ELSEVIER

Computational Materials Science 30 (2004) 337–348

COMPUTATIONAL MATERIALS SCIENCE

www.elsevier.com/locate/commatsci

Overcoming the doping bottleneck in semiconductors

Su-Huai Wei *

National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, CO 80401, USA

Received 12 December 2003; accepted 1 February 2004

Abstract

Application of semiconductors as electric and optoelectronic devices depends critically on their dopability. Failure to dope a material, i.e., to produce enough free charge carriers beyond a certain limit, is often the single most important bottleneck for advancing semiconductor-based high technology. Using the first-principles band structure method, we have studied systematically the general chemical trends of the defect formation and ionization in semiconductors to understand the physical origin of the doping difficulty. New approaches to overcoming the doping limit have been developed. This paper reviews our recent progress and discusses some of the computational issues in defect calculations.

PACS: 61.72.Bb; 61.72.Ji; 61.72.Vv
Keywords: Semiconductor; Doping; Band structure; Theory

1. Introduction

The application of semiconductors as novel electronic devices depends critically on their dopability. Many semiconductor compounds will not be very useful as an electronic material if they cannot be doped [1]. Furthermore, many optoelectronic device applications also require bipolar doping, that is, the ability to dope efficiently a materials both n-type and p-type. However, extensive experimental and theoretical studies show that doping polarity, that is, a material can be doped either n-type or p-type, but not both, exists in many semiconductors, especially in wide-gap semiconductors. For example, it is well known that p-type doping in oxides and nitrides is very difficult [1–3], whereas efficient n-type doping is very difficult to achieve in ZnTe [4,5] and diamond [6]. These doping difficulties have so far hindered the full utilization of these semiconductors as novel electronic materials. It is, therefore, important to understand what causes the doping bottleneck and how to overcome the doping difficulty.

Generally speaking, three main factors could limit dopability. (i) The desired dopant may have a low solubility. (ii) The desired dopant has good solubility, but the defect transition energy level may be too deep, thus, the defect is not ionized at normal operating temperature. (iii) The desired dopant has good solubility and is ionizable, but as

* Tel.: +1-303-3846666; fax: +1-303-3846531.
E-mail address: swei@nrel.gov (S.-H. Wei).

0927-0256/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.commatsci.2004.02.024

the Fermi energy shifts due to the increased carrier density, oppositely charged native defects or defect complexes of the dopant (e.g., DX centers [7] or AX centers [8]) could form, thus limiting further change of the Fermi energy.

Using the first-principles band structure method, we have systematically calculated the defect formation energies and transition energy levels of intrinsic and extrinsic defects and defect complexes in various semiconductors [9–17]. General chemical trends of the defect formation in semiconductors were identified, which enabled us to understand the origin of the "doping limit rule" [9,18] and develop strategies to overcome the doping limit [10,12,14–17]. For example, we show that dopant solubility can be enhanced significantly by expanding the physically accessible range of the dopant atomic chemical potentials. This can be done by using either epitaxial growth techniques [10], such as MBE, or metastable molecules as doping sources [19]. Dopant ionization energy can also be reduced significantly by combining a symmetry-compatible donor with another donor [16] or an acceptor with another acceptor [20] that can form stable defect complexes. Furthermore, we show that bipolarly dopable transparent conducting oxides can also be achieved by exploiting the large disparity between the optical and fundamental band gaps in some of the oxides [11].

The rest of the paper is organized as follows. Section 2 describes salient features of the defect calculation method. Section 3 discusses the effects of chemical potential and the origin of the doping limit rule. Section 4 explains how to determine symmetry and occupation of the defect levels. Sections 5–8 describe strategies to enhance the dopant solubility. Section 9 investigate the effectiveness of co-doping, Sections 10 and 11 describe the strategies for lowering the defect ionization energy, and Section 12 describes how to modify the band structure to improve the dopability. The last section gives a brief summary of the paper.

2. Methods of calculation

In this section, I will discuss some salient computational issues involved in defect calculation. In modern first-principles defect calculation, the defect system is modeled by putting a defect or defect complex at the center of a periodic supercell. For a charged defect, a uniform background charge is added to keep the global charge neutrality of the periodic supercell. All internal structural parameters of the supercell are optimized by minimizing the total energy and quantum mechanical forces. To determine the defect formation energy and defect transition energy levels, we calculate the total energy $E(\alpha, q)$ for a supercell containing the relaxed defect α in charge state q. We also calculate the total energy E(host) of the host for the same supercell in the absence of the defect, as well as the total energies of the involved elemental solids or gases at their stable phases. From these quantities, we can deduce the defect formation energy $\Delta H_f(\alpha, q)$ as a function of the electron Fermi energy [21] E_F as well as on the atomic chemical potentials [22,23] μ_i:

$$\Delta H_f(\alpha, q) = \Delta E(\alpha, q) + \sum_i n_i \mu_i + q E_F, \tag{1}$$

where $\Delta E(\alpha, q) = E(\alpha, q) - E(\text{host}) + \sum_i n_i E(i) + q\epsilon_{VBM}$ (host). E_F is referenced to the valence band maximum (VBM) of the host. μ_i is the chemical potential of constituent i referenced to elemental solid/gas with energy $E(i)$. n_i is the number of elements, and q is the number of electrons transferred from the supercell to the reservoirs in forming the defect cell. The defect transition energy level $\epsilon_\alpha(q/q')$ is the Fermi energy E_F in Eq. (1), at which the formation energy $\Delta H_f(\alpha, q)$ of defect α and charge q is equal to that of another charge q' of the same defect, i.e.,

$$\epsilon_\alpha(q/q') = [\Delta E(\alpha, q) - \Delta E(\alpha, q')]/(q' - q). \tag{2}$$

The band structure and total energy calculations are performed using the local density approximation [24,25] or the generalized gradient approximation (GGA) [26] as implemented by either the all-electron, full-potential, linearized augmented plane wave (FLAPW) method [27] or using the pseudopotential approach [28,29]. The Brillouin zone (BZ) integration for the charge density and total energy calculations is performed using the special k-points [30] or equivalent k-points in the superstructures [31]. Several k-point sampling

methods have been tested to calculate the formation energy of the charged defect and transition energy level. In the first approach the transition energy level is determined by using the total energy calculated with the special k-points [30]. In this case, to be consistent, the energy of the band edges (VBM and the conduction band minimum (CBM)), thus, the band gap, is determined by the average over the same k-points where the total energy is calculated. The advantage for this approach is that the calculation is consistent and, thus, straightforward. By choosing the appropriate k-point sampling, it also partially corrects the LDA band gap error. This approach has been widely used in many studies [15–17,32] and is efficient in correcting the LDA band gap error when the cell size is not large. However, the disadvantage for this approach is that the averaged band edges could depend on the cell size, k-point sampling and the charge state. To avoid these problems, one may have the tendency to use only the Γ point to determine the total energy and transition energy level because the symmetry of the defect and the band gap are well defined at Γ, and both the shallow and deep levels are correctly described. However, the Γ point-alone approach is quite poor in obtaining the converged charge density and total energy if the cell size is not large [33].

We have developed a new mixed scheme to combine the advantages of the special k-point and Γ point-only approaches. In this new scheme, the formation energy of the charged state is given by

$$\begin{aligned}\Delta H_{\mathrm{f}}(\alpha,q) &= \Delta H_{\mathrm{f}}(\alpha,0)\\ &\quad + [E(\alpha,q) - (E(\alpha,0) - q\epsilon_{\mathrm{D}}^{\mathbf{k}}(0))]\\ &\quad - q[\epsilon_{\mathrm{D}}^{\Gamma}(0) - \epsilon_{\mathrm{VBM}}^{\Gamma}(\mathrm{host})] + qE_{\mathrm{F}}, \end{aligned} \tag{3}$$

where the electron/structural relaxation energy (second term on the right hand side) is determined using the special k-points, whereas the single electron defect energy levels with respect to the VBM (third term on the right-hand side) as well as the Fermi energy $E_{\mathrm{F}} = \epsilon_{\mathrm{F}} - \epsilon_{\mathrm{VBM}}^{\Gamma}$ (host) (last term on the right hand side) are evaluated at the Γ point and are aligned using core electron energy levels or average potentials away from the defect. Here, $\epsilon_{\mathrm{D}}^{k}(0)$ and $\epsilon_{\mathrm{D}}^{\Gamma}(0)$ are the defect level at the special k-points (averaged) and at Γ-point, respectively. This approach has been used successfully for studying defects in CdTe [14] and other systems in which the LDA band gap error is already corrected.

It has been argued [34] that the periodic boundary conditions adopted for the supercell may introduce spurious Coulomb interaction between the charged defects in different cells. To estimate the magnitude of this interaction [34], point charges immersed in neutralizing jellium are usually assumed [35]. Attempts to include higher-order terms is difficult, because the higher-order multipoles of the defect charge are not uniquely defined in the supercell approach [36,37]. However, in reality, the charged defect does not have a point charge- (delta-function-) like distribution, especially for shallow defects, which have a relatively uniform charge distribution. Therefore, direct application of the Makov and Payne correction using only point charge [35] often overestimates the effect. To demonstrate this, we have calculated the unscreened Coulomb interaction energy E_{Coul} of a periodic defect charge on a simple cubic lattice site, separated by a distance of L and immersed in a uniform jellium background [15]. The self-interaction of the defect charge is removed. We assume that the defect charge has a normalized Gaussian distribution of the form

$$\rho(r) = \frac{\alpha^3}{\pi^{3/2}} \mathrm{e}^{(-\alpha^2 r^2)}, \tag{4}$$

where $1/\alpha$ defines the width of the charge distribution. In this case, the limit $\alpha \to \infty$ corresponds to the point-charge model in [34], and for $\alpha \to 0$, the charge distribution is fully delocalized. The calculated E_{Coul}, which contains all the high-order interactions, as a function of α is shown in Fig. 1. We see that for $\alpha \to 0$, the Coulomb energy approaches zero. For most defects where $\alpha \ll 0.1$ a.u.$^{-1}$, the Coulomb energy is much smaller than what is expected from a point-charge model [34,35], shown as an horizontal line in Fig. 1. Thus, we assume the Makov and Payne correction are usually not important for most calculations, especially for shallow states.

To check the convergence of the cell size, we systematically increased the cell size and compared the calculated results with the extrapolated results to infinite dimension by assuming a linear

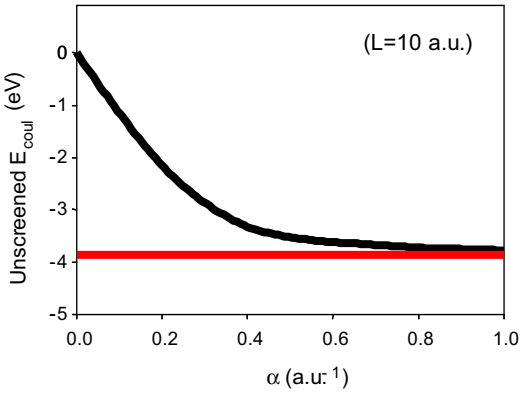

Fig. 1. Unscreened Coulomb interaction energy (in eV) as a function of the defect charge distribution α. The horizontal line gives the value for point charge model. See text for details.

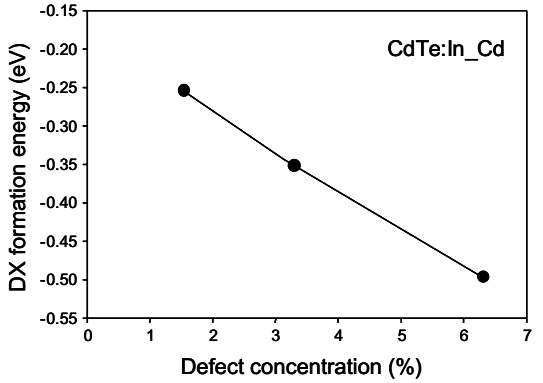

Fig. 2. Calculated DX formation energy of In_{Cd} in CdTe as a function of defect concentration in the supercell approach. The solid dots are the calculated values that are connected by straight lines.

dependence of the defect formation energy on the defect concentration. Fig. 2 shows the calculated formation energy of In_{Cd} DX center in CdTe. We see that the calculated results follow the linear function quite well, indicating that the extrapolated results are fairly well converged.

3. Chemical potential dependence and the doping limit rule

Eq. (1) indicates that the defect formation energy, and consequently, the solubility of the dopants, depend sensitively on the atomic chemical potential, as well as on the electron Fermi energy. This is because in forming the defect, particles are exchanged between the host and chemical reservoirs. For example [14], to form the substitutional defect Na_{Cd} in CdTe one has to take a Na atom from the Na chemical reservoir, put it into the host, and remove a Cd atom from the host and put it into the Cd chemical reservoir. Therefore, the formation energy of Na_{Cd} decreases if the chemical potential of Cd decreases or if the chemical potential of Na increases. Furthermore, to form a positively charged defect $(q > 0)$, one has to put the electrons removed from the defect into an electron reservoir with its characteristic energy ϵ_F. Thus, the positively charged defect will have a higher formation energy in an n-type sample in which the Fermi energy is close to the CBM. Similarly, negatively charged defects $(q < 0)$ will have a higher energy in p-type material in which the Fermi energy is close to the VBM. Therefore, *by adjusting the chemical potential of the dopant or the Fermi energy, one can control the dopant solubility*.

A consequence of the Fermi energy dependence of the defect formation energy is the recently recognized doping limit rule [9,38], which states that materials in which the CBM is too high are difficult to dope n-type, whereas materials in which the VBM is too low are difficult to dope p-type. In other words, a good p-type semiconductor must have a sufficiently small work function, while a good n-type semiconductor must have a sufficiently large electron affinity. This simple rule indicates that the band gap value alone does not determine whether a material is dopable or not, as was previously thought [39]. What matters is not just the band gap, but the relative position of the band edges (the CBM and the VBM) [9]. For example, antimonides or tellurides, which have high VBM (Fig. 3) are easier to be doped p-type, whereas nitrides or oxides, which has very low VBM (Fig. 3) are difficult to be doped p-type. On the other hand, despite the large band gap, the CBM of ZnO is very low [40]. This induces heavy n-type dopability in ZnO. Note that the VBM of ZnO is also lower than GaN due to the large electronegativity of oxygen versus nitrogen; thus,

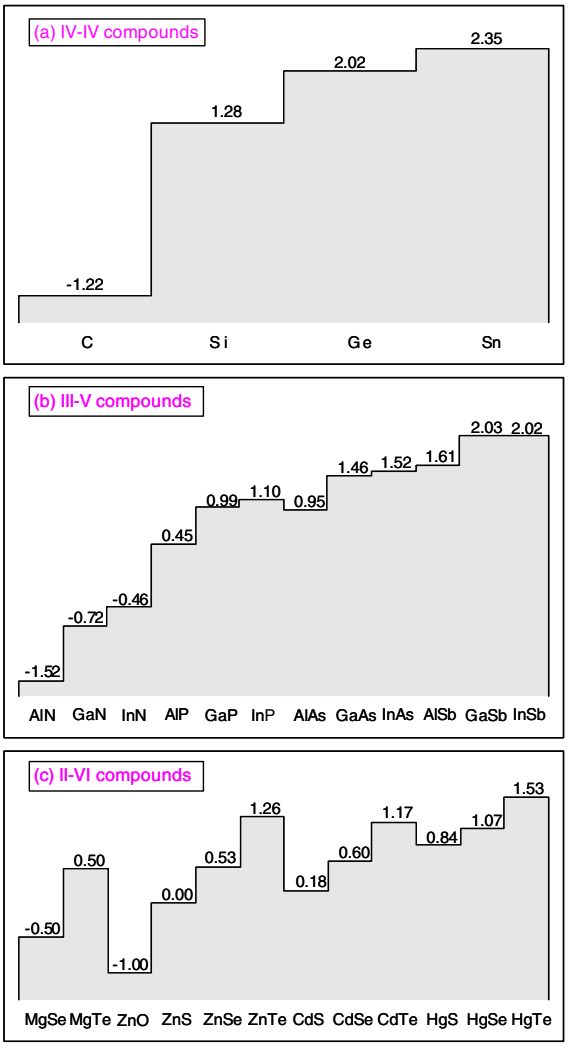

Fig. 3. Calculated natural valence band offsets of all group IV, III–V, and II–VI semiconductors in the diamond or zinc-blende structure.

despite ZnO and GaN having similar lattice constants and band gaps, it will be much more difficult to dope ZnO p-type than to obtain p-type GaN.

Further investigation has shown that the shift of the Fermi energy E_F in a semiconductor due to doping is bounded by E_{pin}^{p} below and E_{pin}^{n} above, and that the pinning energies $E_{pin}^{p,n}$ in a group of similar materials tend to line up [9,38]. The physical origin of this doping limit rule is traced to the spontaneous formation of intrinsic compensating defects as the Fermi energy shifts during the doping process. For example, n-type doping in semiconductors is usually limited by the formation of the compensating cation vacancy when the Fermi energy increases [18]. This also suggests that the dopability of a material depends on the ease of forming compensating intrinsic defects. *A material is easy* (*difficult*) *to be doped* p-*type if the CBM is high* (*low*) *in energy*. For example, diamond is relatively easier to be doped p-type than SiC, although diamond has relatively low VBM. This is because diamond has high CBM and is very difficult to be doped n-type. Another example is between GaN and InN, where p-type doping in InN is more difficult than in GaN, even though InN has higher VBM than GaN (Fig. 3). This is because InN has very low CBM [40,41], and thus can easily be doped n-type. Similar rules exist for n-type doping, i.e., *a material is easy* (*difficult*) *to be doped* n-*type if the VBM is low* (*high*) *in energy*.

It should be noted that under equilibrium growth conditions there are some thermodynamic limits on the achievable values of the chemical potentials μ_i: First, to avoid precipitation of the elemental dopant and the host elements, μ_i are bound by

$$\mu_i \leqslant 0. \tag{5}$$

Second, μ_i are limited to those values that maintain a stable host AC, so

$$\mu_A + \mu_C = \Delta H_f(\mathrm{AC}), \tag{6}$$

where $\Delta H_f(\mathrm{AC})$ is the formation energy of host AC. Finally, to avoid the formation of secondary phases between the dopants X and the host elements, μ_i is limited by

$$n\mu_A + m\mu_X \leqslant \Delta H_f(A_nX_m) \quad \text{or}$$
$$n\mu_X + m\mu_C \leqslant \Delta H_f(X_nC_m). \tag{7}$$

As an example, Fig. 4 plots the calculated chemical potential region accessible under equilibrium growth condition for CdTe:Na in the two-dimensional (μ_{Cd}, μ_{Na}) plane [14], as defined by Eqs. (5)–(7). It shows that because Na forms a very stable compound Na_2Te with Te (with a calculated formation energy of −2.84 eV), the highest possible μ_{Na} at the Cd-rich condition ($\mu_{Cd} = 0$, $\mu_{Te} = \Delta H_f(\mathrm{CdTe}) = -0.79$ eV) is −1.02 eV. Under the

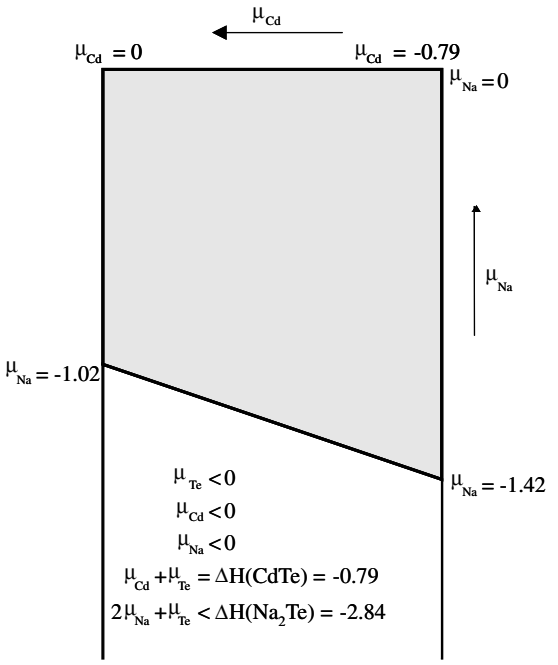

Fig. 4. Calculated available equilibrium chemical potential region for CdTe:Na in the two dimensional (μ_{Cd}, μ_{Na}) plane. The shaded area is forbidden under equilibrium growth condition.

Te-rich condition ($\mu_{Cd} = -0.79$ eV, $\mu_{Te} = 0$), μ_{Na} is further reduced to less than −1.42 eV. Above these chemical potential limits, secondary Na_2Te compound will form, thus stopping the doping process.

For intrinsic defects, the accessible chemical potential region is given by Eqs. (5) and (6) only. For example, for V_{Cd}^0 in CdTe, the lowest formation energy occurs at the Cd-poor condition, whereas for Cd interstitial Cd_i, the lowest formation energy occurs at the Cd-rich condition. For impurities (extrinsic defects), Eq. (7) should also be considered. For example, for Na_{Cd}^0, the calculated $\Delta H_f(Na_{Cd}^0) = 0.45$ eV at $\mu_i = 0$. Therefore, the lowest formation energy occurs at the Cd-poor condition with $\Delta H_f(Na_{Cd}^0) = 0.45 - 0.79 + 1.42 = 1.08$ eV. This is because at the Cd poor condition ($\mu_{Cd} = -0.79$ eV), the highest possible μ_{Na} is −1.42 eV (see Fig. 4). On the other hand, for Cu_{Cd}^0, the calculated $\Delta H_f(Cu_{Cd}^0) = 1.31$ eV at $\mu_i = 0$. Because the formation energy of Cu_2Te is close to zero, the lowest formation energy of Cu_{Cd}^0 at the Cd-poor condition is $\Delta H_f(Cu_{Cd}^0) = 1.31 - 0.79 = 0.52$ eV. Thus, the solubility of Cu in CdTe is much larger than that of Na. This analysis indicates that impurities that do *not* form strong bonds with the host elements have higher solubility than impurities that do form strong bonds with the host elements, contrary to naive thoughts. This is because if the impurity does not form strong bond with the host atom, it will also be less likely to form secondary phases.

4. Symmetry and occupation of single-particle defect state

It is important to know the symmetry and character of the single-particle defect state, because defects with different symmetry and character will behave differently. Also, to modify defect states, only the states with the same symmetry can couple strongly with each other. Another important issue often encountered in the LDA calculations is to estimate the effects of the LDA band gap corrections on the calculated energy states. For example, although both anion vacancy and cation interstitial have the same a_1 defect levels in II–VI semiconductors (see below), anion vacancy has the a_1^v character derived from the valence band, whereas cation interstitial has instead the a_1^c character derived from the conduction band. Thus, the energy level of the cation interstitial is expected to follow closely with the CBM, whereas the energy level of anion vacancy state will not.

For simple extrinsic impurities, one can predict in principle whether a dopant is a donor with a single-particle energy level close to the CBM or an acceptor with a single-particle energy level close to the VBM by simply counting the number of the valence electrons of the dopants and the host elements. For example, in CdTe, one can expect that group-I elements substituting on the Cd site, X_{Cd}^I (X^I = Na, Cu) give acceptors, whereas group-VII elements substituting on the Te site, Y_{Te}^{VII} (Y^{VII} = Cl, Br) give donors. Generally speaking, to produce a shallow acceptor, it is advantageous to use a more electronegative dopant, whereas to produce a shallow donor, it is advantageous to use a less electronegative dopant.

For intrinsic defects, the situation is more complicated. Fig. 5 shows the single-particle energy levels of tetrahedrally coordinated charge-neutral defects in CdTe [14]. Generally speaking, when a high-valence atom is replaced by a low-valence atom (e.g., Cd_{Te}) or by a vacancy (V_{Cd} and V_{Te}), defect states are created from the host valence (v) band states that move upward in energy. The defect states consist of a low-lying singlet a_1^v state and a high-lying threefold-degenerate t_2^v state. Depending on the potential, both t_2^v and a_1^v can be above the VBM. These states are occupied by the nominal valence electrons of the defect plus the valence electrons associated with the neighboring atoms (six electrons if the defect is surrounded by four Te atoms or two electrons if it is surrounded by four Cd atoms). For example, for charge-neutral V_{Cd}, the defect center has a total of $0+6=6$ electrons. Two of them will occupy the a_1^v state and the remaining four will occupy the t_2^v states just above the VBM, so V_{Cd} is an acceptor. On the other hand, if a low-valence atom is replaced by a high-valence atom (e.g., Te_{Cd}), or if a dopant goes to an interstitial site (e.g., Cd_i and Te_i), the a_1^v and t_2^v are pulled down and will remain inside the valence band. Instead, the defect states a_1^c and t_2^c are created from the host conduction band states that move downward in energy. Depending on the potential, both the a_1^c and t_2^c states can be in the gap. For example, for charge-neutral Te_{Cd}, $6+6=12$ electrons are associated with this defect center. Eight of them will occupy the bonding a_1^v and t_2^v states, two will occupy the a_1^c state, and the remaining two will occupy the t_2^c state. Since the partially occupied t_2^c state is close to the CBM, Te_{Cd} is also a donor. For the interstitial defect, Cd_i has two electrons that will fully occupy the a_1^c state and is thus expected to be a donor. The Te_i defect center has six electrons. Two will occupy the a_1^c state and the remaining four will occupy the t_2^c states. Since the partially occupied t_2^c states are closer to the VBM, Te_i is expected to be a deep acceptor.

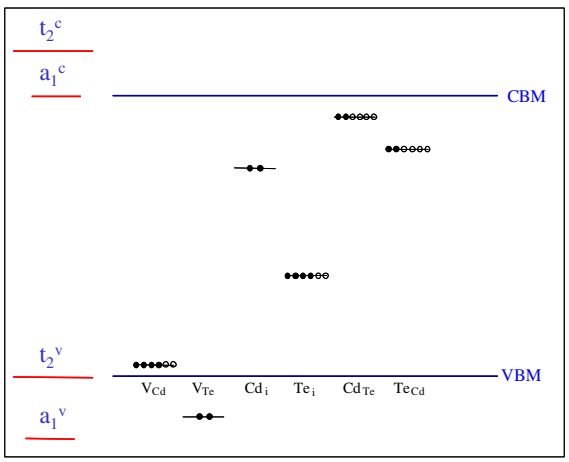

Fig. 5. Single-particle defect levels for the tetrahedrally coordinated neutral intrinsic defects in CdTe. The solid and open dots indicate the occupation of each state. See text for details.

After the general discussions above on the semiconductor defect physics, in the following we will discuss how to design methods to overcome the doping bottlenecks in semiconductors.

5. Optimizing the host chemical potential

A good dopant should avoid either self-compensation or compensation by intrinsic defects. Our investigation has shown that the dominant intrinsic defect that compensates acceptors is cation interstitial or anion vacancy, and the dominant intrinsic defect that compensates donors is cation vacancy [14]. Therefore, to avoid the formation of intrinsic compensating defects, for p-type doping the growth should be carried out at the anion-rich condition where the formation energy of cation interstitial or anion vacancy is relatively large. Under the anion-rich condition, however, substitution at the cation site has the lowest formation energy, whereas substitution at the anion site has the highest formation energy. It is, therefore, beneficial to use cation substitution as p-type dopants. On the other hand, to avoid the formation of compensating acceptors for n-type doping, the growth should be carried out under the cation-rich condition where the formation energy of cation vacancy or anion interstitial is relatively large. However, under the cation-rich condition, substitution on cation site has the highest formation energy, whereas substitution on anion site has the lowest formation energy. Therefore, it is usually

favorable to use anion substitution as n-type dopants if they produce shallow donor levels.

6. Surface-enhanced defects solubility: The case of GaAs:N

As discussed above, what controls the dopant solubility is the dopant chemical potential, μ_A. Therefore, the key to enhancing the solubility of the dopant is to raise the chemical potential and avoid the formation of the precipitates of the dopants. One approach is to use epitaxial growth. Because in epitaxial growth the formation of the secondary phase is often more difficult to form at the surface, therefore, the dopant chemical potential and solubility can be significantly increased. For example, Zhang and Wei [10] show that for isovalent N doping in GaAs, because the bulk GaN is very stable, the highest N chemical potential one can achieve is $\mu_N = -0.95$ eV at As-rich limit, and the lowest formation energy of N in GaAs is 1.64 eV, which leads to the extremely low N solubility of $[N] < 10^{14}$ cm^{-3} at $T = 650$ °C, as observed in melt-grown GaAs:N [42]. However, in epitaxial growth, the GaN precipitate will not form spontaneously at the surface until the chemical potential $(\mu_{As}, \mu_N) = (-0.44$ eV, 0.0). At this chemical potential, the formation energy of N in GaAs is reduced from the original value of 1.64 eV to only 0.24 eV. This gives rise to $[N] \sim 4\%$ at 650 °C. This example shows that the physics of surface-enhanced N solubility resides in the difference between the formation energy of GaN on the GaAs surface and that of relaxed bulk GaN. The difficulty of forming a GaN at the GaAs surface leads to a higher achievable μ_N, thus, a higher epitaxial N solubility.

7. Metastable molecular dopant enhanced solubility: The case of ZnO:N

Another way to increase the chemical potential of the dopant is to use metastable dopant containing molecules, which can deliver dopant at higher chemical potential [19,43]. An example was pointed out recently by Yan et al. [19] who show that in the case of N doping of ZnO at least four different gases can be used, namely N_2, N_2O, NO, and NO_2. When these molecules arrive intact at the growing surface, their respective chemical potentials determine the doping efficiency. The three nitrogen–oxygen molecules are all only metastable with respect to the dissociation into N_2 and O_2, therefore, N chemical potentials in these molecules are much higher than that in N_2. For example, using the LDA and FLAPW method, the calculated minimum formation energy for N_O is 1.2 eV, which occurs at the chemical potential ($\mu_N = 0$ eV, $\mu_{Zn} = -0.2$ eV) if N_2 is used as dopant. Here, we have taken into account the calculated formation energies $\Delta H_f(ZnO) = -3.6$ eV and $\Delta H_f(Zn_3N_2) = -0.6$ eV. If the N chemical potential is instead limited by NO with a calculated formation energy of $\Delta H_f(NO) = +0.8$ eV, the minimum formation energy of N_O is reduced to 0.4 eV, which occurs at ($\mu_N = 2.5$ eV, $\mu_{Zn} = -1.9$ eV). Therefore, using NO as the dopant instead of N_2 can greatly increase the N solubility in ZnO. Notice that in metastable molecule doping, one has to recalculate the constraints given by Eq. (7) when the dopant chemical potential is increased above its equilibrium value. Otherwise, unphysical negative formation energy could be obtained [19,44].

8. Large-size-mismatched p-type doping in ZnO

Although one can incorporate large amounts of N into ZnO through the non-equilibrium approach, it is now clear that the level of N_O is relatively deep, making acceptor-ionization difficult. Experimental evidences also show that N-doped ZnO could be unstable [45]. Other group-V substitutions on oxygen sites produce even deeper acceptors. For example, the calculated As_O acceptor level is deep, at about 930 meV [13] above the VBM, making it impossible for As_O to dope ZnO efficiently p-type. As shown in Section 5, p-type doping can be best obtained under anion-rich growth conditions with substitution on the cation site to improve solubility and reduce compensation. For ZnO, however, we have shown that cation substitution by group-I elements produces

only deep levels because group-I elements tend to occupy interstitial sites [13]. To further pursue p-type ZnO, Limpijumnong et al. [17] have recently developed a theory for large size-mismatched impurities. Guided by strain relief and Coulomb interaction, we show that in ZnO an As_{Zn}–$2V_{Zn}$ complex represents a new class of stable defects with shallow acceptor levels. In the complex, the As atom occupies the Zn antisite, which is energetic enough to spontaneously induce two Zn vacancies. The resulting As_{Zn}–$2V_{Zn}$ complex is an acceptor with both low formation energy and low ionization energy, $\epsilon(0/-) = 0.15$ eV, in good agreement with experiments [46,47], 0.12–0.18 eV. The finding also naturally explains the puzzling experimental observations that oxygen-rich growth/annealing conditions, which would severely suppress the formation of As_O, are required for successful p-type doping.

9. Effects of co-doping/cluster doping

There have been discussions that the use of co-doping [48,49]/cluster doping [44] may enable one to (a) enhance the dopant solubility by lowering the formation energy of the defect complex through interactions (mostly the Coulomb interaction) between the constituents of the defect complex and (b) lower the defect transition energy levels through the coupling between the donor–acceptor states. We have tested these suggestions by calculating the defect formation energies and defect transition energy levels of the defect complexes in CdTe as well as ZnO [14]. Our results show that (a) in general, co-doping/cluster-doping does not lower the formation energies of the defect complexes below that of the corresponding single point defects, because the Coulomb interaction between the donors and acceptors is not sufficient to compensate the energy cost of creating the extra individual point defects. This is especially true in the case of cluster doping (e.g., $4Cu_{Cd} + Cl_{Te}$ in CdTe) where interaction between the close-packed Cu_{Cd} acceptors also increases the formation energy of the cluster. However, in some cases, the formation energy of the defect complexes (e.g., $V_{Cd} + Cl_{Te}$ in CdTe) are not too much higher than the corresponding point defect (e.g., V_{Cd}), but due to the introduction of new chemical species, one may have larger freedom to adjust the chemical potentials during growth. Therefore, co-doping could be useful in increasing dopant incorporation under non-equilibrium growth conditions. (b) The calculated defect transition energy level can increase or decrease, but the effects are usually very small. The main problem for conventional co-doping is that the donor level usually has s-like symmetry and the acceptor level has p-like symmetry, therefore, the level repulsion is very weak. Furthermore, in the case of two donors plus one acceptor (or two acceptors plus one donor), or in the case of cluster doping, the donor–donor (acceptor–acceptor) level coupling can even raise the ionization energy [14]. In the following, we will show other approaches that we have designed to engineer shallow donor levels in wide-gap semiconductors.

10. Reducing ionization energy by combining donor with donor

ZnTe is a wide-gap material that has been doped p-type with free hole concentration [4] of 1×10^{20} cm^{-3}, whereas for n-type ZnTe, the maximum free electron concentration has never exceeded 4×10^{18} cm^{-3} despite a much higher dopant concentration [4,5,50], suggesting that a suitable shallow n-type dopant in ZnTe is yet to be found.

To reduce the ionization energy, we have explored a different and novel idea in which a fully occupied deep donor is used to attract a second partially occupied donor to lower its ionization energy [16]. In particular, we studied a double donor (either Si, Ge, or Sn on the Zn site) paired with a single donor (either F, Cl, Br, or I on the Te site) in ZnTe. Different from the Coulomb binding that exists in charged donor–acceptor complexes in the co-doping approach [48,49], the binding between the two donors results from the level repulsion between the two donor states. The level repulsion significantly reduces the energy of the fully occupied lower level, stabilizing the donor–donor pair, while it increases the energy of the partially occupied upper level, thus reducing the

ionization energy. Notice that because the doubly occupied a_1^{IV}-derived state is *charge neutral*, there is no Coulomb repulsion between the two nominally donor impurities. Furthermore, because the two donor states have the same symmetry and atomic character the level repulsion is very efficient. For example, we find that the formation of a Br_{Te}–Sn_{Zn} pair in ZnTe is exothermic with a binding energy of 0.9 eV. It lowers the electron ionization energy of Br_{Te} by a factor of more than three from 240 to 70 meV, resulting in an effective shallow donor.

11. Reducing ionization energy by combining isovalent defect with donor

It is well known that diamond, which possesses unique physical properties, can be doped p-type relatively easily by boron acceptors, but that efficient n-type doping is very difficult to achieve [6]. Currently, the best n-type diamond has been achieved using *P* as the dopant. However, even in this case, the ionization energy of the impurity, at about 0.5 eV below the CBM [51], is still not shallow enough for practical applications.

The approach that is described in the previous section for improved n-type doping in ZnTe, however, cannot be applied directly to n-type doping in diamond, because diamond has an indirect band gap. The donor levels are derived from the linear combination of the six equivalent Δ minima, which split into levels with a_1, e, and t_2 symmetry, through the inter-valley coupling. In this case, a single deep donor is not sufficient to push all the levels up, because (a) there is no fully occupied single deep donor, and (b) the symmetry lowering of the donor–donor defect pair (usually from T_d to C_{3v}) splits the degenerate t_2 donor level, and can therefore make the lowest split t_2 level even lower than the original t_2 level of an isolated defect, such as substitutional *P*.

Based on the above considerations, we proposed a new method to overcome the high ionization problem in n-type doping of diamond, by combining donor (e.g., N) with isovalent (e.g., Si) impurities [15]. In this case, the level repulsion between isovalent-induced levels and donor-induced levels increases the energy of the partially occupied donor levels, and thus reduces the ionization energy. Moreover, the level repulsion decreases the energy of the fully occupied isovalent levels and stabilizes the isovalent-donor complex. To achieve the most effective level repulsion, we proposed to form a N + 4Si defect cluster, which preserves the T_d symmetry, and thus, avoids the splitting of the defect levels. Our theoretical calculations show that the N + 4Si defect complex in diamond indeed has very shallow donor transition energy level at 0.09 eV below the CBM, whereas the transition energy of isolated N donor is at 1.7 eV below the CBM. This is the shallowest donor level that has ever been found for diamond. We find that this defect has a large binding energy of 3.17 eV, and is therefore stable.

12. Enhancing the dopability by modifying the band structure of the host

One can also modify the crystal and band structure of the host to improve dopability. An example is the design of a bipolarly dopable transparent conducting oxide (TCO). TCO comprises a group of materials with unique physical properties [52]. Despite their usually large band gaps (> 3 eV), making them transparent under normal conditions, TCOs can sustain a high concentration of charged carriers and also maintain a high mobility. Almost all of the well-known TCOs, such as ZnO, In_2O_3, SnO_2, and their alloys, have n-type conductivity. This can be attributed to the fact that the VBM states of these TCOs are mainly O p states with very high workfunction.

Recently, several Cu-containing oxides have been proposed [52] as good candidates for p-type TCOs. This is based on the observations that Cu has shallow, occupied 3d orbitals above the O 2p orbitals. The coupling between the Cu d states and the O p states leads to smaller workfunction (higher VBM) for these Cu compounds than conventional oxides, thus, according to the doping limit rule [10], making it easier to dope them p-type. However, raising VBM also lowers the band gap, and thus reduces transparency. One way to solve this problem is to modify the crystal

structure to reduce the oxygen-mediated d–d coupling between the Cu atoms, thus, enlarging the band gap and enhancing the transparency. This approach has been successful in creating p-type TCOs such as $CuAlO_2$ and $SrCu_2O_2$ [11,52].

However, to have bipolarly dopable TCOs, according to the doping limit rule [10], the materials should have high VBM and low CBM, and thus a small band gap. So how can we have a small band gap material that is also transparent? Using first-principles band structure methods, we have calculated the electronic and optical properties of $CuInO_2$ in the delafossite structure [12]. We find that $CuInO_2$ has a small fundamental direct gap (i.e., the smallest direct band gap) at Γ. The high VBM and low CBM of $CuInO_2$ makes it easy to be doped both p-type and n-type. However, the corresponding dipolar optical transition matrix element at Γ is exactly zero because the two band edge states have the same (even) parity. An important consequence is that absorption *near* the fundamental gap at Γ for $CuInO_2$ is very small and barely increases with energy until transitions at the next critical points take place (which defines an apparent optical band gap). The calculated apparent gap is large for $CuInO_2$, therefore, it is transparent, in good agreement with experiment [53,54]. Thus, the large difference in terms of the energy and transition matrix element between the fundamental band gap and the apparent band gap is the key to achieving the formerly inconceivable combination of good transparency with bipolar dopability.

13. Summary

Using the first-principles band structure method, we have systematically studied the general chemical trends of the defect formation and ionization in semiconductors to understand the physical origin of the doping difficulty. New approaches of overcoming the doping limit have been studied. We show that dopant solubility can be enhanced significantly by optimizing the host elements' chemical potential or by expanding the physically accessible range of the dopant atomic chemical potentials. This can be done by either using epitaxial growth techniques, such as MBE, or a metastable dopant containing molecules as doping sources. We find that, under equilibrium growth conditions, conventional co-doping or cluster doping (mixing donors with acceptors) in general do not reduce the formation energy of the defect complexes below that of a single point defect or lower the defect transition energy level significantly. Instead, by combining a selective donor with a donor, an acceptor with an acceptor, or an isovalent defect with a donor, one can form stable defect complexes and lower the defect ionization energy. Several examples from recent studies are discussed.

Acknowledgements

I would like to thank A. Janotti, S. Limpijumnong, X. Nie, C.-H. Park, D. Segev, and S.B. Zhang for their contribution in this study. This work was supported in part by US Department of Energy, Grant DE-AC36-99-GO10337.

References

[1] G.F. Neumark, Mater. Sci. Eng. R21 (1997) 1.
[2] S.B. Zhang, J. Phys.: Condens. Mat. 14 (2003) R881.
[3] S.J. Pearton, F. Ren, A.P. Zhang, K.P. Lee, Mater. Sci. Eng. R 30 (2000) 55.
[4] I.W. Tao, M. Jurkovic, W.I. Wang, Appl. Phys. Lett. 64 (1994) 1848.
[5] J.H. Chang et al., Appl. Phys. Lett. 79 (2001) 785.
[6] R. Kalish, Diamond Relat. Mater. 10 (2001) 1749.
[7] D.J. Chadi, K.J. Chang, Phys. Rev. Lett. 61 (1988) 873; D.J. Chadi, K.J. Chang, Phys. Rev. B 39 (1989) 10063.
[8] C.H. Park, D.J. Chadi, Phys. Rev. Lett. 75 (1995) 1134; C.H. Park, D.J. Chadi, Phys. Rev. B 55 (1997) 12995.
[9] S.B. Zhang, S.-H. Wei, A. Zunger, J. Appl. Phys. 83 (1998) 3192.
[10] S.B. Zhang, S.-H. Wei, Phys. Rev. Lett. 86 (2001) 1789.
[11] X. Nie, S.B. Zhang, S.-H. Wei, Phys. Rev. B 65 (2002) 075111.
[12] X. Nie, S.B. Zhang, S.-H. Wei, Phys. Rev. Lett. 88 (2002) 066405.
[13] C.H. Park, S.B. Zhang, S.-H. Wei, Phys. Rev. B 66 (2002) 073202.
[14] S.-H. Wei, S.B. Zhang, Phys. Rev. B 66 (2002) 155211.
[15] D. Segev, S.-H. Wei, Phys. Rev. Lett. 91 (2003) 126406.
[16] A. Janotti, S.-H. Wei, S.B. Zhang, Appl. Phys. Lett. 83 (2003) 3522.

[17] S. Limpijumnong, S.B. Zhang, S.-H. Wei, C.H. Park, Phys. Rev. Lett. 92 (2004) 155504.
[18] S.B. Zhang, S.-H. Wei, A. Zunger, Phys. Rev. Lett. 84 (2000) 1232.
[19] Y. Yan, S.B. Zhang, S.T. Pantelides, Phys. Rev. Lett. 86 (2001) 5723.
[20] S.A. Awadalla, A.W. Hunt, K.G. Lynn, H. Glass, C. Szeles, S.-H. Wei, Phys. Rev. B 69 (2004) 075210.
[21] G.A. Baraff, M. Schluter, Phys. Rev. Lett. 55 (1985) 1327.
[22] S.B. Zhang, J.E. Northrup, Phys. Rev. Lett. 67 (1991) 2339.
[23] D.B. Laks, C.G. Van de Walle, G.F. Neumark, P.E. Blochl, S.T. Pantelides, Phys. Rev. B 45 (1992) 10965.
[24] P. Hohenberg, W. Kohn, Phys. Rev. 136 (1964) B864; W. Kohn, L.J. Sham, Phys. Rev. 140 (1965) A1133.
[25] J.P. Perdew, A. Zunger, Phys. Rev. B 23 (1981) 5048.
[26] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865.
[27] S.-H. Wei, H. Krakauer, Phys. Rev. Lett. 55 (1985) 1200; D.J. Singh, Planewaves, Pseudopotentials, and the LAPW Method, Kluwer, Boston, 1994.
[28] Abinit is a common project of the University Catholique de Louvain, Corning Incorporated, and of other contributors (URL http://www.abinit.org).
[29] G. Kresse, J. Furthmuller, Phys. Rev. B 54 (1996) 11169; G. Kresse, J. Furthmuller, Comput. Mat. Sci. 6 (1996) 15.
[30] H.J. Monkhorst, J.D. Pack, Phys. Rev. B 13 (1976) 5188.
[31] S. Froyen, Phys. Rev. B 39 (1989) 3168.
[32] C.G. Van de Walle, S. Limpijumnong, J. Neugebauer, Phys. Rev. B 63 (2001) 245205.
[33] G. Makov, R. Shah, M.C. Payne, Phys. Rev. B 53 (1996) 15513.
[34] G. Makov, M.C. Payne, Phys. Rev. B 51 (1995) 4014.
[35] L.G. Wang, A. Zunger, Phys. Rev. B 66 (2002) R161202.
[36] J. Lento, J.-L. Mozos, R.M. Nieminen, J. Phys.: Condens. Mat. 14 (2002) 2637.
[37] P. Schultz, Phys. Rev. B 60 (1999) 1551; P. Schultz, Phys. Rev. Lett. 84 (2000); L.N. Kantorovich, Phys. Rev. B 60 (1999) 15479.
[38] W. Walukiewicz, Appl. Phys. Lett. 54 (1989) 2094.
[39] J.A. Van Vechten, in: S.P. Keller (Ed.), Handbook of Semiconductors, vol. 3, North Holland, Amsterdam, 1980, p. 1.
[40] S.-H. Wei, A. Zunger, Appl. Phys. Lett. 72 (1998) 2011.
[41] S.-H. Wei, X. Nie, I. Batyrev, S.B. Zhang, Phys. Rev. B 67 (2003) 165209.
[42] I.-H. Ho, G.B. Stringfellow, J. Cryst. Growth 178 (1997) 1.
[43] S.B. Zhang, S.-H. Wei, Yanfa Yan, Physica B 302–303 (2001) 135.
[44] L. Wang, A. Zunger, Phys. Rev. Lett. 90 (2003) 256401.
[45] X. Li et al., J. Vac. Sci. Tech. A 21 (2003) 1342.
[46] Y.R. Ryu, T.S. Lee, H.W. White, Appl. Phys. Lett. 83 (2003) 87.
[47] C. Morhain et al., Phys. Stat. Sol. (b) 229 (2002) 881.
[48] H. Katayama-Yoshida, T. Yamamoto, Phys. Stat. Sol. (b) 202 (1997) 763.
[49] H. Katayama-Yoshida, T. Nishimatsu, T. Yamamoto, N. Orita, J. Phys. Condens. Matter 13 (2001) 8901.
[50] H. Ogawa et al., Jpn. J. Appl. Phys. 33 (Part 2) (1994) L980.
[51] S. Koizumi et al., Science 292 (2001) 1899.
[52] See Reviews in MRS Bull. 25 (8) (2000).
[53] H. Kawazoe, M. Yasukawa, H. Hyodo, M. Kurita, H. Yanagi, H. Hosono, Nature (London) 389 (1997) 939.
[54] H. Yanagi, T. Hase, S. Ibuki, K. Ueda, H. Hosono, Appl. Phys. Lett. 78 (2001) 1583.