Electronic Structure Calculations of Point Defects in Semiconductors

ANDREAS HÖGLUND
Dissertation presented at Uppsala University to be publicly examined in Siegbahnsalen, Ångström Laboratory, Lägerhyddsvägen 1, Uppsala, Friday, June 8, 2007 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

In this thesis point defects in semiconductors are studied by electronic structure calculations. Results are presented for the stability and equilibrium concentrations of native defects in GaP, InP, InAs, and InSb, for the entire range of doping conditions and stoichiometry. The native defects are also studied on the (110) surfaces of InP, InAs, and InSb. Comparing the relative stability at the surface and in the bulk, it is concluded that the defects have a tendency to migrate to the surface.

It is found that the cation vacancy is not stable, but decomposes into an anion antisite-anion vacancy complex. The surface charge accumulation in InAs is explained by complementary intrinsic doping by native defects and extrinsic doping by residual hydrogen.

A technical investigation of the supercell treatment of defects is performed, testing existing correction schemes and suggesting a more reliable alternative. It is shown that the defect level of $2V_{c,-III,\text{con}}$ in the solarcell-material CuIn$_x$Ga$_{1-x}$Se$_2$ leads to a smaller band gap of the ordered defect $\gamma$-phase, which possibly explains why the maximal efficiency for CuIn$_y$Ga$_{1-y}$Se$_2$ has been found for $x=0.3$ and not for $x=0.6$, as expected from the band gap of the $\alpha$-phase.

It is found that Zn diffuses via the kick-out mechanism in InP and GaP with activation energies of 1.60 eV and 2.49 eV, respectively. Explanations are found for the tendency of Zn to accumulate at pn-junctions in InP and to why a relatively low fraction of Zn is found on substitutional sites in InP.

Finally, it is shown that the equilibrium solubility of dopants in semiconductors can be increased significantly by strategic alloying. This is shown to be due to the local stress in the material, and the solubility in an alloy can in fact be much higher than in either of the constituting elements. The equilibrium solubility of Zn in Ga$_y$In$_{1-y}$P is for example five orders of magnitude larger than in GaP or InP.

Keywords: electronic structure calculations, point defects, semiconductor, formation energy, equilibrium solubility limit, thermodynamic equilibrium concentration, transfer levels, negative-U, (110) surface, diffusion, activation energy, solar cells

Andreas Höglund, Department of Physics, Condensed Matter Theory, Box 530, Uppsala University, SE-75121 Uppsala, Sweden

© Andreas Höglund 2007

ISSN 1651-6214
ISBN 978-91-554-6916-0
urn:nbn:se:uu:diva-7926 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7926)
Till mamma och pappa
och callecaleman
och carlasortiz
och tysken
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I  Relative concentration and structure of native defects in GaP
   A. Höglund, C. W. M. Castleton, and S. Mirbt,

II  Managing the supercell approximation for charged defects in semiconductors: Finite-size scaling, charge correction factors, the band-gap problem, and the ab initio dielectric constant
   C. W. M. Castleton, A. Höglund, and S. Mirbt,

III Point defects on the (110) surfaces of InP, InAs, and InSb: A comparison with bulk
   A. Höglund, C. W. M. Castleton, M. Göthelid, B. Johansson, and
   S. Mirbt,

IV Increasing the equilibrium solubility of dopants in semiconductor multilayers and alloys
   A. Höglund, O. Eriksson, C. W. M. Castleton, and S. Mirbt,

V  Diffusion mechanism of Zn in InP and GaP
   A. Höglund, C. W. M. Castleton, and S. Mirbt,
   Submitted to Appl. Phys. Lett.

VI Why does charge accumulate on the surfaces of InAs but not on other III-V semiconductors?
   C. W. M. Castleton, A. Höglund, M. Göthelid, M. Qian, and S.
   Mirbt,

VII The nature of cation vacancies on III-V semiconductor surfaces
   A. Höglund, C. W. M. Castleton, M. Göthelid, and S. Mirbt,
   Submitted to Phys. Rev. B
Ordered defect phases in CuIn$_{1-x}$Ga$_x$Se$_2$ form localized electron paths
A. Höglund, S. Mirbt, and C. Persson,
_In Manuscript._

Equilibrium solubility of dopants and substitutional defects in binary and ternary semiconductor structures: Ga$_x$In$_{1-x}$P and Si$_x$Ge$_{1-x}$
A. Höglund, O. Eriksson, C. W. M. Castleton, and S. Mirbt,
_In Manuscript._

Calculated STM-images of native defects on the (110) surfaces of InP, InAs and InSb
A. Höglund, C. W. M. Castleton, and S. Mirbt,
_In Manuscript._

Calculations of the lattice parameters in metallic and semiconducting multilayer systems: a reference data base
A. Höglund, M. Råsander, P. Souvatzis, and O. Eriksson,
_In Manuscript._

Reprints were made with kind permission from the publishers.
## Contents

1 Introduction in Swedish ........................................... 1
2 Introduction ......................................................... 5
3 Computational methods .......................................... 9
   3.1 Density functional theory ................................. 10
      3.1.1 The Kohn-Sham equations ......................... 11
      3.1.2 The local density approximation ................. 12
   3.2 Computations on solids ................................. 13
      3.2.1 The self-consistent iterative process .......... 14
   3.3 Pseudopotentials ..................................... 15
      3.3.1 The projector augmented wave method ........ 18
      3.3.2 Ultrasoft pseudopotentials .................... 20
   3.4 Forces and ionic relaxation .................................. 21
   3.5 Chemical bonding in solids ............................... 22
      3.5.1 Charge density transfer .......................... 22
      3.5.2 The electron localization function ............. 23
   3.6 The deficiencies of DFT ................................. 24
4 Defects and impurities in semiconductors .................. 27
   4.1 The electrical properties of impurities .......... 28
   4.2 The structure of point defects ....................... 30
   4.3 Theoretical treatment of defects .................... 31
      4.3.1 The defect formation energy ...................... 32
      4.3.2 Normalization of the formation energy ....... 32
      4.3.3 The chemical potential .......................... 33
      4.3.4 Defect density .................................. 33
      4.3.5 Definition of the Charge Transfer Level ...... 35
      4.3.6 The supercell approximation ................... 36
   4.4 Identification of defects ............................... 36
5 Native Defects .................................................. 39
   5.1 GaP ......................................................... 39
   5.2 InP, InAs and InSb ...................................... 45
6 Supercell size effects ......................................... 49
   6.1 Finite size scaling ..................................... 49
   6.2 Native defects of GaP in the infinite limit ....... 52
   6.3 Defect level positions in the infinite limit ...... 52
7 Defects on III-V Semiconductor surfaces .................. 55
   7.1 Simulated STM images from calculations ............ 58
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>Photoelectric threshold</td>
<td>59</td>
</tr>
<tr>
<td>7.3</td>
<td>Native defects on InP, InAs, and InSb (110) surfaces</td>
<td>59</td>
</tr>
<tr>
<td>7.4</td>
<td>Relative stability of defects at the surface and in the bulk</td>
<td>62</td>
</tr>
<tr>
<td>7.5</td>
<td>Reconfiguration of $V_{In}$</td>
<td>65</td>
</tr>
<tr>
<td>7.6</td>
<td>Charge accumulation on InAs surfaces</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>CuIn$_{1-x}$Ga$_x$Se$_2$ semiconductor solar cells</td>
<td>69</td>
</tr>
<tr>
<td>9</td>
<td>Diffusion</td>
<td>73</td>
</tr>
<tr>
<td>9.1</td>
<td>Diffusion of Zn in InP and GaP</td>
<td>74</td>
</tr>
<tr>
<td>10</td>
<td>Dopants in alloys and multilayers</td>
<td>79</td>
</tr>
<tr>
<td>11</td>
<td>The future generation of semiconductor materials</td>
<td>83</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>89</td>
</tr>
</tbody>
</table>
1. Svensk introduktion


Material med en periodisk upprepning av atomstrukturen kallas för kristallina material. De kristallina materialen kan delas in i tre kategorier: metaller, halvledare och isolatorer. Metaller leder elektrisk ström och material i denna kategori är exempelvis järn (Fe) och zinc (Zn). Halvledare leder elektrisk ström om de dopas, vilket kommer att beskrivas nedan, och typiska material är kisel (Si) och indiumfosfid (InP). Isolatorer leder, som namnet antyder, inte ström och är typiskt hård och spröd material så som salter (ex. NaCl) och oxider, som till exempel kvarts (SiO$_2$).

Den här avhandlingen kommer att handla om halvledare. Det är halvledar teknologin som ligger till grund för all modern elektronik; allting som innehåller ett chip bygger på halvledare. (Bardeen, Brattain och Shockley fick Nobelspriset 1956 för skapandet av den första transistorn.) Halvledarmaterialen utgörs av de elementära halvledarna som återfinns i grupp IV i periodiska systemet (exempelvis Si), III-V halvledare (exempelvis InP) och II-VI halvledare (exempelvis ZnO). I grova drag så har nästan alla halvledarmaterial samma kristallstruktur som diamant. Kisel är den halvledare som dominerar halvledarindustrin. De andra typerna, III-V och II-VI halvledararna, kan dock
erbjuder en mängd förbättringar, särskilt för optiska tillämpningar (se Kapitel 11).

Ibland förekommer det avvikelser från den perfekta kristallstrukturen, så kallade punkt-defekter. Dessa kan vara saknade atomer (vakanser), atomer på fel gitterposition (antisites/substitutionella defekter) eller atomer i tomrummet mellan de ordinarie atomerna (interstittella defekter). I periodiska kristaller kan en defekt ha mycket stort inflytande på materialegenskaperna. Defekter kan ha stor påverkan på t.ex. hållfasthet, optiska egenskaper, elektrisk ledningsförmåga, osv. Det är till exempel vakanser som ger färgade transparenta isolatorer dess färg.

Även främmande atomslag eller orenheter kan utgöra defekter, och ofta skråddarsyr man ett materials egenskaper genom att medvetet inplantera dessa. Hela halvledarindustrin och all elektronik bygger på just detta, där man dopar halvledare genom att tillsätta främmande atomer så att det bildas höga (eller låga) kontrollerade koncentrationer av laddningsbärare. (T.ex. indiumfosfid dopat med zinc beskrivs med notationen InP:Zn). Laddningsbärarna kan vara antingen de negativa elektronerna eller positiva elektronhål (en avsaknad elektron kan bete sig som en partikel, likt en bubbla som effektivt sett har positiv laddning). Om dopning resulterar i höga koncentrationer av elektro- ner som laddningsbärare kallas halvledaren för n-dopad, och om den är dopad med elektronhål kallas den p-dopad. Rena halvledare är alltså av litet praktiskt intresse och det krävs att man dopar dessa för att få det önskade elektriska beteendet. Det har även visat sig att man med dopning även kan påverka den joniska ledningsförmågan, vilket är av intresse för tillämpningar inom bränslecellsteknologi, och materialegenskaper som är avgörande för vätebagring i material, vilket är av stor vikt inom energisektorn.

I denna avhandling har jag studerat defekter i halvledare, både de som förekommer naturligt, som ofta är icke-önskvärda, och de som medvetet tillförs materialet. Detta har gjorts med numeriska kvantmekaniska elektronstrukturberäkningar, eller mer exakt med täthetsfunktionalberäkningar, DFT (för vilken Walter Kohn tilldelades Nobelpriset 1998). Den teoretiska bakgrunden och de metoder som använts vid defektberäkningarna kommer att beskrivas i Kapitel 3 och 4. Det arbete som jag har utfört och som presenteras i denna avhandling är följande:

De naturliga defekterna i GaP, InP, InAs och InSb har studerats och dessa resultat ges i Kapitel 5. Genom att finna den mest stabila kofigurationen för alla defekter har vi avgjort vilka defekter som är mest förekommande och i vilka koncentrationer. Detta har gjorts som en funktion av alla möjliga dopningsförhållanden (från starkt p-dopande material till n-dopade) och därigenom har vi även kunnat avgöra vilka defekter som ställer till problem i form av t.ex. kompensation av dopning och icke-ljusemitterande rekombinationer i optoelektroniska komponenter. Även defektternas stabilitet med kemisk obalans har studerats (stökiometri), och från detta kan man se hur man praktiskt kan bli av med, eller i alla fall minimera, ovanstående problem. Bortsett från
dessa tillämpade sidor har arbetet även lett till ökad förståelse av de naturliga defekternas grundläggande struktur och egenskaper.

En beräknings-teknisk undersökning av begränsningarna med att behandla icke-periodiska defekter med större periodiska enheter ges i Kapitel 6. De befintliga korrektionerna har testats och en ny metod har lagts fram. Vi visar att denna ger bättre och noggrannare resultat för exempelvis de tidigare felaktigt beskrivna grunda dopnivåerna.


I princip alla solceller som används är halvledarbaserade. Idag är det främst kisel som används men mycket forskning pågår med fokus på att hitta nya och bättre solcellsmaterial. Detta för att kunna göra solenergin mer kostnadseffektivt och ett mer konkurrenskraftigt alternativ till andra kraftkällor. CuIn$_{1-x}$Ga$_x$Se$_2$ eller CIGS är ett sådant nytt lovande material, där $x=0.3$ har visat sig ge högst effektivitet. CIGS är dock ett väldigt komplext material med defektkoncentrationer så stora att de avservärt kan ändra den kemiska kompositionen. Kapitel 8 sammanfattar vår studie av hur ett defektkomplex, $[2\text{VCu}^{II}\text{ICu}]$, kan komma att förändra bandgapet hos CIGS. Denna förändring är en möjlig förklaring till varför man funnit den högsta effektiviteten för CIGS för $x=0.3$ och inte för $x=0.6$ som förväntas från bandgapsökningen.

Dopning ger halvledare deras önskade elektriska egenskaper men om dopämnena börjar vandra i materialet (diffundera) så leder det till försämrad prestanda. För att en komponent ska vara tillförlitlig och ha lång livslängd så vill man alltså att dopämnena ska förbli på sina substitutionella positioner. Det är därför centrale att förstå hur och vid vilken energi som dopatomerna blir mobila (denna energi kallas för aktiveringsenergin). I Kapitel 9 ges resultaten av våra beräkningar för Zn-diffusion i InP och GaP. Detta ger mekanismen via vilken Zn diffunderar (kick-out mekanismen: Zn$_{III}$+$\text{III}^{I}$$\rightarrow$Zn$_{I}^{II}$) och även aktiveringsenergierna då detta börjar att ske. Vi ger förklaringar till varför en relativt låg andel Zn är aktivt som dopämne i InP, till skillnad från andra III-V halvledare, samt varför Zn har en tendens attackumuleras i gränsskiktet på $pn$-dioder.

Ett ständigt problem när man både försöker öka prestandan och minska storleken på elektroniska komponenter är att åstadkomma allt högre doping-koncentrationer. Bortsett från kompensation av naturliga defekter så finns det
en mättnadskoncentration för ett dopämne i en given halvledare. I Kapitel 10 visar vi hur denna övre gräns kan förhöjas genom att legera med en annan halvledare. Överraskande kan mättnadskoncentrationen i denna legering, vid ett strategiskt val av material, bli avsevärt mycket högre än i något av materia- len som ingår i den. Den fysikaliska förklaringen till hur detta kan vara möjligt finner vi vara relaterad till spänningar i materialet och den ges i kapitel 10.

Slutligen ges en framtidsutsikt i Kapitel 11 där det diskuteras vilka som kan tänkas vara nästa generations halvledarmaterial.

För att kunna förklara halvledares beteende så måste man alltså studera dem på en mikroskopisk nivå med hjälp av kvantmekaniska datorberäkningar. Ironiskt nog kräver alltså förståelsen av halvledare användandet av dem...
2. Introduction

All materials in our surrounding are built up by atoms. Each atom has a certain number of electrons and the behavior of these electrons is described by the area of physics called quantum mechanics. The atomic structure and the inter-atomic bonds, that all these electrons give rise to, in turn determine the properties of a material, such as the hardness, the electric conductivity, the elasticity, the density, the magnetic properties, and even the color and transparency. In this way the quantum mechanical laws, that act on a microscopic level and that can seem difficult and non-intuitive, determine how everything appears and behaves on a macroscopic level. The properties of a material can therefore be calculated with the aid of these laws. All that is needed is to know is what kinds of atoms the material consists of. Remarkably, these kinds of calculations, whose only input is the type and number of the atoms and a few constants of nature (Planck’s constant, the elementary charge, the electron mass, etc.), give excellent agreement with experimental data. Quantum mechanical calculations can, however, quickly become an overwhelming task for systems larger than a few atoms. To study a macroscopic material, containing \( \sim 10^{23} \) atoms, requires that the periodicity of the material is taken advantage of and that the calculations are performed on a computer.

Materials with a periodic repetition of their atomic structure are called crystalline materials. The crystalline materials can be divided into three categories: metals, semiconductors and insulators. Metals conduct electric currents and materials in this category are, for example, iron (Fe) and zinc (Zn). Semiconductors conduct electric currents if they are doped, which will be described below, and typical materials are silicon (Si) and indium phosphide (InP). Insulators, as the name suggests, do not conduct electric currents and are normally hard and brittle materials such as salts (as NaCl) or oxides, for example quartz (SiO\(_2\)).

This thesis will be about semiconductors. Semiconductor technology forms the foundation of all modern electronics; everything that contains a "chip" relies on semiconductors. (Bardeen, Brattain and Shockley received the Nobel prize 1956 for the creating of the first transistor.) Semiconductor materials are grouped into the elemental semiconductors that are found in group IV of the periodic table (for example Si), the III-V semiconductors (for example InP), and the II-VI semiconductors (for example ZnO). Roughly speaking, almost all semiconductors have the same crystal structure as diamond. Silicon is the semiconductor that dominates the semiconductor industry. The other
two groups, the III-V and II-VI semiconductors, can however offer numerous improvements, especially for optical applications (see Chapter 11).

There are sometimes deviations from the perfect crystal structure, so called point-defects. These can be missing atoms (vacancies), atoms at the wrong lattice site (antisite/substitutional defects) or atoms in between the ordinary lattice sites (interstitial defects). In periodic crystals a defect atom can have a very large influence on the material properties. Defects can have a large impact on for example the strength of the material, its optical properties, the electric conductivity, and so on. It is, interestingly, vacancies that give most colored transparent ionic crystals their color.

Defects can also be foreign atoms or impurities, and material properties are often engineered by deliberately inserting these. The entire semiconductor industry relies on this, where semiconductors are doped by inserting other types of atoms so that controlled high (or low) concentrations of charge carriers are formed. (For example, indium phosphide doped with zinc is described with the notation InP:Zn). The charge carriers can either be negative electrons or positive electron holes (the lack of an electron can behave like a particle, like a bubble that effectively has positive charge). If the doping results in high concentrations of electrons as charge carriers, the semiconductor is called n-doped, and if they are doped with holes they are called p-doped. Pure semiconductors are therefore of little practical interest and they must be doped to get the desired electrical behavior. It turns out that through doping one can also affect a materials ionic conductivity, which is of great interest for applications in fuel cell technology, and also the properties central for hydrogen storage, which is of large significance to the energy sector.

In this thesis I have studied defects in semiconductors, both the native defects, that are usually undesirable, and the ones that have been deliberately added to the material. This has been done with the use of numerical quantum mechanical calculations, more specifically by density functional theory, or DFT (for which Walter Kohn was awarded the Nobel prize in 1998). The theoretical background and the methods used for defect calculations will be described in Chapters 3 and 4. The work I will present in this thesis is the following:

The native defects in GaP, InP, InAs, and InSb have been studied and these results are given in Chapter 5. By finding the most stable configuration for all defects we have determined which defects are the most common and in what concentrations. This has been done as a function of all possible doping conditions (from strongly p-type material to strongly n-type) and thereby we have been able to see which defects cause problems such as, for example, self-compensation and non-radiative recombination in optoelectronic components. The defect’s stability with chemical imbalance (stoichiometry) have also been studied, and from this it can be seen how in practice one could get rid off, or at least minimize, the problems above. Apart from these applied aspects,
this work has led to a deeper understanding of the native defects fundamental nature and properties.

A computational technical investigation of the limitations of treating non-periodic defects within larger periodic units is given in Chapter 6. The available corrections for this problem have been tested and a new scheme has been suggested. We show that this new scheme gives more accurate results for, for example, the earlier incorrectly described shallow doping levels.

In Chapter 7 results are presented for the native defects on the surfaces of InP, InAs and InSb. In the same way as for the bulk materials above, the most common defects have been calculated under all possible doping and stoichiometric conditions. We have calculated the relative stability between surface and bulk positions, and from that it can be seen that the defects have a general tendency to be more stable at the surface. We show that the cation vacancy is not stable on these III-V surfaces, and instead an anion antisite, anion vacancy complex is formed spontaneously. Further, an explanation is given as to why InAs, as opposed to other III-V semiconductors, accumulates charge on the surface, which is based on doping by native defects but also residual-hydrogen that binds to the surface.

In principle all solar cells in use are semiconductor based. Today the most common material in use is silicon, but there is a lot of ongoing research with the focus on finding new and better solar cell materials. This is in order to make solar energy more cost efficient and a more competitive alternative to the available power sources. CuIn$_{1-x}$Ga$_x$Se$_2$ or CIGS is one of these new and promising materials, where $x=0.3$ has turned out to give the maximal efficiency. CIGS is, however, a very complex material with defect concentrations large enough to considerably change the chemical composition. Chapter 8 summarizes how a defect complex, [2VCu-IIICu], can come to change the band gap of CIGS. This change is a possible explanation for the fact that the maximal efficiency of CIGS has been found for $x=0.3$ and not for $x=0.6$ as expected from the band gap increase.

Doping gives semiconductors their desired electrical properties, but if the dopants start to move in the material (diffuse) it will lead to a decrease in performance. In order for an electrical component to be reliable and have a long life span, the dopants should ideally remain at their substitutional positions. It is therefore of central importance to understand how and at what energy the dopant atoms become mobile (this energy is called the activation energy). In Chapter 9, our results are given for calculations of Zn diffusion in InP and GaP. This gives the mechanism by which Zn diffuses (the kick-out mechanism: Zn$_{III}$+III$_{III} \rightarrow$Zn$_{I}$) and also the activation energy at which this starts to happen. We give explanations as to why a relatively low fraction of Zn is active as dopant in InP, as compared to other III-V semiconductors, and to why Zn tends to accumulate in $pn$-junctions of InP (the same accumulation is predicted for GaP).
A reoccurring problem when one attempts to both increase the performance and decrease the size of electric components is the achievement of higher doping concentrations. Apart from compensation by native defects, there is a saturation limit for a dopant in a given semiconductor. In Chapter 10 we show how this upper limit can be increased by alloying. Surprisingly, the saturation limit in this alloy can be higher than in either of the constituent elements. The physical explanation to this is found to be related to internal stress and is given in chapter 10.

Finally, an outlook is presented in Chapter 11, where it is discussed what the next generation of semiconductor materials might be.

In order to explain the behavior of semiconductors, one must, as outlined here, study them at a microscopic level with the use of quantum mechanical computer calculations. Ironically the understanding of semiconductors requires the use of them...
3. Computational methods

All materials are composed of atomic nuclei and electrons. The macroscopic material properties that we observe are only dependent on the configuration and structure of these electrons and ions. Knowing only what the type of atoms the material consists of is, in principle, enough to be able to calculate the total wavefunction and energy of the system using the (time independent) Schrödinger equation.

\[
H \Psi_{\text{Tot}}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{R}_1, \mathbf{R}_2) = E \Psi_{\text{Tot}}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{R}_1, \mathbf{R}_2),
\]

(3.1)

where \( \Psi \) is the total wavefunction, \( \mathbf{r}_i \) and \( \mathbf{R}_i \) are the positions of the electrons and ions respectively, and \( H \) is the Hamiltonian for the system.

\[
H = -\sum_i \frac{\hbar^2}{2m_e} \nabla^2 \mathbf{r}_i - \sum_{i,j} \frac{eZ_j}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{R}_j|} + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{r}_j|}
\]

\[
+ \frac{1}{2} \sum_{i \neq j} \frac{Z_i Z_j}{4\pi\varepsilon_0 |\mathbf{R}_i - \mathbf{R}_j|} - \sum_i \frac{\hbar^2}{2M_i} \nabla^2 \mathbf{R}_i,
\]

(3.2)

\( M_i \) and \( Z_i \) are the nucleus mass and charge respectively and \( m_e \) is the electron mass. One can then minimize the total energy with respect to the ionic structure in order to find the equilibrium crystal structure. These calculations are known as \textit{ab initio} calculations since they are based only on the laws of physics and the values of nature constants, such as Planck’s constant \( \hbar \) and the electron charge \( e \). Nothing else is assumed or used as empirical input.

As mentioned in the introduction, it is only possible to solve the Schrödinger equation exactly for small simple systems (the hydrogen atom with spherical symmetry, for example) so for larger systems consisting of \( \sim 10^{23} \) electrons and ions, additional approximations and theorems have to be used.

The first step is the \textit{Born-Oppenheimer approximation} which separates the motion of the ions and the electrons. This is justified by the fact that the mass of a nucleus is much larger than the mass of an electron so that they move on different time scales. From the electron point of view, the ions are stationary and the electron cloud will rearrange itself instantly to any new ionic configuration. Mathematically, the total wavefunction is rewritten as a product of the electron wavefunction and the nuclear wavefunction

\[
\Psi_{\text{Tot}}(\mathbf{r}_1, \ldots, \mathbf{R}_1, \ldots) = \Psi_e(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots) \cdot \chi_{\text{nuc}}(\mathbf{R}_1, \ldots).
\]

(3.3)
Thesemicolon in the electron wavefunction indicates that it is dependent upon the ionic positions as parameters but not as variables. After using the ansatz above the approximation itself lies in setting a small term to zero (basically the kinetic energy of the ions, which is valid at $T=0$ K) so that the total Schrödinger equation can be separated into two equations. The first is the Schrödinger equation for the nuclei and the second the corresponding for the electrons for a given frozen ionic configuration. It is the latter that will be of interest from now on,

$$H'\Psi_e(r_1,\ldots;R_1,\ldots) = E_e\Psi_e(r_1,\ldots;R_1,\ldots),$$

(3.4)

where $H'$ consists of the three first terms of $H$ in Eq. (3.2).

The last, and probably hardest, obstacle to be overcome is the reduction of the many electron equation above to a solvable problem. This is most easily done by the approximation of rewriting Eq. (3.4) as a one-particle equation for an electron moving in an average potential from all the electrons. Early attempts to solve this were made by Hartree under the assumption that the total wavefunction then simply is a product of one electron orbitals, $\Psi(r_1, r_2, \ldots, r_n) = \psi_1(r_1) \psi_2(r_2) \cdots \psi_n(r_n)$ [1]. This approach does not take the Pauli exclusion principle into account since the total wavefunction has to be antisymmetric. An improvement was later made by expressing the total wavefunction as a determinant of one-electron orbitals. The fact that electrons are fermions will now be incorporated automatically since the total wavefunction will both be antisymmetric and become zero if two electrons of the same spin are located at the same position. In this way exchange effects are taken into account. This approach is called the Hartree-Fock method. The Hartree-Fock method is not exact and the difference between the Hartree-Fock energy and the true ground state energy is called the correlation energy. The correlation energy is due to the fact that the Hartree-Fock formalism uses average potentials and that in certain cases the determinant is not a good approximation to the true all electron wavefunction.

3.1 Density functional theory

In density functional theory an attempt is made to make the electron density the central quantity. In a system of $n$ electrons the electron density is defined from the wavefunctions as follows,

$$n(r) = 2^{n/2} \sum_{i=1}^{n/2} |\psi_i(r)|^2.$$  

(3.5)

---

1This problem can be overcome by using a linear combination of determinants in what is called the configuration interaction approach, but it is very computationally heavy [2].

2For a non spin-polarized system.
The energy is rewritten from being the expectation value of $H'$ to a functional only depending on the electron density and not explicitly the wavefunctions. The foundation of DFT rests on two theorems formulated by Hohenberg and Kohn [3].

**Theorem 1 (Uniqueness)** The ground state expectation value of any observable is a unique functional of the exact ground state electron density $n(r)$.

All ground state properties can thus be extracted from the exact electron density. The problem is now only how to find this density. The second theorem is helpful in this matter:

**Theorem 2 (Variational principle)** For a given external potential, the true density $n(r)$ minimizes the total energy functional.

This reduces the very complex problem of finding all physical properties of a system to finding the minimum of the energy with respect to the electron density (which still is not trivial since $n(r)$ is function in three dimensional space). The energy functional is given by the following expression [4]:

$$
E[n(r)] = E_{ke}[n(r)] + E_{nuclear}[n(r)] + E_H[n(r)] + E_{XC}[n(r)] + E_{ion}(R_0)
$$

$$
= 2 \sum_i \int \psi_i \left( -\frac{\nabla^2}{2} \right) \psi_i \, dr + \int V_{ion}(r) \, n(r) \, dr
$$

$$
+ \frac{1}{2} \int \int \frac{n(r) \, n(r')}{|r - r'|} \, dr \, dr' + E_{XC}[n(r)] + E_{ion}(R_0),
$$

where the two first terms are the electronic kinetic energy and the interaction between the nuclei and the electrons. Terms three and five are the coulomb interaction energy between electrons (the Hartree term) and between ions at positions $R_0$. All these terms assume an uncorrelated pair density. That is, the correlation factor, $f(r_1; r_2)$, is equal to zero for the following expression

$$
n_2(r_1, r_2) = n(r_1) n(r_2) \left( 1 + f(r_1; r_2) \right).
$$

The correlation factor is not known exactly and all $f(r_1; r_2)$ nonzero terms are collected in the exchange-correlation functional, $E_{XC}[n(r)]$. This term will have to be approximated as will be described below.

### 3.1.1 The Kohn-Sham equations

The Hohenberg-Kohn theorems show that the electron density can rigorously be made the fundamental quantity of the many-body problem, but they are pure theorems of existence and say nothing of how this exact charge density can be found.
Kohn and Sham showed that there is a way to map the problem of solving Eq. (3.6) to the one of solving a system of noninteracting electrons moving in an effective potential from all the (other) electrons [5]. According to Theorem 2, the true electron density will minimize the total energy, but all all means if finding it are valid. It could be guessed or, as suggested by Kohn and Sham, calculated from a reference system of non-interacting electrons moving in an effective potential. The derivation of their Schrödinger-like one particle equation is founded on the fact that the differential of the energy must be zero since it is a minimum according to theorem 2, $\delta E = 0$. Thus, developing this variation with the full energy functional added (Eq. (3.6)), under the condition that the sum of the density throughout the molecule or solid should be constant and equal to the number of electrons,

$$\int_V n(r) \, dr = n_e$$

finally gives the Schrödinger-like equations called the Kohn-Sham equations:

$$\mathcal{H} \psi_i = \left[ -\frac{1}{2} \nabla^2 + V_{\text{eff}}(r) \right] \psi_i = \varepsilon_i \psi_i$$

$\mathcal{H}$ is the one electron Hamiltonian and $V_{\text{eff}}(r)$ the effective potential in which the electron moves.\(^3\) The effective potential is just a combination of potential terms that originate from the energy functional:

$$V_{\text{eff}}(r) = V_{\text{ion}} + V_H + V_{XC} = V_{\text{ion}}(r) + \int \frac{n(r')}{|r - r'|} \, dr' + \frac{\delta E_{XC}[n(r)]}{\delta n(r)}.$$  \hspace{1cm} (3.10)

Since the electron density is needed to calculate the last two terms, which are the average Coulomb potential from all electrons and the exchange-correlation potential, the Kohn-Sham equations need to be solved self-consistently. The initial electron density is composed from, for example, the atomic densities.

The Kohn-Sham equations can now be solved instead of finding the minimum of Eq. (3.6), and the orbitals $\psi_i$ then give the electron density according to Eq. (3.5) above. These orbitals are often called Kohn-Sham orbitals and in the case of a non spin-polarized system, each of these orbitals contains two electrons. (If spin-related effects, such as collinear magnetism, are under consideration the total density is instead expressed as the sum of the spin-up and spin-down densities $n(r) = n_\uparrow(r) + n_\downarrow(r)$ [6].)

### 3.1.2 The local density approximation

There is no analytically known expression for the exchange-correlation potential, $V_{XC} = \frac{\delta E_{XC}[n(r)]}{\delta n(r)}$, and hence it needs to be approximated. Many methods have been developed for this purpose but the most commonly used is the local density approximation (LDA). It is surprisingly accurate in spite of its

\(^3\)For a detailed derivation see Dreizler and Gross [3].
simplicity; the exchange-correlation energy, $E_{XC}[n(r_0)]$, is approximated to be equal to the exactly known exchange-correlation energy in a homogeneous electron gas of the same electron density $n(r_0)$. In this way the LDA exchange-correlation energy is a function of $n$ rather than $n(r)$,

$$E_{XDA}^{LDA}(n) = \int n(r) \varepsilon_{XC}[n(r)] dr,$$

(3.11)

where $\varepsilon_{XC}$ is the exchange-correlation energy per particle in a homogeneous electron gas. ($\varepsilon_{XC}$ is calculated from, for example, Quantum Monte Carlo methods and then parametrized to a continuous function of the electron density) [7, 8].

LDA is based on the assumption that the electron density is locally almost constant. For electron densities rapidly changing in space this is not a good approximation and attempts have been made to include the dependence of the density gradient, $\nabla n(r)$ (the general gradient approximation, GGA). There also exists a generalization of LDA in which the spin is explicitly taken into consideration, called the local spin density approximation, LSDA.

### 3.2 Computations on solids

The theory outlined so far works for molecules with finitely many electrons. To calculate the bulk properties of a solid, one faces a problem consisting of infinitely many electrons. To be able to solve it, the periodicity of the solids crystal structure must be taken advantage of. The effective potential, $V_{\text{eff}}$, in the Kohn-Sham equation will now be periodic and the one-particle wavefunctions can with the use of Bloch’s theorem [9] be rewritten as

$$\psi_k(r) = u_k(r) \cdot e^{i k \cdot r}.$$

(3.12)

Here $u_k$ is a function with the periodicity of the lattice and the exponential term is a phase term where the $k$ vectors are quantized by the outer periodicity, the number of unit cells in one period. In the case of infinitely repeating unit cells, $k$-space is continuous and all $k$ vectors are allowed in Eq. (3.12).

The periodic function $u_k(r)$ can be expanded in plane waves and hence the entire wavefunction can be expressed in a basis of plane waves.

$$u_k(r) = \sum_G c_{kG} e^{iG \cdot r} \Rightarrow \psi_k(r) = \sum_G c_{kG} e^{i(k+G) \cdot r},$$

(3.13)

where $G$ are the reciprocal lattice vectors. (The summation is only over the reciprocal lattice vectors since $u_k(r + R) = u_k(r)$ but $e^{ik \cdot r} = 1$ only for $k = G$). Each plane wave is an eigenfunction of the kinetic energy operator and is therefore a solution for the empty lattice. Because of this, relatively few plane waves are needed in the expansion of a wavefunction in a weak potential, and hence the plane wave basis is an ideal choice for weak potentials. In principle,
an infinite plane wave basis set is needed in order to be able to expand the wavefunctions exactly but the basis set can be limited since the terms, \( c_{k,G} \), with sufficiently high kinetic energy \( \frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{G}|^2 \) can be omitted because their contribution to the total energy is negligible [10]. This is done through truncating the sum in Eq. (3.13) after some sufficiently large kinetic energy. This energy is called the cutoff energy.

Using Bloch’s theorem the problem of infinitely many electrons has been turned into a problem of infinitely many \( k \)-points. Two things can now be taken advantage of to reduce the computational effort. First, all \( k \)-points in the expansion of the wavefunction, Eq. (3.13), outside the first Brillouin zone can be written as \( \mathbf{k}' = \mathbf{k} + \mathbf{G} \) and since \( e^{i\mathbf{R} \cdot \mathbf{G}} = 1 \) we only have to consider \( k \)-points inside the first Brillouin zone. Further, rotation and inversion symmetries can be used to reduce the subset in \( k \)-space to what is called the irreducible part of the first Brillouin zone. Second, in this subset of \( k \)-space there is still a continuous and therefore infinite number of \( k \)-vectors, but the wavefunctions of closely located \( k \)-points are almost identical and therefore a small region can be sampled by one single \( k \)-point. The electronic part of the total energy can thus be calculated, to a good approximation, using only a discrete number of \( k \)-points. Special sets of \( k \)-points have been derived for doing so with a minimum number of \( k \)-points [11, 12].

The cutoff energy and the \( k \)-point mesh in the irreducible Brillouin zone are parameters of convergence and need to be adjusted to the type of calculation. Convergence errors can always be reduced by using a denser set of \( k \)-points or a higher energy cutoff.

### 3.2.1 The self-consistent iterative process

As mentioned, the electron density is calculated self-consistently: first \( V_{\text{ion}} \) is calculated from the ion positions and types in the crystal, then an initial trial electron density is used to calculate \( V_H \) and \( V_{XC} \). Using the plane-wave basis set for the wavefunctions, the Kohn-Sham equations can be rewritten as

\[
\sum_{G'} \left( \frac{\hbar^2}{2m_e} |\mathbf{k} + \mathbf{G}|^2 \delta_{GG'} + V_{\text{ion}}(\mathbf{G} - \mathbf{G}') + V_H(\mathbf{G} - \mathbf{G}') + V_{XC}(\mathbf{G} - \mathbf{G}') \right) c_{i,k+G'} = \varepsilon_i c_{i,k+G'},
\]

which is a matrix eigenvalue equation were the potentials are in the form of Fourier transforms. Diagonalizing this system and solving it for the vector \( c_{k,G} \) allows us to calculate \( \psi_i \) for the corresponding \( k \)-vector and eigenvalue. After this has been solved for all \( k \)-vectors a new electron density is calculated according to Eq. (3.5). \( V_H \) and \( V_{XC} \) are then constructed again and a new electron density calculated until self-consistency is reached (Fig. 3.1). For numerical reasons the fastest convergence is obtained if a mixing of the old and
new densities is used for calculating the effective potential $V_{\text{eff}}$ instead of just the new density.

### 3.3 Pseudopotentials

When isolated atoms are put together to form a crystalline material, the highest electronic states undergo a large change since these electrons, the valence electrons, will be the ones involved in the chemical bonding. The core electrons which, on the other hand, are low in energy remain essentially unchanged no matter the chemical surrounding and therefore they do not influence the material properties to the same extent. Plane-waves are the eigenfunctions for free electrons and are therefore the obvious choice for a basis set when describing the weakly bound valence electrons. If focus is put on describing the chemical bonding, a natural approximation is then to freeze the core states and solve the Kohn-Sham equation for the valence states in a plane wave basis set. This is the fundamental idea of the pseudopotential approximation. The tightly bound core states are frozen since the solution of the
atomic-orbital-like core states would require impractically many plane-waves in the basis.

Since the hard potential and the core states of the atomic cores hardly change for different calculations, the potential is replaced by a weaker pseudopotential for the pseudo-wavefunctions \( \psi^{PS} \).

The Phillips-Kleinman method is an illustrative example of the basic principles of pseudopotentials [13]. Let \( \psi_c \) and \( \psi_v \) be the exact core and valence states respectively. \( \psi_v \) then solves the Schrödinger equation with eigenvalues \( \varepsilon_v \),

\[
\mathcal{H} \psi_v = \varepsilon_v \psi_v. \tag{3.15}
\]

The pseudo wavefunctions are smooth functions expressed as expansions of planes waves as in Eq. (3.13) above. The pseudo wavefunctions are not orthogonal to the core states so the exact valence states can be related to the pseudo wavefunctions with the part linearly dependent on the the core states subtracted,

\[
\psi_v = \psi_v^{PS} - \sum_c \langle \psi_c | \psi_v^{PS} \rangle \psi_c. \tag{3.16}
\]

Inserting this into Eq. (3.15) gives,

\[
\mathcal{H} \psi_v^{PS} - \sum_c \langle \psi_c | \psi_v^{PS} \rangle \mathcal{H} \psi_c = \varepsilon_v \psi_v^{PS} - \sum_c \langle \psi_c | \psi_v^{PS} \rangle \varepsilon_c \psi_c.
\]

Since \( \mathcal{H} \psi_c = \varepsilon_c \psi_c \) for the core states, this can be rewritten as follows,

\[
\mathcal{H} \psi_v^{PS} = \varepsilon_v \psi_v^{PS} + \sum_c \langle \psi_c | \psi_v^{PS} \rangle (\varepsilon_c - \varepsilon_v) \psi_c. \tag{3.17}
\]

Hence the problem of solving a Schrödinger equation in a hard potential \( \mathcal{H} = -\frac{1}{2} \nabla^2 + V_{eff} \) for the exact valence states \( \psi_v \) can be transformed into the easier problem of solving the Schrödinger equation above in a softer potential,

\[
V^{PS} = V_{eff} + \sum_c (\varepsilon_c - \varepsilon_v) |\psi_c\rangle \langle \psi_c| \tag{3.18}
\]

for the pseudo-wavefunctions \( \psi_v^{PS} \) but with the true energy eigenvalues, \( \varepsilon_v \). This soft potential is called the pseudopotential and it is calculated once and for all, along with the core states, for an isolated atom. Note that this pseudopotential is non-local, i.e. it does not just operate on the wavefunction by the simple multiplication of a function, \( f(r) \).

The Phillips-Kleinman pseudopotential serves as a nice example but for numerical reasons it is complicated and therefore more efficient pseudopotentials have been developed. These share the typical pseudopotential properties that
inside the core region the potential should be soft whereas outside the core region the pseudopotential becomes equal to the effective potential as the core wavefunctions vanish (see Fig. 3.2). Therefore, the pseudo wavefunctions will be equal to the exact all electron wavefunctions outside a certain radius $r_c$ of the atom.

The main purpose of pseudopotentials is to drastically reduce the computational effort and therefore a good pseudopotential should be as soft as possible, meaning that as few plane-waves as possible should be needed for the expansion of the pseudo wavefunction. The pseudopotential should also have the property that although it is generated from a certain atomic configuration, it should be transferable to other problems and hence be as accurate regardless whether it is used in a single atom calculation or in a crystal. The charge density of the pseudo-wavefunctions should, of course, also be as close as possible to the true valence density since this is an important physical property [6].

A group called norm conserving pseudopotentials [14] has been constructed with the third requirement of an accurate charge density in mind. As before, the pseudo-wavefunctions and potential are constructed to be equal to the actual valence wavefunction and the original potential outside the core radius $r_c$, but now it is also a condition that the norm of the pseudo-wavefunctions and
the original wavefunctions should be equal inside of the core radius,
\[ \int_0^{r_c} |\psi_{iPS}(r)|^2 dr = \int_0^{r_c} |\psi_i(r)|^2 dr. \] (3.19)

If, instead, focus is placed on the first condition of constructing as soft a pseudopotential as possible for challenging problems, then the norm condition is removed and the pseudo-wavefunctions are allowed to be softer inside \( r_c \). These potentials are called \textit{ultrasoft potentials} and with the corresponding methods large values of the core radius can be used so that the cutoff energy can be drastically reduced (see Section 3.3.2).

In comparison to full potential methods, the pseudopotentials apparently do not give the total charge density, but only the one produced by the valence electrons. For the same reason, only the energy contribution from the valence states is calculated, which is normally a thousand times smaller than the full electronic energy. Often in physics, energy differences are what is important and then this can actually be an advantage: since the change in energy lies almost entirely in the valence contribution, pseudopotential methods will give more accurate results because of their quick convergence. (Taking the difference of two large quantities can give very large numerical errors, especially if the calculation are computationally so demanding that the results cannot be fully converged).

3.3.1 The projector augmented wave method

The projector augmented wave method [15] (PAW) was constructed out of the linear augmented-plane-wave (LAPW) method and the ultrasoft pseudopotential method (US-PP). It is not a traditional pseudopotential method but rather an all-electron method in the frozen-core approximation. It is based on the transformation between the exact all electron wavefunctions, \( \Psi_n \) and the smooth pseudo wavefunctions, \( \tilde{\Psi}_n \) (\( n \) is here a composite quantum number for the band, k-point and spin)
\[ |\Psi_n\rangle = \mathcal{T} |\tilde{\Psi}_n\rangle. \] (3.20)

This leads to an equivalent Kohn-Sham equation for the pseudo wavefunctions
\[ \mathcal{T}^\dagger H \mathcal{T} |\tilde{\Psi}_n\rangle = \epsilon_n \mathcal{T}^\dagger \mathcal{T} |\tilde{\Psi}_n\rangle. \] (3.21)

When solved, the pseudo wavefunctions are transformed back to the true wavefunctions which are then used to evaluate the total energy [16]. This is the basic outline of PAW and now we only need to define the transformation operator, \( \mathcal{T} \). \( \mathcal{T} \) is expressed in terms of the solutions of the Schrödinger equation for an isolated atom, \( |\phi_i\rangle \), and the soft pseudo version of them, \( |\tilde{\phi}_i\rangle \) (also called all-electron and pseudo partial waves).
\[ \mathcal{T} = 1 + \sum_i (|\phi_i\rangle - |\tilde{\phi}_i\rangle) \langle \tilde{\phi}_i|, \] (3.22)
where $\langle \tilde{p}_i |$ are the projector operators dual to the states $| \tilde{\phi}_i \rangle$, fulfilling $\langle \tilde{p}_i | \tilde{\phi}_i \rangle = \delta_{ij}$, if $i$ and $j$ belong to the same augmentation sphere. (If $\tilde{\phi}_i$ were a complete basis set in the entire space, the corresponding complex conjugates could be used as projectors). The sum in Eq. (3.22) is over all basis partial waves centered on an atom and also over all atoms or, rather, augmentation spheres. Outside the augmentation spheres the pseudo partial waves are defined to be equal to the all electron partial waves such that the second term of $\mathcal{T}$ is equal to zero ($| \tilde{\phi}_i \rangle = | \phi_i \rangle$, for $r > r_c$). As for the pseudopotential methods the all electron and the pseudo wavefunction will then be equal outside the augmentation spheres, as seen from combining Eq. (3.20) and Eq. (3.22)

$$
| \Psi_n \rangle = | \tilde{\Psi}_n \rangle + \sum_{i} \left( | \phi_i \rangle - | \tilde{\phi}_i \rangle \right) \langle \tilde{p}_i | \tilde{\Psi}_n \rangle = | \tilde{\Psi}_n \rangle + \sum_{\alpha} \left( | \Psi^{(1)}_{1\alpha} \rangle - | \tilde{\Psi}^{(1)}_{1\alpha} \rangle \right),
$$

(3.23)

where $\alpha$ indicates summation over all augmentation spheres. Inside the augmentations spheres the pseudo wavefunction is identical to its expansion in pseudo partial waves, $| \tilde{\Psi}_{1\alpha} \rangle$. Therefore these terms will cancel in Eq. (3.23) and the all electron wavefunction will be equal to $| \Psi^{(1)}_{1\alpha} \rangle$ which is the true wavefunction in the frozen core approximation. (The partial waves, $| \tilde{\phi}_i \rangle$ and $| \phi_i \rangle$, are not recalculated in the frozen core approximation).

Using Eq. (3.23) to derive the expressions for the electron density and total energy gives

$$
n(r) = \tilde{n}(r) + \sum_{\alpha} \left( n^{(1)}_{1\alpha}(r - R_{\alpha}) - \tilde{n}^{(1)}_{1\alpha}(r - R_{\alpha}) \right),
$$

(3.24)
where $\mathbf{R}_\alpha$ are the positions of the ions, and

$$E[\tilde{\Psi}_n, \mathbf{R}] = \tilde{E} + \sum_\alpha \left( E^1_\alpha - \tilde{E}^1_\alpha \right). \quad (3.25)$$

$\tilde{E}$ is the contribution from the pseudo wavefunctions and is similar to that of pseudopotential methods.

To compensate for the lack of norm-conservation for the pseudo wavefunctions and maintain the condition of charge neutrality, compensating charges are also introduced in the augmentation spheres. A restriction is put on them to have the same multi-pole moments as the all electron charge density.

For a given accuracy PAW needs fewer plane-waves and hence a lower energy cutoff than norm-conserving pseudopotential methods and is therefore less time consuming. The computational effort is instead more comparable to that of ultrasoft pseudopotentials. PAW has the advantage that is describes materials with large magnetic moments, some transition metals, alkali and alkali-earth metals more accurately than US-PP. For other materials PAW and US-PP give almost indistinguishable results. Other advantages of the PAW method are, for example, that the all electron density and potential is obtained and not just the valence part, non-collinearity for magnetic moments have been implemented and that the frozen core approximation could, in principle, be overcome. Most important, PAW is an exact all-electron method under the assumption that a complete partial wave basis set is used and should then give results identical to those of other all-electron frozen core methods.

### 3.3.2 Ultrasoft pseudopotentials

In the ultrasoft pseudopotential method [17] the condition that the pseudo wavefunctions must have the same norm as the all electron wavefunctions inside the core radius is relaxed. This is accomplished through a generalization of the orthonormality condition of the wavefunctions, $\langle \psi_\alpha | S | \psi_\beta \rangle = \delta_{\alpha\beta}$. This allows the pseudo wavefunctions to be as soft as possible in the core regions, drastically lowering the plane-wave cutoff energy and resulting in a great reduction in computational time in comparison to norm-conserving pseudopotentials. First row elements and transition metals are overwhelmingly time consuming for a norm-conserving pseudopotential treatment since their hard potentials require impractically high energy cutoffs. The total electronic charge is instead ensured to be correct through introducing localized atom-centered augmentation charges. The total electron density is therefore composed of a smooth part extended over the entire unit supercell and a hard part which is only non-zero inside the core radii. The transferability properties of the US-PP are not compromised and remain as good as those of norm-conserving pseudopotentials.

As mentioned earlier, the PAW method is developed from the ideas of LAPW and US-PP and in retrospect the US-PP method can be viewed as
an approximation of the PAW method [18]. If the contributions from the core region to the PAW total energy functional (the second and third term in Eq. (3.25)) are only included as linear approximations the US-PP total energy expression is obtained,

$$E_{\text{Tot}}[\{\psi_i\}, \{R_I\}] = \sum_i \langle \psi_i | -\nabla^2 + V_{NL} | \psi_i \rangle + \frac{1}{2} \int \int \frac{n(r)n(r')}{|r-r'|} \, dr \, dr'$$

$$+ E_{\text{xc}}[n] + \int V_{\text{loc}}^{\text{ion}}(r) n(r) \, dr + U(\{R_I\}), \quad (3.26)$$

where the local and non-local parts of the pseudopotential are incorporated in $V_{\text{loc}}^{\text{ion}}$ and $V_{NL}$ respectively. The difference between US-PP and PAW lies in the pseudization of the augmentation charges. In fact, in the case of having the exact augmentation charges (and no basis set truncations) the US-PP would reproduce the results of PAW. For most materials the results are also very similar for US-PP and PAW, although a slightly lower energy cutoff can be used in US-PP. The exceptions are the materials listed in the previous section, these magnetic materials and transition or alkali metals are more easily and accurately described using PAW. This is because these materials require hard augmentation charges which are difficult and computationally expensive to represent on the regular grid of the US-PP method (PAW instead uses a radial support grid in the augmentation spheres).

### 3.4 Forces and ionic relaxation

So far the task of calculating the electronic contribution to the energy of a fixed ionic geometry has been dealt with. In this section ways to find the ionic configuration with the lowest energy will be discussed. An ion experiencing a net force will move in the direction of the force in order to minimize the energy. The equilibrium configuration of a system will be found when all ions have a net force equal to zero. The force on an ion is simply the derivative of the energy with respect to the ions position, but since the movement of an ion indirectly may change all the terms in the expression for the energy in Eq. (3.6), this can be greatly simplified using the Hellmann-Feynman theorem:

$$F_i = -\frac{dE}{dR_i} = -\frac{d}{dR_i} \sum_i \langle \psi_i | H | \psi_i \rangle$$

$$= -\sum_i \langle \frac{\partial \psi_i}{\partial R_i} | H | \psi_i \rangle - \sum_i \langle \psi_i | \frac{\partial H}{\partial R_i} | \psi_i \rangle - \sum_i \langle \psi_i | H | \frac{\partial \psi_i}{\partial R_i} \rangle$$

$$= -\frac{\partial}{\partial R_i} \sum_i \varepsilon_i \langle \psi_i | \psi_i \rangle - \sum_i \langle \psi_i | \frac{\partial H}{\partial R_i} | \psi_i \rangle = -\frac{\partial E}{\partial R_i}, \quad (3.27)$$

where $H | \psi_i \rangle = \varepsilon_i | \psi_i \rangle$ has been used in the last step and the first term on the last row disappears since the derivative of the normalization constants are zero.
The theorem greatly reduces the computational effort since only the explicit dependence on $\mathbf{R}_i$ has to be considered. In order to calculate the force without it one would have to minimize the energy functional for many ionic configurations infinitesimally close to the one the force is being calculated for (since it’s non-trivial how the exchange-correlation part would change when the ion is moved, for example). However, the theorem is based on the fact that $\psi_i$ are the exact wavefunctions and since the use of approximate wavefunctions in Eq. (3.27) will introduce an error in the calculated forces it is crucial that the electronic configuration is properly converged in each step of the ionic relaxation.

An additional term may also appear in the expression for the force. It originates from the derivative of the basis set with respect to the ion position and it is called the Pulay force. This will introduce an additional error if it is not calculated explicitly. Luckily, the plane-wave basis set is independent of the positions of the ions and the Pulay force will therefore be zero. Caution must be taken though, when the size or shape of the unit cell is changed. This will alter the cutoff energy resulting in artificial energy differences unless the plane-wave basis is large enough for full convergence for both unit cells.

Moving the ions according to the forces will eventually lead to an equilibrium configuration of the system, corresponding to a local minima. There is, however, no guarantee that this is the global minima. (Methods such as simulated annealing or Monte Carlo simulations have been developed to find the global minima for structures, but it can not be guaranteed that the global minimum is found in a finite time). Therefore, there is currently no better way of finding the global minima than the pragmatic method of investigating if different initial geometries relax to the same low energy structure.

3.5 Chemical bonding in solids

Once the electron density has been calculated, the desired material properties can be examined. No matter if the hardness of a material or the reconstruction of a surface is of main interest, the way in which the electrons form bonds is of great importance. There are three main types of bonds in crystalline matter, covalent, ionic and metallic, which divide materials into distinct types (insulator/semiconductors, ionic compounds and metals). It is therefore important to be able to analyze the chemical bonding of a material from the wavefunctions or, in the case of DFT, rather the electron density.

3.5.1 Charge density transfer

One simple yet powerful way to visualize the bonding is to see how electrons are redistributed when atoms are composed into a solid. Letting $\rho_{\text{cryst}}(\mathbf{r})$ be the self-consistent electron density of a fully relaxed configuration of ions and
\( \rho_{\text{atomic}}(r) \) the summed atomic charge distributions for the same ionic structure, this can be quantified as the difference

\[
\Delta \rho(r) = \rho_{\text{cryst}}(r) - \rho_{\text{atomic}}(r).
\] (3.28)

Covalent bonds will be recognized as a charge accumulation in between two atomic cores and they will hence be clearly directional and localized. For ionic bonds, \( \Delta \rho(r) \) will have minima at the positions of one species of atomic cores and maxima on the other species. This corresponds to electrons completely being pulled over by the more electro-negative type of atoms. Finally, metallic bonds will be seen as an overall charge transfer from the ion cores to the interstitial regions, but unlike the covalent bond the electrons are delocalized and evenly distributed without any directional properties.

Plots of the charge density transfer will vary continuously from one of these typical cases to the other. Therefore the method does not make a clear distinction between different types on bonds and cannot provide a strict way of mapping out materials or classifying bonds. Rather, it provides a visualization of the fundamental features; chemical bonds are formed first when atoms are put together to a molecule or a solid. A plot of the charge density transfer therefore, in the simplest way possible, shows how these bonds are formed.

### 3.5.2 The electron localization function

The concept of electron localization is not an exact defined physical quantity but rather an intuitive one. An attempt to introduce a definition for it was done in the Hartree-Fock formalism [19] based on the conditional probability of finding an electron at a position \( r \), under the assumption that an electron of the same spin is located at \( r' \). The first term in the spherically averaged Taylor expansion of this conditional probability was used in a function \( \eta \), renormalized to take values from zero to one. \( \eta \) is called the electron localization function (ELF) and is defined at all positions \( r \) in space. The interpretation in this derivation is that for a strongly localized electron the probability of finding another electron of the same spin nearby is almost zero. If instead the electron is very delocalized there is a relatively large probability of finding a second electron at the reference point.

Later this functional form of the ELF was incorporated into DFT [20]. If \( \psi_i \) denote the Kohn-Sham orbitals and \( \rho \) the charge density, the first term of the Taylor expansion mentioned above takes the form

\[
D(r) = \frac{1}{2} \sum_i |\nabla \psi_i(r)|^2 - \frac{1}{8} \left( \frac{|\nabla \rho(r)|^2}{\rho(r)} \right),
\] (3.29)

where the first term is the kinetic energy density of a number of non-interacting electrons in the Kohn-Sham formalism and the second term is its lower bound, the kinetic energy density for particles behaving like bosons.
The electron localization function remains unchanged,
\[ \eta(r) = \frac{1}{1 + \left( \frac{\rho}{D} \right)^2}, \quad (3.30) \]
where
\[ D_h(r) = \frac{3}{10} (3\pi^2)^{5/3} \rho(r)^{5/3} \quad (3.31) \]
is the value of \( D(r) \) corresponding to a homogeneous electron gas of electron density \( \rho \).

The interpretation of \( D(r) \) in Eq. (3.31) is to be a measure of the excess local kinetic energy due to the Pauli principle. In regions where electrons are alone or in a pair of different spin, there is no repulsion from the Pauli principle and therefore \( D = 0 \) and \( \eta = 1.0 \), corresponding to a perfect localization of electron in that point. In a region of higher excess kinetic energy density, \( D \) will be non-zero and \( \eta < 1 \). At a point of the same kinetic energy as an homogeneous electron gas, for example, \( D = D_h \) and \( \eta = 1/2 \).

The central idea of the ELF is that the repulsion from the Pauli principle, quantified by \( D(r) \), is used as a measure of the electron localization relative to a homogeneous electron gas of the same density (no matter if it is interpreted as a high probability of finding two electrons of the same spin in the same region or as a local kinetic energy density higher than the bosonic one).

The numerical value of \( \eta \) should not be taken as a precise measure of the electron localization but the strength of ELF lies in the topological analysis of its iso-surface. The local maxima of \( \eta \), for example, correspond to ion cores, bonds or anti-bonds. Through this it is possible to clearly distinguish between shared electron interaction (covalent and metallic bonds) and closed-shell interaction (ionic, hydrogen, electrostatic and van der Waals bonds); the former has at least one \( \eta \)-maximum located in-between the ion cores whereas the other types has nothing between the bonding cores. Further, both lone pairs and atomic shells can also be identified from a topological analysis of the ELF.

The ELF therefore provides a possible way to classify bonds in contrast to the charge density transfer method. On the other hand the definition of the latter is unquestionable while there may be some things ambiguous about the definition of electron localization.

### 3.6 The deficiencies of DFT

Although DFT has had a huge success in quantum computations, one must be aware of the situations in which it is not give a good representation of the
physics. To begin with, DFT is only concerned with the ground state properties and this corresponds to no thermal motion and zero temperature. Further, the weakest point in DFT is often the approximation which is made for the exchange-correlation energy, in this case the local density approximation. The variational principle ensures that the energy obtained from DFT with an exact exchange-correlation functional always is bigger than or equal to the true energy, and that it converges down to that true energy. Using LDA, this is not true anymore. Although a global minimum is found, it is not necessarily larger than the true energy. The only thing known is that this value is the best under the approximation and it depends on the approximation whether it is close to the true minimum or not.

Classical examples of errors related to the use of LDA are that iron is predicted to have a paramagnetic fcc crystal structure instead of a ferromagnetic bcc structure, the band gaps of semiconductors are typically 50% too small and some transition metal oxides are calculated to have metallic properties when they, in fact, are insulators.

It must also be noted that the Kohn-Sham equations are found as conditions which have to be fulfilled in order to minimize the energy functional. These are then interpreted as one particle Schrödinger-like equations. Therefore, the Kohn-Sham eigenvalues have no physical significance in a strict sense. However, under the assumption that the exact $V_{xc}$ is used, it has actually been shown that the eigenvalue of the highest occupied Kohn-Sham orbital is equal to the exact ionization energy (Koopman’s theorem) [2].
**4. Defects and impurities in semiconductors**

There are many types of semiconductors: elemental group four (Si, Ge) semiconductors, binary III-V semiconductors (GaAs, InP, AlSb,...), II-VI semiconductors (ZnS, CdTe), oxides and various non-crystalline organic semiconductors. In this work, focus will be put mainly on the III-V semiconductors. These have tetragonally oriented covalent bonds where every atom is four-fold coordinated. The structure is typically of Zincblende type where the bonds, in contrast to the elemental semiconductors, become partly ionic due to the difference in the number of electrons of the group III and V atoms. In this way the group III atoms contribute 3/4 of an electron to each bond and the group five elements 5/4 electrons. This will come in handy when discussing impurities in the next section.

What characterizes semiconductors is that they have a gap of forbidden states above the highest occupied band, which is called the **band gap**. The band gap therefore separates the fully occupied valence bands from the higher conduction bands. Unless electrons are thermally or optically excited to the conduction band, semiconductors will not lead electric current. As mentioned in the introduction, semiconductors will hence be useless in electronic applications without atomic impurities. The electrical properties of semiconductors can be controlled through artificial implantation of foreign atoms. Herein lies the reason to why the functionality of all modern electrical devices, more advanced than a lamp, are based on doped semiconductors.

Everything altering the ideal crystal structure of a semiconductor will, in fact, affect its electronic properties. These disturbances are called **crystal defects** and examples are line defects, dislocations, stacking faults and point defects. The first three can fairly easily be avoided when growing semiconductor crystals while point defects will always occur in solids at thermal equilibrium, no matter if the initial solid was of ideal crystal structure.

Point defects can be divided into **native defects** (or intrinsic defects), formed only from the host atom types, and impurities (or extrinsic defects) consisting of foreign atoms.
4.1 The electrical properties of impurities

An impurity or native defect will alter the electrical properties of a semiconductor if it either acts as a donor, contributing electrons to the conduction band, or an acceptor, creating holes in the valence band (Fig. 4.2 (a)). Whether an atom of a certain type acts as a donor or acceptor is dependent on the location of the defect in the crystal lattice. For example, a group IV atom in a III-V semiconductor substituting a group III atom will only need three electrons to complete the four bonds to its closest neighbors and the last electron will occupy the conduction band and hence act a free charge carrier. In this case the group IV atom will act as a donor. If it is instead located on a group V lattice site, there will be one electron missing and it will act as an acceptor. As will be shown later, the same impurity atom can actually act both as acceptor or a donor depending on the position of the Fermi level. Acceptors have shallow defect levels slightly above the valence band edge and donors have shallow defect levels slightly below the conduction band edge. In this way the positions of the dopant’s defect levels are of great importance for the concentration of charge carriers.

A semiconductor with a considerable amount of acceptors inserted will have a large charge carrier population in the form of positively charged holes in the valence band. It will therefore be said to be p-doped. Correspondingly, an n-doped semiconductor has a large concentration of donors, giving rise to a charge carrier population of electrons in the conduction band. Joining a p- and n-doped semiconductor will form the electrical component called the p-n junction. The physics of the p-n junction is here briefly outlined since this serves as a nice and simple illustration of how impurities and defects are a vital part of the functionality of electronics. When the two parts are put in contact (Fig. 4.2 (a)), the system is not in equilibrium and the lower Fermi level $\varepsilon_F$ in the p-doped part will cause electrons to flow in from the the n-doped part (vice
versa for holes). In this way a layer of a certain thickness around the interface will be totally emptied of charge carriers (Fig. 4.2 (b)). This layer is called the depletion region and it will have a net charge, $q$, from negative acceptor ions on the $p$-doped side and from positive donor ions on the $n$-doped side. This results in an electric field over the interface which will cause a band bending of the valence and conduction bands (Fig. 4.2 (c)). The construction demonstrates the basic function of optoelectronic devices; in a detector, photons of energy $\hbar \omega$ equal to the band gap will excite electrons into the conduction band and cause a measurable voltage between contacts on the $p$- and $n$-doped sides. In a light emitting diode (LED) or a laser, a voltage is instead applied over these contacts, pumping electrons into the conduction band of the $n$-doped part. These electrons will then recombine with holes at the interface, emitting photons of energy $\hbar \omega$ equal to the band gap. In principle, this process will be without heat loss and this is why LED’s are very efficient.

In order to make semiconductor devices as good and fast as possible the number of charge carriers should be made as large as possible. There are different factors that put a limit on the maximum doping concentration. First, impurities are normally most stable at substitutional sites and introducing a dopant atom will raise the number of charge carriers by one (if it is a single donor/acceptor and there is available thermal energy). This is only true up to
a certain concentration, after which the concentration of substitutional impurities is saturated. Additional impurity atoms will instead occupy interstitial sites or the other substitutional sites in compound semiconductors (where it may act as a dopant of the other type, reducing the number of charge carriers).

Second, native defects will, as mentioned, always be present and may occur in concentrations large enough to affect the number of charge carriers. Antisites, interstitials or vacancies all have defect levels in the band gap and depending on their position these will be either acceptors, donors or deep levels. In a material that is attempted to be made strongly \( p \)-doped, for example, a large concentrations of native donors will severely lower the number of holes. This is known as self-compensation. Many new promising semiconductor materials can, in fact, only be doped either \( p \)- or \( n \)-type because of self-compensation. New semiconductor materials that are beneficial in respect to these two issues are therefore of high research interest.

Deep levels, which are defect levels located in the middle of the band gap, have a tendency to interfere with the optical functionality of semiconductors. The electrons in the \( p-n \) junction above would, for example, recombine in two steps, with the deep level as an intermediate, not emitting a photon of energy \( h\omega \) but rather two phonons and the energy would be turned into heat. A high concentration of defects with deep defect levels can thereby lower the efficiency of optical devices considerably.

If native defects are studied and understood these problems may be avoided, and semiconductors could even be engineered in such a way to make native defects work in favor of the wanted electric properties. (CuInSe\(_2\) is for example doped by changing the stoichiometry and thereby the numbers of native defects).

### 4.2 The structure of point defects

A native defect or impurity will behave differently depending on its location in the crystal lattice. An impurity can either replace a host atom at a lattice site, which is known as a substitutional impurity, or be located in an open space in-between the host atoms at an interstitial position. In III-V semiconductors there are two different substitutional sites since the impurity can either replace a group III or a group V atom. Interstitial atoms can, in principle, be located anywhere in-between lattice sites but will in practice always be located at (or close to) the most open regions where they cause the least strain on the surrounding crystal. In III-V semiconductors there are three main interstitial sites: the two tetragonal sites, either surrounded by four group III or V atoms and the hexagonal site. The native defects have the same types of interstitials, with the only difference being that the interstitial atom either can be a group III or group V atom. The other kinds of defects are the antisite defect, where a group III atom substitutes a group V atom at the ordinary lattice site (or vice
versa) and the vacancies, where an atom of either type is missing from the lattice.

For the impurities one can often predict whether it is going to behave like an acceptor or donor from the number of valence electrons. A group II (VI) substituting a group III (V) atom is expected to form a acceptor (donor) level in the band gap, as previously explained. For the native defects, on the other hand, it is not as apparent where their defect levels will be located in the band gap. For both vacancies and the cation antisite defect, the defect levels will ideally consist of a single non-degenerate level ($a^1_v$) and a higher triply degenerate level ($t^2_v$) originating from the valence band. In the neutral charge state these levels will be occupied by the electrons from the nearest-neighbors and also, in the case of the antisite, by the cations valence electrons. The lower and fully occupied $a^1_v$-level will often remain inside the valence band and leave only the triply degenerate $t^2_v$-level in the band gap, occupied by three, five and six electrons for the anion-, cation-vacancy and cation antisite respectively. For the anion antisite and the interstitial defects, the defect levels in the band gap will instead originate from the conduction band. For the interstitials, the valence electrons of the interstitial atom will occupy $a^1_c$ and a higher $t^2_c$ level and for the anion antisite the five electrons of the nearest-neighbors and the five valence electrons will first occupy the $a^1_v$ and $t^2_v$ levels inside the valence band and the remaining two electrons will then occupy the, typically deep, $a^1_c$-level in the band gap (see Ref. [22] for details).

This rough analysis might be interesting to get a better understanding of the underlying nature of the defect levels, but the actual positions of the defect levels are, of course, very much dependent on which III-V semiconductor they are in. Even more important, the symmetry of the defect will be reduced and the degeneracy of the defect level lifted if it lowers the energy of the system (the Jahn-Teller effect). The only way to find the magnitude of this geometrical distortion is by explicit calculation. The next section is devoted to how to calculate the defect levels in an accurate way.

4.3 Theoretical treatment of defects

For defects, either found intrinsically in or extrinsically doped into a semiconductor which is to be used in an electrical device, there are two properties of main interest: their defect levels and their concentrations. The positions and occupancies of its defect levels are of interest because it will determine if the defect act as an acceptor, deep defect or donor, as described above. The defect concentration is obviously of interest because the concentration of charge carriers will be directly proportional to the concentration of the dopants. If the dopant only occurs in small numbers it will be useless no matter how good it might be in other aspects. In this section the theoretical tools for calculating these two properties will be derived.
There are, of course, many other properties of interest, such as the mobility and optical properties of defects and impurities, and these can be derived from that the defect levels and the defect energy (which is used to calculated the concentration). As mentioned before, defects are always present in real materials and they will therefore always also, to various extent, influence all material properties, conduction, color, magnetic behavior and so on.

4.3.1 The defect formation energy

The formation energy of a defect in a semiconductor is simply the energy required to create the defect from a perfect semiconductor crystal. Any particles added or removed in order to do this are considered to come from or being moved to particle reservoirs of certain chemical potentials. The chemical potential is therefore a measure of how expensive it is to add an atom of a specific kind.

\[ \varepsilon_{\text{form}}(q) = E_{\text{Tot}}^{\text{Def}}(q) - E_{\text{Tot}}^{\text{Bulk}} - \sum \Delta n_A \mu_A + q(\varepsilon_F + E_v), \]  

(4.1)

where the first two terms on the right hand side are the total energies of the systems with and without the defect respectively. \( \Delta n \) is the number of particles added to create the defect and \( \mu \) is the chemical potential of the particle reservoir. The number of electrons added is simply the negative value of the charge state of the created defect, i.e. \( -\Delta n_e = +q \) and the chemical potential for electrons is the Fermi level, which conventionally is put to zero at the valence band edge, \( \mu_e = \varepsilon_F + E_v \) (\( E_v \) being the energy of the valence band maximum). The summation is performed over all types of atoms that change upon creating the defect.

4.3.2 Normalization of the formation energy

By subtracting the formation energy of the neutral defect from that of the charged defect we get the normalized formation energy,

\[ \varepsilon_{\text{form}}(q/0) = \varepsilon_{\text{form}}(q) - \varepsilon_{\text{form}}(0) = E_{\text{Tot}}^{\text{Def}}(q) - E_{\text{Tot}}^{\text{Def}}(0) + q(\varepsilon_F + E_v). \]  

(4.2)

Not surprisingly, this implies that the most stable charge state of the defect is not determined by the chemical potentials or by the energy of the perfect crystal, but only by the total energies of structurally relaxed defects and the Fermi
level. Therefore the chemical potentials only alter the formation energies (and concentrations) relative to other types of defects.

4.3.3 The chemical potential

The chemical potentials define the reservoirs from which particles are taken when growing a semiconductor crystal. The value of the chemical potential reflects the abundance of that type of atom. Rigorously, the chemical potential is the partial derivative of the Gibbs free energy with respect to the number of particles, \( \mu = \frac{\partial G}{\partial N} \) [23]. In the case of a homogenous system this reduces to the gibbs free energy per particle. As seen, from Eq. (4.1) the formation energy of a defect is thus dependent on the growth conditions and the chemical balance (stoichiometry) of the crystal through the variable \( \mu \). More precisely, which was already pointed out in the previous paragraph, the formation energy of one type of defect relative to another is dependent on \( \mu \). But there are restrictions to what values the chemical potentials can have. First, the particles in the reservoir must have a lower chemical potential than that of the corresponding bulk material. That is, \( \mu_{\text{cation}} \leq \mu_{\text{Bulk cation}} \) and \( \mu_{\text{anion}} \leq \mu_{\text{Bulk anion}} \), otherwise the more stable bulk material will be formed [24]. These upper limits are referred to as the cation-rich and anion-rich limits. Second, the parameters \( \mu_{\text{cation}} \) and \( \mu_{\text{anion}} \) are not independent of each other. Under the assumption that the semiconductor is grown under equilibrium conditions at constant temperature and pressure, the balance of the chemical reaction gives that \( \mu_{\text{III–V}} = \mu_{\text{cation}} + \mu_{\text{anion}} \) must be fulfilled.\(^1\) If now \( \mu_{\text{cation}} \) is considered to be the only independent variable, the upper limit will be given by the cation-rich case and the lower limit by the anion-rich case,\(^3\)

\[
\mu_{\text{III–V}} - \mu_{\text{anion}} \leq \mu_{\text{cation}} \leq \mu_{\text{Bulk cation}}. \tag{4.3}
\]

For a given defect with a fixed charge state, the formation energy of this defect will be dependent on the cation chemical potential. In addition, unless the defect is neutral, the formation energy will also be dependent on the Fermi energy from which electrons are taken to charge the defect.

4.3.4 Defect density

The total energy of a piece of material with \( n \) defects of a certain type and charge state \( q \) is

\[
E = E_0 + n \cdot \varepsilon_{\text{form}}(q), \tag{4.4}
\]

where the \( E_0 \) is the total energy of the entire crystal without any defects. The entropy from having \( n \) defects in the system is correspondingly

\[
S = S_{\text{config}} + n \cdot s_{\text{form}}, \tag{4.5}
\]

\(^1\)The notation may be clearer from an example: \( \mu_{\text{InAs}} = \mu_{\text{In}} + \mu_{\text{As}} \).
where $S_{\text{config}}$ is the configurational entropy from all the defects and $s_{\text{form}}$ is the individual entropy contribution at each defect site. The first term can be calculated from the statistical argument of having $n$ defects distributed on $N$ possible sites (being the total number of cation sites for an anion antisite, for example),

$$
S_{\text{config}} = k_B \ln \frac{N!}{(N-n)!n!},
$$

(4.6)

where $k_B$ is the Boltzmann’s constant.

In order to find the thermal equilibrium number of defects, the Gibbs free energy $G = E - T \cdot S$ needs to be minimized. Therefore the partial derivative of the entropy with respect to the number of defects must be calculated. Under the assumption that the number of possible sites is much larger than the number of defects, $N \gg n$, and using Stirling’s formula this derivative can be written

$$
\frac{\partial S_{\text{config}}}{\partial n} = k_B \ln \left(\frac{N-n}{n}\right) \approx k_B \ln \left(\frac{N}{n}\right).
$$

(4.7)

Using this expression to find the local minimum of $G$ gives:

$$
\frac{\partial G}{\partial n} = e_{\text{form}}(q) - k_B T \ln \left(\frac{N}{n}\right) - T s_{\text{form}} = 0.
$$

(4.8)

Rearranging terms then gives the final expression,

$$
n = N e^{-e_{\text{form}}/k_B T} \cdot e^{s_{\text{form}}/k_B}.
$$

(4.9)

Replacing $N$ by $N/V$ in Eq. (4.9) gives the defect concentration, $[C] = \frac{n}{V}$.

The formation energy, $e_{\text{form}}$, is straightforward to calculate using ab initio methods according to Eq. (4.1). The formation entropy $s_{\text{form}}$, on the other hand, is more difficult to calculate. It contains all the entropy contributions of a single defect, which include the vibrational entropy, the internal configurational entropy (which depends on the symmetry of the defect) among other terms,

$$
s_{\text{form}} = s_{\text{vib}} + s_{\text{int.conf}} + \cdots.
$$

(4.10)

When it comes to the concentrations of dopants and charge carriers, it is really the orders of magnitude of the numbers that are relevant. Luckily, the exponential of the entropy term in Eq. (4.9) is typically something like seventeen orders of magnitude smaller than the exponential of the formation energy term. Therefore, it is the formation energy of a defect that dictates whether or not it will be present in large numbers. The entropy contribution lies in the range $s_{\text{form}} \in [0, 10k_B]$, where explicit numbers of $5 - 6k_B$ have been calculated for point defects in Si. [25, 26] Since the dependence of the concentration on the entropy is exponential, the entropy contribution is expected to raise the concentration with about two orders, but can to a good approximation be assumed to be the same for similar defects.

34
4.3.5 Definition of the Charge Transfer Level

The charge transfer level (CTL) is defined as the value of the Fermi level for which the most stable charge state of a given defect changes. That is, the charge transfer level $\varepsilon^{q|Q}$ is the value of the Fermi level for which $\varepsilon^{form}(q) = \varepsilon^{form}(Q)$. Written out this gives,

$$E_{\text{Tot Def}}^q(q) + q(\varepsilon_F^{q|Q} + E_v) = E_{\text{Tot Def}}^Q(Q) + Q(\varepsilon_F^{q|Q} + E_v)$$

$$\Rightarrow \varepsilon_F^{q|Q} = \frac{1}{q - Q} \left\{ E_{\text{Tot Def}}^Q(Q) - E_{\text{Tot Def}}^q(q) \right\} - E_v, \quad (4.11)$$

where $\varepsilon_F^{q|Q}$ is given relative to the valence band edge. This defines the thermal ionization energy, which is just the energy needed to excite an electron from the defect level to the conduction band, $E_{\text{ion}}^{th} = E_{\text{gap}} - \varepsilon_F^{q|Q}$. The optical transfer levels, on the other hand, are calculated similarly but since ionic movement is on a much slower time-scale than electronic, the final geometry must be equal to the initial geometry. If there is no significant difference in the geometry for the different charge states, the thermal and optical levels will, however, coincide.

Thermal excitation of a valence electron to a (deep) defect level normally makes the defect level less attractive to a second electron because of Coloumb repulsion. In rare cases the relaxation effects can overcome this repulsion making it more favorable to occupy the defect level by two electrons at the same time (for example by forming a new covalent bond). This is called the negative-U effect. A schematic illustration of this is shown in Fig. 4.4: (a) the Coloumb repulsion ensures a normal ordering of the defect levels. (b) relaxation effects puts $\varepsilon_F^{+|0} > \varepsilon_F^{0|-}$ and the defect will never be stable in the neutral charge state.

![Figure 4.4: Schematic illustration of the negative-U effect.](image)

35
4.3.6 The supercell approximation
Defect calculations can either be performed in clusters or in supercells. Clusters have the obvious disadvantage that the electronic states are never entirely delocalized as in real materials. In supercells, on the other hand, there is an artificial interaction between the defect and its images in the neighboring supercells. The interaction might both be electrostatic and quantum mechanical from the overlap of defect wave-functions. A single defect in an otherwise perfect and infinite crystal has dispersion free and atomic like defect states in the band gap. An example of the dispersion of a deep defect level for different supercell sizes is shown in Fig. 4.5. It can be seen that the dispersion of the defect level decreases with increasing supercell size, as expected.

\textbf{Figure 4.5:} Dispersion of the $P_{Ga}^{0}$ defect level in GaP for the 8, 64, 216 and 512 atom supercells.

For a given supercell size, the dispersion may be even larger for shallow states near the conduction or valence band. For this reason there has been a number of corrections developed in order to perform more accurate calculations already in small supercells. [27, 28]

Another technical aspect of defect calculations is that the introduction of a defect might change the zero-level of the energy in the calculations. To calculate the formation energies in Eq. (4.1) accurately it may therefore be necessary to correct for this by aligning the potential at a point far from the defect in the defect supercell with the corresponding point in the bulk supercell.

4.4 Identification of defects
The influence from defects is more easily observed than the defects responsible for it. The reason why ionic crystals (alkali halides) have color, for example, is the presence of point defects. Therefore there are still many unknown defects connected to typical properties of materials. The collaboration between theory and experiment therefore required to be able identify these defects. The experimental techniques, such as electron paramagnetic resonance (EPR)/electron spin resonance (ESR) spectroscopy, deep-level transient spectroscopy (DTLS), optical detection of magnetic resonance (ODMR) and positron annihilation, are used to probe certain properties of the defect. These results are then compared to the theoretical results to find the possible can-
didate for this behavior. After refined examination, all candidates can, hope-
fully, be ruled out except one. Knowing the defect responsible, makes it rel-
avely easy to either magnify or counteract the material property (depending
on whether it is wanted or not) by controlling the defect concentration.
5. Native Defects

5.1 GaP

GaP is a common binary semiconductor in electrical and optical devices. It has an indirect band gap and therefore it is doped when used in light emitting diodes (LED). GaP:ZnO and GaP:N produces red and green light respectively, for example, and GaInP has some of the widest direct band gaps among III-V semiconductors, apart from the nitrides [29]. Any unwanted point defects present may interfere with the functionality of the materials. In the case of optical devices deep levels are especially problematic since they lead to non-radiative recombination. Because of its low cost and high efficiency, GaP is a constituent in a large part of modern semiconductor electronics and optoelectronics. It is therefore of interest to understand how and to what extent native (intrinsic) defects will influence material properties through self-compensation and non-radiative centers.

The most clearly identified native defect in GaP is so far the phosphorus antisite, $P_{Ga}$. It has a deep defect level inside the bandgap and is expected to cause non-radiative recombination. It has been observed in semi-insulating GaP grown under P-rich conditions using EPR and was concluded to be in a singly positive charge state, $P_{Ga}^+$, and to have tetragonal symmetry [30]. Later, a $P_{Ga}^+$ concentration of $4 \times 10^{16}$ cm$^{-3}$ was measured along with an ionization energy of $1.25 \pm 0.10$ eV relative to the valence band, ascribed to the (+|+2) transfer level [31]. The gallium antisite has not been studied as extensively and the only observation of it was made with EPR after neutron irradiation [32]. The gallium interstitial has been identified in Ga$_{0.98}$Al$_{0.02}$N$_{0.99}$P$_{0.01}$ using ODMR where it was found to occupy both the Ga- and P-surrounded tetragonal sites [33]. In another ODMR study Ga$_{3+}^2$ was observed and believed to be bonded to another but unidentified defect [34]. There has been no reported observation of interstitial phosphorus, $P_i$, in GaP.

Positron annihilation has been used to study vacancies in GaP [35, 36]. No vacancies were found in $p$-type or semi-insulating specimen while a large concentration was found in $n$-type, in which the two studies report vacancy densities of $10^{16}$ cm$^{-3}$ and $10^{17}$ cm$^{-3}$. It was believed that these large concentrations were due to phosphorus vacancies, $V_P$. The gallium vacancy has been identified using EPR. The charge state of vacancy was first concluded to be doubly negative, $V_{Ga}^{2-}$ [37]. That it only could be observed after electron irradiation was taken as an indication that it was not the equilibrium charge state in semi-insulating GaP. Since the observed vacancy was of tetragonal symme-
Figure 5.1: Formation energies as a function of the Fermi-level for all fully relaxed native defects under stoichiometric conditions (\(\mu_{Ga} = -4.06\) eV). Values are only shown for the most stable charge state, the charge state given by the slope of the line. Charge transfer levels are labeled next to the vertical lines indicating their positions. The antisite defects are shown as dotted lines, vacancies as full lines and interstitials as broken lines. The alternative y-axis on the right gives the defect concentration in orders of magnitude at a temperature of 300K.

The native defects in GaP have been studied theoretically with the use of first-principles methods in Paper I (details of the computational approach are given therein). The considered native defects are the phosphorus and gallium antisites (PGa and GaP), the vacancies in either sub-lattice (VGa and VP) and the gallium and phosphorus interstitials, each one in either the Ga- or P-surrounded tetragonal sites (GaP, GaP, PGa and PGa). All these defects have the same density of possible sites to occupy. That is, \(N/V\) in Eq. (4.9) is equal to \(2.5 \times 10^{22}\) cm\(^{-3}\) for all of these defects. Therefore, the relative concentration of the defects can be directly determined from a comparison of formation energies.

The formation energy is a function of the Fermi level, \(\varepsilon_F\). For a certain defect, the formation energy is calculated under the assumption that the defect does not interact with other defects and that the Fermi level position is determined by the global doping conditions. Plotting the formation energy’s dependence on the Fermi-level, as in Fig. 5.1, shows what defect will have the highest concentration under all possible doping conditions (that is, if not explicitly taking the condition of charge neutrality into account). The slope of a curve is given by the charge state of the defect (Eq. (4.1)) and for clarity
only the most stable charge state will be displayed for each defect. The charge transfer levels are indicated by vertical lines, labeled by the two charge states. The concentration estimate on the right hand axis is calculated with a typical formation energy contribution, $S^{\text{form}} = 5k_B$, at a temperature of 300 K.

The relaxation effects are calculated to be small for $P_{\text{Ga}}, G_{\text{aP}}$, and $V_{\text{Ga}}$ (with relaxation energies on the order of 0.1-1.0 eV) and largest for $P_1$ and $V_p$ (with relaxation energies up to 3 eV). $V_p$ undergoes a considerable change of structure and symmetry going from one charge state to another, which is schematically illustrated in Fig. 5.2. This strong relaxation gives rise to three negative-U levels inside the bandgap.

Also $P_{\text{Ga}}$ shows a negative-U behavior. The symmetry of this antisite is tetragonal for all relevant charge states and it is instead the volume relaxation effects that merges the charge transfer levels into one single $(+2|0)$ level. The volume relaxation for $P_{\text{Ga}}^{+2}$ is inwards and by 12.6% of the ideal nearest-neighbor volume of 6.52 Å$^3$, while $P_{\text{Ga}}^0$ will have an outward relaxation by 7.4%.

The results of Fig. 5.1 are presented for perfectly stoichiometric GaP. It is seen from this figure that the most occurring native defects will be $P_{\text{Ga}}^{+2}$ under $p$-type conditions, $V_{\text{Ga}}^{-3}$ under $n$-type conditions and $G_{\text{aP}}^{-2}$ in semi-insulating material. The concentration of $G_{\text{aP}}^{-2}$ is, however, much lower than for the other two defects and it will itself have a much higher concentration in $n$-type GaP than in semi-insulating.

As explained in chapter 3, the defect formation energy is dependent on the chemical potentials. This means that the native defect densities will be dependent on under which conditions the GaP crystal has been grown. $G_{\text{aP}}, G_{\text{a}}, V_{\text{p}}$ will obviously be more favorable in GaP grown under Ga-rich conditions, for example. The formation energies for the extreme cases, the Ga- and P-rich limits, are presented in Figs. 5.3 and 5.4.

Note that since the Fermi level is considered to be fixed by the global doping conditions, negative values of the formation energy are possible. An example...
Figure 5.3: As for Fig. 5.1 but under Ga-rich conditions (μ_Ga = −3.61 eV).

Figure 5.4: As for Fig. 5.1 but under P-rich conditions (μ_Ga = −4.50 eV).
of this is $V_{Ga}^{-3}$ in Fig. 5.3. This means that when GaP is strongly n-doped and the Fermi level put close to the conduction band minimum, $V_{Ga}^{-3}$ will have a very large equilibrium concentration. For every Ga vacancy that is formed, three electrons are required, hence reducing the number of negative charge carriers and lowering the value of the Fermi-level. $V_{Ga}^{-3}$ will therefore be a strong compensator in GaP grown under P-rich conditions.

From these figures it is seen that GaP will be the dominant defect in Ga-rich GaP, with the highest density in n-type GaP, while GaP is most common in p-type. In GaP grown under P-rich conditions $P_{Ga}^{2+}$ and $V_{Ga}^{-3}$ will dominate under p- and n-type doping, respectively. This is essentially the same as for the perfectly stoichiometric case although the lower formation energies give concentrations considerably larger, 15 and 7 orders of magnitude to be exact.

Comparing the P- and Ga-rich cases will give rise to even larger differences in native defect concentrations. This shows that the concentration and stability of the defects have a very sensitive dependence on the chemical potentials and that, for example, self-compensation can be avoided, or at least reduced, by a careful choice of the growth conditions. From Figs. 5.3 and 5.4 it is seen that the most common native defects, mentioned above, all have charge states such that they will compensate and counteract doping. In Ga-rich GaP, $GaP^{-2}$ will compensate attempts to make the material n-doped. In P-rich GaP, $V_{Ga}^{-3}$ will compensate attempts to make the material n-doped and $P_{Ga}^{+2}$ will compensate attempts to make the material p-doped.

Upon comparing the theoretical results to experiments it is seen that $P_{Ga}$ is indeed expected to be a common defect in GaP, especially in the +2 charge state and if grown under P-rich conditions. EPR and ODMR relies on the interaction of an unpaired spin with the surrounding crystal, which means that only defects in charge states with an odd number of electrons occupying the defect levels in the gap can be detected by these methods. Therefore, in the case of $P_{Ga}$ only the singly positive charge state is observable. In p-type GaP it was only seen after optical excitation by $1.25 \pm 0.10$ eV [31], while no excitation was needed in semi-insulating GaP. This is in accordance with the theoretical results presented here, since in semi-insulating GaP the singly positive charge state has a formation energy which is only 0.03 eV higher than the most stable charge states.

That $P_{i}$ has never been observed is simply explained by its large formation energy. That the gallium antisite has not been observed is probably due to that its most stable charge state $GaP^{-1}$ is not paramagnetic and detectable. $GaP^{-1}$, on the other hand is, and predicted from the results presented here it should be possible to detect it in p-type GaP grown under Ga-rich conditions with only little or no excitation.

In both positron annihilation experiments it was believed that the large vacancy concentrations were due to $V_{P}$. In one experiment this conclusion was drawn because of a comparison to a similar study in GaAs. In the other one because it was known that $V_{Ga}$ is in a negative charge state in the entire bandgap.
but since no positron trapping was observed in $p$-type and semi-insulating GaP it was incorrectly concluded that $V_{Ga}$ only occurred in negligible amounts. Their argument did not take the formation energy’s linear dependence on the Fermi-level into account. From Fig. 5.1 it is clear that $V_{Ga}$ is not common in $p$-type and semi-insulating GaP but instead very common in $n$-type. Since the formation energy of $V_p$ is larger than that of $V_{Ga}$ by more than 1 eV in $n$-type GaP, which corresponds to many orders of magnitude for the concentrations, the experimental results are reinterpreted as being mainly due to the large $V_{Ga}$ concentration. The observation of $V_{Ga}^{-2}$ (or $V_{Ga}^0$) after excitation strengthens this conclusion. (Both $V_{Ga}^{-2}$ and $V_{Ga}^0$ are in addition found to have structures very close to the perfectly tetragonal, with deviations of 1.28% and 1.23% from the ideal tetragonal distance between nearest-neighbors, respectively).

In conclusion, an extensive mapping of the stability and structure of the native defects in GaP has been provided over a wide range of physical parameters. The most important aspects of this study can be summarized as:

- The high vacancy concentration experimentally observed has been reinterpreted as being due to $V_{Ga}^{-3}$.
- $P_{Ga}$ was found to be one of the most common native defects in accordance with experiment. It was, for the first time, found to have a negative-U behavior.
- The native defect concentrations should be possible to control through the stoichiometry during growth, to avoid self-compensation, for example.
5.2 InP, InAs and InSb

InP, InAs and InSb are all direct band gap semiconductors. They share the same cation and have pentavalent anions of increasing size in the series $\text{P} \rightarrow \text{As} \rightarrow \text{Sb}$, which gives band gaps of 1.42 eV, 0.42 eV, and 0.24 eV, given in the same order. InP is of great technical importance due to its suiting properties for optoelectronics. InAs is used in quantum well structures and since InSb has the smallest band gap of the III-V semiconductors it is being used for mid-infrared optoelectronics applications.

The formation energies of all the native defects in InP, InAs and InSb are presented in Figs. 5.5, 5.6, and 5.7, respectively. The figures show the extreme stoichiometric limits of cation- and anion-rich material (the formation energies of the perfect stoichiometric materials are the linear average of these).

In In-rich InP (Fig. 5.5), the In interstitial, $\text{In}_i^+$, will be the most common native defect in $p$-type material and both types of vacancies, $V_p$ and $V_{\text{In}}$, will be the very common in $n$-type material. In P-rich InP, the anion antisite, $\text{P}_{\text{In}}^{\text{As}}$, will be the most common defect in $p$-type material and both vacancies are again most common in $n$-type. All these defects are most stable in charge states which make them contribute to self-compensation. The formation energies of these defects in $p$- or $n$-doped InP are also low enough that this contribution should be significant. Finally, it can be concluded that InP, regardless of the stoichiometric conditions, is intrinsically semi-insulating.

Figure 5.5: The most common defects in bulk InP.
Figure 5.6: The most common defects in bulk InAs.

Figure 5.7: The most common defects in bulk InSb.
In In-rich InAs (Fig. 5.6), the In interstitial, $\text{In}_{i:\text{As}}^{+3}$, is the most common native defect for all possible values of the Fermi-level. This means that $\text{In}_{i:\text{As}}^{+3}$ will act as a donor and in the absence of dopants it will donate electrons to the conduction band and make InAs intrinsically $n$-doped. In As-rich InAs, the In interstitial still has a formation energy low enough to be present in relevant concentrations under $p$-type conditions, but the anion antisite, $\text{As}_{\text{In}}$ has the overall lowest formation energy in the entire band gap. Since $\text{As}_{\text{In}}$ is neutral when the Fermi-level is close to the conduction band, it will not make As-rich InAs intrinsically $n$-type but it will be a compensator when attempting to $p$-dope the material.

In In-rich InSb (Fig. 5.7), the In interstitial, $\text{In}_{i:\text{In}}$, is again the most common native defect for all values of the Fermi-level, but unlike InP and InAs it is now most stable at the In surrounded tetragonal site. Like in InAs it will act as a donor and will make In-rich InSb intrinsically $n$-doped. In Sb-rich InSb, the anion antisite, $\text{Sb}_{\text{In}}$, is the most stable in the entire band gap. Since it is always lowest in energy for charge state +2 it will act as a donor and therefore InSb will become intrinsically $n$-doped independent of the stoichiometric conditions.
6. Supercell size effects

Defect calculations can be performed in either clusters or supercells. A cluster models an isolated point defect in a reasonable way. However, clusters do not accurately describe the properties of an infinite bulk material. Furthermore, the cluster must be as large as possible to avoid interactions between the defect and the surface of the cluster. Calculating a defect in a supercell, on the other hand, models the bulk properties in a reasonable way but instead introduces spurious interactions between the defect and its images in the neighboring supercells. The problem can be minimized through the use of as large a supercell as possible but, as will be seen, the errors from using even a 216 atom supercell can still be significant. As a reference, the defect concentrations to which different supercell sizes correspond are given in table 6.1. It is seen that in order to correctly model a strongly doped material, i.e. with a dopant concentration on the order of 10^{18} cm^{-3}, a supercell size of close to 20000 atoms would have to be used. However, it is important to note that the supercells give rise to an ordered array of defects. Even if it is desired to study higher defect concentrations, an ordered defect arrangement is highly unlikely in real materials.

6.1 Finite size scaling

There have been several corrections developed to remove the unwanted defect-defect interactions appearing in relatively small supercells. Examples are the Makov-Payne correction [27] and corrections for the defect level dispersion [39], of which the former is more commonly used. In Paper II it is shown that these are not reliable and that a finite scaling of the form

\[
e_{sc}^{form}(q, L) = e_{\infty}^{form}(q) + \frac{a_1}{L} + \frac{a_3}{L^3}
\]


<table>
<thead>
<tr>
<th>(L)</th>
<th>(6.4 \cdot 10^{21})</th>
<th>(7.9 \cdot 10^{20})</th>
<th>(2.4 \cdot 10^{20})</th>
<th>(9.9 \cdot 10^{19})</th>
<th>(5.1 \cdot 10^{19})</th>
<th>(2.9 \cdot 10^{19})</th>
<th>(2.9 \cdot 10^{18})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L)</td>
<td>8</td>
<td>64</td>
<td>216</td>
<td>512</td>
<td>1000</td>
<td>1728</td>
<td>17576</td>
</tr>
<tr>
<td>%</td>
<td>25.6%</td>
<td>3.2%</td>
<td>1.0%</td>
<td>0.4%</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

Table 6.1: The effective defect concentrations corresponding to different supercell sizes (for a concentration of possible lattice sites of 2.5 \cdot 10^{22} cm^{-3} and a lattice constant of 5.4\AA).
Figure 6.1: Finite size scaling of formation energies for the most common native defects in GaP. L is the length of the supercell side in units of the simple cubic 8 atom supercell side length, so that $\frac{1}{L} = 1$, $\frac{1}{L} = \frac{1}{2}$ and $\frac{1}{L} = \frac{1}{3}$ for the 8, 64 and 216 atom supercells respectively.
is a more accurate correction for the finite size effects. Here $L$ is the linear dimension of the supercell and $\varepsilon_{\text{form}}^\infty(q)$ the formation energy in the limit of an infinite supercell. The second term corresponds to factors depending on the defect-defect distance and the third on factors varying with the volume of the supercell (such as the jellium charge distribution, the charge density etc.). From Eq. 6.1 it is seen that in order to find the formation energy in the infinite limit at least three different supercell sizes must be calculated. The error from performing the scaling excluding the 512 atom supercell and only considering the 8, 64 and 216 atom supercell sizes is shown to be on the order of 0.01-0.1 eV. That is, although the small 8 atom supercell might even be metallic this scheme still fairly accurately predicts the formation energy of the 512 atom supercell.

In Paper II, it is found that the Makov-Payne correction does improve results for some defects but for others it might make them worse, which makes it an unreliable method for correction. On the other hand, what is shown to be surprisingly good is the realignment of the potentials of systems with and without defect. That is, ensuring that the potential in a bulk-like region within the defect supercell is the same as that of the defect free supercell can greatly improve results for a given supercell size. This appears to be the best correction for supercell size problems if it is not possible to calculate three different supercell sizes for the same system and perform the finite size scaling.

Figure 6.2: Formation energies and charge transfer levels of native defects in GaP, scaled to the limit of an infinite supercell size.
6.2 Native defects of GaP in the infinite limit

The extent to which finite size scaling changes the results of typical defect supercell calculations has been investigated for the native defects in GaP in Paper I. Fig. 6.1 displays the scaling behavior of the formation energies for the native defects in GaP (P_i is excluded since it has an increasing formation energy going from the 64 to the 216 supercell size making it even less common in the infinite size limit).

The formation energies extrapolated to the infinite limit have been used to derive the thermodynamic transfer levels as before and they are presented in Fig. 6.2. The scaling raises the formation energies somewhat but leaves the qualitative features essentially unchanged. The latter can be seen by comparing transfer levels in the infinite limit in Fig. 6.2 to those of the uncorrected 216 atom supercell in Fig. 5.1 of the previous chapter.

6.3 Defect level positions in the infinite limit

Calculating the donor and acceptor levels is often problematic since these levels are typically located 5-100 meV from either the conduction band minimum or the valence band maximum. As previously mentioned, typical dopant concentrations correspond to supercell sizes of 20000 atoms or more. When smaller supercells are used, the shallow acceptor or donor levels will be in resonance with either the valence band maximum or conduction band minimum. This is, however a physical effect; increasing
the dopant concentration will cause the defect states to overlap more and be less localized. None the less, if one wants to model reasonable doping concentrations this is undesired and the results must be scaled to the infinite limit. Interestingly, it has been found experimentally that the ionization energy of shallow dopants can vary with the dopant concentration [40, 41]. Whether this is due to the broadening of the defect band which follows from increasing the dopant concentration, or from an actual shift of the dopant level is unclear at this point.

The scaling behavior of typical dopants in InP have been studied in Paper II and are shown in Fig. 6.3. Experimentally, the Zn_{In} acceptor level is found to be 35 meV above the valence band maximum. It is seen from the figure that even for 512 atom supercells this level will incorrectly not be in the band gap. Therefore it is necessary to either scale the results to the infinite limit or to, at least, perform a potential alignment if only one supercell size is considered.

Donor levels are even more problematic than acceptor levels because of the LDA underestimation of the band gap. There has been an ongoing debate whether it is correct to plot the transfer levels up to the LDA value or the experimental value of the conduction band minimum. From Fig. 6.3 it is found that the donor levels in the infinite limit are close to the valence band minimum of LDA when LDA is used and to that of GGA if GGA is used (excluding one of the four data points gives the error bars to the left in the figure). It is therefore concluded that the theoretical band gap should be used for plotting the defect levels. By coincidence, the donor levels from the 64 atom supercells lie close to the experimental valence band minimum. That the misconception that it is correct to use the experimental band gap has been so long lived is possibly due to this coincidence.
7. Defects on III-V Semiconductor surfaces

Even in objects with a relatively large surface area, like this sheet of paper, the majority of atoms are sitting in a bulk like environment. When considering point defects, the three dimensional bulk has a much higher number of possible lattice sites as compared to the two dimensional surface. So unless the material is only a few atom layers thin, it is the bulk defects that will dictate the electric properties of the material. So why is it then interesting to study the surface properties of a material? First, the stability of defects and dopants can be rather different at the surface than in the bulk, which may result in significantly different electric properties of the surface, such as for example surface conduction with an insulating bulk or a 2D electron gas at the surface. Surface reconstructions, band bending etc. can also be the cause of this. Second, the surfaces are important when considering the interaction of the material with the surrounding. An important example is the growth of semiconductors. Since the growth of the material takes place at the surface, a type of the defect that is particularly stable at the surface might be grown in as a meta-stable configuration in the bulk. Finally, as already mentioned, in materials with a high surface area relative to the bulk volume, such as thin-films or nano structures, the surface properties will become of increasing importance.
A surface has qualitatively different properties depending on in which plane the crystal is cleaved. A semiconductor with a tetragonal bonding structure has three important types of surfaces (considering only low index surfaces). These are the (111), (110) and (100) surfaces, which are illustrated in Fig. 7.1. For binary compounds it is seen that the (111) and (100) surfaces will be either cation or anion terminated. This means that if the material is cleaved in order to create a surface, all the dangling bond orbitals will contain either 3/4 or 5/4 electrons. These surfaces are therefore electrically charged and are called polar surfaces. The (110) surface, on the other hand, contains an equal number of anions and cations and is therefore electrically neutral. Typical for III-V semiconductor surfaces is that both the electrons of the cut bonds (3/4+5/4=2) occupy the dangling orbital of the more electronegative anion. The anions will therefore maintain their sp$^3$-like symmetry but now with one fully occupied dangling orbital pointing out from the surface. The cations will instead have no electrons in the fourth orbital and it is therefore more energetically favorable for them to have a planar sp$^2$-configuration. This causes the row of anions to be further out from the surface than the row of cations, as illustrated in Fig. 7.2. We calculate this angle to be 27.5°, 30.5°, and 32.0° for InP, InAs, and InSb, respectively. These angles agree well with those measured by Low-Energy Electron Diffraction (LEED), which are found to be 29.5°, 31.0°, and 31.6°, given in the same order [43].

Because the (110) surface is non-polar, binary III-V and II-VI compound semiconductors which have zincblende crystal structure will naturally cleave along the 110 planes [43]. All the results presented in the following are for (110) surfaces.
Figure 7.3: The most common defects on the (110) InP surface.

Figure 7.4: The most common defects on the (110) InAs surface.
7.1 Simulated STM images from calculations

Scanning tunneling microscopy (STM) is an experimental method for analysis of material surfaces. A conducting tip is swept over the surface and a voltage is applied between the surface and the tip. The measured current will depend on the applied and the electronic structure of the surface at the position of the tip. This can then be used to produce an image of the surface. To a first approximation (the Tersoff-Hamann approach [44]) the tunneling current is proportional to the local density of states at the tip position, integrated in energy interval from the Fermi level \( \varepsilon_F \) up to the applied voltage \( eU \),

\[
I(U, \mathbf{r}_{tip}) \propto \int_{\varepsilon_F}^{\varepsilon_F + eU} n(E, \mathbf{r}_{tip}) dE. \tag{7.1}
\]

For a negative bias voltage electrons will tunnel from the surface to the tip. In this case the tunneling current will be proportional to the density of filled states. This therefore produces an image of the occupied states of the surface. For a positive bias voltage, empty states will be probed instead and this produces an image over the unoccupied states of the surface.

Since the charge density is equal to \( \rho(\mathbf{r}) = e \cdot n(\mathbf{r}) \), where \( e \) is the elementary charge, STM images can be simulated through the calculation of the partial charge density for all the states in the energy interval \([\varepsilon_F, \varepsilon_F + eU]\).
Table 7.1: The photoelectric threshold calculated from first-principles compared to experiment (Å).

<table>
<thead>
<tr>
<th></th>
<th>InP</th>
<th>InAs</th>
<th>InSb</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Phi )</td>
<td>5.69</td>
<td>5.27</td>
<td>4.95</td>
</tr>
<tr>
<td>( \Phi^{Exp.} )</td>
<td>5.69</td>
<td>5.31</td>
<td>4.77</td>
</tr>
</tbody>
</table>

7.2 Photoelectric threshold

The work function of metals is defined as the minimum energy required to remove an electron from the material to a point far outside the surface [45]. Because the Fermi-level separates filled and unfilled levels in metals, the work function can also be considered to be a metal’s electron affinity and ionization energy. In semiconductors, however, the Fermi-level will depend on the doping and, since the work function depends on the Fermi-level, it is more reasonable to use the photoelectric threshold as a characteristic material property in semiconductors. The photoelectric threshold is defined as

\[
\Phi = E^\infty - E_v,
\]

where \( E^\infty \) is the energy difference between the vacuum level and \( E_v \) is the valence band maximum. The photoelectric threshold is nothing else than the ionization energy of the semiconductor [43], but unlike in metals it will not be the same as electron affinity.\(^1\) The photoelectric threshold, as calculated in Paper III, are presented for InP, InAs and InSb and compared to experiment in Table 7.1. From the fairly close agreement it can be confirmed that the valence band maximum is well determined in LDA.

7.3 Native defects on InP, InAs, and InSb (110) surfaces

The possible native defects at III-V semiconductor surfaces are cation and anion vacancies, interstitials and antisites. Taking the two qualitatively different interstitial sites into account, this makes a total of eight defects to consider. In the same way as for bulk defects, the formation energies of these surface defects determine whether or not they will occur in concentrations large enough to influence the material properties. The formation energies are presented as a function of the Fermi level in Figs. 7.3, 7.4, and 7.5. For all three materials the In interstitial (In\(_{11}^+\)) and the anion antisites (P\(_{In}^0\), As\(_{In}^0\), and Sb\(_{In}^0\)) are found the be the types of native defects that have the highest concentration.

\(^1\)The electron affinity of semiconductors is defined as \( \chi = E^\infty - E_c \), where \( E_c \) is the conduction band minimum.
Figure 7.6: The relaxed geometry of the As antisite, $\text{As}_{\text{In}}^0$, on the (110) InAs surface.

at thermodynamic equilibrium. Since the antisites are neutral, they will have an equally high concentration, independent of if the semiconductor is $p$-type, semi-insulating or $n$-type. The singly positive In interstitial, on the other hand, is expected to be more common in $p$-doped semiconductors than in $n$-doped.

The stable geometry of the As antisite, $\text{As}_{\text{In}}^0$, is shown in Fig. 7.6. The geometries of $\text{P}_{\text{In}}^0$ and $\text{Sb}_{\text{In}}^0$ are qualitatively identical. As expected, the anion antisite atom will not have planar sp$^2$-like configuration like the surface cations, but instead a tetragonal sp$^3$ configuration with one fully occupied dangling bond pointing out from the surface in the neutral state. In this way the antisite atom is no different from the other anion atoms at the surface and it will also be located at a similar distance from the bulk of the material as the other surface anions. This is unlike anion antisites in the bulk which typically are most common in the charge state that makes them isovalent to the cation atom they are substituting, i.e. they are most stable in charge state +2. That the neutral charge state of the surface anion antisites are the most stable means that the condition of charge neutrality will not be a restriction and the entire (110) surface could very well be covered with anion antisites. This would be synonymous to an anion terminated (110) surface. Under anion-rich conditions, the formation energies of $\text{P}_{\text{In}}^0$, $\text{As}_{\text{In}}^0$ and $\text{Sb}_{\text{In}}^0$ are all close to zero which means that, in this case, the anion terminated (110) surface is just as stable as the ideal surface. Although the anion antisite is independent of the Fermi-level, like all antisite it depends strongly on the stoichiometry of the system, such that it will have a relatively large concentration for anion rich conditions.

The most stable position for the In interstitial, $\text{In}_{\text{In}}^{+1}$ on the InAs (110) surface is shown in Fig. 7.7. The same position is also the most stable one on the InP and InSb surfaces. In the +1 charge state the In interstitial atom has two valence electrons. It will bond to two surface anions and one surface cation, as shown in the figure. The four electrons of the two anion dangling bonds plus the two valence electrons complete these three bonds, and the fourth bond is
an empty dangling bond pointing out of the surface. Further, the In interstitial will naturally be favored by In rich stoichiometry but as seen from Figs. 7.3, 7.4, and 7.5, it will still be very common in $p$-type material under In poor conditions since a formation energy of less than 0.5 eV corresponds to a very high defect concentration. Comparing the three materials, it is found that the formation energy of In$_{i1}^+$ is the lowest in energy in InP, intermediate in InAs and the highest in InSb. A closer analysis shows that the In-In bonding distance is the reason for this trend. In InSb the difference between the In-In distance and twice the covalent radius of In is 0.24 Å whereas it is only 0.15 Å in InP, which leads to a lower formation energy in InP.

In addition to the anion antisite and the cation interstitial, both type of vacancies are expected to be common in strongly $n$-type InP. V$_{In}^-$ will have the highest concentration under P-rich conditions, and V$_P^-$ the highest concentration under In-conditions (see Section 7.5 for further details). In InP with a stoichiometric composition both types of vacancies are therefore equally common.

If surface conductance is important for an application, it should be noted that $p$-type doping of these surfaces can be strongly self-compensated by In$_{i1}^+$. That is, if the technique for creating the surface allows for In interstitials to form (not cleaving in vacuum, for example), the low formation energy of In$_{i1}^+$ will result in such a high interstitial concentration that it is impossible to make the surface $p$-doped. Similarly, the vacancies at the (110) InP surface should make $n$-doping of the InP surface very difficult. Unlike the In interstitials, V$_P^-$ and V$_{In}^-$ can easily form thermally and therefore there is no way of avoiding them by a suit ing technique of creating the surface.
7.4 Relative stability of defects at the surface and in the bulk

For most materials the bulk contains a much larger number of point defects than the surface, simply because there are more possible sites for the defects in the bulk. However, this comparison is problematic since we’re comparing a concentration per unit volume to a concentration per unit area. It is therefore more meaningful to consider the fraction of defect sites among all the possible sites. Using Eq. 4.9 we can, for a particular defect type, write this fraction as

\[ k = \frac{n}{N} = e^{-\frac{\epsilon_{\text{form}}}{k_B T}} \cdot e^{\frac{S_{\text{form}}}{k_B}}, \]  

(7.3)

where \( n \) is the number of defect sites, \( N \) the number of possible defect sites in the lattice, \( \epsilon_{\text{form}} \) and \( S_{\text{form}} \) the formation energy and entropy of the defect, respectively. Cleaving the semiconductor under consideration gives an surface fraction of defect sites which is equal to that of bulk. That is, \( k_{\text{surf}}/k_{\text{bulk}} = 1 \) immediately after the cleavage if we denote the surface and bulk fraction by \( k_{\text{surf}} \) and \( k_{\text{bulk}} \), respectively. What is relevant is then how this ratio changes when the system goes to thermodynamic equilibrium. The difference in formation entropy between bulk and surface positions of a defect is negligible in comparison the difference in formation energy.\(^2\) Therefore the ratio \( k_{\text{surf}}/k_{\text{bulk}} \) can be written as

\[ \frac{k_{\text{surf}}}{k_{\text{bulk}}} = e^{\frac{\epsilon_{\text{form}}_{\text{bulk}} - \epsilon_{\text{form}}_{\text{surf}}}{k_B T}}. \]  

(7.4)

\(^2\)The largest contribution to the formation entropy is the vibration entropy [46, 47], which to good approximation has been shown to be that same at the surface and in the bulk [48, 49].
Figure 7.9: Comparison of the formation energy of $V_{In}$ to that of the more stable $[P_{In}^\cdot V_{P}]$ complex on the (110) InP surface.

If $k_{Surf}/k_{Bulk}$ is smaller than one for a certain type of defect at thermodynamic equilibrium, that type of defect will have a tendency to migrate to the surface of the material. We find that the ratio $k_{Surf}/k_{Bulk}$ will be less than one for most of the considered defects. In average the native point defects are roughly 0.9 eV lower in energy at the surface than in the bulk of the material. This means that defects are, in general, more favorable at the surface and that an ideal (110) surface should be much more rare than an ideal bulk crystal. For a given native defect it means that it would be more stable at the surface and, migration barriers aside, would tend to migrate to the surface. The reason for this is that defects in the bulk cause strain in the surrounding crystal, at the surface there is no restriction in the direction perpendicular to the surface and bonds can adjust fully to remove residual strain.
Figure 7.10: Comparison of the formation energy of $V_{\text{In}}$ to that of the more stable $[\text{As}_{\text{In}}-V_{\text{As}}]$ complex on the (110) InAs surface.

Figure 7.11: Comparison of the formation energy of $V_{\text{In}}$ to that of the more stable $[\text{Sb}_{\text{In}}-V_{\text{Sb}}]$ complex on the (110) InSb surface.
Figure 7.12: Calculated STM image for $[\text{As}_\text{In}-\text{V}_{\text{As}}]$ complex on (110) InAs surface corresponding to a bias voltage of -2.0 eV. A negative voltage probes filled occupied states and the spots are the filled dangling bonds of the surface anions.

7.5 Reconfiguration of V$_\text{In}$

In Paper VII, it is shown that the cation vacancy at (110) InP, InAs, and InSb surfaces, is not stable to the formation of an anion antisite-anion vacancy complex. That is, the complex is lower in energy than V$_\text{In}$ and since they both have the same atomic configuration the V$_\text{In}$-geometry can relax to that of the complex. The structure of this complex is shown in Fig. 7.8, where it is seen that a neighboring anion in the first layer has moved in to the cation vacancy such that an anion antisite and an anion vacancy are created. This causes a disturbance also in the second surface layer and the overall structure will be asymmetric. In InP, there is no energy barrier between the V$_\text{In}$ structure and the [P$_\text{In}$-V$_\text{P}$] structure and therefore cation vacancies will spontaneously rearrange as complexes, even at a temperature of 0 K. Since the barrier is zero at InP surfaces, it is also expected to be zero or very small at InAs and InSb surfaces (and other III-V semiconductors), such that there will be no cation vacancies at the (110) surfaces. From the physical understanding of this complex, as discussed in Paper VII, there is no reason to believe that this would not be general to all III-V semiconductors.

The formation energies of the cation vacancy and the complex is presented in Figs. 7.9, 7.10, and 7.11 for InP, InAs and InSb, respectively. It is seen that the complex will be lower in energy under all conditions, from strongly $p$-type material to strongly $n$-type material, and also under extreme off-stoichiometric conditions such as cation- and anion-rich material. Since the number and type of atoms are the same in both systems, the stability of the complex obviously does not depend on the stoichiometry (chemical potentials). This is seen from the constant formation energy difference between the complex and V$_\text{In}$ in the figures.

Calculating the STM images, as described previously, for the complex gives very similar results for the three materials considered. The calculated STM image of the singly negative complex at the InAs surface, $[\text{As}_\text{In}-\text{V}_{\text{As}}]^{-1}$, is
shown in Fig. 7.12. The anion antisite is located deeper in the surface than the other surface anions, and therefore these states do not contribute much to the conductivity. Most likely, experiments are not sensitive enough to make out the contribution from the antisite atom and therefore the STM appearance of the complex is very similar to that of the anion vacancy. One important difference is, however, the asymmetry originating from that one of the anion dangling bonds next to the “vacancy” (the one forming a dimer with the antisite) has reduced strength in the STM image. Under close inspection it should therefore, at least in principle, be possible to tell the anion vacancy and the complex apart.

Comparing the formation energy of the complex to that of the anion vacancy (shown in grey in Figs. 7.9, 7.10, and 7.11), it is seen that the complex is more common under anion-rich conditions (or equally common in strongly p-type InP and InAs) and they should also be equally common in n-type material under stoichiometric conditions. Therefore, what has been concluded to be anion vacancies on III-V (110) surfaces experimentally in the past, are under anion-rich conditions and in n-type materials very likely to be antisite-anion vacancy complexes.
7.6 Charge accumulation on InAs surfaces

![Figure 7.13](image)

Figure 7.13: Stable geometries for hydrogen at the InAs (110) surface. $H^{+1}$ bonded to an anion shown to the left and $H^{-1}$ bonded to a cation shown to the right.

Unlike in other III-V semiconductors, InAs has been found to have surface charge accumulation even for cleaned surfaces. InAs is the middle compound in the series InP, InAs, and InSb and while most properties progress monotonically along this series, neither InP nor InSb from charge accumulation layers so the formation on InAs has been hard to explain. In Paper VI we attempt to explain this by two complementary mechanisms. The first is that InAs is intrinsically doped by cation ad-atoms, $In^{+}$, acting as donors. This would be expected to be the case on MBE and irradiation-annealed cleaned surfaces but not on cleaved (110) surfaces. The second is due to adsorption of residual hydrogen. Hydrogen is almost always present in practice as a completely hydrogen free experimental environment is very difficult to achieve. Hydrogen is stable on the surface either in a -1 charge state, bonded to the empty dangling bond of the surface cation, or in a +1 charge state, bonded to a surface anion at the filled dangling bond (see Fig. 7.13). STM measurements of occupied states on III-V (110) surfaces show mainly the filled dangling bonds of the anions. When a hydrogen bonds to an anion this lowers the energy of the dangling bond making it as low in energy as the bulk bonds. Therefore it no longer contributes to the tunneling current and the dangling bond disappears from the STM image. A missing filled dangling bond looks like an anion vacancy and the calculated STM images of anion vacancies and $H^{+1}$ are very hard to tell apart.

Whether the hydrogen atoms bond to anions or cations will depend on the doping conditions and the Fermi-level. The $\epsilon(\pm1)$ charge transfer level of hydrogen is shown at the (110) surfaces of InP, InAs, InSb, and GaAs in Fig. 7.14. The formation energy of hydrogen is lower than those of all native defects (the chemical potential used in Fig. 7.14 is half that of molecular hydrogen, at a realistic hydrogen partial pressure for vacuum chambers and at 300 K), and therefore the Fermi-level will be determined by hydrogen and pinned at the $\epsilon(\pm1)$ charge transfer level. This level lies inside the band gap for InP, InSb, and GaAs, but above the conduction band minimum for InAs. $H^{+1}$ will therefore be a donor in InAs, making the surface $n$-type. The surface
Figure 7.14: $\epsilon$ (+|-) transfer level of hydrogen on III-V surfaces. Plotted up the LDA surface band gap.

charge accumulation on InAs can hence be explained as unintentional doping by donor type defects.
8. CuIn$_{1-x}$Ga$_x$Se$_2$ semiconductor solar cells

Semiconductor solar cells are basically pn-junctions, as described in Chapter 4. The two vital properties in turning light into electric energy are the absorption of light (over the band gap in semiconductors) and the separation of charge carriers. Somewhat simplified, semiconductor solar cells are light emitting diodes working reversely. Today, silicon based solar cell devices dominate the market. The highest efficiency has, however, so far been achieved using GaAs ($\sim$30%). New materials are under development with one of the most promising being CuIn$_{1-x}$Ga$_x$Se$_2$ (CIGS), because of its low cost yet high efficiency (an efficiency of close to 20% has been achieved in a research environment). CIGS is a complex material which is grown as thin layer films and, unlike other solar cell materials, is more efficient in a polycrystalline form than in a mono-crystalline. CIGS also has a large structural tolerance to off-stoichiometry and it contains large numbers of native defects. It has been found that CuInSe$_2$ (CIS) can be made intrinsically doped by native defects through tuning of the stoichiometric conditions [50]. Cu-poor conditions tend to make CIS $p$-type and Cu-rich tends to make it $n$-type. CuGaSe$_2$ (CGS), on the other hand, has unfortunately proven to be difficult to $n$-dope,
even extrinsically [51]. CIGS is, however, not used for \(pn\)-homojunctions but are instead used as absorber material in a more complex ZnO/CdS/CIGS structure. The crystal structure of CIS (and CGS) is a chalcopyrite structure with 50-50 distribution of Cu and In (or Ga) on the cation sub-lattice, as shown in Fig. 8.1.

The optimal band gap regarding the efficiency of (singular junction) solar cell materials lies close to 1.4 eV [52], which is why higher efficiency has been achieved for GaAs (band gap of 1.42 eV) than in Si (band gap of 1.12 eV). Because of this, the maximum efficiency for \(\text{CuIn}_{1-x}\text{Ga}_x\text{Se}_2\) devices would be expected at \(x=0.6\). (Ga also costs a third or less of the price for In, making devices with a relatively large Ga content more cost efficient, which is central to the production of solar cell materials). Unfortunately, this expected increase in efficiency with increasing Ga content has not been observed and there seem to be no increase in efficiency for \(x \geq 0.3\). In Paper VIII we attempt to find an explanation for this unexpected behavior.

CIGS is always found to have high concentrations of native defects. Especially copper vacancies have a remarkably low formation energy. The copper vacancies, \(V_{\text{Cu}}\), can form neutral complexes with In (or Ga) antisites, \(\text{In}^{+2} \text{Cu}(\text{Ga}^{+2} \text{Cu})\), \([2V_{\text{Cu}} + \text{III}_{\text{Cu}}]^0\). It has been suggested that these complexes have concentrations as high as up to 10% and that they can form an ordered defect phase, the \(\gamma\)-phase, which has composition \(\text{Cu}_1\text{In}_5\text{Se}_8\) [50]. We have, us-
Figure 8.3: DOS near the band edges of a 216 site supercell $\alpha$-CIS without (upper panel) and with (lower panel) one $[2V_{\text{Cu}}^- \text{Cu}^+ \text{II}]_0$ defect complex.

Figure 8.4: Charge density difference $\rho_{\text{geo}+0}(\mathbf{r}) - \rho_{\text{geo}+0}^0(\mathbf{r})$ of (a) $[2V_{\text{Cu}}^- \text{Ga}_{\text{Cu}}^-]$ in $\alpha$-CGS, and (b) for 448 atom supercell of $\gamma$-CGS. Due to the small relaxation of the defect site $\rho_{\text{geo}+0}^0(\mathbf{r}) - \rho_{\text{geo}+0}^0(\mathbf{r}) \approx \rho_{\text{geo}-2}^0(\mathbf{r}) - \rho_{\text{geo}-2}^0(\mathbf{r})$, i.e. electrons are excited to the same orbitals at the defect/conduction-band as they are removed from. Cu atoms are shown in yellow, Ga in green and Se in purple.
ing LDA+U with the $\alpha$-phase as reference, studied the band gap of this $\gamma$-phase over the entire range of $x$. The band gap of CIGS is shown in Fig. 8.2 for the $\alpha$-phase (full lines) and the $\gamma$-phase (broken lines). For $\alpha$-CIGS the band gap can be described by an linear expression, $E_{\Gamma}^\alpha = (1 - x)0.96 + x1.68$, whereas for the $\gamma$-CIGS a bowing parameter has to be taken into account, $E_{\Gamma}^\gamma = (1 - x)0.91 + x1.20 + x(1 - x)0.22$.

The defect level of $[2\text{V}_{\text{Cu}}^{+3}\text{Cu}^{0}]$ has been determined both through calculations of the transfer levels and through density of states (DOS) calculations. From Fig. 8.2 it is seen that the $\varepsilon(0|-2)$ transfer level and the defect DOS-peak (as in Fig. 8.3) coincide fairly well. The difference in geometry of the complex is very small for different charge states (0, -1, and -2) corresponding to less than 0.06 eV. Therefore, the thermodynamic transfer levels will coincide with the optical transfer levels. Further, there is no apparent repulsion between the two electrons on the defect level, resulting in that $\varepsilon(0|-) = \varepsilon(0|-2)$. The defect level of an isolated complex will have no dispersion of the defect level since the defect electrons are localized on the defect. This has been controlled by calculating the difference in charge density between the neutral and the doubly negative complex, which shows the orbitals to which electrons are added (or removed) as seen in Fig. 8.4. Increasing the concentration of $[2\text{V}_{\text{Cu}}^{+3}\text{Cu}^{0}]$ will broaden the defect level and the limit of the the $\gamma$-phase, the wide defect band of the complex will constitute the lower part of the conduction band. This can be seen from Fig. 8.2, where the conduction band minimum of $\gamma$-CIGS has the same variation with $x$ as the defect level.
9. Diffusion

The diffusion of defects and the ionic conductivity in a material are of great importance for many applications, such as doping of semiconductors, fuel cells and more. The mobility of impurities can be quite problematic in semiconductor devices if, for example, p-dopants diffuse into the n-type region of the material. The driving force of atomic conduction can either be an external electric field or concentration gradients. Since we will be concerned with the problematic and unwanted mobility of dopants, only the latter will be discussed from here on. The redistribution of defects due to gradient in the defect concentration has been formulated in the phenomenological relation called Fick’s law [9]

\[ J = -D \nabla [C], \] (9.1)

where \( J \) is the rate of ions flowing through a unit area per unit time, \([C]\) is the concentration of the defect in question, and \(D\) is the diffusion constant or diffusivity (in units of \(\text{cm}^2/\text{s}\)). The minus sign simply means that defects will diffuse away from regions of high concentration (Being strict, it is more correct to talk about a gradient of the chemical potential for the defects). Because of the thermal vibration of the ions on their lattice sites, the mobility of diffusing atoms will be higher at elevated temperatures, which is reflected in the temperature dependence of the diffusion constant

\[ D = D_0 e^{-\varepsilon_a/k_B T}, \] (9.2)

where \(T\) is the temperature, \(k_B\) is Boltzmann’s constant, and \(\varepsilon_a\) is the activation energy for the process. \(\varepsilon_a\) is related to the energy that is required to move the diffusing atom through the lattice.

A defect can move in a material by many different mechanisms. There are three categories of these diffusion mechanisms: interchange diffusion where two or more atoms switch places through rotation, substitutional-interstitial diffusion where interstitial atoms are the diffusing species, and vacancy mechanism where the diffusing defect move into the vacancy such that the vacancy becomes the moving species. Substitutional-interstitial diffusion is normally fast and has activation energies of around 2 eV, whereas vacancies diffuse slower and have activation energies in the range of 3 to 5 eV [52].

Since experiments only provide a macroscopic view of defect diffusion, first-principles calculations are an important tool in determining the microscopic process of the diffusion. There are several processes in the three categories of diffusion mechanisms above. Different interchange mechanisms are,
for example, two atoms switching places, or four moving in a ring. Vacancy diffusion naturally depends on the concentration of vacancies in the material. In substitutional-interstitial diffusion, the diffusing atom can either become interstitial by leaving a vacancy behind (the dissociative mechanism) or by being replaced at the substitutional site by a native interstitial atom (the kick-out mechanism). The rate limiting factor for these will therefore be either the concentration of vacancies or the concentration of native interstitials. In the computational approach, many processes must be considered in detail and the concentration of native defects must be taken into account. From the total energies of the defect system the activation energy can be calculated as follows

\[ \varepsilon_a = \varepsilon_{form} + \varepsilon_m, \]  

where \( \varepsilon_{form} \) is the formation energy of the diffusing defect, as derived in a previous chapter, and \( \varepsilon_m \) is the migration barrier of this defect when going from site to site. The migration barrier can accurately be calculated using the nudged elastic band method (NEBM) [54], in which the diffusion path between the initial and end point is divided into a number of intermediate points. The total energy is then calculated at all these intermediate points, allowing the diffusing atom and all the surrounding atoms to relax fully with the restriction that the diffusing atom only can move in a hyper-plane such that it does not move along the diffusion path. The migration barrier, \( \varepsilon_m \), is simply the energy required to overcome the highest point along this total energy curve. It is thereby possible to calculate the activation energy from first-principles. The diffusion process with the lowest activation energy will be the dominating one and it will account for what is experimentally observed on a larger scale. The activation energies can typically be accurately determined from calculations.

### 9.1 Diffusion of Zn in InP and GaP

Zn is one of the most common \( p \)-dopants in III-V semiconductors. Experimentally, Zn diffusion is most often studied by putting a semiconductor in contact with an external Zn source under elevated temperatures. The diffusion profiles are then matched to parametric models of the diffusion mechanism. Because there is certain arbitrariness in this parametric fit, especially regarding the charge states of the involved defects, the experimental literature is often contradictory about the exact diffusion mechanism. Further, the in-diffusion from an external Zn source is not the ideal way of studying the mobility of dopants in semiconductors with respect to device reliability, degradation and shift of the \( pn \)-junction position. This unintentional doping is better studied using a highly Zn doped semiconductor as a source, since the mechanism making Zn mobile (kick-out or dissociative for example) is not likely to be the same in an external Zn source. First-principles calculations therefore have an important
Figure 9.1: Formation energies of Zn in GaP (top) and InP (bottom). ‘i:X’ stands for the tetragonal interstitial position, surrounded by X atoms. The red line gives the activation energy of the kick-out mechanism, under the assumption of a constant kick-out barrier for the different charge states. The blue line is a lower bound for the activation energy of the dissociative mechanism. (Zn_{i:In} is only included to show the migration barrier dependence on the Fermi-level.)
role in finding the active mechanism behind diffusion. In Paper V we have studied the Zn diffusion in InP and GaP.

It is fairly well accepted that Zn diffuses by a substitutional-interstitial mechanism in InP and GaP [55]. Parametric fits to the diffusion profiles have long been done using the dissociative mechanism, although more recent studies are in support of the kick-out mechanism. According to Eq. (9.3), a lower bound of the activation energy is the maximum formation energy of the diffusing species along the diffusion path. Therefore, the formation energy of cation interstitials is of central importance to the kick-out mechanism and, similarly, the formation energy of cation vacancies is central to the dissociative mechanism. From figs. 5.1 and 5.5, the kick-out mechanism seems plausible in the \textit{p}-type materials and the dissociative mechanism seems plausible in \textit{n}-type. A exact treatment will, however, need to take Zn into account. The formation energies of Zn in GaP and InP are presented in Fig. 9.1. Substitutional Zn will be most stable on the cation site in both materials, which is expected since Zn is a \textit{p}-dopant. Its shallow $\varepsilon(0|-)$ acceptor level is found 0.02 eV above the valence band maximum in InP, and 0.12 eV above the valence band maximum in GaP. This is in good agreement with the experimental data of 0.04 eV and 0.07 eV, respectively [56]. Interstitial Zn is most stable at the P surrounded tetragonal interstitial site. In InP the Ga surrounded position is, in fact, not even meta-stable. The charge state of the Zn interstitial will be +2 in \textit{p}-type material. Comparing InP to GaP, it is found that Zn is more stable in InP, both in substitutional and interstitial form. This suggest that Zn would tend to accumulate in regions of InP in GaP-InP heterojunctions. Furthermore, the formation energy of interstitial Zn is almost as low as that of substitutional Zn in \textit{p}-type InP. This can explain why, for large Zn concentrations, a relatively small fraction of the Zn is found on substitutional sites in InP, unlike in all other III-V semiconductors [55].

Since Zn is an acceptor in III-V semiconductors and we are mainly concerned with the out-diffusion of Zn, we will start by looking at Zn diffusion in \textit{p}-type material. In the dissociative mechanism, Zn becomes mobile by leaving a vacancy behind, Zn$_{\text{III}}$+$\text{V}_{\text{III}}$ → Zn$_{\text{II}}$+$+$V$_{\text{III}}$. The lower bound of the activation energy for this mechanism is therefore the sum of the formation energies for interstitial Zn and the cation vacancy (that is, neglecting the migration barrier). This lower bound is shown in Fig. 9.1 by the dashed blue line, and in \textit{p}-doped material the activation energy of the dissociative mechanism will be larger than 4.7 eV in GaP and InP.

In the kick-out mechanism, Zn instead becomes interstitial by being replaced by a cation interstitial, Zn$_{\text{III}}$+$+$$\text{III}$ → Zn$_{\text{II}}$. The negative Zn substitutional and the positive cation interstitial will bind and form a complex, [Zn$_{\text{III}}$+$+$$\text{III}$]. The binding energy of this complex is calculated to be 0.27 eV in InP and 0.48 eV in GaP, and the charge state is found to be +2 in \textit{p}-type material. The energy barriers can be calculated by displacing the atoms according to the diffusion mechanism, and using the NEBM. The diffusion path for both the
Figure 9.2: Diffusion path of interstitial Zn in $p$-type GaP (top) and InP (bottom).
First, Zn\textsubscript{III} is kicked-out of the substitutional site with a barrier of 0.39 eV in InP and 0.62 eV in GaP. Second, Zn\textsuperscript{III+2} moves from interstitial site to interstitial site with a migration barrier of 0.98 eV in InP and 0.61 eV in GaP. The activation energy will be given by the highest energy along the diffusion path and is 1.60 eV in InP and 2.49 eV in GaP. Since these values are significantly smaller than for the dissociative mechanism it is concluded that Zn will diffuse by the kick-out mechanism in the \textit{p}-type materials.

By considering different charge states for the [Zn\textsubscript{III+III\textsubscript{1}}] complex and assuming that the migration barrier is constant for the different charge states, the dependence of the kick-out activation energy on the Fermi-level can be calculated. The result is shown as the red curve in Fig. 9.1. Comparing this to the lower bound of the dissociative activation energy, it is seen that the kick-out mechanism is the energetically favorable one also under semi-insulating conditions in both materials. The \textit{n}-type materials are a bit more problematic because of the LDA band gap problem. In \textit{n}-type InP the kick-out mechanism is likely since the activation energy of the dissociative mechanism will lower only under extreme conditions (an experimental band gap and a migration barrier of less than 0.5 eV). In \textit{n}-type GaP the dissociative mechanism can, however, not be excluded.

In \textit{p}-type InP and GaP, Zn will remain in the interstitial form once kicked-out (see Fig. 9.2). A cation vacancy could cause it to return to the substitutional form but these are rare in the \textit{p}-doped materials (Figs. 5.1 and 5.5). They are, however, very common in the \textit{n}-type materials and Zn would therefore be expected to become mobile in \textit{p}-type material, through the kick-out mechanism, and diffuse until it reaches an \textit{n}-type region where it would go back to the substitutional form through the dissociative mechanism. In \textit{pn}-homojunctions, Zn would therefore be expected to diffuse out from the \textit{p}-type region and into the \textit{n}-type. This is a plausible explanation to the Zn accumulation observed at \textit{pn}-homojunctions [57, 58, 59].
10. Dopants in alloys and multilayers

So far we have looked at point defects in one semiconductor at a time. In real components, however, more complex compounds, such as $A_xB_{1-x}$, are often used to be able to control properties such as the band gap and the lattice match to a substrate by varying $x$. In this section we will look at in what way the equilibrium concentration of native defects and the equilibrium solubility of dopants is affected by the variation of $x$. (From here on we will only discuss in terms of the dopant solubility although the results naturally extends to all substitutional defects).

In Papers IV and IX it is shown that the central parameter that determines the dopant solubility is the local strain of the material. As previously, the equilibrium dopant solubility is calculated through the formation energy. It is found that the formation energies of disordered alloys, ordered compounds and multilayer structures can be explained and predicted by the formation energy of the pure elements of these systems under external strain. For example, the formation energy of a dopant in a multilayer is very close to that of the pure material which makes up this particular multilayer, calculated at the same lattice constant that is found locally in the multilayer. This is found to be no different for ordered and disordered structures, for the studied systems of Zn and Cd dopants in Ga$_x$In$_{1-x}$P. Therefore, the solubility of a dopant can be predicted in any complex semiconductor compound if only the local strain of the compound is known.1

Anything that changes the local strain of a material will therefore also change the equilibrium solubility of dopants. That is everything from applying external strain to forming disordered alloys, ordered compounds, multilayers or growing materials epitaxially, lattice matched to a substrate (in the last case the Fermi-level may come to influence charged dopants in a yet unknown way). But what then determines the formation energy’s dependence on strain? Substituting a host atom for a dopant of equal size and with the same equilibrium bond lengths as the host atom, there is no dependence at all of the formation energy on the strain. This is because, no matter the strain, the bonds to the dopant atom will be exactly as far from the equilibrium bond length as the host atom it is substituting. For a relatively small dopant, the bonds to the dopant atom will be closer to the equilibrium bond length than those of the substituted host atom under compressive strain. This gives a relatively low formation energy. Under expansive strain, the bonds to the dopant atom will

---

1 The local environment is mainly the dopant and its nearest-neighbors, in this case.
instead be further from equilibrium bond length which gives a relatively large formation energy. Therefore, the formation energy of a relatively small dopant atom has a positive slope as a function of strain, as schematically illustrated in Fig. 10.1(b).

Following the same argument, the formation energy’s dependence on the strain of a relatively large dopant will have a negative slope, as shown in Fig. 10.1(c). Therefore, relatively small dopants will have an increasing solubility with increasing compressive strain. Similarly, relatively large dopants will have an increasing solubility with increasing expansive strain. This argument is synonymous to that of strain relief; smaller dopants will release the strain in a compressed material and larger dopants will release the stress in expanded material, and these substitutions are therefore favored by a low formation energy.

The properties of a compound $A_xB_{1-x}$ normally increase (or decrease) monotonically with $x$. That is, properties such as the band gap or the lattice constant will if not linearly then at least steadily change their value from that of the pure material A to that of the pure material B. In Papers IV and IX we show that this is not the case for the formation energy of substitutional defects and dopants. If there is a local strain of the material such that the formation energy is lowered, it is fully possible that this value is even lower than the lowest of the pure materials A and B. The equilibrium solubility of a dopant can hence be higher in $A_xB_{1-x}$ for $x \neq 1, 0$ than in either A or B. For example, it is found that the equilibrium solubility of the acceptor Zn$_{II}$ in a Ga$_x$In$_{1-x}$P multilayer structure with a relative layer thickness corresponding to $x=0.9$ is five orders of magnitude higher than in the highest of InP and GaP. The formation energy’s dependence on strain is shown in Fig. 10.2 for the C$_{IV}$ dopant in multilayered Si$_x$Ge$_{1-x}$ and for the Zn$_{III}$ and Cd$_{II}$ in disordered, CuPt-ordered and multilayered Ga$_x$In$_{1-x}$P.

Since it is now understood how Zn$_{III}$ can have a much greater solubility in Ga$_x$In$_{1-x}$P than in either GaP or InP, conditions can now be formulated for increasing the doping solubility in general. The equilibrium solubility of a dopant in a material A can be raised by strategic alloying or formation of multilayers as follows:
Figure 10.2: Correlation of formation energy of $C_0^{IV}$, $Zn_{III}$ and $Cd_{III}$ and the local lattice constant (left) or the average bond distance (right). Dotted curves show the formation energy dependence of the pure materials. Following the analytical treatment of Ref. [60] the local lattice constants of the multilayers are calculated in the entire range $x = 0 \rightarrow 1$ and the ranges they span are indicated by full lines. Values from multilayers, ordered compounds and disordered alloys are denoted by ‘×’, ‘◦’, and ‘⊿’, respectively.
if the dopant atoms are smaller in size than the host atoms A, one should alloy with a material B which has a smaller lattice constant than material A, $a_0^A > a_0^B$, and a formation energy larger than that of material A, i.e. $\varepsilon_{\text{form}}^A < \varepsilon_{\text{form}}^B$.

if the dopant atoms are larger in size than the host atoms A, one should alloy with a material B which has a larger lattice constant than material A, $a_0^A < a_0^B$, and a formation energy larger than that of material A, i.e. $\varepsilon_{\text{form}}^A < \varepsilon_{\text{form}}^B$.

The condition for the formation energy is necessary since the dopants will occupy the sites where they are the most stable.

Another situation where the lattice constant is altered is close to interfaces. The lattice matching at the interface will cause the one of the materials to experience compressive strain and the other expansive strain, much in the same way as for multilayers. At an increasing distance to the surface the lattice constant will converge to that of the pure material on each side, either through a continuous volume expansion or through a dislocation. In either case, there will be regions under strain close to the interface in which, if the conditions above are fulfilled, dopants and substitutional defects are expected to be the most stable. Under those circumstances there would be a high concentration of the dopant or substitutional defect in question close to the interface. This is a possible explanation for the so often observed accumulation of dopants and defects at interfaces. (For discussion of experimental findings see Papers IV and IX).
11. The future generation of semiconductor materials

Almost all of today’s electronics are semiconductor based. Early devices were based on the elemental semiconductors Si and Ge. Ge was the major semiconductor material in the early 1950s but soon Si came to dominate the scene because of its superior behavior at room temperature, lower price and its pure oxide [52]. Still today almost all electronics are Si based. Si is regarded as the first generation semiconductor. At the end of the 20th century the next step in the development was taken through GaAs and InP. These III-V semiconductors are considered to be the second generation semiconductors. Presently, a lot of effort is put into GaN and SiC which may come to be the third generation of semiconductors for electronics and optronics. Further ahead in the future lies diamond as a semiconductor material, its properties are ideal for the task but so far only simple np- and Schottky-diodes, far from optimal performance, have been developed.

The control of the charge carriers is essential for semiconductor performance. At elevated temperatures intrinsic charge carriers will be generated over the band gap, at which the performance starts to deteriorate. The concentration of these intrinsic charge carriers depend exponentially on the temperature as

\[ n_i = \sqrt{N_C N_V} e^{-E_G / 2k_B T}, \]  

(11.1)

where \( N_C (N_V) \) is the effective density of states at the conduction (valence) band, \( k_B \) the Boltzmann’s constant and \( E_G \) the band gap [52]. This relation is plotted in Fig. 11.1 for a variety of materials [61]. An intrinsic charge carrier concentration on the order of \( 10^{12} \text{ cm}^{-3} \) normally causes a device to break down, it is seen from Fig. 11.1 that this will happen already at \( 100 - 150 \text{°C} \) for Si, at \( 500 \text{°C} \) for GaAs, close to \( 1000 \text{°C} \) for GaN, and around \( 2000 \text{°C} \) for diamond and AlN. The good performance of AlN and diamond is due to their large band gaps. The band gap of AlN is 6.23 eV in the wurtzite structure and 6.0 eV in the zincblende structure, which is even larger than the 5.5 eV band gap of diamond.

The need for wide band semiconductors is increasing in the fields of light emitters, high temperature, frequency or high power devices and microwave power applications. The advantages of replacing Si with wider band gap materials would be a higher efficiency, since the intrinsic charge carriers, as described above, will cause a leakage at elevated temperatures, smaller devices, less cooling systems which reduce the weight and complexity of the entire sys-
tems, and also lower cost. But this requires that also other properties of these wide gap materials are superior to those of Si, such as higher temperature stability, higher chemical stability, higher thermal conductivity, and higher breakdown field.

During recent years silicon carbide, SiC, has been subject to intense research and due to its advantageous properties such as wide band gap, high electric field strength and high thermal conductivity it has been believed the semiconductor material of tomorrow. Especially for use in pn-diodes, junction field effect transistors and thyristors. Complications in producing high quality SiC has however punctured the SiC breakthrough and hindered the efforts in research and development from turning into a full-scale industrial production of SiC devices.

Silicon carbide is an excellent material for devices operating up to about 600°C and 20 GHz. In comparison with the very promising GaN, SiC has the advantage that more research has been carried out on it and therefore the device fabrication is much more developed in terms of growth substrates etc. GaN also contain much more defects than SiC which has been a practical obstacle. The commercialization of SiC started with the production of blue light emitting diodes, for which a wide band gap is necessary.

There are several different polytypes of SiC, with different band gaps: 2.2 eV 3C-SiC, 3.26 eV 4H-SiC and 3.0 eV 6H-SiC, for example. The band gap is 2-3 times that of Si (1.1 eV) and the electrical breakdown field is ten times that of Si. The higher thermal conductivity means that a certain power can be handled by a smaller device. In total there are 150-250 polytypes, but
only 4H-SiC and 6H-SiC can be grown as substrates. Of these two 4H-SiC is more relevant for electronic applications due to its larger band gap and higher charge carrier mobility. As an example of the higher efficiency of wide band semiconductors, the power losses in SiC switches two orders of magnitude lower than those of the corresponding Si component.

Production of SiC crystals are problematic since SiC does not melt, instead it sublimates at temperature above 2000°C. Normal semiconductor growth from melt is therefore not possible (Czochralski process). Unfortunately it is very hard to grow highly pure SiC and this has drastically slowed down the production of SiC devices. Dislocation defects is the largest concern but samples are also typically contaminated with nitrogen and boron. However, SiC has the advantage that it can oxidize to form SiO2 which is a contributing reason to the Si domination in electronics today.

Carbon in the diamond structure is superior to any other semiconductor material due to its extreme properties: high band gap, high breakdown electric field, high carrier mobility and high thermal conductivity. These properties are shown in table 11.1 (data taken from ref. [61]).

Previously, the use of diamond in electronics has not been possible due to high amount of defects and imperfections in the material. Diamond typically contains nitrogen, boron and hydrogen as contaminants and natural diamond normally has concentrations high enough to make it useless for electronic applications. However, through the recent breakthrough in crystal growth it is now possible to produce highly pure diamond by chemical vapor deposition (CVD). Although diamond is still in its infancy as a semiconductor material,
much higher temperatures than in Si have been reached in various components: \(pn\)-junctions have been tested up to 400°C, Schottky diodes operating up to 1000°C and FETs with frequencies up to 80 GHz, for example. The greatest problem to overcome in order for diamond to be used industrially is the production of large area, inexpensive, high quality single crystal diamond substrates. Because of this, large efforts have been put into growing diamond on other lattice matched substrates, such as Ni, Co, Pt, Si, BeO, SiO₂, cubic BN, \(\beta\)-SiC and GaN. There are also difficulties in doping diamond. \(p\)-type doping has satisfactory been achieved by boron but the natural \(n\)-dopants result in a activation energy of 1.6-1.7 eV, which is a deep level and much too high for thermal activation [62]. Phosphorus is another candidate but is defect level is still 0.59 eV below the conduction band.
Acknowledgements

Among the work related people, I would first like to thank my first supervisor Susanne Mirbt and my “second supervisor” Christopher Castleton. Their “good cop/bad cop” feedback on my manuscripts have hopefully taught me a good balance between doing things rigorously and getting things off the table. I would also like to thank my superbest-work-pals; Petros, Anders, Carlitsson och Ola (who was actually the one who introduced me to the group) for making it fun to go to work.

I have had many questions during my years in the group, and for the answers to them I would like to thank all the seniors and also the past and present PhD-students (Olle, Lars, Biplab, Bettan, Erik, Bulten, Love, Thomas, Björn, Moyses, Oscar, Micke and everyone else).

I must also acknowledge the countless number of table hockey games which have given their lives for the sake of my, Petros’ and Anders’ mental health.

Now on to the people outside of work, the ones who made the late nights in the office durable: my old flat-mate Jonas, Krull-Peter, Keizo, Calleman, Kissus Mor-tinnitus, Rägger, Johan-O, Robban, Tjalle, Marcus, Anders J, Oskar D, Pierre, I love you all in an non-gayhomosexual way.

I would also like to thank my childhood friends Peter & Stefan and my brother Tobias for being the ones that I can always count on, no matter what.

So I’m finally done with writing this thesis, but what can possibly go wrong with my girl by my side, thanks for everything Mia.

Last but not the least, I would like to thank two people who spent a lot of time teaching me the things that in the end got me this far, thank you mamma-Kerstin and pappa-Roger.

How bizarre that you spend years doing research, months writing a thesis, but only the last five minutes on throwing together some junk to fill up this page, the one that people actually do read...
Bibliography


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 316

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se

urn:nbn:se:uu:diva-7926

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007