Thickness control of molecular beam epitaxy grown layers at the 0.01–0.1 monolayer level

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2012 Semicond. Sci. Technol. 27 085007
(http://iopscience.iop.org/0268-1242/27/8/085007)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 130.88.0.35
The article was downloaded on 15/08/2012 at 06:33

Please note that terms and conditions apply.
Thickness control of molecular beam epitaxy grown layers at the 0.01–0.1 monolayer level

P Dasmahapatra\(^1\), J Sexton\(^1\), M Missous\(^1\), C Shao\(^2\) and M J Kelly\(^2\)

\(^1\) School of Electrical and Electronic Engineering, The University of Manchester, Sackville Street, Manchester M60 1QD, UK
\(^2\) CAPE, Department of Engineering, University of Cambridge, 9 JJ Thomson Avenue, Cambridge CB1 1EG, UK

E-mail: mjk1@cam.ac.uk

Received 16 April 2012, in final form 13 May 2012
Published 13 June 2012
Online at stacks.iop.org/SST/27/085007

Abstract

Electron tunnelling through semiconductor tunnel barriers is exponentially sensitive to the thickness of the barrier layer, and in the most common system, the AlAs tunnel barrier in GaAs, a one monolayer variation in thickness results in a 300% variation in the tunnelling current for a fixed bias voltage. We use this degree of sensitivity to demonstrate that the level of control at 0.06 monolayer can be achieved in the growth by molecular beam epitaxy, and the geometrical variation of layer thickness across a wafer at the 0.01 monolayer level can be detected.

(Some figures may appear in colour only in the online journal)

There have been many ideas for electronic devices and circuits that use quantum mechanical tunnelling of electrons through thin energy barriers formed by wide-band semiconductors in semiconductor multilayer structures [1]. The route to exploitation of these devices and circuits has been blocked by the extreme sensitivity of the tunnel current to the thickness and to the height of the barrier layer. This has limited the uniformity of device properties across a wafer and the reproducibility of the same properties from wafer to wafer. Six years ago [2] one of us showed that there was a narrow window of opportunity for such devices if the barrier could be precisely an integer number of atomic layers thick. In this paper we show that we can control layers to better than 0.1 monolayers during their growth by molecular beam epitaxy (MBE), and that we can identify and determine variations in layer thickness at the level of 0.01 monolayers.

First, we consider the basic textbook theory of quantum mechanical tunnelling to show how the sensitivity of the current arises. For a beam of electrons of positive energy \(E\) (of individual mass \(m^*\)) incident on a thin barrier of height \(V\) (>\(E\)) and width \(l\), there is a finite probability of transmission of the electrons through the barrier given by [3]:

\[
T(E) = \frac{4k^2K^2}{[(k^2 + K^2)^2 \sinh^2 Kl + 4k^2K^2]},
\]

where

\[
k^2 = 2m^*E/h^2, \quad K^2 = 2m^*(V - E)/h^2.
\]

By integrating the product, \(N(E)T(E)f(E)v(E)\), where \(N(E)\) is the incident density of states, \(f(E)\) is the Fermi factor for the occupation of those states, \(T(E)\) is this transmission probability and \(v(E)\) is the electron velocity, we calculate the electron number current. This simple analysis is the starting point for designing many semiconductor devices, where electrons are the majority carriers; they are in the conduction band of a narrower band-gap semiconductor (e.g. GaAs) and are then incident on a thin layer of a wider band-gap semiconductor (e.g. AlAs) that is sandwiched on both sides by GaAs. In this case the barrier thickness, \(l\), is a few atomic layers, \(V\) is the difference in the conduction band edge energies of the two materials typically a fraction of 1 eV, and the mass is now the conduction-band electron effective mass in GaAs. By feeding in typical values, it is straightforward to show that the increase of the thickness \(l\) by one monolayer typically
reduces the value of $T(E)$ by about a factor of nearly 3 for each tunnelling electron, and so the current is reduced, well verified in experiments, most elegantly with the use of perpendicular magnetic fields [4]. The strong dependence of tunnelling current when the barrier thickness shows an excursion of ±1 ML around a 10 ML nominal barrier thickness is best illustrated in figure 1. Most systems designers want to be able to pre-specify their device performance to within ±10% about an absolute value of current at a fixed bias, meaning that deep sub-monolayer precision of the crystal growth is essential.

There is also another problem. We may try to grow an ideal tunnel barrier with exactly $N$ (say)10 monolayers of AlAs in GaAs, in order to obtain the desired reproducibility. Suppose however that at either interface, there is a monolayer alloy of average composition Al$_x$Ga$_{1-x}$As, with $x = \alpha$ on one side and $x = \beta$ on the other side of the barrier layer. The total layer might be described as being $N + \theta$ (0 < $\theta$ < 1) thick, where $\alpha + \beta = \theta (\pm 0.1)$. How do we describe the tunnelling through such a barrier? Depending on the wavelength of the tunnelling electron, there are two extreme possibilities, and they each produce a quite different answer. The typical wavelength scale for the tunnelling electron ($\lambda = 2\pi/k$) is about 10 nm for a practical device, which in our case is three times longer than the thickness of a typical barrier (here ~3 nm for 10 monolayers of AlAs). The tunnelling electron has a coherence area of the order of this wavelength in diameter, and it samples the local tunnel barrier on that areal scale. This is the sense in which we can describe a tunnel barrier as having a fractional layer thickness. If the Al and Ga are perfectly randomly distributed within the interface layer on less than the same 10 nm length scale of the tunnelling electron wavelength, then the electron will see a uniform barrier of width set by interpolating with an exponential function between the values of $T_N$ and $T_{N+1}$ to get $T_{N+\theta}$ from the sinh function in $T(E)$ above. By contrast, if the Al atoms within each interface alloy monolayer are segregated into local clusters on the length scale of 10 nm or more, the tunnelling electron will see either $N$ or $N+1$ layers and tunnel accordingly, and the total current of the diode will scale by the relative areas of thickness $N$ and $N+1$ layers, and hence a linear interpolation between $T_N$ and $T_{N+1}$. In practice, we could get anything in between these two extremes. Just how serious this is comes from the fact that when there is a factor of 3 in the transmission through layers differing by only one atomic layer, it is easy to obtain as much as a 25% variation through a barrier of order 10 monolayers if both interfaces are half-integer (50%/50%) alloy monolayer [2], entirely dependent on the degree of the presence or absence of clustering on the 10 nm length scale. We must also ensure that $\alpha = \beta = 0$ above to within ±0.05, i.e. that both interfaces of the AlAs tunnel barrier had complete layers of AlAs adjacent to complete layers of GaAs to eliminate the form of variability seen in earlier experiments [6, 7]. There is another aspect to this variability, namely that if the barrier is a ternary alloy, the alloy atoms have to be randomly positioned in the group III or group V sublattice. Any tendency to cluster at the scale of the same 10 nm will this time change the height of the barrier locally. Overall, this is a smaller effect than the integrity of the integer monolayers in the barrier.

There are many elaborate theories of tunnelling in semiconductor multilayers, where $T = T(E, V)$ and $f = f(E, \Theta)$, where $V$ is the bias voltage and $\Theta$ the temperature, and they form the basis of simulators for tunnel device design, such as Silvaco used above, but this simple analysis captures the essence of the sensitivity of tunnelling to barrier properties that is our key concern.

We now describe the asymmetric space layer tunnel (ASPA) diode [5, 6] which is the vehicle for this study. We place a single AlAs tunnel barrier towards one end of the i-region of an n-i-n GaAs structure. This acts as a rectifying diode with some very attractive characteristics as a microwave detector. It has the same basic transfer efficiency (the terminal voltage achieved for a fixed input microwave power), and the same dynamic range, as a Si or GaAs Schottky diode or a planar-doped-barrier (PDB) diode [5, 6]. It has the very desirable property of low sensitivity of tunnel currents to ambient temperature over the whole −50 °C to +100 °C military temperature range. In contrast to the extreme way that both Schottky diodes and planar-doped diodes vary through the strong $T^2$ dependence of their currents which is due to thermionic emission, our ASPAT diode has only a small residual temperature dependence to the extent that thermionic emission over the AlAs barrier is a weak parallel channel for current. Finally, the ASPAT diode has the maximum curvature $T^2 \exp (-1/T)$ dependence of their currents which is due to thermionic emission, our ASPAT diode has only a small residual temperature dependence to the extent that thermionic emission over the AlAs barrier is a weak parallel channel for current. Finally, the ASPAT diode has the maximum curvature of its $T^2$ characteristic near $V = 0$, and so it acts naturally as a zero-bias detector, and has very low noise properties there.

Attempts to establish a manufacturable process to make the ASPAT have foundered on the intrinsic device-to-device fluctuations, as we were seeing the effects of alloy monolayers at the interfaces [7, 8]. As discussed in the previous section this paper focuses on avoiding these alloy interfaces.

A multiple wafer growth run produced three wafers, VMBE2067 A, B and C, following the ex situ pre-calibration

![Figure 1. The forward bias current density from the Silvaco software simulation using their energy band parameters for the tunnel barrier height and the actual mesa area.](image-url)
Figure 2. The map for the wafers A of the current at 0.5 V forward bias on 400 devices across the wafer, showing a gradual increase in current (and a thinning of the barrier by \( \sim 0.06 \) monolayer) across the wafer. The flat of the wafer is at the bottom of the map, and all three wafers, disposed symmetrically on the holder, show the same qualitative and quantitative behaviour. Note that there are ten fields of identical devices, with six fields blank for other purposes.

The key point though for these experiments was the necessity to achieve a precise 10 monolayer AlAs barrier thickness. This was achieved both through the pre-calibration procedure immediately before the growth of the ASPAT wafer, but this time, and more importantly, through the use of a very low growth rate for AlAs to ensure tight control over the thickness and the completion of the GaAs layer before AlAs is introduced and vice versa at the top interface with a layer of AlAs being completed before restarting the GaAs growth. The mechanical Al shutter control (within 0.1 s between opening and shutting) could in theory contribute 10% of one monolayer fluctuation at usual growth rates.

Devices were fabricated on the three 2” n-doped substrates using a full wet-etched process. Epitaxial layers were first etched using a non-selective ortho-phosphoric based etchant \( \text{H}_3\text{PO}_4:\text{H}_2\text{O}_2:\text{H}_2\text{O} \) (3:1:50) to isolate the ASPAT. Front and back alloyed ohmic contacts comprised of 5 nm AuGe, 12.5 nm Ni and 200 nm Au were thermally evaporated and annealed at 415 °C for 1 min. This provided a low (<0.1 \( \Omega \) mm) contact resistance as determined from the transmission line method (TLM) characterization study, and thus its contribution to the overall device resistance is negligible. The full-wafer layout consisted of a total of 16 fields with 6 left intentionally blank for future testing studies. Within the 10 used fields, a total of 456 devices (per field) with mesa dimensions of 100 \( \times \) 100 \( \mu \)m\(^2\) down to 5 \( \times \) 5 \( \mu \)m\(^2\) were available to be manually measured for this study.

We present the results in two parts, first the wafer-to-wafer reproducibility and then the in-wafer uniformity. We do the former by using a Cascade probe station to measure the average and standard deviation of the forward bias current from typically 30 devices from the same position on each of three different wafers, each device having a mesa of 100 \( \mu \)m side. We obtain the same average current of 42.0 mA to within 1% at 0.5 V forward bias from all three wafers, and the standard deviation of 4% within the local block, the left-hand block in the bottom row in figure 2. Note that these variations also include any variation coming from the probe technique and the ohmic contact resistances. This very tight reproducibility

Figure 3. The average \( I-V \) characteristics from the same block (the left-hand block on the lowest row in figure 2) in each of three wafers are almost identical and agree closely with the simulated version for the device as specified.
means that the three wafers have local barrier thicknesses that are equal to within 0.01 monolayer from wafer to wafer!

In figure 2, we show a colour contour of the currents at 0.5 V forward bias for wafer A which shows a qualitatively and quantitatively similar pattern to the other wafers. The absence of variability that would be due to monolayer alloy interfaces is clear from the horizontal colour uniformity. In contrast to previous work the local device-to-device variability is much reduced as expected. The most obvious feature is a 30% gradation of current across the wafer from bottom to top in each case.

This last effect is of great interest, as it reflects the ultimate limitations of the geometry of the growth chamber. During growth the platen that holds the wafers (see below) rotates at \(\sim 30\) rpm, ensuring that the deposition time for a monolayer (\(\sim 6\) s) eliminates any systematic variation in barrier thickness across the wafers, other than a linear and radial effect, actually seen above because of the off-axis geometry of the Knudsen cells within the growth chamber. In the figure, the flat of the wafer is at the bottom, and is nearest the centre of the flat platen that holds the wafers during growth, typically 3 cm from the axis about which the platen rotates during growth. The lower current indicates a thicker barrier. The top of the figure comes from the opposite side of the wafer which rotates at 8 cm from the platen axis, and the higher current indicates a thinner barrier. There is a 20% rise in current across the wafer, and this represents a 0.06 monolayer decrease in barrier thickness (simulations indicate a 300% fall in current for an extra monolayer of AlAs in the barrier—see above): this is a <1% variation across the wafer in the barrier thickness which is intended to be 10 monolayers thick.

We have been able to model this cross-wafer variation. The aluminium beam intensity is rate-determining for the growth of AlAs barriers (as there is an over-pressure of arsenic). From the off-axis position of the Knudsen cell containing Al, and its pointing towards a point on the axis at or below the centre of the platen, we can map out the \(\cos(\theta)\) variation of the beam at any instant at any place on the platen (with \(\theta\) being the angle between the axis of the Knudsen cell and the direction of the beam reaching the substrate), and average this over one cycle to get the dependence of the average Al beam intensity. This can be done analytically (see appendix A), and if the beam is pointed at the axis 10 cm below the platen centre, a 1.5% variation in thickness across the wafer is predicted given the geometry of the growth chamber is in close agreement with the 1% that is observed. Indeed, taking the finite diameter of the Knudsen cell opening into account, the variation is even closer to that observed. This variation can clearly be reduced in a larger growth chamber, and in careful designs of the platen that holds the wafers during growth.

In figure 3, we show the close reproducibility of the diode performance between wafers, and compare it with the simulated behaviour. The only discrepancy is in reverse bias where the spreading field from the diode shape (with a top contact covering only 25% of the mesa top area) plays a role.

In conclusion, we have achieved levels of reproducibility that are compatible with large-scale manufacture, but we still have in-wafer uniformity to deal with. A redesign of the platen may accomplish this. In terms of materials control we have shown the fidelity of the MBE growth process to a very high degree, with both the wafer-to-wafer spot reproducibility and side-to-side wafer uniformity showing the reproducibility approaching 0.01 monolayer precision. The absence of the alloy interface effects seen as noise on the radial variation data in earlier experiments [7, 8] is particularly gratifying. Short of atomic force microscopy, where atoms can be positioned one at a time, we think we have achieved the ultimate level of materials control that can be extracted from the tools used in mainstream microelectronics manufacture.

**Appendix A. Radial variation of the growth rate in an MBE chamber**

In figure A1, we show a schematic of the growth chamber of an MBE system with an on-axis platen holding three wafers, subject to molecular beams of Al and As during the growth of AlAs. We focus on the beam intensity from the Al Knudsen cell, which is off-axis above the platen. The cell also defines the zero of the azimuthal coordinate, and points to a point (P) below the centre of the platen (as shown). As the platen rotates, a point Q at \((r, \phi)\) on the platen will be subjected to a beam intensity given by the cosine of the angle of the centre of the Al beam and the line of sight from the centre of the beam to the point. We average this instantaneous intensity over one cycle of the platen (or over a half cycle by symmetry), to obtain a dependence on the radius \(r\) for this intensity. A point
nearest to the axis sees angles between $\Theta_1$ and $\Theta_2$, while that furthest from the axis sees angles between $\Theta_3$ and $\Theta_4$ in the diagram.

We define a general point on the platen as $P = (-R, 0-H) - (r \cos \phi, r \sin \phi, 0)$, where $R$ and $H$ are defined in the diagram, respectively, as the distance of the cell centre from the axis and the height of the cell above the platen. It is assumed that the cell is aimed at a point on the axis below the platen at a distance $K > H$ from the point level with the cell mouth. The instantaneous intensity for the point $Q$ is given by $\cos \Theta = O \cdot P|O| \cdot |P|$, where $O = (-R, 0, -K)$ and between the cell and the point on the axis at which it is pointed. The intensity is then given by

$$I(r) = \frac{1}{\pi} \int_{0}^{\pi} \frac{|O \cdot P(\phi)|}{|O| \cdot |P(\phi)|} d\phi$$

$$= \frac{1}{\pi} \int_{0}^{\pi} \frac{[R + r \cos \phi + HK]/\sqrt{[(R^2 + K^2) \times \{R + r \cos \phi)^2 + (r \sin \phi)^2 + H^2]]}}{\sqrt{[1 + \cos \phi /(2R + 2HK)]/}}$$

$$= \sqrt{[1 + 2r \cos \phi /(2R + 2HK)]}$$

We call the entire prefactor $\nu = [1/\pi](R^2 + HK)/\sqrt{[(R^2 + K^2)(R^2 + H^2 + r^2)]}$ and

$$I(r) = \nu \int_{0}^{\pi} [1 + \alpha \cos \phi]/\sqrt{(1 + \beta \cos \phi)}]$$

where

$$a = rR/(R^2 + KH) \text{ and } \beta = 2aR/(R^2 + H^2 + r^2)$$

$$= \nu[2(1 - \alpha/\beta)K(\rho)/\sqrt{[1 + \beta]} + 2(\alpha/\beta)\sqrt{[1 + \beta]}E(\rho),$$

where $\rho = \sqrt{[2\beta/(1 + \beta)]}$ and $K(x), E(x)$ are the complete elliptic integrals of the first and second kind.

Using figure values for the MBE growth chamber ($R = 10 \text{ cm}, H = 30 \text{ cm}, K = 40 \text{ cm}$), and with the wafer position varying from 3 to 8 cm from the central axis, we observe the result in figure A2, showing a uniform 1.5% fall in intensity across the wafer, in close agreement with the 1% observed.

Note that if the wafer platen were to be reshaped as the exterior of a cone of half angle $\tan^{-1}(R/H)$ the cross-wafer variation of the angle of incidence range between $\Theta_1$ and $\Theta_4$ would be reduced, and the in-wafer uniformity of the layers would be improved. Uniform heating of such a platen would need to be solved.

References

[6] Syme R T 1993 Microwave detection using GaAs/AlAs tunnel structures GEC J. Res. 11 12–23