Sub-10-nm Fin-Width Self-Aligned InGaAs FinFETs
Alon Vardi and Jesús A. del Alamo, Fellow, IEEE

Abstract—We study the scaling properties of self-aligned InGaAs FinFETs with sub-10-nm fin widths fabricated through a CMOS compatible front-end process. Working devices with fins as narrow as 7 nm, fin aspect ratios in excess of 5, and gate lengths as short as 20 nm have been fabricated using precision dry etching and digital etch. The devices feature self-aligned metal contacts that are 20–30 nm away from the edge of the gate. FinFETs with $L_g = 30$ nm, $W_f = 7$ nm, and channel height of 40 nm exhibit a transconductance of 900 $\mu$S/\mu/m at $V_{DS} = 0.5$ V. When normalized to $W_f$, this is a record value among all III–V FinFETs, indicating that our device architecture makes efficient use of conduction along the fin sidewalls.

Index Terms—III-V, MOSFET, FinFETs.

I. INTRODUCTION

InGaAs is a promising channel material candidate for CMOS technologies beyond the 10 nm node [1]. In this dimensional range, only high aspect-ratio (AR) 3D transistors with a fin or nanowire configuration can deliver the necessary performance. There have recently been multiple demonstrations of InGaAs FinFETs but their performance has been disappointing when contrasted with planar devices. This is particularly the case for very narrow fins [2]–[11].

In this work, we leverage high-precision fin etching and a self-aligned design to demonstrate the most aggressively scaled InGaAs FinFETs to date that also display outstanding electrical characteristics. For this, we have used a Si CMOS-compatible front-end process featuring reactive-ion etching and digital etch that yields high aspect-ratio InGaAs fins with smooth high quality sidewalls [11]. Unique to the process presented here is a scaled gate oxide and greater attention to fin hard-mask definition. The large process margin that we have achieved has allowed us to carry out the first detailed scaling study of self-aligned InGaAs FinFETs where we examine the role of gate oxide thickness, fin width and gate length on important device figures of merit. Fin sidewall quality emerges as a significant issue in need of further technological and fundamental studies.

II. FABRICATION PROCESS

Our fabrication process closely follows CMOS requirements, particularly self-alignment of refractory gate and ohmic contacts, very low thermal budget, extensive use of RIE and an entirely lift-off free process in the front end. Our process is based on a contact-first, gate-last flow developed for self-aligned planar InGaAs Quantum-Well MOSFETs that has yielded record $g_m$ values [12].

Details of the fabrication process are in Ref. 11. The device channel consists of $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ with a thickness $H_c = 40$ nm. Mo/W is first sputtered as contact metal ($\text{R}_{\text{ohm}} = 5 \Omega/\square$), followed by SiO$_2$ CVD. The gate pattern is then defined and the SiO$_2$ and Mo layers are etched by RIE. The top n$^+$ InGaAs cap is subsequently wet-etched in a well-controlled manner that yields an undercut of $\sim$20 nm.

Fins are then patterned using HSQ and RIE etched using a BCl$_3$/SiCl$_4$/Ar chemistry [13]. This yields fins as narrow as 15 nm ($W_f$) with a fin height ($H_f$) of $\sim$200 nm. The fins are highly vertical in the top $\sim$70 nm. To further trim the fin width and smooth the sidewalls, we perform several cycles of digital etch (DE) [14]. Fig. 1 shows test structures featuring 8 nm wide, 170 nm tall fins obtained after 3 DE cycles.

Immediately after the last DE cycle (within $\sim$5 min), a high-k gate dielectric is deposited by ALD. Mo is then sputtered as gate metal and patterned by RIE. The device is finished by via opening and pad formation. This is the only lift-off step at the backend of the process.

In this flow, the final gate length is defined by the recess opening in the SiO$_2$ which can reach below 20 nm. This is also the length of the fin, which is completely covered by

Manuscript received July 6, 2016; accepted July 24, 2016. Date of publication August 1, 2016; date of current version August 23, 2016. This work was supported in part by DTRA under Grant HDTRA1-14-1-0057, in part by Lam Research, in part by NSF under E3S STC grant #0939514, and in part by the Korea Institute of Science and Technology. The review of this letter was arranged by Editor J. Cai.

The authors are with Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: alonva@mit.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2016.2596764
the gate. From extensive calibrations, we have established that the final fin width can be determined by \( W_{\text{HSQ}} \times N_{\text{DE}} \times 2 \) nm, where \( W_{\text{HSQ}} \) is the width of the fin-etch mask and \( N_{\text{DE}} \) is the number of digital etch cycles. This gives results that are accurate within 1 nm. Focus Ion Beam (FIB) imaging has been used to confirm our dimension calibrations. Fig. 2a shows a cut along the fin of a device with \( L_g = 20 \) nm. Fig. 2b shows cross sections of two fins that are 12 and 7 nm wide. Our devices have 10-50 fins in parallel with a 200 nm fin pitch.

A key aspect of our process is that the HSQ that defines the fin etch is kept in place. This makes our FinFETs double-gate transistors with carrier modulation only on the sidewalls. While theoretically inferior to tri-gate designs, practically, the greater simplicity of the process allows us to aggressively scale all device dimensions and implement a self-aligned process. This ultimately results in significantly better performance than prior InGaAs tri-gate MOSFET demonstrations. In addition, for high channel height to fin width aspect ratio (\( AR = H_c/W_f \)), a top gate yields diminishing returns.

Our well controlled process yields many working devices with a wide range of dimensions and drain current that scales properly with number of fins. This has allowed us to carry out detailed scaling studies. Here we present data from three process splits (samples A, B and C). The samples were processed sequentially with each run incorporating further process refinements as well as EOT scaling of the gate dielectric. The inset in Fig. 4b gives the gate oxide composition and EOT of each run. Sample A was the first run carried out with this process flow and was used to calibrate the process parameters. The results of sample B were reported in ref. 11. Sample C is the newest. Unique to it is a highly processed sequentially with each run incorporating further process splits (samples A, B and C). The samples were out detailed scaling studies. Here we present data from three proper with number of fins. This has allowed us to carry

III. RESULTS

The electrical characteristics of a typical device from sample C with \( L_g = 30 \) nm and \( W_f = 7 \) nm (\( AR = 5.7 \)), are shown in Fig. 3. Well-behaved characteristics and good sidewall control are demonstrated. The device \( R_{\text{on}} \) is 320 \( \Omega \cdot \mu \)m and a peak \( g_m \) of 900 \( \mu S/\mu \)m is obtained at \( V_{DS} = 0.5 \) V. Consistent with the double-gate nature of our devices and following common practice, all figures of merit have been normalized by the conducting periphery which, in our case, is two times the channel height. The subthreshold characteristics of the same device (Fig. 3b) indicate a saturated subthreshold swing, \( S_{\text{sat}} \), of 100 mV/dec and DIBL of 90 mV/V at 0.5 V.

In devices with \( L_g = 30 \) nm and \( W_f = 22 \) nm, a peak \( g_m \) of 1500 \( \mu S/\mu \)m is obtained. For \( L_g = 2 \) \( \mu \)m and \( W_f = 22 \) nm, \( S_{\text{lin}} \) at \( V_{DS} = 50 \) mV is as low as 68 mV/dec, indicating a high quality interface between the semiconductor sidewall and the high- \( k \) gate oxide.

IV. SCALING AND BENCHMARKING

Fig. 4 shows the scaling of \( g_m, S_{\text{sat}}, V_T \) and \( S_{\text{lin}} \) as a function of \( L_g \) for transistors with \( W_f = 20-22 \) nm from the different samples. In all samples there is a strong improvement in \( g_m \) as \( L_g \) decreases. EOT scaling also results in a noticeable increase of \( g_m \), \( V_T \) rolloff mitigation and a significant enhancement of \( S_{\text{lin}} \) over the entire \( L_g \) range. As a consequence, \( I_{\text{on}} \) shows a remarkable increase as EOT and \( L_g \) are scaled down. This is classic FinFET scaling behavior that has not experimentally been demonstrated before with InGaAs.

Fig. 5 shows \( R_{\text{on}}, \ g_m, V_T \) and \( S_{\text{lin}} \) as a function of \( L_g \) and \( W_f \) for devices from sample C. \( R_{\text{on}} \) is normalized to gate periphery (2 \( \times \) \( H_c \)) and represents the minimum value that is obtained as \( V_{GS} \) increases (typically for \( V_{GS} - V_T > 0.5 \) V). \( R_{\text{on}} - L_g \) dependencies are linear with an increasing slope as \( W_f \) scales down. This is attributed to mobility degradation. From
As a result of increased sensitivity of VT to Wf, this also suggests degradation of intrinsic transport properties perhaps as a result of increased surface scattering due to sidewall roughness. The importance of sidewall roughness is also evident from the huge improvement in gm of the Wf = 7 nm devices in sample C vs. those in ref. 11 (sample B). A key process enhancement in run C was the use of a thicker adhesion layer for the HSQ fin mask that yielded better edge definition. In our process, a thin layer of Si3N4 is deposited prior to HSQ spin coating. The layer thickness was modified from 2 nm in sample B to 3 nm in sample C. This resulted in improved Si3N4 coverage which provided better support and adhesion to the highly scaled HSQ lines.

gm shows marked improvement as Lg is reduced but degrades as Wf shrinks. Since RsTd is roughly the same for all Wf, this also suggests degradation of intrinsic transport properties, perhaps as a result of increased surface scattering due to sidewall roughness. The importance of sidewall roughness is also evident from the huge improvement in gm of the Wf = 7 nm devices in sample C vs. those in ref. 11 (sample B). A key process enhancement in run C was the use of a thicker adhesion layer for the HSQ fin mask that yielded better edge definition. In our process, a thin layer of Si3N4 is deposited prior to HSQ spin coating. The layer thickness was modified from 2 nm in sample B to 3 nm in sample C. This resulted in improved Si3N4 coverage which provided better support and adhesion to the highly scaled HSQ lines.

VT rolloff (Fig. 5c) shows several trends. First VT shifts to more positive gate voltages as Wf scales down. This is partly a result of mobility degradation with Wf and the presence of a delta-doped layer underneath the channel. For Wf < 15 nm, the observed enhanced VT shift is a result of quantum size effects [16]. The dependence of VT on Lg shows an anomalous behavior in that VT rolloff increases as Wf decreases. This could be the result of line edge roughness which is more prominent in narrow fin devices and becomes increasingly important for long fins [17]. Quantum effects are also expected to enhance the impact of line-edge roughness in sub-10 nm Wf as a result of increased sensitivity of VT to Wf [16]. To further analyze the electrostatic scaling behavior of the devices, Fig. 5d plots Slin, as a function of Lg normalized to the natural electrostatic scaling length. For reference, a red line is plotted to indicate the ideal scaling behavior [18].

All devices show a similar trend regardless of Wf. This is what is expected when the data are plotted in this way. However, Slin is consistently above the ideal value. This suggests that interface states at the sidewalls play a role in limiting the electrostatic control of the channel by the gate.

In Fig. 6a, gm is normalized in the traditional approach, by the conducting gate periphery. This normalization is relevant for performance as the drain current and the intrinsic gate capacitance both scale with gate periphery. This figure leaves clear that while the latest generation of Si FinFETs features sub-10 nm wide fins with AR > 5, until our recent results, the narrowest InGaAs FinFETs in the literature had a fin width of 15 nm and AR of at most 1.8. Our results reported here extend this range to a minimum Wf of 7 nm and AR > 5 while preserving high gm.

A different perspective can be obtained by normalizing gm by the fin width, as in Fig. 5b. This figure of merit is relevant for transistor density as it emphasizes the need to obtain high current out of a small fin footprint. When normalized this way, a significant gap in performance between Si FinFETs and InGaAs FinFETs becomes evident. This stems from the high aspect ratio of Si fins and reveals the significantly greater effectiveness in charge control modulation through the sidewalls in Si FinFETs when contrasted with InGaAs FinFETs. Our sample C devices reported here, improve over prior InGaAs FinFET demonstrations by nearly a factor of 2 and approach Si performance.

V. Conclusions

We demonstrate self-aligned InGaAs FinFETs with extremely narrow fins (down to 7 nm), high channel aspect ratios (as high as 5.7), short gate lengths (down to 20 nm) and improved performance. When scaled by the fin footprint, our transistors improve the state of the art by nearly a factor of two, suggesting effective channel charge control from the sidewalls of very thin, high aspect-ratio fins.

Acknowledgments

Device fabrication was carried out at the Microsystems Technology Laboratories and the Electron Beam Lithography Facility at MIT.
REFERENCES


