First-principles calculations for defects and impurities: Applications to III-nitrides

Chris G. Van de Walle and Jörg Neugebauer

Citation: J. Appl. Phys. **95**, 3851 (2004); doi: 10.1063/1.1682673
View online: http://dx.doi.org/10.1063/1.1682673
View Table of Contents: http://jap.aip.org/resource/1/JAPIAU/v95/i8
Published by the American Institute of Physics.

APPLIED PHYSICS REVIEWS

First-principles calculations for defects and impurities: Applications to III-nitrides

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

Jörg Neugebauer
Universität Paderborn, Fakultät für Naturwissenschaften, Warburger Str. 100, D-33095 Paderborn, Germany

(Received 14 July 2003; accepted 26 January 2004)

First-principles calculations have evolved from mere aids in explaining and supporting experiments to powerful tools for predicting new materials and their properties. In the first part of this review we describe the state-of-the-art computational methodology for calculating the structure and energetics of point defects and impurities in semiconductors. We will pay particular attention to computational aspects which are unique to defects or impurities, such as how to deal with charge states and how to describe and interpret transition levels. In the second part of the review we will illustrate these capabilities with examples for defects and impurities in nitride semiconductors. Point defects have traditionally been considered to play a major role in wide-band-gap semiconductors, and first-principles calculations have been particularly helpful in elucidating the issues. Specifically, calculations have shown that the unintentional *n*-type conductivity that has often been observed in as-grown GaN cannot be attributed to nitrogen vacancies, but is due to unintentional incorporation of donor impurities. Native point defects may play a role in compensation and in phenomena such as the yellow luminescence, which can be attributed to gallium vacancies. In the section on impurities, specific attention will be focused on dopants. Oxygen, which is commonly present as a contaminant, is a shallow donor in GaN but becomes a deep level in AlGaN due to a *DX* transition. Magnesium is almost universally used as the *p*-type dopant, but hole concentrations are still limited. Reasons for this behavior are discussed, and alternative acceptors are examined. Hydrogen plays an important role in *p*-type GaN, and the mechanisms that underlie its behavior are explained. Incorporating hydrogen along with acceptors is an example of codoping; a critical discussion of codoping is presented. Most of the information available to date for defects and impurities in nitrides has been generated for GaN, but we will also discuss AlN and InN where appropriate. We conclude by summarizing the main points and looking towards the future. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1682673]

TABLE OF CONTENTS

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

I. INTRODUCTION

A. Goals

The properties of materials are often controlled by defects and impurities. This is particularly true in the case of semiconductors, where the incorporation of impurities in small concentrations determines the electrical conductivity. The fabrication of *p*-type and *n*-type doped layers underlies the design of virtually all electronic and optoelectronic devices. To achieve such control, comprehensive knowledge of the fundamental processes that control doping is required. In recent years, first-principles calculations have made important contributions to this knowledge.

Thanks to algorithmic developments as well as increases in computer power, first-principles calculations are now reaching unprecedented levels of accuracy in treating increasingly larger systems at the microscopic level. State-of-the-art calculations for solids are based on density-functional theory and produce detailed information about atomic structure (including relaxations), wave functions, charge densities, potentials, and energies. All of these data can be used to elucidate the properties of impurities and point defects, as will be illustrated with many examples in this review. The aspect that we will focus on most closely, and that has proved crucial in relation to studying doping, is the formation energy of point defects. We will describe a formalism that allows calculation of defect and impurity concentrations based on first-principles formation energies; this formalism also addresses the energetics of charge states, and hence the thermodynamic transition levels associated with deep and shallow impurities and defects.

The formalism is entirely general in nature and could be applied to any semiconductor or insulator. The goal of the first part of the article (Sec. II) is to provide an overview of the state-of-the-art methodology for performing first-principles calculations for defects and impurities. The aim is *not* to address the fundamentals of density-functional or pseudopotential theory, or to provide a guide on how to run first-principles computer codes–various excellent reviews are available for that purpose.[1–3] Rather, we intend Sec. II to be useful for a computational theorist getting started in defect calculations, as well as for more experienced practitioners looking for a reference to some of the details that are important in practical calculations. In addition, this section is intended to be accessible to experimentalists who are curious about the background of the computational work or have questions about some of the underlying assumptions.

The methodology described in Sec. II will be illustrated with specific examples in the latter half of the review. We will focus on one specific material system, namely, the III–V nitrides. Two Applied Physics Reviews have recently appeared that cover this set of materials. One, by Vurgaftman and Meyer,[4] presents a comprehensive compilation of band parameters for all of the nitrogen-containing III–V semiconductors that have been investigated to date. The second, by Bhuiyan *et al.*,[5] reviews the growth, characterization, and properties of InN and contains a brief discussion of defects. In the present article we focus on the properties of defects and impurities in III-nitrides. Section III contains results for native defects, while Sec. IV describes impurities. A review of computational studies on GaN by Estreicher and Boucher[6] covered developments through 1995, but much has happened since then.

In the remainder of this Introduction we provide some additional motivation for defect studies in nitrides, an important issue being the realization of adequate doping. We therefore also provide a discussion of doping limitations in semiconductors in general.

B. Doping problems in nitrides

Within the past decade the nitride semiconductors have emerged as a very important materials system because they are uniquely suited to light emission in the green, blue, and UV regions of the spectrum–wavelength regions that were previously not accessible with solid-state light emitters.

n-type doping of nitrides has never been a problem; in fact, as-grown material has often exhibited unintentional *n*-type conductivity, the cause of which was widely debated. *n*-type doping with electron concentrations exceeding 10^{19} cm^{-3} can routinely be achieved. *p*-type doping, however, has traditionally been very difficult. *p*-type doping was first achieved by Amano *et al.* in 1989,[7] who observed that Mg-doped GaN grown by MOCVD (metal-organic chemical vapor deposition) was highly resistive after growth, but could be activated by low-energy electron beam irradiation. Nakamura *et al.*[8] subsequently showed that the Mg activation can also be achieved by thermal annealing at 700 °C under N_2 ambient. Nakamura *et al.* further observed that the process was reversible, with *p*-type GaN reverting to semi-insulating when annealed in a NH_3 ambient, revealing the crucial role played by hydrogen. Since then, hole concentrations on the order of 10^{18} cm^{-3} have been achieved and used in devices. Still, the limited conductivity of *p*-type doped layers constitutes an impediment for progress in device applications.

C. Fundamental causes of doping limitations

When discussing doping of semiconductors, and its inherent limitations and difficulties, a number of factors need to be considered. Here we enumerate them and illustrate them with the example of *p*-type doping of GaN.

1. Solubility

To achieve a high free-carrier concentration, one obviously needs to achieve a high concentration of the dopant impurity. The solubility corresponds to the maximum concentration that the impurity can attain in the semiconductor, under conditions of thermodynamic equilibrium. This concentration depends on temperature, and on the abundance of the impurity as well as the host constituents in the growth environment. Increasing the abundance of the impurity (or its chemical potential, see Sec. II E) 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. For instance, we will see that the solubility of Mg in GaN is limited by formation of Mg_3N_2 (see Sec. II E). In previous work on wide-band-gap semiconductors, it was found that the hole concentration in ZnSe and ZnTe is limited by the solubility of the acceptor impurities (Na, Li, and N).[9]

2. Ionization energy

The ionization energy of a dopant determines the fraction of dopants that will contribute free carriers at a given temperature. 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. This means that a Mg concentration of 10^{20} cm^{-3} only leads to a hole concentration of about 10^{18} cm^{-3}. Ionization energies are largely determined by intrinsic properties of the semiconductor, such as effective masses, dielectric constant, etc. Switching to a different acceptor has therefore no dramatic effect on the ionization energy.

3. Incorporation of impurities in other configurations

In order for Mg in GaN to act as an acceptor, it needs to be incorporated on the gallium site. There has been concern about Mg incorporating in other positions in the lattice, such as an interstitial position, or substituting for a nitrogen atom (essentially an antisite configuration). For GaN:Mg (GaN doped with Mg), we have shown that these other configurations are always much higher in energy, and hence will not form.[10] In other cases, however, such competition may be a serious issue: for instance, while Li on the Ga site in GaN forms an acceptor, Li on an interstitial site is a donor, and because of its small size it is energetically very favorable.[11] This can obviously lead to serious self-compensation.

Another instance of impurities incorporating in undesirable configurations consists of the so-called *DX* centers. The prototype *DX* center is Si in AlGaAs (for a review, see Ref. 12). 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. This has been explained in terms of Si moving off the substitutional site, towards an interstitial position.[13] It has been found that oxygen forms a *DX* center in AlGaN, when the Al content exceeds about 30%.[14] This prediction has been confirmed experimentally.[15]

4. Compensation by native point defects

Native defects are point defects intrinsic to the semiconductor, such as vacancies (missing atoms), self-interstitials (additional atoms incorporated on sites other than substitutional sites), and antisites (in a compound semiconductor, a cation sitting on a nominal anion site, or vice versa). 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. In the case of ZnSe, it was shown that compensation by native defects is not an insurmountable problem.[16] Some degree of compensation is often unavoidable, but this problem is not necessarily more severe in wide-band-gap semiconductors than in, say, GaAs. For GaN, we have found that compensation by vacancies can limit the doping level in some cases: gallium vacancies (V_{Ga}) are acceptors and compensate *n*-type GaN; nitrogen vacancies (V_N) are single donors and compensate *p*-type GaN.

Native defects have sometimes been invoked to play a role that goes beyond compensation, namely, to act as a source of doping. For instance, the frequently observed *n*-type conductivity of as-grown GaN was long attributed to nitrogen vacancies. Nitrogen vacancies indeed act as shallow donors, but their incorporation in *n*-type GaN costs too much energy for them to be present in the large concentrations necessary to explain the observed *n*-type conductivity.[17]

5. Compensation by foreign impurities

This source of compensation may seem rather obvious, but we mention it for completeness, and it often plays a

crucial role: for instance, when doping with acceptors (such as magnesium) in order to obtain *p*-type conductivity, impurities that act as donors (such as oxygen) should be carefully controlled. Such control may be more difficult than is obvious at first sight. For instance, for reasons that will be explained in Sec. IV B 3, the presence in the growth system of a contaminating impurity with donor character may lead to a much larger incorporation of this impurity in *p*-type material than in *n*-type material.

Each and every one of the factors listed here can be explicitly examined using a computational approach, and Secs. III and IV will contain explicit examples of results.

II. METHODOLOGY

Modern first-principles calculations have had a major impact on the understanding of defects and impurities in semiconductors. With the capability to calculate total energies, it became possible to investigate the atomic structure of the defect; i.e., the stable position in the host lattice, the relaxation of the surrounding atoms, as well as the energy along a migration path.[18–20] More recently, formalisms have been developed to use the total energy of the defect to calculate its concentration, under the assumption of thermodynamic equilibrium.[21,22] The same formalism can also be applied to the calculation of impurity solubilities.[9,23] In the following sections we describe this formalism in detail.

A. Concentrations of defects

In thermodynamic equilibrium the concentration c of an impurity, defect, or complex is given by the expression

$$c = N_{\text{sites}} N_{\text{config}} \exp(-E^f/kT) \quad . \tag{1}$$

Here, E^f is the formation energy (see Sec. II B 4), N_{sites} is the number of sites in the lattice (per unit volume) where the defect can be incorporated, k is Boltzmann's constant, and T is the temperature. N_{config} is the number of equivalent configurations in which the defect can be incorporated. For vacancies, antisites, and subsitutional defects $N_{\text{config}}=1$ if no symmetry breaking occurs. If symmetry breaking occurs or if complexes are formed it is the number of inequivalent configurations in which the defect can be incorporated on the same site.

1. Justification for the thermodynamic equilibrium approach

The expression for concentration as a function of formation energy [Eq. (1)] is, strictly speaking, only valid in thermodynamic equilibrium. Growth of semiconductors is obviously a nonequilibrium process. How then do we justify using Eq. (1) and attaching relevance to formation energies? The justification is based on the argument that many growth situations are close enough to equilibrium to warrant the use of the equilibrium approach. An important consideration here is that not *all* aspects of the process need to be in equilibrium in order to justify the use of equilibrium expressions for defects and impurities. What is required is a sufficiently high mobility of the relevant impurities and point defects to allow them to equilibrate at the temperatures of interest.

2. Molecular beam epitaxy (MBE) versus MOCVD growth

MOCVD growth of GaN is carried out at high temperatures (usually between 1000 and 1100 °C). The mobility of various point defects,[24] both on the Ga and on the N sublattice, should be sufficiently high to allow equilibration of the defects and impurities that are being incorporated in the bulk. Under these circumstances, point defects will incorporate in concentrations determined by their formation energies, which, as discussed below, depend on the relative abundance of the various species in the growth environment. MBE growth, on the other hand, is carried out at lower temperatures ($\sim$800 °C), and the assumption of thermodynamic equilibrium is less likely to be satisfied. MBE-grown material may thus in principle exhibit point-defect concentrations that deviate from their equilibrium values.

We do want to make the point that, even if the equilibrium conditions are not met that would justify use of Eq. (1) to derive concentrations, the formation energies defined in Sec. II B 4 are still physically meaningful. Nonequilibrium implies that once certain high-energy defects form, kinetic barriers may preserve them, even if their concentration exceeds the nominal equilibrium value. It should be clear, however, that defects with a *high* formation energy will always be unlikely to form, since a lot of energy needs to be expended in their creation, and the driving force to lowering the energy is large.

3. Surface effects

An exception of the above argument is the creation of defects at the surface: There the defect formation energy may be significantly different from the bulk formation energy, due to structural as well as electronic effects. The structural effects can be due, for instance, to local strains underneath specific features of reconstructed surfaces, as discussed by Tersoff in the case of C incorporation in Si.[25] The electronic effects are related to the band bending that is usually present near semiconductor surfaces. As discussed in Sec. II B, the formation energies of charged defects and impurities depend sensitively on the Fermi level, and near the surface the position of the Fermi level with respect to the band edges can be strongly shifted due to the presence of space-charge layers.

Typically, defect formation energies at the surface are lower, resulting in high defect concentrations at the surface. In cases where complete equilibration within the bulk of the growing material is not accomplished, it may still be possible for limited equilibration to occur within the first few atomic layers beneath the growing surface, where diffusion over short length scales may still be possible. Since the defect formation energy quickly converges to its bulk value when moving the defect from the surface to bulk (see, e.g., Ref. 26) even limited equilibration within the first few atomic layers is sufficient to achieve bulk defect concentrations.

A comprehensive examination of the effects of surfaces on the incorporation of defects and impurities is beyond the scope of the present review. In the nitrides, we are aware of the following studies: Bungaro *et al.*[27] investigated Mg incorporation at GaN(0001) surfaces; Zywietz *et al.* calculated

oxygen on GaN(0001) and $(000\bar{1})$ surfaces;[28] Northrup[29] investigated Be incorporation at GaN(0001) in the presence of indium; and Rosa *et al.* studied Si on GaN(0001).[30] Hydrogen on GaN surfaces, finally, was studied in Ref. 31.

B. Formation energies

1. Techniques for estimating or calculating formation energies

Hartree–Fock based models commonly employ quantum-chemistry approaches that have been successfully applied to atoms and molecules. The main problems with the technique are the neglect of correlation effects and the computational demands: *ab initio* Hartree–Fock methods can only be applied to systems with small numbers of atoms. The reason is that these methods require the evaluation of a large number of multicenter integrals. Simpler semiempirical methods have been developed that either neglect or approximate some of these integrals. The accuracy and reliability of these methods is hard to assess.

Information about point defects and impurities can in principle also be obtained from tight-binding calculations. Tight-binding methods use the fact that within a local basis set the Hamilton matrix elements rapidly decrease with increasing distance between the orbitals. Thus, instead of having to diagonalize the full Hamiltonian matrix most of the matrix elements vanish and only a *sparse* matrix has to be diagonalized. Depending on how the remaining Hamilton matrix elements are determined one can distinguish two main approaches: (i) empirical tight-binding methods and (ii) first-principles tight-binding methods.

An important problem for the empirical tight-binding approach is the choice of parameters, for which there is no consistent prescription. The shortcomings of tight-binding theory were highlighted in early work on point defects in GaN, where the a_1 state of the nitrogen vacancy was found to lie close to the bottom of the conduction band.[32,33] In reality, this state lies near the top of the valence band.[17] The location of this state is determined by the strong interaction between Ga dangling bonds surrounding the nitrogen vacancy; the tight-binding calculations of Refs. 32 and 33, which only took nearest-neighbor interactions into account, failed to include this interaction, resulting in an incorrect positioning of the defect levels. The origin of the failing of a first nearest-neighbor tight-binding method is the extremely ionic character and the (correspondingly) small lattice constant of the group-III nitrides.[17] Second nearest-neighbor interactions therefore play an important role.

First-principles tight-binding methods use local orbitals to *explicitly* calculate the Hamilton matrix elements. The choice of orbitals is critical: instead of the standard local orbitals (e.g., atomic orbitals), specifically designed and extremely localized orbitals are used. Approximations are made in neglecting some of the multi-center integrals and charge self-consistency. For group-III nitrides which are highly ionic, this approximation is not well satisfied. Significant improvement has been found by using a point-charge model to take charge transfer and polarizability into account.[34]

2. Density-functional theory

Density-functional theory (DFT) calculations based on pseudopotentials, a plane-wave basis set, and a supercell geometry are now regarded as a standard for performing first-principles studies of defects in semiconductors. DFT in the local density approximation (LDA)[35] allows a description of the many-body electronic ground state in terms of single-particle equations and an effective potential. The effective potential consists of the ionic potential due to the atomic cores, the Hartree potential describing the electrostatic electron-electron interaction, and the exchange-correlation potential that takes into account the many-body effects. This approach has proven to describe with high accuracy such quantities as atomic geometries, charge densities, formation energies, etc. Most of the results described in Secs. III and IV are based on an implementation of pseudopotential-density-functional theory described in Ref. 3.

An analysis of GaN defect and bulk calculations showed that the Ga $3d$ electrons are not chemically inert but play an important role for the chemical bonding.[36–38] Thus, in general the Ga $3d$ electrons cannot be simply treated as core electrons (which would be computationally less expensive) but have to be explicitly treated as valence electrons.[38] The localized nature of the Ga $3d$ states significantly increases the computational demand, requiring an energy cutoff of at least 60 Ry in the plane-wave expansions. An attractive alternative is to use the so-called ''nonlinear core correction'' $(nlcc)$,[39] in combination with soft Troullier–Martins pseudopotentials[40] for which an energy cutoff of 40 Ry suffices.

The explicit inclusion of Ga $3d$ states as valence states yields demonstrable improvements in the structural properties as well as in the enthalpy of formation. In large part, these improvements can also be achieved by using the $nlcc$.[41] For the electronic structure, however, the benefit of explicitly including Ga $3d$ states is unclear. DFT-LDA places these d states too high in the band structure, causing them to be closer to the valence-band maximum (VBM); $p-d$ repulsion then causes the VBM to be pushed up, leading to a decrease in the band gap. This effect has actually been found to persist in GW calculations.[42] The inclusion of $3d$ states in calculations of band-structure-related properties such as alloy band gaps or deformation potentials may therefore not necessarily be an improvement compared to the use of the $nlcc$, where the effects of d states are only approximated and the anomalous repulsion between d states and the VBM is absent.

Defect and impurity calculations should be carried out at the theoretical lattice constant, in order to avoid a spurious elastic interaction with defects or impurities in neighboring supercells. Since our purpose is to investigate properties for a single, isolated defect or impurity in an infinite solid, the lattice constant of the supercell should correspond to that of the unperturbed host. It has sometimes been suggested that, in the process of relaxing the host atoms around the defect, the volume of the supercell should be relaxed as well. Such a volume relaxation would actually correspond to finding the lattice constant of a bulk system containing an ordered array of impurities at very high concentration. This could result in

a very different lattice constant from the one we are interested in, corresponding to a dilute system.

Well-converged calculations with good-quality pseudopotentials should produce lattice parameters within a few percent of the experimental value. When the Ga $3d$ electrons are explicitly included, with an 80 Ry energy cutoff,[43] we find $a^{\text{th}}=3.193$ Å (compared with $a^{\text{exp}}=3.19$ Å). The calculated c/a ratio is 1.634 (experiment: 1.627), very close to the ideal c/a ratio of $\sqrt{8/3}=1.633$. For zinc-blende GaN, we find $a^{\text{th}}=4.518$ Å, which is (to within 0.002 Å) $\sqrt{2}$ larger than the wurtzite lattice constant. Using the *nlcc* and a 40 Ry energy cutoff, the values are $a^{\text{th}}=3.089$ Å and $c/a=1.633$; and for zinc-blende GaN: $a^{\text{th}}=4.370$ Å, again (to within 0.001 Å) $\sqrt{2}$ larger than the wurtzite lattice constant.

For calculations of defects and impurities in semiconductors within the density-functional approach, use of the local density approximation seems to be well justified. The generalized gradient approximation (GGA) apparently does not offer any advantages, neither for bulk properties[43] nor for formation energies of point defects.[44] The quantitative differences that exist for the latter can be explained in terms of differences in the lattice constant (which gives rise to differences in the band gap) and in calculated formation enthalpies between LDA and GGA.[44,41]

3. Beyond density-functional theory

One shortcoming of the DFT approach is its failure to produce accurate excited-states properties—the band gap is commonly underestimated.[45,46] Research is currently under way to overcome this limitation of density-functional theory. No method is currently available that goes beyond DFT and provides total-energy capability for the large supercell calculations required to investigate defects. Even methods aimed solely at calculating the band structure, such as the *GW* approach,[47–49] are currently prohibitively expensive for large cells.

A promising approach that yields bulk band structures in good agreement with experiment was recently introduced, based on self-interaction and relaxation-corrected (SIRC) pseudopotentials.[50] This approach was used in Ref. 51 to perform calculations of the electronic structure of bulk InN and of various native point defects. Defects can introduce levels in the band gap, and when occupied with electrons these levels contribute to the total energy of the system; it is therefore important to consider the effect of the band-gap error on the calculated properties of defects.

The study of Ref. 51 indicated that the character of the defect-induced states is very similar in SIRC calculations compared to LDA, but conduction-band related states are shifted to higher energies. The SIRC approach currently does not simply allow evaluation of total energies, and therefore the effects of the calculated changes in the band structure on the total energy of the defect were only estimated, without inclusion of selfconsistency. Still, it could be concluded that while total energies may be affected in some instances, at least for InN the shifts do not alter the conclusions based on the DFT calculations.

The SIRC potentials used in Ref. 51 produced a band gap of wurtzite InN of 1.55 eV, which was considered in reasonable agreement with the commonly accepted band-gap value of 1.9 eV. Very recently,[52,53] it has become clear that the band gap of InN is actually only ~ 0.8 eV. This does not affect the conclusions reported in Ref. 51: indeed, if the qualitative conclusions of the LDA conclusions remain valid even when the band gap is increased to 1.55 eV, then they should certainly still apply when the band gap is only ~ 0.8 eV.

Other approaches have recently emerged that go beyond density-functional theory and look promising for addressing properties of defects and impurities. The fixed-node diffusion quantum Monte Carlo method was applied to the study of silicon self-interstitials in Ref. 54. The formation energy of the split-$\langle 110\rangle$ interstitial defect was found to be significantly higher (by 1.6 eV) in the quantum Monte Carlo approach than in LDA. While this is a large number, it can probably be almost completely attributed to the upward shift of defect-induced levels in the band gap. Indeed, a shift of 0.7 eV needs to be applied to the LDA conduction band to bring it into agreement with experiment (and, presumably, the quantum Monte Carlo result). Assuming that the self-interstitial-induced t_2 level undergoes a similar shift, its occupation with two electrons would raise the energy by 1.4 eV. In spite of the lack of selfconsistency, this estimate is close to the calculated difference between LDA and quantum Monte Carlo.

Finally, we mention another promising approach that may overcome the limitations of DFT-LDA, namely, the use of an ''Exact exchange'' Kohn–Sham formalism. It has been demonstrated that this approach can produce high-quality band structures for semiconductors,[55] and in principle it lends itself to a self-consistent evaluation of total energies. Unfortunately, the computational requirements are currently prohibitive, and creative approaches to improve the computational efficiency will be essential in order to apply the method to supercell calculations.

4. Definition of formation energy

The formation energy of a defect or impurity X in charge state q is defined as

$$E^f[X^q]=E_{\text{tot}}[X^q]-E_{\text{tot}}[\text{GaN,bulk}]-\sum_i n_i\mu_i+q[E_F+E_v+\Delta V]. \tag{2}$$

$E_{\text{tot}}[X]$ is the total energy derived from a supercell calculation with one impurity or defect X in the cell, and $E_{\text{tot}}[\text{GaN,bulk}]$ is the total energy for the equivalent supercell containing only bulk GaN. n_i indicates the number of atoms of type i (host atoms or impurity atoms) that have been added to ($n_i>0$) or removed from ($n_i<0$) the supercell when the defect or impurity is created, and the μ_i are the corresponding chemical potentials of these species. Chemical potentials are discussed in detail in Sec. II E; for now, it suffices to know that these chemical potentials represent the energy of the reservoirs with which atoms are being exchanged.

E_F is the Fermi level, referenced to the valence-band maximum in the bulk. Due to the choice of this reference, we need to explicitly put in the energy of the bulk valence-band

maximum, E_v, in our expressions for formation energies of charged states. As discussed in Sec. II C, we also need to add a correction term ΔV, to align the reference potential in our defect supercell with that in the bulk.

To illustrate these concepts, let us provide a specific example. For this we choose a Mg acceptor in GaN, because it will allow us to address a number of relevant issues.

$$E^f[\mathrm{Mg}_{\mathrm{Ga}}^0] = E_{\mathrm{tot}}[\mathrm{Mg}_{\mathrm{Ga}}^0] - E_{\mathrm{tot}}[\mathrm{GaN,bulk}] - \mu_{\mathrm{Mg}} + \mu_{\mathrm{Ga}} - E_{\mathrm{corr}}, \quad (3)$$

$$E^f[\mathrm{Mg}_{\mathrm{Ga}}^-] = E_{\mathrm{tot}}[\mathrm{Mg}_{\mathrm{Ga}}^-] - E_{\mathrm{tot}}[\mathrm{GaN,bulk}] - \mu_{\mathrm{Mg}} + \mu_{\mathrm{Ga}} - [E_F + E_v + \Delta V(\mathrm{Mg}_{\mathrm{Ga}})]. \quad (4)$$

The correction term E_{corr} that appears in the formation energy of $\mathrm{Mg}_{\mathrm{Ga}}^0$ is specific to shallow centers, and is discussed in Sec. II D 3.

In principle, the *free energy* should be used in Eq. (1). Use of the (zero-temperature) formation energy as defined in Eq. (2) implies that contributions from vibrational entropy are neglected. Explicit calculations of such entropies are very demanding, and currently not feasible for the large number of defects to be addressed. These entropy contributions cancel to some extent, e.g., when solubilities are calculated; in general, they are small enough not to affect qualitative conclusions. Experimental and theoretical results for entropies of point defects show that the entropy is typically in the range between 0 and $10\,k$, where k is the Boltzmann constant. A simple estimate based on an Einstein model for the phonon frequencies gives values between 3 and $5k$ for the native defects in GaN.

Entropy effects can play an important role under certain circumstances; for instance, they have been suggested to be responsible for the stabilization of a specific configuration of the Mg–H complex, where the free energy is lowered due to the large entropy associated with a low-energy excitation.[79] In general, however, the inclusion of entropy does not cause any qualitative change in the results.

5. Supercells

The most common approach for performing calculations for impurties and defects is in a supercell geometry. The defect is surrounded by a finite number of semiconductor atoms, and that whole structure is periodically repeated.[56–59] This geometry allows the use of various techniques which require translational periodicity of the system. Provided the impurities are sufficiently well separated, properties of a single isolated impurity can be derived.

A major advantage of the supercell method is that the band structure of the host crystal is well described. Indeed, it should be clear that performing a calculation for a supercell that is simply filled with the host crystal, in the absence of any defect, simply produces the band structure of the host. This contrasts with cluster approaches, where the host is modeled by a finite number of semiconductor atoms terminated at a surface (which is typically hydrogenated, in order to eliminate surface states). Even fairly large clusters typically still produce sizeable quantum confinement effects which significantly affect the band structure, and interactions between defect wave functions and the cluster surface are hard to avoid.

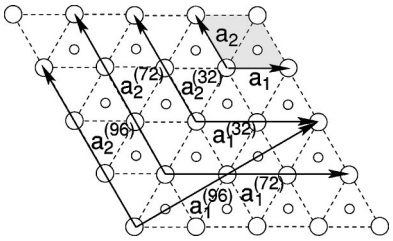

FIG. 1. Top view (along [0001] direction) of the GaN wurtzite structure: small circles represent nitrogen, large circles gallium. The shaded area corresponds to the primitive unit cell. The translation vectors for the primitive unit cell and for the 32-, 72-, and 96-atom supercells are also shown.

An alternative approach that provides a good description of the band structure of the host crystal is based on the Green's function determined for the perfect crystal. This function is then used to calculate changes induced by the presence of the defect.[60] The Green's function approach seems to be more cumbersome and less physically transparent than the supercell technique. Still, it is occasionally used, for instance in the linear muffin-tin orbital calculations of Gorczyca *et al.* for GaN, AlN, and BN.[61,62] However, their implementation of the method only allowed treatment of ideal substitutional defects, without inclusion of relaxations.

In order to study the atomic and electronic structure of an impurity in the GaN crystal, we construct an artificial unit cell (supercell) composed of several primitive GaN unit cells and containing one impurity. The larger the supercell size, the closer our results will be to the case of a single, isolated impurity, because interactions between impurities in neighboring supercells are suppressed. Convergence as a function of supercell size should always be checked. Typical supercells for the wurtzite structure contain 32, 72, or 96 atoms. The 32-atom supercell is composed of eight wurtzite GaN primitive unit cells (each containing four atoms), such that each translation vector of the supercell is doubled from that of the basic unit cell (see Fig. 1). For the 72-atom supercell, the primitive unit cell is repeated three times in each of the basal-plane directions. Both 32- and 72-atom supercells suffer from the problem that the separation between impurities in neighboring supercells is quite different when measured along different directions. The 96-atom supercell avoids this problem by having translation vectors that are mutually perpendicular, leading to a cell with orthorombic symmetry.

For the zinc-blende (ZB) structure, typical supercells contain 32 or 64 atoms. The 32-atom supercell has *bcc* symmetry, while the 64-atom cell is cubic, consisting of the conventional eight-atom cubic unit cell of the zinc-blende structure doubled in each direction.

Within the supercell, relaxation of several shells of host atoms around the impurity or defect is always included. In a 96-atom supercell, relaxing all atoms within a sphere of radius 4.8 Å around a substitutional impurity corresponds to relaxing 46 atoms (seven shells of atoms). In the zinc-blende 64-atom cell, the same relaxation radius leads to at least 44 atoms being relaxed (five shells). These relaxation radii are

typically sufficient to capture all relevant relaxations around a defect; however, some exceptions exist where longer-range relaxations are important, for instance around divacancies in silicon[63] and around gallium interstitials in GaN.[24]

Convergence tests for point defects and impurities indicate that, for zinc blende, 32-atom and 64-atom supercells yield very similar results, indicating convergence. For wurtzite, the *absolute* values of formation energies are not yet converged in a 32-atom cell. The 96-atom cell results are expected to be converged; for substitional impurities, these results (for neutral and singly charged states) are very close to the ZB 64-atom cells.[64]

6. Special k-points

Brillouin-zone integrations are carried out using the Monkhorst–Pack scheme[65] with a regularly spaced mesh of $n\times n\times n$ points in the reciprocal unit cell shifted from the origin (to avoid picking up the Γ point as one of the sampling points). Symmetry reduces this set to a set of points in the irreducible part of the Brillouin zone.

When we describe a defect in a supercell approach, defect–defect interactions between defects in neighboring supercells lead to dispersion of the defect-induced levels in the band gap. A truly isolated defect (corresponding to the limit of an infinitely large supercell) would lead to a flat, dispersionless level. The use of special points actually provides a way of averaging over the defect band that leads to a result that should be very close to the level of the isolated defect. Note that this implies that the Γ point, which is sometimes used for Brillouin-zone integrations due to the resulting numerical simplicity, provides a very poor description since there the defect-defect interaction reaches its maximum.

These arguments show that for a fully occupied defect level the use of special points will lead to a contribution to the total energy that is a good approximation to the value expected for an isolated defect. Care should be taken, however, in cases where the defect level is only partially occupied. Most computational schemes assume a ''metallic'' occupation of electronic levels, meaning that eigenvalues are filled with electrons up to a Fermi level that is calculated to yield the correct total number of electrons in the same. If the partially occupied defect level is the highest occupied electronic level, then this metallic occupation (potentially with some ''smearing'' representing a finite temperature) results in a larger fraction of electrons being placed at $\mathbf{k}$ points where the eigenvalues of the defect level are lower. This unequal occupation of the defect level then produces a poor approximation to the total energy. Indeed, since averaging over the dispersion of the defect level produces the best approximation to the position of the defect level of the isolated defect, one should make sure that states at different $\mathbf{k}$ points corresponding to the same defect level are equally occupied. Doing so has been found to yield measurable improvements in convergence as a function of supercell size.[66]

Convergence tests indicate that for zinc blende, the $2\times2\times2$ sampling yields total energies that are converged to better than 0.1 eV in both 32-atom and 64-atom supercells. For the 32-atom wurtzite supercell, a $2\times2\times2$ set does not yield fully converged results. In the 96-atom wurtzite cell, finally, the $2\times2\times2$ $\mathbf{k}$ point mesh produces converged results, i.e., increasing the $\mathbf{k}$ point sampling changes the energy only by ~0.01 eV.[24]

7. Self-consistent calculation of defect concentrations

In Sec. II A we discussed how concentrations of defects and impurities depend on formation energies, and in Sec. II B 4 we showed how these formation energies are defined. We found that the formation energies depend on atomic and electronic chemical potentials. The atomic chemical potentials reflect the experimental conditions that exist during growth or impurity incorporation, and as such are explicitly variable. However, the *electronic* chemical potential (i.e., the Fermi level), is *not* a free parameter. It is of course a quantity that we experimentally want to influence, specifically by doping of the semiconductor–and within our framework we do that by including dopant impurities. But the Fermi level E_F cannot be *directly* varied; ultimately, it is determined by the condition of charge neutrality. In principle equations such as Eq. (2) can be formulated for every point 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. The solution of this problem amounts to finding the root of a polynomial with $x=\exp(-E_F/kT)$ as the variable.[16]

The notion of calculating point defect concentrations as a function of environmental conditions is, of course, not new. Kröger developed an elaborate formalism and applied it to many solids.[67] The formalism described here differs from Kröger's approach in three major ways: (1) Instead of working with mass-action relations, which always relate to specific defect *reactions* and thus involve *pairs* of defects, we write down equations for formation energies [Eq. (2)] for each defect individually. This greatly simplifies the formalism, makes it more transparent, and still allows for obtaining a self-consistent solution for all the coupled equations, as described above. (2) Instead of working with partial pressures, we prefer to work with chemical potentials, as described in Sec. II E 1. Again, this renders the formalism more transparent and also allows us to clearly identify the effect of an abundance of certain species in the environment, even if equilibration with a gas outside the material cannot be assumed. As discussed in Ref. 31, the use of chemical potentials as variables also results in a reduction of the number of free parameters required to represent a phase diagram. (3) Last but not least, the availability of first-principles calculations to evaluate the key parameters provides us with an enormous advantage over Kröger, who had to infer the value of crucial parameters from limited experimental information. This process often involved serious assumptions, which in turn could affect the results in uncontrolled ways. The unbiased, systematic results for all potential defects provided by state-of-the-art calculations allow us to approach defect problems truly from first principles.

Rather than just showing results for defect concentrations, it is often very 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. For clarity of

presentation the atomic chemical potentials may be set equal to fixed values; a general case can always be addressed by referring back to Eq. (2). We will see that the dependence of formation energies on Fermi level provides immediate insight into the electrical activity (donor or acceptor character) of a defect or impurity, the position of its charge transfer level, and the potential behavior of certain defects as compensating centers.

C. Charge states

Most point defects and impurities can occur in multiple charge states. As shown in Eq. (2), the formation energy depends on the charge state. Formation energies have to be calculated for each relevant charge state. The stable charge state is then the one which has the lowest formation energy for a given Fermi level.

Equation (2) shows that the formation energy of charged impurities takes into account that electrons are exchanged with the Fermi level. The Fermi level E_F is referenced with respect to the valence-band maximum in the bulk, i.e., $E_F=0$ at the top of the valence band (E_v) in bulk GaN. A problem when calculating E_v is that in a supercell approach the defect or impurity strongly affects the band structure. We therefore cannot simply use E_v as calculated in the defect supercell. To solve this problem a two-step procedure is used: (i) The top of the valence band E_v is calculated in bulk GaN by performing a band-structure calculation at the Γ point and (ii) an alignment procedure is used in order to align the electrostatic potentials between the defect supercell and the bulk.

The fact that E_v found for the bulk (e.g., in a defect-free supercell) cannot be directly applied to the supercell with defect can be attributed to the long-range nature of the Coulomb potential and the periodic boundary conditions inherent in the supercell approach. The creation of the defect gives rise to a constant shift in the potential, and this shift cannot be evaluated from supercell calculations alone since no absolute reference exists for the electrostatic potential in periodic structures. The problem is similar to that of calculating heterojunction band offsets,[68] and similar techniques can be used to address these issues.[16] Our preferred method is to align the electrostatic potentials by inspecting the potential in the supercell far from the impurity and aligning it with the electrostatic potential in bulk GaN. This leads to a shift in the reference level ΔV, which needs to be added to E_v in order to obtain the correct alignment. The resulting shifts are taken into account in our expressions for formation energies in Sec. II B 4.

Another issue regarding calculations for charged states is the treatment of the $G=0$ term in the total energy of the supercell. This term would diverge for a charged system; we therefore assume the presence of a compensating uniform background (jellium) and evaluate the $G=0$ term as if the system were neutral.[59] Makov and Payne[69] have pointed out that the energy of this cell will converge very slowly as a function of supercell size, due to the electrostatic interactions between the periodic array of monopoles, which converge only as $1/\epsilon L$ (where L is the linear dimension of the supercell and ϵ is the static dielectric constant). Makov and Payne proposed to add a correction term (essentially the Madelung energy of a lattice of point charges in a dielectric environment) that would lead to a better estimate of the energy of a single isolated defect. Note that since this term scales as q^2, it can become quite sizeable for more highly charged systems.

While the Makov–Payne approach works well for atomic or molecular systems calculated in otherwise empty supercells, it has been found to lead to an overestimate of the correction term for defects in semiconductors.[66,70] An indiscriminate application of the Makov–Payne correction may not necessarily yield a better approximation of the total energy for an isolated defect. The reason has been proposed to be the improved screening that takes place around the defect within the supercell, which effectively reduces the strength of the interactions between defects in neighboring cells. The conclusions of Schwarz[66] were based on calculations for GaAs, and it remains to be seen whether they also apply to GaN (which has a smaller dielectric constant). For now, since more work is clearly needed to better understand these corrections, we refrain from applying them, and the results reported below do not include them.

D. Electronic structure

1. Thermodynamic transition levels versus optical levels

Point defects and impurities almost always introduce levels in the band gap of the semiconductor or near the band edges. The experimental detection of these levels often forms the basis for the identification of the defect or impurity. Calculation of these levels is therefore an important priority. The levels that are of experimental relevance always involve transitions between different charge states of the center. This means that the Kohn–Sham levels that result from a band-structure calculation for the center cannot directly be identified with any levels that are relevant for experiment.

The thermodynamic transition level $\epsilon(q_1/q_2)$ is defined as the Fermi-level position where charge states q_1 and q_2 have equal energy. As the name implies, this level would be observed in experiments where the final charge state can fully relax to its equilibrium configuration after the transition. This type of level is therefore what is observed in DLTS (deep-level transient spectroscopy) experiments, or, in the case of shallow centers, corresponds to the thermal ionization energy, as would be derived from an analysis of temperature-dependent Hall data.

Let us illustrate this concept for the case of a shallow acceptor, such as Mg_{Ga}, in GaN. The relevant charge states here are $q_1=0$ and $q_2=-1$, and the thermodynamic transition level $\epsilon(0/-)$ is usually called the thermal ionization energy or the acceptor ionization energy E_A. By definition

$$E^f[\mathrm{Mg}_{\mathrm{Ga}}^-](E_F=E_A)=E^f[\mathrm{Mg}_{\mathrm{Ga}}^0]. \tag{5}$$

From Eqs. (3) and (3), it then follows that

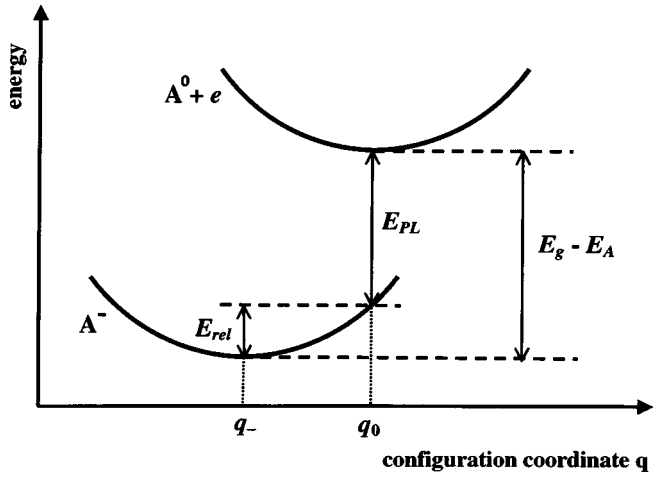

FIG. 2. Schematic configuration coordinate diagram illustrating the difference between thermal and optical ionization energies for an acceptor A. The curve for A^0 is vertically displaced from that for A^- assuming the presence of an electron in the conduction band. E_{rel} is the Franck–Condon shift, i.e., the relaxation energy that can be gained, in the negative charge state, by relaxing from configuration q_0 (equilibrium configuration for the neutral charge states) to configuration q_- (equilibrium configuration for the negative charge state). Configuration coordinate diagrams are discussed in Sec. II D 2.

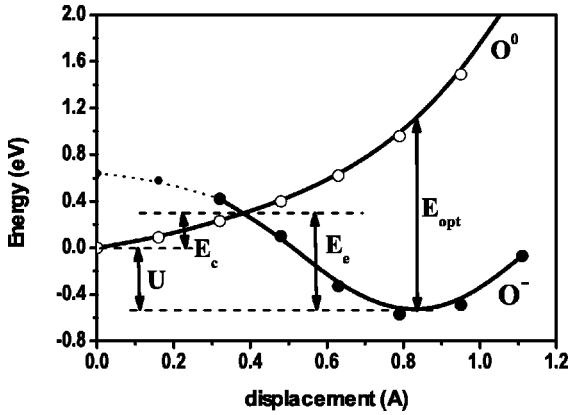

FIG. 3. Calculated configuration coordinate diagram for an oxygen DX center in wurtzite AlN, showing formation energies for an impurity in the neutral (open circles) and negative (closed circles) charge states as a function of displacement along [0001]. The Fermi level is assumed to be located at the bottom of the conduction band, and the zero of energy corresponds to the formation energy of the neutral charge state at the substitutional site. The lines are a guide to the eye. U is the energy gain due to DX center formation. E_{opt} is the optical ionization energy. E_c and E_e are capture and emission barriers for electrons. From Ref. 14.

$$E_A = E^f[\mathrm{Mg}_{\mathrm{Ga}}^-](E_F=0) - E^f[\mathrm{Mg}_{\mathrm{Ga}}^0] = E_{\mathrm{tot}}[\mathrm{Mg}_{\mathrm{Ga}}^-] - E_{\mathrm{tot}}[\mathrm{Mg}_{\mathrm{Ga}}^0] + E_{\mathrm{corr}} - E_v - \Delta V(\mathrm{Mg}_{\mathrm{Ga}}). \tag{6}$$

For purposes of defining the thermal ionization energy, it is implied that for each charge state the atomic structure is relaxed to its equilibrium configuration. The atomic positions in these equilibrium configurations are not necessarily the same for both charge states. Indeed, it is precisely this difference in relaxation that leads to the difference between thermodynamic transition levels and optical levels.

The optical level $\epsilon^{\mathrm{opt}}(q_1/q_2)$ associated with a transition between charge states q_1 and q_2 is defined similarly to the thermodynamic transition level, but now the energy of the final state q_2 is calculated using the atomic configuration of the initial state q_1. The optical level would be observed in experiments where the final charge state cannot relax to its equilibrium configuration after the transition. This type of level is therefore what is observed, for instance, in photoluminescence experiments. Indeed, it is informative to consider the following simplified picture of a photoluminescence experiment, again illustrated for the specific example of a Mg acceptor: The exciting light creates electron-hole pairs. The holes can be trapped at $\mathrm{Mg}_{\mathrm{Ga}}^-$ centers, turning them into $\mathrm{Mg}_{\mathrm{Ga}}^0$. Using our definition of the thermal ionization energy E_A, the equilibrium configuration of the $\mathrm{Mg}_{\mathrm{Ga}}^0 + e$ state (where e is an electron at the bottom of the conduction band) is $E_g - E_A$ higher than the equilibrium configuration of $\mathrm{Mg}_{\mathrm{Ga}}^-$, where E_g is the band gap.

Electrons in the conduction band can then recombine with the hole on the acceptor, as illustrated in Fig. 2. This leads to emission of a photon with energy E_{PL}. During this emission process, the atomic configuration of the acceptor remains fixed–i.e., in the final state, the acceptor is in the negative charge state, but with a structure (configuration coordinate q_0) that is the same as that for the neutral charge state. The difference between the energy of this configuration and that of the equilibrium configuration q_- is the relaxation energy E_{rel} (the Franck–Condon shift). Figure 2 shows that $E_{\mathrm{PL}} = E_g - E_A - E_{\mathrm{rel}}$. If the optical ionization energy, E_A^{opt} is defined as the energy difference between the band gap and the PL line, we find that $E_A^{\mathrm{opt}} = E_g - E_{\mathrm{PL}} = E_A + E_{\mathrm{rel}}$. This simplified picture ignores excitonic effects, etc., but it does show that the ionization energy extracted from an optical measurement should be larger than the thermal ionization energy E_A by an amount E_{rel}. If the atomic configuration in the two charge states is significantly different, E_{rel} can be sizable.

2. Configuration coordinate diagrams

Configuration coordinate diagrams can be very useful in discussing and analyzing the energetics of impurities in different charge states. In fact, we have already employed a configuration coordinate diagram in this review, in the context of our discussion of optical ionization energies in Sec. II D 1. Underlying the idea of the configuration coordinate diagram is the notion that the energy of the defect or impurity depends on its atomic configuration. In many cases, one can identify a single coordinate (or generalized coordinate) that plays the dominant role in the energetics. This could, for instance, be the magnitude of the breathing relaxation (relevant, e.g., for a nitrogen vacancy, where the surrounding Ga atoms can undergo large displacements); or the magnitude of the off-center displacement of an impurity along a specific direction (relevant, e.g., for oxygen in AlGaN or in GaN under pressure, which forms a DX center; see Fig. 3 and Sec. IV A 1).

The energy of the center, in a specific charge state, can be plotted as a function of this coordinate. In schematic figures, this dependence is often shown to be parabolic; indeed,

elastic restoring forces cause the energy to depend quadratically on displacement for small displacements. However, first-principles calculations allow us to explicitly calculate this dependence without making any additional approximations. Both our examples (Figs. 2 and 3) show that the coordinate value for which the minimum in the total energy occurs can be different in different charge states. In the case of the *DX* center, this forms the central feature of the metastability of the center.

By plotting the dependence of energy on coordinate for two different charge states, one can gain immediate insight in the various processes and their energetics that can be observed experimentally. For instance, Fig. 3 illustrates that U is the thermodynamic energy gain due to *DX* center formation; E_{opt} is the optical ionization energy (no relaxation of the final state allowed); and E_c and E_e are capture and emission barriers for electrons.

3. Deep levels versus shallow levels

In the case of a shallow acceptor the level introduced in the Kohn–Sham band structure due to the presence of the impurity is merely a perturbation of the host band structure. This ''acceptor level'' therefore exhibits essentially the same dispersion as the uppermost valence band. In the negative charge state, the acceptor level is filled–but in the neutral charge state, one electron is removed from this level. For a true, isolated acceptor (corresponding to a calculation in a very large supercell), the electron would be removed from the top of the valence band, at the Γ point. But in our finite-size supercells, the electron is actually taken out of the highest occupied Kohn–Sham level *at the special* **k** *points*, where the band energy is lower than at the Γ point. A correction is therefore needed, obtained from the energy differences between the highest occupied state at the Γ point and the special **k** points. The magnitude of this correction can be sizeable; for substitutional Be in GaN it was found to range from ~ 0.2 eV in a 96-atom cell to ~ 0.5 eV in a 32-atom cell.[64] This correction term, which we call E_{corr} and which has been defined as a positive number, needs to be taken into account in all results for neutral acceptors.[71] A similar correction is needed for shallow donors: E_{corr} is then the difference between the lowest occupied state at the Γ point and the special **k** points.

E. Chemical potentials

1. Boundaries and range

The chemical potentials depend on the experimental growth conditions, which can be Ga-rich or N-rich (or anything in between). They should therefore be explicitly regarded as variable in our formalism. However, it is possible to place firm *bounds* on the chemical potentials; these bounds will prove very useful in interpreting the results.

The Ga chemical potential, μ_{Ga}, is subject to an upper bound: under extreme Ga-rich conditions, $\mu_{\mathrm{Ga}}=\mu_{\mathrm{Ga[bulk]}}$. Indeed, in thermodynamic equilibrium the Ga chemical potential cannot be higher than the energy of bulk Ga. If we tried to push it higher, than we would no longer be growing GaN, but instead precipitating bulk Ga. This is indeed what is experimentally observed: MBE growth of GaN is often carried out under Ga-rich conditions, and precipation of Ga on the growing surface can be explicitly observed.[72,73]

Similarly, extreme N-rich conditions place an upper limit on μ_{N} given by $\mu_{\mathrm{N}}=\mu_{\mathrm{N[N_2]}}$, i.e., the energy of N in a N_2 molecule. It should be kept in mind that these chemical potentials, which are free energies, are temperature and pressure dependent.

In addition to the upper bounds defined above, we can also impose lower bounds, using the following expression:

$$\mu_{\mathrm{Ga}}+\mu_{\mathrm{N}}=E_{\mathrm{tot}}[\mathrm{GaN}], \tag{7}$$

where $E_{\mathrm{tot}}[\mathrm{GaN}]$ is the total energy of a two-atom unit of bulk GaN. The upper limit on μ_{Ga} then results in a lower limit on μ_{N}

$$\mu_{\mathrm{N}}^{\mathrm{min}}=E_{\mathrm{tot}}[\mathrm{GaN}]-\mu_{\mathrm{Ga[bulk]}}. \tag{8}$$

Similarly, the upper limit on μ_{N} results in a lower limit on μ_{Ga}

$$\mu_{\mathrm{Ga}}^{\mathrm{min}}=E_{\mathrm{tot}}[\mathrm{GaN}]-\mu_{\mathrm{N[N_2]}}. \tag{9}$$

The total energy of GaN can also be expressed as

$$E_{\mathrm{tot}}[\mathrm{GaN}]=\mu_{\mathrm{Ga[bulk]}}+\mu_{\mathrm{N[N_2]}}+\Delta H_f[\mathrm{GaN}], \tag{10}$$

where $\Delta H_f[\mathrm{GaN}]$ is the enthalpy of formation, which is negative for a stable compound. Our calculated value for $\Delta H_f[\mathrm{GaN}]$ is -1.24 eV (including the $3d$ states as valence states) or -0.48 eV (using the *nlcc*); the experimental value is -1.17 eV.[74] By combining Eq. (8) or Eq. (9) with Eq. (10) we observe that the host chemical potentials thus vary over a range corresponding to the magnitude of the enthalpy of formation of GaN.

The chemical potentials can *in principle* be related to partial pressures, using standard thermodynamic expressions. For instance, the chemical potential for hydrogen atoms in a gas of H_2 molecules is given by

$$2\mu_{\mathrm{H}}=E_{\mathrm{H_2}}+kT\left[\ln\left(\frac{pV_Q}{kT}\right)-\ln Z_{\mathrm{rot}}-\ln Z_{\mathrm{vib}}\right], \tag{11}$$

where $E_{\mathrm{H_2}}$ is the energy of an H_2 molecule, k is the Boltzmann constant, T is the temperature, and p is the pressure. $V_Q=(h^2/2\pi mkT)^{3/2}$ is the quantum volume, and Z_{rot} and Z_{vib} are the rotational and vibrational partition functions. When using such expressions for the chemical potentials one should be careful to verify that equilibrium conditions apply. For instance, when a material is annealed at high temperature under an overpressure of a certain element, it may be appropriate to relate the chemical potential of that element to the partial pressure of the gas, provided the diffusivity of the species in the solid is high enough to ensure equilibration between defects inside the solid and the gas outside. Similarly, MOCVD growth at high temperatures may be close enough to equilibrium to allow relating the chemical potentials relevant for defect and impurity incorporation to the partial pressures of the flowing gases. A clear counterexample is provided by MBE growth: there, no equilibrium can be assumed between the species in the molecular beam and the species inside the growing solid. It therefore would not make

sense to relate the chemical potentials relevant for defects and impurities to a partial pressure in the molecular beam.

2. Impurity solubility

For impurities we also need to consider the corresponding elemental chemical potential μ_X. The lower bound on μ_X is minus infinity, corresponding to the total absence of the impurity from the growth environment. An upper bound on the impurity chemical potential is given by the energy of the elemental bulk phase. However, stronger bounds usually arise due to formation of other solubility-limiting phases. For instance, when Mg is being incorporated in GaN, the Mg can interact with N and form Mg_3N_2. Equilibrium with Mg_3N_2 implies

$$3\mu_{\mathrm{Mg}}+2\mu_{\mathrm{N}}=3\mu_{\mathrm{Mg[bulk]}}+2\mu_{\mathrm{N[N_2]}}+\Delta H_f[\mathrm{Mg_3N_2}], \tag{12}$$

where $\Delta H_f[\mathrm{Mg_3N_2}]$ is the enthalpy of formation of Mg_3N_2. Equation (12) allows us to relate μ_{Mg} to μ_{N}, assuming equilibrium with Mg_3N_2. Combining this information with the expression for the formation energy of Mg_{Ga} [Eq. (3)], we find that the *lowest* formation energy (and hence the *highest* concentration of Mg_{Ga}, i.e., the solubility limit) occurs under *nitrogen-rich* conditions, i.e., when $\mu_{\mathrm{N}}=\mu_{\mathrm{N[N_2]}}$. This may seem obvious, since N-rich conditions indeed make it easier for Mg to be incorporated on a substitutional Ga site. The situation is not always this obvious, however. For instance, when incorporating Si into GaN, the maximum solubility is achieved under Ga-*rich* conditions–in spite of the fact that the substitutional Si donor also incorporates on a Ga site. The reason is that in the case of GaN:Si the solubility-limiting phase is Si_3N_4. Nitrogen-rich conditions, which would make it easier for Si to incorporate on a Ga site, promote the formation of Si_3N_4, thereby suppressing the solubility. The net effect is that Ga-rich conditions are more favorable.[75]

F. Complexes

So far we have implicitly discussed isolated impurities and point defects. It is possible, of course, for defects and impurities to agglomerate and form complexes. The simplest situation is a complex AB consisting of two constituents, A and B. The chemical reaction to form the complex is

$$A+B\Leftrightarrow AB;(E_b), \tag{13}$$

where E_b is the energy gained by this reaction. This binding energy E_b between the constituents can be defined in terms of the formation energies

$$E_b=E^f(A)+E^f(B)-E^f(AB), \tag{14}$$

where the sign has been chosen such that a positive binding energy corresponds to a stable, bound complex.

Merely having a positive binding energy does not imply that a complex will necessarily form. Consider, for instance, the situation where the incorporation of the constituents, as well as the complex, is governed by thermal equilibrium at the growth temperature. In order for the concentration of complexes to be larger than that of either constituent, its formation energy should be lower than both $E^f(A)$ and $E^f(B)$. This immediately implies that the binding energy E_b needs to be greater than the larger of $E^f(A)$ and $E^f(B)$. In other words, unless the binding energy of the complex is large on the scale of the formation energy of the constituents, its concentration will be small. The reason is that complexes generally have a much smaller configurational entropy. For example, a complex consisting of two constituents can be formed in $N_{\mathrm{sites}}N_{\mathrm{config}}$ configurations [see Eq. (1)] whereas if the two constituents are independently formed they can be created in $\propto N_{\mathrm{sites}}^2$ configurations. Thus, to discuss complex formation it is not sufficient to look at the complex binding energy but also at the configurational entropy.

Formation of complexes does not always occur under equilibrium conditions, however. A typical situation occurs when one of the constituents is incorporated and essentially ''frozen in'' during the growth process, and complex formation occurs only during the subsequent cooldown. Let us give a specific example, namely, formation of Mg–H complexes in GaN. The Mg concentration is essentially determined at the growth temperature. As we will see in Sec. IV C 2, the binding energy of the Mg–H complex is significantly smaller than a typical formation energy of the Mg acceptor. Therefore the concentration of Mg–H complexes at the growth temperature is small compared to the Mg concentration. Complexes can form during cooldown, however. A supply of hydrogen is available because hydrogen was incorporated during the growth, or because a reservoir of hydrogen is still available outside the crystal. These hydrogen atoms are highly mobile and can become bound at the substitutional Mg atoms. At sufficiently high temperatures these complexes can also dissociate, but once the temperature is lowered below a certain value any complexes that form will be stable.

Let us illustrate these issues with a specific scenario. We consider the complex AB consisting of two constituents A and B which can undergo the chemical reaction described in Eq. (13). The complex formation can be described by the following mass-action law:

$$\frac{c_A\times c_B}{c_{AB}}=\frac{N_{\mathrm{sites}}}{N_{\mathrm{config}}}\exp(-E_b/kT), \tag{15}$$

where c_A, c_B, and c_{AB} are the respective concentrations. Let us first consider the case where c_A and c_B are equilibrium concentrations according to Eq. (1). Then $c_A=N_{\mathrm{sites}}\times\exp[-E^f(A)/kT]$ and $c_B=N_{\mathrm{sites}}\exp[-E^f(B)/kT]$. Using these relations together with Eq. (15) yields

$$\begin{aligned}c_{AB}&=c_A\times c_B\frac{N_{\mathrm{config}}}{N_{\mathrm{sites}}}\exp(E_b/kT)\\&=N_{\mathrm{config}}N_{\mathrm{sites}}\\&\quad\times\exp\{-[E^f(A)+E^f(B)-E_b]/kT\}.\end{aligned} \tag{16}$$

Using Eq. (14) the above equation reproduces the concentration of AB complexes: $c_{AB}=N_{\mathrm{config}}\times N_{\mathrm{sites}}\times\exp[-E^f(AB)/kT]$.

Let us now consider the case where the number of defects A and B are *fixed* but where at least one of the constitu-

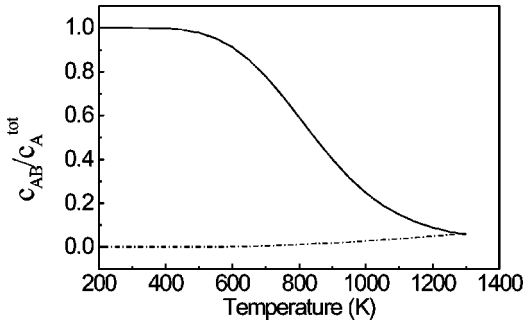

FIG. 4. Ratio of the concentration of complexes AB (Mg–H, in the current example) to the total concentration of A (hydrogen) as a function of temperature. The dash-dotted line represents thermal equilibrium conditions, while the solid line represents results in the event that the total concentrations of A (hydrogen) and B (magnesium) are determined by thermal equilibrium at the growth temperature (1300 K) and kept fixed during cooldown. The formation energies of A and B are taken to be 1 eV, and the binding energy E_b is 0.7 eV, values that apply to the case of Mg–H complexes in GaN. The concentration ratio $c_{AB}/c_A^{\mathrm{tot}}$ is small at the growth temperature but rises rapidly as the temperature is lowered for the case where the total concentrations of A and B are kept fixed.

ents is sufficiently mobile to realize local equilibrium according to Eq. (13). We can then solve Eq. (15) using the condition that $c_A^{\mathrm{tot}}=c_A+c_{AB}$ and $c_B^{\mathrm{tot}}=c_B+c_{AB}$ are kept fixed. In this case, decreasing the temperature leads to an increase in the number of complexes AB and a decrease in the number of isolated defects A and B. To be specific let us apply this to the case where the constituents are Mg^- and H^+, as already mentioned above. Charge neutrality requires that both are formed in roughly the same concentration $[c_{\mathrm{Mg}^-}\approx c_{\mathrm{H}^+}]$ (assuming that the Fermi level is far enough from the band edges to lead to negligibly small free-carrier concentrations). Their formation energies must therefore be the same, i.e., $E^f(\mathrm{Mg}^-)\approx E^f(\mathrm{H}^+)$. As we will see in Sec. IV B 1, first-principles calculations show that $E^f(\mathrm{Mg}^-)\approx E^f(\mathrm{H}^+)\approx 1$ eV.[76] The binding energy of the Mg–H complex is $E_b=0.7$ eV.[76] Let us further assume that during growth ($T\approx 1300$ K) the impurity concentration is in thermodynamic equilibrium and that after growth the sample is cooled down in a fashion that keeps the impurity concentration constant (i.e., the total numbers of Mg and H atoms are kept fixed to their values at the growth temperature). The resulting temperature dependence of the complex concentration c_{AB} (expressed with respect to the total number of atoms of type A, which is being held constant) is shown in Fig. 4. The figure shows that at the growth temperature the complex concentration is more than an order of magnitude smaller than the impurity concentration. However, as the temperature is reduced the complex concentration rises, and below about 500 K only complexes will be formed. In contrast, if the concentrations of A, B, and AB were all determined by thermal equilibrium at each temperature, the ratio of c_{AB} to c_A^{tot} would be significantly smaller. Freezing in the defect or impurity concentration thus significantly enhances complex formation, leading to concentrations of complexes significantly higher than what might be expected from equilibrium arguments.

G. Diffusion

An important question when studying defect concentrations is to specify whether the system is in thermodynamic equilibrium or whether it is governed by kinetic processes. Generally, the system will be in thermodynamic equilibrium if the defects are sufficiently mobile to follow and eliminate gradients in the corresponding chemical potential. Calculations of diffusivities of point defects in the nitrides have only recently been performed.[24]

A powerful tool to study and analyze the mobility of a defect is the calculation of total energy surfaces. Such surfaces provide direct insight into stable configurations and migration paths, and they show the location of saddle points, providing values for migration barriers. Total energy surfaces are also useful for identifying spatial locations where additional local minima (metastable configurations) might occur.

We illustrate these concepts by considering the behavior of an interstitial impurity X_i. In order to construct the corresponding total energy surface, we calculate the total energy of X_i at various locations in the lattice. For each position of X_i, the surrounding host atoms are allowed to relax, resulting in an adiabatic total energy surface. The resulting energy values as a function of the coordinates of the X_i position, $\mathbf{R}_{X_i}$, define the potential energy surface: $E=E(\mathbf{R}_{X_i})$. The energy surface is therefore a function of three spatial dimensions. In order to obtain accurate results, a database of energy values (for a sufficently large number of spatial coordinates) is needed.

Symmetry can be used in the calculation and visualization of the total energy surface.[59] We only need to calculate positions in the irreducible portion of the unit cell. The chosen locations are typically not equally spaced in the crystal, because it is better to increase the density of points near important locations such as local minima and saddle points. Since the energy surface is a function of $\mathbf{R}_{X_i}$, it possesses the full symmetry of the crystal. To make effective use of these symmetries, an analytic description of the surface is essential. This can be achieved through expansion in a basis set with the appropriate symmetry. A fitting of the total-energy results (including the atomic forces) to this set of basis functions is then performed, using a least-squares method. As basis functions, one can use symmetrized plane waves that reflect the full symmetry of the crystal (zinc blende or wurtzite). One can also make use of the physical fact that the interstitial impurity can never approach any host atom too closely (exchange processes, which would carry a very high energy cost, are not included in the total energy surface). To this end, some high-energy values near the host atoms can be added to the calculated data set; these additions does not affect the shape of the energy surface in the relevant regions away from the atoms.

Since the energy surface is a function of three spatial dimensions, it is difficult to present the results in a single plot. For visualization purposes, we need to show a cut of the energy surface, restricting the coordinates to a single plane. Judicious choice of such planes ensures we convey all the essential information, i.e., stable and metastable sites as well as barriers between them. For GaN, the (11–20) and (0001)

planes are good choices. The energy surface can then be displayed as a contour plot or as a perspective plot of the energy (along the z-axis) as a function of coordinates in the plane.

H. Direct comparison with experiment

The first-principles calculations described above can be used to predict experimental results, and also to help interpret experimental observations. It is often desirable to be able to perform a direct comparison between calculated and measured quantities. Such a comparison can be used, for instance, to identify the chemical nature or microscopic structure of a defect. Here we mention two examples that have proven particularly useful for defect studies.

1. Calculation of hyperfine parameters

The first-principles calculations provide explicit information about the wave functions in the system; it is therefore possible to calculate hyperfine parameters, which can be directly compared with the quantities measured by electron paramagnetic resonance. The wave functions have to be obtained from a calculation that explicitly includes spin polarization; i.e., spin-up and spin-down electrons have to be treated independently. It has been shown that it is very important to take contributions to the spin density from all the occupied states into account. One might assume that it would suffice to only include the wave function of the unpaired electron in the calculation of hyperfine parameters, but polarization of the valence states can strongly affect the result; for hydrogen impurities in Si, the difference was as large as a factor of two.[77]

Hyperfine parameters are particularly sensitive to the wave functions in the core region. When using a pseudopotential approach, the wave function in the core region is replaced by a smooth pseudowave function. It has been demonstrated that the information contained in the pseudowave function, in conjunction with information about wave functions in the free atom, is sufficient to calculate hyperfine parameters with a high degree of accuracy. The formalism is described in detail in Ref. 77.

The experimental observation of hyperfine signals usually provides some information about the chemical identity of the atoms in the vicinity of the defect, as well as about the symmetry. The ability to directly compare with calculated values for specific defect configurations then allows an explicit identification of the microscopic structure. Examples can be found in Ref. 77.

2. Calculation of vibrational frequencies

Defects or impurities often give rise to localized vibrational modes (LVM). Light impurities, in particular, exhibit distinct LVMs that are often well above the bulk phonon spectrum. The value of the observed frequency often provides some indication as to the chemical nature of the atoms involved in the bond, but a direct comparison with first-principles calculations can be very valuable.

Evaluating the vibrational frequency corresponding to a stretching or wagging mode of a particular bond can be accomplished by using calculated forces to construct a dynamical matrix. In the case of light impurities, where a large mass difference exists between the impurity and the surrounding atoms, it is often a good approximation to focus on the displacement of the light impurity only, keeping all other atoms fixed. A fit to the calculated energies as a function of displacement then produces a force constant. This approach actually lends itself very well to taking higher-order terms (anharmonic corrections) into account. In the case of an impurity such as hydrogen the anharmonic terms can be on the order of several $100\ \mathrm{cm}^{-1}$, so an accurate treatment is required. The formalism has been described in Ref. 78, and applications to LVMs in nitrides are detailed in Refs. 79 and 80.

3. Calculation of charge transfer levels

The calculation of charge transfer levels was discussed in Sec. II D 1. Thermodynamic transition levels can be derived from experiments such as DLTS or temperature-dependent Hall measurements, while optical levels would be observed in photoluminescence. Comparisons with experiment should be carried out judiciously. For instance, the error in the LDA band gap may obviously affect the results. In the case of shallow levels the results are best expressed by referencing them to the valence-band maximum for acceptors, and to the conduction-band minimum for electrons. For deep levels, it is also often possible to determine whether the states have predominantly valence-band or conduction-band character, as discussed in the case of InN in Ref. 51. Levels that are valence-band derived (for instance, the anion dangling bonds in the case of a cation vacancy) are likely to be only modestly affected by band-gap corrections, while levels that are conduction-band derived (for instance, the cation dangling bonds in an anion vacancy) will likely shift up with the conduction band when band-gap corrections are applied.

I. Accuracy

It is appropriate to ask what the error bar is on the calculated formation energies. One component of the error is a purely numerical error bar, associated with the choice of supercell size, plane-wave-energy cutoff, and **k**-point sampling. The magnitude of this error can be estimated simply by increasing the level of convergence. Well-converged calculations, of the type that will be reported in Secs. III and IV, have numerical error bars that are smaller than 0.1 eV. Accuracies better than 0.01 eV can be achieved in cases where it is considered important, for instance when taking energy differences between configurations of a specific defect in a single charge state (for which other types of errors discussed below would systematically cancel).

Another type of error that, in principle, could be explicitly checked by increasing the supercell size is the one caused by the electrostatic interactions between charged defects induced by the periodic boundary conditions, as discussed in Sec. II C. Unfortunately, the large cell sizes that would be required to perform these checks often pose unreasonable computational demands. This emphasizes the need for a reliable method to correct for these interactions. We estimate that for $1+$ or $1-$ charge states in 96-atom nitride

supercells this error is on the order of 0.1 eV; however, it may become more important for higher charge states.

Another type of error could potentially be associated with the use of pseudopotentials. Explicit comparisons with all-electron calculations have shown, however, that the use of pseudopotentials produces highly accurate results (see, e.g., Ref. 41).

That leaves us with one remaining source of error, namely, density-functional theory itself, as discussed in Sec. II B 3. The magnitude of this error depends on how many electrons reside in defect-related states, and whether those states are affected by the band-gap error. As discussed in Ref. 51, a qualitative picture based on the physics of the defect states can often be very illuminating. For instance, one expects the defect states associated with a cation vacancy in a III–N semiconductor to be derived from nitrogen dangling bonds, which have valence-band character and are hence unlikely to shift significantly when the band gap is corrected. Therefore, a correction of the band gap is expected to have no significant effect on the formation energy, since the energy of electrons residing in those defect states remains largely unchanged.

Many cases where accuracy is important involve comparison of energies of configurations where the number of atoms and the charge state are kept fixed. In that case the systematic errors due to charge-state effects or due to density-functional theory cancel, and the accuracy is determined by the numerical error bar, as discussed above. An important example would be the calculation of migration barriers, which involve the energy difference between the ground state and the saddle point.

Other types of calculated quantities exhibit larger error bars. For instance, transition levels (see Sec. II D 1) are energy differences between different charge states, and hence unavoidably reflect the errors due to electrostatic interactions and due to the band-gap error. A conservative error bar of at least several 0.1 eV should always be assumed.

III. NATIVE POINT DEFECTS

A. Formation energies

First-principles calculations for native point defects in GaN have been reported by several groups. The most comprehensive calculations, for all types of defects and charge states, were reported in Ref. 17. Bogusławski *et al.*[81] also performed calculations for all defects, but did not explicitly report formation energies. They seem to have focused on neutral charge states. Their results are qualitatively similar to those of Ref. 17, although some differences were evident in the interpretation of those results. To the extent that quantitative differences occur they can probably be attributed to the use of the Γ point in the Brillouin-zone integrations, and the neglect of Ga *d* states in the calculations of Ref. 81. Studies for vacancies were also reported in Ref. 82, by Mattila *et al.*,[83] by Mattila and Nieminen,[84] and neutral defects were investigated by Gorczyca *et al.*[62] The formation energies of most of the defects are quite similar in zinc-blende and wurtzite; where exceptions occur they will be explicitly flagged.

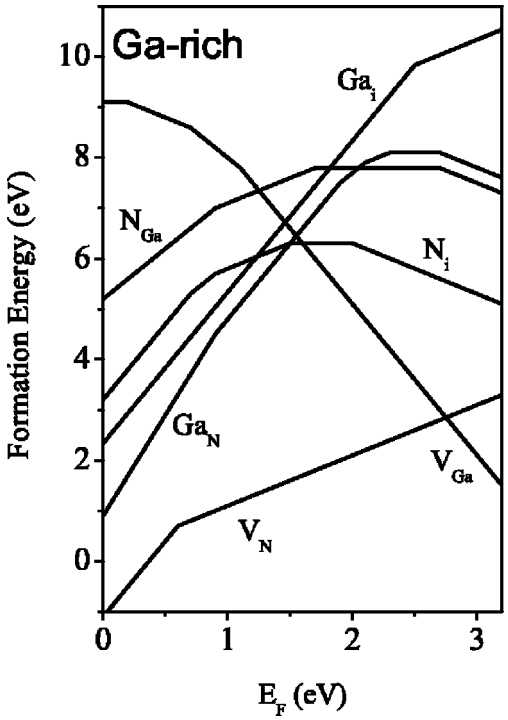

FIG. 5. Formation energies as a function of Fermi level for native point defects in GaN. Ga-rich conditions are assumed. The zero of Fermi level corresponds to the top of the valence band. Only segments corresponding to the lowest-energy charge states are shown. The slope of these segments indicates the charge state. Kinks in the curves indicate transitions between different charge states.

Formation energies for all native point defects in GaN, in all relevant charge states, are shown in Fig. 5. These results were obtained from first-principles calculations;[17,75,82] some of the values in Fig. 5 may differ from the earlier publications due to the fact that calculations for vacancies and self-interstitials were updated using 96-atom supercells.[24] For each charge state of each defect Fig. 5 displays only the line segment that gives rise to the overall lowest energy. Thus, a change in slope of the lines represents a change in the charge state of the defect. The corresponding thermodynamic transition levels, as defined in Sec. II D 1, are illustrated in Fig. 6. Antisites are not included in Fig. 6 be-

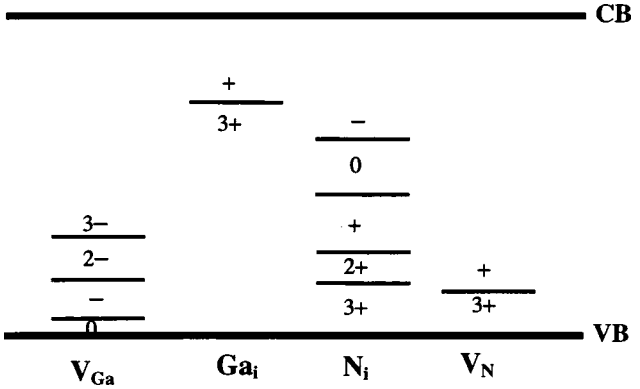

FIG. 6. Thermodynamic transition levels for defects in GaN, determined from formation energies displayed in Fig. 5.

cause we will see that they are unlikely to ever play any role in GaN. We refer to Sec. II I for a discussion of the error bars on the quantities displayed in Figs. 5 and 6.

Before discussing the defects in detail, we point out some general trends. Figure 5 shows that self-interstitial and antisite defects are high-energy defects in GaN, and are thus unlikely to occur during growth. These defects may still be created under nonequilibrium conditions, of course, for instance by electron irradiation or ion implantation. Self-interstitials (along with vacancies) may also play a role in high-temperature diffusion. Overall, however, only vacancies have low enough energies to be present in significant concentrations in GaN. It is also evident from Fig. 5 that the neutral charge states have significantly higher energies than other charge states. An investigation limited only to neutral charge states would therefore not produce a reliable picture of defect formation.

B. Nitrogen vacancies

Nitrogen vacancies (V_N) behave as shallow donors in GaN (notice the dependence on Fermi energy in Fig. 5; an increase of formation energy with E_F is indicative of donors). When purposely created, for instance during irradiation or ion implantation, nitrogen vacancies will increase the electron concentration.[85] However, their high formation energy under n-type conditions makes it very unlikely that nitrogen vacancies would form spontaneously during growth of not-intentionally doped GaN, and hence they cannot be responsible for n-type conductivity.

This conclusion contradicted the conventional wisdom in the nitride community, where for almost 25 years the n-type ''autodoping'' of GaN had commonly attributed to the nitrogen vacancy.[86,87] Of course it also raised the question of what the actual source was of the observed n-type conductivity in not-intentionally doped GaN. Based on first-principles calculations,[17,88] it was proposed that unintentional incorporation of oxygen could explain the conductivity. This possibility had previously been raised in a couple of experimental papers.[89,90] It has now become well accepted that impurities are responsible for the unintentional conductivity, and in particular the role of oxygen has been studied in great detail, as discussed below.

We note that Fig. 5 does show that nitrogen vacancies have a low formation energy in p-type GaN, making them a likely compensating center in the case of acceptor doping.

The electronic structure of the nitrogen vacancy shows a s-like a_1 state lying close to the valence-band maximum, and p-like t_2 states above the conduction-band minimum. The latter are degenerate in the case of zinc-blende GaN, but are split into a singlet and a doublet state in the wurtzite structure. In the neutral charge state one electron would be placed in the t_2 states, but because these are resonant with the conduction band this electron is transferred to the lower-lying conduction-band minimum; this cause the nitrogen vacancy to act as a shallow donor, consistent with experiment.[85]

This result qualitatively differs from previous tight-binding (TB) calculations.[32,33] The main difference is the unusually large splitting between the a_1 and t_2 defect levels which is not correctly reproduced in the TB calculations. This large splitting originates from a peculiar property of GaN: due to the much smaller size of the nitrogen atoms compared to the Ga atoms, the Ga–Ga distance in GaN is comparable to that in bulk Ga, resulting in metalliclike bonds between the Ga atoms surrounding the nitrogen vacancy.[17] This strong interaction between the Ga atoms, which are second-nearest neighbors, explains the large splitting of the a_1 and t_2 defect levels. It also explains the failing of tight-binding calculations in which mainly first-nearest neighbor interactions are taken into account.

The atomic structure of the nitrogen vacancy is characterized by a small outward breathing relaxation, by about 4% of the bond length.[24] In the wurtzite structure the four Ga neighbors can be divided into three equivalent atoms lying in a plane perpendicular to the c axis and one inequivalent Ga atom located along the c axis above the defect. The differences in relaxation among the inequivalent directions are small.

Figure 5 shows that for Fermi levels close to the VBM a charge state other than the single positive charge state is more stable. The occurrence of the $3+$ charge state is associated with a large breathing relaxation, in which the neighboring Ga atoms move outward by almost 15% of the bond length.[24] We note that the $2+$ charge state is never stable; this is characteristic of a negative-U impurity, which is usually associated with a strikingly large lattice relaxation of one of the charge states (here $3+$). The origin of the stability of the $3+$ charge state is similar to what was observed by Northrup and Zhang for the As vacancy in GaAs[91] and by Garcia and Northrup for the Se vacancy in ZnSe:[92] A large outward relaxation of the four atoms surrounding the vacancy raises the energy of the a_1 level and eventually shifts it into the band gap. This rise in the a_1 level can only be accommodated if the level is empty, i.e., in the $3+$ charge state for V_N in GaN.

The transition level $\epsilon(3+/+)$ between the $3+$ and $+$ charge states occurs at 0.59 eV above the valence-band maximum (VBM),[24] somewhat higher than the value of 0.16 eV reported in Ref. 75 or 0.39 eV reported in Ref. 93, presumably due to the use of a larger supercell which allows better relaxation of V_N^{3+}. The low formation energy of the $3+$ charge state under p-type conditions indicates that nitrogen vacancies can be a serious source of compensation in p-type GaN.

C. Gallium vacancies

The gallium vacancy (V_{Ga}) is the lowest energy defect in n-type GaN, where it acts as a triple acceptor. This defect plays a role in donor compensation as well as in the frequently observed yellow luminescence.

The electronic band structure of V_{Ga} shows levels within about 1 eV of the valence-band maximum. In the wurtzite structure these p-like t_2 states are split in a singlet and a doublet state. An outward breathing relaxation occurs: in the $3-$ charge state the three equivalent nitrogen atoms move outward by $\approx 11\%$ of the bond length, while the N atom along the c axis moves outward by $\approx 12\%$.[82]

Gallium vacancies (V_{Ga}^{3-}) have relatively low formation energies in highly doped n-type material (E_F high in the gap); they could therefore act as compensating centers. Yi and Wessels[94] have found evidence of compensation by a triply charged defect in Se-doped GaN.

The Ga vacancy has a deep level (the $2-/3-$ transition level) about 1.1 eV above the valence band.[82,84] Transitions between the conduction band (or shallow donors) and this deep level would therefore result in emission around 2.3 eV (see Sec. II.H.3). The gallium vacancy has therefore been proposed as the source of the ''yellow luminescence.''

1. Yellow luminescence

The yellow luminescence (YL) in GaN is a broad luminescence band centered around 2.2 eV. The YL appears to be a universal feature: It has been observed in bulk GaN crystallites as well as in epitaxial layers grown by different techniques. The intensity can vary over a wide range, with good samples exhibiting almost no YL.

The origins of the YL have been widely debated. Ogino and Aoki[95] proposed a model in which the YL is a transition between a shallow donor and a deep acceptor level, as illustrated in Fig. 7; a variety of experiments have confirmed this model. Proposals for the microscopic nature of the deep level have included a complex between a Ga vacancy (V_{Ga}) and a carbon atom;[95] a N_{Ga} antisite;[96] and an isolated V_{Ga}[82,97] (or a complex between V_{Ga} and oxygen[82]).

At this point in time the gallium vacancy (in isolated form or complexed with an impurity) appears to be the most likely source of the yellow luminescence; evidence is presented below.

2. Experimental confirmation

a. Complexing with donor impurities Gallium vacancies can form complexes with donor impurities in GaN.[82] The $V_{Ga}-Si_{Ga}$ complex has a rather small binding energy, due to its components being only second-nearest neighbors. The $V_{Ga}-O_N$ complex, on the other hand, has a large binding energy (1.8 eV, see Refs. 82 and 84), and can therefore enhance the concentration of Ga vacancies. The electronic structure of this complex is very similar to that of the isolated gallium vacancy, giving rise to a deep level again about 1.1 eV above the valence band. The presence of oxygen can therefore enhance the concentration of Ga vacancies and hence the YL.

An increase in V_{Ga} concentrations has indeed been observed in positron anihilation experiments on oxygen-doped samples (Ref. 98, see next section). The correlation with oxygen is also a likely explanation for the increase in YL intensity in the neighborhood of the interface with the sapphire substrate,[99] where the oxygen concentration is known to be higher.[100,101]

We also note that Reshchikov *et al.*[102] have observed a yellow luminescence band centered at 2.15 eV and a green luminescence band centered at 2.43 eV in a 200 μm thick GaN layer grown by hydride vapor phase epitaxy (HVPE).

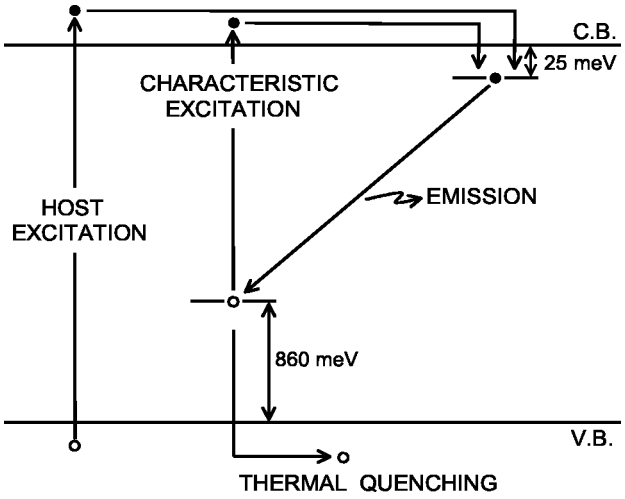

FIG. 7. Schematic illustration of the transition between a shallow donor and a deep acceptor that leads to the yellow luminescence in GaN. Ogino and Aoki[95] placed the deep acceptor level at 860 meV above the valence-band maximum. The calculated $2-/3-$ transition level of the gallium vacancy is located at 1.1 eV above the valence-band maximum.[82] From Ref. 95.

The microscopic origin of these transitions is still unclear, but complexing of V_{Ga} with different types of impurities or defects could be a potential explanation.

The formation of complexes involving Ga vacancies causes a small shift in the transition energy, contributing to a broadening of the luminescence line. Other factors contributing to the width of the line could be strain, proximity to extended defects, and Coulomb effects in the recombination of donor-acceptor pairs with varying separations.

b. Positron annihilation The most direct evidence of a correlation between Ga vacancies and YL has emerged from positron annihilation measurements by Saarinen and coworkers.[103] These experiments provide a direct probe of vacancies in the sample. It was found that the concentration of V_{Ga} correlates with the intensity of the YL, providing direct evidence for the involvement of the V_{Ga} acceptor levels in the YL. Saarinen *et al.* also observed that the concentration of gallium vacancies in MOCVD-grown GaN increased from 10^{16} to 10^{19} cm^{-3} when the V/III molar ratio increased from 1000 to 10000,[104] consistent with V_{Ga} being more favorable in N-rich material. They also found that the YL was suppressed in p-type material, consistent with easier formation of V_{Ga} under n-type conditions.[98] However, n-type doping with oxygen resulted in higher concentrations of V_{Ga} than doping with silicon, consistent with the formation of $V_{Ga}-O_N$ complexes leading to an enhancement in V_{Ga} concentrations.

c. n-type versus p-type GaN Gallium vacancies are more likely to form in n-type than in p-type GaN. This trend is consistent with experimental observations indicating suppression of the YL in p-type material.[98,105–107] Conversely, an increase in n-type doping increases the intensity of the YL.[108–110] Schubert *et al.*[109] also found that the defects giving rise to the YL act as compensating centers, in agreement with the gallium-vacancy model.

d. Ga-rich versus N-rich It is obvious that the concentration of gallium vacancies will be lower in Ga-rich mate-

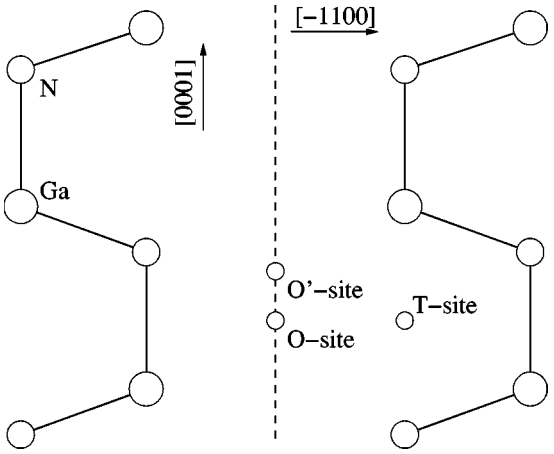

FIG. 8. Schematic representation of atomic positions in the (11-20) plane of wurtzite GaN. The large circles represent Ga atoms, medium circles N atoms. The high-symmetry interstitial sites are indicated: *O* is the octahedral interstitial site, *T* the tetrahedral interstitial site.

rial. The YL was indeed found to be suppressed in samples grown by metal-organic chemical vapor deposition (MOCVD) with higher gallium flow rates.[97,98,105] The YL was also found to be stronger in samples grown under N-stable conditions in molecular beam epitaxy (MBE), consistent with higher V_{Ga} concentrations when the growth is more N-rich.[111]

e. Recombination mechanism Various experiments have linked the YL with a deep level located about 1 eV above the valence band.[96,112,113] This is in agreement with the calculated position of the defect level induced by the Ga vacancy[82] (see Fig. 7). In addition, the calculated pressure dependence of this level is also consistent with experiment.[96]

f. Similarity with SA centers in II–VI *compounds* It is useful to point out the similarity between the YL in GaN and the so-called self-activated (SA) luminescence in *II–VI* compounds. Metal vacancies and their complexes with donor impurities are well known in II–VI compounds (e.g., ZnS, ZnSe). The metal vacancy complexes (the so-called SA centers) exhibit features which are strikingly similar to the YL: recombination between a shallow donorlike state and a deep acceptor state, and a broad luminescence band of Gaussian shape.[114,115]

D. Nitrogen interstitials

The ground state of nitrogen interstitials consists of a split-interstitial configuration in which the N_i forms a N–N bond with one of the nitrogen host atoms, sharing its lattice site.[75] This configuration is strongly energetically preferred over other interstitial positions such as the octahedral (*O*) or tetrahedral (*T*) interstitial sites, due to the large strength of the N–N bond. The nitrogen interstitial can occur in various charge states, as shown in Fig. 5. The N–N bond distance varies from ≈ 1.1 Å in the $3+$ charge (comparable to the bond distance in N_2) to ≈ 1.45 Å in the $1-$ charge state.[24,75]

In the neutral charge state N_i has two singlet states in the band gap, occupied with three electrons. Depending on the position of the Fermi level the nitrogen interstitial may act both as an acceptor or a donor. However, nitrogen interstitials have fairly high formation energies for all Fermi-level positions (Fig. 5). They are thus unlikely to occur in thermal equilibrium. However, their formation may be induced under nonequilibrium conditions, for instance under irradiation.[116,117] They could also play a role in diffusion at high temperatures.

E. Gallium interstitials

Obtaining accurate results for the atomic configuration of the gallium interstitial is difficult due to the fairly large lattice relaxations induced by this defect. In the wurtzite structure, there are two distinct types of interstitial open spaces, as shown in Fig. 8. *T* is the tetrahedral interstitial site (or cage site). This site is equidistant from four Ga and four N atoms. *O* is the octahedral interstitial site. This site is in the ''interstitial channel'' along the *c* axis. The *O* site is equidistant from six Ga and six N atoms. Both sites are obvious candidates for local minima of an interstitial defect. The calculations reported in Refs. 17 and 75 found the octahedral site to be most stable for Ga_i, while it was argued in Ref. 81 that the tetrahedral site was slightly more stable than the *O* site. Based on our recent 96-atom supercells we can now confidently state that the octahedral interstitial site, at the center of the hexagonal channel, is the stable site for Ga_i in all charge states.[24,75] The *T* site is not a local minimum, but plays a role in the diffusion process.

As shown in Fig. 5, the stable charge states are $3+$ and $1+$, meaning that Ga_i always acts as a donor (the lack of stability of the $2-$ charge state gain being characteristic of a ''negative-*U*'' defect). The formation energy of Ga_i^{3+} is lowest (but still higher than that of the nitrogen vacancy) for Fermi-level positions near the valence-band maximum (VBM), i.e., under *p*-type conditions. The $+2$ state is not thermodynamically stable, and the $+1$ charge state is stable at higher Fermi levels, where the high formation energy renders its formation unlikely under equilibrium conditions. However, Ga interstitials can be induced by non-equilibrium processes, such as in the irradiation experiments of Ref. 118.

The gallium interstitial induces two defect levels, a deep donor level and a resonance in the conduction band. In the neutral charge state the deep donor level is doubly occupied and the level in the conduction band singly occupied. Similar as for the nitrogen vacancy this electron will be donated to the bottom of the conduction band where it forms a shallow donor level.

F. Nitrogen antisites

The nitrogen antisite (N_{Ga}) has a singlet and a doublet state in the band gap. The atomic configuration reflects a strong distortion. As in the case of the nitrogen interstitial, the driving force is the tendency to form strong N–N bonds: the nitrogen atom on the Ga site moves toward a N neighbor and forms a N–N bond. In Ref. 75 the N–N bond length was found to be sensitive to the charge state: filling up the defect levels increased the bond length from 1.2 Å in the $2+$ charge state up to 1.5 Å in the $4-$ charge state. The distance to the other three N neighbors was found to be much larger (2.1 Å) indicating that no bond is formed. The nitrogen antisite is

thus characterized only by a single N–N bond. Mattila *et al.*[83] also performed a detailed study of relaxations around the nitrogen antisite in zinc-blende GaN. A metastability has been found in the neutral charge state.[62,83]

The formation energy of the nitrogen antisite, in any of its charge states, is quite high (see Fig. 5); it is thus unlikely that N_{Ga} will form in appreciable concentrations.

G. Gallium antisites

The gallium antisite Ga_N has a singlet and a doublet state in the band gap. The substitutional Ga atom forms four covalentlike bonds to the surrounding nearest neighbor Ga atoms, with a Ga–Ga bond length of ≈ 2.1 Å (in the neutral charge state),[75] much shorter than the bond length in bulk Ga (2.44 Å).[119] In spite of this significant compression the Ga–Ga bond length is still $\approx 12\%$ larger than the bulk Ga–N bond length, indicative of a large strain. Indeed, atomic relaxation lowers the formation energy by nearly 5 eV. However, the formation energy remains too high for this defect to ever occur in appreciable concentrations.

H. Complexes

We already mentioned an example of complex formation between a native defect and an impurity, namely the $V_{Ga}-O_N$ complex. It is also of interest to consider complexes between two native defects. Mattila and Nieminen[84] performed calculations for complexes between nitrogen and gallium vacancies. They found the dominant charge state of the complex to be $2-$, as expected based on the dominant charge states of the constituents: $3-$ for V_{Ga} and $+$ for V_N. The binding energy is sizeable, but the formation energy of the complex did not seem low enough, for any position of the Fermi level, for the complex to ever occur in appreciable concentrations. Put another way: since V_N is unlikely to occur under *n*-type conditions, and V_{Ga} unlikely to occur under *p*-type conditions, no conditions can be identified where both would be favorable enough for a complex to form.

Chadi[120] has proposed that a complex consisting of a nitrogen antisite and a nitrogen vacancy may be important in *p*-type GaN. Such a complex could originate starting from a gallium vacancy by moving a neigboring nitrogen atom to the vacancy site. Chadi focused on the significant lowering in energy that can be achieved by converting V_{Ga}^{3-} into a $[N_{Ga}-V_N]^{3+}$ complex: as much as 3.2 eV when the Fermi level is at the valence-band maximum.[120] However, our calculations show that, under Ga-rich conditions and for $E_F = E_v$, the sum of formation energies of N_{Ga}^{2+} and V_N^{+} still exceeds 5.3 eV. Even a sizeable binding energy and/or a shift to N-rich conditions cannot bring the formation energy of $[N_{Ga}-V_N]^{3+}$ to low enough values for this complex to occur in large enough concentrations to affect the electronic properties.

I. Comparison between GaN and GaAs

It is informative to compare the formation of native point defects in GaN with the situation in a more conventional semiconductor such as GaAs. Two main differences are evident.

First, we noted that in GaN only vacancies have low formation energies. In contrast, in GaAs self-interstitials, antisites, and vacancies all have comparable formation energies.[23,91,121] The high formation energy of antisites and interstitials in GaN can be explained in terms of the large mismatch in the covalent radii of Ga and N. When a *large atom* (Ga) is brought into the crystal (Ga_i, Ga_N) the atoms around the defect have to move away from the defect, leading to large strains. Although atomic relaxation significantly reduces the formation energy (sometimes by several electron volts), the resulting formation energy remains high. Conversely, when a *small atom* (N) is brought into the crystal (N_i, N_{Ga}) the initial bond length is too *long* to form nearest-neighbor bonds. The system seems to prefer to break the symmetry and form low-symmetry configurations: for both the N interstitial and the N antisite we find that structures with one short N–N bond are preferred. Again, however, the large displacements necessary to allow formation of this N–N bond lead to significant strains and sizeable formation energies.

Second, in GaN the defect with the overall lowest formation energy is the nitrogen vacancy, under both Ga-rich and N-rich conditions. If deviations from stoichiometry occur due to point-defect formation, they will therefore always tend toward nitrogen-deficient material (even under N-rich conditions). This was confirmed by explicit self-consistent calculations of point-defect concentrations and stoichiometries in Ref. 121. In contrast, in GaAs point-defect formation energies are more ''balanced,'' and As-rich conditions would indeed lead to As-rich material. The reason for the asymmetry in GaN can be found in the high binding energy of nitrogen molecules, which makes it difficult (or even impossible) for the GaN solid to ever become nitrogen-rich: Nitrogen atoms much prefer to leave the solid and become part of N_2 molecules, rather than incorporate in the solid in the form of nitrogen-rich point defects. This feature is absent in the case of GaAs, where As molecules exhibit only modest binding energies.

J. Native defects in AlN

Studies of native defects in AlN are of interest for two main reasons. First, AlN is being considered as a candidate substrate for epitaxial growth of nitrides; indeed, bulk crystals of AlN are somewhat easier to obtain than crystals of GaN.[122] Second, AlGaN alloys are widely used in device structures, and knowledge of native defects is important for improving the crystal quality. Studies that explicitly address point defects in alloys are rare so far; indeed, the accurate calculation of both alloy and defect properties requires large supercells, and a large number of calculations to address all relevant configurations. Bogusławski and Bernholc[123] focused on identifying trends in the formation energy of nitrogen vacancies in AlGaN alloys, finding a strong dependence on the chemical identities of the nearest neighbors. Lacking detailed information about alloys, the properties of native defects in AlGaN can, as a first approximation, be obtained by interpolating between AlN and GaN.

We will not discuss AlN in as much detail as we have GaN, but confine ourselves to mentioning the principal studies that have been performed along with their main conclusions.

First-principles calculations of formation energies of native defects in AlN were performed by Mattila and Nieminen,[84] Gorczyca *et al.*,[62] Fara *et al.*,[124] and Stampfl and Van de Walle.[44,125] The main conclusions are similar to those for GaN: Self-interstitials and antisites are high in energy in wurtzite AlN, and only vacancies have low enough formation energies to occur in high enough concentrations to affect the electronic properties. An interesting exception occurs in *zinc-blende* AlN, where the Al interstitial (a triple donor) was found to have lower energy than the nitrogen vacancy in *p*-type material.[44] The lower formation energy of Al_i in the zinc-blende phase is probably due to the fact that in the wurtzite phase the interstitial can only strongly interact with *three* N neighbors, while in zinc blende it can form bonds with *four* nitrogens. This explanation is similar to the one proposed in the case of the beryllium interstitial in wurtzite and zinc-blende GaN.[64]

The difference between zinc-blende and wurtzite structures also leads to interesting differences in the case of the nitrogen vacancy, related to the position of the higher-lying defect-induced level:[44,124] In zinc blende, this level is a resonance in the conduction band causing the vacancy to act as a shallow donor, while in wurtzite (which has a larger band gap) the level lies well below the conduction-band edge causing the vacancy to act as a deep donor.

K. Native defects in InN

Indium nitride is the least studied of the group *III*-nitrides. Bulk InN is difficult to prepare due to its low thermal stability; reliable experimental information about the properties of InN is therefore scarce. In fact, it was only recently discovered[52,53] that the band gap of InN is not 1.9 eV, as was long believed, but only ~ 0.8 eV. Not intentionally doped InN has often been found to have very high electron densities–an observation similar to GaN before better doping control of that material was achieved. The unintentional *n*-type conductivity of InN has been attributed to the nitrogen vacancy as well as to the nitrogen antisite.[126,127] Indium-containing nitride alloys are an important constituent in optoelectronic devices: for example, the active layer in short-wavelength light-emitting diodes and laser diodes usually consists of $In_xGa_{1-x}N$. Increasing the In content of the alloy, in principle, makes it possible to extend the emitting light range from UV to red. Knowledge of point defects in InN is once again the first step toward developing an understanding of defects in InGaN alloys.

Few calculations have been performed for native defects in InN. An important problem in InN is the value of the band gap, which is close to zero in DFT-LDA calculations. It should be emphasized that the closure of the gap occurs only near the Γ point, affecting only a very small portion of the Brillouin zone; at the special **k**-points used for reciprocal-space integrations the material still behaves like a semiconductor. Still, a critical examination of the DFT-LDA results is appropriate. Stampfl and Van de Walle[51] addressed this problem by using the self-interaction and relaxation-corrected (SIRC) pseudopotentials.[50] As discussed in Sec. II B 3, the study showed that the conclusions obtained with LDA are not affected by the computational approach.

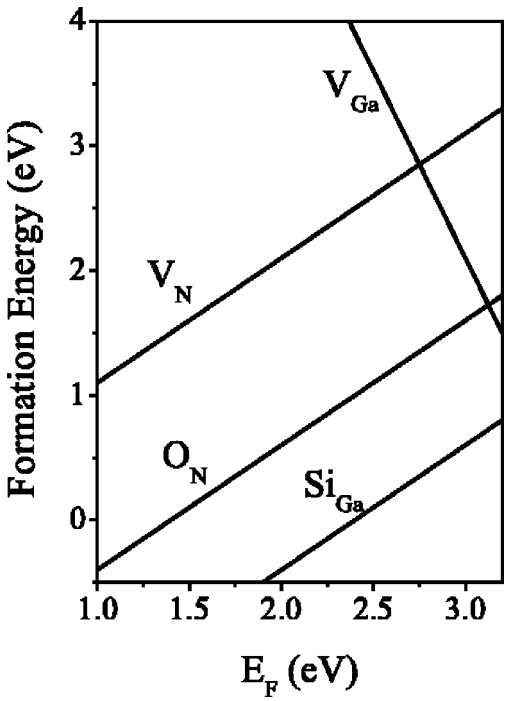

FIG. 9. Formation energy vs Fermi energy for native defects (nitrogen and gallium vacancies) and donors (oxygen and silicon) in GaN. The zero of Fermi energy is located at the top of the valence band. Gallium-rich conditions and equilibrium with Ga_2O_3 and Si_3N_4 are assumed.

The results for point defects are again qualitatively similar to those for GaN: Self-interstitials and antisites are high in energy, and vacancies are mainly important as compensating centers. The nitrogen vacancy, a shallow donor, is the lowest energy native defect in InN. In *n*-type material its formation energy is high (much higher than that of common impurities such as oxygen, silicon, or hydrogen, which all act as donors). Nitrogen vacancies thus do not account for the observed *n*-type conductivity of as-grown InN.

IV. IMPURITIES

A. Donors in GaN

Calculations for extrinsic donors have been performed for silicon, germanium, and oxygen.[84,93,128,129] While carbon could in principle behave as a donor when incorporated on the gallium site, the formation energy for this configuration is very high, and much larger than for incorporation of carbon on the nitrogen site, where it acts as an acceptor. Silicon is the most widely used intentional *n*-type dopant, while oxygen is the most likely candidate for unintentional doping.

Figure 9 summarizes first-principles results for native defects and impurities relevant for *n*-type doping (see Sec. II I for a discussion of error bars). The figure incorporates information from Refs. 82 and 24. We observe that nitrogen vacancies (V_N) have high energies in *n*-type GaN, and are thus unlikely to occur in significant concentrations. This

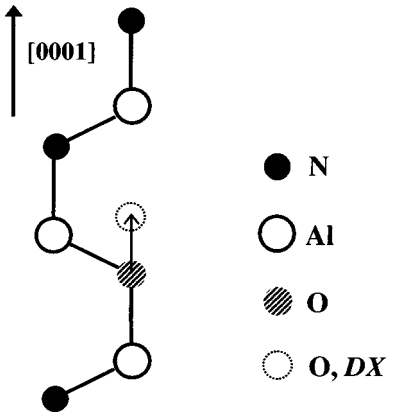

FIG. 10. Schematic illustration of the local environment around an oxygen impurity in wurtzite AlN. The dotted lines indicate the oxygen position in the *DX* configuration.

finding allowed us to conclude that nitrogen vacancies are not responsible for *n*-type conductivity in GaN. In contrast, Fig. 9 shows that oxygen and silicon have relatively low formation energies in *n*-type GaN, and can thus be readily incorporated. Both oxygen and silicon form shallow donors in GaN. The slope of the lines in Fig. 9 indicates the charge state of the defect or impurity: Si_{Ga}, O_N, and V_N all appear with slope $+1$, indicating they are single donors.

1. Oxygen

The suggestion that oxygen can be responsible for *n*-type conductivity in GaN was made by Seifert *et al.*[89] and by Chung and Gershenzon.[90] Still, the prevailing conventional wisdom, attributing the *n*-type behavior to nitrogen vacancies, proved hard to overcome. After first-principles calculations showed that nitrogen vacancies could not explain the observed *n*-type conductivity,[17] more detailed experiments were performed, which confirmed that unintentionally doped *n*-type GaN samples contained concentrations of extrinsic donors (particularly oxygen) high enough to explain the electron concentrations.

Götz *et al.*[130] reported electrical characterization of intentionally Si-doped as well as unintentionally doped samples grown by MOCVD, and concluded that the *n*-type conductivity in the latter was due to silicon. They also found evidence of another shallow donor with a slightly higher activation energy, which was attributed to oxygen. Götz *et al*. also carried out SIMS (secondary-ion mass spectroscopy) and electrical measurements on hydride vapor phase epitaxy (HVPE) material, finding levels of oxygen or silicon in agreement with the electron concentration.[100]

High levels of *n*-type conductivity have always been found in GaN bulk crystals grown at high temperature and high pressure.[131] It has been established that the characteristics of these samples (obtained from high-pressure studies) are very similar to epitaxial films which are intentionally doped with oxygen.[132,133] The *n*-type conductivity of bulk GaN can therefore be attributed to unintentional oxygen incorporation.

The high-pressure experiments have also shown that freezeout of carriers occurs at pressures exceeding 20 GPa.[131,133–135] Originally this observation was interpreted as consistent with the presence of nitrogen vacancies, since the V_N donor gives rise to a resonance in the conduction band, which emerges into the band gap under pressure. However, the observations are also entirely consistent with a "*DX*-like" behavior of the oxygen donor.

The prototype *DX* center is silicon in GaAs, which undergoes a transition from a shallow to a deep center when hydrostatic pressure is applied.[12] First-principles calculations for oxygen in GaN under pressure[14] show that at sufficiently high pressure the oxygen impurity moves off the substitutional site and assumes an off-center configuration (see Fig. 10): a large outward relaxation introduces a deep level in the band gap, and the center actually becomes negatively charged (i.e., it behaves as an acceptor). Alloying with AlN increases the band gap similar to the application of hydrostatic pressure; the behavior of impurities in AlGaN should therefore be similar to that in GaN under pressure. Indeed, first-principles calculations[14,93] show that oxygen will undergo a *DX* transition in $Al_xGa_{1-x}N$ when $x>0.3$, consistent with the observed decrease in *n*-type conductivity of unintentionally doped AlGaN.[15,136,137] A configuration coordinate diagram for oxygen in AlN is shown in Fig. 3. The conclusion is that oxygen cannot be used as a shallow donor in $Al_xGa_{1-x}N$ when $x>0.3$. Even if another donor impurity is used that does not exhibit *DX* behavior, the presence of oxygen in the layer could be detrimental to *n*-type conductivity: Indeed, once oxygen undergoes the *DX* transition it behaves as a deep acceptor, and therefore counteracts the electrical activity of other donors.

2. Silicon

Si_{Ga} is an energetically very stable configuration; the nitrogen substitutional site and the interstitial configurations are energetically unfavorable.[75] This can be understood by noting that silicon has an atomic radius very similar to gallium. Thus, while easily fitting in on a Ga site, it causes a large strain if it replaces a small N atom or goes on an interstitial site.

Several first-principles studies have addressed the issue whether silicon donors undergo a *DX* transition. Park and Chadi[93] reported that Si becomes a *DX* center in $Al_xGa_{1-x}N$ at $x>0.24$. Possible geometries for the Si *DX* center are shown in Fig. 11; Park and Chadi found the α-BB configuration to be most favorable in AlN. Bogusławski and Bernholc[129] found that Si in $Al_xGa_{1-x}N$ would undergo a shallow-deep transition at $x>0.6$. Van de Walle, finally, reported that Si would remain shallow throughout the alloy range.[14]

Experimentally it has been confirmed that Si remains a shallow donor in GaN under pressure up to 25 GPa[135] and in $Al_xGa_{1-x}N$ alloys up to $x=0.44$.[15] For higher Al content, several groups have reported that Si seems to undergo a shallow-deep transition.[138–140] However, even in the deep state the activation energy is still modest. A configuration coordinate diagram based on the experimental work of Zeisel *et al.*[139] is shown in Fig. 12. Further work will be necessary in order to establish whether silicon is a viable dopant in AlGaN with high Al content.

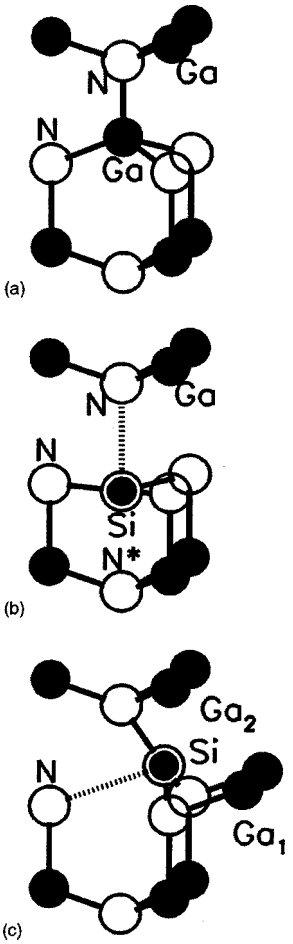

FIG. 11. Schematic illustration of the local environment around a silicon impurity in wurtzite AlN. (a) shows the ideal wurtzite lattice. (b) Illustrates the broken-bond *DX* configuration with bond breaking along the *c* axis (the so-called γ-BB center). (c) Illustrates another broken-bond configuration, with bond breaking along a different direction (the so-called α-BB center). From Ref. 93.

3. Germanium

Park and Chadi[93] found that Ge is a shallow donor in both GaN and AlN, i.e., it does not exhibit a *DX* transition. In contrast, Bogusławski and Bernholc[129] found that Ge in $Al_xGa_{1-x}N$ would undergo a *DX* transition at $x>0.3$. Experimentally, Zhang *et al.*[141] found Ge in $Al_xGa_{1-x}N$ to be a shallow donor for $x<0.2$; the behavior of Ge in AlGaN alloys with $x>0.2$ remains an open question.

B. Acceptors in GaN

1. Magnesium

Magnesium has emerged as the acceptor dopant of choice in GaN. It has been found, however, that hole concentrations obtained with Mg doping are limited.[142,143] In addition, it is well known that Mg-doped GaN grown by MOCVD needs to be subjected to post-growth treatments such as low-energy electron-beam irradiation[7] or thermal annealing[8] in order to activate the acceptors. All of these features have been addressed by first-principles calculations.

Figure 13 shows calculated formation energies for impurities and defects relevant for *p*-type GaN, for Ga-rich conditions (see Sec. II I for a discussion of error bars). The figure incorporates information from Refs. 10, 24, and 144. The Mg acceptor has a low enough formation energy to be incorporated in large concentrations in GaN. For the purposes of the plot, we have assumed Ga-rich conditions (which are actually the least favorable for incorporating Mg on Ga sites), and equilibrium with Mg_3N_2, which determines the solubility limit for Mg. We note that the formation energies for Mg^0_{Ga} and Mg^-_{Ga} intersect for a Fermi level position around 200 meV; this transition level would correspond to the ionization energy of the Mg acceptor. Since the calculated formation energies are subject to numerical error bars of ± 0.1 eV, this value should not be taken as an accurate assessment of the ionization energy. Nonetheless, it is in reasonable agreement with the experimental value of 208 meV

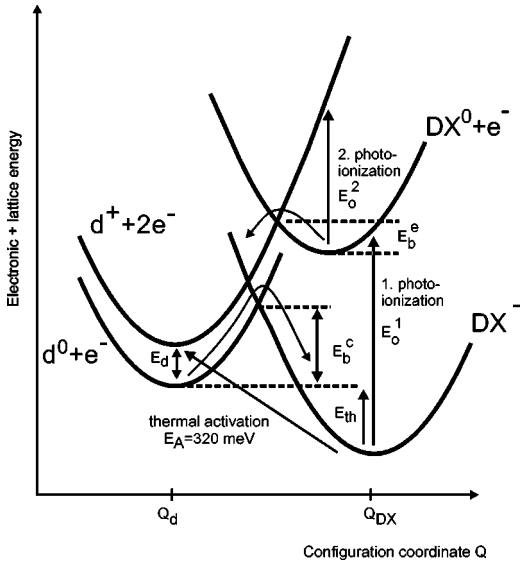

FIG. 12. Configuration coordinate diagram for a Si *DX* center in AlN, based on experimental information. From Ref. 139.

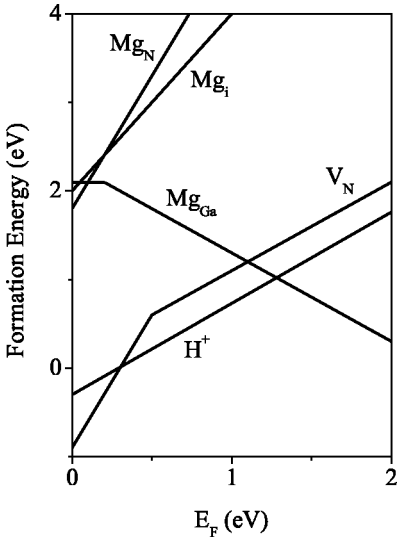

FIG. 13. Formation energy as a function of Fermi level for Mg in different configurations (Ga-substitutional, N-substitutional, and interstitial configuration). Also included are the dominant native defect (V_N) and interstitial H. Gallium-rich conditions and equilibrium with Mg_3N_2 are assumed.

TABLE I. Ionization energies of acceptors in GaN, E_A, in eV, as calculated by first-principles DFT-LDA, and by effective-mass theory (EMT). Experimental values are listed where available. Values are for wurtzite GaN, except where indicated by "(zb)."

	DFT-LDA		EMT		Experiment	
Acceptor	E_A	Reference	E_A	Reference	E_A	Reference
Li_{Ga}	0.39	11	–		–	
	0.16	146	–		–	
Be_{Ga}	0.17	64	0.209	147	0.090	148, 149
	0.06	150	0.187	151	0.150	152
					0.200	153
					0.250	154
C_N	0.26	155	0.230	147	0.230	159
	0.18 (zb)	156	0.152	151	0.215 (zb)	157
	0.65	150				
	0.2	158				
Mg_{Ga}	0.20	164	0.215	147	0.208	145
	0.23	150	0.224	151		
Ca_{Ga}	0.64	11	0.259	147		
	0.62	150	0.302	151		
Zn_{Ga}	0.23	11	0.331	147	0.328	160
	0.33	150	0.364	151		
Cd_{Ga}	0.65	150	0.625	151	0.550	161

determined by Götz *et al.*[145] Other calculated values for the ionization energy of Mg in GaN are included in Table I.

Other positions of Mg in the lattice have been investigated. "Antisite" (Mg_N) configurations and Mg on interstitial sites (Mg_i) were found to have high formation energies.[10] In addition, *AX* center configurations, in which the Mg moves off the substitutional site similar to the donor *DX* centers discussed in Sec. IV A, were also found to be unfavorable.[93] We therefore conclude that Mg overwhelmingly prefers the Ga site in GaN, the main competition being with the formation of Mg_3N_2, which is the solubility-limiting phase. Note that in this discussion we focused on doping limitations related to point defects. It has also been proposed[162,163] that the formation of extended defects, in particular pyramidal inversion domains, causes the compensation of highly Mg-doped material.

Other potential sources of compensation are also illustrated in Fig. 13. The nitrogen vacancy, which had a high formation energy in *n*-type GaN (see Fig. 9) has a significantly lower formation energy in *p*-type material, and could potentially act as a compensating center. However, we also note that hydrogen, when present, has a formation energy much lower than that of the nitrogen vacancy. In growth situations where hydrogen is present (such as MOCVD or HVPE) Mg-doped material will preferentially be compensated by hydrogen, and compensation by nitrogen vacancies will be suppressed.[164] The role of hydrogen in *p*-type doping is discussed below, in Sec. IV C 2.

Figures 5 and 13 show that V_N exhibits a $3+/+$ transition around 0.6 eV (see Sec. III B), associated with a large lattice relaxation in the $3+$ charge state. Compensation by V_N may therefore be responsible for the persistent photoconductivity effects that have been observed in Mg-doped material.[165–167] The nitrogen vacancy also may give rise to the blue lines (around 2.9 eV) commonly observed by photoluminescence in Mg-doped GaN.[93,167]

However, the microscopic origin of the blue luminescence and the source of compensation are still controversial. Lee and Chang[168] proposed a variant of the vacancy model, involving a complex of a Mg interstitial and a nitrogen vacancy, and Reboredo and Pantelides[169] proposed various substitutional-interstitial complexes that can bind hydrogen.

2. Alternative acceptors

For Mg, we concluded that achievable doping levels are mainly limited by the solubility of Mg in GaN. Several groups have investigated other candidate acceptors in GaN, and evaluated them in terms of solubility, shallow versus deep character, and potential compensation due to incorporation on other sites.[11,64,150,155,156,158,170,171] The impurities studied include Li, Na, K, Be, Zn, Ca, Cd, and C. The study of Ref. 11 showed that the incorporation of an acceptor on the substitutional site is governed by the atomic radius, which determines the energy cost of relaxation, and by the bond strength, which can be estimated using the enthalpy of formation of specific compounds. For instance, when Mg is placed on a Ga site, it is surrounded by four N atoms, much like it would be in Mg_3N_2, which has an enthalpy of formation of -4.80 eV (Ref. 172). For Be, the comparable number would be the enthalpy of formation of Be_3N_2, which is -6.11 eV. This would indicate that Be_{Ga} is more strongly bound and should have a lower formation energy than Mg_{Ga}; however, this is countered by the fact that the atomic size of Be is much smaller than that of Mg (which is more closely matched to Ga), leading to a large energy cost due to strain. The final result is that the formation energy of Be is only slightly lower than that of Ga.[11]

Calculated ionization energies are listed in Table I, and calculated formation energies of neutral acceptors are summarized in Table II. The various studies have shown that none of the investigated impurities performs better than Mg in all respects. Only Be has a comparable solubility, and it may have a lower ionization energy. However, Be may suffer from compensation due to incorporation on interstitial sites.

Only Be has emerged as a viable acceptor, exhibiting higher solubility and lower ionization energy than Mg. The possibility that silicon could incorporate on the nitrogen site and act as a shallow acceptor has sometimes been considered; however, the formation energy of Si_N is so high[129,150] (due to the size mismatch discussed in Sec. IV A 2) that its formation is highly unlikely.

The ionization energy for Be reported in Fig. 13 is 170 meV.[64] This value is slightly lower than the value previously calculated value for Mg.[164] We emphasize that the error bar on these values (which we estimate to be at least 100 meV) does not allow drawing firm conclusions about the magnitude of the ionization energy. Still, the similarity of the values for Be and Mg is in line with the expectation that the ionization energy for these shallow acceptors is largely determined by intrinsic properties of the semiconductor, such as effective masses and dielectric constants. Indeed, predictions from effective mass theory[147,151] for ionization energies of substitutional acceptors in wurtzite GaN produce values for Be between 185 and 233 meV–only slightly lower than the calculated value for Mg. Bernardini *et al.*[171] reported a

TABLE II. Formation energies of neutral substitutional acceptors in GaN, in eV, calculated by DFT-LDA.

Acceptor	E^f (Ga-rich)	E^f (N-rich)	Reference
Li_{Ga}		3.14	11
		4.5	146
Be_{Ga}		1.51	64
		2.29	150
C_N	2.62	4.4	155
	3.0	3.9	75
	2.67	3.95	156
	1.1	2.8	129
		4.24	150
	2.6	4.4	62
Mg_{Ga}		2.39	11
		1.40	150
	1.0	0.6	62
Ca_{Ga}		3.74	11
		2.15	150
Zn_{Ga}		2.86	11
		1.21	150
	2.4	0.5	62
Cd_{Ga}		1.60	150

much smaller value (60 meV) for the Be ionization energy. The most likely explanation for this discrepancy is that Bernardini *et al.* did not include the correction term E_{corr}, discussed in Sec. II D 3. This correction lowers the formation energy of the neutral charge state, and hence increases the ionization energy. Neglect of the correction would thus result in artificially low acceptor ionization energies.

While the properties of substitutional Be_{Ga} render it attractive as a shallow acceptor, the calculations have also shown that incorporation of Be on interstitial sites, where it acts as a donor, may lead to self-compensation.[11,64] In order to assess the extent of this problem, and avenues for overcoming it, the diffusivity of interstitial Be was investigated in detail in Ref. 64, using the technique of total energy surfaces outlined in Sec. II G. A large anisotropy in the diffusion was found, with a migration barrier in planes perpendicular to the *c* axis of 1.2 eV, while the barrier for motion along the *c*-axis is 2.9 eV. Complexes between interstitial Be and substitutional Be ($Be_{int}-Be_{Ga}$) were also investigated and found to have a binding energy of 1.35 eV.[64] Northrup[29] has suggested that the problem of compensation by Be interstitials may be overcome by controlling Be incorporation at the surface, in particular in the presence of indium.

Although Be doping of GaN has been reported by various groups, no conclusive results regarding its doping efficiency have been obtained. The experimental situation was discussed in Ref. 64.

Lithium has also been studied as a potential acceptor,[11,146] but the calculated ionization energy is not as low as that of Be, and Li also suffers from compensation by self-interstitials. The migration barriers were investigated in Ref. 146: the results were 1.42 eV for motion in planes perpendicular to the *c* axis, and 1.55 eV for motion along the *c*-axis. The anisotropy is thus much smaller than in the case of Be.

3. Compensation

It may seem obvious that, when acceptor doping is performed, the incorporation of unintentional donor-type impurities should be carefully controlled, since they may cause compensation. As mentioned in Sec. I C 5, such control can be a bit tricky. Consider, for instance a donor species which is known to be present as a contaminant in the growth environment, either introduced through the host-atom sources or emanating from the walls of a chamber. To a good approximation, the chemical potential of this species will therefore be roughly constant, for the specific temperature and pressure corresponding to the growth conditions. Equation (2) shows that the formation energy of the donor species then depends only on the Fermi level. As can be seen in Fig. 9, the formation energy decreases when the Fermi level moves towards the valence band. The donor species will therefore have a much higher tendency to incorporate in *p*-type material than in *n*-type.

It would therefore be dangerous to base an assessment of contaminants (for instance through SIMS studies) solely on data for *n*-type material. An unintentional donor that shows up in fairly low concentrations in *n*-type material may indeed have a much lower formation energy, and hence a higher concentration, in *p*-type material.

C. Hydrogen

Hydrogen has strong effects on the properties of GaN. Many growth techniques, such as metal-organic chemical vapor deposition (MOCVD) or hydride vapor phase epitaxy (HVPE) expose the growing material to large concentrations of hydrogen. The presence of hydrogen has particularly important consequences for *p*-type doping of the material: Hydrogen incorporated during growth leads to passivation of acceptors, and a post-growth processing step is required to render the acceptors electrically active.

A detailed overview of theoretical work on the role of hydrogen in GaN has been given in a recent review paper.[173] We therefore refrain from providing a comprehensive discussion here. Instead, we merely briefly review the most important aspects of hydrogen's behavior, and cite some more recent work that was not included in Ref. 173.

1. Isolated interstitial hydrogen

Isolated interstitial hydrogen behaves as an amphoteric impurity in GaN.[76,144,164,174–176] Figure 14 illustrates the atomic configurations,[176] while Fig. 15 shows the calculated formation energy of hydrogen in various charge states as a function of Fermi level. The formation energy is defined as the energy difference between hydrogen at an interstitial position in GaN, and hydrogen in a reservoir (in this case free H_2 molecules at $T=0$ K).[76] The values in Fig. 15 are obtained from 96-atom supercell calculations for wurtzite GaN using the *nlcc*,[80] but the values are quite close to the 32-atom zinc-blende supercell results of Ref. 76. Error bars are discussed in Sec. II I. One immediate conclusion from Fig. 15 is that the formation energy of hydrogen is lower in *p*-type GaN than in *n*-type GaN, corresponding to a much higher solubility in *p*-type than in *n*-type GaN.

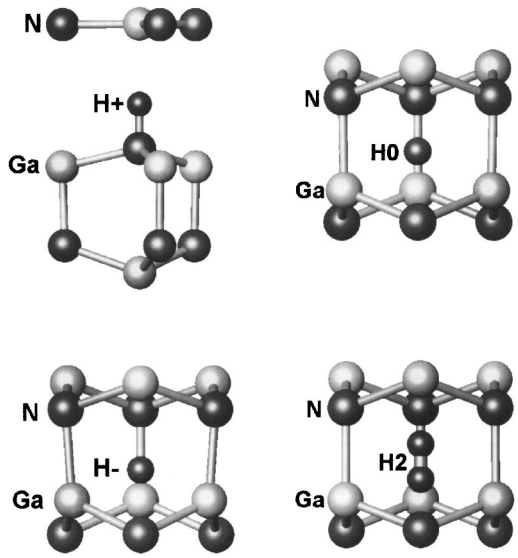

FIG. 14. Configurations of hydrogen in wurtzite GaN obtained from density-functional theory. The projection is orthographic, allowing lattice relaxations to be discerned. H^+ resides at the bond center, while H^0 and H^- are located at the center of the hexagonal channel. The H_2 molecule also resides at the center of the channel, oriented along the c axis. From Ref. 176.

Myers *et al.*[176,177] used first-principles formation energies for H in various configurations to predict solubilities in *p*-type, intrinsic, and *n*-type material, and compared the results with experimental observations. They found good agreement, provided the hydrogen formation energies were adjusted by 0.22 eV.

In *p*-type GaN, H behaves as a donor (H^+); it thus compensates acceptors. The preferred locations for H^+ are at the antibonding site behind a nitrogen atom, or at the bond-center site; in either case H is strongly bonded to the nitrogen atom. The diffusion barrier for H^+ is only 0.7 eV, which indicates a high diffusivity at moderate to high temperatures. In *n*-type GaN, H behaves as an acceptor (H^-); its most stable site is at the antibonding site behind a Ga atom. The migration barrier for H^- is high, corresponding to a very low diffusivity. For Fermi-level positions below ≈ 2.1 eV H^+ is favored; higher Fermi-level positions favor H^-. The neutral charge state is never stable, characteristic of a so-called negative-U center.

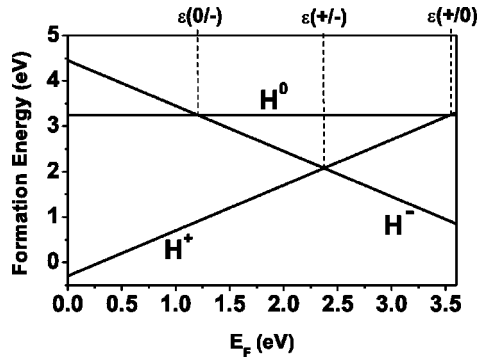

FIG. 15. Formation energies as a function of Fermi level for H^+, H^0, and H^- (solid lines), and for a H_2 molecule (dashed line) in GaN. $E_F = 0$ corresponds to the top of the valence band. The formation energy is referenced to the free H_2 molecules.

2. Acceptor-hydrogen complexes

The behavior of isolated interstitial hydrogen, as discussed in Sec. IV C 1 above, provides crucial information about interaction with impurities. Since both the solubility and the diffusivity of hydrogen in *n*-type GaN are low, hydrogen-donor complexes will rarely form, and we focus on complexes with acceptors.

In *p*-type GaN hydrogen occurs in the positive charge state and is electrostatically attracted to negatively charged acceptors. In the case of Mg acceptors, H sits in an antibonding position behind a N atom which is a neighbor of the acceptor, with a binding energy of 0.7 eV.[76,164,176] In the case of Be acceptors, the H atom favors a bond-center site in the Be-H complex, with a binding energy of 1.8 eV.[64,171]

The hydrogen atom is strongly bonded to a nitrogen atom in these acceptor-hydrogen complexes; as a consequence the vibrational frequency of the complex is representative of a N–H bond. The calculated vibrational frequency is 3360 cm^{-1} (Ref. 76) or 3284 cm^{-1} (Ref. 176) in the harmonic approximation, or 3045 cm^{-1} if anharmonic effects are taken into account.[80] These values agree well with the experimental number of 3125 cm^{-1}, measured by Fourier-transform infrared absorption spectroscopy.[178–180] Alternative structures of the Mg–H complex have been investigated by Fall *et al.*[181] and Limpijumnong *et al.*[79]

Figure 15 compares the formation energy of hydrogen in *p*-GaN with that of the nitrogen vacancy. In the absence of hydrogen, nitrogen vacancies are the dominant compensating centers, and due to charge neutrality the Fermi level will be located near the point where the formation energies of V_N^+ and Mg_{Ga}^- are equal. When hydrogen is present, however, compensation by nitrogen vacancies is suppressed. The Mg concentration is also increased, compared to the hydrogen-free case. This can be understood by inspection of the formation energies in Fig. 15: since the formation energy of hydrogen is lower than that of V_N^+, the Fermi level equilibration point is moved higher in the gap, leading to a lower formation energy and hence higher concentration of Mg. Incorporation of hydrogen is therefore beneficial in two respects: suppression of native defects, and enhancement of the acceptor concentration. Incorporation of hydrogen of course has the downside that complete compensation of the acceptors occurs, hence the need for a post-growth anneal to activate Mg-doped MOCVD-grown GaN.[8,182] The activation can also be achieved by low-energy electron beam irradiation;[7] this process has recently been investigated in more detail, experimentally as well as theoretically.[183]

Since H acts as a donor in *p*-type GaN, the improvement in *p*-type doping is actually an example of successful codoping; we return to this issue in Sec. IV D.

3. Interactions of hydrogen with native defects

Interactions of hydrogen with native point defects in GaN have been studied in Refs. 167, 184 and 185. Since

antisites and self-interstitials are unlikely to form in GaN (see Sec. III) we focus on H interacting with vacancies. This interaction is often described in terms of tying off dangling bonds. This picture does not apply in the case of the nitrogen vacancy, which is surrounded by Ga atoms at a distance of 1.95 Å from the center of the vacancy; a typical Ga–H bond distance is too large for more than one H to fit inside the vacancy. The calculated binding energy of the $(V_N H)^{2+}$ complex, expressed with respect to interstitial H in the positive charge state, is 1.56 eV.[167]

In the case of the gallium vacancy (V_{Ga}) one, two, three, or four H atoms can be accommodated in the vacancy, and levels are removed from the band gap as more hydrogens are attached.[167,184] Distinct N–H bonds are formed, with stretch frequencies between 3100 and 3200 cm^{-1}. Hydrogenated gallium vacancies with one or two H atoms behave in much the same way as the unhydrogenated kind; they may therefore contribute to compensation of donors as well as to the yellow luminescence (see Sec. III C).

4. Hydrogen in AlN and InN

The behavior of hydrogen in AlN is very similar to GaN:[144] H^+ dominates in *p*-type, H^- in *n*-type. Due to the larger band gap of AlN, the solubility of H could be significantly larger than in GaN under both *p*-type and *n*-type conditions. Calculations for H in InN, however, revealed a true surprise:[144] hydrogen in InN behaves exclusively as a donor. I.e., it is not amphoteric as in GaN and AlN, but actually contributes to the *n*-type conductivity of the material. This donor behavior is due to the fact that H^0 and H^- are always higher in energy than H^+, making H+ the only stable charge state for all positions of the Fermi level. This theoretical prediction has been confirmed by experimental studies on MBE-grown samples[186] as well as by investigations of muonium, a pseudoisotope of hydrogen.[187]

D. Codoping

The concept of ''codoping'' involves the incorporation of donors along with acceptors when *p*-type doping is attempted. Codoping has been proposed as an effective way of increasing hole concentrations in *p*-type GaN. Experimentally, codoping with oxygen has been reported to result in high hole conductivities in the case of beryllium-oxygen[188,189] or magnesium-oxygen[190,191] codoping. Yamamoto and Katayama-Yoshida[192] have proposed that complexes consisting of two acceptors and one donor (e.g., Mg–O–Mg) would be effective in enhancing the doping efficiency, and would provide an explanation for the experimental observations. Yamamoto and Katayama-Yoshida[192] performed first-principles calculations, but their arguments were based mainly on trends in the electrostatic (Madelung) energy. Several first-principles studies have now been performed that investigate these proposed complexes in more detail.

Be–O–Be complexes in GaN were investigated in Ref. 64. The formation energy of these complexes is not lower than that of the isolated acceptors (at least if equilibrium with the proper solubility-limiting phases is taken into account). Formation of acceptor-oxygen complexes (which are electrically neutral) is energetically quite favorable, but attaching a second acceptor impurity to such a complex is only marginally favored. Finally, the ionization energy of Be–O–Be was not lower than that of isolated acceptors (within the error bars of the calculations). We also note that similar acceptor-donor-acceptor complexes were investigated by Zhang *et al.* in CdTe,[193] and also found neither to increase the solubility of the acceptor nor improve the shallowness of the acceptor level. We therefore doubt that formation of Be–O–Be or Mg–O–Mg complexes is a viable approach to increasing *p*-type doping of GaN, or provides an explanation for the experimentally observed *p*-type conductivity[188–191] in codoped samples. We cannot exclude the possibility, of course, that the presence of the donor in the growth environment somehow improves the properties of the *p*-type layer in some other fashion, for instance by acting as a surfactant.

The main problem with codoping is that the donors that are incorporated alongside acceptors during the growth cannot be removed from the acceptor-doped layer after the growth. For the layer to be *p*-type, an excess of acceptors still needs to be present, hence the idea of having two acceptors for every donor. However, the proposed acceptor–donor–acceptor complexes do not seem to perform as promised. Going back to the notion of incorporating donors along with acceptors, such compensation during the growth is in fact quite desirable. Indeed, it shifts the Fermi level away from the valence-band edge toward the middle of the gap. This results in a lowering of the formation energy of acceptors (and hence in an increase of the acceptor solubility), as well as an increase in the formation energy of compensating donor-type native defects. However, the compensation by the intentionally introduced donor will persist after growth, and the material will not exhibit *p*-type conductivity.

This problem could be overcome if the donor could be removed from the *p*-type layer after growth. This is obviously not possible with oxygen. It requires that the donor impurity is not too strongly bound and exhibits a sufficiently high diffusivity, so that the donors can be removed from the vicinity of the acceptors during an anneal at modest temperatures (to avoid formation of other compensating defects). These criteria are fulfilled in the case of hydrogen, as described in Sec. IV C 2. The concept of codoping thus works successfully with hydrogen as the codopant.

V. CONCLUSIONS

In the first part of this article, we reviewed the state of the art in computational approaches for calculating defects and impurities in semiconductors from first principles. The methodology is entirely general and can be applied to any material. In the second part of the article, we focused on applications for nitride semiconductors. First-principles theory has played an important role in interpreting and guiding experiments in this rapidly developing field; in fact, in a number of areas theory has led experiment, for instance in the prediction of the behavior of hydrogen and its interactions with acceptor impurities.

Looking toward the future, we can be confident that first-principles computations will continue to play an important role in addressing defects and impurities, not only in the nitrides but also in other semiconductors, including materials such as ZnO (Ref. 194) or transparent *p*-type conductors such as $SrCu_2O_2$.[195] New developments in methodology could make the approach even more powerful. As mentioned in Sec. II B 3, the band-gap problem inherent in density-functional calculations limits the accuracy in some cases, and solving this problem should be an important goal. Other developments may render the exploration of migration paths or of potential configurations for low-symmetry configurations less cumbersome. The latter capability would make it easier to study complexes between point defects and impurities, an area that has been only superficially explored so far.

We will also see increased attention being paid to interactions between point defects or impurities and extended defects. Finally, as mentioned in Sec. II A 3, interactions between point defects or impurities need to be explored in greater detail. Just like point defects in the bulk play an important role in diffusion, point defects at surfaces are likely to affect atomic mobilities at surfaces, and hence play a decisive role in growth. Likewise, a full understanding of impurity incorporation requires comprehensive calculations of the behavior of impurities at and near the surface. Such first-principles calculations can then form the foundation for realistic simulations of the actual growth process.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research, Contract No. N00014-02-C-0433, by the Air Force Office of Scientific Research, Contract No. F4920-00-C-0019, and by the Deutsche Forschungsgemeinschaft. The authors benefited from collaborations and discussions with W. Götz, N. M. Johnson, M. Kneissl, S. Limpijumnong, M. D. McCluskey, J. E. Northrup, L. Romano, and C. Stampfl. C.VdW. is grateful to the Fritz-Haber-Institut, Berlin, for its hospitality, and to the Alexander von Humboldt Foundation for a *US Senior Scientist Award*.

[1] W. E. Pickett, Comput. Phys. Rep. **9**, 115 (1989).
[2] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).
[3] M. Bockstedte, A. Kley, J. Neugebauer, and M. Scheffler, Comput. Phys. Commun. **107**, 187 (1997).
[4] I. Vurgaftman and J. R. Meyer, J. Appl. Phys. **94**, 3675 (2003).
[5] A. G. Bhuiyan, A. Hashimoto, and A. Yamamoto, J. Appl. Phys. **94**, 2779 (2003).
[6] S. K. Estreicher and D. E. Boucher, in *GaN and Related Materials*, edited by S. J. Pearton (Gordon and Breach, New York, 1997), pp. 171–199.
[7] H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, Jpn. J. Appl. Phys., Part 2 **28**, L2112 (1989).
[8] S. Nakamura, N. Iwasa, M. Senoh, and T. Mukai, Jpn. J. Appl. Phys., Part 1 **31**, 1258 (1992).
[9] C. G. Van de Walle, D. B. Laks, G. F. Neumark, and S. T. Pantelides, Phys. Rev. B **47**, 9425 (1993).
[10] J. Neugebauer and C. G. Van de Walle, Mater. Res. Soc. Symp. Proc. **395,** 645 (1996).
[11] J. Neugebauer and Chris G. Van de Walle, J. Appl. Phys. **85**, 3003 (1999).
[12] P. Mooney, in *Deep Centers in Semiconductors*, edited by S. T. Pantelides (Gordon and Breach, New York, 1992), p. 643.
[13] D. J. Chadi and K. J. Chang, Phys. Rev. Lett. **61**, 873 (1988).
[14] C. G. Van de Walle, Phys. Rev. B **57**, R2033 (1998).
[15] M. D. McCluskey, N. M. Johnson, Chris G. Van de Walle, D. P. Bour, M. Kneissl, and W. Walukiewicz, Phys. Rev. Lett. **80**, 4008 (1998).
[16] D. B. Laks, C. G. Van de Walle, G. F. Neumark, P. E. Blöchl, and S. T. Pantelides, Phys. Rev. B **45**, 10 965 (1992).
[17] J. Neugebauer and Chris G. Van de Walle, Phys. Rev. B **50**, 8067 (1994).
[18] G. A. Baraff and M. Schlüter, Phys. Rev. B **28**, 2296 (1983).
[19] Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. Lett. **52**, 1129 (1984).
[20] R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. **54**, 360 (1985).
[21] S. B. Zhang and J. E. Northrup, Phys. Rev. Lett. **67**, 2339 (1991).
[22] D. B. Laks, C. G. Van de Walle, G. F. Neumark, and S. T. Pantelides, Phys. Rev. Lett. **66**, 648 (1991).
[23] J. E. Northrup and S. B. Zhang, Phys. Rev. B **47**, 6791 (1993).
[24] S. Limpijumnong and C. G. Van de Walle, Phys. Rev. B **69**, 035207 (2004).
[25] J. Tersoff, Phys. Rev. Lett. **74**, 5080 (1995).
[26] G. Schwarz, A. Kley, J. Neugebauer, and M. Scheffler, Phys. Rev. B **58**, 1392 (1998).
[27] C. Bungaro, K. Rapcewicz, and J. Bernholc, Phys. Rev. B **59**, 9771 (1999).
[28] T. Zywietz, J. Neugebauer, and M. Scheffler, Appl. Phys. Lett. **74**, 1695 (1999).
[29] J. E. Northrup, Appl. Phys. Lett. **78**, 2855 (2001).
[30] A. L. Rosa, J. Neugebauer, J. E. Northrup, C.-D. Lee, and R. M. Feenstra, Appl. Phys. Lett. **80**, 2008 (2002).
[31] C. G. Van de Walle and J. Neugebauer, Phys. Rev. Lett. **88**, 066 103 (2002).
[32] D. W. Jenkins and J. D. Dow, Phys. Rev. B **39**, 3317 (1989).
[33] D. W. Jenkins, J. D. Dow, and M. H. Tsai, J. Appl. Phys. **72**, 4130 (1992).
[34] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).
[35] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).
[36] V. Fiorentini, M. Methfessel, and M. Scheffler, Phys. Rev. B **47**, 13353 (1993).
[37] K. Karch, F. Bechstedt, and T. Pletl, Phys. Rev. B **56**, 3560 (1997).
[38] J. Neugebauer and Chris G. Van de Walle, Mater. Res. Soc. Symp. Proc. **339,** 687 (1994).
[39] S. G. Louie, S. Froyen, and M. L. Cohen, Phys. Rev. B **26**, 1738 (1982).
[40] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[41] M. Fuchs, J. L. F. Da Silva, C. Stampfl, J. Neugebauer, and M. Scheffler, Phys. Rev. B **65**, 245 212 (2002).
[42] T. Kotani and M. van Schilfgaarde, Solid State Commun. **121**, 461 (2002).
[43] C. Stampfl and C. G. Van de Walle, Phys. Rev. B **59**, 5521 (1999).
[44] C. Stampfl and C. G. Van de Walle, Phys. Rev. B **65**, 155212 (2002).
[45] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[46] L. J. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).
[47] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[48] A. Rubio, J. L. Corkill, M. L. Cohen, E. L. Shirley, and S. G. Louie, Phys. Rev. B **48**, 11 810 (1993).
[49] M. Palummo, L. Reining, R. W. Godby, C. M. Bertoni, and N. Börnsen, Europhys. Lett. **26**, 607 (1994).
[50] D. Vogel, P. Krüger, and J. Pollmann, Phys. Rev. B **55**, 12836 (1997).
[51] C. Stampfl, C. G. Van de Walle, D. Vogel, P. Krüger, and J. Pollmann, Phys. Rev. B **61**, R7846 (2000).
[52] V. Yu. Davydov, A. A. Klochikhin, R. P. Seisyan, V. V. Emtsev, S. V. Ivanov, F. Bechstedt, J. Furthmller, H. Harima, A. V. Mudryi, J. Aderhold, O. Semchinova, and J. Graul, Phys. Status Solidi B **229**, R1 (2002).
[53] J. Wu, W. Walukiewicz, K. M. Yu, J. W. Ager III, E. E. Haller, H. Lu, W. J. Schaff, Y. Saito, and Y. Nanishi, Appl. Phys. Lett. **80**, 3967 (2002).
[54] W.-K. Leung, R. J. Needs, G. Rajagopal, S. Itoh, and S. Ihara, Phys. Rev. Lett. **83**, 2351 (1999).
[55] M. Städele, M Moukara, J. A. Majewski, P. Vogl, and A. Görling, Phys. Rev. B **59**, 10 031 (1999).
[56] R. P. Messmer and G. D. Watkins, in *Radiation Damage and Defects in Semiconductors* (Institute of Physics and Physical Society, London, 1972), No. **16**, p. 255.
[57] S. G. Louie, M. Schluter, and J. R. Chelikowsky, Phys. Rev. B **13**, 1654 (1976).
[58] W. E. Pickett, M. L. Cohen, and C. Kittel, Phys. Rev. B **20**, 5050 (1979).
[59] C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B **39**, 10 791 (1989).
[60] R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. **52**, 1814 (1984).

[61] I. Gorczyca, A. Svane, and N. E. Christensen, Solid State Commun. **101**, 747 (1997).
[62] I. Gorczyca, A. Svane, and N. E. Christensen, Phys. Rev. B **60**, 8147 (1999).
[63] S. Öğüt and J. R. Chelikowsky, Phys. Rev. Lett. **83**, 3852 (1999).
[64] C. G. Van de Walle, S. Limpijumnong, and J. Neugebauer, Phys. Rev. B **63**, 245205 (2001).
[65] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[66] G. Schwarz, Ph.D. Thesis, Technical University, Berlin, 2002.
[67] F. A. Kröger, *The Chemistry of Imperfect Crystals* (North Holland, Amsterdam, 1964).
[68] A. Franciosi and C. G. Van de Walle, Surf. Sci. Rep. **25**, 1–140 (1996).
[69] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).
[70] G. Schwarz and J. Neugebauer (unpublished).
[71] A similar correction was also taken into account in Ref. 11; however, there the correction was applied to the *negative* charge state, rather than the *neutral* charge state. This accounts for some of the differences between the results presented in Ref. 11 and in the present review.
[72] E. J. Tarsa, B. Heying, X. H. Wu, P. Fini, S. P. DenBaars, and J. S. Speck, J. Appl. Phys. **82**, 5472 (1997).
[73] O. Brandt, R. Muralidharan, P. Waltereit, A. Thamm, A. Trampert, H. von Kiedrowski, and K. H. Ploog, Appl. Phys. Lett. **75**, 4019 (1999).
[74] *CRC Handbook of Chemistry and Physics*, 73rd Ed., edited by David R. Lide (CRC Press, Boca Raton, 1992), p. 5–18.
[75] J. Neugebauer and C. G. Van de Walle, in *Festkörperprobleme/Advances in Solid State Physics*, Vol. **35**, edited by R. Helbig (Vieweg, Braunschweig/Wiesbaden, 1996), p. 25.
[76] J. Neugebauer and C. G. Van de Walle, Phys. Rev. Lett. **75**, 4452 (1995).
[77] C. G. Van de Walle and P. E. Blöchl, Phys. Rev. B **47**, 4244 (1993).
[78] C. G. Van de Walle, Phys. Rev. Lett. **80**, 2177 (1998).
[79] S. Limpijumnong, J. E. Northrup, and C. G. Van de Walle, Phys. Rev. Lett. **87**, 205 505 (2001).
[80] S. Limpijumnong, J. E. Northrup, and C. G. Van de Walle, Phys. Rev. B **68**, 075206 (2003).
[81] P. Bogusławski, E. L. Briggs, and J. Bernholc, Phys. Rev. B **51**, 17 255 (1995).
[82] J. Neugebauer and C. G. Van de Walle, Appl. Phys. Lett. **69**, 503 (1996).
[83] T. Mattila, A. P. Seitsonen, and R. M. Nieminen, Phys. Rev. B **54**, 1474 (1996).
[84] T. Mattila and R. M. Nieminen, Phys. Rev. B **55**, 9571 (1997).
[85] D. C. Look, D. C. Reynolds, J. W. Hemsky, J. R. Sizelove, R. L. Jones, and R. J. Molnar, Phys. Rev. Lett. **79**, 2273 (1997).
[86] H. P. Maruska and J. J. Tietjen, Appl. Phys. Lett. **15**, 327 (1969).
[87] M. Ilegems and H. C. Montgomery, J. Phys. Chem. Solids **34**, 885 (1973).
[88] J. Neugebauer and C. G. Van de Walle, in *Proceedings of the 22nd International Conference on the Physics of Semiconductors*, edited by D. J. Lockwood (World Scientific, Singapore, 1995), p. 2327.
[89] W. Seifert, R. Franzheld, E. Butter, H. Sobotta, and V. Riede, Cryst. Res. Technol. **18**, 383 (1983).
[90] B.-C. Chung and M. Gershenzon, J. Appl. Phys. **72**, 651 (1992).
[91] J. E. Northrup and S. B. Zhang, Phys. Rev. B **50**, 4962 (1994).
[92] A. Garcia and J. E. Northrup, Phys. Rev. Lett. **74**, 1131 (1995).
[93] C. H. Park and D. J. Chadi, Phys. Rev. B **55**, 12995 (1997).
[94] G.-C. Yi and B. W. Wessels, Appl. Phys. Lett. **69**, 3028 (1996).
[95] T. Ogino and M. Aoki, Jpn. J. Appl. Phys. **19**, 2395 (1980).
[96] T. Suski, P. Perlin, H. Teisseyre, M. Leszczyński, I. Grzegory, J. Jun, M. Boćkowski, and S. Porowski, Appl. Phys. Lett. **67**, 2188 (1995).
[97] X. Zhang, P. Kung, A. Saxler, D. Walker, T. Wang, and M. Razeghi, Acta Phys. Pol. A **88**, 601 (1995).
[98] J. Oila, V. Ranki, J. Kivioja, K. Saarinen, P. Hautojärvi, J. Likonen, J. M. Baranowski, K. Pakula, T. Suski, M. Leszczynski, and I. Grzegory, Phys. Rev. B **63**, 045205 (2001).
[99] D. M. Hofmann, D. Kovalev, G. Steude, B. K. Meyer, A. Hoffmann, L. Eckey, R. Heitz, T. Detchprom, H. Amano, and I. Akasaki, Phys. Rev. B **52**, 16 702 (1995).
[100] W. Götz, J. Walker, L. T. Romano, and N. M. Johnson, Mater. Res. Soc. Symp. Proc. **449,** 525 (1997).
[101] G. Popovici, W. Kim, A. Botchkarev, H. Tang, H. Morkoç, and J. Solomon, Appl. Phys. Lett. **71**, 3385 (1997).
[102] M. A. Reshchikov, H. Morkoç, S. S. Park, and K. Y. Lee, Appl. Phys. Lett. **78**, 3041 (2001).
[103] K. Saarinen, T. Laine, S. Kuisma, J. Nissilä, P. Hautojärvi, L. Dobrzynski, J. M. Baranowski, K. Pakula, R. Stepniewski, M. Wojdak, A. Wysmolek, T. Suski, M. Leszczynski, I. Grzegory, and S. Porowski, Phys. Rev. Lett. **79**, 3030 (1997).
[104] K. Saarinen, P. Seppälä, J. Oila, P. Hautojärvi, C. Corbel, O. Briot, and R. L. Aulombard, Appl. Phys. Lett. **73**, 3253 (1998).
[105] X. Zhang, P. Kung, D. Walker, A. Saxler, and M. Razeghi, in *Gallium Nitride and Related Materials*, edited by R. D. Dupuis, F. A. Ponce, J. A. Edmond, and S. Nakamura (MRS Symposia Proceedings, Pittsburgh, 1996), Vol. 395, p. 625.
[106] W. Götz, N. Johnson, J. Walker, D. P. Bour, H. Amano, and I. Akasaki, in *Proceedings of the 6th International Conference on SiC and Related Materials*, Kyoto, Japan, Sept. 18–21, 1995, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima, Inst. Phys. Conf. Ser. No 142 (IOP Publishing, Bristol, 1996), p. 1031.
[107] W. Kim, A. Salvador, A. E. Botchkarev, O. Aktas, S. N. Mohammad, and H. Morkoç, Appl. Phys. Lett. **69**, 559 (1996).
[108] N. Kaneda, T. Detchprohm, K. Hiramatsu, and N. Sawaki, Jpn. J. Appl. Phys., Part 2 **35**, L468 (1996).
[109] E. F. Schubert, I. D. Goepfert, and J. M. Redwing, Appl. Phys. Lett. **71**, 3224 (1997).
[110] I.-H. Lee, I.-H. Choi, C. R. Lee, and S. K. Noh, Appl. Phys. Lett. **71**, 1359 (1997).
[111] E. J. Tarsa, B. Heying, X. H. Wu, P. Fini, S. P. DenBaars, and J. S. Speck, J. Appl. Phys. **82**, 5472 (1997).
[112] F. J. Sánchez, D. Basak, M. A. Sánchez-García, E. Calleja, E. Muñoz, I. Izpura, F. Calle, J. M. G. Tijero, B. Beaumont, P. Lorenzini, P. Gibart, T. S. Cheng, C. T. Fozon, and J. W. Orton, MRS Internet J. Nitride Semicond. Res. **1**, 7 (1996).
[113] E. Calleja, F. J. Sánchez, D. Basak, M. A. Sánchez-García, E. Muñoz, I. Izpura, F. Calle, J. M. G. Tijero, B. Beaumont, P. Lorenzini, and P. Gibart, Phys. Rev. B **55**, 4689 (1997).
[114] *Point Defects in Crystals*, edited by R. K. Watts (Wiley, New York, 1977), p. 248ff.
[115] P. J. Dean, Phys. Status Solidi A **81**, 625 (1984).
[116] Q. Zhou, M. O. Manasreh, M. Pophristic, S. Guo, and I. T. Ferguson, Appl. Phys. Lett. **79**, 2901 (2001).
[117] Q. Zhou and M. O. Manasreh, Appl. Phys. Lett. **80**, 2072 (2002).
[118] K. H. Chow, G. D. Watkins, A. Usui, and M. Mizuta, Phys. Rev. Lett. **85**, 2761 (2000).
[119] R. W. G. Wyckoff, *Crystal Structures* (Interscience Publishers, New York, 1963), Vol. 1.
[120] D. J. Chadi, Appl. Phys. Lett. **71**, 2970 (1997).
[121] C. G. Van de Walle, J. E. Northrup, and J. Neugebauer, in *Proceedings of the 4th Symposium on Non-Stoichiometric III-V Compounds*, Asilomar, CA, October 2–4, 2002, edited by P. Specht, T. R. Weatherford, P. Kiesel, T. Marek, and S. Malzer, (Friedrich-Alexander-Universität, Erlangen, Germany, 2002), p. 11.
[122] J. C. Rojo, L. J. Schowalter, R. Gaska, M. Shur, M. A. Khan, J. Yang, and D. D. Koleske, J. Cryst. Growth **240**, 508 (2002).
[123] P. Bogusławski and J. Bernholc, Phys. Rev. B **59**, 1567 (19959).
[124] A. Fara, F. Bernardini, and V. Fiorentini, J. Appl. Phys. **85**, 2001 (1999).
[125] C. Stampfl and C. G. Van de Walle, Appl. Phys. Lett. **72**, 459 (1998).
[126] T. L. Tansley and C. P. Foley, J. Appl. Phys. **60**, 2092 (1986).
[127] M. Sato, Jpn. J. Appl. Phys., Part 2 **36**, L658 (1997).
[128] T. Mattila and R. M. Nieminen, Phys. Rev. B **54**, 16 676 (1996).
[129] P. Bogusławski and J. Bernholc, Phys. Rev. B **56**, 9496 (1997).
[130] W. Götz, N. M. Johnson, C. Chen, H. Liu, C. Kuo, and W. Imler, Appl. Phys. Lett. **68**, 3114 (1996)
[131] P. Perlin, T. Suski, H. Teisseyre, M. Leszczyński, I. Grzegory, J. Jun, S. Porowski, P. Boguslawski, J. Bernholc, J. C. Chervin, A. Polian, and T. D. Moustakas, Phys. Rev. Lett. **75**, 296 (1995).
[132] C. Wetzel, T. Suski, J. W. Ager III, W. Walukiewicz, S. Fisher, B. K. Meyer, I. Grzegory, and S. Porowski, *Proceedings ICPS-23* (World Scientific, Singapore, 1996), p. 2929.
[133] P. Perlin, T. Suski, A. Polian, J. C. Chervin, W. Knap, J. Camassel, I. Grzegory, S. Porowski, and J. W. Erickson, Mater. Res. Soc. Symp. Proc. **449,** 519 (1997).
[134] C. Wetzel, W. Walukiewicz, E. E. Haller, J. W. Ager III, I. Grzegory, S. Porowski, and T. Suski, Phys. Rev. B **53**, 1322 (1996).
[135] C. Wetzel, T. Suski, J. W. Ager III, E. R. Weber, E. E. Haller, S. Fischer B. K. Meyer, R. J. Molnar, and P. Perlin, Phys. Rev. Lett. **78**, 3923 (1997).
[136] Y. Koide, H. Itoh, N. Sawaki, I. Akasaki, and M. Hashimoto, J. Electrochem. Soc. **133**, 1956 (1986).

[137] H. G. Lee, M. Gershenzon, and B. L. Goldenberg, J. Electron. Mater. **20**, 621 (1991).
[138] C. Skierbiszewski, T. Suski, M. Leszczynski, M. Shin, M. Skowronski, M. D. Bremser, and R. F. Davis, Appl. Phys. Lett. **74**, 3833 (1999).
[139] R. Zeisel, M. W. Bayerl, S. T. B. Goennenwein, R. Dimitrov, O. Ambacher, M. S. Brandt, and M. Stutzmann, Phys. Rev. B **61**, 16 283 (2000).
[140] Y. Taniyasu, M. Kasu, and N. Kobayahsi, Appl. Phys. Lett. **81**, 1255 (2002).
[141] X. Zhang, P. Kung, A. Saxler, D. Walker, T. C. Wang, and M. Razeghi, Appl. Phys. Lett. **67**, 1745 (1995).
[142] D. P. Bour, H. F. Chung, W. Gtz, L. Romano, B. S. Krusor, D. Hofstetter, S. Rudaz, C. P. Kuo, F. A Ponce, N. M. Johnson, M. G. Craford, and R. D. Bringans, Mater. Res. Soc. Symp. Proc. **449,** 509 (1997).
[143] L. T. Romano, M. Kneissl, J. E. Northrup, C. G. Van de Walle, and D. W. Treat, Appl. Phys. Lett. **79**, 2734 (2001).
[144] S. Limpijumnong and Chris G. Van de Walle, Phys. Status Solidi B **228**, 303 (2001).
[145] W. Götz, R. S. Kern, C. H. Chen, H. Liu, D. A. Steigerwald, and R. M. Fletcher, Mater. Sci. Eng. B **59**, 211 (1999).
[146] F. Bernardini and V. Fiorentini, Phys. Rev. B **61**, 12 598 (2000).
[147] F. Mireles and S. E. Ulloa, Phys. Rev. B **58**, 3879 (1998).
[148] F. J. Sánchez, F. Calle, M. A. Sánchez-García, E. Calleja, E. Muñoz, C. H. Molloy, D. J. Somerford, J. J. Serrano, and J. M. Blanco, Semicond. Sci. Technol. **13**, 1130 (1998).
[149] D. J. Dewsnip, A. V. Andrianov, I. Harrison, J. W. Orton, D. E. Lacklison, G. B. Ren, S. E. Hooper, T. S. Cheng, and C. T. Foxon, Semicond. Sci. Technol. **13**, 500 (1998).
[150] V. Fiorentini, F. Bernardini, A. Bosin, and D. Vanderbilt, in *Proceedings of the 23rd International Conference on the Physics of Semiconductors*, edited by M. Scheffler and R. Zimmermann (World Scientific, Singapore, 1996), p. 2877.
[151] H. Wang and A.-B. Chen, Phys. Rev. B **63**, 125 212 (2001).
[152] C. Ronning, E. P. Carlson, D. B. Thomson, and R. F. Davis, Appl. Phys. Lett. **73**, 1622 (1998).
[153] T. S. Cheng, S. E. Hooper, L. C. Jenkins, C. T. Foxon, D. E. Lacklison, J. D. Dewsnip, and J. W. Orton, J. Cryst. Growth **166**, 597 (1996).
[154] A. Salvador, W. Kim, Ö. Aktas, A. Botchkarev, Z. Fan, and H. Morkoç, Appl. Phys. Lett. **69**, 2692 (1996).
[155] A. F. Wright, J. Appl. Phys. **92**, 2575 (2002).
[156] L. E. Ramos, J. Furthmüller, L. M. R. Scolfaro, J. R. Leite, and F. Bechstedt, Phys. Rev. B **66**, 075 209 (2002).
[157] D. J. As, U. Köhler, and K. Lishka, Mater. Res. Soc. Symp. Proc. **693,** I2.3.1 (2002).
[158] P. Bogusławski, E. L. Briggs, and J. Bernholc, Appl. Phys. Lett. **69**, 233 (1996).
[159] S. Fischer, C. Wetzel, E. E. Haller, and B. K. Meyer, Appl. Phys. Lett. **67**, 1298 (1995).
[160] B. Monemar, H. P. Gislason, and O. Lagerstedt, J. Appl. Phys. **51**, 640 (1980).
[161] M. Ilegems, R. Dingle, and R. A. Logan, J. Appl. Phys. **43**, 3797 (1972).
[162] M. Leroux, P. Vennéguès, S. Dalmasso, M. Benaissa, E. Feltin, P. de Mierry, B. Beaumont, B. Damilano, N. Grandjean, and P. Gibart, Phys. Status Solidi A **192**, 394 (2002).
[163] P. Vennéguès, M. Benaissa, S. Dalmasso, M. Leroux, E. Feltin, P. De Mierry, B. Beaumont, B. Damilano, N. Grandjean, and P. Gibart, Mater. Sci. Eng. B **93**, 224 (2002).
[164] J. Neugebauer and C. G. Van de Walle, Appl. Phys. Lett. **68**, 1829 (1996).
[165] C. Johnson, J. Y. Lin, H. X. Jiang, M. A. Khan, and C. J. Sun, Appl. Phys. Lett. **68**, 1808 (1996).
[166] J. Z. Li, J. Y. Lin, H. X. Jiang, A. Salvador, A. Botchkarev, and H. Morkoc, Appl. Phys. Lett. **69**, 1474 (1996).
[167] C. G. Van de Walle, Phys. Rev. B **56**, 10 020 (1997).
[168] S.-G. Lee and K. J. Chang, Semicond. Sci. Technol. **14**, 138 (1999).
[169] F. A. Reboredo and S. T. Pantelides, Phys. Rev. Lett. **82**, 1887 (1999).
[170] J. Neugebauer and C. G. Van de Walle, in *Proceedings of the 23rd International Conference on the Physics of Semiconductors*, Berlin, 1996, edited by M. Scheffler and R. Zimmermann (World Scientific Publishing Co. Pte Ltd., Singapore, 1996), p. 2849.
[171] F. Bernardini, V. Fiorentini, and A. Bosin, Appl. Phys. Lett. **70**, 2990 (1997).
[172] *Lange's Handbook of Chemistry*, 13th Ed., edited by J. A. Dean, (McGraw-Hill, New York, 1985).
[173] J. Neugebauer and Chris G. Van de Walle, in *Hydrogen in Semiconductors II*, edited by N. H. Nickel, *Semiconductors and Semimetals* Vol. 61, Treatise editors R. K. Willardson and E. R. Weber (Academic Press, Boston, 1999), p. 479.
[174] A. Bosin, V. Fiorentini, and D. Vanderbilt, Mater. Res. Soc. Symp. Proc. **395**, 503 (1996).
[175] A. F. Wright, Phys. Rev. B **60**, 5101 (1999).
[176] S. M. Myers, A. F. Wright, G. A. Petersen, C. H. Seager, W. R. Wampler, M. H. Crawford, and J. Han, J. Appl. Phys. **88**, 4676 (2000).
[177] S. M. Myers, A. F. Wright, G. A. Petersen, W. R. Wampler, C. H. Seager, M. H. Crawford, and J. Han, J. Appl. Phys. **89**, 3195 (2001).
[178] W. Götz, N. M. Johnson, D. P. Bour, M. D. McCluskey, and E. E. Haller, Appl. Phys. Lett. **69**, 3725 (1996).
[179] H. Harima, T. Inoue, S. Nakashima, M. Ishida, and M. Taneya, Appl. Phys. Lett. **75**, 1383 (1999).
[180] B. Clerjaud, D. Côte, A. Lebkiri, C. Naud, J. M. Baranowski, K. Pakula, D. Wasik, and T. Suski, Phys. Rev. B **61**, 8238 (2000).
[181] C. J. Fall, R. Jones, P. R. Briddon, and S. Öberg, Mater. Sci. Eng. B **82**, 88 (2001).
[182] W. Götz, N. M. Johnson, J. Walker, D. P. Bour, and R. A. Street, Appl. Phys. Lett. **68**, 667 (1996).
[183] S. M. Myers, C. H. Seager, A. F. Wright, B. L. Vaandrager, and J. S. Nelson, J. Appl. Phys. **92**, 6630 (2002).
[184] A. F. Wright, J. Appl. Phys. **90**, 1164 (2001).
[185] A. F. Wright, J. Appl. Phys. **90**, 6526 (2001).
[186] D. C. Look, H. Lu, W. J. Schaff, J. Jasinski, and Z. Lilienthal-Weber, Appl. Phys. Lett. **80**, 258 (2002).
[187] E. A. Davis, S. F. J. Cox, R. L. Lichti, and C. G. Van de Walle, Appl. Phys. Lett. **82**, 592 (2003).
[188] O. Brandt, H. Yang, H. Kostial, and K. H. Ploog, Appl. Phys. Lett. **69**, 2707 (1996).
[189] K. H. Ploog and O. Brandt, J. Vac. Sci. Technol. A **16**, 1609 (1998).
[190] R. Y. Korotkov, J. M. Gregie, and B. W. Wessels, Appl. Phys. Lett. **78**, 222 (2001).
[191] G. Kipshidze, V. Kuryatkov, B. Borisov, Yu. Kudryavtsev, R. Asomoza, S. Nikishin, and H. Temkin, Appl. Phys. Lett. **80**, 2910 (2002).
[192] T. Yamamoto and H. Katayama-Yoshida, in *Proceedings of the 19th International Conference on Defects in Semiconductors*, Aveiro, Portugal, 1997, edited by G. Davies and M. H. Nazaré, Mat. Sci. Forum **258-263** (Trans Tech, Zürich, 1997), p. 1185.
[193] S. B. Zhang, S.-H. Wei, and Y. Yan, Physica B **302**, 135 (2001).
[194] C. G. Van de Walle, Phys. Rev. Lett. **85**, 1012 (2000).
[195] X. Nie, S.-H. Wei, and S. B. Zhang, Phys. Rev. B **65**, 075 111 (2002).