Do we really understand dislocations in semiconductors?

R. Jones

School of Physics, University of Exeter, Exeter, EX4 4QL, UK

Abstract

The properties of dislocations in Si, GaAs and GaN are reviewed. Although, theoretical investigations favour a reconstruction eliminating, or reducing, the electrical activity in each case, with a consequent increase in the barrier for dislocation motion, there are problems in reconciling these results with experiment. It may be that the influence of impurities or point defects cannot be overlooked.

I. INTRODUCTION

Theoretical investigations of dislocations in semiconductors were begun by Hornstra [1] who modelled the perfect 60° dislocation in Ge where the terminating plane lay on the widely separated layers. These dislocations possess a line of atoms with broken bonds and certainly would possess deep levels. Since the weak beam TEM images of Ray and Cockayne [2] found that dislocations in Si were dissociated, interest has switched to partial dislocations lying on the closely spaced glide planes. These, however, are strongly reconstructed and are electrically inactive except possibly for shallow electron and hole traps lying near the band edges [3-5]. Thus, any electrical effects correlated with these dislocations have to come from impurities, or intrinsic point defects, attached to the dislocation line.

One important property which would reflect the strong reconstruction of the core is the
dislocation velocity which is governed by the formation energy $F_k$ and migration energy $W$ of kinks. It seemed that the first calculations using simple interatomic potentials gave estimates of these quantities in reasonable agreement with the experimental values. But more recent \textit{ab initio} calculations have given results in conflict with the available experimental ones and this raises the question as to whether the fundamental ideas about the structure of the core are really correct, and whether impurities and intrinsic point defects can be ignored. This will be discussed in Section 1.

Dislocations in hexagonal GaN are quite unlike those in other III-V materials like cubic GaAs. For example, the screw dislocation in GaN [6] lies along [0001] with a full Burgers vector $b = c[0001]$, whereas in Si or GaAs [2,7,8], it lies along [110] and is dissociated into 30° partials with $b = a/6 \{2\{1\}$. The core structures of the dislocations are also quite different. In the case of GaAs, dislocations are believed to lie on the glide set of planes, as in Si, with a non-stoichiometric core structure consisting of two adjacent chains of chemically identical atoms. Calculations [9] indicate that, in the 90° partial, bonds between atoms of the same type are formed in a similar way to Si. Bonds between like atoms also arise in anti-site defects and these appear to have relatively low formation energies in GaAs, but not in GaN. The reconstruction between like atoms in GaAs is, however, relatively weak. This explains the lower activation energies for dislocation motion in GaAs than, say, in Si, and the reconstruction is easily broken or affected by impurities. On the other hand, the core of the screw dislocation in GaN is stoichiometric with a structure similar to that of the (10\overline{1}0) surface, where, relative to the ideal surface, Ga and N atoms assume positions determined by the transfer of charge from Ga to lone pairs on N. Such a relaxation leads to the removal of deep states from the band gap. Similarly, edge dislocations undergo a relaxation also eliminating deep states from the band gap. Thus, in spite of the differences in bonding, dislocation type and Burgers vector, the electronic structures of the cores are similar, with the elimination of deep states from the gap. However, the mobilities of the dislocations are very different.

Growth of GaN on sapphire by vapour phase epitaxy is often associated with the ap-
pearance of long nanopipes parallel to $c$ which have hexagonal cross sections with uniform diameters ranging from 50–250 Å [10]. The first suggestion was that they were the manifestation of screw dislocations with empty cores as discussed by Frank long ago [11]. However, recent calculations do not support the idea that the core of a screw dislocation with Burgers vector equal to $c$ is open with such a large diameter [12]. On the other hand, Liliental-Weber et al. [13] have found that the density of nanopipes is increased in the presence of impurities, e.g. O, Mg, In and Si and have argued that these impurities decorate the (10$ar{1}$0) walls of the nanopipes inhibiting overgrowth. This points to the importance of impurities for dislocations in this material.

We discuss in the next section the recent theoretical modelling of Si and then discuss dislocations in GaN where it is argued that impurities, and oxygen in particular, have strong interactions with dislocation cores.

II. DISLOCATIONS IN SILICON

In both Si and GaAs, the commonly occurring 60° and screw dislocations lie on $\{111\}$ planes and are dissociated into partials separated by an intrinsic stacking fault [2,7,8]. The dissociation reaction is:

$$\frac{a}{2}[110] \rightarrow \frac{a}{6}[1\overline{2}1] + \frac{a}{6}[2\overline{1}1].$$

(1)

The lowest energy partials are of the glide type [14]. This means that the core lies on the narrowly spaced $\{111\}$ planes and contains lines of atoms with dangling bonds lying almost in the $\{111\}$ glide plane giving rise to the possibility of bond reconstruction (Fig. 1). If the core lay on the widely spaced planes leading to shuffle dislocations, the dangling bonds would be normal to the glide plane and reconstruction leading to the pairing of bonds would be impossible. The two chemical species present in GaAs lead to two types of partial dislocations: the core atoms can be Ga, giving $\beta$ dislocations, or As ones leading to $\alpha$ dislocations.
A screw dislocation dissociates into two 30° partials: one of Ga ($\beta$) type and the other of As ($\alpha$) type. The 60° dislocation dissociates into a 30° and a 90° partial of the same type. In both partials, reconstructed bonds would have to be formed between atoms of the same type.

The dislocation core structures in Si and GaAs have been investigated using AIMPRO: an *ab initio* density functional cluster method [15,16,5,9]. The important conclusion, in agreement with previous work in the case of Si [3,4,17–21], is that the core bonds in Si, and the $\beta$ partial in GaAs, are strongly reconstructed with bond lengths comparable to crystalline values. The $\alpha$ core in GaAs, however, is more weakly reconstructed.

There are several implications arising from a strong reconstruction. Firstly, the dislocations are unlikely to be electrically active: the formation of a covalent bond between core
atoms will lead to a great separation of bonding and anti-bonding states or between filled and empty levels. Any states in the band gap are then likely to be shallow electron or hole traps. Secondly, the reconstruction might not occur if electrically active impurities present in the core possess extra or fewer electrons than the host atom they replace. For example, N or P in Si prefer to remain three-fold coordinated with the remaining valence electrons occupying a lone pair [15,16,22] and in GaAs, a bond does not form for example between a pair of Be impurities within the core of a \( \beta \) dislocation [16,9]. Hence, whenever these impurities diffuse to the core, they cause a change in the reconstruction possibly leading to a pinning effect.

The third implication is that bond reconstruction leads to low dislocation velocities as these bonds have to be broken for the dislocation to advance. Now, it is normally considered, but see below for a heretical view, that dislocations move according to the Hirth-Lothe theory [14] by the creation, diffusion and annihilation, of kinks. The magnitude of the dislocation velocity is then controlled by the kink formation energy, \( F_k \), and the barrier to kink migration \( W \) – assuming that there are no pinning points or strong obstacles along the dislocation line [14]. There is controversy over whether \( W \) is much greater than \( F_k \) [23–27] but strong reconstruction implies that the barrier to dislocation motion is considerable and the brittleness of the semiconductors is then understood to arise from the difficulty of breaking the reconstructed bonds. A recent review [28] gives experimental values of \( W_m \) to lie between 1.2 and 1.8 eV, and \( F_k \) to range between 0.4 and 0.7 eV.

Calculations using AIMPRO have given \( F_k \) and \( W \) to be 0.1 eV and 1.8 eV respectively [5]. The extremely low value of \( F_k \) was attributed to the influence of the cluster surface. However, since these calculations there have been several others which have also produced very low values for \( F_k \). A recent tight binding model [29] has suggested that the ground state structure of the 90° partial consists of a line with periodicity twice that in Fig. 1. However, the energy difference is very small and this result is controversial [30]. The kink formation and migration energies are only 0.12 eV and 1.62 eV respectively. A more recent paper, [31] using a full *ab initio* method gave even smaller values of \( F_k \) of 0.04 eV and \( W \) to
be about 1.1 eV. Clearly, these results are in conflict with the experimental ones.

There are, however, difficulties with the simple Hirth-Lothe model for the dislocation velocity. When the length $L$ of a dislocation segment is less than the average separation between kinks, then the velocity should be controlled by the barrier to form a new double kink and becomes linear in $L$. The activation energy for the dislocation velocity should then switch from $F_k + W$ to $2F_k + W$. Experiments carried out on Si$_{0.9}$Ge$_{0.1}$ epilayers of varying thicknesses from 2500 Å up to $\sim 1$ μm demonstrate that, although the velocity clearly possess a linear length dependence for $L < \sim 1$μm, the activation energy for the dislocation velocity is the same, 2.2 eV, for short and long dislocation segments [32,33].

The critical length, which is related to the kink formation energy, is then at least 1μm and hence $F_k > 0.5$ eV. The authors suggest that there are strong obstacles (jogs?) present with this separation and so the dislocation velocity for long segments is always controlled by the nucleation of double kinks on a pinned segment of length less than the mean kink separation. Thus the activation energy is always $2F_k + W$. If this equals the experimental value of 2.2 eV, then $W$ must be less than 1.2 eV. This model can explain the magnitude of the dislocation velocity which would otherwise be attributed to a very large entropy term. The nature of the strong obstacles separated by $\sim 1$μm is not known.

Recently, a new TEM imaging technique has been developed which allows the stationary and moving dislocation line to be observed enabling $F_k$ and $W$ to be found at specific temperatures [34,35]. These observations showed that kinks on 90° partials were three times as numerous as on 30° partials at 600°C. The authors concluded that for the former partial, $F_k$ is 0.73 eV and is lower by only 0.07 eV (at 420°C) than $F_k$(30). An analysis of the movement of 90° segments showed that $W$ at 130°C is 1.24 eV and 1.7 eV at 600°C. This difference is unlikely to be due to a ‘negative’ entropic contribution to $W$ but is not apparently related to obstacles. Obstacles, or traps for kinks, are observed and separated by about 100 Å at 600°C and release kinks with an activation barrier of 2.4 eV. The velocity of the dislocation segment is then controlled by kinks released by the obstacles and not those thermally generated on the line. Thus the Hirth-Lothe theory is not appropriate and the
depinning of kinks from obstacles control the velocity. This would explain both Maeda’s experiments on the length dependence of the dislocation velocity described above but leaves unexplained the difference between the theoretical and experimental values of $F_k$. A curious feature, is that $F_k \sim 0.7 \text{eV}$ implies that the mean separation between kinks in thermal equilibrium should be at least $1 \mu\text{m}$, but the published TEM picture show kinks separated by only perhaps $100 \text{Å}$.

The high density of obstacles suggests that they are dragged along with the dislocation possibly increasing in density. This must lead to an aging effect although this has not been reported.

Tight binding calculations for the $30^\circ$ partial show that there is a number of configurations for kinks but the lowest has formation energy of $0.35 \text{ eV}$ and a barrier to movement of around $1.5 \text{ eV}$ [29]. Reconstruction defects, or solitons [36], have energies about $1.3 \text{ eV}$ with a low migration barrier of $0.3 \text{ eV}$.

It seems to us, that it is unlikely that the formation energies of the kink on both partials can be so similar unless they possess very similar structures. This cannot be explained by the reconstructed models and we must look to impurities or intrinsic defects to explain the results. One possibility is that oxygen both forms pinning points and prevents kink formation. It is known that oxygen diffuses in bulk with an energy $2.5 \text{ eV}$ but an oxygen dimer can diffuse more rapidly with a barrier around $1.8 \text{ eV}$ [37]. Also, oxygen is likely to be bound to the dislocation line [38] taking advantage of the dilated bonds there. A high density of oxygen dimers can thus be expected along the dislocation core which might be mobile. However, once dimers have aggregated then presumably, they will pin the dislocation line as discussed previously [39,38]. It is tempting to argue that oxygen will lead to an increase in the dislocation activation energy and must be taken into account although a detailed analysis is lacking. A difficulty with this interpretation is that oxygen is not present in SiGe alloys. It may be that interstitial or vacancy clusters form the obstacles detected in the TEM study.

There is evidence that the direct involvement of impurities can increase the dislocation
velocity. The behaviour of dopants is well known [9]; less well studied is the case of hydrogen. The effect of hydrogen introduced by a plasma into plastically deformed Si has been considered by Yamashita et al. [41]. They find the activation energy for the velocity of 60° dislocations to drop from 2.2 eV to 1.2 eV below 540°C. A nitrogen plasma had no effect. It appears that the effect is only observed when the hydrogen is introduced below about 470°C. Above this temperature, no effect is detected. This temperature is close to that when platelets break up either with the loss of hydrogen from the sample or driving it to more stable defects [42]. The most stable hydrogen defect known is V_2H_6 which disappears above 600°C [43]. Recent calculations suggest that H or H_2 readily attack the reconstructed dislocation leading to the formation of hydrogenated solitons which preferentially nucleate mobile soliton-kinks leading to an enhanced velocity [44].

![Diagram of a screw dislocation in GaN](image)

**FIG. 2.** Top view (in [0001]) of the open-core screw dislocation in GaN. The three fold coordinated atoms 1 (Ga) and 2 (N) adopt a hybridization similar to the (1010) surface atoms.

### III. DISLOCATIONS IN GALLIUM NITRIDE

Screw dislocations in GaN are not dissociated and have a Burgers vector equal to \( \mathbf{c} \) and thus lead to a considerable strain in the core. The screw dislocation with a full core [45] possesses severely strained bond lengths distorted by as much as 0.4 Å. Consequently it
is not surprising that such dislocations possess deep gap states ranging from $E_v + 0.9$ to $E_v + 1.6$ eV and shallow states around $E_c - 0.2$ eV. The energy of the dislocation dipole within its unitcell can be reduced if the atoms with the most distorted bonds are removed from the core and inserted into the bulk leading to a screw dislocation dipole with empty cores as shown in Fig. 2.

The atoms on the walls of the core (Fig. 3) adopt three fold coordination similar to those found on the (10\textup{[110]}\textup{0}) surface. Thus Ga (N) atoms develop $sp^2(p^3)$ hybridisations which lower the surface energy and clean the gap [46]. However, in contrast to the (10\textup{[110]}\textup{0}) surface, the open core dislocation appears to possess shallow gap states.

In the open core dislocation, bonds are distorted less than in the full core and by only about 0.1 Å. The calculated line energy is also lower at 4.55 eV/Å. Removing further material from the core resulted in a higher line energy and thus we conclude that the equilibrium diameter is about 7.2 Å. This is very much less than the diameter of nanopipes and we suppose that their formation is not due to Frank’s mechanism.

Threading edge dislocations were modelled by relaxing clusters containing a dislocation with Burgers vector $a[\text{11\textup{[20]}\textup{]}]/3$ as well as a supercell containing a dislocation dipole. The relaxed core is shown in Fig. 4. It consists of a line of Ga and N atoms which have 3-fold coordination and the other core atoms possess bonds strained by less than 0.1 Å. The three-fold coordinated Ga (N) atoms (labelled 1 and 2 in Fig. 4) move in such a way to enhance $sp^2$ and $p^3$ hybridisation respectively. This leads to empty Ga lone pairs pushed towards $E_c$, and filled lone pairs on N atoms lying near $E_v$, in a manner identical to the (10\textup{[110]}\textup{0}) surface. Thus threading edge dislocations are then also electrically inactive except possibly for shallow levels. This geometry for the core has recently been confirmed experimentally using Z–contrast imaging techniques [47].
FIG. 3. Projection of the wall of the open-core screw dislocation.

FIG. 4. Top view (in [0001]) of the relaxed core of the threading edge dislocation. The three fold coordinated atoms 1 (Ga) and 2 (N) adopt a hybridization similar to the (1010) surface atoms.

To summarise, the density functional calculations reveal that the common threading screw and threading edge dislocations in wurtzite GaN are electrically inactive. A comparison of the calculated line energies shows that screw dislocations exist as open-core dislocations whereas edge dislocations are filled. However, the strained and ‘dangling’ bonds present in their cores could permit impurities and intrinsic defects to be trapped there.

There is experimental evidence that oxygen acts as a donor in bulk GaN [48] and total energy calculations show that O sits on a N site [49]. Since the internal surfaces of screw dislocations are very similar to those of the low energy (1010) surface, we investigated [50] the likely surface sites for oxygen replacing N atoms. We found that the energy of a neutral
O\textsubscript{N} defect is 0.8 eV lower at the relaxed (10\overline{1}0) surface. This shows that there is a tendency for O to segregate to the surface. The added oxygen has an additional electron occupying a state near \(E_c\). The defect has therefore a high energy and would attract acceptors resulting in a neutral complex. One possible acceptor, other than added dopants, would be a gallium vacancy (\(V\textsubscript{Ga}\)) which acts as a triple acceptor and has been calculated to have a low formation energy in \(n\)-type GaN \[51,52\]. Consequently, we suppose that the surface oxygen concentration could be sufficiently large, and the oxygen atoms sufficiently mobile, that the three N neighbours of \(V\textsubscript{Ga}\) at the (10\overline{1}0) surface are replaced by O forming the \(V\textsubscript{Ga}-(O\textsubscript{N})\textsubscript{3}\) defect.

Our calculations \[50\] showed that this defect is more stable at the surface than in the bulk by 2.15 eV. Two O neighbours of the surface vacancy lie below the surface and each is bonded to three Ga neighbours, but the surface O is bonded to only two subsurface Ga atoms in a normal oxygen bridge site. The defect is electrically inactive with the O atoms passivating the vacancy in the same way as \(V\textsubscript{H}\textsubscript{4}\) in Si.

The question then arises as to the influence of this defect on the growth of the material. Growth over the defect must proceed by adding a Ga atom to the vacant site but this leaves three electrons in shallow levels near the conduction band resulting in a very high energy. This suggests that the defect can stabilise the surface and thus inhibit growth. From this we can conclude that such defects lead to the formation of nanopipes if we assume that during growth of the epilayers, either nanopipes with very large radii are formed which gradually shrink when their surfaces grow out, or there is a rapid drift of oxygen to a preexisting nanopipe. In either case the concentration of oxygen and \(V\textsubscript{Ga}-(O\textsubscript{N})\textsubscript{3}\) defects increases at the walls of the nanopipe. The maximum concentration of this defect would be reached if 50% (100%) of the first (second) layer N atoms were replaced by O and further growth then would be prevented. It is, however, likely that far less than the maximum concentration is necessary to stabilise the surface and make further shrinkage of nanopipe impossible. Provided oxygen could diffuse to the surface fast enough, the diameter and density of the holes would be related to the initial density of oxygen atoms in the bulk. This model requires
that the walls of the nanopipe are coated with oxygen although the initial stages of formation of the pipe are obscure.

It is also necessary to explain why the tubes have (10\bar{1}0) walls. $V_{Ga-O_{3N}}$ defects are also expected to be stable on stoichiometric (11\bar{2}0) surfaces, but nanopipes with these walls do not arise. We presume this is because of their higher absolute surface energies [46]. Other surfaces, e.g. (0001), (10\bar{1}1), are non-stoichiometric and probably possess small band gaps (or are metallic) and hence screen the attraction between the oxygen donor and the $V_{Ga}$ acceptor, preventing their initial pairing. Thus these defects are less likely to form on these surfaces and their growth is not inhibited.

In conclusion, oxygen clearly has a tendency to segregate to the (10\bar{1}1) surface and to form stable and chemically inert $V_{Ga-O_{3N}}$ defects. These defects increase in concentration when the internal surfaces grow out. When a critical concentration of the order of a monolayer is reached, further growth is prevented. This model leads to nanopipes with (10\bar{1}0) walls coated with GaO and supports the suggestions of Liliental-Weber et al. that nanopipes are linked to the presence of impurities [13].

The $V_{Ga-O_{3N}}$ defect considered above is electrically inactive but defects like $V_{Ga-O_{2N}}$ and $V_{Ga-O_{N}}$ act as single and double acceptors respectively. If these were trapped in the strain field of a dislocation, then we would expect the dislocation to appear electrically active [53]. It may be that these defects are linked to a prominent defect related yellow luminescence [54].

In conclusion, the density functional calculations show that in wurtzite GaN the stress field of threading edge dislocations is likely to trap gallium vacancies and oxygen as well as their complexes. We find that the gallium vacancy and oxygen related defect complexes are electrically active and suggest that they increase the intensity of the yellow luminescence near threading edge dislocations consistent with cathodoluminescence studies [54].
IV. SUMMARY

It is clear that dislocations in Si, in spite of almost 50 years of effort, still possess many unexplained features. The study of dislocations in GaN by contrast is in its infancy. Progress in understanding their properties has been limited because of the difficulty of distinguishing the effects of impurities and point defects from the intrinsic properties of the dislocation core.

I thank Malcolm Heggie and Joachim Elsner for detailed discussions and help with this work.
REFERENCES


