Investigating 50nm channel length MOSFETs containing a dielectric pocket, in a circuit environment

D. Donaghy, S. Hall
Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool, L69 3BX, England
donaghy@liv.ac.uk, *s.hall@liv.ac.uk*

V. D. Kunz, C. de Groot, P.Ashburn
Department of Electronics & Computer Science, University of Southampton, Southampton SO17 1BJ, England

Abstract

The performance potential of a 50nm channel length vertical MOSFET (vMOS) architecture is assessed by numerical simulation and the use of a circuit level model. The vMOS architecture incorporates some novel features, namely a dielectric pocket for suppression of short channel effects and strategies to reduce parasitic overlap capacitance. The model allows quantification of performance limitations arising from drain and source overlap capacitances, which are shown to constitute up to 60% of the total inverter capacitance. The model allows some optimisation of performance from a circuit perspective. Comparison between a production lateral device and the vMOS device is made at the 350nm node although a 50nm channel can be realised at this node for the vMOS. The comparison is made for power supply, $V_{DD} \sim 4V_{th}$. The dual channels and the reduced gate length aid the optimised 50nm vertical structure to operate 40% times faster at, V_{DD}=1V than the 350nm lateral device at V_{DD} = 3V for a fan-out, N=10. For a loaded inverter, the performance advantages increases to a factor of 3.5. The model further predicts that the vertical 50nm inverter at V_{DD} = 1V has an energy-delay product an order of magnitude lower than the 350nm lateral inverter at V_{DD}=3V for a fan-out of 10.

1. Introduction

As technology is scaled down from the deep-submicrometer to decananometer region, device sizes and isolation spacings become comparable to depletion layer widths. To reduce some of the short-channel problems a vertical structure containing a dielectric pocket between the body/drain junction is proposed. Vertical transistors can overcome scaling problems due to lithography resolution [1], whereby decananometer channels can be realised with relaxed lithography as the channel length is determined by the accuracy of ion-implantation or epitaxial growth. The 2001 ITRS roadmap also recognises that vertical structures allow double gate devices thus increasing current drive. For long channel devices there is some scope for reducing the device area while maintaining good off-state characteristics. This is ideal in memory circuits or in ESD/output buffer circuitry. Vertical structures without body contacts can half the transistor area [2]. The use of a dielectric pocket can reduce charge sharing between the body and the gate, thus allowing better threshold voltage control or the channel length to be shortened [3,4]. The pocket serves to suppress high leakage currents due to the parasitic bipolar transistor. Furthermore the pocket prevents boron diffusion into the body during device fabrication and so mitigates bulk punch-through. The vMOS is compared here to a 350nm lateral production device for gate lengths of 50nm and 350nm at the 350nm technology node for the condition that power supply, $V_{DD} \sim 4\ V_{th}$ (threshold voltage).

2. VMOS device modelling

Fig. 1 shows the generic device architecture. A basic process to realize such a device is as follows. Body regions are implanted for pMOS and nMOS followed by growth of an oxide layer to form the dielectric pocket. A layer of polysilicon is then deposited to form the extrinsic drain contact. The turret is then etched. A silicon epitaxial layer is grown over the entire surface followed by gate oxide growth.

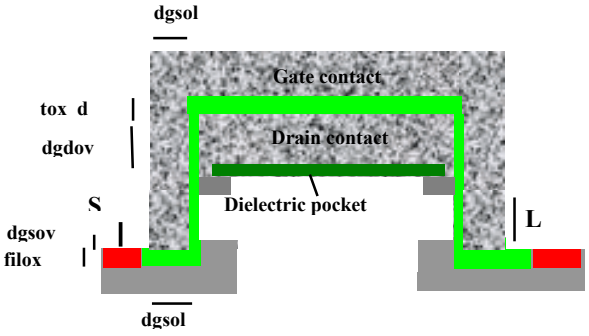

Fig 1: vMOS with dielectric pocket in drain region

A gate electrode is formed by deposition of poly-Si followed by appropriate implant. Dopant is out-diffused from the drain polysilicon layer into the single crystal body to form the intrinsic drain contact.

To obtain model parameters for circuit simulation the device was modelled using the ISE tool with the Van Dort quantum correction model, hydrodynamic model, avalanche and band-to-band tunneling all switched on to provide self-consistent data for short-channel, highly doped devices. We assume the drain encroaches under the dielectric pocket by 25nm and pocket thickness is set initially at 10nm. The turret width is set to a minimum feature size of 350nm. The gate oxide thickness was varied as well as the body doping and the results of V_{th} and I_{off} for the pMOS are summarized in figure 2.

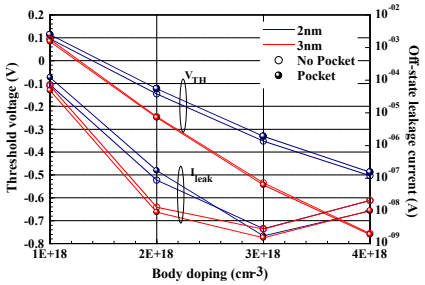

Fig. 2 50nm channel length p-vMOS: Threshold voltage and off-state leakage current versus body doping for 2nm & 3nm gate oxide thickness.

The threshold voltage was determined by linear extrapolation of transconductance, g_m(Vgs) to zero. The off-state current was taken as the current measured at V_{DS} = –1V and V_{GS} = 0V. The leakage current is a minimum for a body doping of $3x10^{18}cm^{-3}$; band-to-band tunnelling becomes increasingly dominant beyond the maximum. A 2nm gate oxide and a body doping of $3x10^{18}cm^{-3}$ result in a threshold voltage of –0.32V and a leakage current of 1nA/μm at a drain to source voltage of –1V. The dielectric pocket increases the threshold voltage by 20mV and reduces the off-state leakage by about a factor of two. Depletion capacitance associated with the drain contact is reduced by lower body doping

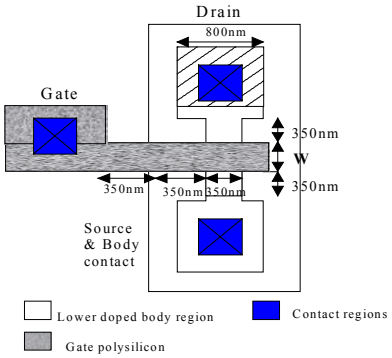

Figure 3. vMOS Plan – view

under the contact as shown by the shaded region in Fig. 3. The gate to source capacitance is reduced by increasing the field oxide above the source [5]. Gate-drain capacitance is reduced by an oxide layer on the turret top under gate poly-Si. The deposited oxide layer on top of the turret serves to minimise gate to drain overlap capacitance caused by the poly-Si gate track over the top of the turret which provides for dual channel operation. The oxide thickness on the side of the poly-Si extrinsic drain contact is set to 5nm and this is explained later.

3. Circuit model

A unity fan-out CMOS inverter circuit as shown in Fig. 4 and parasitic capacitances identified.

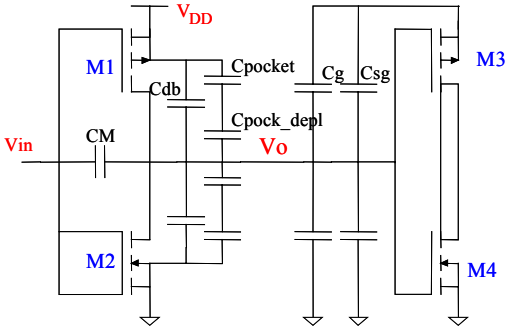

Figure 4. Circuit used as a test bench

V_{th}'s of the lateral device were +0.6/-0.74V. As mentioned above, V_{th} for vMOS devices were obtained from simulations to be +/- 0.3V for channel length L = 50nm and V_{DD} = 1V. For V_{DD} = 3V, V_{th}'s were set the same as the lateral devices. Thus the devices considered operate with $V_{DD} \sim 4\ V_{th}$.

The capacitive components of the device structure were identified and lumped together into one single capacitor, C_T assigned to the output node, V_o. As the gate to source/drain overlap capacitances (C_{gso}/C_{gdo}), are responsible for a significant portion of C_T in a vertical structure, these capacitances were isolated from the total gate source/drain capacitance, (C_{gs}/C_{gd}). Other capacitances considered were the drain to bulk depletion capacitance C_{db}, the dielectric pocket capacitance, C_{pocket} and finally the capacitance due to the drain contact (Fig. 3). The pocket capacitance takes account of the depletion region underneath the oxide layer. The voltage dependence of depletion capacitance is linearised in the usual manner. Note that vertical devices benefit from being operated 'source down' because this reduces the switched depletion capacitance compared to 'drain down' ie source at the top of the turret.

We define a circuit performance metric $\tau = 3\tau_f$ where τ_f is the fall time of V_o following an abrupt input voltage step. The fall time can be calculated by considering the discharge of the effective load capacitance, C_T, via the NMOS pull-down transistor, between 10 and 90% of the full voltage swing as indicated in the model equations below, where i(t) is the transient current through the NMOS and other symbols have their usual meaning.

$$\tau = 3\left(\tau_{fs} + \tau_{fus}\right) = -3C_T(V_O)\int_{0.1V_{DD}}^{0.9V_{DD}} \frac{dt}{i(t)}$$

$$\tau_{fs} = \frac{C_T}{2C_{ox}Wv_{sat}}\left(1 + \frac{L.\xi_{sat}}{V_{DD} - V_{th}}\right)\int_{V_{Dsat}}^{V_H} \frac{1}{(V_{DD} - V_{th})(1 + \lambda V_O)}dV_o$$

$$\tau_{fus} = \frac{C_T L}{2C_{ox}W\mu_{eff}(V_{DD})}\int_{V_{Lt}}^{V_{Dsat}} \frac{1}{(V_{DD} - V_{th})V_O - \frac{V_O^2}{2}}dV_o$$

The latter two equations represent the time the device is saturated and unsaturated respectively. The equations are easily solved numerically. To validate the model, the unity fan-out inverter circuit was simulated using the SpectreS simulator in Cadence v4.4.3 and a production 350nm lateral device SPICE deck. The simulated fall time was equal to 58ps while the calculated fall time was equal to 49ps, an error of 15% that represents reasonable agreement for a comparative study.

4. Results

Figure 5 shows that the 50nm vertical inverter is 62% faster at V_{DD}=1V than the lateral 350nm inverter at V_{DD} = 3V although the VMOS has 3 times more capacitance. The 50nm vertical inverter operating at 1V has 45% less capacitance than the vertical 350nm, inverter at V_{DD} = 3V and is 7 times faster. The capacitances due to the dielectric pocket and intrinsic drain region defined by lateral encroachment under the pocket are seen to be non-critical. However, the dielectric pocket should be as thin as possible to minimise the drain overlap capacitance.

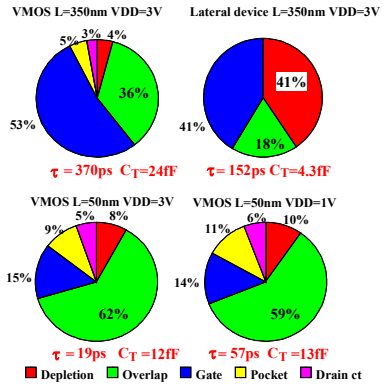

Fig.5. Relative contributions of key capacitances
$L_N = L_P$ = 350nm, 50nm, W_N = 350nm, W_P = 700nm

Overlap capacitances dominate the vertical structures, accounting for over 60% of the total capacitance compared to the dominant depletion and gate capacitance for lateral devices : contributing 41% each to the overall capacitance, C_T. Drain overlap capacitance contributes 43% of the total capacitance. An important component is that of the edge, defined by the thickness of the drain poly-Si contact; *dgdov* in Fig.1. We investigate the influence of this component in Fig.6. The sidewall oxide can be thickened by using poly SiGe or amorphous Si, which will oxidise faster, for the drain contact. We estimate that a thickness of 5nm is realistic and this will reduce C_T by 35% as shown in Fig.6.

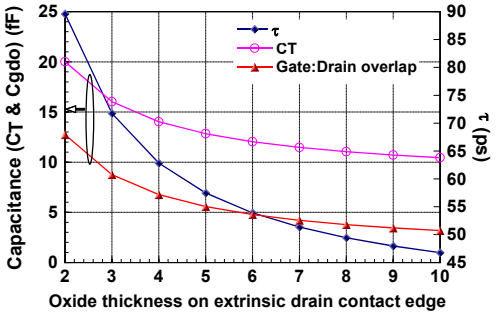

Fig. 6. Effect of varying the oxide thickness on the side of the extrinsic drain contact.
V_{DD}=1V, L_N=L_P=50nm, W_N=350nm, W_P=700nm

Fig.7 shows the influence of the oxide thickness on top of the turret (*tox_d* in Fig. 1) The minimum thickness value for the oxide is seen to be of the order 20nm.

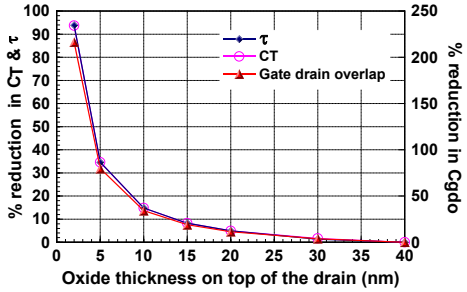

Figure 7. Effect on C_T, Cgdo & τ of reducing the thickness of the oxide on top of the drain
L_N=L_P=50nm, W_N=350nm, W_P=700nm, V_{DD}=1V.

The influence of the thickened oxide in the source (FILOX process [5]) for reducing gate-source capacitance is shown in Fig. 8. The figure shows that the oxide thickness needs to be greater than about 20nm. The capacitance contribution of the drain contact region is only 6%. This increases to 19% if the doping under the contact is set equal to the channel doping, and the speed is also reduced by 19% due to the dominance of this component.

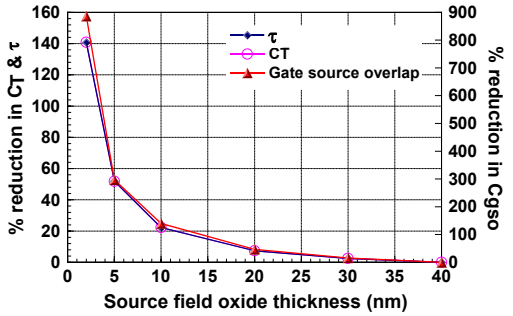

Fig. 8. Influence of the source field oxide thickness.

The effects of increasing fan-out (N) are shown in Figs. 9, 10. From Fig.9, the optimised 50nm inverter at 1V copes better with increased loading than the lateral 350nm inverter. Although the vertical devices have significantly higher capacitance than lateral devices, dual channels and the smaller channel length, enable the 50nm, vertical inverter to operate a very significant 3.5 times faster at V_{DD} = 1V than the 350nm lateral inverter at V_{DD}=3V for a fan-out of 10.

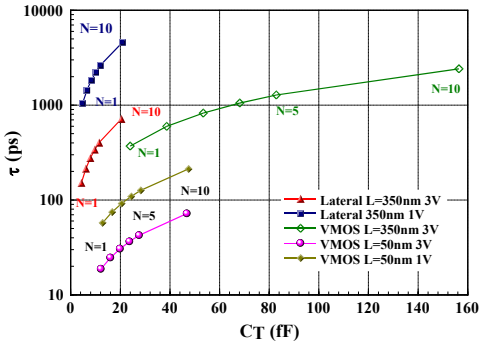

Figure 9. τ versus C_T for lateral and vertical devices for fan 1 to 10.
Wn=350nm and Wp=700nm for all devices.

The lateral inverter at V_{DD}=1V has a delay of approximately 1ns (5ns) for a fan-out of 1 (10) due to its high V_{th} voltage, thus making it impractical for high speed circuits.

Finally, the energy-delay (E-τ) metric is shown in Fig.11. E-τ of the 50nm vertical inverter at V_{DD}=1V is an order of magnitude lower than that of the 350nm lateral inverter at V_{DD}=3V for fan-out of 10. The 50nm VMOS inverter at V_{DD}=1V has 3 times lower energy-delay product than the lateral inverter at the same supply voltage. Note that the high current drive of vMOSTs will bring further advantage at smaller feature size when interconnect loading starts to dominate.

5. Conclusions

The short channel length and dual gates lead to significantly improved performance in speed at 1V supply for vMOS at the 350nm node. Overlap capacitance is the major contributor, representing 60% of effective load C_T. The model indicates that oxide regions in the source and on top of the drain both need to be about 20nm thick so as to sufficiently reduce capacitance. The oxide on the edge of the drain contact needs to be thickened also. The 50nm optimised loaded vertical inverter (V_{DD}=1V) is typically 3.5 times faster than the 350nm inverter (V_{DD}=3V) with an order of magnitude lower energy delay-product for fan-out of 10.

Finally, it should be noted that architecture presented here allows for much lower gate electrode resistance than the sidewall fillet concept of [2]. Our approach gives lower gate resistivity and allows for easier silicidation.

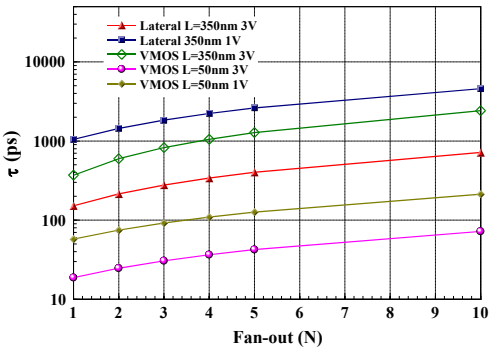

Figure 10. τ versus fan-out for a lateral and vertical devices. W_N=350nm, W_P=700nm

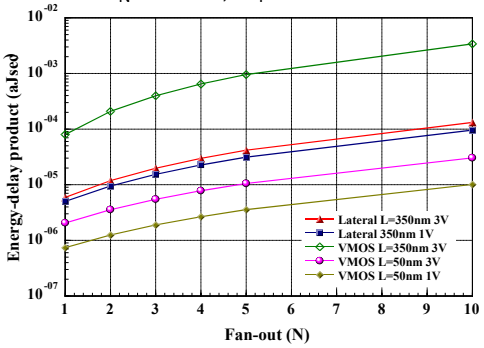

Figure 11. Energy-delay product versus fan-out.

Acknowledgement: The work was funded by UK EPSRC and the EU SIGMOS project.

7. References

[1] M. Yang, M. Carroll & J. C. Sturm, '*25-nm p-Channel Vertical MOSFET's with SiGeC Source-Drains,*', IEEE Electron Device Letters Vol.20 No.6, p.301 (1999)

[2] T. Schulz et al.,*'Short channel vertical sidewall MOSFETs..*, IEEE TED, Vol.48 No.8, p.1783 (2001)

[3] M. Jurczak et al., '*Dielectric Pockets – A New Concept ..*', IEEE TED Vol.48 No.8, p.1770 (2001)

[4] A.C. Lamb, S. Hall et al., '*A 50nm vertical MOSFET concept ..,'* ESSDERC 2001, p.352

[5] V.D. Kunz, C.H. de Groot, and P. Ashburn, Method for local oxidation in vertical semiconductor devices using fillets UK patent application no. 0128414.0, 2001.