МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Национальный исследовательский
Нижегородский государственный университет им. Н.И. Лобачевского

ВВЕДЕНИЕ В ФИЗИКУ ПОЛУПРОВОДНИКОВЫХ ДИОДОВ И МЕТОДЫ ИХ ПРОЕКТИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ

Учебное пособие

Рекомендовано методической комиссией радиофизического факультета для студентов ННГУ, обучающихся по направлениям подготовки 03.03.03 и 03.04.03 «Радиофизика», 02.03.02 «Фундаментальная информатика и информационные технологии», специальностям 10.05.02 «Информационная безопасность телекоммуникационных систем», 11.05.02 «Специальные радиотехнические системы»

Нижний Новгород
2020

УДК 53.082, 538.95
ББК 32.85
В-24
Рецензенты: канд. физ.-мат. наук, доцент **Н.В. Прончатов-Рубцов**
докт. техн. наук, доцент **Е.С. Фитасов**

В-24. ВВЕДЕНИЕ В ФИЗИКУ ПОЛУПРОВОДНИКОВЫХ ДИОДОВ И МЕТОДЫ ИХ ПРОЕКТИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ. Авторы: Е.В. Волкова, А.С. Пузанов, С.В. Оболенский, Е.А. Тарасова: Учебное пособие. – Нижний Новгород: Нижегородский госуниверситет, 2020. – 78 с.

Данное пособие является продолжением цикла учебных пособий по полупроводниковой электронике и содержит информацию о физических основах работы полупроводниковых диодов и особенностях применения высокопроизводительных вычислений для анализа переходных ионизационных процессов и долговременных изменений характеристик полупроводниковых структур при радиационном воздействии.

Пособие предназначено для студентов ННГУ, обучающихся по направлениям подготовки 03.03.03 и 03.04.03 «Радиофизика», 02.03.02 «Фундаментальная информатика и информационные технологии», специальностям 10.05.02 «Информационная безопасность телекоммуникационных систем», 11.05.02 «Специальные радиотехнические системы».

Ответственный за выпуск:
зам. председателя методической комиссии радиофизического факультета ННГУ
д.ф.-м.н., профессор **Е.З. Грибова**

УДК 621.382
ББК 22.344

© Нижегородский государственный
университет им. Н.И. Лобачевского, 2020

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Данная книга является продолжением цикла учебных пособий, предназначенных для студентов 3 – 4–х курсов дневных отделений высших технических учебных заведений, специализирующихся на изучении физики полупроводниковых приборов. В первой книге [1] авторами были изложены основные определения и понятия, касающиеся физики полупроводников, а также приведен расширенный объем сведений из области теории транспорта электронов в полупроводниковых материалах и применения высокопроизводительных вычислений для моделирования указанных эффектов. Данное учебное пособие является логическим продолжением предыдущего и содержит сведения о физических процессах, протекающих вблизи контактов полупроводника с другими материалами, а также об основных физических принципах работы полупроводниковых диодов различных групп (выпрямительных, оптоэлектронных, генераторных). Для лучшего восприятия материала в пособии подробно рассмотрены принципы работы классических приборов, но, помимо этого, приведена информация и о современных гетеронаноструктурных диодах. В отдельной главе приведен обзор современных технологий создания полупроводниковых приборов и обсуждены принципы их проектирования с использованием технологий высокопроизводительных вычислений. Рассмотрена практическая задача анализа переходных ионизационных процессов и долговременных изменений характеристик полупроводниковых структур при радиационном воздействии.

Пособие не заменяет лекций, лабораторных практикумов или учебников. Оно является дополнительным источником сведений, в котором объяснения простых базовых понятий сочетаются с детальными описаниями сложных физических эффектов. В связи с таким подходом к изложению материала пособие может использоваться в качестве как приложения к базовому курсу лекций, так и методической основы спецкурса по физике современных полупроводниковых приборов и наноэлектронике.

Используемая в пособии терминология сочетается с понятийным аппаратом как хорошо известных учебников, так и достаточно редких монографий по физике полупроводников и полупроводниковых приборов [2–8], а также оригинальных методических разработок кафедры квантовой радиофизики и электроники.

ЧАСТЬ 1
ФИЗИКА ЯВЛЕНИЙ НА ГРАНИЦЕ РАЗДЕЛА ТВЕРДЫХ ТЕЛ. ВЫПРЯМЛЯЮЩИЕ ДИОДЫ

В первой части представленного учебного пособия по полупроводниковым диодным структурам и приборам изложены элементарные основы теории контактных явлений. При этом речь пойдет как о контактах между различными материалами, например, полупроводником и металлом, так и о контактах между областями с разным типом проводимости внутри одного полупроводникового кристалла. Физические свойства подобных контактов, называемых также переходами, широко используются для выпрямления тока и лежат в основе работы базовых элементов целого ряда полупроводниковых сверхвысокочастотных устройств и быстродействующих интегральных схем, а также приборов полупроводниковой оптоэлектроники.

Параграфы 1.1. – 1.4. данной части посвящены изучению границ раздела между двумя полупроводниками (если быть более точным, то между областями различной проводимости внутри единого полупроводникового кристалла). В параграфах 1.5, 1.6. рассматривается физика контактных явлений на границах материалов с существенно отличающейся по величине проводимостью (границы раздела металл–полупроводник, металл–диэлектрик–полупроводник). В заключение раздела в параграфе 1.7. рассматриваются общие эквивалентные схемы выпрямительных диодов.

1.1. Классификация границ раздела между двумя полупроводниками

Для обозначения уровня легирования полупроводниковых слоев, т.е. величин концентраций ионов доноров или акцепторов, в данном пособии будем использовать общепринятые сокращения:

- n^{++} и p^{++} – концентрация примеси в полупроводнике более 10^{19} см$^{-3}$;
- n^{+} и p^{+} – ~10^{18}…10^{19} см$^{-3}$;
- n и p – ~ 10^{16}…10^{18} см$^{-3}$;
- n^{-} и p^{-} – ~ 10^{15}…10^{16} см$^{-3}$;
- n^{--} и p^{--} – менее 10^{15} см$^{-3}$.

Буквы *«n»* и *«p»* обозначают, соответственно, электронный и дырочный типы приводимости материала. Слои нелегированного (чистого) полупроводника обозначаются символом *«i»*. Подразумевается, что температура, при которой работает прибор, является комнатной (кроме случаев, где явно указывается иная ситуация), а концентрация основных носителей заряда равна концентрации введенной примеси.

Для обозначения границы раздела полупроводников, которую также часто называют переходом, используют дефис или косую черту, например, n^{+}- n Si переход, p-Si / n-Ge переход. При этом в первом случае предполагается,

что оба слоя полупроводника изготовлены из кремния, а во втором случае - первая область из кремния, а вторая из германия.

В данном пособии мы будем использовать следующую классификацию границ раздела между двумя полупроводниками (таблица 1.1).

Таблица 1.1. Классификация границ раздела

	Гомопереход	Гетеропереход
Изотипный (униполярный) переход	Один и тот же полупроводник и один и тот же тип проводимости по обе стороны от границы раздела (пример: n^+- n и p^+- p переходы)	Разные полупроводники, но один тип проводимости по обе стороны от границы раздела (пример: *n*-AlGaAs/*n*-GaAs и *p*-AlGaAs/*p*-GaAs переходы)
Анизотипный (биполярный) переход	Один и тот же полупроводник, но различные типы проводимости с обеих сторон границы раздела (пример: p - n переход)	Различные полупроводники и различные типы проводимости с обеих сторон границы раздела (пример: p-AlGaAs/n-InP переход)

Примечание: GaAs читается как арсенид галлия, InP – фосфид индия, AlGaAs – тройное соединение алюминий-галлий-мышьяк.

1.2. Изотипный (униполярный) гомопереход

Рассмотрим униполярный гомопереход на примере n^+- n перехода (рис. 1.1 а).

Проведем мысленный эксперимент. Возьмем два образца одного и того же полупроводника, отличающиеся друг от друга степенью легирования. Для каждого из них концентрация ионов доноров равна концентрации электронов, т.е. образцы электронейтральны[1]. Соединим два полупроводниковых образца с различными концентрациями примесей так, чтобы кристаллические решетки образцов точно совпали, а на границе раздела не оказалось бы ни одного дефекта[2].

[1] В данной ситуации неосновные носители заряда можно не учитывать.

[2] Подобное механическое соединение на самом деле невозможно, так как на поверхности полупроводникового образца всегда присутствуют окислы, которые, будучи по своей природе диэлектриками, препятствуют протеканию тока. В процессе изготовления твердотельных приборов никакой «стыковки» кристаллических решеток не происходит, просто различные области единого монокристалла легируются разными примесями (подробнее о технологии роста кристаллов см. [8]).

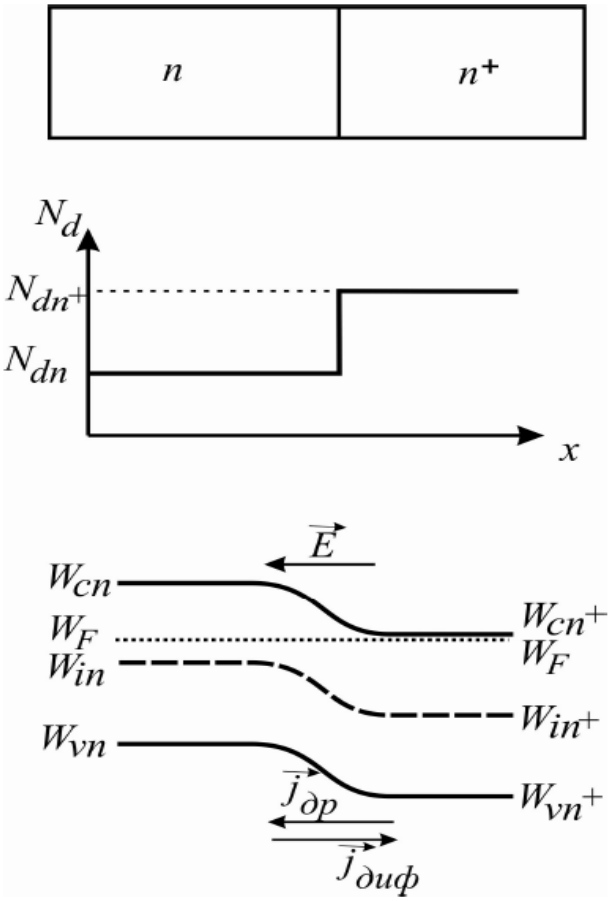


Рис. 1.1. Униполярный n^+- n переход в состоянии равновесия: а) схематическое изображение конструкции; б) график зависимости концентрации доноров от координаты; в) энергетическая зонная диаграмма. W_c, W_v – соответственно, уровни дна зоны проводимости и потолка валентной зоны; W_F – уровень Ферми; W_i – уровень Ферми в собственном полупроводнике (для простоты будем считать, что он лежит в середине запрещенной зоны). $\vec{E}$ – напряженность встроенного электрического поля; $\vec{j}_{др}$, $\vec{j}_{диф}$ – соответственно, плотности дрейфового и диффузионно токов в переходе. На рисунке вторым индексом в обозначениях величин указан тип полупроводника, к которому данные величины относятся. В дальнейшем подобные уточнения мы будем опускать

Так как концентрация электронов в сильно легированной n^+- области больше, чем в n – области (рис. 1.1. б), то возникнет диффузионный поток носителей заряда, который будет направлен справа налево (напомним, что диффузионный *ток* при этом направлен слева направо, поскольку направление тока выбирается по направлению движения положительных зарядов). Поскольку обе части полупроводника изначально нейтральны, то уход электронов из n^+ области означает, что там останется нескомпенсированный положительный заряд ионов доноров. Соответственно, в n-области будут копиться избыточные электроны. В итоге образуется *встроенное электрическое поле*, которое, в свою очередь, вызовет встречный дрейфовый ток носителей заряда. Сначала электроны будут преимущественно переходить из правой области в левую, затем внутреннее поле достигнет такой величины, при которой создаваемый им дрейфовый ток уравновесит ток диффузионный, и будет достигнуто *динамическое равновесие*. Зонная диаграмма перехода в равновесном состоянии показана на рис. 1.1 в.

Обратите внимание, что вектор напряженности электрического поля $\vec{E}$ направлен в сторону роста энергии дна зоны проводимости (в данном случае – от положительных зарядов ионов доноров, обнажившихся в n^+ слое, к отрицательным зарядам электронов, скопившимся в n-слое). Поскольку на дне зоны проводимости кинетическая энергия носителей $W_к=0$, то вклад в полную энергию дает только потенциальная составляющая $U(r)=-e\varphi(r)$, где e – абсолютная величина заряда электрона, $\varphi(r)$ - потенциал. Поскольку $\vec{E}=-\nabla\varphi$, то

вектор напряженности электрического поля будет сонаправлен с $\nabla U(r)$, а значит, направлен в сторону роста энергии дна зоны проводимости[3].

На практике чаще всего используют изотипные переходы с уровнями легирования от 10^{14} до 10^{18} см$^{-3}$. Высота потенциального барьера n^+-n^- - перехода зависит от разницы концентраций между областями перехода и имеет величину 1…10 kT, а электрическое сопротивление, обусловленное таким барьером, меньше сопротивления остальных переходов, которые рассматриваются ниже.

1.3. Анизотипный (биполярный) гомопереход (*p-n* переход)

Электронно-дырочным или p-n переходом называется приконтактная область между частями полупроводника с электронным (*n*) и дырочным (*p)* типами проводимости. Наиболее простой метод получения *p-n* переходов состоит во введении донорной и акцепторной примесей в процессе роста кристалла при эпитаксии, с помощью диффузии или ионного легирования.

В зависимости от характера распределения примесей различают резкий (ступенчатый) и плавный *p-n* переходы. Мы рассмотрим резкий *p-n* переход, в котором значения концентраций донорной N_d и акцепторной N_a примесей изменяются скачком на границе раздела.

1.3.1. Электронно-дырочный переход в равновесном состоянии

Снова проведем мысленный эксперимент: будем соединять в единое целое два полупроводниковых образца с различными типами проводимости так, что кристаллические решетки образцов точно совпадут, а на границе раздела не будет дефектов (рис.1.2).

Так как в *p*-области концентрация дырок (p_p) – основных носителей заряда – значительно больше, чем в *n*-области (p_n), то происходит

1) *диффузия*

дырок в *n*-область, где они окажутся неосновными носителями заряда. Таким образом, в некотором слое *n*-области, примыкающем к границе раздела, увеличивается концентрация неосновных носителей. Следовательно, там будет идти интенсивная

2) *рекомбинация*

электронов и дырок. Поскольку изначально *n*-область полупроводника была электронейтральной (концентрация электронов равнялась концентрации ионов

[3] Мы можем пользоваться соотношениями электростатики, т.к. скорость носителей заряда в полупроводниках на порядки меньше скорости света, т.е. можно считать, что при перемещении электронов и дырок поле мгновенно отслеживает изменение положения зарядов. Эта особенность характерна для полупроводниковых приборов, в отличие от приборов вакуумной электроники.

доноров), а электроны, оказавшиеся вблизи границы, рекомбинируют с дырками, пришедшими из *p*-слоя, то в приграничной области

3) *обнажится некомпенсированный объемный заряд*

положительного знака, обусловленный **ионами** донорной **примеси**. Аналогично, диффузия и рекомбинация электронов будут сопровождаться появлением в *p*-области отрицательного объемного заряда ионов акцепторной примеси (рис. 1.2 в). Таким образом, в приконтактной области появится встроенное электрическое поле, напряженность которого будет постепенно увеличиваться. В свою очередь, это поле вызовет встречный дрейфовый поток носителей заряда и будет препятствовать диффузионному движению. Очевидно, что когда диффузия будет полностью скомпенсирована дрейфом, система придет в состояние равновесия, а рост поля прекратится. При этом в полупроводнике установится определенная напряженность электрического поля и соответствующая ему *контактная разность потенциалов* $U_к$ (рис. 1.2 в).

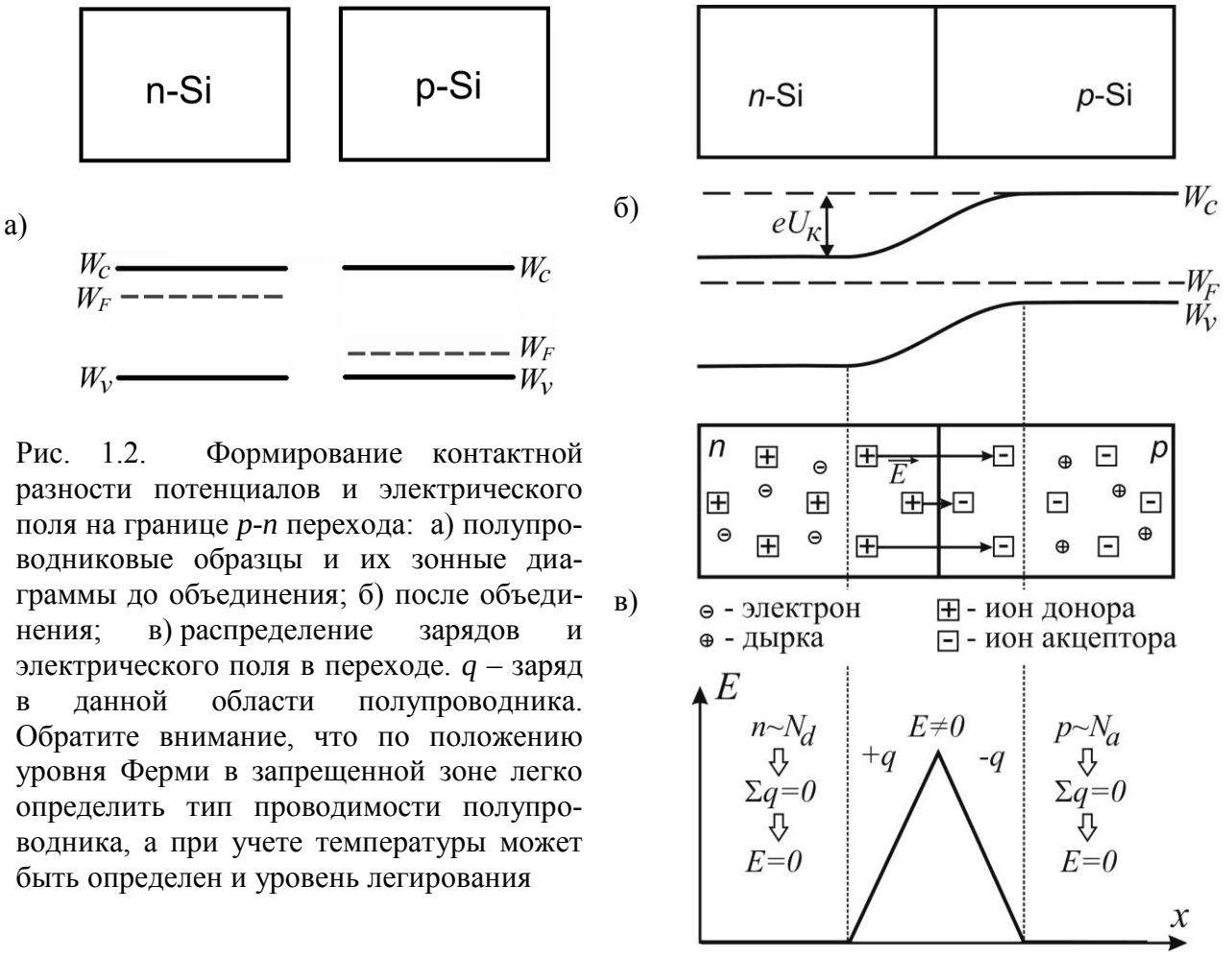


Рис. 1.2. Формирование контактной разности потенциалов и электрического поля на границе *p*-*n* перехода: а) полупроводниковые образцы и их зонные диаграммы до объединения; б) после объединения; в) распределение зарядов и электрического поля в переходе. q – заряд в данной области полупроводника. Обратите внимание, что по положению уровня Ферми в запрещенной зоне легко определить тип проводимости полупроводника, а при учете температуры может быть определен и уровень легирования

Подобное равновесие называют динамическим, поскольку движение носителей заряда прекращается лишь в среднем. Согласно распределению Больцмана, всегда найдутся высокоэнергетические носители заряда, которые смогут преодолеть имеющийся потенциальный барьер $eU_к$. Однако точно такое же количество носителей заряда перейдет в противоположную сторону под действием встроенного электрического поля, поэтому суммарный ток будет

равен нулю ($\Sigma j = 0$). При установлении равновесия происходит выравнивание термодинамических характеристик, в частности, уровней Ферми, в областях с различными типами проводимости, т.е. слева и справа от перехода. Приконтактную область, где имеется электрическое поле, называют *p-n* переходом, *запирающим слоем* или *областью пространственного заряда (ОПЗ)*.

Обратите внимание, что ионы примесей присутствуют во всем полупроводнике (доноры – в *n*-области, акцепторы – в *p*-области), а не только в самом переходе. Однако, вне ОПЗ образец электронейтрален (рис. 1.2.в).

На рис. 1.3 представлены распределения концентрации примеси (*а*), плотности объемного заряда (*б*), напряженности электрического поля (*в*) и потенциала (*г*) в резком несимметричном *p-n* переходе.

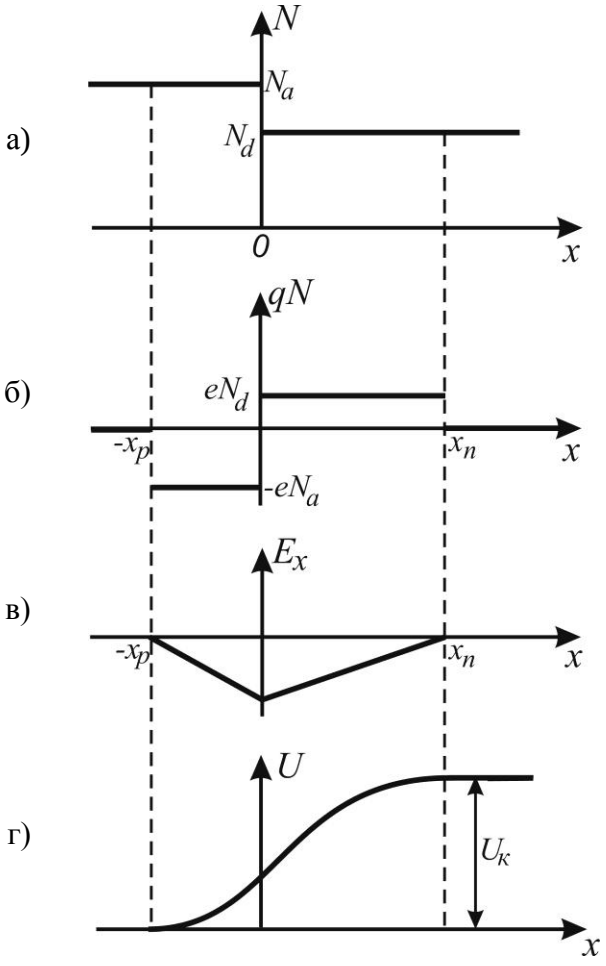


Рис. 1.3. Распределение концентрации примесей (а), плотности объемного заряда (б), поля (в) и потенциала (г) в *p-n* переходе. Обратите внимание, что на данном рисунке справа расположена область n-типа, а слева p-типа, так что электрическое поле, образованное ионами доноров и акцепторов, направлено справа налево ($E_x<0$).

Так как по условию задачи величины концентраций доноров и акцепторов различны, а их суммарные заряды в обедненных электронами и дырками областях равны друг другу ($eN_dx_n = eN_ax_p$), то размеры самих областей различны. Они тем меньше, чем выше уровень легирования, что является следствием закона сохранения заряда

В равновесном состоянии величины концентраций электронов и дырок в невырожденном примесном полупроводнике следующим образом зависят от температуры:

$$n = N_c e^{-\frac{W_c - W_F}{kT}}, \qquad p = N_v e^{-\frac{W_F - W_v}{kT}}. \tag{1.1}$$

Для собственного полупроводника эти концентрации равны:

$$n_i = p_i = N_c \cdot e^{-\frac{W_c - W_i}{kT}} = N_v \cdot e^{-\frac{W_i - W_v}{kT}} . \qquad (1.2)$$

Здесь W_c , W_V , W_i – уровни энергии, соответствующие нижнему уровню зоны проводимости, верхнему уровню валентной зоны, середине запрещенной зоны; W_F – уровень Ферми; N_c, N_V – эффективные плотности квантовых состояний в зоне проводимости и валентной зоне, соответственно; n_i – собственная концентрация носителей заряда.

Энергию середины запрещенной зоны W_{ip} и W_{in} для *p*- и *n*-типа полупроводника и энергию Ферми W_F можно записать через соответствующие потенциалы:

$$W_{ip} = -e\varphi_{ip}, \qquad W_{in} = -e\varphi_{in} , \qquad W_F = -e\varphi_F \qquad (1.3)$$

На основании соотношения (1.2) с учетом выражений (1.3) величины концентрации основных носителей вдали от *p-n* перехода могут быть выражены равенствами

$$p_p = n_i \cdot e^{\frac{\varphi_F - \varphi_{ip}}{\varphi_T}} , \qquad n_n = n_i \cdot e^{\frac{\varphi_{in} - \varphi_F}{\varphi_T}} , \qquad (1.4)$$

где $\varphi_T = \frac{kT}{e}$ - так называемый, тепловой потенциал. Равновесные концентрации основных (n_n, p_p) и неосновных (n_p, p_n) носителей заряда в невырожденных полупроводниках *p*- и *n*-типа связаны соотношением (так называемый, «закон действующих масс»):

$$n_p p_p = n_n p_n = n_i^2 = N_c \cdot N_v \cdot e^{-\frac{W_g}{kT}} ,$$

где W_g – ширина запрещенной зоны полупроводника.

Контактная разность потенциалов определяется соотношением:

$$U_к = \varphi_{in} - \varphi_{ip}. \qquad (1.5)$$

Выразим $U_к$ через равновесные концентрации электронов и дырок. Для этого из выражений (1.4) найдем φ_{ip} , φ_{in} и подставим в уравнение (1.5). После преобразований получим:

$$U_к = \varphi_T ln \frac{n_n p_p}{n_i^2} = \varphi_T ln \frac{N_d N_a}{n_i^2} = \varphi_T ln \frac{n_n}{n_p} = \varphi_T ln \frac{p_p}{p_n} \qquad (1.6)$$

или

$$p_n = p_p \cdot e^{-U_к / \varphi_T} , \qquad n_p = n_n \cdot e^{-Uк / \varphi_T} . \qquad (1.7)$$

Таким образом, высота потенциального барьера *p-n* перехода определяется температурой, концентрацией легирующей примеси и собственной концен-

трацией носителей при заданной температуре (т.е. типом материала). При этом, как видно из зонной диаграммы, величина контактной разности потенциалов в невырожденном *p-n* переходе ограничена значением ширины запрещенной зоны полупроводника W_g. Если же одна область легирована сильно, а другая – слабо, то $U_к < W_g/2$.

1.3.2. Вольт-амперная характеристика p-n перехода

Пусть к электронно-дырочному переходу подключен источник ЭДС таким образом, чтобы потенциальный барьер уменьшился. Такое подключение называется *прямым* и оно соответствует подсоединению источника плюсом к *p*-области и минусом к *n*-области. В этом случае из источника ЭДС в область *n*-типа поставляются электроны, а *p*-типа – дырки. Эти носители компенсируют заряд части доноров и части акцепторов, так что оставшийся в ОПЗ объемный заряд ионов становится меньше, что приводит к уменьшению напряженности электрического поля. В свою очередь, это приводит к тому, что по обе стороны от ОПЗ число основных носителей, которые имеют возможность преодолевать барьер, экспоненциально возрастает. Точно так же возрастает и величина диффузионного тока $j_{диф}$. При этом дрейфовый ток неосновных носителей заряда, направленный в противоположную сторону, практически не меняется, поскольку в сильном поле перехода имеет место режим насыщения дрейфовой скорости носителей заряда. Таким образом, при положительном подключении в *p-n* переходе преобладает диффузионный ток, т.е. $j = j_{диф} - j_{др} > 0$. При этом основные носители переходят в область с противоположным типом проводимости. Такой процесс называется *инжекцией.* Носители проникают через энергетический барьер в области, где они оказываются неосновными и рекомбинируют. Эти избыточные неравновесные носители нарушат электронейтральность полупроводника вблизи перехода и вызывают в равном количестве приток основных носителей из глубины *p*- и *n*-областей. Скорость рекомбинации электронов и дырок конечна, поэтому неравновесные носители могут продвинуться вглубь полупроводника, причем глубина их проникновения значительно превысит толщину ОПЗ. При этом электронейтральность кристалла за пределами области объемного заряда не нарушается.

Таким образом, при приложении внешнего напряжения в прямом направлении в результате инжекции носителей через *p-n* переход будет протекать ток, величина которого будет нарастать с увеличением приложенного напряжения, т.е. с уменьшением высоты барьера перехода.

При *обратном* включении внешнего напряжения (минусом к *p*-области и плюсом к *n*-области) направления внешнего и встроенного полей совпадают, т.е. высота барьера увеличивается. При этом диффузионный поток носителей спадает практически до нуля, а суммарный ток определяется дрейфовым током неосновных носителей заряда и незначителен по величине (поскольку количество неосновных носителей на несколько порядков меньше, чем основных).

Зонные диаграммы *p-n* перехода при разной полярности приложенного напряжения изображены на рис. 1.4.

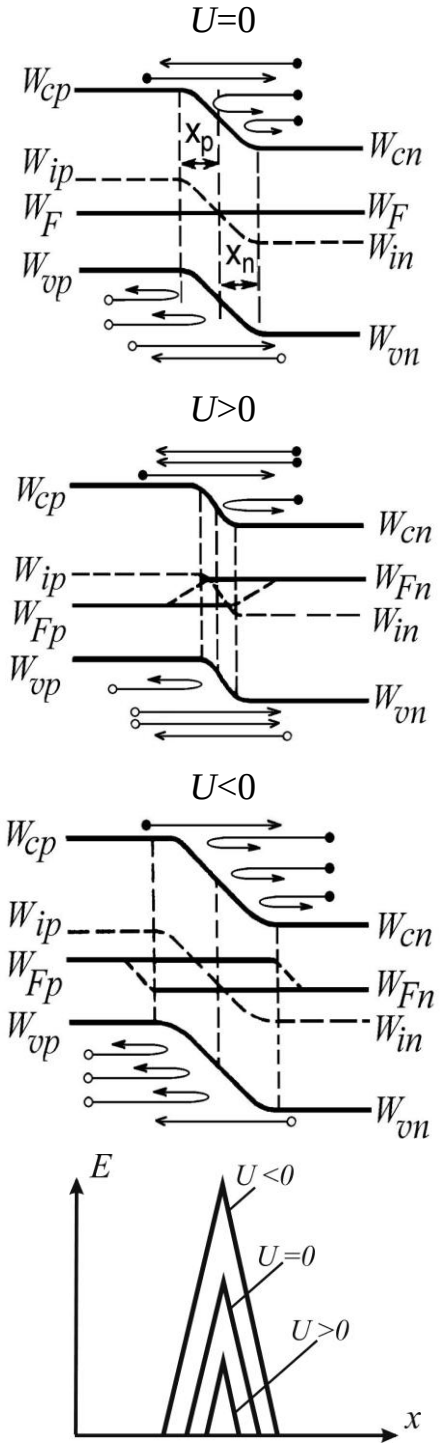


Рис. 1.4. Зонные диаграммы и распределение абсолютной величины напряженности электрического поля E в *p-n* переходе при различных внешних напряжениях U. Важно, что *при изменении внешнего напряжения происходит расширение или сужение области перехода*, так что в нее попадает большее или меньшее количество ионов доноров и акцепторов. Это определяет увеличение или уменьшение напряженности поля в переходе

При неравновесных условиях принято вводить два новых параметра распределения: W_{Fn} – для электронов и W_{Fp}- для дырок. Их значения выбирают таким образом, чтобы для концентраций электронов и дырок при наличии неравновесных носителей оставались справедливыми соотношения (1.1) и (1.4). Величины W_{Fn} и W_{Fp} называют квазиуровнями Ферми для электронов и дырок, соответственно. Таким образом, в невырожденных полупроводниках справедливы соотношения:

$$n = N_C e^{-\frac{W_c - W_{Fn}}{kT}}, \qquad p = N_V e^{-\frac{W_{Fp} - W_v}{kT}}, \tag{1.7а}$$

$$n = n_i \cdot e^{\frac{W_{Fn} - W_i}{kT}}, \qquad p = p_i \cdot e^{-\frac{W_{Fp} - W_i}{kT}}. \tag{1.7б}$$

Ввиду сложности строгого анализа, обычно делают еще ряд допущений, упрощающих решение задачи:

1. Модель электронно-дырочного перехода одномерная, *p*- и *n*-области имеют бесконечную протяженность.
2. В *p*- и *n* -областях примеси распределены равномерно, а на границе раздела значения их концентраций изменяются скачком.

3. Уровень инжекции мал (полагают малым внешнее напряжение).
4. Концентрация неосновных носителей мала по сравнению с концентрацией основных носителей. В этом случае ток вдали от *p*-*n*-перехода будет определяться основными носителями.
5. Электроны и дырки исчезают только вследствие рекомбинации друг с другом. Пренебрегают наличием ловушек для носителей заряда.
6. Генерация и рекомбинация в ОПЗ отсутствуют.
7. Явления, связанные с пробоем перехода, отсутствуют.

При этих допущениях для токов неосновных носителей вне ОПЗ можно записать:

$$j_p = -eD_p \frac{dp}{dx}, \qquad j_n = eD_n \frac{dn}{dx}, \tag{1.8}$$

где D_p, D_n – соответственно, коэффициенты диффузии дырок и электронов.

Одномерные уравнения непрерывности для дырок в *n*-области и электронов в *p*-области при отсутствии генерации можно записать в виде:

$$\frac{\partial p}{\partial t} = -\frac{1}{e} \cdot \frac{\partial j_p}{\partial x} - \frac{p - p_{n0}}{\tau_p}, \qquad \frac{\partial n}{\partial t} = \frac{1}{e} \cdot \frac{\partial j_n}{\partial x} - \frac{n - n_{p0}}{\tau_n}, \tag{1.9}$$

где τ_p и τ_n - времена жизни носителей заряда, p_{n0} и n_{p0} – значения концентрации равновесных дырок в *n*-области и равновесных электронов в *p*-области, соответственно.

При нахождении статической вольт-амперной характеристики необходимо решить уравнение непрерывности для случая, когда концентрация неосновных носителей не меняется во времени:

$$\frac{\partial p}{\partial t} = \frac{\partial n}{\partial t} = 0. \tag{1.10}$$

Из уравнений (1.9) с учетом (1.8) и (1.10) получим:

$$\begin{aligned} &\frac{d^2 p}{dx^2} - \frac{p - p_{n0}}{L_p^2} = 0 \text{ при } x \geq x_n, \\ &\frac{d^2 n}{dx^2} - \frac{n - n_{p0}}{L_n^2} = 0 \text{ при } x \leq -x_p, \end{aligned} \tag{1.11}$$

где $L_p = \sqrt{D_p \tau_p}$ - диффузионная длина дырок в *n*-области, $L_n = \sqrt{D_n \tau_n}$ - диффузионная длина электронов в *p*-области. Граничными условиями являются:

$$p\ (x \to \infty) \to p_{n0}\ ;\ n\ (x \to -\infty) \to n_{p0};$$

$$p(x_n) = p_{n1}\ ;\ n(-x_p) = n_{p1}.$$

Решая уравнения (1.11) при этих условиях, получим:

$$p(x) = p_{n0} + (p_{n1} - p_{n0})e^{-(x-x_n)/L_p},$$
$$n(x) = n_{p0} + (n_{p1} - n_{p0})e^{(x+x_p)/L_n}. \quad (1.12)$$

При приложении внешнего напряжения U в прямом направлении высота потенциального барьера становится равной $e(U_к\text{-}U)$. При этом концентрации неосновных носителей на границе ОПЗ на основании соотношения (1.7) будут выражаться в следующем виде:

$$p_{n1} = p_p \cdot e^{-\frac{e(U_k - U)}{kT}} = p_{n0}e^{\frac{eU}{kT}},$$
$$n_{p1} = n_n \cdot e^{-\frac{e(U_k - U)}{kT}} = n_{p0}e^{\frac{eU}{kT}}. \quad (1.13)$$

Подставляя (1.13) в (1.12), определим электронный и дырочный токи в точках $x=\text{-}x_p$ и $x=x_n$, соответственно:

$$j_n(-x_p) = eD_n \frac{dn}{dx} = \frac{eD_n n_{p0}}{L_n}\left(e^{\frac{eU}{kT}} - 1\right),$$
$$j_p(x_n) = -eD_p \frac{dp}{dx} = \frac{eD_p p_{n0}}{L_p}\left(e^{\frac{eU}{kT}} - 1\right). \quad (1.14)$$

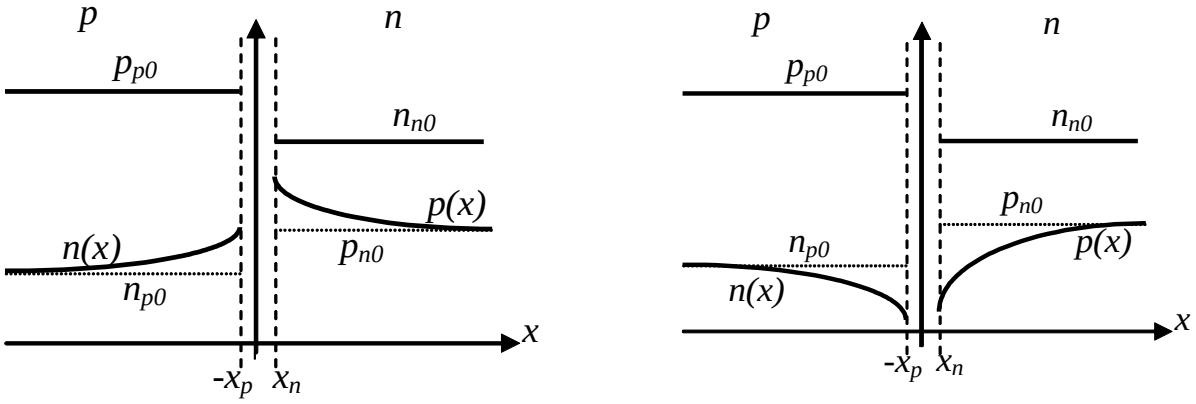
В предположении отсутствия генерации и рекомбинации в ОПЗ плотности токов j_n и j_p в интервале $\text{-}x_p<x<x_n$ не зависят от координаты, т.е.:

$$j_n(x_n) = j_n(-x_p); \qquad j_p(-x_p) = j_p(x_n); \quad (1.15)$$

Полный ток в стационарном режиме во всех сечениях одинаков. Проще всего вычислить ток на границах перехода в точках $x= \text{-}x_p$ или $x=x_n$. С учетом (1.15) плотность полного тока выражается соотношением:

$$j = j_n(-x_p) + j_p(x_n) = \left(\frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p}\right)\left(e^{\frac{eU}{kT}} - 1\right). \quad (1.16)$$

На рис. 1.5 показаны распределения концентраций электронов и дырок в p-n переходе при прямом и обратном смещениях.



а) б)

Рис. 1.5. Распределение величин концентраций электронов и дырок в *p*-*n* переходе: а) при прямом смещении; б) при обратном смещении

Если площадь перехода *S*, то полный ток через переход будет равен $J=jS$. Тогда вольт-амперную характеристику перехода можно записать в виде:

$$J = J_S\left(e^{\frac{U}{\varphi_T}} - 1\right). \tag{1.17}$$

Величина обратного тока перехода с увеличением обратного напряжения стремится к J_S, поэтому величину

$$J_S = \left(\frac{ep_{n0}L_p}{\tau_p} + \frac{en_{p0}L_n}{\tau_n}\right)S \tag{1.18}$$

называют *током насыщения* или *обратным током* p-n перехода. В (1.18) использованы соотношения: $L_n = \sqrt{D_n\tau_n}$, $L_p = \sqrt{D_p\tau_p}$.

Иногда, подчеркивая природу этого тока, его называют *током тепловой генерации* или просто *тепловым током*, т.к. обуславливающие его неосновные носители появляются в нейтральных *p*- и *n*-областях, прилегающих к переходу, за счет тепловой генерации. Эти носители диффундируют к границам перехода, захватываются его полем и переносятся в соседнюю область. Механизм образования теплового тока отражает формула (1.18), в которой p_{n0}/τ_p и n_{p0}/τ_n – скорости генерации неосновных носителей, а SL_pp_{n0}/τ_p и SL_nn_{p0}/τ_n - полное число неосновных носителей, генерируемых в слоях толщиной L_p и L_n за единицу времени. Именно эти носители без рекомбинации доходят до границы перехода и образуют обратный ток.

Можно рассматривать физический смысл выражения (1.18) и с другой стороны: отношения L_p/τ_p и L_n/τ_n имеют смысл скорости диффузии носителей заряда. Тогда формулу (1.18) для тока насыщения можно записать в классическом виде – как произведение заряда частиц, их концентрации, скорости и

площади сечения образца: $J_S = (ep_{n_0} V_p + en_{p_0} V_n)S$. Так как имеется ввиду скорость *диффузионного* движения носителей заряда в областях *за пределами пространственного заряда* перехода, то *ток не зависит* от величины обратного напряжения, подаваемого на переход, хотя напряженность поля перехода и ширина ОПЗ от внешнего напряжения зависеть будут. При попадании неосновных носителей в область перехода они будут подхвачены полем, и, независимо от величины последнего, переброшены на другую сторону ОПЗ. Таким образом, ток в данной ситуации будет ограничиваться сопротивлением областей сбора неосновных носителей и практически не будет зависеть от напряжения. Вольт-амперная характеристика диода, рассчитанная в рамках рассмотренной идеализированной модели, представлена на рисунке 1.10 сплошной линией.

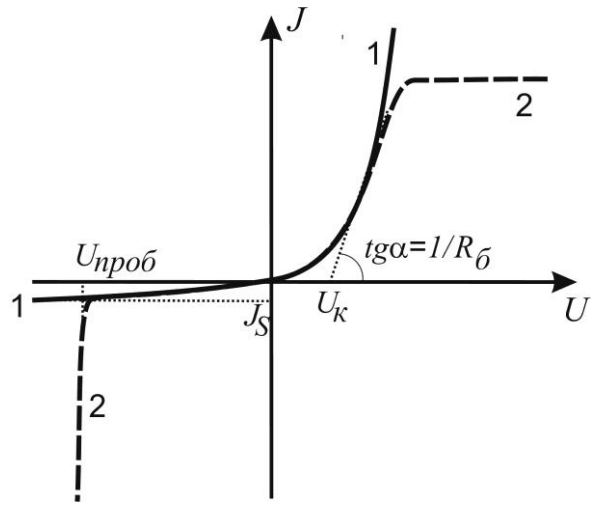


Рис. 1.6. ВАХ *p-n* перехода:
1) идеальная; 2) реальная.

Здесь $R_б$ – сопротивление базы, $U_к$ – контактная разность потенциалов, J_S – обратный ток *p-n* перехода, $U_{проб}$ – напряжение пробоя

Выражение (1.17) хорошо согласуется с экспериментальными данными для диодов, изготовленных из узкозонных полупроводников, например, германия. Можно показать, что с возрастанием энергии запрещенной зоны условие 6 перестает выполняться. Поэтому кремниевые и арсенидгаллиевые диоды не имеют насыщения на обратной ветви вольтамперной характеристики. Отличия от теоретической зависимости наблюдаются при увеличении прямого тока и при достаточно больших обратных смещениях, когда имеет место резкое возрастание обратного тока перехода. Рассмотрим обе ситуации.

У реального диода последовательно с сопротивлением *p-n* перехода имеются сопротивления квазинейтральных областей полупроводника. Область с большей концентрацией примесей (а, следовательно, с большей концентрацией свободных носителей и меньшим сопротивлением) называют эмиттером, а с меньшей – базой; сопротивление эмиттера обычно мало и им пренебрегают. При выводе формулы (1.17) мы неявно предполагали, что все прикладываемое внешнее напряжение падает непосредственно на области перехода. Однако, это справедливо только при малых токах, т.е. при отрицательных и небольших положительных смещениях ($U \leq U_к$). В этих случаях переход существенно обеднен свободными носителями заряда, а сопротивление ОПЗ велико. Соответственно, и падение напряжения на этой области много больше, чем на

других. При больших прямых токах падение напряжения на базе уже соизмеримо с падением напряжения на переходе. С учетом сопротивления базы аналитическое выражение зависимости тока диода от приложенного к нему напряжения может быть представлено в следующем виде:

$$J = J_S \left[e^{\left(U/\varphi_T - JR_б/\varphi_T\right)} - 1 \right], \tag{1.19}$$

где U - напряжение, приложенное к диоду, $R_б$ - сопротивление базы.

Проведя логарифмирование и дифференцирование выражения (1.19), определим дифференциальное сопротивление в произвольной точке вольт-амперной характеристики:

$$R_д = \frac{dU}{dJ} = \frac{1}{\frac{dJ}{dU}} = \frac{\varphi_T}{(J + J_S)} + R_б \ . \tag{1.20}$$

Видно, что при малых токах первое слагаемое в формуле (1.20) много больше второго, т.е. сопротивление диода в этом случае определяется сопротивлением самого *p-n* перехода (в этой области токов реальная и идеальная ВАХ совпадают). При больших токах дифференциальное сопротивление перехода мало и общее сопротивление определяется сопротивлением базы, т.е. зависимость тока от напряжения представляет собой прямую линию, тангенс угла наклона которой равен $1/R_б$. При дальнейшем увеличении прямого напряжения ток прибора выходит на насыщение. Причины этого явления зависят от конструкции прибора и определяются насыщением зависимости скорости носителей заряда в сильных электрических полях, малой концентрацией носителей заряда при слабом легировании полупроводника и разогревом полупроводника протекающим током.

Резкое возрастание обратного тока при увеличении напряжения выше критического, называемого напряжением пробоя $U_{проб}$, может быть вызвано следующими эффектами:

- туннелированием электронов сквозь узкий и высокий потенциальный барьер из валентной зоны в зону проводимости при большой напряженности электрического поля в ОПЗ (эффект Зинера или *туннельный пробой*);
- ударной ионизацией атомов полупроводника в сильном электрическом поле ОПЗ (*лавинный пробой*);
- перебросом электронов из валентной зоны в зону проводимости за счет саморазогрева (*тепловой пробой*).

Преобладание того или иного из упомянутых механизмов зависит от материала полупроводника, конструкции диода и температуры. *Туннельный пробой* наблюдается, как правило, в вырожденных полупроводниках. Толщина ОПЗ в них настолько мала (~10 нм), что при высоком значении напряженности поля в переходе (а значит, при сильном «наклоне» зон) становятся возможными туннельные переходы электронов с занятых состояний в валентной зоне на

свободные состояния зоны проводимости. При этом величина тока экспоненциально возрастает.

Лавинный пробой развивается в случае, когда поле в полупроводнике (или в приборной структуре) настолько велико, что на длине свободного пробега носители набирают энергию, достаточную для ударной ионизации атомов вещества. При этом налетающий горячий электрон, который приобрел энергию больше ширины запрещенной зоны, «выбивает» электрон из валентной зоны. В результате возникает пара противоположно заряженных частиц (электрон и дырка), одна или обе из которых также начинают участвовать в ударной ионизации вместе с исходным электроном. В такой ситуации происходит лавинообразное нарастание числа участвующих в процессе носителей, откуда и название данного типа пробоя.

Наконец, в случае протекания сильного тока полупроводниковая структура разогревается, а, значит, растет и количество термогенерированных носителей заряда, что, в свою очередь, приводит к дальнейшему росту тока. В результате развивается *тепловой пробой*, который приводит к разрушению кристаллической решетки полупроводника и выходу прибора из строя. Отметим, что туннельный и лавинный пробой сами по себе обратимы, однако, нельзя допускать их перехода в пробой тепловой. Величина напряжения пробоя у германиевых и кремниевых диодов с электронно-дырочными переходами может достигать сотен и даже тысяч вольт.

Получим выражение для связи напряжения на диоде с максимальной напряженностью поля в переходе, обуславливающей лавинный пробой. Учтем, что поле при равномерном распределении легирующей примеси зависит от координаты линейно, т.е. график имеет вид треугольника, а потенциал имеет смысл площади под графиком поля. Из геометрической формулы для площади треугольника имеем:

$$E_{\max} = \frac{2(U_к + U)}{d}. \qquad (1.20\ а)$$

Здесь U – абсолютная величина обратного смещения, т.е. $U>0$; d - ширина области пространственного заряда перехода

$$d = \sqrt{\frac{2(U_к + U)\varepsilon\varepsilon_0}{eN_d}}, \qquad (1.20\ б)$$

где ε – диэлектрическая проницаемость полупроводника, ε_0 – диэлектрическая постоянная.

Тогда максимальное поле в переходе

$$E_{\max} = \sqrt{\frac{2eN_d(U_к + U)}{\varepsilon\varepsilon_0}}. \qquad (1.20\ в)$$

Для того, чтобы в структуре начал развиваться лавинный пробой, необходимо, чтобы найденное максимальное значение электрического поля в переходе превысило некоторую определенную для заданного материала величину, т.е. стало бы больше напряженности *поля пробоя* $E_{проб}$. Для оценки этой величины вспомним, что при возникновении лавинного пробоя электрон на длине свободного пробега набирает энергию, необходимую для разрыва валентной связи, т.е. энергию, равную ширине запрещенной зоны W_g. Далее элементарная оценка на основе второго закона Ньютона с учетом эффективной массы электрона и характерного времени свободного пробега $\tau_{своб} \sim 10^{-13}$ с показывает, что электроны будут набирать указанную энергию при напряженности поля $E_{проб} \sim 10^5$ В/см, что для перехода с $N_d \approx N_a \approx 10^{17}$ см$^{-3}$ соответствует напряжению ~ 10 В. Пробой будет развиваться там, где локализовано максимальное поле, т.е. на границе *p*-*n*-перехода. Область, в которой реализуется ударная ионизация, будет иметь толщину порядка длины свободного пробега носителей заряда. При росте внешнего напряжения толщина этой области увеличивается.

В заключение данного раздела отметим два момента, часто вызывающих вопросы у студентов при первом знакомстве с физикой работы полупроводниковых диодов.

1) Обратите внимание, что *«изогнуть» зонную диаграмму «в другую сторону» при подаче прямого смещения нельзя*! Это означает, что даже при подаче большого «открывающего» диод напряжения невозможно реализовать ситуацию, когда уровни энергии электронов в *p*-области стали бы ниже соответствующих уровней в *n*-области. Все подаваемое на диод смещение распределяется между запирающим слоем и базой. Чем большее напряжение мы подаем, тем выше ток, а значит, тем большая часть внешнего смещения падает на области базы и тем меньшая – на области самого перехода. В пределе можно предполагать «выглаживание» потенциального барьера перехода, но не его инверсию.

2) *При увеличении температуры прямой и обратный токи диода будут расти.* Физически это объясняется следующими причинами:

а) увеличение температуры приводит к росту концентрации носителей заряда из-за тепловой генерации;

б) при увеличении температуры уменьшается контактная разность потенциалов. Это легко понять из простых соображений. С ростом температуры уровни Ферми *p* и *n* областей будут стремиться к середине запрещенной зоны полупроводника, т.е. барьер между частями диода, имеющими разный тип проводимости, будет снижаться.

Таким образом, в формуле для ВАХ диода (1.17) основное влияние на зависимость результирующего тока от температуры оказывают $U_к(T)$ и $J_S(T)$, а не $\varphi_T = kT/e$, как может показаться на первый взгляд.

1.3.3. Емкость электронно-дырочного перехода

Всякий *p-n* переход, по существу, представляет собой систему двух полупроводниковых слоев, разделенных областью объёмного заряда. Такая система подобна плоскому конденсатору. Опыт показывает, что полупроводниковые диоды обладают значительной емкостью, которая ограничивает их применение в высокочастотных приборах. Изучение емкости перехода во многих случаях помогает исследовать распределения поля и ионов доноров и акцепторов в ОПЗ, а также более детально объяснить механизм протекания тока.

Найдем ширину потенциального барьера и емкость *p-n* перехода. Помимо допущений, принятых при выводе вольт-амперной характеристики, будем считать, что вся примесь ионизирована и в области объёмного заряда концентрация свободных носителей равна нулю. В этом случае плотность объемного заряда постоянна и определяется только значениями концентраций соответствующих примесей, т.е.

$$\rho_p = -eN_a, \quad \rho_n = eN_d . \tag{1.21}$$

Этот объемный заряд создает электрическое поле, которое проникает в *p*-область на глубину x_p, а в *n*-область - на x_n. Вне слоя пространственного заряда напряженность электрического поля равна нулю, т.е.

$$E_{p_x} = -\left.\frac{d\varphi}{dx}\right|_{x=-x_p} = 0; \quad E_{n_x} = -\left.\frac{d\varphi}{dx}\right|_{x=x_n} = 0 . \tag{1.22}$$

Распределение электростатического потенциала φ в области объемного заряда можно найти, решив уравнение Пуассона:

$$\frac{d^2\varphi}{dx^2} = -\frac{\rho}{\varepsilon\varepsilon_0} , \tag{1.23}$$

где ε - диэлектрическая проницаемость полупроводника.

Интегрируя это уравнение с учетом граничных условий (1.22), получим выражение для напряженности электрического поля в *p*- и *n*-областях:

$$\begin{aligned} E_{p_x} &= -\frac{d\varphi_p}{dx} = -\frac{e}{\varepsilon\varepsilon_0} N_a (x + x_p) \text{ при } -x_p < x < 0, \\ E_{n_x} &= -\frac{d\varphi_n}{dx} = -\frac{e}{\varepsilon\varepsilon_0} N_d (x_n - x) \text{ при } 0 < x < x_n \end{aligned} \tag{1.24}$$

В точке *x=0* из условия непрерывности поля следует, что

$$E_p(0) = E_n(0) . \tag{1.25}$$

Из последнего равенства с учетом (1.24) получим:

$$eN_a x_p = eN_d x_n. \tag{1.26}$$

Это соотношение выражает равенство положительного и отрицательного заряда в ОПЗ, т.е. условие электронейтральности образца.

При приложении к *p-n* переходу внешнего напряжения в соответствии с допущением об отсутствии электрического поля вне слоя объёмного заряда граничные условия для потенциала можно представить следующим образом:

$$\varphi(-x_p) = 0 \quad \text{и} \quad \varphi(x_n) = U_к - U. \tag{1.27}$$

Величина напряжения U в формуле (1.27) положительна при прямом смещении перехода и отрицательна – при обратном.

Интегрируя уравнения (1.24) с учетом граничных условий (1.27), получим

$$\varphi_p = \frac{eN_d}{2\varepsilon\varepsilon_0}(x_p + x)^2 \quad \text{при } -x_p < x < 0;$$

$$\varphi_n = -\frac{eN_d}{2\varepsilon\varepsilon_0}(x_n - x)^2 + U_k - U \quad \text{при } 0 < x < x_n. \tag{1.28}$$

Так как потенциал φ в пределах ОПЗ должен быть непрерывен, то $\varphi_p(0)=\varphi_n(0)$, т.е.

$$\frac{e}{2\varepsilon\varepsilon_0}\left(N_d x_n^2 + N_a x_p^2\right) = U_к - U. \tag{1.29}$$

Используя соотношение (1.29) и равенство (1.26), получим выражение для ширины области объёмного заряда:

$$d = x_n + x_p = \sqrt{\frac{2\varepsilon\varepsilon_0(N_a + N_d)(U_к - U)}{eN_a N_d}}. \tag{1.30}$$

Как видно из выражения (1.30), ширина ОПЗ уменьшается с увеличением прямого (положительного) напряжения и увеличивается при обратном напряжении.

Изменение ширины области объёмного заряда в связи с изменением напряжения приводит к изменению заряда в *p*- и *n*-областях. Поэтому *p-n* переход ведет себя подобно емкости. Эту емкость называют барьерной, т.к. она связана с образованием потенциального барьера между *p*- и *n*-областями.

$$C = \frac{dQ}{dU} = S\sqrt{\frac{e\varepsilon\varepsilon_0 N_a N_d}{2(N_a + N_d)(U_к - U)}} = S\frac{\varepsilon\varepsilon_0}{d}, \tag{1.31}$$

где S - площадь перехода. Обратите внимание, что формула (1.31) совпадает с выражением для емкости плоского конденсатора, хотя в отличие от последнего, в *p-n* переходе заряд является объемным.

В случае резко несимметричного *p-n* перехода (когда концентрация легирующей примеси в одной из частей много больше, чем в другой) переход заключен, в основном, в области со слабым легированием. Например, при $N_a >> N_d$, переход простирается в *n*-область, а величина барьерной емкости не зависит от свойств *p*-области:

$$C = S\sqrt{\frac{e\varepsilon\varepsilon_0 N_d}{2(U_к - U)}} \ . \qquad (1.32)$$

Выражение (1.32) позволяет найти контактную разность потенциалов и концентрацию донорной примеси. График зависимости $\frac{1}{C^2} = f(U)$, изображенный на рис. 1.7, отсекает на оси абсцисс отрезок, равный по величине $U_к$. Если известна зависимость *C=f(U),* то на основании равенства (1.31) можно построить зависимость ширины обеднённой области от приложенного напряжения.

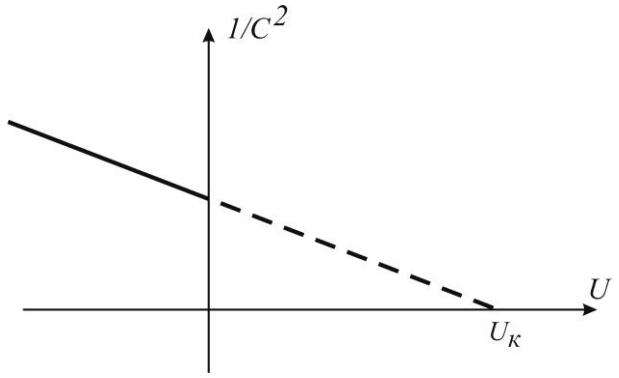


Рис. 1.7. Вольт-фарадная характеристика *p-n* перехода. Продолжив зависимость в область прямых смещений, можно определить контактную разность потенциалов

Барьерная емкость, рассмотренная выше, вносит основной вклад в емкость перехода при обратном смещении. При прямом смещении в емкости преобладает диффузионная емкость, обусловленная изменением распределения концентрации неосновных носителей заряда. По сути, перераспределение концентрации электронов и дырок является перезарядкой обкладок эквивалентного конденсатора. Такая перезарядка будет происходить тем интенсивнее, чем большее прямое смещение приложено к переходу.

Исходя из зависимости (1.13) при прямом смещении может быть получено выражение для диффузионной емкости перехода:

$$C_{dif} = \frac{e^2}{2kT}(L_p p_{n0} + L_n n_{p0})\exp(\frac{e(U + Uк)}{kT}) . \qquad (1.33)$$

При повышении частоты переменного сигнала емкость не успевает перезаряжаться полностью. Величина диффузионной емкости с увеличением частоты сигнала уменьшается по закону $\sim f^{-1/2}$. Вместе с тем диффузионная емкость быстро возрастает с ростом постоянного тока $\sim \exp(\frac{e(U + U_к)}{kT})$. По этим причинам диффузионная емкость играет особенно большую роль на низких частотах и при прямом смещении сравнимом с $U_к$.

1.4. Анизотипный (биполярный) и изотипный (униполярный) гетеропереходы

Гетеропереходы образуются между различными по составу полупроводниками. На рис. 1.8 показана зонная диаграмма перехода между электронным и дырочным полупроводниками с разной шириной запрещенной зоны. Для примера показана гетероструктура, в которой электронный полупроводник имеет большую ширину запрещенной зоны, чем дырочный – *p*-*N*-переход (в литературе более широкозонный полупроводник часто обозначается заглавной буквой).

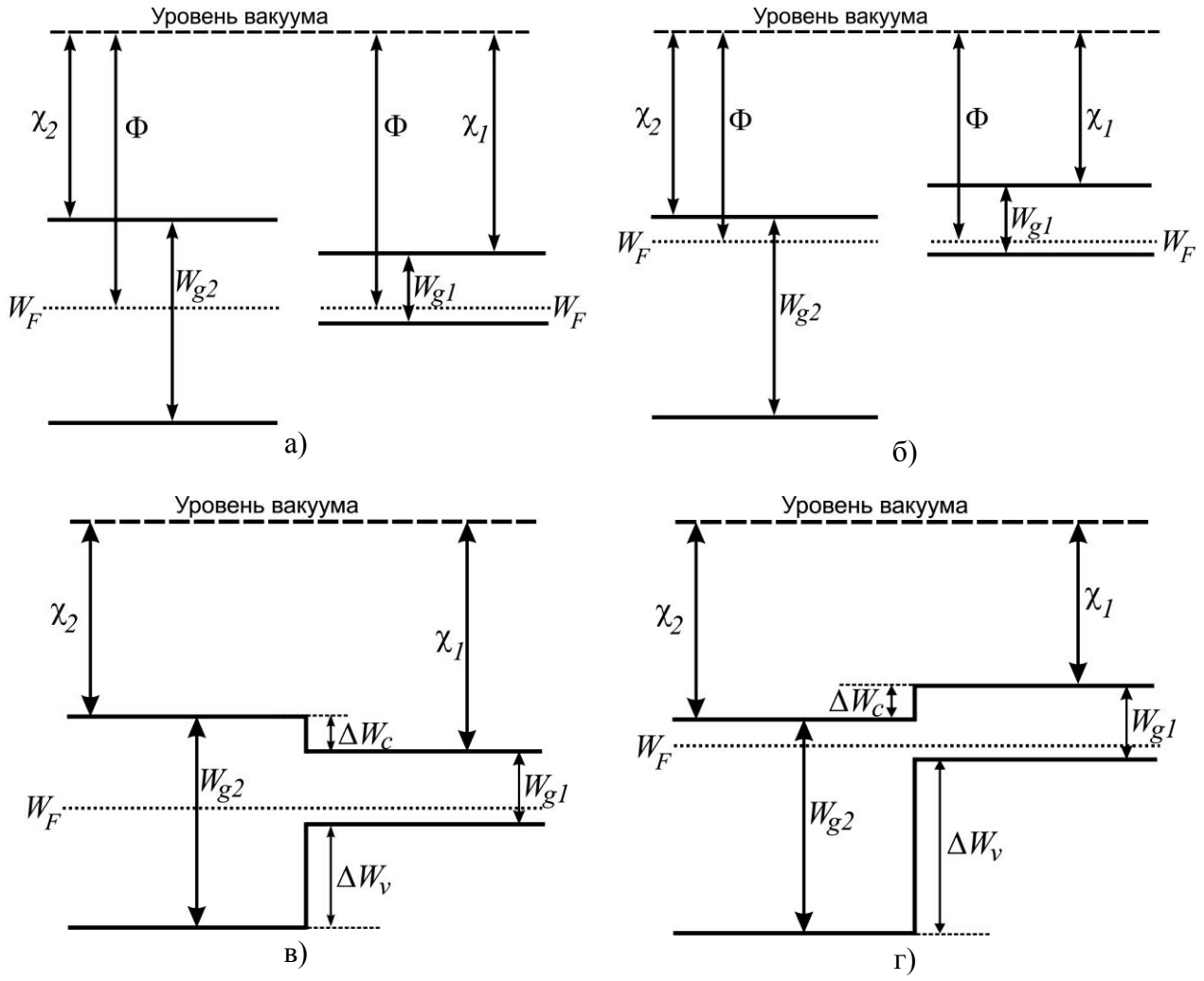


Рис. 1.8. Зонные диаграммы а), б) двух изолированных полупроводников *N* и *p* типов с одинаковыми работами выхода $\Phi_1=\Phi_2=\Phi$; в), г) гетеро- *p*-*N* переходов при условии равенства работ выхода материалов

В отличие от гомоструктурного *p*-*n* перехода, на зонной диаграмме гетероструктурного перехода имеются «*разрывы*», т.е. резкие перепады дна зоны проводимости и/или потолка валентной зоны, величины которых *определяются соотношениями электронного сродства материалов и их ширины запрещенной зоны*. Согласно модели Андерсона, величина «разрыва» дна зоны проводимости

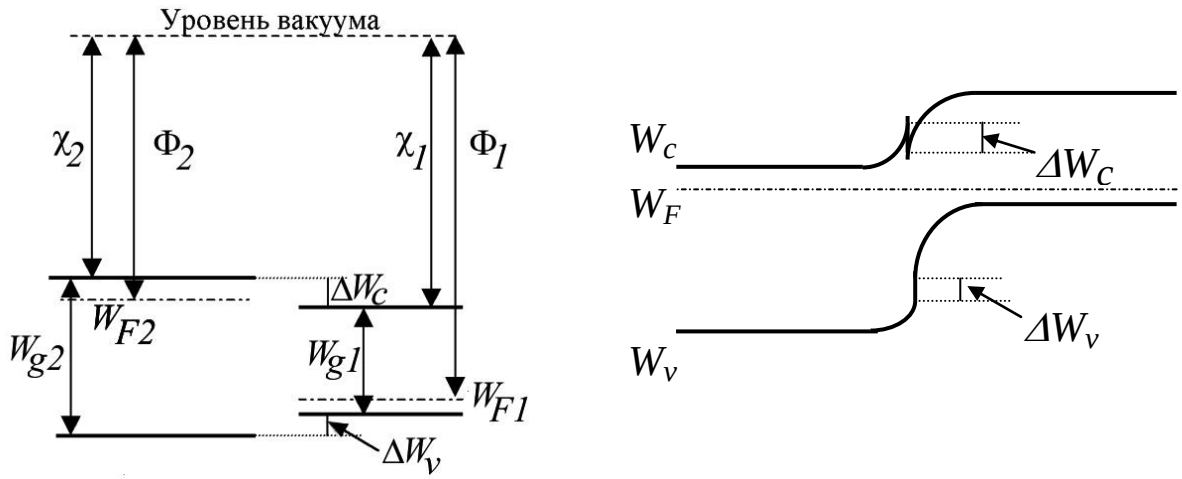
$$\Delta W_c = \chi_1 - \chi_2 , \qquad (1.34)$$

где χ_1 и χ_2 - электронные сродства контактирующих полупроводников. Соответствующий «разрыв» потолка валентной зоны

$$\Delta W_v = \Delta W_g - \Delta W_c , \qquad (1.35)$$

где ΔW_g - разность ширины запрещенных зон полупроводников. На рисунке 1.12 а, б для примера приведены варианты зонных структур, которые могут наблюдаться при одинаковых значениях работы выхода, но различных соотношениях между шириной запрещенной зоны и электронным сродством материалов *p* и *N*-типа.

В реальных ситуациях полупроводниковые слои гетеропереходов отличаются между собой не только электронным сродством и шириной запрещенной зоны, но и работой выхода[4]. В этом случае, как и в гомоструктурном *p-n* переходе, будет происходить перераспределение зарядов и появление *встроенного электрического поля* вблизи границы раздела слоев, т.е. формирование ОПЗ. На зонной диаграмме это отразится появлением изгиба зон (рис. 1.9).



а) б)

Рис. 1.9. Зонные диаграммы а) двух изолированных полупроводников *N* и *p* типов с разными работами выхода Φ_1 и Φ_2 (уровень вакуума принят за начало отсчета энергии); б) *p-N* перехода

При этом, поскольку непосредственно на границе слоев должны остаться разрывы дна зоны проводимости и потолка валентной зоны определенной величины, на зонной диаграмме может появиться характерный пичок (рис. 1.9). Отметим, что в зависимости от конкретной комбинации значений электронного сродства, уровней легирования и ширины запрещенной зоны материалов

[4] Напомним, что под работой выхода понимается энергия, которую необходимо затратить для выхода электрона из твердого тела в вакуум в состояние с кинетической энергией, равной нулю. При этом предполагается, что электрон находился на уровне Ферми. Таким образом, работа выхода в полупроводниках определяется электронным сродством материала и степенью его легирования, поскольку последняя задает расстояние между дном зоны проводимости и положением уровня Ферми. Обратите внимание, что в случае невырожденного полупроводника уровень Ферми расположен в запрещенной зоне, где электронов нет.

гетероструктуры подобный пичок может образоваться в зоне проводимости, в валентной зоне, в обеих зонах или ни в одной их них.

Сходство гетеро- и гомо- *p-n* переходов заключается в том, что электрическое поле в ОПЗ образовано зарядами ионов доноров и акцепторов (для изотипного гетероперехода электрическое поле будет образовано ионами доноров и электронами; см. n^+- *n* переход). Их отличие в том, что на границе раздела гетероперехода химические связи между атомами кристаллической решетки будут напряженными из-за того, что состав материала в этом месте меняется скачком. Поскольку связь представляет собой взаимодействие электронных облаков и ядер атомов, то напряжённая химическая связь обуславливает наличие сильного электрического поля на расстояниях порядка межатомных, составляющих единицы ангстрем. Это сильное поле обуславливает разрывы дна зоны проводимости и потолка валентной зоны.

По аналогии с гомоструктурным *p-n* переходом, ширина областей пространственного заряда гетероперехода при отсутствии внешнего напряжения определяется соотношениями:

$$x_n = \sqrt{\frac{2N_d\varepsilon_1\varepsilon_2(\varphi_2-\varphi_1)}{eN_a(\varepsilon_2N_d+\varepsilon_1N_a)}}\ ,\ x_p = \sqrt{\frac{2N_a\varepsilon_1\varepsilon_2(\varphi_2-\varphi_1)}{eN_d(\varepsilon_2N_d+\varepsilon_1N_a)}}, \qquad (1.36)$$

а емкость гетероперехода - соотношением:

$$C = S\sqrt{\frac{e\varepsilon_1\varepsilon_2N_aN_d}{2(\varepsilon_1N_a+\varepsilon_2N_d)(\varphi_1-\varphi_2-U)}}. \qquad (1.37)$$

Аналогично гомоструктурному *p-n* переходу, ВАХ гетероструктурного *p-n* перехода имеет вид:

$$J = J_S\left(e^{\frac{eU}{AkT}}-1\right), \qquad (1.38)$$

где A – безразмерный коэффициент, характеризующий величину разрыва зон гетероструктуры. Для гомоструктурного *p-n* перехода $A = 1$.

На рис. 1.9 хорошо видно, что высота барьера для электронов и дырок различна. Резкий подъем потолка валентной зоны образует больший гетеробарьер для дырок, чем барьер для электронов в зоне проводимости. В этом случае, отношение электронного и дырочного токов будет иметь вид:

$$\frac{J_{S_n}}{J_{S_p}} = \frac{n_pV_n}{p_nV_p}\exp\left(\frac{\Delta W_V}{kT}\right). \qquad (1.39)$$

Часто $\Delta W_V \approx 10...20kT$, т.е. высота барьера для дырок существенно больше их средней энергии $\frac{3}{2}kT$. Поскольку распределение частиц по энергии носит экспоненциальный характер, то в итоге в подобных переходах электронный ток

на несколько порядков превышает дырочный. Такое явление называют *униполярной инжекцией*.

Рассмотрим теперь *варизонный гетеропереход* (рис. 1.10), т.е. такой гетеропереход в котором химический состав полупроводникового материала изменяется плавно. Например, в структуре $AlAs\text{-}Al_yGa_{1-y}As\text{-}GaAs$ химический состав среднего слоя плавно меняется так, что показатель соотношения Ga и Al в тройном соединении зависит от координаты $y = y(x)$. Это позволяет плавно изменять ширину запрещённой зоны среднего слоя структуры. Этот эффект используется в некоторых приборах для увеличения скорости носителей заряда.

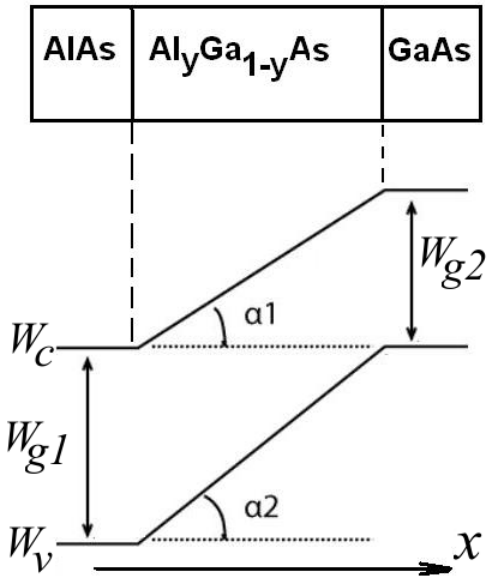


Рис. 1.10. Зонная диаграмма варизонного гетероперехода. В области, где ширина запрещенной зоны меняется, химический состав полупроводникового материала изменяется *плавно* в зависимости от координаты.

Анизотипный и изотипный варизонные переходы отличаются между собой типом зарядов, которые создают дополнительное электрическое поле в переходе. В первом случае поле создается ионами доноров с одной стороны и акцепторов – с другой, а во втором – ионами и свободными носителями заряда противоположного знака (либо донорами и электронами, либо акцепторами и дырками).

Вернемся теперь к резкому гетеропереходу, когда химический состав полупроводника сильно меняется на длинах порядка межатомного расстояния. Чтобы избежать дефектов кристаллической решетки, подберем химический состав слоев таким образом, чтобы постоянные кристаллических решеток совпадали, как, например, в гетеропереходе $GaAs/Ga_{0.7}Al_{0.3}As$.

На рис. 1.11 приведено распределение легирующей примеси и зонная диаграмма для *селективно легированного гетероперехода*. Такой гетеропереход специально изготавливается так, чтобы скачок в распределении концентрации доноров по координате не совпадал с границей гетероперехода, где реализуется разрыв зон.

В силу того, что потенциальная яма на границе подобного гетероперехода достаточно глубока и имеет размеры порядка длины волны электрона, происходит *квантование электронных уровней*. В такой ситуации энергия движения в поперечном слоям ямы направлении (на рисунке 1.11 - по оси *x*) может изме-

ниться только на определенную величину, а значит, рассеяние по данной координате подавлено.

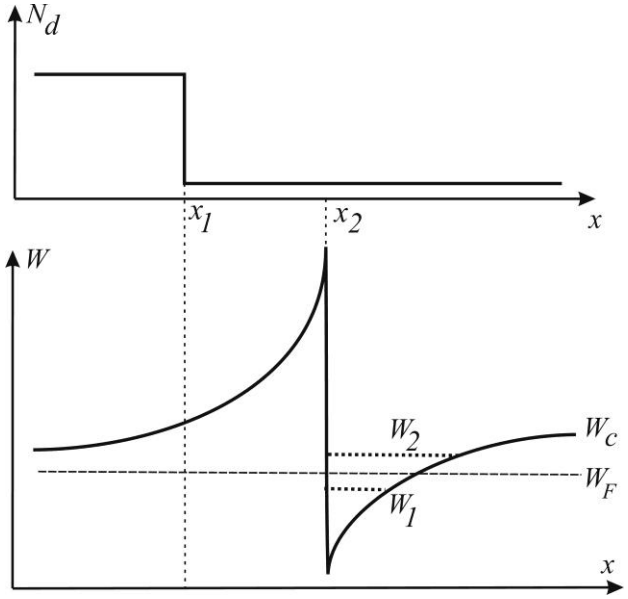


Рис. 1.11. Распределение легирующей примеси (вверху) и зонная диаграмма (внизу) для селективно легированного изотипного гетероперехода. Так как уровень Ферми расположен между двумя квантовыми уровнями W_1 и W_2, то нижний уровень будет заселен электронами значительно больше, чем верхний [3]

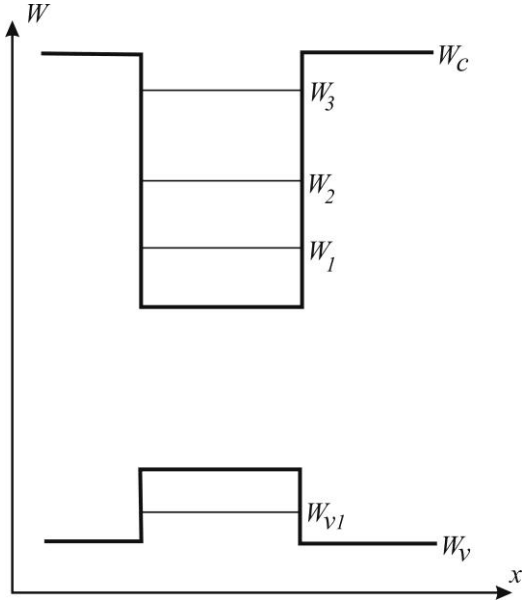


Рис. 1.12. Зонная диаграмма квантовой ямы, образованной двумя резкими гетеропереходами. Благодаря тому, что при изготовлении гетероперехода ширина среднего слоя может легко варьироваться, появляется возможность создания требуемой глубины квантовой ямы, т.е. регулировки количества и положения в ней энергетических уровней

Кроме того, поскольку в такой селективно легированной структуре область квантовой ямы изготовлена в материале с низким уровнем легирования, то рассеяние на ионах легирующей примеси практически отсутствует, а значит, подвижность электронов внутри квантовой ямы высока. Поэтому при движении в направлении, параллельном слоям ямы (по оси *y* или *z*) возможно достижение больших скоростей электронов, таких, как в нелегированном материале. Из-за того, что область сильно легированного материала изготавливается на расстояниях меньше диффузионной длины от квантовой ямы, концентрация электронов в яме будет высока. Подобный слой материала с высокой концентрацией носителей и одновременно с большой их подвижностью часто используется в качестве активной области полупроводниковых приборов, например, в полевых транзисторах с высокой подвижностью электронов (HEMT – high electron mobility transistor).

Другим вариантом образования квантовых ям является создание трёхслойной композиции из двух широкозонных и среднего узкозонного слоев полупроводника (рис. 1.12). Такая технология позволяет добиться реализации в одном полупроводниковом слое квантовой ямы, как для электронов, так и для дырок. Важно, что для реализации среднего слоя необязательно подбирать материал с той же постоянной решетки, как у крайних слоев. Из-за того, что средний слой тонкий, при образовании гетероструктуры он растягивается, так

что на границе раздела не возникают дефекты (псевдоморфный гетеропереход [3]). За счет растяжения возможно варьирование параметров квантовой ямы. Подобные структуры используются для создания оптоэлектронных приборов.

Для треугольной ямы (рис.1.11) потенциал может быть аппроксимирован зависимостью $V(x) = eE_s x$ [3], тогда, решая уравнение Шредингера:

$$\frac{\hbar^2}{2m^*}\frac{d^2\Psi_i}{dx^2} + (W_i - V(x))\Psi_i = 0, \tag{1.40}$$

где $\hbar$ – приведенная постоянная Планка, получаем распределение уровней в яме по энергии (уравнение Эйри):

$$W_i \approx \left(\frac{\hbar^2}{2m^*}\right)^{1/2}\left(\frac{3eE_s\pi(i+3/4)}{2}\right)^{2/3}, \tag{1.41}$$

где для вычисления электрического поля E_S необходимо решить уравнение Пуассона. Для прямоугольной ямы в приближении ее бесконечной высоты положение уровней вычисляется по традиционной формуле:

$$W_n = \frac{n^2\pi^2\left(\frac{\hbar}{d}\right)^2}{2m^*}. \tag{1.42}$$

Важно понимать, что речь идет о квантовании только по одной координате (в нашем случае – по оси *x*). При этом по двум другим направлениям (в плоскости *yz*) электроны двигаются классически, поэтому такой электронный газ принято называть *двумерным*. Классическое движение частиц, как и в обычном, трехмерном, случае, описывается эффективной массой, подвижностью и скоростью насыщения, соответствующей объемному полупроводнику. Дисперсионное соотношение в двумерном электронном газе выглядит следующим образом[5]:

$$W = W_c + W_n + \frac{p_z^2 + p_y^2}{2m^*}, \tag{1.43}$$

где W_c – энергия дна зоны проводимости, W_n – положение уровней квантования по отношению к дну зоны проводимости.

Кроме потенциальных ям с помощью гетеропереходов могут быть созданы потенциальные барьеры для электронов и дырок (рис.1.13).

[5] Предлагаем читателю подумать самостоятельно, в каких приближениях справедлива зависимость такого вида.

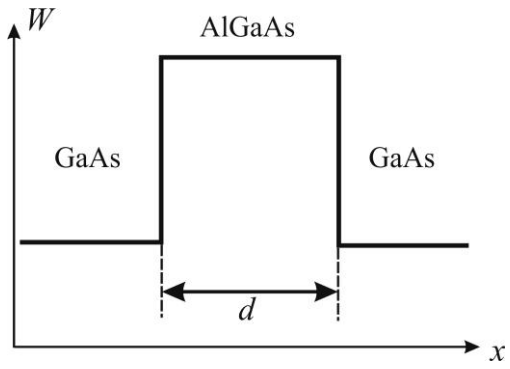


Рис. 1.13. Зонная диаграмма энергетического барьера, образованного двумя резкими гетеропереходами. Благодаря тому, что при изготовлении гетероперехода ширина среднего слоя легко варьируется, можно создавать барьеры с требуемым коэффициентом прохождения электронов за счет туннелирования

Коэффициент прохождения T через прямоугольный туннельно-прозрачный барьер высотой V_0 определяется толщиной барьера d, длиной волны туннелирующего электрона λ и его эффективной массой m^*:

$$T = \frac{1}{1 + \frac{(k_1^2 + k_2^2)^2}{4k_1^2 k_2^2} sh^2 k_2 d}, \qquad (1.44)$$

где $k_1 = \frac{\sqrt{2m^* W}}{\hbar}$, $k_2 = \frac{\sqrt{2m^*(V_0 - W)}}{\hbar}$, $\hbar$ – приведенная постоянная Планка.

Если $k_2 d >> 1$, то $sh(k_2 d) \approx exp(k_2 d)/2$. Тогда выражение (1.44) примет вид

$$T \approx T_0 \exp\left(- 2d \sqrt{\frac{2m^*(V_0 - W)}{\hbar^2}} \right), \qquad (1.45)$$

где $T_0 = \left(\frac{k_1 k_2}{k_1^2 + k_2^2} \right)^2$.

1.5. Структура металл - диэлектрик - полупроводник (МДП)

1.5.1. Идеальная МДП-структура

Рассмотрим структуру металл-диэлектрик-полупроводник (МДП), которая схематически изображена на рис. 1.14.

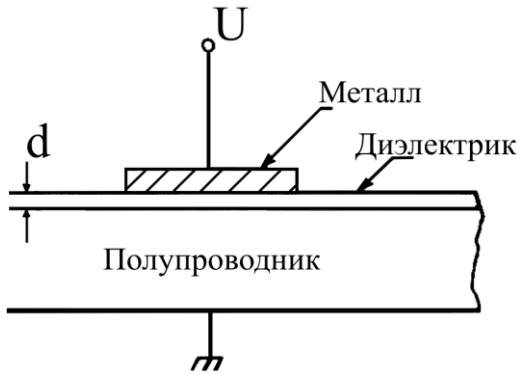


Рис. 1.14. Структура металл-диэлектрик-полупроводник (МДП - структура), где *d*-толщина слоя диэлектрика, а *U*- напряжение, приложенное к металлическому электроду

Зонные диаграммы идеальных МДП-структур при U=0 приведены на рис. 1.15. На рисунке введены следующие обозначения: $-e\varphi_M$ - работа выхода из металла, χ - сродство к электрону полупроводника, W_g - ширина запрещенной зоны, $-e\psi_B$ - разность между уровнем Ферми W_F и серединой запрещенной зоны W_i, φ_b – потенциальный барьер между металлом и диэлектриком, χ_i – сродство к электрону для изолятора.

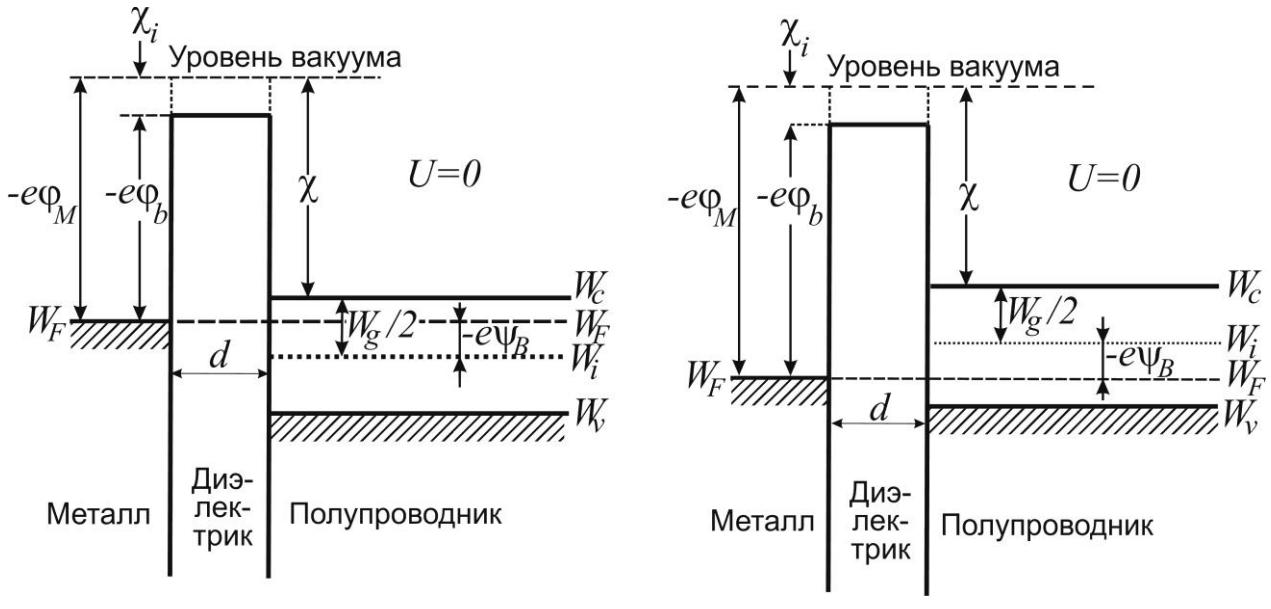


Рис. 1.15. Зонные диаграммы идеальных МДП-структур при U = 0: *а)* — полупроводник *n*-типа; *б)* — полупроводник *p*-типа

Понятие «*идеальная МДП-структура*» определим следующим образом:

1. Работы выхода электронов из металла и полупроводника одинаковы, т.е. разность работ выхода из металла и полупроводника равна нулю. Это означает, что в отсутствие внешнего напряжения $U = 0$ энергетические зоны полупроводника не изогнуты (состояние плоских зон, см. рис. 1.15).
2. При любых напряжениях смещения в структуре могут существовать только заряд в ее полупроводниковой части и равный ему заряд противоположного знака на металлическом электроде, отделенном от полупроводника слоем диэлектрика.
3. При постоянном напряжении смещения отсутствует перенос носителей заряда через диэлектрик, т. е. сопротивление диэлектрика предполагается бесконечно большим.

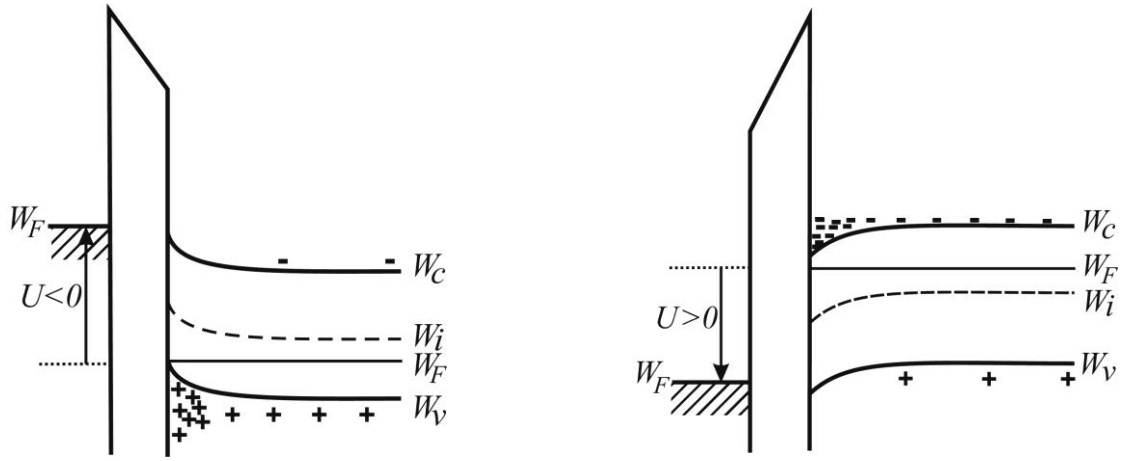
Если к идеальной МДП-структуре приложить напряжение того или иного знака, то на полупроводниковой поверхности появится электрический заряд. При этом возможен один из трёх вариантов состояния равновесия. Рассмотрим их на примере МДП-структуры с полупроводником *p*-типа (рис. 1.16, левый столбец). Если к металлическому электроду структуры приложено отрицательное напряжение ($U < 0$), край валентной зоны у границы с диэлектриком изгибается вверх и приближается к уровню Ферми (рис. 1.16 *а*).

Поскольку в идеальной МДП-структуре сквозной ток равен нулю, уровень Ферми в полупроводнике остается постоянным. Так как концентрация дырок экспоненциально зависит от разности энергий ($W_F - W_V$), такой изгиб зон приводит к увеличению числа основных носителей (дырок) у поверхности полупроводника. Этот режим называется режимом обогащения (аккумуляции). Если к МДП-структуре приложено не слишком большое положительное напряжение ($U > 0$), зоны изгибаются в обратном направлении и приповерхностная область полупроводника обедняется основными носителями (рис. 1.16 *б*). Этот режим называют режимом обеднения или истощения поверхности. При больших положительных напряжениях зоны изгибаются вниз настолько сильно, что вблизи поверхности уровень Ферми пересекает собственный уровень W_i. В этом случае (рис. 1.16 *в*) концентрация неосновных носителей (электронов) у поверхности превосходит концентрацию основных носителей (дырок). Эта ситуация называется режимом инверсии.

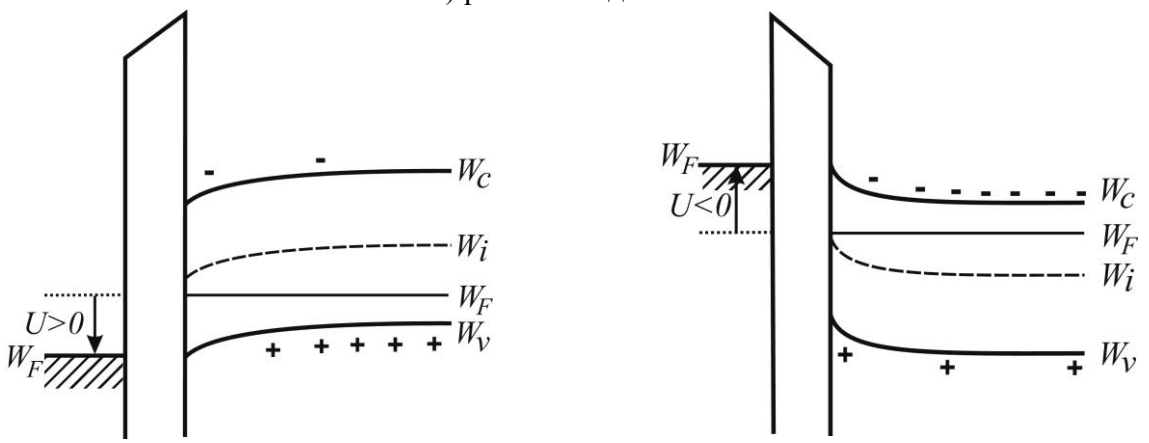
Аналогичное рассмотрение можно провести и для МДП-структуры с полупроводником *n*-типа, при этом все указанные режимы будут осуществляться при напряжениях противоположной полярности (рис. 1.16, правый столбец).

МДП структура с полупроводником *p*-типа МДП структура с полупроводником *n*-типа

а) режим обогащения



б) режим обеднения



в) режим инверсии

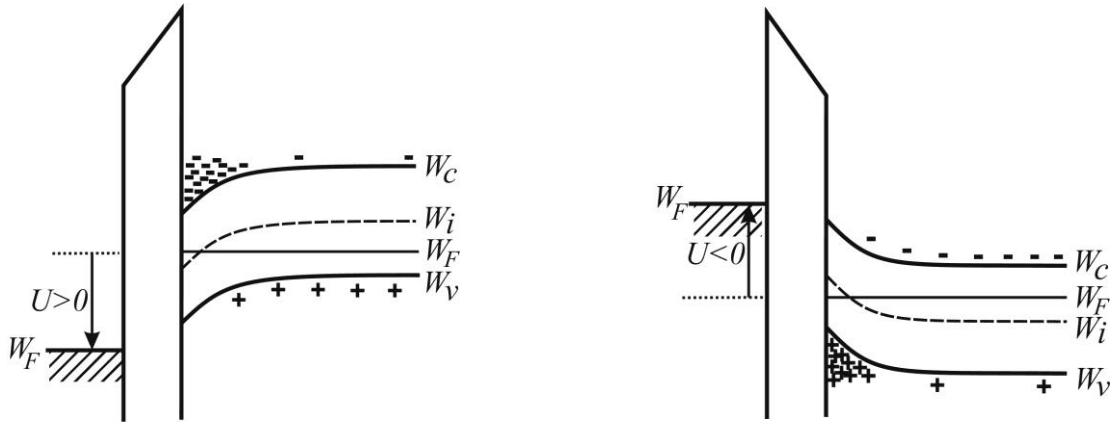


Рис. 1.16. Зонные диаграммы идеальных МДП – структур с полупроводниками p- (левый столбец) и n-типов (правый столбец) при подаче внешнего напряжения: *а*) режим аккумуляции; *б*) – режим обеднения; *в*) режим инверсии. Важно, что тангенс угла наклона дна зоны проводимости в слое диэлектрика и в прилегающем полупроводнике имеют одинаковый знак. Это объясняется тем, что нормальное к поверхностям раздела слоев электрическое поле создается зарядами, расположенными в полупроводнике и на поверхности металла (в идеальном диэлектрике зарядов нет). Поэтому вектор напряженности электрического поля имеет одно и то же направление в диэлектрике и прилегающем к нему слое полупроводника

1.5.2. Емкость МДП – структуры

Вольт-фарадная характеристика МДП-структуры приведена на рисунке 1.21:

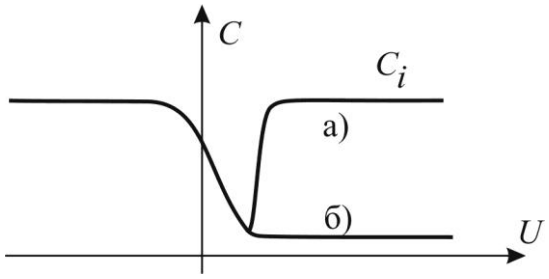


Рис. 1.17. Вольт-фарадная характеристика МДП–структуры на низкой частоте (а) и высокой частоте (б)

Поясним этот график на примере структуры металл – диэлектрик - полупроводник *p*-типа. При подаче значительного отрицательного потенциала на металл реализуется режим аккумуляции основных носителей заряда (дырок). Емкость в этом случае почти неизменна: $C_i \approx \varepsilon\varepsilon_0 S/d$, где d – ширина диэлектрика. При убывании отрицательного смещения вблизи поверхности образуется обедненная область, действующая как диэлектрик. Полная ёмкость при этом убывает. Далее ёмкость проходит через минимум и снова возрастает при образовании вблизи поверхности инверсионного слоя электронов при подаче положительного смещения. Отметим, что возрастание ёмкости в области положительных смещений зависит от способности электронов следовать за изменениями приложенного переменного сигнала. Это возможно лишь при низких частотах, когда скорость генерации - рекомбинации неосновных носителей достаточна для изменения заряда электронов в инверсионном слое в соответствии с изменением сигнала, на котором производятся измерения. В области же более высоких частот измерительного сигнала в правой части характеристики не наблюдается увеличения ёмкости (рис. 1.17 б).

1.6. Контакт металл - полупроводник

1.6.1. Зонная диаграмма

Чтобы показать, как формируется потенциальный барьер вблизи границы металла с полупроводником, имеющим другую работу выхода, предположим вначале, что материалы электрически нейтральны и изолированы друг от друга. На рис. 1.18 а представлена энергетическая зонная диаграмма для полупроводника *n*-типа, работа выхода из которого ($-e\varphi_n$) меньше, чем работа выхода из металла ($-e\varphi_m$). Именно этот случай приводит к возникновению потенциального

барьера в полупроводнике, называемого барьером Шоттки. Кроме того, будем считать, что заряженные поверхностные состояния[6] отсутствуют.

Если металл и полупроводник теперь электрически соединить друг с другом, то часть электронов перейдет из полупроводника в металл. При этом уровни Ферми W_F в обоих материалах сравняются - рис. 1.18 б. Вследствие такого перехода электронов в зазоре между металлом и полупроводником возникнет электрическое поле, и энергетические зоны у поверхности полупроводника изогнутся, как показано на рисунке 1.22 б. При дальнейшем уменьшении толщины вакуумного зазора энергетический барьер, образованный потенциалом этого зазора, становится туннельно-прозрачным для электронов, и им можно пренебречь (рис. 1.18 в).

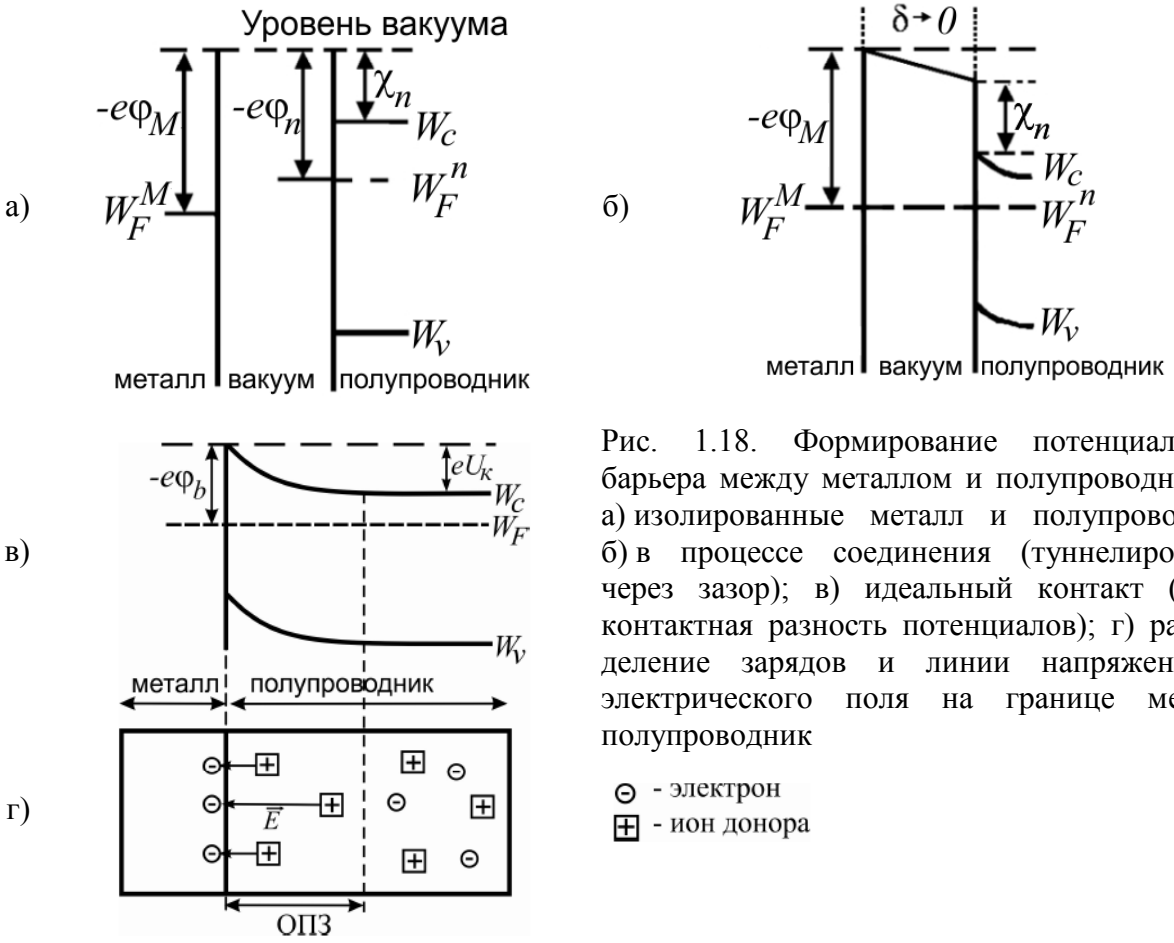


Рис. 1.18. Формирование потенциального барьера между металлом и полупроводником: а) изолированные металл и полупроводник; б) в процессе соединения (туннелирование через зазор); в) идеальный контакт ($U_к$ – контактная разность потенциалов); г) распределение зарядов и линии напряженности электрического поля на границе металл-полупроводник

⊖ - электрон
⊞ - ион донора

В действительности, идеализированная картина формирования барьера Шоттки (рис. 1.18 а-в) может нарушаться по целому ряду причин. Во-первых, из-за сохранения промежуточного слоя толщиной порядка нескольких межатомных расстояний (рис. 1.18 б). Во-вторых, из-за наличия поверхностных состояний в полупроводнике. Если плотность поверхностных состояний доста-

[6] Очевидно, что на границе раздела полупроводника с какой-либо средой (в том числе, вакуумом) происходит нарушение периодичности потенциала кристаллической решетки. В связи с этим в данной области появляются электронные уровни в запрещенной зоне, которые принято называть п*оверхностными состояниями.* Они могут захватывать электроны или дырки, приобретая, таким образом, отрицательный или положительный заряд.

точно велика, то заряд, связанный с ними, может эффективно экранировать полупроводник от электрического поля в промежуточном слое. Поэтому величина заряда в обедненной области и высота барьера могут не зависеть от работы выхода металла. Третья причина, по которой в рис. 1.18 в. вносятся коррективы, связана с влиянием сил изображения (т.е. с тем, что электроны, приближающиеся к барьеру, заряжают поверхность раздела равным и противоположным по знаку зарядом, который притягивает электрон и, таким образом, изменяет профиль потенциального барьера, при этом снижая его высоту). Такое изменение формы барьера известно как эффект Шоттки.

Форма потенциального барьера зависит от распределения заряда в обедненной области. Потенциальная энергия W в области объёмного заряда удовлетворяет уравнению Пуассона:

$$\Delta W = \frac{e\rho}{\varepsilon\varepsilon_0}, \tag{1.46}$$

где ε- диэлектрическая постоянная, ρ- плотность электрического заряда. В одномерном случае, когда все величины зависят только от координаты x, отсчитываемой вглубь полупроводника, для идеального барьера получаем:

$$\frac{d^2W}{dx^2} = \frac{e^2}{\varepsilon\varepsilon_0}(N_d - n(x)), \tag{1.47}$$

где N_d и $n(x)=N_d\,exp(-W(x)/kT)$ (распределение Больцмана) – соответственно, концентрации ионизированных доноров и свободных электронов. В объеме полупроводника при этом полный электрический заряд равен нулю в силу условия электронейтральности. Введя безразмерную потенциальную энергию $w(x) = W(x)/kT$ и дебаевскую длину $l_D = \sqrt{\frac{\varepsilon\varepsilon_0 kT}{e^2 N_D}}$, преобразуем (1.47) к виду:

$$\frac{d^2w}{dx^2} = \frac{1}{l_D^2}\left(1 - e^{-w(x)}\right). \tag{1.48}$$

с граничными условиями $w(0) = \frac{-e(\varphi_m - \varphi_n)}{kT} = \frac{eU_k}{kT}$, $w(\infty)=0$, $U_к>0$ – контактная разность потенциалов. Для типичных параметров полупроводников при комнатной температуре дебаевская длина экранирования во много раз больше постоянной решетки.

При $w(0)>>1$, если пренебречь изгибом зон в переходной области, где концентрации электронов и доноров сравнимы (так называемое приближение полного обеднения), форма барьера может быть определена из уравнения (1.48) без экспоненциального слагаемого (учет его необходим лишь в области $w(x)\le 1$):

$$w(x) \approx \frac{1}{2}\left(\frac{x}{l_D} - \sqrt{\frac{2eU_k}{kT}}\right)^2. \tag{1.49}$$

Получающийся барьер параболической формы известен как барьер Шоттки.

1.6.2. Теория процессов переноса зарядов

Перенос заряда через контакт металл-полупроводник осуществляется главным образом основными носителями в отличие от *p-n* переходов, где электрический ток обусловлен неосновными носителями. Это позволяет применять диоды Шоттки вплоть до терагерцового диапазона частот, так как скорость перераспределения основных носителей заряда характеризуется временем релаксации Максвелла $\tau_M = \frac{\varepsilon\varepsilon_0}{\sigma}$ и для типичных значений диэлектрической проницаемости ε и удельной проводимости σ лежит в пределах $10^{-11}\ldots10^{-14}$ с. На рис. 1.19 показаны четыре основных транспортных процесса при прямом смещении: эмиссия электронов из полупроводника над барьером в металл, квантово-механическое туннелирование через барьер, рекомбинация в области пространственного заряда, рекомбинация в нейтральной области.

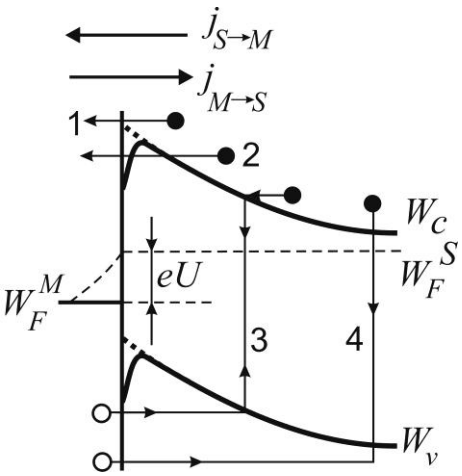


Рис. 1.19. Процессы токопереноса в контакте металл-полупроводник при прямом смещении. W_F^M, W_F^S – соответственно, квазиуровни Ферми в металле и полупроводнике; U – внешнее напряжение; $j_{S\to M}$, $j_{M\to S}$ – соответственно, плотности токов из полупроводника в металл и обратно. Стрелками обозначены следующие процессы:

1. эмиссия электронов из полупроводника над барьером в металл;
2. квантово-механическое туннелирование через барьер;
3. рекомбинация в области пространственного заряда;
4. рекомбинация в нейтральной области (инжекция дырок)

Кроме того, определенный вклад в полный ток может быть связан с токами утечки через ловушки на границе металл-полупроводник через периферийные области контакта, где в силу неоднородности возможны сильные краевые электрические поля. Хотя полностью исключить влияние указанных факторов, по-видимому, невозможно, тем не менее, современный уровень технологии позволяет изготовить диоды с барьером Шоттки, у которых механизм надбарьерного токопереноса является доминирующим, а их поведение вполне соответствует теоретическим представлениям.

Надбарьерное прохождение электронов из полупроводника в металл можно разделить на два этапа. Первый - дрейфово-диффузионный выход электронов из объема полупроводника к его поверхности. Второй - эмиссия электронов в металл. Оба процесса прохождения действуют последовательно, но, как правило, ток через контакт ограничивается одним из них. В соответствии с диффузионной теорией определяющим является первый процесс,

согласно диодной (или термоэмиссионной) – второй [2, 4]. Ниже мы ограничимся рассмотрением основных положений теории термоэлектронной эмиссии, адекватно описывающей процессы токопереноса в полупроводниках с высокой подвижностью электронов, таких, например, как кремний или арсенид галлия, и диффузионной теории.

Транспортные свойства контакта металл/полупроводник определяются соотношением толщины области пространственного заряда и длины свободного пробега электронов (которая в свою очередь определяется процессами рассеяния на фононах, примесных центрах и других нарушениях кристаллической структуры). Если длина свободного пробега много больше толщины ОПЗ, можно воспользоваться так называемой диодной теорией выпрямления (Бете, 1942 г.) *в приближении термоэлектронной эмиссии*.

Предположим, что

- ✓ величина изгиба зон (т.е. высота барьера для электронов, движущихся из полупроводника в металл) $|eU_к| >> kT$;
- ✓ область, определяющая термоэлектронную эмиссию, находится в термодинамическом равновесии (иными словами, положение квазиуровня Ферми в полупроводнике не изменяется вплоть до границы с металлом);
- ✓ протекание электрического тока не нарушает этого равновесия.

Ток через диод Шоттки представляет собой разность между током из металла в полупроводник и противоположным ему током, причем металл и полупроводник характеризуются каждый своим квазиуровнем Ферми. Величина тока в этом случае зависит только от высоты барьера и не зависит от его формы.

При приложении смещения U для электронов, покидающих полупроводник, высота барьера изменяется на eU, тогда как для электронов, движущихся в противоположном направлении, величина барьера меняется мало (только за счет эффекта Шоттки). Иными словами, концентрация электронов на границе раздела со стороны полупроводника растет как $e^{(eU/kT)}$ при увеличении напряжения U, а со стороны металла практически постоянна. Электроны на полупроводниковой стороне контакта находятся в термодинамическом равновесии с электронами в объеме полупроводника. Их концентрация выражается соотношением:

$$n(0) = n_0 \cdot e^{-\frac{e(U_к - U)}{kT}} = N_C \cdot e^{-\frac{W_C - W_F}{kT}} \cdot e^{-\frac{e(U_к - U)}{kT}} = N_C \cdot e^{-\frac{e(\varphi_b - U)}{kT}}, \quad (1.50)$$

где n_0 – равновесная концентрация электронов в глубине полупроводника, φ_b – потенциал барьера для электронов, идущих из металла в полупроводник, $N_C = 2\left\{\frac{2\pi m^* kT}{h^2}\right\}^{\frac{3}{2}}$ – эффективная плотность состояний в зоне проводимости.

Для полупроводников со сферическими изоэнергетическими поверхностями эти электроны имеют изотропное распределение по скоростям. Число электронов, падающих на единицу площади границы раздела в единицу времени, в соответствии с элементарной кинетической теорией равно $n(0)\cdot\langle\upsilon\rangle/4$, где $\langle\upsilon\rangle$ - средняя по абсолютной величине тепловая скорость электронов в полупроводнике. Тогда:

$$j_{\mathrm{M}\to\mathrm{S}} = \frac{eN_c\langle \mathrm{x}\rangle}{4} e^{-\frac{e\left(\varphi_b - U\right)}{kT}}, \qquad (1.51)$$

где $j_{\mathrm{M}\to\mathrm{S}}$ - плотность электрического тока из металла в полупроводник. Здесь индекс $\mathrm{M}\to\mathrm{S}$ соответствует направлению движения положительных зарядов (т.е. направлению вектора плотности тока). Поток электронов при этом направлен в противоположную сторону, т.е. из полупроводника в металл.

Плотность тока в обратном направлении не зависит от смещения (если пренебречь любой возможной полевой зависимостью φ_b). При нулевом смещении $j_{\mathrm{M}\to\mathrm{S}} = j_{\mathrm{S}\to\mathrm{M}}$, следовательно,

$$j_{\mathrm{S}\to M} = \frac{eN_c\langle x\rangle}{4} e^{-\frac{e\varphi_b}{kT}}. \qquad (1.52)$$

Учитывая, что полный ток $j = j_{\mathrm{M}\to S} - j_{\mathrm{S}\to\mathrm{M}}$, получаем

$$j = \frac{eN_c\langle \upsilon\rangle}{4} e^{-\frac{e\varphi_b}{kT}} \left[e^{\frac{eU}{kT}} - 1 \right]. \qquad (1.53)$$

При максвелловском распределении по скоростям:

$$\langle x\rangle = \sqrt{\frac{8kT}{\pi m^*}}. \qquad (1.54)$$

Подставляя (1.54) и выражение для N_c в (1.53), получаем окончательное выражение для вольт-амперной характеристики диода Шоттки в модели термоэлектронной эмиссии:

$$j = j_s \left[e^{\frac{eU}{kT}} - 1 \right], \qquad (1.55)$$

$$j_s = A^* T^2 e^{-\frac{e\varphi_b}{kT}}. \qquad (1.56)$$

Здесь

$$A^* = 4\pi m^* ek^2 / h^3 \quad - \tag{1.58}$$

эффективная постоянная Ричардсона. Для напряжений смещения, больших $3kT/e$, единицей в квадратных скобках в (1.55) можно пренебречь, и тогда плотность тока будет пропорциональна $e^{(eU/kT)}$.

Реальная вольт-амперная характеристика имеет вид:

$$j = j_S \left[e^{\frac{eU}{nkT}} - 1 \right], \tag{1.59}$$

где n>1 - почти постоянная величина. Обычно её называют фактором (или коэффициентом) неидеальности диода. (ранее применительно к гетероструктурам на с.26 этот коэффициент уже вводился).

Отклонение ВАХ от идеальной при больших токах связано с падением напряжения смещения на последовательно включенном сопротивлении R (например, сопротивлении нейтральной области полупроводника R_s), которое всегда присутствует в реальных диодах. По этой причине фактическое падение напряжения на барьере Шоттки будет меньше напряжения на внешних выводах диода. При этом плотность тока будет пропорциональна $(e^{\frac{e(U - JR)}{nkT}} - 1)$, где J - ток через диод.

Ёмкость диода Шоттки находится так же, как ёмкость резконеоднородного p-n перехода, и имеет вид:

$$C = S\sqrt{\frac{e\varepsilon\varepsilon_0 N_d}{2(U_k - U)}}. \tag{1.60}$$

Остановимся теперь на основах так называемой диффузионной теории выпрямления (Давыдов, Шоттки, Пекар, 1939 г.), применимой при длинах свободного пробега малых по сравнению с толщиной обеднённого слоя. В этом случае ток определяется выражением

$$j_{nx} = \sigma E_x + (-e) \cdot \left(- D \frac{dn(x)}{dx} \right), \tag{1.61}$$

где $\sigma = en\mu$ - проводимость, μ - подвижность носителей заряда, $j_{nx}=const$ – плотность электронного тока.

Кроме того, на поверхности $\varphi(0)=U_к-U$.

Из (1.61) с учетом соотношения Эйнштейна $\mu=eD/kT$ получим:

$$\frac{dn(x)}{dx} - \frac{e}{kT} \cdot \frac{d\varphi}{dx} \cdot n(x) - \frac{j_{nx}}{\mu kT} = 0. \tag{1.62}$$

Распределение электростатического потенциала $\varphi(x)$ связано, конечно, с распределением концентрации $n(x)$, однако, формально мы можем рассматривать (1.62) как линейное дифференциальное уравнение первого порядка для

неизвестной функции $n(x)$, считая $\frac{d\varphi}{dx}$ заданной функцией x. Иначе говоря, ограничимся линейной моделью. В этом случае общее решение уравнения (1.62) имеет вид

$$n(x) = n_0 e^{\frac{e\varphi(x)}{kT}} - \frac{j_{nx}}{\mu kT} \int_x^\infty e^{\frac{e}{kT}[\varphi(x)-\varphi(\xi)]} d\xi, \qquad (1.63)$$

где $n_0 = n(x \to \infty) = \text{const}$.

На поверхности полупроводника (т.е. при $x=0$) концентрация электронов равна:

$$n(0) = n_0 e^{-\frac{eU_k}{kT}} \cdot e^{\frac{eU}{kT}} - \frac{j_{nx}}{\mu kT} \int_0^\infty e^{\frac{e}{kT}[\varphi(0)-\varphi(\xi)]} d\xi. \qquad (1.64)$$

В рассматриваемом случае потоки электронов из металла в полупроводник и обратный (каждый в отдельности) много больше результирующего потока $(1/e)j_{nx}$. При этом на границе $x=0$ сохраняется практически равновесная концентрация электронов, т.е.

$$n(0) = n_0 e^{-\frac{eU_k}{kT}}. \qquad (1.65)$$

Разрешая (1.64) относительно j_{nx} и используя (1.65), получим

$$j_{nx} = j_S \left[e^{\frac{eU}{kT}} - 1 \right], \qquad (1.66)$$

где

$$j_S = \frac{n(0)\mu kT}{\int_0^\infty e^{\frac{e}{kT}[\varphi(0)-\varphi(\xi)]} d\xi}. \qquad (1.67)$$

При вычислении интеграла в (1.67) ограничимся линейным членом в разложении $\varphi(\xi)$ в ряд по степеням ξ

$$\varphi(\xi) \approx \varphi(0) + \left(\frac{d\varphi}{d\xi} \right)_{\xi=0} \xi. \qquad (1.68)$$

Тогда

$$\int_0^\infty e^{\frac{e}{kT}[\varphi(0)-\varphi(\xi)]} d\xi \approx \int_0^\infty e^{-\frac{e}{kT}\left(\frac{d\varphi}{d\xi}\right)_{\xi=0}} d\xi = \frac{kT}{e} \frac{1}{\left(\frac{d\varphi}{dx}\right)_{x=0}}. \quad (1.69)$$

Подставляя (1.69) в (1.67), получим

$$j_S = en(0)\mu\left(\frac{d\varphi}{dx}\right)_{x=0}. \quad (1.70)$$

В (1.70) можно не учитывать слабую зависимость $\left(\frac{d\varphi}{dx}\right)_{x=0}$ от приложенного напряжения.

Таким образом, в случае диффузионной теории выпрямления ВАХ качественно имеет вид, изображенный на рис. 1.6, однако, при той же высоте потенциального барьера ток насыщения по величине значительно меньше, чем в модели термоэлектронной эмиссии.

1.6.3. Омический контакт

Омическим контактом называют контакт металл — полупроводник, сопротивление которого пренебрежимо мало по сравнению с объемным сопротивлением полупроводника. Один или несколько омических контактов присутствуют во всех без исключения полупроводниковых приборах. Важно, что хороший омический контакт не должен приводить к существенному изменению характеристик прибора, а падение напряжения на таком контакте при пропускании тока должно быть мало по сравнению с падением напряжения на активной области прибора.

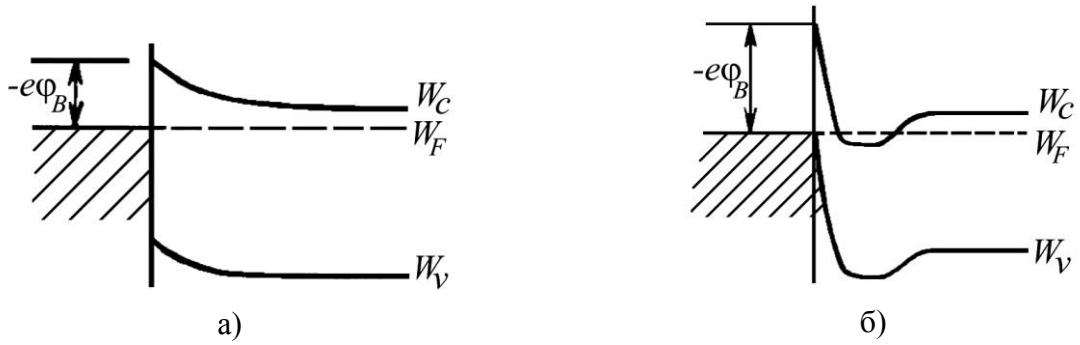
Рассмотрим сначала удельное сопротивление контакта, определяемое как обратная величина от производной плотности тока по напряжению. Наиболее важной характеристикой контакта является сопротивление при нулевом смещении (величина, обратно пропорциональная тангенсу угла наклона ВАХ):

$$R_c = \left(\frac{dJ}{dU}\right)^{-1}_{U=0}. \quad (1.71)$$

Малые значения R_c достигаются при малой ширине барьера (высокой степени легирования) или его малой высоте. Именно из этих соображений исходят при изготовлении омических контактов (рис. 1.20 а, б).

Для широкозонных полупроводников трудно изготовить контакт с малой высотой барьера. Кроме того, используемые металлы не всегда имеют достаточно малую работу выхода. В таких случаях для изготовления омических контактов создают дополнительный высоколегированный подслой на поверхности полупроводника (рис. 1.20 б). Использование такого подслоя позволяет уменьшить ширину барьера до толщины, когда протекание тока в основном

происходит за счет туннелирования «сквозь» барьер. Таким образом, если в некотором твердотельном приборе контакт между металлом и полупроводником должен быть невыпрямляющим (т.е. он не должен влиять на характеристики устройства), то под слоем металла обязательно содержится слой высоколегированного полупроводника, формирующий омический контакт. Обратите внимание, что на зонной диаграмме омические контакты, как правило, не отображаются – там приводится лишь активная область прибора.



а) б)

Рис. 1.20. Омические контакты с малой высотой барьера (*а*) и высокой степенью легирования (*б*)

1.7. Эквивалентные схемы диодов

Для расчета параметров полупроводниковых приборов часто используют метод эквивалентной схемы, который заключается в замене единого устройства набором стандартных элементов. В случае пассивных приборов для такой замены используются конденсаторы, сопротивления и индуктивности. В усилительных устройствах, например, транзисторах, в эквивалентную схему необходимо включать генераторы тока и/или напряжения. Цель подобной замены – сведение задачи о расчете характеристик полупроводниковых элементов к известной задаче анализа токов в радиотехнической цепи с помощью уравнений Кирхгофа. Вычисление сопротивлений полупроводниковых слоев обычно производят на основе известной проводимости этих слоев, т.е. исходя из концентрации носителей заряда и их подвижности.

Выше мы уже отмечали, что в ОПЗ полупроводника есть электрическое поле и содержится мало свободных носителей заряда, поэтому такая область отчасти похожа на слой диэлектрика в конденсаторе. Отличие состоит в том, что через эту область течет ток, что учитывается в эквивалентной схеме с помощью сопротивления, включенного параллельно конденсатору.

Общий вид эквивалентных схем диодов на основе n^+-*n* перехода, *p*-*n* перехода, барьера Шоттки и МДП структуры изображен на рис. 1.21.

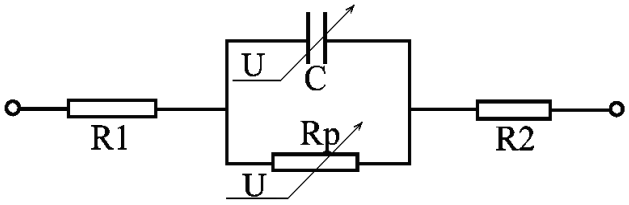


Рис. 1.21. Эквивалентная схема диодов на основе *p-n* перехода, барьера Шоттки и МДП- структуры

Здесь C – управляемая напряжением емкость перехода; U – величина напряжения, приложенного непосредственно к переходу (т.е. без учета падения напряжения на сопротивлениях R_1 и R_2); R_1, R_2 – последовательные сопротивления областей полупроводникового материала по обе стороны от перехода, включая сопротивления металлических контактов; R_p – управляемое напряжением сопротивление перехода или барьера Шоттки (определяется из ВАХ диодов).

Обычно одно из последовательных сопротивлений R_1, R_2 значительно меньше другого, например, в p^+-n переходе или контакте металл-полупроводник. В такой ситуации меньшим сопротивлением пренебрегают. Большее сопротивление, по сути, является сопротивлением базы диода.

При обратном смещении сопротивление перехода R_p увеличивается, так что сопротивлениями R_1 и R_2 часто можно пренебречь. При небольшом прямом смещении, когда внешнее напряжение сравнимо с контактной разностью потенциалов, следует учитывать все сопротивления, а при больших прямых смещениях обычно существенным является только сопротивление базы.

Из-за того, что при больших прямых смещениях емкость диода шунтирована малым сопротивлением перехода R_p, ее можно не учитывать. При прямых напряжениях сравнимых с контактной разностью потенциалов обычно учитывают диффузионную емкость, а при обратных напряжениях – барьерную.

В случае МДП-структуры на эквивалентной схеме остаются лишь два элемента C и R_2, включенные последовательно. Последнее означает, что постоянный ток через диэлектрик не протекает.

ЧАСТЬ 2

СУПЕРКОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ПОЛУПРОВОДНИКОВЫХ ДИОДОВ С УЧЕТОМ РАДИАЦИОННОГО ВОЗДЕЙСТВИЯ

Радиационное воздействие принято разделять на корпускулярное и электромагнитное ионизирующие излучения. Корпускулярное образуется элементарными частицами: протонами, нейтронами, электронами, альфа-частицами; а электромагнитное - это рентгеновское и гамма-излучение, а также излучение оптического диапазона. Основными источниками радиации являются: ядерный взрыв (нейтроны и гамма-кванты), ядерные энергетические установки (нейтроны и гамма-кванты), космическое пространство (протоны и электроны), рентгеновские и гамма-установки, а также ускорители электронов и протонов. Для изготовления интегральных схем используют ускоритель ионов [9, 10].

Воздействие радиации приводит к образованию различного рода структурных дефектов и ионизации полупроводниковых материалов. Начало исследований в области радиационной физики твердотельных приборов приходится на 60-е года XX века и обусловлено активным внедрением диодов, транзисторов и интегральных схем на их основе в военные и космические системы радиолокации и связи. Обзоры первых исследований в этой области науки и техники представлены в классических монографиях [9, 10].

В это время (1960-е и 70-е годы) появились первые аналитические модели реакции полупроводниковых приборов на радиационное воздействие. Однако такого рода модели плохо описывали нелинейные эффекты, связанные с высоким уровнем инжекции и фотовозбуждения неравновесных носителей заряда в полупроводниковых структурах при радиационном воздействии. Кроме того, практически не учитывалось влияние тепловых эффектов, возникающих как за счет кратковременного увеличения тока приборов на порядок величины и более непосредственно после ионизирующего воздействия, так и за счет долговременного снижения амплитуды токов приборов, вызванного радиационным дефектообразованием в структуре. Сложно учесть и влияние электромагнитных полей высокой напряженности, которые, взаимодействуя с электрическими проводами и кабелями внутри блоков оборудования, могут возбуждать импульсы напряжения с амплитудой в десятки и сотни вольт [11].

Несмотря на указанные недостатки, приближенные аналитические модели полупроводниковых диодов и транзисторов применяются и в настоящее время, особенно для расчетов интегральных схем с высокой степенью интеграции, т.е. большим количеством указанных элементов на кристалле – 10^9 транзисторов и более. Важным преимуществом аналитических моделей является простота и, следовательно, высокая «скорость» расчета параметров приборов благодаря использованию характеристик в виде алгебраических зависимостей. Последние устанавливают связь между параметрами полупроводниковых материалов: количеством, предельно возможной скоростью электронов, геометрическими размерами приборов и их электрическими характеристиками.

Это крайне важно для технологов и конструкторов, поскольку позволяет определять характерные размеры, взаимное расположение и внутреннюю структуру таких фрагментов интегральных схем как инвертор, триггер, усилитель, генератор и т.д. Иными словами, это дает возможность перейти от геометрических и электрических параметров слоев полупроводника и контактов к электрическим параметрам функциональных блоков интегральной схемы.

Основной проблемой при таком подходе являются погрешности при определении коэффициентов в алгебраических уравнениях, связывающих электрические и конструктивные характеристики диодов и транзисторов, а также низкая точность самих моделей. Для коррекции и уточнения моделей приходится использовать экспериментальную проверку результатов расчетов, так что процесс разработки интегральной схемы растягивается во времени, а затраты на его проведение резко возрастают. Каждая новая экспериментальная итерация, т.е. изготовление пробной партии интегральных схем, увеличивает сроки и стоимость конечного продукта в несколько раз. Таким образом, крайне важно иметь более точные модели диодов и транзисторов, позволяющие проводить оптимизацию их конструкции без изготовления. Однако подобное требует несравненно больших компьютерных ресурсов. Это крайне сложный и до конца не отлаженный процесс, если речь идет о перспективных транзисторах на новых гетеронаноструктурных материалах. Ситуация еще более усложняется, когда возникает необходимость разработки радиационно-стойких интегральных схем: дополнительные требования порождают дополнительные «степени свободы» при проектировании и число параметров, которые требуется оптимизировать, значительно возрастает.

Существующий сегодня стандартный промышленный подход основан на так называемых библиотеках элементов, в том числе и радиационно-стойких, т.е. их стандартных конструкций, описанных в документации, которую изготовитель интегральных схем предоставляет дизайн-центрам. Эти центры могут быть территориально расположены по всему миру и занимаются только разработкой электрической конструкции интегральных схем на уровне готовых блоков, а изготовлением схемы занимается фабрика, которая не отвечает за саму конструкцию, а обязуется только изготовить и правильно соединить блоки в целую схему, согласно предоставленному ей проекту дизайн-центра. Указанная технология известна во всем мире, как «foundry», и аналогична строительству дома из бетонных плит: ограниченность разновидностей плит сильно ограничивает возможности строителей. Указанное соображение особенно важно, если речь идет о перспективных интегральных схемах. Например, при разработке аналоговых усилителей и генераторов субтерагерцового диапазона частот (100….1000 ГГц), которые необходимы для перспективных видов военной, космической и иной специализированной электронной техники, будут возникать существенные трудности. При этом конструкторская информация, предоставляемая фабрике-изготовителю, обладает значительной интеллектуальной стоимостью, так что делиться ею с конкурентами нежелательно. Кроме того, подобная информация позволяет определить параметры разрабатываемого устройства, что так же крайне нежелательно при производстве военной техни-

ки, тем более, радиационно-стойких интегральных схем, относящихся к специфической области применения.

Все сказанное выше приводит нас к необходимости разработки собственных средств проектирования и изготовления интегральных схем, особенно их наиболее важных полупроводниковых элементов – диодов и транзисторов. Следует прибегать к помощи уточненных (и усложненных) математических моделей, основанных на численном решении системы дифференциальных уравнений, описывающих движение электронов как поток заряженной жидкости, т.е. использующих гидродинамическое приближение для решения задачи. Такой подход позволяет проектировать диоды и транзисторы значительно точнее, а затем, для расчета интегральной схемы в целом, можно по-прежнему пользоваться функциональными моделями[7]. Преимуществом указанного подхода является максимальное использование технологических возможностей фабрики-изготовителя, так как, кроме стандартных блоков, интегральная схема может быть построена и на специальных, вновь разработанных, блоках, что улучшит значения ее электрических параметров (быстродействия, частоты функционирования, усиления, выходной мощности и т.п.). Особенностью такого подхода является необходимость значительно более тесного контакта между дизайн-центром и специалистами фабрики-изготовителя, позволяющего на уровне отдельных технологических операций определять возможность и целесообразность новых конструктивных решений, модификаций технологического процесса и особенностей межоперационного контроля параметров изготавливаемой микросхемы. При таком уровне взаимодействия разработчиков и технологов возникает масса возможностей улучшения параметров интегральных схем, что требует серьезной компьютерной оптимизации для выявления наиболее целесообразных вариантов прежде, чем будет организована дорогостоящая и длительная во времени экспериментальная проверка. Из-за огромного числа ключевых конструктивных и технологических параметров – 100…1000 и более – полноценная процедура оптимизации может быть проведена *только с использованием суперкомпьютера.*

Интерес отечественных специалистов в области разработки систем автоматизированного проектирования к высокопроизводительным вычислениям отражается как в лавинообразном увеличении числа публикаций, так и в издании в последние годы большого числа монографий по данному вопросу на русском языке [12-25]. Особенностью проектирования радиационно-стойких интегральных схем является необходимость включения в процедуру моделирования операций, связанных с анализом влияния процессов ионизации и дефектообразования, которые крайне чувствительны к наличию границ раздела материалов с разными химическими составами, плотностями, а значит, и коэффициентами поглощения радиационного излучения. Неравновесные процессы, возникающие на подобных границах при облучении интегральных схем, приводят к усилению радиационного воздействия в одних слоях и его

[7] Функциональные модели – модели высокого уровня, описывающие основную функцию того или иного блока аппаратуры; например, у усилителя в качестве параметров будут фигурировать только коэффициенты усиления и шума. Детали конструкции и технологии изготовления в таких моделях, как правило, не фигурируют.

ослаблению в других. Поэтому процедура оптимизации параметров современных радиационно-стойких приборов, изготавливаемых на гетеронаноструктурах, должна проводиться с помощью двух- и трехмерных гидродинамических моделей, учитывающих пространственную неоднородность влияния радиационного воздействия. Зачастую требование повышения радиационной стойкости противоречит требованию улучшения электрических параметров интегральных схем - приходится искать компромисс, причем, решение может лежать не только в области технологии и конструкции полупроводниковых приборов. Следует выбрать правильную компоновку радиотехнических систем и схемотехническое решения на уровне блоков оборудования, размещенных на печатных платах[8], а также определить оптимальное положение полупроводниковых приборов на плате.

2.1. Особенности создания современных полупроводниковых наногетероструктур диодов и транзисторов

С момента изобретения транзисторов в конце сороковых годов XX века встал вопрос об их эффективности. Например, изготовление усилителей электрических сигналов требует вполне определенного уровня усиления используемых транзисторов. Чем этот уровень выше, тем меньше последовательных каскадов транзисторов следует применять, тем проще, технологичней и дешевле интегральная схема. Для генераторов важно иметь заданную выходную мощность сигнала, для цифровых схем – определенную логику обработки и преобразования входного сигнала в выходной.

Любой из указанных типов интегральных схем, как правило, содержит от нескольких единиц (терагерцовые усилители сигнала), до нескольких миллиардов (процессоры) транзисторов. Чем меньше транзисторов, тем легче сделать схему, т.е. выдержать требуемые параметры при заданной рынком стоимости продукта. При предъявлении высоких требований к электрическим параметрам, что характерно для военной и иной специализированной техники, обычно стоимость продукта не столь важна. Работа на пределе технологических и конструктивных возможностей оборудования определяет, получится ли вообще интегральная схема или на используемом оборудовании невозможно ее изготовить. Иными словами, если процент выхода годных схем будет стремиться к нулю, то никакие деньги не позволят выпустить такие сложные схемы на данном оборудовании. Решением может являться покупка новой технологической «линейки», т.е. по сути нового завода с конвейером по производству полупроводниковых приборов.

Основная причина ограничения возможностей завода - это технологические допуска при изготовлении наиболее важных электродов приборов, напри-

[8] Печатная плата – конструктивное решение в виде диэлектрической пластины с нанесенными проводящими (металлическими) дорожками-проводами к которым припаяны диоды, транзисторы, интегральные схемы и другие элементы. Изготавливается методами фотолитографии, позволяющими «печатать» проводящие дорожки аналогично печати фотографий на фотобумаге. Примером может служить материнская плата компьютера.

мер, затворов полевых транзисторов, а также зазоров между электродами и иными конструктивными элементами как внутри, так и между соседними элементами. Указанная проблема может быть решена с использованием специальных видов литографического оборудования, которое формирует специальные технологическое маски на поверхности интегральной схемы, аналогично черным и белым полосам на фотографической пластине или пленке, используемой в фотоаппаратах. Дифракция света или иного излучения, например, электронного луча, ограничивает характерные размеры металлических проволочек, которые возможно нанести на поверхность интегральной схемы и использовать в качестве электродов диодов и транзисторов. На сегодня предельно возможным размером указанных проволочек является 10 нм, т.е. 10^{-8} м. Меньшие размеры изготавливаются только в единичных уникальных лабораториях, б*о*льшие – 100…500 нм и выше – доступны почти всем фабрикам в мире. Характерные размеры (топологические нормы) порядка 25…100 нм – это предмет гордости ведущих мировых производителей интегральных схем. Использование столь коротких электродов (10…100 нм) обусловлено стремлением снизить время пролета электронов под затвором транзистора: именно это время требуется, чтобы выключить ток в приборе (см. рис. 2.1). Указанное позволяет использовать такие приборы в интегральных схемах и аппаратуре СВЧ и КВЧ[9] диапазонов частот.

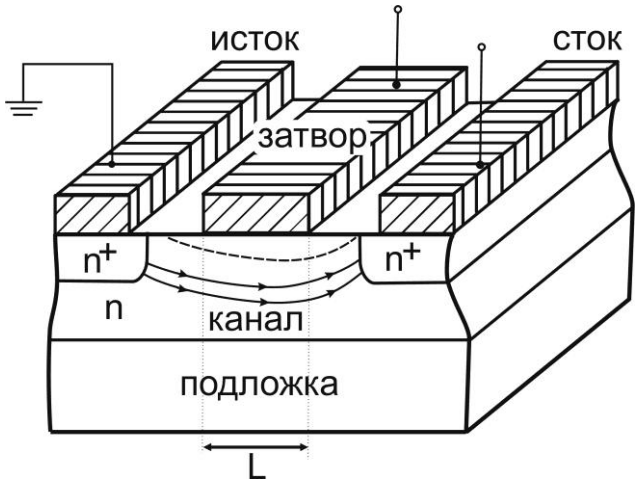


Рис. 2.1. Схематическое изображение конструкции классического полевого транзистора с управляющим барьером Шоттки. L – длина затвора: от 10 до 500 нм в зависимости от типа и назначения транзистора. При подаче электрического напряжения на электроды истока и стока возникшее электрическое поле заставит электроны двигаться в канале прибора в направлении от истока к стоку, как показано на рисунке стрелками. Меняя отрицательное напряжение на затворе, можно в той или иной мере отталкивать электроны от затвора, т.е. управлять величиной тока исток-сток

Схожим образом при закрывании водопроводного крана необходимо некоторое конечное время, чтобы из-под его поршня ушла вся вода. Пока есть вода – есть ток; поршень вытеснил воду – ток закончился. Открывая и закрывая краны - транзисторы, мы управляем электрическими сигналами, можем их преобразовать, усиливать и обрабатывать. Чем быстрее удается открывать и

[9] Стандартный термин сверхвысокочастотные (СВЧ) и крайневысокочастотные (КВЧ) монолитные интегральные схемы, т.е. схемы, изготовленные на одном полупроводниковом кристалле и использующие специальные напыленные на поверхность кристалла волноводы, фильтры и иные электрические элементы для передачи сигналов таких высоких частот без потерь от одного каскада интегральной схемы к другому.

закрывать такие «краны», тем выше частота сигнала и тем больше возможностей у разработчиков радиотехнического оборудования. Процессор будет быстрее производить расчеты, интегральная схема быстрее и точнее обрабатывать и преобразовывать сигнал. Радиолокатор сможет использовать более высокие частоты, т.е. меньшие длины волн. Это поможет точнее различать детали радиолокационной картины и использовать «радиозрение», т.е. визуализировать объекты с использованием радиочастот, как это раньше было сделано в инфракрасном, оптическом и рентгеновском диапазонах. Возможность реализации подобного «радиовидения» крайне важна для целого спектра прикладных задач. Среди них, например, борьба с терроризмом (в терагерцовом диапазоне частот можно обнаружить оружие и запрещенные вещества под одеждой и в багаже автомобиля), медицинская диагностика (возможно исследование особенностей функционирования органов человеческого организма, прозрачных для других видов излучения), диагностика материалов и конструкций (поиск скрытых трещин) и многие другие.

Преодолеть указанные выше ограничения литографического оборудования можно, используя другое геометрическое направление: если нельзя использовать очень короткие электроды, то следует использовать крайне тонкие полупроводниковые слои, которые, как оказалось, создавать несколько легче. Для этих целей применяются установки молекулярно-пучковой эпитаксии, когда в вакууме от сильно нагретого вещества за счет сублимации[10] формируется поток (пучок) молекул, которые оседают на полупроводниковой пластине, создавая тончайшие слои. Имея несколько таких источников вещества, которые могут закрываться и открываться с помощью быстро поворачивающихся специальных заслонок, возможно нанесение на пластину слоев различных атомов. Причем возможно образование соединений с произвольной пропорцией химических элементов.

Важно, что используя исходные пластины, вырезанные из кристаллических кремния, арсенида галлия, фосфида индия, сапфира или алмаза, удается вырастить на них кристаллические полупроводниковые слои максимально хорошего качества, т.е. почти не содержащие дефектов, препятствующих протеканию электронов или замедляющих их движение. А это повышает предельные частоты работы приборов. Изменяя температуру источников вещества в ростовом реакторе, удается формировать такие слабые его потоки, что за 1 секунду на пластине вырастает всего один монослой атомов кристалла. Варьируя химический состав наносимых гетерогенных слоев, можно, например, создать проводящий слой материала, зажатый между диэлектрическими слоями. Таким образом можно создавать сверхкороткие, то есть сверхбыстродействующие, приборы.

Отдельно следует упомянуть принципиальную возможность изготовления слоев материала в несколько раз тоньше, чем длина волны де Бройля электронов в полупроводнике. Это открывает поистине гигантские возможности для проектирования приборов. Формирование указанных слоев позволяет

[10] Сублимация – испарение вещества в вакууме, минуя жидкую фазу.

реализовать неклассические способы управления потоком носителей заряда. В частности, через активную область прибора за счет эффекта туннелирования можно пропускать лишь некоторое определенное количество электронов. Причем, направленно изменять число проходящих частиц можно путем вариации электрического потенциала[11]. Кроме того, можно в несколько раз увеличить скорость электронов за счет подавления их столкновений с кристаллической решеткой[12], создавать полупроводниковые лазеры с квантовыми ямами и т.п.

Основной проблемой при проектировании указанных приборов на гетеронаноструктурах является большое число их параметров. Для создания диода или транзистора необходимо изготовить десятки и даже сотни полупроводниковых, металлических и диэлектрических слоев с определенной толщиной и химическим составом, а также указать количество электронов, их подвижность и коэффициент диффузии, и иные параметры, определяющие способность носителей заряда ускоряться в электрическом поле и тормозиться при выходе из него. Также необходимо контролировать скорости взаимодействия электронов и дырок, приводящие к их взаимному уничтожению – рекомбинации – и т.п. В отличие от структуры классического прибора, изображенного на рисунке 2.1, обычное количество слоев типичного современного гетеронанотранзистора достигает 10 и более; для каждого слоя следует задать от 10 до 30 параметров, так что оптимизация интегральной схемы, содержащей всего 10 таких транзисторов – это крайне ресурсоемкая задача, которую в полном объеме можно решать только с использованием суперкомпьютера. Разумеется, можно разбить задачу на несколько этапов – с использованием сложной модели рассчитывать один типичный транзистор, а интегральную схему, состоящую, как правило, из большого числа подобных транзисторов, моделировать с помощью упрощенной модели. Но в таком случае оптимизация интегральной схемы будет проходить по упрощенному сценарию, что не позволит создавать оптимальные конструкции, особенно если речь идет о схемах с высоким уровнем радиационной стойкости.

2.2. Особенности контроля параметров полупроводниковых структур, диодов и транзисторов

Как указано выше, для моделирования физических процессов в полупроводниковых приборах необходимо с достаточной точностью знать параметры полупроводниковых слоев гетеронаноструктур диодов и транзисторов. Несмотря на то, что технологические процессы изготовления приборов на сегодняшний день хорошо развиты, необходимо контролировать результаты изготовления как полупроводниковых структур, так и приборов в целом. Это связано с тем, что сложнейшие технологические установки, изготавливающие современ-

[11] В гетеробиполярных транзисторах с туннельным эмиттером и туннельно-резонансных диодах
[12] В полевых транзисторах с двумерным электронным газом типа HEMT

ные гетеронаноструктуры требуют особой процедуры компьютерного управления, которая в идеале должна корректироваться в процессе изготовления на основе контроля параметров изготавливаемых слоев «in situ», т.е. внутри технологической установки в процессе изготовления. Часто это слишком сложно, дорого или долго. В результате структуры изготавливаются с определенными погрешностями, что обуславливает снижение параметров будущих полупроводниковых приборов и интегральных схем. Компенсация промахов на этапе изготовления структур возможна при своевременной коррекции конструкции диодов и транзисторов. Например, избыточное количество электронов в проводящем слое, например канале полевого транзистора, может быть компенсировано снижением его толщины, что позволит получить оптимальные токи транзистора, и прибор не будет перегреваться из-за избыточной проводимости.

Это важно еще и потому, что процесс изготовления, как правило, осуществляется в несколько этапов различными фирмами – гетеронаноструктуры изготавливают в одной, а интегральные схемы на их основе делают в другой. В связи с этим возникают операции как входного контроля параметров слоев полупроводниковых структур, так и межоперационного и выходного контроля параметров диодов, транзисторов и интегральной схемы. Таким образом, при изготовлении радиационно-стойкой аппаратуры в целом необходим тесный контакт между разработчиками и технологами различных фирм, изготавливающих структуры, интегральные схемы и радиотехническое оборудование. При этом от кристалла до готового прибора должно быть организовано сквозное проектирование.

Особенностью контроля параметров полупроводниковых слоев является проблема определения их качества с точки зрения отсутствия дефектов кристаллической решетки – вакансий, т.е. нехватки атомов в узлах решетки, или междоузлий – их избытка, когда есть лишние атомы, искажающие химические связи в кристалле. И то и другое, а также более сложные комплексы указанных дефектов и различного рода вредных примесей, случайно попавших в образец, образуют стабильные (и не очень) нарушения правильной кристаллической структуры материала и существенно препятствуют протеканию электронов в активной области прибора. На практике о качестве полупроводниковой структуры можно судить по результатам измерения подвижности носителей заряда, т.е. скорости частиц, достигаемой ими в единичном поле. Сравнение с табличными данными, приведенными в литературе для полупроводниковых кристаллов близких к идеальным, позволяет делать выводы о степени дефектности выращенных слоев.

Важным является и определение концентрации электронов в объеме структуры, так как произведение концентрации и скорости носителей задает сопротивление полупроводникового слоя. Согласно закону Ома, исходя из сопротивления и приложенного к металлическим контактам напряжения можно определить ток в диоде и/или транзисторе, а следовательно, и его вольтампер-

ную характеристику – зависимость тока прибора от поданного напряжения[13]. Эту характеристику в дальнейшем используют для расчетов параметров и моделирования процессов функционирования интегральных схем.

Одним из основных методов контроля параметров полупроводниковых слоев является измерение сопротивления, а вернее вольтамперной характеристики исследуемого образца полупроводниковой структуры, на которой созданы металлические электроды. К электродам припаиваются провода измерительного стенда, и производится измерение. Возникает своего рода парадокс: для моделирования вольт-амперной характеристики полупроводникового диода или транзистора, который предполагается изготовить на купленной полупроводниковой структуре, вначале необходимо изготовить «тестовые» электроды и произвести входной контроль – измерить вольтамперную характеристику тестового диода с целью проверки качества и точности изготовления полупроводниковой структуры.

Конечно, для облегчения измерений «тестовые» электроды изготавливаются по простейшему варианту технологического маршрута и таким образом, чтобы результаты измерений можно было расшифровать оптимальным способом, дающим минимальную погрешность. Однако на практике все не так просто, как хотелось бы. Как упоминалось выше, современные структуры содержат большое количество полупроводниковых слоев разного химического состава – это необходимо, чтобы будущий диод или транзистор хорошо функционировал. Электрический ток протекает сразу по нескольким слоям, поэтому расшифровка результатов измерений сопротивления образца – непростая задача. Непонятно, какой именно из слоев определяет измеренное сопротивление. Для уточнения результатов используют измерения в магнитных полях разной напряженности[14], образец охлаждают и/или нагревают, освещают светом с различной длиной волны и т.п. И, конечно, сложнейшая задача расшифровки результатов измерений не обходится без компьютерных расчетов.

После этапа изготовления интегральной схемы происходит анализ ее качества и отбраковка образцов, параметры которых выходят за рамки технического задания. Здесь также измеряются вольт-амперные, переходные и другие электрические характеристики интегральных схем и их составных частей. Это позволяет сопоставить результаты входного контроля параметров полупроводниковых структур и выходного контроля параметров приборов. Конечно, в процессе изготовления интегральной схемы осуществляют межоперационный контроль процесса изготовления, который позволяет корректировать и нивелировать погрешности изготовления на более поздних этапах. Так как процедура

[13] В общем случае количество протекающих в диоде и/или транзисторе электронов, а также их скорость зависят от поданного напряжения, так как электроны обычно вбрасываются в прибор через потенциальный барьер имитирующего контакта. Высота барьера управляется поданной разностью потенциалов. Из-за этого и некоторых других факторов зависимость тока от напряжения будет нелинейной. Данную зависимость используют при описании прибора и называют его вольтамперной характеристикой (ВАХ). Для ее детального прогнозирования проводится моделирование процесса движения электронов в активной области прибора, что является сложной самосогласованной задачей, так как электроны заряжены, и их количество и местоположение изменяют электрическое поле, в котором они двигаются.

[14] Имеется в виду эффект Холла.

изготовления интегральных схем содержит от нескольких десятков до нескольких сотен субпроцедур, межоперационный контроль и его расшифровка с помощью компьютерного моделирования, а также прогноз будущих параметров интегральной схемы и определение возможностей компенсации технологических ошибок являются важными задачами компьютерной оптимизации. Большое количество параметров структур требует большого объема расчетов, что обусловливает необходимость применения высокопроизводительных компьютерных средств.

Отдельно следует упомянуть о современных средствах электрических измерений. Как правило, для снижения погрешности измерений, возникающей вследствие электрических шумов в измерительных цепях, электромагнитных наводок, связанных с действием источников паразитных электрических сигналов (радио, телевидения, передатчиков сотовой связи и т.п.), используют усреднение нескольких (до сотни) измерений одной и той же зависимости. Измерения производятся в автоматическом режиме под управлением компьютера. Это достаточно долгий (до нескольких часов) процесс, так что при использовании нескольких измерительных стендов, определяющих те или иные параметры исходной полупроводниковой структуры и/или изготовленной интегральной схемы, требуется несколько дней, чтобы получить исчерпывающую информацию о разрабатываемом приборе. Использование компьютерного моделирования в реальном времени позволит снизить объем измерений, так как модель будет «подсказывать» измеряющему устройству в каком диапазоне входных параметров (например, напряжений) следует производить наибольшую детализацию измерений. Где для анализа проблемного слоя структуры следует увеличить точность и детализацию измерений, но при этом затратить больше времени, а где достаточно быстрых, но не очень точных измерений.

2.3. Особенности методов моделирования полупроводниковых диодов, транзисторов и интегральных схем

Исходными параметрами физико-технологических моделей являются технологические режимы. Как отмечалось выше, выходные параметры технологических операций измеряются и обрабатываются на высокопроизводительных рабочих станциях, что позволяет корректировать технологический процесс в режиме реального времени. Выходными параметрами физико-технологических моделей являются электрофизические параметры структур, такие как глубина залегания рабочей области активных элементов, времена жизни и подвижности электронов и дырок после проведения технологических операций диффузии примесей, ионного легирования и т.д. Зачастую технологический процесс, в том числе и радиационное воздействие на полупроводники (как полезное технологическое, так и вредное, например, в космическом про-

странстве) анализируют с помощью статистического метода Монте-Карло[15]. Это позволяет выделить важнейшие факторы, определяющие параметры процесса, среди различных и плохо поддающихся анализу явлений. Физико-технологические модели используются как для оптимизации технологических процессов, так и для получения исходных данных, необходимых при создании физико-топологических моделей активных элементов интегральных схем (рис. 2.2).

Исходными параметрами физико-топологических моделей являются геометрические размеры областей активных элементов и физические характеристики слоев: концентрация и подвижность электронов, времена, характеризующие процессы их рекомбинации и движения в приборе, набора кинетической энергии в электрическом поле и ее потери и др. Выходными параметрами являются электрические и эксплуатационные характеристики активных элементов: вольтамперные и вольтфарадные характеристики, напряжение пробоя, токи утечки, коэффициенты усиления, времена переключения и др.

Схемотехнические модели связывают функционирование отдельных активных элементов в единую электрическую схему. Исходными данными для них являются выходные параметры физико-топологических моделей. Наконец, модели системного уровня позволяют проектировать аппаратуру из функциональных блоков, каждый из которых описан набором параметров (например, таблицей истинности для цифровых схем и коэффициентом усиления для аналоговых).

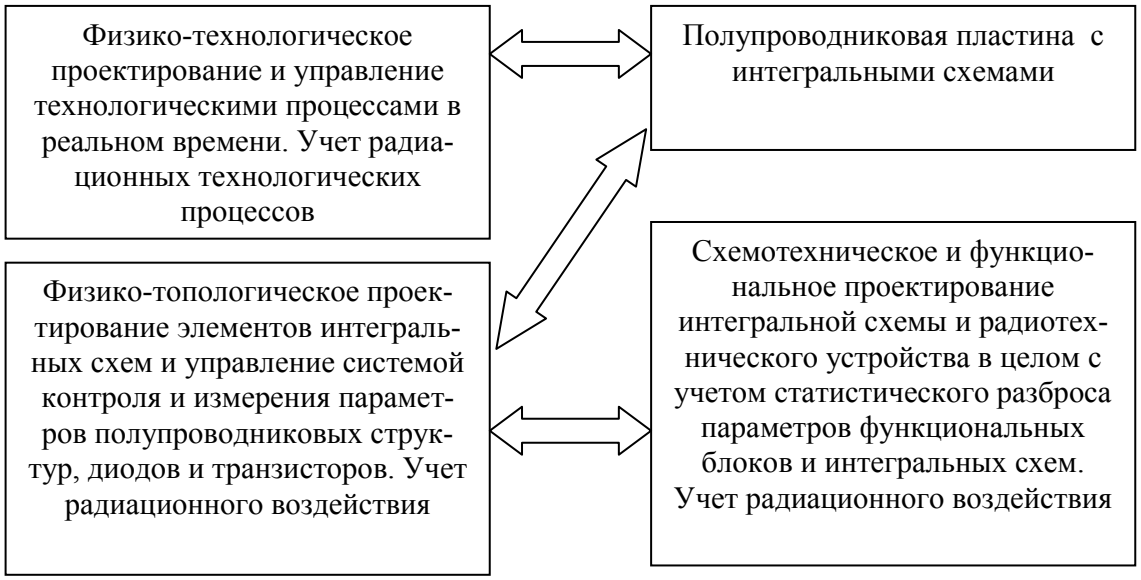


Рис. 2.2. Схема проектирования полупроводниковых диодов, транзисторов и интегральных схем

[15] Статистический метод Монте-Карло применяется для моделирования возникновения радиационных дефектов при облучении протонами, ионами и нейтронами. Для расчетов применяют приближение твердых шаров – моделируют случайные процессы выбивания и расталкивание атомов в кристаллической решетке полупроводника по аналогии с процессами движения и столкновения шаров на бильярдном столе. В силу большого количества движущихся одновременно атомов (до 10^4 штук) задача достаточно ресурсоемка и требует использования высокопроизводительных вычислений

Для всех типов моделей необходима формализация задачи, приводящая в общем случае к системе нелинейных уравнений большой размерности. Так, например, простейшая система уравнений переноса носителей заряда в полупроводниковых приборах – так называемая диффузионно-дрейфовая модель – включает три уравнения в частных производных: уравнение Пуассона и уравнения непрерывности для электронов и дырок. Более сложная квазигидродинамическая модель содержит восемь уравнений в частных производных: уравнение Пуассона, уравнения непрерывности, уравнения баланса энергии и импульса для электронов и дырок, а также уравнение теплопроводности. Для небольшой двумерной пространственной сетки 32×32 узла получаем 3072 дифференциально-алгебраических уравнения для диффузионно-дрейфовой модели и 8192 дифференциально-алгебраических уравнения для квазигидродинамической модели.

Характерным временным масштабом диффузионно-дрейфовой модели является время жизни неосновных носителей заряда, лежащее, обычно, в микросекундном диапазоне. Характерным временным масштабом квазигидродинамической модели являются времена релаксации энергии и импульса носителей заряда, лежащие в субпикосекундном диапазоне.

Предположим, что для решения одного уравнения на каждом временном шаге требуется 1000 операций с плавающей точкой. Тогда для моделирования работы полупроводникового прибора в течение одной миллисекунды в диффузионно-дрейфовом приближении будет произведено порядка 10^9, а в квазигидродинамическом – 10^{17} операций с плавающей точкой.

При этом для учета радиационных эффектов в исходной системе уравнений необходимо менять значения параметров, которые, в свою очередь, рассчитываются с использованием еще более ресурсоемкого метода частиц на основе статистической процедуры Монте-Карло. Другим примером является схемотехническое моделирование. В простейшем случае каждый элемент электрической схемы описывается дифференциальным уравнением, а каждый узел – алгебраическим уравнением. Тогда получается, что моделирование функционально законченного узла интегральной схемы требует решения на каждом временном шаге $10^3 \ldots 10^4$ дифференциально-алгебраических уравнений. По вычислительной сложности данная задача аналогична физико-топологическим моделям одного транзистора.

На физико-технологическом уровне также существуют вычислительно сложные задачи, связанные с расчетом диффузии примесей, окислением, процессами ионной имплантации и др. Тенденция к увеличению объема вычислений, диктуемая возрастающей сложностью адекватных математических моделей, требует применения радикальных мер повышения производительности, основной из которых в настоящее время является распараллеливание численных алгоритмов. Использование стандартных библиотек программ решения систем дифференциальных, нелинейных и линейных алгебраических уравнений, многие из которых содержат эффективные параллельные алгоритмы, определяет применение данного подхода в области высокопроизводительных вычислений на супер-ЭВМ в качестве базового.

Очевидно, что эффективность численных вычислений определяется в решающей степени алгоритмом решения математической задачи. Применение супер-ЭВМ наиболее целесообразно при использовании максимально сложных и точных моделей. На практике всегда присутствует неконтролируемый разброс исходных технологических параметров, который снижает требования к точности моделей и обусловливает применение упрощенных аналитических и численных методов решения. Однако сама возможность учета статистических разбросов стимулирует разработку более сложных моделей полупроводниковых приборов.

Применение высокопроизводительных вычислений необходимо как на этапе физико-технологического и физико-топологического проектирования самих полупроводниковых элементов и интегральных схем, так и для разработки автоматизированных систем управления технологическими процессами реального времени, т.е. для решения достаточно разнородных задач. При этом наличие супер-ЭВМ не отрицает моделирования на персональных компьютерах. Более того, модели должны быть различными, разработка универсальной модели невозможна и нецелесообразна ввиду ее крайней сложности и неэффективности.

Таким образом, супер-ЭВМ и высокопроизводительные рабочие станции прочно заняли свою нишу в области проектирования и изготовления изделий микроэлектроники, позволяя создавать новые и совершенствовать существующие полупроводниковые приборы и интегральные схемы. Пожалуй, именно в области микроэлектроники особенно ярко проявляется взаимосвязь между развитием вычислительной техники и моделированием: большие вычислительные мощности стимулируют разработку более сложных моделей, которые, в свою очередь, позволяют создавать более мощные суперкомпьютеры.

2.4. Численные методы решения задачи переноса носителей заряда в полупроводниковых приборах при воздействии проникающих излучений

В данном разделе рассмотрены вопросы нормировки и выбора базиса переменных системы уравнений переноса носителей заряда, сведение системы уравнений переноса носителей заряда к дифференциально-алгебраической системе уравнений, методы решения системы дифференциально-алгебраической уравнений и перспективы использования параллельных вычислений.

2.4.1. Нормировка и выбор базиса переменных системы уравнений переноса носителей заряда

При построении численных алгоритмов удобно пользоваться системой уравнений переноса носителей заряда, приведенной к безразмерному виду. Это позволяет не только уменьшить мантиссы обрабатываемых чисел до приемлемых величин, пригодных для цифровой обработки, что особо актуально для вычислительных архитектур, использующих числа одинарной точности

(например, технология параллельных вычислений CUDA корпорации NVidia [12-15]), но также избавиться от некоторых постоянных размерных коэффициентов в уравнениях модели. Нормировочные коэффициенты, используемые в работе, приведены в таблице 2.1.

Таблица 2.1. Нормировка переменных уравнений переноса носителей заряда

Модель переноса			Нормируемая величина	Условное обозначение	Нормировочный коэффициент
Квазигидродинамическая модель	Электротепловая модель	Диффузионно-дрейфовая модель	Пространство	x	x_0
			Время	t	$t_0 = \frac{x_0^2}{\varphi_0 \mu_0}$
			Подвижность	μ_n, μ_p	$\mu_0 = \max\left(\mu_n(x), \mu_p(x)\right)$
			Концентрация	n, p, N_d, N_a	$N_0 = \max\left(N_d(x), N_a(x)\right)$
			Потенциал	φ	$\varphi_0 = \frac{k_B T_0}{q}$
			Плотность тока	j_n, j_p	$j_0 = \frac{q \mu_0 N_0 \varphi_0}{x_0}$
			Температура	T	T_0
			Энергия	W_n, W_p	$W_0 = \frac{3}{2} k_B T_0$
			Плотность потока энергии	S_n, S_p	$S_0 = \frac{\mu_0 n_0 W_0 \varphi_0}{x_0}$

Искомыми функциями при решении системы уравнений переноса носителей заряда являются зависимости от координат и времени

- в диффузионно-дрейфовом приближении: потенциала электрического поля и концентрации электронов и дырок;
- в электротепловом приближении: потенциала электрического поля, концентрации электронов и дырок и температуры кристаллической решетки;
- в квазигидродинамическом приближении: потенциала электрического поля, концентрации электронов и дырок, температуры кристаллической решетки, средней энергии электронов и дырок

при заданных зависимостях электрофизических параметров полупроводника, начальных и граничных условиях.

Множество исходных функций системы уравнений переноса носителей заряда принято называть базисом переменных. В диффузионно-дрейфовом приближении распространены три базиса:

- $\{\varphi, n, p\}$ – потенциал электрического поля, концентрации электронов и дырок;
- $\{\varphi, \varphi_n, \varphi_p\}$ – потенциал электрического поля, квазиуровни Ферми для электронов и дырок;
- $\{\varphi, \Phi_n, \Phi_p\}$ – потенциал электрического поля, экспоненты квазиуровней Ферми для электронов и дырок.

В электротепловом и квазигидродинамическом приближениях принято использовать базисы $\{\varphi, n, p, T\}$ и $\{\varphi, n, p, T, W_n, W_p\}$, $\{\varphi, n, p, T, W_n, W_p, \upsilon_n, \upsilon_p\}$, соответственно.

2.4.2. Сведение системы уравнений переноса носителей заряда к дифференциально-алгебраической системе уравнений

Любая система дифференциальных уравнений в частных производных параболического и эллиптического типов, к числу которых относится семейство систем уравнений переноса носителей заряда в полупроводниках, может быть сведена к системе дифференциально-алгебраических уравнений вида

$$M\frac{du}{dt} = f(u,t) \qquad (2.1)$$

путем дискретизации пространственной части уравнений в частных производных. В выражении (2.1) u – вектор нормированных искомых переменных, M – матрица массы, f – векторная функция нормированных пространственно-дискретизированных правых частей системы уравнений переноса носителей заряда, t – нормированное время.

В настоящее время применяются 3 метода пространственной дискретизации:

- конечных объемов (FVM);
- конечных элементов (FEM);
- конечных разностей (FDM).

Дискретизация определяется формой записи системы дифференциальных уравнений в частных производных: для дифференциальной формы записи дискретизация осуществляется при помощи конечно-разностных методов; интегральная форма записи дискретизируется при помощи метода конечных объемов, а вариационная форма – при помощи метода конечных элементов.

Методы дискретизации, их преимущества и недостатки, широко рассмотрены в литературе. Применительно к одномерной квазигидродинамической модели переноса носителей заряда, рассматриваемой ниже, явным преимуществом будет обладать метод конечных разностей. Основной акцент сделаем на дискретизации плотностей токов и потоков энергии носителей заряда. Поэтому без ущерба общности обсуждения дальнейших результатов исключим из системы уравнения баланса импульса электронов и дырок, а также уравнение теплопроводности кристаллической решетки.

На отрезке [0, 1] введем неравномерную пространственную сетку с узлами $x = \{x_i\}$, $i = 0, 1, \ldots, N-1$, согласованную с распределением легирующей примеси: вблизи границ раздела и контактов полупроводникового прибора шаг сетки делаем более мелким. Определим значения базисных функций $\{\varphi, n, p, W_n, W_p\}$ в следующем виде: $u = \{u_i\}$, $u_{5i} = \varphi_i$, $u_{5i+1} = n_i$, $u_{5i+2} = p_i$, $u_{5i+3} = (W_n)_i$, $u_{5i+2} = (W_p)_i$. Тогда векторная функция нормированных пространственно-дискретизированных правых частей системы уравнений переноса носителей заряда в квазигидродинамическом приближении записывается в виде

$$f_{5i}=\left(\frac{2\varepsilon_0\varphi_0}{qN_0L_0^2}\right)\frac{\varepsilon_{i+1/2}\dfrac{u_{5i+5}-u_{5i}}{x_{i+1}-x_i}-\varepsilon_{i-1/2}\dfrac{u_{5i}-u_{3i-5}}{x_i-x_{i-1}}}{x_{i+1}-x_{i-1}}+\left(\left(N_d\right)_i-\left(N_a\right)_i-u_{5i+1}+u_{5i+2}\right),$$

$$f_{5i+1}=2\frac{\left(j_n\right)_{i+1/2}-\left(j_n\right)_{i-1/2}}{x_{i+1}-x_{i-1}}-R_i+G_i,$$

$$f_{5i+2}=-2\frac{\left(j_p\right)_{i+1/2}-\left(j_p\right)_{i-1/2}}{x_{i+1}-x_{i-1}}-R_i+G_i,$$

$$f_{5i+3}=\frac{1}{u_{5i+1}}\left(2\frac{\left(S_n\right)_{i+1/2}-\left(S_n\right)_{i-1/2}}{x_{i+1}-x_{i-1}}+\frac{\left(\left(j_n\right)_{i+1/2}+\left(j_n\right)_{i-1/2}\right)\left(u_{5i-5}-u_{5i+5}\right)}{}-\right.$$
$$\left.-\frac{u_{5i+1}\left(u_{5i+3}-1\right)}{\tau_{Wn}}-R_iu_{5i+3}+G_iW_e-u_{5i+3}f_{5i+1}\right),\qquad(2.2)$$

$$f_{5i+4}=\frac{1}{u_{5i+2}}\left(-2\frac{\left(S_p\right)_{i+1/2}-\left(S_p\right)_{i-1/2}}{x_{i+1}-x_{i-1}}+\frac{\left(\left(j_p\right)_{i+1/2}+\left(j_p\right)_{i-1/2}\right)\left(u_{5i-5}-u_{5i+5}\right)}{}-\right.$$
$$\left.-\frac{u_{5i+2}\left(u_{5i+4}-1\right)}{\tau_{Wp}}-R_iu_{5i+4}+G_iW_h-u_{5i+4}f_{5i+2}\right),$$

$$R_i=\left(u_{3i+1}u_{3i+2}-n_i^2\right)\left(\frac{1}{\tau_p\left(u_{3i+1}+n_i\right)+\tau_n\left(u_{3i+2}+n_i\right)}+C_nu_{3i+1}+C_pu_{3i+2}\right),$$

а матрица массы становится диагональной с элементами $m_{5i,5i} = 0$, $m_{5i+1,5i+1} = 1$, $m_{5i+2,5i+2} = 1$, $m_{5i+3,5i+3} = 1$, $m_{5i+4,5i+4} = 1$.

Дискретные аналоги плотностей токов и потоков энергии электронов и дырок определяются в ячейках сетки на основе аппроксимации Шарфеттера-Гуммеля (SG) [16], позволяющей сохранить монотонность численной схемы независимо от величины шага пространственной дискретизации и имеющей ряд особенностей для квазигидродинамического приближения [17]

$$\left(j_n\right)_{i+1/2}=\begin{cases}\dfrac{u_{5i+8}-u_{5i+3}+u_{5i}-u_{5i+5}}{x_{i+1}-x_i}\dfrac{\left(\left(\mu_n\right)_{i+1}u_{5i+6}-\left(\mu_n\right)_iu_{5i+1}\times h_n\left(u_{5i},u_{5i+3},u_{5i+5},u_{5i+8}\right)\right)}{1-h_n\left(u_{5i},u_{5i+3},u_{5i+5},u_{5i+8}\right)} & h_n(\ldots)\neq1\\ \dfrac{\left(\mu_n\right)_{i+1}u_{5i+6}u_{5i+8}-\left(\mu_n\right)_iu_{5i+1}u_{5i+3}}{x_{i+1}-x_i} & h_n(\ldots)=1\end{cases},$$

$$\left(j_p\right)_{i+1/2}=\begin{cases}\dfrac{u_{5i}-u_{5i+5}-u_{5i+9}+u_{5i+4}}{x_{i+1}-x_i}\dfrac{\left(\left(\mu_p\right)_{i+1}u_{5i+7}-\left(\mu_p\right)_iu_{5i+2}\times h_p\left(u_{5i},u_{5i+4},u_{5i+5},u_{5i+9}\right)\right)}{1-h_p\left(u_{5i},u_{5i+4},u_{5i+5},u_{5i+9}\right)} & h_p(\ldots)\neq1\\ \dfrac{\left(\mu_p\right)_iu_{5i+2}u_{5i+4}-\left(\mu_p\right)_{i+1}u_{5i+7}u_{5i+9}}{x_{i+1}-x_i} & h_p(\ldots)=1\end{cases},$$

$$\left(S_n\right)_{i+1/2}=\begin{cases}\dfrac{u_{5i+8}-u_{5i+3}+u_{5i}-u_{5i+5}}{x_{i+1}-x_i}\dfrac{\left(\left(\mu_n\right)_{i+1}u_{5i+6}u_{5i+8}-\left(\mu_n\right)_iu_{5i+1}u_{5i+3}\times h_n\left(u_{5i},u_{5i+3},u_{5i+5},u_{5i+8}\right)\right)}{1-h_n\left(u_{5i},u_{5i+3},u_{5i+5},u_{5i+8}\right)} & h_n(\ldots)\neq1\\ \dfrac{\left(\mu_n\right)_{i+1}u_{5i+6}u_{5i+8}^2-\left(\mu_n\right)_iu_{5i+1}u_{5i+3}^2}{x_{i+1}-x_i} & h_n(\ldots)=1\end{cases},\qquad(2.3)$$

$$\left(S_p\right)_{i+1/2}=\begin{cases}\dfrac{u_{5i}-u_{5i+5}-u_{5i+9}+u_{5i+4}}{x_{i+1}-x_i}\dfrac{\left(\left(\mu_p\right)_{i+1}u_{5i+7}u_{5i+9}-\left(\mu_p\right)_iu_{5i+2}u_{5i+4}\times h_p\left(u_{5i},u_{5i+4},u_{5i+5},u_{5i+9}\right)\right)}{1-h_p\left(u_{5i},u_{5i+4},u_{5i+5},u_{5i+9}\right)} & h_p(\ldots)\neq1\\ \dfrac{\left(\mu_p\right)_iu_{5i+2}u_{5i+4}^2-\left(\mu_p\right)_{i+1}u_{5i+7}u_{5i+9}^2}{x_{i+1}-x_i} & h_p(\ldots)=1\end{cases},$$

$$h_n\left(u_{5i},u_{5i+3},u_{5i+5},u_{5i+8}\right)=\exp\left(-2\frac{u_{5i+8}-u_{5i+3}+u_{5i}-u_{5i+5}}{u_{5i+8}+u_{5i+3}}\right),$$

$$h_p\left(u_{5i},u_{5i+4},u_{5i+5},u_{5i+9}\right)=\exp\left(2\frac{u_{5i}-u_{5i+5}-u_{5i+9}+u_{5i+4}}{u_{5i+9}+u_{5i+4}}\right).$$

Отметим, что если положить равенство средних энергий электронов и дырок тепловой энергии носителей заряда в каждом узле расчетной сетки $u_{5i+3}\equiv u_{5i+4}\equiv 1$ и, следовательно, постоянные их подвижности $\mu_i\equiv\mu_{i+1}\equiv\mu_{i+1/2}$, то выражения для плотностей токов переходят в традиционный для диффузионно-дрейфовой модели вид

$$\left(j_n\right)_{i+1/2}=\begin{cases}\left(\mu_n\right)_{i+1/2}\dfrac{u_{5i}-u_{5i+5}}{x_{i+1}-x_i}\dfrac{\left(u_{5i+6}-u_{5i+1}\exp\left(u_{5i+5}-u_{5i}\right)\right)}{1-\exp\left(u_{5i+5}-u_{5i}\right)} & u_{5i+5}\neq u_{5i}\\ \left(\mu_n\right)_{i+1/2}\dfrac{u_{5i+6}-u_{5i+1}}{x_{i+1}-x_i} & u_{5i+5}=u_{5i}\end{cases},$$

$$\left(j_p\right)_{i+1/2}=\begin{cases}\left(\mu_p\right)_{i+1/2}\dfrac{u_{5i}-u_{5i+5}}{x_{i+1}-x_i}\dfrac{\left(u_{5i+7}-u_{5i+2}\exp\left(u_{5i}-u_{5i+5}\right)\right)}{1-\exp\left(u_{5i}-u_{5i+5}\right)} & u_{5i+5}\neq u_{5i}\\ \left(\mu_p\right)_{i+1/2}\dfrac{u_{5i+2}-u_{5i+7}}{x_{i+1}-x_i} & u_{5i+5}=u_{5i}\end{cases}. \qquad (2.4)$$

2.4.3. Методы решения системы дифференциально-алгебраической уравнений

Любая гидродинамическая модель переноса носителей заряда в полупроводнике включает уравнение Пуассона, определяющее алгебраическую компоненту дискретизированной задачи (2.1). Матрица массы дифференциально-алгебраических уравнений является вырожденной, что определяет «бесконечную жесткость» задачи (2.1). При построении разностных схем для жестких систем предъявляются повышенные требования к устойчивости решения – явные схемы для решения жестких задач требуют очень мелкого шага интегрирования и поэтому практически неприменимы.

2.4.3.1. Неявные итерационные схемы

Среди неявных схем широко распространены методы Рунге-Кутты (IRK – implicit Runge-Kutta methods). Любой неявный метод Рунге-Кутты для перехода на новый временной слой требует решения системы нелинейных алгебраических уравнений при помощи итераций ньютоновского типа. Для s-стадийного неявного метода Рунге-Кутты минимальное число возникающих нелинейных систем *s* соответствует диагонально неявным методам (DIRK – diagonal implicit Runge-Kutta methods). Именно они чаще всего и используются на практике.

2.4.3.2. Многошаговые методы

Неявные многошаговые методы (формулы дифференцирования назад) положены в основу популярных программ Гира. Коэффициенты многошаговых методов подбираются так, чтобы q-шаговый метод имел точность $O(\tau^q)$. Однако

можно показать [19], что неявные многошаговые методы с $q > 2$ теряют свойство безусловной устойчивости. При $q > 6$ неявные многошаговые методы становятся абсолютно неустойчивыми. Этим неявные многошаговые методы сильно уступают неявным методам Кунге-Кутты. Тем не менее, неявные многошаговые методы первого и второго порядков (неявный метод Эйлера и неявное правило трапеций) реализованы в системе автоматизированного проектирования изделий микроэлектроники TCAD Sentaurus фирмы Synopsys [20] для решения задачи переноса носителей заряда. Также на базе методов Гира реализован комплекс программ оценки радиационного воздействия на изделия микроэлектроники DIODE-2D разработки НИЯУ МИФИ [8].

2.4.3.3. Безитерационные схемы

Наличие итераций сильно усложняет использование неявных методов Рунге-Кутты и многошаговых методов, так как к проблемам устойчивости добавляется проблема сходимости итерационного процесса при решении систем нелинейных алгебраических уравнений. Альтернатива, которая обходит эту трудность – методы типа Розенброка (ROS) и Розенброка-Ваннера (ROW) [18, 19]. Формально эти схемы неявные, но итераций в них не возникает и число арифметических действий для перехода на новый временной слой фиксировано и заранее известно (как в явных схемах). За это безусловное преимущество эти схемы получили название явно-неявных или полунеявных.

Формулы перехода на новый временной слой однопараметрического семейства одностадийных схем Розенброка имеют вид [21]

$$\begin{gathered} \hat{u} = u + \tau \operatorname{Re} k, \\ \left(M - \alpha\tau f_u'(u,t)\right)k = f\left(u, t + 0.5\tau\right), \end{gathered} \tag{2.5}$$

где $f_u' \equiv \dfrac{\partial f}{\partial u}$ – матрица Якоби, τ – шаг по времени, u – решение на текущем временном слое, $\hat{u}$ – решение на новом временном слое, α – числовой параметр, определяющий свойства схемы. При $\alpha = 0$ это явная схема, имеющая точность $O(\tau)$. Этот вариант схемы практически непригоден для расчета жестких задач. При $\alpha = 0{,}5$ получается известная схема «с полусуммой». Она имеет точность $O(\tau^2)$ и безусловно устойчива. При $\alpha = 1$ имеем неявный метод Эйлера. Помимо безусловной устойчивости он имеет хорошее качественное поведение численного решения (за счет L1-устойчивости). Однако неявный метод Эйлера имеет невысокую точность $O(\tau)$, что препятствует его применению.

Описанные выше схемы вещественны. Однако существует одна комплексная схема этого семейства с $\alpha = \dfrac{1+i}{2}$, которая обладает уникальными свойствами [18]: точность $O(\tau^2)$, L2-устойчивость и, соответственно, безусловная устойчивость. Эта схема обладает высокой надежностью и пригодна для расчетов с сильной жесткостью. В литературе ее принято называть одностадийной схемой Розенброка с комплексным коэффициентом (CROS).

Качественно поведение одностадийной схемы Розенброка с комплексным коэффициентом можно описать следующим образом. Для дифференциальной подсистемы делается один шаг точности $O(\tau^2)$, а для алгебраической – одна ньютоновская итерация. Для получения точности $O(\tau^2)$ на дифференциальной подсистеме в правой части (2.2) необходимо использовать момент $t + 0{,}5\tau$. Однако если использовать этот момент в алгебраической подсистеме, то ньютоновские итерации сходятся именно к этому моменту, а не к нужному значению $t + \tau$. Поэтому для алгебраической подсистемы необходимо использовать в правой части (2.2) момент $t + \tau$, что обеспечивает общий второй порядок точности.

Выбор шага интегрирования является важной задачей. С одной стороны, шаг должен быть достаточно малым для того, чтобы обеспечивать заданную точность вычислений; с другой стороны – достаточно большим, чтобы избежать бесполезной вычислительной работы. Традиционным является метод уменьшения шага вдвое, который использовал еще Рунге в своих пионерских работах. Пусть u_2 результат последовательного расчета двух шагов τ, w – результат расчета одного большого шага 2τ. Тогда норма погрешности err оценивается по формуле Ричардсона [18, 19]

$$err = \frac{1}{2^p - 1}\|u_2 - w\|. \qquad (2.6)$$

Затем величина нормы погрешности сравнивается с заданной величиной допустимой погрешности tol, что позволяет вычислить оптимальный шаг как $\tau\left(\frac{tol}{err}\right)^{\frac{1}{p+1}}$. На практике, однако данное выражение обычно помножают на «гарантийный фактор» $fac < 1$ для предотвращения резких изменений шага интегрирования. В данной работе $fac = 0{,}25^{\frac{1}{p+1}}$ и выражение для нового шага интегрирования

$$\tau_{new} = \tau\left(\frac{1}{4}\frac{tol}{err}\right)^{\frac{1}{p+1}}. \qquad (2.7)$$

Таким образом, если $err < tol$, два вычисленных шага считаются принятыми, и решение продолжается исходя из u_2. В противном случае оба шага отбрасываются, и вычисления повторяются с шагом τ_{new}.

2.4.4. Использование параллельных вычислений

На основе вышеизложенного подхода написаны компьютерные программы решения системы уравнений переноса носителей заряда в полупроводнике в диффузионно-дрейфовом и квазигидродинамическом приближениях. На начальном этапе вычислительный алгоритм был реализован на центральном процессоре в однопоточном режиме. Анализ его производительности показал, что основное время вычислений тратится на решение системы линейных

алгебраических уравнений. С ростом числа узлов расчетной сетки N время вычислений увеличивалось пропорционально N^3.

Практика моделирования полупроводниковых приборов позволяет выделить следующие причины для применения расчетных сеток с большим числом узлов:

- наличие в современных приборах областей с резким изменением электрофизических параметров материала: гетеропереходов (скачек диэлектрической проницаемости), резких p-n-переходов (скачек уровня легирования) и т.п., приводит к тому, что имеет место скачок напряженности электрического поля, и, как следствие, резкое изменение электростатического потенциала, концентрации электронов и дырок и их средних энергий. Поэтому корректное моделирование переноса носителей заряда в таких областях возможно лишь при уменьшении шага пространственной дискретизации, что приводит к общему увеличению числа узлов расчетной сетки;
- уменьшение топологических норм изготовления активных элементов микросхем до значений менее 1 мкм повлекло за собой усиление взаимовлияния различных областей их топологии, что существенно изменяет и электрические характеристик элементов. Ранее этими эффектами можно было пренебречь без потери точности моделирования, но для корректного моделирования современных элементов необходимы двумерная и даже трехмерная постановки, а значит увеличение числа узлов расчетной сетки;
- внедрение приборов со сложной латеральной топологией, например, полевых транзисторов с V-образной канавкой затвора. В этом случае для построения корректной численной модели используется более сложная и мелкая сетка пространственной дискретизации уравнений, что обеспечивается увеличением числа ее узлов.

Также увеличение предельных рабочих частот полупроводниковых приборов и интегральных схем, обеспечиваемое сокращением размеров их рабочих областей, требует уменьшения временного шага интегрирования системы уравнений переноса носителей заряда при моделировании высокочастотных и переходных процессов.

Все вышеизложенные факторы приводят к увеличению времени расчета статических, высокочастотных и переходных характеристик перспективных полупроводниковых приборов и интегральных схем. Сокращение времени моделирования может быть достигнуто оптимизацией численного алгоритма решения системы дифференциально-алгебраических уравнений (2.1), применением высокопроизводительных ЭВМ и распараллеливанием вычислений, прежде всего, операций линейной алгебры.

В течение многих лет одним из основных методом повышения производительности персональных компьютеров и рабочих станций было увеличение тактовой частоты центрального процессора, которая с 1978 г. до 2005 г. возросла с 4,77 МГц (процессор Intel 8086) до 3,8 ГГц (процессор Intel Pentium 4), то есть примерно в 1000 раз. Однако из-за фундаментальных физических ограничений (рост потребляемой мощности и тепловыделения, быстро приближающийся физический предел размера транзистора) в последние годы производи-

тели центральных процессоров оказались перед необходимостью поиска замены этому традиционному источнику быстродействия.

В 2005 г. ведущие производители центральных процессоров стали предлагать процессоры с двумя вычислительными ядрами вместо одного. До этого времени только мощные серверы и суперкомпьютеры использовали многопроцессорные и многоядерные вычислительные архитектуры. В последующие годы тенденция увеличения числа ядер центрального процессора закрепилась и в настоящее время на рынке персональных компьютеров представлены 2-, 4- и 6-ядерные процессорные системы. В будущем число ядер центрального процессора будет только возрастать. Программная реализация вычислений на многоядерных и многопроцессорных архитектурах реализуется на основе протокола MPI (Message Passing Interface).

В отличие от центральных процессоров графические процессоры были изначально ориентированы на параллельные вычисления, что обусловлено особенностями обработки графической информации при выводе ее на дисплей. Однако специфическая модель программирования через графические программные интерфейсы OpenGL и DirectX была слишком ограничивающей для большинства разработчиков приложений для научных расчетов.

В ноябре 2006 г. корпорация NVidia выпустила первую видеокарту GeForce 8800 GTX с поддержкой технологии CUDA. В отличие от предыдущих поколений графических процессоров, в которых вычислительные ресурс подразделялись на вершинные и пиксельные шейдеры, в архитектуру CUDA включен унифицированный шейдерный конвейер, позволяющей программе, выполняющей вычисления общего назначения, задействовать любое арифметически-логическое устройство, входящее в состав графического процессора. Так как планировалось, что новое семейство графических процессоров будет использоваться для вычислений общего назначения, то арифметически-логические устройства были сконструированы с учетом требований стандарта IEEE к арифметическим операциям над числами с плавающей точкой одинарной (а позднее и двойной) точности и разработан набор команд, ориентированный на вычисления общего назначения, а не только на графику. Также исполняющим устройствам графических процессоров был разрешен произвольный доступ к памяти для чтения и записи, а также доступ к программно-управляемой кеш-памяти.

Для облегчения доступа разработчиков программного обеспечения к ресурсам технологии CUDA в 2007 г. корпорация NVidia выпустила компилятор расширенной версии языка C/C++, получивший название CUDA C. Несколько позднее был выпущен компилятор CUDA Fortran.

Сравнение роста производительности графических и центральных процессоров представлено на рисунке 2.3. В настоящее время производительность «настольных суперкомпьютеров» на базе специализированных графических карт Tesla K40 корпорации NVidia достигла 4,3 Тф/с [15].

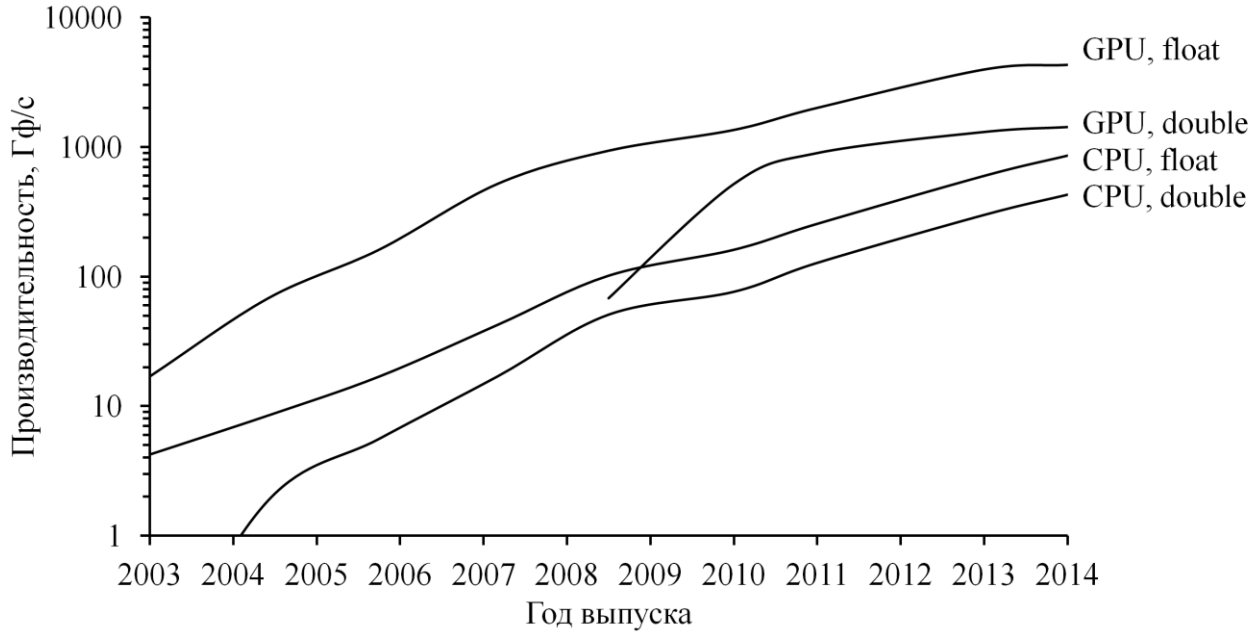


Рис. 2.3. Сравнение роста производительности графических и центральных процессоров за последнее десятилетие: GPU, float – графический процессор одинарная точность, GPU, double – графический процессор двойная точность, CPU, float – центральный процессор одинарная точность, CPU, double – центральный процессор двойная точность [12-15, 22]

Максимальное ускорение, которое можно получить от распараллеливания программы, дается законом Амдала [12]

$$S = \frac{1}{1 - P + \frac{P}{N}}, \tag{2.8}$$

где P – доля вычислений, которая может быть идеально распараллелена на N независимых потоков (нитей).

Таким образом, учитывая данные, приведенные на рисунке 2.3, выражение (2.8), то обстоятельство, что число физических ядер современных графических процессоров в разы превосходит количество ядер центральных процессоров, а скорость переключения между выполняемыми нитями на графическим процессоре на порядки превосходит аналогичный показатель для центральных процессоров, задача распараллеливания вычислений с применением графических процессоров для научных и инженерных расчетов является актуальной.

2.5. Результаты моделирования

В данном разделе рассмотрены результаты моделирования с применением высокопроизводительных вычислений на базе технологии CUDA корпорации NVidia.

2.5.1. Решение уравнения Пуассона с применением технологии CUDA массивно-параллельных вычислений

На предварительном этапе для оценки эффективности распараллеливания вычислений и погрешности расчетов отдельно решалось двумерное уравнение Пуассона. Решение данной задачи также имеет самостоятельное значение, так как уравнение Пуассона описывает распределение потенциала электрического поля в задачах электростатики и стационарное распределение температуры, например, в изделиях микроэлектроники.

В качестве тестовой задачи рассматривалось уравнение

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=\frac{\left(x-\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2-2\sigma^2}{\sigma^4}\exp\left(-\frac{\left(x-\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2}{2\sigma^2}\right) \tag{2.9}$$

в области $x, y \in [0, 1]$ с периодическими граничными условиями, имеющее точное решение

$$\varphi_0(x,y)=\exp\left(-\frac{\left(x-\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2}{2\sigma^2}\right). \tag{2.10}$$

На рисунке 2.4 приведена зависимость погрешности численного решения тестовой задачи (2.9) в сравнении с точным решением по формуле (2.10) от числа узлов расчетной сетки для различных алгоритмов. Видно, что увеличение числа узлов расчетной сетки уменьшает погрешность решения уравнения Пуассона при применении алгоритма быстрого преобразования Фурье. Погрешность численного расчета решения задачи (2.9) методом простой итерации зависит от числа узлов расчетной сетки и количества итераций. При заданном числе узлов имеет место снижение погрешности до некоторой константы с ростом числа итераций. Общей закономерностью является требование увеличения числа итераций для обеспечения заданной точности решения с ростом числа узлов расчетной сетки, что объясняется ухудшением обусловленности матрицы.

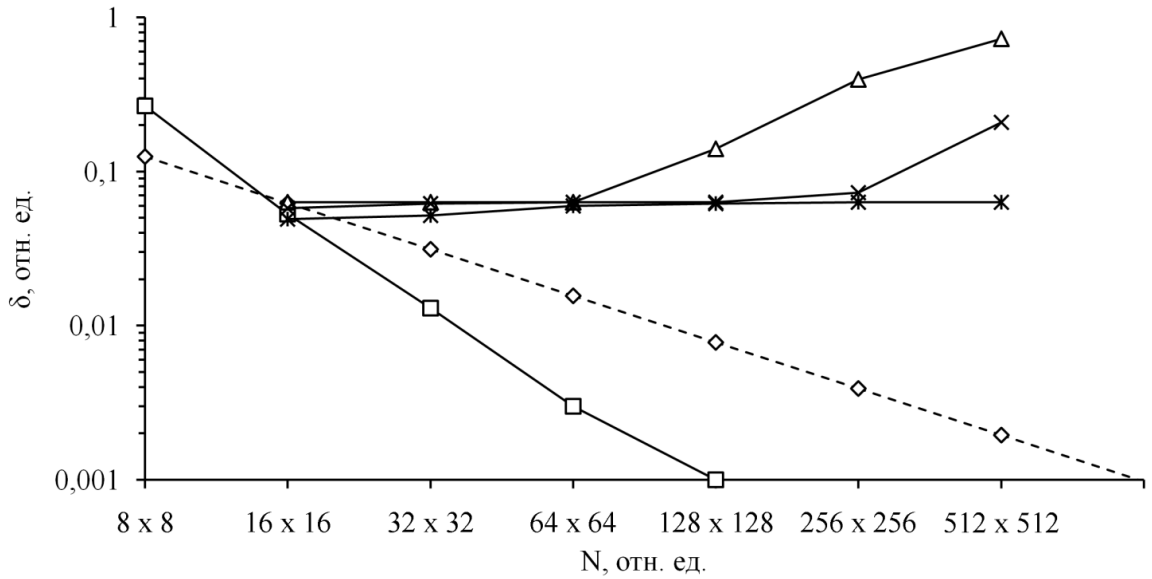


Рис. 2.4. Зависимость погрешности численного решения тестовой задачи от числа узлов расчетной сетки: ◊ – оценка погрешности $1/N$, □ – метод преобразования Фурье, Δ – метод простой итерации (2000 итераций), × – метод простой итерации (20000 итераций), ж – метод простой итерации (200000 итераций)

На рисунке 2.5 приведена зависимость времени решения задачи (2.9) от числа узлов расчетной сетки для различных алгоритмов. Видно, что применение быстрого преобразования Фурье для решения уравнения Пуассона позволяет резко снизить время вычисления по сравнению с более общим итерационным алгоритмом. К сожалению, метод решения уравнения Пуассона на основе быстрого преобразования Фурье накладывает жесткие ограничения на граничные условия задачи, которые должны быть нулевыми или периодическими, что сужает применимость данного алгоритма. В целом увеличение производительности при решении уравнения Пуассона на графической карте составляет около 10 раз. Для тестирования использовалась система Intel® Core™ 2 Duo CPU T6500 @ 2.1 GHz GeForce GT 120M 1.25 GHz.

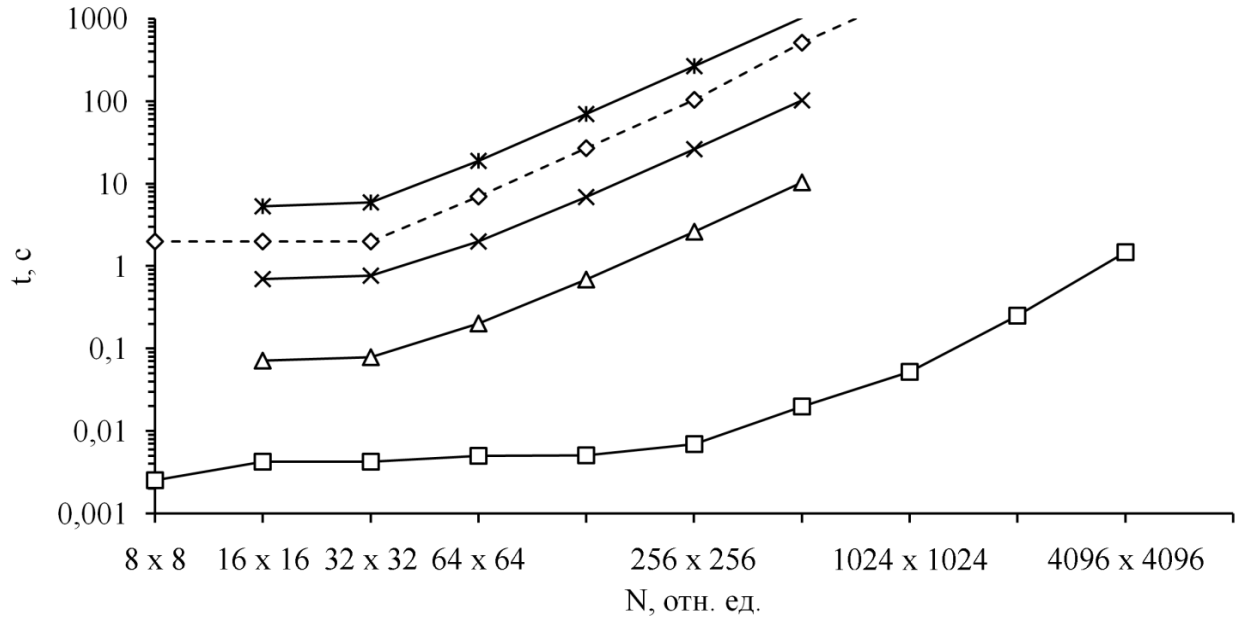


Рис. 2.5. Зависимость времени решения тестовой задачи от числа узлов расчетной сетки: ◊ – расчет методом простой итерации на центральном процессоре (20000 итераций), □ – метод преобразования Фурье, Δ – метод простой итерации (2000 итераций), × – метод простой итерации (20000 итераций), ж – метод простой итерации (200000 итераций)

2.5.2. Решение системы уравнений переноса носителей заряда в полупроводниковых приборах с применением технологии CUDA массивно-параллельных вычислений

Для исследования возможностей повышения производительности вычислений при применении технологии CUDA программы решения системы уравнений переноса носителей заряда в полупроводниковых приборах в диффузионно-дрейфовом и квазигидродинамическом приближениях были модернизированы. Наиболее вычислительно емкая процедура решения системы линейных алгебраических уравнений на каждом шаге интегрирования системы дифференциально-алгебраических уравнений была перенесена на графический процессор.

Характерными особенностями матрицы Якоби системы дифференциально-алгебраических уравнений, полученной из системы дифференциальных уравнений в частных производных являются большая размерность и разреженность. Это обусловливает применение итерационных методов для решения системы линейных алгебраических уравнений такого рода задач, например метода бисопряженных градиентов (BiCG). Так как изначально версия программы, целиком выполняемая на центральном процессоре, использовала LU-разложение для решения системы линейных алгебраических уравнений, для корректного сравнения была разработана программа, целиком выполняемая на центральном процессоре и использующая метод бисопряженных градиентов для решения системы линейных алгебраических уравнений. Таким образом, в тестировании участвовали три версии программы (таблица 2.2).

Таблица 2.2. Описание тестируемых программ

Номер версии	Метод решения системы линейных алгебраических уравнений	Используемый для расчетов процессор
1	исключения Гаусса (LU-разложение)	центральный
2	бисопряженных градиентов	центральный
3	бисопряженных градиентов	графический

Тестирование программ проводилось на двух персональных компьютерах с различной производительностью, основные характеристики которых представлены в таблице 2.3.

Таблица 2.3. Характеристики персональных компьютеров, используемых для тестирования программ

Номер ПК	Характеристики центрального процессора (ЦП) и оперативной памяти (ОЗУ)	Характеристики видеокарты
1	Intel Core 2 Duo, 2.8 ГГц, 2.8 ГГц ОЗУ – 2 Гб	NVidia GeForce 9500 GT
2*	Intel Core i7-2600, 3.4 ГГц, 3.7 ГГц ОЗУ – 8 Гб	NVidia GeForce GTX 560

* Производительность центрального и графического процессоров выше

В качестве объекта моделирования была рассмотрена кремниевая диодная структура с топологией p-области: $N_a = 10^{16}$ см$^{-3}$, $L_p = 50$ мкм; n-области: $N_d = 10^{16}$ см$^{-3}$, $L_n = 50$ мкм и электрофизическими параметрами кремния: $\mu_{0n} = 1440$ см^2/(В·с), $\tau_n = 10^{-7}$ с, $\mu_{0p} = 480$ см^2/(В·с), $\tau_p = 10^{-7}$ с. Коэффициенты Оже-рекомбинации предполагали равными $C_n = 1{,}1 \cdot 10^{-30}$ см^6/с и $C_p = 0{,}3 \cdot 10^{-30}$ см^6/с.

2.5.2.1. Исследование стационарного решения

Результаты моделирования воздействия стационарного ионизирующего излучения на тестовую диодную структуру, приведенные на рисунке 2.6, сравнивали с данными работы [8]. Качественный ход полученных зависимостей ионизационного тока от темпа генерации неравновесных носителей заряда совпадает с литературными данными. Имеющие место отличия результатов объясняются большим числом узлов расчетной сетки ($N = 8\ldots1024$ в данной работе против $N = 70$ в работе [8]) и меньшей нормой невязки, при которой решение считалось достигнутым ($\delta = 10^{-7}$ в данной работе против $\delta = 10^{-2}$ в работе [8]).

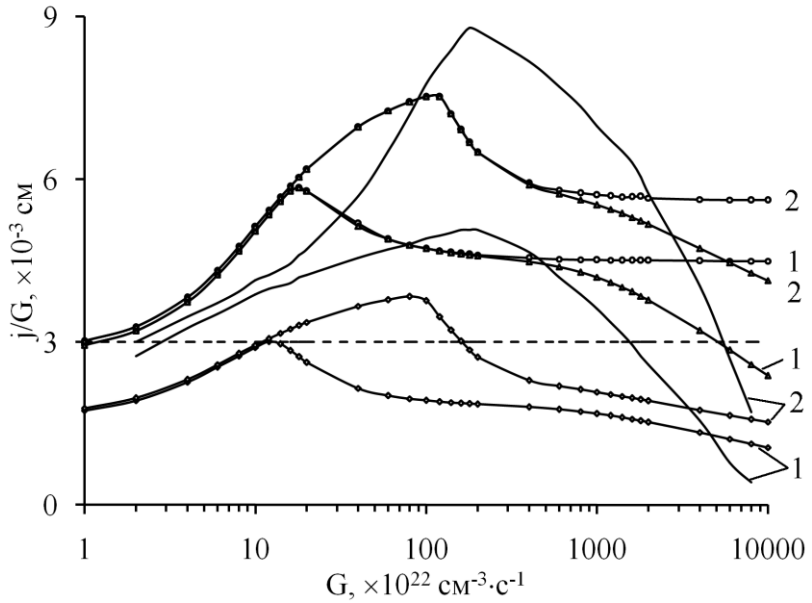


Рис. 2.6. Зависимость отношения нормированного на темп генерации плотности тока к темпу генерации электронно-дырочных пар от темпа генерации электронно-дырочных пар: 1 – U = 0 В, 2 – U = -5 В;

- - - – аналитическая модель [8];

—— – результаты численного моделирования [8];

—o— – $\mu_{0n} = \mu_{0p} = 832{,}5$ см^2/(В·с), $C_n = C_p = 0$ см^6/с;

—Δ— – $\mu_{0n} = \mu_{0p} = 832{,}5$ см^2/(В·с), $C_n = 1{,}1 \cdot 10^{-30}$ см^6/с, $C_p = 0{,}3 \cdot 10^{-30}$ см^6/с;

—◊— – $\mu_{0n} = 1440$ см^2/(В·с), $\mu_{0p} = 480$ см^2/(В·с), $C_n = 1{,}1 \cdot 10^{-30}$ см^6/с, $C_p = 0{,}3 \cdot 10^{-30}$ см^6/с

Задача решалась методом установления. Характерные времена решения в зависимости от числа узлов расчетной сетки приведены на рисунке 2.7.

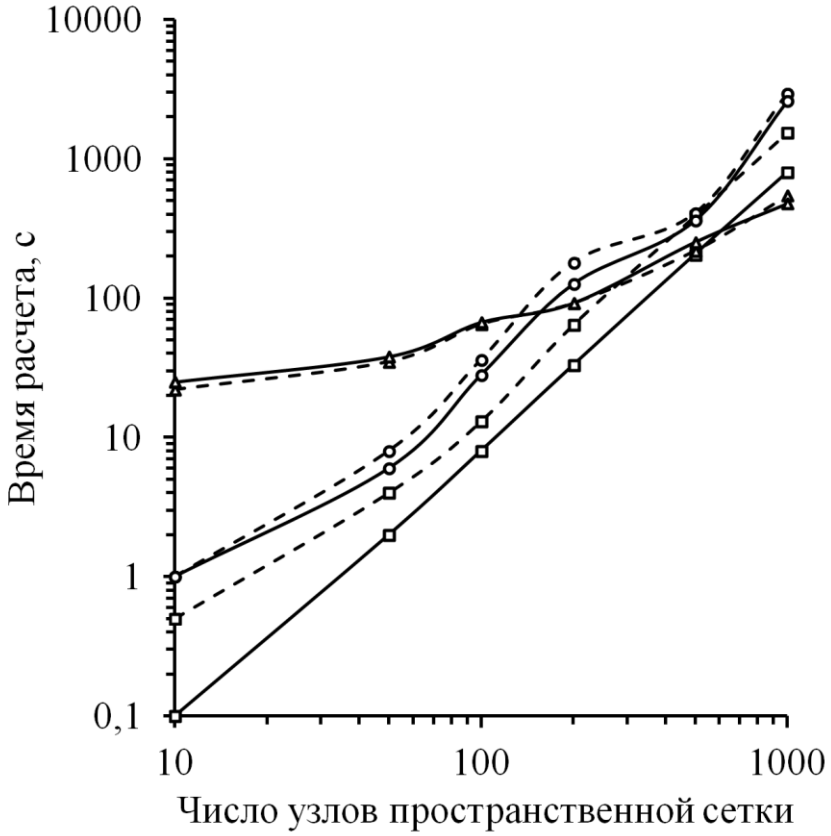


Рис. 2.7. Зависимость времени расчета при решении тестовой задачи от числа узлов пространственной сетки для трех версий программы и двух персональных компьютеров с различной производительностью: (- - -) – персональный компьютер 1; (—) – персональный компьютер 2; □ – версия программы 1; ○ – версия программы 2; Δ – версия программы 3

Из результатов, приведенных на рисунке 2.7, видно следующее:

- заметное сокращение времени расчета от применения распараллеливания вычислений на графическом процессоре при решении рассматриваемой нами тестовой задачи наблюдается при $N > 512$;
- с увеличением числа узлов пространственной сетки (при $N > 512$) прирост времени выполнения программы, полностью выполняемой на центральном процессоре, значительно больше, чем для программы, часть операций которой выполняется на графическом процессоре, независимо от метода решения системы линейных алгебраических уравнений (прямого или итерационного);
- обработка данных на графических процессорах с одинарной точностью не приводит к их заметному искажению по сравнению с обработкой на центральных процессорах с двойной точностью.

2.5.2.2. Исследование переходных процессов

Результаты моделирования воздействия биэкспоненциального импульса ионизирующего излучения при различных температурах кристаллической решетки полупроводника приведены на рисунке 2.8. При проведении расчетов учитывали температурные зависимости концентрации собственных носителей заряда в полупроводнике и подвижностей электронов и дырок.

Характерный временной шаг интегрирования системы дифференциально-алгебраических уравнений составил порядка 10^{-13} с на фронте импульса излучения с последующим увеличением до единиц наносекунд на его спаде. Аналогичный расчет на основе явной численной схемы с большей погрешностью требует временного шага интегрирования порядка 10^{-15} с. Отметим, что с уменьшением длительности фронта импульса ионизирующего излучения, например, при рассмотрении воздействия лазерного импульса наносекундной длительности, эффективность одностадийной схемы Розенброка с комплексным коэффициентом будет только возрастать.

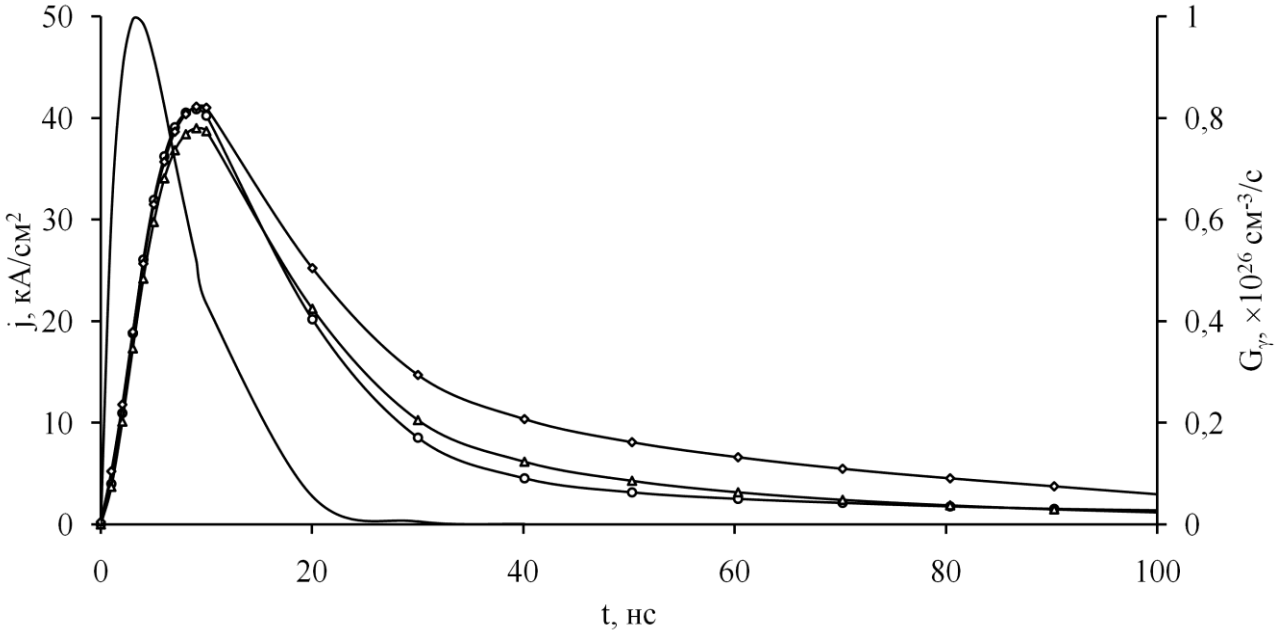


Рис. 2.8. Зависимость плотности ионизационного тока тестовой диодной структуры от времени для температур: —о— T = 300 K; —Δ— T = 560 K; —◊— T = 680 K; — – форма импульса ионизирующего излучения

Из графика видно, что увеличение температуры кристаллической решетки тестовой диодной структуры с 300 К до 700 К практически не меняет максимальное значение плотности мгновенной составляющей ионизационного тока, величина которого определяется максимальной интенсивностью излучения. Однако с ростом температуры увеличивается вклад запаздывающей компоненты ионизационного тока, приводящей, в конечном счете, к развитию радиационно-стимулированного лавинно-теплового пробоя.

ЗАКЛЮЧЕНИЕ

Сфера применения параллельных вычислений, в частности на графических процессорах, в вычислительной физике неуклонно расширяется. Это позволяет решать как уже существующие задачи с большей детализацией, так и ставить принципиально новые вычислительно сложные задачи. Представленная в данной работе одномерная квазигидродинамическая модель переноса носителей заряда в полупроводниковых приборах легко может быть обобщена на двумерный и трехмерный случай. При этом, по-видимому, более эффективной будет являться пространственная дискретизация методом конечных объемов, обладающая свойством естественной консервативности. Отметим, что расширение математической модели (введение новых уравнений, увеличение размерности задачи) не требует изменения постановки и решения системы дифференциально-алгебраических уравнений, а рост числа узлов расчетной сетки только увеличит эффективность параллельного алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. Волкова Е.В. Пузанов А.С., Оболенский С.В. Применение параллельных вычислений в задачах моделирования транспорта электронов в полупроводниках в условиях радиационного воздействия: учебное пособие. – Нижний Новгород, Издательство Нижегородского госуниверситета, 2014. – 60 с.
2. Зи С.М. Физика полупроводниковых приборов. – М. Мир, 1981. – 567 с.
3. Пожела Ю. Физика быстродействующих транзисторов. – Вильнюс: Мокслас, 1989. – 264 с.
4. Степаненко И.П. Основы Микроэлектроники. – М.: Лаборатория базовых знаний, 2001. – 488с.
5. Шур М. Современные приборы на основе арсенида галлия. – М.: Мир, 1991. – 632 с.
6. Полевые транзисторы на арсениде галлия / ред. Ди Лоренцо Д.В., Канделуола Д.Д. – М.: Радио и связь, 1988. – 496 с.
7. Бонч-Бруевич В.Л., Калашников С.Г. Физика полупроводников. – М.: Наука, 1990. – 559 с.
8. Технология СБИС / Пирс К., Адамс А., Кац Л., Цай Дж., Сейдел Т., Макгиллис Д.; под редакцией Зи С. – М: Мир, 1986. – 404 с.
9. Ricketts L.W. Fundamentals of Nuclear Hardening of Electronic Equipment. – Wiley-Interscience, 1972. – 548 P.
10. Коршунов Ф.П., Гатальский Г.В., Иванов Г.М. Радиационные эффекты в полупроводниковых приборах. – М.: Наука и техника, 1978. – 232 с.
11. Рикетс Л.У., Бриджес Дж.Э., Майлетта Дж. Электромагнитный импульс и методы защиты. – М.: Атомиздат, 1979. – 328 с.
12. Воеводин В.В. Вычислительная математика и структура алгоритмов. – М.: МГУ, 2010. – 168 с.
13. Гергель В.П. Высокопроизводительные вычисления для многопроцессорных многоядерных систем. – М.: МГУ, 2010. – 544 с.
14. Корняков К.В., Кустикова В.Д., Мееров И.Б., Сиднев А.А., Сысоев А.В., Шишков А.В. Инструменты параллельного программирования в системах с общей памятью. – М.: МГУ, 2010. – 272 с.
15. Линев А.В., Боголепов Д.К., Бастраков С.И. Технологии параллельного программирования для процессоров новых архитектур – М.: МГУ, 2010. – 160 с.
16. Боресков А.В., Харламов А.А. Основы работы с технологией CUDA. – М.: ДМК Пресс, 2011. – 232 с.
17. Сандерс Дж., Кэндрот Э. Технология CUDA в примерах: Введение в программирование графических процессоров. – М.: ДМК Пресс, 2011. – 232 с.
18. Гречников Е.А., Михайлов С.В., Нестеренко Ю.В., Поповян И.А. Вычислительно сложные задачи теории чисел. – М.: МГУ, 2012. – 312 с.
19. Боресков А.В., Харламов А.А. Марковский Н.Д., Микушин Д.Н., Мортиков Е.В., Мыльцев А.А., Сахарных Н.А., Фролов В.А. Параллельные вы-

числения на GPU. Архитектура и программная модель CUDA. – М.: МГУ, 2012. – 336 с.
20. Гергель В.П. Современные языки и технологии параллельного программирования. – М.: МГУ, 2012. – 408 с.
21. Антонов А.С. Технологии параллельного программирования MPI и OpenMP. – М.: МГУ, 2012. – 344 с.
22. Лыкосов В.Н., Глазунов А.В., Кулямин Д.В., Мортиков Е.В., Степаненко В.М. Суперкомпьютерное моделирование в физике климатической системы. – М.: МГУ, 2012. – 408 с.
23. Якобовский М.В. Введение в параллельные методы решения задач. – М.: МГУ, 2013. – 328 с.
24. Стронгин Р.Г., Гергель В.П., Гришагин В.А., Баркалов К.А. Параллельные вычисления в задачах глобальной оптимизации – М.: МГУ, 2013. – 280 с.
25. Рутм Г., Фатика М. CUDA Fortran для ученых и инженеров. – М.: ДМК Пресс, 2014. – 364 с.

Екатерина Валерьевна **Волкова**
Александр Сергеевич **Пузанов**
Сергей Владимирович **Оболенский**
Елена Александровна **Тарасова**

ВВЕДЕНИЕ В ФИЗИКУ ПОЛУПРОВОДНИКОВЫХ ДИОДОВ И МЕТОДЫ ИХ ПРОЕКТИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ

Учебное пособие

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Нижегородский государственный университет им. Н.И. Лобачевского».
603950, Нижний Новгород, пр. Гагарина, 23.