Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

А. Е. Почтенный

ФИЗИКА

В 6-ти частях

Часть 6

Квантовые электронные свойства твердых тел

**Тексты лекций для студентов
химико-технологических специальностей**

Минск 2014

УДК 537.1(075.8)
ББК 22.31я73
П65

Рассмотрены и рекомендованы редакционно-издательским советом Белорусского государственного технологического университета

Рецензенты:
заведующий кафедрой физики полупроводников БГУ
кандидат физико-математических наук *В. Ф. Стельмах*;
заведующий кафедрой информационно-измерительной техники БГПА
доктор физико-математических наук *В. Б. Яржембицкий*

Почтенный, А. Е.
П65 Физика. В 6 ч. Ч. 6. Квантовые электронные свойства твердых тел: тексты лекций для студентов химико-технологических специальностей / А. Е. Почтенный. – Минск: БГТУ, 2015. – 93 с.

Пособие предназначено для студентов химико-технологических специальностей. В издании рассмотрены вопросы динамики и статистики электронов в твердых телах, в первую очередь в полупроводниках; принципы функционирования ряда полупроводниковых приборов – диодов, биполярных и полевых транзисторов, приборов с зарядовой связью, лазеров; технология полупроводников и полупроводниковых приборов.

УДК 537.1(075.8)
ББК 22.31я73

− Но это противоречит здравому смыслу! − возразил Филби.
− А что такое здравый смысл? − спросил Путешественник во Времени.

Герберт Уэллс, «Машина времени»

ВВЕДЕНИЕ

Профессия инженера имеет очень солидный возраст. Египетские пирамиды вызывают уважение к инженерам-строителям, работавшим тысячи лет назад. С тех пор строительная наука, возможно, усовершенствовалась, но, в общем, она понятна с точки зрения бытового здравого смысла − каменный дом долговечнее деревянного, стальная конструкция прочнее деревянной и т. д.

Однако ситуация резко меняется, когда мы вторгаемся в области знаний, связанные с микроскопическими свойствами вещества, например, в полупроводниковую электронику. Здесь повседневный здравый смысл начинает нас подводить, и совершенно понятно почему – потому, что электроны в твердом теле не подчиняются классическим законам. Они подчиняются квантовой механике и квантовой статистике, и для анализа работы полупроводниковых приборов мы должны выработать новый здравый смысл, основанный на фундаментальных представлениях этих разделов физики.

Мы живем на бытовом и техническом уровне в мире твердотельной электроники – мобильные телефоны, компьютеры, связь через Интернет, телевизоры и радиоприемники, принтеры и копировальные аппараты работают с использованием полупроводниковых приборов. Поскольку вы уже знакомы с основами и некоторыми положениями квантовой механики и квантовой статистики, мы можем разобраться с физической базой и элементарными принципами работы подобной техники. Для этого предстоит сделать некое усилие – усвоить специфический язык этой области знаний, который называется «зонной моделью». Тогда, допустим, при обсуждении работы микросхем КМОП – логики в ваших электронных часах вы сможете сказать: «Я понимаю, о чем идет речь».

1. КЛАССИЧЕСКАЯ ЭЛЕКТРОННАЯ ТЕОРИЯ

Первая электронная теория проводимости была предложена П. Друде в 1900 г. − всего через три года после открытия Дж. Дж. Томсоном электрона. Несмотря на то что эта теория или, точнее, модель оказалась принципиально неверной, она представляет не только исторический интерес по двум причинам: во-первых, в ней вводится ряд понятий и соотношений, которые либо в неизменном, либо в подкорректированном виде входят и в современные теории проводимости; во-вторых, модель Друде позволяет быстро построить наглядную картину явления и грубо оценить характеристики, точное определение которых требует сложного анализа.

В модели Друде предполагается, что электрический ток в веществе обусловлен движением электронов, оторвавшихся от атомов и ставших свободными, а наличие электрического сопротивления вызвано столкновениями электронов с положительными ионами, оставшимися в узлах кристаллической решетки.

Если приложенная к образцу разность потенциалов создает внутри образца однородное электрическое поле с напряженностью ξ, то свободный электрон в таком поле приобретет постоянное ускорение:

$$a = \frac{F}{m} = \frac{e\vec{\xi}}{m},$$

где F − суммарная сила, действующая на электрон; m − масса электрона; e − заряд электрона. В результате такого равноускоренного движения электрон за время t приобретет *дрейфовую скорость* υ_D, равную

$$\upsilon_D = at = \left(\frac{e\vec{\xi}}{m}\right)t.$$

Эта скорость будет увеличиваться до тех пор, пока электрон не потеряет ее в результате столкновения с ионом, после чего электрическое поле снова начнет разгонять электрон.

В этом случае для вычисления средней скорости дрейфа достаточно рассмотреть движение электрона в течение промежутка времени τ между двумя столкновениями. Такой промежуток времени называется *временем релаксации*. Тогда дрейфовую скорость можно оценить как

$$\upsilon_D = a\tau = \left(\frac{e\vec{\xi}}{m}\right)\tau = \mu\vec{\xi}, \tag{1.1}$$

где величину

$$\mu = \frac{\upsilon_D}{\vec{\xi}} \tag{1.2}$$

называют *дрейфовой подвижностью*, или просто *подвижностью,* и в модели Друде она равна

$$\mu = \frac{e\tau}{m}. \tag{1.3}$$

Соотношение (1.2), в отличие от (1.3), является определением подвижности и не связано с конкретной моделью проводимости.

Рассчитаем теперь силу тока в образце, которая по определению равна заряду q, проходящему в единицу времени через поперечное сечение образца. Допустим, образец имеет форму цилиндра с площадью поперечного сечения S.

За время t через это сечение пройдет N электронов, содержащихся в отрезке этого цилиндра длиной $l = \upsilon_D \cdot t$. Если ввести понятие *концентрации* свободных электронов n как их количества в единице объема V:

$$n = \frac{N}{V}, \tag{1.4}$$

то силу тока можно вычислить как

$$I = \frac{dq}{dt} = \frac{d(eN)}{dt} = \frac{d(enV)}{dt} = \frac{d(enSl)}{dt} = \frac{d(enSv_D t)}{dt} = enS\upsilon_D,$$

так как величины e, n, S и υ_D являются константами. Плотность тока будет равна

$$j = \frac{I}{S} = en\upsilon_D, \tag{1.5}$$

и в рамках модели Друде

$$j = \left(\frac{e^2\tau n}{m}\right)\vec{\xi}. \tag{1.6}$$

Величину

$$\sigma = \frac{e^2 \tau n}{m} \tag{1.7}$$

называют *удельной проводимостью*. Тогда формула (1.6), представленная в виде

$$j = \sigma \vec{\xi}, \tag{1.8}$$

является обычным законом Ома, если только σ, а значит, в соответствии с (1.7) и время релаксации τ не зависят от поля. Такое предположение выглядит неправдоподобным. Скорее от поля не зависит расстояние Λ, проходимое электроном между двумя столкновениями и называемое *длиной свободного пробега*, которое связано с τ соотношением

$$\tau = \frac{\Lambda}{\left(\upsilon_D + \upsilon_T\right)},$$

где υ_T - тепловая скорость движения электронов. С другой стороны, мы знаем, что закон Ома часто выполняется. Последняя формула показывает, что это возможно, если дрейфовая скорость электронов пренебрежимо мала по сравнению с тепловой.

Попробуем сделать численные оценки для металлов, типичное значение подвижности электронов в которых составляет $5 \cdot 10^{-3}$ м^2/В·с. Тогда при напряженности поля 1 В/м

$$\upsilon_D = \mu \vec{\xi} = 5 \cdot 10^{-3} \text{ м/с}.$$

Тепловую скорость электронов оценим из соотношения

$$\frac{1}{2} m \left(\upsilon_T\right)^2 = \frac{3}{2} kT,$$

где k − постоянная Больцмана; T − абсолютная температура. Тогда при комнатной температуре тепловая скорость

$$\upsilon_T = \left(\frac{3kT}{m}\right)^{1/2} = \left(\frac{3 \cdot 1{,}38 \cdot 10^{-23} \cdot 300}{9{,}1 \cdot 10^{-31}}\right)^{1/2} \cong 10^5 \text{ м/с},$$

то есть действительно намного больше (в сто миллионов раз) дрейфовой. Приведенный расчет показывает, что в металлах закон Ома должен выполняться с точностью до одной миллионной доли процента.

Столь малое изменение скорости электрона во внешнем поле по сравнению с его равновесной тепловой скоростью позволяет использовать при описании электропроводности равновесные статистические функции распределения.

Интересно, а к какой величине тока приводят столь малые по сравнению с тепловой дрейфовые скорости электронов? Если на каждый атом металла приходится один свободный электрон, то объемная концентрация электронов

$$n = \frac{\rho N_{\text{A}}}{M},$$

где ρ − плотность; N_{A} − число Авогадро; M − молярная масса. Приблизительные оценки дают концентрацию свободных электронов в металлах около 10^{29} м$^{-3}$, что при $\vec{\xi}$ = 1 В/м в соответствии с (1.2) и (1.5) приводит к значению плотности тока порядка 10^8 А/м^2.

Используя (1.3), мы можем привести формулу (1.7) к виду

$$\sigma = en\mu, \tag{1.9}$$

который является наиболее употребительным выражением для удельной проводимости, справедливым не только в рамках модели Друде.

Сформулируем теперь те предположения, которые лежат в основе модели Друде:

1) электроны проводимости считаются свободными, то есть не взаимодействуют с ионами в узлах кристаллической решетки, а только упруго сталкиваются с ними и не взаимодействуют друг с другом;

2) соударения электронов с ионами беспорядочные, следовательно, после каждого соударения дрейфовая скорость обращается в ноль («потеря памяти»);

3) все электроны движутся с одной и той же тепловой скоростью, равной среднеквадратичной скорости в распределении Максвелла − Больцмана.

Пока еще наших знаний недостаточно, чтобы критически оценить справедливость этих предположений. Правда, по поводу третьего предположения уже сейчас мы можем сказать, что оно требует улучшения, а именно учета распределения электронов по скоростям, что и было в свое время сделано Г. Лоренцем, но не привело к существенному изменению модели. Однако, не оценивая сами предположения, мы можем рассмотреть выводы из модели Друде и прийти к заключению о ее справедливости.

Один из таких выводов − расчет электронной теплоемкости. Согласно закону Дюлонга и Пти, молярная теплоемкость C твердых тел, обусловленная колебаниями атомов, равна $3R$, где R = 8,31 Дж/моль·К − универсальная газовая постоянная. Такой будет теплоемкость диэлектриков, не имеющих свободных электронов.

Металлы же должны обладать дополнительной теплоемкостью, обусловленной наличием свободных электронов. Если на каждый атом металла приходится хотя бы один свободный электрон, то в одном моле металла содержится как минимум $N = 6\cdot10^{23}$ моль$^{-1}$ таких электронов, обладающих суммарной энергией $E = (3/2)kTN_{\text{А}}$, что дает добавку $C_{\text{эл}}$ к молярной теплоемкости, равную

$$C_{\text{эл}} = \frac{dE}{dT} = \frac{3}{2}kN_{\text{А}} = \frac{3}{2}R,$$

то есть молярная теплоемкость металла

$$C_{\text{эл}} = 3R + \frac{3}{2}R = \frac{9}{2}R = 4{,}5R,$$

что в полтора раза больше молярной теплоемкости диэлектриков. Приведенные в таблице данные свидетельствуют, что ничего подобного на опыте не наблюдается.

Металл	$C_{\text{м}}$, Дж/(моль К)	Диэлектрик	$C_{\text{м}}$, Дж/(моль К)
Алюминий	24,35	Карбид кремния	26,65
Медь	24,52	Сера	23,64
Золото	25,23	Германий	23,4

Это не единственный пример экспериментально наблюдаемых противоречий модели Друде. Основными из них являются:

1) отсутствие предсказанной электронной теплоемкости;

2) отсутствие магнитной восприимчивости, обусловленной свободными электронами;

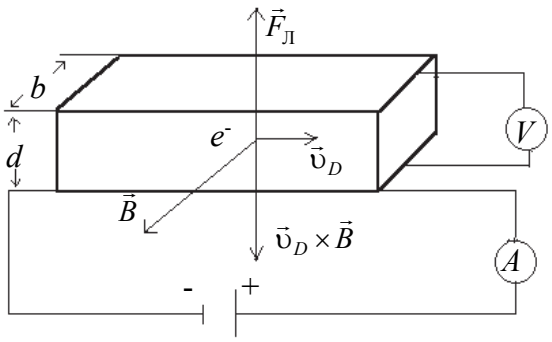
3) экспериментальное обнаружение положительного знака заряда носителей тока в ряде металлов и полупроводников.

Как мы увидим в дальнейшем, основные причины недостатков модели Друде в том, что она не учитывает квантовый характер как динамики, так и статистики электронов.

Третье из отмеченных выше противоречий вызывает вопрос о том, как можно измерить знак заряда носителей. Такие измерения основаны на *эффекте Холла*, с которым необходимо ознакомиться. Пусть образец исследуемого материала, по которому течет электриче-

ский ток, помещен в магнитное поле с индукцией B, перпендикулярное направлению движения носителей (рисунок).

Если знак заряда носителей отрицательный, то они будут двигаться направо, и действующая на эти носители со стороны магнитного поля сила Лоренца будет отклонять их к верхней грани образца, на которой начнет накапливаться избыточный отрицательный заряд. Между верхней и нижней гранями образца возникнет электрическое поле с напряженностью ξ_H и разность потенциалов U_H. Такая разность потенциалов называется холловской, а само явление ее возникновения − эффектом Холла.



Механизм возникновения эффекта Холла

Электроны проводимости отклоняются магнитным полем к верхней грани образца, на которой вследствие этого накапливается отрицательный заряд, и между верхней и нижней гранью образца возникает разность потенциалов. А что изменится, если знак заряда носителей будет не отрицательным, а положительным?

Величина холловской разности потенциалов $U_H = \xi_H d$ будет расти до тех пор, пока кулоновская сила $F_К = e\xi_H$, действующая на носители со стороны холловского поля ξ_H, не сравняется с силой Лоренца $F_Л = e\upsilon_D B$, после чего наступит равновесие. Условие равновесия можно записать в виде

$$e\xi_H = \frac{eU_H}{d} = e\upsilon_D B,$$

откуда

$$U_H = \upsilon_D B d.$$

Если выразить дрейфовую скорость из (1.5) через силу тока I, протекающего по образцу, и учесть, что $S = bd$, то окончательно для холловской разности потенциалов получим:

$$U_H = \frac{IB}{enb} = \frac{R_H IB}{b}, \qquad (1.10)$$

где величина

$$R_H = \frac{1}{en} \tag{1.11}$$

называется *коэффициентом Холла* и зависит только от заряда и концентрации носителей.

Если знак заряда носителей положительный, то на верхней грани образца будет накапливаться положительный заряд, то есть знак холловской разности потенциалов изменится, а все приведенные выше вычисления останутся теми же самыми.

Таким образом, измерив холловскую разность потенциалов и протекающий через образец ток, а также зная величину магнитного поля и ширину образца, мы можем определить как знак заряда, так и концентрацию носителей.

Отметим, что мы рассмотрели самый простой случай – эффект Холла в материале с одним типом носителей.

2. ЭЛЕКТРОННАЯ ЭНЕРГЕТИЧЕСКАЯ СТРУКТУРА ТВЕРДЫХ ТЕЛ

2.1. Квантовое описание электронов в твердых телах

Как уже отмечалось в предыдущем разделе, корректное описание динамики электрона в твердом теле должно быть квантовым. Из всех возможных формулировок квантовой механики мы будем использовать только одну, основанную на уравнении Шрёдингера:

$$-\left(\frac{\hbar^2}{2m}\right)\Delta\Psi + U\Psi = \frac{(i\hbar)\partial\Psi}{\partial t}, \tag{2.1}$$

где $\hbar = 1{,}05\cdot10^{-34}$ Дж·с − постоянная Планка; m − масса частицы; U − потенциальная энергия частицы; i – «мнимая единица» ($i^2 = -1$); Ψ − волновая функция частицы; Δ − оператор Лапласа, в декартовых координатах, равный $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$.

Если полная энергия частицы E со временем не меняется, то *временное* уравнение Шрёдингера (2.1) сводится к *стационарному* уравнению Шрёдингера:

$$-\left(\frac{\hbar^2}{2m}\right)\Delta\Psi + U\Psi = E\Psi\,. \tag{2.2}$$

Волновая функция Ψ частицы может зависеть от времени и координат частицы и должна быть однозначной, непрерывной и конечной, а также иметь непрерывную и конечную производную. Произведение волновой функции Ψ на комплексно ей сопряженную Ψ^* представляет собой плотность вероятности обнаружения частицы около заданной точки.

Обратите внимание, что уравнение Шрёдингера содержит две неизвестные величины − волновую функцию и полную энергию частицы, значения которых, удовлетворяющие уравнению, называют соответственно *собственными функциями* и *собственными значениями энергии*.

При рассмотрении одной частицы, в частности одного электрона, речь идет об *одночастичном*, или, соответственно, *одноэлектронном* уравнении Шрёдингера. Твердое тело состоит из большого числа частиц, которые к тому же взаимодействуют между собой, поэтому для

его описания необходимо *многочастичное* уравнение Шрёдингера, найти точное решение которого нельзя никогда. Значит, нужно уметь находить такие приближения, которые лучше всего соответствовали бы описываемому явлению.

Поскольку химическая связь обусловлена только валентными электронами, то первое приближение, которое мы сделаем, − это разделим *валентные электроны* и *ионы решетки* на две независимые системы. Тогда в уравнении Шрёдингера для твердого тела

$$-\left(\frac{\hbar^2}{2m}\right)\Delta\varphi + U\varphi = E\varphi\,, \tag{2.3}$$

где $\varphi = \varphi(x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N)$ − функция $3N$ переменных, то есть координат всех N атомов твердого тела. Потенциальную энергию можно представить в виде суммы ионной $U_{\text{ии}}$ и электронной (имеются в виду валентные электроны) $U_{\text{ээ}}$ компонент, слагаемых, описывающих взаимодействие электронов с ионами $U_{\text{эи}}$, а также электронов и ионов с внешними полями $U_{\text{вн}}$:

$$U = U_{\text{ээ}} + U_{\text{ии}} + U_{\text{эи}} + U_{\text{вн}}. \tag{2.4}$$

Электронная компонента потенциальной энергии представляет собой сумму всех энергий кулоновского взаимодействия электронов между собой:

$$U_{\text{ээ}} = \frac{1}{2}\sum_{i\neq j}\frac{e^2}{\left(\vec{r}_i - \vec{r}_j\right)}, \tag{2.5}$$

где r − радиус-векторы, задающие положения электронов (подумайте сами, откуда взялся множитель 1/2). Ионная компонента − сумма энергий взаимодействия всех ионов с координатами R между собой. Каждую из таких энергий мы представим в виде функции расстояния между ионами, тогда

$$U_{\text{нн}} = \frac{1}{2}\sum_{k\neq l} U_{\text{нн}}\left(\vec{R}_k - \vec{R}_l\right). \tag{2.6}$$

Электрон-ионное взаимодействие соответственно

$$U_{\text{эн}} = \sum_{i,k} U_{\text{эн}}\left(\vec{r}_i - \vec{R}_k\right), \tag{2.7}$$

где множитель 1/2 отсутствует (почему?).

Поскольку ионы колеблются около положения равновесия в узлах кристаллической решетки, ион-ионное и ион-электронное взаимодей-

ствие можно разделить на часть U^0, описывающую взаимодействие с неподвижными ионами, и поправку к ней, учитывающую колебания:

$$U_{\text{ии}} = U^0_{\text{ии}} + U_{\text{иф}},\ U_{\text{зи}} = U^0_{\text{зи}} + U_{\text{эф}}, \qquad (2.8)$$

где происхождение индекса «ф» связано со словом *фонон* − квант колебаний решетки кристалла. Подобное разделение приносит две выгоды. Во-первых, при описании взаимодействия неподвижных ионов часто можно, не решая уравнение Шрёдингера, получить важную информацию, используя свойства *симметрии* кристаллической решетки. Во-вторых, оно позволяет ввести очень мощную концепцию *элементарных возбуждений*, или *квазичастиц*, которая в настоящее время является, пожалуй, основной в теории твердого тела.

Любое отклонение от равновесия в твердом теле может рассматриваться как возбуждение, энергия которого распределяется по всему телу. Появление этой энергии формально может быть описано как возникновение различных, в зависимости от природы возбуждения, квазичастиц. Если возбуждение невелико, этих частиц немного и взаимодействием между ними можно пренебречь, что существенно упрощает решение задач конкретного описания свойств твердого тела. Так, например, введение представления о фононах позволяет без труда описать теплоемкость кристаллов.

Однако и теперь еще подстановка выражений (2.4)−(2.8) в уравнение Шрёдингера (2.3) не дает возможности решить это уравнение. Дальнейшие упрощения связаны с тем, какое физическое явление мы хотим описать. Во многих случаях эти упрощения направляют на то, чтобы считать независимыми, во-первых, волновые функции электронной и ионной подсистемы, а во-вторых, волновые функции отдельных электронов. В этих случаях волновая функция системы частиц может быть представлена как произведение волновых функций, составляющих эту систему элементов (совсем, как в теории вероятностей − вероятность одновременного осуществления нескольких независимых событий равна произведению вероятностей осуществления каждого из этих событий в отдельности).

«Расцепление» электронной и ионной систем достигается либо заменой взаимодействия электронов с ионами взаимодействием электронов с равномерно распределенным по объему кристалла средним зарядом ионов (*континуальная модель*), либо фиксацией положения ионов (*адиабатическое приближение*), что позволяет представить волновую функцию кристалла как

$$\varphi = \varphi_{\text{э}} \cdot \varphi_{\text{и}},$$

где $\varphi_э$ и $\varphi_и$ – соответственно волновые функции электронной и ионной подсистем. Электронную волновую функцию можно упростить, заменив электрон-электронное взаимодействие взаимодействием каждого отдельного электрона с усредненным кулоновским потенциалом всех остальных электронов (*приближение Хартри – Фока*), что сводит многоэлектронную волновую функцию $\varphi_э$ к произведению *одноэлектронных* волновых функций, и тем самым вместо многоэлектронного уравнения Шрёдингера мы получаем набор одноэлектронных уравнений, решать которые гораздо проще. Общее представление о том, какие приближения дают описание различных электронных свойств твердого тела, можно составить на основании таблице.

Учитываемое взаимодействие электронов	Описываемые свойства
Без взаимодействия	Теплоемкость электронного газа; диа- и парамагнетизм свободных электронов; электроны проводимости в металлах и многих полупроводниках
Электрон в переменном внешнем поле	Диэлектрическая проницаемость электронного газа
Электрон в периодическом поле ионов решетки	Образование энергетических зон; различие металлов и диэлектриков
Электрон в периодическом поле ионов решетки и постоянном внешнем поле	Возможность электропереноса; пробой Зинера
Электрон-фононное	Электропроводность
Электрон-фотонное	Поглощение и отражение света, дисперсия, поляризация света
Электрон-фотонное во внешнем магнитном поле	Магнетооптика
Электрон-фонон-электронное	Сверхпроводимость

Аналогичную таблицу для ионной подсистемы можете составить сами в качестве упражнения. Не забудьте при этом продумать не только возможные взаимодействия, но и те реальные физические явления, которые описываются с учетом наличия (или, наоборот, отсутствия) тех или иных взаимодействий.

2.2. Электронные состояния

Несмотря на невообразимую сложность решения уравнения Шрёдингера для твердого тела, физики придумали достаточно простые

и эффективные методы описания физических свойств кристаллов. Среди этих методов почетное место занимает *зонная модель*, на которой и базируется, в частности, электронная теория полупроводников и полупроводниковых приборов.

Если мы рассмотрим изолированный атом, допустим, кремния, то электроны в нем, как мы уже знаем, могут обладать только вполне определенными значениями энергии, или, как говорят, могут находиться только на вполне определенных энергетических уровнях, одинаковых во всех атомах одного сорта. Что произойдет, если мы сблизим два таких атома? Взаимодействие между атомами расщепит каждый энергетический уровень на два, один из которых ляжет чуть ниже исходного, а другой − чуть выше. Если мы соберем вместе N атомов, то каждый энергетический уровень отдельного атома, смещаясь, образует систему N уровней, «зазор» между которыми будет тем больше, чем сильней взаимодействие, то есть чем ближе друг к другу атомы. Разрешенные энергии электронов в такой системе будут теперь образовывать не отдельные уровни, а *электронные энергетические зоны*, содержащие столько уровней, сколько атомов входит в систему. Набор разрешенных значений энергии валентных электронов, полученный из расщепления самых верхних заполненных в атоме уровней E_0, называют обычно *валентной зоной*, а набор самых нижних возбужденных состояний, полученный из нижнего пустого уровня E_1, − *зоной проводимости*. Качественно картина образования зон в твердом теле представлена на рис. 2.1 − *кривой Вильсона*, на горизонтальной оси которой отложено расстояние между соседними атомами (*период кристаллической решетки*) b. На рис. 2.1 показано образование только верхней заполненной (валентной) и нижней пустой зон, хотя зон будет столько, сколько имеется уровней энергии электронов в атоме вещества.

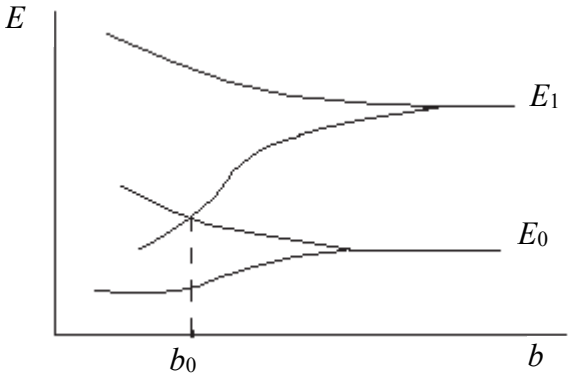


Рис. 2.1. Кривая Вильсона показывает расщепление энергетических уровней при сближении атомов

Представление об энергетических зонах позволяет классифицировать твердые тела по их электрическим свойствам. Допустим, на уровне E_0 в атоме находится один электрон (как, например, в щелочных металлах). Поскольку принцип Паули разрешает находиться на

одном уровне двум электронам с противоположной ориентацией спинов, N-валентных электронов в кристалле займут в соответствии с принципом минимума энергии нижние $N/2$ уровней в валентной зоне, а верхняя половина зоны останется пустой. Во внешнем электрическом поле электроны хотели бы двигаться, да некуда − все состояния заняты (позже мы рассмотрим этот вопрос корректно). Вот если бы электрону удалось выскочить в верхнюю половину зоны, тогда он без труда, используя пустые уровни, смог бы реализовать свою охоту к перемене мест. Нужна ли электрону для этого какая-то энергия? В принципе да, но... Ширина зоны составляет единицы электронвольт, а число уровней в зоне очень велико − допустим, в образце объемом 1 см^3 порядка 10^{23}. Значит, энергетический зазор между соседними уровнями в зоне будет порядка 10^{-23} эВ, а это много меньше тепловой энергии kT не только при комнатной температуре (около 25 мэВ), но даже при температуре жидкого гелия, и этой тепловой энергии вполне достаточно, чтобы электрон переходил внутри зоны с одного уровня на другой уровень. Следовательно, материал с частично пустой валентной зоной будет хорошим проводником тока − металлом.

А если на уровне E_0 в атоме находятся уже два электрона? Тогда валентная зона окажется полностью заполненной и электропереноса по ней не будет. Электрону, чтобы двигаться, придется преодолеть энергетический зазор между потолком валентной зоны и дном зоны проводимости, который называется *запрещенной зоной*, а вероятность такого перехода невелика. Такой материал был бы обязательно диэлектриком (изолятором), если бы не еще одно «но». Посмотрите внимательно на рис. 2.1 − если расстояние между атомами меньше, чем b_0, валентная зона и зона проводимости перекрываются и образуют единую частично заполненную зону и материал снова будет металлом (хорошая идея: взять изолятор, сильно сжать и превратить его в металл − это один из примеров того, что называется переходом Мотта). Значит, чтобы материал был *металлом*, необходимо, чтобы его валентная зона или изначально была заполнена частично, или перекрывалась с зоной проводимости − примеров металлов второго типа множество, скажем, бериллий, магний, цинк и т. д. Если же валентная зона заполнена полностью и отделена от зоны проводимости запрещенной зоной, то материал является *диэлектриком*.

Однако в нашу классификацию не укладываются такие материалы, как *полупроводники*. Дело в том, что между полупроводниками и диэлектриками нет качественного различия, а есть только количественное. Полупроводниками называют такие диэлектрики, у которых

ширина запрещенной зоны относительно невелика, так что количество попавших из валентной зоны в зону проводимости электронов (*свободных электронов*) достаточно, чтобы эти вещества худо-бедно проводили электрический ток (обратите внимание, что у металлов все валентные электроны являются свободными − другой вопрос, все ли они будут участвовать в электропроводности?). Границу между полупроводниками и диэлектриками не установишь. Например, у антимонида индия (InSb) ширина запрещенной зоны 0,18 эВ, а у фосфида галлия (GaP) − 2,25 эВ. Самые популярные в микроэлектронике полупроводники − кремний, германий и арсенид галлия (GaAs) − имеют ширину запрещенной зоны 1,15, 0,65 и 1,38 эВ соответственно. Так что в качестве условной границы между полупроводниками и диэлектриками можно взять максимальную ширину запрещенной зоны полупроводников около 2,5 эВ.

Впрочем, как мы увидим в дальнейшем, практическая ценность полупроводников определяется не столько величиной запрещенной зоны, сколько возможностью их *легировать*, то есть очень сильно менять электрические свойства, вводя очень малое количество примесей.

Для того чтобы получать полезные результаты, качественной картины мало − нам нужна количественная модель электронных энергетических зон, и мы в качестве таковой будем использовать очень удобную *модель Фейнмана*, обычно не встречающуюся (незаслуженно) в литературе по физике твердого тела.

Для построения этой модели необходимо слегка расширить наше знакомство с аппаратом квантовой механики. При этом минимальные усилия, как мы постепенно убедимся, принесут максимум пользы. Первое, что мы сделаем, − перепишем уравнение Шрёдингера (2.1) в операторной форме:

$$H\psi = i\hbar\left(\frac{\partial\psi}{\partial t}\right), \tag{2.9}$$

где $H = -\frac{\hbar^2}{2m}\Delta + U$ называется оператором Гамильтона, или гамильтонианом, и является оператором полной энергии частицы. В квантовой механике каждой физической величине сопоставляется оператор. Может, это и удивительно, но не удивительнее того, что в классической механике каждой физической величине соответствует число, просто к последнему факту мы привыкли и говорим, что масса футбольного мяча 450 г, не задумываясь о том, как это странно, что массу можно охарактеризовать числом.

В соответствии с принципом суперпозиции волновую функцию системы можно представить в виде линейной комбинации полного набора собственных функций:

$$\psi = \sum_j \psi_j \,, \tag{2.10}$$

где мы пренебрегли численными коэффициентами.

Если пространственная и временная части волновой функции независимы, то волновую функцию можно представить в виде

$$\Psi = \varphi(t)\varphi(r), \;\; \psi_j = \varphi_j(t)\varphi_j(r), \tag{2.11}$$

где φ не зависит от координат, а ϕ − от времени. Тогда с учетом (2.10) и (2.11) уравнение (2.9) преобразуется к виду

$$\sum_j \varphi_j H \varphi_j = i\hbar \sum_j \varphi_j \left(\frac{d\varphi_j}{dt} \right). \tag{2.12}$$

Дальнейшие наши усилия будут направлены на то, чтобы в (2.12) вообще избавиться от пространственной составляющей ϕ(r). Для этого умножим обе части (2.12) на одну из волновых функций ϕn, принадлежащих полному набору собственных функций ϕj, и проинтегрируем получившееся выражение по объему:

$$\sum_j \varphi_j \int \varphi_n H \varphi_j dV = i\hbar \sum_j \frac{d\varphi_j}{dt} \int \varphi_n \varphi_j dV \,. \tag{2.13}$$

Поскольку волновые функции из полного набора ортогональны по отношению друг к другу, и если они к тому же нормированы на единицу, то есть

$$\int \varphi_n \varphi_j dV = \begin{cases} 1, & \text{если} \quad n = j \\ 0 & \text{если} \quad n \neq j \end{cases},$$

соотношение (2.13) после введения обозначения $\int \varphi_n H \varphi_j dV = H_{nj}$ сводится к виду

$$i\hbar \frac{d\varphi_n}{dt} = \sum_j H_{nj} \varphi_j \,, \tag{2.14}$$

где H_{nj} описывает взаимодействие состояния n с состоянием j.

Давайте теперь сообразим, чего мы добились. Исключив пространственную часть волновой функции, мы не сможем уже узнать вероят-

ность обнаружения электрона в окрестности какой-то точки − но нам это и не надо! Мы теперь можем получить ответ на вопрос, с какой вероятностью электрон в момент времени t находится в состоянии с номером n. Допустим (см. рис. 2.2), у нас есть одномерная цепочка атомов («одномерный кристалл»), в которой расстояние между соседями равно b (период решетки). Волновая функция φ_n определяет вероятность обнаружения электрона в момент времени t на узле с номером n (такую волновую функцию называют *узельной*). Допустим, электрон на узле n не взаимодействует больше ни с какими узлами, кроме родного. Тогда все H_{nj} при $n \neq j$ в (2.14) равны нулю и мы имеем уравнение

$$i\hbar \frac{d\varphi_n}{dt} = H_{nn}\varphi_n = E_n\varphi_n,$$

решением которого является гармоническая функция:

$$\varphi_n = K_n \cdot \exp\left(\frac{-iH_{nn}t}{\hbar}\right) = K_n \cdot \exp\left(\frac{-iE_n t}{\hbar}\right),$$

где E_n − энергия электрона на узле n, часто реально не означает ничего, кроме выбора нуля энергии. Следовательно, вероятность обнаружить электрон на узле n равна

$$\psi_n^* \psi_n = |K_n|^2$$

и не зависит от времени. В общем-то, результат тривиальный: если нет взаимодействия с другими узлами, то электрон, раз попав на узел n, навсегда останется в нем. Интереснее будет, когда мы учтем взаимодействие.

Начнем с самого простого случая и учтем взаимодействие только с ближайшими соседями, имеющими номера $n - 1$ и $n + 1$. Константа взаимодействия с узлами $n - 1$ и $n + 1$ одинакова, т. к. расстояние b до них одинаково, и мы просто обозначим ее $-A$, где знак «минус» говорит о том, что взаимодействие ослабляет связь электрона с узлом n, и его энергия должна иметь знак, противоположный знаку E. Тогда (2.14) будет выглядеть как

$$i\hbar \frac{d\varphi_n}{dt} = E_n\varphi_n - A\varphi_{n-1} - A\varphi_{n+1}. \qquad (2.15)$$

Такие же уравнения мы получим для любого другого узла − электрону одинаково вероятно оказаться около любого атома, ни один из атомов никаких преимуществ не имеет, так как цепочка строго перио-

дична (наш «кристалл» идеален).

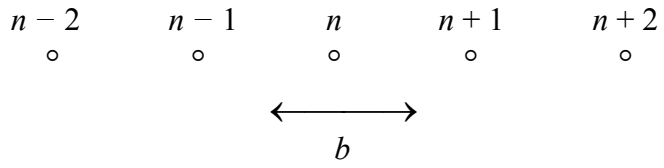


Рис. 2.2. Одномерная периодическая цепочка атомов

Решение уравнения (2.15) будем искать опять в виде

$$\varphi_j = K_j \cdot \exp\left(\frac{-iEt}{\hbar}\right), \tag{2.16}$$

где E – искомая энергия, а индекс j принимает три значения: $n - 1$, n, $n + 1$. Подставим (2.16) в (2.15) и получим:

$$EK_n = E_n K_n - \mathrm{A}\left(K_{n-1} + K_{n+1}\right). \tag{2.17}$$

Если атом n находится в точке с координатой x, то координаты соседних атомов будут $x + b$ и $x - b$. Тогда, рассматривая коэффициенты K_j как функции этих координат, мы можем записать (2.17) в виде

$$EK\left(x\right) = E_n K\left(x\right) - \mathrm{A}\left[K\left(x+b\right) + K\left(x-b\right)\right], \tag{2.18}$$

которая является разностным уравнением, очень похожим на дифференциальное, решение которого и поищем в традиционном виде

$$K(x) = \exp(ikx). \tag{2.19}$$

С учетом (2.19) мы можем свести (2.18) к уравнению

$$E\exp(ikx) = E\exp(ikx) - A\{\exp[ik(x + b)] + \exp[ik(x - b)]\},$$

разделив которое на $\exp(ikx)$, получим:

$$E = E_n - A[\exp(ikb) + \exp(-ikb)].$$

Используя формулу Эйлера

$$\exp(ia) + \exp(-ia) = 2\cos(a),$$

получим окончательно для энергии электрона на узле n

$$E = E_n - 2A\cos(kb). \tag{2.20}$$

Что же это такое получилось? А вот что. Если электрон на изолированном узле (в изолированном атоме) имел энергетический уровень E_n, то в результате взаимодействия его энергетический уровень пре-

вратился в энергетическую зону шириной $4A$. Иначе говоря, электрон теперь может иметь энергию в интервале от E_n -$2A$ до E_n +$2A$ и не может иметь энергию за пределами этого интервала. Если, например, $E_n = E_1$, т. е. энергии первого возбужденного состояния, то в линейной цепочке мы получим ту самую зону проводимости (см. рис. 2.3), о которой говорили раньше:

$$E = E_1 - 2A\cos(kb).$$

Изображенная на рис. 2.3 кривая зависимости энергии E от k называется *зонной структурой*, или *дисперсионной кривой*. Почему? Чтобы ответить на этот вопрос, надо выяснить смысл величины k, с которой мы уже встречались в предыдущем разделе, вводя ее как $(2mE)^{1/2}/h$.

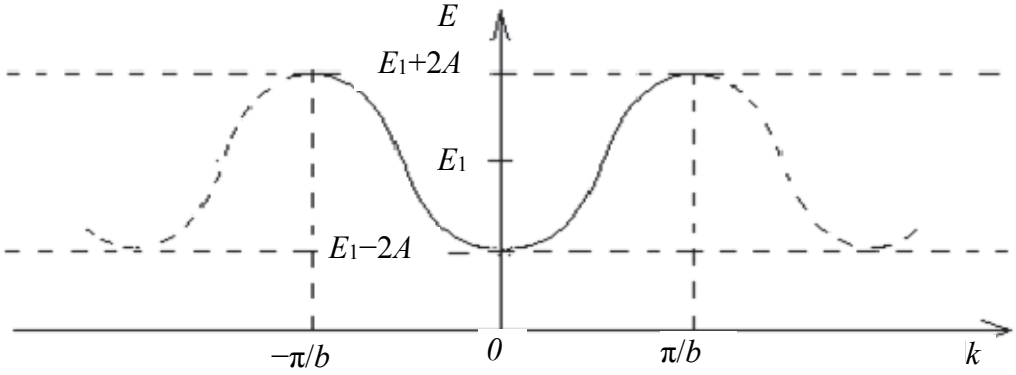


Рис. 2.3. Дисперсионная кривая для электрона в одномерной решетке

В более общем виде можно записать

$$k = \frac{\left[2m\left(E-U\right)\right]^{1/2}}{\hbar}, \qquad (2.21)$$

если учесть потенциальную энергию частицы. Величина (E-U) − это кинетическая энергия, тогда числитель в (2.21) не что иное как импульс частицы p:

$$k = \frac{p}{h} = \frac{2\pi p}{\hbar} = \frac{2\pi}{\lambda}, \qquad (2.22)$$

где λ − длина волны де Бройля. Значит, k − это просто-напросто волновое число электрона, а $\hbar k$ − его импульс, то есть кривая на рис. 2.3 фактически дает зависимость энергии электрона от его импульса, или *закон дисперсии*. Обратите внимание, что из (2.20) получается закон дисперсии не такой, как для свободного электрона, у которого зависимость E от k параболическая:

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m}. \qquad (2.23)$$

Сам характер зависимости (2.20) показывает, что все интересующие нас значения энергии сосредоточены в интервале волновых чисел от $-\pi/b$ до π/b, а за пределами этого интервала нет ничего нового − косинусоида просто повторяет сама себя. Такой интервал волновых чисел в физике твердого тела часто называют *первой зоной Бриллюэна* (это не энергетическая зона!).

Если учесть взаимодействие не только с ближайшими соседями, закон дисперсии изменится. Он вообще не универсален и зависит от сделанных нами допущений. Как мы убедимся, рассматривая зонную структуру реальных полупроводников, это не обязательно косинусоида, и она даже не обязательно симметрична. Зато всегда верно, что все разрешенные состояния спрятаны в первой зоне Бриллюэна, и заботиться о других значениях волнового числа незачем.

В качестве полезного упражнения попробуйте учесть взаимодействие с двумя ближайшими соседями, если константы взаимодействия с первым и вторым соседом относятся друг к другу в одном случае, как 2.83, а в другом − это отношение равно 4.

2.3. Туннелирование электронов в решетке

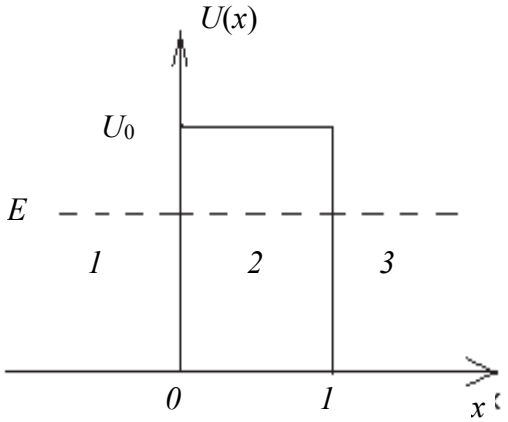


Рис. 2.4. Прохождение частицы через потенциальный барьер

Уже фейнмановская модель показывает, что электроны в твердом теле описываются только в рамках квантово-механических представлений, и всякое использование аналогий из классической механики чревато серьезными ошибками. Чтобы прояснить еще раз этот вопрос, рассмотрим другое представление об электронных энергетических зонах, также редко встречающееся в учебной литературе. Сперва вкратце вспомним, в чем заключается и как описывается туннельный эффект. Явление это, не имеющее никакого аналога в классической механике, представляет собой проникновение частиц сквозь потенци-

альный барьер. Если свободная частица с энергией E налетает слева на прямоугольный потенциальный барьер высотой $U_0 > E$ и шириной l (рис. 2.4), то ее волновая функция в областях 1, 2, 3 может быть записана как

$$\psi_1 = A_1 e^{ik_1 x} + B_1 e^{-ik_1 x},\ \psi_2 = A_2 e^{k_2 x} + B_2 e^{-k_2 x},\ \psi_3 = A_3 e^{ik_3 x} + B_3 e^{-ik_3 x}, \quad (2.24)$$

где $k_1 = k_3 = \dfrac{\sqrt{2mE}}{\hbar}$, $k_2 = \dfrac{\sqrt{2m(U_0 - E)}}{\hbar}$.

Поскольку в области *3* нет отраженной волны, а в области *2* − экспоненциально нарастающей, следует положить $B_3 = 0$, $A_2 = 0$. Используя свойства непрерывности волновой функции и ее производной, можно записать:

$$\psi_1(0) = \psi_2(0),\ \ \psi_1'(0) = \psi_2'(0),\ \ \psi_2(1) = \psi_3(1),\ \psi_2'(1) = \psi_3'(1). \quad (2.25)$$

Вероятность прохождения электрона (или любой другой частицы) сквозь барьер, т. е. из области *1* в область *3*

$$D = \left(\frac{k_3}{k_1}\right) \cdot \left(\frac{|A_3|^2}{|A_1|^2}\right) = \frac{|A_3|^2}{|A_1|^2},$$

после подстановки (2.24) в (2.25) получается в виде

$$D = D_0 \cdot \exp\left\{\frac{-2l}{\hbar}\sqrt{2m(U_0 - E)}\right\}. \quad (2.26)$$

Любой не слишком резкий барьер произвольной формы $U(x)$ может быть разбит на много прямоугольных, тогда коэффициент прохождения такого барьера

$$D = D_0 \cdot \exp\left\{\frac{-2}{\hbar}\int_{x_1}^{x_2}\sqrt{2m(U(x) - E)}dx\right\}, \quad (2.27)$$

где x_1 и x_2 − координаты точек входа частицы под барьер и выхода из-под барьера соответственно.

Таким образом, частица обладает ненулевой вероятностью туннелирования сквозь барьер, сильно возрастающей при уменьшении высоты и ширины барьера, а также массы частицы. Точно также частица, пролетающая над барьером, может отразиться и вернуться в область *1*. Графически зависимость вероятности туннелирования от соотношения между энергией частицы и высотой барьера представлена на рис. 2.5.

Пусть теперь частица налетает на двугорбый симметричный потенцальный барьер (рис. 2.6). В этом случае при некоторых (так называемых квазистационарных) значениях энергии частицы, два из которых, E_0 и E_1 (а их может быть существенно больше в зависимости от параметров барьера) показаны на рис. 2.6 и 2.7, коэффициент прохождения барьера равен единице.Иначе говоря, такой барьер абсолютно прозрачен для частиц с квазистационарными значениями энергии. Если же взять N-горбый периодический барьер, то каждый квазистационарный уровень расплывается в зону, содержащую $N - 1$ близко расположенных уровней (см. рис. 2.8.), и частицы со значениями энергии, попадающими в эти зоны, туннелируют через такую одномерную решетку совершенно свободно, с единичной вероятностью (рис. 2.9).

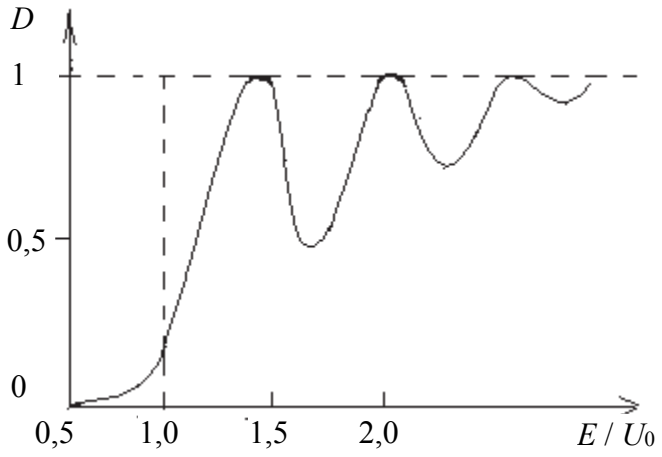


Рис. 2.5. Коэффициент прохождения частицы через прямоугольный потенциальный барьер

Вот такие «прозрачные окна» и являются электронными энергетическими зонами в периодической решетке, то есть в кристалле. Если, например, E_0 − верхний заполненный уровень в изолированном атоме, а E_1 − нижний пустой, то при образовании из таких атомов кристалла ΔE_0 будет валентной зоной, а ΔE_1 − зоной проводимости.

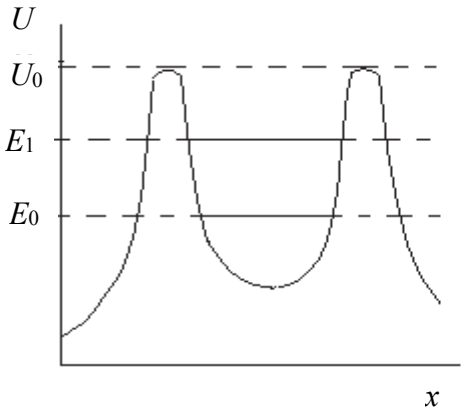


Рис. 2.6

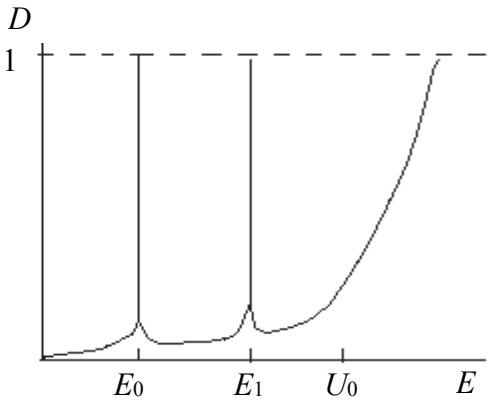


Рис. 2.7

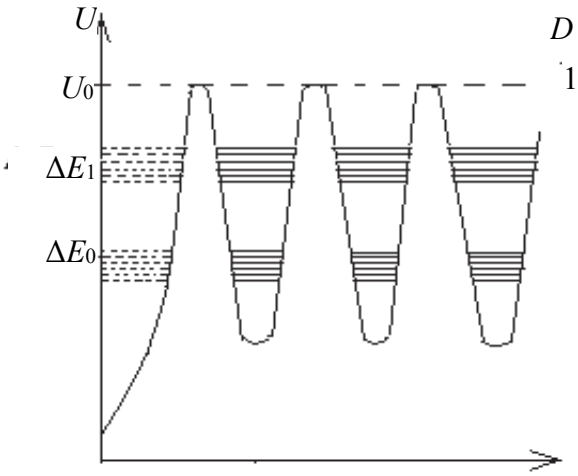


Рис. 2.8

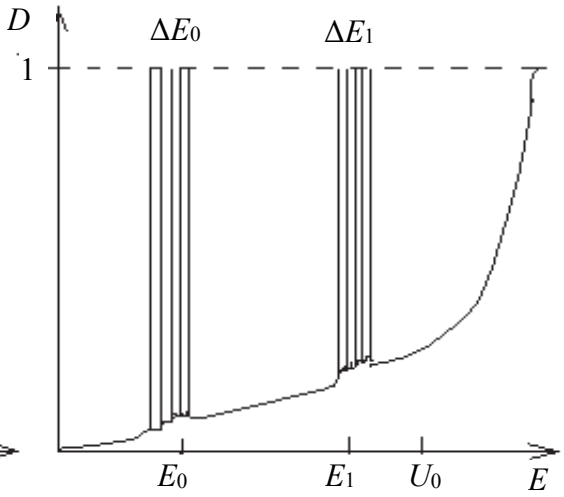


Рис. 2.9

Такое рассмотрение вызывает сильное подозрение, что электрон в периодической решетке не должен испытывать никаких соударений с атомами, а будет двигаться, как говорят, без рассеяния (его будут нести на себе «волны вероятности», благополучно огибая все рифы). Как мы увидим скоро, так оно и оказывается.

2.4. Динамика электронов

Ознакомившись с энергетическими зонами, мы можем сделать следующий шаг к пониманию электронных свойств твердых тел − построить динамику электрона в кристаллической решетке. Ясно, что строить динамику можно различными способами, и мы с Вами поступим наиболее утилитарным, который в итоге позволит нам описывать электронные свойства полупроводников почти во всех практически важных случаях, но абсолютно непригоден, например, для металлов или имеющих узкие энергетические зоны молекулярных кристаллов. Способ этот называется *методом эффективной массы* и основан на хитроумной уловке: вначале вычисляется скорость электрона как квантового объекта, а потом используется обычная ньютоновская динамика.

Когда мы говорим, что электрон − квантовый объект, то в первую очередь имеем в виду, что он обладает волновыми свойствами, и если имеет энергию E, то его распространение характеризуется циклической частотой

$$\omega = \frac{E}{\hbar}. \qquad (2.28)$$

Скорость переноса энергии в любой волне − это групповая скорость

$$\upsilon = \frac{d\omega}{dk},$$

где $k = 2\pi/\lambda$ − волновое число. Тогда с учетом (2.28) скорость электрона

$$\upsilon = \frac{1}{\hbar} \cdot \frac{\partial E}{\partial k}. \qquad (2.29)$$

Например, для свободного электрона, у которого

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m},$$

соотношение (2.29) дает

$$\upsilon = \frac{1}{\hbar} \cdot \frac{2\hbar^2 k}{2m} = \frac{\hbar k}{m} = \frac{p}{m},$$

то есть обычную скорость частицы с массой m и импульсом p.

Если мы начнем разгонять электрон электрическим полем с напряженностью &, то за время dt он приобретет энергию dE, равную совершенной за это время полем работе:

$$dE = Fdx = e\,\&\,\upsilon dt = e\,\&\,\frac{1}{\hbar} \cdot \frac{\partial E}{\partial k} dt,$$

откуда

$$\frac{\partial k}{\partial t} = \frac{e\,\&}{\hbar}. \qquad (2.30)$$

Теперь, используя (2.29), найдем ускорение:

$$a = \frac{d\upsilon}{dt} = \frac{1}{\hbar} \cdot \frac{d}{dt}\left[\frac{\partial E}{\partial k}\right] = \frac{1}{\hbar} \cdot \frac{\partial^2 E}{\partial k^2} \cdot \frac{\partial k}{\partial t},$$

что с учетом (2.30) можно записать как

$$a = \frac{e\,\&}{\hbar^2}\left(\frac{\partial^2 E}{\partial k^2}\right),$$

или, вспомнив, что $e\mathcal{E} = F$ − внешняя сила, действующая со стороны электрического поля и разгоняющая электрон, еще интересней:

$$F = \left(\frac{\hbar^2}{\partial^2 E / \partial k^2} \right) a, \tag{2.31}$$

что внешне очень напоминает формулу $F = ma$. Если величину

$$m^* = \frac{\hbar^2}{\partial^2 E / \partial k^2} \tag{2.32}$$

назвать *эффективной массой* электрона, то мы получим некий аналог второго закона Ньютона:

$$F = m^* a,$$

в котором, конечно, величина m^* не отражает ни инертных, ни гравитационных свойств электрона.

Попробуем применить (2.32) для свободного электрона, у которого $E = \hbar^2 k^2 / 2m$, и, следовательно,

$$\frac{\partial E}{\partial k} = \frac{\hbar^2 k}{m}, \quad \frac{\partial^2 E}{\partial k^2} = \frac{\hbar^2}{m}.$$

Тогда

$$m^* = \frac{\hbar^2}{\partial^2 E / \partial k^2} = \frac{\hbar^2}{\hbar^2} m = m,$$

то есть для свободного электрона мы получаем обычную массу, значит, в этом случае наша формула (2.31) работает, как положено.

Теперь рассмотрим одномерную решетку, где в соответствии с моделью Фейнмана

$$E = E_1 - 2A \cos(kb), \tag{2.33}$$

и, следовательно, скорость в соответствии с (2.29)

$$\upsilon = \frac{1}{\hbar} \cdot \frac{\partial E}{\partial k} = \frac{2Ab}{\hbar} \sin(kb), \tag{2.34}$$

а эффективная масса

$$m^* = \frac{\hbar^2}{\partial^2 E / \partial k^2} = \frac{\hbar^2}{2Ab^2 \cos(kb)}. \tag{2.35}$$

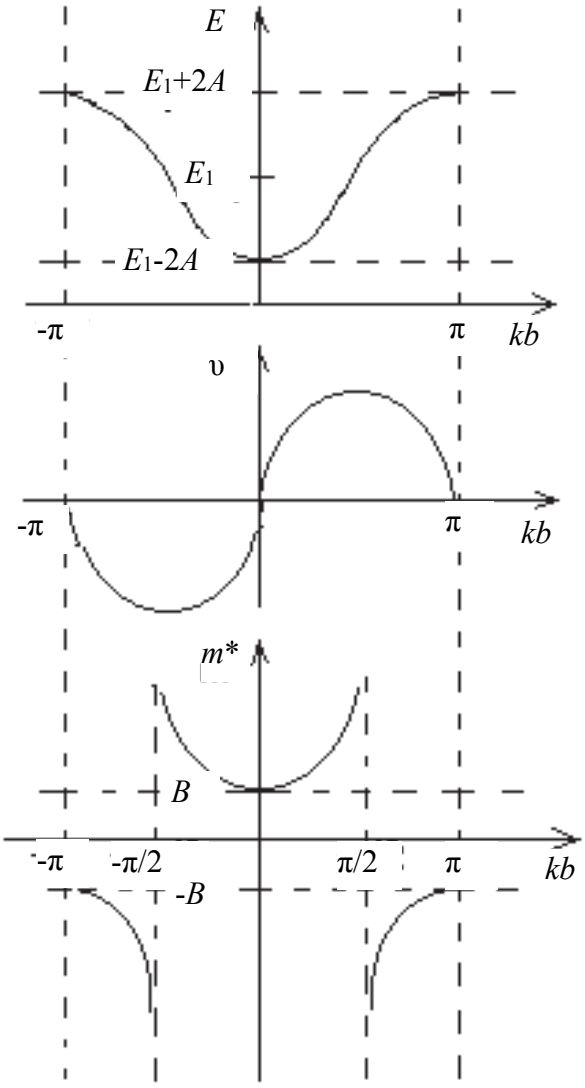


Рис. 2.10. Энергия, скорость и эффективная масса электрона в первой зоне Бриллюэна

В интервале от $k = \pi/2b$ до $k = \pi/b$ эффективная масса отрицательна, то есть электрон будет разгоняться навстречу действующей силе. Если мы перейдем к трехмерному случаю, то все еще больше запутается. Во-первых, эффективная масса будет зависеть от направления, то есть одно и то же электрическое поле, приложенное по разным осям, создаст разное ускорение. Это уже само по себе достаточно скверно, но может быть еще хуже − поле, приложенное вдоль оси x, может вызвать ускорение вдоль оси y. Как сказал бы математик, эффективная масса является тензором, но для нас с вами это означает, что понятием эффективной массы надо пользоваться аккуратно и даже осторожно.

Первый вопрос, в котором нам поможет понятие эффективной массы − это вопрос о том, какие электроны вносят свой вклад в проводимость. Выражение для плотности тока

$$j = en\upsilon \tag{2.36}$$

справедливо только, если у всех электронов одинаковые скорости, что сомнительно. Теперь у нас есть формула (2.34), которая однозначно утверждает, что υ зависит от волнового числа k, а, следовательно, в соответствии с (2.33) и от энергии. Значит, мы должны вычислять плотность тока как

$$j = \int e\upsilon(E)\,dn\,. \tag{2.37}$$

Однако мы можем попробовать формально использовать соотношение (2.36), введя понятие эффективной концентрации электронов $n_{эфф}$,

которые являются свободными, не взаимодействуют с решеткой и обеспечивают весь перенос заряда. Для таких электронов мы запишем

$$j = e\upsilon n_{\text{эфф}} = ea\tau n_{\text{эфф}} = \frac{e^2 \,\&\, \tau}{m} n_{\text{эфф}} . \qquad (2.38)$$

В свою очередь (2.37) можно представить в виде:

$$j = \int ea(E)\tau dn = \int e\frac{eA\,\&}{m^*}\tau dn = \frac{e^2 \,\&\, \tau}{\hbar^2}\int \frac{\partial^2 E}{\partial k^2} dn .$$

Поскольку для одномерного случая $dn = dk/\pi$, то мы получаем

$$j = \frac{e^2 \,\&\, \tau}{\pi\hbar^2}\int \frac{\partial^2 E}{\partial k^2} dk ,$$

что после сопоставления с (2.38) дает нам выражение для эффективной концентрации электронов:

$$n_{\text{эфф}} = \frac{m}{\pi\hbar^2}\int \frac{\partial^2 E}{\partial k^2} dk , \qquad (2.39)$$

которое, конечно, справедливо только при нулевой температуре, так как никакой заселенности уровней мы не учитывали; в данном случае это совершенно неважно. Интеграл мы должны вычислить для всех волновых чисел, которыми обладают электроны. Значит, если все уровни от самого нижнего до E_M заполнены, то интеграл в (2.39) надо брать в пределах от $-k_1$ до k_1 (рис. 2.11), то есть

$$n_{\text{эфф}} = \frac{m}{\pi\hbar^2}\left[\left(\frac{\partial E}{\partial k}\right)_{k=k_1} - \left(\frac{\partial E}{\partial k}\right)_{k=-k_1}\right],$$

или, поскольку косинус − четная функция,

$$n_{\text{эфф}} = \frac{2m}{\pi\hbar^2}\left(\frac{\partial E}{\partial k}\right)_{k=k_1} , \qquad (2.40)$$

а в явном виде

$$n_{\text{эфф}} = \frac{4mAb}{\pi\hbar^2}\sin\left(k_1 b\right). \qquad (2.41)$$

Полученное соотношение (2.41) очень любопытно. Во-первых, если $k_1 = 0$, то $n_{\text{эфф}} = 0$ и тока нет – ну, это тривиально, если зона пуста и нет носителей, то откуда вообще возьмется ток?

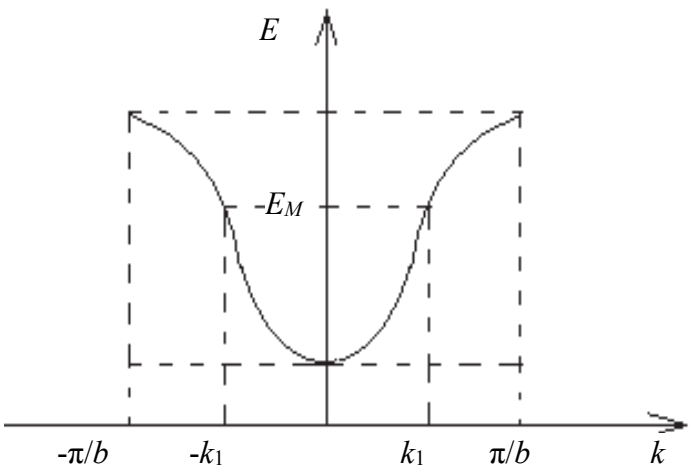


Рис. 2.11. Эффективная концентрация электронов определяется граничной энергией E_M и граничным волновым числом k_1

Но если $k_1 = \pi/b$, то есть зона вся заполнена, то тока опять нет! А когда же проводимость будет самой большой? Когда $k_1 = \pi/2b$, то есть $E_M = E_1$ − когда зона заполнена ровно наполовину (типичный случай металла с одним электроном проводимости на каждый атом). Следует отметить, что соотношение (2.40) носит более общий характер, чем (2.41) (подумайте сами, почему).

Еще одна возможность, которую нам предоставляет эффективная масса − это введение понятия дырки, и мы эту возможность не упустим. Как вы знаете, дырка − это пустое место с положительным зарядом, и такой полностью самостоятельной частицы не существует. Зачем вводится понятие дырки, вам объясняли еще в школе, но почему это можно делать?

Давайте уберем из полностью заполненной зоны, по которой, как мы только что доказали, ток не течет, немного электронов, так что величина $k_1 b = \pi$ уменьшилась на α. Тогда в цепочке появится ток с плотностью

$$j = \frac{e^2 \,\&\, \tau}{m} n_{\text{эфф}} = \frac{4e^2 \,\&\, \tau A b}{\pi \hbar^2} \sin(\pi - \alpha),$$

а поскольку $\sin(\pi - \alpha) = \sin\alpha$, то это такой же ток, который создавало бы такое же количество электронов у дна зоны. Вот только посмотрите на (2.35) − электроны у потолка зоны имеют отрицательную эффективную массу. Значит, исчезновение электрона с отрицательной массой приводит к возникновению тока, однако не проще ли говорить о появлении положительно заряженной дырки? Тогда если в каком-то материале проводимость обусловлена, например, двумя зонами, в одной из которых мало электронов с подвижностью μ_e и концентрацией n_e, а в другой мало дырок (много электронов) с подвижностью μ_h и концентрацией n_h, то удельная проводимость такого материала (так называемая амбиполярная проводимость)

$$\sigma = e\, n_e\, \mu_e + e\, n_h\, \mu_h. \qquad (2.42)$$

Не правда ли, очень похожа на параллельное соединение проводников? И действительно, в этом случае говорят о *параллельных каналах* проводимости просто потому, что у нас суммируются токи в этих «каналах».

2.5. Зонная структура реальных полупроводников

При вычислении концентраций электронов и дырок в полу − проводниках и соответствующей им энергии Ферми совершенно безразлично, являются ли эффективные массы скалярами или тензорами, так как в последнем случае в плотность состояний достаточно подставить усредненный «скалярный эквивалент» тензора эффективной массы. Однако ряд явлений в полупроводниковых материалах определяется именно фактической формой поверхностей постоянной энергии в трехмерном k-пространстве. Мы до сих пор для простоты полагали, что полупроводники обладают простой кубической структурой. На самом же деле это не так. Например, наиболее широко используемые в микроэлектронике кремний и германий имеют кристаллическую решетку типа алмаза, совершенно не похожую на простую кубическую. Вследствие этого, дисперсионные кривые для разных направлений будут различными. Кроме того, для каждого фиксированного направления зонная структура заметно сложнее, чем это предсказывается простой моделью Фейнмана (см. рис. 2.12). Например, в валентной зоне существует *три типа дырок* (h − тяжелые, l − легкие, s − отщепленные), что связано, вообще говоря, с тем, что валентная зона формируется электронами как s-состояний атомов, так и p-состояний, и эти электроны к тому же взаимодействуют между собой. В зоне проводимости подобные фокусы тоже происходят, но электрические свойства определяются фактически только одной дисперсионной кривой, минимум которой зато не совпадает с центром первой зоны Бриллюэна. Вследствие этого, полупроводники такого типа называются *непрямозонными*. Происхождение этого названия связано с процессами поглощения света в полупроводниках.

Фонтон, обладающий достаточной для возбуждения электорна энергией, обладает в то же время весьма малым импульсом (попробуйте сами оценить численно, что это означает, связывая ширину зоны Бриллюэна с постоянной решетки).

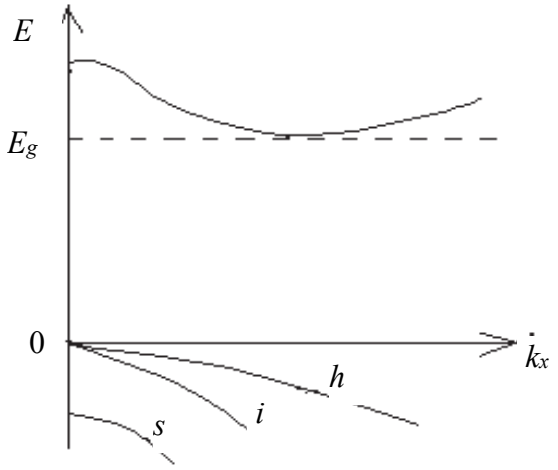


Рис. 2.12. Структура зон для одного из направлений в кремнии

Поэтому на рис. 2.12 вызванный поглощением света электронный переход будет изображаться практически вертикальной линией, или, как говорят, оптические переходы являются *прямыми*. Следовательно, фотоны не могут возбудить электроны из максимума валентной зоны в минимум зоны проводимости. Поэтому, кстати, ширина запрещенной зоны, определенная по краю полосы собственного поглощения в кремнии или германии, будет больше табличной.

После выполнения всех необходимых процедур усреднения эффективная масса и время релаксации в конечном итоге оказываются изотропными. Соответственно, измеренная в разных направлениях проводимость оказывается одной и той же. Другое важное отклонение зонной структуры от идеальной наблюдается в ряде соединений элементов третьей и пятой группы таблицы Менделеева. У таких соединений, типичным и промышленно важным представителем которых является арсенид галлия, в зоне проводимости появляется дополнительная долина (см. рис. 2.13).

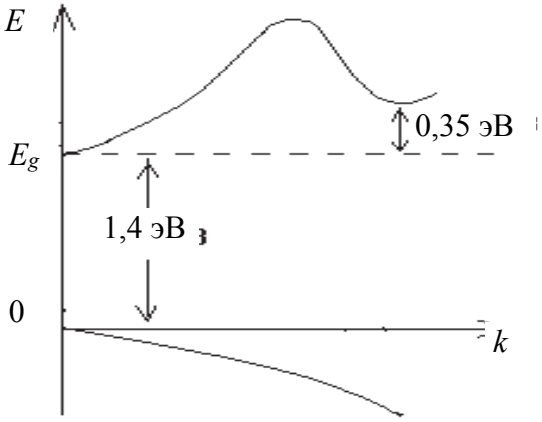


Рис. 2.13. Зонная структура арсенида галлия

Электроны в нижней долине GaAs имеют эффективную массу $0{,}067m_0$ (m_0 − масса покоя электрона), а в верхней − $0{,}35m_0$. Появление второй долины связано со значительным вкладом ионной компоненты в полную энергию химической связи в таких материалах (вспомните задачу из раздела, посвященного модели Фейнмана!). Наличие дополнительной долины проявляется в сильных электрических полях, составляя основу функционирования очень интересных электронных приборов − диодов Ганна. В заключение следует отметить, что детали зонной структуры, как правило, несущественны при разработке подавляющего большинства полупроводниковых устройств.

3. ЭЛЕКТРОНЫ В МЕТАЛЛАХ

Статистика свободных электронов

Мы уже знаем, что в твердом теле энергетические уровни отдельных атомов превращаются в зоны, или наборы уровней. На каждом из таких уровней принцип Паули позволяет разместиться только двум электронам с противоположной ориентацией спинов, поэтому различные электроны в кристалле будут иметь различную энергию, а значит, видимо, и различную скорость. Кроме того, сами уровни вряд ли будут расположены через равномерные промежутки по шкале энергий, вспомним, хотя бы электрон в прямоугольной яме. Интересно было бы знать, как эти уровни распределены по энергиям и как зависит от энергии скорость электрона. Зачем? Причин для такого любопытства множество, но одна из них сразу бросается в глаза. Мы хотим описать электрический ток в твердом теле, а плотность тока (по определению), как нам известно из первой главы,

$$j = en\upsilon_D. \tag{3.1}$$

Если электроны с разными энергиями имеют разные скорости, то есть $\upsilon_D = \upsilon_D(E)$, то совершенно непонятно, какую скорость подставлять в (3.1). Надо, конечно, поступить так: взять узкий интервал энергий dE, в пределах которого скорость практически не меняется, и если концентрация электронов, имеющих энергии, лежащие в этом интервале, равна dn, то можно по формуле (3.1) посчитать плотность тока, создаваемого этими электронами:

$$dj = e\upsilon_D(E)dn, \tag{3.2}$$

чтобы затем, проинтегрировав (3.2) по энергиям, получить полную плотность тока, создаваемого всеми электронами. Однако для этого надо знать в явном виде функции $\upsilon_D(E)$ и $n(E)$. Первую задачу (динамическую) мы уже решили, а второй задачей займемся сейчас.

Прежде всего, мы должны ввести понятие *плотности состояний* $g(E)$ − это, по определению, количество электронных энергетических уровней, или, как чаще говорят, количество электронных состояний, приходящихся на единицу объема твердого тела и на единичный интервал энергий. С использованием функции $g(E)$ величина dn в (3.2) может быть представлена как

$$dn = 2g(E)f(E,T)dE, \tag{3.3}$$

где f − вероятность того, что на данном уровне есть электрон, зависящая от энергии и температуры, а множитель 2 просто говорит о том, что на уровне может быть 2 электрона с противоположной ориентацией спинов.

Теперь для вычисления тока нам понадобилась в явном виде функция $g(E)$. А где ее взять? В общем случае ее приходится определять экспериментально, и эксперименты эти непростые − позже мы с ними ознакомимся. Однако в простых случаях, которых нам хватит на половину курса, плотность состояний можно рассчитать теоретически, используя для твердого тела *модель Зоммерфельда*, придуманную, вообще-то, для металлов. В этой модели весь кристалл считается одной бесконечно глубокой прямоугольной потенциальной ямой, в пределах которой электроны проводимости свободны. Иначе говоря, теория Зоммерфельда − это теория свободных электронов.

Для вычисления плотности состояний удобнее всего воспользоваться понятием *фазового пространства*, то есть объединенного абстрактного пространства координат и импульсов. Для одной частицы такое пространство будет иметь шесть измерений, соответствующих трем координатам и трем проекциям импульса частицы. Элемент объема такого пространства

$$d\Gamma = dx\, dy\, dz\, dp_x\, dp_y\, dp_z.$$

На одно электронное состояние в таком пространстве приходится объем, равный h^3 , как это следует из соотношения неопределенностей Гайзенберга:

$$\Delta p_x\ \Delta x \geq h,$$

$$\Delta p_y\ \Delta y \geq h,$$

$$\Delta p_z\ \Delta z \geq h,$$

что после перемножения дает

$$(\Delta\Gamma)_{\min} = (\Delta x\, \Delta y\, \Delta z\, \Delta p_x\, \Delta p_y\, \Delta p_z)_{\min} = h^3.$$

Тогда число электронных состояний в объеме пространства

$$dV = dx\, dy\, dz$$

и в интервале импульсов от p до $p + dp$ будет равно

$$\frac{4\pi p^2 dp d\upsilon}{h^3},$$

что на единицу объема дает число электронных состояний

$$\frac{4\pi p^2 dp}{h^3}.$$

Поскольку мы рассматриваем свободные электроны, то их энергия − это просто кинетическая энергия, связанная с импульсом как

$$E = \frac{p^2}{2m}.$$

Значит,

$$p^2 = 2mE,$$

$$dp = \left(\frac{2m}{E}\right)^{1/2} \cdot \frac{dE}{2},$$

и число электронных состояний в единице объема равно

$$\frac{4\pi 2mE\sqrt{2m}dE}{2h^3\sqrt{E}} = 2\pi\left(\frac{2m}{h^2}\right)^{3/2}\sqrt{E}dE,$$

а в единице объема и единичном интервале энергий, что и является плотностью состояний

$$g(E) = 2\pi\left(\frac{2m}{h^2}\right)^{3/2}\sqrt{E}. \qquad (3.4)$$

Такова *плотность состояний в модели Зоммерфельда*, полученная в приближении свободных электронов. Таким образом, чем выше мы поднимаемся по шкале энергий, тем гуще расположены энергетические уровни. А как они заселены? Чтобы ответить на этот вопрос, нам нужна вероятность заполнения уровней, задаваемая функцией распределения $f(E,T)$, каковой для электронов является *функция распределения Ферми − Дирака:*

$$f(E,T) = \frac{1}{\exp\left(\frac{E - E_F}{kT}\right) + 1}, \qquad (3.5)$$

где E_F − *энергия Ферми*, или, как часто говорят, *уровень Ферми*. Чтобы дать определение уровня Ферми, подставим в (3.5) $E = E_F$ и получим $f(E,T) = 1/2$. Значит, *уровень Ферми − это энергетический уровень, вероятность заполнения которого* 1/2.

В идеализированном случае T = 0 соотношение (3.5) приводит к значениям функции распределения:

$$f\left(E,T\right)=\begin{cases}1 & \text{при } E<E_F,\\ 0 & \text{при } E>E_F,\end{cases} \tag{3.6}$$

то есть при абсолютном нуле температур все разрешенные состояния ниже уровня Ферми заполнены электронами, а все состояния выше уровня Ферми пусты. Иначе говоря, *при T = 0 уровень Ферми − это граница на шкале энергий, отделяющая заполненные и пустые уровни энергии электронов*. Физический смысл (3.6) очевиден − при $T = 0$ электроны заполняют самые низкие энергетические уровни. Однако по принципу Паули каждое состояние может быть занято только одним электроном, поэтому уровни до некоторого максимального заняты с единичным числом заполнения, а для вышележащих уровней числа заполнения равны нулю. Рис. 3.1 иллюстрирует заполнение электронных уровней в металлах при нулевой температуре в приближении Зоммерфельда. Используя (3.3) и (3.4), мы можем записать выражение для концентрации свободных электронов:

$$n=\int dn=\int_0^\infty 2g\left(E\right)f\left(E,T\right)dE=4\pi\left(\frac{2m}{h^2}\right)^{3/2}\int_0^\infty\sqrt{E}f\left(E,T\right)dE, \tag{3.7}$$

которое при T = 0 примет вид, в соответствии с (3.6),

$$n=4\pi\left(\frac{2m}{h^2}\right)^{3/2}\int_0^{E_F}\sqrt{E}dE=\frac{8\pi}{3}\left(\frac{2m}{h^2}\right)^{3/2}E_F^{3/2},$$

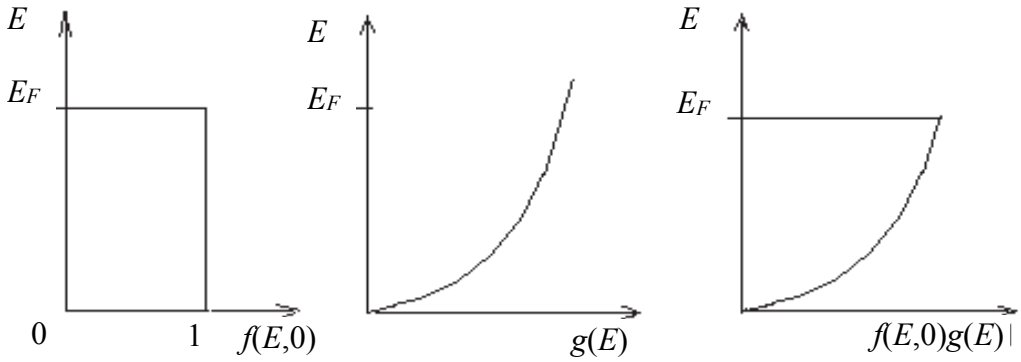


Рис. 3.1. Функция распределения Ферми − Дирака, функция плотности электронных состояний и произведение этих двух функций при абсолютном нуле температур

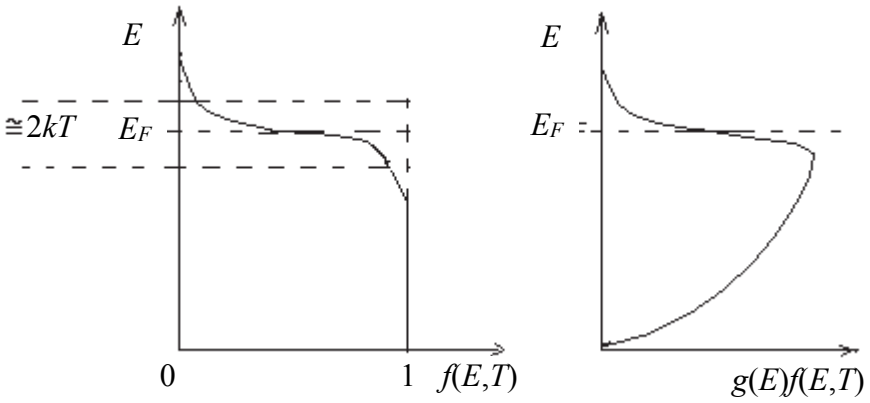


Рис. 3.2. Энергетическое распределение свободных электронов при ненулевой температуре

А что будет при ненулевой температуре? В соответствии с (3.5) резкая ступенька на зависимости $f(E)$ в окрестности уровня Ферми станет более плавной (см. рис. 3.2), причем функция Ферми − Дирака будет существенно отличаться от нуля и единицы только на расстояниях порядка kT от уровня Ферми, то есть в полосе энергий шириной около $2kT$ вблизи E_F. Это означает, что полученная выше формула (3.8) и возможные следствия из нее будут хорошо выполняться при условии $kT \ll E_{F0}$, то есть при температурах

$$T \ll T_0 = E_{F0}/k, \tag{3.9}$$

где T_0 называют *температурой вырождения*. Если условие (3.9) выполняется, то электронный газ называется вырожденным, в противном случае − невырожденным. Оценив численное значение $T_0 \cong 10^5$ К, мы убеждаемся, что электронный газ в металлах остается вырожденным при любых температурах вплоть до точки плавления. Но что это означает? Чтобы прояснить этот вопрос, рассмотрим свойства невырожденного электронного газа. В таком газе практически для всех энергий E будет выполняться условие $E - E_F \gg kT$, а значит, единицей в знаменателе выражения (3.5) можно пренебречь, и функция распределения примет вид

$$f\left(E,T\right) \cong \exp\left(\frac{E_F - E}{kT}\right),$$

то есть станет классической функцией распределения Максвелла − Больцмана. Таким образом, *вырожденный электронный газ является сугубо квантовым* и подчиняется статистике Ферми − Дирака, а невырожденный можно считать классическим.

Теперь становится ясно, что пользоваться какими-либо классическими представлениями об электронах в металле, как мы пытались делать это в первой главе, просто бесполезно. Рассмотрим физические следствия из квантового характера статистики электронов на примере электронной теплоемкости металлов, разрешив тем самым загадку, которая оказалась не «по зубам» классической теории.

Какие электроны в металле могут поглощать тепловую энергию? Ясно, что не те, которые располагаются существенно ниже уровня Ферми, так как, поглотив энергию порядка kT, электрон оказался бы на уже занятом уровне, а это запрещено принципом Паули. В состояниях же существенно выше (придайте, пожалуйста, сами количественный смысл слову «существенно») уровня Ферми электронов просто нет, то есть энергию поглощать некому. Таким образом, поглощают тепловую энергию только те электроны, которые имеют энергию в полосе шириной порядка $2kT$ около уровня Ферми, а эти электроны составляют долю около $2kT/E_F$ от общего количества электронов (то, что мы сейчас делаем − это, конечно, только полуколичественная оценка). При комнатной температуре $kT \cong 25$ мэВ, $2kT/E_F \cong$ $\cong 0{,}05/8$, то есть меньше 0,01. Это означает, что вклад в теплоемкость вносят менее одного процента свободных электронов, потому-то электронная теплоемкость и не обнаруживается на опыте.

Приведенные выше рассуждения справедливы и для других процессов, связанных с поглощением и переносом энергии, таких, к примеру, как теплопроводность и электропроводность − внешние возмущения воздействуют только на электроны близ уровня Ферми, подобно тому, как шторм на море треплет только корабли на поверхности, не затрагивая глубоководных обитателей.

4. ЭЛЕКТРОНЫ И ДЫРКИ В ПОЛУПРОВОДНИКАХ

4.1. Собственные полупроводники

Теперь, вооружившись минимально необходимой теорией, мы можем приступить к изучению электрических свойств особенно интересующих нас материалов − полупроводников. При этом мы начнем с чистых материалов, которые обычно называют *собственными полупроводниками*. Мы уже знаем, что это диэлектрики, ширина запрещенной зоны в которых имеет значения от нескольких сотен миллиэлектронвольт до двух-трех электронвольт, что много больше тепловой энергии kT, составляющей при комнатной температуре около 25 мэВ. При нулевой температуре валентная зона такого материала вся заполнена электронами, а зона проводимости пуста, и никакого тока нет. Если полупроводник подогреть, некоторые электроны, получившие достаточное количество тепловой энергии, перейдут из валентной зоны в зону проводимости и смогут обеспечить проводимость. Давайте попробуем посчитать концентрацию этих электронов, которые мы обычно называем свободными. Мы уже знаем, что эту концентрацию можно записать как

$$n_e = \int g(E) f(E,T) dE, \tag{4.1}$$

где нужно вычислить интеграл по всем энергиям от дна до потолка зоны проводимости. $f(E,T)$ − это просто функция распределения Ферми − Дирака, которая нам известна. А вот где взять вид функции плотности состояний $g(E)$? Здесь нам поможет то, что функция распределения экспоненциально, то есть очень быстро убывает с энергией, а значит, свободные электроны будут преимущественно скапливаться у дна зоны проводимости. Вспомнив модель Фейнмана, мы сообразим, что поскольку у дна зоны значения kb близки к нулю, то выражение

$$E = E_1 - 2A\cos(kb)$$

можно разложить в ряд Тейлора по степеням kb в окрестности точки $kb = 0$ (здесь мы пользуемся одномерной моделью, но этого нам достаточно).

Ограничиваясь членами ряда до второго порядка включительно, получим:

$$E = E_1 - 2A + Ak^2 b^2 . \tag{4.2}$$

Вспомнив, что эффективная масса

$$m^* = \frac{\hbar^2}{\partial^2 E / \partial k^2} = \frac{\hbar^2}{2Ab^2},$$

перепишем (4.2) в виде

$$E = E_1 - 2A + \frac{\hbar^2 k^2}{2m^*}. \qquad (4.3)$$

Первые два слагаемых в (4.3) фактически дают нам просто уровень отсчета энергии, поэтому, полагая $E_1 - 2A = 0$, (4.3), можно представить просто как

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

что полностью идентично выражению для энергии свободного электрона, если только в качестве его массы взять эффективную массу m^*. А это означает, что для описания плотности электронных состояний у дна зоны проводимости мы можем воспользоваться моделью Зоммерфельда.

Что еще мы забыли? А то, что, кроме свободных электронов в зоне проводимости, у нас появились и дырки в валентной зоне. Поскольку они будут скапливаться у потолка валентной зоны, то все, что мы получили для свободных электронов, справедливо и для дырок. Только в (4.1) интеграл надо брать по энергиям от дна до потолка валентной зоны, а применяя модель Зоммерфельда, использовать не эффективную массу электрона m_e^*, а эффективную массу дырки m_h^*. И еще одна маленькая хитрость. Поскольку в (4.1) $f(E,T)$ − функция распределения по энергиям *электронов*, для *дырок* в (4.1) мы должны подставить в качестве функции распределения $1 - f(E,T)$. Поняли, почему?

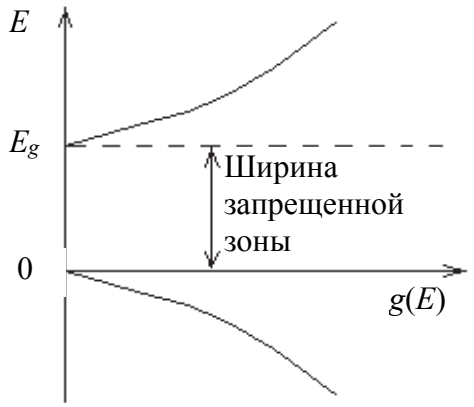


Рис. 4.1. Плотность состояний электронов и дырок в собственном полупроводнике

Поскольку мы хотим одновременно рассматривать две зоны, нам надо выбрать какую-то удобную шкалу энергий и в дальнейшем всегда ее придерживаться. Обычно в качестве нуля на шкале энергий выбирают потолок ва-

лентной зоны, а ширину запрещенной зоны обозначают E_g (от английского *gap* − щель). Тогда плотности состояний (см. рис. 4.1) для электронов и дырок равны

$$\left.\begin{aligned} g(E) &= 4\pi\frac{\left(2m_e^*\right)^{3/2}}{h^3}\left(E-E_g\right)^{1/2}, \\ g(E) &- 4\pi\frac{\left(2m_h^*\right)^{3/2}}{h^3}\left(-E\right)^{1/2}, \end{aligned}\right\} \tag{4.4}$$

И, соответственно, концентрации свободных электронов и дырок

$$\left.\begin{aligned} n_e &= 4\pi\frac{\left(2m_e^*\right)^{3/2}}{h^3}\int\limits_{E_g}^{\infty}\left(E-E_g\right)^{1/2}f(E,T)\,dE, \\ n_h &= 4\pi\frac{\left(2m_h^*\right)^{3/2}}{h^3}\int\limits_{-\infty}^{0}\left(-E\right)^{1/2}\left[1-f(E,T)\right]dE, \end{aligned}\right\} \tag{4.5}$$

где мы положили вместо энергии потолка зоны проводимости ∞, учитывая резкое убывание концентрации электронов с энергией, и по этой же причине для дырок вместо энергии дна валентной зоны можно положить -∞.

Если мы подставим в (4.5) функцию Ферми − Дирака

$$f(E,T)=\frac{1}{\exp\left(\dfrac{E-E_F}{kT}\right)+1},$$

то получим интегралы, не вычисляемые в квадратурах. Однако нетрудно сообразить, что уровень Ферми в собственном полупроводнике расположен где-то в середине запрещенной зоны (очень скоро мы найдем его точное положение), а значит, для всех электронов в зоне проводимости и валентной зоне (вспомните численные значения E_g!) $E - E_F \gg kT$, то есть функцию распределения можно представить в больцмановском виде:

$$f(E,T)=\exp\left(-\frac{E-E_g}{kT}\right). \tag{4.6}$$

Интересно, как следует назвать газ свободных электронов в полупроводнике − вырожденным или невырожденным?

Вот теперь все вычисляется! Подставим (4.6) в первый интеграл из (4.5), введем переменную $x = (E - E_g) / kT$ и, учитывая, что

$$\int_0^\infty x^{1/2} e^{-x} dx = \sqrt{\pi}/2,$$

получим концентрацию электронов в зоне проводимости:

$$n_e = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2} \exp\left(-\frac{E_g - E_F}{kT}\right),$$

или, если ввести обозначение

$$N_C = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{3/2},$$

$$n_e = N_C \exp\left(-\frac{E_g - E_F}{kT}\right). \qquad (4.7)$$

Еще немного помучившись с вычислениями, Вы получите из второй формулы в (4.5) концентрацию дырок:

$$n_h = N_v \exp\left(-\frac{E_F}{kT}\right), \quad N_v = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2}. \qquad (4.8)$$

Теперь мы с вами можем без труда рассчитать концентрации носителей тока в собственном полупроводнике, если только будем знать численные значения ширины запрещенной зоны и эффективных масс (они определяются экспериментально, а как − об этом позже). Нужные вам значения приведены в табл. 4.1, где буквой m_0 обозначена масса свободного электрона ($9{,}1 \cdot 10^{-31}$ кг).

Обратите внимание, что для эффективной массы дырок в германии и кремнии приведены по три значения − она различна для различных направлений в кристаллической решетке. Для вычисления концентрации дырок в этих материалах можете использовать *среднее геометрическое* этих трех масс (то есть кубический корень из их произведения).

Еще одна величина, необходимая нам для расчетов по формулам (4.7) и (4.8), − энергия Ферми E_F. Ее мы можем рассчитать, учитывая,

что в собственном полупроводнике концентрации свободных электронов и дырок одинаковы. Согласны? Если несогласны, предлагаю аргументы на выбор. Во-первых, *каждый* появляющийся в зоне проводимости электрон порождает *одну* дырку в валентной зоне. Во-вторых, не забывайте о законе сохранения заряда − наш материал, что бы там внутри ни происходило, должен оставаться электронейтральным. Убедились? Тогда приравняйте n_e и n_h и, используя (4.7) и (4.8), получите

$$E_F = \frac{E_g}{2} + \frac{3}{4}kT\ln\left(\frac{m_h^*}{m_e^*}\right), \tag{4.9}$$

то есть уровень Ферми в собственном полупроводнике действительно, как мы и предполагали выше, близок к середине запрещенной зоны, немного смещаясь вверх при повышении температуры. Полученный результат (4.9) доказывает применимость к собственным полупроводникам статистики Больцмана вместо статистики Ферми − Дирака. Еще одно существенное предположение, которое мы делали, заключается в том, что достаточно учесть только электроны и дырки на краях зон. Попробуйте убедиться сами в справедливости этого предположения.

Таблица 4.1.

Свойства собственных полупроводников при *T* = 300 K

Полупроводник	E_g, эВ	Эффективные массы		Подвижности, 10^4 м^2/В·с	
		m_e^*/m_0	m_h^*/m_0	μ_e	μ_h
Ge	0,67	0,12	0,04 0,28 0,08	3600	1800
Si	1,11	0,26	0,16 0,50 0,24	1500	500
GaAs	1,40	0,067	0,65	8500	400
GaP	2,25	0,35	0,5	150	140
InP	1,30	0,08	0,2	4600	150
InSb	0,17	0,013	0,18	70000	1000
CdS	2,5	0,27	0,07	340	18

4.2. Примесные полупроводники

Как мы уже отмечали, практическая ценность полупроводников обусловлена тем, что введение в них малого количества примеси (примерно одного атома на миллион атомов полупроводника) вызыва-

ет очень большое изменение проводимости. Химизм легирования полупроводников Вам хорошо известен, и нет смысла его обсуждать снова. Наша задача заключается в переводе этого химизма на язык зонной теории с тем, чтобы можно было количественно описывать электрические свойства примесных полупроводников. Для такого перевода в первую очередь необходимо оценить энергии ионизации примесных атомов.

Когда мы внедряем в полупроводник типа кремния или германия атом химического элемента V группы таблицы Менделеева (атом фосфора, мышьяка или сурьмы), то лишний пятый валентный электрон примесного атома, не встроившийся в систему ковалентных связей кристалла − матрицы, оказывается слабо связанным со своим родным атомом. А это означает, что такой электрон находится далеко от атомного остатка примеси, образуя вместе с ним нечто, очень похожее на атом водорода. Тогда, используя, как говорят, *водородоподобное приближение*, мы можем оценить энергию ионизации такой *донорной* примеси. Для этого нам достаточно даже боровской модели атома водорода, в соответствии с которой момент импульса электрона может иметь только значения

$$m\upsilon r = n\hbar, \tag{4.10}$$

где n − любое целое число, начиная с единицы; r − радиус электронной орбиты; υ − скорость движения электрона по орбите. Используя второй закон Ньютона в виде

$$\frac{m\upsilon^2}{r} = \frac{1}{4\pi\varepsilon_0\chi}\cdot\frac{e^2}{r^2}, \tag{4.11}$$

где χ − относительная диэлектрическая проницаемость легируемого материала, и, учитывая, что полная энергия электрона равна сумме кинетической и потенциальной, а именно,

$$E = \frac{m\upsilon^2}{2} - \frac{1}{4\pi\varepsilon_0\chi}\cdot\frac{e^2}{r}, \tag{4.12}$$

из (4.10) − (4.12) получим

$$E = -\frac{me^4}{32\pi^2\varepsilon_0^2\chi^2\hbar^2 n^2},$$

откуда энергия ионизации

$$\Delta E = E_\infty - E_1 = \frac{me^4}{32\pi^2\varepsilon_0^2\chi^2\hbar^2}. \qquad (4.13)$$

Используя численные значения диэлектрической проницаемости, получим энергию ионизации примесей в германии (χ = 15,4) $\Delta E \cong 10$ мэВ и в кремнии (χ = 12) $\Delta E \cong 30$ мэВ. Применяя концепцию дырки, мы получим точно такие же значения энергии ионизации *акцепторов* (от английского *accept* − принимать) − атомов химических элементов III группы (бор, алюминий, галлий, индий).

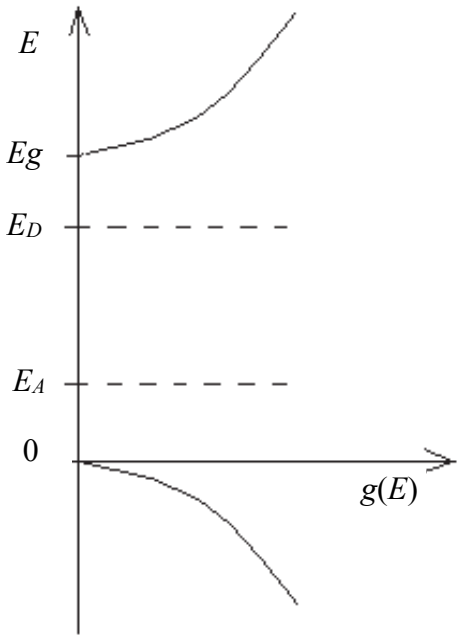


Рис. 4.2. Примесные уровни в полупроводнике

Посмотрев теперь на рис. 4.2, вы поймете, что же означают на языке зонной модели рассчитанные нами энергии − это энергетические зазоры между примесными уровнями и краями разрешенных зон, то есть для доноров $\Delta E = E_g - E_D$, а для акцепторов $\Delta E = E_A$. Конечно, полученные нами энергии − только оценочные значения, а точные значения можно получить из эксперимента. Экспериментальные методы мы будем обсуждать позже, а пока воспользуемся известными данными, приведенными в табл. 4.2.

Все эти энергии ионизации имеют величину порядка kT при комнатной температуре (учитывая это и посмотрев на рис. 4.2, вы сразу поймете, почему такие примеси называют «мелкими»), а значит, активно поставляют носители заряда − электроны в зону проводимости и дырки в валентную зону.

Чтобы больше не возвращаться к водородоподобной модели, отметим, что если из (4.10)−(4.12) рассчитать величину r для мелких примесей в кремнии и германии, то можно убедиться: армейский анекдот о том, что между ядром и электронами находится воздух не всегда бессмыслен.

Прежде чем двигаться дальше, давайте оговорим следующее: во-первых, мы будем рассматривать только *слабо легированные* полупроводники, в которых примесные уровни локальны и не образуют при-

месную зону. Примесные уровни только поставляют носители тока, а по ним самим никакого электропереноса нет, иначе наша теория не сработает (сильно легированные полупроводники больше напоминают по своим электронным свойствам металлы). Во-вторых, давайте сообразим, что будет, если в полупроводнике есть и доноры, и акцепторы. В этом случае принцип минимума энергии заставит электроны с гораздо большей вероятностью «опуститься» на акцепторные уровни, а не «подняться» в зону проводимости − такое явление называют *компенсацией* примесей, а отношение концентрации тех примесей, которых меньше, к концентрации тех примесей, которых больше, называют *степенью компенсации* К. Например, если концентрация доноров n_D больше, чем концентрация акцепторов n_A, то К = n_A/n_D, и полупроводник называют полупроводником *n*-типа (от английского «*negative*» − отрицательный). Догадываетесь, почему? В противном случае К = n_D/n_A, и полупроводник будет *p*-типа (*positive* − положительный). Ту примесь, которой больше, называют *основной*, а которой меньше − *неосновной*. Точно такую же терминологию применяют и к носителям – в полупроводнике *n*-типа основными носителями будут электроны, а неосновными − дырки, в полупроводнике *p*-типа наоборот. В большинстве полупроводниковых приборов используются *слабо компенсированные* полупроводники (К « 1), в которых неосновные примеси − это те, от которых просто не удалось очистить исходный материал.

Вот теперь мы можем начать вычислять концентрации носителей тока. А, собственно говоря, что значит начать? Мы их можем просто записать!

Таблица 4.2

Свойства примесных полупроводников при T = 300 К ($kT \cong 25$ мэВ)

Тип	Примесь	Энергия ионизации примеси, мэВ	
		Ge	Si
Доноры	Сурьма (Sb)	9,6	39
	Фосфор (P)	12,0	45
	Мышьяк (As)	12,7	49
Акцепторы	Индий (In)	11,2	160
	Галлий (Ga)	10,8	65
	Бор (B)	10,4	45
	Алюминий (Al)	10,2	57

$$\left.\begin{aligned} n_e &= 2\frac{\left(2\pi m_e^* kT\right)^{3/2}}{h^3}\exp\left(-\frac{E_g - E_F}{kT}\right) = N_c e^{-\left(E_g - E_F\right)/kT}, \\ n_h &= 2\frac{\left(2\pi m_h^* kT\right)^{3/2}}{h^3}\exp\left(-\frac{E_F}{kT}\right) = N_\upsilon e^{-E_F/kT}. \end{aligned}\right\} \quad (4.14)$$

Вы сомневаетесь? А разве при выводе этих формул что-нибудь говорилось о том, есть в материале примеси или их нет? Говорилось только о том, что уровень Ферми должен располагаться далеко от разрешенных зон. И пока это так, формулы (4.14) справедливы. А где же в формулах (4.14) «спрятаны» примеси? Конечно, в величине E_F! Поэтому все, что нам остается сделать дальше, − это найти положение уровня Ферми и проанализировать его поведение при разных концентрациях примесей и разных температурах. Это удобнее всего сделать, используя закон сохранения заряда. Считая, что примеси могут быть только однократно ионизованными, запишем

$$n_e + n_A^- = n_h + n_D^+, \quad (4.15)$$

где n_A^- и n_D^+ − концентрации ионизованных акцепторов и доноров соответственно. Если полные концентрации акцепторов и доноров соответственно n_A и n_D, то без всяких комментариев можно записать:

$$\left.\begin{aligned} n_A^- &= n_A f\left(E,T\right) = n_A \frac{1}{e^{\left(E_A - E_F\right)/kT} + 1}, \\ n_D^- &= n_D\left[1 - f\left(E,T\right)\right] = n_D \frac{1}{e^{\left(E_F - E_D\right)/kT} + 1}. \end{aligned}\right\} \quad (4.16)$$

Теперь, найдя из формул (4.14) − (4.15) и табл. 4.1 и 4.2 все, что нам нужно, мы можем рассчитать проводимость примесного полупроводника. Однако необходима осторожность! (Вспомните, что мы только что говорили о применимости формул (4.14)). Давайте все-таки разберемся с нашими результатами подробнее. Прежде всего, полупроводник будет, скорее всего, слабо компенсированным (пусть для определенности, *n*-типа). Тогда в (4.15) явно можно пренебречь слагаемыми n_A^- и n_h, оставив только

$$n_e = n_D^+$$

или

$$N_C \exp\left(-\frac{E_g - E_F}{kT}\right) = \frac{n_D}{e^{(E_F - E_D)/kT} + 1}.$$

Пока примеси очень мало, уровень Ферми все еще где-то около середины запрещенной зоны, то есть (см. рис. 4.2)

$$\left.\left|\frac{E_F - E_D}{kT}\right| >> 1,\ \ \frac{E_F - E_D}{kT} < 0,\ \text{при}\ n_D^+ \cong n_D\right\}.$$

Тогда, обозначив $N_C \exp(-E_g/kT) = C$ (это константа), получим:

$$C \exp\left(\frac{E_F}{kT}\right) \cong n_D,$$

откуда

$$E_F \cong kT \ln\left(\frac{n_D}{C}\right), \tag{4.17}$$

то есть энергия Ферми довольно слабо (логарифмически) растет с увеличением концентрации доноров. Посмотрим теперь, что происходит с изменением температуры. Ситуация изображена на рис. 4.3 и качественно понятна.

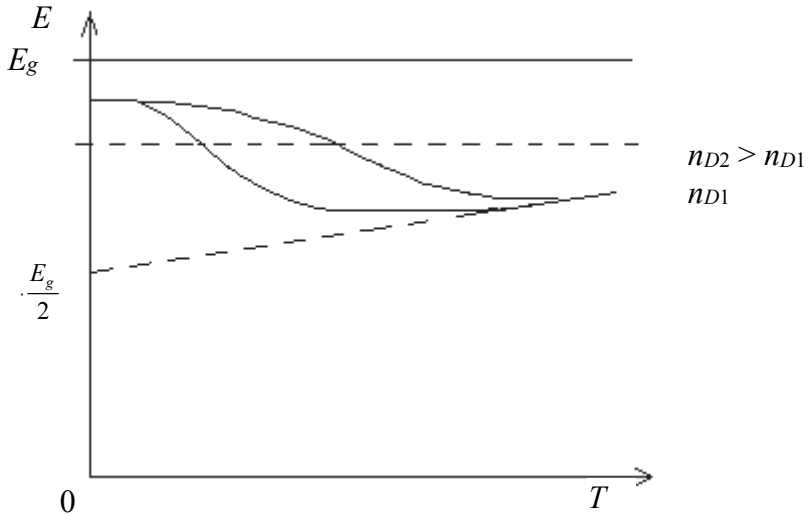


Рис. 4.3. Температурная зависимость положения уровня Ферми в примесном полупроводнике

При нулевой температуре уровень Ферми расположен точно посередине между донорными уровнями и дном зоны проводимости. При очень высокой температуре все примеси ионизованы и больше не

могут увеличивать концентрацию носителей, резервуар же валентных электронов гораздо больше и по сравнению с примесями неисчерпаем, поэтому полупроводник становится практически собственным. Ясно, что чем больше концентрация примеси, тем при большей температуре это произойдет. Штриховая линия, начинающаяся от значения $E_g/2$ при $T = 0$, отражает поведение уровня Ферми в собственном полупроводнике (вспомните предыдущий раздел). Аналогичный рисунок для полупроводника *p*-типа Вы можете сделать сами в качестве упражнения.

Таким образом, нашими формулами при очень низких температурах явно пользоваться нельзя, при очень больших (правда, нереальных) − тоже. Значит, нам всегда нужно внимательно выяснить применимость теории, и если она применима − нет проблем! Тогда нам не нужно больше ничего, кроме табл. 4.1 и 4.2, для расчета проводимости полупроводника.

4.3. Рассеяние носителей заряда

До сих пор мы рассматривали только концентрацию носителей заряда в полупроводнике. Однако удельная проводимость материала σ зависит не только от концентрации носителей n, но и от их подвижности μ. Для подвижности мы располагаем пока только классическим соотношением:

$$\mu = \frac{e\tau}{m}.$$

Ясно, что с учетом наших новых знаний мы должны заменить массу носителя m на его эффективную массу m^*. А что делать со временем релаксации τ?

Обычно вопрос о времени релаксации решается с использованием кинетического уравнения Больцмана, с чем при желании вы можете ознакомиться по литературе из рекомендуемого списка. Однако необходимо отметить, что никому пока не удалось рассчитать τ из первых принципов, не обращаясь к экспериментальным данным. Поэтому мы проанализируем этот вопрос только качественно, выяснив наиболее важный с практической точки зрения аспект − температурную зависимость подвижности.

Вспомним, что причиной рассеяния являются не соударения носителей заряда с атомами в узлах кристаллической решетки, а взаимодействие носителей с *дефектами*: тепловыми колебаниями атомов

решетки (фононами), нейтральными и ионизованными примесями, дислокациями, границами зерен в поликристалле, таким страшным дефектом кристалла, как его поверхность и т. д.

Рассматривая только процессы, происходящие вдали от поверхности тщательно выращенного монокристалла, мы можем ограничиться случаями рассеяния носителей на точечных дефектах − нейтральных и ионизованных (заряженных).

Прежде всего, вспомним, что время релаксации связано с длиной свободного пробега Λ, тепловой скоростью υ_T и дрейфовой скоростью υ_D соотношением

$$\tau = \frac{\Lambda}{\upsilon_D + \upsilon_T} \cong \frac{\Lambda}{\upsilon_T},$$

причем $\upsilon_T \sim T^{1/2}$. Тогда

$$\tau \sim \Lambda T^{-1/2}. \qquad (4.18)$$

Разумно предположить, что длина свободного пробега обратно пропорциональна *площади эффективного сечения S* рассеивающего центра (как в молекулярно-кинетической теории газов).

В случае нейтральных дефектов радиусом площади эффективного сечения можно считать амплитуду тепловых колебаний A, пропорциональную квадратному корню из энергии тепловых колебаний E_T, которая, в свою очередь, пропорциональна температуре T. Таким образом,

$$\Lambda \sim 1/S \sim 1/A^2 \sim 1/E \sim T^{-1},$$

и время релаксации τ_0 при рассеянии на нейтральных точечных дефектах, т. е. на фононах и незаряженных примесных атомах, в соответствии с формулой (4.18)

$$\tau_0 \sim T^{-3/2}, \qquad (4.19)$$

вследствие чего подвижность носителей при таком механизме рассеяния должна убывать с температурой по такому же закону.

В случае заряженных точечных дефектов, которыми будут являться ионизованные примесные атомы (а какие, интересно, конкретно для электронов и какие для дырок?), величина эффективного сечения рассеяния будет определяться электрическим полем дефекта и температурой материала. Действительно, столкновение носителя с ионом произойдет только тогда, когда тепловая энергия носителя (т. е. его кинетическая энергия) сравняется с потенциальной энергией отталкивания между носителем заряда и ионом:

$$\frac{Ze^2}{4\pi\varepsilon_0\chi r} = \frac{3}{2}kT,$$

откуда площадь эффективного сечения рассеяния

$$S \sim \tau^2 \sim 1/T^2,$$

соответственно длина свободного пробега

$$\Lambda \sim 1/S \sim T^2,$$

и время релаксации τ_u при рассеянии на ионизованных дефектах в соответствии с (4.18)

$$\tau_u \sim T^{3/2} \tag{4.20}$$

При наличии обоих механизмов рассеяния результирующее время релаксации следует находить по правилу:

$$\frac{1}{\tau} = \frac{1}{\tau_0} + \frac{1}{\tau_u}. \tag{4.21}$$

(Вам ничего не напоминает эта формула?).

Поскольку все вышесказанное одинаково справедливо как для электронов, так и для дырок, в случае амбиполярной проводимости

$$\sigma = \frac{e^2\tau_e n_e}{m_e^*} + \frac{e^2\tau_h n_h}{m_h^*}. \tag{4.22}$$

Обратите внимание, что подвижность зависит от температуры степенным образом, в то время как концентрация носителей в полупроводниках − экспоненциальным. Поэтому в полупроводниках (в отличие от металлов) температурная зависимость проводимости определяется, как правило, температурной зависимостью концентрации носителей. Из этого правила, естественно, есть исключения. Например, при температурах, достаточно высоких для истощения примесных уровней, но слишком низких для эффективной тепловой генерации собственных носителей, температурная зависимость проводимости обусловлена в основном температурной зависимостью подвижности. В этой области температур проводимость полупроводника может убывать при нагревании.

4.4. Рекомбинация

Обратимся снова к уравнениям (4.7) и (4.8), задающим концентрации электронов в зоне проводимости и дырок в валентной зоне полупроводника вне зависимости от того, собственный он или примесный.

Если перемножить эти концентрации, то из (4.7) и (4.8) получим:

$$n_e n_h = N_c N_v \exp\left(-\frac{E_g}{kT}\right) = 4\left(\frac{2\pi kT}{h^2}\right)^3 \left(m_e^* m_h^*\right)^{3/2} \exp\left(-\frac{E_g}{kT}\right). \qquad (4.23)$$

В этом выражении крайне интересно отсутствие уровня Ферми E_F. В собственном полупроводнике $n_e = n_h = n_i$ (буква i означает английское слово «*intrinsic*» − собственный). Тогда из (4.23) следует, что для данного конкретного полупроводника, будь он собственный или примесный, выполняется соотношение

$$n_e n_h = n_i^2. \qquad (4.24)$$

Допустим, мы вводим в собственный полупроводник донорную примесь. При этом концентрация электронов в зоне проводимости возрастет, и это естественно. Но концентрация дырок в валентной зоне тогда в соответствии с (4.24) должна уменьшиться. Не странно ли?

Не странно, если внимательно разобраться с понятием теплового равновесия электронов и дырок в полупроводнике. Не забывайте, что понятие температуры имеет строгий смысл, только если вещество действительно находится в равновесии с окружающими телами. Если мы хотим описать процесс электропроводности, то мы можем только приблизительно пользоваться равновесными функциями распределения с большой степенью точности.

Когда термодинамическое равновесие в полупроводнике существует, именно температура определяет количество и спектры электронов, дырок, фононов и фотонов внутри твердого тела. Фононы и фотоны подчиняются закону распределения Бозе − Эйнштейна, а электроны − закону распределения Ферми − Дирака.

В каждый момент времени происходит поглощение фотонов теплового излучения или фононов, рожденных колебаниями решетки, приводящее к возбуждению электронов в состояния с более высокой энергией. Например, электроны могут переходить в зону проводимости из валентной зоны или с примесных уровней − осуществляется *тепловая генерация* носителей. В то же время каждый электрон может перейти в незанятые состояния с более низкой энергией. Например, электрон из зоны проводимости может перейти на пустые уровни в валентной зоне или на примесях − осуществляется процесс *рекомбинации* носителей, обратный процессу генерации. В состоянии термодинамического равновесия эти противоположные процессы должны совпадать по скорости − количество генерированных в единицу вре-

мени носителей должно совпадать с количеством рекомбинировавших – этого требуют законы термодинамики. Таким образом, тепловое равновесие является *динамическим* (очень похоже на равновесие насыщенного пара над поверхностью жидкости, не правда ли?).

С ростом концентрации примеси растет концентрация носителей одного типа, например, электронов в зоне проводимости в случае донорной примеси. Однако тогда будет расти и вероятность «встречи» электрона с дыркой, т. е. вероятность рекомбинации, и концентрация дырок будет уменьшаться, что и объясняет соотношение (4.24).

Таким образом, если примесные добавки приводят к увеличению n_e, они должны одновременно *во столько же раз* (а не на ту же величину) уменьшить n_h и наоборот. При рассмотрении свойств полупроводников и полупроводниковых приборов полезно знать, что смещение уровня Ферми на величину kT приводит к увеличению концентрации носителей одного типа в e = 2,718 раз и уменьшению концентрации носителей другого типа во столько же раз.

Отметим, что соотношение (4.24) – частный случай закона действующих масс, известного из термодинамики и широко применяемого в химии.

5. ПОЛУПРОВОДНИКОВЫЕ СТРУКТУРЫ

5.1. Контакт полупроводников. *p-n*-переход

Первый значительный прогресс в полупроводниковой электронике связан с использованием контакта двух примесных полупроводников с различным типом проводимости. Такой контакт называют *электронно-дырочным* переходом, или *p-n-переходом*.

Если привести в контакт полупроводники *p*- и *n*-типа, то разность концентраций электронов и дырок в этих материалах приведет к возникновению диффузионных потоков − дырки начнут диффундировать из полупроводника *p*-типа в полупроводник *n*-типа, а электроны в противоположном направлении. Такой процесс диффузии сопровождается переносом не только массы, но и заряда. Перенесенные заряды создают электрическое поле, препятствующее диффузии, так что между полупроводниками возникает разность потенциалов U_k, которая называется *контактной*. При установлении равновесия, как это следует из общих термодинамических принципов, должны уравняться химические потенциалы, т. е. уровни Ферми этих полупроводников, что и отражено на рис. 5.1. Область толщиной *d* в окрестности *p-n*-перехода содержит мало носителей заряда (подумайте сами, почему!), поэтому называется *обедненной областью*. В то же время она содержит неподвижные заряды − отрицательно заряженные акцепторы слева от контакта и положительно заряженные доноры справа от контакта, напоминая заряженный конденсатор. Действительно, *p-n*-переход обладает емкостью, которая называется *барьерной* и будет рассмотрена чуть позже. А пока займемся вычислением контактной разности потенциалов, чтобы затем на основе этого расчета найти толщину обедненного слоя.

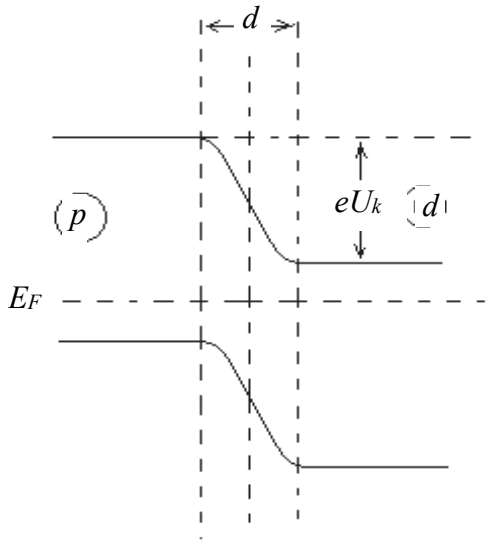


Рис. 5.1. Энергетическая диаграмма равновесного состояния *p-n*-перехода

Пусть толщина обедненного слоя в *p*-области равна d_p, а в *n*-области соответственно d_n, т. е.

$$d = d_p + d_n. \tag{5.1}$$

Поскольку исходные материалы были электронейтральными и за пределами обедненной области они таковыми и остались, из закона сохранения заряда следует, что отрицательный заряд p-области обедненного слоя равен положительному заряду n-области обедненного слоя. Если площадь контакта равна S, а концентрации акцепторов и доноров в соответствующих полупроводниках равны n_A и n_D, то, полагая все примеси ионизованными, можно записать:

$$en_A\, Sd_p = en_D\, Sd_n,$$

откуда следует выражение

$$n_A\, d_p = n_D\, d_n, \tag{5.2}$$

связывающее толщины обедненных слоев с концентрацией примесей.

В соответствии с уравнениями Максвелла напряженность электрического поля $\vec{\varepsilon}$ подчиняется соотношению

$$\nabla \cdot \vec{\varepsilon} = \frac{\rho}{\chi\varepsilon_0}, \tag{5.3}$$

где ρ − объемная плотность заряда, χ − относительная диэлектрическая проницаемость среды (в данном случае полупроводника), $\varepsilon_0 = 8{,}86\cdot10^{-12}$ Ф/м − электрическая постоянная.

В рассматриваемом случае электрическое поле меняется только вдоль оси x, перпендикулярной контакту полупроводников, т. е. (5.3) сводится к виду

$$\frac{\partial\varepsilon}{\partial x} = \frac{\rho}{\chi\varepsilon_0}. \tag{5.4}$$

Тогда, выбирая ноль на оси x в месте контакта, для напряженности поля в p-области получим

$$\varepsilon_p = \int_{-d_p}^{x}\left(\frac{\rho}{\chi\varepsilon_0}\right)dx = \int_{-d_p}^{x}\left(-\frac{en_A}{\chi\varepsilon_0}\right)dx = -\frac{en_A}{\chi\varepsilon_0}\left(x + d_p\right), \tag{5.5}$$

т. е. напряженность поля меняется по линейному закону от нуля на границе обедненной области (при $x = -d_p$) до значения $(-en_Ad_p/\chi\varepsilon_0)$ в месте контакта (при $x = 0$).

Поскольку напряженность поля связана с потенциалом φ соотношением

$$\vec{\varepsilon} = -\nabla\varphi, \tag{5.6}$$

а в одномерном случае $|\nabla\varphi| = \partial\varphi/\partial x$, распределение потенциала в обедненной p-области можно вычислить из соотношения

$$\& = -\frac{\partial\varphi}{\partial x}. \qquad (5.7)$$

Если через φ_1 обозначить потенциал на границе обедненного p-слоя (при $x = -d_p$), а через φ_0 − потенциал в месте контакта полупроводников (при $x = 0$), то из (5.7) получим:

$$\varphi_0 - \varphi_1 = \int_{\varphi_1}^{\varphi_0} d\varphi = \int_{-d_p}^{0} \frac{en_A}{\chi\varepsilon_0}\left(x + d_p\right)dx = \frac{en_A}{2\chi\varepsilon_0} d_p^2. \qquad (5.8)$$

Аналогично для n-области получим:

$$\&_n = \int_{d_n}^{x} \left(\frac{en_D}{\chi\varepsilon_0}\right)dx = \frac{en_D}{\chi\varepsilon_0}\left(x - d_n\right), \qquad (5.9)$$

причем, как показывают соотношения (5.2), (5.5) и (5.9), при $x = 0$ $\varepsilon_n = \varepsilon_p$, как и следовало ожидать. Если потенциал в точке $x = d_n$ обозначить φ_2, то из (5.7) и (5.9) получим:

$$\varphi_2 - \varphi_0 = \int_{\varphi_0}^{\varphi_2} d\varphi = -\int_{0}^{d_n} \frac{en_D}{\chi\varepsilon_0}\left(x - d_n\right)dx = \frac{en_D}{2\chi\varepsilon_0} d_n^2. \qquad (5.10)$$

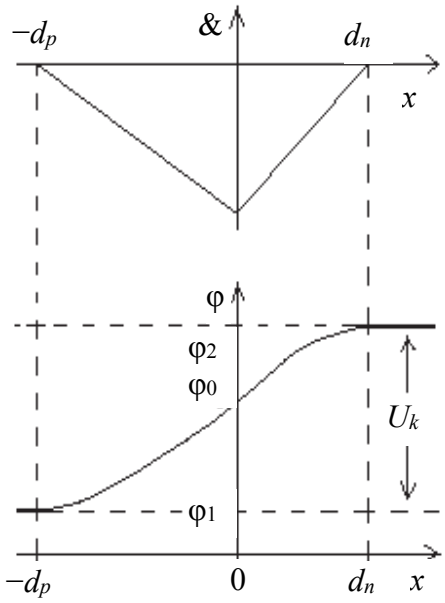


Рис. 5.2. Распределение напряженности поля и потенциала в области p-n-перехода

Тогда контактная разность потенциалов (см. рис. 5.2)

$$U_k = \varphi_2 - \varphi_1 = \frac{e}{2\chi\varepsilon_0}\left(n_A d_p^2 + n_D d_n^2\right). \qquad (5.11)$$

Используя (5.2), выражение (5.11) можно свести к виду (проделайте необходимые вычисления сами!)

$$U_k = \frac{e}{2\chi\varepsilon_0} \cdot \frac{n_A n_D}{\left(n_A + n_D\right)} d^2, \qquad (5.12)$$

откуда

$$d = \left[\frac{2\chi\varepsilon_0 U_k}{e} \cdot \frac{\left(n_A + n_D\right)}{n_A n_D}\right]^{1/2}. \qquad (5.13)$$

Таким образом, толщина обедненного слоя тем больше, чем ниже кон-

центрация примесей, причем из (5.2) следует, что глубина проникновения электрического контактного поля больше в тот полупроводник, концентрация примесей в котором меньше.

Приведенный расчет справедлив для резкого *p*-*n*-перехода, в котором концентрация примесей на границе между *p*- и *n*-полупроводниками меняется практически скачкообразно. Если же изменение концентрации примесей *n* в переходе происходит плавно, так, что зависимость этой концентрации от координаты *x* можно описать линейным законом:

$$n(x) = ax$$

с коэффициентом пропорциональности *a*, то расчеты по изложенной выше методике приведут к результату:

$$d = \left(\frac{12 U_k \chi \varepsilon_0}{ea} \right)^{1/3}. \qquad (5.14)$$

Попробуйте получить этот результат сами, используя, возможно, еще какое-нибудь упрощающее предположение относительно концентраций примесей в *p*- и *n*-полупроводниках.

В заключение отметим, что типичная толщина обедненного слоя в *p*-*n*-переходе составляет доли микрона, а типичное значение контактной разности потенциалов – около 0,3 В для германия и около 0,5 В для кремния.

5.2. Полупроводниковый диод

Фактически конструкция из двух полупроводников с различным типом проводимости, образующих *p*-*n*-переход, представляет собой готовый полупроводниковый прибор − *диод*. Чтобы разобраться с тем, как он работает, рассмотрим текущий через *p*-*n*-переход электронный ток (для дырочного тока получится все то же самое, в чем легко убедиться, перевернув рис. 5.1 вверх ногами). Если рассмотреть внимательно рис. 5.1, то бросается в глаза, что слева от *p*-*n*-перехода (в *p*-области) электронов проводимости очень мало, а справа (в *n*-области) очень много. Зато переходить электроны слева направо могут совершенно беспрепятственно, а для переходов справа налево электронам необходимо «запрыгивать» со дна зоны проводимости *n*-области на дно зоны проводимости *p*-области, то есть преодолевать потенциальный барьер высотой eU_k. Ток, текущий слева направо, обозначим просто I_0, а через $I_1 - \exp(-eU_k/kT)$ обозначим ток, текущий справа налево, учитывая за-

даваемую распределением Больцмана вероятность преодоления электронами потенциального барьера в контакте. В отсутствие внешнего электрического поля суммарного тока через p-n-переход нет, то есть

$$I_0 = I_1 \exp\left(-\frac{eU_k}{kT}\right). \qquad (5.15)$$

Приложим теперь к *p-n*-переходу напряжение U плюсом к *p*-области, а минусом − к *n*-области. Такое напряжение, называемое *прямым напряжением*, понизит потенциальный барьер на величину eU, и текущий справа налево ток станет равен $I_1\exp[-e(U_k - U)/kT]$, то есть увеличится. Ток же, текущий слева направо, не изменится и останется равным I_0 . В результате в *p-n*-переходе возникнет суммарный ток:

$$I = I_1 \exp\left[-\frac{e(U_k - U)}{kT}\right] - I_0. \qquad (5.16)$$

Используя (5.15), соотношение (5.16) можно привести к виду

$$I = I_0\left[\exp\left(\frac{eU}{kT}\right) - 1\right], \qquad (5.17)$$

где величину I_0 называют обратным током диода.

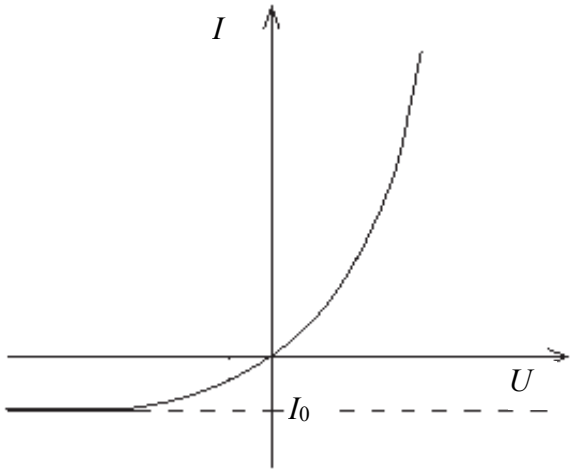


Рис. 5.3. Теоретическая вольтамперная характеристика *p-n*-перехода, описываемая формулой (5.17)

Если к *p-n*-переходу приложить обратное напряжение плюсом к *n*-области, а минусом к *p*-области, то, думаю, вы догадываетесь сами, что получится снова соотношение (5.17), только в нем будет $U < 0$ и, соответственно, $I < 0$.

Выражение (5.17) называется *уравнением Шокли*, или просто *уравнением выпрямителя*, и описывает вольтамперную характеристику полупроводникового диода (см. рис. 5.3).

Прежде всего следует осознать, что природа тока, текущего через *p-n*-переход, принципиально иная, чем у тока в однородном полупроводнике, где электрическое поле просто заставляет электроны двигаться в каком-то направлении. В *p-n*-переходе же при прямом смещении в *p*-область хлынет поток

электронов, для которых резко снизился «стерегущий» их потенциальный барьер eU_k. p-область заполняется электронами, которые здесь являются чужаками − неосновными носителями. Точно так же n-область заполняется инжектированными из p-области дырками. В итоге оказывается, что устройство с p-n-переходом, которое называют полупроводниковым диодом, работает на инжектированных неосновных носителях, а электрическое поле, в отличие от однородного материала, в диоде не только упорядочивает движение носителей, но и (самое главное!) генерирует их за счет инжекции в тех областях, где в отсутствие поля таких носителей почти не было. Безусловно, попавшие «за границу» неосновные носители, например, электроны в p-области, далеко от перехода не уйдут − они исчезнут, рекомбинировав с основными носителями, но, впрочем, проходя до гибели расстояния в 1 мм, т. е. в тысячу раз больше, чем толщина переходной области. Таким образом, ток в каждой из областей диода является двухкомпонентным − представляет собой сумму токов основных и инжектированных неосновных носителей – и неоднородным (см. рис. 5.4): ток основных носителей из-за рекомбинации уменьшается по мере приближения к p-n-переходу, а ток инжектированных неосновных носителей по той же причине убывает по мере удаления от p-n-перехода.

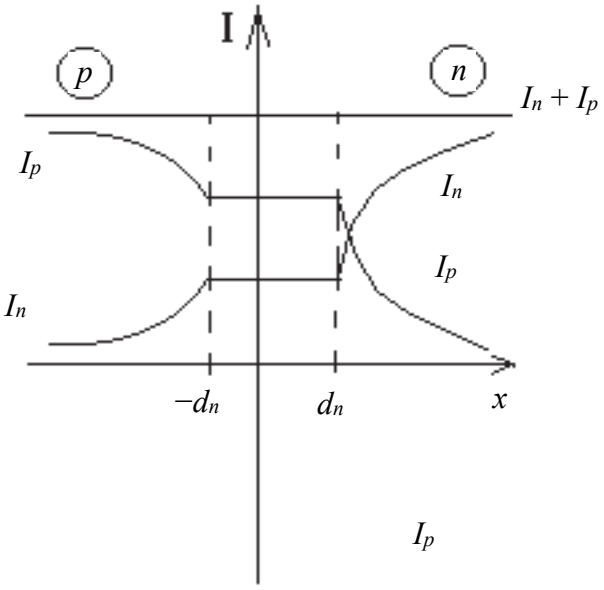


Рис. 5.4. Компоненты полного тока в диоде

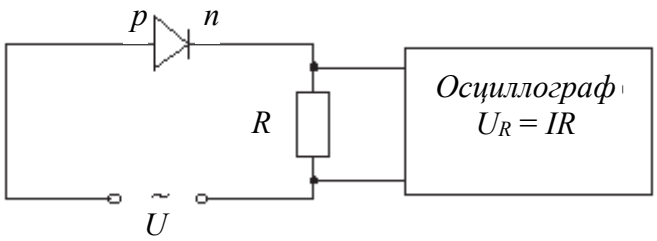


Рис. 5.5. Принципиальная схема однополупериодного выпрямителя

Как и во всех других случаях, когда внешнее поле влияет не только на перенос, но и на генерацию носителей (типичный пример − вакуумный диод), ток в полупроводниковом диоде не подчиняется закону Ома, а подчиняется придуманному Шокли уравнению выпрямителя.

Графически это выражается в том (см. рис. 5.3), что вольтамперная характеристика, в отличие от линейной омической, не представляет собой прямой линии, т. е. является нелинейной. Соответственно и элементы электронных схем, не подчиняющиеся закону Ома, называют нелинейными.

Именно нелинейные элементы позволяют реализовывать фантастическое многообразие электронных устройств − от элементарной (хотя и не совсем уж простенькой) бытовой аппаратуры до сложнейшего научного и технологического оборудования. Простейшее из таких устройств − знакомый Вам однополупериодный выпрямитель на одном-единственном полупроводниковом диоде, изображенный на рис. 5.5. Осциллограф в этой схеме будет измерять пропорциональное току в цепи падение напряжения на резисторе R, равное $U_R = IR$, и на экране мы увидим фактически зависимость тока в цепи от времени (рис. 5.6).

В те полупериоды, когда напряжение U от источника создает прямое смещение на диоде, ток в цепи (прямой ток диода) большой. В полупериоды, создающие обратное смещение диода, ток в цепи (обратный ток диода) очень маленький − на рис. 5.6 этот ток сильно преувеличен, в чем нетрудно убедиться, подставив числа в уравнение выпрямителя (не говоря уже о том, что реально обратный ток еще раз в сто меньше, чем это следует из уравнения выпрямителя, из-за процессов рекомбинации в обедненной области).

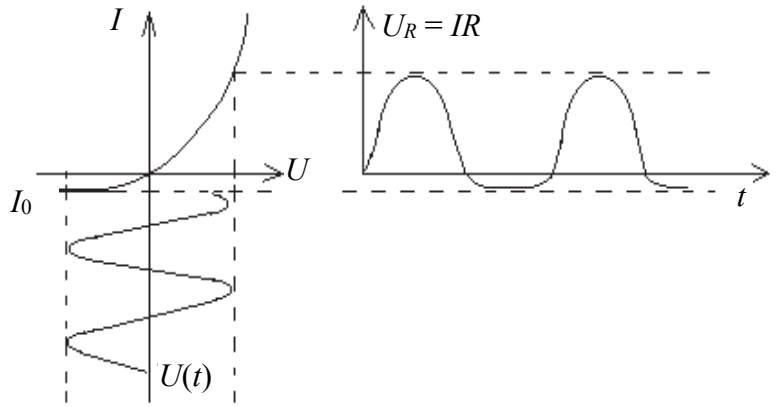


Рис. 5.6. Графики напряжений на входе и выходе однополупериодного выпрямителя

В итоге мы вместо переменного тока имеем ток, текущий в одном направлении, то есть постоянный, правда, пульсирующий. Частоту этих пульсаций увеличивают вдвое, используя вместо одного диода четыре (диодный мостик), затем эти пульсации сглаживают фильтром, состоящим, например, из резистора и электролитического конденсатора большой емкости. А затем... затем чуть позже, когда мы узнаем, что такое стабилитрон.

5.3. Биполярный транзистор

23 декабря 1947 года в Bell Telephone Laboratories заработал первый транзистор. Это событие в итоге оказало даже большее влияние на повседневную жизнь, чем создание атомной бомбы. Малая энергоемкость в расчете на один бит информации, сравнимая с энергопотреблением нейронов головного мозга, и исключительная долговечность транзисторов привела к революции в области электронных средств связи и к созданию быстродействующих ЭВМ с большим объемом памяти. Поэтому неудивительно, что создатели транзистора − Джон Бардин, Уолтер Браттейн и Уильям Шокли − были удостоены Нобелевской премии по физике в 1956 г. Собственно, Нобелевская премия была присуждена им даже не за изобретение транзистора как таковое, а за реализацию целой исследовательской программы. Как говорил Бардин в своей Нобелевской лекции, «общая цель программы состояла в том, чтобы, как можно глубже разобраться в явлениях, наблюдаемых в полупроводниках, причем не эмпирически, а объяснить их на основе атомной теории». Но, хотя такая цель и не ставилась, в виду постоянно имелась возможность создания полупроводникового триода, или усилителя, что и было блестяще реализовано.

Попробуем разобраться с принципом работы транзистора на примере самой популярной в учебниках (но не на практике) схемы включения − схемы с общей базой (рис. 5.7).

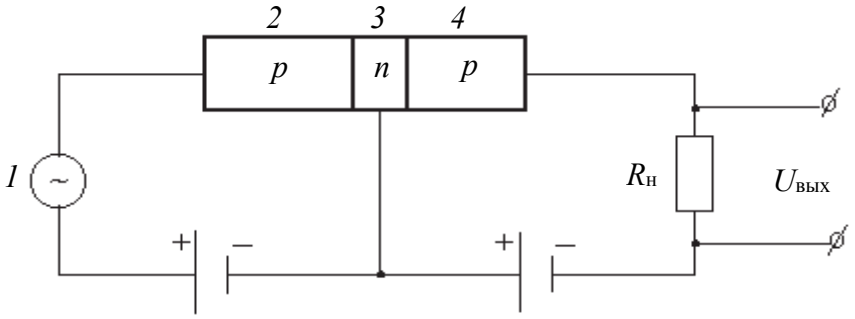


Рис. 5.7. Транзистор типа *p-n-p*, включенный по схеме с общей базой:
1 − источник сигнала; *2* − эмиттер; *3* − база; *4* − коллектор;
$R_н$ − сопротивление нагрузки в коллекторной цепи;
$U_{вых}$ − выходное напряжение схемы

Биполярный транзистор состоит из трех областей, называемых эмиттером, базой и коллектором и образующих два *p-n*-перехода, которые обычно именуют эмиттерным и коллекторным. На рис. 5.7 приведен *p-n-p*-транзистор, хотя с равным основанием будет работать и

n-p-n-транзистор − не забудьте только поменять полярность источников напряжения.

Эмиттерный переход включен в прямом направлении, и поток дырок из эмиттера беспрепятственно хлынет в базу. Коллекторный переход включен в обратном направлении.., но ведь для пришедших из эмиттера дырок он не помеха! Надо только, чтобы толщина базы была существенно меньше диффузионной длины дырок, составляющей величину около 1 мм. И тогда практически все пришедшие из эмиттера дырки попадут в коллектор и будут благополучно участвовать в протекании коллекторного тока. Отношение α тока коллектора $I_к$ к току эмиттера $I_э$

$$\alpha = I_к/I_э \qquad (5.18)$$

называется коэффициентом передачи тока эмиттера. Этот коэффициент близок к единице и составляет величину 0,98–0,99 и больше.

А теперь задумаемся, что же мы получили? Вроде бы ничего ценного. Почти весь ток от источника *1* на рис. 5.7 поступает в нагрузку $R_н$, и коэффициент усиления тока в нашем устройстве чуть меньше единицы (это и есть величина α). Так что здесь усиливается?

Оказывается, схема с общей базой усиливает напряжение. Действительно, эмиттерный переход включен в прямом направлении, поэтому имеет очень малое сопротивление. Значит, даже малое напряжение от источника сигнала *1* создаст во входной (эмиттерной) цепи заметный ток. Почти весь этот ток передается в выходную цепь как коллекторный ток. Но ведь коллекторный переход включен в обратном направлении, и его сопротивление очень велико. Значит, и нагрузочный резистор $R_н$ может иметь достаточно большое сопротивление. (Уяснили? Если нет, вспомните закон Ома для замкнутой цепи.) Протекая через высокоомный резистор $R_н$, ток коллектора создает на нем выходное напряжение, существенно превышающее напряжение источника сигнала *1*. Таким образом, транзистор, включенный по схеме с общей базой, представляет собой усилитель напряжения. А поскольку входной и выходной токи почти равны друг другу, такая схема является и усилителем мощности.

Итак, вся прелесть транзистора заключается в том, что он имеет низкое входное и высокое выходное сопротивления, обеспечиваемые разным включением эмиттерного и коллекторного переходов. Теперь мы можем с пониманием расшифровать название «транзистор». Это сокращение английских слов «*transfer resistor*» − преобразование сопротивления.

Если вы заглянете внутрь любого бытового электронного устройства, то увидите, что транзисторы в усилительных каскадах соединены не так, как на рис. 5.7, а так, как на рис. 5.8 − по схеме с общим эмиттером. Для разнообразия на рис. 5.8 изображен транзистор типа *n-p-n*. Это, кстати, более распространенный тип кремниевых транзисторов (не забудьте задать вопрос «Почему?», когда будете изучать курс технологии!), и наиболее распространенный маломощный транзистор КТ-315 относится именно к этому типу. Кроме того, на рис. 5.8 приведено и схемное обозначение транзистора. Обратите внимание на направление стрелочки, которой снабжен эмиттерный вывод − это направление указывает тип транзистора.

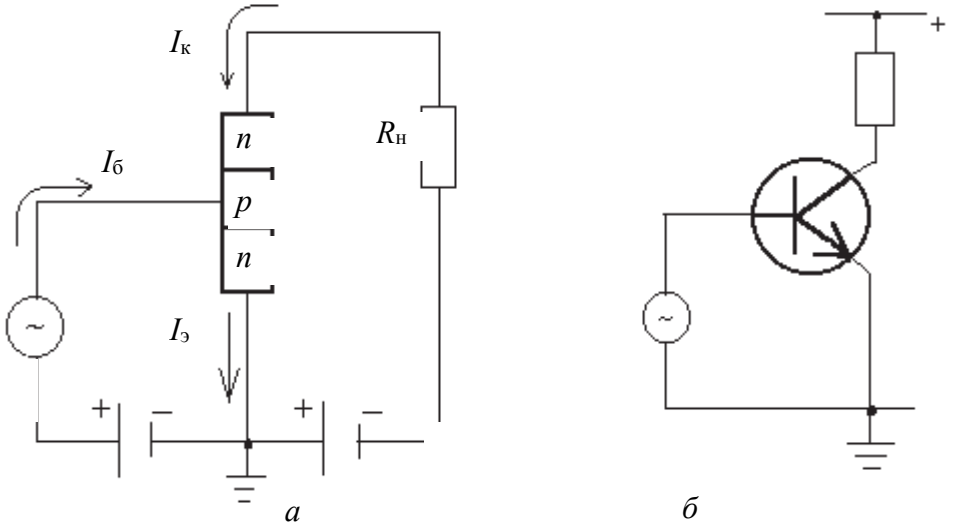


Рис. 5.8. Транзистор, включенный по схеме с общим эмиттером:
а − условная схема; *б* − принципиальная схема

В схеме с общим эмиттером по-прежнему большая часть тока поступает в коллектор через эмиттерный переход, включенный в прямом направлении, и выражение (5.18) остается справедливым. В то же время (см. рис. 5.8)

$$I_э = I_б + I_к. \qquad (5.19)$$

Исключая из формул (5.18) и (5.19) ток эмиттера $I_э$, получим соотношение, связывающее токи коллектора и базы:

$$I_к = \frac{\alpha I_б}{1-\alpha} = \beta I_б, \qquad (5.20)$$

где величина

$$\beta = \frac{\alpha}{1-\alpha} \qquad (5.21)$$

называется коэффициентом передачи тока базы. Если, например, $\alpha = 0{,}99$, то $\beta \approx 100$, то есть мы получим усиление тока базы в сто раз. Таким образом, транзистор, включенный по схеме с общим эмиттером, представляет собой усилитель тока.

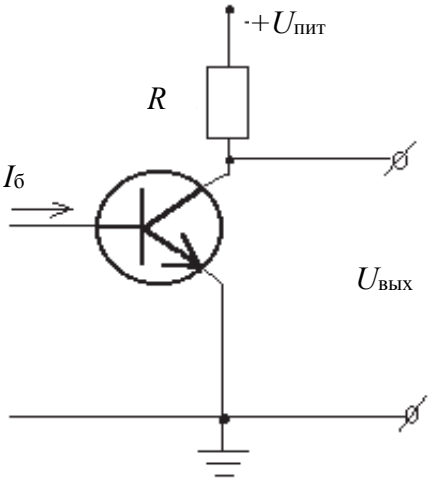


Рис. 5.9. Транзистор как логический элемент

До сих пор мы говорили об использовании транзисторов в качестве усилительных приборов, то есть об аналоговых устройствах. Однако в компьютерах нужны принципиально иные устройства − цифровые, или логические. Эти устройства должны иметь два устойчивых состояния, которые называют логическим нулем и логической единицей, а также должны легко переключаться из одного устойчивого состояния в другое. Транзистор позволяет легко реализовать такое устройство. Рассмотрим схему, приведенную на рис. 5.9 и очень похожую на усилитель тока. Если ток базы $I_б = 0$, то нет и тока коллектора (см. рис. 5.9). Транзистор в этом случае, как говорят, «закрыт», то есть имеет очень большое (в идеале бесконечно большое) сопротивление, во всяком случае, гораздо большее, чем сопротивление резистора R. Тогда практически все напряжение питания $U_{пит}$ падает на транзисторе, и $U_{вых} = U_{пит}$ (допустим, +5 вольт, что по современным стандартам соответствует логической единице). Если же создать некоторый ток базы, появится ток в цепи коллектора. Транзистор «откроется», и его сопротивление станет существенно меньше сопротивления резистора R (конечно, надо разумно выбрать номинал резистора R − где-то около нескольких килоом). Все (конечно, *почти* все) напряжение $U_{пит}$ придется на резистор R, и $U_{вых} = 0$, что соответствует логическому нулю. Вот на таких элементах с двумя устойчивыми состояниями и работают компьютеры.

В завершение этого раздела хочу обратить ваше внимание на следующее обстоятельство. Иногда приходится слышать (к счастью, редко), что транзистор − это полупроводниковый аналог вакуумного триода. Не верьте этому. Принцип работы биполярного транзистора, управляемого током, не имеет ничего общего с принципом работы вакуумного триода, управляемого электрическим полем. Все сходство

между этими устройствами ограничивается тем, что оба они «трехногие», то есть имеют по три электрода, и оба являются усилителями электрической мощности.

В то же время существует большое семейство полупроводниковых приборов, которые по своему принципу действия (по крайней мере, в схемотехнической части) находятся с вакуумными лампами в гораздо более близком родстве. Эти приборы называются *полевыми транзисторами*, и к их изучению мы скоро приступим.

5.4. МДП-структуры

Поговорим еще немного о контактах. Когда мы приводим различные материалы в соприкосновение, возникает множество любопытных эффектов − в этом мы уже убедились на примере *p-n*-перехода. Рассмотрим сейчас специфическую трехслойную структуру, содержащую металл, диэлектрик и полупроводник − МДП-структуру (рис. 5.10). Будем для простоты считать, что до объединения всех трех материалов их уровни Ферми E_F совпадали. Кроме того, для определенности выберем полупроводник *n*-типа.

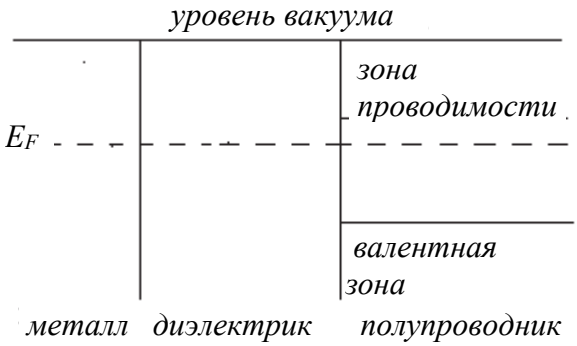


Рис. 5.10. Энергетическая диаграмма МДП-структуры в тепловом равновесии

Приложим к такой структуре напряжение − плюсом на металл, а минусом на полупроводник. В результате энергетическая диаграмма примет вид, изображенный на рис. 5.11, − электрическое поле создаст искривление зон в полупроводнике. В результате уровень Ферми в полупроводнике около контакта с диэлектриком станет еще ближе к дну зоны проводимости и дальше от потолка валентной зоны. Следовательно, вблизи границы раздела диэлектрик-полупроводник концентрация электронов проводимости в полупроводнике увеличится, а концентрация дырок уменьшится. Иначе говоря, в полупроводнике на границе с диэлектриком образуется слой, *обогащенный* основными носителями.

Это явление нетрудно понять и без всяких зонных диаграмм. Положительный потенциал на металле оттолкнет положительные дырки и притянет отрицательные электроны к границе полупроводник-диэлектрик.

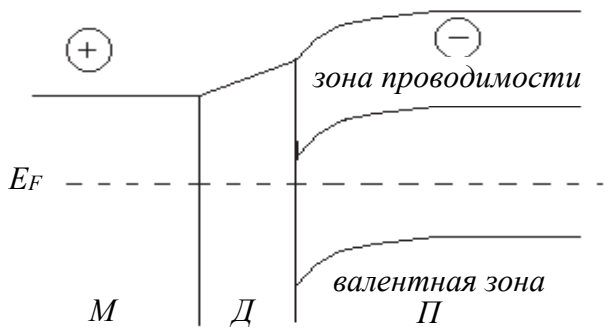


Рис. 5.11. Энергетическая диаграмма МДП-структуры при подаче на металл положительного потенциала

Деваться притянутым электронам некуда − ток через диэлектрик не течет, поэтому в полупроводнике образуется и сохраняется обогащенный слой.

Приложим теперь к металлу «минус», а к полупроводнику «плюс», т. е. обратное напряжение. Все произойдет наоборот − электроны оттолкнутся от границы раздела, а дырки притянутся. В полупроводнике у границы с диэлектриком образуется *обедненный* слой. Чем обедненный? Конечно, основными носителями. Энергетическая диаграмма, иллюстрирующая явление обеднения, приведена на рис. 5.12. Пока все это достаточно тривиально, но посмотрите внимательно на рис. 5.12. По мере увеличения обратного напряжения на МДП-структуре дно зоны проводимости в полупроводнике близ границы с диэлектриком становится все дальше от уровня Ферми, а потолок валентной зоны − все ближе. В один прекрасный момент, точнее, при одном прекрасном напряжении уровень Ферми станет ближе к валентной зоне, чем к зоне проводимости. Однако на языке статистики это означает, что концентрация дырок станет больше концентрации электронов проводимости. Вот это фокус! Полупроводник *n*-типа только за счет внешнего электрического поля стал полупроводником *p*-типа. Такое явление называется *инверсией*.

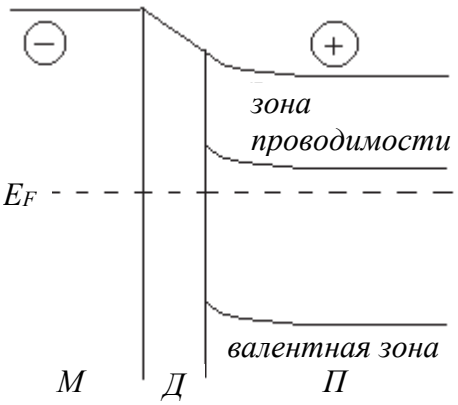


Рис. 5.12. Энергетическая диаграмма МДП-структуры, иллюстрирующая явление обеднения

Толщина слоя, в котором могут происходить обогащение, обеднение или инверсия, рассчитывается аналогично тому, как мы с Вами это делали для толщины обедненного слоя в *p*-*n*-переходе, и имеет типичное значение порядка 1 мкм. Давайте проведем некоторые прикидки. Подвижность электронов в кремнии при комнатной температуре составляет 0,15 м^2/В·с (см. таблицу свойств собственных полупро-

водников). Если мы создадим в МДП-структуре электрическое поле с напряженностью 1000 В/м (это очень слабое поле!), то электроны приобретут дрейфовую скорость 150 м/с и пробегут расстояние в 1 мкм примерно за 7 нс. Время же тепловой генерации носителей может составлять несколько секунд.

Теперь представьте себе, что мы прикладываем к МДП-структуре достаточное для инверсии обратное напряжение импульсами длительностью, допустим, около 1 мс. Примерно за 0,001% этого времени электроны вытолкнутся из приповерхностной области, а дырки за счет тепловой генерации еще не успевают появиться. Практически все время действия импульса в граничном слое находится крайне мало носителей, поэтому такой режим работы МДП-структуры называется *режимом глубокого обеднения*.

Конечно, явления инверсии и глубокого обеднения очень интересны с точки зрения физики, но какая от них может быть польза, если через МДП-структуру не течет ток? Пользу можно извлечь, если пропустить ток вдоль границы раздела полупроводник-диэлектрик! Тогда можно построить наиболее современные полупроводниковые приборы − так называемые МОП-транзисторы и ПЗС − приборы с зарядовой связью. Об этих приборах, отражающих сегодняшний уровень микроэлектроники, и пойдет речь.

5.5. МОП-транзистор

Сокращение МОП означает «металл-окисел (или оксид, если вам так больше нравится) − полупроводник». МОП-транзистор представляет собой МДП-структуру, в которой в качестве диэлектрика выступает слой диоксида кремния, что очень удобно с точки зрения технологии. МОП-транзисторы обычно подразделяют на транзисторы с *индуцированным каналом* и транзисторы со *встроенным каналом*. МОП-транзистор с индуцированным каналом схематически изображен на рис. 5.13. МДП-структура состоит из ме-

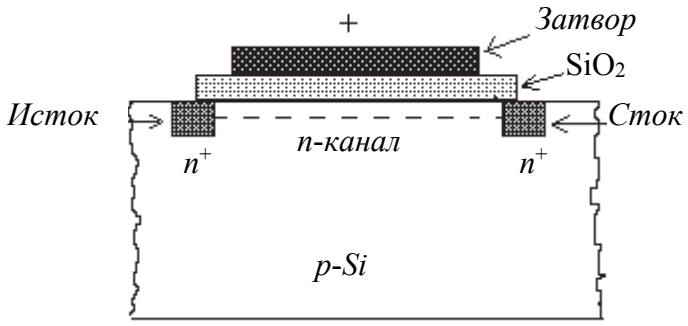


Рис. 5.13. Схематическое изображение МОП-транзистора с индуцированным каналом

таллического затвора, диэлектрического слоя диоксида кремния и полупроводника − кремния, в данном случае, для определенности, *p*-типа. Типичные размеры такого транзистора − единицы микрон в плоскости поверхности полупроводниковой пластины и не более 1 мкм в направлении, перпендикулярном поверхности, толщина оксидного слоя 0,02–0,25 мкм. У границы оксидного слоя в кремнии созданы сильнолегированные области *n*-типа, к которым подсоединяются проводники − исток и сток.

Если между стоком и истоком подключить источник напряжения плюсом к стоку, а на затвор относительно истока напряжение не подавать, то ток по кремнию практически течь не будет, так как в полученной n^+-*p*-n^+ − структуре хотя бы один *p*-*n*-переход всегда будет включен в обратном направлении. Если же на затвор подать положительное относительно истока напряжение $U_{зи}$, то по мере увеличения этого напряжения в полупроводнике образуется вначале обедненный слой (что, естественно, не способствует увеличению тока), а затем при достижении некоторого *порогового напряжения* $U_{пор}$, произойдет (внимание!) инверсия. В полупроводнике у границы раздела с диэлектриком создается инверсный слой, в данном случае *n*-типа, который и является индуцированным (созданным внешним электрическим полем) каналом. По этому каналу охотно потечет ток, и, думаю, вы без труда сообразите сами, что этот ток будет тем больше, чем больше напряжение между затвором и истоком (почему?). Указанная зависимость тока стока изображена на рис. 5.14 и называется переходной, или передаточной, характеристикой МОП-транзистора. Если же *n*-канал в *p*-кремнии мы создадим не электрическим полем, а непосредственно легированием донорными примесями при изготовлении прибора, то получим МОП-транзистор со встроенным каналом. Работать он будет чисто на явлении обеднения. При увеличении отрицательного потенциала на затворе относительно истока концентрация электронов в n-канале уменьшается и, соответственно, уменьшается ток стока. Наконец, при некотором напряжении $U_{отс}$, которое называется напряжением отсечки (знакомый термин, не правда ли?), обедненная область захватит весь *n*-канал,

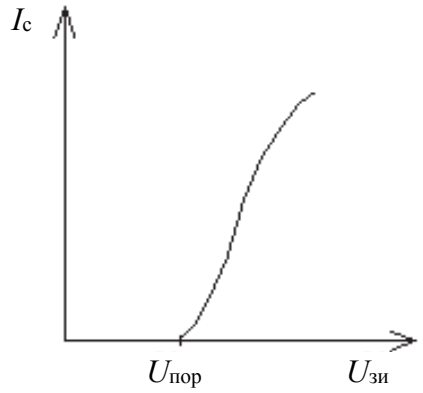


Рис. 5.14. Переходная (стокозатворная) характеристика МОП-транзистора с индуцированным *n*-каналом

и транзистор закроется, т. е. ток стока практически исчезнет. Переходная характеристика МОП-транзистора со встроенным каналом изображена на рис. 5.15 и, я думаю, Вы уже заметили, что она полностью аналогична переходной характеристике полевого транзистора с управляющим *p*-*n*-переходом.

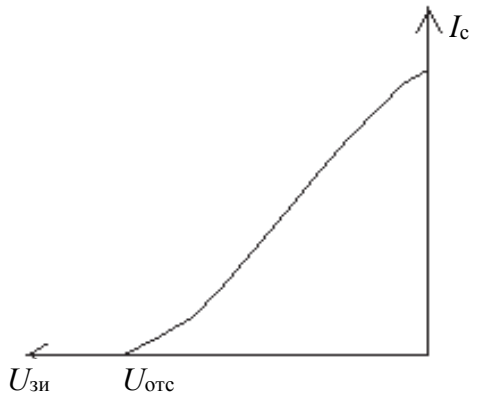


Рис. 5.15. Переходная характеристика МОП-транзистора со встроенным каналом *n*-типа

Канал в МОП-транзисторе отделен от затвора слоем диэлектрика (часто вместо «МОП-транзистор» говорят «транзистор с изолированным затвором»), поэтому такой транзистор обладает очень большим входным сопротивлением. Это обстоятельство вместе с большей технологической простотой дает МОП-транзисторам преимущество по сравнению с биполярными транзисторами, которые, в свою очередь, имеют преимущество в быстродействии. Чем закончится эта «битва богов и титанов», пока не совсем ясно, однако есть такая «вариация на тему МОП-транзисторов», которая поистине достойна удивления и восхищения. Речь идет о так называемой КМОП-логике – комплементарной МОП-логике. Слово «complement» по-английски означает «дополнение» и не имеет никакого отношения к прекрасному слову «compliment».

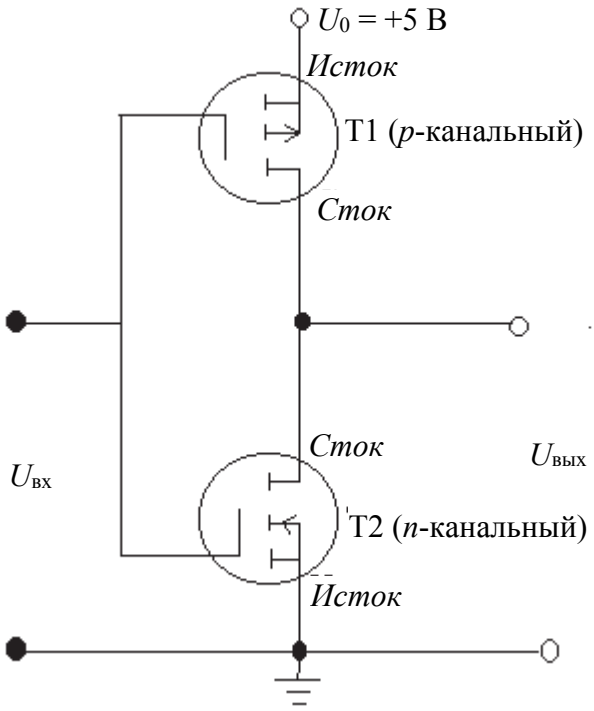


Рис. 5.16. КМОП-инвертор на базе транзисторов с индуцированным каналом

Так называемый КМОП-инвертор схематически изображен на рис. 5.16 и работает следующим образом. Когда входное напряжение $U_{вх}$ = 0 (на входе логический ноль), транзистор T2 закрыт, и его сопротивление очень велико. Транзистор T1 же при этом имеет отрицательный потенциал на затворе относительно

истока, в нем индуцируется *p*-канал, этот транзистор открыт и имеет очень маленькое сопротивление.

Значит, все напряжение U_0 будет падать на транзисторе Т2, и $U_{вых} = U_0 = +5$ В (на выходе логическая единица). Подадим теперь на вход $U_{вх} = +5$ В (логическую единицу). На транзисторе Т1 станет $U_{зи} = 0$, и он закроется. Транзистор Т2, напротив, откроется, т. к. для него $U_{зи} = +5$ В. Теперь все напряжение U_0 падает на транзисторе Т1, и $U_{вых} = 0$ (на выходе логический ноль). Ну что ж, инвертор как инвертор. Но! Всегда, когда один из транзисторов открыт, другой − закрыт и суммарный ток стока этой комплементарной «сладкой парочки» равен току закрытого транзистора, т. е. обратному току *p-n*-перехода, что составляет около 50 нА.

Токи же затворов в МОП-транзисторе вообще мизерны. Тогда при стандартном напряжении питания $U_0 = 5$ В, рассеиваемая КМОП-инвертором мощность составит 0,25 мкВт, что на несколько порядков меньше, чем у любых других аналогичных устройств. Именно крайне малая потребляемая мощность и сделала КМОП-логику такой знаменитой. Поэтому, когда Вы смотрите на электронные часы, работающие годы от крохотной батарейки, знайте − если бы не КМОП-логика, таких часов не существовало бы!

5.6. Приборы с зарядовой связью

Приборы с зарядовой связью, которые обычно называют просто ПЗС, имеют в своей основе МОП-структуру, работающую в режиме глубокого обеднения. Внешнее сходство с МОП-транзисторами является большим достоинством ПЗС, потому что их производство не требует смены технологии и, следовательно, существенных капиталовложений.

Глубокое обеднение подразумевает, что устройство должно работать в динамическом (импульсном) режиме, характерные времена которого мы рассматривали раньше.

Единичным элементом ПЗС служит трехэлектродная ячейка (рис. 5.17), имеющая три электрода, отделенных слоем оксида от полупроводника − в данном случае для определенности выбран *n*-кремний. В момент времени t_1 на электрод *1* подается отрицательный потенциал, создающий обедненную область в приповерхностном слое кремния. Как показывает распределение поверхностного потенциала в кремнии у границы с диэлектриком (рис. 5.17), под электродом *1* в кремнии создается потенциальная яма.

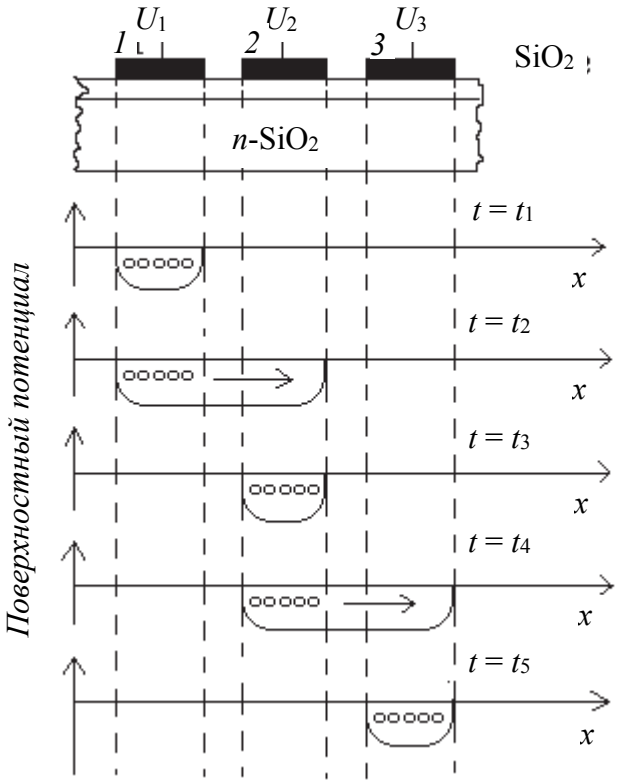


Рис. 5.17. Принципиальное устройство и режимы работы ПЗС

Для того чтобы ПЗС приносил пользу, эту яму надо заполнить дырками путем какого-то внешнего воздействия. Например, эти дырки можно инжектировать через смещенный в прямом направлении p-n-переход или создать путем освещения, в общем, они откуда-то должны появиться.

После того как потенциальная яма под электродом *1* заполнена дырками, надо подать такой же отрицательный потенциал на электрод *2* (рис. 5.17, момент времени t_2). Потенциальная яма станет в два раза шире, и дырки начнут всю ее заполнять за счет диффузии. Если теперь не слишком быстро выключать напряжение U_1, то к моменту времени t_3, когда $U_1 = 0$, все дырки вытолкнутся в потенциальную яму под электродом *2*.

Можно ли снова накапливать дырки под электродом *1*? Пока еще нет, надо убрать сначала дырки из-под электрода *2*, для чего проделаем прежний фокус − подадим напряжение U_3 (момент t_4), а затем выключим U_2 (момент t_5). Теперь мы переместили пакет заряда от электрода *1* к электроду *3* и можем под электродом *1* накапливать следующий пакет.

Если мы создадим из таких трехэлектродных ячеек решетку, в которой каждый третий электрод соединен с остальными (рис. 5.18), то получим устройство, хранящее и перемещающее пакеты зарядов. Если решетка имеет, например, 3000 электродов, то в ней можно одновременно накапливать 1000 пакетов заряда. Некоторые варианты применения таких устройств очевидны, например, запоминаю-

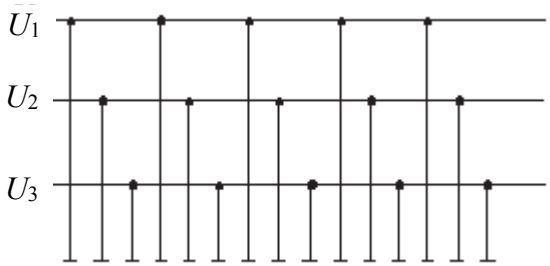


Рис. 5.18. Решетка электродов в ПЗС

щие устройства для стековой памяти или линии задержки. Другие применения ПЗС не столь очевидны. К примеру, ПЗС можно использовать для обработки видеосигнала. Для этого надо изготовить плоскую решетку с прозрачными электродами и сфокусировать на эту поверхность изображение. Падающий свет будет генерировать в кремнии электрон-дырочные пары с концентрацией, пропорциональной интенсивности падающего света. В течение некоторого времени, называемого *периодом интегрирования*, на решетку подается только напряжение U_1, в соответствующих ячейках накапливаются дырки (или электроны, если Вы используете *p*-кремний и положительные потенциалы на электродах), т. е. происходит запись изображения в виде пакетов заряда. Затем следует период считывания, в течение которого с решетки ПЗС происходит съем информации путем переключения напряжений U_1, U_2, U_3. Таким образом, мы фактически имеем видикон, т. е. передающую «трубку», преобразующую оптическое изображение в электрический сигнал.

Огромное достоинство ПЗС − исключительная дешевизна, основной недостаток при использовании в качестве запоминающих устройств − потеря информации при отключении источника питания (так называемые ПЗУ − постоянные запоминающие устройства, из них не сделаешь).

Обдумав физические принципы работы ПЗС, попробуйте сформулировать сами, какие и почему существуют ограничения в этих приборах на скорость передачи заряда, на величину накапливаемого заряда, а также на размер ячейки (т. е. фактически на габариты устройств).

5.7. Неравновесные носители заряда

При отклонении распределения электронов по разрешенным состояниям от равновесного в полупроводнике возникает много интересных явлений. С такими отклонениями от термодинамического равновесия мы сталкиваемся в любом биполярном полупроводниковом приборе. Однако и в однородном полупроводнике можно создать в заметном количестве неравновесные носители, например, облучая его фотонами достаточной энергии (а что значит «достаточной»?). При поглощении электронами таких фотонов будут образовываться избыточные по отношению к равновесным электроны и (или) дырки, что приведет к увеличению удельной проводимости материала. Такое явление называется *фотопроводимостью*.

Рассмотрим развитие этого процесса во времени, или, как обычно говорят, его *кинетику*, полагая для простоты, что освещение (не обязательно в видимой области спектра) создает (генерирует) один тип носителей, допустим, электроны в зоне проводимости, со скоростью A. Этот процесс будет сопровождаться обратным процессом рекомбинации, скорость которого пропорциональна концентрации Δn избыточных носителей, в данном случае электронов, которые обычно называют *фотоносителями*. Тогда изменение концентрации неравновесных фотоносителей в единицу времени:

$$\frac{d(\Delta n)}{dt} = \mathrm{A} - B(\Delta n). \tag{5.22}$$

Разделив в этом уравнении переменные и проинтегрировав его с учетом начального условия $\Delta n = 0$ при $t = 0$, получим:

$$\Delta n = (A/B)\cdot(1 - e^{-Bt}). \tag{5.23}$$

Поскольку из соображений размерности удобно ввести обозначение $\tau = 1/B$, соотношение (5.23) обычно записывают в виде

$$\Delta n = A\tau\,(1 - e^{-t/\tau}), \tag{5.24}$$

а величину τ называют *временем жизни* неравновесных носителей (не путайте с временем релаксации в выражениях для подвижности, которое тоже обычно обозначают τ!).

При $t \to \infty$ величина Δn стремится к установившемуся значению

$$\Delta n_m = A\tau. \tag{5.25}$$

Скорость генерации фотоносителей A пропорциональна интенсивности света I падающего на полупроводник (вообще говоря, монохроматического) и коэффициенту поглощения α:

$$A = \gamma\alpha I, \tag{5.26}$$

а коэффициент пропорциональности γ называют *квантовым выходом*, так как он представляет собой число носителей заряда (или пар носителей), образуемых одним квантом света, если интенсивность света I измерять числом квантов в секунду. Обычно квантовый выход γ не превышает единицу. Тогда установившаяся концентрация неравновесных носителей:

$$\Delta n = \gamma\alpha I\tau. \tag{5.27}$$

Видите, от чего она зависит?

Если же прекратить освещение полупроводника, то уравнение (5.22) примет вид

$$\frac{d(\Delta n)}{dt} = -B(\Delta n) = -\frac{\Delta n}{\tau}, \tag{5.28}$$

а его решение с учетом начальных условий $\Delta n = \Delta n_m$ при $t = 0$:

$$\Delta n = \Delta n_m e^{-t/\tau}. \tag{5.29}$$

Изменение со временем удельной проводимости $\Delta\sigma$, представляющей собой разность удельных проводимостей освещенного и затемненного материала, будет в соответствии с (5.24) и (5.29) описываться формулами (объясните точно, почему):

$$\Delta\sigma = \Delta\sigma_m \left[1 - \exp(t/\tau)\right] \tag{5.30}$$

при освещении и

$$\Delta\sigma = \Delta\sigma_m \exp(-t/\tau) \tag{5.31}$$

при прекращении освещения.

Посмотрите внимательно на формулы (5.30) и (5.31), дайте на их основе определение величины τ и предложите экспериментальную методику определения времени жизни неравновесных носителей в полупроводнике.

5.8. Диффузионный и дрейфовый ток

До сих пор мы рассматривали только ток, возникающий исключительно в результате действия внешнего электрического поля, под влиянием которого носители заряда приобретают скорость дрейфа. Такой ток называется *дрейфовым*, и именно он доминирует в металлах, обладающих равномерным по объему распределением носителей. Не так обстоит дело в полупроводнике, так как в нем очень легко получить в разных местах разную концентрацию носителей заряда, создав, например, градиент температур, или освещая часть полупроводникового образца, или неоднородно легируя полупроводник. При этом образуется градиент концентраций носителей заряда *gradn*, и, как следствие, возникает *диффузионный ток*, плотность которого $\vec{J}$ описывается обычным законом диффузии:

$$\vec{J} = -eDgradn, \tag{5.32}$$

где D − коэффициент диффузии. Как показал Эйнштейн при анализе броуновского движения, коэффициент диффузии и подвижность связаны соотношением

$$D = kT\mu/e, \tag{5.33}$$

которое и называется *соотношением Эйнштейна*. Формулы (5.32) и (5.33) применимы как к электронам, так и к дыркам, надо лишь быть аккуратным при выборе знаков.

Суммарный ток в полупроводнике будет складываться из дрейфового и диффузионного тока (уточните сами это утверждение, не забывая, что плотность тока − векторная величина). В состоянии теплового равновесия не существует *результирующего* потока носителей заряда в неоднородно легированном кристалле, хотя могут существовать равные по величине и противоположные по направлению диффузионные и дрейфовые потоки. Если тепловое равновесие нарушить, возможность компенсации диффузионного и дрейфового тока утрачивается и возникает результирующий поток носителей.

Предположим, что в зоне проводимости находятся избыточные электроны при таких условиях, что движение этих электронов обусловлено диффузией, а не электрическим полем. В конце концов все они рекомбинируют, и доля электронов, которая до рекомбинации пройдет при беспорядочном движении расстояние L, будет порядка $\exp(-L/L_e)$. Характерную длину L_e называют *диффузионной длиной электронов*. Она равна

$$L_e = (\tau_e D_e)^{1/2}. \tag{5.34}$$

Аналогичное соотношение справедливо и для диффузионной длины дырок L_h:

$$L_h = (\tau_h D_h)^{1/2}. \tag{5.35}$$

Диффузионные длины электронов и дырок чрезвычайно важны для работы биполярных полупроводниковых приборов, действие которых основано на инжекции неосновных носителей в часть полупроводникового кристалла.

5.9. Экспериментальные методы исследования полупроводников

Попробуем теперь подвести итог: что нужно для описания электрических свойств полупроводника, а именно его удельной проводи-

мости? Какие параметры необходимо измерять экспериментально, чтобы сказать: да, я знаю о проводимости все. Если вы усвоили материал предыдущих разделов, то сразу сможете ответить на этот вопрос и перечислить такие параметры. Это подвижности электронов и дырок, эффективные массы электронов и дырок, ширина запрещенной зоны и (для примесных полупроводников) энергии ионизации примесей. В случае неравновесных процессов следует добавить время жизни неравновесных носителей.

Именно перечисленные выше параметры, которые являются *параметрами зонной теории*, и были приведены в таблицах свойств собственных и примесных полупроводников. Теперь нам следует обсудить вопрос о том, как эти параметры могут быть определены экспериментально.

При измерении подвижности первое, что приходит в голову, − использовать само определение подвижности как скорости дрейфа носителей υ_D при единичной напряженности $\mathcal{E}$ электрического поля:

$$\mu = \upsilon_D/\mathcal{E}. \qquad (5.36)$$

Тогда если в плоском образце толщиной d, к которому приложено напряжение U, создать у одной поверхности пакет носителей, который затем в электрическом поле $\mathcal{E} = U/d$ будет дрейфовать через образец, и измерить *время пролета* носителей через образец $t = d/\upsilon_D$, то тем самым можно определить подвижность как

$$\mu = d^2/tU. \qquad (5.37)$$

Все методы измерения подвижности такого типа называются *времяпролетными* и отличаются друг от друга способами создания пакета носителей. Пакет носителей может создаваться инжекцией носителей из металлического контакта, как в классическом опыте Шокли-Хейнса; световым импульсом; импульсным пучком высокоэнергетичных электронов и т. д.

На рис. 5.20 приведена схема для случая возбуждения носителей фотонами, имеющими энергию большую, чем ширина запрещенной зоны полупроводника. Основная доля фотонов при этом поглощается в тонком приповерхностном слое полупроводника, генерируя как электроны, так и дырки. Следовательно, меняя напряженность «тянущего» напряжения, можно измерить время пролета как электронов, так и дырок, регистрируя с помощью запоминающего осциллографа интервал времени между срабатыванием оптического затвора и возникновением максимума тока в нагрузочном резисторе R. Длитель-

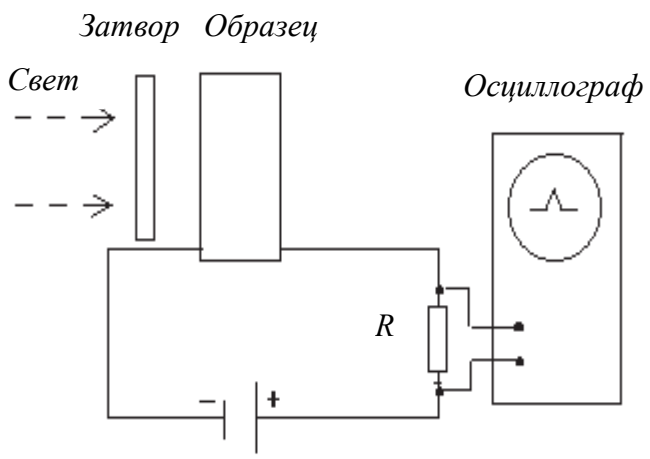


Рис. 5.20. Схема времяпролетного метода измерения подвижности носителей при использовании фотовозбуждения

ность светового импульса при этом должна быть много меньше, чем время пролета, поэтому (убедитесь численно сами!) затвор должен быть электрооптическим (работающим на эффекте Керра или эффекте Поккельса) либо в качестве источника света следует использовать импульсный лазер.

Другой экспериментальный метод определения подвижности основан на одновременном измерении удельной проводимости и постоянной Холла. В первом разделе мы с Вами получили соотношение для постоянной Холла $R_H = 1/en$. Поскольку в случае одного типа носителей $\sigma = en\mu$, мы можем сразу записать

$$\mu = \sigma R_H. \tag{5.38}$$

Однако соотношение (5.38) справедливо только для металлов и вырожденных полупроводников, т. е. для материалов, у которых все электроны двигаются с одинаковой скоростью, равной скорости Ферми. В общем случае постоянная Холла

$$R_H = \gamma/en, \tag{5.39}$$

где безразмерная величина γ, называемая холл-фактором, зависит от комбинации процессов рассеяния, эффективных в данных условиях, и от того, как меняется время релаксации с энергией электрона. Как правило, холл-фактор близок по величине к единице. Например, $\gamma = 3\pi/8$, если преобладает рассеяние на фононах и электронный газ не вырожден. Когда существенно рассеяние на заряженных центрах, γ для изотропной зоны близко к 1,9; с другой стороны, γ может уменьшиться до 0,7, если изоэнергетические поверхности существенно отклоняются от сферической формы.

Таким образом, в подобных экспериментах измеряется величина

$$\mu_H = \gamma\mu = \sigma R_H, \tag{5.40}$$

которая называется *холловской подвижностью* в отличие от μ, называемой *дрейфовой подвижностью*. Для многих полупроводников

в настоящее время имеются экспериментальные данные только по холловской подвижности, что создает очевидные трудности.

В полупроводниках со смешанной электронно-дырочной проводимостью постоянная Холла

$$R_H = \frac{\gamma}{e} \cdot \frac{\mu_h^2 n_h - \mu_e^2 n_e}{\mu_h n_h + \mu_e n_e}, \tag{5.41}$$

знак постоянной Холла может быть положительным и отрицательным.

Для собственных полупроводников $n_e = n_h = n$, и формула (5.41) принимает вид

$$R_H = \frac{\gamma}{en} \cdot \frac{\mu_h - \mu_e}{\mu_h + \mu_e}, \tag{5.42}$$

то есть в области собственной проводимости знак постоянной Холла определяется знаком заряда носителей, подвижность которых выше. Обычно такими носителями являются электроны (почему?). Поэтому, например, в примесном дырочном полупроводнике при переходе к собственной проводимости холловская ЭДС проходит через нуль и изменяет знак.

Измерение *эффективной массы* представляет собой гораздо более серьезную проблему. Фактически единственный метод измерения эффективной массы – наблюдение *циклотронного резонанса*.

Заряженная частица с зарядом q, например, электрон, движущаяся со скоростью υ в магнитном поле с индукцией B, испытывает на себе действие силы Лоренца:

$$\vec{F}_n = q\vec{\upsilon} \cdot \vec{B}.$$

Поскольку эта сила всегда направлена перпендикулярно вектору скорости, она создает только нормальное ускорение $a_n = \upsilon^2/R$, где R – радиус кривизны траектории частицы. Тогда, если магнитное поле перпендикулярно вектору скорости, уравнение движения электрона имеет вид

$$e\upsilon B = m\upsilon^2/R,$$

откуда

$$\upsilon/R = eB/m. \tag{5.43}$$

Траектория движения электрона в магнитном поле будет представлять собой окружность, а циклическая частота движения электрона по этой окружности

$$\omega = \frac{2\pi}{T} = \frac{2\pi\upsilon}{2\pi R} = \frac{\upsilon}{R},$$

откуда, сопоставляя это выражение с (5.43), получим:

$$\omega = eB / m. \qquad (5.44)$$

Если электрон (или дырка) двигаются в твердом теле, в качестве m в (5.44) должна (вспомните, почему!) фигурировать соответствующая эффективная масса. Если образец из исследуемого вещества поместить, кроме постоянного магнитного поля с индукцией B, в переменное электромагнитное поле, то при частоте этого переменного поля, удовлетворяющей условию (5.44), возникает резонанс, который и называется циклотронным резонансом, и энергия переменного электромагнитного поля будет поглощаться образцом. Тогда в соответствии с (5.44) эффективные массы носителей m^* связаны с резонансной *циклотронной частотой* $\omega_ц$ соотношением

$$\omega_ц = eB / m^*, \qquad (5.45)$$

откуда и могут быть найдены эффективные массы носителей в полупроводнике.

Схема эксперимента по циклотронному резонансу представлена на рис. 5.21. Электромагнитная волна с частотой в десятки ГГц от СВЧ-генератора по волноводу *1* поступает в циркулятор, направляющий эту волну в волновод *2*, содержащий образец исследуемого материала. Пройдя через образец, волна отражается от специального отражателя, снова проходит через образец и поступает в циркулятор, который по

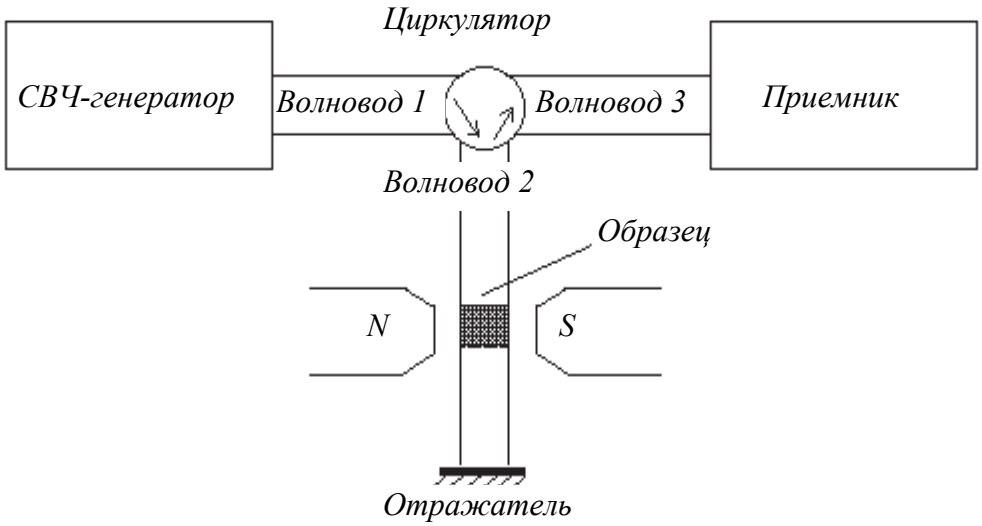


Рис. 5.21. Схема эксперимента по циклотронному резонансу

волноводу *3* направляет ее в приемник СВЧ-излучения. Образец помещен между полюсами электромагнита, регулируя ток в обмотке которого, можно изменять индукцию магнитного поля *B*. При значениях *B*, удовлетворяющих условию (5.45), будет наблюдаться поглощение образцом мощности СВЧ-волны, регистрируемое приемником.

На рис. 5.22 приведены результаты подобного эксперимента для германия, имеющего такую же зонную структуру, что и кремний. На спектре циклотронного резонанса наблюдаются два резонансных пика, соответствующих дыркам, и два электронных пика. Попробуйте догадаться сами, куда пропал третий дырочный пик и откуда взялся второй электронный. А вот вопрос посложней − попробуйте сообразить, как отличить электронные пики от дырочных (подсказка: просто по картинке это сделать нельзя, нужно какое-то экспериментальное ухищрение). Получить спектр циклотронного резонанса довольно трудно. Как и при любом резонансе, пик поглощения не наблюдается, если затухание велико, то есть если большинство носителей испытывают рассеяние прежде, чем повернет на угол хотя бы в один радиан. Так как число соударений в единицу времени − есть величина, обратная времени релаксации τ, условие резонанса имеет вид

$$\omega_ц > 1/\tau. \qquad (5.46)$$

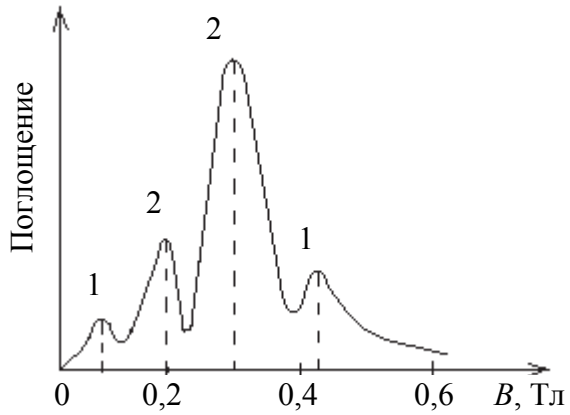


Рис. 5.22. Спектр циклотронного резонанса в германии. Пики 1 соответствуют дыркам, а пики 2 – электронам. Частота СВЧ-излучения ω = 24 ГГц

Это условие налагает жесткие ограничения на возможность наблюдения циклотронного резонанса в диапазоне СВЧ: τ должно быть больше, чем 10^{-10} с. Это достигается только при температуре жидкого гелия (около 4 К) и лишь в некоторых полупроводниках высокой чистоты, такой, чтобы можно было пренебречь рассеянием на ионизованных и нейтральных примесях, доминирующим при такой низкой температуре. Именно при таких условиях и получен приведенный на рис. 5.22 спектр германия.

Другой метод наблюдения основан на применении инфракрасного излучения вместо излучения СВЧ, а также сильного магнитного поля,

которое можно создать либо в импульсном режиме (с индукцией приблизительно до 100 Тл), либо с помощью сверхпроводящего магнита (до 12 Тл). В этом случае условие резонанса можно выполнить даже при комнатной температуре ($\tau = 10^{-13}$ с). Правда, поскольку с импульсными магнитами трудно работать из-за потерь на индукцию, циклотронный резонанс в инфракрасном диапазоне при комнатной температуре позволяет измерить эффективную массу носителей, не превышающую $0{,}2m_0$ (рис. 5.23).

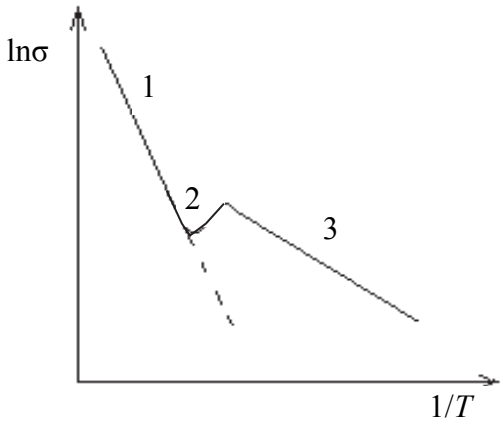


Рис. 5.23. Температурная зависимость удельной проводимости примесного полупроводника

Перейдем теперь к методам определения энергетических параметров зонной модели − *ширины запрещенной зоны* полупроводника и *энергий ионизации примесей*.

Наиболее простой метод измерения ширины запрещенной зоны основан на изменении проводимости и температуры. Проводимость полупроводника можно представить в виде

$$\sigma = e(n_e\,\mu_e + n_h\,\mu_h),$$

а для собственного полупроводника, у которого

$$n_e = n_h = n_i,$$

в еще более простом виде

$$\sigma = en_i(\mu_e + \mu_h) = \sigma_0 \exp(-E_g / 2kT), \qquad (5.47)$$

где величина σ_0 зависит от температуры, но степенным образом, т. е. температурная зависимость проводимости будет в основном определяться экспонентой в выражении (5.47). График зависимости $\ln\sigma$ от величины $1/T$, обратной температуре, будет представлять собой прямую линию с тангенсом угла наклона

$$\mathrm{tg}\alpha = -E_g / 2k, \qquad (5.48)$$

что и позволяет определить экспериментально ширину запрещенной зоны E_g.

У примесного полупроводника температурная зависимость проводимости имеет более сложный вид (рис. 5.23). При высоких темпе-

ратурах (область 1) преобладающей будет собственная проводимость, и остаются справедливыми выражения (5.47) и (5.48). По мере охлаждения полупроводника вероятность межзонных переходов уменьшается, и в некотором промежуточном температурном интервале (область 2), когда примесные уровни не ионизованы, а собственной проводимости практически нет, температурная зависимость проводимости определяется температурной зависимостью подвижности носителей. При дальнейшем охлаждении полупроводника происходит вымораживание примесной проводимости (область 3), и температурная зависимость проводимости описывается соотношением

$$\sigma \sim \exp(-E_i / 2kT), \tag{5.49}$$

а график зависимости $\ln\sigma$ от $1/T$ вновь представляет собой прямую линию, но уже с тангенсом угла наклона:

$$\mathrm{tg}\alpha = -E_i / 2kT, \tag{5.50}$$

откуда можно определить энергию ионизации примеси E_i.

Еще более простой способ определения ширины запрещенной зоны основан на исследовании спектра оптического поглощения полупроводника. Фотоны с энергией $h\nu$, меньшей, чем ширина запрещенной зоны, поглощаться полупроводником не могут, поэтому спектр поглощения полупроводника имеет ярко выраженный край полосы поглощения, соответствующий условию

$$h\nu = h\nu_0 = E_g. \tag{5.51}$$

При этом, однако, следует учесть, что соотношение (5.51) справедливо только для прямозонных полупроводников, как это уже обсуждалось раньше. В противном случае $h\nu_0 > E_g$ и спектр поглощения не позволяет определить точное значение ширины запрещенной зоны без дополнительных экспериментальных ухищрений, выходящих за рамки нашего курса.

Совершенно аналогично (и с теми же ограничениями) ширина запрещенной зоны может быть определена на основе спектра фотопроводимости полупроводника.

Таким образом, рассмотренные нами теоретические положения и экспериментальные методы позволяют составить достаточно полную картину электрических свойств полупроводников, позволяющую осмысленно создавать различные электронные устройства на основе таких материалов.

6. ИЗЛУЧЕНИЕ И ЛАЗЕРЫ

Mehr Licht!
Предсмертные слова Гёте

Концепция равновесия – это научный (или, по крайней мере, методический) принцип, столь же мощный, как и законы сохранения энергии или импульса. Когда мы говорим, что электроны распределены в соответствии с функцией Ферми − Дирака, то утверждаем сразу *две* вещи:

- система находится в состоянии равновесия;
- у системы есть определенная температура.

Нельзя говорить о температуре одного электрона о температуре выведенной из равновесия системы, так как не существует такой величины T, которую можно подставить в $f(E, T)$ и получить правильное распределение электронов по энергии. Правда, вы мне можете возразить и сказать, что мы уже много раз пользовались равновесными функциями распределения при описании неравновесного процесса электропроводности – ведь система электронов в этом случае выводится из равновесия внешним электрическим полем. Но вспомните: электрическое поле вызывает дрейфовую скорость, которая составляет всего лишь миллионную долю процента от тепловой скорости. Это означает, что *очень маленькое* отклонение от равновесия создает *очень большие* токи, поэтому в слабых полях электропроводность – *почти* равновесное явление.

В лазерах все иначе. Хотя распределение заселенности уровней в рабочем теле лазера сильно возмущено, тем не менее тут присутствует равновесие особого рода. Правда, абсолютная температура системы при этом отрицательна, но ничего страшного, ведь она «горячее» самой высокой положительной температуры. Странно? Давайте разберемся подробней.

Пусть у нас есть система (см. рис. 6.1), содержащая только два энергетических уровня: E_1 и E_3 (индекс 2 нам понадобится позже). В больцмановском приближении заселенности уровней E_1 и E_3, то есть количества электронов N_1 и N_3 с энергиями E_1 и E_3, связаны соотношением

$$N_3 = N_1 \cdot \exp\left(-\frac{E_3 - E_1}{kT}\right), \tag{6.1}$$

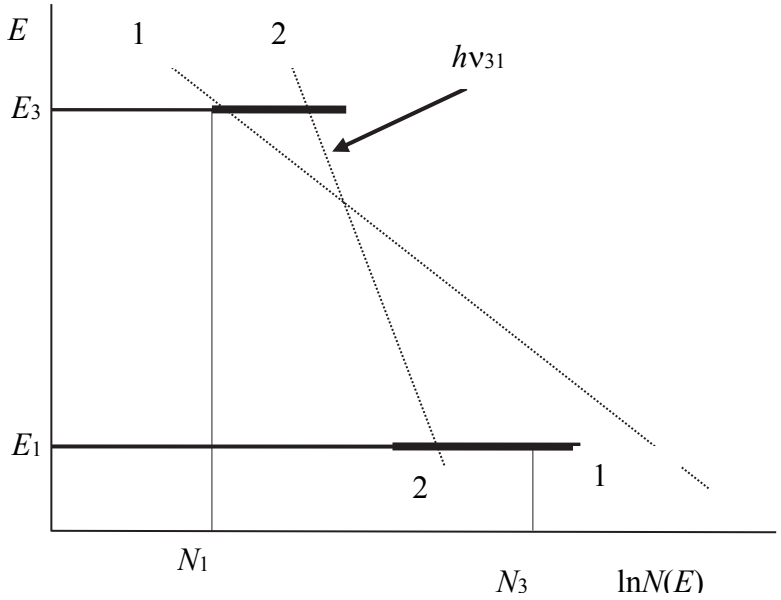


Рис. 6.1. Распределение электронов

где k – постоянная Больцмана, T – абсолютная температура.

Что реально может означать подобная двухуровневая система? Это могут быть любые два электронных энергетических уровня атома, молекулы или твердого тела, выбранные из множества разрешенных уровней.

Если наша система находится в равновесии, то число переходов электронов в единицу времени с уровня E_1 на уровень E_3 равно числу переходов с уровня E_3 на уровень E_1 (такое утверждение называется *принципом детального равновесия*). Будем считать (это зависит от конкретного материала), что все переходы с уровня E_3 на уровень E_1 являются излучательными, то есть сопровождаются излучением фотонов.

Такое излучение, обусловленное самопроизвольным «падением» электронов с верхних уровней на нижние, называется *спонтанным излучением*. При спонтанном излучении испускаются фотоны с частотой ν_{31}, удовлетворяющей условию

$$h\nu_{31} = E_3 - E_1, \tag{6.2}$$

где h – постоянная Планка. (Кстати, какой физический закон требует, чтобы выполнялось это условие?). При переходах электронов с уровня E_1 на уровень E_3 происходит поглощение фотонов с такой же частотой, и процесс так и называется – *поглощение*. Таким образом, тепловое равновесие в нашей системе поддерживается порхающими в ней фотонами, которые беспрерывно излучаются и поглощаются при электронных переходах. Это пронизывающее равновесную систему излучение называется тепловым, и его спектральная плотность $\rho(\nu)$, то есть объемная плотность энергии фотонов, приходящаяся на единичный интервал частот вблизи заданной частоты ν, описывается знаменитой *формулой Планка*:

$$\rho(\nu) = \frac{8\pi n^3 h\nu^3}{c^3} \cdot \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1}, \qquad (6.3)$$

где c – скорость света в вакууме; n – показатель преломления вещества, в котором развиваются описанные нами события.

Число электронных переходов в единицу времени (скорость переходов), соответствующих поглощению, пропорционально количеству электронов на нижнем уровне и спектральной плотности излучения, соответствующей частоте ν_{31}, то есть равно $B_{13}N_1\rho(\nu_{31})$. Коэффициент пропорциональности B_{13} обычно называют *коэффициентом поглощения*. Скорость электронных переходов, соответствующих спонтанному излучению, пропорциональна количеству электронов на верхнем уровне, то есть равна $A_{31}N_3$, где A_{31} – *коэффициент спонтанного излучения*. Коэффициенты B_{13} и A_{31} характеризуют вероятности процессов поглощения и спонтанного излучения соответственно.

Казалось бы, теперь мы можем приравнять друг к другу скорости процессов поглощения и спонтанного излучения и найти связь между вероятностями этих процессов. Однако все ли возможные процессы в нашей двухуровневой системе мы учли? Например, не могут ли электроны на верхнем уровне поглощать фотоны с частотой ν_{31}? Такая мысль на первый взгляд выглядит дикой, но… ведь разрешено все, что не запрещено, не так ли? А кто может запретить такой процесс? Конечно, законы сохранения энергии и импульса. Давайте попробуем записать закон сохранения энергии для такого процесса. Чтобы энергия действительно сохранялась, такое соотношение должно выглядеть следующим образом:

$$E_3 + h\nu_{31} = E_1 + h\nu_{31} + h\nu_{31}.$$

Интересно, не правда ли? Энергия при таком процессе будет сохраняться, если электрон на верхнем уровне поглотит *один* фотон, перейдет на нижний уровень и испустит *два* таких же фотона. Импульс при таком процессе тоже может сохраняться – проверьте, пожалуйста, это сами в качестве маленького упражнения. Итак, такой процесс возможен. Он называется *вынужденным излучением*. Скорость вынужденного излучения будет равна $B_{31}N_3\rho(\nu_{31})$, где B_{31} называют *коэффициентом вынужденного излучения*.

Вот теперь мы можем в соответствии с принципом детального равновесия приравнять скорости электронных переходов снизу вверх и сверху вниз:

$$B_{13}N_1\rho(\nu_{31}) = A_{31}N_3 + B_{31}N_3\rho(\nu_{31}). \qquad (6.4)$$

Полученное соотношение (6.4) с учетом формул (6.1)–(6.3) можно преобразовать к виду

$$\rho(\nu_{31}) = \frac{A_{31}}{B_{13}\cdot\exp\left(\frac{h\nu}{kT}\right) - B_{31}}. \qquad (6.5)$$

Сравнение формул (6.3) и (6.5) показывает, что для вероятностей электронных переходов должны выполняться условия

$$B_{13} = B_{31}, \qquad (6.6)$$

$$A_{31} = B_{31}\cdot\frac{8\pi n^3 h\nu^3}{c^3}. \qquad (6.7)$$

Полученные нами соотношения называются *соотношениями Эйнштейна*, а вся изложенная выше теория – *теорией излучения Эйнштейна*.

Описанная выше ситуация реализуется без какого бы то ни было внешнего излучения следующим образом. Поскольку наша двухуровневая система обладает ненулевой температурой, в соответствии с уравнением (6.1) всегда есть вероятность переходов электронов с уровня E_1 на уровень E_3. Появившиеся на уровне E_3 электроны рано или поздно падают на уровень E_1, создавая спонтанное излучение фотонов с энергией $h\nu_{31}$. Эти фотоны порождают вынужденное излучение фотонов с той же энергией. В системе устанавливается равновесие, описываемое как уравнением (6.1), так и уравнением (6.4). Этому равновесию, а конкретнее уравнению (6.1), соответствует штриховая линия 1–1 на рис. 6.1.

Давайте теперь посветим на такую систему внешним излучением с энергией фотонов $h\nu_{31}$. Это вовсе не обязательно должно быть монохроматическое излучение, пусть это будет излучение обычной лампы накаливания со сплошным спектром, лишь бы в нем присутствовали фотоны с указанной выше энергией. Такие фотоны, во-первых, будут поглощаться двухуровневой системой, а во-вторых, создавать в ней вынужденное излучение фотонов с той же энергией. В итоге заселенность уровней E_1 и E_3 электронами изменится. Увеличение числа поглощенных фотонов уменьшит заселенность нижнего уровня. При этом, правда, увеличится и количество переходов электронов с верхнего уровня на нижний, но, хотя вероятности вынужденных переходов

вверх и вниз абсолютно одинаковы, переходов с E_1 на E_3 будет больше, поскольку на нижнем уровне находится значительно больше электронов. Иными словами, в результате преобладающим эффектом будет поглощение света, которое мы постоянно наблюдаем в природе. В поле внешнего излучения установится равновесие, уравнение (6.1) для которого показано на рис. 6.1 штриховой линией 2–2, соответствующей более высокой, чем для линии 1–1, температуре. Ну конечно, внешнее излучение нагревает нашу систему. Если мы будем увеличивать интенсивность внешнего излучения, температура системы будет становиться все больше и больше, но уравнение (6.4) показывает, что она все равно будет иметь конечную величину при любой, сколь угодно большой, интенсивности внешнего излучения. Процессы поглощения всегда будут преобладать над процессами вынужденного излучения.

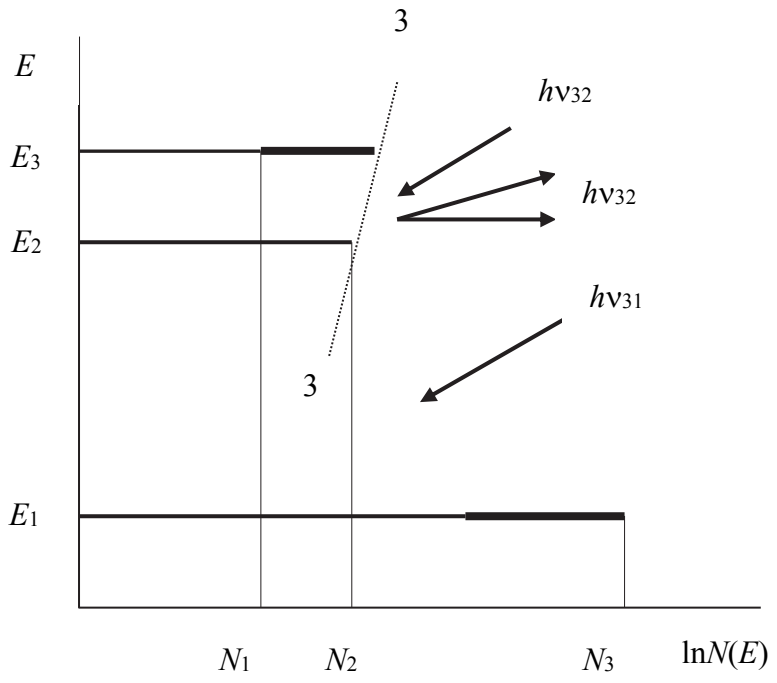


Рис. 6.2. Распределение электронов в инверсной структуре

Ситуация, однако, может резко измениться, если в нашей системе будет присутствовать третий электронный уровень с энергией E_2, большей, чем E_1, но меньшей, чем E_3 (рис. 6.2). Если мы будем освещать такую трехуровневую систему только фотонами с энергией $h\nu_{32}$, то основным откликом системы на это воздействие будет поглощение этих фотонов. Если же мы будем освещать систему фотонами с энергией $h\nu_{31}$, то, как рассматривалось выше, мы увидим поглощение этих фотонов. Но при увеличении интенсивности этого освещения (его называют устоявшимся термином *накачка*) заселенность электронами самого верхнего из трех уровней будет возрастать, и в конце концов станет больше, чем заселенность уровня E_2. Такое явление называют

инверсией заселенности. Посмотрев внимательно на линию 3–3 (рис. 6.2) и формулу (6.1), мы с удивлением обнаружим, что температура системы, состоящей из уровней E_2 и E_3, стала отрицательной, причем эта отрицательная температура больше самой большой положительной. Впрочем, это не самое интересное свойство системы с инверсной заселенностью уровней. Пусть на эту систему будут падать фотоны с энергией $h\nu_{32}$. Поскольку заселенность уровней с энергией E_3 больше, чем уровней с энергией E_2, процессы вынужденного излучения будут преобладать над процессами поглощения, то есть из такой системы будет вылетать фотонов с энергией $h\nu_{32}$ больше, чем падать на нее. Мы будем наблюдать явление, которое по-английски называется Light Amplification by Stimulated Emission of Radiation (усиление света за счет испускания вынужденного излучения). Значит, мы получим усилитель света, который по первым буквам в английском наименовании явления называется *лазер* (laser). Поскольку в спектре излучения ламп накачки присутствуют не только фотоны с энергией $h\nu_{31}$, но и фотоны с энергией $h\nu_{32}$, лазер будет работать не только как усилитель, но и как генератор оптического излучения, причем это излучение будет монохроматическим.

Если вспомнить, что фотоны являются бозе-частицами, то можно предсказать еще два интересных свойства лазерного излучения. Если фермионы подчиняются принципу Паули и не собираются в количестве больше одного в каждом квантовом состоянии, то бозоны, напротив, ведут себя по принципу «чем нас больше соберется, тем нам будет веселее». Если в системе появляется бозон в каком-то квантовом состоянии, то сразу повышается вероятность появления в системе второго бозона в таком же квантовом состоянии. Появление второго бозона резко повышает появление такого же третьего, и так далее по нарастающей. Процесс развивается очень быстро, поэтому все возникшие бозоны излучаются практически одновременно, то есть лазерное излучение является когерентным (кстати, а можно ли создать, допустим, электронный лазер?). Кроме того, поскольку одним из квантовых состояний фотона является состояние его поляризации, то лазерное излучение поляризовано.

Итак, что же требуется для того, чтобы заработал лазер? Во-первых, нужна трехуровневая система. Понятно, что в ней не обязательно должно быть ровно три уровня (да это и невозможно) – пусть их будет хоть три миллиона. Далее, в этой системе необходимо создать инверсную заселенность уровней. Это уже проблема посложнее,

и обычно она решается не столь прямолинейно, как описывалось выше. Тем не менее, в настоящее время известно множество лазерных сред, то есть веществ, которые могут испускать лазерное излучение – твердых, жидких и газообразных. Однако же найти самую первую такую среду было, конечно, весьма непросто. Первая такая удачная попытка была осуществлена в 1954 году одновременно в Физическом институте Академии наук (СССР) Николаем Басовым и Александром Прохоровым и в Колумбийском университете (США) Чарлзом Таунсом, за что эти трое исследователей были удостоены в 1964 году Нобелевской премии по физике. В качестве рабочего вещества в этом первом устройстве использовался аммиак. Правда, излучал он не свет, а радиочастотные волны (microwave), и был назван Таунсом мазером (помните, откуда берется подобное название?). Благодаря способности усиливать радиоволны мазеры сразу начали применяться в радиотелескопах. Первый именно лазер был создан в 1960 году американским физиком Теодором Мейманом. Рабочим телом в этом лазере был стержень из рубина. Рубин представляет собой окись алюминия, содержащую атомы хрома, которые и придают рубину красивый цвет и, как следствие, ювелирную ценность. Именно переходы в атомах хрома создают лазерное излучение рубина буквально по такой примитивной схеме, которая изображена на рис. 6.2.

Мощным импульсом света от газоразрядной ксеноновой лампы накачки (рис. 6.3) электроны в атомах хрома (а точнее, ионах Cr^{3+}) возбуждаются с уровней основного состояния на уровни накачки, на которых создается инверсная заселенность по отношению к промежуточным уровням, соответствующим состоянию с энергией E_2 на рис. 6.2. Вынужденные переходы электронов с уровней накачки на промежуточные вызывают излучение света в красном диапазоне спектра на длине волны около 690 нм. Накачка осуществляется в импульсном режиме, соответственно и излучение рубинового лазера импульсное. С двух концов рубинового стержня располагаются плоские зеркала, одно из которых полупро-

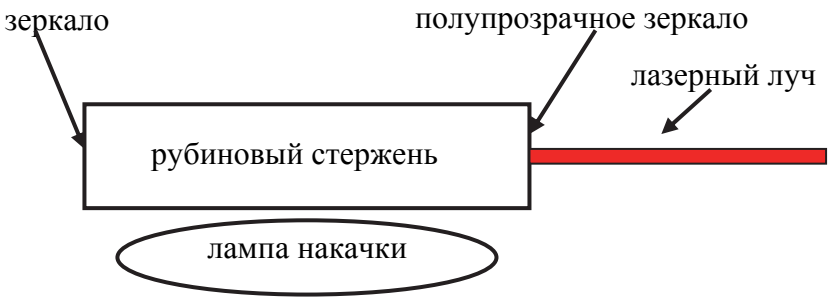


Рис. 6.3. Принципиальная схема рубинового лазера

зрачное. Отражаясь поочередно от этих зеркал, генерированные в рубине фотоны опять и опять проходят через стержень, создавая все новые фотоны. Процесс нарастает лавинообразно до тех пор, пока наконец световой импульс не выйдет наружу через полупрозрачное зеркало.

Рубиновый кристалл в такой схеме может быть заменен другим веществом, содержащим подходящие для излучения атомы, причем этим веществом не обязательно должно быть твердое тело. Еще в 1960 г. Али Джаван, американский физик иранского происхождения, создал первый газовый лазер. Сейчас весьма распространены маломощные газовые лазеры на основе смеси гелия с неоном, излучающие на длине волны 632 нм и используемые для исследовательских и метрологических целей, и мощные лазеры на основе CO_2, излучающие в инфракрасном диапазоне на длине волны 10,6 мкм и используемые для технологических целей. Оба эти типа лазеров излучают свет, в отличие от рубинового лазера, не в импульсном, а в непрерывном режиме.

В 1966 г. Б. Степановым, А. Рубиновым и В. Мостовниковым в Институте физики АН БССР (Минск) были созданы первые жидкостные лазеры на основе органических красителей, которые благодаря низкой стоимости и возможности использования в одном лазере набора красителей (как патроны в барабанном револьвере), излучающих на различных длинах волн, получили широкое распространение.

Совершенно особый класс лазеров представляют собой полупроводниковые лазеры. Эти лазеры отличаются тем, что они миниатюрны, легко управляются электронным способом, работают при низких напряжениях, устойчивы к механическим воздействиям, имеют большой срок службы и, что очень важно, изготавливаются по стандартной технологии микроэлектроники, то есть пригодны для массового производства и поэтому дешевы.

Основная идея, лежащая в основе работы полупроводниковых лазеров, очень проста. Она состоит в использовании излучательной рекомбинации электронно-дырочных пар в полупроводниках, которая и приводит к лазерной генерации. Для этого нам необходим материал с множеством электронов в зоне проводимости, стремящихся ее покинуть, и в то же время с множеством дырок в валентной зоне, готовых принять эти электроны. Ясно, что однородный полупроводник для этих целей не годится, так как в нем не реализуются одновременно оба указанных условия. Однако высокую концентрацию электронов можно получить в сильно легированном полупроводнике *n*-типа, а высокую концентрацию дырок – в сильно легированном полупроводнике *p*-типа. Если такие два полупроводника соединить, то в обла-

сти *p-n*-перехода можно одновременно реализовать оба условия, необходимых для лазерной генерации.

На рис. 6.4 изображен контакт сильно легированных полупроводников *p*- и *n*-типа в тепловом равновесии и при прямом напряжении (то есть плюсом на p-области) на *p-n*-переходе. При подаче на *p-n*-переход прямого напряжения около перехода появляется область перекрытия, где велики концентрации как электронов, так и дырок. Эта область называется активной, так как именно здесь происходит излучательная рекомбинация электронно-дырочных пар. Чтобы поддерживать процесс излучения, необходимо компенсировать убыль электронов и дырок, пропуская через устройство электрический ток и тем самым инжектируя в активную область новые носители. Таким образом, конструктивно полупроводниковый лазер весьма похож на обычный полупроводниковый диод (по-английски он так и называется laser diode); схематически он изображен на рис. 6.5.

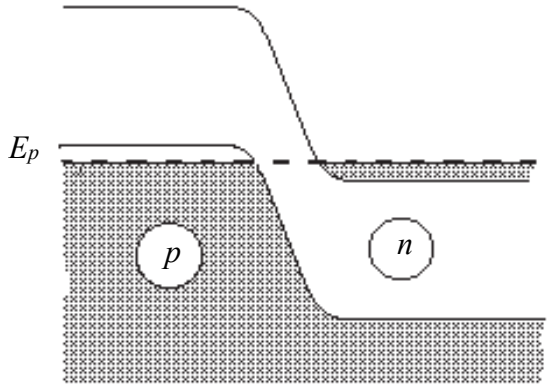


Рис. 6.4. Зонная структура полупроводникового лазера

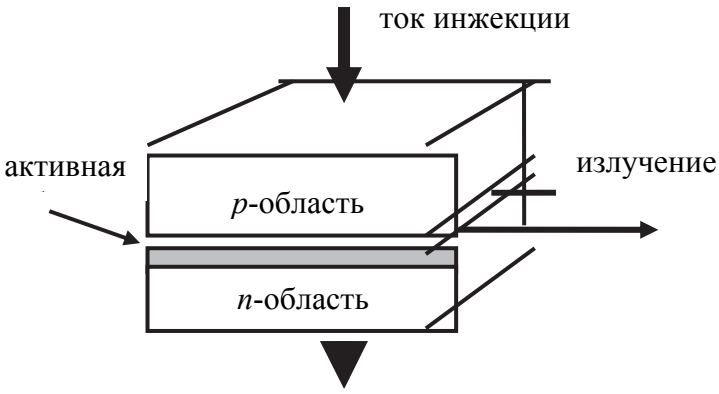


Рис. 6.5. Разрез устройства полупроводникового лазера

За четыре десятилетия с момента создания лазеров они нашли широкое применение в самых различных областях человеческой деятельности. Мощное излучение лазеров используется в промышленных и медицинских технологиях. Узкая направленность лазерного луча позволяет измерить расстояние до Луны с точностью до сантиметра; голография начала по-настоящему развиваться только с применением монохроматического лазерного излучения. Без лазерных проигрывателей компакт-дисков и без компьютерных CD ROOM'ов мы уже не представляем себе повседневную жизнь. На очереди – развитие лазерных средств высококачественной передачи информации типа телевизионных сигналов.

ОГЛАВЛЕНИЕ

Учебное издание

Почтенный Артем Евгеньевич

ФИЗИКА

В 6-ти частях

Часть 6

КВАНТОВЫЕ ЭЛЕКТРОННЫЕ СВОЙСТВА ТВЕРДЫХ ТЕЛ

Тексты лекций

Редактор *К. В. Великода*
Компьютерная верстка *К. В. Великода*
Корректор *К. В. Великода*

Издатель:
УО «Белорусский государственный технологический университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/227 от 20.03.2014.
Ул. Свердлова, 13а, 220006, г. Минск.