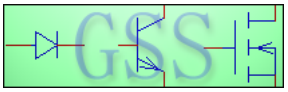
GSS User's Guide



Version 0.46.00

Copyright Notice of the GSS software

The GSS software, is covered by the following BSD (or MIT) type license:

Copyright Notice of the GSS User's Guide

The document itself, is covered by the GNU Free Documentation License.

Contents

Chapter 1 Introduction to GSS Software

1.1 Overview

GSS (General-purpose Semiconductor Simulator) is an **open source** 2D device simulator for the numerical simulation of semiconductor devices. Based on the well established methods of drift-diffusion (DD) and energy balance (EB), GSS can calculate the intrinsic physical variables of a semiconductor device (such as potential and carrier concentrations), as well as terminal currents and voltages. With a carefully designed device model and well calibrated parameters, GSS can be used to reliably predict the electrical characteristics of real semiconductor devices prior to manufacture.

1.1.1 History

The original author of the code, Gong Ding, has been developing GSS since 2004 and first published the code into the public domain in 2006. GSS continues to be actively developed and maintained by Gong Ding and Li Yisuo. In 2008, the **GeniEDA Corp** was founded to offer greater support to GSS users and to further develop the code as a serious alternative to commercially available products. At the same time, the **GENIUS** project was initiated to develop an open-sourced 3D parallel process and device simulator.

1.1.2 Why Chosen GSS

Here are some great reasons for choosing GSS over other available programs:

- GSS open source is free, but it is most definitely not a toy. At its present stage of development, GSS offers around 70 to 80% of the functionality to be found in the leading commercial products. Since the software structure is so well organized and incorporates a number of advanced technologies, GSS has the potential to become a world class 2D simulator.
- The source code is made available to user, thus rendering the calculation of numerical solutions transparent and open to detailed inspection. Furthermore, there are no limitations in modifying and redistributing the code except as defined by the BSD license which covers this work.
- The technical support for this product is based at source code level, which offers more possibilities when seeking solutions to different problems.
- The physical model, as well as the numerical arithmetic, is described in detail in this manual. It is strongly recommended and intended that serious users should read this material carefully.

1.1.3 Basic Numerical Arithmetic

Currently, GSS offers four solvers with different levels to meet the different demands of device simulation.

Level 1 DDM Traditional device such as diodes, bipolar transistors and long gate (> 1 µm) MOS transistors can be numerically analyzed by level one DD solver. For this solver, Poisson's equation and both the electron and hole current continuity equations are solved self-consistently by a full Newton's scheme.

Level 2 DDM For a power transistor or device, Joule heating can not be neglected. In order to take this effect into account, the level two DD solver considers an extra lattice temperature equation. At this level, GSS is self-consistently solving four equations.

Level 3 EBM GSS simulates the behavior of deep submicron devices such as advanced bipolar and CMOS transistors by solving the electron and hole energy balance equations self-consistently with the other device equations in a level three energy balance solver. Up to six equations are solved in a fully coupled model by Newton's scheme. Effects such as carrier heating and velocity overshoot are accounted for and their influence on device behavior can be analyzed.

Quantum DDM For simulation of deep submicron and nanometer MOS devices, the density gradient model (which based on the lowest moments of the Wigner Function) is integrated into GSS. For this model, three basic DD equations plus two quantum potential equations are solved consistently.

1.1.4 Dynamic Loadable Library for Physical Models

Through GSS's highly flexible interface, the user can add support for new materials or modify the default physical models to their own configuration. Instead of hard coding the physical models and parameters into one binary file, GSS loads information about materials from a separate dynamic loadable library (a shared object format for the Linux/Unix system). This mechanism has several advantages:

- Run time efficiency is superior when compared to reading in information through a script interpreter[1].
- Adding support for new materials is clearly delineated and separate from other areas of the code, thus making debugging infinitely easier.
- Automatic differentiation can be employed to reduce the workload. Please refer to the following paragraph.
- Intellectual property can be shared with others without exposing proprietary models or critical parameters by pre-compiling it into a dynamic library.

 Note:

Currently, GSS does not offer a way for adjusting default parameters from the input deck. In order to do this, it remains necessary to modify the material library source code and then recompile it. Since the source code is well organized and easy to read, this task is normally relatively easy. Alternatively, the user could design an interface for the material library itself.

1.1.5 Mesh Generation and Refinement

GSS employs the Triangle [1] mesh generator to model arbitrary device geometries and complex surface topographies in a simulation using an unstructured mesh of triangular elements. This initial mesh can then be further refined during the

[1] Silvaco ATLAS uses "C script" interpreter to support user defined model.

solution process based on potential or impurity concentration. Wherever the potential or impurity concentrations vary by more than a specified tolerance over a triangular element, it will be refined by sub-division into smaller triangles. This flexibility makes the modeling of complex devices and structures possible.

Electrodes can be placed anywhere in the device structure. Impurity distributions can be created by user-specified analytic functions, or generated by more accurate **Implant** and **Diffusion** processes (this feature has been supported since Version 0.46).

1.1.6 Automatically Differentiation

Jacobian Matrix Generated by AD

When employing Newtons' iteration method to solve the governing nonlinear equations arising from the drift-diffusion model, the first derivative of the vector equations (the Jacobian matrix) needs to be calculated. In the past, symbolic differentiation could be performed by hand. However, this process tends to be tedious and unsuitable (in the authors experience) for creating bug free code.

NO Derivatives

Since version 0.46, GSS employs automatic differentiation (AD) to relieve this burden. All code has been rewritten in the AD format. This enables advanced arithmetic to be integrated into the GSS code with affordable effort. Complex physical models can now be implemented into the material library with the help of the AD tool provided. No derivatives need to be explicitly written down[2], which reduces the workload by at least 75

 Note:

Generating a Jacobian matrix by AD takes 10-30% more time than hand written code. However, this cost is not considered expensive when one considers the great convenience it provides overall.

1.2 Features And Capabilities

1.2.1 Comprehensive Set of Models

GSS can provide a comprehensive set of physical models, including:

- Drift-diffusion transport models.
- Energy balance transport models.
- Density gradient quantum transport models.
- Lattice heating and heat sinks can be considered with both DD and EB model.
- DC, AC small-signal, and full time-dependency simulation.
- Fermi-Dirac and Boltzmann statistics.
- Advanced mobility models.
- Ohmic, Schottky, insulating contacts and floating metal gate.
- SRH, radiative, Auger, and surface recombination.
- Impact ionization.

[2] Both Synopsys SenTaurus and Silvaco ATLAS require physical model as well as its derivatives provided by user.

- Optoelectronic interactions with finite element electromagnetic solver.
- Band-to-band tunneling.
- Graded and abrupt heterojunctions.

Furthermore, since GSS is an open-source code, the user can easily add their own physical models. The authors will be pleased to assist users with this task.

1.2.2 Numerical Implementation

GSS employs a range powerful numerical techniques, including:

- Accurate and robust discretization techniques.
- Full Newton nonlinear iteration strategy.
- Exact generation of Jacobian matrix by AD.
- The stability of Newton's iteration arithmetic is ensured by powerful line search or trust region method, plus sophisticated damping strategy.
- Efficient solvers, both direct and krylov space based, can be chosen for linear subproblems.
- Dynamic memory allocation, hence no explicit limit to problem size.

1.2.3 Circuit Level Mixed-Type Simulation

Unlimited Mixed-Type Simulation

NGSPICE, the open-source implementation of SPICE3, provides an interface for GSS to perform mixed type circuit simulation. An interface is provided that enables GSS/NIGSPICE to exchange data by TCP/IP protocol via a network. Each GSS process simulates one transistor and is controlled by NGSPICE. This mechanism supports unlimited numerical devices[3] run in a parallel model.

1.3 Brief Overview of this Manual

[3] The max number of numerical device is theoretically limited by the 65535 ports of TCP/IP protocol.

PART I. Introduction to Semiconductor Physics

A Guide For Readers

This chapter briefly summarize results in solid-state physics that are useful in the numerical simulation of semiconductor devices, which we hope would save the reader from checking various scattered sources. Only the results will be presented with brief comments, due to the nature of this book as a technical manual. The readers are strongly recommended to read textbooks on solid-state physics and semiconductor physics for a thorough treatment on these topics. The author believes that the knowledge on the following topics are essential to users of device simulators:

- Band structure theory, where material properties such as band-gap, effective mass and density of states (DOS) are derived;
- Distribution function for carriers at equilibrium (Fermi-Dirac distribution), and
- Transport equations of carriers, which encapsulate the various conservation laws in the motion of carriers in non-equilibrium situations.

All transport models discussed in this book originate from the Boltzmann transport equation (BTE), with simplifications to various extents, and thus are all classical transport models. Quantum transport models are not discussed. The three components combine to provide a complete theory on semiconductor physics.

Chapter 2 Energy Band Structure Of Semiconductor

Energy band structure is one of the mostly complex contents in solid state physics. We only give simplified introduction here. Currently the energy band theory has two models, The tight banding model and the near free electron approximation. The tight banding model builds in the atomic energy level's crystal field stretch. Near free electron approximation bases on free electron's periodic field disturbance, which is more straightforward in $E \sim k$ explanation.

2.1 Tight Binding Model

Tight binding model (tight-binding, TB), is also called linear combination of atomic orbital, LCAO. This model bases on such physical image that believes the molecular's, crystal also belongs to molecular, electronic state is similar as the electronic state in the crystal it composes. The molecular orbits is compose of the linear combination of crystal atomic orbits. When atomic orbits compose of molecular orbits, the orbits number is not changed. Molecular orbits' energy can be higher, lower or equal to atomic orbits' energy level, which are called anti bonding orbit, bonding orbit and non bonding orbit. We need to mention here that in molecular calculation, quantum mechanics perturbance theory are generally used. Because of the difficulty on Schrödinger equation solving, a complicate molecular has no analytical solution[1]. In this case, the disturbance theory is generally used. The perturbance method's basis is to select some states and believe that the real states can be represented with the linear combination of this states. Then by putting the linear combination on the unknown question to fix the basic state coefficients. In molecular calculation[2], if we select the atomic state as the basic state we have the linear combination of atomic orbital, LCAO.

Here, we give a preliminary impression through the simple H_2 molecular. ψ is used to represent H atom outside nuclear electron wave function, because two atoms are completely equal, according to the atomic orbit linear combination model, the H_2 molecular orbit should have the following form:

$$\varphi_+ = C_+(\psi_A + \psi_B) \tag{2.1}$$
$$\varphi_- = C_-(\psi_A - \psi_B) \tag{2.2}$$

Normally we call φ_+ as bonding state, φ_- as antibonding state. For bonding state, the electron clouds is in between the two nucleus and attracted simultaneously by two nucleus, energy level is lower. For antibonding state the density state of electron cloud in between the nucleus are small, energy level is higher, shown below Figure (2.1).

Now considering a crystal made of n atoms. There are $10^{22} \sim 10^{23}$ atoms in every centimeter cube of the crystal. So normally n is a big number. When n atoms has long distance among each others, there is no crystal, every atom's energy level are

[1] the most complicate molecular, which has analytical solution, is H_2^+
[2] In molecular calculation, atomic linear combination of atomic orbits are generally used. eg. famous software Gaussian, which can calculate around 100 atoms molecular structure.

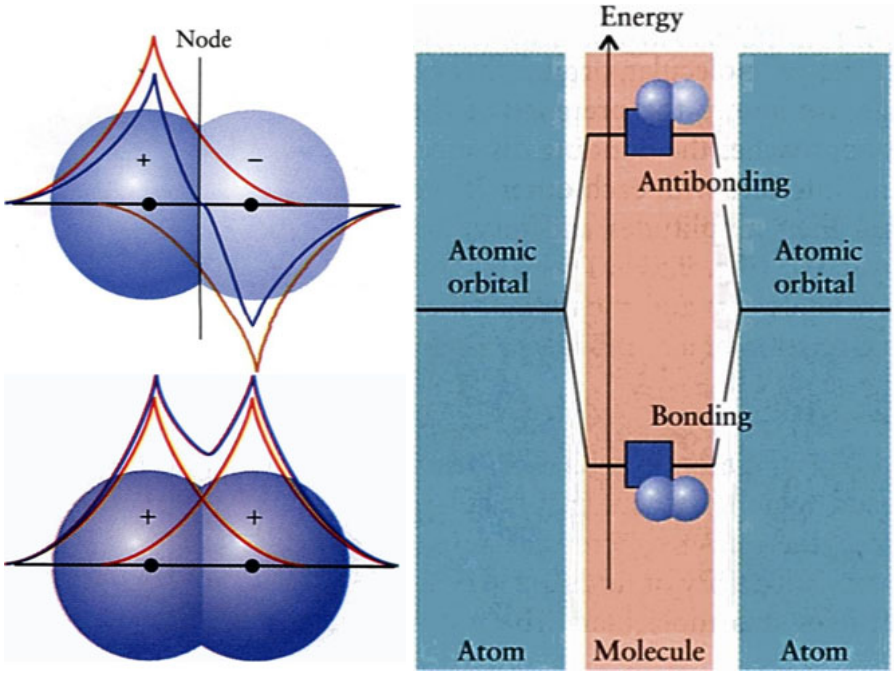


Figure 2.1: H_2 molecular orbit and its energy level

isolated as atoms. They are all n degree digenerated[3]. When n atoms com close to each other and form crystal, every atom is affected by the surrounding atom potential field. Now the real electron orbit is expressed as the linear combination of n atoms, in consequence every atom's energy level split to n levels, shown as Figure (2.2), These n energy levels turn to be a energy band, electrons are not belong to specific atom, but shared by the whole crystal. The each spliced energy band is called permitted band (permitted band can be overlapped shown in Figure (2.3)). Between permitted band, there is no energy level, which is called band gap. Internal electrons were in low energy level, sharing movement is weak, their energy split is very small, energy band is narrow. External electrons, especially valence electrons have strong sharing characteristics. They are similar as free electrons, called near free electrons, energy split is severe, energy band is wide.

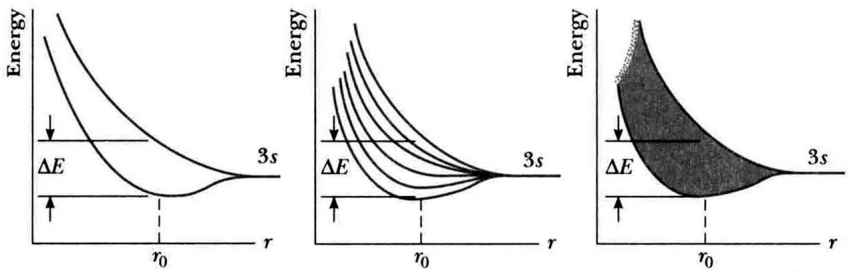


Figure 2.2: Splitting of $3s$ energy levels as two, six, and N atoms come close to form a crystal.

Please pay attention that every energy band includes energy level number is related to isolated atomic energy level. We have to consider the atomic energy degeneration.For example s energy level does not have degeneration,after n atoms combine to be a crystal, s energy level split to n close energy level and form an

[3] Temperarily we don't calculate the self degeneration of atom energy level

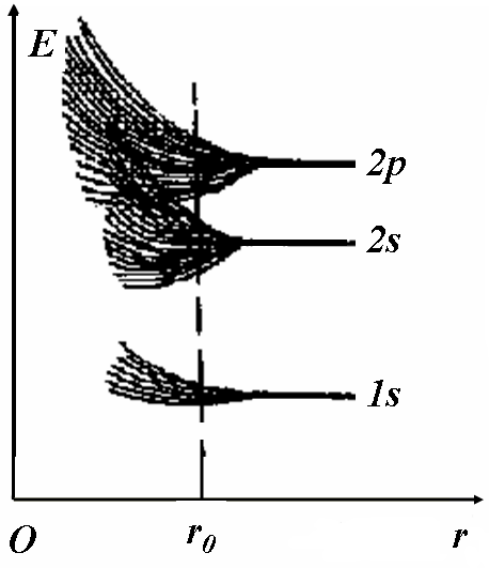


Figure 2.3: Energy band overlap illustration

energy band. Atom's p energy level is 3 level degenerated, which splits to $3n$ close energy level. The real crystal is composed with big number n close energy levels, so every energy band can be treated as continuous, called quasi continuous. We have to point out that real crystal energy band can be more complicate. Crystal energy band may not be corresponding to isolated atoms. Some atomic's external s and p electron form complicate energy level for better bonding. For example IV column Carbon, Silicon and etc. atoms all have 4 valance electrons $2s$ electrons, $2p$ electrons. In this case only two p electrons are used to bond, however all 4 electrons are going to bond with other atoms and form stable crystal in reality. Accordingly we use diamond as an example, electron does not use carbon atom basic state as the base state, whereas it uses sp^3 complex bonding as the base state:

$$\varphi_1 = \frac{1}{2}(\psi_{2s} + \psi_{2px} + \psi_{2py} + \psi_{2pz}) \tag{2.3}$$

$$\varphi_2 = \frac{1}{2}(\psi_{2s} + \psi_{2px} - \psi_{2py} - \psi_{2pz}) \tag{2.4}$$

$$\varphi_3 = \frac{1}{2}(\psi_{2s} - \psi_{2px} + \psi_{2py} - \psi_{2pz}) \tag{2.5}$$

$$\varphi_4 = \frac{1}{2}(\psi_{2s} - \psi_{2px} - \psi_{2py} + \psi_{2pz}) \tag{2.6}$$

To form orbit complex, it needs some energy. However after complex orbit is formed, bond number increases, and due to electron cloud density increases at the corner of tetragon, the bond strength increases, the energy decrease is enough to satisfy the complex formation's energy.

We give electron's filling rule at 0 k temperature below. For normal temperature, the rule follow Fermi-Dirac statistic distribution. Based on Pauli's non accommodating principle and lowest energy rule, electrons will fill the energy level from the lowest energy level to higher energy level. Every energy level can have two spin direction electrons. Energy band filling can be split into full band, valance band and vacant band. Full band is to describe the energy band which is filled out with electrons. Full band does not participate the conducting process. Valance band is formed by valance electrons' energy level split, valance band energy level is high, which can be either filled out or not. Vacant band is corresponding to separate

atom stimulated energy level, normally there is no electron filling in. In metals, valance band acts as conduction band for conducting purpose. In semiconductor, we need to stimulate the valance band electron to the lowest vacant band to conduct current, so the lowest energy level vacant band is called conduction band.

First we give metal, Li's, energy level filling diagram. Li has 2 $1s$ electrons and 1 $2s$ electron, accordingly, after n atoms form metal crystal, $2n$ $1s$ electrons fill in the lowest energy energy band, and n $2s$ electrons will fill half of the second energy band, shown as Figure (2.4). Based on conducting theory, semi filled energy band can conduct current, accordingly metal Li is a good conductor.

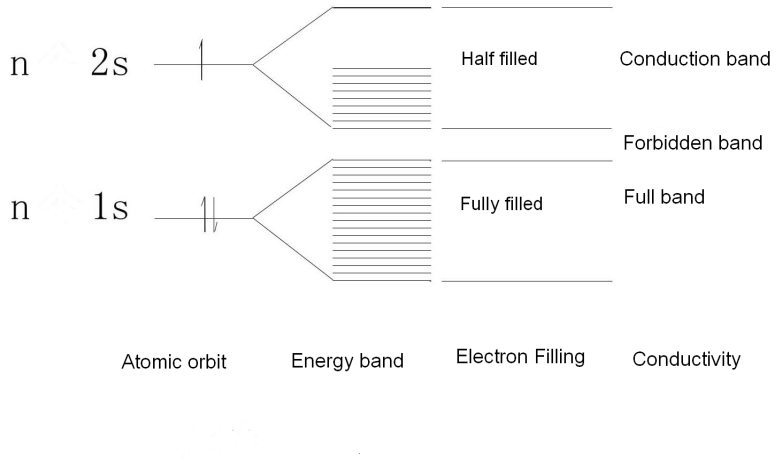


Figure 2.4: Energy band diagram of metal lithium and the filling of electron at the energy levels.

For silicon atomic crystal, it was mentioned before that Silicon extra layer's 4 electron adopt sp^3 complex bonding, Previous $4n$ sp^3 orbit splits to $2n$ bonding orbits and $2n$ anti-bonding orbits. So $4n$ valance electrons fill out all the bonding orbits,and there is no electron in anti-bonding orbits. Based on conducting theory, full band and empty band can not conduct current. Accordingly if there is no thermal emission factor, silicon can not conduct current, shown as Figure (2.5).

We have to explain here that H belongs to I group's alkaline metal, but H can only form molecular crystal with Van der waals force. This is mainly due to H's first ionization energy is as high as 13.6 eV, which is much bigger than alkaline metal, around $4-5$ eV. In gas phase, alkaline metal has dual atomic molecular similarly. But in solid state, alkaline metal atom's extra layer electrons can be easily ionized to form free electrons in metal crystal. The ionization energy can be balanced by the share bonding energy loss. However H_2 molecular's electron is bonded around the nucleus, which can not form sharing electron state. Nevertheless, theocratically H can form metal hydrogen by using cryogenic high pressure accumulation to force the electrons separate from the fetter.

2.2 One Dimension Near Free Electron Model

This section is going to discuss the near free electron model. In crystal, electron is between free electron and fettered electrons. Isolated atom's electrons move around its nuclei and other electrons' potential field. Free electrons are not affected by any outside field. In crystal, electrons are periodically located at nuclei

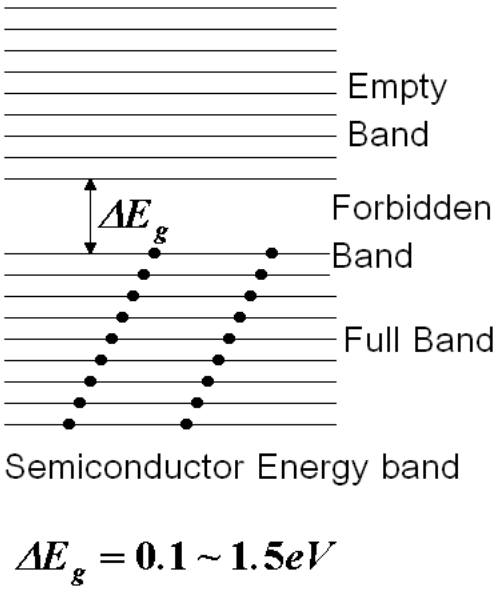


Figure 2.5: Semiconductor energy band diagram and electron filling energy level

potential and large amount of uniform potential. Near free electron model is single electron approximation theory, which treats every electron's movement as independent election in a equalized potential field.

Near free electron model means electrons are not fettered by nuclei, and can move inside the whole solid body. Electrons are called shared electrons. Shared electrons' movement rule is similar as free electrons. Free electrons are illustrated below.

De ploit first propose wave particle duality characteristics for all the micro particles. Free particles all have wave length, frequency, momentum, energy following the relationship below

$$\mathbf{p} = m_0\mathbf{v} = \hbar\mathbf{k} \tag{2.7}$$

$$E = \frac{1}{2}\frac{p^2}{m_0} = \hbar\nu \tag{2.8}$$

with certain momentum and certain energy's free particle is similar as frequency as ν and wave vector as $\mathbf{k}$ plane wave, the relationship between these two are the same as photon and light wave.

One dimension particle's wave function satisfy steady state Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) = E\psi(x) \tag{2.9}$$

Its solution is plane wave.

$$\psi_k(x) = Ae^{ikx} \tag{2.10}$$

$$E_k = \frac{\hbar^2k^2}{2m_0} \tag{2.11}$$

For wave vector $\mathbf{k}$'s moving status, free electron's energy, momentum both have certain value. Accordingly wave vector $\mathbf{k}$ can describe free electron's moving status. The difference is $\mathbf{k}$ represents free electrons different state. Figure (2.6) is free electron's $E \sim k$ curve,is a parabolic curve. Because wave vector $\mathbf{k}$ is continuous, free electron's energy is continuous spectrum, from 0 to infinity.

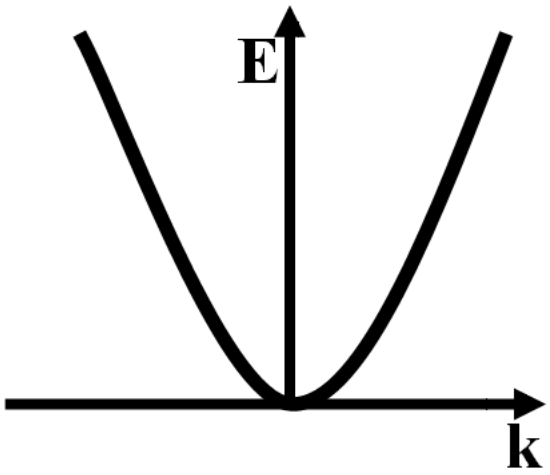


Figure 2.6: Free energy's $E \sim k$ relationship

In crystal electrons move in nuclei potential and other huge amount of electron's average potential field $V(x)$, basically it satisfies

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{2.12}$$

if we can solve this equation, we can obtain electron's wave function in the crystal. However $V(x)$ is very complicate in real crystal. Accordingly normally we use pseudo potential to replace $V(x)$. Pseudo potential is a method to use an artificial field to replace the real potential field, in order to simplify calculation. Accordingly, near free electron model's accuracy is totally dependent on pseudo potential parameters, and because it is single electron, it was very important during the early research of material science, nowadays tight bonding is more popular.

No matter it is real potential field or pseudo potential field, $V(x)$ must satisfy crystal periodical condition.

$$V(x) = V(x + na) \tag{2.13}$$

Where n is integer and a is crystal constant.

F.Bloch proved that those wave functions satisfy Equation (2.12) must have following format

$$\psi_k(x) = u_k(x)e^{ikx} \tag{2.14}$$

$u_k(x)$ is a periodic function which has same period as crystal matrix, as $u_k(x) = u_k(x + na)$. Generally we call Equation (2.14)'s wave function as Bloch wave function.

If we compare the Bloch wave function and free electron wave function, we find only periodic modulation amplitude $u(x)$ is replaced by fixed amplitude A. Electrons can also be at any position of the crystal, this is the history of near free electrons. For specific wave vector k's fix electron, its probability at certain crystal cell $\mid \psi_k^*(x)\psi_k(x) \mid = \mid u_k^*(x)u_k(x) \mid$ changes with x, but for every basic cell's corresponding position, its probability distribution is fixed.

By using perturbance theory, we can obtain Figure (2.7)'s one dimension crystal $E \sim k$ relationship, detail can be referred to Huang Kun's solid state physics.

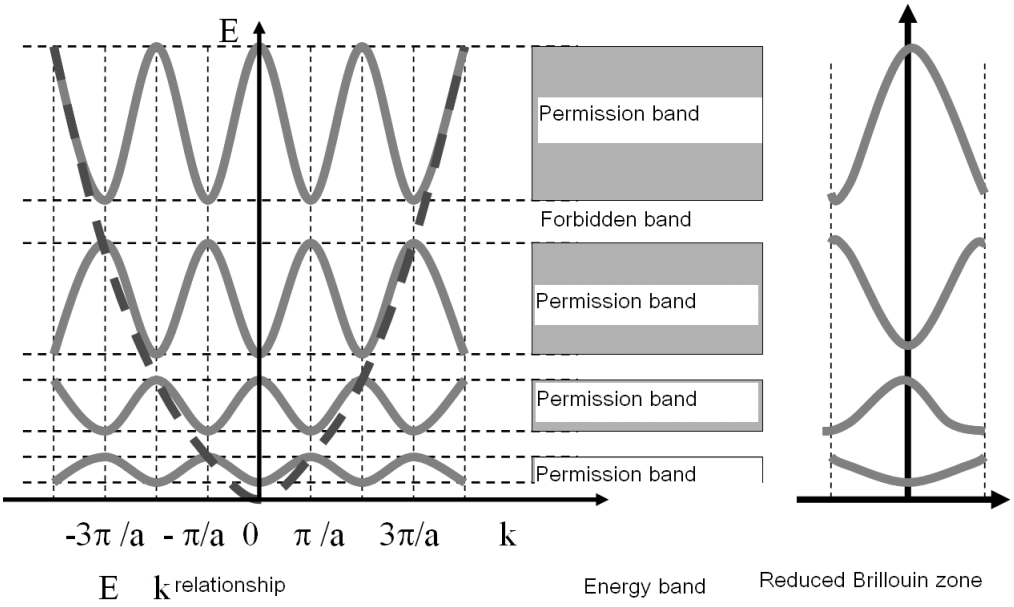


Figure 2.7: One dimension crystal's $E \sim k$ relationship

First, when $k = \frac{2\pi n}{a}$, energy turns to be non-continuous and form a series of permit band and band gap, this result is the same as tight bonding theory. Attention, energy $E(k)$ is k's multiple value function, if we want to know the electron's energy, we have to point out which energy band and its wave vector the electron has. Second, at the same energy band, electron state can be expressed with share movement of wave vector However because crystal has translation symmetry, the k to make sure the electron's status is not single. The following proves

$$k' = k + \frac{2\pi n}{a} \tag{2.15}$$

and k express the same electron state. Equation (2.14) can be written as

$$\psi_{k'}(x) = u_k(x)\exp\left(-i\frac{2\pi n}{a}x\right)\exp\left(i(k+\frac{2\pi n}{a})x\right) \tag{2.16}$$

Because

$$u_k(x)\exp\left(-i\frac{2\pi n}{a}x\right) = u_k(x+na)\exp\left(-i\frac{2\pi n}{a}(x+na)\right) \tag{2.17}$$

which means $u_k(x)\exp\left(-i\frac{2\pi n}{a}x\right)$ is still crystal periodic function, so Equation (2.16) can be written as

$$\psi_{k'}(x) = u_{k'}(x)\exp\left(ik'x\right) \tag{2.18}$$

where $u_{k'}(x) = u_k(x)\exp\left(-i\frac{2\pi n}{a}x\right)$. It means at the same energy band, every other $2\pi/a$'s k describe the same electron state.

Since in the same energy band k repeat with period $2\pi/a$, energy $E(k)$ is also k's period function, we have

$$E(k) = E(k+\frac{2\pi n}{a}) \tag{2.19}$$

then we can take $-\frac{\pi}{a} \sim \frac{\pi}{a}$ region (First Brillouin region)'s k value to express electron's energy state then use other region combine to first Brillouin region, shown as Figure (2.7)'s simple Brillouin region's $E \sim k$ relationship.

For limit length one dimension crystal, we still need to consider certain boundary condition, for Equation (2.12) we introduce boundary condition, which can let k take the following value

$$k = \frac{n}{L} 2\pi \tag{2.20}$$

N is the crystal cell number, $L = Na$ is the total length of the crystal. This result means the wave vector k can be a series of isolated value, and the status inside Brilloin region are uniformly distributed. Every possible k value has length $2\pi/L$.

For three dimension situation, we have

$$\left.\begin{aligned} k_x &= \frac{2\pi n_x}{L_x} \quad (n_x = 0, \pm 1, \pm 2 \cdots) \\ k_y &= \frac{2\pi n_y}{L_y} \quad (n_y = 0, \pm 1, \pm 2 \cdots) \\ k_z &= \frac{2\pi n_z}{L_z} \quad (n_z = 0, \pm 1, \pm 2 \cdots) \end{aligned}\right\}$$

now every k at k space's volume is

$$\frac{(2\pi)^3}{L_x \cdot (L_y \times L_z)} = \frac{(2\pi)^3}{V}$$

where V is crystal's volume. Accordingly, in k space k value's density is $V/(2\pi)^3$.

Because N's a big number, k is also very dense, which can be considered as quasi continuous. We can prove the every energy band has Nk states, because every state can take 2 opposite spin direction electrons, every energy band can take $2N$ electrons.

For real three dimensional crystal, Brillouin region is more complicate. For example Silicon crystal structure is face center cubic matrix, its first Brillouin region is a fourteen face body, shown in Figure (2.8). And three dimension situation is different from one dimension that the energy of different band may not be necessarily separated. It is possible to overlap each other. Please attention that tight bonding theory gives similar conclusion. Figure (2.9) shows that silicon's valance band and conduction band. We can clearly see that Silicon at Γ point has two band with different energy. In valance band there are three energy bands overlapped each other. Band gap is not at Γ point,whereas it is at ΓX axis.

The good thing is semiconductor numerical simulation avoids the complicate energy band structure most of the cases. Because those which can contribute to current are electrons and holes concentrated under the bottom of conduction band or at the bop of the valance band. So we only need to consider the energy band structure at these two areas. After introducing the effective mass concept, the bottom of conduction band and the top of valance band structure can be simply replaced by effective mass.

2.3 Single Electron's Movement in Semiconductor

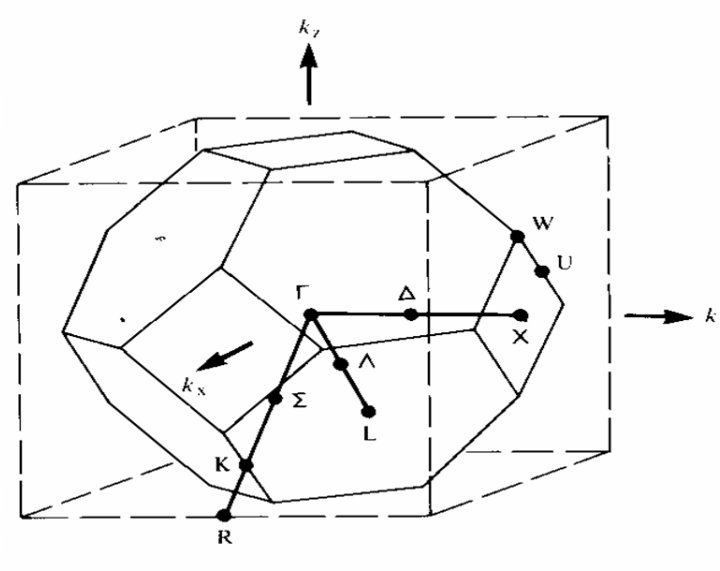


Figure 2.8: Silicon's first Brillouin region

2.3.1 Average electron speed in semiconductors

In semiconductor real problems, electron's movement is treated as canonical particles. For example in transport process, when electron's free distance is far longer than basic cell length, electron can be treated as a canonical particle. Through quantum mechanics calculation, canonical particle's speed can be written as

$$\mathbf{v} = \frac{1}{\hbar}\nabla_k E(k) \tag{2.21}$$

The formulae above means for electrons inside crystal, quasi canonical moving speed $\mathbf{v}$ is dependent on the $E(k) \sim k$ relationship under its condition,if we know $E(k) \sim k$ relationship,we can obtain its moving speed, and construct moving equation.

2.3.2 Effective mass

Although real crystal's band diagram is very complicate, for semiconductor, the useful electrons are located only at the top of valance band and the bottom of conduction band. Accordingly it would be enough if we know the $E(k) \sim k$ relationship around the extreme value.

By using Taylor progression expansion, we can obtain the $E(k) \sim k$ approximation around the extreme value. We use one dimension case as example, assume conduction band's bottom is at wave number $k = 0$, Expand $E(k)$ Taylor progression around $k = 0$, we have

$$E(k) = E(0) + \left(\frac{\mathrm{d}E}{\mathrm{d}k}\right)_{k=0} k + \left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)_{k=0} k^2 + \cdots \tag{2.22}$$

Because when $k = 0$ energy has extreme low value, its first order derivative is 0, we have

$$E(k) - E(0) = \left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)_{k=0} k^2 \tag{2.23}$$

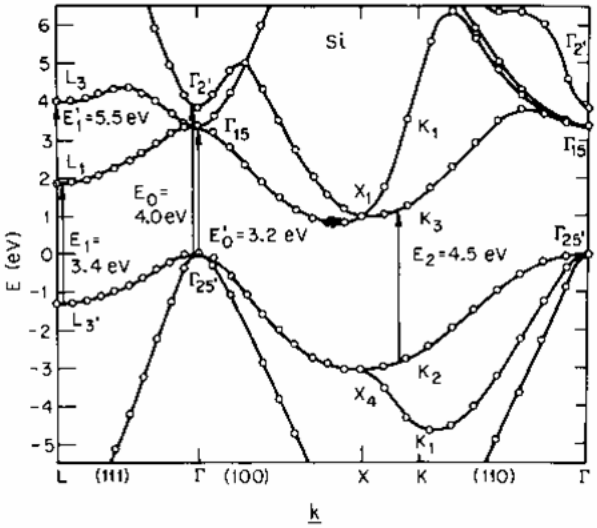


Figure 2.9: Silicon's energy band structure

For given semiconductor, energy's second order derivative should be a certain number. Let

$$\frac{1}{\hbar^2}\left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)_{k=0} = \frac{1}{m_n^*} \tag{2.24}$$

Put Equation (2.24) into Equation (2.23), we have the bottom of the band's $E(k)$

$$E(k) - E(0) = \frac{\hbar^2 k^2}{2m_n^*} \tag{2.25}$$

Equation (2.25) and Equation (2.11) free electrons' $E(k) \sim k$ relationship are similar. What is different is Equation (2.11)'s m_0 is electrons' inertia mass, here the m_n^* is the effective mass at the bottom of the band. Because $E(k) > E(0)$, the effective mass of the bottom of the conduction band is positive.

Similarly assume the top of the band is at $k = 0$, we can also have

$$E(k) - E(0) = \left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)_{k=0} k^2 \tag{2.26}$$

let

$$\frac{1}{\hbar^2}\left(\frac{\mathrm{d}^2E}{\mathrm{d}k^2}\right)_{k=0} = \frac{1}{m_n^*} \tag{2.27}$$

top of the band $E(k)$ is

$$E(k) - E(0) = \frac{\hbar^2 k^2}{2m_n^*} \tag{2.28}$$

m_n^* is called the effective mass at the top of the band. Because at the top of band $E(k) < E(0)$, m_n^* is negative.

From Equation (2.25) and Equation (2.28) we notice, after introducing effective mass, if we can measure its quantity, We can make sure about the $E(k) \sim k$ relationship near energy band extreme value.

The discussion above is all based on one dimension condition, at three dimension condition, because energy band is not symmetrical, the effective mass is expressed as tensor format.

2.3.3 Outside force and electron acceleration

In reality, many semiconductor devices works at certain voltage, there is electric field inside semiconductor. The electrons experiences not only the periodic potential's effect, but also the electric field from the outside electric field. The periodic potential has already replaced by effective mass, the outer electric field will affect the electrons' speed variation.

With outside force $\mathbf{F}$ on the electron, after $\mathrm{d}t$ time, the momentum of the electron will increase

$$\mathrm{d}E = \mathbf{F} \cdot \mathbf{v}\mathrm{d}t = \mathbf{F} \cdot \frac{1}{\hbar}\nabla_k E \mathrm{d}t \tag{2.29}$$

Following electron energy E and $\mathbf{k}$'s function relationship, $\mathbf{k}$ will certainly changes $\mathrm{d}\mathbf{k}$. Accordingly electron's energy after $\mathrm{d}t$ time will increase:

$$\mathrm{d}E = \nabla_k E \cdot \mathrm{d}\mathbf{k} = \nabla_k E \cdot \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t}\mathrm{d}t = \left(\frac{\hbar \mathrm{d}\mathbf{k}}{\mathrm{d}t}\right) \cdot \left(\frac{1}{\hbar}\nabla_k E\right)\mathrm{d}t \tag{2.30}$$

Compare the previous two formulae, we have

$$\frac{\mathrm{d}(\hbar\mathbf{k})}{\mathrm{d}t} = \mathbf{F} \tag{2.31}$$

Equation (2.31) is the electron moving status with the outer force. For example constant outside field, electron will move with uniform speed in $\mathbf{k}$ space. Equation (2.31) has similar format as Newton rule, only use $\hbar\mathbf{k}$ to replace canonical mechanics' momentum. Accordingly $\hbar\mathbf{k}$ has equivalent momentum format, called quasi momentum. without outside force status $\mathbf{k}$ does not change, so quasi momentum does not change.

Accordingly the outside force leads $\mathbf{k}$ to change with time, so that the electron speed follows the variation of the time, it means electron accelerates. $\mathbf{a} = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}$. In one dimension case, based on Equation (2.21)'s acceleration is

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{1}{\hbar^2}\frac{\mathrm{d}^2 E}{\mathrm{d}k^2}\mathbf{F} \tag{2.32}$$

By using effective mass, the formulae above can be written as

$$\mathbf{F} = m^*\mathbf{a} \tag{2.33}$$

In this case, the electron's acceleration under outside force has the same format as the canonical mechanics.

We introduce single electron inside crystal's movement under outside field. It seems that the problem has already been solved. If we reconsider the electron's movement under different scattering mechanism, we can describe electron's transportation problem. However semiconductor macro phenomenon depends on huge number of particles. The huge number of electrons inside the crystal leads to computational difficulty. So in the following pages, we are going to adopt macro partial model to decrease the tracing number of particles. But he calculation load is still huge. So we need to deal with this problem in macroscopic way and have basic continuous materials' Boltzmann transport equation.

Chapter 3 Balanced Electron Statistic Distribution

3.1 Electron's Fermi-Dirac Statistic Distribution

From this chapter, we are not only limited on single particle's description. And by using statistical rule to describe huge number of electrons inside the crystal. Based on quantum statistics, spinning $\frac{1}{2}$ electron follow Fermi-Dirac statistic distribution. The probability of energy level E is occupied is

$$f_{FD}(E) = \frac{1}{1 + \exp(\frac{E - E_F}{k_b T})} \tag{3.1}$$

$f_{FD}(E)$ is called electron's Fermi distribution function, it is a distribution function which is used to describe the electron's distribution under thermal stability. k_b is Boltzmann constant, T is the systematical thermal temperature.

If given $E \sim k$ relationship, through transformation, we can obtain Fermi distribution function with momentum as the self viable, it is used generally in semiconductor quantum mechanics calculation to fix the boundary condition. For example the bottom of conduction band, we have

$$E(k) = \frac{h^2 k^2}{2m^*} \tag{3.2}$$

now, Fermi distribution function can be expressed as

$$f_{FD}(k) = \frac{4\pi m^* k_b T}{h^2} \ln\{1 + \exp[-\frac{1}{k_b T}(\frac{h^2 k^2}{8\pi^2 m^*} - E_F)]\} \tag{3.3}$$

Formulae (3.1) and Formulae (3.3)'s E_F is called Fermi energy level, in semiconductor, it is just related to temperature, semiconductor material, doping concentration and system zero point selection. E_F is a very important parameter, when E_F is given, electron at different energy level's statistical distribution are totally fixed. E_F can be fixed by material energy band's occupied energy level number equals to electron number. Discuss later.

We integrate semiconductor's huge amount of electron as a thermal system. Quantum calculation theory proves that Fermi level E_F is the system's chemical potential μ. Because the system at thermal stability has uniform chemical potential, the electron system at thermal stability has a fixed fermi energy level.

From Formulae (3.1), when $T = 0\text{K}$,

$$f_{FD}(E) = \begin{cases} 1 & E < E_F \\ 0 & E > E_F \end{cases} \tag{3.4}$$

so at 0K, energy level less than E_F are all occupied by electrons. And energy level large than E_F are all vacant. Accordingly in 0K, E_F is the boundary of whether energy level is occupied by an electron.

When $T > 0$K, as Formulae (3.4) shows, the probability of energy level equal to E_F is 0.5.

$$f_{FD}(E) \begin{cases} > \frac{1}{2} & E < E_F \\ = \frac{1}{2} & E = E_F \\ < \frac{1}{2} & E > E_F \end{cases} \tag{3.5}$$

Figure (3.1) gives Fermi function under different temperature. We can see, at $T > 0$ fermi function and 0K fermi function's difference is only several k_bT closed to E_F.

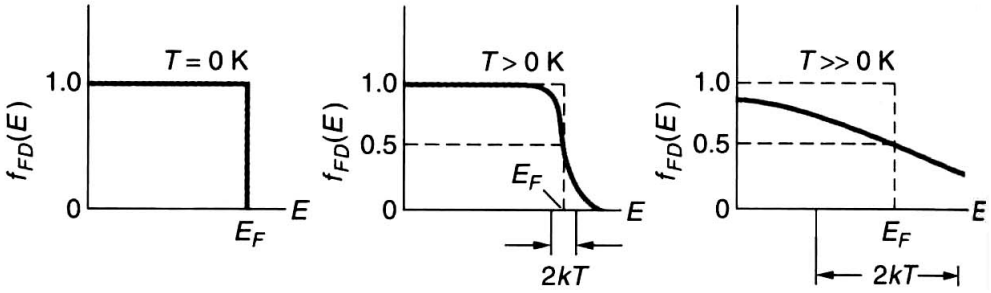


Figure 3.1: Fermi distribution and temperature relationship

The probability of electron occupying energy level $5k_bT$ higher than E_F is < 0.007, which is very small and almost empty. Energy level which is $5k_bT$ lower than E_F has probability > 0.993 for electron occupation. So we know the quantum state are filled with electrons.

Normally we can consider at temperature, which is not too high, energy larger than E_F's energy level does not have electrons, energy level less than E_F's energy levels are occupied by electrons. The probability of electron occupy E_F level at any temperature is always 0.5. Normally we consider E_F marking the electrons' filling level.The higher E_F is the more high energy level can be occupied by electrons.

Attention for specific energy level's electron density is decided by Fermi level E_F and state density g together. In semiconductor, because electron are partially filled at the bottom of conduction band, the valance band top are occupied by holes. Which means 50% filling probability's E_F is normally inside the band gap, shown below Figure (3.2). Because the energy band's limitation, E_F itself is not occupied by electron, which means E_F has 0 energy state density. The electron density detail calculation will be discussed in the third section.

3.2 Boltzmann Distribution Function

Because Fermi integral does not have analytical solution except several special cases. The application is not convenient. The good thing is in most cases, we can simply Fermi distribution function to Boltzmann distribution function.

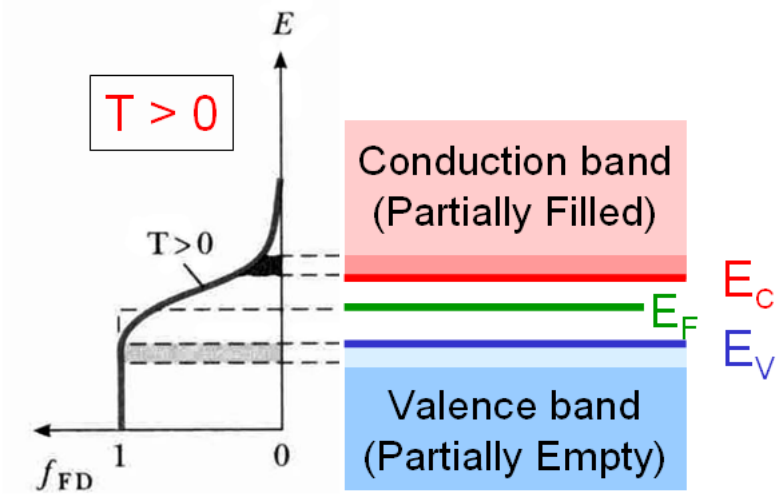


Figure 3.2: The Fermi energy level of intrinsic semiconductor

When $E - E_F \gg k_bT$, $\exp(\frac{E-E_F}{k_bT}) \gg 1$, so

$$1 + \exp(\frac{E-E_F}{k_bT}) \approx \exp(\frac{E-E_F}{k_bT})$$

now fermi distribution function can be approximately written as

$$f_{FD}(E) \approx f_B(E) = \exp(-\frac{E-E_F}{k_bT}) \tag{3.6}$$

as typical Boltzmann distribution. Attention when $E - E_F \gg k_bT$, the probability of energy level is occupied by electrons are very small, and the main difference between Fermi distribution and Boltzmann distribution is the previous one affected by Pauli non accommodating principle's limitations. When $E - E_F \gg k_bT$ two or more electrons can occupy same electron state, the probability of this kind of things happening is very small. Accordingly Pauli non accommodating principle's influence on result is very small, two distribution function's results has almost no difference.

3.3 Hole's Distribution Function

In order to differentiate, we are going to use $f_n(E)$ to express electron distribution function. Since $f_n(E)$ express energy as E's electron state occupied probability, then $1 - f_n(E)$ is the probability of that state being occupied, the hole distribution:

$$f_p(E) = 1 - f_n(E) = \frac{1}{1 + \exp(\frac{E_F - E}{k_bT})} \tag{3.7}$$

Similarly, when $E_F - E \gg k_bT$, the formulae above's 1 in denominator can be neglected, we have hole's Boltzmann distribution function.

$$f_p(E) = \exp(-\frac{E_F - E}{k_bT}) \tag{3.8}$$

Generally, semiconductor's E_F is inside the band gap and has distance bigger than k_bT to the bottom of conduction band and top of the valance band. So

the electron in conduction band and hole in valance band follows Boltzmann distribution. Generally we call those electron systems which satisfy Boltzmaan distribution as non degenerate systems. The corresponding semiconductor turns to be non degenerate semiconductor.

But under high doping and low temperature condition, E_F can be close or even go into the conduction band or valance band. In this case we have to go back to fermi distribution. At room temperature, for silicon, the doping must be higher than 10^{19}cm^{-3}, so that we need to consider fermi distribution, But for GaAs, when doping concentration arrives 10^{17}cm^{-3} we need to consider using fermi distribution as shown in Figure (3.5) and Figure (3.6).

3.4 Carrier's Concentration

Now we discuss the carriers concentration under semiconductor thermal stability. We know carriers occupying quantum state, and every other unit time quantum state number, state density $g(E)$. As discussed before, in k space the electron status distribution is uniform, the state density is $2V/(2\pi)^3$. Here we consider two opposite spinning directions' state.Based on the specific $E \sim k$ relationship, we can get energy band's state density from k space's state density. Because the carriers in semiconductor are mostly under the bottom of the conduction band and the top of valance band, we only need to consider this part's $E \sim k$ relationship and state density. Assume at the bottom of conduction band $k = 0$, and we have parabolic $E \sim k$ relationship:

$$E = E_c + \frac{\hbar^2 k^2}{2m_n} \tag{3.9}$$

Assume sphere equal energy face, energy inside $E \sim E + \mathrm{d}E(k \sim k + \mathrm{d}k)$'s state density is

$$\mathrm{d}z = \frac{2V}{(2\pi)^3} \times 4\pi k^2 \mathrm{d}k \tag{3.10}$$

put into Equation (3.9), Let k expressed by E, we have

$$\mathrm{d}z = \frac{V\left(2m_n\right)^{2/3}}{2\pi^2\hbar^3}\left(E - E_c\right)^{1/2}\mathrm{d}E \tag{3.11}$$

This way, we have the state density at the bottom of the conduction band:

$$g_c\left(E\right) = \frac{\mathrm{d}z}{\mathrm{d}E} = \frac{V\left(2m_n\right)^{2/3}}{2\pi^2\hbar^3}\left(E - E_c\right)^{1/2} \tag{3.12}$$

Similarly, we can have the state density close to the top of the valance band:

$$g_v\left(E\right) = \frac{V\left(2m_p\right)^{2/3}}{2\pi^2\hbar^3}\left(E_v - E\right)^{1/2} \tag{3.13}$$

Here although we assume the sphere equalized energy face, for other conditions, by introducing state density effective mass, we still have Equation (3.12) and Equation (3.13).

After we have the state density, we consider the energy level is continuous inside the energy band, carriers' density can be written as integral format:

$$n_0 = \frac{1}{V}\int_{E_c}^{+\infty} g_c(E) f_n(E)\,\mathrm{d}E \tag{3.14}$$

$$p_0 = \frac{1}{V}\int_{-\infty}^{E_v} g_v(E) f_p(E)\,\mathrm{d}E \tag{3.15}$$

3.4.1 Non degenerate semiconductor carriers' concentration's relationship with Fermi energy level

For non degenerate semiconductor, using Boltzmann distribution, put state density Equation (3.12) and distribution Formulae (3.6) into Equation (3.14), we have

$$n_0 = \int_{E_c}^{+\infty} \frac{(2m_n^*)^{2/3}}{2\pi^2\hbar^3}(E-E_c)^{1/2}\exp(-\frac{E-E_F}{k_bT})\,\mathrm{d}E \tag{3.16}$$

the integral above have analytical solution, by using variable replacement we have

$$n_0 = 2\Big(\frac{m_n^* k_b T}{2\pi\hbar^2}\Big)^{3/2}\exp(-\frac{E_c-E_F}{k_bT}) = N_c\exp(-\frac{E_c-E_F}{k_bT}) \tag{3.17}$$

in the above formulae,

$$N_c \equiv 2\Big(\frac{m_n^* k_b T}{2\pi\hbar^2}\Big)^{3/2} \tag{3.18}$$

is called conduction band bottom effective state density.

Similarly we can obtain valance band's hole concentration's expression

$$p_0 = \int_{-\infty}^{E_v} \frac{(2m_p^*)^{2/3}}{2\pi^2\hbar^3}(E_v-E)^{1/2}\exp(-\frac{E_F-E}{k_bT})\,\mathrm{d}E \tag{3.19}$$

we have

$$p_0 = 2\Big(\frac{m_p^* k_b T}{2\pi\hbar^2}\Big)^{3/2}\exp(-\frac{E_F-E_v}{k_bT}) = N_v\exp(-\frac{E_F-E_v}{k_bT}) \tag{3.20}$$

in the formulae above

$$N_v \equiv 2\Big(\frac{m_p^* k_b T}{2\pi\hbar^2}\Big)^{3/2} \tag{3.21}$$

is called valance band top effective state density.

Formulae (3.17) and Formulae (3.20) express the carrier's concentration n_0, p_0 and E_F's relationship. Because E_F is not only related to temperature but also to the semiconductor doping energy level and doping concentration density. Accordingly even for the same semiconductor material, because of the dopants difference and doping concentration's difference at the same temperature electron and hole's concentration can be very different. But from Equation (3.17) and Equation (3.20) we have n_0 and p_0's product

$$n_0 p_0 = N_c N_v\exp(-\frac{E_c-E_v}{k_bT}) = N_c N_v\exp(-\frac{E_g}{k_bT}) \tag{3.22}$$

which does not contain E_F, accordingly it is not related to doping[1], when the material is fixed, it is just the function of temperature. When the temperature is not changed $n_0 p_0$'s product keeps constant. The detail number is decided by material's energy band parameter. So conduction band electron concentration and valance band hole concentration are limited to each other.

[1] We do not consider the band gap narrowing effect here

3.4.2 Degenerate semiconductor concentration's relationship with Fermi energy level

For degenerated semiconductor, following Fermi distribution, similarly we can obtain the relationship between electron/hole's concentration and E_F.

$$n_0 = N_c F_{1/2}(\eta_n) \tag{3.23}$$

$$p_0 = N_v F_{1/2}(\eta_p) \tag{3.24}$$

in above formulae

$$\eta_n = \frac{E_F - E_c}{k_b T}$$
$$\eta_p = \frac{E_v - E_F}{k_b T}$$

$F_{1/2}(\eta_s)$ is 1/2 order fermi integration:

$$F_{1/2}(\eta_s) = \frac{2}{\sqrt{\pi}} \int_0^{+\infty} \frac{\eta^{1/2}}{1 + \exp(\eta - \eta_s)} \, \mathrm{d}\eta \tag{3.25}$$

The fermi integration does not have analytical solution, normally we adopt numerical method to calculation. When accuracy request is high, we can use chebyshev multiple items interpolation. GSS used a simple method:

$$F_{1/2}(\eta) = \frac{2\sqrt{\pi}}{3\sqrt{\pi} a^{-3/8} + 4\exp(-\eta)}$$
$$a = \eta^4 + 33.6\eta(1 - 0.68\exp(-0.17(\eta+1)^2)) + 50$$

In η from $-\infty$ to $+\infty$'s variation, the analytical above formulae's result and accurate value's error is less than 0.4%. In this case another advantage is we can obtain $F'_{1/2}$'s analytical expression, which is very useful for the later Newton iteration solving E_F.

Till now we already have n_0, p_0 and E_F's relationship, because the previous two equations have three variables, we need to add electric neutralization condition to make the complete equation set.

3.4.3 Intrinsic semiconductor carriers' concentration

For intrinsic situation, normally we can use Boltzmann distribution, electric neutralization condition is $n_0 = p_0$, by using Equation (3.17) and Equation (3.20), we can solve the intrinsic semiconductor fermi level as:

$$E_i = E_F = \frac{1}{2}(E_c + E_v) + \frac{1}{2} k_b T \ln \frac{N_v}{N_c} \tag{3.26}$$

The formulae above's first item is at the center of the band gap, the second item of GeÂřÂćSiÂřÂćGaAs is around $k_b T$ level, which is far less than the first item in room temperature. It means E_i locates at the middle of the band gap approximately, following Boltzmann distribution condition. Define

$$\Psi_{\text{intrinsic}} = -\frac{1}{q} E_i \tag{3.27}$$

as the intrinsic fermi potential, this quantity is used for electric potential's reference in semiconductor simulation.

Put Equation (3.26) into Equation (3.17) or Equation (3.20), we have intrinsic carriers' concentration n_i as

$$n_i = n_0 = p_0 = (N_c N_v)^{1/2} \exp(-\frac{E_g}{2k_b T}) \tag{3.28}$$

and compared with Equation (3.22), we have semiconductor carriers concentration production's another expression:

$$n_0 p_0 = n_i^2 \tag{3.29}$$

which is any non degenerate semiconductor's carrier concentration product, which equals to corresponding temperature's intrinsic carrier's concentration square and independent to doping concentration. By using intrinsic carrier concentration n_i, intrinsic fermi energy level E_i can transform n_0, p_0's expression Equation (3.17) and Equation (3.20) to

$$n_0 = n_i \exp(\frac{E_F - E_i}{k_b T}) \tag{3.30}$$

$$p_0 = n_i \exp(\frac{E_i - E_F}{k_b T}) \tag{3.31}$$

sometimes these two formulae are more convenient.

3.4.4 Band gab narrowing effect and effective intrinsic carrier concentration

In this section we only discussing single ionized shallow energy level doping Doped semiconductor's N type doping, called Donor. Normally donor energy level E_D is close to E_c, the donor electron is easy to jump into conduction band, which lead to increase of conduction electron density and high conductivity. Simultaneously fermi energy level moves to conduction band shown in Figure (3.3).

When doping semiconduction is P type dopant, we call it acceptor, normally acceptor energy level E_A is close to E_v, because acceptor is lack of electrons. The valance band is easy to jump into acceptor energy level and form holes in the valance band, which leads to hole concentration increase and conductivity increase. Simultaneously fermi energy level turns to valance band shown in Figure (3.4).

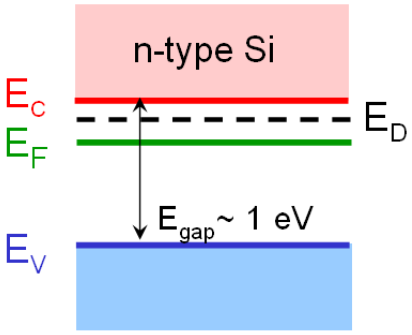


Figure 3.3: Donor energy level

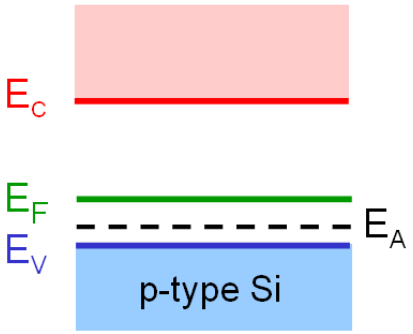


Figure 3.4: Acceptor energy level

In most semiconductor devices, there is always region doping concentration higher than 10^{18}cm^{-3}, these highly doped region is very important for deciding the device's characteristics. In high doping area, the carriers' concentration must adopt

fermi distribution and we need to consider the band shape variation due to the high doping.

During high doping, there are two main factors affecting electron and hole's state density: first is the function between carriers and between carriers and ionized dopant and the second is uniformity of dopant and the dopants' wave function overlap's effect on energy band tain and dopant energy band. These two factors together will lead to band gap narrowing.

If we consider Fermi distribution function and band gap narrowing effect, the carriers' concentration's expression is comparatively complicate. For simplification, we can introduce intrinsic carrier concentration concept. In this way we can can relatively accurate result much easier.

Assume under high doping concentration, valance band's top $E_{v,eff}$, conduction band's bottom $E_{c,eff}$ carrier's concentration can be expressed as:

$$\begin{aligned} n &= N_c F_{1/2}\left(\frac{E_F - E_{c,eff}}{k_b T}\right) \\ p &= N_v F_{1/2}\left(\frac{E_{v,eff} - E_F}{k_b T}\right) \end{aligned} \tag{3.32}$$

Similar to define intrinsic carrier concentration, define intrinsic carrier concentration as:

$$n_{ie}^2 = n_i^2 \beta \exp\left(\frac{\Delta E_g}{k_b T}\right) \tag{3.33}$$

ΔE_g is the band gap width's variation value, β is the factor to represent the degeneration.

$$\beta = \mathrm{F}_{1/2}\left(\frac{E_F - E_{c,eff}}{k_b T}\right) / \exp\left(\frac{E_F - E_{c,eff}}{k_b T}\right)$$

now, Equation (3.30) and Equation (3.31) turn to be

$$n_0 = n_{ie} \exp(\frac{E_F - E_i}{k_b T}) \tag{3.34}$$

$$p_0 = n_{ie} \exp(\frac{E_i - E_F}{k_b T}) \tag{3.35}$$

Equation (3.33) needs to calculate fermi integral, in reality it is very inconvenient, normally we use empirical formulae to fit it.

After introducing effective intrinsic carrier concentration, within certain wide doping concentration, carrier's concentration can still be expressed with Boltzmann distribution. For total doping concentration less than $8 \times 10^{18}\mathrm{cm}^{-3}$ and doping balance degree is less than 10%, we adopt effective intrinsic carrier concentration method for better result, its error is less than 10%. For higher doping concentration and high doping balanced material, its result are not very good.

3.4.5 Doped semiconductor carriers' concentration

Based on the result of previous section, for not very high doped cases, using effective intrinsic carrier concentration with Boltzmann distribution can have very good result, otherwise we are going back to fermi distribution.

Because semiconduction simulation consider's temperture is normally around$300K$and doping level is not very high, we can consider dopants are fully ionized, the electric neutral condition is relatively easy:

$$n_0 + N_A = p_0 + N_D \tag{3.36}$$

N_A and N_D are donor and acceptor doping concentrations. Together Equation (3.34) and Equation (3.35), we gave

$$n_0 = \frac{(N_D - N_A) + \sqrt{(N_D - N_A)^2 + 4n_{ie}^2}}{2} \tag{3.37}$$

$$p_0 = \frac{-(N_D - N_A) + \sqrt{(N_D - N_A)^2 + 4n_{ie}^2}}{2} \tag{3.38}$$

$$E_F = E_i + k_b T \ln\left(\frac{n_0}{n_{ie}}\right) = E_i - k_b T \ln\left(\frac{p_0}{n_{ie}}\right) \tag{3.39}$$

The previous three formulae can be used to calculate steady state uniform semiconductor's internal properties. Besides, the ohmic contact between semiconductor and metal, we consider semiconductor keeps steady state. The previous three formulae can be used to calculate the electron, hole concentration and static electric potential at this point, so we can fix the boundary condition at this point.

For highly doped and low temperature condition, we have to use degenerate fermi distribution, simultaneously for more accurate result, we need to consider incomplete ionization. In this case effective donor concentration and acceptor concentration are

$$N_D^+ = \frac{N_D}{1 + g_D \exp(\frac{E_F - E_D}{k_b T})} \tag{3.40}$$

$$N_A^+ = \frac{N_A}{1 + g_A \exp(\frac{E_A - E_F}{k_b T})} \tag{3.41}$$

g_D and g_A are donor and acceptor energy's degeneration degree, for Ge, Si or GaAs, $g_D = 2$, $g_A = 4$. E_D and E_A are donor and acceptor energy level.

Now the electric neutralization condition is

$$n_0 + N_A^+ = p_0 + N_D^+ \tag{3.42}$$

To fix E_F, n_0 and p_0 we need to solve Equation (3.42) together with Equation (3.32). Since no analytical solution exists, we can only use Newton iteration method to get the numerical result. The detail calculation process can be found in "??", on page ??.

In the end, experiments measured Silicon, GaAs's fermi level and temperature, doping relationship is shown in Figure (3.5) and Figure (3.6).

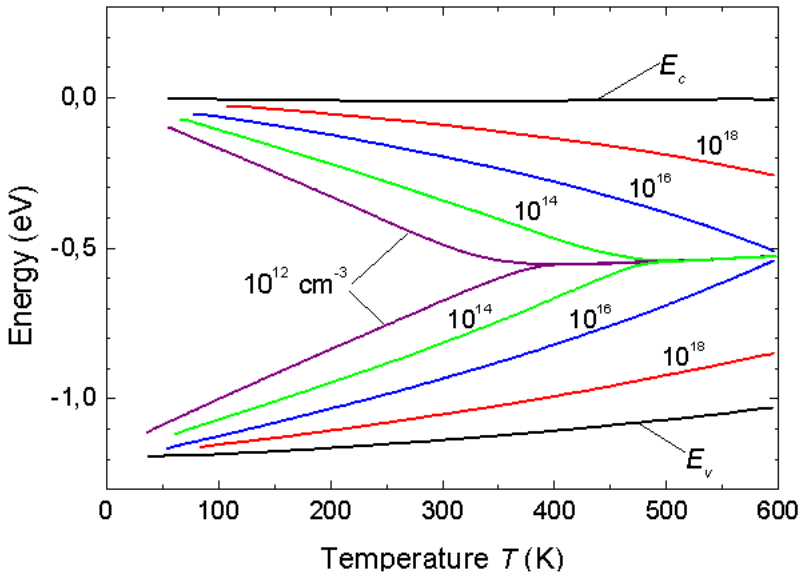


Figure 3.5: Silicon fermi level and doping concentration curve

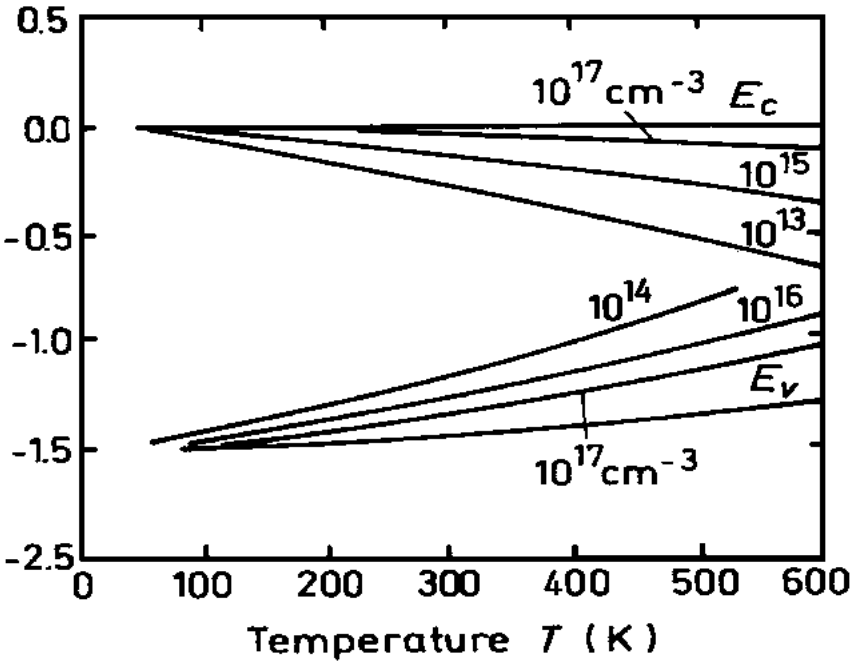


Figure 3.6: GaAs fermi level and doping concentration curve

Chapter 4 Non-Equilibrium Carriers

4.1 Quasi Fermi Level

In last chapter, we discussed the problems with the thermal stability assumption. In this thermal stability's carrier concentration is called balanced carrier concentration. If it is non degenerate semiconductor, electron, hole's production satisfy[1]

$$n_0 p_0 = N_c N_v \exp\left(-\frac{E_g}{k_b T}\right) = n_i^2 \tag{4.1}$$

Attention thermal stability is not absolute motionless sate, electrons inside semiconductor are still moving continuously. However under thermal balance the electron jumping from valance band to conduction band is the same as the electrons jumping from the conduction band to the valance band, which keeps the carrier's generation and recombination balance.

However, semiconductor's thermal stability is conditional. If we put outside force on the semiconductor to destroy the thermal balance condition, it is called non equilibrium state. Now Equation (4.1) is not correct. For example the light emission semiconductor's carrier has big recombination, which is far more than the equilibrium state. The carrier more than the balanced state is called surplus carrier. When the light stops, non equilibrium state can not be sustain. Following half life time around micro second's exponential degradation to zero. This half life time degradation is called non balanced carrier's life time. For different semiconductor material, different material process condition, surplus carriers life time can vary in a big scope. For silicon, around tens of micro seconds, for GaAs, only several nano second.

when semiconductor's electron system is at balance state. The whole semiconductor has uniform fermi energy level, electron, hole are described by it. Under non degenerated condition follows Equation (3.17) and Equation (3.20). Under degenerated conditions it follows Equation (3.23) and Equation (3.24). Because there is uniform fermi energy level E_F, in thermal stability state, semiconductor's electron and hole concentration product must satisfy Equation (4.1). Accordingly there is uniform fermi energy level, which is similar as balance state.

When outside's impact destroy the thermal stability, there is no uniform fermi energy level. However in general conditions, because semiconductor's surplus carrier's life time is at $10^{-8} \sim 10^{-3}$ second scope, the carrier and crystal's energy transfer's collision's relaxation time (in order to differentiate from momentum relaxation time, we call it energy relaxation time.) is around 10^{-10} second and below. Accordingly surplus carriers will collide with the crystal matrix many times from the generation to recombination, and fully exchange energy. If the crystal thermal capacity is relatively big, we can consider conduction band electron and crystal matrix or valance band hole and crystal matrix are independent from thermal stability and share the same temperature [2]. This kind of state similar to thermal stability state is called quasi thermal stability state. Certainly, now

[1] for simplification, we use intrinsic carrier concentration to express. In reality all use effective intrinsic carrier concentration.

[2] It is only suitable for drift diffusion model. In fluid dynamic model, electron, hole temperature are not equal to crystal temperature, and thermal carriers exist.

at the conduction band electron system and valance band hole system does not have thermal stability, so there is no uniform fermi energy level. But the quasi thermal stability can consider conduction band electron system itself and valance band itself, following Boltzmann distribution or Fermi distribution, separately have their own fermi energy level. The corresponding electron fermi level is E_{F_n} and hole fermi level is E_{F_p}. Besides, now the fermi energy level is localized value. At different semiconductor area, surplus carrier concentrations are different, fermi levels are different.

After introduction fermi level, non stable state carrier concentration can be approximate with similar formulae, if the carrier concentration is not very high so that E_{F_n} and E_{F_p} go into conduction band or valance band, we can use Boltzmann distribution:

$$n = N_c \exp\left(-\frac{E_c - E_{F_n}}{k_b T}\right) = n_0 \exp\left(\frac{E_{F_n} - E_F}{k_b T}\right) = n_i \exp\left(\frac{E_{F_n} - E_i}{k_b T}\right) \tag{4.2}$$

$$p = N_v \exp\left(-\frac{E_{F_p} - E_v}{k_b T}\right) = p_0 \exp\left(\frac{E_F - E_{F_p}}{k_b T}\right) = n_i \exp\left(\frac{E_i - E_{F_p}}{k_b T}\right) \tag{4.3}$$

Known carrier concentration, we can use the previous formulae to fix fermi level E_{F_n} and E_{F_p}'s position.

$$E_{F_n} = E_c + k_b T \ln \frac{n}{Nc} \tag{4.4}$$

$$E_{F_p} = E_v - k_b T \ln \frac{p}{Nv} \tag{4.5}$$

Now we need to fix the electron concentration and hole concentration's productÂčâĹń

$$np = n_0 p_0 \exp\left(\frac{E_{F_n} - E_{F_p}}{k_b T}\right) = n_i^2 \exp\left(\frac{E_{F_n} - E_{F_p}}{k_b T}\right) \tag{4.6}$$

We can see E_{F_n} and E_{F_p}'s difference directly reflects np and $n_0 p_0$'s difference, which represent the semiconductor's distance from thermal stability. The bigger the difference from etcher, the higher the semiconductor is from the thermal stability. If there is no difference, there is uniform fermi energy level, so semiconductor is at the thermal stability state.

4.2 Carriers Recombination and Generation

Because semiconductor internally interacts with each other, which leads to certain electron and hole'e existence, and solely decided by fermi level's position. But the stability does not mean the electron and hole do not jump from different energy levels. As discussion before, from microscopic point of view, valance electron jump to conduction band electron form electron hole pair. Simultaneously, conduction band electron continuously jump back to valance band to recombine with holes and disappear. Stability state means the semiconductor internally different energy level's jumping is balanced. The jumping impact on generation and recombination does not lead to system macroscopic effects. However when the semiconductor is affected by outside force, the carrier concentration are not equal to stable concentration, the surplus generation and recombination will be reflected to macroscopic phenomenon.

In semiconductor the carrier's recombination has many different path, but there are two main types: band to band electron hole direct recombination and through band gap recombination center's recombination. The previous one is the meet between electron and hole. Electron jump from the conduction band's certain state to valance band's hole's state. Simultaneously release the energy (photon or

phonon). The recombination is called direct band to band recombination. shown as figure Figure (4.1)(a) illustrate: it is also possible that two electron (two holes) recombine with one hole (one electron), one of the electron will transfer the energy to another electron (hole). This kind of process is called Auger process, shown in figure Figure (4.2)(b) illustrate, this kind of recombination is called band to band Auger recombination. Another type of carrier move through a defect or doping center and are captured. And then the center recapture the reverse carrier so as to finish the recombination process. Simultaneously release surplus energy. This kind of recombination is called recombination through recombination center. This kind of defect or doping center is called recombination center. The released energy can be photon or phonon, shown as Figure (4.1)(a) illustrate, and also can take Auger process, shown as Figure (4.2)(a).

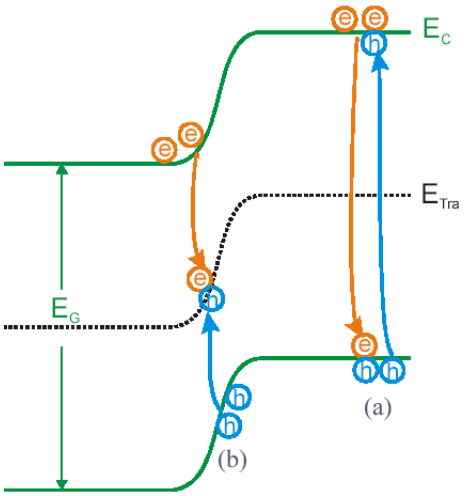


Figure 4.1: Direct recombination and doping center recombination

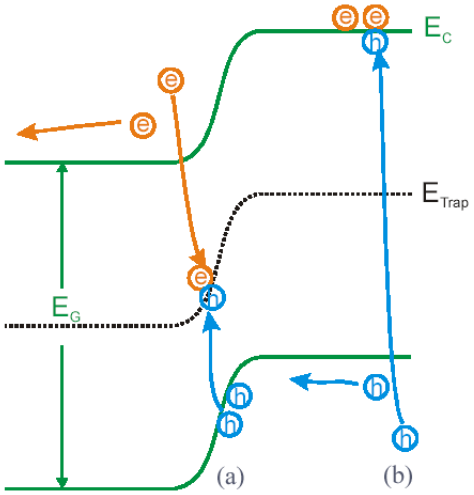


Figure 4.2: Auger recombination

4.2.1 Band to band Direct recombination

For band to band direct recombination we can assume recombination rate R proportional to carrier concentration n and p as

$$R = rnp \tag{4.7}$$

where r is recombination coefficient, it is not related to n and p, because during thermal stability we have

$$G_0 = R_0 = rn_0p_0 \tag{4.8}$$

then the surplus carrier's net recombination rate isÂčâĹń

$$U = R - G_0 = r(np - n_0p_0) \tag{4.9}$$

Assume the non equilibrium carrier $n = n_0 + \Delta n$, $p = p_0 + \Delta p$ and $\Delta n = \Delta p$, we have

$$U = r(n_0 + p_0)\Delta p + r(\Delta p)^2 \tag{4.10}$$

If there is only direct recombination, we can obtain surplus carrier's life time:

$$\tau = \frac{\Delta p}{U} = \frac{1}{r(n_0 + p_0 + \Delta p)} \tag{4.11}$$

Carrier's life time τ, is dependent on recombination probability r. Based on intrinsic light absorption data, and combining with the theoretic calculation we have r's value. Based on theocratical calculation in room temperature intrinsic silicon's $r = 10^{-11}\text{cm}^2/\text{s}$ and $\tau = 3.5\text{s}$. However real silicon material's carrier lifetime is much lower than the data, the maximum life time is only several milliseconds. This case means, for silicon life time is not dependent on direct band to band recombination, there must be other recombination mechanisms dominating the carrier's life time. It is discussed below about recombination through recombination center.

4.2.2 Auger indirect recombination

Different from direct recombination, band to band Auger recombination process is a process with three particles. Accordingly the recombination rate R will not be proportional to np, but proportional to n^2p or np^2. For two electrons and one hole process

$$R_{an} = r_{an}n^2p \tag{4.12}$$

For two holes and one electron process

$$R_{ap} = r_{ap}p^2n \tag{4.13}$$

$r_{an}$and$r_{ap}$are called Auger recombination coefficient.

On the contrary of Auger electron-hole recombination process, there is impact ionization process. A high energy electron (or hole)collide another electron in valance band and activate the electron into conductor band to form a pair of electron-hole pair, simultaneously jump to the bottom of the conduction band (or jump to the top of the valance band). This impact ionization's probability is proportional to high energy or hole's concentration. In non degenerate condition, it is proportional to total concentration n or p, now its generation rate is:

$$G_{an} = g_{an}n \tag{4.14}$$
$$G_{ap} = g_{ap}p \tag{4.15}$$

In thermal stability, we should have$R_{an0} = G_{an0}$,$R_{ap0} = G_{ap0}$ so we have

$$g_{an} = r_{an}n_i^2 \tag{4.16}$$
$$g_{ap} = r_{ap}n_i^2 \tag{4.17}$$

$$U_{Auger} = r_{an}(pn^2 - nn_i^2) + r_{ap}(np^2 - pn_i^2) \tag{4.18}$$

Assume non balance carrier $n = n_0 + \Delta n$, $p = p_0 + \Delta p$ and $\Delta n = \Delta p$, we have the life time of carrier

$$\tau_n = \tau_p = \frac{1}{(n_0 + p_0 + \Delta p)(r_{an}n + r_{an}p)} \tag{4.19}$$

Because band to band Auger recombination is a three body problem, its probability is not high when carrier concentration is not high. Generally Auger band to band recombination is important at narrow band gap semiconductor and high temperature conditions.

4.2.3 Recombination through recombination center

In semiconductor the defect or dopants introduced some deep energy level center. Although at room temperature its impact on conductivity is not obvious, their existence can help carriers recombination. Normally this kind of dopant and defect center is effective recombination center.

Recombination center's impact on carriers can be described as four period: electron/hole capturing, electron stimulation (electron is thermal stimulated from recombination energy level E_t to the conduction band.) or hole stimulation (Hole is thermal stimulate to valance band from E_t.

Although these four processes happens simultaneously, only after the recombination center captured an electron and a hole, the electron hole recombination process is finished. Shockley, Read and Hall first discussed this situation, their theory is called SRH recombination theory.

Let N_r be recombination center concentration,n_r as the concentration of recombination centers, which capture electrons, $p_r = N_r - n_r$, is those without capturing electron recombination center concentration.

$$R_n = r_n n(N_r - n_r) \tag{4.20}$$
$$R_p = r_p p n_r \tag{4.21}$$
$$G_n = e_n n_r \tag{4.22}$$
$$G_p = e_p(N_r - n_r) \tag{4.23}$$

r_n and r_p are called recombination center's electron/hole capturing coefficients. e_n and e_p are called electron and hole generation coefficients.

Obviously, during thermal stability we have $R_{n0} = G_{n0}$ and $R_{p0} = G_{p0}$. We can lead to

$$\begin{cases} e_n = r_n n_l \\ e_p = r_p p_l \end{cases} \tag{4.24}$$

In above formulae

$$n_l = N_c \exp\left(-\frac{E_c - E_t}{k_b T}\right) = n_i \exp\left(\frac{E_t - E_i}{k_b T}\right) \tag{4.25}$$
$$p_l = N_v \exp\left(-\frac{E_t - E_v}{k_b T}\right) = n_i \exp\left(\frac{E_i - E_t}{k_b T}\right) \tag{4.26}$$

Recombination center's net capture rate for electrons and holes are

$$U_n = r_n n(N_r - n_r) - e_n n_r \tag{4.27}$$
$$U_p = r_p p n_r - e_p p_r \tag{4.28}$$

During stability state from $U = U_n = U_p$ we have

$$n_r = N_r \frac{r_n n_l + r_p p_l}{r_n(n + n_l) + r_p(p + p_l)} \tag{4.29}$$

Put the above formulae into U_n or U_p we have

$$U = \frac{(pn - n_i^2)}{\tau_p(n + n_l) + \tau_n(p + p_l)} \tag{4.30}$$

$$\tau_n = \frac{1}{r_n N_r} \tag{4.31}$$
$$\tau_p = \frac{1}{r_p N_r} \tag{4.32}$$

Recombination rate U can also be written as

$$U = \frac{pn - n_i^2}{\tau_p \left[n + n_i \exp\left(\frac{E_t - E_i}{k_b T} \right) \right] + \tau_n \left[p + n_i \exp\left(-\frac{E_t - E_i}{k_b T} \right) \right]} \tag{4.33}$$

In the above formulae, when $E_t \sim E_i$, U turns to be maximum. Accordingly those deep energy level close to the band gap center is the most effective recombination center. For example Cu, Fe, Au and etc. in Silicon.

4.2.4 Carrier's impact ionization

In strong electric density, electron and hole are accelerated in the electric field, they can have very big kinetic energy. Those energy larger than $\frac{3}{2}E_g$'s electrons and holes can knock the electron in valance band out during the collision with crystal lattice to form conduction electron and simultaneously form a hole. From energy band's point of view, it means high energy electron or hole stimulate the electron from valance band to conduction band to form electron hole pair, shown as Figure (4.3). The electron and hole from collision can move to the reverse directions and stimulate another collision to form next generation electron-hole pair. Carrying this process, carriers number can increase qualitatively, this effect of carrier breading is called avalanche effect. Because of avalanche effect, large number of carriers are generated in unit period, current increases sharply so that breakdown happens. This is called avalanche breakdown mechanism, shown as Figure (4.4). Avalanche breakdown normally happens at reverse bias pn junction.

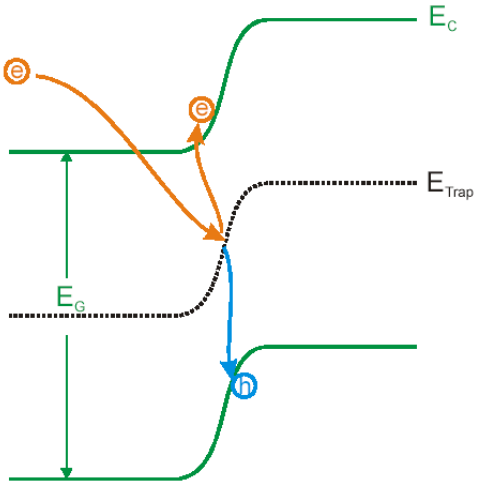


Figure 4.3: Carrier's impact ionization

Avalanche breakdown not only needs strong electric field, but also certain thickness of of electric field. Because the carriers' kinetic energy's increase needs certain acceleration region. If the strong electric field region is too small, particles leave the region before accelerate to high energy, which can lead impact ionization. Avalanche breakdown will not happen.

4.2.5 Carrier's band to band tunneling

For diode, if pn junction's both side has not very high concentration and there is no sharp variation, reverse breakdown carriers' generation is dominated by

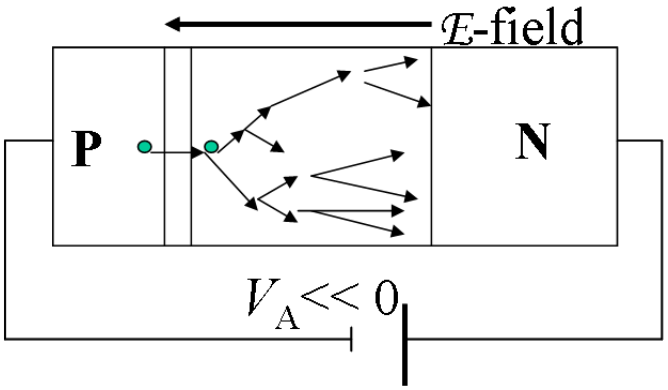


Figure 4.4: Diode's avalanche mechanism

impact ionization. For highly doped diode, when reverse biased electric field less than impact ionization requested electric field, the breakdown possibly happens because of carriers band to band tunneling.

In relatively high reverse bias, those electrons in the top of pn junction's p region valance band can has higher energy than those electrons at the bottom of conduction band. shown as Figure (4.5), which leads to that those electrons at the valance band of p region can tunnel to the conduction band of n region with quantum mechanic tunneling effect. The tunneling effect generates a hole in valance band and simultaneously form a electron and conduction band. When reverse bias voltage increases to V_B, as the tunneling rate increases, this tunneling electrons reach certain quantity, leads to very big reverse current, which makes pn junction breakdown. Because Zerner first explain electric dielectric breakdown phenomenon, it is also called Zener breakdown.

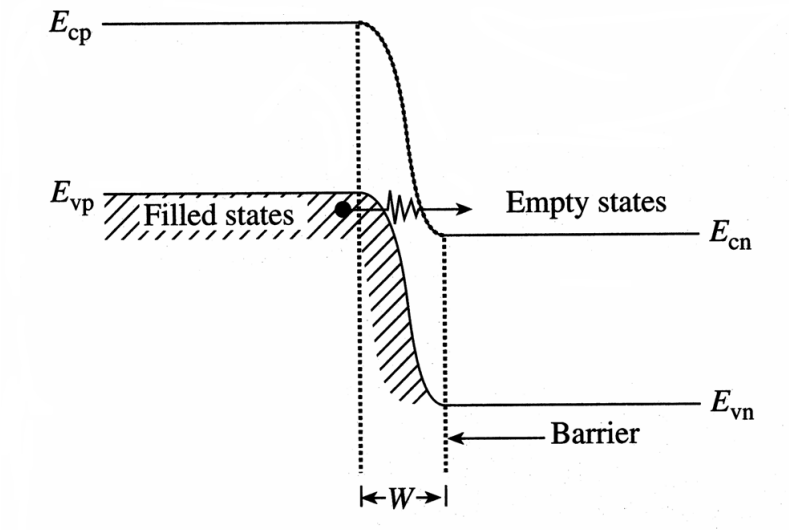


Figure 4.5: Zener diode tunneling process

Based on quantum mechanics theory, electron with energy E's probability of passing energy barrier W is

$$p = \exp\left[-\frac{2}{\hbar}\int_0^W \sqrt{2m_e\,(U-E)}dx\right] \tag{4.34}$$

Approximate the potential barrier in Figure (4.5) to be triangle with height E_g and width w. So that uniform potential barrier internal electric field F is uniform, and we have $U = qFx$. First assume electron's initial energy as zero, after calculation we have the tunneling rate as

$$p = \exp\left[-\frac{4\sqrt{2m_e}}{3\hbar}\frac{E_g^{2/3}}{qF}\right] \tag{4.35}$$

This result illustrates that the tunneling rate is closely related to the barrier region's electric field. After considering the electron's initial energy should be electric field's function, we have the formulae [2], this is GSS internal tunneling model parameters.

Generally tunneling happens when both sides of the pn junction are highly doped. And the transit region is very narrow. Its breakdown voltage $V_B < 4E_g/q$. For example some special zener diode have stable voltage after the tunneling no matter how much current go through, which is generally used for voltage stabilization. Closed to NMOS source electrode, when n^+ type's source region and background doping forms sharp junction, high voltage can also lead to tunneling phenomenon.

Different from Avalanche breakdown, tunneling breakdown does not happen at certain V_B suddenly. As electric field increase, tunneling rate increases continuously, device's current increases accordingly. So it is called soft breakdown.

Chapter 5 Carrier Transport Equation

In this chapter the carrier transport models used in device simulation are discussed. The relationship between various models are illustrated in Figure (5.1). We shall confine ourselves to the semi-classical transport models originated from the Boltzmann transport equation. Quantum transport models are beyond the scope of this book.

We shall start from the Boltzmann Transport Equation (BTE) first, and then proceed to derive the two most widely used models, namely the drift-diffusion (DD) model and the hydrodynamic (HD) model, by taking the low-order moment equations. Since a number of simplifying assumptions have to be used to obtain the DD and HD models, they have limited applicability. It is essential for users of device simulators to appreciate the limitations of each models, which we shall enumerate and discuss.

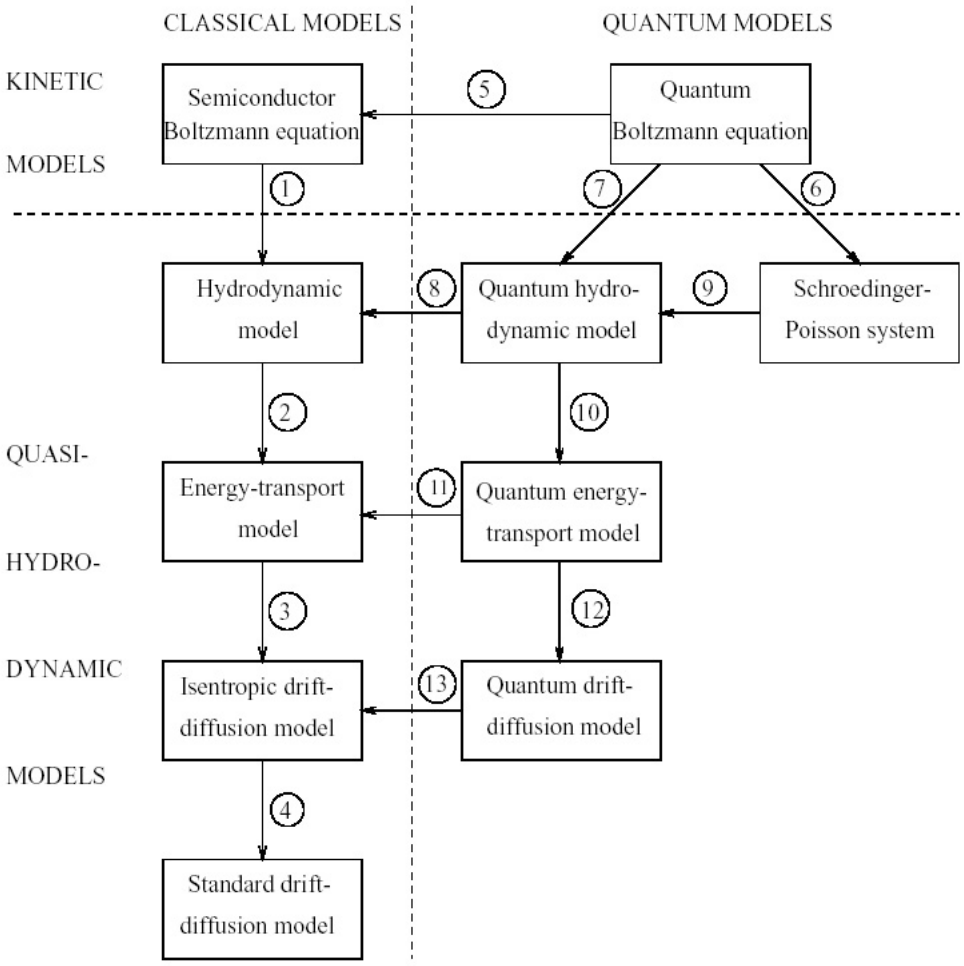


Figure 5.1: Relationship between common semiconductor transport models.

5.1 Boltzmann Transport Equation

A piece of semiconductor material is said to be under thermal equilibrium when there is no external field, and the temperature is uniform. The probability of finding an electron occupying an energy state $E(\mathbf{k})$ is, according to Fermi-Dirac distribution

$$f_0 = \frac{1}{1+\exp\left[\dfrac{E(\mathbf{k})-E_F}{k_bT}\right]}. \tag{5.1}$$

For non-degenerate semiconductor, it is usually adequate to use the simpler Boltzmann distribution function

$$f_0 = \exp\left[-\frac{E(\mathbf{k})-E_F}{k_bT}\right]. \tag{5.2}$$

The system deviates from equilibrium in the presence of external field or temperature gradients, hence the distribution function of electrons deviates from f_0. Let $f(\mathbf{k},\mathbf{r},t)$ be the non-equilibrium distribution function for electrons, the number of electrons in the small volume $\mathbf{r}\sim\mathbf{r}+\mathrm{d}\mathbf{r}$ in real space, $\mathbf{k}\sim\mathbf{k}+\mathrm{d}\mathbf{k}$ in reciprocal space, and at time t is

$$\mathrm{d}N(\mathbf{k},\mathbf{r},t) = 2f(\mathbf{k},\mathbf{r},t)\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}, \tag{5.3}$$

where the coefficient 2 accounts for the spin degeneracy. We turn to examine the dynamic equation that governs the evolution of the distribution function $f(\mathbf{k},\mathbf{r},t)$.

After a short period of time $\mathrm{d}t$, the number of electrons in the same volume element at time $t+\mathrm{d}t$ becomes

$$\mathrm{d}N(\mathbf{k},\mathbf{r},t+\mathrm{d}t) = 2f(\mathbf{k},\mathbf{r},t+\mathrm{d}t)\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}. \tag{5.4}$$

For small $\mathrm{d}t$, we expand the above equation as a Taylor series, and take the lowest order terms

$$\mathrm{d}N(\mathbf{k},\mathbf{r},t+\mathrm{d}t) = 2\left[f(\mathbf{k},\mathbf{r},t)+\frac{\partial f}{\partial t}\mathrm{d}t\right]\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r} \tag{5.5}$$

Obviously, the number of electrons in the volume $\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}$ increases at the rate

$$2\frac{\partial f}{\partial t}\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}. \tag{5.6}$$

Therefore the change in electron number is mainly due to the change in distribution function. In the following, we discuss the two processes leading to the changes in distribution function, namely, the drift process and the scattering process.

5.1.1 Drift Process

The semi-classical motion of electrons, in the absence of collisions, causes the distribution function to evolve in both $\mathbf{k}$ space and r space, whose rate is denoted by $\left(\dfrac{\partial f}{\partial t}\right)_d$. We know that at $t+\mathrm{d}t$, the electron at position $\mathbf{r}$ came from $\mathbf{r}-\mathbf{v}\mathrm{d}t$, where $\mathbf{v}$ is the electron velocity. Similarly, the electron with wave vector $\mathbf{k}$ originally has the wave vector $\mathbf{k}-(\mathrm{d}\mathbf{k}/\mathrm{d}t)\mathrm{d}t$, where $(\mathrm{d}\mathbf{k}/\mathrm{d}t)$ represents the acceleration of the electron under external field. As a result, the number of electrons in the volume $\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}$ increases by

$$\begin{aligned} 2\left(\frac{\partial f}{\partial t}\right)_d\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r} &= 2\left[f(\mathbf{k}-\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t}\mathrm{d}t,\mathbf{r}-\mathbf{v}\mathrm{d}t,t)-f(\mathbf{k},\mathbf{r},t)\right]\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}/\mathrm{d}t \\ &= -2(\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t}\cdot\nabla_k f+\mathbf{v}\cdot\nabla_r f)\mathrm{d}\mathbf{k}\mathrm{d}\mathbf{r}. \end{aligned} \tag{5.7}$$

5.1.2 Scattering Process

Electrons are constantly being scattered, which causes abrupt changes in the wave vector $\mathbf{k}$ of the scattered electrons. The electron distribution function changes accordingly. In the pervious section, it is seen that the distribution function deviates from the equilibrium distribution f_0 due to acceleration of electrons under external field. In the contrary, the scattering process tends to restore the equilibrium distribution. In fact, the distribution function can reach steady-state only in the presence of scattering. The contribution of scattering to evolution of the distribution function is denoted $\left(\frac{\partial f}{\partial t}\right)_s$.

There are a dozen of different scattering mechanisms in semiconductor. The scattering rate due to various mechanisms have very different functional form, but are generally function of the electron wave vector k. The detailed calculation of scattering process is almost always performed using Monte Carlo simulation, due to the enormous complexity. In other types of simulators, relaxation time approximation is used, where the scattering term is approximated by

$$\left(\frac{\partial f}{\partial t}\right)_s = -\frac{f - f_0}{\tau(\mathbf{k})}, \tag{5.8}$$

where f_0 is the Fermi distribution in equilibrium state, and $\tau(\mathbf{k})$ the relaxation time. It is obvious that when the external field is removed, the scattering process causes the distribution function exponentially decays towards f_0.

The carrier mobility used in numerical simulation are derived from the carrier relaxation time, with further approximations, where its dependence on $\mathbf{k}$ is replaced by a simpler dependence on carrier energy E, or even treated as a constant. The calculation of carrier mobility will be discussed in a separate section.

5.1.3 Boltzmann Transport Equation

We can now assemble the complete Boltzmann Transport Equation as follows

$$\frac{\partial f}{\partial t} = -\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} \cdot \nabla_k f - \mathbf{v} \cdot \nabla_r f - \frac{f - f_0}{\tau(\mathbf{k})}. \tag{5.9}$$

We learnt from Equation (2.31), that

$$\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = \frac{\mathbf{F}}{\hbar}. \tag{5.10}$$

Therefore, we need an appropriate expression of the force due to external field $\mathbf{F}$ to describe the carrier transport in semiconductor, which we will describe in the following section.

5.2 Electromagnetic Field in Semiconductor

The charged carriers in semiconductor devices are driven by the electromagnetic field in the device, and the force exerted on an electron is given by Lorentz force law

$$\mathbf{F} = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{5.11}$$

On the other hand, the electromagnetic field induced by these charged carriers is described by, in general, the Maxwell's equations, which must be coupled to

the transport equations described in the previous section. In semiconductor devices, we can make some simplifications on the Maxwell's equations, and use the equivalent d'Alembert's equations

$$\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu \mathbf{J} \tag{5.12}$$

$$\nabla^2 \psi - \frac{1}{c^2}\frac{\partial^2 \psi}{\partial t^2} = -\frac{\rho}{\varepsilon} \tag{5.13}$$

$$\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \psi}{\partial t} = 0 \tag{5.14}$$

where $\mathbf{A}$ is the magnetic vector potential, ψ the electric The constants used include the local speed of light c, the permittivity ε and magnetic permeability μ. When substituted into Equation (5.11), we have

$$\mathbf{F} = e(-\nabla\psi - \frac{\partial \mathbf{A}}{\partial t} + \mathbf{v} \times \nabla \times \mathbf{A}). \tag{5.15}$$

We now shall estimate the three terms in Equation (5.15) and assess their relative importance. Let the characteristic length L and characteristic length τ be related by the typical carrier velocity of $10^5 \mathrm{m} \cdot \mathrm{s}^{-1}$

$$v = \frac{L}{\tau} \ll c, \tag{5.16}$$

and by estimating the derivatives using the appropriate characteristic lengths, one obtains

$$\frac{A}{L^2} - \frac{1}{c^2}\frac{A}{\tau^2} = -\mu J \tag{5.17}$$

$$\frac{\psi}{L^2} - \frac{1}{c^2}\frac{\psi}{\tau^2} = -\frac{\rho}{\varepsilon} \tag{5.18}$$

$$\frac{A}{L} + \frac{1}{c^2}\frac{\psi}{\tau} = 0 \tag{5.19}$$

From Equation (5.19) we have

$$A = -\frac{L}{c^2 \tau}\psi, \tag{5.20}$$

which simplifies Equation (5.15) to

$$F = e(\frac{\psi}{L} + \frac{L^2}{c^2\tau^2}\frac{\psi}{L} - \frac{vL}{c^2\tau}\frac{\psi}{L}) \tag{5.21}$$

Since the last two terms in Equation (5.21) are smaller than the first term by a factor in the order of $\left(\frac{v}{c}\right)^2 \ll 1$, they can be safely dropped. The expression for external force on carriers is therefore reduced to Equation (5.22). Similarly, inspecting Equation (5.18), one sees that to obtain ψ, it is sufficient to solve Poisson's equation Equation (5.23).

$$\mathbf{F} = -e\nabla\psi \tag{5.22}$$

$$\nabla^2 \psi = -\frac{\rho}{\varepsilon} \tag{5.23}$$

The above discussion only considered field induced by the electric charge in the device, we now turn to consider the effect of electromagnetic interference. Assume

that the intensity of incident electromagnetic wave is 100 $\mathrm{W}\cdot\mathrm{cm}^{-2}$, which is a very strong EM radiation. The corresponding electric field is $\mathbf{E} = 1.94\times10^{4}\ \mathrm{V}\cdot\mathrm{m}^{-1}$, and the energy deposited is in the order of 10^{-6} W, in a transistor with 1 μm gate length. The internal electric field under 5 V operation voltage, in comparison, is as high as $5\times10^{6}\ \mathrm{V}\cdot\mathrm{m}^{-1}$. As a result, the effect of electromagnetic interference is negligible in the calculation of semiconductor device characteristics. Even in microwave range, the wavelength is orders of magnitudes higher than the dimensions of semiconductor devices. The micrometer-sized devices are extremely inefficient as antenna. Therefore, the EM interference couples to the semiconductor devices through the surge current induced in the PCB traces.

5.3 Drift-Diffusion Model

We shall derive the drift-diffusion model from the Boltzmann equation here and highlight the simplifying assumptions. It is hoped that this would help illustrate the applicability and limitations of the drift-diffusion model.

Firstly, the evolution of distribution function, which is usually a delicate balance between the between the drift and scattering processes, is a smooth process. We therefore have

$$\frac{\partial f}{\partial t} \ll -\frac{\mathbf{F}}{\hbar}\cdot\nabla_k f - \mathbf{v}\cdot\nabla_r f, \tag{5.24}$$

and the Boltzmann transport equation Equation (5.9) can be simplified to

$$\frac{\mathbf{F}}{\hbar}\cdot\nabla_k f + \mathbf{v}\cdot\nabla_r f = -\frac{f-f_0}{\tau}. \tag{5.25}$$

Secondly, when the electric field is not too strong, the distribution function deviates only slightly from the equilibrium distribution f_0. With $\nabla f \approx \nabla f_0$, $\nabla_k f \approx \nabla_k f_0$, we have

$$-\frac{f-f_0}{\tau} = \frac{\mathbf{F}}{\hbar}\cdot\nabla_k f_0 + \mathbf{v}\cdot\nabla_r f_0. \tag{5.26}$$

Further, recall that the distribution function at equilibrium is really the Fermi-Dirac distribution function

$$f_0 = \frac{1}{1+\exp\left[\dfrac{E_c(x,\mathbf{k})-E_F(x)}{k_b T(x)}\right]}. \tag{5.27}$$

Assuming a single valley, parabolic, spherical band structure, $E_c(x,k)$ can be written as

$$E_c(x,\mathbf{k}) = E_{c_0} - q\psi(x) + \frac{\hbar^2\mathbf{k}^2}{2m^*}, \tag{5.28}$$

where $\psi(x)$ is the electrostatic potential due to the external field. We can now proceed to evaluate the gradients of f,

$$\nabla f = f_0(1-f_0)\nabla\left(\frac{q\psi + E_F}{k_b T}\right); \tag{5.29}$$

$$\nabla_k f = -f_0(1-f_0)\frac{\hbar^2\mathbf{k}}{m^* k_b T}. \tag{5.30}$$

Since

$$\mathbf{F} = -q\mathbf{E} = q\nabla\psi(x); \tag{5.31}$$

$$\mathbf{v} = \frac{\hbar\mathbf{k}}{m^*}, \tag{5.32}$$

we can simplify Formulae (5.26) assuming uniform temperature in the device, or $\nabla T = 0$,

$$f = f_0 - \tau f_0(1 - f_0)\frac{\mathbf{v}}{k_b T} \cdot \nabla E_F. \tag{5.33}$$

Finally, current density is calculated as the product of the group velocity and the distribution function of electrons, integrated in the reciprocal space

$$\mathbf{J}_n = -\frac{q}{4\pi^3}\int_{V_k} \mathbf{v} \cdot f d\mathrm{k}. \tag{5.34}$$

Substitute Equation (5.33) into Equation (5.34), we obtain

$$\mathbf{J}_n = -\frac{q}{4\pi^3}\int_{V_k} \mathbf{v} \cdot \left[f_0 - \tau f_0(1 - f_0)\frac{\mathbf{v}}{k_b T} \cdot \nabla E_F\right] d\mathrm{k}. \tag{5.35}$$

Since f_0 has even symmetry in k-space, while $\mathbf{v}$ has odd symmetry, we have

$$\int_{V_k} \mathbf{v} \cdot f_0 dk = 0, \tag{5.36}$$

and hence

$$\mathbf{J}_n = -\frac{q}{4\pi^3}\int_{V_k} -\mathbf{v} \cdot \tau f_0 \left(1 - f_0\right) \frac{\mathbf{v}}{k_b T} \cdot \nabla E_F dk. \tag{5.37}$$

Since

$$E_F = -q\phi_n \tag{5.38}$$

where ϕ_n is the electron Fermi potential, we have

$$n = \frac{1}{4\pi^3}\int_{V_k} f_0 dk; \tag{5.39}$$

$$\langle\tau\rangle = \frac{m^*}{k_b T}\frac{\int_{V_k} \mathbf{v} \cdot \mathbf{v}\tau f_0 \left(1 - f_0\right) dk}{\int_{V_k} f_0 dk}, \tag{5.40}$$

we obtain the final expression for electron current density $\mathbf{J}_n$

$$\mathbf{J}_n = -q^2 n\nabla\phi_n \frac{\langle\tau\rangle}{m^*} = -q\mu_n n\nabla\phi_n, \tag{5.41}$$

where electron mobility μ_n is defined as

$$\mu_n = q\frac{\langle\tau\rangle}{m^*}. \tag{5.42}$$

Similarly we have for hole current density

$$\mathbf{J}_p = -q\mu_p p\nabla\phi_p. \tag{5.43}$$

It is seen from equation Equation (5.41) and Equation (5.43) that if carrier concentration and mobility are treated as constants, current density is directly proportional to the gradient of Fermi potential. In order to write the current density as the sum of drift current and diffusion current, which requires knowledge on the gradient of carrier density, one need an explicit expression of carrier density. In general, Fermi-Dirac distribution should be use for this purpose, but we shall use the mathematically simpler Boltzmann distribution here, with the help of effective intrinsic carrier concentration n_{ie},

$$\begin{aligned} n &= n_{ie} \exp\left[\frac{q}{k_b T}\left(\psi - \phi_n\right)\right] \\ p &= n_{ie} \exp\left[\frac{q}{k_b T}\left(\phi_p - \psi\right)\right], \end{aligned} \tag{5.44}$$

we find the Fermi potentials

$$\begin{aligned} \phi_n &= \psi - \frac{k_b T}{q} \ln\left(\frac{n}{n_{ie}}\right) \\ \phi_p &= \psi + \frac{k_b T}{q} \ln\left(\frac{p}{n_{ie}}\right). \end{aligned} \tag{5.45}$$

Substituting into Equation (5.41) and Equation (5.43) we have

$$\begin{aligned} \mathbf{J}_n &= q\mu_n n\mathbf{E} + qD_n \nabla n - q\mu_n n\left(\frac{k_b T}{q}\nabla \ln n_{ie}\right) \\ \mathbf{J}_p &= q\mu_p p\mathbf{E} - qD_p \nabla p + q\mu_p p\left(\frac{k_b T}{q}\nabla \ln n_{ie}\right), \end{aligned} \tag{5.46}$$

where $\mathbf{E} = -\nabla\psi$, and diffusivity D is related to mobility μ according to the Einstein relation $D_n = \frac{k_b T}{q}\mu_n$ and $D_p = \frac{k_b T}{q}\mu_p$. The last term in each of the above two equations represents current due to the gradient in intrinsic carrier concentration. In a homogeneous semiconductor device with uniform temperature, $\nabla \ln n_{ie} = 0$, and the last term vanishes. This leads to the familiar expressions of drift-diffusion current density

$$\begin{aligned} \mathbf{J}_n &= q\mu_n n\mathbf{E} + qD_n \nabla n \\ \mathbf{J}_p &= q\mu_p p\mathbf{E} - qD_p \nabla p \end{aligned} \tag{5.47}$$

We shall summarize the limitations of the drift-diffusion model. The drift-diffusion model is derived from the Boltzmann transport equation, but uses the equilibrium carrier distribution function to approximate the actual distribution function. This is a reasonable approximation at low field, but breaks down at high electric field. Carriers accelerates under high electric field, and scattering is not sufficiently strong to bring carrier temperature back to the lattice temperature. As a result, these hot carriers have carrier temperature higher than the lattice temperature used in Equation (5.27). Additionally, the distribution function becomes highly asymmetric if the high-field region is very short (comparable to carrier mean-free path). This asymmetry is of course not considered in the drift-diffusion model either. Therefore, the drift-diffusion is only applicable when the electric field is not too high, and the device dimension is not too small.

5.4 Hydrodynamic Model

We take the Boltzmann transport equation, multiply on the both sides $\Phi_1 = 1$, $\Phi_2 = \hbar k$ and $\Phi_3 = \hbar^2 k^2 / 2m^*$, then integrate each equation throughout k-space.

The resulting equations essentially dictates the carrier continuity , conservation of momentum, and conservation of energy, respectively.

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\mathbf{u}) = \left(\frac{\partial n}{\partial t}\right)_{coll} \tag{5.48}$$

$$\frac{\partial m_n^* n\mathbf{u}}{\partial t} + \nabla \cdot (m_n^* n\mathbf{u}\mathbf{u} + nkT_n) = -en\mathbf{E} + \left(\frac{\partial m_n^* n\mathbf{u}}{\partial t}\right)_{coll} \tag{5.49}$$

$$\frac{\partial n\omega_n}{\partial t} + \nabla \cdot (n\omega_n\mathbf{u} + nkT_n\mathbf{u}) = -en\mathbf{E} \cdot \mathbf{u} - \nabla(-\kappa_n \nabla T_n) + \left(\frac{\partial n\omega_n}{\partial t}\right)_{coll}. \tag{5.50}$$

The independent variables are n is electron density, $\mathbf{u}$ the average electron velocity and T_n the electron temperature. Additionally we have the internal energy of electron gas $\omega_n = \frac{3}{2}kT_n + \frac{1}{2}m_n^* u^2$, and the heat flux $-\kappa_n \nabla T_n$ where κ_n is electron thermal conductivity.

Similarly we have the hydrodynamic equations for holes

$$\frac{\partial p}{\partial t} + \nabla \cdot (p\mathbf{v}) = \left(\frac{\partial p}{\partial t}\right)_{coll} \tag{5.51}$$

$$\frac{\partial m_p^* p\mathbf{v}}{\partial t} + \nabla \cdot (m_p^* p\mathbf{v}\mathbf{v} + pkT_p) = ep\mathbf{E} + \left(\frac{\partial m_p^* p\mathbf{v}}{\partial t}\right)_{coll}, \tag{5.52}$$

$$\frac{\partial p\omega_p}{\partial t} + \nabla \cdot (p\omega_p\mathbf{v} + pkT_p\mathbf{v}) = ep\mathbf{E} \cdot \mathbf{v} - \nabla(-\kappa_p \nabla T_p) + \left(\frac{\partial p\omega_p}{\partial t}\right)_{coll} \tag{5.53}$$

where p is the hole density, $\mathbf{v}$ the average hole velocity, T_p the hole temperature, $\omega_p = \frac{3}{2}kT_p + \frac{1}{2}m_p^* v^2$ the hole internal energy, κ_p the hole thermal conductivity.

In Equation (5.48) and Equation (5.51), the collision terms can be expressed in terms of carrier generation and recombination,

$$\left(\frac{\partial n}{\partial t}\right)_{coll} = \left(\frac{\partial p}{\partial t}\right)_{coll} = G - R. \tag{5.54}$$

According to Baccarani and Wordeman, the collision term in the momentum and energy conservation equations takes the following forms with the relaxation time approximation,

$$\left(\frac{\partial m_n^* n\mathbf{u}}{\partial t}\right)_{coll} = -\frac{m_n^* n\mathbf{u}}{\tau_p^n} \qquad \tau_p^n = m_n^* \frac{\mu_{n0}}{e}\frac{T_n}{T_0} \tag{5.55}$$

$$\left(\frac{\partial m_p^* p\mathbf{v}}{\partial t}\right)_{coll} = -\frac{m_p^* p\mathbf{v}}{\tau_p^p} \qquad \tau_p^p = m_p^* \frac{\mu_{p0}}{e}\frac{T_p}{T_0} \tag{5.56}$$

$$\left(\frac{\partial n\omega_n}{\partial t}\right)_{coll} = -\frac{n\omega_n - \frac{3}{2}nkT_0}{\tau_\omega^n} \quad \tau_\omega^n = \frac{m_n^*}{2}\frac{\mu_{n0}}{e}\frac{T_0}{T_n} + \frac{3}{2}\frac{\mu_{n0}}{ev_{ns}^2}\frac{T_nT_0}{T_n + T_0} \tag{5.57}$$

$$\left(\frac{\partial p\omega_p}{\partial t}\right)_{coll} = -\frac{p\omega_p - \frac{3}{2}pkT_0}{\tau_\omega^p} \quad \tau_\omega^p = \frac{m_p^*}{2}\frac{\mu_{p0}}{e}\frac{T_0}{T_p} + \frac{3}{2}\frac{\mu_{p0}}{ev_{ps}^2}\frac{T_pT_0}{T_p + T_0} \tag{5.58}$$

where T_0 is the lattice temperature, μ_{n0} and μ_{p0} the electron and hole mobility, v_{ns} and v_{ps} the saturate velocity of carriers. This model accounts for the phonon scattering and ionized impurity scattering.

Starting from the hydrodynamic model, we can arrive at the drift-diffusion model with some further simplifications. If the first two terms in the electron momentum conservation equation are dropped, one obtains

$$\mathbf{J}_n = -en\mathbf{u} = \frac{e^2}{m_n^*}\tau_p^n n\mathbf{E} + \frac{e}{m_n^*}kT_L\tau_p^n\nabla n. \tag{5.59}$$

Similarly, for holes, one has

$$\mathbf{J}_p = \frac{e^2}{m_p^*}\tau_p^p p\mathbf{E} - \frac{e}{m_p^*}kT_L\tau_p^p\nabla p. \tag{5.60}$$

If one defines the carrier mobility as

$$\mu_n = \tau_p^n\frac{e}{m_n^*} \tag{5.61}$$

$$\mu_p = \tau_p^p\frac{e}{m_p^*} \tag{5.62}$$

we see the familiar formulae of drift-diffusion current density

$$\mathbf{J}_n = e\mu_n n\mathbf{E} + e\mu_n(\frac{kT_L}{e})\nabla n \tag{5.63}$$

$$\mathbf{J}_p = e\mu_p p\mathbf{E} - e\mu_p(\frac{kT_L}{e})\nabla p. \tag{5.64}$$

5.5 Carrier Mobility

Carrier mobility has pivotal importance in the study of carrier transport in semiconductors. From the discussion in the two proceeding sections, it is obvious that carrier mobility is closely related to the effective mass and the scattering rate of carriers. Higher effective mass or higher scattering rate would lead to lower carrier mobility. The carrier effective mass has been discussed in "Energy Band Structure Of Semiconductor", on page 12. In this section, we shall discuss a few dominant scattering processes in semiconductor devices, and their contribution to the carrier mobility.

Before we enumerate the various scattering mechanisms, we shall first consider the problem of how these scattering mechanisms combine to give the total scattering rate, and hence the combined carrier mobility. Assuming the scattering mechanisms are independent of each other, the total scattering rate is simply the sum of that of the individual scattering processes. However, from Equation (5.40) and Equation (5.42), one sees that the calculation of mobility requires knowledge of the carrier distribution function. The exact shape of the distribution function is not available in either the drift-diffusion model or the hydrodynamic model, which makes the rigorous calculation of carrier mobility impossible. In practice, we consider the carrier mobility in the near equilibrium situation. The mobility due to each individual scattering mechanism is first calculated from the respective scattering rates, and the total mobility is calculated using Matthiessen's rule

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \ldots \tag{5.65}$$

where μ is the total mobility, while μ_1, μ_2 etc. are the mobility due to individual scattering mechanisms.

Phonon scattering is one of the most dominant scattering mechanism in semiconductor devices. The band structure derived in "Energy Band Structure Of Semiconductor", on page 12 assumes that atoms in the crystal are all stationary, in which case carriers can propagate with Bloch wave function without scattering.

However, the thermal motion of atoms cause them vibrate around the equilibrium position on the lattice, which alters the crystal potential and hence the band structure. Intuitively, one can think of electrons travelling along a bumpy conduction band, with humps and dips induced by slight displacement of atoms in the lattice. The lattice vibration propagate in the lattice as phonon. As temperature increases, the number of phonon increases (stronger lattice vibration), and the phonon scattering rate increases. The approximate relationship between the phonon limited mobility in bulk silicon and the temperature is

$$\mu_{phonon} \propto T^{-3/2} \tag{5.66}$$

Ionized impurity scattering is another important scattering mechanism in semiconductors. The ionized dopants are charged and deflect carriers that pass by. The scattering rate depends on not only the density of ionized impurities, but the carrier energy as well. With higher concentration of ionized impurities, the scattering rate is higher. On the other hand, electrons with higher kinetic energy pass by the ion more quickly, and is hence less probable to be scattered. Since the average kinetic energy of electrons is directly related to the carrier temperature, the ionized impurity scattering strongly depends on temperature. With the increase in temperature, scattering reduces, and the ionized impurity limited mobility increases as

$$\mu_{coulomb} \propto T^{3/2} \tag{5.67}$$

As one lowers the temperature, the phonon scattering becomes weaker, and the ionized impurity scattering becomes dominant. This temperature dependence is especially pronounced in highly doped semiconductors. Figure (5.2) shows the dependence of electron mobility in bulk silicon, when both phonon scattering and ionized impurity scattering are considered.

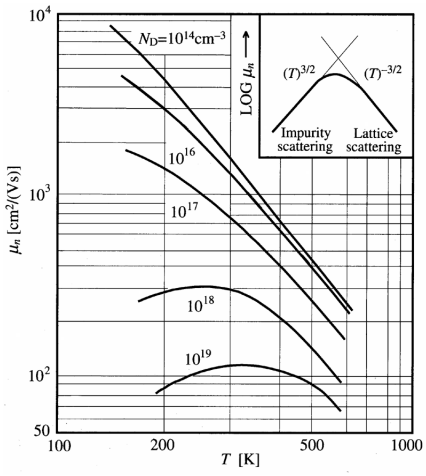


Figure 5.2: Temperature dependence of electron mobility of silicon.

In practice, empirical models are developed to describe the carrier mobility at various doping concentration, temperature, and electric field. Contribution from various scattering mechanisms are all incorporated in these unified models [3].

Apart from the two dominant mechanisms described above, there are a few other scattering processes that are important in some specific situations.

Electron-hole scattering is important in power electronics devices, because these devices often operate in high-level injection where carrier concentration exceeds

the doping concentration. Advanced mobility models such as the Philips model includes the contribution of electron-hole scattering [4][5].

At very low temperature dopants are not completely ionized, and those not ionized are electrically neutral. These neutral impurities are perturbation to the periodic crystal potential, and are scatterers to carriers as well. The effect of scattering from these neutral defects is observable only when phonon scattering and ionized impurity scattering are both very weak.

In MOSFET transistors, most current flows through the inversion layer at the semiconductor/insulator interface. The inversion carrier mobility is significantly different from that in the bulk semiconductor, and is important to surface-type devices such as the MOSFETs. Since carriers are confined in the very thin inversion layer, quantum mechanical effects is strong, which changes the density of states and hence the carrier mobility. When one increase the electric field perpendicular to the interface, e.g. by increasing the gate voltage, the inversion carriers are confined in a narrower inversion layer. As a result, the density of states in the inversion layer increases; carriers see more states to scatter to, and carrier mobility decreases. Another interface related scattering mechanism is the surface roughness scattering. With increasing perpendicular electric field, the carriers are confined closer to the interface, which leads to higher surface roughness scattering rate. Separate mobility model are developed to include the above effects related to the perpendicular E-field near the semiconductor/insulator interface.

Generally, when electric field is low, carrier drift velocity is directly proportional to the driving electric field, where the carrier mobility serves as the linear coefficient. However, under high electric field (parallel to the direction of electron transport), this linear relationship breaks down. Carrier mobility decreases with increasing field, while beyond a critical field, carrier velocity saturates to a constant value. Carrier mobility must therefore be adjusted to account for this velocity saturation effect under high parallel electric field. Figure (5.3) and Figure (5.4) show the relationship between carrier velocity and driving electric field for silicon and GaAs. Note that the carrier velocity in GaAs reaches a maximum at E-field of 4 $\mathrm{kV} \cdot \mathrm{cm}^{-1}$, beyond which velocity declines. This leads to negative differential resistance in GaAs, which is useful in generating microwave oscillations.

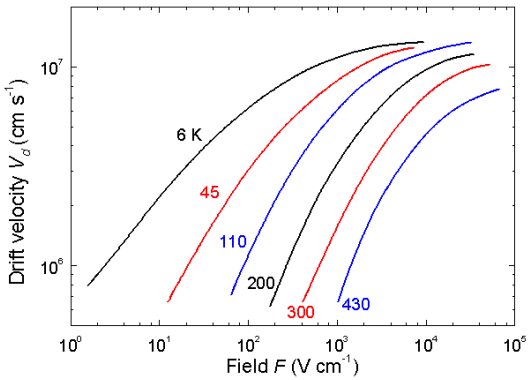


Figure 5.3: Silicon drift speed and electric field relationship

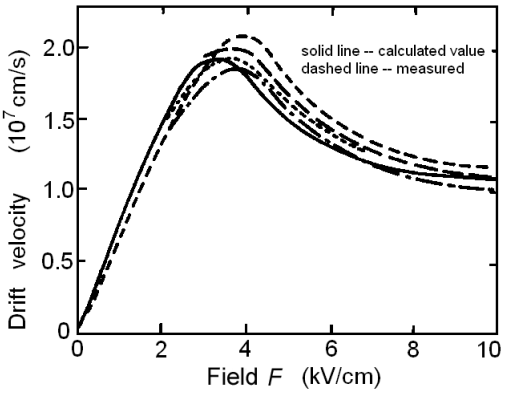


Figure 5.4: GaAs drift speed and electric field relationship

5.6 Constants in Semiconductors

In this section, we shall discuss two important constants in semiconductors, namely the Debye length and the dielectric relaxation time.

Debye length was first used in plasma physics to characterize the range in which a charge perturbation has effects on other particles. Debye length in semiconductor physics carries a similar significance. Consider a large piece of semiconductor with n-type doping, the potential in the sample is governed by the Poisson's equation

$$\varepsilon \frac{\mathrm{d}^2 \psi}{\mathrm{d}x^2} = \frac{\varepsilon}{q} \frac{\mathrm{d}^2 E_c}{\mathrm{d}x^2} = -q\left(N_D - n\right). \tag{5.68}$$

We used the conduction band minimum as the reference potential, while holes are ignore due to the low concentration

$$E_c\left(x\right) = E_c\left(\psi = 0\right) + q\psi\left(x\right). \tag{5.69}$$

Since

$$n = N_c \cdot \exp\left(-\frac{E_c - E_F}{k_b T}\right), \tag{5.70}$$

the Poisson's equation can be written as

$$\frac{\varepsilon}{q} \frac{\mathrm{d}^2 E_c\left(x\right)}{\mathrm{d}x^2} = -q\left(N_D - N_c \cdot \exp\left(-\frac{E_c - E_F}{k_b T}\right)\right). \tag{5.71}$$

In order to simplify the above equation, we write the conduction band as

$$E_c\left(x\right) = E_c^0 + \Delta E_c\left(x\right), \tag{5.72}$$

where E_c^0 is the conduction band energy at equilibrium, which is a constant, ΔE_c is the band bending due to the disturbance. Noting that we have $N_D = n_0$ at equilibrium,

$$\begin{aligned} N_c \cdot \exp\left(-\frac{E_c - E_F}{k_b T}\right) &= N_c \cdot \exp\left(-\frac{E_c^0 - E_F}{k_b T}\right) \cdot \exp\left(-\frac{\Delta E_c\left(x\right)}{k_b T}\right) \\ &= N_D \cdot \exp\left(-\frac{\Delta E_c\left(x\right)}{k_b T}\right) \end{aligned} \tag{5.73}$$

The Poisson's equation is now simplified to

$$\frac{\mathrm{d}^2 \Delta \mathrm{E_c}}{\mathrm{dx}^2} = -\frac{q^2 N_D}{\varepsilon} \cdot \left[1 - \exp\left(-\frac{\Delta E_c\left(x\right)}{k_b T}\right)\right]. \tag{5.74}$$

When we are not too far from equilibrium, we have $\Delta E_c \ll k_b T$, and the above equation can be linearized as

$$\frac{\mathrm{d}^2 \Delta E_c}{\mathrm{d}x^2} = \frac{q^2 N_D}{\varepsilon} \cdot \frac{\Delta E_c(x)}{k_b T} \tag{5.75}$$

Analytical solution is now possible, where Debye length is defined as the characteristic length

$$L_{Db} = \sqrt{\frac{\varepsilon k_b T}{q^2 N_D}}. \tag{5.76}$$

The change in conduction band is therefore

$$\Delta E_c(x) = \Delta E_c(x=0) \cdot \exp\left(-\frac{x}{L_{Db}}\right). \tag{5.77}$$

One can easily obtain the expression of Debye length for p-type semiconductors

$$L_{Db} = \sqrt{\frac{\varepsilon k_b T}{q^2 N_A}} \tag{5.78}$$

The relationship between Debye length and doping concentration is shown in Figure (5.5). High doping concentration leads to very small Debye length. In the meshing stage of semiconductor simulation, the Debye length is an important parameter. In order to capture the spatial variation of carrier concentration, the dimension of the mesh grids in critical device regions must be less than the Debye length.

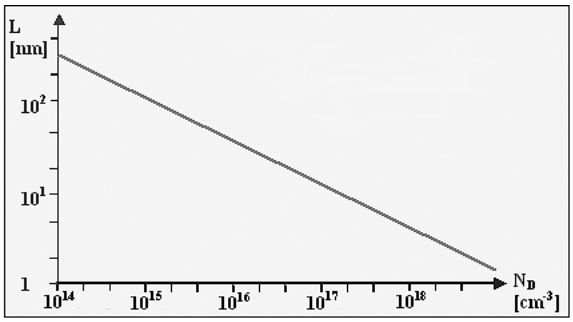


Figure 5.5: Debye length as a function of doping concentration.

While the Debye length determines the length scale in which a charge disturbance is effective, the other scaling constant, namely the dielectric relaxation time, determine the time scale in which majority carriers can respond to a charge perturbation. In order to obtain the dielectric relaxation time, we start again from the electrostatic equation.

$$\frac{dE(x,t)}{dx} = -q\frac{\Delta\rho(x,t)}{\varepsilon}, \tag{5.79}$$

where $\Delta\rho$ is the change in carrier concentration, and ΔE is the resulted change in electric field. We then turn to the continuity equation, and note that in the low field, near-equilibrium case, carrier generation-recombination can be ignored, and the incremental carrier $\Delta\rho$ contributes little current. The continuity equation thus contains only the drift current term

$$\frac{\partial \Delta\rho}{\partial t} = -q\mu\rho\frac{\partial E(x)}{\partial x}. \tag{5.80}$$

Combining the continuity equation and the electrostatic equation, we have

$$\frac{\partial \Delta\rho}{\partial t} = -\frac{q\mu\rho}{\varepsilon} \cdot \Delta\rho. \tag{5.81}$$

Solving this equation we obtain

$$\Delta\rho\,(x,t) = \Delta\rho\,(x,0) \cdot \exp\left(-\frac{t}{\tau_d}\right) \tag{5.82}$$

where the dielectric relaxation time is defined as

$$\tau_d = \frac{\varepsilon}{q\mu\rho_0}. \tag{5.83}$$

Assuming $\rho_0 = N_D$ for n-type doped semiconductor, and $\rho_0 = N_A$ for p-type doped semiconductor, we can relate the Debye length, the Einstein relation and the dielectric relaxation time with

$$\tau_d = \frac{L_{Db}^2}{D}. \tag{5.84}$$

In the numerical simulation of semiconductor devices, the dielectric relaxation time is an upper bound to the time steps. Discretizing Equation (5.81) yields

$$\Delta\rho\,(\Delta t) = \Delta\rho\,(0) - \frac{\Delta t}{\tau_d}\Delta\rho\,(0)\,, \tag{5.85}$$

where Δt is the time step, and $\Delta\rho$ is the resulting update to the carrier concentration. Obviously, if $\Delta t > \tau_d$, $\Delta\rho$ will change sign at every step, and the numerical solution will oscillate with amplitude $\Delta\rho$, which is unphysical. For semiconductors with high carrier mobility, τ_d is small. For example GaAs has $\mu = 6000\text{cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$, so we have $\tau_d \approx 10^{-15}$s when doping level is 10^{18}cm^{-3}. The time steps can not exceed this value in numerical simulation if explicit time-discretization methods. Such short time steps is intolerable in practice, therefore implicit discretizations, which offers guaranteed numerical stability, must be used.

Chapter 6 Semiconductor Contact Interface

In semiconductor manufacturing, we can not avoid semiconductor contact with other material. For example by using vacuum evaporation, sputter deposition and other method's semiconductor surface to form metal film and leading to semiconductor metal contact or semiconductor surface SiO_2 film, leading to semiconductor insulator interface. Semiconductor numerical software has to be able to deal with these kinds of interface.

6.1 Semiconductor and Metal Contact

Metal and semiconductor's contact has two type. Super low contact resistance contact and similar as PN junction's single direction Schottky contact. In semiconductor device and integrate circuit manufacturing, these two type of contacts are generally used.

6.1.1 Semiconductor and metal contact potential barrier

First let us explain the work function concept. The valance band electron inside metal needs to have very high energy in order to jump out the metal, in another word, the outside system has to do some work in order to bring out the electron. This work's average value is electron's overflow work, or work function. Metal's electron work function Φ_M is defined as

$$\Phi_M = E_0 - E_F \tag{6.1}$$

E_0 is metal surface vacuum static electron's energy, E_F is electron's Fermi energy level. Obviously work function represents the bundling ability of metal. For semiconduction, electron work function Φ_S is defined as

$$\Phi_S = E_0 - E_F \tag{6.2}$$

Because E_F changes with the variation of doping concentration level, Φ_S is related to doping concentration. Besides, semiconductor electron affinity is defined as

$$\chi = E_0 - E_c \tag{6.3}$$

Because metal and semiconductor's work function are different, there will be contact potential difference when they are contacted. Simultaneously there is space charge at one side of the semiconductor, the corresponding energy band starts to bend.

Semiconductor	χ(eV)	Φ_S(eV)					
		N-Type $N_D$cm^{-3}			P-Type $N_A$cm^{-3}		
		10^{14}	10^{16}	10^{18}	10^{14}	10^{16}	10^{18}

Si	4.17	4.494	4.375	4.256	4.951	5.070	5.147
Ge	4.00	4.297	4.180	4.061	4.377	4.495	4.571
GaAs	4.07	4.289	4.170	4.050	5.206	5.325	5.444

Table 6.1: Semiconductor work function: from GSS database

Metal	Φ_M(eV)	Metal	Φ_M(eV)	Metal	Φ_M(eV)
Au	5.47 < 100 >	Al	4.06 < 110 >	Cu	4.48 < 110 >
Pt	5.93 < 111 >	Pb	4.25 polycr	W	4.55 polycr

Table 6.2: Some metal material's work function[6]

Consider metal and n type semiconductor's contact, and set $\Phi_M > \Phi_S$, when they are not contacted, the energy band is shown as Figure (6.1), Because $\Phi_M > \Phi_S$, semiconductor's fermi energy is higher than metal's fermi energy, when they are closely contacted, semiconductor's electron will flow to metal, so that the surface of metal has negative charge, semiconductor surface forms positive space charge layer, producing inner electric field pointing from semiconductor to metal. This field will stop electron's further movement inside metal, so that the whole system reaches stability. Now the metal and semiconductor's Fermi level is at the same level, shown in figure Figure (6.2). Then metal and semiconductor have contact potential difference, although metal has transition layer, due to electron density inside metal are several order bigger than semiconductor, the transition layer is at several atomic distance level, which can be neglected usually. So we can consider the contact potential is all at the semiconductor's space charge layer. From figure Figure (6.2), we know semiconductor's conductor band electron has to pass the potential barrier $qV_{bi} = \Phi_M - \Phi_S$. Besides semiconductor potential barrier is due to energy band bending up, electron density is much less than balanced concentration. Accordingly it is a high resistance region, which is also called electron stop layer. Also the electron in the metal has to overcome potential barrier $\Phi_B = \Phi_M - \chi$ to reach the other side of semiconductor. This electron stopping layer is first studied by Schottky. So it is also called Schottky contact. Φ_B is called Schottky barrier height. Schottky barrier height is not related to doping concentration, it is only related to semiconductor and metal material.

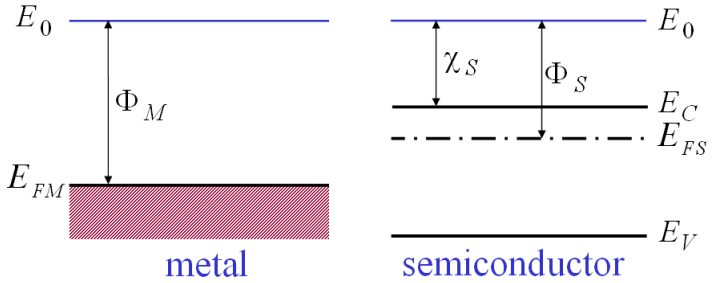


Figure 6.1: Energy level before contact

If n type semiconduction's work function is bigger than metal, $\Phi_S > \Phi_M$, electron will from from metal to semiconductor when they are contacted. There will be negative space charge region in side semiconductor and the space charge region's

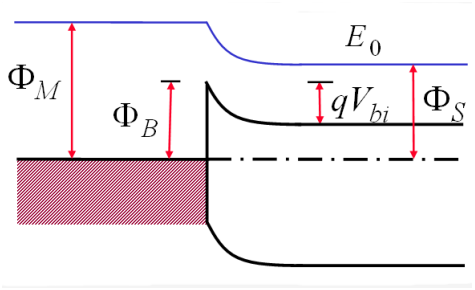


Figure 6.2: Energy level after contact

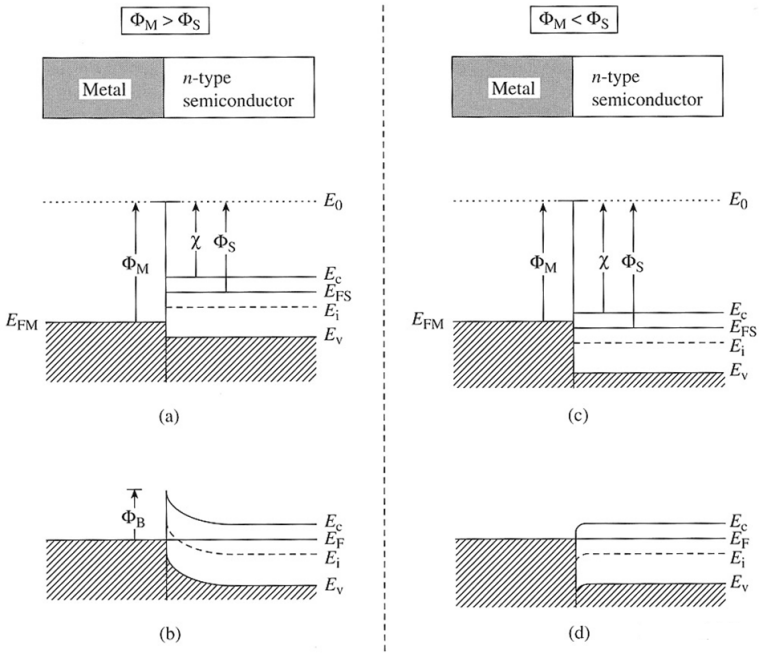


Figure 6.3: Different work function's metal and n type semiconductor's contact

energy band will bend down shown as Figure (6.3). This condition's conduction band electron and metal's electron need not overcome potential barrier to reach the other side, and the semiconductor interface space charge region has much more electron contraction than balanced concentration. This electron deposition region's conductivity is very good, also called anti electron stopping layer.

When metal and p type semiconductor contact, the condition forming stopping layer and anti stopping layer is just reverse to n type semiconductor. When $\Phi_S > \Phi_M$, hole will flow to metal and let metal surface be positively charged, semiconductor surface has negative space charge layer, energy band bends down, shown as Figure (6.4). It is hole stopping layer, hole needs to overcome potential barrier $\Phi_S - \Phi_M$ to reach the metal, and holes inside the metal, empty state with Fermi level as E_{FM}, needs to overcome $\Phi_B = \chi + E_g - \Phi_M$ to reach semiconductor. Here hole exchange means electrons inside metal exchange with semiconductor valance band electrons. Electrons from metal to semiconductor's conduction band top needs to increase energy $\Phi_S - \Phi_M$, and the electron from semiconductor needs to increase energy $\Phi_B = \chi + E_g - \Phi_M$ to reach E_{FM}. If metal contact p type semiconductor with $\Phi_S < \Phi_M$, similarly we know the energy band bend up to form hole anti stopping layer.

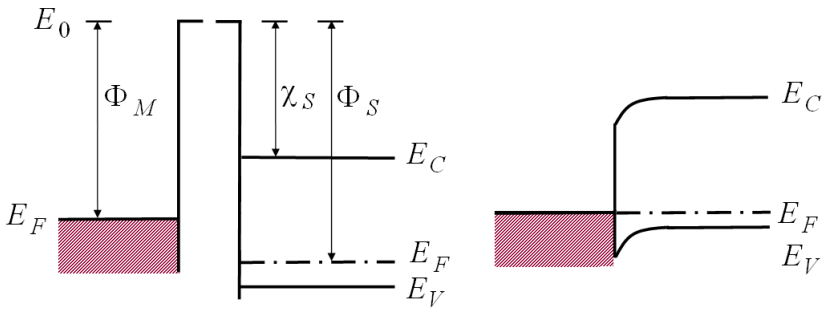


Figure 6.4: Metal and p type semiconductor's contact$\Phi_S > \Phi_M$

6.1.2 Schottky contact's current relationship

Experiment proves that metal semiconductor contact formed Schottky potential is similar as pn junction's rectifying characteristics. Figure (6.5) gives a metal and n type semiconductor formed ideal Schottky barrier. Assume electron density inside the barrier is 0, we can obtain potential's width and inner electric field value from potential height and semiconductor doping concentration.

$$\frac{d^2\phi}{dx^2} = -\frac{qN_D}{\varepsilon_s\varepsilon_0} \tag{6.4}$$

Consider Schottky interface at $x = 0$, boundary condition is $\phi(0) = 0$, $\phi(W) = V_{bi}$ and $E(W) = -\nabla\phi(W) = 0$, W is Schottky barrier width, shown in Figure (6.5). Solve the above partial differential equation, we can have potential barrier width

$$W = \sqrt{\frac{2\varepsilon_s\varepsilon_0 V_{bi}}{qN_D}} \tag{6.5}$$

Potential barrier's electric potential distribution

$$\phi(x) = V_{bi} - \frac{qN_D}{2\varepsilon_s\varepsilon_0}(W - x)^2 \tag{6.6}$$

But in application, normally we use triangle to approximate potential barrier, then

$$\phi(x) = V_{bi}\left(1 - \frac{x}{W}\right) \tag{6.7}$$

Figure (6.6) shows metal and n type semiconductor formed Schottky contact electron and hole's four different moving method: (A)represents net electron flow overcomes potential barrier, important for Schottky junction's IV characteristics; (B)represents electron tunnels through quantum mechanic effect to go through the barrier, normally it is considered together with Schottky potential barrier. Used to correct (A)'s current. (C)represents electron's recombination, discussed previously. (D)represents hole goes through metal overcome barrier $E_g - \Phi_M$ into semiconductor's process, called minority carrier injection. It follows same injection rule with (A), and normally leads to very small current. For (A)'s process there are two theories, thermionic electron injection theory and diffusion theory. We are going to introduce simply, the detail explanation can be found in [7].

Thermionic electron injection theory is proposed by Bethe. For high mobility carriers, the average mean free path is big and potential barrier is thin, electron's collision inside the barrier is neglected, then whether electron can go through the barrier is limited by barrier height. When electron's kinetic energy is higher than

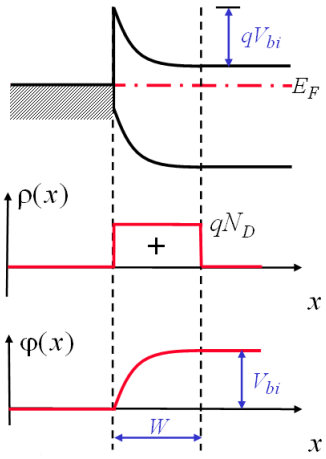


Figure 6.5: Ideal Schottky contact's potential distribution

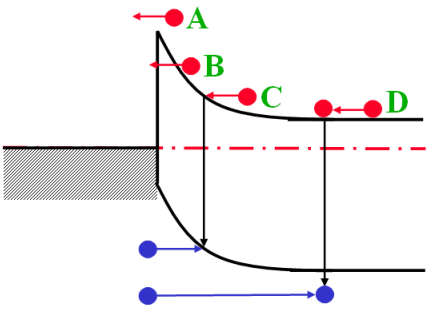


Figure 6.6: n type Schottky contact's current characteristics

potential barrier, electron can freely go through the barrier to the other side. Consider with voltage bias V_A's condition, potential is used to increase E_F, the electron current from semiconductor to metal is composed of electrons with energy high than $E_F + \Phi_B$'s directional movement. After simply leading we have:

$$\begin{aligned} J_{S\to M} &= q\int_{E_F+\Phi_B}^{\infty} v_x dn = \frac{4\pi m_n^* q k_b^2}{h^3} T^2 e^{-\Phi_B/k_bT} e^{qV_A/k_bT} \\ &= A^* T^2 e^{-\Phi_B/k_bT} e^{qV_A/k_bT} \end{aligned} \tag{6.8}$$

However the electron current from metal to semiconductor keeps constant, equal to electron current from semiconduction to metal at $V_A = 0$.

$$J_{M\to S} = J_{S\to M}\Big|_{V_A=0} = A^* T^2 e^{-\Phi_B/k_bT} \tag{6.9}$$

So the total current is

$$J_{tot} = J_{S\to M} - J_{M\to S} = A^* T^2 e^{-\Phi_B/k_bT}\left(e^{qV_A/k_bT} - 1\right) \tag{6.10}$$

A^* is effective Richardson coefficient,

$$A^* = 120(m_n^*/m_0)\ \mathrm{A/cm^2 \cdot K^2} \tag{6.11}$$

Because parameter involves electron's effective mass at the interface m_n^*, so it is related to material's band structure, most of the cases, A^* has to be decided by experiments.

Diffusion theory is proposed by Schottky, consider current is limited by semiconductor region's drift diffusion movement.

$$J = q\mu_n n\frac{dE_F}{dx} = \mu_n N_c k_b T e^{-E_c/k_b T}\frac{d}{dx}\left(e^{E_F/k_b T}\right) \tag{6.12}$$

Express space charge region's potential with triangle potential, solve the partial differential equation above, we can have the current relationship:

$$J = qN_c\mu_n F_{\max} e^{-\Phi_B/k_b T}\left(e^{qV_A/k_b T} - 1\right) \tag{6.13}$$

$F_{\max}$ is potential barrier's electric field density

$$F_{\max} = \sqrt{\frac{2qN_D\left(V_{bi} - V_A\right)}{\varepsilon_s\varepsilon_0}} \tag{6.14}$$

Then we have the current expression as

$$J = qN_c\mu_n\sqrt{\frac{2qN_D\left(V_{bi} - V_A\right)}{\varepsilon_s\varepsilon_0}}\, e^{-\Phi_B/k_b T}\left(e^{qV_A/k_b T} - 1\right) \tag{6.15}$$

In diffusion theory, current increase's limitation is not how electron overcome the potential barrier, but the semiconductor's mobility. So it is suitable for low mobility semiconductor material to form Schottky contact.

In fact most of the semiconductor material as Si, Ge, GaAs mobility are all high, thermionic electron injection theory is suitable. Only some low mobility material for example Cu_2O, amorphous silicon and CVD's CdS poly film and etc, diffusion theory are suitable.

Figure (6.7) shows the Schottky barrier's variation with different bias. Forward bias lowers potential height, more electrons move from semiconductor to metal, shown in Figure (6.7)'s (b); reverse bias increase the potential height, electron will move from semiconductor to metal more difficultly. shown in Figure (6.7)'s (c). However metal to semiconductor's thermionic electron injection is stable, formation is shown in Figure (6.7)'s (d)'s reverse current.

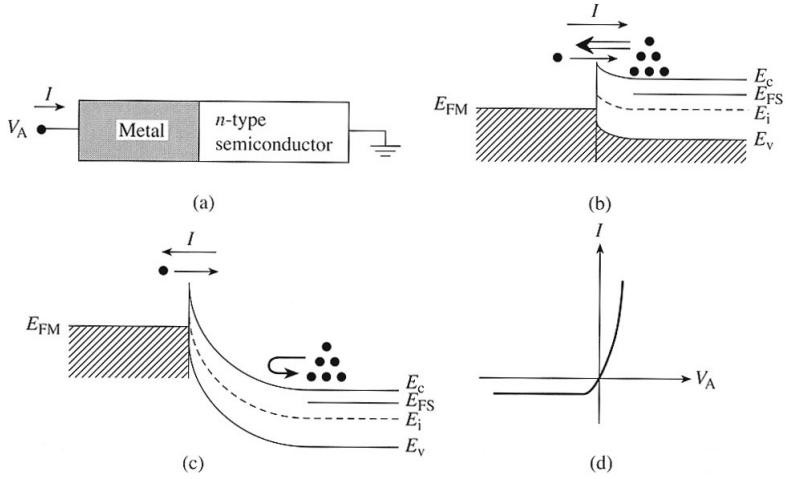


Figure 6.7: Schotty diode's forward and reverse IV characteristics

In real application,we always find Schottky junction's IV characteristic is different from Figure (6.7)'s ideal model, it is because the barrier height decreases as the bias increases and the tunneling effects influence.

Mirror force is the main factor leading to Schottky barrier decreases. A electron outside metal will induct a same value positive charge, the inducted charge's attraction to this electron is called mirror force. Now electron is not only affected by potential barrier but also the mirror force, shown in Figure (6.8) When elec-

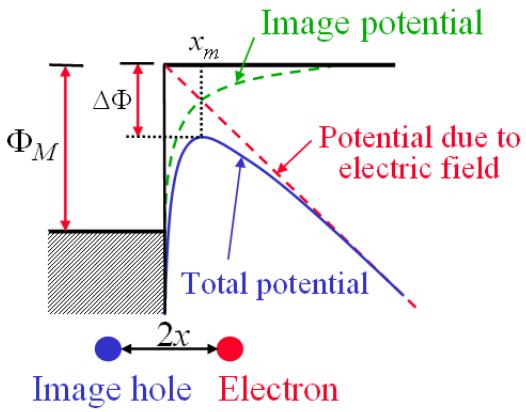


Figure 6.8: Mirror force leading potential barrier decrease

tron is x distance from potential barrier, the inducted charge at $-x$ position, the electron's mirror force is

$$f_{image}(x) = -\frac{q^2}{4\pi\varepsilon_s\varepsilon_0 (2x)^2} = -\frac{q^2}{16\pi\varepsilon_s\varepsilon_0 x^2} \tag{6.16}$$

Corresponding mirror potential

$$\phi_{image}(x) = -\frac{q}{16\pi\varepsilon_s\varepsilon_0 x} \tag{6.17}$$

The electron potential on electron is the sum of mirror potential and electric field potential. Assume triangle potential barrier, the electric field E is a constant, we have

$$\phi(x) = -\frac{q}{16\pi\varepsilon_s\varepsilon_0 x} - Ex \tag{6.18}$$

Solve the derivative of the formulae above, we have the potential barrier's maximum point.

$$x_m = \sqrt{\frac{q}{16\pi\varepsilon_s\varepsilon_0 E}} \tag{6.19}$$

Potential barrier decrease

$$\Delta\Phi = \phi(x_m) = \sqrt{\frac{qE}{4\pi\varepsilon_s\varepsilon_0}} \tag{6.20}$$

Accordingly when electric field is strong, potential barrier decreases a lot. Reverse voltage is high, the mirror force's effect will be obvious, which leads to reverse current increase.

For quantum mechanical tunneling current, the canonical method can not describe, normally we use empirical formulae to correct. The correction of barrier decrease due to tunneling current is

$$\Delta\Phi = \alpha E^{\gamma} \tag{6.21}$$

α and γ's value can be found in [8].

6.1.3 Ohmic contact

Ohmic contact is the low contact resistance metal semiconductor contact. This contact can have relatively big current and the contact's voltage drop is negligible. The ideal case to form ohmic contact to is to form anti stopping layer, then the contact will not form electron or hole potential barrier. It forms a high conduction region. But because the limitation of surface state and metal material, ti is difficult to make the ohmic contact. Normally we use highly doped regions through implantation and diffusion at the semiconductor surface to form thin Schottky contact with high tunneling current to realize it.

Figure (6.9)(a) illustrates a n type Schottky contact, when potential barrier is enough thin the electron can go through the barrier freely, but the hole's diffusion is limited by potential barrier, which is equal to low surface recombination rate. This contact only let electron go through. If we adopt figure Figure (6.9)(b)'s structure, and doped high recombination rate centers. Recombination center can sustain the carrier concentration balance. This contact electron and hole can both go through. This is because the assumption of surface recombination rate to be infinity. The hole arrived to the surface will be recombined, the lost electron can be balanced by tunneling electrons from metal to semiconductor, which is similar to hole go through the interface.

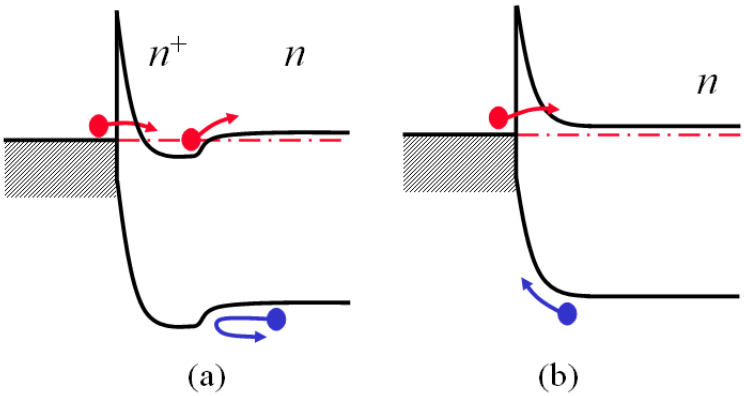


Figure 6.9: With low and high surface recombination rate's ohmic contact

6.2 Metal-Oxide-Semiconductor Structure

Metal-oxide-semiconductor contact (MOS) is CMOS device's basic structure. CMOS devices dominates current integrate circuit's production. Accordingly MOS structure is carefully studied. We only give simple introduction here.

Consider ideal MOS structure shown in Figure (6.10), assume it satisfy the following condition (1) metal and semiconductor's work function difference is 0; (2) the insulator layer does not have any charge absolute insulate. (3) the insulator and semiconductor interface does not have any interface state. Then we have the energy level relationship:

$$\Phi_M = \Phi_S = \chi + E_c - E_F \tag{6.22}$$

We discuss the ideal MOS structure's metal and semiconductor under vertical electric field below, semiconductor surface layer's electric potential and electric charge distribution status.

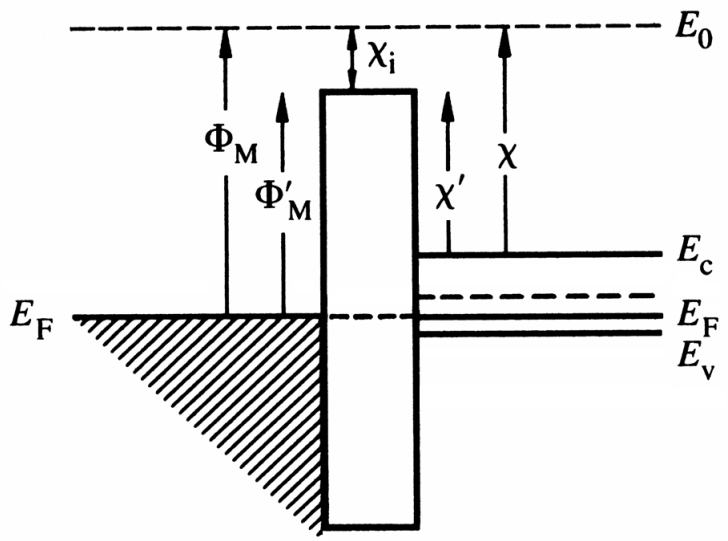


Figure 6.10: Ideal MOS structure

We notice, MOS structure is actually a capacitor device. So when metal and semiconductor is biased, the insulator's each side with metal and semiconductor will be charged. The charge will have reverse sign and the charge distribution status is different. In the metal, the free electron density is very high, charge is basically distributed at one atomic layer's thickness; however the semiconductor's carrier density is much lower, the charge distribution is at a certain thickness layer, which is called space charge region. In space charge region, from the surface to the inner body the electric field decreases. In another word the potential in space charge region will change continuously, so that the semiconductor surface has potential difference to the internal semiconductor body, simultaneously the energy band starts to bend.

Consider p type semiconductor's ideal MOS structure, semiconductor side is grounded, metal is biased with V_G. Energy band variation is

$$E_{F,metal} - E_{F,semiconductor} = -qV_G \tag{6.23}$$

Attention, electron energy is $-qV$, when $V)G < 0$, electron energy increases, surface potential is negative, surface energy band bend up, shown in figure Figure (6.11). At thermal balanced condition, semiconductor fermi energy should keep constant, so band top will turns to Fermi energy level as it is close to surface. simultaneously the valance band's hole concentration increases to form accumulation layer. From the figure we can notice hole's concentration increase when it goes to the surface. It means the hole is mainly distributed close to the interface. When metal and semiconductor is positively biased $V_G >$, surface potential is positive, the surface energy band bends down, shown in Figure (6.12). Then the closer to the surface, the further away the Fermi energy level to the valance band top. The hole concentration decreases accordingly. At close to the surface certain distance region, the top of valance band is much lower than Fermi level. According to Boltzmann distribution, we know the surface hole concentration will be much less than that of the internal body. The surface layer's negative charge is almost equal to acceptor concentration N_A, surface layer's this status is called depletion.

When the positive bias between metal and semiconductor keep increasing, the energy band at the surface will bend more down, shown in Figure (6.13). Then the surface Fermi level can be higher than band gap's middle energy E_i. Then the

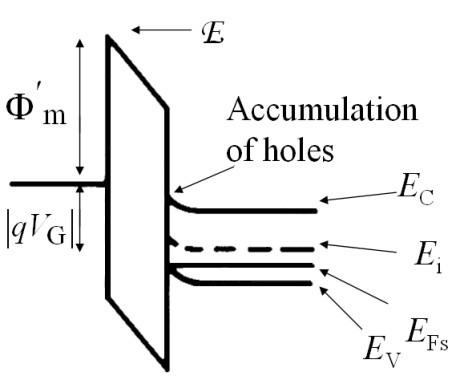


Figure 6.11: PMOS with bias $V_G <$ 0's energy band diagram

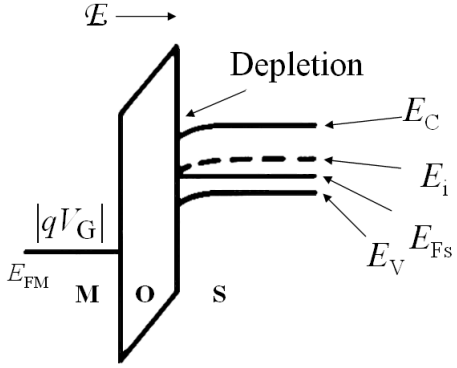


Figure 6.12: PMOS with bias $V_G >$ 0's energy band diagram

Fermi level to conductor band bottom is closer than the distance to valance band, the surface electron concentration will be higher than the hole concentration to form reverse type of doing in the surface, called inverted layer. From Figure (6.13) we notice, inversion layer is very close to the surface. From the inversion layer to the body there is one depletion layer inside. In this circumstance, semiconductor space charge layer's negative charge is composed of two parts, one parts is depletion layer is dominated by acceptor charge N_A, the other part is inverted layer's electrons. The latter one is close to the surface. When surface electron concen-

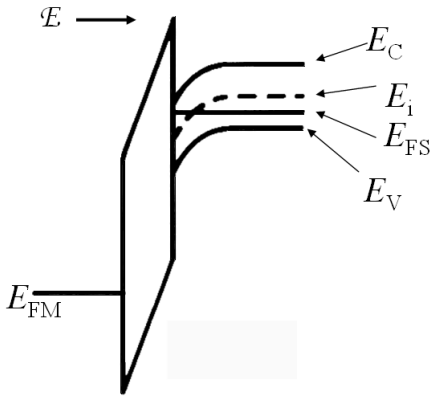


Figure 6.13: PMOS with bias$V_G \gg$ 0's energy band diagram

tration is equal to inner body hole concentration, which satisfy the energy band relationship $E_{i,surface} - E_{i,bulk} = 2[E_F - E_{i,bulk}]$, surface has enough electron concentration to conduct, which is called strong inversion, the corresponding voltage is called threshold voltage V_T.

For n type semiconductor, the accumulation, depletion and inversion layer's voltage is just reverse. When metal and semiconductor is positively biased the surface electrons accumulate; with negative bias between metal and semiconductor, the semiconductor surface deplete, when the absolute voltage value increase, the surface invert to hole dominate region shown in Figure (6.14).

In fact MOS structure still need consider some non ideal factors, mainly included three correctons: generally metal side's work function is different from semiconductor side's work function. Then it is same as ideal MOS structure with additional potential barrier between metal and semiconductor; if consider stable charge in-

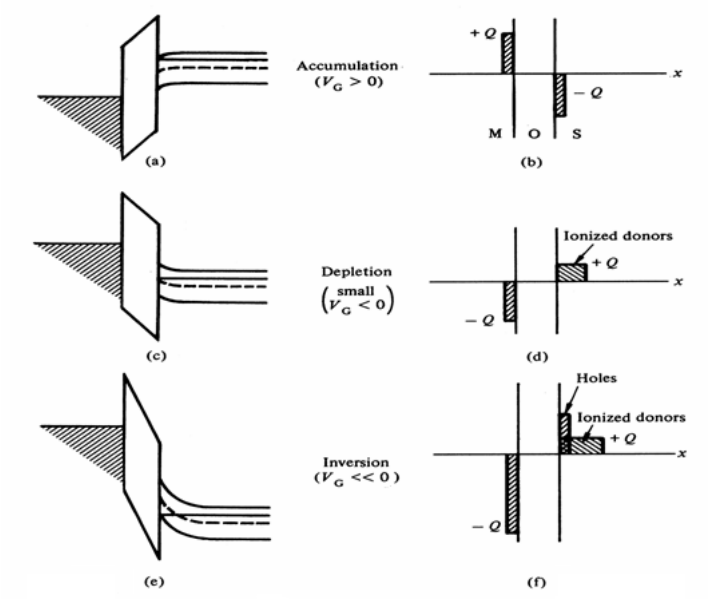


Figure 6.14: NMOS with bias energy band diagram

side the oxide, Poisson equation needs to have stable charge's influence; Silicon and silicon dioxide's interface state needs also to be considered, Poisson equation's interface condition needs to be corrected as additional area electric charge's material interface.

From canonical theory point of view, MOS structure is not complicate. If we don't not consider inversion two dimension electron gas and electron tunneling through oxide process, silicon dioxide interface is a solid wall for carriers. The electric field is totalled described by Poisson equation. Theoretically with certain carriers' distribution function we can have semiconductor side's physical variable's analytical formulae, we are going to discuss some theocratical solution in MOS device's numerical cases.

6.3 Semiconductor hetero-junction

Semiconductor hetero-junction is composed by two different semiconductor materials. Because the hetero-junction is composed of two different semiconductor with different band gap width E_g and other different physical characteristics, so it has certain special performance. Well use and control the performance can make some valuable semiconductor devices. For example hetero bipolar junction transistor (HBJT), high electron mobility transistor (HEMT) and etc. have very big application value in micro wave and high frequency domain.

Hetero-junction both side conductor material type is different is called hetero-type hetero-junction, structure is nP or Np, the capital letter represents the material with wider band gap; when both side of semiconductor have same type of doping, it is called same type hetero-junction nN or pP. In fact hetero-junction contact band structure is similar to semiconductor metal contact. The following Figure (6.15) illustrates the Np type hetero-junction energy band diagram. Its feature is interface has energy band edge split peak. Assume $\Phi_1$and Φ_2, $\chi_1$and χ_2 are N type and p type semiconductor's work function and electron affinity, $\Delta E_c$and ΔE_v are their conductor band bottom energy difference and valance band top energy difference. When both side contact to pn junction, because $\Phi_2 > \Phi_1$, $E_{F1}$is

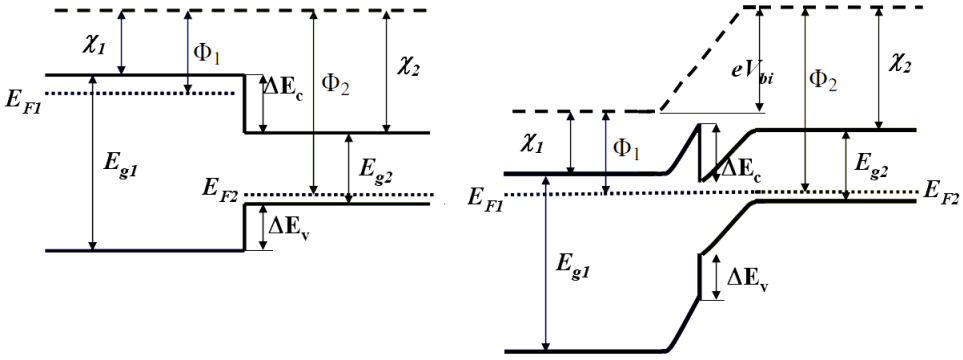


Figure 6.15: Np type hetero-junction energy band

higher than E_{F2}, so electron shifts from n type material to p type material until E_{F1}=E_{F2}. At the contact place p type material side there is a negative charge layer, n type region there is a positive charge area with the same charge amount. In this space charge region, the potential increases from p type to n type, energy band bends. Interface point P type semiconductor valance band bend down, n type semiconductor conduction band bend up. However $\Delta E_c = \chi_1 - \chi_2$ remains the same. There is no spike.

Figure (6.16) shows the thermal stable state's Nn hetero-junction energy band diagram. Same type hetero-junction interface energy band is also not continuous. The different from hetero type hetero-junction is same type heter-junction narrow band gap side's space charge region is electron accumulated layer, the wide band gap region is depleted layer. The hetero type hetero-junction both side's space charge regions are depleted regions.

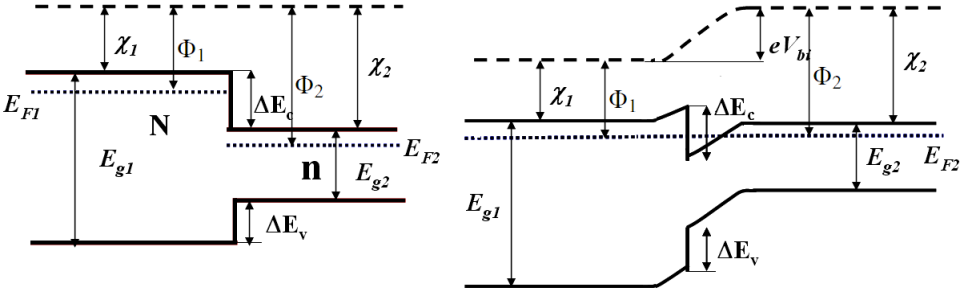


Figure 6.16: Nn type hetero-junction energy band

Hetero-junction's current characteristics is very complicate problem, there are different models for different conditions, for example diffusion model, injection model, injection - recombination model, tunneling - recombination model and etc. However there is not uniform theory to describe hetero-junction's IV characteristics. As limited by canonical theory, GSS only realizes the similar model as Schottky contact's thermionic model, the detail realization is in "Hetero-junction", on page 131.

PART II. Semiconductor Drift Diffusion Model

Reader's Guide

Since Schokley gives the basic semiconductor theory, people are used to use analytical method to analyze the semiconductor devices. Analytical models are based on certain assumptions and analysis, give certain mathematical description on physics and electrical characteristics. Analytical model is convenient and proved to be efficient by many experiments. However it is difficult to get the analytical solution. So normally it is only used to deal with one dimensional problems. And a lot of assumptions are needed, accordingly the accuracy is not high, sometimes it is hard to represent the psychical effects.

As the very large scale integrated circuit's (VLSI) development, this one dimensional analysis can not satisfy the request. Two dimension and three dimension's analysis turns to be on the schedule. Numerical model evolves accordingly. Compared to analytical model, numerical model starts from the basic semiconductor models, following the geometry structure boundary, constructing strict mathematical model, and then using numerical method to get the device characteristics. Obviously numerical model is more accurate than analytical model. Solving procedure is also more complicate than analytical method. In small dimension device domain, complicate carriers' transportation process can only be solved by numerical method.

Semiconductor device numerical method mainly has Monte Carlo (MC) method and continuous material method. MC needs to calculate the Boltzmann transportation equation of semiconductor, the calculation expenditure is huge. And if the high energy carriers distribution has relations or have low concentration at certain area, MC method has different results with different simulations, which can not satisfy the device simulation request. The later solution is to use the Boltzmann transportation equation's low order equations. By omitting the high order items in the equation, we can get a series of complicate mechanic models, including fluid dynamics model and drift diffusion model. These models have low expenditure of calculation than Boltzmann transportation equation, which leads to the possibility of using the numerical simulation on semiconductor industry.

In this part, we are going to illustrate how GSS realize the whole drift diffusion model. First we are going to introduce the detail of drift diffusion equation and its parameters. Because numerical solutions needs to discrete the model on certain meshes, we are going to describe the meshing structure in GSS. After have meshes, it is important to control the equations' discretion on the meshes. After we finish the discretization, partial differential equations is turned to be ordinary differential

equations. In the end it is how to deal with large scale non linear equation set. It is the scope of numerical method, which we only give simple illustration.

From user's point of view, semiconductor device numerical simulation is the process of constructing device structure, using appropriate physics models and corresponding mathematical abstraction, then by using numerical method software with specific process parameters, eg. geometry dimensions, electrical parameters, to calculate and obtain the characteristics of the device.

After finishing the drift diffusion model, we are going to do case study. Step by step discuss the GSS model construction and how to select different parameters in different cases and how to analyze the result with GSS.

Chapter 7 Basic Governing Equations

Since Gummel's work, the drift-diffusion model has been widely used in the semiconductor device simulation. It is now the defacto industry standard of this field.

The original DD model can be achieved by following approximation from hydrodynamic model:

- Light speed is much faster than carrier speed.
- All the collision is elastic.
- Bandgap does not change during collision.
- Carrier temperature equals to lattice temperature and keeps equilibrium.
- The gradient of driving force should keep small.
- Carrier degenerate can be neglected.

Some improvements have been applied to DD model for extend its capability. These "patches" of course make things complex, but they can deal with real problems.

This chapter contains the DD model and its variations used by GSS code for describing semiconductor device behavior as well as physical based parameters such as mobility, recombination rate and son on.

7.1 Level 1 Drift-Diffusion Equation

Level 1 Drift-Diffusion (DDML1) is the fundamental solver of GSS code for lattice temperature keeps constant though out the solve procedure.

The primary function of DDML1 is to solve the following set of partial differential equations, namely Poisson's equation, along with the hole and electron continuity equations:

Poisson's Equation

$$\nabla \cdot \varepsilon \nabla \psi = -q\left(p - n + N_D^+ - N_A^-\right) \tag{7.1}$$

where, ψ is the electrostatic potential, which specified the vacuum level throughout GSS code. This setting makes the description of metal-oxide-semiconductor contact and heterojunction easier. n and p are the electron and hole concentration, N_D^+ and N_A^- are the ionized impurity concentrations. q is the magnitude of the charge of an electron.

The relationship of conduct band E_c, valence band E_v and vacuum level ψ is:

$$E_c = -q\psi - \chi - \Delta E_c \tag{7.2}$$

$$E_v = E_c - E_g + \Delta E_v \tag{7.3}$$

Here, χ is the electron affinity. E_g is the bandgap of semiconductor. ΔE_c and ΔE_v are the bandgap shift caused by heavy doping or inner strain force.

Further more, the relationship of vacuum level ψ and intrinsic Fermi potential $\psi_{\text{intrinsic}}$ is:

$$\psi = \psi_{\text{intrinsic}} - \frac{\chi}{q} - \frac{E_g}{2q} - \frac{k_b T}{2q} \ln\left(\frac{N_c}{N_v}\right) \tag{7.4}$$

The reference 0 eV of energy is set to intrinsic Fermi level of equilibrium state in the GSS code.

Continuity Equation

The continuity equations for electrons and holes are defined as follows:

$$\begin{cases} \dfrac{\partial n}{\partial t} = \dfrac{1}{q}\nabla \cdot \mathbf{J}_n - (U - G) \\ \dfrac{\partial p}{\partial t} = -\dfrac{1}{q}\nabla \cdot \mathbf{J}_p - (U - G) \end{cases} \tag{7.5}$$

where $\mathbf{J}_n$ and $\mathbf{J}_n$ are the electron and hole current densities, U and G are the recombination and generation rates for both electrons and holes.

Drift-Diffusion Equation

The current densities $\mathbf{J}_n$ and $\mathbf{J}_n$ are expressed in terms of the level 1 drift-diffusion model here.

$$\begin{cases} \mathbf{J}_n = q\mu_n n\mathbf{E}_n + qD_n\nabla n \\ \mathbf{J}_p = q\mu_p p\mathbf{E}_p - qD_p\nabla p \end{cases} \tag{7.6}$$

where μ_n and μ_p are the electron and hole mobilities. $D_n = \dfrac{k_b T}{q}\mu_n$ and $D_p = \dfrac{k_b T}{q}\mu_p$ are the electron and hole diffusivities, according to Einstein relationship.

Effective Electrical Field

$\mathbf{E_n}$ and $\mathbf{E_p}$ are the effective driving electrical field to electrons and holes, which related with local band diagram. The band structure of heterojunction has been taken into account here [9].

$$\mathbf{E}_n = \frac{1}{q}\nabla E_c - \frac{k_b T}{q}\nabla\left(\ln(N_c) - \ln(T^{3/2})\right) \tag{7.7}$$

$$\mathbf{E}_p = \frac{1}{q}\nabla E_v + \frac{k_b T}{q}\nabla\left(\ln(N_v) - \ln(T^{3/2})\right) \tag{7.8}$$

The lattice temperature keeps uninform throughout DDML1, the above temperature gradient item takes no effect in fact.

By substituting drift-diffusion model into the current density expressions, and combining with Poisson's equation, the following basic equations for DDML1 are obtained:

$$\begin{cases} \dfrac{\partial n}{\partial t} = \nabla \cdot \left(\mu_n n\mathbf{E}_n + \mu_n \dfrac{k_b T}{q}\nabla n\right) - (U - G) \\ \dfrac{\partial p}{\partial t} = -\nabla \cdot \left(\mu_p p\mathbf{E}_p - \mu_p \dfrac{k_b T}{q}\nabla p\right) - (U - G) \\ \nabla \cdot \varepsilon\nabla\psi = -q(p - n + N_D^+ - N_A^-) \end{cases} \tag{7.9}$$

DDML1 is suitable for PN diode, BJT transistor and long gate MOSFET simulation. It is robust, and runs pretty fast for real work. The detailed discretization scheme can be found at "GSS First Level DDM Solver", on page 113.

7.2 Level 2 Drift-Diffusion Equation

The Level 2 DD model considers the influence of lattice temperature by solving the extra thermal equation simultaneously with the electrical equations. Also, the formula of drift-diffusion equation should be modified according to [10].

The electron diffusion current in DDML1 can be written as:

$$\mathbf{J}_{n,diff} = \frac{k_b T}{q}\mu_n q \nabla n = k_b T \mu_n \nabla n \tag{7.10}$$

Temperature Gradient Modification

But for DDML2, it has the form of

$$\mathbf{J}_{n,diff} = \mu_n k_b (T\nabla n + n\nabla T) \tag{7.11}$$

The hole diffusion current should be modified in the same manner.

$$\mathbf{J}_{p,diff} = -\mu_p k_b (T\nabla p + p\nabla T) \tag{7.12}$$

Heat Flow Equation

The following heat flow equation is used:

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot \kappa \nabla T + \mathbf{J} \cdot \mathbf{E} + (E_g + 3k_b T) \cdot (U - G) \tag{7.13}$$

where ρ is the mass density of semiconductor material. c_p is the heat capacity. κ is the thermal conductivity of the material. $\mathbf{J} \cdot \mathbf{E}$ is the joule heating of current. $(E_g + 3k_b T) \cdot (U - G)$ is lattice heating due to carrier recombination and generation.

From above discussion, the governing equations for DDML2 are as follows:

$$\begin{cases} \dfrac{\partial n}{\partial t} = \nabla \cdot \left(\mu_n n \mathbf{E}_n + \mu_n \dfrac{k_b T}{q} \nabla n + \mu_n \dfrac{k_b \nabla T}{q} n \right) - (U - G) \\ \dfrac{\partial p}{\partial t} = -\nabla \cdot \left(\mu_p p \mathbf{E}_p - \mu_p \dfrac{k_b T}{q} \nabla p - \mu_p \dfrac{k_b \nabla T}{q} p \right) - (U - G) \\ \nabla \cdot \varepsilon \nabla \psi = -q \left(p - n + N_D^+ - N_A^- \right) \\ \rho c_p \dfrac{\partial T}{\partial t} = \nabla \cdot \kappa \nabla T + \mathbf{J} \cdot \mathbf{E} + (E_g + 3k_b T) \cdot (U - G) \end{cases} \tag{7.14}$$

This model can be used as power transistor simulation as well as breakdown simulation. Unfortunately, nearly all the physical parameters are related with temperature. They should be considered during self consistent simulation, which greatly slows down the speed. The DDML2 solver runs 50-70% slower than DDML1. However, it seems no convergence degradation happens in most of the case. The discretization scheme can be found at "GSS Second Level DDM Solver", on page 117.

7.3 Level 3 Energy Balance Equation

Since version 0.45, the Energy Balance Model [11] is introduced into GSS code for simulating short channel MOSFET. This is a simplification of full hydrodynamic (HD) model[1] [12]. The current density expressions from the drift-diffusion model are modified to include additional coupling to the carrier temperature. Also, reduced carrier energy conservation equations, which derived from second order moment of Boltzmann Transport Equation, are solved consistently with drift-diffusion model. The simplification from HD to EB makes sophisticated Scharfetter-Gummel discretization still can be used in the numerical solution, which ensures the stability.

Current Equation for EBM

The current density $\mathbf{J}_n$ and $\mathbf{J}_p$ are then expressed as:

$$\mathbf{J}_n = q\mu_n n \mathbf{E}_n + k_b \mu_n (n\nabla T_n + T_n \nabla n) \tag{7.15}$$

$$\mathbf{J}_p = q\mu_p p \mathbf{E}_p - k_b \mu_p (p\nabla T_p + T_p \nabla p) \tag{7.16}$$

[1] GSS-0.37 implemented full HD model, but it runs very slow. More over, HD is unstable for dual carrier simulation due to machine round-off error.

where, T_n and T_p are electron and hole temperature, respectively. The difference between above equations and carrier density equations in DDML2 is lattice temperature replaced by carrier temperature.

Energy Balance Equations

In addition, the energy balance model includes the following electron and hole energy balance equations:

$$\frac{\partial (n\omega_n)}{\partial t} + \nabla \cdot \mathbf{S}_n = \mathbf{E}_n \cdot \mathbf{J}_n + H_n \tag{7.17}$$

$$\frac{\partial (p\omega_p)}{\partial t} + \nabla \cdot \mathbf{S}_p = \mathbf{E}_p \cdot \mathbf{J}_p + H_p \tag{7.18}$$

where, ω_n and ω_p are electron and hole energy. For HD model, the carrier energy includes thermal and kinetic terms $\omega_c = \frac{3}{2}k_b T_c + \frac{1}{2}m^* v_c^2$, but only thermal energy for EB model $\omega_c = \frac{3}{2}k_b T_c$. Here c stands for n or p. $\omega_0 = \frac{3}{2}k_b T$ is the carrier equilibrium energy, for carrier temperature equals to lattice temperature.

$\mathbf{S}_n$ and $\mathbf{S}_p$ are the flux of energy:

$$\begin{aligned} \mathbf{S}_n &= -\kappa_n \nabla T_n - (\omega_n + k_b T_n)\frac{\mathbf{J}_n}{q} \\ \mathbf{S}_p &= -\kappa_p \nabla T_p - (\omega_p + k_b T_p)\frac{\mathbf{J}_p}{q} \end{aligned} \tag{7.19}$$

The heat conductivity parameter for carriers can be expressed as:

$$\kappa_c = (\frac{2}{5} + \gamma)\frac{{k_b}^2}{q} T_c \mu_c c \tag{7.20}$$

where c stands for n and p, respectively. The constant parameter γ equals -0.7 in the GSS software.

The H_n and H_p are the rate of net loss of carrier kinetic energy:

$$\begin{aligned} H_n = {} & (R_{Aug} - G) \cdot \left(E_g + \frac{3k_b T_p}{2}\right) - \frac{3k_b T_n}{2}(R_{SHR} + R_{Dir} - G) \\ & - \frac{n(\omega_n - \omega_0)}{\tau_n} \end{aligned} \tag{7.21}$$

$$\begin{aligned} H_p = {} & (R_{Aug} - G) \cdot \left(E_g + \frac{3k_b T_n}{2}\right) - \frac{3k_b T_p}{2}(R_{SHR} + R_{Dir} - G) \\ & - \frac{p(\omega_p - \omega_0)}{\tau_p} \end{aligned} \tag{7.22}$$

where τ_n and τ_p are energy relaxation times for electrons and holes, respectively. The R_{Aug}, R_{SHR} and R_{Dir} are different recombination mechanisms referred in "Carrier Recombination", on page 75.

Lattice Heat Equation for EBM

At last, the lattice heat flow equation should be rewritten as:

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot \kappa \nabla T + H \tag{7.23}$$

where

$$H = R_{SHR} \cdot \left(E_g + \frac{3k_b T_p}{2} + \frac{3k_b T_n}{2}\right) + \frac{n(\omega_n - \omega_0)}{\tau_n} + \frac{p(\omega_p - \omega_0)}{\tau_p} \tag{7.24}$$

The carrier energy is mainly contributed by joule heating term $\mathbf{E_c} \cdot \mathbf{J_c}$, and heating (cooling) due to carrier generation (recombination) term. The carriers exchange

energy with lattice by collision, which described by energy relaxation term τ_{ω_c}. This model is suitable for sub-micron MOS (channel length 1 ~ 0.1 μm) and advanced BJT simulation. However, the computation burden of EB method is much higher than DD. And the convergence of EB solver is difficult to achieve, which requires more strict initial value and more powerful inner linear solver. The discretization scheme can be found at "GSS Third Level EBM Solver", on page 119.

From above discussion, all the governing equations of DD/EB method is elliptical or parabolic. From mathematic point of view, does not like hyperbolic system[2], the solution of elliptical or parabolic system is always smooth. The required numerical technique is simple and mature for these systems. As a result, the DD and EB method is preferred against full hydrodynamic method.

7.4 Quantum Modified Drift-Diffusion Equation

Today's microelectronic devices are so small that quantum mechanical effects are important. The classical drift-diffusion and hydrodynamical equations, can not face the challenge of nanometer device. Many quantum models are developed in recent years. The full quantum models, like Schrödinger-Poisson method and similar models, often lead to numerical complexity.

The density-gradient (DG) theory [13] [14], which is less detailed than full quantum models, dealing only in coarse-grained information and not providing explicit connections to more fundamental physics [15]. However, it is able to predict both the terminal characteristics and the density distribution with comparable result with Schrödinger-Poisson method for the device down to 10 nm [16]. The DG method is formulated in terms of partial differential equations and therefore tractable with the numerical methods commonly used for classical device simulation, while other quantum models involving complex eigenvalue problems.

DG-DDM Governing Equations

The governing equations of DG-DDM[17][18] are listed as below:

$$\begin{cases} \nabla\cdot\varepsilon\nabla\psi = -q\left(p-n+N_D^+-N_A^-\right)-\rho_s \\ \dfrac{\partial n}{\partial t} = \dfrac{1}{q}\nabla\cdot\mathbf{J}_n-(U-G) \\ \dfrac{\partial p}{\partial t} = \dfrac{1}{-q}\nabla\cdot\mathbf{J}_p-(U-G) \\ E_{qc}-E_c = -\dfrac{\hbar^2\gamma_n}{12m_n^*}\left[\nabla^2\left(\dfrac{E_{Fn}-E_{qc}}{k_bT}\right)+\dfrac{1}{2}\left(\nabla\dfrac{E_{Fn}-E_{qc}}{k_bT}\right)^2\right] \\ E_{qv}-E_v = \dfrac{\hbar^2\gamma_p}{12m_p^*}\left[\nabla^2\left(\dfrac{E_{qv}-E_{Fp}}{k_bT}\right)+\dfrac{1}{2}\left(\nabla\dfrac{E_{qv}-E_{Fp}}{k_bT}\right)^2\right] \end{cases} \tag{7.25}$$

where, γ_n and γ_p are the fitting parameters to make the result of DG-DDM consistence with Poisson-Schrödinger equation. E_{qc} is the quantum conduction band. E_{qv} is the quantum valence band. The quantum potential of electron and hole are defined as:

$$\begin{cases} \Lambda_n = \dfrac{E_{qc}-E_c}{q} \\ \Lambda_p = \dfrac{E_{qv}-E_v}{q} \end{cases} \tag{7.26}$$

Current Equation Modification

The current equation of DG-DDM keeps the same as Equation (7.6). However,

[2] One have to face discontinuous problem, i.e. shock wave.

the driving force of carrier should be modified as:

$$\mathbf{E}_n = \frac{1}{q}\nabla E_c - \frac{k_b T}{q}\nabla\left(\ln(N_c) - \ln(T^{3/2})\right) + \nabla\Lambda_n \tag{7.27}$$

$$\mathbf{E}_p = \frac{1}{q}\nabla E_v + \frac{k_b T}{q}\nabla\left(\ln(N_v) - \ln(T^{3/2})\right) + \nabla\Lambda_p \tag{7.28}$$

Express Quantum Potential with $\sqrt{n}$ and $\sqrt{p}$

Here gives another form of quantum potential equation which is useful in the numerical discretization. When Boltzmann statistics holds, the density of electrons and holes can be described as follows:

$$n = n_0 \exp\left(\frac{E_{Fn} - E_{qc}}{k_b T}\right) \tag{7.29}$$

$$p = p_0 \exp\left(\frac{E_{qv} - E_{Fp}}{k_b T}\right) \tag{7.30}$$

In the above equations, n_0 and p_0 are constants. Setting them to arbitrary value, i.e. 1, will not affect final result. By substitution Equation (7.29) into Equation (7.25), the following equation can be obtained.

$$\Lambda_n = -\frac{\hbar^2\gamma_n}{12qm_n^*}\left[\nabla^2 \ln n + \frac{1}{2}(\nabla \ln n)^2\right] \tag{7.31}$$

One can note that

$$\begin{aligned}\frac{1}{2}\left[\nabla^2 \ln n + \frac{1}{2}(\nabla \ln n)^2\right] &= \nabla^2 \ln\sqrt{n} + \left(\nabla \ln\sqrt{n}\right)^2 \\ &= \nabla\cdot\left(\frac{\nabla\sqrt{n}}{\sqrt{n}}\right) + \left(\frac{\nabla\sqrt{n}}{\sqrt{n}}\right)^2 \\ &= -\frac{1}{n}\nabla\sqrt{n}\cdot\nabla\sqrt{n} + \frac{\nabla^2\sqrt{n}}{\sqrt{n}} + \frac{1}{n}\left(\nabla\sqrt{n}\right)^2 \\ &= \frac{\nabla^2\sqrt{n}}{\sqrt{n}}\end{aligned} \tag{7.32}$$

As a result, the quantum potential of electrons can be rewritten as:

$$\Lambda_n = -\frac{\hbar^2\gamma_n}{6qm_n^*}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{7.33}$$

Similar expressions exist for the quantum potential of holes

$$\Lambda_p = \frac{\hbar^2\gamma_p}{6qm_p^*}\frac{\nabla^2\sqrt{p}}{\sqrt{p}} \tag{7.34}$$

The discretization scheme of DG-DDM equations is listed in "??", on page ??.

7.5 Bandgap Parameters

The bandgap parameters, including bandgap E_g, effective density of states in the conduction band N_c and valence band N_v, and intrinsic carrier concentration n_{ie}, are the most important and fundamental physical parameters for semiconductor material [7].

Effective Density of States

Effective density of states in the conduction and valence band are defined as

follows:

$$N_c \equiv 2\left(\frac{m_n^* k_b T}{2\pi\hbar^2}\right)^{3/2} \tag{7.35}$$

$$N_v \equiv 2\left(\frac{m_p^* k_b T}{2\pi\hbar^2}\right)^{3/2} \tag{7.36}$$

The temperature dependencies of effective density of states is fairly simple:

$$N_c(T) = N_c(300\text{ K})\left(\frac{T}{300\text{ K}}\right)^{1.5} \tag{7.37}$$

$$N_v(T) = N_v(300\text{ K})\left(\frac{T}{300\text{ K}}\right)^{1.5} \tag{7.38}$$

Bandgap

The bandgap in GSS is expressed as follows:

$$\begin{aligned} E_g(T) &= E_g(0) - \frac{\alpha \cdot T^2}{T+\beta} \\ &= E_g(300) + \alpha\left[\frac{300^2}{300+\beta} - \frac{T^2}{T+\beta}\right] \end{aligned} \tag{7.39}$$

Bandgap Narrowing due to Heavy Doping

When bandgap narrowing effects due to heavy doping takes place [19], the band edge shifts:

$$\Delta E_g = \frac{E_{bgn}}{2k_b T}\left[\ln\frac{N_{total}}{N_{ref}} + \sqrt{\left(\ln\frac{N_{total}}{N_{ref}}\right)^2 + 0.5}\right] \tag{7.40}$$

For silicon, $\alpha = 4.73\times10^{-4}$eV/K, $\beta = 6.36\times10^{2}$K, $E_{bgn} = 9\times10^{-3}$eV, $N_{ref} = 1.0\times10^{17}\text{cm}^{-3}$.

The intrinsic concentration should be modified:

$$n_{ie} = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right)\cdot\exp(\Delta E_g) \tag{7.41}$$

Since the carrier current Equation (7.7) and Equation (7.8) involves the energy level of conduction band N_c and valence band N_v, the bandgap shift should be attributed to them. The bandgap narrowing is attributed half to the conduction band and another half to the valence band as default:

$$E_c' = E_c - \frac{1}{2}\Delta E_g \tag{7.42}$$

$$E_v' = E_v + \frac{1}{2}\Delta E_g \tag{7.43}$$

7.6 Carrier Recombination

Three recombination mechanisms are considered in GSS at present, including Shockley-Read-Hall, Auger, and direct (or radiative) recombination. The total recombination is considered as the sum of all:

$$U = U_n = U_p = U_{\text{SRH}} + U_{\text{dir}} + U_{\text{Auger}} \tag{7.44}$$

where U_{SRH}, U_{dir} and U_{Auger} are SRH recombination, direct recombination and Auger recombination, respectively.

SRH Recombination

Shockley-Read-Hall (SRH) recombination rate is determined by the following for-

mula:

$$U_{\mathrm{SRH}} = \frac{pn - n_{ie}^2}{\tau_p[n + n_{ie}\exp(\frac{\mathbf{ETRAP}}{kT_L})] + \tau_n[p + n_{ie}\exp(\frac{-\mathbf{ETRAP}}{kT_L})]} \tag{7.45}$$

where τ_n and τ_p are carrier life time, which dependent on impurity concentration [20].

$$\tau_n = \frac{\mathbf{TAUN0}}{1 + N_{total}/\mathbf{NSRHN}} \tag{7.46}$$

$$\tau_p = \frac{\mathbf{TAUP0}}{1 + N_{total}/\mathbf{NSRHP}} \tag{7.47}$$

The parameter $\mathbf{ETRAP} = E_t - E_i$, where E_t is the energy level for the recombination centers and E_i is the intrinsic Fermi Energy.

Auger Recombination

The Auger recombination is a three-carrier recombination process, involving either two electrons and one hole or two holes and one electron. This mechanism becomes important when carrier concentration is large.

$$U_{\mathrm{Auger}} = \mathbf{AUGN}(pn^2 - nn_{ie}^2) + \mathbf{AUGP}(np^2 - pn_{ie}^2) \tag{7.48}$$

Where **AUGN** and **AUGP** are Auger coefficient for electrons and holes. The value of Auger recombination U_{Auger} can be negative some times, which refers to Auger generation.

Direct Recombination

The direct recombination model expresses the recombination rate as a function of the carrier concentrations n and p, and the effective intrinsic density n_{ie}:

$$U_{\mathrm{dir}} = \mathbf{DIRECT}(np - n_{ie}^2) \tag{7.49}$$

The default value are listed below:

	Unit	Silicon	GaAs	Ge
ETRAP	eV	0	0	0
DIRECT	$\mathrm{cm^3s^{-1}}$	1.1e-14	7.2e-10	6.41e-14
AUGN	$\mathrm{cm^6s^{-1}}$	1.1e-30	1e-30	1e-30
AUGP	$\mathrm{cm^6s^{-1}}$	0.3e-30	1e-29	1e-30
TAUN0	s	1e-7	5e-9	1e-7
TAUP0	s	1e-7	3e-6	1e-7
NSRHN	$\mathrm{cm^{-3}}$	5e16	5e17	5e16
NSRHP	$\mathrm{cm^{-3}}$	5e16	5e17	5e16

7.7 Mobility Models

Carrier mobility is one of the most important parameters in the carrier transport model. The DD model itself, developed at early 1980s, is still being used today due to advanced mobility model enlarged its ability to sub-micron device.

Mobility modeling is normally divided into: low field behavior, high field behavior and mobility in the (MOS) inversion layer.

The low electric field behavior has carriers almost in equilibrium with the lattice.

The low-field mobility is commonly denoted by the symbol μ_{n0}, μ_{p0}. The value of this mobility is dependent upon phonon and impurity scattering. Both of which act to decrease the low field mobility. Since scattering mechanism is depended on lattice temperature, the low-field mobility is also a function of lattice temperature.

The high electric field behavior shows that the carrier mobility declines with electric field because the carriers that gain energy can take part in a wider range of scattering processes. The mean drift velocity no longer increases linearly with increasing electric field, but rises more slowly. Eventually, the velocity doesnąt increase any more with increasing field but saturates at a constant velocity. This constant velocity is commonly denoted by the symbol v_{sat}. Impurity scattering is relatively insignificant for energetic carriers, and so v_{sat} is primarily a function of the lattice temperature.

Modeling carrier mobilities in inversion layers introduces additional complications. Carriers in inversion layers are subject to surface scattering, extreme carrier-carrier scattering, velocity overshoot and quantum mechanical size quantization effects. These effects must be accounted for in order to perform accurate simulation of MOS devices. The transverse electric field is often used as a parameter that indicates the strength of inversion layer phenomena.

It can be seen that some physical mechanisms such as velocity overshoot and quantum effect which can't be described by DD method at all, can be taken into account by comprehensive mobility model. The comprehensive mobility model externs the application range of DD method. However, when the EB method (which accounts for velocity overshoot) and QDD method (including quantum effect) are used, more calibrations are needed to existing mobility models.

7.7.1 Analytic Mobility Model

In the GSS code, Analytic Mobility model [3] [21] is the default low field mobility model for all the material. It is an concentration and temperature dependent empirical mobility model expressed as:

$$\mu_0 = \mu_{min} + \frac{\mu_{max}\left(\frac{T}{300}\right)^{\alpha} - \mu_{min}}{1 + \left(\frac{T}{300}\right)^{\beta}\left(\frac{N_{total}}{N_{ref}}\right)^{\gamma}} \tag{7.50}$$

where $N_{total} = N_A + N_D$ is the total impurity concentration.

Other parameters for Si GaAs and Ge are listed below:

	Unit	Silicon: N	Silicon: P	GaAs: N	GaAs: P	Ge
μ_{min}	$cm^2 \cdot (V \cdot s)^{-1}$	55.24	49.70	0.0	0.0	Si
μ_{max}	$cm^2 \cdot (V \cdot s)^{-1}$	1429.23	479.37	8500.0	400.0	Si
α	–	-2.3	-2.2	-1.0	-2.1	Si
β	–	-3.8	-3.7	0.0	0.0	Si
γ	–	0.73	0.70	0.436	0.395	Si
N_{ref}	cm^{-3}	1.072e17	1.606e17	1.69e17	2.75e17	Si

In the actual implement, the Analytic mobility model is modified for taking high field effects (carrier velocity saturation effects) into account. For silicon material,

Caughey-Thomas expression [3] is used for this modification:

$$\mu_n = \frac{\mu_{0,n}}{\left[1+\left(\frac{\mu_{0,n}E_{//}}{V_{sat}}\right)^2\right]^{1/2}} \tag{7.51}$$

$$\mu_p = \frac{\mu_{0,p}}{1+\left(\frac{\mu_{0,p}E_{//}}{V_{sat}}\right)} \tag{7.52}$$

where $E_{//}$ is the electric field parallel to current flow. V_{sat} is the saturation velocities for electrons or holes. They are computed by default from the expression:

$$V_{sat}(T) = \frac{2.4\times10^7}{1+0.8\cdot\exp\left(\frac{T}{600}\right)} \tag{7.53}$$

For GaAs material, another expression is used [22]:

$$u_n = \frac{\mu_{0,n}+\frac{V_{sat}}{E_{//}}\left(\frac{E_{//}}{E_{0,N}}\right)^4}{1+\left(\frac{E_{//}}{E_{0,N}}\right)^4} \tag{7.54}$$

$$u_p = \frac{\mu_{0,p}+\frac{V_{sat}}{E_{//}}\left(\frac{E_{//}}{E_{0,P}}\right)^4}{1+\left(\frac{E_{//}}{E_{0,P}}\right)^4} \tag{7.55}$$

where $V_{sat}(T) = 11.3\times10^6 - 1.2\times10^4 T$ is the carrier saturation velocities for GaAs. $E_{0,N} = 4.0\times10^3$ V/cm and $E_{0,P} = 1.0\times10^6$ V/cm are the reference electrical field for electrons and holes, respectively. The negative differential property of carrier mobility is described by this mode. When electric field increases in this model, the carrier drift velocity ($\mu E_{//}$) reaches a peak and then begins to decrease at high fields due to the transferred electron effect.

When using this model for GaAs MESFET device simulation, the negative differential property may cause the drain output characteristics (current vs. voltage) exhibit an unrealistic oscillation behavior. Another model to describe high field effects developed by Yeager [23] can be used.

$$\mu = \frac{V_{sat}}{E_{//}}\tanh\left(\frac{\mu_0 E_{//}}{V_{sat}}\right) \tag{7.56}$$

It can be loaded by Hypertang keyword in PMIS statement.

Due to the widely usage of Si material, there are many mobility models existing for silicon. The following paragraphs described some more (complex) mobility modes for silicon which had been implemented into GSS. However, for other semiconductor materials in the GSS's material database, only one mobility model in the same formula as Analytic is provided at present.

7.7.2 Philips Mobility Model

Another low field mobility model implemented into GSS is the Philips Unified Mobility model [4][5]. This model takes into account the distinct acceptor and donor

scattering, carrier-carrier scattering and carrier screening, which is recommended for bipolar devices simulation.

The electron mobility is described by the following expressions:

$$\mu_{0,n}^{-1} = \mu_{Lattice,n}^{-1} + \mu_{D+A+p}^{-1} \tag{7.57}$$

where $\mu_{0,n}$ is the total low field electron mobilities, $\mu_{Lattice,n}$ is the electron mobilities due to lattice scattering, μ_{D+A+p} is the electron and hole mobilities due to donor (D), acceptor (A), screening (P) and carrier-carrier scattering.

$$\mu_{Lattice,n} = \mu_{max}\left(\frac{T}{300}\right)^{-2.285} \tag{7.58}$$

$$\mu_{D+A+p} = \mu_{1,n}\left(\frac{N_{sc,n}}{N_{sc,eff,n}}\right)\left(\frac{N_{ref}}{N_{sc,n}}\right)^{\alpha} + \mu_{2,n}\left(\frac{n+p}{N_{sc,eff,n}}\right) \tag{7.59}$$

The parameters $\mu_{1,n}$ and $\mu_{2,n}$ are given as:

$$\mu_{1,n} = \frac{\mu_{max}^2}{\mu_{max} - \mu_{min}}\left(\frac{T}{300}\right)^{3\alpha-1.5} \tag{7.60}$$

$$\mu_{2,n} = \frac{\mu_{max} \cdot \mu_{min}}{\mu_{max} - \mu_{min}}\left(\frac{300}{T}\right)^{1.5} \tag{7.61}$$

where $N_{sc,n}$ and $N_{sc,eff,n}$ is the impurity-carrier scattering concentration and effect impurity-carrier scattering concentration given by:

$$\begin{aligned} N_{sc,n} &= N_D^* + N_A^* + p \\ N_{sc,eff,n} &= N_D^* + N_A^* G(P_n) + \frac{p}{F(P_n)} \end{aligned} \tag{7.62}$$

where N_D^* and N_A^* take ultra-high doping effects into account and are defined by:

$$\begin{aligned} N_D^* &= N_D\left(1 + \frac{1}{C_D + \left(\frac{N_{D,ref}}{N_D}\right)^2}\right) \\ N_A^* &= N_A\left(1 + \frac{1}{C_A + \left(\frac{N_{A,ref}}{N_A}\right)^2}\right) \end{aligned} \tag{7.63}$$

The screening factor functions $G(P_n)$ and $F(P_n)$ take the repulsive potential for acceptors and the finite mass of scattering holes into account.

$$G(P_n) = 1 - \frac{0.89233}{\left[0.41372 + P_n\left(\frac{m_0}{m_e}\frac{T}{300}\right)^{0.28227}\right]^{0.19778}} + \frac{0.005978}{\left[P_n\left(\frac{m_e}{m_0}\frac{T}{300}\right)^{0.72169}\right]^{1.80618}} \tag{7.64}$$

$$F(P_n) = \frac{0.7643P_n^{0.6478} + 2.2999 + 6.5502\frac{m_e}{m_h}}{P_n^{0.6478} + 2.3670 - 0.8552\frac{m_e}{m_h}} \tag{7.65}$$

The P_n parameter that takes screening effects into account is given by:

$$P_n = \left[\frac{f_{cw}}{3.97 \times 10^{13} N_{sc,n}^{-2/3}} + \frac{f_{BH}}{\frac{1.36 \times 10^{20}}{n+p}\left(\frac{m_e}{m_0}\right)} \right]^{-1} \left(\frac{T}{300}\right)^2 \tag{7.66}$$

Similar expressions hold for holes. The default parameters for Philips model are listed as below:

	Unit	Silicon: N-TYPE	Silicon: P-TYPE
μ_{min}	$cm^2 \cdot (V \cdot s)^{-1}$	55.2	44.90
μ_{max}	$cm^2 \cdot (V \cdot s)^{-1}$	1417.0	470.5
α	–	0.68	0.719
N_{ref}	cm^{-3}	9.68e16	2.23e17
C_D	-	0.21	0.21
C_A	-	0.5	0.5
$N_{D,ref}$	cm^{-3}	4.0e20	4.0e20
$N_{A,ref}$	cm^{-3}	7.2e20	7.2e20
m_e	m_0	1.0	-
m_h	m_0	-	1.258
f_{cw}	-	2.459	2.459
f_{BH}	-	3.828	3.828

In the actual code, Philips model is corrected by Caughey-Thomas expression for taking high field velocity saturation effects into account. This model can be loaded by Philips keyword in the PMIS statements.

7.7.3 Lombardi Surface Mobility Model

Along insulator-semiconductor interfaces, the carrier mobilities can be substantially lower than in the bulk of the semiconductor due to surface scattering. If no surface degradation is considered, the drain-source current may exceed about 30% for MOS simulation.

GSS uses Lombardi Surface Mobility model [24] for accounting surface degradation:

$$\mu_s^{-1} = \mu_{ac}^{-1} + \mu_{sr}^{-1} \tag{7.67}$$

where μ_{ac} is mobility degraded by surface acoustical phonon scattering. μ_{sr} is mobility degraded by surface roughness scattering.

$$\mu_{ac,n} = \frac{3.61 \times 10^7}{E_\perp} + \frac{1.70 \times 10^4 N_{total}^{0.0233}}{\left(\frac{T}{300}\right)\sqrt[3]{E_\perp}} \tag{7.68}$$

$$\mu_{ac,n} = \frac{1.51 \times 10^7}{E_\perp} + \frac{4.18 \times 10^3 N_{total}^{0.0119}}{\left(\frac{T}{300}\right)^{0.9}\sqrt[3]{E_\perp}} \tag{7.69}$$

$$\mu_{sr,n} = \frac{3.58 \times 10^{18}}{E_{\perp}^{\gamma_n}} \tag{7.70}$$

$$\mu_{sr,p} = \frac{4.10 \times 10^{15}}{E_{\perp}^{\gamma_p}} \tag{7.71}$$

where

$$\gamma_n = 2.58 + \frac{6.85 \times 10^{-21}(n+p)}{N_{total}^{0.0767}} \tag{7.72}$$

$$\gamma_p = 2.18 + \frac{7.82 \times 10^{-21}(n+p)}{N_{total}^{0.123}} \tag{7.73}$$

where $E_{\perp}$ is the components of electric field perpendicular to the current direction which stands for the distance to the insulator-semiconductor interface.

The Lombardi model is not used alone. Instead, it is the composition of other comprehensive mobility models such as Lucent model in next paragraph.

7.7.4 Lucent High Field Mobility Model

The Lucent Mobility model [25] is an all-inclusive model which suitable for MOS simulation. This model incorporates Philips Unified Mobility model and the Lombardi Surface Mobility model, as well as accounting for high field effects. For low longitudinal field, the carrier mobility is given by Matthiessen's rule:

$$\mu_0 = \left[\frac{1}{\mu_b} + \frac{1}{\mu_{ac}} + \frac{1}{\mu_{sr}}\right]^{-1} \tag{7.74}$$

where μ_b is bulk mobility comes from (slightly modified) Philips model, μ_{ac} and μ_{sr} keep the same as Lombardi model.

Finally, for accounting high field effects, the total mobility is modified using the expressions as Caughey-Thomas model:

$$\mu_n = \frac{2\mu_{0,n}}{1 + \left[1 + \left(\frac{2\mu_{0,p}E_{//}}{V_{sat}}\right)^2\right]^{1/2}} \tag{7.75}$$

$$\mu_p = \frac{\mu_{0,p}}{1 + \left(\frac{\mu_{0,p}E_{//}}{V_{sat}}\right)} \tag{7.76}$$

Lucent is an accurate model recommended for MOS devices. The only shortcoming is its heavy computational burden. This model can be loaded by Lucent keyword in the PMIS statements.

7.7.5 Hewlett-Packard High Field Mobility Model

It is reported that Hewlett-Packard mobility model [26] achieves the same accuracy as Lucent model with relatively small computational burden in the MOS simulation.

This model also takes into account dependence on electric fields both parallel and perpendicular to the direction of current flow.

$$\mu_n = \frac{\mu_{\perp,n}}{\sqrt{1 + \frac{\left(\frac{\mu_{\perp,n}E_{//}}{V_{c,n}}\right)^2}{\frac{\mu_{\perp,n}E_{//}}{V_{c,n}} + \gamma_n} + \frac{\mu_{\perp,n}E_{//}}{V_{s,n}}}}$$
$$\mu_p = \frac{\mu_{\perp,p}}{\sqrt{1 + \frac{\left(\frac{\mu_{\perp,p}E_{//}}{V_{c,p}}\right)^2}{\frac{\mu_{\perp,p}E_{//}}{V_{c,p}} + \gamma_p} + \frac{\mu_{\perp,p}E_{//}}{V_{s,p}}}} \tag{7.77}$$

where $\mu_{\perp,n}$ and $\mu_{\perp,p}$ can be expressed as below:

$$\mu_{\perp,n} = \begin{cases} \mu_{0,n} & \text{if } N_{total} > N_{ref} \\ \dfrac{\text{mun0}}{1 + \dfrac{E_\perp}{E_{ref,n}}} \end{cases}$$
$$\mu_{\perp,p} = \begin{cases} \mu_{0,p} & \text{if } N_{total} > N_{ref} \\ \dfrac{\text{mup0}}{1 + \dfrac{E_\perp}{E_{ref,p}}} \end{cases} \tag{7.78}$$

The default value for N_{ref} is $5{\times}10^{17}$ cm^{-3}. If the above conditions are not satisfied, then $\mu_{\perp,n} = \mu_{0,n}$ and $\mu_{\perp,p} = \mu_{0,p}$, where $\mu_{0,n}$ and $\mu_{0,p}$ are the low field mobility values calculated by Analytic model.

The default value of Hewlett-Packard mobility model are:

	Unit	Silicon: N-TYPE	Silicon: P-TYPE
mun0	$\text{cm}^2 \cdot (\text{V} \cdot \text{s})^{-1}$	774.0	-
mup0	$\text{cm}^2 \cdot (\text{V} \cdot \text{s})^{-1}$	-	250
$V_{c,n}$	$\text{cm} \cdot \text{s}^{-1}$	4.9e6	-
$V_{c,p}$	$\text{cm} \cdot \text{s}^{-1}$	-	2.928e6
$V_{s,n}$	$\text{cm} \cdot \text{s}^{-1}$	1.036e7	-
$V_{s,p}$	$\text{cm} \cdot \text{s}^{-1}$	-	1.2e7
γ_n	-	8.8	-
γ_p	-	-	1.6
N_{ref}	cm^{-3}	5e17	5e17
$E_{ref,n}$	$\text{V} \cdot \text{cm}^{-1}$	5.5e5	-
$E_{ref,p}$	$\text{V} \cdot \text{cm}^{-1}$	-	2.78e5

Hewlett-Packard mobility model can be loaded by HP keyword in the PMIS statement.

7.7.6 Mobility Model used for EB

We should notice here, all the above mobility models are developed under the framework of DD method. Since DD is an approximate model for semiconductor, the difference between DD model and real device is corrected by mobility models! These mobility model contains some physical model that DD does not consider. For example, the high field correction has already contains the effect of hot carriers. The surface mobility for MOSFET not only considers the mobility degrade due to surface roughness, but also contains the effect caused by carrier concentration decrease due to quantum well in inverse layer. These corrections extended the application range of DD model, also make the mobility model rather complex.

When the physical model is more accurate, the carrier mobility model can be less complicated. Thus, the mobility models suitable for DD model may NOT be suitable for energy balance model. There are some mobility models developed special for energy balance model [11][27]. However, they have not be implemented into GSS yet.

7.8 Impact Ionization

The generation rate of electron-hole pairs due to the carrier impact ionization (II) is generally modeled as [7]:

$$G^{II} = \alpha_n \frac{|\,\mathbf{J}_n\,|}{q} + \alpha_p \frac{|\,\mathbf{J}_p\,|}{q} \tag{7.79}$$

where α_n and α_p are electron and hole ionization coefficients, related with electrical field, material and temperature.

Selberherr Model

Selberherr gives an empirical formula [10], which is the default model used by GSS:

$$\alpha_{n,p} = \alpha^{\infty}_{n,p}(T) \exp\left(-\frac{E^{Crit}_{n,p}}{E_{n,p}}\right) \tag{7.80}$$

where $E_{n,p}$ is the magnitude of driving fields. When EdotJ model is used, $E_{n,p}$ can be given by:

$$E_n = \frac{\mathbf{E}\cdot\mathbf{J}_n}{|\mathbf{J}_n|}, \qquad E_p = \frac{\mathbf{E}\cdot\mathbf{J}_p}{|\mathbf{J}_p|} \tag{7.81}$$

and for GradQf model:

$$E_n = \left|\nabla\phi_{F_n}\right|, \qquad E_p = \left|\nabla\phi_{F_p}\right| \tag{7.82}$$

where $E^{Crit}_{n,p} = \dfrac{E_g}{q\lambda_{n,p}}$, for which $\lambda_{n,p}$ are the optical-phonon mean free paths for electrons and holes given by:

$$\begin{aligned} \lambda_n(T) &= \lambda_{n,0}\cdot\tanh\left(\frac{E_{op}}{2k_bT}\right) \\ \lambda_p(T) &= \lambda_{p,0}\cdot\tanh\left(\frac{E_{op}}{2k_bT}\right) \end{aligned} \tag{7.83}$$

in the above expressions, E_{op} is the optical-phonon energy. $\lambda_{n,0}$ and $\lambda_{p,0}$ are the phonon mean free paths for electrons and holes at 300 K.

The temperature dependent factors α_n^∞ and α_p^∞ are expressed as:

$$\alpha_n^\infty = \alpha_{n,0} + \alpha_{n,1} \cdot T + \alpha_{n,2} \cdot T^2$$
$$\alpha_p^\infty = \alpha_{p,0} + \alpha_{p,1} \cdot T + \alpha_{p,2} \cdot T^2$$

The default parameters used for Selberherr model:

	Unit	Silicon	GaAs	Ge
$\lambda_{n,0}$	cm	1.04542×10^{-6}	3.52724×10^{-6}	6.88825×10^{-7}
$\lambda_{p,0}$	cm	6.32079×10^{-7}	3.67649×10^{-6}	8.39505×10^{-7}
E_{op}	eV	6.3×10^{-2}	3.5×10^{-2}	3.7×10^{-2}
$\alpha_{n,0}$	cm^{-1}	7.030×10^{5}	2.994×10^{5}	1.55×10^{7}
$\alpha_{n,1}$	$cm^{-1} \cdot K^{-1}$	0.0	0.0	0.0
$\alpha_{n,2}$	$cm^{-1} \cdot K^{-2}$	0.0	0.0	0.0
$\alpha_{p,0}$	cm^{-1}	1.528×10^{6}	2.215×10^{5}	1.00×10^{7}
$\alpha_{p,1}$	$cm^{-1} \cdot K^{-1}$	0.0	0.0	0.0
$\alpha_{p,2}$	$cm^{-1} \cdot K^{-2}$	0.0	0.0	0.0

Valdinoci Model

GSS has another Valdinoci model for silicon device which has been reported to produce correct temperature dependence of breakdown voltage of junction diodes as high as 400K [28]. It can be loaded by specification Valdinoci in the PMIS statements.

	Silicon: N-TYPE		Silicon: P-TYPE	Unit
A0N	4.3383	**A0P**	2.376	V
A1N	-2.42×10^{-12}	**A1P**	1.033×10^{-2}	$V \cdot K^{-A2X}$
A2N	4.1233	**A2P**	1.0	-
B0N	0.235	**B0P**	0.17714	V
B1N	0.0	**B1P**	-2.178×10^{-3}	K^{-1}
C0N	1.6831×10^{4}	**C0P**	0.0	$V \cdot cm^{-1}$
C1N	4.3796	**C1P**	9.47×10^{-3}	$V \cdot cm^{-1} \cdot K^{-C2X}$
C2N	1.0	**C2P**	2.4924	-
C3N	0.13005	**C3P**	0.0	$V \cdot cm^{-1} \cdot K^{-2}$
D0N	1.233735×10^{6}	**D0P**	1.4043×10^{6}	$V \cdot cm^{-1}$
D1N	1.2039×10^{3}	**D1P**	2.9744×10^{3}	$V \cdot cm^{-1} \cdot K^{-1}$
D2N	0.56703	**D2P**	1.4829	$V \cdot cm^{-1} \cdot K^{-2}$

The electron impact ionization rate for Valdinoci model reads:

$$\alpha_n = \frac{E_{//}}{a_n(T) + b_n(T) \exp\left(\frac{d_n(T)}{E_{//} + c_n(T)}\right)} \tag{7.84}$$

where

$$a_n(T) = \mathbf{A0N} + \mathbf{A1N} \cdot T^{\mathbf{A2N}}$$
$$b_n(T) = \mathbf{B0N} \cdot \exp(\mathbf{B1N} \cdot T)$$
$$c_n(T) = \mathbf{C0N} + \mathbf{C1N} \cdot T^{\mathbf{C2N}} + \mathbf{C3N} \cdot T^2$$
$$d_n(T) = \mathbf{D0N} + \mathbf{D1N} \cdot T + \mathbf{D2N} \cdot T^2$$

Similar expressions hold for holes. The parameters for Valdinoci model are listed in the table.

Generation by Tunneling

The carrier generation by band-band tunneling G^{BB} is also considered by GSS, which can be expressed as: [2]

$$G^{BB} = 3.5 \times 10^{21} \cdot \frac{E^2}{\sqrt{E_g}} \cdot \exp\left(-22.5 \times 10^6 \cdot \frac{E_g^{3/2}}{E}\right) \tag{7.85}$$

where E is the magnitude of electrical field.

7.9 Fermi-Dirac Statistics

In general, the electron and hole concentrations in semiconductors are defined by Fermi-Dirac distributions and density of states:

$$n = N_c \, \mathrm{F}_{1/2}(\eta_n) \tag{7.86}$$
$$p = N_v \, \mathrm{F}_{1/2}(\eta_p) \tag{7.87}$$

The η_n and η_p are defined as follows:

$$\eta_n = \frac{E_{F_n} - E_c}{k_b T} = \mathrm{F}^{-1}_{1/2}\left(\frac{n}{N_c}\right) \tag{7.88}$$
$$\eta_p = \frac{E_v - E_{F_p}}{k_b T} = \mathrm{F}^{-1}_{1/2}\left(\frac{p}{N_v}\right) \tag{7.89}$$

where E_{F_n} and E_{F_p} are the electron and hole Fermi energies. The relation ship of Fermi energy and Fermi potential is $E_{F_n} = -q\phi_n$, $E_{F_p} = -q\phi_p$.

Evaluate Inverse Fermi Integral

$\mathrm{F}^{-1}_{1/2}$ is inverse Fermi integral of order one-half. Joyce and Dixon gives its approximation analytic expression in the year of 1977 [29], which can be given by:

$$F^{-1}_{1/2}(x) = \begin{cases} \log(x) + ax + bx^2 + cx^3 + dx^4 & x < 8.463 \\ \left(\left(\frac{3\sqrt{\pi}}{4}x\right)^{3/4} - \frac{\pi^2}{6}\right)^{1/2} & otherwise \end{cases} \tag{7.90}$$

where

$$a = 0.35355339059327379$$
$$b = 0.0049500897298752622$$
$$c = 1.4838577128872821 \times 10^{-4}$$
$$d = 4.4256301190009895 \times 10^{-6}$$

In the GSS code, the η_n and η_p are derived from carrier concentration by Joyce-Dixon expression.

For convenience, we introduce floowing two parameters as referred by [30]:

$$\gamma_n = \frac{\mathrm{F}_{1/2}(\eta_n)}{\exp(\eta_n)} \tag{7.91}$$
$$\gamma_p = \frac{\mathrm{F}_{1/2}(\eta_p)}{\exp(\eta_p)} \tag{7.92}$$

The carrier concentration for Fermi statistics and Boltzmann statistics can be described uniformly by:

$$n = N_c \gamma_n \exp(\eta_n) \tag{7.93}$$

$$p = N_v \gamma_p \exp(\eta_p) \tag{7.94}$$

where $\gamma_n = \gamma_p = 1$ for Boltzmann statistics, and less than 1.0 for Fermi statistics.

DD Equation with Fermi Statistics

Consider the drift-diffusion current Equation (7.6), when the carrier satisfies Fermi statistics and forces zero net current in equilibrium state, one can get the modified current equation, for which the Einstein relationship:

$$D_n = \frac{k_b T}{q} \mu_n \tag{7.95}$$

$$D_p = \frac{k_b T}{q} \mu_p \tag{7.96}$$

should be replaced by:

$$D_n = \frac{k_b T}{q} \mu_n \mathrm{F}_{1/2}(\eta_n) / \mathrm{F}_{-1/2}(\eta_n) \tag{7.97}$$

$$D_p = \frac{k_b T}{q} \mu_p \mathrm{F}_{1/2}(\eta_p) / \mathrm{F}_{-1/2}(\eta_p) \tag{7.98}$$

where $\mathrm{F}_{-1/2}$ is the Fermi integral of order minus one-half. The corresponding current equation for electrons is

$$\mathbf{J}_n = \mu_n (q n \mathbf{E}_n + k_b T \lambda_n \nabla n) \tag{7.99}$$

where

$$\lambda_n = \frac{\mathrm{F}_{1/2}(\eta_n)}{\mathrm{F}_{-1/2}(\eta_n)} \tag{7.100}$$

The Fermi integral has an useful property:

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{F}_\nu(\eta) = \mathrm{F}_{\nu-1}(\eta) \tag{7.101}$$

From the above property, one can derive two useful derivatives:

$$\frac{\partial}{\partial n} \eta_n(n) = \frac{\lambda_n}{n} \tag{7.102}$$

$$\frac{\partial}{\partial n} \gamma_n(n) = \frac{\gamma_n}{n}(1 - \lambda_n) \tag{7.103}$$

With the two derivatives, Equation (7.99) can be rewritten into the following equivalent formula:

$$\mathbf{J}_n = \mu_n (q n \mathbf{E}_n + k_b T \nabla n - n k_b T \nabla (\ln \gamma_n)) \tag{7.104}$$

The last term is the modification to Einstein relationship, which can be combinated into potential term. As a result, the current Equation (7.6) keeps unchanged, but the effective driving force should be modified as:

$$\mathbf{E}_n = \frac{1}{q} \nabla E_c - \frac{k_b T}{q} \nabla \left(\ln(N_c) - \ln(T^{3/2}) \right) - \frac{k_b T}{q} \nabla (\ln \gamma_n) \tag{7.105}$$

The same formula exists for holes:

$$\mathbf{E}_p = \frac{1}{q} \nabla E_v + \frac{k_b T}{q} \nabla \left(\ln(N_v) - \ln(T^{3/2}) \right) + \frac{k_b T}{q} \nabla (\ln \gamma_p) \tag{7.106}$$

As a conclusion, when Fermi statistics is considered, the formula of DD method keeps unchanged, only an extra potential term should be considered. However, Fermi statistics also effect the implement of Ohmic boundary condition, please refer to "??", on page ??.

Chapter 8 Mesh Techniques in TCAD

By using numerical technique to solve partial differential equations (PDEs), first we need to divide the computational region into finite sub domains. So that we can discretize the PDEs in these sub domain to form an approximate algebraic system. And then, we can obtain the approximate solution of original PDEs by solving the algebraic system numerically. Here meshing is the technique to divide the computational region.

8.1 Semiconductor Physical Model and Numerical Model

Since GSS is a two dimension numerical simulation software, for real simulation work, we have to simplify the device to get the two dimension model. From physical entity to numerical model, we need to abstract device's main factors and neglect the unimportant factors. This process needs knowledge and experience, normally requires expertise. For users, numerical model for most of the devices exist already and can be used as a template.

Physical Device and Numerical Model

Figure (8.1) shows the SOI-CMOS transistors in large scale integration circuit [31]. If we want to use two dimension semiconductor software to do the simulation, one of the transistors should be picked out, and be simplified to two dimensional planar structure as shown in Figure (8.2).

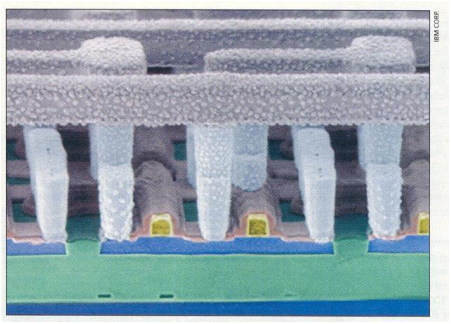


Figure 8.1: Real SOI-CMOS transistor

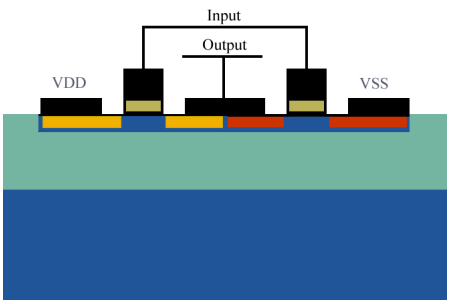


Figure 8.2: SOI-CMOS transistor calculation model

8.2 Semiconductor Simulation's Request on Mesh

After obtain the numerical device model, we have to mesh it. The method for meshing decides the following possible numerical processes. And the mesh quality decides the convergence speed and even whether converge or not[1].

Structured Mesh

There are two kinds of mesh: structured mesh and unstructured mesh. Structured

[1] See appendix: How to conquer convergence problems.

mesh has (or can be projected to) regular mesh lines in cartesian coordinate system, shown as Figure (8.3). the advantage of structured mesh is that it is easy to be generated and one can use simple yet efficient finite difference method on it. The disadvantage is that the mesh is not flexible. It is difficult to fit the device with complicate boundary shape with structured mesh. And mesh densification often leads to surplus mesh nodes.

Unstructured Mesh

Unstructured mesh allows mesh point to be arranged disorderly. In two dimension condition, it is obtained from triangle or quadrangle discretion, shown as Figure (8.4). Unstructured mesh can conserve the complicate boundary shape. Another advantage is the mesh can be locally refined at some critical region, and keeps coarse for unimportant area. However unstructured mesh is more difficult to realize than structured mesh. Further more, only finite volume or finite element method can be used, for which the memory requirement is higher than finite difference method.

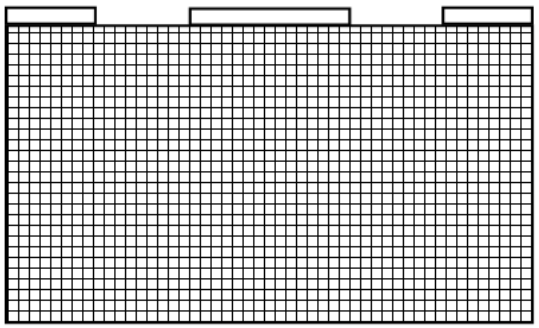


Figure 8.3: Structured mesh

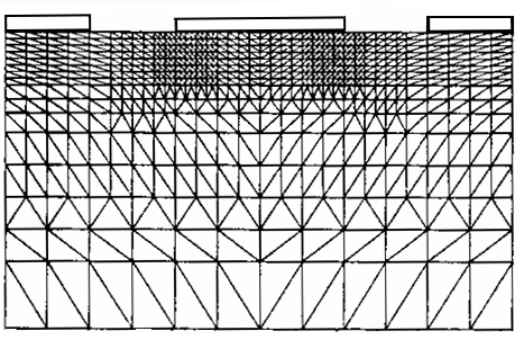


Figure 8.4: Unstructured mesh

From Structured Mesh to Unstructured Mesh

In early days, two dimension semiconductor numerical simulation software only support simple diode, bipolar transistor devices and etc. The geometry of all those devices can be considered as simple rectangle. Accordingly, structured mesh can discretize the whole region efficiently. However as the semiconductor process develops, MOS structure turns to be the main stream device. Unfortunately it is not so easy to build a MOS structure. It is composed of silicon, silicon dioxide insulation and electrode regions. Further more, the gate silicon dioxide boundary has bird peak like shape due to process concerns. In this case, structured mesh can not describe the complicate device structure.

Besides, the accuracy of semiconductor numerical simulation is strongly dependent on mesh construction. Obviously more dense mesh should bring to more accuracy. However the calculation loading will increase square proportional to mesh nodes

number[2]. In order to solve this conflict, we can use dense mesh in the area we concerned and coarse mesh in the area we don't care much. Generally in the region where potential and carriers have high gradients, we need dense mesh (See"??", on page ??), whereas in smooth region, we can use coarse mesh. So that we could balance both accuracy and computational efficiency.

Accordingly unstructured mesh is generally used in semiconductor numerical simulation. Several famous software, eg. PISCES, MINIMOS and etc. all adoped triangle based unstructured mesh. GSS also supports triangle mesh.

Delaunay Mesh

There are some different methods to generate unstructured mesh, The most popular method is called Delaunay method [32] [33]. This method is very adaptive to complicate boundary, and can generate high quality triangles. However, Delaunay generated mesh is isotropic. Some physical processes, such as thermal conduction and diffusion, do not dependent on direction, are situate to be discretized on this kind of unstructured mesh.

Quadtree Mesh

In semiconductor device, since the current has certain fixed direction, if the mesh node direction follows the current direction, we can decrease the split error [34] which cased by direction mismatch. By using unstructured mesh generate by Quadtree technology, the above requirement can be satisfied [35]. This technology turns to be the main stream in semiconductor numerical simulation. The disadvantage of Quadtree technology is its low adaptiveness to complicate boundary conditions. It often generates triangle with bad quality at curved boundary.

Figure (8.5) and Figure (8.6) show the mesh of NMOS transistor generated by SGFramework and MEDICI, which based on Delaunay Quadtree method, respectively.

Triangle Mesh Generator: the Tradeoff

GSS uses Triangle [1] as its mesh generator, which is based on Delaunay method. Accordingly the mesh in GSS is basically isotropic and has good boundary fitting. However, with assistant mesh lines, we can give certain order to the mesh nodes and obtain mesh topology between Delaunay and Quadtree meshes.

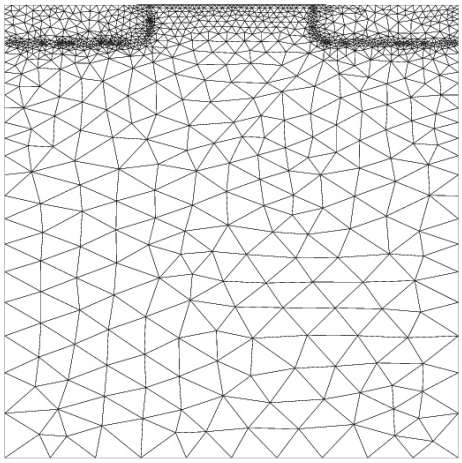


Figure 8.5: Isotropic unstructured mesh, generated by SGFramework

8.3 GSS Mesh Data Structure

Unstructured mesh needs smart data structure and algorithm to manage. In whole GSS development, many efforts are spent on developing efficient and stable

[2] Please read appendix: GSS's memory and CPU request.

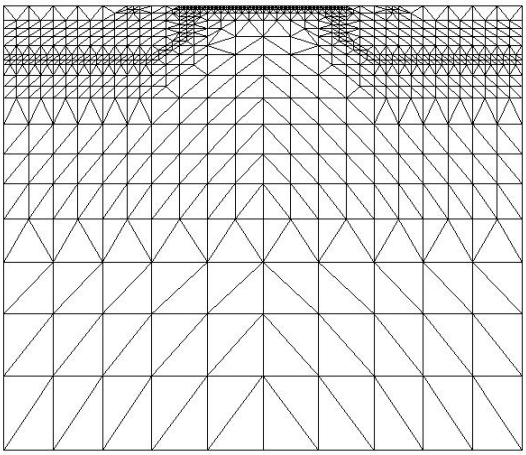


Figure 8.6: Quadtree unstructured mesh, generated by MEDICI

internal mesh functions.

Voronoi Diagram

After mesh generation, we have the computational region be divided into triangles. In order to implement the finite volume method (FVM), GSS needs to build Voronoi diagram from triangle mesh since FVM uses Voronoi diagram as its control volume. The Voronoi diagram is the regions surrounded by each edge's vertical half separation line. Accordingly, the triangles and its Voronoi diagram overlap each other, shown in Figure (8.7).

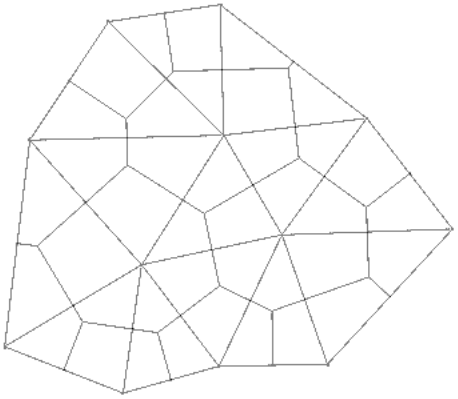


Figure 8.7: Voronoi unit and triangle dual mesh

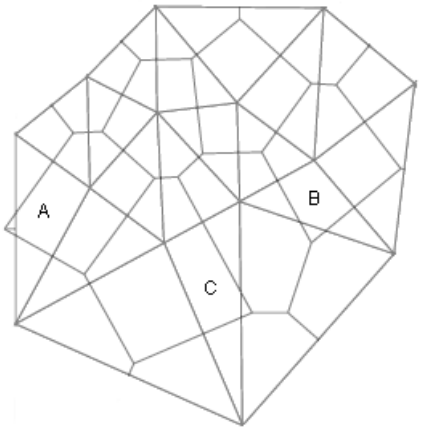


Figure 8.8: Voronoi unit contains obtuse triangle

Special Process of Obtuse Triangle

Since the vertex of Voronoi diagram is the circumcircle's center of certain triangle, here is a special case for obtuse triangle. For acute or vertical triangle, the center of its circumcircle is inside it (or locate at its edge), which has no problem. However for obtuse triangle the circumcircle's center is outside, shown as triangle A, B and C in Figure (8.8). For this instance, the distance between the center to corresponding edge of the obtuse triangle should be negative, shown as $d3$ in Figure (8.10), to keep the consistent of Voronoi diagram to triangle mesh.

Mesh Quality

The mesh quality is another thing we should concern. A triangle with obtuse angle or very shape acute angle, will lead to large numerical error in PDE discretion, which should be avoid. Fortunately, the mesh generator use by GSS is very

outstanding, which can guarantee that there is no obtuse triangle on the boundary and the minimum angle is larger than 20°. Whereas quadtree method often generates a lot of obtuse triangles near the boundary, which affects the accuracy of numerical simulation.

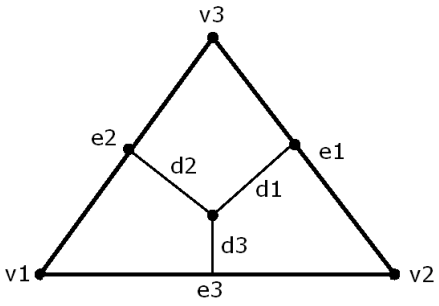


Figure 8.9: Triangle unit

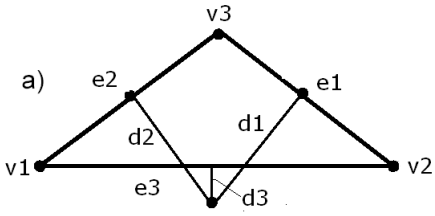


Figure 8.10: Obtuse triangle unit

Mesh Data Structure

The topology of unstructured mesh is mainly composed of nodes, triangles and Voronoi cells. The node is a geometry point plus the boundary mark. Here is the brief of node data:

```
struct   Node          // node
{
  double  x,y;         // x and y coordinate
  int     bc_index;    // the index to boundary of the node
  int     zone_index;  // the zone index;
};
```

Triangle data includes vertex index to node, length of three edges, the degree of three angles, circumcircle's center position, the distance from circumcircle's center to each edge[3], triangle area and boundary condition and so on.

```
struct Tri             // triangle
{
  int     node[3];     // three nodes, local index
  double  edge_len[3]; // the length of 3 edges: a,b and c
  double  angle[3];    // the degree of 3 angles: A, B and C
  double  xc,yc;       // the location of circle center
  double  d[3];        // the distance from circumcircle center to 3 edges
  double  s[3];        // the area of region separated by da, db and dc
  int     bc[3];       // the boundary condition index of 3 edge
  double  area;        // the area of triangle
  int     zone_index;  // the zone index
};
```

The data structure for Voronoi cell is the most complex, since neighbor information should be stored here. A Voronoi cell should know how many neighbors it has, and the geometrical information about each of its neighbors.

```
struct VoronoiCell     // VoronoiCell
{
  double  x,y;         // location of cell center
  int     nb_num;      // the number of neighbors
  int     *nb_array;   // the index of neighbor nodes
  int     *inb_array;  // inverse index of neighbor nodes
  double  *elen;       // the array of the length of cell'e boundary edge
  double  *ilen;       // the array of the distance to it's neighbor nodes
  double  *angle;      // the angle of it's neighbor to horizontal line
  double  area;        // the area of cell
  int     *celledge;   // the cell's boundary edge index array
  int     bc_index;    // the index to boundary of the node
};
```

[3] One should care with obtuse triangle. The distance from circumcircle's center to the longest edge should be negative.

After building the data structure for triangles and Voronoi cells, we are ready to discrete PDEs on the mesh.

Mesh Node Reordering

Although the order of node for an unstructured mesh is not important from topological view, which means by exchanging the order of two different nodes will not affect the topology structure. However reasonable order can decrease the band width of matrix generated by later PDE discretization. The decrease of the band width of matrix can reduce the non zero filling for LU factorization. For Krylov space iteration method, it helps to improve the efficiency of pre- conditioner, so as to accelerate the solving.

GSS use breadth first searching algorithm to reorder the mesh nodes. After reordering, the band width of matrix decreases sharply. Figure (8.11) and Figure (8.12) show the before and after reordering band width of the matrix generated by problem "??", on page ??.

The basic step of breadth first searching algorithm is shown below:

```
1.  Calculate mesh's topology structure,
    initialize the list, clean node visit mark

2.  Assign left bottom node's order to be 0,
    insert it into list end, set it to be visited

3.  When (list is not empty)
    {
      Eject the list's top node as current node, order increase 1
      Look for current node's neighbor node
      If(neighbor node is not visited before)
      {
         Mark this node as visited, insert it to list end
      }
    }

4.  Update all node's index
```

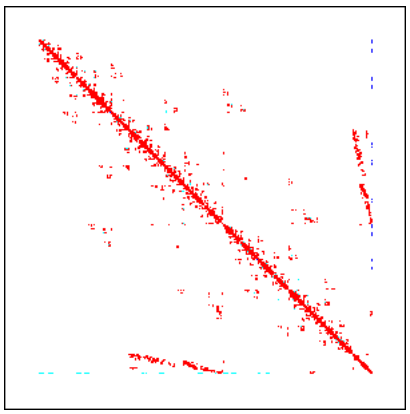


Figure 8.11: Initial matrix band width

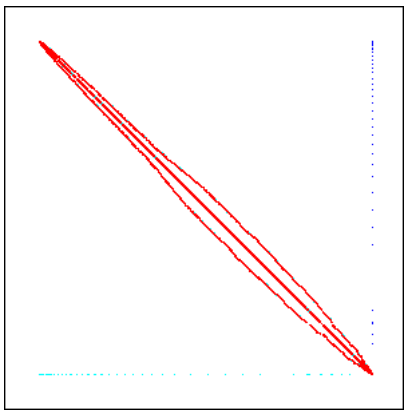


Figure 8.12: Reordered matrix band width

8.4 Finite Volume Discretion of Derivative Operator

8.4.1 Gradient of Scaler Field

Assume there is scaler Φ defined at the field. We are going to discretize $\nabla\Phi$ on mesh shown as Figure (8.13). Generally, numerical method needs only gradient along the edge of triangle, which is defined at the center of the edge:

$$\left.\frac{\partial\Phi}{\partial r}\right|_{m,n} = \frac{\Phi_m - \Phi_n}{|\mathbf{r}_m - \mathbf{r}_n|} \tag{8.1}$$

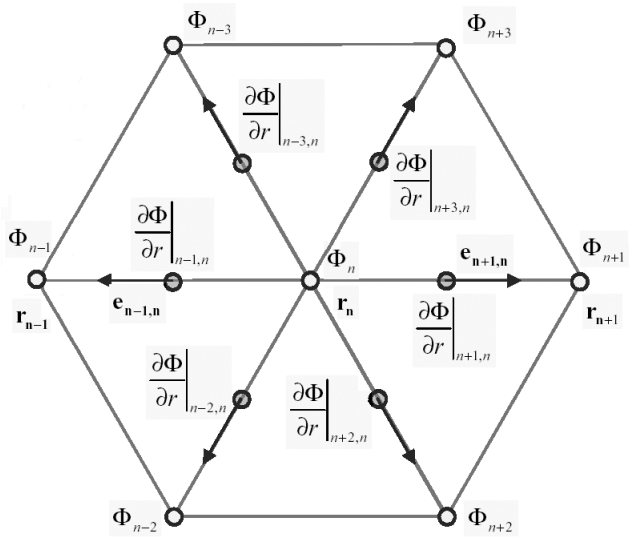


Figure 8.13: Directional gradients' FVM discretion

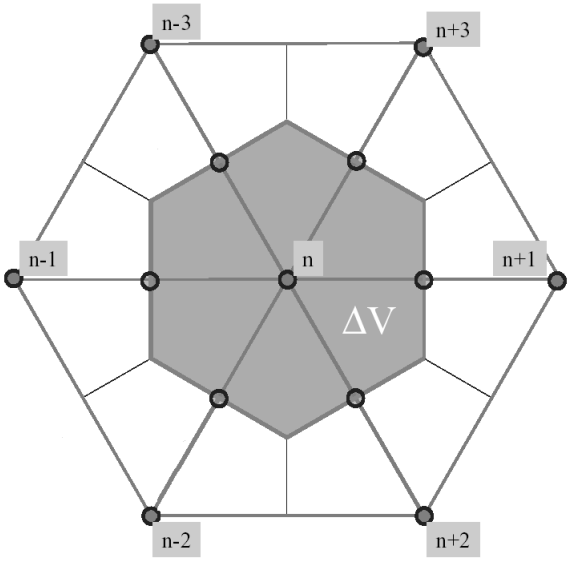


Figure 8.14: FVM discretion of $\nabla\Phi$

However, for certain instance, we need to calculate $\nabla\Phi$ at the node n. For example, when Φ represents the electrostatic potential, and we needs to get electric field $\mathbf{E} = -\nabla\Phi$ at the node n. Here $\nabla\Phi$ can be given by Green-Gauss formulae or least squares method.

Construct Gradient by Green-Gauss Formulae

Start from Green-Gauss formulae, the average value of $\nabla\Phi$ over a control volume can be illustrated as:

$$\Phi_x = \frac{1}{\Delta V}\int_{\Delta V}\Phi_x dV = \frac{1}{\Delta V}\oint \Phi dy \tag{8.2}$$

$$\Phi_y = \frac{1}{\Delta V}\int_{\Delta V}\Phi_y dV = \frac{1}{\Delta V}\oint \Phi dx \tag{8.3}$$

where ΔV represents the area of the two-dimensional control volume, shown as Figure (8.14). The contour integral along the face of control volume can be discretized to

$$\Phi_x = \frac{1}{\Delta V}\sum_{m=1}^{N}\frac{\Phi_n + \Phi_m}{2}\Delta y_{nm} \tag{8.4}$$

$$\Phi_y = \frac{-1}{\Delta V}\sum_{m=1}^{N}\frac{\Phi_n + \Phi_m}{2}\Delta x_{nm} \tag{8.5}$$

where m is n's neighbor node, Δx_{nm} and Δy_{nm} denote the increments of x and y along the control volume face. If Φ is linear function at the region, Green-Gauss formulae can get exact result.

Construct Gradient by Least-squares Construction

The least-squares gradient construction is obtained by minimize the sum of the squares of the differences between neighboring values $m = 1, N$ and values extrapolated from the node n under consideration to the neighboring locations:

$$\sum_{m=1}^{N}\omega_{nm}^2 E_{nm}^2 \tag{8.6}$$

Error function E can be given by

$$E_{nm}^2 = (-d\Phi_{nm} + \Phi_x dx_{nm} + \Phi_y dy_{nm})^2 \tag{8.7}$$

where $d\Phi_{nm} = \Phi_m - \Phi_n$, $dx_{nm} = x_m - x_n$, $dy_{nm} = y_m - y_n$, ω is a weighting factor. When the partial derivative is 0, Equation (8.6) reaches its minimum:

$$\frac{\partial \sum_{m=1}^{N}\omega_{nm}^2 E_{nm}^2}{\partial \Phi_x} = 0 \tag{8.8}$$

and

$$\frac{\partial \sum_{m=1}^{N}\omega_{nm}^2 E_{nm}^2}{\partial \Phi_y} = 0 \tag{8.9}$$

through straight-forward algebra, the two formulae above can be written as:

$$a_n\Phi_x + b_n\Phi_y = d_n \tag{8.10}$$
$$b_n\Phi_x + c_n\Phi_y = e_n \tag{8.11}$$

where a_n, b_n and c_n can be expressed as:

$$a_n = \sum_{m=1}^{N} \omega_{nm}^2 \mathrm{d}x_{nm}^2 \tag{8.12}$$

$$b_n = \sum_{m=1}^{N} \omega_{nm}^2 \mathrm{d}x_{nm}\mathrm{d}y_{nm} \tag{8.13}$$

$$c_n = \sum_{m=1}^{N} \omega_{nm}^2 \mathrm{d}y_{nm}^2 \tag{8.14}$$

$$d_n = \sum_{m=1}^{N} \omega_{nm}^2 \mathrm{d}\Phi_{nm}\mathrm{d}x_{nm} \tag{8.15}$$

$$e_n = \sum_{m=1}^{N} \omega_{nm}^2 \mathrm{d}\Phi_{nm}\mathrm{d}y_{nm} \tag{8.16}$$

Since parameter a_n, b_n and c_n are only related to mesh topology structure, they can be pre-computed for saving CPU time when this gradient needs to be evaluated many times.

Note that the determinant of this system is given by:

$$\mathrm{DET} = \mathrm{ac} - \mathrm{b}^2 \tag{8.17}$$

For $\omega_{nm} = 1$, the determinant corresponds to a difference in quantities of the order $O(\mathrm{d}x^4)$, which may lead to ill-conditioned systems. However if inverse distance weighting $\omega_{nm}^2 = 1/(\mathrm{d}x_{nm}^2 + \mathrm{d}y_{nm}^2)$ is used, then DET scales to $O(1)$, condition number is much better.

After having $\nabla\Phi$, the gradients along triangle edge can be redefined as below, however it is rarely used.

$$\left.\frac{\partial\Phi}{\partial r}\right|_{m,n} = \frac{\mathbf{r}_m - \mathbf{r}_n}{|\mathbf{r}_m - \mathbf{r}_n|}\nabla\Phi \tag{8.18}$$

8.4.2 Vector field's divergence

Assume discretize the divergence of vector field $\mathbf{A}$ on the cell shown as Figure (8.15). The average integral value of $\nabla \cdot \mathbf{A}$ over control volume can be expressed by Gauss theory as:

$$\frac{1}{\Delta V}\int_V \nabla \cdot \mathbf{A}\mathrm{d}V = \frac{1}{\Delta V}\oiint_{\Delta S} \mathbf{A} \cdot \mathbf{dS} \tag{8.19}$$

Sum the contour integral and discretize it, we have

$$\frac{1}{\Delta V}\int_V \nabla \cdot \mathbf{A}\mathrm{d}V = \frac{1}{\Delta V}\oiint_{\Delta S} \mathbf{A} \cdot \mathbf{dS} \approx \frac{1}{\Delta V}\sum_{m=1}^{N} \mathbf{A}_{m,n} \cdot \mathbf{e}_{m,n}\mathrm{d}S \tag{8.20}$$

where $\mathrm{d}S$ is the outer face of the control volume, $\mathbf{e}_{m,n}$ is the face normal vector, $\mathbf{A}_{m,n}$ is the value of $\mathbf{A}$ at the face of the control volume. Generally, $\mathbf{A}_{m,n} \neq (\mathbf{A}_m + \mathbf{A}_n)/2$, different problem needs different numerical scheme to construct $\mathbf{A}_{m,n}$. If we need to calculate fluid dynamic problem, we need some complicated reconstruction i.e. Roe scheme. For semiconductor problem, normally we adopt Scharfetter-Gummel discretion scheme.

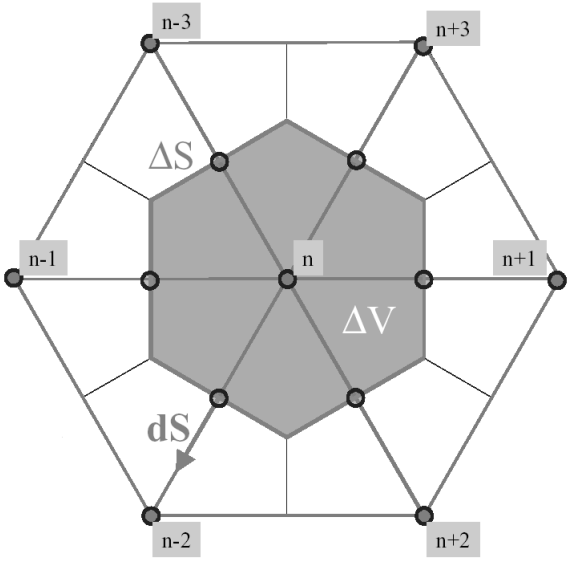


Figure 8.15: Divergence's FVM discretion

Especially, if vector $\mathbf{A}$ can be illustrated as scaler field's gradients $\mathbf{A} = \nabla\Phi$, then the discretion of Laplas operator $\nabla^2 = \nabla \cdot \nabla$ can be written as

$$\frac{1}{\Delta V}\int_V \nabla\cdot\nabla\Phi dV \approx \frac{1}{\Delta V}\sum_{m=1}^{N}\nabla\Phi\Big|_{m,n}\cdot\mathbf{e}_{m,n}dS \approx \frac{1}{\Delta V}\sum_{m=1}^{N}\frac{\partial\Phi}{\partial r}\Big|_{m,n}dS \tag{8.21}$$

8.4.3 Vector field's curvature

Assume vector $\mathbf{A}$ is defined in the region, only consider two dimension problem, for electromagnetic example, magnetic vector $\mathbf{H}$ with two dimensional TM mode, where $\mathbf{H}_z$ is 0.

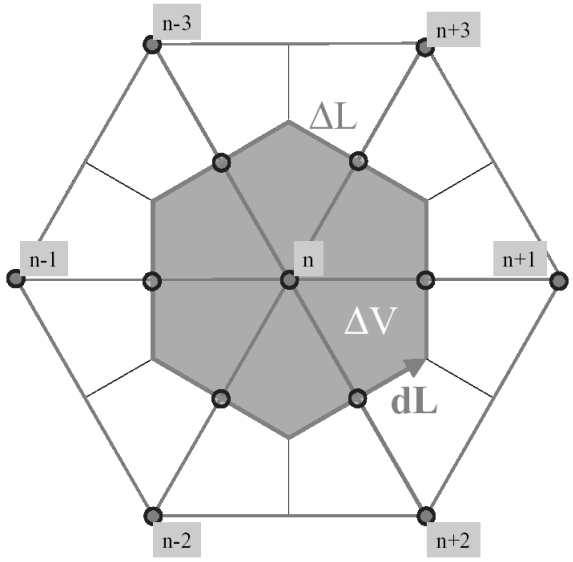


Figure 8.16: Curling's FVM discretion

Over control volume ΔV in Figure (8.16), the curling of $\mathbf{A}$ can be discretized as

$$\frac{1}{\Delta V}\int_V \nabla\times\mathbf{A}\Big|_n \cdot \mathbf{e}_z \mathrm{d}V = \frac{1}{\Delta V}\oint_{\Delta L}\mathbf{A}\cdot\mathbf{dL} \approx \frac{1}{\Delta V}\sum_{m=1}^{N}\mathbf{A}_{m,n}\cdot\mathbf{e}_{\perp,m,n}\Delta L_m \tag{8.22}$$

where $\mathbf{dL}$ represents the vector along the face of control volume. $\mathbf{e}_{\perp,m,n}$ is the tangential unit vector along the face. ΔL_m is the length of *mth* face. $\mathbf{A}_{m,n}$ is the value of $\mathbf{A}$ at the face. Simultaneously we should use some complicate reconstruction technique.

Chapter 9 Numeric Method of Drift-Diffusion Model

9.1 Variable Scaling

For the numerical simulation of semiconductor, variable quantity can be differed much. For example, electrostatic potential is normally several V, but carrier density is normally at 1×10^{18} cm^{-3} level. Without preprocessing, we can not get accuracy solution when Poisson's equation combined with Carrier continuous equation. Since the computer double type floating number has only 16 digit number, when the potential and carrier density difference is more than 16 magnitudes, the variation of potential will disappear in the huge number of carrier variation. In this case we need to scale the variables.

Although variables non-unitization is the normal method, GSS tries a new method. Semiconductor simulation involves many parameters. Most of them have units, by using non-unitization is not straightforward, and very easy to make mistakes. So I leave the unit and use a set of basic unit to replace the original unit.

There are many physical units. However we only need 5 independent units actually. International Unit system (SI) selects mechanical length, mass, time, and electrical current strength and thermal temperature as its basic units. GSS select length, time, potential, charge quantity and temperature as basic unit. Other units are leaded units.

Inside GSS, physical variables still take units such as cm and s. However, the unit has its own definition, shown below:

	Unit	Unit define value	Annotation
Basic Unit	cm	$\max^{-1/3}$(Doping)	Length unit can be re-defined by user
Basic Unit	s	10^{12}	Using pico second as the basic unit
Basic Unit	V	1.0	Same as SI definition
Basic Unit	C	$1.0/e$	Electron charge as 1
Basic Unit	K	1.0/300	Use 300K as basic temperature unit
Leaded Unit	m	100 cm	
Leaded Unit	μm	1×10^{-4} cm	
Leaded Unit	J	C · V	
Leaded Unit	kg	J/m^2 · s^2	J = C · V = kg · m^2/s^2
Leaded Unit	eV	1.0 · V	Electron charge as 1
Leaded Unit	A	C/s	
Leaded Unit	mA	1×10^{-3} A	

In this case, semiconductor control equation and parameters does not change their physical formula, eg. carrier density can still be written as $n = 1.0 \times 10^{18}$ cm^{-3}, but by selecting max(Doping) $= 1.0 \times 10^{18}$ cm^{-3}, we can let the value of n be 1.0.

In GSS material parameter database, a lot of parameters are given with natural unit, which leads to efficient compilation.

9.2 FVM Discretion of Poisson's Equation

Before solving DDM equations, here introduce finite volume solution of Poisson's equation on two dimensional mesh first. The following equation is going to be solved:

$$\nabla \cdot \varepsilon \nabla \psi = -\rho \tag{9.1}$$

Boundary condition includes the first, the second, the third type and interface between different materials.

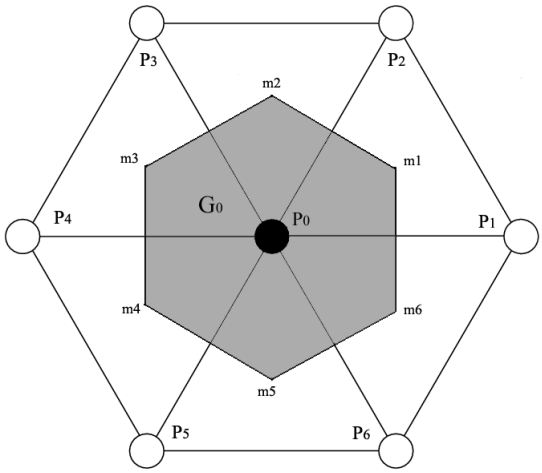


Figure 9.1: Poisson equation internal points discretion

Discretion of Internal Point

Suppose P_0 is the internal point of a Voronoi cell G_0, shown as Figure (9.1), $P_1 \sim P_6$ and P_0 are the neighbors, $m_1 \sim m_6$ is triangle $P_0P_iP_{i+1}$ (Suppose $P_7 = P_1$, same below)'s surrounding circle center. Integrating over G_0, we get

$$\iint_{G_0} \nabla \cdot \varepsilon \nabla \psi \, \mathrm{dV} = -\iint_{G_0} \rho \, \mathrm{dV} \tag{9.2}$$

by using Green's formula, above Equation (9.2) can be rewritten as

$$\int_{\partial G_0} \varepsilon \nabla \psi \cdot \mathrm{d}\mathbf{S} = -\iint_{G_0} \rho \, \mathrm{dV} \tag{9.3}$$

where ∂G_0 is G_0's boundary. We notice:

$$\int_{\partial G_0} \varepsilon \nabla \psi \cdot \mathrm{d}\mathbf{S} = \sum_{i=1}^{6} \int_{\overline{m_i m_{i+1}}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} \approx \sum_{i=1}^{6} \frac{\overline{m_i m_{i+1}}}{\overline{P_0 P_{i+1}}} \varepsilon_{i+1} \left[\psi\left(P_{i+1}\right) - \psi\left(P_0\right)\right] \tag{9.4}$$

Among them, n is ∂G_0's normal vector, ε_{i+1} is the value of ε along the face of control volume $\overline{m_i m_{i+1}}$, for non-uniform material we can use linear interpolation to get it:

$$\varepsilon_{i+1} = \frac{\varepsilon(P_0) + \varepsilon(P_{i+1})}{2} \tag{9.5}$$

In most of the situations, we can consider charge density ρ does not charge a lot in control volume, accordingly the right hand side integral of Formulae (9.3) can be written as

$$-\iint_{G_0} \rho \, \mathrm{dV} \approx -\rho V(G_0) \tag{9.6}$$

where $V(G_0)$ is the control volume G_0's area.

By above discussing, the finite volume discrete scheme at internal point P_0 can be written as:

$$\sum_{i=1}^{6} \frac{\overline{m_i m_{i+1}}}{\overline{P_0 P_{i+1}}} \varepsilon_{i+1} \left[\psi(P_{i+1}) - \psi(P_0)\right] = -\rho V(G_0) \tag{9.7}$$

Discretion of Boundary Point

Then we need to consider finite volume discrete about boundary point, shown as Figure (9.2):

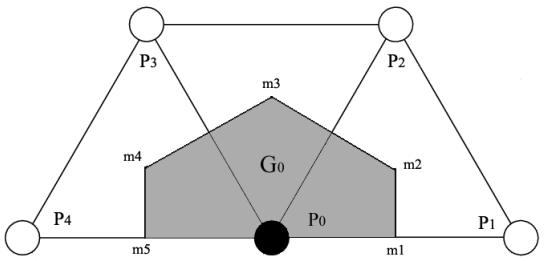


Figure 9.2: Poisson equation boundary points discretion

Suppose P_0 is the boundary point, now control volume is $P_0 m_1 m_2 m_3 m_4 m_5 P_0$. If P_0 follows first type boundary condition, for example $\psi = \psi_0$, then we can simply force

$$\psi(P_0) = \psi_0 \tag{9.8}$$

If P_0 follows second or third boundary condition type, for example

$$\frac{\partial \psi}{\partial n} + k\psi = r \tag{9.9}$$

we need go back to Green's formulae. Similar as Equation (9.3), the discretion equation of boundary point P_0 can be written as

$$\int_{\partial G_0} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} = \sum_{\mathrm{i}=1}^{4} \int_{\overline{\mathrm{m_i m_{i+1}}}} \varepsilon_{\mathrm{i}+1} \frac{\partial \psi}{\partial \mathrm{n}} \mathrm{dS} + \int_{\overline{P_0 m_1}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} + \int_{\overline{P_0 m_5}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} \tag{9.10}$$

The first item of right hand side of Equation (9.10) can be discreted as the same mentioned by Equation (9.4). For the last two items, we need to start from boundary condition Equation (9.9). We obtain the integral only contains ψ by

removing the normal vector derivatives. Suppose k and r are constant, ψ on $\overline{P_0m_1}$ and $\overline{P_0m_5}$'s linear distribution can be integrated with trapezoid formulae calculation, then we have

$$\begin{aligned}\psi(m_1) &= \frac{\psi(P_0)+\psi(P_1)}{2}\\ \psi(m_5) &= \frac{\psi(P_0)+\psi(P_4)}{2}\end{aligned} \tag{9.11}$$

Integrate through $\overline{P_0m_1}$ can be written as

$$\begin{aligned}\int_{\overline{P_0m_1}} \varepsilon\frac{\partial\psi}{\partial n}\mathrm{dS} &= \int_{\overline{P_0m_1}} \varepsilon(r-k\psi)\,\mathrm{dS}\\ &= \overline{P_0m_1}\varepsilon(P_0)\left[r-k\frac{\psi(P_0)+\psi(m_1)}{2}\right]\\ &= \overline{P_0m_1}\varepsilon(P_0)\left[r-k\frac{3\psi(P_0)+\psi(P_1)}{4}\right]\end{aligned} \tag{9.12}$$

similarly

$$\int_{\overline{P_0m_5}} \varepsilon\frac{\partial\psi}{\partial n}\mathrm{dS} = \overline{P_0m_5}\varepsilon(P_0)\left[r-k\frac{3\psi(P_0)+\psi(P_4)}{4}\right] \tag{9.13}$$

Put Equation (9.12) and Equation (9.13) in Equation (9.10), we get the finite volume discrete equation at the boundary points.

Especially, if the boundary is Neumann type:

$$\frac{\partial\psi}{\partial n} = 0 \tag{9.14}$$

Obviously, boundary does not need to be processed. During programming, this kind of boundary point can be processed the same as internal point.

Discretion of Interface Point

If we meet the interface of two different materials as show in Figure (9.3), where P_0, P_1, P_4 are the points located on the interface, P_0's control volume lies on both materials. s_1, s_2 are the center points of $\overline{P_0P_1}$ and $\overline{P_0P_4}$. Interface $\overline{s_1s_2}$ separate

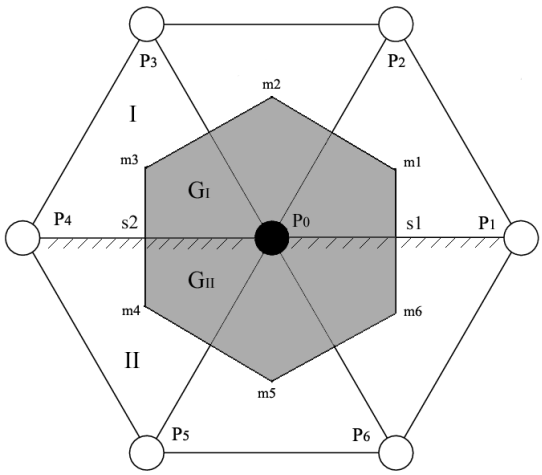


Figure 9.3: Boundary points' Poisson equation discretion

control volume G as G_I and G_{II} parts. By integrating at both G_I and G_{II}, we get

$$\int_{\overline{s_1 m_1 m_2 m_3 s_2}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} + \int_{\overline{s_2 s_1}} \varepsilon_{\mathrm{I}} \frac{\partial \psi}{\partial \mathrm{n}} \mathrm{dS} = - \iint_{\mathrm{G_I}} \rho \mathrm{dV} \tag{9.15}$$

$$\int_{\overline{s_2 m_4 m_5 m_6 s_1}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} + \int_{\overline{s_1 s_2}} \varepsilon_{\mathrm{II}} \frac{\partial \psi}{\partial \mathrm{n}} \mathrm{dS} = - \iint_{\mathrm{G_{II}}} \rho \mathrm{dV} \tag{9.16}$$

From above circle integrals, the interface $\overline{s_1 s_2}$ is integrated twice, with dielectric constant ε selected with region G_I and G_{II}'s value, respectively. Sum up two formulae above, and use the relationship at interface:

$$\varepsilon_I \frac{\partial \psi}{\partial n} - \varepsilon_{II} \frac{\partial \psi}{\partial n} = \sigma \tag{9.17}$$

we have

$$\int_{\partial G_I + \partial G_{II}} \varepsilon \frac{\partial \psi}{\partial n} \mathrm{dS} + \sigma \overline{s_1 s_2} = - \iint_{G_I + G_{II}} \rho \mathrm{dV} \tag{9.18}$$

The following things are simple, Formulae (9.4) can be used here to do the discretion.

Obviously, the discretion of Poisson's equation on unstructured mesh is very convenient. Boundary condition and material interface can be processed in flexible way. As soon as meshing supporting function be well designed, discretion process is easier than finite difference method[1].

The finite volume discrete at last form a closed linear algebra equation set $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is a band sparse coefficient matrix, $\mathbf{x}$ is the solution vector of ψ, right hand side vector $\mathbf{b}$ is formed by source ρ and some boundary item. Regarding numerical solution to linear equations, please refer to later chapters.

9.3 Numerical Scheme of 1D DDM Equations

In this section, we are going to introduce numerical discretion scheme of DDM equations in one dimensional. This can be a good fundamental about the actual numerical arithmetic used in GSS, since although it is one dimension condition, most of the concepts can be put into high dimension in a straightforward way.

Basic DDM Governing Equation in 1D

Suppose ψ, n, p as basic variables, semiconductor material is uniform, and temperature inside is uniform and unchanged, DDM equations in 1D can be described as:

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{q}{\varepsilon}(N_D - N_A + p - n) \tag{9.19}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - R \tag{9.20}$$

$$\frac{\partial p}{\partial t} = \frac{-1}{q}\frac{\partial J_p}{\partial x} - R \tag{9.21}$$

$$J_n = q\mu_n\left(-n\frac{\partial \psi}{\partial x} + V_T\frac{\partial n}{\partial x}\right) \tag{9.22}$$

$$J_p = q\mu_p\left(-p\frac{\partial \psi}{\partial x} - V_T\frac{\partial p}{\partial x}\right) \tag{9.23}$$

where $V_T = \dfrac{k_b T}{q}$ is the thermal electrical constant.

Mesh Discretion in 1D

First we us N points to discrete calculation region into $N-1$ parts, leading mesh

[1] The boundary discretion is a complex and boring work in finite difference method, since it not only dependent on physical boundary condition but also where is the boundary.

distance to Δx, as shown in Figure (9.4).

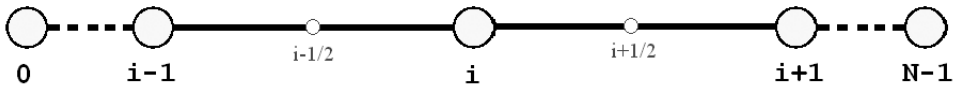


Figure 9.4: One dimension discrete mesh points

The Discreted DD Equation at Point i

For simple illustration, here we only consider steady-state problem, which means time partial difference item is zero. At mesh point i, let Poisson's equation discretized with three-point-center-difference, continuous equation can be descretized by using current value at half point $i-1/2$ and $i+1/2$, yet we get:

$$\frac{\psi_{i-1}-2\psi_i+\psi_{i+1}}{\Delta^2 x} = -\frac{q}{\varepsilon}\left(N_D - N_A + p_i - n_i\right) \tag{9.24}$$

$$\frac{1}{q}\frac{J_{n,i+1/2}-J_{n,i-1/2}}{\Delta x} - R_i = 0 \tag{9.25}$$

$$\frac{-1}{q}\frac{J_{p,i+1/2}-J_{p,i-1/2}}{\Delta x} - R_i = 0 \tag{9.26}$$

where $R_i = R(n_i, p_i)$ is the carrier recombination item only contains carrier density. Since the 1D discretion of Poisson's equation by three-point-center-difference is a canonical method, we don't want to explain it here. As a result, the key of DDM numerical discretion is to get half point current at $i+1/2$ and $i-1/2$.

Half Point Current

Based on Equation (9.22), electron current at half point $i+1/2$ is given by:

$$J_{n,i+1/2} = q\mu_{i+1/2}\left(-n_{i+1/2}\frac{\psi_{i+1}-\psi_i}{\Delta x} + V_T \nabla n\Big|_{i+1/2}\right) \tag{9.27}$$

Linear Formula of Half Point Current

A straightforward linear difference method can be written as:

$$n_{i+1/2} = \frac{n_i + n_{i+1}}{2} \tag{9.28}$$

$$\nabla n\Big|_{i+1/2} = \frac{n_{i+1}-n_i}{\Delta x} \tag{9.29}$$

The pity is numerical experiment shows this simple discretion format has strong confinement. It can only be stable used when drift current is much smaller than diffusion current. In another word, this format can only be used for small mesh distance, weak electric field and low carrier concentration condition.

Scharfetter-Gummel Formula of Half Point Current

In order to give stable solution at high drift current, we usually use Scharfetter-Gummel scheme to discretize semiconduct current equation. This scheme can be naturally leaded from the equation of electron's current in region $[x_i, x_{i+1}]$:

$$J_n = q\mu\left(n(x)E + V_T\frac{dn}{dx}\right) \tag{9.30}$$

where $E = -\frac{\psi_{i+1}-\psi_i}{\Delta x}$ is the electrical field intensity. Suppose in the region, except carrier density, other quantity including electric intensity, carrier mobility and current intensity keep unchanged, the previous formulae can be re-written to electron density related ordinary differential equation:

$$\frac{dn}{dx} + \frac{E}{V_T}n = C_0 \tag{9.31}$$

Attention, current J_n here is unknown, and sum into constant value C_0. Solving this ordinary differential equation symbolically, we have

$$n(x) = \frac{V_T}{E} C_0 + C_1 \exp\left(-\frac{E}{V_T} x\right) \tag{9.32}$$

where C_0 and C_1 are constant coefficient. Put electron density at point i and $i+1$ as boundary condition into Equation (9.32), we can solve C_0 and C_1 out. Then we have electron density distribution at $[x_i, x_{i+1}]$:

$$n(x) = n_i \left[1 - g(x)\right] + n_{i+1} g(x) \tag{9.33}$$

where

$$g(x) = \frac{1 - \exp\left(\frac{\psi_{i+1} - \psi_i}{V_T} \frac{x - x_i}{\Delta x}\right)}{1 - \exp\left(\frac{\psi_{i+1} - \psi_i}{V_T}\right)} \tag{9.34}$$

is the growth function.

After having the analytic carrier density distribution Equation (9.33), we can get the carrier density n at half point $i+1/2$ as well as density gradients ∇n there. The discretion of electron current at the half point can then be got by substitution n and ∇n into Equation (9.27).

By using the same process, we can obtain discretion scheme for hole current. However, some simpler method exists here.

Pay attention to hole current Equation (9.23), using the following formulae to replace

$$\begin{cases} J_n \to J_p \\ n \to p \\ \mu_n \to \mu_p \\ k_b \to -k_b \end{cases}$$

then the equation will be the same as electron current Equation (9.22). Accordingly, the discretion formula of hole current can be got by simple replacement to the discretion formula of electron current.

For simple illustration, we introduce two assistant function, their function plot is shown below Figure (9.5) and Figure (9.6):

$$\text{aux1}(x) = \frac{x}{\sinh(x)} \tag{9.35}$$

$$\text{aux2}(x) = \frac{1}{1 + e^x} \tag{9.36}$$

Half point carrier density, carrier gradient and current can be expresses by this

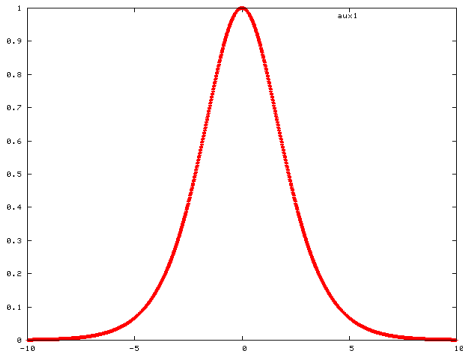


Figure 9.5: aux1(x) function

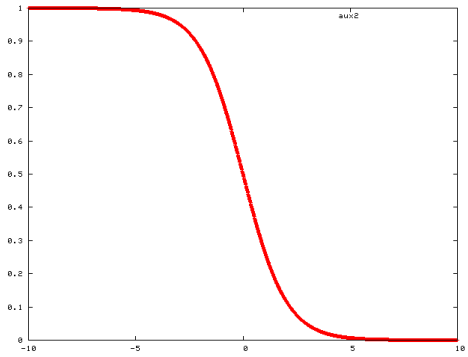


Figure 9.6: aux2(x) function

two assistant function as:

$$n|_{mid} = n_i \,\mathrm{aux2}\left(\frac{\psi_i - \psi_{i+1}}{2V_T}\right) + n_{i+1}\,\mathrm{aux2}\left(\frac{\psi_{i+1} - \psi_i}{2V_T}\right) \tag{9.37}$$

$$\nabla n\Big|_{mid} = \mathrm{aux1}\left(\frac{\psi_i - \psi_{i+1}}{2V_T}\right)\frac{n_{i+1} - n_i}{\Delta x} \tag{9.38}$$

$$J_n|_{mid} = q\mu_n|_{mid}\left(n|_{mid}\frac{\psi_i - \psi_{i+1}}{\Delta x} + V_T \nabla n\Big|_{mid}\right) \tag{9.39}$$

$$p|_{mid} = p_i \,\mathrm{aux2}\left(\frac{\psi_{i+1} - \psi_i}{2V_T}\right) + p_{i+1}\,\mathrm{aux2}\left(\frac{\psi_i - \psi_{i+1}}{2V_T}\right) \tag{9.40}$$

$$\nabla p\Big|_{mid} = \mathrm{aux1}\left(\frac{\psi_i - \psi_{i+1}}{2V_T}\right)\frac{p_{i+1} - p_i}{\Delta x} \tag{9.41}$$

$$J_p|_{mid} = q\mu_p|_{mid}\left(p|_{mid}\frac{\psi_i - \psi_{i+1}}{\Delta x} - V_T \nabla p\Big|_{mid}\right) \tag{9.42}$$

$$\tag{9.43}$$

If we don't need to obtain half point carrier density and other middle variables, current can also be directly expressed by the following formulae:

$$J_n|_{mid} = \frac{qV_T\mu_n}{\Delta x}\left[n_{i+1}\,\mathrm{B}\left(\frac{\psi_{i+1} - \psi_i}{V_T}\right) - n_i\,\mathrm{B}\left(\frac{\psi_i - \psi_{i+1}}{V_T}\right)\right] \tag{9.44}$$

$$J_p|_{mid} = \frac{qV_T\mu_p}{\Delta x}\left[p_i\,\mathrm{B}\left(\frac{\psi_{i+1} - \psi_i}{V_T}\right) - p_{i+1}\,\mathrm{B}\left(\frac{\psi_i - \psi_{i+1}}{V_T}\right)\right] \tag{9.45}$$

where $\mathrm{B}(x) = \dfrac{x}{e^x - 1}$ is the Bernouli function.

Mobility at Half Point

In the discretion above, we need half point mobility value. In general, weak field mobility can be expressed as

$$\mu = \mu(N_A, N_D, n, p, T) \tag{9.46}$$

We can compute mobility directly at half point, by interpolation N_A and N_D. However, the interpolation usually cause large numerical error. Instead, we can interpolation point's mobility to half point, for example linear interpolation:

$$\mu|_{mid} = \frac{\mu_i + \mu_{i+1}}{2} \tag{9.47}$$

Or more physical, consider mobility inverse's relaxation time as a linear function, leads to the following interpolation method:

$$\mu^{-1}|_{mid} = \frac{1}{2}\left(\frac{1}{\mu_i} + \frac{1}{\mu_{i+1}}\right) \tag{9.48}$$

Two interpolation methods have little difference in real application in term of result, accordingly we can choose simple calculation linear interpolation method. After obtain half point weak field mobility, we use the electric field at the half point to get the mobility correction of strong electric field. Then we insert current's discretion into Equation (9.25) and Equation (9.26), we finish the discretion of DDM equations.

After the discretion, a large scale nonlinear algebraic equations have been brought out. It is the nonlinear solver's task to carry out the solution and will be discussed later. User can find a 1D diode simulation code on our web site, which is a good demonstration to this section.

9.4 Discussion about Convectional-Diffusion System

Drift-diffusion model's current equation is typical convection-diffusion equation. This kind of equation has both convection items and diffusion items: at diffusion conditions, convection can be dominating, which is like fluid, or diffusion can be dominating, which is close the thermal conductance.

9.4.1 Numerical scheme for convectional problem

Model Equation of Convectional Problem

In teaching materials, normally we use first order linear wave transport equation as convection problem's model equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{9.49}$$

where $a > 0$ is wave's speed.

First Order Upwind Scheme

Convection problem's numerical simulation is always very complicated. For linear wave equation, the simplest scheme is the first order upwind scheme, which comes from straightforward consideration: since wave comes from the upstream, the difference format should use the upstream information.

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0 \tag{9.50}$$

When $a\dfrac{\Delta t}{\Delta x} \leqslant 1$, it is a stable scheme. However the 1st order upwind scheme's numerical dissipation is very serious. In real application, it could be more than the physical dissipation, so that we could have completely wrong physical image. Figure (9.7) shows the initial triangle wave which adopts $a\dfrac{\Delta t}{\Delta x} = 0.5$ calculation result. Only 20 steps later, the peak value decreases almost half. Because the numerical dissipation of the 1st order upwind scheme is too strong, its numerical results are not accepted by the world.

Lax-Friedrichs Scheme

Lax-Friedrichs is another 1st order scheme with even heavier numerical dissipation:

$$\frac{u_i^{n+1} - \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right)}{\Delta t} + a\frac{u_{i+1}^n - u_{i-1}^n}{\Delta x} = 0 \tag{9.51}$$

Second Order Schemes

All typical 2nd order formulae including 2nd order Lax-Wendroff scheme, 2nd order upwind scheme and Bean-Warming scheme have the same problem: although

numerical dissipation decreases, one suffers from numerical dispersion, which turns to have fake oscillation where the solution gradient is large.

Here is the 2nd order upwind scheme:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\frac{3u_i^n - 4u_{i-1}^n + u_{i-2}^n}{2\Delta x} = 0 \tag{9.52}$$

Figure (9.8) shows the fake numerical oscillation at discontinuous region caused by numerical dispersion of 2nd order upwind scheme.

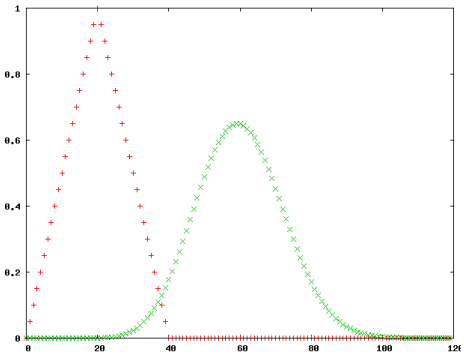


Figure 9.7: Numerical dissipation of 1st order upwind scheme

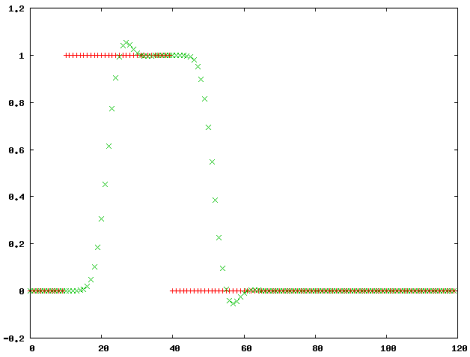


Figure 9.8: Numerical oscillation of 2nd order upwind scheme

Numerical Dissipation and Dispersion

Numerical dissipation (also called numerical viscosity) is generally known in last century 50's. In late 60's Hirt proposes inspiring stable criterion theory, which uses the sign of even order local truncation item to judge difference scheme's stability [36]. A stable numerical scheme must have positive even order local truncation item, and thus be dissipated, for example different type of upwind scheme; or all the even order local truncation items are zero, and thus zero dissipated, eg. leapfog scheme. A dissipation scheme will dissipate system energy, removing wave peak and wave valley. However zero dissipation scheme will not consume system energy itself. These two types of schemes have different application field. Dissipation scheme is generally used in fluid dynamic simulation. Because fluid itself has viscosity, if we control the additional numerical viscosity to certain extent, the influence can be neglected. Zero dissipation scheme is more useful in electromagnetic field calculation. For example FDTD method uses leapfog scheme [37], because Maxwell equations are linear and single wave speed (light speed c) wave functions with no viscosity, numerical viscosity's introduction is equivalent as solving dissipative materials' electromagnetic transport problem, which is not tolerable in many conditions.

As computational hydrodynamics developing, the scheme accuracy turns to be higher as numerical dissipation turns to be weaker. CFD expert Harten proposes a paper in 1983 on 'high resolution' slogan. Simultaneously he proposed famous TVD scheme [38] [39]. But we find after numerical dissipation decreases, the numerical dispersion problem comes out. Numerical dispersion normally will not lead to calculation failure (it could lead to non-linear instability in certain problem), but it will affect physical wave speed. Because partial differential equation's accurate solution can be treated as the summation of series waves with different wave number (reference to Fourier transformation), however numerical dispersion changes these waves' speed. When there is high gradients, the waves separate each other, and the numerical oscillation cames out.

Then people realize a good hydrodynamic scheme needs suitable dissipation and keeps dispersion as small as possible. The standard is that the numerical dissi-

pation can depress the numerical dispersion's side effect. Accordingly Harten's slogan is changed to 'high resolution, non oscillation', and he proposed ENO (essentially non-oscillatory) format [40] [41], so that to decrease or remove the numerical dispersion's parasitic oscillation and non-linear instability.

Numerical scheme with zero dissipation can not help to depress numerical dispersion leading oscillation. However in electromagnetic theory, wave speed is constant light speed c, although leapfog format will lead to light speed variation, the shift keeps the same in all the simulation region. Accordingly it will not lead to oscillated numerical result. However, there are more proposals about even less numerical dispersion scheme [42].

After discussing numerical dissipation and dispersion, we can back to convection problem. Obviously we need a 'high resolution, non oscillation' method. Here 'high resolution' means at least 2nd order accuracy, and it should not have numerical oscillation. We can find suitable solution in CFD world: TVD or ENO format. Here SG format gives another way: center difference plus artificial viscosity item (will discussed below).

Figure (9.9) is initial value with discontinuity at $x = 20$, while wave speed is $a = 1$ moving toward right direction. Considering the wave shape when $t = 30$, theocratical solution is non transformed wave moving, with discontinuity at $x = 50$. Figure (9.10) gives TVD scheme and 2nd order upwind scheme's comparison. We notice that TVD scheme dissipation is relatively small and without oscillation phenomenon.

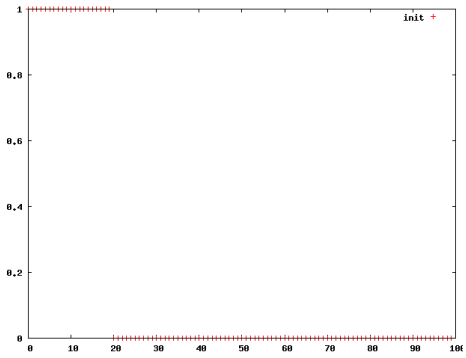


Figure 9.9: initial wave with discontinuity

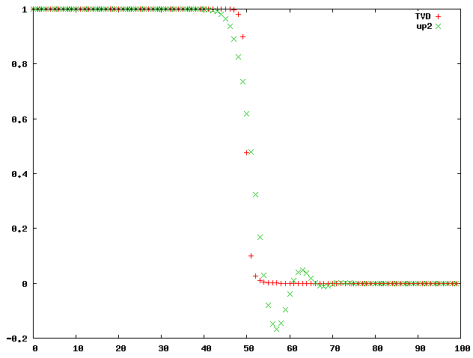


Figure 9.10: TVD scheme vs 2nd order upwind scheme

9.4.2 Numerical scheme for diffusion problem

Model Equation of Diffusion Problem

The model equation of diffusion problem is the 2nd order thermal transfer equation.

$$\frac{\partial u}{\partial t} = b\frac{\partial^2 u}{\partial x^2} \tag{9.53}$$

Center Difference in Space Discretion

Since the physical quantity diffuses to every direction in diffusion problem, it is natural to adopt center difference method. For diffusion problem, normally we consider it is easy to solve, because its physics basement is dissipated. Using center difference scheme to discretize 2nd order derivative item will have 2nd order accuracy and enough numerical dissipation. Accordingly there is no difficulty on numerical discretion.

Diffusion system's trouble is that the time step for satisfying stable condition is normally very small. We need to use implicit format for time discretion in most

of the case. Consider the center difference scheme for model equation:

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}=b\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{\Delta x^2} \tag{9.54}$$

Stability condition is $b\frac{\Delta t}{\Delta x^2}\leqslant\frac{1}{2}$. When Δx is relatively small, the limitation for time step is very strict. So diffusion problem normally adopts implicit format, for example, Crank-Nicolson format, which is 2nd order accurate in time.

Implicit format in Time Discretion

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}=\frac{b}{2}\left(\frac{u_{i+1}^n-2u_i^n+u_{i-1}^n}{\Delta x^2}+\frac{u_{i+1}^{n+1}-2u_i^{n+1}+u_{i-1}^{n+1}}{\Delta x^2}\right) \tag{9.55}$$

In convection-diffusion mixed problem, the limitation to diffusion item is normally stricter than convection item. Semi-implicit format is sometimes be used in early period, which means using implicit format for diffusion item and explicit format for convection item. But after the improvement of calculation method, currently it tends to adopt both implicit format.

9.4.3 Scharfetter-Gummel Scheme

We introduced famous Scharfetter-Gummel scheme in the previous section, brief calls SG scheme. This is a generally used scheme in convection diffusion problems, however not the only one. Because SG's own characteristics, it is suitable in semiconductor simulation and plasma discharge simulation. Here we are going to give a more essential discussion.

Characteristic of SG Scheme

We rewrite the current discretion formation as below:

$$J_n|_{mid}=\frac{qV_T\mu_n}{\Delta x}\left[n_{i+1}\,\mathrm{B}\left(\frac{\psi_{i+1}-\psi_i}{Vt}\right)-n_i\,\mathrm{B}\left(\frac{\psi_i-\psi_{i+1}}{Vt}\right)\right] \tag{9.56}$$

Noticing that

$$\lim_{x\to0}\mathrm{B}(x)=\lim_{x\to0}\frac{x}{e^x-1}=1 \tag{9.57}$$

$$\lim_{x\to+\infty}\mathrm{B}(x)=\lim_{x\to+\infty}\frac{x}{e^x-1}=0 \tag{9.58}$$

$$\lim_{x\to-\infty}\mathrm{B}(x)=\lim_{x\to-\infty}\frac{x}{e^e-1}\sim-x \tag{9.59}$$

When $\psi_i=\psi_{i+1}$, electric field is 0, there is no drift movement, current is dominated by diffusion. The corresponding SG scheme is

$$J_n|_{mid}=\frac{qV_T\mu_n}{\Delta x}(n_{i+1}-n_i) \tag{9.60}$$

This is a three-point-center-difference discretion scheme (only one side), which is suitable for discretion diffusion problem.

When $\psi_i\gg\psi_{i+1}$, diffusion can be neglected, the extreme condition SG scheme becomes:

$$J_n|_{mid}=q\mu En_{i+1} \tag{9.61}$$

On the other hand, when $\psi_i\ll\psi_{i+1}$, there is

$$J_n|_{mid}=q\mu En_i \tag{9.62}$$

Obviously, Formulae (9.61) and Formulae (9.62) represents current is composed of electron drifting movement from node with low electrostatic potential to node with high electrostatic potential. Here we show the SG format's upwind character.

In another point of view, weight function $g(x)$ represents two nodes' electron density contribution, figure Figure (9.11) shows different $\Delta\psi/V_T$ conditions' normalized $g(x)$ function curve. We can find that when $\Delta\psi/V_T$ is relatively large, half point electron density is more weighted by upstream.

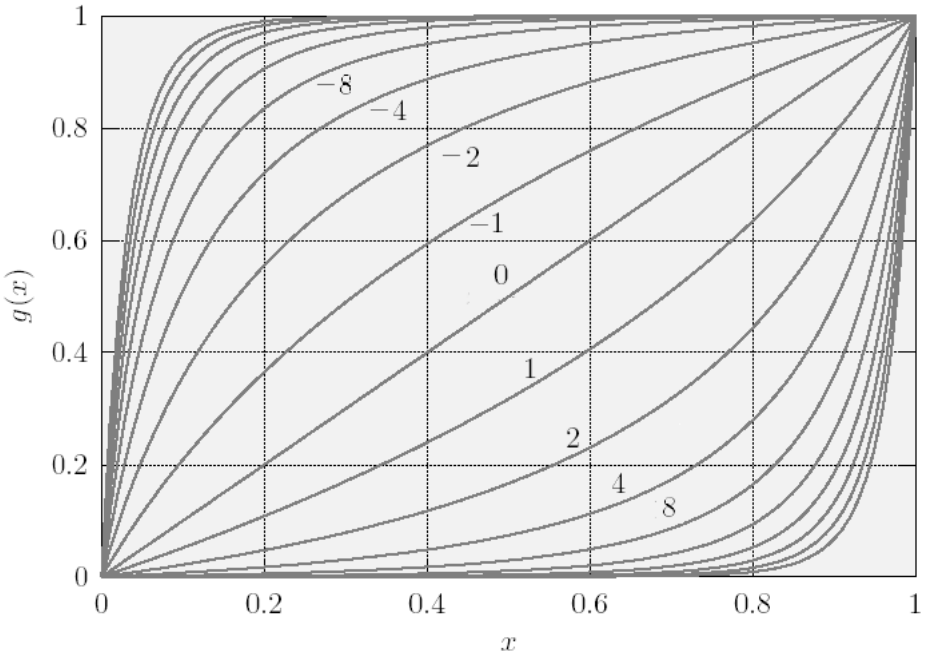


Figure 9.11: $g(x)$function

Essential Property of SG Scheme

We can prove further, SG scheme can be written as center difference plus one artificial diffusion item. Assume $a = qD$, $b = -\frac{qE}{k_bT}$, current equation can be written as

$$j = -abn + a\frac{dn}{dx} \tag{9.63}$$

now continuous equation is still $\frac{dj}{dx} = R$ Making this equation discretion on uniform mesh, assume a, b and R are constant, we have

$$\left(1 - \frac{b\Delta x}{2}\right)n_{i+1} - 2n_i + \left(1 + \frac{b\Delta x}{2}\right)n_{i-1} = a^{-1}\Delta x^2 R \tag{9.64}$$

After introducing both side boundary condition, the above three diagonal equation set can be solved by chasing method:

$$n_i = A + B\frac{\left(1 + \frac{b\Delta x}{2}\right)^i}{\left(1 - \frac{b\Delta x}{2}\right)^i} - \frac{\Delta x R}{ab}i \tag{9.65}$$

where A and B are boundary condition deciding constant. Obviously if Reynolds number $b\Delta x/2 > 1$, when the footnote i changes from odd to even or vice versa, n_i's value will vibrating. One of the solution methods is to introduce artificial

diffusion item D_a, now the current equation is discretized to

$$\begin{aligned} j_{1+1/2} &= \left(-abn + a\frac{dn}{dx} + D_a\frac{dn}{dx}\right)\Big|_{i+1/2} \\ &= -ab\left(\frac{n_i + n_{i+1}}{2}\right) + a\left(\frac{n_{i+1} - n_i}{\Delta x}\right) + D_a\left(\frac{n_{i+1} - n_i}{\Delta x}\right) \end{aligned} \tag{9.66}$$

The purpose of introducing artificial diffusion item is to let new Reynolds number

$$\frac{b\Delta x}{2\left(1 + D_a/a\right)} \tag{9.67}$$

keep less than 1. A suitable choice is

$$D_a = a\left(\alpha_{i+1/2}\coth\alpha_{i+1/2} - 1\right) \tag{9.68}$$

where

$$\alpha_{i+1/2} = \frac{\psi_{i+1} - \psi_i}{2V_T} \tag{9.69}$$

now the discretion equation can be written as

$$\mathrm{B}\left(2\alpha_{i+1/2}\right)n_{i+1} - \left[\mathrm{B}\left(2\alpha_{i-1/2}\right) + \mathrm{B}\left(-2\alpha_{i+1/2}\right)\right]n_i + \mathrm{B}\left(-2\alpha_{i-1/2}\right)n_{i-1} = a^{-1}\Delta x^2 R_i \tag{9.70}$$

$\mathrm{B}(x)$ is still Bernouli function. This is SG discretion scheme.

Disadvantage of Artificial Diffusion

The artificial diffusion introduced by SG scheme will not cause serious problem in one dimensional calculation, since diffusion direction and drift direction is always the same in 1D. But at high dimensional condition, since diffusion toward all different directions, so besides drift direction, there is virtual diffusion effect, called crosswind diffusion, also the more drifting the more artificial diffusion we will introduce. In MOS simulation, when current goes through the channel, the crosswind effect leads to artificial diffusion current vertical to channel, which may induce relatively large numerical error.

9.4.4 Define your own format

Up till now, SG format is not mystery anymore. It can be treated as a self consistent convection diffusion scheme with 2nd order accuracy. Maybe someone will consider to make their own scheme to replace SG, i.e. to reduce artificial diffusion.

For example Formulae (9.60) and Formulae (9.62) can also be used in DDM discretion. For the above formula, the diffusion item can have 2nd order when center difference is used, convection item's upwind discretion is only first order accurate. In order to improve convection item's discretion accuracy, someone goes to upwind basement – CFD society – for higher order upwind scheme. In theory, it is all right. Many high accuracy schemes, eg. TVD and ENO can be found in CFD world. The trouble is afterwards, which is going be discussed in detail later. After DDM discretion, the PDEs of DDM lead to large scale nonlinear equations. One has to build Jacobian matrix to form the Newton iteration. Things for SG scheme is relatively easy, the Jacobian matrix can be calculated accurately. However for those high order upwind scheme, since adopting complicate reconstruction and limiter technique, there is almost no way to construct accurate Jacobian matrix by human[2]. We can only use approximate matrix instead (CFD also suffered with

[2] The automatically differentiation may be the hope.

this problem). Using approximate matrix will lead to Newton iteration convergence speed decrease to linear order, for certain disadvantage conditions, it could lead to divergence. As a result, high accurate schemes are good for lab usage, for example looking for new algorithm.

Since GSS as a general semiconductor software has high request on calculation speed and convergence need, accordingly we adopt simple and efficient SG format[3].

9.5 GSS First Level DDM Solver

This section will introduce numerical discretion method use in GSS for solving first level DDM equations. Based on the discretion of 1D DDM and 2D Poisson's equation we had introduced, the 2D DDM discretion is almost there.

Compact Format of DDM Equations

For convenience, introducing solve vector $\mathbf{Q}$, flux vector $\mathbf{F}$ and source vector $\mathbf{S}$ here:

$$\mathbf{Q} = \begin{pmatrix} 0 \\ n \\ p \end{pmatrix}, \mathbf{F} = \begin{pmatrix} \varepsilon \nabla \psi \\ \frac{1}{q}\mathbf{J}_n \\ -\frac{1}{q}\mathbf{J}_p \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \rho \\ G - R \\ G - R \end{pmatrix}$$

Write basic DDM Equation (7.9) as compact format:

$$\frac{\partial \mathbf{Q}}{\partial t} = \nabla \cdot \mathbf{F} + \mathbf{S} \tag{9.71}$$

Do integration over Voronoi cell i as we have done in Equation (9.2) discretion of Poisson's equation, and adopt Green theorem, we have

$$\int_{\Omega_i} \frac{\partial \mathbf{Q}_i}{\partial t} \mathrm{d}V = \sum_e \mathbf{F}_e l_e + \int_{\Omega_i} \mathbf{S}_i \mathrm{d}V \tag{9.72}$$

Where Ω_i is the Voronoi cell belongs to i, $\mathbf{F}_e$ is flux at the face edge e of Voronoi cell, l_e is the length of the face edge e.

Volume Integration of Solution Variable

The variable $\mathbf{Q}$ in the integral express at left hand side of Equation (9.72) can be brought out by assuming electron and hole density is constant in the cell. Although it is a rough approximation, that is the best so far. Now the integration becomes:

$$\int_{\Omega_i} \frac{\partial Q_i}{\partial t} \mathrm{d}V = \frac{\partial Q_i}{\partial t} \Delta V_{\Omega_i} \tag{9.73}$$

where, ΔV_{Ω_i} is the area of Voronoi cell i.

Volume Integration of Source Term

We temporarily forget the generation of carrier here. The charge ρ and recombination rate are considered to be constant inside voronoi cell. Then the source item $\mathbf{S}$ integration can also be written as:

$$\int_{\Omega_i} \mathbf{S}_i \mathrm{d}V = \mathbf{S}_i \Delta V_{\Omega_i} \tag{9.74}$$

Evaluate Flux Function

The remaining difficulty is how to evaluation the flux function $\mathbf{F}_e$ at the face of Voronoi cell.

[3] TVD style AUSM+ scheme had been tried in GSS version 0.3x, which brings less numerical dissipation than SG scheme. However, it is easy to divergence.

Since discretion method of Poisson's equation had been described before, we focus on carrier continuity equation's flux function, the discretion of electron current density and hole current density in 2D.

In real code, flux at the face of Voronoi cell is obtained by scanning all the triangles [43]. Since Voronoi diagram and triangles are overlapped each other shown in Figure (8.7). Scanning all the triangles is equivalent as scanning all the Voronoi cells. It is going to be discussed later that this method will lead to convenience for calculating electric field and current compared to calculating them inside the Voronoi cell.

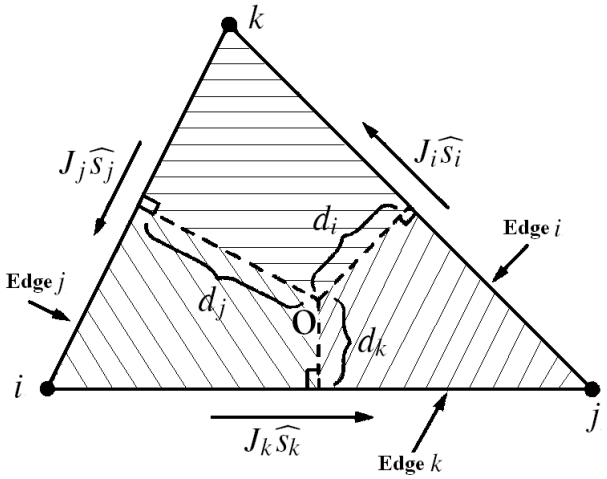


Figure 9.12: Triangle unit

The following current density can be electron current density or hole current density, shown in Figure (9.12). The node i, j, k is the vertex of triangle as well as the center of Voronoi cell. Thus, the triangle can be divided into 3 parts belongs to each Voronoi cell, which represented with different type of shade. Single triangle's flux contribution to Voronoi cell i, j, k is:

$$\mathbf{F}_i = J_j d_j - J_k d_k \tag{9.75}$$

$$\mathbf{F}_j = J_k d_k - J_i d_i \tag{9.76}$$

$$\mathbf{F}_k = J_i d_i - J_j d_j \tag{9.77}$$

Where $\widehat{s_k}$ is edge k's unit vector. d_k is the flux cross section for side k. $J_k = \mathbf{J} \cdot \widehat{s_k}$ is an average value of the projection of current density vector $\mathbf{J}$ onto triangle edge k between nodes i, j. J_i, J_j has similar meaning, they are obtained through S-G discretion scheme.

To obtain current density J_k along triangle edge k, we need to assume electrical field and current density changes slowly along the edge, so that we can treat them as constant value. Then the current transport along triangle edge can be treated as one dimensional problem.

Now Equation (9.39) – Equation (9.42) or Equation (9.44) and Equation (9.45) all can be used to calculate the current density. Taking Equation (9.44) and Equation (9.45) as an example, current J_k flows along edge k can be written as:

$$J_k^n = \frac{q\mu_n V_T}{L_k}\left[n_j \mathrm{B}\left(\frac{\psi_j - \psi_i}{V_T}\right) - n_i \mathrm{B}\left(-\frac{\psi_j - \psi_i}{V_T}\right)\right] \tag{9.78}$$

$$J_k^p = \frac{q\mu_p V_T}{L_k}\left[p_i \mathrm{B}\left(\frac{\psi_j - \psi_i}{V_T}\right) - p_j \mathrm{B}\left(-\frac{\psi_j - \psi_i}{V_T}\right)\right] \tag{9.79}$$

where, L_k is the length of edge k, μ_n and μ_p are mobility at the center of edge k.

The previous formulae represents the flux's conservation. Using $J_k d_k$ as an example, it is the current which goes out of the Voronoi cell i and goes into the Voronoi cell j. Accordingly the flux for the Voronoi i is negative, while for Voronoi j is positive and absolute value are the same.

By scanning all the triangles with following this way and sum all of the flux along triangle edges, we can go into desired flux function for every Voronoi cell.

9.6 Mobility Implementation in 2D

From the previous section we notice, current density expression at the face of 2D Voronoi cell is similar as 1D condition. But the treatment of mobility μ for 2D condition is more difficult from 1D. This is because electric field direction and current direction are the same or opposite for 1D. However, electric vector and current density vector can have an angle in 2D case. Because high field mobility model and surface mobility model need the parallel and vertical electric field component to the direction of current (see "Mobility Models", on page 76), we have to be careful to deal with it.

How to calculate mobility value on triangle edge center is the difficulty in 2D semiconductor simulation. GSS software calculates triangle vertex's mobility and then uses interpolation method Equation (9.47) to obtain mobility at triangle edge center. Noticing $\mu = \mu(N_A, N_D, n, p, E_{//}, E_{\perp})$, since the impurity concentration and carrier concentration for the node are known, the difficulty is how to solve $E_{//}$ and $E_{\perp}$.

EJ Model

Theoretically, calculating $E_{//}$ and $E_{\perp}$ needs to have node's $\mathbf{E}$ and carrier current density vector $\mathbf{J}$, and then do the projection with the following formulae:

$$E_{//} = \frac{\max(0, \mathbf{E} \cdot \mathbf{J})}{|\mathbf{J}|} \tag{9.80}$$

$$E_{\perp} = \frac{|\mathbf{E} \times \mathbf{J}|}{|\mathbf{J}|} \tag{9.81}$$

Mobility Evaluation in Voronoi Cell

We can use many method to obtain $E_{//}$ and $E_{\perp}$. If we solve inside the Voronoi cell, noticing Equation (5.41) and Equation (5.43), current density can be illustrated with quasi-Fermi potential's gradients:

$$\mathbf{J}_n = -q\mu_n n \nabla \phi_n \tag{9.82}$$

$$\mathbf{J}_p = -q\mu_p p \nabla \phi_p \tag{9.83}$$

Here, taking electron's mobility as an example: we can use least square method on the Voronoi diagram (please refer to "Finite Volume Discretion of Derivative Operator", on page 93) to compute potential and electron quasi-Fermi potential's gradients. Then we have the electric field vector and current density vector in the Voronoi cell as the Equation (9.80) and Equation (9.81) can be used to calculate $E_{//}$ and $E_{\perp}$ (μ_n in $\mathbf{J}_n$ of Equation (9.82) can be canceled during the calculation). After having $E_{//}$ and $E_{\perp}$, we can obtain node's electron mobility.

This method's shortcoming is Voronoi's least square method involves too many neighbor nodes. Normally a Voronoi cell have 5 ~ 7 neighbor nodes, since the mobility is the interpolation function between two neighbor Voronoi's mobility. So totally more than 10 neighbor nodes are involved. It is quite troublesome for later Jacobian matrix calculation. And the matrix bandwidth is relatively large.

Mobility Evaluation in Triangle

In order to decrease programmer's work load, Laux suggests gradients' calculation

inside triangle. From finite element analysis we known, when ψ at each triangle vertex is known, $\nabla\psi$ can be treated as constant value inside triangle:

$$E_x = -\psi_x = -\frac{(y2-y3)\psi_1 + (y3-y1)\psi_2 + (y1-y2)\psi_3}{2\Delta} \tag{9.84}$$

$$E_y = -\psi_y = -\frac{(x3-x2)\psi_1 + (x1-x3)\psi_2 + (x2-x1)\psi_3}{2\Delta} \tag{9.85}$$

Where ψ_1, ψ_2 and ψ_3 are the ψ at the vertex of triangle, Δ is triangle area.

For calculating the current vector inside a triangle, we use weighted interpolation method [43], shown as Figure (9.12). First, as discussed in the last section, we consider there exists current vector $\mathbf{J}$ inside triangle. The current density along the triangle edge which evaluated by S-G discretion scheme is the projection of $\mathbf{J}$ on the triangle edge. Since only the current density along the triangle edges are known, we have to construct $\mathbf{J}$ by J_i, J_j and J_k. Because S-G discretion format is non-linear, this current vector $\mathbf{J}$ is not easily obtained as electric field vector, and might not have single solution. There are three different interpolation methods to construct current vector $\mathbf{J}$:

$$\mathbf{J}_{jk} = \frac{1}{\sin^2 i}\left[\left(J_j + \cos i \cdot J_k\right)\hat{s}_j + \left(J_k + \cos i \cdot J_j\right)\hat{s}_k\right] \tag{9.86}$$

$$\mathbf{J}_{ki} = \frac{1}{\sin^2 j}\left[\left(J_k + \cos j \cdot J_i\right)\hat{s}_k + \left(J_i + \cos j \cdot J_k\right)\hat{s}_i\right] \tag{9.87}$$

$$\mathbf{J}_{ij} = \frac{1}{\sin^2 k}\left[\left(J_i + \cos k \cdot J_j\right)\hat{s}_i + \left(J_j + \cos k \cdot J_i\right)\hat{s}_j\right] \tag{9.88}$$

Where, J_i, J_j and J_k are current along triangle edge. $\mathbf{J}_{jk}$, $\mathbf{J}_{ki}$ and $\mathbf{J}_{ij}$ are current vector $\mathbf{J}$ constructed by different linear interpolation. Laux uses a weighted average method to obtain the final current vector $\mathbf{J}$. For example in Figure (9.12), current vector $\mathbf{J}$ is considered to be average value of $\mathbf{J}_{jk}$ and $\mathbf{J}_{ki}$ with different weights in small triangle Δioj related to edge k:

$$\mathbf{J}_k = \frac{d_i \cdot \mathbf{J}_{ki} + d_j \cdot \mathbf{J}_{jk}}{d_i + d_j} \tag{9.89}$$

Since E_x and E_y are known, Equation (9.80) and Equation (9.81) can be used to calculate $E_{//}$ and $E_\perp$ with Equation (9.89) through edge k, finally we can obtain mobility along edge k. The mobility along other two edges can be composed with the similar way. Please pay attention that during the S-G current calculation at the triangle edge, we need the mobility information, and electrical field evaluation needs S-G current. It seems that it is a coupled process. In fact, mobility only affected by current direction, not amplitude. Accordingly we can assume any mobility value during S-G current calculation, and normally select mobility value which scales current density at the order of $O(1)$ for improving the evaluation accuracy of $\mathbf{E}\cdot\mathbf{J}$ and $\mathbf{E}\times\mathbf{J}$. After finishing mobility evaluation, we can scale the S-G current to its real value.

In this way, $E_{//}$ and $E_\perp$ calculation only involves 3 points, the band width of Jacobian matrix can be greatly reduced and simultaneously calculation accuracy can still be ensured. The shortcoming of this method is that when current density is very low, calculating $\mathbf{E}\cdot\mathbf{J}$ and $\mathbf{E}\times\mathbf{J}$ may brings large numerical error, which may affects convergence (for equilibrium state, since current density is 0, $\mathbf{E}\cdot\mathbf{J}$ and $\mathbf{E}\times\mathbf{J}$ are meaning less).

ESimple Model

When the accurate requirement is not so strict, there is another simple method shown in Figure (9.13) The mobility for weighting current density from node 1 to

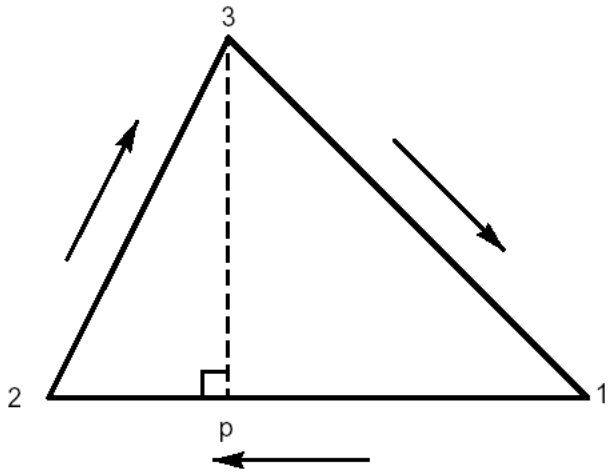


Figure 9.13: Solving $E_{//}$ and $E_{\perp}$ by simple method

node 2 can use following parallel and vertical electric field expressions:

$$E_{//} = \frac{|\psi_2 - \psi_1|}{d_{12}} \tag{9.90}$$

$$E_{\perp} = \frac{|\psi_3 - \psi_p|}{d_{3p}} \tag{9.91}$$

this method's advantage is easy to program, calculation speed is fast, convergence performance is better. However simulation result is heavily related to mesh in this situation: when current flow is mainly along triangle edge, the result is relatively accurate. If we can not secure the previous condition, we need to go back to Equation (9.80) and Equation (9.81).

GSS DDML1E/DDML2E/EBML3E solvers are based on above discussion. Simultaneously support Laux's method and the simplification method, which is suitable for either accuracy first and speed first condition. In the later chapter, these two methods are called EJ model and ESimple model. ESimple model is set as the default model for mobility evaluation.

9.7 GSS Second Level DDM Solver

GSS Level 2 DDM solver considers thermal effect during device simulation. Beside one crystal thermal transportation equation has been added to system, expression of current density has many variations, shown in Equation (7.14). Here correspondingly basic variables include ψ, n, p and crystal temperature T.

Current Equation in DDML2

Rewrite the electron and hole current density as below, we consider the SG discretion scheme:

$$J_n = \mu_n(qE + k_b\nabla T)n + \mu_n k_b T\nabla n \tag{9.92}$$

$$J_p = \mu_p(qE - k_b\nabla T)p - \mu_p k_b T\nabla p \tag{9.93}$$

Assumption before SG Discretion

Semiconductor thermal conduct are normally very fast, for example silicon's thermal conductivity is one third as copper's. And because the semiconductor device dimension is very small, device internal temperature gradients will not be large. As a result, we can assume the temperature difference between two neighbor nodes is not large, and temperature's gradients is treated as constant, besides the electric

field and mobility constant assumption:

$$D_n = \frac{\mu_n k_b T}{q} \approx const$$
$$D_p = \frac{\mu_p k_b T}{q} \approx const$$
$$E \approx const$$
$$\nabla T \approx const$$

SG Discretion of DDML2 Current Equation

We use the electron current density discretion as an example to deduct Scharfetter-Gummel format with temperature. Now electron density equation can be written as :

$$\frac{\mathrm{d}n}{\mathrm{d}x} + \frac{\mu_n (qE + k_b \nabla T)}{qD_n} n = \frac{J_n}{qD_n} \tag{9.94}$$

Let

$$a = \frac{\mu_n (qE + k_b \nabla T)}{qD_n} = \frac{E}{V_T(T)} + \frac{\nabla T}{T} \tag{9.95}$$

$$C_0 = \frac{J_n}{qD_n} \tag{9.96}$$

The general solution of ordinary differential Equation (9.94) can be written as:

$$n = \frac{C_0}{a} + C_1 e^{-ax} \tag{9.97}$$

substitution n_i, n_j into above equation as boundary condition, then

$$n = n_i [1 - \mathrm{g}(x)] + n_j \mathrm{g}(x) \tag{9.98}$$

where

$$\mathrm{g}(x) = \frac{1 - e^{-ax}}{1 - e^{-a\Delta x}} \tag{9.99}$$

After we obtain analytic expression of electron density, the electron density's gradients at middle point $x = \Delta x/2$ can be solved out. Then we can obtain SG discretion scheme of electron current equation.

Because afterwards we need to use carrier density at middle point, we put the SG format as the following:

$$n|_{mid} = n_i \,\mathrm{aux2}(\alpha_n) + n_j \,\mathrm{aux2}(-\alpha_n) \tag{9.100}$$
$$\nabla n\Big|_{mid} = \mathrm{aux1}(\alpha_n) \frac{n_j - n_i}{\Delta x} \tag{9.101}$$
$$J_n|_{mid} = q\mu_n|_{mid} \left(n|_{mid} \frac{\psi_i - \psi_j}{\Delta x} + \frac{k_b T|_{mid}}{q} \nabla n\Big|_{mid} + \frac{k_b n|_{mid}}{q} \nabla T\Big|_{mid} \right) \tag{9.102}$$
$$p|_{mid} = p_i \,\mathrm{aux2}(-\alpha_p) + p_j \,\mathrm{aux2}(\alpha_p) \tag{9.103}$$
$$\nabla p\Big|_{mid} = \mathrm{aux1}(\alpha_p) \frac{p_j - p_i}{\Delta x} \tag{9.104}$$
$$J_p|_{mid} = q\mu_p|_{mid} \left(p|_{mid} \frac{\psi_i - \psi_j}{\Delta x} - \frac{k_b T|_{mid}}{q} \nabla p\Big|_{mid} - \frac{k_b p|_{mid}}{q} \nabla T\Big|_{mid} \right) \tag{9.105}$$
$$\tag{9.106}$$

where

$$\alpha_n = \frac{\psi_i - \psi_j}{2V_T|_{mid}} + \frac{T_j - T_i}{2T|_{mid}} \tag{9.107}$$
$$\alpha_p = \frac{\psi_i - \psi_j}{2V_T|_{mid}} - \frac{T_j - T_i}{2T|_{mid}} \tag{9.108}$$

All involved temperature can be expressed with linear interpolation, for example:

$$T|_{mid} = \frac{T_i + T_j}{2} \tag{9.109}$$

$$\nabla T\Big|_{mid} = \frac{T_j - T_i}{\Delta x} \tag{9.110}$$

$$V_T|_{mid} = \frac{k_b T|_{mid}}{q} \tag{9.111}$$

Discretion of Thermal Transfer Equation

In DDML2, GSS will solve crystal thermal transfer equation:

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot \kappa \nabla T + \mathbf{J} \cdot \mathbf{E} + (E_g + 3k_b T) \cdot (U - G) \tag{9.112}$$

Temporarily we don't consider transient problem and carrier generation, $(E_g + 3k_bT) \cdot U$ can be considered constant and directly integrated inside the control volume,

With the basement from Poisson's equation, discretion of $\nabla \cdot \kappa \nabla T$ is not a problem. The most difficult part is joule heating item $\mathbf{J} \cdot \mathbf{E}$.

Here we introduce GSS adopted discretion method. Put joule heating item as:

$$\mathbf{J} \cdot \mathbf{E} = -(\mathbf{J}_n + \mathbf{J}_p) \cdot \nabla \psi \tag{9.113}$$

Consider $\mathbf{J}_n \cdot \nabla \psi$'s integration inside Voronoi cell. Assume potential variation inside a Voronoi cell is small, we have the following approximation:

$$\begin{aligned}
\int_{\Omega_i} \mathbf{J}_n \cdot \nabla \psi \mathrm{d}V &= \int_{\Omega_i} \nabla \cdot (\mathbf{J}_n \psi)\, \mathrm{d}V - \int_{\Omega_i} \psi \nabla \cdot \mathbf{J}_n \mathrm{d}V_i \\
&\approx \oint_{\partial\Omega_i} \mathbf{J}_n \psi \cdot \mathrm{d}\mathbf{S} - \psi_i \oint_{\partial\Omega_i} \mathbf{J}_n \cdot \mathrm{d}\mathbf{S} \\
&= \sum_j \frac{\psi_i + \psi_j}{2} J_{mid} \Delta L_j - \psi_i \sum_j J_{mid} \Delta L_j \\
&= \sum_j \frac{\psi_i - \psi_j}{2} J_{mid} \Delta L_j
\end{aligned} \tag{9.114}$$

Then, through Gauss formulae and an approximation we avoid to integration the $\mathbf{J} \cdot \mathbf{E}$ over the Voronoi cell, yet the final discretion format is very simple. This discretion format is independently thought out by the author, but afterwards, the author also saw a similar paper in 1994 [44].

9.8 GSS Third Level EBM Solver

GSS third level EBM solver needs to solve up to 6 equations: Poisson's equation, electron and hole continuous equation, electron and hole energy balance equation and crystal lattice thermal transfer equation. Since the discretion scheme of Poisson's equation and crystal transfer equation is similar to the previous two sections, we won't discuss it again. We focus on numerical discretion scheme about continuous equation and energy balance equation.

SG Discretion of Current Equation

We had mentioned hole equation can be obtained by replace k_b in the electron equation with $-k_b$. We only rewrite the electron current Equation (7.15) as below:

$$\mathbf{J}_n = -q\mu_n n \nabla \left(\psi - \frac{k_b T_n}{q} \right) + q D_n \nabla n \tag{9.115}$$

Following SG discretion procedure, we rewrite it as an equation with electron density n:

$$\frac{dn}{dx} - \frac{1}{\Delta x}\left[\frac{\frac{q}{k_b}\Delta\psi - \Delta T_n}{T_n(x)}\right]\cdot n = \frac{J_n}{qD_n} \tag{9.116}$$

Assume electric field density and electron temperature's gradients are constant. We solve the above equation with known electron density of two sides. After some deduction, we can SG discretion scheme of electron current equation [45].

$$J_n = \frac{qD_n}{\Delta x}\frac{T_j - T_i}{\ln T_j - \ln T_i}\left[\mathrm{B}(\alpha)\frac{n_j}{T_j} - \mathrm{B}(-\alpha)\frac{n_i}{T_i}\right] \tag{9.117}$$

where

$$\alpha = \frac{\ln T_j - \ln T_i}{\Delta T}\left[\frac{q}{k_b}\Delta\psi - 2\Delta T\right] \tag{9.118}$$

SG Discretion of Energy Balance Equation

Regarding energy balance equation's discretion, we use the similar deduction of SG discretion procedure, rewrite electron energy flow as Equation (7.19):

$$S_n = -\left(\frac{5}{2}+\gamma\right)\frac{k_b^2}{q}T_n\mu_n n\nabla T_n - \frac{5}{2}k_b T_n\frac{J_n}{q} \tag{9.119}$$

Rewrite the above formulae as ordinary differential equation with variable T_n:

$$\frac{dT_n}{dx} + \frac{\frac{5}{2}}{\left(\frac{5}{2}+\gamma\right)}\frac{J_n}{qD_n n(x)}\cdot T_n = -\frac{S_n}{\left(\frac{5}{2}+\gamma\right)k_b D_n n(x)} \tag{9.120}$$

Electron temperature at both side is know as the boundary condition.

The most reasonable method is solving Equation (9.117) and Equation (9.120) self-consistently. Unfortunately, the solution will be infinite series. So we need to do some approximation here. For example assume the electron temperature gradients is constant when we solve the current Equation (9.117). But obviously when we solve the energy flow Equation (9.120), this assumption is destroyed. Currently there are several different methods to solve the energy flow's discretion. Tang's method shows electron density are separately solved from current equation and energy flow equation, however, the difference of electron density from two solution are very large [46]. While Forghieri's method directly assume electron density as exponential function distribution [45] during the discretion of Equation (9.120). In 1994, Choi proposed a new energy flow equation discretion scheme, electron density is solved from current Equation (9.117) and substitution it into energy flow Equation (9.120). It decreases the mismatch between two, the new discretion format can help to decrease electron temperature calculation error and enhance the convergence [47].

GSS adopts Choi's discretion scheme, the electron energy flow is:

$$S_n = -\left(\frac{5}{2}+\gamma\right)\frac{k_b D_n}{\Delta x}\frac{\Delta T}{\ln T_j - \ln T_i}\left[\mathrm{B}(\alpha)\frac{\mathrm{B}(\Phi)}{\mathrm{B}\left(\tilde{\Phi}\right)}n_j - \mathrm{B}(-\alpha)\frac{\mathrm{B}(\Phi)}{\mathrm{B}\left(\tilde{\Phi}\right)}n_i\right] \tag{9.121}$$

where

$$\begin{aligned}\tilde{\Phi} &= \frac{\ln T_j - \ln T_i}{\Delta T}\left(\frac{q}{k_b}\Delta\psi - \Delta T\right) - \ln\frac{n_j}{n_i}\\ \Phi &= \frac{\frac{5}{2}}{\left(\frac{5}{2}+\gamma\right)}\tilde{\Phi}\end{aligned} \tag{9.122}$$

We notice discretion of current equation and energy balance equation all involves the following formulae:

$$\frac{\Delta T}{\ln T_j - \ln T_i} \tag{9.123}$$

When $\lim_{T_i \to T_j}$, we need special treatment.

In the realization of EBML3E solver, the selection of independent variables needs more attention. Regarding the additional energy balance equation (represented with F_{ω_n} and F_{ω_p}), the author first consider electron temperature T_n and hole temperature T_p as independent variables. However, for region with small electron density, there is $\frac{\partial F_{\omega_n}}{\partial n} \gg \frac{\partial F_{\omega_n}}{\partial T_n}$, which means that Jacobian matrix loses diagonal domination. It always leads to convergence problem. So referred from CFD, put the independent variable as carrier density and temperature's product nT_n and pT_p. In this way we can secure Jacobian matrix's diagonal domination, and have better convergence property.

9.9 GSS Quantum Corrected DDM Solver

Dr. Andreas Wettstein published a series of papers [48][49][17], carefully studied DG-DDM's discretion scheme. His format although is not perfect in mathematic, it is very effective in real practice. This scheme is integrated into Dessis software later on.

DG-DDM Discretion Scheme

For DG-DDM Equation (7.25), the discretion for Poisson's equation as well as continuous equation keeps the same as canonical DDM equations, we won't discuss them again here. We focus on discretion scheme of quantum potential equation caused by the gradients of electron density. Hole's quantum potential equation can also refer to this treatment. Integrate Equation (7.33) in control volume, and use Gauss formulae we have:

$$\int_{\Omega_i} \Lambda_n \mathrm{d}V = -b_n \oint_{\partial\Omega_i} \frac{\nabla\sqrt{n}}{\sqrt{n}} \mathrm{d}S \tag{9.124}$$

where $b_n = \dfrac{\hbar^2}{6qm_n^*}$

We notice

$$n = n_0 \exp\left(\frac{E_{Fn} - E_{qc}}{k_b T}\right) = n_0 \exp\left(-\Phi\right) \tag{9.125}$$

where $\Phi = \dfrac{E_{qc} - E_{Fn}}{k_b T}$. Put this formulae into Equation (9.124), we have

$$\begin{aligned} \Lambda_n \Delta V_{\Omega_i} &= -b_n \sum_j \left(\frac{\sqrt{n_j} - \sqrt{n_i}}{d_{ij}} \cdot \frac{1}{\sqrt{n_i}}\right) \cdot \Delta L_j \\ &= b_n \sum_j \left(1 - \exp\left(\frac{\Phi_i - \Phi_j}{2}\right)\right) \cdot \frac{\Delta L_j}{d_{ij}} \end{aligned} \tag{9.126}$$

Here j is the sum of all the neighbor nodes of i, ΔL_j is the corresponding face of Voronoi cell to node j , d_{ij} is the distance from node i to node j.

Dr. Wettstein at first uses Equation (9.126) as discretion scheme to electron quantum potential equation. But the exponential items in Equation (9.126) easily

lead to divergence. So he points out that when exponential item is positive, we should use Taylor series expansion to avoid numerical overflow, which helps to improve the numerical stability. The corrected discretion scheme is:

$$\Lambda_n \Delta V_{\Omega_i} = \begin{cases} b_n \sum_j \left(1 - \exp\left(\frac{\Phi_i - \Phi_j}{2}\right)\right) \cdot \frac{\Delta L_j}{d_{ij}} & for\ \Phi_i < \Phi_j \\ b_n \sum_j \left(\frac{\Phi_j - \Phi_i}{2} - \frac{(\Phi_j - \Phi_i)^2}{8}\right) \cdot \frac{\Delta L_j}{d_{ij}} & otherwise \end{cases} \tag{9.127}$$

Non-conservation Formula

Wettstein scheme does not have a conservation formula. For example when $\Phi_i < \Phi_j$, the incense of quantum potential Λ_n of Voronoi cell i is obviously not equal to the decrease of quantum potential Λ_n for Voronoi cell j. I have discussed with Dr. Wettstein. His answer is that for hydrodynamics, this conservation is not tolerable, but quantum potential is not a conservation quantity. And the application of this scheme is good.

Low Carrier Concentration Problem

In GSS, DG-DDM realization meets another problem. The author select basic independent variable as ψ, n, p, E_{qc} and E_{qv} (the last two variables are equal to Λ_n and Λ_p). Noticing that the quantum potential equation contains E_{Fn} which has the following formula:

$$E_{Fn} = -q\phi_n = -q\left(\psi - \frac{k_b T}{e} \ln\left(\frac{n}{n_{ie}}\right)\right)$$

During Jacobian matrix evaluation, the derivative of E_{Fn} over electron density have item like $\frac{1}{n}$. In low electron density region, the value of $\frac{1}{n}$ is huge, which leads loss of Jacobian matrix diagonal domination. The condition number becomes worse, leading to bad convergence in the end (similarly, hole has the same problem). In energy balance model realization, author solved the similar problem by reselecting the independent variables. But this time, we can not bypass it. Silvaco cooperation's ATLAS software has the similar problem. They assume a QMINCONC variable: when carrier density is lower than this value, there is no consideration of quantum effect [50]. GSS does not change the equation, but when carrier density is lower than intrinsic carrier density, Jacobian matrix evaluation does not consider partial differential item of electron (hole) density. This will lead to inaccurate Jacobian matrix, which leads to slow convergence.

The author currently does not have good idea to solve above problems. Please feel free to send email to the author if you have any solutions or suggestions.

9.10 Discretion the Carrier Generation Term

Impact ionization is an important phenomenon of semiconductor device, which leads directly to device breakdown. For power device, the unwanted impact ionization may limit the maximum voltage and current. How to optimize the device geometry and doping profile for high voltage endurance is an interesting task.

Unfortunately, impact ionization is always difficulty in numerical simulation. It is very easy to meet the convergence problem since breakdown current usually increases exponentially. For further discussion please refer to "??", on page ??.

Rewrite impact ionization item G^{II} mentioned in Equation (7.79) as:

$$G^{II} = \alpha_n(E)|J_n| + \alpha_p(E)|J_p| \tag{9.128}$$

Based on finite volume method, we need to calculate integral of G^{II} over control volume Ω_i. Please note the driving electric field $\mathbf{E}$ is required for evaluating α_n and α_p. For example, Equation (7.80) is α_n expression for Selberherr model.

Here we calculate G^{II} again inside the triangle, shown in Figure (9.14).

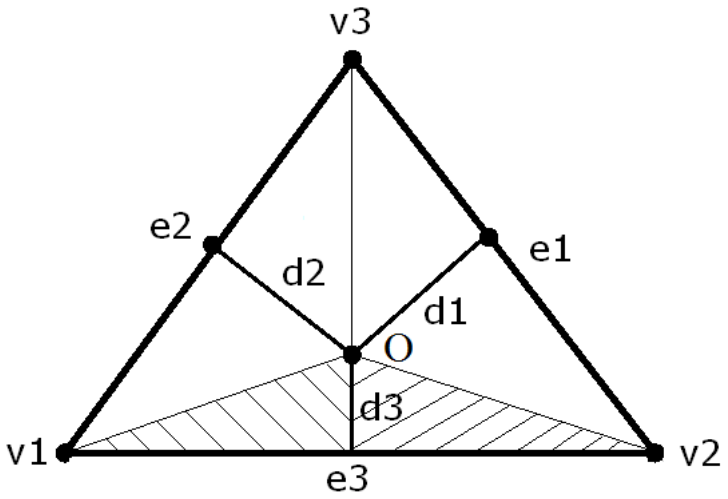


Figure 9.14: Impact ionization item's treatment inside triangle unit

ESide Impact Ionization Model

There are many choices for electric field magnitude and current evaluation. The simplest and most non physical method is both electric field and current definition on the triangle edge. This method is called ESide model afterwards. G^{II} is considered as fixed value inside every shade area of Figure (9.14). The electric field directly comes from electrostatic potential gradients of node v1 and v2. Here current J_n and J_p are obtained directly by S-G scheme.

This method has stable numerical performance due to its simplicity. In early semiconductor numerical simulation softwares, it was generally used. However, its numerical simulation result is always higher than the real breakdown voltage. Simultaneously, simulation is very dependent on mesh. Accordingly it is not recommended to use ESide model nowadays. The author puts it in to GSS because whenever we meet convergence problem, ESide could be the last choice as a reference.

EVector Impact Ionization Model

Another improving method is to define the electric field inside triangle, called as EVector model. This model uses Equation (9.84) and Equation (9.85) to evaluate E_x and E_y inside a triangle, assuming electric field keeps constant inside the whole triangle. Current still obtained directly by S-G scheme. EVector model is more accurate than Eside model, however it still has non-physical part.

EdotJ Impact Ionization Model

The EdotJ model, first appeared in Laux's paper in 1985 [51], is the most physical model as well as the most complicate model. This model is similar with his mobility implementation [43]. In fact, when we realize Laux mobility model, its impact ionization model has no more difficulty. In EdotJ model, electric field is defined the same as what we mentioned for mobility EJ model:

$$E_{//} = \frac{\max(0, \mathbf{E} \cdot \mathbf{J})}{|\mathbf{J}|} \tag{9.129}$$

It clearly shows that only current parrel to electric field can lead to impact ionization. Also, current J_n and J_p in Equation (9.128) is obtained from complicated interpolation method Equation (9.89). Its only disadvantage is bad convergence, calculation is easily diverge.

GradQf Impact Ionization Model

There is another method for electric field intensity evaluation, called GradQf model, which uses the gradients of Fermi potential as the driving force of impact ionization:

$$E_n = |\nabla\phi_n| \tag{9.130}$$

$$E_p = |\nabla\phi_p| \tag{9.131}$$

where Fermi potential ϕ_n and ϕ_p are defined in Equation (5.45).

In GradQf model, current along triangle edge should not be directly used, otherwise it will lead to oscillation result. Author uses Equation (5.41) and Equation (5.43) to obtain the current. The discretion format is:

$$|J_n| = \mu_n n|_{mid} \frac{|\phi_{n,v_2} - \phi_{n,v_1}|}{\overline{v_1 v_2}} \tag{9.132}$$

$$|J_p| = \mu_p p|_{mid} \frac{|\phi_{p,v_2} - \phi_{p,v_1}|}{\overline{v_1 v_2}} \tag{9.133}$$

The numerical result is smooth.

Here we can do some more simplify. When impact ionization takes evident place, which means where is strong enough electric field existing. The drift current driving by electric field will be much stronger than diffusion current. If we omit diffusion current and only consider drift current, the final result difference is less than 10%. Accordingly when the request is not strict, the previous two formulae can be simplified as:

$$|J_n| = \mu_n n|_{mid} \frac{|\psi_{v_2} - \psi_{v_1}|}{\overline{v_1 v_2}} \tag{9.134}$$

$$|J_p| = \mu_p p|_{mid} \frac{|\psi_{v_2} - \psi_{v_1}|}{\overline{v_1 v_2}} \tag{9.135}$$

For left part of shade area of Figure (9.14), which belongs to node v_1's Voronoi volume, impact ionization integration at this region can be written as:

$$\int_{shadow} G^{II} dV = \frac{d3}{4} \left[\alpha_n \mu_n n|_{mid} \left(|\phi_{n,v_2} - \phi_{n,v_1}|\right) + \alpha_p \mu_p p|_{mid} \left(|\phi_{p,v_2} - \phi_{p,v_1}|\right)\right] \tag{9.136}$$

Triangle's other region can follow this treatment.

In GSS, GradQf model is the default impact ionization model. It has good convergence and high accuracy for diode and bipolar transistors. But when current moving does not follow electric field direction, eg. "??", on page ??, GradQf model can not gives correct result. For this situation, EdotJ model should be used.

Band-band Tunneling

Tunneling leading carrier generation G^{BB} has following format:

$$G^{BB} = \alpha \cdot \frac{E^2}{\sqrt{E_g}} \cdot \exp\left(-\beta \cdot \frac{E_g^{3/2}}{E}\right) \tag{9.137}$$

It only involves the magnitude of electric intensity **E**. The discretion is relatively easy. The electric field intensity can be calculated with Equation (9.84) and Equation (9.85).

9.11 Boundary Condition Processing

From the theory of partial differential equations, we know that the PDEs can converge to physical solution when correct boundary conditions are given.

The boundary condition specification is a very important and complicated question in semiconductor simulation. The boundary condition should be flexibly

selected based on different device structure and simulation state. GSS supports many boundary conditions, roughly be divided into electrodes and interfaces.

Electrode boundaries include

```
Ohmic contact electrode
Schottky contact electrode
Gate contact electrode
Simple gate contact electrode
```

Each electrode can be contacted to external lumped-parameter devices, other electrodes or SPICE circuit. And we can select voltage stimulation or current stimulation to electrode. Current stimulation model is suitable for high inject simulation and/or when current is a multiple-value function of voltage, for which the voltage stimulation will lead to branch.

Besides, GSS supports material interfaces, including

```
Semiconductor - Oxide interface
Semiconductor - Metal interface
Semiconductor - Semiconductor interface, heterogenous junction
Semiconductor - Semiconductor interface, homogenous junction
Neumann Boundary
```

9.11.1 Neumann boundary

Neumann boundary exists on the surface of semiconductor body or artificially introduced boundary far from active region. For this boundary type, carrier will not be able to go through the boundary, and electric field perpendicular to the interface is zero. There is $\hat{n} \cdot \nabla \psi = 0$.

For boundary, Voronoi cell is only half as the complete inner cell, shown in the Figure (9.15).

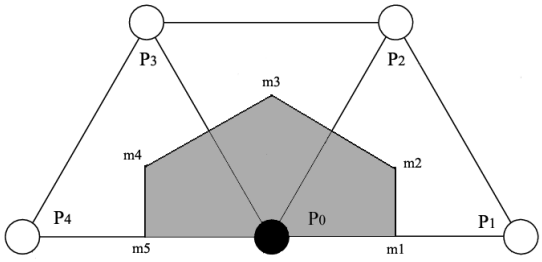


Figure 9.15: Boundary's voronoi

Due to the Neumann boundary condition, the flux function $\mathbf{F}$ on the segment $\overline{m_1 m_5}$

$$\mathbf{F} = \begin{pmatrix} \varepsilon \nabla \psi \\ \frac{1}{q}\mathbf{J}_n \\ -\frac{1}{q}\mathbf{J}_p \end{pmatrix}$$

gives no contribution to Voronoi cell P_0. As a result, we can safely skip this boundary.

We need to pay attention that we can not set all the boundary as Newmann boundary condition. In this case the device is "floating". Since the electrostatic

potential ψ does not have a reference "zero", it can be any value. For example, add any function ψ' satisfying $\nabla^2\psi' = 0$ to ψ, the original Poisson's equation will still hold. From mathematical point of view, in order to give ψ a unique solution, we have to give the first or the third boundary condition some where. From physical point of view, we have to give at least one electrode boundary condition of a device.

9.11.2 Ohmic contact electrode

Ohmic electrode is the most commonly used electrode boundary. It is implemented as Dirichlet boundary conditions, where electrostatic potential ψ, electron concentration n, and hole concentrations p are fixed. Minority and majority carrier quasi-Fermi potentials are equal to the applied bias of the electrode.

$$\phi_n = \phi_p = V_{app} \tag{9.138}$$

where, the relationship of Fermi potential and Fermi level is $E_{Fn} = -q\phi_n$, $E_{Fp} = -q\phi_p$.

For non-degenerated carriers satisfying Boltzmann statistics, starting from charge balance condition:

$$n + N_A = p + N_D \tag{9.139}$$

The relationship of Fermi potential and carrier concentration are given by Equation (4.2) and Equation (4.3). We substitution them into above equation will yield:

$$n = \frac{N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_{ie}^2}}{2} \tag{9.140}$$

$$p = \frac{N_A - N_D + \sqrt{(N_D - N_A)^2 + 4n_{ie}^2}}{2} \tag{9.141}$$

and

$$\psi_{intrinsic} = \phi_n + \frac{k_b T}{q}\ln\left(\frac{n}{n_{ie}}\right) = \phi_p - \frac{k_b T}{q}\ln\left(\frac{p}{n_{ie}}\right) = V_{app} + \frac{k_b T}{q}\operatorname{asinh}\left(\frac{N_D - N_A}{2n_{ie}}\right) \tag{9.142}$$

Noticing in GSS, relationship of ψ and $\psi_{intrinsic}$ are given by Equation (7.4). The electrostatic potential in the real code is

$$\psi = V_{app} + \frac{k_b T}{q}\operatorname{asinh}\left(\frac{N_D - N_A}{2n_{ie}}\right) - \frac{\chi}{q} - \frac{E_g}{2q} - \frac{k_b T}{2q}\ln\left(\frac{N_c}{N_v}\right) \tag{9.143}$$

For degenerate condition, we need to consider Fermi Statistics. GSS solves the following equations:

$$\begin{cases} \mathrm{F}_\psi(\psi, n, p) = N_c \mathrm{F}(\eta_n) + N_A^+ - N_v \mathrm{F}(\eta_p) - N_D^+ = 0 \\ \mathrm{F}_n(\psi, n, p) = n - N_c \mathrm{F}(\eta_n) = 0 \\ \mathrm{F}_p(\psi, n, p) = p - N_v \mathrm{F}(\eta_p) = 0 \end{cases} \tag{9.144}$$

where, N_A^+ and N_D^+ are the effect doping concentration where incomplete ionization are considered. Other variables are shown below:

$$\eta_n = \frac{-q\phi_n - E_c}{k_b T} = \frac{-qV_{app} - E_c}{k_b T} \tag{9.145}$$

$$\eta_p = \frac{E_v + q\phi_p}{k_b T} = \frac{E_v + qV_{app}}{k_b T} \tag{9.146}$$

$$E_c = -q\psi - \chi \tag{9.147}$$

$$E_v = E_c - E_g \tag{9.148}$$

For above three equations, only ψ, n and p are independent variables. By using Newton's iteration method, we can solve Equation (9.144) numerically for ψ, n and p. Attention that F_ψ should not be written as $F_\psi = n + N_A^+ - p - N_D^+$. Although it is the same in maths, however during Newton iteration, the diagonal item of its Jacobian matrix 0, which is easy to precondition failure.

Total current density flow out of ohmic electrode is the sum of current flux of all the boundary Voronoi cells, shown in figure Equation (9.16). Since the carrier

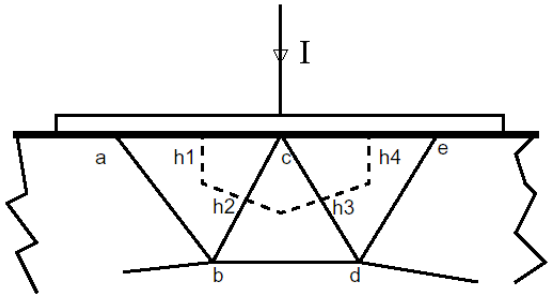


Figure 9.16: Ohmic contact electrode's total current density

density of Ohmic boundary keeps unchanged, current flow into ohmic boundary is the same as current flow out of ohmic boundary. In Figure (9.16), current density flow out of Voronoi cell c is

$$J = h_1 J_{ca} + h_2 J_{cb} + h_3 J_{cd} + h_4 J_{ce} \tag{9.149}$$

where displacement current density and conduction current density are both considered. Using j_{ca} as an example:

$$J_{ca} = J_{n,ca} + J_{p,ca} + \varepsilon_s \frac{\partial E_{ca}}{\partial t} \tag{9.150}$$

Because GSS is a 2D model, the third dimensional depth Z.Width are defined, so that we can transfer current density to current. As a result, the external circuit can be considered and work together with GSS.

9.11.3 Schotkey contact electrode

Schottky contact electrode's boundary needs metal's work function **WORKFUNC**. Electric potential's definition is the following:

$$\psi = V_{app} - \mathbf{WORKFUNC} \tag{9.151}$$

Here electric potential is still Dirichlet boundary condition.

It is worthy to mention that because interface recombination rate can not be infinity, ϕ_n, ϕ_p will not be equal to V_{app}. In order to obtain carrier density equation,

we introduce Schottky interface's ejection current density [52]:

$$J_{sn} = qv_{sn}(n_s - n_{eq}) \tag{9.152}$$

$$J_{sp} = qv_{sp}(p_s - p_{eq}) \tag{9.153}$$

where J_{sn} and J_{sp} are current density through Schottky interface. n_s and p_s are electron, hole density. n_{eq} and p_{eq} is electron, hole density with assumption of surface infinite recombination rate.

$$n_{eq} = N_c \exp\left(\frac{-q\Phi_B}{k_b T}\right) \tag{9.154}$$

$$p_{eq} = N_v \exp\left(\frac{-E_g + q\Phi_B}{k_b T}\right) \tag{9.155}$$

where Φ_B is potential height. v_{sn} and v_{sp} are surface recombination velocity:

$$v_{sn} = \frac{A_n^* T^2}{qN_c} \tag{9.156}$$

$$v_{sn} = \frac{A_p^* T^2}{qN_v} \tag{9.157}$$

Where A_n^* and A_p^* are electron and hole effective Richardson coefficients separately. In order to consider mirror force correction and tunneling effects on potential decrease, GSS has the following correction [8]:

$$\Delta\Phi_B(E) = \sqrt{\frac{qE}{4\pi\varepsilon_{semi}}} + \alpha \cdot E^{\gamma} \tag{9.158}$$

Where, E is interface electric intensity's absolute value. α and γ's typical value can be found in [8]. So Equation (9.154) and Equation (9.155)'s electron potential correction is $\Phi_B - \Delta\Phi_B$, hole potential correction is $\Phi_B + \Delta\Phi_B$.

After obtain current density through interface, electron and hole density's value can be obtain from continuous equation. And Schottky electrode's total current density value is equal to the sum of every voronoi's ejection current density and displacement current density.

9.11.4 Semiconductor insulator interface

Semiconductor insulation interface is quite general. For example MOS structure, SOI structure and etc. GSS software supports semiconductor insulation layer interface, and we provide a relative easy method for MOS device gate electrode.

In fact, semiconductor insulator interface for carriers is a solid wall. For Continuous equation, semiconductor insulator interface is Neumann type boundary. All important interface character will be represented by Poisson equation.

For semiconductor insulator interface, Poisson equation's boundary condition is

$$\varepsilon_s \frac{\partial \psi}{\partial n} - \varepsilon_i \frac{\partial \psi}{\partial n} = \sigma \tag{9.159}$$

where ε_s and ε_i are semiconductor and insulator's dielectric constant separately. n is semiconductor to insulator's normal vector. At the insulator side, if there is metal or poly silicon to form gate electrode, then gate and insulator contact forms gate electrode boundary condition. In GSS, gate electrode's electric potential is

$$\psi = V_{app} - \textbf{WORKFUNC} \tag{9.160}$$

This expression accords to Schottky electrode's electric potential boundary condition.

Because MOS device gate oxide thickness is normallly thin, it is difficult for most of the cases. And this think structure is also a challenge for mesh. Normally it leads to too many nodes. GSS provides simplified gate electrode boundary condition. Users only need to select gate oxide thickness d, dielectric constant ε_i, gate electrode work function, **WORKFUNC**, and does not need to show gate oxide layer's modeling. But this model will not take care of oxide fix charge. In figure Figure (9.17), put Poisson equation's boundary condition to Equation (9.159) oxide layer side's directional derivative can be approximated as

$$\varepsilon_i \frac{\partial \psi}{\partial n} = \varepsilon_i \frac{V_{app} - WORKFUNC - \psi}{d} \tag{9.161}$$

So, semiconductor insulator interface silicon part boundary condition is

$$\varepsilon_i \frac{V_{app} - WORKFUNC - \psi}{d} - \varepsilon_s \frac{\partial \psi}{\partial n} = \sigma \tag{9.162}$$

This is the third type boundary condition.

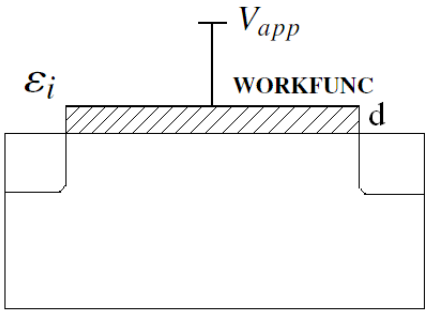


Figure 9.17: Simple MOS Gate Electrode

9.11.5 External circuit for electrode

Up to now, GSS has three electrode boundary conditions. Ohmic electrode and Schottky electrode can let current flow in-out for both steady state and transient state. But MOS gate dielectric only have displacement current when biased with a variational source.

In order to put semiconductor device at suitable circuit condition, GSS assigns electrode with simple external circuit structure. Ohmic electrode and Schottky electrode can be voltage drive or current driven. However, MOS gate electrode can only have voltage source. Figure (9.18) shows the external circuit for electrodes. When the electrode is voltage driven, the lumped element R, C and L defined by user are considered. And for current driven situation, the electrode only has a current source.

In GSS software, all the electrodes have an additional external circuit current-voltage equation. For this we introduced third dimension width Z.Width to transform current density to current.

After having two types of external circuit structure, GSS software can simulation simple circuit. The influence of parasitical R, C and L caused by inter-connection of a transistor can be directly simulated by GSS. And GSS can take care of typical bipolar or MOS amplification circuit without any problems. For more complicate circuit, GSS can work with SPICE software to do mixed type simulation. Please refer to GSS extension chapter.

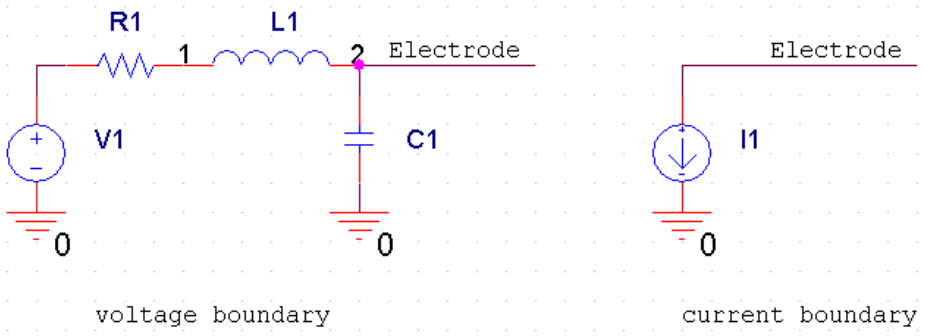


Figure 9.18: Electrode with voltage driven or current driven

9.11.6 Thermal boundary condition

GSS second level solver supports temperature field calculation. We need to introduce corresponding thermal transfer boundary condition. GSS requires two types of boundary conditions for thermal conduction equation: one for interface between two materials and another for thermal conduction at device surface.

The thermal conduction between two material can be performed to i.e. silicon dioxide and silicon interface, which is important for thermal analysis of SOI device. Suppose there is no thermal source on the boundary, the governing equation has the formula as bellow:

$$\kappa_1 \frac{\partial T}{\partial n} - \kappa_2 \frac{\partial T}{\partial n} = 0 \tag{9.163}$$

However at device surface and electrode interface, we allow thermal exchange with external environment and introduce thermal exchange coefficient h to represent the exchange speed. The thermal exchange boundary condition is

$$\frac{\partial T}{\partial n} = h\left(T_{ext} - T\right) \tag{9.164}$$

where T_{ext} is the environment temperature, which is fixed during the simulation.

The default thermal exchange coefficient h for Neumann boundary is zero for default. However electrode usually has a large thermal exchange coefficient h, the default value is set to the thermal exchange coefficient as silicon to copper.

Please attention, we should not set all the boundary's thermal exchange coefficient to zero, which means the system is thermal insulated. For a thermal insulated system has internal thermal source, steady state analysis is obviously not possible to have stable result.

9.11.7 Carrier temperature boundary condition

GSS third level solver involves energy balance equations, which require boundary conditions for electron and hole temperature. GSS uses following assuming: At device surface, the carrier temperature has a zero gradient perpendicular to the interface, which is a Neumann type boundary condition; For the boundary condition at electrode, we force carrier temperature equal to lattice temperature:

$$T_n = T_p = T \tag{9.165}$$

The above assuming is something rough, which requires surface and electrode to be far away from active region. We need to consider this and try to satisfy it during modeling.

9.11.8 Hetero-junction

Heterogenous junction is very complicate in physics, especially the current through heterogenous junction. GSS as the canonical simulation software, can only support heterogenous junction description to certain extent. Heterogenous interface's electrode potential condition is two material interface's Poisson equation. Figure Figure (9.19) shows heterogenous junction's energy band diagram. We can see that except vacuum energy level, conduction band, valance band and intrinsic Fermi energy all have discontinuous effect in the interface. Because electric potential can not have sudden change, adopting vacuum potential as semiconductor's Poisson equation's variable is the most suitable case.

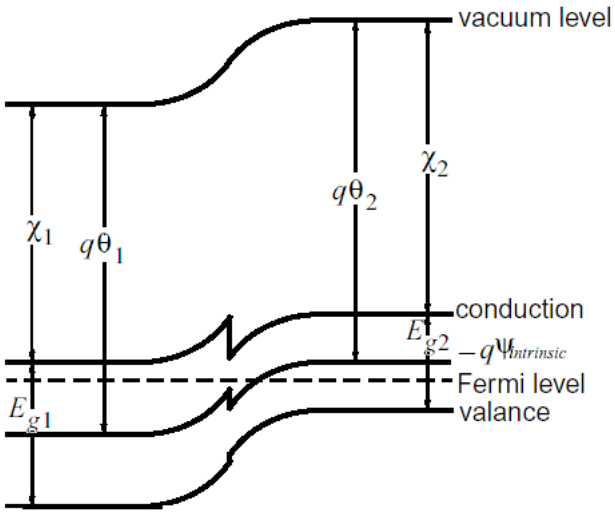


Figure 9.19: Heterogenous junction energy band diagram

Figure Figure (9.19) shows vacuum potential and intrinsic Fermi potential's relationship.

$$\psi = \psi_{vacuum} = \psi_{intrinsic} - \theta \tag{9.166}$$

where, θ represents energy band parameter

$$\theta = \frac{\chi}{q} + \frac{E_g}{2q} + \frac{k_b T}{2q} \ln\left(\frac{N_c}{N_v}\right) \tag{9.167}$$

AT the boundary, Poisson equation satisfy boundary condition

$$\varepsilon_{s1}\frac{\partial \psi}{\partial n} - \varepsilon_{s2}\frac{\partial \psi}{\partial n} = 0 \tag{9.168}$$

Here we do not consider interface state leading surface charge.

Heterogenous junction's current adopts k. Hess and G.J. Iafrate proposed model [53]. Assume electron jumping from material 1 to material 2 through potential

barrier, which means $E_{c2} > E_{c1}$, now the electron current can be represented as:

$$J_{n,1->2} = A \cdot T_1^2 \cdot \exp\left(\frac{E_{Fn1} - E_{c2}}{k_b T_1}\right) \tag{9.169}$$

$$J_{n,2->1} = A \cdot T_2^2 \cdot \exp\left(\frac{E_{Fn2} - E_{c2}}{k_b T_2}\right) \tag{9.170}$$

where A is Richardson coefficient. Hole current has the similar format.

9.11.9 Boundary condition for DG-DDM

Here we discuss the boundary condition for quantum potential equations of density gradient model.

In practical, quantum effect usually only happens at certain limited region, for example, the MOS inversion layer or quantum wells of resonant tunneling diode (RTD). The quantum region needs to be small enough to march the wave length of electron and the quantum effect will decreases sharply outside this small region. As a result, the quantum effect far away from quantum region is not important.

The electrodes and surface boundaries are usually far away from quantum regions. The boundary condition for quantum potential is not very critical. Here we can assume quantum conduction band and quantum valance band energy at electrode equal to conduction band and valance band energy, respectively.

$$\begin{aligned} \Lambda_n &= \frac{E_{qc} - E_c}{q} = 0 \\ \Lambda_p &= \frac{E_{qv} - E_v}{q} = 0 \end{aligned} \tag{9.171}$$

And for surface boundary, we assume Neumann type boundary condition: $\partial_n \Lambda_n = \partial_n \Lambda_p = 0$.

For MOS, the quantum effect happens at Si/SiO_2 interface. The boundary condition for quantum potential at this interface needs careful consideration. For a simple MOS structure we mentioned in "Semiconductor insulator interface", on page 128, since SiO_2 layer is not explicitly build, we need the truncated formulae of Equation (9.124) at the interface. According to the result quantum WBK equation, electron density into SiO_2 will degrade with the following relationship [54]:

$$n(x) = n_0 \exp\left(-2x/x_{np}\right) \tag{9.172}$$

where

$$x_{np} = \frac{\hbar}{\sqrt{2 m_{nox} \Phi_{Bn}}} \tag{9.173}$$

is the characteristic penetration depth of electron obtained from the WKB equation. Here $m_{nox} = 0.4 m_0$ is the effective mass in SiO_2. $\Phi_{Bn} = 3.15\text{eV}$ is the electron potential barrier height. As a result, Si/SiO_2 interface has the following relationship:

$$b_{nox} \nabla \sqrt{n} = -\frac{b_{nox}}{x_{np}} \sqrt{n_0} \tag{9.174}$$

where

$$b_{nox} = \frac{\hbar^2}{6 q m^*_{nox}} \tag{9.175}$$

And $m^*_{nox} = 0.14m_0$ is electron effective mass in SiO_2.

Holes should have similar relationship. However currently the data for hole is not enough, we can only make sure $\Phi_{Bp} = 4.10\text{eV}$. The author assume $m_{pox} = 0.4m_0$, $m^*_{pox} = 1.0m_0$. After having detail data we will do correction.

For constructing SiO_2 with poly silicon gate device, we can treat SiO_2 as wide band semiconductor material and solve DG-DDM equations on Si bulk, SiO_2 layer and poly silicon gate consistently, keeping quantum potential continuous for each interface.

Besides, for device with quantum wells deposited by different band gap materials, which also has quantum potential continuity at the material interface.

9.12 Transient Simulation

When space discretization finished, DDM equations are converted into large scale ordinary differential equations (ODEs):

$$\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}t} = F(\mathbf{Q}) \tag{9.176}$$

For steady-state simulation, there is $\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}t} = 0$. However, if transient simulation is required, we must consider suitable time discretization method. We had already mentioned in "Constants in Semiconductors", on page 52 that explicit method always has a strict time step limitation, usually at femto-second level, which is limited in real application. Practical code must adopt absolute stable implicit algorithm, so that we can use relatively large time step. Furthermore, due to the high stiffness of ODEs arising from space discretion of semiconductor equations, the time discretion algorithm should not only A stable but also L stable [55].

A Stable

First, we should choose ODE discretization scheme with A Stable. The most simple and famous formulas here are first order implicit Euler (EB) method and second order Crank-Nichloson (CN) method:

$$\frac{y_{n+1} - y_n}{\Delta t} = f\left(y_{n+1}\right) \tag{9.177}$$

and

$$\frac{y_{n+1} - y_n}{\Delta t} = \frac{1}{2}\left(f\left(y_n\right) + f\left(y_{n+1}\right)\right) \tag{9.178}$$

For model equation:

$$y' = \lambda y\,, \quad \mathrm{Re}(\lambda) \le 0 \tag{9.179}$$

consider stability functions of both methods:

$$R_{\mathrm{EB}}(z) = \frac{1}{1-z} \tag{9.180}$$

and

$$R_{\mathrm{CN}}(z) = \frac{1 - \frac{z}{2}}{1 + \frac{z}{2}} \tag{9.181}$$

where, $z = \Delta t\lambda$, and Δt is the time step.

The previous two formulae both satisfy $|R(z)| \leq 1$, which means they are absolute stable, with no restrictions to the time step.

L Stable

However for stiff problems, which satisfies $\lambda \gg \Delta t$, resulting in very large $z = \Delta t\lambda$, the result is completely changed. Here we take z to its limit, which turns to be infinity, the stability function of 1st order implicit Euler method becomes:

$$\lim_{z\to-\infty} R_{\mathrm{EB}}(z) = \lim_{z\to-\infty} \frac{1}{1-z} = 0 \tag{9.182}$$

and for 2nd order Crank-Nichloson method:

$$\lim_{z\to-\infty} R_{\mathrm{CN}}(z) = \lim_{z\to-\infty} \frac{1+\frac{z}{2}}{1-\frac{z}{2}} = -1 \tag{9.183}$$

Since the exact solution of Equation (9.179) is $\lim_{z\to-\infty} e^z = 0$, it is obvious that Crank-Nicholson has problems.

We notice $y^{n+1} = R(z)y^n$, which means 2nd order Crank-Nichloson method will lead to oscillation result, which is demonstrated in Figure (9.20). However, the implicit Euler method keeps correct for this limitation.

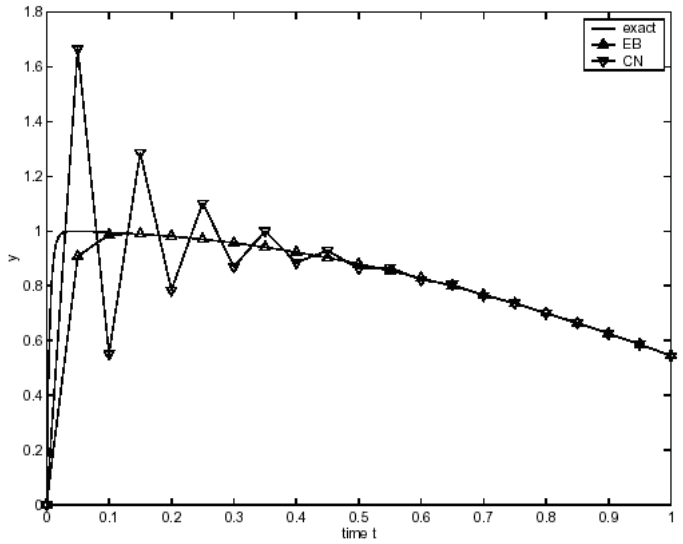


Figure 9.20: numerical result of model equation $y' = -200(y - \cos t)$, $y(0) = 0$

Define a scheme is L-stable if $\lim_{z\to-\infty} R(z) = 0$. The conclusion is although 2nd order Crank-Nichloson method is A stable, but not L stable. Numerical experiment did prove it will lead to oscillation in semiconductor simulation. So it can not be adopted in our code.

Although implicit Euler method both satisfies A stability and L stability, it is a 1st order accurate algorithm, which has relatively large local truncated error. Since the error will be accumulated during every time step, the 1st order method can not ensure long time simulation accurate. For practical, we should find some higher order method.

BDF2 and TR-BDF2

Second order algorithm satisfying the A and L stable includes backward differentiation formulae 2 (BDF2) and trapezoidal BDF2 (TR-BDF2). BDF2 is a two step algorithm, which needs relatively larger memory size:

$$3y^{n+1} - 4y^n + y^{n-1} = 2\Delta t f(y^{n+1}) \tag{9.184}$$

TR-BDF2 is a single step format, yet slightly more accurate than BDF2. However, for each time step it requires two Newton iterations, computation load is much higher than BDF2:

$$\begin{aligned} y^* &= y^n + \frac{\Delta t}{4}(f(y^n) + f(y^*)) \\ 3y^{n+1} - 4y^* + y^n &= \Delta t f(y^{n+1}) \end{aligned} \tag{9.185}$$

CFD code often chooses BDF2 algorithm against TR-BDF2. First it is because the burden of hydrodynamics calculation is heavy. It is reasonable to buy speed with big memory size. The second reason is the numerical viscosity introduced by truncated error of BDF2 often helps smooth the numerical solution. PISCES-IIB adopts TR-BDF2 algorithm [56]. That is because at 1980s memory size is very limited. The PISCES code was requested to be run on machine with only 8MB memory. However, MEDICI code, which is the commercial version of PISCES-IIB, uses BDF2 method.

GSS adopts both Euler and BDF2 algorithm. Because BDF2 needs current and previous result, implicit Euler method should be used for the first time step calculation. After that, BDF2 can start to work.

Since the truncated error of implicit Euler method is $O(\Delta t)$, the first time step should keep small enough to avoid large truncated error. After BDF2 is activated, truncated error is $O(\Delta^2 t)$. Time step can be larger. Further more, GSS has an automatic time step selection algorithm for speed up the transient simulation and control the error. Please refer to the next section.

Because the time step can be variational, BDF2 needs the following correction:

$$\frac{1}{t_{n+1} - t_{n-1}}\left(\frac{2-r}{1-r}y^{n+1} - \frac{1}{r(1-r)}y^n + \frac{1-r}{r}y^{n-1}\right) = f\left(y^{n+1}\right) \tag{9.186}$$

where

$$r = \frac{t_n - t_{n-1}}{t_{n+1} - t_{n-1}} \tag{9.187}$$

9.13 Automatic Time Step Control

The local truncated error (LTE) based automatic time step control is a widely used technique in ODE numerical solution. Its first application in semiconductor simulation can be retrospected to BANK etc in 1985 [?]. Due to the importance of transient simulation, GSS also adopts this technique.

LET of Implicit Euler Method

For estimating the LET of 1st order implicit Euler method, its semi-discrete scheme can be written down:

$$\frac{x_{n+1} - x_n}{h_n} = f(x_{n+1}) \tag{9.188}$$

where $h_n = t_{n+1} - t_n$. The LTE of this scheme is:

$$\mathrm{LTE} = \frac{x_{n+1} - x_n}{h_n} - \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{h_n^2}{2}\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + O\left(\frac{\mathrm{d}^3 x}{\mathrm{d}t^3}\right) \tag{9.189}$$

Clearly, for getting the LTE, one should evaluate the second order derivative of $\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}$. For this purpose, we linear interpolate the predict value of $n+1$ time step x_{n+1}^p from x_{n-1} and x_n:

$$x_{n+1}^p = \left(1 + \frac{h_n}{h_{n-1}}\right)x_n - \frac{h_n}{h_{n-1}}x_{n-1} \tag{9.190}$$

Noticing that from Taylor series, the difference between x_{n+1} and x_{n+1}^p is

$$x_{n+1} - x_{n+1}^p = \frac{h_n (h_n + h_{n-1})}{2} \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + O\left(\frac{\mathrm{d}^3 x}{\mathrm{d}t^3}\right) \tag{9.191}$$

Thus, the LTE can be expressed as:

$$\mathrm{LTE}(BE) = \frac{h_n}{h_n + h_{n-1}} \left(x_{n+1} - x_{n+1}^p\right) \tag{9.192}$$

LET of BDF2 Method

For BDF2 scheme, we can get the LET by the same procedure [57]. The LTE of its semi-discrete scheme is

$$\mathrm{LTE} = -\frac{h_n^2 (h_n + h_{n-1})}{6} \frac{\mathrm{d}^3 x}{\mathrm{d}t^3} + O\left(\frac{\mathrm{d}^4 x}{\mathrm{d}t^4}\right) \tag{9.193}$$

here, we need to construct a second order predict value of x_{n+1}^p. As a result, the previous value x_n, x_{n-1} and x_{n-2} are used to do a second-order polynomial interpolation:

$$x_{n+1}^p = c_1 x_n + c_2 x_{n-1} + c_3 x_{n-2} \tag{9.194}$$

where

$$\begin{aligned} c_1 &= 1 + \frac{h_n (h_n + 2h_{n-1} + h_{n-2})}{h_{n-1} (h_{n-1} + h_{n-2})} \\ c_2 &= -\frac{h_n (h_n + h_{n-1} + h_{n-2})}{h_{n-1} h_{n-2}} \\ c_3 &= \frac{h_n (h_n + h_{n-1})}{h_{n-2} (h_{n-1} + h_{n-2})} \end{aligned}$$

Here we also use Taylor series to expend $x_{n+1} - x_{n+1}^p$:

$$x_{n+1} - x_{n+1}^p = \frac{h_n}{6} (h_n + h_{n-1}) (h_n + h_{n-1} + h_{n-2}) \frac{\mathrm{d}^3 x}{\mathrm{d}t^3} + O\left(\frac{\mathrm{d}^4 x}{\mathrm{d}t^4}\right) \tag{9.195}$$

As a result, the LTE can be expressed as:

$$\mathrm{LTE}(BDF2) = \frac{h_n}{h_n + h_{n-1} + h_{n-2}} \left(x_{n+1} - x_{n+1}^p\right) \tag{9.196}$$

Time Step Control Based on LET

The time step control should satisfy the LTE of each time step be limited in a certain level:

$$\mathrm{LTE} < \mathrm{E}_{user} \tag{9.197}$$

In the application, E_{user} can be expressed in terms of a relative error tolerance and an absolute error tolerance parameter ε_r and ε_a, respectively.

$$\mathrm{E}_{user} = \varepsilon_r \left|x_{n+1}\right| + \varepsilon_a \tag{9.198}$$

The default value in GSS code are $\varepsilon_r = 10^{-3}$, $\varepsilon_a = 10^{-4}$.

The time step can be controlled by considering the relative error of the allowable error E_{user} and the actual local error LTE.

$$r = \frac{\mathrm{LTE}}{\mathrm{E}_{user}} = \frac{C_{k+1} h_n^{k+1} x^{(k+1)}(t_n)}{C_{k+1} h_{allowable}^{k+1} x^{(k+1)}(t_n)} = \left(\frac{h_n}{h_{allowable}}\right)^{k+1} \tag{9.199}$$

where k is the order of ODE scheme, for implicit Euler $k = 1$, and for BDF2 $k = 2$.

The relative error r is estimated by following strategy in the GSS code. Since the Poisson's equation does not time relatively, it is not considered in LET estimation. The time related continuation equations of DD model are estimated for r:

$$r = \left[\frac{1}{N}\sum\left(\frac{\mathrm{LET}\,(f_n, f_p)}{\varepsilon_r\,|n, p| + \varepsilon_a}\right)^2\right]^{1/2} \tag{9.200}$$

For convenient, defining $r_{LE} = r^{-\frac{1}{k+1}}$ here. The time step selection criterion in GSS code is now described.

- If $r_{LE} < 0.9$, then the current result is rejected. GSS uses new time step $h_n^* = h_n \cdot r_{LE}$ to re-calculate the solution for t_{n+1}.
- For values of $r_{LE} \geq 0.9$, the current solution is considered acceptable and the new time step is taken to be $\min(h_n \cdot r_{LE}, 2h_n, h_{max})$. Where h_{max} is the maximal time step defined by user.

Note:

Since the prediction of x_{n+1}^p needs previous values, the automatic time step control can only start at third time step for Euler method and forth time step for BDF2. Besides, the predict value can be a good initial value of x_{n+1}, which can improve the convergence and save computation time.

9.14 Nonlinear Solver: Newton's Iteration Method

When space, time are discretized, semiconductor drift diffusion model forms nonlinear linear algebra equation set. Semiconductor numerical method's last step is to solve this up to thousands order's non-linear equation set.

In history, semiconductor model has non coupling method Gummel method [58] and coupling method Newton two methods. Gummel method does not solve all the equations together, it uses iterate gradually technique. First assume carrier concentration is constant, solve Poisson equation, then put Poisson equation's result into two continuous equation to solve the carrier's concentration. Keep doing iteration until convergence. Gummel method is a poor method, which needs small memory size and is also fast in certain circumstances. But this non coupling algorithm needs many iterations when equations' coupling is strong. For example current is mainly contributed by voltage drive drift current, which needs a lot of iteration to converge. On the other hand, Newton method considers all the equation sets together, which has better stability and fast convergence speed, each iteration cost is high. Newton method's disadvantage is that it needs to construct and save complicate Jacobian matrix. Every iteration needs to solve huge linear equation set. Fortunately as computer memory and performance's improvement, solving huge linear equation set requested memory and time turns to be acceptable. So Newton method is generally used. GSS only provide full coupled Newton method's support. Generally, non-linear equation solving is to search as solution vector $\mathbf{x}$ to let $f(\mathbf{x}) = \mathbf{0}$. Non-linear equation's solving is always a difficult problem. From method point of view, there is stable iteration method, fastest decrease method and Newton method. Currently the best non-linear solution method is not Newton method, although Newton method has

different improved versions. Here we introduces Newton method and its improved version, for further understanding non-linear equation's solution theory, please refer to [59].

9.14.1 Line Search method

Newton method needs to calculate $f(\mathbf{x})$ and its gradients $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$, through solving linear equation

$$\mathbf{g}^T(\mathbf{x}) \cdot \mathbf{p}_k = -f(\mathbf{x}_k) \tag{9.201}$$

we obtain $\mathbf{p}_k$Ît$f(\mathbf{x}_k)$'s Newton decrease vector. Newton method's iteration is

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k \tag{9.202}$$

Iterate until certain convergence condition is satisfied. For example

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leqslant \beta \|\mathbf{x}_k - \mathbf{x}^*\| \tag{9.203}$$

where β is certain positive constant.

Newton iteration's convergence and initial value is correlated to $\mathbf{g}(\mathbf{x})$. If initial value is inside real solution's convergence region. But for many conditions accurately calculating $\mathbf{g}(\mathbf{x})$ is still very difficult. And accurate initial value is not easy to obtain. In order to improve Newton method's generality, there are two different techniques: linear search and trust region, which be help to improve convergence performance [59][60]. Linear search method is essentially a one dimensional minimization problem, shown in figure Figure (9.21). We notice formulae Equation

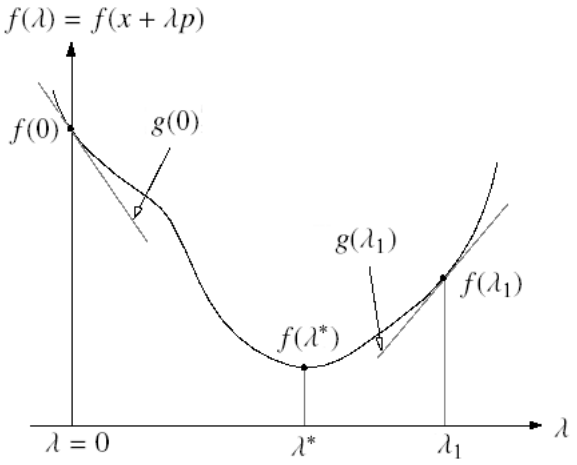


Figure 9.21: Linear search method's illustration diagram

(9.201) solved $\mathbf{p}_k$ represents $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_k$ decrease direction. Rewrite Equation (9.202) as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda \mathbf{p}_k \tag{9.204}$$

where, λ is a tunable parameter, representing $\mathbf{p}_k$ direction's searching step. If $\lambda = 1$, it is basic Newton method. And typical linear searching method uses three times multi nominal interpolation so that $f(\mathbf{x})$ reaches minimum value through $\mathbf{p}_k$ direction.

In fact algorithm defines slope $g(\lambda) = \mathbf{g}^T(\mathbf{x}_k + \lambda \mathbf{p}_k) \cdot \mathbf{p}_k$. Linear search normally starts from $\lambda = \lambda_1$ą$\lambda = \lambda_2$'s function value $f_1 = f(\mathbf{x}_k + \lambda_1 \mathbf{p}_k)$, $f_2 = f(\mathbf{x}_k + \lambda_2 \mathbf{p}_k)$ and slope $g_1 = g(\lambda_1)$, $g_2 = g(\lambda_2)$, forming third order multiple nominal

$$p(\lambda) = a(\lambda - \lambda_1)^3 + b(\lambda - \lambda_1)^2 + c(\lambda - \lambda_1) + d \tag{9.205}$$

where

$$a = \frac{-2\left(f_2 - f_1\right) + \left(g_1 + g_2\right)\left(\lambda_2 - \lambda_1\right)}{\left(\lambda_2 - \lambda_1\right)^3}$$
$$b = \frac{3\left(f_2 - f_1\right) - \left(2g_1 + g_2\right)\left(\lambda_2 - \lambda_1\right)}{\left(\lambda_2 - \lambda_1\right)^2}$$
$$c = g_1$$
$$d = f_1$$

So by solving $p\left(\lambda\right)$'s minimum value we can obtain

$$\lambda = \lambda_1 + \frac{-b + \sqrt{b^2 - 3ac}}{3a} \tag{9.206}$$

Now we need $a \neq 0$ and $b^2 - 3ac > 0$. If the previous condition is not satisfied, we need to use f_1, f_2 and g_1 for 2nd order interpolation to obtain

$$p\left(\lambda\right) = b\left(\lambda - \lambda_1\right)^2 + c\left(\lambda - \lambda_1\right) + d \tag{9.207}$$

Now $p\left(\lambda\right)$'s minimum value requests

$$\lambda = \lambda_1 - \frac{c}{2b} \tag{9.208}$$

Because we used f_1, f_2 and g_1 to construct $p\left(\lambda\right)$, $b = 0$ shows $p\left(\lambda\right)$ is not a linear function.

When λ is fixed, we still need a set of mechanism to judge whether interpolation's λ is acceptable. Armijo's judgement is relatively convenient and generally used [59], which needs

$$\begin{cases} f\left(\mathbf{x}_k + \lambda \mathbf{p}_k\right) \leqslant f\left(\mathbf{x}_k\right) + \alpha\lambda \mathbf{g}^T\left(\mathbf{x}_k\right)\mathbf{p}_k \\ \left|\, \mathbf{g}^T\left(\mathbf{x}_k + \lambda \mathbf{p}_k\right)\mathbf{p}_k \,\right| \leqslant \left|\, \beta \mathbf{g}^T\left(\mathbf{x}_k\right)\mathbf{p}_k \,\right| \end{cases} \tag{9.209}$$

where$0 < \alpha < \beta < 1$, α represent the upper limit of λ from function value point of view, β represents the lower limit for λ. GSS's non-linear solver default value $\alpha = 10^{-4}$, $\beta = 0.9$. Normally it starts from $\lambda_1 = 0$, $\lambda_2 = 1$. If λ is accepted, this iteration is finished, otherwise decrease λ_2 to search a acceptable λ.

The previous discussion is for single non-linear equation. If for equation set, there is similar conclusion. Now single equation's gradients $\mathbf{g}(\mathbf{x})$ turns to be equation set's Jacobian matrix. And λ needs to cater each equation. From the discussion above, searching with direction $\mathbf{p}_k$ is always considered to be accurate. So gradients $\mathbf{g}(\mathbf{x})$ or equation set's Jacobian matrix needs to be accurate or approximately accurate, otherwise searching method is easy to fail.

9.14.2 High speed decrease method

Before introducing Trust Reign method, we must discuss about fastest decrease method, which is proposed by French scientist Cauchy, also called Cauchy method. Assume non-linear equation $f(\mathbf{x})$ gradients is $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$, target function is continuous and has derivative at close to $\mathbf{x}_k$. So expand $f(\mathbf{x})$ with Taylor expansion at $\mathbf{x}_k$.

$$f(\mathbf{x}) = f\left(\mathbf{x}_k\right) + \mathbf{g}_k^T\left(\mathbf{x} - \mathbf{x}_k\right) + o\left(\|\mathbf{x} - \mathbf{x}_k\|\right) \tag{9.210}$$

É è$\mathbf{x} - \mathbf{x}_k = \mathbf{d}_k$čňÉÏÊ¡£ÉÒÔÐťşÉ

$$f(\mathbf{x}) = f\left(\mathbf{x}_k\right) + \mathbf{g}_k^T\mathbf{d}_k + o\left(\|\mathbf{d}_k\|\right) \tag{9.211}$$

If $\mathbf{g}_k^T \mathbf{d}_k < 0$, then $\mathbf{d}_k$ is the decreasing direction, which leads $f(\mathbf{x}_k + \mathbf{d}_k) < f(\mathbf{x}_k)$. In order to reach fastest decrease speed, we need $\mathbf{g}_k^T \mathbf{d}_k$ to be the minimum value. We can prove that only when $\mathbf{d}_k = -\mathbf{g}_k$, $\mathbf{g}_k^T \mathbf{d}_k$ has the minimum value. So we called $-\mathbf{g}_k$ the fastest decrease direction. So the fastest decrease method's iteration is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{g}(\mathbf{x}_k) \tag{9.212}$$

Although this method is easy, it decreases very slowly, we can prove its convergence speed is linear.

9.14.3 Trust Region method

Trust region needs no accurate Jacobian matrix $\mathbf{J}$, consider the following Newton iteration

$$\mathbf{J} \cdot \Delta \mathbf{x}_k = -f(\mathbf{x}_k) \tag{9.213}$$

If Jacobian is not accurate, then we use $\mathbf{B} = \mathbf{J} + \lambda \mathbf{I}$ to replace Jacobian , where λ is a very big value. Obviously $\mathbf{J}$'s value is covered by $\lambda \mathbf{I}$, $\Delta \mathbf{x}_k$ goes through $-f(\mathbf{x}_k)$'s direction. Newton iteration now turns to be fastest decrease method, and the step $\| \Delta \mathbf{x}_k \|$ turns to be very small. We can prove that if the step is enough small the iteration can be converged at $\mathbf{J}$. This is Trust region method.

In reality, trust region method assume $\mathbf{x}_k$ as the current iteration point. Then use $\mathbf{x}_k$ as center, δk as radius's close sphere region solve a sub problem. Where δk is decided by the specific problem, which always needs to be given.

Assume $\mathbf{d}_C$ is the fastest decrease searching direction, then the testing step $\mathbf{d}_k$ is

$$\mathbf{d}_k = \mathbf{d}_C + \lambda (\mathbf{d}_N - \mathbf{d}_C) \tag{9.214}$$

We can see testing step is composed by fastest decrease method and Newton method. If Jacobian matrix is odd, testing step will turns to be fastest decrease method. In searching process, we need to solve testing step $\mathbf{d}_k$ in the trust region. It means at every iteration testing step needs to satisfy $\| \mathbf{d}_k \| < \delta k$. So $\lambda \in [0, 1]$ and choose testing step satisfying $\| \mathbf{d}_k \| < \delta k$'s maximum value. After obtaining $\mathbf{d}_k$, we use an evaluation function $\mathbf{d}_k$ to decide whether it it acceptable or not. Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$, or decrease the trust region radius. In real practice, initial δk value is not important. If initial δk is too small trust region method is easily converged to regional optimized solution.

9.14.4 Jacobian matrix's construction

Newton method needs to construct Jacobian Matrix. This step in GSS cost the author a lot of energy. We need to consider DDML1's equation set.

$$\begin{aligned} F_\psi(\psi, n, p) &= 0 \\ F_n(\psi, n, p) &= 0 \\ F_p(\psi, n, p) &= 0 \end{aligned} \tag{9.215}$$

Its Jacobian matrix has the following format

$$\begin{Bmatrix} \frac{\partial F_\psi}{\partial \psi} & \frac{\partial F_\psi}{\partial n} & \frac{\partial F_\psi}{\partial p} \\ \frac{\partial F_n}{\partial \psi} & \frac{\partial F_n}{\partial n} & \frac{\partial F_n}{\partial p} \\ \frac{\partial F_p}{\partial \psi} & \frac{\partial F_p}{\partial n} & \frac{\partial F_p}{\partial p} \end{Bmatrix} \tag{9.216}$$

For N nodes problem, DDML1 equation set has $3N$ equations. And Jacobian matrix has $3N \times 3N$ order. But because differential symbol's regional characteristics, every nodes equation is only related to its neighbor nodes. During solving derivative, every nodes' equation needs only process for its own node or its neighbor nodes. So Jacobian matrix is sparse. Generally, every nodes' neighbor has $5 \sim 7$, corresponding matrix width is 3*(neighbor nodes number+1), approximately around $18 \sim 24$. Drift diffusion model discretion leads non-linear equation set easy to solve difference. In GSS code, Jacobian matrix is manually written, which cost a lot of energy to make sure the accuracy. So GSS suggest to use Line search method. In fact calculating this method leads to fast convergence. Iteration number is normally less than 10. Trust region method's calculation normally needs longer time than Line search method. But for certain, when the Jacobian matrix is almost singular, the problem is solved more efficiently.

If there is a new algorithm needs to be insert to GSS, first we need to adopt self difference to replace manual difference to obtain Jacobian matrix, called Matrix-Free method [61]. Put Newton method solved linear equation set

$$f'(\mathbf{x}) \cdot \Delta \mathbf{x}_k = -f(\mathbf{x}_k) \tag{9.217}$$

left hand side as difference approximation

$$f'(\mathbf{x}) \cdot \Delta \mathbf{x}_k \approx \frac{f(\mathbf{x}_k + h\Delta \mathbf{x}_k) - f(\mathbf{x}_k)}{h} \tag{9.218}$$

where h is a tunable parameter. Matrix-Free method does not use obvious formulae to construct Jacobian matrix. In real practice, it is easier. But because it only can obtain matrix and solution vector's product, we can only use iteration method to solve the equation set. Its disadvantage is very time consuming and difference method is still not accurate, sometimes it leads to convergence problem. Generally when new algorithm is proven to be effective, we can use manual code to calculate Jacobian matrix more accurately, so that to increase the speed.

9.14.5 DDM equation set accurate convergence criterion

Non linear equation set $f(\mathbf{x}) = 0$'s criterion has two illustration method: absolute convergence and relative convergence. Absolute convergence is $f(\mathbf{x})$'s mode value less than a certain value.And relative convergence means the solutions $\mathbf{x}$ between two steps are less than certain fixed value ε_r:

$$\frac{\| \mathbf{x}_{k+1} - \mathbf{x}_k \|}{\| \mathbf{x}_k \|} < \varepsilon_r \tag{9.219}$$

These two convergence criterions are all used in practice. When semiconductor outside bias is relatively low, the current through it is less, minority carrier's quantity has almost no contribution to current value. Now adopting absolute convergence criterion is suitable. If we force minority carrier's relative variation less than certain value, we need more iterations, or maybe the machine accuracy will affect so that it can not be reached. But when inject current is big, because numerical error leads to current continuous equation's absolute convergence difficult to meet, relative error is not related to current. Now relative convergence condition is more useful. When we do forward IV curve calculation for diode, we can find that low bias GSS report is absolute converge, after going into conduction region GSS reports relative convergence. GSS software request drift diffusion equation absolute convergence criterion:

Equation type	Convergence Criterion (2-Norm)

Poisson Equation	10^{-29}C/μm
Electron and Hole continuous equation	5×10^{-18}A/μm
Thermal conduction equation	10^{-11}W/μm
Electron and hole's energy balance equationąą	10^{-18}W/μm
Electrode boundary condition's outside circuit equation	10^{-9}V

Table 9.1: Absolute convergence criterion

For relative convergence request is all variables 2-Norm two steps iterations' relative variation ε_r less than 10^{-5}, simultaneously with the calculation accuracy, equation set must satisfy relax 4 order's absolute criterion. When either absolute convergence or relative convergence is met, GSS consider Newton iteration is converged.

9.15 Linear solver: Krylov Subspace Method

Non-linear iteration's every step needs to solve a linear equation set, which consume a lot to time during GSS solving. Linear equation set's solving method has direct method (LU decomposition). Fix point iteration algorithm (Gauss-Seidel iteration, super relax iteration and so on) and Krylov subspace iteration method (conjugate gradients class, minimum residual class).

Semiconductor device model equation set discretion formed coefficient matrix is sparse not band shape. LU decomposition can not make use of it. Calculation is at $O(N^3)$ level. Fix point iteration algorithm converges very slowly. Currently it is not recommended to use. Krylov subspace's algorithm only asks matrix vector to product, for this type of problem, calculation load is at $O(KMN)$ level, where K is iteration order, normally tens of iteration will lead to convergence. M is matrix's bandwidth. Between 18 ~ 24. From here we know Krylov type algorithm can decrease the computation speed to the extreme. For linear equation set solving, there are two problems need to be balanced: first is matrix condition number, second is floating calculation cutting error. Big condition number will lead to LU algorithm fail and increase Krylov type algorithm's iteration number. Through variable scaling, put most of the carrier's concentration around 1 will decrease condition number, but because semiconductor device electron-hole product is a constant, this will lead to minority carrier's absolute value turns to be smaller, which leads to minority carrier concentration sensitive to floating calculation cutting error. So variable's scaling needs to be controlled in certain scope. Generally, drift diffusion model's linear equation set first choose minimal residual type's Transport Free Quasi-Minimal Residual (TFQMR) and Generalized Minimal Residual (GMRES) method. Their solving convergence process is smoother. Conjugate Gradient Squared method (CGS), Bi-Conjugate Gradient (BICG), Bi-Conjugate Gradient Stabilized (BCGS) and etc.'s convergence speed is fast. But without minimize the residual, convergence process is not reasonable, and has strong vibration. Calculation process density's middel value can be negative. Drift diffusion model although is resistant to negative density, it still needs additional process. When problem scale is relatively small (less than 1000 nodes), we can select LU method. Current commercial software, Medici integrates LU and CGS method, Dessis adopts LU and TFQMR method. GSS internal linear solver PETSC[62] includes LU method and almost all the Krylov subspace methods for users to choose.

9.15.1 Conjugate direction method

Conjugate direction method is conjugate gradients type method's basement. We are going to introduce it in this chapter. More details can be referred to [?][?].

Assume linear equation set is

$$Ax = f \tag{9.220}$$

where $A \in R^{n,n}, f \in R^n$, A symmetric and normal. In order to further illustrate the problem, first we introduce vector conjugate regarding matrix concept.

Given a symmetrical normal matrix B, called any two non-zero vector x, y are regarding matrix B conjugate (or B normal cross), if

$$x^T By = 0 \tag{9.221}$$

Now assume for formulae Equation (9.220)'s matrix A, there is n non zero vector $p_1, p_2, \cdots, p_n$, satisfying

$$(p_i, Ap_j) = 0, \ \forall i, \ j, \ i \neq j \tag{9.222}$$

then $p_1, p_2, \cdots, p_n$ regarding A normal cross or A conjugate. Obviously these n non zero vector is linear non correlated. So if x^* is equation Equation (9.220)'s solution, x_0 is any vector, $x^* - x_0$ can be linear combination of these p_i.

$$x^* - x_0 = \sum_{k=1}^{n} \alpha_k p_k \tag{9.223}$$

where α_k are constant. Equation's both sides time A, we have

$$f - Ax_0 = A(x^* - x_0) = \sum_{1}^{n} \alpha_k A p_k \tag{9.224}$$

Let $r_0 = f - Ax_0$, do inner product for r_0 and p_k, together with Equation (9.222), we have

$$\alpha_k = \frac{(r_0, p_k)}{(Ap_k, p_k)}, \ k = 1, 2, \cdots, n \tag{9.225}$$

If

$$x_k = x_0 + \sum_{i=1}^{k} \alpha_i p_i, \ k = 1, 2, \cdots, n \tag{9.226}$$

then

$$x_k = x_{k-1} + \alpha_k p_k, \ k = 1, 2, \cdots, n \tag{9.227}$$

From Equation (9.227) we know, if we can find n normal cross vector regarding A, $p_1, p_2, \cdots, p_n$, then select a initial point x_0, we can start from Equation (9.227)'s iteration to obtain equation set Equation (9.220)'s solution. Where α_k is given by Equation (9.225). Assume we don't consider iteration process' cutting error, after n step's iteration we can have the accurate solution.

The following problem is how to fix n normal cross vector $p_1, p_2, \cdots, p_n$ of A. According to Gram-Schmidt normal cross method, For r_0 we use the following

deduction, which can lead to A normal cross n vectors.

$$\begin{aligned} p_1 &= r_0 \\ p_{k+1} &= Ap_k - \sum_{i=1}^{k} \beta_{i,k} p_i \\ \beta_{i,k} &= \frac{(A^2 p_k, p_i)}{(Ap_i, p_i)} \end{aligned} \tag{9.228}$$

Consider Equation (9.225)-Equation (9.228) we have the conjugate direction method's step below:

Step1: initialization

$$r_0 = f - Ax_0 \qquad p_1 = r_0 \tag{9.229}$$

Step2: $k = 1, \cdots, n$

$$\begin{aligned} \alpha_k &= \frac{(r_0, p_k)}{(Ap_k, p_k)} \\ x_k &= x_{k-1} + \alpha_k p_k \\ \beta_{i,k} &= \frac{(A^2 p_k, p_i)}{(Ap_i, p_i)}, \ i = 1, 2, \cdots, k \\ p_{k+1} &= Ap_k - \sum_{i=1}^{k} \beta_{i,k} p_i \end{aligned} \tag{9.230}$$

[?] gives expand subspace theory. From this theory we can directly visualize conjugate direction method. We can prove that solving linear equation set problem is similar as solving target function.

$$J(x) = \frac{1}{2} x^T Ax - f^T x \tag{9.231}$$

minimum value problem.

Expand subspace theory let x_0 to be randomly selected initial point, $p_1, p_2, \cdots, p_n$ is non zero conjugate A vector. If solution series x_k is generated from conjugate direction method Equation (9.229) Equation (9.230), then x_k is at line $x_{k-1} + \alpha_k p_k$ and line group $x_0 + Span\{p_1, p_2, \cdots, p_k\}$ direction, leading minim function $J(x)$ value.

Expand subspace theory shows conjugate method's every iteration step finds higher dimensions minimum point. Then obviously after calculate n iterations, we find function $J(x)$'s minimum point at R^n. Figure 9.22 shows conjugate direction method at three dimension space's searching process. First step start from x_0, through p_1 direction search and find minimum point x_1. Second step, from x_1, through p_2 direction search to have minimum point of p_1, p_2 fixed plane. Third step, starting from x_2, through p_3 direction search to have the minimum point from p_1, p_2, p_3 fixed plane.

9.15.2 Conjugate gradient method

Conjugate gradients method is based on conjugate direction method. The difference from conjugate direction method is n conjugate direction's fixing method is different. In conjugate direction method, conjugate direction is predefined. Apply Gram-Schmidt normal cross method, in conjugate gradients method, conjugate

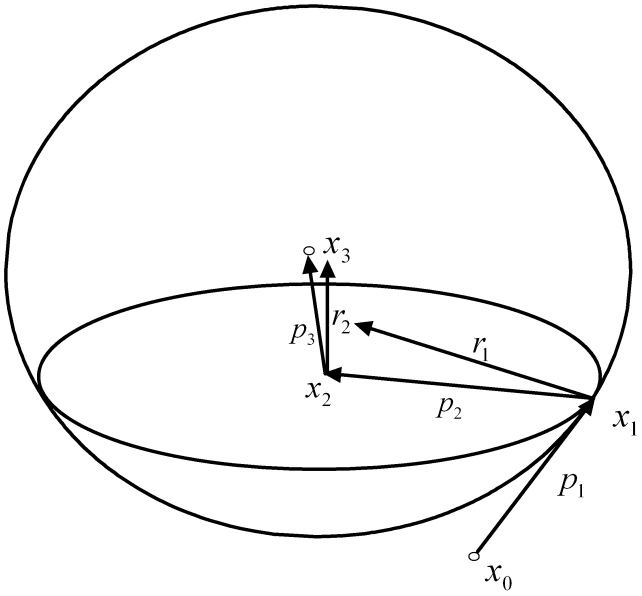


Figure 9.22: Conjugate direction method searching illustration

direction vector's update uses every iteration's gradients vector. In detail in conjugate gradients algorithm, conjugate direction vector's update formulae is

$$p_{k+1} = r_k - \beta_{k+1} p_k \tag{9.232}$$

where $r_k = f - Ax_k$. As shown in figure 9.22, p_2 is fixed by r_1 and p_1, p_3 is fixed by r_2 and p_2.

Conjugate gradients method's algorithm step is shown below:

Step 1: initialization

$$r_0 = f - Ax_0 \qquad p_1 = r_0 \tag{9.233}$$

Step 2: $k = 1, 2, \cdots$,

$$\begin{aligned} \alpha_k &= \frac{(r_k, p_k)}{(Ap_k, p_k)} \\ x_k &= x_{k-1} + \alpha_k p_k \\ r_k &= b - Ax_k \\ \beta_{k+1} &= \frac{(Ap_k, r_k)}{(Ap_k, p_k)} \\ p_{k+1} &= r_k - \beta_{k+1} p_k \end{aligned} \tag{9.234}$$

When linear equation set coefficient matrix's condition number is relatively big, conjugate gradients algorithm's convergence speed is very slow. In order to increase the convergence speed, during solving equation, we need to treat the original equation. This treatment method is generally called preprocessing technique.

Preprocessing's basic concept is to transfer the original equation to a equal equation set.

$$\tilde{A}x = \tilde{f} \tag{9.235}$$

And equation set Equation (9.235)'s coefficient matrix is the biggest, minimum feature value is far less than original equation's corresponding part, so that it can accelerate conjugate gradients algorithm convergence. Normally preprocessing technique has poly nominal preprocessing and incomplete factor decomposition method. Here we introduce incomplete factor decomposition method mainly.

Incomplete factor decomposition method's starting point is to select a matrix M, which is very close to A^{-1}, put the original equation equal to

$$MAx = Mf \tag{9.236}$$

The simplest method is to select M as A's diagonal part's reverse. And mark $M^{1/2} = diag[1/\sqrt{a_{11}}, 1/\sqrt{a_{22}}, \cdots, 1/\sqrt{a_{nn}}]$čň Write Equation (9.236)as

$$\tilde{A}\tilde{x} = \tilde{f} \tag{9.237}$$

where $\tilde{A} = M^{1/2}AM^{1/2}, \tilde{x} = M^{-1/2}x, \tilde{f} = M^{1/2}f$. After using conjugate gradients method to solve $\tilde{x}$, it is easy to obtain x. For a big type of matrix from partial differential equation discretion, the previous method can really decrease A's condition number.

9.15.3 Double conjugate gradient method

In previous introduced conjugate direction method algorithm, we need linear equation set coefficient matrix A to be symmetrical and normal. Double conjugate gradient method is suitable for coefficient matrix A not as normal matrix. This method's every step has two searching directions $p, \bar{p}$, which are conjugate to A and satisfying

$$\begin{aligned} \bar{p}_i^T A p_j = p_i^T A \bar{p}_j = 0, \qquad & i \neq j \\ \bar{r}_i^T r_j = r_i^T \bar{p}_j, \qquad & i \neq j \\ \bar{r}_i^T p_j = r^T \bar{p}_j^T, \qquad & j < i \end{aligned} \tag{9.238}$$

Double conjugate gradients method's algorithm step is given below:

Step1: initialization

$$p_0 = r_0 \qquad \bar{p}_0 = \bar{r}_0 \tag{9.239}$$

Step 2čž$k = 1, 2, \cdots,$

$$\begin{aligned} \alpha_k &= \frac{\bar{r}_k^T r_k}{(\bar{p}_k^T A p_k)} \\ r_{k+1} &= r_k - \alpha_k A p_k \\ \bar{r}_{k+1} &= \bar{r}_k - \alpha_k A^T \bar{p}_k \\ \beta_k &= \frac{\bar{r}_{k+1}^T r_{k+1}}{\bar{r}_k^T r_k} \\ p_{k+1} &= r_{k+1} + \beta_k p_k \\ \bar{p}_{k+1} &= \bar{r}_{k+1} + \beta_k \bar{p}_k \\ x_{k+1} &= x_k + \alpha_k p_k \\ \bar{x}_{k+1} &= \bar{x}_k + \alpha_k \bar{p}_k \end{aligned} \tag{9.240}$$

Conjugate gradients method is not suitable for non symmetrical system. This is because we can not make every step's residual vector cross normal. Double

conjugate method adopts another method, by using two crossed series to replace cross normal's residual vector. The scarification is that it may not provide the minimum value.

In real practice, double conjugate gradients method convergence process can be non stable. And even leads to breakdown. This condition can use look first strategy to solve. This method will lead to algorithm realization difficulty. Breakdown problem can also adopt other brutal method to solve, such as GMRES method.

9.15.4 GMRES

In real engineering problem, linear equation set's coefficient matrix is not symmetrical and normal. They are structured sparse matrix. Solving these equations is the focus of current research. There are several algorithms already. But theocratically every algorithm has shortcoming. Now we introduce general minimum residual method's basic knowledge. In GMRES we need to use Arnoldi process, so we give arnoldi process first.

For any given vector r_0, mark $K_m = Span\{r_0, Ar_0, \cdots, A^{m-1}r_0\}$, Arnoldi process in fact is to construct space K_m's normal cross base procedure.

Arnoldi process

1. define $v_1 = v/ \parallel v \parallel_2$.

2.

$$\begin{aligned} &for: j = 1, 2, \cdots, m \\ &\quad w = Av_j \\ &\quad h_{i,j} = v_i^T w,\ w = w - h_{i,j}v_i,\ i = 1, 2, \cdots, j \\ &\quad do\{ \\ &\qquad h_{j+1,i} = \parallel w \parallel_2 \\ &\quad \}while(h_{j+1,i} \neq 0) \\ &\quad v_{j+1} = w/h_{j+1,j} \\ &end \end{aligned} \tag{9.241}$$

define matrix $V_m = [v_1, v_2, \cdots, v_m]$čňmatrix H_m is

$$H_m = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2,m-1} & h_{2m} \\ 0 & h_{32} & \cdots & h_{3,m-1} & h_{3m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & h_{m,m-1} & h_{mm} \end{pmatrix} \tag{9.242}$$

$$\bar{H}_m = \begin{pmatrix} H_m \\ h_{m+1,m}e_m^T \end{pmatrix} \tag{9.243}$$

where $e_m^T = (0, 0, \cdots, 1)$.

GMRES algorithm step is shown below:

Step 1: initialization

choose $x_0 \in R^n$ and calculate

$$r_0 = b - Ax_0 \qquad v_1 = r_0/ \parallel r_0 \parallel \tag{9.244}$$

Step 2: ű$j = 1, 2, \cdots, k$, calculate until satisfying

Using Arnoldi process to obtain $v_i, i = 1, 2, \cdots, k$

step 3: solving least square problem

$$\min_{y \in R^n} \| \beta e_1 - \bar{H}_k y \| \tag{9.245}$$

obtain's result isy_k.

Step 4: Calculate$x_k = x_0 + V_k y_k$.

In GMRES method, how to solve minimum $\| \beta e_1 - \bar{H}_k y \|$ is a key question, he detail method can be referred to[?]ąč

In fact, it is difficult to compare GMRES and BICG. GMRES method's minimized residual, but calculation load is big, which needs big memory size. BICG method does not minimize residual, but its accuracy is similar as GMRES. BICG method needs to solve twice of matrix vector product, which costs less memory, but its stability is not as good as GMRES.

Chapter 10 Functional Extension of GSS

10.1 AC Small Signal Model

Besides basic steady-state and transient simulation functions, GSS also supports small signal AC sweep as the post processing after a steady-state simulation. In real application, small signal AC sweep is used to evaluate the band width of amplifier, cut frequency of filter and so on.

Since small signal AC sweep is the post processing of the steady-state simulation, user should first get the status of semiconductor device under certain DC bias before input the small sine signal. The typical amplitude of the small signal is 0.0026 V, so that we can consider the status of semiconductor device is not significantly changed from DC situation. By using Taylor expansion, the effect of the small signal is linearized and solved. The detail procedure is given below.

Suppose we bias the electrode as:

$$V = V_0 + \tilde{V}e^{j\omega t} \tag{10.1}$$

where V_0 is the biased DC voltage, $\tilde{V}$ is the amplitude of the small sine signal, ω is the angular frequency of this signal. The DC calculation result satisfies following semiconductor equations:

$$F_\psi(\psi, n, p) = \nabla \cdot \varepsilon \nabla \psi + q(p - n + N_D - N_A) = 0 \tag{10.2}$$

$$F_n(\psi, n, p) = \frac{1}{q}\nabla \cdot \mathbf{J}_n - U = \frac{\partial n}{\partial t} = 0 \tag{10.3}$$

$$F_p(\psi, n, p) = -\frac{1}{q}\nabla \cdot \mathbf{J}_p - U = \frac{\partial p}{\partial t} = 0 \tag{10.4}$$

and AC small signal's solution can be written as

$$\psi_i = \psi_{i0} + \tilde{\psi}_i e^{j\omega t} \tag{10.5}$$

$$n_i = n_{i0} + \tilde{n}_i e^{j\omega t} \tag{10.6}$$

$$p_i = p_{i0} + \tilde{p}_i e^{j\omega t} \tag{10.7}$$

where ψ_{i0}, n_{i0} and p_{i0} are mesh point i's DC result. And $\tilde{\psi}$, $\tilde{n}_i$ and $\tilde{p}_i$ are corresponding AC signal's value. Generally speaking, they are all complex numbers. Substitute (10.5)-(10.7) into semiconductor basic equations. Based on small signal approximation, expand Taylor expansion to the first order, the three semiconductor equations has the following expression:

$$F_\psi(\psi, n, p) = F_\psi(\psi_0, n_0, p_0) + \frac{\partial F_\psi}{\partial \psi}\tilde{\psi}e^{j\omega t} + \frac{\partial F_\psi}{\partial n}\tilde{n}e^{j\omega t} + \frac{\partial F_\psi}{\partial p}\tilde{p}e^{j\omega t} = 0 \tag{10.8}$$

$$F_n(\psi, n, p) = F_n(\psi_0, n_0, p_0) + \frac{\partial F_n}{\partial \psi}\tilde{\psi}e^{j\omega t} + \frac{\partial F_n}{\partial n}\tilde{n}e^{j\omega t} + \frac{\partial F_n}{\partial p}\tilde{p}e^{j\omega t} = j\omega\tilde{n}e^{j\omega t} \tag{10.9}$$

$$F_p(\psi, n, p) = F_p(\psi_0, n_0, p_0) + \frac{\partial F_p}{\partial \psi}\tilde{\psi}e^{j\omega t} + \frac{\partial F_p}{\partial n}\tilde{n}e^{j\omega t} + \frac{\partial F_p}{\partial p}\tilde{p}e^{j\omega t} = j\omega\tilde{p}e^{j\omega t} \tag{10.10}$$

We notice $F(\psi_0, n_0, p_0) = 0$ satisfies DC solution, so we can obtain the linear equations about AC small signal at semiconductor region:

$$\begin{bmatrix} \frac{\partial F_\psi}{\partial \psi} & \frac{\partial F_\psi}{\partial n} & \frac{\partial F_\psi}{\partial p} \\ \frac{\partial F_n}{\partial \psi} & \frac{\partial F_n}{\partial n} - j\omega & \frac{\partial F_n}{\partial p} \\ \frac{\partial F_p}{\partial \psi} & \frac{\partial F_p}{\partial n} & \frac{\partial F_p}{\partial p} - j\omega \end{bmatrix} \begin{bmatrix} \tilde{\psi} \\ \tilde{n} \\ \tilde{p} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{10.11}$$

For the electrode, the following equations should be satisfied:

$$\begin{cases} \tilde{\psi} = \tilde{P} \\ \tilde{n} = 0 \\ \tilde{p} = 0 \end{cases} \tag{10.12}$$

where $\tilde{P}$ is the potential at the electrode. The relationship of potential $\tilde{P}$ and application voltage $\tilde{V}$ is decided by the electrode. GSS's default electrode circuit diagram is shown in Figure (10.1). Accordingly we need another equation to

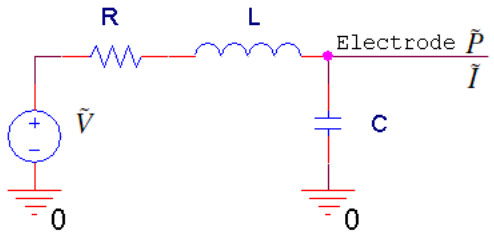


Figure 10.1: The electrode in AC sweep

describe the lumped elements:

$$\tilde{V} - (Z_1 Y_2 + 1)\tilde{P} = Z_1 \tilde{I} \tag{10.13}$$

where $Z_1 = R + j\omega L$, $Y_2 = j\omega C$ are the lumped impedance and conductance of the electrode. $\tilde{I}$ is the current injected into the electrode.

The matrix of linear equations (10.11) is the Jacobian matrix for the DC simulation minus $j\omega$ for main diagonal. If we have a linear solver supports complex number, the problem is solved. However if we have only have a real linear solver, another step is required. Write (10.11) as real format:

$$\begin{bmatrix} J & -D \\ D & J \end{bmatrix} \begin{bmatrix} X_R \\ X_I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{10.14}$$

where J is the Jacobian matrix in DC condition, X_R and X_I are AC solution's real and imagine part. D is a diagonal matrix

$$D = \begin{bmatrix} 0 & & \\ & -\omega & \\ & & -\omega \end{bmatrix} \tag{10.15}$$

In GSS code, AC solver shares the same Jacobian matrix with DDML1E solver, accordingly AC solver's calling must follow DDML1E steady-state solving. Each

solving of Equation (10.14), we can obtain semiconductor's respondence to a sine signal for a certain frequency. In the application we may need to sweep the frequency domain characteristics of the device for a wide bandwidth. From (10.15) we know (10.14)'s ω is not at the main diagonal, Accordingly when ω is high, matrix will lose diagonal domination, condition number becomes worse obviously. Generally, when sweep frequency is close to cutoff frequency, convergence turns to be difficult [63].

10.2 Circuit Level Mixed-type Simulation

GSS is a device level simulation software, which can only deal with single transistor. Since version 0.45, GSS designs an interface to SPICE software to expand GSS's application in circuit simulation area. In real application, the circuit simulation is under the control of SPICE, the less important devices are simulated directly by SPICE compact model, while GSS is used to simulate some key devices for accurate numerical result. When multiple numerical devices in circuit are required to be simulated at device level, SPICE can control many GSS processes at the same time, with each simulates one device's IV characteristic.

SPICE software is originally developed by UC Berkeley with Fortran language in 1972. The first version SPICE2G is published to public domain in 1975[1]. In 1985, Berkeley rewrote SPICE by C language, the final version SPICE3f5 is published in 1994. Afterwards, Berkeley stopped the development. Today, SPICE has became the defacto industry standard for circuit simulation. We can find many commercial versions of SPICE all over the world, which little difference with each other.

Because the previous SPICE has dozens years history, some of the codes are not compatible to modern compiler. SourceForge supports NGSPICE project, which is to keep SPICE up-to-date. The NGSPICE developers corrected many bugs of SPICE3F5 and added some new features [64]. At the same time, SourceForge supports GNUCap (GNU Circuit Analysis Package) proejct, which is a C++ software to do advance digital and analog mix simulation. Its circuit component is compatible with SPICE model and simultaneously provides good characteristics. GSS currently can work with NGSPICE for circuit-device mixed mode simulation. In the near future, GSS might be accepted by GNUCap as a plug-in for semiconductor device simulation. In mixed mode development, a lot of help obtained from NGSPICE/GNUCap development team, we appreciate them here.

Now we first introduce how to build circuit equations. And based on that, we will give the core arithmetic of NGSPICE. In the end, we will introduce the interface of GSS to NGSPICE and the necessarily changes at GSS end.

10.2.1 Circuit nodal analyze method

Currently commercial version of SPICE can analyze circuit with thousands of components. People are always amazed by its power. Here we are going to discuss the basic rule of this powerful software, which is fairly simple.

All circuit analysis course will mention Kirchhoff's two principles: nodal current conservation law and branch voltage conservation law. In practice, using node to describe circuit's topology structure is more easier than using the directed loop.

SPICE software uses nodal current conservation law to construct circuit equation. Each electrical component is considered as a branch of the circuit, with its endpoint connected to another branch at circuit node. The circuit equation is

[1] It is recognized as the first open source software.

constructed by branch current voltage characteristics. Since the basic independent variable is the voltage for each node, this method is called nodal analysis method [65].

Here uses a simple circuit on Figure (10.2) for example: Based on Kirchhoff nodal

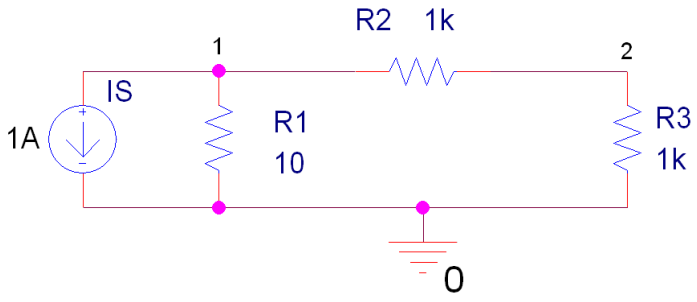


Figure 10.2: Node analysis method circuit diagram

current conservation law, the flow in and flow out current of each node should be the same. We can obtain equations for node 1 and node 2:

$$-I_s + \frac{V_1}{R_1} + \frac{V_1 - V_2}{R_2} = 0 \tag{10.16}$$

$$\frac{V_2 - V_1}{R_2} + \frac{V_2}{R_3} = 0 \tag{10.17}$$

For simplification, we transform the equation to matrix format:

$$\begin{bmatrix} \frac{1}{R_1} + \frac{1}{R_2} & -\frac{1}{R_2} \\ -\frac{1}{R_2} & \frac{1}{R_2} + \frac{1}{R_3} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} I_s \\ 0 \end{bmatrix} \tag{10.18}$$

Where matrix at left hand side is called circuit's conductance matrix since every item has the dimension as conductance. The current source in the circuit is placed at right hand side. By solving this matrix, we can obtain the voltage of each node in the circuit.

Computer can build conductance matrix in an automatical way. Computer will scan each device in the circuit. When it scans the current source I_s between node 1 and ground, it will evaluate current scaler at node 1 on the right side of equation. When it scans resister 2, assuming its two nodes are $R+$ and $R-$, the relationship of current and the modal voltage can be expressed as:

$$I_{R+} = \frac{V_{R+} - V_{R-}}{R_2} \tag{10.19}$$

$$I_{R-} = \frac{V_{R-} - V_{R+}}{R_2} \tag{10.20}$$

As a result, the conductance matrix of resister is:

$$\begin{bmatrix} \frac{\partial I_{R+}}{\partial V_{R+}} & \frac{\partial I_{R+}}{\partial V_{R-}} \\ \frac{\partial I_{R-}}{\partial V_{R+}} & \frac{\partial I_{R-}}{\partial V_{R-}} \end{bmatrix} = \begin{bmatrix} \frac{1}{R_2} & -\frac{1}{R_2} \\ -\frac{1}{R_2} & \frac{1}{R_2} \end{bmatrix} \tag{10.21}$$

After we have built conductance matrix of R2, computer will insert it into the circuit conductance matrix (the matrix at left hand side of Equation (10.18)) by the global node index of $R+$ and $R-$, which is similar as finite element analysis'

stiffness matrix building. Scanning resistor 1 and resister 3 are disposed comparatively. However, they both have one endpoint grounded. Since the ground potential is consistently 0 as a fixed boundary condition of the circuit, accordingly we only need to calculate $\frac{\partial I}{\partial V}$ for non-grounded endpoint. After scanning all the devices, the circuit equation (10.18) is constructed.

For one device with N endpoints, each node's current can be written as:

$$I_i = I\left(V_1, V_2 \cdots V_N\right) \tag{10.22}$$

Then the transfer matrix is

$$\begin{bmatrix} \frac{\partial I_1}{\partial V_1} & \frac{\partial I_1}{\partial V_2} & \cdots & \frac{\partial I_1}{\partial V_N} \\ \frac{\partial I_2}{\partial V_1} & \frac{\partial I_2}{\partial V_2} & \cdots & \frac{\partial I_2}{\partial V_N} \\ \vdots & & & \\ \frac{\partial I_N}{\partial V_1} & \frac{\partial I_N}{\partial V_2} & \cdots & \frac{\partial I_N}{\partial V_N} \end{bmatrix} \tag{10.23}$$

For any circuit including resistor and current sources (without voltage sources), theocratically by following the method above: each resistor's conductance matrix inserts into circuit conductance matrix' corresponding position, and put the current source to corresponding node equation's right hand side, we can construct the circuit equation. The work after is to solve a linear system.

Nodal analysis method has some limitation, which can not describe a voltage source. Because voltage source's current has no relationship with its voltage. Old version of SPICE gives every voltage source a small resister around $1 \times 10^{-12}\ \Omega$ and force the voltage and current through the resister has certain relationship. New version of SPICE adopts modified nodal analysis method, which allows each voltage source has one more current variable.

Real circuit will not only contain resistor. Capacitor and inductor can be neglected in static calculation. But they are necessary parts for transient circuit. Capacitor and inductor's current voltage relationship includes time derivative. We need certain technique to build the conductance matrix. In reality, the method below can be used for any time derivative related devices.

Capacitor's current voltage relationship:

$$I_C = C\frac{dV_C}{dt} \tag{10.24}$$

The formulae above is a difference equation, SPICE use numerical integration to solve the formulae above. The below is 1st order and 2nd order numerical integration.

$$I_C^{n+1} = \frac{C}{\Delta t^n} V_C^{n+1} - \frac{C}{\Delta t^n} V_C^n \tag{10.25}$$

$$I_C^{n+1} = \frac{2C}{\Delta t^n} V_C^{n+1} - \frac{2C}{\Delta t^n} V_C^n - I_C^n \tag{10.26}$$

Real SPICE code even provides higher order accurate numerical integration method. Their general format is :

$$I_C^{n+1} = \underbrace{\frac{C}{b_{-1}\Delta t^n}}_{g_{eq}} V_C^{n+1} \underbrace{-\frac{a_0 \cdot C}{b_{-1}\Delta t^n} V_C^n - \frac{b_0}{b_{-1}} I_C^n - \frac{b_1}{b_{-1}} I_C^{n-1} - \ldots - \frac{b_k}{b_{-1}} I_C^{n-k}}_{I_{eq}} \tag{10.27}$$

So capacitor's equal circuit diagram is shown (10.3), current voltage relationship can be written as:

$$I_C^{n+1} = g_{eq} \cdot V_C^{n+1} + I_{eq} \tag{10.28}$$

Or according to nodal analysis method, written in matrix method :

$$\begin{bmatrix} +g_{eq} & -g_{eq} \\ -g_{eq} & +g_{eq} \end{bmatrix} \cdot \begin{bmatrix} V_1^{n+1} \\ V_2^{n+1} \end{bmatrix} = \begin{bmatrix} -I_{eq} \\ +I_{eq} \end{bmatrix} \tag{10.29}$$

Inductor device's current voltage relationship:

$$V_L = L\frac{dI_L}{dt} \tag{10.30}$$

In SPICE, inductor can also be illustrated by numerical integration:

$$V_L^{n+1} = \underbrace{\frac{L}{b_{-1}\Delta t^n}}_{r_{eq}} I_L^{n+1} \underbrace{-\frac{a_0 L}{b_{-1}\Delta t^n} I_L^n - \frac{b_0}{b_{-1}} V_L^n - \frac{b_1}{b_{-1}} V_L^{n-1} - \ldots - \frac{b_k}{b_{-1}} V_L^{n-k}}_{V_{eq}} \tag{10.31}$$

Inductor's equal circuit diagram is shown in figure (10.4), which is similar as an voltage source with inner resistor. The inductor's current voltage relationship can be written as:

$$V_L^{n+1} = r_{eq} \cdot I_L^{n+1} + V_{eq} \tag{10.32}$$

Because the existence of voltage source, we have to use voltage source's current as variable. Inductor's matrix format in modified nodal method is :

$$\begin{bmatrix} 0 & 0 & +1 \\ 0 & 0 & -1 \\ +1 & -1 & -r_{eq} \end{bmatrix} \cdot \begin{bmatrix} V_1^{n+1} \\ V_2^{n+1} \\ I_L^{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ V_{eq} \end{bmatrix} \tag{10.33}$$

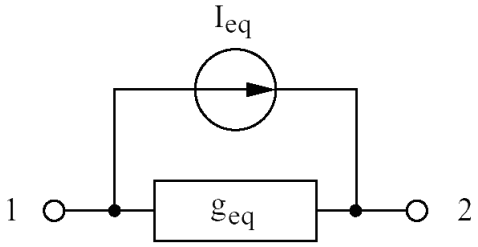


Figure 10.3: Capacitor equal circuit diagram

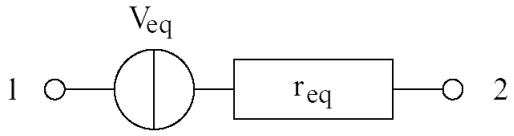


Figure 10.4: Inductor's equal circuit diagram

Generally, current voltage relationship for complicate device is nonlinear. The typical example is a semiconductor diode. The previous linear devices can be treated as the special case for nonlinear devices. We start from more general conditions, deduct circuit solution method. Assume the circuit equation we need to solve is:

$$\mathbf{I}(\mathbf{V}) = 0 \tag{10.34}$$

Where, $\mathbf{V}$ represents the node's voltage scaler, $\mathbf{I}$ represents for each node's current value. If circuit has N nodes, $\mathbf{V}$ and $\mathbf{I}$ all have N factors. By using Newton iteration solving the formulae above, the iteration process is :

$$\mathbf{V}^{n+1} = \mathbf{V}^n - J^{-1}(\mathbf{V}^n)\,\mathbf{I}(\mathbf{V}^n) \tag{10.35}$$

Where, n is the iteration serial number

$$J(\mathbf{V}) = \begin{bmatrix} \frac{\partial I_1}{\partial V_1} & \frac{\partial I_1}{\partial V_2} & \cdots & \frac{\partial I_1}{\partial V_N} \\ \cdots & & & \cdots \\ \frac{\partial I_N}{\partial V_1} & \frac{\partial I_N}{\partial V_2} & \cdots & \frac{\partial I_N}{\partial V_N} \end{bmatrix}$$

In each iteration, current voltage scaler $\mathbf{V}^n$ is known, accordingly the next step's voltage scaler $\mathbf{V}^{n+1}$ can be obtain from (10.35).

SPICE solves the formulae (10.35)'s transformation, we call Jacobian matrix J as conductance matrix G. Formulae (10.35) can be written as:

$$G^n\mathbf{V}^{n+1} = G^n\mathbf{V}^n - I^n \tag{10.36}$$

Since conductance matrix G is corresponding to the linear conductance when the voltage of non-linear device is at $\mathbf{V}^n$. The right hand side's $G^n\mathbf{V}^n - I^n$ is corresponding to equal current source when non linear device voltage is $\mathbf{V}^n$. SPICE needs to obtain each device's liner conductance and equal current source before each iteration. This part information is given by device model. (10.36)'s iteration is end until it is converges.

10.2.2 GSS's circuit mix mode module

Device and circuit mix mode simulation can adopt two method. The first method is to mix the circuit equation and semiconductor equation together to forma a equation set, which is called couping method [66]. The second method is two level iteration method: The outer part is circuit equation iteration, and in each iteration, the device simulation unit are given the node voltage to iterate till convergence. And then calculate the device's transfer matrix and equal current source feedback to outer circuit [67]. Coupling method's total iteration number is less, but its biggest problem is that the number of device is limited. Because all the equation needs to be solved together. For example 10 semiconductor devices, each semiconductor device has 3000 control equation. Then the matrix turns to be huge. The matrix iteration calculation follows $O(n^2)$ increase. When the matrix is very big, each iteration's calculation is very big. Generally when problem is small, for example only one semiconductor device need to be simulated, coupling method has its advantage. Two level iteration method separate problems, which is suitable for large scale problem solution. Although it takes more iterations, it avoids to deal with huge matrix, which decrease the computation for each iteration. Simultaneously, second order iteration method can be used to parallel computation, each numerical device can take a single CPU, SPICE only needs to wait for the slowest device to be calculated for the circuit iteration.

GSS and NGSPICE adopts second layer iteration method. In reality, I introduced a NGSPICE numerical device NDEV. Its is basically a TCP/IP network interface. Shown in figure (10.5), it sends node voltage information to GSS. After GSS receives the signal and calculated the transfer matrix and equal current source, it will send back to NDEV [68][69].

In order to be suitable for SPICE work, GSS needs to provide device's transfer matrix and equal current information. From (10.36) we know, the most important is how to deal with device transfer matrix $G = \frac{\partial i}{\partial V}$. A simple thought is to use the finite difference approximation, for transfer matrix's each factor

$$G_{eq} = \frac{\partial i}{\partial V} \approx \frac{i(V_0 + \Delta V) - i(V_0)}{\Delta V} \tag{10.37}$$

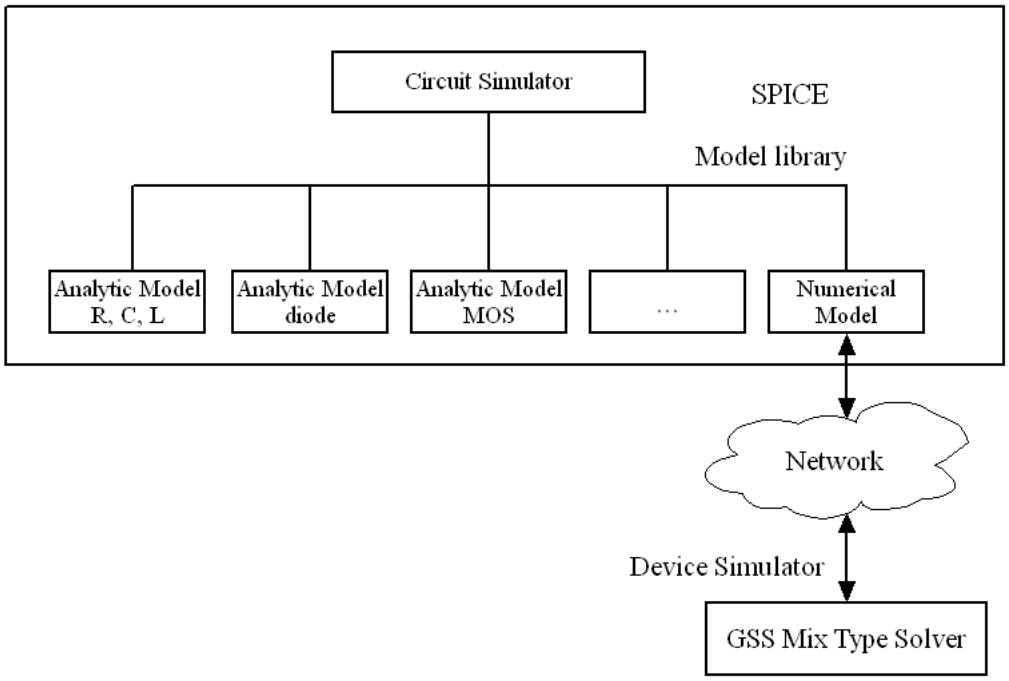


Figure 10.5: Mix mode simulation layer structure

This method although is very easy to realize, when $i(V_0 + \Delta V)$ and $i(V_0)$ are very close, the error is very big. And it is also a problem to fix ΔV.

Based on analytical method to fix G needs a lot of efforts, but it is more accurate. So it is still worthy [57][70]. Put the GSS solved semiconductor equation set as the following format:

$$\mathbf{F}(\mathbf{w}, V) = 0 \tag{10.38}$$

where $\mathbf{w}$ represents semiconductor basic variable, including potential ψ, electron density n, hole density p and temperature T; V illustrates the outer voltage at certain electrode. Noticing semiconductor basic variable $\mathbf{w}$ is dependent on outer voltage, accordingly we can write as $\mathbf{w} = \mathbf{w}(V)$. Device electrode's current can be written as $i = I(\mathbf{w})$, accordingly electrode current does not show dependency to the outer voltage. Adopting series derivative solving method we can explain G_{eq} as

$$G_{eq} = \frac{\partial i}{\partial V} = \frac{\partial I}{\partial \mathbf{w}} \cdot \frac{\partial \mathbf{w}}{\partial V} \tag{10.39}$$

Because electrode current's expression is known, $\frac{\partial I}{\partial \mathbf{w}}$ can be deducted directly. In order to obtain $\frac{\partial \mathbf{w}}{\partial V}$, derivating

$$\mathbf{J}_{\mathbf{w}} \frac{\partial \mathbf{w}}{\partial V} + \frac{\partial \mathbf{F}}{\partial V} = 0 \tag{10.40}$$

Where $\mathbf{J}_{\mathbf{w}}$ is semiconductor control equation set's Jacobian matrix, at GSS iteration we already obtained. $\frac{\partial \mathbf{F}}{\partial V}$ can be obtained through device electrode boundary control equation's symbol differential. Accordingly thourhg solving the prevoius formulae's linear equation set we can obtain $\frac{\partial \mathbf{w}}{\partial V}$. After obtaining $\frac{\partial I}{\partial \mathbf{w}}$ and

$\frac{\partial \mathbf{w}}{\partial V}$, we can solve G_{eq} through (10.39). For multiple electrodes device, transfer matrix has following format:

$$G = \begin{bmatrix} \frac{\partial I_1}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_1} & \frac{\partial I_1}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_2} & \cdots & \frac{\partial I_1}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_N} \\ \cdots & & & \cdots \\ \frac{\partial I_N}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_1} & \frac{\partial I_N}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_2} & \cdots & \frac{\partial I_N}{\partial \mathbf{w}}\frac{\partial \mathbf{w}}{\partial V_N} \end{bmatrix} \tag{10.41}$$

After obtaining transfer matrix, the equal current cource can be deducted from (10.36).

During the above deduction, GSS has another benefit. The GSS always use the previous step's solution for the next step's iteration initial value. But after obtaining $\frac{\partial \mathbf{w}}{\partial V}$, the next iteration initial value can be used with better approximation:

$$\mathbf{w}^{n+1} = \mathbf{w}\left(V^n\right) + \left(\frac{\partial \mathbf{w}}{\partial V}\right)^n \Delta V \tag{10.42}$$

This will help for convergence.

10.3 Device IV curve's automatic scan

Calculating device's IV curve normally adopts voltage scan's method: gradually increasing electrode voltage, for every voltage value calculating the current through the device, and then draw the IV curve. This method is easy to be considered, but in real application there are problems. Mentioned in "??", on page ??, at diode's forward IV curve diagram (??)'s high partial voltage part, current and voltage is exponential related. A small increase in voltage leads to big current variation, which means that it is difficult to converge by using the previous voltage's solution. "??", on page ?? snapback phenomenon shows multiple current value function at the same voltage. Voltage scan will not converge.

In order to solve the two problems above, we can scan the current. For diode's forward conduction, increase big current only means a small variation of voltage. Then Newton iteration's initial value is much better. For GGMOS's snapback point, current is still single value function, which removes a lot of of multiple solution problems.

In conclusion, when IV curve is smooth (parallel to voltage axis), it is suitable to use voltage scan; for IV curve is sharp (parallel to current line), it is suitable for current scan. Device IV curve automatic scan function can shift from this two, which avoids human interference [71].

Device IV curve automatic scan adopts circuit structure shown in (10.6). One voltage source is linked to device port which needs to be scanned through a tunable resistor. From circuit analysis we know, one voltage source with inner resister R is a line with slope $K = 1/R$, called loading line. When $R \to 0$, slope $K \to \infty$ represents loading line is perpendicular to voltage axis, similar to ideal voltage source situation, shown in (10.7)(a) line; When $R \to \infty$, slope $K \to 0$ represents loading line is perpendicular to current axis, similar as ideal current source, shown in figure (10.7) (b) line. Accordingly by tuning R [2] we can make loading line be perpendicular to device IV curve, shown in figure (10.7) (c) line.

Device IV curve automatic scan algorithm secure each scan point, loading line and IV line perpendicular state. The algorithm is shown in figure (10.8). At P pooint,

[2] allow negative resistance

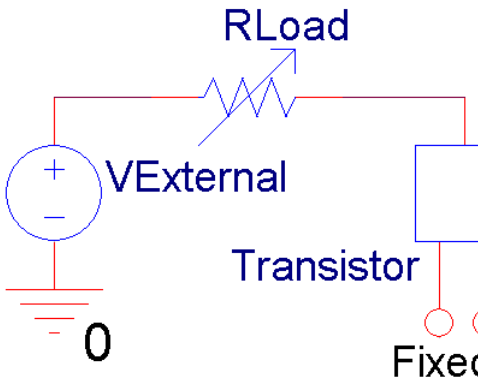


Figure 10.6: Device IV curve automatic scan circuit structure

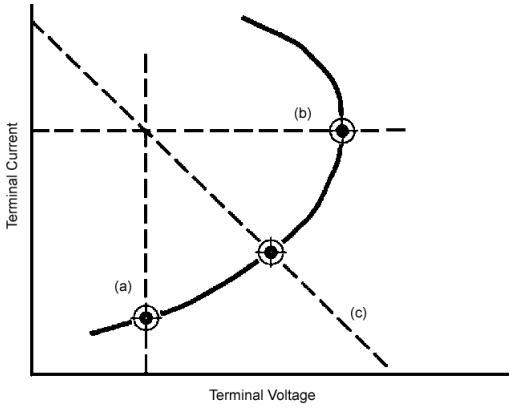


Figure 10.7: Device's IV curve and loading line

GSS calculates device current node's $I = I(V)$ and IV curve's slope $K = \partial I/\partial V$. Assume loading line's resistance $RLoad = -1/K$, perpendicular to current IV line. Then changing outer source voltage V_{ext}, similar as moving loading line through parallel to voltage axis δV_{ext}. Assuming loading line and IV curve meet at Q point, GSS calculates Q point's position, simultaneously calculate Q point's IV curve's slope. Then evaluate the loading line resistance again, like rotating an angle at the Q point to make it perpendicular to the IV curve. Now the loading line and voltage axis's meeting point turns to be V_{ext}^{new}, which means after loading line resistance change, outer voltage has to be set as V_{ext}^{new} to let the loading line and IV curve meet at Q point. Do the loop like the description above until we finish the scanning of whole IV curve.

It is worthy to mention that at the snapback point voltage value's step needs to change sign. Figure (10.9) shows how to fix snapback point: after snapback point, IV curve's slope S1 and S2 have different sign. The product is negative; The line T, connecting 1 and 2 points, has slope value bigger than S1's slope absolute value. Satisfying the previous two condition, we can judge the snapback point at the IV curve, we need to change the step's sign.

In the end I want to introduce the disadvantage of IV line automatic scan. Because it needs to introduce an additional resistor (The resistor is very big at high slope IV curve), electrode with big resistor leads to more problem condition numbers, accordingly IV curve automatic scan may lead to severe convergence problem.

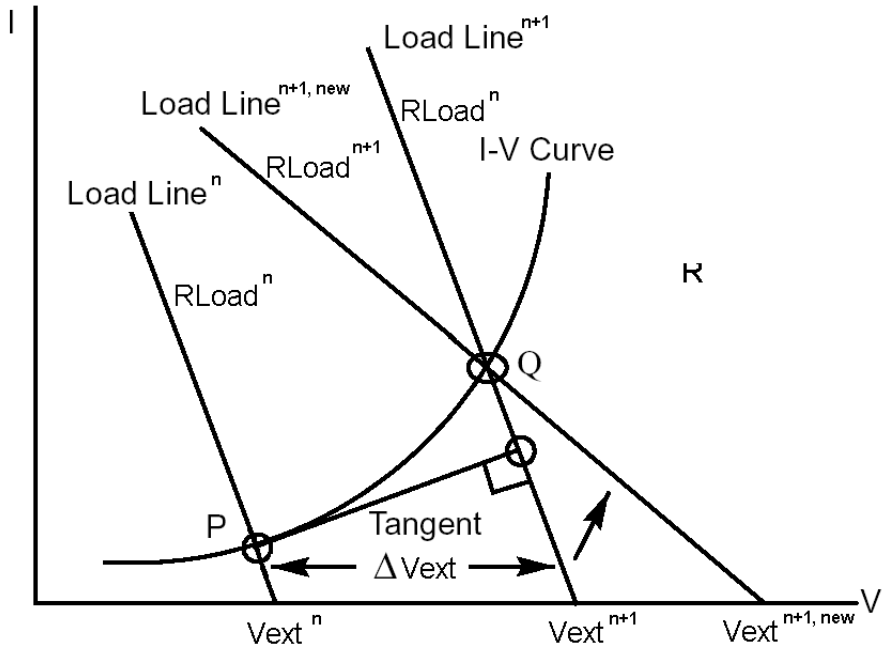


Figure 10.8: Device IV curve automatic scan process

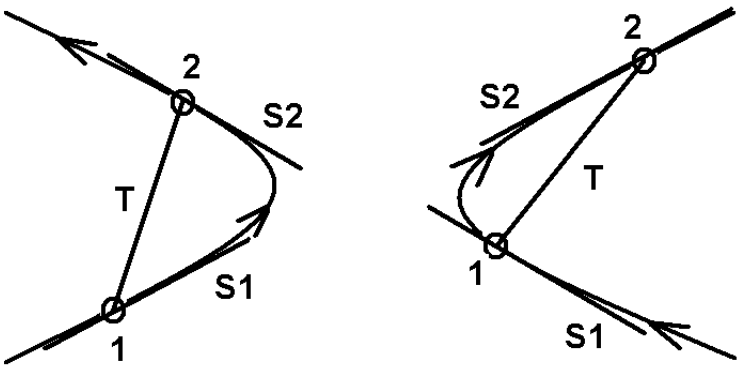


Figure 10.9: Snapback point's judgement

Bibliography

[1] Jonathan Richard Shewchuk. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In Applied Computational Geometry: Towards Geometric Engineering, volume 1148, pages 203–222. Springer-Verlag, May 1996. 8, 90

[2] J. J. Liou. Modeling the Tunneling Current in Reverse-Biased p/n Junctions. Solid-State Electronics, volume 33, pages 971–972, 1990. 40, 85

[3] D. M. Caughey and R. E. Thomas. Carrier Mobilities in Silicon Empirically Related to Doping and Field. Proc. IEEE, volume 55, pages 2192–2193, 1967. 50, 77, 78

[4] D. B. M. Klaassen. A Unified Mobility Model for Device Simulation - I. Solid-State Electronics, volume 35, pages 953–959, 1992. 51, 78

[5] D. B. M. Klaassen. A Unified Mobility Model for Device Simulation - II. Solid-State Electronics, volume 35, pages 961–967, 1992. 51, 78

[6] J. Hölzl and F. K. Schulte. Work Functions of Metals. Springer-Verlag, 1979. 56

[7] S. M. Sze. Physics of Semiconductor Devices, 2nd edition. New York: John Wiley & Sons, 1981. 58, 74, 83

[8] J. M. Andrews and M. P. Lepselter. Reverse Current-Voltage Characteristics of Metal-Silicide Schottky Diodes. Solid-State Electronics, volume 13, pages 1011–1023, 1970. 61, 128

[9] U. Lindefelt. Current-Density Relations for Nonisothermal Modeling of Degenerate Heterostructure Devices. J. Appl. Phys., volume 75, number 2, pages 958–966, Jan. 1994. 70

[10] S. Selberherr. Analysis and Simulation of Semiconductor Devices. Springer-Verlag, 1984. 70, 83

[11] Yu Zhiping, D. Chen, L. So, and R. W. Dutton. PISCES-2ET and Its Application Subsystems. Technical report, Integrated Circuits Laboratory, Stanford University, 1994. 71, 83

[12] Andreas Aste and R. Vahldieck. Time-domain simulation of the full hydrodynamic model. Int. J. Numer. Model., volume 16, pages 161–174, 2003. 71

[13] M. G. Ancona and H. F. Tiersten. Macroscopic physics of the silicon inversion layer. Phys. Rev. B, volume 35, number 15, pages 7959–7965, May, 1987. 73

[14] M. G. Ancona and G. J. Iafrate. Quantum correction to the equation of state of an electron gas in a semiconductor. Phys. Rev. B, volume 39, number 13, pages 9536–9540, May, 1989. 73

[15] C.S. Rafferty, B. Biegel, Z. Yu, M.G. Ancona, J. Bude, and R.W. Dutton. Multi-dimensional quantum effect simulation using a density-gradient model and script-level programming techniques. In Proc. SISPAD, pages 137–140, 1998. 73

[16] E. Lyumkis, R. Mickevicius, O. Penzin, B. Polsky, K. El Sayed, A. Wettstein, and W. Fichtner. Density Gradient Transport Model for the Simulations of Ultrathin, Ultrashort SOI under Non-Equilibrium Conditions. In IEEE International SOI Conference, 2002. 73

[17] A. Wettstein, O. Penzin, and E. Lyumkis. Integration of the Density Gradient Model into a General Purpose Device Simulator. VLSI Design, volume 15, number 4, pages 751–759, 2002. 73, 121

[18] Ren-Chuen Chen and Jinn-Liang Liu. A quantum corrected energy-transport model for nanoscale semiconductor devices. J. Comput Phys., volume 204, pages 131–156, 2005. 73

[19] J. W. Slotboom. The pn Product in Silicon. Solid-State Electronics, volume 20, pages 279–283, 1977. 75

[20] D. J. Roulston, N. D. Arora, and S. G. Chamberlain. Modeling and Measurement of Minority-Carrier Lifetime Versus Doping in Diffused Layers of $n^+ - p$ Silicon Diodes. IEEE Trans. Electron Devices, volume ED-29, pages 284–291, 1982. 76

[21] S. Selberherr. Process and Device Modeling for VLSI. Microelectron. Reliab., volume 24, number 2, pages 225–257, 1984. 77

[22] J. J. Barnes, R. J. Lomax, and G. I. Haddad. Finite-Element Simulation of GaAs MESFET's with Lateral Doping Profiles and Sub-Micron Gates. IEEE Trans. Electron Devices, volume ED-23, pages 1042–1048, 1976. 78

[23] H.R. Yeager and R.W. Dutton. Circuit-simulation models for the high electron-mobility transistor. IEEE Trans. Electron Devices, volume 33, pages 682–692, 1986. 78

[24] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi. A Physically Based Mobility Model for Numerical Simulation of Nonplanar Devices. IEEE Trans. Computer-Aided Design, volume 7, number 11, pages 1164–1170, 1988. 80

[25] M. N. Darwish, J. L. Lentz, M. R. Pinto, P. M. Zeitzoff, T. J. Krutsick, and H. H. Vuong. An Improved Electron and Hole Mobility Model for General Purpose Device Simulation. IEEE Trans. Electron Devices, volume 44, number 9, pages 1529–1538, 1997. 81

[26] K.M. Cham, S-Y Oh, D. Chin, and J.L. Moll. Computer-Aided Design and VLSI Device Development. Kluwer Academic Publishers, 1986. 81

[27] Avant! Corporation. Medici User's Manual, June, 2001. 83

[28] M. Valdinoci, D. Ventura, M.C. Vecchi, M. Rudan, G. Baccarani, F. Illien, A. Stricker, and L. Zullino. Impact-ionization in silicon at large operating temperature. In International Conference on Simulation of Semiconductor Processes and Devices, Sept. 1999. 84

[29] W. B. Joyce and R. W. Dixon. Analytic Approximation for the Fermi Energy of an Ideal Fermi Gas. Appl. Phys. Lett., volume 31, pages 354–356, 1977. 85

[30] Yu Zhiping and R. W. Dutton. SEDAN III - A Generalized Electronic Material Device Analysis Program. Technical report, Stanford Electronics Laboratory Technical Report, July 1985. 85

[31] L. Geppert. semiconductors. 1999 technology analysis and forecast. IEEE Spectrum, volume 36, pages 52–56, 1999. 88

[32] L. Paul Chew. Guaranteed-quality mesh generation for curved surfaces. In SCG '93: Proceedings of the ninth annual symposium on Computational geometry, pages 274–280, New York, NY, USA, 1993. ACM Press. 90

[33] Peter Su and Robert L. Scot Drysdal. A Comparison of Sequential Delaunay Triangulation Algorithms. In Proceedings of the Eleventh Annual Symposium on Computational Geometry, pages 61–70. ACM Press, 1995. 90

[34] Dan X. Yang. Mesh generation and information model for device simulation. PhD thesis, Stanford University, 1994. 90

[35] Zakir Hussain Sahul. Grid and geometry servers for semiconductor process simulation. PhD thesis, Stanford University, 1996. 90

[36] C. W. Hirt. Heunstic stability theory for finite difference equations. J. Compute Phys., volume 2, pages 339–355, 1968. 108

[37] K.S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propagat, volume 14, pages 302–307, 1966. 108

[38] A. Harten. High resolution schemes for hyperbolic conservation laws. J. Comput Phys., volume 49, pages 357–393, 1983. 108

[39] A. Harten, P. D. Lax, and Van Leer. On upstream difference and Godunov-type schemes for hyperbolic conservation laws. SIAM Review, volume 25, pages 35–61, 1983. 108

[40] Ami Harten, Stanley Osher, Björn Engquist, and Sukumar R. Chakravarthy. Some results on uniformly high-order accurate essentially nonoscillatory schemes. Applied Numerical Mathematics, volume 2, number 3–5, pages 347–377, October 1986. 109

[41] Ami Harten and Stanley Osher. Uniformly high-order accurate nonoscillatory schemes. SIAM Journal on Numerical Analysis, volume 24, pages 279–309, April 1987. 109

[42] C. Manry, S. Broschat, and J. Schneider. Higher-order FDTD methods for large problems. J. Appl. Comput. Electromag. Soc., volume 10, number 2, pages 302–307, 1995. 109

[43] Steven E. Laux and Robert G. Byrnes. Semiconductor device simulation using generalized mobility models. IBM J. Res. Dev., volume 29, number 3, pages 289–301, May 1985. 114, 116, 123

[44] Akira Kato, Mitsutaka Katada, Toyoharu Kamiya, Toyoki Ito, and Tadashi Hattori. A Rapid, Stable Decoupled Algorithm for Solving Semiconductor Hydrodynamic Equations. IEEE Trans. Computer-Aided Design, volume 13, number 11, pages 1425–1428, 1994. 119

[45] A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, and M. Rudan. A New Discretization Strategy of the Semiconductor Equations Comprising Momentum and Energy Balance. IEEE Trans. Computer-Aided Design, volume CAD-7, pages 231–242, 1988. 120

[46] T. W. Tang. Extension of the Scharfetter-Gummel algorithm to the energy balance equation. IEEE Trans. Electron Devices, volume ED-31, pages 1912–1914, 1984. 120

[47] Woo-Sung Choi, Jae-Gyung Ahn, and Young-June Park. A Time Dependent Hydrodynamic Device Simulator SNU-2D With New Discretization Scheme and Algorithm. IEEE Trans. Computer-Aided Design, volume 13, number 7, pages 899–908, 1994. 120

[48] A. Wettstein. Quantum Effects in MOS Devices. PhD thesis, Hartung-Gorre, Karlsruhe, 2000. 121

[49] A. Wettstein, A. Schenk, and W. Fichtner. Quantum Device-Simulation with the Density Gradient Model on Unstructured Grids. IEEE Trans. Electron Devices, volume 48, number 2, pages 279–284, Feb. 2001. 121

[50] SILVACO. ATLAS Userąs Manual, March 2, 2007. 122

[51] Steven E. Laux and Bertrand M. Grossman. A General Control-Volumn Formulation for Modeling Impact Ionization in Semiconductor Transport. IEEE Trans. Computer-Aided Design, volume CAD-4, number 4, pages 520–526, Octoble 1985. 123

[52] C. R. Crowell and S. M. Sze. Current Transport in Metal-Semiconductor Barriers. Solid-State Electronics, volume 9, pages 1035–1048, 1966. 128

[53] K. Hess and G. J. Iafrate. Modern Aspects of Heterojunction Transport Theory. Elsevier Science Publications, 1987. 131

[54] Seonghoon Jin, Young June Park, and Hong Shick Min. A numerically efficient method for the hydrodynamic density-gradient model. In Simulation of Semiconductor Processes and Devices, 2003. 132

[55] A. De Mari. An accurate numerical one-dimensional solution of the p-n junction under arbitrary transient conditions. Solid-State Electronics, volume 11, pages 1021–1053, 1968. 133

[56] M. R. Pinto, C. S. Rafferty, H. R. Yeager, and R. W. Dutton. PISCES-II - supplementary report. Technical report, Integrated Circuits Laboratory, Stanford University, 1985. 135

[57] Kartikeya Mayaram. CODECS: A Mixed-Level Circuit and Device Simulator. Memorandum No. UCB/ERL M88/71, Berkeley: University of California, 1988. 136, 156

[58] H. K. Gummel. A self-consistent iterative scheme for one-dimensional steady state transistor calculations. IEEE Trans. Electron Devices, volume 11, pages 455–465, 1964. 137

[59] J. E. Dennis Jr. and Robert B. Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, 1983. 138, 139

[60] D. G. Luenberger. Linear and Nonlinear Programming. Addison-Wesley, 2nd edition edition, 1984. 138

[61] Peter N. Brown and Youcef Saad. Hybrid Krylov methods for nonlinear systems of equations. SIAM J. Sci. Stat. Comput., volume 11, pages 450ÍC–481, 1990. 141

[62] Argonne National Laboratory. PETSC User's Manual, 2006. 142

[63] S. E. Laux. Techniques for Small-Signal Analysis of Semiconductor Devices. IEEE Trans. Electron Devices, volume ED-32, pages 2028–2037, Oct. 1985. 151

[64] Nenzi Paolo. NGSPICE User Manual. http://ngspice.sourceforge.net, 2005. 151

[65] William J. McCalla. Fundamentals of Computer-Aided Circuit Simulation. Boston: Kluwer Academic Publishers, 1993. 152

[66] Mayaram Kartikeya and Donald O. Pederson. Coupling Algorithms for Mixed-Level Circuit and Device Simulation. IEEE Trans. Computer-Aided Design, volume II, number 8, pages 1003–1012, 1992. 155

[67] Yu Zhiping, Robert W. Dutton, and Hui Wang, editors. A Modularized, Mixed IC Device/ Circuit Simulation System, Japan, April 1992. Proceedings of the Synthesis and Simulation Meeting and International Exchange. 155

[68] Thomas L. Quarles. Adding Devices to SPICE3. Memorandum No. UCB/ERL M89/45. Berkeley: University of California, 1989. 155

[69] Thomas L. Quarles. SPICE3 Implementation Guide. Memorandum No. UCB/ERL M89/44, Berkeley: University of California, 1989. 155

[70] David Alan Gates. Design-Oriented Mixed-level Circuit and Device Simulation. Memorandum No. UCBERL M93/51, Berkeley: University of California, 1993. 156

[71] R.J.G. Goossens, S.G. Beebe, Yu Zhiping, and R.W. Dutton. An Automatic Biasing Scheme for Tracing Arbitrarily Shaped I-V Curves. IEEE Trans. on Computer-Aided Design, volume 41, pages 310–317, March 1994. 157