Electronic materials theory: Interfaces and defects

Chris G. Van de Walle[a)]
Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304

(Received 9 June 2003; accepted 17 June 2003; published 2 September 2003)

An overview of developments in materials theory is presented, with an emphasis on first-principles calculations. Examples are given from the fields of heterojunction interfaces and point defects in semiconductors. Predictive theories of materials are shown to be increasingly important for understanding but also designing materials and structures. © *2003 American Vacuum Society.*
[DOI: 10.1116/1.1599867]

I. INTRODUCTION

The experimental advances in electronic materials over the past decades have been accompanied by a remarkable increase in the ability to predict structural and electronic properties from first principles. Basic theory, along with modeling and simulation, has always been instrumental in understanding materials. Only recently, however, has the capability emerged to accurately predict properties based solely on the composition of the material, without any fitting to experimental quantities. Such a description must be based on a quantum-mechanical treatment, i.e., a solution of the Schrödinger equation for the system of atomic constituents. The seemingly impossible task of solving this vast many-body problem was rendered feasible by the development of density functional theory (DFT), an achievement for which Walter Kohn received the Nobel Prize in Chemistry in 1998.[1] This approach reduces the problem to a one-particle Schrödinger equation, with all many-body aspects folded into an effective potential. The exact form of this potential is unknown, but approximations such as the local-density approximation (LDA)[1] have been remarkably productive.

Other advances have also greatly enhanced the ability to tackle large systems. For instance, the properties of many electronic materials are largely determined by the valence electrons; an efficient way to eliminate the core electrons from the problem is provided by the use of pseudopotentials.[2] State-of-the-art pseudopotentials are generated using only information about the atom without any fitting to experiment. Most problems require calculations of not only electronic wave functions, but also atomic positions in a structure. An important advance in this respect was the development of the Car–Parrinello method,[3] which allows simultaneous optimization of the electronic and atomic degrees of freedom. The ability to move atoms allows performing first-principles molecular dynamics, as well. We also note that the tremendous increase in computer power that has become available over the last few decades has reshaped the field: Twenty years ago, calculations for systems with two atoms in the unit cell required a mainframe whereas, these days, systems with several hundred atoms can be calculated on a desktop machine.

In this article, I will focus on two areas in which these theoretical and computational advances have had a major impact: Heterojunction interfaces, and defects in semiconductors. Both are intimately connected to the high-quality growth techniques that have enabled a host of electronic devices. First-principles techniques have also been instrumental in the theory of surfaces; the interested reader may consult the article by Feibelman in this volume,[4] as well as a review article by Duke.[5] The American Vacuum Society (AVS) has played an important role in stimulating the development of these techniques, and AVS-sponsored meetings have been a key forum in which progress has been discussed. Many of the pioneering developments have been published in the Journal of Vacuum Science and Technology, which has played a major role in disseminating the information.

II. HETEROJUNCTIONS

When two semiconductors are joined at a planar interface, they form a heterojunction. The valence and conduction bands exhibit discontinuities (Fig. 1), and these band offsets can be used to tailor the distribution and flow of carriers in a layer structure. In 1963, Herbert Kroemer proposed that double heterostructures could be used to confine carriers, an essential feature for semiconductor lasers; Kroemer received the 2000 Nobel Prize in Physics for his work.[6] Heterojunctions are also commonly used in electronic devices such as bipolar transistors and high electron mobility transistors (HEMTs). In fact, the fundamental studies of two-dimensional electron gases that led to the discoveries of the integer and fractional quantum Hall effect rely entirely on carriers being confined near a GaAs/AlGaAs heterojunction.[7,8]

The magnitude of the band discontinuities is the key quantity characterizing a heterojunction. For instance, the conduction-band offset between well and barrier layers determines the degree of confinement of electrons in a quantum well. In bipolar transistors, the inclusion of a heterojunction enables achieving a higher emitter efficiency and a higher current-amplification factor—a concept also proposed by Kroemer. And, in HEMTs, the channel is formed by a two-dimensional electron gas near a heterojunction. The design of all of these devices requires knowledge of the band offsets at the heterojunction. Band offsets can be measured experimentally, but accurate measurements are quite difficult, particularly for polar interfaces or in cases where the materials

[a)]Electronic mail: vandewalle@parc.com

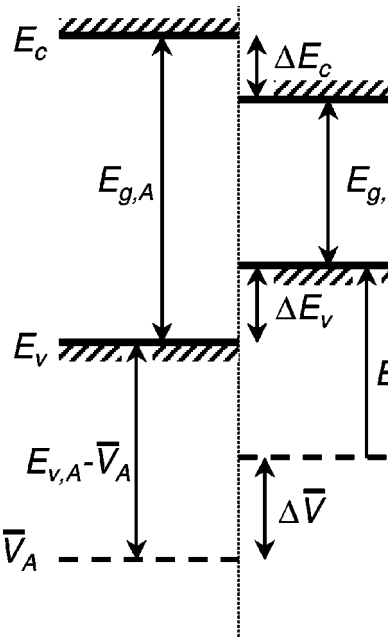

FIG. 1. Schematic illustration of the band lineup between semiconductors A and B. The positions of valence and conduction bands are indicated, all referenced to their appropriate reference level $\overline{V}$ in each semiconductor. The difference between $\overline{V}_A$ and $\overline{V}_B$ determines the band lineup.

exhibit a lattice mismatch. Each materials combination or strain configuration, in principle, requires another measurement. There has, therefore, been a strong driving force for computational predictions of band-offset values. The high numbers of citations received by papers presenting computational results for band offsets are a testimony to the fact that device designers actively rely on these numbers.

Over the years, a number of different models have been proposed for predicting band offsets. Perhaps the simplest is the electron affinity rule,[9] which obtains the conduction-band offset by taking the difference between the electron affinities of the two semiconductors. The accuracy of this model is limited, mainly because electron affinities are *surface* quantities, and a surface constitutes a very severe perturbation to the crystal: At the surface, the electron density spills out into a vacuum, often resulting in substantial dipoles. In contrast, most interfaces between two semiconductors constitute a relatively gentle perturbation. Using surface-related quantities to obtain a heterojunction band offset is therefore inappropriate.

In addition, heterojunction interfaces may exhibit features that are beyond what can be described by an electron affinity. Pseudomorphic interfaces between lattice-mismatched semiconductors are an excellent example. Modern growth techniques allow growing a thin layer of one semiconductor on top of a second semiconductor with a different lattice constant, in such a way that the first semiconductor assumes the in-plane lattice constant of the underlying layer. An excellent example is growth of SiGe alloys on top of silicon, a materials combination that is now commercially used in electronic devices. The strain that results from bringing the in-plane lattice constants into alignment results in shifts of valence and conduction bands, which in turn affect the band offsets. Theories of band discontinuities need to be capable of including such effects.

The AVS-sponsored conferences on Physics and Chemistry of Semiconductor Interfaces, started in 1974, have been a major forum in which theories of heterojunctions have been discussed. The Conference Proceedings, published in the Journal of Vacuum Science and Technology, provide excellent documentation of the progress in the field.[10]

Reviews of theoretical approaches for predicting band offsets have been given in Refs. 11 and 12. Here, I will focus on a brief review of first-principles calculations and related model theories.

A. First-principles calculations

We start with two semiconductors A and B and bring them together at an interface (Fig. 1). The first issue is to determine the atomic structure of the interface. Even at an ideal, abrupt interface, some relaxation of the atoms may take place in the vicinity of the junction. First-principles calculations of forces indicate in which direction the atoms need to be moved in order to minimize the energy of the system as a function of atomic positions. Such determinations of the atomic structure can also be applied when the layers are pseudomorphically strained. The determination of strains based on the macroscopic elastic theory has been shown to be accurate even for very small thicknesses of the semiconductor, down to a few atomic layers.[13]

We are then in a position to calculate the lineup of band structures at this interface. This problem can be divided into an interface-specific part and a bulk part, as illustrated in Fig. 1. A bulk calculation yields the band structure of semiconductor A relative to a reference level, usually an average of the electrostatic potential $\overline{V}_A$. Similarly, the band structure of semiconductor B is referenced to a level $\overline{V}_B$. For instance, the position of the valence band, E_v, is a distance $E_v - \overline{V}$ above the position of the average electrostatic potential. The band-offset problem then consists of determining the difference in average electrostatic potentials between A and B, $\Delta \overline{V}$. This procedure is similar to one practiced in x-ray photoemission spectroscopy (XPS) where, typically, the separation between two representative core levels is measured across the interface, and independent measurements on bulk samples are performed to obtain the energy separation between the valence-band maximum and the core levels in each material. The core-level separation is then used to line up the valence bands and obtain the band offsets.

$\Delta \overline{V}$ *cannot* be obtained from bulk calculations alone, because there is no absolute reference for the average potential in an infinite solid. This problem is due to the long-range nature of the Coulomb interaction, which causes the average potential of an infinite system to be ill defined.[14] A band structure calculation for an individual solid thus cannot provide information about the absolute position of the average potential. $\Delta \overline{V}$ can be obtained from first-principles calculations for an A/B interface. The algorithms typically assume periodicity, which can be maintained by considering a super-

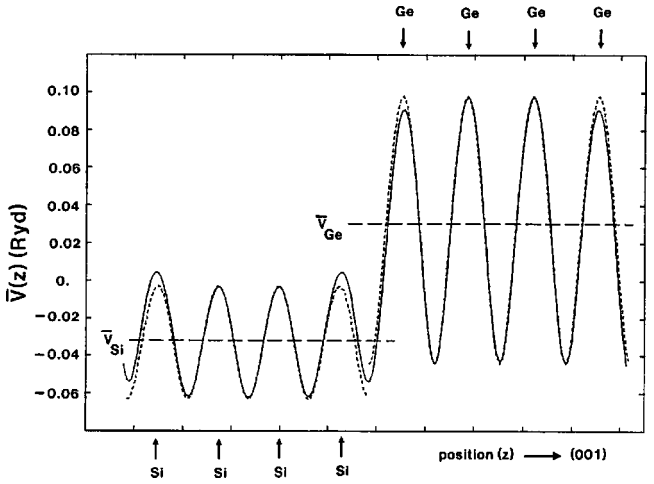

FIG. 2. Variation of the planar-averaged potential $\overline{V}(z)$ across a Si–Ge (001) interface, calculated in a $4+4$ superlattice consisting of unstrained Si and strained Ge. The dashed lines show the corresponding potentials for bulk Si and (strained) bulk Ge (shifted so their average values coincide with $\overline{V}_{\mathrm{Si}}$ and $\overline{V}_{\mathrm{Ge}}$). The bulk potentials are seen to coincide with $\overline{V}(z)$ already at a distance of one atomic layer away from the interface. The shift between $\overline{V}_{\mathrm{Si}}$ and $\overline{V}_{\mathrm{Ge}}$ determines the band lineup. (From. Ref. 13.)

lattice in which layers of the two semiconductors are periodically repeated.[15] Results for isolated interfaces can be obtained if the layers are sufficiently thick to ensure adequate separation between adjacent interfaces; in practice, four to six atomic layers typically suffice since charge densities and potentials converge rapidly to their bulk value away from the interface.[16] It is then possible to identify a ''bulklike'' region in the middle of each superlattice layer, where the value of the average potential can be determined. This is illustrated in Fig. 2, for the example of a Si–Ge interface.

Density functional theory does not guarantee that the calculated band structure is accurate. A well known consequence of this deficiency is the failure of DFT-LDA to produce the correct band gaps. Corrections beyond DFT-LDA may therefore be necessary to obtain the actual band positions with respect to the average electrostatic potential. $\Delta\overline{V}$ itself depends only on the charge density of the heterojunction, and as such is a ground-state property that is reliably given by DFT; the corrections to DFT-LDA are therefore limited to the positions of the energy levels in the band structure, obtained from bulk calculations for the individual materials. Going beyond the LDA, but staying within DFT, for instance by using the generalized gradient approximation, does not provide a systematic improvement of the band structure.[17] Going beyond DFT, the so-called GW method has been shown to produce band structures that compare favorably with experiment. GW calculations for a large number of semiconductors were reported in Ref. 18. It was found that the positions of the *valence band* were generally quite accurate within DFT. The average DFT error in the valence-band offset was found to be ~ 120 meV. DFT results for valence-band offsets are therefore generally reliable to within about 0.1 eV; if higher accuracy is required, corrections based on GW calculations for bulk materials can be added.

A large number of groups have carried out first-principles calculations for band offsets at a wide variety of heterojunctions. Space does not permit a discussion of other developments, such as the linear response theory,[14] which have significantly contributed to our understanding of heterojunction band offsets. In Sec. II B, I briefly discuss the development of model theories for band offsets, which in many cases have been strongly influenced by the information extracted from first-principles calculations. Such models are widely used in the design of semiconductor devices.

B. Model theories

We have to accept that the average electrostatic potential obtained from a band structure calculation for a bulk semiconductor is not known on an absolute energy scale. It is defined only to within an arbitrary constant, which can be fixed by making specific assumptions about the boundary conditions. In order to solve the heterojunction problem, an obvious approach is to specify the boundary condition to be exactly that at the semiconductor interface, i.e., to perform a calculation in which both semiconductors are present and joined at the junction, as described in Sec. II A. This entails performing a calculation for each interface, which is usually quite time consuming; also, even though the calculations provide quantitative answers, they do not directly provide any information about the *mechanisms* that determine the lineups. Various model theories have therefore been developed to address the problem.

What all of the models have in common is that they attempt to associate a reference level with each semiconductor, the reference level being an intrinsic property of the bulk semiconductor; band offsets then follow from simply lining up the reference levels. Most models employ the concept of an interface dipole, with the magnitude and importance attached to the dipole varying widely between different theories. Tersoff, who received the 1988 *Peter Mark Award* from AVS for his work on surfaces and interfaces, pointed out that these different point of views can largely be reconciled by realizing that the magnitude of the dipole depends on the choice of reference.[11] Some theories *a priori* choose a reference so that during the lineup process, the interface dipole will be minimal. Other theories choose a different reference (often associated with the particular calculational technique), and then find that a dipole arises that drives the system toward a particular lineup. The latter approach may have the advantage that it is more similar to treatments of metal/semiconductor junctions, where charge transfer is clearly an important driving force. However, within the field of semiconductor heterojunctions both types of model theories seem capable of success.

1. Intrinsic reference levels without a need for additional dipoles

Anderson's electron-affinity rule[9] was one of the first models for band offsets. The limitations of this rule were

discussed herein. Nonetheless, the fundamental idea is a valuable one, if it were only possible to define some kind of ''intrinsic'' electron affinity, which would ignore surface effects and only take the ''bulk'' contribution into account.[19]

Harrison's theory[20] derived intrinsic reference levels for bulk semiconductors, based on atomic energy levels, in the context of the linear combination of atomic orbitals (LCAO) theory. Frensley and Kroemer[21] took information obtained from band-structure calculations and used it to construct a model for band lineups. They relied on establishing reference levels based upon values of the potential at interstitial sites. The accuracy of the approach was limited but it definitely drove the field in the direction of establishing intrinsic reference levels based on bulk calculations.

Another example is the ''model-solid theory,'' [16,22] which is based on an analysis of the self-consistent charge distribution around a large number of semiconductor interfaces, and a comparison with various possible models to generate such a charge density. A simple superposition of neutral atomic charge densities produced a very good approximation to the self-consistent charge density. Using neutral atoms as a building block has an important advantage: Since each building block is neutral, and has no dipole nor quadrupole, the average potential in a system consisting of a superposition of these building blocks is completely determined by the average potential in a single building block. One can then calculate values of the average potential on a common energy scale for all semiconductors, and band offsets can be directly obtained by taking differences between entries in a table.[22] The approach works well for nonpolar interfaces, and additional dipoles at nonpolar interfaces can be evaluated based on electrostatic theory. The model-solid theory also lends itself very well to incorporation of strain effects,[22] using *deformation potentials* to evaluate the resulting shifts in the band structure.

2. Alignment of reference levels driven by dipoles: Charge neutrality levels

This class of models devotes little attention to what the initial, ''zeroth-order'' reference levels are, and stresses that local charge neutrality generates interface dipoles that drive the system toward a particular type of lineup. The idea is that the proximity of another material at the interface induces a distribution of states in the gap of the semiconductor.[11,23] At a metal/semiconductor interface, these states would be related to tails of metal wave functions, and the states are commonly referred to as metal-induced gap states (MIGS).[24] These induced states can carry a certain amount of charge, depending on the fraction of them that are filled. The ''charge neutrality level'' (CNL) in a semiconductor is similar to the Fermi level in a metal: local charge neutrality is maintained by filling states up to the CNL. In general, the CNL will be close to the metal Fermi energy at a metal–semiconductor junction. At a heterojunction, the band lineup is (approximately) determined by lining up the CNLs for the two semiconductors. Indeed, if the levels are not lined up, charge would flow between the two materials, which would set up a dipole that would restore the initial situation.[11] Tersoff proposed a specific way of calculating the CNL, as the branch point in the complex band structure;[11] Cardona and Christensen proposed another implementation, based on dielectric midgap energies.[25]

III. POINT DEFECTS, IMPURITIES, AND DOPING IN SEMICONDUCTORS

Controlled introduction of impurities forms the basis of much of semiconductor technology; indeed, *p*-type (acceptor-doped) and *n*-type (donor-doped) layers and the junctions between them control carrier confinement, carrier flow, and ultimately the device characteristics. Commonly used semiconductors such as Si and GaAs can be doped both *p*- and *n*-type. However, constraints on doping may still limit device performance. For instance, bipolar *npn* transistors would benefit from an increase in the *p*-type doping in the base. Also, the shrinking size of Si field-effect transistors requires higher doping densities, with As donors exhibiting deactivation when the doping increases above $\sim 3\times 10^{20}$ cm^{-3}.

Wide-band-gap semiconductors such as ZnSe, GaN, and ZnO have exhibited the most severe doping problems. For a long time, it was considered impossible to dope ZnSe and GaN *p*-type; when breakthroughs finally occurred (by Park *et al.* for ZnSe,[26] and by Amano *et al.*[27] and Nakamura *et al.*[28] for GaN) they revolutionized the field, rapidly leading to demonstrations of light-emitting diodes and subsequently laser diodes. These blue and UV light emitters are destined to have a huge impact: Besides applications in optical storage and biological agent detection, they enable the development of solid-state white-light sources, which are starting to play an important role in illumination.

The progress in experimental control of doping has gone hand in hand with important advances in theoretical understanding. Many of these developments were deliberated if not presented at AVS-sponsored meetings such as the PCSI conferences.[29] It has long been known that intrinsic point defects can play a critical role in doping. In fact, the conventional wisdom held that point defects such as the nitrogen vacancy in GaN would spontaneously form in large concentrations, dope the material *n*-type, and make it impossible to achieve conversion to *p*-type. This explanation was quite convenient but also largely unverifiable experimentally, because intrinsic point defects are very difficult to detect. We were able to show, based on first-principles calculations, that point defects play a far less important role than previously assumed. For instance, we found that the concentration of nitrogen vacancies in *n*-type GaN[30] is too low to be consistent with the observed electrical conductivity of many samples. This forced the community to turn its attention to other sources of conductivity, in particular, the unintentional incorporation of extrinsic dopants. Based on calculations, we proposed that oxygen is the prime candidate in GaN. This suggestion initially met with a lot of resistance: Growth techniques such as molecular-beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) generally em-

ploy an ultraclean environment, and the assumption was that contamination by impurities was unlikely. Upon careful examination, however, sources of oxygen contamination were indeed identified, such as the lining of plasma sources in the MBE systems, and the presence of water in the ammonia source gas used in MOCVD. Once the culprit was known, measures could be implemented to avoid the contamination, enabling enhanced control over doping levels.

Another interesting example has occurred in the case of ZnO. Again, the commonly observed *n*-type conductivity of this material was traditionally attributed to native point defects, in particular oxygen vacancies. First-principles calculations showed that this explanation was untenable, but unlike the case of GaN no obvious candidate impurity was available to explain doping through unintentional incorporation. A first-principles investigation of hydrogen as an impurity in ZnO proved very informative: Although hydrogen in other semiconductors usually acts only as a compensating center, always *reducing* the prevailing conductivity, hydrogen in ZnO acts as a shallow donor, i.e., as a *source* of conductivity.[31] Since hydrogen is often unintentionally present during growth or processing, knowledge of its electrical behavior is very important. Following the theoretical prediction several experimental groups have verified the shallow-donor nature of hydrogen in ZnO.[32–34]

In general, five fundamental causes for doping limits can be identified; in the following sections, we illustrate these with specific instances where theory was able to solve important problems.

A. Solubility

In order to achieve a high free-carrier concentration, one obviously first needs to incorporate a high concentration of dopants. The equilibrium concentration of an impurity is given by

$$c = N_{\text{sites}} \exp^{-E^f/k_B T}, \tag{1}$$

where E^f is the *formation energy*, N_{sites} is the number of sites the impurity can be incorporated upon, k_B is the Boltzmann constant, and T the temperature. Equation (1) shows that impurities with a *high* formation energy occur in *low* concentrations. Equilibrium is assumed in Eq. (1); while most growth techniques are quite close to equilibrium (as determined by the mobility of point defects), kinetic limitations sometimes do occur. Even then, however, the magnitudes of formation energies are useful indicators of which impurities or defects are more likely to form.

The formation energy is not a constant but depends on the growth conditions.[35,36] For example, the formation energy of an oxygen donor in GaN is determined by the relative abundance of O, Ga, and N atoms, as expressed by the chemical potentials μ_{O}, μ_{Ga}, and μ_{N}, respectively. These chemical potentials are treated as variables. If the O donor is charged (as is expected when it has donated its electron), the formation energy depends further on the Fermi level, E_F, which acts as a reservoir for electrons. Forming a substitutional O donor requires the removal of one N atom and the addition of one O atom; the formation energy is therefore:

$$E^f(\text{GaN:O}_\text{N}^q) = E_{\text{tot}}(\text{GaN:O}_\text{N}^q) - E_{\text{tot}}(\text{GaN, bulk}) - \mu_{\text{O}} + \mu_{\text{N}} + qE_F. \tag{2}$$

First-principles calculations allow explicit derivation of $E_{\text{tot}}(\text{GaN:O}_\text{N}^q)$, the total energy derived for a system containing substitutional O on a N site. q is the charge state of the O donor. Similar expressions apply to other impurities and to the various native point defects. The Fermi level E_F is not an independent parameter, but is determined by the condition of charge neutrality. In principle, equations such as Eq. (2) can be formulated for every native defect and impurity in the material; the complete problem (including free-carrier concentrations in valence and conduction bands) can then be solved self-consistently, imposing charge neutrality. However, it is instructive to plot formation energies as a function of E_F in order to examine the behavior of defects and impurities when the doping level changes. An example is shown in Fig. 3.

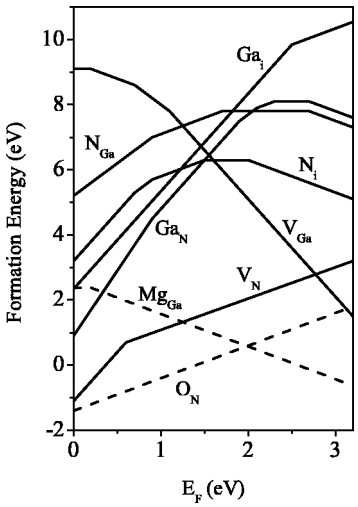

FIG. 3. Formation energies as a function of Fermi energy for native point defects and representative dopants (oxygen and magnesium) in GaN. The zero of Fermi energy is located at the top of the valence band, and Ga-rich conditions are assumed.

On the issue of solubility, Eq. (2) shows that the formation energy (and, hence, the concentration) of the impurity depends on the abundance of the impurity as well as the host constituents in the growth environment, as expressed by the chemical potentials.[29] Increasing the abundance of the impurity does not necessarily increase the concentration of impurities incorporated in the solid, because it may become more favorable for the impurity to form a different phase. In the case of oxygen in GaN, the solubility of oxygen is limited by formation of Ga_2O_3; in the case of Mg in GaN, the solubility-limiting phase is Mg_3N_2.

B. Ionization energy

The ionization energy of a dopant determines the fraction of dopants that will contribute free carriers at a given tem-

perature. A high ionization energy limits the doping efficiency: For instance, the ionization energy of Mg in GaN (around 200 meV) is so large that at room temperature only about 1% of Mg atoms are ionized. Ionization energies are largely determined by intrinsic properties of the semiconductor, such as effective masses, dielectric constant, etc.

C. Incorporation of impurities in other configurations

Impurities behave as proper dopants only when they are incorporated on a specific lattice site. For instance, in order for Mg in GaN to act as an acceptor, it needs to be incorporated on the gallium site. Incorporation on other lattice sites such as an interstitial position or a substitutional nitrogen site actually leads to *donor* behavior. For GaN doped with Mg, these configurations are energetically unfavorable; but in the case of a light atom, such as Li or Be, competition between substitutional and interstitial incorporation is a serious problem.[37]

Another instance of impurities incorporating in undesirable configurations consists of the so-called *DX* centers. The prototype *DX* center is Si in AlGaAs: In GaAs and in AlGaAs with low Al content, Si behaves as a shallow donor, but when the Al content exceeds a critical value, Si behaves as a deep level. Chang and Chadi explained this behavior in terms of Si moving off the substitutional site towards an interstitial position.[38] This ground-breaking work has also made it easier to recognize similar phenomena in other materials systems, such as the behavior of oxygen as a *DX* center in AlGaN.[39] The knowledge that oxygen becomes a deep compensating center in AlGaN when the Al concentration exceeds 40% has guided the interpretation of many experimental results.[40]

D. Compensation by native point defects

Native defects are point defects intrinsic to the semiconductor, such as vacancies, self-interstitials, and antisites. Native defects have frequently been invoked to explain doping problems in semiconductors. For instance, the problem of achieving *p*-type ZnSe was long attributed to self-compensation by native defects: It was hypothesized that every attempt to incorporate acceptors would be accompanied by the spontaneous generation of large numbers of native defects, acting as donors. First-principles calculations have shown that compensation by native defects is not an insurmountable problem.[41] Some degree of compensation is often unavoidable,[42] but this problem is not necessarily more severe in wide-band-gap semiconductors than in conventional semiconductors such as GaAs.

In some cases, native defects have been invoked as a source of doping, for instance in the case of unintentional *n*-type conductivity observed in GaN. As is evident from Fig. 3, the formation energy of nitrogen vacancies is too high under *n*-type conditions for them to be present in the large concentrations necessary to explain the observed unintentional conductivity. However, nitrogen vacancies may act as compensating centers in *p*-type GaN; and gallium vacancies can compensate *n*-type material.

Native point defects play an important role in self-diffusion as well as impurity diffusion. Calculated formation energies and migration barriers of vacancies and self-interstitials (starting with the pioneering work of the groups of Pantelides[43] and Joannopoulos[44]) have been instrumental in developing the models for diffusion processes that are extensively used in the microelectronics industry.

E. Compensation by foreign impurities

Although this source of compensation may seem obvious, we mention it for completeness: For instance, when incorporating acceptors in order to obtain *p*-type conductivity, impurities that act as donors should obviously be kept out of the growth system. ''Codoping,'' the intentional introduction of donors along with acceptors, has sometimes been advocated as a means of enhancing *p*-type conductivity. Indeed, these donors tend to shift the Fermi-level higher in the gap, which leads to a lower formation energy (and, hence, higher concentration) of the acceptor (see Fig. 3). Figure 3 also shows that this Fermi-level shift increases the formation energy of native defects that would act to compensate acceptors. Unfortunately, this shift in Fermi level persists after growth is completed, resulting in highly resistive material. Codoping can therefore only succeed if these donors can be removed from *p*-type layer in a postgrowth treatment. This can be accomplished if *hydrogen* is the codopant. Although the incorporation of hydrogen is often unintentional (e.g., in MOCVD or gas-source MBE), hydrogen actually plays a beneficial role during the growth through this codoping effect,[45] and can be removed in an electron-beam treatment[27] or a thermal anneal.[28] It was the discovery of these activation procedures that led to the rapid acceleration of device development in the nitride semiconductors.

Hydrogen can, of course, also play a beneficial role in passivating defects. This passivation plays an essential role in improving the electronic properties of amorphous and polycrystalline silicon. In addition, the degree of perfection required of Si/SiO_2 interfaces for silicon integrated circuits can only be achieved because of hydrogen passivation. Surprisingly, passivation with deuterium has been found to be significantly more stable than with hydrogen, both at Si surfaces[46] and at Si/SiO_2 interfaces.[47] This huge isotope effect has been explained in terms of the qualitatively different overlap between Si—H *versus* Si—D local vibrational modes and the silicon phonon spectrum.[48]

IV. CONCLUSIONS AND OUTLOOK

A. Interaction between theory and experiment

In this review, I have attempted to highlight several areas in which electronic materials theory and, in particular, first-principles calculations, have contributed to progress in materials and device physics. Such progress depends on a strong interaction between theory and experiment, an interaction that has always been fostered by AVS. This link is often facilitated by a direct comparison between calculated and experimentally measured quantities. For semiconductor

interfaces, the band offsets can obviously be directly compared, but atomic structure is a more difficult issue, because even the best microscopies may not be able to pinpoint positions and produce chemical identification of the atoms at a buried interface. Comparisons between simulated and measured cross-sectional scanning-tunneling microscopy images can be quite helpful.

In the area of point defects, calculations of local vibrational modes provide direct contact with infrared absorption or Raman spectroscopy, and hyperfine parameters[49] allow a direct comparison with experiments such as electron paramagnetic resonance (EPR), Mössbauer spectroscopy, or perturbed angular correlation (PAC). Such interactions between theory and experiment are often instrumental in providing a microscopic identification of a defect, which may not be possible on the basis of a measurement alone.

B. Progress in theoretical and computational methods

I have focused on first-principles calculations based on DFT. While tremendously powerful, DFT has its limitations, particularly in the treatment of excited states. Other techniques, such as quantum Monte Carlo, are proving their value in addressing problems that are beyond the reach of DFT.

In principle, the energy that enters in Eq. (1), or is minimized in determining atomic structures of interfaces, should be the *free energy*, i.e., entropy effects should be included. In general, these are small enough not to affect qualitative conclusions, but they can be essential in achieving quantitative accuracy.[50] Evaluating vibrational entropies is, in principle, a huge task, because it involves calculating the entire phonon spectrum. Fortunately, techniques such as thermodynamic integration are being developed that will enable us to more rigorously address properties at finite temperatures.

Computations are also evolving in the direction of multiscale modeling. Most problems indeed require addressing length and time scales that are many orders of magnitude beyond what can currently be accomplished in first-principles calculations. Multiscale modeling aims for the seamless integration of first-principles techniques with approaches more suited to other length and time scales.

C. Interfaces

A very exciting aspect of studies of interfaces is the prospect of *engineering* heterojunction band offsets. As reviewed in Ref. 12, such offset modification may be accomplished by tailoring the growth or by inclusion of interlayers. Theory can play an important role by providing guidance as to which structures can actually be achieved, given thermodynamic and/or kinetic limitations, and predicting the corresponding electronic structure. Most first-principles calculations to date have *assumed* one or a small set of specific structures, and produced a band-offset value for those structures. A future challenge will be to *calculate* (as opposed to *assume*) the geometries and interfacial structures. This type of investigation is a lot more demanding, since total-energy calculations have to be carried out for a large number of possible structures. The Car–Parrinello approach, which allows simultaneous optimization of atomic and electronic degrees of freedom, is tremendously helpful in obtaining the detailed atomic positions for a given interfacial structure. But many different structures may exist, for instance when different cation/anion ratios lead to distinct stoichiometries near a compound semiconductor interface, or when atomic mixing occurs over several layers. Each of these structures requires a separate calculation. Developing a systematic approach for evaluating the energetics of such a large number of structures is very desirable.

Instead of studying the thermodynamics of a large number of possible structures, one can also envision obtaining the actual interfacial structure by performing explicit modeling of the growth process. Explicit growth simulations are outside the capability of current first-principles simulations, since both the time and length scales are many orders of magnitude larger than can be achieved within current computational limits. Multiscale modeling is essential here, for instance with kinetic Monte Carlo methods that employ parameters obtained from first principles.[51]

Another emerging area involves interfaces between semiconductors that exhibit spontaneous and/or piezoelectric polarization. In the III-nitride system, first-principles calculations have been ahead of experiment in quantitatively assessing these effects,[52] but the extent to which the presence of the polarization fields may affect the atomic structure or defect formation has not yet been addressed.

D. Defects and impurities at interfaces

Even the most sophisticated growth techniques cannot produce an ideal, abrupt, structurally perfect interface. Indeed, just like in the bulk, thermodynamics predicts that a certain concentration of defects is unavoidable at an interface. In fact, defect formation may be enhanced near interfaces: Computations have shown that the formation energy of point defects can be significantly lower near a surface, and push impurity concentrations well above the solubility limit.[53] Walukiewicz[42] and Duke and Mailhiot[54] have discussed how defect formation and atomic rearrangements at the metal/semiconductor interface may affect Fermi-level pinning. These examples illustrate that interface-specific defect phenomenona can significantly affect materials properties, and the capability to perform large-scale computations will make it easier to address these issues in the future.

The occurrence of atomic mixing at heterovalent polar interfaces, which was first discussed in detail at the PCSI conferences, is an interesting example of defect formation at heterojunctions. As pointed out by Harrison[55] and by Martin,[56] polar surfaces and interfaces of compound semiconductors are invariably reconstructed from planar geometries. For instance, an analysis of the electrostatic potential at a Ge/GaAs(001) junction shows that the ideal interface would be energetically unstable; a certain amount of disorder is necessary for stabilizing the structure. Various types of reconstructions can lead to very different band offsets at the

interface; the problem was addressed in detail by Grant and Harrison.[57] Full first-principles calculations are demanding, because of the large interface cells required; however, studies of this type could contribute in important ways to experimental progress in these ''difficult'' materials systems.

E. First-principles materials design

Even though space constraints forced us to focus on only a narrow range of areas, I hope to have provided persuasive evidence that theory has played an increasingly influential role in the development of electronic materials, and that AVS has provided a nurturing environment in which these contributions are stimulated and recognized. The role of theory is bound to expand now that first-principles computations are truly capable of accurately predicting materials properties. The impact will be particularly strong in the realm of nanoscience, where predictive modeling will be crucial to understanding and controlling structures at the smallest length scales.

ACKNOWLEDGMENTS

This work was supported in part by ONR (Contract No. N00014-02-C-0433). Thanks are due to the many collaborators the author has had the pleasure of working with over the years, in particular R. Martin, S. Pantelides, and J. Neugebauer.

[1]W. Kohn, ''Nobel Lecture: Electronic structure of matter-wave functions and density functionals,'' Rev. Mod. Phys. **71**, 1253 (1999).
[2]M. Cohen, ''Pseudopotentials and total energy calculations,'' Phys. Scr. T **1**, 5 (1982).
[3]R. Car and M. Parrinello, ''Unified approach for molecular dynamics and density-functional theory,'' Phys. Rev. Lett. **55**, 2471 (1985).
[4]P. J. Feibelman, ''Surface theory moves into the real world,'' J. Vac. Sci. Technol. A, these proceedings.
[5]C. B. Duke, ''Semiconductors surface reconstruction: the structural chemistry of two-dimensional surface compounds,'' Chem. Rev. (Washington, D.C.) **96**, 1237 (1996).
[6]H. Kroemer, ''Nobel Lecture: Quasielectric fields and band offsets: Teaching electrons new tricks,'' Rev. Mod. Phys. **73**, 783 (2001).
[7]K. von Klitzing, ''The quantized Hall effect,'' Rev. Mod. Phys. **58**, 519 (1986).
[8]H. L. Stormer, ''Nobel Lecture: The fractional quantum Hall effect,'' Rev. Mod. Phys. **71**, 875–889 (1999).
[9]R. L. Anderson, Solid-State Electron. **5**, 341 (1962).
[10]C. B. Duke, ''Twenty years of semiconductor surface and interface structure determination and predictions: The role of the annual conferences on the Physics and Chemistry of Semiconductor Interfaces,'' J. Vac. Sci. Technol. **11**, 1336 (1993).
[11]J. Tersoff, ''The theory of heterojunction band lineups,'' in *Heterojunctions: A Modern View of Band Discontinuities and Applications*, edited by G. Margaritondo and F. Capasso (North–Holland, Amsterdam, 1987).
[12]A. Franciosi and C. G. Van de Walle, ''Heterojunction band offsets engineering,'' Surf. Sci. Rep. **25**, 1 (1996).
[13]C. G. Van de Walle and R. M. Martin, ''Theoretical study of Si/Ge interfaces,'' J. Vac. Sci. Technol. B **3**, 1256 (1985).
[14]S. Baroni, R. Resta, A. Baldereschi, and M. Peressi, ''Can we tune the band offset at semiconductor heterojunctions?,''in *Spectroscopy of Semiconductor Microstructures*, edited by G. Fasol, A. Fasolino, and P. Lugli (Plenum, London, 1989), pp. 251–271.
[15]W. E. Pickett, S. G. Louie, and M. L. Cohen, ''Ge–GaAs (110) interface: A self-consistent calculation of interface states and electronic structure,'' Phys. Rev. Lett. **39**, 109 (1977).
[16]C. G. Van de Walle and R. M. Martin, ''Theoretical calculations of semiconductor heterojunction discontinuities,'' J. Vac. Sci. Technol. B **4**, 1055 (1986).
[17]C. Stampfl and C. G. Van de Walle, ''Density-functional calculations for III–V nitrides using the local density approximation and the generalized gradient approximation,'' Phys. Rev. B **59**, 5521 (1999).
[18]X. Zhu and S. G. Louie, ''Quasiparticle band structure of thirteen semiconductors and insulators,'' Phys. Rev. B **43**, 14142 (1991).
[19]J. A. Van Vechten, ''Ionization potentials, electron affinities, and band offsets,'' J. Vac. Sci. Technol. B **3**, 1240 (1985).
[20]W. A. Harrison, ''Elementary theory of heterojunctions,'' J. Vac. Sci. Technol. **14**, 1016 (1977).
[21]W. R. Frensley and H. Kroemer, ''Prediction of semiconductor heterojunction discontinuities from bulk band structures,'' J. Vac. Sci. Technol. **13**, 810 (1977).
[22]C. G. Van de Walle, ''Band lineups and deformation potentials in the model-solid-theory,'' Phys. Rev. B **39**, 1871 (1989).
[23]C. Tejedor and F. Flores, ''A simple approach to heterojunctions,'' J. Phys. C **11**, L19 (1977).
[24]S. G. Louie, J. R. Chelikowsky, and M. L. Cohen, ''Theory of semiconductor surface states and metal–semiconductor interfaces,'' J. Vac. Sci. Technol. **13**, 790 (1976).
[25]M. Cardona and N. Christensen, ''Band offsets in tetrahedral semiconductors,'' J. Vac. Sci. Technol. B **6**, 1285 (1988).
[26]R. M. Park, M. B. Troffer, C. M. Rouleau, J. M. DePuydt, and M. A. Haase, ''*p*-type ZnSe by nitrogen atom beam doping during molecular beam epitaxial growth,'' Appl. Phys. Lett. **57**, 2127 (1990).
[27]H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, ''*p*-type conduction in Mg-doped GaN treated with low-energy electron beam irradiation (LEEBI),'' Jpn. J. Appl. Phys., Part 2 **28**, L2112 (1989).
[28]S. Nakamura, T. Mukai, M. Senoh, and N. Iwasa, ''Thermal annealing effects on *p*-type Mg-doped GaN films,'' Jpn. J. Appl. Phys., Part 2 **31**, L139 (1992).
[29]O. F. Sankey and R. W. Jansen, ''Prediction of equilibrium concentrations of defects and factors that influence them,'' J. Vac. Sci. Technol. B **6**, 1240 (1988).
[30]J. Neugebauer and C. G. Van de Walle, ''Atomic geometry and electronic structure of native defects in GaN,'' Phys. Rev. B **50**, 8067 (1994).
[31]C. G. Van de Walle, ''Hydrogen as a cause of doping in ZnO,'' Phys. Rev. Lett. **85**, 1012 (2000).
[32]S. F. J. Cox, E. A. Davis, S. P. Cottrell, P. J. C. King, J. S. Lord, J. M. Gil, H. V. Alberto, R. C. Vilão, J. Piroto Duarte, N. Ayres de Campos, A. Weidinger, R. L. Lichti, and S. J. C. Irvine, ''Experimental confirmation of the predicted shallow donor hydrogen state in zinc oxide,'' Phys. Rev. Lett. **86**, 2601 (2001).
[33]K. Shimomura, K. Nishiyama, and R. Kadono, ''Electronic structure of the muonium center as a shallow donor in ZnO,'' Phys. Rev. Lett. **89**, 255505 (2002).
[34]D. M. Hofmann, A. Hofstaetter, F. Leiter, H. Zhou, F. Henecker, B. K. Meyer, S. B. Orlinskii, J. Schmidt, and P. G. Baranov, Phys. Rev. Lett. **88**, 045504 (2002).
[35]S. B. Zhang and J. E. Northrup, ''Chemical potential dependence of defect formation energies in GaAs: Application to Ga self-diffusion,'' Phys. Rev. Lett. **67**, 2339 (1991).
[36]C. G. Van de Walle, D. B. Laks, G. F. Neumark, and S. T. Pantelides, ''First-principles calculations of solubilities and doping limits: Li, Na, and N in ZnSe,'' Phys. Rev. B **47**, 9425 (1993).
[37]J. Neugebauer and C. G. Van de Walle, ''Chemical trends for acceptor impurities in GaN,'' J. Appl. Phys. **85**, 3003 (1999).
[38]D. J. Chadi and K. J. Chang, ''Theory of the atomic and electronic structure of DX centers in GaAs and $Al_xGa_{1-x}As$ alloys,'' Phys. Rev. Lett. **61**, 873 (1988).
[39]C. G. Van de Walle, ''DX center formation in wurtzite and zinc-blende AlGaN,'' Phys. Rev. B **57**, 2033 (1998).
[40]M. D. McCluskey, N. M. Johnson, C. G. Van de Walle, D. P. Bour, M. Kneissl, and W. Walukiewicz, ''Metastability of oxygen donors in AlGaN,'' Phys. Rev. Lett. **80**, 4008 (1998).
[41]D. B. Laks, C. G. Van de Walle, G. F. Neumark, and S. T. Pantelides, ''Role of native defects in wide band-gap semiconductors,'' Phys. Rev. Lett. **66**, 648 (1991).
[42]W. Walukiewicz, J. Vac. Sci. Technol. B **6**, 1257 (1988).

[43]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, ''Microscopic theory of atomic diffusion mechanisms in silicon,'' Phys. Rev. Lett. **52**, 1814 (1984).
[44]Y. Bar-Yam and J. D. Joannopoulos, ''Barrier to migration of the silicon self-interstitial,'' Phys. Rev. Lett. **52**, 1129 (1984).
[45]J. Neugebauer and C. G. Van de Walle, ''Role of hydrogen in doping of GaN,'' Appl. Phys. Lett. **68**, 1829 (1996).
[46]P. Avouris, R. E. Walkup, A. R. Rossi, T. C. Shen, G. C. Abeln, J. R. Tucker, and J. W. Lyding, ''STM-induced H atom desorption from Si(100): Isotope effects and site selectivity,'' Chem. Phys. Lett. **257**, 148 (1996).
[47]K. Cheng, J. Lee, Z. Chen, S. A. Shah, K. Hess, J.-P. Leburton, and J. W. Lyding, ''Fundamental connection between hydrogen/deuterium desorption at silicon surfaces in ultrahigh vacuum and at oxide/silicon interfaces in metal-oxide semiconductor devices,'' J. Vac. Sci. Technol. B **19**, 1119 (2001).
[48]C. G. Van de Walle, ''Hydrogen in silicon: Fundamental properties and consequences for devices,'' J. Vac. Sci. Technol. A **16**, 1767 (1998).
[49]C. G. Van de Walle and P. E. Blöchl, ''First-principles calculations of hyperfine parameters,'' Phys. Rev. B **47**, 4244 (1993).
[50]C. G. Van de Walle and J. Neugebauer, ''Role of hydrogen in surface reconstructions and growth of GaN,'' J. Vac. Sci. Technol. B **20**, 1640 (2002).
[51]P. Kratzer, E. Penev, and M. Scheffler, ''First-principles studies of kinetics in epitaxial growth of III–V semiconductors,'' Appl. Phys. A: Mater. Sci. Process. **75**, 79 (2002).
[52]F. Bernardini and V. Fiorentini, ''Polarization fields in nitride nanostructures: 10 points to think about,'' Appl. Surf. Sci. **166**, 23 (2000).
[53]J. Tersoff, ''Enhanced solubility of impurities and enhanced diffusion near crystal surfaces,'' Phys. Rev. Lett. **74**, 5080 (1995).
[54]C. B. Duke and C. Mailhiot, ''A microscopic mode of metal semiconductor contacts,'' J. Vac. Sci. Technol. B **3**, 1170 (1985).
[55]W. A. Harrison, ''Theory of polar semiconductor surfaces,'' J. Vac. Sci. Technol. **16**, 1492 (1979).
[56]R. M. Martin, ''Atomic reconstruction at polar interfaces of semiconductors,'' J. Vac. Sci. Technol. **17**, 978 (1980).
[57]R. W. Grant and W. A. Harrison, ''Dipoles at polar heterojunction interfaces,'' J. Vac. Sci. Technol. B **6**, 1295 (1988).