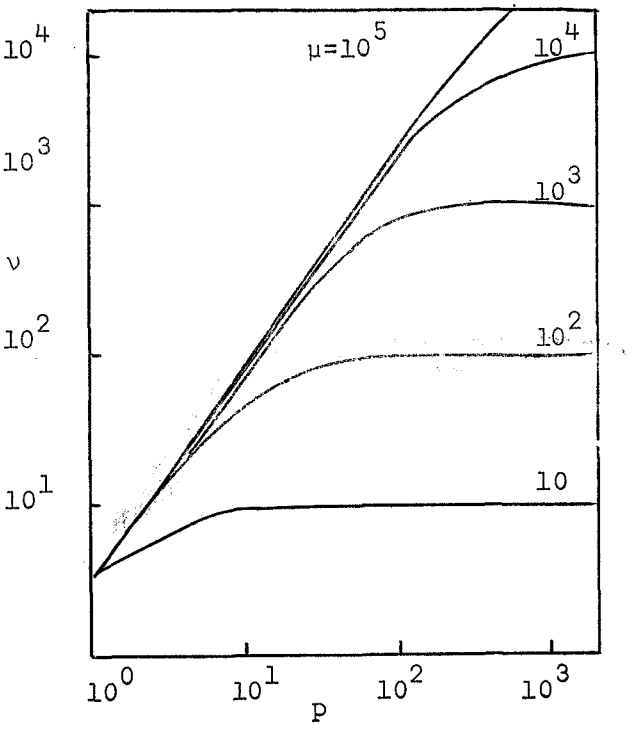


Figure 1.58. Enhancement of magnetic induction by cylindrical rods as a function of the ratio p=length of rod/rod diameter, with the permeability μ as parameter.

the Hall effect generator in the last paragraph. Thin magnetic films have been shown to give rise to much larger angles of rotation [52]. This study proposed thin films of iorn, nickel, and cabott, along with a gallium arsenide junction laser and silicon photodiodes and phototransistors as an optical source and detector, respective-

ly, as a system to measure extremely low magnetic flux densities of about 1×10^{-6} gauss or lower. According to Weiss, an indium arsenide Hall generator can measure an alternating magnetic field of about 6×10^{-6} gauss, and a d-c magnetic field of 6×10^{-4} gauss. Generally, if long term stability (several hours) is required, the lower limit of a Hall generator seems to be about 2×10^{-3} gauss, with reasonable accuracy and reproducibility. The above figure for the Faraday effect may be achievable, but the study indicated did not discuss noise problems and practical measurement problems and it is doubtful that fields as low as 1×10^{-6} gauss can be measured with any degree of confidence.

While on the subject of magnetic field measurements, it seems worthwhile to mention that another useful way to measure small magnetic fields very accurately is by the use of nuclear magnetic resonance. This technique utilizes the fact that a sample of a magnetic material will absorb energy if it is placed in an alternating high frequency magnetic field superimposed on a steady d-c magnetic field. This absorption will occur at one specific frequency of the alternating field and this frequency is proportional to the magnitude of the d-c magnetic field as well as the magnetic moment involved in the absorption process. Nuclear magnetic resonance is thus a good technique for investigating magnetic materials as well as for measuring magnetic fields. With this technique one can measure fields in the range of 3 gauss to 20 gauss with an accuracy of one part in 10^6 [53]. While the sensitivity of this technique is not as good as those previously mentioned, the accuracy is outstanding.

This concludes the discussion of ferromagnetic materials. It is obvious that they are both interesting and useful as sensors. There exists a great deal of information concerning magnetic materials, and ferromagnetic materials in particular. For a good up-to-date review of the theory and practical aspects of these materials the reader is referred to the volumes by Beckowitz and Kneller [54].

1.6.4 Antiferromagnetic Materials

The diffraction of neutrons has confirmed the early hypothesis that the magnetic moments of ions adjacent sublattices of magnetic materials can indeed be anti-parallel, and that the resulting net magnetic moment can be zero [55]. This type of antiferromagnetic material is quite similar to that of ferromagnetism, with certain important differences. Most antiferromagnetic materials are compounds with a good deal of ionic bonding. The cations are separated by much larger distances than those in ferromagnetic materials, and the molecular field must be to a large extent due to the an interaction with non-magnetic anions. This point will not be discussed here and the reader is referred to Reference 55.

It can be shown that the molecular field in an antiferromagnetic material is negative [56], and the field and magnetic moments of all the sublattices must be taken into account. The procedure is more complicated than for ferromagnetic materials, but the techniques are basically the same in the phenomenological theory. It can be shown that there exists a critical temperature above which

the material has no spontaneous magnetization. However below this temperature the effect of an applied field on the magnetization, and thus the resulting magnetic susceptibility, has a unique behavior . The temperature which distinguishes these regions of different behavior is called the Neel temperature, T_N, of antiferromagnetic materials. Above this critical temperature the magnetic susceptibility can be described by

$$\chi_m = \frac{2C}{T + \Theta}, \quad T > T_N \tag{1.167}$$

where C is the Curie constant given by $C = T_c/\alpha$ and T_c is the Curie temperature given by Equation (1.165). Θ and T_N are different quantities and Θ is called the paramagnetic temperature for an antiferromagnetic material. Both T_N and Θ are related to C, but their ratio is independent of C. It can be shown that

$$\frac{\Theta}{T_N} = \frac{1 + B}{1 - B} \tag{1.168}$$

where B is a positive constant which is less than unity. Thus Θ/T_N is greater than unity. This ratio is a measure of the importance of next-nearest nighbor interactions in a magnetic material. Table 23 gives a list of several antiferromagnetic materials and shows Θ, T_N, Θ/T_N, and the ratio of susceptibilities at Θ and T_N. Also Figure 1.59 illustrates the behavior of magnetic susceptibility for a typical antiferromagnetic material. Figure 1.60 further shows the distinguishing features of the magnetic susceptibility for paramagnetic, ferromagnetic, and antiferromagnetic materials.

Table 23 Data for some Antiferromagnetic Compounds[a]

Compound	Crystal Structure	T_N(°K)	θ(°K)	$\frac{\theta}{T_N}$	C_{mole}	$\frac{\chi_p(\Theta)}{\chi_p(T_N)}$
MnO	f.c.c.	122	610	5.0	4.40	0.69
FeO	f.c.c.	185	570	3.1	6.24	0.77
CoO	f.c.c.	291	280	0.96	3.0	--
NiO	f.c.c.	515	--	--	--	0.67
MnS	f.c.c.	165	528	3.2	4.30	0.82
MnF_2	rutile	74	113	1.5	4.08	0.75
FeF_2	rutile	85	117	1.4	3.9	0.72
CoF_2	rutile	40	53	1.3	3.3	--
NiF_2	rutile	78	116	1.5	1.5	--
MnO_2	rutile	86	--	--	--	0.93
Cr_2O_3	rhombohedral	307	1070	3.5	2.56	0.76
α-Fe_2O_3	rhombohedral	950	2000	2.1	4.4	--
FeS	hex. layer	613	857	1.4	3.44	--
$FeCl_2$	hex. layer	24	-48	-2.0	3.59	<0.2
$CoCl_2$	hex. layer	25	-38.1	-1.5	3.46	∿0.6
$NiCl_2$	hex. layer	50	-68.2	-1.4	1.36	--
$FeCO_3$	complex	57	--	--	--	0.25
$CuCl_2.2H_2O$	rhombohedral	4,3	5	1.16	--	--
$FeCl_2.4H_2O$		1.6	2	1.2	3.61	--
$NiCl_2.6H_2O$		5.3	--	--	--	--

aFor a list of references see T. Nagamiya, K. Yosida, and R. Kubo, Advan. Phys. 4, 1 (1955). See also R. D. Pierce and S. A. Friedberg, J. Applied Phys. 32, 66S (1961); T. Haseda and E. Kanda, J. Phys. Soc. Japan 12, 1051 (1957).

1.6.5 Ferrimagnetic Materials

Ferrimagnetic materials are ferromagnetic materials for which the spontaneous magnetization at T=0°K does not correspond to parallel alignment of the magnetic moments of the individual paramagnetic ions [13]. This is true even in light of the experimental evidence that these individual ions have their normal magnetic moments. For example, if all the magnetic moments of magnetite ($Fe_3\ O_4$) were parallel, then each $Fe_3\ O_4$ unit cell should contribute 14 Bohr magnetons to the total spontaneous magnetization. The observed experimental value is only about four. However, if it is hypothesized that the moments of the Fe ions are antiparallel to each, then this discrepancy can be accounted for. Neutron diffractions experiments verify that this hypothesis is indeed correct.

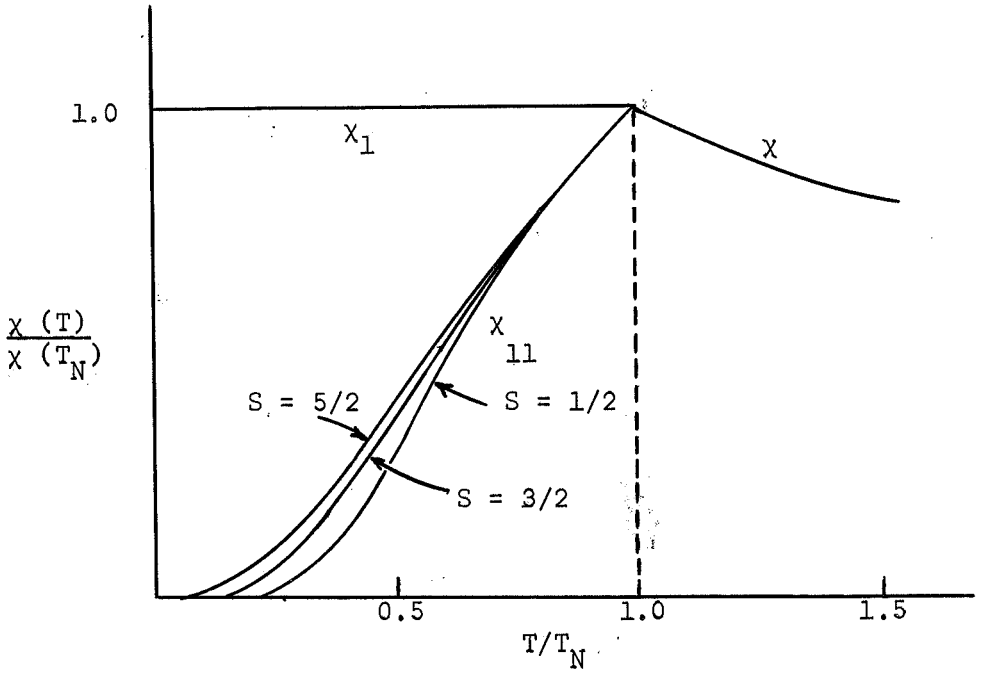


Figure 1.59. The susceptibility of an antiferromagnetic material as a function of temperature in reduced units.

Theoretically, the same interactions which lead to antiferromagnetism are responsible for ferrimagnetic behavior. However, for ferrimagnetic materials the net magnetic moment of the sublattics is non-zero, and a spontaneous magnetization results. This is the basic difference between ferrimagnetic and antiferromagnetic materials. The spontaneous magnetization vanishes at some finite temperature which is called the Neel ferrimagnetic temperature, T_{FN}. Some authors call this temperature the ferrimagnetic Curie temperature [13]. Generally, no one Curie constant can be defined since each sublattice has a different number of magnetic ions whose magnetic moments differ. For a material such as magnetite which has two sublattices the Neel ferrimagnetic temperature is given by

$$T_{FN} = \alpha \, (C_A \, C_B)^{1/2} \tag{1.169}$$

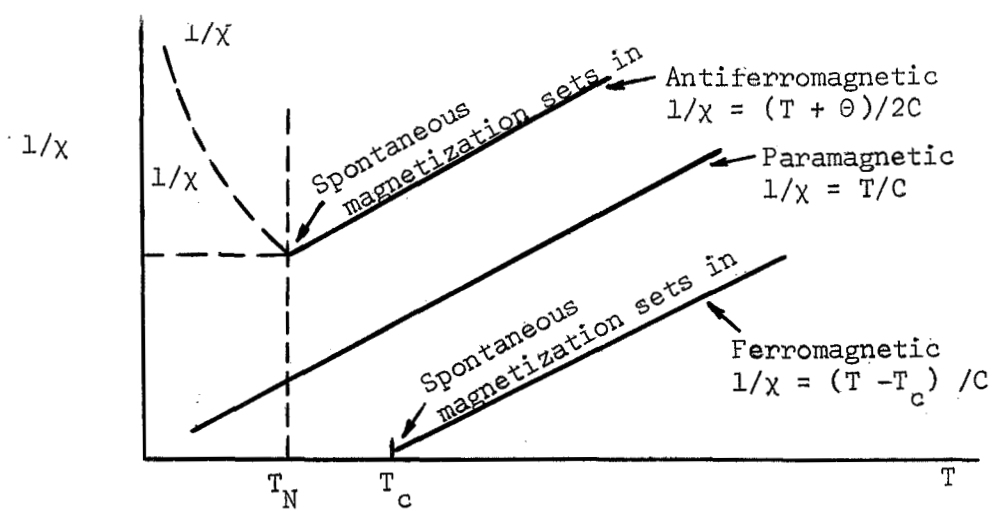


Figure 1.60. Comparison of the behaviors of paramagnetic susceptibility as a function of temperature in antiferromagnetic, ideal paramagnetic, and ferromagnetic materials.

where C_A and C_B are given by Equation (1.159), but with n and m replaced by n_A, m_A and n_B, m_B, respectively. It can be further shown that the susceptibility above T_{FN} is given by the following eqation for a material with two sublattices [13]

$$X_m = \frac{(C_A + C_B)\ T - 2\alpha C_A\ C_B}{T^2 - T_{FN}^2}\ ,\ T > T_{FN} \qquad (1.170)$$

The curvature of X_m versus T is characteristic of a ferrimagnetic material. Figure 1.61 shows experiment data for magnetite, and the unique behavior of the susceptibility of this material is quite apparent.

The most common, and probably the most important, class of ferrimagnetic materials are knwon as ferrites. These materials can be described by the chemical formula $XO \cdot F_{e2}O_3$, where X is a metallic cation of valence two. Typical divalent metals used in ferrite compounds are Mn, Fe, Co, Ni, Cu, Zn, Mg and Cd. Quite often, in commercial applications, mixed ferrites are used which enhance certain desirable properties and suppress other undesirable ones. The mixed systems of Ni-Zn, Mn-Zn, and Mg-Zn have been extensively studied and considerable information is available for these systems [13] . The commercial ferrites are ceramic semiconductors, with very high electrical resistivities. This is quite different from the electrical behavior of most ferromagnetic materials which have metallic conduction. Thus the ferrite are a unique class of magnetic materials in this respect, and their usefulness arises mainly from the fact that they are poor electrical conductors.

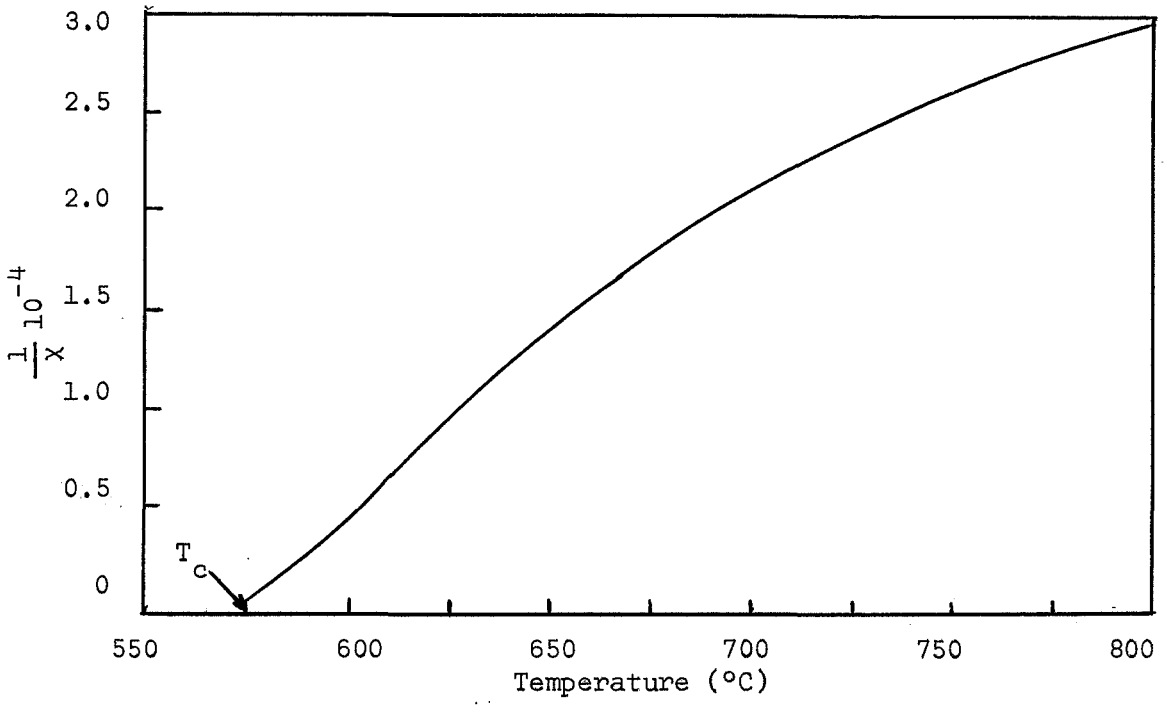


Figure 1.61. Reciprocal susceptibility of magnetite, $FeO.Fe_2O_3$, above the Curie temperature.

The application of ferrites in the electronics and electrical areas are numerous, both in the analog and digital fields. Their use covers the requency range from very low frequencies to well into the microwave regions of the spectrum. Since ferrites have large permeabilities they can be used to construct small size inductors. Their small conductivities allow inductors constructed of ferrites to be used at high frequencies. Ferrites are also useful as transformer materials since their low conductivity limits eddy current loss. Ferrite rods can be used as magnetic core delay lines with small size and high characteristic impedance in video systems. In the microwave area ferrites were the first materials developed which exhibited non-reciprocal signal propagation characteristics of

sufficient magnitude to make them useful in communications systems. This unique property is again related to the interaction of the materials electron spin system with applied magnetic fields. This effect has provided isolators, switches, circulators, and modulators for single-side-band signals without which the microwave engineer would be at a servere handicap. Ferrites have been used in traveling wave tube and kylstron signal sources. Also, one should not forget the ferrite rod antenna, which is used almost exclusively as a receiving antenna in commercial a-m and f-m radio receivers. The use of essentially square hysteris loop ferrite materials have furnished devices for use in signal wave-shaping, timing, gating, modulation, and information storage (memory) which are almost exclusively in computers which have been constructed since the early 1960's. Ferrites have also found application in the field of magnetic recording.

With all their many and varied applications in electronics, the ferrite materials do not seem to have been applied to the sensor area very often. It is not certain what advantages ferrites would have over other materials used in sensor applications. Probably their most unique property could be in their switching characteristics due to the hysteris loop. This could lead to their adaption to digital sensors. One specific example of this has been in the development of a high speed ferrimagnetic pressure sensor [57]. The purpose for this work was to build a sensor whose response time is compatible with present high speed data acquisition systems. The pressure sensor used a ferrite core operated as a high-repetition-rate pulse transformer. The amplitude of the output pulse from the

device is a function of pressure (see section 1.6.6). The switching time of this device was about 0.7 μ sec, which is three orders of magnitude smaller than the response time of other types of pressure sensors. The usable pressure range of the transducer was 0-20,000 psi. The output voltage for the particular device studied varied from about 500 mv at 0 psi to 200 mv at 20,000 psi. This variation was linear and yields as sensitivity of 15 μ volt/psi. A typical diaphram pressure sensor using a silicon strain gage bridge would have an output of about 35 μv/psi. The pressure sensor operated from -55°C to + 120°C with a ± 2% change in linearity. It could be operated up to 400°C, but the deviation from linearity increased significantly. Another feature of this type of pressure sensor is that its characteristics are insensitive to radiation which makes it attractive for space applications.

Other direct applications of ferrites as sensors seem to be uninvestigated. They have been used in conjunction with Hall effect generators in much the same manner in which ferromagnetic materials have been used [32]. It appears that the use of ferrites as digital sensors in the manner in which the pressure sensor discussed in the last paragraph was used is probably the most promising aspect of ferrites in sensors.

1.6.6 Magnetostriction

The spontaneous magnetization of a ferromagnetic or ferrimagnetic material is accompanied by an elongation or contraction of the crystal in the direction of the magnetization. This phenomena is called magnetostriction. The phenomena can be described by the

equation [18]

$$\varepsilon_i = Q_{ij} H_j \qquad (1.171)$$

where ε_i are the strain components, H_j the magnetic field intensity components, and the Q_{ij} are called the magnetostriction coefficients. A converse effect also exists whereby the state of magnetization can be changed by applying an external stress. An example of this was illustrated when the ferrite pressure sensor was discussed in the last section. Figure 1.62 shows the strain versus magnetic field intensity for several ferromagnetic metals. Recall that one oersted corresponds to 1 gauss of magnetic flux density in free space. Notice that in Figure 1.62 , the strain saturates as the magnetic field intensity is increased. The saturation value of strain is called the saturation magnetostriction, and is the number quoted in the literature. For the ferromagnets, it can be seen that the saturation magnetostriction is very small, generally less than 100 μ strain. However, some of the ferrites have large values of saturation magnetostriction. For example, cobalt ferrite ($C_oO.\ Fe_2O_3$) has a value of around 800 x 10^{-6} in/in [58] , while a specimen of polycrystalline MnBi has been found to have a value of about -250 x 10^{-6} in/in. Of course the magnetic field intensity required to produce saturation of the magnetostriction is large, being on the order of 22,000 oersted for mn Bi [44] . Magnetostriction in ferrite materials has been investigated only to a very limited extent.

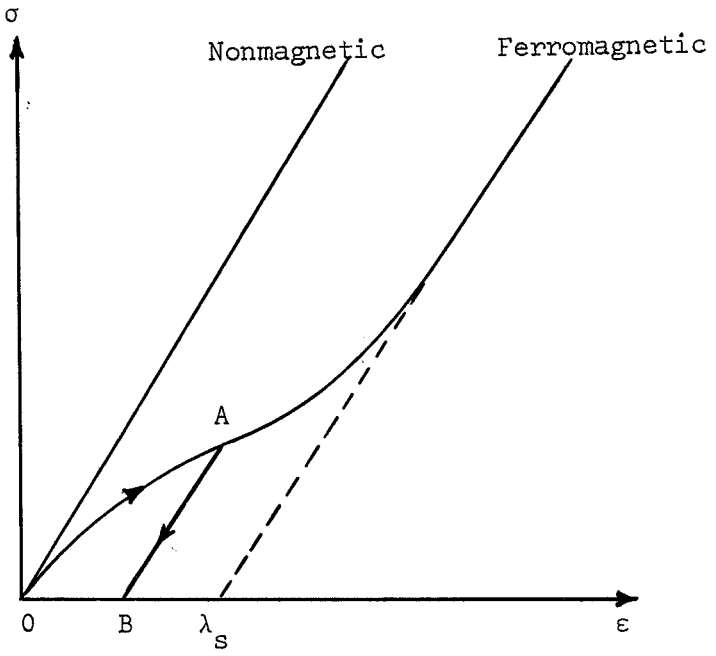


Figure 1.62. Stress-strain relation for ferromagnetic material compared with that for a nonmagnetic material of equal stiffness.

Magnetostrictive effects causes the normal stress strain relation obtained in other elastic materials to become non-linear. This is illustrated in Figure 1.63 . When a stress is applied to a magnetostrictive material, the resulting total strain is composed of two parts,

$$\varepsilon = \varepsilon_e + \varepsilon_m \tag{1.172}$$

where ε_e is the ordinary elastic strain, and ε_m is a "magnetic strain" which is caused by the reorientation of magnetic domains in the crystal. In Figure 1.62 , λ_s denotes the saturation magnetostriction. Note that there is a hysterisis effect in the stress-strain curve, which is due to ε_m.

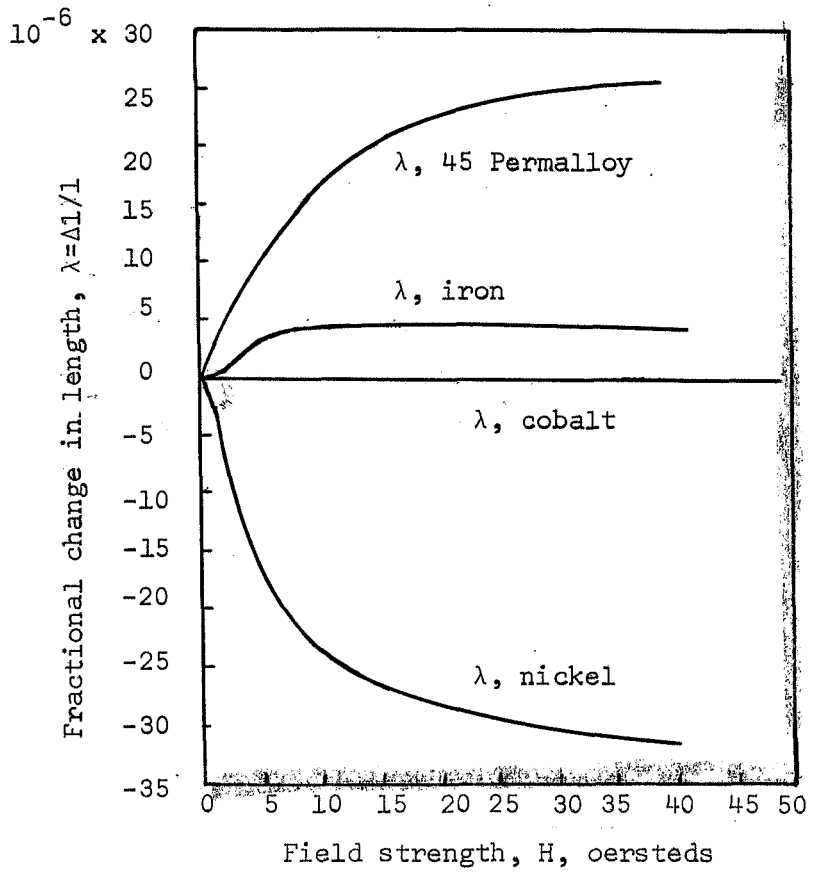


Figure 1.63. Magnetostriction of some common materials, showing expansion or contraction.

The effects of stress on the magnetization properties of a ferromagnetic material is illustrated in Figure 1.64. Here the B-H curve is shown for permallory (Ni - Fe alloy) with applied stress as a parameter. The effect is significant, with the most sensitive portion of the curves occuring at small values of magnetic field intensity.

Another good examples of the use of the magnetostrictive effect has been in the development of a stress sensor designed to tensile test specimens of various materials while under the influence of large hydrostatic pressures [59]. This sensor was used to study the mechanical properties of materials under large hydrostatic

pressures. Thus a given material would be submected to hydrostatic pressures of up to 15,000 kg/cm^2, and their stress-strain relationship measured while biased hysrostatically. This places some severe restrictions on the pressure sensors. The sensor was made from a 13% alluminum-ion alloy cylinder with a-c bridge arrangement for detection. A dummy sensor was used to compensate for temperature and hydrostatic pressure variations. With this system, output voltages of between 50-300 mv could be obtained for tensil loads of from 25-125 kg, with a bridge

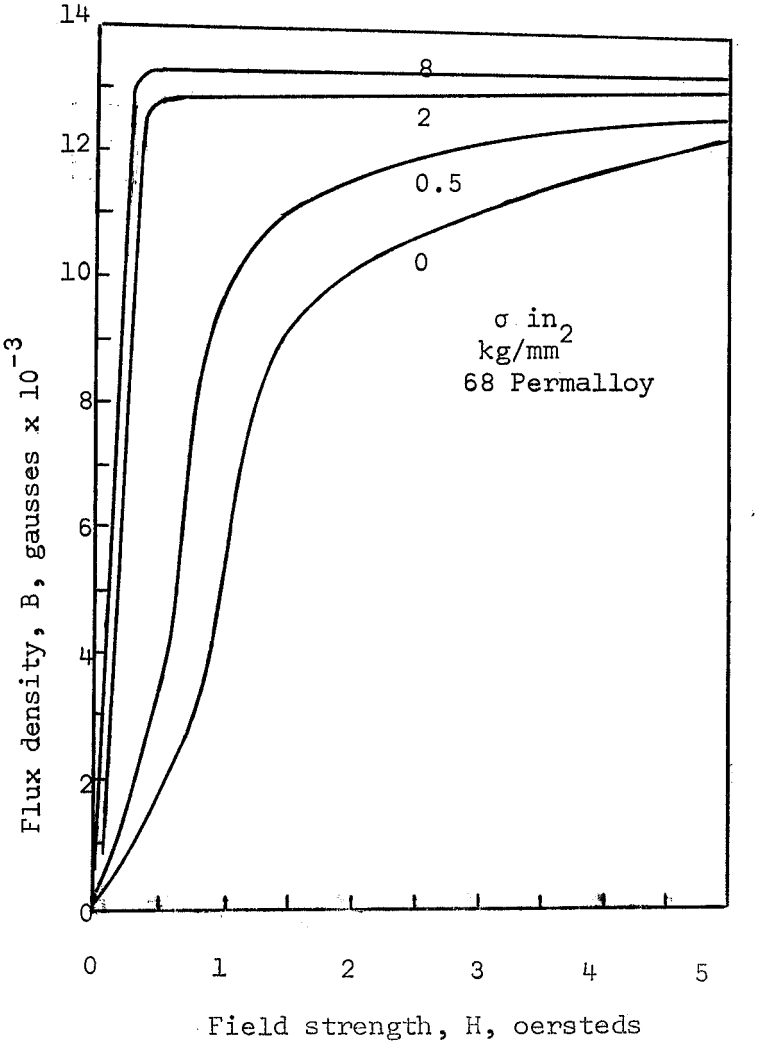


Figure 1.64. The influence of stress on the magnetization of Permalloy.

excitation of 10 volts. The bridge could calibrated to have a constant response (output voltage versus load) curve independent of hydrostatic pressures up to 13,400 kg/cm^2. The sensor provided excellent reproducibility and stability. The linearity of the response curve was poor and some hysterisis was observed.

1.6.7 Pyromagnetic Effect

The pyromagnetic effect is the effect of temperature on the spontaneous magnetization of a ferromagnetic or antiferromagnetic material. It is directly analogous to the pyroelectric effect discussed in section 1.5.7. The pyromagnetic coefficient, q_p, is defined by

$$q_p = \frac{dM_s}{dT} \tag{1.173}$$

where M_s is the spontaneous magnetization and T is temperature.

Recently, a sensor application of the pyromagnetic effect has been proposed, and a feasibility study made [60]. The application proposed is that the pyromagnetic effect can be used to develop an infrared detector, in much the same manner as has been done for the pyroelectric effect. A pyromagnetic material is placed in a magnetic circuit with a pickup loop wound around the magnetic core. Any thermally absorbed incident radiation causes the temperature to change with time, and thus a voltage will be induced in loop due to the change in the spontaneous magnetization with changing temperature. The pyromagnetic voltage is expected to be the largest as the Curie temperature (or the Neel temperature) is approached from lower temperatures, since here M_s changes

most rapidly with temperature. Sensors based on this principle were tested. Among the materials used were gadolinium, iron-rhodium alloys, and iron-manganese-aresenic alloys. The study concluded that the promagnetic effect yielded infrared sensors which are competitive with other broadband, room temperature detectors, such as the thermocouple and pyroelectric type sensors, in terms of detectivity, N. E. P., and responsivity (typically 3 volts/watt). The pyromagnetic detectors appear to be faster than these other previously mentioned detectors. A typical pyromagnetic sensor had a response time of less than a millisecond.

Another conclusion based on the results of the above study indicates that the pyromagnetic effect offers a valuable research tool for studying magnetic phase transitions in materials.

List of References

1) F. J. Blatt, Physics of Electronic Conduction in Solids, McGraw Hill Book Company, New York, N. Y., Chapter 5, 1968.

2. A. C. Smith, J. F. Janak, and R. B. Adler, Electronic Conduction in Solids, McGraw Hill Book Company, New York, N. Y., Chapter 7, 1967.

3) A. H. Wilson, The Theory of Metals, Cambridge University Press, New York, N. Y. Chapter 8, 1965.

4) R. A. Smith, Wave Mechanics of Crystalline Solids, Chapman and Hall Ltd., London, England, Chapter 10, 1961.

5) G. S. Koerber, Properties of Solids, Prentice Hall, Inc., New Jersey, Chapter 5, 1962.

6) G. T. Meaden, Electrical Resistance of Metals, Plenum Press, New York, Chapter 3, 1965.

7) R. A. Smith, Semiconductors, Cambridge University Press, London, England, Chapter 5, 1964.

8) B. Donovan, Elementary Theory of Metals, Pergamon Press, New York, Chapters 7-8, 1967.

9) D. Long, Energy Bands in Semiconductors, Interscience Publishers, New York, Chapter 2, 1968.

10) T. M. Rice and D. B. McWhan, "Metal-Insulator Transition in Transition Metal Oxides", I. B. M. Jour. Res. and Dev., vol. 14, May 1970, pp. 251-257.

11) D. Adler and H. Brooks, "Theory of Semiconductor-To-Metal Transistions", Phys. Rev., vol. 155, March 1967, pp. 826-840.

12) F. C. Brown, The Physics of Solids, W. A. Benjamin, Inc., New York, Chapter 7, 1967.

13) C. Kittel, Introduction to Solid-State Physics, John Wiley and Sons, Inc., New York, 3rd edition, Chapter 12, 1967.

14) J. C. Anderson, Dielectrics, Reinhold Publishing Corp.,New York, 1964.

15) A. R. Von Hippel, Dielectric Materials and Applications, John Wiley and Sons, Inc. New York, 1954.

16) J. C. Burfoot, Ferroelectrics, An Introduction to the Physical Principles, D. Van Nostrand Co. Ltd., London, England, Chapter 6, 1967.

17) E. Fatuzzo and W. J. Merz, Ferroelectricity, John Wiley and Sons, Inc., New York, N. Y., Chapter 4, 1967.

18) S. Bhagavantam, Crystal Symmetry and Physical Properties, Academic Press, Inc., New York, N. Y., Chapter 11, 1966.

19) J. M. Ziman, "Some Remarks on Current Theories of the Transport Properties of Metals, With Reference to the Effects of Pressure," in Physics of Solids at High Pressures, Academic Press, New York, N. Y. 1965, pp. 171-183.

20) Donald Long, Energy Bands in Semiconductors, Interscience Publishers, New York, N. Y., 1968, pp. 61-62.

21) J. J. Wortman and F. T. Wooten, "Research on Piezojunction Sensors", Technical Report AFAL-TR-67-160, Air Force Avionics Laboratory, Wright Patterson AFB, Ohio.

22) K. Bulthuis, "Effect of High Uniaxial Pressure Along the Main Crystallographic Axes For Silicon and Germanium", Phillips Research Reports, vol. 23, 1968, pp. 25-47.

23) M. A. Littlejohn and C. Cross, "Orientation Dependence of Gage Factor in Electron-Irradiated Silicon", Transactions of I.S.A., vol. II, 1969 pp. 1-7.

24) R. Bowman, "Semiconductor Sensor For Pressure Measurement at Cryogenic Temperatures", I.S.A. Conference Preprint no. 16. 14-3-66, New York, N. Y., Oct., 1966.

25) E. E. Pittelli and W. Rindner, "Simplified Theory and Transducer Palentialities of the Pinch Effect in Semiconductors", NASA Technical Note, NASA TN D-5130, April 1969.

26) E. Rashba, V. Romanov, I. Boiko, and I. Zhadko, "Electrical Pinch in Elastically Deformed Germanium", Phys. Stat. Sol., vol 16, 1966, pp. 43-56.

27) W. Wang, Solid-State Electronics, McGraw Hill Book Co., Inc., New York, N. Y. Chapter 7, 1966.

28) W. P. Wason, Physical Acoustics, vol. 1-Part A, Academic Press, New York, N. Y. 1964, pp. 169-270.

29) L. W. Canfil and W. C. Nieberding, "Sensitivity of Quartz Piezo-electric Pressure Transducers as a Function of Temperature From 20°K to 477°K", NASA Technical Memorandum, NASA TM X-1479, December, 1967.

30) H. D. P. Neubert, Instrument Transducers, Oxford University Press, London, England, Chapter 4, 1963.

31) H. W. Katz, Solid-State Magnetic and Dielectric Device, John Wiley and Sons, Inc. London, England, Chapter 5, 1959.

32) H. Weiss, Structure and Application of Galvanomagnetic Devices Pergammon Press, Oxford, England, 1969, pp. 170-176.

33) S. Katoaka, H. Yamada, and H. Fryisada, "Integrated InSb Magnetoresistive Displacement Transducers", Solid-State Sensors Symposium Conference Record, June 1970, pp. 56-61.

34) R. H. Bube, Photoconductivity of Solids, John Wiley and Sons, Inc., New York, N. Y., Chapter 3, 1960.

35) R. H. Bube, "Photoconductivity", article in Physics and Chemistry of II-VI Compounds, John Wiley and Sons, Inc. New York, N. Y. 1967, pp. 657-705.

36) J. T. Houghton and S. D. Smith, Infrared Physics, Oxford University Press, London, England, Chapter 5, 1966.

37) M. L. Schultz and G. A. Morton, "Infrared-Sensitive Extrinsic Germanium and Germanium-Silicon Alloy Photoconductors", in Photoelectronic Materials and Devices, D. Van Nostrand Co., Inc. Princeton, N. J., Chapter 3, 1965.

38) E. L. Stelzer, J. L. Schmit, and O. N. Tufte, "Mercury Cadmium Telluride as an Infrared Detector Materials", I.E.E.E. Trans. E. D., vol 16, 1969, pp. 880-884.

39) R. H. Bube, "Photoconductors", in Photoelectronic Materials and Devices, D. Van Nostrand Co., Inc., Princeton, N. J. Chapter 2 1965.

40) A Rose, Photoconductivity and Allied Problems.

41) R. H. Bube, Photoconductivity", R. C. A. Engineer, vol. 5, 1959 pp. 29-32.

42) J. S. Blakemore, SEmiconductor Stateistics, Pergamon Press, New York, N. Y., Chapters 4-7, 1962.

43) A. F. Joffe, Physics of Semiconductors, Adacemic Press, New York, N. Y. 1960, pp. 271-287.

44) E. J. Scheibner, "Solid-State Physical Phenomena and Effects, Part I," IRE Trans. on Component Parts, vol 8, 1961, pp. 133-151.

45) J. Tauc, Photo and Thermoelectric Effects in Semiconductors, Pergamm Press, New York, N. Y., 149, 1962.

46) F. D. Rossi, "Thermoelectricity and Thermoelectric Power Generation", Solid State Electronics, vol.11, 1968, pp. 833-868.

47) E. H. Putley, The Hall-Effect and Semiconductor Physics, Dover Publications, New York, N. Y., Chapter IV, 1960.

48) R. Kubo and T. Nagamiua, Solid-State Physics, McGraw Hill Book Co., New York, N. Y., Part III, Chapter 2, 1969.

49) F. Seitz, The Modern Theory of Solids, McGraw Hill Book Co., New York, N. Y. 1940, pp. 599-603.

50) J. M. Ziman, Principles of the Theory of Solids, Cambridge University Press, London, England, Chapter 10, 1964.

51) J. E. Besancon et al, "Ferromagnetic Transducers", Proceedings of Conference on Megagammon Press, Oxfored, England , 1969 pp. 170-176.

52) R. B. Yarborough and F. S. Greene, Jr., "An Investigation of the Use of Faraday Rotation for the Measurement of Magnetic Fields", NASA Contract Report NASA-CR 73259, April 1968.

53) R. J. Blume, "Direct Quartz Crystal Control of a Low-Level Pound-Knight-Watkins Spin Magnetometer", Rev.Sci. Instr.,29, July 1958, pp. 574-577.

54) A. E. Beckowitz and E. Kneller, Magnetism and Metallurgy, vols. I and II, Academic Press, New York, N. Y. 1962.

55) A. H. Morrish, The Physical Principles of Magnetism, John Wiley and Sons, Inc., New York , N. Y. Chapter 8, 1965.

56) J. S. Blakemore, Solid-State Physics, W. B. Saunders Co., Philadelphia, Pa.,Chapter 5, 1969.

57) F. F. Stuckl, "A High-Speed Ferrimagnetic Microtransducer", Third Congress of Internation Federation of Automatic Control, London, England, Paper 7, June, 1966.

58) C. W. Richards, Engineering Materials Science, Wadsworth Publishing Company, Inc., Belmont, Calif., 1961, pp. 436-445.

59) A. Oguchi and S. Yoshida, "A Magnetostrictive Load Cell For Use Under High Hydrostatic Pressures", Jap. Jour. Appl. Phys., vol. 7, June 1968, pp. 672-678.

60) R. M. Walser, et al, "The Feasibility of Infrared Pyromagnetic Detectors", Solid-State Sensors Symposium Conference Record, June 1970, pp. 79-81.

2.0. DEVICE PROPERTIES AND JUNCTION SENSORS

The object of this section is to review the physical principles of solid state devices and to discuss their application as solid state sensors. The discussion is limited to devices composed of two or more dissimilar materials and whose operation depends on the properties of the interface region between the dissimilar materials. Solid state sensors which depend on some bulk effect for their sensor properties and which nearly have ohmic contacts for measuring this bulk effect are not considered in this section but are considered in the section on material properties.

The classical example of a solid state junction sensor is a semiconductor p-n junction. These find extensive use as optical radiation detectors, as temperature sensors, and as stress sensors. The properties of the interface region between the p- and n-type semiconductors is of fundamental importance to the use of p-n junctions as sensors. The semiconductor p-n junction provides the best example of the type of solid state sensors to be discussed in this section.

The material in this section is classified according to the particular type of solid state device. For example a discussion of the p-n junction constitutes one of the major subsections of this part of the report. The various uses to which a particular type of device may be put and the properties of the device when used as a sensor are discussed under the various types of devices.

In additon to the general properties of individual devices used as sensors, a section is also included on the use of sensors in large

arrays. This is becoming a very important part of the use of solid state sensors expecially in the optical area, where information is frequently contained in a two-dimensional picture which must be detected. Because of the ease with which integrated circuit technology can be carried over to solid state sensor arrays, the use of solid state sensors in arrays has distinctive advantages over other types of sensor arrays.

2.1. Semiconductor p-n Junctions

The most basic solid state device is the p-n junction composed of a p-type semiconductor region adjoining an n-type semiconductor region. P-n junctions can be formed in a large number of semiconductor materials. Certainly the semiconductor materials which have so far found wide use have been those in which p-n junctions can be formed. The class of materials in which one can readily form p-n junctions includes the Group IV elements of Si, Ge, and the Group III-V elements such as GaAs, and InAs as well as a few of the Group II-VI elements. For most of the Group II-VI semiconductors, it has not been possible so far to produce p-n junctions because of the creation of compensating defects for one particular type of conductivity. This has severely limited device applications of the Group II-VI elements.

For the Group IV elements and for the Group III-V elements there exists in most cases a fairly well established technology for producing p-n junctions and therefore p-n junctions make ideal solid state sensors provided they have a desirable response to the force which is to be sensed. As far as present applications are concerned, the semiconductor p-n junction is probably the most important solid state sensor. The characteristics of a p-n junction are influenced by many types of external stimuli. The most important of these are light, particle irradiation, temperature, and mechanical stress. Its ability to respond to all of these stimuli makes it useful in a variety of sensing applications, but this also leads to problems when one wishes to sense one particular type of stimuli in the presence of other types of stimuli. The p-n junction is also one of the most thoroughly understood semiconductor devices, which is another important attribute in considering solid state devices for sensor applications.

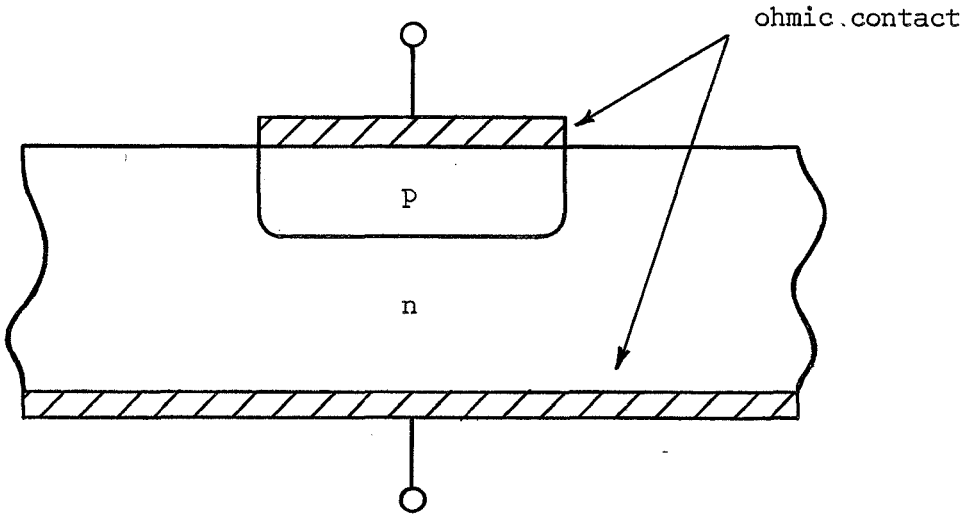
2.1.1. Basic Theory

The basic theory of p-n junctions is very well understood and there are many good texts on this subject ([1] - [4]). The object of the present section is to review the importance features of p-n junctions and to present the basic equations and models needed to understand the use of p-n junction as solid state sensors.

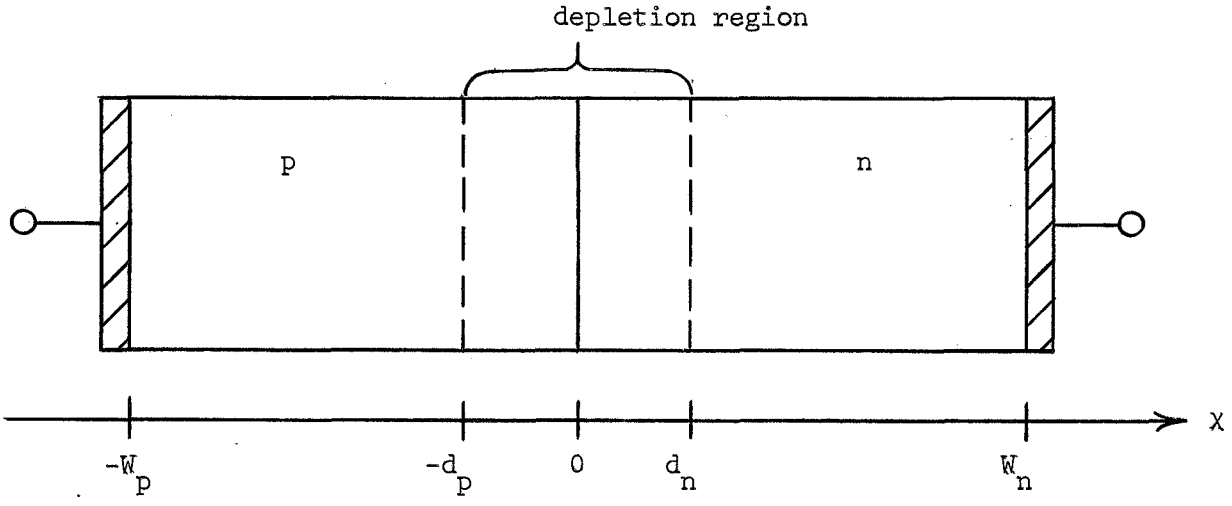
2.1.1.1 Basic Properties

The basic configuration of a p-n junction is shown in Figure 2.1 . This consists of a p-type region adjoining an n-type region with ohmic contacts to the n- and p-type regions. In equilibrium is well known that there exists a depletion region near the interface of the two regions. The free carrier density within this depletion region is not zero but is small compared with the free carrier density within the bulk of the semiconductor regions. The depletion region consists essentially of a region filled with bound charge, positive charge on the n-type side and negative charge on the p-type side. Because of the uncompensated charge within the depletion region, there is an electric field and an electric potential variation within the depletion region. The difference in potential across the depletion region equals the difference in Fermi levels of two unconnected p- and n-type regions with similar doping levels. For most diodes this equilibrium junction potential V_o varies from about 1/2 the energy band gap expressed in volts to a few tenths of a volt above the band gap voltage depending on the doping on either side of the junction.

The properties of a p-n junction are governed by the semiconductor device equations which in one-dimension are



a. Typical diffused or alloyed p-n junctions



b. Basic one-dimensional model

Figure 2.1. Basic p-n Junction

$$\frac{\partial n}{\partial t} = - U + G_{ex} + \frac{1}{q} \frac{\partial J_n}{\partial x} \tag{2.1}$$

$$\frac{\partial p}{\partial t} = - U + G_{ex} - \frac{1}{q} \frac{\partial J_p}{\partial x} \tag{2.2}$$

$$J_n = q\mu_n n E + q D_n \frac{\partial n}{\partial x} \tag{2.3}$$

$$J_p = q\mu_p p E - q D_p \frac{\partial p}{\partial x} \tag{2.4}$$

$$\frac{\partial E}{\partial x} = \frac{q}{E} (p - n + N_D^+ - N_A^-). \tag{2.5}$$

In these equations, E is the electric field, n and p are the free electron and hole densities, J_n and J_p are the current densities, μ_n and μ_p are the carrier mobilities, U is the net internal recombination-generation rate, G_{ex} is the generation rate of hole-electron pairs due to some external source, N_D^+ and N_A^- are the ionized donor and acceptor charge densities and $D_n = (kT/q)\ \mu_n$, $D_p = (kT/q)\ \mu_p$. These normal equations assume that the magnetic field is negligable. Thus, additional terms must be added to consider the possible applications of devices as magnetic field sensors.

2.1.1.2 Junction Space Charge Region Properties

The depletion region properties of a p-n junction are readily calculated from the approximation of no free carriers, and for an abrupt junction with N_A and N_D doping levels, [11-4]

$$V_J = \frac{q}{2\varepsilon} (N_D\, d_n^2 + N_A\, d_p^2), \tag{2.6}$$

$$d_n^2 = \frac{2\varepsilon V_J}{q} \quad \frac{N_A}{N_D} \quad \frac{1}{N_A + N_D}, \tag{2.7}$$

$$d_p^2 = \frac{2\varepsilon V_J}{q} \quad \frac{N_D}{N_A} \quad \frac{1}{N_A + N_D}, \tag{2.8}$$

where V_J is the voltage across the depletion region with the p-type region taken at zero potential. The above equations are valid both at equilibrium and with an applied potential if V_J is the difference between the equilibrium voltage and the applied voltage, i.e.

$$V_J = V_o - V_a, \tag{2.9}$$

where

V_o = equilibrium junction potential

$$= \frac{kT}{q} \ln \left(\frac{N_D N_A}{n_i^2} \right), \tag{2.10}$$

and V_a is the voltage applied to the termials of the diode. In considering junction breakdown and avalanche multiplication, the maximum electric field within the depletion region is useful and this is given by

$$E_{max} = \frac{q}{\varepsilon} \sqrt{\frac{2\varepsilon V_J}{q} \quad \frac{N_A N_D}{N_A + N_D}}. \tag{2.11}$$

For typical semiconductors and doping levels, the depletion region width is in the range of 10^{-4} cm with values ranging from 10^{-5} cm to 10^{-3} cm. The maximum electric field within a typical depletion region is 10^{4} to 10^{5} v/cm.

A final parameter which is useful in characterizing the depletion region is the junction capacitance. This is given as

$$C = A \left(\frac{q\varepsilon}{2V_J} \frac{N_A N_D}{N_A + N_D} \right)^{1/2}. \tag{2.12}$$

2.1.1.3. Current Densities at p-n Junction

The potential across the depletion region establishes the carrier concentrations at the edges of the depletion region. These are known to be given by the relationships

$$\frac{n_n (d_n)}{n_p (-d_p)} = \frac{p_p (-d_p)}{p_n (d_n)} = \exp (q V_J/kT), \tag{2.13}$$

where the n or p subscript indicates the n- or p-type region. For small values of applied forward bias these can readily be shown to give

$$n_p(-d_p) = n_{po} \exp (q V_a/kT), \tag{2.14}$$

$$p_n (d_n) = p_{no} \exp (q V_a/kT), \tag{2.15}$$

where n_{po} and p_{no} are the equilibrium values of electron and hole concentrations in the n- and p-type regions. For large values of

applied voltage the carrier density expressions approach the forms

$$n_p(-d_p) \rightarrow n_i \exp\ (\ qV_a/2kT), \tag{2.16}$$

$$P_n\ (d_n) \rightarrow n_i \exp\ (\ qV_a/2kT). \tag{2.17}$$

As indicated by these expressions the applied voltage determines the carrier concentrations at the edges of the depletion region. The diode current is influenced by both the carrier concentration and by the transport of the carriers away from the edge of the depletion region. The transport of carriers away from the depletion region is by drift and diffusion processes. The major factors influencing this process are the carrier mobilities (or diffusion constants), the lifetimes τ_n and τ_p of the excess carriers, and any possible doping gradients present in the semiconductor. Carrier diffusion is the dominant transport mechanism in most cases and for pure diffusion the current density expressions are

$$J_n = \frac{qD_n}{L_n}\ [n_p\ (-d_p) - n_{po}], \tag{2.18}$$

$$J_p = \frac{qD_p}{L_p}\ [p_n\ (-d_n) - p_{no}], \tag{2.19}$$

where L_n and L_p are the diffusion lengths given by

$$L_n = \sqrt{D_n\ \tau_n}\ ,\ L_p = \sqrt{D_p\ \tau_p} \tag{2.20}$$

In terms of the applied voltage the currents become:

(a) Low applied voltage

$$J = J_n + J_p = J_s(e^{q V_a/kT}-1),$$

$$J_s = \frac{q D_n n_{po}}{L_n} + \frac{q D_p p_{no}}{L_p}, \qquad (2.21)$$

$$= \left(\frac{q D_n}{L_n p_{po}} + \frac{q D_p}{L_p n_{no}}\right) n_i^2,$$

(b) High applied voltage

$$J \to J_s e^{qV_a/2kT},$$

$$J_s = \left(\frac{q D_n}{L_n} + \frac{q D_p}{L_p}\right) n_i . \qquad (2.22)$$

In the above equations it has been assumed that $L_n >> W_n$ and $L_p >> W_p$ so that few carriers reach the ohmic contacts. If this is not the case then the following replacements should be made in the above equations:

$$L_n \to L_n \coth (W_n/L_n),$$

$$L_p \to L_p \coth (W_p/L_p). \qquad (2.23)$$

The simple diffusion model is found to be inadequate to describe the characteristics of real diodes at very large applied bias, at very small applied bias and at reverse bias. First at very large bias ohmic voltage drops in the bulk become important and the voltage

which is actually applied to the p-n junction is reduced by the ohmic voltage drop. In many cases this effect can be approximated by an ideal diode in series with a resistance. Also at high forward bias, the diffusion constants and lifetimes become dependent on the injection level and these effects must be considered for an accurate model at large forward bias. The effects are probably not very important for sensor applications since most sensors operate at low power levels.

The major modifications of the diffusion model which must be made at low forward bias and at reverse bias arise from the generation and recombination of excess monority carriers within the depletion region. This gives rise to an additional current component J_u which must be added to the diffusion current components so that

$$J = J_u + J_n + J_p. \tag{2.24}$$

A fairly accurate approximation to J_u can be taken as [5]

$$J_u = \frac{qn_i^2 W_J [\exp(q V_a/kT)-1]}{(\tau_{po} n_1 + \tau_{no} p_1) + (\tau_{po} + \tau_{no}) n_i \exp(q V_a/2kT)} . \tag{2.25}$$

where τ_{po}, τ_{no}, n_1, and p_1 are constants characteristic of the recombination centers [6] and W_J is the width of the junction depletion region. The depletion region width is a function of applied voltage and for most devices has the variation

$$W_J = W_{JO} \left(\frac{V_o - V_a}{V_o}\right)^m \qquad (2.26)$$

where m is a constant in the range of 0.5 to 0.33, and W_{JO} is the equilibrium width of the depletion region. For an ideal abrupt junction m = 0.5 while m = 0.33 for an ideal linear graded junction. For small values of forward bias and for reverse bias, the generation-recombination current dominates the current for most diodes.

The combined theoretical I-V characteristics for a typical diode under forward bias are shown in Figure 2.2. As can be seen from the figure there are several changes in slope as the forward bias is increased. For many diodes all regions can be distinguished. Typical experimental I-V curves for Si, Ge, and GaAs diodes are shown in Figure 2.3. By considering these curves and the band gaps of 0.66 ev for Ge, 1.11 ev for Si, and 1.43 for GaAs, some important general conclusions can be drawn concerning the forward I-V characteristics of p-n junctions. For a given applied voltage it is noted that the lower band gap material gives the larger current. This arises mainly from the exponential dependence of diffusion current on n_i^2 as seen by Equation (2.22). Also it is noted that the wider the band gap the more important is the generation-recombination current. For Ge, J_u makes a much smaller contribution to the total current than for GaAs. In the reverse bias direction the same general trends hold. Ge diodes show almost an ideal current saturation as predicted by the diffusion current while Si, GaAs and other wide band gap materials show a voltage dependent reverse bias current from the generation of excess carriers within the depletion region.

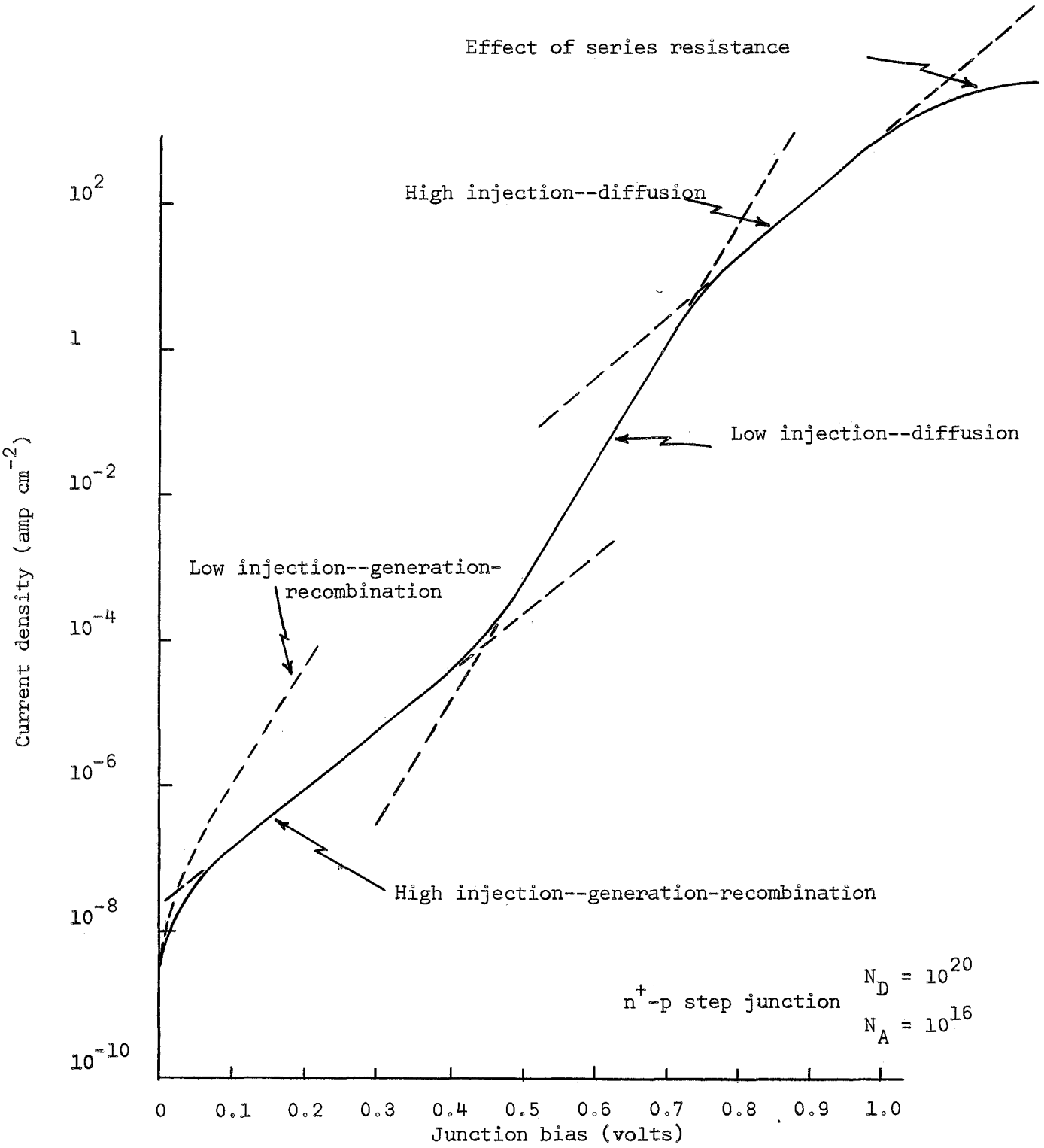


Fig. 2.2. Typical theoretical characteristic of p-n junction diode [7]

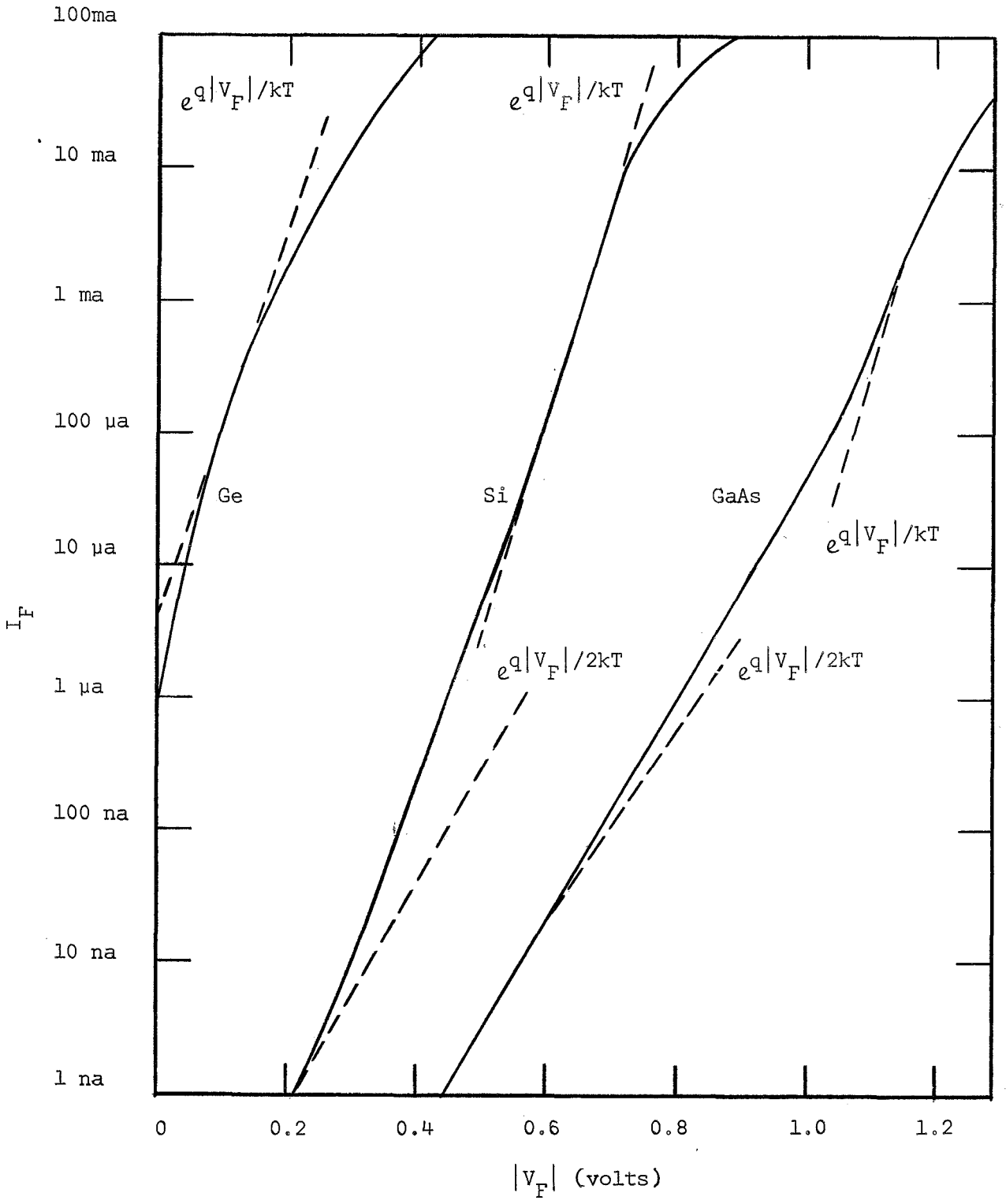


Figure 2.3. Experimental I-V Characteristics of Ge, Si, and GaAs diodes [4]

As discussed above the energy band gap of the semiconductor is one of the major parameters determing the details of the I-V characteristics of a diode. Other factors which are of somewhat lesser importance are minority carrier mobility and lifetime. In general for a given applied voltage the ideal diffusion current (see Equation (2.22)) is directly proportional to the square root of the mobility and inversely proportional to the square root of the lifetime. The dependence on mobility and lifetime is then much weaker than the exponential dependence on band gap. For typical semiconductors, mobility varies from several thousand to a few hundred $cm^2/v.$ sec depending on the semiconductor and type doping. Lifetime has an even wider variation, ranging from the nsec range to say 100 μsec. The longer lifetimes are found in indirect band gap materials such as Ge and Si while short lifetimes are normally found in the direct band gap materials such as GaAs and many other III-V materials.

2.1.1.4. Avalanche Multiplication At Large Reverse Bias

At very large reverse bias all p-n junctions diodes show a very large increase in current which is not predicted by the above models. This breakdown phenomena is known to be due to charge carrier multiplication in the high fields present within the depletion region. This charge carrier multiplication is known to arise from two effects - either Zener multiplication or avalanche multiplication. Zener multiplication is a valence band to conduction band tunneling transition, while avalanche multiplication is the creation of an electron-hole pair by a carrier which has acquired an energy which is larger than the band gap

energy. The Zener effect is known to dominate in diodes which have low breakdown voltages (less than about 7-8 volts) which the avalanche effect dominates in high breakdown voltage diodes. Both of these effects have been studied in great detail and can be predicted with considerable accuracy.

The typical dependence of reverse current on voltage near breakdown is shown in Figure 2.4 . The increase in current near breakdown can be characterized by a multiplication factor M such that

$$J = J_{ideal} M, \tag{2.27}$$

where J_{ideal} is the current which would flow in the absence of the multiplication processes. This is the diffusion and generation-recombination current of Equation (2.24). Empirically it has been found that a good approximation to M can be obtained with an equation of the form [8]

$$M = \frac{1}{1 - (\frac{V}{V_B})^m} , \tag{2.28}$$

where V_B is the breakdown voltage at which point the current becomes very large and m is a constant in the range of 1 to 4 for most devices. M can be considered as the average factor by which a current carrier is multiplied in traversing the depletion region because of the creation of electron-hole pairs.

Detailed calculations can be made for the breakdown region [9]. These will not be considered here in detail because the avalanche multiplication region does not appear to offer many advantages for diodes sensor operation. The breakdown voltage has been found

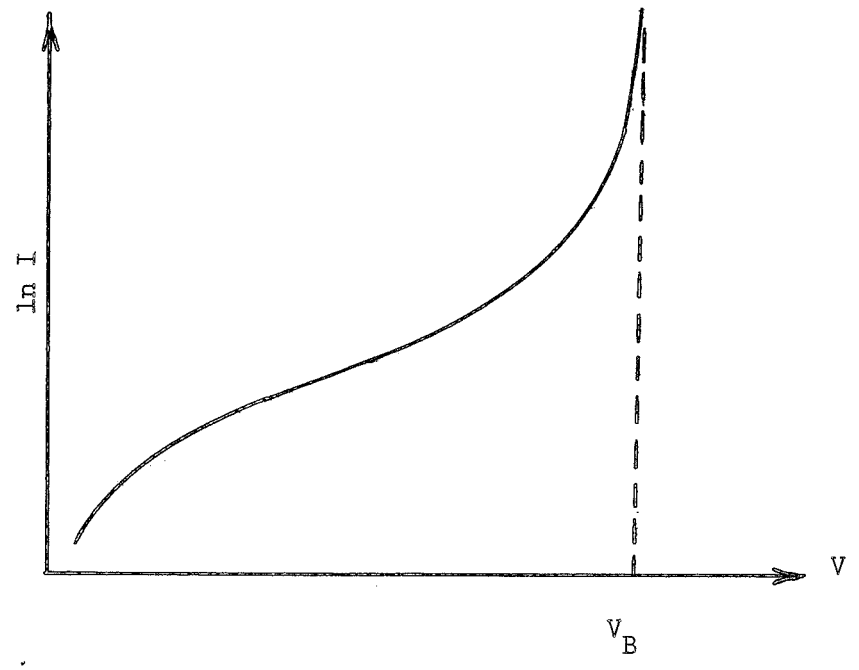


Figure 2.4. Typical dependence of reverse current on voltage near breakdown.

experimentally to be fairly insensitive to temperature, mechanical stress, magnetic field and other external forces. One exception to the above statement is in the area of optical sensors where operation in the avalanche region can lead to increased sensitivity because of the internal multiplication process. This is discussed further in the section on optical p-n junction sensors.

2.1.1.5. Tunnel Diode Properties

The previous paragraphs have reviewed the major properties of normal p-n junctions which are useful in sensor applications. The tunnel diode is a special type of p-n junction with unique properties

which have been explored for sensor applications. The major physical difference between normal p-n junctions and the tunnel diode is the doping level of the p- and n-type regions. In the tunnel diode the regions are degenerately doped such that the equilibrium Fermi level is below the valence band edge on the p-side and above the conduction band on the n-side. This is shown in Figure 2.5 (a) for the equilibrium case. As a bias is applied to the diode in either the forward or reverse direction, carriers can tunnel through the depletion region and this gives an additional tunnel current component which is not present in normal p-n junctions. Normal diodes have no empty energy states across the depletion region into which the carriers can tunnel at small forward bias.

If the depletion region is then sufficiently narrow, carriers can tunnel through the depletion region at small values of forward and reverse bias. As the forward bias is increased to the point where $V > V_n + V_p$, as shown in Figure 2.5 (c), there are no available states into which the carriers can tunnel and the tunnel current must be reduced to zero. Qualitatively then it can be argued that the tunnel current in the forward direction should initially increase with bias, then reach some maximum value and finally decrease to zero. The tunnel current is in addition to the normal diffusion and generation-recombination current which is present in a diode at forward bias. Typical experimental I-V characteristics exhibiting the above features are shown in Figure 2.6 . The curves show the familiar negative resistance characteristic of the tunnel diode. The maximum current is known as the peak current I_p and the minimum current as the valley current I_V.

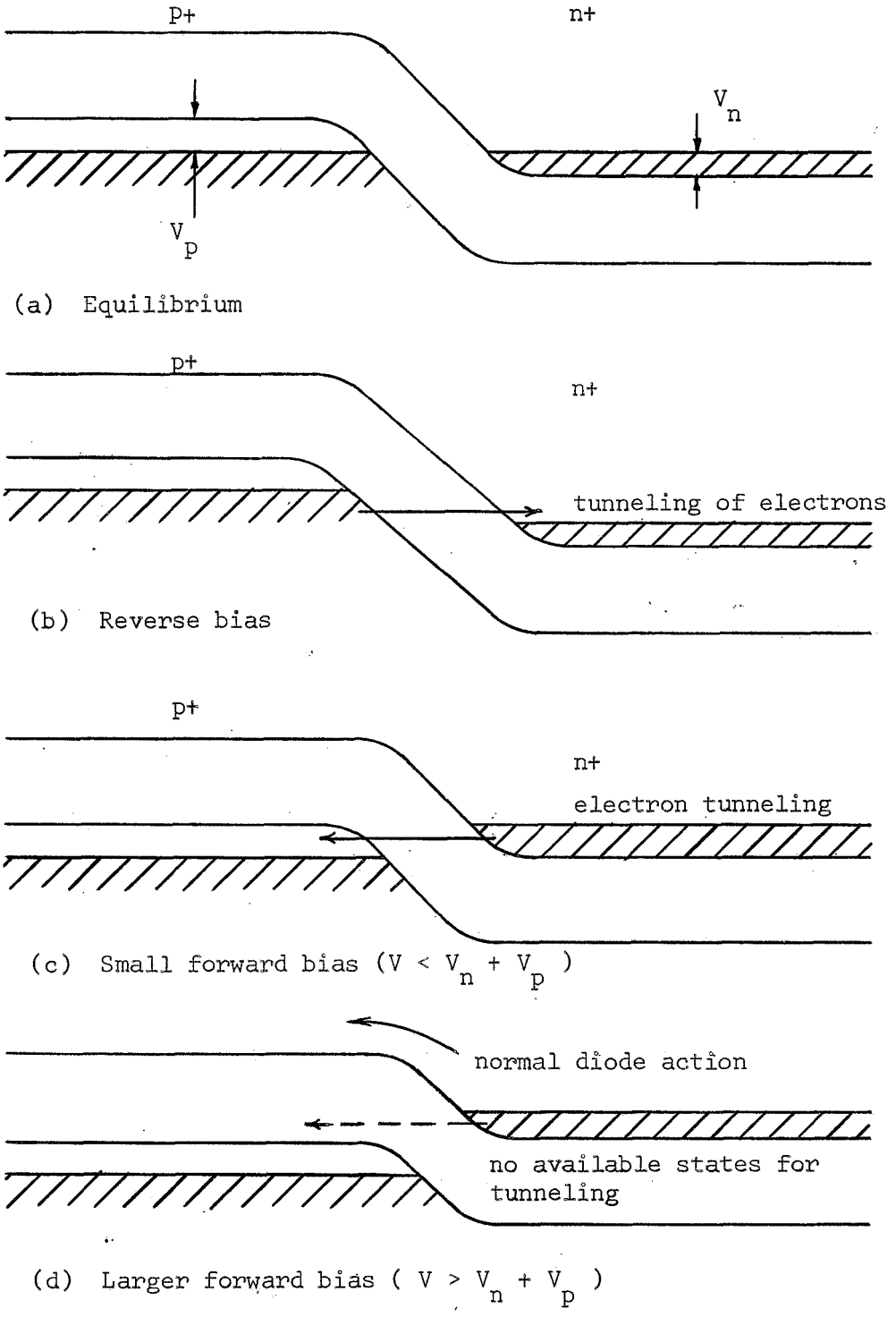


Figure 2.5. Energy bands for tunnel diode

The important factors determining the tunnel current are the tunneling probability and the density of available states into which the carriers can tunnel. A simple model of these effects leads to an expression for tunneling current of [9]

$$I_t = A \exp\left(\frac{-\pi m^{1/2} E_g^{3/2}}{2\sqrt{2}\, q \hbar E}\right) V (V_n + V_p - V)^2, \qquad (2.29)$$

where A is a constant, m is the electron effective mass, E_g is the band gap energy, and E is the average electric field within the depletion region. This simple model describes the major features of the tunneling current in diodes with appropriate choices of the constants in the equation.

The theoretical tunneling current goes to zero for $V > V_n + V_p$ and only the normal diode currents should be observed for voltages larger than $V_n + V_p$. Practical tunnel diodes show an additional

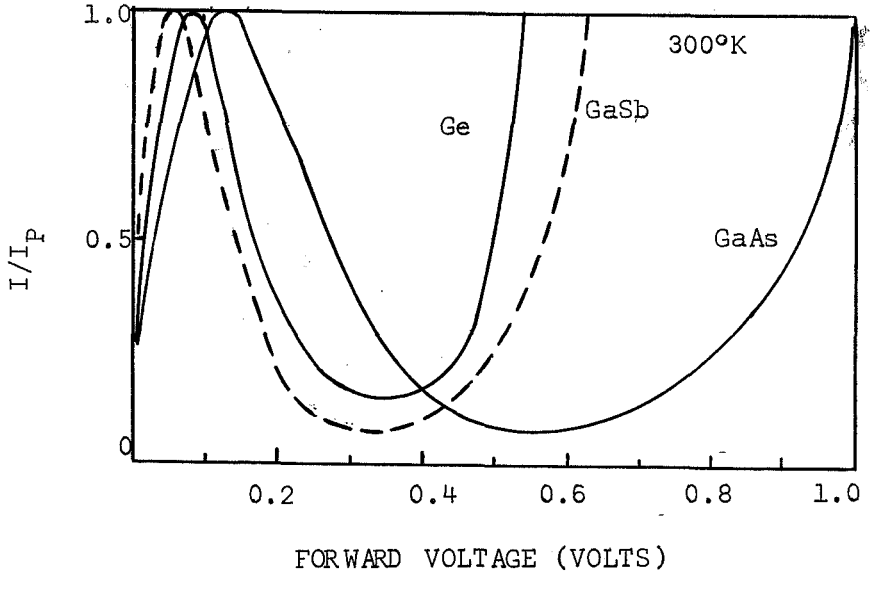


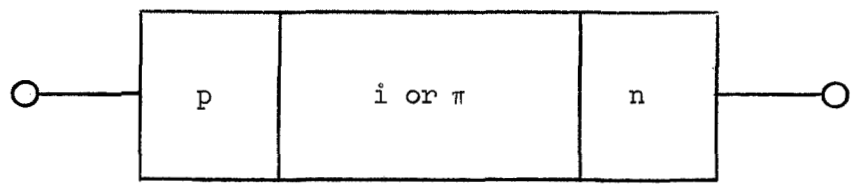
Figure 2.6 Typical current-voltage characteristics of Ge, GaSb, and GaAs tunnel diodes at 300°K. The ratio I_p/I_v is 8/1 for Ge and 12/1 for GaSb and GaAs tunnel diodes.

current for $V > V_n + V_p$ which is known as the excess current. This excess current arises from two sources: (1) the non abrupt nature of the band edges at degenerate doping levels and the tailing off of these bands into the band gap and (2) carrier tunneling by means of energy states within the band gap. These effects cause the valley current to be larger than for the ideal case. The effect on the peak current is less pronounced.

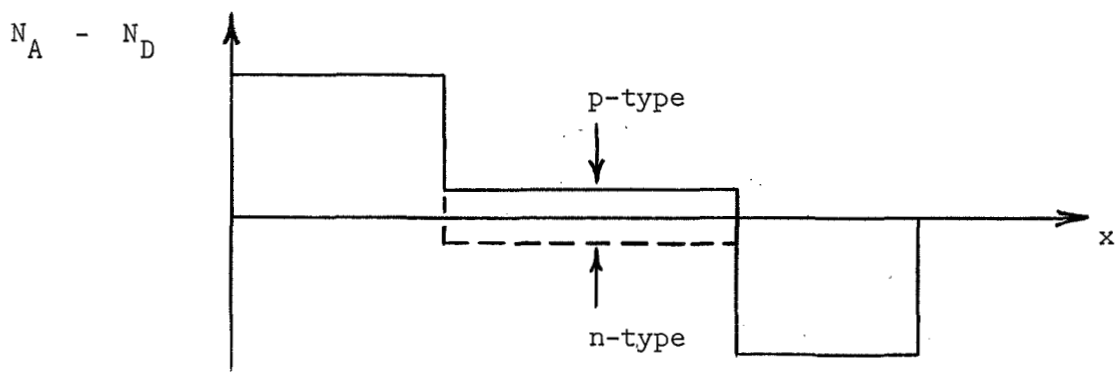
Additional complications arise in the analysis of tunneling current for indirect band gap semiconductors such as Si and Ge. Because of the requirement for momentum conservation as well as energy consevation, direct tunneling can not occur as assumed in the model leading to Equation (2.31). The difference in momentum between the valence and conduction band edges must be supplied by scattering agents such as phonons or impurities. This complicates the mathematical model, but the major features of the tunnel diode characteristic are similar to the simple model discussed above.

2.1.1.6. p-i-n Diode Properties

The p-i-n or p-π-n diode is a special type of junction diode in which there is an intrinsic or lightly doped semiconductor region seperating heavily doped p- and n-type regions. The impurity distribution for such a diode is shown in Figure 2.7. Because of the difficulty in achieving exactly intrinsic material the center region is normally slightly n-type or slightly p-type. If the intrinsic region is narrow compared with a diffusion length, the forward bias characteristic is similar to a normal p-n junction in the high



(a)



(b)

Figure 2.7. P-I-N Junction diode. (a) Structure (b) Impurity distribution

injection region, i.e. [10, 11]

$$J \simeq J_o \, (e^{q\,V/2kT}-1), \tag{2.30}$$

where J_o is a constant. At very low currents the resistance of the device can be very large because of the high resistance of the center region. Under forward bias the conductivity is modulated by the injected carriers and the resistance of the center region decreases rapidly with applied bias. The non-equilibrium carrier densities are confined mainly to the center lightly doped region where they recombine giving rise to the forward current.

Most important uses, especially sensor application, of p-i-n diodes operate under reverse bias conditions. The p-i-n diode exhibits a large depletion region width under reverse bias. For example with Si and a center region doping level of $10^{13}/cm^3$ (about 10^3 Ω . cm material) the zero bias depletion region width is about 10 μm and at a reverse bias of 100 volts, the depletion region width is about 100 μm. Thus at a reverse bias voltage of 100 volts a region 100 μm in width can be completely depleted for a doping density of $10^{13}/cm^3$. The major features of the p-i-n diode under reverse bias are the wide depletion region and the small capacitance due to the wide depletion region. The reverse bias current is due essentially to the generation of excess electron-hole pairs within the depletion region as expressed by the equation

$$J_{rev} \simeq \frac{q\, n_i^2\, W_J}{\tau_{po}\, n_1 + \tau_{no}\, P_1} \quad , \tag{2.31}$$

where W_J is the depletion region width which under large bias will essentially equal the width of the lightly doped center region. The wide depletion region of the p-i-n diode is the important characteristic of such a diode for optical sensors and particle sensors.

2.1.1.7. Noise

The term noise refers to uncontrollable spontaneous fluctations in the current passing through a device or in the voltage appearing across a device. Noise is a very important practical consideration

in all types of sensors since it determines the lower limit at which a sensor can be used to sense small quantities.

Noise in most solid state devices can be classified into thermal noise, shot noise, and flicker or 1/f noise. The thermal noise is due to the random motion of charge carriers arising from their thermal velocity. The open circuit mean square voltage $\langle V_t^2 \rangle$ due to thermal noise is $\langle V_t^2 \rangle = 4\ kTBR$, where R is the real part of the impedence across which the voltage is measured, and B is bandwidth in Hz used in the measurement system. The thermal noise can also be expressed in terms of a mean square noise current as

$$\langle i_t^2 \rangle = 4\ kT\ BG, \tag{2.32}$$

where G is the real part of the device admittance.

In most semiconductor devices the shot noise constitutes the major noise source. The current in a device consists of the flow of discrete electrons and the random emission of these carriers within any device gives rise to a random amplitude modulation of the current. At low and intermediate frequencies the shot noise is independent of frequency and

$$\langle i_s^2 \rangle = 2\ q\ BI, \tag{2.33}$$

where I is the total device current.

The flicker or 1/f noise is distinquished by its spectral distribution which is proportional to $(1/f)^{\alpha}$ where α is generally close to unity. This noise is believed to be due to surface effects. It becomes important in devices at low frequencies.

2.1.1.7 Fabrication Techniques for p-n Junctions

By far the most important practical technique for producing semiconductor p-n junctions is diffusion where the impurity diffuses into the semiconductor at high temperature. The source of the impurity may be either a gas stream passing over the semiconductor or some form of solid source deposited onto the semiconductor. With many compound semiconductors, the deposited solid source is the preferred approach because the deposited film can act not only as a source of impurities for diffusion but also as a protective coating to prevent decomposition of the semiconductor at the high temperautres required for diffusion.

In addition to diffusion, alloying also finds important uses in making special diode. This is especially true with tunnel diodes where it has not proven possible to obtain the large doping levels and abrupt transitions from n to p type material required for the tunneling phenomena with diffusion. Alloying is also employed with many III-V materials where the simplicity of the technique makes it advantageous.

The p-i-n diode presents special fabrication difficulties, but three techniques have been employed. These are (1) the epitaxial growth of a high resistivity layer on a heavily doped layer followed by diffusion or epitaxial growth of the remaining heavily doped layer, (2) the diffusion of p and n regions into a high resistivity substrate, and (3) the ion-drift (with lithium for example) method of introducing a highly compensated near intrinsic region [12, 13]. Of the three techniques the ion-drift technique is capable of producing the

highest resistivity regions over large depths and this technique is used almost exclusively in fabricating solid state particle detectors.

2.1.1.8. Dynamic Properties of Devices

The previous paragraphs have given a brief review of the static properties of p-n junctions. No mention has so far been made of the high frequency dynamic properties of p-n junctions. For example temperature and stress sensors would normally be limited by thermal time constants or by mechanical time constants as opposed to the inherent response of the semiconductor junction. In certain sensor applications such as optical detection of short light pulses the high frequency properties of the p-n junction are important. In these cases the dynamic properties of p-n junctions are discussed in the following sections on the specific applications of p-n junctions to sensors.

2.1.2. Optical Sensors

2.1.2.1. Introduction

Optical detectors using p-n junctions are one of the most used solid state sensors. As discussed in this section, optical sensors refers to sensor sensitive to radiation around the visible region of the electromagnetic spectrum. This covers the wavelength region from say 50 μm to around 0.1 μm and covers photon energies of about 0.025 ev to about 10 ev. This covers the far infrared to the vaccum ultraviolet region. At longer wavelengths p-n junctions become ineffective sensors because of the lack of physical phenomena for creating excess carriers within the semiconductor. At very short wavelengths photons behave more like particles than like waves in their interaction with materials. Semiconductor p-n junctions are sensitive to high energy photons or gamma rays and are used for sensors in this region. Because of the many similarities between high energy photons and other particles such as electrons, the discussion of the use of p-n junctions as sensors of high energy photons is contained in Section 2.1.5 on Radiation Sensors.

The object of the p-n junction optical sensor is to convert the electromagnetic energy of the photons into a voltage or current which can then be measured. The physical phenomena which is responsible for the response of p-n junctions to light is the creation of electron-hole pairs by the incident photons. An energy greater than the band gap energy is needed to excite an electron from the valence band to the conduction band so the energy band gap of the semiconductor

determines the upper limit of wavelength for which a particular semiconductor p-n junction can be used as a sensor. The normal semiconductors Si, Ge, and GaAs are sensitive over the visible region of the spectrum and into the ultraviolet region but they can not be used very far into the infrared. For the infrared region narrow band gap materials such as InAs (0.33 ev), InSb (0.16 ev), HgTe (0.02 ev), or contributions of these materials must be used.

2.1.2.2. Basic Optical Properties of p-n Junction

A p-n junction employed as an optical sensor is sketched in Figure 2.8. The incident light flux is taken as I. The magnitude of the voltage or current response of a p-n junction depends on several factors. First the response depends on the intensity of light which actually enters the semiconductor and this is some fraction TI of the light intensity incident on the semiconductor. The remaining part RI is reflected from the surface. After entering the semiconductor, the photons with energy larger than the band gap are essentially 100% effective in producing electron-hole pairs provided the semiconductor is sufficiently thick. The average distance traveled by a photon before producing an electron hole pair is a very strong function of wavelength or energy and can vary from 1 cm or so near the band edge to 10^{-6} cm at energies considerably above the band gap energy. The external sensor response of either voltage or current is determined by the fraction of created electron-hole pairs which are separated by the electric field of the junction depletion region. This fraction is a complicated function of several material and device parameters. The most important

of these are the light absorption coefficient, the minority carrier lifetime and mobility, the surface recombination of carriers, the junction depth and total thickness of the detector, as well as any applied terminal voltage. These various factors are discussed in more detail in the following paragraphs.

The reflectance of a semiconductor surface is know to be related to the refractive index $\bar{n}$ and absorption constant k by the expression

$$R = \frac{(1 - \bar{n})^2 + k^2}{(1 + \bar{n})^2 + k^2} \tag{2.34}$$

The absorption coefficient α is related to k by the expression

$$\alpha = \frac{4\pi k}{\lambda} . \tag{2.35}$$

Both $\bar{n}$ and α are functions of the wavelength. Figure 2.9 shows the typical reflectivity for silicon in the region above the band gap [14, 15]. Below the band gap of 1.1 ev there is a region of fairly constant reflectivity at about 30%. Above the band gap the reflectivity increases to about 70% at an energy of about 5 ev and then shows a slow decrease at larger energies. The reflectivity peaks are related to the energy band structure of the semiconductor. The reflectivity shown in Figure 2.9 is typical in major features of the group IV elements as well as the Group III-V and IV-VI elements.

Neglecting multiple reflections within the semiconductor (this is valid in the present considerations since most of the light is absorbed within the semiconductor) the light intensity transmitted

into the semiconductor is

$$IT = I\ (1-R). \tag{2.36}$$

For silicon p-n junctions, the light intensity within the semiconductor then ranges from about 70% to 30% of the incident intensity. The amount of reflected light can be reduced by the use of antireflection coatings on the surface of the semiconductor. This is discussed in more detail later in this section.

The photons present inside the semiconductor give rise to the creation of excess electron-hole pairs. The generation rate of these carriers is given by

$$G = \alpha\ N\ e^{-\alpha x}, \tag{2.37}$$

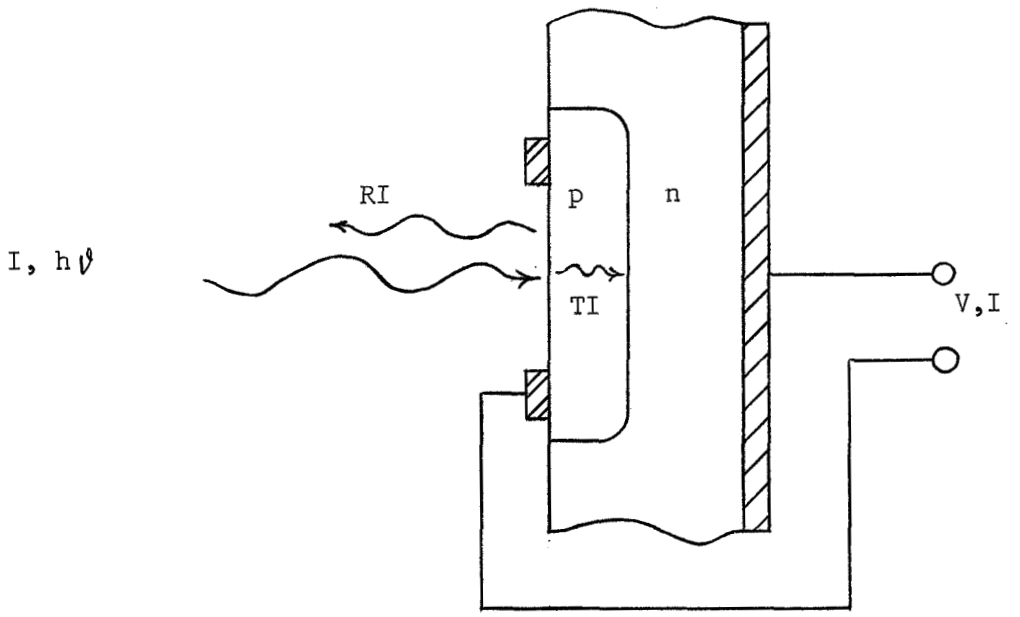


Figure 2.8. p-n Junction used as optical sensor

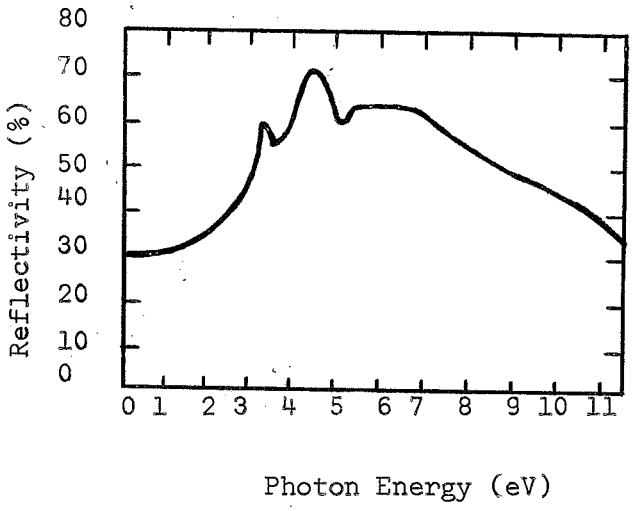


Figure 2.9. Reflectivity of Si for photon energies above the band gap energy

where N is the density of photons per unit area per unit time. This gives the number of excess carriers generated per unit time per unit volume assuming that each absorbed photon gives rise to one excess electron-hole pair. At very large photon energies (larger than about 3.4 ev for Si) each photon creates on the average more than one electron-hole pair. In this case the above expression must be multiplied by the internal quantum yield Q which is a rapidly increasing function of energy at large photon energies. For Si, Q is around 2.0 at photon energies of 5 ev.

The dependence of absorption coefficient on photon wavelength for three semiconductors is shown in Figure 2.10 [9]. In each case there is a sharp increase in α above the band edge. The sharpest rise is exhibited by direct band gap materials such as GaAs.

The created excess carriers within the semiconductor are governed

by the semiconductor device equations (Equations (2.1) - (2.5)). A model for the p-n junction type of optical detector is shown in Figure 2.11. The width of the junction depletion region is taken as W_J and the widths of the undepleted p- and n-type regions are W_p and W_n respectively. The front surface on which the radiation is incident is assumed to have some recombination levels which are characterized by a surface recombination velocity S. The back surface is assumed to be an ohmic contact or a contact with a high surface recombination velocity so that no excess carriers can exist at the back surface. Assuming uniform doping within the p- and n-type regions and assuming that diffusion currents dominate, the excess carriers, $\bar{n}$ and $\bar{p}$, are described by the equations:

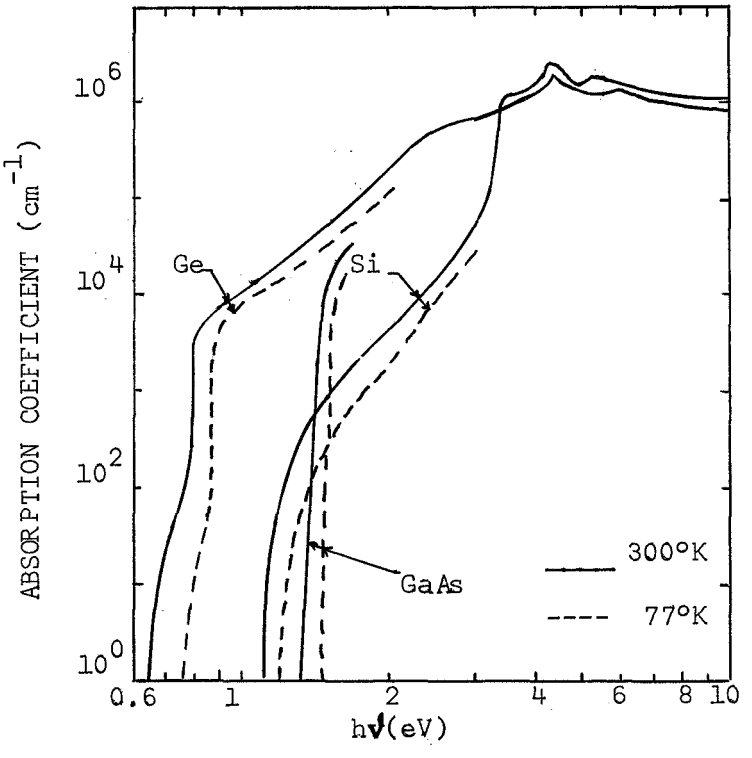


Figure 2.10. Measured absorption coefficients near and above the fundamental absorption edge of pure Ge, Si, and GaAs [9].

(a) p-type region

$$\frac{d^2\bar{n}}{dx^2} - \frac{\bar{n}}{L_n^2} = -\frac{\alpha N}{D_n} e^{-\alpha x}, \tag{2.38}$$

with boundary conditions

$$\bar{n}\,(W_p) \simeq n_{po}(e^{qV/kT}-1), \tag{2.39}$$

$$\bar{n}\,(o)S = D_n \left.\frac{d\bar{n}}{dx}\right|_{x=0}. \tag{2.40}$$

(b) n-type region

$$\frac{d^2\bar{p}}{dx^2} - \frac{\bar{p}}{L_p^2} = -\frac{\alpha N}{D_p} e^{-\alpha x}, \tag{2.41}$$

with boundary conditions

$$\bar{p}\,(W_p + W_J) \simeq p_{no}\,(e^{qV/kT}-1), \tag{2.42}$$

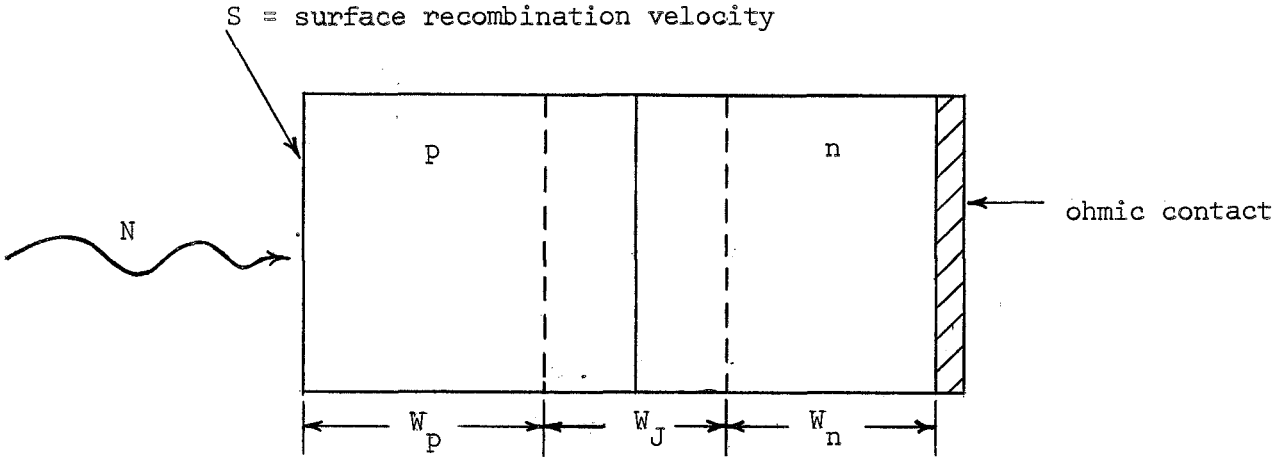


Figure 2.11. Model for p-n junction optical detector

$$\bar{p}\ (W_p + W_J + W_n) = 0. \tag{2.43}$$

The solutions to these equations give fairly complex functions for the excess carriers. The important characteristic of the sensor is the terminal characteristics relating current to voltage. The diffusion currents crossing the junction are related to the carrier densities by the equation

$$J_n = q\ D_n\ \frac{dn}{dx}\ \Big|_{x = W_p}, \tag{2.44}$$

$$J_p = q\ D_p\ \frac{dp}{dx}\ \Big|_{x = W_p + W_J}. \tag{2.45}$$

The total diffusion current which is the sum of these two quantities can be expressed finally as

$$J_d = J_n + J_p = J_o\ (e^{qV/kT}-1) - q\ \alpha(L_n\ F_n + L_p\ F_p)\ N, \tag{2.46}$$

where

$$J_o = \frac{q\ D_n\ n_{po}}{L_n}\ \frac{\sinh(\frac{W_p}{L_n}) + \frac{SL_n}{D_n}\ \cosh(\frac{W_p}{L_n})}{\cosh(\frac{W_p}{L_n}) + \frac{SL_n}{D_n}\ \sinh(\frac{W_p}{L_n})}$$

$$+\ \frac{q\ D_p\ p_{no}}{L_n}\ \coth(\frac{W_n}{L_p}), \tag{2.47}$$

$$F_n = \frac{e^{-\alpha W_p}\left[\sinh\left(\frac{W_p}{L_n}\right) + \frac{SL_n}{D_n}\cosh\left(\frac{W_p}{L_n}\right)\right] - \alpha L_n\left(1 + \frac{S}{\alpha D_n}\right)}{(1 - \alpha^2 L_n^2)\left[\cosh\left(\frac{W_p}{L_n}\right) + \frac{SL_n}{D_n}\sinh\left(\frac{W_p}{L_n}\right)\right]}$$

$$+ \frac{\alpha L_n}{(1 - \alpha^2 L_n^2)} e^{-\alpha W_p}, \qquad (2.48)$$

$$F_p = \frac{\coth\left(\frac{W_n}{L_p}\right) - e^{-\alpha W_n}\cosh\left(\frac{W_n}{L_p} - \alpha L_p\right)}{(1 - \alpha^2 L_p^2)} e^{-\alpha(W_p + W_J)} \qquad (2.49)$$

The first term on the right side of Equation (2.46) is just the normal diffusion current of a p-n junction as discussed in Section 2.1.1 in the absence of any optical radiation. The second term on the right which is directly proportional to N represents the current due to the excess carriers created by the light.

The above discussion has considered only the regions of the device external to the depletion region. Carriers created by the light within the depletion region also contribute to the current. The transit time across the depletion region is normally small compared to the lifetime so essentially every carrier generated within the depletion region is separated by the field within the depletion region and contributes to the current. The space charge region contribution is found by integrating the generation rate across the depletion region and gives

$$J_{sc} = J_{gr} - q\, N\, e^{-\alpha W_p}\,(1-e^{-\alpha W_J}), \tag{2.50}$$

where J_{gr} is the normal generation-recombination current of a p-n junction as given by Equation (2.27).

The total current then for the p-n junction sensor is

$$J_T = J_{diode} - q\,\alpha\bar{L}\,N, \tag{2.51}$$

$$= J_{diode} - J_{photo}$$

where J_{diode} is the normal voltage dependent diode current given by the expression

$$J_{diode} = J_o\,(e^{qV/kT}-1) + \frac{q\, n_i\, W_J[\exp(qV/kT) - 1]}{(\tau_{po}\, n_1 + \tau_{no} p_1) + (\tau_{po} + \tau_{no})\exp(qV/2kT)}, \tag{2.52}$$

and

$$\bar{L} = L_n F_n + L_p F_p + \frac{e^{-\alpha W_p}(1 - e^{-\alpha W_J})}{\alpha}. \tag{2.53}$$

The factor $\bar{L}$ gives an effective distance over which the generation of excess carriers by the optical photons contribute to the junction current. Similarly $L_n F_n$ and $L_p F_p$ give effective lengths in the p- and n-type regions over which the generation contributes to the junction current. It is obvious from Equation (2.51) that $\alpha \bar{L} \leq 1$ since all of the generated carriers must be collected by the junction to have this quantity equal unity.

The general equations for the generated photo current are fairly complex functions of the absorption coefficient, the widths of the various regions and the surface recombination velocity. It is possible to get an understanding of the important parameters by considering several limiting cases. First consider the case where the excess carrier generation rate is fairly uniform over W_p, W_J and to a depth of at least L_p within the n-type region. This corresponds to a small value of α as ocurs for example at energies just above the band gap energy. Several possibilities can still occur even for this simplification. If the p-region width is larger than a diffusion length, carriers created to a distance of L_n on the p-side and a distance of L_p on the n-side will be collected by the junction and

$$\bar{L} \simeq L_n + L_p + W_J. \tag{2.54}$$

This represents an upper limit on the effective length over which carrier collection by the junction can take place. On the other hand if the width of the p-type region is less than a diffusion length

$$\bar{L} \leq W_p + L_p + W_J, \tag{2.55}$$

where the equality holds if surface recombination is negligable. For large surface recombination, W_p above becomes multiplied by the factor αL_n. The above case where W_p is less than a diffusion length is the case for most detectors so that near the band gap energy

$$J_{photo} \simeq q\,\alpha\,(W_p + L_p + W_J)\,N. \qquad (2.56)$$

This indicates that the photo current is directly proportional to the absorption coefficient near the band gap where α is small.

As the absorption coefficient increases at large photon energies, the absorption becomes nonuniform and concentrated toward the surface. At large values of α such that $\alpha W_p >> 1$, essentially all absorption occurs within the surface p-type region very near the surface. In this case surface recombination is a very important factor in determining the fraction of the carriers collected by the junction. Under the assumption of $W_p << L_n$ and $\alpha W_p >> 1$, $\bar{L} \simeq L_n F_n$ and the photo current expressions becomes

$$J_{photo} \simeq q\,\frac{(1 + \frac{S}{\alpha D_n})}{(1 + \frac{SW_p}{D_n})}\,N. \qquad (2.57)$$

This is a decreasing function for increasing values of α and for $\alpha D_n >> S$ approaches the limit

$$J_{photo} \simeq \frac{qN}{(1 + \frac{SW_p}{D_n})}, \qquad (2.58)$$

which is independent of absorption coefficient. The ideal photo current of qN is then reduced by the presence of the surface recombination.

From the above limiting cases a general picture of the dependence of photocurrent on absorption coefficient and wavelength can be constructed. For photon energies just above the band gap the photocurrent is small because the absorption is small and the photocurrent

increases essentially proportionally to the increase in absorption coefficient. The photocurrent reaches a maximum when most of the photons are absorbed within the active region of the junction which is essentially $W_p + W_J + L_p$. The maximum photocurrent should then occur for wavelengths such that $(W_p + W_J + L_p)\ \alpha \simeq 1$. For larger values of absorption coefficient, the created carriers are closer to the surface where surface recombination can reduce the fraction collected at the p-n junction. At large values of α or high photon energies the photocurrent is expected to approach a constant independent of wavelength. The amount of decrease in photocurrent at high energies from the peak value depends strongly on the surface recombination.

The above conclusions and statements are based upon keeping N constant and independent of wavelength. Because surface reflectance is wavelength dependent N also changes with wavelength and these changes are in addition to those considered in the above paragraph.

The terminal I-V characteristic as described by Equation (2.51) is the normal diode current shifted along the current axis by the amount of the photocurrent. This is shown in Figure 2.12. The generated photocurrent can be measured at low bias voltages since

$$I_{sc} \simeq - I_{ph} = - qA\ \alpha\ \overline{L}\ N, \tag{2.59}$$

where I_{sc} is the short circuit current. If the power falling on the active area of the diode is P then the photon flux density N is given by the expression

$$N = \frac{P}{A\ h\nu}\ (1 - R), \tag{2.60}$$

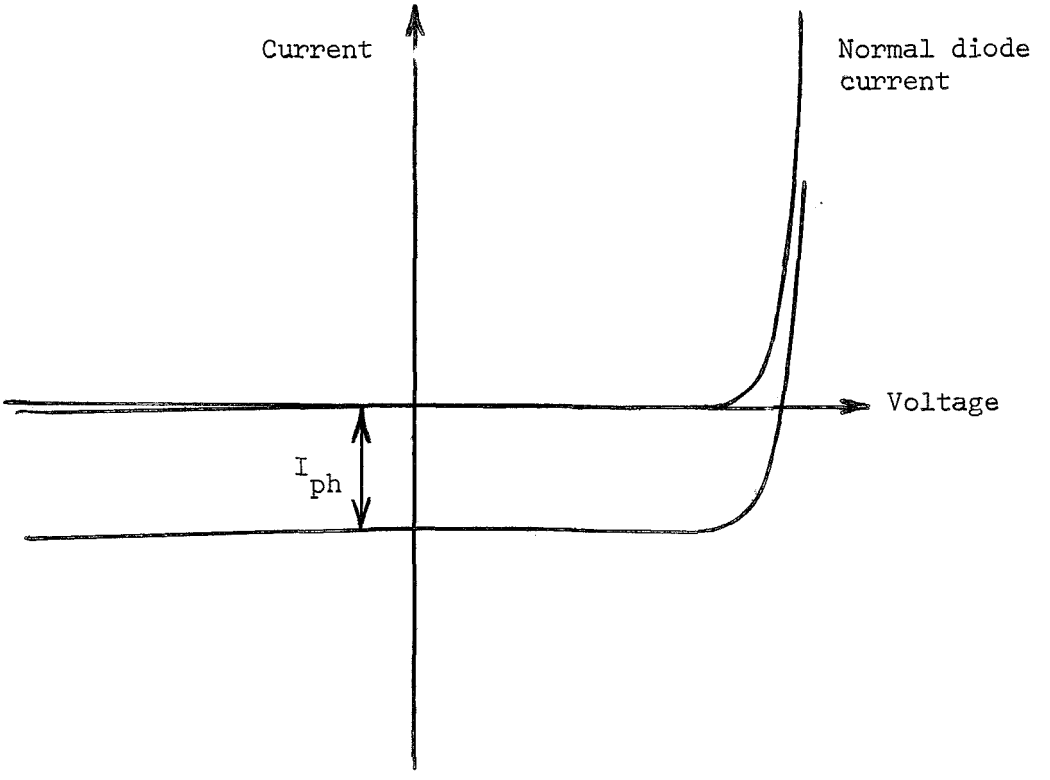


Figure 2.12. Sketch of terminal I-V characteristic of photodiode.

where ν is the photon frequency. Then

$$I_{sc} \cong I_{ph} = - q \frac{P}{h\nu} (1 - R) \alpha \bar{L}. \qquad (2.61)$$

An effective quantum efficiency can be defined as

$$\eta = \frac{I_{ph}}{q} \frac{h\nu}{P} = (1 - R) \alpha \bar{L}. \qquad (2.62)$$

This is the ratio of collected carriers to photons in the incident light beam. The maximum value which this effective quantum efficiency can have is obviously unity which would correspond to a negligable reflection and a 100% efficiency in collecting the carriers by the

junction. For silicon as shown in Figure 2.9 , R is 0.3 or larger so the effective quantum efficiency should be less than 0.7. Figure 2.13 shows the measured effective quantum efficiency as a function of photon wavelength for typical Si and Ge photodiodes [16]. Accounting for the reflection losses shows that most of the internally generated carriers are collected by the p-n junctions in the region where maximum response is observed. Most of the decreased response at short wavelengths is due to the increased reflectance in this region.

An internal quantum efficiency η_i can be defined as

$$\eta_i = \frac{\eta}{(1 - R)} = \alpha \overline{L}. \qquad (2.63)$$

This gives the fraction of photons which create electron-hole pairs and which subsequently are separated by the junction and contribute to the photocurrent. According to the theoretical model developed in the preceeding paragraphs this quantity should become fairly independent of wavelength at long wavelengths where most of the absorption is near the surface. Figure 2.14 shows this quantity as determined by Gutkin, et. al [17] for GaAs p-n junctions. The curves of part (a) show the response as a function of wavelength for illumination on the p-side while part (b) shows similar data for devices with a thin n-type region near the surface. In each case, in the region from about 2.8 eV to about 5.0 eV the internal response is almost independent of energy. This is consistent with the model predictions at large energies. Also the large variation in this region with temperature is consistent with lifetime being the controlling factor. Within this energy range the reflectance is changing so the overall efficiency would be decreasing

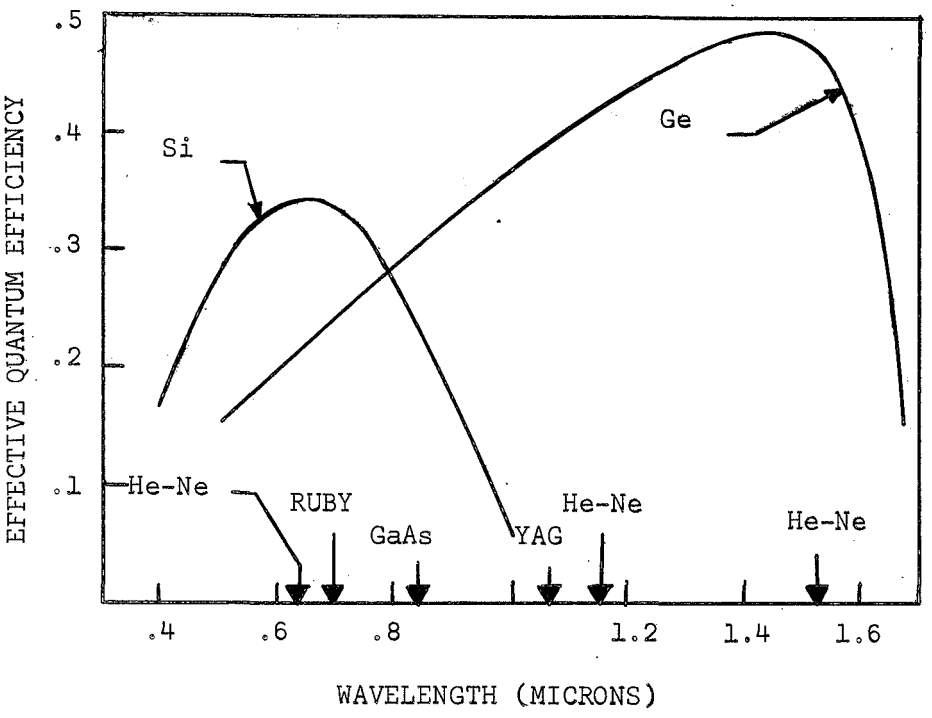


Figure 2.13. Effective quantum efficiency η as a function of wavelength for a silicon and a germanium photodiode. (Laser lines within this wavelength range are indicated)[16].

within this energy range. When surface recombination is low as in silicon devices with an SiO_2 layer on the surface, the internal quantum efficiency should be high and approaching unity at short wavelengths.

The response of p-n junction sensors in the untraviolet region requires special discussion because of the additional physical mechanism leading to an increased internal quantum yield. At large photon energies (exceeding about 3 eV in Si) each incident photon creates on the average more than one electron-hole pair. At about 5 eV in Si for example each photon creates about 2 electron-hole pairs [18]. This

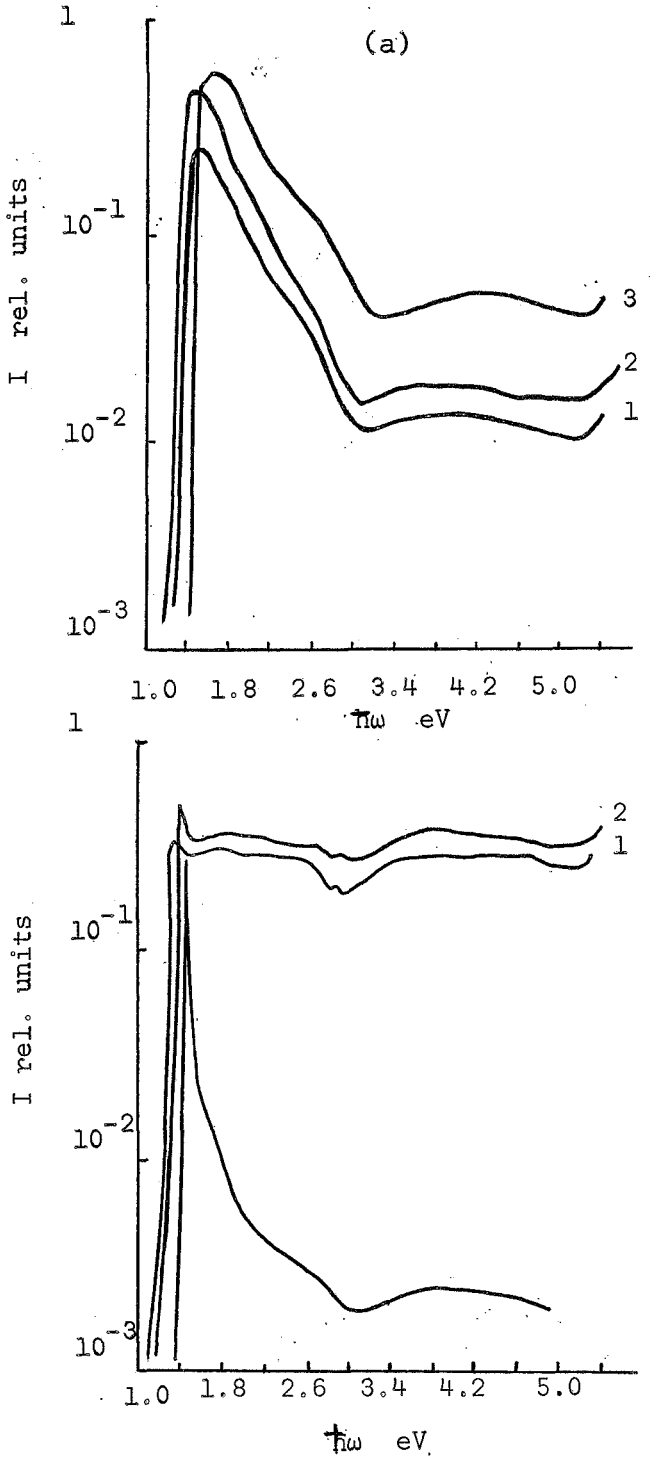


Figure 2.14. Internal efficiency or response of GaAs p-n junctions with (a) illuminated p-surface and (b) illuminated n-surface. (1) 370°K (2) 293°K (3) 90°K, after Gutkin et. al. [17].

effect tends to cause the quantum efficiency of detectors to increase at high photon energies. The response of two silicon photodetectors in the vacuum ultraviolet region from 7 eV to 19 eV is shown in Figure 2.15 . For these particular devices the minimum efficiency occurred between 11 and 12 eV and was about 0.04 for one diode and about 0.14 for the other diode. These curves combined with those of Figure 2.13 show that silicon p-n junctions have a useful photocurrent response from the band gap energy far out into the vacuum ultraviolet.

Optical sensors using p-n junction may be operated in either a current mode or a voltage mode. If the device is operated at zero volts or at some reverse bias voltage, the sensor acts basically as a current generator with the generated current proportional to the incident light flux. From Equation (2.51) and (2.52)

$$J_{rev} \simeq - J_o - \frac{q\, n_i\, W_J}{(\tau_{po}\, n_1 + \tau_{no}\, p_1)} - q\, \alpha\, \bar{L}\, N. \qquad (2.64)$$

The first two terms represent the reverse current of a normal p-n junction while the third term is the collected photocurrent. The lower limit on detected photon flux N is determined by the reverse bias current since the photocurrent must be larger than the reverse bias current in order to be detected. As a general rule the reverse bias current increases as temperature increases so operation at low temperatures is required for the detection of low light intensities. Also the reverse bias current tends to increase as the semiconductor band gap is reduced. This makes it difficult to detect low light

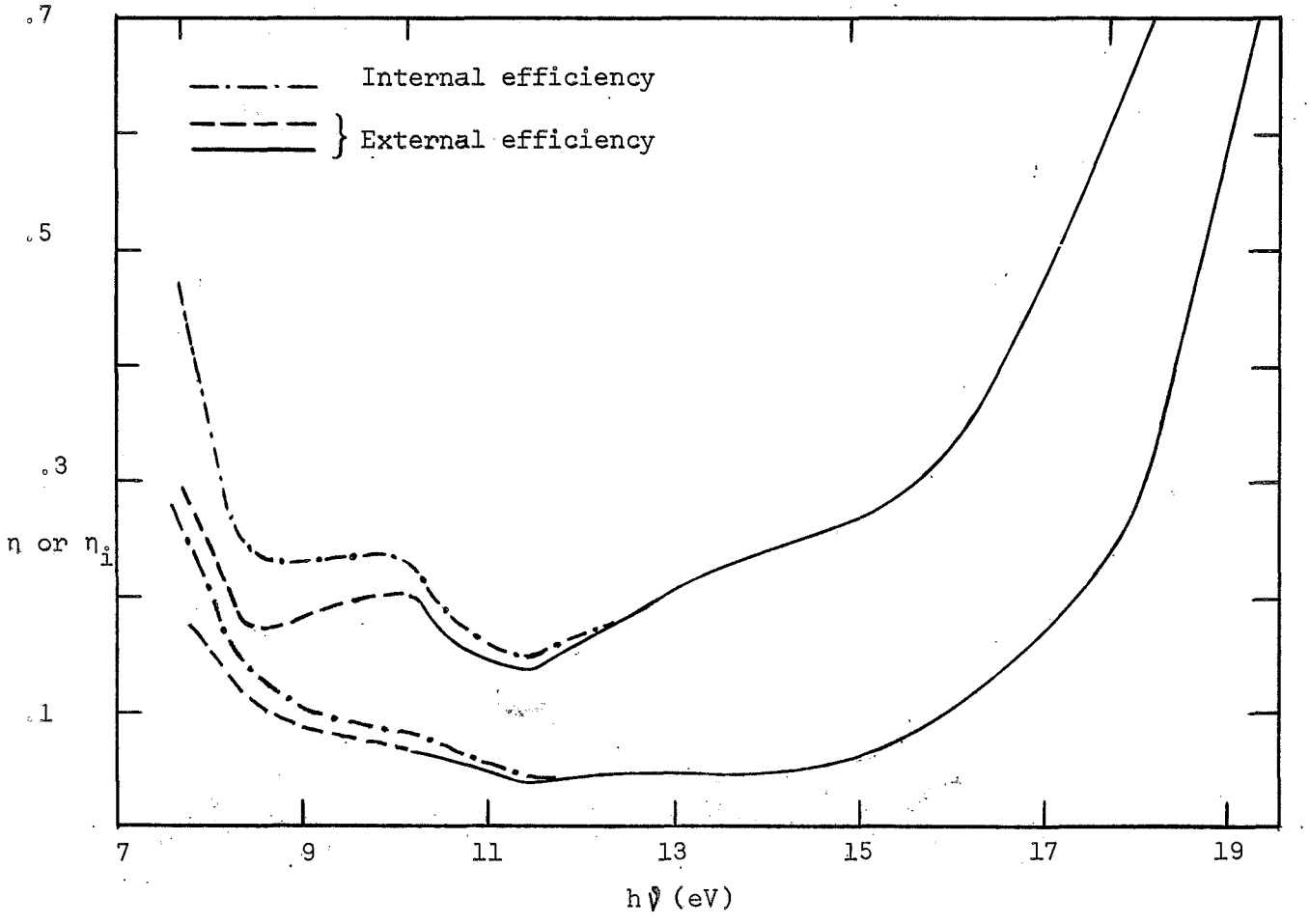


Figure 2.15. Measured and calculated values of diode sensitivity. Both the measured quantum efficiency η and the calculated internal quantum η_i efficiency are shown. After Morse, et.al. [19]

intensities in the infrared region unless cooling is also employed to reduce the reverse bias current. The problem of signal detectivity is discussed in more detail in a later part of this section.

In the voltage mode of operation the open circuit voltage can be used as a measure of the incident light flux. At low forward voltages the generation-recombination current tends to dominate and the photovoltage as given by Equation (2.51) - (2.53) is

$$V_{oc} \simeq \frac{2kT}{q} \ln \left(\frac{\tau_{po} + \tau_{no}}{2 n_i W_J} q \alpha \bar{L} N \right), \tag{2.65}$$

or

$$V_{oc} \propto \frac{2kT}{q} \ln N. \tag{2.66}$$

Thus the open circuit voltage depends approximately logarithmically on the photon flux. In most applications this is probably a disadvantage of the voltage mode of operation. In applications, however, where the incident photon flux can vary over several orders of magnitude, the logarithmical dependence might be used to advantage. At large values of forward bias the normal diffusion current dominates and

$$V_{oc} \simeq \frac{kT}{q} \ln \left(\frac{q \alpha \bar{L} N}{J_o} \right). \tag{2.67}$$

In either case the dependence is logarithmic on photon flux.

Figure 2.16 shows experimental data on open circuit voltages for a junction optical sensor as a function of illumination intensity [20].

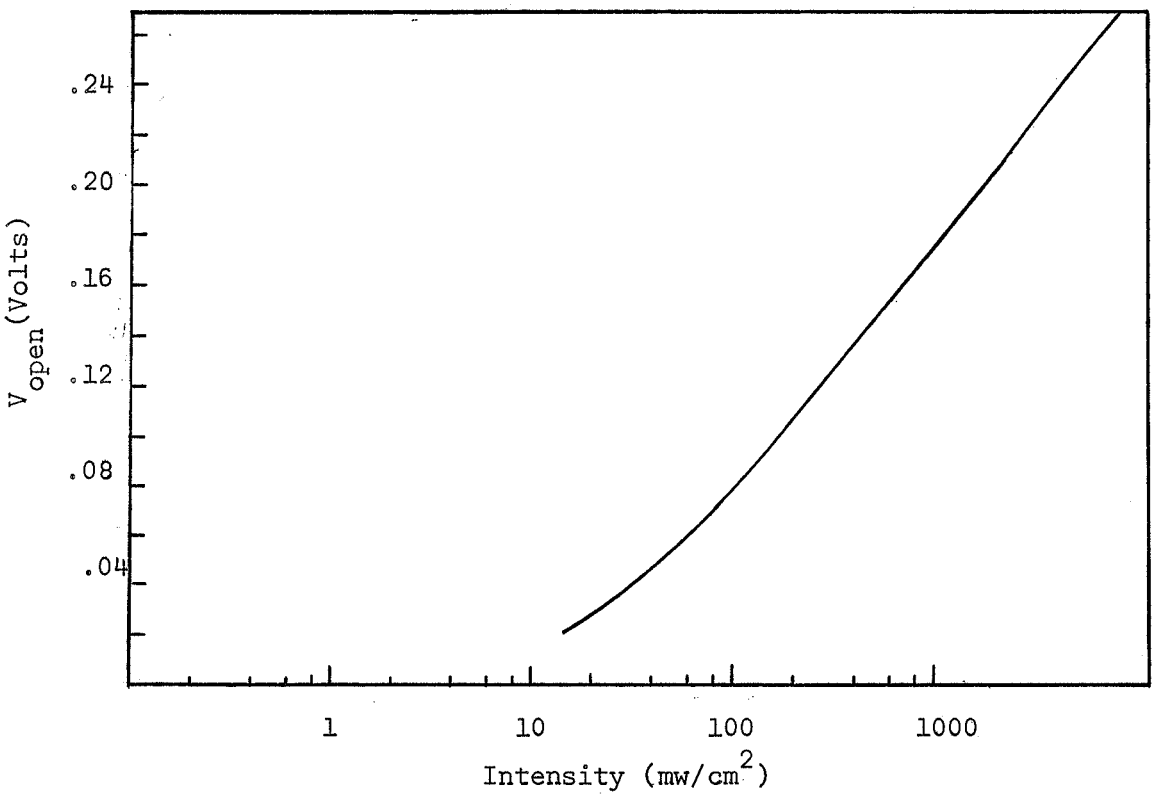


Figure 2.16. Open circuit voltage vs. illumination intensity [20].

The straight line on the graph, except at very low voltage levels, indicates the logarithmic dependence of the open circuit voltage on the light flux.

2.1.2.3 p-i-n Photodiode

The p-i-n type of junction photodiode requires special mention because the depletion region thickness can be tailored to optimize the sensitivity range and the frequency response of the device. A typical structure of a p-i-n photodiode is shown in Figure 2.17 [21]. In a practical device the i region will be slightly n- or p-

type. The p-type region is assumed to be very thin compared with a diffusion length. Figure 2.17 (b) shows a sketch of the exponentially decreasing light intensity within the semiconductor. The basic difference between the p-i-n diode and the normal p-n junction diode is the width of the depletion region. If the i region is close to an intrinsic region, the depletion region at fairly small reverse bias voltages will extend all the way across the i-region. The equations previously developed for the p-n junction (Equations (2.51) - (2.53) for example) can be applied to the p-i-n diode if W_J is taken as the width of the depletion layer which is approximately equal to the width of the lightly doped i region for most values of reverse bias. When surface recombination is small and the p region is so thin that little absorption occurs within this region, the photocurrent is approximately

$$J_{photo} \simeq q\, N \left[1 - \frac{e^{-\alpha W_J}}{(1 + \alpha L_p)}\right]. \qquad (2.68)$$

For a large photoresponse it is desirable to have αW_J ($\simeq \alpha W_i$) large so that most of the light is absorbed within the wide depletion region.

The ability to tailor the spectral response of a p-i-n detector is shown in Figure 2.18 [22]. This shows the relative spectral response as a function of wavelength for a GaAs p-i-n type detector at reverse bias voltage of 0 and -30 volts. This particular device was an n^+-n-p^+ device but a p^+-p-n^+ device should behave similarly. The explanation for the observed behavior is as follows. At zero reverse bias, the depletion region is relatively narrow and is located deep within the semiconductor at the n-p^+ junction (or i-n boundary in

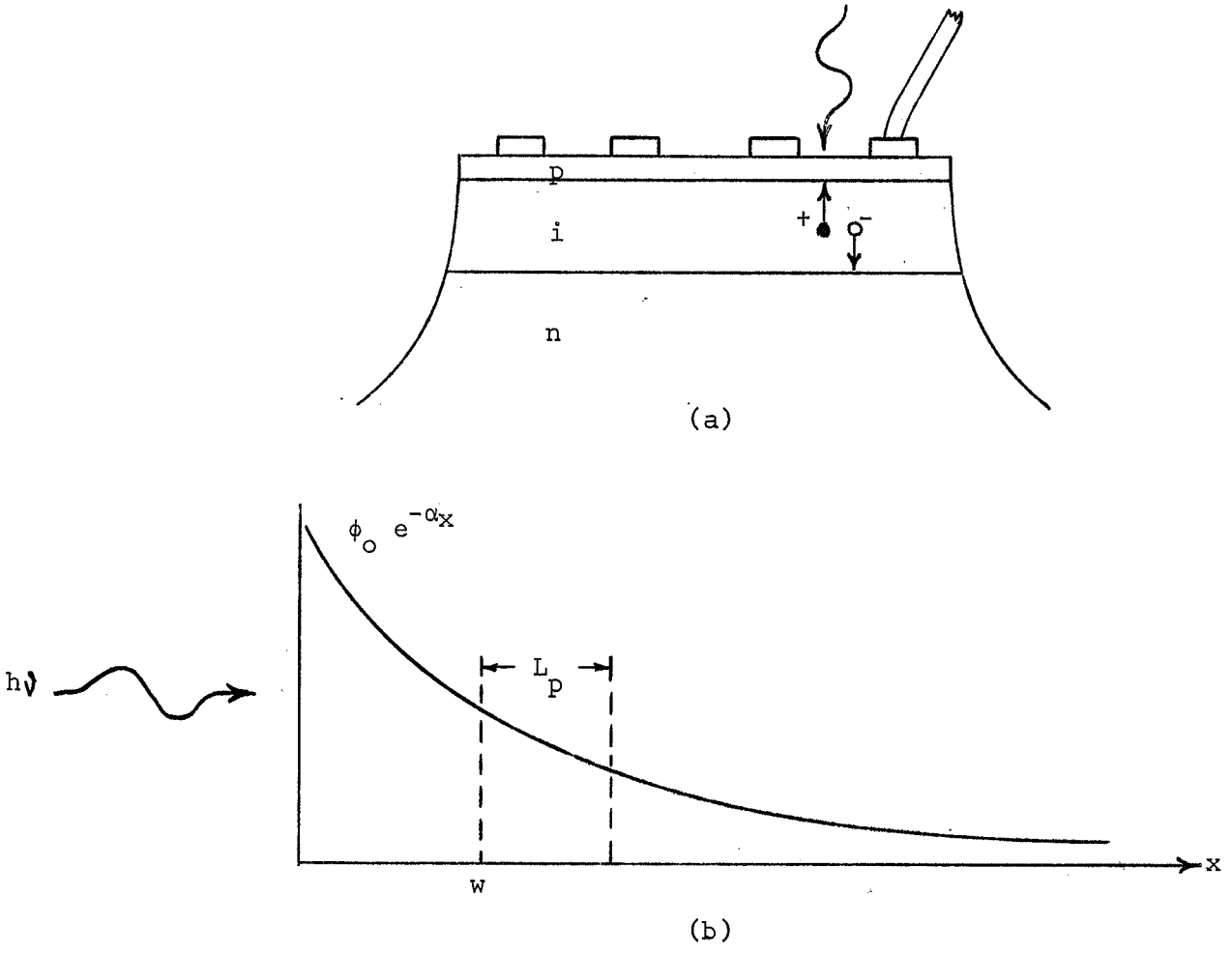


Figure 2.17. Semiconductor p-i-n photodiode. (a) Basic structure (b) Light intensity versus distance (w is depletion layer width)

Figure 2.16 for the p-i-n device). The response increases in the 9,000 to 8,800 A° region because of the increased absorption within the semiconductor at energies above the band gap energy. The decrease response (V_B = 0) for wavelengths shorter than about 8,800 A° is due to the increased absorption near the surface and the decreased fraction of the light able to reach the depletion region. At a reverse bias of -30 volts, there is a very wide depletion region, and even though

the absorption is near the surface it is still within the wider depletion region and is able to contribute to the collected photocurrent.

This example illustrates that it is possible to fabricate fairly narrow band photodetectors if the junction depletion region is deep within the semiconductor. For such a case, the semiconductor region between the depletion region and the surface acts as an optical filter preventing short wavelength light from reaching the active junction region. For such a narrow band detector the minority carrier lifetime should be short such as in the direct band gap III-V materials in order to obtain narrow band response. On the other hand for a detector with

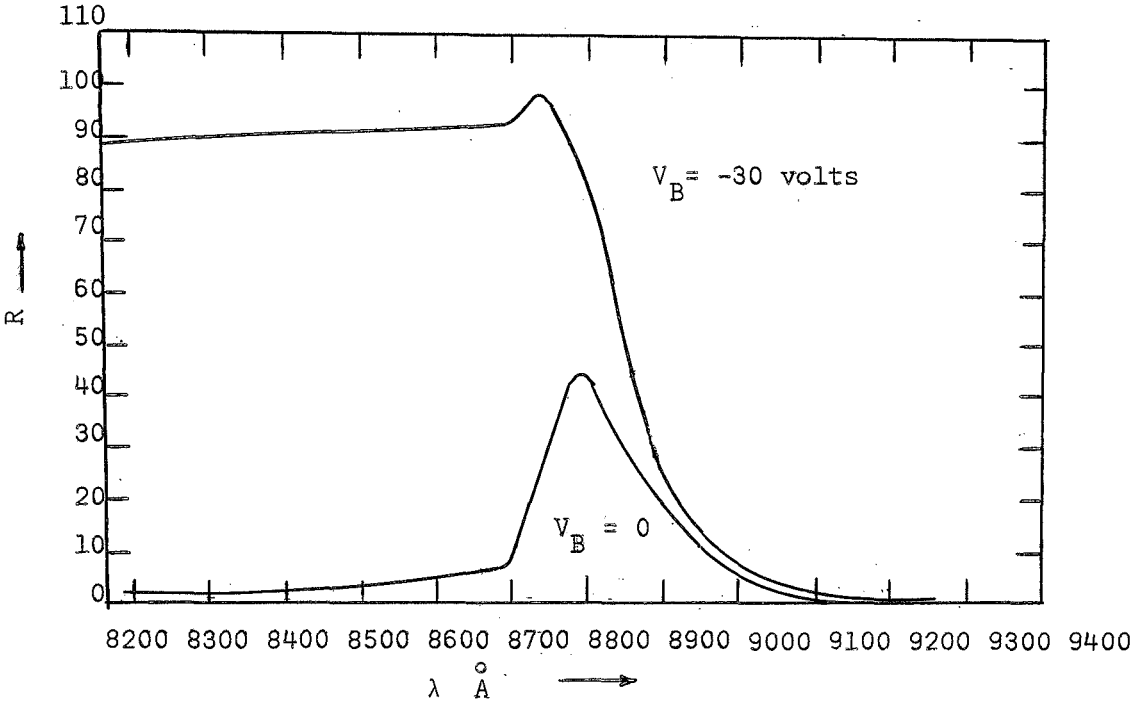


Figure 2.18. Experimental results on change in relative response (R) with bias (V_B) for n+-n-p+ type GaAs detector [22].

a broad wavelength response, the depletion region should be close to the surface and wide. Also the lifetime in both the bulk and at the surface should be long for a broad band detector.

2.1.2.4. Avalanche Photodiode

The avalanche photodiode consists of a normal photodiode operated at such a large reverse bias that avalanche multiplication takes place. Within the avalanche region each carrier passing through the high field of the space change region is increased by some factor due to avalanche multiplication. This occurs, not only to the normal thermal current component, but also occurs to the carriers generated by the incident photons. The avalanche photodiode thus has an internal gain mechanism and is the most promising solid state photodetector with an internal gain.

The cross-section of a typical planar avalanche photodiode is shown in Figure 2.19 [23] [24]. The device consists of a standard shallow n+ diffusion over the entire junction area plus a deeper n-type guard ring diffusion around the edge of the shallow diffusion. The purpose of the guard ring diffusion is to suppress edge breakdown resulting from junction curvative and to give more uniform breakdown over the entire junction area. The elimination of microplasmas or small areas where the breakdown voltage is less than that of the whole junction is one of the major problems in fabricating good avalanche photodiodes. Microplasmas can be minimized by using substrate material of uniform doping concentration, low dislocation density and by designing the active area to be no larger than is

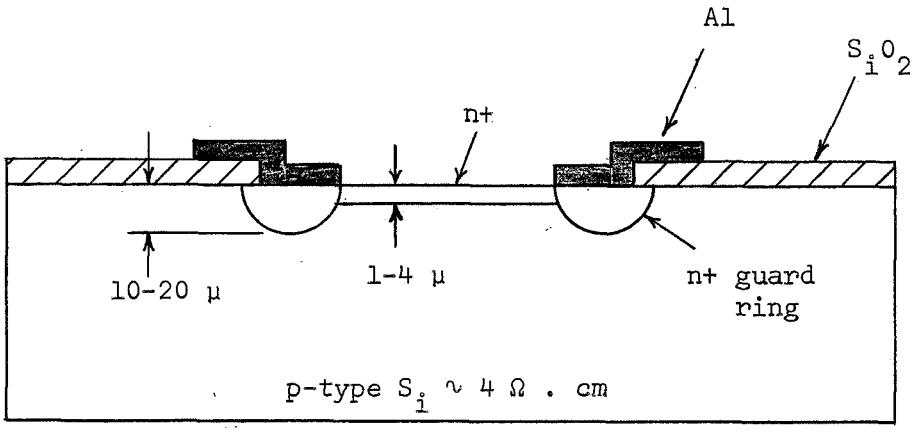


Figure 2.19. Cross-sectional view of planar photodiode (Ref. [23][24]).

needed to accomodate the incident light beam.

The ratio of total junction current, I to the current I_p which flows in the absence of avalanche multiplication is given by the factor M as

$$M = \frac{I}{I_p} = \frac{I_{ph} + I_d}{I_{pho} + I_{do}} = \frac{1}{1 - \left(\frac{V-IR}{V_B}\right)^n}, \qquad (2.69)$$

where I_d is the multiplied dark current, I_{ph} is the multiplied photocurrent, and R is the series resistance of the diode. The dependence of M on voltage has been discussed in Section. 2.1.1.4 (see Equation (2.30)). The additional factor here is the presence of the IR term to account for a finite series resistance. A typical set of I-V characteristics for a Ge photodiode at various values of I_p are shown in Figure 2.20 .

Under the condition that $I_{pho} >> I_{do}$ and $IR << V_B$, the multiplication can be written as

$$M \simeq \frac{I}{I_p} = \frac{}{1 - \left(\frac{V}{V_B}\right)^n \left(1 - \frac{nIR}{V}\right)} \quad . \tag{2.70}$$

The maximum multiplication will occur near $V = V_B$ and if we call this value of multiplication M_{opt} this becomes

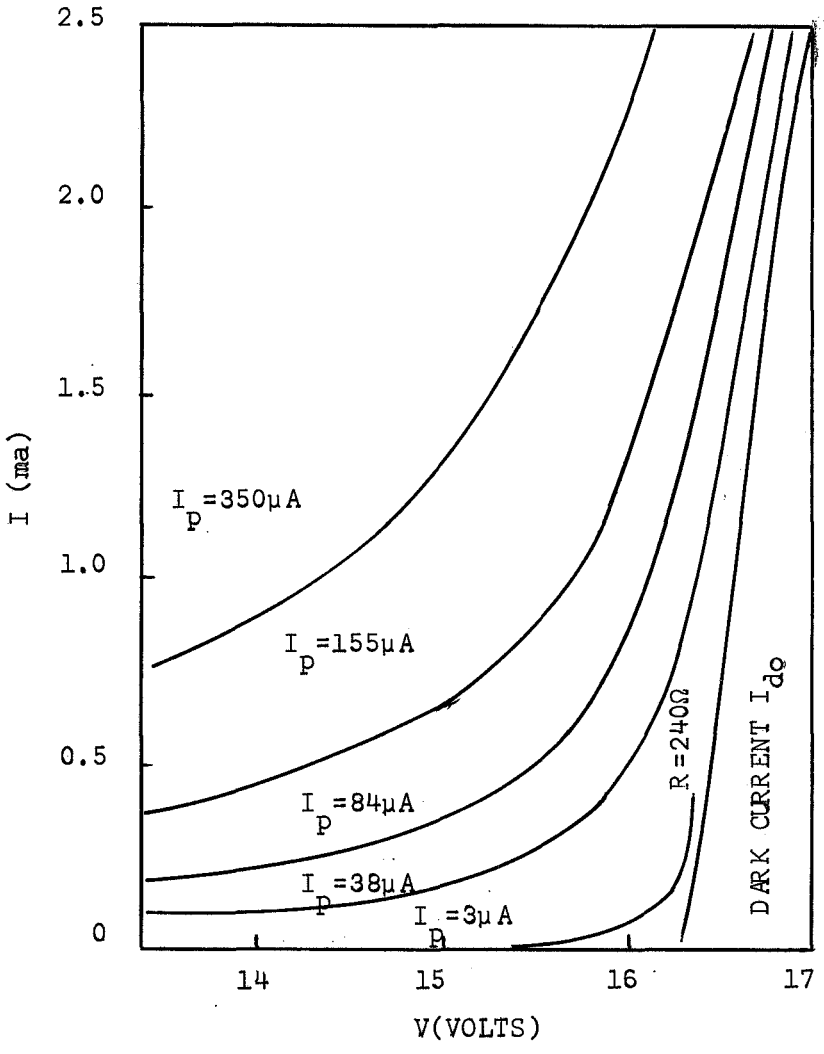


Figure 2.20. Static current-voltage characteristics of a Ge photodiode for various primary currents [16].

$$M_{opt} \simeq M\Big|_{V\to V_B} \simeq \frac{1}{\frac{nIR}{V_B}} = \frac{I}{I_p} . \qquad (2.71)$$

Solving the above equation for the ratio I/I_p gives

$$\frac{I}{I_p} = M_{opt} \simeq \sqrt{\frac{V_B}{nI_pR}} . \qquad (2.72)$$

This gives the maximum photomultiplication and this is found to depend inversely upon the square root of I_pR. Thus to obtain large multiplication I_p and R should both be as small as possible. The above equation shows that the maximum multiplication decreases as the normal photocurrent I_p increases. Experimental data verifying this is shown in Figure 2.21 [25]. As seen in the figure, the maximum multiplication which can be achieved decreases with increasing I_p or correspondingly with increasing light intensity. At low light levels it appears that multiplications on the order of 100 can be achieved. This is limited as can be seen from Equation (2.71) by the series resistance and the dark currents.

2.1.2.5. Frequency Response of Photodiodes

The maximum frequency at which a photodiode will respond to light changes is an important parameter especially in the use of photodiodes in detecting light pulses or in various types of optical communication systems. Consider first the normal p-n junction type of photodiode such as shown in Figure 2.11. If the depletion region is sufficently thin, the frequency response will be limited by the diffusion time

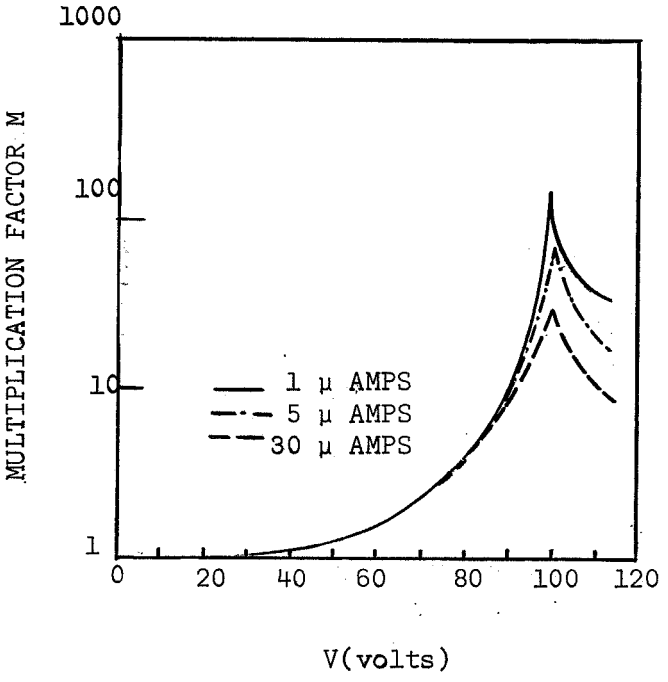


Figure 2.21. Typical multiplication vs. bias for avalanche photodiode. The different curves correspond to different dc photocurrents measured at low reverse voltage [25].

of carriers to the depletion region. The frequency response is then somewhat dependent on the wavelength. At wavelengths close to the band edge the carriers of importance are those created within the p-type region and those created within a distance L_p of the depletion region on the n-type side. The longest diffusion time will be that associated with holes in the n-type region and the 3-dB frequency will be approximately

$$f_{3dB} \simeq \frac{2.4\, D_p}{2\pi\, L_p^2}. \tag{2.73}$$

At short wavelengths the absorption is confined largely to the p-type region near the surface. In this case the frequency response will be

limited by the diffusion of carriers across the p-type region or

$$f_{3dB} \simeq \frac{2.4\ D_n}{2\pi\ W_p^2} \quad . \tag{2.74}$$

For a typical p-n junction, the maximum response of the junction occurs when $W_P\ \alpha \simeq 1$. Using this in the above expression gives

$$f_{3dB} \simeq 0.4\ \alpha^2\ D_n. \tag{2.75}$$

High frequency response then requires a high absorption coefficient or small W_p and a large diffusion constant or correspondingly a large mobility. As an example of the magnitude of the cutoff frequency consider silicon with $\alpha = 10^4$/cm and $D_n = 10$ cm^2/sec. Then

$$f_{3dB} \simeq 4 \times 10^8/\text{sec} = 400\ \text{MHz}. \tag{2.76}$$

Consider next the frequency response of the p-i-n type of photo-detector shown in Figure 2.17. For this type of detector the frequency response is limited by the carrier transit time t_r through the depletion region. Then

$$f_{3dB} \simeq \frac{2.4}{2\ \pi\ t_r} \simeq 0.4\ \frac{v_{s1}}{W}\ , \tag{2.77}$$

where W is the width of the depletion region and v_{s1} is the saturation velocity of the carriers. For a silicon device with $W = 10^{-3}$ cm and $v_{s1} \simeq 10^7$ cm/sec one finds

$$f_{3dB} \simeq 4 \times 10^9/\text{sec} = 4\ \text{Ghz}. \tag{2.78}$$

This is an order of magnitude larger than the cutoff frequency for the standard p-n junction.

For a given modulation frequency and for a given radiation wavelength, one must consider both the wavelength response and the frequency response in making the optimum choice of a detector. Figure 2.22 shows data on the most optimum operating ranges for Ge and Si photodetectors [16]. As expected Ge must be used at long wavelengths while Si becomes the most optimum detector at short wavelengths.

Finally consider the frequency response of the avalanche photodetector. The response is limited by the avalanche multiplication process and by the carrier transit time within the depletion region. The current response within the avalanche zone can be written as [26]

$$i = i_s \frac{M}{1 + j \omega M \tau_1}, \tag{2.79}$$

where i_s is the current without multiplication and τ, is a time constant characteristic of the avalanche process. The 3 dB frequency in then given by the expression

$$f_{3dB} = \frac{1}{2 \pi M \tau_1}. \tag{2.80}$$

Both theoretical and experimental values which have been reported for τ_1 are between 10^{-11} sec and 10^{-12} sec [26]. The above equation shows that carrier multiplication can only be achieved for a particular diode at the expense of the frequency response. The gain-bandwidth product becomes independent of the multiplication factor.

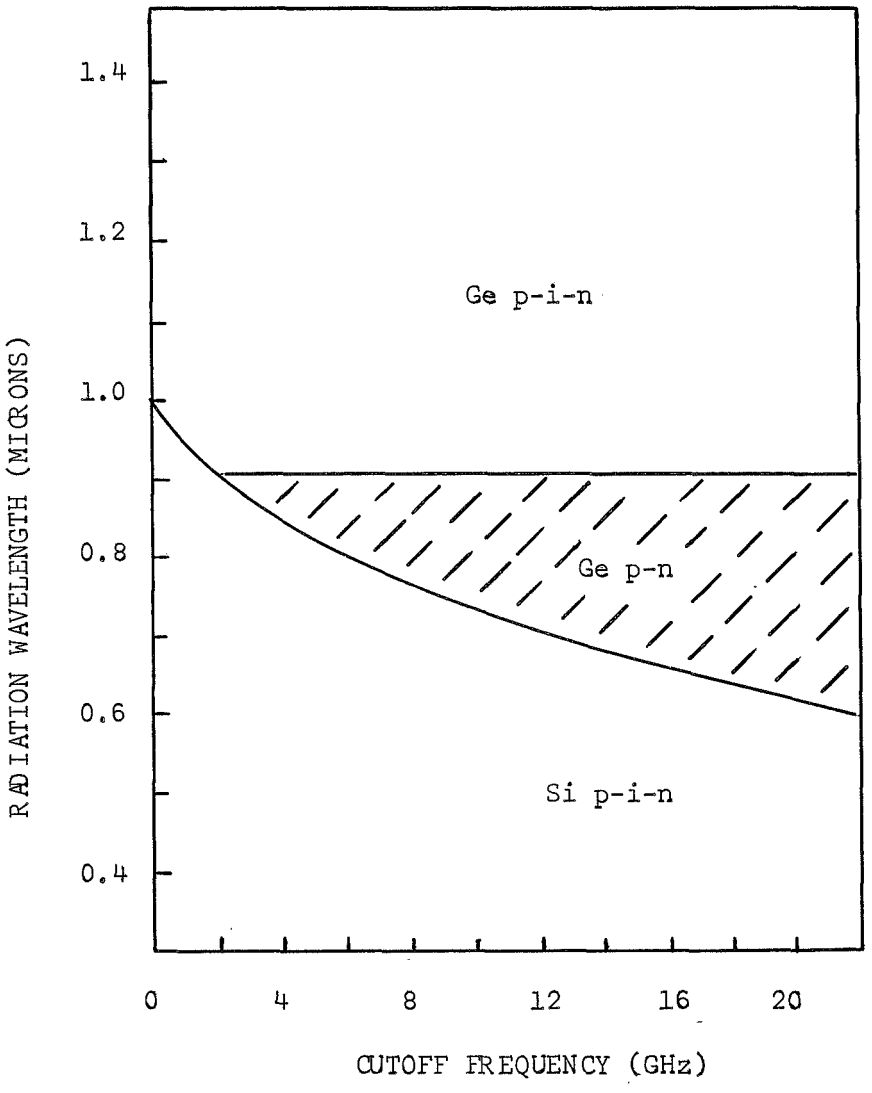


Figure 2.22. Radiation wavelength versus cutoff frequency operating region for Ge and Si photodiodes ([9] page 668).

2.1.2.6. Noise Properties

The major sources of noise in photodiodes are the thermal noise and shot noise. An equivalent circuit showing the noise contribution is shown in Figure 2.23. In the figure R is the device series resistance and C is the device capacitance. Similar noise sources are present in the avalanche photodiode with the exception that the avalanche process increases the shot noise as well as increasing the signal. For an avalanche junction with equal ionization coefficients, it has been shown that the mean square shot noise is [16]

$$\langle i_s^2 \rangle = 2\ q\ B\ M^3\ I_p, \qquad (2.81)$$

where I_p is the photocurrent at low reverse bias voltages. The signal power on the other hand increases as M^2, so the signal-to-noise ratio decreases with increasing multiplication. This is shown in Figure 2.24 for a Ge avalanche photodiode.

There are several figure of merits used to describe the performance of photodectors. One of these is the signal-to-noise ratio. Noise is the ultimate limit to low level sensitivity of a sensor so it is very important to strive for high signal-to-noise ratio. The detector shown in Figure 2.24 has a singal-to-noise ratio at low multiplications of about 20dB. Another figure of merit is the noise equivalent power (NEP) which is defined as the root mean square value of the sinusoidally modulated radiant power falling on a photodetector which gives rise to an rms signal voltage equal to the rms noise voltage from the detector. A final figure of merit which is frequently employed for infrared detectors

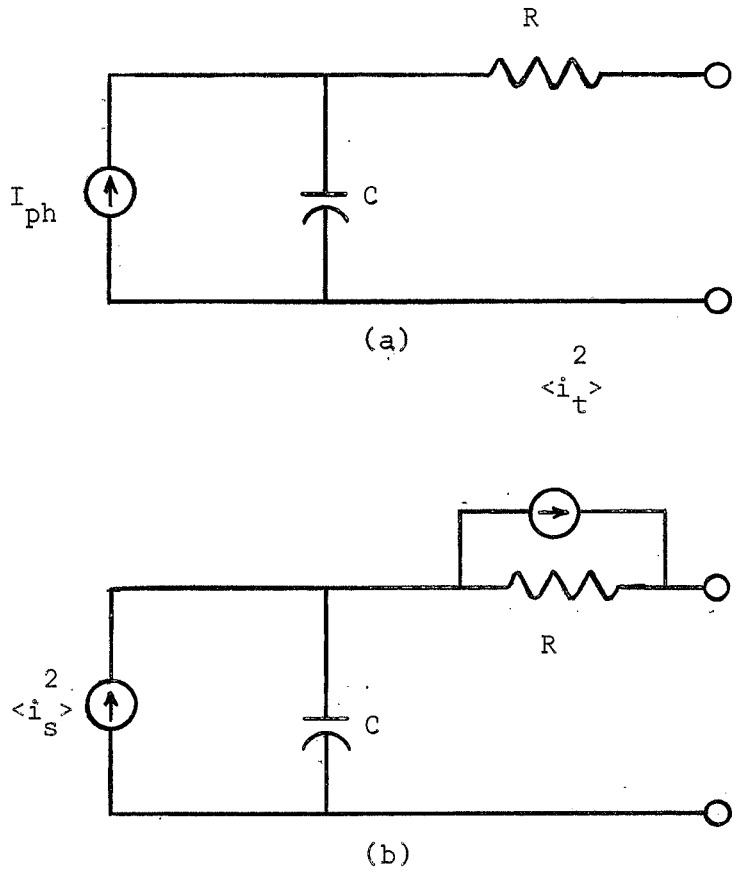


Figure 2.23. (a) Photodiode equivalent circuit. (b) Noise equivalent circuit.

is the detectivity D* defined as

$$D^* = \frac{A^{1/2} B^{1/2}}{NEP}, \qquad (2.82)$$

where A is the area of the photodetector and B is the bandwith. The detectivity depends on the type radiation, i.e. whether it is monochromatic or from a black body source, as well as depending on the modulation frequency.

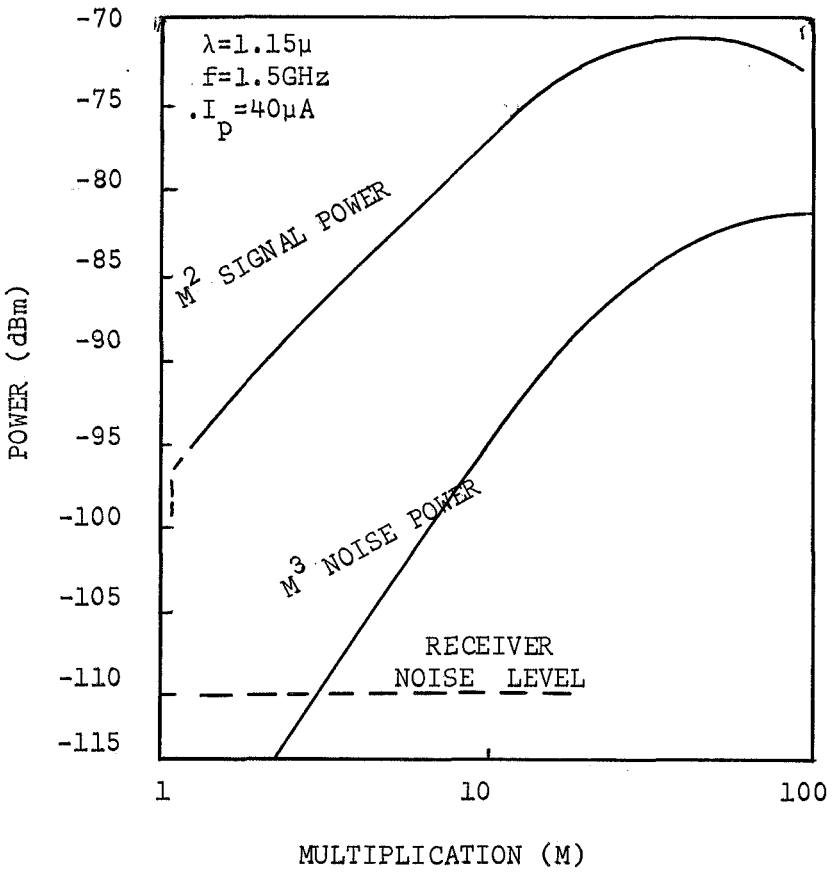


Figure 2.24. Signal and Noise power output of a Ge avalanche photodiode. Both are measured with respect ot 1 mW power level [9].

2.1.3. Particle Sensors

In addition to optical radiation sensors, the use of p-n junctions as particle detectors is one of the most common sensor applications of p-n junctions [27],[28]. As used in this report the term particle sensors is used to include high energy photons or gamma rays as well as material particles such as electrons or alpha particles. The fundamental physical effect responsible for the use of semiconductor junctions as particle detectors is the creation of electron-hole pairs by high energy particles within a semiconductor. Experimentally it is found that the average energy to create an electorn-hole pair is essentially independent of the type of high energy particle producing the pairs and is also essentially independent of the energy of the particle. For silicon this energy is 3.55 ev $\pm$ .05 ev. This means that the number of electron-hole pairs created in the stoping of a high energy particle in a semiconductor is directly proportional to the energy of the particle. For example a 1 Mev particle in Si will produce about 2.8×10^5 electron-hole pairs. For heavy particles at low energies, the number of created pairs is not exactly linear with energy because of a threshold value for pair production. For example α particles in silicon have a threshold value of about 1 keV. As long as the particles total energy is much larger than the threshold energy the total charge produced is essentially linear with total particle energy.

To produce a solid state particles sensor then requires some means to collect the created electron-hole pairs and the p-n junction

or p-i-n junction is commonly employed. In order to measure the total energy of an incident particle, all the created charge must be collected. The range of 1 MeV β particles in silicon is approximately 2 mm so carrier collection must occur over a fairly large distance. The p-i-n diode with its wide depletion region is ideally suited for use as a particle detector, the primary requirement being that the width of the space charge region must exceed the range of the particle to be detected. Such detectors are normally produced in Si or Ge by the Lithium drift technique [12, 13].

The basic principles of operation of the particle detector are similar to the optical radiation detector. As shown in Figure 2.25 for the p-i-n detector, an incident particle creates a series of electron hole pairs along its track. Within the depletion region, the electron-hole pairs are separated by the electric field and constitute a current pulse which flows externally for each incident particle. Some carriers are also created within the surface p+ region and as in an optical detector the width of this region must be less than a diffusion length to efficiently collect the carriers created within this region. Each incident particle creates a current pulse which can be integrated to obtain the total charge within the pulse and thus to obtain a measure of the total energy of the particle. The time response of the detector is determined by the carrier transit time and external measurement circuits.

In additon to detectors for measuring total particle energy, solid state p-n junctions can also be used in particle identification systems. For identifying the nature of a particle use is made of the fact that

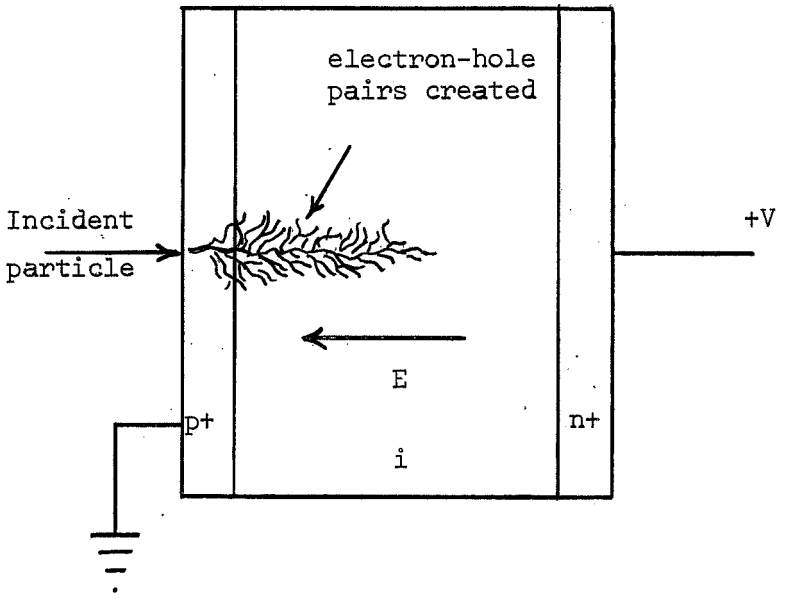


Figure 2.25. Illustration of p-i-n particle detector.

$$E \frac{dE}{dx} \propto m Z^2, \tag{2.83}$$

Where E is the particle energy, m is the nuclear mass and Z is the atomic number. Thus by measuring E and $\frac{dE}{dx}$ particles with different values of the $m Z^2$ product can be detected. By passing a particle through a very thin p-n junction detector, such that the energy loss within the detector is small compared with the total particle energy, an estimate of dE/dx can be obtained. In actual practice of course $\delta E/d$ is measured where δE is the energy loss within the narrow detector of thickness d. The combination of a detector for measuring dE/dx followed by a p-i-n detector for measuring total energy can then be used to obtain a value of mZ^2 for the particle. The typical thickness

of a dE/dx detector is between 25 and 250 μm depending on the application [27].

2.1.4 Temperature Sensors

The basic dependence of p-n junction diode properties on temperature makes it a useful temperature sensor. The forward current of a p-n junction as discussed in Section 2.1.1 can be written as

$$J = J_u + J_n + J_p \tag{2.83}$$

where

$$J_u \simeq \frac{q\, n_i^2\, W_J [\exp(qV/kT) - 1]}{(\tau_{po}\, n_1 + \tau_{no} p_1) + (\tau_{po} + \tau_{no})\, n_i \exp(qV/2kT)} \tag{2.84}$$

$$J_n + J_p = \begin{cases} \left(\dfrac{q\, D_n}{L_n\, P_p} + \dfrac{q\, D_p}{L_p\, N_n}\right) n_i^2 \,[\exp(qV/kT) - 1\,], \\ \qquad \text{at low injection} \\ \left(\dfrac{q\, D_n}{L_n} + \dfrac{q\, D_p}{L_p}\right) n_i \exp(qV/2kT), \\ \qquad \text{at high injection.} \end{cases} \tag{2.85}$$

The exact temperature dependence depends on the operating current range. Consider first a diode operated in the forward bias region below high injection but at a large enough current so that generation-recombination current is negligable. Then

$$J \simeq \left(\frac{q D_n}{L_n P_p} + \frac{q D_p}{L_p N_p} \right) n_i^2 \exp (qV/kT). \tag{2.86}$$

The temperature dependence of the current arises from several terms. The most important terms are the exp(qV/kT) term and n_i^2 where

$$n_i^2 \propto T^3 e^{-E_g/kT}. \tag{2.87}$$

In addition D_n/L_n and D_p/L_p have some temperature dependence. These terms can be written as

$$\frac{D_n}{L_n} = \sqrt{\frac{kT}{q} \frac{\mu_n}{\tau_n}},$$

and (2.88)

$$\frac{D_p}{L_p} = \sqrt{\frac{kT}{q} \frac{\mu_p}{\tau_p}}.$$

Since μ_n and μ_p are temperature dependent these have some slow variation with temperature. For example in lightly doped silicon $\mu_n \propto T^{-2.5}$ and $\mu_p \propto T^{-2.7}$.

The major temperature dependence in the current density comes from the exponential terms and neglecting the other temperature variations gives

$$J \simeq J_o \exp[(qV - E_g)/kT]. \tag{2.89}$$

If the device is operated at a constant current, the voltage is then

$$V = -\frac{kT}{q} \ln \frac{J_o}{J} + \frac{E_g}{q}. \tag{2.90}$$

The junction voltage is then linearly dependent on the temperature. As long as $V < E_g/q$, $J < J_o$ and the term $\ln (J_o/J)$ is a positive quantity. Then

$$\frac{\partial V}{\partial T} = -\frac{k}{q} \ln (J_o/J) \tag{2.91}$$

This indicates that at a constant current the junction voltage decreases with increasing temperature. The slope of the V vs T curve can be changed somewhat by varing the current with small currents giving the largest change in voltage with temperature. Figure 2.26 shows experimental data on the temperature dependence of the voltage of a p-n junction at a constant current [29]. The

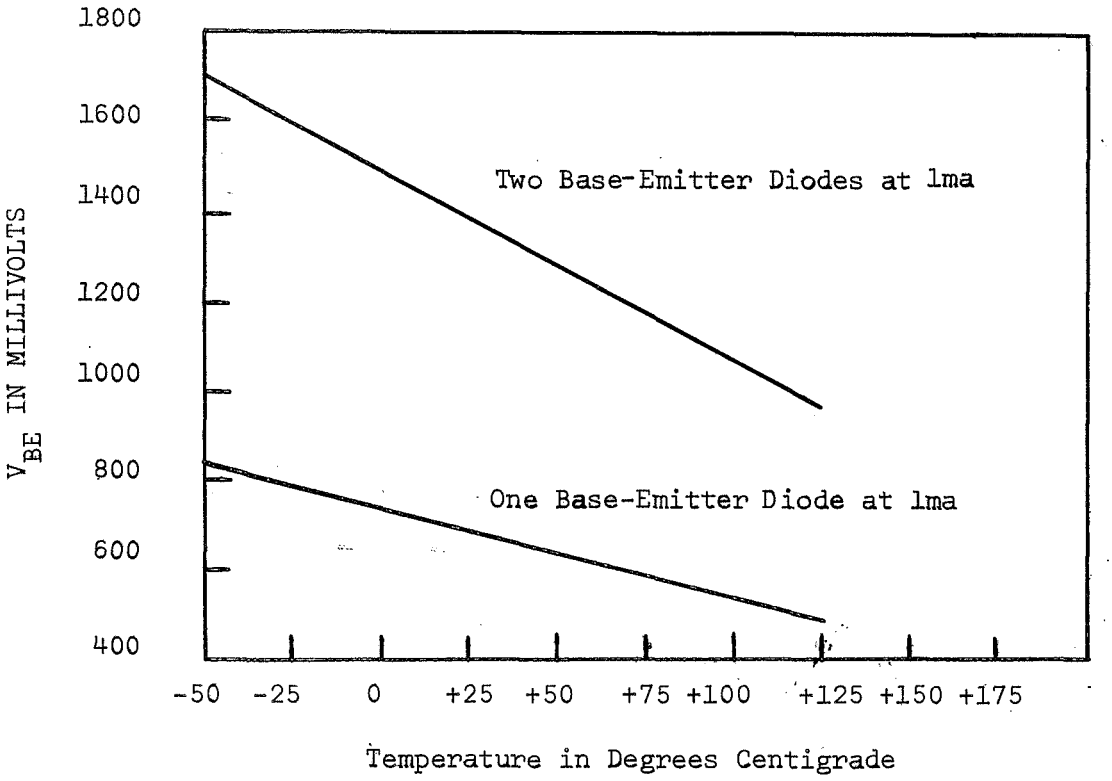


Figure 2.26. Variation of p-n junction voltages with temperature at a constant forward current.

particular data is on an emitter-base junction in an integrated circuit. As can be seen from the figure, the junction voltage decreases essentially linearly with temperature over the range of -50°C to + 125°C as predicted by the theory.

At low forward voltages or at low temperatures, the junction generation-recombination current dominates and in this case

$$J \simeq \frac{q\, n_i\, W_J}{\tau_{po} + \tau_{no}} \exp(qV/2kT). \qquad (2.92)$$

In terms of the temperature dependence

$$J \simeq J_o \exp[(qV - E_g)/2kT] \qquad (2.93)$$

where again the $T^{3/2}$ temperature dependence of n_i has been neglected. In the above case

$$V = - \frac{2kT}{q} \ln\left(\frac{J_o}{J}\right) + \frac{E_g}{q} . \qquad (2.94)$$

This has the same form as Equation (2.90) except for the factor of 2. Thus a diode can be used as a linear temperature sensor with operation in the ideal current region or within the generation-recombination current region. A change in operation between these two regions with temperature will give a dependence which is not linear with temperature.

Diodes made from GaAs have been found to be especially good temperature sensors in the region between 0°K and room temperature [30]. Figure 2.27 shows the voltage across a GaAs diode at 10 μ A forward current as a function of temperature from 0°K to 300°K. The dependence is almost exactly linear except at very low temperatures.

In principle the reverse diode current could be used as a measure of the temperature. However the reverse current is very uncontrollable and this would probably not be a very practical sensor. The temperature dependence of the breakdown voltage however does offer some possible sensor applications in temperature measurement. The breakdown voltage is controlled by either Zener tunneling or by avalanche multiplication. For diodes with breakdown voltages less than about 5.5 volts, the Zener

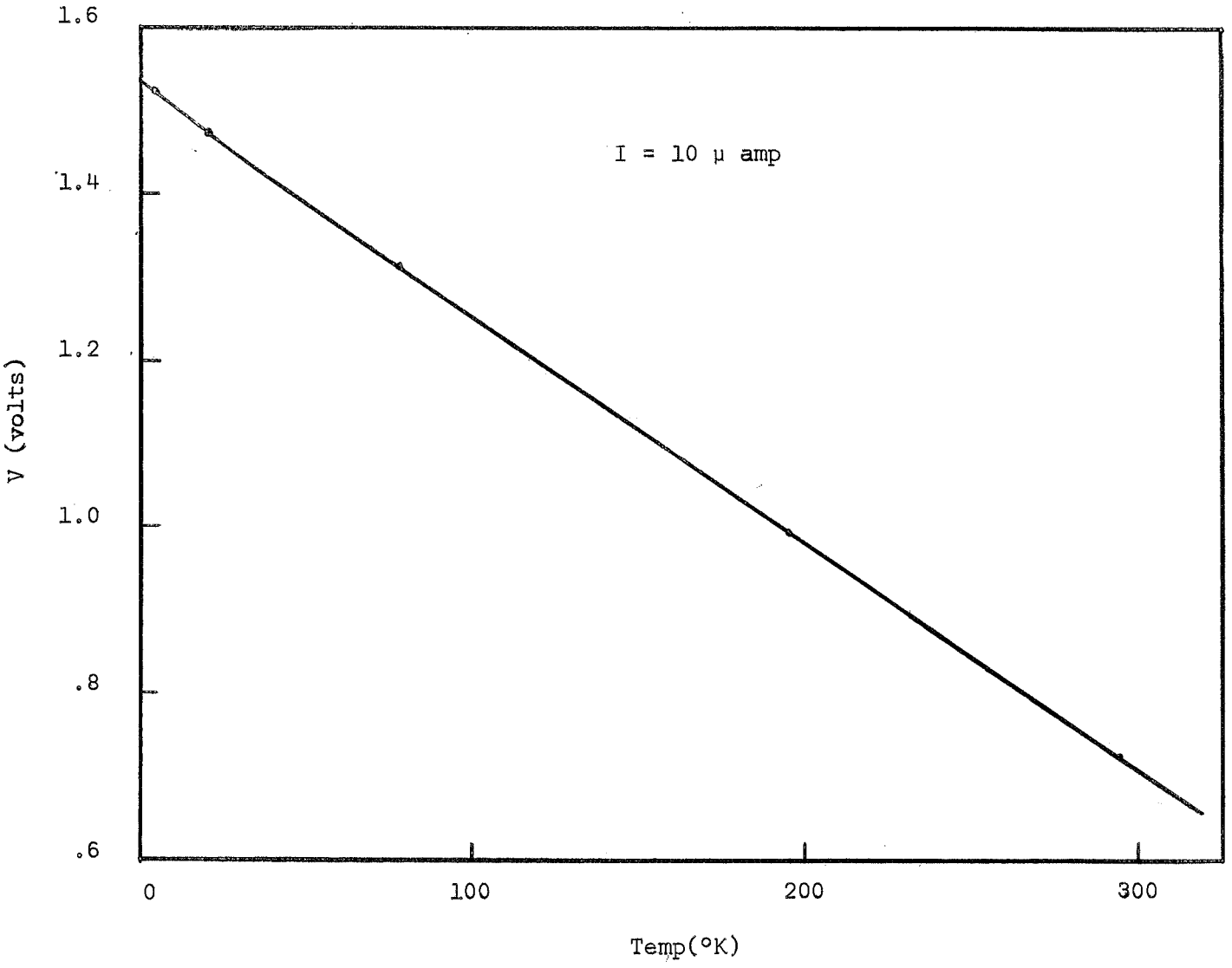


Figure 2.27 Variation of forward voltage for a GaAs diode with temperature [30] .

effect dominates while for diodes with larger breakdown voltages avalanche multiplication dominates. The temperature dependence of the avalanche phenomena is of opposite sign to that of the Zener phenomena so the temperature dependence of the breakdown voltage reverses sign at about 5.5 volts. Figure 2.28 shows typical experimental data on the Zener diode temperature coefficient, defined as

$$\alpha_z = \frac{1}{V_B} \frac{dV_B}{dT}, \qquad (2.95)$$

as a function of zener breakdown voltage [31]. As seen in the figure this can be either positive or negative. Shown in Figure 2.29 is the actual reverse bias voltage of a diode operated at a constant reverse bias current of 100 μ A. This particular diode shows the increased breakdown voltage with temperature which is characteristic of the avalanche phenomena. There is about a 2 mV/°C increase in V_B with temperature. This experimental value is in good agreement with the curves of Figure 2.28.

Both the forward voltage drop across a diode and the reverse breakdown voltage can be used as a temperature sensor. Under either operating condition the voltage change is on the order of a few millivolts per °C. The magnitude of the response in each case can be adjusted over a limited range by adjusting the operating current of the diode or by adjusting the breakdown voltage.

2.1.5. Stress Sensors

This section contains a discussion of the effects of mechanical stress or strain on the characteristics of p-n junctiond diodes. The observations of Rindner [32] [33] and Sikorski [34] of order of magnitude changes in the I-V characteristics of p-n junctions when stressed has stimulated considerable interest in possible applications of p-n junction devices as stress sensors. The major effect which is observed with Ge and Si p-n junctions under anisotropic stress is a large increase in the junction current at any given voltage level. These changes are shown in Figure 2.30 for the forward bias region of a silicon mesa diode with compressive forces from 3 to 34 grams [35]. Changes have also been observed in the reverse breakdown voltage of .

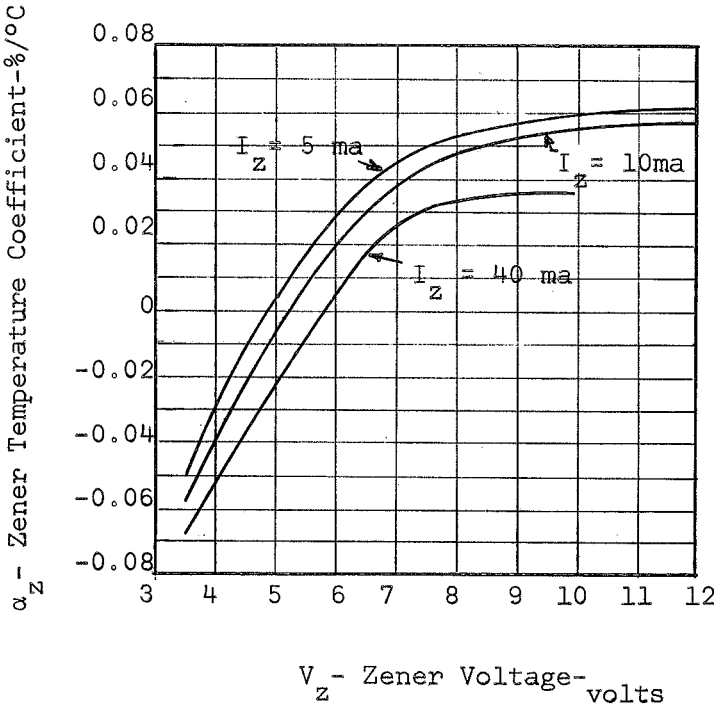


Figure 2.28. Typical zener diode temperature coefficient vs breakdown voltage [31].

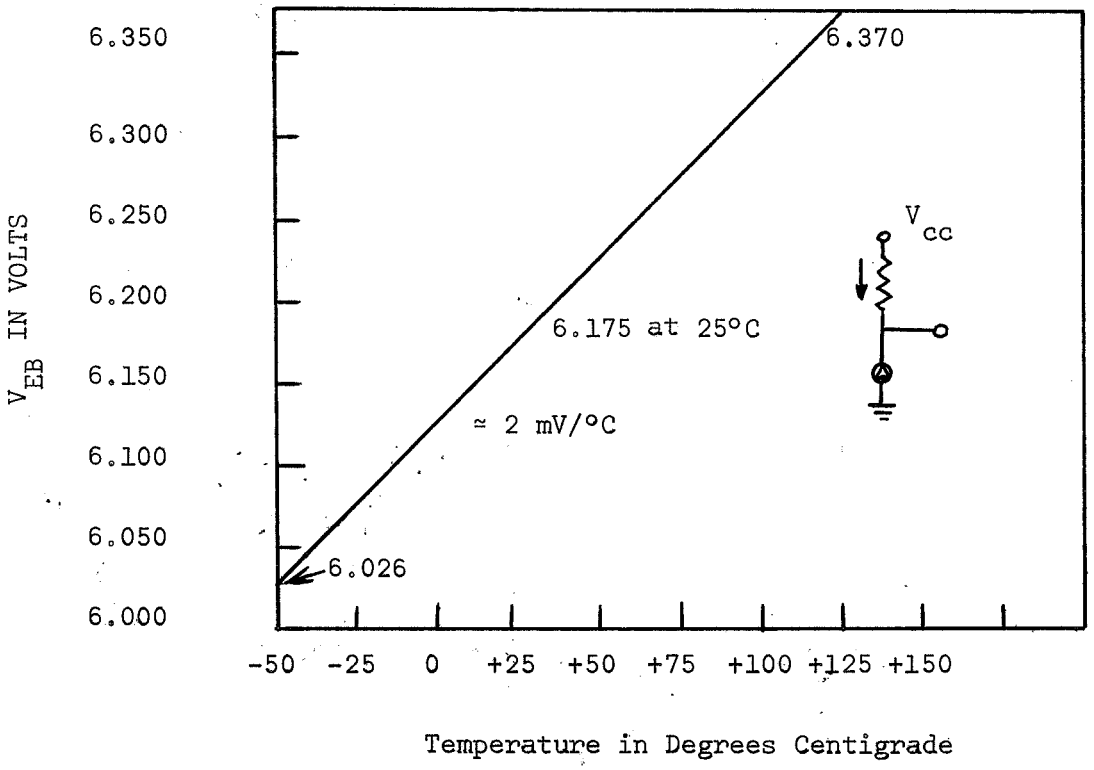


Figure 2.29. Reverse emitter-base breakdown voltage vs. temperature for integrated small geometry transistor [29].

stressed p-n junctions [36]. The effects observed with anisotiopic or uniaxial stress are considerable larger than the effects observed under hydrostatic stress [37] and as shown in Figure 2.30 can result in as much as two orders of magnitude change in the current for a given voltage.

The major effects giving rise to stress induced changes in diode properties are well understood at present. Consider again the equation for current in a forward biased p-n junction as

$$J = J_u + J_n + J_p, \tag{2.96}$$

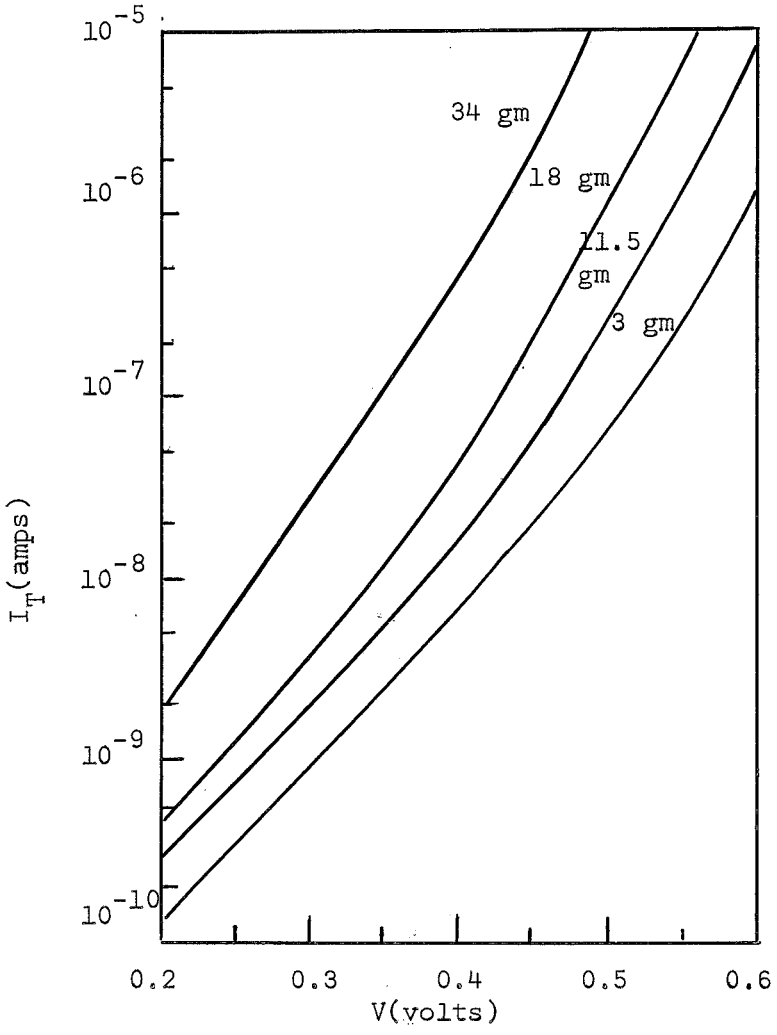


Figure 2.30. Forward-biased I-V characteristic of a mesa diode subjected to stress (the applied force is shown). [35]

where

$$J_u \simeq \frac{q\, n_i^2\, W_J\, [\exp(qV/kT) - 1]}{(\tau_{po}\, n_1 + \tau_{no}\, p_1) + (\tau_{po} + \tau_{no})\, n_i \exp(qV/2kT)}\ , \tag{2.97}$$

$$J_n + J_p \simeq \sqrt{qkT}\ \left(\frac{1}{P_p}\ \sqrt{\mu_n/\tau_n} + \frac{1}{N_n}\ \sqrt{\mu_p/\tau_p}\right)\ n_i^2\ (e^{qV/kT}-1). \tag{2.98}$$

Stress induced changes in the J-V characteristic can rise from several factors. These are n_i, the intrinsic carrier density; n_1 and p_1 which are parameters characteristic of the recombination center; τ_{no} and τ_{po}, the lifetimes in heavily doped material; and μ_n or μ_p, the minority carrier mobilities. All of these factors are probably sensitive to some extent to mechanical stress. The mobilities are known to be stress dependent from piezoresistance measurements on bulk materila. The mobility changes, however, are much too small to account for the large changes observed in stressed p-n junctions.

The factors τ_{no}, τ_{po}, n_1, p_1, τ_n, and τ_p are all related to the impurity centers responsible for recombination in the diode. The minority carrier lifetime can be written as

$$\tau = \tau_{no}\ \left(\frac{p_o + p_1}{p_o + n_o}\right) + \tau_{po}\ \left(\frac{n_o + n_1}{p_o + n_o}\right), \tag{2.99}$$

where p_o and n_o are the equilibrium carrier densities. The values τ_n or τ_p are obtained by using p_o and n_o values in the above equation which apply to either the p-type material or the n-type material respectively. The important quantities are then τ_{no}, τ_{po}, n_1, and p_1 which characterize the trap level. These quantities depend on the position of the defect level with respect to the conduction

and valence bands as well as the capture coefficients and desities of the defects. It is known that stress shifts the energy band edges of the conduction and valence bands and thus must change to some extent n_1 and p_1. Extensive measurements on monority carrier lifetime is Ge and Si has shown that indeed there are changes in minority carrier lifetime under stress [38]. Figure 2.31 shows some experimental measurements on the change in lifetime is silicon as a function of stress in the <100> direction. Similar results were found for stress in other directions, and similar results were observed for Ge. Minority carrier lifetime was found to change by as much as a factor of 3 at stress levels of 10^{10} dynes/cm^2. These effects on minority carrier lifetime have been explained by a model based on stress induced changes in the parameters τ_{no} and τ_{po} with stress.

While the effects discussed above are certainly not negligable at large stress levels, they are not large enough to explain the major stress effects on p-n junctions which are at least an order of magnitude larger than the stress induced changes in mobility and lifetime. The remaining factor is the intrinsic carrier concentration n_i and this has been shown to be the major factor responsible for the stress effects on p-n junctions [35] [39]. The factor n_i^2 in Equation (2.98) is normally expressed in terms of the energy band gap E_g as

$$n_i^2 = N_c N_v \exp(-E_g/kT) \qquad (2.100)$$

where N_c and N_v are conduction and valence band effective densities of states. Based on this equation, a change in band gap of approximately

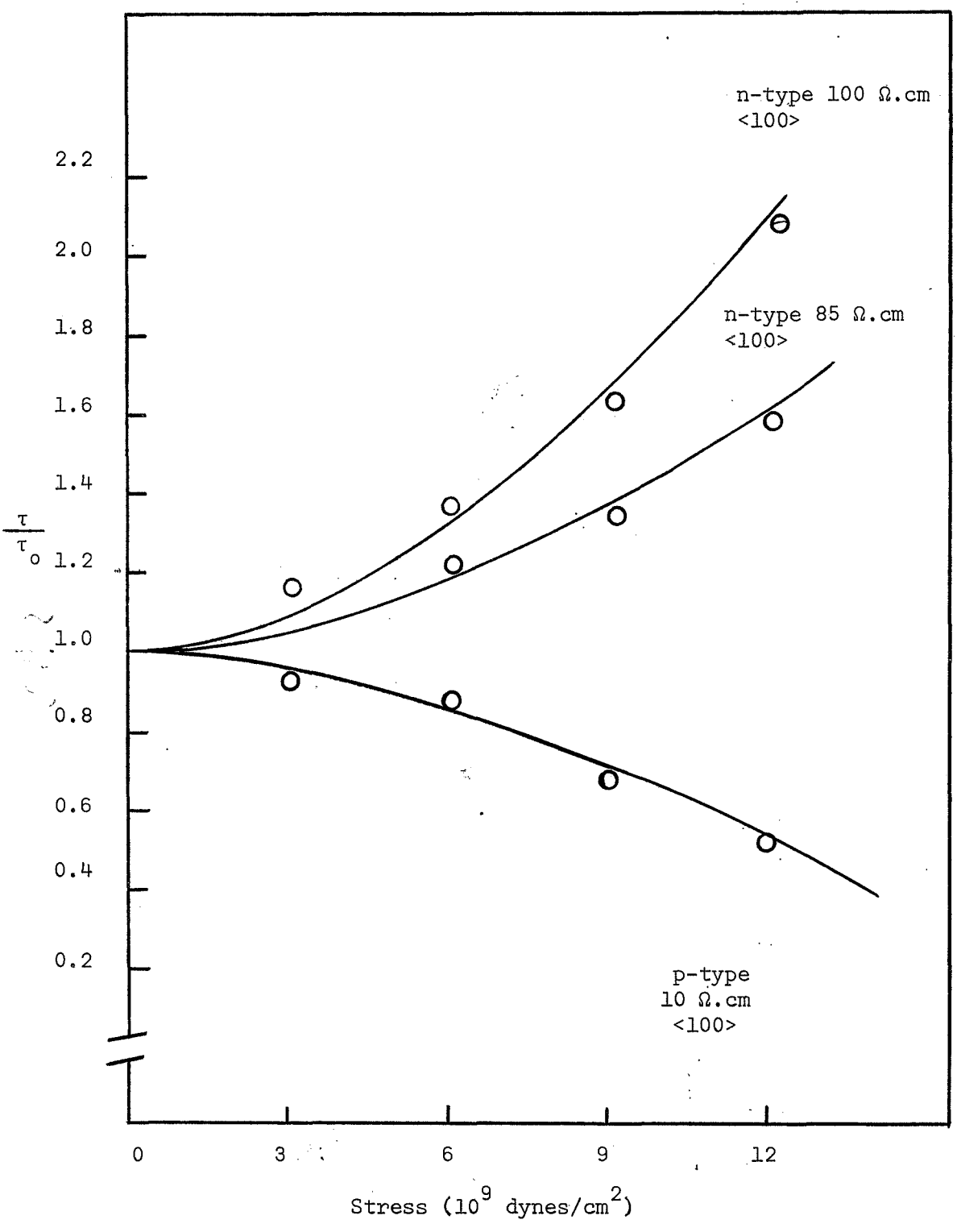
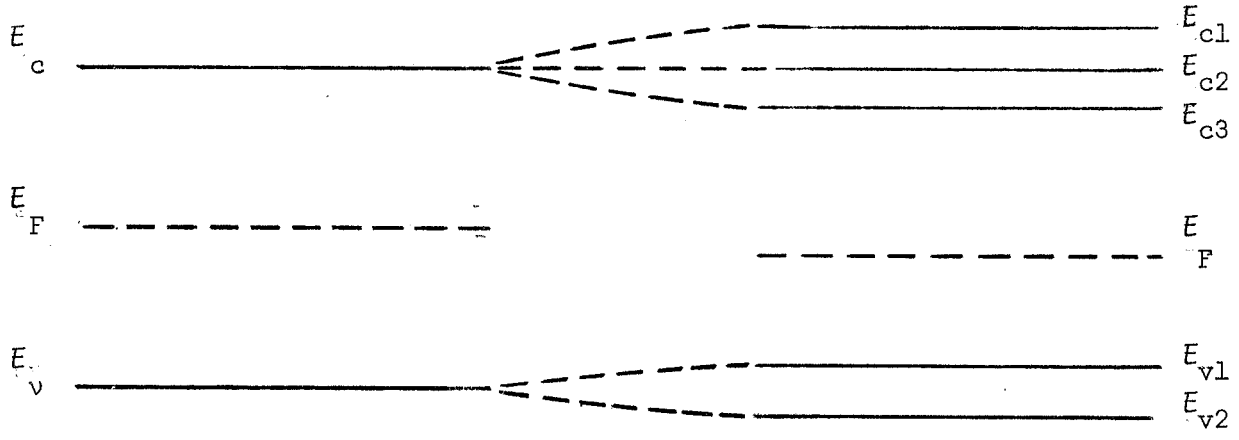


Figure 2.31. Typical changes in minority carrier lifetime with stress for silicon (Stress in <100> direction) [37].

2.3 (kT) or 0.08 ev at room temperature results in an orders of magnitude change in n_i^2 and consequently an order of magnitude change in current density. Because of the exponential dependence on band gap energy this term is the most sensitive to stress and is found to be the dominant stress dependent term at large stresses. This stress effect on minority carrier density and the subsequent stress effects on the I-V characteristic of diodes will be referred to as the piezojunction effect.

In silicon and germanium devices subjected to uniaxial stress the physical situation is slightly more complex than a simple change in E_g in Equation (2.100) with stress. These complications arise from the multiple valley nature of the conduction and valence bands in these materials. In silicon for example the conduction band actually consists of six energy minima oriented along the <100> directions of k-space. In unstressed silicon all of these minima are at the same energy so the conduction band is six fold degenerate. Under a general stress these energy minima shift in energy, resulting, in the most general case, in three sets of energy minima at slightly different energies. The valence band in silicon is also two-fold degenerate and under a general stress this splits into two valence band energy levels. These effects are illustrated in Figure 2.32. Under a general stress the minority carrier density or n_i^2 must be calculated on the basic of three closely spaced conduction bands and two closely spaced valence bands. Such a calculation gives [39]

$$\frac{n_i^2}{n_{io}^2} = \sum_{\substack{i=1,2,3 \\ j=1,2}} \frac{1}{6} \exp[-(\Delta E_{ci} - \Delta E_{vj})], \qquad (2.101)$$



a. No stress b. General stress

Figure 2.32. Effect of a general stress on energy bands of silicon

where each term $(\Delta E_{ci} - \Delta E_{vj})$ corresponds to the changes in energy of the conduction band at energy E_{ci} and the valence band at energy E_{vj}, and n_{io}^2 is the unstressed value of the square of the intrinsic carrier density.

For germanium the situation is similar but slightly different because in this case there are eight conduction band minima in the <111> directions. In the most general case the conduction band splits under stress into four closely spaced conduction band levels instead of the three shown in Figure 2.32. For germanium then

$$\frac{n_i^2}{n_{io}^2} = \sum_{\substack{i=1,2,3,4 \\ j=1,2}} \frac{1}{8} \exp[-(\Delta E_{cj} - \Delta E_{vj})]. \qquad (2.102)$$

For GaAs and other III-V direct band gap semiconductors the conduction band consists of a single energy minimum which is shifted in energy under stress but is not split into multiple levels.

Regardless of the details of the conduction and valence bands for particular semiconductors, the major effect of high stresses on semiconductor p-n junction devices arises from the stress induced shifts in the energy bands and the effects of these shifts on the minority carrier density in the semiconductor. The operation of a p-n junction depends on the minority carrier densities as opposed to normal bulk conduction devices which depend on majority carriers. Piezoresistance arises from the effects of stress on majority carriers and more importantly on majority carrier mobility while stress effects in p-n junctions arise from stress effects on minority carrier densities as well as stress effects on mobility. The changes in minority carrier density with stress can be much larger than mobility cahnges accounts for the large changes observed in stressed p-n junctions.

To determine the effects of stress on the n_i^2 product requires a knowledge of the stress effects on the conduction and valence band energy levels. These effects are well known and are described by deformation potential theory. The energy level shifts of the conduction band are approximately linear with strain and for silicon are [40]

$$\Delta E_{c1} = \Xi_d \varepsilon + \Xi_u \varepsilon_1, \tag{2.103}$$

where Ξ_d and Ξ_u are deformation potential coefficients, $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$, and ε_i is the strain components which are related to the stress components as discussed in Section 1.5.1. The changes for

ΔE_{c2} and ΔE_{c3} are obtained by replacing ε_1 by ε_2 and ε_3 respectively in the above equation. For Ge the expressions are slightly more complex and are given by the expression [40] [41]:

$$\begin{aligned}
\Delta E_{c1} &= (\Xi_d + \Xi_u/3)\ \varepsilon + \Xi_u\ (\varepsilon_4 + \varepsilon_5 + \varepsilon_6)\ /3, \\
\Delta E_{c2} &= (\Xi_d + \Xi_u/3)\ \varepsilon + \Xi_u\ (\varepsilon_4 - \varepsilon_5 - \varepsilon_6)\ /3, \\
\Delta E_{c3} &= (\Xi_d + \Xi_u/3)\ \varepsilon + \Xi_u\ (-\varepsilon_4 + \varepsilon_5 - \varepsilon_6)\ /3, \\
\Delta E_{c4} &= (\Xi_d + \Xi_u/3)\ \varepsilon + \Xi_u\ (-\varepsilon_4 - \varepsilon_5 + \varepsilon_6)\ /3.
\end{aligned} \tag{2.104}$$

For direct band gap materials such as GaAs, only the hydrostatic component of strain is important and

$$\Delta E_c = \Xi_d\ \varepsilon. \tag{2.105}$$

Theoretical expressions for the effects of stress on the valence levels are also available. All of the Group IV semiconductors as well as the Group III-V materials have similar valence band structures and the valence band changes are [40] [42]

$$\begin{aligned}
\Delta E_{v1,2} = D_d\ \varepsilon \pm \{ (\tfrac{2}{3} D_u)^2\ (\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 - \varepsilon_1\ \varepsilon_2 - \varepsilon_1\ \varepsilon_3 \\
-\varepsilon_2\ \varepsilon_3) + \tfrac{1}{3} (D_u')^2\ (\varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2)\ \}^{1/2},
\end{aligned} \tag{2.106}$$

where $\pm$ signs correspond to the two valence levels.

Typical values of the deformation potential coefficients which are needed in the above equations are given below [40]:

Deformation Potential	Germanium (eV)	Silicon (eV)	Gallium Arsenide (eV)
D_u	3.15	2.04	3.15
D_u'	6.06	2.68	5.63
Ξ_u	19.2	11.0	___
$D_d - (\Xi_d + \Xi_u/3)$	4.82	-1.44	___
$D_d - \Xi_d$	___	___	-8.7

These values can be used to calculate the change in minority carrier density which is expressed by the factor γ .

Going back to the current density expression, the forward biased current of a p-n junction can be expressed as [35]

$$I \simeq I_{io}\,\gamma\, \exp(qV/kT) + I_{or}\sqrt{\gamma}\, \exp(qV/2kT), \qquad (2.107)$$

where the stress dependence is contained in the factor γ , and

$$I_{oi} = A\sqrt{qkT}\left(\frac{1}{P_p}\sqrt{\mu_n/\tau_n} + \frac{1}{N_n}\sqrt{\mu_p/\tau_p}\right) n^2_{io}, \qquad (2.108)$$

$$I_{or} = \frac{q\, An_{io}\, W_J}{(\tau_{po} + \tau_{no})} \quad . \tag{2.109}$$

The first term gives the ideal diffusion current while the second term gives the generation-recombination current.

Theoretical values of the factor γ which gives the stress dependence are shown in Figure 2.33 for silicon with assumed stress in the [100], [011], and [111] directions, or with hydrostatic stress [43]. At stress levels of 10^{10} dynes/cm^2 or above, very large changes in γ are predicted. These are consistent with the observed changes on stressed diodes. As can be seen from Equation (2.107) the ideal diode current is proportional to γ while the generation-recombination current is proportional to $\sqrt{\gamma}$ [35]. The ideal curent is thus much more sensitive to stess and piezojunction diodes utilizing this stress sensitive effect should be operated at voltage levels where the ideal current dominates. The typical effect of stress on a diode has previously been shown in Figure 2.30 , where both the ideal current at large forward bias is evident and at low forward bias the current is essentially generation-recombination controlled.

One can define a current gauge factor for the piezojunction effect in much the same way as for a strain gauge as

$$GF = \frac{\Delta I}{I_o \varepsilon} = \frac{\gamma - 1}{\varepsilon} , \tag{2.110}$$

where ε is the uniaxial strain and γ is a function of strain, ε, as well as function of the stress orientation. The theoretical model predicts that the [100] direction in silicon is the most sensitive

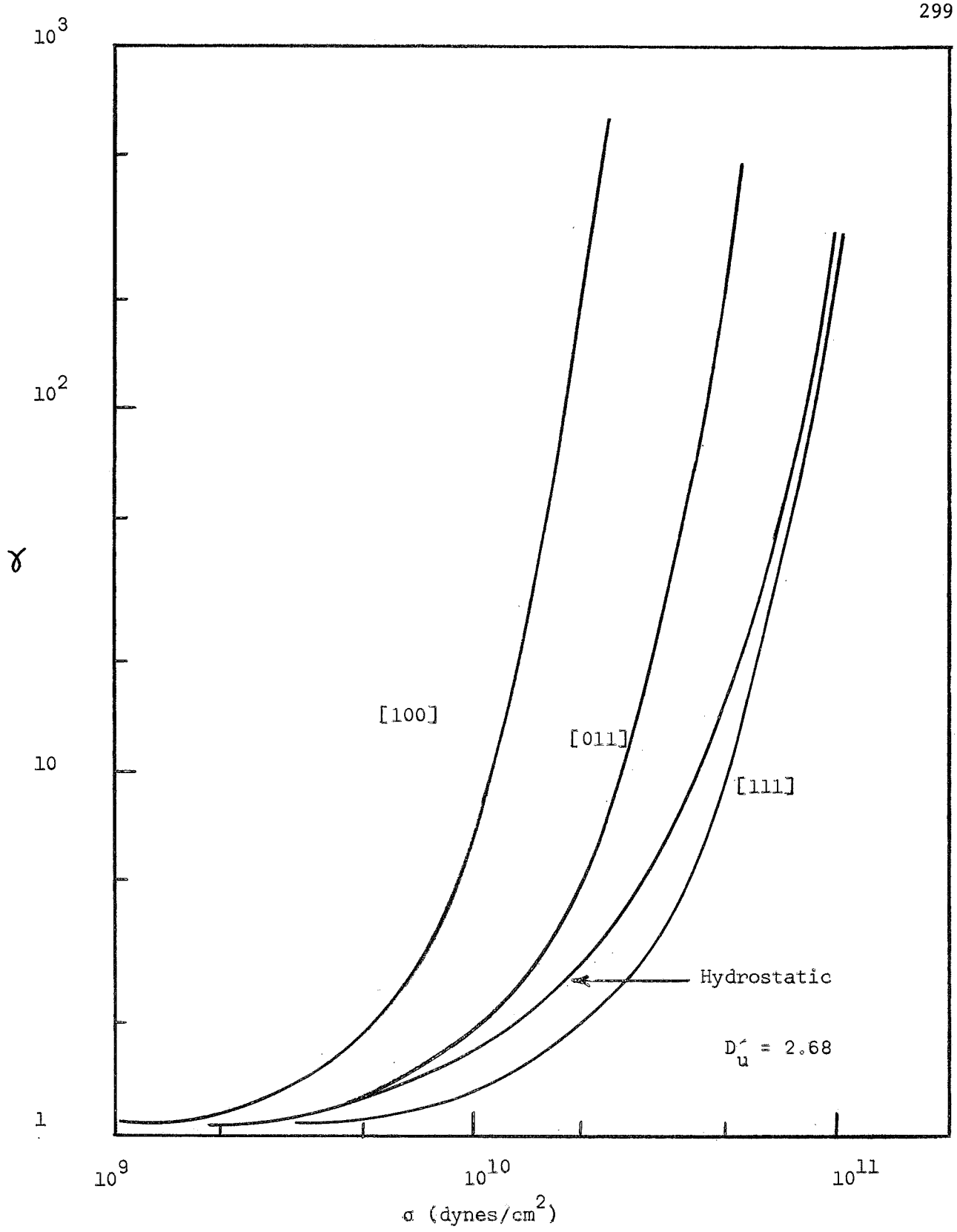


Figure 2.33. Theoretical ratio of stressed to unstressed minority carrier density for a Hydrostatic, [100], [011], and [111] uniaxial, compressive stress for silicon diodes [43].

to stress and this has been experimentally verified [40]. As can be seen from the stress dependence of γ, the above gauge factor is very stress dependent. At very large stress levels, γ becomes essentially exponentially dependent or stress. This can be seen from Equation (2.101) or (2.102). Under any uniaxial stress condition at least one of the $\Delta\varepsilon_{ci} - \Delta\varepsilon_{vj}$ factors will be negative and at large stress levels this factor will dominate giving

$$\gamma = A \exp(C_1 \sigma) = A \exp(C_2 \varepsilon), \tag{2.111}$$

Where A, C_1 and C_2 depend on the stress orientation. Then at large stress levels

$$\frac{\Delta I}{I\varepsilon} \rightarrow \frac{\exp(C_2 \varepsilon)}{\varepsilon} . \tag{2.112}$$

This is basically an exponentially increasing function of stress.

If one is interested only in changes in stress around some bias stress level, a better gauge factor is

$$GF = \frac{1}{I} \frac{dI}{d\varepsilon} = C_2. \tag{2.113}$$

The factor C_2 is the decrease in energy between the lowest conduction level and the highest valence level divided by kT, i.e.

$$C_2 = \frac{D_{vu} - \Xi_{c1}}{kT} \tag{2.114}$$

where $D_{\nu u}$ is the deformation potential of the highest valence level and Ξ_{c1} is the deformation potential for the lowest conduction level. To obtain an estimate of the sensitivity, $D_{\nu u} - \Xi_{c1}$ is on the order of 10 eV so that

$$\frac{1}{I}\ \frac{dI}{d\varepsilon} \sim \frac{10\ eV}{0.026\ eV} \sim 400. \tag{2.115}$$

This can be compared to the gauge factor of a normal semiconductor strain gauge which is on the order of 100. Thus the piezojunction effect is more sensitive than a silicon strain gauge by about a factor of 4 at large stress levels.

One can also define a voltage sensitivity factor by the expression

$$\text{Voltage sensitivity} = \frac{1}{V}\ \frac{dV}{d\varepsilon} \tag{2.116}$$

which is to be evaluated at constant current. At large stress levels this has the value

$$\frac{1}{V}\ \frac{dV}{d\varepsilon} \simeq \frac{kTC_2}{V}\ . \tag{2.117}$$

For a bias voltage of 0.5 volts this has a value of about 20. This is considerably smaller than the piezoresistance effect.

The above calculations have illustrated one of the problems in comparing junction stress sensors with piezoresistance. This is the problem of defining a sensitivity or gauge factor for the nonlinear device. Piezoresistance gauges are normally specified in terms of the resistance sensitivity to strain, although the current sensitivity or voltage sensitivity as calculated above would all give the same value.

With a nonlinear I-V device the sensitivity depends on how the quantity is evaluated. In this work the term current or voltage sensitivity will be used to characterize the piezojunction effect. These closely correspond to practical bias conditions where a device is either operated at constant current and the voltage changes used as the signal or it is operated at a constant voltage and the current changes used as the signal. In any nonlinear device it is extremely important to specify exactly how any specified sensitivity is measured or calculated.

For stress levels up to 2×10^9 dynes/cm^2, it has been shown that a second order expansion of γ is sufficient to describe the piezojunction effect as [44]

$$\gamma \simeq 1 + A\,\sigma + B\,\sigma^2, \tag{2.118}$$

where for [100[stress

$$A = -1.9 \times 10^{-11}\ \text{cm}^2/\text{dyne},$$
$$B = 11.4 \times 10^{-21}\ \text{cm}^4/\text{dyne}^2. \tag{2.119}$$

The experiments used to obtain the parameters were performed on a p+-n junction where piezoresistive effects are small and the results are in good agreement with the theoretical model. Using these values, the piezojunction current gauge factor at low stress is then

$$\frac{1}{I}\ \frac{\Delta I}{\varepsilon} = 24.7. \tag{2.120}$$

This is a factor of 4 to 5 less than the gauge factor of a silicon piezoresistance strain gauge. Thus it can be concluded that the piezojunction diode offers no advantages as far as sensitivity is concerned at low stress levels. The stress level must be around 10^9 dynes/cm^2 or above to have a device which is more sensitive than the silicon strain gauge.

It is also of interest to note the dependence of the stress sensitivity on T^{-1} in Equations (2.113) and (2.114). This is similar to the temperature dependence of gauge factor for silicon strain gauges.

There are several practical problems which must be overcome in using the piezojunction effect. Most of these arise from the large stress levels needed to observe the effect. The fracture stress of silicon is somewhere above 10^{10} dynes/cm^2, but the stress levels for optimum piezojunction operation can easily be within an order of magnitude of the fracture stress. Thus it is easy to over stress the device leading to permanent damage and irreversible effects. Also to achieve such high stress levels at reasonable forces, very small areas must be stressed. Much of the research in this area has been done with the use of small diamond indenter points for applying stress. This approach has been used with some success in constructing devices but is not a very reproducible technique. Another approach has been to fabricate a p-n junction on the end of a silicon needle. Sensors using these devices have been fabricated as accelerometers and pressure sensors [45]. The above approaches have the problem that they are difficult to fabricate plus the problem that only a small part of the total junction area is stressed. When only a small part of the

area is stressed, the devices are not as sensitive as is theoretically possible plus the fact that the spreading resistance around the small stressed area is a problem leading to decreased sensitivity [39].

Probably the best approach to a good practical device is a small area mesa diode with essentially the entire junction area stressed. A device of 1 mil (10^{-3} inches) in diameter requires a total force of about 5 grams to give an average stress level of 10^9 dynes/cm^2. Somewhat smaller mesas can be fabricated so this represents a practical approach at least as far as the size of the total force required is concerned.

In addition to the stress sensitivity of the forward bias current of a p-n junction, the reverse breakdown voltage has also been shown to change with large uniaxial stress. In silicon it has been found both experimentally and theoretically that the breakdown voltage V_B decreases with stress and that [41]

$$\frac{\Delta V_B}{V_B} \simeq - (1 \times 10^{-12} \text{ cm}^2/\text{dyne})\, \sigma, \qquad (2.121)$$

where σ is the magnitude of the uniaxial stress. This represents a much smaller percentage change in V_B than the percentage change is resistance due to piezoresistance. In Ge an effect of comparable magnitude is observed although the effect changes sign at large stress levels.

While the breakdown voltage sensitivity of a diode is small, the current sensitivity of a device biased near breakdown can be quite large. The current near breakdown is

$$I = MI_o = \frac{I_o}{1 - \left(\frac{V}{V_B}\right)^m} \tag{2.122}$$

where I_o is the reverse bias current in the absence of multiplication. Then

$$\frac{1}{I}\frac{dI}{d\varepsilon} = -\,m\,M\left(\frac{V}{V_B}\right)^m \frac{1}{V_B}\frac{d\,V_B}{d\varepsilon} \tag{2.123}$$

$$\simeq -\,m\,M\,\frac{1}{V_B}\frac{dV_B}{d\varepsilon}\,.$$

Thus near breakdown the current sensitivity is a factor of mM times the voltage sensitivity. This can easily be on the order of 100 or 1000, so the current sensitivity can be theoretically about 350 for mM = 10^3.

Stress effects on the tunneling current component have also been extensively studied [46] [47]. The peak tunneling current is found in many diodes to decrease while the valley current increases with increasing stress. The observed effects have been found to be directly related to changes in energy band structure with stress. The tunneling current has been discussed in Section 2.1.1 and Equation (2.31) gives for the current

$$I_t = A\exp\left(\frac{-\,\pi\, m^{1/2}\, E_g^{3/2}}{2\sqrt{2}\;\; q\hbar\, E}\right) V\,(V_n + V_p - V)^2. \tag{2.124}$$

From this the peak current can be expressed as

$$I_p = \frac{4}{27} A(V_n + V_p)^3 \exp\left(\frac{-\pi\, m^{1/2}\, E_g^{3/2}}{2\sqrt{2}\; q\hbar\, E} \right). \qquad (2.125)$$

For small changes in the energy band gap this can be expressed as

$$I_p = I_{po} \exp\left(- \frac{\Delta E_g}{E_t} \right), \qquad (2.126)$$

where

$$E_t = \frac{4\sqrt{2}}{3\pi} \frac{q\hbar\, E}{m^{1/2}\, E_g^{1/2}}, \qquad (2.127)$$

and I_{po} is the unstressed value of peak tunneling current. The above equation is only exact for a direct band gap material with a single conduction band minima and a single valence band maxima, or for a material under a hydrostatic stress. It should be approximately valid however for real materials at low stress levels since at low stresses the important quantity is the "center-of-mass" shift of the energy levels and this is the same for a uniaxial stress as for a hydrostatic stress.

The relative sensitivity of a tunnel diode to stress as compared with the forward biased diode current can be obtained by comparing Equation (2.126) with Equation (2.100). The normal diode current depends exponentially on $\Delta E_g/kT$ while the tunnel current depends exponentially on $\Delta E_g/E_t$. The tunneling current sensitivity can be expressed as

$$\frac{1}{I_p}\frac{dI_p}{d\varepsilon} = -\frac{1}{E_t}\frac{dE_g}{d\varepsilon}. \qquad (2.128)$$

Thus if $E_t < kT$, the tunnel diode is more sensitive to stress than is the normal diffusion current. An estimate of E_t using the following parameters, $m \simeq m_o$, $E_g = 1eV$, $E = 10^6$ volts/cm, gives $E_t \simeq 0.017$ /eV. This is slightly smaller than kT at room temperature, but considering the approximate nature of the values of E and m used in the calculations it appears that the sensitivity of the tunnel diode is approximately the same or slightly larger than that of the diffusion current.

It appears that it might be possible to fabricate a tunnel diode with $E_t < kT$ and thus with a sensitivity exceeding that of the normal p-n junction. The GaSb tunnel diode is the most sensitive tunnel diode which as been experimentally reported [46] [47]. Nathan and Paul observed a decrease in peak tunnel current by a factor of about 4×10^{-3} at a hydrostatic stress level of about 10^{10} dynes/cm^2. This gives a stress sensitivity of

$$\frac{1}{I_p}\frac{dI_p}{d\sigma} \simeq 563 \times 10^{-12} \text{ cm}^2/\text{dyne}. \qquad (2.129)$$

For a uniaxial stress this would correspond to one-third of this value or to about 187×10^{-12} cm^2/dyne. This can be compared to the peizoresistance π_{11} of n-type silicon of 102×10^{-12} cm^2/dyne. Thus the GaSb tunnel diode is slightly more sensitive than the piezoresistance of n-type silicon and slightly less sensitive than the theoretical

predictions at large stress for the diffusion current. Other tunnel diodes have shown lower sensitivity so they will be less sensitive than the piezoresistance effect. Sikorshi [48] has reported slightly larger sensitivity values on GaSb than those discussed above. Typical changes in tunnel diode characteristics with stress for small values of applied voltage are shown in Figure 2.34.

The valley current of tunnel diodes has also been found to be sensitive to stress [47]. Changes in the valley current are less stress sensitive than the peak current and thus are not of much practical significance.

The tunnel diode is a negative resistance device and can consequently be employed as an amplifier. The basic sensitivity to stress of the tunnel diode can be combined with the amplification to obtain quite large voltage sensitivities. Values of several thousand have been reported by Sikorshi [48]. Operated in the amplified mode the tunnel diode provides a very small stress sensor which has an overall sensitivity which is more than an order of magnitude larger than piezoresistance gauge factors.

An important consideration in any sensor is signal-to-noise ratio. There does not appear to be any reported work on the noise properties of piezojunction devices or stress sensitive tunnel diodes. The limiting noise is probably shot noise as with most junction devices. It appears that there is a pressing need for a comparative study of the noise properties of stressed p-n junction devices and piezoresistance devices.

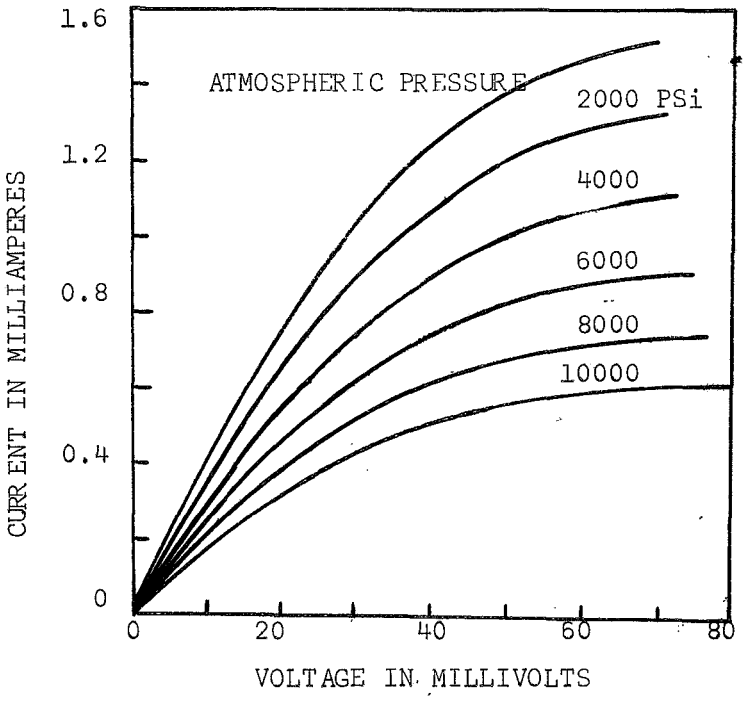


Figure 2.34 Forward I-V characteristics, limited to the first positive resistance region of an n-type GaSb tunnel diode as a function of hydrostatic pressure.

2.1.6 Magnetic Field Sensors

The effects of a magnetic field on p-n junction devices can be grouped into two categories. First, the magnetic field can change the electronic energy states of the semiconductor, and second, the field acts on the current carriers and modifies the transport equations. The second effect is due basically to the Lorentz force on the current carriers and in the presence of a magnetic field the current density equation (Equations (2.3) and (2.4)) are modified to

$$\vec{J}_n = q \mu_n n \vec{E} + q D_n \nabla n + r_n \mu_n \vec{J}_n \times \vec{B}, \qquad (2.130)$$

$$\vec{J}_p = q \mu_p p \vec{E} - q D_p \nabla p + r_p \mu_p \vec{J}_p \times \vec{B}, \qquad (2.131)$$

where r_n and r_p are the Hall coefficients which are on the order of unity. In addition to adding Hall terms to the current density equations, a magnetic field changes slightly the mobility and diffusion constants in the above expressions.

The first effect of magnetic fields on p-n junctions to be discussed arises for a shift of the energy levels in a magnetic field. Under an intense magnetic field the conduction and valence bands are known to split into a series of magnetic sub-bands. The conduction band edge points are increased by the amount

$$\Delta E_c = 1/2 \hbar \omega_c = \frac{q\hbar}{2 m^*} B, \qquad (2.132)$$

where m* is the effective mass and B is the magnetic field. This is a very small shift except in large magnetic fields and for small effective masses. A similar expression also applied to the lowering of the valence bands. The effect of the magnetic sub-bands is best observed in low effective mass materials such as InSb and at low temperatures. The best observations have been on tunnel diodes because of the high sensitivity of the tunnel current to energy band changes. Typical changes in the tunneling characteristics of an InSb diode at 77°K are shown in Figure 2.35 [49]. The tunneling current has been shown to vary as [50]

$$\frac{I(B)}{I(0)} = \frac{B}{B_o} \quad \frac{\cosh\left(1/2\ g \frac{m^*}{m_o} \frac{B}{B_o}\right)}{\sinh\left(\frac{B}{B_o}\right)}, \tag{2.133}$$

where B_o is a nomalization constant which is in the range of 20 to 30 k gauss for InSb diodes. To observe significant changes in the tunneling current the field must be on the order of B_o. Consequently the effects of magnetic field on the energy band structure are only observed at large magnetic fields.

At magnetic fields below 10 k gauss the most important magnetic field effects on diodes arise from the field induced changes in the transport equation. As previously mentioned these arise from magnetic field changes in mobility, diffusion constant, and in the additional Hall term involving $\vec{J} \times \vec{B}$ in the transport equations. The magnitude of the Hall effect is proportional to μB. For a magnetic field of 10 k gauss and a mobility of 10^3 $cm^2/v.$ sec this product has a value of 0.1, so the Hall effect is a small perturbation except in

high magnetic fields or for materials with a very large mobility such as InSb.

Consider first the effect of the Hall term on the I-V characteristic of a diode. This effect can lead to a nonuniform distribution of current across a diode as can be seen with the aid of Figure 2.36 . The diode current flow in the magnetic field gives rise to Hall voltages in the p- and n-type material of opposite sign. The Hall voltage polarities are idicated in Figure 2.36 by $\pm$ signs. On side A_1 in Figure 2.36 the Hall voltage adds to the applied junction potential while it subtracts on side A_2. This leads to an enhanced current flow on one side of the junction (side A_1) and a reduced current flow on the other side. This effect has been analyzed

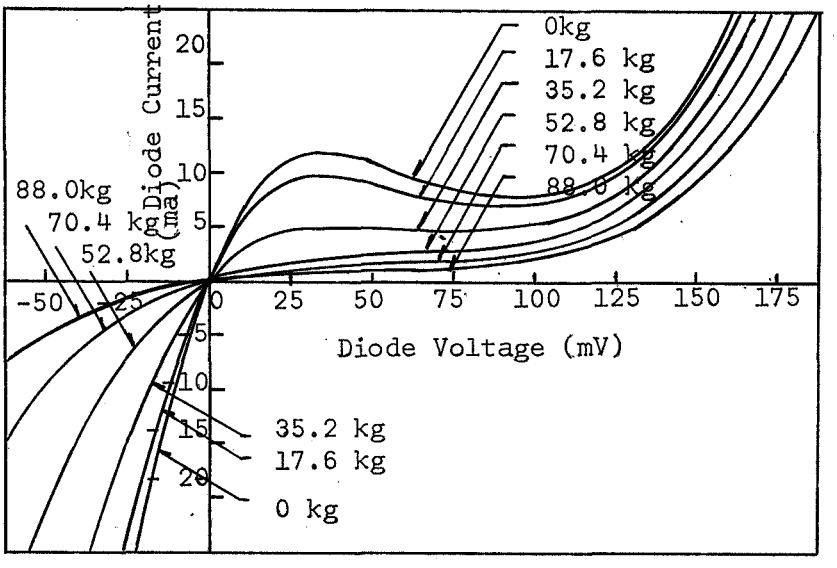


Figure 2.35. Current-voltage characteristics at 77°K for the InSb tunnel diode in different magnetic fields perpendicular to the current flow across the p-n junction [49].

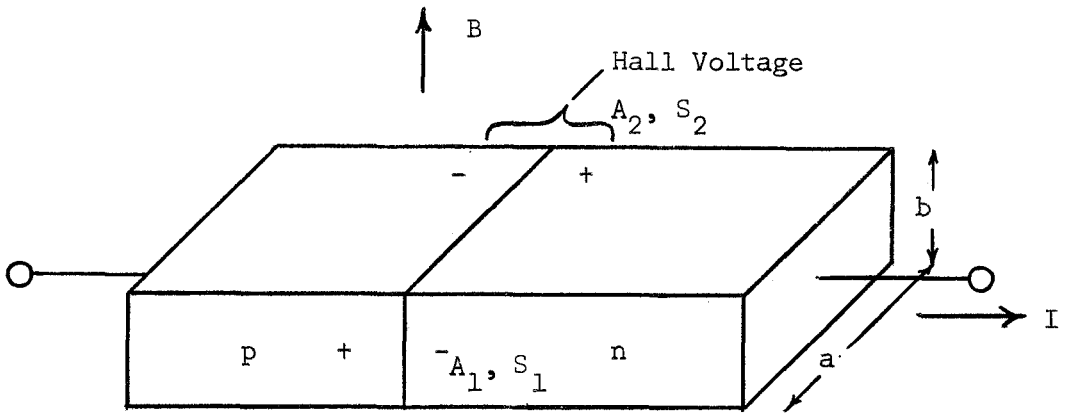


Figure 2.36. p-n Junction showing current flow and magnetic field

in detail by Garfinkel and Engeler [51]. Calculated current distributions across the junction are shown in Figure 2.37 for various values of V_H $\beta = qV_H/kT$ where V_H is the combined Hall voltage across the sample given by the expression

$$V_H = B \left(\frac{r_n}{qn} + \frac{r_p}{qp}\right) \frac{I}{b} \quad . \tag{2.134}$$

As can be seen from the figure when the combined Hall voltage is on the order of kT/q or larger, there is a very significant change in the current distribution across a p-n junction. Series resistance which is neglected in the calculations leading to Figure 2.37 has the effect of reducing the nonuniformity of the injected current [51].

The redistribution of current however does not produce a first order change in the terminal I-V characteristics. This can be seen as follows. For small magnetic fields the increase in current on one side of the junction is equal to the decrease in current on the other

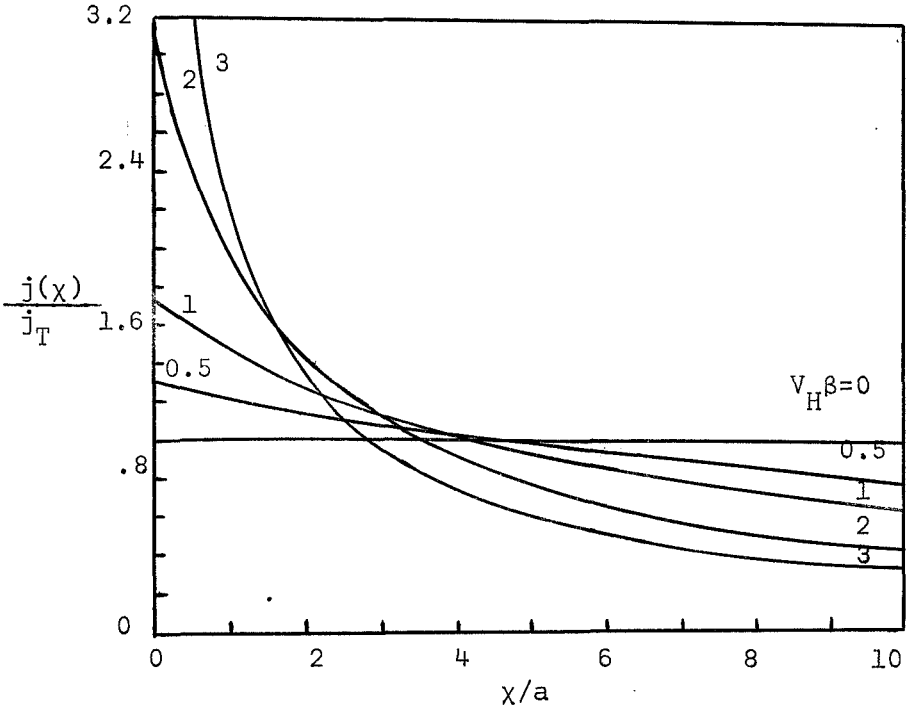


Figure 2.37. Calculated current distribution for a p-n junction in a magnetic field (series resistance neglected) [51].

side of the junction and thus the total current remains constant. When $\beta V_H \gtrsim 1$ there will be a net change in total current but this change is proportional to β^2 so there is no linear change in device properties due to this redistribution of current.

The redistribution of current at a p-n junction combined with surface recombination can lead to a first order change in diode properties with magnetic field. This can again be understood with the aid of Figure 2.36. Assume that the two sides of the p-n junction have different surface recombination velocities S_1 and S_2. The current redistribution at the junction causes recombination at surface A_1 to be enhanced while recombination at surface A_1 is reduced. If the surface recombination velocities are not equal, i.e. $S_1 \neq S_2$, there will be a net increase or decrease in the total junction

current. If $S_1 > S_2$ the current increases while $S_1 < S_2$ leads to a decreased current. This is a first order effect which is directly proportional to the magnetic field. The physical geometry of most diodes is quite complex, especially for diffused diodes in the region of the junction near the surface. For this reason it is very difficult to qualitatively describe the effects of surface recombination on device characteristics in a magnetic field. However the effect, if present, should vary linearly with magnetic field and can lead to either an increase or decrease in current with magnetic field.

In a long thin semiconductor bar, surface recombination can be combined with bulk recombination to give an effective fillamant carrier lifetime. For surface recombinations which differ on two sides of the bar, a magnetic field causes a linear change in net surface recombination. This in turn will lead to a linear change in fillament lifetime with magnetic field. Since the diffusion length is related to lifetime by the expression $L = \sqrt{D\tau}$, the diffusion length will also show a first order linear change with magnetic field. Misra and Srivastava have also argued that there should be a linear change in minority carrier diffusion constant with magnetic field because of the basic tensor nature of the diffusion constant [52]. Both of these effects give rise to first order linear changes in diffusion length with magnetic field which can be expressed as

$$\frac{\Delta L}{L} = A\ B + C\ B^2 + -- \qquad (2.135)$$

where A describes the linear dependence which will dominate at low magnetic fields. Experimental evidence tends to confirm the fact that there is indeed a linear change in minority carrier diffusion length with magnetic field [52] [53]. Decreases of about 10% in diffusion length have been observed at fields of 10K gauss and these changes appear to be linear with magnetic field [53].

Normal p-n junctions as commercially produced show small changes in I-V characteristics with magnetic fields. Parshad and Mehta have investigated silicon and germanium diodes and find small changes in the I-V characteristics [54] [55]. Essentially no change was observed in the forward bias current of silicon diodes at fields of 8 K gauss. Changes of as much as 30% were observed in germanium diodes at the same field but the changes were observed to vary as B^2 and not directly with magnetic field.

Normal p-n junctions are not extremely sensitive to small changes in diffusion length or to saml1 changes in surface recombination and this appears to be the reason they are relatively insensitive to magnetic fields. Recognizing this problem, Karahushan and Stafeev [53] [56] have investigated a type of junction diode which is very sensitive to small changes in diffusion length and consequently they have observed large changes in diode properties with magnetic fields. The basic diode structure is essentially a long or wide base diode constructed from a germanium fillament as shown in Figure 2.38. The wide base region of the diode is a high resistivity region of 50 Ω. cm Ge and is in the range of 0.1 cm in length. The diode is essentially a p-n junction with a long high resistivity region forming the lightly

doped side of the junction.

Under forward biased conditions, the diode has considerable conductivity modulation of the high resistivity region and this is the unique feature of this particular diode. The properties of such a diode are very strongly dependent on the minority carrier diffusion length and consequently on magnetic field. This can be understood quantatively as follows. Suppose for example that a magnetic field decreases the diffusion length slightly. This reduces the conductivity modulation of the lightly doped region. This in turn causes a larger voltage drop across the bulk resistance and a smaller fraction of the applied voltage appears across the junction depletion region. This in turn causes a reduced injection and a further reduction in

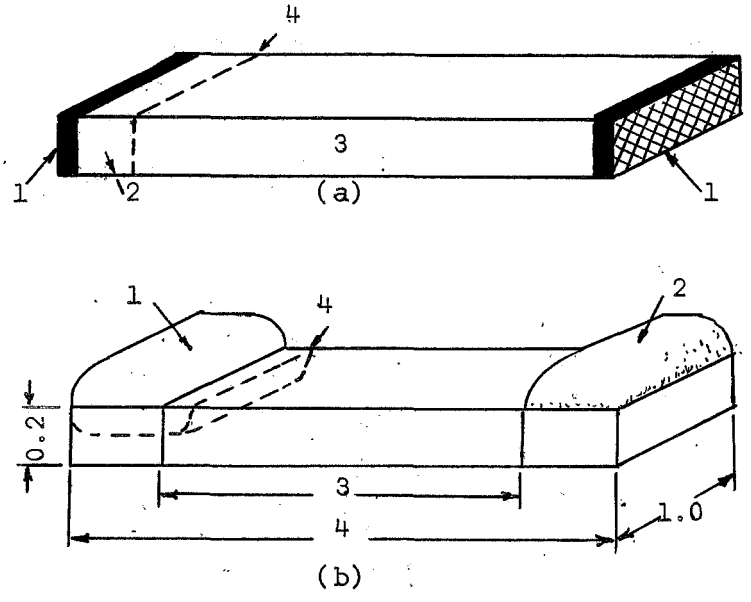


Figure 2.38. General view of a magnetodiode. a. Ideal magnetodiode. 1) metal; 2) p-type semiconductor; 3) n-type semiconductor; 4) p-n junction. b. Test specimen. 1) indium; 2) tin; 3) n-type Ge; 4) p-n junction.

conductivity modulation. This again reduces the junction voltage and the process continues until a new steady state condition is reached. Thus a small change in diffusion length can lead to a large change in terminal I-V characteristics.

The I-V characteristics of the wide base, high resistivity diode have been analyzed in detail by Stafeev [57], and at high injection

$$I \simeq I_c \exp\left(\frac{qV}{CkT}\right), \tag{2.136}$$

where I_c is a complex function of the diode parameters,

$$C = 2\ \frac{b + \cosh(d/L)}{b + 1}, \tag{2.137}$$

$$b = \mu_n/\mu_p,$$

and d is the length of the high resistivity region. The change in current due to a change in diffusion length has also been evaluated as

$$\frac{1}{I}\ \frac{\Delta I}{\Delta B} = \frac{d}{L}\left(\frac{qV}{CkT} + \frac{L}{d} - 1\right)\ \frac{1}{L}\ \frac{\Delta L}{\Delta B}. \tag{2.138}$$

For large values of $\frac{d}{L}$, the sensitivity is essentially linearly proportional to the length of the high resistivity region. The factor multiplying $\frac{1}{L}\ \frac{\Delta L}{\Delta B}$ can easily be 10 or larger so the current sensitivity of such a diode can be a factor of 10 or more larger than the diffusion length sensitivity.

The I-V characteristics of a diode in varying magnetic fields are shown in Figure 2.39. At an applied voltage of about 3 volts, the current is seen to change by about an order of magnitude for a magnetic field of 10 K gauss, which is about a factor of 10 more sensitive than the diffusion length sensitivity. As can also be seen from the figure, at a constant current the voltage changes by a farily large amount with magnetic field. In equation form this is

$$\frac{1}{V}\frac{\Delta V}{\Delta B} = \frac{d}{L}\left(\frac{qV}{CkT} + \frac{L}{d} - 1\right)\frac{CkT}{qV}\frac{1}{L}\frac{\Delta L}{\Delta B}. \qquad (2.139)$$

Typical experimental values of voltage sensitivity are 100 μv/gauss to 300 μv/gauss with values as high as 800 μv/gauss observed [53] [56].

In summary then it can be stated that significant changes in p-n junction characteristics have been observed with magnetic field. Magnetic field sensitive diodes require a special structure which makes use of conductivity modulation as well as normal minority carrier injection. The phenomena responsible for the effect appears to be a reduction of minority carrier diffusion length with magnetic field. The observed effects are linear in magnetic field as opposed to magnetoresistance which varies as the square of the magnetic field. Based upon the experimental data, it appears that the useful magnetic field range for these diodes is from a lower limit of around 1 gauss to fields aroun 10 K gauss.

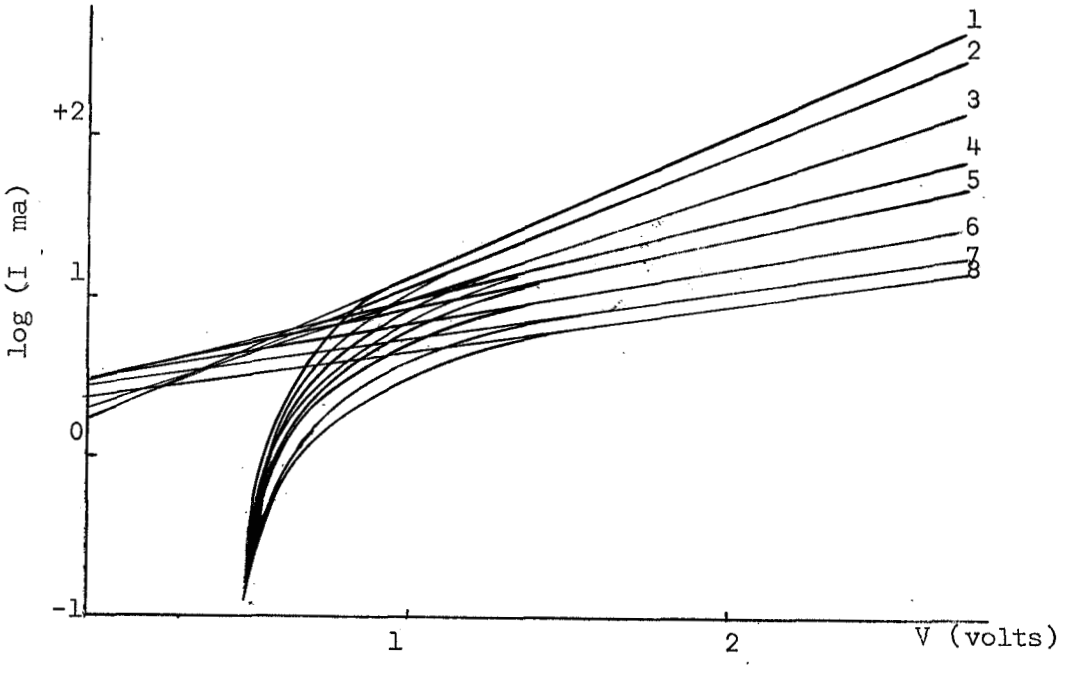


Figure 2.39. Forward branches of I-V characteristics of magnetodiode in different magnetic fields. 1) B=0, 2) B=2.5 kgauss 3) B=5.3 kgauss, 4) B=7.8 kgauss, 5) B=10.1 kgauss, 6) B=12.6 kgauss, 7) B=19.5 kgauss, 8) B=23 kgauss.

2.2. Semiconductor Heterojunctions

Semiconductor heterojunctions are formed by joining two dissimilar semiconductors such as Ge and Si. The heterojunction is thus more general than the p-n junction and in a certain sense the normal p-n junction can be considered as a special class of heterojunctions. In this discussion the term heterojunction will be used to refer only to those devices composed of two different semiconductor materials. The two semiconductors forming a heterojunction in general have different energy band gaps, different work functions or electron affinities, different lattice constants, different thermal expansion coefficients, different dielectric constants and many other lesser important differences.

The ability to form a junction from semiconductors with different energy band gaps has some important advantages for special types of solid state devices. Heterojunctions were originally thought to offer very good possibilities for emitter-base junctions of transistors because in principle such an emitter can have a very large beta. Because of practical difficulties however in fabricating such junctions and additional complications of interface states this promise has not been fulfilled. The most interesting sensor application of heterojunctions appears to be in the optical sensor area and in the stress sensor area. For optical sensors, detectors can be constructed which have essentially a flat response over some band of wavelengths and falls off rapidly at either end of the band. The center wavelength of the detection band as well as the band width can in principle be adjusted by suitable choices of materials. Heterojunctions have been shown to be quite sensitive to stress

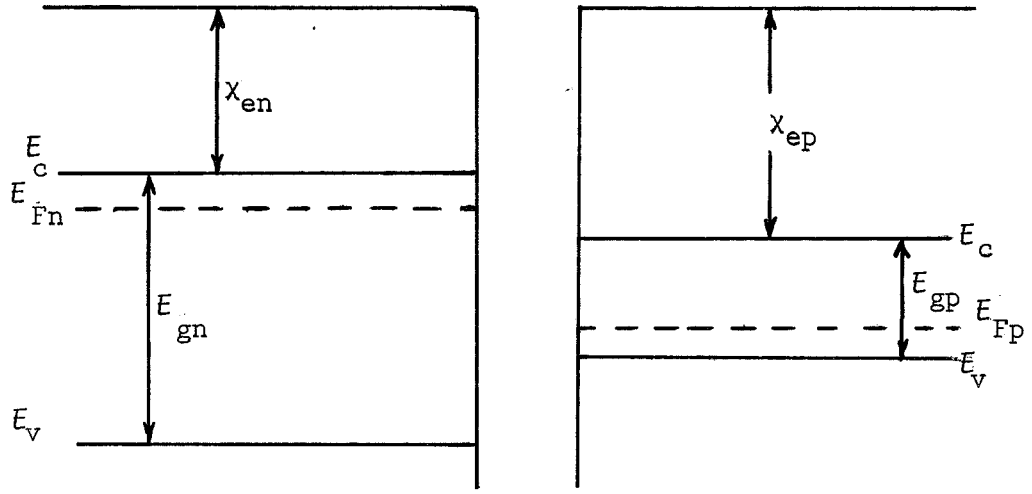
when one of the materials is peizoelectric and the other is not. This introduces an additional physical mechanism by which the potential across a p-n junction can be effected by stress. Such piezoelectric-heterojunction diodes have shown considerable promise for stress sensors.

The objective of the present section is to review the basic physical theory of heterojunction diodes and to discuss possible sensor application of these devices.

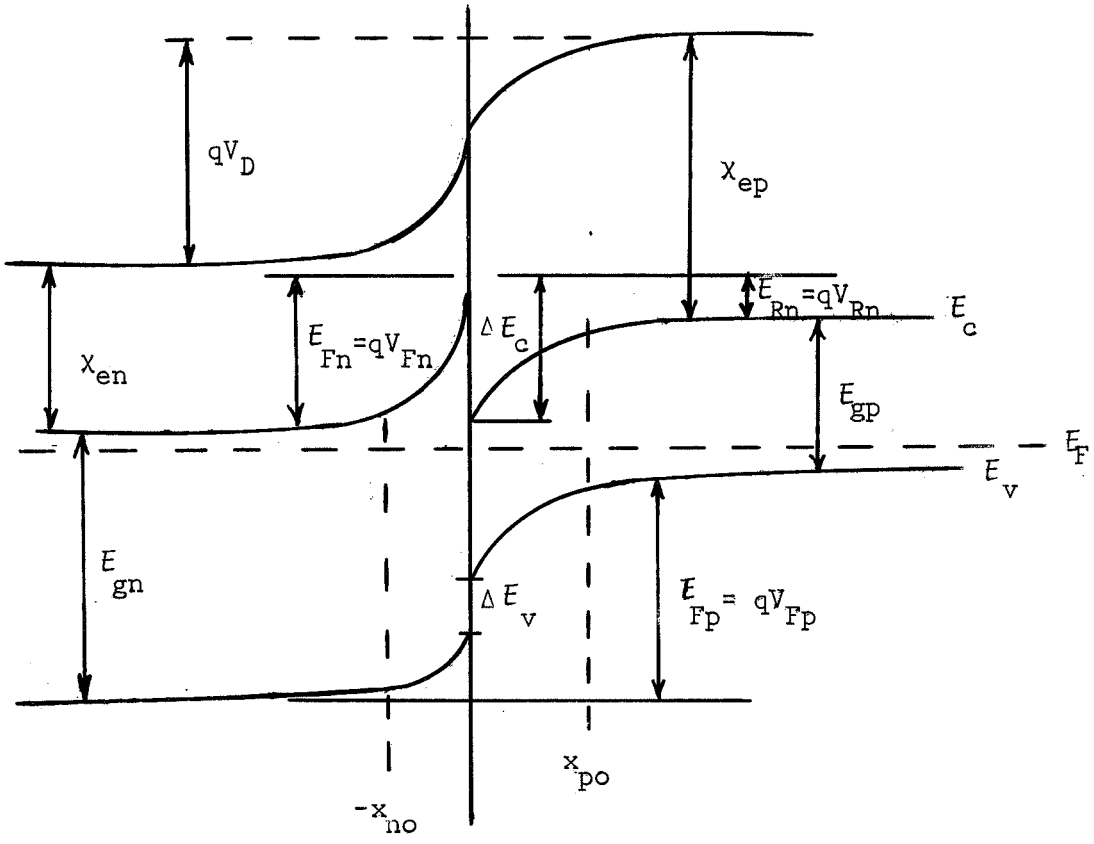
2.2.1. Basic Theory

The theoretical analysis of heterojunctions is considerably more complicated than that of the simple p-n junction. This arises from several facts. First one has twice as many fundamental parameters because of the two different materials. There are at least two distinct cases of an ideal p-n heterojunction as well as different cases of n-n and p-p heterojunctions. All of these cases will not be covered here in detail but rather the more important cases will be considered. In addition to these fundamental material differences, the problems of lattice mismatch and interface states make a detailed theoretical analysis extremely difficult.

The theoretical models which have been most extensively studied for the heterojunction and which have been most successful in describing the observed properties are based on an abrupt energy band picture. This model assumes that there is an abrupt transition at the hetero-junction interface from one material to the other. The energy band diagram of an abrupt n-p heterojunction is shown in Figure 2.40 [57] [58]. For illustration purposes the band gap of the n-type material



(a) Before contact



(b) After contact

Figure 2.40. Energy band diagram of n-p abrupt heterojunction with a positive reverse barrier [57][58].

is assumed to be larger than that of the p-type material. If the Fermi levels of the two materials differ before contact, a space charge region is formed at the interface after contact with the material with the higher initial Fermi level (n-type material) becoming positively charged and the other material (p-type) becoming negatively charged. The total electrostatic potential across the space charge region is V_D which is equal to the potential difference between the initial Fermi levels in the two materials as in a normal p-n junction. The major difference in the enrgy band picture for the heterojunction and that of a normal diode is the discontinuities ΔE_c and ΔE_v in the conduction and valence bands at the interface.

Consider now the barriers to carrier flow across the interface of the heterojunction in Figure 2.40 . For holes there is a barrier for flow from the p-type side to the n-type side of E_{FP} which is very similar to the barrier at a normal p-n junction. Hole flow from the n-side to the p-side encounters no potential barrier. For electrons however there is a potential barrier for flow in either direction. For flow from the n-side to p-side the barrier is E_{Fn} and for flow in the opposite direction the barrier is E_{Rn}. This is quite different from that of a normal p-n junction. The height of the barrier depends on the relative discontinuities in the bands at the interface as well as the energy band gap differences.

In addition to the band picture illustrated in Figure 2.40, two other situations can arise as shown in Figure 2.41. In case (a) the height of the reverse barrier is negative and electron flow across the barrier is controlled essentially by the hieght of the forward barrier E_n. In case (b) the energy bands are such that there is no

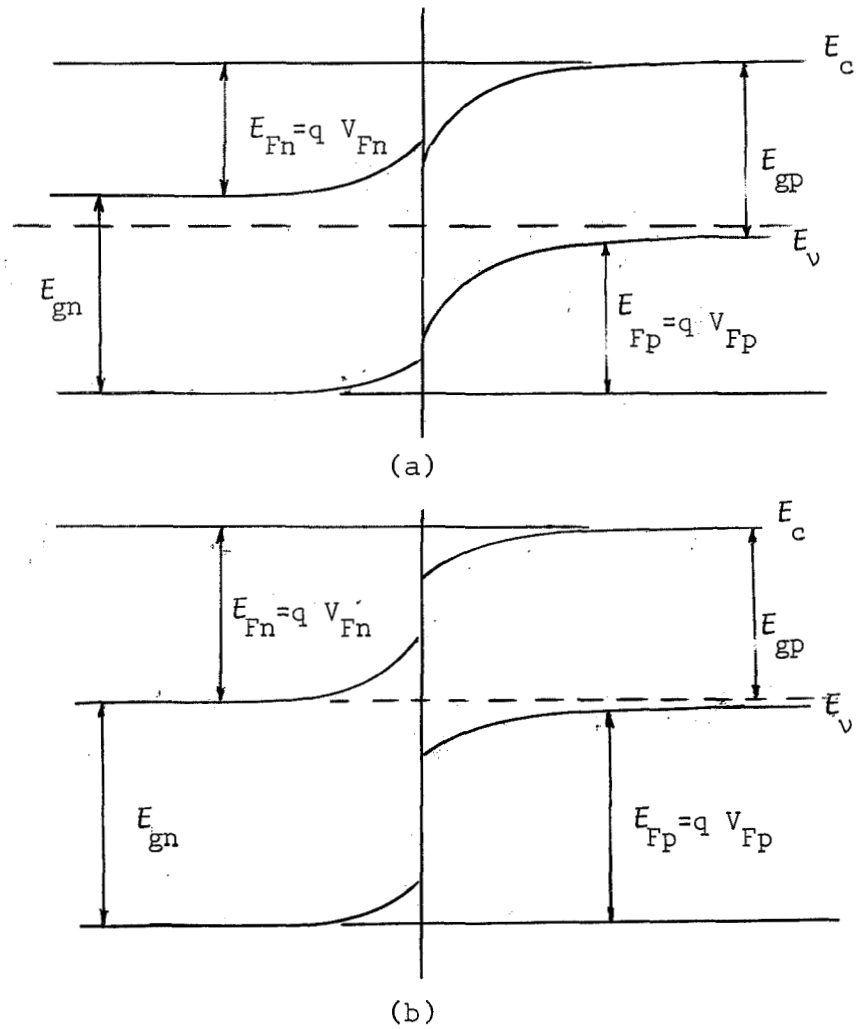


Figure 2.41. Energy band diagrams of n-p heterojunctions with (a) negative reverse barrier and (b) no reverse barrier.

reverse barrier to electron flow. As far as the conduction band is concerned in a p-n heterojunction there are three different cases shown in Figures 2.40 and 2.41 which will be referred to as (1) the positive reverse barrier case (2) the negative reverse barrier case and (3) the zero reverse barrier case.

In each case of the n-p heterojunction depicted in Figures 2.40 and 2.41 , the valence band is similar and exhibits no reverse barrier to hole flow. Just as for the conduction band three cases

of (1) a positive reverse barrier, (2) a negative reverse barrier and (3) zero reverse barrier can occur for the valence band depending on the energy bands before formation of the junction.

The barrier which is most important in determining the current flow depends on the doping of the junction as well as the energy gaps of the two materials. If the wide band gap material is n-type as shown in Figures 2.40 and 2.41 , the electron barrier is the controlling barrier. The hole barrier is considerably larger and under most circumstances the hole current would be negligable. On the other hand if the wide band gap material is p-type, the controlling barrier becomes the hole barrier.

In additon to p-n heterojunctions, n-n and p-p heterojunctions can be formed. A typical n-n heterojunction is shown in Figure 2.42 . The total voltage drop across the interface space charge region is $V_{B1} + V_{B2}$. This arises from the formation of a depletion region in one material and an accumulation region in the other material. As seen in the figure there is a barrier for current flow in both directions with the magnitude of the barrier being qV_{B1} in one direction and $\Delta E_c - qV_{B2}$ in the opposite direction. By comparing Figure 2.42 with Figure 2.40 , it is seen that the basic shape of the conduction energy bands is the same in both cases. The major difference is the location of the Fermi level with respect to the conduction band of the right hand material in each case.

The above discussions have been in terms of ideal interfaces. In practical heterojunctions there are at least two factors which can

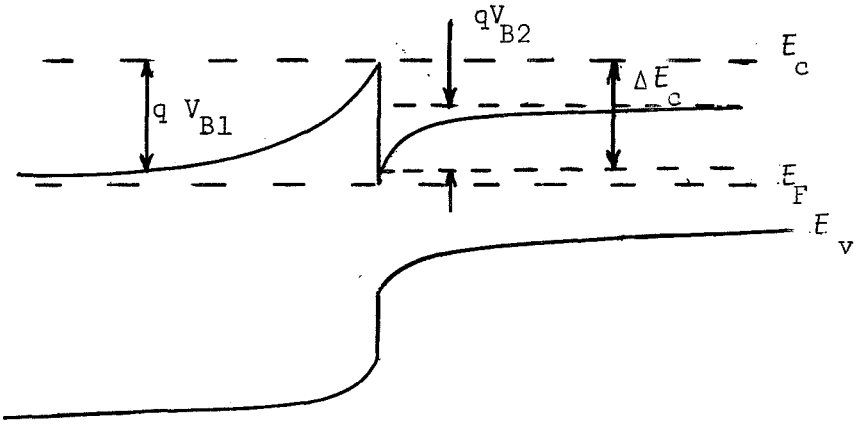


Figure 2.42. Illustration of energy band diagram of n-n heterojunction.

modify the ideal energy band diagrams. First there is never an exactly abrupt transition from one material to another. Most devices are made at relatively high temperatures and some interdiffusion of the materials results. This must give rise to some grading of the material properties such as the band gaps near the interface. In the extreme case of considerable interdiffusion, the band gap would be expected to exhibit a gradual transition from one material to the other and the abrupt transition would be lost. The expected energy band diagram for a graded band gap situation is shown in Figure 2.43 . In the graded band gap picture the energy barriers are similar to those of a normal p-n junction except that the electron barrier E_{Fn} differs from the hole barrier E_{FP}. Even when the amount of band gap grading is small it can still have a large influence on the energy barrier at the interface. Consider for example the energy band diagram of Figure 2.40 . Any small amount of interdiffusion of the two materials will directly effect

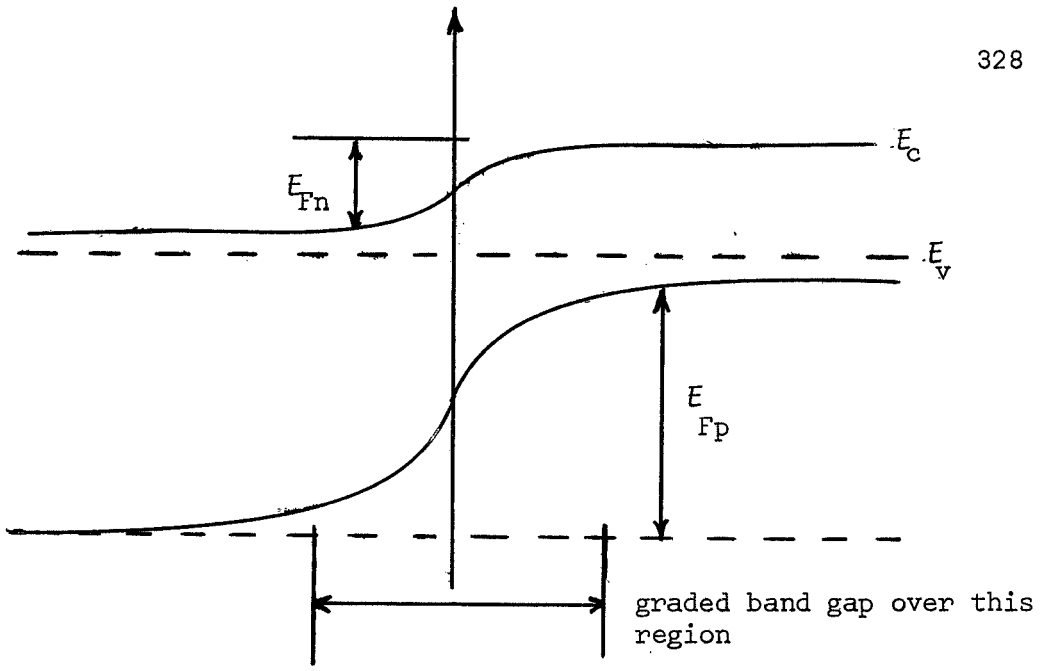


Figure 2.43. Expected energy band diagram for a graded band gap p-n heterojunction.

the band gaps at the interface. This in turn changes the energy band discontinuities ΔE_c and ΔE_v and directly changes both the forward and reverse energy barrier. It is thus impossible to theoretically predict the exact value of the barrier discontinuity and the value can change considerably from device to device.

In addition to the interdiffusion of the two materials, the lattice constants of the two materials in general are not identical and there is some lattice stress and distortion at the interface. In general this leads to additional energy states at the interface located within the energy band gap. This will trap charge causing a discontinuity in the electric field at the interface as well as act as recombination centers.

The evaluation of the properties of the space charge region of the heterojunction is not as straightforward as for the normal p-n junction.

Consider a p-n heterojunction such as shown in Figure 2.40 . The Ge-GaAs heterojunction is believed to be of this type. If the discontinuity ΔE_c is not too great for this type of diode, a simple depletion region is formed on each side of the junction. If ΔE_c is too large, the conduction band in the narrow band gap material will approach closely to the Fermi level and an electron inversion layer forms near the interface. The existance of an inversion layer greatly complicates the analysis and this case will not be considered here. Consideration is given only to the case where a depletion layer exists on each side of the interface. In this case the potential is discribed by Poisson's equation as

$$\frac{d^2V}{dx^2} = \begin{cases} \frac{-q}{\varepsilon_n} N_D & -x_n \leq x < o \\ \frac{q}{\varepsilon_p} N_A & 0 < x < x_p, \end{cases} \tag{2.140}$$

where N_D and N_A are the doping densities on the n- and p- side of the junction. This is solved subject to the boundary conditions on the field of

$$E = \begin{cases} E_n & x = 0^- \\ 0 & x = -x_n \end{cases} \tag{2.141}$$

$$E = \begin{cases} E_p & x = 0^+ \\ 0 & x = x_p \end{cases} \tag{2.142}$$

where E_n and E_p are the values of the electric fields on either side of the interface. For a normal p-n junction $E_n = E_p$. For a heterojunction these can differ however because of the presence of interface

states. In general these are related by the expression

$$\varepsilon_p E_p = \varepsilon_n E_n + Q_I, \tag{2.143}$$

Where Q_I is the interface charge density per unit area.

The solution of Equation (2.140) subject to the boundary condition on electric field gives

$$E_n = (2q N_D/\varepsilon_n)^{1/2} V_n^{1/2},$$

$$E_p = (2q N_D/\varepsilon_p)^{1/2} V_p^{1/2}, \tag{2.144}$$

where V_n and V_p are the voltage drops in the n- and p- type regions. The sum of V_n and V_p must equal the equilibrium junction potential V_o minus the applied voltage V_a. This leads to two equations describing the depletion region of

$$(V_n/V_o) + (V_p/V_o) = (1 - V_a/V_o),$$

and (2.145)

$$(V_n/V_o)^{1/2} + Q_I/(2q N_D \varepsilon_n V_o)^{1/2} = (N_A \varepsilon_p/N_D \varepsilon_n)^{1/2} (V_p/V_o)^{1/2}$$

The solution of these two equations gives V_n and V_p as a function of V_a the applied voltage. These can be calculated in general but the solution is quite complex. The following graphical technique illustrates the important feature of the solution. If the two equations are plotted

as functions of $(V_n/V_o)^{1/2}$ and $(V_p/V_o)^{1/2}$, Equation (2.144) describes a circle and Equation (2.145) describes a straight line. This is shown in Figure 2.44 where the intersection of the two curves gives the correct solution. From this figure the effect of changing various device parameters on the space charge regions can be seen. The only simple case occurs when $Q_I = 0$ and $N_A \varepsilon_p = N_D \varepsilon_n$. In this case, $V_n = V_p = (V_o - V_a)^{1/2}$ and half of the depletion region voltage occurs on each side of the interface. In the slightly more general case where just $Q_I = 0$, one finds

$$V_n = [N_A \varepsilon_p/(N_A \varepsilon_p + N_D \varepsilon_n)] (V_o - V_a) \qquad (2.146)$$

$$= K_n (V_o - V_a),$$

$$V_p = [N_A \varepsilon_n/(N_A \varepsilon_p + N_D \varepsilon_n)] (V_o - V_a) \qquad (2.147)$$

$$= K_p (V_o - V_a),$$

where K_n and K_p are the fractions of the total voltage appearing on each side of the interface.

The above discussion in connection with Q_I has been in terms of a constant surface charge at the interface. In general the interface charge depends on the position of the trap energy with respect to the Fermi level. This means that the value of Q_I must in general be considered as a function of the interface potential or $Q_I = Q_I(V_n)$. This case requires a detailed knowledge of the interface states and goes

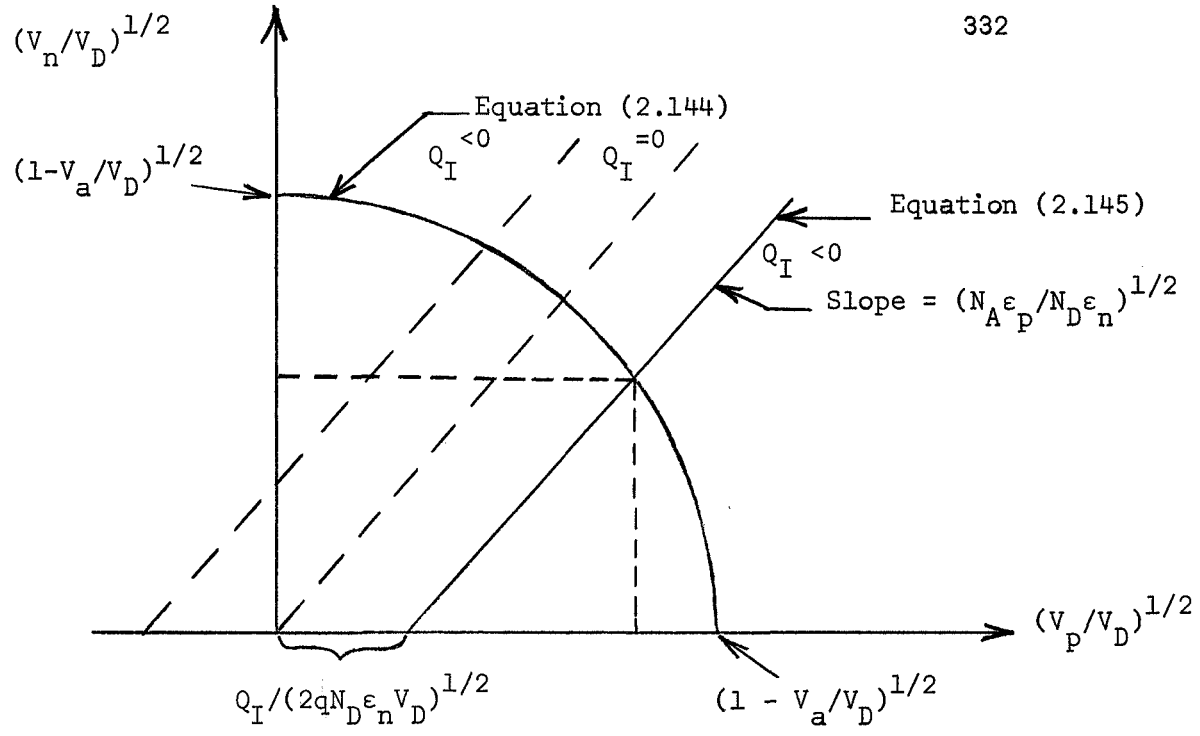


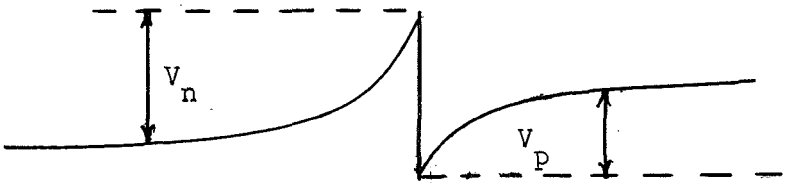
Figure 2.44. Graphical solution of voltage equations for heterojunction.

beyond the purpose of this discussion. For the purpose of this work Q_I will be considered as a constant.

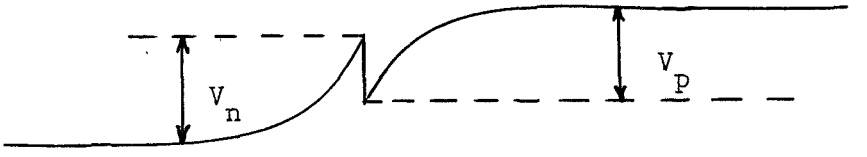
In addition to the properties of the space charge region the J-V characteristics are of interest. The three types of electron barriers which can appear in heterojunctions are shown in Figure 2.45. A p-n junction can have either of these types of barriers while an n-n heterojunction typically has the positive reverse barrier. Assuming that the current is barrier controlled in each of these three cases, the current density can be expressed as [57]-[59]

(1) Positive reverse barrier

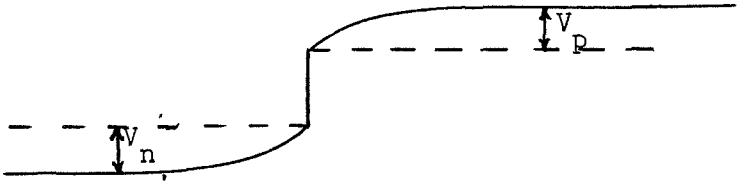
$$J_n = J_o \exp[-q (V_n + V_a)/kT] [\exp (qV_a/kT) -1], \qquad (2.148)$$



(a) Positive reverse barrier



(b) Negative reverse barrier



(c) Normal barrier

Figure 2.45. Illustration of three types of electron barriers in a heterojunction.

(2) Negative reverse barrier

$$J_n = J_o \exp(-qV_o/kT)\,[\exp(qV_a/kT) - 1], \qquad (2.149)$$

$$J_n = J_{on}\,[\exp(qV_a/kT) - 1],$$

(3) Normal barrier

$$J_n = J_o \exp(-qV_o/kT)\,[\exp(qV_A/kT) - 1], \qquad (2.150)$$

$$J_n = J_{on} [\exp (qV_a/kT) - 1],$$

where in each case V_o is the total equalibrium barrier height and J_o is a constant depending on the material parameters, but which does not depend exponentially on any of the barrier heights. The expressions for current for the negative reverse barrier and the normal barrier cases depend on total barrier heigh and on applied voltage in the same manner as for normal p-n junctions. The positive reverse barrier case is the most interesting in terms of unique J-V characteristics. In the forward bias direction

$$J_{nf} \simeq J_o \exp (-q V_n/kT). \qquad (2.151)$$

In the case of no interface states this becomes

$$J_{nf} \simeq J_o \exp (-K_n V_o/kT) \exp (qK_n V_a/kT), \qquad (2.152)$$

which increases exponentially with applied voltage but with a slope of $q K_n/kT$. In the reverse bias direction

$$J_{nr} \simeq - J_o \exp [-q (V_n + V_a) /kT] \qquad (2.153)$$

$$\simeq - J_o \exp (-K_n V_o/kT) \exp (-qK_p V_a/kT), \qquad (2.154)$$

where the second expression results when the surface charge density is negligible. Remembering that V_a is negative under reverse bias,

it is seen that the current depends exponentially on applied voltage in both directions.

The positive reverse barrier case with an n-n heterojunction is slightly different from the p-n junction because of the presence of the accumulation space charge region. Sze has shown that for this case the current density expression becomes [9]

$$J_n = J_{no} \left(1 - \frac{V_a}{V_o}\right) \left[\exp \left(\frac{q\, V_a}{kT}\right) - 1 \right]. \qquad (2.155)$$

Under reverse bias this gives a linear J-V characteristic.

The valence bands of semiconductor heterojunctions exhibit the same three types of barriers and similar expression can be written for hole currents. The expressions can be obtained from Equations (2.148) - (2.150) by replacing J_n with J_p, and by interpreting the energy barriers as the valence band energy barriers.

In general the total current at an interface is the sum of the electron and hole current injected across the space charge regions plus any recombination within the space charge region. The injected currents have been discussed above so now consider the space charge generation-recombination current given in general by

$$J_{g\text{-}r} = q \int_{-x_n}^{x_p} U \, dx, \qquad (2.156)$$

$$= q \int_{-x_n}^{x_p} \frac{(pn - p_o n_o)\, N_t}{(n_1 + n)/C_p + (p_1 + p)/C_n} \, dx, \qquad (2.157)$$

where the Shockley-Read theory has been used to describe the recombination centers. For a heterojunction the most important recombination states are likely to be located near the interface so we can write $N_t = N_{ts}\,\delta(x)$, where N_{ts} is the density of recombination centers per unit area at the interface. With this approximation and the fact that $pn = p_o\, n_o \exp(q\, V_a/kT)$ within the space charge region

$$J_{g\text{-}r} \simeq q\, N_{ts}\, \frac{p_o\, n_o\, [\exp(qV_a/kT) - 1]}{[n_1 + n(0)]/C_p + [p_1 + p(0)]/C_n}, \quad (2.158)$$

where n (0) and p (0) are the values of n and p at the interface. If the equilibrium Fermi level is several kT from the recombination level the denominator can be simplified and the above becomes

$$J_{g\text{-}r} \simeq q\, N_{ts}\, \frac{p_o\, n_o\, [\exp(qV_a/kT) - 1]}{n(0)/C_p + p(0)/C_n}. \quad (2.159)$$

The development can not be carried further without a specific model which can be used to evaluate n (0) and p (0). As a specific example consider the positive reverse barrier p-n junction as shown in Figure 2.40. For this case n (0) and p (0) are determined by the narrow band gap material and

$$n(0) = \frac{p_o\, n_o}{N_A} \exp(q\, V_{po}/kT) \exp(q\, K_n\, V_a/kT),$$

$$p(0) = N_A \exp(-q\, V_{po}/kT) \exp(q\, K_p\, V_a/kT), \quad (2.160)$$

where V_{po} is the equilibrium value of depletion region voltage across the p-type region. It is recalled that K_n and K_p are the fractions

of the applied voltage appearing across the n- and p-type regions. Two slightly different voltage dependences can be observed in Equation (2.159) depending on whether $n(0)/C_p$ or $p(0)/C_n$ is the larger quantity. These two cases give

Case 1. $n(0)/C_p \gg p(0)/C_n$

$$J_{g-r} \simeq q N_{ts} C_p N_A \exp(-q V_{po}/kT) \exp(-q K_n V_a/kT) \times [\exp(q V_a/kT) - 1] \quad (2.161)$$

Case 2. $n(0)/C_p \ll p(0)/C_n$

$$J_{g-r} \simeq q N_{ts} C_n \frac{p_o n_o}{N_A} \exp(q V_{po}/kT) \exp(-q K_p V_a/kT) \times [\exp(q V_a/kT) - 1]. \quad (2.162)$$

At large forward bias these become

Case 1. $$J_{g-r} \simeq J_{ro} \exp(-q V_{po}/kT) \exp(q K_p V_a/kT), \quad (2.163)$$

Case 2. $$J_{g-r} \simeq J_{ro} \exp(-q V_{no}/kT) \exp(q K_n V_a/kT), \quad (2.164)$$

where J_{ro} may differ for the two cases but does not depend on any of the energy barriers. An examination of the other two cases of a

negative reverse barrier and no reverse barrier shows that each of these also leads to a generation-recombination current of the form of either Equation (2.163) or (2.164). The value of J_{ro} for each type of junction depends on the details of the junction and the recombination levels. These equations show that the generation-recombination current for all three cases depends on the individual barrier heights i.e. on V_{no} or V_{po}. The injected currents on the other hand for the negative reverse barrier and for the normal barrier were found to depend only on the total energy band gap. It is interesting to note that if the heterojunction is approximately symmetrical so that $K_n = K_p = 1/2$, the generation-recombination current varies as exp $(q V_a/2kT)$. This is identical to the corresponding voltage dependence of generation-recombination current in normal p-n junctions.

For wide band gap semiconductor heterojunctions such as occur with the II-VI materials or many III-V materials, the generation-recombination current is likely to be the dominant current. Generation-recombination current is a real problem with GaAs p-n junctions and with other wide band gap materials and certainly one expects more recombination center with heterojunctions than with normal p-n junctions.

2.2.2. Optical Sensors

Semiconductor heterojunctions offer some very interesting possibilities for optical sensor applications. Consider the p-n heterojunction shown in Figure 2.46 where the wide bandgap semiconductor is located near the surface. The wide bandgap material may be either n- or p- type but p-type is shown for illustrative purposes. At photon energies below the

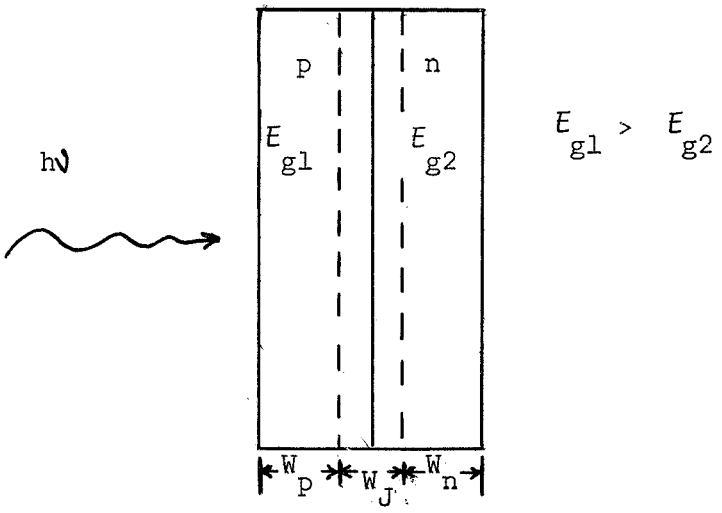


Figure 2.46. Heterojunction optical sensor with semiconductors of two different band gaps.

bandgap energy of either material, there is little absorption in either material and little photoresponse. As the photon energy approaches the bandgap energy of the n-type material, absorption will begin to occur within the n-type material. Carriers created within the space charge region and to a depth of one diffusion length on the n-type side of the junction are then collected by the junction and consitute the photocurrent. The photocurrent can then be written as

$$J_{photo} = q\,\alpha_n\,\bar{L}\,N, \tag{2.165}$$

where α_n is the absroption coefficient in the n-type material, N is the photon density per unit time reaching the junction interface and $\bar{L}$ is the effective depth over which carrier collection occurs. This is similar to the absorption of a normal p-n junction with an extremely thin p-type region. In terms of the material parameters (see Section 2.1.2 for a more complete discussion)

$$\overline{\alpha L} = \alpha_n F_p + (1 - e^{-\alpha_n x_n}), \qquad (2.166)$$

where x_n is the width of the depletion region on the n side of the interface and

$$F_p = \frac{\coth(\frac{W_n}{L_p}) - e^{-\alpha_n W_n} \operatorname{csch}(\frac{W_n}{L_p}) - \alpha_n L_p}{(1 - \alpha_n^2 L_p^2)} e^{-\alpha x_n}. \qquad (2.167)$$

When $\alpha_n (x_n + L_p) >> 1$, essentially all carriers created within the n-type material are collected by the junction and $\overline{\alpha_n L} \to 1$. For photon energies below E_{gp} the operation can then be summarized as follows. The photocurrent increases rapidly for energies above E_{gn} and approaches a constant response when essentially all the photons are absorbed either within the depletion region or within one diffusion length of the depletion region.

When the photon energy becomes larger than E_{gp} absorption occurs within the surface p-type region. This reduces the absorption within the n-type region. If the p-type region is very thin compared with a diffusion length the response will remain constant or decrease slightly due to surface recombination. On the other hand if the width of the p-type region is large compared with a diffusion length the response will fall off rapidly due to the increased absorption within the p-type region. For energies above E_{gp} the response is given by Equation (2.165) with

$$\overline{\alpha L} = \alpha_p F_n + \alpha_n F_p + (1 - e^{-\alpha x_p - \alpha x_n}), \qquad (2.168)$$

where

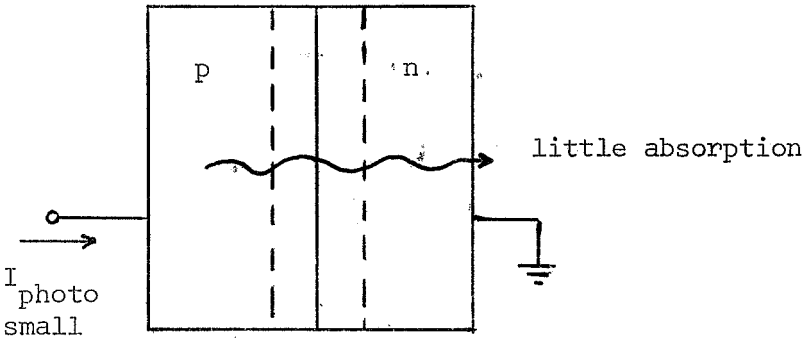
$$F_p = \frac{\coth\left(\frac{W_n}{L_p}\right) - e^{-\alpha_n W_n} \operatorname{csch}\left(\frac{W_n}{L_p}\right) - \alpha_n L_p}{(1 - \alpha_n^2 L_p^2)} e^{-\alpha_n x_n - \alpha_p (x_p + W_p)}, \tag{2.169}$$

$$F_n = \frac{e^{-\alpha_p W_p} \left[\sinh\left(\frac{W_p}{L_n}\right) + \frac{SL_n}{D_n} \cosh\left(\frac{W_p}{L_n}\right)\right] - \alpha_p L_n \left(1 + \frac{S}{\alpha_p D_n}\right)}{(1 - \alpha_p^2 L_n^2)\left[\cosh\left(\frac{W_p}{L_n}\right) + \frac{SL_n}{D_n} \sinh\left(\frac{W_p}{L_n}\right)\right]}$$

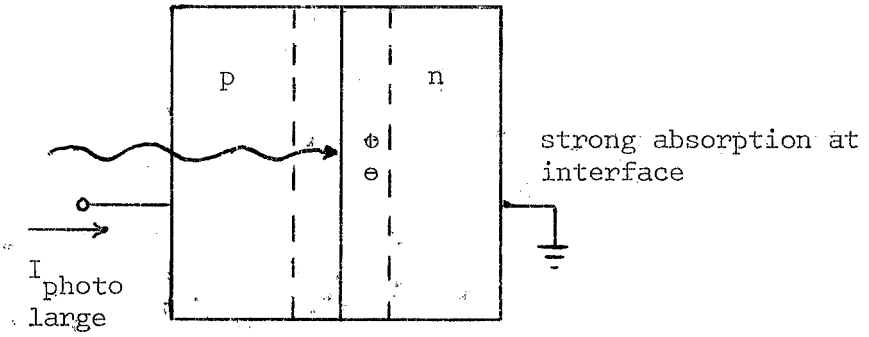
$$+ \frac{\alpha_p L_n}{(1 - \alpha_p^2 L_n^2)} e^{-\alpha_p W_p}. \tag{2.170}$$

As in the case of the normal p-n junction these are fairly complex expressions.

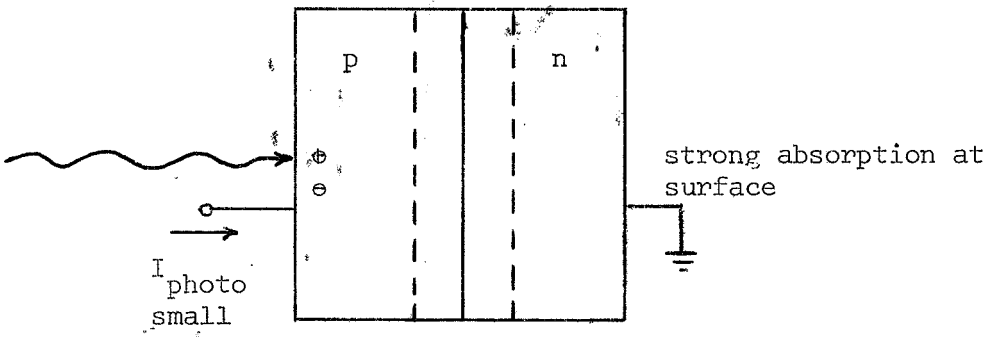
The most unique heterojunction optical sensor is obtained when the width of the surface region (p-region in Figure 2.46) is wide compared with a diffusion length. Photon energies above E_{gp} are strongly absorbed near the surface and are not able to contribute to the photoresponse. Such a heterojunction has a band of wavelengths corresponding to the region of energies between the two bandgaps to which it is sensitive and the response falls off rapidly on either side of this region. The response in the sensitive region can be fairly constant because the carriers are created very near the interface within the space charge region where the collection efficiency is very high. Conditions existing within the three photon energy regions are illustrated in Figure 2.47. The measured photoresponse of a Ge-



a. Photon energy $< E_{gn}$



b. Photon energy $> E_{gn}$ but $< E_{gp}$



c. Photon energy $> E_{gp}$

Figure 2.47. Illustration of heterojunction response to different wavelengths for $E_{gp} > E_{gn}$.

GaAs heterojunction as a function of photon energy is shown in Figure 2.48 [60]. This response was obtained by having the light incident on the wide band gap GaAs material. The energy band between the bandgaps of the two semiconductors is from 0.63 ev of Ge to 1.4 ev of GaAs. The slow increase in the response from about 0.6 ev to about 0.8 ev is due to the slowly increasing absorption coefficient of Ge within this region of energy. A direct band gap material with a more rapidly increasing absorption coefficient would show a sharper increase with energy. For the heterojunction response shown in Figure 2.48 the band of photon energies for significant response extends from about 0.85 ev to about 1.45 ev. Shown in Figure 2.49 is a second example of the photoresponse of a GaAs -GaP heterojunction detector [61]. Again the almost constant response is seen in the energy region between the band gaps of the two semiconductor materials.

The heterojunction detector could be used with the narrow band gap material toward the incident light. Such a detector however would have a smaller response than a normal p-n junction of the narrow band gap semiconductor and consequently appears to be of little practical use.

For a narrow band detector, the width of the surface semiconductor layer should be large compared with a diffusion length or it should have a large surface recombination velocity so few carriers which are absorbed near the surface can reach the junction depletion retion. On the other hand if one wishes to use a heterojunction as a solar energy converter, a thin surface layer is needed so that all available photons can contribute to the current. Heterojunctions have been shown to

have high theoretical efficiencies as solar cells. Theoretical efficiencies of 30% are predicted for InP - GaAs and InP-GaP heterojunction as compared with a theoretical efficiency of 24% for silicon p-n junctions [62]. These large efficiencies however have yet to be demonstrated experimentally.

The most extensively studied heterojunction photodetector appears to be the CdS solar cell. This is believed to be a heterojunction formed by n-type CdS and Cu_2S dorming a very thin p-type layer on the surface of the CdS. The CdS has a band gap of 2.4 ev while the Cu_2S is believed to have a band gap near 1.0 ev [63],[64] . As the CdS solar cell is normally operated, the light is incident on the narrow band gap material as opposed to the situation illustrated in Figure 2.47 and discussed in detail above. In this sense the CdS heterojunction solar cell does not

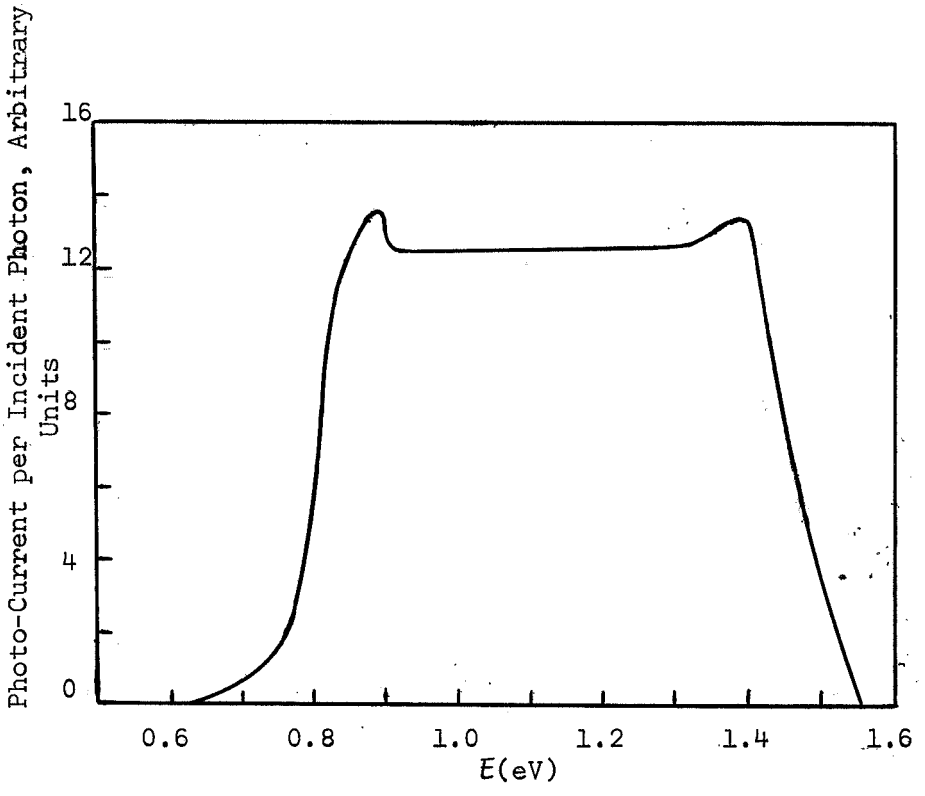


Figure 2.48. Short-circuit photocurrent of a GaAs-Ge p-n heterojunction per incident photon vs. photon energy [60].

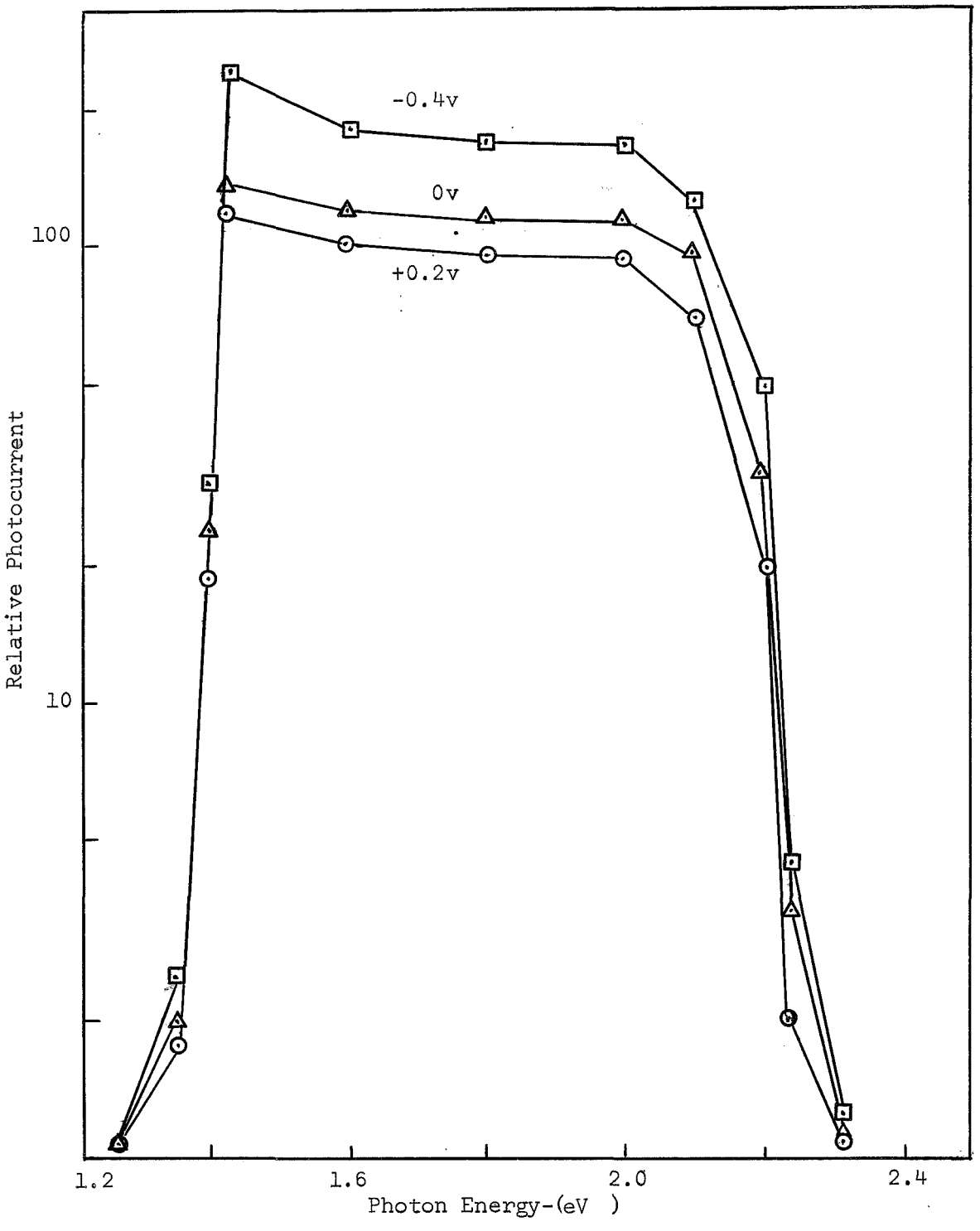


Figure 2.49. Relative photoresponse current spectrum for device 42 GaAs-GaP [61]

operate in the most efficient manner. Efficiencies of as much as 8% have been achieved in CdS heterojunction solar cells [63].

The terminal J-V characteristic of a heterojunction when generation-recombination currents are negligible is given by the equation

$$J = J_o \exp[-q(V_i + V_a)/kT] [\exp(q V_a/kT) -1] - q\alpha\overline{LN}, \tag{2.171}$$

for the positive reverse barrier case and

$$J = J_o [\exp(q V_a/kT) -1] - q\alpha\overline{LN}, \tag{2.172}$$

for the negative reverse barrier and normal barrier cases, where V_i is either V_n or V_p depending on whether the controlling barrier is the conduction band or valence band barrier. The above expressions are very similar to the corresponding equations for normal p-n junctions. The detectors may be employed in either a short circuit current mode where the current is directly proportional to the incident flux or in an open circuit voltage mode where the open circuit voltage is proportional to the log of the incident photon flux. If the generation-recombination current is dominant over the ideal current then the expression for generation-recombination must replace the injected current terms in Equations (2.171) and (2.172). Neglecting series resistance the short circuit current is just the photocurrent and doesn't depend on the current flow mechanism in the diode. For this reason this mode of operation is probably the most desirable mode in the heterojunction optical sensor.

In summary concerning the heterojunction optical sensors, it can be stated tha they offer unique sensor possibilites which are not available with normal p-n junctions. The most important of these unique features is the ability to form detectors which have a very nearly constant response over a band of photon energies and falls off rapidly on either side of the band. By controlling the composition of the semiconductors composing the heterojunction, the position of the band as well as the width of the band can be controlled in principle over the entire energy band spectrum of the available semiconductors. For the Group IV and III-V semiconductors this covers energies from about 0.16 ev of InSb to about 2.24 ev of GaP. The II-VI materials provide slight extensions to higher and lower energies.

2.2.3. Stress Sensors

Semiconductor heterojunctions have been found to be very sensitive to stress. The stress effects show up as shifts in the terminal I-V characteristics under either compressive or tensional stress. The changes appear to be linear at low stress levels.

Theoretically stress can effect the diode properties in several ways. First mobility or effective mass changes are expected to be present as in any simiconductor under stress. In terms of the current density equations for heterojunctions (Equations (2.148)-(1.150)) this would have an effect on the J_o terms. The experimentally observed effects are considerably larger than the piezoresistance effect so this does not appear to be the major stress sensitive parameter. A second stress effect arises from the stress induced energy band shifts.

Consider the expression for heterojunction current for either the negative reverse barrier or the normal barrier cases which is of the form

$$J = J_o \exp(-q\ V_o/kT)\ [\exp(qV/kT) - 1]. \qquad (2.173)$$

It is recalled that V_o is the total barrier height for dominant current flow and may be either the electron or hole barrier. Stress can change this total barrier height V_o by changing the energy band gap. This effect is exactly analogous to the stress effects occuring in normal p-n junctions as discussed in detail in Section 2.1.4. The stress dependence of V_o for the p-n heterojunctions shown in Figure 2.41 follows the stress dependence of energy band gap in the narrow band semiconductor forming the heterojunction. For a direct band gap material or for low stress levels, the change in band gap is essentially the hydrostatic deformation potential multiplied by the strain or

$$V_o \rightarrow V_o + \Xi_H\ \varepsilon, \qquad (2.174)$$

where Ξ_H is the hydrostatic deformation potential.
Then under stress

$$J = J_o \exp(-q\ V_o/kT)\ [\exp(q\ V/kT) - 1]\ \exp(q\ \Xi_H \varepsilon/kT). \qquad (2.175)$$

The current gauge factor is then

$$\frac{1}{I}\ \frac{dI}{d\varepsilon} = \frac{q\ \Xi_H}{kT}. \qquad (2.176)$$

Typical hydrostatic deformation potentials for most semiconductors are between 1 ev and 5 ev [40],[41],[65] and may have either a positve or negative sign depending on the semiconductor. This effect gives a current gauge factor of

$$\frac{1}{I}\frac{dI}{d\varepsilon} \sim 40 - 200. \qquad (2.177)$$

This energy band gap effect is comparable to the corresponding effect in normal p-n junctions and this effect does not use any additional unique features of the heterojunction. As with the normal p-n junction piezojunction effect, the effects at high stress are nonlinear and the multiple valley nature of the semiconductor must be considered. This will not be done in detail here because the bandgap changes in the heterojunction are similar to the p-n junction discussed in Section 2.1.4.

There is a final unique means by which stress can change the properties of heterojunctions which is not present in normal p-n junctions. This is through the peizoelectric effect [66] [67]. Assume that one of the materials composing the heterojunction is peizoelectric and the other is not piezoelectric. Within the piezoelectric material

$$D = \varepsilon_s E + P = \varepsilon_s E + e\,\varepsilon \qquad (2.178)$$

where P is the polarization ε_s is the semiconductor permittivity and e is the piezoelectric coefficient. Within the non-piezoelectric material there is no stress dependent polarization. Thus the piezo-

electric effect introduces a stress dependent discontinuity in the electric field at the heterojunction interface. If both materials are piezoelectric but with different piezoelectric coefficients there will still be a net discontinuity at the interface which can be described for a p-n heterojunction by the equation

$$\varepsilon_p E_p + e_p \varepsilon = \varepsilon_n E_n + e_n \varepsilon \tag{2.179}$$

or

$$\varepsilon_p E_p = \varepsilon_n E_n + (e_n - e_p) \varepsilon, \tag{2.180}$$

where ε_n and ε_p are the permittivities of the n- and p-type materials. Comparing this equation with Equation (2.143) it is seen that the piezoelectric effect is similar as far as interface electric fields are concerned to an interface charge density. If both interface states and surfaces changes are present then

$$\varepsilon_p E_p = \varepsilon_n E_n + Q_I + (e_n - e_p) \varepsilon. \tag{2.181}$$

The net result of the piezoelectric effect on the heterojunction space charge regions is to cause a displacement of the space charge regions. The total voltage across the space charge region is determined by the equilibrium Fermi levels and is independent of the piezoelectric effect. The piezoelectric effect increases the electric field on one side of the interface and decreases it on the other side causing a displacement of the the space charge regions. This is llustrated in

Figure 2.50, where the piezoelectric material is assumed to be the n-type material. Under a stress the total charge per unit area within the space charge region is not zero but equals the net polarization.

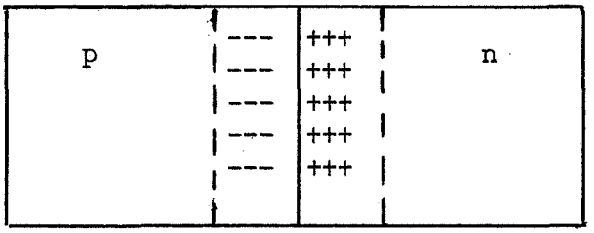
The important effect which the piezoelectric phenomena has on the space charge region is the shift in the fractions of the applied voltage which appears across the n- and p-type regions. As previously noted the effect of a polarization in Equation(2.180) is equivalent to an interface surface charge. Consequently the total voltages V_n and V_p across the two space charge regions can be shown graphically in terms of the applied voltage and polarization as in Figure 2.51. This is identical to Figure 2.44 except that now the shift of the straight line curve along the horizontal axis is proportional to the net polarization instead of surface state density. A change in polarization consequently changes both V_n and V_p while keeping the sum of these constant. The two mathematical equations describing V_n and V_p are

$$(V_n/V_o) + (V_p/V_o) = (1 - V_a/V_o), \tag{2.182}$$

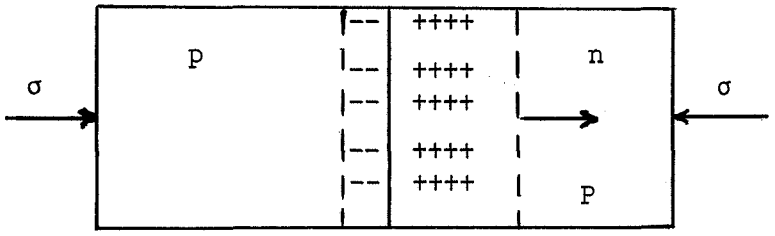
$$(V_n/V_o)^{1/2} + P/(2qN_D\varepsilon_n V_o)^{1/2} = (N_A\varepsilon_p/N_D\varepsilon_n)^{1/2}(V_p/V_o)^{1/2}, \tag{2.183}$$

where

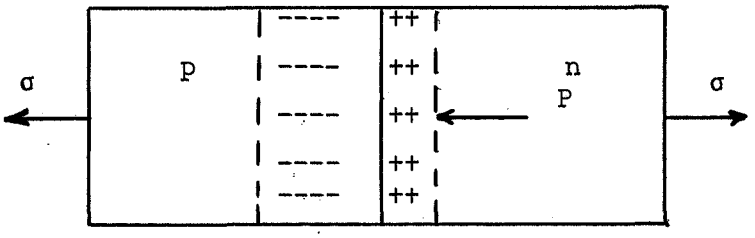
$$P = (e_n - e_p)\,\varepsilon. \tag{2.184}$$



a. No stress



b. Positive applied stress (compression)



c. Negative applied stress (tension)

Figure 2.50. Illustration of the effect of a piezoelectric semiconductor on the space charge regions of a heterojunction, n-type material assumed to be piezoelectric.

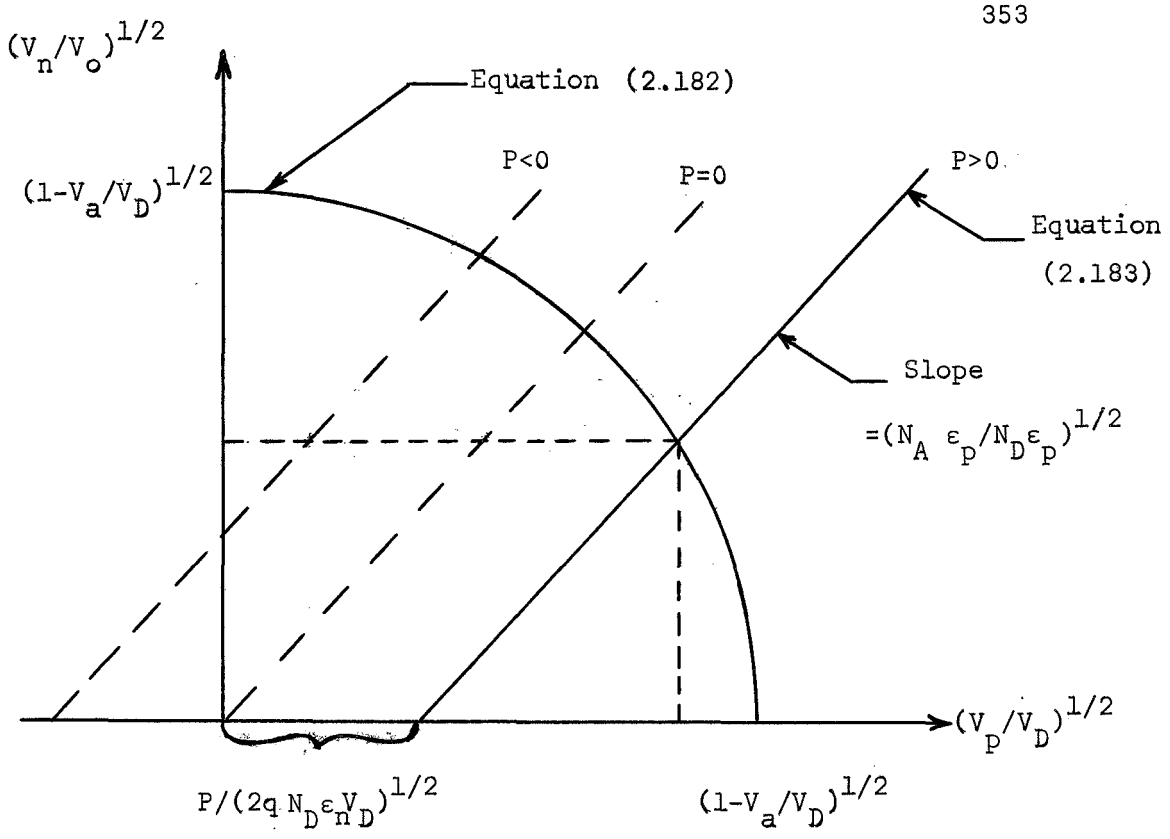


Figure 2.51. Graphical solution of heterojunction voltage equations for a net piezoelectric effect (Q_I=0 for this case).

The exact solution for V_n and V_p is complicated, so only small signal variations in P will be considered here. For small variations one can write [64]

$$V_n = V_{no} - \alpha P, \tag{2.185}$$

$$V_p = V_{po} + \alpha P, \tag{2.186}$$

where V_{no} and V_{po} are the equilibrium values of voltages across the n- and p- type regions as given by Equation (2.146) and (2.147) and α is a constant given by the expression

$$\alpha = 2\,[(V_o - V_a)/2q]^{1/2}\,\left[\frac{N_A\,\varepsilon_p\,N_D\,\varepsilon_n}{(N_D\,\varepsilon_n + N_A\,\varepsilon_p)^3}\right]^{1/2}. \qquad (2.187)$$

For the special case of the symmetrical heterojunction ($N_D\,\varepsilon_n = N_A\,\varepsilon_p$) the above reduces to

$$\alpha = \left[\frac{V_o - V_a}{4q\,N_D\,\varepsilon_n}\right]^{1/2}. \qquad (2.188)$$

A numberical calculation will now illustrate the magnitude of this peizoelectric effect. For illustration purposes consider the symmetrical junction and take $V_o - V_a = 0.5$ volts, $N_D = 10^{16}/cm^3$ and $\varepsilon_n = 4\,\varepsilon_o$. Then

$$\alpha = 1.48 \times 10^7 \text{ V. cm}^2/\text{coul}. \qquad (2.189)$$

For most materials the piezoelectric constant is around 10^{-5} coul/cm^2 per unit strain, so

$$\alpha P = \alpha(e_n - e_p)\,\varepsilon = (148 \text{ volts})\,\varepsilon. \qquad (2.190)$$

This indicates that a strain of 10^{-3} results in a shift in V_n and V_p of about 0.148 volts. This is a significant voltage shift and if the heterojunction current depends on the individual barrier heights this will result in large changes in the diode current.

The above example has illustrated the changes in V_n and V_p which occur under stress due to the piezoelectric effect. In order for this to effect the terminal I-V characteristics, the current must depend on V_n or V_p. Looking at the injected current expressions for the three types of barriers in Equations (2.148) -(2.150) it is seen that J depends on V_n or V_p only for the positive reverse barrier case. In the other two cases J depends only on the total barrier height and the piezoelectric effect does not change the total barrier height. Thus according to the ideal theory only the positive reverse barrier type of heterojunction should be sensitive to stress through the piezoelectric effect. The simplest type of heterojunction which has a positive reverse barrier and is piezoelectric is the Ge-GaAs heterojunction. However no experimental measurements appear to have been made on the stress sensitivity of this type of heterojunction. The theory gives at large forward bias for an n-type barrier

$$J_n = J_o \exp(-q\, V_n/kT) \tag{2.191}$$

$$J_n = J_o \exp(-K_n\, qV_o/kT) \exp(K_n\, qV_a/kT) \exp(q\, \alpha P/kT). \tag{2.192}$$

From this the current and voltage gauge factors can be evaluated as

$$\frac{1}{I}\,\frac{dI}{d\varepsilon} = \frac{q}{kT}\,\alpha(e_n - e_p), \tag{2.193}$$

$$\frac{1}{V_a}\,\frac{dV_a}{d\varepsilon} = \frac{1}{K_n\, V_a}\,\alpha(e_n - e_p). \tag{2.194}$$

Using the values previously used for illustration purposes (plus $V_a \simeq 0.5$ v, $K_n = 1/2$ these become

$$\frac{1}{I}\frac{dI}{d\varepsilon} = 5700,$$

$$\frac{1}{V}\frac{dV}{d\varepsilon} = 590. \qquad (2.195)$$

These are quite large gauge factors as compared with the piezoresistance effect and are also considerably larger than the piezojunction effect of normal p-n junctions arising from energy bandgap charges.

Most of the experimental work which has been done on heterojunction stress sensors has been with evaporated n-type CdSe, p-type Se heterojunctions [67],[68],[69]. Both of these semiconductors are piezoelectric but the combination results in a net piezoelectric polarization. Shown in Figure 2.52 are reported experimental I-V curves for these diodes under compression and tension. Shown in Figure 2.53 is detailed data on the voltage charge at constant current as a function of strain. The data indicates that the response is quite linear with strain at low strain levels and the device is quite sensitive. An experimental voltage gauge factor of 320 is obtained from the data and a current gauge factor of about 600 is obtained.

The energy band structure of the CdSe-Se heterojunctions discussed above are believed to be of the normal barrier height type as illustrated in Figure 2.41 (b). According to the injected current model this type of barrier should not be stress sensitive since the current depends only

on the total barrier height. Experimentally however such heterojunctions are quite stress sensitive as discussed above so there is obviously a discrepancy which has not been discussed in the literature. One possible reason for the stress sensitivity of the CdSe-Se junctions may be the fact that the current is controlled by generation-recombination levels at the junction interface and not controlled by the ideal injected current. The approximate expressions for generation-recombination current as presented in Equation (2.163) and (2.164) both depend on the individual barrier heights V_n or V_p. To account for dielectric

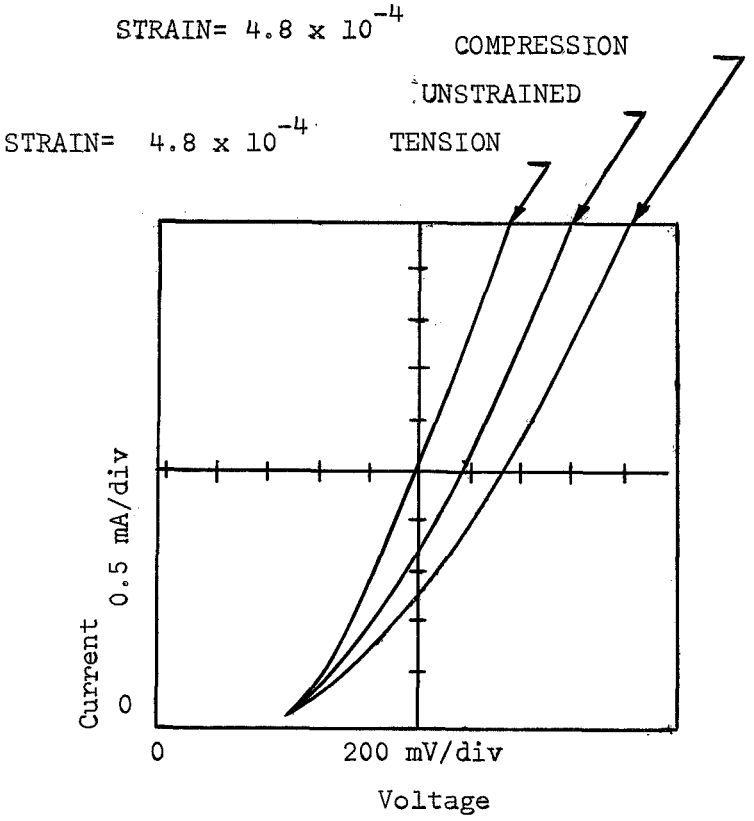


Figure 2.52. Effect of strain on forward I-V characteristics of CdSe - Se heterojunction [67].

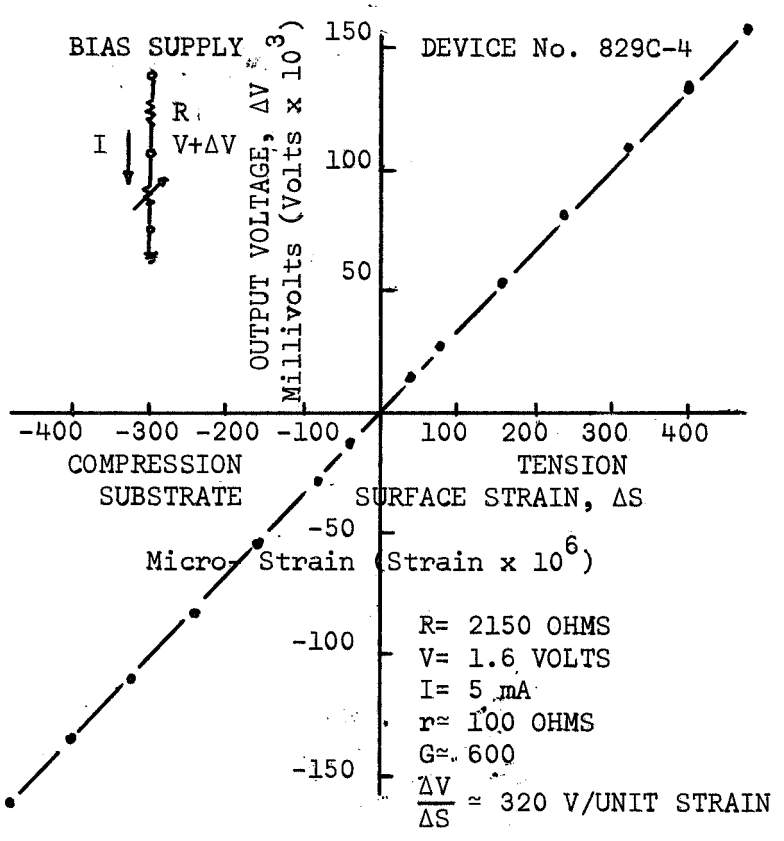


Figure 2.53. Voltage-strain characteristic of CdSe -Se heterojunction [67].

polarization in these expressions V_{no} and V_{po} should be replaced by $V_{no} - \alpha P$ and $V_{po} + \alpha P$ respectively. Considering the case 2 generation-recombination current of Equation (2.164) in the presence of a piezo-electric effect

$$J_{g-r} = J_o \exp(-K_n\, q\, V_o/kT) \exp(q\, K_n\, V_a/kT) \exp(q\, \alpha P/kT). \tag{2.196}$$

This equation has exactly the same dependence on applied voltage and polarization P as does equation (2.192) for the injected current of the positive reverse barrier heterojunction. Thus a heterojunction dominated by generation-recombination current at the junction interface can exhibit the same stress dependence due to the piezoelectric effect as does the ideal positive reverse barrier diode. This may provide the explanation of the stress sensitivity of the CdSe-Se heterojunction.

From the previous discussion it can be seen that the piezoelectric heterojunction is very sensitive to stress. At low stress levels it appears to be one of the most sensitive physical effects which have been observed for stress sensors. At large stress levels the piezojunction effect of normal p-n junctions can approach the observed sensitivity of the CdSe-Se heterojunctions but at low stress levels the heterojunction is superior. Another advantage of the CdSe-Se heterojunction sensor is that it can be made in thin film form by evaporation. It may however suffer from the reproducibility and stability problem of most thin film devices. There does not appear to be any reported data on these problems.

The frequency response of the heterojunction stress sensor should be quite high. Limiting factors are the mechanical response of the sensor, the response of the piezoelectric polarization, and the response of the junction. The frequency response has been reported to extend up to at least 2MHz.

In summary it can be stated that the piezoelectric heterojunction appears to be a very potentially useful stress sensor. The large sensitivity exhibited by such devices at low stress levels is an

especially important feature of these junctions. The stress sensitivity arises from a unique combination of effects whereby one stress sensitive physical phenomena, the piezoelectric effect, is combined with the current-voltage properties of a heterojunction to produce a very sensitive stress sensor.

2.2.4. Other Heterojunction Sensors

The previous sections have explored the most important sensor applications of heterojunctions which have developed to date. These optical sensors and stress sensors take advantage of unique properties of the heterojunction to produce sensors with features which can not be achieved by standard p-n junctions. In principle there is no reason why heterojunctions can not be used in a variety of other sensor applications such as temperature sensors or particle detectors. In these other areas, however, heterojunctions do not appear to offer any unique advantages over normal p-n junctions. For temperature as an example the current is exponentially related to 1/T as in a normal p-n junction. While the temperature sensitivity may be slightly different than for a normal p-n junction it is not sufficiently different to make it competitive with normal p-n junctions. The area of optical sensors and stress sensors appear to be the only areas in which the heterojunction offers unique sensor functions.

2.3. Other Semiconductor p-n Junction Devices

The basic semiconductor p-n junction devices have been discussed in the two previous sections. There are other more complicated junction devices which can and have been used in sensor applications. The most important of these are bipolar transistors, p-n-p-n switches and junction field effect transistors. The object of the present section is to briefly discuss the use of these more complicated devices as sensors. All of these devices contain two or more p-n junctions as opposed to one simple p-n junction discussed in the two previous sections.

2.3.1. Bipolar Transistors

The addition of a second reverse biased p-n junction within a diffusion length or less of a forward biased p-n junction gives a three terminal bipolar transistor which is capable of current amplification. A cross section of a typical n-p-n transistor is shown in Figure 2.54. Most of the minority carriers injected into the base by the emitter-base junction are collected by the collector so the majority of carriers flow from the emitter to the collector. The base current is composed of carriers injected into the emitter plus carriers which recombine in the base plus any non-ideal generation-recombination current within the space charge regions. The basic function of the bipolar transistor is to amplify the base current. The current amplification is characterized by the transistor beta defined as

$$\beta = \frac{I_c}{I_b} \quad , \tag{2.197}$$

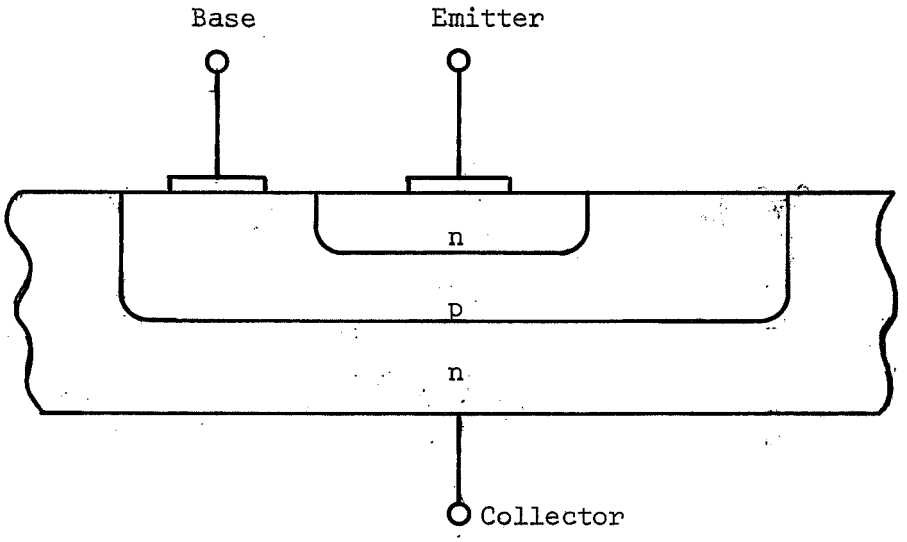


Figure 2.54. Cross sectional view of bipolar transistor.

with typical values ranging from 50 to several hundred.

In practically all sensor applications of transistors the operation is essentially that of a single junction sensor followed by the internal gain of the transistor. The most common use of a transistor as a sensor is the optical phototransistor sensor. In this operation no external connection is made to the base and the transistor is opeated as a two terminal device as shown in Figure 2.55. Operation of the device proceeds as follows. The incident photons on the emitter region create electron-hole pairs which are seperated by the field at the emitter p-n junction. For the n-p-n device of Figure 2.55, the positive charge ends up in the base region and the negative charge within the emitter region. These charges tend to forward bias the emitter junction causing minority carriers to be injected into the base and collected at the collector contributing to the

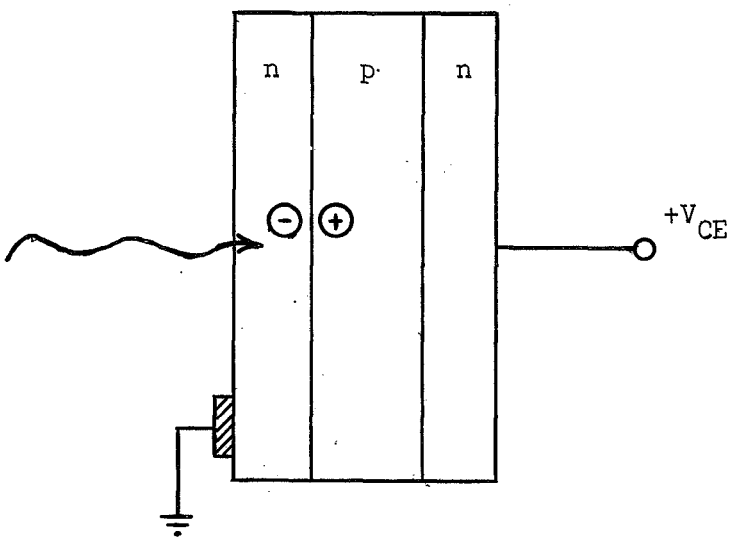


Figure 2.55. Operation of phototransistor

collector current. The positive charge in the base created by the incident photon is equivalent to adding a positive charge to the base by means of an external base contact. Thus the p-n junction photo-current gives an equivalent base current to the transistor which in turn is amplified and a larger collector current appears. This process can be expressed as

$$I_b = I_{ph} \tag{2.198}$$

$$I_c = \beta I_b = \beta I_{ph},$$

where I_{ph} is the normal photocurrent of a p-n junction and has been expressed in Section 2.1 as (see Equation (2.61))

$$I_{ph} = q \frac{P}{h\nu} (1 - R) \overline{\alpha L}, \qquad (2.199)$$

where P is the incident power. Thus a constant incident power on a phototransistor acts like a constant base current. Figure 2.56 shows the collector current vs. collector voltage characteristics of a typical phototransistor. The curves are for constant incident power densities and as can be seen from the figure the characteristics are very similar to transistor characteristics at constant base current.

As discussed above, the phototransistor is essentially a photo diode internally coupled with a transistor which provides additional gain. As such the phototransistor has essentially all the features and problems of a normal p-n junction photo sensor followed by a transistor to provide gain. As with any transistor, the internal gain is a function of current level with the gain decreasing both at low light levels and at high light levels. Since β is also temperature dependent, the photocurrent for a phototransistor is also temperature dependent. The major advantage of the phototransistor over a simple diode is the higher sensitivity provided by the internal transistor gain.

Transistor have also been investigated for use as stress sensors [34],[39]. Again as with the optical transistor sensor the basic mode of operation is that of a stress sensitive p-n junction coupled with the internal gain provided by the transistor. The normal method of employing a transistor as a stress sensor is illustrated in Figure

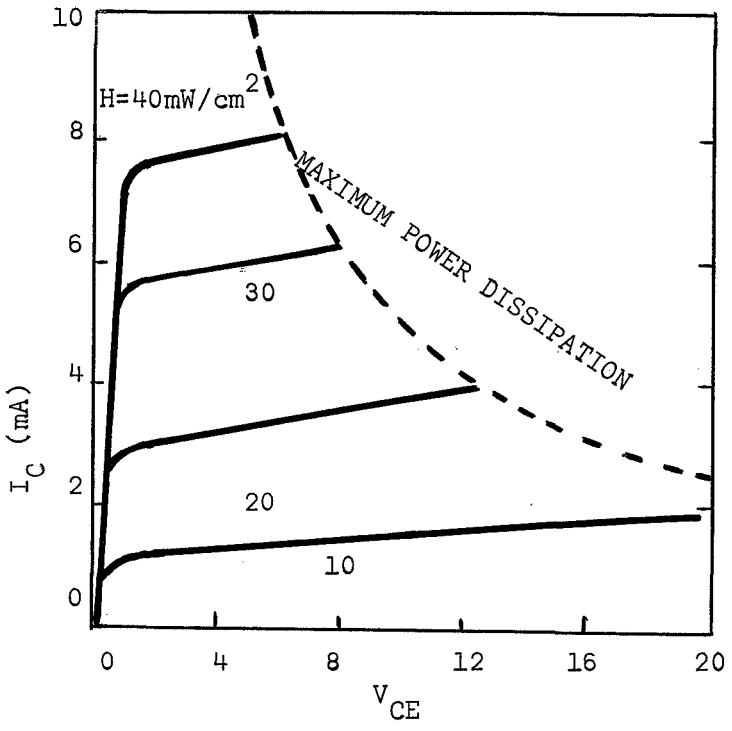


Figure 2.56. Typical collector current vs. collector voltage characteristics of a phototransistor [31].

2.57. The stress is normally applied to the emitter base junction by means of a small stressing point (or indenter point). This gives a very anisotropic stress field in the transistor. If the stress is applied so that the largest force occurs on the emitter side of the p-n junction, the stress induced diode current has been shown to effect mainly the base current and not the collector current [39]. The stress induced changes in emitter-base diode properties are then amplified by the transistor beta. Shown in Figure 2.58 are experimentally measured changes in collector current at constant base current and constant emitter current as a function of force on a point stressing a transistor [40]. The curves are essentially changes in alpha and beta for the transistor normalized to their unstressed

values. The experimental observations are in good agreement with the theory of stress induced diode behavior followed by the gain of the transistor.

In order to use the internal gain of a transistor with the stress effects of the emitter base junction, the applied stress must be anisotropic. If a uniform uniaxial stress is applied to a transistor the emitter-base I-V characteristics are changed as for a normal p-n junction but the beta of the transistor remains essentially unchanged. This makes it very difficult to reproducibly build stress sensitive transistors. The small stressed point has not proven to be

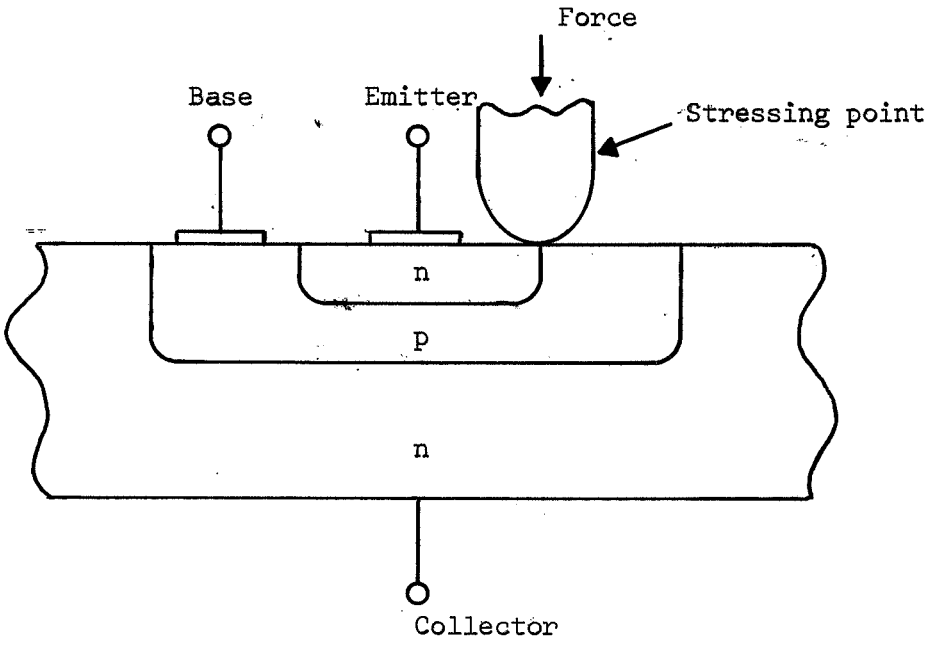


Figure 2.57. Illustration of transistor used as stress sensor.

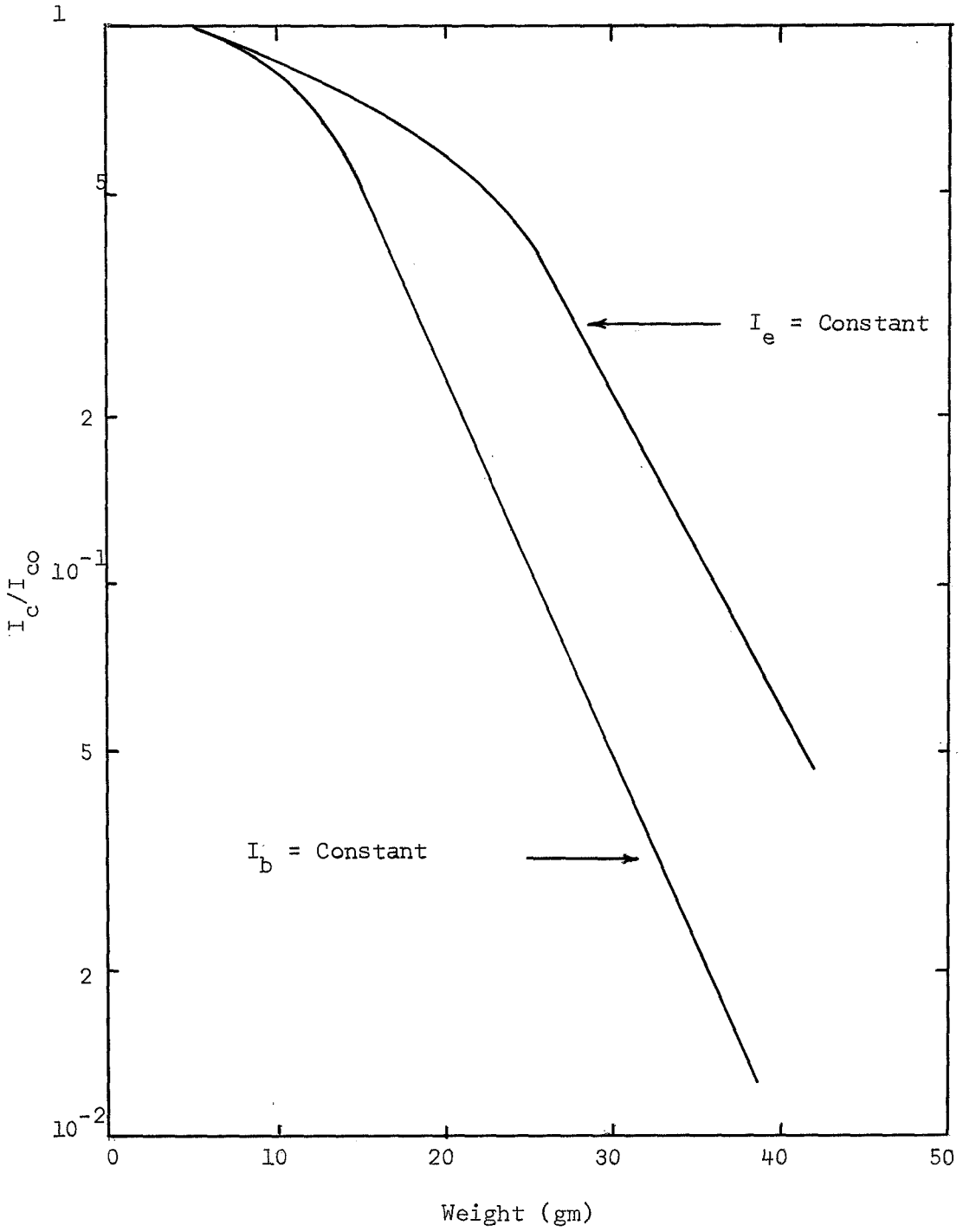


Figure 2.58. Ratio of stressed to unstressed collector current of a 2N1958 n-p-n transistor with constant base current and constant emitter current [40].

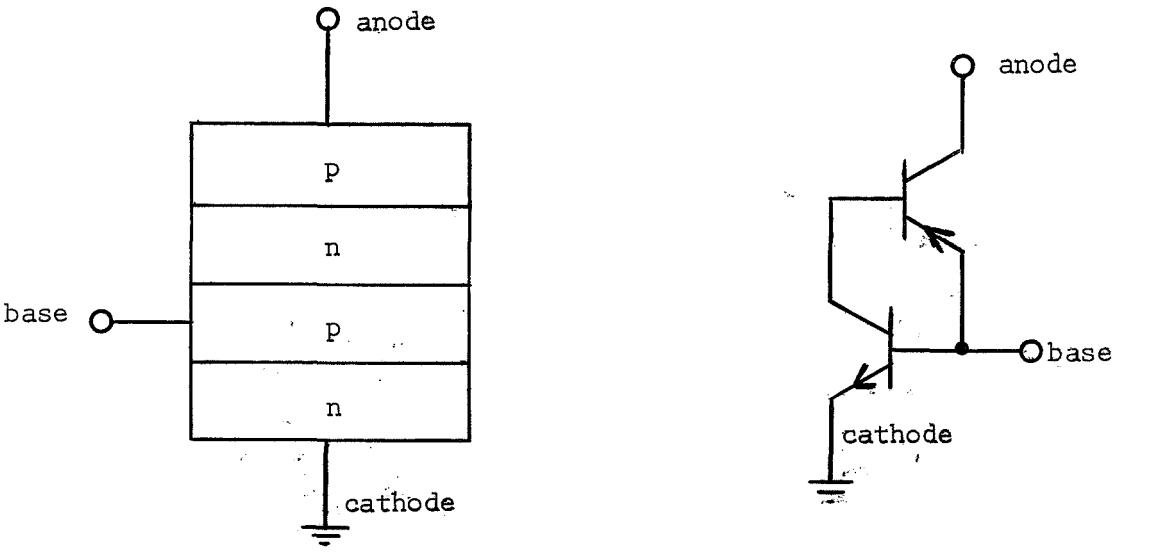
a reliable method of stressing a p-n junction. The advantage of the stressed transistor over a stressed p-n junction is the internal gain of the transistor. Essentially the same gain as achieved by the stressed transistor, however, can be obtained by connecting a transistor in parallel with a stressed p-n junction and operating the combination at a constant current. As far as a stress sensor is concerned the disadvantage of the transistor appear to outweight the advantages considering the fact that the same gain can be achieved by a stage of amplification.

The two areas discussed above are the two major areas in which transistors have been investigated for sensor applications. Basically the advantage of the transistor sensor is the fact that it provides amplification of the basic sensor properties of p-n junctions. In general this makes the transistor about two orders of magnitude more sensitive than a normal p-n junction. A transistor could in fact be operated in the avalanche region to provide even higher gains than obtained by the basic minority carrier gain phenomena. It should always be remembered, however, that the basic sensor is the p-n junction and the increased sensitivity can also be obtained by employing a p-n junction sensor followed by a transistor amplifier. The seperation of sensor and amplifier functions can give greatly increased flexibility, especially for stress sensors. The properties of the transistor sensor should always be compared with the properties of a p-n junction followed by a transistor amplifier. In this light the advantages of the transistor become much less significant.

2.3.2. P-N-P-N Devices

The p-n-p-n device is a more complex solid state device than the transistor requiring three p-n junctions. Any phenomena which effects the I-V characteristics of any of the p-n junctions will obviously effect the overall terminal characteristics. The properties of the p-n-p-n device are known to be equivalent to an n-p-n transistor connected to a p-n-p transistor as shown in Figure 2.59. The object of this section is not to discuss in detail the theory of operation of this device but rather to discuss one type of use of this device as a sensor.

The p-n-p-n device has a very large internal gain and exhibits a negative resistance region. The terminal I-V characteristics are sketched in Figure 2.60. The stable operating regions are a low current high voltage region and a high current low voltage region. The voltage at which the device switches from one state to the other is influenced by the base current I_b. As shown in the figure, increasing base current causes the device to switch states at decreasing voltage levels. Since the terminal characteristics are sensitive to the base current any change in this current will effect the characteristics. Such changes can be due to light or stress in the same manner that the base current in a transistor can be changed by these external forces. Photons incident on the lower p-n junction in Figure 2.59 will result in a junction current which is directly equivalent to a base current as discussed in the previous section in connection with the transistor. Also mechanical stress can change the gain of the n-p-n transistor in the equivalent cirucuit of Figure 2.59 (b). Either of these processes will influence the external I-V characteristics and the external I-V characteristics of the



a. Basic structure

b. Equivalent structure

Figure 2.59. P-N-P-N Device and equivalent n-p-n, p-n-p transistor model.

p-n-p-n device can be employed as a stress sensor.

One of the unique features of p-n-p-n devices is the ability to form very simple oscillator circuits with the addition of a resistor and capacitor as shown in Figure 2.61. The p-n-p-n device acts essentially as a switch giving a high impedance while the capacitor charges to the breakover voltage V_B. At this point the diode switches to its low impedance state, discharging the capacitor and finally converts to its high impedance state when the current drops below the minimum holding current. The period of oscillation is determined by the time required for the R-C circuit to charge to the voltage V_B. Any external force such as light or mechanical stress which changes V_B or effectively changes I_b will also change the frequency of oscillation. Thus a simple circuit such as this can be

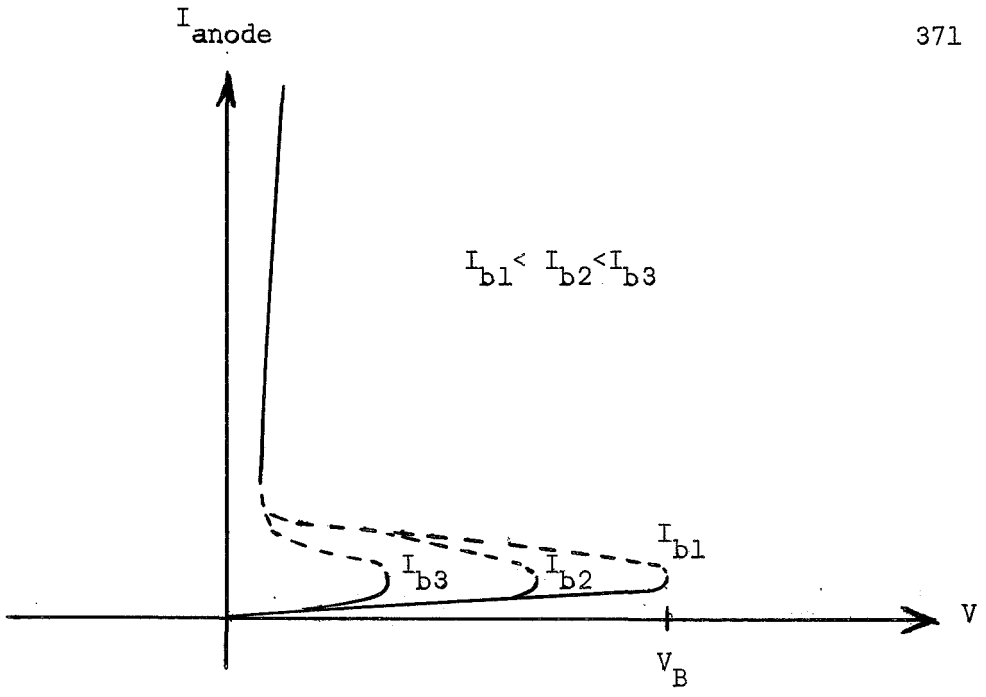


Figure 2.60. Sketch of p-n-p-n device characteristics.

used as a sensor with a digital output. The output frequency of such an oscillator used as a force sensor is shown in Figure 2.62 [70]. In this case the force is applied to a small area of the p-n-p-n device. As can be seen from the data in the figure fairly large frequency shifts can be observed with changing stress. The effect is nonlinear with force as shown in Figure 2.62. However, no attempts have been made to optimise such p-n-p-n sensors. The effects of light on a p-n-p-n device in a circuit such as shown in Figure 2.61 should give similar frequency shifts since this essentially changes the base current.

The above example of the use of a p-n-p-n device under mechanical stress as a digital output sensor provides a unique and interesting use of a solid state device as a sensor. Such use however involves certain limitations and disadvantages as well as advantages. The major advantages are simplicity, small size, and digital output. At first sight these would appear to be major advantages over the simple p-n junction sensor. Here again however as with the transistor it must be remembered that the basic sensor element is one p-n junction and the other junctions simply modify the form of the single junction response. The same type of digital output can be obtained by combining a simple p-n junction sensor with an oscillator where the osccillation frequency is sensitive to the diode properties. Such an oscillator gives a digital output which is just as stress sensitive as a p-n-p-n device [45].

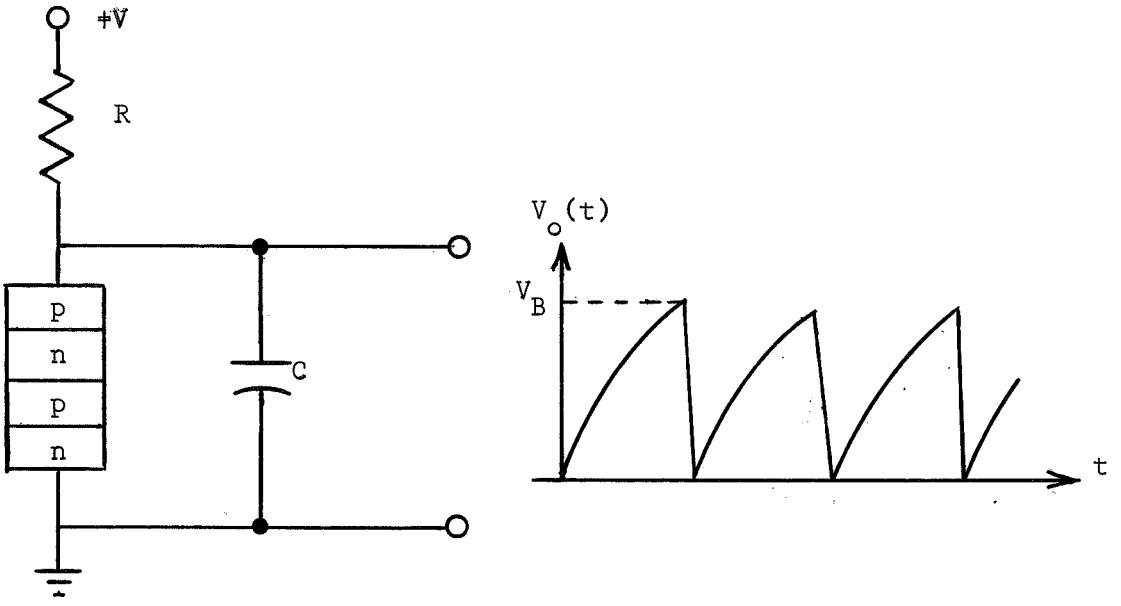


Figure 2.61. Simple oscillator circuit using p-n-p-n device.

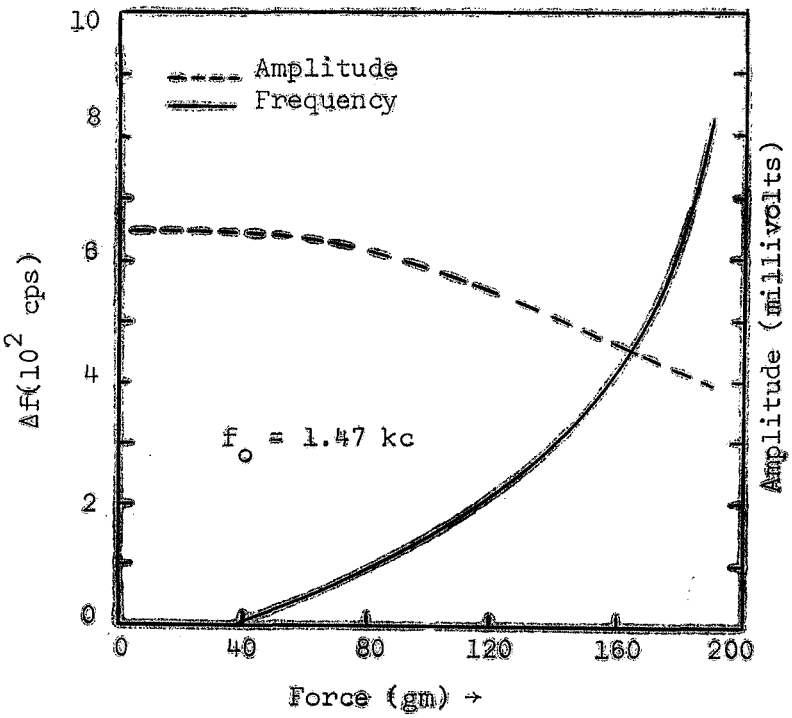


Figure 2.62. Oscillator amplitude and frequency change (Δf) as a function of force for a p-n-p-n oscillator under stress [70].

Since the basic stress sensor in a p-n-p-n device is still a simple p-n junction there is little advantage to the use of such a device as a sensor as opposed to a simple p-n junction followed by signal conditioning circuits. Since the p-n-p-n device provides signal conditioning functions it is not appropriate to compare it directly to a p-n junction sensor. The p-n-p-n device however should be compared to a simple junction sensor followed by equivalent signal conditioning circuits. When this is done there is little advantage to the p-n-p-n device. It will still probably have slightly fewer components, however there are also disadvantages of the p-n-p-n device. The major disadvantage is that of making reproducible devices.

There is always some variation between p-n-p-n devices as well as some variation in the sensor properties of junctions. With a device combining both functions, the problems are compounded. Also by combining sensor and signal conditioning functions some flexibility in design is lost. The p-n-p-n sensor must represent a compromise between an ideal p-n-p-n device and an ideal sensor device, so that one does not have as much design freedom as if the functions are performed seperately. Thus it appears that there is little if any practical advantage to a p-n-p-n device over a simple junction sensor followed by appropriate signal conditioning circuits.

2.3.3. Junction Field Effect Transistors

The junction field effect transistor is a widely used solid state device which has found little use as a sensor. This is due in large part to the fact that it is basically a majority carrier device. The devices discussed in all previous sections depend on minority carriers for their fundamental operating principles. The minority carrier density is many orders of magnitude smaller than the majority carrier density. Consequently the magnitude of change in carrier density needed for a significant change in minority carrier density is orders of magnitude smaller than is required for the same percentage change in majority carrier density. As the junction field effect transistor is normally used, it is only sensitive to effects which influence the majority carriers through either their density or mobility.

The typical cross section of a junction field effect transistor is shown in Figure 2.63. For a device with a uniformly doped channel

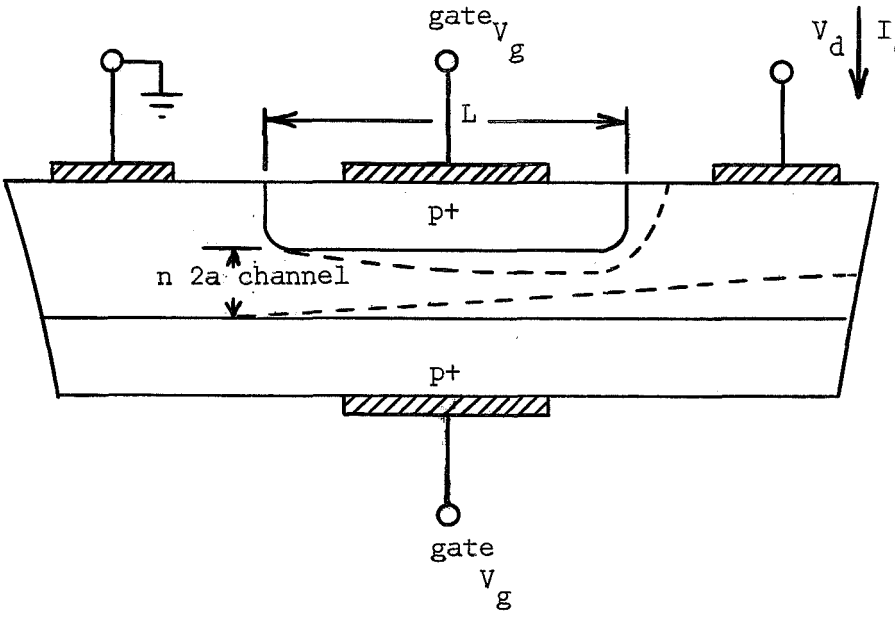


Figure 2.63. Sketch of junction field effect transistor.

the terminal I-V characteristics of an ideal device are described by the equation

$$I_d = G_o V_p \left[\frac{V_d - V_g}{V_p} - \frac{2}{3}\left(\frac{V_d - V_g}{V_p}\right)^{2/3} - \left(\frac{-V_g}{V_p}\right) + \left(\frac{-V_g}{V_p}\right)^{3/2}\right], \qquad (2.200)$$

for $V_d \leq V_g + V_p$,

$$I_d = \frac{G_o V_p}{3}\left[1 + 3\frac{V_g}{V_p} + 2\left(\frac{-V_g}{V_p}\right)^{3/2}\right], \qquad (2.201)$$

$$\text{for } V_d \geq V_g + V_p,$$

where G_o is the unmodulated channel conductance, i.e.

$$G_o = \frac{2 Z a}{\rho L} = \frac{2 q \mu_n N_D Z a}{L}, \qquad (2.202)$$

Z is the device dimension into the page in Figure 2.63 and V_p is the device pinch-off voltage given by the expression

$$V_p = \frac{a^2 q}{2\varepsilon} N_D. \qquad (2.203)$$

The terminal characteristics as described by Equation (2.200) and (2.201) are shown in Figure 2.64. As can be seen from the above expressions, the terminal I-V characteristics depend on the device dimensions and on the mobility μ_n and doping density N_D of carriers within the channel.

Consider first the possible use of a junction FET as a stress sensor. Assuming that the major stress sensitive parameter is the carrier mobility, the current gauge factor is calculated as

$$\frac{1}{I_d}\frac{dI_d}{d\varepsilon} = \frac{1}{\mu_n}\frac{d\mu_n}{d\varepsilon}. \qquad (2.204)$$

This is the same as the gauge factor of a normal semiconductor strain gauge. Since the device is non-linear the voltage gauge

factor (evaluated at constant current) can differ appreciably from the current gauge factor. In fact the voltage gauge factor can be quite large in the current saturation region where the current is essentially constant. Practical devices do not show ideal current saturation as shown in Figure 2.64 but the curves actually have some finite slope in the region where $V_d \geq V_g + V_p$. Operation within this region can be characterized by the drain resistance r_d defined as

$$\frac{1}{r_d} = \frac{dI_d}{dV_d} . \tag{2.203}$$

The voltage gauge factor is then

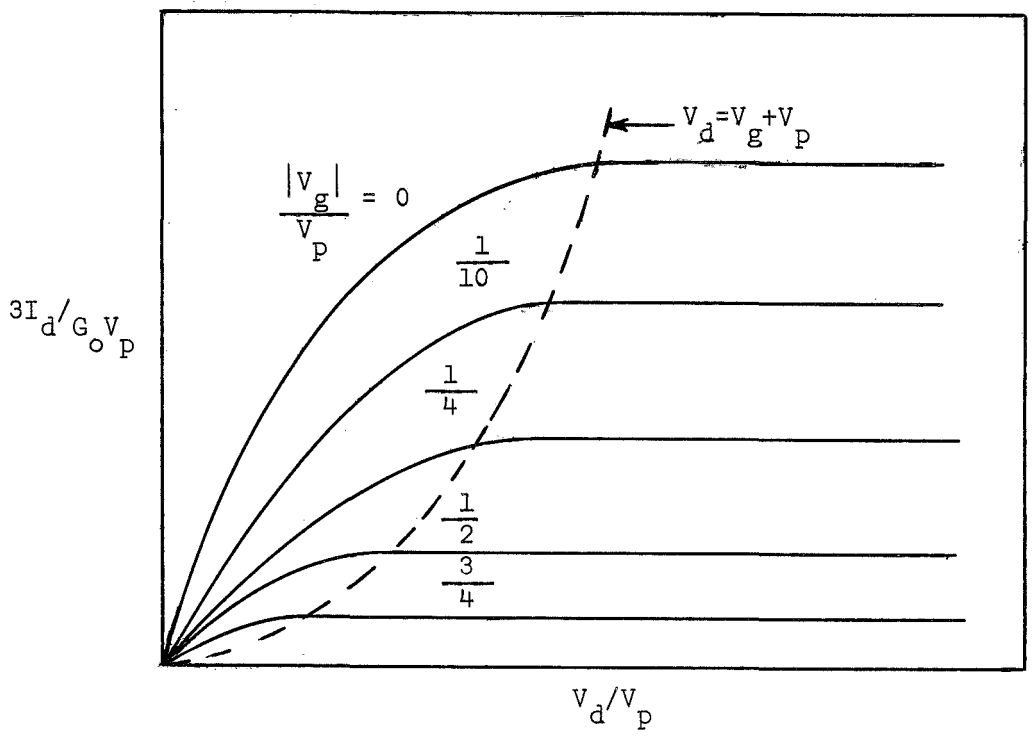


Figure 2.64. Terminal characteristics of junction field effect transistor.

$$\frac{1}{V_d}\frac{dV_d}{d\varepsilon} = \frac{r_d I_d}{V_d}\frac{1}{\mu_n}\frac{d\mu_n}{d\varepsilon} \quad . \qquad (2.204)$$

Typical values might be $r_d = 5 \times 10^4\ \Omega$, $I_d = 5 \times 10^{-3}$ amp, $V_d = 10$ volts. For these values $r_d I_d/V_d = 25$, and in this example the voltage gauge factor is 25 times larger than the piezoresistance effect. The increased voltage sensitivity arises basically from the internal gain present in the junction field effect device and equivalent results can be achieved by combining a normal strain gauge with the gain of a seperate field effect transistor.

Consider next the possible use of the junction FET as a photodetector. Photons incident on the device of Figure 2.63 will create electron-hole pairs which can be seperated by the junction depletion region. Thus a gate photocurrent can be observed in a junction FET. This current will also add to the channel current and constitute a distributed current source along the channel. For typical junction FET dimensions the saturation drain current at a small gate bias voltage is normally in the 10^{-3} amp range. To appreciably effect the terminal properties the photocurrent must be a significant fraction of this current. This requires a fairly large incident photon density to achieve such large photocurrents. The junction FET will be sensitive to low light levels if it is operated at low current levels. This can be achieved by either operating at low drain voltages or by operating at large gate voltages near cutoff. For operation at any voltage, it has been shown that the terminal I-V characteristics are the same as Equations (2.200) and (2.201) if I_d is replaced by $I_d - I_{ph}/2$, where I_{ph} is the total junction photocurrent [71]. As a photodetector

then the junction FET is equivalent to the circuit shown in Figure 2.65. The device is equivalent to a normal junction FET with a photodiode connected between the gate and drain. Because of the distributed nature of the junction FET only one half of the photocurrent is effective in changing the terminal characteristics. The drain voltage sensitivity to light can be large when the FET is operated in the current saturation region.

From the equivalence of Figure 2.65 it is seen that results equivalent to that of the junction FET photodector can be achieved by a normal photodetector followed by an FET amplifier. The only advantage of the FET then is its smaller size achieved by combining both function into one device. This advantage, however, is more than offset in most applications by the loss of freedom in the design of the photodector and FET. It is unlikely that one can achieve optimum designs for both the FET and photodetector with the same device parameters.

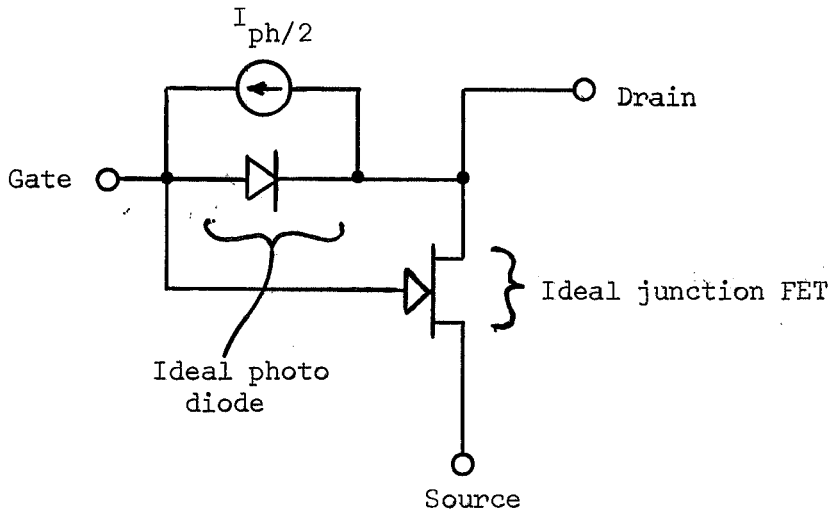


Figure 2.65. Equivalent circuit of junction FET employed as a photodetector.

2.4. Metal - Semiconductor Devices

Metal-semiconductor junctions consist of a metal deposited, usually by vacuum evaporation, on a clean semiconductor surface. Such a metal layer may form either an ohmic or non-ohmic contact to the semiconductor. Only the non-ohmic or metal-semiconductor Schottky barrier diode will be considered in this work. Ohmic metal contacts are necessary for all solid state devices but they do not exhibit useful sensor functions.

The most important sensor application area for Schottky barrier diodes appears to be optical sensors. In this area they offer some unique sensor applications which are not possible with other solid state devices. By proper choice of metal and semiconductor, a wide range of barrier heights can be achieved. The most promising optical sensor area appears to be in the infrared region where low barrier heights are required.

The other sensor area which has been investigated to a lesser extent is the stress sensor area. Metal semiconductor diodes have been found to be stress sensitive and some features appear useful as stress sensors.

2.4.1. Basic Theory

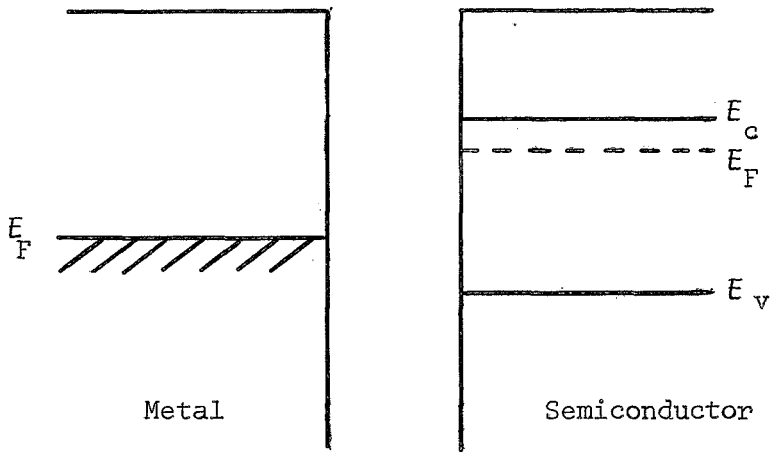
The theory of metal semiconductor devices has now advanced to the point where all major properties of these devices are fairly well understood [9]. Several recent publications have unified many of the previous models which were valied only over certain limited operating ranges [27],[73]. There are basically two types of metal semiconductor contacts, either the non-ohmic rectifying contact or the ohmic contact.

Only the non-ohmic rectifying contact is considered in this work. Such a contact is characterized by a depletion type of space charge region between the metal and semiconductor. The semiconductor may be either n- or p-type. The energy band diagram of a metal n-type semiconductor contact is shown in Figure 2.66. The energy band diagram near the interface is characterized by the barrier height $q\,\phi_{Bm}$ between the metal and semiconductor and by the potential V_o between the bulk of the semiconductor and the metal. Ideally the barrier height would be determined by the difference in work function between the metal and semiconductor. The possible presence of interface states and charges at the interface complicates the situation and the barrier height must be determined experimentally. For a given semiconductor with a constant surface state density it has been shown that the barrier height should be related to the metal work function ϕ_m by an equation of the form [9]

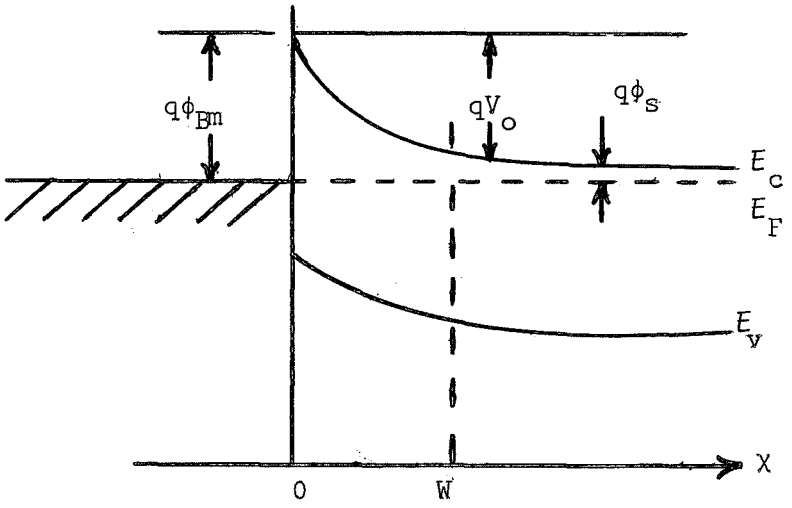
$$\phi_{Bm} = C_2\,\phi_m + C_3, \tag{2.205}$$

where C_2 and C_3 are constants. For most semiconductors an approximate relationship of this form is experimentally observed. For n-type silicon, $C_2 = 0.27 \pm .05$, $C_3 = -0.55 \pm 0.22$ eV [9]. For other semiconductors the constants are slightly different. From the above it is seen that the barrier height can be adjusted over a certain range by different choices of metal.

The barrier height is also slightly dependent on the electric field at the metal-semiconductor interface. This effect arises from the image force on an electron near the metal and is the Schottky



a. Before contact



b. After contact

Figure 2.66. Energy band diagram for metal semiconductor device.

barrier lowering effect. This barrier lowering is illustrated in Figure 2.67 by an expanded view of the energy bands near the interface. The amount of barrier lowering at the interface is related to the electric field at the interface E_s by the expression

$$\Delta\phi = \sqrt{q\, E_s/4\pi\varepsilon}. \qquad (2.206)$$

For typical fields the barrier lowering is less than 0.1 eV. Although this is relatively small it can still have a large effect on the current-voltage characteristic because of the exponential dependence on barrier height. The barrier height ϕ_{Bm} discussed above includes the barrier lowering effect. The quantity ϕ_m will be used to indicate the barrier height which would exist in the absence of barrier lowering. Barrier heights for several metals on Si, Ge, and GaAs semiconductors are shown in Table 2.1 [9].

Table 2.1

Measured Schottky Barrier Heights for Selected Semiconductors [9]

Semiconductor	Metal	Barrier Heights ϕ_{Bm} (Volts) I-V	C-V	Photo
n-Si (Chem)	Au	0.79	0.80	0.78
p-Si (Chem)	Au	0.25		
n-Si (Chem)	Mo	0.59	0.57	0.56
n-Si (Chem)	W	0.67	0.65	0.65
n-Si (Back Sputtering)	PtSi	0.85	0.86	0.85
p-Si (Back Sputtering)	PtSi	0.20		

Table 2.1 (continued)

n-Ge (Vac Cleave)	Au		0.45	
	Al		0.48	
n-Ge	W	0.48		
n-GaAs (Vac Cleav)	Au		0.95	0.90
	Pt		0.94	0.86
	Be		0.82	0.81
	Ag		0.93	0.88
	Cu		0.87	0.82
	Al		0.80	0.80
	Al(770)		0.88	
p-GaAs (Vac Cleav)	Au		0.48	
	Au(770)		0.46	
	Pt(770)		0.48	0.42
	Ag(770)		0.44	
	Cu(770)		0.52	
	Al		0.63	0.50
	Al(770)		0.61	
n-GaAs	W	0.71	0.77	0.80

Values are given for three types of measurements and the agreement is very good between the different techniques.

The most basic property of the metal semiconductor junction is the voltage dependent depletion layer width. For a uniformly doped semiconductor, Poisson's equation is easily solved and the results give [9]

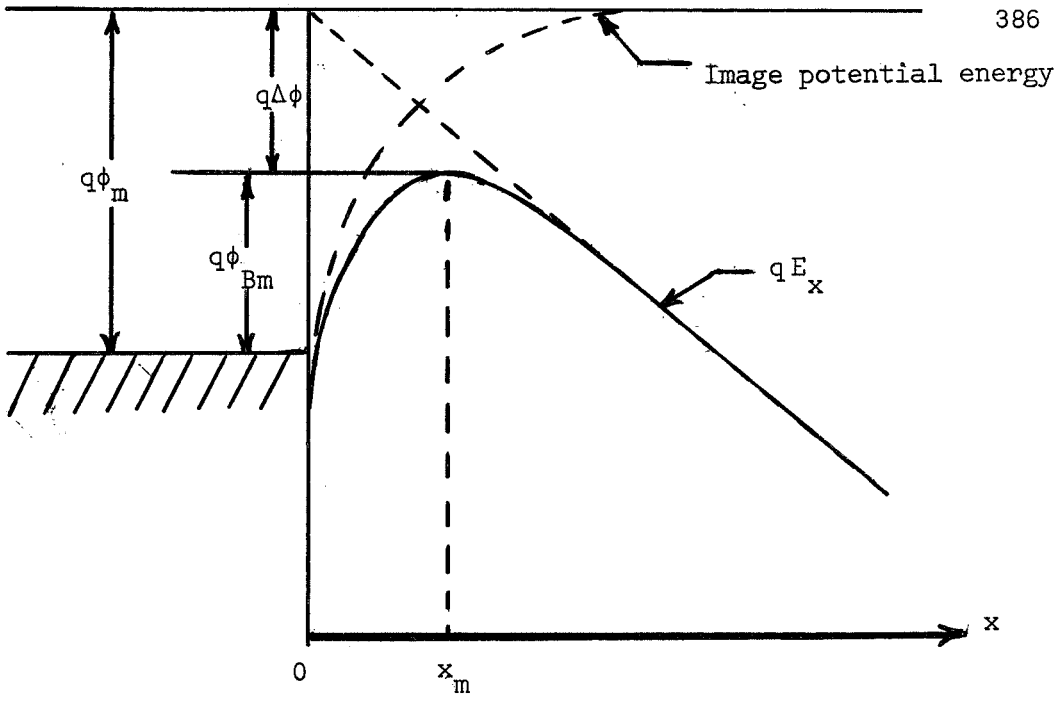


Figure 2.67. Energy band diagram near the interface showing Schottky barrier lowering.

$$W = \sqrt{\frac{2\varepsilon_s}{q\,N_D}\left(V_o - \frac{kT}{q} - V_a\right)}\quad , \tag{2.207}$$

$$E_s = E\big|_{x=0} = \sqrt{\frac{2\,q\,N_D}{\varepsilon_s}\left(V_o - \frac{kT}{q} - V_a\right)}, \tag{2.208}$$

where ε_s is the semiconductor permittivity, W is the width of the depletion layer, and V_a is the applied terminal voltage to the metal semiconductor diode. Another fundamental parameter is the depletion

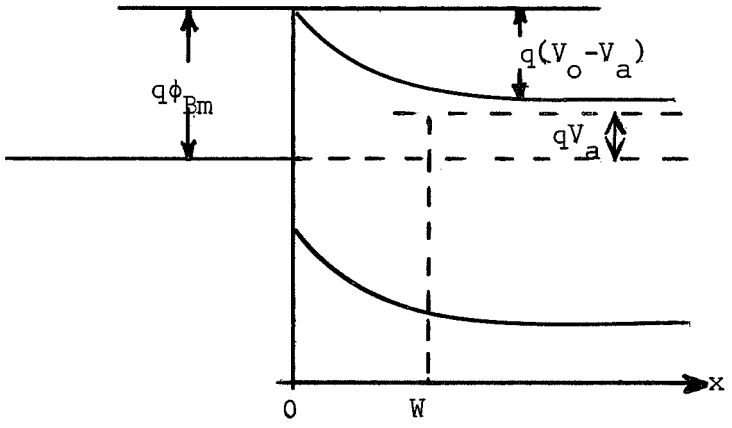
region capacitance per unit area given by the expression

$$C = \left[\frac{2}{q\,\varepsilon_s\,N_D}(V_o - V_a)\right]^{-1/2}. \qquad (2.209)$$

These relationships are similar to the corresponding relationships for abrupt p-n junctions. The electric field at the interface as given by Equation (2.208) is the value of electric field which should be used in the Schottky barrier lowering term of Equation (2.206).

The most important device property of the metal-semiconductor barrier is the terminal current-voltage characteristic under forward and reverse bias. For an n-type semiconductor, forward bias occurs when the semiconductor is biased negatively with respect to the metal. Energy band diagrams under both forward and reverse bias are sketched in Figure 2.68. Under forward bias the barrier for electron flow from the semiconductor to the metal is reduced while the metal-to-semiconductor barrier remains constant (except for small changes in $\Delta\phi$). Under reverse bias the semiconductor-to-metal barrier for electron flow increases.

Consider the forward bias case. There are basically two ways for carriers to cross the barrier from the semiconductor into the metal. First they can be thermally excited to a sufficiently high energy to go over the top of the barrier. Second, carriers with insufficient energy to go over the barrrier may tunnel through the barrier. These two processes are illustrated in Figure 2.69. Consider first the thermionic current over the top of the barrier. There are essentially two models for this current. The first model considers the current



b) Reverse bias

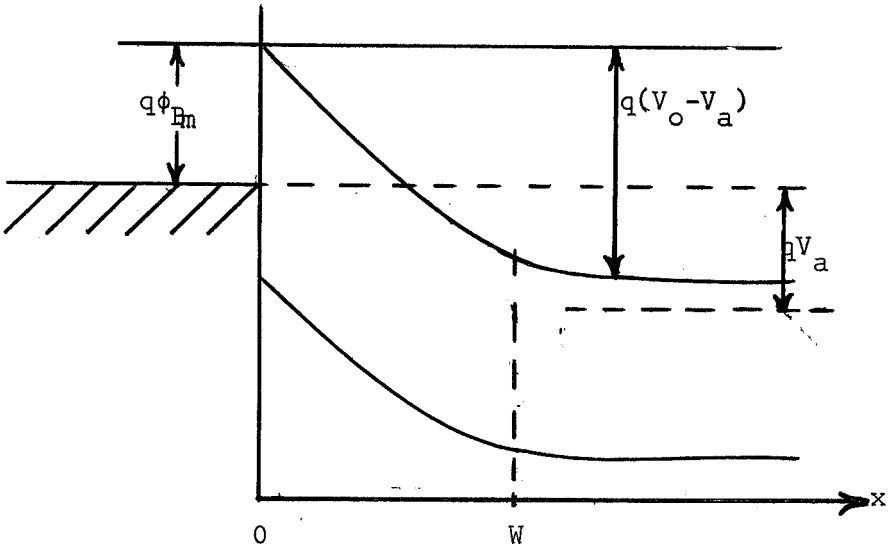


Figure 2.68. Energy band diagrams for metal semiconductor barrier under forward and reverse bias.

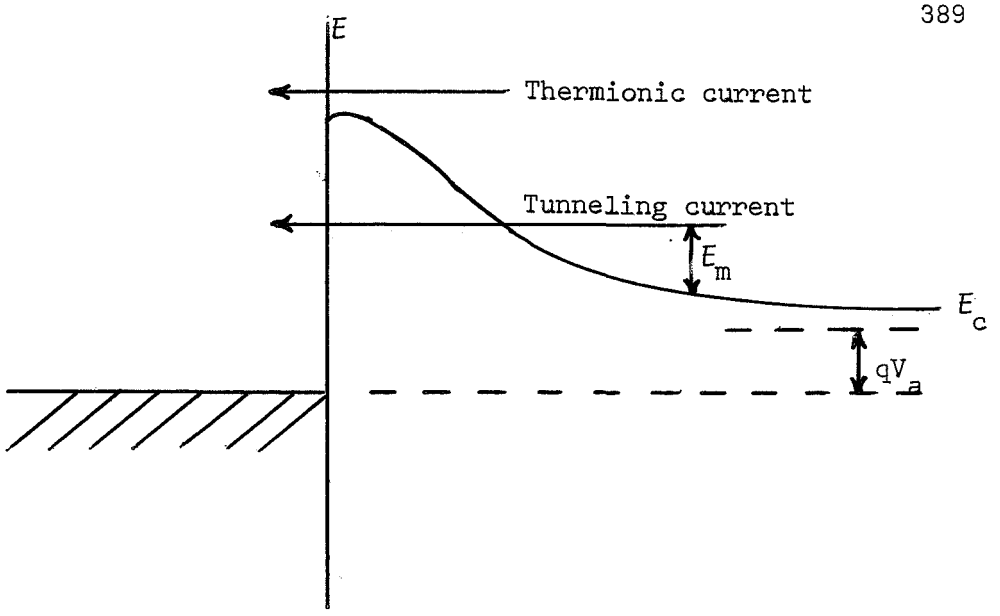


Figure 2.69. Illustration of thermal emission and tunneling currents for metal semiconductor barrier.

flow to be emission limited. This means that the current is determined entirely by the density of carriers moving toward the barrier with sufficient energy to get over the barrier. The second model considers the current flow as diffusion limited, and is based on a solution of the transport equation considering drift and diffusion currents. The emission limited model is valid when the carriers undergo essentially no collisions while traversing the space change region while the diffusion limited model is valid when the carriers undergo many collisions while traversing the space charge region. A unified theory has been reported by Crowell and Sze [72] which combines these two approaches. The predicted terminal current voltage characteristic for thermionic emission is

$$J_{th} = \frac{qN_c v_R}{1 + v_R/v_D} \exp(-q\,\phi_{Bm}/kT)\,[\exp(q\,V_a/kT) - 1], \quad (2.210)$$

where

$$v_R = \frac{A^* T^2}{q\,N_c}, \quad (2.211)$$

$$v_D^{-1} = \int_{x_m}^{W} \frac{q}{\mu_n kT} \exp\left[\frac{-qV(x)}{kT}\right] dx \quad (2.212)$$

$$\simeq \frac{1}{\mu_n} \left[\frac{\varepsilon_s}{q(V_o - V_a)\,8\pi\,N_D}\right]^{1/2} \left\{1 - \exp\left[\frac{-2q\,(V_o - V_a)}{kT}\right]\right\}.$$

In the above equations, N_c is the conduction band density of states and A^* is Richardson's constant, or

$$A^* = \frac{4\pi\,q\,m^*\,k^2}{h^3}. \quad (2.213)$$

For $v_R \ll v_D$ the above reduces to the ideal emission limited case and for $v_R \gg v_D$ it reduces to the ideal diffusion limited case. The current density can be written in the form

$$J_{th} = J_m \exp(-q\,\phi_{Bm}/kT)\,[\exp(q\,V_a/kT) - 1], \quad (2.214)$$

where at small applied voltages J_m is essentially independent of the applied voltage.

Next consider the tunneling current illustrated in Figure 2.69. The tunneling current is found to be strongly peaked about some energy identified as E_m in Figure 2.69. This peak in tunneling current as a function of electron energy arises from two effects. First the density of carriers available to tunnel decreases exponentially with increasing energy. Second the tunneling probability for each carrier is a very rapidly increasing function of energy as energy increases because of the decreasing width and height of the barrier. The product of these two trends leads to a fairly narrow energy range over which most of the tunneling occurs. The energy E_m can be anywhere within the energy range from the edge of the conduction band to the top of the energy barrier i.e. $0 < E_m < q(V_o - V_a)$. A single energy parameter E_{oo} determines the position of the energy peak. The equations describing this are

$$\frac{E_m}{E_b} = [\cosh(E_{oo}/kT)]^{-2}, \tag{2.215}$$

$$E_{oo} = \frac{qh}{4\pi} \left(N_D/m\varepsilon\right)^{1/2}, \tag{2.216}$$

where E_b is the total barrier height from semiconductor-to-metal, or $E_b = q(V_o - V_a)$. For $E_{oo} < 0.1$ kT the energy E_m is essentially at the top of the barrier and the current is emission limited. For $E_{oo} > 10$ kT the energy E_m is essentially at the bottom of the barrier and the tunneling occurs predominately at the bottom of the energy barrier.

The tunneling current has been analyzed in detail by Crowell and Rideout [73]. The theoretical I-V characteristics for the tunneling current are of the form

$$J_{tu} = \left(\frac{J_{ma}}{J_m}\right) J_m \exp[-q(\phi_{Bm} - \phi_s)/E_o] \exp(q V_a/E_o)[1 - \exp(-q V_a/kT)], \tag{2.217}$$

where E_o and J_{ma}/J_m are temperature dependent quantities but are voltage independent, and $J_m = A^* T^2$. For normal values of barrier height, E_{oo} and temperature, the parameter J_{ma}/J_m is between 1 and 10. The parameter E_o is a complex function of the barrier parameters but has the following limiting forms:

$$\begin{aligned} E_o &\simeq kT \text{ for } E_{oo} < 0.1\ kT \\ E_o &\simeq E_{oo} \text{ for } E_{oo} > 10\ kT. \end{aligned} \tag{2.218}$$

The first limit corresponds to the thermionic emission case while the second limit corresponds to tunneling at the bottom of the energy barrier. The above model reduces to the barrier emission limited thermionic model for $E_{oo} \ll kT$. From the above it is seen that the forward bias current depends essentially exponentially on applied voltage for either the thermionic emission current or the tunneling current. The metal-semiconductor diode current will be given by either Equation (2.210) or (2.217) depending on whether the current is emission limited or tunneling limited.

The above discussion has considered the metal-semiconductor diode currents under forward bias. Consider now the reverse bias case which for an n-type semiconductor has the metal negatively biased with respect to the semiconductor as illustrated in Figure 2.68. As in the forward biased case, the current can be either barrier emission limited or tunneling limited. For thermionic emission under reverse bias

$$J_{th} \simeq J_m \exp(-q\ \phi_{Bm}/kT)\ [\exp(q\ V_a/kT) - 1], \qquad (2.219)$$

where ϕ_{Bm} is the total barrier reduced by the Schottky barrier lowering effect; i.e. $\phi_{Bm} = \phi_m - \Delta\phi$. In the absence of Schottky barrier lowering the reverse current would saturate as in a normal p-n junction. The major voltage dependence of reverse current comes from the barrier lowering and under reverse bias

$$J_{th} \simeq - J_m \exp(-q\ \phi_m/kT) \exp\left(\frac{q}{kT}\ \sqrt{q\ E_s/4\pi\ \varepsilon_s}\right), \qquad (2.220)$$

where E_s is related to the applied voltage by Equation (2.208).

Tunneling can also occur under reverse bias and can be the controlling factor for the reverse bias current. For small reverse bias, such that $0 < V_a < \phi_m - V_o$, the model of Conwell and Rideout applies and the tunneling reverse bias current is

$$J_{tu} \simeq -\left(\frac{J_{ma}}{J_m}\right) J_m \exp(-q\ V_o/E_o) \exp\left[-q\ V_a\left(\frac{1}{kT} - \frac{1}{E_o}\right)\right], \qquad (2.221)$$

which has an exponential dependence on applied voltage. For large reverse bias the tunneling is determined by carrier tunneling from below the Fermi level within the metal into the semiconductor. This approaches the Fowler-Nordheim field emission model which gives

$$J_{FN} = C_1 E_s^2 e^{-E_o/E_s}, \tag{2.222}$$

where C_1 and E_o are constants given by the equations

$$C_1 = \frac{q^3}{2\pi h(\phi_m + \phi_F)} (\phi_F/\phi_m)^{1/2}, \tag{2.223}$$

$$E_o = \frac{4}{3q} (2m/\hbar^2)^{1/2} \phi_m^{2/3} \tag{2.224}$$

where ϕ_F is the energy difference between the Fermi level and the conduction band edge in the metal. The total reverse bias current is given by the sum of the thermionic emission current and the tunneling current.

The above discussions have presented several models for current flow in metal semiconductor devices. In general for small barrier heights and/or high temperatures the thermionic emission current dominates in both the forward and reverse bias directions. The tunneling current is only important for large barrier heights, at low temperatures or for very large reverse bias voltages. For most Si, Ge, and GaAs diodes, which have been most extensively studied, the current transport mechanism around room temperature is due to thermionic emission. For such diodes

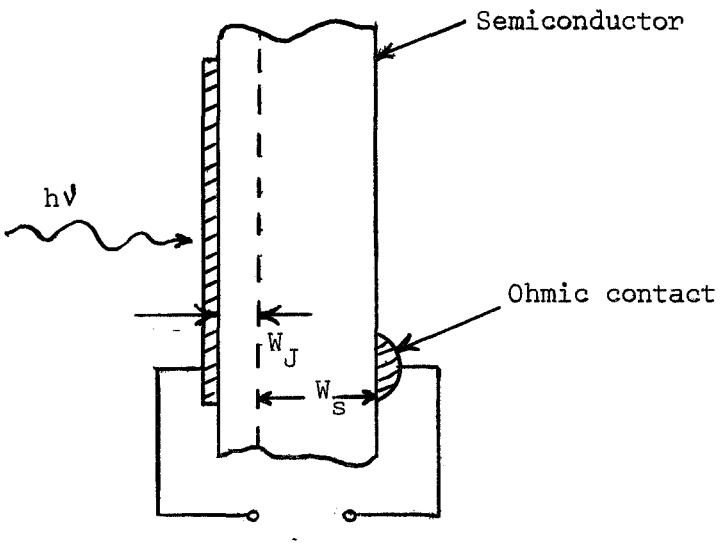
the important current density expressions are Equations (2.210) and (2.220).

2.4.2 Optical Sensors

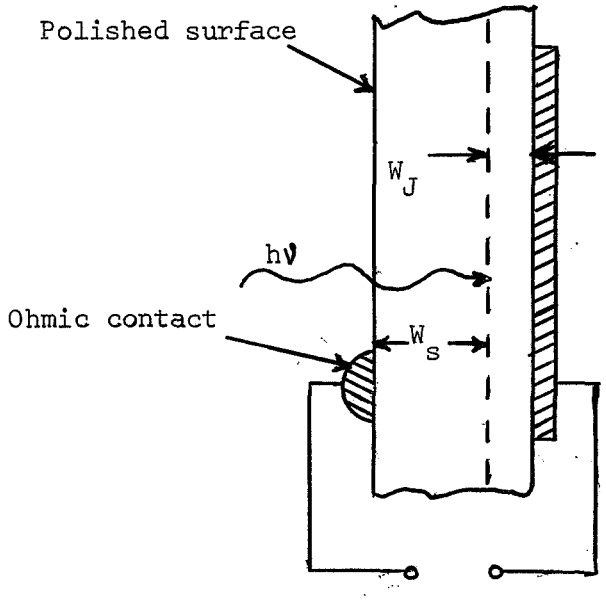
As optical sensors, metal-semiconductor devices offer unique sensors which can not be achieved with normal p-n junctions or hetero-junctions. Metal semiconductor diodes have been shown to have useful optical detection properties in the infrared as well as the visible and into the ultraviolet region of the spectrum.

There are several different modes of operation of metal semiconductor optical detectors. First the detector can be operated with the photons incident on the metal surface or on the semiconductor surface. These two modes of operation are illustrated in Figure 2.70. As other modes of operation, the device can be operated as a detector at energies below the bandgap of the semiconductor or at energies above the bandgap energy. For discussion purposes the operation is divided into the low energy or infrared detectors where $h\nu < E_g$ and the high energy or optical and ultraviolet detectors where $h\nu > E_g$.

For low energy detectors there is little absorption within the semiconductor and few electron-hole pairs created within the semiconductor. The physical mechanism by which detection occurs is the excitation of carriers within the metal. Carriers excited to an energy within the metal which is larger than the metal semiconductor barrier energy can get over the barrier and contribute to a photocurrent. This is illustrated in Figure 2.71. The range of photon energies for which this type of detector exhibits a useful photocurrent is the energy range



(a) Metal surface



(b) Semiconductor surface

Figure 2.70. Illustration of two modes of incident photon energy.

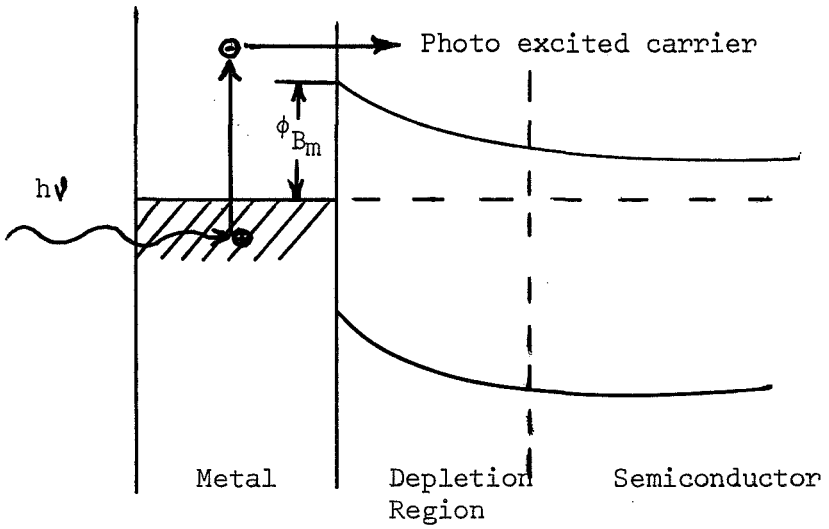


Figure 2.71. Excitation of carriers within metal for metal-semiconductor junction detector.

from the barrier height to the band gap energy or $q\,\phi_{Bm} < h\nu < E_g$. The magnitude of the observed photocurrent depends on several factors. First a large fraction of the incident photon energy can be reflected from the surface of the metal. For bulk metal films in the infrared region as much as 98% or 99% of the incident radiation is reflected. The films used with metal-semiconductor detectors are normally thin films in the range of 50°A to 2000A°. The optical properties of thin films can be considerably different from the optical properties of bulk films. In any case most of the photons are going to be reflected at the surface and only a small percentage of the light enters the metal perhaps 2% to 10% of the incident light. Over a band of wavelengths the surface reflectance can be greatly reduced (to 10% or less) by the use of antireflection coatings.

The photons which enter the metal are capable of exciting electrons to energies sufficient to contribute to the photocurrent. The fraction of light absorbed within the metal depends on the absorption coefficient. For typical metals in the infrared region $\alpha \sim 10^6$/cm or slightly less. Thus for a 100 A° film about 73% of the light is absorbed within the film assuming that bulk properties apply to this thin a film. Thus films of a few hundred A° absorb most of the light which enters the film.

Another factor of importance in determining the photoresponse is the fraction of photoexcited carriers which get over the energy barrier. Only carriers excited to an energy larger than ϕ_{Bm} can get over the barrier and only a fraction of these will have sufficient momentum perpendicular to the surface to surmount the barrier. Cohen, et. al. have shown on the basis of a simple model for the excited carriers that the fraction of excited carriers getting into the semiconductor is [74]

$$y_o = \frac{(h\nu - q\,\phi_{Bm})^2}{8h\nu\,E_F} , \qquad (2.224)$$

where E_F is the Fermi level in the metal above the conduction band edge. To illustrate to magnitude of this assume that $h\nu$ = 1 ev, $q\,\phi_{Bm}$ = 0.5 eV, and E_F = 4 eV. These are typical numbers for an infrared detector. These give $y_o = 7.8 \times 10^{-3}$ which predicts that slightly less than 1% of the excited carriers cross the barrier and contribute to the current. The model of Cohen, et. al. assumes that the metal is sufficiently thin that all excited carriers can contribute to the current. The excited carriers are actually hot carriers within the metal and can only travel some average distance before losing their energy in creating phonons.

For metals this mean free path ℓ_p is on the order of a few hundred angstroms with silver having the largest value of about 570 A°. Only those carriers which are excited within a distance equal to the phonon mean free path can contribute to the current. This gives rise to an additional reduction in photocurrent if the film thickness is larger than the phonon mean free path.

The combination of the above effects leads to a qualitative understanding of the photoresponse of metal semiconductor devices in the infrared spectral region. The most important limitations on the measured efficiency are the reflectance and the internal efficiency as given by Equation (2.224). Except for very thin film thickness most of the light getting into the metal is absorbed within the metal. If the light is incident on the metal surface the metal film thickness must be less than the mean free path of hot electrons within the metal. This means a film thickness of no more than a few hundred angstrons. If the light is incident through the semiconductor, the metal can be as thick as desired and the active region of the metal will be near the semiconductor interface to a depth into the metal of the hot electron mean free path. With light incident through the semiconductor the peak absorption is at the interface so this appears to be best means of operation for a simple metal semiconductor device. If an antireflection coating is used, however, this is more easily done on the metal surface, so in this case using light incident on the metal appears to be the best mode of operation. From combining the effects of reflectance and internal efficiency the dependence of external quantum efficiency on photon energy is

$$\eta = K \frac{(h\nu - q\,\phi_{Bm})^2}{h\nu}, \qquad (2.225)$$

where the constant K accounts for reflection and absorption within the metal film. The photocurrent density corresponding to the above efficiency can be written as

$$J_{ph} = \frac{qP}{A\ h\nu}\ \eta, \tag{2.226}$$

where P/A is the power per unit area incident on the detector.

The most extensively studied infrared detectors using metal semiconductor devices appear to be Au on p-type Si and Ge [74]. The previous discussion has concerned an n-type semiconductor but similar effects occur with metal p-type devices. For Au on Si the barrier height is 0.28 eV and for Au on Ge it is 0.01-0.06 eV. Because of the low barrier height to Ge it is a possible detector out to about 10 μ in the infrared. Shown in Figure 2.72 is $\sqrt{\eta\ h\nu}$ as a function of photon energy for Au p-type silicon diodes at 77°K [75]. From Equation (2.225) it is seen that the model gives

$$\sqrt{\eta\ h\nu} = \sqrt{K}\ (h\nu - q\ \phi_{Bm}). \tag{2.227}$$

This linear relationship to photon energy is seen in Figure 2.72 at low photon energies. Curves are shown for two metal film thicknesses of 60 A° and 2000 A°. The 60 A° film is seen to give a much larger efficiency with K values of 25%/eV and 1%/eV for the two films. For the 60 A° film an external efficiency of 1% is observed at a photon energy of about 0.4 eV or about 3 μ wavelength. The data shown in the figure was obtained by having the light incident through the semiconductor. The observed efficiency for the 60 A° film is about two orders of

magnitude larger than the predictions of the model discussed above based on surface reflection and internal efficiency. The efficiency for the 2000 A° film is much closer to the rough calculation of 10^{-2}% efficiency calculated above. This is a rather fortunate occurance for possible sensor use of metal semiconductor infrared sensors. The enhanced response of the detector for thin metal films is apparently due to the fact that the optical and perhaps electronic properties of thin metal films differ from bulk films. In connection with the work shown in Figure 2.72, Cohen, et. al. studied the absorption in thin gold films and found the amount of light which is absorbed within the film to be

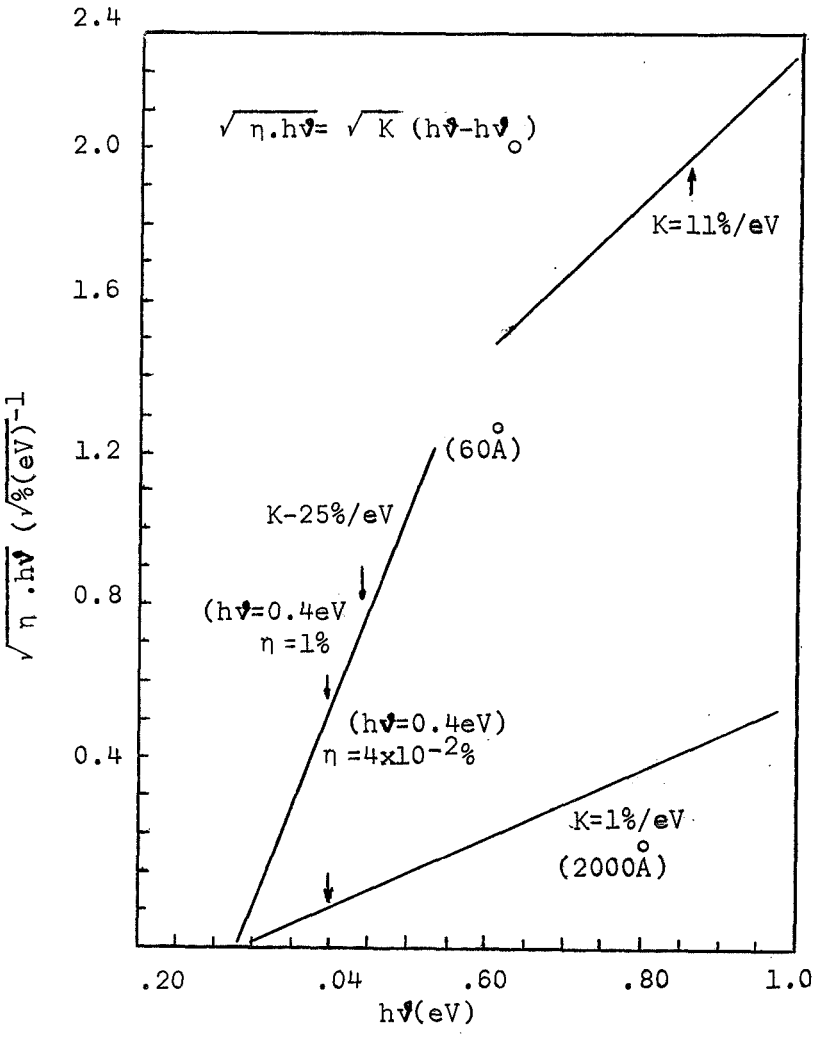


Figure 2.72. $\sqrt{\eta\, h\nu}$ vs hv for 60 A° and 2000 A° Au-p-type Si diodes at 77°K.

dependent on the film thickness.

They found that about 25% of the light is absorbed in 40 A° to 75 A° films but that only about 5% of the light is absorbed in 2000 A° films. This accounts for some but not all of the enchanced efficiency of thin films. The internal efficiency for carriers crossing the barrier also appears to be larger for thin films with thicknesses less than the mean free path which is not too surprising. Regardless of the origin of the effect, metal semiconductor detectors with thin metal films appear to be about a factor of 25 more efficient than thick films.

Shown in Figure 2.73 is similar external efficiency data for Au p-type Ge diodes at 5°K. Again it is seen that the thin 40 A° film is about 25 times more efficient than the 2000 A° film. The curves also indicate that the energy barrier height depends on the metal thickness, being smaller for the thin metal film. This may also account for some of the enhanced efficiency of the thin film. At $h\nu$ = 0.12 eV or around 10μ wavelength the observed efficiency is 1.4% for the 40 A° film and 3×10^{-2}% for the 2000 A° film.

The above results show that is possible to obtain useful infrared sensors using metal-semiconductor diodes. Taking 1% efficiency as being the useful lower limit, the Au p-type Ge detector at 5°K is useful out to 10 μ wavelengths and the Au p-type Si detector at 77°K is useful out to 3 μ wavelengths. As with other infrared detectors the devices must be cooled so that the barrier height is much larger than kT. The efficiency of the metal-semiconductor infrared detector is strongly dependent on the photon energy as given by Equation (2.225). This is not too objectionable however as long as the response can be fitted by

a simple model such as Equation (2.225) and as long as the constants of the model are easily measured.

The previous sections have discussed the operation of metal semiconductor infrared detectors at photon energies below the bandgap energy. The following discussion considers detectors at photon energies larger than the bandgap. For most semiconductors this is in the visible and ultraviolet spectral regions. As soon as the photon energy exceeds

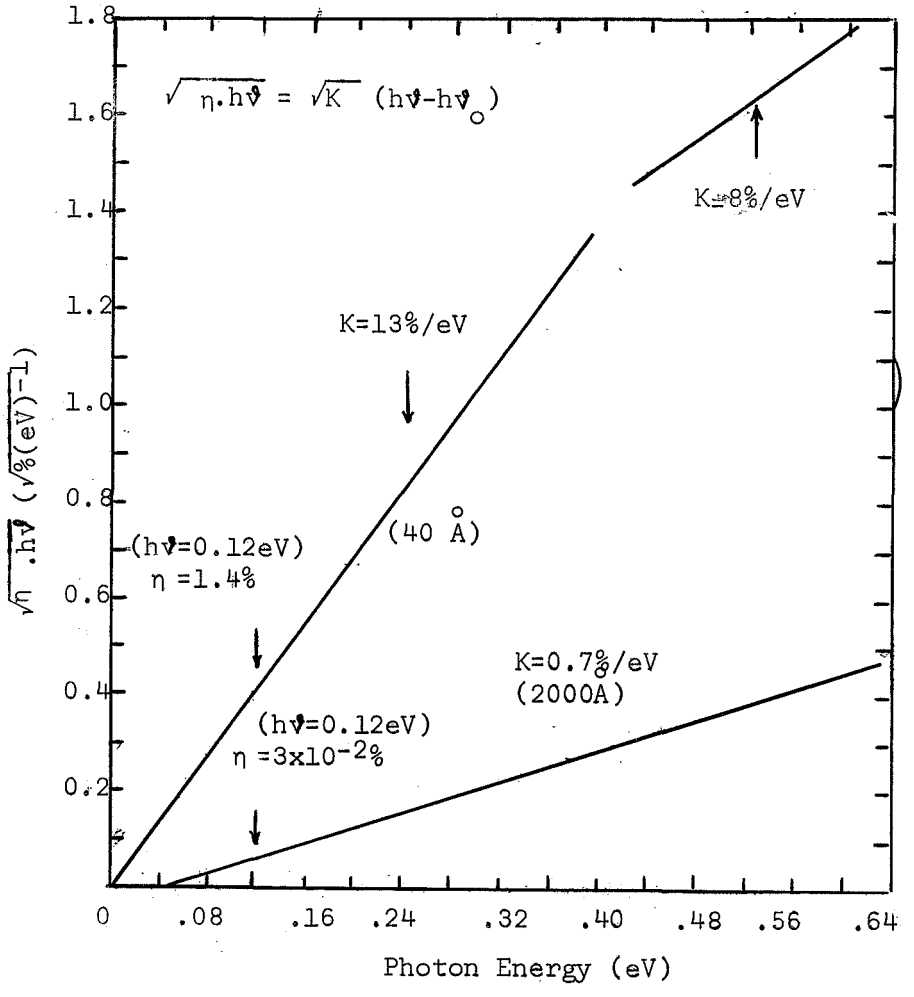


Figure 2.73. $\sqrt{\eta\ h\nu}$ vs hv for 40 A° and 2000 A° Au-p-type Ge diodes at 5°K.

the semiconductor bandgap energy, hole-electron pairs will be created by any light within the semiconductor and this constitutes an additional source of carriers for a measured photocurrent. Any photons absorbed within the metal will still give rise to an injected current at the metal which will add to any current due to carriers generated within the semiconductor. The present discussion considers only the current due to carriers generated within the semiconductor.

The simplest means of exciting the diode is to have the light incident on the semiconductor as shown in Figure 2.70, where W_s is the width of the semiconductor region and W_J is the depletion region width. For this arrangement carriers created within the depletion region and within the semiconductor within a diffusion length of the depletion region will be collected by the space charge region and constitute the photocurrent. This situation is essentially identical as far as absorption and minority carrier transport are concerned to a p-n junction type of photodiode with one region infinitesimally small. For example this is equivalent to the p-n junction device shown in Figure 2.11 with $W_n = 0$, so that one has only the semiconductor region near the surface and the depletion region. From this analogy the equations describing the photocurrent for the above case of metal-semiconductor device are easily obtained from those for the p-n junction as

$$J_{ph} = -q\,\overline{\alpha L}\,N\,, \tag{2.228}$$

where N is the photon flux which enters the semiconductor and is related to the incident power by the expression

$$N = \frac{P}{Ah\nu} (1 - R), \tag{2.229}$$

and

$$\overline{L} = L F + \frac{e^{-\alpha W_s}(1 - e^{-\alpha W_J})}{\alpha} \tag{2.230}$$

with

$$F = \frac{e^{-\alpha W_s}[\sinh (W_s/L) + \frac{SL}{D} \cosh(W_s/L)] - \alpha L (1 + \frac{S}{2D})}{(1 - \alpha^2 L^2)[\cosh (W_s/L) + \frac{SL}{D} \sinh (W_s/L)]}$$

$$+ \frac{\alpha L}{(1 - \alpha^2 L^2)} e^{-\alpha W_s} \tag{2.231}$$

In the above equation, L is the minority carrier diffusion length within the semiconductor, D is the diffusion constant and S is the surface recombination velocity. The operation is essentially identical to a p-n junction with the junction deep within the semiconductor. The various possible modes of operation for a p-n junction photodiode have been discussed in detail in Section 2.1 and only a brief review will be given here. As the photon energy just exceeds the bandgap energy the photocurrent increases because of the increased absroption within the semiconductor. For large photon energies the response normally decreases. This is due to two effects. First the carriers are created closer to the surface as α increases and the internal collection efficiency decreases. The rate at which the response decrease depends

on the width of the semiconductor region and the surface recombination velocity. If the region is wide compared with a diffusion length or if S is very large, a useful photo response is observed only for photon energies within a narrow energy band above the band gap energy. For $W_s >> L$ the peak response occurs at approximately $\alpha L = 1$. The possible use of metal semiconductor devices as narrow band detectors has been discussed theoretically by Sizelove and Love [76]. Their theoretical model predicts for metal-GaAs devices narrow band responses over about 0.02 μ wavelengths around 0.9 μ wavelength with peak internal efficiencies of around 15%. There does not appear to be any experimental evidence concerning such narrow band metal-semiconductor detectors.

When the metal semiconductor device is operated with the light incident on the semiconductor as discussed above it acts very similar to a p-n junction sensor with a junction depth equal to the thickness of the semiconductor. Thus there appears to be little advantage of this arrangement over a p-n junction detector.

Consider next the operation of a metal semiconductor detector with the light incident on the metal surface as shown in Figure 2.70 (a). As with the infrared detector one of the major problems with this arrangement is surface reflectance. Unless the metal is very thin (less than a few hundred anstroms) and an antireflection coating is used, very little of the incident light gets through to the semiconductor to create electron-hole pairs. To compete with the efficiencies obtainable with p-n junctions an antireflection coating is needed. In a study of ZnS antireflection coatings on a Au-Si diode, Schneider has shown that a theoretical transmission of around 80% can be ahieved into the semiconductor at a wave-

length of $\lambda = 6328$ A° [77]. It is reasonable to assume that equivalent transmission efficiencies can be achieved at other wavelengths by appropriate antireflection coatings.

The photons which are transmitted into the semiconductor and absorbed there produce electron-hole pairs which can be collected by the depletion region and contribute to the photocurrent. The generation, diffusion and collection of carriers within the semiconductor is similar to a p-n junction detector with one semiconductor region negligable small. For example the operation is equivalent to the p-n junction detector of Figure 2.11 with $W_p = 0$. From this analogy the photocurrent density is

$$J_{ph} = - q \alpha \overline{L} N, \tag{2.232}$$

where

$$\overline{L} = L F + \frac{(1 - e^{-\alpha W_J})}{\alpha} \tag{2.233}$$

$$F = \frac{\coth(W_s/L) - e^{-\alpha W_s} \cosh(W_s/L) - \alpha L}{(1 - \alpha^2 L^2)} e^{-\alpha W_J}, \tag{2.234}$$

$$N = \frac{P}{Ah\nu} \quad T = \frac{P}{Ah\nu} \quad (1 - R - A), \tag{2.235}$$

and A is the fraction of light absorbed within the metal film and within any antireflection coating. The equation for $\overline{L}$ is complex but has simple limiting forms. For photon energies just above the band gap energy the creation of electron-hole pairs is fairly uniform through-

out the semiconductor and $\overline{L} \simeq L + W_J$, and the photocurrent begins to increase directly with α for increasing photon energies. At large photon energies essentially all absorption will be within the space charge region and the internal quantum efficiency should approach unity as $\overline{L} \to 1/\alpha$ and

$$J_{ph} \to -q\,N = -\frac{qP}{A\,h\nu}\,(1 - R - A). \qquad (2.236)$$

The short wavelength external efficiency should then be limited only by the fraction of incident light which gets into the semiconductor. With an antireflection coating as previously discussed, Schneider was able to obtain an external efficiency of 70% at a wavelength of 6328 A° with a Si detector [77].

The external or quantum efficiency of a Cr - GaAs metal semiconductor detector is shown in Figure 2.74 over the wavelength range from 3000 A° to 8000 A° [78]. This covers the visible and ultraviolet photon energy range from about 1.6 eV to about 5 eV. As can be seen from the data the efficiency varies by about a factor of three over this range of wavelengths. The efficiency is about 10^{-1} to 2×10^{-1} which is smaller than that which can be achieved with p-n junctions but the response is more constant over a broader wavelength range than can be achieved with p-n junction detectors.

The frequency response of the metal semiconductor detector is limited by the RC time constant of the device and the transit time of carriers across the depletion region. The carrier transit time is

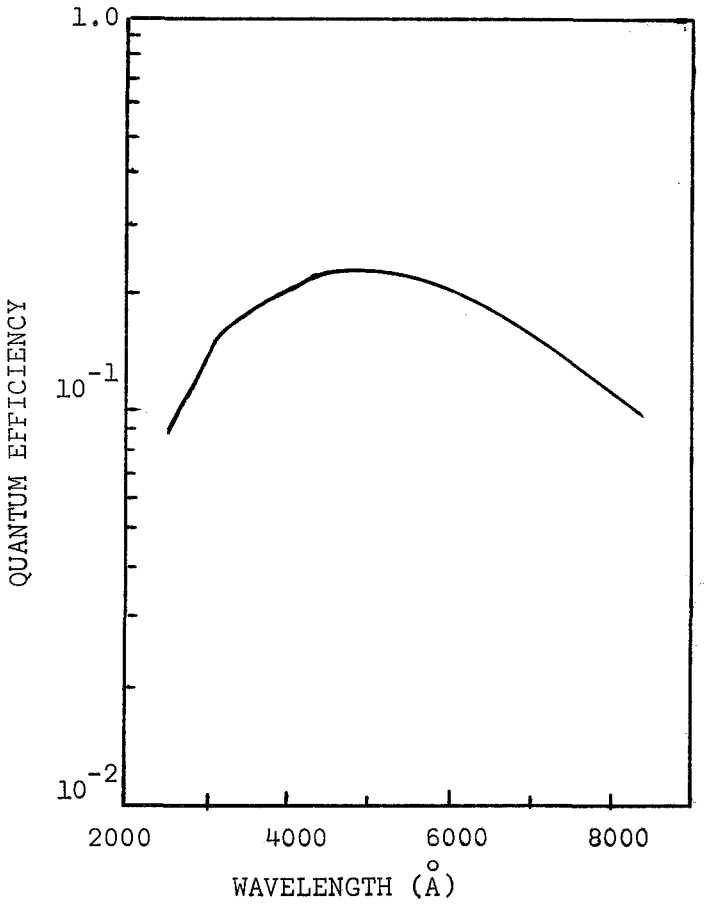


Figure 2.74. Efficiency of a Cr-GaAs metal semiconductor diode as a function of photon wavelength [78].

$$\tau \simeq \frac{W_J}{v_s}, \tag{2.237}$$

where v_s is the carrier saturation velocity which in on the order of 10^7 cm/sec for most semiconductors. For a depletion region width of 5μ the cutoff frequency neglecting the RC time constant would be

$$f_o \simeq \frac{v_s}{W_J} = 20 \text{ GHz}. \tag{2.238}$$

Schneider has measured a cutoff frequency of 22 GHz for a device with W_J = 5 μ in good agreement with the transit time estimation [77]. The high cutoff frequency of the detector as discussed above is one of the real advantages of the metal semiconductor detector. This frequency limit applies to the use of the detector in either the infrared, or visible spectral regions.

As with a p-n junction photodetector, the metal semiconductor device can be operated in the avalanche multiplication region to obtain an enhanced signal due to the avalanche multiplication. Operation is very similar to the p-n junction avalanche photodetector and similar limitations apply with regard to maximum signal multiplication and with respect to noise. Current gains of as much as 35 have been reported [9] for silicon devices with a 100 A° layer of PtSi metal film and at a wavelength of 6328 A°.

In summary then the metal semiconductor optical detectors appear to be especially useful as infrared detectors and are possible useful in the visible and ultraviolet spectral regions. By an appropriate choice of metal semiconductor combination the long wavelength response can be extended far into the infrared. Detectors with efficiencies of 1% or larger have been demonstrated at wavelengths out to 10 μ. These have also been achieved using Si and Ge for which well established technologies exist. The usefulness of metal semiconductor detectors in the visible and ultraviolet spectral regions arises from the fact that they are not limited by internal collection efficiency since all carriers are created within the depletion region. The efficiency is rather limited by the transmission into the semiconductor through the

metal film. By the use of thin films and antireflection coatings these sensors can be made to be competitive in efficiency with normal p-n junction sensors. Metal semiconductor detectors have very good frequency properties with cutoff frequencies of 22 GHz having been experimentally demonstrated. Although presently in the development stage the avalanche metal semiconductor detector appears to be especially promising as a high speed photodetector for visible and ultraviolet radiation.

2.4.3. Stress Sensors

There has been little reported work on the effects of mechanical stress on metal semiconductor devices. The terminal I-V characteristics of diodes at low forward and reverse bias are of the form

$$J = J_o \exp(-q\,\phi_{Bm}/E_o) \exp(q\,V_a/E_o)\,[1 - \exp(-q\,V_a/kT)], \quad (2.239)$$

where J_o is a constant which may depend slightly on the barrier heights if the current flow is diffusion limited as discussed in connection with Equation (q.210). In the above, E_o in general is temperature dependent and depends on the barrier hieght ϕ_{Bm}. The only simple case to discuss is the case of pure thermionic emission over the barrier where $E_o = kT$ and J_o is a constant. This case will be used to discuss the possible stress sensitivity of a metal semiconductor diode.

A large number of devices meet the above criterior and have

$$J \simeq J_o \exp(-q\,\phi_{Bm}/kT)\,[\exp(q\,V_a/kT) - 1]. \quad (2.240)$$

For this case the major stress dependent term is likely to be the

barrier height ϕ_{Bm}. Since ϕ_{Bm} is difficult to theoretically predict it is equally difficult to predict the stress dependence of the barrier height. According to the model of Cowley and Sze [9],

$$q\,\Delta\phi_{Bm} = (1 - C_2)\,\Delta E_g, \tag{2.240}$$

where C_2 is a constant discussed in connection with Equation (2.205) and ΔE_g is a band gap change. The above equation does not account for degenerate energy bands as are present in Ge and Si but at low stress levels this equation may give a rough estimation of the stress sensitivity. For Si, $(1 - C_2) = 0.73$ so this would predict a stress sensitive current for metal Si diodes of the form

$$J = J_s \exp(-q\Delta\phi_{Bm}/kT)\,[\exp(q\,V_a/kT) - 1],$$

$$= J_s \exp(-0.73\,\Delta E_g/kT)\,[\exp(q\,V_a/kT) - 1], \tag{2.241}$$

where J_s is the unstressed value of saturation current. This is slightly less sensitive than a p-n junction since at low stress levels the current in a p-n junction varies as $\exp(-\Delta E_g/kT)$. Thus based upon this simple model one would expect an exponential dependence of current on stress.

Rideout and Crowell have reported on the experimental stress sensitivity of Au - Potassium tatalate (KT_2O_3) diodes [79]. These diodes were stressed with a steel stylus of 0.25 mm radius and the resulting I-V characteristics are shown in Figure 2.75. These characteristics

are very similar to those obtained with stressed p-n junctions. They attributed the stress sensitivity to a barrier height reduction as discussed above and found

$$\Delta\phi_{Bm} \simeq (40 \times 10^{-12} \text{ eV. cm}^2/\text{dyne})\ P\ (0), \qquad (2.242)$$

where P(0) is the maximum stress on the diode. There do not appear to be any values of the stress dependence of band gap in KT_2O_3 with which to compare this number. This change is about a factor of 4 larger than the effective band gap reduction in Si at large stress levels. Potassium tantalate has a 3.5 eV band gap which is 3 times that of Si so the change Si.

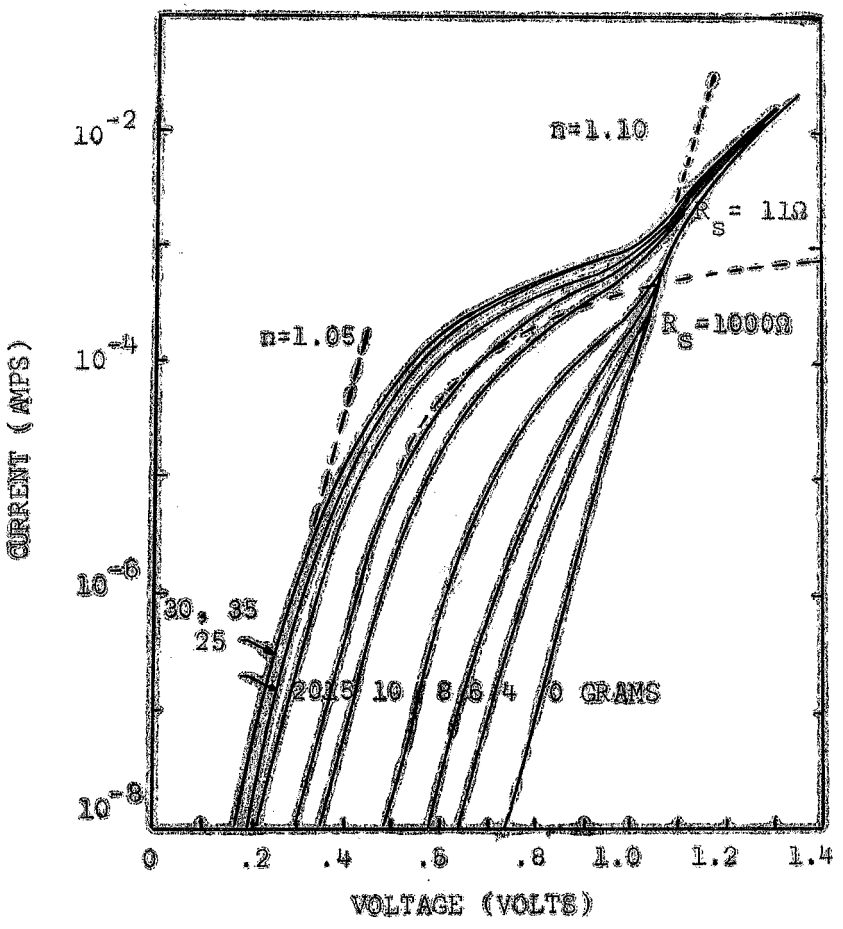


Figure 2.75. Forward-bias voltage-current characteristic of 0.25 mm radius Au-$KTaO_2$ diode for various forces. Steel stylus radius, 0.25 mm [79].

in band gap with stress might be expected to be larger in KT_2O_3 than in Si.

Stress effects on tungsten -n-type silicon diodes have also been investigated [80], [81]. As formed, these diodes have been reported to be relatively insensitive to stress, at least at stress levels below 10^8 dynes/cm^2. The data which has been reported on these diodes has been on diodes which have had their stress sensitivity increased by either γ-irradiation, by mechanical damage, or by gold doping. All of these techniques are known to introduce deep defect levels. Thus the results reported on these diodes probably involve the stress sensitivity of these defect levels. It was found that each of these techniques could be used to obtain diodes which were stress sensitive at 10^7 dynes/cm^2 which is two orders of magnitude lower stress than required for significant band gap changes. The observed I-V characteristics followed closely $I \alpha V^n$ with n between 1 and 2. The diodes studied were obviously not metal semiconductor diodes in the normal sense. The devices were found to be quite stress sensitive and under certain condtions the stress sensitivity was fairly temperature independent. These effects may be worthy of further study although they are not conventional metal semiconductor diodes.

The small amount of work which has been done on stress effects in metal-semiconductor diodes indicates that they are stress sensitive to some extent. This is to be expected from the stress dependence of the semiconductor energy band structure. The absence of a complete understanding of the magnitude of the metal semiconductor barrier height prevents a theoretical prediction of the magnitude of the stress effect.

The stress dependence of the barrier height is not likely, however, to vary more rapidly with stress than the variation of band gap within the semiconductor. Although no temperature studies have been reported for stress effects on diodes exhibiting ideal emission current the theory predicts similar temperature problems as for stressed p-n junction. Metal semiconductor stress sensors are likely to have the same problems as p-n junction stress sensors. One possible reason for studying them is that one may find a combination of material such as Au-KT_2O_3 which has a higher stress sensitivity than does the p-n junction sensor. The devices studied by Okamotom et. al. can not be described by conventional metal semiconductor theory and there still remains a question as to why the devices are stress sensitive. The stress sensitivity is more likely a bulk effect than a junction effect.

2.4.4. Other Sensors

The previous sections have discussed the areas which appear to be most promising for sensor applications of metal semiconductor devices. Other sensors are possible but they do not appear to offer any advantage over other junction devices. The emission current is exponentially related to temperature at forward bias as

$$J = J_o \exp[-q(\phi_{Bm} - V_a)/kT]. \qquad (2.243)$$

Thus a metal semiconductor device can in principle be used as a temperature sensor when operated at a constant current just as a p-n junction. As such a sensor however it does not appear to offer any major advantages over p-n junctions. Thus the well established technology for p-n

junctions makes then more attractive as temperature sensors.

No magnetic field effects appear to have been reported on metal semiconductor devices. Based upon the operating principles of the diodes they would not be expected to be very sensitive to magnetic fields.

2.5. Metal Insulator Semiconductor (MIS) Devices

The devices considered in previous sections (with the exception of junction FET's) have been minority carrier devices. The operation of these devices depends fundamentally on the properties of minority carriers within the semiconductor. The use of these devices as sensors depends in most cases on the fact that minority carrier densities are small and changes in minority carrier properties due to external stimuli are much larger percentage changes than the corresponding changes for majority carriers.

The present section is concerned with a different class of devices the metal insulator semiconductor devices which make use of majority carrier properties as opposed to minority carriers. The operation of MIS devices depends on the properties of a very thin conductive channel at the surface of a semiconductor. One of the advantages of MIS devices over other bulk effect sensors is the fact that a relatively small amount of total charge is involved in the channel. The typical channel charge in a Si MIS device is on the order of 10^{11}-10^{12} charges/cm^2. To have the same amount of charge a conventional bulk effect device of doping level $10^{15}/cm^3$ would have to have a thickness of 10^{-3} to 10^{-4} cm. Thus one of the unique features of MIS devices is their equivalence to very small bulk effect devices or sensors. This small size coupled with the fact that the properteis of MIS devices are sensitive to the properties of the insulating layer on the semiconductor give them interesting possible sensor applications.

The present state of sensor application of MIS devices can be classified as in the development stage. Several sensors have been pro-

posed and investigated but few of these ideas have been carried to the point of producing competitive sensors.

2.5.1. Basic Theory

This section reviews the basic operating principles of metal insulator semiconductor capacitors and devices. This includes devices made on single crystal substrates as well as thin film MIS devices. The basic cross sectional structure of a standard MIS device is shown in Figure 2.76. An n-type channel device is shown constructed on an p-type semiconductor. The important physical parts of the device are the semiconductor region to which ohmic source and drain contacts are made, the insulating layer, and the gate which controls the conductance or resistance of the channel. For an MIS capacitor instead of a transistor, the source and drain contacts are missing and the structure forms a two terminal capacitor between gate and substrate. The structure shown in Figure 2.76 is typical of an MIS device fabricated on silicon where the insulating layer is SiO_2. There are several different physical configurations for thin film MIS devices, one of which is shown in Figure 2.77. Again the important parts are the semiconductor with ohmic source and drain contacts, the insulating layer, and the metal gate. While there are minor differences between thin film MIS devices and devices made on bulk semiconductor materials the fundamental operating principles and equations describing their operation are very similar. Consequently both types of devices are considered in this section.

An understanding of the operation of MIS devices requires an understanding of surface space charge regions within a semiconductor.

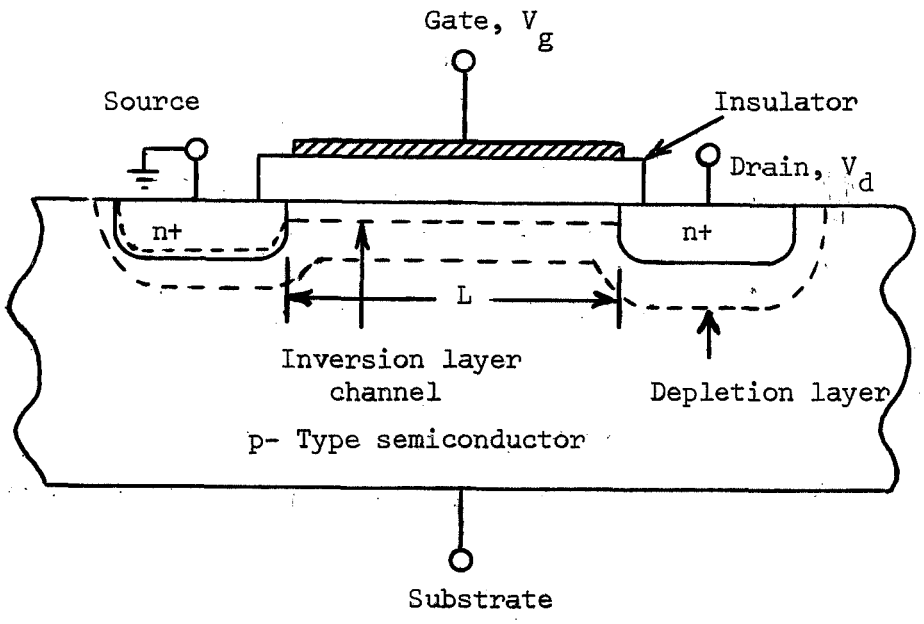


Figure 2.76. Standard MIS device structure

The types of surface space change layers which can exist in a p-type semicinductor are illustrated in Figure 2.78. First when a negative potential exists on the gate with respect to the semiconductor, positive charges are attracted toward the surface and form a thin accumulation layer of holes near the surface. A positive potential between gate and semiconductor forces mobile holes away from the semiconductor surface forming a depletion region as shown in Figure 2.78 (b). The depletion region has a net negative charge due to the bound impurity atoms. At the same time that holes are forced from the surface, electrons are attracted toward the surface and for sufficiently large positive voltages on the gate a thin inversion layer or channel of electrons forms at the surface as shown in Figure 2.78 (c). The existence of the inversion layer depends on the insulator being a very high resistivity

region. If the resistance of the insulator is too low the electrons will simply flow through the insulator and no inversion layer can be formed.

The behavior of the surface space charge layers are determined by Boltzmann statistics and Poisson's equations. The resulting general solutions are quite complicated but simplifying approximations are possible for the description of MIS transistor operation. One of the important quantities associated with the surface space charge layers is the width of the depletion layer at the onset of the formation of the conductive channel. This depends on the substrate doping level N_A as

$$W_{dm} = \left[\frac{4\varepsilon kT}{q^2 N_A} \ln(N_A/n_i)\right]^{1/2}. \qquad (2.244)$$

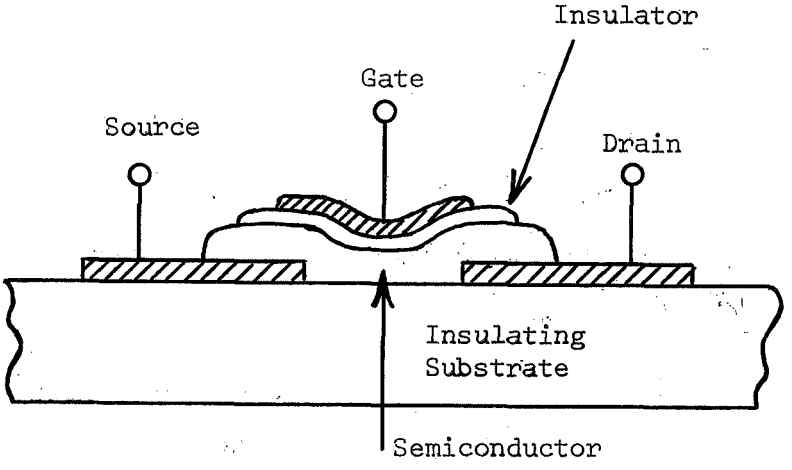
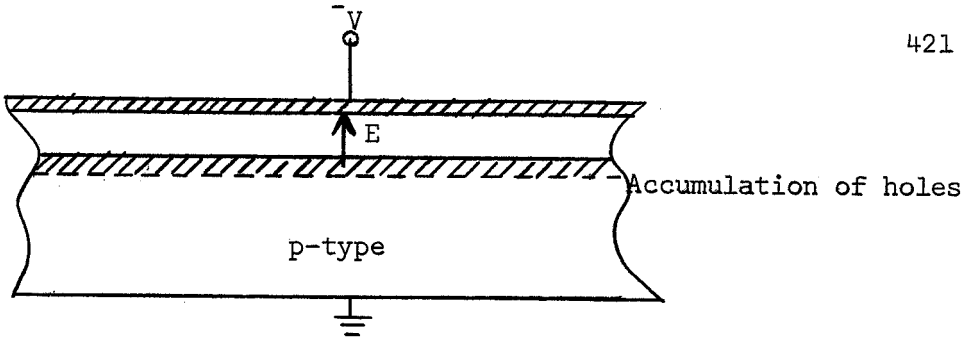
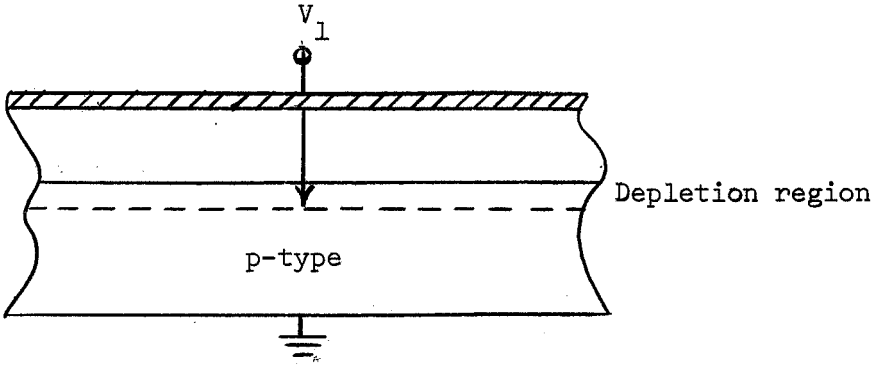


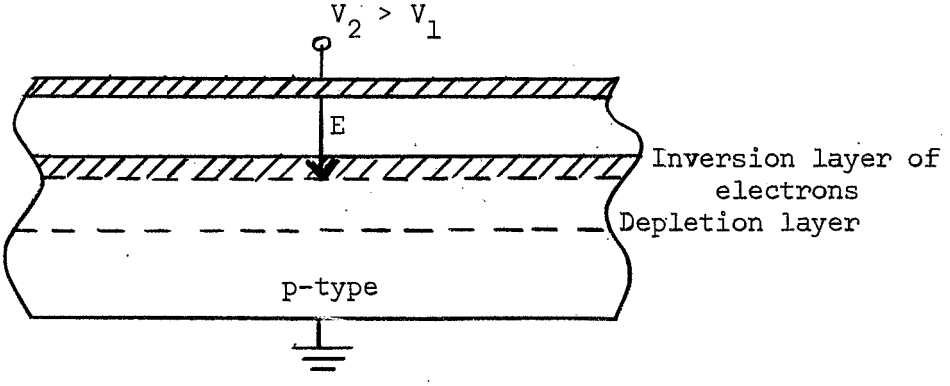
Figure 2.77. One possible variation of MIS geometry for thin film device.



a. Accumulation space charge layer



b. Depletion space charge layer



c. Depletion and inversion space charge layers

Figure 2.78. Types of surface space charge regions for an MIS device with an p-type semiconductor.

This is the maximum width of a depletion region which can exist in equilibrium at a semiconductor surface. It is also found that the width of any inversion layer is considerably less than the width of the depletion layer so the above equation represents a good approximation to the total width of the equilibrium surface space charge layer. The onset of strong inversion or channel formation occurs when the electron density becomes equal to the impurity doping level of the semiconductor. Since the potential across the depletion region is related to the width by the expression

$$V_d = \frac{q\,N_A}{2\varepsilon}\,W_d^2 \quad , \tag{2.245}$$

the maximum potential across the depletion region in equilibrium is

$$V_{do} = \frac{2kT}{q}\,\ln\,(N_A/n_i). \tag{2.246}$$

This is also the potential at the surface of the semiconductor just at the onset of channel formation.

The simplest MIS device is a two terminal voltage dependent capacitor formed between the semiconductor and the metal gate. A qualatiative understanding of the capacitance of such a device can be obtained with the aid of Figure 2.78. In general the capacitance of the device can be considered as the capacitance of two capacitors in series, the insulator capacitance C_I and a capacitance associated with the semiconductor C_s. Then

$$C = \frac{C_I\,C_s}{C_s + C_1}. \tag{2.247}$$

At negative gate voltages the device acts essentially like a parallel plate capacitor with the insulating layer as the dielectric. The semiconductor capacitance is large. The measured capacitance per unit area is then essentially the insulator capacitance given by the expression

$$C_I = \frac{\varepsilon_I}{W_I}, \tag{2.248}$$

where W_I is the insulator thickness. As the potential between metal and semiconductor is increased, the surface accumulation layer is reduced and at some point the field within the semiconductor is zero. Ideally this occurs at zero applied volts. The extent of any small space charge variation within the semiconductor then decays exponentially with distance with the Debye length giving the characteristic decay length. The semiconductor then acts as a capacitor with width equal to the Debye length and the terminal capacitance is

$$C_{FB} = \frac{C_I}{1 + \frac{\varepsilon_I}{\varepsilon_s}\frac{L_D}{W_I}}, \tag{2.249}$$

where

$$L_D = \sqrt{\varepsilon_s \, kT/q^2 N_A} \tag{2.250}$$

and ε_s is the semiconductor dielectric constant. As the terminal voltage goes positive the depletion region forms within the semiconductor and the semiconductor acts like a capacitor of thickness equal to the

width of the depletion region.

When the inversion layer forms, the terminal capacitance is not a unique quantity but becomes dependent on the measurement frequency. At high frequencies the semiconductor acts as a parallel plate capacitor with width W_{dm} which is the width of the depletion region. In this case

$$(C_{min})_{h.f.} = C'_{min} = \frac{C_I}{1 + \frac{\varepsilon_I}{\varepsilon_s}\frac{W_{dm}}{W_I}} , \qquad (2.251)$$

where W_{dm} is given by Equation (2.244). If the capacitance is measured at very low frequencies a different effect occurs. At very low frequency charges are able to flow across the depletion region and no part of the ac measuring potential can be supported across the depletion region. In this case at large positive bias and at low frequencies one again observes simply the oxide capacitance. The ideal C-V curves are shown in Figure 2.79. Both the low frequency and high frequency curves are shown as well as a third curve which is a non-equilibrium or transient capacitance curve. This is the initial capacitance which is observed if the device voltage is changed very rapidly positively. It takes a finite time for the inversion layer to form and before this forms only a depletion layer exists within the semiconductor. The minimum capacitance C_{min} occurs approximately at the point where the heavy inversion or channel begins to form.

In actual devices the C-V properties are modified somewhat from the ideal curves shown in Figure 2.79 because of two major effects. These are interface states and insulator charges. These are present

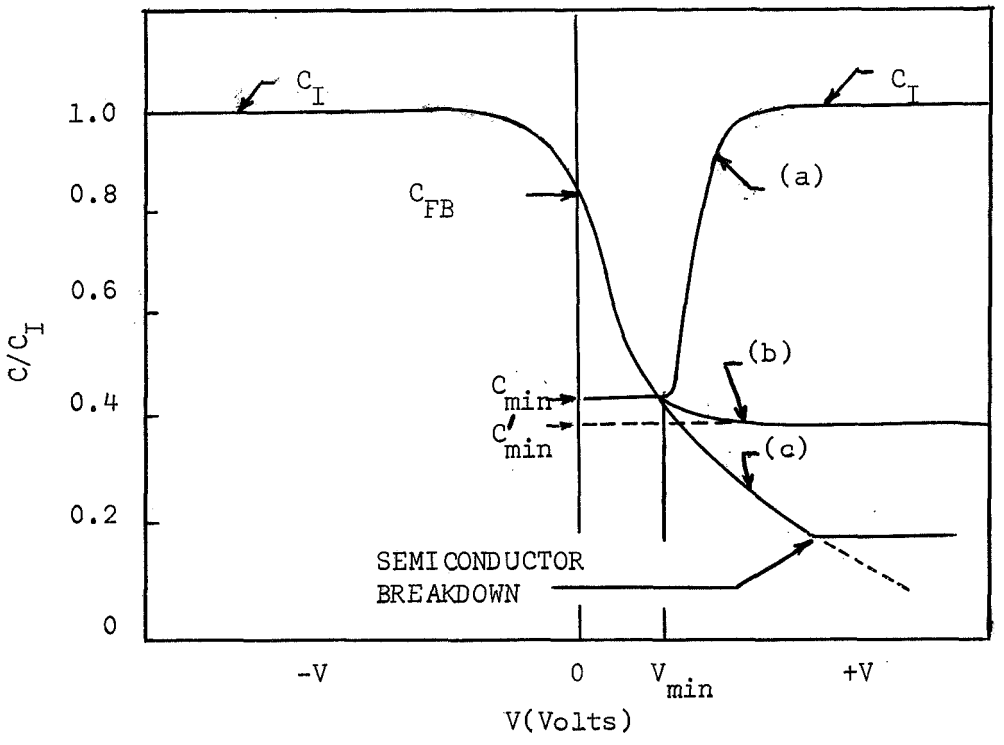


Figure 2.79. Ideal MIS capacitance - voltage curves. (a) Low frequency, (b) high frequency, (c) monequilibrium or transient case.

to some extent in all practical devices. Interface states are energy levels located at the insulator semiconductor interface within the band gap of the semiconductor. These have two effects on the C-V characteristic. First the charge within the interface states causes a parallel shift of the C-V curve along the voltage axis. Second if the C-V measurement frequency is low enough the surface state charge can respond to the measurement frequency and give an additional capacitance term which causes a distortion of the ideal C-V curve. The insulator charges are charged species either ions or electrons

within the insulator. These also shift the C-V curve along the voltage axis, and if they are mobile, lead to a time and temperature dependent instability of the C-V curve. The total charge within the insulator can also charge with time and bias because of the injection of charge into the insulator from either the semiconductor or metal. This effect is pronounced for insulator films with small bandgaps.

The least harmful of all the above effects is a shift of the flat band voltage from zero by the amount

$$V_{FB} = \phi_{ms} - \frac{Q_{ss}}{C_I} - \frac{1}{C_I} \int_o^{W_I} \frac{x}{W_I} \rho_I (x)\, dx, \qquad (2.252)$$

where Q_{ss} is the surface state charge, ρ_I is the insulator charge density and ϕ_{ms} is any difference in work function between metal and semiconductor. If V_{FB} is time, temperature, and bias independent the effects of V_{FB} not being zero are not too important provided it is not excessively large.

It has proven to be very difficult to find insulator-semiconductor combinations which have very low insulator charge densities and surface state charges and which have stable values of these parameters under bias. The only really successful materials combination with respect to these parameters has been thermally grown oxide on silicon. With this system it is possible to now produce stable MIS devices with flat band voltages approaching the ideal. Part of this success is due to the very wide band gap of SiO_2 which reduces charge injection effects to a negligable level. Thin film devices have been especially plagued with insulator and interface stability problems.

The MIS transistor differs from the MIS capacitor in the addition of two ohmic contacts which are used to make contact to the surface inversion layer. The inversion layer forms a conductive path between the source and drain ohmic contacts which can be controlled by the bias on the metal or gate electrode. Useful operation of the MIS device is then always within the inversion region where both a depletion and inversion layer exists within the semiconductor. The conductive charge per unit area within the channel can be written approximately as

$$Q = C_I(V_g - V - V_T), \tag{2.253}$$

where V is the channel potential and V_T is a threshold voltage required to form the depletion region and reach the onset of channel formation. Contributions to V_T come from four major sources; work function difference between metal and semiconductor, insulator charge, surface state charge, and depletion region charge. Expressed mathematically

$$V_T = V_{FB} + V_{do}$$

$$= \phi_{ms} - \frac{Q_{ss}}{C_I} - \frac{1}{C_I}\int_o^{W_I} \frac{x}{W_I}\rho_I(x)\,dx + V_{do}. \tag{2.253}$$

The threshold voltage V_T is the gate-to-channel or gate-to-source voltage which must exist in order to establish the inversion layer.

When a drain-to-source potential is applied as shown in Figure 2.80 current flows along the channel and a potential drop exists along the channel making the gate-to-channel voltage $V_g - V$ in Equation (2.253) a function of distance along the channel. The channel charge

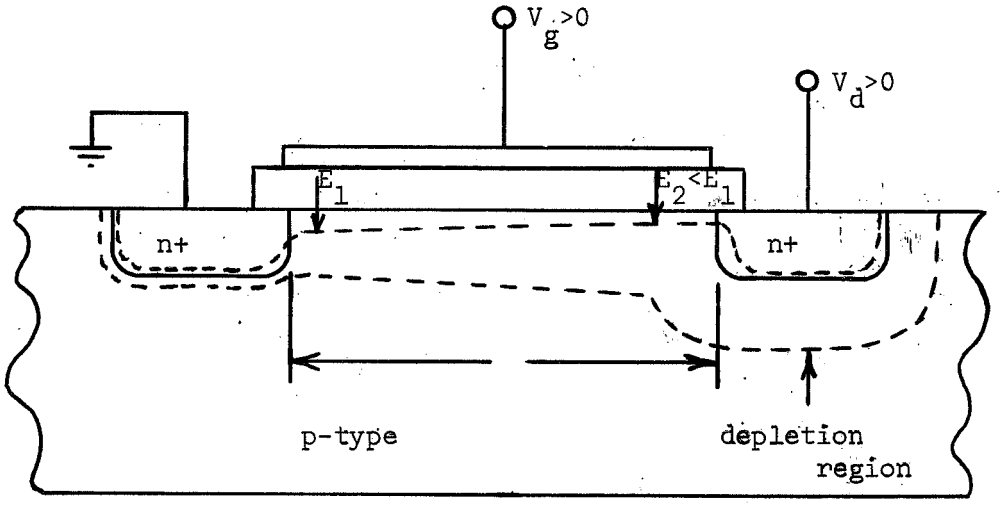


Figure 2.80. Sketch of MIS device with an applied drain-to-source voltage.

density then becomes nonuniform along the length of the channel. A depletion region seperates the conductive channel from the substrate so the total current is constant along the channel. This requirement combined with Ohms Law for current flow along the channel gives the equation

$$I_d = \mu_n C_I Z(V_g - V - V_T) \frac{dV}{dx} , \qquad (2.254)$$

where μ_n is the average carrier mobility within the channel and Z is the width of the device into the page in Figure 2.80. The above equation neglects the increased charge within the depletion region in going from source-to-drain [9]. Including this adds another voltage dependent term to the equation which greatly complicates the mathematics without adding any important new physical insight into the operation of MIS devices. For devices on very lightly doped substrates the above

equation is approximately correct.

Assuming that the source is at zero potential, Equation (2.254) can be integrated between source and drain to give

$$I_d = \mu_n C_I(Z/L) [(V_g - V_T) V_d - (1/2) V_d^2]. \tag{2.255}$$

The above equation must be restricted to the region of $V_d < V_g - V_T$ since for larger drain voltages the channel is reduced to zero near the drain and Equation (2.253) no longer describes the channel charge. For larger drain voltages the drain current becomes saturated at its value when $V_d = V_{ds} = V_g - V_T$ so a complete theoretical description of the MIS device is

$$I_d = \begin{cases} \mu_n C_I(Z/L)[(V_g - V_T)V_d - (1/2) V_d^2] \\ \qquad \text{for } V_d < V_{ds} = V_g - V_T \\ (1/2) \mu_n C_I(Z/L)(V_g - V_T)^2 \\ \qquad \text{for } V_d > V_{ds}. \end{cases} \tag{2.256}$$

The theoretical current voltage characteristics are shown in Figure 2.81. The region of operation with $V_d > V_{ds}$ is the current saturation region. Actual MIS devices show some finite drain resistance within the satureation region instead of the ideal saturation of the theoretical model.

As far as possible sensor applications are concerned there are

several important features of the operation of MIS devices which should be noted. First the MIS field effect ransistor is basically a majority carrier device. The terminal properties depend directly on the carrier mobility, the insulator capacitance and on the threshold voltage. The insulator capacitance and threshold voltage depend on the properties of the insulator while the mobility is basically a property of the semiconductor. Thus MIS sensors can be built around effects which influence either the semiconductor or insulator properties. Both types of sensors have been

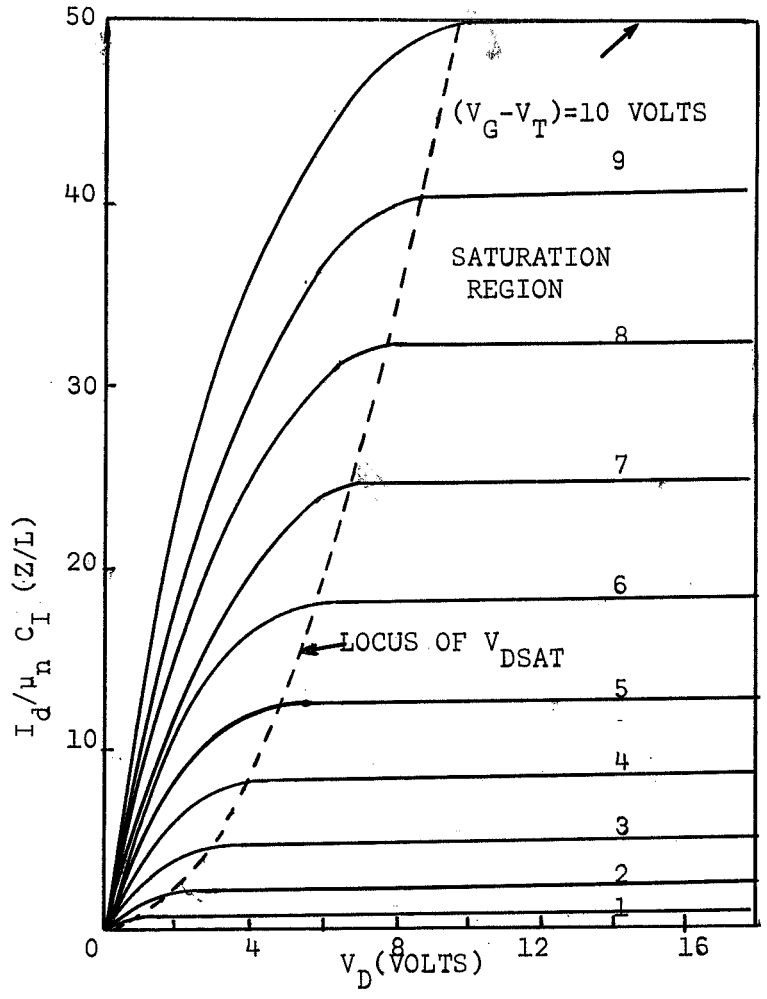


Figure 2.81. Theoretical characteristics of ideal MIS field effect transistor.

proposed and investigated but the most interesting MIS sensors appear to be those using changes in the insulator properties due to external forces which are to be detected.

2.5.2. Optical Sensors

There appears to have been little investigation of the use of MIS devices as optical sensors. Practical insulators used in MIS devices show little absorption within the visible region, so useful sensors would make use of the absorption within the semiconductor. The problem of getting light into the semiconductor with the MIS device is similar to the problem with the metal semiconductor. The metal must be thin and antireflection coatings used to achieve high efficiencies in getting light into the semiconductor. Photons which are absorbed within the semiconductor to a depth equal to the space charge region plus one carrier diffusion length will be collected by the space charge region and constitute a photocurrent across the surface space charge region. For a simple MIS capacitor this will modify the surface potential and consequently change the C-V characteristic. The experimentally observed effect of light on the C-V characteristics of a silicon device is shown in Figure 2.82. As seen in the figure the major effect occurs in the inversion region where the measured capacitance depends on the light intensity. Increasing light intensity has somewhat the same effect on the C-V characteristic as decreasing the C-V measurement frequency. With light the low frequency type of C-V curve can be observed at higher measurement frequencies than is possible without the light. This is due to the added carrier generation in the depletion region which effectively lowers the impedance of the depletion region and prevents

the ac measurement frequency from appearing across the depletion layer.

As the data of Figure 2.82 indicates the inversion region capacitance could be used as a light sensor. To obtain large changes in capacitance, very large levels of illumination are required. This may be partly due to reflection and absorption at the metal but in general the C-V characteristics are not expected to be very sensitive to light. Consequently the use of the MIS capacitance does not appear to be a very useful light sensor.

Consider next the effect of light on the properties of an MIS transistor such as shown in Figure 2.80. Photon generated carriers within the depletion region and within a diffusion length of the depletion region will cross the depletion region and add to the channel current. At very low channel currents this may be a significant fraction

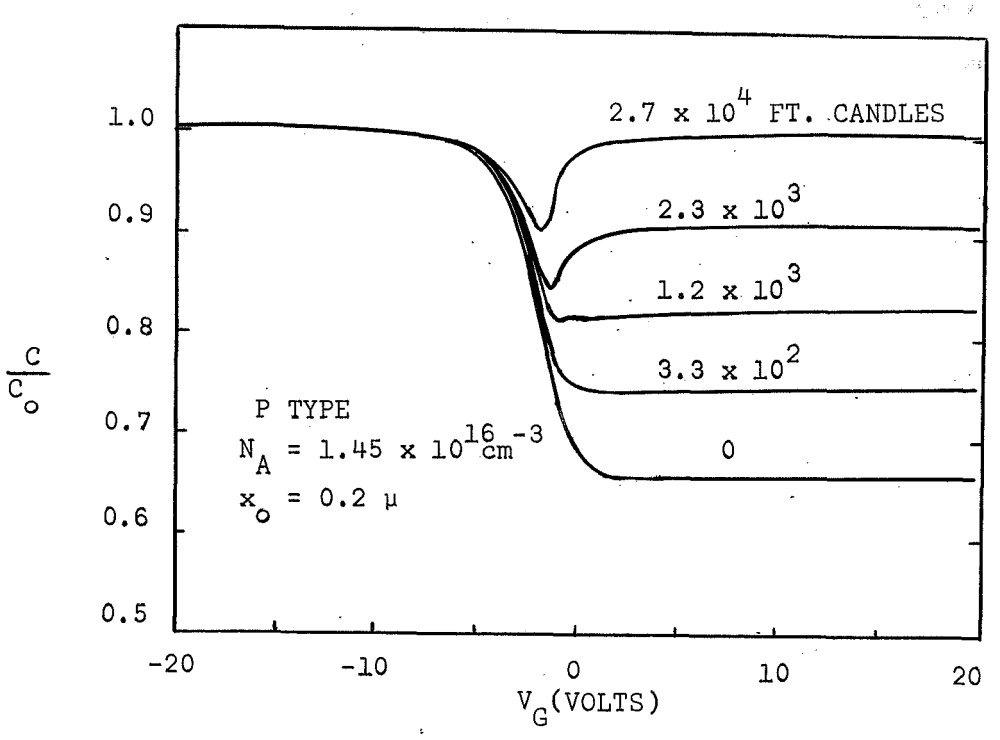


Figure 2.82. Effect of illumination on C-V characteristics of silicon MIS device [82].

of the total current and may be detected. The collected photocurrent constitutes a distributed photocurrent along the channel in much the same manner as for a junction FET as discussed in Section 2.3.3. Including the distributed photocurrent in the device equation leads to the same equation as Equation (2.256) with I_d replaced by $I_d - I_{ph}/2$, where I_{ph} is the total depletion region photocurrent [71]. As a photodetector then, the MIS device is equivalent to the circuit shown in Figure 2.83. The depletion region photocurrent appears as a current source between the drain and substrate of an otherwise ideal MIS device. The magnitude and wavelength response of I_{ph} is limited in much the same manner as a metal semiconductor photodiode because of reflection and absorption within the metal gate.

Because of the equivalence seen in Figure 2.83 to an ideal MIS device and a seperate photodiode, the MIS photodetector does not appear to be extremely useful as an optical sensor. The device has one possible advantage. If the MIS device is operated at constant drain current the MIS photodetector can give fairly large drain voltage changes because of the high output impedance of the device in saturation. This is essentially using the inherent voltage gain of the MIS device coupled to a photosensitive diode and similar response can be obtained with seperate components. However the small size obtained by combining both sensor and active gain device into one sensor might be of some advantage in certain applications such as large sensor arrays. In certain types of devices it should also be possible to eliminate the metal gate and eliminate the reflection and absorption present in this metal. This can be done on n-channel devices with a large enough

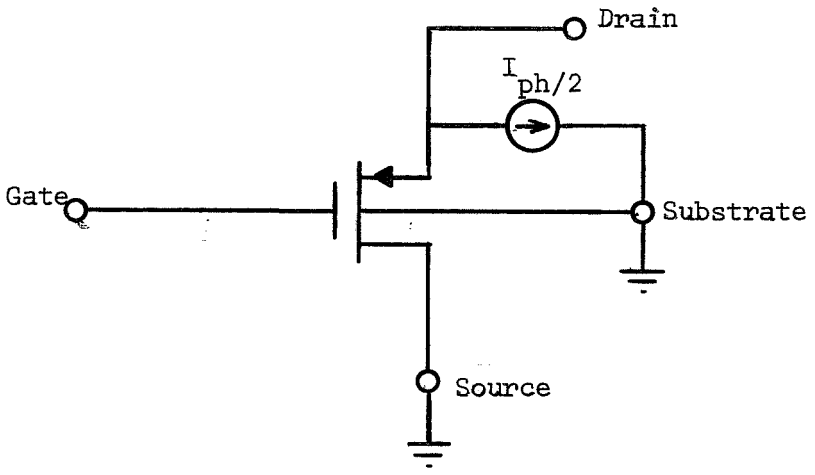


Figure 2.83. Equivalent circuit of MIS transistor used as a photodetector.

positive oxide charge to establish the channel in the equilibrium case. In this case though, the depletion and inversion layer acts essentially as a p-n junction diode as far as the photocurrent is concerned and such a device appears to offer little if any advantages over a p-n junction photodiode.

2.5.3 Stress Sensor

The area of stress sensors appears to be one of the most promising areas for sensor applications of MIS devices. In terms of the terminal device equations (Equation (2.256)) stress can have an influence through two factors. First I_d depends directly on the mobility μ_n and the mobility is known to be stress dependent from piezoresistance. Second stress can change the threshold voltage V_T. This is especially important in devices where either the insulator or semiconductor is piezo-

electric. These effects will be considered seperately in the following discussion although they may both be present in MIS devices.

The piezoresistance effect or stress dependent mobility effect in MIS devices differs in some important aspects from bulk piezoresistance. The most important difference arises from the fact that the surface inversion layer is basically a two dimensional conductive layer. Current flow is restricted to the plan of the inversion layer as opposed to normal three-dimensional conduction properties. The most general form of the two-dimensional resistivity of such a layer is

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix} , \tag{2.257}$$

where the resistivity matrix must conform to the symmetry of the inversion layer plane. In a similar manner one can write a two dimensional conductivity or mobility matrix as

$$\mu = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix} , \tag{2.258}$$

assuming that the anisotropic conductivity arises from an anisotropic mobility.

Piezoresistance coefficients can also be defined for an inversion layer and the most general form of the piezoresistance tensor is [83]

$$\Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} & \Pi_{14} \\ \Pi_{21} & \Pi_{22} & \Pi_{24} \\ \Pi_{41} & \Pi_{42} & \Pi_{44} \end{pmatrix} . \tag{2.259}$$

For a two dimenstional inversion layer Π is a 3 x 3 matrix as opposed to the 6 x 6 matrix for a bulk semiconductor. These coefficients have the usual meaning as discussed in Section 1 of this report. Measurements have been reported on the piezoresistance coefficients of silicon p-type inversion layers [84]. Results of these measurements are shown in Table 2.2 along with the corresponding Π coefficients of bulk p-type silicon. The plan in each case refers to the plane of the inversion layer. While there are differences between the piezoresistance of the bulk and an inversion layer, the piezoresistance coefficients are of the same order of magnitude as for bulk p-type material. As an example for a (111) inversion plane with both stress and current flow along the same direction it is found that

$$\frac{\Delta\rho_{111}}{\rho} = (44.7 \times 10^{-12} \text{cm}^2/\text{dyne})\ \sigma_{111}. \tag{2.260}$$

As a function of strain in the <111> direction this gives

$$\frac{\Delta\rho_{111}}{\rho} = 84.0\ \varepsilon_{111}. \tag{2.261}$$

For this stress condition the current gauge factor for an MIS transistor is

$$\frac{1}{I_d}\ \frac{dI_d}{d\varepsilon} = \frac{1}{\mu}\ \frac{d\mu}{d\varepsilon} = -\ \frac{1}{\rho}\ \frac{d\rho}{d\varepsilon} = -\ 84.0 \tag{2.262}$$

Table 2.2. Piezoresistance Coefficients for Silicon p-type inversion layers (10^{-12} cm^2/dyne)

Plane	Inversion Layer	Bulk
(111)	$\begin{pmatrix} 44.7 & -15.3 & 0 \\ -15.3 & 44.7 & 0 \\ 0 & 0 & 60.00 \end{pmatrix}$	$\begin{pmatrix} 71.8 & -66.3 & 0 \\ -66.3 & 71.8 & 0 \\ 0 & 0 & 138.1 \end{pmatrix}$
(100)	$\begin{pmatrix} -1.0 & 23.8 & 0 \\ 23.8 & -1.0 & 0 \\ 0 & 0 & 127.8 \end{pmatrix}$	$\begin{pmatrix} 6.6 & -1.1 & 0 \\ -1.1 & 6.6 & 0 \\ 0 & 0 & 138.1 \end{pmatrix}$
(110)	$\begin{pmatrix} -23.8 & 38.2 & 0 \\ -14.3 & 58.0 & 0 \\ 0 & 0 & 67.7 \end{pmatrix}$	$\begin{pmatrix} 6.6 & -1.1 & 0 \\ -1.1 & 71.8 & 0 \\ 0 & 0 & 183.1 \end{pmatrix}$

This is similar to the resistance gauge factor of piezoresistance for bulk silicon. Since the MIS transistor has a high output impedence in saturation the voltage gauge factor can be quite high limited only by the output impedance of the transistor. The voltage gauge factor is

$$\frac{1}{V_d}\ \frac{dV_d}{d\varepsilon} = -\ \frac{r_d}{V_d}\ \frac{dI_d}{d\varepsilon} = \frac{I_d\, r_d}{V_d}\ \frac{1}{\rho}\ \frac{d\rho}{d\varepsilon}\ . \qquad (2.263)$$

As a typical example consider the case of

$I_d = 10^{-3}$ amp, $V_d = 5$ volts, and $r_d = 10^4\Omega$. Then

$$\frac{1}{V_d}\ \frac{dV_d}{d\varepsilon} = 50\ \frac{1}{\rho}\ \frac{d\rho}{d\varepsilon}\ . \qquad (2.264)$$

Thus the voltage gauge factor in this example is 50 times the current gauge factor. This is just another example of the use of the inherent gain present in an active device to increase the voltage sensitivity.

No measurements similar to those reported for p-type inversion layers appear to have been reported for n-type inversion layers. The behavior of n-type layers would not be expected however to be very different and the piezoresistance would be expected to be on the same order of magnitude as for bulk piezoresistance in n-type material.

In summary then concerning the effect of stress on carrier mobility in MIS devices it can be stated that the current gauge factor has approximately the same magnitude as the gauge factor of bulk piezoresistance. The orientation dependence is somewhat different from that of the bulk effect due to the two-dimensional nature of the inversion layer. The

voltage gauge factor of an MIS device can be larger than that of bulk piezoresistance but this arises from the gain present in the MIS transistor.

The remaining stress effect to be considered is the stress dependence of threshold voltage V_T. In the current saturation region, a change in threshold voltage of amount ΔV_T results in a change in drain current of

$$\Delta I_d = - \mu_n C_I (Z/L)(V_g - V_T) \Delta V_T. \tag{2.265}$$

In terms of a current gauge factor this is

$$\frac{1}{I_d} \frac{dI_d}{d\varepsilon} = - \frac{2}{(V_g - V_T)} \frac{dV_T}{d\varepsilon} . \tag{2.266}$$

One straightforward way in which V_T can be changed by stress is through the use of a piezoelectric material as either the insulator or semiconductor. When stressed, a piezoelectric material forms a surface charge density equal to the normal component of the electric displacement induced at the surface. If the insulator of an MIS device is piezoelectric a surface charge layer forms under stress just inside the insulator at the semiconductor insulator interface. This is just the normal piezoelectric effect but because of the presence of the semiconductor under the insulator this will induce a charge within the semiconductor. The peizoelectric effect is then equivalent to a change in interface surface charge density and this changes V_T directly. If the semiconductor is piezoelectric, stress induces

a surface charge at the interface just inside the semiconductor, again giving rise to a stress dependent interface surface charge density. These two possible effects are illustrated in Figure 2.84. At first it would appear that these two cases are completely equivalent and that one type of sensor should be as good as the other. There is, however, one important difference. The surface charge within the insulator can remain uncompensated only for times short compared with the dielectric relaxation time of the insulator. This can be fairly long but the piezoelectric insulator case can not exhibit a dc response just as normal piezoelectric materials do not exhibit dc responses. In the piezoelectric semiconductor case the compensating charge is the channel charge and this is the detected stress induced charge. Thus the piezoelectric semiconductor case will respond to dc stress assuming of course that stable piezoelectric MIS transistors can be fabricated.

Thin film CdS and CdSe MIS transistors meet the requirements of being piezoelectric semiconductors and these are also the most extensively investigated types of MIS thin film devices. MIS transistors composed of these materials with SiO as the insulator have been experimentally shown to be stress sensitive through the piezoelectric effect [84],[85]. The stress induced surface charge density can be written in general as [84]

$$\Delta Q_{ss} = [e]\,[\varepsilon]\cdot[n], \tag{2.267}$$

where [e] is the semiconductor piezoelectric matrix, [ε] is the strain

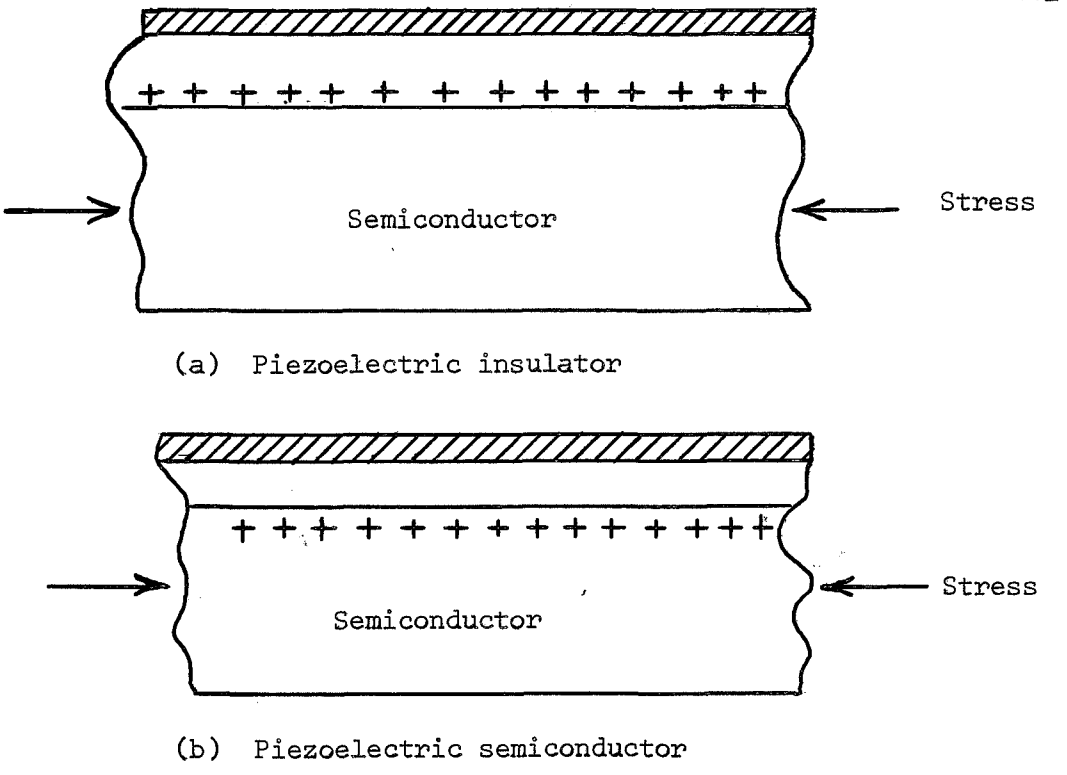


Figure 2.84. Stress dependent surface charge density of piezoelectric MIS device.

matrix, and [n] is a unit normal to the surface. If the piezoelectric material is assumed to be CdS or CdSe with the c-axis perpendicular to the surface as normally occurs in vacuum evaporated thin films of these materials and the applied stress is along a crystal axis, then the above matrix equation reduces to

$$\Delta\phi_{ss} = e_{31}\,\varepsilon = d_{31}C_{11}\varepsilon \tag{2.268}$$

where ε is either ε_{11} or ε_{22}, the strain along the crystal axis. This gives rise to a threshold voltage change of

$$\Delta V_T = \frac{\Delta Q_{ss}}{C_I} = \frac{d_{31}\, C_{11}}{C_I}\ \varepsilon. \tag{2.268}$$

The current gauge factor is then

$$\frac{1}{I_d}\ \frac{dI_d}{d\varepsilon} = -\ \frac{2}{(V_g - V_T)}\ \ \frac{d_{31}\, C_{11}}{C_I}\ . \tag{2.269}$$

For a CdS device with a 1000 A° thick oxide insulating layer and with $(V_g - V_T) = 1$ volt the theoretical sensitivity can be evaluated as

$$\frac{1}{I_d}\ \frac{dI_d}{d\varepsilon}\ \simeq 280. \tag{2.270}$$

This in only slightly larger than the piezoresistance gauge facotr of silicon. CdSe has a smaller piezoelectric coefficient and should thus exhibit a smaller gauge factor. Devices with properties close to the theoretical values have been experimentally reported [84][85]. From Equation (2.269) it would appear that the gauge factor becomes infinitely large as the channel pinch off conduction is approached or as $V_g \rightarrow V_T$. Other factors not considered however limit the response and the largest reported gauge factors for CdS devices are 400 to 800 [86].

From the general device equation for MIS devices (Equation (2.256)) it is seen that a change in V_T is exactly equivalent as far as terminal device characteristics are concerned to a change in device gate voltage. Further the change in threshold voltage of Equation

(2.268) (or more fundamentally Equation (2.267) divided by C_I) is simply the piezoelectric voltage induced across a piezoelectric material of thickness equal to the insulator thickness. Thus the performance of the piezoelectric MIS transistor is equivalent to a normal piezoelectric stress sensor followed by an MIS transistor. This equivalence is shown in Figure 2.85. The importance of the piezoelectric MIS transistor must then be considered in the light of the fact that similar resulsts can be ahieved by a conventional piezoelectric sensor followed by an MIS transistor amplifier. The MIS piezoelectric transistor has one advantage. It can respond to dc stress if the semiconductor is piezoelectric while the conventional piezoelectric sensor can not. If dc response is then the overiding consideration, the piezoelectric transistor has a very definite advantage. However such devices which must use semicondictors other than silicon, are plagued by instability problems which have so far not been satisfactorily solved. It is difficult to find materials which are good piezoelectric materials and good semiconductors so the MIS piezoelectric transistor involves compromises with both features. With seperate sensor and amplifier functions as in the equivalent circuit of Figure 2.85 each part can be chosen for most optimum operation.

The voltage sensitivity of the MIS piezoelectric transistor can be large-for example an order of magnitude or more larger than the current sensitivity. This arises from the inherent gain present in the active MIS device. Except for giving a dc response, the piezoelectric MIS transistor appears to have no major advantage and several disadvantages over a conventional piezoelectric sensor followed by an MIS amplifier.

2.5.4 Magnetic Field Sensors

The MIS device characteristics are sensitive to magnetic fields basically through the Hall effect. The most extensively studied device has been the silicon MIS device with SiO_2 as the insulating layer. The discussion in this section is therefore in terms of this type of device. Similar effects should be observable in other MIS devices. To observe the Hall effect with an MIS transistor requires the addition of two Hall electodes placed perpendicular to the drain-to-source current flow. The structure of an MIS Hall effect device is shown in Figure 2.86. The magnetic field is assumed to be applied perpendicular to the plane of the conductive channel. From the figure it is seen that the arrangement is equivalent to a standard Hall type sample with the conduction region being the channel. Current flow along the channel from drain-to-source in the presence of a magnetic field produces a Hall

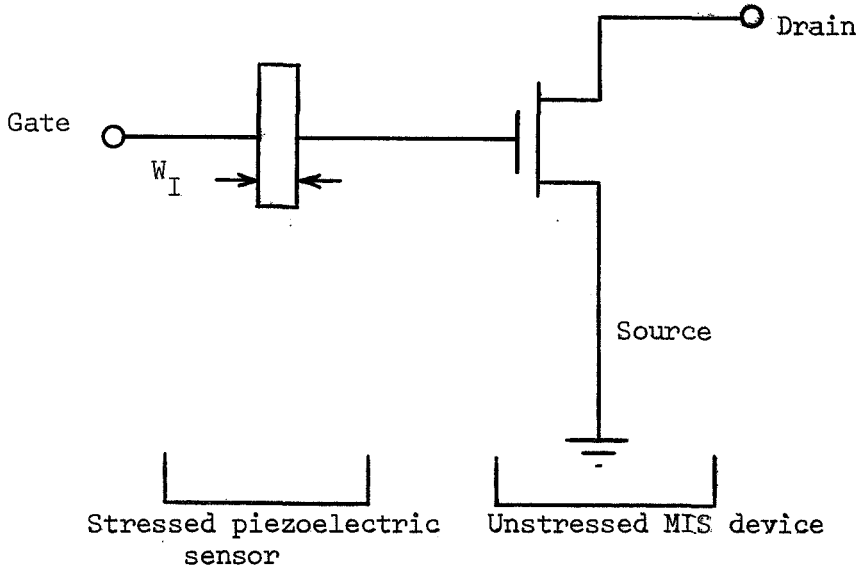
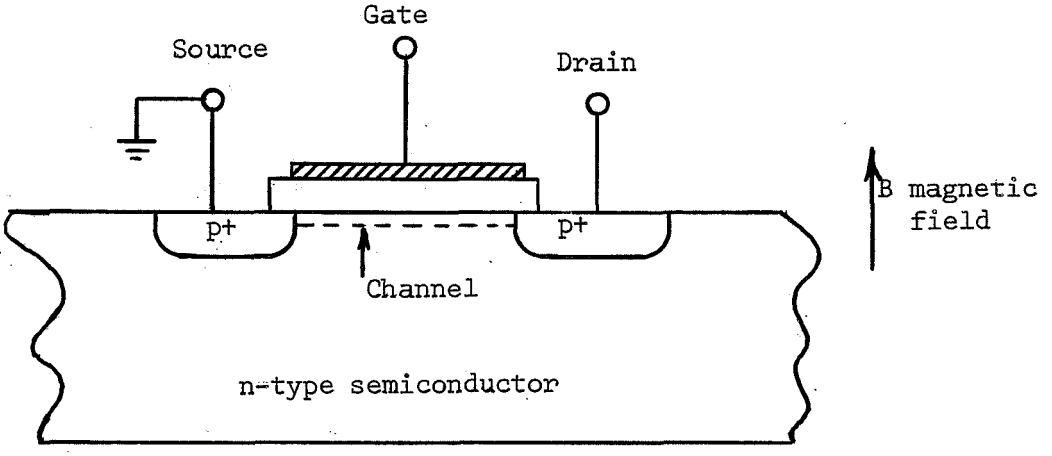
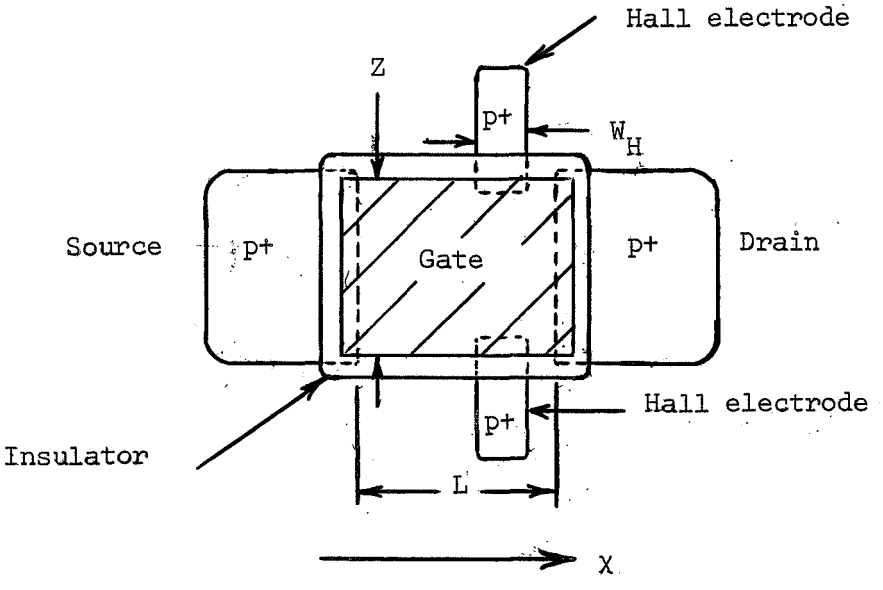


Figure 2.85. Equivalence circuit of MIS piezoelectric transistor used as a stress sensor.



(a) Side view



(b) Top view

Figure 2.86. Structure of MIS Hall effect device.

field perpendicular to both the current flow and the magnetic field. This can be detected as a Hall voltage or current at the Hall electrodes.

Two modes of operation will be considered. These are operation at low drain voltages where the channel charge density is essentially uniform along the channel and operation of the MIS device within the current saturation region. First at low drain voltages the ratio of Hall field to field along the channnel is

$$\frac{E_H}{E_x} = \mu B, \tag{2.271}$$

which produces an open circuit Hall voltage of

$$V_H = \mu B\left(\frac{Z}{L}\right) V_d. \tag{2.272}$$

In terms of drain current and gate voltage this can be expressed as

$$V_H = \frac{I_d}{C_I (V_g - V_T)} B. \tag{2.273}$$

This is all essentially identical to a standard Hall device. One difference however is the fact that with the MIS device the channel resistance or conductance is controlled by the gate voltage, consequently for a given current the magnitude of the Hall voltage depends on gate voltage as in Equation (2.273). To remain within the linear operating region I_d must be restricted so that

$$I_d << (1/2)\ \mu\ C_I(Z/L)\ (V_g - V_T)^2. \tag{2.274}$$

If we take as typical values for p-type silicon μ=150 $cm^2/V.$ sec, V_d = 1 volt, Z/L = 0.1, the Hall voltage sensitivity is estimated from Equation (2.272) as

$$V_H/B = 0.15 \text{ mV/kgauss.} \tag{2.275}$$

This small value is due to the low mobility of p-type silicon. Thus to achieve high sensitivities requries high mobility materials. When operated as a linear Hall element, the MIS device is equivalent to conventional Hall effect magnetic sensors and has the same sensitivity.

The high output impedance of the MIS transistor when operated in the saturation region can be used to increase the voltage sensitivity of the MIS Hall sensor [87]. When operated at saturation, the channel conductance is a function of distance along the channel. From the general MIS device equations, it is found that the Hall field is

$$E_H = \mu B E_x = \frac{\mu B(V_g - V_T)}{2L\sqrt{1 - x/L}} \tag{2.276}$$

This makes the Hall voltage a function of position along the channel and ideally

$$V_H = E_H Z = \frac{\mu B Z(V_g - V_T)}{2L\sqrt{1 - x/L}} . \tag{2.277}$$

This equation becomes infinite as $x \to L$ or at the pinch off region of the channel. The above derivation does not include the effects of

the ohmic source and drain contacts at x=0 and x=L. These require any measured Hall voltage to actually go to zero at these points. A plot of the Hall voltage as a function of distance as given by Equation (2.277) is shown in Figure 2.87. Also shown are qualitative sketches of the shorting effects of the end electrodes. A more complete analysis of the two dimensional Hall field verifies the major features of the simple model presented above [87]. Physically it is impossible as well as undesirable to locate the Hall contacts at x=L. If we choose x=0.75 L as a reasonable position for the Hall contacts then

$$V_H(x=0.75L) = \mu B(Z/L) (V_g - V_T). \qquad (2.278)$$

This can be taken as a good estimation of the Hall voltage which can be achieved by an MIS device operated in saturation. For typical silicon parameters ($V_g - V_T$ = 10 volts) this gives V_H values of a few mV.kgauss, which is still small compared with good Hall device of InSb.

If both the drain electrode and the Hall electrode are operated at saturation voltages, the current flowing through the Hall electrodes can be forced to flow through a high drain impedance and the voltage sensitivity can be increased. The circuit configuration for this type of operation is shown is Figure 2.88. The resistance values are those used by Fry and Hoey [87] in their measurements on silicon devices. An estimation of the increase in Hall electrode current due to a magnetic field is given by the equation

$$\Delta I_H \simeq \mu B(\frac{W_H}{Z}) I_d. \qquad (2.279)$$

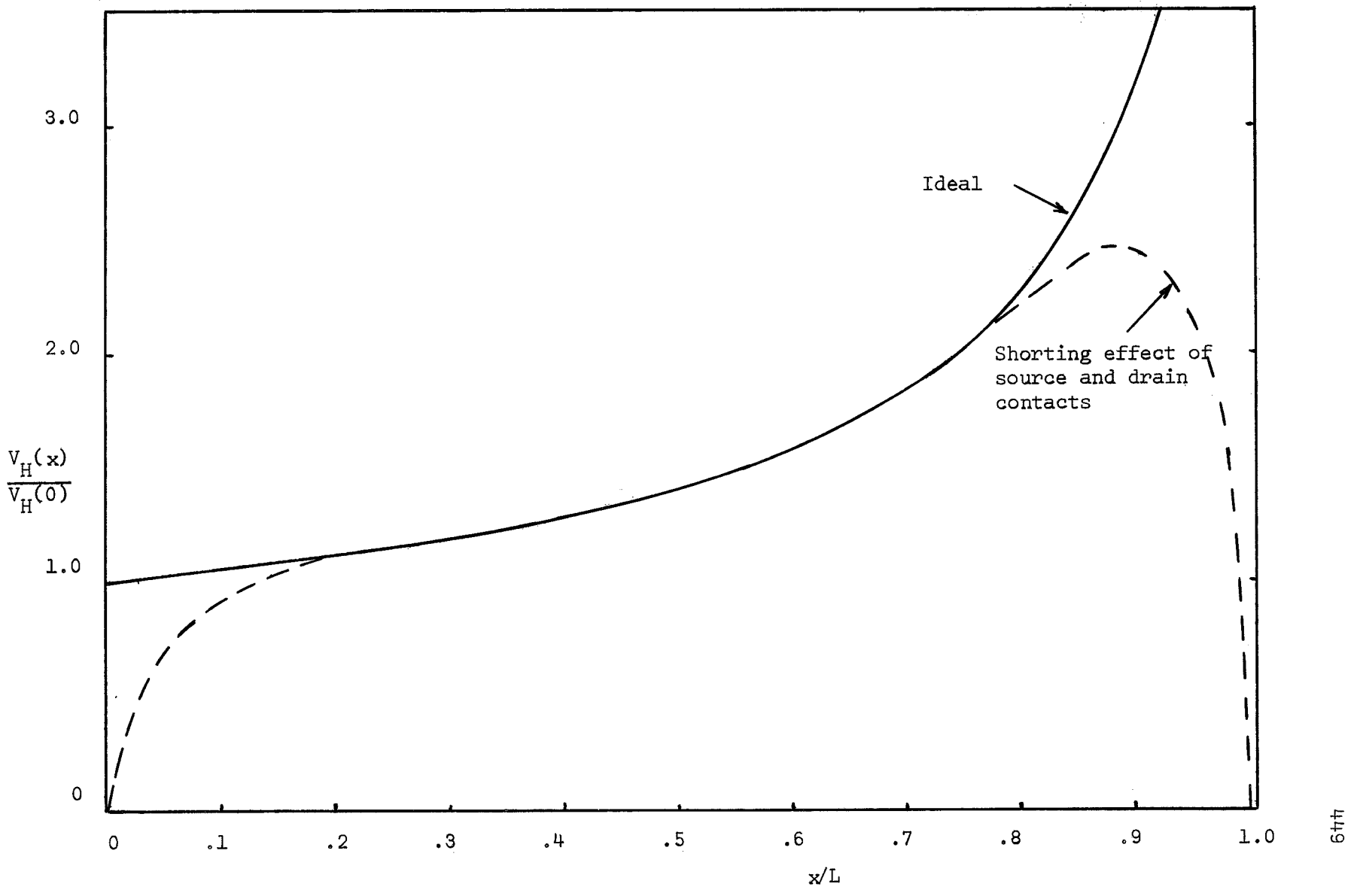


Figure 2.87. Dependence of Hall voltage on position along channel.

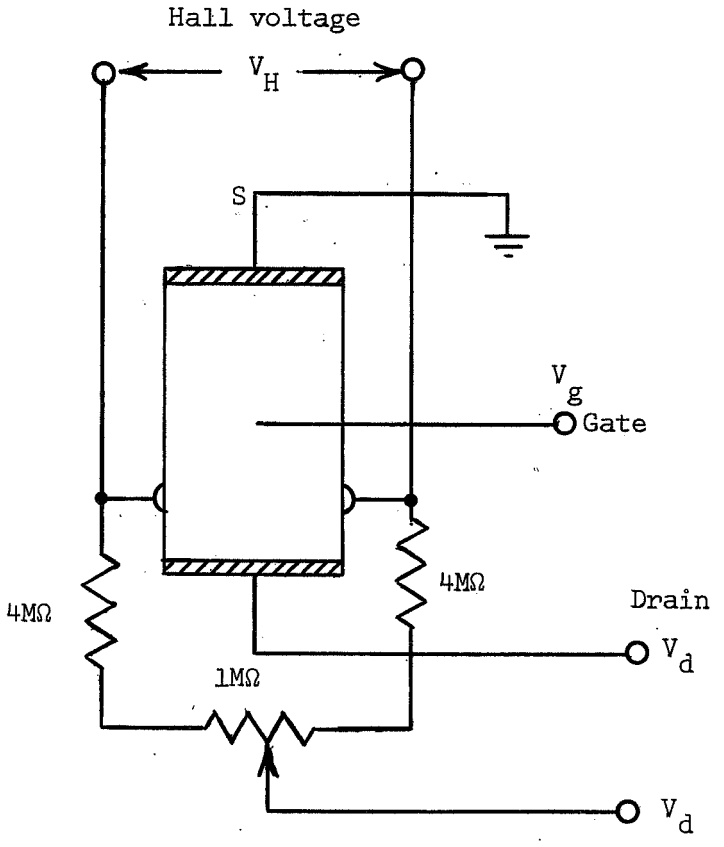


Figure 2.88. Operation of MIS Hall device with Hall electrodes at saturation for increased sensitivity [87].

This arises from assuming that in saturation no Hall voltages can be supported between the Hall electrodes. This current flows through the load resistance and gives a measured Hall voltage for the balanced arrangement of Figure 2.88 of

$$V_H \simeq \mu B\left(\frac{W_H}{Z}\right) I_d R_L. \tag{2.280}$$

For the values $\mu = 150$ cm^2/V.sec, $I_d \doteq 10^{-4}$ amp, $W_H/Z = 0.2$, $R_L \doteq 9$ MΩ which are typical of the values given by Fry and Hoey, the above becomes

$$V_H/B = 0.27 \text{ volt/kgauss} \tag{2.281}$$

The experimentally observed sensitivity for the above conditions was slightly less than a factor of two larger than the above calculated value. Thus it appears that Equation (2.280) is a reasonably good estimation of the sensitivity when the Hall electrodes are operated at saturation.

The voltage sensitivity when operated in the current saturation mode in Figure 2.88 is high as the above example shows. However the basic physical phenomena is the Hall effect. The high sensitivity is achieved by using the high output impedance of the MIS device and the inherent voltage gain present in the MIS device. It is unfair to compare the MIS Hall device with a simple Hall element sensor because of the gain present in the MIS device. A normal Hall magnetic field sensor followed by an MIS voltage amplifier is essentially equivalent to the MIS Hall device of Figure 2.86.

A major disadvantage of the MIS Hall sensor which combines both sensor and amplifier is that it is not possible in general to optimize both effects simultaneously. The most sensitive Hall devices are made in high mobility materials such as InSb while Si makes the best MIS devices. The performance of an InSb Hall sensor followed by an MIS device used as an amplifier will far exceed the sensitivity of a combined Si MIS Hall sensor. Unless size and not sensitivity is an overriding consideration, the MIS Hall sensor does not appear to have an advantage over a seperate Hall effect sensor and MIS amplifier. Noise is a problem which has not been investigated with the MIS Hall sensor. If it could be shown to have lower noise than a normal Hall sensor and MIS amplifier then this would be a definite advantage. However this remain to be investigated.

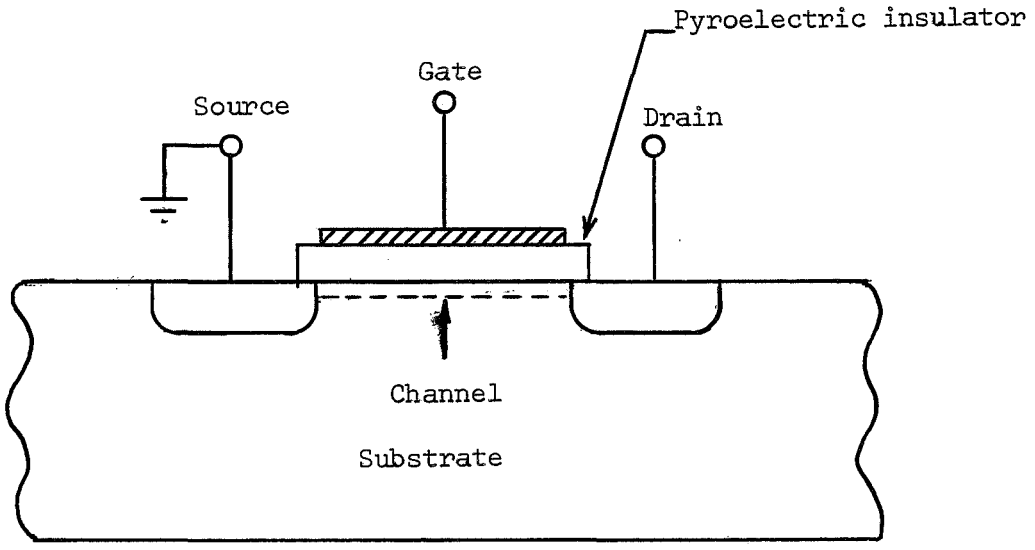
2.5.5 Other Sensors

The major sensor areas which have been investigated for MIS devices are discussed in the previous sections. There are a few other minor sensor applications which should be briefly discussed. First consider possible uses of MIS devices as temperature sensors. Temperature can effect the characteristics of an MIS device through either the mobility or the threshold voltages. With normal MIS devices these effects are not very large and are not easily predicted or controlled. Normal MIS devices then do not appear to be very useful temperature sensors. One exception to this however is a proposed MIS temperature sensor using a pyroelectric material as the insulator [88]. The proposed device configuration is shown in Figure 2.89 along with an equivalent circuit. A pyroelectric material is a material with a temperature dependent polarization. One example of such a material is $BaTiO_3$. The frequency response of the pyroelectric affect is limited by the dielectric relaxation time in a similar manner to the piezoelectric effect. If the temperature of a piezoelectric material is charged by an amount ΔT a surface charge appears on the material of the amount

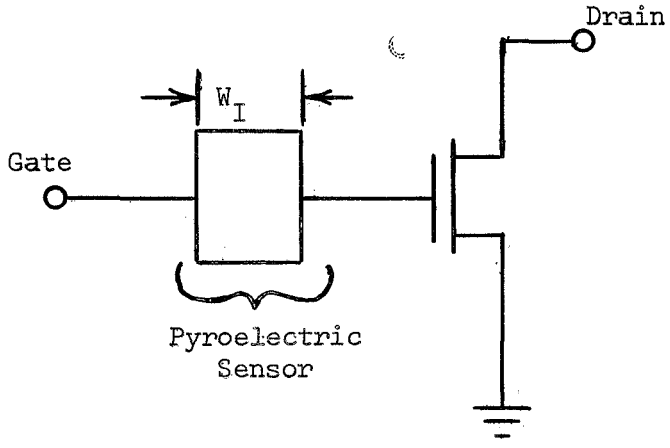
$$\Delta Q_{ss} = p\Delta T, \tag{2.282}$$

where p is the pyroelectric constant. If the insulator of an MIS device is pyroelectric this will give rise to a threshold voltage change of

$$\Delta V_T = \frac{\Delta Q_{ss}}{C_I} = \frac{p\Delta T}{C_I}, \tag{2.283}$$



(a) Device configuration



(b) Equivalent circuit

Figure 2.89. Pyroelectric MIS temperature sensor

and an output drain current change of

$$\frac{\Delta I_d}{\Delta T} = \frac{p \; g_m}{C_I} , \qquad (2.284)$$

where g_m is the MIS device transconductance. Using parameters appropriate to Si and $BaTiO_3$, a sensitivity of

$$\frac{\Delta I_d}{\Delta T} \sim 0.15 \text{ amp/°C} \qquad (2.285)$$

has been estimated [88].

The equivalent circuit of Figure 2.89 (b) shows that the pyroelectric MIS temperature sensor is equivalent to a pyroelectric sensor followed by a seperate MIS voltage amplifier. Thus to make the combined device competitive with the seperate components it must have unique advantages. Size is possibly one advantage. In most applications however this is probably more than offset by the problems in building the combined device. Thus the high sensitivity of the pyroelectric MIS device arises from the voltage gain of the MIS device and similar results can be achieved with seperate pyroelectric sensor and amplifier devices.

A second somewhat unique sensor which has been investigated using silicon MIS devices is a micrometeoroid detector [89]. In this application a large area MIS (SiO_2 as the insulator) capacitor is used. The device configuration is shown in Figure 2.90. Two inch diameter capacitors have been used with oxide thicknesses of 2,000 to 10,000 A°. When

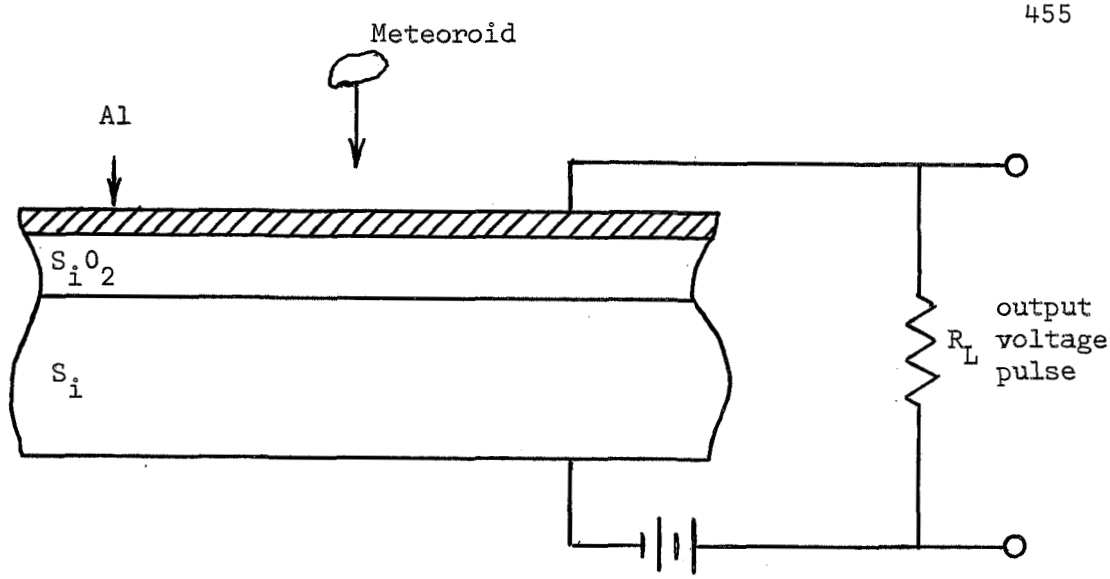


Figure 2.90. MIS (or MOS) capacitor type meteroroid detector.

the MIS device is hit by a micrometeroroid, destructive breakdown occurs in the vicinity of the point of penetration. The capacitor is momentarily shorted out at the point of micrometeoroid impact. The conducting path at the point of micrometeroid impact is found to dissipate in less than 1 μ sec allowing the capacitor to recharge through the load resistor. In this manner a pulse is generated in the external circuit for each micrometeoroid impact. Each impact results in the destruction of a small area of the capacitor around the point of impact. One of the big advantages of the MOS sensor as opposed to other capacitor sensors is the very thin oxide layer. This makes it possible for very small micrometeoroids to penetrate the device and give output pulses. This example is an application of well established silicon technology to a unique sensor problem. The results achieved indicate that the MOS capacitor makes a very good micrometeroroid sensor.

A final possible area for the use of MIS devices as sensors is in the area of chemical sensors or environment sensors. This involves the sensing of the component gases of the atmosphere or component gases of some other environment. There appears to have been little research in this area so the following statements are somwhat conjective based on the known properties of MIS devices. The first important idea is the fact that the MIS device is sensitive to very small surface charge densities. Charge densities of $10^{11}/cm^2$ to $10^{12}/cm^2$ are easily detected. Assuming an interatomic spacing of 5° A there are 4×10^{14} atoms/cm^2 on the surface. Thus a surface charge density of about 10^{-3} of an atomic layer is required for large changes in device current. The second important idea is the fact that gasses of various types are absorbed onto the surface of materials. The idea behind the MIS chemical sensor is that these two effects can be combined to produce an MIS device which is sensitive to gases absorbed onto the surface of the insulating layer.

There are many problems to be overcome before practical MIS chemical sensors can be achieved. First one must have an insulator or possible thin metal film which will selectively absorb the gas to be detected. Second the process should be reversible with no permanant change in the MIS device characteristic. Whether all these problems can be solved and sensitive MIS chemical sensor fabricated remains to be seen. This does however appear to be an area of sensor research worthy of an extended investigation.

2.6 Sensor Arrays

The previous sections have been concerned with the properties of individual sensors. In some sensor applications one-and two-dimensional arrays of sensors are needed. The use of large two-dimensional arrays of optical sensors is becoming very important in information handling and processing systems. The present section discusses some of the unique problems associated with sensor arrays.

The problems with sensor arrays are somewhat different from the problems of individual sensors. First one must be able to produce a large number of sensors with almost identical characteristics and to produce them with high yield. This problem is identical to that faced by all large scale integrated circuits. The second major problem is that of addressing or scanning the large number of sensors. The large sensor arrays are essentially large scale integrated circuits and they have the same problems of reproducibility and yield as do other large scale integrated circuits. The solution of these problems for integrated circuits will also provide the needed solutions for sensor arrays.

Silicon and germanium are the only materials wiht which serious attempts have been made at large two-dimensional sensor arrays. Silicon is obviously the most logical choice because of the well established technology. One dimensional sensor arrays have been made in some other materials such as InSb, InAs, and HgCdTe. The disussion is this section concerns basically arrays of junction sensors. Sensor arrays can be made using bulk phenomena such as photoconductivity in materials such as HgCdTe [90]. The major problems of sensor arrays are common to either junctions sensors or bulk effect sensors. Junction sensors

have an advantage in their small size and reduced interaction between sensor elements.

2.6.1 Optical Sensors

Most of the research in optical sensor arrays has been directed toward silicon image converter systems. Silicon p-n junction sensors have a useful response to light of wavelengths from about 0.4 μ to about 1.0 μ. This covers the visible spectrum and into the near infrared. Image converter systems are essentially large (100 x 100 or 200 x 200) two-dimensional arrays of closely spaced individual sensor elements. The sensors may be either p-n junctions or phototransistors or could in fact be metal semiconductor junctions. The characteristics of individual sensors such as are used in sensor arrays have been discussed in previous sections.

One of the major problems with sensor arrays is that of scanning the individual sensor elements. There are basically two techniques which are employed for scanning. One uses an electron beam for periodically sampling the output of each sensor and the other technique uses an array of X and Y connecting stripes such that one individual X line and one individual Y line uniquely defines one sensor elements. These two techniques will be referred to as electron beam scanning and XY scanning. The exact sensor configuration depends on the scanning technique employed.

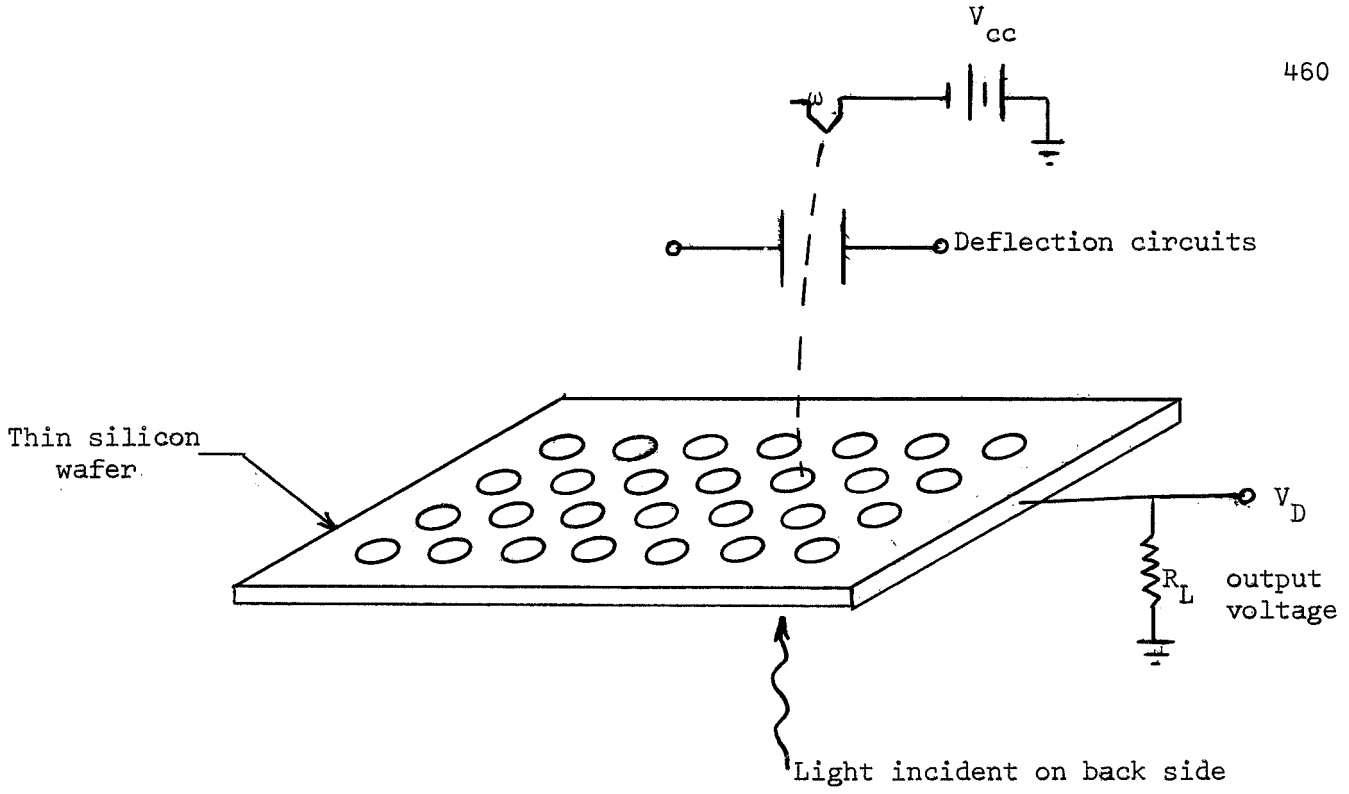
The simplest scanning technique and consequently the technique with the highest yield is the electron beam scanning. An electron beam represents the simplest technique known for electronically locating one spot within a two-dimensional plane. The use of electron beam

scanning with an array of diode sensors is shown in Figure 2.91. The basic sensor array consists of closely spaced p-n junctions as shown in Figure 2.91 (b). With electron beam scanning the light is normally incident on the silicon wafer on the side opposite the p-n junctions. This requires that the semiconductor wafer be less than a diffusion length in thickness. With this condition met and with a low surface recombination velocity on the surface, the carriers will diffuse across the wafer and be collected by the p-n junction.

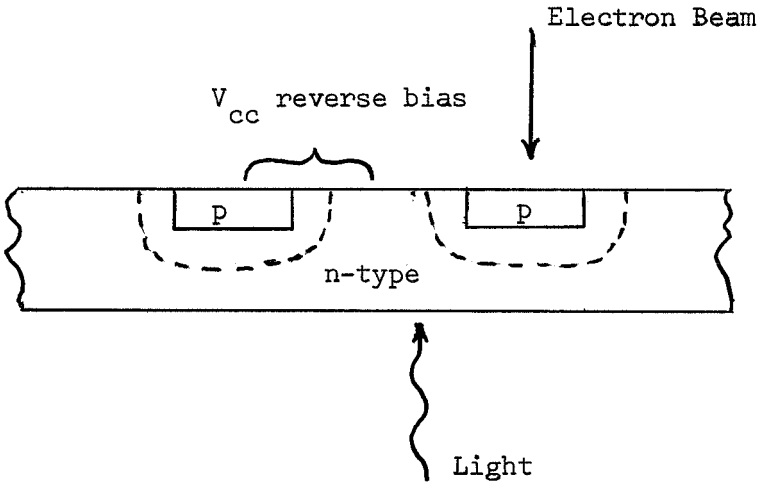
The operation of each individual sensor can be understood with the aid of Figure 2.91 (b). When the electron beam is on an individual sensor, electrons are deposited on the p-type material. This results in the buildup of negative charge on the p-type region and reverse biases the p-n junction. If the electron beam is on an individual diode for a sufficient time, the diode becomes reverse biased to V_{cc} the electron beam acceleration voltage. In the normal scanning operation, the electron beam remains on each individual element for a short time but sufficiently long to establish the reverse bias of V_{cc} across the diode. During the time the electorn beam is off any individual element the reverse bias voltage decays back toward zero. The current tending to reduce the bias voltage comes from the normal reverse bias diode current and from the collected photocurrent. For proper sensor operation the photocurrent must exceed the normal reverse bias current. During a time τ the charge collected by the junction is

$$Q_{ph} \simeq I_{ph} \tau. \tag{2.286}$$

This results in a voltage change across the diode of



(a) Two-dimensional array



(b) Individual sensors

Figure 2.91. Electron beam scanning of diode optical sensor array.

$$\Delta V_J = \frac{I_{ph}\,\tau}{C_J}, \tag{2.287}$$

where C_J is the diode capacitance. For actual diodes I_{ph} and C_J are slightly voltage dependent but these complicating effects will be neglected in this simple analysis. If τ is the scan time of the electron beam across the entire sensor array, the above equation gives the voltage change across an individual diode from the time the electron beam left a diode to its return to the diode. When the electron beam returns to the diode it returns the reverse bias voltage to V_{cc}. This results in a transient current through the diode and through the external resistor R_L in Figure 2.91 (a). Ideally the peak output voltage is then

$$V_o)_{peak} \simeq \Delta V_J = \frac{I_{ph}\,\tau}{C_J}. \tag{2.288}$$

The peak output voltage is then directly proportional to the photon current and the scanning time τ. Further, since I_{ph} is proportional to the photon flux incident on the semiconductor surface in the vicinity of the sensor, the peak output voltage is porportional to the photon flux. This technique of sensor operation is known as the stored charge, or integrated flux, or pulsed mode technique. The output is proportional to the total charge collected at the p-n junction even though the sensor is only sampled periodically. The results in a very efficient sensor array which is sensitive to low light levels. The sensor array is not restricted to diodes but the individual sensors may be phototransistors.

When operated in the stored charge mode the individual sensors give

a linear output with incident photon flux over a wide range of flux values. The response is however limited at both low light levels and at high light levels. At low light levels there are two limiting effects, first the photon current must be larger than the reverse bias current. Second for any detection system there is a minimum detectable output voltage change due to noise in the system. Thus the lower limit of photocurrent is

$$I_{ph} > I_{rev} + \frac{V_{min} C_J}{\tau}, \tag{2.289}$$

where I_{rev} is the reverse bias current and V_{min} is the minimum detectable output voltage. The upper limit on photocurrent is determined by the maximum allowed change in diode voltage. This is obviously V_{cc}. Thus the combined limits on photocurrent are

$$I_{rev} + \frac{V_{min} C_J}{\tau} < I_{ph} < \frac{V_{cc} C_J}{\tau} \tag{2.290}$$

Since I_{ph} is proportional to photon flux these show that there are upper and lower limits to light intensity over which linear operation is possible. If I_{rev} is negligible and $V_{min} \simeq 10^{-2} V_{cc}$, the linear region extends over about two order of magnitude in light intensity. The limits also depend on scan time τ and can be changed by changing the scan time.

Consider next the XY scanning technique A typical section of an XY scanned array of phototransistors is shown in Figure 2.92. [91]. The individual transistors are manufactured essentially identically to integrated circuit transistors. The major difference is that the collectors of the transistors are organized into Y lines with a common

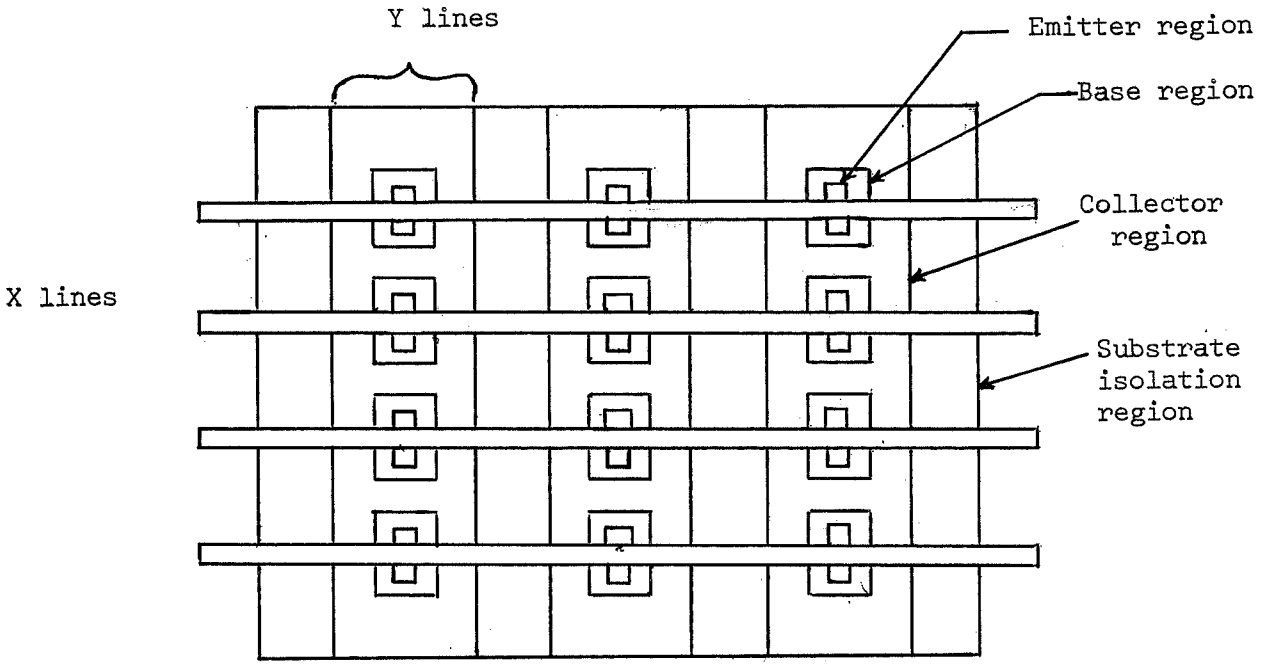


Figure 2.92. XY scanned phototransistor array [91].

collector for all transistors in a row along the Y axis. Horizontal X lines which make ohmic contact only to the emitter regions then completes the XY matrix. A voltage applied between any collector Y line and an emitter X line applies a bias to one unique phototransistor. The array is scanned by periodically applying voltage pulses along the X and Y lines. With XY scanning the indicent light can be on the surface containing the sensor diodes leading to slightly more efficient operation at short wavelengths.

The stored charge mode of operation can also be used with XY scanning as well as with phototransistors. With phototransistors the photocurrent appearing in Equations (2.286) through (2.290) should be interpreted as the photocurrent collected by the junction multiplied

by the transistor beta. The experimentally observed response of an XY scanned phtototransistor in the charge storage mode is shown in Figure 2.93 as a function of photon flux [91]. The peak output voltage as predicted by Equation (2.288) is essentially linear with photon flux over about two orders of magnitude in photon flux. The saturation predicted by Equation (2.290) is also seen at both low and high flux levels. It is noted that large scan times are required for low light levels and short scan times are required for high light levels. The scan time can be adjusted to compensate for changing values of average light intensity. Typical scan rates for a 200 line system range from 5 Hz to 100 Hz. A rate of about 60 Hz has been found to be optimum for XY scanning of silicon phototransistors under normal room illumination [91]. The minimum detectable signal at 60 Hz scanning rate has been found to be about 10^{-4} watts/cm^2.

Some general conclusions can be drawn concerning electron beam and XY scanning. Each technique has certain advantages and disadvantages. First electron beam scanning is simple in terms of external circuits required for scanning. Simple voltage ramps are needed to scan across a two dimensional plane. XY scanning requires complex scanning circuits. Electron beam scanning requires a hybird vaccum solid state technology while XY scanning is completely solid state. Certainly the all solid state approach is the perferred approach. There still remain, however, problems with achieving the high yields required for the XY scanning. With electron beam scanning a single defective sensor is not too serious a problem. It produces either a small light or dark spot on the array. With XY scanning the most pronounced defects are entire

scanning lives missing. A defective driving line renders an entire line of the matrix defective and these result in very pronounced defects in the sensor array. To obtain a good XY scanned array requires 100% yield on all scanning lines. To achieve this yield on large 100 x 100 or 200 x 200 sensor arrays has so far been extremely difficult.

Most optical sensor array research has been directed toward the visible spectral region. Some research however has been reported on infrared sensor arrays using germanium [92] and InAs [93]. Since the technology for these materials is more difficult than that for silicon the development of sensor arrays in these materials is likely to be much slower.

For silicon sensor arrays it appears that most of the problems are well along the way to being solved - especially for electron beam scanned arrays. The solution of the remaining problems for XY scan-

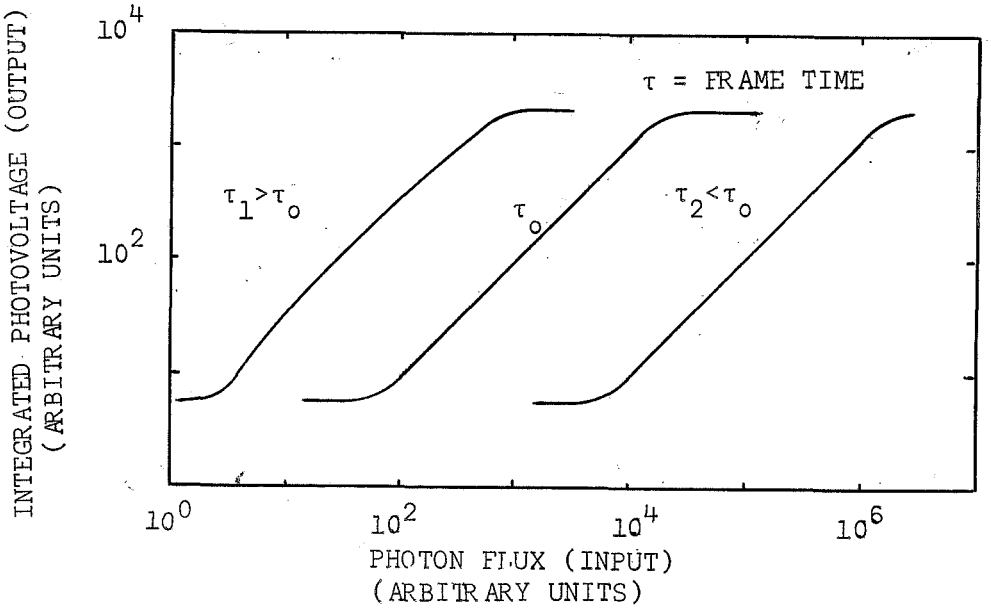


Figure 2.93. Output response of XY scanned phototransistor in the charge storage mode [91].

ned arrays are closely related to large scale integration problems. Because of the large effort being expended by industry in these, yield problems will likely be solved in the near future and inexpensive optical sensor arrays of 200 x 200 or larger elements will become available.

2.6.2 Other Sensors

The requirements for large sensor arrays appears to be essentially unique to optical sensors. In some cases other arrays of sensors are required such as stress sensor arrays or temperature sensor arrays. Stress and temperature do not normally vary greatly over small distances so there is little use for large arrays of small sensor elements as in optical sensors. There is no fundamental reason why large stress sensor arrays could not be built with integrated circuit techniques. However there does not appear to be any great use for such sensor arrays. If the area of accoustoelectric effects should become very important there might then be a demand for large stress sensor arrays. Another possible area might involve large arrays of magnetic sensors for such future devices as magnetic bubble circuits and memories. With integrated circuit techniques well established techniques are available for producing solid state sensor arrays. Should the need arise for large arrays of stress, temperature or magnetic field sensors it appear that they can be as easily built as large optical sensor arrays.

LIST OF REFERENCES

1. A. B. Phillips, Transistor Engineering, McGraw-Hill Book Co., Inc., New York, 1962.

2. J. Lindmayer and C. Y. Wrigley, Fundamentals of Semiconductor Devices, D. Van Nostrand Co., Inc., New York, 1964.

3. W. W. Gartner, Transistors, Principles, Design and Applications, D. Van Nostrand Co., Inc., New York, 1960.

4. A. S. Grove, Physics and Technology of Semiconductor Devices, John Wiley and Sons, Inc., New York, 1967.

5. J. R. Hauser, "An Approximation for Generation-Recombination Current in P-N Junctions", Proc. IRE vol. 53, July 1965, pp. 743-744.

6. C. T. Sah, R. N. Noyce, and W. Schokley, "Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics", Proc. IRE, vol 45, Sept. 1957, pp. 1228-1243.

7. R. M. Burger, "Integrated Silicon Device Technology, Volume VIII Diodes", Technical Documentary Report No. ASD-TDR-63-316 Contract AF 33(657) - 10340, Research Triangle Institute, Durham, North Carolina, March 1966, AD 630 894.

8. S. L. Miller, "Ionization Rates for Holes and Electrons in Silicon", Phys. Rev., vol 105, Feb. 1957, pp. 1246-1249.

9. S. M. Sze, Physics of Semiconductor Devices, Wiley-Interscience, John Wiley and Sons, New York, 1969.

10. A Herlet, "Bestimmung der diffusionslange L unter der inversionsdichte n.aus den durchlasskennlinie von legierten silizum - flacher-gleichrichterm". fur Angewandte Physik, vol. 9, no. 4, April 1957, pp. 155-158.

11. E. D. Graham, "Computer Techniques for Accurate Solid State Device Analysis", Ph.D. Thesis, September 1969, N. C. State University, Electrical Engineering Department.

12. E. M. Pell, "Ion Drift in an p-n Junction", J. Appl. Phys., vol. 31, February, 1960, pp. 291-302.

13. J. W. Mager, "Characteristics of p-i-n Junction Produced by Ion-Drift Techniques in Silicon", J. Appl. Phys., vol. 33, September, 1962, pp. 2894-2902.

14. M. Neuberger and S. J. Welles, "Silicon", EPIC Report, Hughes Aircraft Company, Culver City, California, October 1969, Contract F 33615-68-C-1125.

15. H. R. Philipp and E. A. Taft, "Optical Constants of Silicon in the Region 1 to 10 eV", Phys. Rev., vol. 120, no. 1, October 1, 1960, pp. 37-38.

16. H. Melchior and W. T. Lynch, "Signal and Noise Response of High Speed Germanium Avalanche Photodiodes", IEEE Trans. on ED, vol. ED-13, no. 12, December 1966, pp. 829-838.

17. A. A. Gutkin, E. M. Magerramov, D. N. Nasledov, and V. E. Sedov, "Spectral Characteristics of GaAs p-n Junctions in the Near Ultraviolet Region", Soviet Physics - Solid-State, vol. 8, no.3, September 1966, pp. 570-573.

18. V. S. Vavilov, "On Photo-ionization by Fast Electrons in Germanium and Silicon", J. Phys. Cehm. Solids, vol. 8, January 1959, pp. 223-226.

19. A. L. Morse, W. F. Crevier, G. L. Weissler, and D. B. Medved, "Vacuum Ultraviolet and Solid State Physics, Technical Report No. USC-Vac UV-101, 15 June 1965, Contract NsG-178-61, N66-17303.

20. R. J. Schwartz and C. W. Kim, "P-I-N Thermo-Photo-Voltaic Diode", School of Electrical Engineering, Purdue University, August 1967, Joint Services Electronic Program OND N00014-67-A 00226-004, AD 658 902.

21. R. P. Rietz, "High Speed Semiconductor Photodiodes", Rev. Sci. Int., vol. 33, no. 9, September 1962, pp. 994-998.

22. D. B. Medved and G. P. Rolik, "Electronic Tuning of Narrow Band Detectors", Appl. Phys. Letters, vol. 10, no. 8, 15 April 1967, pp. 213-214.

23. L. K. Anderson, P. G. McMullin, L. A. D'Asaro, and A. Goetzberger, "Microwave Photodiodes Exhibiting Microplasma-Free Carrier Multiplication", Appl. Phys. Letters, vol. 6, no. 4, 15 February 1965, pp. 62-64.

24. R. D. Baertsch, "Noise and Ionization Rate Measurements in Silicon Photodiodes", IEEE Trans. Electron Devices, vol. ED-13, no. 12, December 1966, p. 987.

25. H. Kroger, "High-Grain Low-Noise Photodetectors", Sperry Rand Research Center, Scientific Report for period 1 April 1965 - 31 March 1966, Contract No. N600(19)63823, AD 632 441.

26. R. Kuvas and C. A. Lee, "Quasistatic Approximation for Semiconductor Avalanches", J. Appl. Phys., vol. 41, no. 4, 15 March 1970, pp. 1743-1755.

27. W. K. Hofker, "Semiconductor Detector for Ionizing Radiation", Philips Technical Review, vol. 27, no. 12, 1966, pp. 323-336.

28. J. M. Hollander and I Perlman, "The Semiconductor Revolution in Nuclear Radiation Counting", Science, vol. 154, 7 October 1966, pp. 84-93.

29. T. F. Prosser, "An Integrated Temperature Sensor-Controller", IEEE J. of Solid-State Circuits, vol. SC-1, no. 1, September 1966, pp. 8-13.

30. Lake Shore Cryotronics, Inc. P. O. Box 214, Hamburg, New York 14075.

31. Texas Instruments Inc., "Preferred Semiconductors and Components" Handbook, 1 September 1969.

32. W. Rindner, "Anisotropic Strain Effects in p-n Junctions", Bull. Am. Phys. Soc., vol. 7, January 1962, pp. 65-66.

33. W. Rindner and I. Braun, "Resistance of Elastically Deformed Shallow p-n Junctions II", J. Appl. Phys., vol. 34, July 1963, pp. 1958-1970.

34. M. E. Sikorski, P. Andreatch, A. Grieco, and H. Christensen, "Transistor Microphone", Rev. Sci. Instr., vol. 33, October 1962, pp. 1130-1131.

35. J. J. Wortman and J. R. Hauser, "Effect of Mechanical Stress on p-n Junction Device Characteristics. II. Generation-Recombination Current", JAP, vol. 37, no. 9. August 1966, pp.3527-3530.

36. A. Goetzberger and R. H. Finch, "Lowering the Breakdown Voltage of p-n Junctions by Stress", J. Appl. Phys., vol. 35, June 1964, pp. 1851-1854.

37. H. H. Hall, J. Bardeen and G. L. Pearson, "The Effects of Pressure and Temperature on the Resistance of p-n Junctions in Germanium", Phys. Rev., vol. 84, October 1951, pp. 129-132.

38. B. J. Sloan and J. R. Hauser, "The Effects of Uniaxial Compressive Stress on Minority Carrier Lifetime in Silicon and Germanium", J. Appl. Phys., vol. 41, July 1970, pp. 3504-3508.

39. J. J. Wortman, J. R. Hauser, and R. M. Burger, "Effect of Mechanical Stress on p-n Junction Device Characteristics", J. Appl. Phys., vol. 35, July 1964, pp. 2122-2131.

40. J. J. Wortman, "Effects of Mechanical Strain on p-n Junctions", NASA Contractor Report, NASA CR-275, August 1965.

41. J. R. Hauser and J. J. Wortman, "Some Effects of Mechanical Stress on the Breakdown Voltage of p-n Junctions", J. Appl. Phys., vol. 37, no. 10, September 1966, pp. 3884-3892.

42. G. E. Pikus and G. L. Bir, "Effect of Deformation on the Hole Energy Spectrum of Germanium and Silicon", Soviet Physics - Solid State, vol. 1, no. 11, May 1970, pp. 1502-1517.

43. C. D. Parker, "A Feasibility Study of a Miniature Solid-State Pressure Transducer", Final Report, NASA Contract NAS1-7489, Research Triangle Institute, September 1968.

44. L. K. Monteith and J. J. Wortman, "Characterization of p-n Junctions under the Influence of a time Varying Mechanical Strain", AFOSR Final Scientific Report on Air Force Contract No. F44 620-67-c-0048, "Effect of Stress Waves on p-n Junctions", Research Triangle Institute, January 1970.

45. J. J. Wortman and L. K. Monteith, "Semiconductor Mechanical Sensors", IEEE Trans. on Elec. Dev., vol. ED-16, no. 10, October 1969, pp. 855-860.

46. M. I. Nathan and W. Paul, "The Pressure Dependence of Tunneling in Esaki Diodes", Int. Conf. Semiconductor Phys. Prague, 1960, p. 209.

47. W. Rindner, A. Garfein, E. Pittelli, and A. Inannini, "A Miniature Electromechanical Tunnel Diode Transducer", Solid-State Electronics, vol. 13, April 1970, pp. 457-471.

48. M. E. Sikorski, "Use of p-n Junctions Semiconductor Transducers in Pressure and Strain Measurements", in Physical Acoustics, Edited by W. P. Mason, vol. 1, part B., Academic Press, New York, 1964.

49. A. R. Calawa, R. H. Rediker, B. Lax and A. L. McWhorter, "Magneto-Tunneling in InSb", Phys. Rev. Lett., vol. 5, no. 2, July 15, 1960, pp. 55-57.

50. P. N. Butcher, J. A. Hulbert, and K. F. Hulme, "Indium Antimonide Tunnel Diodes in High Magnetic Fields", Phys. Chem. Solids, vol. wl, 1961, pp. 320-324.

51. M. Garfinkel and W. E. Engeler, "Magnetic Field Effect on the Current Distribution in p-n Junctions", J. Appl. Phys., vol. 36, no. 6, June 1965, pp. 1877-1882.

52. Meheshawar Misra and G. P. Srivastava, "Diffusion of Minority Carriers in the Base Region of an Alloy Junction Transistor in the Presence of a Magnetic Field", J. Appl. Phys., vol. 39, no. 4, March 1968, pp. 2127-2131.

53. E. I. Karakushan and V. I. Stafeev, "Magnetodiodes" Soviet Physics-Solid State, vol. 3, no. 3, September 1961, pp. 493-498.

54. R. Parshad and S. C. Mehta, "Effect of Magnetic Field on the V-I Characteristic of Solid State Diodes", Indian J. Pure Appl. Phys., vol. 5, January 1967, pp. 23-25.

55. S. C. Mehta and R. Parshad, "Breakdown Phenomena in Silicon p-n Junctions Under Magnetic Fields", J. Appl. Phys., vol. 41, no. 2, February 1970, pp. 760-764.

56. E. I. Karakushan and V. I. Stafeev, "Large Area Magnetodiodes", Soviet Physics-Solid State, vol. 3, no. 7, January 1962, pp. 1476-1482.

57. R. L. Anderson, "Experiments on Germanium-Gallium Arsenide Heterojunctions", Solid-State Elec., vol. 5, Sept-Oct. 1962, pp. 341-351.

58. S. S. Perlman, "P-N Heterojunctions with Experiments on Germanium-Gallium Arsenide Structures", Ph.D. Thesis, Carnegie Institute of Technology, Elec. Engr. Dept., May 1963.

59. S. S. Perlman and D. L. Feucht, "P-N Heterojunctions", Solid-State Elec., vol. 7, December 1964, pp. 911-923.

60. R. L. Anderson, "Experiments on Ge-GaAs Heterojunctions", Solid-State Electronics, vol. 5, Sept.-Oct. 1962, pp. 341-351.

61. R. L. Anderson, G. Zeidenbergs, and M. Davis, "Heterojunctions Devices", Technical Report No. RADC-TR-68-70, March, 1968 Syracuse University, Dept. of Electrical Engineering, AD 669 365.

62. A. K. Sreedhar, B. L. Sharma, and R. K. Purohit, "Efficiency Calculations of Heterojunction Solar Energy Converters", IEEE Trans. on Elec. Dev., vol. ED-16, no. 3, March 1969, pp. 309-312.

63. F. A. Shirland, "The History, Design, Fabriaction and Performance of CdS Thin Film Solar Cells", Advanced Energy Conversion, vol. 6, no. 4, October-December 1966, pp. 201-222.

64. E. R. Hill, "A Model for the CdS Solar Cell", Proc. of the 5th Photovoltaic Specialists Meeting, vol. II, January 1966, N66-17333.

65. M. Neuberger, "II-VI Semiconducting Compounds Data Tables", EPIC Report, Hughes Aircraft Co., Culver City, California, October 1969.

66. R. M. Moore, "The Effect of a Field-Independent Polarization Discontinuity on Heterojunction Characteristics", IEEE Trans. on ED, vol. ED-16, no. 2, February 1969, pp. 186-196.

67. R. M. Moore and C. J. Busanovich, "The Heterode Strain Sensor: An Evaporated Heterojunction Device", IEEE Trans. on ED, vol. Ed-16, no. 10, October 1969, pp. 850-855.

68. Robert M. Moore, and Charles J. Busanovich, "The Feasibility of a Heterode Pressure Transducer", Proc. of IEEE, vol. 57, no. 8, August 1969, pp. 1433-1434.

69. R. M. Moore and C. J. Busanovich, "An Evaporated Heterojunction Diode Strain Sensor", Proc. of IEEE, vol. 57, no. 4, April 1969, pp. 735-736.

70. Research Triangle Institute, "Feasibility Study of Piezotransistor Accelerometers", Annual Report Contract NASr-222, NASA, September 1965.

71. S. G. Bandy, "Design, Fabrication, and Evaluation of a Silicon Junction Field-Effect Photodetector", Solid-State Electronics Laboratory, Stanford University, Report SU-SEL-69-075, Dec. 1969, Technical Report No. 4828-2, Contract OEG 0-8-071112-2995.

72. C. R. Crowell and S. M. Sze, "Current Transport in Metal-Semiconductor Barriers", Solid-State Elec., vol. 9, November/December 1966, pp. 1035-1048.

73. C. R. Crowell and V. L. Rideout, "Normalized Thermionic-Field (T-F) Emission in Metal-Semiconductor (Schottky) Barriers", Solid-State Elec., vol. 12, February 1969, pp. 89-105.

74. Jerome Cohen, Juri Vilms, and Rober J. Archer, "Investigation of Semiconductor Schottky Barriers for Optical Detection and Cathodic Emission", Hewlett-Packard Company, Final Report 15 November 1967-November 1968, Contract No. F19628-68-C-0090, AD 682 522.

75. Jerome Cohen, Juri Vilms, Robert J. Archer, "Investigation of Schottky-Barriers for Optical Detection and Cathodic Emission", Final Report Contract No. F 19628-69-C-0148, 1 February 1969- 31 May 1969, Hewlett-Packard Company, Palo Alto, California, AD 691 283.

76. John R. Sizelove and John A. Love, III, "Narrow Spectral Response Schottky Photodiodes", Applied Optics, vol. 6, no. 10, October 1967, pp. 1777-1780.

77. M. V. Schneider, "Schottky Barrier Photodiodes with Antireflection Coatings", Bell System TEch. J., vol 45, November 1966, pp. 1611-1938.

78. R. D. Baertsch and J. R. Richardson, "An Ag-GaAs Schottky-Barrier Ultraviolet Detector", J. Appl. Phys., vol. 40, January 1969, pp. 229-235.

79. V. L. Rideout and C. R. Crowell, "Pressure Sensitivity of Gold-Potassium Tantalate Schottky Barrier Diodes", Applied Phys. Letters, vol. 10, no. 11, 1 June 1967, pp. 329-332.

80. H. Okamoto, H. Ariyoshi, and G. Mizushima, "Observation on Stress Effect in Schottky Barrier Diodes", Solid-State Elec., vol. 12, May 1969, pp. 441-444.

81. K. Chino, H. Okamoto, H. Ariyoshi, and Y. Mizushima, "Stress Effect of Gold-Doped and Gamma-Irradiated Schottky-Barrier Diodes", Japan J. Appl. Phys., vol. 8, no. 4, April 1968, pp. 502-503.

82. A. S. Grove, B. E. Deal, E. H. Snow, and C. T. Sah, "Investigation of Thermally Oxidized Silicon Surfaces Using Metal-Oxide-Semiconductor Structures", Solid-State Elec., vol. 8, February 1965, pp. 145-163.

83. D. Colman, R. T. Bate, and J. P. Mize, "Mobility Anisotropy and Piezoresistance in Silicon p-Type Inversion Layers", J. Appl. Phys., vol. 39, no. 4, March 1968, pp. 1923-1931.

84. R. S. Muller and J. Conragan, "A Metal-Insulator-Piezoelectric Semiconductor Electromechanical Transducer", IEEE Trans. on ED, vol. ED-12, no. 11, November 1965, pp. 590-595.

85. J. R. Fiebiger and R. S. Muller, "Electrical Performance of Metal Insulator-Piezoelectric Semiconductor Transducers", J. Appl. Phys., vol. 38, no. 4, 15 March 1967, pp. 1948-1955.

86. J. Conragan and R. S. Muller, "Piezoelectric Field-Effect Transistor Strain Transducers", 1970 Solid State Sensors Symposium, Conference Record Extended Abstracts, Minneapolis, Minnesota, June 18-19, 1970.

87. P. W. Fry and S. J. Hoey "A Silicon MOS Magnetic Field Transducer of High Sensitivity", IEEE Trans. on Ed, vol. ED-16, no. 1, January 1969, pp. 35-39.

88. F. T. Wooten, "A Proposed Ultra-Sensitive Miniature Temperature Sensor", Proc. of IEEE, vol. 55, no. 4, April 1967, p. 564.

89. R. P. Donovan, J. H. Siviter, and L. K. Monteith, "MOS Capacitors as Micrometeoroid Detectors", paper presented at 137th National Meeting of the Electrochemical Society, May 10-15, 1970, Los Angeles, California.

90. William Vaughan, "A New Material for IR Detectors", Optical Spectra, vol. 4, February 1970, pp. 57-61.

91. Westinghouse Electric Corp., Staff of Science and TEchnology Aerospace Division, Integrated Electronic Systems, Prentice-Hall, Inc. New Jersey, 1970.

92. M. A. Schuster and J. M. Saboe, "A Monolithic Germanium Imaging Mosaic", Institute of Electrical and Electronics Engineers Annual Region III Convention, 7th , Cocoa Beach, Fla., Nov. 18-20, 1968, Record, A69-25307.

93. W. E. Davern, R. E. Glusick, C. W. Kim, M. E. Seymour, and R. D. Stewart, "Solid State Image Sensor Research" Final Technical Report, General Electric Co., March 1969, Contract No. NAS 12-131, N69-36922.

3.0 STATE-OF-THE-ART FOR SOLID-STATE SENSORS

The impact of contemporary solid state materials and technology on sensor development appears diverse and overwhelming. However much of the activity can be characterized as research and development and one can only estimate the potential output. To obtain a realistic assessment of solid state sensors it is important that the overview contain a reasonable balance between accomplishment and activity. To achieve and maintain this balance may be impossible. Therefore, the authors have attempted an organization of the survey which provides a contrast between activity and accomplishment. In addition, an assessment will be provided which admittedly reflects the view point of the authors.

One of the major difficulties with the organization of the survey was the association of various transducers and transduction principles. As an example, acceleration, displacement, and velocity could be considered individually. The development of transducers to measure each mechanical variable could be documented. In fact an excellent treatment of transducers has been organized in this manner and has been very convenient [1]. However, this approach tends to emphasize the relationship between measurements and transducer design and operation.

In many instances the critical feature of transducer design and operation is the housing required for a particular measurement. For example, each of the mechanical transducers listed above can be designed using the strain-gauge transduction principle. However, the design and operation of each transducer is distinct. One should certainly consider semiconductor strain-gauge as the transduction element and recognize

that semiconductor strain gauges represent a contribution of solid state to transducer development. In addition, thin-film, wire, metal-foil and many other strain gauges may be used. Therefore it appeared more reasonable to the authors to compare and contrast strain-gauges as opposed to housing design to accomodate various strain gauges. Pursuing an organizational philosophy which emphasizes transduction principles can result in numerous catagories. However we have attempted to cover only two major catagories in detail. Mechanical and optical transducers have been greatly influenced by solid state technology and represent a logical choice. Admittedly this has proved convenient for the authors and may tend to obscure other transducer development. However, we have attempted to minimize this effect with the detailed treatment of various bulk and junction properties suitable for transducers in Sections 1 and 2.

3.1 Transduction Principles for Mechanical Transducers

Mechanical transducers generally respond to forces and motion. The essence of a mechanical transducer is the sensing element or that which responds to the force or motion and the transduction principles which converts the mechanical variable into an electrical output. The mechanical variables of primary interest are displacement, strain, velocity and acceleration.

As mentioned in the previous section the emphasis will be primarily on the transduction principles. For strain transducers the sensing element and transduction principle may be one and the same. However for an accelerometer the sensing element is the seismic mass and the transduction principle may be strain-gauge, piezoelectric, servo, reluctance, potentiometric or photoelectric.

The transduction principles which will receive most attention are strain-gauge, semiconductor junctions (p-n, MIS, Schottky), and piezoelectricity. In some instances only a few serious attempts to apply the principle have been reported. However the information is included to provide insight into the development activities associated with solid state sensors and transducers.

3.1.1 Strain-Gauges

Mand solid-state materials and device properties are sensitive to strain. Historically, the change in resistance of a strand of wire provides the foundation for strain gauging and continues as a reference for assessing the performance of more recent strain gauge principles. It is not our intention to review the literature on wire resistance strain gauges for this is done in a most satisfactory manner. [2],[3],[4],[5],[6] However some of the properties of wire materials which are used in strain gauges will provide the basis for comparison. The major factors are strain

will provide the basis for comparison. The major factors are strain sensitivity, temperature effects, accuracy, linearity, noise, size, packaging, fatigue, hysteresis and circuit requirement.

3.1.1.1 Metal Wire Strain Gauges

The most often quoted feature of a strain gauge is the gauge factor. As discussed in Sections 1 and 2 the numbers can be confusing. This confusion arises primarily due to the lack of an accepted definition of gauge factor for a non-linear device. In this section, incremental gauge factors will be used and for gauges whose resistance changes linearly with strain the conventional gauge factor $\frac{\Delta R}{R_o \Delta \ell / \ell_o} = \frac{1}{R_o} \frac{\Delta R}{\varepsilon}$ and incremental gauge factor $\frac{1}{R} \frac{dR}{d\varepsilon}$ are the same.

The gauge factor for some representative wire compositions are listed in Table 1. The major feature is that the values do not vary over a large range. Considering only incremental changes in dimensions due to changes in strain the gauge factor for metal wire gauges can be calculated. The analysis yields a value of approximately 1.6. Including resistivity changes with strain accounts for most of the variations from 1.6 for the materials listed in Table 1.

Positive and negative gauge factors are possible and the sign depends upon the resistance change with strain. Since gauge factor is defined as $\frac{1}{R_o} \frac{\Delta R}{\varepsilon} = \frac{R - R_o}{R_o} \frac{\ell_o}{\ell - \ell_o}$, where R_o and ℓ_o are the unstrained values of resistance and length, the sign of the gauge factor will be determined by the ratio $\frac{R - R_o}{\ell - \ell_o} = \frac{\Delta R}{\Delta \ell}$. In tension $\ell > \ell_o$ and $\Delta \ell > o$. Therefore a positive gauge factor results from an increase in resistance with increasing tensile strain. The availability of negative gauge factors are limited for wire gauges, however for semiconductor strain gauges positive and negative gauge factors are available and can prove useful.

ALLOY	NOMINAL COMPOSITION	GAUGE FACTOR	RESISTIVITY (Ω m') $\times 10^{-8}$	TEMP. COEFFICIENT (per °C) $\times 10^{-4}$	THERMAL COEFFICIENT OF LINEAR EXPANSION $\frac{(\text{inch per °C})}{\text{inch}} \times 10^{-6}$
Constantan		2.1	48		16
Ferry	45 Ni, 55 Cu,	2.2	45	$\pm$0.2	13
Advance		2.1	45		15
Karma	75 Ni, 20 Cr, etc.	2.1	125	0.2	10
Nichrome	80 Ni, 20 Cr,	2.5	100	1	13
Iso-elastic	36 Ni, 8 Cr, 0.5 Mo, etc.	3.6	105	1.75	4
Alloy 479	92 Pt, 8 W	4.7	62	2.4	
Nickel		-12	6.5	68	13
Platinum		4.8	10	40	9

Table 1 Properties of Wire Materials for Use in Strain Gauges [2]

While the sensitivity to strain is of primary importance, variations in gauge resistance with temperature are equally important in most instances. Assuming a thermally controlled enviornment the primary concern would be loss of sensitivity. However, many measurements must be obtained in an environment where temperature changes must be tolerated. In addition there is the possibility of internal heating for large currents through the wire. Factors which must be considered are the temperature coefficient of gauge factor and the temperature coefficient of expansion. If the gauge is bonded to a specimen, the difference in thermal coefficient of expansion between the wire and the specimen will introduce a strain which will be temperature dependent.

It is difficult to summarize the temperature effects for wire strain gauges, however reasonable estimates of the expected behavior can be obtained. Representative values for the temperature coefficient of resistivity are included in Table 1. One cannot be absolutely sure what these values mean. As pointed out by Hines, manufacturers are concerned with tolerances for these alloys [3]. The value of $\pm 0.2 \times 10^{-4}$ ohm/ohm.°C for constantan defines an acceptable region of performance between 20°C and 100°C. Each melt of a particular alloy will have a different temperature coefficient of resistivity, therefore these specifications can represent only linear tolerance limits.

The gauge factor can also change with temperature. Assuming temperature variations can be reduced or controlled to an acceptable value, the change in gauge factor will result in a loss in sensitivity. A nominal value for the temperature coefficient of gauge factor is 10^{-6}/°C over approximately 200°C temperature range [2]. The temperature

coefficient of gauge factor is more complicated for a bonded gauge due to the temperature-strain behavior of the adhesive. Some aspects of this problem will be considered after introducing the concept of apparent strain.

Expansion of the wire gauge with increasing temperatures can be characterized by a thermal coefficient of linear expansion. Values for certain metal alloys near room temperature are given in Table 1. If a wire strain gauge made of the materials in Table 1 is bonded to specimen materials, one is also interested in the thermal coefficient of expansion for some common materials such as those shown in Table 2 [3].

Silica (Quartz Glass)	+ 0.5 ppm/°C
Stainless Stell	+ 11 $(\frac{\text{ppm}}{^\circ\text{C}})$
Aluminum	+ 23 ppm/°C
Aluminum Bronze	+ 17 ppm/°C
Magnesium	+ 25 ppm/°C

Table 2 Thermal Coefficient of Expansion for Some Common Materials [2]

Assuming the gauge is bonded to a specimen at a given temperature T, a strain is introduced in the gauge when the temperature is different from T, . This results in a bonded resistance different from the unbonded resistance and can present difficulties in achieving bridge balance without applied strain. In addition, this strain is temperature dependent and results in an apparent strain with temperature variations. As an example of the apparent strain due to a thermal coefficient of expansion mismatch consider a constantan gauge on a stainless steel specimen. The apparent strain would be -5 microstrain /°C.

Before summarizing temperature effects it is important to realize that much effort has been devoted to studying self-compensation techniques for temperature variations of strain gauges. In general any change in resistance not associated with the applied strain is unwanted. If the change in resistance ΔR is assumed to be due to 1) the applied strain ε, 2) temperature coefficient of resistance ρ_T and 3) apparent strain ε_t, a linear approximation to the change in resistance due to a change in any or all of the three variables listed is

$$\Delta R = \frac{\partial R}{\partial \varepsilon}\Big|_{\varepsilon=0}\Delta\varepsilon + \frac{\partial R}{\partial T}\Big|_{T=T_R}\Delta T + \frac{\partial R}{\partial \varepsilon_t}\Big|_{\varepsilon_t = \varepsilon_{T_R}}\Delta\varepsilon_t \qquad (3.1)$$

where T_R is the intended operating temperature and ΔT is a small variation about T_R and ε_{T_R} is the strain at the temperature T_R resulting from the bonding procedure. Defining $R_o = R(0, T_R, \varepsilon_{TR})$

$$\frac{\Delta R}{R_o} = \frac{1}{R_o}\left.\frac{\partial R}{\partial \varepsilon}\right|_{\varepsilon=0}^{\Delta\varepsilon} + \frac{1}{R_o}\left.\frac{\partial R}{\partial T}\right|_{T=T_R}\Delta T + \frac{1}{R_o}\left.\frac{\partial R}{\partial \varepsilon_t}\right|_{\varepsilon_{t=\varepsilon_{T_R}}}^{\Delta\varepsilon_t}$$

(3.2)

$$\cong GF \quad \Delta\varepsilon + \rho_T \Delta T + GF \quad \Delta\varepsilon_t \tag{3.3}$$

where the defintion of gauge factor $GF = \frac{1}{R_o}\frac{\Delta R}{\Delta \varepsilon}$ and the temperature coefficient of resistance $\rho_T = \frac{1}{R_o}\frac{\Delta R}{\Delta T}$ have been used. In addition we have assumed that the resistance of the gauge changes linearly with strain. The apparent strain ε_t is due to a mismatch in thermal expansion or

$$\varepsilon_t = \left(\frac{\Delta \ell}{\ell}\right)_s - \left(\frac{\Delta \ell}{\ell}\right)_g \tag{3.4}$$

where $\left(\frac{\Delta \ell}{\ell}\right)_s$ is the change in the linear dimension of the specimen and $\left(\frac{\Delta \ell}{\ell}\right)_g$ is the change in the linear dimension of the gauge. Using the defintion of thermal expansion coefficient $\alpha_s' = \left(\frac{1}{\ell}\frac{\Delta \ell}{\Delta T}\right)_s$ and $\alpha_g = \left(\frac{1}{\ell}\frac{\Delta \ell}{\Delta T}\right)_g$, the apparent strain for small temperature variations at T_R can be written as

$$\varepsilon_t = (\alpha_s - \alpha_g)\,\Delta T. \tag{3.5}$$

Therefore

$$\frac{\Delta R}{R_o} = GF \quad \Delta\varepsilon + \rho_T \Delta T + GF \quad (\alpha_s - \alpha_g)\,\Delta T \tag{3.6}$$

The results of Equation (3.6) suggests that self-compensation of

temperature effects should be considered. The basic idea is to balance the changes in temperature coefficients of resistivity against the apparent strain. The total unit change in resistance of a bonded gauge as a function of temperature with zero applied strain can be expressed as

$$\left(\frac{\Delta R}{R}\right) = \rho_T \Delta T + \varepsilon_t \quad GF \tag{3.7}$$

$$= \rho_T \Delta T + (\alpha_s - \alpha_g) \quad GF \quad \Delta T.$$

Ideally one desires a $\frac{\Delta R}{R}$ that is invariant with ΔT. Therefore

$$(\alpha_s - \alpha_g) = -\frac{\rho_T}{GF} \tag{3.8}$$

can conceptually provide the desired features.

Practically speaking this criteria can be met at only one temperature value and due to the non-linearity in these parameters the compensation may apply for only a limited temperature range typically less than 100°C span. As a quick check one might consider a constantan strain gauge and determine that self-compensation is practical near room temperature for specimens whose thermal coefficient of linear expansion is approximately 7.5 ppm/°C. At first glance this approach may appear naive since the temperature variations for these parameters are generally unknown. However, the apparent strain and bonding process have received considerable attention and are under sufficient control that temperature effects as low as ± 1 microstrain/°C are achievable for a moderate temperature range and a particular choice of gauge and specimen materials [3].

The above discussion is considerably simplified and one must resort to careful calibration and characterization of wire strain gauges to gain confidence in strain measurements. A major factor that has been mentioned casually is the bonding agent. One is interested in transmitting the true strain from the specimen to the wire gauge. Inadequate bonding gives rise to non-linearity, hystersis and fatigue for the strain gauge. To discuss each of these factors in detail is not our intention, however, Figure 3.1 is an example of the bonding problems one might encounter. Obviously the out of date adhesive is unacceptable. However the most important feature is the apparent curing of the adhesive during the initial test. In addition the gauge may exhibit non-linearities and hystersis as shown in Figure 3.2. This behavior leads one to question the accuracy of a strain gauge output. Therefore extensive and careful bonding procedures must be a integral part of any bonded strain gauge measurement [5].

Overall accuracy and sensitivity of strain gauge measurements can be misleading. Basically one is interested in a detectable signal output from a strain gauge and with what accuracy does this output represent the true strain. Neither of these concepts are easily characterized. In the first case, signal to noise of the strain gauge and read-out circuits are important. In the second case calibration and linearity of the resistance change with strain must be considered. Other factors which may be considered are self-heating of the strain gauge, ambient temperature variations, junction theromovoltaic effects in the lead wires and photovoltaic effects in some instances (mainly semiconductors).

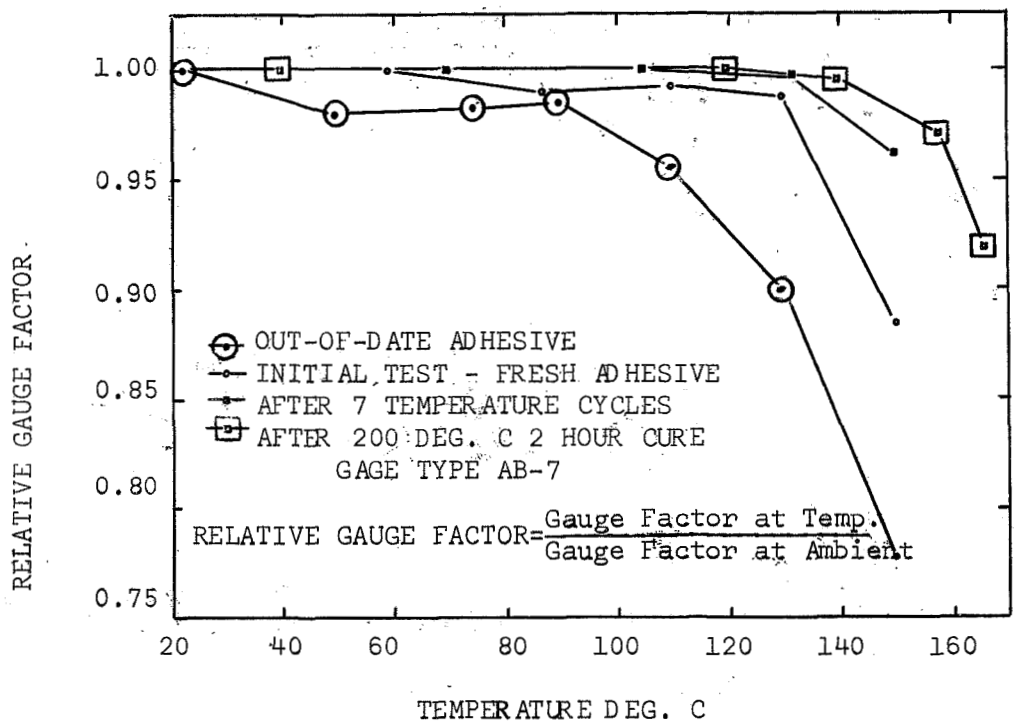


Figure 3.1 Temperature effects on gauge factor [4].

However, these later effects can be considered as experimental control and only the order of magnitude will be considered.

For a wire strain gauge the gauge factor is approximately 2. The end point linearity of the resistance change will be approximately 0.01% of full scale for strains from 0 to 1000 microstrains [7]. Therefore, assuming the gauge factor has been verified and hystersis has been reduced to an acceptable range the measured change in resistance is directly proportional to the strain where the constant of proportionality is $\frac{1}{G.F.R_o}$. Thus the remaining concern is how small a ΔR can be measured and will the measuring circuit introduce non-linearities.

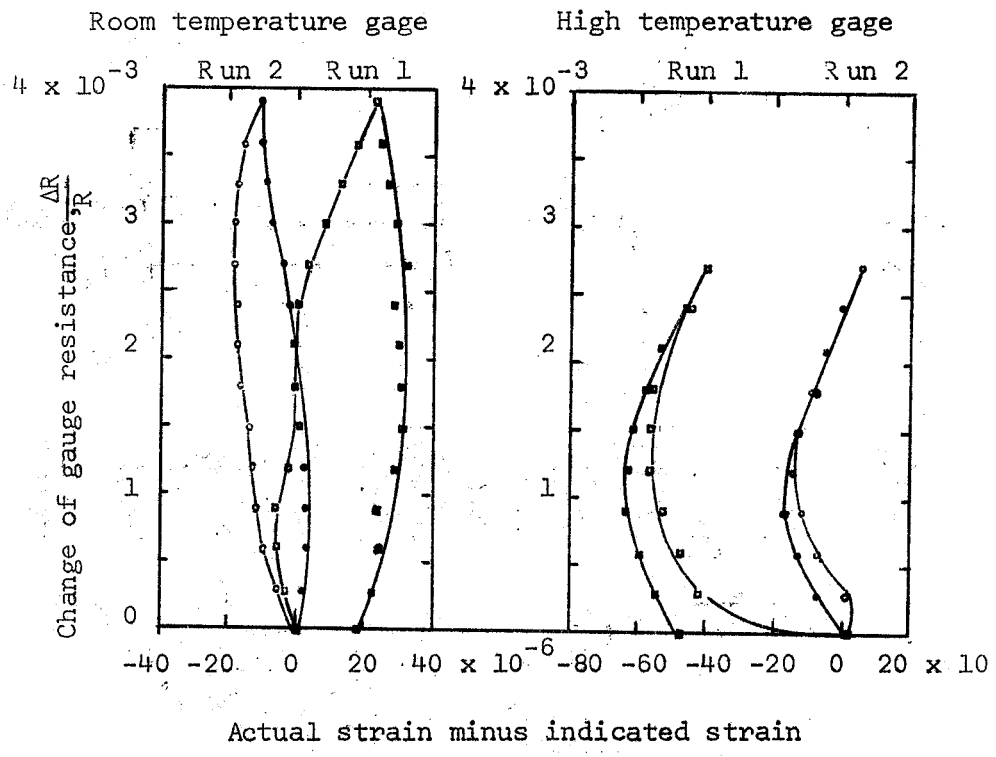


Figure 3.2. Nonlinearity of two gauges at room temperature [6].

Most often a bridge circuit is used to measure the resistance change. For discussion purposes consider the bridge in Figure 3.3 where only one element is a strain gauge. The voltage difference between points 1 and 2 is the bridge output and proportional to ΔR. If the bridge is initially balanced such that $V_o = 0$ and $R_2 R_4 = R_1 R_3$, the output voltage for a finite resistance change ΔR_1 is

$$V_o = \frac{V_i \Delta R_1 R_3}{(R_2 + R_3)(R_1 + R_4 + \Delta R)} \qquad (3.9)$$

To facilitate the discussion consider the case where $R_1=R_2=R_3=R_4=R$ and

$$V_o = \frac{V_i \, \Delta R/R}{2(2 + \Delta R/R)} \quad . \qquad (3.10)$$

It is obvious that although the change in resistance with strain may be linear, the bridge output may not be linear with ΔR. However for a gauge factor of approximately 2 and strains below 1% the difference between a linear response and the actual voltage output is approximately 1%. The situation is drastically changed for strain higher gauge factors such as semiconductors. This problem will be treated later. Therefore the bridge circuit will introduce no more than a 1% error in relating the voltage output measurement to strain assuming the output voltage can be measured with an accuracy better than 1%.

To measure the bridge output voltage to better than 1% for strains from 0 to 1% strain is impossible. To recognize the limitation consider Equation (3.10) and a gauge factor of 2. Then

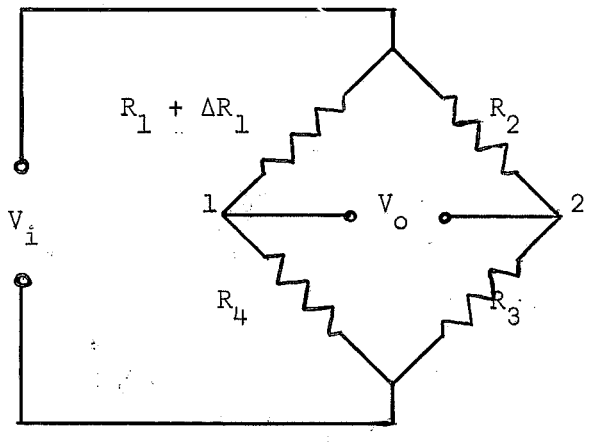


Figure 3.3 Resistance bridge.

$$\frac{V_o}{V_i} = \frac{\varepsilon}{2(1 + \varepsilon)} \qquad (3.11)$$

where

$(\frac{\Delta R}{R\varepsilon}$ = gauge factor) has been used and ε is the strain in inches/inch. For a 1% strain

$$\frac{V_o}{V_i} \simeq 5 \times 10^{-3}$$

and we must detect less that 1 part in 10^4 of the excitation voltage. In other words for a 1 volt excitation the output is 5 mv. Without concern for details this voltage level can easily be measured. In addition the impedance as seen by the excitation source is R and the current thru each branch of the bridge is approximately $\frac{V}{2R}$. Typically the wire gauge resistance will be approximately 100 ohms with a resistance tolerance of $\pm$ 0.5% and the current thru each element of the bridge is approximately 5 ma. The power dissapated in each element is approximately 2.5×10^{-3} watts. Thus 1% strain measurements are encouraging.

At strains approaching 10^{-6} numerous problems begin to affect the measurement. First consider an ideal situation where temperature variations can be ignored and one can precisely obtain an initial balance where all elements have identical resistance values. One is then faced with $\frac{V_o}{V_i} = 5 \times 10^{-7}$ for one microstrain. Using 1 volt excitation one must measure 1/2 microvolt. Using an instrument like the Keithley nanovoltmeter Model 148, the measurement appears promising. However, one must consider the voltmeter in some detail to assure that

operational procedures are met. For example the 1 μ volt range requires a source impedance (bridge impedance) less than 10^3 ohms while on the 0.1 microvolt scale a source impedance less than 100 ohms is required. Therefore for practical resistance values for the strain gauges, 1 micro-volt full scale appears to be a reasonable order of magnitude for a state of the art measurement.

Of course one must consider the non-ideal factors which may complicate or obscure the measurement. One factor is the noise input from the bridge to the measuring instrument. Johnson or thermal noise will predominate in the wire strain gauge. The rms noise voltage is given by

$$V_n = \sqrt{4\ kT\ R\ \Delta f} \qquad (3.12)$$

where k is Boltzman's constant, T is the absolute temperature, R is the resistance of bridge, and Δf is the overall bandwidth of the measurement. For a bridge with R=100 ohms and using the 1 micro volt scale on the Keithley instrument which has a bandwidth of approximately 1 hz the rms noise voltage is 1.29×10^{-9} volts. Therefore, noise will not be the limiting factor.

As usual, temperature effects play an important role in the measurement. As mentioned previously temperature compensation the order of 1 microstrain/°C is attainable. Using circuit compensation techniques approximately 0.1 microstrain/°C can be obtained. The details of how this is accomplished will add little to our discussion therefore one is referred to the literature [2],[3],[7],[8]. If the temperature

can be controlled to 0.1°C the apparent strain due to temperature variations will not limit the measurement assuming one is successful in achieving the ultimate compensation limits.

A factor which may seriously affect the measurement is thermal emf due to thermocouples formed by the metal-metal interface between the strain gauge and the lead wires. If the gauge is constantan and the leads copper, the thermal emf is approximately 40 micro volts/°C [9]. This voltage is due to thermal gradients and one must keep the temperature difference between the gauge and lead wires below 0.001 °C to obtain the measurement without inteference from thermal emf's.

Finally, the problem of initial bridge balance must be considered. If the measurement is to be an absolute value the bridge must be balanced such that the output voltage is much less than 1/2 microvolt with zero strain. To a first approximation the ratio of the resistance tolerance to the desired value of the resistance will be equal to the ratio of the bridge output voltage to the excitation voltage. Therefore for 0.5% tolerance for 100 ohm resistors and a 1 volt excitation the offset voltage or bridge output at zero strain would be 50 micro volts. Therefore one must obtain matched elements for the bridge or provide a zero adjust in parallel with one of the bridge elements. To balance the bridge to 1 part in 10^7 may prove to be a forminable if not impossible task therefore one could resort to measuring a change in the bridge output. The Keithley nanovolt D. C. Amplifier Model 140 is capable of resolving 0.1 micro volt changes about 50 micro volts. Therefore the measurement of 1 microstrain appears feasible.

The discussion to this point has been somewhat hypothetical and with justification. It is not easy to obtain estimates for strain gauge performance near 1 microstrain. Therefore, the discussion is intended as a summary of problems one encounters in defining the sensitivity of strain gauges. From this discussion detecting 1 micro strain appears feasible. In fact measuring one microstrain appears feasible. Although, 0.1 microstrain may be detected, measurement appears to be another problem. At least from the literature reviewed for this survey there appears to be a barrier at 0.1 microstrain.

As the strain level increases, 100 microstrain appear routine. One might be interested in an estimate of accuracy for routine measurements. According to Neubert [2] 5% accuracy for a routine, however carefully executed measurement is feasible for static measurements while 10% is given for dynamic tests.

3.1.1.2 Semiconductor Strain Gauges

The discussion of wire strain gauges is intended as a basis for comparison with semiconductor strain gauges. Most of the discussion pertaining to semiconductor strain gauges will be limited to unique features when compared to wire strain gauges. Possibly the most striking feature is the gauge factor. Whereas the gauge factor for wire gauges depended primarily upon dimensional changes with strain, the semiconductor strain gauge depends primarily upon a change in resistivity with strain. These strain induced changes are termed piezoresistance and have been treated in detail in Section 1. The salient features of the gauge factor can be identified in Figure 3.4.

In principal a wide range of gauge factors both positive and negative are available. The dependence on crystal orientation provides another variable which has proven useful.

Gauge factor variations with temperature in Figure 3.5 would be acceptable for many applications. Unfortunately the gauge factor is not the primary problem. In fact the gauge factor exhibits nearly a linear decrease with increasing temperature over the range from -50°C to + 125°C of approximately 0.2%/°C [10]. Therefore the gauge factor at 100°C would decrease approximately 15% from the 25°C value. If 2×10^{-4} ohm-cm p-type silicon is considered the gauge factor would be reduced from 120 at 25°C to approximately 100 at 100°C. If this were the only problem with silicon strain gauges one could truly be optimistic.

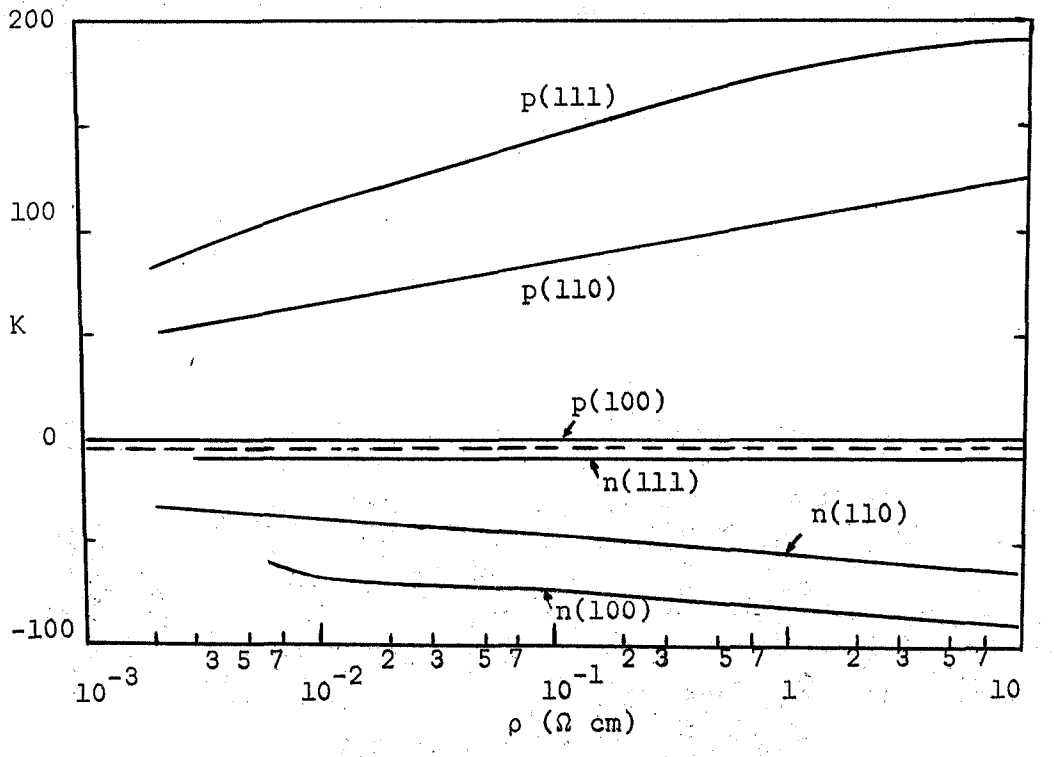


Figure 3.4. Gauge-factor K of p and n silicon for different crystal orientations as function of resistivity ρ [11].

However, the overall performance of a p-type silicon strain is not satisfactory in all respects. To illustrate the major difficulties consider the comparison of several properties of typical p-type silicon and constantan strain gauge wire in Table 3. For the present discussion the items of interest are the temperature coefficient of resistivity, the gauge factor and the linear thermal expansion coefficient. In self-temperature compensation is considered, one observes that p-type silicon does not offer much promise. The low coefficient of expansion ($\sim 4 \times 10^{-6} \frac{\text{inch/inch}}{^\circ\text{C}}$) and the high gauge factor practically eliminates compensation for apparent strain. Typical values of apparent strain from

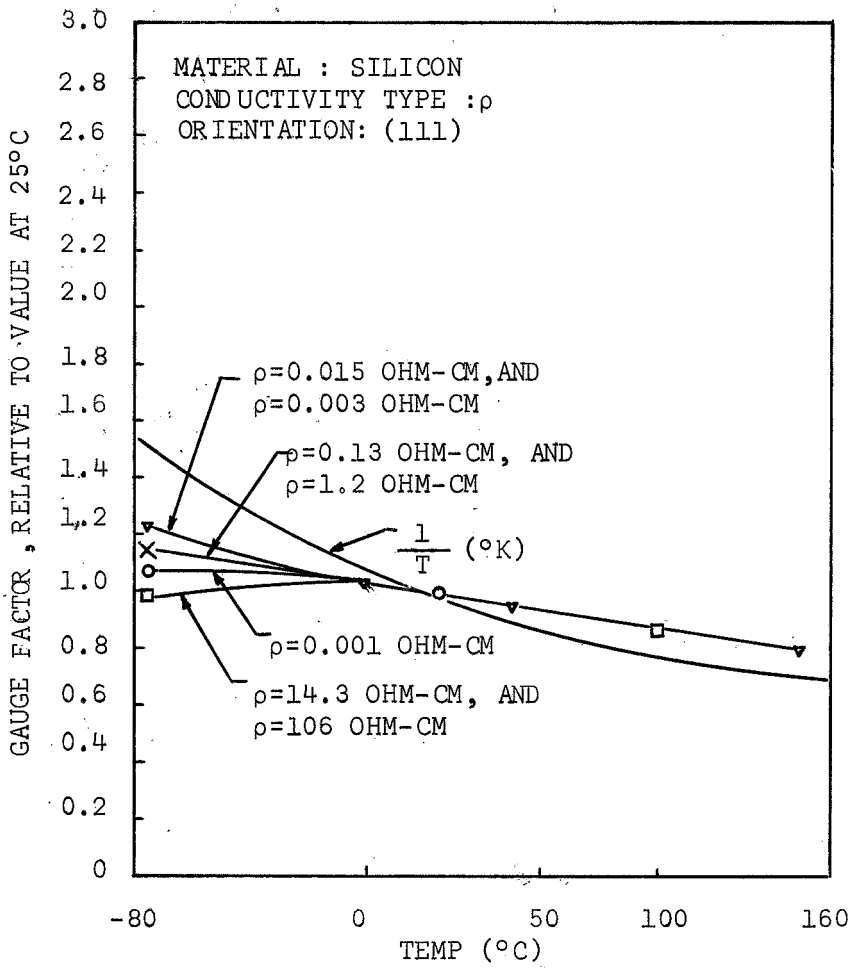


Figure 3.5. Gauge factor as a function of temperature [10].

Table 3 Properties of Silicon Strain Gauge and Wire Resistance Strain Gauge [11]

	Known Values		Semi-conductor strain gauges (silicon P [111])	Electrical resistance strain gauges (constantan)
Electrical properties	Gauge Factor		130	2
	Resistivity	10^{-5} Ω cm	1700	5
	Temperature coeff. of the resistivity	10^{-4}/°C	10	-0.24
	Apparent strain on steel	10^{-6}/°C	15.4	-15
	Resistance	Ω	100 to 10000	100 to 1000
	Linearity region	10^{-3}	-0.5 to + 3	± 100 to ± 200
Mechanical	Wire meterial		single-crystal semi-conductor	many crystalled metal
	Modulus of elasticity	10^3 kp/mm^2	19	15
	Max. elastic strain	10^{-3}	2 to 5	1 to 3
	Min. curvature radius	mm	1	0.1
	Linear Expansion Coefficient	$10^{-6}(\frac{\text{inch/inch}}{°C})$	4	16
	Number of load-cycles at ± 1 x 10^{-3} alternating strain		>10	>10
Geometric properties	Cross-section of the wire	mm^2	50 to 200 x 10^{-4}	3 x 10^{-4}
	Measuring length	mm	1 to 10	1 to 500
Miscellaneous	Common max. Measuring temperature	°C	250	320

50°F to 150°F and for 0.018 ohm-cm p-type silicon-bonded strain gauge are given in Table 3. Apparent strain of a gauge bonded to steel as a function of temperature is shown in Figure 3.6.

One of the major difficulties with semiconductor strain gauges is the large variation of resistance with temperature as shown in Figure 3.7. If a bridge circuit is used to detect resistance changes the unwanted thermally induced variations may be characterized by a linear thermal coefficient of resistance for small temperature variations. For 0.1 Ω-cm p-type silicon at 23°C the thermal coefficient of resistance is 40×10^{-4} /°C [32]. Therefore a temperature variation of 1°C results in a $\frac{\Delta R}{R_o} = 4 \times 10^{-3}$. For a full bridge with a 1 volt excitation, this thermal variation yields approximately a 4 millivolt signal. Clearly, this is unwanted and severely limits the sensitivity of the strain gauge.

The linearity of p-type silicon with a [111] orientation was a primary consideration in the early development of semiconductor strain gauge [10]. To characterize the resistance change with strain a power series of the form

$$\frac{\Delta R}{R_o} = F_1 \varepsilon + F_2 \varepsilon^2 + F_3 \varepsilon^3 + \ldots . \qquad (3.13)$$

is convenient. The resistance change is not symmetrical about zero strain as shown in Figure 3.8. However a region of linearity can be obtained to a prescribed accuracy. As one might expect the value of the constants F_1 F_2 & F_3 depends upon the resistivity of p-type silicon. Thus the range of linearity will vary with resistivity. Typical values for the constants F_1 and F_2 as a function of resistivity for p-type silicon are shown in Figure 3.9. For the 0.02 ohm-cm gauge considered

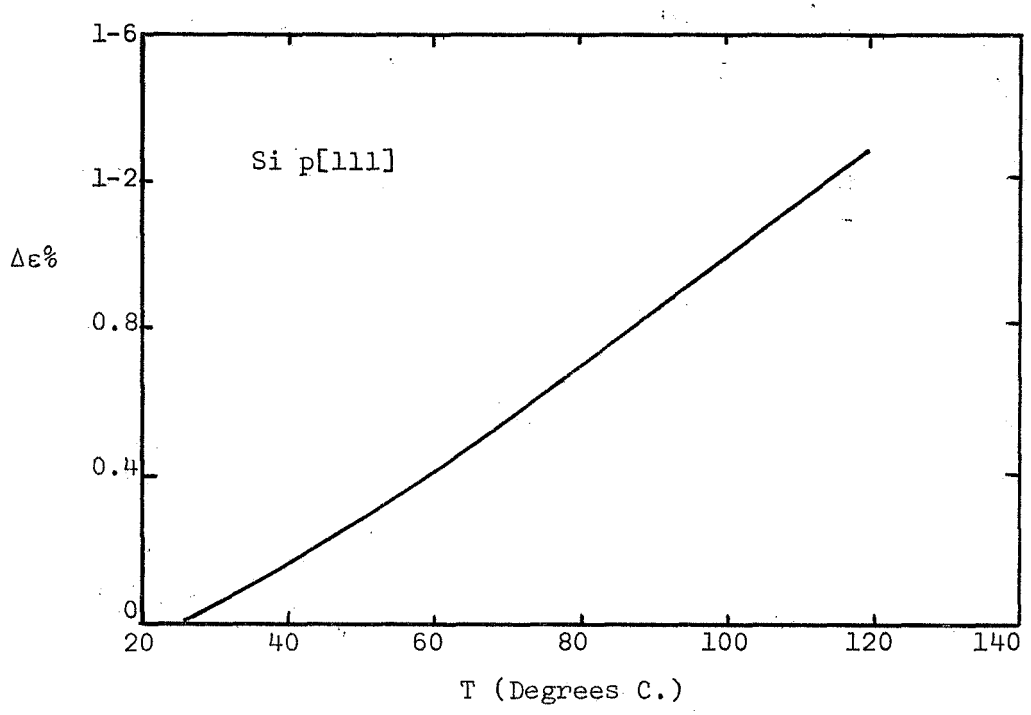


Figure 3.6. Apparent strain $\Delta\varepsilon$ of a gauge bonded to steel as function of the temperature T. [11]

above, the second order term is 1% of the linear term at approximately $\pm$ 300 microstrain. The non-linearity is clearly evidence in Figure 3.10 where the second order term is 10% the linear term and occurs at approximately $\pm$ 3 x 10^{-3} strain.

For a bonded gauge the region of linearity may be uncertain. If the bonding results in a compressive strain in the silicon with zero strain applied to the specimen, the resistance change with applied strain will vary about this built-in strain. For example the response in Figure 3.10 would be shifted along the horizontal axis by the amount of the built-in strain. Therefore, considerable caution must be excercised in the bonding process to insure a linear response even over a range of $\pm$ 100 microstrain. The concern for linearity both before

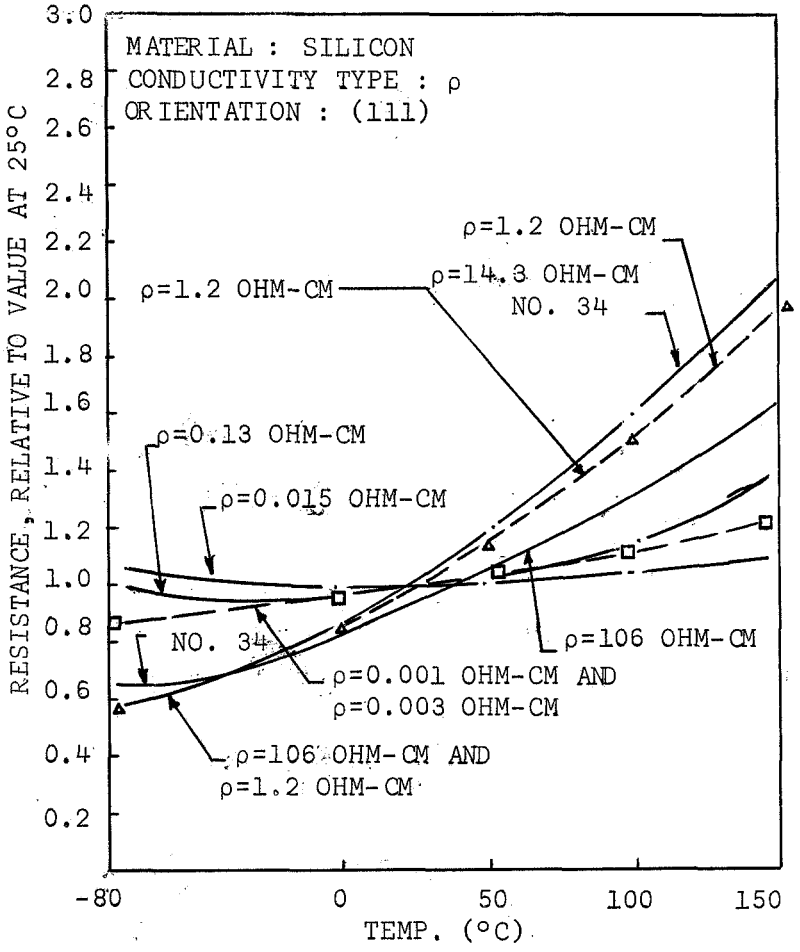


Figure 3.7. Resistance as a function of temperature [10].

and after bonding, led to the use of low resistivity p-type silicon the order of 2×10^{-4} ohm-cm. The reason for this choice can be easily understood if one estimates the linear range for a lightly doped p-type silicon. For 10 ohm-cm the second order term is 1% of the linear term at approximately 25 microstrain.

Non-linearities introduced by the detection circuitry can be characterized using the results in Equation (3.10). For example consider a p-type silicon gauge with a (111) orientation and a resistivity

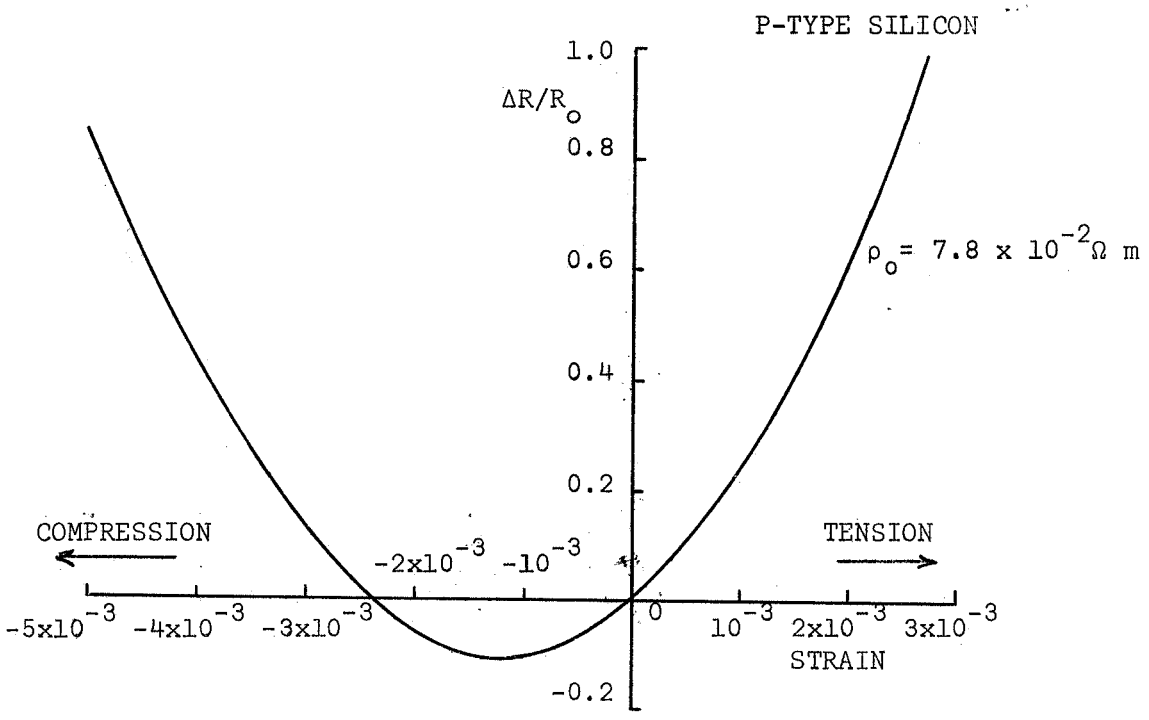


Figure 3.8. Relative resistance change as a function of strain of lightly doped p-type silicon at room temperature. [2]

of 2×10^{-4} ohm-cm. The gauge factor is approximately 120 near room temperature. Ignoring temperature variations and possible built-in strains the bridge output would be

$$\frac{Vo}{Vi} = \frac{60\ \varepsilon}{(2 + 120\ \varepsilon)} . \qquad (3.14)$$

If one assumes that the desired linear response is 30 ε, then the percent difference between the linear output and non-linear output at a particular strain is

$$\Delta\left(\frac{V_o}{V_i}\right) = \left(1 - \frac{1}{1 + 60\ \varepsilon}\right) \times 10^{+2}\%. \qquad (3.15)$$

A 1% difference occurs at approximately ± 170 microstrain. This should be compared to the value of ± 300 microstrain for the non-linearity of $\Delta R/R$. Since these non-linear effects occur at approximately the same value of strain they cannot be treated independently.

Returning to the relationship given in Equation (3.13) and using the appropriate values of F_1 and F_2 for 2×10^{-4} ohm-cm p-type silicon one obtains

$$\frac{V_o}{V_i} = \frac{30\,\varepsilon + 1000\,\varepsilon^2}{(1 + 60\,\varepsilon + 2000\,\varepsilon^2)} \qquad (3.16)$$

Assuming that 30 ε is the desired linear output, the percent difference of Equation (3.16) from linearity is

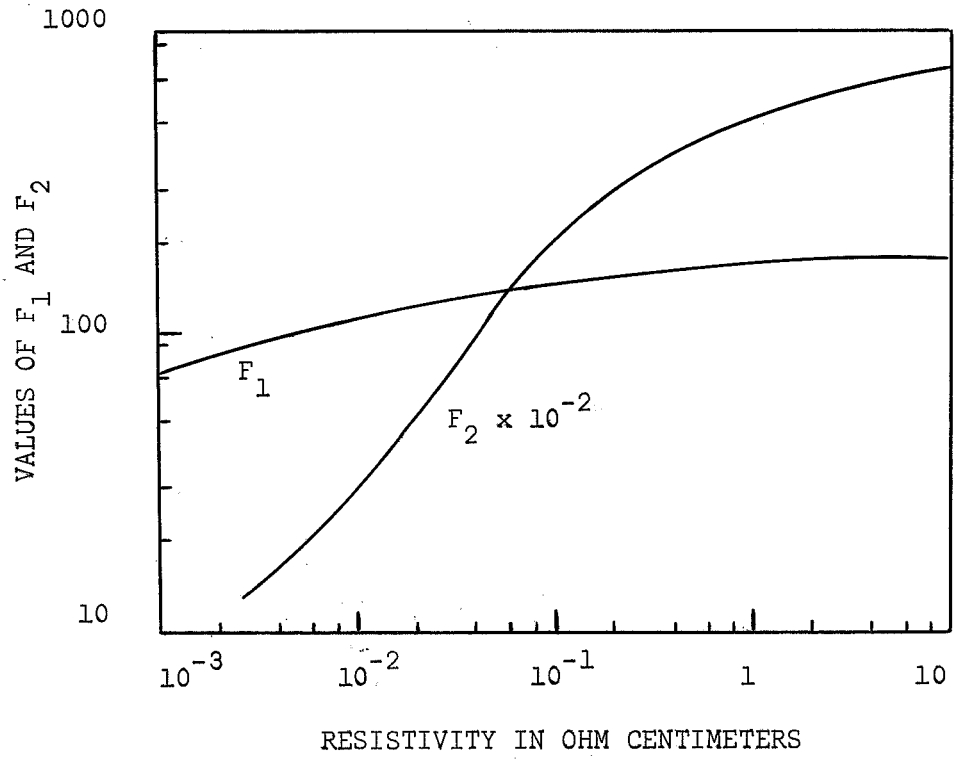


Figure 3.9. Typical values of F_1 and F_2 for unbonded p-silicon gauges [7].

$$\Delta\left(\frac{V_o}{V_i}\right) = \left(1 - \frac{1 + (100/3)\,\varepsilon}{(1 + 60\,\varepsilon + 2000\,\varepsilon^2}\right) 10^2\%. \qquad (3.17)$$

A 1% difference occurs with a compressive strain of -1.3×10^{-2} and with a tensile strain of $+ 3.8 \times 10^{-4}$. Unfortunately the compressive strain value is seriously in error because of the neglect of higher order terms in the polynomial approximation for $\Delta R/R$. However the tensile strain of 3.8×10^{-4} is an appropriate estimate for the range of linearity for the semiconductor strain gauge and uncompensated bridge output. Due to the self compensating feature of p-type silicon in a bridge circuit, the linearity is extended by at least a factor of 2.

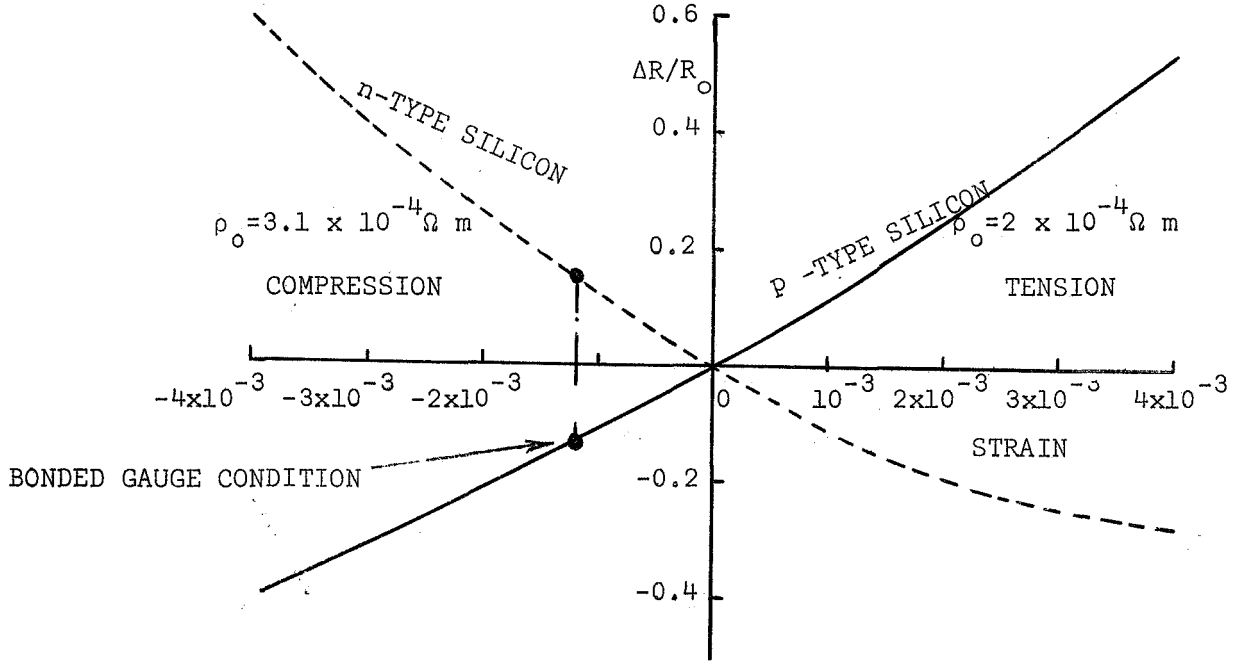


Figure 3.10. Relative resistance change as a function of strain of heavily doped p- and n-type silicon at room temperature. [2]

Material	Apparent Strain, μ"/"|°C
Quartz	2.3
InVar	3.4
1018 Steel	13.5
17-4 PH Stainless Steel	13.7
316 Stainless Steel	18.5
302 Stainless Steel	20
2024-T4 Aluminum	26.8
7075-T6	28.6
A-10 Magnesium	40

Table 4 Apparent Strain for 0.018 ohm-cm p-type Silicon Strain Gauge Bonded to the Listed Materials. [8]

Numerous methods have been devised to compensate for non-linear affects and for temperature variations for p-type semiconductor strain gauges. The various approaches depend upon additional elements in parallel or in series with the strain gauge and upon the method of bridge excitation with constant voltage or constant current. These compensation techniques will not be considered in detail in this report, however extensive treatments can be found in the literature [12],[13], [2],[7],[8],[14]. What we will briefly discuss is the success of these various techniques. Using compensation, the linearity range can be increased approximately an order of magnitude to $\pm$ 1.5 x 10^{-3} strain. [15] However, a sacrifice in gauge sensitivity is usually the trade-off. In addition temperature compensation can result in approximately an order of magnitude reduction in $\Delta R/R$ variations with temperature. However, these improvements do not guarantee the full potential of semiconductor strain gauges. The ultimate use of semiconductor strain must take full advantage of the increased sensitivity as well as the sophisticated technology and circuitry available with integrated circuits. A detailed discussion of these attributes will be postponed until later. The general position of our assessment to this point is that silicon p-type strain gauges provide a large signal output for a given strain thus a hope for increased resolution and sensitivity. However temperature effects limit sensitivity to approximately the same range as metal wire gauges with considerably more effort devoted to compensation. A 1 microstrain sensitivity is still an ambitions effort while a 1 microstrain resolution is more promising.

Although p-type silicon was preferred in the early 1960's, n-type silicon has received considerable attention in the interim. The major factor which focused attention on n-type gauges was the use of an n-type gauge to compensate non-linearity and temperature variations for a p-type gauge.

However, a careful examination indicates that n-type silicon can perform with p-type silicon with two possible exceptions. For a given resistivity the gauge factor for n-type silicon is approximately 1/2 that for p-type (111) silicon and the non-linearity in tension is excessive for some applications. The gauge factor variation with temperature, the resistance variation with temperature and the resistance variation with strain are shown in Figures 3.11, 3.12, and 3.13 for comparison with p-type silicon. Certainly, the resistance and gauge factor variations with temperature are similar for n- and p-type silicon.

N-type silicon can be used to compensate both non-linearity and resistance variations with temperature for p-type gauges. The compensation for p-type gauges is derived with a p-type gauge and n-type gauge in conjugate arms of a bridge. The negative gauge factor and positive temperature coefficient of resistance provide a rather unique opportunity. Consider the bridge arrangement in Figure 3.14. If the bridge is initially balanced (i.e. $R_1R_3=R_2R_4$), the ratio of the voltage out to voltage in to a first approximation is

$$\frac{V_o}{V_i} = \frac{R_3\,\Delta R_1 - R_4\,\Delta R_2}{(R_1 + R_4 + \Delta R_1)(R_2 + R_3 + \Delta R_2)} . \qquad (3.18)$$

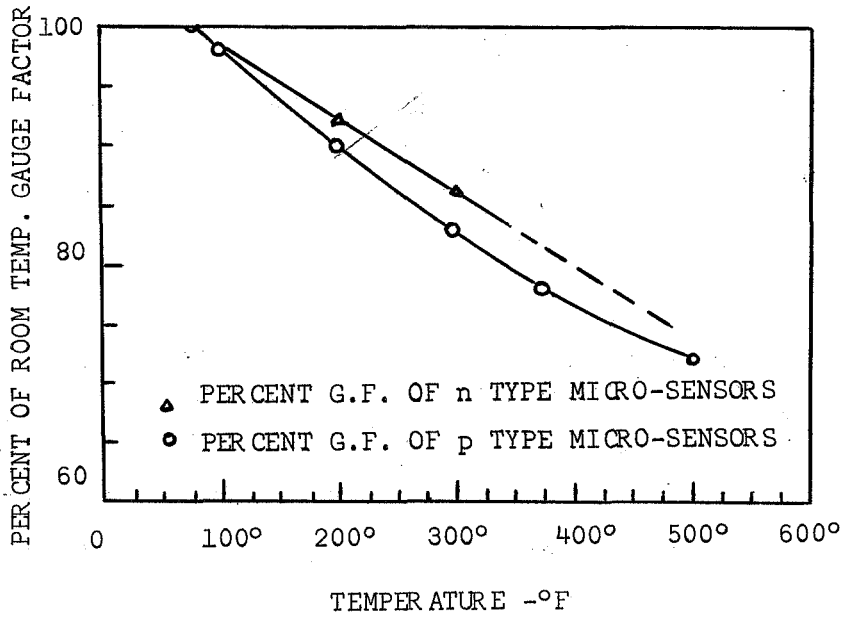


Figure 3.11. Gauge factor vs temperature for n- and p-types (unbounded) tested in tension. [15]

For convenience consider $R_1 = R_2 = R_3 = R_4$ however $\Delta R_1 \neq \Delta R_2$. This simply means that at the intended operating temperature and with zero applied strain all the resistance values are the same. Of course, there will be some bridge unbalance and this could be included. However the problem is not significantly different from bridge unbalance for wire strain gauges. Without loss of generality Equation (3.18) is reduced to

$$\frac{V_o}{V_i} = \frac{\frac{\Delta R_1 - \Delta R_2}{R}}{4 + 2\left(\frac{\Delta R_1 + \Delta R_2}{R}\right) + \frac{\Delta R_1 \, \Delta R_2}{R^2}} \quad . \qquad (3.19)$$

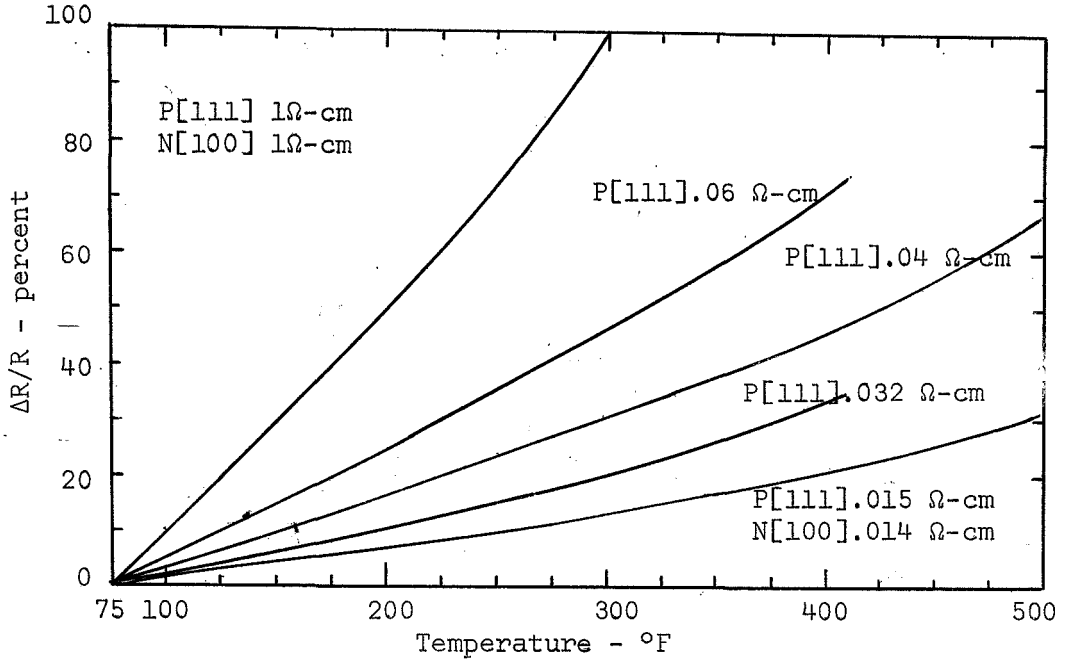


Figure 3.12. ΔR/R vs. temperature for selected resistivities and types of semiconductor elements. Note:Gauges are unbonded. [15]

The last term in the denominator can be neglected with respect to the middle term for most cases and

$$\frac{V_o}{V_i} = \frac{\frac{\Delta R_1 - \Delta R_2}{R}}{2(2 + \frac{\Delta R_1 + \Delta R_2}{R})} . \qquad (3.20)$$

A comparison with Equation (3.10) suggests the major contributions to compensation.

First the denominator term which produces a non-linear bridge response should be examined. For most instances build-in strain and temperature effects can be neglected and only non-linearities with strain need be considered. Thus $\frac{\Delta R_1}{R} = GF_1 \quad \varepsilon$ and $\frac{\Delta R_2}{R} = GF_2 \varepsilon$. Since R_2 is an n-type strain gage GF_2 is negative. Thus

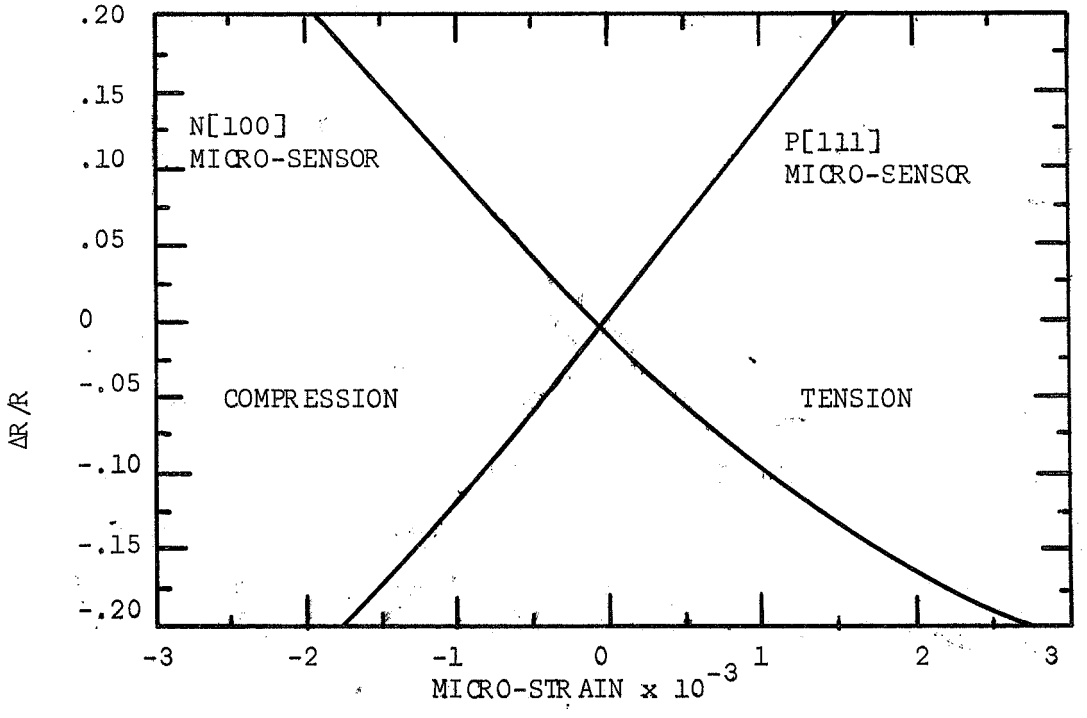
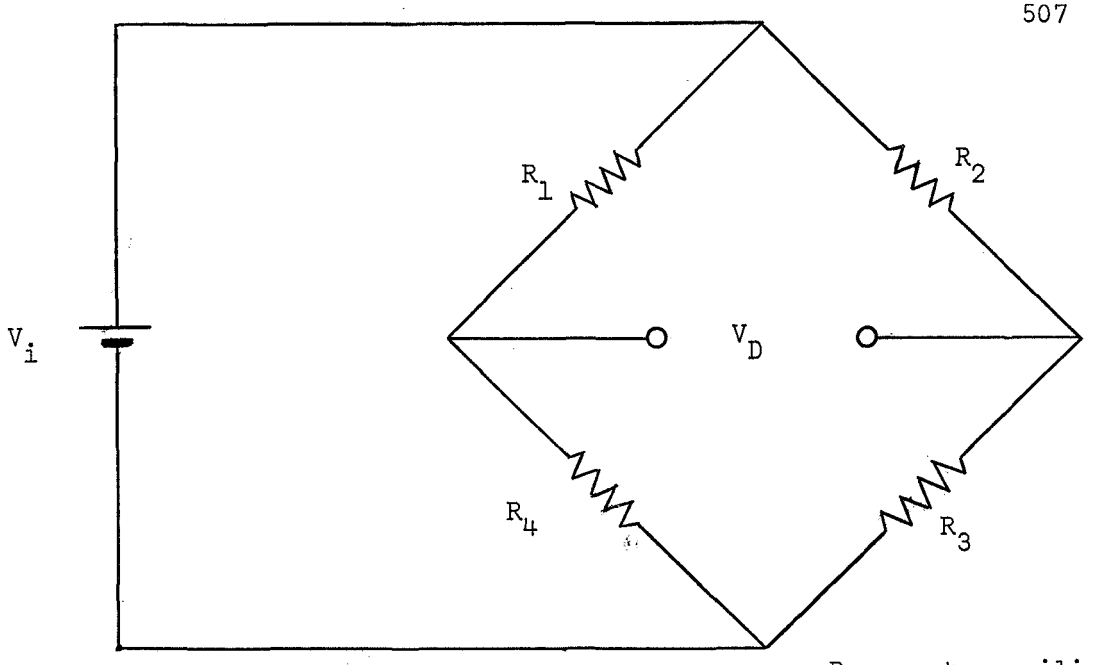


Figure 3.13. ΔR/R vs. strain for n- and p- types. [15]



R_1 = p-type silicon gauges

R_2 = n-type silicon gauges

Figure 3.14. Compensated bridge using p-type and n-type gauge elements.

$$\frac{\Delta R_1 + \Delta R_2}{R} = (GF_1 - GF_2)\,\varepsilon \qquad (3.21)$$

and the bridge non-linearity is reduced at high strain levels. The linear range would be extended by approximately a factor of 2 in our example since $R_1 = R_2$.

Secondly the numerator of Equation (3.20) clearly offers compensation possibilities. Considering the general change in resistance due to temperature coefficient of resistance, due to thermal expansion mismatch between the gauges and the specimen and due to strain

$$\frac{\Delta R}{R} = GF\,\varepsilon + \rho_T\,\Delta T + (\alpha_s - \alpha_g)\,GF\,\Delta T \qquad (3.22)$$

where GF is the appropriate gauge factor, ρ_T is the temperature coefficient of resistance, α_s (α_g) is the thermal expansion coefficient for the specimen (gauge) and ΔT a small temperature variation about the intended temperature. Using Equation (3.22) in Equation (3.20) and neglecting bridge non-linearity

$$\frac{V_o}{V_i} = \frac{(GF_1 - GF_2)\,\varepsilon + (\rho_{T_1} - \rho_{T_2})\,\Delta T + (GF_1 - GF_2)\,(\alpha_s - \alpha_g)\,\Delta T.}{4} \tag{3.23}$$

For self-temperature compensation one should attempt to satisfy

$$(\rho_{T_1} - \rho_{T_2}) = (\,GF_2 - GF_1\,)\,(\alpha_s - \alpha_g) \tag{3.24}$$

realizing that ρ_{T_1}, ρ_{T_2} and GF_1 are positive while GF_2 is negative in our example. Since $\alpha_s > \alpha_g$ and both are positive the right hand side of Equation (3.24) will be negative. Thus $\rho_{T_2} > \rho_{T_1}$ offers compensation. Fortunately, a low resistivity p-type silicon gauge which has rather respectable performance can be used with a higher resistivity n-type gauge for temperature compensation.

Thirdly, non-linearity compensation for gauge factor can be accomplished to a degree. To illustrate the compensation consider the resistance change due to strain for p-type silicon as

$$\frac{\Delta R}{R} = F_1\,\varepsilon + F_2\,\varepsilon^2 \tag{3.25}$$

and for n-type silicon

$$\frac{\Delta R}{R} = - F_3 \varepsilon + F_4 \varepsilon^2. \qquad (3.26)$$

The above relationships are for illustration and higher ordered terms should be considered for detailed results. However, the salient features can be developed by considering Equations (3.25) and (3.26) in the strain dependent portion of Equation (3.23). Neglecting the apparent strain

$$\frac{V_o}{V_i} = \frac{(GF_1 - GF_2)\varepsilon}{4} = \frac{(F_1 + F_3)\varepsilon + (F_2 - F_4) \varepsilon^2}{4} . \qquad (3.27)$$

Therefore the second order term is reduced by the conjugate bridge arrangement of the p-type and n-type gauges. In addition the apparent gauge factor is increased.

In general the measurement problem with semiconductor gauges are similar to those with wire strain gauges. In principle, the limiting factor in determining sensitivity is noise. Again Johnson or shot noise is expected although with semiconductors "1/f" or "excess noise" is a possibility. The estimate of shot noise in Equation (3.12) is approximately 10^{-9} volts for a gauge resistance of 100 ohms, at 23°C and a 1 hz bandpass in the measurement. The contribution of "1/f" noise can exceed shot noise especially below 10^3 Hz. Since "1/f" noise correlates with surface affects it is difficult to generalize. However, 10^{-8} volts or greater provides a very crude estimate for the noise voltage from a semiconductor strain gauge element of 100 ohms at 23°C and a 1 hz bandpass for the measurement.

A variety of bridge circuits have been proposed for detecting the resistance change with applied strain and to compensate for linearity and temperture affects [7]. Detailed analysis of bridge design will not be considered in this survey. However, one factor which relates to bridge design should be mentioned. To this point, the discussion has considered discrete semiconductor strain gauge elements. With the sophistication of the planar technology, diffused strain gauges offer some interesting possibilities. In transducer design, bonding the strain gauge to a sensing element such as a diaphragm may limit the transducer performance. For example the bonding agent may not withstand temperatures in excess of 200°C without fatigue, creep or failure. In addition the thermal expansion mismatch must be considered where sensitivity is a primary concern. With diffused strain gauges in a silicon diaphragm the bonding problem is eliminated, at least for the strain gauge. A transducer can now be designed which depends almost exclusively on the housing and diaphragm. Thus the emphasis can be shifted from the transduction element to the sensing element.

3.1.2 Junction Strain Gauge

A variety of strain sensitive junction devices have been investigated. [16],[17],[18],[19] Many of the salient features have been considered in Section 2. In this section, junction devices will be considered as transduction principles. Operational characteristics, advantages, disadvantages and unique features of the various junction devices will be discussed.

3.1.2.1 P-n Junction (Piezojunction)

The piezojunction effect is a rather general transduction principal for mechanical sensors. As described in Section 2 the response of p-n junctions to stress is evidenced in various ways depending upon the junction configuration, mode of operation and stressing technique. The major feature of the piezojunction effect is that at a constant voltage applied to a p-n junction the current changes exponentially with strain in the junction region. As discussed in Section 2.15 the sensitive region of operation is for stress levels in excess of 10^9 dynes/cm^2. For practical reasons, uniaxial stresses are encountered. These high stress levels dictate unique constraints for using the piezojunction effect.

First, the sensing element must provide a mechanical advantage if the mechanical stimuli are to be realistic. For example a pressure of 1000 atmospheres is approximately the equivalent of 10^9 dynes/cm^2. Also, one gram force is 980 dynes and requires a concentration of this force over an area of 10^{-6} cm^2 to achieve a 10^9 dynes/cm^2 stress. This is roughly a strain of 0.1 percent or 1000 microstrain in silicon. There are at least five approaches currently used for achieving these strains levels in silicon p-n junctions:

1) A phonograph needle pressed near the junction of a planar device,

2) A mesa device or multiple mesa device stressed by a post or column,

3) A semiconductor needle with a p-n junction on the apex stressed by pressing the needle against a hard surface,

4) A thin ($\sim 5 \times 10^{-3}$ cm) semiconductor chip, containing the junction bonded to a metal bar which is stressed, and

5) Bending bulk material, containing the junctions in configurations such as cantilevers, diaphragms and beams.

The first and second technique have been used in transducers. However, these transducers suffer two major disadvantages. They are difficult to manufacture (identor point alignment near the junction is required) and they are fragile. The third technique eliminates the alignment however the fabrication of the p-n junction in the tip of the needle is not a routine technology. The fourth technique is akin to the strain gauge in that bonding severly limits the use of large strains. While the fifth technique eliminates the bonding problem, bulk silicon is subject to low stress fracture.

The piezojunction effect has certain attributes for transducer applications. These are

1) It utilizes a well developed technolgoy.
2) The physical mechanisms are well understood.
3) It results in small size and low power transducers.
4) It has a differential current gage factor of approximately 400 for silicon.
5) It can result in high resonant frequency devices.
6) It responds to static as well as dynamic strains.

The major disadvantage of the piezojunction effect are the temperature sensitivity of the p-n junction device and the bias stress that is required to obtain a large differential current gauge factor. In the later case the device must operate near the fracture strength of the material and results in fragility.

The piezojunction effect in silicon diodes and transistors has been used in transducers [20],[21]. However the problems encountered in achieving the transduction are not clearly identified when compared to strain gauges. For example, it is well known that the piezojunction

effect with stress in non-linear. However the degree of non-linearity depends not only upon the operating range of strain but also the bias strain. This certainly adds an additional variable to the transducer design.

Another factor which has not been considered to our knowledge is the noise level in any of the p-n junction devices. To estimate the order of magnitude consider a p-n junction with a bias stress of σ and a small sinusodidal stress $\Delta\sigma$ about σ_I that we wish to detect. The current change due to $\Delta\sigma << 10^9$ dynes/cm^2 for a (100) uniaxial stress in a silicon p-n junction is

$$\frac{\Delta I_s}{I_{so}} = A\Delta\sigma \text{ , where } I_{so} \text{ is the diode current at the bias [22]} \qquad (3.28)$$

stress and ΔI_s is the change in current due to a change in stress. Using an estimated value for $A = -1.9 \times 10^{-11}$ cm^2/dyne yields for a tensile strain $\Delta I_s = -1.9 \times 10^{-11} I_{so} \Delta\sigma$. The mean-square value of a noise current due to noise current generator is [23]

$$\overline{I_n^2} = 2q\, I_d\, \Delta f \qquad (3.29)$$

where I_d is the diffusion or drift current of a p-n junction, q the electronic charge and Δf the bandwidth of the system used to detect the current. Here we have assumed that the junction current is the "ideal" "current". If generation-recombination current predominates the form for $\frac{\Delta I_s}{I_s}$ and $\overline{I_n^2}$ would change however the general result is similar. The I_d in Equation (3.29) is the junction current at the

bias stress and the RMS noise current is

$$\langle I_n \rangle = (2q\, I_{so}\, \Delta\, f)^{1/2}. \tag{3.30}$$

If $\Delta\sigma$ is a sinusoidal stress ΔI_s will be a sinusoid for $\Delta\sigma < 10^9$ hence the RMS value is

$$\langle \Delta I_s \rangle = \frac{9.5 \times 10^{-12}}{\sqrt{2}}\, I_s .\, \Delta\sigma. \tag{3.31}$$

Therefore the signal to noise ratio is

$$\frac{S}{N} = \frac{9.5 \times 10^{-12}\, I_{so}\, \Delta\sigma}{(4q\, I_{so}\, \Delta f)^{1/2}} . \tag{3.32}$$

Typically the current will be measured with a system bandwidth less than 10^3 Hz and the stress values of interest are between 10^7 and 10^9 dynes/cm^2. The diode current at the bias stress may range from 0.1 to 10 ma. Using the worst case estimates of $\Delta f = 10^3$ Hz, $I_{so} = 10^{-4}$ amps. and $\Delta\sigma = 10^7$ dynes/cm^2, the signal to noise ratio is

$$S/N = 3.7 \times 10^{-4} \left(\frac{I_{so}}{\Delta f}\right)^{1/2} \Delta\sigma \simeq 37. \tag{3.22}$$

therefore, it appears that shot noise will not limit the piezojunction effect.

However "1/f" noise may present a problem especially for low frequency tranducers. The "1/f" contribution usually predominates below 10^3 Hz and the above estimate provides a best case. In fact, noise power measurements at 100 Hz indicate that the "1/f" contri-

bution can be 10 times the shot noise contribution [24]. However, it has also been shown that noise generation depends upon the device characteristics which depend upon surface effects [25]. Therefore it is difficult to estimate the contribution of "1/f" noise to total noise power. That the total "1/f" noise power could be at least and order of magnitude greater than shot noise for a frequency range of 0 to 10^3 Hz does not appear unreasonable. Therefore the results of Equation (3.33) could be overly optimistic.

Another factor which does not appear in most discussions of transducers utilizing the piezojunction effect is the unique housing and packaging problems resulting from the need for a bias stress. To illustrate the problem consider a diaphragm pressure sensor where the tranduction element is a p-n junction. The deflection of a clamped diaphragm with differential pressure variations across the diaphragm has been extensively investigated and is well understood. However, if the piezojunction is a semiconductor needle mounted as shown in Figure 3.15 the diaphragm is contrained at the center. In addition the bias stress maintains a tensile strain at the bonded interface between the diaphragm and the housing. Surely these factors complicate the packaging problem. In fact they also complicate the reliability problems. However, in the tradition of bonded strain gauges, piezojunction transducers can be realized although there limitations are not understood.

Packaging problems are under investigation and considerable progress has been made. The anodic bonding technique may represent the most significant factor in transducer design since the diaphragm [26].

In addition piezojunction elements may be planar devices in a silicon diaphragm which eliminates another problem. However, it is much too early to refer to these accomplishments as state-of-the art.

Finally, it should be recognized that one of the most unique features of the piezojunction is its size. Whereas strain gauges are the order of millimeters a piezojunction can be the order of microns. Therefore the piezojunction can provide additional spatial resolution. In addition the piezojunction transduction element provides greater flexibility in the geometry of the sensing element. For example, a diaphragm the order of millimeters in diameter are now feasible with the transduction occuring over only a small portion near the center thereby taking full advantage of the diaphragm transducer.

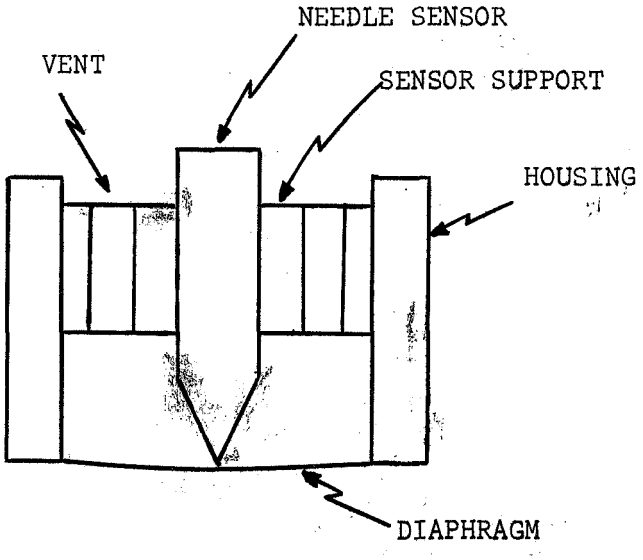


Figure 3.15. Pressure transducer configuration utilizing the needle sensor. [20]

3.1.2.2 Tunnel Diode

The previous section is intended as a rather general discussion of p-n junctions however it is difficult to include all p-n junctions in a single catagory. Tunnel diodes are unique in some respect and warrant a separate but brief discussion. The detailed discussion of stress affects included in Section 2.1.5 point to some of these unique features.

In general all the advantages of the p-n junctions are applicable to tunnel junctions with one reservation. The technology of tunnel diode fabrication is well developed in the sense that devices can be manufactured. However silicon tunnel diode technology relative to silicon planar diode technology is rather in its infancy. In addition, promising tunnel diode transducers are either germanium, gallium arsenide or gallium antimonide. The technology of etching, diffusing, contacting and passivating these materials is not nearly as advanced as for silicon. However, for research purposes one can consider the technology as adequate.

The performance of tunnel diodes can be somewhat confusing. Gauge factors are reported from 100 to 30,000 [27]. Gauge factors for the tunneling currents in silicon and germanium are given by Kiggins and Milnes for hydrostatic stress [28]. Values range from -80 to -160 for germanium and from 20 to 80 for silicon. Gauge factors for uniaxial stress would be approximately 1/3 these values. Using the stress sensitivity given in Section 2.1.5 for gallium antimonide and bulk modulus of 0.56×10^{12} dynes/cm^2 results in a differential current gauge factor for hydrostatic stress of approximately 300. Again uniaxial stress will result in approximately 1/3 this value. The gallium antimonide gauge factor is the largest for tunneling currents.

The confusing factor arises when the tunnel diode is operated in the amplifier mode and the current output is measured with changes in strain. Sikorski discusses this mode of operation and defines a differential voltage gauge factor [29]. A gauge factor for silicon tunnel diode of approximately 30,000 is given. One should be careful in the interpretation of this result. In reality the peak current sensitivity has been amplified by approximately the gain of the tunnel diode amplifier or approximately 1000. Therefore one should not compare the gauge factor of 30,000 to that of 100 for piezoresistance. A realistic comparison would be minimum detectable stain or signal to noise at a given strain value.

One very significant advantage of the tunnel diode is the temperature sensitivity. Kiggins gives a tunneling current temperature coefficient of 4×10^{-4}/°C and a temperature coefficient of gauge factor of less than 12×10^{-4}/°C [28]. These values are for a germanium tunnel diode with a gauge factor in excess of 125 for hydrostatic stress. These are very respectable characteristics and should provide encouragement for transducer design using tunnel diodes.

3.1.2.3 Other Junction Devices

The stress sensitivity of junction devices could be a lengthly discourse, however there are some limitations in terms of unique devices and unique transduction of strain. For example, Section 3.1.2.1 discusses the p-n junction in a general way. There is a purpose in that a single p-n junction, a transistor, and other multiple p-n junctions are basically the same. The piezo-transistor is a stress sensitive p-n junction which provides a stress sensitive current that

is amplified by the transistor action.

However, there are at least two other types of junctions that should be mentioned. The two types are heterojunctions, and metal-insulator-semiconductors. Whereas the previous discussion in Section 3 has concerned transduction principles either available in commercial transducers or considered in advanced development, the junctions mentioned above must be considered as research projects.

Of the two junctions, the heterojunction provides a new dimension in junction strain gauges. Whereas all the junction devices considered are based upon single crystal technology, the heterojunction devices utilize thin film technology. However, the new flexibility will not be cheap. Thin film active devices have proven illusive in terms of reproducibility, reliability and active device performance. However, the hetrode strain sensor reported by Moore [30] offers some unique advantages provided the technology can be established.

Strain sensitivities as large as 500 volts/unit strain are reported while a linear voltage-strain characteristic indicates a voltage gauge factors near 200 [30]. In addition, differential resistance gauge factors between 600 and 10^3 were observed. These gauge factors represent the largest values known to the authors. So that we are not misunderstood, we have compared only differential gauge factors for primary stress effects. Amplification with tunnel diodes or piezo transistors is not a primary stress effect.

One additional feature which should be considered is the flexibility resulting from thin film devices. With other junction devices the strain transducer is a bulk material which must be bonded to the element

which senses the mechanical variable, such as a diaphragm. With thin film devices the strain transducer can be deposited on a variety of substrates and possibly eliminate the problem of bonding. In principal one has a wider choice of diaphragm material since the stiffness of the gauge does not limit the diaphragm performance.

For the metal-insulator-semiconductor junction there are two major features that should be considered. First, the gauge factor can be misleading due to the gain of the field effect transistor. Therefore one should consider the primary stress effect of the piezoelectric portion of the device and the amplification separately in evaluating the characteristics. Thus the maximum gauge factor of 10^4 is somewhat misleading [31].

However, there is one major advantage of the piezoelectric field effect transistor strain transducer over a piezoelectric transducer. The influence of the change in polarization in the insulator with static stress can be measured as a change in channel resistance for the piezoelectric FET. Thus static as well as dynamic strains can be detected. Conventional piezoelectric devices are basically dynamic transducers. However the deciding factor of performance and realiability have not been proven with the piezoelectric FET and one must await further research and development in this area before critical assessment is possible.

3.1.3 Thin Film Strain Gauges

Thin film technology is a very important part of solid-state technology. In principal it offers the ultimate in obtaining various materials with varying electrical properties. However, the promise

has not been entirely realized. Possibly, the most advanced thin film technology is for metals. Much of the activity has been concerned with contacts to semiconductors. However, thin film strain gauges have been considered [33].

Realistically thin films of metal should be compared to metal foil gauges. However the importance of thin films as strain gauges are uncertain at best and the differences between rolled foils and evaporated thin films is not documented. Therefore we have chosen not to introduce metal-foil
addition the discussion concerning thin films will be restricted to a few isolated examples to merely assess activity in this rather broad area.

There appears to be little reported activity on evaporated thin film gauges. One cannot help but wonder as to the reasons why so little appears in the open literature. Considering the history of thin films the most logical reason would be the inability to control the properties of the thin films. However foils appear to be acceptable. This paradox cannot be resolved in this document. With the new evaporation techniques such as electron-beam and ion sputtering, the possibility of thin film strain gauges may not be quit so remote. This is especially true for alloy systems which are not suitable for processing as a foil or wire.

Although the above discussion is merely speculation, thin film semiconductors strain gauges have been realized using high-vaccum evaporated germanium [34]. Table 5 gives a rough comparison of the characteristics of the the thin film strain gauge to a bulk semiconductor gauge and a metal wire gauge. The at random crystal orientation for the micro crystals

Characteristics Gauge	Gauge factor	Linearity 1000 X 10^{-6}%	Temperature coefficient of resistance %/°C
Metallic resistance wire strain	2	0.1	0.002
Semiconductor bulk strain	100	1.0	0.2
Germanium-film strain	30	0.2	0.02

Table 5 Comparison of Deposited Semiconductor, Resistance Wire and Bulk Semiconductor Strain Gauges [34]

which are likely to occur in an evaporated film of germanium leads to self-temperature compensation of the gauge resistance and gauge factor. Also the linearity appears to be acceptable to 1000 microstrain. Even though the characteristics are acceptable and the thin film process quite flexible, evaporated thin film strain gauges are not commonplace. Again we must wonder as to the reasons.

3.1.4 Piezoelectric Transduction

All of the transduction principles discussed thus far are essentially the same. Strain modulates the conductivity of the material or device. However, the piezoelectric effect transforms a mechanical force into electrical polarization. In turn the electric polarization

results in opposite polarity of charge at two surfaces of the piezoelectric element. In most instances this charge is measured as a voltage across the piezoelectric element as a capacitor.

A major advantage of piezoelectric transduction is the high mechanical input impedance which results in negligible motion under a load. Thus the notion of force transduction is reasonably accurate. The stiffness of the piezo-element results in high natural mechanical resonant frequencies. In fact, piezoelectric elements are operated at frequencies as high as 10^9 Hz [35],[36]. For mechanical transducers the natural frequency of the piezo-element is likely to be less than 10^6 Hz. However, the mechanical resonance of the housing and contacting is likely to reduce the resonance to $10 - 50 \times 10^3$ Hz.

The major disadvantages of piezoelectric transduction are 1) the lack of steady-state response 2) the high electrical output impedance and 3) maximum operating temperature is likely to be near 200°C. In addition the sensitivity of piezoelectric ceramics may degrade with time.

The material properties which characterize piezoelectric transduction are included in Section 1.5.5. Of fundamental importance is the efficiency with which the mechanical energy is transformed into electrical energy. The coupling co-efficient k can be used to estimate the sensitivity of the piezoelectric transduction element. Typically these values range from 10% for quartz to 76% for rochelle salt. As shown by Wilner [38] the signal power output P_s from a self-generating transducer for a sinusoidal mechanical input with a peak mechanical energy E_p is

$$P_s = 2 \Pi f k^2 E_p \qquad (3.34)$$

where f is the frequency of the mechanical input. Of course we have assumed the transducer to be a linear system and have ignored many problems that may reduce the sensitivity such as housing design, noise, cable capacitance and environmental effects. Nevertheless, the idealized sensitivity is proportional to k^2 for frequencies well below the natural resonances of the piezo-element. In addition the increasing sensitivity with increasing frequency provides an indication that self-generating systems are ideally suited for time varying mechanical stimuli whereas strain gauge sensitivities are independent of frequency for the ideal case. Further comparison of sensitivities will be postponed until Section 3.1.6.

From Equation (3.34) the major disadvantage of piezoelectric transduction is quite clear. There is essentially no d.c. response. In addition the output signal is developed across the capacitance of the piezo-element and the measuring circuits must have a relatively high input impedance to provide a detectable voltage signal level. These factors can be illustrated by a simple equivalent circuit in Figure 3.16, where the charge q is developed across the capacitance C of the piezo-element, C_s is the stray capacitance (including housing and cables) and R is the input impedance of the voltage measuring instrument. The complete solution for the voltage ν with a sinusoidal charge variation $q = Q_o \sin \omega t$ is

$$\nu = V_1 \sin(\omega t + \tan^{-1} \frac{1}{\omega RC_t}) - V_2 e^{-t/RC_t t} \qquad (3.35)$$

with

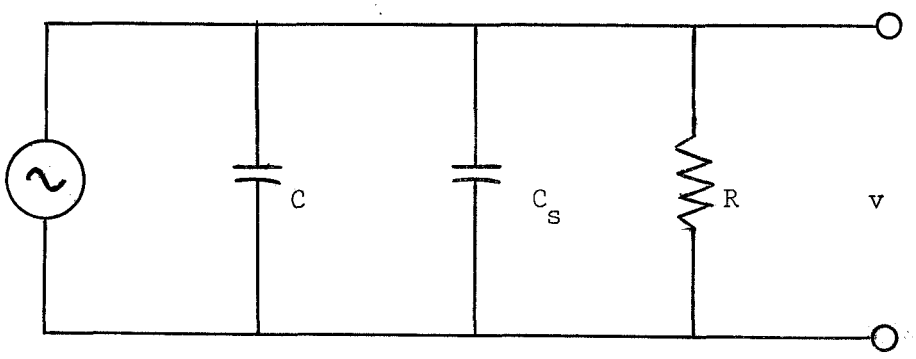


Figure 3.16. Simple equivalent circuit for piezoelectric transduction.

$$V_1 = \frac{Q_o/C_t}{[1 + (\frac{1}{\omega RC_t})^2]^{1/2}}, \quad C_t = C + C_s$$

and

$$V_2 = \frac{Q_o/C_t}{(\frac{1}{\omega RC_t}) + \omega RC_t}$$

In the equivalent circuit the shunt resistance of the piezo-element has been ignored since it is typically greater than 10^{10} ohms. However, the shunt resistance obviously limits the value of the input resistance to the instrumentation. From Equation (3.35) the second term involving an exponential function is a transient. To illustrate the points of interest this term can be ignored and only the steady state component considered. Considering only the magnitude of the voltage, the maximum value occurs when $C >> C_s$ and $\omega RC_t >> 1$. Thus

the stray capacitance must be kept to a minimum. For a value of C= 25 pf and $R=10^8$ the second contraint is satisfied for f >> 60 Hz.

The environmental factors of primary importance for piezolectric elements are the variations of piezoelectric sensitivity, dielectric constant and leakage resistance with temperature. Since quartz is widely employed in piezoelectric transducers, the variations of the d-coefficient, the relative dielectric constant and the volume resistivity with temperature are shown in Figures 3.17, 3.18 and 3.19. The 0.016% /°C decrease in piezoelectric sensitivity with increasing temperature does not appear to be a serious limitation. The increasing relative dielectric constant effectively decreases the sensitivity of piezo-transduction. However, the decrease is approximately 1% /°C and will not be a important factor for reasonable temperature variations.

The decrease in leakage resistance with temperature as shown in Figure 3.19 does present a problem in piezoelectric transduction. While the Curie temperature for quartz is 576°C, the rather rapid decrease in leakage resistance limits the operating range to approximately 200°C. The insulation resistance decreases by approximately a factor of 10 from 0°C to 200°C thus reducing the maximum input impedance to the read out circuits.

Although quartz has been considered as the example, piezoelectric ceramics and semiconductors have received considerable attention in the past 20 years [39],[40]. The major advantage of the piezoelectric ceramics are the increased coupling coefficient and piezoelectric sensitivity. For illustration the coupling coefficient may range from 9% to 67%

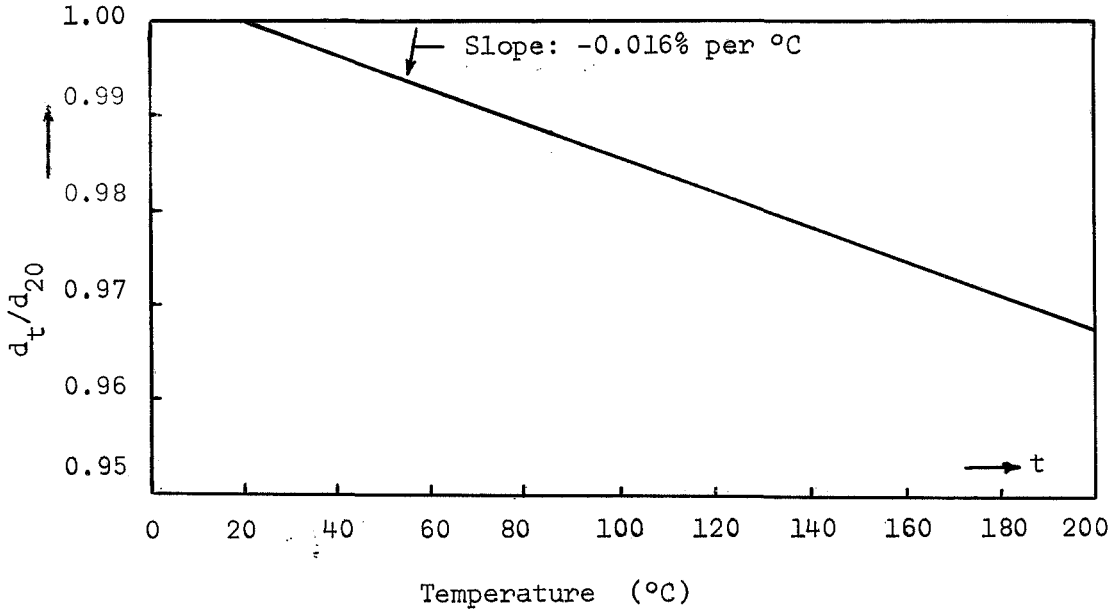


Figure 3.17. Variation of the d_{11} coefficient of quartz with temperature, referred to d_{11} at 20°C. [37]

for lead zirconate-titanate compounds operated in the thickness extension mode while the piezoelectric sensitivity ranges from 90 to 320 x 10^{-12} coulombs/newton. For barium titinate compounds corresponding ranges are 28% to 50% and 80 to 190 x 10^{-2} coulombs/newton [37].

The large coupling coefficients and piezoelectric sensitivity are offset somewhat by the larger relative dielectric constant. From Equation (3.35) the voltage sensitivity may be defined as

$$v_{max} = \frac{Q_o}{C} = \frac{d_{11}\,\sigma_1}{C_A} \text{ volts} \qquad (3.36)$$

where C_A is the capacitance per unit area of the piezo-element d_{11} the piezoelectric sensitivity for the thickness expander mode and a

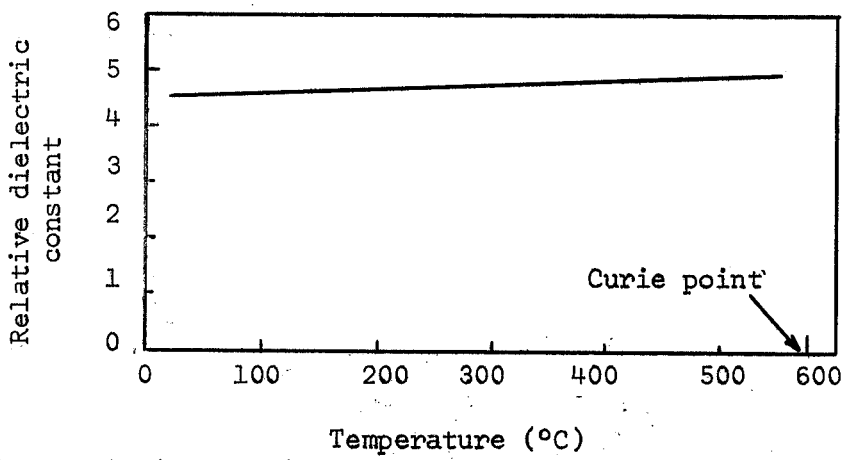


Figure 3.18. Variation of dielectric constant of quartz at higher temperature. [37]

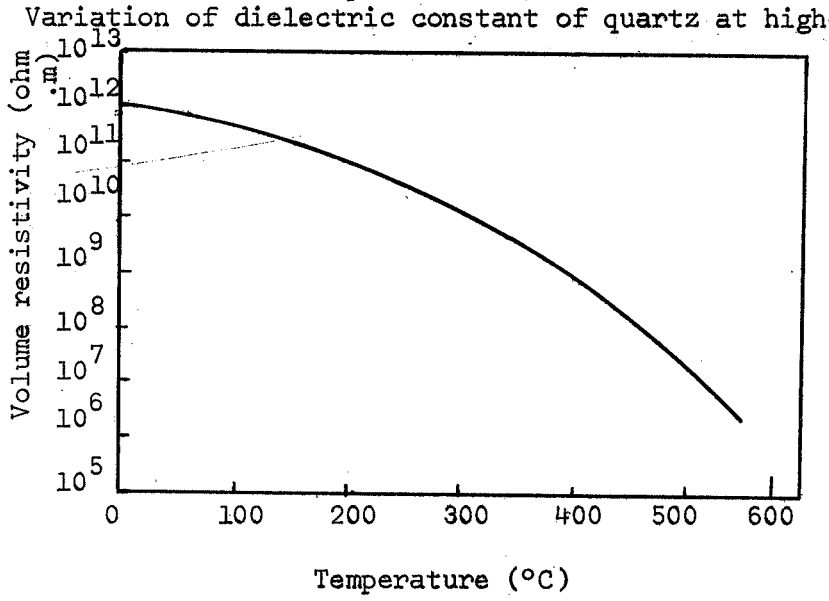


Figure 3.19. Variation of resistivity of quartz at higher temperatures. [37]

uniaxial stress σ_1 ($\frac{\text{newtons}}{m^2}$) applied in the thickness direction. The relative dielectric constants range from 500 to 2000 as compared to approximately 4.5 for quartz [39]. Some of the capacitance per unit area may be reduced by reducing the thickness of the piezo-element. Nevertheless, the capacitance of ceramic elements is about a factor of 10 larger than for quartz elements.

A major disadvantage of ceramic peizo-elements arises due to the fact that most are also ferroelectric. As such, alignment of the spontaneous polarization within domains is required before appreciable piezoelectric transduction can be observed. This is referred to as "poling". As a result many ceramic piezo-elements are more temperature sensitive tha quartz elements. Part of this is explained by the pyro-electricity which is in strong evidence in ceramics such as lead-zirconate-titanate. The temperature sensitivity is somewhat complicated by aging effects. In essence the domains are radomized with thermal energy. Clearly this is an undesirable effect. As a result the temperature range for most piezo-electric ceramics is limited to less than 250°C although the Curie temperature approaches 400°C for some lead-ziroconate-titanate compositions.

3.1.5 Magnetic Transduction

There are two magnetic transduction principles of primary concern. The magnetic properties of solids such as ferromagnetism and ferrimagnetism are sensitive to stress. Also, Hall effect transduction has been considered for mechanical transducers. Both transduction principles have received only limited attention and will not be discussed in detail.

The change in magnetic properties with stress is primarily due to reverse magnetostriction. A slight distortion of crystal structure or dimensions introduced by an applied stress will change the distribution of nuclear and electron charges and spins. Thus the permeability of ferromagnetic and ferrimagnetic materials responds to an applied stress. This effect has been considered in various forms.

Ferrimagnetic transduction which utilizes square loop ferrite cores similar to digital memory cores has been reported [41]. The transducer was primarily a pressure transducer with a range of 0-20,000 psi. The transduction is accomplished with essentially a pulse transformer where the amplitude of the output pulse is proportional to the input pulse and the permeability of the squre-loop ferrite core. The core is wound with two coils, a drive or input winding and an output winding. The major feature of this transduction principle is the high speed capabilities (switching or pulsing intervals of 0.7 μ sec) and the compatibility with data processing equipment usually associated with digital equipment.

Detection of pressure shock waves has been reported using ferromagnetic transducers [42]. The principle of operation is essentially the same as for the square loop ferrite core transducer. However the transduction is an irreversible process in this particular instance. Initially there is a remanent magnetization in one side of magnetic core and a sense winding about the other side of the magnetic core or path. The shock wave destroys the ferromagnetism in the magnetic circuit and reduces the magnetic flux from the remanent magnetization to essentially zero. The electrical energy observed in the output pulse is approximately 10^3 larger than the stored magnetic energy hence the authors consider the effect as a transducer. Obviously, the major disadvantage of this transducer is the one-shot capability. In addition the output is essentially digital in nature. Possibly, a careful study would yield a characterization of the conversion of mechanical to electrical energy hence a characteristic of the shock-wave.

The permeability of ferrites in a microwave cavity has also been considered as a transduction principle [43]. As the ferrite is stressed the permeability of the ferrite changes and results in a shift in the reasonant frequency of the cavity. As an accelorometer a frequency shift of approximately 10^6 Hz/g was observed. The non-linear output and the stability of the cavity with temperature represent two of the major difficulties with this particular transduction principal.

Hall effect transduction is a distinctly different principle. The details are discussed in Section 1 under electrical conductivity. Thus the transduction principle is derived from charge carrier transport. In Section 1 the Hall voltage was defined as illustrated in Figure 3.20 where

$$V_H = K I_c \text{ Sin } \theta, \tag{3.37}$$

where K is a constant which includes geometry and basic material parameters, I_c is the input control current, B is the magnetic flux density and θ is the angle between the surface of the Hall sample and the magnetic flux.

As one might expect the angular relationship in Equation (3.37) can be used for mechanical transducers. Basically, the transduction involves mechanical displacement (ideally angular although linear displacements in non-uniform magnetic fields can be detected [44]) to a Hall voltage.

As an angular transducer the major advantages are the absence of slip rings for electrical leads (the magnetic field can be part of the motor), long life, high speeds of operation and low power.

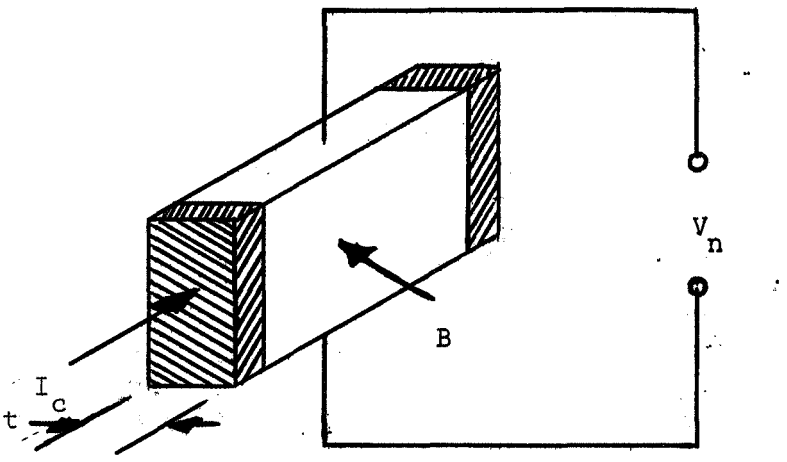


Figure 3.20. Illustration of Hall voltage.

Typical power requirements are 28 volts d. c. at 20 milliamperes. The frequency range is from 60 to 20,000 Hz. However errors in measuring angular displacement are as large as 1° - 30' [45]. Major contributors to the error are temperature effects, magnetic instabilities, misalignment voltages, noise and amplifier errors.

If the sensing element in a mechanical transducer results in a substantial displacement either angular or linear, Hall effect can be used as the transduction principle. For example, a spring-mass type of accelerometer consisting of a loaded cantilever where the Hall effect device is at the end of the cantilever has been constructed [45]. A seismic mass actually provides the load for the cantilever. The Hall sample is placed between the poles of a magnet. Displacement in the magnetic field results in a change in the Hall voltage which is essentially linear for accelerations from 1 to 10^{-3} g's.

Although specific transducers have been realized, the major considerations must be the unique feature derived from using Hall effect transduction. From the examples reported to date, there is insufficient evidence to indicate a significant impact of Hall transduction on mechanical transducer design. From the limited information one of the problem areas appears to be the source of uniform, stable, large magnetic fields.

3.1.6 Summary for Mechanical Transducers

To summarize the impact of solid state technology on the design of mechanical transducers, the trends in solid state technology should be briefly mentioned. During the '50's the emphasis was on discrete device performance. As a result, semiconductor material research was the primary concern in understanding and extending the range of diodes, transistors and associated components. In addition, the emphasis on size so that all discrete components occupied the same order of magnitude of volume was evident.

As transistor electronics equalled and then surpassed vacuum tube electronics in versatility and performance, the emphasis shifted to integrated circuits. Here the primary concerns were reliability, standardization and repeatable technology. With the introduction of the planar silicon devices, solid state technology took on a new dimension in complexity. The large device densities achieved by integrated circuits naturally leads one to think of systems. How can one interconnect large arrays of solid state devices on a single silicon chip to provide a desired function?

Large scale integration will become a reality during the 70's. So the time span we must consider begins with solid state material science of the 50's and extends at least to LSI of the 70's. The main reason for mentioning these trends is that transducer development based upon solid state technology is likely to parallel the concept of device, circuit and system. As a result, the deciding facts in transducer development are likely to be: 1) Is the technology available to realize the tranducer? 2) Can the desired performance characteristics such as response time, signal to noise and impedance be obtained with a given transduction principle? 3) Is the sensing element such as a diaphragm, cantilever or seismic mass a conventional design? 4) What unique housing problems are likely to be encountered? 5) How much signal conditioning is required? 6) Can integrated circuits be an integral part of the transducer system? and 7) Can self calibration or accuracy estimates be included in the system design?

The above questions cannot be answered in full for mechanical transducers, however an attempt to state our present position in these matters will provide a constrast of transduction principles and identify areas of promising research and development. Concerning technology, there appears to be adequate technology for the most part. One exception is thin film technology which has met with only limited success. The major problem is control of electrical properties. R. F. sputtering adds a new dimension and further research utilizing sputtered films should be encouraged.

Silicon technology has been widely used to achieve a variety of mechical transducers [47],[48],[49],[50]. In fact the variety of devices

conceived and constructed utilizing the detailed knowledge and technology can be viewed as a problem rather than an asset. The major difficulty is the sorting of the various transduction principles into their appropriate catagories and assessing their true merits in a particular application. The over used concept of sensitivity whether it is gauge factor, pressure sensitivity or acceleration only serves to obscure the much needed unifying measure of performance. Much of the work we present in this document is aimed at this problem. However we readily admit that we have not achieved our objective. However this problem will be discussed in more detail in Section 4. The main reason for introducing the matter here is to illustrate that technology is not limiting the development of silicon transduction elements.

One faces a somewhat different situation when materials other than silicon are considered. In most instances a new technology must be developed peculiar to the material. The best example is gallium arsenide. Early attempts to translate silicon technology for purification, diffusing and etching simply did not work. With recent advances in a technology developed specifically for gallium arsenide, devices and transducers are feasible. However, the development cost was enormous in terms of man hours and money and cannot be duplicated for every material available. Therefore, technology will seriously impede the development of transducers obtained with materials that are not ameanable to simple processes for controlling electrical properties and geometries.

The characteristics of the more promising transduction principles for contemporary transducer design have been considered in earlier sections. It may prove useful to attempt a few generalizations based

upon the material in these sections and to provide our assessment of the problems and opportunities. There is little doubt that the silicon strain gauge is the most significant contribution of contemporary solid state technology to mechanical transduction. However its utility is limited. For precision strain measurements the wire strain gauge remains a primary choice. For dynamic acceleration measurements piezoelectric transduction is superior. Judging the performance of silicon strain gauges as to what is best in a particular situation can prove discouraging. One should consider the natural attributes of silicon strain gauges not offered by other transduction principles. With a strain gauge signal conditioning can be less of a problem for moderate strains (greater than 10 microstrains). Diaphragm silicon strain gauge accelerometers are superior to piezoelectric accelerometers for frequencies less than 100 Hz. Diaphragm silicon strain gauge pressure transducers appear feasible in the 1 mm to 1 micron of Hg range where few pressure transducers offer sufficient sensitivity.

One major attribute that has not been fully exploited is the use of silicon diaphragms with diffused strain gauges. One might term these as integrated piezoresistive transducers [51]. The entire bridge can be diffused into the diaphragm and it is at least conceivable that signal conditioning with transistor amplifiers could be included. Thus one advances from a simple transduction element to a transducer circuit. This concept could be applied to strain gauges, accelerometers or pressure transducers. One major advantage which could be realized is a reduced cost in the overall system. It may be possible to provide diaphragm transducers with outputs suitable for oscilloscopes,

digital displays or conventional voltmeters. Temperature compensation, linearity compensation and system calibration could be designed into the transducer circuit. Of course this is highly speculative but there appear to be no fundamental limitations in implementing into a single silicon wafer the circuit elements and amplifiers presently required. To do this economically would require a certain degree of standardization.

One of the major concerns about silicon strain gauges has been temperature compensation. However, the recent work of Gross provides encouragement [52]. The performance of electron irradiated silicon strain gauges offers significant improvements with respect to temperature variations in resistance. To utilize irradiation compensation with all the concepts discussed above does not appear promising, however the diffused bridge arrangement would still apply.

Finally, we should consider the junction transduction principles discussed earlier. Again, the natural attributes of these devices are emphasized. With the possible exception of heterojunctions, sensitivity is not a primary consideration. For the p-n junction size is the primary attribute. For the metal-insulator-silicon devices, the d.c. response for piezoelectric insulators is the major factor. One should not consider the gain available with transistors or field effect devices as a primary attribute. Fabrication and reliability of active devices for transduction has been a problem. Although, many of the problems can and will be solved, external gain using integrated circuits is a more reasonable solution. Junction devices will find applications, however these applications will likely require those attributes not available with either silicon strain gauges or other more conventional transduction principles.

3.2 Optical Transduction

During the past decade, the impact of solid state technology on optical transduction has been enromous. First, the space era was to a large extent supported by the development of the silicon solar cell. Following rather natural trends improvements upon the performance of silicon solar cells were given high priority. As a result, a variety of concepts evolved. The use of wide-band gap materials with homojunctions and heterojunctions have received considerable attention in attempts to increase the efficiency of converting sunlight into electrical energy. Another consideration was the relatively small active area of silicon solar cells even in large arrays. Attempts to increase the power output per unit area of the array emphasized thin film technology. All of these attempts to improve upon silicon solar cells have been only qualified successes. For overall performance silicon solar cells remain the primary solar energy converter for space applications. However, the tremendous effort devoted to solar cells has greatly influenced optical transduction even though silicon has not been replaced.

In this review, we will not be primarily concerned with the solar cell. In fact, the solar cell will be considered as a photodiode without reference to solar energy conversion. The primary reason for this approach is to reinforce the transduction principles and not the peculiarties of achieving a particular sensor or transducer.

Another factor which has greatly accelerated optical transduction principles is the laser. Coherent optics has been considered for communications, surgery, cutting, electronic memories and a variety of

special applications. In many of these applications a suitable detector of the light intensity is required. Since laser radiation now covers the range from the ultraviolet well into the infrared, optical transduction was emphasized over a rather broad wavelength spectrum. However, we will not pursue the details of sensors suitable for detecting laser radiation. Consistent with our primary objective only the unique optical transduction principles developed for laser detection will be considered.

To organize this section optical transduction has been divided into three parts. First a brief discussion of photoemission will be directed toward photomultiplier's. This is introduced primarily for comparison and as background information to provide a basis for assessing solid state impact or optical transduction. Secondly, bulk effects such as photoconductivity will be discussed. Finally, photojunctions will be considered. In each of these sections a division with respect to ultraviolet visible and infrared radiation appears natural. The authors fully realize that some overlap will occur and infrared transduction for a particular material or device may be less efficient than visible light transduction. However we may choose to consider the infrared transduction. This will place somewhat of a burden on the reader to be flexible in terms of our choices and to provide his own interpretation for the primary importance of a particular transduction principle.

Finally, it must be stated that the author's consider this section to be somewhat incomplete. One need only consider the literature for the past ten years to understand the difficulty in attempting complete-

ness. We have highlighted the areas that have been productive or appear promising. It is quite likely that our judgment is fallable. However, the rather general discussions in Sections 1 and 2 are intended to be sufficient background such that an understanding of the material in those sections provides the insight for one to arrive at an independent judgement. Therefore our position for the summary of optical transduction based upon contemporary solid state technology should not be too disconcerting.

3.2.1 Photoemission

Photoemission was not treated in either Section 1 or 2. This omission was by design. Einstein's concept of the photoelectric effect representing the foundation of photoemission as a photocathode or photocell occurred in the 1930's. Therefore photoemission represents the tradition in optical transduction and requires only a brief introduction.

Coupling of the photocathode with a dynode system resulted in the photomultiplier. The photocathode emits electrons as a result of optical stimuli with the proper energy. The emitted electrons are directed to a dynode where secondary electrons are created by the energy loss of the impinging electron. This process is repeated many times with resulting electron amplification. Thus a photomultiplier is a photocathode with gain. The gain can be from 10^6 to 10^7. One of the major advantages of a photomultiplier is that the amplification of the photocathode emission using dynodes results in less noise than conventional amplification of a photodiode output. Thus one is able

to detect lower light intensities with the photomultiplier than with a photodiode and conventional amplification.

Since our intent is assessment of transduction principles, the photomultiplier will not be considered in detail. However the photocathode will receive considerable attention as the transduction element of a photomultiplier. To characterize a photocathode, spectral response, sensitivity, and dark current must be considered. Each of these characteristics should be defined before proceeding.

The spectral response is usually presented with the photoemission current plotted on the ordinate and the wavelength of incident light on the abscissa. The response or sensitivity is usually related to incident light energy rather than absorbed light energy. Therefore, factors such as reflection and transmission are ignored. The ordinate or photoemission yield may be in units of ampere/watt or electrons/photon. The later unit is a guantum yield. The abscissa may be either wavelength or energy. One may find any combination of units of the four possibilities.

The sensitivity of a photocathode will be in terms of the photoemitted current per unit incident power in the light flux. For monochromatic light the spectral response provides the sensitivity at each wavelength provided absolute sensitivity is obtained as opposed to relative sensitivity which is often plotted. However, the maximum sensitivity is usually given in terms of milliamps/watt at the wavelength of the peak response.

In some applications the photocathode responds to a light source with a distributed spectral emission. A tungsten lamp is a classic example. In this case, the photoemitted current is

$$I = q \int_{\tau_{min}}^{\tau_{max}} I_p(\tau)\, \eta(\tau)\, d\tau \qquad (3.38)$$

where $I_p(\tau)$ is the incident photon flux spectrum, $\eta(\tau)$ is the quantum efficiency as a function of wavelength and the limits of integration extend over the range of wavelengths where both the light source and photocathode spectrum are non-zero.

The dark current from a photocathode provides an unwanted signal and as such may be regarded as a limiting factor in determining the sensitivity of a photocathode. Typical values may range from 10^{-12} - 10^{-15} amp/cm^2. The expression for thermionic emission is

$$J = A_o\, T^2\, e^{-(E_a + \mu)/kT} \qquad (3.39)$$

where J is current density, T is absolute temperature, k is Boltzmann's constant, μ is the energy difference between the bottom of the conduction band and the Fermi level, A_o is a constant and E_a is the electron affinity of the photocathode material. The work function which determines the cuttoff frequency for photoemission is greater than the sum of the electron affinity and μ. Therefore as one reduces the work function the thermionic emission increases. This will be an important factor for near infrared photoemission.

To provide the brief introduction to photoemission only representative photocathodes in the ultraviolet, visible and infrared will be considered. However, each region of the spectrum presents unique problems and will be treated in separate sections.

3.3.1.1 Ultraviolet Photoemission (2000 A°- 3500 A°)

Although a number of conventional photocathode materials extend into the ultraviolet few appear to span the range from 2000 A° - 3500 A° or from approximately 3 ev to 6 ev. Furthermore, the reported response data are often obscured by the transmission of the enclosure. Of the conventional photocathodes the cesium-antimony photocathodes with three spectral responses designated as S-13; S-19 and S-21 may be considered. The most promising of these is the S-13 which has a fused-silica enclosure. The spectral sensitivity of the S-13, Cs-Sb photocathode, with enclosure of fused silica is shown in Figure 3.21. The maximum response occurs at 4400 A° with a typical radiant sensitivity of 47.7 milliamperes/watt and a typical quantum efficiency of 13% [53]. At 2000 A° the sensitivity has dropped to approximately 29 milliampers/watt with a quantum efficiency of 17%. Typical values are quoted since the performance of each photocathode will only approximate the above description. Typical S-13 photocathode dark emission at 25°C is 4×10^{-15} amperes/cm^2.

Using the dark current as limiting the value, a monochromatic photon flux of approximately 10^5 $\frac{\text{photons}}{cm^2\text{-sec}}$ could be detected in the range from 2000 A° to 4000 A°. For 2000 A° wavelength of the incident flux the incident power density would be 10^{-13} watts/cm^2. One would be faced with measuring 2.9×10^{-15} amps/cm^2. This clearly exceeds the range of routine measurement. With a photomultiplier, current amplification of 10^6 results in a current level that is well within the range of electrometer capabilities. Therefore one of the critical factors in photoemission transduction is amplification with adequate

signal to noise.

The concept of a photomultiplier as a detector of individual quanta of incident radiation is rather wide spread. Exactly how this might be accomplished is not clear from the above discussion. It would appear that the lower limit of 10^5 $\frac{\text{photons}}{\text{cm}^2\text{-sec}}$ represents the minimum detectable light flux. In fact this is true if steady state current output is measured. The photocathode steady state photoemission current (I) is

$$I = I_p \bar{\eta} q \qquad (3.40)$$

where I_p is the photon flux rate in photons/cm^2-sec, $\bar{\eta}$ is the number of photo electrons per incident photon (quantum yield) and q is the electronic charge. Thus the photoemission current is directly proportional to the incident flux rate and limited by the dark current.

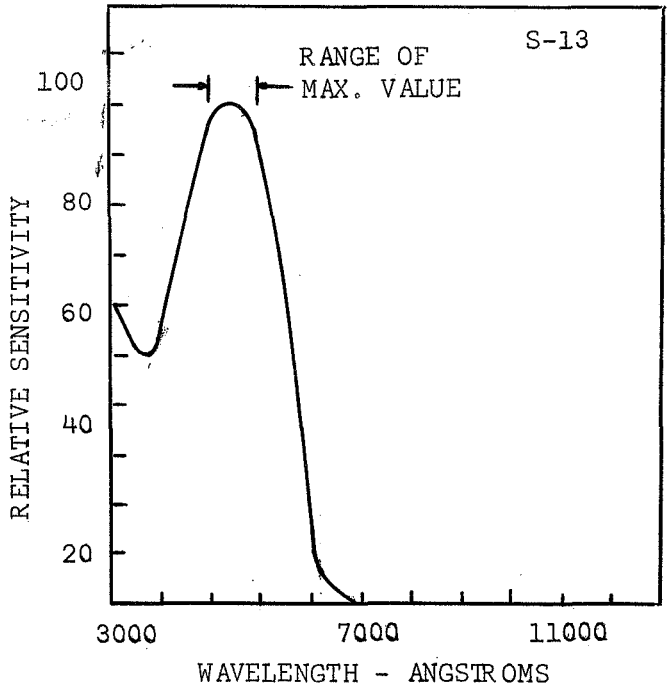


Figure 3.21. Relative spectral response characteristic for S-13 photocathode [53].

For illustration of how quanta are detected consider the following photocathode arrangement. The photocathode has a spectral response of S-13 with a dark current of 4×10^{-15} amperes/cm^2. The photocathode is arranged as shown in Figure 3.22. Assume a light flux rate of ϕ_p photon/sec. Consider a single photoemitted electron. The current contribution in the external circuit I (t) will be

$$I(t) = -\frac{q}{\tau_t} \quad o < t < \tau_t \tag{3.41}$$

where τ_t is the time of flight for the electron between the cathode and anode and t=0 corresponds to the instant of electron escape from the photocathode surface. To calculate τ_t assume that the electron escapes with zero kinetic energy. Therefore the drift time will be determined by the acceleration of the electron in the electric field between the cathode and anode. Under these assumptions the transit time of the electron is

$$\tau_t = \sqrt{\frac{2L^2 m}{q V}} \quad \text{seconds}, \tag{3.42}$$

where L is the spacing between the cathode and anode, m is the electron mass, and V is the applied voltage. For a spacing L = 1 cm and an applied voltage of 500 volts, the transit time is 1.5×10^{-9} seconds. Thus the instantaneous current during the electron transit time is 1.1×10^{-10} amperes. Without amplification this pulse of 1.5 nanoseconds is clearly beyond state-of-the art detection. However with the current gain of a photomultiplier the current value could

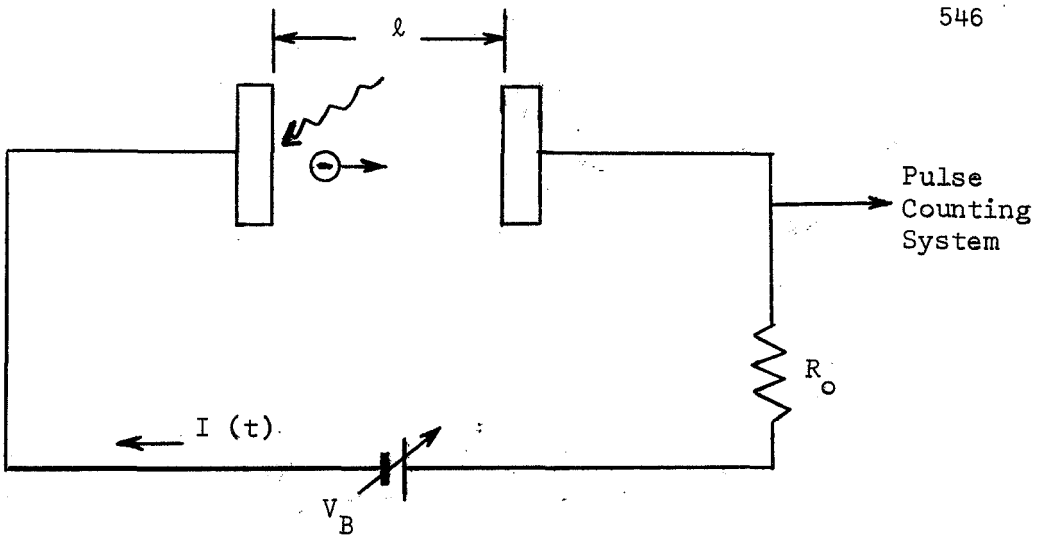


Figure 3.22. Simple model for analyzing pulsed response of photo cathode.

be increased to 1.1×10^{-4} amperes. This boarders on state of-the-art in integrated circuit performance and definitely appears promising. However there would be an upper limit for counting quanta. The flux rate should be much less than the pulse detector response time. Assuming a response time near 10^{-9} seconds, requires the photon flux rate to be much less than 10^{9} photon/sec which corresponds to an incident power of approximately 10^{-9} watts. Therefore quanta counting is limited to very low incident flux rates. There appears to be a region of incident photon flux rate from 10^{-9} to 10^{-12} photon/sec where either steady-state or pulse currents can be detected.

3.2.1.2 Ultraviolet Photo Emission (1050 A° - 2000 A°)

For the ultraviolet in the range from 1050 A° to 2000 A° the window or enclosure problem becomes difficult. LiF can be used since

the optical absorption edge is 1050 A° (∿ 12 ev). However, the transmission is useable over most of the spectrum from 1050 A° to 2000 A°. The photocathodes in this range can be one of the alkali halides. The spectral response for those with relatively high quantum yields are shown in Figure 3.23. Since these materials have relatively large band gaps, thermionic emission should be less than for the S-13 (CsSb) photocathode. Of the akali halide CsI, has been judged to be the most promising in the 1050 A° to 2000 A° range [54].

Although we have not found data which clearly establishes the limiting factors for alkali halide photocathodes, there also appears to be no fundamental reason why their performance as a quanta detector or as a detector of a photon flux rate should not be similar to the performance of the S-13 photocathode in the 2000 A° - 3500 A° range. However, the problems of calibration of light sources, transmission windows, vacuum enclosures, and material processes may preclude reliable performance characteristics. Therefore it would be meaningless to speculate further on photoemission transduction in the range from 1050 A° to 2000 A°.

3.2.1.3 Visible Photoemission (3500 A° - 7000 A°)

Photoemission is best suited for optical transduction in the visible spectrum. Most practical photocathodes have a peak in sensitivity and quantum efficiency in the range from 3500 A° to 7000 A°. Over the visible range there are two units of sensitivity usually given. The radiant sensitivity (σ_λ) applies to a single wavelength usually given at the peak of the spectral response. Typical values of quantum efficiency usually correspond to the same wavelength. How-

ever, in many instances the photocathode must respond to a light source which has a continous emission wavelength over the entire visible spectrum. In this case a sensitivity based upon the response to such a light source is useful.

With a calibrated light source, usually a tungsten lamp operated at a particular characteristic temperature photoemission in terms of microamperes per lumen in often given. This characteristic is termed luminous sensitivity. The lumen is a unit of spectral power and is based upon the response of the eye. Therefore the lumen is specifically related to visible light flux.

There are at least 13 types of photocathodes that are suitable for photoemission transduction in the visible. The data for these various materials are rather common. A variety of applications have

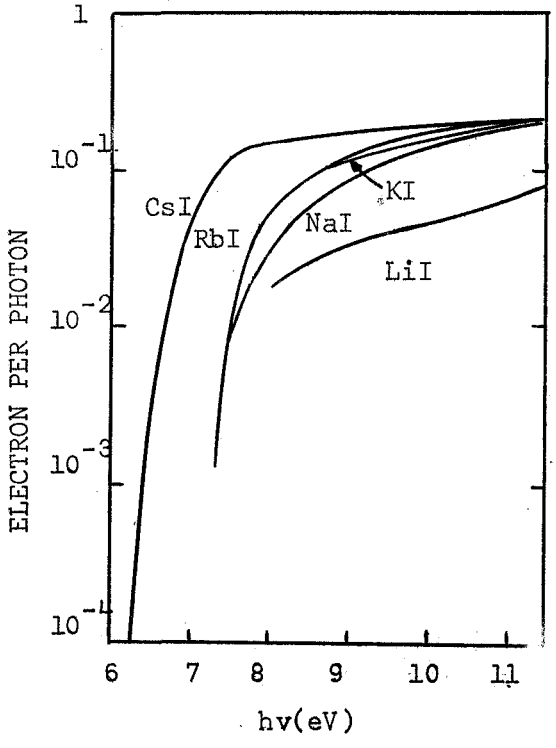


Figure 3.23. Spectral response curves of alkali iodides. [54]

been found for these photocathodes including image converters, television camera tubes, photomultipliers and photodiodes. Therefore, Table 6 with characteristics for these 13 photocathodes is included for later use in comparing other types of optical transduction.

Except for the concept of luminous sensitivity, the performance of a photocathode in the visible spectrum is similar to the description of the S-13 in Section 3.2.1.1. Of course the actual numbers vary depending upon the particular photocathode material, however the concept of quanta detector and flux rate detection are identical. One factor which does influence these concepts is the wide variety of light sources and filters available for use in the visible. Very careful calibration of the sensitivity can be routine when typical values are not satisfactory.

The limiting factor for steady state current measurement of photoemitted electrons is the dark current which can range from 10^{-9} to 10^{-15} amperes/cm^2 at room temperature. Thus the minimum detectable light flux may vary by a factor of 10^6. Typical values in Table 6 are rather optimistic. The incident power in watt/cm^2 for monochromatic light of wavelength λ and energy E_λ in election volts is given by

$$P_\lambda = q\, E_\lambda\, I_p \left(\frac{\text{watts}}{cm^2}\right) \qquad (3.43)$$

where I_p is the incident photon flux in $\frac{\text{photons}}{cm^2\text{-sec}}$.

Device S-Number[a]	Photocathode Type and Envelope	Conversion Factor[b] (k) (lumens/watt)	Typical Luminous Sensitivity[c] (μa/lumen)	λ_{max} (angstroms)	Typical Radiant Sensitivity[d] (ma/w)	Typical Quantum Efficiency[e] %
S-1	Ag-O-Cs Lime-glass bulb	93.9	25	8000	2.35	0.36
S-3	Ag-O-Rb Lime-glass bulb	286	6.5	4200	1.86	0.55
S-4	Cs-Sb Lime-glass bulb	977	40	4000	39.1	12
S-5	Cs-Sb 9741 glass bulb	1252	40	3400?	50.1?	18?
S-8	Cs-Bi Lime-glass bulb	755	3	3650	2.26	0.77
S-9	Cs-Sb Semi-Transparent, Lime-glass bulb	683	30	4800	20.5	5.3
S-10	Ag-Bi-O-Cs Semi-Transparent, Lime-glass bulb	508	40	4500	20.3	5.6
S-11	Cs-Sb Lime-glass bulb	804	60	4400	48.2	14
S-13	Cs-Sb Semi-Transparent, Fused-silica bulb	795	60	4400	47.7	13
S-19	Cs-Sb Fused-silica bulb		40		~ 22	~ 11
S-20	Sb-K-Na-Cs (Multialkali) Semi-Transparent, Lime-glass bulb	428	150	4200	64.2	18
S-21	Cs-Sb Semi-Transparent, 9741 glass bulb	779	30	4400	23.4	6.6
	CsK-Sb Pyrex-glass bulb	1140	77	3850	88	28

[a] The S-number is the designation of the spectral-response characteristic of the device and includes the transmission of the device envelope.

[b] k is the conversion factor from amperes/lumen to amperes/watt at the wavelength of peak sensitivity.

[c] The luminous sensitivity is for the photocathode for 2870°K color temperature test lamp.

[d] The radiant sensitivity is at the wavelength of maximum response.

[e] 100 % quantum efficiency implies one photoelectron per incident quantum, or e/hv=λ/12,395, where λ is expressed in angstrom units. Quantum efficiency at λ_{max} is computed by comparing the radiant sensitivity at λ_{max} with the 100 % quantum efficiency expression above.

Table 6 Spectral Response and Related Photocathode Characteristics [53]

If the minimum detectable photon flux rate is limited by dark current

$$I_p \text{ (min)} = \frac{I_D}{qn_\lambda} \left(\frac{\text{photons}}{\text{cm}^2\text{-sec}}\right) \tag{3.44a}$$

where I_D is the dark current in amperes/cm^2 and n_λ is the number of electrons per photon or quantum yield of photoemission. Thus the minimum detectable power is

$$P_\lambda \text{ (min)} = \frac{E_\lambda \ I_D}{\eta_\lambda} \left(\frac{\text{watts}}{\text{cm}^2}\right). \tag{3.44b}$$

The quantum yield can be related to the absolute sensitivity at a given wavelength by

$$\sigma_\lambda = \frac{\eta_\lambda}{E_\lambda} \left(\frac{\text{amperes}}{\text{watt}}\right). \tag{3.44}$$

where E_λ is the photon energy in electron volts. Thus the minimum detectable power is

$$P_\lambda \text{ (min)} = \frac{I_D}{\sigma_\lambda} \left(\frac{\text{watts}}{\text{cm}^2}\right). \tag{3.45}$$

For a given dark current, the minimum detectable power is proportional to the absolute sensitivity at a given wavelength.

To illustrate the use of Equation 3.45 consider the S-20 photocathode response in Figure 3.24. The photocathode is a multialkali of antimony-phosphorous-sodium-cesium (S_b-K - N_a - C_s). A typical value of dark current of 0.3×10^{-15} amperes/cm^2 has been reported [53].

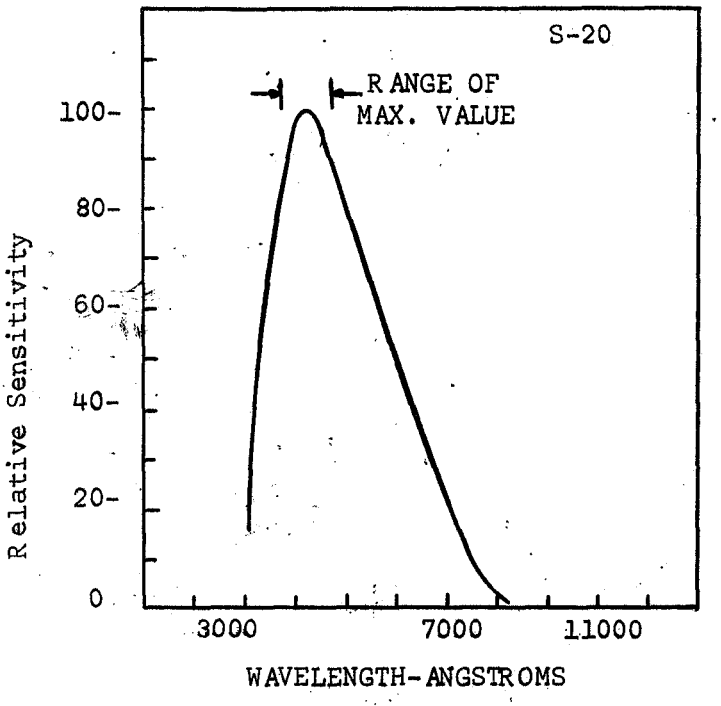


Figure 3.24. Relative spectral response characteristic for S-20 photocathode [53].

Using the radiant sensitivity in Table 6 and the reported dark current yields a minimum detectable power at 4200 A° of 4.7×10^{-15} watt/cm^2 (2.9×10^4 $\frac{e.v}{sec\text{-}cm^2}$) or a flux rate of 10^4 $\frac{photons}{sec\text{-}cm^2}$. At 7000 A° the minimum detectable power drops to approximately 1/4 the value at the peak. Thus the S-20 appears to offer a reasonable performance throughout the visible.

The luminous sensitivity of the S-20 response is the largest of the typical values reported. This is partly due to the high quantum efficiency and also due to the general shape of the spectral response. Therefore it was considered as the appropriate standard for comparing the performance of visible light transduction in subsequent sections.

3.2.1.4 Infrared Photoemission

The traditional photoemitter in the infrared is the S-1 silver oxygen-cesium (A_g-O-C_s) photocathode. The relative spectral response is shown in Figure 3.25. The major features are the relatively low quantum efficiency and the rather broad spectral response from the ultraviolet with the proper window to the near infrared. The quantum efficiency falls off rapidly for energies less than 1.5 ev ($\sim$ 8000 Å) as shown in Figure 3.26. However, in some instances a response to 0.73 ev (17,000 Å) has been reported. Therefore in principal, S-13 can be used from 6 ev (2000 Å) to 0.73 ev (17,000 Å).

A major disadvantage of the S-1 photocathode is the relatively large dark current which may exceed 10^{-12} amperes/cm^2. However, the large dark current for photocahtodes capable of photoemission in the near infrared should be expected. Considering Equation (3.39) for thermionic emission, the dark current depends exponentially on the sum of the electron affinity E_a of the material and the Fermi level measured from the conduction band. The photoemission threshold will exceed this sum. Therefore, the dark current increases exponentially as the photoemission threshold increases to long wavelengths. At 8000 Å the minimum detectable power for a dark current of 10^{-12} amperes/cm^2 is 4.25×10^{-10} watt/cm^2. This is a factor of 10^5 larger than for the S-20 photocathode in the visible. However, the dark current can be reduced orders of magnitude by cooling the photocathode. In principal the dark current can be decreased by cooling the photcathode until the limiting factor is either statistical fluctuations in the dark emission, extraneous light or thermal noise.

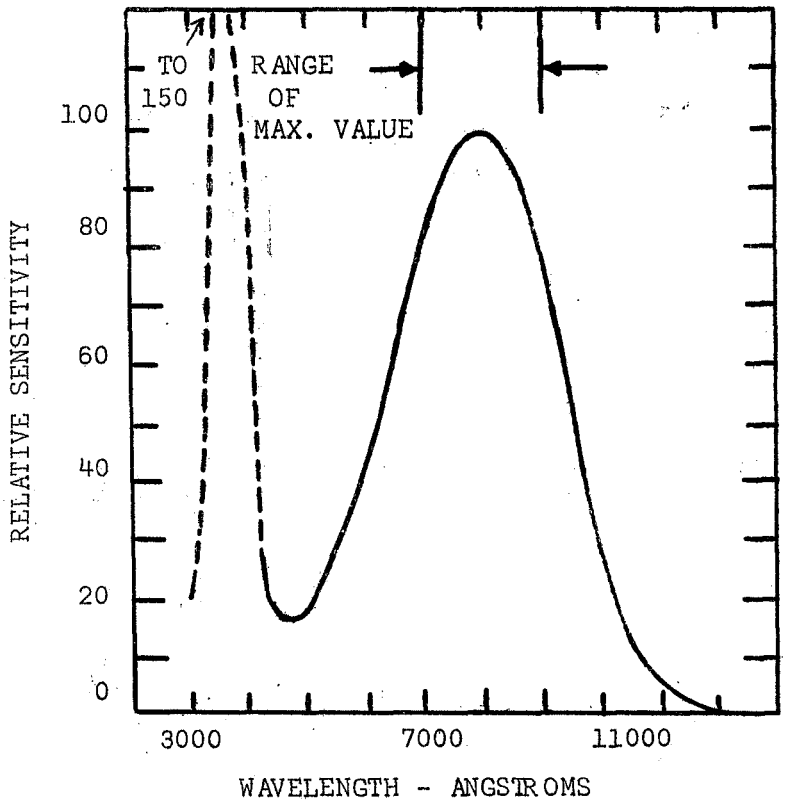


Figure 3.25. Relative spectral response characteristic for S-1 photocathode [53].

In summary there is little choice in infrared photocathodes. The performance in terms of radiant sensitivity, quantum efficiency and minimum detectable power does not match the performance of photocathodes operated in the visible. Although, photoemission in the near infrared is developing with present research, it is likely that the photocathodes will be cooled to reduce thermionic emission.

3.2.2 Photoconductivity

Photoconductivity is the modulation of the bulk electrical conductivity of a solid with photo-excitation of charge carriers. The absorption processes which result in excess charge carriers are

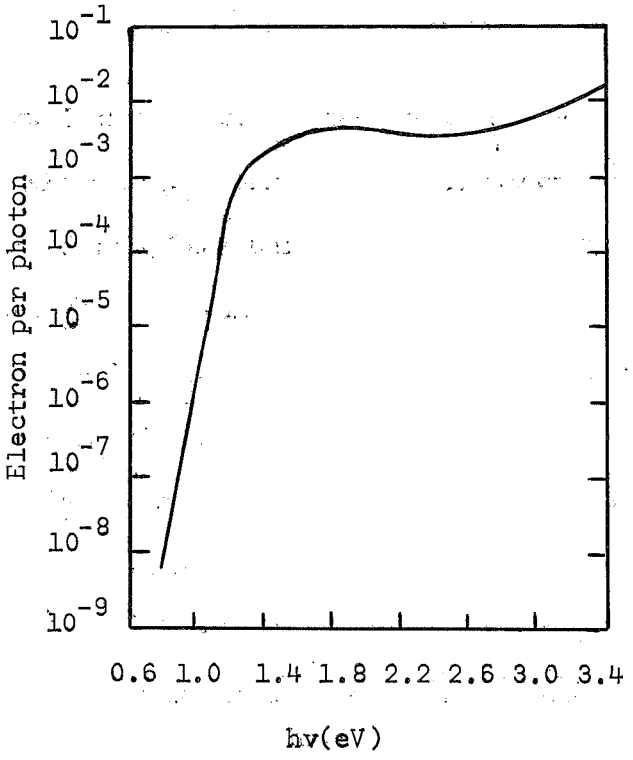


Figure 3.26. Spectral response curve of Ag-O-Cs cathode showing long wavelength threshold. [56]

discussed in Section 1. Also included is a description of how the conductivity depends upon the wavelength of the incident light. The normalized relative response for a number of photoconductors are shown in Section 1. Therefore we will not duplicate this effort by analyzing the spectral response of a particular material. Instead, the factors which characterize the sensitivity at a given wavelength will be considered. In the visible range the primary factors are specific sensitivity and photoconductivity gain. In the infrared the concept of detectivity is more convenient. These measures of sensitivity will be defined in the appropriate sections.

Another factor previously introduced in Section 1 which provides insight into photoconductivity is the response time. This concept depends on material parameters and can be treated in a rather general way. The salient features are: 1) the response time cannot be shorter than the free carrier lifetime and 2) the response time may be lengthened either thru sensitizing centers (recombination centers) or trapping centers. The details of these processes are treated in Section 1 and unimportant for our present discussion. However the free carrier lifetime is possibly the most crucial factor for characterizing photoconductivity. Often the importance of lifetime is obscured by the concepts of sensitivity. However in the following discussion the emphasis will be on relating the performance of photoconductivity to the free carrier lifetime.

3.2.2.1 Ultraviolet Photoconductors

There appears to be very little effort devoted to photoconductors in the ultraviolet. Possibly the best understood and most extensively investigated is zine sulfide. [57] However, the spectral response covers only a very limited portion of the spectrum. In fact, this may be one of the serious limitations for ultraviolet photoconductors. Many of the wide band gaps II-VI compounds are direct band semiconductors. Ideally there would be little absorption for incident radiation with an energy less than the band-gap energy. However the absorption rises rapidly as the incident photon energy exceeds the band gap energy. Typically the absorption coefficient for direct band gap materials rises from essentially zero to approximately 10^5 cm^{-1} near the band gap energy [58]. Incident photons with energy greater than band gap energy

are absorbed in a thin layer near the surface of the material. As discussed in Section 1 the surface recombination of excess carriers shortens their lifetime such that conductivity modulation becomes inefficient.

Another factor which characterizes wide band gap materials is the difficulty of obtaining intrinsic electrical properties. By intrinsic electrical properties, we mean the exclusion of impurities or imperfections as the primary contributor to conductivity, optical absorption and photoconductivity. This problem can be appreciated by estimating the intrinsic carrier concentration at 300°K. Using a value of conductivity ($\sigma = 10^{-10}$ ohm-cm) and mobility ($\mu \simeq 100$ $\frac{cm^2}{volt\ sec}$) corresponds to a carrier concentration less than 10^7 cm^{-3} [59]. Therefore the contribution from impurities must be much less than 10^7 cm^{-3} or impurity concentration of one part in 10^{16}. This is orders of magnitude beyond the state-of-the art. Even the values used to obtain the above estimate must be used under tenuous assumptions. However, the point to be made is that wide band gap materials will be extremely difficult to obtain as intrinsic photoconductors.

The incorporation of impurities in II-VI compound semiconductors may be turned to an advantage. Certain impurities give rise to an extrinsic photoconductivity which proves useful. The details of sensitizing photoconductors will be postponed until the next section. However, the effect has proved very useful for photoconductors in the visible spectrum and there appears to be no fundamental reason why the same approach will not work in the ultraviolet. In fact, a copper-doped zinc sulfide has been considered [60]. The major feature is a

broadening of the photoresponse to longer wavelengths. Since the photo-activation of an electron from a copper site in the host zinc sulfide is less than band gap energy the result appears consistent with an extrinsic photoconductivity.

Although zinc sulfide has been extensively studied there is very little evidence which offers encouragement as a practical photoconductor. The free carrier lifetime can be used as preliminary judgement. As we will show in the Section 3.2.2.2 free carrier lifetimes less than 10^{-6} seconds do not provide encouragement. A lifetime of 10^{-9} seconds has been reported for zinc sulfide [61]. Therefore it appears that other materials should be considered for ultraviolet photoconductors. However zinc sulfide has provided a convenient material to introduce some of the anticipated difficulties in obtain ultraviolet photoconductors.

Before proceeding to the visible spectrum a bit of speculation appears in order. Although the problems of achieving an ultraviolet photoconductor appear forminable, many of the reported electronic properties of diamond are encouraging. Diamond is an indirect band gap semiconductor and as such does not have the very sharp increase in absorption near the band gap energy. However the bandgap is 5.4 ev and the absorption coefficient rises from 10 to 10^{4} for photon energies from 5.3 ev to 6.6 ev [62]. However the absorption can be extended to longer wavelengths or smaller energies by impurities or electron irradiation [63],[64]. In addition carrier lifetime of 10^{-2} seconds have been reported [65]. Lifetime obtained from photoconductive decay of 9 seconds has also been reported [66]. In addition

photoconductivity has been observed for photon energies from 2 to 6 ev [67]. Although the above data appear encouraging, contact problems and rapid trapping of photo-excited carriers may seriously impede the applicability of diamond. However additional research certainly appears warranted.

3.2.2.2 Visible Photoconductors

In the visible portion of the optical spectrum cadmium sulfide (CdS) represents state-of-the-art in photoconductive transduction. A great deal is known about CdS however a complete review of CdS properties is beyond the scope of this effort. Our attention will be primarily on specific sensitivity, photoconductive gain, speed of response, maximum performance and signal to noise ratio. Each of these measures of performance will be defined and representative values for CdS will be given. For extensive treatments of photoconductivity, one is referred to the works by Bube [68],[69],[78].

The spectral response for CdS is shown in Section 1. The salient features are the peaks in response near the middle of the visible spectrum for intrinsic photoconductivity and the relatively narrow response. The response may be extended to longer wavelengths by introducing impurities. The spectral response in Figure 3.27 indicates that a trace of iodine in CdS extends the spectral response to 8000 Å. By various annealing procedures this response can be altered [71]. Indications are that enhancement in the long wavelength region near 8000 Å can be increased to a value equivalent to the intrinsic photoconductivity. However, the short

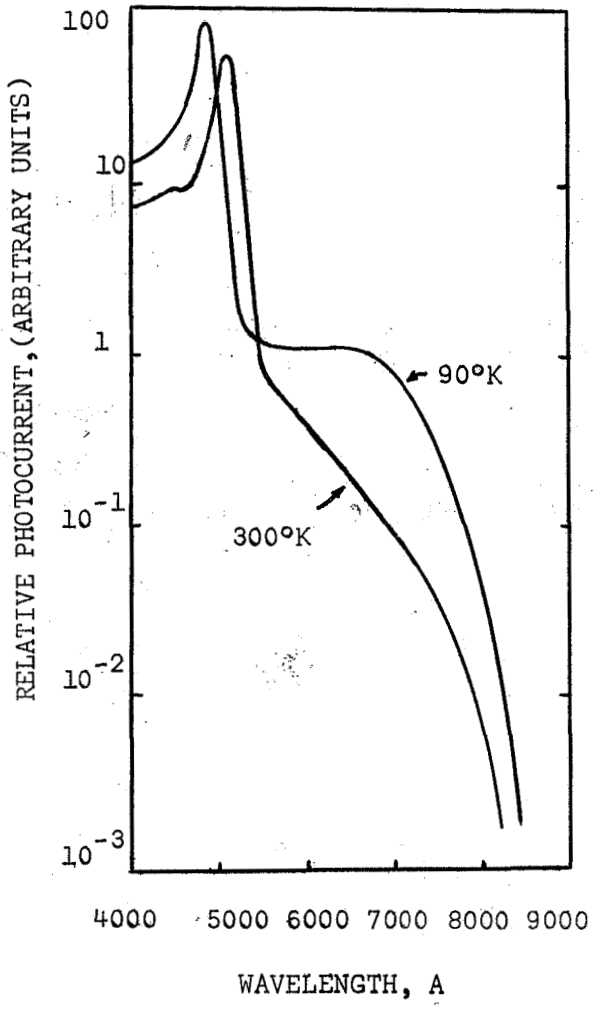


Figure 3.27. Spectral response of photoconductivity for a photosensitive CdS crystal with a trace of iodine. [69]

wavelength response shows a decrease in sensitivity. One can interpret this as a degrading of the near surface photoconductivity. Without pursuing the details, the important feature is that CdS can be considered as a photoconductor for the entire visible spectrum. In the following discussions the spectral sensitivity will not be considered in detail. However, the measures of performance imply wavelength dependence.

A photoconductor converts light energy into electrical energy and one of the measures of sensitivity is the change in electrical conductivity due to absorbed power in a unit volume. The units of electrical conductivity are $(\text{ohms} - \text{cm})^{-1}$ and the absorbed power is in units of watts/cm^3. Assuming the photoconductivity varies linearly with applied field and light intensity, the specific sensitivity is independent of geometry, applied field and incident light intensity. Under these conditions and further assuming that one type carrier predominates the specific sensitivity S_p is defined as

$$S_p = \frac{\Delta I}{\frac{V}{P}} \ell^2 \tag{3.46}$$

where ΔI is the photocurrent, V is the applied voltage, ℓ is the electrode spacing and P is the absorbed radiant power density. For an ideal photoconductor where the absorbed radiant power is uniform throughout the volume of the photoconductor

$$\frac{\Delta I}{V} = \Delta Y = \frac{\Delta \sigma A}{\ell} \tag{3.47}$$

where ΔY is the change in conductance (ohms^{-1}), $\Delta\sigma$ is the change in conconductivity $(\text{ohm} - \text{cm})^{-1}$, A is the area and ℓ the length between electrodes of the photoconductor. Using Equation (3.47), Equation (3.46) results in

$$S_p = \frac{\Delta \sigma}{P/v} \tag{3.48}$$

where v is the volume of the photoconductor. Thus the specific

sensitivity for photoconductors is somewhat analogous to responsivity for photoemission.

The incident light intensity (I_p) decreases with distance into the photoconductor as

$$I_p(x) = I_p(o)\, e^{-\alpha x} \tag{3.49}$$

where α is the absorption coefficient in cm^{-1} and the incident light intensity $I_p(o)$ is in photons/cm^2-sec. Thus the number of photons absorbed per unit volume per second is $\alpha I_p(x)$. There are two factors which should be introduced for convenience. If one is interested in minimum detectable incident light flux, the precise meaning of $I_p(o)$ must be introduced. For a given light flux external to the detector there will be some reflection from the surface. This concept is not included in Equation (3.49). This loss should be included in obtaining absolute measures of performance. However we are interested in orders of magnitude and will ignore reflection and assume $I_p(o)$ to be the external light intensity.

The generation of carriers per unit volume is

$$g_E(x) = \alpha I_p(x) \tag{3.50}$$

where $g_E(x)$ is a function of position and in units of electrons/cm^3-sec. It simplifies matters if we assume a uniform generation rate throughout the volume of the photoconductor. Our approximation for the generation rate of carriers is

$$g_E = \alpha I_p(o) \tag{3.51}$$

or that all the incident light flux is absorbed in a distance $\frac{1}{\alpha}$. Again this provides order of magnitude approximations.

The change in conductivity with absorbed radiation can be related to material parameters by

$$\Delta\sigma = \Delta n q \mu + n q \Delta\mu \tag{3.52}$$

where n is the free carrier density, e is the electronic charge, μ is the free carrier mobility, Δn is the change in free carrier density and $\Delta\mu$ is the change in free carrier mobility due to the absorbed radiant power. Under steady state conditions the excess carrier density Δn is related to the generation rate (g_E) resulting from absorbed radiant power and the lifetime of free carriers (τ) by

$$\Delta n = g_E \Delta\tau + \Delta g_E \tau. \tag{3.53}$$

Therefore Equation (3.52) becomes

$$\Delta\sigma = g_E q\mu\Delta\tau + q\mu\tau\Delta g_E + nq\Delta\mu. \tag{3.54}$$

For our purposes only the second term will be considered. Bube discusses the contributions of the first and third terms [69]. We will ignore these effects in this development and assume that the change in conductivity corresponds to a change in free carrier generation rate. Under these assumptions

$$S_p = \frac{q\mu\tau}{P/v} \Delta g_E \tag{3.55}$$

In this simple development the specific sensitivity is directly proportional to the free carrier lifetime. Data given in Table 7 presented

Material	Specific Sensitivity cm^2/ohm watt	Electron Life-time, sec.	Response Time at 10^{-4} ft-c, sec
Impure CdS (Contains recombination centers such as Ni)	10^{-9}	10^{-10}	——
Pure CdS (No spectrographically detectable impurities)	10^{-5}	10^{-6}	——
Sensitive CdS (Contains deliberately incorporated impurities such as Cl and Cu)	10^{-1}	10^{-2}	10
Special Sensitive CdS (Contains only traces of iodine impurity)	10^{-1}	10^{-2}	10^{-1}

Table 7 Properties of Various Cadmium Sulfide Photoconducting Materials [69]

by Bube tends to verify this simple concept [69]. The useful photoconductors are those with a specific sensitivity of 10^{-1} cm^2/ohm-watt and correspondingly have an electron lifetime of 10^{-2} seconds.

Another measure of sensitivity is the photoconductive gain which is analogous to quantum yield for photoemission. This gain is the measure of the number of times a photo excited electron traverses the length between electrodes during the free carrier lifetime. Actually as the elctron exists the photoconductor at the anode another takes its place at the cathode. The number of times this occurs for

each photo excitation represents a gain. The transit time (T_t) of the electron is

$$T_t = \frac{\ell}{\mu E} \tag{3.56}$$

where μ is the mobility in $\frac{cm^2}{volt\text{-}sec}$, E is the elective field in volts/cm and ℓ is the distance between electrodes in cm. Thus the photoconductivity gain is

$$G = \frac{\tau}{T_t} = \frac{\mu\tau V}{\ell^2} \tag{3.57}$$

where τ is the free carrier lifetime. Of course the above development depends upon near ohmic contacts and uniform electric field. Using representative values for CdS gains as high as 10^4 to 10^5 are not uncommon. Again we notice that τ is a figure of merit for the measure of sensitivity. (Actually $\mu\tau$ usually appear however the variation in lifetime τ is more diverse for materials than for mobility μ.)

For a practical photoconductor the response time is an important consideration. Given a change in light intensity how long does it take for the photoconductor to reach a new steady-state value of conductivity? It should be obvious that the response time is related to the free carrier lifetime. If the intensity is suddenly reduced to zero the excess carriers would remain for their lifetime. Thus for intrinsic photoconductors or those where the free carrier is lost through recombination the response time is the free carrier lifetime.

In actual practice the response time is observed to be much longer than the free carrier lifetime. This is explained by trapping centers which acquire a portion of the photo-excited carriers when the intensity is increased and results in a longer rise time. Trapped carriers must be released from traps when the intensity is decreased thus lengthening the decay time. The observed decay time has been observed to have the simple relation

$$\tau_o = \tau \frac{n_t}{n} \tag{3.58}$$

where τ_o is the observed decay time, τ is the free carrier lifetime, n_t is the density of trapped carriers within about kT of the steady-state Fermi level during excitation and n is the density of free carriers [72]. Considering the values of lifetime given in Table 7 the response time for CdS would be longer than 10^{-2} seconds. In fact, the response times given in Table 7 for CdS are a factor of 10^3 longer than the free electron lifetime.

From the discussion to this point there appear to be no fundamental limitations to sensitivity of CdS photoconductors. In fact the photoconductive gain may be increased by increasing the voltage or reducing the electrode spacing. However, there are physical processes which limit the photoconductivity gain. The primary consideration is space charge limited current which arises due to injection of carriers at high fields. This process limits the maximum gain to

$$G_{max} = \frac{\tau_o}{RC} M \tag{3.59}$$

where τ_o is the observed response time, RC is the dielectric relaxation of the dark material and M is a constant. [73] The constant M is determined by the ratio of field injected space charge in traps to the number of photo-excited carriers in traps. Values for M are usually near unity. Thus we are forced to increasing τ_o to increase the maximum gain. However the increase in gain is at the expense of the response time.

The ultimate limit for the sensitivity of a photoconductor is the minimum detectable signal resulting from absorbed radiant power. Ideally this limit will be determined by the signal to noise ratio for the photoconductor. The signal to noise ratio for photoconductors has been considered theoretically [74],[75]. Rose obtains an expression for the mean squared noise current (I_n^2) of a trap free photoconductor

$$I_n^2 = 4F\,(Gq)^2\,\Delta f \qquad (3.60)$$

where F is the total number of carriers per second excited by an incident light flux, G is the photoconductivity gain and Δf is the bandwidth required to observe the response time of the photoconductor. ($\Delta f = \frac{1}{2\,\tau_o}$). Without traps the response time is equal to the free carrier lifetime. Substituting the photocurrent (I)

$$\Delta I = qFG \qquad (3.61)$$

into Equation (3.60) yields

$$I_n^2 = 4\,\Delta I\,qG\,\Delta f. \qquad (3.62)$$

Thus the signal to noise ratio for the photoconductor is

$$\frac{\Delta I}{I_n} = 1/2\left(\frac{F}{\Delta f}\right)^{1/2} = 1/2\ (2F\ \tau)^{1/2} = \sqrt{2}\ N^{1/2} \qquad (3.63)$$

where N is the steady state number of free electrons with lifetime τ and generated by the photon flux of F excitations per second. The factor $N^{1/2}$ is also the signal to noise ratio for the incident photon flux [75]. Thus the signal-to-noise ratio for the photocurrent is essentially limited by the statistical fluctuations in the photon flux.

The photocurrent noise considered above is a generation-recombination noise source for a trap free photoconductor. The frequency band width of importance was $\Delta f = \frac{1}{2\tau}$ where τ is the free carrier lifetime. If free carrier trapping is important the bandwith Δf should be changed to accomodate the extended response time. However, the noise per unit bandwidth for the trap free case is

$$\frac{I_n^2}{\Delta f} = 4\ q\Delta I\ G, \qquad (3.64)$$

and the noise per unit bandwidth with trapping is

$$\frac{I_n^2}{\Delta f} = 4\ q\Delta I\ G\left(\frac{n_t}{n}\right)^2 \qquad (3.65)$$

where $n_t < n$ and n_t is the density of electrons in traps and n is the free carrier density. Thus for low density traps the noise

contribution per unit bandwidth is less for trapping than for trap free photocurrents. However as trapping increases the noise per unit bandwidth increases to that given in Equation (3.64).

At low applied fields photocurrent noise is not the limiting factor. Johnson noise or thermal noise will determine the minimum detectable signal. The photocurrent noise exceeds Johnson noise for an applied field (E)

$$E > \frac{kT}{qL_D} \tag{3.66}$$

where L_D is the distance a free carrier diffuses during its lifetime [75]. With typical values of $10^{-4} \lesssim L_D \lesssim 10^{-1}$ cm the transition from Johnson noise to photocurrent noise occurs in the range of 10^{-1} to 10^{2} volt/cm near room temperature. The mean-squared noise current per unit bandwidth for Johnson noise is

$$\frac{I_n^2}{\Delta f} = \frac{4kT}{R} \tag{3.67}$$

where k is Boltzmanns constant, T is the temperature in degrees Kelvin, and R is the resistance of the photoconductor. The ratio of Equation (3.64) to Equation (3.67) is G eV/kT where V is the applied voltage. For sufficiently high applied voltages or above the transition field in Equation (3.66) the photocurrent noise exceeds Johnson Noise.

One noise source neglected to this point is "1/f" noise. Rose and Shulman suggest that "1/f" noise is contact noise for CdS photoconductors and by obtaining ohmic contacts the contribution from "1/f" or excess noise can be reduced if not eliminated [75],[76].

The evidence cited is that for ohmic contacts with indium or gallium on CdS. The shot noise is orders of magnitude lower than for silver contacts which are presumably non ohmic. Possibly the most significant evidence to support their point of view is the shot noise measurements by Shulman [77]. The results compare favorably with a model based upon shot noise and approach within approximately a factor of 2 the photocurrent noise.

Considering only the photocurrent noise, an estimate of the theoretical minimum detectable radiant power can be obtained. The signal to noise ratio from Equation (3.63) is

$$\left(\frac{\text{Signal}}{\text{Noise}}\right) = \frac{S}{N_v} = \left(\frac{F}{4\Delta f}\right)^{1/2}. \tag{3.68}$$

Considering a $S/N_v = 1$ as the limiting factor

$$F_{min} = 4\ \Delta f \tag{3.69}$$

where Δf is related to the response of the photoconductor. For the trap free photoconductor of CdS

$$\Delta f = \frac{1}{2\tau} \simeq 50\ \text{sec}^{-1} \tag{3.70}$$

Thus the minimum total number of excitations per second F_{min} would be 200. With a typical distance of 10^{-2} cm between electrodes and assuming uniform excitations throughout the volume of the photoconductor the absorbed flux per unit area per second is 200 $\frac{\text{photon}}{\text{cm}^2 - \text{sec}}$ for 1 cm^2 surface area. This is approximately 8×10^{-17}

watts/cm at 5000Å. Of course this assumes an external quantum efficiency of unity or that all the incident photons are absorbed. Therefore the above estimate is an absolute limit and as such is extremely optimistic. Using an electron mobility of 200 cm^2/volt sec results in a photoconductive gain [68]

$$G = \frac{\mu \tau V}{\ell^2} = 2 \times 10^4 \text{ V.} \tag{3.71}$$

Therefore the photocurrent for this example is

$$\Delta I = F \ell G = 6.4 \times 10^{-13} \text{ V amps.} \tag{3.72}$$

The maximum gain is usually limited by space-charge limited current hence the applied voltage in Equations (3.71) and (3.72) will have limited values.

An estimate for the maximum applied voltage can be obtained from the concept that space-charge limited currents becomes important when [69]

V_{max} C= total charge in crystal

$$= qN \tag{3.73}$$

$$= qF\tau$$

Assuming an external quantum efficiency η_λ and a photon flux of I_p $\frac{\text{photons}}{cm^2 - \text{sec}}$

$$V_{max} = \frac{q F \tau}{C} = \frac{q \eta_\lambda I_p \tau \ell}{\varepsilon} .$$

Using typical values of $\varepsilon = 8.85 \times 10^{-13}$ $\frac{\text{farads}}{\text{cm}}$, $\tau = 10^{-2}$ sec, $\ell = 10^{-2}$ cm and assuming η_τ near unity

$$V_{max} = 1.8 \times 10^{-11} I_p A. \qquad (3.75)$$

For a absorbed flux of 200 $\frac{\text{photons}}{\text{cm}^2 - \text{sec}}$ over a 1 cm^2 area the maximum applied voltage is 3.6×10^{-9} volts. This result is unrealistic and yields a photocurrent of approximately 10^{-21} amps. It is obvious that we are unable to take advantage of the photoconductor S/N ratio with the trap free response time used in this example.

From a practical viewpoint the major falacy in the above example is using the response time of a trap free photoconductor for this extremely low light intensity. It is extremely difficult to achieve semiconductors with trap densities below 10^{12} cm^{-3}. Therefore the filling and emptying of these traps for low values of photo-excited carrier densities extends the response time orders of magnitude longer than the free carrier lifetime. In addition to lengthening the response time, the traps permit higher values of applied voltage before injected space charge limited current becomes important. For a trap density near 10^{12} cm^{-3} uniformly distributed in energy throughout the band gap of CdS the response time at 10^{-9} foot candles or 4×10^3 photons-cm^{-2} -sec^{-1} at 5000 Å wavelength is 7 hours. Thus the presence of traps extends the response time by a factor of 10^6. At 10^{-9} foot candles for the above example the maximum applied field is 0.2 volts and the photo-current is approximately 10^{-13} amps. This is within state-of-the- art measurements, however the response time is several hours. Thus the lower limit in sensitivity is not a very practical situation.

In the above treatment the dark current has not been considered. To detect the steady state photocurrent, the dark current should be much smaller than the photocurrent. For the above example CdS in excess of 10^{14} ohm-cm resistivity is required. Ohmic contacts are likely to be a significant problem. Non-ohmic contacts will introduce additional noise and influence the space charge limited current [76].

That the above exercise is somewhat unrealistic should not be too disconcerting. The problems one will encounter in attempting to achieve the ultimate sensitivity of a CdS photoconductor are made quite clear in this exercise. In addition we have ignored the electronics that must be used for these measurements and achieving ultimate sensitivity will depend critically on noise sources in the measuring circuitry.

Light intensity measurements between 10^{8} and 10^{10} $\frac{\text{photons}}{\text{cm}^2\text{-sec}}$ at 5000 Å appear to be typical low intensity applications. To relate CdS performance characteristics to infrared photoconductors, a specific detectivity (defined in Section 3.2.2.3) of 5×10^{14} cm Hz 1/2 watt near 5000 Å appears to be achievable [76a]. The major limitation in these instances is still the response time which maybe three to six orders of magnitude longer than the free carrier lifetime or from seconds to hours. Thus the price for low light intensity measurements with photoconductors is the long response time. In many practical applications long response times are unacceptable.

3.2.2.3 Infrared Photoconductors

Infrared transduction may be generally divided into three catagories: 1) photoconductivity 2) photovoltaic, and 3) thermal. Semiconductors and solid state technology have made it possible using these transduction

principles to span the entire infrared portion of the electromagnetic spectrum from 0.75 microns to 1000 microns. (The long wavelength limit is somewhat arbitrary). This accomplishment is the result of extensive research and development over the past twenty-five years. In this section we choose to only introduce infrared photoconductors. By doing so detracts somewhat from the accomplishments in infrared technology although later sections will introduce junction devices. Neglecting thermal detectors of infrared radiation surely obscures their relationship to photoconductors. However, the reader is referred to a wealth of literature which clearly treats the fundamentals of infrared detectors [78],[79],[80].

To cover all the photoconductors reported in the past twenty-five years is beyond the scope of our primary objective. However, we will attempt to focus on the more important materials that truly represent photoconductive transduction in the infrared. To accomplish this objective will require a treatment of two types of photoconductors classified as intrinsic and extrinsic. This distinction will not be considered in detail in this section. The subject is treated in Section 1. However, a few distinctions between these two processes will become apparent in developing the performance characteristics of each.

Since photoconductors are quanta detectors, the energy of the photons must be larger than a certain value before photoexcition of charge carriers occurs. Therefore we may anticipate some of the problems for infrared photoconductors. For an intrinsic photoconductor this energy is the band gap energy. For the extrinsic photoconductor this energy is the transition energy from the impurity level to the

appropriate band. We should remember that kT = 0.026 e. v. at room temperature and corresponds to a wavelength of 48.5 microns. Thus at room temperature thermally excited carriers will result in very insensitive infrared transduction.

One is usually interested in the maximum performance for a particular infrared detector. Of course this involves the usual concept of sensitivity and signal to noise. Visible photoconductors were limited either by internal noise sources or by fluctuations in the incident photon flux. For infrared photoconductors the primary concern is background radiation fluctuations. Other noise sources such as generation-recombination, Johnson, 1/f and amplifier noise are still present and in some instances are of greater intensity than background noise. However, the background noise is the lower limit in noise intensity for most applications of infrared detectors. In space applications where the background temperatures may be quite low statistical fluctuations in the incident photon flux may represent the limiting noise.

To compare optical transduction for infrared photoconductors with varying geometries and materials the concept of signal to noise ratio must be considered in some detail. Before proceeding some definitions will prove useful. The following definitions closely follow the work of Jones [81]. The detector input is radiant power and the output is an electrical signal. The radiant power is usually expressed as watts. The electrical signal may be either a voltage or a current. One must also be concerned about the method for providing the input power. If a chopped light source is used and the output measured as

the rms voltage of the fundamental frequency, the same measure should be applied to the input power.

To arrive at a characteristic of optical transduction which may be used as a measure of sensitivity the following concepts are introduced:

1) Responsivity (R) - The ratio of output to input in units of volts/watt. This implies a linear relationship; however an incremental responsivity may also be used for non-linear transduction.

2) Noise Equivalent Power (NEP) - The ratio of rms voltage of the noise and the responsivity in rms volts per rms watts.

3) Detectivity (D) - The reciprocal of the Noise Equivalent Power.

If the radiation flux density is P_λ watts/cm^2, A is the area of the detector in cm^2, N is the detector noise in rms volts and V is the detector signal in volts, the following relations hold;

$$R = \frac{V}{P_\lambda A} \quad \frac{\text{volts}}{\text{watt}} \tag{3.76}$$

$$NEP = \frac{NP_\lambda A}{V} \quad \text{watts} \tag{3.77}$$

$$D = \frac{V}{NP_\lambda A} \quad \text{watts}^{-1}. \tag{3.78}$$

When the noise source is black body radiation as with most infrared detectors the rms voltage of the noise is proportional to square root of the detector area and the square root of the noise bandwidth. To obtain a measure of performance independent of area and bandwidth of the electronic measurement (Δf) a specific detectivity (D*) has been defined

as

$$D^* = D\ (A\ \Delta f)^{1/2}\ \text{cm} - H_z^{1/2}/\text{watt}. \tag{3.79}$$

Wherever possible this figure of merit will be used. Finally it should be mentioned that the limiting noise may not be radiation noise in every case. In the event that another noise source such as Johnson noise or "1/f" noise predominates D* should be reconsidered. Using the definition in Equation 3.79 without regard to the noise source can be misleading.

The performance of an ideal photoconductor limited by statistical fluctuations in the background radiation is useful for comparison. If we define the nose equivalent power P_e for our ideal detector as the signal power producing a detector output equal to that produced by all background sources of fluctuation affecting the detector then [79]

$$P_e = 7.08 \times 10^{-18}\ \frac{T^{3/2}}{\lambda_\varepsilon}\ (\varepsilon A \Delta f)^{1/2}\ \{\int_{xo}^{\infty}\ \frac{x^2\ e^x}{(e^x - 1)^2}\ dx\}^{1/2} \tag{3.80}$$

where T is temperature in degrees Kelvin, λ_o is the longest wavelength to which the detector responds, ε is the emissivity of the detector, A is the detector area, Δf is the bandwidth for amplification, and the integral is over the Planck photon density with energies greater than the excitations energy of the photoexcited electrons. Using the definitions for detectivity and specific detectivity the performance of an ideal photoconductor limited by background radiation can be obtained. A plot of this result for 300 °K background is shown in Figure 3.28. The Planck black body distribution function for 300°K

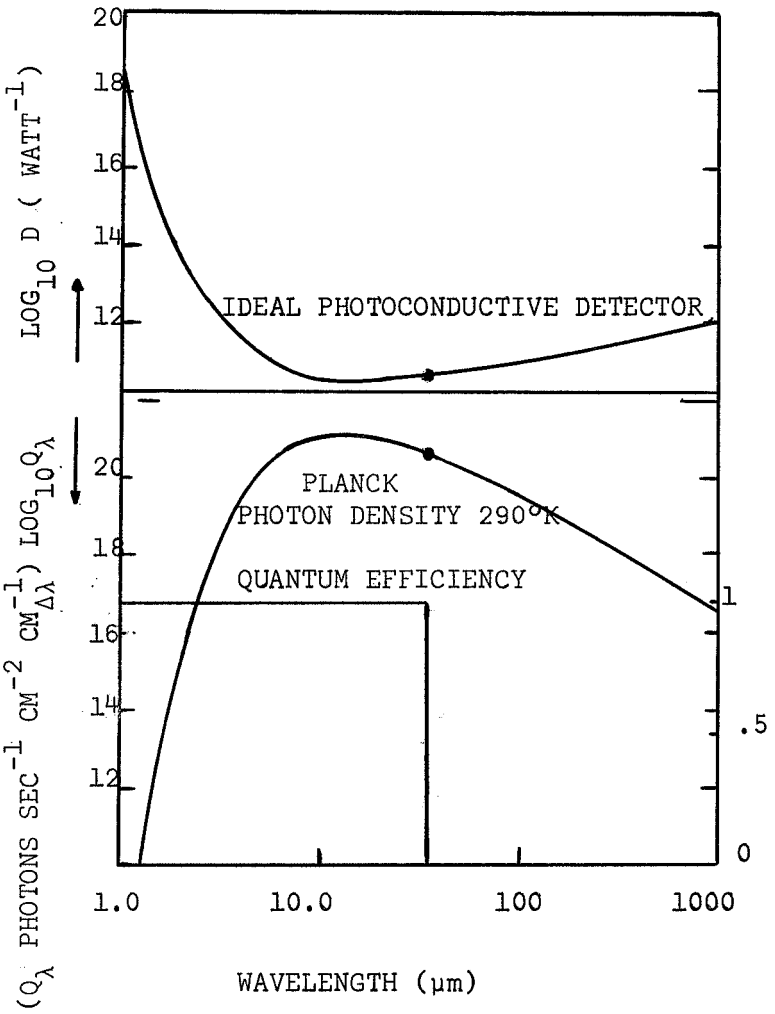


Figure 3.28. Detectivity of ideal detectors. The Planck photon distribution for 290°K black-body radiation and the detectivities for ideal thermal and photoconductive detectors exposed to this background are shown. The dot on the curve for photoconductive detectors indicates the detectivity of a detector of quantum efficiency (or spectral response) given by the lowest curve. [79]

is also plotted in Figure 3.28. The major feature is the minimum in specific detectivity occuring near 10 microns and the maximum in the Planck photon density. As the long wavelength cutoff of the detector

increases the total blackbody radiation capable of exciting electrons increases and the noise equivalent power increases or the detectivity decreases. Beyond the peak the photon density from black body radiation increases the noise equivalent power by a relatively small amount due to the decreasing background photon density. Later the concept of an ideal photoconductive detector will prove useful in comparing the performance of the various materials used throughout the infrared.

The intensity of the incident photon flux I_o which we wish to detect decreases with distance in the photoconductor as

$$I_p(x) = I_p(o)\, e^{-\alpha x} \tag{3.81}$$

where α is that part of absorption related to free carrier excitation. If most of the incident flux is to be absorbed and result in free carriers the photoconductor thickness (ℓ) should be at least greater than $1/\alpha$. For intrinsic photoconductors the absorption coefficient near the band-edge absorption is the order of 10^4 cm^{-1}. Therefore thin films approximately one micron thick may be used for intrinsic photoconductors.

In extrinsic or impurity type photoconductors the absorption coefficient is

$$\alpha = \sigma N_t \tag{3.82}$$

where σ is the optical cross section for the impurity level and N_t is the density of impurity atoms. The density of impurities is limited by their solid solubility in the host material. As a result the absorption coefficient for exciting the impurity levels is much

lower than for intrinsic photoconductors. A value of $\alpha \simeq 0.14$ cm^{-1} has been given for gold in germanium [78]. For this case, the extrinsic photoconductor should be approximately 10 cm thick.

The essential features of the infrared photoconductor are identical to the visible photoconductor. The generation of carriers is assumed to be uniform throughout the volume of the photoconductor and equal to

$$g_E = \alpha I_p (o). \tag{3.83}$$

The excess carrier density (Δn) in steady state is

$$\Delta n = g_E \tau \tag{3.84}$$

where τ is the free carrier lifetime. Consistent with our previous definitions we may consider a noise source due to carrier density fluctuations either induced by the lattice or by random arrival of background photons. This will be a generation-recombination mean square fluctuation in current given by

$$I_n^2 = 4\Delta I G q \Delta f = 4 \frac{\Delta I^2}{F} \Delta f. \tag{3.85}$$

consistent with Equation (3.62).

Thus the signal to noise is

$$S/N_v = \frac{\Delta I}{I_n} = 1/2 \left(\frac{F}{\Delta f}\right)^{1/2} = F/2 \left(\frac{1}{F\Delta f}\right)^{1/2}. \tag{3.86}$$

To indroduce the dependence of S/N_v on the material parameters and on the incident radiation consider

$$F = g_E A \ell. \tag{3.87}$$

Thus

$$S/N_v = \frac{g_E A_\ell}{2} \left(\frac{1}{g_E A\ell\Delta f}\right)^{1/2} = \frac{\alpha I_p^{(o)}}{2} \left(\frac{\tau A^{1/2}\ell^{1/2}}{\Delta n \Delta f}\right)^{1/2}. \tag{3.88}$$

Assuming the photoconductor thickness is equal to $1/\alpha$

$$S/N_v = \frac{I_p(0)\ \alpha^{3/4}}{2} \left(\frac{\tau}{\Delta n}\right)^{1/2} \left(\frac{A^{1/2}\ 1/2}{\Delta f}\right) \tag{3.89}$$

$$= \frac{\phi_p(0)\ \alpha^{3/4}}{2} \left(\frac{\tau}{\Delta n}\right)^{1/2} \left(\frac{1}{A\Delta f}\right)^{1/2}$$

where ϕ_p (o) is the total incident flux rate in photons/sec. To convert to a detectivity we introduce the energy per photon as hc/λ where h is Plancks constant c is the speed of light and λ is the wavelength of the incident photons. Thus the detectivity is

$$D = \frac{\lambda\ \alpha^{3/4}}{2hc} \left(\frac{\tau}{\Delta n}\right)^{1/2} \left(\frac{1}{A\Delta f}\right)^{1/2} \tag{3.90}$$

or the specific detectivity is

$$D^* = \frac{\alpha^{3/4}\ \lambda}{2hc} \left(\frac{\tau}{\Delta n}\right)^{1/2}. \tag{3.91}$$

In this simple model we see some of the features of the more accurate and elaborate models. To increase the detectivity one can increase the free carrier lifetime. However we are reminded that the response time also lengthens. We may choose to reduce Δn to n_o the dark carrier

concentration. Thus our earlier comments regarding thermally activated carriers becomes relevant. To achieve a maximum D* the dark free carrier concentration should be reduced by cooling the detector. Of course one need only cool the detector until D* approaches that for an ideal photoconductor.

The major feature missing in Equation (3.91) is the proper wavelength dependence for the specific detectivity. For a practical photoconductor the absorbed radiation for equal intensities at all wavelengths will follow the dependence of absorption coefficient on wavelength. Thus what has been done above is only a relative value which should apply only at a given wavelength. In addition the cut off in photoresponse at the band gap energy is not apparent from the simple model. However these factors are apparent in Figure 3.29 from the plot of specific detectivity from 1 to 40 microns for various infrared photodetectors [82]. Also included in Figure 3.29 are some photovoltaic detectors which will be discussed in a later section.

The major feature in Figure 3.29 are the near ideal performance of the photoconductors over the range from 1 to 40 microns. At the shorter wavelength from 1 to 6 microns intrinsic photoconductors predominate while at longer wavelengths extrinsic photoconductors have proven superior. As expected all the detectors are cooled to 77 °K or below to approach the ideal photoconductor. At the shorter wavelengths it is possible to operate near 300°K with a sacrifice in detectivity. For the indium antimonide photoconductor the detectivity decreases two orders of magnitude as the temperature is increased from 77°K to 295°K.

The response time of a photoconductor is also an important characteristic in many applications. Typical values for a variety of photoconductors are shown in Table 8. (Also shown are other typical values for detectivity, useful wavelength and resistance). The response times cover the range from 40 nanoseconds to 3.3 milliseconds or five orders of magnitude. The shorter response times are for extrinsic photoconductivity in germanium. The high degree of structural purity and the short free carrier lifetime are primarily responsible.

The longest response times are for the lead salts. The unique feature of lead sulfide and lead selinide is their performance at or near room temperature. The detectivity can be increased by cooling

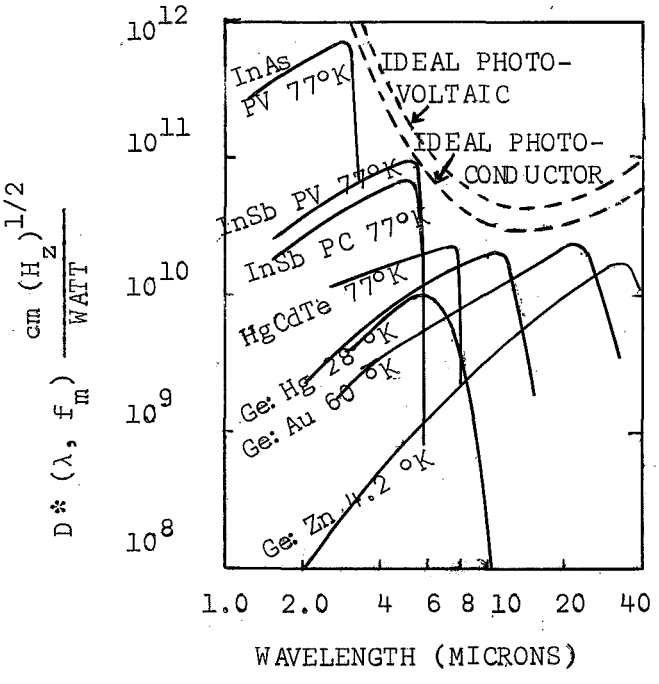


Figure 3.29. The spectral detectivity of a number of low temperature infrared detectors. Typical frequency of the noise measurement was 1000 Hz [82].

Type	Operating temp.(°K)	Useful wavelength range (μm)	Wave-length of max. detectivity	Detectivity D* (or D) at optimum wavelength	Response time(μsec)	Resistance
Si-PV	295	<1.2	0.9	2.5×10^{12}	1000	450kΩ
Tl_2S-PC layer	295	<1.2	0.9	2.2×10^{12}	500	5MΩ
Ge-PV	295	<1.8	1.6	4×10^{11}		
PbS-PC layer	295	<3	2.5	8×10^{10}	300	500kΩ
PbS-PC layer	193	<3.4	2.6	4×10^{11}	3500	
PbSe-PC layer	295	<4.5	3.5	10^9	2	
PbSe-PC layer	77	<7	5	10^{10}	40	
PbTe-PC layer	77	<5.2	4	2.7×10^9	25	30MΩ
InAs-PV	295	<3.7	3.4	7×10^9	1	20Ω
InAs-PV	196	<3.5	3.2	7×10^{10}	1	50Ω
InAs-PV	77	<3.4	3.0	3×10^{11}	2	<10MΩ
InSb-PC	295	<7.8	6	3×10^8	0.05	5Ω
InSb-PC	77	<5.5	5	5×10^{10}	5-10	2000Ω
InSb-PV	77	<5.5	5.4	7×10^{10}	<1	
Ge-Au-IPI	60	<9	5	3×10^{10}	<1	1MΩ
Ge-Hg-IPI	30	<14	12	3×10^{10}	0.06	>1MΩ
Ge-Cu-IPI	15	<30	25	4×10^{10}	0.1	>3000Ω
Ge-Zn-IPI	4	<40	35	2×10^{10}	∿ 1	∿ 1MΩ
Ge-Ge-IPI	4	<125	104	3×10^{11}	0.04	∿ 10kΩ
Ge-B-IPI	4	<140	108	2×10^{11}	0.1-0.01	∿ 10kΩ

PV, photovoltaic; PC, photoconductive; IPI impurity photoionization.

Table 8 Examples of Photoconductive Detectors [79]

however the response time increases as well. The exact reasons for this behavior is not well documented however free carrier trapping at low temperatures is a likely explanation.

To illustrate the measurement using a photoconductor consider idium antimonide. The photocurrent ΔI is given by

$$\Delta I = F\ q\ G \tag{3.92}$$

where F is the total number of excitations per second and G is the photoconductive gain. Using Equations (3.57) and (3.74)

$$\Delta I = \frac{E q \mu \tau V_{max}}{\ell^2} = \frac{(F\ q\ \tau)^2 \mu}{\varepsilon A \ell} \tag{3.93}$$

where the maximum applied voltage is determined by the space-charge limited current for a trap free photoconductor. This assumption will be discussed after the estimate of photocurrent. From the definition of noise equivalent power an estimate of the minimum detectable photocurrent considering only background noise limitations may be obtained. The noise equivalent power may be equated to an incident flux I_p by

$$NEP = \eta_\lambda\ P_\lambda(min)\ A \tag{3.94}$$

where η_{τ_λ} is the external quantum efficiency, P_{τ_λ} (min) is the minimum detectable incident radiant power density and A is the detector area. Using the defintion of specific dectivity

$$NEP = \frac{A^{1/2}\ \Delta f^{1/2}}{D^*} = n_\lambda\ P_\lambda\ (min)\ A \tag{3.95}$$

For a monochromatic light flux density I_p ($\frac{photons}{cm^2 - sec}$) with an energy E_λ (electron-volts) the incident radiant power density is

$$P_\lambda \text{ (min)} = I_p \text{ (min)} \; q\, E_\lambda \tag{3.96}$$

and

$$\eta_\lambda \; I_p \text{ (min)} \; A = \frac{A^{1/2} \; \Delta f^{1/2}}{D^* q \; E_\lambda} \quad . \tag{3.97}$$

The incident flux is related to the excitations by

$$F = \eta_\lambda \; I_p \; A \tag{3.98}$$

assuming a quantum efficiency near unity and that all incident photons are absorbed uniformily in the volume of the photoconductor. Therefore the photocurrent for this simple model is

$$\Delta I = \frac{\Delta f \; \mu \tau^2}{(D^* \; E_\lambda)^2 \; \varepsilon \ell} \tag{3.99}$$

If one chooses to measure the photocurrent within the response time τ the bandwidth of the measuring equipment must be

$$\Delta f \sim \frac{1}{\tau} \tag{3.100}$$

and under this condition the photocurrent is

$$\Delta I = \frac{\mu \tau}{(D^* \; E_\lambda)^2 \varepsilon \ell} \quad . \tag{3.101}$$

Using typical values of $\mu = 78{,}000 \; \frac{cm^2}{v\text{-}sec}$, $\tau = 10^{-5}$ seconds, $D^* = 5 \times 10^{10}$ $(cm\text{-}Hz)^{1/2}$ - watt, the length between electrodes $\ell = 10$ microns, the dielectric constant $\varepsilon = 1.5 \times 10^{-13} \; \frac{farads}{cm}$ and the energy at the maximum detectivity $E_{\tau \simeq}$ 1/4 electron volt, yields a

value of photocurrent of approximately 10^{-5} amps. Detecting 10 micro amps with a band with of 10^{5} Hz presents a significant challenge however the measurement is state-of-the-art. A unity bandwith measurement would result in a significant decrease in the minimum detectable incident power density however a photocurrent of 10^{-10} amps must be detected. This is a routine measurement with a response time approching 1 second. The trap free assumption appears reasonable for 10^{-5} amps photocurrent with a I_p (min) of approximately 10^{11} $\frac{\text{photons}}{\text{cm}^2 - \text{sec}}$. However at 10^{-10} amps the assumption is not justified. However, one can operate infrared photoconductors under a variety of conditions. Full utilization of maximum detectivity and minimum response time also appear feasible.

3.2.3 Photo-Junction Transduction

A variety of junction devices have been used as photodetectors. The primary examples are p-n homojunctions either diodes or transistors, Schottky barrier diodes, avalanche diodes, p-i-n diodes, heterojunctions and metal-insulator-semiconductor devices. A theoretical and phenomenological description of these various junction devices are included in Section 2. We will avoid duplication by considering only performance characteristics such as responsitivity and detectivity.

The operation of photojunction devices depends upon the absorption of the incident photons in the depletion region or within a minority carrier diffusion length of the depletion region. For the hole-electron pairs created in the depletion region, the charge pair separate in the applied field and drift to the appropriate boundary. The pairs created

outside the depletion region and within a minority carrier diffusion length of the depletion region separate by diffusion. If the minority carrier diffuses to the boundary it will drift across the depletion region. Hole-electron pairs created beyond a minority carrier diffusion length from the depletion region ultimately recombine with appropriate carriers without reaching the depletion region. In a photojunction device nearly all the applied voltage is dropped across the depletion region and only those charge carriers which traverse a portion or all of the depletion region contribute to the photocurrent.

From the rather simple introduction two major features of photojunction devices become apparent. The response time of the device to changes in light intensity may be as short as the transit and diffusion times for the carriers in a particular material. In addition the maximum response with wavelength should occur when most of the photoexcitations are within a diffusion length of the depletion region.

The detectivity for photojunctions implies either a measure of noise or adequate knowledge of the noise sources. If the signal to noise ratio varies as the reciprocal of the square root of the device area the specific detectivity D^*_λ is independent of area. This criteria is predicated on background noise limitations. However as discussed in Section 2 junction devices are often limited by other noise sources such as 1/f or excess noise. In this case the reported values for D^*_λ must be carefully examined before meaningful comparisons can be made.

In this section silicon devices will receive most of the attention. Since state-of-the-art is our primary goal this choice is certainly justified. Only in the infrared is silicon lacking in performance. Although the emphasis is on silicon other materials will be introduced

where appropriate. In many instances meaningful comparisons will not be possible. However we will emphasize the inherent advantages and difficulties of the research and development for the non-silicon devices. In most instances we will take for granted well established silicon technology.

3.2.3.1 Ultraviolet Photojunctions

There have been relatively few investigations of junctions devices as UV detectors. [83],[84],[85] To characterize the work to date we will consider two regions of the ultraviolet spectrum. The near ultraviolet (200 $\mathring{A}$ - 3500 $\mathring{A}$) where silicon photodiodes will perform although the detectivity is rather low and the far ultraviolet 7-20 ev where the photoresponse increases with increasing photon energy. The behavior in each of these regions can be understood by considering four factors; the absorption coefficient as a function of wavelength shown in Figure 3.30, the reflectivity as a function of wavelength shown in Figure 3.31, the yield or electron-hole pairs per absorbed photon and the surface recombination of excess carriers.

The maximum response of a photojunction device without gain would occur when all the incident photons are absorbed within a diffusion length of the junction and all the photoexcited electrons enter the depletion region. There are two factors which reduce the efficiency in the UV. Some of the incident photons are reflected (30 to 70%) from the surface reducing the efficiency or detectivity. In the ultraviolet the absorption coefficient is greater than but approximately equal to 10^5 cm^{-1}. Therefore most of the excess carriers are created within 1000 $\mathring{A}$ of the surface. Normally the carriers recombine

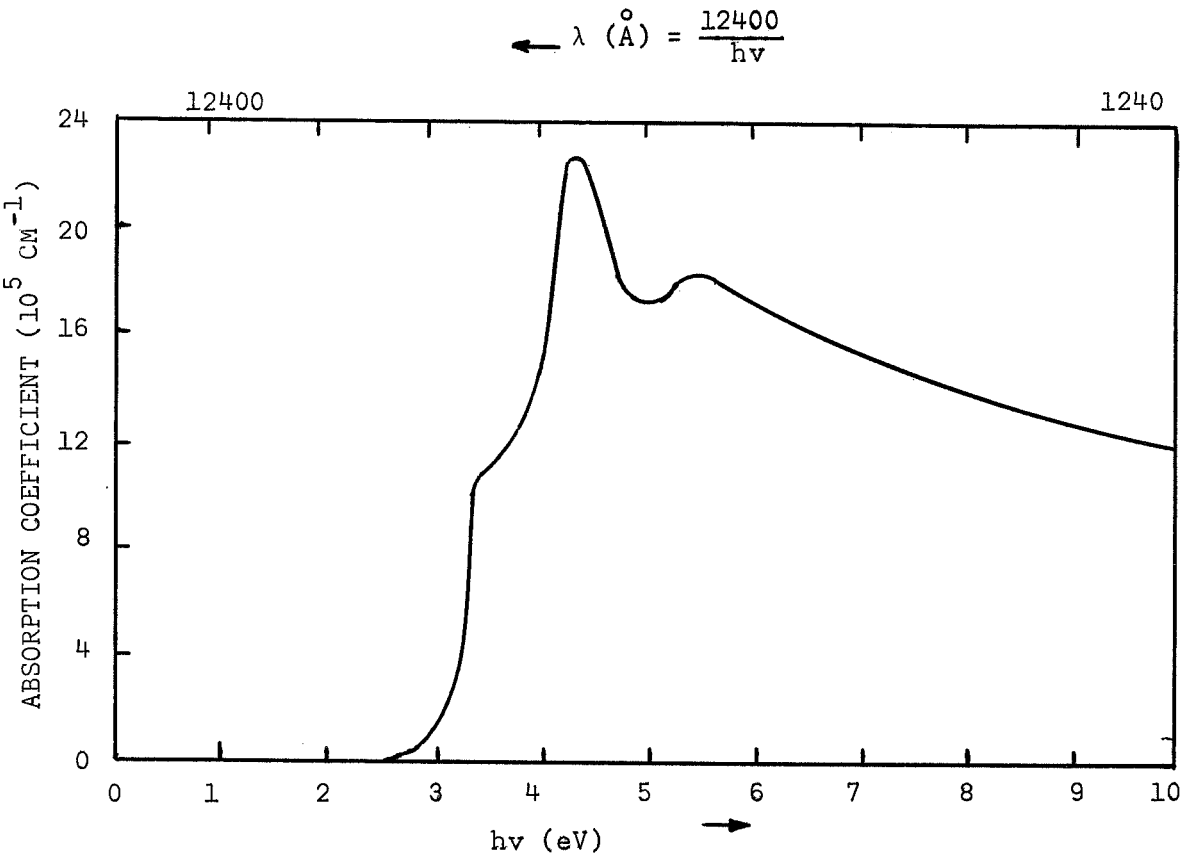


Figure 3.30. Absorption coefficient of silicon as a function of wavelength. [83]

near the surface at a faster rate than in the bulk. Therefore the number of carriers which diffuse to the depletion region will decrease as the absorption coefficient increases. This effect is more pronounced the deeper the junction.

Silicon junction devices do respond to ultraviolet radiation in the 2000 - 3500 Å range as shown in Figure 3.32. The rapid fall off in this range can be explained by the above discussion. The initial decrease near the visible appears to correlate with the increase in

reflectivity while the decrease for wavelengths shorter than 4000 Å appears to correlate with the rapid increase in absorption coefficient. Therefore the responsivity hence the spectral detectivity are limited by the reflectivity and quantum yield of the device. To improve the performance of p-n junctions in the UV special attention must be given to anti-reflecting coatings and to surface properties. Because of the need for low surface recombination velocity silicon is the obvious choice.

Another device which offers some promise in the 2000 - 3500 Å range is the Schottky barrier. The quantum efficiency for a gallium

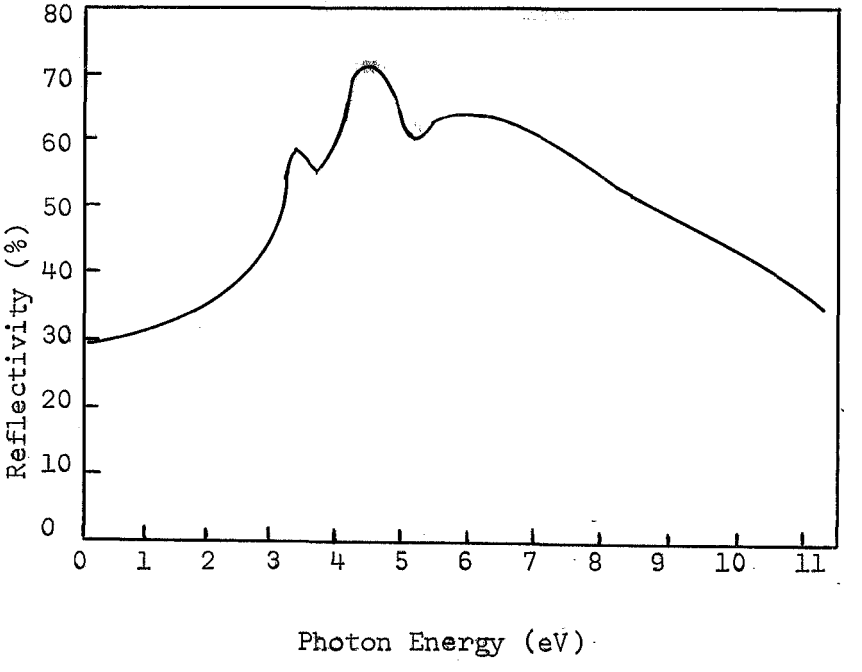


Figure 3.31. Normal incidence reflectivity as a function of photon energy for single crystal, etched silicon at 300°K. [86]

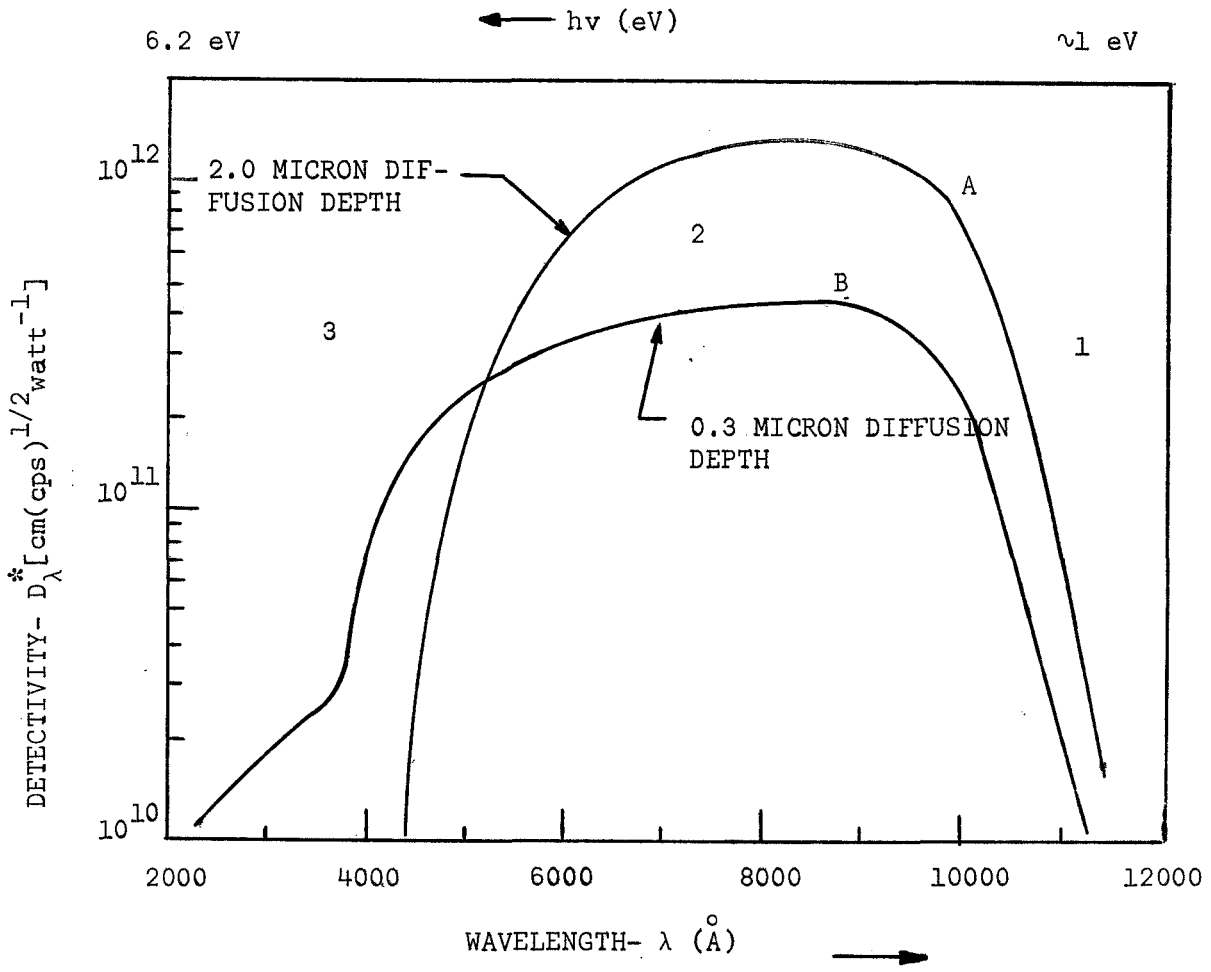


Figure 3.32. Spectral responsivity of typical silicon junction devices. [32]

arsenide with chromium field plate Schottky-barrier photodiode has one distinct advantage over the p-n junction in that the depletion region extends to the surface of the semiconductor. Since the UV is absorbed within 1000 Å most of the created hole-electron pairs should contribute to the photocurrent. Collection efficiencies in excess of 90% have been reported [87]. However there is one distinct disadvantage. The field plate is a metal and represents an optical

barrier in the UV. The metal field plate will reflect and absorb the incident photons and reduce the responsivity and detectivity of the Schottky-barrier photodiode. The narrow optical transmission window of silver thin films near 3225 Å has been used to obtain rather narrow band response [87]. However imporvement of the broad response for the Cr-GaAs device will require additional research into the optical properties of conducting films in the ultraviolet.

In the 7-20 ev range the results cited in Section 2 appear to offer promise for semiconductor junction devices. Photodiodes have been operated in this range and have proven to be more sensitive than thermopiles [83]. Absolute measure of detectivity were not given however the results do warrant further consideration. The increased yield with internal gain provides increased sensitivity. There is some question about the reflectivity of silicon in this range however a gradual decrease over a portion of the range provides additional encouragement.

3.2.3.2 Visible Photojunctions

There has been an enormous effort devoted to photodiodes and solar cells which may be used to detect photons in the visible portion of the optical spectrum. In this section we will briefly discuss the various devices and materials presently considered as research and development. The concept of minimum detectable light flux and spectral response will be considered to characterize the photodiode performance. One will readily recognize that the emphasis given to silicon is not in proportion to its importance.

The major activities in the development of visible photodetectors may be divided into four catagories 1) Schottky-barrier diodes, 2) heterojunctions, 3) P-I-N photovoltaic diodes and 4) p-n junctions. Each of these devices is treated in detail in Section 2. Before proceeding to a discussion of the differences and problems associated with these devices and the materials and technology used to fabricate them, the similarities should be considered.

In general, the photocurrent can be approximated by

$$J_p = q \alpha \overline{L} N \tag{3.102}$$

where α is the absorption coefficient $\overline{L}$ is an effective length accounting for the collection of ionized hole-electron pairs, N is the absorbed photon flux. The dependence of these parameters on device and material properties has received detail attention in Section 2 and here we will attempt to generalize with simplifying assumptions. Near the peak responsivity $\alpha \overline{L} \gtrsim 1$ which merely states that most of the photons are absorbed within $\overline{L}$ and contribute to the photocurrent. The absorbed photon flux is less than the incident flux due to reflectivity. Normally the reflectivity for most semiconductors considered will vary less than a factor of two throughout the visible. The reflectivity may be included in a general way by designating the ratio of incident flux ϕ_p to absorbed flux N as

$$\frac{N}{\phi_p} = \gamma_\lambda \tag{3.103}$$

where γ_λ is a loss factor including reflectivity. Under these assumptions near the peak in responsivity

$$J_p \simeq q \gamma_\lambda \phi_p. \tag{3.104}$$

Maximum performance of these devices will depend upon the dark current for time indepent flux and upon the noise sources in the devices for time varying flux. For the time independent flux a minimum detectable flux may be estimated from

$$\phi_p \simeq \frac{J_D}{q \gamma_\lambda} \tag{3.105}$$

where J_D is the diode dark current. Dark current for a junction device may vary over a wide range, however we will choose 10^{-6} amp/cm^2 for discussion. Assuming $\gamma_\lambda \simeq 1$ results in a detectable flux $\phi_p < 6 \times 10^{12}$ $\frac{\text{photons}}{\text{cm}^2 - \text{sec}}$. Considering state-of-the art silicon diodes the actual current measurement may approach 10^{-12} amps which can present difficulties.

When the diode current is a modulated photocurrent described by

$$I_p = I_{po} + I_{p1} e^{j\omega t} \tag{3.106}$$

where I_{po} is much greater than the dark current I_D and I_{p1} is the magnitude of the time varying photocurrent, the photodiode performs essentially as a noiseless detector. The limiting factor in this case is generation-recombination noise due to statistical fluctuations in the incident photon flux. The mean square value of noise current is

$$\langle I_\eta^2 \rangle = 4 q I_{po} \Delta f \tag{3.107}$$

where Δf is the bandwidth of the measuring circuit. The signal to noise ratio is

$$\left(\frac{I_{pl}^2}{2\langle I_\eta^2\rangle}\right)^{1/2} = \frac{S}{N_v} = \left(\frac{I_p^{2\,1/2}}{4\,q\,I_{p_o}\,\Delta f}\right)^{1/2}. \tag{3.108}$$

Defining the modulation index $m = \frac{I_{pl}}{I_{po}}$ and using Equation (3.102) the signal to noise ratio is

$$S/N_v = \left(\frac{m^2 \alpha \bar{L}\, NA}{8\,\Delta f}\right)^{1/2} \tag{3.109}$$

or proportional to the statistical fluctuations in the absorbed photon flux $N^{1/2}$.

Although the unit bandwidth signal to noise ratio in Equation (3.109) appears optimistic we must reexamine Equation (3.108) as I_{po} approaches the dark current. Considering the dark current I_D as the limiting shot noise source Equation (3.108) becomes

$$\frac{S}{N_v} = \left(\frac{I_p^{2\ 1/2}}{2\,q\,I_D\,\Delta f}\right)^{1/2} \tag{3.110}$$

or using Equation (3.102) and $\alpha\,\bar{L} \simeq 1$ at the peak in responsivity

$$S/N_v = \left(\frac{q}{4\,I_D\,\Delta f}\right)^{1/2} N\,A. \tag{3.111}$$

Therefore the minimum detectable flux is

$$N_m\,A > \left(\frac{4\,I_D\,\Delta f}{q}\right)^{1/2}. \tag{3.112}$$

At flux rates approaching the values obtained from Equation (3.112), dark current noise replaces statistical fluctuations in the photon flux as the limiting factor in detectivity. For a dark current of 10^{-12} amperes and a unit bandwidth the minimum detectable flux rate is 5×10^{3} $\frac{\text{photons}}{\text{sec}}$. Assuming the flux uniformly distributed over a typical device area of 10^{-6} cm^2 yields a flux density of 5×10^{9} $\frac{\text{photons}}{cm^2 - \text{sec}}$. For photons with energy near 2 ev the noise equivalent power is approximately 10^{-15} watts. Applying the conventional definition of specific detectivity yields a value near 10^{12} cm $(Hz)^{1/2}$/watt. This compares favorably with the results for the silicon photojunction in Figure 3.32.

A major advantage of junction photodiodes is the short response times approaching 10^{-10} seconds. To fully utilize this response time requires a measurement bandwidth of approximately 10^{10} Hertz. In this case the minimum detectable flux is

$$N_m A > 2 \times 10^{5} \left(\frac{I_D}{q}\right)^{1/2} \qquad (3.113)$$

or approximately 5×10^{8} $\frac{\text{photons}}{\text{sec}}$ for a dark current of 10^{-12} amperes.

Returning to the various devices, we will consider Schottky barrier diodes first. Much of the work has been devoted to large area solar cells [88],[89]. Gallium arsenide was chosen over silicon in an attempt to obtain a better match between the solar emission spectrum and the material absorption spectrum. The merits of this consideration are emphasized by Figure 3.33. The importance of

gallium-arsenide or gallium-arsenide alloys is more pronounced at higher temperatures. However, the results in Figure 3.33 neglect reflection and leakage currents in the device. Thus we must return to technology to convert the ideal to reality. In practice silicon solar cells may achieve 14-15 % efficiency while gallium-arsenide thin film cells achieve 5% efficiency.

The spectral response of Schottky barrier cells depends upon absorption and reflection of incident photons by the barrier layers. Spectral response for platinum and cuprous selenide on gallium arsenide are shown in Figure 3.34. Similar data for the same

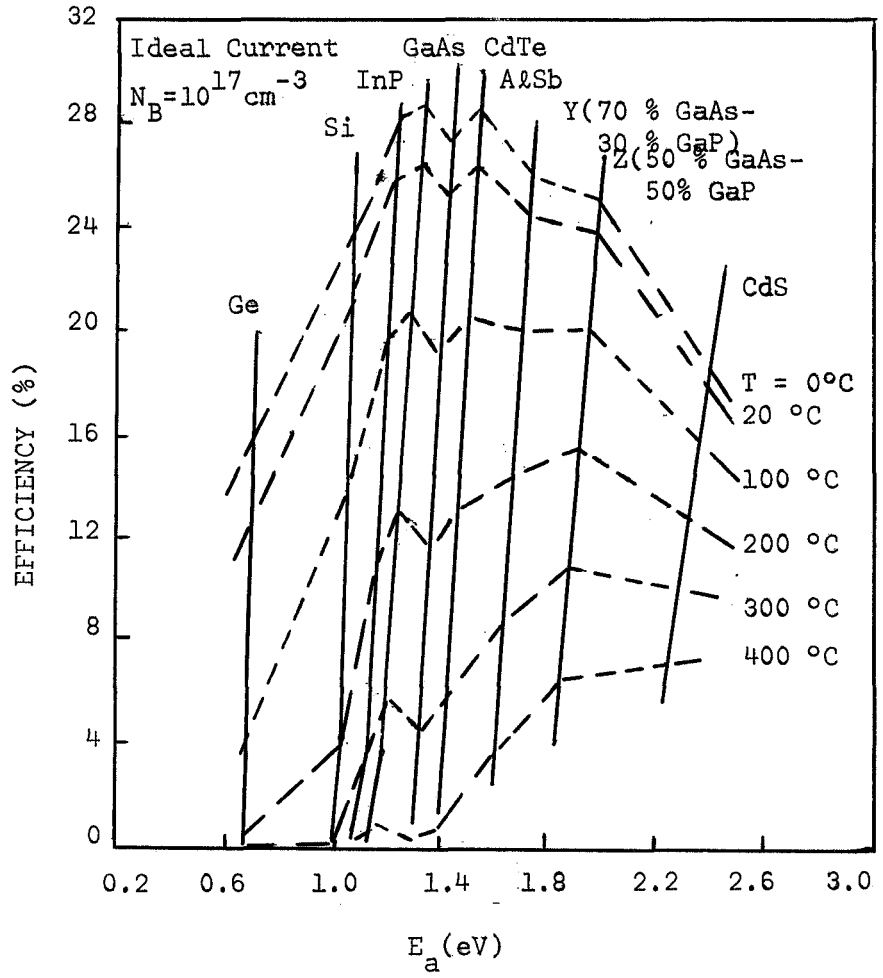


Figure 3.33. Conversion efficiency as a function of energy gap for ideal current-voltage characteristics. [90]

barriers on silicon are shown in Figure 3.35. Also shown for comparison is the response of a diffused p-n junction.

The major advantages of Schottky barrier devices are: 1) In principal a photdiode can be constructed on any semiconducting material. 2) The absorption of photons near the surface does not result in reduced sensitivity as for the diffused p-n junction since the field extends to the surface. 3) The technology is inherently simple. 4) Large area thin film structures are feasible. 5) Arrays are rather simple extensions of the single device.

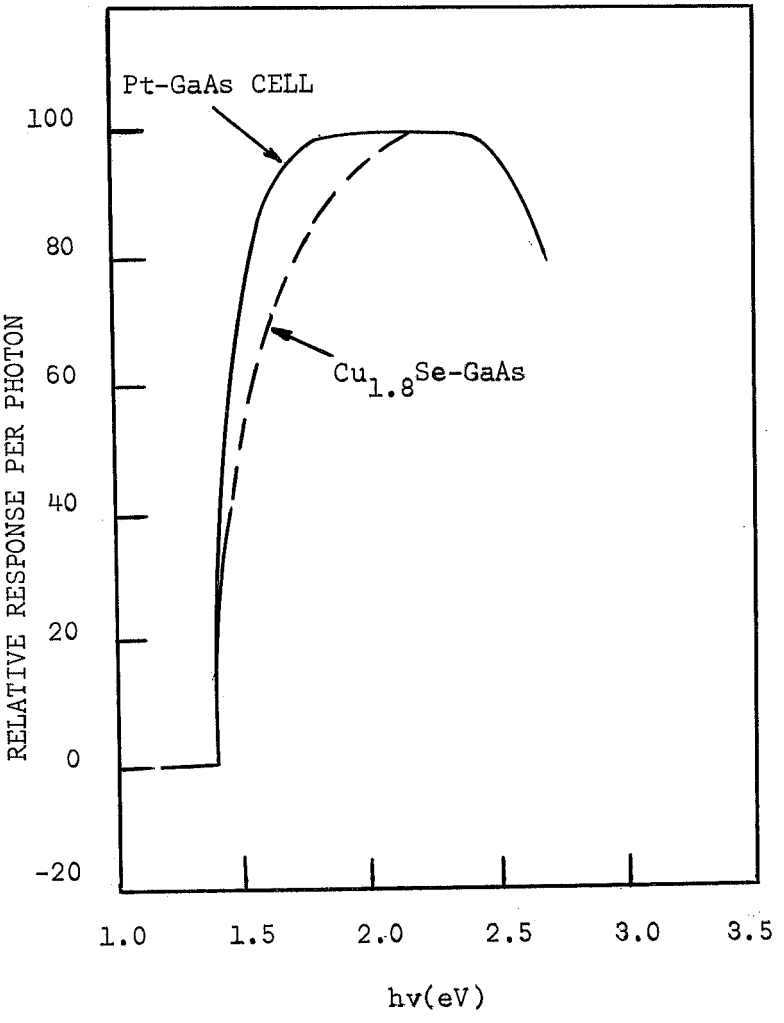


Figure 3.34. Spectral response of Pt and $Cu_{1.8}Se$ barrier cells. [88]

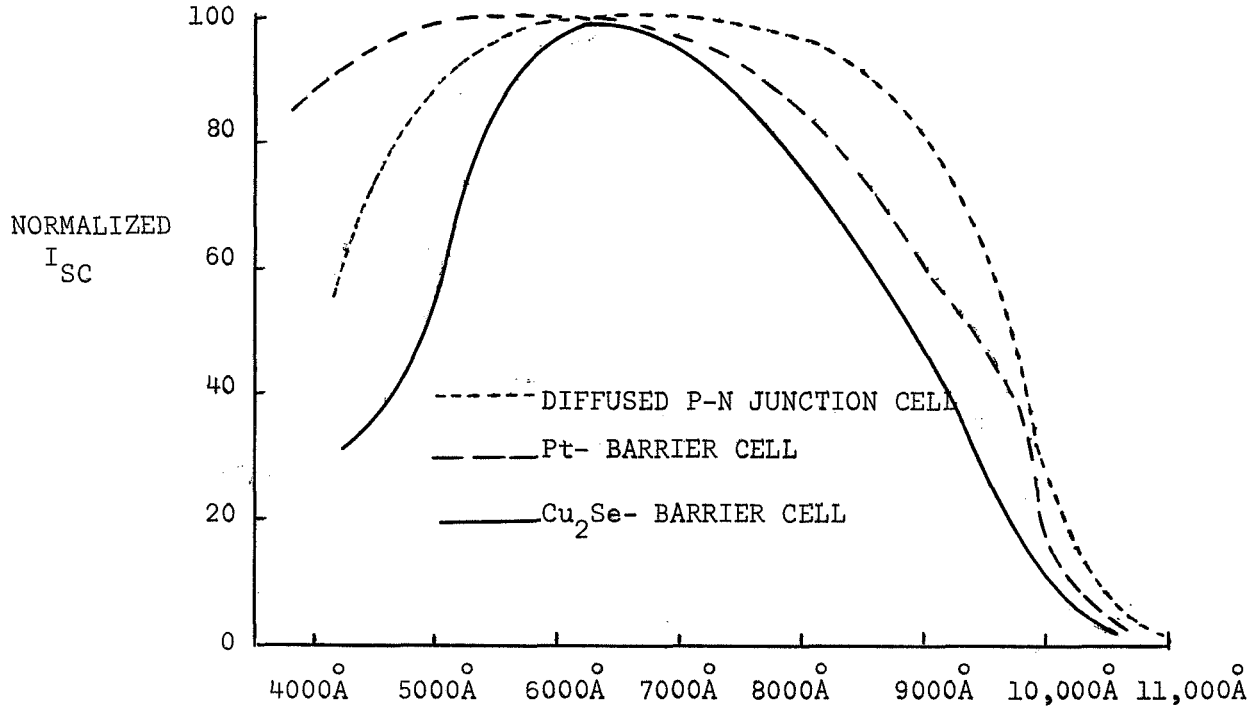


Figure 3.35. Spectral responses of various types of Si cells. [89]

The major disadvantages are 1) The photon flux must penetrate the barrier electrode. 2) Antireflection coatings are desirable for many applications. 3) Finding the appropriate barrier material which results in the appropriate barrier height to ensure low leakage and yet provides acceptable optical properties is pretty much a trial and error process. 4) Technology beyond silicon and to some extent gallium arsenide is essentially undeveloped.

The classic heterojunction is the cadmium sulfide solar cell. Excellent reviews of early work and contemporary characteristics are given by Shirland [92] and Rappaport [93]. In both articles the major justification for heterojunction work is large power to weight ratio for conversion of solar energy. However, Rappaport indicates that advances in silicon solar cell performance and large arrays are proceeding at least in pace with if not exceeding the development of larger area heterojunction arrays. However the payoff for optical transduction in the visible is rather impressive. In Table 9 are listed the materials considered to be most important in the visible. The efficiency of conversion of sunlight may be considered as a relative figure of merit. For a particular application, not associated with solar energy conversion, any of the various materials may prove adequate.

In general a heterojunction may be obtained with any two of the above materials in essentially a sandwich structure with a resulting p-n junction. However CdS and CdTe solar cells are usually formed by a surface reaction with copper. Ignoring these details we can generalize somewhat. The various materials are either direct or indirect band gap where the appropriate designation may be inferred by the information in Table 10. The absorption coefficient near the band edge is significantly higher for direct bandgap materials hence the thickness required to absorb most of the incident photons of energy near to but greater than band gap energy is less for direct bandgap materials. This difference is important in at least two cases. Thin film devices appear to be more reasonable with direct band gap materials. Rather sharp spectral response in terms of selective wavelength sensitivity

Material	Bandgap Energy eV	Theoretical Efficiency %	Experimental Max. Sunlight Efficiency %	References
Si	1.11	22	14	5
InP	1.25	24	6.7	13
GaAs	1.40	26	11	15
CdTe	1.45	23	7.2	16
GaP	2.23	19	1	17
CdS	2.4	18	8	18

Table 9 Efficiencies of Some Solar Cell Material [93]

is more feasible with direct band gap semiconductor heterojunctions.

The spectral response of a heterojunction photodiode in principle can be varied considerably depending upon the materials used in constructing the device. A representative response is shown in Figure 3.36 for the geometrical arrangement also shown in the same figure. Region number 1 is assumed to be a direct band gap semiconductor of thickness $\Delta\ell$ and band gap energy E_1. Region 2 is assumed to be a direct band gap semiconductor of thickness Δt and band gap energy E_2. For direct band gap materials the absorption coefficient rises rapidly from near unity to near 10^5 cm^{-1} as the energy of the incident photons varies from slightly

Material	Bandgap Energy in eV.	Thickness (microns)	% Utilization (estimated)
Si	1.11	100	85
InP	1.25	1	95
GaAs	1.40	2	90
CdTe	1.45	10	90
CdS	2.4	1	Anomalous behavior
GaP	2.23	10 - 100	90

Table 10 Minimum Film Thickness Required for Various Materials [93]

less than bandgap energy to slightly greater than bandgap energy. Consider the situation where $\Delta\ell >> \Delta t$ and $E_1 > E_2$. Assuming that $\Delta\ell >> 1/\alpha_1$ where α_1 is the absorption coefficient in region 1, there will be little photo response for wavelengths such that $\lambda_1 \lesssim \frac{h_c}{E_1}$ where h is Planck's constant and c is the speed of light. As the wavelength increases beyond λ_1 the absorption is reduced in region 1 and the photons reach the depletion region near the interface between the materials. Since region 2 is a direct bandgap material and the photon energy is greater than the bandgap energy most of the photons will be absorbed near the junction and contribute to the short circuit photocurrent. Since the absorption coefficient remains near 10^5 cm^{-1} until

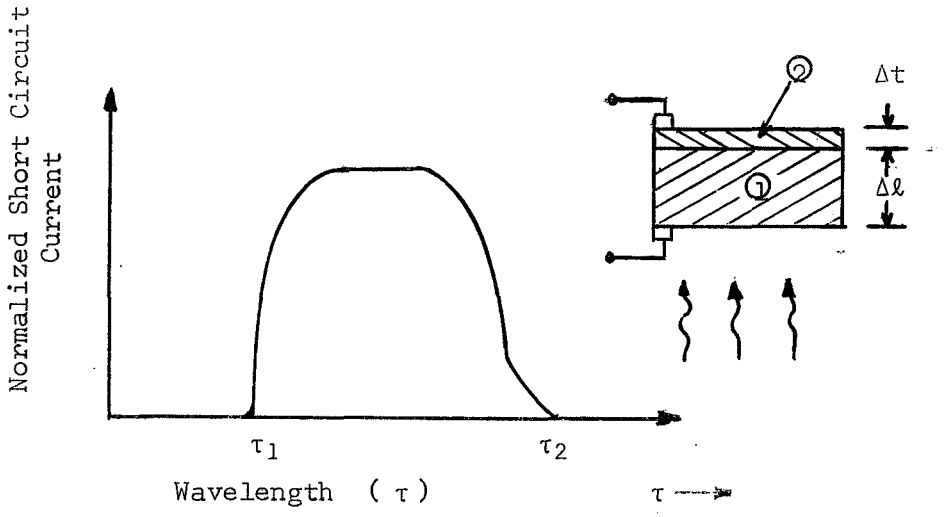


Figure 3.36. Pass band response of heterojunction.

the photon energy is less than the bandgap energy the response remains relative constant between λ_1 and λ_2 where $\lambda_2 \approx \frac{hc}{E_2}$. In principal the response bandwidth may be varied by proper choice of materials. However in practice it still remains a challenge to realize heterojunctions from a variety of materials. To appreciate this point, a recent publication deals with heterojunction solar energy conversion and calculates theoretical maximum attainable efficiency for ten heterojunctions of common semiconductors such as Si, GaAs, In P and GaP. For comparison only two experimental values are reported [94]. In addition the efficiency is only half that for silicon homojunctions.

The major advantages for heterojunctions are: 1) The spectral response is relatively constant over the response bandwidth. 2) Selectivity for the response bandwidth appears feasible. 3) Large area devices or arrays have been realized.

The main disadvantages for heterojunctions are: 1) The technology is often unreliable and difficult to reproduce. 2) Materials cost is usually greater than for silicon. 3) Junction properties are not well understood for the many possible materials combinations. However, much of the work with heterojunctions has been directed to solar cells and many of the disadvantages are rapidly disappearing. Although heterojunction technology and performance are improving, Rappaport [93] does not forsee a strong challenge for silicon solar cells in the near future. Therefore, one should look for uniqueness when considering heterojunctions for optical transduction.

Silicon photocells are usually diffused junction type. Figure 3.37 shows the physical configuration for a typical solar cell and a photodiode. To maintain low source impedance operation of the solar cell the contact to the n-region in this example covers the entire back surface. The p-region contact ususally consists of a soldered ridge the length of the cell. Further reduction of series resistance can be obtained using a grid of contacts extending along the p region normal to the solder ridge. These precautions in reducing series impedance are not usually required for photodiode operation.

When photons are absorbed and carriers generated within a minority carrier diffusion length of a p-n junction its electrical characteristics are altered. A rather complete description of how the characteristics change are included in Section 2. The essence of the change is a shift in characteristics as illustrated in Figure 3.38 where the I-V plot with and without incident photons are shown. With reverse bias the photocurrent depends upon excess carrier generation and collection by the depletion region. This is the normal mode of operation for a photodiode.

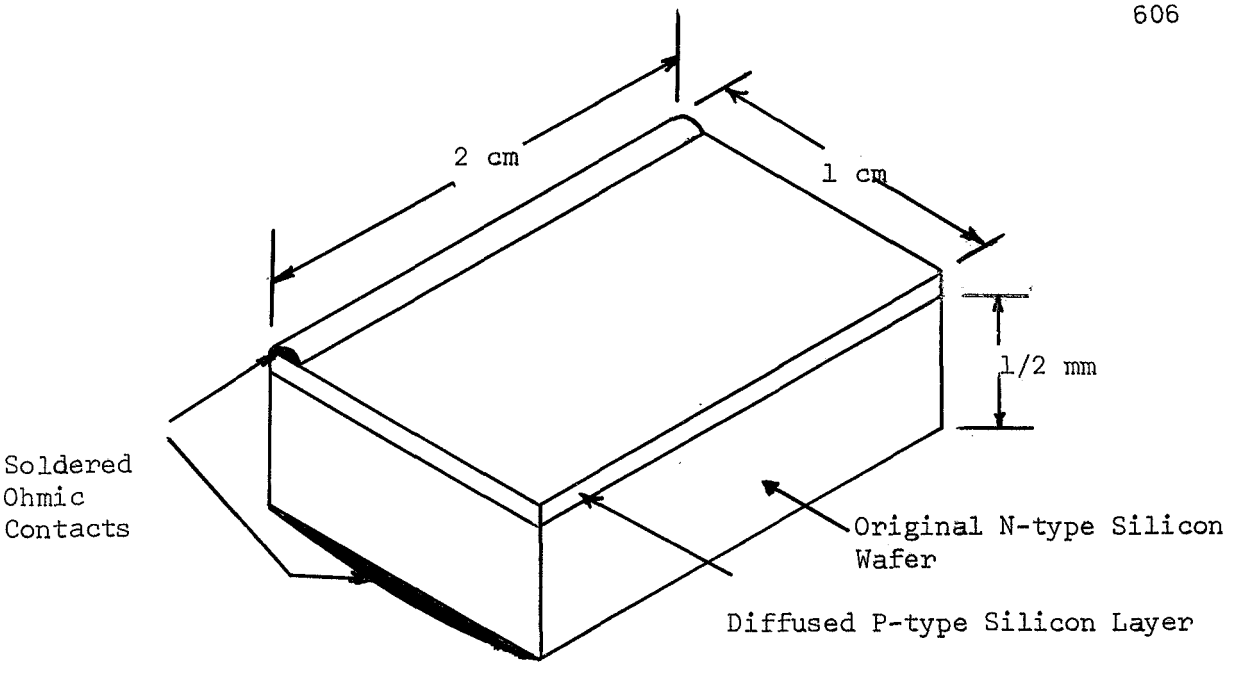


Figure 3.37 Silicon solar cell [96]

If the junction is open circuited without bias in the presence of visible light a voltage exists between the terminals which may approach the magnitude of the internal barrier potential. With a finite impedance the load may receive power from the junction as a solar cell. This operation is illustrated by the fourth quadrant in Figure 3.38.

The characteristics of solar cells and photodiodes depend upon contact geometry, the resistivity of the starting material and the processing used to form the p-n junction. In addition, much concern for nuclear radiation sensitivity has influenced the design of silicon solar cells. Much of the early work considered n+ on p silicon cells due to their superior radiation resistance. However p on n have received considerable attention with the discovery that lithuim drifted solar cells are more radiation resistant than convention n+ on p cells [96].

Concern for radiation resistance and solar energy conversion focused attention on 0.1 to 10 ohm-cm starting material. Concern for preserving minority carrier lifetime hence long diffusion lengths emphasized the technology of junction formation in particular diffusion. As a result junction depths between 0.3 and 0.9 microns have been realized with energy conversion efficiencies between 10 - 15% [97].

Surface reflectance is important in optimizing solar cell performance. Anti-reflection coatings improve cell efficiency by 15 to 30%. The coating may be Krylon at low temperatures or evaporated lithuim oxide. Sapphire windows are often employed to reduce the nuclear radiation intercepting the solar cell in a space application.

A primary factor in the operation of photodiode or photovoltaic cell is the efficiency with which the generated hole-electron pairs contri-

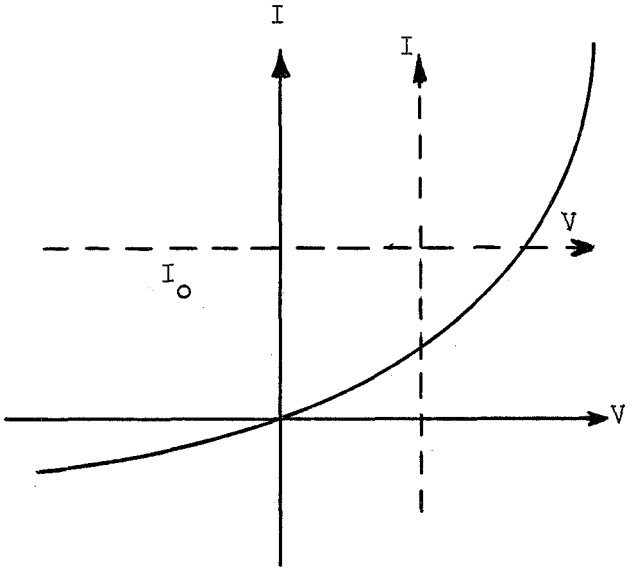
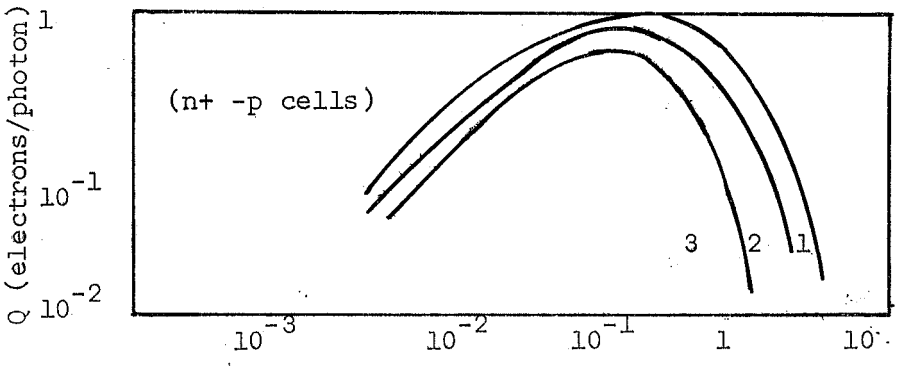


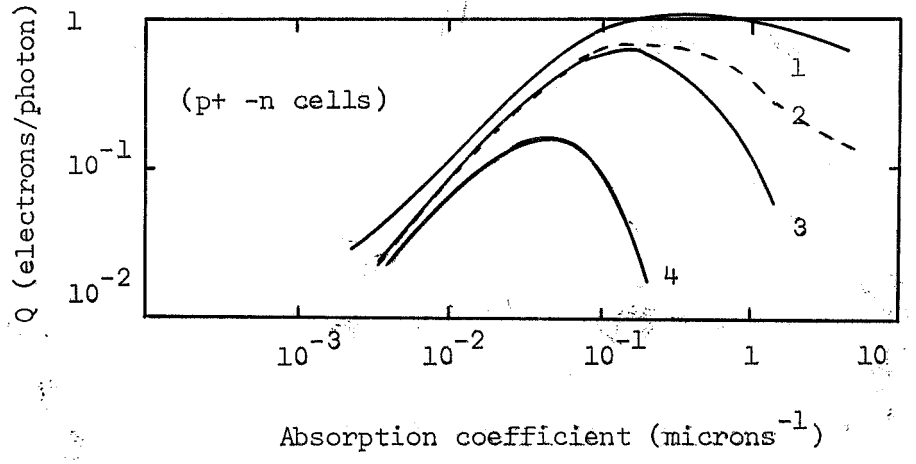
Figure 3.38 Effect of illumination on junction's . The characteristic of the dark junction is referred to the solid axes. The characteristic of the illuminated junction is referred to the dotted axes.

bute to junction current. To describe this efficiency analytically one must consider the thickness of the p-type and n-type regions, the minority carrier lifetimes, diffusion constants, surface recombination of excess carriers, absorption coefficient and reflectance. The details of such an analysis is included in Section 2. The essence of this analysis relates the collection efficiency to the density of excess minority carriers within a diffusion length of the depletion region. Thus the collection efficiency should be a function of junction depth. The results shown in Figure 3.39 indicate a definite dependence. As the wavelength of incident photons decreases the absorption coefficient increases such that most of the photoexcitations occur in the p^+ region. Since the depletion region does not extend very far into a degenerate or highly doped region the primary method for collection is thru minority carrier diffusion. From the results in Figure 3.39 the minority carrier diffusion length is between 1 and 10 microns. Certainly the 25 micron junction depth is well beyond the minority carrier diffusion length.

The junction depths given in Figure 3.39 are approximate so that the difference in n+-p (n+on p) and p+-n (p+on n) may result in part from uncertainties in junction depth. However a difference is expected for short wavelengths and large absorption coefficient ($\sim 10^4$ cm^{-1}) resulting from longer diffusion length for electrons in the p+ region as compared to holes in the n+ region. The results in Figure 3.39 are consistent with this intepretation where representative collection efficiencies are plotted as a function of absorption coefficient. In the short wave length and large absorption coefficient region the collection efficiency of the p+-n cells are clearly superior to the n+-p cells.



curve	junction depth (μ)
1	1
2	2
3	3



curve	junction depth (μ)
1	1
2	5
3	10
4	25

(Only absorbed photons are counted; the electrons are those of the short circuit current.)

Figure 3.39 Quantum efficiency, Q, of silicon solar cells [98].

The dependence of photovoltaic cell response on wavelength with constant light intensity at each wavelength is shown in Figure 3.40. Also shown is the relative response for equal number of photons at each wavelength. The response is normalized to unity for each set of data. The sharp cutoff at approximately 1.1 ev corresponds to the bandgap. For ultraviolet radiation with photon energies greater than 3 ev the hole-electron pairs are generated near the surface. Since the surface usually exhibits a high surface recombination velocity and acts as a minority carrier sink, the response is expected to fall off in the UV.

The sensitivity of a silicon photodiode to a photon flux in the visible is illustrated by Figure 3.41. With a reverse bias the current is nearly linear with photon flux. As indicated in Equation 3.105 the lower limit of detectable flux is established by the dark leakage current. Considering a lower limit for reverse leakage current density in a silicon p-n junction of 10^{-6} amp/cm^2 results in a minimum flux of approximately 10^{12} $\frac{\text{photons}}{\text{cm}^2 - \text{sec}}$ near the peak in spectral response. For a sinusoidally varying flux the minimum detectable value approaches 10^9 photon/cm^2-sec and a specific detectivity near 10^{12} cm $(\text{Hz})^{1/2}$/watt. Specific detectivity as a function of wavelength for a typical silicon photojunction device is shown in Figure 3.32.

The major advantages of silicon photodiodes are 1) The available technology is absolutely superior to that for any other material. 2) Large arrays have been produced. 3) Material cost is low compared to high quality III-V compound semiconductors. 4) The electrical properties of silicon are well understood. 5) There is a rich history associated with solar cells which can be applied for optical transduction in the visible.

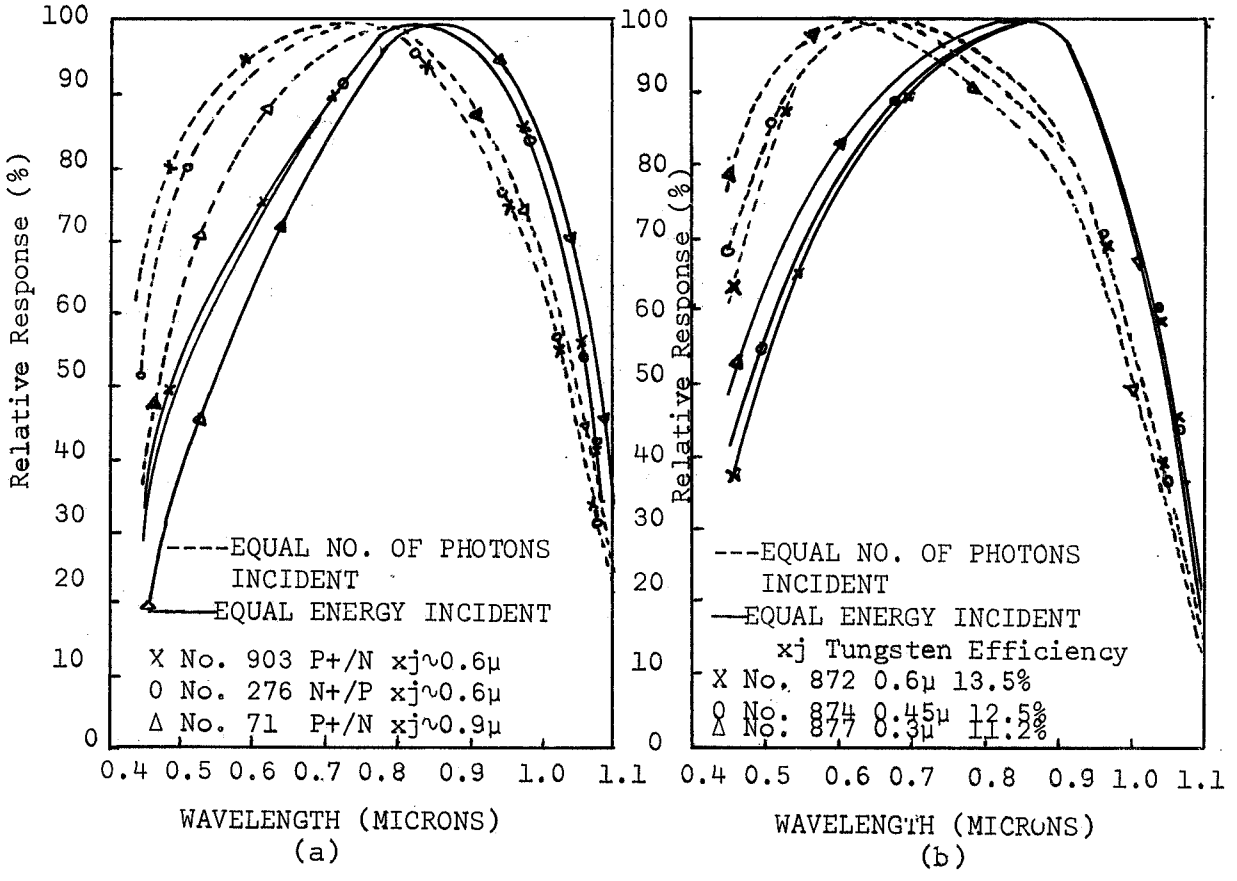


Figure 3.40 (a) Spectral response for p +/n cells No. 903, No. 71 and n + /p cell No 276. Symbols for identification. Measured points each 0.02μ (b) Variation of spectral response of p + /n cells as junction depth is decreased. Symobls for identification. Measured points each 0.02 μ. [97]

There is only one major disadvantage with silicon photodiodes in the visible region. At room temperature noise sources associated with the dark current limits the specific detectivity and background limited performance cannot be achieved. In principle a photodiode from a material with a bandgap near 1.6 ev or a threshold of response nearer the long wavelength visible region (∿ 7500 Å) would have a lower value of dark current and a larger specific detectivity. In practice this can be a

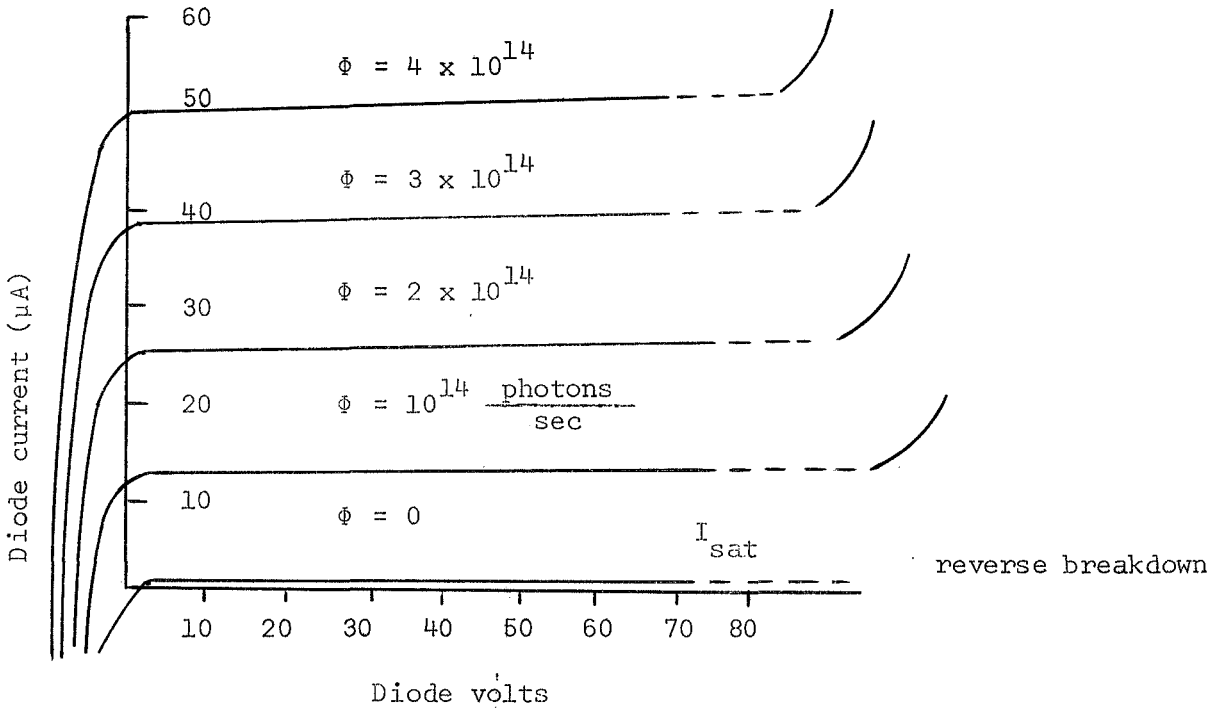


Figure 3.41 V-I-Φ characteristics of a silicon photodiode [99].

significant challenge. The primary problem is the dominance of generation-recombination currents in wide bandgap semiconductor devices. Improvements for silicon and wider bandgap semiconductors must rely upon higher quality materials.

3.2.3.3 Infrared Photojunctions

Photojunctions in the infrared are not nearly as well developed as silicon in the visible. However, there has been considerable effort devoted to the 1-15 micron range with varying, degrees of success. In the 1-5 micron range indium arsenide (InAs) and idium antimonide photovoltaic devices have been particularly successful while the 8-14 micron

range lead-tin-telluride has achieved marginal performance. Beyond 15 microns the photovoltaic devices are not particularly attractive when compared to the performance of extrinsic photoconductors (germanium: gold) considered in Section 3.2. In this section the peculiar aspects of infrared photojunctions and the performance of InAs, InSb, and (P_b S_n T_e) will be considered.

The most significant problems for infrared sensors are the noise sources and the reverse saturation current which limit the sensitivity. Usually the device must be cooled below room temperature and often approaching 77°K to achieve optimum performance. As illustrated in Figure 3.38 the photojunction may be characterized by

$$I = I_s(e^{qV/kT}-1) - I_p \tag{3.114}$$

where I is the total current from the device, I_s is the dark reverse saturation current, V is the applied voltage and I_p is the excess photocurrent. The photocurrent is characterized by

$$I_p = q\ N\ A\ \alpha\ \overline{L} \tag{3.115}$$

consistent with Equation (3.102). If the device is reverse biased the total current (Is + Ip) gives rise to shot noise as discussed in Section 3.2.2.2.

The photovoltaic mode of operation is more common for infrared sensors. In this case the device is operated without bias and the change in "open-circuit" voltage is sensed across an external load

resistor. From Equation (3.114) the open circuit voltage is

$$V_{oc} = \frac{kT}{q} \ln \left(\frac{I_p}{I_s} + 1\right). \tag{3.116}$$

In this mode of operation the noise source associated with the device is best described by a differential resistance (R_d)

$$\frac{1}{R_d} = \frac{dI}{dV}\Big|_{V=0} \tag{3.117}$$

determined thru Equations (3.114) and (3.115). Typically this resistance is less than 500 ohms for the devices we will consider. The mean square noise voltage is given by

$$\langle V_n^2 \rangle = 4\ kT\ R_d\ \Delta f. \tag{3.118}$$

The noise contribution from the external load resistance will also be described by Equation (3.118) and since it must be larger than R_d the noise source due to R_d may be neglected. We have neglected 1/f noise which may predominate at low frequency operations.

For a time independent photon flux, Equation (3.116) can be used to estimate the performance characteristics of the photovoltaic device. When $I_p << I_s$ the open circuit voltage is given by

$$V_{oc} = \frac{kT}{q} \frac{I_p}{I_s} \tag{3.119}$$

or the photovoltaic device is linear with incident light intensity.

Assuming one can measure one microvolt and using a reverse leakage current density of 10^{-8} amps/cm^2 at 100 °K yields an internal photocurrent density of 1.2×10^{-13} amps/cm^2. Furthermore, assuming unit quantum efficiency, an absorption coefficient of 10^5 cm^{-1} and an effective length $\bar{L} = 10^{-4}$ cm yields a photon flux of $7.5 \times 10^5 \frac{\text{photons}}{\text{cm}^2\text{-sec}}$ or for 1 ev photons a power density of $1.2 \times 10^{-13} \frac{\text{watts}}{\text{cm}^2}$. When $I_p \simeq 0.1\ I_s$ the open circuit voltage one must measure is approximately 0.8 millivolts. Considering the same values of quantum efficiency, absorption coefficient, effective length $\bar{L}$ and photon energy yields a photon flux of $1.8 \times 10^9 \frac{\text{photons}}{\text{cm}^2 - \text{sec}}$ and a power density of $3 \times 10^{-10} \frac{\text{watts}}{\text{cm}^2}$.

The ultimate performance is determined by the noise sources in the measurement. As mentioned previously the load resistance provides one noise source while 1/f contribution from the device is expected at low frequencies. Considering a photocurrent (I_p).

$$I_p = I_{p_o} + I_{p_{max}} \sin \omega t \tag{3.120}$$

where I_{p_o} is a d.c. component and $I_{p_{max}}$ is the a.c. amplitude, an estimate of signal to noise ratio for the photovoltaic device can be obtained. Assuming the limiting noise source to arise from the load resistor R_L the mean square noise voltage is

$$\langle V_\eta^2 \rangle = 4kT\ R_L\ \Delta f \tag{3.121}$$

where Δf is the measurement bandwidth. The signal to noise ratio is

$$\frac{V_{oc}(rms)}{V_n(rms)} = \frac{S}{N_v} = \frac{\frac{kT}{\sqrt{2}\,q}\,\frac{J_{pmax}}{J_s}}{(4\,kT\,R_L\,\Delta f)1/2} \quad . \tag{3.122}$$

For a signal to noise ratio equal to unity

$$J_{pmax} = \left(\frac{8\,q^2\,R_L\,\Delta_f}{kT}\right)^{1/2} J_S. \tag{3.123}$$

Considering typical values of $J_s = 10^{-8}$ amps/cm^2 at T = 100°K and $R_L = 10^4$ ohms yields $J_{pmas} = 1.2 \times 10^{-14}$ $\frac{amps}{cm^2}$. Considering the parameters from the above example, the minimum detectable photon flux is approximately 7.5×10^4 $\frac{photons}{cm^2\text{-sec}}$ or for 1 ev photon an incident power density of 1.2×10^{-13} $\frac{watts}{cm^2}$. A typical area of 10^{-4} cm^2 results in a noise equivalent power (NEP) of 1.2×10^{-17} watts, or a specific detectivity of 1.2×10^{-15} cm Hz$^{1/2}$/watt. This is approximately three orders of magnitude below the detectivity for an ideal detector at 1 micron in Figure 3.28. One can do much better in approaching an ideal detector performance near 10 microns where the background flux density peaks. As shown in Figure 3.29 photovoltaic devices do approach background limited performance as opposed to our example where detector noise sources limit the performance.

Specific results for infrared detectors indicate that background limited performance can be obtained however the technology is far from routine. In the range from 1-5 microns InAs and InSb have been successful as shown in Figure 3.29. In fact InAs arrays have been realized with an element density approaching 200 per inch with rather impressive specific detectivity [100],[101]. The devices are p-on-n where the

larger electron mobility is utilized in the p region contribution to photocurrent. The devices are typically 5 x 7 mils with reverse breakdown voltages in excess of 25 volts. Typical reverse leakage current at 100°K is near 10^{-8} amps/cm^2. The dynamic resistance at zero bias and 300°K is approximately 270 ohms. At 300°K the leakage current is the order of 1 amp/cm^2. Thus the need for cooling becomes apparent since the detectivity depends upon the leakage current density.

The responsivity and specific detectivity for typical InAs photovoltaic devices at 100°K and 300°K are shown in Figure 3.42 and 3.43 respectively. At the peak near 3 microns the responsivity is 10^7 $\frac{\text{volts}}{\text{watts}}$ and the specific detectivity 1.5 x 10^{12} cm Hz$^{1/2}$/watt at a 500 Hz chopping frequency for the incident flux. For these particular devices a large 1/f noise contribution up to 10^3 Hz was observed. Therefore one must be careful in the specific detectivity. The near ideal performance at 100°K somewhat negates this caution.

In the 8-14 micron range lead-tin-telluride has been a primary consideration for photovoltaic devices. Diffused junction devices of n on p with junction depths from 2-4 microns have been reported [102]. Excessive leakage current appears to limit their performance where generation- recombination predominates. Current densities the order of 10 amps/cm^2 and 1/2 volt bias with the diode at 77°K appear to be typical. As a result the specific detectivity is near 10^9 cm Hz$^{1/2}$/watt with an external quantum efficiency of approximately 0.1 at 10.6 microns. One of the unique features of the ternary systems is the change in bandgap with the alloy ratios. For Pb-Sn-Te the peak in response shifts from 8 microns to 9.4 microns with a change from 15% to 17% tin.

Another example of a ternary system receiving considerable attention is mercury-cadmium-telluride (Hg-Cd-Te) [103]. Theoretically, the band gap can be varied from 1.6 ev to 0 ev. However, it is doubtful that the full range will prove feasible. The present interest is in short response

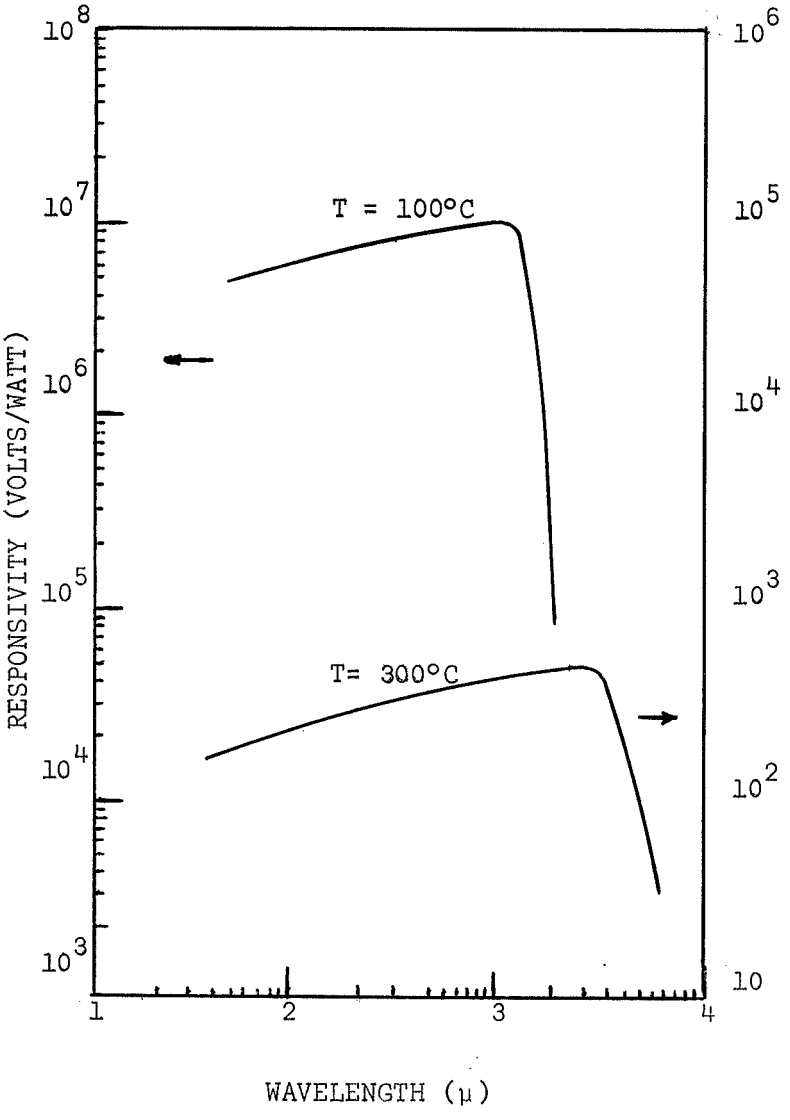


Figure 3.42 Spectral response for InAs diode. [101]

time and high performance at room temperature. Specific detectivities at the peak in responsivity near 1.5 microns are 10^{10} cm-Hz$^{1/2}$/watt at 295°K and only increases to about 1.2 x 10^{10} cm-Hz$^{1/2}$/watt at 77°K. Open circuit responsivities exceed 10^3 $\frac{\text{volts}}{\text{watt}}$ with detector resistance on the order of tens of kilohms. The response times were the order of micro-seconds. The performance at room temperature can prove to be a significant asset in many applications where cooling is not feasible.

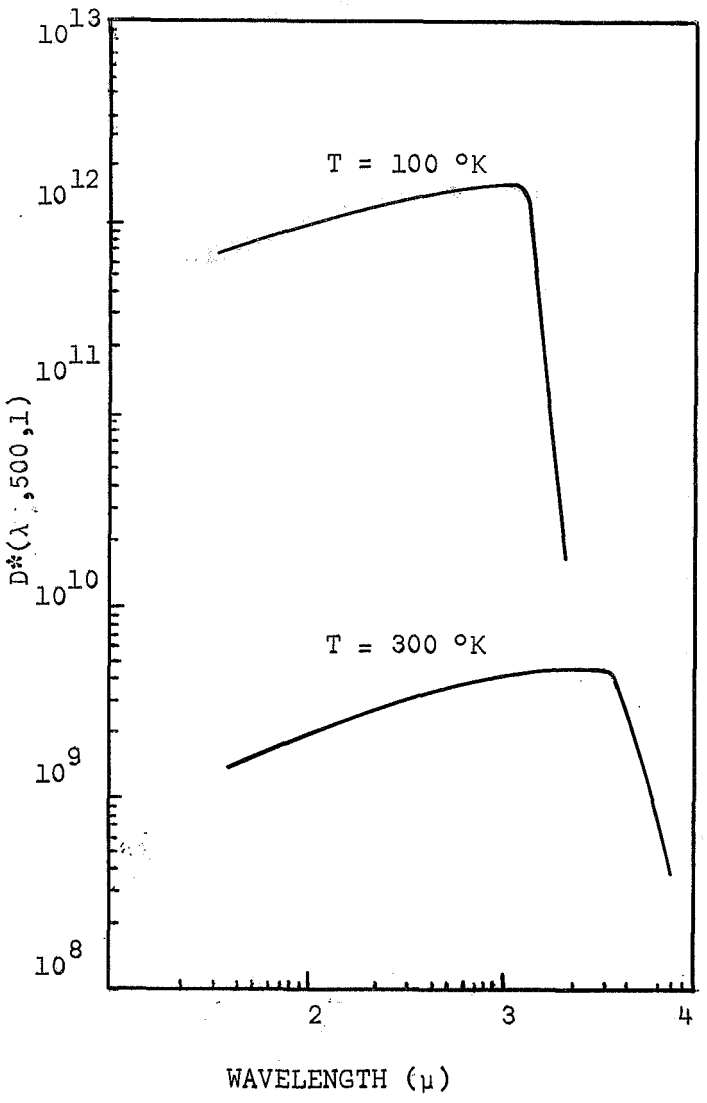


Figure 3.43 D* versus wavelength for InAs diode [101]

3.2.3.4 Summary of Optical Transduction

Optical transduction has been influenced by contemporary solid state as much as any other transduction principle. Photomultiplier and photoconductors predate depletion layer devices however much of the present understanding of photoconductors such as CdS was a spin-off from the intense efforts of the '50's and 60's.

The major advantages of photoemission and photoconductor are 1) internal gain approaching 10^6 and 2) the variety of spectral characteristics available. The major differences between these principles is the response time. To obtain the photoconductive gain there must be a sacrifice in response time. For photoconductors response times in excess of 10^{-3} seconds are typical. Photomultipliers response times are near 10^{-9} seconds or frequency response in the GHz range.

The frequency response of junction sensors is high being limited by transit time and diffusion time effects. Typically the response extends to several GHz. They can be tailored somewhat to give maximum response at various wavelengths by the use of compound semiconductors or by various metal-semiconductor (M-S) combinations. Avalanche multiplication can in principle be used with both p-n junctions and M-S devices to increase the sensitivity by a factor of 10-50. Higher multiplications decrease frequency response and noise increases faster than the signal thus the detector SNR decreases with multiplication. Junction devices are ideally suited for large sensor arrays which are becoming so important. On the negative side the response of all devices is wavelength dependent. Also, no method is known for providing large internal gain as in the photomultiplier.

P-n junctions, metal-semiconductor junctions, and heterojunctions appear to be the only useful photojunctions. Junction FET's and MIS devices have been studied but the response is equivalent to a p-n junction combined with an FET transistor. There's no advantage over the separate combination of these devices. Phototransistors are equivalent to a photodiode combined with a transistor to achieve gain. With silicon the combination can be made almost as easily as a p-n junction and there is some advantage. With other materials its probably not worth the effort to build phototransistors just to amplify the signal.

To summarize the impact of solid state technology on optical transduction the following assessments of photojunctions in the infrared, ultraviolet and visible are included. The p-n junction and metal-semiconductor sensors are most useful in the infrared region. The long wavelength limit is determined for p-n junctions by the semiconductor band gap and for metal-semiconductor devices by the metal-semiconductor work function. InAs, InSb, and In(As,Sb) are the longest wavelength junction devices studied. InSb goes out to about 6 μ. These diodes have to be cooled to liquid N_2 to obtain useful diode properties. Internal quantum efficiencies close to unity are readily achieved at energies larger than the bandgap. Surface reflectance is an important factor limiting the external quantum efficiency especially at short wavelengths. Arrays of InAs detectors have been fabricated.

Useful M-S infrared detectors with external quantum efficiencies larger than 1% have been demonstrated with Au-Si devices out to about 3μ and with Au-Ge out to about 10 μ. The metal-semiconductor infrared detector appears to offer many possibilities in the infrared especially if one can use a low work function material with silicon. Large detector arrays

in the infrared would then be feasible. Arrays with materials other than silicon are going to be slow in coming to practical use.

With special care junctions can be used in the ultraviolet region. The decreased response of p-n junctions in the ultraviolet is due to surface recombination reducing the internal efficiency and due to increased surface reflection. External efficiencies of 0.1 or larger have been demonstrated for Si p-n junctions from 7 to 20 eV. Above about 15 eV the efficiency increases because of more than one electron-hole pair created by each incident phonon. With proper attention to surface passivation to minimize recombination and perhaps antireflection coatings to minimize reflection, p-n junctions should be useful in the ultraviolet. Because of the requirement of low surface recombination, silicon is the obvious choice.

It has also been demonstrated that Cr-GaAs metal semiconductor devices are useful over the range 1.6 eV to 5 eV with a quantum efficiency of 0.1 to 0.2.

In the visible region Ge, Si, and GaAs appear to be the leading materials. Ideally the sensor material should have a bandgap close to the lower end of the visible region say 1.7 eV or so. From band gap energy alone GaAs is the best choice. However this small advantage is offset by the more difficult technology of GaAs. Silicon is the dominant visible p-n junction sensor. Large arrays of these are also feasible. The response of silicon is within about a factor of 2 over the range 0.4 μ to 1μ, covering the visible and near infrared.

Heterojunctions are interesting because they can be made to respond to a band of energies or wavelengths and give a fairly constant

response over the band. The response can be more uniform than for a normal - p-n junction and should have a quantum efficiency close to unity.

Metal semiconductor devices can be used as visible detectors but have no significant advantage over silicon.

REFERENCES

1) H. N. Norton, Handbook of Transducers for Electrical Measuring Systems, Prentice-Hall, Englewood Cliffs, New Jersey (1969).

2) H. K. P. Neubert - Strain Gauges, MacMillan-London- (1967).

3) F. F. Hines & L. J. Weymouth, "Practical Aspects of Temperature Effects on Resistance Strain Gages", in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York, (1962) p. 148-168.

4) I. G. Scott, "Some Temperature Effects on Resistance Strain Gages" in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York (1962) p. 73-92.

5) P. K. Stein, "Adhesives: How they Determine and Limit Strain Gage Performance", in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York (1962) p. 45-72.

6) R. L. Bloss, "Characteristics of Resistance Strain Gages" in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York, (1962) p. 123-142.

7) T. L. Ward - "Recent Developments in the Application of Semiconductor Strain Gages to Physical and Mechanical Measurements", 20th Annual ISA Conference, ISA Preprint 17; 11-4-65, Oct. 4, 1965, 20 pages.

8) J. Dorsey, "Temperature Compensation for Semiconductor Strain Gages". Instrument Society of America 1962 Proceedings, ISA Preprint 37.2.62.1 (1962), 9 pages.

9) Handbook of Chemistry and Physics 48th Edition, Chemical Rubber Company, Cleveland, Ohio (1968).

10) E. D. Padgett, "Silicon Piezoresistive Devices" in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York (1962) pp. 1-20.

11) C. Rohrback and N. Czaika, "On the Mechanism, Properties and Application of Semiconductor Transducers" translated from Materialpruafung, vol. 4, (1962) (translation 38 pages)

12) W. P. Mason, J. J. Forst, and L. M. Tornillo, "Semiconductor Stain Transducers", in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages Academic Press, New York, (1962),pp. 109-120.

13) A. D. Kurtz, "Adjusting Crystal Characteristics to Minimize Temperature Dependence", in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York, (1962), pp. 259-272.

14) E. Franks, "Strain Indicator for Semiconducto Gages" in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York, (1962) p. 283-306.

15) J. C. Sanchez and W. V. Wright, "Recent Developments in Flexible Silicon Strain Gages", in M. Dean and R. D. Douglas, Semiconductor and Conventional Strain Gages, Academic Press, New York, (1962) pp. 330-346.

16) H. H. Hall, J. Bardeen, and G. L. Pearson, "The Effects of Pressure and Temperature on the Resistance of p-n junctions in Germanium", Physical Review, vol. 84, Oct. 1951, pp. 129-132.

17) W. Rindner, and I. Braun, "Resistance of Elastically Deformed Shallow p-n Junctions II", J. Appl. Phys., vol. 34, July, 1963, pp. 1958-1970.

18) M. E. Sikorski, P. Andreatch, A. Grieco, and H. Christensen, "Transistor Microphone" Rev. Sci. Instr., vol. 33, Oct. 1962, pp. 1130-1131.

19) J. J. Wortman, J. R. Hauser, and R. M. Burger, "Effect of Mechanical Stress on p-n Junctions Device Characteristics", J. Appl. Phys. vol. 35, July 1964, pp. 2122-2131.

20) J. J. Wortman and L. K. Monteith, "Semiconductor Mechanical Sensors", IEEE Transactions on Electron Devices, vol. 16, Oct. 1969, pp. 855-860.

21) W. Rindner, G. Doering and R. Worson, "Structural and Operational Characteristics of Piezotransistors and Applied Devices", Solid State Electronics, vol. 8, March 1965, pp. 227-240.

22) L. K. Monteith and J. J. Wortman, "Response of p-n Junction to Time Varying Mechanical Strains", Submitted for Publication.

23) B. Shneider and M. J. O. Strutt, "Theory and Experiments on Shot Noise in Silicon p-n Junction Diodes and Transistors", Proceeding of the IRE, vol. 47, April 1959, pp. 546-554.

24) A. U. MacRae, "1/f Noise From Vacuum Cleaved Silicon", Journal of Applied Physics, vol. 33 (1962), pp. 2570-2572.

25) L. K. Monteith, "Correlation of I-V Characteristic with Noise for Ion Drifted p-i-n Junction Particle Detectors", Rev. of Sci. Instru., vol. 35 (1964), pp. 388-392.

26) J. J. Wortman, Private Communication

27) W. P. Mason, "Semiconductor Devices as Pressure Transducers", Electronics, February 23, 1962, pp. 35-39.

28) T. R. Kiggins and A. G. Milnes, "A Solid State Tunnel Diode Strain Gage", ISA Preprint No. 43.3.63 (1963), 9 pages.

29) M. E. Sikorski, "Use of p-n Junction Semiconductor Transducers in Pressure and Strain Measurements" in W. P. Mason, Physical Acoustics, vol. I Part B, Academic Press, New York, N. Y. (1964).

30) R. M. Moore, C. J. Busanovick, "The Heterode Strain Sensor; An Evaporated Heterojunction Device", IEEE Transactions on Electron Devices", vol. 16, (1969), pp.850-855.

31) J. Conragan, R. S. Muller, "Piezoelectric Field-Effect Transistor Strain Transducers" The 1970 Solid State SEnsor Symposium, Minneapolis, Minn. June 1970.

32) W. M. Bullis, F. H. Brewer, C. D. Kolstad and L. J. Swartzenduber, "Temperature Coefficient of Resistivity of Silicon and Germanium Near Room Temperature", Solid State Electronics vol. 11, (1968) pp. 639-646.

33) R. Badertscher and H. Diamond, "High Gage Factor Thin Film Strain Gages and Associated Transducer Design", The 1968 Solid State Sensors Symposium, Sept. 12-13, Minneapolis, Minn. IEEE Catalog #68C-55-ED, (1968) pp. 11-12.

34) O. Watanabe, K. Hatakuyama, and T. Shioda, "Germanium-Film Strain Gage with Self-Compensation of Temperature", Electrical Engineering in Japan, vol. 87, (1967) pp. 101-112.

35) W. P. Mason, Physical Acoustics, vol. I, Part A, Academic Press, New York (1964).

36) N. F. Foster, "The Application of Piezoelectric Semiconductors to the Fabrication of High-Frequency Ultrasonic Transducers", Transactions of the Metallurgical Society of AI ME, vol. 230, (1964) pp. 1503-1507.

37) H. D. P. Neubert, Instrument Transducers, Oxford University Press, London, (1963).

38) L. B. Wilner, "Sensitivity Comparison of Transduction Mechanisms", Rev. of Scientific Instruments, vol. 36, (1965) pp. 693-695.

39) C. S. Brown, R. C. Kell, R. Taylor, & L. A. Thomas, "Piezo-Electric Materials" The Institution of Electrical Engineers, Proc. V PtB (Electronics and Communications Engr.) Jan. 1962, pp. 99-144.

40) G. Arlt, and P. Quadflug, "Piezoelectricity in III-V Compounds with a Phenomenalogical Analysis of the Piezoelectric Effect", Phys. Stat. Sol. vol. 25, (1968), pp. 323-330.

41) F. F. Stucki, "A High-Speed Ferrimagnetic Microtransducer", Sensors and Measuring Systems, Session 15 Paper 15.A, June 21, 1966 - International Federation.

42) J. E. Besancon, J. L. Champstier, Y. Leclanche, J. Vedel, and J. P. Plantevin, "Ferromagnetic Transducers", Conference on Megagauss Magnetic Field Generation by Explosives and Related Experiments, Trascati Italy, September 21-23, 1965, Proceedings.

43) E. F. Nitka and T. K. Ishii, "Microwave Ferrite Acceleration Sensors", IEEE Transactions on Electron Devices, vol. ED-16, (1969) pp. 845-849.

44) M. Nalecz and Z. L. Warsza, "Hall Effect Transducers for Measurement of Mechanical Displacements", Solid State Electronics, vol. 9, (1966), pp. 485-495.

45) R. S. Davidson and R. D. Gourlay, "Applying the Hall Effect to Angular Transducers", Solid State Electronics, vol. 9, (1966), pp. 471-484.

46) N. Nalecz and H. Ziomechki, "A Hall Effect Accelerometer", Journal of Franklin Institute, vol. 276, July 1963, pp. 14-25.

47) A. S. Platon, H. A. Ricci, "Wind Tunnel Tests of Solid State Strain Gages", Ballistic Research Laboratory Memorandum Report #1928 , July, 1968 (N68-35247)

48) J. J. Wortman, "Semiconductor Piezojunction Transducers" NASA Contractor Report #NASA, CR-1089 June, 1968.

49) R. Bowman, "Research and Development of a New Technique for Sensing Pressure", NASA Contractor Report #NASA CR 96717 (N68-34734).

50) G. L. Vick, "A Feasibility Study of a Solid State Acceleration Sensing Technique", Technical Report Nr. FDL-TDR-64-55 AF Contract No. AF 33(657)-10671 Teledyne Systems Corporation, Hawthorne, Calif. (AD 610548).

51) E. J. Evans etal, "Improved Temperature Compensation for Integrated Piezoresistive Transducers", Picatinny Arsenal, Dover, New Jersey (AD 66621).

52) C. Gross, "The Use of Irradiated Semiconductor Strain Gages in the Precise Measurement of Low-Level Strains", Advances in Instrumentation vol. 24, (1968) p. 571-576.

53) RCA Photomultiplier and Image Tubes, RCA Publication PIT-700 (1965).

54) A. H. Sommer, Photoemissive Materials, John Wiley, Inc. New York (1968) - p. 188.

55) B. Kazan and M. Knoll, Electronic Image Storage, Academic Press, New York (1968).

56) A. H. Sommer, Photoemissive Materials, John Wiley Inc. New York, N. Y. (1968) p. 146.

57) S. J. Czyzak, D. C. Reynolds etal, "On the Properties of Single Cubic Zinc Sulfide Crystals", Optical Society of America, Journal vol. 44, Nov. 1954, pp. 864-867.

58) N. A. Vlasenko, "Study of the Absorption Spectrum of Zinc Sulfide", Optics and Spectroscopy, vol. 7, Oct. 1959, pp. 320-324.

59) F. A. Kroger, "Some Optical and Electrical Measurements on Blue Fluorescent ZnS-Cl Single Crystals", Physica, vol. 22 (1956), pp. 637-643.

60) I. Broser and H. J. Schulz, "A Comparative Study of Infrared Luminescence and Some Other Optical and Electrical Properties of ZnSiCu Single Crystals", Electro chemical Society Journal, vol. 108, June 1961, pp. 545-548.

61) W. Lehmann, "Mean Lifetime of Free Electrons in Electroluminescent Zinc Sulfide", American Physical Society, Bulletin, vol. 3, (1958), p. 45.

62) H. R. Philipp and E. A. Taft, "Kramers-Kronig Analysis of Reflectance Data for Diamond", Physical Review, vol. 136, Nov. 1964, p. A1445-1448.

63) F. A. Raal, "A New Absorption Band in Diamond and its Likely Cause", Physical Society Proceedings, vol. 71, May 1958, pp.846-847.

64) H. B. Dyer and P. Ferdinando, "The OPTICAL Absorption of Electron-Irradiated Semiconducting Daimond", British Journal of Applied Physics, vol. 17, March 1966, pp. 419-420.

65) D. C. Reynolds, "Magneto-Optical Properties of Semiconducting Solids", I.S.A. Transactions Vol. 4, Jan. 1965, pp. 34-45.

66) C. Johnson, etal, "Photo Effects and Related Properties of Semi-conducting Diamonds", J. of Physics and Chemistry of Solids, vol. 25, Aug. 1964, pp. 827-836.

67) A. Halpern and J. Levinson, "Photoconductivity of UV Excited Diamonds", Int. Conference on Semiconductor Physics, Proceedings, 7th Paris 1964. Edited by M. Hulin, Academic Press, New York, (1964).

68) R. H. Bube, Photoconductivity of Solids, John Wiley, New York, 1960.

69) R. H. Bube, "Photoconductors" in Photoelectronic Materials and Devices edited by S. Larach, Van Nostrand, New York, 1965, pp. 100-139.

70) R. H. Bube, "Photoconductivity" in Physics and Chemistry of II-VI Compounds, edited by Aven and Prener, John Wiley, New York (1967), pp. 659-705.

71) R. H. Bube and L. A. Barton, "Achievement of Maximum Photoconductivity Performance in Cadmium Sulfide Crystals", RCA Review, Vol. 20, (1959) pp. 564-598.

72) A. Rose,"An Outline of Some Photoconductive Processes", RCA Review, vol. 12, (1951) pp. 362-414.

73) A. Rose and M. A. Lampert, "Gain-Bandwidth Product for Photoconductors", RCA Review, vol. 20 (1959) pp. 57-68.

74) A. Rose, "Performance of Photoconductors", in Photoconductivity Conference, edited by R. G. Breckenridge, B. R. Russell and E. E. Hahn, John Wiley, New York, (1955) p. 3-48.

75) A. Rose, "Noise Currents" in Photoelectronic Materials and Devices edited by S. Larach, Van Nostrand, New York (1965), pp. 222-238.

76) C. Shulman, R. W. Smith and A. Rose, "Some Noise Observations on CdS Crystal Photoconductors", Physical Review, vol. 92, (1953) p. 857.

76A) P. W. Kruse, L. D. McGlanchlin and R. B. McQuistan, Elements of Infrared Technology", John Wiley, New York (1962).

77) C. T. Sherlman, "Measurement of Shot Noise in CdS Crystals", Physical Review, vol. 98, (1955), p. 384-386.

78) R. L. Petritz, "Fundamentals of Infrared Detectors, "Proceedings of the IRE, vol. 47, (1959) pp. 1458-1967.

79) E. H. Putley, "Solid State Devices for Infrared Detection", Journal of Scientific Instruments, vol. 43, (1966), pp. 857-868.

80) R. F. Potter, and W. L. Eisenmean, "Infrared Photodetectors: A Review of Operational Detectors", Applied Optics, vol.1, (1962), pp. 567-574.

81) R. C. Jones, "Phenomenological Description of the Response and Detecting Ability of Radiation Detectors, "Proceedings of the IRE, vol 47, (1959) pp. 1495-1502.

82). E. L. Stelzer, J. L. Schmit, and O. N. Tufte, "Mercury Cadmium Telluride as an Infrared Detector Material", IEEE Transactions on Electron Devices, vol. ED-16, (1969) pp. 880-884.

83) A. L. Morse, W. F. Crevier, G. L. Weissler, and D. B. Medved, "Far Ultraviolet Response of Silicon p-n Junction Photodiodes", NASA Contract NsG-178-61 (N66-17303) 15 June 1965.

84) R. D. Baertsch and J. R. Richardson, "An Ag-Ga-As Schottky-Barrier Ultraviolet Detector, Journal of Applied Physics, vol. 40, (1969), pp. 229-235.

85) J. A. Tuzzolino, "Silicon Photodiode Vacuum Ultraviolet Detector", Review of Scientific Instruments, vol 35, (1964), pp. 1332-1335.

86) H. R. Philipp and E. A. Taft, "Optical Constants of Silicon in the Range 1 to 10 ev", Physical Review, vol 120, Oct. 1960, pp. 37-38.

87) R. D. Baertsch and J. R. Richardson, "Ag-Ga-As Schottky-Barrier Ultraviolet Detector", Journal of Applied Physics, vol. 40, (1969) pp.229-235.

88) P. Volid, D. M. Perkins, S. G. Ellis, R. R. Addiss, W. Hui and G. Noel, "GaAs Thin-Film Solar Cells", IEEE Transactions on Electron Devices, vol ED-14 (1967), pp. 26-30.

89) D. A. Crossby, R. G. Gill, A . Spielman and M. Wolf, "Thin-Film, GaAs Photovoltaic Solar Energy Cells" NASA Contract NAS 3-8510, Radio Corporation of America, Princeton, N. J. (N 68-13216).

90) J. J. Wysocki and P. Rappaport, "Effects of Temperature on Photovoltaic Solar Energy Conversion", Journal of Applied Physics, vol. 31 (1960) pp. 571-578.

91) M. V. Schneider, "Schottky Barrier Photodiodes with Antireflection Coating" Bell System Technical Journal, vol. 45, November 1966, pp. 1611-1638.

92) F. A. Shirland, "The History, Design, Fabrication and Performance of CdS Thin Film Solar Cells, Advanced Energy Conversion, vol. 6, (1966) pp. 201-221.

93) P. Rappaport, "The History and State-of-the Art of the Solar Cell", RCA Laboratories, Princeton, New Jersey(N68-28739). pp. 803-828.

94) A. K. Sreedhar, B. L. Sharma, and R. K. Purohit, "Efficiency Calculations of Heterojunction Solar Energy Converters", IEEE Transactions on Electron Devices, vol ED-16 (1969) pp. 309-312.

95) J. J. Wysocki, P. Rappaport, E. Davidson, R. Hand, and J. J. Loferski, "Lithium Doped Radiation Resistant Silicon Solar Cells", Applied Physics Letters, vol. 9, (1966) p. 44-46.

96) P. Rappaport, "The Photovoltaic Effect and Its Utilization", RCA Review, vol. 20, Sept. 1959, pp 373-397.

97) D. A. Iles, "The Present Status of Silicon Solar Cells,"IRE Transactions on Military Electronics, vol. MIL-6, January 1962, pp. 5-14.

98) J. J. Loferski and J. J. Wysocki," Spectral Response of Photo-voltaic Cells," RCA Review, vol. 22, March 1961, pp. 38-55.

99) I. Wunderman, "Photon Coupled Pairs-Characteristics and Applications of Devices Which Use a Light Beam as the Signal Carrier", Air Force Contract #AF33)657)-9772 to HP Associates, Palo Alto, Calif. November, 1963.

100) W. E. Davern, etal, "Solid State Image Sensor Research",Report prepared under NASA Contract No. NAS 12-131 by General Electric Company, Electronics Laboratory, Syracuse, New York, Phase II Technical Report, March 1969.

101) W. E. Daver etal, "Solid State Image Sensor Research", NASA Contract NAS 12-13) by General Electric Company, Electronics Laboratory, Syracuse, New York, Fianl Technical Report, March 1969.

102) W. Ross, R. Lee, and R. J. Eddington, "Preparation and Properties of Lead-Tin-Telluride Photodiodes" Solid State Electronics, vol. 13, (1970) pp. 75-81.

103) A. N. Kohn and J. J. Schlickman, "1-2 Micron (Hg,Cd) Te Photodetectors", IEEE Transactions on Electron Devices, vol. ED-16 (1969) pp. 885-890.

4.0 Conclusions and Recommendations

It is difficult to draw specific conclusions for such a voluminous treatment of solid state sensors and transducers. Oddly enough there is a tendency to recommend further investigation into the areas we did not treat in the preceeding sections. In part this inclination is tempered by the fact that many of the omitted topics are areas of contemporary research with a degree of uncertainty concerning the applicability to sensor development.

Our conclusions and recommendations are generally divided into catagories which essentially summarize the emphasis in Sections 1 and 2. For example stress sensors are considered separately for bulk effects and for junction devices. Finally a general conclusion and recommendation concerning the overall development of solid state sensors is included.

4.1 Stress Sensors - Bulk Effects

There is a good understanding of the basic mechanisms which determine the effects of mechanical stress on bulk materials. Such phenomena have been used extensively in mechanical sensors of all types. Generally, the piezoresistance effect in semiconductors has a high stress sensitivity; but has accompanying large thermal variations in its properties. Piezoresistance in metals has low sensitivity and low thermal variations. These distinquishing features seem to prevail in all materials systems studied to date. The piezoelectric effect offers very large stress sensitivities, but its application is limited to dynamic measurement applications. These three bulk mechanisms should be the first considered for any present or anticipated mechanical sensor application.

The possibility of finding a new material system or a new bulk material phenomena which will drastically improve upon the properties of presently available sensors does not appear promising. However, there are two areas where this pessimism might not be warranted. The transition metal compounds have been found to have electrical properties which are highly stress dependent. Although little work has been done in this area, these materials would seem to offer extremely large sensitivities over some ranges of stress. They also appear to have applications as "switches" which can be turned on and off with applied stress or pressure. The technology and theoretical understanding of the properties of this class of materials has been advanced to the point where their use as sensors could be exploited.

Another class of materials which should be considered as stress sensors is the ferrimagnetic materials, better known as ferrites. Basically,

these materials are semiconductors or semi-insulators. One can expect that the inherent properties of these materials, such as electrical conductivity or remanant magnetization, would be as sensitive to stress as a non-magnetic semiconductor. Little has been done to verify this statement. Ferrimagnetic materials can be used as memory devices, and sensors using this mode of operation would be compatible with high-speed data processing techniques. This type of sensor could be useful in areas where large amounts of data need to be processed, retrieved and monitored, such as in hospitals and large manufacturing plants. Practical sensors operating in this mode and using ferrimagnetic materials have been realized.

The use of thin film stress sensors should not be overlooked. The general tendency of users seems to be to over look thin films because of previously encountered problems such as reproducibility and repeatability of thin film properties. However thin film technology has advanced considerably in the last three to four years, and many of these problems are being overcome. Thin films offer the advantages of small size and weight, economy, and ruggedness. Using thin film technology, sensors can be passivated and protected from extreme enviornments, where factors such as humidity, unwanted chemicals and optical illumination can be devastating to the basic sensor performance.

One other area which has the ability to improve stress sensor performance is the use of controlled impurities. The feasibility of such a technique has been demonstrated with the use of electron irradiation to improve silicon strain gage thermal variations. This technique could be extended to the more standard chemical diffusion techniques, and many impurities appear to offer properties similar to electron irradiation produced defects.

4.2 Stress Sensors - Junction Devices

The phenomena underlying stress effects in junction devices appears to be fairly well understood. Metal-semiconductor devices have not been extensively studied; however, there is no reason to believe that they will be much better than p-n junctions as stress sensors. In terms of increased stress sensitivity the heterojunction sensor appears to warrant further research. The question of signal-to-noise properties of junction sensors needs to be investigated. Junction devices have an advantage over piezoresistance in sensitivity but whether or not this can be used depends on the relative noise properties of the two devices. Little work has been done on this problem.

Considerable past research work has been done on stress sensors using transistors of either the junction or field effect type. There is no evidence either theoretical or experimental to indicate that these devices are more stress sensitive than a seperate stress sensor followed by an active device to achieve gain. In practically all cases there appears to be definite advantages to the use of separate sensor and amplification devices.

Suggested device stress sensor research areas are the study of heterojunction devices for devices with increased stress sensitivity and the study of fundamental noise properties for junction devices to see if the increased sensitivity can be used to measure low stress levels.

4.3 Optical Sensors - Bulk Effects

Photoconductivity is the principle bulk material phenomena used for optical sensors. A wide variety of intrinsic and extrinsic photoconductors are available with good spectral response from the near ultraviolet to the far infrared. However, many of the materials used as bulk infrared detectors are difficult to prepare controllably and reproducibly. The influence of impurities on the spectral response is not completely understood nor well investigated.

It appears that thin film techniques, and especially cathodic sputtering, offers some promise as a means for the preparation of many good photoconductor materials. Specifically, co-sputtering of the II-VI alloy compounds, such as HgCdTe and PbSnTe, could be a convenient and simple technique for the preparation of these otherwise-difficult-to-prepare materials.

The use of sandwich structures of photoconductors and ferroelectric insulators have recently(Oct. 1970) been shown to exhibit optical memory effects which could be very important in the field of optical image processing. This relatively unexplored area should be thoroughly investigated.

Optical effects in ferrites could also be used to present optical memory effects. Very little work has been uncovered to indicate whether ferrites are optically sensitive. However, optical excitation in these insulators and semi insulators should be a fundamental process.

4.4 Optical Sensors - Junction Devices

This is an area in which solid state sensors are firmly established. The major problem areas for solid state sensors are in the infrared region of the spectrum and in the development of a sensor to compete with the internal gain of the photomultiplier. In the area of optical sensors the following research areas are proposed: (1) Study the use of metal-semiconductor detectors for the infrared spectral region. (2) Continue the development of p-n junction technology of narrow band semiconductors for the infrared spectral region. (3) Develop p-n junction detectors which are optimized for the ultraviolet spectral region. (4) Develop heterojunction detectors for use as band sensitive detectors and as increased efficiency solar cells.

4.5 Temperature Sensors - Bulk Effects

Most bulk properties of solid state materials vary with temperature. However, because of simplicity and economy, the electrical resistance is the principle bulk phenomena used in temperature sensing. Both metals and semiconductors have been used in this respect, and such sensors are available which cover the range from cryogenic temperatures to several hundred degrees centigrade. Many temperature standards are simply calibrated metal or semiconductor resistors. Perfectly linear response is achievable only in metals over a limited temperature range.

One area which is attractive for research in temperature sensors is that of the transition metal compounds. These materials offer the possibility of obtaining thermal switches over a wide range of temperatures. This could be achieved,for example, by controlling the composition of the transition metal oxides, such as vanadium oxide.

4.6 Temperature Sensors - Junction Devices

Junction devices appear adequate for most sensor applications from 0°K to a few hundred degrees centigrade. The major problem area appears to be high temperature sensors. Junctions devices are not likely to solve this problem because of the difficulty of achieving p-n junctions in wide bandgap semiconductors. The use of p-n junction sensors at low temperature may warrant further study.

4.7 Magnetic Sensors - Bulk Effects

Magnetoresistance and the Hall effect are bulk phenomena used as magnetic field sensors. The use of ferromagnetic rods as "field concentrators" to enhance the sensitivity of Hall generators has been an important advancement in the measurement of very low magnetic fields intensities. Such instruments are theoretically capable of resolving magnetic flux densities of less than 10^{-5} gauss. It does not seem likely that one would be able to improve on this limit. Superconductivity has not been exploited to any extent in this measurement area, and presently does not seem to offer any unique advantages. However, as progress is made, superconductivity could prove to be a useful phenomena for magnetic field measurements.

There does appear to be the possibility that thin film magnetic field sensors could be made which would offer ease of fabrication and economy. Also conventional ferromagnetic magnetometers could be constructed in thin film form. This would reduce the size and expense of these instruments and would yield a process compatible with integrated circuit processing so that the complete magnetic and much of the electronics could be fabricated in chip form.

4.8 Magnetic Field Sensors - Junction Devices

In solid state magnetic sensors the Hall effect reigns supreme. Solid state devices can be used to amplify the signal but no effect of magnetic field on junction devices has been shown to be as sensitive as the Hall effect.

Although superconductor devices have not been covered in this report they have been found to be very sensitive to magnetic fields. They require cooling to very low temperatures which is most applications is a big disadvantage.

Integrated sensors using Hall sensors and integrated circuits for amplifying and signal conditioning should be investigated.

4.9 Chemical Sensors

This is an area of sensors which is becoming more important with the emphasis on environmental control. Very little has been said in this report on this type of sensor. It appears that there may be some very important applications of solid state materials and devices in this area. Thin film devices and MIS type devices are quite sensitive to rather small changes in surface properties. Gases absorbed on the surface and coupled to the electronic properties of the films may constitute a means of sensing the chemical environment.

Proposed research areas are (1) explore the use of thin films as chemical sensors. (2) Explore the use of MIS type field effect transistor as chemical sensors.

4.10 Integrated Sensors

Because of the nature of the properties of solid state materials and devices, solid state sensors are in general sensitive to more than one external influence. Temperature is a very troublesome parameter to which all solid state sensors are sensitive to some degree. In sensors not designed for temperature measurements this is a very undesirable effect. The area of solid state provides techniques not only for sensors but for conditioning the environment of the sensor as well. For example, temperature stabilized electronic integrated circuit blocks have been demonstrated and fabricated which keep the temperature of some component or circuit at a constant value independent of the external environment. This approach to temperature insensitive solid state sensors should not be over looked. Active research should be undertaken in the development of temperature stabilized sensors. It may be feasible to develop a temperature stabilized integrated circuit block which can be used with a variety of solid state sensors.

Solid state technology also provides techniques for conditioning the signal output from sensors. High gain integrated circuits and even analog-to-digital converters can be achieved in very small size integrated circuits. In many cases these can be packaged in the sensor housing along with the solid state sensor. When the sensor is located some distance from the observation point there are in most cases very definite advantages to integrating signal conditioning circuits as close as possible to the sensor. Most of the research on using devices such as bipolar transistors and MOS transistors as sensors are attempts to achieve signal amplification or conditioning at the sensor. In practically all cases it is more

advantageous to separate the sensor and signal conditioning functions but to locate them physically as close to each other as possible.

Solid state sensor technology should not be looked at simply in terms of the sensor alone. Some of the most significant sensor improvements probably lie in the integrated solid state sensor approach whereby signal conditioning circuits are located at the sensor.

4.11 Sensor Research and Development

General conclusions to be drawn from our activities are difficult to quantify, however certain attributes of sensor research and development have become apparent. From NASA technical documentation obtained through computerized research facilities at North Carolina Science and Technology Research Center and from numerous manual searchers approximately 2000 documents were scanned. All of these documents were in one way or another related to the broad area of solid state sensors and transducers. However the overwhelming number were given only a cursory examination. Approximately 500 documents were ultimately examined in some detail. From this experience it became apparent that a number of novel ideas concerning solid state sensors had envolved during the past two decades however most of the work was related to extending, stabilizing or further characterizing the performance of a few concepts. For example, a wealth of literature has been devoted to the problem of temperature effects on the performance of silicon strain gages.

In addition to the extensive literature concerning the development of now traditional solid state sensors, there is a substantial effort with the objective of applying well developed technology to novel transducer concepts. A primary example is the effort to achieve a solid state vidicon utilizing silicon diode arrays.

The two examples given above for the primary activity in solid state sensor development are typical of activities with a high probability of success. If a primary objective for a particular sensor appears feasible by either extending the performance of a well established device

or utilizing a well developed technology, past experience indicates that the additional effort will be rewarding. There is a danger that is approach can be self defeating since fundamental limitations are always present although not well defined in many instances. Ultimate stress sensitivity for a piezoresistive strain gage is a primary example where little quantitative information concerning signal to noise performance is available.

To provide sensors beyond contemporary solid state technology will require one or all of the following: (1) new materials such as amorphous semiconductors, (2) new phenomena such as metal-semiconductor transitions or (3) new sensor conepts such as digital transducers. Each of these areas (new materials, new phenomena and new sensor concepts) is an extensive research area and will not progress very rapidly unless there is a large demand for a particular sensor uniquely associated with the research. Following this line of reasoning, the development of solid state sensors is likely to follow the development and understanding of new materials and technology for larger markets for semiconductor devices.