IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

The microscopic origin of the doping limits in semiconductors and wide-gap materials and recent developments in overcoming these limits: a review

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2002 J. Phys.: Condens. Matter 14 R881

(http://iopscience.iop.org/0953-8984/14/34/201)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.96
The article was downloaded on 18/05/2010 at 12:24

Please note that terms and conditions apply.

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **14** (2002) R881–R903 PII: S0953-8984(02)23469-4

TOPICAL REVIEW

The microscopic origin of the doping limits in semiconductors and wide-gap materials and recent developments in overcoming these limits: a review

S B Zhang

National Renewable Energy Laboratory, Golden, CO 80401, USA

Received 5 March 2002
Published 15 August 2002
Online at stacks.iop.org/JPhysCM/14/R881

Abstract

This paper reviews the recent developments in first-principles total energy studies of the phenomenological equilibrium 'doping limit rule' that governs the maximum electrical conductivity of semiconductors via extrinsic or intrinsic doping. The rule relates the maximum equilibrium carrier concentrations (electrons or holes) of a wide range of materials to their respective band alignments. The microscopic origin of the mysterious 'doping limit rule' is the spontaneous formation of intrinsic defects: e.g., in n-type semiconductors, the formation of cation vacancies. Recent developments in overcoming the equilibrium doping limits are also discussed: it appears that a common route to significantly increase carrier concentrations is to expand the physically accessible range of the dopant atomic chemical potential by non-equilibrium doping processes, which not only suppresses the formation of the intrinsic defects but also lowers the formation energy of the impurities, thereby significantly increasing their solubility.

(Some figures in this article are in colour only in the electronic version)

1. Introduction: what could lead to doping difficulties?

Semiconductors are at the heart of modern electronics and optoelectronics. What makes semiconductors so unique, among other things, is the unique role of the various defects. Without defects, many of the electronic properties of semiconductors would be not much different from those of insulators. Defects in semiconductors can significantly enhance the performance of the hosts. However, they can also introduce undesired effects that could put severe restrictions on or even damage the physical properties of the hosts.

Conductivity via extrinsic dopant is an example. A small fraction of donor or acceptor impurities, say parts per million, is often enough to produce detectable electrical activity in conventional semiconductors. Depending on the properties of the host materials, however, further increase of the impurity concentration can either lead to further increase of the

conductivity, or have no apparent effect, or actually cause a decrease in the conductivity. This problem is particularly severe for wide-gap optoelectronic materials such as for group III nitrides [1] and for II–VI compounds [2] where often at least one type of doping (either n or p) is very difficult if not impossible. Lack of free charge carriers is a serious problem against the utilization of full potential of these important optoelectronic materials, as to date free carrier transport is still the dominant vehicle for delivering/passing information among devices. Even worse, a bulk of device applications is bipolar in nature, that relies critically on the ability to dope the material both sufficiently p- and n-type. With the fast pace in miniaturization of the electronic world and optoelectronics, the doping difficulties become increasingly acute. Solutions to the problem even if partial would not only satisfy our curiosity in material physics but also would promote the advances of modern technologies and in many aspects serve our society. In this article, I shall concentrate on some of the theoretical studies by state-of-the-art first-principles total-energy calculations in the last 15 years or so and emphasize some of the recent developments and emerging trends.

The issue of doping in itself deals with a complex problem of enhancing conductivity. A number of factors can contribute to the final measured conductivity. For example, the mobility of the carriers depends on the lifetime of the carriers and the effective mass, that is a bulk property of the host material. Clearly, the carrier lifetime will be greatly affected by defect scattering and/or by non-radiative centres present in the host. Most important, however, is the carrier concentration. As discussed above, without enough carriers, the device made of the material would lose its functionality. Even if one considers only the difficulties in terms of generating enough free carriers, there are several scenarios that must be considered:

(i) structural instability associated with the dopant,
(ii) dopants being on the wrong lattice positions or forming undesired clusters,
(iii) self-compensation by the creation of intrinsic defects,
(iv) exceptionally low dopant solubility and
(v) too large impurity ionization energy.

These considerations so far, however, are studied mostly on the individual material/impurity basis. In section 4, I shall discuss the general trends that have recently emerged and been recognized.

1.1. Dopant specific instability

The best-known examples of dopant-specific instabilities associated with an impurity are the DX and AX centres in III–V and II–VI semiconductors. DX stands for a complex of a donor (D) and an unknown (at the time of discovery) intrinsic defect (X) [3], whereas AX stands for an acceptor (A) and an intrinsic defect (X). The intrinsic defects (X) in the two cases, of course, may not have any connection. The formation of the DX or AX limits the maximum concentration of free carriers in the host material. Several models for the DX centres have been proposed and subsequently rejected. Finally, a negative-U model [4–8] was proposed that appears to explain all the major experimental observations and has been the model for the DX centres for the last 13 years. In this model, the donor impurity undergoes a large Jahn–Teller distortion along one of the [111] directions. The DX centre can thus also be viewed as an interstitial–vacancy complex centred at the donor. This breaks one of the four tetrahedral bonds with two electrons and creates two dangling bonds which now can hold up to four electrons. As results, not only can a donor hold its own one extra electron, it accepts one more electron from another donor to become stable. Since the proposal of the negative-U model, it has also been found that generic (therefore intrinsic) instability might exist in semiconductors [9, 10].

For the AX centres, the double-broken-bond (DBB) model [10–14] is today's prevailing model. In the DBB model, two *fcc* nearest neighbour anion atoms are displaced towards each other to form a new bond by breaking their original bonds with the cations along the [110] zigzag chain. This creates two cation dangling bonds. Two of the electrons of the broken bonds go to the newly formed anion–anion bond whereas the other two are donated to the electron reservoir (the Fermi level). Because the AX always acts as a double donor, it is highly undesirable in p-type materials. In a number of cases, in fact, DBB formation has been found causing doping limitation (or total failure of doping) in II–VI compounds. However, in other systems with doping difficulties, studies so far have not been able to demonstrate the DBB as the cause of the problems. In addition, an AX-like defect has also been suggested to give rise to the n-type doping limit in heavily doped Si [15].

1.2. Unintentional dopants/unintentional doping effects

Unintentional dopants, of course, could be an important cause for concern. For example, Van de Walle [16] recently pointed out that hydrogen could be the source of free electrons in ZnO. While being good for n-type conductivity, it could cause problems in p-type ZnO, as H could eliminate the holes. In general, however, hydrogen is an amphoteric impurity that passivates either donor [17] or acceptor [18]. The problem with unintentional doping effects is that the intentional dopants may not occupy the desired atomic sites and may behave qualitatively different from one's expectation. One example [19] is the group-I and group-V impurities in ZnO. For p-type II–VI compounds, the desire is to have group-I atoms occupy the group-II sites, and to have group-V atoms occupy the group-VI sites. In reality, the formation energy of the group-I dopant at the interstitial sites could be significantly lower than that at the group-II substitutional sites. The formation energy of group-V dopant at the group-II sites could also be lower than that at the group-VI sites, at least for ZnO under Zn-poor growth conditions. Both will result in severe compensations or even n-type conductivity. Similar antisite effects for group-V impurities in ZnSe have also been suggested [20]. Another unexpected doping effect could occur when the dopant has significant size difference from the host atoms, e.g. when using nitrogen as a dopant in ZnSe [21] and ZnO [22]. Instead of one N substituting one group-VI atom, a N_2 molecule could replace one group-VI atom. This would once again cause n-type conductivity when the original intention was to dope the II–VI host p-type.

1.3. Self-compensation due to spontaneous formation of intrinsic defects

Self-compensation happens because the formation of self-compensating defects involves charge transfer from (to) the dopant to (from) the defects. Baraff and Schlüter [23, 24], Jansen and Sankey [25] and Zhang and Northrup [26–28] have studied the self-compensation mechanisms in III–V and II–VI semiconductors. In particular, in the work on Si-doped GaAs [27], Northrup and Zhang found that the (3−)-charge Ga vacancy is responsible for the compensation of Si at modest Si concentrations. At higher Si concentrations, however, the amphoteric nature of the Si dopant in III–V semiconductors takes over. More and more Si occupy the As sites instead of the Ga sites. The work of Garcia and Northrup [29], on the other hand, deals with p-type ZnSe. It was found that the complex formed between the acceptor (As) and the Zn interstitial is responsible for the compensation of p-type doping in ZnSe. Ramamoorthy and Pantelides [30] studied the formation of complexes between As donors and vacancies in Si, that hinders the further increase of free-electron concentration. In all these cases, charge transfer from the dopants to the defects and the subsequent Coulomb binding play an important role for them to attract each other, forming complexes. The more recent

work of Poykko *et al* [31] on ZnSe and Tsur and Riess [32] on binary oxides provided further support to the self-compensation mechanism as a plausible cause for doping difficulties.

1.4. Solubility limit of the dopants

In certain respects, solubility for substitutional impurity is a measure of the chemical similarity and size difference between a dopant atom and the host atom it intends to replace. The more alike they are, the easier it is for the dopant to replace the host atom. However, if no impurity exists in nature that closely resembles the host atom, it will be difficult. In this regard, Neumark [33], Laks *et al* [34] and more recently Wei and Zhang [35] suggested that the difficulty in doping p-type II–VI semiconductors is due primarily to the lack of adequate dopants with reasonable solubility. Van de Walle *et al* [36] also suggested that the difficulty with p-type GaN lies in the low solubility of Mg, the commonly used acceptor. Another noteworthy point is that while a certain impurity may have reasonably high solubility, the fraction on the desired atomic sites can still be very low. Beryllium doping of GaN is [37] one such example. Several other examples [19–22] have been discussed in section 1.2 above.

1.5. Ionization energy of the dopants

The ionization energy is the energy required to free electrons or holes into the conduction or valence bands from their respective bounded impurity states. According to the Boltzmann statistics, the larger the ionization energy is, the harder to ionize the impurity. In order to ionize effectively at room temperature, the impurity ionization energy has to be comparable to $kT = 26$ meV or less. An impurity can also be a resonance with its energy level either above the conduction band edge (n-type) or below the valence band edge (p-type). In such cases, the charge carriers in the level could ionize instantly, giving 100% ionization efficiency. It is, however, noteworthy that due to the impurity effective Coulomb attraction, instant ionization may not actually happen. Instead, the electron (hole) forms a shallow impurity bounded exciton with an energy level a few milli-electron volts below (above) the band edges.

2. The physics and chemistry of point defects and complexes

The equilibrium concentration of a given defect is determined by its formation energy and by the growth temperature. Several factors may affect the formation energy: the doping, the concentration of other defects co-existing in the host material and the stoichiometry of the host, that is often described in terms of partial pressures of the individual components present in the growth chamber. A general description of defect chemistry is given in [38]. In first-principles total energy calculations, however, it is quite frequent that one describes the defect formation energy in terms of the atomic chemical potentials [26, 34] and the Fermi energy (which is the chemical potential of the electrons) [23]. The advantage of employing the chemical potentials is that one can easily separate one defect from others, and distinguish the effects due to different physical properties reflected in different atomic chemical potentials. One can calculate the Fermi energy by the charge neutrality requirement among the various defects and dopants and calculate the atomic chemical potentials via the equation of states of the various gaseous phases [27].

In the following, I shall demonstrate by examples how the host atomic chemical potential μ_{host}, the impurity atomic chemical potential μ_I and the Fermi energy ε_F affect the defect formation energy ΔH_f. These are important concepts that hold the key to understand and to overcome the doping difficulties.

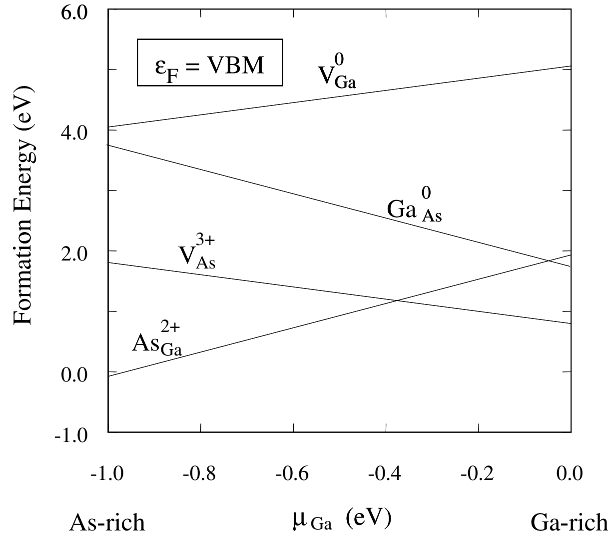

Figure 1. Calculated defect formation energy in GaAs as a function of the Ga atomic chemical potential, μ_{Ga}. The Fermi energy is at the valence band maximum (VBM). (Reproduced with permission from the American Institute of Physics.)

2.1. Dependence of the defect formation energy on the chemical potentials

Consider first charge-neutral defects. The formation energy of such defects often depends on the atomic chemical potentials of the host atoms. For example, to form a cation vacancy in a binary compound, one cation atom is removed from the host material and is placed in the atomic 'reservoir' of energy, μ_C. The formation energy is thus

$$\Delta H_f(\mathrm{V}_C^0) = E_{tot}(\mathrm{V}_C^0) - E_{tot}(0) + \mu_C \qquad (1)$$

where $E_{tot}(\mathrm{V}_C^0)$ is the total energy of the host crystal having one vacancy and $E_{tot}(0)$ is the total energy of the host without any defect. Figure 1 shows [26, 28] a few calculated formation energies of native defects in p-type GaAs as a function of the Ga chemical potential. One sees that the Ga-on-As antisite ($\mathrm{Ga}_{\mathrm{As}}$) and the As vacancy ($\mathrm{V}_{\mathrm{As}}$) are easier to form in Ga-rich conditions, while in As-rich conditions the As-on-Ga antisite ($\mathrm{As}_{\mathrm{Ga}}$) and the Ga vacancy ($\mathrm{V}_{\mathrm{Ga}}$) are instead easier to form. To suppress compensation by intrinsic defects, it is always advantageous to prepare materials at the chemical potentials that *maximize* the formation energies of the undesirable defects.

2.2. Dependence of the defect formation energy on the Fermi energy

The formation energy of a positively charged defect D^+ is equal to the energy of a neutral defect D^0, minus the energy $\varepsilon(0/+)$ required to ionize the D^0 to form D^+, plus the energy of the ionized electron residing in the electron reservoir (=Fermi energy). Thus,

$$\Delta H_f(\mathrm{D}^+) = \Delta H_f(\mathrm{D}^0) - \varepsilon(0/+) + \varepsilon_F. \qquad (2)$$

As shown in figure 2, the higher the Fermi energy is, the more energy is needed to form D^+. So donors (that produce electrons in the reaction $\mathrm{D}^0 \rightarrow \mathrm{D}^+ + \mathrm{e}^-$) are more difficult to form in electron-rich (n-type) materials. Similarly, for acceptors, the formation energy *decreases* as ε_F increases,

$$\Delta H_f(\mathrm{A}^-) = \Delta H_f(\mathrm{A}^0) + \varepsilon(-/0) - \varepsilon_F. \qquad (3)$$

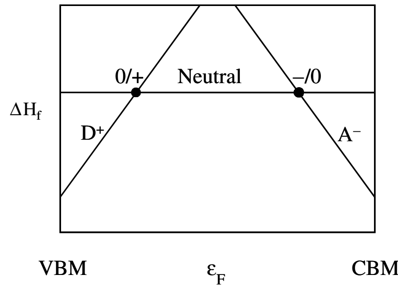

Figure 2. A schematic plot showing the effect of the defect charge state on its formation energy, as a function of the Fermi energy, ε_F. (Reproduced with permission from the American Institute of Physics.)

So acceptors (that produce holes in the reaction $A^- + h^+ \rightarrow A^0$) are more difficult to form in hole-rich (p-type) materials.

These simple considerations show that

(a) If one dopes a material intentionally n-type via some donor impurity, as ε_F moves up in the gap, the formation energy of native acceptors $\Delta H_f(A^-)$ decreases. At some points, the formation energy is so low that some native acceptors (e.g. the cation vacancy) could form *spontaneously*, thus negating the effect of the intentionally introduced donors.

(b) If one dopes a material intentionally p-type via some acceptor impurity, as ε_F moves down in the gap, the formation energy of positive donors $\Delta H_f(D^+)$ decreases. At some point the formation energy is so low that some native donors (e.g. the anion vacancy) could form spontaneously, thus negating the effects of the intentionally introduced acceptors.

Thus, the spontaneous formation of native defects determines the maximum and minimum Fermi energies (which will be termed the pinning energies) over which ε_F may vary under equilibrium.

In general, the formation energy of a defect (or impurity) α of charge q in a host material, e.g. a ternary $A_lB_mC_n$ for generality, is given [26] by

$$\begin{aligned}\Delta H_f(q,\alpha) &= E_{tot}(q,\alpha) - E_{tot}(0) + \sum_{s=A,B,C} n_s\mu_s + n_I\mu_I + q\varepsilon_F \\ &= \Delta E_{tot}(q,\alpha) + \sum_{s=A,B,C} n_s\mu_{s,solid} + n_I\mu_{I,solid} + q\varepsilon_{VBM} \\ &\quad + \sum_{s=A,B,C} n_s\mu_s + n_I\mu_I + q\varepsilon_F \end{aligned} \tag{4}$$

where $\Delta E_{tot}(q,\alpha) = E_{tot}(q,\alpha) - E_{tot}(0)$, n_s is the number of the sth atoms and n_I is the number of the impurity (I) atoms being transferred to the atomic reservoirs to form the defect. The μ are the atomic chemical potentials. Usually, the atomic chemical potential can vary over a certain range, as mentioned above, with the upper limit equal to the energy of the elemental solid (or gas), $\mu_{s,solid}$. This happens because if $\mu_s > \mu_{s,solid}$, the elemental solid will spontaneously form, that hinders any further increase of μ_s. For convenience, in equation (4) we have set μ_s (old) $= \mu_{s,solid} + \mu_s$ (new) and μ_I (old) $= \mu_{I,solid} + \mu_I$ (new) so that μ_s (new) and μ_I (new) $\leqslant 0$. Similarly, we have set ε_F(old) $= \varepsilon_{VBM} + \varepsilon_F$ (new). The Fermi energy at which two different charge states, q and q', of the same defect α have the same formation energy, $\Delta H_f(q,\alpha) = \Delta H_f(q',\alpha)$, defines the defect transition energy, $\varepsilon(q/q')$. Hence

$$\varepsilon(q/q') = [\Delta E_{tot}(q,\alpha) - \Delta E_{tot}(q',\alpha)]/(q'-q). \tag{5}$$

Also, for the host material to be thermodynamically stable, it also requires [26]

$$l\mu_A + m\mu_B + n\mu_C = \mu(\mathrm{A}_l\mathrm{B}_m\mathrm{C}_n) = \Delta H(\mathrm{A}_l\mathrm{B}_m\mathrm{C}_n) \quad (6)$$

where $\Delta H(\mathrm{A}_l\mathrm{B}_m\mathrm{C}_n)$ is the formation enthalpy of the host. Equation (6) reduces the number of independent μ variables by one. For example, for the defects in GaAs in figure 1, μ_{Ga} is used as the independent variable. Thus, by equation (6) one has $\Delta H(\mathrm{GaAs})$ (which is the As-rich limit) $\leqslant \mu_{\mathrm{Ga}} \leqslant 0$ (which is the Ga-rich limit).

3. Background for first-principles defect calculations

Since the groundbreaking work of Baraff and Schlüter [23, 24] on native defects and defect complexes in GaAs, first-principles studies of the various defects have played an increasingly important role in the understanding of the physical and chemical properties of defects in general. The great advantages of the first-principles calculations are the elimination (or more precisely a great reduction) of the need to rely on empirical parameters, and the deeper insights the first-principles methods can provide that have never been possible in the past. The effects of atomic relaxation are fully taken into account in studies after Baraff and Schlüter. In most first-principles calculations, one takes the supercell approach adopted from earlier surface calculations where one constructs a fictitious periodic system: each unit cell contains one defect or defect complex. One then uses the standard first-principles approaches, for example the pseudopotential [39–42] or the linearized augmented plane wave (LAPW) approaches [43, 44], to calculate the eigenenergies and total energies, using the density functional approach under the local density approximation (LDA) [45, 46] with various forms of exchange–correlation functionalities or more recently the general gradient approximation (GGA) [47]. The total energy of the defect system is calculated by using a special k-point scheme for integration over the entire Brillouin zone [48, 49]. Often, and in particular for our applications, the defects are charged. To deal with charged defects that form an infinite array in the periodic system, a charge background, in the form of a jellium, must be added [50], so the total energy per unit cell does not become infinite. This, however, introduces some errors. Alternatively, one can apply a real-space approach to the defect problems [51]. The advantage of the real-space approach is, of course, the elimination of the spurious supercell–supercell interactions. On the other hand, one needs to passivate the dangling bonds and eliminate other surface effects intrinsic to any finite-size clusters. In addition, the following points are worth mentioning:

(i) Errors due to the jellium background

Makov and Payne [52] have demonstrated that the errors can be corrected up to $\mathrm{O}(L^{-5})$ where L is the dimension of the supercell. In many of the calculations, however, only first-order correction (L^{-3}) is included assuming a point charge at the centre of the defect or complex. Clearly, this is a rather crude estimate, as the real defect states will have their wavefunctions reasonably spread over several atomic distances. A somewhat better approximation is to distribute the charge into a collection of point charges, especially for high charge state defects such as (3+) or (3−), etc. In the work of Zhang and Northrup on GaAs [26], for example, the effect due to the triple charge on the Ga vacancy was estimated with a charge distribution over four nn As atoms, i.e. $(-3/4)$ per As.

(ii) The importance of electronic energy reference

It is customary that almost all theoretical studies refer to the valence band maximum (VBM) as the electron energy reference (cf equation (4)). However, in the LDA calculation, the energy

zero, whose absolute value is meaningless for a periodic infinite system, will have to be set in some way, often by convenience. Thus, one must make sure that the energy zero is indeed reset to the VBM by $\varepsilon_F(\text{old}) = \varepsilon_{VBM} + \varepsilon_F(\text{new})$, or serious errors can occur as ε_{VBM} can be as large as several electron volts. This problem was realized back in the late 1980s in the study of the DX centres under pressure and in GaAlAs alloys [6]. A closely related issue is how to find the ε_{VBM} for finite-size unit cells where every eigenvalue of the bulk could be strongly perturbed by the presence of the periodic array of defects. In the work of Zhang and Northrup [26], a region in the unit cell was chosen to best represent the bulk in the presence of the defects. Average self-consistent potential was calculated for the defect cell and aligned with the average potential of the same cell but without any defect. The rationale was that the potential is a local quantity that will converge to the bulk value with cell size much faster than the eigenvalues. For large enough unit cells, of course, this alignment term approaches zero. According to my experience with ordinary semiconductors, the alignment term is usually small, to within 0.1 eV or less for supercell size equal to or larger than 64 atoms. However, for systems with high lying d bands, it can be somewhat larger, 0.3 eV, in particular, for $CuInSe_2$ in a 32-atom cell [53].

(iii) One-particle defect levels: which way to calculate is better?

In the supercell approach, it is inevitable that the calculated defect states, whether shallow or deep, show some kind of dispersion. It appears that within the framework of the supercell approach, it is best to calculate the *average* one-particle level positions with respect to the *average* band edges (VBM or the conduction band minimum (CBM)) at the special k-points used in the total calculation [54]. There are at least two advantages of this.

(a) Because the total energy is calculated with a Brillouin zone sum over the special k-points, the one-particle levels calculated this way are consistent with the defect transition energy levels (or defect levels) calculated from the total energy difference between two different charge states, *when* the Franck–Cordon (FC) shift [55] is negligible. Even if the FC shift is not negligible, this approach will at least reflect the actual shift given by atomic relaxations.

(b) It also provides the correct one-particle positions for shallow defects. This is rather natural from (a) above, as shallow defects do not have any significant FC shifts.

Recently, accurate determination of the shallow level positions becomes an important issue, especially in the context of doping. So far, however, in most defect calculations, the defect levels were plotted against the VBM *at* Γ. Clearly, using such an approach, the calculated acceptor levels will almost always be shallower than they actually are, as the VBM at Γ is typically several tenths of an eV higher than the average VBM over the special k-points. Alternatively, one may calculate the one-particle defect levels at Γ with respect to the VBM at Γ to estimate the defect level position. In a number of cases, in my experience, the two approaches yield similar results. The average-VBM approach has the advantage of being consistent with the total energy calculation, whereas the everything-Γ approach does not. From the above discussions, it appears that one should also present the defect formation energies, in addition to the defect levels, with respect to the *average* LDA bandgap, that for finite cells is larger than the minimum gap at Γ. This would reduce systematically the LDA gap errors on the defect formation energies (see below), although in a rather unexpected way.

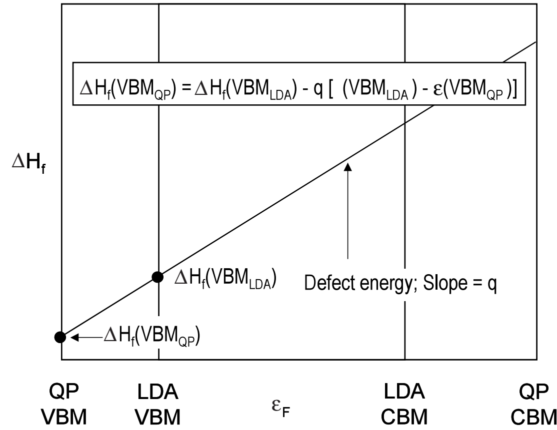

Figure 3. A schematic plot facilitating the discussion of the GW quasiparticle correction to the LDA defect formation energy.

(iv) The LDA corrections

The errors in the defect formation energy and deep level position associated with the underestimation of the LDA bandgap are fundamental errors that cannot be easily accounted for unless one goes beyond the LDA. Even so, some general trends emerge from past experiences with LDA calculations and from examining the LDA corrections by the GW quasiparticle approaches [56, 57].

First, one can expect that the error in the LDA total energy, E_{tot}, for closed-shell defects is reasonably small because the basic principles of the density functional theory guarantee [45] accurate ground state properties. Closed shell here means that all the defect states derived from the valence band states are fully occupied whereas all the defect states derived from the conduction band states are fully empty. For example, the defect states of an anion vacancy should be completely empty, as there are cation dangling-bond states near the CBM. This leads to V_{As}^{3+} in GaAs and V_{O}^{2+} in ZnO. Conversely, the defect states of a cation vacancy should be completely filled, as these are anion dangling-bond states near the VBM. This leads to V_{Ga}^{3-} in GaAs and V_{Zn}^{2-} in ZnO. Indeed, LDA calculations of reconstructed semiconductor surfaces where all individual atoms form closed shells and collectively satisfy the electron-counting model [58] yields highly reliable results.

Second, GW quasiparticle calculations provide an estimate of the LDA errors on the band edge states. A closer examination of the procedure described in figure 3 reveals that both the valence-band-derived states and the conduction-band-derived states need corrections, not just the latter. In particular, the LDA value, $E_{tot}(q, \alpha) - E_{tot}(0) + \sum_{s=A,B,C} n_s \mu_{s,solid}$ (=the inclined line in figure 3), is reasonably correct because the first term here involves only a closed-shell defect whereas the second and third terms involve only the ground states of the bulk. One thus only needs to correct the VBM to which the Fermi level ε_F is referenced, as graphically depicted in figure 3. Namely,

$$\Delta H_f(\mathrm{VBM}_{QP}) = \Delta H_f(\mathrm{VBM}_{LDA}) - q[\varepsilon(\mathrm{VBM}_{LDA}) - \varepsilon(\mathrm{VBM}_{QP})] \qquad (7)$$

where q is the charge state of the defect. Because the term in the brackets $\delta\varepsilon_{VBM} = \varepsilon(VBM_{LDA}) - \varepsilon(VBM_{QP})$ is usually larger than zero [56, 57], a general trend emerges from this discussion: LDA correction will lower the formation energy of positively charged, and raise that of negatively charged, closed-shell defects at the VBM. The magnitude of the

corrections depends on $\delta\varepsilon_{VBM}$. For C, Si, Ge, AlAs and GaAs the corrections are [56, 57] only a couple of tenths of an eV at Γ and are even smaller for the k-point averaged VBM (see the discussions in (iii) above), but for more ionic materials such as LiCl it could be larger than 1 eV [56].

Third, for shallow impurities, the charge-neutral state usually has an open shell whereas the charged ones do not. In these cases, one should preserve the relative positions of the impurity levels with respect to the band edge states, as they probably both share the same LDA errors. Hence, in the case of a single acceptor, this requires us to move $\varepsilon(-/0)$ down by $\delta\varepsilon_{VBM}$, or, according to equation (5), to move $\Delta E_{tot}(0)$ up by $\delta\varepsilon_{VBM}$. In the case of a single donor, this requires us to move $\varepsilon(0/+)$ up by $\delta\varepsilon_{CBM}$, or to move $\Delta E_{tot}(0)$ up by $\delta\varepsilon_{CBM} = \varepsilon(\mathrm{CBM}_{QP}) - \varepsilon(\mathrm{CBM}_{LDA})$. For deep-level open-shell defects, however, there are no such simple rules to follow. Instead, one has to determine the fraction of the state being either valence or conduction band state derived, via, for example, a projection scheme [59].

4. The phenomenological models that predict the doping limits

Previous studies on transition-metal impurities in semiconductors established [60, 61] the universality of the energetic positions of the deep levels with respect to the vacuum level (the so-called 'vacuum pinning rule'). Similarly, a 'doping limit rule' was observed and discussed based on an amphoteric defect model by Walukiewicz [62–65]. The work of Tokumitsu [66], Ferreira *et al* [67] and more recently Zhang *et al* [68] established a broader base for the 'doping limit rule' in a wide range of semiconductors, showing that there are common and surprisingly simple principles that cut across failure to dope in different material classes such as group-IV, III–V and II–VI compounds. According to Zhang *et al* [68], doping failure is not related to the *size* of the bandgap [69] *per se*, but rather to the position of the VBM with respect to a pinning energy, $\varepsilon_{pin}^{(p)}$, for p-type conductivity, and the position of the CBM with respect to a pinning energy, $\varepsilon_{pin}^{(n)}$, for n-type conductivity. In other words,

(a) a material for which $\varepsilon_{pin}^{(n)} \ll \varepsilon_{CBM}$ cannot be doped n-type;
(b) a material for which $\varepsilon_{pin}^{(p)} \gg \varepsilon_{VBM}$ cannot be doped p-type.

From the previous discussions, equilibrium doping reaches its limit when there are sufficient spontaneously generated defects that compensate the intentional dopants. The net maximum concentration $N^{(n/p)}(T, \varepsilon_F)$ of free carriers (electrons or holes) in a semiconductor is determined [70], in the single, parabolic band approximation, by the position of the pinning Fermi energy,

$$N^{(n/p)}[T, \varepsilon_F^{(n/p)}] = \frac{1}{2\pi^2}[2m^*_{(n/p)}]^{3/2} \int_0^\infty \varepsilon^{1/2}\,\mathrm{d}\varepsilon/[\exp(\beta(\varepsilon - \varepsilon_F^{(n/p)})) + 1] \quad (8)$$

where $\beta = 1/kT$ is the temperature factor, and m^* is the carrier effective mass. Given the measured maximum electron or hole concentration, $N_{max}^{(n/p)}$, one may obtain [68] $\varepsilon_{pin}^{(n)}$ and $\varepsilon_{pin}^{(p)}$ simply by inverting equation (8).

Figure 4 shows the values of $\varepsilon_{pin}^{(n)}$ and $\varepsilon_{pin}^{(p)}$ obtained from the measured maximum carrier concentrations in various III–V compounds. In this figure, the VBMs are aligned. While the data for $\varepsilon_{pin}^{(p)}$ are scattered within a relatively small range of 0.5 eV, the data for $\varepsilon_{pin}^{(n)}$ are scattered over a considerably larger range of 1.2 eV, showing no trend. Walukiewicz, however, showed (as later summarized in [65]) that a better correlation could be obtained if one includes the band offsets in the alignment in figure 4. (The most recent calculated valence band offsets can be found in [71].) This is equivalent to aligning individual bulk band diagrams with respect to

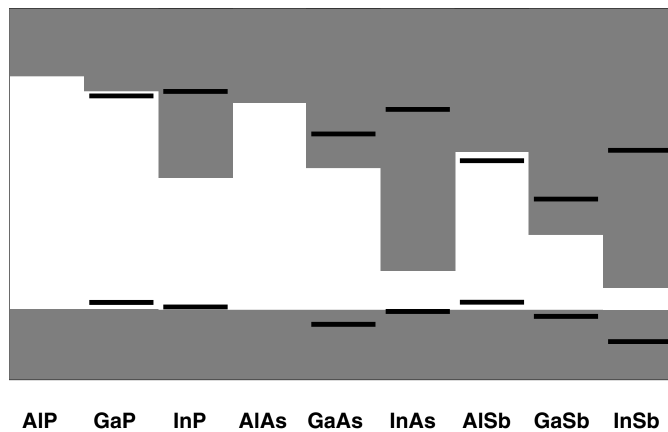

Figure 4. The calculated pinning energies (by inverting equation (8)), $\varepsilon_{pin}^{(n)}$ and $\varepsilon_{pin}^{(p)}$, for conventional III–V compounds. The VBMs are lined up. Notice the large scattering in $\varepsilon_{pin}^{(n)}$.

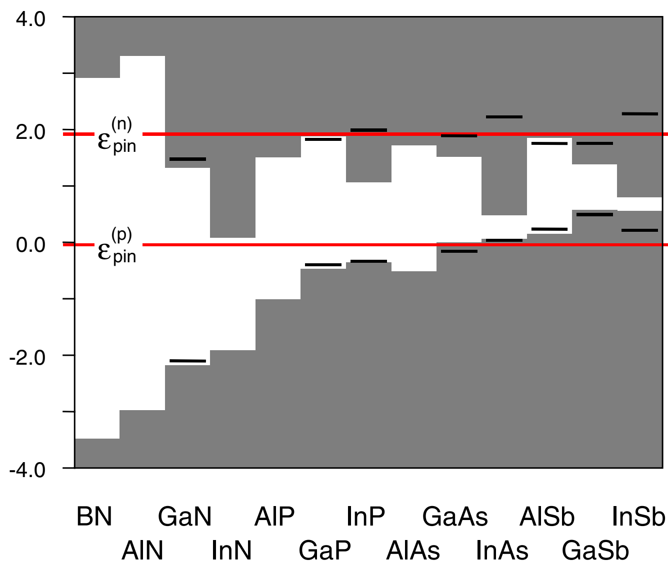

Figure 5. Same as in figure 4 with, however, also group III nitrides. The bands are aligned with respect to the calculated valence band offsets with the energy zero at the VBM of GaAs. The average pinning energies are also shown for $\varepsilon_{pin}^{(n)}$ and $\varepsilon_{pin}^{(p)}$, respectively. (Reproduced with permission from the American Institute of Physics.)

an absolute energy reference, for example to an absolute vacuum level, as shown in figure 5 for the III–V compounds. Here, except for p-type GaN, the scatters in $\varepsilon_{pin}^{(n)}$, and separately in $\varepsilon_{pin}^{(p)}$, are both approximately 0.5 eV. Results for II–VI and I–III–VI_2 ternary compounds are shown in figure 6. This remarkably simple rule permits one to predict rather accurately whether a material can be doped or cannot be doped a certain type, merely by positioning its band energies in a diagram like figures 5 or 6 with similar materials that have known doping properties.

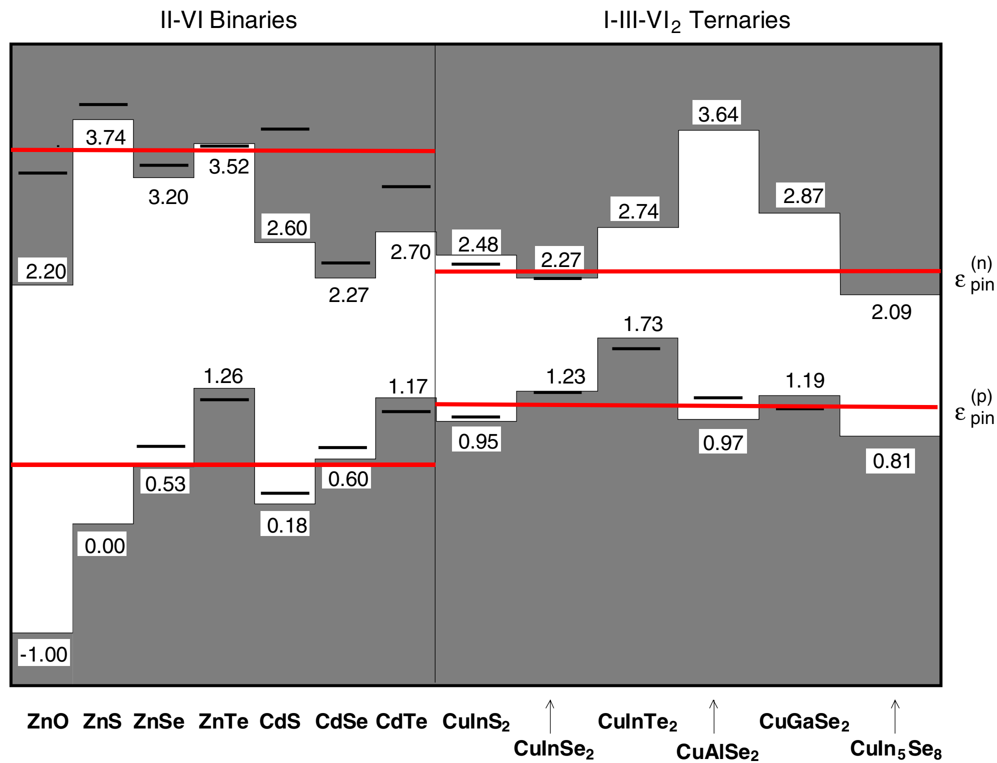

Figure 6. Same as in figure 5 but, however, for the II–VI and I–III–VI$_2$ compounds. The energy zero is at the VBM of ZnS. (Reproduced with permission from the American Institute of Physics.)

5. The microscopic origin of the 'doping limit rule'

In searching for the microscopic origin of the 'doping limit rule', Zhang *et al* [72] have recently considered two types of defect:

(i) the intrinsic cation vacancy V_{cation} and
(ii) the extrinsic DX centre in n-type III–V semiconductors.

In case (i), the pinning energy is determined by the condition at which the defect formation energy equals zero

$$\Delta H_f[q, \alpha, \varepsilon_F = \varepsilon_{pin}^{(n)}] = 0. \tag{9}$$

Calculation for the cation vacancy indicated [26] that the highest defect level is the $(2-/3-)$ transition level near the VBM. For n-type doping, however, ε_F is always closer to the CBM than to the VBM. Hence, the cation vacancy will most likely have a charge state, $q = -3$. In order to reach the maximum $\varepsilon_{pin}^{(n)}$, the cation chemical potential should also be at its maximum, $\mu_{cation} = 0$. Hence, from equations (9) and (4) with $n_s = 1$ and $\mu_{s,solid} = E_{tot}$ (cation solid), one has

$$\varepsilon_{pin}^{(n)} = [\Delta E_{tot}(-3, V_{cation}) + E_{tot}(\text{cation solid})]/3. \tag{10}$$

In case (ii), spontaneous formation of the DX centre of charge q' from the precursor donor (d) state of charge q takes place when the Fermi energy ε_F is at the defect transition energy

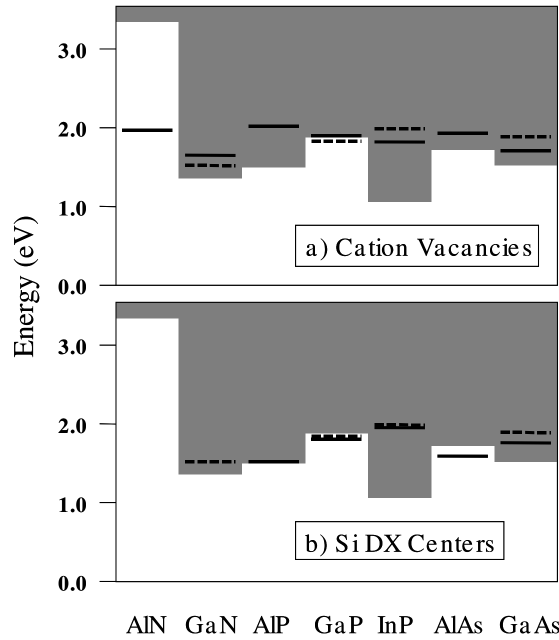

Figure 7. First-principles (solid lines) versus experimental (dashed lines) n-type pinning energies in seven III–V compounds for (a) cation vacancies and (b) silicon DX centres. The energy zero is the VBM of GaAs [72].

$\varepsilon(q/q')$. By equation (5), one has

$$\varepsilon_{pin}^{(n)} = [\Delta E_{tot}(q,\alpha) - \Delta E_{tot}(q',\alpha)]/(q'-q). \tag{11}$$

The DX centre (and its precursor donor state) can in principle present in three different charge states $q = (+, 0, -)$, with the shallow donor (+/0) level near the CBM. Depending on the host, however, the (+/−) level can be either above the (+/0) level (thus, a positive-U system) or below the (+/0) level (a negative-U system). Zhang *et al* [72] have calculated the DX assuming the donor impurity is a Si atom. For the negative-U system (where two electrons in the same defect orbital attract each other)

$$\varepsilon_{pin}^{(n)} = \varepsilon(+/-) = [\Delta E_{tot}(-1, \mathrm{DX}) - \Delta E_{tot}(+1, \mathrm{Si_{Ga}})]/2. \tag{12}$$

For the positive-U system (where the two electrons, instead, repel each other)

$$\varepsilon_{pin}^{(n)} = \varepsilon(0/-) = \Delta E_{tot}(-1, \mathrm{DX}) - \Delta E_{tot}(0, \mathrm{Si_{Ga}}). \tag{13}$$

Figure 7 shows the calculated pinning energies: seven for the vacancies and five for the DX centres. The lower number for the DX centres is because they are unstable in GaN and AlN by the calculation of Zhang *et al* [72] and by that of Park and Chadi [14]. The following can be seen from figure 7.

(i) There is a good quantitative agreement between the calculated $\varepsilon_{pin}^{(n)}$ and those deduced from experiment. This means that the calculated $\varepsilon_{pin}^{(n)}$ can be used to predict the maximum n-type carrier density. For example, it explains why n-type doping in GaP is much less effective than in InP, because $\varepsilon_{pin}^{(n)}(\mathrm{V}_{cation})$ is considerably higher than the CBM in InP, but not in GaP. It also explains why n-type doping is possible in GaN, but impossible in AlN because $\varepsilon_{pin}^{(n)}(\mathrm{V}_{cation})$ is in the gap.

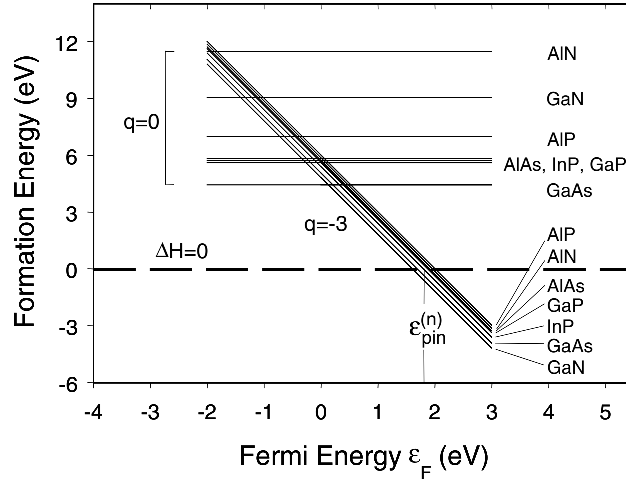

Figure 8. The calculated cation vacancy formation energies for charge state $q = 0$ and -3, as a function of an absolute Fermi energy ε_F in seven III–V compounds. The zero of the ε_F is set at the VBM of GaAs [72].

(ii) There is a clear tendency that $\varepsilon_{pin}^{(n)}$ line up for the cation vacancies in figure 7(a). The variance σ_p of the seven calculated $\varepsilon_{pin}^{(n)}$ values is less than 0.4 eV. Note that had one ignored the valence band offsets, however, the variance here would be an order of magnitude larger, 3.3 eV. Remarkably, good line-up exists only for charge state q that corresponds to an electronic closed shell. For the cation vacancies, this occurs at $q = -3$ for the III–V compounds and at $q = -2$ for the II–VI compounds. The $q = -1$ DX centres can also be viewed as closed-shell defects. While a similar trend of alignment is also seen in figure 7(b) for the Si-induced DX centres with $\sigma_p = 0.4$ eV, it is expected that the DX level positions will vary with the chemical identity of the actual impurity, and are thus not truly intrinsic. While no systematic study for the DX centres is yet available, calculations for the AX centres by different impurities indeed showed [10] drastically different pinning energies. As such, it is most likely that the intrinsic defects, such as the cation vacancies, hold the key to the observed ε_{pin} alignments.

Zhang *et al* [72] presented a qualitative discussion on the physical origin of the $\varepsilon_{pin}^{(n)}$ alignment. First, let us separate the total energy of the vacancy into a sum of its occupied eigenvalues plus other contributions (=electron–electron double counting, exchange–correlation, and ion–ion term): $E_{tot} = \Sigma_i \varepsilon_i + F$. For bulk zincblende semiconductors, six electrons occupy the VBM (ε_{VBM}). For the charge-neutral vacancy in which N electrons are removed from the valence band edge ($N = 3$ in III–V compounds and 2 in II–VI compounds), $(6 - N)$ electrons occupy the VBM-derived t_2 defect level. Thus, using equation (4) with $n_s = 1$, $\mu_s = 0$ and $\mu_{s,solid} = E_{tot}$ (cation solid), one has

$$\Delta H_f(q = 0, \mathrm{V}_{cation}) = E_{tot}(q = 0, \mathrm{V}_{cation}) - E_{tot}(0) + E_{tot}(\text{cation solid})$$
$$= (6 - N)\varepsilon_{t2} - 6\varepsilon_{VBM} + F(\mathrm{V}_{cation}) - F(0) + E_{tot}(\text{cation solid}). \quad (14)$$

For a charged vacancy ($q \neq 0$), it was found [72] that $\varepsilon(q/0) \approx \varepsilon_{t2}$, so $\Delta H_f(q \neq 0, \mathrm{V}_{cation}) \approx \Delta H_f(q = 0, \mathrm{V}_{cation}) + q(\varepsilon_F - \varepsilon_{t2})$ (cf equation (2) but with $q > 1$). Thus

$$\Delta H_f(q, \mathrm{V}_{cation}) \approx [6(\varepsilon_{t2} - \varepsilon_{VBM}) + F(\mathrm{V}_{cation}) - F(0) + E_{tot}(\text{cation solid})]$$
$$- (q + N)\varepsilon_{t2} + q\varepsilon_F. \quad (15)$$

Depending on q and N, the vacancy formation energy $\Delta H_f(q, \text{V}_{cation})$ can vary from material to material, as shown by the $\sim$7 eV spread of the horizontal lines for $q = 0$ in figure 8. However, for closed-shell vacancies ($q = -N$), the $(q + N)\varepsilon_{t2}$ term in equation (15) vanishes. This considerably reduces the large material dependence of $\Delta H_f(q, \text{V}_{cation})$, as shown by the closely bunched inclined lines for $q = -3$ in figure 8. The pinning energy $\varepsilon_{pin}^{(n)}$ is the Fermi energy at which $\Delta H_f\ (q = -3, \text{V}_{cation}) = 0$ (cf equation (9)). The spread in $\varepsilon_{pin}^{(n)}$ (on the horizontal axis in figure 8) is only 0.4 eV, thus nearly independent of the host semiconductors.

6. Going beyond the equilibrium doping limits

The above discussions demonstrate clearly the intrinsic thermodynamic limitations to the ability to dope semiconductors. To overcome these fundamental limitations, there are currently at least two possible approaches.

(a) Designing new materials with desired doping properties

One such example is the recent success in fabricating p-type transparent conductive oxides (TCOs). Only five years ago, there was no such thing as p-type TCOs. With the recent developments, however, not only were several Cu-based p-type TCOs discovered [73–76] but also the physical mechanism that led to p-type conductivity was explained by Nie *et al* [77].

(b) Going beyond the equilibrium doping

Several experiments [15, 22, 78–83] have led the way in this direction in the past a few years, but little theoretical study has followed suit due to the difficulties in dealing with non-equilibrium processes. From the previous discussions, going beyond the equilibrium theory implies that, due to some kinetic limitations, the concentration of the charge compensating defects can be suppressed to significantly below the equilibrium value. Often, this requires low temperature growth. More importantly, the solubility of the dopant has to be significantly enhanced in a number of important cases, e.g. p-type ZnO:N [22, 79], where, even without the compensation by intrinsic defects, a desirable carrier concentration cannot be reached due to the exceedingly low equilibrium solubility.

6.1. Surface-enhanced nitrogen solubility in dilute GaAs nitrides

Our first understanding of the non-equilibrium impurity solubility came from the study of isoelectronic doping of GaAs, or alloying, by nitrogen. For this reason, here I shall discuss in some detail how the surface processes determine the N solubility. The general principles apply to impurities that provide charge carriers as well, although the details could be different.

Alloying semiconductors with large lattice mismatch is of fundamental interest. Not only are the physical properties of such alloys expected to be qualitatively different from conventional alloys; their growth often poses serious challenges because the alloys are often thermodynamically unstable. With a few per cent of N, its conduction band edge can be significantly lower than that of GaAs [84–86]. As such, $GaAs_{1-x}N_x$ is a potential candidate for long wavelength lasers and as an absorber for high efficiency tandem solar cells. A graded GaAsN layer is also a natural buffer layer for the epitaxial growth of cubic-phase GaN on GaAs, thus holding promise for high quality integration between two important optoelectronic materials. However, equilibrium solubility of N in bulk GaAs is exceedingly low ([N] $< 10^{14}$ cm^{-3} at $T_{growth} = 650\,^{\circ}$C) [87, 88], due to the formation of a fully relaxed,

secondary GaN phase, and yet single-phase epitaxial films grown at $T = 400$–$650\,^{\circ}$C with [N] as high as $\sim$10% have been reported [89–94].

The incorporation mechanism of excessive nitrogen in GaAs was largely unknown until the work of Zhang and Wei [95]. In their work, they suggested that surface effects suppress the formation of secondary GaN phase during epitaxial growth. Indeed, a key factor in fabricating high [N] homogeneous GaAs:N films is to eliminate the formation of GaN precipitates. As such, there exists a new region of the atomic chemical potentials ($\mu_{\rm Ga}$, $\mu_{\rm As}$, $\mu_{\rm N}$), available for epitaxial growth but not available for equilibrium growth. As a result, the calculated maximum solubility [N] at typical growth temperatures is a few per cent, e.g. 4% at $T_{growth} = 650\,^{\circ}$C instead of $<10^{14}$ cm^{-3} by equilibrium theory.

(a) Equilibrium solubility

From equation (4), the formation energy of a charge-neutral defect in GaAs:N is given by

$$\Delta H_f = \Delta E_{tot}(0,\alpha) + n_{\rm Ga}\mu_{{\rm Ga},solid} + n_{\rm As}\mu_{{\rm As},solid} + n_{\rm N}\mu_{\rm N,N_2} + n_{\rm Ga}\mu_{\rm Ga} + n_{\rm As}\mu_{\rm As} + n_{\rm N}\mu_{\rm N} = C + n_{\rm Ga}\mu_{\rm Ga} + n_{\rm As}\mu_{\rm As} + n_{\rm N}\mu_{\rm N}, \qquad (16)$$

where C is a constant and

$$\mu_{\rm Ga} \leqslant 0, \qquad \mu_{\rm As} \leqslant 0 \qquad \text{and} \qquad \mu_{\rm N} \leqslant 0. \qquad (17)$$

Equation (6) further requires that $\mu_{\rm Ga} + \mu_{\rm As} = \Delta H$ (GaAs), where the calculated GaAs formation enthalpy $\Delta H({\rm GaAs}) = -0.62$ eV. Thus, the defect formation energies in GaAs:N are functions of only two independent variables, ($\mu_{\rm As}$, $\mu_{\rm N}$), satisfying (see figure 9)

$$-0.62 \text{ eV} \leqslant \mu_{\rm As} \leqslant 0, \qquad \text{and} \qquad \mu_{\rm N} \leqslant 0. \qquad (18)$$

Physically, less negative $\mu_{\rm As}$ (or $\mu_{\rm N}$) corresponds to more As-(or N-) rich growth conditions, and vice versa. Spontaneous formation of the secondary bulk GaN phase, however, puts a further restriction on the chemical potentials, namely

$$\mu_{\rm Ga} + \mu_{\rm N} \leqslant \mu_{\rm GaN} = \Delta H({\rm GaN}) \qquad (19)$$

where the calculated GaN formation enthalpy $\Delta H({\rm GaN}) = -1.57$ eV. Because of the secondary phase formation, the upper limit for $\mu_{\rm N}$ is not $\mu_N^{max} = 0$ as suggested by equation (17), but $\mu_N^{max} = -1.57 + 0.62 = -0.95$ eV $\ll 0$. (In deriving μ_N^{max}, we have used $\mu_{\rm Ga}^{min} = \Delta H$ (GaAs) $- \mu_{\rm As}^{max} = -0.62$ eV.) These conditions define the 'original region' in figure 9. Nitrogen substitution is a special case of equation (16):

$$\Delta H_{sub} = C - \mu_{\rm N} + \mu_{\rm As}, \qquad (20)$$

where $C = E_{tot}(N_{\rm As}) - E_{tot}(0) - E_{tot}({\rm N_2}$ molecule$)/2 + E_{tot}$ (As solid). The higher the $\mu_{\rm N}^{max}$ (and the lower the $\mu_{\rm As}^{min}$), the lower the minimum ΔH_{sub}^{min} is. The calculated ΔH_{sub}^{min} is 1.64 eV $\gg KT_{growth} \sim 0.1$ eV, which accounts for the low [N] in equilibrium-grown GaAs:N.

(b) Surface-enhanced solubility

In epitaxial growth, the relaxed GaN phase may not form even if $\mu_{\rm N}$ exceeds $\mu_{\rm N}^{max}$. This is because the secondary phase formation typically requires

(1) the accumulation of a GaN layer on the surface and
(2) that the energy of the accumulated layer is high enough to facilitate the nucleation of dislocations.

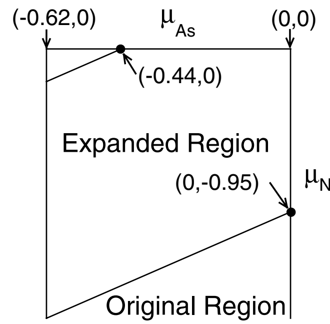

Figure 9. The physically accessible region of the chemical potentials, $(\mu_{\text{As}}, \mu_{\text{N}})$, is shown for GaAs:N. The 'original region' is defined by equations (18) and (19), while the 'expanded + original' region is defined by equations (18) and (21) [95].

In other words, the maximum N chemical potential is set not by equation (19) but by

$$\mu_{\text{Ga}} + \mu_{\text{N}} \leqslant \mu_{\text{N-}rich}^{Surface}, \tag{21}$$

where $\mu_{\text{N-}rich}^{Surface}$ is an effective formation enthalpy of the surface GaN layer. Note that this surface layer may not have the bulk structure of GaN. When μ_{N} reaches this surface-determined μ_{N}^{max}, exchanging N at the surface substitutional site with the N reservoir should cost no energy. In other words, $\Delta H_{sub}^{Surface}$ equals zero. Thus, to find nitrogen solubility in epitaxial GaAs:N films, one simply solves

$$\Delta H_{sub}^{Surface}(\mu_{\text{As}}^{min}, \mu_{\text{N}}^{max}) = 0 \tag{22}$$

to find $(\mu_{\text{As}}^{min}, \mu_{\text{N}}^{max})$. One then plugs this set of $(\mu_{\text{As}}^{min}, \mu_{\text{N}}^{max})$ into equation (20) to determine ΔH_{sub}^{min} for substitutional N *inside* the bulk.

Figure 10 shows the results along the N- and As-rich boundaries in figure 9 calculated by Zhang and Wei [95]. A (001)-(2 × 4) surface reconstruction (inset of figure 10) was used. It shows ΔH_{sub} for the surface (labelled as 1 and 3D, respectively), subsurface (3C) and bulk sites, respectively. Between the surface sites, site 1 has lower energy. When ΔH_{sub} (site1) = 0, $(\mu_{\text{As}}^{min}, \mu_{N}^{max}) = (-0.44\ \text{eV}, 0.0)$, at which ΔH_{sub}^{min} (bulk) is significantly reduced from the original value of 1.64 eV to 0.24 eV. Using Boltzmann statistics, this leads to a nitrogen concentration [N] $\sim$ 4% at 650 °C. The corresponding epitaxially accessible atomic chemical potential region $(\mu_{\text{As}}, \mu_{\text{N}})$ is the (expanded + original) region in figure 9.

6.2. Uncovering the mystery of p-type ZnO

Zinc oxide (ZnO) is an emerging material for short wavelength optoelectronic devices [96–98]. To finally realize its device applications, however, an important issue is to fabricate both high quality p-type and n-type ZnO films. As discussed in section 4, ZnO can be easily doped high quality n-type, but is difficult to dope p-type. Nitrogen, a good p-type dopant for other II–VI semiconductors [99, 100], has long been considered as a possible candidate for p-type ZnO [101], but various efforts to realize this goal have been frustrated [102, 103]. Among other things, low N solubility has contributed to the inability to achieve p-type ZnO.

Recently, Joseph *et al* [79] reported that one could get p-type ZnO but either with high hole concentration and poor mobility or with low hole concentration and high mobility. The realization of the high hole concentration ZnO films was attributed to a 'codoping' phenomenon [78] that requires simultaneous presence of two dopants, in this case Ga

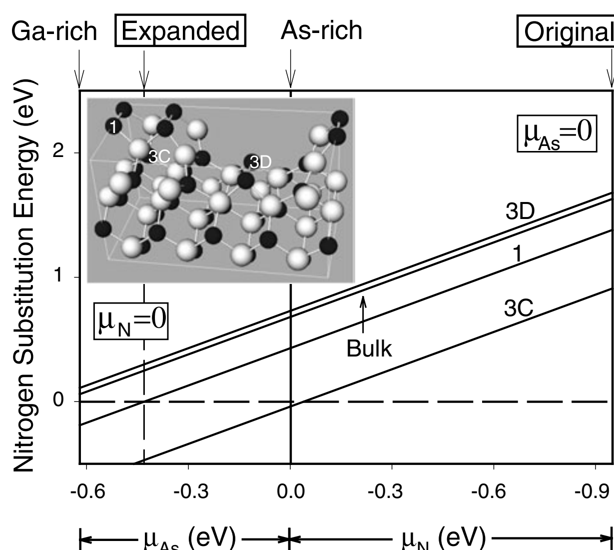

Figure 10. The nitrogen substitutional energy ΔH_{sub} in GaAs:N is shown as a function of (μ_{As}, μ_N). The inset is the GaAs(2 × 4) surface, indicating the various substitutional sites. Black circles denote the anion atoms whereas light circles denote the Ga atoms [95].

and N. The concept of codoping, however, has so far not been supported by first-principles calculations [104]. On the other hand, previous discussion on the surface-enhanced N solubility in GaAs (see section 6.1) suggests that the chemical potential of the dopant could be significantly increased, hence so would be the dopant solubility, in a non-equilibrium growth process. This opens the door for engineering dopant sources—an area that has not attracted much attention in the past partly because there are no significant alternative dopant sources. Recently, Yan *et al* [22] pointed out that in the case of N doping of ZnO (or other *oxides*), there is indeed a unique and unusual opportunity. There are at least four different gases, namely N_2, NO, NO_2 and N_2O, that can be used, in addition to the electron-cyclotron-resonance (ECR) plasma source in [79]. If these molecules arrive intact at the growing surface, it is their respective chemical potentials that determine the doping efficiency. Clearly, the growth conditions need to be set up accordingly to ensure that the chosen species is available and arrives intact at the growing surface.

Under equilibrium conditions, the concentration of a point defect or an impurity is mainly related to its formation energy, which depends on the chemical potentials of the host and relevant impurity atoms as defined by the appropriate reservoirs (cf equation (4)). The chemical potentials of Zn and O satisfy $\mu_{Zn} + \mu_O = \Delta H(\text{ZnO})$ (cf equation (6)), where $\Delta H(\text{ZnO})$ is the formation enthalpy of ZnO and $\Delta H(\text{ZnO}) \leqslant \mu_O \leqslant 0$. For the present purposes, the relevant defects are the substitutional N (N_O) and $N_2[(N_2)_O]$ both at the O sites. N_O is an acceptor and $(N_2)_O$ is a double donor. In the final analysis, *it is the relative concentrations of these two defects that control the doping type*.

In the paper by Yan *et al* [22], they drew a sharp distinction between the two gases used by Joseph *et al* [79], namely N_2 and N_2O, and the two alternative gases, namely NO and NO_2. The molecules of the first two gases contain *pairs* of N atoms, whereas the alternative gases contain *single* N atoms. Let us first consider the formation of the desirable N_O defects. It is clear that N_2 and N_2O are not the best options because energy must be supplied to break their N–N bonds. In contrast, NO and NO_2 molecules can be incorporated directly in the growth surface to form

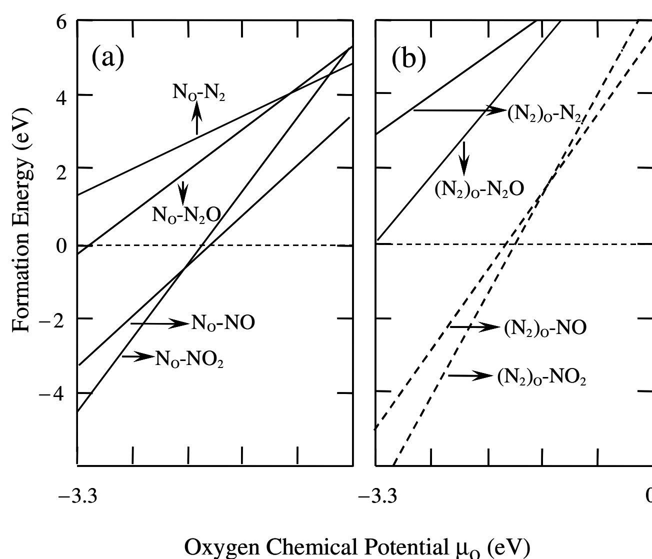

Figure 11. The calculated defect formation energies as a function of the oxygen chemical potential, μ_O, for (a) N_O and (b) $(N_2)_O$ in ZnO: $\mu_O = -3.3$ eV corresponds to the Zn-rich limit whereas $\mu_O = 0$ corresponds to the O-rich limit. In both cases, the effects of the different N atomic chemical potential (i.e. from gas-phase N_2, N_2O, NO and NO_2 molecules) are depicted [22].

N_O defects, taking advantage of the fact that O atoms also take part in the growth process as host atoms. They arrived at the same conclusion when examining the formation of the undesirable $(N_2)_O$ defect. Now it is N_2 and N_2O that can be incorporated directly in the growth surface to form $(N_2)_O$. In contrast, NO and NO_2 molecules provide single N atoms and can produce $(N_2)_O$ centres only by the less likely process of two N atoms arriving simultaneously at the same site on the growth surface (once a single N_O is incorporated into the bulk, it is unlikely that it can be found by a diffusing N atom because the migration energy is high, 3 eV).

Yan *et al* [22] presented the numerical results and a detailed analysis for the production of the desirable N_O defects. Figure 11(a) shows the formation energy of N_O. The Fermi energy is assumed to be at the VBM corresponding to optimal p-type doping conditions. The difference between N_2/N_2O and NO/NO_2 is very clear, i.e. the use of NO/NO_2 leads to significantly reduced formation energies for N_O because it does not entail an energy supplement to break the N–N bonds. The negative formation energies of N_O at the Zn-rich conditions indicate that NO or NO_2 molecules can be incorporated spontaneously to form the N_O defects. For comparison purposes, figure 11(b) shows the formation energies of $(N_2)_O$. The solid lines control the formation of $(N_2)_O$ when N_2 or N_2O is used. It is clear that if N_2O gas is used and the molecules arrive intact at the growing surface, a high concentration of $(N_2)_O$ centres will be observed at the Zn-rich conditions. The dashed lines in figure 11(b) show that when NO or NO_2 molecules are used, the formation energies of $(N_2)_O$ are higher than N_O (see figure 11(a)) at the O-rich conditions, but lower at the Zn-rich conditions. The lower formation energies of $(N_2)_O$ at the zinc-rich conditions indicate clearly that a non-equilibrium doping process is essential to achieve p-type ZnO by nitrogen.

These results allow for a discussion of the experimental data [79]. First, note that the growth conditions are Zn rich because when Zn and O atoms are ablated from the target, O atoms may form O_2 molecules through simple collisions, whereas Zn atoms would have to nucleate Zn metal somewhere in the chamber and have other Zn atoms find that nucleus for continued

precipitation (the calculated binding energies of Zn_2 and O_2 molecules are 0.2 and 6.9 eV, respectively, showing that Zn atoms do not form molecules at the temperature of interest).

(1) N_2O gas without electron-cyclotron-resonance (ECR) plasma

Under these conditions, n-type ZnO films with very high carrier concentrations (4.5×10^{20} cm^{-3}) were obtained. Figure 11(b) shows that the formation energy of $(N_2)_O$ is nearly zero at the Zn-rich limit, which accounts for the very high level of n-type doping. Without ECR, the N_2O molecules do not crack in the gas phase, and therefore do not form N_O defects.

(2) N_2O gas with ECR plasma on

Under this condition, p-type ZnO films with very low carrier concentration (2.0×10^{10} cm^{-3}), but high mobility (1.9×10^3 cm^2 V^{-1} s^{-1}), were obtained. In this case, the ECR plasma provides energy that drives the reaction $N_2O \leftrightarrow NO + N$ in the *forward* ($\rightarrow$) direction. Because the energy for breaking an NO molecule (6.6 eV) is 2.2 eV higher than that for breaking an N_2O molecule (4.4 eV) [105], further dissociation of NO can generally be neglected [106]. Once the reaction products, NO and N, enter the chamber, the reaction $N_2O \leftrightarrow NO + N$ would go in reverse ($\leftarrow$). In addition, N atoms would also form N_2 molecules. The net result is a mixture of NO, N_2 and N_2O. The present results show that NO will introduce low formation energy N_O, whereas N_2O will introduce low formation energy $(N_2)_O$. A calculation of the branching ratios at the temperature of interest is needed in order to tell which concentration would be higher. The experimental data show high resistivity and slightly p-type doping, which suggests that the branching ratios are such that NO wins slightly.

(3) N_2O gas with ECR plasma source and additional Ga source

Under these conditions, p-type material with very high carrier concentration (4×10^{19} cm^{-3}), but very low mobility (0.07 cm^2 V^{-1} s^{-1}), was obtained. This is similar to the previous case. In this case, however, when NO and N entered the chamber, Ga was also found. N and Ga atoms can form GaN molecules (binding energy is 3.4 eV), reducing the concentration of N that could reconvert NO to N_2O. Consequently more NO molecules survive and reach the growing surface, leading to p-type ZnO with high carrier concentration. Clearly, the wt% of Ga_2O_3 in the ZnO target is important because it controls the branching ratios. The poor mobility is probably the result of GaN precipitates in the film.

(4) N_2 gas with ECR plasma on

Under these conditions, n-type material with high carrier concentration (1.3×10^{19} cm^{-3}), but low mobility (1.3 cm^2 V^{-1} s^{-1}), was obtained. Because the dissociation energy for N_2 molecules, 9.9 eV, is 5.5 eV higher than that for N_2O molecules (4.4 eV) [105], the ECR that cracks N_2O molecules cannot crack N_2 molecules. However, the ECR source can transfer energy to N_2 molecules and change the chemical potential of the molecules. In this case, the effective chemical potential of N_2 is $\mu'_{N_2} = \mu_{N_2} + \Delta E$, where μ_{N_2} is the chemical potential of N_2 molecules without the ECR source and ΔE is the energy transferred from the ECR source. In the experiment of Joseph *et al* [79], the ECR can crack the N_2O molecules. Assuming that ΔE is close to the dissociation energy of N_2O molecules, i.e. 4.4 eV, the formation energy of $(N_2)_O$ from the N_2 molecules (shown in figure 11(b)) should be shifted downward about 4.4 eV, giving a negative formation energy at the Zn-rich conditions. This explains the growth of highly n-type ZnO under these conditions.

7. Summary

In summary, I have discussed the various doping difficulties in semiconductors and wide-gap materials. A number of important issues regarding first-principles total energy defect calculations are also addressed. I then reviewed the recent developments in first-principles total energy calculations of the phenomenological equilibrium 'doping limit rule', addressing its microscopic origin. In particular, the calculations establish the connection between the n-type 'doping limits' and the spontaneous formation of native defects, namely, the cation vacancies. New directions to overcome the equilibrium doping limits are suggested and discussed. These include the isovalent doping of GaAs by nitrogen and the p-type doping of ZnO by nitrogen. The importance of significantly increasing the maximum impurity chemical potential is emphasized.

Acknowledgments

I thank S-H Wei, Y Yan, S T Pantelides and A Zunger for their invaluable contributions in the course of these studies. This work was supported by the US DOE-SC-BES under contract no DE-AC36-99GO10337.

References

[1] Pearton S J, Ren F, Zhang A P and Lee K P 2000 *Mater. Sci. Eng.* R **30** 55
[2] Neumark G F 1997 *Mater. Sci. Eng.* R **21** 1
[3] Lang D V and Logan R A 1977 *Phys. Rev. Lett.* **39** 635
[4] Chadi D J and Chang K J 1988 *Phys. Rev. Lett.* **61** 873
[5] Chadi D J and Chang K J 1989 *Phys. Rev.* B **39** 10 063
[6] Zhang S B and Chadi D J 1990 *Phys. Rev.* B **42** 7174
[7] Zhang S B 1991 *Phys. Rev.* B **44** 3417
[8] Chadi D J 1994 *Phys. Rev. Lett.* **72** 534
[9] Wei S H, Zhang S B and Zunger A 1993 *Phys. Rev. Lett.* **70** 1639
[10] Chadi D J 1999 *Phys. Rev.* B **59** 15 181
[11] Park C H and Chadi D J 1995 *Phys. Rev. Lett.* **75** 1134
[12] Park C H and Chadi D J 1995 *Phys. Rev.* B **52** 11 884
[13] Park C H and Chadi D J 1996 *Phys. Rev.* B **54** R14 246
[14] Park C H and Chadi D J 1997 *Phys. Rev.* B **55** 12 995
[15] Chadi D J, Citrin P H, Park C H, Adler D L, Marcus M A and Gossmann H-J 1997 *Phys. Rev. Lett.* **79** 4834
[16] Van de Walle C G 2000 *Phys. Rev. Lett.* **85** 1012
[17] Zhang S B and Chadi D J 1990 *Phys. Rev.* B **41** 3882
[18] Reboredo F A and Pantelides S T 1999 *Phys. Rev. Lett.* **82** 1887
[19] Park C H, Zhang S B and Wei S-H 2002 *Phys. Rev.* B **66** at press
[20] Kwak K W, Vanderbilt D and King-Smith R D 1994 *Phys. Rev.* B **50** 2711
Kwak K W, Vanderbilt D and King-Smith R D 1995 *Phys. Rev.* B **52** 11 912
[21] Cheong B-H, Park C H and Chang K J 1995 *Phys. Rev.* B **51** 10 610
[22] Yan Y, Zhang S B and Pantelides S T 2001 *Phys. Rev. Lett.* **86** 5723
[23] Baraff G A and Schlüter M 1985 *Phys. Rev. Lett.* **55** 1327
[24] Baraff G A and Schlüter M 1986 *Phys. Rev.* B **33** 7346–8
[25] Jansen R W and Sankey O F 1989 *Phys. Rev.* B **39** 3192
[26] Zhang S B and Northrup J E 1991 *Phys. Rev. Lett.* **67** 2339
[27] Northrup J E and Zhang S B 1993 *Phys. Rev.* B **47** 6791
[28] Northrup J E and Zhang S B 1994 *Phys. Rev.* B **50** 4962
[29] Garcia A and Northrup J E 1995 *Phys. Rev. Lett.* **74** 1131
[30] Ramamoorthy M and Pantelides S T 1996 *Phys. Rev. Lett.* **76** 4753
[31] Poykko S, Puska M J and Nieminen R M 1998 *Phys. Rev.* B **57** 12 174

[32] Tsur Y and Riess I 1999 *Phys. Rev.* B **60** 8138
[33] Neumark G F 1989 *Phys. Rev. Lett.* **62** 1800
[34] Laks D B, Van de Walle C G, Neumark G F, Blochl P E and Pantelides S T 1992 *Phys. Rev.* B **45** 10 965
[35] Wei S-H and Zhang S B 2002 *Phys. Status Solidi* b **229** 305
[36] Van de Walle C G, Stampfl C and Neugebauer J 1998 *J. Cryst. Growth* **189/190** 505
[37] Van de Walle C G, Limpijumnong S and Neugebauer J 2001 *Phys. Rev.* B **63** 245205
[38] Kroger F A 1973 *The Chemistry of Imperfect Crystals* 2nd edn (Amsterdam: North-Holland)
[39] Hamann D R, Schlüter M and Chiang C 1979 *Phys. Rev. Lett.* **43** 1494
[40] Troullier N and Martins J L 1991 *Phys. Rev.* B **43** 1993
[41] Vanderbilt D 1990 *Phys. Rev.* B **41** 7892
[42] Ihm J, Zunger A and Cohen M L 1979 *J. Phys. C: Solid State Phys.* **12** 4409
[43] Wei S-H and Krakauer H 1985 *Phys. Rev. Lett.* **55** 1200
[44] Singh D J 1994 *Planewaves, Pseudopotentials, and the LAPW Methods* (Boston, MA: Kluwer)
[45] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[46] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[47] Perdew J P and Wang Y 1992 *Phys. Rev.* B **45** 13 244
[48] Chadi D J and Cohen M L 1973 *Phys. Rev.* B **8** 5747
[49] Monkhorst H J and Pack J D 1976 *Phys. Rev.* B **13** 5188
[50] Van de Walle C G, Denteneer P J H, Bar-Yam Y and Pantelides S T 1989 *Phys. Rev.* B **39** 10 791
[51] Ogut S and Chelikowsky J R 1999 *Phys. Rev. Lett.* **83** 3852
[52] Makov G and Payne M C 1995 *Phys. Rev.* B **51** 4014
[53] Zhang S B, Wei S-H, Zunger A and Katayama-Yoshida H 1998 *Phys. Rev.* B **57** 9642
[54] Zhang S B and Wei S-H 2002 *Appl. Phys. Lett.* **80** 1376
[55] Ziman J M 1972 *Principles of the Theory of Solids* 2nd edn (Cambridge: Cambridge University Press) p 275
[56] Hybertsen M S and Louie S G 1986 *Phys. Rev.* B **34** 5390
[57] Zhang S B, Tomanek D, Cohen M L, Louie S G and Hybertsen M S 1989 *Phys. Rev.* B **40** 3162
[58] Pashley M D 1989 *Phys. Rev.* B **40** 10 481
[59] Janotti A, Zhang S B, Wei S-H and Van de Walle C G 2002 *Phys. Rev. Lett.* **89** at press
[60] Caldas M, Fazzio A and Zunger A 1984 *Appl. Phys. Lett.* **45** 671
[61] Langer J M and Heinrich H 1985 *Phys. Rev. Lett.* **55** 1414
[62] Walukiewicz W 1989 *Phys. Rev.* B **39** 8776
[63] Walukiewicz W 1993 *Mater. Res. Soc. Symp. Proc.* **300** 421
[64] Walukiewicz W 1994 *Inst. Phys. Conf. Ser.* **141** 259
[65] Walukiewicz W 2001 *Physica* B **302** 123
[66] Tokumitsu E 1990 *Japan. J. Appl. Phys.* **29** L698
[67] Ferreira S O, Sitter H, Faschinger W, Krump R and Brunthaler G 1995 *J. Cryst. Growth* **146** 418
[68] Zhang S B, Wei S-H and Zunger A 1998 *J. Appl. Phys.* **83** 3192
[69] Van Vechten J A 1980 *Handbook of Semiconductors* vol 3, ed S P Keller (Amsterdam: North-Holland) p 1
[70] Kittel C 1986 *Introduction to Solid State Physics* 6th edn (Singapore: Wiley) pp 202–3
[71] Wei S-H and Zunger A 1998 *Appl. Phys. Lett.* **72** 2011
[72] Zhang S B, Wei S-H and Zunger A 2000 *Phys. Rev. Lett.* **84** 1232
[73] Kawazoe H, Yasukawa M, Hyodo H, Kurita M, Yanagi H and Hosono H 1997 *Nature* **389** 939
[74] Yanagi H, Inoue S, Ueda K and Kawazoe H 2000 *J. Appl. Phys.* **88** 4159
[75] Ueda K, Hase T, Yanagi H, Kawazoe H, Hosono H, Ohta H, Orita M and Hirano M 2001 *J. Appl. Phys.* **89** 1790
[76] Yanagi H, Hase T, Ibuki S, Ueda K and Hosono H 2001 *Appl. Phys. Lett.* **78** 1583
[77] Nie X, Wei S-H and Zhang S B 2002 *Phys. Rev. Lett.* **88** 066405
[78] Yamamoto T and Katayama-Yoshida H 1999 *Japan. J. Appl. Phys.* **38** L166
[79] Joseph M, Tabata H and Kawai T 1999 *Japan. J. Appl. Phys.* **38** L1205
[80] Korotkov R Y, Gregie J M and Wessels B W 2001 *Appl. Phys. Lett.* **78** 222
[81] Katayama-Yoshida H, Nishimatsu T, Yamamoto T and Orita N 2001 *J. Phys.: Condens. Matter* **13** 8901
[82] Vailionis A, Glass G, Desjardins P, Cahill D G and Greene J E 1999 *Phys. Rev. Lett.* **82** 4464
[83] Glass G, Kim H, Desjardins P, Taylor N, Spila T, Lu Q and Greene J E 2000 *Phys. Rev.* B **61** 7628
[84] Weyers M, Sato M and Ando H 1992 *Japan. J. Appl. Phys.* **31** L853
[85] Neugebauer J and Van de Walle C G 1995 *Phys. Rev.* B **51** 10 568
[86] Wei S-H and Zunger A 1996 *Phys. Rev. Lett.* **76** 664
[87] Ho I-H and Stringfellow G B 1997 *J. Cryst. Growth* **178** 1
[88] Zhang S B and Zunger A 1997 *Appl. Phys. Lett.* **71** 677

[89] Wolford D J, Bradley J A, Fry K and Thompson J 1984 *Proc. 17th Int. Conf. on the Physics of Semiconductors* ed J D Chadi and W A Harrison (New York: Springer) p 627
[90] Weyers M and Sato M 1993 *Appl. Phys. Lett.* **62** 1396
[91] Kondow M, Uomi K, Kitatani T, Watahiki S and Yazawa Y 1996 *J. Cryst. Growth* **164** 175
[92] Qiu Y, Nikishin S A, Temkin H, Faleev N N and Kudriavtsev Y A 1997 *Appl. Phys. Lett.* **70** 3242
[93] Bi W G and Tu C W 1997 *Appl. Phys. Lett.* **70** 1608
[94] Uesugi K and Suemune I 1998 *J. Cryst. Growth* **189/190** 490
[95] Zhang S B and Wei S-H 2001 *Phys. Rev. Lett.* **86** 1789
[96] Tuttle J R *et al* 1995 *Prog. Photovoltaics Res. Appl.* **3** 235
[97] Delahoy A E and Cherny M 1996 *Mater. Res. Soc. Symp. Proc.* **426** 467
[98] Yu P *et al* 1996 *Proc. 23rd Int. Conf. on Physics of Semiconductors (Berlin)* vol 2, p 1453
[99] Park R M *et al* 1990 *Appl. Phys. Lett.* **57** 2127
[100] Ohkawa K, Karasawa T and Mitsuyu T 1991 *Japan. J. Appl. Phys.* **30** L152
[101] Kobayashi A, Sankey O F and Dow J D 1983 *Phys. Rev.* B **28** 946
[102] Sato Y and Sato S 1996 *Thin Solid Films* **281/282** 445
[103] Minegishi K *et al* 1997 *Japan. J. Appl. Phys.* **36** L1453
[104] Zhang S B, Wei S-H and Yan Y 2001 *Physica* B **302/303** 135
[105] Cleland T A and Hess D W 1989 *J. Electrochem. Soc.* **136** 3103
[106] Date L *et al* 1999 *Surf. Coatings Technol.* **116–119** 1042