Complementary III-V Heterojunction Tunnel FETs
Monolithically Integrated on Silicon

C. Convertino\textsuperscript{a}, H. Schmid\textsuperscript{a}, L. Czornomaz\textsuperscript{a}, H. Riel\textsuperscript{a}, S. Sant\textsuperscript{b}, A. Schenk\textsuperscript{b} and K.E. Moselund\textsuperscript{a}

\textsuperscript{a} IBM Research Zurich, 8803 Rüschlikon, Switzerland
\textsuperscript{b} Integrated Systems Laboratory, ETH Zürich, Switzerland

The tunnel FET (TFET) is considered as one of the most promising devices for ultra-low power operation, and it is clear that heterojunction devices are required to achieve simultaneously steep slope and high on-current ($I_{on}$). However, technologically, heterojunction TFETs are much more complex than a silicon MOSFET. Our focus has been on dense monolithic integration of complementary III-V heterojunction TFETs on silicon, which may eventually evolve into a hybrid technology platform. In this paper we will give a general overview of the development of TFETs, discuss the challenges and opportunities both at the individual device level as well as in terms of technology development. In particular we will focus on the role of defects on device performance.

**Introduction**

Conventional scaling is hampered by the inability to further reduce operating voltages. The main reason for this is the limitation on the subthreshold swing of a MOSFET which determines the abruptness of the on-off transition. This is an inherent limitation of the MOSFET device physics and cannot be remedied by the introduction of novel materials or architectures, but remains an underlying physical limitation, hence scaling of operating voltages beyond 0.5 V poses a problem for the MOSFETs due to their fixed turn-on subthreshold swing, SS, which is given by,

$$SS = \ln(10) \frac{k_BT}{q} \left(1 + \frac{C_d}{C_{ox}}\right)$$

where $k_B$ is Boltzmann’s constant, $T$ is the temperature, $q$ the elementary charge, $C_d$ the depletion layer capacitance and $C_{ox}$ the gate-oxide capacitance. At 300 K this results in an ideal value of 59.5 mV/dec, which is usually referred to as the 60 mV/dec limit of the MOSFET. Hence, there is a fundamental limit to how steep the turn-on of the MOSFET can be; independent of device design or charge carrier mobility in the material. Actual scaled devices usually have slightly worse $SS$ due to non-optimal electrostatics. Tunnel FETs based on band-to-band tunneling (BTBT), however, are based on different operating mechanisms and may achieve sub-thermal swings.

The surface tunnel transistor (STT) was first proposed and demonstrated by T. Baba in 1992 [1]. In this seminal paper the gated p-i-n structure was proposed along with a discussion of the operating mechanism and guidelines as to doping densities in the
different regions were proposed. The STT was proposed as an alternative three-terminal electronic device, the benefit of a potentially steeper subthreshold slope, however, was not discussed - likely because this was not yet seen as a limitation for CMOS scaling. The fabricated STT was based on a GaAs/AlGaAs mesa structure very similar to those employed later by groups at Notre Dame [2] and Penn State [3]; transistor operation was demonstrated at 77K and at room temperature, but for these first devices electrical performance was not impressive. This implementation was followed three years later by a paper from Reddick and Amaratunga, in 1995, which showed the first silicon based STT [4]. Another decade went by, before suddenly a number of experimental tunnel FET papers appear and by then it was realized that the subthreshold swing in these devices is not limited by kT/q. Silicon based TFETs were published with [5], and some of the best tunnel FETs based on carbon nanotubes were demonstrated in [6], which for the first time showed an actual sub-thermionic swing. Bhuwalka et al. demonstrated vertical Si-SiGe tunnel FET [7]. Since then we have seen the emergence of a plethora of tunnel FET geometries, initially predominantly silicon-based, as this simplifies the implementation substantially and recently more and more shifting to III-Vs as it was realized that a heterojunction with a discontinuity in either the valence or conduction band was required to achieve simultaneously a steep subthreshold swing and a high on-current.

Very recently, experimental tunnel FETs showing sub-thermionic swing over a significant current range were demonstrated [8], showing that it is possible to beat CMOS performance. However, technologically heterojunction TFETs are much more complex than a silicon MOSFET, and in particular establishing a complementary process is a challenge. Our own work initially started on Si nanowire (NW) TFETs, from here we moved on to heterojunction devices based on InAs/Si for the p-channel device and InAs/GaSb for the n-channel device.

Our own work started on harvested Vapor-Liquid-Solid (VLS) grown silicon nanowires [9], providing a homojunction TFET by in-situ n^-i-p doping, which gradually developed into vertical nanowire geometries including an InAs source on a silicon channel, fabricated first by a combination of selective epitaxy of nanowires followed by etching [10], and then followed by a more mature concept based on template-assisted-selective epitaxy [11] [12] [13]. Most recently we demonstrated, what we believe to still be the only scalable complementary III-V heterojunction tunnel FET technology platform, with InAs/Si p-TFETs and InAs/GaSb n-TFETs integrated laterally on a silicon substrate [14].
The experimental work on tunnel FETs was of course accompanied by an equally great effort on the simulation side. Initial simulation works were very optimistic with respect to achievable performance; as the predicted performance failed to materialize in actually fabricated devices, this led to an evaluation of the tunneling models used. For silicon based devices the use of non-local tunneling models is required, whereas for the less well established III-V, thorough calibration of the tunneling parameters was required to achieve accurate modeling of experimental results [15]. It also became clear that careful device design and a stringent control of defects is required to realize the promise of TFETs experimentally.

**Device Fabrication**

In this work we will focus mainly on our most recent devices, the laterally integrated complementary InAs/Si p-TFETs and InAs/GaSb n-TFETs. A schematic of the process flow is shown in Figure 2. The p⁺-Si drain for the p-TFET is defined by boron diffusion doping through openings in an oxide mask. The hole concentration is mid-10¹⁹ cm⁻³ throughout the thickness of the Si layer, determined by secondary ion mass spectroscopy (SIMS) analysis. This correlates well with four-point-probe measurements on pieces from the same Silicon-on-Insulator (SOI) wafer used for calibration of the diffusion doping process, which resulted in average active carrier concentrations in the low 10¹⁹ cm⁻³ range. The device features are defined by e-beam lithography using a hydrogen silsesquioxane (HSQ) mask and dry-etching of the SOI layer using HBr. The template is fabricated by depositing ~50 nm-thick SiO₂, and an opening at one end is created by patterning with polymethylmethacrylate (PMMA) and etching the oxide using buffered hydrofluoric acid (BHF). The Si which is to be replaced by InAs is etched in 2% TetraMethyl Ammonium Hydroxide (TMAH) at 75°C leaving a hollow oxide nanostructure, which we refer to as the template. This anisotropic etch leaves smooth (111) planes in the remaining Si. Metal-organic chemical vapor deposition (MOCVD) is used to selectively grow an InAs source within the template which is in-situ n-doped (2x10¹⁸ cm⁻³) using Si₂H₆. From previous results [12] we know that the InAs material quality is high, with electron Hall mobilities of 5400 cm²/Vs in non-intentionally doped material (4x10¹⁷ cm⁻³).

For the n-channel TFET instead the entire device structure is grown consisting of an n⁺InAs drain is grown by MOCVD at 550°C with doping of about 2x10¹⁸ cm⁻³.
Afterwards, an in-situ unintentionally doped InAs channel is grown followed by an unintentionally doped segment of GaSb.

Once the semiconductor structure is grown the template oxide is stripped in BHF, and a gate stack is deposited, consisting of 20 cycles Al₂O₃ followed by 33 cycles HfO₂ deposited at 250°C, and a gate metal of about 50nm sputtered tungsten (W) or ALD deposited TiN. This results in an equivalent oxide thickness (EOT) of ~ 1.75 nm.

**P-channel TFET**

The cross-section of the Si nanowire is about 17x27 nm², and the InAs nanowire cross section is 30x32 nm², the difference in dimensions is due to the use of a diluted HF-dip right before growth which etches the inner walls of the template. Minor thickness variations (±1-2 nm) might occur from one device to the next, as a result of variations in the oxidation process thinning down the SOI wafer. For normalization purposes an effective width, \( W_{\text{eff}} = 100 \text{ nm} \) is used, which assumes a gate-all-around (GAA) geometry and is intermediate between the values on the Si and InAs side. As the tunneling path is...
non-local, both sides of the junction matters, and a normalization simply taking the smallest geometry is not justifiable.

First measurements were carried out at this point and showed relatively poor on-current and SS with large variation between individual devices. A contact alloying step (5 min at 300°C in Ar) was performed, which improved device performance. In separate tests, transmission line measurement (TLM) structures on Si showed a strong improvement in contact resistance following the annealing step. We believe this to be due to NiSi formation visible in Figure 3.c.

The initial high contact resistance on the highly doped p⁺ Si is expected to be caused by the existence of an atomically thin boron-rich oxide at the Si interface, a known common side-effect of some types of diffusion doping [4]. This boron-rich layer is not removed in the HF steps, but is penetrated during the annealing step, as can be seen in the SEM picture showing the formation of NiSi regions underneath the Ni contact.

In Figure 3.b remnants of an interface oxide on the InAs contacts are observed and no formation of a Ni-InAs alloy is visible, thus this might still limit the $I_{on}$. However, in this case it is likely due to rapid re-oxidation of the InAs surface during the transfer to the evaporation chamber. InAs contact resistance is generally fairly low. In the past we have measured around $10^{-6}$ $\Omega$ cm$^2$ to low-doped InAs on dedicated contact test structures, and with sulphur-passivation prior to the metal evaporation which is known to reduce the contact resistance [5], we measured $9.5 \times 10^{-8}$ $\Omega$ cm$^2$. The physical gate length of the devices measured here is about 8-900nm.

![Figure 3 SEM images of devices. (a) cross-section of a pTFET device along the channel, the heterojunction is visible with an inclination according to the (111) plane resulting from TMAH etching. (b) The InAs-Ni interface shows the presence of an interface oxide, which seems to be partially perforated. (c) The Si-Ni interface shows Ni-alloy formation.](image_url)
Cross-sections of the (d) InAs nanowire (NW) segment constituting the TFET source, and (e) the Si NW channel. (f) Top view of device illustrating the lateral integration approach. The triangle on the Si side indicates the position of the Si p+ doping mask. Color is applied on top of SEM to highlight the individual regions, figure from [17].

N-channel TFET

An example of a InAs/GaSb n-channel TFET is shown in Figure 4. In comparison with the p-TFET it can be noted that the GaSb segment is again reduced compared to the InAs, this is because the GaSb is slightly attacked when the template is removed by HF. In the case of the n-TFET the silicon does not constitute part of the active device but is simply there for nucleation purposes. The entire device structure is grown in one run, and the gate lengths for the n-TFET are shorter at about 150nm physical gate length. The metal used for the n-TFET is TiN, whereas for the p-TFET it is W, we used both metals interchangeably and did not observe a significant impact on device properties.

Figure 4 SEM images of n-TFET devices. (a) False-colored SEM top-view of InAs/GaSb TFET after gate-stack patterning. The gate-recess is visible as a shaded blue area. The gate-length is about 150nm, with 50nm overlap to the GaSb source. Cross-sections of the gated (b) InAs NW segment constituting the TFET channel, and (c) the GaSb NW source. (d) Top-view image of a completed device after lift-off of metal contacts.

Electrical characterization

Transfer characteristics of both p-TFET and n-TFET are shown in Figure 5 for different $V_{DS}$ bias levels. The $I_{on}$ of the p-TFET is 4μA/μm, whereas for the all III-V n-TFET it is about an order of magnitude larger. We have been optimizing the InAs/Si TFET through several generations, which is evidenced in the excellent average slope and good turn-on characteristics. On the other hand, this is the first Template-Assisted Selective Epitaxy (TASE) implementation of an InAs/GaSb TFET, hence the gate stack and doping profiles are not optimized which results in a poor slope, and the drop-off in $I_{on}$ at large gate bias, $V_{GS}$, is a result of the onset of GaSb depletion.
Figure 5 Room temperature and DC transfer characteristics of p-TFET and n-TFET devices for different values of drain bias $|V_{DS}| = 50$ mV, 0.3 & 0.5 V. Black curves represent the gate leakage, which is negligible in both cases, figure from [16].

Diode characteristics $I_D(V_{DS})$ measured on the TFETs are shown in Figure 6. The reverse branch corresponds to the output characteristic of the tunnel FETs. Previously, we have observed negative-differential resistance (NDR) regions in our tunnel diodes [18] [19], but not in the tunnel FETs. The explanation for this different behavior is as follows: In the case of the InAs/Si junction, high carrier concentrations on both sides of the junction are required to observe the NDR. Simulation shows that this condition is only met for gate voltages above ~1V, i.e. outside of the present measurement range, see [17] for more detail.

Figure 6 Diode characteristics of p-TFET and n-TFET at 300K, the reverse branches correspond to the output characteristic of the TFET, figure from [16].

The lack of negative differential resistance (NDR) observed in n-TFETs, as opposed to p-i-n diodes fabricated on the same chip (not shown here), is attributed due to the presence of interface traps, $D_{it}$, in the gate stack which effectively suppresses the NDR.
Low-temperature characterization

We carried out temperature sweeps for both devices down to 125 K and corresponding transfer characteristics are shown in Figure 7. In both cases only a fairly small $I_{on}$ dependence is shown, in the case of the n-TFET dominated by GaSb depletion as previously mentioned. The subthreshold swing, however, show a very strong temperature dependence, due to the presence of traps.

Traps have been identified as the main impediment of TFETs when it comes to achieving sub-thermionic slopes [20]. The nature of the traps is likely very different depending on device geometry and dimensions. For example, in the present case we believe the SS of the InAs/GaSb n-TFET to be dominated by $D_{it}$ from the non-optimized gate stack. In the InAs/Si p-TFET, however, the majority of the gate overlaps the Si channel, which is expected to have relatively low oxide $D_{it}$, whereas the large lattice mismatch of 11% is expected to contribute significant trap densities at the heterojunction.

Figure 7 Transfer characteristics for p-TFET and n-TFET as a function of temperature from room temperature down to 125 (150) K, for a $V_{DS}$ bias of 0.5V, figure from [16].
We have carried out extensive TCAD simulation of our InAs/Si devices [21], using simulation parameters carefully calibrated to first tunnel diodes and then full TFETs. As seen in Figure 8 this results in excellent matching of experiment and simulation at both room-and low-temperature. By turning on and off various trap mechanisms (Figure 8.B) we can observe that although band-to-band-tunneling is responsible for the $I_{on}$ at high $V_{GS}$, the entire transient is dominated by traps, which in our case originate at the InAs/Si heterojunction, whereas oxide $D_{it}$ only plays a minor role. More importantly, it can also be observed that during the entire transient it is really trap-assisted mechanisms which dominate. Since these are thermionic processes they result in a slope similar to the 60mV/dec limit of a MOSFET. If no traps were present the subthreshold swing at the higher current levels relevant for transport would be steeper, whereas since the on-set of BTBT is more gradual at the lower current levels SS in this case would actually be less abrupt. In [21] specific device optimizations to improve performance are investigated.

**benchmarking and outlook**

A comparison of the performance of our different generations of TFETs is shown in Figure 9 – all correspond to p-channel devices. Here it can be seen that our first silicon homojunction TFETs based on VLS grown nanowires achieved reasonably low subthreshold swing, even by today’s standard. However, current levels were substantially lower than that which is achievable by the use of the III-V/Si heterojunction devices.

The main difference between the vertical and lateral InAs/Si TFETS is the shrinking of the nanowire dimensions going from about 100nm in cross-section to below 30nm. This results in a much-improved electrostatic control. Comparison with the simulation
data, shows that the larger vertical devices are limited by the traps at the oxide interface at room temperature, and only at lower temperatures by the heterojunction traps. In the more scaled lateral devices however, the oxide interface is less important, and performance is completely dominated by the heterojunction.

Figure 9 Subthreshold swing as a function of drain bias at room temperature, comparing the different generations of TFETs from our group. The black line with stars corresponds to silicon homojunction devices with 5nm HfO2, the red and blue lines correspond to vertical InAs/Si TFETs with Al2O3/HfO2 gate dielectric of different thickness, whereas the green line with filled circles represents the lateral InAs/Si p-TFET with an intermediate thickness Al2O3/HfO2 gate dielectric.

In Figure 10 we compare the SS vs. $I_D$ for the present work with that extracted from a number of experimental references. A few other good experimental devices are not included because they either did not include an SS vs. $I_D$ plot or $I_D$ values were not normalized to effective width. The purpose is not to discuss the merits of each one; those are all among the very best experimental TFETs in their own right, but with different merits and challenges. The majority of fabricated devices lie outside of this chart because slopes are too large or current levels much too low for practical use.
Figure 10 Subthreshold swing as a function of drain current, extracted from a number of published TFET works [8, 14, 22-26]. Values are extracted from publication figures so minor variations might appear. In all cases we have chosen $V_{DS}$ values close to 0.3 or 0.5 V.

Among all TFETs, those from the Lund group [27] stand out as having consistently achieved sub-thermionic subthreshold swings. This is extremely encouraging showing that it is possible to beat CMOS performance using tunnel FETs. Key to achieving this, as shown by their work, is the combination of extremely scaled geometries, bandgap engineering and gate alignment. As it is evident from Figure 10, our devices show state-of-the art performance, but they are limited by heterojunction defects. However, the true merit of our devices lie in being fully scalable and integrated in a CMOS-like process flow, which has been our target from the onset. In [3] we have established the basis for a complementary Tunnel FET platform, by combining these p-channel TFETs with n-channel InAs/GaSb devices.

**Conclusion**

In this work we have given an overview of TFET developments in general and presented our own work on InAs/Si Tunnel FETs which are monolithically integrated on Si using TASE - a technology which allows for VLSI compatible integration of III-V heterojunction devices directly on Si. The device scaling afforded us by this approach down to sub-30 nm, is shown to improve average SS compared to our previous works on vertical InAs/Si TFETs. Further reduction of SS is limited by defects, in our case particularly at the heterojunction.

**Acknowledgments**

We would like to acknowledge M. Hopstaken for SIMS measurements, N. Bologna from EMPA for TEM support, as well as technical assistance from A. Olziersky, U. Drechsler and S. Karg, and technical discussions with W. Riess, L. Selmi and A. Ionescu. We would also like to acknowledge partial financial support from the EU via FP7 project E2SWITCH under grant agreement No. 619509 and European Project STEEEPER under grant agreement no 257267.
References
