RZ 3952 (#ZUR1807-070) 31/07/2018
Electrical Engineering 4 pages

Research Report

High Performance Quantum Well InGaAs-On-Si MOSFETs With sub-20 nm Gate Length For RF Applications

C. B. Zota, C. Convertino, Y. Baumgartner, M. Sousa, D. Caimi and L. Czornomaz

IBM Research – Zurich
8803 Rüschlikon
Switzerland

This is the accepted version of the article published by IEEE: C. B. Zota, C. Convertino, Y. Baumgartner, M. Sousa, D. Caimi and L. Czornomaz, "High Performance Quantum Well InGaAs-On-Si MOSFETs With sub-20 nm Gate Length For RF Applications," Proc. IEEE Internat. Electron Dev Meeting, 39.4 (2018).
10.1109/IEDM.2018.8614530

Research
Africa • Almaden • Austin • Australia • Brazil • China • Haifa • India • Ireland • Tokyo • Watson • Zurich

High Performance Quantum Well InGaAs-On-Si MOSFETs With sub-20 nm Gate Length For RF Applications

C. B. Zota, C. Convertino, Y. Baumgartner, M. Sousa, D. Caimi and L. Czornomaz
*IBM Research GmbH Zürich Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland
E-mail: zot@zurich.ibm.com

Abstract—We demonstrate RF-compatible quantum well InGaAs MOSFETs integrated on Si substrates, with L_G down to 14 nm and a Si CMOS compatible RMG fabrication flow. Devices exhibit simultaneously extrapolated f_t and f_{max} of 370 and 310 GHz, respectively, the highest reported combined f_t/f_{max} for III-V MOSFETs on Si. This is enabled by the scaled L_G, g_m of 1.75 mS/µm, 8 nm source and drain spacers and raised source and drain extensions maintaining low access resistance. The use of the InP/$In_{0.75}Ga_{0.25}As$/InP quantum well offers three times higher electron mobility and a 60% increase of g_m, compared to reference devices.

I. Introduction

High-electron mobility III-V semiconductors, such as the $In_xGa_{1-x}As$ system, have been widely investigated as replacements for Si in CMOS technology, offering increased drive current at reduced V_{DD} [1][2]. They are also currently being used as the channel in state-of-the-art HEMTs for high-speed electronics, as well as considered for future mm-wave applications [3]. Integration of III-V layers on Si substrates presents challenges and opportunities in both cases. For the former, integration is necessary in order to utilize the infrastructure and technology established by Si CMOS [4]. For the latter, successful integration would enable e.g. combined RF and digital signal processing capabilities, reducing cost and enabling new functionalities, improving upon what is currently supplied by Si RF-CMOS technology [5]. Integration of HEMTs on Si is challenging due to in part having to transfer a complex heterostructure onto a Si substrate with minimal loss of electron mobility, in part due to that standard HEMT fabrication utilizes a flow which is typically not compatible with Si CMOS. In this work, we present InGaAs-on-Si quantum well MOSFETs for RF-applications, utilizing a Si CMOS compatible replacement metal gate (RMG) fabrication flow and achieving record RF performance for III-V-on-Si FETs. Using this highly scalable technology, we also demonstrate ultra-scaled RF-devices with L_G down to 14 nm, among the shortest ever reported for this type of device.

II. Fabrication

Fig. 1 shows a cross-sectional schematic of the fabricated InGaAs-on-Si quantum well RF-MOSFET. Fabrication follows a Si CMOS compatible RMG process flow with self-aligned raised source and drain (RSD) epitaxially grown contacts [6]. Compared to our previous work, we introduce here an InP/$In_{0.75}Ga_{0.25}As$/InP quantum well (QW) in the channel, an enhanced SiN_x spacer process, as well as RF-optimized device structure and layout. The fabrication proceeds with integration of the QW heterostructure on a Si wafer by direct wafer bonding, a technique compatible with large-area Si substrates as well as 3D sequential integration [7]. Reference devices without the InP top barrier layer, as well as FinFET RF-devices are also fabricated (the latter formed by a fin dry etch of the channel). Subsequently, the dummy gate is deposited and patterned, and the 8 nm SiN_X spacers are formed by atomic layer deposition (ALD) and reactive ion etching (RIE). RSD spacers extensions are then formed by several cycles of controlled oxidation and etching, which form a cavity under the spacers as well as remove the InP top barrier in the contact regions. Following, metal-organic chemical-vapor deposition (MOCVD) regrowth of 25 nm n+ InGaAs (N_D ~ 1e19 cm^{-3}) RSD contacts is performed, which fills the cavities under the spacers to form the RSD extensions. An interlayer dielectric layer (ILD) is then deposited and planarized by CMP, followed by stripping of the dummy gate. Subsequently, the Al_2O_3 & HfO_2 (EOT ~1 nm)/TiN high-k metal gate is deposited by ALD, Fig. 2(a) shows a cross-sectional STEM image of the device at this stage. Following, W is sputtered and patterned as the gate top-metal for reduced gate resistance. Next, a second ILD is deposited, and the M1 contacts are deposited and patterned. For devices with more than two gate fingers, a second metal level is deposited to connect the sources of the device. Fig. 2(b) shows a cross-sectional SEM image of a fabricated device with two gate fingers, as well as a top-view SEM image in Fig. 2(c), showing the layout of the device.

Fig. 3 shows a HRTEM image of the channel region, including an energy-dispersive X-ray spectroscopy (EDX) map of the QW heterostructure. From the EDX map, the dimensions of the QW are determined to be 2 nm InP/10 nm InGaAs/20 nm InP. The InGaAs layer shows high crystalline quality, as well as a near-perfect InP/InGaAs interfaces.

III. Results

Devices are first characterized under DC conditions. Fig. 4 shows output characteristics of devices with L_G = 20, 60 and 120 nm, respectively. The presence of the 2 nm InP top barrier decreases resilience against short channel effects (SCE) (g_d of references devices without the top barrier is 50% lower), but L_G = 20 nm devices nevertheless show relatively healthy output behavior. The on-resistance, R_{ON}, for the three devices is 400, 470 and 525 Ωµm, respectively. Transfer characteristics for the L_G = 20 nm device is shown in Fig. 5. Peak transconductance, $g_{m,peak}$, reaches 1.25 mS/µm at V_{DS} = 0.5 V, as well as maximally 1.3 mS/µm at V_{DS} = 0.9 V. The g_m peaks at V_{GS} = 0

V, which is optimal for RF-applications due to improved gate oxide reliability. Devices without the InP top barrier peak instead at 0.3 V, which indicates a reduced influence of the interface traps (i.e. Fermi level pinning) in the former. Fig. 6 further shows $g_{m,peak}$ versus L_G for devices with and without the 2 nm InP top barrier. In scaled devices, i.e. those operating in the quasi-ballistic regime, thus scaling with the transmission, rather than the mobility, $g_{m,peak}$ is ~60% higher with the top barrier [8]. For long-channel devices, i.e. those operating close to the drift-diffusion regime, $g_{m,peak}$ is approximately 3x higher using the top barrier, indicating a similar difference in electron mobility between the two types of devices. We calculate the mobility from the slope of R_{ON} versus L_G, as shown in Fig. 6, by approximating the oxide capacitance. 1500 and 500 cm^2/Vs is obtained for devices with and without the top barrier, respectively. This difference is assumed to be caused by reduced surface roughness and oxide defect scattering using the top barrier. From the y-axis intercept, the extrinsic resistance $R_{ext} \approx 400$ Ωμm is obtained. This parameter can be further analyzed considering the schematic of the total contributions to $R_S = R_{ext}/2 = R_C + R_A + R_{Sp}$ shown in the inset of Fig. 8. From TLM measurements (Fig. 8) we determine $R_C = 75$ Ωμm and $R_A = 25$ Ωμm, and deduce the spacer resistance $R_{Sp} = 100$ Ωμm. Device performance is thus limited by the R_{Sp}. Nevertheless, R_S is comparable to that of state-of-the-art HEMTs, which utilize modulation doping to obtain very low R_{Sp}, but incur a penalty in the vertical direction due to the presence of an InAlAs barrier in the contact regions [9].

RF-characterization was performed up to 45 GHz with a Keithley vector network analyzer. De-embedding using on-chip open and short structures was performed up to (but excluding) M1. Fig. 9 shows a gain plot of a two-finger, $L_G = 20$ nm device exhibiting cutoff frequency $f_t = 370$ GHz and maximum oscillation frequency $f_{max} = 310$ GHz at $V_{DS} = 0.9$ and $V_{GS} = 0$ V. Here, f_t and f_{max} are extrapolated at -20 dB/decade using a hybrid-π small signal model with excellent fit to the measured S-parameters [10]. This represents the highest combined f_t and f_{max} for a III-V MOSFET on Si. Fig. 10 and 11 show f_t and f_{max}, respectively, versus L_G. The inset of Fig. 10 shows a STEM image of the channel region in a $L_G = 14$ nm device, one of the shortest L_G III-V RF-MOSFETs or HEMT devices ever fabricated, enabled by the highly scalable process flow presented in this work. f_t peaks at 370 GHz for $L_G = 20$ nm, while f_{max} peaks at 360 GHz at $L_G = 35$ nm due to lower gate resistance, R_G. Fig. 12 shows g_m and g_d at $V_{DS} = 0.9$ V, obtained from the hybrid-π model at $f = 10$ GHz. g_m peaks at 1.75 mS/μm for $L_G = 30 – 35$ nm and is somewhat reduced for shorter L_G due to short channel effects. The g_m frequency dispersion is minimal, indicating only minor influence of oxide border traps compared to other reports for III-V MOSFETs. This is explained by an increased tunneling distance from the channel to the border trap due to the InP top barrier. Fig. 13 further shows the voltage gain $A_V = g_m/g_d$ for both FinFET and planar RF-devices. Planar devices exhibit $A_V = 5$ at optimal L_G, while FinFETs exhibit $A_V = 20$ to 30 at similar L_G. This shows a potential advantage of FinFETs for RF-applications. FinFETs fabricated here, however, show reduced $f_t/f_{max} = 150/150$ GHz, due to increased parasitic capacitances coming from insufficient fin spacing scaling.

Fig. 14 shows C_{gs} and C_{gd} versus L_G. C_{gs} contains contributions from the parasitic capacitances $C_{gs,par}$ (primarily between the RSD and the gate, as well as the gate metal and the S/D W plugs), the oxide capacitance C_{OX} as well as the quantum capacitance C_Q, which becomes significant for very scaled EOT such as in this work. Thus, as can be seen, scaling L_G from 20 to 14 nm offers only a small reduction of C_{gs}, 0.7 to 0.65 fF/μm, due to C_{gs} being dominated by $C_{gs,par}$ and C_Q. Together with the reduction of g_m due to SCE, this explains why peak f_t is obtained at $L_G = 20$ rather than 14 nm. Fig. 15 shows f_t and g_m versus the distance d between the source/drain and the gate, from 200 to 1900 nm, at $L_G = 20$ nm. In this range, $C_{gs} + C_{gd}$ is reduced by 30% due to decoupling of the parasitic capacitance between the gate metal and the source/drain W plugs, while g_m is reduced from 1.45 to 1.05 mS/μm due to increased R_A. Peak f_t is obtained at $d = 400$ nm due to an optimal combination of $C_{gs} + C_{gd}$ and g_m.

Fig. 16 shows a benchmark of f_t and f_{max} for III-V-on-Si MOSFETs as well as state-of-the-art Si RF-CMOS [11]-[16]. Dashed traces show geometric means, $\sqrt{f_t \times f_{max}}$. The devices shown in this work represent the first demonstration of a Si CMOS compatible III-V technology clearly outperforming state-of-the-art Si RF-CMOS.

IV. Conclusions

We have demonstrated quantum well InGaAs RF-MOSFETs integrated on Si substrates using a Si CMOS-compatible self-aligned RMG process flow with L_G down to 14 nm. $L_G = 20$ nm devices exhibit extrapolated $f_t = 370$ GHz and $f_{max} = 310$ GHz, the highest reported for III-V MOSFETs on Si. This is the first demonstration of III-V-on-Si clearly outperforming Si RF-CMOS, showing that III-V's could make a significant impact in this field.

Acknowledgment

This work was funded by Horizon 2020 grant agreement numbers 688784 (INSIGHT) and 687931 (REMINDER). The authors gratefully acknowledge the support of the BRNC operations team as well as the MIND group.

References

[1] H. Riel, et al., MRS Bulletin, vol. 39, no. 8, p. 668, 2014.
[2] X. Sun et al., in VLSI Techn. Symp., T3-4, 2017.
[3] X. Mei et al., IEEE Electron Device Lett., vol. 36, no. 4, p. 327, 2015.
[4] S.-H. Kim et al. IEDM Tech. Dig., p. 429, 2013.
[5] E.-Y. Jeong et al., in VLSI Techn. Symp., T11-2, 2017.
[6] C. Zota et al., in VLSI Techn. Symp., T15-5, 2018.
[7] L. Czornomaz et al., IEDM Tech. Dig., p. 23.4.1, 2012.
[8] R. Kim et al., IEEE Trans. Electron Devices, vol. 7, no. 6, p. 787, 2008.
[9] J. Wu et al., IEEE Electron Device Lett., vol. 39, no. 4, p. 472, 2018.
[10] I. Kwon et al., IEEE Trans. MTT., vol. 50, no. 6, p. 1503, 2002.
[11] J. Singh et al., in VLSI Tech. Dig., p. 31, 2017;
[12] B. Sell et al., IEDM Tech. Dig., p. 685, 2017.
[13] A. Leuther et al., Proc. EuMA, p. 130, 2017.
[14] S. Johansson et al., IEEE Electron Device Lett., vol. 35, no. 5, p.518, 2014.
[15] C. Zota et al., IEEE Trans. Electron Devices, vol. 61, no. 12, p.4078, 2014.
[16] D.-H. Kim et al., IEEE Electron Device Lett., vol. 34, no. 2, p. 196, 2013.

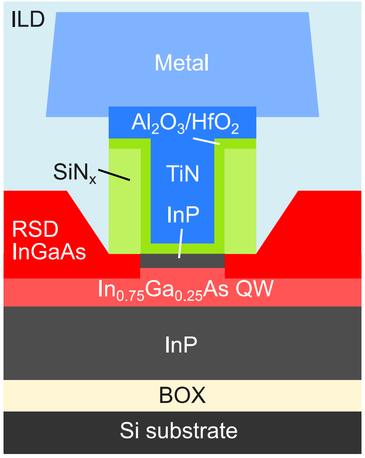

Fig. 1. Schematic cross-section of a fabricated device, with a HKMG, raised source and drain epi, SiN_x spacers and a quantum well InGaAs channel.

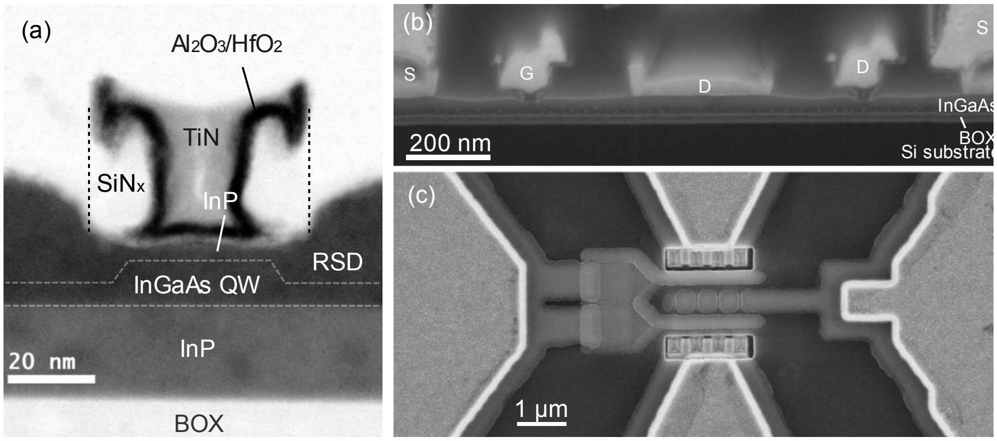

Fig. 2. (a) Cross-sectional STEM pictograph of a fabricated device with L_G = 18 nm, before deposition of the M1 level. (b) Cross-sectional SEM pictograph of fabricated devices after M1. (c) Top-view SEM pictographs of a fabricated device with two gate fingers. Devices with 4 gate fingers were also fabricated, in which case an additional metal line connected the sources together.

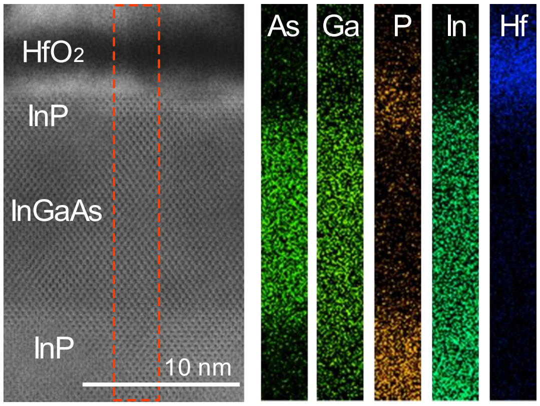

Fig. 3. HRTEM pictograph and EDX map of the channel region. 2 nm InP/10 nm $In_{0.75}Ga_{0.25}As$/20 nm InP is determined for the channel heterostructure. High crystal quality is observed for the InGaAs layer, as well as the InP/InGaAs interface, which gives rise to the strong increase of mobility as compared to devices without the InP top barrier.

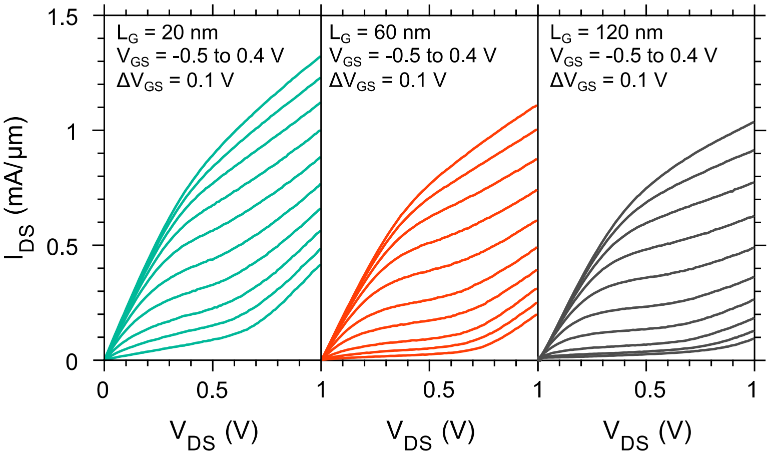

Fig. 4. Output characteristics of fabricated devices with L_G = 20, 60 and 120 nm. The 20 nm device shows significant short-channel effects at high bias due to the presence of the 2 nm InP top barrier. The on-resistances are 400, 470 and 525 Ωµm, respectively.

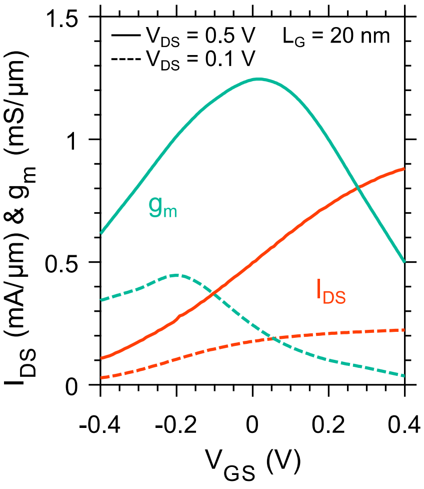

Fig. 5. Transfer characteristics of a device with L_G = 20 nm (same as in Fig. 4 and 11). Peak transconductance of 1.25 mS/µm is obtained at V_{DS} = 0.5 V.

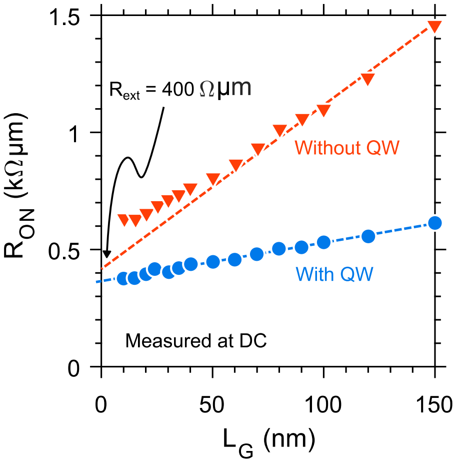

Fig. 6. Average values of the on-resistance versus gate length, for devices both with and without the 2 nm InP top barrier. The extrinsic resistance R_{ext} is approximately equal, but the mobility is 1500 cm^2/Vs and 500 cm^2/Vs, with and without the InP, respectively.

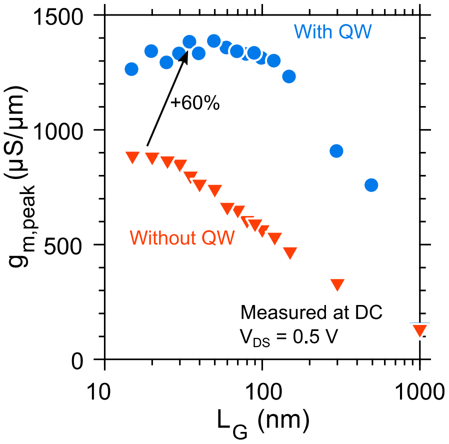

Fig. 7. Peak transconductance for devices with and without the 2 nm InP top barrier. The increased mobility yields a 60% improvement at scaled L_G for devices with the QW, and a 3x improvement at long L_G, i.e. in the drift-diffusion regime, corresponding to the increase of mobility.

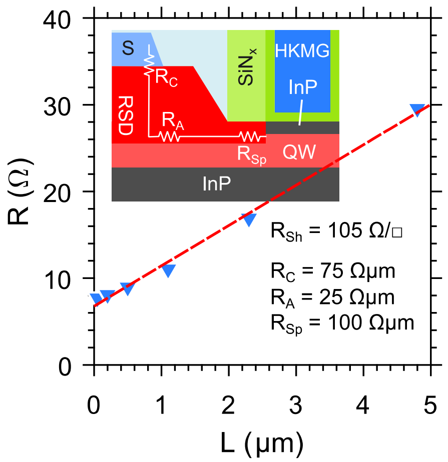

Fig. 8. TLM measurements of van der Pauw structures. Inset shows a schematic of the contact region with associated contributions to the total extrinsic resistance $R_{ext}/2 = R_C + R_A + R_{Sp}$.

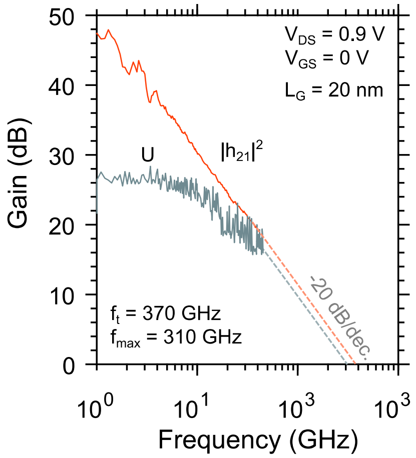

Fig. 9. Gain plot for the highest performing device. f_t and f_{max} are obtained from -20 dB/decade extrapolations confirmed by a small signal model with excellent fit to the measured S-parameters.

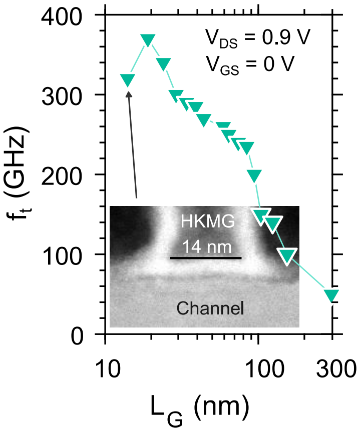

Fig. 10. f_t versus gate length. f_t peaks at 370 GHz for L_G = 20 nm. Inset shows a STEM image of the channel for the L_G = 14 nm device, which is among the shortest reported for an RF-MOSFET or HEMT.

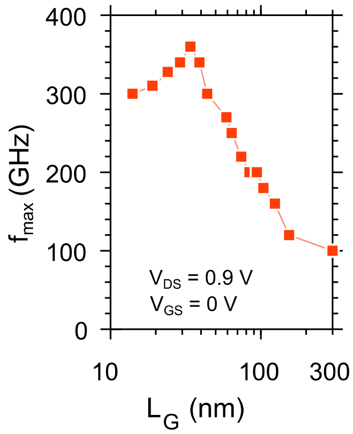

Fig. 11. f_{max} versus gate length. f_{max} peaks at 360 GHz for L_G = 35 nm. The reduction of f$_{max}$ for shorter L_G is due to an increase of R_G, from ~30 Ω at maximum f_{max} to ~60 Ω at L_G = 14 nm.

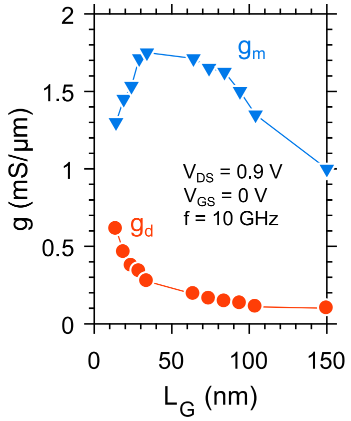

Fig. 12. g_m and g_d at V_{DS} = 0.9 V measured at 10 GHz versus L_G. For g_m, the frequency dispersion is minimal indicating only a minor impact from border traps. g_m peaks at 1.75 mS/µm but is reduced at shortest L_G due to short channel effects.

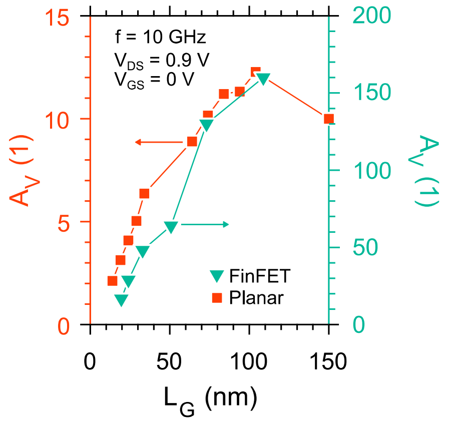

Fig. 13. Voltage gain $A_V = g_m/g_d$ for planar and FinFET RF-devices. At peak f_t, A_V is 4-5 for the planar devices, while being ~20 for the FinFETs due to strong reduction of g_d in the latter. However, FinFETs exhibit f_t/f_{max} = 150 /150 GHz due to increased parasitics.

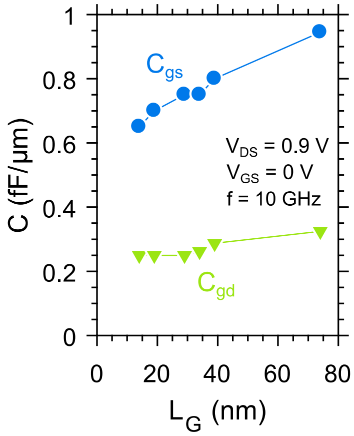

Fig. 14. C_{gs} and C_{gs} versus L_G. C_{gs} contains parasitic capacitances, as well as the oxide and quantum capacitances. Scaling of C_{gs} (through C_{ox}) with L_G is limited for this reason.

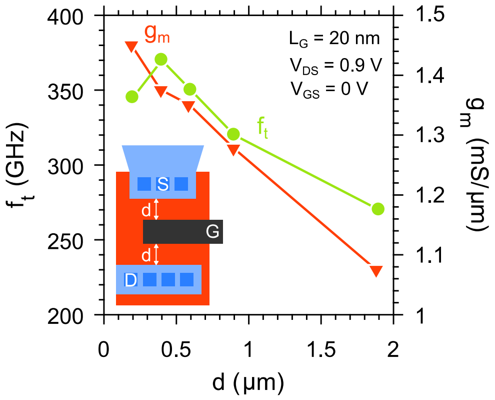

Fig. 15. f_t and g_m versus distance between source/drain and gate. In this range, g_m is reduced from 1.45 to 1.05 due to increased access resistance, while $C_{gs} + C_{gd}$ is reduced by 30%. The optimum between capacitance and g_m is found at d = 400 nm.

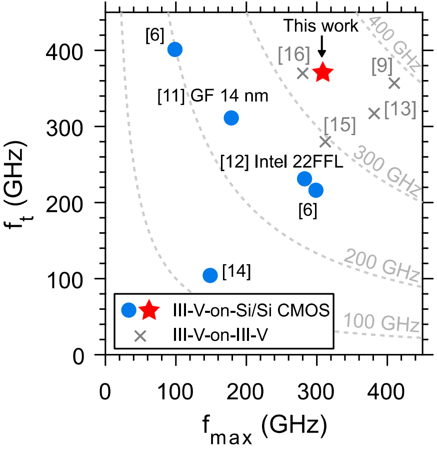

Fig. 16. Benchmark of III-V RF-MOSFETs as well as state-of-the-art Si RF-CMOS. The devices presented here for the first time clearly outperform Si RF-CMOS, as well as exhibit the highest combined f_t/f_{max} for a III-V-on-Si MOSFET.