Final Report for ONR Contract N00014-84-C-0396
S. T. Pantelides and J. Tersoff, Principal Investigators
1992



The research that was performed under the terms of this contract in the years 1988-1992 falls in two categories: A. Bulk properties, point defects and impurities, and B. Surface morphology and growth. Brief descriptions of the accomplishments are given below. More detailed accounts can be found in the attached published papers.

## A. BULK PROPERTIES, POINT DEFECTS, AND IMPURITIES

### Impurity Diffusion in Silicon

The atomistic mechanisms that mediate dopant impurity diffusion in Si have been debated for many years. Experimental data cannot distinguish different mechanisms in an unambiguous way. We have carried out a variety of first-principles calculations in order to determine the relative importance of different mechanisms. In particular, we explored the role of vacancies and self-interstitials. We computed the corresponding activation energies for several dopant impurities and compared them with experimental diffusion activation energies. We found that for boron, phosphorus, and arsenic both vacancies and self-interstitials contribute with activation energies in the observed range, whereas for antimony only vacancies contribute with an activation energy in the experimentally observed range In addition, we carried out a systematic study of non-equilibrium diffusion, e.g. when excess self-interstitials are injected, and were able to account for several experimental observations. This work is described in the attached papers A and B.

### Hydrogen in Si

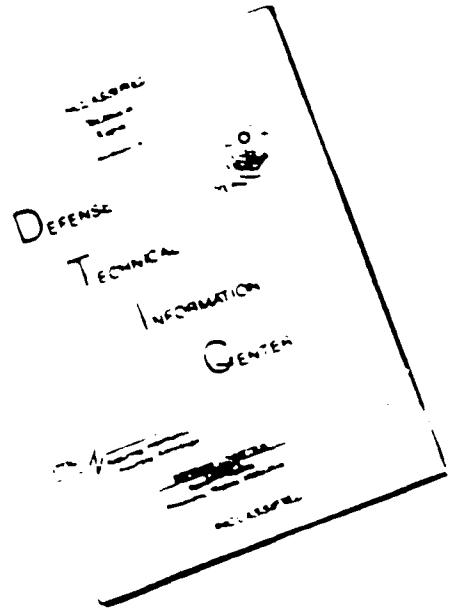
Hydrogen in Si generated a significant amount of attention because a large number of experimental data did not yield a consistent picture. Major questions were the charge state of diffusing H in n-type or p-type Si, how H passivates shallow impurities, and the atomic configurations of H-impurity pairs. Semiempirical theories or even first-principles theories that looked only at a few chosen configurations did not resolve the issues but rather contributed to the puzzles. We carried out a set of comprehensive calculations that provided a systematic description of H in Si. We identified different diffusion paths for the different charge states and showed that H is likely to exhibit negative-U behavior, i.e. it's favored to be either positive or negative, never neutral. We determined the total-energy surfaces for H in the vicinity of impurities such as B and P and found that H rotates around B with an activation energy of 0.2 eV. Independently, experiments established that indeed H rotates about B with such an activation energy. Finally, we used the total-energy surfaces of H in Si to compute a diffusion constant in agreement with experimental data.. This work is described in the attached papers C-G.



### Auger Recombination Rates

Calculations of Auger recombination rates are quite difficult because the involve summations over two initial and two final states. Available calculations involved many approximations. Calculated carrier life times in Si were typically too large compared with observed values, suggesting that phonon-assisted processes may play an important role. We performed very detailed calculations without any compromising approximations,

# DISCLAIMER NOTICE




THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

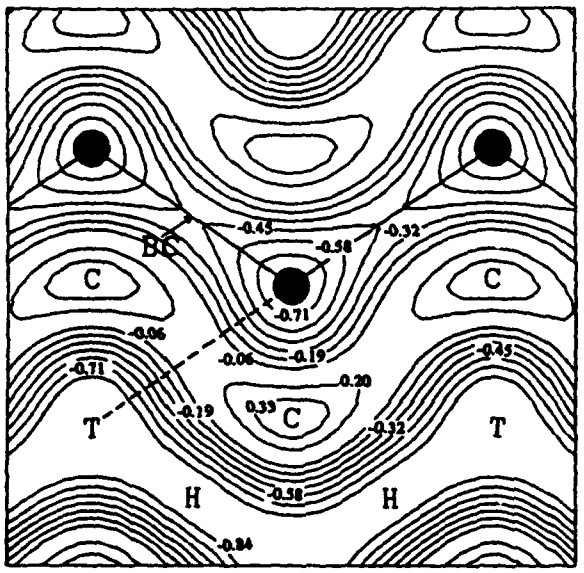



FIG. 1. Total-energy contour plot depicting the migration of a neutral B interstitial through the Si crystal. The labeled sites are $T$ (tetrahedral), $H$ (hexagonal), BC (bond center), and $C$ (at the center of a rhombus formed by three adjacent Si atoms and the nearest $T$). The energy difference between contours is 0.13 eV. The dashed line is the kick-out pathway.

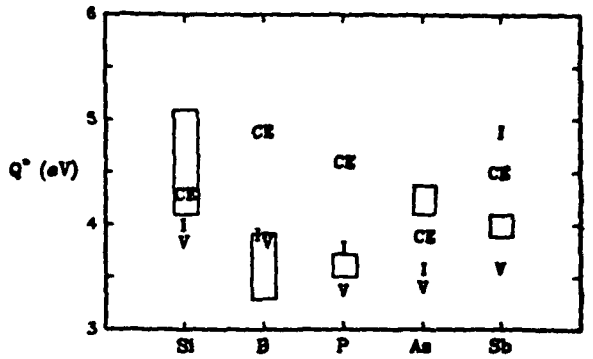



FIG. 2. The calculated activation energies under equilibrium conditions for vacancy-mediated ($V$), interstitial-mediated ($I$), and concerted-exchange (CE) mechanisms for Si self-diffusion and various impurities. The boxed areas are a selected range of experimental results from Ref. 13. The self-diffusion CE activation energy is from Ref. 4.

mediated mechanisms, $Q_j^*$ is the sum of the formation and migration energies for the diffusing species.

For defect-mediated mechanisms, the first task of theory is to determine which defect or complex leads to diffusion with the smallest $Q_j$. For vacancy-mediated ($V$-mediated) diffusion, we find that the relevant diffusing species is the vacancy for B and the impurity-vacancy ($XV$) pair for P, As, and Sb.[10,11] For interstitial-mediated ($I$-mediated) diffusion, a global total-energy surface was needed in order to determine the precise migration pathways and the diffusing species.[12] A contour plot of such a surface in the (110) diamond-structure crystal plane for a neutral B interstitial ($B_I$) is shown in Fig. 1. We find that the energetically preferred diffusion pathway in all cases is the kick-out process,[10,11] so that the diffusing species is the interstitial impurity atom.

The determination of $Q_j^*$ for the CE mechanism necessitates mapping out the entire exchange path and identifying the lowest-energy saddle point. Pandey carried out such a task in the case of self-diffusion (Si-Si exchange), but the reoptimization of the entire path for impurity-Si exchange is an unduly demanding computational task. For our purposes, it was adequate to obtain an upper bound for the saddle-point energy.[11]

The calculated activation energies for $V$-, $I$-, and CE-mediated mechanisms for substitutional B, P, As, and Sb diffusion under equilibrium conditions are shown graphically in Fig. 2. A selected range of experimental values are shown as well.[13] For comparison, we also show the corresponding activation energies for self-diffusion. In all cases, the activation energies are those for neutral species. Species of different charge states have been found to contribute with only slightly different activation energies at the temperatures of interest.[10] We note that for B, P, and As, the $V$- and $I$-assisted-mechanism activation energies are the same within the accuracy of the calculation (0.4 eV is the maximum difference). Because the theoretical CE activation energies are only upper bounds, we view them as comparable to those of defect-mediated mechanisms. Only in the case of Sb does the large difference between activation energies allow us to conclude that the $V$-assisted mechanism dominates, in agreement with conclusions drawn from experimental data.[3] In summary, our calculated activation energies, being in the same range as experimental values, confirm the reliability of our theoretical methods but do not establish the relative importance of the various mechanisms, with the exception of Sb. For such a task, it would be necessary to determine the values of the respective preexponentials, which is currently unfeasible. Nevertheless, we show in the remainder of this paper that theoretical calculations combined with experiments involving injection of excess point defects allow a number of important definitive conclusions to be made.

Excess point defects can be injected into the bulk by surface treatments. For example, evidence has accumulated that oxidation injects self-interstitials whereas nitridation injects vacancies.[2,3,14,15] Under these conditions, the dopant diffusion coefficient in buried impurity layers is either retarded or enhanced. Conclusions derived from such experiments have, however, widely conflicted, largely because they were based on unsatisfactory assumptions. The most serious shortcoming was the implicit assumption that the concentrations of *vacancies* and *self-interstitials* determine the dopant diffusion coefficient. As we made clear earlier in this paper, the relevant concentrations are generally those of either the impurity-vacancy pair ($XV$) or the interstitial impurity ($X_I$). The key element of a correct theory of nonequili-

brium diffusion is to recognize that the diffusion coefficient $D$ is still given by a sum of terms of the form of Eq. (1) where the diffusivities $d_j$ have the same values as in equilibrium. It is then the task of theory to determine the correct expressions for all the relevant $C_j$. We present here the essential elements of such a theory under injection of interstitials (the theory for vacancy injection is completely analogous).

We first determine the effect of excess point defects on the CE mechanism for which $D_j$ is proportional to the concentration of substitutional impurity atoms, $C_{X_s}$. Point-defect injection occurs at comparatively low levels so that it affects $C_{X_s}$, and hence $D_{CE}$, only minimally through the formation of $XV$ pairs or $X_I$. As a result, if the CE mechanism is the dominant impurity diffusion mechanism, the diffusion coefficient *cannot be retarded by moderate levels of injection.* But this is in contradiction with experimental observation.[2,3] We therefore do not discuss the CE mechanism any further in the following treatment.

We now turn to the effects of defect injection on defect-mediated diffusion and, specifically, on the concentrations of the relevant diffusing species, the $XV$ pair or $X_I$. For this purpose, it is essential to identify *all* the reactions that govern the defect concentrations. These are as follows:

$$S \rightleftarrows V \text{ and } S \rightleftarrows I, \tag{3}$$

$$I + X_s \rightleftarrows X_I, \tag{4}$$

$$V + X_s \rightleftarrows XV, \tag{5}$$

$$I + XV \rightleftarrows X_s, \tag{6}$$

$$I + V \rightleftarrows \square, \tag{7}$$

$$V + X_I \rightleftarrows X_s. \tag{8}$$

Reactions (3) are written schematically and indicate that a free or internal surface $S$ can produce or absorb both vacancies and interstitials independent of each other. In reaction (7), the symbol $\square$ represents bulk Si.

Starting with the seven reactions given above, one can immediately obtain four expressions of the form

$$\frac{\partial C_j}{\partial t} = \sum_\alpha g_{j\alpha} - \sum_\alpha r_{j\alpha} C_j \tag{9}$$

for each of the four defect species $j$ ($V$, $I$, $XV$, $X_I$). The terms $g_{j\alpha}$ are generation rates and $r_{j\alpha}$ are recombination frequencies. At steady state, we require $\partial C_j/\partial t = 0$ for all four species. The resulting four equations comprise a complete and exact set which can in principle yield solutions for all relevant concentrations.

By invoking a series of approximations, analytical expressions for the concentrations can be obtained which manifest the essential physical results.[11] The concentrations of $C_{X_i}$ and $C_{XV}$ *under injection of self-interstitials are*

$$C_{X_i} = C_{X_i}^* + K_4 C_{X_s} C_I', \tag{10}$$

$$1/C_{XV} = 1/C_{XV}^* + \beta C_I'/C_{X_s} C_V^*, \tag{11}$$

where $C_I'$ denotes the injected interstitial concentration. $K_4$ is the equilibrium constant of reaction (4) and $\beta$ is a constant containing the reverse rate constant of reaction (7) divided by the forward rate constant of reaction (6).

Combining Eqs. (1), (10), and (11), the total diffusion coefficient has the form

$$D = D_I + D_V, \tag{12}$$

where

$$D_I = D_I^* + D_I' \tag{13}$$

and

$$1/D_V = 1/D_V^* + 1/D_V', \tag{14}$$

where the subscript $I$ ($V$) denotes the $I$- ($V$-) assisted diffusion component. The primes denote the nonequilibrium contribution to diffusion. $D_I'$ and $1/D_V'$ are proportional to $C_I'$ and $1/C_V'$, respectively, as in the corresponding terms in Eqs. (10) and (11). $D_I'$ and $D_V'$ are activated with activation energies given by

$$Q_I' = E_{inj} - \Delta E + E_m, \tag{15}$$

where $E_{inj}$ is the activation energy of the interstitial injection process, $\Delta E$ is the energy difference between $X_I$ and $X_s + I$ as in reaction (4), and $E_m$ is the migration energy of $X_I$; and

$$Q_V' = (E_{inj} + E_{I:XV}^{barrier}) - (E_F^V + E_{V:X_s}^{barrier}) + E_m, \tag{16}$$

where $E_{I:XV}^{barrier}$ ($E_{V:X_s}^{barrier}$) is the energy barrier to recombination of interstitials (vacancies) with $XV$ pairs ($X_s$), $E_F^V$ is the formation energy of vacancies, and $E_m$ is the $XV$ migration energy.

We now discuss three distinct cases. If, under equilibrium, the $I$ component is dominant, interstitial injection leads to an enhanced total diffusion coefficient of the form of Eq. (13), where the activation energy of $D'$ is given by Eq. (15). Hill[16] measured the diffusion coefficients of B, P, and As under interstitial injection and found that they obey Eq. (13). This finding immediately suggests that these impurities diffuse predominantly via an interstitial mechanism. Furthermore, we have calculated the activation energy for these impurities from Eq. (15) using our theoretical values for $\Delta E$ and $E_m$ and an experimental value for $E_{inj}$ extracted from Ref. 17. The results, as well as Hill's measured values, are given in Table I. The excellent agreement between two values corroborates the conclusion that these impurities diffuse primarily assisted by interstitials.

On the other hand, if, under equilibrium, the $V$ component is dominant, interstitial injection can lead to either diffusion retardation or enhancement, depending

TABLE I. Activation energies for diffusion mediated exclusively by interstitials under interstitial injection [theory, Eq. (15)] and measured activation enthalpies under oxidation conditions (experiment from Hill, Ref. 16). All quantities are in eV.

| Species | $Q_i$ (theory) | $Q'$ (experiment) |
|---|---|---|
| B | 2.6 | 2.3 |
| P | 2.5 | 2.4 |
| As | 2.3 | 2.3 |

upon the level of injection. At low levels of injection, the $I$ component remains small while the $V$ component is retarded according to Eq. (14). It is generally believed that Sb diffusion under interstitial injection is such a case,[13] but data are usually reported at a single temperature. We predict that temperature-dependent data would obey Eq. (14) so that a *reciprocal Arrhenius plot* (i.e., $1/D$ vs $1/T$) would be appropriate to extract an activation energy for $D_i^*$ to be compared with Eq. (16). Such data would provide a test of our theory and assess the conclusion that Sb diffuses predominantly by a $V$ mechanism. If interstitial injection were to occur at high levels, the term $D_i^*$ will ultimately overwhelm all other contributions and enhanced diffusion with an activation energy given by Eq. (15) would be observed.

Lastly, if, under equilibrium, the $V$ and $I$ components are comparable, interstitial injection can lead to either a net enhancement or a net retardation, according to Eqs. (12)–(14). Temperature-dependent plots would be rather complex but under certain conditions the new curve will *cross* the corresponding equilibrium curve. Such crossing is a definitive experimental signature of a dual mechanism. Unfortunately, temperature-dependent data of diffusion under vacancy or interstitial injection, though highly desirable, are scarce.

We acknowledge many helpful discussions with J. Bernholc and P. M. Fahey. This work was supported in part by ONR contract No. N000014-84-C-0396.

[1]D. Mathiot and J. C. Pfister, J. Appl. Phys. 55, 3518 (1984).

[2]D. A. Antoniadis and I. Moskowitz, J. Appl. Phys. 53, 6788 (1982).

[3]P. Fahey, G. Barbuscia, M. Moslehi, and R. W. Dutton, Appl. Phys. Lett. 46, 784 (1985).

[4]K. C. Pandey, Phys. Rev. Lett. 57, 2287 (1986).

[5]S. M. Hu, J. Appl. Phys. 57, 1069 (1985).

[6]P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964); W. Kohn and L. J. Sham, Phys. Rev. 140, A1133 (1965). The exchange and correlation potentials are based on work by D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. 45, 566 (1980), as parametrized by J. Perdew and A. Zunger, Phys. Rev. B 23, 5048 (1981).

[7]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. 43, 1494 (1979).

[8]Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B 30, 1844 (1984).

[9]Atomic configurations of high symmetry require only two special $k$ points.

[10]See R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. 54, 360 (1985), for the specific details regarding the defect-assisted mechanisms. The impurity-vacancy-pair migration energy is taken from the experiments reported in M. Hirata, M. Hirata, and H. Saito, J. Phys. Soc. Jpn. 27, 405 (1969).

[11]See C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides (to be published) for more details.

[12]Chris G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. 60, 2761 (1988). Following the methodology outlined in this paper, we have used six stars of reciprocal-lattice vectors and the total energy of the interstitial at fifteen sites throughout the crystal to generate the total-energy surface.

[13]From P. M. Fahey, P. B. Griffin, and J. D. Plummber, Rev. Mod. Phys. (to be published).

[14]S. M. Hu, J. Appl. Phys. 45, 1567 (1974).

[15]T. Y. Tan, U. Gösele, and F. F. Morehead, Appl. Phys. A 31, 97 (1983).

[16]C. Hill, in *Semiconductor Silicon 1981*, edited by H. R. Huff, J. R. Kriegler, and Y. Takeishi (Electrochemical Society, New York, 1981), p. 988.

[17]S. M. Hu, Appl. Phys. Lett. 27, 165 (1975).

carrying out for the first time the appropriate multiple sum over the Brillouin zone. We found that in electron Auger lifetimes are much smaller than previously thought and compare well with experimental values, thus obviating the need to invoke phonon assistance. This work is described in the attached paper H.

Doping Levels and Compensation in ZnSe

It has long been known that ZnSe is easily doped n-type but very difficult to dope p-type. In contrast, ZnTe is easily doped p-type, but not n-type. A number of explanations have been proposed, e.g., native defects are such that they trap the would-be free holes in ZnSe or the would-be free electrons in ZnTe; large lattice relaxation makes the would-be shallow acceptors deep in ZnSe; other configurations of the impurity, such as interstitial, may have deep levels that capture the free carriers; solubility may be low altogether. Neither experiment nor theory could determine which one of these possible mechanisms actually operate and why such similar materials behave so differently. We first computed the concentrations of native defects and showed that they are too small to matter. Then we constructed a comprehensive theory in terms of which we could actually calculate both solubilities and doping levels, including the Fermi level. Through first-principles calculations we implemented the theory for ZnSe and found that the major problem for doping is one of solubility, because of formation of other competing phases. The existence of interstitial impurities played a secondary role. We found that nitrogen would be the most effective dopant, as found independently by experiments. This work is described in the attached papers I-K

Forces on dopant impurities at semiconductor interfaces

Diffusion of dopants across interfaces is a central problem in device fabrication. The potential barrier felt by a neutral impurity at an interface depends simply upon the enthalpies in the respective materials. However, the potential felt by a *charged* impurity or defect, such as a dopant, is a more subtle problem. While it was recognized that a charged impurity would respond to electric fields, it was unclear what role the band lineup played, since for electrons the lineup plays a role analogous to an electric field. This remained an outstanding puzzle until we provided a rigorous solution. We showed that the donor or acceptor level of the defect plays much the same role for the ion that the band edge plays for an electron or hole in defining the effective potential. This work is described in the attached paper L.

Carbon Impurities in Silicon

The predominant defect in silicon is dissolved (substitutional) carbon. Substitutional carbon also forms complexes with other defects. Our calculations provided the first theoretical perspective on the energetics of carbon defects and defect reactions in silicon. Moreover, we showed that the accepted experimental value for the enthalpy of substitution was incorrect. Not only did our calculation give a different value, but a reanalysis of the measured data confirmed this result. This is particularly important since the enthalpy of substitution controls the solubility of carbon in silicon, and hence the driving force for precipitation (a major source of imperfections in commercial silicon wafers). This work is described in the attached paper M.

(B)

Reprinted from: Physical Review B, Vol. 40, No. 8 September 15 1989, Part 1

# Mechanisms of dopant impurity diffusion in silicon

C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides

## B. SURFACE MORPHOLOGY AND GROWTH

### Steps on Si (001) Surfaces

In the last few years there has been great interest in surface steps, because steps are crucial in controlling growth. Steps on Si(001) can be either one or two atomic layers high. Single-layer steps lead to antiphase boundaries in GaAs grown on Si, while double steps do not. The competition between single and double steps on Si(001) has therefore been studied intensely. It was believed that stress-related interactions between steps led to an abrupt transition, from single steps on surfaces oriented within a critical angle of (001), to double steps for orientations above this critical angle. Yet we showed that the behavior is actually quite different. As the surface orientation is varied, the transition from single steps to double steps takes place through a "devil's staircase" of intermediate phases, consisting of ordered mixtures of single and double steps. Moreover, we showed that the dominant thermal excitation of double steps is break-up into single-step pairs. As a result, there is an unanticipated critical point in the surface phase diagram, above which the thermodynamic distinction between single and double steps disappears. Thus the temperature as well as the angle is crucial in controlling anti-phase domain formation in growth of GaAs on Si.

We also identified an entirely new surface phase defined by wavy steps. A certain density of steps is energetically favorable, because steps create stress domains. For very flat surfaces, where the step density becomes too low, the steps spontaneously become wavy, effectively increasing the step density without the need to nucleate new steps. This behavior has been observed experimentally.

The above work is described in detail in the attached papers N-P.

### Surface segregation and bulk ordering in alloys

Semiconductor alloys are generally expected to be almost perfectly disordered. It was therefore quite a surprise when, in 1985, ordering was seen in MBE-grown samples of Si-Ge and III-V alloys. Such bulk ordering is thermodynamically unfavorable; yet it quickly proved to be as ubiquitous as it was mysterious. Ordering is important since it changes the band gap and other properties.

The surprising explanation came from studies of surface segregation in Si-Ge alloys, in conjunction with the experimental work by LeGoues et al. We found that the dimerization at a Si or Ge (001) 2×1 surface creates large stresses down to the fourth layer, driving a remarkable *lateral* segregation in the third and fourth layers. The "bulk" ordering arises when the pattern of stress-driven surface segregation is frozen into each layer as it is buried by the next layer. Experiments directly confirmed the surface origin of the ordering, highlighting the importance of a microscopic understanding of growth. This work is described in detail in the attached paper Q.

### Growth of strained layers

Strained layers are becoming increasingly important in semiconductor technology, so the question of their growth and stability has acquired some urgency. Epitaxial strained layers tend to break up into islands; yet the growth of uniform layers is an es-

sential step in the fabrication of many semiconductor devices. For the simplest such case, Ge on Si(001), the Ge grows up to three atomic layers as a uniform film, but any additional Ge forms islands. By a series of calculations of film energy, we identified the principal factor stabilizing the three-layer film. It is energetically favorable for the local distortions associated with surface dimerization to lie in a soft material, and Ge is softer than Si. For the first three layers, this energy gain is enough to offset the strain energy from the mismatch between Ge and Si. Thus the thickness of the stable film is determined by the depth of the elastic distortions associated with the surface reconstruction. This work is described in detail in the attached paper R.

PHYSICAL REVIEW B VOLUME 40, NUMBER 8 15 SEPTEMBER 1989-I

# Mechanisms of dopant impurity diffusion in silicon

C. S. Nichols, C. G. Van de Walle,* and S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598*
(Received 20 April 1989)

We present a comprehensive investigation of dopant diffusion in silicon under equilibrium and nonequilibrium concentrations of intrinsic point defects. Using first-principles total-energy calculations combined with available experimental data, we seek to resolve a series of outstanding controversies regarding the diffusion mechanisms of B, P, As, and Sb in silicon. We find that, under equilibrium conditions, vacancies and interstitials mediate the diffusion of all dopants with comparable activation energies, except Sb, for which the interstitial component has a high activation energy. Under nonequilibrium conditions, e.g., under injection of excess point defects, we derive the relevant expressions for the activation energy for a variety of possible diffusion mechanisms and injection conditions. Under oxidation, the calculated values are in excellent agreement with the available experimental data. In addition, theory and experiment suggest that the concerted exchange mechanism, involving no point defects, plays only a minor role in dopant diffusion.

## I. INTRODUCTION

Diffusion of impurity or host atoms through crystalline solids has been studied extensively for many years.[1,2] There are two basic types of mechanisms by which substitutional dopant impurities can diffuse. Diffusion can be mediated either by native point defects, such as vacancies and self-interstitials, or by an intrinsic mechanism that occurs spontaneously (i.e., in the absence of defects) in the bulk. In addition, there have been some suggestions that "extended defects" may play a role in high-temperature diffusion processes.[3] While experimental and theoretical approaches have considered these atomistic mechanisms of diffusion, most evidence for or against a particular mechanism has been indirect; for example, the comparison between an impurity profile measured experimentally and one derived analytically or numerically. In recent years, however, both experimental and theoretical tools have been developed with which diffusion can be studied from a microscopic, atomistic viewpoint directly. Despite these advances, there is no general consensus regarding the relative contributions of the various mechanisms to impurity diffusion.

It has been observed that a number of surface processing conditions[4] alter the bulk point-defect concentration in Si. Oxidation, for example, has been shown to inject excess self-interstitials, while nitridation of the surface injects excess vacancies.[5] Although the details of such processes are not completely understood, measurement of dopant diffusion coefficients under equilibrium concentrations of point defects and under oxidation or nitridation conditions affords the possibility of discriminating between the mechanisms responsible for dopant diffusion. However, the theories underlying the interpretation of such experiments are incomplete, relying for the most part on assumptions whose validity is uncertain.[6] For example, in oxidation experiments, Antoniadis and Moskowitz[7] observed that the P diffusion coefficient is enhanced with respect to its equilibrium value. They concluded that P diffuses via a *dual vacancy-interstitial mechanism*, although the microscopics of the interstitial mechanism were not specified. Also under interstitial injection, Fahey *et al.*[8] found an identical P diffusion enhancement. However, under injection of excess vacancies they observed a retardation of the P diffusion coefficient with respect to its equilibrium value. From these two experiments, they concluded that P diffusion is almost *exclusively mediated by self-interstitials.* No attempt has been made to assess the contribution of any intrinsic mechanism in these experiments.

Numerical solution of the coupled system of equations governing diffusion in Si offers relatively inexpensive and quick insight into possible mechanisms. However, the complexity of the relevant equations has made solution of the full problem unfeasible, and the consequent simplifying assumptions made are often unrealistic.[9,10] In one such simulation, the authors concluded that P diffuses exclusively by a *vacancy* mechanism,[11] in conflict with the nonequilibrium experimental conclusions above.

First-principles calculations have also recently addressed the problem of impurity diffusion pathways and mechanisms.[12,13] The work of Ref. 12 considered only the migration of aluminum as an interstitial and did not address the issue of a vacancy mechanism. The work of Ref. 13 focused on equilibrium conditions and considered defect-mediated mechanisms. This work established that native *point* defects mediate impurity diffusion with activation energies comparable to experimental values, obviating the need for "extended defects." No definitive conclusion was reached in this work regarding the dominance of one point-defect species over the other, however. Finally, although thorough first-principles calculations have demonstrated that a concerted exchange (CE) mechanism is energetically comparable to defect mechanisms for self-diffusion,[14] no quantitative support that this mechanism is relevant for dopant impurity diffusion has been offered. An excellent overview of the various experiments, simulations, and theoretical calculations probing

# Mechanisms of Equilibrium and Nonequilibrium Diffusion of Dopants in Silicon

C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides
*IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 23 November 1988)

We have developed a theory of impurity diffusion in silicon under equilibrium and nonequilibrium concentrations of point defects. The results of first-principles calculations of several key quantities are combined with this theory and compared to experimental data. We find that vacancies and self-interstitials mediate the equilibrium diffusion of B, P, and As with comparable activation energies, but interstitials are dominant. Sb diffusion, on the other hand, is mediated primarily by vacancies. We also find that the direct-exchange mechanism plays only a minor role for all dopants studied.

PACS numbers: 61.70.Bv, 66.30.Jt, 71.55.Ht

The microscopic mechanisms of dopant-impurity diffusion in Si are of intrinsic scientific interest and also form the cornerstones of modeling programs for the design and fabrication of devices. Despite extensive research on the subject, there exists no general consensus regarding the relative contributions of the various mechanisms. For example, some authors analyze or numerically fit experimental data and conclude that phosphorus (P) diffusion is primarily assisted by vacancies,[1] whereas others conclude that P diffusion is mediated in part or primarily by self-interstitials.[2,3] More recently, Pandey[4] extrapolated his results for self-diffusion and suggested that the concerted exchange (CE), which requires no intrinsic defects, may be the dominant mechanism for dopant-impurity diffusion. Dopant diffusion experiments under injection of excess concentrations of point defects have offered promise for unraveling these controversies.[2,3] However, theories underlying the interpretation of such experiments are unsatisfactory, relying for the most part on assumptions whose validity is uncertain.[5]

In this Letter, we present the main results of an extensive theoretical investigation of the energetics of diffusion for several dopant impurities (B, P, As, and Sb) in Si under both equilibrium and nonequilibrium conditions. In the first part, we report the theoretical activation energies for equilibrium diffusion and compare with available experimental data. In the second part, we present a systematic theory of impurity diffusion under injection of excess point defects. In particular, we derive expressions for the activation energies of diffusion and predict the expected form for the diffusion coefficient. Combining available experimental data with our theoretical results allows us to draw a number of definitive conclusions. We conclude that B, P, and As all have substantial self-interstitial diffusion components, that Sb is primarily assisted by vacancies, and that the exchange mechanism plays only a minor role in substitutional-dopant diffusion.

In the calculations we use density-functional theory,[6] the local-density approximation, and norm-conserving pseudopotentials.[7] The supercell method is used to solve the relevant Schrödinger equation following the methodology of Ref. 8. Convergence with respect to all variables has been carefully studied. In particular, it was found necessary to use plane waves up to a kinetic energy of 20 Ry for the wave functions and potentials (plane waves above 10 Ry are included in second-order Löwdin perturbation theory), up to 32-atom supercells, and up to three special $k$ points in the irreducible wedge of the Brillouin zone.[9] Unless otherwise stated, relaxation of the surrounding Si network was calculated for every location of the impurity or defect. We estimate our total error to be less than 1 eV, depending upon the particular impurity and the particular atomic configuration. The majority of the error comes from the local-density-approximation uncertainty in the defect- and impurity-related levels in the energy gap. Nevertheless, this scheme has a proven reliability in calculating a number of properties of semiconductors.

The diffusion coefficient $D$ is a sum of contributions of the form

$$D_J = C_J d_J / C_{X_s}, \tag{1}$$

where $C_J$ is the concentration of the defect $J$ whose long-range migration effects diffusion of the substitutional dopant, $d_J$ is the corresponding diffusivity, and $C_{X_s}$ is the concentration of substitutional impurities. For the CE mechanism, the pertinent defect is the substitutional impurity itself, whereas for defect-mediated diffusion, this species needs to be identified (e.g., it can be either the isolated vacancy, the impurity-vacancy pair, etc.). In all cases, under equilibrium conditions, $D_J$ can also be written in the Arrhenius form

$$D_J^* = D_{J,0}^* \exp(-Q_J^*/kT), \tag{2}$$

where $D_{J,0}^*$ is the preexponential which contains a variety of factors, $Q_J^*$ is the activation energy, $k$ is the Boltzmann's constant, and $T$ is the temperature. The asterisk denotes equilibrium quantities. For the CE mechanism, $Q_J^*$ is the energy needed to place a pair of atoms (the impurity atom and one of its nearest-neighbor Si atoms) at the saddle point of the exchange path. For defect-

dopant diffusion is given in the review by Fahey *et al.*[15]

We recently published a brief account of an extensive theoretical study of dopant diffusion mechanisms in Si.[16] In this paper, we give a more extensive and detailed exposition of that work. In particular, we use first-principles state-of-the-art calculations to investigate vacancy-, interstitial-, and CE-mediated diffusion pathways of substitutional B, P, As, and Sb in Si. Activation energies for equilibrium conditions are calculated and compared to available experimental work. In addition, we discuss a systematic framework for impurity diffusion under nonequilibrium concentrations of point defects. Expressions for the activation energies of diffusion in terms of theoretically and experimentally available numbers are found and predictions for the expected form of the diffusion coefficient are given. Using these results, in conjunction with available experimental data, we can discriminate between the different mechanisms. We find that P, As, and B diffusion have substantial interstitial components, while Sb diffusion is vacancy dominated. Theoretical results and experimental data suggest that the CE mechanism has a limited role in dopant diffusion. In large part, we confirm the picture advocated by Fahey and co-workers[8] with regard to the point-defect mechanisms.

## II. METHODOLOGY

Our calculations are based on Hohenberg-Kohn density-functional theory, the Kohn-Sham local-density approximation (LDA) for exchange and correlation,[17] and norm-conserving pseudopotentials[18] for the electron-ion interaction. The relevant Schrödinger equation is solved to obtain the total energy by a momentum-space formalism[19] in a supercell geometry.[20] We describe the essential aspects of this methodology below. A useful and more comprehensive discussion can also be found in a recent review by Pickett.[21]

For the electron-electron interactions, density-functional theory, within the LDA, is utilized. The LDA consists of the assumption that the exchange and correlation energy at a point r is a function of the electron density only at r. The approximation is considered valid for systems with slowly varying electron densities. In practice, in semiconductors the LDA has proven remarkably successful. But it is by now well known that the LDA predicts conduction-band states and levels derived mostly from conduction-band states to be too low in energy. This is the major source of error in the methodology.

The pseudopotentials used in our calculations are generated according to the Hamann-Schlüter-Chiang scheme.[18] More details concerning the specifics of the Si potential and results of test calculations carried out for this potential are given in Ref. 22. For B impurities, we use a pseudopotential first discussed by Denteneer *et al.*[23] The cutoff radii of the B pseudopotential were adjusted so as to minimize the basis-set size, but still faithfully describe the properties of B impurities in Si. The convergence properties of this pseudopotential are fully discussed in Ref. 23. For the donor impurities treated here, P, As, and Sb, we used the pseudopotentials as tabulated by Bachelet, Hamann, and Schlüter.[24]

The solution of the relevant Schrödinger equation in a supercell geometry has been extensively exploited in many calculations, including superlattice geometries, defects, and amorphous semiconductors. In the present work, we follow the methodology of Bar-Yam and Joannopoulos[20] for combining density-functional theory and the pseudopotential approximation in a supercell framework. The supercell approach artificially introduces periodicity by translating a unit cell, which contains the defect or impurity, along its three direct-lattice vectors until all of space is filled. Convergence of the unit-cell size is achieved when the defects in neighboring cells interact by less than some desired tolerance (as manifested by the dispersion of the defect levels). Furthermore, enough neighbors of the impurity are required so as to obtain accurate relaxations. For example, an impurity atom at the bond-center (BC) position causes a large disruption of the crystalline network such that relaxation of at least two shells of neighbors are important.

The wave functions and potentials are expanded in a plane-wave basis. Convergence of the basis-set size was extensively and thoroughly tested. Plots of the total-energy difference between As impurities in two different sites in the crystal are shown in Fig. 1. Figure 1(a) shows

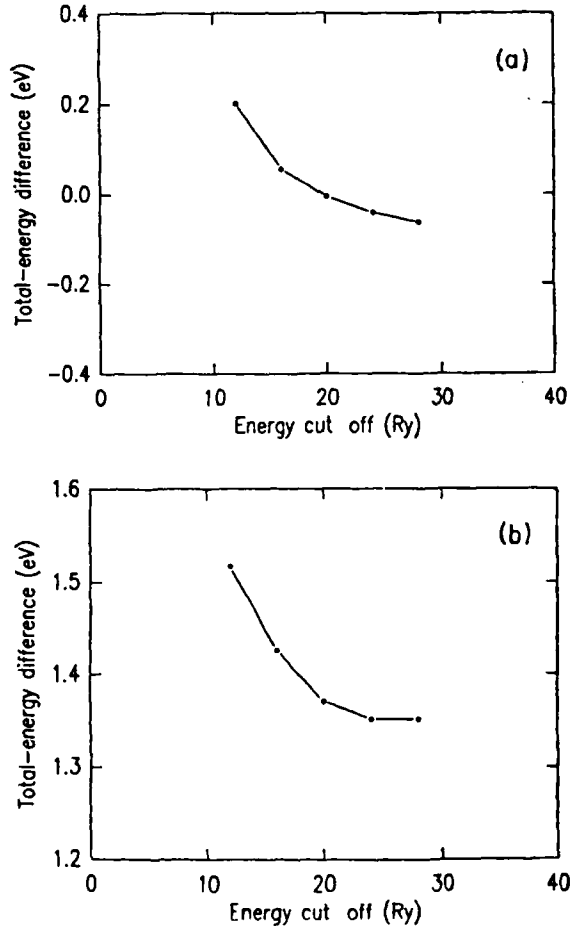



FIG. 1. Convergence of the total-energy differences in an 8-atom supercell as a function of the basis-set cutoff energy. (a) Total-energy difference between As at the $T$ and $H$ sites. (b) Total-energy difference between As at the BC and $T$ sites.

the total-energy difference between a neutral As atom at the high-symmetry tetrahedral ($T$) and hexagonal ($H$) sites as a function of the energy cutoff. The calculations were performed in an 8-atom supercell, which is sufficient for basis-set–convergence studies. The abscissa is the kinetic energy $E$ of the plane waves used in diagonalizing the Hamiltonian. The points plotted are for the highest-energy plane waves included in any basis set; plane waves from 0 to $E/2$ are utilized in the exact diagonalization process, while those between $E/2$ and $E$ are treated in second-order Löwdin perturbation theory.[25] We introduce the notation $(E_1;E_2)$ to indicate the two energy cutoffs. From Fig. 1(a) we conclude that the total-energy difference has achieved convergence at a cutoff of (10;20) Ry to within 0.05 eV of its final value. We have also tested the Löwdin perturbation theory for this particular total-energy difference by computing the total energy at the cutoff (20;20) Ry. Results of these calculations show that at (10;20) Ry, the additional plane waves cause deviations only on the order of 0.03 eV.

Figure 1(b) depicts a somewhat different test case. The total-energy difference between neutral As at the $T$ and (BC) sites is shown. At a cutoff of (10;20) Ry, the total-energy difference is within 0.02 eV of its fully converged value. Test calculations have also been performed in 16- and 32-atom cells [up to (12;24) Ry] and show that the cutoff (10;20) Ry is sufficient for reliable conclusions. We have also performed similar tests for P and Sb and observe that these impurities have convergence properties analogous to As.

We have performed supercell-size convergence tests for all impurities in 8-, 16-, and 32-atom cells. The maximum error in total-energy differences encountered in scaling from 16- to 32-atom cells is found to be 0.3 eV for P, As, and Sb impurities in Si. For B impurities, the maximum change in total-energy differences between any two atomic configurations is much smaller, only 0.1 eV. We therefore use 32-atom cells throughout these calculations, such that the distance between defects in neighboring cells is 9.4 Å.

Integrations over the Brillouin zone to obtain the charge density are performed using a special-points scheme. The special points are generated according to the algorithm of Monkhorst and Pack.[26] In 32-atom cells, a sampling of two special points has proven adequate for high-symmetry configurations. Lower-symmetry configurations require a larger, but equivalent, set.

Unless otherwise explicitly stated, relaxation of the surrounding Si host network is calculated for every location of the impurity or defect. Hellman-Feynman forces are not obtained in the present calculations. Instead, we relax the first-neighbor shell of atoms to at least three different positions, and use the resulting total energies to fit a parabola to obtain the minimum-energy distance. All relaxations reported in this work are radially away (or towards) the defect and hence are symmetry preserving. Tests have indicated that other relaxations have only minor effects on the total energy. For configurations which cause little distortion of the network (e.g., $H$ or $T$), only the first shell of neighbors is relaxed. For configurations which cause severe disruption (e.g., BC), two shells of neighbors are relaxed.

In order to test all elements of this methodology, we have calculated the relaxation of the Si host surrounding neutral substitutional impurities. The four Si neighbors of substitutional B, which has a covalent radius $\approx 75\%$ that of Si,[27] experience an inward "breathing"-mode relaxation of 0.2 Å, with an accompanying decrease of the total energy of 0.8 eV. Despite its small size with respect to Si and its tendency of form threefold-coordinated molecules, B remains at the nominal substitutional site (to within 0.1 Å). Substitutional P, with a covalent radius slightly smaller than Si, is found to cause no distortion of the host network. Both As and Sb have larger covalent radii than Si and cause an outward "breathing"-mode distortion of the four neighboring Si atoms. The Si-As interatomic distance is calculated to be 2.43 Å and the Si-Sb interatomic distance is 2.54 Å. The Si-As distance is in excellent agreement with extended x-ray-absorption fine-structure (EXAFS) measurements,[28] which give 2.41 Å. The results for P and As are also in good agreement with previous total-energy calculations on positively charged substitutional donors.[29]

We estimate our total error to be less than 1 eV, depending, of course, upon the atomic configuration and the particular impurity. The majority of the error comes from the LDA uncertainty in the defect- and impurity-related levels in the energy gap. Overall, this scheme has a proven reliability in calculating bulk properties of semiconductors, reconstruction of semiconductor surfaces, and general properties of defects.[21]

## III. EQUILIBRIUM DIFFUSION

### A. Background and atomistic mechanisms

Under either equilibrium or nonequilibrium conditions, the diffusion coefficient $D$ is given by a sum of contributions of the form

$$D_i = \frac{C_i d_i}{C_X} , \tag{1}$$

where $C_i$ is the concentration of the defect $i$ whose long-range migration effects diffusion of the substitutional dopant, $d_i$ is the corresponding diffusivity, and $C_X$ is the total concentration of impurities.

For the CE mechanism, the pertinent defect is the substitutional impurity itself, whereas for defect-mediated diffusion this species needs to be identified. For impurity diffusion mediated by vacancies, it is convenient to identify two distinct limits. In the first limit, impurity-vacancy binding is weak, so that when a vacancy positions itself next to an impurity, the two switch places, the vacancy migrates away, and the impurity awaits the arrival of another vacancy in order to migrate another step. Such a process is the same as that for self-diffusion; i.e., the Si vacancy is the relevant diffusing species. The impurity diffusion activation energy is equal to the sum of the vacancy-formation energy and the migration energy of either a Si atom or an impurity atom into a vacant adjacent

site, whichever is larger. Thus, the impurity diffusion activation energy for this mechanism is equal to or larger than the activation energy of self-diffusion. In the second limit, impurity-vacancy binding is strong, so that when a vacancy positions itself next to an impurity atom, migration of the *pair* occurs. After exchanging positions, the vacancy moves away from the dopant atom around a sixfold ring to at least a third-neighbor position. It can then return by a different path, placing itself next to the impurity. The vacancy and the impurity exchange and the process repeats itself. The net activation energy is the sum of the pair-formation and migration energies and can be less than the activation energy for self-diffusion.

Self-interstitial-mediated diffusion can occur in a variety of ways. Two processes believed important are what we call "coordinated push" of a self-interstitial on a substitutional impurity along the bonding direction towards a Si neighbor[13] [see Figs. 2(a)–2(c)] and the kick-out mechanism[13] [see Figs 2(d)–2(f)]. The coordinated push mechanism, similar to the vacancy-mediated mechanisms described above, has two extreme limits, depending upon the binding energy of the impurity–self-interstitial pair. In either extreme limit, at the saddle point of the coordinated push process [Fig. 2(b)] the impurity atom is at the BC site, while the neighboring Si atom is pushed off its substitutional site towards the channel $T$ site. The impurity ends up in the next substitutional site, its former Si neighbor now in the channel as a self-interstitial [Fig. 2(c)]. The original self-interstitial is a nearest neighbor of the impurity. In the weak-binding limit, the new self-interstitial can readily migrate away and, to move to the next substitutional site, the impurity must await the arrival of another self-interstitial. In this limit, then, it is the self-interstitial which mediates long-range migration of the substitutional impurity. The activation energy for diffusion is the sum of the formation energy of the substitutional impurity–self-interstitial pair plus the migration energy of the impurity over the BC saddle point. The activation energy may thus be equal to or larger than the activation energy for self-diffusion, depending upon the differences in migration energy. In the strong-binding limit, the new self-interstitial remains bound to the substitutional impurity and the former can then execute exactly the same coordinated push. In the strong-binding limit, it is the self-interstitial–impurity *pair* which effects long-range migration of the impurity. The activation energy for the strong-binding limit may be less than the activation energy for self-diffusion as a result of binding. The second interstitial process involves the kick-out of the substitutional impurity [Fig. 2(e)] into the low-electron-density channel in which it migrates with a rather small barrier. After migration along the channel, the impurity kicks back into a substitutional site, ejecting a Si atom into the channel. The diffusing species is thus the interstitial impurity.

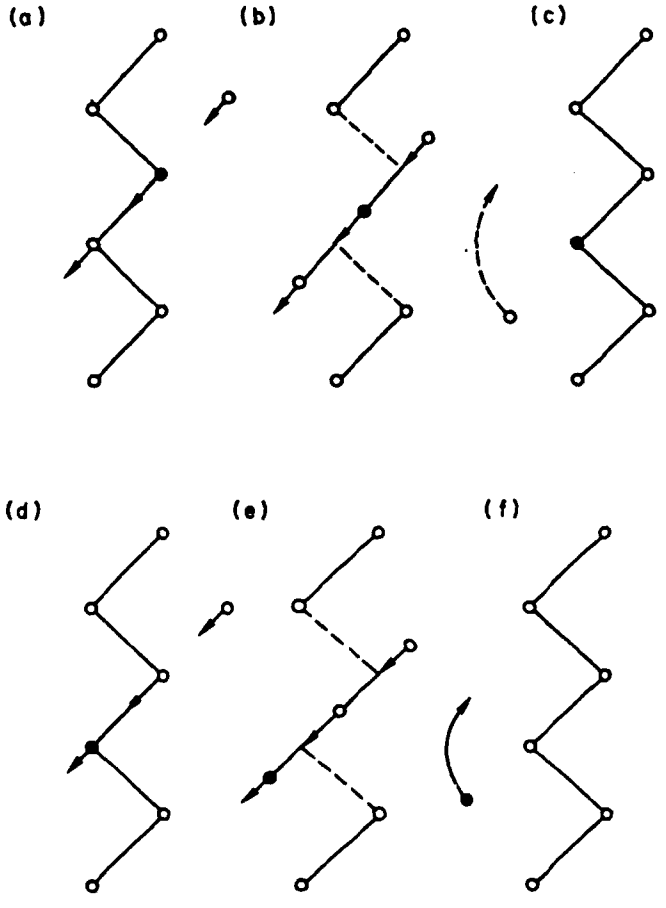



FIG. 2. Schematic diagram showing two interstitial-mediated diffusion processes considered in this paper. (a)–(c) Schematic diagram of the (110) plane in diamond structure Si depicting the "coordinated push" diffusion process. Panel (a) depicts the initial positions of all atoms, while panel (c) depicts the final positions. Panel (b) shows the saddle-point configuration. (d)–(f) The same crystal plane in Si, but now depicting the kick-out diffusion process. Panel (d) depicts the initial atomic positions, while panel (f) shows the final positions. Panel (b) depicts the saddle-point configuration.

For all mechanisms, under equilibrium conditions, the individual diffusion coefficients $D_i$ can also be written in the Arrhenius form

$$D_i^* = D_{i,0}^* \exp(-Q_i^*/k_B T) . \quad (2)$$

The preexponential $D_{i,0}^*$ contains a variety of factors, including the entropy of diffusion. $Q_i^*$ is the activation energy, $k_B$ is Boltzmann's constant, and $T$ is the temperature. The asterisks denote equilibrium quantities. For the CE mechanism, $Q_i^*$ is the energy required to place an impurity atom and one of its Si neighbors at the exchange-path saddle point. For defect-mediated mechanisms, $Q_i^*$ is the sum of the formation and migration energies for the diffusing species.

The determination of $Q_i^*$ for the CE mechanism necessitates mapping out the entire exchange path and identifying the lowest-energy saddle point. Pandey[14] has carried out such a task in the case of self-diffusion (Si-Si exchange), but the reoptimization of the entire impurity-Si exchange is an unduly demanding computational exercise. For our purposes, it was adequate to obtain an upper bound for the saddle-point energy. We assumed the same path as for the Si-Si exchange, calculated the to-

tal energy of the saddle point, and included the relaxation of the neighboring atoms by using Pandey's calculated relaxation for pure Si (0.75 eV).

In general, for defect-mediated diffusion, the defects responsible for diffusion may exist in several charge states, each of which can contribute to diffusion. For the temperatures at which diffusion experiments are performed (900–1100 °C), even for relatively high doping levels (e.g., $10^{18}$–$10^{19}$ $cm^{-3}$), the Fermi level is at midgap. Previous calculations[13] on dopants in Si have shown that all charge states have roughly the same formation energies for interstitial impurities or impurity-vacancy pairs when the Fermi level is near midgap (to within approximately 0.3 eV). The energetics of diffusion are therefore relatively insensitive to the dopant charge state. In this paper, we report results for neutral species only.

Formation energies for any impurity-defect complex are always defined in our calculations with respect to the substitutional impurity in the absence of any defects. The formation energy for an impurity-vacancy pair ($XV$) in an $N$-atom cell is defined as

$$E_f(XV) \equiv E(XV) - E(X_s) + \frac{1}{N} E_{\text{bulk}} , \tag{3}$$

where $E(XV)$ is the calculated total energy per supercell containing an impurity-vacancy pair, $E(X_s)$ is the total energy per supercell containing a substitutional impurity, and $E_{\text{bulk}}$ is the total energy per supercell of pure bulk Si. The formation energy for an interstitial impurity ($X_i$) in an $N$-atom cell is defined as

$$E_f(X_i) \equiv E_f(I) + [E(X_i) - E(X_s\text{-}I)] , \tag{4}$$

where $E_f(I)$ is the formation energy of a Si self-interstitial, $E(X_i)$ is the total energy per supercell containing an interstitial impurity, and $E(X_s\text{-}I)$ is the total energy per supercell containing a (substitutional impurity–$I$) pair. The formation energy of an $I$ in pure Si is defined as

$$E_f(I) \equiv E(I) - \frac{N+1}{N} E_{\text{bulk}} , \tag{5}$$

where $E(I)$ is the total energy per supercell containing a Si interstitial. Note that it makes no difference in finding the formation energy of an interstitial impurity which position we use for the Si interstitial, as long as we use a consistent choice throughout Eq. (4). Note also in Eq. (4) that the second term is just the energy of exchanging an interstitial Si and a substitutional impurity.

For defect-mediated pathways, the first task of theory is to determine which defect or complex leads to diffusion with the smallest $Q_i^*$. For all impurities, the formation energies of the impurity-vacancy complexes are calculated with appropriate relaxations, etc. However, because a supercell beyond present computer capacity is required to obtain the migration energy, we have taken these values from experiment (for P, As, and Sb) (Ref. 30) or used a simple estimate (for B). See Table I for the actual values used. For vacancy-mediated ($V$-mediated) B diffusion, we find that there is a relatively small binding energy between the impurity and the defect, where the binding energy of an impurity-defect complex is defined as the difference between the formation energy of the defect in pure Si and the formation energy of the impurity-defect complex (see Table I). In particular, we note that the binding energy of the B$V$ pair is smaller than our estimated migration energy. Given the error bars of our calculation, we cannot discern between a simple vacancy mechanism or a B$V$-pair mechanism. For all other impurities, however, the relevant diffusing species is the impurity-vacancy pair.

For interstitial-mediated pathways, a global total-energy surface depicting the interactions between $N$ Si atoms and a single impurity in an $N$-atom supercell was

TABLE I. Calculated activation energies ($Q^*$) under equilibrium conditions for B, P, As, and Sb diffusion as well as Si self-diffusion. Listed also are the separate contributions to $Q^*$ and the binding energies ($E_b$) of impurity-vacancy pairs. All quantities are in eV and all species are in their neutral charge state. The experimental binding energies are from Ref. 15 and the Si CE value is from Ref. 14.

| Species | $Q^*$ | $E_f$ | $E_m$ | $E_b$(theor) | $E_b$(expt) |
|---|---|---|---|---|---|
| $B_i$ | 3.9 | 3.9 | 0.0 | | |
| B$V$ | a | 3.0 | 1.0 | 0.5 | |
| B(CE) | 4.9 | | | | |
| $P_i$ | 3.8 | 3.0 | 0.8 | | |
| P$V$ | 3.4 | 2.5 | 0.94 | 1.0 | 1.04 |
| P(CE) | 4.6 | | | | |
| $As_i$ | 3.6 | 3.2 | 0.4 | | |
| As$V$ | 3.4 | 2.3 | 1.07 | 1.2 | 1.23 |
| As(CE) | 3.9 | | | | |
| $Sb_i$ | 4.9 | 4.7 | 0.2 | | |
| Sb$V$ | 3.6 | 2.3 | 1.28 | 1.2 | 1.44 |
| Sb(CE) | 4.5 | | | | |
| $Si_i$ | 4.0 | 3.6 | 0.4 | | |
| Si$V$ | 3.8 | 3.5 | 0.3 | | |
| Si(CE) | 4.3 | | | | |

[a]Since $E_m > E_b$, the B$V$ pair is not a stable diffusing species. See the text for further explanation.

required in order to ascertain the lowest-energy diffusion pathways and the migrating species. Such a surface results from the collection of the total energies of $N+1$ atoms competing for $N$ substitutional sites with the relaxation of the surrounding host crystal included for each configuration. The details of generating a total-energy surface are described elsewhere.[22] Because the total-energy surface is a function which depends upon the three spatial dimensions, it is easiest to examine a slice through a given crystal plane. The energy surface can be displayed either as a contour plot (the contours depicting constant energies) or as a perspective plot (the in-plane coordinates are the chosen diamond-structure coordinates and the third axis is the energy). We emphasize that for either type of plot, the relaxation of the host crystal is included in obtaining the total energy, but the positions marked as atoms serve only as a template for identifying positions in the crystal plane.

A contour plot of the total-energy surface depicting the diffusion of a neutral B atom through the diamond-structure (110) plane is shown in Fig. 3(a). A perspective plot for the same species is shown in Fig. 3(b). In constructing both figures, the zero of energy was chosen at the saddle point of the B kick-out process [see Fig. 3(a) and below]. In the perspective plot, regions of different energy have been color-coded: the lowest-energy regions are red, follow by blue, with the highest-energy regions in green. The lowest-energy migration pathway for neutral B is along the low-electron-density channels; indeed, there is virtually no barrier to migration from the $H$ to the $T$ site, etc. If, instead of moving along the channel, the B atom continues in a ⟨111⟩ direction towards a substitutional Si atom, a kick-in process is initiated. The impurity climbs up energy contours in the direction of the saddle point. The saddle point for this process is roughly two-thirds of the way from the $T$ site to the substitutional site. Although it is not shown in Figs. 3(a) and 3(b), the relaxation of the two Si atoms along the [111] direction ahead of the B atom is accounted for in constructing the total-energy surface, as are relaxations of the second neighbors. The energy barrier to the kick-in process via this pathway is $\approx 1$ eV. Once the B atom has passed the kick-in saddle point, its energy decreases towards the substitutional site, while the furthest Si atom is kicked out into the channel. The reverse of the kick-in process depicted in the total-energy surfaces is the kick-out process, and it represents the lowest-energy mechanism by which substitutional B becomes an interstitial. In sum, the diffusing species is thus the interstitial B atom. We find qualitatively similar results for all dopants studied; the diffusing species for interstitial-mediated pathways is always the interstitial impurity which is created by the kick-out process.

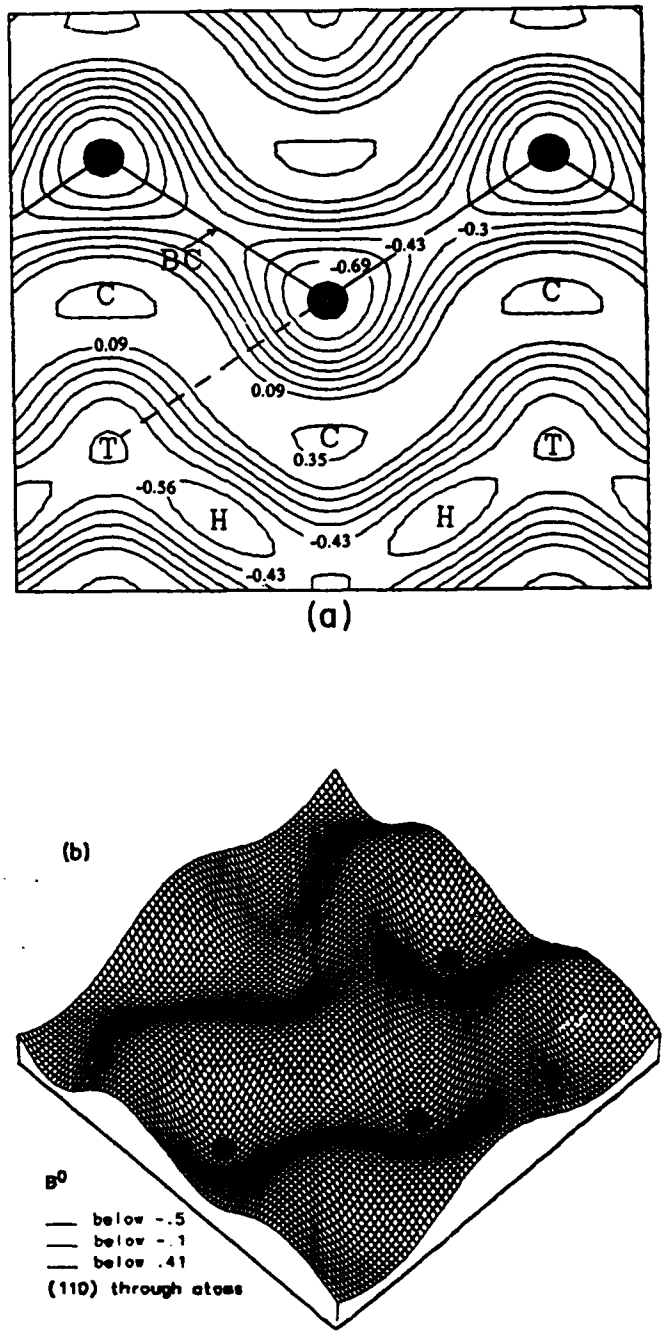



FIG. 3. Total-energy surface plots. (a) Total-energy contour plot depicting the migration of a neutral B interstitial through the Si crystal. The labeled sites are $T$ (tetrahedral), $H$ (hexagonal), BC (bond-center) and $C$ (at the center of a rhombus formed by three adjacent Si atoms and the nearest $T$). The energy difference between contours is 0.13 eV. The dashed line is the kick-out pathway. The values of the contours near the channel regions are $\approx 0.2$ eV higher than those reported in our previous publication (Ref. 16). This is a consequence of generating the surface with a higher plane-wave cutoff and does not change any of our conclusions based on that previous figure. (b) Perspective plot of the same process. The areas colored red are lowest in energy, blue are intermediate, followed by the highest-energy regions in green. Relaxations of the host atoms are not indicated in the figure, but are taken into account in the total-energy calculations.

### B. Results and discussion

The calculated activation energies for CE-, $V$-, and $I$-mediated mechanisms for substitutional B, P, As, and Sb diffusion under equilibrium conditions are shown graphically in Fig. 4. A selected range of experimental values is shown as boxed areas in Fig. 4 as well. Actual values for

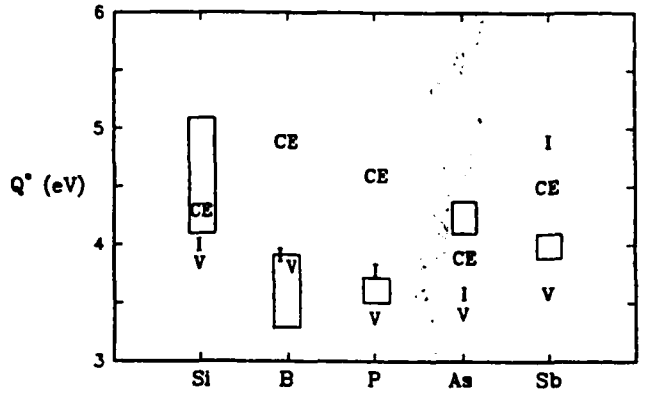



FIG. 4. The calculated activation energies under equilibrium conditions for vacancy-mediated ($V$), interstitial-mediated ($I$), and concerted exchange (CE) mechanisms for Si self-diffusion and various impurities. The boxed areas are a selected range of experimental results from Ref. 20.

$Q_i^*$ and the contributions to it are given in Table I. For comparison, the calculated values for Si self-diffusion are also shown (the CE value is that reported in Ref. 14). The calculated values presented in Fig. 4 and Table I for interstitial-mediated diffusion are those for *impurity* interstitial diffusion, effected by the kick-out mechanism. Calculated activation energies for both extreme binding limits of the coordinated push mechanism are significantly larger than the activation energies for the kick-out mechanism, thereby ruling out the impurity–self-interstitial pair and the self-interstitial as the relevant diffusing species. For B, P, and As, the differences between $V$- and $I$-type mechanisms are smaller than our estimated error bar (0.4 eV is the maximum difference). Furthermore, because the CE activation energies are upper bounds, we view them as comparable to defect-assisted pathways. Only for Sb impurities is one mechanism dominant: the $V$-assisted pathway is a full 1.3 eV lower than the $I$-assisted pathway, so that the vacancy mechanism prevails, in agreement with conclusions drawn from experimental data.[8]

From our calculations we can also obtain the binding energies of the various impurity-vacancy complexes The binding energies of some impurity-vacancy pairs have been measured experimentally[15] and are included in Table I along with the theoretical values. We note for P, As, and Sb that the impurity-vacancy binding energies are slightly larger than the pair-migration energies, suggesting that it is at least energetically feasible for such complexes to migrate. None of these impurities is in the strong-binding limit and we therefore consider the range of activation energies between the simple vacancy mechanism and the impurity-vacancy pair as the error bar for the vacancy-mediated process.[31] However, as pointed out in the preceding subsection, the binding energy of B$V$ pairs is rather small (0.5 eV smaller than our estimated migration energy), so that if B diffuses by a vacancy mechanism, it is probably effected by the isolated vacancy, as opposed to the pair mechanism. We discuss this point more fully in the next section. The relative magnitudes of the binding and migration energies also have important consequences for the electrical deactivation of heavily doped Si.[32]

In summary, our calculated equilibrium activation energies, being in the same range as experimental values, establish the reliability of our calculational approach, but do not establish the relative importance of the various mechanisms, with the exception of Sb. In order to do this within the confines of equilibrium diffusion, calculations of the various preexponentials are required. Reliable entropy calculations are not currently possible, however. Nonetheless, we will demonstrate in the remainder of this paper that, within a suitable framework for nonequilibrium diffusion, theoretical calculations of various barrier heights, etc., combined with experiments involving injection of excess point defects, do allow a number of definitive conclusions to be drawn.

## IV. NONEQUILIBRIUM DIFFUSION

### A. Theory of externally stimulated diffusion

Under equilibrium conditions, the point defects responsible for diffusion are created thermally. However, evidence has accumulated that surface treatments selectively inject point defects into the bulk. Oxidation, for example, injects self-interstitials, while direct nitridation injects vacancies.[5] Under injection conditions, the dopant diffusion coefficient has been observed to be either enhanced or retarded. Such experiments have, however, led to widely conflicting conclusions regarding the dominant diffusion mechanism, largely because of contradictory assumptions. Virtually all attempts at constructing a theory for diffusion under point-defect injection have relied on unclear postulates, mainly with regard to the relevant diffusing species. Most theories assume that the concentrations of vacancies and self-interstitials determine the dopant diffusion coefficient. A review and critique of the early work may be found in Ref. 6. As we showed in Sec. III of the present paper, for P, As, and Sb, it is certainly true that the relevant diffusing species are either the interstitial impurity or the impurity-vacancy pair. For B, the interstitial impurity, the isolated vacancy, or the impurity-vacancy pair may control long-range dopant migration.

A correct theory of nonequilibrium diffusion rests upon the realization that the diffusion coefficient $D$ is still given by a sum of expressions of the form of Eq. (1) with the individual diffusivities the same as under equilibrium. The task, then, of such a theory is to determine the relevant species concentrations, taking account of their creation, annihilation, and interaction with the other species present. In this subsection, we present the details of such a theory under self-interstitial injection (the case for vacancy injection is entirely analogous). In the second part of this subsection, these results are combined with experimental results to assess the contribution of the various mechanisms to the diffusion of dopant impurities.

Nonequilibrium dopant diffusion experiments provide a convenient framework within which the contribution of the CE mechanism can be investigated. The CE diffusion

coefficient is proportional to the fraction of impurities that are substitutional, $C_{X_s}$. Thus, point-defect injection can only affect the CE mechanism through changes in $C_{X_s}$; e.g., through formation of impurity-vacancy pairs or interstitial impurities. However, injection occurs at low levels relative to $C_{X_s}$, so that it affects $C_{X_s}$ and hence $D_{CE}$ only minimally. Therefore, if the CE is the dominant mechanism, the diffusion coefficient should be either unchanged, or, if the injection process is efficient, so that diffusion may proceed by a vacancy- or interstitial-mediated mechanism, then *enhanced* diffusion should be observed. Under no circumstances should diffusion be retarded by moderate rates of injection. This is in contradiction to the experimental observation that B, P, As, and Sb all show significant retardation under one or another defect-species injection.[7,8] The CE mechanism therefore cannot be the dominant impurity diffusion mechanism and we do not discuss it any further in the following.

The task remaining is thus to determine the effects of point-defect injection on point-defect-mediated diffusion. This is accomplished by finding the concentrations of the relevant diffusing species which may be the impurity interstitial, the impurity-vacancy pair, or, for the case of B, the Si vacancy. The concentrations of all species are governed by the following *complete* set of reactions:

$$S \rightleftarrows V , \tag{6}$$

$$S \rightleftarrows I , \tag{7}$$

$$I + X_s \underset{k_1'}{\overset{k_1}{\rightleftarrows}} X_I , \tag{8}$$

$$V + X_s \underset{k_2'}{\overset{k_2}{\rightleftarrows}} XV , \tag{9}$$

$$I + XV \underset{k_3'}{\overset{k_3}{\rightleftarrows}} X_s , \tag{10}$$

$$V + X_I \underset{k_4'}{\overset{k_4}{\rightleftarrows}} X_s , \tag{11}$$

$$I + V \underset{k_5'}{\overset{k_5}{\rightleftarrows}} \square . \tag{12}$$

Reactions (6) and (7) are schematic and represent the independent thermal generation or annihilation of interstitials or vacancies by free or internal surfaces, $S$. Rate constants for the various reactions are denoted $k_1$, $k_1'$, etc. In Eq. (12), $\square$ represents bulk Si. From first-principles calculations,[33] we have found significant barriers for the reverse reactions (10)–(12), indicating that the bulk plays only a small role in supplying point defects.

Following experimental measurements of the dopant diffusion coefficient under point-defect injection, we distinguish two time domains. All data to date show on the time scale of $\lesssim 1$ hour nonconstant diffusion coefficients: $D$ is initially equal to the equilibrium value, attains an extremum value, and finally decreases or increases to some final steady-state value. During the transient period, under injection of either vacancies or self-interstitials, *all* impurities are expected to show enhanced diffusion. This phenomenon is a consequence of the increased concentration of $I$ or $V$ which, because the concentration of $X_s$ is many orders of magnitude larger than that of $XV$ or $X_I$, drives Eqs. (8) or (9), respectively, further to the right, increasing the concentration of the diffusing species. In addition, we have calculated a barrier of $\approx 1$ eV for $I$-$V$ recombination,[33] indicating that the point defects do not readily recombine. Observation of the initial transient behavior initially led Antoniadis and Moskowitz[7] to suggest the existence of such a recombination barrier. From their data, they estimate a barrier of $\approx 1.4$ eV. Enhanced diffusion is observed experimentally[8] for B, P, As, and Sb under interstitial injection, and for As and Sb under vacancy injection on the short-time scale. P under vacancy injection does, however, show consistently retarded diffusion for all times measured. It is not clear why this should be the case.

We now turn to consideration of the diffusion problem under point-defect injection at steady state. From the seven reactions listed above [Eqs. (6)–(12)], four expressions for the time rate of change of the concentrations of $I$, $V$, $XV$, and $X_I$ may be readily obtained:

$$\frac{\partial C_I}{\partial t} = g_{\text{th}}^I - r_{\text{th}}^I C_I + g_{\text{inj}} - k_1 C_I C_{X_s} + k_1' C_{X_I} - k_3 C_I C_{XV} + k_3' C_{X_s} - k_5 C_I C_V + k_5' C_{\text{Si}} = 0 , \tag{13}$$

$$\frac{\partial C_V}{\partial t} = g_{\text{th}}^V - r_{\text{th}}^V C_V - k_2 C_V C_{X_s} + k_2' C_{XV} - k_4 C_V C_{X_I} + k_4' C_{X_s} - k_5 C_I C_V + k_5' C_{\text{Si}} = 0 , \tag{14}$$

$$\frac{\partial C_{XV}}{\partial t} = k_2 C_V C_{X_s} - k_2' C_{XV} - k_3 C_I C_{XV} + k_3' C_{X_s} = 0 , \tag{15}$$

$$\frac{\partial C_{X_I}}{\partial t} = k_1 C_I C_{X_s} - k_1' C_{X_I} - k_4 C_V C_{X_I} + k_4' C_{X_s} = 0 , \tag{16}$$

where $g_{\text{th}}^I$ ($g_{\text{th}}^V$) is a thermal surface generation rate for $I$ ($V$), $r_{\text{th}}^I$ ($r_{\text{th}}^V$) is a thermal surface recombination frequency for $I$ ($V$), $g_{\text{inj}}$ is the surface injection rate for interstitials, $C_{\text{Si}}$ is the concentration of Si lattice sites, and all other terms are as defined above. We have excluded any spatial dependence in the concentrations for reasons of simplification. This amounts to assuming that the impurity profile is flat in the region of interest. The resulting set of four equations is a complete and exact set, which can, in principle, be solved for all the relevant concentrations. This result is singular and contrasts strikingly with previous work wherein differing sets of equations were written down which were either incomplete or taken as self-evident.[6]

In practice, a series of approximations are required be-

fore analytical results can be obtained. For each species, the dominant generation and recombination term or terms are identified and are then used to solve for the concentrations. The general guiding principle in determining the dominant terms is based on a knowledge of the relative barrier heights as found from our first-principles calculations. For example, our calculations show that bulk generation of $I$ and $V$ defects through Frenkel-pair formation requires at least 8 eV, making this an unlikely source of either point defect. We extend this result to the reverse of reactions (10) and (11). Furthermore, because the binding energies of $XV$ or $X_s$-$I$ pairs are relatively small ($\approx 1$ eV, see Table I), we assume that Eqs. (8) and (9) are in local equilibrium. Finally, for the case of self-interstitial injection, the annihilation of $X_I$ by vacancies (the minority species) is likely to be a second-order effect, and thus we ignore it.

From these general considerations, the point-defect concentrations are

$$C_I = \frac{g_{\rm th}^I + g_{\rm inj}}{r_{\rm th}^I} = C_I^* + C_I' , \tag{17}$$

$$\frac{1}{C_V} = \frac{r_{\rm th}^V + k_3 C_I}{g_{\rm th}^V} = \frac{1}{C_V^*} + \frac{k_3 (C_I^* + C_I')}{g_{\rm th}^V} . \tag{18}$$

The asterisks denote equilibrium quantities, while the primes denote just that component due to the nonequilibrium process. We note in passing that the often-quoted relationship

$$C_I^* C_V^* = C_I C_V \tag{19}$$

used in the analysis of injection experiments *does not hold* for the general case under consideration here. Hu has discussed the shortcomings and fallacies associated with the assumption of Eq. (19) more extensively, and we refer the interested reader to that article.[6]

Using Eqs. (17) and (18), and the general considerations outlined above for choosing the dominant terms for generation and recombination, we find, for the nonequilibrium concentrations of $C_{X_i}$ and $C_{XV}$,

$$C_{X_i} = C_{X_i}^* + \frac{k_1 C_{X_s} C_I'}{k_1'} , \tag{20}$$

$$\frac{1}{C_{XV}} = \left[ 1 + \frac{k_3 (C_I^* + C_I')}{r_{\rm th}^V} \right] \left[ \frac{1}{C_{XV}^*} + \frac{k_3 C_I'}{k_2 C_{X_s} C_V^*} \right] . \tag{21}$$

The species whose concentration is enhanced by interstitial injection, $C_I$ or $C_{X_i}$, exhibit a simple additive dependence on the injected species. On the other hand, the species which may be annihilated by interstitial injection, $C_V$ or $C_{XV}$, show an inverse dependence on the interstitial concentration. We will return to this point later.

Combining Eqs. (1), (20), and (21), the total diffusion coefficient is of the form

$$D = D_I + D_V , \tag{22}$$

where

$$D_I = D_I^* + D_I' \tag{23}$$

and

$$\frac{1}{D_V} = \frac{1}{D_V^*} + \frac{1}{D_V'} , \tag{24}$$

where the subscript $I$ ($V$) denotes the $I$- ($V$-) assisted component of diffusion. In principle, each component may consist of several terms. For example, B diffusion assisted by vacancies may be mediated by either the isolated vacancy, the B-$V$ pair, or by both. However, for all other impurities studied here, the $I$- or $V$-assisted mechanism is mediated by a single species.

Activation energies for the nonequilibrium part of diffusion may be obtained by combining Eqs. (1) and (2) with the relevant species concentration. The latter quantity is expressed in terms of the various barrier heights, etc. Specifically, $D_I'$ is activated with an activation energy

$$Q_I' = E_{\rm inj} - \Delta E + E_m , \tag{25}$$

where $E_{\rm inj}$ is the activation energy of the interstitial injection process, $\Delta E$ is the energy difference between $X_i$ and $X_s$-$I$ as in reaction (8), and $E_m$ is the migration energy of $X_i$. The first term follows from experiments performed by Hu[5] in which it was observed that interstitial injection results in the growth of stacking faults and that this process is activated. We infer that the interstitial injection process is itself also activated and therefore the $I$ concentration is given by

$$C_I' \propto e^{-E_{\rm inj}/k_B T} . \tag{26}$$

The second term follows from

$$\frac{k_1}{k_1'} \propto e^{\Delta E/k_B T} . \tag{27}$$

In the limit that the vacancy concentration is unperturbed from its equilibrium value, then Eq. (21) may be simplified to

$$\frac{1}{C_{XV}} = \frac{1}{C_{XV}^*} + \frac{k_3 C_I'}{k_2 C_{X_s} C_V^*} . \tag{21'}$$

From Eq. (21'), $D_V'$ is activated with an activation energy

$$Q_V' = (E_{\rm inj} + E_{I\text{-}XV}^{\rm barrier}) - (E_f^V + E_{V\text{-}X_s}^{\rm barrier}) + E_m , \tag{28}$$

where $E_{I\text{-}XV}^{\rm barrier}$ ($E_{V\text{-}X_s}^{\rm barrier}$) is the energy barrier to recombination of interstitials (vacancies) with $XV$ pairs ($X_s$), $E_f^V$ is the thermal formation energy of vacancies, and $E_m$ is the $XV$ migration energy. The various terms in the activation energy arise as above or from the following expressions:

$$k_3 \propto e^{E_{I\text{-}XV}^{\rm barrier}/k_B T} , \tag{29}$$

$$k_2 \propto e^{E_{V\text{-}X_i}^{\rm barrier}/k_B T} , \tag{30}$$

and

$$C_V^* \propto e^{E_f^V/k_B T} . \tag{31}$$

Equation (28) has a simple physical interpretation. The first term in parentheses is the energy required to annihilate the diffusing $XV$ pairs, while the second term represents the energy required to produce the pairs initially. The overall activation energy is the competing cost of these two terms, in addition to the migration energy.

Diffusion coefficients measured under point-defect injection at several temperatures may appear misleadingly complex. From Eqs. (23) and (24), both diffusion coefficients can display non-Arrhenius behavior which obscures determination of an activation energy. To extract meaningful information from temperature-dependent data, it is necessary to isolate the equilibrium and nonequilibrium contributions to the total diffusion coefficient. If diffusion is mediated by self-interstitials, then under interstitial injection the nonequilibrium diffusion coefficient ($D_I'$) is obtained by subtracting the (known) equilibrium diffusion coefficient ($D_I^*$) from the total diffusion coefficient ($D_I$). The two possible resulting forms for $D_I'$ are shown schematically in Figs. 5(a) and 5(b). Such figures, which assume Arrhenius behavior for the nonequilibrium diffusion contribution, clearly assume only one species contributes to interstitial-mediated diffusion.

On the other hand, if diffusion is mediated by vacancies, then under interstitial injection the diffusion coefficient displays an inverse behavior [Eq. (24)]. This dependence naturally suggests introducing an *inverse* Arrhenius plot of $1/D_V$ versus $1/T$. The nonequilibrium diffusion coefficient ($D_V'$) is thus determined from rearranging Eq. (24),

$$\frac{1}{D_V'} = \frac{1}{D_V} - \frac{1}{D_V^*} . \tag{24'}$$

The two possible forms for measured diffusion coefficients are shown in Figs. 6(a) and 6(b). From any of these four plots, Figs. 5(a), 5(b), 6(a), and 6(b), the contribution due solely to the nonequilibrium process can be isolated and, hence, a corresponding activation energy that can be compared to either Eq. (25) or (28) may be found.

### B. Results and discussion

We now turn to a discussion of three distinct cases. Consider the instance in which, under equilibrium, the $I$

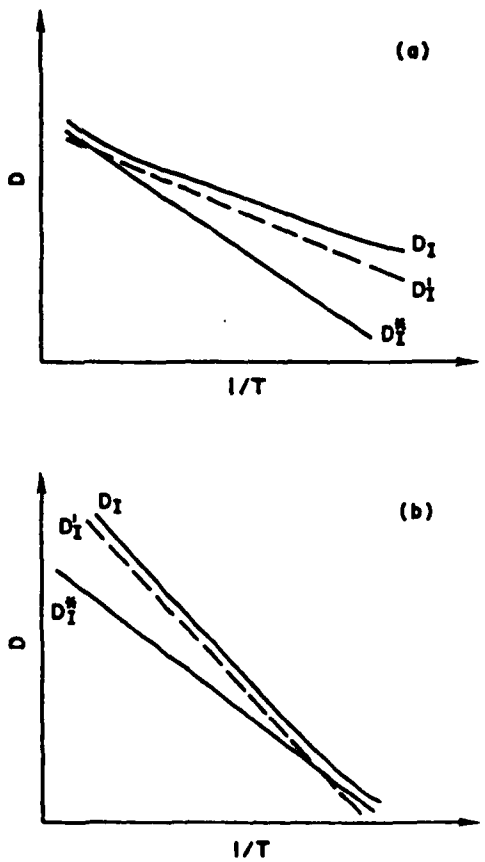



FIG. 5. (a) and (b) Arrhenius plot for interstitial-mediated diffusion under interstitial injection. $D_I^*$ denotes the equilibrium diffusion coefficient, $D_I$ the total diffusion coefficient under nonequilibrium conditions, and the dashed line is the nonequilibrium contribution ($D_I'$)

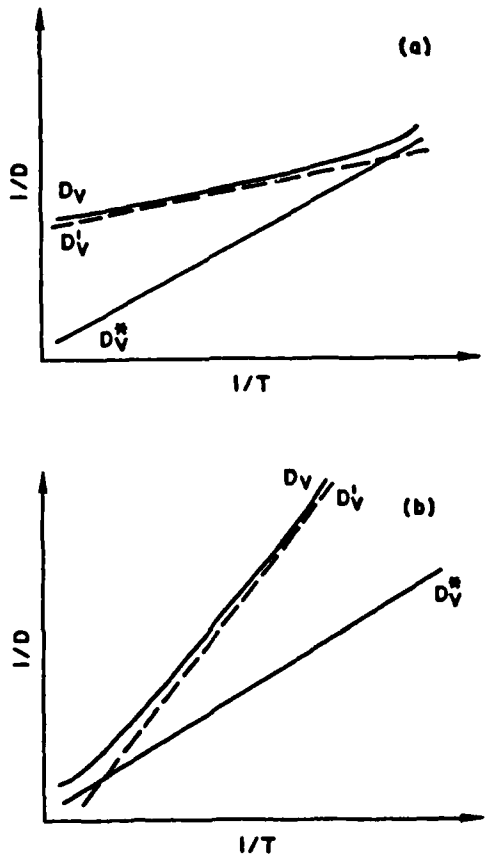



FIG. 6. (a) and (b) Reciprocal Arrhenius plot for vacancy-mediated diffusion under interstitial injection. $D_V^*$ denotes the equilibrium diffusion coefficient, $D_V$ the total diffusion coefficient under nonequilibrium conditions, and the dashed line is the nonequilibrium contribution ($D_V'$).

TABLE II. Activation energies ($Q'$) for diffusion mediated exclusively by interstitials under interstitial injection (theory) and measured activation energies under oxidation conditions (experiment from Hill, Ref. 34). All quantities are in eV.

| Species | $Q'$(theor) | $Q'$(expt) |
|---|---|---|
| B | 2.6 | 2.3 |
| P | 2.5 | 2.4 |
| As | 2.3 | 2.3 |

component is dominant. Interstitial injection leads to enhanced diffusion of the form

$$D = D^* + D' , \tag{32}$$

where the activation energy of $D'$ is as given in Eq. (25). The measured diffusion coefficient may have either of the forms depicted in Fig. 5(a) or 5(b). Hill[34] has measured the diffusion coefficients of B, P, and As under interstitial injection and found that they indeed obey Eq. (32) and are similar to that depicted in Fig. 5(a). This observation suggests therefore that these impurities diffuse principally by an interstitial mechanism. In order to further test this possibility, we have calculated the activation energy for these impurities under interstitial injection [Eq. (25)] using our calculated values for $\Delta E$ and $E_m$ and a value for the interstitial injection energy, $E_{\text{inj}}$, extracted from the data of Ref. 35. These results, along with Hill's experimental values, are given in Table II. The excellent agreement between the two sets of values corroborates the conclusion that these impurities diffuse primarily assisted by interstitials.

Using a damaged layer created by Ar-ion implantation as the interstitial source, Bronner and Plummer[36] observed P diffusion enhancement over a limited temperature range. They measured a dependence for the diffusion coefficient very similar to Fig. 5(a), as was found by Hill, but did not take data at temperatures high enough to discern any curvature in the Arrhenius plot. Furthermore, they did not determine $E_{\text{inj}}$. But, using their measured activation energy for the total diffusion coefficient, 1.3 eV, we may infer a value for $E_{\text{inj}}$. Because the total diffusion-coefficient curve and the nonequilibrium contribution to it are essentially parallel over the limited temperature range investigated, solving Eq. (25) for $E_{\text{inj}}$ and inserting our theoretical values, we find $E_{\text{inj}} = 1.1$ eV, a value which can be tested experimentally.

We noted in the preceding subsection on equilibrium diffusion that the B$V$ pair has a rather small binding energy compared to its migration energy. This means that it is more likely that the isolated vacancy, rather than the pair, effects long-range migration of B, if indeed a vacancy mechanism is appropriate. In turn, then, the diffusion coefficient of B is determined not by the B$V$ concentration, but by the isolated $V$ concentration [Eq. (18)]. Under interstitial injection, the isolated vacancy concentration is less than or equal to the equilibrium concentration, so that either no change in the diffusion coefficient is observed or diffusion is retarded. This is clearly in contradiction to Hill's experiments, which show enhancement. We cite this finding as further proof that B migration is mediated predominantly by Si self-interstitials.

If equilibrium diffusion is mediated primarily by vacancies, interstitial injection can lead to either enhanced or retarded diffusion depending upon the level of injection. At low injection levels, which may be achieved by oxidation at moderate temperatures, the majority $V$ component is retarded according to Eq. (24), while the $I$ component remains small, with a diffusion coefficient given by Eq. (23). There are a number of data which support the contention that Sb diffusion does indeed exhibit retardation under interstitial injection.[8] Based on the theory presented in the preceding subsection, we predict that temperature-dependent data would obey Eq. (24), so that a reciprocal Arrhenius plot would be required in order to find an activation energy for $D_V'$ to be compared with Eq. (28). The measured diffusion coefficient under such conditions would have a form schematically similar to either Fig. 6(a) or 6(b). Such experiments, which have not been performed to date, would provide a test of our theory and furnish a basis by which the conclusion that Sb diffusion is vacancy dominated can be assessed. At high interstitial injection levels, the interstitial contribution will ultimately overwhelm all other terms, and enhanced diffusion with an activation energy given by Eq. (25) would be observed. However, data are usually reported at a single temperature, so that no crossover from retardation to enhancement has been observed.

Lastly, if both the $I$ and $V$ components under equilibrium conditions are comparable, plots of the diffusion coefficient under interstitial injection would be rather complex. Enhanced *or* retarded diffusion could, according to Eqs. (23) and (24), be observed. However, examining the various possible combinations of Figs. 5(a) and 5(b) with Figs. 6(a) and 6(b) yields, under certain conditions, a unique diffusion coefficient which *crosses* the corresponding equilibrium curve. This is depicted schematically in Fig. 7. Such a crossing provides a definitive experimental signature that the $I$ and $V$ components contribute with comparable magnitudes. Temperature-

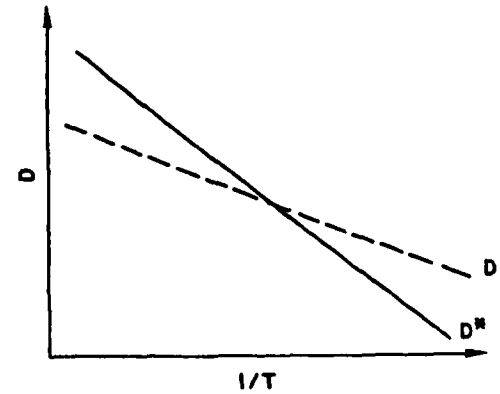



FIG. 7. Predicted schematic form for the diffusion coefficient if, under equilibrium, the $I$ and $V$ components are comparable.

dependent data of diffusion under point-defect injection, which may be utilized to search for this crossing, are currently not available.

## V. CONCLUSIONS

In this paper we have developed a consistent framework for understanding dopant diffusion in semiconductors on a microscopic level. This framework provides important insight into the dynamics of diffusion both energetically (through expressions or calculated values for the activation energy) and pictorially (through the use of a total-energy surface). The expressions for the activation energies under various assumptions regarding the diffusion mechanism and the injected species of point defect also provide a readily accessible link between first-principles theoretical calculations and experiment.

Specifically, we conclude that B, P, and As diffusion is mediated predominantly by interstitials, whereas Sb diffusion is mediated primarily by vacancies. In large part, we confirm the conclusions of Fahey and coworkers[8] with respect to the point-defect mechanism. However, we contradict the finding of Mathiot and Pfister[7] with regard to P diffusion. The discrepancy between the two conclusions can be traced to the fact that, due to the symmetry of the relevant equations, the simulations of Mathiot and Pfister could only discern that one of the two point defects is dominant. Their choice of vacancies over interstitials was based solely on comparisons of impurity diffusion with conclusions regarding self-diffusion which exhibit the same ambiguity. In other words, the equations governing self-diffusion are symmetric with respect to the choice of vacancies or interstitials and this ambiguity propagates to the choice of a dominant defect for impurity diffusion.

The methodology presented here is by no means limited to the study of dopants in Si, but can foreseeably be applied to other impurities in semiconductors or even in metallic systems. Furthermore, the expressions developed for the activation energies can be used to probe the various external processing conditions themselves.

## ACKNOWLEDGMENTS

One of us (C.S.N.) would like to acknowledge many helpful conversations with P. M. Fahey and S. M. Hu. This work was supported in part by U.S. Office of Naval Research (ONR) Contract No. N00014-84-C-0396.

---

*Present address: Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, NY 10510.

[1]Fick's laws were first written down in 1855. A Fick, Pogg. Ann. **94**, 59 (1855). The first application to diffusion in solids seems to date from H. Braune, Z. Phys. Chem. **110**, 147 (1924).

[2]See, for example, *Atomic Diffusion in Semiconductors*, edited by D. Shaw (Plenum, London, 1973), for further details concerning the basic principles of diffusion.

[3]See, for example, W. Frank, U. Gösele, H. Mehrer, and A. Seeger, in *Diffusion in Crystalline Solids*, edited by G. E. Murch and A. S. Nowick (Academic, Orlando, FL, 1984), p. 63.

[4]The literature documenting the effects of surface processing conditions on the bulk point-defect concentration is extensive. See, for example, S. M. Hu, Appl. Phys. Lett. **51**, 308 (1987), and the appropriate references therein.

[5]These conclusions are based on growth or shrinkage of extrinsic stacking faults. See, for example, S. M. Hu, J. Appl. Phys. **45**, 1567 (1974); S. Mizuo, T. Kusaka, A. Shintani, M. Nanba, and H. Higuchi, *ibid.* **54**, 3860 (1983); Y. Hayafuji, K. Kajiwara, and S. Usui, *ibid.* **53**, 8639 (1982).

[6]S. M. Hu, J. Appl. Phys. **57**, 1069 (1985).

[7]D. A. Antoniadis and I. Moskowitz, J. Appl. Phys. **53**, 6788 (1982).

[8]P. Fahey, G. Barbuscia, M. Moslehi, and R. W. Dutton, Appl. Phys. Lett. **46**, 784 (1985).

[9]B. J. Mulvaney and W. B. Richardson, Appl. Phys. Lett. **51**, 1439 (1987), and references therein.

[10]F. F. Morehead and R. F. Lever, Appl. Phys. Lett. **48**, 151 (1986).

[11]D. Mathiot and J. C. Pfister, J. Appl. Phys. **55**, 3518 (1984).

[12]G. A. Baraff, M. Schlüter, and G. Allan, Phys. Rev. B **50**, 739 (1983).

[13]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. **54**, 360 (1985).

[14]K. C. Pandey, Phys. Rev. Lett. **57**, 2287 (1986).

[15]P. M. Fahey, P. B. Griffin, and J. D. Plummer, Rev. Mod. Phys. **61**, 289 (1989).

[16]C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. **62**, 1049 (1989).

[17]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965). The exchange and correlation potentials are based on work by D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980), as parametrized by J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[18]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[19]J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).

[20]Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B **30**, 1844 (1984).

[21]W. Pickett, Computer Phys. Rep. (to be published).

[22]C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B **39**, 10 791 (1989).

[23]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. B **39**, 10 809 (1989).

[24]G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 433 (1982).

[25]P.-O. Löwdin, J. Chem. Phys. **19**, 1396 (1951).

[26]H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[27]F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry* (Wiley, New York, 1980).

[28]A. Erbil, W. Weber, G. S. Cargill, and R. F. Boehme, Phys. Rev. B **34**, 1392 (1986).

[29]M. Scheffler, Physica B+C **146B**, 176 (1987).

[30]M. Hirata, M. Hirata, and H. Saito, J. Phys. Soc. Jpn. **27**, 405

(1969).
[31]See the article by A. D. Le Claire, in *Physical Chemistry*, edited by W. Jost (Academic, New York, 1970), Vol. X, p. 261, for a thorough discussion of correlation effects (as manifested by the binding strength between an impurity and a vacancy) in diffusion, especially with regards to impurity diffusion mediated by vacancies.
[32]S. T. Pantelides and C. S. Nichols (unpublished).
[33]C. S. Nichols and S. T. Pantelides (unpublished).
[34]C. Hill, in *Semiconductor Silicon 1981*, edited by H. R. Huff, J. R. Kriegler, and Y. Takeishi (Electrochemical Society, New York, 1981), p. 988.
[35]S. M. Hu, Appl. Phys. Lett. **27**, 165 (1975).
[36]G. B. Bronner and J. D. Plummer, J. Appl. Phys. **61**, 5286 (1987).

# Structure and Properties of Hydrogen-Impurity Pairs in Elemental Semiconductors

P. J. H. Denteneer,[a] C. G. Van de Walle,[b] and S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 3 January 1989)

A variety of experiments have revealed several puzzling properties of hydrogen-impurity pairs. For example, H atoms passivate the electrical activity of some impurities, whereas they induce electrical activity in others; they appear to tunnel around some impurities but not around others. We report first-principles pseudopotential-density-functional calculations for several hydrogen-impurity complexes and unravel the origins and intricacies of the rich behavior of H bound to different substitutional impurities in Si and Ge.

PACS numbers: 61.70.Bv, 66.30.Jt, 71.55.Ht

Over the years experimental observations have unveiled a very diverse role for hydrogen atoms in semiconductors containing impurities. In virtually all cases, H atoms are found to form pairs with substitutional impurities but their effect on electrical activity has been puzzling.[1-7] In some cases, as for example substitutional boron or phosphorus in Si, H passivates the electrical activity of the impurity.[4-7] In other cases, as for example substitutional Si in Ge, H converts a normally inactive impurity into a shallow acceptor.[2,3] Alternatively, this amphoteric effect of H on the electrical activity of impurities can be described[3] by stating that sometimes the complex behaves as a substitutional atom that lies one column to the *left* of the impurity in the Periodic Table, e.g., the (H,Si) complex in ultrapure Ge, whereas in other cases the complex behaves as a substitutional atom on column to the *right* of the impurity in the Periodic Table, e.g., the (H,B) complex in Si. Suggestions for the origins of this unusual behavior have been made on the basis of semiempirical calculations,[8] but the conclusions were only tentative.

A second question that has been debated extensively over the years is whether H is tunneling around the impurity as opposed to occupying a particular site close to the impurity. For example, certain experimental evidence led to the belief that H tunnels around Si and C in Ge,[2] but subsequent experiments showed that a static model with trigonal symmetry was more appropriate for the acceptor complexes.[3] In contrast, Muro and Sievers[9] found evidence of tunneling hydrogen in the hydrogen-beryllium acceptor complex in Si. The experimental findings were satisfactorily accounted for by the dynamic tunneling model of Ref. 2. On the other hand, there is no evidence that H tunnels around Be in Ge. No theoretical understanding of the conditions that favor tunneling is available.

A third question that attracted considerable attention is the specific atomic configuration of H-impurity pairs. Most of the attention so far has focused on the (H,B) pair in Si. A large number of theoretical calculations has been reported contrasting the properties of only a few configurations.[10-13] Though the configuration having H in one of the four Si—B bonds is favored on the basis of total-energy calculations, the results are not definitive because no search has been made for the global total-energy minimum with full relaxation of the host crystal. Also, it is generally believed that the (H,Be) complex in Si consists of an H atom tunneling around Be between four equivalent antibonding (*AB*) sites on the extension of Si—Be bonds. There is no experimental or theoretical evidence, however, that establishes this over other possible paths.

All of the above questions regarding the interaction of H with substitutional impurities in semiconductors can be addressed *simultaneously* by calculating the total-energy surfaces for an H atom around each specific impurity and by a concomitant examination of the corresponding energy levels in the energy gap. In this Letter we report the results of such a study for three qualitatively different hydrogen-impurity complexes. The main conclusions are as follows: Acceptor impurities such as B or Be bind an H atom rather strongly at several symmetrically equivalent sites in their immediate vicinity. Barriers for H motion around the impurity between such sites are small by comparison with the binding energy, so that motion around the impurity can occur either thermally or quantum mechanically (tunneling), depending on subtle differences between the complexes. In contrast, isovalent impurities, such as Si in Ge, bind an H atom very weakly, and the barrier for possible motion around the impurity is significantly larger than the binding energy so that the resulting pairs are static. The effect of H on the electrical activity of the impurity in each case follows naturally from the bonding properties of the complexes.

The calculations are carried out using the first-principles pseudopotential-density-functional method. The method is well documented[14] and has been shown to accurately reproduce and predict ground-state properties of semiconductors. Its successful application to defects and defect complexes is documented in Refs. 15–17. We use periodically repeated supercells to describe the host crystal (including the substitutional impurity) in which H resides. In order to include all relevant relaxations of

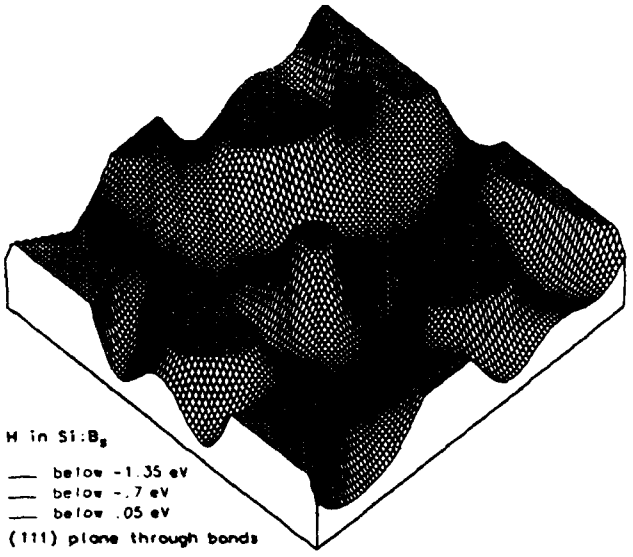



FIG. 1. Energy surface for an H atom in the (111) plane through three bond-minima (*BM*) positions in Si:$B_s$. The plane does not contain atoms, but the unrelaxed lattice position of the B atom is located just 0.4 Å outside the plane in the center of the red ring. The contours are color coded in three different ranges for presentation purposes. For clarity, the surface is cut off at an energy value of 0.05 eV, resulting in the green plateaus. The zero of energy is chosen at the tetrahedral interstitial site.

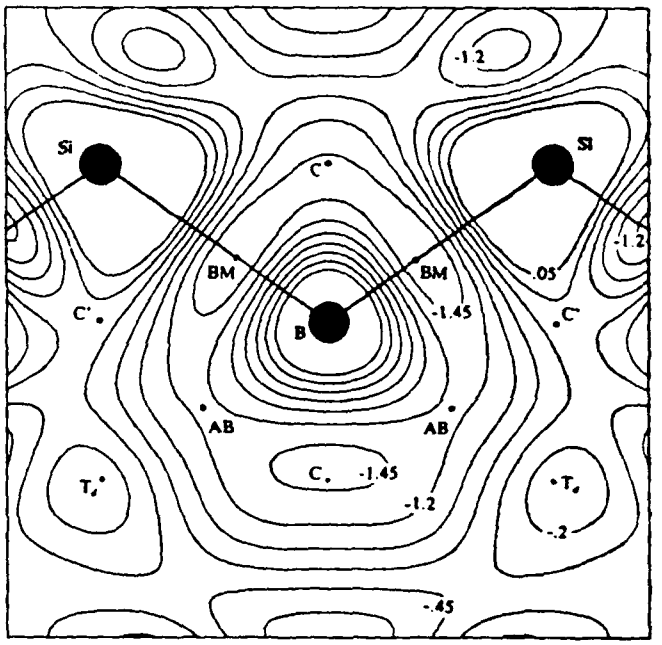



FIG. 2. Contour plot of the energy surface for an H atom in the (110) plane in Si:$B_s$. Big dots indicate (unrelaxed) atomic positions; bonded atoms are connected by solid lines. The substitutional boron atom occupies the center of the plot. Positions of special interest are indicated (see text). Sites denoted *C* and *C'* are equivalent if the B atom in the middle is replaced by a Si atom. The unit of energy is eV and the spacing between contours is 0.25 eV. Close to the atoms contours are not shown above an energy value of 0.05 eV. All relevant relaxations are taken into account to calculate total energies, but the relaxations of the host-crystal atoms are not shown in the figure because they are different for different positions of H.

the host crystal for all of the H positions considered it is necessary to use supercells of up to 32 atoms.[18] We find that most properties of the complexes are described accurately when we use expansions of the wave functions in plane waves with kinetic energy up to 12 Ry.[19] In order to calculate energy barriers with an accuracy of $\lesssim 0.1$ eV, kinetic energy cutoffs of up to 20 Ry in 32-atom cells are used. Two to four special **k** points (depending on the symmetry of the H position) are used to integrate over the first Brillouin zone of the 32-atom cell, which is found to induce negligible error bars on calculated energy differences. Complete energy surfaces for an H atom in the neighborhood of a substitutional impurity in either Si or Ge are obtained by making use of the symmetry of the host crystal.[16,18]

The main result of our calculations is that both (H,B) and (H,Be) in Si exhibit a low-energy shell around the impurity, primarily going through sites close to the center of a Si-impurity bond (bond minimum or *BM* site) and sites labeled *C* (midway between any two of the impurity's nearest neighbors). The low-energy shell is clearly visible as a ring in the total-energy surface for an H atom in the (111) plane shown in Fig. 1. In the contour plot of Fig. 2 for H in the (110) plane only half a ring containing the *BM* and *C* sites is visible. The lower part of Fig. 2 contains antibonding sites (*AB*), which are clearly saddle points, another *C* site, and the tetrahedral interstitial site ($T_d$), which is a local maximum. The *AB* site is 0.5 eV higher in energy than the *BM* site and can only be mistaken for a minimum if only sites for H along the ⟨111⟩ direction are considered.[11] The result that the *AB* site is a saddle point definitively rules out as the stable site for H.

The energy surface for (H,Be) in Si is qualitatively the same as for (H,B) in Si. In each case a low-energy path through *BM* and *C* sites is available. In the case of (H,B) the *BM* site is the global minimum with a site close to *C* being the saddle point for motion of H, whereas in the case of (H,Be) the roles of *BM* and *C* are reversed. More specifically, for (H,B) the saddle point is 0.2 eV higher in energy than the *BM* site, whereas for (H,Be) the *C* site is 0.1 eV lower than the *BM* site. For (H,Be) the *AB* site is 0.4 eV higher than the *C* site and again a saddle point. In the lowest-energy (*BM*) configuration for (H,B) the Si and B atoms closest to H relax outward by the large amounts of 0.24 and 0.42 Å, respectively. Second-nearest-neighbor relaxations are also significant in this configuration. In contrast, the lowest-energy (*C* site) configuration for the (H,Be) complex only involves a small relaxation of Be of 0.14 Å

away from H.

The *BM* configuration for (H,B) is in agreement with a wealth of experimental observations,[20-23] although sometimes a slightly off-axis position close to the bond center is proposed for H.[22] Also the majority of theoretical calculations appear to agree now on a configuration similar to the *BM* configuration.[10,12,13] Furthermore, our calculated vibrational frequency of the H stretching mode for the *BM* configuration of $1830 \pm 100$ cm$^{-1}$ is in good agreement with the experimental value[21] of 1903 cm$^{-1}$. Similar experimental information for the (H,Be) complex is presently not available, but since all of the features of the microscopic structure of the (H,B) complex are in excellent agreement with experimental observations, we can be confident of our description of the (H,Be) complex.

In contrast to the case of (H,B) and (H,Be) in Si, where we find a low-energy region surrounding the impurity, in the case of (H,Si) in Ge the total-energy surfaces of H in various charge states are virtually identical to the surfaces one obtains in the pure material without low-energy regions restricted to the neighborhood of the impurity.[16] This is to be expected since Si and Ge are very similar. For the three charge states considered (positive, neutral, and negative) the global energy minima for H in Ge:Si, are the bond-centered site for $H^+$ and $H^0$, and a site close to the $T_d$ site (displaced from $T_d$ over 0.2 Å toward Si) for $H^-$. Although Si and Ge are very similar and one would not expect the isovalent impurity Si in Ge to be able to bind H, the (H,Si) complex in Ge has a positive binding energy.[24] The binding energies for the three minimum configurations turn out to be very small, but consistently positive (i.e., the complex is bound); we find $E_b = 20$, 28, and 52 meV for $H^+$, $H^0$, and $H^-$, respectively. Since barriers for movement of H around the Si impurity are much larger than these binding energies (e.g., for $H^-$ there is a saddle point for possible motion of H at the hexagonal interstitial site with a barrier of 0.35 eV), the H atom cannot move around the Si impurity while still being bound. We will return to the question of motion of H around impurities later on in the paper.

Regarding the effect of the H atom on the electrical activity of substitutional impurities, we arrive at the surprising result that in all cases the H-impurity pair has an energy level that is virtually identical to the level of an H atom at the same site *without* the impurity. Whether the impurity is deactivated or activated by H is merely a consequence of the specific site that H occupies near the impurity. In the case of B and Be, H is located in the region close to the impurity (containing *BM* and *C* sites). For such positions the H-related level occurs at midgap.[16] The electron of H drops in the empty acceptor level and reduces the activity of the impurity by one unit: The (H,B) complex is completely inactive and the (H,Be) complex is a single acceptor. For the (H,Si) complex in Ge the influence of H on electrical activity depends on the Fermi-level position, since the Fermi-level position determines which charge state and site are favored. We find that for *p*-type Ge (Fermi level close to the top of the valence bands) $H^+$ is 0.2 eV lower in energy than $H^-$, which is 0.2 eV lower in energy than $H^0$. Therefore, in *p*-type Ge, H acts as a donor, just like in *p*-type Si.[16] As a consequence a (H,Si) complex in *p*-type Ge would behave as a donor (this is, of course, a hypothetic case since H would first pair with the acceptors before pairing with isovalent Si impurities). In *n*-type Ge, $H^-$ close to $T_d$ is the lowest-energy state. In ultrapure Ge, in which (H,Si) complexes have been observed, the Fermi level is effectively located in the middle of the gap. In that case, $H^-$ close to $T_d$ is the lowest-energy state. For a position of H close to $T_d$ an H-related level is found below the top of the valence bands. The level will be doubly occupied leaving a hole in the top of the valence band. Therefore, the (H,Si) complex with H close to the $T_d$ site acts as an acceptor in agreement with the experimental observation in ultrapure Ge.[2]

We now turn to the question of motion of the H atom in H-impurity pairs. As we saw above, in the (H,Si) complex in Ge, H cannot move around the impurity since the binding energy of (H,Si) is much smaller than any barrier H would have to overcome. However, in both the cases of (H,B) and (H,Be) in Si, the H atom is firmly bound with a binding energy of about 1 eV (referenced with respect to a dissociated state of isolated ionized acceptors and neutral H atoms in Si and with the Fermi level close to the top of the valence bands). From the energy surfaces discussed above we already saw that barriers for motion of H around the impurity are small: 0.2 eV for (H,B) and 0.1 eV for (H,Be). Such barriers can easily be overcome when H is moving thermally. Very recently, in experiments using the optical dichroism of the H-B absorption bands under uniaxial stress, an activation energy of 0.19 eV was found for H motion from one *BM* site to another, in agreement with our calculated result.[23]

We now consider the possibility that H would tunnel around the substitutional impurity. Such tunneling may occur because of the small mass of the H atom. The much heavier Si or impurity atoms do not participate in the quantum-mechanical process, and merely define the potential in which the light particle moves. These potential wells should be calculated by keeping the host crystal atoms fixed at the positions they have for the initial lowest-energy configuration. For tunneling to occur, the resulting potential must have two or more identical or similar wells separated by small barriers.[25]

In the case of (H,Be), the global minimum is at the *C* site, with little relaxation of the host. Tunneling between equivalent *C* sites can occur if there is a path that does not require motion of the host atoms and the corresponding barrier is small. Tunneling through the *BM* site is not possible because, with the host atoms frozen, we find a barrier of 2.4 eV. We have, however, identified a tun-

neling path going through an *AB* site, with a barrier of 0.4 eV. An estimate of tunneling frequencies in a one-dimensional model has shown that such a barrier is consistent with the possibility of tunneling in this system.

In the case of (H,B), the global minimum is at the *BM* site, which requires large relaxations of the neighboring B and Si atoms. With H located at one bond center, these relaxations are such that the adjacent bond centers are high in energy and thus do *not* provide a potential well for the H atom to tunnel to. Thus, tunneling between equivalent *BM* sites is not possible.

In conclusion, our theoretical calculations reveal that H occupies different sites when it pairs with different impurities, and that the nature of the site determines both the electrical activity of the pair and the possibility of thermal and quantum-mechanical motion around the impurity.

This work was supported in part by the U.S. Office of Naval Research under Contract No. N00014-84-C-0396. One of us (P.J.H.D.) acknowledges support from IBM Netherlands, N.V.

---

[a]Present address: Physics Department, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands.

[b]Present address: Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, NY 10510.

[1]E. E. Haller, in Proceedings of the Third International Conference on Shallow Impurities in Semiconductors, Linköping, Sweden, 1988 (to be published).

[2]E. E. Haller, B. Joos, and L. M. Falicov, Phys. Rev. B **21**, 4729 (1980).

[3]J. M. Kahn, R. E. McMurray, Jr., E. E. Haller, and L. M. Falicov, Phys. Rev. B **36**, 8001 (1987).

[4]C. T. Sah, J. Y. C. Sun, and J. J. T. Tzou, Appl. Phys. Lett. **43**, 204 (1983).

[5]K. Bergman, M. Stavola, S. J. Pearton, and J. Lopata, Phys. Rev. B **37**, 2770 (1988).

[6]J. I. Pankove, D. E. Carlson, J. E. Berkeyheiser, and R. O. Wance, Phys. Rev. Lett. **51**, 2224 (1983).

[7]N. M. Johnson, Phys. Rev. B **31**, 5525 (1985).

[8]J. Oliva and L. M. Falicov, Phys. Rev. B **28**, 7366 (1983).

[9]K. Muro and A. J. Sievers, Phys. Rev. Lett. **57**, 897 (1986).

[10]G. G. DeLeo and W. B. Fowler, Phys. Rev. B **31**, 6861 (1985).

[11]L. V. C. Assali and J. R. Leite, Phys. Rev. Lett. **55**, 980 (1985); **56**, 403 (1986).

[12]A. Amore Bonapasta, A. Lapiccirella, N. Tomassini, and M. Capizzi, Phys. Rev. B **36**, 6228 (1987).

[13]K. J. Chang and D. J. Chadi, Phys. Rev. Lett. **60**, 1422 (1988).

[14]J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979); P. J. H. Denteneer, Ph. D. thesis, Eindhoven University of Technology, 1987 (unpublished).

[15]Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. Lett. **52**, 1129 (1984).

[16]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **60**, 2761 (1988); C. G. Van de Walle, F. R. McFeely, and S. T. Pantelides, Phys. Rev. Lett. **61**, 1867 (1988).

[17]E. Kaxiras and K. C. Pandey, Phys. Rev. Lett. **61**, 2693 (1988).

[18]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. B (to be published).

[19]Plane waves with kinetic energy up to 6 Ry are included exactly; those with kinetic energy between 6 and 12 Ry, in perturbation theory according to P. O. Löwdin, J. Chem. Phys. **19**, 1396 (1951).

[20]A. D. Marwick, G. S. Oehrlein, and N. M. Johnson, Phys. Rev. B **36**, 4539 (1987).

[21]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Appl. Phys. Lett. **50**, 1086 (1987); Phys. Rev. B **37**, 8313 (1988).

[22]M. Stutzmann, Phys. Rev. B **35**, 5921 (1987); M. Stutzmann and C. P. Herrero, Appl. Phys. Lett. **51**, 1413 (1987); C. P. Herrero and M. Stutzmann, Solid State Commun. (to be published).

[23]M. Stavola, K. Bergman, S. J. Pearton, and J. Lopata, Phys. Rev. Lett. **61**, 2786 (1988); preliminary accounts of these experimental results as well as our calculated results for the (H,B) complex were reported in Proceedings of the Fifteenth International Conference on Defects in Semiconductors, Budapest, Hungary, 1988 (to be published).

[24]The binding energy is defined here as the energy difference between configurations in which H in a specific charge state occupies the same site in Ge:Si, and in pure Ge.

[25]E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1970), p. 74.

(D)

# Theory of hydrogen diffusion and reactions in crystalline silicon

Chris G. Van de Walle,* P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 21 November 1988)

The behavior of hydrogen in crystalline silicon is examined with state-of-the-art theoretical techniques, based on the pseudopotential-density-functional method in a supercell geometry. Stable sites, migration paths, and barriers for different charge states are explored and displayed in total-energy surfaces that provide immediate insight into these properties. The bond-center site is the global minimum for the neutral and positive charge states; in the negative charge state, the tetrahedral interstitial site is preferred. The positive charge state is energetically favorable in *p*-type material, providing a mechanism for passivation of shallow acceptors: electrons from the H atoms annihilate the free holes, and formation of H-acceptor pairs follows compensation. Also addressed are the issues of molecule formation and hydrogen-induced damage. A number of different mechanisms for defect formation are examined; hydrogen-assisted vacancy formation is found to be an exothermic process.

## I. INTRODUCTION

The topic of hydrogen (H) in semiconductors has recently attracted a great deal of interest. From a fundamental point of view, it is attractive to study the interaction between H, the most elementary atom, and silicon (Si), the prototypical semiconductor. The role of H in crystalline semiconductors has also emerged as an important technological problem: its effects have recently most dramatically been observed in the passivation of shallow impurities. In this paper we will concentrate on the behavior of hydrogen itself as it diffuses through a Si crystal. The information about stable sites and charge states obtained here is essential for understanding not only isolated hydrogen, but also its reactions with other impurities.

Hydrogen has been known for a long time to saturate dangling bonds at surfaces, vacancies, and grain boundaries, and to passivate deep-level defects, such as those due to transition-metal impurities.[1] In cases where deep levels are detrimental for device properties, their elimination by hydrogenation is of great benefit. The fact that hydrogen can also passivate shallow impurities has only been appreciated more recently.[1-9] Shallow levels determine the doping of the semiconductor, which determines its characteristics in device operation; accidental passivation of these impurities can cause outright failure of the device. On the other hand, one can envision applications in which intentional passivation of certain areas of a device could be an integral part of the fabrication process. Since hydrogen is present, intentionally or not, during many of the processing steps for fabricating modern semiconductor devices, its potential effects, whether harmful or beneficial, should be thoroughly understood.

There exist indeed a wide variety of ways in which H can penetrate Si, a number of which are discussed in Ref. 10. They include crystal growth, high-temperature permeation,[11] ion implantation,[12] chemomechanical polishing,[13] wet etching, unintentional hydrogenation during ion bombardment, plasma etching, boiling in water, and surface exposure to monatomic H. The latter can be achieved by placing the sample in or downstream from a microwave plasma.[4,14] If the sample is shielded from the plasma, this is the preferred way for introducing H under the best controlled conditions, avoiding any additional damage in the material.

Passivation of the electrical activity of *p*-type silicon was first observed in metal-oxide-semiconductor (MOS) capacitors by Sah *et al.*,[2] who suggested H as the probable cause. Subsequent experiments by Pankove *et al.*[3] and by Johnson[4] unambiguously showed the correlation between H and acceptor profiles, and established the existence of H-acceptor pairs. Initially, passivation of shallow donors was thought to be nonexistent[2,3] or very weak.[5] Recently, however, conclusive evidence has been provided for passivation of samples doped with P, As, and Sb, in which a reduction of up to 80% was observed in carrier concentrations.[6] This passivation, while dramatic, is still not as complete as can be obtained in *p*-type samples.

Apart from its role in interacting with existing defects and impurities, hydrogen has recently been shown to *induce* defects as well.[15] Extended defects (described as "platelets") in the near-surface region were observed after hydrogenation, and correlated with the presence of large concentrations of H.

A number of authors[2-5,7,9] have offered interpretations of the passivation data seeking to unravel the underlying mechanisms. Attempts to explain the observed phenomena led to a number of contradictory assumptions regarding the nature of the charge states of H along its diffusion path, and hence about the H-impurity reactions that can occur. Particular models were advanced for the structure of the hydrogen-impurity complexes that are a result of passivation. The electronic structure of these complexes is such that all impurity levels are removed from the band gap. A complete understanding of the passivation process can only be obtained, however, by considering

the reactions that lead to H-impurity pairing. This requires a description of the behavior of H as an isolated impurity, an aspect that was first stressed in a paper by one of the present authors.[16] On the basis of available data, it was proposed that H has a donor level in the band gap. Accordingly, passivation of $p$-type material is due to compensation, i.e., the electron from the H annihilates a free hole, and $H^+$ is formed. Pairing of the $H^+$ and the negatively charged acceptor then follows compensation. The present calculations will confirm this suggestion, but will also show that the behavior of H is more complex, and depends upon the doping of the host material. Brief accounts of some of the major results of this work have been published elsewhere.[17,18]

Until recently, no experimental observations were available for isolated paramagnetic hydrogen centers. However, a large amount of experimental effort has been devoted to the study of muonium, a pseudoisotope of H. The muon-spin-rotation technique allows the measurement of the hyperfine splitting of muonium, by studying the spin precession in a magnetic field.[19] Two paramagnetic forms of $\mu$ have been observed: the so-called "normal" muonium (Mu), with an isotropic hyperfine interaction, and "anomalous" muonium ($Mu^*$), with trigonal symmetry and a strong anisotropy of its hyperfine tensor. Normal muonium is usually associated with the tetrahedral interstitial site ($T$). Recently, anomalous muonium was shown to be located at the bond center.[20] This site was actually suggested by Cox and Symons,[21] based on chemical arguments. It has to be noted that the muonium lifetime is only 2.2 $\mu$s, and its mass is $\frac{1}{9}$ that of H. Even though electronic properties do not depend on mass or lifetime, the observed behavior of muonium may differ from that of H, and conclusions about muonium do not necessarily apply to H. Nonetheless, we will see that certain of our results are in good agreement with the experimental observations on muonium. An extensive overview of the field of muonium in semiconductors has recently been compiled by Patterson.[19]

There has been one recent report of a paramagnetic hydrogen state, with indirect evidence that it would be associated with the bond center. Gordeev *et al.*[22] observed by ESR a paramagnetic state due to H in Si, called the *AA9* center. They also showed that the characteristics of *AA9* are similar to those of anomalous muonium ($Mu^*$). Since $Mu^*$ is now known to be associated with the bond center[20] (a fact not appreciated in Ref. 22), this provides indirect evidence for the presence of a paramagnetic H state at the bond-center site.

Over the past ten years a number of theoretical studies were carried out that were aimed at determining the location and properties of H in Si. Many of these studies implicitly assumed that H would retain its atomic character in its interactions with bulk Si, i.e., no strong binding to the crystalline network would occur, and H would favor interstitial locations where the interaction with the Si charge density would be minimal. This point of view led to the neglect of relaxation of the network in many of the earlier studies of the location of H in the Si crystal. It is now known, and will emerge very clearly from the present study, that relaxation of the host crystal around the H impurity is an essential feature of the interaction; most of the essential physics is missed when relaxation is not allowed. For instance, the global energy minimum for H in the positive and neutral charge states occurs at the bond-center position, i.e., midway between two Si atoms, provided these atoms are allowed to relaxed outward over a significant distance in order to accommodate the H atom. If no relaxation is allowed, H cannot insert into the bond.

An overview of the literature has been included in a recent review by Patterson.[19] In the following we will not attempt a complete listing, but rather point out some relevant features and deficiencies of previous work. Among the first theoretical investigations were the extended-Hückel-theory cluster calculations of Singh *et al.*[23] Empirical-pseudopotential Green's-function calculations were carried out by Rodriguez *et al.*[24] Mainwood and Stoneham,[25] using the semiempirical Hartree-Fock–based method of complete neglect of differential overlap (CNDO), addressed the possibility of different charge states for the H. This issue was also addressed in the work of Johnson *et al.*,[5] where empirical tight-binding theory was used to derive the stable site for H in pure Si.

We also mention the empirical-pseudopotential supercell calculations of Pickett *et al.*,[26] even though they were carried out for H at the tetrahedral interstitial site in Ge, not Si. Their band structures showed a H-induced deep donor state more than 6 eV below the valence-band maximum. In contrast, recent calculations[27] using *ab initio* norm-conserving pseudopotentials have shown that H at $T$ in Ge induces a level just below the valence-band maximum, very similar to the situation in Si. The erroneous result of Pickett *et al.* can be ascribed to lack of self-consistency, and/or the use of empirical pseudopotentials. Starting from this result, it was argued that a spin-polarized treatment was necessary, which would introduce a shift in the defect level of up to 0.5 Ry, bringing it closer to the gap region. We will show in Sec. II E that this is incorrect, and that spin polarization has only a minor effect on the energy-level structure.

Katayama-Yoshida and Shindo[28] actually carried out spin-density-functional calculations for H at the tetrahedral interstitial site in Si. They found a defect state in the upper part of the band gap. In our calculations (also including spin polarization) this state is close to and just below the top of the valence band. The result of Ref. 28 may be due to an insufficiently converged basis set.

A wide variety of cluster calculations have been applied to the problem. Besides the CNDO listed above, we mention the MNDO (modified neglect of diatomic overlap) method, used by Corbett *et al.*,[29] and minimal-basis-set Hartree-Fock calculations by Sahoo *et al.*[30] More recent cluster calculations have included relaxation of the Si atoms: Estreicher[31] has used the method of partial retention of diatomic differential overlap (PRDDO) and *ab initio* minimal-basis-set Hartree-Fock calculations, and Deák and co-workers[32] have applied the MINDO/3 (modified intermediate neglect of differential overlap) method.

The results obtained from these cluster calculations display wide variations and inconsistencies, which illustrates the inadequacy of many of these methods to treat the problem at hand. The CNDO, MNDO, PRDDO, and MINDO/3 calculations are based on methods taken from quantum chemistry, which were developed to produce good results for molecules. Their application to solid-state problems, in which the semiconductor host is modeled by a cluster, is usually not justified. Very few cluster calculations test for convergence as a function of cluster size, or examine the effect of the termination of the cluster (usually with H atoms) and possible interactions with the defect states. Furthermore, the Hamiltonians used in those calculations contain parameters which are usually fitted to molecular properties. Certain aspects of local bonding may therefore be well reproduced, but there is no guarantee that the specific solid-state aspects of the interaction of the defect with a crystalline environment can be predicted. Systematic studies to investigate these problems in cluster calculations are very rare, and the few accounts that have been published are far from encouraging. For instance, Deák and Snyder[33] concluded that MNDO, CNDO, and MINDO/3 all have serious difficulty in producing the band structure of the host lattice (Si is found to be metallic in most of their cyclic-cluster calculations), and that calculated ground-state properties for defects may be subject to significant errors. One should therefore apply great caution in applying results from such calculations to the analysis of solid-state properties. The cluster calculations of Estreicher[31] seem to have been tested most carefully for some of the potential problems mentioned above.

In contrast with most previous approaches, in this study we have used state-of-the-art theoretical methods which were developed with the explicit purpose of studying a wide variety of properties of solid-state materials.[34] These techniques will be described in Sec. II. Section III contains our results for a single H atom in crystalline Si; they are most clearly displayed in the form of total-energy surfaces, which provide immediate insight into stable sites and low-energy paths for different charge states. We will also explore the stability of the different charge states in intrinsic, *p*-type and *n*-type material. Section IV deals with interactions between several H atoms, including molecule formation and mechanisms for defect formation. Section V contains a brief summary.

## II. METHODS

The calculational procedure used in this work is based on density-functional theory in the local-density approximation[35] (LDA) and *ab initio* norm-conserving pseudopotentials.[36] The total energy is calculated using a momentum-space formalism:[37] wave functions and potentials are expanded in plane-wave basis sets, and integrations over the first Brillouin zone are performed using the special-points algorithm.[38] A thorough description of the theoretical approach can be found in Ref. 39. Here, the properties of different charge states of H in Si are studied in a supercell geometry.[40] We carried out spin-polarized calculations for a number of representative configurations; the major conclusion is that the deviations from the spin-averaged calculations are small. Our results for total energies are most clearly displayed in the form of total-energy surfaces; we have used a novel technique to generate such surfaces, taking the full symmetry of the host crystal into account. We now proceed to describe and analyze each of these features in more detail.

### A. Pseudopotentials

For Si we use a pseudopotential generated according to the Hamann-Schlüter-Chiang scheme,[36] with cutoff radii of 0.99, 1.49, and 1.11 a.u., respectively, for *s*, *p*, and *d* potentials. The *d* potential is generated using an ionized configuration: $s^1p^{0.75}d^{0.25}$. Test calculations carried out with this pseudopotential for Si in the diamond structure yield a theoretical lattice constant of 5.41 Å, and a bulk modulus of 0.94 Mbar (to be compared with the experimental values of 5.43 Å and 0.99 Mbar). At an energy cutoff of (6;12) Ry, at which most of our calculations for lattice relaxations are carried out, the theoretical lattice constant is 5.42 Å. [The notation $(E_1;E_2)$ Ry means that plane waves with kinetic energy up to $E_2$ Ry are included in the expansions of wave functions and potentials; waves with kinetic energy up to $E_1$ Ry are included in an exact diagonalization of the Hamiltonian matrix, while those with kinetic energy between $E_1$ and $E_2$ Ry are included in second-order Löwdin perturbation theory.[41] We always choose $E_2=2E_1$.] These results indicate the reliability of the Si pseudopotential for structural studies.

The band structure produced by this pseudopotential is in satisfactory agreement with experiment, except for the well-documented problem of local-density-functional theory that the band gap is too small. We calculate a conduction-band minimum along the $\Gamma$–$X$ direction, at about 0.8 times the distance to the $X$ point, at 0.48 eV above the top of the valence band. The implications for our defect calculations will be discussed later.

For H we have simply used the Coulomb potential. Test calculations have shown that no gain in convergence properties is obtained by using a pseudopotential, and that the $1/r$ divergence of the Coulomb potential near the core presents no difficulties.

### B. Plane-wave basis set

We have performed extensive tests to establish the convergence as a function of the plane-wave basis set which is used for expanding wave functions and potentials. Typical plots showing convergence of total-energy differences as a function of energy cutoff are shown in Fig. 1. The ordinate shows the energy difference between two reference configurations in an eight-atom cell. (The eight-atom cell size is generally two small for extracting meaningful results, but is adequate for a study of the plane-wave convergence.) Figure 1(a) is for the energy difference between $H^-$ at the $T$ site and $H^-$ at the hexagonal interstitial site ($H$). Figure 1(b) is for the energy difference between $H^-$ at the $T$ site and $H^-$ at the bond-center site (including only first-neighbor relaxation). The negative charge state was chosen to avoid problems with

level occupations. Indeed, the H atom induces a level near the band gap in the band structure, which has to be left unoccupied (half-occupied) in the positive (neutral) charge state. The occupation for the negative charge state is most easily accomplished, since it only requires filling up all levels with two electrons. Charge states are discussed in more detail in Sec. II F. The abscissa of both plots in Fig. 1 is the energy cutoff $E$ used in the plane-wave expansions: plane waves with kinetic energy up to $E/2$ are included in an exact diagonalization of the Hamiltonian matrix, while those with kinetic energy between $E/2$ and $E$ are included in second-order Löwdin perturbation theory.[41]

In Fig. 1(a) the energy difference is converged (within 0.01 eV of its final value) at a cutoff of (9;18) Ry. Figure 1(b), for the bond center, shows that at (9;18) Ry the energy difference is within 0.15 eV of its final value, which is only reached above (18;36) [note the different scale of panels 1(a) and 1(b)]. We thus see that the convergence properties depend on the position of the H atom. In fact, as we will see later, one can distinguish two regions in the Si crystal in which the H impurity shows distinctly different behavior. The first region is that of high electron density, including the bond-center site ($B$), the sites $C$ and $C'$ (at the center of a rhombus formed by three adjacent Si and the nearest $T$), etc. The location of these sites is illustrated in Fig. 2. The second region consists of

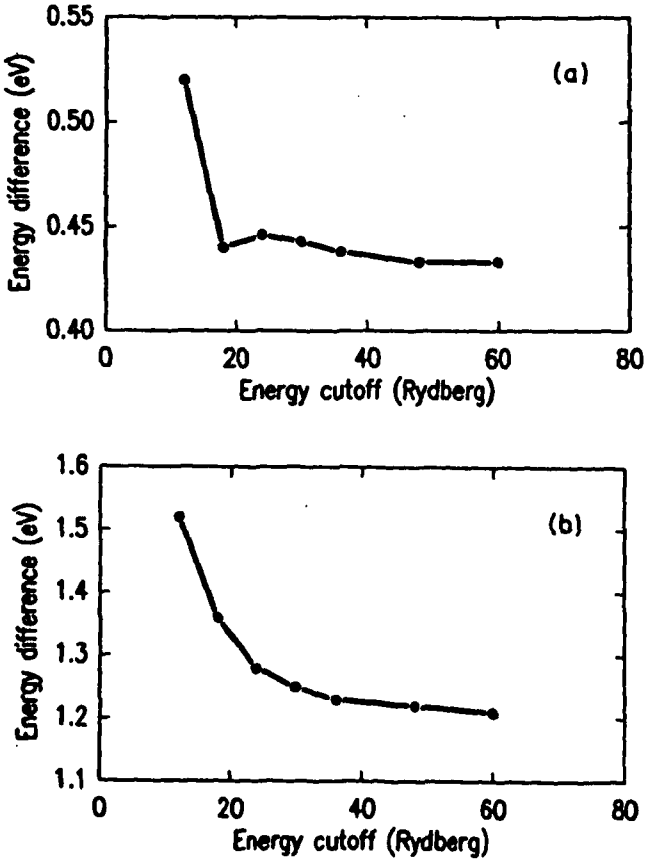



FIG. 1. Convergence of total-energy differences between reference configurations in an eight-atom cell, as a function of energy cutoff. (a) is for the energy difference between $H^-$ at $T$ and at $H$. (b) is for the energy difference between $H^-$ at $T$ and at $B$.

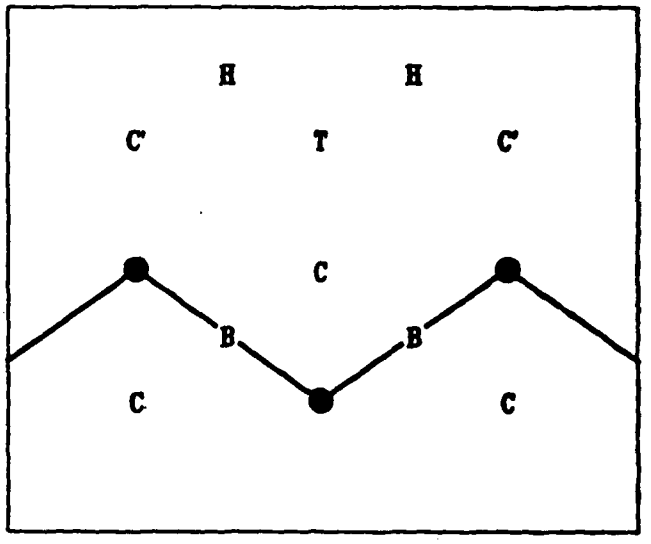



FIG. 2. Schematic illustration of the (110) plane through the atoms in the Si crystal, with labels for relevant high-symmetry positions. $T$ is the tetrahedral interstitial site, $H$ is the hexagonal interstitial site, $B$ the bond center, and $C$ (and $C'$) is at the center of a rhombus formed by three adjacent Si atoms and the nearest $T$ site.

the low-electron-density "channels" and includes the high-symmetry tetrahedral ($T$) and hexagonal ($H$) interstitial sites. Energy differences between $H$ positions located within the same region generally converge quite fast [cf. Fig. 1(a)], while those between positions in different regions [cf. $B$ and $T$, in Fig. 1(b)] are slower in convergence. Reference configuration (b) was chosen for this test because it presents the extreme "worst case" in terms of convergence properties; energy differences between other configurations have consistently been found to converge faster than presented in Fig. 1(b). We have also carried out test calculations in 16- and 32-atom supercells up to (12;24) Ry that confirm that the behavior as a function of cutoff is the same for all cell sizes.

We have found that inclusion of plane waves with kinetic energy up to (6;12) Ry (i.e., waves up to 6 Ry included in exact diagonalization, between 6 and 12 Ry in second-order perturbation theory) is sufficient for obtaining the general features of the energy surfaces. Since some of the configurations (e.g., the $B$ site) are particularly sensitive to the energy cutoff, all of the calculations necessary to derive energy differences for those sites were also carried out at the higher cutoff of (9;18) Ry. This cutoff corresponds to a basis set of $\sim$5500 plane waves in the 32-atom cells. The values for energy differences and barriers that will be quoted all result from these high-cutoff calculations. This may lead to small quantitative differences with values that were quoted in earlier brief accounts of the present work.[17,18]

Finally, we have also examined the reliability of Löwdin second-order perturbation theory[41] for deriving total-energy differences in this system, by comparing results obtained at $(E/2;E)$ Ry with those obtained at $(E;E)$, i.e., without any perturbation theory, for various values of the cutoff $E$. We found that the deviations were always smaller than 0.1 eV [e.g., only 0.07 eV at (6;12)

Ry], and decreased rapidly to zero when the cutoff was increased [being smaller than 0.01 eV at (15;30) Ry], showing that the values converge to the same limit at infinite cutoff.

### C. Supercells

To study the atomic and electronic structure of an impurity in the crystalline environment, while still preserving the translational symmetry of the system required for our theoretical formalism, we artificially introduce periodicity by constructing a supercell in which the impurity is surrounded by a sufficiently large number of Si atoms. We typically use supercells containing 32 Si atoms, such that the distance between neighboring defects if 9.4 Å. The convergence as a function of supercell size was tested by performing calculations on supercells containing 8, 16, and 32 Si atoms. For energy differences between H positions in the same (high- or low-electron-density) region, the 16-atom cell was found to suffice; energy differences between different locations for the same charge state were converged to within 0.1 eV, and differences between different charge states to within 0.2 eV. For energy differences between H positions in high- versus low-density regions, however, the error bar in a 16-atom cell is larger. Compared with a 32-atom cell, deviations of up to 0.4 eV may occur in the energy differences. These deviations can be attributed to the larger extent of the defect wave functions, as observed in plots of the charge density (see Sec. III B), which causes more significant interactions between neighboring 16-atom supercells for H at the bond center.

The 32-atom cells also facilitate the extraction of band positions for an isolated defect, since the dispersion of this level (caused by interactions with defects in neighboring supercells) is less than 0.5 eV in this case. The position of the level that would correspond to an isolated defect (i.e., without dispersion) was determined by taking a weighted average over the band positions at the special points. It changed by less than 0.1 eV when the cell size was increased from 16 to 32 atoms.

Dispersion of the defect levels due to interactions between neighboring supercells places an error bar on the derived position for any defect level. A more important source of uncertainty, however, is due to the intrinsic deficiencies of the local-density approximation (LDA), particularly the fact that the LDA predicts conduction bands and hence conduction-band-derived energy levels to be too low. We will therefore refrain from quoting specific results for positions of energy levels in the band gap. We note, however, that a qualitative distinction between various positions of the H-induced level can still be made. We also note that, while the absolute position of the defect level is uncertain, its relative motion induced by displacements of the impurity or by changes in the charge state is quite reliable. These observations will allow us to derive conclusions about the deep levels induced by hydrogen, as described later in the paper. Only in the section where we discuss the relative stability of various charge states as a function of the Fermi-level position will we be confronted with the limitations of the LDA.

For each position of the impurity, we need to let the neighboring Si atoms relax to find the lowest-energy configuration. Relaxation of two shells of Si atoms surrounding the H impurity is included in the full calculations. The need to relax two shells of Si atoms was another reason to carry out the calculations in 32-atom supercells, since 16-atom cells are too small to include anything but first-neighbor relaxations in a meaningful way. Second-neighbor relaxations can lower the energy by several tenths of an eV, e.g., for H at the bond center. Relaxation of further shells causes less than 0.1 eV change in the total energy, as was checked with a Keating model.[42]

### D. Special points

Integrations over the first Brillouin zone are performed using the special-points scheme.[38] In the 32-atom cells, two special points in the irreducible part of the zone are used for trigonal symmetry situations (e.g., H on the extension of a Si-Si bond), and equivalent larger sets for lower-symmetry configurations. We test the convergence as a function of the special-point sample as follows: If we increase the parameter $q$ in the Monkhorst-Pack[38] scheme from 2 to 4, the number of special points generated in the irreducible part of the Brillouin zone increases from two to seven for the $T$ site, and from two to twelve for the $B$ site. Even though the absolute value of the total energy changes significantly when the larger $k$-point set is used, all energy *differences* between different sites change by less than 0.05 eV.

### E. Spin polarization

In this work we report local-density-functional results for total energies and defect levels; spin polarization,[28] which affects only the neutral charge state (with an unpaired electron), was not included. We established the validity of this approach by carrying out self-consistent spin-density-functional calculations, which are much more time consuming, at selected sites. For the bond-center position the inclusion of spin polarization has very minor effects: the total energy goes down by less than 0.02 eV, and the defect level is split by only 0.04 eV.

The deviation from the spin-averaged results is expected to be largest for H at the $T$ site, where the crystal charge density reaches its lowest value so that the impurity is most "free-atom"-like. (Note that the $T$ is *not* a stable site for $H^0$ in Si, as we will see in the next section.) It is worthwhile to point out here, for the purposes of the present study, namely the derivation of total energies, what the effect is of neglecting spin polarization in the LDA calculation for the free H atom. The total energy deviates from the spin-polarized value by $\sim 0.9$ eV. The error can be associated with the absence of exchange splitting, which would lower the occupied level. In the solid, such exchange splittings are known to be substantially reduced from the free-atom values; this was observed in calculations for transition-metal impurities.[43] These qualitative arguments were confirmed by the full spin-polarized calculations. For $H^0$ at the $T$ site, we

found that inclusion of spin polarization lowered the total energy only by 0.1 eV. The defect level was split into a spin-up and a spin-down level, which were separated by 0.37 eV. These results are consistent with spin-polarized linearized muffin-tin orbital (LMTO) Green's-function calculations for H in Si.[44]

The overall conclusion is that the effects of spin polarization on the total energy are very small. They are therefore not included in the calculations that lead to the total-energy surfaces presented in the next section. We will, however, show contour plots of spin densities, which provide valuable information about the electronic structure of the impurity at different sites.

### F. Charge states

The calculation of charge states requires careful treatment, since the LDA pseudopotential expressions for the total energy are all derived assuming charge neutrality in the unit cell.[37,39] Such neutrality is indeed necessary to avoid divergence of the long-range Coulomb terms. Taken individually, the $G=0$ terms of the electron-ion, electron-electron (Hartree), and ion-ion interactions are infinite. A finite result is obtained, however, by appropriate combination of terms, leading to two well-defined and finite contributions: (1) the Ewald energy, which is the energy of a periodic array of positive point ions in a uniform neutralizing (negative) background, and (2) the so-called $\alpha Z$ term,[37] which represents the Fourier component for $G=0$ of the electron-(pseudo)ion interaction.

Our approach for performing the calculations on a charged system is as follows: We define the occupation of the electronic energy levels to represent the system that we want to study (i.e., positively or negatively charged, by taking out an electron, or putting in an extra electron with respect to the neutral system). The charge density is then calculated from the wave functions of the occupied states, and all summations in the total-energy terms, as well as the generation of a new potential in the self-consistent process, are carried out with this charge density. However, the $G=0$ terms (i.e., the Ewald and $\alpha Z$ terms) are always calculated for the *neutral* system (the charge being determined by the ionic charges in the supercell). Neutrality is essential here, because a non-neutral system would surely lead to diverging terms. It is also the appropriate approach to the physics problem, for the following reasons.

Since charge neutrality is a fundamental requirement, all calculations should really be set up with a number of electrons that exactly equals the number of positive charges in the unit cell. Since the latter is determined by the structure, use of the Ewald and $\alpha Z$ terms for the *neutral* system is appropriate. The neutrality condition leads, strictly speaking, to the requirement of the presence of an additional charge that would compensate the extra charge in our system. Specifically, if we put an extra electron on a defect, then we should have a hole (compensating positive charge, or absence of an electron) present in the same unit cell—and very far removed from the extra electron, so as not to lead to spurious interaction terms. This setup is generally impractical, since it leads to a requirement of very large supercells, and makes separation of the terms in the total energy that are related to the defect (and not to the compensating charge) very difficult. Nevertheless, it was used by Vanderbilt and Joannopoulos[45] in a study of defects in Se, in conjunction with an elaborate scheme for specifying level occupations.

Since the compensating charge is not supposed to interact with the charge on the defect, and basically only serves to maintain charge neutrality for the calculation of $G=0$ terms—something which we impose anyway—we can take the shortcut of leaving it out of the calculation altogether. By doing this, we are neglecting a Madelung-type term which would describe the interaction between (screened) positive and negative charges, and which would vanish in the limit of an infinite supercell. This approach will therefore be justified if the results are shown to be converged as a function of supercell size. In that case, the supercell is large enough to avoid spurious interactions between charged defects in neighboring cells. A test as a function of supercell size is therefore essential, and has been carried out in this study with satisfactory results.

As a final check on the procedure, we have examined one test case in which the "strict" application of charge neutrality was obeyed, by having two oppositely charged defects present in the supercell. Our calculations on individual defects in a 32-atom cell (to be described in more detail in the next section) established that at the bond-center site H is most stable in the positive charge state, while at the tetrahedral interstitial site the negative charge state is favored. We also obtained values for the total energies for each of these configurations. We then proceed to construct a 32-atom supercell in which both defects are present at the same time: one impurity at a bond center, $B$ (with appropriate relaxation of the surrounding lattice), the other at the tetrahedral interstitial site ($T$). The minimum-energy electronic structure for this arrangement should put the H at $B$ in the positive charge state, and the H at $T$ in the negative charge state. This is indeed what we find by analyzing the charge density. This supercell is now overall neutral (since it contains one positive and one negative defect), and therefore the calculation strictly follows the treatment of $G=0$ terms, as discussed above. Such an arrangement of defects therefore follows the scheme proposed by Vanderbilt[45] for performing calculations for charged states. We want to check whether the total energy obtained from this calculation is equal to the sum of the total energies obtained from separate calculations for the individual defects. Spurious interactions between the positive and negative charges within the unit cell (and with neighboring cells) may be present, of course; assuming, however, that the defects are sufficiently well separated so that the only interaction would be through a screened Coulomb interaction, the resulting changes in the total energy would be quite small. That allows us to obtain a value for the total energy for a pair of (to a good approximation) noninteracting defects. This total energy turns out to be the same (within 0.1 eV) as the sum of total energies obtained from calculations for individual defects, in which the prescription outlined above was followed. This agree-

ment confirms the validity of our prescription for charge states.

### G. Energy surfaces

To study the behavior of an impurity (in a particular charge state) in a semiconductor, one needs to know the total energy of many different configurations, in which the impurity is located at different sites in the host crystal. For each position of the impurity, the surrounding atoms should be completely relaxed. The resulting energy values as a function of the coordinates of the impurity define an energy surface: $E=E(\mathbf{R}_{\text{imp}})$. Notice that this function does not depend on the coordinates of the host atoms; that is because for each position $\mathbf{R}_{\text{imp}}$ an energy minimization procedure has been performed (i.e., relaxation) that determines what the coordinates of the host atoms are. Once the function is known, it immediately provides information about stable sites, migration paths, and energy barriers along these paths.

A function such as $E=E(\mathbf{R}_{\text{imp}})$ that depends on three dimensions is difficult to visualize. Symmetries of this object can play an important role in simplifying both the calculational task and the conceptual understanding. The crucial point here is the realization that the energy surface possesses the full symmetry of the crystal. To make effective use of this symmetry, an analytic description of the surface is essential. We achieve this through expansion in a basis set with the appropriate symmetry. A natural choice for a basis set with the full symmetry of the lattice is a set of symmetrized plane waves. Such a description is used, for instance, in self-consistent plane-wave basis-set calculations to represent charge densities; experience indicates that relatively few coefficients suffice to adequately describe the overall features of the function. A progressively better description can be obtained by including more symmetrized plane waves.

In order to test the representation, we worked with a large data base of energy values (more than 16 locations of H in the lattice). For each test we selected a particular subset of these, containing $m$ values, and used this set to determine the expansion coefficients when $n$ basis functions (symmetrized plane waves) were included in the expansion of the energy surface. With $m \geq n$, a least-squares-fitting procedure was used. We found that a minimum of six basis functions is required to represent the features of the surface; the quality of the representation could be judged by taking the predicted energy value at a data point that was not included in the subset of $m$ points, and comparing it with the value that was independently calculated from first principles. Increasing the number of basis functions from eight to ten led to energy changes in the relevant areas of the surface of less than 0.05 eV. (Near the atoms, the surface rises very rapidly; relative variations with the number of basis functions may be larger in these regions, but are of no consequence for the physical behavior). We conclude that eight to ten calculated data points suffice to determine the expansion coefficients.

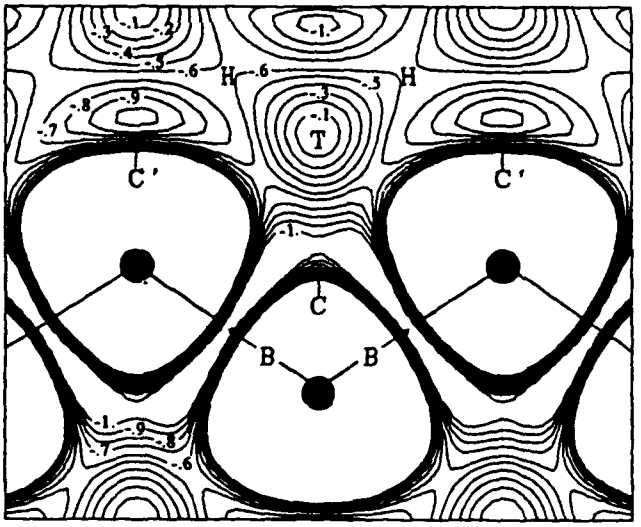
While it is impossible to pictorially represent the energy surface as a function of all three dimensions, our choice of data points and our fitting procedure assure that we take the full three-dimensional character into account. For visualization, we restrict the coordinates of the H impurity to a particular plane [e.g., the (110) plane through the atoms]. The energy surface can then be displayed as a contour plot (the curves presenting lines of constant energy), or as a perspective plot of the energy (along the vertical axis) as a function of the coordinates in the plane. Both types of plot will be used here. Note that the Si relaxations for each position of the impurity atom are different but are not displayed in these figures.

In Figs. 3–5 we show contour plots and perspective plots for different charge states of H in Si, with the impurity coordinates restricted to the (110) plane through the atoms. In all plots the (arbitrary) zero of energy is at the $T$ site. The contour plots are self-explanatory; high-symmetry sites (cf. Fig. 2) have been included for easy inspection. The perspective plots have been color-coded to allow straightforward identification of the relevant regions. Red regions present the lowest-energy values, blue is intermediate, and green is for the highest energies. The plots should be interpreted as a perspective view of a landscape, in which the low-lying regions ("valleys") represent the most favorable positions for the impurity. The plateaus around the perfect-crystal Si atomic sites are not real: when the H atom approaches any Si atom too closely, the energy rises rapidly; this gives rise to a very steep "mountain" in the surface, which would obscure everything behind it in a perspective plot. We have therefore cut off these mountains at a value listed as the upper limit of the green regions in the plots. A quantitative discussion of these plots will be given in the next section.

## III. RESULTS FOR A SINGLE HYDROGEN IMPURITY IN CRYSTALLINE SILICON

The energy surfaces for H in the positive, neutral, and negative charge states, as depicted in Figs. 3–5, exhibit a number of common features. In all three charge states there are two distinct regions in which the H atoms exhibit significantly different behavior. First, there is the region of high electron density, which includes the $B$ (bond-center) site and the $C$ site (at the center of a rhombus formed by three adjacent Si and the nearest $T$). In this region the nearby Si atoms relax strongly. For example, when the H atom is placed at the bond-center site, the adjacent Si atoms relax out by 0.4 Å for a net gain in energy of more than 4 eV. If no relaxation were included, the red low-energy region in Fig. 3(b) would completely disappear. Figure 6 shows the energy surface calculated for a rigid Si lattice; the bond-center and other positions in the high-density region are indeed high in energy now. Coming back to the case in which relaxation *is* included, we find that in the high-density region a H-induced defect level occurs in the upper part of the energy gap; it is identified as a state formed out of an antibonding combination of Si orbitals. The second region consists of the low-electron-density "channels" and includes the high-symmetry tetrahedral ($T$) and hexagonal ($H$) interstitial sites. Here, the Si atoms in the vicinity of

H relax very little if at all. Furthermore, a H-related level now occurs just below the top of the valence bands. The precise position of the defect levels changes only by $\sim 0.1$ eV as a function of charge state. We now discuss the various charge states and their relative stability in more detail.




(a)

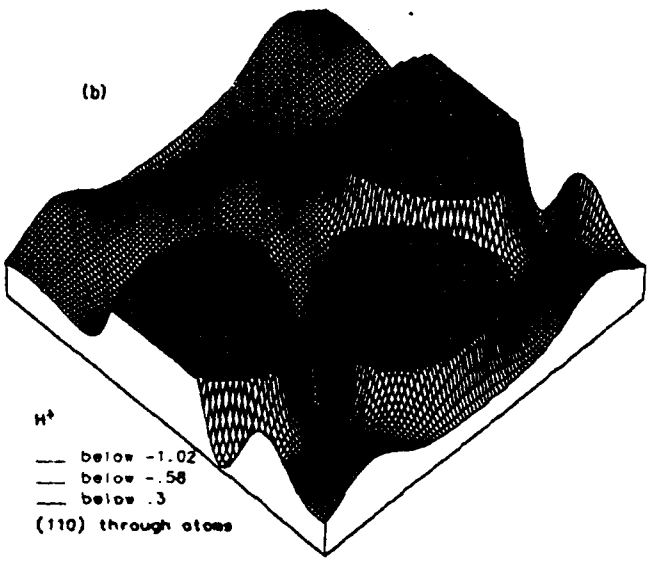
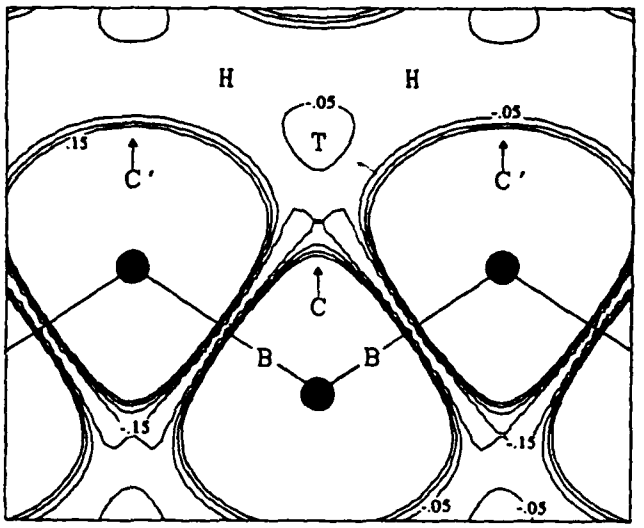



FIG. 3. (a) Contour plot and (b) perspective plot of the energy surface for $H^+$ in a (110) plane through the Si atoms. The zero of energy is arbitrarily chosen at $T$. The black dots represent Si atoms at their unrelaxed positions; the relaxations (which are different for different H positions) are not shown but are taken into account in the total-energy calculations. In (a) the contour interval is 0.1 eV. The color coding of the perspective plot in (b) is indicated in the figure: the energy values below $-1.02$ eV are shown in red; between $-1.02$ and $-0.58$ eV in blue; and between $-0.58$ and 0.3 eV in green. The surface is cut off at an energy value of 0.3 eV.

### A. Positive charge state

Figure 3 shows the energy surface in the (110) plane for a positively charged H ($H^+$). The global minimum is at the bond center ($B$) site, symmetrically located between two Si atoms. In contrast, the energy of $H^+$ in the low-density region is more than 0.5 eV higher (the bond center is 1.2 eV lower in energy than the $T$ site). Of course, the state $H^+$ in the low-density region actually does not occur, because the H-related level which must be kept empty lies inside the valence bands. Note that the positive charge state does *not* imply that H occurs as a bare proton; at the bond center, the missing charge is actually taken from the region near the Si atoms, corresponding to the state occurring in the band gap. In a simplified picture of combination of orbitals on the H and




(a)

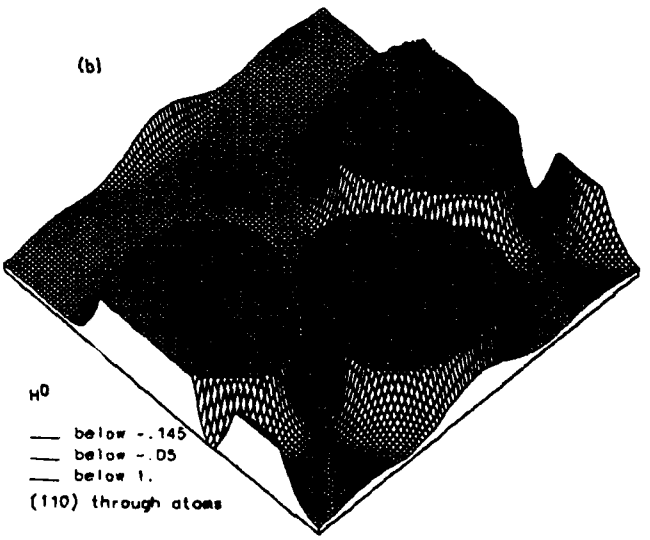
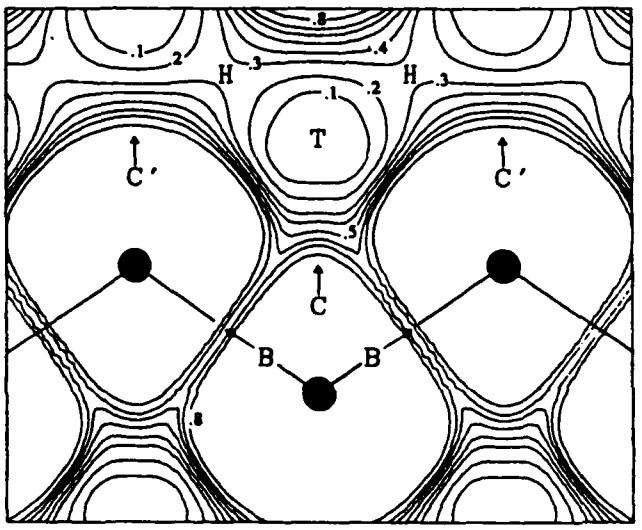



FIG. 4. (a) Contour plot and (b) perspective plot of the energy surface for $H^0$ in a (110) plane through the Si atoms. See caption of Fig. 3.

the neighboring Si atoms, this state can be considered to be formed out of an antibonding combination of Si orbitals; viewed as a state of the defect complex, it is effectively nonbonding in character, since it has a node through the H atom. For H at the bond center, our use of the notation $H^+$ therefore implies that the actual defect is a complex formed by the H and the surrounding Si, the electron being removed from an antibonding combination of Si orbitals rather than from the H itself.

A migration path in the (110) plane can be traced between the bond-center positions; the barrier along this path is ~0.2 eV high. This path can clearly be seen as the red region winding its way around the Si atoms in Fig. 3(b). The saddle point occurs very close to the point indicated with *C* in the contour plot; the points *C'* are symmetry-related points along equivalent paths perpendicular to the plane of the figure. At the saddle point in the (110) plane, H is located 1.25 Å away from the *T* site. The Si atom below it, on the line through *T* and the saddle point, relaxes down by 0.16 Å to make the Si-H distance equal to 1.63 Å. Since we cannot show the energy surface as a function of all three dimensions, the (110) plane and the indicated migration path should only be considered as a representative example. We have also studied the behavior in various other planes. Figure 7 shows the energy surface in a (110) plane parallel to the plane in Fig. 3 and lying halfway between equivalent planes through the atoms. In particular, we are interested in the behavior around the *M* site, which is midway between two *C* sites [only one of which lies in the (110) plane]. Corbett *et al.*[29] proposed this site as the minimum-energy location for neutral H in Si. In our energy surfaces for $H^+$, we find it to be at approximately the same energy as the bond center *B*, with no barrier between the two. Migration along a path involving the *M* sites still involves a barrier of ~0.2 eV. As can be seen in Fig. 7, the *M* point also lies on a line perpendicular to the Si—Si bond, connecting the bond center with the neighboring hexagonal interstitial site; all points between *B* and *M* on this line have approximately the same energy. For these "buckled" configurations, the Si-H distance remains almost constant (equal to 1.6 Å), due to appropriate relaxation of the Si atoms. In all cases, the H prefers to be symmetrically located with respect to two Si atoms.

Figure 8, finally, shows a cross section which contains the *B*, *M*, and *H* sites, but also goes through the atoms. The flatness of the surface along the line from *B* to *M* is once again evident. It can also be observed that the ener-




(a)

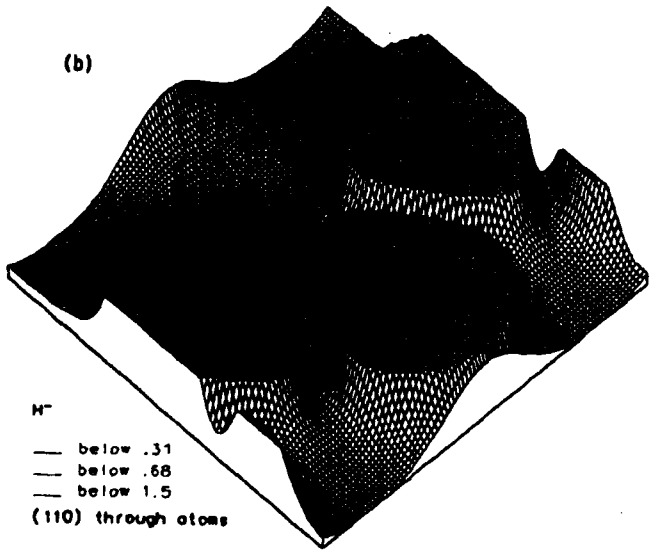



FIG. 5. (a) Contour plot and (b) perspective plot of the energy surface for $H^-$ in a (110) plane through the Si atoms. See caption of Fig. 3.

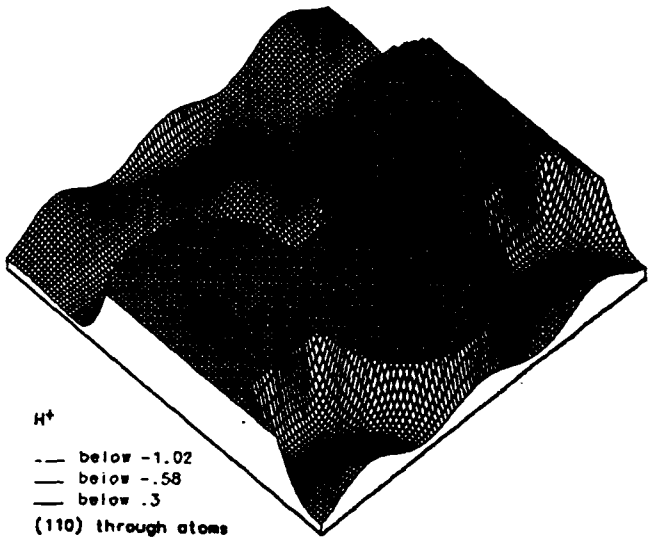
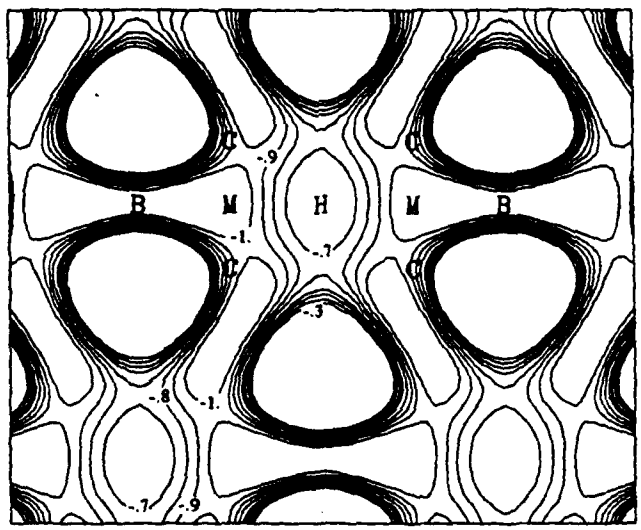



FIG. 6. Perspective plot of the energy surface for $H^+$ in a (110) plane through the Si atoms. To generate the values for this plot (unlike all others) the Si atoms were kept fixed in their rigid lattice positions. Comparing with Fig. 3, we see that the low-energy regions have disappeared.

gy surface rises steeply along the bond direction ([111]).

For H at the bond center itself, in the positive charge state, the neighboring Si atoms move out over 0.41 Å, to make the Si-H distance equal to 1.59 Å. This distance is slightly larger than the Si—H bond length in molecules such as $SiH_4$, where it is 1.48 Å. This is understandable since H at the bond center is bonded to *two* Si atoms, forming a three-center bond. The second neighbors move by 0.07 Å; the distance between first and second neighbors is equal to 2.31 Å.

These motions of the Si atoms are quite large, and must involve a significant energy cost. To estimate this raise in energy due to strain, we have performed a calculation of the Si atoms in the positions described above, but in the absence of the H impurity. The total energy is 1.55 eV higher than for the lattice in equilibrium. This means that the energy gained due to bonding between H and Si must be greater than 1.55 eV, in order for the bond-center configuration to be stable. This "cost of relaxation" can also be interpreted in the following fashion: If a situation could be created in which outward motion of the Si atoms would levy no energy cost, the bond-center configuration for $H^+$ would be more stable by 1.55 eV, compared to the situation in crystalline Si. This ob-




(a)

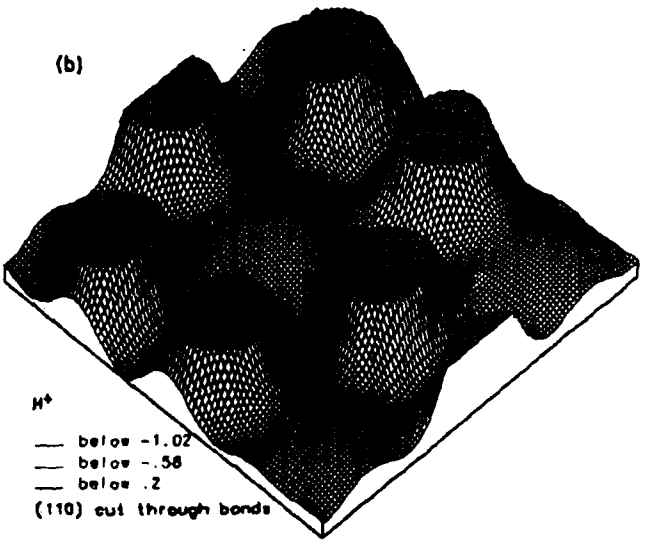
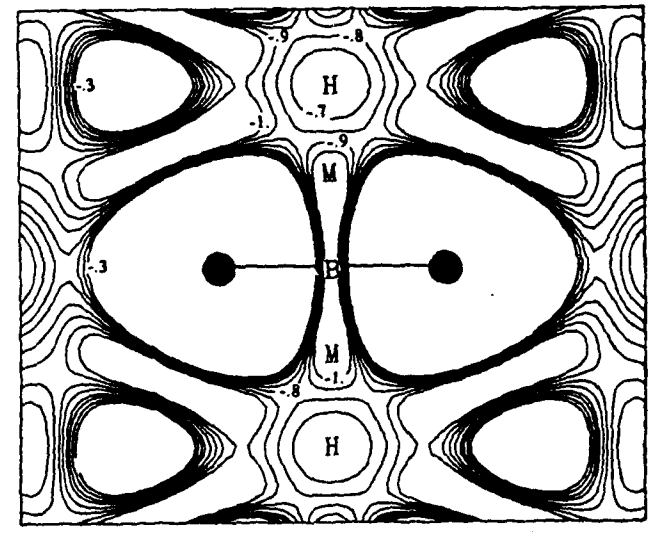



FIG. 7. (a) Contour plot and (b) perspective plot of the energy surface for $H^+$ in a (110) plane through the sites *B*, *C*, *H*, and *M*. This plane is parallel to the plane of Fig. 3, and midway between equivalent planes through the atoms. The *M* point is located midway between a bond center and the nearest hexagonal interstitial site. The zero of energy is arbitrarily chosen at *T* (not in the plot). See caption of Fig. 3.




(a)

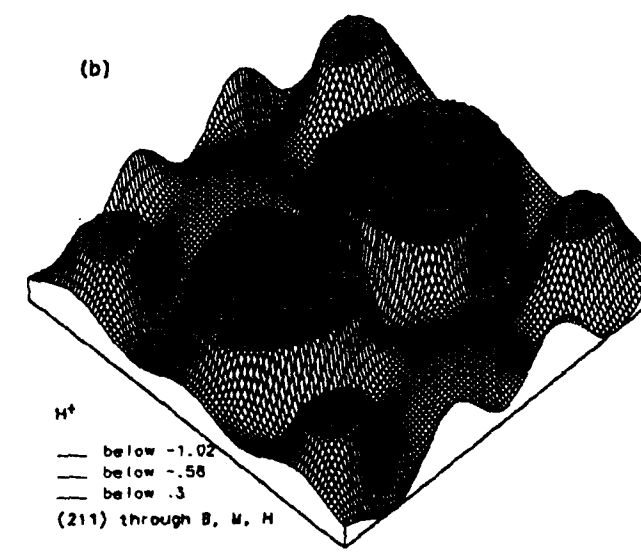



FIG. 8. (a) Contour plot and (b) perspective plot of the energy surface for $H^+$ in a (211) plane through the Si atoms, containing the sites *B*, *H*, and *M*. The *M* point is located midway between a bond center and the nearest hexagonal interstitial site. The zero of energy is arbitrarily chosen at *T* (not in the plot). See caption of Fig. 3.

servation might be relevant for amorphous Si, in which bond distortions are readily allowed.

### B. Neutral charge state

For neutral H the same features and relative positions of extrema can be recognized as in the case of $H^+$, including a global minimum at the bond center. For H at the bond center the neighboring Si atoms move out over 0.45 Å, to make the Si-H distance equal to 1.63 Å, i.e., slightly larger than in the positive charge state. The second-neighbor relaxation is the same as for the positive charge state. The energy cost due to Si motion, as described at the end of Sec. III A, is 1.73 eV for the case of the relaxations appropriate for neutral H. As in the case of $H^+$, we have examined carefully whether there is any tendency for $H^0$ to preferentially bind to one of the Si neighbors, leading to an asymmetric configuration, as suggested by DeLeo *et al.*[46] In contrast to Ref. 46, we find that the symmetric situation is lowest in energy. The saddle point of the migration path in the (110) plane is again located on the line between *C* and *T*, but closer to *T* than in the case of $H^+$: H is 0.60 Å away from *T* now. Relaxation of the Si atoms is negligible for H at this site. The energy is less than 0.2 eV higher than at the bond center. Figure 9(a) shows the charge density in a (110) plane for neutral H at the bond center. A concentration of charge around the impurity is immediately obvious. Most of this charge in the bond region is related to H-induced levels buried in the valence band. It is interesting to also examine the spin density which results from a spin-polarized calculation, as described in Sec. II E. Figure 9(b) shows the difference between the spin-up and spin-down densities. This figure is remarkably similar to one that would result from plotting the charge density associated with the H-induced defect level in the band gap (this being the level that is occupied with one, e.g., spin-up, electron in the neutral charge state). It is clear that this density corresponds to an antibonding combination of Si orbitals, with mainly *p*-type character. Notice that virtually no spin density is to be found at the bond center itself. These observations can be relevant for interpretation of muon-spin-resonance experiments.[19]

The charge density for neutral H at the *T* site is shown in Fig. 10(a). Note that the *T* site is *not* a stable site for $H^0$ in Si, but it is educational to inspect Fig. 10 and compare it with Fig. 9. The difference between spin-up and spin-down densities is displayed in Fig. 10(b). Once again, it corresponds closely to the density associated with the H-induced defect level, which is now below the top of the valence band. This density is now clearly associated with an *s*-like state centered on the impurity.

Turning back to the energy surface for $H^0$, we note that the path through the region of high electron density is favored (as for $H^+$), but the low-density path is only 0.2 eV higher. Thus, neutral H seems to be able to move rather freely through the network with very small energy barriers. We note that the *T* site is a local maximum of the energy surface for $H^0$. Moving from *T* towards a substitutional site, the energy first decreases and then increases in the [111] direction. However, the lowest energy in this antibonding direction (less than 0.1 eV lower than at *T*) does not correspond to a local minimum, but to a saddle point, i.e., the energy can be lowered by moving the H off the [111] direction. The same conclusion holds for the hexagonal interstitial (*H*) site, which lies in

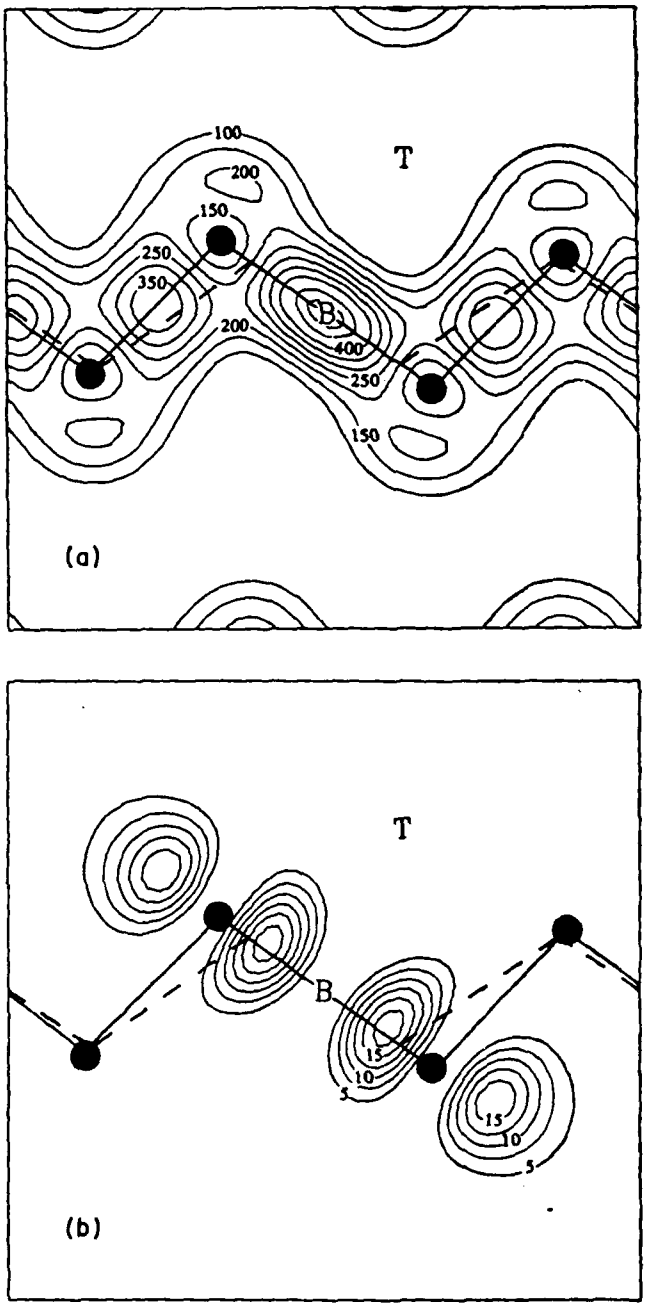



FIG. 9. (a) Contour plot of the charge density in the (110) plane through the atoms for neutral H at the bond center. The Si atoms in their relaxed positions are indicated with black dots and connected with solid lines. Dashed lines connect the unrelaxed atomic positions. The contour interval is 50; units are electrons per unit cell (for a supercell containing one H and 32 Si atoms). (b) Contour plot of the difference between spin-up and spin-down densities in the (110) plane through the atoms for neutral H at the bond center. The contour interval is 2.5 electrons/(unit cell).

the [111] direction halfway between two $T$ sites. The $H$ site is a local minimum along the [111] direction, but only a saddle point when considered in three dimensions. Similar conclusions regarding antibonding and hexagonal sites hold for the positive charge state, where the energy differences are more pronounced. It is interesting to note that the instability of the antibonding site also occurs for H around a boron acceptor in Si,[47] eliminating this site as a candidate for the structure of B-H complexes that result from passivation (see further).

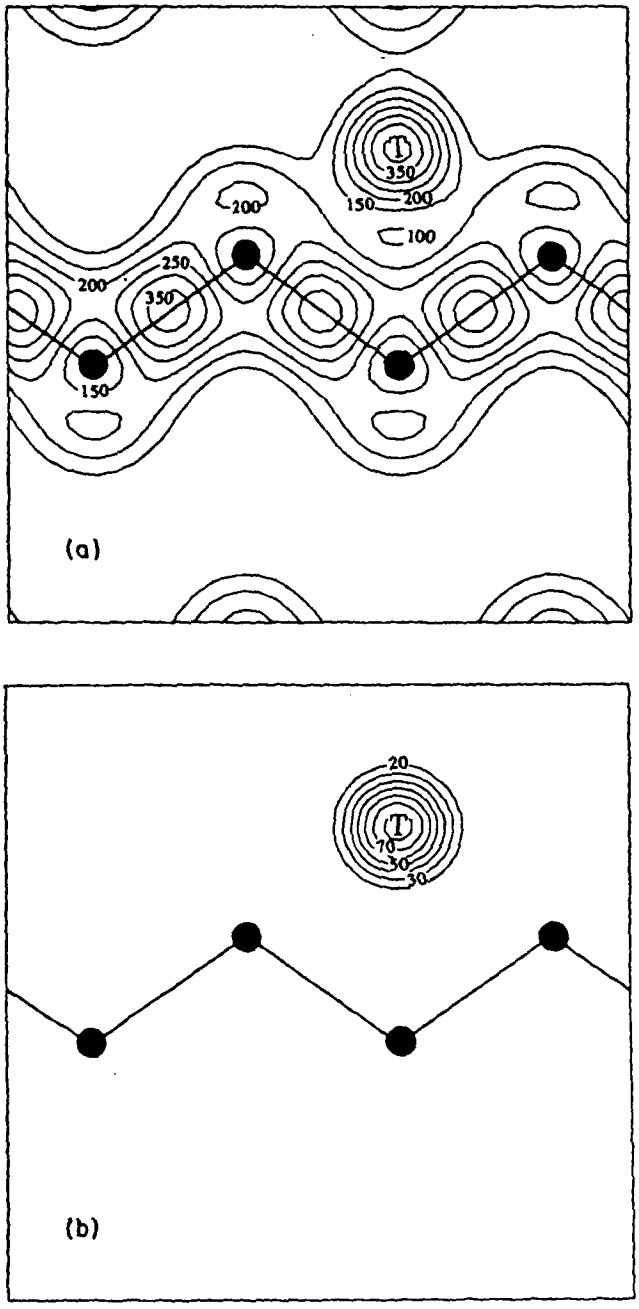



FIG. 10. (a) Contour plot of the charge density in the (110) plane through the atoms for neutral H at $T$. The Si atoms are indicated with black dots; no relaxation occurs. The contour interval is 50; units are electrons per unit cell (for a supercell containing one H and 32 Si atoms). (b) Contour plot of the difference between spin-up and spin-down densities in the (110) plane through the atoms for neutral H at $T$. The contour interval is 10 electrons/(unit cell).

### C. Negative charge state

The negative charge state distinctly differs from $H^+$ and $H^0$ in that it is now the low-electron-density regions of the crystal which provide the most stable sites for the impurity. This can be understood by realizing that the energy cost of placing a second electron in the level in the gap (which was the trademark of the high-density sites) becomes too high, and it is more favorable to move the H to locations where the induced defect level occurs at lower energies. The $T$ site is now the lowest in energy, with the energy rising sharply outside the low-density regions. In particular, the $B$ site is now more than 0.5 eV higher in energy than the $T$ site. The barrier for migration along a path through the low-density region and going through the hexagonal interstitial site is 0.25 eV.

The negative charge state is thus the only one for which the $T$ site is a stable site (local and global minimum in the energy surface). The charge density associated with this state is quite similar to that depicted in Fig. 10. This is the position for which the analysis of Altarelli and Hsu applies, showing why the H level is expected to be deep and not effective-mass-like.[48]

### D. Relative stability of different charge states

We now examine the *relative* energies of the different charge states, in order to determine the lowest-energy state. To alter the charge state, electrons must be taken from or removed to a reservoir; the Fermi level determines the energy of electrons in this reservoir. The relative energies therefore depend on the position of the Fermi level. Figure 11 shows the relative formation energies for different charge states, as a function of Fermi-level position. To simplify the plot, we only show the formation energies for the impurity positions which correspond to the global minimum for a particular charge state, i.e., $B$ for $H^+$ and $H^0$, and $T$ for $H^-$. Figure 11(a) shows the values directly obtained from the LDA calculations. As pointed out above, these suffer from an uncertainty in the position of the defect level. Rigorous calculational schemes which could eliminate these uncertainties by going beyond the LDA are presently prohibitively complex and too computationally demanding to apply to defect calculations. We have therefore applied a very simple *a posteriori* correction, amounting to a rigid shift of the defect level together with the conduction bands, to bring the band gap into agreement with experiment. The result of this procedure is shown in Fig. 11(b). The energies are shifted now, according to the number of electrons present in the level. We stress that Fig. 11 is not intended to display quantitative results, but merely to provide a qualitative indication of the stability of different charge states.

In $p$-type material (Fermi level at the top of the valence band), the lowest-energy state is $H^+$ in the high-density region; thus, $H^+$ diffuses via the high-density path and exhibits donorlike behavior. These conclusions are

unambiguous and independent of any error bars in our LDA calculations. This result confirms the suggestion that the passivation of *p*-type material is a direct result of compensation, i.e., electrons from neutral H atoms annihilate the free holes in the valence band.[16] Pairing between $H^+$ and ionized acceptors follows compensation. The structure of the hydrogen-impurity complexes that result from this pairing will be addressed in a forthcoming publication.[47]

From Fig. 11 we see that our calculations predict H to be a negative-*U* impurity, much like the Si self-interstitial.[49] In *p*-type material the stable state is $H^+$ in the high-density region; as the Fermi level is raised, however, the stable state becomes $H^-$ in the low-density region. $H^0$ is not the stable state for any Fermi level. However, the uncertainty in the LDA energy levels (and in our simple correction procedure) makes the error bar too large to unambiguously exclude the occurrence of $H^0$.

### E. Vibrational frequencies

The frequencies of the hydrogen stretching mode for $H^+$ and $H^0$ at the bond-center site are calculated in a 32-atom cell at a cutoff of (10;20) Ry. The H atom is moved from the equilibrium position in the [111] direction (towards the Si atoms) over distances of 2% and 4% of the Si—Si equilibrium bond length. Since the proton is much lighter than the Si atoms, it is a good approximation to assume that the Si atoms do not move on the time scale of the H vibration. Relaxation of the host crystal is therefore kept fixed to that of the equilibrium position. The energy differences obtained from the calculations for different H positions are fitted to a parabola, from which the vibrational frequency can be determined. This procedure leads to 2210 $cm^{-1}$ for $H^+$ and 1945 $cm^{-1}$ for $H^0$. The error bar on these values is $\pm 100$ $cm^{-1}$. Experimental values[50,51] for stretching modes involving a single H atom in hydrogenated amorphous or crystalline Si range between 2000 and 2200 $cm^{-1}$. It has often implicitly been assumed that such stretching modes involve single Si—H bonds (such as for H tying off a dangling bond at a vacancy). The present results show, however, that bond-centered H in crystalline Si gives rise to similar frequencies.

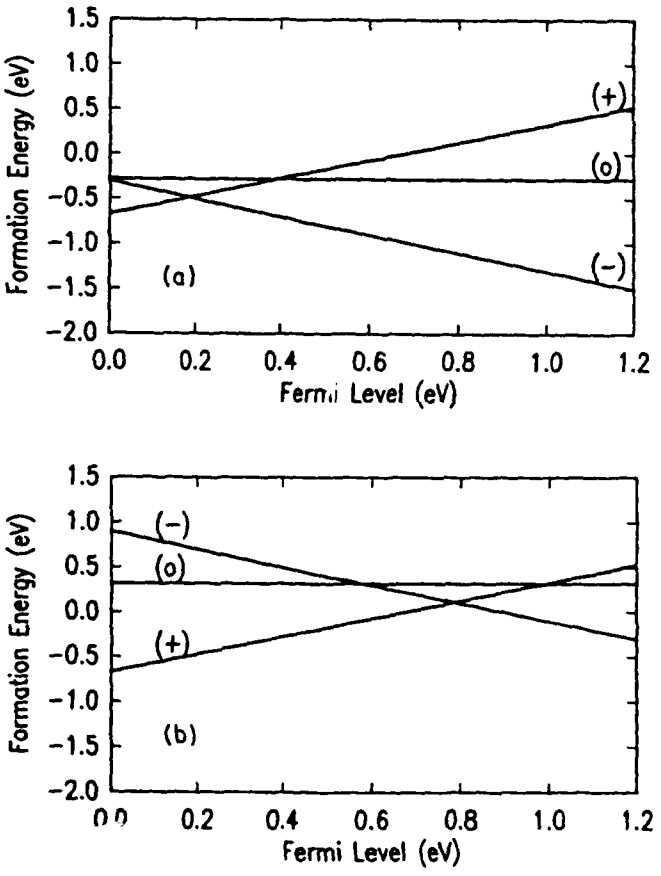



FIG. 11. Relative formation energies for different charge states of a H interstitial impurity in Si. (a) shows the straight LDA values, while (b) results from applying a simple correction scheme to the energy levels (see text). The zero of energy is arbitrarily chosen as the energy of $H^0$ at *T*. This figure is not intended to display quantitative results, but merely to provide a qualitative indication of the stability of different charge states.

### F. Discussion

A large amount of experimental information has been accumulated in recent years based on observations of interactions of H with shallow impurities. Interpretations of the data were often based on contradicting assumptions, as pointed out by Pantelides.[16] The comprehensive theoretical description provided in the present study now allows a coherent interpretation of all the data. We will also discuss results from experiments which directly address the problem of the location of H (or muonium) in the Si crystal.

#### *1. Passivation of shallow impurities*

It is known that the final result of the passivation mechanism in *p*-type material is the formation of neutral acceptor-H pairs, as observed in infrared spectroscopy measurements,[3,4,52] Raman studies,[53] and ion-channeling measurements.[54,55] The structure of these pairs will be addressed in a separate publication.[47] In order to understand the formation of these H-acceptor pairs, however, one needs to know the nature of the charge states of H along its diffusion path, which will determine which hydrogen-impurity reactions can occur. The assumptions that had been made previously were often contradicting and mutually inconsistent. Pantelides[16] showed that the only way to account for all the available data in *p*-type material was for H to have a deep donor level in the band gap. This conclusion has now been confirmed by the present theoretical results.

Hydrogen atoms in *p*-type material (where the Fermi level is below the hydrogen's donor level) prefer the positive charge state and will lose their electrons; these electrons can annihilate the holes through a mechanism of direct *compensation*. *Pairing* of hydrogen with acceptors is *not necessary* for compensation and passivation of *p*-type material, as has clearly been shown in recent experiments by Johnson and Herring.[56] Once $H^+$ has been formed, however, its high mobility and Coulombic attrac-

tion to negatively charged acceptor impurities will readily lead to the formation of acceptor-hydrogen pairs: $H^+ + B^- \rightarrow (HB)^0$ (where boron has been chosen as a typical acceptor). The pair formation is therefore a *consequence* of passivation in *p*-type material.

The term compensation is often presumed to imply that the stable state of the system is such that the atoms which act as donors are spatially separated from the acceptors which they compensate. This occurs, for instance, in the case of compensation by counterdoping, e.g., adding phosphorus to boron-doped Si. The term compensation, however, in general applies to any situation in which electrons from donor atoms annihilate free holes, and as such correctly describes the hydrogenation of *p*-type material. Because of the high mobility of the H species, the final experimentally observed situation will usually be such that H is paired with acceptors; pairing is only absent during the initial (transient) phase of hydrogenation, or at a temperature sufficiently high to dissociate H-acceptor pairs. Under those conditions, compensation is the accurate description for the state of the system. Our major goal in stressing the compensation aspect is to make clear that *pairing* is not essential for passivation, and that, indeed, the reaction of pair formation can only be correctly understood if compensation is considered to be the initial step. It should be clear that calculations in which only the structure of the resulting H-acceptor pairs is addressed cannot have any bearing on the issue of compensation as the initial step in the passivation mechanism. The statements by Chang and Chadi,[57] claiming that compensation is not involved in the passivation, are therefore unfounded, since they are inferred solely from an analysis of the already formed H-B pair.

The sequence of events in which pair formation follows compensation is essential for understanding a wide variety of experimental results, which will be summarized below.

(1) Since the diffusing species in *p*-type material is positively charged, electric fields are expected to significantly influence the diffusion properties. The observed electric field dependence[7,9] of hydrogen neutralization of shallow acceptors follows immediately, without having to invoke participation of free holes in the reaction.[7]

(2) When the *p*-type (B-doped) material is counterdoped, making it effectively *n* type, H diffusion is retarded and H-acceptor pairing is suppressed; the final H concentration is 2 orders of magnitude smaller than in *p* type.[4] Our theory shows indeed that if the Fermi level is raised the $H^+$ concentration will decrease. Neutral or negative H will not react with $B^-$ the way $H^+$ does, and the final concentration of (BH) pairs will be significantly lower. A thin *n*-type overlayer was also observed to block the penetration of hydrogen.[7,8] Once again, no $H^+$ can be formed in this layer. If $H^-$ is formed, it is kept out of the *p*-type substrate by the electric field in the depletion region. If $H^0$ would be formed, it would not as readily pair up with $B^-$ as $H^+$ does.

(3) Reverse bias of the junction formed by an $n^+$-type overlayer on a *p*-type substrate during hydrogenation results in a suppression of neutralization in the space-charge layer. The actual experiments were carried out with deuterium, an isotope of hydrogen which is more readily detectable with secondary-ion mass spectrometry.[7] The concentration of deuterium in the space-charge layer can greatly exceed the boron concentration, without neutralization occurring. These observations are consistent with molecule formation in that region, and a suppression of (BH) pairing due to the absence of $H^+$. Similar results from experiments by Tavendale[9] were explained as due to field drift of a positively charged species under an electric field. The fact that this positively charged species is $H^+$ (and not free holes) has recently been unambiguously established by Johnson and Herring,[56,58] who carefully analyzed the variation with depth of the H concentration in *p-n* junctions. Their results show that H must have a deep donor level, not far from midgap.

(4) Recent experiments by Johnson and Herring[56] have also provided direct support for the compensation mechanism. By carrying out electrical measurements in real time during hydrogenation they were able to directly study the migrating species, rather than having to infer its properties from formation kinetics of various H-related complexes. At 300 °C a temperature at which any $(HB)^0$ complexes should be completely dissociated, they still observed a sharp increase in the resistance upon hydrogenation. These observations must be due to the indiffusion of $H^+$ and compensation.

Johnson and Herring[59] have also analyzed deuterium concentrations in uniformly doped *n*-type material, as well as epitaxial layers of varying *n*-type doping on a single substrate. They concluded that H can occur in a negative charge state, with an acceptor level close to the donor level found in the experiments described above. While not as conclusive yet as the results for *p*-type materials, these observations do lend support to our prediction that $H^-$ is the stable charge state in *n*-type material. Further experimental work is required to test our prediction that H is actually a negative-*U* impurity.

### 2. *Location of H (and muonium) in the Si crystal*

Let us now turn to experiments in which the location of H in the lattice was the object of investigation. A number of ion-channeling experiments have been performed in order to determine the location of hydrogen in pure and doped Si. Once again, deuterium (D) is used, this time in order to take advantage of a nuclear reaction for detection. Picraux and Vook[60] found that D would be located predominantly in a single interstitial site 1.6 Å along a [111] direction from a Si atom in the antibonding direction. A major problem of the technique is the introduction of lattice damage due to the ion beam, and the resulting attachment of D to these defects. The observed D positions are therefore likely not those in pure, but in damaged Si, and may not only correspond to atomic, but also to molecular H. This problem has been addressed in careful experiments by Nielsen,[12] in which beam-induced damage was kept to a minimum. He found 80% of deuterium atoms to be located close to bond-center sites, while 20% are close to tetrahedral sites. The occurrence

of D at the bond center was ruled out by Nielsen on the ground of older theoretical calculations.[23,25,29] His low-temperature results are consistent, however, with a significant fraction of D located at the bond-center site, which emerges as the lowest-energy position for $H^+$ and $H^0$ from the present study.

As mentioned in the Introduction, a wealth of experimental information has been generated from muon-spin-resonance experiments. Our results for the behavior of neutral H in Si are in general agreement with the observations on the paramagnetic center. Muonium has been found to diffuse very rapidly in Si,[61] in agreement with the low barriers found in our total-energy surface for $H^0$. Recently, "anomalous muonium" has been unambiguously identified as occupying a bond center,[20] in agreement with the global minimum that emerges from our calculations for $H^0$. The so-called "normal muonium" is usually associated with the tetrahedral interstitial site.[61] Our energy surface for $H^0$ shows that the $T$ site is not a stable site. The bond-center site is the only local minimum in this surface (to an accuracy of $\sim 0.1$ eV; a barrier of 0.1 eV would, however, be far too small to confine the muon anyway, given its large zero-point motion). However, other locations around $T$ (in the low-density region of the crystal) may account for the observed signal, with the muonium tunneling rapidly between different sites. Such sites, while not being global minima of the energy surface, are the only locations accessible to the muon which do not require the large relaxations of the Si host atoms necessary for a bond-center position.[62] On the time scale of the muon lifetime, such relaxations may be sufficiently slow to effectively trap the muon in the low-density regions of the crystal, where relaxation of the host atoms is negligible.

These observations lead us to the following remarks. Our calculated results and energy surfaces correspond to zero temperature, and a static approximation; the mass of the particles does not enter into this description. At finite temperatures, phonon displacements of the Si atoms will create a continuously varying potential environment for the hydrogen atom; its insertion into the bond center, and diffusion along the migration paths shown above, will necessarily be coupled to the motion of the Si atoms. Even at zero temperature, the zero-point motion of the very light H atom will have significant amplitude. In principle, the total energy surfaces and information about relaxation obtained above can form the basis of an analysis in which the quantum nature of the particle is taken into account. We do not address this issue any further here.

## IV. RESULTS FOR INTERACTIONS OF SEVERAL HYDROGEN ATOMS

### A. $H_2$ molecules

First, we examine how two neutral H atoms may combine and form a $H_2$ molecule in the Si crystal. We have found the minimum-energy position for the molecule straddling the tetrahedral interstitial site, oriented in the $\langle 100 \rangle$ direction, with the atoms separated by 0.86 Å (to be compared with 0.75 Å in vacuum). This configuration is illustrated in Fig. 12. At the hexagonal interstitial site, which would lie on a migration path, the energy of the molecule is 1.1 eV higher. The binding energy of $H_2$ (as compared with isolated neutral H atoms at their lowest interstitial position, i.e., at the bond center) is $2 \pm 0.5$ eV per molecule, or $\sim 1$ eV per atom. This binding energy applies to the case where $H_2$ is formed out of two isolated neutral H atoms. If instead the molecule were formed out of one $H^0$ and one $H^+$ (a possibility suggested by Johnson and Herring[58]), the binding energy would be lowered by the energy difference between $H^+$ and $H^0$. From Fig. 11, we see that in $p$-type material this difference can be up to 1 eV. If the molecule is formed out of (or dissociated into) $H^0$ and $H^+$, this result may explain the observation of Johnson and Herring that the binding energy of the molecule is lower than the diffusion barrier.

### B. H-induced defects

Another phenomenon that involves the cooperative interaction of several H atoms with the Si lattice is related to the recent observation[15,63] that hydrogenation can induce microdefects in a region within $\sim 1000$ Å from the surface. Care was taken to eliminate radiation damage that could result from direct exposure to the plasma during hydrogenation. The defects, studied with transmission electron microscopy (TEM), have the appearance of platelets along $\{111\}$ crystallographic planes, range in size from 50 to 100 Å, and exhibit no net Burgers vector. They cannot be categorized as intrinsic Si defects, such as dislocation loops or stacking faults. Some elastic-strain contrast was observed around the defects. The thickness of the platelets is comparable to a single $\{111\}$ Si plane. By correlating the density of platelets with the deuterium concentration, one or two H atoms per Si—Si bond are present. Furthermore, Raman measurements[15] showed spectral features at 1960 and 2100 $cm^{-1}$, which were at-

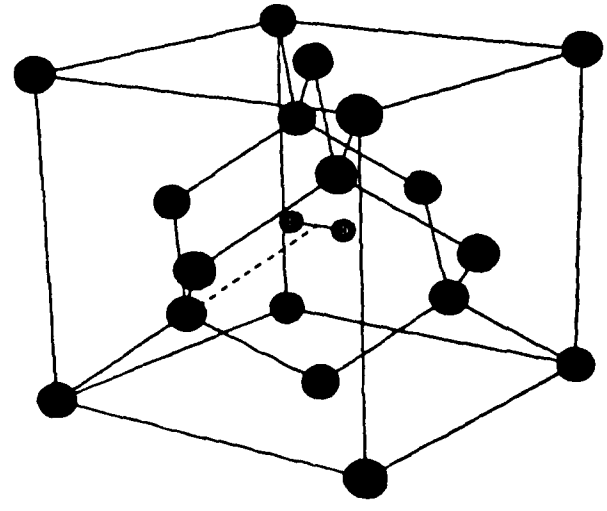


FIG. 12. Schematic illustration of the minimum-energy position of a $H_2$ molecule in the Si crystal: located at the tetrahedral interstitial site and oriented along $\langle 100 \rangle$.

tributed to H incorporated in the Si crystal.

We have examined several possibilities (some of which were mentioned in Ref. 63) for the structure of these platelets, by performing total-energy calculations in a superlattice geometry; edge effects at the platelet boundary are thus neglected. First, we explore the situation in which one H atom is inserted into each of the Si—Si bonds of a {111} plane, as schematically illustrated in Fig. 13(a). Recalling that H in the bond-center position requires large relaxations of the neighboring Si atoms, it might be expected that the presence of H in a particular bond center would favor the introduction of other H in nearby bond centers. As discussed in Sec. III E, the vibrational frequency associated with such a bond-center configuration (1945 $cm^{-1}$ for $H^0$) is close to the experimentally observed frequencies (1960 and 2100 $cm^{-1}$); the Raman measurements alone are therefore not sufficient to exclude a model in which H is bonded not to one, but two Si atoms. However, we can eliminate this model as a candidate for the defects by inspecting the total energy. The problem is now two dimensional; we assume that no in-plane relaxation occurs. The first plane of Si atoms near the bond center moves out over 0.45 Å; the second plane relaxes by 0.09 Å. These values are very similar to those obtained for relaxation near a single bond-centered H. For the relaxed configuration we find that the energy per H is more than 0.5 eV higher than it is for the isolated impurity, i.e., the formation of this type of extended defect is clearly unfavorable.

Another possibility for extended defect formation is the insertion of *two* H atoms in each Si—Si bond, i.e., the formation of two Si—H bonds out of each Si—Si bond. It is essential to place the H atoms off the Si-Si axis in order to find a favorable configuration, as illustrated in Fig. 13(b). A representative position is for the H atoms at two *M* sites associated with each Si—Si bond. The energy per H atom is now similar to that for isolated atoms. However, this indicates that this structure would be unstable to $H_2$ molecule formation. We conclude that these proposed configurations are energetically not favorable.

We have therefore examined a different type of mechanism, based on the removal of Si atoms from the defect region, with the resulting dangling bonds tied off by H atoms. This mechanism is based on our calculated result that H atoms can assist Frenkel-pair creation. In a perfect crystal the creation of a Frenkel pair (vacancy-interstitial pair) normally costs about 8 eV.[64] If, however, a sufficient number of H atoms are available in the immediate neighborhood of a particular Si atom, Frenkel-pair formation can actually be exothermic with a slight gain of energy. In the final configuration a self-interstitial is emitted while four H atoms saturate the dangling bonds of the vacancy. The calculated energy gain for the process in which a neutral interstitial H atom passivates a dangling bond is ~2.2 eV per Si—H bond.[65] This value is obtained by comparing the total energy of a fully saturated vacancy (i.e., four H atoms tying off the dangling bonds) with the sum of the energies of (a) a vacancy in which only three dangling bonds are saturated by H, and (b) an isolated $H^0$ at its most favorable site in the lattice. This energy value was confirmed in a superlattice calculation modeling an extended defect in which a double row of Si atoms was removed in a {111} plane, with all dangling bonds tied off by H, as illustrated in Fig. 13(c). The energy gain per Si—H bond is equal to that calculated at a single vacancy.

These theoretical results for the interaction of several H atoms lead us to the following conclusions. On the basis of energetic considerations, $H_2$ molecules are the

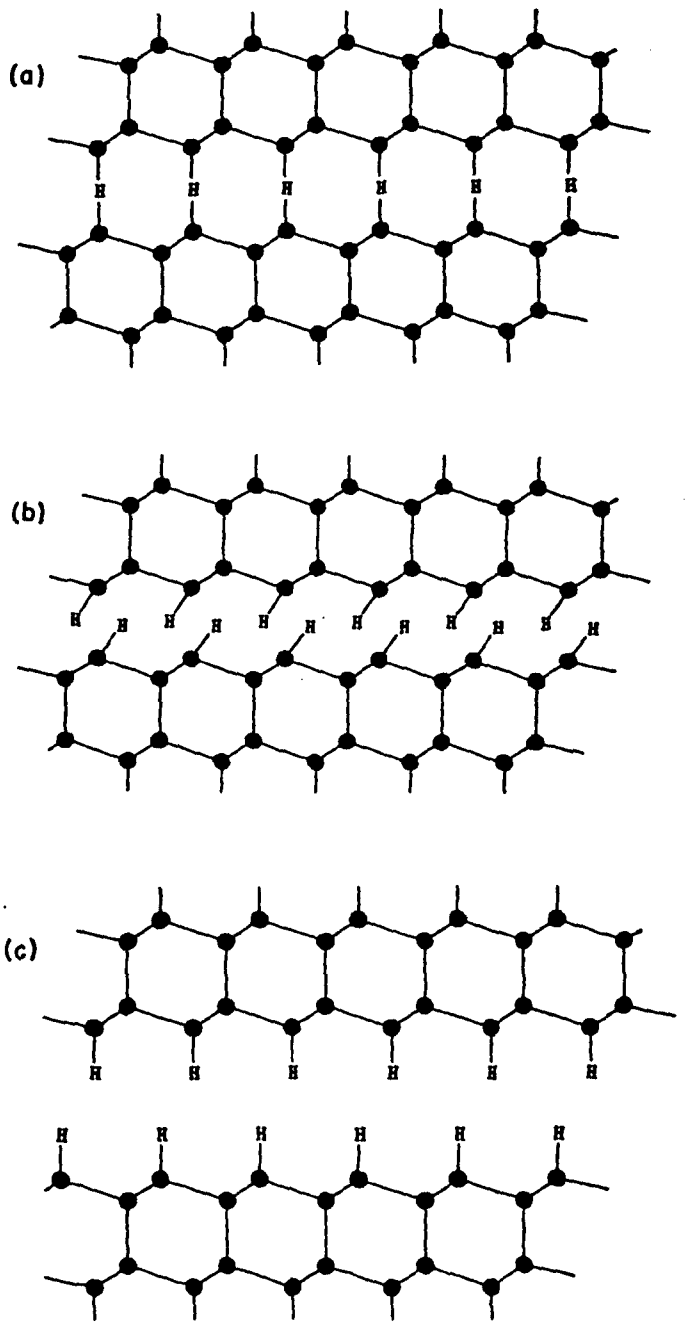



FIG. 13. Schematic illustration of possible structures for H-induced defects ("platelets") in crystalline Si. In (a) single H atoms are situated in the bond centers of [111] Si—Si bonds. In (b) each Si—Si bond is replaced by two Si—H bonds, with the H located off axis. In (c) a double layer of Si atoms has been removed through a vacancy-formation process described in the text, and the resulting dangling bonds are tied off by H.

preferred state for several neutral H atoms in pure crystalline Si. Kinetic considerations also suggest that H-assisted Frenkel-pair creation would be a rare event. However, H-assisted ejection of threefold- or twofold-coordinated Si atoms is kinetically more favorable, such that enlargement of a preexisting defect is likely. The particular atomistic processes that lead to defect nucleation and enlargement cannot be described in more detail at this point; however, the energetic arguments given above for defect formation and extension suggest the vacancy-formation mechanism is likely to be involved in the observed hydrogen-induced damage.

## V. SUMMARY

This work provides a comprehensive description of the diffusion and reactions of H in crystalline Si, based on the first-principles pseudopotential-density-functional method. Hydrogen as an impurity shows distinctly different behavior as a function of its charge state, as exemplified by the total-energy surfaces that we generated (Figs. 3–5). $H^+$ and $H^0$ prefer the high-electron-density regions of the crystal, with a global energy minimum at the bond-center site. $H^-$ prefers the low-electron-density region and has its lowest energy at the tetrahedral interstitial site. The vibrational frequencies for the H stretching modes at the bond center (1945 $cm^{-1}$ for $H^0$ and 2210 $cm^{-1}$ for $H^+$) are very close to measured frequencies for single H atoms in *a*-Si:H or crystalline Si.

The stability of different charge states depends on the Fermi-level position: $H^+$ is favored in *p*-type material, providing a straightforward mechanism for passivation of *p*-type Si through compensation and subsequent pair formation. The calculations for *n*-type material produce $H^-$ as the stable charge state, and indicate hydrogen would be a negative-*U* impurity, but within the error bar $H^0$ cannot be excluded.

$H_2$ molecules are the most stable state for H in crystalline Si in the absence of other defects. Hydrogen can also induce defects; we have discussed a mechanism for extended defect formation through spontaneous Frenkel-pair generation.

## ACKNOWLEDGMENTS

We appreciate helpful discussions with C. Herring, N. M. Johnson, and R. F. Kiefl. This work was supported in part by the U.S. Office of Naval Research under Contract No. N00014-84-C-0396.

*Present address: Philips Research Laboratories, 345 Scarborough Road, Briarcliff Manor, NY 10510.

[1]S. J. Pearton, J. W. Corbett, and T. S. Shi, Appl. Phys. A **43**, 153 (1987).

[2]C. T. Sah, J. Y. C. Sun, and J. J. T. Tzou, Appl. Phys. Lett. **43**, 204 (1983); J. Appl. Phys. **54**, 5864 (1983).

[3]J. I. Pankove, D. E. Carlson, J. E. Berkeyheiser, and R. O. Wance, Phys. Rev. Lett. **51**, 2224 (1983); J. I. Pankove, R. O. Wance, and J. E. Berkeyheiser, Appl. Phys. Lett. **45**, 1100 (1984); J. I. Pankove, P. J. Zanzucchi, C. W. Magee, and G. Lucovsky, *ibid.* **46**, 421 (1985).

[4]N. M. Johnson, Phys. Rev. B **31**, 5525 (1985).

[5]N. M. Johnson, C. Herring, and D. J. Chadi, Phys. Rev. Lett. **56**, 769 (1986); **59**, 2116 (1987); N. M. Johnson and C. Herring, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 277.

[6]K. Bergman, M. Stavola, S. J. Pearton, and J. Lopata, Phys. Rev. B **37**, 2770 (1988); in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 281.

[7]N. M. Johnson, Appl. Phys. Lett. **47**, 874 (1985).

[8]J. I. Pankove, C. W. Magee, and R. O. Wance, Appl. Phys. Lett. **47**, 748 (1985).

[9]A. J. Tavendale, D. Alexiev, and A. A. Williams, Appl. Phys. Lett. **47**, 316 (1985).

[10]C. H. Seager, R. A. Anderson, and J. K. G. Panitz, J. Mater. Res. **2**, 96 (1987).

[11]A. Van Wieringen and N. Warmoltz, Physica **22**, 849 (1956).

[12]B. Bech Nielsen, Phys. Rev. B **37**, 6353 (1988).

[13]A. Schnegg, H. Prigge, M. Grundner, P. O. Hahn, and H. Jacob, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 291; it should be noted that more recent work has indicated that impurities other than H may be involved in the observed phenomena.

[14]N. M. Johnson and M. D. Moyer, Appl. Phys. Lett. **46**, 787 (1985).

[15]N. M. Johnson, F. A. Ponce, R. A. Street, and R. J. Nemanich, Phys. Rev. B **35**, 4166 (1987).

[16]S. T. Pantelides, Appl. Phys. Lett. **50**, 995 (1987).

[17]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **60**, 2761 (1988).

[18]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 253.

[19]B. D. Patterson, Rev. Mod. Phys. **60**, 69 (1988).

[20]R. F. Kiefl, M. Celio, T. L. Estle, S. R. Kreitzman, G. M. Luke, T. M. Riseman, and E. J. Ansaldo, Phys. Rev. Lett. **60**, 224 (1988).

[21]S. F. J. Cox and M. C. R. Symons, Chem. Phys. Lett. **126**, 516 (1986).

[22]V. A. Gordeev, Yu. V. Gorelkinskii, R. F. Konopleva, N. N. Nevinnyi, Yu. V. Obukhov, and V. G. Firsov (unpublished); Yu. V. Gorelkinskii and N. N. Nevinnyi, Pis'ma Zh. Tekh. Fiz. **13**, 105 (1987) [Sov. Tech. Phys. Lett. **13**, 45 (1987)].

[23]V. A. Singh, C. Weigel, J. W. Corbett, and L. M. Roth, Phys. Status Solidi B **81**, 637 (1977).

[24]C. O. Rodriguez, M. Jaros, and S. Brand, Solid State Commun. **31**, 43 (1979).

[25]A. Mainwood and A. M. Stoneham, J. Phys. C **17**, 2513 (1984).

[26]W. E. Pickett, M. L. Cohen, and C. Kittel, Phys. Rev. B **20**,

5050 (1979).
[27]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. 62, 1884 (1989).
[28]H. Katayama-Yoshida and K. Shindo, Phys. Rev. Lett. 51, 207 (1983).
[29]J. W. Corbett, S. N. Sahu, T. S. Shi, and L. C. Snyder, Phys. Lett. 93A, 303 (1983).
[30]N. Sahoo, K. C. Mishra, and T. P. Das, Hyperfine Interact. 32, 601 (1986).
[31]S. Estreicher, Phys. Rev. B 36, 9122 (1987).
[32]P. Deák. L. C. Snyder, J. L. Lindström, J. W. Corbett, S. J. Pearton, and A. J. Tavendale, Phys. Lett. A 126, 427 (1988).
[33]P. Deák, L. C. Snyder, R. K. Singh, and J. W. Corbett, Phys. Rev. B 36, 9612 (1987); P. Deák and L. C. Snyder, *ibid.* 36, 9619 (1987).
[34]R. M. Martin, in *Festkörperprobleme (Advances in Solid State Physics)*, edited by P. Grosse (Vieweg, Braunschweig, 1985), Vol. XXV, p. 3.
[35]P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964); W. Kohn and L. J. Sham, *ibid.* 140, A1133 (1965); exchange and correlation potentials are based on the data from D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. 45, 566 (1980), as parametrized by J. Perdew and A. Zunger, Phys. Rev. B 23, 5048 (1981).
[36]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. 43, 1494 (1979).
[37]J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C 12, 4409 (1979).
[38]A. Baldereschi, Phys. Rev. B 7, 5212 (1973); D. J. Chadi and M. L. Cohen, *ibid.* 8, 5747 (1973); H. J. Monkhorst and J. D. Pack, *ibid.* 13, 5188 (1976).
[39]P. J. H. Denteneer, Ph.D. thesis, Eindhoven University of Technology, 1987, available from the author upon request.
[40]Following the methodology of Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B 30, 1844 (1984).
[41]P. O. Löwdin, J. Chem. Phys. 19, 1396 (1951).
[42]P. N. Keating, Phys. Rev. 145, 637 (1966); G. A. Baraff, E. O. Kane, and M. Schlüter, Phys. Rev. B 21, 5662 (1980).
[43]A. Zunger, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull (Academic, New York, 1986), Vol. 36, pp. 275–464.
[44]F. Beeler, Ph.D. thesis, University of Stuttgart, 1986.
[45]D. Vanderbilt and J. D. Joannopoulos, Phys. Rev. B 27, 6311 (1983).
[46]G. G. DeLeo, M. J. Dorogi, and W. B. Fowler, Phys. Rev. B 38, 7520 (1988).
[47]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, the following paper, Phys. Rev. B 39, 10 809 (1989).
[48]M. Altarelli and W. Hsu, Phys. Rev. Lett. 43, 1346 (1979).
[49]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. 52, 1814 (1984).
[50]M. H. Brodsky, M. Cardona, and J. J. Cuomo, Phys. Rev. B 16, 3556 (1977).
[51]H. J. Stein, Phys. Rev. Lett. 43, 1030 (1979).
[52]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Phys. Rev. B 37, 8313 (1988).
[53]M. Stutzmann, Phys. Rev. B 35, 5921 (1988).
[54]A. D. Marwick, G. S. Oehrlein, and N. M. Johnson, Phys. Rev. B 36, 4539 (1987).
[55]B. B. Nielsen, J. U. Andersen, and S. J. Pearton, Phys. Rev. Lett. 60, 321 (1988).
[56]N. M. Johnson and C. Herring, in *Proceedings of the Third International Conference on Shallow Impurities in Semiconductors*, Linköping, 1988, IOP Conf. Ser. (IOP, London, 1989).
[57]K. J. Chang and D. J. Chadi, Phys. Rev. Lett. 60, 1422 (1988).
[58]N. M. Johnson and C. Herring, Phys. Rev. B 38, 1581 (1988).
[59]N. M. Johnson and C. Herring, in *Proceedings of the 15th International Conference on Defects in Semiconductors*, Budapest, 1988 (Trans Tech, Aedermannsdorf, 1989).
[60]S. T. Picraux and F. L. Vook, Phys. Rev. B 18, 2066 (1978).
[61]B. D. Patterson, E. Holzschuh, R. F. Kiefl, K. W. Blazey, and T. L. Estle, Hyperfine Interact. 17-19, 599 (1984).
[62]G. D. Watkins, in *Proceedings of the 15th International Conference on Defects in Semiconductors*, Budapest, 1988 (Trans Tech, Aedermannsdorf, 1989).
[63]F. A. Ponce, N. M. Johnson, J. C. Tramontana, and J. Walker, in *Proceedings of the Microscopy of Semiconducting Materials Conference*, Inst. Phys. Conf. Ser. No. 87, edited by A. G. Cullis (Hilger, London, 1987), p. 49.
[64]Y. Bar-Yam and J. D. Joannopoulos, J. Electron. Mater. 14a, 261 (1985); R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, *ibid.* 14a, 269 (1985).
[65]In contrast, J. M. Baranowski and J. Tatarkiewicz [Phys. Rev. B 35, 7450 (1987)] proposed that the interstitial H be put on the backside of the dangling bond. We find that the energy gain for such an arrangement is only 1.3 eV.

PHYSICAL REVIEW B VOLUME 39, NUMBER 15 15 MAY 1989-II

# Microscopic structure of the hydrogen-boron complex in crystalline silicon

P. J. H. Denteneer,* C. G. Van de Walle,† and S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 21 November 1988)

The microscopic structure of hydrogen-boron complexes in silicon, which result from the passivation of boron-doped silicon by hydrogen, has been extensively debated in the literature. Most of the debate has focussed on the equilibrium site for the H atom. Here we study the microscopic structure of the complexes using parameter-free total-energy calculations and an exploration of the entire energy surface for H in Si:B. We conclusively show that the global energy minimum occurs for H at a site close to the center of a Si—B bond (*BM* site), but that there is a barrier of only 0.2 eV for movement of the H atom between four equivalent *BM* sites. This low energy barrier implies that at room temperature H is able to move around the B atom. Other sites for H proposed by others as the equilibrium sites are shown to be saddle points considerably higher in energy. The vibrational frequency of the H stretching mode at the *BM* site is calculated and found to be in agreement with experiment. Calculations of the dissociation energy of the complex are discussed.

## I. INTRODUCTION

The role that hydrogen plays in semiconductors has become the subject of intense research[1,2] following the discovery that hydrogen is able to passivate the electrical activity of shallow acceptors in silicon. This passivation effect is of considerable importance for technological reasons. The properties of electronic devices are largely determined by the presence and activity of shallow impurity levels and passivation of their activity by omnipresent (accidentally or intentionally) hydrogen would alter the properties of those devices in an uncontrollable way as long as the passivation mechanism is not thoroughly understood. The passivation effect was first suggested by Sah *et al.*[3] in an inventive analysis of experiments on metal-oxide-semiconductor (MOS) capacitors. The connection between hydrogen and boron (as the prototypical acceptor-type impurity) concentrations was soon established in studies of the passivation effect under controlled experimental conditions by Pankove *et al.*[4] and Johnson.[5] This discovery supplemented the understanding of the role of hydrogen in semiconductors, which was previously known to be the saturation of dangling bonds at defects, surfaces, and interfaces, or passivation of *deep* levels in the energy gap, e.g., those due to transition-metal impurities. At first, the passivation effect was found to be considerably smaller in case of silicon doped with donor-type impurities (*n* type).[6] Recently, however, it was found that also in *n*-type material there is a strong passivation effect, although still not as strong as in *p*-type material.[7]

A large number of experiments was performed to elucidate the fundamental reactions underlying the passivation mechanism and they generally claimed to support each other. For some time, however, the analysis of these experiments contained contradictory assumptions regarding the charge state of H. A step forward in the understanding of the passivation mechanism was made in Ref. 8, in which one of the present authors suggested that hydrogen is a deep donor in silicon and was able to account for a large portion of the experimental observations. Assuming that H is a deep donor in Si, passivation in *p*-type material would come about in two steps: (1) compensation, i.e., the annihilation of free holes associated with the ionized acceptors by the electrons of the H atoms, and (2) formation of a neutral complex (or pair) out of a negatively charged acceptor and a positively charged H atom. We stress that the first step already establishes passivation and that the second step is only the logical consequence of the first step. On the basis of first-principles total-energy calculations, Van de Walle *et al.*[9] conclusively showed that H indeed acts as a donor in *p*-type material, confirming the proposed passivation mechanism. This conclusion could be reached from calculations for H in different charge states in pure Si. Questions pertaining to the nature and quantitative properties of the hydrogen-acceptor complex were not addressed in that work.

Soon after the hydrogen-acceptor complexes were discovered, a controversy arose regarding their microscopic structure. Pankove *et al.*,[4] on the basis of infrared spectroscopy of boron-doped Si (Si:B), proposed that H would be inserted in a Si—B bond with the substitutional B pushed out toward the plane of three neighboring Si atoms. This configuration was confirmed in theoretical calculations by DeLeo and Fowler,[10] who used a semiempirical cluster method. These authors also reproduced the measured vibrational frequency of the H stretching mode. However, Assali and Leite,[11] using a method very similar to the one DeLeo and Fowler employed, proposed a site for the H atom on the extension of a Si—B bond, the so called antibonding site. Using a spring-constant model they too were able to reproduce the measured H vibrational frequency, although DeLeo and Fowler[12] found a very different frequency if H were to be at the antibonding site. Based on tight-binding–model calculations for the hydrogenated vacancy in pure Si, Baranowski and Tatarkiewicz[13] speculated

that H would occupy a site on the extension of a B—Si bond (backbonding site), forming a Si($p$)—H($s$) bond. Hartree-Fock cluster calculations were used by Amore Bonapasta *et al.*,[14] who found a position near the center of a Si—B bond as the equilibrium site for H.

Experimental investigations into the microscopic structure of hydrogen-acceptor complexes (in which the acceptor usually is boron) have included infrared measurements and Raman studies of the H vibrational frequency,[4,5,15–17] ion-channeling measurements of the lattice location of H and the displacement from the substitutional site of B,[18–21] the perturbed-angular-correlation technique to explore hydrogen-indium pairs in Si,[22] x-ray-diffraction studies of the lattice relaxation due to passivation,[23] and uniaxial-stress studies of the H-stretching mode.[24] Generally, the picture emerges from these studies that H dominantly occupies a site near the center of a Si—B bond, although smaller percentages are seen to reside at antibonding or tetrahedral interstitial sites.[19,20,22] The latter observations, however, could also be connected with damage induced by H. The vibrational frequency of the H-stretching mode is found to be 1903 $cm^{-1}$ for low temperatures[16,17] ($I \sim 5$ K). We will discuss some of the results in these papers in more detail in Sec. III, where the theoretical results of the present paper are given.

In previous theoretical work[10–14,25,26] only a limited set of possibilities for the equilibrium site of the H atom was considered. Since it is to be expected that *anytime* the H atom is located close to the B atom it will remove the electrically active level from the gap, it is necessary to study the entire total-energy surface for H in B-doped Si in order to determine the favored site. Furthermore, since the energy differences between configurations in which H occupies different sites are small, there is a need for accurate calculations of such energy differences. Most of the theoretical approaches above use either a cluster model, usually without studying the effect of enlarging the cluster or the effect of terminating the cluster in different ways, and/or semiempirical Hamiltonians that contain a number of parameters that have been fitted to reproduce the properties of *molecules*. If tests are performed one invariably finds (see, e.g., Ref. 25) that these methods are unable to reproduce the properties of even simple bulk semiconducting crystals. When the techniques are used for small clusters to simulate defects in crystals, quite often some of the results are in agreement with either experiment or more sophisticated calculations. Typically, however, other results may be in serious error. In general, the lack of tests of convergence and accuracy renders most predictions of such calculations as questionable. In this work, we use a parameter-free method of calculating total energies, the pseudopotential-density-functional method (see Sec. II), which has proven to be very reliable in calculating and predicting properties of a wide variety of semiconducting systems, such as bulk solids, surfaces, interfaces, and localized and extended defects. Furthermore, we test all of our results for convergence and accuracy with respect to numerical approximations involved. Finally, we have developed a way to visualize the entire energy surface for a H interstitial atom in B-doped Si similar to the method used by some of the present authors in a study of H in pure Si.[9]

The remainder of the paper is organized as follows: In Sec. II we discuss calculational details of our method that are especially pertinent to the present study, as well as tests of how the results depend on the inevitable numerical approximations involved. In Sec. III the results of our approach are presented and compared with available experimental data. Finally, we summarize the paper in Sec. IV.

## II. CALCULATIONAL DETAILS

The Hamiltonian in the Kohn-Sham equations[27] for the valence electrons in a crystal is constructed using norm-conserving pseudopotentials[28] to describe the interaction between atomic cores (nuclei plus core electrons) and valence electrons. For the exchange and correlation interaction we use the local-density approximation (LDA) to the exchange and correlation functional that was parametrized by Perdew and Zunger[29] from the Monte Carlo simulations of an electron gas by Ceperley and Alder.[30]

We solve the Kohn-Sham equations by expanding all functions of interest (one-electron wave functions, potentials, etc.) in plane waves and solving the resulting matrix eigenvalue problem. This procedure is iterated until a self-consistent solution is obtained, i.e., until the effective potential for the valence electrons that enters the Hamiltonian equals the effective potential that is calculated from the wave functions that are solutions for this Hamiltonian. From the self-consistent one-electron energies and wave functions the total energy of the crystal is most conveniently calculated in momentum space.[31,32] This pseudopotential-density-functional method is a "first-principles" method in that it contains no adjustable parameters derived from experiment. This method has been very successful in calculating and predicting the ground-state properties of a wide variety of semiconducting systems.[33]

We calculate the total energy for a silicon crystal with a substitutional boron atom and an interstitial hydrogen atom for a large number of inequivalent sites of the H atom. For every position of the H atom that we consider, the atoms of the Si:B host crystal are allowed to relax by minimizing the total energy with respect to the host-crystal atomic coordinates. Relaxations up to second-nearest neighbors are investigated as to their importance.

As the method in general is well documented, we will discuss only the calculational details that are especially pertinent to the present study.

### A. Norm-conserving pseudopotentials

For Si and B norm-conserving pseudopotentials are generated according to the scheme of Ref. 28. We use the degrees of freedom that one has in generating such pseudopotentials to our advantage by carefully choosing core cutoff radii $r_c$ (outside of which true and pseudo-wave-functions are identical[28]). These cutoff radii can be chosen such that a pseudopotential is generated whose Fourier transform converges more rapidly in $q$ space, implying that a smaller number of plane waves will be re-

quired to describe the pseudopotential.[34] Generally, moving $r_c$ outward improves the pseudopotential in the above respect. However, moving $r_c$ outward deteriorates the description of the atom by the pseudopotential. Cutoff radii are chosen such that a reasonable balance between both effects is found. The Si pseudopotential is the same as used in previous work and is described elsewhere.[9,35] The pseudopotential for B is newly generated and is discussed here in more detail. We generate pseudopotentials for angular-momentum components $l=0$ and 1 only. The cutoff radii for $l=0$ and 1 are 1.10 and 1.18 a.u., respectively. These $r_c$ are somewhat larger than those used in Ref. 36 (1.0 and 0.9 a.u. for $l=0$ and 1, respectively). The generated pseudopotential is tested by calculating the equilibrium lattice constant $a_{eq}$ and bulk modulus $B_0$ of boron phosphide (BP) in the zinc-blende structure for consecutively larger values of the kinetic-energy cutoffs $E_1$ and $E_2$, which determine the numbers of plane waves in the expansion of the wave functions (plane waves with kinetic energy up to $E_2$ are included in the calculation, those between $E_1$ and $E_2$ in second-order Löwdin perturbation theory;[37] we invariably choose $E_2=2E_1$). In the following, we will use the notation $(E_1;E_2)$ to denote the choice of cutoffs. The calculations are performed both for the newly generated B pseudopotential as well as for the one that is tabulated in Ref. 38. For phosphorus we use in both cases the tabulated pseudopotential of Bachelet, Hamann, and Schlüter[38] (to be called the BHS pseudopotential). The Fourier transform of the P pseudopotential falls off more rapidly for large $q$ than the Fourier transform of the B pseudopotential. Therefore the convergence with respect to kinetic-energy cutoff will be determined by the B pseudopotential. For each choice of energy cutoffs, $a_{eq}$ and $B_0$ are calculated by computing the total energy of BP at five lattice constants ranging between $-5\%$ and $+5\%$ of the experimental lattice constant.[39] The results are fitted to Murnaghan's equation of state for solids, which contains $a_{eq}$ and $B_0$ as parameters.[40]

We combine the results for $a_{eq}$ and $B_0$ in Fig. 1. The single points in Fig. 1 ($a_{eq}=4.56$ Å and $B_0=1.66$ Mbar) are results obtained in Ref. 36 using a pseudopotential for B and P very much like the BHS pseudopotential and an energy cutoff of 20 Ry (no Löwdin perturbation technique was used in their calculation). Our results indicate that the results of Ref. 36 have not entirely converged with respect to increasing the energy cutoff. The main conclusion to be drawn from Fig. 1 is that the newly generated B potential results in virtually the same $a_{eq}$ and $B_0$ as found with the BHS pseudopotential, but that it converges faster to these values than with the BHS pseudopotential. Both converged values for $a_{eq}$ (4.48 and 4.49 Å for the new and BHS pseudopotential, respectively) are in fair agreement with the lattice constant of 4.538 Å that is found experimentally.[41] The calculated bulk moduli of 1.62 and 1.68 Mbar for the new and BHS pseudopotential, respectively, cannot be compared with any experimental result. Therefore, we have reached our goal of generating a norm-conserving pseudopotential that can be represented by fewer plane waves than the one so far available, while it still accurately describes a B atom in a solid-state environment.

To illustrate the point that the cutoff radii $r_c$ cannot be pushed out too far, we mention that the converged result for $a_{eq}$ using a potential for B generated by choosing the $r_c$ to lie at radii for which the outermost maxima of the radial wave function for the respective $l$ values occur ($r_c=1.52$ and 1.56 a.u. for $l=0$ and 1, respectively) is 4.34 Å. The percentage of deviation from the experimental value is more than 3 times as large as for the two other pseudopotentials.

For hydrogen we did not use a pseudopotential, although it is possible to generate one. Instead we use the

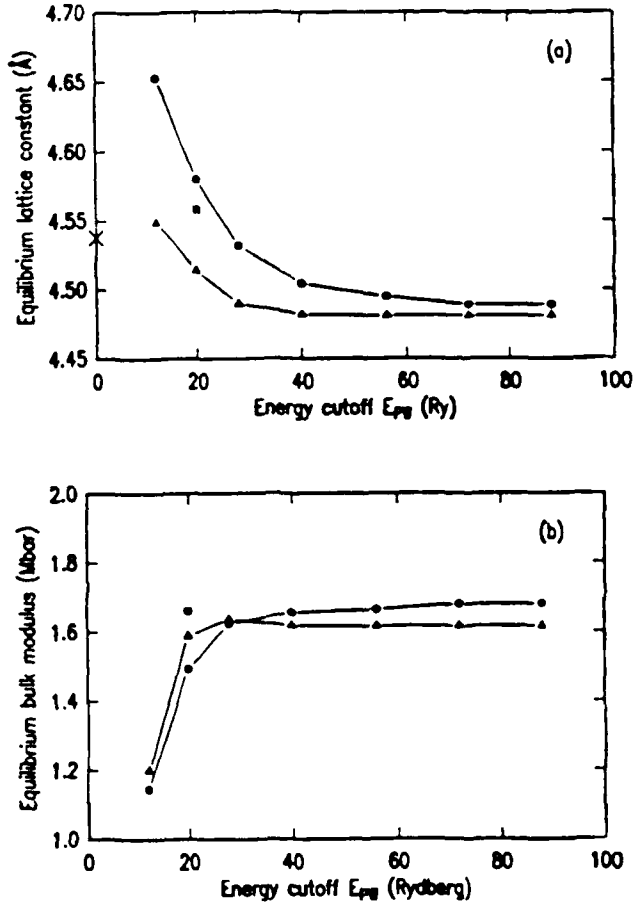



FIG. 1. Convergence of ground-state properties of BP as a function of kinetic energy cutoff $E_{pw}$ (determining the number of plane waves in the expansion of the wave functions) for two different pseudopotentials for boron. The dots represent results obtained using the tabulated pseudopotentials for B and P from Ref. 38, whereas the triangles represent results obtained using a newly generated pseudopotential for B and the tabulated pseudopotential from Ref. 38 for P. The solid squares represent results obtained in Ref. 36 using pseudopotentials for B and P very similar to the pseudopotentials in Ref. 38. Plane waves with kinetic energy up to $\frac{1}{2}E_{pw}$ are included exactly in the calculation, and those between $\frac{1}{2}E_{pw}$ and $E_{pw}$ in second-order perturbation theory (Ref. 37). (a) Equilibrium lattice constant $a_{eq}$ of BP (in Å). The cross on the vertical axis denotes the experimental lattice constant (Ref. 41). (b) Equilibrium bulk modulus $B_0$ of BP (in Mbar).

exact $1/r$ Coulomb potential of the proton. In this we follow our earlier work[9,35] and we refer to those papers for a more detailed discussion.

We note that Fig. 1 is not instrumental in determining the energy cutoffs that will be sufficient for the problems to be addressed in this paper. Those cutoffs depend on the properties and accuracy one is interested in and can only be determined by explicitly calculating those properties for consecutively larger cutoffs. This will be discussed in more detail in Sec. II D. Figure 1 *does* show qualitatively that these properties may be obtained at lower cutoffs by using the newly generated B potential as compared to the (standard) BHS pseudopotential.

### B. Supercells

To model simple and complex defects we use supercells that are periodically repeated. We investigate how calculated properties depend on supercell size and we determine when they become independent of supercell size (within a desired accuracy). As in previous work[9,35,42] we use supercells of 8, 16, and 32 atoms in which defects are separated by 5.43, 7.68, and 9.41 Å, respectively.

In addition to the finite separation between defects, another artifact particularly pertinent to defect calculations in general arises from using a (finite-size) supercell. Defect levels that show no dispersion for a truly isolated defect do have dispersion when using finite-size supercells. This is, however, not a big problem in the present calculation. The substitutional B and interstitial H atoms together exactly supply the four valence electrons of the Si atom that has been replaced by the substitutional B atom. Therefore an equal number of bands is filled as in the case of pure Si. Therefore, a H-related defect level, which is found to be located in the energy gap exactly as in the case of H in pure Si (See Ref. 35 and also Sec. III A) is unoccupied. Even if a large dispersion of this level causes it to drop into the valence bands for certain points in the first Brillouin zone (1BZ), the level can be left unoccupied when it is properly identified [this identification can be done in a variety of ways: (1) the charge density associated with the defect level is localized and correlated with the position of H; (2) by comparing the band structure of Si with a substitutional B atom (Si:B) with and without the H atom; (3) the H-related defect level will move significantly with respect to the other bands if the band structure is calculated with the H atom at a different position].

The dispersion of the H-related defect level for H in Si:B is about 2.0, 1.1, and 0.6 eV for the 8-, 16-, and 32-atom cells, respectively. See Sec. III A for a further discussion of these levels.

### C. Brillouin-zone integrations

In two distinct stages of the calculation of the total energy, an integration over the 1BZ has to be performed: (1) calculation of the valence charge density from the one-electron wave functions, and (2) calculation of the band-structure energy term from the one-electron energies.[32] Both integrations are replaced by summations over special k points in the irreducible part of the 1BZ (IRBZ).[43,44] It has been established in many calculations that by using only a very small number of k points (between 1 and 10) very accurate total-energy *differences* can be obtained. In general, one has to test for every application how many k points are sufficient for a certain accuracy. Such tests are reported below.

We employed the general Monkhorst-Pack (MP) scheme[44] to generate special points sets with their parameter $q$ equal to 2. The number of special points generated with this choice of $q$ depends on the position of the H atom in the unit cell. It is also different for the different supercell sizes that we use. When H is located at a general position on the extension of a Si—B bond, $q=2$ results in two, five, and two special points for the 8-, 16-, and 32-atom cell, respectively. For less symmetric H positions this number can be as high as 16 in the 16-atom cell and 4 in the 32-atom cell. The following test was executed to determine the accuracy that is obtained with the $q=2$ choice for special points in the MP scheme: We calculate the total-energy difference between configurations in which H occupies a position near the center of a Si—B bond and one in which H is located on the extension of a Si—B bond. These two reference configurations are defined only for the purpose of carrying out meaningful tests of the Brillouin-zone integrations (this subsection) and the dependence of results on supercell size and basis-set size (next subsection). They should not be confused with the fully relaxed configurations that will be described later. In the first configuration [to be called the bond-minimum (*BM*) reference configuration] the H atom and the Si and B atoms constituting the bond in which H is located are allowed to relax their position in order to find the minimum-energy configuration. In this *BM* reference configuration the Si and B atoms relax outward by 0.24 and 0.42 Å, respectively. In the second configuration [to be called the antibonding (*AB*) reference configuration] only the H and B atoms are relaxed. In this configuration the H atom has a distance of 1.32 Å from the B atom, which relaxes inward (away from H and towards a Si atom) by 0.09 Å. The relaxation of B is an artifact springing from the fact that the Si atoms are kept fixed. In the fully relaxed *AB* configuration the four Si neighbors of B relax inward because of the smaller size of the B atom (see Sec. III B). Although we do not allow all atoms to relax, these reference configurations are certainly sufficiently close to the fully relaxed configurations to make tests meaningful. In the 16-atom cell using energy cutoffs $(E_1;E_2)=(6;12)$ Ry, we find an energy difference of 0.316 eV for $q=2$. By choosing $q=4$, we enlarge the number of k points in the 1BZ by a factor of 8 and find 30 special points in the IRBZ. For $q=4$ the above energy difference drops to 0.306 eV. In the 32-atom cell we obtain an energy difference of 0.287 eV using $q=2$ (two points in the IRBZ), whereas $q=4$ (15 points in the IRBZ) yields 0.286 eV. We conclude that the $q=2$ choice is good enough to give energy differences between configurations with different H positions and different relaxations with an accuracy of about 0.01 eV. This is slightly better than in the earlier work on H in pure Si,[35]

since here we always integrate over a set of completely filled states. Finally, in the 8-atom cell the $q=2$ choice is not as good as in the 16- and 32-atom cells. Tests show that $q=4$ (10 points in the IRBZ) provides the same accuracy as $q=2$ in the larger cells. The 8-atom cell, however, will only be used to test the convergence of energy differences with respect to increasing the energy cutoffs (see next subsection). For that purpose the $q=2$ choice is sufficient.

### D. Energy cutoffs and supercell size

Calculations using the pseudopotential-density-functional method and a plane-wave basis set are generally performed with a choice of energy cutoffs $(E_1;E_2)$ for which calculated results still depend on this choice ($E_2$ is the kinetic-energy cutoff for plane waves included in the calculation; those with kinetic energy between $E_1$ and $E_2$ are included using second-order Löwdin perturbation theory[37]). For a given accuracy the size of the computational problem (i.e., rank of matrices to be diagonalized) is proportional to the volume of the unit cell, whereas processing time and memory usage are cubic and quadratic, respectively, in these sizes. Only for very small unit cells the usual computational limitations (central-processor-unit time and memory usage) allow one to fully converge the calculations with respect to increasing $E_1$ and $E_2$. One therefore has to make a careful study of the dependence on cutoffs in order to come to a judicious choice and quantitatively reliable results.

As indicated in Sec. II A, the choice of supercell size can also affect calculated energies, because if defects in neighboring cells are too close one is modeling a system with interacting defects. Here we present a study of the dependence on energy cutoffs and supercell size of the energy difference between the *BM* and *AB* reference configurations described in the preceding subsection. Table I and Fig. 2 show the results. In Fig. 2 we see that the three curves for the three supercell sizes are very well

TABLE I. Energy difference (in eV) between situations in which hydrogen occupies the bond-minimum (*BM*) and antibonding (*AB*) reference configurations (see text) as a function of energy cutoffs $(E_1;E_2)$ in (Ry) and as a function of number of atoms in the supercell. The results for the 8-atom cell are only used to study the dependence on energy cutoff since they have not been fully converged with respect to enlarging the mesh used in the k-space integrations (see text).

| $(E_1;E_2)$ (Ry) | 8 atoms | 16 atoms | 32 atoms |
|---|---|---|---|
| (6;12) | 0.481 | 0.316 | 0.287 |
| (8;16) | 0.518 | 0.358 | 0.333 |
| (10;20) | 0.554 | 0.399 | 0.370 |
| (12;24) | 0.586 | 0.433 | 0.400 |
| (14;28) | 0.602 | 0.451 | |
| (16;32) | 0.607 | 0.471 | |
| (18;36) | 0.610 | 0.475 | |
| (20;40) | 0.615 | | |
| (22;44) | 0.621 | | |
| (24;48) | 0.625 | | |
| (26;52) | 0.628 | | |

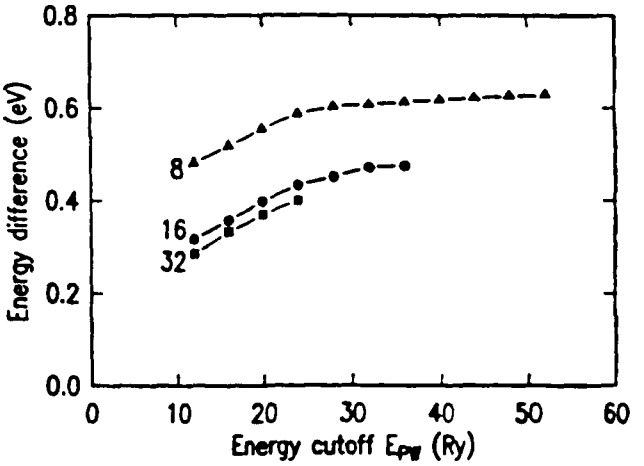



FIG. 2. Convergence of energy difference between the *BM* and *AB* reference configurations (see text) in which H occupies two different sites close to substitutional B in Si, as a function of kinetic-energy cutoff $E_{pw}$ (see caption of Fig. 1) and of supercell size. Supercells used contain, besides the H atom, 8, 16, or 32 host-crystal atoms. The results for the 8-atom cell are only used to further probe the dependence of the energy difference on $E_{pw}$ and are not fully converged with respect to enlarging the mesh used in the k-space integrations (see text).

behaved; they have the same (regular) form and are merely shifted with respect to each other by an almost constant amount. The curves for 16- and 32-atom cells do not differ by more than 0.03 eV. The 8-atom–cell curve shows that the behavior as a function of cutoff is the same as for the larger cells and convergence is eventually reached. The 8-atom–cell curve is not converged with respect to the number of k points used in the Brillouin-zone integrations ($q=2$ was used; see preceding subsection), which is unimportant for the present purpose of testing the dependence of energy differences on energy cutoff. For $E_2=36$ Ry we consider the energy difference to be converged, since the changes resulting from using higher cutoffs are very small compared to other numerical approximations employed (e.g., the Brillouin-zone integrations described in the preceding subsection).

We further study the energy-cutoff dependence of calculated energy differences by examining a larger set of positions for the H atom. The different sites considered here lie in the (110) plane and are depicted in Fig. 3. We use the 32-atom cell and all atoms up to second-nearest neighbors of the H atom are allowed to relax. In addition, the Si neighbors of the B atom are always allowed to relax. Table II summarizes the results. For the purpose of discussing Table II and following results, we find it useful to subdivide the different positions for the H atom into three regions. In region I the valence-electron density is very high (e.g., the *BM* site) and putting a H atom there will induce large relaxations of the crystal. In region II the electron density is lower but still considerable (e.g., the *AB*, *BB*, *C*, and *C'* sites); consequently, relaxations of the crystal are also still considerable. In region III the electron density is very small ($T_d$ and $H'$ sites) and the H atom will not induce much relaxation. Of course, one always has the relaxation of the Si neighbors of the B atom because of the small-

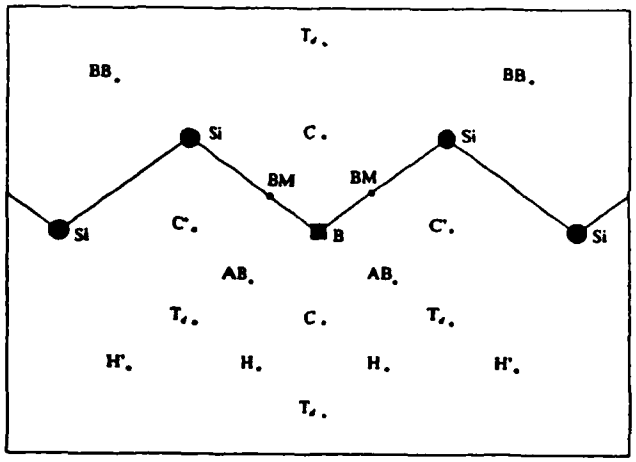



FIG. 3. Location in the (110) plane, containing a zig-zag chain of Si atoms and a substitutional B atom, of sites often referred to in the text. *BM* denotes the bond-minimum site, *AB* the antibonding site, *BB* the backbonding site, $T_d$ the tetrahedral interstitial site, and *H* and *H'* are (inequivalent) hexagonal interstitial sites. The *C* and *C'* sites are equivalent in pure Si, but not in the presence of a substitutional B atom.

Regarding convergence with respect to increasing the energy cutoffs, we make the following observation: energy differences between sites in the same region change by less than 0.05 eV by going from cutoffs (6;12) Ry to cutoffs (10;20) Ry and therefore may be considered fairly well converged at (6;12) Ry. In these calculations the relaxations are determined at the lower cutoffs and kept fixed for the higher cutoffs so that variations of energy differences are due solely to the change in cutoffs. Energy differences between sites in different regions change by about 0.1 eV when the combination of sites is region I–region II. This observation is useful if one wants to extrapolate calculated energy differences to very high energy cutoffs, which because of computational limitations cannot be handled together with large supercells. Tables I and II together provide means of extrapolating to higher cutoffs in order to obtain reliable quantitative estimates for energy barriers. We observe from Table I that the amount of change in going from cutoffs (6;12) Ry to cutoffs (10;20) Ry is about the same as that of going from (10;20) Ry to the converged values that we consider reached at (18;36) Ry. Therefore, calculations of energy differences between two sites at (6;12) and (10;20) Ry allow one to extrapolate to the converged energy differences. Using Table II we find that the *BM* site is 0.48 eV lower than the *AB* site and 0.29 eV lower than the *C* site. One should not apply such extrapolations to energy differences between sites in regions I and III (e.g., *BM* and $T_d$ sites) before a table like Table I for sites in regions I and III is calculated.

TABLE II. Energies (in eV) of situations in which hydrogen occupies different sites (see text and Fig. 3) in Si:B as a function of energy cutoffs $(E_1;E_2)$. As the zero of energy, the energy of the global energy minimum (*BM* site) is chosen. Energies are calculated in a 32-atom cell including relaxation up to second-nearest neighbors of the hydrogen atom. δ is the difference between the (6;12)- and (10;20)-Ry calculations.

| Site | (6;12) Ry | (10;20) Ry | δ (eV) |
|---|---|---|---|
| *BM* | 0.00 | 0.00 | 0.00 |
| *AB* | 0.26 | 0.37 | 0.11 |
| *BB* | 0.97 | 1.10 | 0.13 |
| *C* | 0.11 | 0.20 | 0.09 |
| *C'* | 1.36 | 1.44 | 0.08 |
| *H'* | 1.06 | 1.26 | 0.20 |
| $T_d$ | 1.61 | 1.85 | 0.24 |

Considering the above results, we come to the following choice of supercell size and energy cutoffs that we will use to calculate total energies for a large number of different H positions: We use 32-atom cells and energy cutoffs of (6;12) Ry. The use of the 32-atom cell allows us to take relaxations up to second-nearest neighbors of the H atom into account. Furthermore, the (artificial) dispersion of the H-related defect level in the gap is manageable, although a larger dispersion is not a big problem for the neutral H-B pair in Si as discussed in Sec. II B. The energy cutoffs (6;12) Ry are large enough to obtain qualitatively correct energy differences between different positions of the H atom, whereas it is still possible to calculate energies for a large number of different positions, including those that destroy all point-group symmetry of the system. It is necessary to calculate the energy for a large number of different H positions to get a picture of the entire energy surface for H in Si:B. For cases of special interest the energy difference can also be found in a *quantitatively* reliable way by using higher cutoffs and extrapolation, as shown above.

Occasionally, for positions of H for which the system has very low symmetry, the total-energy difference with a position for which the system has higher symmetry, but that lies in the same density region, is calculated in a 16-atom cell. This difference is then assumed to be the same in the 32-atom cell.

### E. Energy surfaces

It is very illuminating to combine the results of total-energy calculations for different positions of an impurity atom in a host crystal into an energy surface $E(\mathbf{R}_{\text{imp}})$ with the position of the impurity atom $\mathbf{R}_{\text{imp}}$ as the coordinate (note that this does not exclude the possibility that the host crystal contains other impurities). Such a surface provides immediate insight in the migration pathways, migration barriers, and stable sites for the impurity atom.

Quite generally, the observation can be made[9] that the function $E(\mathbf{R}_{\text{imp}})$ has the complete symmetry of the host crystal (without the tracer impurity), i.e., for any operation $\mathcal{R}$ of the space group of the host crystal structure, we have

$$E(\mathbf{R}_{\text{imp}})=E(\mathcal{R}\mathbf{R}_{\text{imp}})\ . \tag{1}$$

For instance, in a pure Si crystal, positions $\mathbf{R}_{\text{imp}}$ of a H atom in the center of different Si—Si bonds will render the same total energy, if all the appropriate relaxations are taken into account. Of course, different atoms relax for different bond-centered (BC) sites, since the Si atoms

forming the bond in which the H atom resides will relax most strongly. However, the relaxations for two different BC sites are connected by the same symmetry operation that connects the two sites. To obtain the energy surface $E(\mathbf{R}_{\text{imp}})$ we now proceed as follows: The function $E(\mathbf{R}_{imp})$ is expanded in a basis set of functions that all have the symmetry of the host crystal. The expansion coefficients are obtained by a least-squares fit to calculated values $E(\mathbf{R}_{\text{imp},i})$ for different positions $\mathbf{R}_{\text{imp},i}$ $(i=1,\ldots,N)$. By varying the degree to which the problem is overdetermined (where overdetermined means that the number of calculated data points, $N$, is larger than the number of symmetry functions, $M$, in the expansion), one can check the stability and, thus, the reliability of the fit.

For host crystals with a high degree of translational symmetry, a suitable set of basis functions is the set of symmetrized plane waves $\Phi_l(\mathbf{r})$:

$$\Phi_l(\mathbf{r})=\sum_{m=1}^{N_l} e^{i\mathbf{K}_m^{(l)}\cdot\mathbf{r}}, \tag{2}$$

where the $\mathbf{K}_m^{(l)}$ are vectors of the reciprocal lattice that corresponds to the Bravais lattice of the crystal. For each $l$, the $N_l$ vectors $\mathbf{K}_m^{(l)}$ transform into each other under operations of the crystallographic point group.

In previous work on H in pure Si,[9,35] typically eight symmetrized plane waves and 10 calculated points $E(\mathbf{R}_{\text{imp},i})$ were sufficient to obtain stable energy surfaces. However, for the problem we are addressing in this paper, the behavior of a H atom in a boron-doped Si crystal, the translational symmetry is essentially lost, and symmetrized plane waves are a less obvious choice of basis functions for the expansion of the energy surface. A possible solution to this problem would be to add a set of localized functions, e.g., Gaussians centered on the atoms, to the basis set or use a basis set consisting completely of localized functions. The disadvantage of such an approach is that a more complicated (nonlinear) fitting problem is encountered, since also the decay constants that appear in the Gaussians need to be fitted. We have chosen the following approach: In the same spirit as used in the supercell approach discussed in Sec. II B, we use as basis functions for the expansion of the energy surface symmetrized plane waves of a supercell. In this way, periodicity is restored so that symmetrized plane waves are suitable basis functions, but the repeat distances can be chosen so large that the region around the substitutional impurity atom that we are interested in is not affected by impurities in neighboring cells. By studying the behavior of the total energy when the H atom is moved away from the B atom, and comparing this with the case of $H^+$ in pure Si, we establish (see Sec. III C) that the influence of the B atom has disappeared at a distance of about 2.1 Å from the B atom. Therefore, to describe the energy surface around a B atom, it is allowed to assume that it has the symmetry of the 8-atom supercell, which has repeat distances of 5.43 Å in three perpendicular directions. This, in turn, implies that the symmetrized plane waves $\Phi_l(\mathbf{r})$, with $\mathbf{K}_m^{(l)}$ reciprocal-lattice vectors belonging to the (simple-cubic) lattice of the 8-atom cell, are suitable functions to expand the surface in. We would like to stress that this choice of supercell is independent of the choice of supercell one uses in calculating the total energies $E(\mathbf{R}_{\text{imp},i})$. For the latter purpose one needs supercells of 32 atoms to take into account all relevant relaxations of the host crystal, as argued before.

Using this approach, the total energy still has to be calculated for a large number of different positions $\mathbf{R}_{\text{imp},i}$ of the H atom. We have found that about 40 inequivalent sites in the 8-atom cell are needed to get a good description of the energy surface. This number is consistent with the number of points (ten) typically used in fitting the energy surface for H in pure Si, the diamond structure of which has a unit cell 4 times as small. Typically, 25 symmetrized plane waves are used in the fit of the energy surface of H in Si:B. Results of this procedure will be shown below.

## III. RESULTS AND DISCUSSION

### A. Electronic structure

The band structure for the Si crystal with the H-B complex closely resembles that of Si with a substitutional B atom; there is no acceptorlike level in the gap showing that the acceptor is passivated. We note that a supercell calculation of the band structure of Si with a substitutional B atom, but without the H atom, will only produce an acceptorlike level in the gap if very large supercells are used. The hydrogenic state corresponding to such a shallow level is known to extend over several tens of angstroms and can therefore not be described by small supercells. Indeed, we do not find such a level in calculations without the H atom with supercells of up to 32 atoms. We do find a level near the gap that behaves almost identically to the level found in the case of H in pure Si;[35] this level is therefore related to H. We find that the wave function associated with this level is mostly localized around the position of the H atom and that the position of this level in the gap moves when the H atom is moved. As already discussed in Sec. II B, we note that our use of supercells induces defect levels to have dispersion. To obtain a dispersionless level from our calculations, we take a weighted average of the defect-level position over the special $\mathbf{k}$ points for which the band structure is calculated during the total-energy calculation (more symmetric $\mathbf{k}$ points carry less weight because they map onto fewer points in the 1BZ). The position of this level depends on the location of the H atom and roughly two cases may be distinguished. If the H atom is in one of the regions of high or intermediate electron density (regions I and II as defined in Sec. II D), the H-related defect level is located slightly above the bottom of the conduction bands. If the top of the valence bands is chosen as the zero of the energy scale, the bottom of the conduction bands of Si with one substitutional B atom is found to be at 0.46 eV (an underestimation of the experimental energy gap of 1.17 eV as is usual in LDA calculations). For H at the *BM*, *AB*, *C*, and *C'* sites (see Fig. 3), the defect level is at 0.50, 0.56, 0.62, and 0.53 eV, respectively. If the H atom is in the low-density region (region III), the defect level appears as a resonance slightly below the top

of the valence bands. For H at the $T_d$ and $H'$ sites of Fig. 3, the defect level is at $-0.37$ and $-0.09$ eV, respectively. If the H atom is located in region III, it is not bound to any atom and acts as an acceptor. The position of the defect level is sensitive to the energy cutoffs used; the quoted results were obtained using 32-atom cells and cutoffs of (6;12) Ry and have only qualitative value. One should also bear in mind here that it is a well-known deficiency of the LDA that, while the valence bands of a semiconductor are well described, the conduction bands, and also conduction-band-related levels, are not in agreement with experiment. This problem has recently been overcome for bulk solids by including many-body corrections.[45] Since for defect calculations this solution involves a prohibitive computational effort, it has not yet been applied to such calculations, which are already very demanding by themselves.

In the self-consistent calculation of the total energy, the H-related level is always unoccupied, since the substitutional B and interstitial H atoms together exactly supply the four valence electrons of the Si atom that has been replaced by the B atom, so that only the "pure-Si"-like bands are occupied if the defect level is in the conduction bands. If the defect level is just below the top of the valence bands, it is still left unoccupied, since for the k points at which the band structure is calculated during the self-consistency process the defect level usually lies between the valence- and conduction-band levels. If it lies below the top valence-band level, we leave it unoccupied artificially to obtain a consistent comparison with the total-energy calculations for H at the other sites.

### B. Relaxation of the host crystal

In this subsection we present results for the relaxation of the host crystal (Si:B) for some characteristic positions of the H atom. For every position the total energy is minimized with respect to the positions of the atoms in the host crystal.

We first mention that in the absence of the H atom the four Si neighbors of the B atom relax toward the B atom in a "breathing-mode"-type relaxation, whereas the B atom shows a very slight tendency to become threefold coordinated by moving towards a plane with three Si neighbors (it moves less than 0.1 Å). Both for neutral B ($B^0$) and negatively charged B ($B^-$) the relaxation of the Si neighbors is 0.21 Å, reducing the Si—Si bond distance of 2.35 Å by 9%. The energy gain of this relaxation is 0.9 eV. The relaxation results in a Si—B distance of 2.14 Å, which is very close to the sum of covalent radii of Si and B (1.17 and 0.90 Å, respectively). It is interesting to compare this result for the Si—B distance with an experimental result from x-ray-diffraction measurements of the lattice contraction in B-doped Si. To make the comparison, some assumptions have to be made, the validity of which is not easily assessed. We first assume that the lattice contraction is solely caused by the difference in covalent radii of Si and B (in general, there is also a, possibly competing, electronic contribution caused by the pressure dependence of the band-gap edges[46]). Using our result of 2.14 Å for the Si—B distance and following the simple argument of Shih *et al.*,[47] the "natural-bond" length defined in Ref. 48 for a Si—B bond becomes 2.07 Å. If we now use Végard's law for the average bond length in B-doped Si (with pure Si and "zinc-blende" BSi as extreme structures), we may extract the contraction coefficient $\beta$, defined by

$$\Delta a/a = \beta C_B , \tag{3}$$

where $C_B$ is the boron concentration and $\Delta a/a$ is the relative change in average lattice constant. We find $\beta = -4.8\times10^{-24}$ cm$^3$/atom, which is in agreement with the experimental results of $\beta = -(6\pm2)\times10^{-24}$ cm$^3$/atom (see the references in Ref. 23).

The relaxation of the host crystal in the presence of a H atom is most appreciable if H resides in the *BM* site (see Fig. 3). This site is located in a Si—B bond slightly displaced from the bond center toward the B atom. We distinguish it from the geometrical bond center (BC), which was found to be the global energy minimum for $H^+$ in Si in previous work.[9] The *BM* site is the global energy minimum for H in Si:B (see the next subsection). For H at this site the neighboring Si and B atoms relax outward (as measured from their ideal lattice positions) by 0.24 and 0.42 Å, respectively. The smaller outward relaxation of the Si atom is easily explained by the fact that it would relax inward by 0.21 Å if the H atom was absent. Put differently, the above relaxations allow for close to ideal H—Si and H—B distances since they result in a H—Si distance of 1.65 Å and a H—B distance of 1.36 Å. For comparison, we mention that for $H^0$ ($H^+$) in the BC site in pure Si the two Si atoms forming the bond relax outward by 0.45 Å (0.41 Å), resulting in a H—Si distance of 1.63 Å (1.59 Å). Typical H—B distances in $B_2H_6$ (diborane) are 1.20 Å for H in a terminating bond and 1.34 Å for H in a bridging bond.[49] The second-nearest Si neighbors of the H atom in the *BM* site relax outward along the original bond axes by 0.05 Å if they are bonded to the Si neighbor of H and relax inward along the original bond axes by 0.14 Å if they are bonded to the B neighbor of H. These relaxations result in Si—Si and Si—B bond distances of 2.33 and 2.11 Å, respectively, which are very close to the Si—Si distance in pure Si (2.35 Å) and the Si—B distance in Si:B (2.14 Å). The gain in energy of these relaxations compared to the configuration in which H occupies the exact bond-center site and all other atoms occupy their ideal lattice positions is calculated to be 3.2 eV.

Our calculated relaxed configuration for the *BM* site is in qualitative agreement with the results of previous work using a variety of methods.[10,14,26] Notable differences are as follows. In Ref. 14 the H atom was found to reside closer to Si than to B. The outward relaxation of the B atom of 0.58 Å found by DeLeo and Fowler[10] (which we extract from their Fig. 1) significantly exceeds our result of 0.42 Å, which, in turn, is larger than the experimental result of 0.28±0.03 Å from ion-channeling measurements.[19] The error estimate of the experimental value results from the analysis of the data and does not include the inherent insensitivity of the channeling method, which is about 0.1 Å.[18]

For H at the *AB* site (see Fig. 3), which is a minimum along the ⟨111⟩ axis, but a saddle point of the entire en-

ergy surface (see next subsection), the H—B distance is 1.32 Å. The B atom hardly moves from its substitutional site (less than 0.05 Å towards H) and the three Si neighbors relax toward B by 0.14 Å. Our calculated H—B distance is in between those found in Refs. 10 and 11, where very different distances of 1.19 and 1.8 Å, respectively, were found using similar semiempirical cluster calculations.

If H is positioned at the *C* site (Fig. 3), the B atom does not move from its substitutional site. This may again be explained by the fact that the H—B distance in this case is close to ideal (1.36 Å). Note that this is different from the case of $H^+$ in pure Si, where the distance is smaller than the ideal H—Si distance of ~1.6 Å. In that case an appreciable relaxation of the Si atom away from the H atom results. For H at the *C* site in Si:B, the inward relaxation of the two Si atoms bonded to B and next to H (see Fig. 3) is obstructed by the presence of H and is only 0.05 Å, whereas the two Si atoms bonded to B but far away from H (in the plane perpendicular to that of Fig. 3) have the same inward relaxation as for the *BM* and *AB* sites discussed above. The minimum energy for H along the line connecting the *C* site and the B substitutional atom is not at the *C* site, but slightly displaced (0.24 Å) from it toward the B atom. For that position the B atom *does* relax away from the H atom to restore the preferred H—B distance of 1.36 Å.

Finally, if H is put at the tetrahedral interstitial site ($T_d$) or hexagonal interstitial site (*H* or *H'* in Fig. 3) the only relaxation is a "breathing-mode" relaxation of 0.21 Å of the Si atoms bonded to the B atom. This is exactly the same relaxation as in the complete absence of the H atom (see above), which is consistent with the earlier finding[9] that there is no appreciable relaxation for H at the $T_d$ or *H'* sites in pure Si.

From the results of first-principles total-energy calculations presented here, it can be inferred that the relaxations of the host crystal are roughly determined by the tendency of two neighboring atoms to be separated by some preferred distance. The preferred distance is roughly determined by the sum of covalent radii (for H a covalent radius of 0.43 Å has to be used). However, there are also deviations from this general behavior, e.g., the H-obstructed relaxation of two Si neighbors of B for H at the *C* site. In any case, the examples described above can be considered as a data base to allow for an efficient search for the configurational energy minimum for an arbitrary H position.

### C. Energy surface for H in Si:B

The effect of introducing a substitutional boron impurity in the silicon crystal on the behavior of H is clearly demonstrated in Fig. 4. We compare the energy of a H atom in Si:B with the energy of a positively charged H ($H^+$) atom in pure Si for various positions of the H atom along the line connecting two bonded Si and B atoms (two Si atoms in the case of pure Si). We note that with $H^+$ in pure Si we do not mean a bare proton in pure Si; the notation is a mere shorthand for the fact that one electron is left out of the system.[35] The other electrons are still allowed to distribute themselves self-consistently

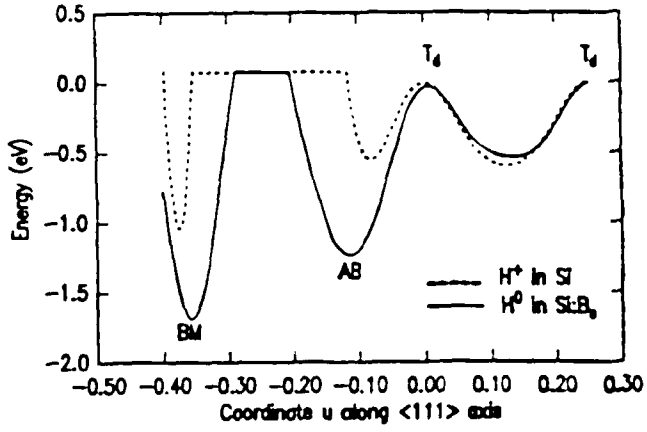



FIG. 4. Energy for positions of the hydrogen atom along the ⟨111⟩ axis for $H^+$ in pure Si (dashed line) and for $H^0$ in Si:B (solid line). For all positions of the H atom, coordinates of the host-crystal atoms have been relaxed to minimize the energy. The curves have been truncated at 0.08 eV for positions very close to the Si (in pure Si) or B atom (in Si:B) at $u = -0.25$. The smaller truncated region for B reflects that H can approach the B atom closer than the Si atom.

according to Schrödinger's equation. The line connecting two bonded atoms we call the ⟨111⟩ axis and the position along this line is given by the single coordinate *u*; a coordinate *u* means that the position has Cartesian coordinates $(u,u,u)$ in units of the Si diamond-structure lattice constant of 5.43 Å. A coordinate $u = -0.5$ denotes the unrelaxed Si atomic position, $u = -0.25$ the unrelaxed B atomic position, and $u = 0$ and 0.25 are $T_d$ sites. The comparison with $H^+$ in pure Si is the most meaningful comparison that one can make, because H behaves as a donor in *p*-type material and will give up its electron to annihilate the free holes resulting from the ionized acceptor. (This does *not* imply that H behaves as a bare proton everywhere in *p*-type Si; just as for H at the bond-center position in pure Si,[35] in the H-B complex the missing electron is not removed from the immediate neighborhood of the H atom, but from a region extending past the neighboring atoms.) The two curves have been obtained from the energy surfaces for the two cases ($H^+$ in Si and H in Si:B) by extracting the energy values for coordinates along the ⟨111⟩ axis. The energy scales have been aligned at the distant $T_d$ site, $u = 0.25$. It is clear from Fig. 4 that the influence of the substitutional B atom does not stretch out further than $u = -0.03$, corresponding to 2.1 Å from the B atom. Beyond that point the curves coincide to within better than 0.1 eV (which is about the estimated error of energy calculation and fit together). The above observation justifies the use of symmetrized plane waves with the periodicity of the 8-atom (conventional) unit cell of the diamond structure as basis functions for the expansion of the energy surface. We repeat (see Sec. III E) that this observation does not imply that it is sufficient to do the total-energy calculations in a supercell of eight atoms. This procedure for the expansion of the energy surface is satisfactory if one is interested in this surface in the neighborhood of the B atom (see Sec. II E). Further away from the B atom, the surface is identical to the one for $H^+$ in pure Si (see Fig. 4; we have also established this for H positions that are not on the ⟨111⟩

axis). Figure 4 also shows that B acts as an attractor to the H atom, since the bond-centered and antibonding minima are lowered and moved towards the B atom. From the three-dimensional and contour plots of the energy surface in the complete (110) plane containing the ⟨111⟩ axis (to be discussed below with Figs. 5 and 6), it follows that the *BM* site is an actual (and even global) minimum, whereas the *AB* site represents a saddle point. There is no energy barrier between the *AB* site and an equivalent *BM* site that is not located along this ⟨111⟩ axis.

In Figs. 5(a) and 5(b) we show three-dimensional plots of the energy surface for H in Si:B for positions of H in the (110) plane (containing a chain of atoms as in Fig. 3) and the (111) plane through three bond-minima sites, respectively. Figure 5(a) shows the low-energy region (in red) around the B atom. The region does not contain the *AB* and *BB* sites on both extensions of the Si—B bond; these sites appear as saddle points of the energy surface. From Fig. 5(b) it is clear that the low-energy region extends all around the B atom, which is located slightly out of the (111) plane, which is shown in Fig. 5(b).

In Fig. 6(a) we show a contour plot of the energy surface for H in the (110) plane in Si:B. It shows most of the salient features of the complete energy surface, which cannot be shown in one picture since the energy is a function of three independent coordinates. The *BM* site is the global minimum, whereas we see again that the *AB* site is a saddle point. In Fig. 6(b) we show exactly the same part of the energy surface for the case of $H^+$ in pure Si. From the comparison we see that the H atom gets trapped close to the B atom and has no low-energy pathway to migrate away from the B atom. The H atom can move between equivalent *BM* sites around the B atom by passing over an energy barrier close to the *C* site (between the *C* site and the B atom) of only 0.2 eV. Of course, for this to happen the relaxation of the host crystal has to adjust accordingly. There is no barrier between the *BM* and *C* sites. The low-energy barrier implies that at room temperature the H atom will be able to move around the B atom between the four equivalent *BM* sites. Very recently, in experiments using the optical dichroism of the H-B absorption bands under uniaxial stress, an activation energy of 0.19 eV was found for H motion from one *BM* site to another.[24] This activation energy is in excellent agreement with our calculated barrier of 0.2 eV.

We find that the *BM* site is 0.48 eV lower than the *AB* site and 0.29 eV lower than the *C* site. The energy difference between *BM* and *AB* sites of 3.12 eV, obtained in Ref. 14 from Hartree-Fock calculations, we consider to be very unreliable. A final observation from Fig. 6(a) is that the *C* and *C'* sites, which are completely equivalent in pure Si, are not only symmetrically inequivalent (e.g., *C* is at 1.36 Å from the B atom, *C'* at 1.92 Å), but that they differ in energy by the large amount of 1.2 eV. This site inequivalence in the neighborhood of a substitutional impurity leads us to a brief discussion of the accuracy with which ion-channeling experiments are able to determine the site of hydrogen.[18,20] The analysis of ion-channeling experiments involves a statistical average over the possible substitutional sites for the impurity B atom.[50] After such an average, the energy surface for a H atom in Si:B has the complete symmetry of the diamond structure of pure Si. This implies that, for instance, the *C* and *C'* sites are considered to be completely equivalent in the analysis of ion-channeling experiments. The same holds for the *AB* and *BB* sites (see Fig. 3), which in our calcula-

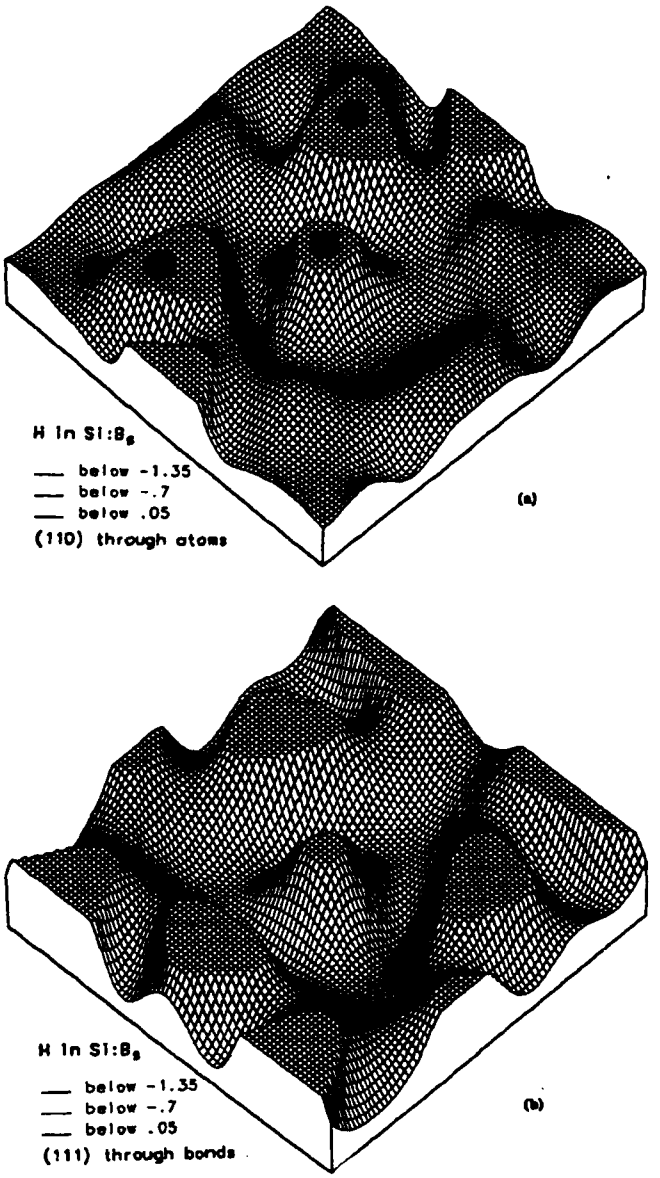



FIG. 5. Energy surface for a hydrogen atom in Si with one substitutional boron atom in (a) a (110) plane containing a chain of atoms, and (b) a (111) plane through three equivalent bond-minima (*BM*) positions. The black dots represent Si atoms and the pink dot the B atom. The plane in (b) does not contain atoms, but the unrelaxed lattice position of the B atom is located just 0.4 Å outside the plane in the center of the surface. Atoms are shown at their unrelaxed positions since they relax differently for different positions of the H atom, but relaxations are taken into account in the total-energy calculations. The energy is below −1.35 eV in the red region, between −1.35 and −0.7 eV in the blue region, and between −0.7 and 0.05 eV in the green region. The surface is cut off at an energy value of 0.05 eV. The zero of energy is chosen at the tetrahedral interstitial site.

tion differ by about 0.7 eV in energy (the *AB* site being the lower-energy site). Therefore, ion-channeling experiments are able to discriminate between sites that remain inequivalent when averaging over the possible substitutional sites for the B atom, e.g., *BM* and *AB* sites. They cannot discriminate, however, between, e.g., *AB* and *BB* sites. On account of this, the conclusion from these experiments that H resides predominantly in a Si—B bond (a Si—Si bond can be excluded since a large displacement from the substitutional site of the B atom is also observed[18]) is indisputable, but the further detailing of percentages of H at other sites[20] is not necessarily relevant to the microscopic structure of the H-B complex. Observation in ion-channeling experiments of H at other sites is most likely related to defects which may be located far away from the B atom.

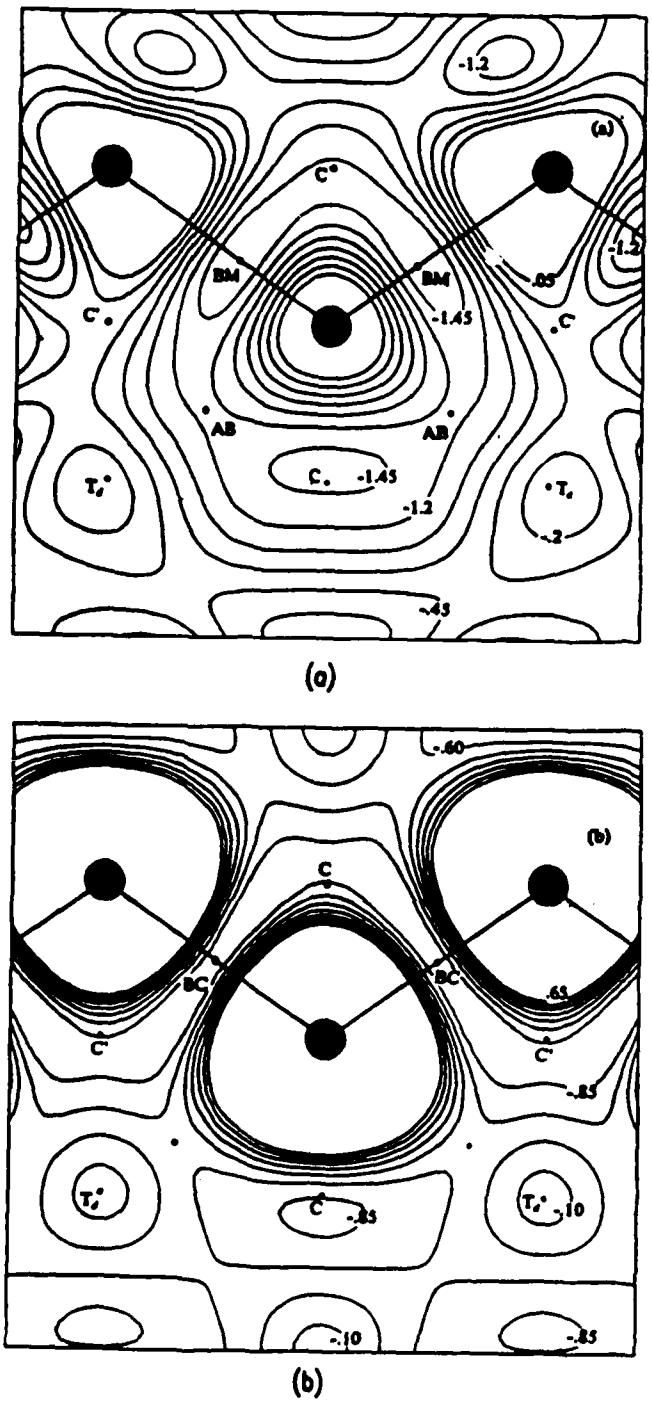



FIG. 6. Contour plots of the energy surface of a H atom in the (110) plane in boron-doped and pure silicon. Large dots indicate (unrelaxed) atomic positions; bonded atoms are connected by solid lines. Positions of special interest are indicated (cf. Fig. 3). The unit of energy is eV and the spacing between contours is 0.25 eV. Close to the atoms contours are not shown above a certain energy value. (a) $H^0$ in Si:B. The boron atom occupies the center of the plot. Highest contour shown is 0.05 eV. (b) $H^+$ in pure Si. Highest contour shown 0.65 eV.

In Fig. 7 we present contour plots of the energy surface in a few other planes, showing that the *BM* site is indeed the global energy minimum and that there is a spherical shell-like region (with some holes in it) at a radial distance of about 1.3 Å from the B atom, for which the energy is between $-1.45$ and $-1.7$ eV (with respect to the energy at a far $T_d$ site). Thus the H atom can move around adiabatically on this shell with an energy barrier at a site closer to the *C* site of only 0.2 eV.

### D. Hydrogen vibrational frequencies

Because infrared measurements of the hydrogen vibrational frequency have been an important source of experimental information on the H-B complex,[4,15,16] it is worthwhile to make a connection with that work by calculating the vibrational frequency for the H-stretching mode. We have done this for a number of different sites for the H atom that all have been proposed as the equilibrium site for the H atom on account of theoretical calculations.

The sites for which we calculated the frequency of the H-stretching mode are the *BM* and *AB* sites already discussed extensively above, as well as the backbonding (*BB*) site shown in Fig. 3. For H in the latter site, the H—Si distance is again 1.60 Å, while the Si atom closest to H relaxes toward the B atom by 0.3 Å. For each of the three sites, we determine the minimum-energy configuration by allowing up to eight atoms around the H atom as well as the H atom itself to relax. Subsequently, we move the H atom away from its equilibrium position in directions corresponding to a stretching mode over distances of 2% and 4% of a Si—Si bond length. The relaxation of the host crystal is now kept as in the minimum-energy configuration.[51] The procedure described above induces energy changes of typically up to 30 meV. These energy differences $\Delta E$ are fitted to a parabola $\Delta E \simeq \frac{1}{2} f u_H^2$, where $u_H$ is the displacement of the H atom and $f$ the force constant of the stretching mode. If $f$ is expressed in units of eV/Å$^2$, the wave number $\kappa$ for the stretching mode is given in units of cm$^{-1}$ by

$$\kappa = \frac{1}{2\pi}\left[\frac{f(\text{eV}/\text{Å}^2)}{938.25}\right]^{1/2} \times 10^5 \text{ cm}^{-1}, \qquad (4)$$

where we have taken the vibrating object to be a proton (with rest mass $m_p c^2 = 938.25$ MeV). A very similar procedure to the one described here was used successfully by Kaxiras and Joannopoulos[52] to calculate vibrational frequencies of H atoms saturating dangling bonds at Si and Ge (111) surfaces.

In Table III we summarize our results and list the results of previous theoretical calculations using a variety of methods. From varying the number of calculated points used in the parabolic fit and from calculations at lower energy cutoffs, we estimate the error bar on our calculated frequencies to be 100 $cm^{-1}$. Also, results obtained from the same calculations in a 16-atom cell fall within this error bar. Considering the error bar, our result for the H vibrational frequency at the *BM* site is in fair agreement with the low-temperature (5 K) experimental results[16,17] of 1903 and 1907 $cm^{-1}$. The agreement with the result obtained at 273 K (1870 $cm^{-1}$) (Ref.

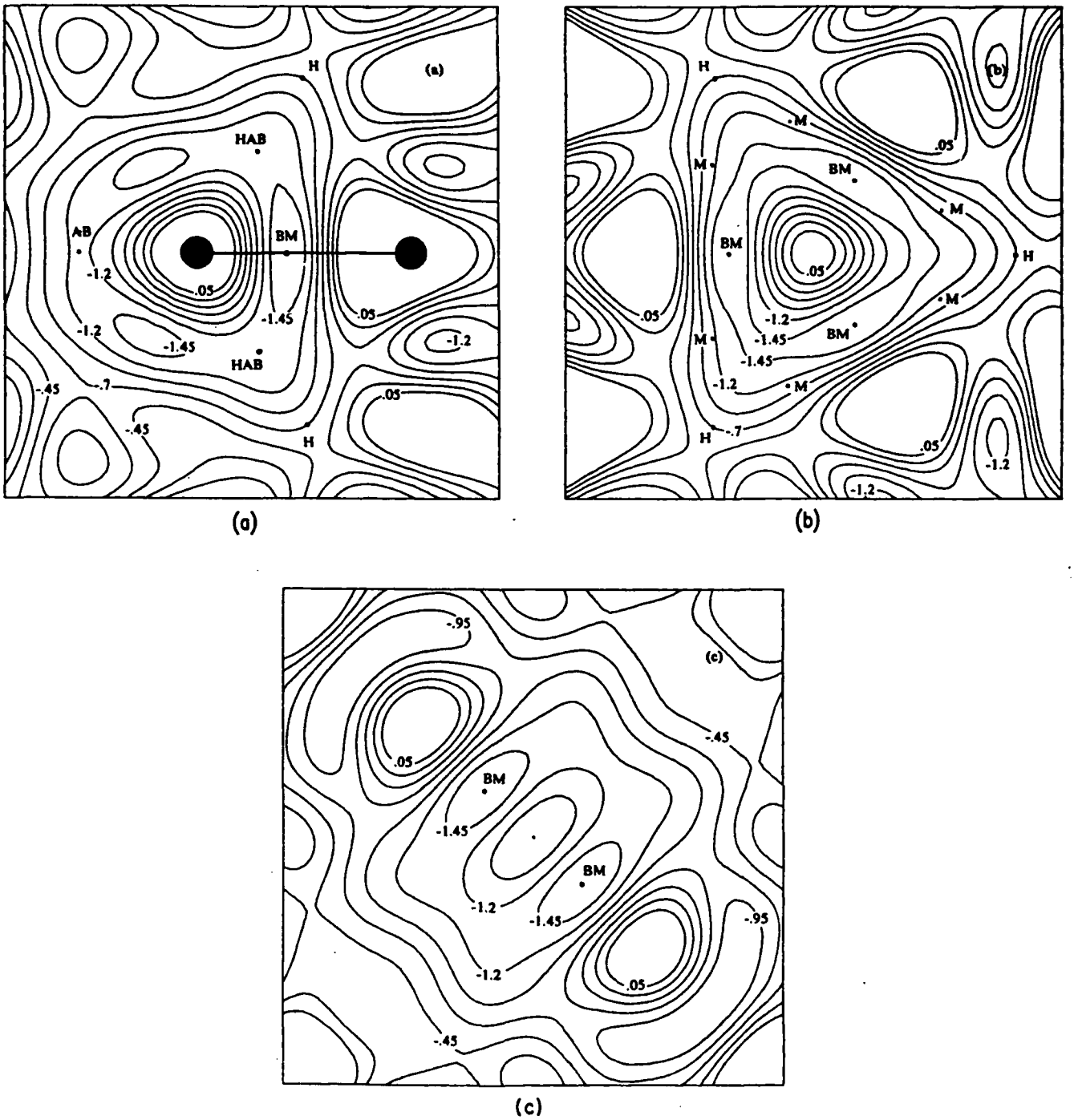



FIG. 7. Contour plots of the energy surface of a neutral H atom in various planes in Si:B [see Fig. 6(a) for the (110) plane]. Indicators are the same as in Fig. 6. (a) $(2\bar{1}1)$ plane containing one B—Si bond (B atom on the left). *HAB* denotes the hexagonal antibonding site, a saddle point of the energy surface about halfway between the hexagonal interstitial site *H* and the B atom. (b) (111) plane through three bond minima (*BM* sites). The *M* sites lie halfway between two *C* sites, one of which is in the (110) plane (see Figs. 3 and 6). In this plane there is a ringlike low-energy region around the B atom (that is not located in this plane). The perspective plot for this plane is shown in Fig. 5(b). (c) (001) plane through two bond minima (*BM* sites). This plane is perpendicular to the (110) plane of Fig. 6(a).

TABLE III. Calculated wave numbers $\kappa_{stretch}$ (in $cm^{-1}$) of vibrational frequencies of hydrogen-stretching modes for hydrogen in the bond-minimum (*BM*), antibonding (*AB*), and backbonding (*BB*) sites in Si:B compared with previous theoretical calculations using a variety of methods.

| Site | Present result $\kappa_{stretch}$ | Previous theoretical calculations $\kappa_{stretch}$ |
|---|---|---|
| *BM* | 1830 | 1880,[a] 1820[b] |
| *AB* | 1680 | 1000,[b] 1870,[c] 2590[d] |
| *BB* | 1590 | |
| Expt. | 1903[e] | |

[a]Reference 10.
[b]Reference 26.
[c]Reference 11.
[d]Reference 12.
[e]Reference 16.

15) is even better, but that is not the appropriate number to compare with.

In view of the error bar of 100 $cm^{-1}$, the results for the *BM* and *AB* sites are not that different, and would not supply strong enough evidence for one to conjecture that the infrared data exclude the *AB* site as the equilibrium site for H. Previous authors did make this claim on account of their finding that the H vibrational frequency is very different for the *BM* and *AB* sites. The peculiar fact occurred, however, that their results for the *BM* site are in general agreement with experiment, but DeLeo and Fowler[12] find the result for the *AB* site to be much larger (2590 $cm^{-1}$), whereas Chang and Chadi[26] find it to be much smaller (1000 $cm^{-1}$) than the experimental value. These authors did not discuss the accuracy of their calculated result. We stress that, of course, the *AB* site can be ruled out as the equilibrium site for the H atom because of the fact that it is a saddle point of the energy surface and 0.48 eV higher in energy than the *BM* site, without a barrier between the two sites [see Fig. 6(a)]. The result for the *AB* site of 1870 $cm^{-1}$ in Ref. 11 was obtained by fitting a force-constant model to the experimental value that was known at that time. Not too much value must be attached to this result.

It is interesting to compare our results of Table III with the vibrational frequency for $H^0$ ($H^+$) in a BC site (geometrical bond center) in pure Si. For that case we calculate a frequency of 1945 (2210 $cm^{-1}$). This is much larger than the result for the *BB* site in Table III. Since the frequency for the *BM* site is in between those obtained for a H atom close to one Si atom (at the *BB* site in Si:B) and a H atom in between two Si atoms (at the BC site in Si), we infer that at the *BM* site there is still a fair amount of bonding between H and B besides the bonding between H and Si. The fact that the H atom is also bonded to the B atom (which is a modification of the original description of the bond-centered configuration by Pankove *et al.*[4]) can also be inferred from the fact that the H atom can easily move around the B atom between equivalent *BM* sites, as discussed in the preceding subsection.

Infrared frequencies that have been associated with stretching modes involving a single H atom in hydrogenated amorphous Si and hydrogenated crystalline Si range between 2000 and 2200 $cm^{-1}$.[53,54] The fact that our results of 1945 and 2210 $cm^{-1}$ (for $H^0$ and $H^+$ at the bond-centered site in pure Si, respectively) are close to these frequencies and also to the vibrational frequency of H saturating a dangling bond at a Si (111) surface (2085 $cm^{-1}$) is remarkable. We also observe that the *AB*-site frequency of 1680 $cm^{-1}$ resulting from a stretching H—B bond is apparently greatly modified with respect to frequencies of 2560 $cm^{-1}$ (for terminal H bonding) and 1985 $cm^{-1}$ (for bridge bonding) found in diborane.[4]

To conclude this subsection we briefly discuss the side bands in the infrared transmission spectra that were recently found for the H-Al and H-Ga complexes in Si (for the H-B complex the side bands are not resolved, but they are expected to be there).[16] Stavola *et al.*[16] suggested that these side bands are the result of a low-frequency excitation, for which they proposed two possibilities. The first possibility is the tunneling of H between different but equivalent *BM* sites. This must be considered unlikely, because of the rather large adjustments in the relaxation of the host crystal that have to happen for these sites to be indeed equivalent. This explanation was more recently abandoned by Stavola and coworkers.[24] The second possibility is that the H atom resides slightly off the *BM* site and off the ⟨111⟩ axis; the vibration would then be modeled by that of a quasilinear molecule, which is known to have side bands.[55] In that case, the configuration would resemble that of oxygen bridging a Si—Si bond.[56] We have investigated this possibility by positioning the H atom slightly off axis from the *BM* site on the ⟨111⟩ axis. While keeping the relaxation of the surrounding crystal fixed, the energy remained constant for small displacements (<0.1 Å) of the H atom in several directions. We did not allow the surrounding crystal to adjust its relaxation, but this can only lower the energy. However, by moving the H atom off axis, we change the symmetry of the system considerably, and the changes in energy that we obtain fall within the accuracy of our calculations. We conclude that an off-axis position for the H atom is very well possible, but cannot be quantitatively assessed by our calculations.

### E. Dissociation energy of the H-B complex

It is found experimentally that at temperatures above 150 °C the conductivity of the passivated samples starts to recover and can eventually be restored completely.[4,5] It seems natural to attribute the increase in conductivity to the dissociation of the H-B complexes and subsequent diffusion of H out of the passivated region. As we saw earlier, H-induced passivation arises from compensation followed by pair formation. By the same token, one can consider the dissociation reaction:

$$(BH)^0 \rightarrow B^- + H^+ , \quad (5)$$

This reaction, however, produces no free holes (i.e., the material is still compensated) and leaves open the possibility of reformation of the pairs by the reverse reaction.

Restoration of the conductivity requires an additional reaction, for instance,

$$H^+ \rightleftarrows H^0 + h^+ , \tag{6}$$

where $h^+$ denotes a free hole. Reaction (6), which denotes an electronic process, equilibrates very fast. The relative amounts of $H^+$ versus $H^0$ are determined by the position of the Fermi level. If the conditions are such that $H^0$ is overwhelmingly favored over $H^+$, the following reaction would apply:

$$(BH)^0 \rightarrow B^- + H^0 + h^+ . \tag{7}$$

Since B is a shallow acceptor, we can assume that its preferred state (after dissociation of the H-B complex) is an ionized $B^-$.

We see that there is no unique, single reaction describing the dissociation of the H-B pair. Nevertheless, we have calculated dissociation energies associated with *specific* dissociation reactions. By dissociation energy we mean the energy difference between the initial and final configurations of the breakup reaction. Quite generally, one can define for any reaction that can occur in two directions the reaction activation energies for the forward and reverse reactions. The dissociation energy as defined above is also precisely the difference between the two reaction activation energies associated with the dissociation reaction.

The dissociation energy associated with reaction (5) is found by calculating the following total energies:[51] (i) $E(\mathrm{HB})$, the total energy of the fully relaxed Si:B crystal with H at the *BM* site; (ii) $E(B^-)$, the total energy of the fully relaxed Si crystal with a substitutional $B^-$; (iii) $E(H^+)$, the total energy of a fully relaxed Si crystal with a $H^+$ at the BC site; and (iv) $E(\mathrm{Si})$, the total energy of a pure Si crystal. The dissociation energy $E_d$ may now be defined as

$$E_d = -E(\mathrm{HB}) + E(B^-) + E(H^+) - E(\mathrm{Si}) . \tag{8}$$

This dissociation energy does not depend on the Fermi level, because no electrons or holes are involved in reaction (5). This formula results in $E_d = 0.59$ eV.

To calculate the dissociation energy based on reaction (7), one has to use the following formula:

$$E_d = -E(\mathrm{HB}) + E(B^-) + E(H^0) + E(h^+) - E(\mathrm{Si}) , \tag{9}$$

where $E(h^+)$ is the energy of a free hole, for which we take minus the Fermi energy $E_F$ (a hole is the absence of an electron). The dissociation energy in (9) does depend on the Fermi level because it involves reaction (6). From (9) we find a dissociation energy $E_d = 1.09\ \mathrm{eV} - E_F$.

Experimentally, one can determine a dissociation energy if the dissociation is governed by first-order kinetics, i.e., the rate of change with time of the number of pairs $N$ is proportional to $N$:

$$\frac{dN}{dt} = -\nu N , \tag{10}$$

where $\nu$ is the dissociation rate constant. This assumption would have to be tested by demonstrating a linear dependence of $\ln(N)$ on time for several temperatures [$\ln(N) = -\nu t$]. For $\nu$, an Arrhenius-type temperature dependence is usually assumed:

$$\nu = \nu_0 e^{-E_A/kT} , \tag{11}$$

where $\nu_0$ is an attempt frequency and $E_A$ the activation energy. If the assumption of first-order kinetics holds, the measured temperature dependence of $\nu$ allows one to extract the activation energy $E_A$. This activation energy is the energy barrier that must be overcome for the breakup to occur [for instance, the forward-reaction activation energy of reaction (5)]. It *does not* correspond to a dissociation energy in the sense of the energy difference between the pair and the isolated breakup products.

The experimental procedure just described has not been carried out. Wichert *et al.*[22] followed the simplified procedure of isochronal annealing in which they *assumed* first-order kinetics. They extracted an activation energy of 1.3 eV in the case of H-In pairs. They estimated that the H-B pairs would break up with a smaller activation energy. We note, however, that it has been found[57] that first-order kinetics is not obeyed, so that this number may not be particularly meaningful and cannot be compared with theoretical values. A more sophisticated analysis of the data would be needed to extract energies that can be compared with theory.

Finally, we mention that one may define the binding energy of the H-B complex as the difference between the energy of the H-B complex in a Si crystal and the sum of the energies of a (neutral) H atom in free space and of a (neutral) B substitutional in Si. This binding energy is more of a conceptual quantity, contrary to the dissociation energy discussed above (which, however, is often called a binding energy as well). According to this definition a binding energy of 3.31 eV is obtained.[58]

## IV. CONCLUSIONS

The study of the total-energy surface for H in Si:B using the first-principles pseudopotential-density-functional method presented in this paper conclusively shows that a H-B complex is formed in which the H atom occupies a site close to the center of a Si—B bond (*BM* site). This complex is the net result of the passivation mechanism that removes the shallow-acceptor level from the gap, thereby neutralizing the electrical activity of boron-doped silicon. Other sites that were previously proposed to be equilibrium sites for H by others are shown to be saddle points of the energy surface that are higher in energy by at least 0.48 eV. We find that the H atom can move between four equivalent *BM* sites over a spherical shell-like region with an energy barrier of only 0.2 eV.

The calculated vibrational frequency for the H-stretching mode centered on the *BM* site is in good agreement with infrared and Raman experiments. The occurrence of sidebands in the infrared spectrum can be qualitatively understood since H can reside slightly off the bond axis from the *BM* site.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Office of Naval Research under Contract No. N00014-84-C-0396. One of us (P.J.H.D.) acknowledges support from IBM Netherlands N.V. We further acknowledge useful discussions with Dr. A. D. Marwick.

*Present address: Physics Department, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands.

†Present address: Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, NY 10510.

[1]S. J. Pearton, J. W. Corbett, and T. S. Shi, Appl. Phys. A **43**, 153 (1987).

[2]An overview of recent work may be found in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), pp. 229–309.

[3]C. T. Sah, J. Y. C. Sun, and J. J. T. Tzou, Appl. Phys. Lett. **43**, 204 (1983); J. Appl. Phys. **54**, 5864 (1983).

[4]J. I. Pankove, D. E. Carlson, J. E. Berkeyheiser, and R. O. Wance, Phys. Rev. Lett. **51**, 2224 (1983); J. I. Pankove, R. O. Wance, and J. E. Berkeyheiser, Appl. Phys. Lett. **45**, 1100 (1984); J. I. Pankove, P. J. Zanzucchi, and C. W. Magee, *ibid.* **46**, 421 (1985).

[5]N. M. Johnson and M. D. Moyer, Appl. Phys. Lett. **46**, 787 (1985); N. M. Johnson, Phys. Rev. B **31**, 5525 (1985).

[6]N. M. Johnson, C. Herring, and D. J. Chadi, Phys. Rev. Lett. **56**, 769 (1986); **59**, 2116 (1987); N. M. Johnson and C. Herring, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 277.

[7]K. Bergman, M. Stavola, S. J. Pearton, and J. Lopata, Phys. Rev. B **37**, 2770 (1988).

[8]S. T. Pantelides, Appl. Phys. Lett. **50**, 995 (1987).

[9]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **60**, 2761 (1988).

[10]G. G. DeLeo and W. B. Fowler, Phys. Rev. B **31**, 6861 (1985).

[11]L. V. C. Assali and J. R. Leite, Phys. Rev. Lett. **55**, 980 (1985); **56**, 403 (1986).

[12]G. G. DeLeo and W. B. Fowler, Phys. Rev. Lett. **56**, 402 (1986).

[13]J. M. Baranowski and J. Tatarkiewicz, Phys. Rev. B **35**, 7450 (1987).

[14]A. Amore Bonapasta, A. Lapiccirella, N. Tomassini, and M. Capizzi, Phys. Rev. B **36**, 6228 (1987).

[15]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Appl. Phys. Lett. **50**, 1086 (1987).

[16]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Phys. Rev. B **37**, 8313 (1988).

[17]M. Stutzmann, Phys. Rev. B **35**, 5921 (1987); M. Stutzmann and C. P. Herrero, Appl. Phys. Lett. **51**, 1413 (1987).

[18]A. D. Marwick, G. S. Oehrlein, and N. M. Johnson, Phys. Rev. B **36**, 4539 (1987).

[19]A. D. Marwick, G. S. Oehrlein, J. H. Barrett, and N. M. Johnson, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 14, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 259.

[20]B. B. Nielsen, J. U. Andersen, and S. J. Pearton, Phys. Rev. Lett. **60**, 321 (1988).

[21]Th. Wichert, H. Skudlik, H.-D. Carstanjen, T. Enders, M. Deicher, G. Grübel, R. Keller, L. Song, and M. Stutzmann, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 265.

[22]Th. Wichert, H. Skudlik, M. Deicher, G. Grübel, R. Keller, E. Recknagel, and L. Song, Phys. Rev. Lett. **59**, 2087 (1987).

[23]M. Stutzmann, J. Harsanyi, A. Breitschwerdt, and C. P. Herrero, Appl. Phys. Lett. **52**, 1667 (1988).

[24]M. Stavola, K. Bergman, S. J. Pearton, and J. Lopata, Phys. Rev. Lett. **61**, 2786 (1988).

[25]P. Deák, L. C. Snyder, R. K. Singh, and J. W. Corbett, Phys. Rev. B **36**, 9612 (1987); P. Deák and L. C. Snyder, *ibid.* **36**, 9619 (1987).

[26]K. J. Chang and D. J. Chadi, Phys. Rev. Lett. **60**, 1422 (1988).

[27]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).

[28]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[29]J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[30]D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[31]J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).

[32]P. J. H. Denteneer, Ph.D. thesis, Eindhoven University of Technology, 1987, available from the author upon request.

[33]S. G. Louie, in *Electronic Structure, Dynamics, and Quantum Structural Properties of Condensed Matter*, edited by J. T. Devreese and P. E. van Camp (Plenum, New York, 1985).

[34]Y. Bar-Yam, S T. Pantelides, and J. D. Joannopoulos, Phys. Rev. B **39**, 3396 (1989).

[35]C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, the preceding paper, Phys. Rev. B **39**, 10 791 (1989).

[36]R. M. Wentzcovitch, M. L. Cohen, and P. K. Lam, Phys. Rev. B **36**, 6058 (1987).

[37]P. O. Löwdin, J. Chem. Phys. **19**, 1396 (1951).

[38]G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).

[39]Two special points (see Refs. 43 and 44) are used in these calculations to integrate over the first Brillouin zone.

[40]F. D. Murnaghan, Proc. Nat. Acad. Sci. U.S.A. **30**, 244 (1944).

[41]*Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, edited by O. Madelung, M. Schulz, and H. Weiss (Springer, Berlin, 1982), Gp. 3, Vol. 17, Pt. a.

[42]Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B **30**, 1844 (1984).

[43]A. Baldereschi, Phys. Rev. B **7**, 5212 (1973); D. J. Chadi and M. L. Cohen, *ibid.* **8**, 5747 (1973).

[44]H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[45]M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986); R. W. Godby, M. Schlüter, and L. J. Sham, *ibid.* **37**, 10 159 (1988).

[46]J. A. Vergés, D. Glötzel, M. Cardona, and O. K. Andersen, Phys. Status Solidi B **113**, 519 (1982).

Here the two-dimensional integral is performed over the saddle surface $S_{ij}$ separating cells $i$ and $j$, and the three-dimensional integral is performed over the cell $V_i$ at $\mathbf{X}_i$. The probability distribution is given by

$$P(\mathbf{x}) = Z^{-1}\exp[-F(\mathbf{x},T)/k_BT], \tag{5}$$

where $Z$ is a normalization constant and $F(\mathbf{x},T)$ is the free energy of the system with the impurity constrained to a position $\mathbf{x}$. This free energy is in general given by

$$F(\mathbf{x},T) = -k_BT\ln\lambda_T^{-(3N-6)}\int dq^{3N-3}\delta[\mathbf{x}-\mathbf{x}(\mathbf{q})]\exp\left[-\frac{1}{k_BT}V(\mathbf{q})\right], \tag{6}$$

where $\mathbf{x}(\mathbf{q})$ represents the functional dependence of the impurity's position relative to the host-atom positions, $V(\mathbf{q})$ is the potential energy depending on the positions $\mathbf{q}=(\mathbf{q}_1,\ldots,\mathbf{q}_N)$ of all $N$ atoms in the crystal, and $\lambda_T$ is the thermal de Broglie wavelength. The total translational degree of freedom has been excluded in the integration.

The above formalism is exact. For a practical implementation we introduce only two approximations. First, we take $\Gamma=\Gamma^0$, setting the so-called efficiency factor equal to 1. Efficiency factors have been studied for model solids and have in many cases been found to change the results for the diffusion constant by only a small fraction.[1,2,11,14] The second approximation of this approach is the way we calculate $F(\mathbf{x},T)$. We note that

$$F(\mathbf{x},T) = F(\mathbf{x},0) - \int_0^T dT\,S(\mathbf{x},T), \tag{7}$$

where $S(\mathbf{x},T)$ is the entropy of the system *with the impurity fixed at* $\mathbf{x}$. We propose to approximate $F(\mathbf{x},T)$ by its value at $T=0$. The free energy $F(\mathbf{x},0)$ is the total energy obtained with the impurity at $\mathbf{x}$ and all other atoms relaxed. This approximation corresponds to the assumption that the vibrational frequencies of the *host* atoms depend only weakly on the position of the impurity, when the latter is fixed. The second term in Eq. (7) can actually be calculated in a rather straightforward but time-consuming manner in the local harmonic approximation.[15] Such calculations (to be discussed below) show, however, that this contribution to $F(\mathbf{x},T)$ is small. In many cases of interest this term can therefore be neglected, making the method very practical.

The present approach reduces the many-body problem of treating the diffusing particle together with all the particles in the embedding crystal to the problem of diffusion of a single particle in a three-dimensional effective potential. This effective potential has the full space-group symmetry of the crystal, and can, therefore, be expanded in symmetrized plane waves.[7] Thus, calculations of the total energy at only a few selected sites can be used to determine an analytic form for the complete effective potential.

The total-energy surface, which we will use to extract the diffusion constants for $H^+$, has been obtained from state-of-the-art electronic structure calculations using the local-density approximation and *ab initio* norm-conserving pseudopotentials.[7] The proton is treated, like all other nuclei, as a classical particle. The impurity is placed in a supercell with 32 Si atoms and the atomic positions up to second-nearest neighbors are relaxed. The total energy is calculated for eight inequivalent positions of the impurity. These values have been used to determine the expansion coefficients for a suitable set of symmetrized plane waves. The resulting analytical expression provides the free-energy surface $F(\mathbf{x},0)$ from which we calculate diffusion coefficients. The relevant total-energy surfaces have been published in Ref. 7.

In order to test the approximation of neglecting the host-atom entropy [second term of Eq. (7)], we calculated the vibrational frequencies of the host atoms. Even though this can, in principle, be done with *ab initio* calculations, we have used here a generalized Keating potential[16] for the Si-Si interactions and a suitably chosen Morse potential for the H-Si interactions.[17] The resulting entropy, obtained within the local harmonic approximation, was found to differ by $-1k_B$ between the most dissimilar sites [e.g., the bond-center (BC) and tetrahedral sites] and significantly less between rather similar sites. At 1000 K these differences translate into corrections of less than $-0.1$ eV for $F(\mathbf{x},T)$. That is precisely the level of accuracy with which $F(\mathbf{x},0)$ can be calculated from first principles.[7] We expect this to be a conservative estimate for other impurities as well because of the relatively large relaxations of neighboring Si atoms when H is at the BC position. Thus a time-consuming calculation of this term from first principles is not warranted. We did test, however, the effect of 0.1-eV uncertainty in $F(\mathbf{x},0)$ on the final diffusion coefficients and results are given below.

For the calculation of the diffusion coefficient, we first determine the minima $\mathbf{X}_i$ in the total-energy surface. These form the so-called diffusion-site lattice. A Wigner-Seitz-like construction then yields the volumes $V_i$, related to the sites $\mathbf{X}_i$, and the saddle surfaces $S_{ij}$ as the faces of the Wigner-Seitz polyhedra. For the positively charged state of H there exists only one set of equivalent minima $\mathbf{X}_i$. These lie in a disk-shaped region of almost degenerate sites centered at the BC position.[18] H atoms can hop from bond to bond via the so-called $C$ site, which is located midway between two second-nearest Si neighbors. The Wigner-Seitz cell of the diffusion-site lattice formed by the BC sites has the shape of a rhombohedron (see Fig. 1). Every face con-

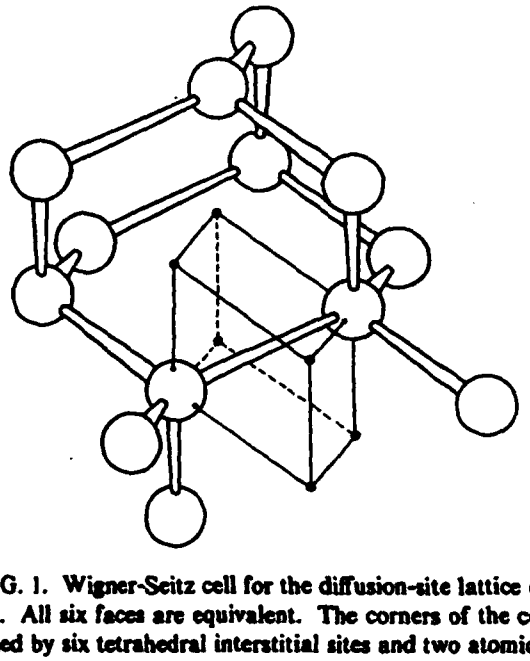


FIG. 1. Wigner-Seitz cell for the diffusion-site lattice of $H^+$ in Si. All six faces are equivalent. The corners of the cell are formed by six tetrahedral interstitial sites and two atomic sites. $C$ sites are located in the center of the faces. The hexagonal interstitial sites lie on the edges of the cell, midway between two neighboring tetrahedral interstitial sites.

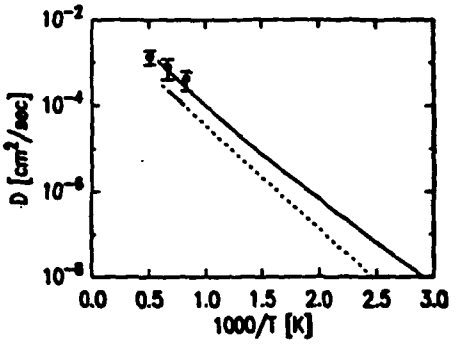



FIG. 2. Diffusion coefficients for $H^+$ in $c$-Si as a function of inverse temperature. Solid line: present calculation for $H^+$; solid circles, theoretical results of Buda *et al.* (Ref. 9); dotted line: $D=9.41\times10^{-5}\exp[(-0.48\ \text{eV})/k_BT]$ as obtained by van Wieringen and Warmoltz (Ref. 19) (the solid part indicates the actual temperature range of the experiments).

tains one $C$ site as the only saddle point. The occupation and the saddle-surface probability are obtained from integrals over the Wigner-Seitz polyhedron and one of its faces, respectively.

Our results for the diffusion constant are shown in Fig. 2. They agree with the theoretical results of Buda *et al.*[9(a)] within their error bars, and with the experimental results of van Wieringen and Warmoltz[19] within a factor of 3, which is within the expectations for the accuracy of the method and the accuracy of the measurements.

The number of distinct diffusion pathways can always be obtained from the number of inequivalent saddle points on the saddle surface. Their relative contribution can be obtained from partial integrals over the saddle surface. The total-energy surface of $H^+$ in Si has only equivalent minima and only identical saddle points. Thus, according to our definition, we have only one pathway. At higher temperatures, however, the impurity need not pass through total-energy minima or saddle points, giving rise to a variety of trajectories. For example, the impurity may diffuse by cutting through bonds or it may merely "rub" against the bond without actually crossing it. Anharmonic effects in the effective potential, which are evident from the curvature in the Arrhenius plot of Fig. 2, enhance the possibility of trajectories that avoid the bond region at high temperatures.

Our approach does not seek to determine the actual trajectories. Instead, it determines the diffusion coefficient as an integral over all possible trajectories. The approach assumes implicitly that the motion of H is randomized after each saddle-surface crossing. Thus, it neglects the effect of dynamical correlations, i.e., trajectories in which successive saddle-surface crossings occur in a correlated fashion. In contrast, Buda *et al.* carry out a time integral over the trajectories that actually occur. The two "paths" identified by them are, in our terminology, two dynamical trajectories. The excellent agreement between the results in Ref. 9(a) and ours seems to indicate that dynamical correlations, even though extremely important for the time evolution of the system, do not substantially affect the value for the diffusion constant. This result is consistent with earlier findings.[14]

The accuracy of our calculation depends of course on the accuracy of the energy surface used as input. The error bar is related to the accuracy of the individual total-energy calculations, as well as to the number of calculated points used for the analytic representation of the energy surface. Both contributions have been estimated to be on the order of 0.1 eV.[7] As we saw above, the correction from the host-atom entropy is even smaller. In order to see how sensitive our results are to these uncertainties, we have artificially introduced changes of the order of 0.1 eV in the regions where they count the most, i.e., the stable site and the saddle surface. The resulting changes in the activation barrier are comparable to the changes in the total-energy surface. The preexponential, obtained from the Arrhenius plot at high temperatures, however, is surprisingly insensitive and varies only by a factor of 5. This is understandable because at high temperatures the impurity explores a large region of phase space and is therefore insensitive to small local changes in the total-energy surface.

In conclusion, we have shown that diffusion constants can be calculated with considerable accuracy from static total-energy calculations. The technique is applicable to systems with low or high activation barriers and is valid over a wide temperature range. We have applied this approach to the calculation of the diffusion constant of

H in Si. Our results compare well with experiment and recent calculations of Buda *et al.*,[9(a)] which describe the time evolution of all particles without approximations.

This work was supported in part by the Office of Naval Research Contract No. N00014-84-0396. We are grateful to R. Car and J. Tersoff for helpful discussions.

[1]C. H. Bennett, in *Diffusion in Solids: Recent Developments*, edited by A. S. Nowick and J. J. Burton (Academic, New York, 1975), p. 73.

[2]G. Jacucci, in *Diffusion in Crystalline Solids*, edited by G. E. Murch and A. S. Nowick (Academic, New York, 1984), p. 429.

[3]M. J. Gillan, J. H. Harding, and R.-J. Tarento, J. Phys. C **20**, 2331 (1987).

[4]A. M. Stoneham, Phys. Scr. **T25**, 17 (1989).

[5]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. **52**, 1814 (1984); **54**, 360 (1985).

[6]K. C. Pandey, Phys. Rev. Lett. **57**, 2287 (1986).

[7]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **60**, 2761 (1988); C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B **39**, 10791 (1989).

[8]C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. **62**, 1049 (1989).

[9](a) F. Buda, G. L. Chiarotti, R. Car, and M. Parinello, Phys. Rev. Lett. **63**, 294 (1989; (b) (private communication).

[10]D. Chandler, J. Chem. Phys. **68**, 2959 (1977).

[11]A. F. Voter, Phys. Rev. Lett. **63**, 167 (1989); A. F. Voter and J. D. Doll, J. Chem. Phys. **82**, 80 (1985).

[12]G. Vineyard, J. Phys. Chem. Solids **3**, 121 (1957).

[13]One must chose a coordinate x relative to the positions of the host atoms in order to exclude the overall translational degree of freedom. If the generalized coordinate $x(\mathbf{q})$ depends only linearly on the atomic positions $\mathbf{q}$, the reduced mass $\mu$ is given by

$$\frac{1}{\mu} = \sum_{i}^{3N} \left( \mathbf{n} \frac{\partial x}{\partial q_i} \right)^2 \frac{1}{m_i},$$

where $\mathbf{n}$ is the normal vector of the saddle surface.

[14]G. DeLorenci and G. Jacucci, Phys. Rev. B **33**, 1993 (1985).

[15]R. LeSar, R. Najafabadi, and D. J. Srolovitz, Phys. Rev. Lett. **63**, 624 (1989).

[16]D. Vanderbilt, S. H. Taole, and S. Narasimhan, Phys. Rev. B **40**, 5657 (1989).

[17]P. E. Blöchl, C. G. Van de Walle, and S. T. Pantelides (to be published).

[18]The analytic form of the total-energy surface actually has minima that are slightly away from the BC sites. However, the energy difference from the BC site is too small to be resolved. We will still refer to the BC site as the stable site, because equilibration between those sites is too rapid to change the picture.

[19]A. van Wieringen and N. Warmoltz, Physica (Utrecht) **22**, 849 (1956).

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Office of Naval Research under Contract No. N00014-84-C-0396. One of us (P.J.H.D.) acknowledges support from IBM Netherlands N.V. We further acknowledge useful discussions with Dr. A. D. Marwick.

*Present address: Physics Department, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands.

†Present address: Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, NY 10510.

[1]S. J. Pearton, J. W. Corbett, and T. S. Shi, Appl. Phys. A **43**, 153 (1987).

[2]An overview of recent work may be found in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), pp. 229–309.

[3]C. T. Sah, J. Y. C. Sun, and J. J. T. Tzou, Appl. Phys. Lett. **43**, 204 (1983); J. Appl. Phys. **54**, 5864 (1983).

[4]J. I. Pankove, D. E. Carlson, J. E. Berkeyheiser, and R. O. Wance, Phys. Rev. Lett. **51**, 2224 (1983); J. I. Pankove, R. O. Wance, and J. E. Berkeyheiser, Appl. Phys. Lett. **45**, 1100 (1984); J. I. Pankove, P. J. Zanzucchi, and C. W. Magee, *ibid.* **46**, 421 (1985).

[5]N. M. Johnson and M. D. Moyer, Appl. Phys. Lett. **46**, 787 (1985); N. M. Johnson, Phys. Rev. B **31**, 5525 (1985).

[6]N. M. Johnson, C. Herring, and D. J. Chadi, Phys. Rev. Lett. **56**, 769 (1986); **59**, 2116 (1987); N. M. Johnson and C. Herring, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 277.

[7]K. Bergman, M. Stavola, S. J. Pearton, and J. Lopata, Phys. Rev. B **37**, 2770 (1988).

[8]S. T. Pantelides, Appl. Phys. Lett. **50**, 995 (1987).

[9]C. G. Van de Walle, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **60**, 2761 (1988).

[10]G. G. DeLeo and W. B. Fowler, Phys. Rev. B **31**, 6861 (1985).

[11]L. V. C. Assali and J. R. Leite, Phys. Rev. Lett. **55**, 980 (1985); **56**, 403 (1986).

[12]G. G. DeLeo and W. B. Fowler, Phys. Rev. Lett. **56**, 402 (1986).

[13]J. M. Baranowski and J. Tatarkiewicz, Phys. Rev. B **35**, 7450 (1987).

[14]A. Amore Bonapasta, A. Lapiccirella, N. Tomassini, and M. Capizzi, Phys. Rev. B **36**, 6228 (1987).

[15]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Appl. Phys. Lett. **50**, 1086 (1987).

[16]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Phys. Rev. B **37**, 8313 (1988).

[17]M. Stutzmann, Phys. Rev. B **35**, 5921 (1987); M. Stutzmann and C. P. Herrero, Appl. Phys. Lett. **51**, 1413 (1987).

[18]A. D. Marwick, G. S. Oehrlein, and N. M. Johnson, Phys. Rev. B **36**, 4539 (1987).

[19]A. D. Marwick, G. S. Oehrlein, J. H. Barrett, and N. M. Johnson, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 14, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 259.

[20]B. B. Nielsen, J. U. Andersen, and S. J. Pearton, Phys. Rev. Lett. **60**, 321 (1988).

[21]Th. Wichert, H. Skudlik, H.-D. Carstanjen, T. Enders, M. Deicher, G. Grübel, R. Keller, L. Song, and M. Stutzmann, in *Defects in Electronic Materials*, Materials Research Society Symposia Proceedings Vol. 104, edited by M. Stavola, S. J. Pearton, and G. Davies (Materials Research Society, Pittsburgh, PA, 1988), p. 265.

[22]Th. Wichert, H. Skudlik, M. Deicher, G. Grübel, R. Keller, E. Recknagel, and L. Song, Phys. Rev. Lett. **59**, 2087 (1987).

[23]M. Stutzmann, J. Harsanyi, A. Breitschwerdt, and C. P. Herrero, Appl. Phys. Lett. **52**, 1667 (1988).

[24]M. Stavola, K. Bergman, S. J. Pearton, and J. Lopata, Phys. Rev. Lett. **61**, 2786 (1988).

[25]P. Deák, L. C. Snyder, R. K. Singh, and J. W. Corbett, Phys. Rev. B **36**, 9612 (1987); P. Deák and L. C. Snyder, *ibid.* **36**, 9619 (1987).

[26]K. J. Chang and D. J. Chadi, Phys. Rev. Lett. **60**, 1422 (1988).

[27]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).

[28]D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[29]J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[30]D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[31]J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).

[32]P. J. H. Denteneer, Ph.D. thesis, Eindhoven University of Technology, 1987, available from the author upon request.

[33]S. G. Louie, in *Electronic Structure, Dynamics, and Quantum Structural Properties of Condensed Matter*, edited by J. T. Devreese and P. E. van Camp (Plenum, New York, 1985).

[34]Y. Bar-Yam, S. T. Pantelides, and J. D. Joannopoulos, Phys. Rev. B **39**, 3396 (1989).

[35]C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, the preceding paper, Phys. Rev. B **39**, 10 791 (1989).

[36]R. M. Wentzcovitch, M. L. Cohen, and P. K. Lam, Phys. Rev. B **36**, 6058 (1987).

[37]P. O. Löwdin, J. Chem. Phys. **19**, 1396 (1951).

[38]G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).

[39]Two special points (see Refs. 43 and 44) are used in these calculations to integrate over the first Brillouin zone.

[40]F. D. Murnaghan, Proc. Nat. Acad. Sci. U.S.A. **30**, 244 (1944).

[41]*Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, edited by O. Madelung, M. Schulz, and H. Weiss (Springer, Berlin, 1982), Gp. 3, Vol. 17, Pt. a.

[42]Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B **30**, 1844 (1984).

[43]A. Baldereschi, Phys. Rev. B **7**, 5212 (1973); D. J. Chadi and M. L. Cohen, *ibid.* **8**, 5747 (1973).

[44]H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[45]M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986); R. W. Godby, M. Schlüter, and L. J. Sham, *ibid.* **37**, 10 159 (1988).

[46]J. A. Vergés, D. Glötzel, M. Cardona, and O. K. Andersen, Phys. Status Solidi B **113**, 519 (1982).

[47]K. Shih, W. E. Spicer, W. A. Harrison, and A. Sher, Phys. Rev. B 31, 1139 (1985).
[48]E. A. Kraut and W. A. Harrison, J. Vac. Sci. Technol. B 3, 1267 (1985).
[49]L. S. Bartell and B. L. Caroll, J. Chem. Phys. 40, 1135 (1965).
[50]J. H. Barrett, Phys. Rev. B 3, 1527 (1971).
[51]These calculations were done in a 32-atom cell with energy cutoffs of (10;20) Ry and two special k points.
[52]E. Kaxiras and J. D. Joannopoulos, Phys. Rev. B 37, 8842 (1988).
[53]M. H. Brodsky, M. Cardona, and J. J. Cuomo, Phys. Rev. B 16, 3556 (1977).
[54]H. J. Stein, Phys. Rev. Lett. 43, 1030 (1979).
[55]W. R. Thorson and I. Nakagawa, J. Chem. Phys. 33, 994 (1960).
[56]D. R. Bosomworth, W. Hayes, A. R. L. Spray, and G. D. Watkins, Proc. R. Soc. London, Ser. A 317, 133 (1970).
[57]M. Stavola (private communication).
[58]Since the crystal calculations do not include spin-polarization effects, we calculate the energy of the H atom also without spin polarization. Our calculated number is quantitatively rather unreliable since a calculation in a crystal is compared with an atomic calculation and these are very different regarding approximations that are made. This is illustrated by the fact that if the two crystal calculations are performed with the lower cutoffs (6;12) Ry [as opposed to (10;20) Ry for the quoted value], the binding energy becomes 2.75 eV.

PHYSICAL REVIEW B VOLUME 41, NUMBER 6 15 FEBRUARY 1990-II

# Microscopic structure of the hydrogen-phosphorus complex in crystalline silicon

P. J. H. Denteneer
*Faculty of Science, Electronic Structure of Materials, Catholic University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*

C. G. Van de Walle
*Philips Laboratories, North American Philips Corporation, 345 Scarborough Road, Briarcliff Manor, New York 10510*

S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 8 August 1989)

The existing discrepancy between theoretical models and experimental results for hydrogen-donor complexes in crystalline silicon is resolved using first-principles pseudopotential-density-functional calculations for the hydrogen-phosphorus pair. In the configuration which is the global energy minimum, H is located on the extension of a P–Si bond on the Si side, with the Si-H pair relaxing away from P by 0.6 Å, leaving the P atom threefold coordinated. The calculated stretching and wagging vibrational frequencies associated with this configuration are in accord with experiment.

The structure and properties of hydrogen-impurity complexes in semiconductors have been studied intensively in the last few years using both experimental and theoretical methods.[1–12] For the hydrogen-boron complex in silicon, which is the prototypal hydrogen-acceptor complex that has been studied most elaborately, a consistent picture has emerged (see, e.g., Ref. 5 and references therein). In the equilibrium configuration of the complex, the H atom resides inside a Si–B bond, forming a three-center bond. Also for the qualitatively different (H,Be) pair in Si and (H,Si) pair in Ge, theory has provided satisfactory explanations of the experimental results as well as new insights (see Ref. 6 and references therein). The H atom in the (H,Be) pair is able to tunnel around the Be atom because its lowest-energy location is close to the *C* site (midway between two Si atoms bonded to Be), where the relaxation of the surrounding Si atoms is small. The H atom in the (H,Si) complex in Ge is located close to a tetrahedral interstitial ($T_d$) site.

In contrast, the structure of hydrogen-donor complexes, e.g., (H,P) in Si, has so far not been determined conclusively. Experiments[8] have shown that all H-donor pairs in Si have similar infrared absorption spectra, suggesting that H is *not* bonded to the donor. The observation of a nondegenerate stretching mode around 1560 $cm^{-1}$ and a doubly degenerate wagging mode around 810 $cm^{-1}$ suggests that the center has trigonal symmetry. Theoretical models have so far not reproduced these frequencies. In Ref. 7, a model was proposed in which H is located on the extension of a P–Si bond on the side of Si. Using empirical tight-binding calculations this "AB (antibonding) of Si" configuration was found to be lower in energy than the "AB of P" configuration. The frequency for the H stretching mode was calculated to be 2145 $cm^{-1}$, which is very different from the experimentally determined value of 1555 $cm^{-1}$.[8] In a subsequent calculation by the same group, but using the more reliable first-principles pseudopotential-density-functional method, the configuration was qualitatively confirmed.[10] However, in the latter calculation the stretching mode was found to be at 400 $cm^{-1}$.

Recently, a number of groups using various kinds of cluster calculations[4,11,12] have proposed a configuration similar to the one in Ref. 10 with the distinction, however, that the Si atom closest to H relaxes from its lattice site towards H to become almost coplanar with its three nearest-neighbor Si atoms. Estreicher *et al.*[11] discuss the inherent difficulties in calculating vibrational frequencies to within a reasonable accuracy using quantum-mechanical cluster calculations and do not attempt to calculate any frequency. DeLeo and Fowler[4] and Amore Bonapasta *et al.*[12] calculate a H stretching frequency of 2150 $cm^{-1}$, again in disagreement with experiment.

Summarizing, it can be said that theoretical studies so far have not been able to put forward a microscopic model for the (H,P) complex that can be conclusively identified as the one that is experimentally observed.

In this paper, we present results of accurate first-principles calculations for the (H,P) pair. We determine the lowest-energy configuration and show that this configuration is responsible for stretching and wagging vibrational frequencies that are in agreement with experiment. We have successfully used the pseudopotential-density-functional method before in studies of H in pure Si and of various complexes in Si and Ge.[5,6,9] If the calculations are properly converged with respect to all the numerical approximations involved, the method is very reliable in determining defect configurations. In particular, total-energy differences between different defect configurations can be calculated to within an accuracy of 0.05–0.1 eV and typical H vibrational frequencies can be calculated with an accuracy of about 100 $cm^{-1}$ (Ref. 5). For details of calculations in which such accuracy is achieved we refer to Refs. 5 and 9. In the present study, we closely examine various configurations with trigonal symmetry (see below), as well as the regions close to the *C*

and $C'$ sites (i.e, H midway between two Si atoms bonded to P and H midway between P and a next-nearest-neighbor Si, respectively).[5] The $C$ and $C'$ sites are at least 1.5 eV higher in energy than the lowest-energy configuration; we will not consider them further. The configurations with trigonal symmetry, which can be classified according to the order in which the H, Si, and P atoms are found along a ⟨111⟩ axis (H-Si-P, Si-H-P, and Si-P-H, respectively), are optimized by relaxing up to nine atoms according to the Hellmann-Feynman forces on these atoms. These forces can be calculated with the same level of accuracy as total energies from the self-consistent solutions of the Schrödinger equation for the valence electrons.[13] In order to optimize the configurations we move the atoms in the direction of the calculated forces until the forces become negligible, thereby minimizing the total energy.

We find that each of the three trigonal-symmetry configurations, including appropriate relaxations of all the atoms, constitutes a local minimum of the total-energy surface. Furthermore, the three minima are very close in energy: they all lie in an energy range of only 0.5 eV (see Fig. 1). These small energy differences open the way for the occurrence of metastable states of the complex.

Now we describe the two local minima and one global minimum configurations mentioned above. Of these three, the configuration highest in energy is the one in which H resides between a Si and P atom forming a bond. We call this configuration "BC (LLR of P)" since it involves a very large lattice relaxation (LLR) of the P atom (BC stands for bond-center site). The P atom relaxes outward (away from H) by 1.22 Å, whereas the Si atom relaxes outward by only 0.10 Å. The H-Si distance in this configuration is 1.50 Å, similar to the H-Si distances found in molecules, e.g., $SiH_4$, and at a hydrogenated vacancy. The H atom breaks the Si−P bond and saturates the Si dangling bond; this allows for the large relaxation of P through the plane of its three neighboring Si atoms to a position where it is threefold coordinated. The charge density for this configuration is shown in Fig. 2(a) and displays a lone pair on the P atom pointing in the direction of the nearest $T_d$ site on the line Si-H-P. The H−Si bond that is formed has a calculated stretch frequency of 1900 $cm^{-1}$, much larger than the observed frequency. In the other local minimum configuration, which we call AB of P, the H atom is located very close to the $T_d$ site closest to the P atom. The energy of AB of P is only 0.10 eV lower than that of BC (LLR of P) (see Fig. 1). In this configuration, none of the atoms relax appreciably from their ideal lattice position, resulting in a H-P distance of almost an undistorted Si−Si bond length (2.35 Å). The calculated H stretch frequency for this configuration is 570 $cm^{-1}$, much smaller than the observed frequency. Finally, the global energy minimum configuration is the one called AB of Si (LLR of Si). It has an energy 0.35 eV

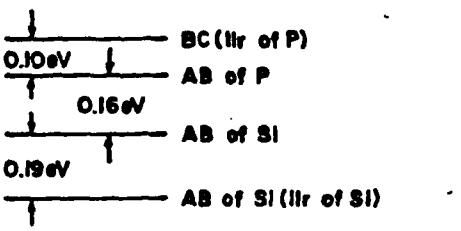



FIG. 1. Relative energies of different configurations with trigonal symmetry for (H,Si,P) complexes in silicon. AB stands for antibonding site, BC for bond-center site, and LLR for large lattice relaxation. A more detailed description of the four configurations is given in the text (see also Fig. 3).

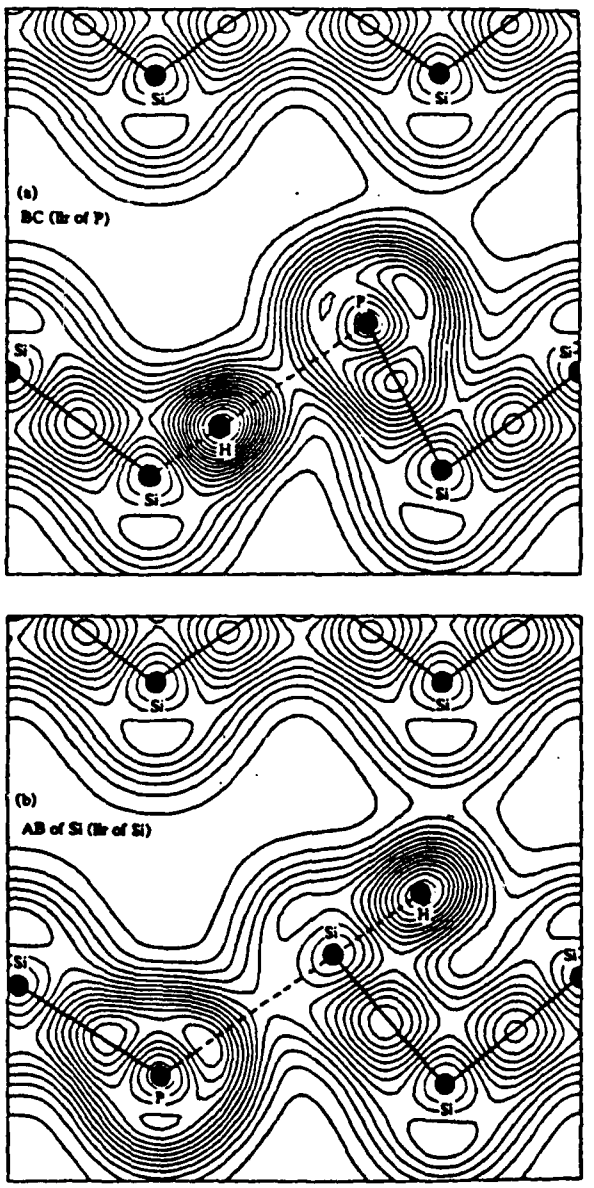



FIG. 2. Total valence charge density in the (110) plane for (a) the BC (LLR of P) and (b) the AB of Si (LLR of Si) configurations for a (H,Si,P) complex in Si. The black dots indicate atomic positions and the straight lines connect bonded atoms. The broken Si−P bond and the Si−H bond are indicated by dashed lines. The contour spacing is 1.87 $e/\Omega$, where $\Omega$ is the unit cell volume of bulk Si (which contains 8 electrons in bulk Si). The lowest-density contour shown (in the channels between the two atomic chains) is 2.32 $e/\Omega$ and the highest-density contour shown (around the H atomic position) is 34.1 $e/\Omega$ in (a) and 28.5 $e/\Omega$ in (b). The maximum density in a Si−Si bond is 24.0 $e/\Omega$.

lower than AB of P, and H is located close to the $T_d$ site of a Si atom bonded to P. This Si atom relaxes outward by 0.59 Å (leaving the P atom threefold coordinated; see Fig. 3). The P atom relaxes by the small amount of 0.14 Å (in the direction of the Si relaxation, contrary to the results of cluster calculations). The H-Si distance is 1.66 Å, which is somewhat larger than a typical value for a H−Si bond distance (see above), indicating a slight weakening of the bond. The H stretch frequency is therefore expected to be lower than for a typical Si−H bond. Indeed, we calculate a frequency of 1460 $cm^{-1}$, which in view of the error bar on calculated frequencies discussed above, is in agreement with the experimental number of 1555 $cm^{-1}$. Also the calculated frequency of the H wagging mode of 740 $cm^{-1}$ is in agreement with the experimental result of 809 $cm^{-1}$. The agreement of both calculated frequencies with experiment, taken together with the fact that the AB of Si (LLR of Si) configuration has the lowest energy of all configurations studied justifies the identification of the experimentally observed complex with this AB of Si (LLR of Si) configuration.

In Fig. 2(b), we show the valence charge density of the (H,P) pair in the AB of Si (LLR of Si) configuration. The P−Si bond is effectively broken and a lone-pair-like density, which is a remnant of the previous P−Si bond, is extending in the direction of the former bond. All the valence electrons of P are accounted for in this way. The Si atom has gone from an $sp^3$ bonding configuration to an $sp^2$ bonding configuration with respect to its three Si neighbors. The surplus electron of Si (which does not have to go in a P−Si bond) pairs with the H electron to form a Si−H bond. Indeed, the charge density between Si and H is very similar to the one found in the case of H saturating a Si dangling bond. Bonding is indicated by the fact that the charge density around the H atom is clearly modified from the spherical form it has when Si and H are far apart (see, e.g., the charge density for the AB of Si configuration in Ref. 10).

For the sake of completeness and to make the connection with the results of other work, we mention that if we do not allow for relaxation of the Si neighbors of the Si atom between H and P, this Si atom relaxes outward by only 0.19 Å. This results in a AB of Si (without *large* lattice relaxation of Si) configuration which is still lower in energy by 0.16 eV than the AB of P configuration (see Fig. 1), but higher in energy by 0.19 eV than the AB of Si (LLR of Si) configuration. For this AB of Si configuration, which is similar to the one found in Ref. 10, the H-Si distance is 2.1 Å, much larger than a typical H−Si bond distance, and the corresponding H stretch frequency is calculated to be 600 $cm^{-1}$. The H wagging mode for this configuration has a calculated frequency of 600 $cm^{-1}$ as well, indicating the absence of H bonding. The configuration that we find to be lowest in energy is almost the same as the one found in Refs. 4, 11, and 12. In those calculations, the Si atom relaxes by an amount between

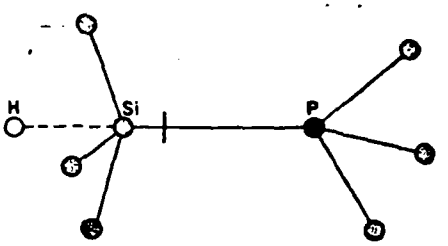



FIG. 3. Schematic representation of the AB of Si (LLR of Si) configuration, which is the lowest-energy configuration for a (H,Si,P) complex in Si (see also Fig. 1). One Si atom has relaxed from its lattice position (indicated by a vertical bar) by 0.59 Å towards H and is only 0.19 Å away from being coplanar with its three Si neighbors.

0.6 and 0.8 Å and a Si-H distance between 1.4 and 1.5 Å is found, which is smaller than our value of 1.66 Å. Consequently, those calculations render a much larger stretching mode frequency of about 2150 $cm^{-1}$. More recently, Chadi *et al.*[14] repeated the calculations of Ref. 10 and found similar results to those presented here by us.

Both configurations with large lattice relaxations discussed above are reminiscent of recently proposed models for the *EL*2 and *DX* defect centers in GaAs.[15,16] In the case of *EL*2, it is proposed that an As antisite can be induced by optical excitation to move by about 1.3 Å from its lattice position to a metastable configuration.[15] In the case of the *DX* center, a Si donor in GaAs may move 1.2 Å from its lattice site.[16] In both cases, the configuration with a large lattice relaxation is inherently associated with a simple point defect and can be provoked to materialize. In the subject of our present study, it is the H atom with its one unpaired electron that is able to promote different bonding environments for the simple substitutional P donor involving large lattice relaxations of either a P or Si atom. In this way, the P atom can yield to its natural tendency to be threefold coordinated. The configuration with a large lattice relaxation of Si is found to be lowest in energy. We suggest that such complexes with large lattice relaxations be further investigated experimentally by means of ion-channeling techniques to confirm our findings.

In conclusion, we have shown on the basis of first-principles calculations of total energy that the configuration with H at an antibonding position of a Si neighbor of P, in which this Si atom relaxes by 0.6 Å, can be identified with the complex that is experimentally observed. In doing so, the discrepancy between results of earlier theoretical studies and experiments is resolved.

This work was supported in part by the U.S. Office of Naval Research under Contract No. N00014-84-C-0396. One of the authors (P.J.H.D.) thanks the IBM Research Division for hospitality during part of the execution time of the work presented here.

[1]S. J. Pearton, J. W. Corbett, and T. S. Shi, Appl. Phys. A 43, 153 (1987).

[2]E. E. Haller, in *Proceedings of the Third International Conference on Shallow Impurities in Semiconductors, Linköping, 1988*, edited by B. Monemar, IOP Conf. Ser. (Institute of Physics and The Physical Society, London, 1989), p. 425.

[3]M. Stavola, S. J. Pearton, J. Lopata, and W. C. Dautremont-Smith, Phys. Rev. B 37, 8313 (1988).

[4]G. DeLeo and W. B. Fowler, in Hydrogen in Semiconductors, edited by J. I. Pankove and N. M. Johnson (unpublished); Bull. Am. Phys. Soc. 34, 834 (1989).

[5]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. B 39, 10809 (1989).

[6]P. J. H. Denteneer, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. 62, 1884 (1989).

[7]N. M. Johnson, C. Herring, and D. J. Chadi, Phys. Rev. Lett. 56, 769 (1986).

[8]K. Bergman, M. Stavola, S. J. Pearton, and J. Lopata, Phys. Rev. B 37, 2770 (1988).

[9]C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B 39, 10791 (1989).

[10]K. J. Chang and D. J. Chadi, Phys. Rev. Lett. 60, 1422 (1988).

[11]S. K. Estreicher, L. Throckmorton, and D. S. Marynick, Phys. Rev. B 39, 13241 (1989).

[12]A. Amore Bonapasta, A. Lapiccirella, N. Tomassini, and M. Capizzi, Phys. Rev. B 39, 12630 (1989).

[13]M. T. Yin and M. L. Cohen, Phys. Rev. B 26, 3259 (1982).

[14]D. J. Chadi *et al.* (private communication).

[15]J. Dabrowski and M. Scheffler, Phys. Rev. Lett. 60, 2183 (1988); D. J. Chadi and K. J. Chang, *ibid.* 60, 2187 (1988).

[16]D. J. Chadi and K. J. Chang, Phys. Rev. B 39, 10063 (1989).

# Accurate interband-Auger-recombination rates in silicon

D. B. Laks and G. F. Neumark
*Division of Metallurgy and Materials Science, Columbia University, New York, New York 10027*

S. T. Pantelides
*IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598*

(Received 26 February 1990)

Band-to-band Auger recombination is the dominant recombination mechanism in silicon at high carrier concentrations. Previous calculations found Auger rates too small to account for experiment. These calculations, however, contained uncontrolled approximations. We calculate accurate Auger recombination rates in both *n*-type and *p*-type silicon, avoiding approximations made in all prior Auger work. Our calculations show that Auger recombination is an order of magnitude stronger than previously thought. Our results for *n*-type Si agree well with experimental lifetimes. In contrast, a phonon-assisted mechanism is indicated for *p*-type Si. This conclusion can be understood based on details of the band structure.

## INTRODUCTION

Electron and hole lifetimes are a key factor in semiconductor physics and technology. The study of these lifetimes is complicated by the variety of carrier-recombination mechanisms that determine them. Recombination mechanisms can be divided into two categories: defect and band to band. Defect recombination can be reduced by avoiding deep-level impurities that act as recombination centers. Band-to-band processes, which are present even in a perfect crystal, provide the ultimate limit to long lifetimes. The two main band-to-band recombination mechanisms are radiative and Auger recombination (AR). In AR an electron recombines with a hole and the energy of recombination is transferred to another electron or hole (Fig. 1).[1,2] In silicon and other indirect-band-gap semiconductors, where radiative recombination is inefficient, band-to-band AR dominates at high carrier concentrations. AR is important for technology as well: It competes with radiative recombination, reducing the efficiency of semiconductor lasers,[3] and it shortens the carrier diffusion lengths, reducing the efficiency of semiconductor solar cells.[4] By converting excess electron-hole pairs to excited carriers, AR plays a role in laser annealing of indirect-band-gap semiconductors.

Many authors have calculated Auger rates in a variety of semiconductors. Calculated Auger rates for silicon[5,6] (without phonon assistance) were an order of magnitude lower than experimental rates.[7,8] Because of this discrepancy, the observed recombination was attributed to a phonon-assisted mechanism, in which the carriers emit or absorb phonons during the Auger transition.[9,10] Calculations for silicon that included phonon-assisted transitions[11] were somewhat more successful in comparison with experiment. Careful examination, however, reveals potentially compromising approximations in *all* of these calculations (both with and without phonons). Examples include dropping a summation over the reciprocal lattice, and using model energy bands and wave functions. The validity of these approximations went untested.

In this paper we describe accurate calculations of the "pure" (no-phonon) Auger recombination rate in both *n*- and *p*-type silicon. We use accurate energy bands and wave functions and perform all summations until they are numerically converged. The Auger rate contains an eight-dimensional surface integral, which we evaluate over a cubic mesh. Furthermore, we have performed detailed convergence studies of all of the parameters that enter the calculation (e.g., the size of the mesh). This provides a quantitative measure of the accuracy of our results. Our theoretical recombination rates agree very well with experiment for heavily doped *n*-type silicon over the entire temperature range; theoretical rates for *p*-type silicon are much smaller than experiment. These

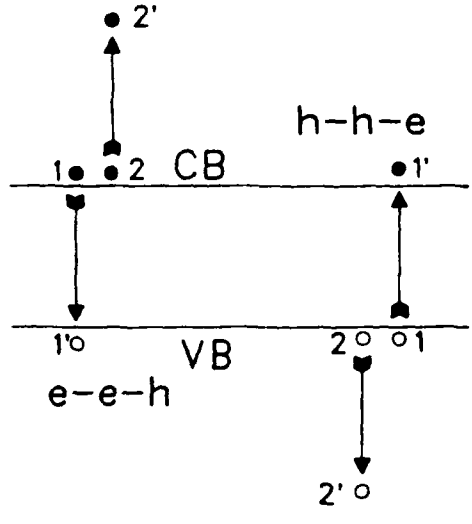



FIG. 1. *e-e-h* and *h-h-e* Auger recombination. CB and VB are conduction- and valence-band edges, respectively.

results suggest that pure AR is the dominant recombination mechanism in $n$-type Si, but that phonon-assisted recombination is in fact dominant in $p$-type Si. The band structure of silicon provides a simple explanation of this difference.

A brief summary of our results has been presented previously.[12,13] Here we give a detailed discussion of both our methods and results.

## AUGER RATES AND THEIR MEASUREMENT

In this section we describe the relationship between Auger rates and lifetimes, and two of the experimental techniques used to measure them. This will provide a better understanding of the relation between the experimental and theoretical results.

Auger recombination involves either two electrons and a hole (electron-electron-hole, or *e-e-h* AR) or two holes and an electron (hole-hole-electron, or *h-h-e* AR). The interaction is mediated by the Coulomb repulsion between the like particles. The different AR mechanisms are characterized by nature of the electron and hole wave functions: in band-to-band AR, all particles are in the bulk bands; in defect AR, one or more particles are bound to crystal defects;[14] and in excitonic AR, the recombination is enhanced by the presence of electron-hole correlations.[15–17] In pure AR, the initial and final electronic states must conserve both energy and momentum. Momentum conservation can be relaxed by phonon emission or absorption (phonon-assisted AR) (Fig. 2). We have investigated the simplest of these processes—pure band-to-band AR.

The *e-e-h* Auger rate, $R$, is proportional to $n^2p$, where $n$ and $p$ are the electron and hole concentrations, respectively. [This is because an *e-e-h* Auger transition requires two electrons and a hole; the occupation probability for electrons (holes) is proportional to $n$ ($p$) using Boltzmann statistics.] Accordingly, the *e-e-h* Auger coefficient, $C_n$, is defined by $R=C_n n^2 p$. In heavily doped $n$-type silicon, where *e-e-h* AR is the dominant recombination mechanism, the hole lifetime is determined by $dp/dt=-R$. Using the definition of $C_n$, the hole lifetime is given by $\tau^{-1}=R/p=C_n n^2$. The $n^{-2}$ dependence of the hole lifetime is the hallmark of AR. (It is assumed here that the carriers are nondegenerate.) The relations for heavily doped $p$-type silicon, where *h-h-e* AR dominates, are completely parallel: the *h-h-e* Auger coefficient is defined by $R=C_p p^2 n$, and $\tau^{-1}=R/n=C_p p^2$. In the following, we will discuss the case of *e-e-h* recombination; *h-h-e* relations can be obtained by and interchanging the pairs $(n,p)$ and $(e,h)$.

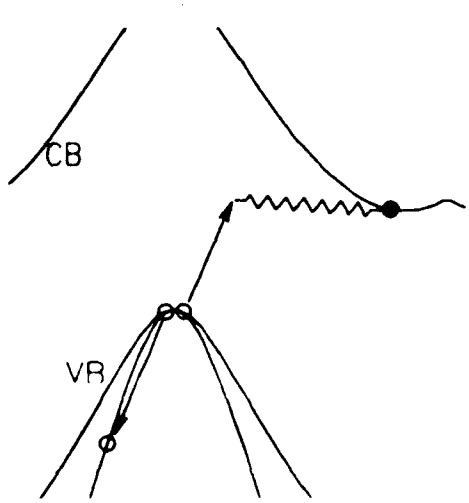



FIG. 2. Phonon-assisted Auger recombination.

The simplest way to determine Auger coefficients[7] is from measured carrier lifetimes [$C_n=(\tau n^2)^{-1}$]. The hole lifetime is measured as a function of electron concentration in a series of heavily doped $n$-type samples. Excess electron-hole pairs are created optically. After the excitation is removed, the hole lifetime is determined by monitoring the weak luminescence produced by the sample. The excited carrier concentration is much smaller that the dopant concentration, so that $n$ remains constant during carrier recombination. In the region where AR is the dominant recombination mechanism, $\tau$ will be proportional to $n^{-2}$, and $C_n$ will be given by the proportionality constant. This technique has two advantages. First, because the electron concentration is the equilibrium value, it can be measured easily, producing more accurate Auger rates. Second, $C_n$ and $C_p$ can be measured independently, by repeating the experiment with $n$-type and $p$-type materials. The one disadvantage of this method is that all measurements are made on heavily doped samples. Chemical impurities distort the band structure of doped material and the Auger rate may not be the same as for pure material.

The second technique measures the Auger rate in intrinsic material.[8] Here we have $n=p$, and the Auger rate becomes $R=(C_n+C_p)n^3$. As in the first method, excess electron-hole pairs are created by an external excitation, and luminescence is used to track the decay of the carrier concentration. With this method, however, the excess carrier concentration is no longer small compared to the equilibrium value. (AR is usually seen only for carrier concentrations much higher than the intrinsic values.) Hence, the total carrier concentration will decrease as a function of time during the luminescence decay. The Auger rate is determined by fitting $n(t)$ to the solution of the nonlinear differential equation

$$R=-dn/dt=(C_n+C_p)n^3+S(n)\,,$$

where $S(n)$ is the rate for all other (significant) recombination processes. This method can only measure the combined Auger rate $C_n+C_p$. The technique is complicated, because of the need to measure excited carrier concentration, and to include all other significant recombination processes in $S(n)$. Fortunately, the Auger rates in silicon as measured by either method are in agreement. This demonstrates that heavy doping does not affect AR in silicon.

## BASIC THEORY

In this section we will present the equations that are used in our theoretical calculations.

The total rate of pure AR is[18]

$$R=2\frac{2\pi}{\hbar}\frac{V^3}{(2\pi)^9}\int_{BZ}\int_{BZ}\int_{BZ}\int_{BZ}|M|^2 f(E_1)f(E_2)[1-f(E_{1'})][1-f(E_{2'})]$$
$$\times\delta(E_1+E_2-E_{1'}-E_{2'})\delta(\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'}-\mathbf{k}_{2'})d\mathbf{k}_1\,d\mathbf{k}_2\,d\mathbf{k}_{1'}\,d\mathbf{k}_{2'}\,, \tag{1}$$

where $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$, and $\mathbf{k}_{2'}$ are the crystal momenta of the electrons and holes, and $E_1$, $E_2$, $E_{1'}$, and $E_{2'}$ are their energies (see Fig. 1). $M$ is the Auger matrix element (see below), $\delta(E)$ and $\delta(\mathbf{k})$ are the energy- and momentum-conserving $\delta$ functions, and $f(E)$ is the probability that an electron is occupying the state with energy $E$. (For $h$-$h$-$e$ AR, $f(E)$ is the occupation probability for a hole.) The $k$ integrals span twelve dimensions; the momentum-conserving $\delta$ function can be used to eliminate the integration over $\mathbf{k}_2$, leaving a nine-dimensional integral. Contracting $\delta(E)$ reduces $R$ to

$$R=2\frac{2\pi}{\hbar}\frac{V^3}{(2\pi)^9}\int_S\frac{|M|^2 f(E_1)f(E_2)[1-f(E_{1'})][1-f(E_{2'})]}{|\nabla_k(E_1+E_2-E_{1'}-E_{2'})|}\,. \tag{2}$$

Here $S$ is an eight-dimensional surface in $\mathbf{k}$ space defined by $E_1+E_2=E_{1'}+E_{2'}$ and $\mathbf{k}_1+\mathbf{k}_2=\mathbf{k}_{1'}+\mathbf{k}_{2'}$. The term in the denominator is the nine-dimensional gradient of $E_1+E_2-E_{1'}-E_{2'}$ with respect to $(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_{1'})$.

The Auger matrix element is given by

$$M=\int\int\phi^*_{\mathbf{k}_1}(\mathbf{r}_1)\phi^*_{\mathbf{k}_2}(\mathbf{r}_2)v(\mathbf{r}_1-\mathbf{r}_2)\phi_{\mathbf{k}_{1'}}(\mathbf{r}_1)\phi_{\mathbf{k}_{2'}}(\mathbf{r}_2)d\mathbf{r}_1\,d\mathbf{r}_2+\text{exchange term}\,,$$
$$v(r)=\sum_{\mathbf{G}}\int_{BZ}\frac{d\mathbf{q}}{(2\pi)^3}v(\mathbf{q}+\mathbf{G})e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\,, \tag{3}$$

where $\phi_{\mathbf{k}}(\mathbf{r})$ is the wave function of the electron (or hole) with wave vector $\mathbf{k}$, $v(\mathbf{r})$ is the screened Coulomb potential,[19] and $\mathbf{G}$ is a reciprocal-lattice vector. The exchange term is found by changing $\phi_{\mathbf{k}_1}(\mathbf{r}_1)$ to $\phi_{\mathbf{k}_2}(\mathbf{r}_1)$ and $\phi_{\mathbf{k}_2}(\mathbf{r}_2)$ to $\phi_{\mathbf{k}_1}(\mathbf{r}_2)$. The $q$ integration runs over the first Brillouin zone (BZ). $v(\mathbf{q}+\mathbf{G})$ is given by the product of the dielectric function and the Coulomb potential in reciprocal space:

$$v(\mathbf{q}+\mathbf{G})=\sum_{\mathbf{G}'}\epsilon^{-1}_{\mathbf{G},\mathbf{G}'}(\mathbf{q})\frac{4\pi e^2}{|\mathbf{q}+\mathbf{G}'|^2}\,, \tag{4}$$

where $\epsilon_{\mathbf{G}',\mathbf{G}'}(\mathbf{q})$ is the dielectric function of the material.[20] Using a diagonal approximation for $\epsilon^{-1}$ reduces $v(\mathbf{r})$ to

$$v(\mathbf{r})=\int\frac{d\mathbf{q}}{(2\pi)^3}\frac{4\pi e^2}{\epsilon(\mathbf{q})q^2}e^{i\mathbf{q}\cdot\mathbf{r}}\,, \tag{5}$$

where the integral now runs over all space.

Because of the periodicity of the crystal lattice, we can put the Auger matrix elements into a form that is easier to evaluate. The wave functions can be expanded in a Fourier series:

$$\phi_{\mathbf{k}}(\mathbf{r})=\frac{1}{\sqrt{V}}\sum_{\mathbf{G}}A(\mathbf{k}+\mathbf{G})e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\,. \tag{6}$$

Using the Fourier expansion of all four wave functions and Eq. (5), the matrix element becomes

$$M=\frac{4\pi e^2}{V}\sum_{\mathbf{G}_1}\sum_{\mathbf{G}_2}\sum_{\mathbf{G}_{1'}}\sum_{\mathbf{G}_{2'}}A^*(\mathbf{k}_1+\mathbf{G}_1)A^*(\mathbf{k}_2+\mathbf{G}_2)B(\mathbf{k}_{1'}+\mathbf{G}_{1'})A(\mathbf{k}_{2'}+\mathbf{G}_{2'})$$
$$\times\frac{1}{|\mathbf{k}_{1'}+\mathbf{G}_{1'}-\mathbf{k}_1-\mathbf{G}_1|^2\epsilon(\mathbf{k}_{1'}+\mathbf{G}_{1'}-\mathbf{k}_1-\mathbf{G}_1)}+\text{exchange term}\,. \tag{7}$$

$\mathbf{G}_1$, $\mathbf{G}_2$, $\mathbf{G}_{1'}$, and $\mathbf{G}_{2'}$ are all reciprocal-lattice vectors. $B$ is used to represent the Fourier components of the wave function with momentum $\mathbf{k}_{1'}$ to indicate that it is in a different band than the other wave functions. All other band indices are suppressed. Applying momentum conservation and substituting $\mathbf{G}=\mathbf{G}_1-\mathbf{G}_{1'}$ gives

$$M=\frac{4\pi e^2}{V}\sum_{\mathbf{G}}\frac{1}{\epsilon(\mathbf{k}_1-\mathbf{k}_{1'}+\mathbf{G})|\mathbf{k}_1-\mathbf{k}_{1'}+\mathbf{G}|^2}\sum_{\mathbf{G}_1}A^*(\mathbf{k}_1+\mathbf{G}_1)B(\mathbf{k}_{1'}+\mathbf{G}_1-\mathbf{G})$$
$$\times\sum_{\mathbf{G}_2}A^*(\mathbf{k}_2+\mathbf{G}_2)A(\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'}+\mathbf{G}_2+\mathbf{G})+\text{exchange term}\,. \tag{8}$$

In this expression the summations over $\mathbf{G}_1$ and $\mathbf{G}_2$ are independent. As a result, the number of terms in each matrix element is reduced from $N^3$ to $2N^2$, where $N$ is the number of reciprocal-lattice vectors used in the calculation. Note that because $\epsilon$ is a function of $\mathbf{k}+\mathbf{G}$, it cannot be factored from the G summation.

## REVIEW OF PRIOR AUGER THEORY

Theoretical study of Auger recombination in semiconductors dates back to the pioneering work of Beattie and Landsberg.[1] In 1958 they investigated the rate of Auger recombination in InSb. To perform the integrals analytically, they made several major approximations in Eqs. (2) and (8). These approximations were designed for a narrow-band-gap direct-transition semiconductor ($E_g$ =0.18 eV in InSb), where all of the k vectors are very near the band edge. During the last two decades these calculations were extended to other semiconductors, including Ge and Si,[5,6] GaAs[21-24] GaSb,[21] and InP.[23,24] Other authors calculated phonon-assisted Auger recombination rates in silicon,[11] and compound semiconductors.[15,25,26]

For silicon, Huldt[5] estimated the no-phonon Auger coefficient as $C_n=0.2\times10^{-31}$ cm$^6$sec$^{-1}$ and predicted that $C_p$ is much smaller than $C_n$, A later calculation by Hill and Landsberg[6] found $C_n=0.12\times10^{-31}$ cm$^6$sec$^{-1}$. On the experimental side, Dziewior and Schmid[7] deduced minority carrier lifetimes from luminescence decay in highly doped $n$-and $p$-type silicon at 77, 300, and 400 K. Their experimental Auger coefficients, $C_n=2.8\times10^{31}$ cm$^6$sec$^{-1}$ and $C_p=0.99\times10^{-31}$ cm$^6$sec$^{-1}$, are much larger than the theoretical values. This led to the suggestion that phonon-assisted Auger recombination was responsible for the experimental rates.[9,10] Support for this thesis came from the temperature dependence of the Auger rate. The measured Auger rates $C(T)$ remain nearly unchanged in the temperature range $T=4$ to 400 K.[7,8] Huldt *et al.*[27] fitted these values to an expected analytical $C(T)$ for both no-phonon and phonon-assisted AR. With the given analytic forms a good fit could be made for the phonon-assisted case but not for the no-phonon case.[10] [The form used for phonon-assisted recombination was $C(T)=C_1\coth(\hbar\omega/2kT)$ where $\omega$ is the frequency of the phonon emitted during recombination. The form used for no-phonon AR was $C(T)=C_2\sqrt{T}\exp(-E_{\mathrm{Th}}/kT)$ where $E_{\mathrm{Th}}$ is the threshold energy and is defined in the Results and Discussion section.] A subsequent calculation of the phonon-assisted Auger rate in silicon by Lochmann and Haug[11] found $C_p$ in good agreement with experiment, but $C_n$ was still four times too small. They concluded that phonon-assisted AR, not pure AR, is the dominant recombination mechanism in highly doped silicon. This conclusion was extended to other indirect-band-gap semiconductors as well.[11]

Careful examination shows that all existing Auger rate calculations—for both pure and phonon-assisted mechanisms and for any of the materials studied—retained many of the approximations of the original Beattie and Landsberg work,[1] even though they were never tested. The underlying assumptions of Beattie and Landsberg's approximations is that all of the k vectors in Eqs. (2) and (8) will be near the band edge, which is true only for a small direct band gap. These approximations may not be valid for wide band gap or indirect-transition semiconductors, where $\mathbf{k}_{2'}$ will be far from the band edge.

Before discussing the particulars of the approximations, it should be noted that there are two categories of theoretical papers under discussion. The difficulty of evaluating the quantities in Eqs. (2) and (8) has been met using two strategies. The first strategy is to introduce as many approximations as are needed to determine the Auger rate. This is the approach used by most authors. The second strategy is to give up on evaluating the full Auger rate in Eq. (2) and concentrate instead on determining the Auger matrix elements alone [Eq. (8)] using fewer approximations.[21-24,28] Approximations used in papers in the first category include the following

In the integration over k [Eq. (2)]:

(1) Of the eight dimensions in the k-space integral, the matrix elements are integrated over, at most, six dimensions. (The reduction occurs because of other approximations, not all mentioned here, that are made.)

(2) Model ($\mathbf{k}\cdot\mathbf{p}$ or parabolic) band structures are used to find the energy conservation surface.

In the matrix element calculation [Eq. (8)]:

(3) The first summation over the reciprocal-lattice vectors (G) is dropped. This approximation is often called neglecting "umklapp terms."

(4) The electron and hole wave functions are taken from $\mathbf{k}\cdot\mathbf{p}$ perturbation theory.

(5) The dielectric function, $\epsilon$, which is in fact a function of k, is either replaced by the static dielectric constant, $\epsilon_0$, or left out entirely.

(6) Some authors do not calculate both the direct and the exchange terms of the matrix elements.

The papers that evaluate only the matrix elements make fewer approximations in Eq. (8). In particular, Brand and Abram[22] appear to be the only authors before this work to include both the sum over G and $k$ dependence of $\epsilon$. (They also use an empirical pseudopotential for the wave functions.) But, as noted above, these calculations are done for only a few matrix elements. As a result, these papers do not provide any estimate of the rate of Auger recombination, which is the quantity of interest. Of the previous evaluations of the total Auger rate, Beattie's work on InSb (Ref. 29) is probably the most accurate. Beattie used the Monte Carlo method to integrate the matrix elements over a full six dimensions in k space. The use of $\mathbf{k}\cdot\mathbf{p}$ band structure and wave functions, a static dielectric constant, and the omission of the sum over G are all appropriate for k vectors very near the band edge. These approximations may be adequate for InSb, where $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$, and $\mathbf{k}_{2'}$ are all near the band edge.

## DESCRIPTION OF PRESENT WORK

We have performed a thorough calculation of the rate of pure *e-e-h* and *h-h-e* Ar in silicon. We do not use any approximations of unknown consequence; in particular all of the approximations described in the previous sec-

tion are avoided. In addition we have verified the accuracy of our numerical approximations through extensive convergence studies.

For the band structure [in Eq. (2)] and the wave functions [in Eq. (8)] we use empirical pseudopotentials.[30] (Nonlocal corrections and spin-orbit splitting were not included; both of these produce very small corrections to the bands that contribute to the total AR rate.) We chose an empirical potential over a first-principles pseudopotential because the latter produces errors in the band gap and the dispersion of the energy bands. In Eq. (2) the integration over k space is performed numerically over an eight-dimensional cubic grid without factoring any of the terms from the integrand. The integration is performed over all regions of k space where the integrand is non-negligible. Auger transitions between the light- and heavy-hole valence bands and the bottom conduction band are included. For the *h-h-e* process the split-off valence band was used as well. Thus our results include all of the 27 different possible *h-h-e* transitions with holes 1, 2, and 2' in the heavy-hole, light-hole, and split-off bands. Fermi-Dirac statistics are used to describe the occupation probabilities for the majority carriers [$f(E_1)$, $f(E_2)$, and $f(E_{2'})$] and Boltzmann statistics are used for the minority carriers [$f(E_{1'})$]. This corresponds to the physical conditions of the experiment[7] to which we compare our results. In the matrix element equation [Eq. (8)], all of the summations over the reciprocal lattice have been retained. The $q$ dependence of the dielectric function is included in the form of Nara and Morita.[31] In our earlier work on AR in silicon[12,13] we used, in addition to the dielectric screening, a Thomas-Fermi screening factor, $\lambda$, for the free-electron screening. Since then, however, Burt[32] has pointed out that Thomas-Fermi screening is static, while the screening in Auger transitions is dynamic.[33] Since the frequencies in Auger transitions (1 eV) are much larger than the plasma frequency of the free carriers (0.1 eV), $\lambda=0$ is probably a better approximation. In this paper we present our results using $\lambda=0$. This results in Auger rates that are about 25% larger than those presented in our previous paper. This difference is of the order of the experimental errors of measurement, and does not affect any of our conclusions.

Because our calculations are a radical departure from previous work in the field, it is interesting to see which of the corrections that we have included are most significant. To this end we have checked some of the approximations that have appeared in previous Auger calculations by performing the same calculation (in silicon) both with and without each approximation. Of those tested, the two worst approximations are the neglect of the sum over the reciprocal lattice ("umklapp" terms), and the use of a static dielectric constant in the matrix elements. Each of these approximations decrease the total Auger transition rate by an order of magnitude. (Neglecting the dielectric screening altogether increases the rate by an order of magnitude.) These approximations may be better in the direct-band-gap materials, which have been the focus of much of the recent theoretical Auger work. Nonetheless, our results serve as a clear warning that such approximations should not be taken for granted, or buried in equations, as has often been done in the past. The use of Fermi-Dirac statistics (as opposed to Boltzmann statistics) increases the recombination rate when the carriers are degenerate. For example, at $T=77$ K, $C_n$ increases by one-third between $n=10^{18}$ cm$^{-3}$ and $10^{20}$ cm$^{-3}$. The effect of the statistics on $C_p$ is even greater, because the effective density of states of the valence bands is lower than that of the conduction band. Because almost all of the previous Auger rate calculations use either $\mathbf{k}\cdot\mathbf{p}$ or parabolic band structures, we attempted to estimate the importance of accurate band structures. To this end we evaluate the Auger rate using parabolic bands (with the same effective masses as our pseudopotentials). Parabolic bands were used only in the evaluation of the energies in the statistical functions $f(E)$, [Eq. (2)]; the full band structure was used for the energy conservation condition. Nevertheless, the total Auger rate was off by 50%. We did not attempt to evaluate the influence of accurate wave functions on our results. Brand *et al.*[21] compared pseudopotential and 15 band $\mathbf{k}\cdot\mathbf{p}$ wave functions in their matrix-element formula and got similar results in either case. (Much of the previous Auger work, however, used only four band $\mathbf{k}\cdot\mathbf{p}$ wave functions, which are less accurate.)

We will now describe the method used to perform the k-space integration [Eq. (2)]. The key to making the eight-dimensional surface integral tractable is to restrict our attention to those regions where the integrand is non-negligible. The occupation probability functions, $f(E)$, guarantee that these regions occur when $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_{1'}$ fall near their respective band edges. We have restricted the integration over k to those regions of k-space satisfying $E_{\rm sum}=(E_1-E_c)+(E_2-E_c)-(E_{1'}-E_v)\le E_{\rm cut}$ for *e-e-h* AR, and $E_{\rm sum}=-(E_1-E_v)-(E_2-E_v)+(E_{1'}-E_c)\le E_{\rm cut}$ for *h-h-e* AR. This choice is based on the fact that the total occupation probability (using Boltzmann statistics) is proportional to $\exp(-E_{\rm sum}/kT)$. The values of $E_{\rm cut}$ used in our calculations range from 200 to 350 meV, depending on the temperature. For silicon, which has an indirect band gap with the conduction-band minimum at $\mathbf{k}=0.85$ (all k vectors are units of $2\pi/A$, where $A$ is the lattice constant), there are several different regions where the integrand is non-negligible. For *h-h-e* recombination there are six regions, in which $\mathbf{k}_1$ and $\mathbf{k}_2$ are holes near the center of the Brillouin Zone and $\mathbf{k}_{1'}$ is in the valley near one of the six conduction-band minima. Because all six regions are equivalent, the calculation for *h-h-e* recombination need be done for only one of the regions, and multiplied by 6. The situation for *e-e-h* recombination is more complicated; here there are three inequivalent types of regions. In the first type, $\mathbf{k}_1$ and $\mathbf{k}_2$ are the same conduction-band valley [for example $\mathbf{k}_1=\mathbf{k}_2=(0.85,0,0)$]. There are six regions of this type. In the second type, $\mathbf{k}_1$ and $\mathbf{k}_2$ are in orthogonal valleys [$\mathbf{k}_1=(0.85,0,0)$ and $\mathbf{k}_2=(0,0.85,0)$]. There are 12 regions of this type. In the third type, of which there are three regions, $\mathbf{k}_1$ and $\mathbf{k}_2$ are in opposite valleys [$\mathbf{k}_1=(0.85,0,0)$ and $\mathbf{k}_2=(-0.85,0,0)$]. The contributions to the Auger rate from the first two types of regions are about the same size. The contribution from the

third type of region is 2 orders of magnitude smaller than that of either of the first two, and is not included in any of the results presented here.

To produce the mesh for the k-space integration, a three-dimensional cubic mesh is generated for each of $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$, and $\mathbf{k}_{2'}$. The side of each cube has length $\Delta k$. The pseudopotential energies for the appropriate bands (three valence bands for holes, one conduction band for electrons) are evaluated at each grid point of the four meshes. The Cartesian product of these four meshes produces a twelve-dimensional mesh of the form $(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_{1'},\mathbf{k}_{2'})$. Applying momentum conservation to fix $\mathbf{k}_{2'}$ reduces the mesh to nine dimensions $(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_{1'},\mathbf{k}_1+\mathbf{k}_2-\mathbf{k}_{1'})$. The grid points form the vertices of a collection of 9 cubes in k space. The set of 9 cubes that intersect the energy-conservation surface $(E_1+E_2=E_{1'}-E_{2'})$ define a mesh over the surface. To determine whether a cube intersects the energy-conservation surface, $E_1+E_2-E_{1'}-E_{2'}$ is evaluated at the 512 vertices of each 9 cube; if this quantity crosses zero between any two vertices, then the cube is placed on a list of cubes that make up the final integration mesh. Next, we construct the wave functions that are needed for the evaluation of the matrix elements. The list of cubes that intersect the energy-conservation surface is used to find the grid points at which the wave functions are required. First, we shift the grids by $\Delta k/2$ in each dimension, so that the grid points now lie at the center of the cubes, rather than at the vertices. The pseudopotential wave function of a point in the $\mathbf{k}_1$ grid is evaluated if that value of $\mathbf{k}_1$ occurs as the first component of the coordinates $(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_{1'},\mathbf{k}_{2'})$ of the center of one of the 9 cubes on the list. The same is done for $\mathbf{k}_2$, $\mathbf{k}_{1'}$, and $\mathbf{k}_{2'}$. Next the matrix elements are calculated at the center of each cube using Eq. (8) and the wave functions. The pseudopotential energies are used to evaluate the Fermi (or Boltzmann) functions. The advantage of using a cubic mesh is clear. The number of points in each dimension of the mesh is proportional to $N=1/\Delta k$. The number of cubes in the grid over the energy conservation surface ( and thus the number of points at which the Auger matrix element need be calculated) is of order $N^8$: the number of points in each of the grids over $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_{1'}$, and $\mathbf{k}_{2'}$, is proportional to $N^3$. Thus the total number of wave functions evaluated is about $4N^3$, which is much smaller than the $N^8$ matrix elements that they determine. In a typical calculation, we used fewer than 1 000 wave functions to calculate 300 000 matrix elements.

We performed careful convergence tests on all of the numerical cutoffs in our calculations. Each parameter was tested separately (with all other parameters held constant), and the tests were performed for all of the calculations (both *e-e-h* and *h-h-e* recombination and at all temperatures). We achieved convergence to within 1% in nearly all cases. (For some low temperature *h-h-e* results only 10% convergence was achieved.) Convergence tests were performed for the following numerical parameters: the number of plane waves used in the energy-band calculation; the number of reciprocal-lattice vectors used in the matrix-element sums; the size of the k-space mesh ($\Delta k$) used in the integration; the energy cutoff ($E_{cut}$).

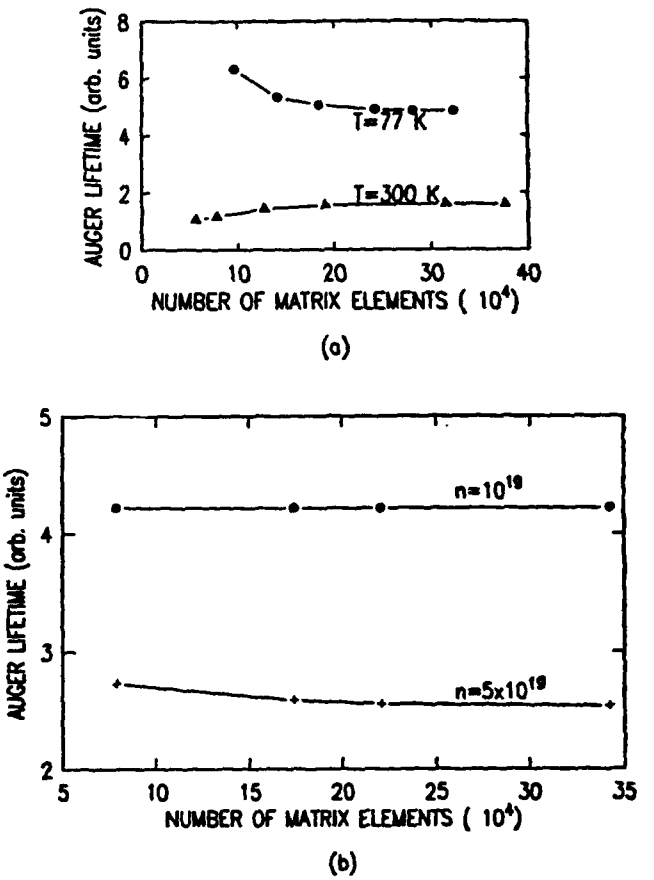



FIG. 3. (a) Convergence of the *e-e-h* Auger lifetime when the mesh size ($\Delta k$) is changed with all other parameters held constant. (b) Convergence of the *e-e-h* Auger lifetime when the energy cutoff ($E_{cut}$) is changed with all other parameters held constant.

Two of the convergence curves (for $\Delta k$ and $E_{cut}$) are shown in Fig. 3.

## RESULTS AND DISCUSSION

We compare our results (Fig. 4) with the experimental lifetimes of Dziewior and Schmid,[7] who measured the minority-carrier lifetimes in heavily doped *n*-type and *p*-type silicon. We chose this experiment for comparison, because it gives the simplest and most direct measurement of the Auger lifetimes. The authors also give separate lifetimes for both *e-e-h* and *h-h-e* AR, over a broad temperature range. (They also measure the Auger rate at $T=4$ K, but we did not extend our calculations to such low temperatures.) A more complex experiment by Svantesson and Nilsson[8] produced very similar results for highly excited intrinsic silicon.

The differences between our results for *e-e-h* and *h-h-e* AR are striking. For *e-e-h* recombination, the theoretical and experimental lifetimes are in very good agreement at both high and low temperatures. For *h-h-e* recombination, in contrast, the theoretical lifetimes are an order of magnitude slower than the experimental values, at best.

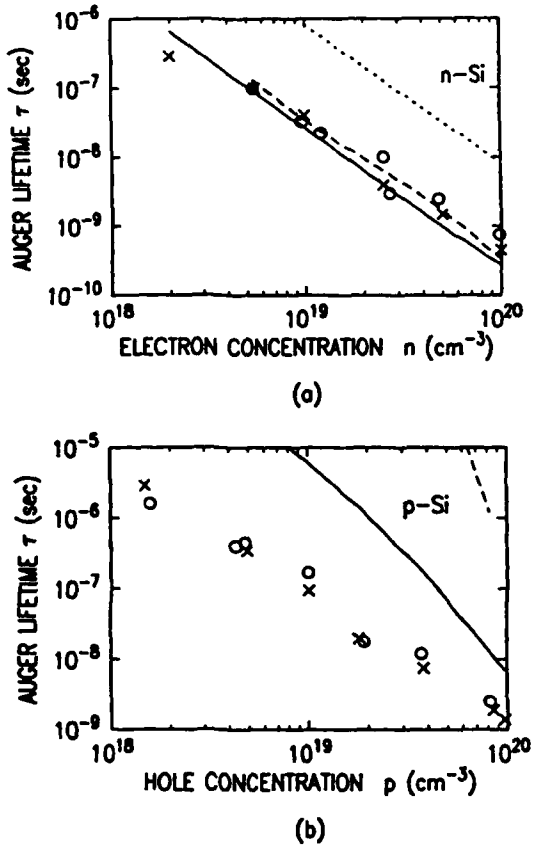



FIG. 4. Experimental and theoretical Auger lifetimes. Experimental results of Dziewior and Schmid (Ref. 7) indicated by circles ($T=77$ K), and $X$'s ($T=300$ K). Theoretical results indicated by dashed ($T=77$ K) and solid lines ($T=300$ K). For comparison, the theoretical results of Hill and Landsberg (Ref. 6) (for $n$-type Si) are indicated by a dotted line ($T=300$ K).

The temperature dependence of the theoretical rates differs as well. While the *e-e-h* rates are nearly temperature independent, the *h-h-e* rates increase rapidly with temperature. The experimental rates show a very weak temperature dependence for both *e-e-h* and *h-h-e* AR. Our pure AR results account for the observed lifetimes in $n$-type silicon but not in $p$-type silicon.

These results are consistent with the assertion of Huldt[5] that pure *e-e-h* Auger transitions are much more probable than pure *h-h-e* transitions. Huldt's theoretical *e-e-h* rate is, however, an order of magnitude smaller than experiment. Later calculations by Hill and Landsberg[6] produced similar results. The authors cautioned, however, that because of the uncertainties in the calculations, their rates were not definitive. These theoretical results led to the conclusion that phonon-assisted Auger recombination dominates in both $n$-type and $p$-type silicon. The weak temperature dependence of the Auger lifetimes was cited as additional evidence of the role of phonon-assisted processes; Haug[9] found that this temperature dependence could be fitted by phonon-assisted mechanism (where the temperature enters through the statistical probability of phonon emission), but not by a pure AR mechanism (where the temperature enters through the occupation probabilities of the electrons; see below). When Lochmann and Haug[11] calculated the phonon-assisted Auger rate for both *e-e-h* and *h-h-e* transitions, they obtained good agreement for *h-h-e* AR, but their results for *e-e-h* recombination were still four times smaller than the experimental values. The point we wish to emphasize here is that the assertion that pure AR is not important in silicon rests entirely on the results of prior theory. But our work shows that, in an accurate theory, both the magnitude and the temperature dependence of the pure *e-e-h* rate agree very well with experiment. We now reopen the question of which recombination mechanism dominates in silicon, pure or phonon-assisted AR? In prior theory the phonon-assisted rates are faster, but both the pure and phonon-assisted calculations used approximations that we have already demonstrated to be invalid. Besides, the theoretical values for the pure and phonon-assisted mechanisms were obtained using different approximations. In particular, the phonon-assisted Auger recombination rate calculation[11] does not include dielectric screening ($\epsilon=1$), while the pure Auger rate calculations[5,6] use static dielectric screening ($\epsilon=12$). Because the Auger rate depends on $\epsilon^{-2}$, using the same $\epsilon$ in both theories would give a pure Auger rate an order of magnitude larger than the phonon-assisted rate. To our knowledge, no one has calculated both pure and phonon-assisted AR rates using consistent approximations. Ideally, we should answer this question by calculating the phonon-assisted AR rates in silicon to the same degree of accuracy as our pure AR calculation, but accurate phonon-assisted rates, which involve an additional integration over phonon momenta, are beyond the limits of present computational capabilities. Instead we present a physical argument to explain why pure AR should dominate in $n$-type silicon, and phonon-assisted AR in $p$-type silicon.

The key to understanding the differences between *e-e-h* and *h-h-e* AR and the relation of pure AR to phonon-assisted AR lies in the concept of recombination thresholds. The thresholds are a consequence of the energy- and momentum-conservation conditions that must be satisfied by the initial and final electronic states. As mentioned above, these conditions determine an eight-dimensional surface in $\mathbf{k}$ space, and Auger transitions can occur only for configurations that lie on this surface. The largest contributions to the Auger rate are those for which $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{K}_{1'}$ are nearest to their respective band edges: otherwise, the statistical function $f(E_1)f(E_2)[1-f(E_{1'})]$ will be vanishingly small. $f(E_1)f(E_2)[(1-f(E_{1'})]$ obtains its maximum value for the configuration that has $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_{1'}$ at the band edges, but the energy-conservation surface need not contain this configuration. The recombination threshold is the configuration on the energy conservation surface that has the maximum value of $f(E_1)f(E_2)[1-f(E_{1'})]$. Using Boltzmann statistics for $f(E)$, we have

$$f(E_1)f(E_2)[1-f(E_{1'})] = \frac{n^2p}{N_c^2N_v}e^{[-(E_1-E_c)-(E_2-E_c)+(E_{1'}-E_v)]/kT}$$

for *e-e-h* AR, and

$$f(E_1)f(E_2)[1-f(E_{1'})] = \frac{np^2}{N_c N_v^2} e^{[(E_1-E_c)+(E_2-E_v)-(E_{1'}-E_c)]/kT}$$

for *h-h-e* AR. (Note that the statistical functions introduce, besides the explicit exponential dependence, a $T^{-9/2}$ dependence through the presence of $N_c$ and $N_v$.) In either case, $f(E_1)f(E_2)[1-f(E_{1'})]$ is proportional to $\exp(-E_{\text{sum}}/kT)$. ($E_{\text{sum}}$ is defined in the previous section.) Thus, when Boltzmann statistics is applicable, the threshold configuration is simply the configuration that has the minimum value of $E_{\text{sum}}$. We call this configuration the Boltzmann threshold configuration. The threshold energy, $E_{\text{Th}}$, is defined as this minimum value of $E_{\text{sum}}[E_{\text{Th}}=E_{\text{sum}}(\mathbf{k}_{\text{Th}})$, where $\mathbf{k}_{\text{Th}}$ is the threshold configuration]. When Fermi-Dirac statistics applies, $f(E_1)f(E_2)[1-f(E_{1'})]$ will not, in general, achieve its maximum value at the minimum value of $E_{\text{sum}}$. Here the threshold configuration will depend on the carrier concentrations, and a threshold energy cannot be defined. Nonetheless, if the carriers are not strongly degenerate, the difference between the two configurations will be small. We will assume that this condition holds, and will use the Boltzmann threshold configuration in our discussions. (This approximation does not enter our calculations, where we use Fermi-Dirac statistics and calculate the Auger rates without direct reference to the threshold.)

Because the total Auger rate is very sensitive to the value of $E_{\text{Th}}$, the relative importance of pure and phonon-assisted AR depends on the difference in thresholds between the two processes. Huldt[5] estimated that for *e-e-h* recombination in silicon $0 \le E_{\text{Th}} \le 52$ meV, but that the *h-h-e* recombination threshold was so large that phonon-assisted recombination was likely to dominate. In the work of both Huldt and of Hill and Landsberg,[6] a threshold of zero was used for calculation of the pure *e-e-h* rate. If the threshold is in fact zero, it is unlikely that phonon-assisted recombination, a *second-order* process, dominates over pure *e-e-h* recombination, a *first-order* process. Although accurate thresholds are crucial to determining the dominant mechanism, there are no other theoretical investigations of the threshold for pure AR in silicon. We have evaluated $E_{\text{Th}}$ for both *e-e-h* and *h-h-e* recombination as the minimum value of $E_{\text{sum}}$ for all of the transitions included in our calculations. We find thresholds of 8 meV for *e-e-h* AR and 76 meV for *h-h-e* AR. Inspection of the band structure of silicon confirms that is easy to find *e-e-h* transitions near the band edge [Fig. 5(a)] but not *h-h-e* transitions [Fig. 5(b)]. (The figures show Auger transitions in which the k vectors are all restricted to a single dimension. In our evaluation of $E_{\text{Th}}$ we have included k vectors in all dimensions.) These thresholds explain why the theoretical *e-e-h* rate is far larger than the *h-h-e* rate, and why the theoretical *h-h-e* rates have a much stronger temperature dependence than the *e-e-h* rates. We can also understand why phonon-assisted AR dominates in *p*-type silicon, and pure AR in *n*-type silicon. For *h-h-e* AR, where the pure Auger mechanism has a high threshold, a phonon-assisted process, even though it is second order, can compete by reducing the threshold. For *e-e-h* AR, where the pure AR threshold is already very low, there is no advantage for phonon-assisted recombination.

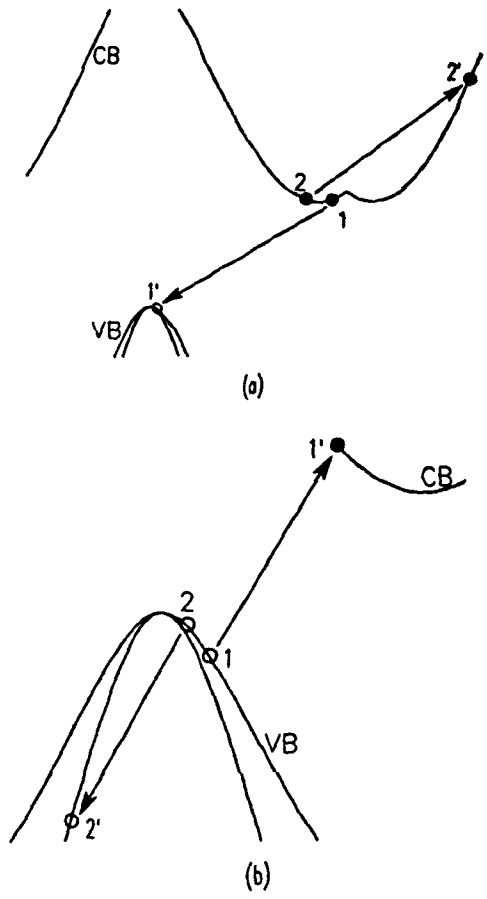



FIG. 5. (a) A possible *e-e-h* transition for silicon. (The top three valence bands and the one lowest conduction band are shown in the ⟨100⟩ direction, using a repeated zone scheme. The apparent cusp in the conduction band is caused by a band crossing at the *X* point.) (b) A possible *h-h-e* Auger transition.

One experiment was conducted to test which type of AR dominates in *n*-type silicon. It was claimed that the results demonstrate that phonon-assisted AR dominates; in fact, the results are inconclusive. Abakumov and Yassievich[34] examined theoretically the effect of uniaxial stress on the AR rate in silicon. If unaxial stress is applied along one of the ⟨100⟩ axes of the crystal, the two conduction-band minima along the stressed axis will be lowered with respect to the four other conduction-band minima. As a result, conduction electrons move from the other valleys to the two valleys that lie along the stressed axis. If enough pressure is applied, these two valleys will eventually contain all of the conduction electrons in the crystal, tripling their population. Abakumov and Yassievich assume that pure Auger transitions can occur only between electrons in the same conduction-band valley. With this assumption they conclude that the application

of uniaxial pressure should triple the pure Auger rate. Subsequently, Grekhov and Delimova[35] measured the experimental pressure dependence of the AR rate in silicon at room temperature. They found the recombination rate to be independent of applied pressure, even when the pressure-induced difference in the conduction-band minima was 40 meV. They concluded that the observed recombination is phonon assisted. (The authors claim that the phonon-assisted transition rate is not affected by uniaxial stress.) The crucial point in this analysis is the assumption that pure AR can only occur between electrons in the same conduction-band valley. Our calculations disprove this assumption; at room temperature the largest contribution to the pure Auger rate comes from the transitions involving electrons in the orthogonal valleys. We have repeated the analysis of Abakumov and Yassievich using the two components of $C_n$ (for electrons in the same and in orthogonal valleys) from our calculation in place of the single Auger coefficient (for electrons in the same valley) used by Abakumov and Yassievich. We find that—even in the case when all of the electrons are transferred to the two preferred valleys—the rate of pure AR is increased by only 25%, which is within the range of experimental uncertainty. In addition, the pressure-induced change in the band structure has an unknown effect on the pure Auger rate, and may well offset the increase caused by the redistribution of the electrons. Also, it has not been demonstrated that the phonon-assisted Auger is independent of applied pressure. We maintain that, because of these arguments, the results of this experiment are inconclusive.

One further point to be explained is the behavior of the Auger coefficients at lower carrier concentrations. The experimental signature of AR is the carrier dependence of the Auger lifetimes, $\tau^{-1}=C_n n^2$ for *e-e-h* AR, or $C_p p^2$ for *h-h-e* AR. For heavily doped *n*-type Si, the Auger coefficient, $C_n$ is constant at $2.7\times10^{-31}$ cm$^6$sec$^{-1}$ when $n$ is above $5\times10^{18}$ cm$^{-3}$, but jumps suddenly to about $2\times10^{-30}$ cm$^6$sec$^{-1}$ when $n$ is below this value.[36–38] The behavior of $C_p$ is similar. (According to Yablonovitch and Gmitter,[38] the carrier dependence of the *e-e-h* Auger rate is best described by an $n^{1.65}$ law, which has been explained either by equilibrium population effects,[39] or by a combination of band-to-band and trap-Auger recombination.[40]) Our calculations for *n*-type Si correctly predict the high-density Auger rates, but not the sudden increase of $C_n$ below $5\times10^{18}$ cm$^{-3}$. The dramatic change in the Auger coefficient is hard to explain. One possible explanation is that degeneracy reduces the Auger rate at high carrier concentrations. But the effects of degeneracy should be minimal at $n=5\times10^{18}$ cm$^{-3}$ and should be more noticeable in the $n=10^{19}$ cm$^{-3}$ to $10^{20}$ cm$^{-3}$ range, where, in fact $C_n$ is constant. Our calculations show that Fermi-Dirac statistics causes only a minor change in $C_n$ even at $n=10^{20}$ cm$^{-3}$, and that it increases the recombination rate. Another possible explanation is that the effects of heavy doping on the band structure diminish the rate of AR. This can be ruled out because experimental measurements of $C_n$ produce the same results in highly excited intrinsic silicon[8] as in heavily doped material.[7]

A more likely explanation of the experimental results is that the Auger rate is not reduced when $n$ is large, but enhanced when $n$ is small. The enhancement of the recombination rate can come from two sources: excitonic AR and AR through electrons in shallow levels. In either case the enhancement would end when the material undergoes a phase transition (caused by increased carrier screening). The abrupt nature of the change suggests that a phase transition is in fact present; the effects of degeneracy, band-structure shifts, and the like, should produce a gradual change that becomes more marked as $n$ increases. Excitonic Auger processes involving actual excitons have been investigated theoretically by Hangleiter,[16,17] and Auger transition enhancement by electron-hole plasma interactions by Takeshima.[15] The increase in the Auger rate predicted by either of these papers is enough to account for the observed change in $C_n$. The excitonic Auger mechanism would be suppressed at higher carrier concentrations, where electron screening would be large enough to nullify the electron-hole attraction. The other possible source of enhancement is AR through donor electrons in shallow levels (or holes in acceptor states for *h-h-e* recombination). These electrons have localized wave functions that are spread over a much larger region of k space than the thermal distribution of conduction-band electrons. Bound electron AR, in combination with pure AR, would dominate below the metal-insulator transition, where there are bound-electron states available. Above the metal-insulator transition these bound states disappear and only pure Auger transitions can occur. Indeed, the change in the Auger coefficient occurs almost exactly at the metal-insulator transition in both *n*-type and *p*-type silicon.

In summary, we have presented an accurate method of calculating pure Auger recombination rates in semiconductors. Applying this method to silicon produces very good agreement with the experimental lifetimes in *n*-type material. We conclude that pure AR dominates in *n*-type silicon and phonon-assisted recombination dominates in *p*-type silicon. Our calculations also show that many of the approximations that have become standard in Auger theory are unreliable. We address the question of the sudden increase in the Auger coefficients below $n=5\times10^{18}$ cm$^{-3}$, and suggest that it is caused by either excitonic or bound-electron AR.

## ACKNOWLEDGMENTS

We are grateful to M. G. Burt (British Telecom, Ipswich, United Kingdom), A. Hangleiter (Universitat Stuttgart, Stuttgart, Germany), and E. Yablonovitch (Bell Communications Research, Red Bank, NJ), for valuable discussions. One of us (D.B.L.) acknowledges support from IBM. This work was supported in part by U.S. Office of Naval Research (ONR) Contract No. N00014-84-C-0396 and a New York State Center for Advanced Technology (CAT) Program Grant to Columbia University.

[1]A. R. Beattie and P. T. Landsberg, Proc. R. Soc. London Ser. A 249, 16 (1958).
[2]P. T. Landsberg, Solid State Electron. 30, 1107 (1987).
[3]M. Takeshima, J. Appl. Phys. 58, 3846 (1985).
[4]M. A. Green, IEEE Trans. Electron Devices ED-31, 671 (1984).
[5]L. Huldt, Phys. Status Solidi A 8, 173 (1971).
[6]D. Hill and P. T. Landsberg, Proc. R. Soc. London Ser. 347, 547 (1976).
[7]J. Dziewior and W. Schmid, Appl. Phys. Lett. 31, 346 (1977).
[8]K. G. Svantesson and N. G. Nilsson, J. Phys. 12, 5111 (1979).
[9]A. Haug, Solid State Commun. 28, 291 (1978).
[10]A. Haug and W. Schmid, Solid State Electron. 25, 665 (1978).
[11]W. Lochmann and A. Haug, Solid State Commun. 35, 553 (1980).
[12]D. B. Laks, G. F. Neumark, A. Hangleiter, and S. T. Pantelides, Phys. Rev. Lett. 61, 1229 (1988).
[13]D. B. Laks, G. F. Neumark, A. Hangleiter, and S. T. Pantelides, in *Shallow Impurities in Semiconductors 1988,* Inst. Phys. Conf. Ser. 95, edited by B. Monemar (Institute of Physics, Bristol, 1989), p. 515.
[14]A. M. Stoneham, *Theory of Defects in Solids* (Oxford University Press, London, 1975), p. 539.
[15]M. Takeshima, Phys. Rev. B 28, 2039 (1983).
[16]A. Hangleiter and R. Hacker, in *Proceedings of the Eighteenth International Conference on the Physics of Semiconductors,* edited by O. Engström (World Scientific, Singapore, 1987), p. 907.
[17]A. Hangleiter, Phys. Rev. B 37, 2594 (1988).
[18]B. K. Ridley, *Quantum Processes in Semiconductors* (Clarendon, Oxford, 1982), p. 268.
[19]J. Callaway, *Quantum Theory of the Solid State* (Academic, New York, 1974), p. 585.
[20]The frequency dependence of $\epsilon$ is not included because it is not important in homopolar semiconductors like silicon.
[21]S. Brand and R. A. Abram, J. Phys. C 17, L201 (1984).
[22]S. Brand and R. A. Abram, J. Phys. C 17, L571 (1984).
[23]M. G. Burt, S. Brand, C. Smith, and R. A. Abram, J. Phys. C 17, 6385 (1984).
[24]M. G. Burt and C. Smith, J. Phys. 17, L47 (1984).
[25]W. Bardyszewski and D. Yevick, J. Appl. Phys. 57, 4820 (1985).
[26]W. Bardyszewski and D. Yevick, J. Appl. Phys. 58, 2713 (1985).
[27]L. Huldt, N. G. Nilsson, and K. G. Svantesson, Appl Phys. Lett. 35, 776 (1979).
[28]P. Scharoch and R. A. Abram, Semicond. Sci. Technol. 3, 973 (1988).
[29]A. R. Beattie, J. Phys. C 18, 6501 (1985).
[30]J. R. Chelikowsky and M. L. Cohen, Phys. Rev. B 10, 5095 (1974).
[31]H. Nara and A. Morita, J. Phys. Soc. Jpn. 21, 1852 (1966).
[32]M. G. Burt (private communication).
[33]M. G. Burt, J. Phys. C 14, 3269 (1981).
[34]V. N. Abakumov and I. N. Yassievich, Fiz. Tekh. Poluprovodn. 11, 1302 (1977) [Sov. Phys.—Semicond. 11, 766 (1977)].
[35]I. V. Grekhov and L. A. Delimova, Fiz. Tekh. Poluprovodn. 14, 897 (1980) [Sov. Phys.—Semicond. 14, 529 (1980)].
[36]Yu Vaitkus and V. Grivitiskas, Fiz. Tekh. Poluprovodn. 15, 1894 (1981) [Sov. Phys.—Semicond. 15, 1102 (1981)].
[37]J. G. Fossum, R. P. Mertens, D. S. Lee, and J. F. Nijs, Solid State Electron. 26, 569 (1983).
[38]E. Yablonovitch and T. Gmitter, Appl. Phys. Lett. 49, 587 (1986).
[39]A. Haug, J. Phys. C 21, L287 (1988).
[40]P. T. Landsberg, Appl. Phys. Lett. 50, 745 (1987).

# Role of Native Defects in Wide-Band-Gap Semiconductors

D. B. Laks,[1,2,a] C. G. Van de Walle,[3] G. F. Neumark,[1] and S. T. Pantelides[2]
[1]*Columbia University, New York, New York 10027*
[2]*IBM T. J. Watson Research Center, Yorktown Heights, New York 10598*
[3]*Philips Laboratories, Briarcliff Manor, New York 10510*
(Received 4 October 1990)

Wide-band-gap semiconductors typically can be doped either *n*-type or *p*-type, but not both. Compensation by native defects has often been invoked as the source of this difficulty. Using first-principles total-energy calculations we show that, for ZnSe and diamond, native-defect concentrations are too low to cause compensation. For nonstoichiometric ZnSe, native defects compensate *both* *n*-type and *p*-type material; thus deviations from stoichiometry cannot explain why ZnSe can be doped only one way. In the absence of a generic mechanism, specific dopants should be examined case by case.

PACS numbers: 71.55.Gs, 61.70.Bv, 72.20.Jv

Wide-band-gap semiconductors (such as ZnSe, ZnS, CdS, ZnTe, BN, or diamond) have ideal band gaps for optical applications using blue or green light, including semiconductor lasers and light-emitting diodes. There is, however, a fundamental problem with these materials: It is difficult, if not impossible, to make diamond and ZnTe *n*-type, and to make the rest *p*-type.[1-3] The simplest explanation[1,4-7] suggested for this phenomenon is that native defects compensate, say, acceptors in ZnSe. Because of the wide band gap, some of the energy needed to form a native donor defect can be recouped when electrons from defect levels in the gap recombine with holes at the Fermi level in *p*-type material. The spontaneous formation of native defects would thus prevent the Fermi level from moving below a fixed value that is determined by the formation energies and electronic levels of the native defects, independent of the dopant and of how the material was prepared. This picture has some very appealing features. It would explain why doping problems occur in all wide-band-gap materials, and are less severe in medium-gap materials such as CdTe. It would also explain why the difficulty in producing *p*-type (or *n*-type) material is universal, appearing for all growth and doping techniques, and for all dopants. That these materials can be doped *n*-type and not *p*-type, or vice versa, can be explained if the native defects with the lowest formation energy are donors in some materials and acceptors in others. For example, Jansen and Sankey[7] have suggested that the native-defect mechanism can account for the difference between ZnSe, which can be made *n*-type, and ZnTe, which can be made *p*-type, even though the two materials are strikingly similar in other ways. There is, however, no direct evidence to either confirm or deny the role of native defects in wide-band-gap semiconductors.

In this Letter we report on theoretical determinations of native-defect concentrations in ZnSe. The underlying calculations attain for the first time the level of accuracy that has so far been practical only for materials like Si and GaAs. We find that (1) the native-defect concentrations are too low to be a significant source of compensation in stoichiometric ZnSe; (2) undetectably small deviations from stoichiometry can produce large concentrations of native defects. We find that the defects formed depend on whether the sample is *n*-type or *p*-type, but always compensate. Hence deviations from stoichiometry cannot explain why ZnSe can be doped *n*-type but not *p*-type, because they are as likely to compensate *n*-type material as *p*-type. We have further determined native-defect concentrations in diamond, and find again that compensation by native defects is insignificant. In the absence of a generic mechanism, potential dopants need to be examined case by case.

Our determination of defect concentrations is based on calculating the total energy of each defect using density-functional theory (DFT) and the local-density approximation (LDA), norm-conserving pseudopotentials and supercells.[8] These techniques have been very successful in elucidating defect properties in Si (Refs. 9 and 10) and GaAs.[11] Applying the same tools to ZnSe, however, presents a problem. Zinc contains a fully occupied band of 3*d* electrons, which are tightly bound to the nucleus, and yet fall within the valence bands of ZnSe. Standard defect calculations are performed with a plane-wave basis set, which would require far too many plane waves to represent the *d* states. If the *d* states are treated as core states of the pseudopotential, it is not necessary to represent them in the basis set; unfortunately, this procedure is unacceptable because it does not correctly represent the properties of ZnSe.[12] To treat the *d* states properly, and still be able to perform supercell calculations, we use an all-new mixed-basis-set program, similar in spirit to that of Louie, Ho, and Cohen.[13,14] The usual plane-wave (PW) basis set is supplemented by a set of pseudoatomic tight-binding (TB) functions situated on each zinc atom. Calculations of the bulk properties of ZnSe (and other semiconductors) show that this scheme describes material properties very well: The predicted lattice constant and bulk modulus agree with experiment to within 1% and 10%, respectively.

For defect calculations, the convergence of the results with respect to basis sets and supercell size was checked to ensure overall accuracy of better than 0.5 eV.[15] An additional uncertainty is introduced by the local-density approximation, which is well known to underestimate band gaps. In $p$-type material, this uncertainty is negligible because all levels in the energy gap are empty. The uncertainty in $n$-type material is larger and can be estimated from the error in the band gap itself. In our discussion of the results for $n$-type materials, we assumed the worst-case values.

Calculations were performed for all native point defects: $Zn_i$, $Se_i$ (interstitials), $V_{Zn}$, $V_{Se}$ (vacancies), $Zn_{Se}$, and $Se_{Zn}$ (antisites) in a variety of charge states; 29 different cases were examined, and detailed results will be published elsewhere. Calculations for these native defects have been reported earlier by Jansen and Sankey,[7] using more approximate techniques.[16] We will refer to their results where appropriate.

For a compound semiconductor like ZnSe, the formation energies and hence the concentrations of native defects are a function of the stoichiometry of the material. The stoichiometry itself is related to the chemical potentials of the constituents of the compound, in our case Zn ($\mu_{Zn}$) and Se ($\mu_{Se}$) atoms. The two chemical potentials are constrained by the condition that (in equilibrium) their sum must equal the total energy of a two-atom unit of perfect ZnSe ($\mu_{ZnSe}=\mu_{Zn}+\mu_{Se}$). (We use the total energy of a perfect ZnSe cell at $T=0$ K for $\mu_{ZnSe}$.) Given the Zn and Se chemical potentials, the formation energy of each native defect is well defined and can be derived from a supercell calculation as follows. The total energy of a supercell for the $i$th defect containing $N$ Zn atoms and $M$ Se atoms ($E_i$) is calculated. The defect formation energy is then

$$E_i - N\mu_{Zn} - M\mu_{Se} = E_i - (N-M)\mu_1 - (N+M)\mu_2 = \varepsilon_i - n_i\mu_1 ,$$

where $\mu_1=(\mu_{Zn}-\mu_{Se})/2$, $\mu_2=(\mu_{Zn}+\mu_{Se})/2$ (a constant), $n_i=N-M$, and $\varepsilon_i=E_i-(N+M)\mu_2$. $n_i$ is the number of extra Zn atoms that must be added to form the defect (+1 for $V_{Se}$, −2 for $Se_{Zn}$, etc.), independent of the size of the supercell. Using this prescription, all of the defect formation energies, and hence their concentrations ($C_i$), are unique functions of $\mu_1$. The concentrations, in turn, determine the stoichiometry. In practice, however, it is more convenient to fix the stoichiometry first, and then determine $C_i$. To do this we write $C_i$ in terms of the total energies and entropies ($S$) of formation as

$$C_i = e^{S/k_B} e^{-(\varepsilon_i - n_i\mu_1)/k_BT} = e^{S/k_B - \varepsilon_i/k_BT} y^{n_i} = a_i y^{n_i} ,$$

where $y=\exp(\mu_1/k_BT)$. The stoichiometry parameter is

$$X = -\tfrac{1}{2}\sum_i n_i C_i = -\tfrac{1}{2}\sum_i n_i a_i y^{n_i}$$

($X=0$ for perfect stoichiometry, and $X>0$ for Se rich). To find defect concentrations as a function of stoichiometry, one simply chooses a value of $X$ (and the temperature) and solves for $y$. (The problem is essentially finding a root of a polynomial, which can be done quickly and easily using standard algorithms.)

Defect concentrations are a function of formation energies and entropies. We have checked that our results are insensitive to the value of the entropies in the range $S=(0\text{–}10)k_B$. By comparison, a recent accurate calculation[17] of the formation entropy of the Si self-interstitial found a formation entropy of $(5\text{–}6)k_B$ for the ground state. The Si self-interstitial represents an extreme case in that the ground-state configuration has low symmetry, which accounts for half of the formation entropy. It is therefore highly unlikely that the entropies for native defects in ZnSe or diamond could be larger than $10k_B$. Similarly, the defect formation energies are high enough that, even with a generous estimate of the atomic relaxation energies, the concentrations remain very low. Relaxations are calculated explicitly for the dominant defects in $p$-type ZnSe and found to be less than 0.6 eV. The concentrations of other defects remain small even if relaxations up to 2 eV are assumed.

Figure 1 shows the concentrations of minority carriers produced by native defects for $p$-type stoichiometric ZnSe. The results shown are for material with $10^{18}$ cm$^{-3}$ dopants. The dominant native defects are $Zn_i^{2+}$, $V_{Zn}^{0}$, and $Se_{Zn}^{2+}$. At molecular-beam-epitaxy- (MBE-) growth temperatures ($T=600$ K) the concentration of minority carriers produced is less than $10^{12}$ cm$^{-3}$. For material grown at higher temperatures, excess native defects will recombine during cooling, unless the sample is rapidly quenched.[18]

We have also calculated native-defect concentrations

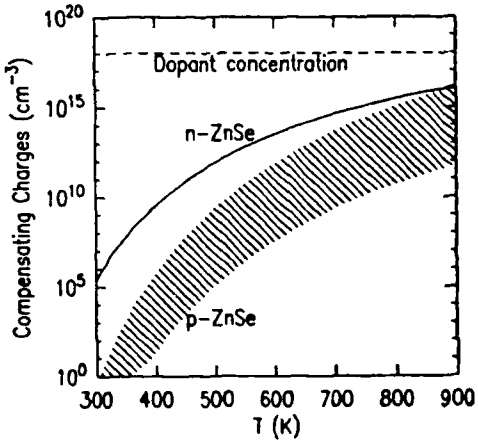



FIG. 1. Concentration of minority carriers produced by all native defects in stoichiometric $p$-type and $n$-type ZnSe. (The range of values shown for $p$-type ZnSe is bounded by assuming an entropy of $10k_B$ per defect for an upper bound and $0k_B$ for a lower bound.)

in $n$-type ZnSe (Fig. 1). The dominant defects are $V_{Zn}^{2-}$ and $Zn_{Se}^{-}$. Well-conducting $n$-type ZnSe can be easily produced; thus it is an experimental fact that native defects do not compensate $n$-type doping. As shown in Fig. 1, native-defect concentrations in $n$-type ZnSe are comparable to, if not greater than, defect concentrations in $p$-type.[19] This is additional proof that native-defect compensation cannot explain why $p$-type ZnSe is harder to grow than $n$-type.

To further support our conclusions, we have derived native-defect concentrations for diamond from the first-principles defect energies of Bernholc *et al.*[20] The doping level is again $10^{18}$ cm$^{-3}$. At a chemical-vapor-deposition-growth temperature of 1100 K, the number of holes produced in $n$-type diamond by native defects is at most $2\times10^{13}$ cm$^{-3}$ (Fig. 2). Clearly, the concentrations of native defects in both stoichiometric ZnSe and diamond are far too low to produce significant compensation.

Jansen and Sankey have estimated native-defect concentrations in ZnSe and ZnTe.[7] They concluded that native-defect compensation could explain why ZnSe prefers to be $n$-type and ZnTe prefers to be $p$-type. However, their results were reported for a very high temperature ($T=1658$ K), and thus do not apply to the question of compensation for material that is grown at 600 K and never thermally annealed at higher temperatures. At lower temperatures, their results also show a low concentration of native defects. Furthermore, for ZnSe their numbers indicate that compensating native-defect concentrations are *lower* in $p$-type material than in $n$-type.

Our conclusion that the concentrations of native defects in stoichiometric ZnSe are very low does not mean that native-defect compensation in ZnSe never occurs. If the sample is grown with even a slight deviation from perfect stoichiometry, the concentration of native defects will necessarily be very large, even at $T=0$ K.[21] Because the density of atomic sites in ZnSe is $4\times10^{22}$ cm$^{-3}$ a deviation from stoichiometry as small as $10^{-4}$ implies a defect concentration of about $10^{18}$ cm$^{-3}$. We find that the native defects that accommodate deviations from stoichiometry are always those that compensate the majority carriers. For $p$-type ZnSe, the dominant defect is $Zn_i$ in Zn-rich material, and $Se_{Zn}$ in Se-rich material; we find that both are double donors. For $n$-type ZnSe, the dominant (acceptor) defects are $Zn_{Se}$ and $V_{Zn}$ for Zn and Se rich materials, respectively. Similar results were found by Jansen and Sankey.[7] This defect structure is much richer than that used in many previous analyses of native defects in II-VI semiconductors.[5] The difficulty in producing $p$-type ZnSe cannot be explained by deviations from stoichiometry because any deviation that compensates $p$-type doping would compensate $n$-type doping equally well.

The deviations from stoichiometry that we are discussing are too small to measure experimentally, which precludes a *direct* confirmation of our predictions. There is, however, indirect evidence to verify one of our predictions, namely, that the zinc vacancy is the dominant native defect in $n$-type Se-rich ZnSe. As-grown bulk ZnSe samples are highly compensated, and must be annealed in a Zn-rich atmosphere to be made well conducting. One known cause of this compensation is large numbers of "self-activated" (acceptor) centers, which are donor-$V_{Zn}$ pairs.[22] This shows that zinc vacancies are a prominent defect in as-grown $n$-type ZnSe. Furthermore, analysis of the Zn-Se phase diagram suggests that ZnSe grown under equilibrium conditions from a melt is Se rich. Thus, our results for Se-rich $n$-type ZnSe provide a natural explanation of the occurrence of self-activated centers in ZnSe.

Having settled the native-defect compensation issue quantitatively, we now reexamine the notion that native-defect compensation increases with the width of the band gap. Let us restate the standard argument for this trend: For $p$-type material, imagine a prototypal compensating native donor defect that, when neutral, introduces one electron into a state in the gap; the formation energy for this defect, $E^0$, is assumed not to depend on the width of the band gap. The energy gained by transferring the electron from the level in the gap ($E_L$) to the Fermi level ($E_F$) should, in contrast, increase with the width of the gap; thus the net energy needed to form compensating defects, $E^0-(E_L-E_F)$, should decrease as the band gap increases. The flaw in this argument is that it assumes that the level in the gap ($E_L$) and $E^0$ are independent of one another. Actually, the level in the gap is *defined* by $E_L=E^0-E^+$, where $E^++E_F$ is the (Fermi-level-dependent) energy of formation of the posi-

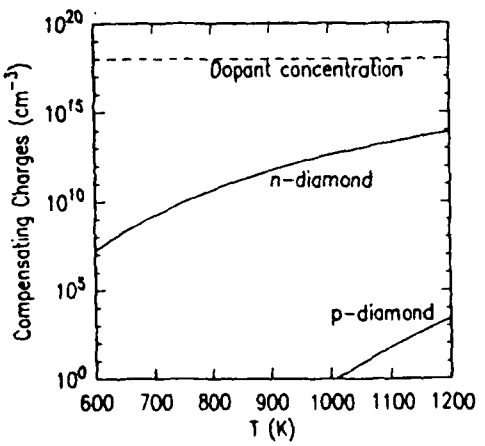



FIG. 2. Concentration of minority carriers produced by all native defects in $n$-type diamond. (The results for $n$-type diamond use the worst-case correction for the LDA band-gap error: The conduction-band edge is taken at the experimental value without shifting up any of the defect levels. True defect concentrations are probably much lower.)

tive charge-state defect. Using this definition, we find that the net energy required to create a compensating defect is $E^0-(E_L-E_F)=E^++E_F$, independent of the energy of formation of the neutral defect. We see that native-defect compensation will increase with the width of the band gap if and only if $E^++E_F$ *decreases* with increasing band gap. The existence of such a trend has not been convincingly established.

Having eliminated native defects as a generic source of compensation in wide-band-gap materials, it is fruitful to identify problems associated with specific dopants. We are studying the technologically important case of $Li_{Zn}$, a promising acceptor in ZnSe.[23] In a separate publication, we will report on the properties of Li impurities in ZnSe, including possible defect reactions.

In conclusion, we have shown that native defects alone cannot be responsible for difficulties in doping the wide-band-gap semiconductors ZnSe and diamond. Native-defect concentrations in MBE-grown stoichiometric ZnSe are too low to compensate. Deviations from stoichiometry in ZnSe *do* produce large numbers of native defects which, however, compensate *n*-type as well as *p*-type material.

We are very grateful to P. Blöchl for many fruitful suggestions, and for making his unpublished work available to us. We are indebted to D. Vanderbilt for his iterative diagonalization program. We acknowledge helpful conversations with R. Bhargava, J. M. DePuydt, T. Marshall, J. Tersoff, and G. D. Watkins. D.B.L. acknowledges support from an IBM Graduate Fellowship. This work was supported in part by NSF Grant No. ECS-89-21159 and ONR Contract No. N00014-84-0396.

---

[a]Present address: Solar Energy Research Institute, Golden, CO 80401.

[1]Y. S. Park and B. K. Shin, in *Electroluminescence*, edited by J. I. Pankove, Topics in Applied Physics Vol. 17 (Springer, Berlin, 1977), p. 133.

[2]R. Bhargava, J. Cryst. Growth **59**, 15 (1982).

[3]G. F. Neumark, Phys. Rev. Lett. **62**, 1800 (1989).

[4]G. Mandel, Phys. Rev. **134**, A1073 (1964).

[5]A. K. Ray and F. A. Kröger, J. Electrochem. Soc. **125**, 1348 (1978).

[6]Y. Marfaing, Prog. Crystal Growth Charact. **4**, 317 (1981).

[7]R. W. Jansen and O. F. Sankey, Phys. Rev. B **39**, 3192 (1989).

[8]C. G. Van de Walle, P. J. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B **39**, 10791 (1989), and references therein.

[9]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. **52**, 1814 (1984).

[10]Y. Bar-Yam and J. D. Joannopoulos, J. Electron. Mater. **14A**, 261 (1985).

[11]G. A. Baraff and M. Schlüter, Phys. Rev. Lett. **55**, 1327 (1985).

[12]S. H. Wei and A. Zunger, Phys. Rev. B **37**, 8958 (1988).

[13]S. Louie, M. Ho, and M. Cohen, Phys. Rev. B **19**, 1774 (1979).

[14]The eigenvalue problem was solved using the iterative diagonalization scheme of R. Natarajan and D. Vanderbilt, J. Comput. Phys. **82**, 218 (1989).

[15]For supercells, 16-, 32-, and 64-atom cells were checked. In most cases, 32-atom cells were adequate. The plane-wave cutoff was typically 9 Ry.

[16]The *d* states were treated as part of the core.

[17]P. Blöchl (private communication).

[18]Defects in ZnSe move readily even at $T=400$ K, so that kinetic barriers do not prevent the attainment of thermal equilibrium. [G. Watkins, in Proceedings of the International Conference on Science and Technology of Defect Control in Semiconductors, Yokohama, 1989 (to be published)].

[19]Occupied defect levels in the gap were shifted upward by the LDA band-gap error. The true defect concentrations in *n*-type ZnSe may actually be higher than those shown in Fig. 1.

[20]J. Bernholc, A. Antonelli, T. M. Sole, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. **61**, 2689 (1988).

[21]We refer only to deviations from stoichiometry accommodated by native-point defects. Deviations from stoichiometry due to higher-dimensional defects, precipitates, and substitutional dopants are not included.

[22]R. K. Watts, *Point Defects in Crystals* (Wiley, New York, 1977), p. 252.

[23]J. Depuydt, M. Haase, H. Cheng, and J. Potts, Appl. Phys. Lett. **55**, 1103 (1989).

PHYSICAL REVIEW B VOLUME 45, NUMBER 19 15 MAY 1992-I

# Native defects and self-compensation in ZnSe

D. B. Laks*
*Division of Metallurgy and Materials Science, Columbia University, New York, New York 10027 and IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598*

C. G. Van de Walle†
*Philips Laboratories, Briarcliff Manor, New York 10510*

G. F. Neumark
*Division of Metallurgy and Materials Science, Columbia University, New York, New York 10027*

P. E. Blöchl‡ and S. T. Pantelides
*IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 26 August 1991; revised manuscript received 27 December 1991)

Wide-band-gap semiconductors typically can be doped either *n* type or *p* type, but not both. Compensation by native point defects has often been invoked as the source of this difficulty. We examine the wide-band-gap semiconductor ZnSe with first-principles total-energy calculations, using a mixed-basis program for an accurate description of the material. Formation energies are calculated for all native point defects in all relevant charge states; the effects of relaxation energies and vibrational entropies are investigated. The results conclusively show that native-point-defect concentrations are too low to cause compensation in stoichiometric ZnSe. We further find that, for nonstoichiometric ZnSe, native point defects compensate *both* *n*-type and *p*-type material; thus deviations from stoichiometry cannot explain why ZnSe can be doped only one way.

## I. INTRODUCTION

Wide band-gap semiconductors, such as ZnSe, ZnTe, ZnS, and diamond, have potential technological applications, especially for optical devices involving green or blue light.[1-3] Despite decades of research, many problems remain, mostly related to doping difficulties; some wide-band-gap materials can easily be made *n* type but not *p* type; others can be made *p* type, but not *n* type.[4] The cause of this difficulty remains a puzzle. At least five different explanations have been suggested,[5-9] but there is no firm evidence for any of them. One of the oldest and most popular explanations is that the doping of wide-band-gap semiconductors is compensated by native point defects.[2-12] According to this mechanism, the wide band gap could promote the formation of compensating native point defects because the formation energy of the defect is offset by the energy gained when electrons are transferred between the defect's electronic state in the gap and the Fermi level. For example, *p*-type doping may be compensated by defects that introduce electrons in levels near the conduction band. When the electrons drop from the level in the gap to the Fermi level (which is near the valence-band edge), the net formation energy for the compensating defect would be reduced by nearly the width of the band gap. This mechanism would be universal: it is independent of the dopant and the growth method used. The native point defect properties would directly determine the behavior of the material. A wide-band-gap semiconductor would tend to be *n* type if the dominant native point defects introduce full states near the conduction-band edge. It would be *p* type if the dominant defects introduce empty electronic states near the valence bands.

Our goal is to examine the native-point-defect mechanism using first-principles theoretical techniques based on density-functional theory and *ab initio* pseudopotentials. We will study native point defects in ZnSe, which is the wide-band-gap semiconductor that has received the most attention in the past decade. ZnSe can be grown *n* type, but only limited progress has been made growing *p*-type material.[13,14] Theoretical tools have been very useful in elucidating the properties of common semiconductors such as Si and GaAs.[15-18] Much less has been done for ZnSe, or any of the other II-VI semiconductors. For these materials, the plane-wave pseudopotential method, the standard for semiconductor defect calculations, does not work well. This is because the *d* electrons of the group-II metals are too tightly bound to be represented as valence electrons with a plane-wave basis set. In all previous pseudopotential calculations for ZnSe,[6,12,19] the zinc 3*d* electrons were treated as core states. Using this method, Jansen and Sankey[12] suggested that native-point-defect compensation is the cause of doping difficulties in ZnSe and ZnTe and on the same basis explained why ZnTe (which prefers to be *p* type) is different from ZnSe (which prefers to be *n* type). Unfortunately, the "*d*-in-the-core" pseudopotential approach is inaccurate: it cannot predict the experimental bulk properties of ZnSe,[20] and is therefore highly suspect for defect calculations.

We solve the *d*-electron problem by using a mixed-basis scheme, which adds to the plane-wave basis a set of tight-binding functions that can represent the *d* electrons as valence states. The mixed-basis scheme is implemented in a program that is efficient enough for large-scale defect calculations. Our defect calculations are the first for a II-VI semiconductor that include a proper treatment of the *d* electrons, and reach the level of accuracy previously attained for Si and GaAs. We calculate the formation energies of all native point defects in ZnSe. Using these formation energies we derive upper bounds for the defect concentrations. The results show clearly that native defect compensation in *stoichiometric* ZnSe is insignificant. Additional support for this conclusion is provided by calculations of native-point-defect concentrations in another wide-band-gap semiconductor, namely diamond. Here we derive the concentrations from published native-point-defect energies.[21] In *nonstoichiometric* ZnSe, native-point-defect compensations *will* occur, but will compensate *n*-type as well as *p*-type material. Deviations from stoichiometry, therefore, do not explain why it is easy to make *n*-type but not *p*-type ZnSe. Our results clearly indicate that native defects are not responsible for self-compensation in ZnSe and thus impose no *intrinsic* limitation on the ability to obtain both *n*-type and *p*-type conduction.

This paper is organized as follows: In Sec. II we describe the details of our mixed basis scheme. By relying on fast-Fourier-transform (FFT) routines and the convolution theorem,[22] total-energy calculations for a defect (which require a cell with a large volume) can be performed efficiently. A description of our test calculations follows; these establish the credibility of our theoretical methods. In Sec. III we describe our own total-energy calculations for the native point defects, and a discussion of the structure of each defect. Because ZnSe is a compound semiconductor, the formation energy of a single defect is not well defined. In Sec. IV we show how chemical potentials can be related to stoichiometry, yielding an unambiguous definition of formation energies in terms of the calculated total energies. Defect concentrations can then be obtained as a function of temperature, stoichiometry, and the Fermi level of the crystal. We then present our calculated native-point-defect concentrations (Sec. V), which show clearly that the native point defects do not affect the doping of ZnSe. We also show that the same is true of diamond. Having shown quantitatively that native point defects are not responsible for compensation, we present a qualitative analysis of whether native-point-defect concentrations increase with the width of the band gap

## II. THE MIXED-BASIS METHOD

In this section we describe our implementation of the mixed basis method. Our formalism is based on density functional theory in the local density approximation (LDA)[23] using *ab initio* pseudopotentials.[24] In traditional pseudopotential calculations with a plane wave basis set, the bulk of the computation consists of solving the eigenvalue problem for the Hamiltonian matrix. The mixed-basis scheme produces a much smaller Hamiltonian matrix by replacing many high-frequency plane waves with a few tight-binding functions. The price paid for the smaller Hamiltonian is that introducing tight-binding functions complicates the matrix elements and the integration of the charge density. We handle the added complications of the tight-binding functions by expanding them over the reciprocal lattice. The tight-binding functions $\phi$ are written as

$$\phi_{\mathbf{k}i}(\mathbf{r}) = \sum_{\mathbf{G}} \phi_i(\mathbf{k}+\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \tag{1}$$

where G is a reciprocal-lattice vector. Functions of this form automatically have the correct translational symmetry (Bloch's theorem). There are two principal advantages to the reciprocal-space expansion. One is that the choice of tight-binding functions is not restricted to any particular analytic form, such as Gaussians or Slater orbitals. This will allow us to use so-called pseudoatomic wave functions as basis functions, as discussed below. The second is that the reciprocal-space expansion is particularly well suited for treating the tight-binding functions and the plane waves in a unified fashion. (Note that the exponentials in the expansion for $\phi$ are simple plane waves.) This simplifies the calculation of the matrix elements between tight-binding functions and plane waves, as well as the calculation of the charge density. The scheme is similar to that used by Louie, Ho, and Cohen.[25] The programs are all new, and both the programs and the algorithms were carefully optimized to make large-scale defect calculations possible. We now discuss the details of our methods, using the work of Louie, Ho, and Cohen[25] as a starting point. A detailed description of the evaluation of the various matrix elements is presented in the Appendix. Further details on the method are given elsewhere.[26]

### A. Basis set

Although a mixed basis allows great freedom in the choice of the basis set, our approach has been to keep our ZnSe calculations as similar as possible to the plane-wave calculations for Si and GaAs. Consequently, we use tight-binding functions to represent only the rapidly changing part of the zinc *d* orbitals. All other contributions to the wave functions are represented by plane waves. In this way we recover most of the advantages of a pure plane-wave calculation, and reduce the effort needed to choose and optimize the basis set. We have performed ZnSe calculations using two different forms for the zinc 3*d* orbitals: Gaussians and pseudoatomic wave functions. The pseudoatomic functions are the 3*d* pseudo-wave functions for the isolated zinc atom, as calculated by the program that generates the atomic pseudopotentials. We multiply the pseudoatomic basis functions by a smooth radial cutoff function that goes to zero for large *r* to remove the long range tail (Fig. 1). (Basis functions with long range tails become numerically unstable as the overlap between basis functions on different sites causes the basis set to become linearly dependent.) Using Gaussians requires at least two basis functions for each

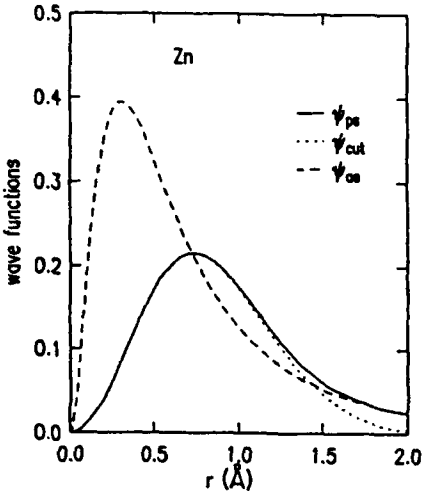



FIG. 1. Zn 3*d* wave function and pseudoatomic basis functions. The all-electron wave function is given by the dashed line. The pseudoatomic wave function is given by the solid line. The basis function (pseudoatomic wave function multiplied by a cutoff function) is given by the dotted line.

zinc *d* orbital (ten functions per zinc atom to represent five *d* orbitals). With the pseudoatomic basis functions, one orbital per state suffices (five functions per zinc atom). The total energy of a perfect ZnSe unit cell calculated with one set of pseudoatomic orbitals is lower than that calculated with two sets of Gaussians, even when the decay constants of the Gaussians are optimized. We conclude that the pseudoatomic basis functions provide a good description of the *d* states in the solid.

The reciprocal-space grid represents a parallelepiped placed about the origin in G space. Because of the shape of the parallelepiped, rotations that are symmetry operations for the crystal will map some G vectors that are inside the FFT grid into G vectors that are outside the grid, and vice versa. As a result, the functions represented on the FFT grid will no longer have the correct rotational symmetry. To correct this problem we set the Fourier coefficients to zero for all G vectors that lie outside the largest sphere that fits inside the FFT grid. This substantially reduces the number of nonzero G vectors. For example, a simple cubic lattice has a FFT grid that is a cube and the ratio of the volume of the inscribed sphere to the volume of the grid is $\pi/6=0.5236$. For other lattices, the ratio is even smaller. To avoid unnecessary storage, the program maps the full FFT G-vector grid onto a smaller G-vector grid containing only the points in the sphere. This grid is mapped back onto the FFT grid whenever an FFT is needed. No such reduction is possible for the real-space grid; the full FFT grid must be used in this case. Because of the asymmetry between real and reciprocal space, it is advantageous to store the functions and perform operations in reciprocal space whenever possible. Thanks to the convolution theorem and high-speed FFT routines, this can be done in most places.

The tight-binding functions used in our basis set are not orthogonal; as a result, a generalized eigenvalue problem must be solved to find the eigenstates of the Hamiltonian matrix. This is commonly done using the Cholesky decomposition.[27] Using this technique requires simultaneous storage of three matrices the size of the Hamiltonian matrix. In addition, the mixed-basis set with a large number of plane waves can suffer from overdetermination problems: the basis set may become nearly linearly dependent, which makes the generalized eigenvalue problem ill-conditioned. To avoid this problem and to save storage space, we have made the tight-binding functions orthogonal to the plane waves.[12] Using the reciprocal-space expansion of the tight-binding functions, this is done simply by setting the Fourier components of the tight-binding functions to zero for every reciprocal-lattice vector that is included in the plane-wave part of the basis set. Because the tight-binding functions are now orthogonal to all of the plane waves, and because the plane waves themselves are mutually orthogonal, the overlap matrix of the Cholesky decomposition is reduced to $n_{TB}\times n_{TB}$, where $n_{TB}$ is the number of tight-binding functions in the basis set.

Although the mixed-basis scheme reduces the size of the basis set by several orders of magnitude, for a supercell calculation the Hamiltonian matrix is still on the order of $2000\times2000$. To reduce the computation time for the eigenvalue problem, we use group theoretical methods to block diagonalize the Hamiltonian matrix.[28] We also use an iterative diagonalization scheme to solve the eigenvalue problem.[29]

### B. Test calculations

We performed a great number of calculations to test the reliability and accuracy of the programs, the basis set, and the pseudopotentials. Test calculations were performed for the two-atom unit cell of ZnSe and for a series of supercells. Because these are the first accurate pseudopotential calculations for a II-VI semiconductor, special care was devoted to these tests.

The two-atom cell tests were performed for three basic material parameters: the lattice constant $a_{lat}$, the bulk modulus $B$, and the transverse optical (TO) phonon frequency, $\nu_{TO}$. The lattice constant and the bulk modulus are then derived from a fit of $E_{Tot}$ $(a_{lat})$ for five or six different lattice constants to the Murnaghan equation of state.[30] We calculated more than 50 sets of Murnaghan equation fits, determining the lattice constant and the bulk modulus as we changed different calculational parameters. In these tests we varied such things as the form of the tight-binding functions (either pseudoatomic functions or Gaussians with different radii), the plane-wave cutoff, the size of the FFT grid, the local component, and the cutoff radii of the pseudopotentials. In all of our tests, the lattice constant was predicted to within a few percent of experiment, and the bulk modulus to within 30%. The ability of these tests to reproduce small energy differences (about 1 meV) guarantees the accuracy of our defect calculations. Based on these tests, we have chosen for our defect calculations a basis set of all plane waves

with energies up to 9 Ryd, and one set of five pseudoatomic basis functions per zinc atom. The calculated lattice constant and bulk modulus are 5.65 Å and 0.62 Mbar, compared with the experimental values of 5.67 Å and 0.63 Mbar.

The first-principles norm-conserving pseudopotentials used in this work were generated according to the Hamann-Schlüter-Chiang scheme.[24] The $s$, $p$, and $d$ cutoff radii were 1.6, 1.56, and 1.01 a.u. for the zinc potential and 1.40, 1.40, and 1.51 a.u. for the selenium potential. The Zn $d$ cutoff radius lies beyond the maximum of the zinc $3d$ wave function (which is at 0.56 a.u.). We tested the effects of this large cutoff radius on the pseudopotential by comparing it to a pseudopotential with a zinc $d$ cutoff radius, 0.5 a.u., within the wave-function maximum. In a comparison of the bulk properties of ZnSe, the only one that was affected by the change in cutoff radius was the zone-center TO phonon frequency. The calculated TO phonon frequency using the smaller radius was 26.2 meV (=6.33 THz) compared with experimental values of 25–26 meV;[31] using the larger cutoff radius induces a 10% error in the calculated frequency. In our defect calculations, we used the larger core radius, which produces a smoother pseudopotential and pseudoatomic function. (This allows us to use a smaller FFT grid.) We have confirmed in supercell defect tests that increasing the core radius changes defect formation energies by less than 0.1 eV. We have also calculated the band structure of ZnSe, and find agreement to within 0.25 eV with the band structure calculated using all-electron methods.[32] Figure 2 shows our calculated band structure. We note that the band gap is smaller than its experimental value, due to the well-documented LDA error. The implications of this deficiency for our defect calculations will be discussed where appropriate.

We also performed a series of defect supercell tests to check the effects of the basis set, pseudopotentials, and FFT grid on defect formation energies. The error bar due to the combined effects of plane-wave cutoff, the FFT grid size, and the pseudopotential is *0.1–0.2 eV*. We have also checked our results with respect to supercell convergence. Comparative tests were performed for 8-, 16-, and 32-atom supercells. We calculate the cell-size correction between an $N_2$-atom supercell and an $N_1$-atom supercell as

$$\Delta(N_2/N_1)=E^{N_2}_{\text{defect}}-E^{N_2}_{\text{perfect}} -E^{N_1}_{\text{defect}}+E^{N_1}_{\text{perfect}}\ , \tag{2}$$

where $E^N_{\text{defect}}$ and $E^N_{\text{perfect}}$ are the calculated total energies of an $N$-atom supercell with a defect and a perfect $N$-atom supercell, respectively. Cell-size corrections were calculated for the zinc vacancy and the zinc interstitial in different charge states. (As discussed later, these two defects are the most abundant native defects in *stoichiometric p-type* ZnSe.) Two trends emerge from these calculations. One is that the defects in the neutral charge state are well converged in a 32-atom cell, but that the charged defects (either positive or negative) may have cell-size errors of up to 0.4 eV. The second is that the correction terms are positive when going from a smaller to a larger supercell, indicating that the true defect-formation energies are larger than those calculated in the 32-atom supercell. (Cell-size corrections are not included in our results, however.) Since our main conclusion will be that the defect-formation energies are too large to allow for substantial compensation, the supercell tests strengthen our results by showing that the true formation energies are likely to be even larger than our calculated values.

## III. RESULTS FOR INDIVIDUAL NATIVE POINT DEFECTS IN ZnSe

Using the methods described above, we have calculated the energies of all of the basic native point defects in ZnSe: $Zn_i$, $Se_i$ (interstitials), $V_{Zn}$, $V_{Se}$ (vacancies), $Zn_{Se}$, and $Se_{Zn}$ (antisites). Interstitial energies were calculated

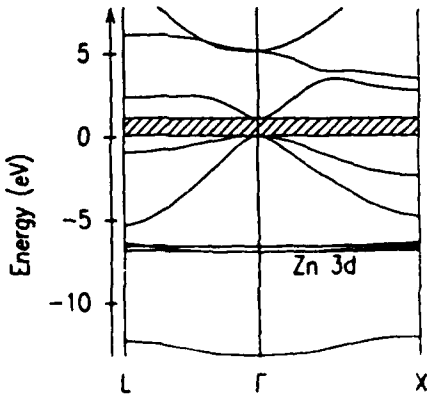



FIG. 2. ZnSe band structure. Calculated ZnSe band structure along high-symmetry directions in the Brillouin zone. The band gap is hatched. [Note that the theoretical (LDA) value of the direct band gap is 1 eV, compared to the experimental value of 2.7 eV. LDA problems are discussed in the text.] Note the set of narrow bands associated with the zinc $3d$ electrons.

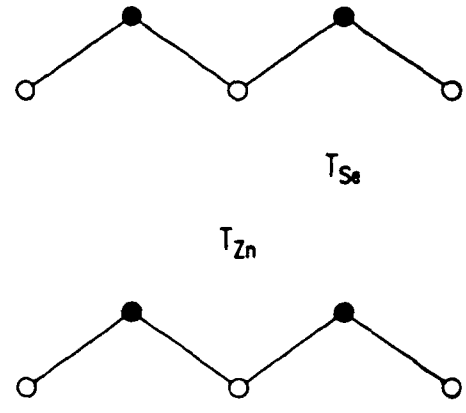



FIG. 3. Location of the two tetrahedral interstitial sites in ZnSe. Zinc atoms are represented by solid circles, selenium atoms by open circles. The $x$ axis is in the [110] direction and the $y$ axis in the [001] direction. Each tetrahedral site is surrounded by four atoms of the same type (two of which are out of the plane of the figure and are not shown).

at the two tetrahedral interstitial sites in ZnSe. The one site ($T_{Zn}$) is tetrahedrally surrounded by four Zn atoms while the other site ($T_{Se}$) is surrounded by four Se atoms (Fig. 3). These two sites are the most favorable interstitial sites for large atoms because they are surrounded by a large empty space. In fact, the nearest-neighbor configuration of the tetrahedral interstitial sites is the same as that of the atomic sites of the perfect crystal. Separate calculations were performed for the different charge states of each defect. Our calculated energies for each charge state of each defect ($E_l$) are presented in Table I.

### A. Relaxation

The formation energy of a point defect can be reduced by relaxation of the atoms surrounding the defect. The lattice-relaxation energy is the energy difference between the unrelaxed defect (all atoms around the defect in their ideal lattice sites) and the relaxed defect. To find these relaxations, we must map out the total energy as a function of the positions of the surrounding atoms, and find the energy minimum. This is an arduous task because the relaxation of each atom is a function of the relaxation of the others. For practical applications, we limit the number of degrees of freedom in our relaxation calculations. The minimum of the total energy is found by calculating the total energy with different relaxations, and fitting the total-energy surface to a parabolic form about the minimum. We limit our calculations to symmetric relaxations in which each shell of atoms relaxes by the same amount ("breathing-mode" relaxations). A possible cause of nonsymmetric relaxation is the Jahn-Teller effect, which occurs when a degenerate electronic state in the band gap is partially filled with electrons.[33] Our pri-

TABLE I. Native-point-defect energies, in eV. $E_l$ is the calculated energy for a supercell containing the defect in a 32-atom cell geometry (excluding relaxation energy). The energy of a perfect ("bulk") 32-atom supercell is −27 363.522 eV. $\epsilon_l = E_l - (N_{Zn} + N_{Se})E_{ZnSe}$; $\epsilon_l$ includes the appropriate shift for charge states (referred to a state with the Fermi level at the top of the valence band). $E_l$ and $\epsilon_l$ individually should *not* be interpreted as carrying physical meaning; in particular, they depend on the pseudopotential and on the choice of reference for the Fermi level. $F_l$ is the formation energy for the defect in stoichiometric $p$-type ZnSe (doped with $10^{18}$ cm$^{-3}$ Li) at $T = 600$ K; it is based upon specific reference energies for individual Zn and Se atoms, calculated by solving the complete set of reaction equations, as described in the text. $F_l$ includes the relaxation energy (which is set to 1 eV where not calculated). $R_l$ is the calculated relaxation energy. Although $F_l$ is a physically meaningful energy, it should not be construed as *the* formation energy of a single defect; as explained in the text, such a concept is not defined in a compound semiconductor.

| Defect | Charge | $n_l$ | $E_l$ | $\epsilon_l$ | $F_l$ | $R_l$ |
|---|---|---|---|---|---|---|
| [illegible] | 2− | 1− | −25 906.228 | 598.343 | 2.20 | 0.00 |
| [illegible] | 1− | 1− | −25 908.114 | 598.378 | 2.09 | |
| [illegible] | 0 | 1− | −25 909.881 | 598.532 | 1.81 | |
| $Zn_i$ ($T_{Se}$) | 0 | 1+ | −28 809.490 | −590.856 | 3.87 | |
| $Zn_i$ ($T_{Se}$) | 1+ | 1+ | −28 813.860 | −592.538 | 2.97 | 0.43 |
| $Zn_i$ ($T_{Se}$) | 2+ | 1+ | −28 818.018 | −594.007 | 1.80 | 0.34 |
| $Zn_i$ ($T_{Zn}$) | 0 | 1+ | −28 810.117 | −591.483 | 3.24 | |
| $Zn_i$ ($T_{Zn}$) | 1+ | 1+ | −28 814.075 | −592.753 | 2.82 | 0.22 |
| $Zn_i$ ($T_{Zn}$) | 2+ | 1+ | −28 817.795 | −593.784 | 2.16 | 0.20 |
| $V_{Se}$ | 0 | 1+ | −27 099.998 | −591.585 | 3.14 | |
| $V_{Se}$ | 1+ | 1+ | −27 102.872 | −592.381 | 2.55 | |
| $V_{Se}$ | 2+ | 1+ | −27 105.503 | −592.935 | 2.21 | |
| $Se_i$ ($T_{Zn}$) | 2− | 1− | −27 609.480 | 604.089 | 6.94 | |
| $Se_i$ ($T_{Zn}$) | 1− | 1− | −27 613.571 | 602.530 | 5.59 | |
| $Se_i$ ($T_{Zn}$) | 0 | 1− | −27 617.084 | 601.550 | 4.83 | |
| $Se_i$ ($T_{Zn}$) | 1+ | 1− | −27 620.356 | 600.810 | 4.30 | |
| $Se_i$ ($T_{Zn}$) | 2+ | 1− | −27 623.412 | 600.287 | 3.98 | |
| $Se_i$ ($T_{Zn}$) | 3+ | 1− | −27 626.280 | 599.950 | 3.86 | |
| $Se_i$ ($T_{Zn}$) | 4+ | 1− | −27 628.975 | 599.787 | 3.91 | |
| $Zn_{Se}$ | 2− | 2+ | −28 542.164 | −1183.563 | 6.46 | |
| $Zn_{Se}$ | 1− | 2+ | −28 546.076 | −1185.014 | 5.22 | |
| $Zn_{Se}$ | 0 | 2+ | −28 549.775 | −1186.252 | 4.20 | |
| $Zn_{Se}$ | 1+ | 2+ | −28 553.036 | −1187.051 | 3.61 | |
| $Zn_{Se}$ | 2+ | 2+ | −28 556.086 | −1187.640 | 3.61 | 0.62 |
| $Se_{Zn}$ | 2− | 2− | −26 159.399 | 1199.827 | 6.96 | |
| $Se_{Zn}$ | 1− | 2− | −26 163.706 | 1197.699 | 5.01 | |
| $Se_{Zn}$ | 0 | 2− | −26 167.784 | 1195.739 | 3.29 | |
| $Se_{Zn}$ | 1+ | 2− | −26 171.377 | 1194.295 | 2.06 | |
| $Se_{Zn}$ | 2+ | 2− | −26 174.689 | 1193.132 | 1.95 | 0.16 |

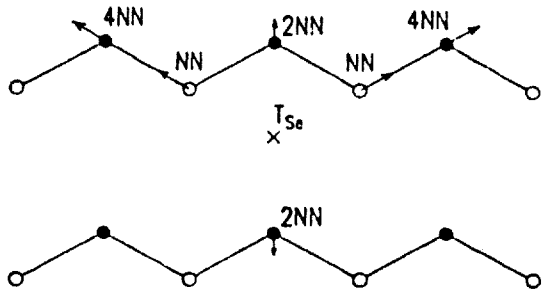



FIG. 4. Relaxations around the $T_{Se}$ site. Outward relaxations are shown for the first-, second-, and fourth-nearest neighbors (NN, 2NN, and 4NN, respectively) around a zinc interstitial on the $T_{Se}$ site. Four of the fourth-nearest neighbors relax in the same direction as the first-nearest neighbors. The magnitude of the relaxations is exaggerated for clarity.

mary objective is to study the behavior of defects in doped ZnSe, where defect states in the gap will either be completely full (in *n*-type material) or completely empty (in *p*-type material). Consequently, we will calculate relaxations only for defects that do not have partially filled states in the gap. For these cases, no Jahn-Teller relaxation will occur.

For substitutional site defects, we have relaxed the shell of four nearest-neighbor atoms. (The nearest-neighbor distance in ZnSe is 2.45 Å.) We have found the relaxations to be small in all cases (smaller than 0.2 Å), and the second-nearest-neighbor relaxation should be even smaller. (The second-nearest-neighbor distance is 4.01 Å.) For the tetrahedral interstitial sites, we relaxed both the first- (consisting of four atoms) and the second-neighbor (six atoms) shells simultaneously. For the relaxation of the nearest-neighbors around a tetrahedral interstitial site ($T_{Zn}$ and $T_{Se}$), it turned out to be important to also include relaxations of fourth-nearest-neighbor atoms that are located on a line through the tetrahedral interstitial site and the first neighbors (Fig. 4). The reason is that a breathing relaxation of the nearest neighbors will change the length of the bond to these fourth-nearest neighbors. Since bond-stretching forces are larger than bond-bending forces, it is energetically more favorable to move the fourth-nearest neighbors outwards. This effect was not included in previous calculations.[12]

As a rule, the calculated relaxations were small: the largest relaxation energy that we found was about 0.6 eV and the typical relaxation distance was 0.1 Å, which is only 4% of the ZnSe bond length of 2.54 Å. Our calculated relaxations are listed in Table II. We will now describe our results for the individual native defects.

TABLE II. Calculated relaxations for native point defects in ZnSe. Calculated energies $E_{relax}$ and relaxations of nearest (NN) and next-nearest (NNN) neighbors. A positive relaxation indicates relaxation outward from the defect. All relaxations in the table are symmetric. For relaxations about interstitial defects, the fourth-nearest neighbors relaxed by the same amount as the NN's.

| Defect | $E_{relax}$ (eV) | Relaxation (Å) NN | NNN |
|---|---|---|---|
| $Zn_i^{+}$ ($T_{Se}$) | 0.43 | 0.11 | 0.05 |
| $Zn_i^{2+}$ ($T_{Se}$) | 0.34 | 0.06 | 0.09 |
| $Zn_i^{+}$ ($T_{Zn}$) | 0.22 | 0.09 | 0.03 |
| $V_{Zn}^{2-}$ | 0.0 | 0.0 | |
| $Zn_{Se}^{2+}$ | 0.62 | 0.2 | |
| $Se_{Zn}^{2+}$ | 0.16 | 0.1 | |

### B. Zinc self-interstitial

We start with the zinc self-interstitial ($Zn_i$). The neutral zinc interstitial has two electrons occupying a single level in the band gap. The possible charge states are therefore 0, 1+, and 2+, making the defect a double donor in *p*-type material. The zinc interstitial in ZnSe is a particularly interesting defect because it was the first isolated native interstitial directly observed in a semiconductor.[34] Using optically detected magnetic resonance, Rong and Watkins identified the isolated zinc self-interstitial in the 1+ charge state. The defects were produced by electron irradiation of ZnSe at a temperature of 4.2 K. They found that the interstitial occupied the $T_{Se}$ site, and that there were no asymmetric relaxations of either the nearest-neighbor Se atoms or the second-nearest-neighbor Zn atoms. They also found the transition level from the 1+ to the 2+ charge state to occur when the Fermi level is at 1.9 eV above the valence-band edge. (This energy is the thermodynamic level in the gap.) The interstitials were observed to be mobile at temperatures above 260 K.[35] Although experiment can determine the site of the defect and its symmetry, the magnitudes of the relaxations and their energies must be determined from theory. In our calculations for the zinc self-interstitial, we have performed relaxations for the $T_{Se}$ site in the 2+ and the 1+ charge states and for the $T_{Zn}$ site in the 1+ charge state. The calculated relaxations are listed in Table II. The calculated valence-charge-density contours for the $T_{Se}$ site interstitial (in the 1+ charge state) are shown in Fig. 5. Including relaxations, the energies of the 1+ charge state at the two interstitial sites are the same to within the accuracy of our calculations. Rong and Watkins actually also found a signal which they tentatively identified as the Zn self-interstitial at the $T_{Zn}$ site. Although this defect is not stable, its energy may be only slightly higher than that of the self-interstitial at the $T_{Se}$ site. We calculate (including relaxation energies) a value of 1.4 eV for the level in the gap between 2+ and 1+ interstitial on the $T_{Se}$ site. The agreement with experiment is reasonable, in light of the large errors in the band gap inherent in LDA. For the 1+ $T_{Se}$ site, Van de Walle and Laks[36] have calculated the values of the hyperfine parameters for the central Zn atom, and the first- and third-nearest-neighbor Se atoms. The hyperfine calculations included the relaxations of the neighboring atoms. The agreement between the theoretical and experimental hyperfine parameters[34] is very good. This confirms both the experimental identification of the defect and the accuracy of the calculated relaxations.

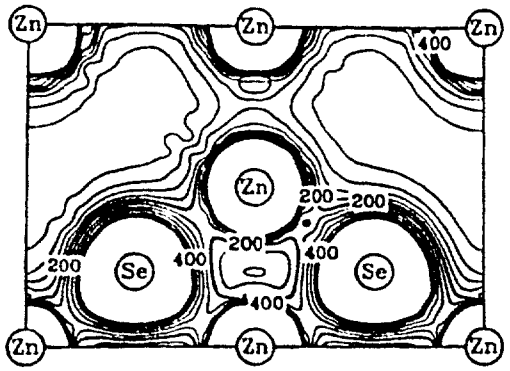



FIG. 5. Valence-charge density of the Zn self-interstitial. Contour plot of the valence-charge density around a zinc interstitial at the $T_{Se}$ site, in the 1+ charge state. Relaxations of neighboring atoms are included. The $x$ axis is along the [110] direction and the $y$ axis along the [001] direction. The interstitial atom is at the center of the plot. The charge density is given in units of electrons per 32-atom cell volume (=728.2 Å$^3$) and the contour spacing is 40.

### C. Zinc vacancy

The other native point defect in ZnSe that has been positively identified is the zinc vacancy. This defect was also observed in electron-irradiated ZnSe at low temperatures by Watkins.[37] The neutral zinc vacancy has a threefold-degenerate level in the band gap (with a capacity of six electrons), of which four are occupied. The possible charge states are 1− and 2−, making the vacancy an acceptor in $n$-type ZnSe. Watkins observed the 1− charge state using electron paramagnetic resonance and found that it undergoes a Jahn-Teller distortion. The 1− vacancy fills five electrons out of the six electron states in the gap. The remaining hole is localized by the Jahn-Teller effect on one of the four nearest-neighbor Se atoms. The Se atom with the hole moves in toward the vacancy and the symmetry of the defect is lowered from tetrahedral (point group $T_d$) to trigonal ($C_{3v}$). The energy lowering from the Jahn-Teller relaxation is estimated by Watkins[35] to be 0.35 eV. The level in the gap between the 2− and the 1− charge states is found to be at 0.66 eV above the valence-band edge. We have calculated the relaxation for the 2− charge state, which is expected to be symmetric. No relaxation (to within 0.02 Å) was found for the nearest neighbors. We did not explicitly calculate the low-symmetry relaxations for the 1− charge state. To compare with experiment, we can look at the level in the gap for the 2− to (unrelaxed) 1− transition, which we find at −0.03 eV. Adding in Watkins's estimate of the Jahn-Teller energy of the 1− vacancy (0.35 eV) brings the level to 0.32 eV. Taking the LDA deficiency into account, this value is once again in reasonable agreement with experiment (0.66 eV).

### D. Other defects

There is no direct experimental evidence about the Se interstitial or the Se vacancy. The neutral Se interstitial has four electrons in a threefold-degenerate level. The possible charge states range from 4+ to 2−. Of the two tetrahedral interstitial sites, the $T_{Zn}$ site is preferred. The neutral Se vacancy has a single level in the gap that is fully occupied by two electrons. The possible charge states are 1+ and 2+. We find that the formation energy of either of these defects is so high that they do not play any important role in ZnSe.

Nothing is known experimentally about the two antisite defects, either. Both the neutral Zn-on-Se antisite and the Se-on-Zn antisite have two electrons in a threefold-degenerate level in the gap. Possible charge states range from 2+ to 4−. For the neutral Se on Zn, we find a large lattice relaxation, in which the antisite lowers its energy by about 0.7 eV by moving about 1 Å along the ⟨111⟩ direction toward a tetrahedral interstitial site. This relaxation is favorable because it lowers the energy of the electrons in the states in the gap; it does not occur for the 2+ charge state, where the states in the gap are empty. This relaxation is similar to that found theoretically for the As-on-Ga antisite defect in GaAs.[38,39] The large lattice relaxations of the antisite in GaAs have explained the puzzling properties of the defect known as $EL2$. The occurrence of a similar relaxation in ZnSe may also be observable experimentally.

## IV. DETERMINATION OF DEFECT CONCENTRATIONS

In this section we will describe how to determine the concentration of the native point defects from their calculated formation energies. Determining defect concentrations for a compound semiconductor is more difficult than for an elemental system, where the total energy of a single bulk atom is well defined. In the latter case the formation energy of a native point defect can be unambiguously determined from an $N$-atom defect supercell calculation: the defect formation energy is the difference between the calculated supercell energy and $N$ times the energy of a single bulk atom. In the case of a Si self-interstitial, for instance, an extra Si atom is placed inside the crystal. This Si atom can be thought of as taken "from the surface," a process which does not change the nature of the surface; the crystal simply becomes one atom larger, and the reference energy is simply the energy of a bulk Si atom, which can be determined from a bulk calculation. This analysis cannot be applied to a compound semiconductor like ZnSe. Here, the energy of a *pair* of zinc and selenium atoms is well defined, but the energy of a *single* zinc or selenium atom depends on interactions between the crystal and its external environment. Hence energies and concentrations of the native defects will also depend on the environment.

Well-defined energies for defects in a compound semiconductor, such as ZnSe, can be determined in one of two ways. The first way is to define the energies of reactions that conserve the relative number of zinc and selenium atoms. For example, the reaction energy for forming a *pair* of zinc and selenium interstitials can be defined in the same way as the formation energy of a single silicon self-interstitial. This is true because removing a pair of zinc and selenium atoms from the surface does not

change the nature of the surface. The energy of a pair of zinc and selenium atoms can be determined from a bulk calculation. The second way to define defect formation energies is to introduce an external reservoir of zinc atoms. Zinc atoms may be added to the crystal from the reservoir, or removed from the crystal and added to the reservoir. The energy of zinc atoms in the reservoir is constant, and in thermal equilibrium with the crystal. The reservoir allows us to assign an energy to the zinc atoms. Since the sum of the zinc and selenium energies is determined by the total energy of the perfect ZnSe cell, the zinc energy determines the selenium energy. This, in turn, allows us to determine the formation energy of any defect.

The one problem with the latter prescription is that we must choose a zinc reservoir and calculate its energy. The choice of the reservoir depends on the conditions under which the crystal is grown. Instead of limiting our choice to a single reservoir energy, we picture a reservoir in which we can change the energy of the zinc atom to any value that we choose. Or, equivalently, we can set the difference between the zinc and selenium energies, $\delta E$, to be any value that we want. The formation energy for the $i$th defect $F_i$ can then be expressed as

$$\begin{aligned} F_i &= E_i - N_{\mathrm{Zn}} E_{\mathrm{Zn}} - N_{\mathrm{Se}} E_{\mathrm{Se}} \\ &= E_i - (N_{\mathrm{Zn}} + N_{\mathrm{Se}}) E_{\mathrm{ZnSe}} - (N_{\mathrm{Zn}} - N_{\mathrm{Se}}) \delta E \\ &= \epsilon_i - n_i \delta E \ . \end{aligned} \tag{3}$$

Here $E_i$ is the total energy of the supercell for the $i$th defect, containing $N_{\mathrm{Zn}}$ zinc atoms and $N_{\mathrm{Se}}$ selenium atoms,

$$\delta E = (E_{\mathrm{Zn}} - E_{\mathrm{Se}})/2 \ ,$$

$$E_{\mathrm{ZnSe}} = (E_{\mathrm{Zn}} + E_{\mathrm{Se}})/2 \ ,$$

$$\epsilon_i = E_i - (N_{\mathrm{Zn}} + N_{\mathrm{Se}}) E_{\mathrm{ZnSe}} \ ,$$

and $n_i = N_{\mathrm{Zn}} - N_{\mathrm{Se}}$. $E_{\mathrm{ZnSe}}$ is determined from a calculation of the energy of a perfect ZnSe supercell. (At $T=0$ K, $E_{\mathrm{Zn}}$ and $E_{\mathrm{Se}}$ can be identified with the chemical potentials for Zn and Se.) $n_i$ is the number of extra Zn atoms that must be added to form the defect ($1+$ for $V_{\mathrm{Se}}$, $2-$ for $\mathrm{Se_{Zn}}$, etc.), independent of the size of the supercell.

From the formation energy of the defect and its entropy $S$ we can determine its fraction $C_i$ by

$$\begin{aligned} C_i \equiv M_i^{\mathrm{defect}} / M_i^{\mathrm{site}} &= e^{S/k_B} e^{-F_i/k_B T} \\ &= e^{S/k_B} e^{-(\epsilon_i - n_i \delta E)/k_B T} \\ &= e^{S/k_B - \epsilon_i/k_B T} y^{n_i} = a_i y^{n_i} \ . \end{aligned} \tag{4}$$

where $M_i^{\mathrm{defect}}$ and $M_i^{\mathrm{site}}$ are the total number of defects and the total number of defect sites in the crystal, $y = \exp(\delta E/k_B T)$, and $a_i = \exp(S/k_B - \epsilon_i/k_B T)$. The concentration of defects per unit volume is found from the fraction by multiplying by the site concentration, which, for ZnSe, is $2.2 \times 10^{22}$ cm$^{-3}$. (The site concentration is the number of atoms *of each type* per unit volume, not the *total* number of atoms per unit volume.)

The stoichiometry parameter is defined as

$$X \equiv \frac{M_{\mathrm{Se}} - M_{\mathrm{Zn}}}{M_{\mathrm{Se}} + M_{\mathrm{Zn}}} = -\tfrac{1}{2} \sum_i n_i C_i = -\tfrac{1}{2} \sum_i n_i a_i y^{n_i} \ . \tag{5}$$

$M_{\mathrm{Zn}}$ and $M_{\mathrm{Se}}$ are the total numbers of zinc and selenium atoms in the crystal. $X=0$ for perfect stoichiometry, and $X>0$ for Se-rich. The factor of $\frac{1}{2}$ enters this equation because the stoichiometry parameter is defined by dividing by the total number of atoms in the crystal, while the fractions are divided by the number of sites of each type. The stoichiometry parameter defined in this way only takes into account the deviations from perfect stoichiometry due to native point defects. In real crystals, deviations from stoichiometry may also be present because of higher dimensional defects, surfaces, and precipitates. Substitutional impurities are counted as the host species that they replace, since this replacement does not directly introduce native defects.

This formulation allows us to determine native-point-defect formation energies and concentrations for any value of $\delta E$. In practice, it is more convenient to fix the stoichiometry parameter $X$ and determine from $X$ the value of $\delta E$. We can do this simply by solving for $y$ given $X$, using Eq. (5). The problem is essentially finding a root of a polynomial, which can be done quickly and easily using standard algorithms.[22] For our purposes, it will be clearer to talk about the stoichiometry $X$ rather than the value of $\delta E$. However, we stress that our approach is quite generally valid for describing a system in equilibrium with other solids or gases which impose certain conditions on the chemical potential and thus determine $\delta E$.

The defect-formation energies for charged defects depend on the Fermi level. (The Fermi level is used here as the chemical potential of the electrons.) Consider a reaction in which a neutral defect $D^0$ with formation energy $E^0$ is ionized to its positive charge state $D^+$ with energy $E^+$:

$$D^0 \rightarrow D^+ + e^- \ . \tag{6}$$

The energy of this reaction is $E^0 - (E^+ + E_F)$ where $E_F$ is the Fermi level. We can treat the combination of $D^+ + e^-$ as a single entity with energy $E^+ + E_F$. The quantity $E^+$ is the energy of the charge-state defect when the Fermi level is at zero. To follow the convention of choosing the zero of the Fermi level in a semiconductor at the valence-band maximum, we must change $E^+$ to $E^+ + E_V$ and $E_F$ to $E_F - E_V$, $E_V$ being the valence-band maximum. This can be generalized to any charge state: the defect energy for charge state $m$ is $E^{(m)} + mE_F$, where we change $E^{(m)}$ to $E^{(m)} + mE_V$ to place the zero of the Fermi level at the top of the valence band.

Dealing with charged defects thus requires that we know the energy of the top of the valence band in the defect cell. The quantity that we want is (the energy of) the top of the bulk valence band in the defect cell. One cannot simply use the $\mathbf{k}=0$ band structure of the defect supercell because it includes the distortion of the band structure in the immediate vicinity of the defect. We should use the valence-band energy far away from the defect, which would correspond to a pure bulk calculation.

However, there is no absolute reference for potentials or eigenvalues in supercell calculations;[40] we therefore need additional information in order to "line up" the bulk band structure with the defect supercell. Here we use the "model solid" theory of Van de Walle and Martin,[41,42] which allows us to calculate the average electrostatic potential of a system of atoms on an absolute energy scale. Since the defect supercell and the bulk supercell in general contain a different set of atoms, the average electrostatic potentials will be shifted; the magnitude of the shift is predicted by the model solid theory. We have verified, by inspection of the locally averaged self-consistent electrostatic potential in the defect cell, that the model-solid lineup indeed provides an adequate description of the potential shift.

## V. RESULTS AND DISCUSSION

In this section we present our calculated native-point-defect concentrations in ZnSe and discuss the general question of native-point-defect compensation in wide-band-gap semiconductors. As described in the preceding section, we need to know the formation energy and entropy of the defect to determine its concentration. Entropy calculations are unnecessary because our results are insensitive to values in the range $0 \le S \le 10k_B$ (see Fig. 6). By comparison, a recent accurate calculation[43,44] of the formation entropy of the Si self-interstitial found a formation entropy of $(5-6)k_B$ for the ground state. The Si self-interstitial represents an extreme case in that the ground-state configuration has low symmetry, which accounts for half of the formation entropy. It is therefore highly unlikely that the entropies for native point defects in ZnSe could be larger than $10k_B$.

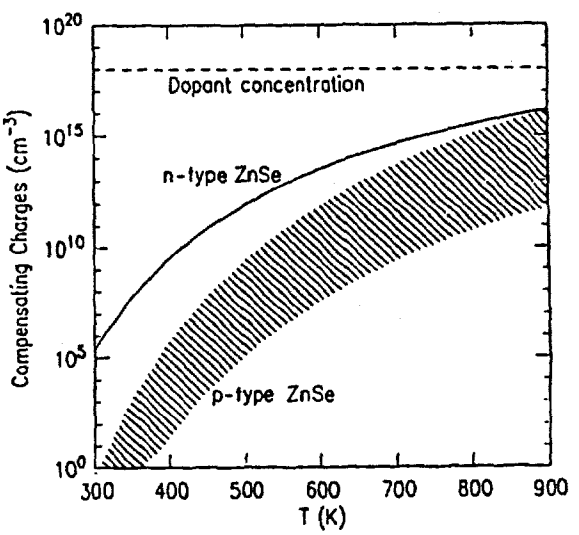



FIG. 6. Native-point-defect compensation in stoichiometric ZnSe. For *p*-type ZnSe the net number of electrons produced by all native point defects is shown. For *n*-type material, the net number of holes is shown. The range of values shown for *p*-type ZnSe is bounded by assuming relaxations of 1 eV and entropy of 10 $k_B$ per defect for an upper bound and $k_B$ for a lower bound. For *n*-type ZnSe, the uncertainty of the results is increased by the LDA band-gap error, and no error estimate is included.

We have explicitly calculated relaxation energies for the defects which are dominant in *p*-type ZnSe. As shown in Table I, the relaxation energies are all smaller than 0.7 eV, which is of the same order as calculated relaxations in other semiconductors including Si,[15,16] diamond,[21] and GaAs.[38,39,45] The defect-formation energies of other defects are high enough that, even with a generous estimate of the atomic relaxation energies (we assume 1 eV), the concentrations remain very low. Even a relaxation of 2 eV does not change our results (that is, the native-point-defect concentrations are still too low to compensate in stoichiometric material).

It is important to assess to what extent the LDA band-gap problem affects the formation energies. The band-gap problem has no direct effect on the concentrations of defects in *p*-type material, where electron levels in the band gap are empty. For *n*-type material, the position of occupied electron levels in the gap is uncertain due to the LDA band-gap error. We will treat this uncertainty by using the worst-case value of the defect energy.

Figure 6 shows the concentrations of minority carriers produced by native point defects for *p*-type stoichiometric ZnSe. The individual native-point-defect concentrations are given in Table III. The error bar is determined by allowing the formation entropy to range from 0 to 10 $k_B$. The results shown are for material with $10^{18}$ cm$^{-3}$ dopants. The dopants are used to determine the position of the Fermi level. As the temperature increases, the Fermi level moves closer to the middle of the band gap. (This is because the intrinsic carrier concentration increases with temperature.) This effect slows the increase of the defect concentrations with increasing temperature. (Jansen and Sankey,[12] in their determination of the defect concentrations at $T=1658$ K, set the Fermi level at the valence-band edge. Taking into account the shift of the Fermi level with temperature would substantially lower their concentrations.) The dominant native point defects are $Zn_i$ (a double donor) and $V_{Zn}$ (an acceptor). At molecular-beam epitaxy growth temperatures ($T=600$ K) the concentration of minority carriers produced is less than $10^{12}$ cm$^{-3}$. For ZnSe grown at higher temperatures and not rapidly quenched, excess native point defects will be annihilated during cooling, as long

TABLE III. Native-point-defect concentrations in stoichiometric *p*-type ZnSe, at $T=600$ K. Only defects with concentrations greater than $10^5$ cm$^3$ are shown. A formation entropy of 5 $k_b$ is assumed for each defect.

| Defect | Charge | Concentration (cm$^{-3}$) |
|---|---|---|
| $Zn_i$ ($T_{Se}$) | 2+ | $2.48\times10^{9}$ |
| $V_{Zn}$ | 0 | $2.14\times10^{9}$ |
| $Se_{Zn}$ | 2+ | $1.46\times10^{8}$ |
| $Se_{Zn}$ | 1+ | $1.71\times10^{7}$ |
| $V_{Zn}$ | 1− | $8.70\times10^{6}$ |
| $V_{Zn}$ | 2− | $1.17\times10^{6}$ |
| $Zn_i$ ($T_{Zn}$) | 2+ | $2.21\times10^{6}$ |
| $V_{Se}$ | 2+ | $8.58\times10^{5}$ |

as the defects are free to move. The temperature that determines the native-point-defect concentration is that at which the defects become immobile. The dominant native point defects in *p*-type material, $V_{Zn}$ and $Zn_i$, are known experimentally to be mobile at temperatures above 400 and 260 *K*, respectively.[35] At 400 K the native-point-defect concentrations in *p*-type ZnSe are at most $10^5\ cm^{-3}$.

We have also determined the concentrations of native defects in *n*-type ZnSe (Fig. 6 and Table IV). The dominant native point defects are $V_{Zn}^{2-}$ and $Zn_{Se}^{-}$. It is an experimental fact that *n*-type ZnSe can be produced much more easily than *p* type. If native-point-defect compensation were the cause, we would expect that defect concentrations would be much larger in *p*-type material than in *n* type. Instead, we find that defect concentrations are actually somewhat larger in *n*-type ZnSe. This is an additional proof that native point defects do not compensate *p*-type ZnSe. (For *n*-type ZnSe the levels in the band gap were shifted up by the LDA band-gap error, which increases the defect-formation energies. Actual defect concentrations may be higher than shown; this would further support the notion that native-point-defect compensation is no greater in *p*-type ZnSe than in *n* type.) We conclude that, in stoichiometric ZnSe, native-point-defect compensation will be insignificant.

To further support our conclusions, we have derived native-point-defect concentrations for diamond (Fig. 7) based on the first-principles defect energies of Bernholc *et al.*[21] The doping level is again $10^{18}\ cm^{-3}$. The calculations show that only the vacancy is found in significant concentrations. Experimentally, diamond is easy to make *p* type but difficult to make *n* type. In our results for *n*-type material, we once again made a worst-case assumption about the LDA error (the LDA band-gap error is about 1 eV here): the Fermi level was shifted rigidly with the conduction-band edge, but the levels in the gap for the vacancy were not shifted at all. This assumption significantly increases the defect concentrations. The true concentrations are probably much smaller still. At a chemical-vapor-deposition-growth temperature of 1100 K, the number of holes produced in *n*-type diamond by native point defects is at most $2\times10^{13}\ cm^{-3}$. Clearly, the concentrations of native point defects in both stoichiometric ZnSe and diamond are far too low to produce significant compensation.

TABLE IV. Native-point-defect concentrations in *n*-type ZnSe. Same conditions as Table III.

| Defect | Charge | Concentration ($cm^{-3}$) |
|---|---|---|
| $V_{Zn}$ | 2− | $1.37\times10^{13}$ |
| $Zn_{Se}$ | 1− | $5.23\times10^{12}$ |
| $Zn_{Se}$ | 0 | $1.26\times10^{12}$ |
| $Zn_{Se}$ | 2− | $3.54\times10^{11}$ |
| $V_{Se}$ | 0 | $3.65\times10^{9}$ |
| $Zn_i$ ($T_{Zn}$) | 0 | $5.07\times10^{8}$ |
| $Zn_{Se}$ | 1+ | $6.19\times10^{7}$ |
| $V_{Se}$ | 1+ | $1.70\times10^{5}$ |

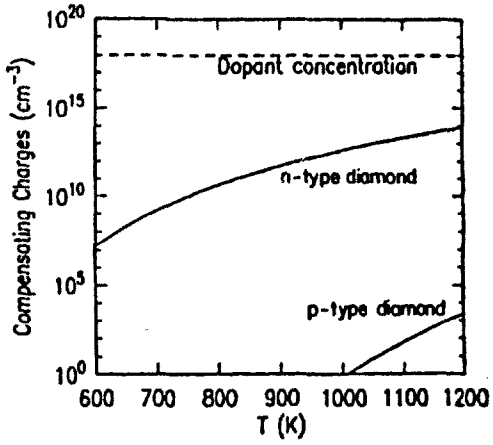



FIG. 7. Native-point-defect compensation in diamond. The native-point-defect concentrations are shown for both *n*-type and *p*-type diamond containing $10^{18}$ dopants. For *n*-type diamond, a worst-case treatment of the LDA band-gap error is used; the Fermi level is shifted up by the band-gap error, while the defect levels in the gap are not shifted. This gives an upper bound on the defect concentrations; actual defect concentrations in *n*-type diamond are probably much lower.

Jansen and Sankey[12] have calculated the formation energies of native defects in ZnSe and ZnTe, using pseudopotentials that treat the Zn *d* electrons as frozen-core states. From their defect-formation energies they derived defect concentrations as a function of temperature and stoichiometry. To derive defect concentrations from their calculated energies, Jansen and Sankey impose the stoichiometry parameter as an external constraint. This is equivalent to using our own method with an unknown chemical potential that produces the same stoichiometry. Their results exhibit the same trends as our own, although the actual defect concentrations are different, due in part to their approximate treatment of the *d* electrons. They also find the zinc interstitial and the selenium-on-zinc antisite defect to be the dominant defects in *p*-type ZnSe, which are both donors. In *n*-type ZnSe, they find that the zinc vacancy and the selenium antisite are dominant (both acceptors).

Jansen and Sankey suggest that their results explain why ZnSe prefers to be *p* type. Their calculated defect concentrations for *n*-type ZnTe are higher than those for *n*-type ZnSe, while their defect concentrations for *p*-type ZnSe are higher than those for *p*-type ZnTe. Based on these results, Jansen and Sankey propose that native point defects hamper the doping in *p*-type ZnSe and *n*-type ZnTe. Careful examination reveals that this conclusion is doubtful. For ZnSe, their numbers indicate that native-point-defect concentrations are 3,000 times *lower* in *p*-type material than in *n* type. Thus, if anything, native-point-defect compensation should prevent the growth of *n*-type ZnSe. Furthermore, their results were reported for a very high temperature ($T$=1658 K), and do not apply to the question of compensation for material that is grown at 600 K and never thermally annealed at higher temperatures. At the lower temperature, the native-point-defect concentrations derived from their

calculated energies are only $\sim 10^8$ cm$^{-3}$, even lower than our own predictions, and far too small to compensate doping.

Our conclusion that the concentrations of native point defects in stoichiometric ZnSe are very low does not mean that native-point-defect compensation in ZnSe never occurs. If the sample is grown with even a slight deviation from perfect stoichiometry, the concentration of native point defects will necessarily be very large, even at $T=0$ K (assuming that deviations from stoichiometry are accommodated by native point defects alone). Because the density of atomic sites in ZnSe is $4\times10^{22}$ cm$^{-3}$, a deviation from stoichiometry as small as $10^{-4}$ implies a defect concentration of about $10^{18}$ cm$^{-3}$. Our major conclusion for nonstoichiometric material is that the native point defects that accommodate deviations from stoichiometry are *always* those that compensate the majority carriers. For *p*-type ZnSe, the dominant defect is $Zn_I$ in Zn-rich material, and $Se_{Zn}$ in Se-rich material; we find that both are double donors. For *n*-type ZnSe the dominant (acceptor) defects are $Zn_{Se}$ and $V_{Zn}$ for Zn- and Se-rich materials, respectively. Similar results were found by Jansen and Sankey.[12]

This defect structure is much richer than that used in many previous analyses of native point defects in II-VI semiconductors. Ray and Kröger,[10] for example, studied the properties of ZnSe as a function of Zn partial pressure, and analyzed their results in terms of only two native defects: $V_{Zn}$ (an acceptor) in Se-rich material and $V_{Se}$ (a donor) in Zn-rich material. Their model predicts that Zn-rich material should be *n* type and Se-rich material *p* type. Our results show that this model is oversimplified; changing the stoichiometry from Zn-rich to Se-rich will not convert *n*-type ZnSe to *p*-type ZnSe. Instead, the greater the deviation from stoichiometry in either direction, the greater the level of compensation.

Having addressed the native-point-defect compensation issue quantitatively, we now reexamine the notion that native-point-defect compensation increases with the width of the band gap. Let us state precisely the standard argument for this trend: For *p*-type material, imagine a prototypal compensating native-donor defect that, when neutral, introduces one electron into a state in the gap; the formation energy for this defect, $E^0$, is assumed not to depend on the width of the band gap. The energy gained by transferring the electron from the level in the gap $E_L$ to the Fermi level $E_F$ should, in contrast, increase with the width of the gap; thus, the net energy needed to form compensating defects, $E^0-(E_L-E_F)$, should decrease as the band gap increases. The flaw in this argument is that it assumes that $E_L$ and $E^0$ are independent of one another. Actually, the level in the gap is *defined by* $E_L=E^0-E^+$, where $E^++E_F$ is the (Fermi-level dependent) energy of formation of the positive charge-state defect (Fig. 8). Substituting this definition into the formula for the net energy of compensation, we find

$$E^0-(E_L-E_F)=E^0-(E^0-E^+-E_F)=E^++E_F\ ,$$

*independent of the energy of formation of the neutral defect.* We see that native-point-defect compensation will increase with the width of the band gap if and only if $E^++E_F$ *decreases* with increasing band gap. For this to be true, additional assumptions would have to be made about how the formation energy of the charged defect changes as the band gap widens. In particular, there is no *a priori* reason to assume that $E^+$ would be lower in wide-band-gap materials. The first-principles results reported in this paper definitely show that, whatever the qualitative trends may be, the native-point-defect concentrations in stoichiometric ZnSe and in diamond are far too small to be a source of compensation.

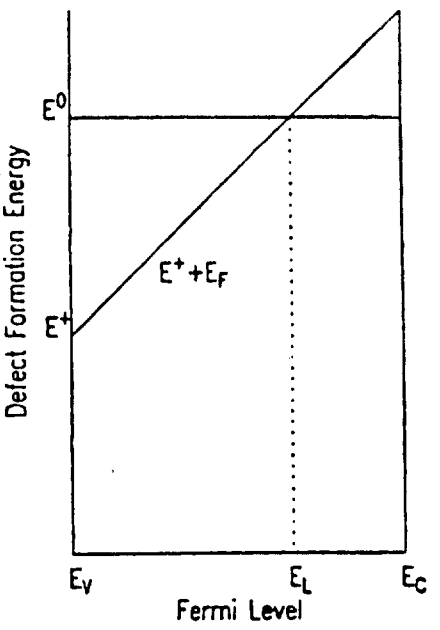



FIG. 8. Level in the gap for a donor defect. Total energy as a function of the Fermi level for the positive and neutral charge state of a prototypal donor defect. The level in the gap ($E_L$) is the value of the Fermi level at which the two charge states have the same energy.

## VI. CONCLUSIONS

We have described a mixed-basis pseudopotential total-energy scheme which is fast and efficient enough for supercell calculations (Sec. II). These programs are capable of accurately describing the structural properties of ZnSe, including the important effects of the zinc 3*d*-electron states. We use these techniques to examine native-point-defect compensation in ZnSe; we calculate the total energies of the native point defects in ZnSe (Sec. III) and show how to extract defect concentrations from these energies (Sec. IV). We have shown that native-point-defect concentrations are very low in stoichiometric ZnSe; in nonstoichiometric material, both *n*- and *p*-type doping would be compensated (Sec. V). These results indicate that native-point-defect compensation is *not* responsible for the doping problems in ZnSe (and other wide-band-gap semiconductors). Efforts at understanding these problems should be aimed at investigating the behavior of individual dopant impurities.

## ACKNOWLEDGMENTS

We are indebted to D. Vanderbilt for his iterative diagonalization program. We acknowledge helpful conversations with R. Bhargava, J.M. DePuydt, T. Marshall, J. Tersoff, and G.D. Watkins. D.B. Laks acknowledges partial support from IBM. This work was supported in part by NSF grant No. ECS-89-21159 and ONR Contract No. N00014-84-0396.


## APPENDIX

The mixed-basis set results in a much smaller Hamiltonian matrix than a plane-wave basis, but requires more effort for matrix element evaluation. This appendix describes the techniques used to calculate various matrix elements. Setting up the Hamiltonian matrix for the Kohn-Sham equations requires evaluation of three types of matrix elements:[46] (1) kinetic energy and overlap, (2) local potential, $V^L$ (pseudopotential plus screening potentials), and (3) nonlocal pseudopotential.

Each type of matrix element must be evaluated between (1) two plane waves (PW-PW), (2) two tight-binding functions (TB-TB), and (3) a plane wave and a tight-binding function (TB-PW). The PW-PW matrix elements are evaluated in the same way as they are in standard calculations with a pure plane-wave basis set.[46] We will limit our discussion to (TB-TB) and (TB-PW) matrix elements.

A tight-binding function centered on atomic site T of the crystal's unit cell can be written as

$$\phi_{\mathbf{k}}(\mathbf{r})=\frac{1}{\sqrt{\Omega}}\sum_{\mathbf{R}} f(\mathbf{r}-\mathbf{R}-\mathbf{T})e^{i\mathbf{k}\cdot(\mathbf{R}+\mathbf{T})}\,, \tag{A1}$$

where $\Omega$ is the unit cell volume, R is a direct lattice vector, and $f$ is a localized real function (the pseudoatomic Zn $3d$ wave function in this work), which is of the form

$$f(\mathbf{r})=f(r)Z_{lm}(\hat{\mathbf{r}}) \tag{A2}$$

where $Z_{lm}$ is a Bethe Kubic-harmonic function. The tight-binding function may be Fourier transformed to give

$$\phi_{\mathbf{k}}(\mathbf{G})=e^{-i\mathbf{G}\cdot\mathbf{T}}f(\mathbf{k}+\mathbf{G})\,, \tag{A3}$$

$$f(\mathbf{K})=\frac{4\pi(-i)^l}{\Omega}Z_{lm}(\hat{K})\int_0^\infty j_l(Kr)f(r)dr\,, \tag{A4}$$

where $j_l$ is a spherical Bessel function.

The number of TB-TB matrix elements is proportional to $n_{\mathrm{TB}}^2$. (For our supercell calculations, $n_{\mathrm{TB}}$ is typically 80.) The TB-TB matrix elements require an integration over the crystal unit cell, making their evaluation a numerically intensive process. Instead of simply summing over a real-space grid, it is more efficient to perform these operations in reciprocal space.

### 1. Local matrix elements

The overlap and the kinetic-energy matrix elements are calculated in the manner of Louie, Ho, and Cohen.[25] The TB-TB matrix elements of the local potential are complicated by the presence of the $V^L(\mathbf{r})$ term. We can write the matrix element in reciprocal space as:

$$H_{ij}^L=\sum_{\mathbf{G}}\sum_{\mathbf{G}'}\phi_i^*(\mathbf{G})V^L(\mathbf{G}-\mathbf{G}')\phi_j(\mathbf{G}')\,. \tag{A5}$$

This formula is of order $n_G^2$, where $n_G$ is the number of reciprocal-lattice vectors, and its evaluation is prohibitively expensive. To convert this expression into a more convenient form, we define

$$F_i(\mathbf{G}')=\sum_{\mathbf{G}}\phi_i(\mathbf{G})V^L(\mathbf{G}'-\mathbf{G})\,, \tag{A6}$$

so that

$$H_{ij}^L=\sum_{\mathbf{G}'}F_i^*(\mathbf{G}')\phi_j(\mathbf{G}')\,. \tag{A7}$$

We can now use the convolution theorem to evaluate $F_i(\mathbf{G}')$:

$$F_i(\mathbf{r})=\phi_i(\mathbf{r})V^L(\mathbf{r})\,. \tag{A8}$$

This procedure is very efficient because the real-space operations are limited to $n_{\mathrm{TB}}$ sets of multiplications and $2n_{\mathrm{TB}}$ FFT's [for the convolution in Eq. (A8)]. The only operations that are performed $n_{\mathrm{TB}}(n_{\mathrm{TB}}+1)/2$ times are a multiplication and a summation over the reciprocal lattice [Eq. (A7)]. Only a small amount of extra computer memory is needed to store one copy of the function $F_i$.

### 2. Nonlocal matrix elements

The nonlocal pseudopotential TB-TB matrix elements can, in principle, be evaluated by applying projection operators to the reciprocal-space expansion of the tight-binding functions. Instead, we take advantage of the localized nature of the tight-binding functions by using the so-called on-site approximation. In the on-site approximation, the nonlocal pseudopotential acts only on tight-binding functions on the same site as the potential. This approximation is well justified because both the nonlocal potential and the tight-binding functions used in our calculations are very short ranged. Thus the product of the nonlocal potential on one site and the tight-binding function on a different site will be extremely small. The on-site approximation reduces the nonlocal matrix element to a single radial integral.[25]

The nonlocal matrix elements involve one additional complication. The on-site approximation is based on the assumption that the tight-binding functions are short-ranged functions. This is true only for the full tight-binding functions; the tight-binding functions used in our calculations are orthogonalized to the plane waves by setting their low-frequency Fourier components to zero. This orthogonalization leads to tight-binding functions with long-range oscillations, for which the on-site approximation is no longer valid. To correct for this we write the tight-binding functions in the following form:

$$|\phi_i\rangle=|\phi_i^F\rangle-|\phi_i^C\rangle \tag{A9}$$

where $|\phi_i\rangle$ is the actual tight-binding function of the basis set, $|\phi_i^F\rangle$ is the "full" tight-binding function before

orthogonalization, and $|\phi_i^C\rangle$ consists of the orthogonalization terms. The $|\phi_i^C\rangle$ are just the components of the full tight-binding function for all reciprocal-lattice vectors in the plane-wave basis.

$$|\phi_i^C\rangle = \sum_{|\mathbf{k}+\mathbf{G}|^2 < E_{\text{cut}}} \phi_i^F(\mathbf{k}+\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} . \tag{A10}$$

In this form, the nonlocal matrix element becomes

$$\begin{aligned} H_{ij}^{NL} &= \langle \phi_i | V | \phi_j \rangle \\ &= \langle \phi_i^F | V | \phi_j^F \rangle - \langle \phi_i^C | V | \phi_j^F \rangle \\ &\quad - \langle \phi_i^F | V | \phi_j^C \rangle + \langle \phi_i^C | V | \phi_j^C \rangle , \end{aligned} \tag{A11}$$

where the nonlocal potential is now represented by $V$. The on-site approximation can now be applied to the first term. The on-site approximation is also applied to the second term:

$$\langle \phi_i^C | V | \phi_j^F \rangle = \frac{1}{\Omega} \sum_{|\mathbf{k}+\mathbf{G}|^2 < E_{\text{cut}}} \phi_i^{F*}(\mathbf{k}+\mathbf{G}) \int e^{-i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{r}+\mathbf{T}_\beta)} V_{l,\beta}(r) f_l(r) Z_{lm}(\hat{\mathbf{r}}) d\mathbf{r} . \tag{A12}$$

The $\mathbf{T}_\beta$ term appears because the tight-binding function is centered about a specific atomic site, while the plane wave is defined with respect to the origin. By applying both the angular-momentum expansion of the plane waves and the orthogonality of the Kubic-Harmonic functions, we can reduce this expression to

$$\langle \phi_i^C | V | \phi_j^F \rangle = \frac{4\pi(-i)^l}{\Omega} \sum_{K^2 < E_{\text{cut}}} e^{-i\mathbf{K}\cdot\mathbf{T}_\beta} \phi_i^{F*}(\mathbf{K}) Z_{lm}(\hat{\mathbf{K}}) \int r^2 V_{l,\beta}(r) j_l(Kr) f_l(r) dr , \tag{A13}$$

where $\mathbf{K}=\mathbf{k}+\mathbf{G}$. We are again left with a set of radial integrals in real space. The third term is of the same form as (the complex conjugate of) the second term. The fourth term is treated as an expansion in terms of plane waves:

$$\begin{aligned} \langle \phi_i^C | V | \phi_j^C \rangle = \sum_{\mathbf{G}} \sum_{\mathbf{G}'} & \phi_i^{F*}(\mathbf{k}+\mathbf{G}) \langle \mathbf{k}+\mathbf{G} | V^{NL} | \mathbf{k}+\mathbf{G}' \rangle \\ & \times \phi_j^F(\mathbf{k}+\mathbf{G}') . \end{aligned} \tag{A14}$$

Because both summations are limited to reciprocal-lattice vectors in the plane-wave basis set, evaluation of the fourth term takes about the same amount of time as the nonlocal plane-wave matrix elements.

### 3. TB-PW matrix elements

We now describe the calculation of the TB-PW mixed-basis terms. These terms are relatively simple. The overlap and kinetic-energy matrix elements are all zero because of the orthogonalization of the tight-binding functions to the plane waves. The local-potential matrix elements between the $i$th plane wave and the $j$th tight-binding function are

$$\begin{aligned} H_{ij}^{L} &= \frac{1}{\Omega} \int_{uc} e^{-i(\mathbf{k}+\mathbf{G}_i)\cdot\mathbf{r}} V^L(\mathbf{r}) \phi_j(\mathbf{r}) d\mathbf{r} \\ &= \frac{1}{\Omega} \int_{uc} e^{-i(\mathbf{k}+\mathbf{G}_i)\cdot\mathbf{r}} F_j(\mathbf{r}) d\mathbf{r} \\ &= F_j(\mathbf{G}_i) . \end{aligned} \tag{A15}$$

The function $F_j$ is the same function that was introduced in the calculation of the local TB-TB matrix elements [Eq. (A6)]. Thus we get all of the local TB-PW matrix elements without any extra calculations. We see here the convenience of expanding the tight-binding functions in reciprocal space. For the nonlocal TB-PW matrix elements we will use the on-site approximation and an appropriate correction term once more:

$$\begin{aligned} H_{ij}^{NL} &= \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j \rangle \\ &= \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j^F \rangle - \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j^C \rangle . \end{aligned} \tag{A16}$$

The $\langle \mathbf{k}+\mathbf{G}_i |$ represents the $i$th plane wave. The two terms here are just the same as the third [Eq. (A13)] and fourth [Eq. A14)] terms in the TB-TB nonlocal matrix elements described above, where

$$\phi_i^F(\mathbf{k}+\mathbf{G}) = \delta_{\mathbf{G},\mathbf{G}_i} . \tag{A17}$$

Thus, these terms do not require any extra computation either.

*Present address: National Renewable Energy Laboratory, Golden, CO 80401.
†Present address: Xerox Palo Alto Research Center, Palo Alto, CA 94304.
‡Present address: IBM Zurich Laboratories, Rüschlikon, Switzerland.

[1]R. A. Reynolds, J. Vac. Sci. Technol. A 7, 269 (1989).
[2]Y. S. Park and B. K. Shin, *Topics in Applied Physics* (Springer, New York, 1977), Vol. 17, p. 133.
[3]H. Hartmann, R. Mach, and B. Selle, Curr. Top. Mater. Sci. 9, 1 (1982).
[4]R. Bhargava, J. Cryst. Growth 59, 15 (1982).
[5]S. Y. Ren, J. D. Dow, and S. Klemm, J. Appl. Phys. 66, 2065 (1989).
[6]D. J. Chadi and K. J. Chang, Appl. Phys. Lett. 55, 575 (1989).
[7]G. F. Neumark, Phys. Rev. Lett. 62, 1800 (1989).
[8]G. F. Neumark, J. Appl. Phys. 51, 3383 (1980).
[9]G. Mandel, Phys. Rev. 134, A1073 (1964).
[10]A. K. Ray and F. A. Kröger, J. Electrochem. Soc. 125, 1348 (1978).

orthogonalization, and $|\phi_i^C\rangle$ consists of the orthogonalization terms. The $|\phi_i^C\rangle$ are just the components of the full tight-binding function for all reciprocal-lattice vectors in the plane-wave basis.

$$|\phi_i^C\rangle = \sum_{|\mathbf{k}+\mathbf{G}|^2 < E_{\text{cut}}} \phi_i^F(\mathbf{k}+\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} . \tag{A10}$$

In this form, the nonlocal matrix element becomes

$$\begin{aligned} H_{ij}^{NL} &= \langle \phi_i | V | \phi_j \rangle \\ &= \langle \phi_i^F | V | \phi_j^F \rangle - \langle \phi_i^C | V | \phi_j^F \rangle \\ &\quad - \langle \phi_i^F | V | \phi_j^C \rangle + \langle \phi_i^C | V | \phi_j^C \rangle , \end{aligned} \tag{A11}$$

where the nonlocal potential is now represented by $V$. The on-site approximation can now be applied to the first term. The on-site approximation is also applied to the second term:

$$\langle \phi_i^C | V | \phi_j^F \rangle = \frac{1}{\Omega} \sum_{|\mathbf{k}+\mathbf{G}|^2 < E_{\text{cut}}} \phi_i^{F*}(\mathbf{k}+\mathbf{G}) \int e^{-i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{r}+\mathbf{T}_\beta)} V_{l,\beta}(r) f_i(r) Z_{lm}(\hat{\mathbf{r}}) d\mathbf{r} . \tag{A12}$$

The $\mathbf{T}_\beta$ term appears because the tight-binding function is centered about a specific atomic site, while the plane wave is defined with respect to the origin. By applying both the angular-momentum expansion of the plane waves and the orthogonality of the Kubic-Harmonic functions, we can reduce this expression to

$$\langle \phi_i^C | V | \phi_j^F \rangle = \frac{4\pi(-i)^l}{\Omega} \sum_{K^2 < E_{\text{cut}}} e^{-i\mathbf{K}\cdot\mathbf{T}_\beta} \phi_i^{F*}(\mathbf{K}) Z_{lm}(\hat{\mathbf{K}}) \int r^2 V_{l,\beta}(r) j_l(Kr) f_i(r) dr , \tag{A13}$$

where $\mathbf{K}=\mathbf{k}+\mathbf{G}$. We are again left with a set of radial integrals in real space. The third term is of the same form as (the complex conjugate of) the second term. The fourth term is treated as an expansion in terms of plane waves:

$$\begin{aligned} \langle \phi_i^C | V | \phi_j^C \rangle = \sum_{\mathbf{G}} \sum_{\mathbf{G}'} \phi_i^{F*}(\mathbf{k}+\mathbf{G}) \langle \mathbf{k}+\mathbf{G} | V^{NL} | \mathbf{k}+\mathbf{G}' \rangle \\ \times \phi_j^F(\mathbf{k}+\mathbf{G}') . \end{aligned} \tag{A14}$$

Because both summations are limited to reciprocal-lattice vectors in the plane-wave basis set, evaluation of the fourth term takes about the same amount of time as the nonlocal plane-wave matrix elements.

### 3. TB-PW matrix elements

We now describe the calculation of the TB-PW mixed-basis terms. These terms are relatively simple. The overlap and kinetic-energy matrix elements are all zero because of the orthogonalization of the tight-binding functions to the plane waves. The local-potential matrix elements between the $i$th plane wave and the $j$th tight-binding function are

$$\begin{aligned} H_{ij}^{L} &= \frac{1}{\Omega} \int_{uc} e^{-i(\mathbf{k}+\mathbf{G}_i)\cdot\mathbf{r}} V^L(\mathbf{r}) \phi_j(\mathbf{r}) d\mathbf{r} \\ &= \frac{1}{\Omega} \int_{uc} e^{-i(\mathbf{k}+\mathbf{G}_i)\cdot\mathbf{r}} F_j(\mathbf{r}) d\mathbf{r} \\ &= F_j(\mathbf{G}_i) . \end{aligned} \tag{A15}$$

The function $F_j$ is the same function that was introduced in the calculation of the local TB-TB matrix elements [Eq. (A6)]. Thus we get all of the local TB-PW matrix elements without any extra calculations. We see here the convenience of expanding the tight-binding functions in reciprocal space. For the nonlocal TB-PW matrix elements we will use the on-site approximation and an appropriate correction term once more:

$$\begin{aligned} H_{ij}^{NL} &= \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j \rangle \\ &= \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j^F \rangle - \langle \mathbf{k}+\mathbf{G}_i | V^{NL} | \phi_j^C \rangle . \end{aligned} \tag{A16}$$

The $\langle \mathbf{k}+\mathbf{G}_i |$ represents the $i$th plane wave. The two terms here are just the same as the third [Eq. (A13)] and fourth [Eq. A14)] terms in the TB-TB nonlocal matrix elements described above, where

$$\phi_i^F(\mathbf{k}+\mathbf{G}) = \delta_{\mathbf{G},\mathbf{G}_i} . \tag{A17}$$

Thus, these terms do not require any extra computation either.

*Present address: National Renewable Energy Laboratory, Golden, CO 80401.
†Present address: Xerox Palo Alto Research Center, Palo Alto, CA 94304.
‡Present address: IBM Zurich Laboratories, Rüschlikon, Switzerland

[1]R. A. Reynolds, J. Vac. Sci. Technol. A 7, 269 (1989).
[2]Y. S. Park and B. K. Shin, *Topics in Applied Physics* (Springer, New York, 1977), Vol. 17, p. 133.
[3]H. Hartmann, R. Mach, and B. Selle, Curr. Top. Mater. Sci. 9, 1 (1982).
[4]R. Bhargava, J. Cryst. Growth 59, 15 (1982).
[5]S. Y. Ren, J. D. Dow, and S Klemm, J. Appl. Phys. 66, 2065 (1989).
[6]D. J. Chadi and K. J. Chang, Appl. Phys. Lett. 55, 575 (1989).
[7]G. F. Neumark, Phys. Rev. Lett. 62, 1800 (1989).
[8]G. F. Neumark, J. Appl. Phys. 51, 3383 (1980).
[9]G. Mandel, Phys. Rev. 134, A1073 (1964).
[10]A. K. Ray and F. A. Kröger, J. Electrochem. Soc. 125, 1348 (1978).

[11]Y. Marfaing, Prog. Cryst. Growth Charact. 4, 317 (1981).
[12]R. W. Jansen and O. F. Sankey, Phys. Rev. B 39, 3192 (1989).
[13]H. Cheng, J. M. DePuydt, J. E. Potts, and M. A. Haase, J. Cryst. Growth 95, 512 (1989).
[14]R. M. Park, M. B. Troffer, C. M. Rouleau, J. M. DePuydt, and M. A. Haase, Appl. Phys. Lett. 57, 2127 (1990).
[15]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. 52, 1814 (1984).
[16]Y. Bar-Yam and J. D. Joannopoulos, J. Electron. Mater. 14A, 261 (1985).
[17]G. A. Baraff and M. Schlüter, Phys. Rev. Lett. 55, 1327 (1985).
[18]C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. B 39, 10791 (1989), and references therein.
[19]T. Oguchi, T. Sasaki, and H. Katayama-Yoshida, in *Impurities, Defects, and Diffusion in Semiconductors: Bulk and Layered Structures*, edited by D. J. Wolford, J. Bernholc, and E. E. Haller, MRS Symposia Proceedings No. 163 (Materials Research Society, Pittsburgh, 1990), p. 81.
[20]S.-H. Wei and A. Zunger, Phys. Rev. B 37, 8958 (1988).
[21]J. Bernholc, A. Antonelli, T. M. Del Sol, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. 61, 2689 (1988).
[22]W. H. Press, B. P. Flannery, S. A. Teukolsy, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1986).
[23]P. Hohenberg and W. Kohn. Phys. Rev. 136, B864 (1964).
[24]D. R. Hamann, M. Schluter, and C. Chiang, Phys. Rev. Lett. 43, 1494 (1979).
[25]S. G. Louie, K.-M. Ho, and M. L. Cohen, Phys. Rev. B 19, 1774 (1979).
[26]D. B. Laks, Ph.D. thesis, Columbia University, 1990.
[27]R. S. Martin and J. H. Wilkinson, Numer. Math. 11, 99 (1968).
[28]J. C. Slater, *Quantum Theory of Molecules and Solids* (McGraw Hill, New York, 1965), Vol. 2.
[29]R. Natarajan and D. Vanderbilt, J. Comput. Phys. 82, 218 (1989).
[30]F. D. Murnaghan, Proc. Natl. Acad. Sci. USA 30, 244 (1944).
[31]H. E. Gumlich, D. Theis, and D. Tschierse, in *Numerical Data and Functional Relationships in Science and Technology*, edited by O. Madelung, Landolt-Börnstein, New Series, Group III, Vol. 17, Pt. b (Springer-Verlag, Berlin, 1982), p. 126.
[32]A. Continenza, S. Massidda, and A. J. Freeman, Phys. Rev. B 38, 12 996 (1988).
[33]J. Calaway, *Quantum Theory of the Solid State* (Academic, New York, 1974).
[34]F. Rong and G. D. Watkins, Phys. Rev. Lett. 58, 1486 (1987).
[35]G. D. Watkins, in *Defect Control in Semiconductors, Proceedings of the International Conference on the Science and Technology of Defect Control in Semiconductors, Yokohama, 1989*, edited by K. Sumino (North-Holland, Amsterdam, 1990), p. 933.
[36]C. G. Van de Walle and D. B. Laks, in *Proceedings of 20th International Conference on the Physics of Semiconductors, Thessaloniki, 1990*, edited by E. M. Anastassakis and J. D. Joannopoulis (World Scientific, Singapore, 1990), p. 722.
[37]G. D. Watkins, in *Radiation Defects in Semiconductors 1976*, Proceedings of the International Conference on Radiation Effects in Semiconductors, edited by N. B. Urli and J. W. Corbett, IOP Conf. Proc. No. 31 (Institute of Physics and Physical Society, London, 1977), p. 95.
[38]J. Dabrowski and M. Scheffler, Phys. Rev. Lett. 60, 2183 (1988).
[39]D. J. Chadi and K. J. Chang, Phys. Rev. Lett. 60, 2187 (1988).
[40]L. Kleinman, Phys. Rev. B 24, 7412 (1981).
[41]C. G. Van de Walle and R. M. Martin, Phys. Rev. B 35, 8154 (1987).
[42]C. G. Van de Walle, Ph.D. thesis, Stanford University, 1986.
[43]P. E. Blöchl and S. T. Pantelides (unpublished).
[44]P. E. Blöchl, D. B. Laks, S. T. Pantelides, E. Smargiassi, R. Car, W. Andreoni, and M. Parrinello, in Proceedings of the 20th International Conference on the Physics of Semiconductors (Ref. 36).
[45]D. J. Chadi and K. J. Chang, Phys. Rev. Lett. 61, 873 (1988).
[46]J. Ihm, A. Zunger, and M. L. Cohen. J. Phys. C 12, 4409 (1979).

PHYSICAL REVIEW B VOLUME 47, NUMBER 15 15 APRIL 1993-I

# First-principles calculations of solubilities and doping limits: Li, Na, and N in ZnSe

Chris G. Van de Walle
*Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304*

D. B. Laks*
*National Renewable Energy Laboratory, Golden, Colorado 80401*

G. F. Neumark
*Division of Metallurgy and Materials Science, Columbia University, New York, New York 10027*

S. T. Pantelides
*IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 1 September 1992)

We present a comprehensive theoretical approach to determine concentrations of dopant impurities in semiconductors. The formalism is applied to the problem of acceptor doping in ZnSe. Formation energies and concentrations of impurities and native defects are expressed as a function of chemical potentials, for which experimentally accessible ranges are calculated. We show that limitations in the achievable hole concentrations can be explained by two mechanisms: one is the competition between various substitutional and interstitial configurations (compensation), the other is the solubility limit imposed by formation of other phases. Nitrogen is most promising among the dopants examined.

## I. INTRODUCTION

Limits to semiconductor doping have been widely discussed both in III-V and II-VI compounds. In wide-band-gap semiconductors the problem is particularly acute because typically one type of conduction ($n$-type or $p$-type) is very difficult to obtain. Detailed understanding of these phenomena has been lacking. In this paper we present a formalism that allows the determination of defect concentrations, impurity solubilities, and doping levels. It includes a unifying treatment of the various interactions of the dopant with the host lattice (in substitutional or interstitial sites), the role of native defects, and the factors that determine solubility. The key quantities that enter this formulation can be obtained from first-principles electronic-structure calculations.

Our formalism entails the following steps.

(1) Calculation of the total energies of all native defects and of the various configurations that can be assumed by the impurity in the crystal, including lattice relaxations and different charge states.

(2) Application of thermodynamics to express the relevant equilibrium concentrations at the temperature of interest, and determination of the resulting Fermi level from the condition of overall charge neutrality for all impurity configurations, native defects, and free carriers.

At this juncture the results remain functions of two chemical potentials (one for the host crystal, which controls the stoichiometry, and one for the impurity) which are free parameters to be fixed by growth conditions. The physical meaning of these chemical potentials and the way in which they enter the formalism will be discussed in detail. Thermodynamics imposes bounds on the experimentally accessible range of these chemical potentials; the bounds result from the last step of the formalism:

(3) Calculation of the heats of formation of competing phases that can be formed out of the constituents (i.e., the impurity and the component elements of the semiconductor).

By imposing these bounds we obtain limits on impurity concentrations, i.e. we can calculate solubilities.

We illustrate the approach with the technologically important example of ZnSe, in which $n$-type doping poses *no* difficulties, but well-conducting, reproducible $p$-type doping has been very hard to achieve.[1] Despite some impressive recent experimental advances,[2–5] the cause of the doping problem has remained unclear. Lithium was the first dopant to yield reproducible, well-conducting $p$-type ZnSe, with a net acceptor concentration of about $10^{17}$ cm$^{-3}$.[2,4] More recently, N doping up to $10^{18}$ cm$^{-3}$ was achieved and led to the fabrication of a blue semiconductor laser.[6] In the case of Li, our results will demonstrate quantitatively the competition between substitutional and interstitial impurity configurations,[7] and identify a regime where the desired substitutional form dominates (earlier work[8] that proposed this competition as the source of compensation did not recognize the existence of such different regimes). Our results will also show that there is a second overriding cause that limits doping, namely the overall solubility which is constrained by the formation of a $Li_2Se$ phase.[9] These conclusions agree with experimental observations on Li-doped ZnSe; more importantly, they provide guidelines for optimizing

growth conditions.

Our investigations of Na indicate qualitative similarities to Li, but significant quantitative differences which render Na unsuitable for $p$-type doping of ZnSe. Indeed, its low solubility explains the failure of doping attempts with Na.[10] Nitrogen, finally, does not exhibit a substitutional/interstitial competition and, in addition, has the highest solubility.

One of the strengths of the formalism is that it treats native point defects (vacancies, self-interstitials, and antisites) and dopant impurities on an equal footing, allowing us to investigate whether native defects can form a significant source of compensation.[11] We find that under appropriate growth conditions the native defect concentration is usually so low as to be unimportant. We previously arrived at this conclusion from a study of native defect concentrations as a function of Fermi-level position, in which the exact nature of the dopant impurities was left unspecified.[12] Our current results confirm that native defects do *not* form a generic source of compensation in ZnSe. We also present more detailed information on defect concentrations under various growth conditions.

The present investigation of acceptor impurities in ZnSe relates to an experimental problem of high current interest due to the impact on a blue semiconductor laser; however, we stress that the formalism is a general one that can be applied to the study of doping in any semiconductor system.

## II. METHODS

### A. Total-energy calculations

In this section we describe how to calculate concentrations of defects and impurities in the semiconductor. In order to obtain quantitative results, one needs reliable values for the total energies of defects and impurities; we have obtained such values from first-principles calculations. The calculations are based on density-functional theory in the local-density approximation,[13] and *ab initio* pseudopotentials.[14] Scalar relativistic effects are included in the pseudopotentials, but spin-orbit splitting is neglected; our calculated bands are therefore averages over the states which would be split due to spin-orbit interactions. The spin-orbit splitting can be introduced as a perturbation. The Fermi-level positions which we will discuss should still be interpreted as referred to the top of the valence band ($\Gamma_8$). We use a mixed-basis approach, ensuring an accurate description of the structural properties by explicitly including the $d$ states of the Zn atoms. The basis set contains plane waves with kinetic energy up to 9 Ry, and pseudoatomic orbitals on the Zn and N atoms.[12] In order to achieve a proper description of Li and Na we implement a nonlinear core exchange-correlation correction.[15]

The defect calculations are performed in a supercell geometry, with 32-atom supercells providing adequate accuracy. Relaxations of up to two shells of neighbors are included. Additional details about the calculational approach are given in Ref. 16. We have used this approach to obtain total energies for the dopant impurities (Li, Na, and N) which are the subject of this study, in their various configurations in the lattice. Our calculations are typically carried out for the charged (positive or negative) state of the dopant. To avoid divergence of the long-range Coulomb terms, the G=0 terms in the total energy are always calculated for a neutral system. A justification of this self-consistent approach to treat charge states of impurities was given in Ref. 17. Because of the extended nature of the wave function the neutral charge state of a shallow impurity is difficult to treat within the supercell formalism; instead, we use experimental activation energies[18] to determine the formation energies of the neutral charge state. Finally, we make use of our previously calculated[12,16] energy values for native defects in all relevant charge states.

### B. Formation energies, concentrations, and chemical potentials

The equilibrium concentration of an impurity or defect $D_i$ is given by

$$[D_i] = N_{\text{sites}} \exp\left(-\frac{E_{\text{form}}(D_i)}{kT}\right), \tag{2.1}$$

where $N_{\text{sites}}$ is the appropriate site concentration [e.g., for substitutional Li ($\text{Li}_{\text{Zn}}$), $N_{\text{sites}}$ is the number of substitutional Zn sites in the crystal, $2.2\times10^{22}\text{cm}^{-3}$], and $E_{\text{form}}$ is the formation energy. The energy appearing in Eq. (2.1) is a Gibbs free energy, which should include a pressure-dependent term; however, this term can be neglected for the solid phase. The Gibbs free energy also contains an entropy contribution; these terms are generally small,[12] and they also tend to cancel when comparing relative free energies.[19] The assumption of thermodynamic equilibrium, which underlies the formalism, is expected to be satisfied, particularly in light of the high mobility of various defects and impurities studied here.[12]

Before we give a general definition of the formation energy of an impurity or defect in the compound semiconductor, we illustrate the concept with the example of a Li atom on a substitutional Zn site:

$$E_{\text{form}}(\text{Li}_{\text{Zn}}^{-}) = \mathcal{E}(\text{Li}_{\text{Zn}}^{-}) - \mu_{\text{Li}} + \mu_{\text{Zn}} - E_F. \tag{2.2}$$

$\mathcal{E}(\text{Li}_{\text{Zn}}^{-})$ is the calculated energy of a supercell containing the $\text{Li}_{\text{Zn}}^{-}$ impurity, minus the energy of a reference cell containing the pure bulk semiconductor. These energies are obtained from first-principles calculations, which are described in Sec. II A. The other terms in Eq. (2.2) contain chemical potentials, the physical significance of which we now discuss in some detail.

$\mu_{\text{Li}}$ is the chemical potential of Li. This term enters because the formation energy is the *difference* between the energy of Li as an impurity, and its energy in a reference state. The reference corresponds to a reservoir of Li atoms, whose energy (at $T$=0) by definition is the chemical potential. This chemical potential depends on the abundance of Li under the relevant growth conditions.[20] For an element in thermal equilibrium with the gas phase, the chemical potential can be related to the partial pressure of the gas;[20] for an ideal gas with partial pressure

$p$ one has $\mu = \mu^0 - kT \ln p$. In the literature one often finds studies of defect concentrations as a function of partial pressures. We prefer to work with chemical potentials for the following reasons: (a) Chemical potentials are thermodynamically defined as energy values, which can be directly related to the energies which we calculate from first principles. (b) Although the assumption of thermodynamic equilibrium is likely to be satisfied within the solid, allowing the use of expressions such as Eq. (2.1), it is uncertain to what extent equilibrium is established between the solid and a surrounding gas under experimental conditions such as molecular-beam epitaxy (MBE). Knowledge of the chemical potential in the gas may therefore not necessarily reflect the relevant chemical potential for the solid. (c) Even if thermodynamic equilibrium with the gas is assumed, the relationship between chemical potential and gas pressure is not well known since the gas sources used in MBE do not obey simple ideal gas laws. While this precludes a quantitative determination of chemical potentials in terms of experimentally accessible quantities, we will see that the chemical potentials are subject to rigorous bounds that can be directly related to experimental conditions.

The Zn chemical potential $\mu_{Zn}$ appears in Eq. (2.2) because, in order to make room for the substitutional impurity, a Zn atom has to be removed to its reservoir. It is very important to realize that $\mu_{Zn}$ should be treated as a variable; indeed, in a compound semiconductor only the *sum* of the chemical potentials of the constituents is fixed, and equal (at $T$=0) to the energy of a two-atom unit of the material:

$$\mu_{Zn} + \mu_{Se} = \mu_{ZnSe}. \tag{2.3}$$

In an elementary semiconductor, this condition would uniquely determine the value of the chemical potential; additional freedom exists, however, in a compound semiconductor. We will therefore explicitly present our results as a function of chemical potentials. Equation (2.3) fixes $\mu_{Se}$ once $\mu_{Zn}$ is chosen; alternatively, $\mu_{Se}$ could be chosen as the free variable, leading to a fixed $\mu_{Zn}$.

The last term in Eq. (2.2) is the Fermi level $E_F$, i.e., the energy of the reservoir delivering the electron responsible for the negative charge on the impurity.

In general the total energy $E_{tot}(D_i)$ for a defect $D_i$ will be determined from a calculation for a supercell containing $n_i^{Zn}$ Zn atoms, $n_i^{Se}$ Se atoms, and $n_i^{Li}$ Li atoms (we continue to use Li as a sample impurity, but the formulas are valid for a general impurity). The defect formation energy $E_{form}(D_i)$ is then

$$\begin{aligned} E_{form}(D_i) &= E_{tot}(D_i) - n_i^{Zn}\mu_{Zn} - n_i^{Se}\mu_{Se} \\ &\quad - n_i^{Li}\mu_{Li} - n_i^e E_F \\ &= \mathcal{E}(D_i) - \Delta n_i \mu_{Zn} - n_i^{Li}\mu_{Li} - n_i^e E_F, \end{aligned} \tag{2.4}$$

$$\mathcal{E}(D_i) = E_{tot}(D_i) - n_i^{Se}\mu_{ZnSe}, \tag{2.5}$$

$$\Delta n_i = n_i^{Zn} - n_i^{Se}, \tag{2.6}$$

where $n_i^e$ is the number of excess electrons in the defect, and $\Delta n_i$ is the number of extra Zn atoms that must be added to form the defect (e.g., +1 for $Zn_i$ and $V_{Se}$, −2 for $Se_{Zn}$, etc.). Here, we treat $\mu_{Zn}$ as an independent variable and use Eq. (2.3) to remove $\mu_{Se}$ from the expression for $E_{form}(D_i)$; alternatively, we could treat $\mu_{Se}$ as independent and eliminate $\mu_{Zn}$.

### C. Self-consistent solution

An expression based on Eq. (2.4) can be written down for all configurations of the impurity, in their various charge states, as well as for all native defects. Once the formation energy is known, the concentration of a specific defect or impurity can be obtained from Eq. (2.1). At this point, all concentrations are still functions of the chemical potentials ($\mu_{Zn}$ and $\mu_{Li}$), as well as of the Fermi level ($E_F$). The chemical potentials, as explained above, are independent parameters; we will therefore express all our results as functions of these chemical potentials. The Fermi level, however, is *not* an independent variable, since it is determined by the condition of charge neutrality:

$$\text{net charge} = 0 = p - n - \sum_i n_i^e [D_i], \tag{2.7}$$

where $p$ and $n$ are the hole and electron densities, respectively. These free-carrier densities are determined from the standard semiconductor equations. The charge conservation equation provides for an interaction between the concentrations of all charged defects through their influence on the Fermi level. For example, a positively charged defect produces extra free electrons that raise the Fermi level; the higher Fermi level, in turn, increases the concentrations of all negatively charged defects and lowers the concentrations of all positively charged defects. As pointed out by Zhang and Northrup,[21] this "negative feedback" reduces the sensitivity of the final results to possible inaccuracies in our first-principles energies. Using this prescription, all of the defect formation energies, and hence the concentrations $[D_i]$, are unique functions of $\mu_{Zn}$, $\mu_{Li}$, and the temperature $T$.

The choice of the chemical potential $\mu_{Zn}$ also determines the stoichiometry; the stoichiometry parameter $X$ can be defined as

$$X = \frac{N_{Se} - N_{Zn}}{N_{Se} + N_{Zn}} = \frac{-\sum_i \Delta n_i [D_i]}{2N_{sites}}, \tag{2.8}$$

where $N_{Zn}$ and $N_{Se}$ are the total numbers of Zn and Se atoms in the crystal. Only deviations from stoichiometry due to native defects are included here. $X$ is positive for Se-rich material and negative for Zn-rich material. In this paper, we express all our results in terms of chemical potentials. Alternatively, we could present the results as a function of the stoichiometry parameter, but because of the one-to-one correspondence between chemical potential and stoichiometry no new information would be obtained.

### D. Bounds on the chemical potentials

Now we discuss how the relevant range of the chemical potentials is determined. For this purpose one has to consider the various phases that can be formed out

of the constituents.[19,20] For instance, $\mu_{Zn}$ is bounded from above by the energy of a Zn atom in Zn metal: $\mu_{Zn}^{max} = \mu_{Zn(bulk)}$. Indeed, if one would try to raise $\mu_{Zn}$ above this level, Zn metal would be preferentially formed. Similarly, $\mu_{Se}$ has an upper bound imposed by bulk Se. Furthermore,

$$\mu_{ZnSe} = \mu_{Zn(bulk)} + \mu_{Se(bulk)} + \Delta H_f(ZnSe), \quad (2.9)$$

where $\Delta H_f$(ZnSe) is the heat of formation of ZnSe ($\Delta H_f$ is negative for a stable compound). Combined with Eq. (2.3) this expression can be used to impose a *lower* bound on th' Zn chemical potential, given by $\mu_{Zn}^{min} = \mu_{Zn(bulk)} + \Delta H_f$ (ZnSe). A lucid discussion of similar arguments, in the context of surface recontructions, has been given in Ref. 19. The Zn chemical potential can thus vary over a range corresponding to the heat of formation of ZnSe.

To find an upper bound on the chemical potential of the dopant we explore the various compounds that the impurity can form in its interactions with the system. For Li, a possible upper bound on $\mu_{Li}$ is of course imposed by Li (bulk) metal. However, the most stringent constraint arises from the compound $Li_2Se$, which leads to the following constraint on the chemical potentials:

$$2\mu_{Li} + \mu_{Se} = \mu_{Li_2Se} = 2\mu_{Li(bulk)} + \mu_{Se(bulk)} + \Delta H_f(Li_2Se). \quad (2.10)$$

Numerical results for the heats of formation, as well as practical applications of the bounds on the chemical potentials, will be given in the following section.

## III. RESULTS AND DISCUSSION

### A. Lithium

#### 1. Configurations of Li in the lattice

We have analyzed various possible configurations and charge states of the lithium impurity in the lattice. The substitutional acceptor $Li_{Zn}^{-}$ induces virtually no relaxation of the surrounding host atoms. For the lithium interstitial ($Li_i^+$), which is a shallow donor, we find the $T_d$ site surrounded by Se atoms ($T_d^{Se}$) to be 0.2 eV lower in energy than the $T_d^{Zn}$ site. For the interstitials, the energy gained by relaxation of the host atoms is smaller than 0.1 eV. We have also studied other interstitial positions, allowing us to estimate that the barrier for migration of the interstitial is less than 0.5 eV (i.e., a Li interstitial can move readily, even at room temperature). Finally, we have also investigated Li on a substitutional Se site, but found this configuration to have a prohibitively large formation energy.

#### 2. Contour plots of total Li concentration

Our results are presented in the form of contour plots, which allow us to explicitly show the dependence on the chemical potentials $\mu_{Zn}$ and $\mu_{Li}$. As explained in Sec. II C, there is no explicit dependence on Fermi energy, since it is determined by charge neutrality. Figure 1(a) shows a contour plot for the total concentration of Li in ZnSe, at $T = 600$ K, which is a typical temperature in MBE growth of ZnSe:Li.[2,4]

We first discuss the contour lines themselves. The total Li concentration ([Li]) increases with increasing $\mu_{Li}$, because it becomes more favorable for the impurity to dissolve in the semiconductor as the energy of the reservoir rises. Similarly, [Li] increases with *decreasing* $\mu_{Zn}$, which is the energy of the reservoir to which Zn needs to be removed in order to accommodate Li on Zn sites.

#### 3. Competition between interstitials and substitutionals

The formation energy for Li in a substitutional location was given in Eq. (2.2). For the interstitial site, where Li is a shallow donor, we have

$$E_{form}(Li_i^+) = \mathcal{E}(Li_i^+) - \mu_{Li} + E_F. \quad (3.1)$$

$\mathcal{E}(Li_i^+)$ is the calculated energy of an interstitial Li at its most stable site, which is at the tetrahedral interstitial site surrounded by Se atoms. Inspection of Eqs. (2.2) and (3.1) reveals that as the Fermi level moves down (i.e., as the material becomes increasingly $p$-type), the formation energy of the acceptor species rises, whereas the formation energy of the donor species goes down. This predicts the existence of a limiting Fermi-level position (maximum hole concentration), which can be obtained by equating the two formation energies. Attempts to push the Fermi level lower would result in preferential formation of donors, which would push the Fermi level back up. Incorporation of additional Li leaves the Fermi level unchanged, as each substitutional acceptor is immediately compensated by an interstitial donor.

The position of the Fermi level (at 600 K) is shown in Fig. 1(b); Li interstitials are responsible for the flattening of the contour lines on the right-hand side of the plot. For a fixed value of $\mu_{Zn}$, the Fermi level saturates as $\mu_{Li}$ is raised, even though the total Li concentration still increases [see Fig. 1(a)]. If no interstitials could form, the contour lines would continue to rise with the same slope as in the left-hand side of the plot. The interstitials cause compensation and limit the achievable hole concentration.[7] Their presence has been experimentally observed.[22,23] A contour plot of the Li interstitial concentration is shown in Fig. 1(c).

The position at which the Fermi level saturates due to interstitial compensation still depends on the Zn chemical potential, as can be noted in Fig. 1. Our results differ markedly from those of Ref. 8, where it was concluded that compensation by Li interstitials would always dominate. The authors of Ref. 8 did not recognize that the level of compensation depends on the Zn chemical potential, and hence on the growth conditions. This dependence explains the experimental observation that the degree of compensation by Li interstitials varies widely in different samples.[2] Our results actually provide a guideline for optimizing the growth conditions: low values of $\mu_{Zn}$ lead to lower compensation, as well as higher $Li_{Zn}$ concentrations.

*4. Bounds on chemical potentials — solubilities*

In order to determine solubility limits, we need to use the information about bounds on the chemical potentials discussed in Sec. II D. The bounds on the Zn chemical potential are shown as the horizontal lines in Fig. 1. For Li, the chemical potential is limited by formation of the compound $Li_2Se$. Formation of $Li_2Se$ on the growing ZnSe surface in MBE has actually been experimentally observed in the case of heavy Li doping.[24] The compound $Li_2Se$ leads to the line with slope +2 in Fig. 1, which was defined in Eq. (2.10). The point where this line intersects the lower bound on $\mu_{Zn}$ is given by $\mu_{Li}^0 = \mu_{Li(bulk)} + \frac{1}{2}\Delta H_f(Li_2Se)$. Our calculated heats of formation for the various compounds are listed in Table I. For comparison, we also list experimental values. The deviations are in line with the expected accuracy of the method.

Our calculated contours, together with the bounds on the chemical potentials, provide important insights in the ability to dope ZnSe with Li. We note that, over much of the range of the Li and Zn potentials, the maximum Li concentration is slightly higher than $10^{18}$ cm$^{-3}$. The fact that the slope of the contours in this region coincides with the slope of the $Li_2Se$ boundary in Fig. 1(a) is accidental, caused by the fact that in this region the removal of *one* Zn atom leads to the incorporation of *two* Li atoms (one substitutional and one interstitial). The highest Li concentration (and lowest Fermi level, i.e., highest hole concentration) occurs in the lower right-hand corner of the accessible region, for $\mu_{Zn} = \mu_{Zn}^{min}$ and $\mu_{Li} = \mu_{Li}^0$. The

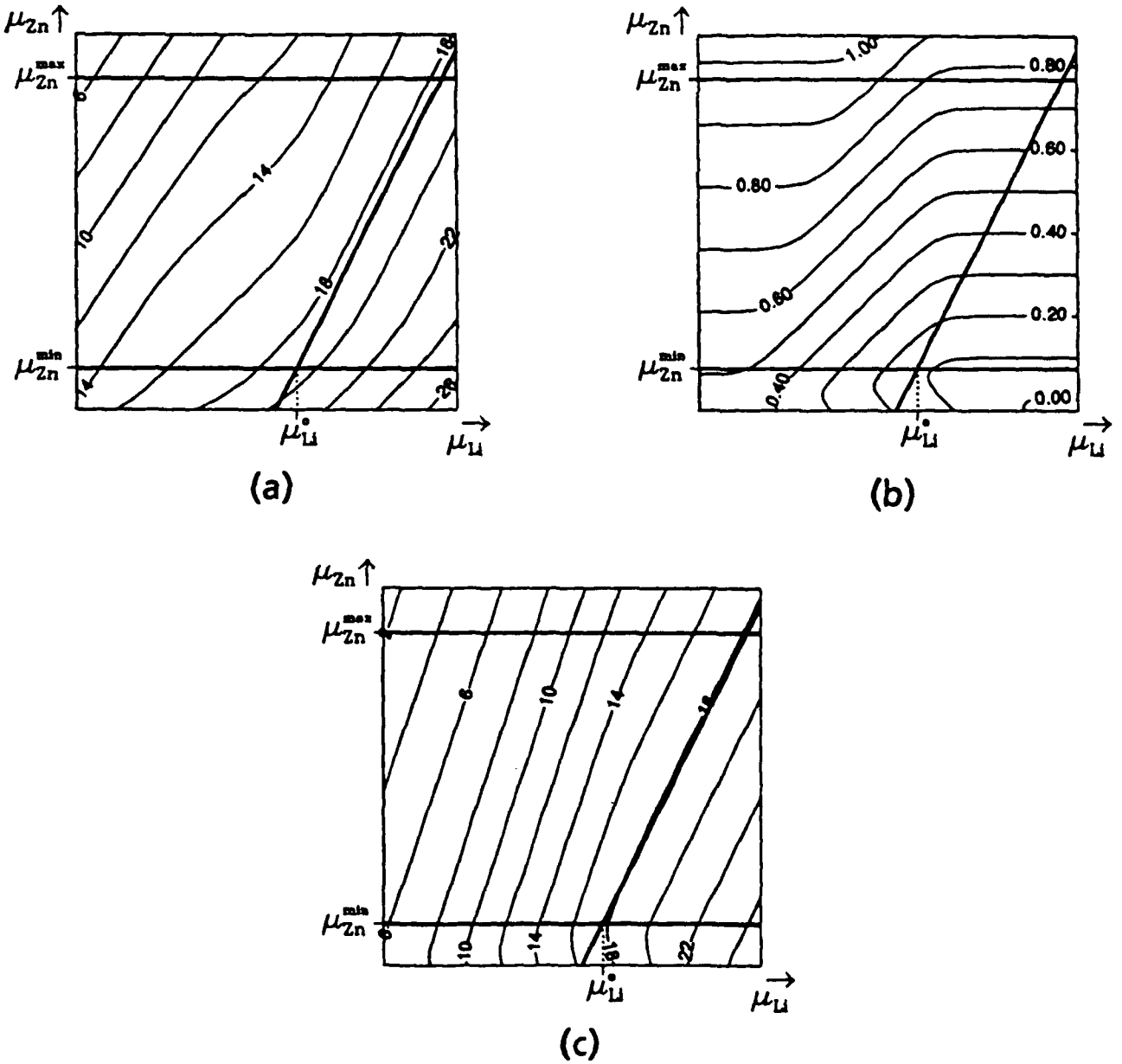



FIG. 1. Contour plots of (a) $\log_{10}$ [Li], where [Li] is the total Li concentration in cm$^{-3}$, (b) Fermi level (in eV, referred to the top of the valence band), and (c) $\log_{10}$ [$Li_i$], where [$Li_i$] is the interstitial Li concentration in cm$^{-3}$, at 600 K in ZnSe:Li, as a function of Zn and Li chemical potentials. Solid lines indicate bounds on $\mu_{Zn}$ and $\mu_{Li}$.

TABLE I. Theoretical and experimental (Ref. 25) heats of formation (in eV per formula unit) for various materials containing Zn, Se, Li, Na, and N. Also listed is the minimum formation energy for the neutral substitutional acceptor in ZnSe, and the corresponding minimum Fermi-level position (in eV, referred to the top of the valence band), at 600 K.

| | Solubility-limiting compound | $\Delta H_f^{theor}$ | $\Delta H_f^{expt}$ | $E_{form}^{min}$ | $E_F$ |
|---|---|---|---|---|---|
| ZnSe | | -1.39 | -1.69 | | |
| ZnSe:Li | $Li_2Se$ | -4.12 | -3.96 | 0.46 | 0.13 |
| ZnSe:Na | $Na_2Se$ | -3.13 | -3.54 | 1.08 | 0.44 |
| ZnSe:N | $Zn_3N_2$ | | -0.24 | 0.38 | 0.09 |

corresponding formation energy of the *neutral* acceptor, and the self-consistently determined Fermi level are also listed in Table I. At this point of highest Li incorporation, the total Li concentration is $1.7 \times 10^{19}$ cm$^{-3}$; fewer than 3% of these Li atoms occur in the form of interstitials.

### 5. Discussion

Our calculated differences in formation energies and heats of formation have an estimated error margin of ±0.1 eV. At a temperature of 600 K, 0.12 eV roughly corresponds to an order of magnitude in concentration. Also, contours with values of [Li] higher than $10^{19}$ cm$^{-3}$ are probably inaccurate because Eq. (2.1) is only valid for dilute concentrations; however, these contours fall outside the physically accessible range anyway. While these uncertainties should be kept in mind when considering plots such as Fig. 1, the qualitative and even quantitative insights are still clear. Some additional conclusions can be drawn. First, even though all native point defects were explicitly included in the calculations, their concentrations are very small over the whole of the accessible range in Fig. 1. The effect of native defects is noticeable for low $\mu_{Zn}$ values, causing bending of the contour lines; however, their concentration would only become important if $\mu_{Zn} < \mu_{Zn}^{min}$, which is physically not allowed. The dominant native defect is the $Se_{Zn}$ antisite, which is a donor. Figure 2 shows a contour plot of the $Se_{Zn}^{2+}$ concentration. At the point of highest Li incorporation, the concentration is $[Se_{Zn}]=2.6 \times 10^{17}$ cm$^{-3}$, which is two orders of magnitude smaller than the Li concentration. Clearly the native defect concentration is too low to play any significant role in compensation. However, the concentration may be high enough to be detectable experimentally. Other native defects have concentrations significantly smaller (by more than four orders of magnitude) than the $Se_{Zn}$ antisite.

The contour plots presented here were made for a temperature of 600 K, which is typical for MBE growth of ZnSe. The qualitative features of the plots do not change when we change the temperature (within physically reasonable limits). To illustrate the quantitative effect of temperature changes, as we lower the temperature from 600 to 500 K, we find that the total Li concentration is reduced by a factor of 5; the concentration of interstitial Li drops by more than an order of magnitude; and the concentration of the dominant native defect ($Se_{Zn}$) drops by almost two orders of magnitude.

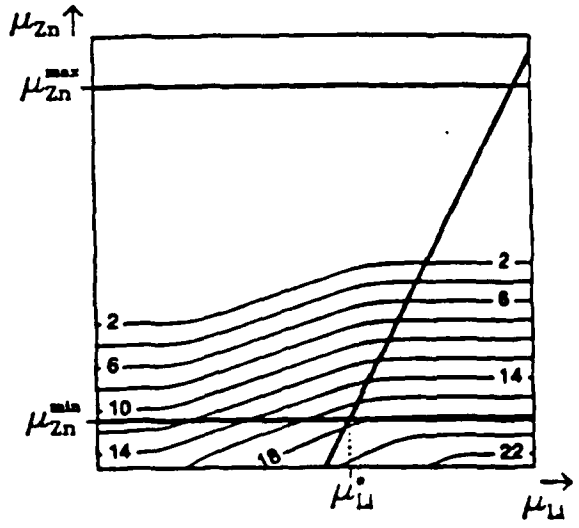



FIG. 2. Contour plot of $\log_{10}[Se_{Zn}^{2+}]$, the Se antisite concentration in cm$^{-3}$, at 600 K in ZnSe:Li, as a function of Zn and Li chemical potentials. Solid lines indicate bounds on $\mu_{Zn}$ and $\mu_{Li}$.

A final point relates to doping of ZnSSe alloys with Li (alloys containing 6% S are commonly used to obtain *lattice matching with GaAs substrates): since $Li_2S$ is* even more stable than $Li_2Se$ (larger $|\Delta H_f|$), the bound on $\mu_{Li}$ in the ZnSSe:Li system will lie even lower, leading to reduced solubility in the alloy.

### 6. Complex formation

So far we have only talked about *isolated* point defects and impurities. In principle we should also consider complexes. Although our formalism is general enough to include any possible complexes, an exhaustive treatment is computationally prohibitive. Inspection of expressions for formation energies actually shows that a complex will only occur in appreciable concentrations (i.e., concentrations on the order of or larger than those of the individual defects out of which it is formed) if the binding energy exceeds the larger of the two formation energies of the individual components of the complex. This consideration makes it less likely that complexes would play an important role.

The only complex we have investigated as part of the current study is one consisting of a Li interstitial and a Li substitutional.[26] Formation of such complexes seems plausible, since the interstitial is quite mobile, and the acceptor and donor are Coulombically attracted. Details of the structure will be published elsewhere.[27] The binding energy of this complex is ~0.3 eV. This value is small enough so that these complexes are largely dissociated at a growth temperature of 600 K (in other words, their concentration is small compared to the concentration of the individual components, as discussed in the preceding paragraph). If we assume, however, that the con-

centration of Li substitutional and Li interstitial atoms is determined at the growth temperature, and remains fixed as the sample is cooled down, then the concentration of $Li_{Zn}$-$Li_i$ pairs will increase as the temperature is lowered. The presence of such complexes should be taken into account in analyses of Fermi level positions and carrier concentrations at room temperature and below.[23,27]

### B. Sodium

We now address Na, another column-I impurity which has been considered as an acceptor dopant in ZnSe.[10] The contour plots for the ZnSe:Na system are shown in Fig. 3. They are qualitatively similar to those in Fig. 1, but exhibit important quantitative differences. The relevant bound on the Na chemical potential is imposed by the compound $Na_2Se$. The most important result is that the solubility of substitutional Na is significantly lower than that of Li — the maximum concentration obtained from the contour plot is lower than $10^{16}$ cm$^{-3}$. At these lower concentrations, very few Na interstitials are present; we also find that the barrier for migration of the Na interstitial is much higher than for $Li_i$. Experimental doping attempts with Na have been unsuccessful;[10] our results clearly show that the solubility limit is the culprit, rather than, e.g., compensation due to foreign impurities in the source.

### C. Nitrogen

Finally, we discuss N in ZnSe. Nitrogen on a substitutional Se site ($N_{Se}$) is a shallow acceptor. The surround-

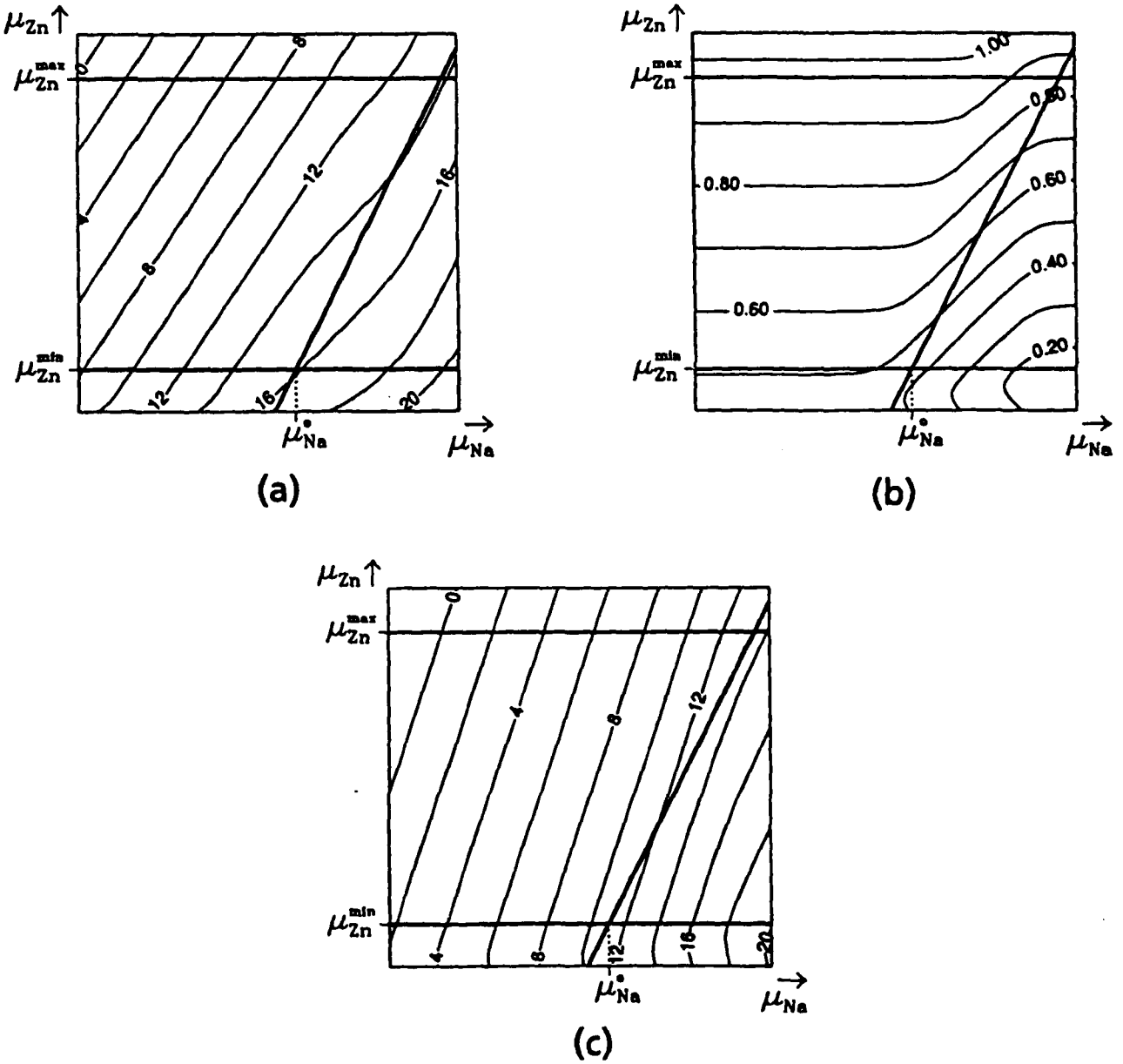



FIG. 3. Contour plots of (a) $\log_{10}$ [Na], where [Na] is the total Na concentration in cm$^{-3}$, (b) Fermi level (in eV, referred to the top of the valence band), and (c) $\log_{10}$ [$Na_i$], where [$Na_i$] is the interstitial Na concentration in cm$^{-3}$, at 600 K in ZnSe:Na, as a function of Zn and Na chemical potentials. Solid lines indicate bounds on $\mu_{Zn}$ and $\mu_{Na}$.

ing Zn atoms undergo a significant inward relaxation, reducing the Zn-N distance to 2.1Å. This distance is very close to the Zn-N distance in the compound $Zn_3N_2$.[28] We have also investigated other configurations, such as the substitutional Zn site and interstitial sites, and found those to be much higher in energy than the substitutional Se site. Thus, N does not suffer from the substitutional/interstitial competition associated with the column-I elements, so that the saturation of the Fermi level which we observed in Fig. 1(b) does not occur here. In this work, we have not investigated any relaxation of the impurity away from the ideal lattice site.[29] According to Ref. 29, in the case of N such relaxations would not interfere with the shallow acceptor character of the dopant; if any relaxations do occur, they would therefore simply lead to a lower formation energy (and hence enhanced concentration) of the shallow acceptor state.

Two bounds on the N chemical potential arise in this case: $N_2$ molecules and the $Zn_3N_2$ compound. The compound $Zn_3N_2$ has the bixbyite structure,[28] which contains 80 atoms in the unit cell. This exceeds the capabilities of state-of-the-art first-principles calculations; we have therefore resorted to calculating a higher-symmetry structure, whose energy closely approximates that of the real compound. With regard to the other bound, our application of $N_2$ molecules as a solubility-limiting phase does not imply that we assume equilibrium between ZnSe:N and $N_2$ gas outside. Rather, we envision formation of some condensed phase involving $N_2$, such as in a void or in a chemisorbed state. Because of the difficulty in obtaining converged results for the $N_2$ molecule with an acceptable basis set, the energy difference between $N_2$ and $Zn_3N_2$ was taken from experiment.

Our results are displayed in Fig. 4. The bending of the contour lines in the upper part of Fig. 4(b) is due to native defects. Indeed, the N concentration is very low here [less than $10^{12}$ cm$^{-3}$; see Fig. 4(a)], and a small concentration of native defects suffices to pin the Fermi level. However, native defects play only a minor role if the right conditions (chemical potentials) are present for high N dopant concentrations. At the point of highest N incorporation, the calculated N concentration is $6.4\times10^{19}$ cm$^{-3}$.

The native defect concentration once again increases as we approach the lower end of the accessible region (low $\mu_{Se}$, i.e., Zn-rich conditions). The dominant native defect is the Zn interstitial; its concentration as a function of chemical potentials is shown in Fig. 5. The compensation due to this native defect is still small enough not to pose any threat to the doping. We have verified that this conclusion remains true even if our calculated formation energy for the native defect would be off by several 0.1 eV. The reason the results are not very sensitive to such inaccuracies is the "negative feedback" mechanism discussed in Sec. II C, acting through the coupling of all defect and impurity concentrations via the charge neutrality condition. In addition, the $Zn_i$ concentration falls off rapidly (faster than the N concentration) as the Se chemical potential is raised, away from its lower bound. Other native defects have concentrations four orders of magnitude smaller than the Zn interstitial. Although our calculations indicate Zn interstitials should be present in N-doped samples in concentrations high enough for experimental observation, other factors have to be taken into account. One such factor is the high mobility of the Zn interstitial,[30] which may cause it to move into the substrate or towards the surface. It is also conceivable that Zn interstitials (donors) would form complexes with substitutional N acceptors.

Once again, we have investigated the effect of temperature on our results. Lowering the temperature from 600 to 500 K decreases the total N concentration by a factor of 4; simultaneously, the concentration of Zn interstitials drops by a factor of 20.

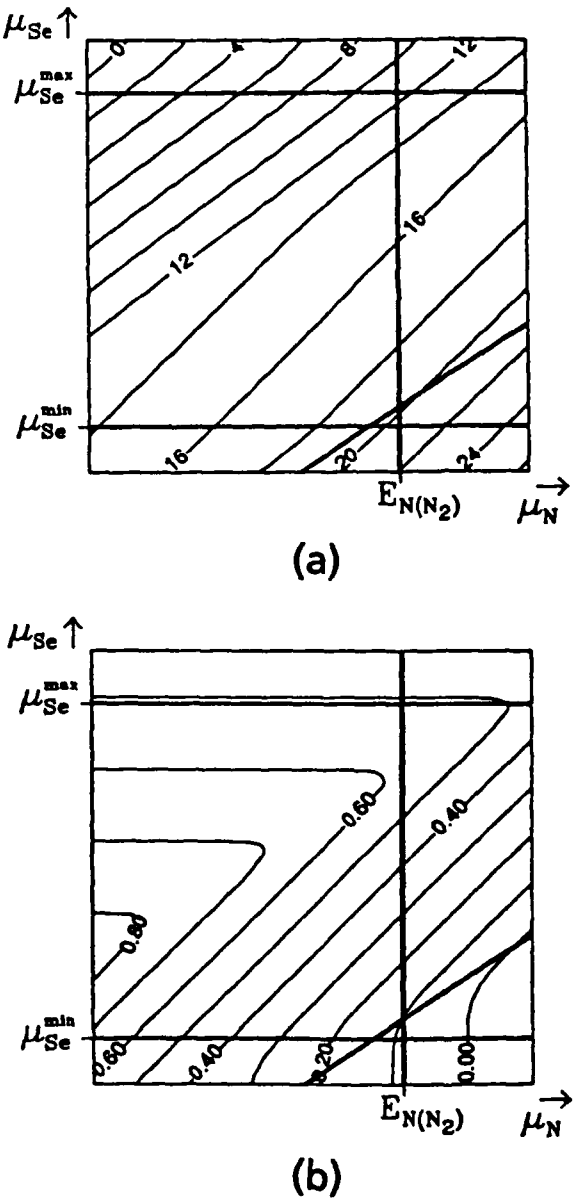



FIG. 4. Contour plots of (a) $\log_{10}$ [N], where [N] is the total N concentration in cm$^{-3}$, and (b) Fermi level (in eV, referred to the top of the valence band) at 600 K in ZnSe:N. Since N is substitutional on a Se site, $\mu_{Se}$ (rather than $\mu_{Zn}$) is chosen as the variable here. Solid lines indicate bounds on $\mu_{Se}$ and $\mu_N$.

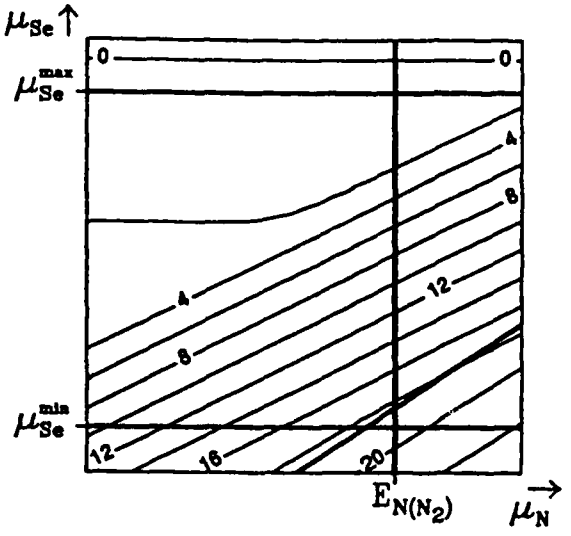



FIG. 5. Contour plot of $\log_{10}[Zn_i^{2+}]$, the Zn interstitial concentration in $cm^{-3}$, at 600 K in ZnSe:N, as a function of Se and Li chemical potentials. Solid lines indicate bounds on $\mu_{Se}$ and $\mu_N$.

A comparison with Fig. 1 show that N has a solubility significantly higher than Li, which is consistent with experimental results. The failure of nitrogen doping starting from $N_2$ is due to the large kinetic barrier for breaking up the molecule; a plasma source or other technique for obtaining N in an atomic state, or at least $N_2$ in an excited state, is required.[5] Once one succeeds in incorporating atomic (as opposed to molecular) nitrogen into the lattice, N should act as a good acceptor, allowing hole concentrations high enough for useful device applications.

## IV. SUMMARY

We have presented a formalism that enables us to calculate impurity concentrations and doping levels in semiconductors. The technologically important case of acceptor doping in ZnSe was discussed in detail; however, the formalism is quite general in nature and can be applied to any semiconductor and any impurity for which reliable first-principles calculations can be carried out. The computed total energies of impurities and defects allow us to write down formation energies as a function of the atomic chemical potentials and of the Fermi level; the latter is then determined by imposing charge neutrality.

The results are presented in the form of contour plots, which reflect the dependence on chemical potentials. Although the latter are free parameters, which vary with the growth conditions, they are subject to thermodynamic bounds corresponding to formation of other phases (e.g., formation of $Li_2Se$ in the case of ZnSe:Li). Imposing these bounds determines the maximum achievable impurity incorporation. In addition, our results provide insight in how variations in growth conditions can promote incorporation of the dopant in the desirable configuration.

For acceptors in ZnSe, we have reached the following conclusions: Although Li suffers from a competition between interstitial and substitutional configurations, appropriate growth conditions can be chosen to suppress interstitial formation. The limited solubility of Li (imposed by formation of $Li_2Se$) is a more severe obstacle to the success of Li as a $p$-type dopant. Sodium suffers from this type of solubility problem to an even greater extent. Nitrogen, finally, emerges as the best choice among the dopants examined here.

## ACKNOWLEDGMENTS

Part of this work was carried out while one of us (C.G.V.d.W.) was at Philips Laboratories, Briarcliff Manor, NY. We are grateful to D. A. Cammack, J. Gaines, F. Greidanus, T. Marshall, R. M. Martin, and J. Tersoff for suggestions and discussions. We are indebted to D. Vanderbilt for his iterative diagonalization program. This work was supported in part by ONR Contract No. N00014-84-0396.

*Present address: IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

[1] R. N. Bhargava, J. Cryst. Growth 86, 873 (1988).
[2] M. A. Haase, H. Cheng, J. M. DePuydt, and J. E. Potts, J. Appl. Phys. 67, 448 (1990).
[3] J. Ren, K. A. Bowers, B. Sneed, D. L. Dreifus, J. W. Cook, Jr., J. F. Schetzina, and R. M. Kolbas, Appl. Phys. Lett. 57, 1901 (1990).
[4] T. Marshall and D. A. Cammack, J. Appl. Phys. 69, 4149 (1991).
[5] R. M. Park, M. B. Troffer, C. M. Rouleau, J. M. DePuydt, and M. A. Haase, Appl. Phys. Lett. 57, 2127 (1990).
[6] M. A. Haase, J. Qiu, J. M. DePuydt, and H. Cheng, Appl. Phys. Lett. 59, 1272 (1991).
[7] G. F. Neumark, J. Appl. Phys. 51, 3383 (1980).
[8] T. Sasaki, T. Oguchi, and H. Katayama-Yoshida, Phys. Rev. B 43, 9362 (1991).
[9] Recently, a similar phenomenon has been experimentally observed in Zn-doped InP, where the solubility is limited by formation of $Zn_3P_2$; L. Y. Chan, K. M. Yu, M. Ben-Tzur, E. E. Haller, J. M. Jaklevic, W. Walukiewicz, and C. M. Hanson, J. Appl. Phys. 69, 2998 (1991).
[10] H. Cheng, J. M. DePuydt, J. E. Potts, and M. A. Haase, J. Cryst. Growth 95, 512 (1989).
[11] G. Mandel, Phys. Rev. 134, A1073 (1964).
[12] D. B. Laks, C. G. Van de Walle, G. F. Neumark, and S. T. Pantelides, Phys. Rev. Lett. 66, 648 (1991).
[13] P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964); W. Kohn and L. J. Sham, *ibid.* 140, A1133 (1965).
[14] D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. 43, 1494 (1979).
[15] S. G. Louie, S. Froyen, and M. L. Cohen, Phys. Rev. B 26, 1738 (1982).
[16] D. B. Laks, C. G. Van de Walle, G. F. Neumark, and S. T. Pantelides, Phys. Rev. B 45, 10965 (1992).
[17] C. G. Van de Walle, P. J. H. Denteneer, Y. Bar-Yam, and

S. T. Pantelides, Phys. Rev. B 39, 10791 (1989).
[18]Activation energies: Li, 114 meV [J. L. Merz, K. Nassau, and J. W. Shiever, Phys. Rev. B 8, 1444 (1973)]; Na, 128 meV [H. Tews, H. Venghaus, and P. J. Dean, *ibid.* 19, 5178 (1979)]; N, 110 meV [K. Shahzad, B. A. Khan, D. J. Olego, and D. A. Cammack, *ibid.* 42, 11 240 (1990)].
[19]G.-X. Qian, R. M. Martin, and D. J. Chadi, Phys. Rev. B 38, 7649 (1988); N. Chetty and R. M. Martin, *ibid.* 45, 6089 (1992).
[20]F. A. Kröger, *The Chemistry of Imperfect Crystals* (North-Holland, Amsterdam, 1964), pp. 136 and 628.
[21]S. B. Zhang and J. E. Northrup, Phys. Rev. Lett. 67, 2339 (1991).
[22]M. A. Haase, J. M. DePuydt, H. Cheng, and J. E. Potts, Appl. Phys. Lett. 58, 1173 (1991).
[23]T. Marshall, in *Proceedings of the 7th Trieste Semiconductor Symposium on Wide-Band-Gap Semiconductors*, edited by C. G. Van de Walle (North-Holland, Amsterdam, in press).
[24]Z. Zhu, H. Mori, M. Kawashima, and T. Yao, J. Cryst. Growth 117, 400 (1992).
[25]*Lange's Handbook of Chemistry*, 12th ed., edited by J. A. Dean (McGraw-Hill, New York, 1979).
[26]G. F. Neumark and C. R. A. Catlow, J. Phys. C 17, 6087 (1984).
[27]C. G. Van de Walle (unpublished).
[28]R. W. G. Wyckoff, *Crystal Structures*, 2nd ed. (Interscience, New York, 1964), Vol. 2.
[29]D. J. Chadi and K. J. Chang, Appl. Phys. Lett. 55, 575 (1989); D. J. Chadi, *ibid.* 59, 3589 (1991).
[30]G. D. Watkins, in *Defect Control in Semiconductors*, edited by K. Sumino (Elsevier Science, Amsterdam, 1990), p. 933.

# Forces on Charged Defects in Semiconductor Heterostructures

J. Tersoff
*IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 9 April 1990)

The forces on dopant impurities and other defects are crucial in determining defect motion and diffusion in semiconductor heterostructures. Impurity ions and other charged defects feel electrostatic forces, just as electrons do. However, at heterojunction interfaces, electrons feel additional forces associated with band-edge discontinuities. Here, the analogous forces for ions and charged defects are derived, and given a simple physical interpretation. The donor or acceptor level is found to play much the same role for the ion that the band edge plays for the electron or hole in defining the effective potential. These forces can in general be spatially discontinuous, because of their dependence on charge state.

PACS numbers: 66.30.Jt, 66.30.Lw, 68.35.Fx

In the last decade, there has been an explosion of interest in semiconductor heterostructures and other compositionally modulated structures, because of their usefulness in electronic devices. Originally, attention centered on the properties of electrons and holes in such structures; the central theoretical problem was then the determination of the band-edge discontinuities at semiconductor interfaces.[1] Now that such devices are a reality, however, attention is turning to the factors which determine the growth and stability of these structures.

Diffusion is often a limiting factor in the design and manufacture of heterostructures, especially diffusion of dopants and other impurities.[2] Impurity diffusion is of great interest in its own right, and has been the focus of intense theoretical study.[3] But in the context of a heterostructure, such diffusion raises a new theoretical issue, which is closely analogous to the band offset problem for electrons. Specifically, one needs to determine what forces act on charged defects at semiconductor interfaces, or in compositionally graded structures. Here, we derive a simple and intuitive solution to this problem.

For a neutral impurity, the only driving force (other than entropic) would be the gradient of the enthalpy;[2] so the enthalpy, which depends on position through the local composition and strain, plays the role of a potential. For a charged impurity, as for an electron, there is an additional driving force, the electric field. But in addition, at a semiconductor interface, an electron or hole sees an abrupt change in the potential, corresponding to the conduction-band or valence-band discontinuity. (Throughout, the word "potential" refers to potential energy, not electrostatic potential, unless specifically stated.) The question, then, is what force the ion sees at the interface.

Intuitively, one might consider that there is some electrostatic dipole at the interface, and that the ion will see a force determined by the product of its charge and this dipole.[4] However, it is now generally appreciated[1] that it is impossible, even in principle, to uniquely define the dipole at a heterojunction (except relative to an arbitrary reference interface). Our goal, then, is to define an effective local potential for an ion, or for any charged defect, referring only to physical observables. In particular, we must evaluate the energy required for a charged defect to cross an interface, without any explicit reference to a dipole at the interface.

To do this, we can imagine moving the ion through the material in two steps: moving the *neutral* defect or impurity, and then ionizing it and returning the electron or hole to the starting position. In this way we reduce the problem to two well-understood problems: the motion of a neutral impurity or defect, and the motion of an electron or hole.

Consider a donor $D$, which could be a substitutional impurity such as a dopant, or more generally a vacancy, an interstitial, or any other sort of defect. We begin with the defect in its thermodynamic reference state, $D_{\rm ref}$, and then consider inserting it into the semiconductor in its neutral charge state, ionizing it, and so forth, writing a sequence of conversions which conserve energy:

$$\begin{aligned} D^0_{\rm ref} &\rightarrow D^0_A - H^0_A \\ &\rightarrow D^+_A - I_A + e_A - H^0_A \\ &\rightarrow D^+_A - I_A + E^c_A - \mu - H^0_A . \end{aligned} \tag{1}$$

Here $H^0_A$ is the enthalpy solution for the neutral impurity in semiconductor $A$, or more generally the enthalpy of formation of the neutral defect from its reference state. In the case of a shallow donor, by the neutral state we mean the donor with the electron bound in its $1s$ hydrogenic orbit. This first step, formation of the defect, typically results in an energy deficit, since the enthalpy of formation is usually positive.

The second step in Eq. (1), ionization, entails a further energy deficit of $I_A$, the ionization energy of the neutral defect in semiconductor $A$. This step yields an electron $e_A$ at the conduction minimum in $A$. The final step makes use of the role of the Fermi level $\mu$ as a reservoir of electrons and holes, to convert the extra electron $e_A$ into a corresponding energy $E^c_A - \mu$, where $E^c_A$ is the

position in energy of the conduction-band minimum of $A$. (As always in the grand-canonical ensemble, particle number and charge are not conserved microscopically, only statistically.)

Of course, a similar equation applies for acceptors, where all of the charges are reversed, the electron is replaced with a hole, and the valence-band maximum replaces the conduction minimum. Double donors and multiple charge states are discussed below.

Note that the subscript $A$ may be taken to fully specify the local semiconductor, including its composition, strain, and local electrostatic potential. Thus just as $H_A^0$ acts as a local potential for the neutral defect, so the corresponding effective potential for the charged donor from (1) may be written

$$V_{\text{eff}}^{+} = H_A^0 + I_A - (E_A^c - \mu). \tag{2}$$

The conduction-band minimum $E^c$ includes the effects of any electrostatic fields associated with band bending, etc., as well as compositional effects (band discontinuities at interfaces). It is important to remember that there are compositional effects in graded devices, just as at abrupt interfaces; in fact, such grading can be viewed as a sequence of interfaces between regions of slightly different composition.

The potential for the ion is seen to be determined entirely by local quantities, i.e., quantities which depend only on the local composition and strain, *except* for the conduction-band minimum. The latter is not a local quantity, in that it depends on the distribution of charges far away; but it is uniquely defined at every point for any system in which the electrons are in equilibrium, i.e., where the Fermi level $\mu$ is well defined.

The only quantities needed to deduce the force from Eq. (2) are the band-edge discontinuity, the defect ionization energy, and the enthalpy of the neutral defect. There are all experimentally measurable quantities, so, in principle, no further theoretical inputs are needed.

We can place a rather simple physical interpretation on this effective potential for charged defects. The quantity $E_A^c - I_A$ is just the donor level $E_A^{0/+}$. Substituting into Eq. (1), one can easily verify that when this energy level falls below the Fermi level $\mu$, i.e., when $E_A^{0/+} - \mu$ is negative, the defect's energy is lower in its neutral state. But when this energy falls above $\mu$, the defect will prefer the positive state.

Moving the charged donor is analogous to moving the neutral defect, plus a hole at the level $E_A^{0/+}$. As this level moves up or down in response to changes either in the electrostatic fields, the local composition, or the strain, there is a corresponding change in the energy of the defect. The result is an effective potential

$$V_{\text{eff}}^{+} = H_A^0 - (E_A^{0/+} - \mu). \tag{3}$$

This can be derived simply by substituting the definition of $E^{0/+}$ into (2). But in addition, it can be interpreted as the effective potential for the neutral defect, plus a term corresponding simply to shifting the hole state up and down in energy (the sign would, of course, be positive instead for an electron). The electronic contribution to this, for the case of an acceptor, is shown in Fig. 1.

In the special case of a shallow donor impurity, the ionization potential is quite small and may be neglected here. In that case, the extra potential seen by the ion is simply the negative of the conduction-band edge. It is as if the conduction-band edge represented a real potential for an electron, while the donor, being positive, simply sees the opposite potential.

For a deep donor, where the ionization potential is a crucial term in the effective potential (2), no such simple interpretation is possible in general. But there may be cases in which the donor level $E^{0/+}$, or some other defect level, has only a very small discontinuity across the heterojunction. Such behavior has been discussed in particular in the case of transition-metal impurities in compound semiconductors.[5] In that case, the donor, although charged, would see at most a weak potential discontinuity across a heterojunction; to a first approximation, it would only see the "real" electrostatic fields

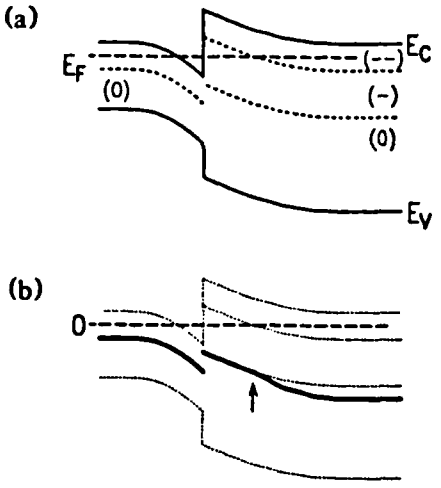



FIG. 1. (a) The band diagram for a semiconductor heterojunction, where both sides are $n$ type. Solid lines are valence maximum and conduction minimum, as indicated. The dashed line is the Fermi level. Dotted lines are single and double acceptor levels $E_A^{0/-}$ and $E_A^{-/--}$, where signs in parentheses indicate the charge state for the Fermi level in the indicated region. Very deep acceptor levels are chosen for visual clarity. (b) The heavy solid line is the electronic contribution to the effective potential, as in Eqs. (3) and (4), but for acceptors; the contribution from $H^0$ is omitted, so this is just the difference between the potential for the charged and neutral defects. The horizontal dashed line is the zero of the potential. Light dotted lines are copied from the band diagram (a), to illustrate the relation of the effective potential to the band diagram. Note that the potential is discontinuous at the interface. An arrow indicates a kink in the potential (discontinuity in force) where the double acceptor level crosses the Fermi level.

associated with band bending in the semiconductor.

Many defects can exist in multiple charge states. This possibility arises whenever one of the defect levels (e.g., $E_A^{0/+}$) falls in the band gap, so that it can be either above or below the Fermi level. The generalization to other charge states is straightforward. For example, for a double donor, following the same sort of analysis as above, one finds

$$V_{\text{eff}}^{++} = V_{\text{eff}}^{+} - (E_A^{+/++} - \mu), \tag{4}$$

where $E_A^{+/++} = E_A^{\epsilon} - I_A^{+}$, $I^{+}$ being the second ionization energy of the defect.

A defect may therefore have several effective potentials, one for each charge state. The defect will assume whichever charge state minimizes the energy, i.e., whichever gives the lowest effective potential. As the defect moves through regions of different electrostatic potential, strain, or composition, two potentials (e.g., $V^0$ and $V^+$) may cross, equivalent to the level (in that case $E^{0/+}$) crossing the Fermi level. At that point, the defect changes charge state, and the force (though not the potential) changes discontinuously. Variations in charge state with defect position may lead to a host of interesting effects.[6]

The electronic contribution to the effective potential (i.e., omitting the enthalpy term $H^0$) is shown for the case of an acceptor in Fig. 1. Note the discontinuity in the potential at the interface, and in the force where the charge state changes. Because we have assumed a constant composition and strain in each region for Fig. 1, the defect levels track the band edges within each region.

Finally, we consider some subtleties which were glossed over above. The first concerns the extent to which quantities such as the conduction-band minimum can be defined locally in an inhomogeneous system. This might seem to pose a serious problem here. For example, the ionization energy plays a crucial role in the derivation; but in a narrow quantum well, the lowest level into which an electron can be placed is raised, relative to the nominal conduction minimum, by a confinement energy. Thus the ionization energy is not strictly a local quantity.

Fortunately, if we go back to the physical processes used to derive Eq. (1), we see that the dependence of the various terms on such effects cancel. Thus $E_A^{\epsilon}$, etc., can be interpreted in the usual way, as corresponding to the values for an infinite, homogeneous system having the composition, strain, and electrostatic potential of the specified point in the real inhomogeneous system.

This may be illustrated by two simple examples. First, consider a deep donor, whose wave function is very localized. The neutral-state enthalpy depends only on the local composition and strain, and is unaffected by the fact that the impurity is in a narrow well. The ionization energy is increased by the confinement energy of the ionized electron. But the energy regained when returning the electron to the Fermi level is increased by exactly the same amount, so there is no net confinement effect. Alternatively, consider a very shallow donor. The ionization energy is negligible; but the enthalpy of the neutral impurity is increased by the confinement energy of the bound hydrogenic level, which is essentially identical to the confinement energy of the ionized electron, so again one has a complete cancellation.

Another subtlety neglected above involves the reference state for defining the enthalpy of the neutral defect. Consider for concreteness a substitutional P in a Si-Ge heterostructure. Moving the P from the Si into the Ge necessarily entails moving one Ge atom into the Si. In principle, this could be done by switching Si and Ge atoms at kinks in an interface step, with zero change in energy. But since we are dealing with a nonequilibrium process, there is no justification for assuming a particular final geometry. The actual final position of the Ge atom in the Si will depend on the diffusion mechanism. So there is a possible contribution to the enthalpy difference equal in magnitude to the enthalpy of substitution of one atom of $A$ in $B$.

Fortunately, for semiconductors which are well enough lattice matched to be useful in heterostructures, this enthalpy of substitution is typically exceedingly small, so we are justified in neglecting it. (For example, the enthalpy of substitution of Si in Ge, or vice versa, is estimated to be 10–20 meV.[7] For cation interchange in GaAs-AlAs, it is even smaller.) However, cases such as Ge-GaAs, where the two semiconductors are from different columns and so the substitution is not isovalent, present an interesting problem which is beyond the scope of this discussion.

Also, the neutral-defect enthalpy $H^0$ is not strictly local: Because of the finite range of the defect's strain field, the enthalpy depends on the host composition and strain over a finite-size region around the defect. However, the defect strain is typically large only for its nearest neighbors, making the potential nonlocal only on a 2–3-Å length scale, an effect which we can safely neglect in discussing diffusion, etc.

Finally, it might be tempting to assume that the enthalpy of a substitutional impurity, say, P, would be similar in Si and Ge, so the ion would feel an effective potential corresponding to the band discontinuity. However, without accurate measurements or calculations, such an assumption would be unjustified. It is equally possible, and perhaps even more plausible, that the ion has similar enthalpies, so the neutral enthalpy in the two hosts would differ according to the band discontinuity.

In conclusion, we have shown that an ion (or any charged defect) in a heterostructure experiences an effective potential, in addition to the enthalpy of the neutral defect, which is given simply by the donor or acceptor level (with the appropriate sign), relative to the Fermi level; or for higher charge states, by the sum of the

intervening levels. This result provides the remaining term, missing in previous work,[2] which is needed for a correct and complete treatment of diffusion in heterostructures.

It is a pleasure to acknowledge S. M. Hu for introducing me to this problem, and F. Stern for helpful comments. I am also grateful for discussions with R. Lever, who independently reached similar conclusions. This work was supported in part by ONR Contract No. N00014-84-C-0396.

[1]J. Tersoff, in *Heterojunction Band Discontinuities: Physics and Device Applications*, edited by G. Margaritondo and F. Capasso (North-Holland, Amsterdam, 1987).

[2]S. M. Hu, Phys. Rev. Lett. 63, 2492 (1989).

[3]R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. 54, 360 (1985); C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. 62, 1049 (1989).

[4]S. M. Hu has advocated this approach in a recent report (unpublished), which provided the motivation for the present work.

[5]J. Tersoff and W. A. Harrison, Phys. Rev. Lett. 58, 2367 (1987), and references therein.

[6]S. T. Pantelides, A. Oshiyama, R. Car, and P. J. Kelly, Phys. Rev. B 30, 2260 (1984).

[7]P. C. Kelires and J. Tersoff, Phys. Rev. Lett. 63, 1164 (1989), and references therein.

# Carbon Defects and Defect Reactions in Silicon

J. Tersoff
*IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 3 January 1990)

The energies of carbon defects in silicon are calculated, using an empirical classical potential, and used to infer defect properties and reactions. Substitutional carbon is found to react with silicon interstitials, with the carbon "kicked out" to form a (100) split interstitial. This interstitial can in turn bind to a second substitutional carbon, relieving stress, in three configurations with similar energies. The results here accord well with a variety of experimental data, including defect structures, activation energies for defect motion, and coupling to strain. A discrepancy with the accepted values for carbon solubility in silicon suggests a reinterpretation of the experimental data.

PACS numbers: 61.70.Rj, 61.70.Bv, 61.70.Yq, 71.45.Nt

In recent years there has been tremendous progress in the theoretical understanding of both dopant and native defects in silicon.[1] However, no comparable study has been made of isovalent impurities. Yet carbon is a ubiquitous and important impurity in silicon, exhibiting a wealth of interesting configurations.[2]

As a first step towards a fuller theoretical understanding, extensive calculations of carbon defects in silicon have been performed, using an empirical classical potential[3] to model the atomic interactions. Specific issues addressed here include the solubility of C in Si, its diffusion, the formation, migration, and structural properties of interstitial C, and the binding of C interstitials to substitutional C, to form C complexes. The results, summarized in Tables I and II, are in accord with experimental data for a striking variety of properties, confirming the value of the present simple approach for initial studies of this challenging system.

As expected, the lowest-energy form of C in Si is found to be substitutional C. The calculated activation energy for substitutional diffusion is a bit less than 4 eV, in reasonably good agreement with experiment.[4] Comparison of substitutional and interstitial energies indicates that a Si self-interstitial can "kick out" substitutional C, or form a C interstitial in the (100) split configuration, precisely the process which has been observed.[5] The interstitial's migration energy, and its elastic coupling tensor, are in good agreement with experiment.[5]

The interstitial can in turn bind with another substitutional C; the compressive stress of the interstitial C tends to cancel the tensile stress of the substitutional C. The predicted structures for this complex accord with those observed by Song *et al.*[6]

The enthalpy of solution for substitutional C has been measured[7,8] as 2.3 eV. Here the energy is calculated as 1.6 eV, an apparent discrepancy. However, simple considerations discussed below suggest that the experiment can be more consistently interpreted as giving an energy

TABLE I. Formation energy (in eV) of defects in silicon containing one carbon atom, and of their two-C complexes with a substitutional carbon. The silicon vacancy is also included, to show its interaction with substitutional C. Labels of two-C complexes are explained in text.

| Defect | Energy of defect | Energy of complex | Label of complex |
|---|---|---|---|
| Substitutional | 1.6 | | |
| Si vacancy | 3.7 | 5.0 | |
| Interstitials[a] | | | |
| *B* | 5.3 | 5.1 | CSC |
| | | 6.4 | CCS |
| *S* | 4.6 | 5.1 | SCSC |
| | | 5.2 | SCCS |
| | | 6.3 | SSCC |
| *X* | 5.9 | | |
| *T* | [3.8[b]] | 7.4 | |
| *H* | 6.7 | | |

[a]Labels *B*, *S*, *X*, *T*, and *H* denote bond centered, (100) split, exchange [i.e., (110) split], tetrahedral hollow site, and hexagonal hollow site; see Ref. 1 for structures.

[b]This small value is believed to be an artifact of the short cutoff distance used; see Ref. 13 for discussion.

TABLE II. Calculated activation energies, and strain coupling constants, for C defects in Si (in eV); and experimental values for comparison.

| Property | Calculation | Experiment |
|---|---|---|
| Energy of substitution | 1.6 | 1.5,[a] 2.5,[b] 1.7[c] |
| Interstitial migration | $\geq 0.7$[d] | 0.9[e] |
| Diffusion | 3.9 | 3.1[f] |
| Interstitial C $A_{ij}$ | | |
| $A_{11}$ | 8 | 7[e] |
| $A_{22}$ | −1 | 0 |
| $A_{33}$ | −7 | −7 |

[a]Reference 8.

[b]Reference 7.

[c]From reevaluation of Ref. 7; see text and Ref. 12.

[d]0.9 eV if include estimated barrier of 0.2 eV between *S* and *B* interstitials; see text.

[e]Reference 5.

[f]Reference 4.

of 1.7 eV, in good agreement with the present calculation.

The empirical classical potential used here to calculate the energies has been presented elsewhere.[3] It begins with potentials derived earlier for elemental Si and C; parameters describing Si-C interactions are determined from the elemental parameters by an interpolation scheme. This approach is necessarily less accurate than "first-principles" methods, and neglects electronic degrees of freedom. However, by simplifying the calculations, it permits us to get a broad view of the possible defects and reactions, tying together a large body of experimental data.

Extensive tests have confirmed the suitability of this method for treating point defects, including isovalent impurities.[3] In particular, results for C and SiC have been compared[3] with "state-of-the-art" quantum-mechanical calculations of Bernholc and co-workers.[9,10] The present method is rather successful in treating point defects in those materials, including antisite defects. There is thus ample reason to expect comparable accuracy for C in Si.

The potential here differs from that described earlier[3] only in a small change of the parameters for carbon. The parameters[11] used here are constrained to reproduce the energy of the vacancy in diamond, as calculated by Bernholc *et al.*,[9] at the expense of a poorer description of graphite, which was deemed less relevant for the present application. (In the defects studied here, three-coordinated C atoms have no opportunity for $\pi$ bonding. This is similar to the vacancy in diamond, but in contrast to graphite.) In addition, since we are not concerned with dynamical simulations, where the potential must go smoothly to zero with distance, the potential is here abruptly truncated at 2.5 Å, consistent with the original nearest-neighbor-only picture.[11] This short cutoff leads to problems only in the case of the tetrahedral interstitial, discussed below.

For consistency, we refer in Table I and throughout to the free energy of formation, $E - N_{\mathrm{Si}}\mu_{\mathrm{Si}} - N_{\mathrm{C}}\mu_{\mathrm{C}}$. The chemical potentials $\mu_{\mathrm{Si}}$ and $\mu_{\mathrm{C}}$ are $-4.63$ and $-7.70$ eV here, determined by equilibrium with Si (cohesive energy 4.63 eV/atom with the present potential), and with SiC (12.33 eV per formula unit). All structures are fully relaxed in a cubic cell 16.3 Å on a side (216 atoms without defects), with periodic boundary conditions, at zero temperature.

The natural place to begin is with substitutional C, since this is the simplest and most common carbon defect in Si. From elementary statistical mechanics,[12] the equilibrium concentration is expected to be $5\times10^{22}\exp(-\Delta/kT)$ cm$^{-3}$. Here $5\times10^{22}$ cm$^{-3}$ is simply the atomic density of pure Si, and $\Delta$ is the energy of substitution per atom. The energy of an isolated substitutional C impurity in Si is calculated to be $\Delta = 1.6$ eV; see Table I.

Bean and Newman[7] reported an energy of C in Si of $2.3 \pm 0.3$ eV. The discrepancy of 0.7 eV with the present results would be considered acceptable even for a first-principles calculation. However, the experimental value was determined by a fit to the solubility data, which yielded a concentration of $3.5\times10^{24}\exp(-2.3/kT)$ cm$^{-3}$. Although this result is still cited as the definitive measurement of carbon solubility, I know of no discussion of the unexpectedly large prefactor.

The actual measured solubility at high temperature (where the measurement should be most reliable), in combination with the theoretical prefactor, yields an energy of substitution of 1.7 eV, in good agreement with the present theoretical result. Moreover, the resulting solubility curve lies within the scatter of the experimental data over the entire temperature range, and actually improves the fit in the more reliable high-temperature range.

It therefore seems reasonable to propose a tentative reinterpretation of this experiment, as consistent with a solubility of $5\times10^{22}\exp(-1.7\ \mathrm{eV}/kT)$ cm$^{-3}$. In fact, an earlier experiment by Newman and Wakefield[4] was interpreted[8] as giving an energy of substitution of 1.5 eV, in excellent agreement with the present result of $\Delta = 1.6$ eV.

The calculated interaction between substitutional carbons is repulsive, with first- and second-neighbor interaction energies of 1.3 and 0.3 eV, respectively. Thus precipitation of substitutional carbon is not expected in the absence of structural defects which could relieve strain.

Substitutional C is found to have a more complex interaction with the Si vacancy. The "nearest-neighbor" interaction, i.e., for C on one of the four threefold sites, is repulsive: 0.2 eV. The C dangling-bond energy is of the order of 1 eV larger than that for Si,[9] which would imply a 1-eV repulsive interaction with the vacancy. However, this is largely cancelled by the energy gained from partial relief of the strain, when the C sits on the less-constrained threefold-coordinated site.

The second-neighbor interaction between substitutional C and the vacancy (shown in Table I) is, however, attractive: $-0.3$ eV. This binding results simply from the partial release of strain, because the vacancy's neighborhood is more easily deformed than the perfect crystal. This result suggests how defects such as internal surfaces, where steric constraints are weakened, can serve as centers for the nucleation of SiC precipitates.

From Table I, the lowest-energy C defects in Si, after the substitutional, are the low-coordination interstitials: the (100) split interstitial, and after that the bond-centered interstitial. The structures of these defects are shown in Fig. 1. (The tetrahedral interstitial is calculated to have an even smaller energy; however, this is apparently[13] an artifact of the short cutoff.)

The formation energy for an interstitial, starting from a substitutional C, from Table I is $4.6 - 1.6 = 3.0$ eV. This is less than the calculated energy of any Si self-interstitial.[3] Thus Si self-interstitials (e.g., generated by irradiation) should react with substitutional C in an exothermic "kick-out" process, forming interstitial C in the

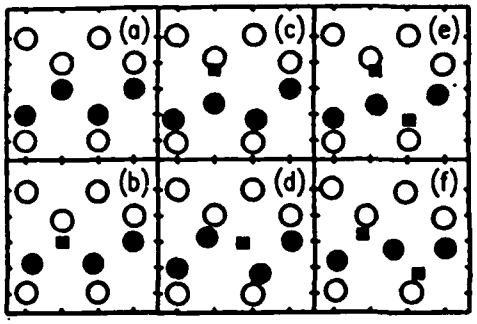



FIG. 1. Relaxed structure of selected C defects in Si. A $(1\bar{1}0)$ plane is shown, with the vertical and horizontal axes corresponding to the [001] and [110] directions, respectively. Each figure is $8\times8$ Å$^2$. Axis tick marks are chosen to correspond to ideal positions of Si atoms. Solid symbols are atoms in the plane of the figure; each open symbol corresponds to two atoms out of the plane, one in front and one behind. Circles are Si, smaller squares are C. (a) Pure Si, for reference. (b) Substitutional C in Si. Note inward displacement of neighboring Si. (c) Interstitial C in (001) split configuration. (d) Interstitial C in bond-centered configuration. Note displacement of C from [111] axis connecting its two Si neighbors. (e) Complex of substitutional C and C split interstitial, denoted SCSC in Table I. (f) Complex of substitutional C and C bond-centered interstitial, denoted CSC in Table I.

(100) split configuration. This reaction has been observed experimentally by Watkins and Brower,[5] with the resulting C interstitials having the expected structure.

The migration energy of the C interstitial, as well as its formation energy, can be estimated from Table I, and compared with experiment. One typically assumes[1] that the saddle point for interstitial migration is the next higher-energy interstitial configuration, in this case the bond-centered interstitial, giving a migration energy of $5.3-4.6=0.7$ eV. However, the split and bond-centered interstitials are both found to be (meta)stable minima, not saddle points, so the calculated migration energy should be $>0.7$ eV.

Watkins and Brower[5] have measured the interstitial migration barrier as 0.9 eV. This is consistent with the calculation if the barrier between the split and bond-centered interstitials is 0.2 eV. In fact, Song *et al.*[6] have studied a closely analogous system, the two-C substitutional-interstitial complex, and find precisely this behavior. Both the split and bond-centered configurations are found to be (meta)stable, with a barrier of 0.2 eV between them. This consistency provides strong, though indirect, evidence that the difference between split and bond-centered interstitial energies in Table I is rather accurate.

It is worth noting that in first-principles calculations of bond-centered interstitial energies, it is often found necessary due to practical constraints to consider only symmetry-preserving relaxations.[1] However, such a calculation for C in Si gives a formation energy of 6.6 eV, 1.3 eV higher than the value in Table I. Such a large value would qualitatively alter the conclusions for interstitial migration.

To address the formation energy of the split interstitial, we note that the activation energy for diffusion should be the sum of the formation energy of the interstitial from the substitutional, $4.6-1.6=3.0$ eV, plus the interstitial migration energy, 0.9 eV, giving a diffusion activation energy of 3.9 eV. This is in rather satisfactory agreement with the experimental value[4] of 3.1 eV. (Even first-principles defect calculations quote uncertainties of 0.5 eV or more.)

A powerful tool in identifying defects of low symmetry is the analysis of their stress-induced alignment. This alignment gives information on the elastic coupling tensor, $B_{ij}=dE/d\epsilon_{ij}$, or more precisely, on its traceless part $A_{ij}=B_{ij}-\frac{1}{3}\delta_{ij}\mathrm{Tr}B$. For the C interstitial, Watkins and Brower[5] found that $A_{11}=7$ eV, $A_{22}=0$ eV, and $A_{33}=-7$ eV. (Other elements are zero by symmetry.) Using the same orientation convention, the calculated components are $A_{11}=8$ eV, $A_{22}=-1$ eV, and $A_{33}=-7$ eV, in good agreement with experiment. (Even the agreement for $A_{22}$ should be considered good, since the relevant energy scale here is 7 eV. It is fortuitous that subtracting the hydrostatic component results in a number near zero.)

Finally, we consider the interaction of interstitial C to substitutional C. Table I gives the energies of several intuitively reasonable configurations for the substitutional-interstitial complex, based on the low-energy split and bond-centered interstitials.

The labeling of the two-carbon complexes in Table I is intended to be intuitive. For the bond-centered case, the interstitial atom is bonded to two neighbors. The label indicates the atoms along the chain, with S for silicon; e.g., CSC means a Si interstitial bonded to two C, as in Fig. 1(f). The split interstitial has two central atoms, each threefold coordinated. The label *WXYZ* means that the two central atoms are *X* and *Y*; *W* is C if any neighbor of *X* besides *Y* is carbon, otherwise it is S. Thus SCSC means that the two central atoms are C and Si, the C has all Si neighbors, while the Si has two C neighbors (and one Si), as in Fig. 1(e).

From Table I, three of the complexes have calculated energies distinctly lower than the rest, with about 1 eV binding energy. The two lowest-energy defects, labeled SCSC and CSC, have been identified by Song *et al.*[6] as the two configurations of a bistable complex, with nearly identical energies, in agreement with the calculation. The defect structures are shown in Figs. 1(e) and 1(f). The third low-energy configuration, SCCS, has not been observed.

The strong binding of interstitial C to substitutional C can be easily understood as arising (at least in part) from the relief of stress. The split and bond-centered interstitials are both under considerable compression, $\frac{1}{3}\mathrm{Tr}B=-8$ and $-18$ eV, respectively. In contrast, substitutional C, being smaller than Si, is under a large

tensile stress of 16 eV. The interstitial and substitutional can bind in complexes with much smaller stress than the individual defects, since the stresses are of opposite sign and tend to cancel. For the complexes labeled SCSC and CSC in Table I, the calculated stresses are only 8 and 12 eV, respectively.

In conclusion, by calculating the energies of a large number of possible C defects, we provide an overview of the expected properties of C in Si. The resulting picture is in excellent accord with a wide body of experimental data, including defect structures and reactions, activation energies for diffusion and for interstitial migration, and even the elastic coupling tensor for the low-symmetry C interstitial. For the solubility of C in Si, where a modest discrepancy exists with experiment, we propose that the experimental data can be more consistently reinterpreted as supporting the results of the present work.

It is a pleasure to acknowledge P. M. Fahey for helpful discussions, and R. M. Tromp for comments on the manuscript. This work was supported in part by ONR Contract No. N00014-84-C-0396.

---

[1]See, for example, R. Car, P. J. Kelly, A. Oshiyama, and S. T. Pantelides, Phys. Rev. Lett. 52, 1814 (1984); 54, 360 (1985); G. A. Baraff and M. Schluter, Phys. Rev. B 30, 3460 (1984); Y. Bar-Yam and J. D. Joannopoulos, Phys. Rev. B 30, 2216 (1984); C. S. Nichols, C. G. Van de Walle, and S. T. Pantelides, Phys. Rev. Lett. 62, 1049 (1989).

[2]For some reviews of recent work on C in Si, see, for example, *Oxygen, Carbon, Hydrogen, and Nitrogen in Crystalline Silicon*, edited by J. L. Mikkelsen, Jr., *et al.*, MRS Symposium Proceedings No. 59 (Materials Research Society, Pittsburgh, 1986).

[3]J. Tersoff, Phys. Rev. B 39, 5566 (1989). For further details on the potentials for pure Si and pure C, see, respectively, J. Tersoff, Phys. Rev. B 38, 9902 (1988), and Phys. Rev. Lett. 61, 2879 (1988).

[4]R. C. Newman and J. Wakefield, in *Metallurgy of Semiconductor Materials*, edited by J. B. Schroeder (Interscience, New York, 1961), pp. 15 and 201.

[5]G. D. Watkins and K. L. Brower, Phys. Rev. Lett. 36, 1329 (1976).

[6]L. W. Song, X. D. Zhan, B. W. Benson, and G. D. Watkins, Phys. Rev. Lett. 60, 460 (1988).

[7]A. R. Bean and R. C. Newman, J. Phys. Chem. Solids 32, 1211 (1971).

[8]Reference 7 reports that the analysis of data of Ref. 4 implies an energy of substitution of 1.5 eV.

[9]J. Bernholc, A. Antonelli, T. M. Del Sole, Y. Bar-Yam, and S. T. Pantelides, Phys. Rev. Lett. 61, 2689 (1988).

[10]C. Wang, J. Bernholc, and R. F. Davis, Phys. Rev. B 38, 12752 (1988).

[11]The parameters used here for carbon, in the potential and with the notation of Ref. 3, are as follows: $A=1544.8$ eV, $B=389.63$ eV, $\lambda=3.4653$ Å$^{-1}$, $\mu=2.3064$ Å$^{-1}$, $\beta=4.1612\times10^{-6}$, $n=0.99054$, $c=19981$, $d=7.0340$, $h=-0.33953$, and $\chi=0.9972$. As discussed in J. Tersoff, Phys. Rev. B 37, 6991 (1988), and in Ref. 3, the parameters $R$ and $S$ were not systematically optimized, a significant shortcoming of the present approach; instead, an abrupt cutoff $R=S=2.5$ Å was used for both C and Si, as discussed in the text.

[12]C. Kittel and H. Kroemer, *Thermal Physics* (Freeman, San Francisco, 1980), 2nd ed. [Equation (86) on p. 80 should be rederived using (80) rather than (83) to avoid the erroneous factor of $e^{-1}$.] This simple result is valid for substitutional impurities in the limit of low concentration, neglecting vibrational effects.

[13]As discussed elsewhere (Ref. 11), the somewhat arbitrary cutoff with distance is the most problematic aspect of the present potential. In the ideal tetrahedral interstitial, the second neighbors are only 15% more distant than the first neighbors, which creates particular problems. Increasing the cutoff to 2.9 Å raises the energy of the tetrahedral interstitial to over 10 eV; no other defect is affected even by the same order of magnitude by the cutoff. One can therefore only say that the energy should be between 3 and 10 eV, and is thus probably much larger than for the low-coordination interstitials.

# Nature of the Step-Height Transition on Vicinal Si(001) Surfaces

E. Pehlke and J. Tersoff
*IBM Research Division, IBM T. J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 18 March 1991)

The Si(001)2×1 surface is expected to undergo a phase transition from single- to double-atomic-height steps with increasing angle of miscut. Here we show that this transition is quite different than previously believed, involving something like a "devil's staircase" of transitions in a mixed phase consisting of a complex sequence of single and double steps. Even at low angles, where only single steps occur, the areas of 2×1 and 1×2 regions are unequal, in agreement with recent experimental results.

PACS numbers: 68.35.Bs, 64.80.Gd

Steps play a crucial role in growth at semiconductor surfaces. There has been particular interest in the role of steps of single versus double atomic height on Si(001), since single-height steps necessarily lead to antiphase boundaries in III-V semiconductors grown on Si [1]. Moreover, a series of papers by Alerhand and others [2–7] has revealed that steps on Si(001) exhibit fascinating behavior, including most notably a phase transition with increasing angle of surface miscut along [110].

Here we show that the nature of this phase transition, and the dependence of miscut angle generally, is rather different than previously believed. In particular, there is neither an abrupt transition from single- to double-height steps with angle [2–4] nor a coexistence between spatially separated regions of single- and double-height steps [5]. Instead, as the angle increases past a critical value, pairs of single-height steps collapse into double-height steps in a complex pattern, so that at zero temperature the surface undergoes a cascade of transitions resembling a "devil's staircase" [8]. In addition, even for small angles, where only single-height steps occur, the sizes of the 1×2 and 2×1 terraces are unequal. This explains the surprising recent measurements of Tong and Bennett [7].

The competition between single and double steps was first analyzed by Chadi [9], who identified the step atomic structures shown in Fig. 1. Because of the symmetry of the dimerized Si(001)2×1 surface, there are two distinct types of single-height steps, denoted $S_A$ and $S_B$. Single-height steps separate regions of 2×1 and 1×2 periodicity, so on vicinal surfaces one cannot have two $S_A$ steps without an intervening $S_B$ step, or vice versa. Since there is no corresponding restriction on double-height steps, only the lower-energy type (denoted $D_B$) need be considered [9].

Alerhand *et al.* [2] pointed out that, in order to fully understand the competition between single- and double-height steps, one must include the elastic interaction between the steps. Because of the anisotropic stress of the 2×1 surface, the energy of a flat Si(001) surface can always be lowered by introducing single-height steps, which break up the surface into domains of alternating orientation of the dimerization [6].

Single-height steps remain energetically preferred for small miscut angles. But because of the strong mutual repulsion between single steps, when the step spacing becomes too small (i.e., at large miscut angles) the energy of a pair of single steps becomes higher than that of a $D_B$ step, leading to a surface phase transition [2]. This picture was further elaborated by Poon *et al.* [4], and by Bartelt, Einstein, and Rottman [5].

From this extensive body of previous work, it is well established that there are two principal interactions between steps on Si(001), which follow directly from elementary elastic theory and the symmetry of the surface [2,4,6]. First, the anisotropic stress of the surface due to the 2×1 dimerization leads to a force monopole acting on the $S_A$ and $S_B$ step edges [6] and, consequently, a logarithmic dependence of the interaction energy on step separation [10]. Second, step edges in general may give rise to a force dipole, and for Si(001) this dipole is significant for $S_B$ and $D_B$ steps. The dipole-dipole interaction energy shows an $l^{-2}$ dependence on step separation $l$ [11]. A simple elastic model incorporating these interactions has proven to be very successful in describing the response of 2×1 and 1×2 domain sizes to an externally applied strain [6,12], and in reproducing the results of detailed atomistic simulations [4].

Here we use the same elastic model, augmented by

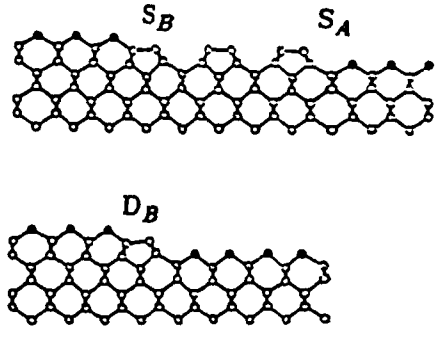



FIG. 1. Side view of the crystal structure of steps of single ($S_B$ and $S_A$) and double ($D_B$) atomic height on vicinal Si(001) surfaces. A $[1\bar{1}0]$ projection is shown. Horizontal bonds are dimers of 2×1 terrace; solid circles denote dimers of 1×2 terrace, which are normal to the plane of the figure.

atomistic simulations, to show that steps on Si(001) exhibit an even richer and more complex behavior than previously recognized. We begin our analysis by considering surfaces miscut at small angles from the (001), where only single-height steps are expected. Recently Tong and Bennett [7] found that the areas of 2×1 and 1×2 terraces were unequal even at small angles. This implies either that $S_A$ and $S_B$ steps are not equally spaced, as has been universally assumed, or else that $D_B$ steps are present even at small angles.

In order to understand this behavior, we calculate the energy of a surface with a given miscut angle, and hence a given step density, as a function of the width of the 2×1 terrace. (Here 2×1 and 1×2 refer respectively to the terraces with dimer bonds perpendicular and parallel to the step edge.) The results are shown in Fig. 2(a) for a relatively small (but otherwise arbitrary) angle. The terrace-size asymmetry is immediately apparent.

Before addressing the physical mechanism responsible for the asymmetry, we should mention some details of the calculation. As we are in effect extending the work of Poon *et al.* [4], we use the same potential model, that of Stillinger and Weber [13], to permit direct comparison. However, as discussed below, the *qualitative* results may be understood from a rather general perspective, and do not depend on the precise potential. The effect of the potential on *quantitative* results is also discussed below.

Because of the long periodicity of the steps, and the corresponding depth of the strain fields, accurate numerical relaxations require enormous cells when performed with traditional methods [4]. We have therefore implemented a more efficient approach, where a few layers of atoms are coupled to a semi-infinite elastic substrate. We have used 6–8 layers of atoms, and a continuum elastic substrate incorporating the full cubic anisotropy of Si. As it is convenient to have the substrate oriented along (001), an additional double-height step is included to simulate the vicinal surface on the flat substrate. This step is fixed in a rigid ideal geometry, so that it exerts no force. Details of this method will be given elsewhere. Comparison with calculations performed as in Ref. [4] indicate that the error from these approximations is less than 1 meV/$a$ ($a$ = 3.84 Å).

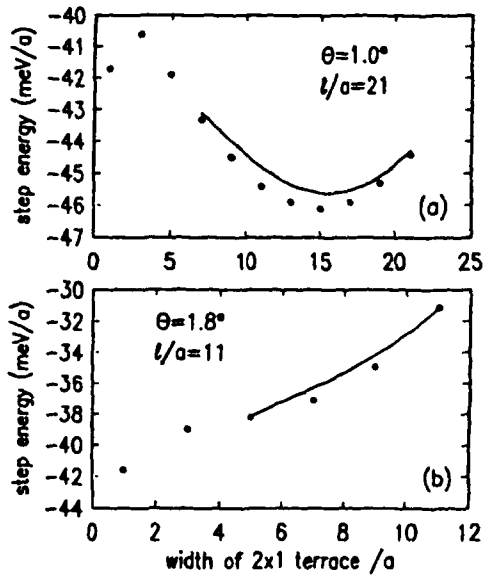



FIG. 2. Energy of a pair of $S_A$-$S_B$ steps for two fixed miscut angles $\theta$ corresponding to step separations $l$ vs the width of the minority terrace. The left-most data point corresponds to a $D_B$ step, representing the smallest possible terrace width. The right-most point is the symmetric case of equally wide 2×1 and 1×2 terraces (i.e., $p$=0). Dots are results from numerical relaxations of Si atoms interacting via the empirical Stillinger-Weber potential. Solid line is elastic continuum model, Eq. (1).

Our results, the dots in Fig. 2(a), show two local minima. The higher local minimum (left-most point) corresponds to a $D_B$ step. The configuration of minimum energy corresponds to a pair of single-height steps, with the $S_A$ step considerably displaced from the midpoint between the two neighboring $S_B$ steps (right-most dot is midpoint). Only a small barrier separates the single- and double-height configurations.

The asymmetric terraces can be understood easily within the elastic model mentioned above. Let $2l$ be the distance between two steps of the same kind, with $a$ being the surface lattice constant, and $(1-p)l$ being the size of the 2×1 terrace, so $p$ describes the asymmetry between 1×2 and 2×1 terrace sizes. Adopting Poon's notation, the parameters $\lambda_\sigma$ and $\lambda_d$ describe the force monopole and force dipole terms, with the remaining local contributions to the step energy included in a constant term $\lambda_0^{(S_A+S_B)}$. Then the energy of the step pair is

$$E=\lambda_0^{(S_A+S_B)}-2\lambda_\sigma \ln\left[\frac{l}{\pi a}\cos\frac{p\pi}{2}\right] +\lambda_d\left[\frac{a}{2l}\right]^2-(3\lambda_\sigma\lambda_d)^{1/2}\frac{a}{l}\tan\frac{p\pi}{2}. \quad (1)$$

Using values corresponding [4] to the Stillinger-Weber potential, Eq. (1) gives the solid curves in Fig. 2. These continuum results agree well with the full atomistic calculations for step separations $\geq 6a$. Since steps are never this close at the angles discussed in this paper, we can safely use the elastic model without further discussion.

For equal terrace sizes ($p$=0) the final term in Eq. (1) drops out, giving the expression used in Ref. [4]. However, the energy can be lowered by moving the $S_A$ step into a region of increasing displacements induced by the rebonding dipole at the $S_B$ step. For small angles this displacement of the $S_A$ step is roughly constant, approximately $5a$.

With increasing angle of miscut the $D_B$ minimum in Fig. 2 becomes deeper, until at a characteristic angle $\theta_c$ the energy of the $D_B$ minimum becomes lower in energy than the $S_A$-$S_B$ minimum, giving the transition discussed by Alerhand *et al.* and Poon *et al.* Including the terrace

asymmetry changes this angle only slightly, from $\theta_c = 1.1°$ to $\theta_c = 1.3°$.

At still higher angles, the local minimum associated with single-height steps disappears entirely, as seen in Fig. 2(b). At such angles $S_A$-$S_B$ steps are not even metastable with respect to step displacement, but collapse spontaneously into $D_B$ steps.

We now turn to the more complex issue of the nature of the step-height phase transition, which has been the subject of recent controversy [2,3,5]. Previous analyses have invariably begun with the *assumption* that one is dealing with a transition from a phase of pure single-height steps to one of pure double-height steps. As noted by Bartelt, Einstein, and Rottman [5], in this case instead of an abrupt transition at $\theta_c$ there should occur an interval of miscut angles $\theta_1 \leq \theta \leq \theta_2$ where the two phases coexist (spatially separated) in thermodynamic equilibrium. The critical angles $\theta_1$ and $\theta_2$ are determined by a common tangent construction [5], which for the present parameters gives $\theta_1 = 0.7°$ and $\theta_2 = 2.0°$.

However, Alerhand *et al.* [3] argued that such coexistence of phases would require faceting of the surface, and hence substantial mass transport, which might not be kinetically allowed. If one allows local equilibration but not long-range diffusion, an abrupt transition from single- to double-height steps with angle would be expected.

Let us first examine the assumption that the phases remain pure. In that case two-phase coexistence (surface faceting) clearly gives the lowest energy, regardless of whether such a state is kinetically accessible in experiments. If, beginning from such a state, we then move one pair of $S_A$-$S_B$ steps into the $D_B$ region, we find that the energy is lowered. Thus there exists a mixed phase of lower energy than any combination of pure phases.

To show that this conclusion is rather general, we examine a simplified model. Consider two types ($a$ and $b$) of steps with dipolar interactions; i.e., the interaction energy between any two steps $i$ and $j$ is proportional to $(\lambda_a^i \lambda_b^j)^{1/2}/L_{ij}^3$, where $L_{ij}$ is the distance between the steps, and $\lambda_a^i$ can take two different positive values, one for $a$ steps and one for $b$ steps. These represent $D_B$ steps and $S_A$-$S_B$ pairs, neglecting the internal degree of freedom of the pair spacing. Then, independent of the choice of parameters, the phase with alternating $a$ and $b$ steps is lower in energy than spatially separated $a$ and $b$ phases.

So far we have only shown that what was believed to be the lowest-energy step arrangement is in fact unstable against formation of a mixed phase. To identify the structure of the minimum-energy mixed phase, we consider a given angle of miscut, and generate all possible sequences of single- and double-height steps up to a given periodicity. For each such sequence, we calculate the energy within the elastic continuum model, minimizing the energy with respect to all step positions. We then choose the step sequence with the lowest relaxed energy.

The results are shown in Fig. 3 for periodicities up to five double steps. In every case, the energy of the mixed sequence identified is lower than the energy for two-phase coexistence or for either pure phase. For angles higher than about 1.5°, the 2×1 area becomes rather small, and periods longer than five are needed to describe the low-energy mixed phases.

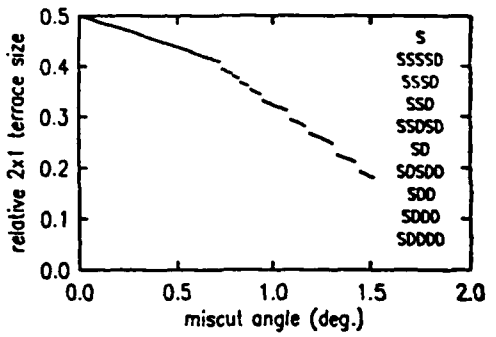



FIG. 3. Fraction of surface with 2×1 dimerization vs miscut angle. Patterns with periods up to five double steps are included, permitting treatment of angles up to about 1.5°. Each of the ten line segments corresponds to a distinct phase; the step orderings in the respective phases are given in the inset, where "S" denotes a *pair* of single-height steps and "D" a $D_B$ step, "SD" stands for a periodic arrangement of $S_B$-$S_A$-$D_B$ steps in that order, etc. Note that all orderings maximize the distance between single-step pairs.

The pattern of sequences in Fig. 3 is rather simple. The fraction of double steps increases monotonically with angle. For any given number of double steps, the remaining single-step pairs stay as far apart as possible, due to their strong mutual repulsion. For example, the sequence 11122 is missing in Fig. 3, while 11212, being slightly lower in energy, can be observed.

The average length of a 2×1 terrace (i.e., the total 2×1 terrace area divided by the number of single-height step pairs) remains approximately constant in the region of mixed phases, while it is, of course, rapidly decreasing with increasing angle in the pure single-height phase. Thus the behavior of the mixed phase is qualitatively reminiscent of the coexisting single- and double-height phases. Consequently, the staircase in Fig. 3 deviates only by a small amount from the almost linear curve (linear in $\tan\theta$) which one would expect in case of two-phase coexistence.

The result looks like a devil's staircase: If we could treat longer and longer periodicities, we would expect to see finer and finer structure. In fact, in the simplified model mentioned above of two steps with dipolar interactions, if we constrain the steps to be equally spaced, the model reduces to one which has been rigorously proven to exhibit a true devil's staircase of transitions [8].

The above results are strictly valid only at zero temperature; due to the small energy differences involved, the complicated ordered structures will be destroyed at temperatures where equilibration is feasible. Nevertheless, our conclusions remain relevant at higher temperatures. For the pure single-height steps which occur at small mis-

cut angles, finite temperatures lead to thermal meandering; but this meandering takes place in a nearly symmetric potential, and so should result in little change in the average step position (and hence in the average $2\times1$ terrace size). Therefore the $T=0$ theory is directly applicable to the low-angle data in Tong and Bennett's experiment [7].

At higher angles, since steps already mix at $T=0$, this should remain true *a fortiori* for finite temperatures. Thus our conclusion that the equilibrium state is mixed rather than faceted or pure, with the $2\times1$ terrace area varying continuously with angle, remains true at all temperatures. However, as discussed by Alerhand *et al.* [2], at higher temperatures the free energy of the single-height phase is lowered by step meandering, shifting the transitions to higher angles.

Before ending, we should return to the question of quantitative accuracy. The interactions of steps on Si(001) are well described by the elastic model; but the values of the parameters in the model are not accurately known. The values used here were obtained in Ref. [4] from a specific empirical atomistic model [13]. However, we have calculated the stress anisotropy for this model, and find it to be a factor of 2 smaller than the most accurate available value [14]. This results in the crucial parameter $\lambda_\sigma$ being a factor of 4 too small. Use of more accurate parameters in the elastic model would probably shift the transitions in Fig. 3 to higher angles.

Finally, we note that at any finite temperature, along any infinite step edge in equilibrium there are necessarily both $S_A$-$S_B$- and $D_B$-like regions. Thus intermixing is expected even in the direction parallel to the steps. Figure 2 suggests a way to describe such step meandering, consistently allowing for both $S_A$-$S_B$ and $D_B$ steps within a unified one-dimensional model Hamiltonian. One simply replaces the harmonic term in the Alerhand-Poon Hamiltonian [2,4] by the full potential of Fig. 2. Preliminary Monte Carlo simulations for this model, for angles in the transition region, confirm that along the step edge there are $S_A$-$S_B$- and $D_B$-like regions. This intermixing has also been previously inferred experimentally by Tong and Bennett from their scattering profiles.

We are grateful to P. A. Bennett for stimulating discussions, and for providing us with his results prior to publication. This work was supported in part by ONR Contract No. N00014-84-C-0396.

[1] H. Kroemer, in *Heteroepitaxy on Silicon,* edited by J. C. C. Fan and J. M. Poate, MRS Symposia Proceedings Vol. 67 (Materials Research Society, Pittsburgh, 1986), p. 3.
[2] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, Phys. Rev. Lett. **64**, 2406 (1990).
[3] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, Phys. Rev. Lett. **66**, 962 (1991).
[4] T. W. Poon, S. Yip, P. S. Ho, and F. F. Abraham, Phys. Rev. Lett. **65**, 2161 (1990).
[5] N. C. Bartelt, T. L. Einstein, and C. Rottman, Phys. Rev. Lett. **66**, 961 (1991).
[6] O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopoulos, Phys. Rev. Lett. **61**, 1973 (1988); D. Vanderbilt, O. L. Alerhand, R. D. Meade, and J. D. Joannopoulos, J. Vac. Sci. Technol. B **7**, 1013 (1989).
[7] X. Tong and P. A. Bennett, Phys. Rev. Lett. **67**, 101 (1991).
[8] P. Bak and R. Bruinsma, Phys. Rev. Lett. **49**, 249 (1982).
[9] D. J. Chadi, Phys. Rev. Lett. **59**, 1691 (1987).
[10] V. I. Marchenko, Pis'ma Zh. Eksp. Teor. Fiz. **33**, 397 (1981) [JETP Lett. 33, 381 (1981)].
[11] V. I. Marchenko and A. Y. Parshin, Zh. Eksp. Teor. Fiz. **79**, 257 (1980) [Sov. Phys. JETP **52**, 129 (1980)].
[12] F. K. Men, W. E. Packard, and M. B. Webb, Phys. Rev. Lett. **61**, 2469 (1988).
[13] F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).
[14] R. D. Meade and D. Vanderbilt, in *Proceedings of the Twentieth International Conference on the Physics of Semiconductors,* edited by E. M. Anastassakis and J. D. Joannopolous (World Scientific, Singapore, 1990), p. 123; (unpublished).

# Phase Diagram of Vicinal Si(001) Surfaces

E. Pehlke[a] and J. Tersoff

*IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598*

(Received 22 May 1991)

Vicinal Si(001) surfaces are believed to undergo a phase transition between single and double atomic height steps as either temperature or angle of miscut is varied. Here we calculate the full temperature-angle phase diagram, which is found to be quite different than previously believed. In particular, there is a critical point above which there is no phase transition at all. The results appear to explain the rather continuous behavior seen in a variety of experiments.

PACS numbers: 68.35.Bs, 64.80.Gd, 68.35.Md

Surface steps are crucial in determining the growth and shape of crystals, and there has recently been intense interest in understanding the thermodynamics of steps, e.g., bunching, faceting, and step-height transitions [1,2]. In particular, steps on vicinal Si(001) surfaces miscut towards [110] exhibit a fascinating transition from single to double atomic height steps. Yet there is considerable controversy concerning the nature or even the existence of a phase transition for this surface [3,4]. Theoretical treatments have predicted a first-order phase transition with temperature and with angle of miscut from (001) [5–8]; yet experiments find only a continuous variation of all observable quantities [9–11].

Here, by calculating the full temperature-angle phase diagram, and including a more complete and accurate description of the fundamental thermal excitations of the system, we reconcile the predicted existence of a phase transition with the continuous behavior observed experimentally. We show that there is a thermodynamic critical point in the surface phase diagram, above which there is no phase transition with angle. If surface equilibration only occurs at temperatures above the critical point, then the phase transitions predicted theoretically should not be experimentally observable. In addition, the nature of the transitions is such that they should be far more difficult to identify in experiments than previously believed, even if they occur in an accessible temperature range.

It is well known from different experiments (see, e.g., references in [5]) that at small miscut angles the Si(001) surface consists of terraces of alternating $1\times2$ and $2\times1$ dimerization. These terraces are separated by single atomic height steps, which are denoted [12] $S_A$ and $S_B$ according to whether the dimerization on the upper terrace is perpendicular or parallel to the step edge, respectively. (On vicinal surfaces such steps must occur in $S_A$-$S_B$ pairs, which we collectively call $S$ steps.) At larger miscut angles double atomic height steps (denoted $D_B$) dominate [13,14], and the surface approaches a single domain structure, consisting of dimers parallel to the step edges ($1\times2$ dimerization). Alerhand *et al.* [5] showed that this transition results from the elastic interaction between steps [15,16] which favors single height steps at large step-step separations (small angles of miscut), and double height steps at smaller separations.

The role of temperature has so far been included only as a contribution to the free energy of single height steps from meandering. Alerhand *et al.* [5], and later Poon *et al.* [8], calculated the free energy of meandering $S_B$ steps on a single-height-stepped surface. (Meandering of the $S_A$ is believed to be negligible.) They employed a one-dimensional model Hamiltonian including kink-energy terms and a harmonic potential (so that the $S_B$ step energetically prefers a position in the middle between the two neighboring $S_A$ steps).

Comparing the free energy of $S$ steps with the energy of straight $D_B$ steps, Refs. [5] and [8] concluded that there is a first-order phase transition with angle of miscut at any temperature, from a pure $S$ phase to a pure $D$ phase. However, experiments to date have not observed the abrupt transition predicted. Instead, only a continuous variation with angle [9,10] and temperature [11] has been observed.

There are two crucial elements missing in previous theoretical treatments of the surface at finite temperature. The first element is a correct identification of the zero-temperature structure. We recently showed that the transition from single to double steps with increasing angle is not abrupt; rather, it takes place through a (presumably infinite) sequence of phases consisting of distinct ordered mixtures of double ($D$) and pairs of single ($S$) height steps [17].

The second missing element is a comprehensive description of step meandering. The meandering of isolated $S_B$ steps has already been treated in detail [5,8]. However, a double step may be viewed as a bound pair of single steps ($S_A$ and $S_B$). At finite temperature, the $S_B$ step of this pair may meander, breaking up the double step locally. This excitation has been proposed based on reflection high-energy electron diffraction experiments of Tong and Bennett [9], and seen in scanning tunneling microscopy experiments of Wierenga, Kubby, and Griffith [14]. And it is this excitation which blurs the distinction between single and double steps at high temperature, leading to a critical point in the phase diagram.

Before presenting results we briefly sketch our procedure. Rather than considering a single-step pair, we must consider at least two pairs of steps, in order to describe the tendency of the surface to form phases consist-

ing of alternating single and double steps [17]. Such a set of two pairs is shown in Fig. 1. The spacing of these steps, in the absence of meandering, can be described by four parameters: $l$, $l'$, $d$, and $L$. Here $L$ is the overall periodicity, which is related to the surface miscut angle $\theta$ and surface lattice constant $a$ by $L/a=\sqrt{2}/\tan\theta$; $l$ and $l'$ are the widths of the $2\times1$ terraces enclosed by two neighboring $S_A$ and $S_B$ steps; and $d$ is the distance between the $S_A$ steps. For meandering steps, we can so specify the spacing along any given atomic row in the [110] direction, i.e., perpendicular to the step.

The calculation of the interaction energy $v(l,l',d,L)$ is based on an elastic model [5,8,15,16,18], which has been widely and successfully employed to treat this surface. We include both the force monopole due to the anisotropy of the surface stress and the force dipole due to local rebonding at the $S_B$ and $D_B$ step edges. For the interaction parameters we take the values derived by Poon *et al.* [8]. However, one must bear in mind that these elastic parameter values were obtained by fitting to atomistic simulations which used an empirical model [19], so they may not be quantitatively accurate. Thus while the results here reliably describe the topology of the phase diagram, the actual temperatures and angles at which the transitions take place are rough estimates.

The geometry of a $D_B$ step is essentially that of an $S_A$-$S_B$ pair separated by about $1.5a$. In fact, by choosing the spacing to be $1.57a$, the long-range interaction field of the $S_A$-$S_B$ pair becomes equivalent to that of a $D_B$ step for the parameter values used here. Thus, with respect to the interaction with other steps, the $D_B$ step may be treated simply as a bound pair of single steps. We need only add to the elastic model a short-ranged (contact) interaction between single steps to give the correct $D_B$ step energy.

However, unlike earlier treatments, to describe the binding and unbinding of single-step pairs our elastic model must accurately reproduce the interaction of steps at atomic distances. We do this by broadening surface forces with a Lorentzian of width $a$, retaining the full complexity of the resulting cumbersome expressions. We have explicitly verified the accuracy of this treatment at all step separations.

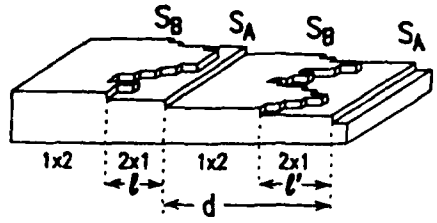



FIG. 1. Schematic drawing of a vicinal Si(001) surface with alternating straight $S_A$ and meandering $S_B$ single atomic height steps. The direction of dimerization is rotated by 90° on consecutive terraces. By $1\times2$ we denote terraces with Si dimer bonds parallel to the $S_A$ step edge. When the $S_B$ approaches the $S_A$ step (separation $1.5a$) a local portion of $D_B$ step is formed, as depicted in the left part of the figure.

The step meandering occurs in units of $2a$ parallel and perpendicular to the step edge, preserving the local atomistic structure of the steps [20]. The energy of a configuration of two meandering step pairs with total length $2Na$ parallel to the steps, and terrace sizes $l_i$ and $l_i'$ at the $i$th position along the step edge, is given by the Hamiltonian

$$H=\sum_{i=1}^{N}[\lambda_\perp|l_{i+1}-l_i|+2\epsilon_c(1-\delta_{l_{i+1},l_i}) +\lambda_\perp|l_{i+1}'-l_i'|+2\epsilon_c(1-\delta_{l_{i+1}',l_i'}) +2v(l_i,l_i',d,L)]. \quad (1)$$

Here $\lambda_\perp$ denotes the energy per length of the intervening $S_A$ step, and $\epsilon_c$ is the corner energy of the kink. We use the values proposed in Ref. [8].

The free energy per $1\times1$ surface unit cell for a fixed separation $d$ of the $S_A$ steps is calculated in the usual way from the maximum eigenvalue $\lambda_{max}$ of the transfer matrix:

$$f(T,\theta,d)=-\frac{k_BT}{2(L/a)}\ln\lambda_{max}(T,\theta,d). \quad (2)$$

Here $k_B$ is Boltzmann's constant, and the factor of 2 in the denominator of the prefactor is due to the unit step of meandering being 2 times the $1\times1$ surface lattice constant. In equilibrium the free energy is minimized with respect to the $S_A$ step separation,

$$f(T,\theta)=\min_d f(T,\theta,d). \quad (3)$$

Note that, technically speaking, due to this minimization our model is effectively not one dimensional. It is the elastic interaction perpendicular to the step edges that leads to the existence of ordered structures of $S$ and $D$ steps, and to the corresponding phase transitions.

Because of the added complexity of treating finite temperature, we restrict ourselves to structures of up to two step pairs. The value of $d$ that minimizes $f(T,\theta,d)$ describes the extent to which these step pairs differ. If $d=L/2$, then the two step pairs are statistically equivalent; any deviation is a signature of the alternating $SD$ phase

We start the calculation of $\lambda_{max}$ with a mean-field type of estimate, disregarding correlations of neighboring step pairs by assuming $p(l_i,l_i')\approx p(l_i)p'(l_i')$, where $p(l_i,l_i')$ is the probability of simultaneously having terrace sizes $l_i$ and $l_i'$. In a final step this result is improved by vector iteration with the full transfer matrix, allowing for additional anticorrelation of the terrace widths $l_i$ and $l_i'$. However, the corresponding correction of the free energy away from the phase transition is quite small.

We can immediately get a qualitative picture of the nature of the phase transition here from Fig. 2, which shows the dependence of the free energy $f(T,\theta,d)$ on $d$, i.e., on

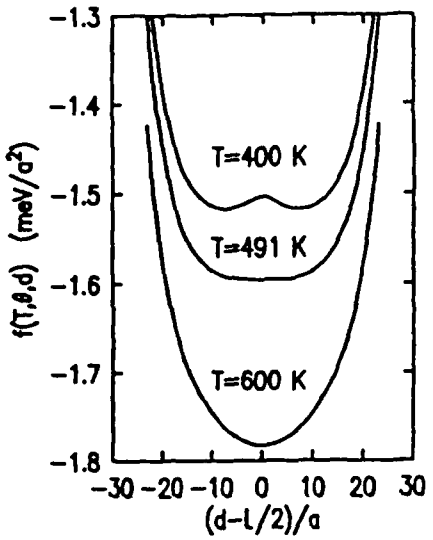



FIG. 2. Free energy per (1×1) surface unit cell for a given separation of the $S_A$ steps, vs the deviation of this separation $d$ from the symmetric (equidistant) value $d=L/2$. The miscut angle is $\theta=1.45°$ and the temperatures were chosen to be below, near, and above the critical point in the phase diagram.

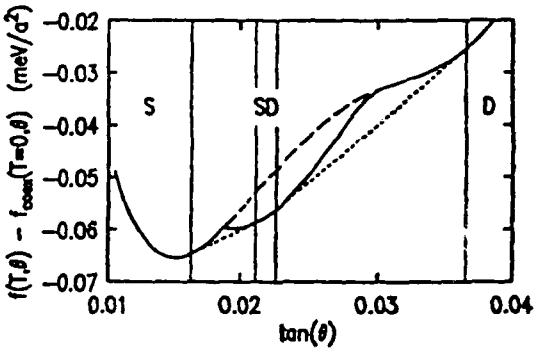



FIG. 3. Free energy per surface unit cell vs $\tan\theta$ for $T\approx 0.9T_c$, and Gibbs's construction. For ease of viewing, a linear function of $\tan\theta$ (arbitrarily chosen to equal the $T=0$ coexistence curve between pure $S$ and $D$ phases) has been subtracted from the data. Solid line: free energy $f(T,\theta)$. Dashed line: free energy $f(T,\theta,d=L/2)$ for equidistant $S_A$ steps, i.e., suppressing $SD$ alternation. Dotted line: (formal) Gibbs's construction for coexistence of $S$ and $SD$ or of $SD$ and $D$ phases. Existence regions for the pure phases are marked. See text for the correct physical interpretation of coexistence regions.

the degree of step alternation, at different temperatures. At finite temperature, because of step meandering, the distinction between $S$ and $D$ steps is not unambiguous. However, at low temperature the minimum of $f$ in Fig. 2 occurs for $d\neq L/2$, i.e., for adjacent step pairs alternating between $S$-like and $D$-like. The latter has also been verified by direct inspection of the probability distributions for the 2×1 terrace sizes of both step pairs.

As the temperature rises, the distinction between $S$ and $D$ becomes smaller, and so the (thermally averaged) elastic energy gained by $SD$ alternation falls; meanwhile the $SD$ configuration becomes less favorable for entropic reasons. At the highest temperature in Fig. 2, entropy clearly wins, and the lowest $f$ occurs for the symmetric configuration.

To explain the procedure for constructing the complete phase diagram, the angle dependence of $f(T,\theta)$ is shown in Fig. 3. At low angles we observe a symmetric (i.e., $d=L/2$) phase of $S$ steps, at high angles a symmetric phase of steps of predominantly $D$ character, and in between the asymmetric $SD$ phase. As we explicitly allow for periodicities only up to two step pairs, Gibbs's construction formally gives two coexistence regions: one of $S$ and $SD$, and one of $SD$ and $D$ phases. However, from our earlier more detailed study of the $T=0$ case [17], we know that there really is no coexistence of phases. Instead, these coexistence regions have to be interpreted as (quite good) approximations to those parts of the phase diagram where the more complicated ordered phases (length $\geq 3$ step pairs) of mixed $S$ and $D$ steps occur.

The resulting phase diagram is shown in Fig. 4. The open circles have been determined as described above, i.e., by Gibbs's construction at each temperature as depicted in Fig. 3. Note that the boundaries of the pure $S$, $D$, and $SD$ phases agree well with the earlier $T=0$ ("devil's staircase") results [17], shown as squares. Some other points near $T_c$ were derived in a different way. For example, the diamonds were obtained from temperature scans. However, at temperatures above $T_c\approx 490$ K the curves of free energy versus $\tan(\theta)$ are convex, and $d$ equals $L/2$ for all values of $\theta$. Thus there is no phase transition above this critical temperature.

This picture of the phase transition implies a very different interpretation of experimental results. The freeze-in temperature of step structures on Si(001) is

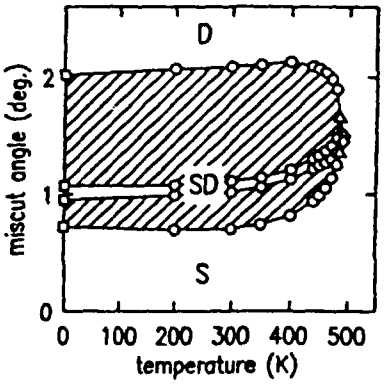



FIG. 4. Phase diagram for vicinal Si(001) surfaces. The thin center region corresponds to an $SD$ phase, and the outside region to the symmetric phase, which at low temperature may be interpreted as $S$ and $D$ phases. In the intervening regions, more complicated mixed ordered phases are expected. The boundaries of the pure-phase regions agree well with results of an earlier more elaborate treatment at $T=0$, represented by squares.

generally believed to be around 800 K [9,21]. If the result that $T_c \approx 490$ K is even roughly correct (or is too high), then the experiments measure surfaces equilibrated above the critical temperature, where there is, in fact, no phase transition. This would reconcile the theoretical predictions of a phase transition with the experimental observations of only continuous behavior.

Because of uncertainty in the values of the parameters which enter the elastic model, we cannot rule out the possibility that $T_c$ could be above 800 K. However, the phase transition might still be extremely hard to observe experimentally. Even at $T=0$, the transition takes place through a quasicontinuous sequence of weak first-order transitions [17], so that properties such as surface energy or terrace asymmetry should vary in a nearly continuous manner. At higher temperatures, this will be all the more true. Thus in the presence of experimental noise, there might be no observable qualitative difference between the behavior above and below the critical temperature.

We are grateful to D. P. DiVincenzo for discussions, and to P. A. Bennett and W. Ranke for providing preprints of their work prior to publication. This work was supported in part by ONR Contract No. N00014-84-C-0396.

---

[a]Present address: Fritz-Haber-Institut, Faradayweg 4-6 D-1000 Berlin 33, Germany.

[1] E. D. Williams and N. C. Bartelt, Science 251, 393 (1991).
[2] N. C. Bartelt, T. L. Einstein, and E. D. Williams, Surf. Sci. Lett. 240, L591 (1990).
[3] G. Kochanski, Bull. Am. Phys. Soc. 36, 910 (1991).
[4] R. Kariotis, M. B. Webb, and M. G. Lagally (unpublished).
[5] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, Phys. Rev. Lett. 64, 2406 (1990).
[6] N. C. Bartelt, T. L. Einstein, and C. Rottman, Phys. Rev. Lett. 66, 961 (1991).
[7] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, Phys. Rev. Lett. 66, 962 (1991).
[8] T. W. Poon, S. Yip, P. S. Ho, and F. F. Abraham, Phys. Rev. Lett. 65, 2161 (1990).
[9] X. Tong and P. A. Bennett, Phys. Rev. Lett. 67, 101 (1991).
[10] E. Schröder-Bergen and W. Ranke (unpublished).
[11] C. E. Aumann, J. de Miguel, R. Kariotis, and M. G. Lagally, Bull. Am. Phys. Soc. 36, 909 (1991).
[12] D. J Chadi, Phys. Rev. Lett. 59, 1691 (1987).
[13] J. E. Griffith, G. P. Kochanski, J. A. Kubby, and P. E. Wierenga, J. Vac. Sci. Technol. A 7, 1914 (1989).
[14] P. E. Wierenga, J. A. Kubby, and J. E. Griffith, Phys. Rev. Lett. 59, 2169 (1987).
[15] V. I. Marchenko, Pis'ma Zh. Eksp. Teor. Fiz. 33, 397 (1981) [JETP Lett. 33, 381 (1981)].
[16] V. I. Marchenko and A. Y. Parshin, Zh. Eksp. Teor. Fiz. 79, 257 (1980) [Sov. Phys. JETP 52, 129 (1980)].
[17] E. Pehlke and J. Tersoff, Phys. Rev. Lett. 67, 465 (1991).
[18] O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopoulos, Phys. Rev. Lett. 61, 1973 (1988).
[19] F. H. Stillinger and T. A. Weber, Phys. Rev. B 31, 5262 (1985).
[20] B. S. Swartzentruber, Y. W. Mo, R. Kariotis, M. G. Lagally, and M. B. Webb, Phys. Rev. Lett. 65, 1913 (1990).
[21] F. K. Men, W. E. Packard, and M. B. Webb, Phys. Rev. Lett. 61, 2469 (1988).

# Sinuous Step Instability on the Si(001) Surface

J. Tersoff and E. Pehlke[a]

*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*

(Received 4 November 1991)

On slightly miscut Si(001) surfaces, straight steps are predicted to be unstable against the formation of long-wavelength undulations. These undulations lower the energy, by, in effect, reducing the size of the stress domains; they are thus analogous to the spontaneous step formation proposed by Alerhand *et al.* However, step undulations are expected to be kinetically favored, and therefore to preempt spontaneous step formation. Moreover, they lead to an unexpected distinct thermodynamic phase in the surface phase diagram at small angles.

PACS numbers: 68.35.Bs, 68.35.Md

Steps on vicinal Si(001) surfaces have been intensely studied, especially since Alerhand *et al.* predicted such remarkable effects as spontaneous formation of steps [1], and a transition in step height with angle of miscut [2–4]. However, theoretical analyses to date have universally assumed that these steps are straight, except for random thermal meandering [1–7]. Yet recently, Tromp and Reuter [8], using low-energy electron microscopy (LEEM), observed steps on rather flat Si(001) surfaces to be sinuous rather than straight on a submicron length scale.

Here we show that, for sufficiently low step densities, straight steps are unstable against long-wavelength distortions, leading to a new phase transition on this surface. The cause is the interaction between surface stress domains. These results lead to a new picture of the structure and phase diagram of vicinal surfaces, and offer a natural explanation for the remarkable observation of Tromp and Reuter.

Alerhand *et al.* first recognized the importance of steps in creating stress domains on Si(001)2×1, and showed that a surface with sufficiently low step density could reduce its energy by introducing extra steps [1]. Given the strength of their argument, the failure to observe such extra steps has been a puzzle. The results here finally resolve this puzzle—step undulations can relieve stress and hence preempt the formation of extra steps.

Moreover, such undulations are kinetically preferred. There is a large barrier to nucleating extra steps, but little barrier to step undulations. Also, during either growth or sublimation (e.g., while heat cleaning), step flow places severe kinetic constraints on the step geometry. Unlike spontaneous formation of up-and-down steps, step undulations are compatible with step flow.

We begin by recalling the relevant features of the Si(001)2×1 surface, and of the continuum elastic model which has been successfully used to describe step interactions on this surface [1–5]. For unreconstructed Si(001), the surface lattice constant in the [110] direction is $a=3.84$ Å. For a surface miscut by an angle $\theta$ in the [110] direction, the separation between equally spaced single-layer steps is

$$L=a/2\sqrt{2}\tan\theta. \quad (1)$$

The Si(001) surface exhibits a 2×1 reconstruction in which pairs of atoms form dimers. Because of the atomic geometry, at single-layer steps the dimerization necessarily rotates by 90°, from 2×1 to 1×2 or vice versa. If the dimers on the upper terrace are perpendicular to the step edge, the step is called [5] $S_A$, or if parallel, $S_B$.

Because the stress is anisotropic and the domain rotates 90° at a step, the stress is discontinuous at the step. Using the known [9] stress tensor of the surface, we take the divergence of the stress to obtain the force on a step, referred to as a "force monopole" [1,6]. The elastic energy of the steps is then $-\frac{1}{2}\int d^2x\, d^2x'\, \chi_{ij}(\mathbf{x}-\mathbf{x}')f_i(\mathbf{x})f_j(\mathbf{x}')$, where $f_i$ is the force density at the surface, and $\chi$ is the elastic Green's function of the surface. We calculate the Green's function numerically for a semi-infinite geometry, using the full cubic anisotropy with the experimental elastic constants. For sinusoidal steps, the Fourier transform of the force density can be calculated analytically. The integral for the elastic energy then transforms into a reciprocal-lattice sum, which is performed numerically.

The only other property needed to describe the steps is an energy per length for each type of step ($S_A$ or $S_B$), reflecting a "local" energy in addition to the energy of the strain field. We do not include any "corner energy" [3], so that we can treat the continuum limit without considering the microscopic distribution of kinks in the meandering steps.

We omit thermal and entropic effects here. These have been extensively discussed already [2–4]. At large step separations, the steps meander about their minimum-energy positions. This meandering has a short correlation length; thus while it results in a renormalization of the local energies [2], it should not qualitatively affect the long-wavelength properties studied here.

We restrict consideration here to equally spaced identical sinusoidal steps, as shown in Fig. 1. Besides simplifying the elastic calculation, this allows the steps to be fully characterized by two numbers: the period $\lambda$ and amplitude $A$ of the sine wave. No distinction need be made here between $S_A$ and $S_B$ steps; because of the symmetrical step pattern assumed, only the sum of their energies enters. These restrictions are discussed further below.

For straight steps the energy $E_s$ can be calculated analytically, giving the well-known [1,6] logarithmic

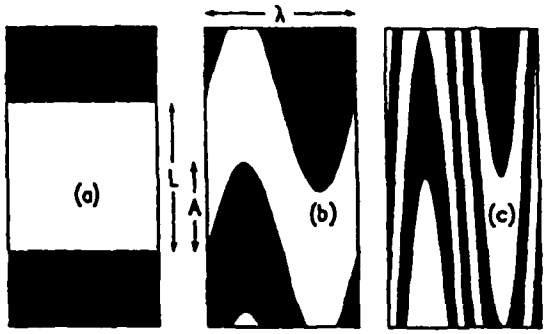



FIG. 1. Pattern of equally spaced sinusoidal steps used here. Black and white regions correspond to 2×1 and 1×2 domains, which are separated by single-layer steps. Step spacing $L$, wavelength $\lambda$, and amplitude $A$ are indicated. (a) $A=0$, (b) $A=0.6L$, (c) $A=3L$. $L$ and $\lambda$ are the same in all three figures.

dependence on step separation $L$, $E_s=C_1-C_2\ln(L/S)$, where $C_2$ reflects the strength of the interaction, and $C_1$ characterizes the local energy of the step. Our value of $C_2$ is 29 meV/$a$, considerably larger than that suggested previously [1], mainly because of more accurate recent calculations of the stress anisotropy [9]. $C_1$ here represents an average of $S_A$ and $S_B$ local energies; its numerical value depends upon the (arbitrary) choice of $S$, and we adopt the convention $S=\pi a$ chosen by Alerhand *et al.* [1]. Since the actual step energy for Si is not well known, we somewhat arbitrarily choose a local energy such that $C_1=58$ meV/$a$. The effect of this choice and of other approximations is discussed below.

The energy per area, $E_s/L$, has a minimum at the step separation

$$L_0=S\exp(1+C_1/C_2)\,, \tag{2}$$

giving $L_0\approx 63a$ for the parameter values used here. [For the small angles of interest here, in discussing energy per area it is not necessary to distinguish between surface area and the projection of that area onto the (100) plane, or between $\theta$ and $\tan\theta$.] As was first pointed out by Alerhand *et al.* [1], a sufficiently flat surface could lower its energy by spontaneously forming additional steps to decrease the step separation to $L_0$. $L_0$ thus provides a second natural length scale in addition to the step separation $L$ imposed by the miscut.

We now turn to the properties of sinusoidal steps like those in Fig. 1. Figure 2 shows the step contribution to the surface energy, on a surface with step separation $L=1000a$, corresponding to a miscut of 0.02°. Contours of constant energy are shown as a function of the amplitude $A$ and wavelength $\lambda$ of the undulations. Only contours with energy lower than that for straight steps are shown, so straight steps are unstable over the entire region within the outermost contour.

The behavior is surprisingly complex. There is a shallow local minimum in energy for straight steps, i.e., $A=0$. Increasing the amplitude raises the energy up to a

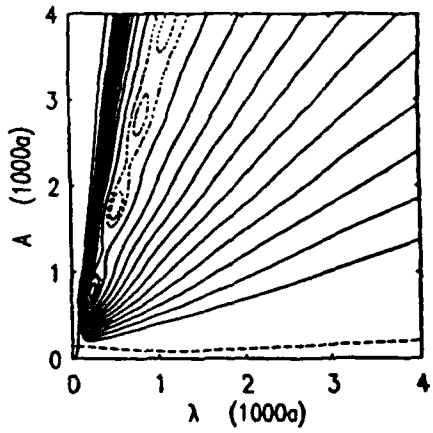



FIG. 2. Contours of constant step energy per surface area, as a function of amplitude $A$ and wavelength $\lambda$, for $L=1000a$. Only contours with energy lower than $-0.11$ meV/$a^2$, the energy for straight steps at this separation, are shown. There are a series of local minima along the line $\lambda\approx 0.3A$, and a weak local minimum along the line $A=0$. The energy has a local maximum with respect to $A$ along a ridge indicated by the dashed line. Successive contours differ in energy by 0.02 meV/$a^2$, with some supplemental dotted contours to better show the minima.

ridge indicated by the dashed line in Fig. 2. For still larger $A$, the energy drops, and a series of local minima are clearly seen, falling nearly along a line defined by $\lambda\approx 0.3A$.

To show the behavior along the minima more clearly, for each value of $A$ we minimize the energy with respect to $\lambda$, and plot the resulting energy and wavelength in Fig. 3. In all cases studied, the first minimum with respect to

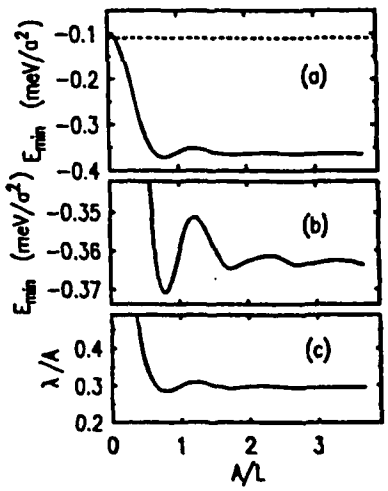



FIG. 3. Properties of steps on a surface with step separation $L=1000a$. For each amplitude $A$, the wavelength $\lambda$ is that which minimizes the energy. (a) Energy per surface area vs reduced amplitude $A/L$. Amplitude is scaled by $L$ to emphasize nearly perfect periodicity. Dotted line is energy for straight steps, for comparison. (b) Same as (a), on different scale to show oscillations. (c) Reduced wavelength $\lambda/A$ vs $A/L$.

$A$ was the deepest, but there were subsequent small oscillations about an asymptotic value. The period of these oscillations is simply the step spacing $L$, suggesting that the oscillations are due to a preference for having a specific alignment of the extrema of different steps. The ratio $\lambda/A$ remains virtually constant beyond the first minimum in Fig. 3, consistent with the nearly straight trough in the energy surface seen in Fig. 2. The approach to an asymptotic value represents an approximate scaling relationship: The energy depends primarily on $A/\lambda$, with corrections due to the discrete step structure with period $L$.

Finally, to obtain an overview of the behavior of the surface, we calculate the minimum-energy step shape for a range of $L$. In Fig. 4, we see that, for step separations of about $200a$ or less, the straight steps have lower energy. For separations less than $150a$, we could not even find a local minimum with respect to $\lambda$ and $A$. However, for step separations larger than $200a$, straight steps can lower their energy by developing undulations. At large step separations the energy appears to be approaching that of the minimum-energy surface, i.e., of the surface with step separation $L_0$.

The amplitude $A$ of the minimum-energy steps, shown in Fig. 5, scales nearly perfectly with $L$ as $A \approx 0.8L$, over the entire range $L > 200a$ where wavy steps are favored. The wavelength $\lambda$ actually decreases with increasing $L$, so $A/\lambda$ increases with $L$.

From Fig. 4, we see that the surface should undergo a phase transition with respect to angle of miscut, from a phase of straight steps to one of wavy steps. Surfaces with intermediate miscut should (if kinetics allow) facet into regions with $L \approx L_0$ and very flat regions of large $L$. Such faceting would still be compatible with step flow, and has apparently been observed by Tromp and Reuter [8]. Alternating up-and-down facets of miscut $L = L_0$ might have slightly lower energy; but like extra up-and-down steps, such up-and-down faceting would be incompatible with step flow.

We can get a semiquantitative understanding of the formation of step undulations in a rather simple way. In Fig. 1, we see that for large $A$ substantial portions of the surface are covered with nearly straight steps at a spacing much smaller than $L$. Intuitively, we expect that the step undulations form in order to decrease the step spacing to a value closer to the minimum-energy spacing $L_0$.

The length of wavy steps (composed on an atomic scale of rectilinear segments) is increased by a factor of $1+4A/\lambda$; so we can think of the characteristic step spacing as being reduced roughly by that factor to $L/(1+4A/\lambda)$. If we assume that the energy is minimized when this characteristic spacing approaches $L_0$, we would expect that $L/(1+4A/\lambda) \approx L_0$, i.e.,

$$A/\lambda \approx (L - L_0)/4L_0. \quad (3)$$

In Fig. 5(b), this linear relationship is included as a dotted line. (It appears as a curve due to the logarithmic scale.) The actual calculated results are seen to correspond rather well to the crude prediction (3), confirming our picture of the driving mechanism here.

Finally, it is important to address the limitations of the present study. Any inaccuracy in the stress anisotropy and in the local step energy $C_1$ simply changes the overall energy scale and the length $L_0$. A moderate change in the energy scale has no affect on our conclusions. While we have only studied one value of $L_0$, it is clear that step waviness should in general occur whenever $L$ becomes much larger than $L_0$. We also note that the value of $L_0$ here is fortuitously close to that inferred by Tromp and Reuter.

We have treated the case of no applied external strain. However, even a modest external strain can significantly affect the stress-domain patterns at small miscut [1,10]. Small strains can easily occur accidentally in experi-

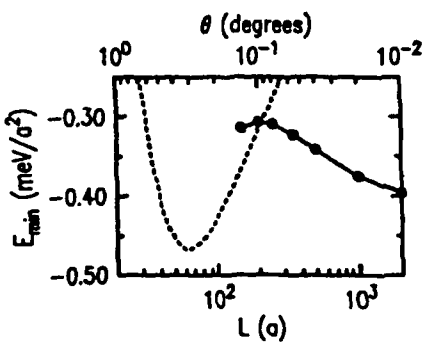



FIG. 4. Step energy per surface area vs step separation $L$ (bottom scale), or angle of miscut (top scale). Dots correspond to sinusoidal steps, whose amplitudes $A$ and wavelengths $\lambda$ are those which minimize energy; solid curve is a spline fit to guide the eye. Dotted curve is the corresponding energy for straight steps.

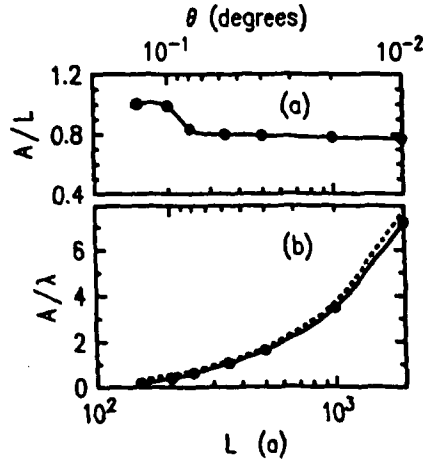



FIG. 5. Properties of minimum-energy steps vs step separation $L$ (bottom scale), or angle of miscut (top scale). Each point corresponds to a point in Fig. 4. (a) Amplitude as fraction of $L$. (b) Dimensionless amplitude $A/\lambda$, along with linear relationship of Eq. (3) (dotted line).

ments, and so should be considered before attempting a detailed comparison with measurements.

We have also assumed a specific shape for the steps, based on analytic convenience. Thus our variational calculation in the parameters $\lambda$ and $A$ actually provides an upper bound on the energy of the wavy phase. This is enough to guarantee our central result, the instability of straight steps. Moreover, the assumed sinusoidal shape is physically reasonable, and is qualitatively consistent with experimental observations [8]; so it seems highly unlikely that a more accurate shape would greatly affect the overall behavior, see Fig. 4.

It would certainly be of interest to determine the actual step shape which minimizes the energy. In particular, meandering of the $S_B$ steps is favored, since it creates segments of $S_A$ step, which are believed to have rather small local energy. Meandering of $S_A$ steps creates higher-energy $S_B$ segments. Thus we expect $S_B$ steps to have undulations of larger amplitude. Aside from the shapes of the individual steps, more complicated patterns of steps are possible, which would not repeat every two steps. Also, the presence of "kissing site" defects [11], associated with antiphase boundaries in the dimerization, appears to cause significant deviations from ideal behavior [8].

The only apparent discrepancy between theory and experiment [8] is the failure to observe the predicted large values of $A/\lambda$ for large $L$. Step flow kinetics would tend to suppress large $A/\lambda$, especially given the rather weak dependence of energy on $\lambda$ in Fig. 2. Large $A/\lambda$ could also be disfavored if the step undulations are coherent only over small patches, as in the experiments, due to defects. And at very large $L$, even small external strains could affect the results.

Spontaneous step formation [1] has the advantage that only the low-energy $S_A$ steps are created. However, even if this should prove to be the structure of lowest energy, it might not be kinetically accessible. A sequence of $S_A$ steps necessarily has an up-and-down pattern. Step flow would quickly eliminate such steps, leaving only the monotonic sequence of steps associated with the miscut. Such step flow occurs not only while growing by vapor deposition, but also during sublimation while heat cleaning the surface [8]. At temperatures low enough to suppress sublimation, the energetic barrier to spontaneous step formation might be prohibitive.

In contrast, step waviness can reduce the elastic energy without interfering with step flow. And the phenomena observed experimentally [8] are all in accord with this picture, including step flow during high-temperature cleaning, coherent step undulations over large areas, and apparently even faceting into regions of more closely spaced straight steps and widely spaced wavy steps. Thus very flat Si(001) surfaces provide a window onto an unexplored regime with a wealth of fascinating new phenomena.

We are grateful to R. M. Tromp for communicating results prior to publication, and for valuable discussions. This work was supported in part by ONR Contract No. N00014-84-C-0396.

---

[a]Current address: Fritz-Haber-Institut, Faradayweg 4-6, D-1000 Berlin 33, Germany.

[1] O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopoulos, Phys. Rev. Lett. 61, 1973 (1988).

[2] O. L. Alerhand, A. N. Berker, J. D. Joannopoulos, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, Phys. Rev. Lett. 64, 2406 (1990).

[3] T. W. Poon, S. Yip, P. S. Ho, and F. F. Abraham, Phys. Rev. Lett. 65, 2161 (1990).

[4] E. Pehlke and J. Tersoff, Phys. Rev. Lett. 67, 465 (1991); 67, 1290 (1991).

[5] D. J. Chadi, Phys. Rev. Lett. 59, 1691 (1987).

[6] V. I. Marchenko and A. Y. Parshin, Zh. Eksp. Teor. Fiz. 79, 257 (1980) [Sov. Phys. JETP 52, 129 (1980)].

[7] The possibility that straight steps might be energetically unstable was previously suggested by X. Tong and P. A. Bennett, Phys. Rev. Lett. 67, 101 (1991). However, they had in mind much smaller step separations, where the effect discussed here does not occur.

[8] R. M. Tromp and M. C. Reuter, following Letter, Phys. Rev. Lett. 68, 820 (1992); R. M. Tromp (private communication).

[9] R. D. Meade and D. Vanderbilt, in *Proceedings of the Twentieth International Conference on the Physics of Semiconductors*, edited by E. M. Anastassakis and J. D. Joannopolous (World Scientific, Singapore, 1990), p. 123; M. C. Payne, N. Roberts, R. J. Needs, M. Needels, and J. D. Joannopoulos, Surf. Sci. 211/212, 1 (1989).

[10] F. K. Men, W. E. Packard, and M. B. Webb, Phys. Rev. Lett. 61, 2469 (1988).

[11] B. S. Swartzentruber, Y. W. Mo, and M. G. Lagally, Appl. Phys. Lett. 58, 822 (1991).

# Equilibrium Alloy Properties by Direct Simulation: Oscillatory Segregation at the Si-Ge(100) 2×1 Surface

P. C. Kelires[a] and J. Tersoff

*IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598*

(Received 6 February 1989)

We study surface and bulk equilibrium in Si-Ge alloys by direct simulation. The composition at a reconstructed (100) surface varies with depth in a complex oscillatory way. Lateral ordering occurs even in the fourth layer, driven by the local stress field. The bulk phase diagram is well described by regular solution theory.

PACS numbers: 68.35.Dv, 61.55.Hg, 64.75.+g

Theoretical understanding of the equilibrium properties of semiconductor alloys has progressed rapidly in recent years,[1] spurred in part by the increasing importance of such alloys in electronic devices. However, until now these studies have been restricted to homogeneous bulk systems; and the theoretical methods employed, such as the cluster-variation method[2] (CVM), offer little immediate prospect of going beyond such systems.

Because of the important role of reconstructed surfaces in modern epitaxial growth techniques, it is particularly desirable to understand how the properties of alloys are modified in such inhomogeneous environments. Here we report what is apparently the first calculation of equilibrium segregation at a semiconductor surface, using a variant of the grand-canonical Monte Carlo method introduced by Foiles[3] to study metal surfaces.

We find surprising results for the 2×1 dimer reconstruction of the (100) surface of an Si-Ge alloy. Strong *oscillatory* variations with depth are found in the equilibrium composition profile near the surface. Even the fourth layer shows striking deviations from bulk behavior, with a marked inequivalence between the two atoms in the unit cell, which reflects the strain induced by the reconstruction.

The method used here is a type of direct simulation. Simulations which allow each site to be either Si or Ge, with an Ising-type Hamiltonian, are quite standard; but the evaluation of the appropriate effective interactions, incorporating the effects of strain implicitly, has only been feasible in the simple bulk situation. On the other hand, methods such as molecular dynamics (MD), which permit arbitrary atomic displacements, and which incorporate strain explicitly, cannot, in practice, reach equilibrium for solid solutions because of the large barriers to atomic diffusion and the short simulation times which are feasible.

Foiles[3] pointed out that the advantages of both approaches could be combined by using a continuous-space Monte Carlo (MC) algorithm, which incorporates two kinds of MC "moves." The simulation includes small random atomic displacements, as well as moves which convert Si atoms into Ge and vice versa, allowing compositional equilibration. This approach may be viewed as a specialized case of the grand-canonical Monte Carlo method, which has been extensively discussed.[3]

An equilibrium distribution is obtained in the usual way, accepting trial moves with a probability

$$\exp[(\mu_{\mathrm{Si}}\Delta n_{\mathrm{Si}}+\mu_{\mathrm{Ge}}\Delta n_{\mathrm{Ge}}-\Delta U)/kT]$$

(but not more than 1), where $\Delta U$ is the change in potential energy due to the move, $\mu$ is the chemical potential for a given species, and $\Delta n$ is the change in the number of atoms of that species. The total number of atoms remains fixed, so only the difference $\Delta\mu=\mu_{\mathrm{Si}}-\mu_{\mathrm{Ge}}$ is relevant.

In order to make the simulation tractable, we use an empirical interatomic potential[4] to model the interactions. Although less accurate than state-of-the-art quantum-mechanical calculations, this approach has been extensively tested.[4-6] It is well suited to the present problem because it describes both the surface dimerization and the elastic properties reasonably well.

Before addressing the surface problem, we illustrate the method by considering the bulk Si-Ge phase diagram, which has also been treated recently by Qteish and Resta.[7] Since the primary driving force for segregation here is the atomic size mismatch, this problem provides a stringent test of our ability to equilibrate both the spatial and chemical degrees of freedom simultaneously and consistently.

We map out the phase boundaries in the natural way. For each temperature, we equilibrate a periodically repeated cubic cell of 216 atoms, at zero pressure, over a range of values of the chemical-potential difference $\Delta\mu$. The presence of a first-order transition with $\Delta\mu$ indicates a miscibility gap at that temperature, and the alloy compositions just before and after the transition represent the miscibility limits. Such a calculation is illustrated in Fig. 1.

Because of the finite cell size, the cell can fluctuate between Si-rich and Ge-rich phases. As a result, the transition is broadened, and the average cell composition is, strictly speaking, a continuous function of $\Delta\mu$ at all tem-

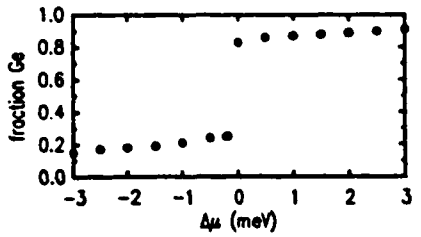



FIG. 1. Alloy composition vs chemical-potential difference (relative to an arbitrary origin) at 150 K.

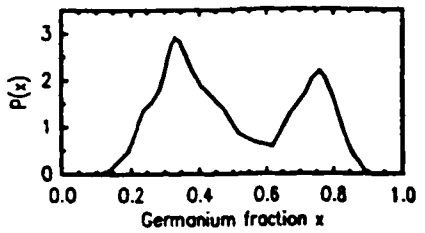



FIG. 2. Probability $P(x)$, in arbitrary units, for the simulation cell to have an instantaneous Ge fraction $x$, plotted vs $x$. Curve shown is for 160 K, at a value of $\Delta\mu$ around the transition value.

peratures. As seen in Fig. 1, the cell used here is large enough that the transition may be quite abrupt. Nevertheless, just below the minimum temperature for complete miscibility, $T_c$, such fluctuations can interfere with an accurate identification of the miscibility limits, or even of the existence of a transition.

In order to avoid this problem, we examine not only the average cell composition, but the probability distribution for this composition. As shown in Fig. 2, for a temperature just below $T_c$, this distribution is bimodal, as the cell fluctuates between phases. At a value of $\Delta\mu$ where the cell will be found in either of the two phases with nearly equal probability, the average compositions of the respective metastable phases give very good estimates of the miscibility limits. Moreover, the mere existence of such a bimodal distribution confirms that the temperature is below $T_c$. Above $T_c$, however broad the distribution, it should be unimodal.

Detailed simulations as a function of $\Delta\mu$ at a series of temperatures result in the phase diagram shown in Fig. 3. There is a small but systematic asymmetry in the phase diagram. Extrapolating by eye gives $T_c$ around 165 or 170 K.

For comparison, regular solution theory predicts[8] that $T_c = 2\Delta H/k$, where $\Delta H$ is the enthalpy of mixing per atom for the 50-50 alloy, and $k$ is the Boltzmann constant. Since the interatomic potential used here gives[4] an enthalpy of mixing (for the perfectly random alloy at $T=0$) of 7.3 meV/atom, $T_c$ is predicted to be 170 K.

Thus the simulation results are in excellent agreement with regular solution theory, as expected for this nearly ideal solution. This agreement, in fact, provides a strong test of the accuracy of the MC equilibration. If the atomic positions did not relax sufficiently in response to atom switching, the effective mixing enthalpy (and hence $T_c$) would be much higher.

The enthalpy of mixing here is about 30% lower than that calculated by Qteish and Resta.[7] This difference may be due to inaccuracies in our empirical potential, such as the failure to describe the observed[9] "bowing" of the elastic constants with alloy composition. On the other hand, it may be due to the fact that fully relaxed geometries are used here, whereas only partial relaxation was included in Ref. 7. In either case, the associated uncertainty is only 2-3 meV/atom, and so will prove unimportant in the surface studies below, where the relevant energy scale is an order of magnitude larger.

Alloy surfaces have been extensively studied for metals,[3,10,11] but not for semiconductors. Experimentally, it is now possible to measure the composition site by site, using atom-probe techniques or low-energy electron diffraction.[10] Not only is enrichment of one constituent generally observed at metal surfaces, but in some cases [e.g., Pt-Ni or Pt-Rh(100)] the composition varies nonmonotonically with depth.[10] Theoretical approaches[11] for metal surfaces (other than that of Ref. 3) have generally not included the local strain effects associated with atomic size mismatch or surface reconstruction, and so are not applicable to the case of Si-Ge alloys.

The Monte Carlo approach described above is now applied to the problem of a semiconductor surface. The simulation cell used here is a 24-layer (100) slab, with 288 atoms per cell (12 per layer) periodically repeated in two dimensions. The two surfaces are prepared in the $2\times1$ dimer reconstruction,[12] which is known to occur (with minor variations) for both Si and Ge(100).

We begin by considering a relatively low temperature, 300 K. It is presumably not possible experimentally to equilibrate the alloy at such a low temperature, but this case provides a natural starting point for discussing higher temperatures.

For simplicity, the surface lattice constant is fixed at the pure Si value. Since Si-Ge alloys are often grown

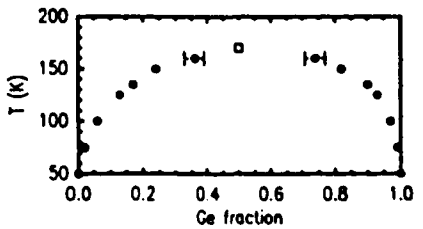



FIG. 3. Calculated phase diagram of Si-Ge alloy. Solid circles are pointed on the phase boundary, below which segregation occurs. Statistical error bars are negligible except where shown. The open square is the result of regular solution theory, based on calculated enthalpy of mixing for the perfectly random 50-50 alloy at $T=0$.

epitaxially on Si substrates, this case is as physically relevant as that in which the alloy takes its natural lattice constant. In any case, tests indicate that, at temperatures above $T_c$, such small variations in the surface lattice constant have little effect. At lower temperatures, the epitaxial constraint suppresses segregation,[1,13] permitting us greater freedom in the choice of alloy composition. A convenient value of $\Delta\mu$ is used here, which yields (with the constrained surface lattice constant) a bulk alloy composition of about 50% Ge.

The results of the simulation are summarized in Fig. 4, which shows the site-by-site composition in equilibrium. The most obvious effect is the strong segregation of Ge to the surface. Ge has a lower surface energy than Si, about 0.07 eV/atom lower for the (100) 2×1 with this potential, so segregation of a layer of Ge to the surface reduces the enthalpy.

A surprising feature of the results is that the Ge concentration is strongly *reduced* in the second layer, relative to the bulk. Thus the concentration profile at the surface is oscillatory. Stranger still, the third and fourth layers show strong deviations from bulk composition, tending towards one Si and one Ge on the respective sites of the 2×1 cell at low temperatures.

The striking behavior of these deeper layers can be easily understood, using the concept of an atomic stress tensor,[14] or more specifically its trace, which defines a local compression. Heuristically speaking, some atoms may be viewed as under compression, if their bonds are shorter than the sum of covalent radii, while other atoms are under tension. This can be quantified by considering a uniform expansion of the system. Then by analogy with the macroscopic pressure, we define an atomic compression

$$p_i = -dE_i/d\ln V, \tag{1}$$

where $E_i$ is the energy of atom $i$, and $V$ is the volume. This compression can be converted into units of pressure by dividing by an appropriate atomic volume.

The decomposition of the total energy into atomic contributions is, in principle, not unique, but may often be made in practice. In particular, this decomposition is specified explicitly in the definition of the interatomic potential used here.[4] The value of such a decomposition in understanding atomic-scale behavior of complex systems has already been demonstrated.[15]

Examining the (100) 2×1 surface of pure Si in this way, we find that the second layer is under a large compression, about 0.4 eV/atom, corresponding roughly to a pressure of 30 kbar. Substituting a Ge atom, which is larger than Si, would obviously tend to raise the energy of site under compression, but would lower the energy of a site under tension.

Since the logarithm of the volume ratio of bulk Ge and Si is 0.12, the above compression can be converted into an estimate of the energy gained or lost by substituting Ge for Si, by multiplying the compression (1) by 0.12. For the second layer, this yields an estimate of 0.04 eV/atom, which is very significant on the scale of thermal energies.

In the fourth layer, the two atoms per cell are inequivalent. One atom is directly below the dimer, and is under a compression of about 0.3 eV, while the other is between dimers, and is under a tension of similar magnitude. Thus the former site is driven towards being pure Si at low temperatures, and the latter towards pure Ge. A similar effect is seen in the third layer, but is a bit weaker.

For the surface dimer layer, the compression is very weak, consistent with the relatively unconstrained geometry. The surface composition is driven by the reduction of surface energy associated with the dangling bond, and not by atomic compression.

Understanding the room-temperature equilibrium structure, and the effects which cause it, we now wish to get an overview of the behavior with increasing temperature. For this purpose we adopt a somewhat simpler approach. Instead of calculating the chemical potential at each temperature for the bulk composition of interest, we fix the number of Ge and Si atoms. Only moves which switch a Ge and a Si simultaneously are considered, thus permitting diffusion while conserving the cell composition.

Figure 5 shows the resulting site-by-site composition, for the first four layers of a slab with 50-50 composition,

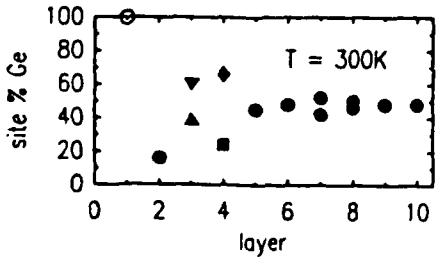



FIG. 4. Composition vs layer number, for fixed slab composition at 300 K, as described in text. For layers with inequivalent sites, both sites are shown.

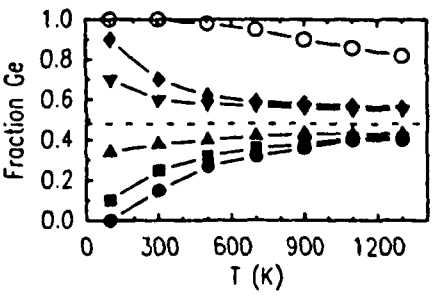



FIG. 5. Composition of individual sites vs temperature, for fixed slab composition, as described in text. As in Fig. 4, open circles are surface layer, filled circles are second layer, triangles and inverted triangles are third layer, and diamonds and squares are fourth layer.

as a function of temperature. The enhancement of Ge concentration at the surface remains strong even at high temperature because of the large lowering of the surface energy. Composition variations in deeper layers are reduced, but still moderately large. Even the lateral composition modulation in the fourth layer is almost 20%, easily large enough to observe experimentally, up to the highest temperatures considered, and probably up to the melting point.

In conclusion, we have demonstrated that it is feasible to calculate equilibrium properties of semiconductor alloys by direct simulation, even for the case of a reconstructed surface. This approach is applicable to a range of important problems, including segregation at semiconductor interfaces, grain boundaries, and other defects.

It is a pleasure to acknowledge discussions with D. P. DiVincenzo, P. Fahey, G. B. Stephenson, and A. Zunger. This work was supported in part by ONR Contract No. N00014-84-C-0396.

---

[a]Current address: Physics Department, University of Crete, 71409 Iraklion, Crete, Greece.

[1]See, for example, D. M. Wood and A. Zunger, Phys. Rev. Lett. **61**, 1501 (1988), and references therein.

[2]R. Kikuchi, J. Chem. Phys. **60**, 1071 (1974).

[3]S. M. Foiles, Phys. Rev. B **32**, 7685 (1985). For a review of the grand-canonical Monte Carlo method, see *Statistical Mechanics Part A: Equilibrium Techniques*, edited by B. J. Berne (Plenum, New York, 1977), Chaps. 4 and 5.

[4]J. Tersoff, Phys. Rev. B **39**, 5566 (1989). For additional background on this approach, see also J. Tersoff, Phys. Rev. Lett. **56**, 632 (1986); Phys. Rev. B **37**, 6991 (1988).

[5]J. Tersoff, Phys. Rev. B **38**, 9902 (1988).

[6]J. Tersoff, Phys. Rev. Lett. **61**, 2879 (1988).

[7]A. Qteish and R. Resta, Phys. Rev. B **37**, 1308 (1988); **37**, 6983 (1988).

[8]D. R. Gaskell, *Introduction to Metallurgical Thermodynamics* (McGraw-Hill, New York, 1973).

[9]V. T. Bublik, S. S. Gorelik, A. A. Zaitsev, and A. Y. Polyakov, Phys. Status Solidi (b) **66**, 427 (1974).

[10]M. Ahmad and T. T. Tsong, J. Vac. Sci. Technol. A **3**, 806 (1985), and references therein; Y. Gauthier, Y. Joly, R. Baudoing, and J. Rundgren, Phys. Rev. B **31**, 6216 (1985).

[11]F. F. Abraham, N. Tsai, and G. M. Pound, Surf. Sci. **83**, 406 (1979); Ph. Lambin and J. P. Gaspard, J. Phys. F **10**, 2413 (1980); R. N. Barnett, U. Landman, and C. L. Cleveland, Phys. Rev. B **28**, 6647 (1983).

[12]R. E. Schlier and H. E. Farnsworth, J. Chem. Phys. **30**, 917 (1959).

[13]J. L. Martins and A. Zunger, Phys. Rev. Lett. **56**, 1400 (1986).

[14]V. Vitek and T. Egami, Phys. Status Solidi (b) **144**, 145 (1987).

[15]P. C. Kelires and J. Tersoff, Phys. Rev. Lett. **61**, 562 (1988).

# Stress-induced layer-by-layer growth of Ge on Si(100)

J. Tersoff
*IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York 10598*
(Received 10 January 1991)

Several experiments have found that Ge initially grows layer by layer on the Si(100)2×1 surface, up to a thickness of 3 atomic layers. Further growth occurs via islands. Here, model calculations show that layer-by-layer growth is stabilized for up to 3 layers because it reduces the strain energy associated with the surface dimerization.

The growth of silicon, germanium, and their alloys from the vapor is surprisingly complex. Several fascinating issues have arisen recently involving steps and domains,[1,2] alloy ordering and surface segregation,[3–5] and island formation.[6–8] Intriguingly, all of these issues revolve around the role of surface stress and strain. Here we show that an outstanding puzzle in the growth of Ge on Si, the critical thickness of 3 Ge layers for the onset of island growth, is also attributable to the role of surface stress and strain.

Several experiments[6,7,9] have found that Ge grows layer by layer on the Si(100)2×1 surface, up to a thickness of 3 atomic layers. Further growth occurs via islands, which are initially coherent despite the 4% mismatch in lattice constants.[6,7] Such islands hinder the subsequent growth of sharp interfaces, e.g., for heterojunction devices or superlattices, so it is important to understand the forces which stabilize the desirable layer-by-layer growth for the first 3 layers.

The Ge islands which form after 3 layers exhibit interesting and unexpected behavior, such as coherent Stranski-Krastanow growth[6,7] and complex faceting.[10] However, here our concern is with the flat Ge film wetting the Si substrate between islands. We therefore consider the islands only as reservoirs of Ge, which determine the Ge chemical potential $\mu$. Our goal, then, is to determine the equilibrium film thickness as a function of $\mu$, and to identify the physical mechanism determining this thickness.

Some care is required in posing the problem of film thickness as one of equilibrium thermodynamics. In fact, the growth of Ge on Si is necessarily a nonequilibrium process, since in equilibrium on a substrate of pure Si, all the Ge would dissolve into the substrate. However, at typical growth temperatures of 500–700 °C bulk diffusion is negligible; so it seems reasonable to begin by ignoring intermixing between Ge film and Si substrate. This issue is discussed further below. There is still considerable surface diffusion above 500 °C, though, as indicated, e.g., by the motion of steps in response to stress.[11] Therefore, for sufficiently slow growth rates, an equilibrium will exist between the Ge film and the Ge islands, maintained by surface diffusion.

To determine the equilibrium film thickness, let $U_n$ denote the energy per 1×1 cell of a Si(100) substrate plus $n$ layers of Ge, terminated with the 2×1 dimer reconstruction. The energy required per atom to add an $n$th layer from a reservoir of Ge at chemical potential $\mu$ is $E_n-\mu$, where

$$E_n = U_n - U_{n-1}. \tag{1}$$

The system seeks to minimize its total energy including the reservoir, i.e., to minimize $U_n - n\mu$, so the condition for stability is that

$$E_n - \mu = 0. \tag{2}$$

(Entropy plays little role here as discussed below.) If $E_n < \mu$, a film of $n-1$ layers will grow to $n$ layers, while if $E_n > \mu$, a film of $n$ layers will shrink to $n-1$ layers.

Figure 1 shows our central result, $E_n$ vs $n$, for a modified Keating model.[12] The model and its motivation are described in detail below; but first we focus on the results, and their implications for film growth.

Given $\mu$, the equilibrium number of layers of Ge for this model can be read directly from Fig. 1 according to Eq. (2), by noting where the layer energy $E_n$ crosses the line $E_n = \mu$. So before going further, one must determine the appropriate range of $\mu$. For large Ge islands whose strain is almost fully relieved by misfit dislocations, the chemical potential approaches that of bulk Ge, which we choose as our reference value $\mu = 0$. At the opposite extreme, if the

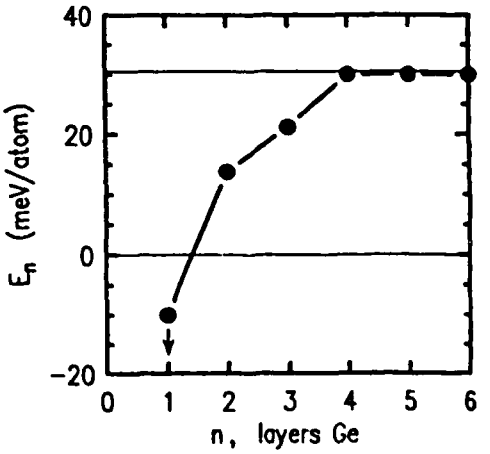



FIG. 1. Energy per atom $E_n$ to add an $n$th layer of Ge on a Si(100) substrate, calculated with a modified Keating model. The arrow schematically suggests a lowering of the energy, for the first layer only, by effects neglected in this model, as discussed in text. Horizontal lines show the physically appropriate range of chemical potential.

islands are wide and fully coherent, then the chemical potential approaches that of bulk Ge biaxially strained to the Si lattice constant, around 30 meV/atom; but $\mu$ for the island is generally below this due to elastic relaxation of the island.[6,8] These bounds on $\mu$ are shown as horizontal lines in Fig. 1; the true value for a given surface in equilibrium should lie somewhere in between.

In the early stages of growth, where islands are coherent,[6,7] $\mu$ approaches the theoretical upper bound. Then from Fig. 1, three layers are expected. This is precisely the regime in which 3 layers of Ge have been observed experimentally.[6,7,9] Thus the results of Fig. 1 account for the central experimental observation; the underlying mechanism is discussed in detail below.

For late-stage growth, where islands are large and presumably nearly free of strain, $\mu$ should approach the lower bound. In that case, from Fig. 1 wetting by only a single layer of Ge is predicted in equilibrium. Such single-layer films have not to my knowledge been reported. This may be simply because the film thickness in this regime of high nominal coverage has not been studied. However, in addition it may be difficult to attain equilibrium in this regime except by halting growth and annealing.

During growth which is not sufficiently slow for full equilibrium, the surface may contain a mixture of islands of different degrees of strain. For still more rapid growth, a significant number of isolated atoms or clusters may be present, which could in effect drive up $\mu$ beyond the upper bound of Fig. 1 (to the extent that it is meaningful to speak of $\mu$ having a value at all in this case). This could lead to a film thicker than 3 layers, but in such a regime one cannot escape the necessity of considering kinetics explicitly.

For the entire relevant range of $\mu$, one finds that the surface will be wetted by at least 1 layer of Ge. In fact, a surface dangling bond has lower energy cost for Ge than for Si, by perhaps 50 meV or more,[4] an effect neglected in the Keating model. So the first point in Fig. 1 should be displaced downward considerably, as suggested schematically by the arrow. Thus wetting is expected regardless of other details, simply because Ge has a much lower surface energy than Si. (The interface energy between Si and Ge is negligible on this scale.[4])

The energy differences between films of 1, 2, or 3 layers are of order 10 meV/atom, whereas at 600 °C the thermal energy $kT$ is 75 meV. Nevertheless, there should be little thermal fluctuation in thickness. The film thickness cannot vary without forming steps, and these are of too high energy to be thermally generated except with large terraces. But for a large terrace, the differences in energy between 1, 2, and 3 layers will be correspondingly large, suppressing thermal fluctuations.

At this point, we have seen that calculated energies of Ge films can account for the experimentally observed film thickness. The physical origin of this multilayer wetting can be understood by noting that there are two primary differences between Ge and Si that are relevant here. First, Ge has a larger lattice constant, and second, it has softer elastic moduli. To separate the contributions of these effects, the calculation of Fig. 1 is repeated twice: once changing the substrate lattice constant to that of Ge, so that the only difference between film and substrate is the smaller elastic moduli of the film; and once changing the substrate elastic moduli to be close to those of the film, so that the only difference is the larger equilibrium lattice constant for the film. These two cases are shown in Fig. 2 as diamonds and squares, respectively. The sums of the respective values are shown as open circles. These are quite close to the original values of Fig. 1, shown as filled circles, confirming that the film's strain energy can be unambiguously decomposed into these two contributing factors.

The difference in elastic moduli favors wetting by 2 layers of Ge, as seen from the diamonds in Fig. 2. This is easily understood. Because of the dimer reconstruction, the Si(100) surface is under considerable atomic-scale strain, especially in the first 2 layers.[13] Thus one can save energy by substituting a softer material in those layers. Deeper in the bulk, the strain is small and little energy is gained by a thicker Ge film.

The difference in atomic size gives a somewhat more complex effect, as seen from the open squares in Fig. 2. The most notable aspect is that the second layer is made less favorable for Ge, consistent with earlier suggestions that the second layer is under local compression.[4] On the other hand, Ge in the third layer is slightly favored, making 3 rather than 2 layers of Ge the equilibrium thickness in the upper range of $\mu$ when both factors are included. For thick films, each additional layer merely adds a layer of bulk strained Ge.

The conclusion to be drawn is that the primary effect leading to a 3-layer Ge film is the energy gained by having a softer material in the near-surface region, where the strains associated with the reconstruction are large. However, the coupling of the Ge size difference to the surface stresses is also significant, and without this effect the Ge film would be only 2 layers thick.

Having explained the experimental results and inferred the physical mechanism at work, we now return to the de-

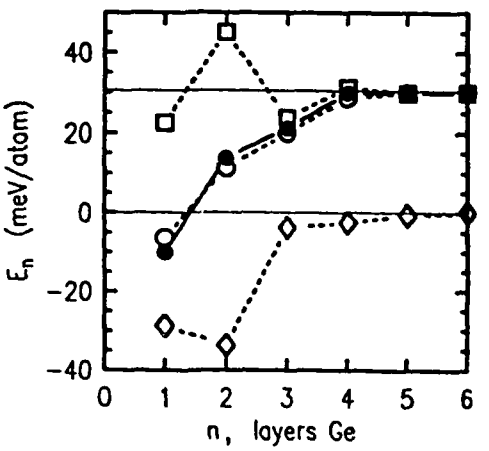



FIG. 2. Energy per atom $E_n$ to add an $n$th layer of Ge on a Si(100) substrate, as in Fig. 1 (solid circles), plus the same calculation for two related models: where the substrate has its lattice constant modified to be the same as Ge (diamonds), or its elastic constants modified to be like Ge (squares). Open circles are sum of diamonds and squares for each $n$.

tails of the model. In the present model, the energy is

$$E=\sum_i\left\{\sum_{j=1}^{4}\frac{\alpha_{ij}}{a_{ij}^2}(x_{ij}^2-\tfrac{1}{16}a_{ij}^2)^2+\sum_{j,k>j}^{4}\frac{2\beta_i}{a_{ij}a_{ik}}(\mathbf{x}_{ij}\cdot\mathbf{x}_{ik}-\tfrac{3}{16}a_{ij}a_{ik}\cos\theta_i)^2\right\}. \quad (3)$$

Here $\mathbf{x}_{ij}$ is the vector connecting atoms $i$ and $j$. Each pure material is described by three parameters: its lattice constant $a_i$, and its elastic parameters $\alpha_i$ and $\beta_i$. Thus these parameters take on only two values in Eq. (3), depending on whether atom $i$ is Si or Ge. For bonds between Si and Ge, the parameters are assigned the geometric mean of their elemental values: $a_{ij}=(a_ia_j)^{1/2}$, and $\alpha_{ij}=(\alpha_i\alpha_j)^{1/2}$. For the pure materials this is simply the familiar Keating model[12] for the elastic energy, if we take all $\theta_i$ as the tetrahedral bond angle $\cos^{-1}(-\frac{1}{3})$.

For the present calculation, this potential is modified relative to the Keating model in the following way. Recognizing that there is some rehybridization for the threefold coordinated surface atoms, which can have an important effect on surface stress,[14] we allow $\theta_i$ to take on a different value for these atoms, denoted $\theta_s$.

To determine the appropriate value of $\theta_s$, the surface stress for Si(100)2×1 is calculated as a function of $\theta_s$, and compared with parameter-free quantum-mechanical calculations of Payne *et al.*[15] and of Meade and Vanderbilt[16] using the local-density approximation (LDA) for correlation and exchange. The surface stress tensor is defined as

$$\sigma_{ij}^{\text{surf}}=\frac{1}{A}\frac{dE^{\text{surf}}}{d\varepsilon_{ij}}. \quad (4)$$

Here $E^{\text{surf}}$ is the surface energy, $A$ the surface area, and $\varepsilon$ is the two-dimensional strain. Thus a positive value corresponds to tensile stress. Table I gives results for $\sigma_\parallel$ and $\sigma_\perp$, the stress components parallel and perpendicular to the surface dimers. The average stress $(\sigma_\parallel+\sigma_\perp)/2$ is seen to be rather sensitive to the value of $\theta_s$.

Physically, one expects that the threefold surface atoms will have a tendency towards $sp^2$ bonding,[14] favoring more open bond angles (i.e., a more negative value of $\cos\theta_s$) and hence a more compressive stress. This is consistent with the fact that an unmodified Keating potential ($\cos\theta_s=-\frac{1}{3}$) gives much too tensile a stress, compared with the LDA calculations.[15,16] In fact, the results of Payne *et al.*[15] are reproduced fairly well by the Keating model with $\cos\theta_s=-\frac{2}{3}$, while those of Meade and Vanderbilt[16] are similar to the Keating model with $\cos\theta_s=-\frac{1}{2}$. Thus either of these values seems more realistic than the unmodified Keating potential.

In Fig. 3, the calculation of Fig. 1 is repeated for these three values of $\theta_s$. The results suggest that the stability of the 3-layer film is relatively insensitive to the choice of $\theta_s$, except that for very large values of $\theta_s$ ($\cos\theta_s\lesssim-\frac{2}{3}$) only a 1-layer film is stable. This is easily understood in terms of the results of Fig. 2 and of Table I. Large values of $\theta_s$ lead to a more compressive stress. This in turn exacerbates the compression in the second layer, making it more unfavorable for Ge. When this penalty outweighs the gain from having additional layers with softer elastic moduli, there is no energy lowering from film thickening beyond 1 layer. (This interpretation has been explicitly verified by repeating the calculation of Fig. 2 with $\cos\theta_s=-\frac{2}{3}$.)

Since the results depend somewhat on $\theta_s$, it seemed reasonable to choose this parameter to fit the LDA calculations. We chose[17] to fit the result of Meade and Vanderbilt, giving a value of $\cos\theta_s=-0.48$. However, as seen from Fig. 3, our conclusions remain valid for essentially all reasonable values of $\theta_s$, except at the extreme of large compressive surface stress.

Finally, the problem of intermixing of the film and substrate deserves more detailed consideration than we can give it here. There is a considerable driving force for interdiffusion, beyond the usual entropic considerations. First, intermixing lowers the strain energy. Second, in an alloy additional energy can be gained by arranging the Si and Ge so as to compensate for the stresses associated with the surface reconstruction. This effect can be rather large on the present energy scale, of order 30 meV/atom even in the fourth layer.[4] Thus any intermixing will considerably complicate the problem by allowing such effects to come into play. However, such intermixing is necessarily kinetically determined, and so is beyond the scope

TABLE I. Calculated surface stress for Si(100)2×1 surface, in eV/(1×1 cell), parallel and perpendicular to the dimers ($\sigma_\parallel$ and $\sigma_\perp$), and their sum and difference, which measure the net tension and anisotropy. Results are for the modified Keating model (see text), with various values of the parameter $\theta_s$, and for the LDA results of Refs. 15 and 16.

| | $\sigma_\parallel$ | $\sigma_\perp$ | $\sigma_\parallel+\sigma_\perp$ | $\sigma_\parallel-\sigma_\perp$ |
|---|---|---|---|---|
| Ref. 15 | 0.7 | −2.0 | −1.3 | 2.7 |
| $\cos\theta_s=-2/3$ | 0.9 | −2.1 | −1.2 | 3.0 |
| $\cos\theta_s=-1/2$ | 1.4 | −1.0 | 0.4 | 2.4 |
| $\cos\theta_s=-1/3$ | 1.7 | 0.1 | 1.8 | 1.6 |
| Ref. 16 | 1.6 | −0.9 | 0.7 | 2.5 |
| $\cos\theta_s=-0.48$ | 1.5 | −0.8 | 0.7 | 2.3 |

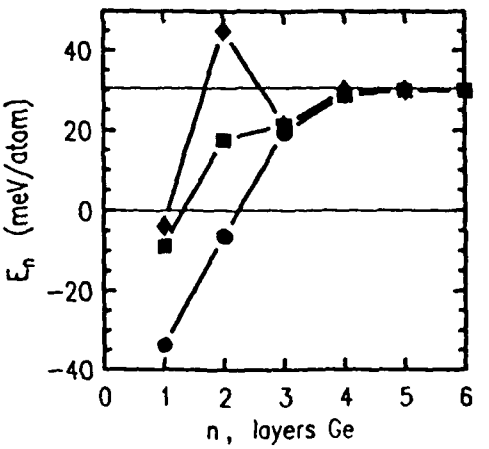



FIG. 3. Energy per atom $E_n$ to add an $n$th layer of Ge on a Si(100) substrate, as in Fig. 1, but for three different values of the parameter $\theta_s$: $\cos\theta_s=-\frac{1}{3}$ (circles), $\cos\theta_s=-\frac{1}{2}$ (squares), and $\cos\theta_s=-\frac{2}{3}$ (diamonds).

of our quasiequilibrium analysis.

Experimentally, Copel *et al.*[18] have observed that some intermixing occurs when Ge is deposited at 500 °C, but that intermixing is suppressed when the film is deposited at room temperature, even when subsequently annealed at 500 °C. This is consistent with the idea that there is considerable surface diffusion, but not bulk diffusion, at this temperature, since during growth intermixing can occur by what is essentially surface diffusion.

In conclusion, the energies of thin Ge films on Si(100)2×1 are reduced, relative to Si(100) plus biaxially strained bulk Ge, by having the strain from the surface dimerization fall in a material of smaller elastic moduli. In addition, the local stresses associated with surface dimerization favor having the larger Ge atoms in the third layer. The net result accounts for the observed 3-layer Ge film thickness in the initial stages of epitaxial growth. The results also predict that in true equilibrium between islands and film, when the islands become large and their strain is relieved by dislocations, the Ge film should shrink to a single atomic layer.

This work was supported in part by ONR Contract No. N00014-84-C-0396. Discussions with D. P. DiVincenzo and R. M. Tromp are gratefully acknowledged.

[1]O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopolous, Phys. Rev. Lett. **61**, 1973 (1988); O. L. Alerhand, A. N. Berker, J. D. Joannopolous, D. Vanderbilt, R. J. Hamers, and J. E. Demuth, *ibid.* **64**, 2406 (1990).

[2]T. W. Poon, S. Yip, P. S. Ho, and F. F. Abraham, Phys. Rev. Lett. **65**, 2161 (1990).

[3]A. Ourmazd and J. C. Bean, Phys. Rev. Lett. **55**, 765 (1985).

[4]P. C. Kelires and J. Tersoff, Phys. Rev. Lett. **63**, 1164 (1989).

[5]F. K. LeGoues, V. P. Kesan, S. S. Iyer, J. Tersoff, and R. Tromp, Phys. Rev. Lett. **64**, 2038 (1990).

[6]D. J. Eaglesham and M. Cerullo, Phys. Rev. Lett. **64**, 1943 (1990).

[7]F. K. LeGoues, M. Copel, and R. M. Tromp, Phys. Rev. B **42**, 11690 (1990).

[8]D. Vanderbilt and L. K. Wickham (unpublished).

[9]M. Asai, H. Ueba, and C. Tatsuyama, J. Appl. Phys. **58**, 2577 (1985); H.-J. Gossman, L. C. Feldman, and W. M. Gibson, Surf. Sci. **155**, 413 (1985).

[10]Y.-W. Mo, D. E. Savage, B. S. Swartzentruber, and M. G. Lagally, Phys. Rev. Lett. **65**, 1020 (1990).

[11]F. K. Men, W. E. Packard, and M. B. Webb, Phys. Rev. Lett. **61**, 2469 (1988).

[12]P. N. Keating, Phys. Rev. **145**, 637 (1966).

[13]J. A. Appelbaum and D. R. Hamann, Surf. Sci. **74**, 21 (1978).

[14]R. D. Meade and D. Vanderbilt, Phys. Rev. Lett. **63**, 1404 (1989).

[15]M. C. Payne, N. Roberts, R. J. Needs, M. Needels, and J. D. Joannopoulos, Surf. Sci. **211/212**, 1 (1989).

[16]R. D. Meade and D. Vanderbilt (unpublished).

[17]This more recent work includes refinements which are believed to make it more accurate than the earlier calculation [R. D. Meade (private communication)].

[18]M. Copel, M. C. Reuter, M. Horn von Hoegen, and R. M. Tromp, Phys. Rev. B **42**, 11682 (1990).