VISTA Status Report
June 2009

H. Ceric, P. Hehenberger, G. Milovanovic, V. Sverdlov, M. Vasicek, S. Selberherr

Institute for Microelectronics
TU Wien
Güßhausstraße 27-29/E360
A-1040 Wien, Austria
Contents

1 Copper Microstructure Impact on Evolution of Electromigration Induced Voids 1
   1.1 Introduction ......................................................... 1
   1.2 Modeling of Grain Boundaries .................................... 1
   1.3 Void Evolution ..................................................... 2
   1.4 Simulation Results ............................................... 2
   1.5 Conclusion ......................................................... 4

2 Do NBTI-Induced Interface States Show Fast Recovery? 5
   2.1 Introduction ......................................................... 5
   2.2 Experimental Methodology ........................................ 5
   2.3 OFIT versus CCP .................................................. 6
   2.4 Analysis of the OFIT Technique .................................. 6
       2.4.1 Dependence on Gate Voltage Low-Level ..................... 6
       2.4.2 Hysteresis due to Stress ...................................... 7
       2.4.3 Hysteresis-Free Area ........................................... 7
       2.4.4 Frequency Scalability ......................................... 7
       2.4.5 Lower Temperature ............................................... 8
   2.5 Extrapolation of Oxide Trap Contribution ....................... 8
   2.6 Simulation of The Charge-Pumping Current ..................... 9
   2.7 Results ............................................................. 10
   2.8 Conclusions ......................................................... 10

3 Nonparabolicity Effects in Quantum Cascade Lasers 11
   3.1 Introduction ......................................................... 11
   3.2 Theoretical Model ................................................. 11
   3.3 Results and Discussion ........................................... 12
   3.4 Conclusion ......................................................... 14

4 Valley Splitting in Thin Silicon Films from a Two-Band \( k \cdot p \) Model 16
   4.1 Introduction ......................................................... 16
   4.2 Two-Band \( k \cdot p \) Model ........................................... 16
   4.3 Subband Dispersion in [001] Thin Silicon Films ............... 17
4.4 Valley Splitting in a Magnetic Field ......................................................... 18
4.5 Valley Splitting in a Point Contact ............................................................ 18
4.6 Valley Splitting by Shear Strain ................................................................. 19
4.7 Conclusion .................................................................................................. 19

5 Consistent Higher-Order Transport Models for SOI MOSFETs ............. 20

5.1 Introduction ................................................................................................. 20
5.2 Model .......................................................................................................... 20
5.3 Results ......................................................................................................... 21
5.4 Conclusion .................................................................................................. 23
1 Copper Microstructure Impact on Evolution of Electromigration Induced Voids

We study the impact of microstructure on nucleation and evolution of electromigration induced voids. The grain boundaries are described with a comprehensive model which includes the dynamics of mobile and immobile vacancies in dependence of mechanical stress. The surface of an evolving void is modeled by a three-dimensional Level-Set algorithm. Simulations have shown that the constellation of grain boundaries determines the electromigration failure behavior seriously.

1.1 Introduction

Contemporary integrated circuits are often designed using simple and conservative design rules to ensure that the resulting circuits meet reliability goals. This precaution leads to reduced performance for a given circuit and metallization technology. An ultimate hope of integrated circuits designers today is to have a computer program at hand, which predicts the behavior of thin film metallizations under any imaginable condition. Due to intensive development in the last decade, physical models of electromigration have reached a level of maturity, which enables predictions of failure behavior. The cause of failure is always an electromigration induced void in the interconnect structure. We present our recent development in modeling of void nucleation and void evolution.

A new model for grain boundary physics is applied as extension and refinement of the standard electromigration continuum model. The model differentiates between mobile and immobile vacancies. Immobile vacancies are captured at grain boundaries and triple points, causing a build-up of tensile stress. High tensile stress leads to nucleation of intrinsic voids. These voids evolve through the interconnect causing a resistance change and, occasionally, a complete failure.

Different approaches have been applied to model evolving void surfaces in the last ten years. However, all of these models lack an appropriate description of the void development process, neglecting relevant physical phenomena that lead to interconnect failure. Moreover, these models are only suitable for simulations of simple two-dimensional interconnects and cannot realistically describe the void evolution mechanisms in modern complex interconnect structures. In this paper a three-dimensional Level-Set module is applied to simulate the evolving void surface. The site of void nucleation and the morphology of the evolving void accurately reproduce experimental observations.

1.2 Modeling of Grain Boundaries

The diffusion of point defects inside the grain boundary is faster compared to grain bulk diffusion due to the fact [1] that a grain boundary generally exhibits a larger diversity of point defect migration mechanisms. Moreover, formation energies and migration barriers of point defects are in average lower than those for lattice.

In 1951, Fisher published his nowadays classic paper [2] presenting the first theoretical model of grain boundary diffusion. That pioneering paper, together with concurrent measurements by Turnbull and Hoffman [3, 4], initiated quantitative studies of grain boundary diffusion in solids. The fact that grain boundaries in metals provide high-diffusivity paths was known already in the 1930s, mostly from indirect and qualitative experiments. For example, the enhanced rates of sintering, creep, discontinuous precipitation, and other processes and reactions in polycrystalline samples were attributed to an accelerated atomic transport along grain boundaries.

The grain boundary model used in this work is designed as an extension of a continuum electromigration model [5, 6]. The grain boundary is treated as a separate medium with the capability of absorbing and releasing vacancies (Fig. 1). Vacancies are trapped from both neighboring grains with the trapping rate \( \omega_T \) and released to these grains with a release rate \( \omega_R \).

![Figure 1: The fluxes \( J_{\nu,1} \) and \( J_{\nu,2} \) change the concentration of mobile vacancies \( (C_{v}^{m}, C_{v}^{2}) \) and immobile vacancies \( (C_{v}^{im}) \).](image)

The vacancy concentration from both sides of the grain boundary is denoted as \( C_{v}^{m} \) and \( C_{v}^{2} \). Correspondingly, fluxes are calculated as

\[
J_{\nu,1} = \omega_T (C_{v}^{eq} - C_{v}^{im}) C_{v}^{1} - \omega_R C_{v}^{im},
\]

\[
-J_{\nu,2} = \omega_T (C_{v}^{eq} - C_{v}^{im}) C_{v}^{2} - \omega_R C_{v}^{im}.
\]
The vacancies are captured at the grain boundary and converted into immobile vacancies \((C_v^{im})\). This is mathematically expressed as
\[
\frac{\partial C_v^{im}}{\partial t} = \frac{J_{v,1} - J_{v,2}}{\delta} = -\frac{\partial C_v^1}{\partial t} + \frac{\partial C_v^2}{\partial t},
\]
where \(\delta\) is the grain boundary thickness. By combining equations eq.(1), eq.(2), and eq.(3) with the vacancy balance equation, the following equation is obtained.
\[
\frac{\partial C_v^\alpha}{\partial t} = -\text{div} J_{v,\alpha} + \frac{1}{\tau} \left( C_{v}^{eq} - C_{v}^{im} \left( 1 + \frac{2\omega R}{\omega T(C_v^1 + C_v^2)} \right) \right),
\]
\(\alpha = 1\) and \(\alpha = 2\) indicate the left and the right side of the grain boundary, respectively. \(J_{v,\alpha}\) is the vacancy transport driven by electromigration and a stress gradient. \(\tau\) is expressed by
\[
\frac{1}{\tau} = \frac{\omega T(C_v^1 + C_v^2)}{\delta}.
\]
The capacity of the grain boundary to accept trapped vacancies is expressed by the stress-dependent equilibrium concentration [5, 7]
\[
C_{v}^{eq} = C_{v}^0 \exp \left( \frac{\sigma_{nn} \Omega}{k_B T} \right), \quad \sigma_{nn} = \vec{n} \cdot \sigma \cdot \vec{n},
\]
where we assume a unique equilibrium vacancy concentration \(C_{v}^0\) in stress free copper, in the grain’s bulk, and in the grain boundaries. \(\vec{n}\) is the normal to the grain surface and \(\sigma\) is the stress tensor.

1.3 Void Evolution

When a sufficiently high stress level is reached at some interfacial spot, where a flaw can be assumed, the failure development enters the next phase and a different modeling ansatz must be applied. Here, we have an evolving void surface shaped by two dynamic forces: the chemical potential gradient and the electron wind.

The development of fatal voids, i.e. voids which lead to a significant resistance increase, or even completely sever the line, is the ultimate cause for the electromigration induced interconnect failure [8]. Therefore, the understanding and prediction of electromigration failure behaviour can only be achieved through a detailed study of the void evolution mechanisms.

Including both contributions, electromigration and chemical potential-driven surface diffusion, gives the total surface vacancy flux \(J_s = \vec{J}_s \cdot \vec{l}\), where \(\vec{l}\) is the unit vector tangential to the void surface [9]
\[
\vec{J}_s = -D_s \left( e Z^* E_s + \Omega \nabla_s \left( \frac{\vec{\sigma} \cdot \vec{E}}{2} - \gamma_s \kappa \right) \right).
\]

1.4 Simulation Results

The site of void nucleation and the morphology of the evolving void accurately reproduce experimental observations. The simulation is started by assuming a grain boundary network (c.f. Fig. 2) in the studied dual-damascene interconnect.
The vacancy release rate $\omega_R$ and vacancy the trapping rate $\omega_T$ are chosen in such a way that during simulation following the conditions are fulfilled:

$$\frac{2 \omega_R}{\omega_T (C_1^v + C_2^v)} \ll 1,$$

and

$$1 < \tau = \frac{\delta}{\omega_T (C_1^v + C_2^v)} < 2.$$

With these conditions the model for immobile vacancies eq.(1)-eq.(6) behaves analogously to a classical Rosenberg-Ohring term [11], which was already successfully applied in [12] where $\tau = 1$ s and $\tau = 2$ s is used.

All parameters for simulation are set according to experiments published in [13]. The line width is 0.18 $\mu$m, the applied current density is 1.5 MA/cm$^2$, and the temperature is 300 $^\circ$C. Barrier and capping layer are Ta/TaN and SiCN, respectively. SiOC is used as interlayer dielectricum. The continuum equations eq.(4) are solved, until a stress threshold ($\sigma_{th}$) for void nucleation is reached at some triple point (c.f. Fig. 3). At this triple point an initial, small spherical void is set and the Hamilton-Jacobi equation eq.(9) is solved.

For stresses $\sigma < \sigma_{th}$ an energy barrier exists between the void embryo and a stable-growing void. If the stress is above the threshold value ($\sigma > \sigma_{th}$), the free energy monotonically decreases with void volume and the energy barrier vanishes. If we now assume an adhesion free patch with a radius of 10 nm (about 20 atoms), we obtain $\sigma_{th} \approx 344$ MPa [5].

After the void is nucleated at the triple point (c.f. Fig. 4), which is a natural free adhesion patch, it moves in electric current direction. The void encounters the second grain boundary and transforms its shape (c.f. Fig. 5). During evolution the void remains attached to the copper/capping layer interface which is also a fast diffusivity path. The strength of adhesion between copper and SiCN determines the speed of the void evolution. By choosing other capping materials and corresponding technology processes which increase adhesion, the speed of the void is reduced and the interconnect life time is increased.

It can be clearly seen in Fig. 5 that the presence of a grain boundary induces a void surface movement toward the inner part of interconnects. Here the grain boundary acts as fast diffusivity path. This result is consistent with numerous experimental observations and explains the fact that bamboo microstructures are prone to failures caused by slit-like voids [14–16].

In the further development this shape change is increased leading to a high increase in interconnect resistance (c.f. Fig. 6). The results of simulation and experiments demonstrate a decisive impact of the microstructure on the failure development [17, 18].
Copper Microstructure Impact on Evolution of Electromigration Induced Voids

The void moves and encounters the second grain boundary.

The shape change is enhanced as the void drifts toward the via.

Figure 7: FIB cross-section of interconnect after an electromigration test [13] (courtesy of Dr. Lucile Arnauld).

The experimental result presented in Fig. 7 clearly shows that the applied models together with the assumed microstructure reproduce the experimentally observed void dynamics.

The distance between the grain boundaries and the angle between the grain boundaries and the copper/capping layer interfaces are crucial for void evolution. Triple points determine the void nucleation site and during the evolution the void is shaped by the grain boundaries it encounters. In order to investigate the exact correlation between the microstructural properties and the void dynamics the statistical simulation using a large number of automatically generated microstructures is necessary.

1.5 Conclusion

The ultimate cause of electromigration failure is an intrinsic void. In this work we study void nucleation and evolution for the case, when the interconnect interface to the capping layer is the dominant material transport path and the void nucleation site is at the intersection of copper grain boundary and capping layer. A detailed, physically based grain boundary model is applied to model the void nucleation process. Void evolution is simulated using a three-dimensional Level-Set algorithm. The simulation results are successfully verified through comparison with experimental observations.
2 Do NBTI-Induced Interface States Show Fast Recovery? 
A Study Using a Corrected On-The-Fly Charge-Pumping Measurement Technique

Data obtained by the recently developed on-the-fly charge-pumping technique has suggested a fast initial degradation and recovery of interface states during negative and/or bias temperature stress, contrary to previously published results. By revising the analysis of the measurement setup, fast interface state creation and recovery are revealed as artifact due to a different amount of oxide traps seen during the stress and relaxation phases. From this analysis we conclude that data gathered during stress and recovery phases must not be directly compared. By properly taking the contribution of (slow) oxide charges into account, which leads to a spurious increase of the charge-pumping current during the stress phase, we demonstrate that no fast initial degradation and no fast recovery of interface states occurs. Nevertheless, the charge-pumping signal is sensitive to the continuous switching of the gate voltage into accumulation, which also accelerates interface state recovery, albeit at a slower rate. We finally conclude that both the fast initial degradation and the fast initial recovery seem to be due to oxide charges. Therefore these oxide charges need to be considered. By performing simulations with our device simulator Minimos-NT using a modified Shockley-Read-Hall model it was possible to reproduce the effect of these oxide charges. For this purpose a temperature and field activated tunneling process is assumed and results in proper agreement of measurement and simulation. A correction scheme for the on-the-fly charge-pumping measurement technique is then presented.

2.1 Introduction

The last years have seen a lively debate on whether interface states and/or oxide charges are responsible for the degradation observed during negative and/or positive bias temperature stress (NBTI, PBTI). Experimental differentiation between oxide and interface states is extremely challenging due to the rapid recovery of the degradation setting in as soon as the stress is removed. In particular, it has been observed that when after NBTI stress the device is positively biased, a considerable part of the recoverable component is lost [19–22]. Until recently, this has been explained by the detrapping of holes [20,21], while interface states have been assumed to only change their occupancy but do not recover.

2.2 Experimental Methodology

Unfortunately, experimental validation of the fast recovery behavior is challenging because the technique that is conventionally used to directly measure interface states is the charge-pumping (CCP) technique, which inherently relies on a bias switch into accumulation. Consequently, it is unclear whether the often observed weak recovery in CP data is a consequence of the fact that interface states do not recover or whether this is an artifact of the measurement technique brought about by the strong bias switch.

A quite striking result obtained with on-the-fly charge-pumping (OFIT) measurements is that in contradiction to conventional charge-pumping (CCP) measurements, OFIT data suggest a considerable amount of fast initial recovery of interface states. We note that this fast initial recovery is not explicitly measured, but is only inferred from the differences between the last stress and the first recovery measurement. Nevertheless, this misconception is a fundamental dilemma and a major issue for our understanding of BTI. Clarification of this matter is the prime requisite for the development of a reliable model. In the following we try to resolve the issue whether interface states do recover quickly (< 1 s) or not.
The lately developed OFIT technique [24–26] has refuted the debate. As illustrated in Fig. 8, the basic difference between OFIT and CCP [27–30] is that the low-level $V_{\text{G,low}}$ of the CP pulse is simultaneously used as a stress condition (for NBTI), while the actual CP measurement is performed by quickly switching back and forth between accumulation $V_{\text{acc}}$ and stress $V_{\text{stress}}$. Consequently, an issue we will get back to later, the low-levels are different during stress and recovery/reference measurements.

### 2.3 OFIT versus CCP

As in conventional CP measurements, care has to be taken that parasitic tunneling currents and geometry effects do not pollute the measured charge-pumping current $I_{\text{cp}}$. The first problem is even more severe in the OFIT technique since there the low level gate voltage equals the stress voltage, resulting in excessive tunneling in thin oxides. In order to avoid these problems we also use large-area devices with thick oxides (30 nm). As shown in Fig. 9, the measured $I_{\text{cp}}$ during stress and recovery are qualitatively identical for three completely different technologies (30 nm thick SiO$_2$, 3.5 nm thin SiO$_2$, and SiON).

Quite remarkably, continuous application of OFIT pulses (as well as CP measurements) has a dramatic impact on both the stress and the recovery characteristics. In particular, with 3 measurements per decade, $I_{\text{cp}}$ is quasi constant during recovery, while up to 100 measurements per decade result in approximately 80% recovery of $I_{\text{cp}}$. Another fact worth mentioning is that the first OFIT measurement point during stress is already responsible for at least 30% of the total degradation. Likewise, the first measurement taken during recovery at 0.1 ms already shows 30% recovery while the rest of the recovery depends basically on the number of measurements per decade.

### 2.4 Analysis of the OFIT Technique

In order to deepen our understanding of the method we performed constant base-level CP measurements ($V_{\text{Base}} = 2 \text{ V}$) using a gradually increasing pulse amplitude $\Delta V_{\text{G}}$. Until reaching the desired stress-level starting from $-1 \text{ V}$ down to $-18 \text{ V}$ the pulse slopes have to be kept constant to obtain comparable results. Constant pulse slopes ensure that the upper and lower energy boundaries of the active energy interval remain unchanged, although $\Delta V_{\text{G}}$ increases [23]. Given the requirements of a constant slope and a constant duty cycle, increasing the pulse amplitude $\Delta V_{\text{G}}$ leads to larger pulses. This means that the rise and fall times have to be adapted at every voltage step. Since it is inevitable to change both the pulse width and also the rise and fall times one has to ask for the potential pitfalls: Are OFIT-data obtained during stress and relaxation comparable? If that is not the case, is there some possibility to correct this nonconformity? These questions will be examined in the following.

Starting with Fig. 10 the two large arrows pointing up and down again reveal some important aspects of the temporal evolution of the pulses during a CP measurement. The charge-pumping current $I_{\text{cp}}$ at stress conditions ($V_{\text{G,low}} \leq -3 \text{ V}$) differs a lot when compared to relaxation ($V_{\text{G,low}} = -1 \text{ V}$). The further the NBTI stress increases the higher the $I_{\text{cp}}$-signal becomes. This can be partly attributed to the desired effect of using the measurement setup to also stress the device. However, this cannot fully account for the observed behavior.

#### 2.4.1 Dependence on Gate Voltage Low-Level

Under the assumption that only interface states $N_{\text{it}}$ contribute to $I_{\text{cp}}$, $I_{\text{cp}}$ should actually become independent of $V_{\text{G,low}}$ as soon as the strong inversion regime is reached. This $I_{\text{cp,0}}$ is marked by the dashed line in Fig. 10. However, as demonstrated previously [31, 32], $I_{\text{cp}}$ continues to increase, albeit at a much slower rate. This increase with $\Delta V_{\text{G}}$ is routinely attributed to slower oxide traps.
an increased offset for the next pulse. The total hysteresis is nearly no stress. The deeper the device is stressed into an additional hysteresis (marked with reversed. Evolution of the rising pulse edges gives the upper branch of the curve is traversed. When the experiment is repeated at a lower frequency the contribution to oxide traps. Only a negligible amount of interface states constituting $\Delta I_{cp}$ can be identified for different frequencies but equal temperature. Following these results at least part of the defects must vary with temperature or frequency. For better comparability, the data at 12.5 kHz are scaled to the reference frequency ($f_{ref} = 125$ kHz).

2.4.3 Hysteresis-Free Area

As displayed in the inset in Fig. 10 the very first pulses are almost free of stress (no hysteresis, $\Delta I_{cp} = 0$) and hence the deviation of $I_{cp}$ from $I_{cp,0}$ is entirely due to oxide traps. Only a negligible amount of interface states $\Delta N_{it}$ are created by the measurement process. The hysteresis-free area will be discussed in more detail in the next section.

2.4.4 Frequency Scalability

When the experiment is repeated at a lower frequency (see bottom of Fig. 11), one finds that the interface state contribution can be scaled to the reference frequency ($f_{ref} = 125$ kHz) [31]. This is compatible with the fact that the stress duration is practically independent of frequency. On the other hand, the recoverable oxide trap contribution to $I_{cp}$ depends on frequency, consistent with the idea that the lower the frequency (corresponding to more time per pulse) the more oxide traps can contribute to $I_{cp}$.

\[ \Delta N_{it} = I_{cp} = I_{cp}^{\text{it}} + I_{cp}^{\text{ot}} \] [29, 33]. So, regardless of the amount of degradation, $I_{cp}$ varies as function of $V_{G,\text{low}}$. This fact has to be taken into account for a meaningful comparison of stress and relaxation CP data.

\[ \Delta N_{it} = I_{cp} = I_{cp}^{\text{it}} + I_{cp}^{\text{ot}} \] [29, 33]. So, regardless of the amount of degradation, $I_{cp}$ varies as function of $V_{G,\text{low}}$. This fact has to be taken into account for a meaningful comparison of stress and relaxation CP data.

**Figure 10**: Charge-pumping current $I_{cp}$ for different pulse amplitudes as observable in constant base-level CP measurements with $V_{\text{base}} = 2$ V and a gradually increasing pulse amplitude $\Delta V_G$ from $V_{G,\text{low}} = -1$ V down to $V_{G,\text{low}} = -17$ V. The $I_{cp}$ shows a significant hysteresis. If the $I_{cp}$ is evaluated at the falling pulse edge, the lower branch of the curve is traversed. Evolution of the rising pulse edges gives the upper branch. However, the contribution of slow oxide states and an additional hysteresis (marked with $\Delta I_{cp}$) are clearly visible for increasing pulse amplitudes. This implies that depending on the pulse amplitude, $I_{cp}$ will contain contributions in addition to the interface states. Provided only interface states are available, $I_{cp}$ should be independent of the pulse amplitude (dashed line of $I_{cp,0}$).

\[ \Delta N_{it} = I_{cp} = I_{cp}^{\text{it}} + I_{cp}^{\text{ot}} \] [29, 33]. So, regardless of the amount of degradation, $I_{cp}$ varies as function of $V_{G,\text{low}}$. This fact has to be taken into account for a meaningful comparison of stress and relaxation CP data.

\[ \Delta N_{it} = I_{cp} = I_{cp}^{\text{it}} + I_{cp}^{\text{ot}} \] [29, 33]. So, regardless of the amount of degradation, $I_{cp}$ varies as function of $V_{G,\text{low}}$. This fact has to be taken into account for a meaningful comparison of stress and relaxation CP data.
2.4.5 Lower Temperature

Finally, at a low temperature (displayed at the top of Fig. 11) practically no hysteresis is introduced (no NBTI stress) and also the oxide trap contribution is reduced, consistent with the idea that these traps are due to a thermally activated tunneling mechanism \([34]\) rather than elastic (and thus temperature-independent) hole tunneling \([19]\).

2.5 Extrapolation of Oxide Trap Contribution

As demonstrated above, during an OFIT measurement we have to expect a distortion of \(I_{cp}\) due to oxide charges and due to the creation of defects during the low-level. In order to analyze this distortion we proceed as follows:

We determine \(V_{G,low}\) to be the lowest value of \(V_{G,low}\) at which no hysteresis is observed. We then use the data set \(V_{G,low} > V_{G,low}\) to extrapolate the impact of oxide charges \(\Delta N_{ot}\) down to the stress-level. It is not possible to obtain this information from the stress pulse because of the contribution of both parts: \(\Delta N_{it}\) and \(\Delta N_{ot}\). Quite remarkably, the data can be fit by a quadratic polynomial, consistent with our NBTI experiments where we also observe a quadratic (\(E_{ox}^2\)) dependence of the hole-trapping component \([34–36]\). The hole-trapping theory developed in \([34]\) was applied to our data and excellent agreement was obtained. The difference between the actual signal \((I_{cp}^{it} + I_{cp}^{ot})\) and the extrapolated curve in Fig. 12 finally gives \(\Delta N_{it}\).

In Fig. 12 and Fig. 13 the extraction algorithm for \(\Delta N_{ot}\) and \(\Delta N_{it}\) is demonstrated. Stress and relaxation pulse responses both consist of two branches, one going down (fall) and one going up (rise) as marked by arrows. In the falling branch, \(V_{G,low}\) varies from 0 V to \(-17\) V. In the rising branch, \(V_{G,low}\) varies from \(-17\) V to 0 V. Only pulses with constant \(I_{cp}^{rise} - I_{cp}^{fall}\) (or even without a hysteresis, i.e. \(I_{cp}^{rise} = I_{cp}^{fall} = 0\)) are suitable to create an extrapolation guess for higher \(V_{G,low}\). This ‘safe window’ ranges from 0 V to \(-8\) V where both branches are indistinguishable.

The extracted components for different temperatures and frequencies are given in Fig. 14. The additionally created oxide traps \(\Delta N_{ot}\) depend on frequency as well as on temperature and clearly show \(V_{G,low} \sim E_{ox}^2\) behavior. The hysteresis due to additionally created traps \(\Delta N_{it}\) is independent of frequency but strongly dependent on temperature.

![Figure 12: Charge-pumping current \(I_{cp}\) for the stress pulse \((V_{stress} = -17\) V\) and the relaxation pulse \((V_{relax} = -8\) V\), shown in Fig. 10. To unravel the contribution of oxide charges and additional interface states we look at the difference \(I_{cp}^{rise} - I_{cp}^{fall}\). In the range \(-8\) V < \(V_{G,low} < 0\) V, this difference is constant, implying no additional creation of interface states. From this ‘safe window’ we extrapolate to the minimum low-level to estimate the contribution due to oxide charges. Note that the first branches \(I_{cp}^{fall}\) of the stress and relaxation pulse differ from each other due to pre-stress pulses between \(V_{G,low} = -8\) V and \(V_{G,low} = -17\) V. In fact, when using fresh devices for each measurement all \(I_{cp}^{fall}\) would coincde.](image1)

![Figure 13: Top: Lower temperatures simplify the extrapolation due to the absence of degradation. Here the full range of pulse amplitudes can be used to verify the extrapolation down to deep inversion. Very good conformity is obtained, again indicating the absence of additional interface states in deep inversion at low temperatures. Bottom: Noise complicates this procedure at low frequencies. Data are scaled to \(f_{ref} = 125\) kHz.](image2)
Do NBTI-Induced Interface States Show Fast Recovery?

2.6 Simulation of The Charge-Pumping Current

To approximately account for the temperature and field activated tunneling process, we use a modified Shockley-Read-Hall (SRH) model with our device simulator Minimos-NT [37]. The SRH-capture-rates are multiplied by

\[
\exp \left( \frac{E_{ox}^2}{E_{ox,ref}^2} \right) \exp \left( -\frac{\Delta E_B}{k_B T} \right)
\]

where \( E_{ox} \) is the electric field in the oxide, \( E_{ox,ref} \) is a reference value, \( \Delta E_B \) the multiphonon emission barrier and \( T \) the temperature. \( \Delta E_B \) can be characterized by a Gaussian distribution with the mean energy \( \Delta E_B \). When setting the parameters some points need to be considered in order to end up with a physically appropriate model:

1. The first exponential factor models the bias dependence. It is very sensitive to changes of \( E_{ox,ref} \) due to the squared exponent, leading to a very small range of valid \( E_{ox,ref} \) values.

2. When setting the barrier \( \Delta E_B \) too low, the oxide traps contribute to the interface trap signal as the second factor approaches unity. Setting \( \Delta E_B \) too high leads to very low rates, effectively eliminating the contribution of oxide traps.

Simulation results are depicted in Fig. 15. As the simulation treats the CP measurement process as stress-free, no additional interface traps due to stress are created and only the oxide-charge part is visible. With the thermally activated barrier the increasing \( \Delta I_{cp} \) can be described.

3. Lastly, the distribution of \( \Delta E_B \) determines the dependence of \( I_{cp} \) on \( V_{G,low} \). Increasing the mean of the distribution at \( \Delta E_B \) increases the mean capture/emission-time constants. Since with constant-slope pulses higher pulse amplitudes \( \Delta V_G \) require longer pulse durations, increasing the mean \( \Delta E_B \) shifts the point from which a significant contribution of oxide traps \( \Delta N_{ox} \) can be observed to higher pulse amplitudes. On the other hand, broadening the distribution of \( \Delta E_B \) (increasing the variance) also broadens the distribution of time constants, observable as broadening the range of \( V_{G,low} \) where \( I_{cp} \) increases. A quadratic behavior as observed in the experiments, Fig. 15, can be well reproduced with a broad Gaussian distribution (\( \Delta E_B,mean = 1 \text{ eV}, \Delta E_B,\sigma = 0.5 \text{ eV} \)).
2.7 Results

Based on the previous results we are now able to better understand the charge-pumping current $I_{cp}$ measured during the OFIT sequence. The presence of additional charges contributes to the signal when the pulse amplitude $\Delta V_G$ is increased. A large spread of time constants slower than that of the interface states is necessary to explain the results. By assuming oxide traps with a thermally distributed activated barrier one is able to explain the measurement results with good accuracy. Whereas interface states are independent of the electric field and account for $I_{cp}$ at low ($10$ kHz) and high frequencies ($1$ MHz) likewise due to their small time constants, the oxide traps with their assumed barrier are by far slower, only affecting $I_{cp}$ at lower frequencies i.e. $10$ kHz.

The particularly troublesome part is the application of the OFIT technique during the stress phase, where both oxide traps $\Delta N_{ot}$ and additionally created interface states $\Delta N_{it}$ add to $I_{cp}$. These contributions are absent during the initial reference measurements and during the OFIT recovery measurements both taken at $V_{G,low} = V_{relax}$. This has fundamental consequences on OFIT measurements: Initially, a reference $I_{cp}$ is recorded. Following this reference measurement, the gate voltage low-level $V_{G,low}$ is switched to stress $V_{stress}$. Due to the much larger $\Delta V_G$ now a significant contribution of $I_{ox}^{cp}$ is obtained.

Furthermore, with the large pulse amplitude, additional interface states are created, which is the intended effect of this OFIT measurement. However, without this the additional increase in $I_{cp}$ due to oxide traps must not be attributed to interface states created by degradation. Consequently, we need to correct for $I_{ox}^{cp}$ in the measurement data. Using the mentioned extrapolation method of $\Delta N_{ot} = AE_{ox}^{2}$ we see that the $30\%$ initial increase in $I_{cp}$ is entirely due to oxide traps. The corrected last stress value in Fig. 16 is identical to the first value at the recovery, leading to the conclusion that no fast interface state recovery occurs.

2.8 Conclusions

We have discussed the dynamics of interface state creation in constant base-level charge-pumping measurements using a gradually increasing pulse amplitude $\Delta V_G$ (from inversion into deep inversion). The charge-pumping current is not constant once reaching the inversion regime, but increases due to slow oxide traps. From this analysis we conclude that data gathered during stress and recovery phases must not be directly compared. A correction scheme for the OFIT measurement technique was necessary. By properly taking the contribution of these traps into account, we demonstrate that no fast initial degradation and no fast recovery of interface states occurs. Nevertheless, the CP signal is sensitive to continuous switches of the gate voltage into accumulation which also accelerates interface state recovery, albeit at a slower rate. We finally conclude that both the fast initial degradation and the fast initial recovery are very likely to be due to oxide charges which are created by a thermally activated process.
3 Nonparabolicity Effects in Quantum Cascade Lasers

We calculate electron-LO phonon and interface roughness scattering rates in a GaAs/Al$_x$Ga$_{1-x}$As quantum cascade laser taking into account conduction subband nonparabolicity. In this work we investigate the Al concentration and the $k_i$ dependence of nonparabolicity effects. It is shown that the subband nonparabolicity may increase the electron-LO phonon scattering rates significantly, but the scattering rates are not qualitatively different from those in the parabolic approximation. Especially, we show that nonparabolicity leads to noticeable changes even for transitions involving electrons at the bottom of the subband and that this behaviour follows from the phonon wave vector and the electron phonon overlap.

3.1 Introduction

Over the past several years, solid-state lasers based on intersubband transitions in semiconductor heterostructures have proved to be very promising candidates for practical sources of radiation, particularly in the midinfrared region [38]. While the possibility of optical amplification by intersubband transitions in biased superlattices had already been predicted in 1971 [39], a successfully working quantum cascade laser (QCL) has first been reported in 1994 [40].

Understanding the physics in these heterostructures is necessary to design and fabricate new QCLs with specific properties. In this work we adress the question how the subband nonparabolicity affects the electron-LO phonon scattering rates as well as the interface roughness scattering rates in a QCL subjected to barrier characteristics.

GaAs/Al$_x$Ga$_{1-x}$As active regions grown on GaAs substrates are ideal candidates for this approach as Al$_x$Ga$_{1-x}$As is nearly perfectly lattice matched to the GaAs substrate. Frequently used theoretical models describing the electron properties in a semiconductor approximate the band structure by parabolic conduction bands. Deviations from the parabolic model are not considered in quantum well (QW) calculations based on effective mass approaches [41].

Here we investigate the influence of conduction band nonparabolicity on the intersubband scattering rates due to electron-LO phonon interactions and interface roughness, considering transitions with different initial energies. In order to determine for which situations the subband nonparabolicity can not be neglected, we show results of scattering rates for different $k_i$ and make comparisons with the parabolic approximation. Our results demonstrate the significance of nonparabolicity effects in heterostructure lasers even for transitions where the initial state is at the bottom of the subband, particularly for scattering rates due to electron-LO phonon interaction, in situations of large electronic confinement. We identify the physical origins of the changes induced by conduction subband nonparabolicity in dependence on the Al content and the in-plane wave vector $k_i$ to be found in terms of the phonon wave vector and the electron phonon overlap.

3.2 Theoretical Model

We consider electrons in the conduction band of a QCL in an electric field applied in the direction perpendicular to the heterointerfaces. In order to describe band nonparabolicity we follow the model proposed by Nag and Mukhopadhyay [42]. The nonparabolic $E - k$ relation is given by

$$E - V(x) = \frac{\hbar^2 k^2}{2m^*(x)} (1 - \gamma(x)k^2)$$

(12)

where $\gamma(x)$ is the nonparabolicity parameter, $m^*(x)$ is the effective mass, and $V(x)$ is the conduction band offset.

The envelope function in the QW structure is given by

$$\Psi_{i,k}(x) = \frac{1}{\sqrt{A}} e^{ik_\parallel x} \Phi_i(z)$$

(13)

where $k_\parallel$ is the in-plane wave vector, $A$ is the cross-sectional area of the QW structure, and $\Phi_i(z)$ is the electron envelope function of the $i$th state. The effective mass equation for the QW structure is given by

$$\left[ -\frac{\hbar^2}{2} \frac{\partial^2}{\partial z^2} \frac{1}{m^*_\parallel(z,E)} \frac{\partial}{\partial z} + V(z) + \frac{\hbar^2 k_\parallel^2}{2m^*_\parallel(z,E)} - eFz \right] \Phi_i(z) = E_i(k_i) \Phi_i(z)$$

(14)

where the nonparabolic effective masses along the perpendicular and parallel directions are defined as [43]

$$m^*_\perp(x,E) = \frac{m^*(x)}{\alpha(x)} \left[ 1 - (1 - 2\alpha(x))^{1/2} \right]$$

(15)

$$m^*_\parallel(x,E) = \frac{m^*(x)}{(1 - 2\alpha(x))^{1/2}}$$

(16)
where $\alpha = 4\gamma(x)m^*(x)(E - V(x))/\hbar^2$.

The Hamiltonian describing the electron-phonon interaction can be written in the form [44]

$$\hat{H}_{ep} = \sum_{i,j,k,k'} \sum_{q} (F_{ij}^{k,k',q} c_{i,k}^\dagger b_{j,k}^q + F_{ij}^{k,k',q} c_{i,k} b_{j,k}^q) + \text{h.c.}$$

(17)

with

$$F_{ij}^{k,k',q} = F_{q}^{LO} \int \psi_{i,k}(x) e^{iqx} \psi_{j,k}(x) dx$$

(18)

where $q$ is the phonon wave vector, and $c_{i,k}^\dagger (b_{j,k})$ and $c_{i,k} (b_{j,k})$ denote creation and annihilation operators of the electron (phonon), respectively. $F_{q}^{LO}$ represents the coupling factor for the electron-LO phonon interaction

$$F_{q}^{LO} = -\frac{e}{q} \left( \frac{\hbar \omega_{LO}}{2V e_0} (\epsilon_\infty - \epsilon_\infty) \right)^{1/2}$$

(19)

where $\epsilon_\infty$ and $\epsilon_\infty$ are high-frequency and static dielectric constants, $V$ is the volume and $e$ is the electronic charge. The electron-LO phonon scattering rate for an electron initially in a state $i$ to the final state $j$ is obtained from the Fermi Golden Rule

$$W_{ij}^\pm (k||) = \frac{2\pi}{\hbar} \sum_{k'} |\langle j|\hat{H}_{ep}|i\rangle|^2 \delta(E_j - E_i) \pm \hbar \omega_q$$

(20)

by evaluating the matrix element in eq.(20) with the electron-phonon interaction Hamiltonian given by (17)

$$W_{ij}^\pm (k||) = -\frac{2\pi}{\hbar} \sum_{k',q} \left[ \frac{\hbar \omega_{LO}}{2V} (\epsilon_\infty - \epsilon_\infty) \right]$$

$$\times \int \int \frac{e^2}{k'^2 + k^2 - 2k'k||} \cos \theta + q^2$$

$$\times \Phi_j^*(z) e^{iqz} \Phi_i(z) dz \delta \left( \frac{1}{2} \pm \frac{1}{2} + N_{LO} \right)$$

$$\times \delta(E_j - E_i) \pm \hbar \omega_q$$

(21)

$\pm$ denoting emission and absorption processes, and $N_{LO}$ is the phonon occupation number given by

$$N_{LO} = \frac{1}{e^{\hbar \omega_{LO}/k_B T} - 1}$$

(22)

The roughness of the interfaces in a QW leads to spatial fluctuations in the width of the well, and consequently to fluctuations of the confinement energy. These fluctuations of the quantization energy act as a fluctuating potential for the motion of confined carriers [45].

The randomness of the interface is described by a correlation function at the in-plane position $r = (x, y)$ which is usually taken to be Gaussian with a characteristic height of the roughness $\Delta$, and a correlation length $\Lambda$ representing a length scale for fluctuations of the roughness along the interface [46], such that

$$\langle \Delta(r) \Delta(r') \rangle = \Delta^2 e^{-|r-r'|^2/\Lambda^2}$$

(23)

The perturbation in the potential $V(z)$ due to a position shift $\Delta(r)$ is given by

$$\delta V = V[z - \Delta(r)] - V(z) \approx -\Delta(r) \frac{dV(z)}{dz}$$

(24)

For the $i$th interface, which is centered about the plane $z_{iL}$ and extends over the range $[z_{iL}, z_{iR}]$, the scattering matrix element can be defined as

$$M_{ji,I} = \left\langle j \big| \delta V \left( \frac{z - z_{iL}}{z_{iR} - z_{iL}} \right) \right| i \right\rangle$$

$$= \frac{\varphi_{ji,I}}{A} \int \Delta(r) e^{i(k|| - k'i)} r$$

(25)

where $|j\rangle$ and $|i\rangle$ denote the final and initial wave functions, respectively. $\varphi_{ji,I}$ is defined as

$$\varphi_{ji,I} = \int \Phi_j^*(z) \frac{dV}{dz} \text{rect} \left( \frac{z - z_{iL}}{z_{iR} - z_{iL}} \right) \Phi_i(z) dz$$

(26)

and the rectangular function reads

$$\text{rect}(z) = \begin{cases} 1, & |z| \leq 0.5 \\ 0, & |z| > 0.5 \end{cases}$$

The expectation value of the square of the matrix element is given by

$$\langle |M_{ji,I}|^2 \rangle = \frac{\varphi_{ji,I}^2}{A^2} \int \langle \Delta(r') \Delta(r) \rangle e^{i(k|| - k'i)} (r-r') dr'dr$$

(27)

Making use of eq.(23) and Fermi’s Golden Rule, the interface roughness induced scattering rates are given by [47]

$$W_{ij}^{IR} (k||) = \sum_{I} \left( \frac{m^* \Delta^2 \Lambda^2}{\hbar^3} \right) |\varphi_{ji,I}|^2 \frac{\delta(E_j - E_i)}{k_B T}$$

$$\times \int_0^{\pi} e^{-\frac{1}{2} (k' - k||)^2 \Lambda^2/4} d\theta$$

(28)

where $k'$ and $k||$ are the final and initial wave vectors, respectively, and $\theta$ is the scattering angle.

### 3.3 Results and Discussion

We consider a GaAs/Al$_x$Ga$_{1-x}$As QCL design which comprises a three-level sheme [48]. The conduction band profile for an electric field of 48 kV/cm is given in Fig. 17.
The conduction band offset is determined as a function of the Al concentration $x$ by the standard approximation [49]

$$\Delta E_c = \begin{cases} 0.75x \text{ eV}, & x \in [0, 0.45] \\ 0.75x + 0.69(x - 0.45)^2 \text{ eV}, & x \in [0.45, 1] \end{cases}$$

The expression for the nonparabolicity parameters in terms of the energy gap were obtained from Ref. [50]

$$\gamma(x) = \frac{\hbar^2}{2m^*(x)E_g(x)}$$

where the energy bandgap is $E_g(x) = (1.424 + 1.247x)m_0$ in the range $0 \leq x \leq 0.45$, and $E_g(x) = 1.900 + 0.125x + 0.143x^2$ for $x > 0.45$. The material parameters used in our calculations are for Al$_x$Ga$_{1-x}$As [51], the effective mass $m^*(x) = 0.067 + 0.083x$, the dielectric constants $\varepsilon_S = 13.18 - 3.12x$ and $\varepsilon_{\infty} = 10.89 - 2.73x$, and the LO-phonon energies $\hbar\omega_{LO} = 36.25 - 6.55x + 1.79x^2$ meV. The characteristic height and the correlation length of the interface roughness are taken as $\Delta = 2.83$ Å and $\Lambda = 70$ Å.

In Fig. 18 we show the calculated electron-LO phonon scattering rates for the intersubband transition $2 \rightarrow 1$ as a function of the Al content for $k_\perp = 0$ nm$^{-1}$, $k_\parallel = 0.1$ nm$^{-1}$, and $k_\parallel = 0.2$ nm$^{-1}$. The scattering rates with the inclusion of subband nonparabolicity are represented by solid lines and the dashed lines are for the parabolic band approximation. In general, the scattering rates are increased due to effects of nonparabolicity, except for intersubband transitions for large aluminum concentration, but the scattering rates are otherwise not qualitatively different from those in the parabolic band approxim-
The variations of the electron-LO phonon scattering rates due to nonparabolicity are for transitions at the bottom of the subband of almost the same order as for transitions with higher initial energies. In the region of low Al concentration the nonparabolicity may cause a variation of about 8% for transitions initially at \( k_\parallel = 0 \text{ nm}^{-1} \), ~10% at \( k_\parallel = 0.1 \text{ nm}^{-1} \) and ~11% at \( k_\parallel = 0.2 \text{ nm}^{-1} \).

For larger aluminum content the effects due to subband nonparabolicity for transitions initially at the bottom of the subband are even higher than for initial conditions away from it.

The nonparabolicity induced increase in the density of final states and of the quantum confinement, resulting in a larger electron-phonon overlap, and the influence of the subband on the phonon wave vector are mainly responsible for the variations in the transition rates. Since these variations are negligible for large \( x \), the confinement decreases with the increase of the Al concentration and the effects of the nonparabolicity on the overlap integral become weaker. This is illustrated in Fig. 19, where the ratio of the nonparabolic and parabolic electron-phonon overlap for the intersubband transition 2 → 1 is shown. Since the electron-phonon interaction has a strong dependence on the phonon wave vector (\( \propto 1/q \)) it is crucial to investigate its behavior due to nonparabolicity in order to understand the influence of nonparabolic subbands on the electron-LO phonon scattering rates. Fig. 20 shows that for nonparabolic subbands the phonon wave vector is larger than for parabolic subbands and that this effect becomes more pronounced for regions with higher in-plane wave vector of the electron \( k_\parallel \). Thus the electron couples more weakly to the phonon for nonparabolic subbands yielding a decrease of the electron-LO phonon scattering rates, which is more affected in case of larger phonon wave vectors. Although in situations of weak confinement the role of the phonon wave vector becomes more important, the decrease of scattering rates resulting from larger phonon wave vectors is compensated by the electron-phonon overlap integral. Hence these rates are generally larger for nonparabolic subbands. This explains why the variations in the electron-LO phonon scattering rates due to inclusion of nonparabolicity with higher initial electron in-plane vectors are of almost the same order as for \( k_\parallel = 0 \text{ nm}^{-1} \) and in some situations of weak confinement even smaller.

The calculated scattering rates due to interface roughness as a function of Al concentration are illustrated in Fig. 21. We show results for transitions where the initial state is at the bottom of the subband and away from it. In general, we observe that the variations on the interface roughness induced scattering rates due to nonparabolicity are approximately constant for all values of \( x \). Numerically estimated we obtain variations of the order of 2-3 %. The effects of subband nonparabolicity on scattering rates due to electron-LO phonon interaction and interface roughness are better illustrated in Fig. 22 where we display the ratio of the parabolic and nonparabolic scattering rates.

### 3.4 Conclusion

In conclusion, we have calculated the scattering rates of the intersubband transition 2 → 1 due to electron-LO phonon interaction and interface roughness in a GaAs/Al\(_x\)Ga\(_{1-x}\)As QCL at three different initial state conditions, namely at the bottom of the subband \( k_\parallel = 0 \) and...
We find that the electron-LO phonon scattering rates are in general increased, except in situations of low confinement, which is equivalent to a large aluminum concentration. In particular, in situations of low aluminum concentration the results indicate the importance of including nonparabolicity in the conduction band due to a noticeable increase of the electron-LO phonon scattering rates even for transitions at the bottom of the subband, as the variations are of almost the same order as for initial states with larger electron in-plane vectors. We have shown that these subband nonparabolicity effects can be understood in terms of the electron-phonon overlap and the phonon wave vector.
4 Valley Splitting in Thin Silicon Films from a Two-Band $k\cdot p$ Model

4.1 Introduction

The rapid increase in computational power and speed of integrated circuits is supported by the continuing size reduction of semiconductor devices’ feature size. Thanks to constantly introduced innovative changes in the technological processes the miniaturization of MOSFETs institutionalized by Moore’s law successfully continues. The 45nm MOSFET process technology by Intel [52] and recently introduced 32nm technology [53] involve new high-k dielectric/metal gates and represents a major change in the technological process since the invention of MOSFETs. Although alternative channel materials with a mobility higher than in Si were already investigated [54], [55], it is believed that Si will still be the main channel material for MOSFETs beyond the 32nm technology node. With scaling apparently approaching its fundamental limits, the semiconductor industry is facing critical challenges. New engineering solutions and innovative techniques are required to improve CMOS device performance. Strain-induced mobility enhancement is one of the most attractive solutions to increase the device speed. It will certainly maintain its key position among possible technological innovations for the future technology generations. In addition, new device architectures based on multigate structures with better electrostatic channel control and reduced short channel effects will be developed. A multi-gate MOSFET architecture is expected to be introduced for the 22nm technology node. Combined with a high-k dielectric/metal gate technology and strain engineering, a multi-gate MOSFET appears to be the ultimate device for highspeed operation with excellent channel control, reduced leakage currents, and low power budget. Confining carriers within a thin film reduces the channel dimension in transversal direction, which further improves gate channel control. At the same time the search for post-CMOS device concepts has accelerated. Spin as a degree of freedom is promising for future nanoelectronic devices and applications. A concept of a racetrack memory recently proposed in [56] is based on the controlled domain wall movement by spin-polarized current in magnetic nanowires. Silicon, the main element of microelectronics, possesses several properties attractive for spintronics applications. Silicon is composed of nuclei with predominantly zero spin and is characterized by small spin-orbit coupling. In a recent ground-breaking experiment coherent spin transport through an undoped silicon wafer of 350$\mu$m length was demonstrated [57]. The experiment was possible due to a unique injection and detection technique of polarized spins delivered through thin ferromagnetic films. Spin coherent propagation at such long distances makes the fabrication of spinbased switching devices likely already in the near future. Spin-controlled qubits may be thought of as a basis for upcoming logic gates. However, the conduction band of silicon contains six equivalent valleys, which is a source of potentially increased decoherence. For successful applications the degeneracy between the valleys must be removed and become larger than the spin Zeeman splitting. Shubnikov-de-Haas measurements in an electron system composed of thin silicon films in Si-SiGe heterostructures reveal that the valley splitting is small [58]. At the same time, recent experiments on the conductivity measurements of point contacts created by confining a quasi-twodimensional electron system in lateral direction with the help of additional gates deposited on the top of the silicon film demonstrate a splitting between equivalent valleys larger than the spin splitting [58]. In this work we demonstrate that a large valley splitting in the confined electron system can be induced by a shear strain component. Our analysis is based on the two-band $k\cdot p$ model for the conduction band in silicon. The parabolic band approximation usually employed for subband structure calculations of confined electrons in Si inversion layers is insufficient in ultra-thin Si films. The two-band $k\cdot p$ model includes strain and is shown to be accurate up to energies of 0.5eV. This model can therefore be used to describe the subband structure in thin silicon films, where the subband quantization energy may reach a hundred meV.

We first describe the subband structure in a thin unstrained silicon film. We demonstrate that the peculiarities of the subband dispersion obtained within the two-band $k\cdot p$ model result in a linear dependence of the valley splitting on the magnetic field. We show that a large valley splitting is observed in experiments on conduction quantization through a quantum point contact in [110] direction, but the splitting is suppressed in [100] point contacts. Finally, we demonstrate that the valley splitting is greatly enhanced in films strained in [110] direction.

4.2 Two-Band $k\cdot p$ Model

The closest band to the lowest conduction band $\Delta_1$ near its minimum is the second conduction band $\Delta_2$. These two bands are degenerate exactly at the X point. Since the minimum of the lowest conduction band in unstrained silicon is only $k_0 = 0.15\cdot 2\pi/a$ away from the X point, where a is the lattice constant of unstrained silicon, the two bands must be included on equal footing in order to describe the dispersion around the minimum. More distant bands separated by larger gaps are included in the second order $k\cdot p$ perturbation theory [62], which results in the following two-band $k\cdot p$ Hamiltonian:
4 Valley Splitting in Thin Silicon Films from a Two-Band k·p Model

4.3 Subband Dispersion in [001] Thin Silicon Films

For [001] silicon films the confinement potential gives an additional contribution $U(z)l$ to the Hamiltonian eq.(30). In the effective mass approximation described by eq.(30) with the coefficient in front of $\sigma_x$ set to zero, the confining potential $U(z)$ is known to quantize the six equivalent valleys of the conduction band of bulk silicon into the four-fold degenerate primed and the two-fold degenerate unprimed subband ladder. In ultra-thin films the unprimed ladder is predominantly occupied and must be considered. The term with $\sigma_x$ in eq.(30) couples the two lowest conduction bands and lifts the two-fold degeneracy of the unprimed subband ladder. The additional unprimed subband splitting, or the valley splitting, can be extracted from the Shubnikov-de-Haas oscillations and is typically in the order of a few tens $\mu eV$ [58]. However, the valley splitting is greatly enhanced in a laterally confined two-dimensional electron gas [58]. The valley splitting is usually addressed by introducing a phenomenological intervalley coupling constant at the silicon interface [63]. Here we investigate the valley splitting based on the two-band k·p model eq.(30) without introducing any additional parameters. We approximate the confining potential of an ultrathin silicon film by a square well potential with infinite potential walls. This is sufficient for the purpose to analyze the valley splitting in a quasi-two-dimensional gas due to interband coupling. Generalization to include a self-consistent potential is straightforward though numerically involved [64]. Because of the two-band Hamiltonian, the wave function $\Psi$ is a spinor with the two components $[0 \uparrow]$ and $[1 \downarrow]$. For a wave function with space dependence in a form $e^{ik_zz}$ the coefficients $A_0$ and $A_1$ of the spinor components are related via the equation $H\Psi = E(k_z)\Psi$. For a particular energy $E$ there exist four solutions $k_z(i = 1, \ldots, 4)$ for $k_z$ of the dispersion relation eq.(31), so the spatial dependence of a spinor component $\alpha$ is in the form $\sum_{i=1}^{4} A_{\alpha,i} e^{ik_zz}$. The four coefficients are determined by the boundary conditions that both spinor components are zero at the two film interfaces. This leads to the following equations:

$$\tan\left( k_0 t \frac{k_0 t}{2} \right) = \frac{k_2}{\sqrt{k_2^2 + \eta^2} \pm \eta} \cdot \tan\left( k_0 t \frac{k_0 t}{2} \right)$$

where $\eta = m_t |\delta|/\hbar^2 k_0^2$. The value of $k_2$ becomes imaginary at high $\eta$ values. Then the trigonometric functions in eq.(33) are replaced by the hyperbolic ones. Special care must be taken to choose $\alpha$ the correct
branch of $\sqrt{k_x^2 + \eta^2}$ in eq.(34): the sign of $\sqrt{k_x^2 + \eta^2}$ must be alternated after the argument becomes zero. Introducing $y_n = (k_1 - k_2)/2$, eq.(33) can be written in the form:

$$\sin(y_n k_0 t) = \pm \eta y_n \sin\left(\frac{1 - \eta^2 - y_n^2}{1 - \eta^2 - y_n^2} k_0 t\right)$$

(35)

For small values of the parameter $\eta$ we obtain from eq.(35) the following dispersion relation for the unprimed subbands $n$:

$$E_n^\pm = \frac{\hbar^2}{2m_1} \left(\frac{\pi n}{t}\right)^2 + \frac{\hbar^2}{2m_2} \left(\frac{\pi n}{k_0} + \frac{\pi \kappa}{2m_2} \sin(k_0 t)\right)^2$$

(36)

$$\pm \left(\frac{\pi n}{k_0} + \frac{\pi \kappa}{2m_2} \sin(k_0 t)\right)^2$$

eq.(36) demonstrates that the unprimed subbands are not necessarily degenerate and degeneracy is preserved only, when shear strain is zero and either $k_x = 0$ or $k_y = 0$.

4.4 Valley Splitting in a Magnetic Field

For zero shear strain the Landau levels in an orthogonal magnetic field $B$ are determined from eq.(36) using the Bohr-Sommerfeld quantization conditions:

$$E_{m}^{1,2} = \hbar \omega_c \left(m + \frac{1}{2}\right) \frac{\pi}{4 \arctan \left(\sqrt{m(1,2)/m(2,1)}\right)}$$

(37)

$$m_{(1,2)} = \left(\frac{1}{m_e} \pm \frac{1}{M} \left(\frac{\pi n}{k_0} \right)^2 \sin(k_0 t) / k_0 t [1 - (\pi n/k_0 t)^2]\right)^{-1}$$

(38)

and

$$\omega_c = \frac{eB}{\sqrt{m_1 m_2} c}$$

is the cyclotron frequency, $e$ is the electron charge, and $c$ is the speed of light. According to eq.(37), the valley splitting $|E_{m}^1 - E_{m}^2|$ is linear regarding the magnetic field. In order to obtain the linear dependence, two conditions must be satisfied:

(i) no shear strain and

(ii) the cyclotron energy is smaller than the

subband quantization energy. Both conditions are satisfied in a biaxially stressed silicon film of 10nm thickness on SiGe used in [58] for magnetic fields as strong as 1T. It follows from eq.(37) and eq.(39) that the valley splitting can be several tens $\mu eVs$, which is consistent with the experiment [58].

4.5 Valley Splitting in a Point Contact

We consider a point contact in [110] direction realized by confining an electron system of a thin silicon film laterally by depleting the area under additional gates. Without strain the low-energy effective Hamiltonian in the point contact can be written as:

$$H_{(1,2)} = \frac{\hbar^2 k_x^2}{2m_{(2,1)}} + \frac{\hbar^2 k_y^2}{2m_{(1,2)}} + \frac{1}{2} \kappa x'^2 + V_b$$

(39)

where the primed variables are along the [110] and [1-10] axes, the effective masses are determined by eq.(39), $\kappa$ is the spring constant of the point contact confinement potential $V(x') = \kappa x'^2 / 2$ in [1-10] direction, and $V_b$ is a gate voltage dependent conduction band shift in the point contact [65]. The dispersion relation of propagating modes within the point contact is written as:

$$E_{p}^{(1,2)} = \frac{\hbar^2 k_x^2}{2m_{(2,2)}} + \hbar \omega_{(1,2)} \left(\frac{p + \frac{1}{2}}{90}\right) + V_b$$

(40)

where $\omega_{(1,2)}^2 = \kappa/m_{(1,2)}$. Since the energy minima of the two propagating modes with the same $p$ are separated, they are resolved in the conductance experiment through the point contact as two distinct steps. The valley splitting is $\Delta E_p = \hbar |\omega_1 - \omega_2|$. The difference in the
effective masses eq.(39) and, correspondingly, the valley splitting can be greatly enhanced by reducing the effective thickness $t$ of the quasi-two-dimensional electron gas which is usually the case in a gated electron system, when the inversion layer is formed. In a [100] oriented point contact without strain the effective Hamiltonian is

$$H^\pm = \frac{\hbar^2 (k_x^2 + k_y^2)}{2mt} + \left( \frac{\pi n}{k_0 t} \right)^2 \left| \frac{k^2 k_x k_y}{M} \right| \sin(k_0 t) \left| 1 - \left( \pi n / k_0 t \right)^2 \right|^{1/2} + \kappa x^2.$$ (41)

Due to symmetry with respect to $k_y$ the subband minima in a point contact are always degenerate. For this reason the valley splitting in [100] oriented point contacts is greatly reduced.

### 4.6 Valley Splitting by Shear Strain

It follows from eq.(36) that shear strain induces a valley splitting linear in strain for small shear strain values and depends strongly on the film thickness [66]:

$$\Delta E_n = 2 \left( \frac{\pi n}{k_0 t} \right)^2 \frac{D_{xy}}{k_0 t} \left| 1 - \left( \pi n / k_0 t \right)^2 \right| \sin(k_0 t).$$ (42)

For higher strain values eq.(33) must be solved numerically. Results shown in Fig.3 demonstrate that valley splitting can be effectively controlled by adjusting the shear strain and modifying the effective thickness $t$ of the electron system. Uniaxial stress along [110] channel direction, which induces shear strain, is already used by industry to enhance the performance of MOSFETs. Therefore, its application to control valley splitting does not require expensive technological modifications.

### 4.7 Conclusion

The unprimed valley structure in (001) silicon thin films has been analyzed within the two-band $k \cdot p$ model. It is shown that the two-fold degeneracy of the unprimed subbands can be lifted leading to the so-called valley splitting which is proportional to the strength of the perpendicular magnetic field. The valley splitting can be enhanced in $<110>$ oriented point contacts, while it is suppressed in a $<100>$ point contact. Finally, the valley splitting can be controlled and made larger than the Zeeman splitting by shear strain. This makes silicon very attractive for spintronic applications.

![Figure 25: Splitting between the unprimed subband energies, or the valley splitting, in a 6.5 nm thick silicon film as a function of shear strain. The splitting values are normalized to the energy of the ground subband without strain. The value $\eta = 1$ corresponds to the shear strain value $\varepsilon_{xy} = 0.016$. The value of valley splitting may alternate its sign, in accordance to eq.(36).](attachment:figure25.png)
5 Consistent Higher-Order Transport Models for SOI MOSFETs

We have developed a two-dimensional non-parabolic macroscopic transport model up to the sixth order. To model higher-order transport parameters with as few simplifying assumptions as possible, we apply an extraction technique from Subband Monte Carlo simulations followed by an interpolation within these Monte Carlo tables through the whole inversion layer. The impact of surface-roughness scattering as well as quantization on the transport parameters is inherently considered in the Subband Monte Carlo data. These tables are used to model higher-order mobilities as well as the macroscopic relaxation times as a function of the effective field and the carrier temperature. We have studied the influence of the inversion layer concentration on higher-order transport parameters within high fields and show the behavior of these parameters in a quantized system of a UTB SOI MOSFET.

5.1 Introduction

For engineering applications macroscopic transport models based on Boltzmann’s transport equation (BTE) like the drift-diffusion (DD) model or the energy transport model (ET) are very efficient compared to the time consuming Monte Carlo (MC) simulation [67]. However, with further decrease of the device dimensions into the deca-nanometer regime both the DD, and the ET model become more and more inaccurate [68]. Investigations have demonstrated [69, 70] that higher-order macroscopic models can cover the gate length range of 100 nm down to about 25 nm. In the analysis of macroscopic models up to the sixth order, it is essential to describe the transport parameters, namely the carrier mobility $\mu_0$, the energy-flux mobility $\mu_1$, the energy relaxation time $\tau_1$, the second-order energy-flux mobility $\mu_2$, and the second-order energy relaxation time $\tau_2$ with as few simplifying assumptions as possible. A rigorous study of the behavior of these parameters in the bulk case has been already carried out using bulk MC tables [69]. The use of this bulk data set for modeling transport in MOSFET devices is problematic due to the importance of surface roughness scattering and quantization in the inversion channel. In [71] surface roughness scattering on the carrier mobility has been investigated using the semiempirical Matthiesen rule, but a rigorous study has not been performed yet.

In order to take these important inversion layer effects as well as non-parabolic bands for high fields into account, we have developed a Subband Monte Carlo (SMC) [72, 73] table based 2D electron gas six moments transport model. It is now possible to study the influence of the inversion layer effects on higher-order transport parameters.

5.2 Model

In order to derive 2D higher-order macroscopic models like the six moments model one has to multiply the BTE

$$\partial_t f + v \cdot \nabla_x f - \frac{F}{\hbar} \cdot \nabla_k f = (\partial_t f)_{\text{coll}},$$

with special weight functions [74] and integrate over the k-space. $f(r, k, t)$ is the distribution function, $F$ is the driving force, and $v(k)$ is the group velocity in the BTE. The moments in 2D space are defined as

$$x(r, t) = \frac{2}{(2\pi)^2} \int X(k)f(r, k, t)d^2k = n(r, t)\langle X(k) \rangle,$$

with $x(r, t)$ as the macroscopic values together with the microscopic counterpart $X(k)$. $n(r, t)$ is the carrier concentration. Substituting the microscopic weights with $\epsilon^i(k)$ and $v(k)\epsilon^i(k)$ one will obtain the even, scalar-valued moments, and the odd, vector-valued moments, respectively. For the six moments model $i$ is in the range of $i \in [0, 2]$. The scattering operator of the BTE is modeled using the macroscopic relaxation time approximation [75]. For the even moments the macroscopic relaxation time ansatz looks like

$$\partial_t \langle X \rangle_{\text{coll}}^\text{even} \approx -n\langle \epsilon^i \rangle - \langle \epsilon^i \rangle_0 \frac{\tau_1}{\tau_i},$$

and for the odd moments

$$\partial_t \langle X \rangle_{\text{coll}}^\text{odd} \approx -n\langle \epsilon^i \rangle_0 \frac{\tau_1}{\tau_i}.$$

$\langle \epsilon^i \rangle_0$ are the even moments in equilibrium. Further approximations during the derivation are the diffusion approximation, which states that the anti-symmetric part of the distribution function is much smaller than the symmetric one [76], the modeling of the tensorial components, and the closure relation [77]. The transport model will be formulated in terms of $\epsilon^i \equiv \langle \epsilon^i \rangle$ and $V_i \equiv \langle v\epsilon^i \rangle$ [69]. The general conservation equation reads

$$\partial_t (nw_i) + \nabla \cdot (nV_i) - iF \cdot nV_{i-1} = -n\frac{w_{i+1} - w_{eq}}{\tau_i},$$

whereas the general fluxes are defined as

$$nV_i = -\frac{\mu_iH_{i+1}}{2q} \left( \nabla (nw_{i+1}) - nFw_{i+1} \frac{2+iH_i}{H_{i+1}} \right),$$

$H_i$ are the non-parabolicity factors, and are defined in the 2D space as

$$U_i = \frac{1}{2} \text{tr} (U_i) = \frac{1}{2} w_i H_i.$$
with $U_i = (v \otimes p e^i)$ as the energy tensors \([77]\). The even moments $w_i$ of the six moments model are expressed as $k_B T_{n_i}$, $2 (k_B T_{n_i})^2 \beta$, and $6 (k_B T_{n_i})^3 \beta^3$. $\beta$ is the kurtosis and denotes the deviation from a heated Maxwellian distribution function
\[
\beta = \frac{1}{2} \frac{w_2}{w_1^2}.
\] (50)
$\epsilon$ is an integer in the range $\epsilon \in [0, 3] \ [69]$. To take the quantization as well as surface roughness scattering into account we extract the transport parameters $\tau_i$ and $\mu_i$ from a self-consistent coupling between a SMC simulator and a Schrödinger-Poisson (SP) solver. In the SMC simulator we consider non-parabolic bands, quantization effects, phonon induced scattering as well as surface roughness scattering \([72, 78]\). The SP solver incorporates the quantum confinement in inversion layers. Our device simulator Minimos-NT \([37]\) calculates the effective field through the channel of a device and extracts higher-order transport parameters from the SMC tables. The mobilities as well as the relaxation times are now a function of the effective field and the carrier temperature $\mu_i(E_{\text{eff}}, T_n), \tau_i(E_{\text{eff}}, T_n)$.

5.3 Results

As an example device a fully depleted SOI MOSFET with a Si film thickness of 4 nm in direction, and a donor doping concentration of $10^{20}$ cm$^{-3}$ in the source and the drain regions, and with an acceptor doping in the channel of $10^{16}$ cm$^{-3}$ has been investigated. An electric field is applied in \langle 100 \rangle direction. In Fig. 26 we show higher-order mobilities as a function of the inversion layer concentration $N_{\text{inv}}$. We point out that in bulk MOSFETs for the low-field case the mobility of our simulation fits the measurement data of Takagi \([79]\) while a significant reduction is observed in the quantized 4 nm channel region \([80]\). Furthermore it is demonstrated that for high lateral fields and high $N_{\text{inv}}$ the mobilities converge to the same value. This is not the case for the relaxation times. In Fig. 27 and Fig. 28 the energy relaxation times and the second-order energy relaxation time as a function of $N_{\text{inv}}$ are plotted for different lateral fields, respectively. For high $N_{\text{inv}}$ and for low fields the relaxation times increase compared to the high lateral field case where the relaxation times are constant. This can be explained with Fig. 29 where the first subband occupation as a function of $N_{\text{inv}}$ of the unprimed, primed, and double primed valley for lateral fields of 50 kV/cm, and 200 kV/cm is shown. Due to the fast increase of the occupation number of the first subband in the unprimed valley at 50 kV/cm compared to the high-field case, where the occupation is constant, the change in the higher-order relaxation times increases as well for high $N_{\text{inv}}$. Fig. 30 shows effective fields for different bias points through an SOI MOSFET with a gate length of 40 nm. With the effective fields and the SMC tables one can model higher-order transport parameters through a device, as presented in Fig. 31. A profile of the carrier, the energy-flux, and the second-order energy flux mobilities has been pointed out.

Fig. 26: The higher-order mobilities as a function of the inversion layer concentration $N_{\text{inv}}$ for different lateral fields $E_{\text{abs}}$ are plotted. For high fields the difference of the mobilities decreases. For low fields in a bulk MOSFET the carrier mobility is equal to the measurement data of Takagi.

Fig. 27: The energy relaxation time as a function of the inversion layer concentration for different lateral fields is shown. For a field of 50 kV/cm and a high $N_{\text{inv}}$, $\tau_1$ increases compared to high-fields, where the energy relaxation time is more or less constant.

Fig. 28: The second-order temperature as a function of the inversion layer concentration for different lateral fields is shown. For a field of 50 kV/cm and a high $N_{\text{inv}}$, $\tau_2$ increases compared to high-fields, where the energy relaxation time is more or less constant.
Figure 28: The second-order relaxation time is shown as a function of the inversion layer concentration for different lateral fields. The increase of $\tau_2$ at 50 kV/cm is as well higher than at the other fields. The change at high $N_{inv}$ and low fields of the second-order relaxation time compared to the other fields is higher than in the energy relaxation time.

Figure 29: The first subband occupation of the unprimed, primed, and double primed valley as a function of the inversion layer concentration for fields of 50 kV/cm and 100 kV/cm. Due to the light mass of the unprimed valley in transport direction the subband occupation number is higher than in the other valleys.

Figure 30: The effective field throughout the whole device for drain voltages of 0.1 V, 0.2 V, and 0.3 V. With the effective field and an interpolation between SMC tables, higher-order transport parameters can be modeled throughout the whole device.

Figure 31: The carrier, the energy-flux as well as the second-order energy flux mobilities are plotted through a 40 nm gate length SOI MOSFET. A drain voltage of 0.3 V has been applied.

voltages due to the deviation of the distribution function from a Maxwellian distribution. The maximum peak is at the channel drain junction where hot electrons from the channel meet cold electrons from the drain. Fig. 33 shows velocity profiles calculated with the DD, ET, and the six moments (SM) models of an SOI MOSFETs with a gate lengths of 40 nm and 60 nm. In the 40 nm device the ET model considerably overestimates the velocity (and thus the drain current) with a velocity three times as high as the DD model, whereas for the 60 nm device it is just two times as high [68]. For large devices the velocity of the ET and the SM converge to the value of the DD model. In Fig. 34 we point out that with further increase of the gate lengths the difference between the output currents decreases. For large devices all transport models yield the same result.
5 Consistent Higher-Order Transport Models for SOI MOSFETs

Figure 34: Output characteristics calculated with the ET and the SM model for SOI MOSFETs with 40 nm, 60 nm, 80 nm, and 100 nm gate lengths. A gate voltage of 1 V is applied. For long channel devices the ET and the SM model yield the same result.

Figure 33: Velocity profile calculated with the DD, ET, and the SM model for a 40 nm and 60 nm device. The velocity of the ET and SM model is smaller in the 60 nm than in the 40 nm device. DD predicts velocities independent from the gate length. Due to the quantization and surface roughness scattering the velocity in the DD model is smaller than the saturation velocity ($\approx 10^7$ cm/s) in the bulk.

5.4 Conclusion

We have studied the behaviour of higher-order transport parameters in inversion layers. A method based on an interpolation between SMC tables for modeling a two-dimensional electron gas has been developed. This approach allows the investigation of UTB SOI devices including the influence of surface-roughness scattering and quantization within higher-order moment models. A generalized set of equations has been used to derive a novel 2D six moments model. First the influence of different inversion layers on higher-order transport parameters within high fields has been shown. We found a significant change in the relaxation times for high inversion layer concentrations and low fields compared to high fields. Furthermore higher-order mobilities converge for high inversion layer concentrations to the same value. Second we have pointed out the importance of using higher-order moment models for the modeling of devices with a gate length in the deca-nanometer regime.
References


