Scattering-limited and ballistic transport in a nano-CMOS circuit

Ismail Saad\textsuperscript{a,b,*}, Michael L.P. Tan\textsuperscript{b,c}, Aaron C.E. Lee\textsuperscript{b}, Razali Ismail\textsuperscript{b}, Vijay K. Arora\textsuperscript{b,d}

\textsuperscript{a} School of Engineering & IT, Universiti Malaysia Sabah, 88999 Kota Kinabalu, Sabah, Malaysia
\textsuperscript{b} Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
\textsuperscript{c} Electrical Engineering Division, Engineering Department, University of Cambridge, 9 J.J. Thomson Avenue, Cambridge CB3 0FA, UK
\textsuperscript{d} Department of Electrical and Computer Engineering, Wilkes University, Wilkes-Barre, PA 18766, USA

\textbf{A R T I C L E  I N F O}

Keywords: Ballistic mobility Nano MOSFET Saturation velocity Quantum confinement

\textbf{A B S T R A C T}

The mobility and saturation velocity in the nanoscale metal oxide semiconductor field effect transistor (MOSFET) are revealed to be ballistic; the former in a channel whose length is smaller than the scattering-limited mean free path. The drain-end carrier velocity is smaller than the ultimate saturation velocity due to the presence of a finite electric field at the drain. The current–voltage characteristics of a MOSFET are obtained and shown to agree well with the experimental observations on an 80 nm channel. When scaling complementary pair of NMOS and PMOS channels, it is shown that the length of the channel is proportional to the channel mobility. On the other hand, the width of the channel is scaled inversely proportional to the saturation velocity of the channel. The results reported may transform the way the ULSI circuits are designed and their performance evaluated.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

The metal oxide semiconductor field effect transistor (MOSFET) is a vehicle for the design of integrated circuits both for digital and analog applications. In a nanoscale channel, the ballistic mobility and saturation velocity play an important role in limiting the MOSFET performance. There is ongoing debate over the interdependence of mobility that is controlled by momentum-randomizing scattering events and saturation velocity that depends on the streamlined motion of electrons [1–5]. Most investigators tend to converge on the theme that the saturation velocity does not sensitively depend on the low-field mobility. However, there does not appear to be a consensus on the factors limiting the current and velocity. In the work reported, the mobility degradation due to quantum confinement effect is described and utilized in evaluating nanoscale MOSFET performance. Consequently, the ultimate saturation velocity due to intense longitudinal electric field is shown to be ballistic with magnitude comparable to the Fermi velocity in strongly inverted channel. This ballistic paradigm, both for the ohmic mobility and drain velocity limited by the ultimate intrinsic velocity is utilized in modeling current–voltage (I–V) characteristics. The theory replicates well all the observed features in experiments on an 80 nm channel. When applied to complementary pair of NMOS and PMOS, new design rule is proposed.

2. Quantum confinement

The channel in a nanoscale MOSFET is indeed a quantum one that is constrained by the gate electric field forming an approximately linear quantum well [4]. Contrary to the belief of many naive researchers, no heating of the electrons by the gate is possible. Rather, the gate confines the electron to length $Z_{GM}$ comparable to the de Broglie wavelength. The average distance of an electron from the interface ($z = 0$) is

$$Z_{GM} = \frac{2(2.338E_0)}{q\hbar}$$

with

$$E_0 = \left(\frac{\hbar^2 q^2 z_0}{2m^*}\right)^{1/3}$$

and

$$\mu_t \approx \frac{V_G + V_T}{\varphi_{ox}}.$$  

where $2.338 \times E_0$ is the ground state energy, $\mu_t$ is the transverse electric field in strong inversion regime. $2.338$ is the first zero of the Airy function appropriate for a triangular quantum well, $V_T$ is the threshold voltage, $\varphi_{ox}$ the thickness of the gate oxide and $m^* = m_0 = 0.916 m_e$ is the longitudinal effective mass. The electrons do not reside at the Si/SiO$_2$ interface as wave-function vanishes there due to the quantum-confinement effect. In the nano-MOSFET, the gate oxide is a few nm ($\varphi_{ox} = 1.59$ nm in our case). The distance...
\( z_{QM} \) of the electrons that is also a few nm cannot be neglected. The gate capacitance \( C_g \) per unit area is smaller than \( C_{ox} \) and is given by
\[
C_g = \frac{\varepsilon_{ox}}{\varepsilon_{oxeff}} = \frac{C_{ox}}{1 + (1/3)(z_{QM}/t_{ox})}.
\]

3. Ballistic saturation velocity

The carrier drift due to the channel electric field driving the electrons has been explored in a number of works (see Ref. [1] for a review and Ref. [6] for sample experimental results). In the archival work of Arora [1], the ballistic transport, although not specifically mentioned by that name, was predicted in the presence of a high electric field. The theory developed was for non-degenerate bulk semiconductors that gave saturation velocity comparable to the thermal velocity. Recently, the theory has been extended to embrace all dimensions under both degenerate and non-degenerate conditions [5]. The ballistic nature of the velocity is apparent from Fig. 1 that streamlines the randomly oriented intrinsic velocity \( \nu_i \) vectors in a very high electric field.

The saturation velocity is the weighted average of the magnitude of the carrier velocity \( |v| \) with averaging weight equal to the probability given by the Fermi–Dirac distribution multiplied by the two-dimensional density of states. This results in saturation velocity being intrinsic velocity \( \nu_{i2} \) for a two-dimensional distribution given by
\[
\nu_{i2} = \nu_{th2} \frac{3j(w2)_f}{3j(w2)_r}
\]
with
\[
\nu_{th2} = \sqrt{\frac{2\varepsilon_{k} k T}{m_{i2}}}
\]
where \( j(w2)_r \) is the Fermi integral of order \( j \) with \( \eta_{F} = (E_{F} - E_{c})/k_{B} T \)
as the normalized Fermi energy with respect to the quantized band-edge \( E_{c} = E_{c}\text{sat} + 2.338 \times E_{0} \). \( E_{c}\text{sat} \) is the bulk conduction band-edge that is lifted by the zero-point energy \( 2.338 \times E_{0} \) due to the quantum confinement (QC) effect. \( T \) is the temperature and \( m_{i2} \) is the transverse effective mass in ellipsoidal band structure of silicon. In \( (100) \) configuration, only the lower two valleys with conductivity effective mass are occupied in the quantum limit. Eq. (4) shows that the ultimate intrinsic saturation velocity \( \nu_{i2} \) in a semiconductor is the two-dimensional thermal velocity \( \nu_{th2} \) only in a non-degenerate regime. In a degenerately induced MOSFET, the intrinsic saturation velocity is limited to the Fermi velocity \( \nu_{F} \) that in turn depends on surface carrier concentration \( n_{s} \):
\[
\nu_{i2}^{\text{deg}} = \frac{2}{3} \nu_{F}.
\]

The numerical work of Eq. (4) indicates that the electrons make a transition from non-degenerate to degenerate statistics as gate voltage is increased.

4. I–V characteristics of a nano-MOSFET

The current–voltage (I–V) characteristics of a nano-MOSFET can be obtained from the velocity-field characteristics as follows:
\[
I_{D} = \frac{C_{ox} \phi_{0} W/L}{2} [(1/2) V_{D} - V_{G}] 
= \frac{2}{1 + (V_{G}/|V|)}
\]
where \( \phi_{0} \) is the low-field mobility, \( W \) the channel width, \( L \) the channel length and \( V_{G} = V_{GS} - V_{Th} \). The drain voltage at the onset of saturation \( V_{Dsat} \) is less than \( V_{Dsat1} \), appropriate for carrier reaching full saturation (\( x = 1 \) below). \( V_{Dsat1} \) is given by
\[
V_{Dsat1} = V_{c} (\sqrt{1 + 2V_{G}/|V| - 1})
\]
\[
V_{c} = \frac{V_{sat}}{L} \mu_{0}
\]

The drain-to-saturation velocity ratio \( x = v_{D}/v_{sat} \) is plotted in Fig. 2 as a function of drain voltage \( V_{D} \) starting from the saturation point \( V_{Dsat} \) to \( V_{D} = 1.0 \text{ V} \). As the drain electric field becomes larger due to increased drain voltage, \( x \) tends to advance towards its ultimate value of 1. In the scenario with \( x = 1 \), the drain voltage is the ultimate saturation velocity. However, when \( x < 1 \), the drain voltage \( v_{D} \) is smaller than \( v_{sat} \) due to the presence of the finite electric field at the drain [5]. The curve at \( V_{GS} = 0.7 \text{ V} \) (upper one in Fig. 2) is representative of a long-channel behavior as the ratio \( V_{GS}/V_{D} \) is smaller for a long channel. That is the reason that \( x = 1 \) model works well for a long channel and current in the saturation region has a flat slope, making channel conductance equal to zero. However, in a nanoscale channel, as seen in the I–V characteristics of Fig. 3 for a 80 nm channel, the channel conductance is non-zero due to the increase in the value of \( x \) as the drain voltage is increased. The current at the onset of saturation region is
\[
I_{Dsat} = \nu_{c} C_{L} (V_{GS} - V_{T} - V_{Dsat} W_{sat} W, \quad V_{D} > V_{Dsat}.
\]

\( I_{Dsat} \) depends only on the gate width and not sensitively on the channel length. Hence, channel length modulation due to pinchoff point moving in the channel is a mistaken identity that does not exist either for the long or short channel. In fact, if the binomial expansion of \( V_{Dsat1} \) is performed under the long-channel approximation, the carrier concentration at the drain is always finite [5,8].

The I–V characteristics for an 80 nm channel of Fig. 3 with \( x = 1 \) (solid lines) terminate at the velocity-saturation limited current giving zero transconductance. However, when \( x < 1 \), solid lines branch out at \( V_{D} = V_{Dsat} \) to dotted line with \( x \) changing from

\[ \text{Fig. 1. Random intrinsic velocity vectors } \nu_i \text{ in a zero-field transforming to streamlined vectors in a very high electric field.} \]

\[ \text{Fig. 2. The ratio } x = v_D/v_{sat} \text{ of the drain velocity as a function of drain voltage beyond the onset of saturation } V_D > V_{Dsat} \text{ for } V_{GS} = 0.7 \text{ (topmost curve), } 0.8, 0.9, 1.0, 1.1 \text{ and } 1.2 \text{ (bottom curve).} \]
In a nanoscale channel, it is highly likely that the mobility and saturation velocity are scattering-independent and hence ballistic. The pseudo channel length modulation in the absence of quasi-pinchoff and finite output conductance is attributed to the drain velocity increasing towards the ultimate saturation velocity as electric field increase due to increase in the drain voltage. New CMOS scaling law that scales width with the saturation velocity and the length with the low-field Ohmic mobility is proposed.

5. Nano-CMOS design

In a CMOS circuit design, switching speed is enhanced when the current \( I_{Dn} \) in NMOS and PMOS transistors are the same as given in Eq. (7) with

\[
V_{Gn(p)} = V_{CSn(p)} - V_{TN(p)} \quad \text{and} \quad V_{CSn(p)} = \frac{V_{DSn(p)}}{\mu_{n(p)} L_{n(p)}}. \tag{10}
\]

In the fabrication process, the oxide thickness \( t_{ox} \) can be made the same for both transistors and threshold voltage can be adjusted so that \( V_{thn} = V_{thp} \). As length is the smallest dimension in the photolithographic process, the designers tend not to sacrifice the smallest dimensions attained and hence keep the channel length same for both channels. In a traditional design based on Ohm’s law \( (V_{DS} = V_{G}) \), the width is scaled so \( (W/W_{th}) \) product is same for current to be equal in both channels. In order to account for the lower mobility of PMOS, the width ratio of p-channel to n-channel is scaled as

\[
\frac{W_p}{W_n} = \frac{\mu_{thn}}{\mu_{thp}}. \tag{12}
\]

The above design is based on the validity of Ohm’s law that obviously does not hold good for nano-scale channels. When velocity saturation is considered, the saturation current of Eq. (9) requires \( (W/W_{th}) \) to be matched for both n- and p-channels. In the triode region \( 0 \leq V_D \leq V_{Dsat} \), the current expression of Eq. (7) including \( V_T \) must also match to complement I-V characteristics of both channels. This leads to

\[
\frac{W_p}{W_n} = \frac{V_{Satn}}{V_{Satp}}, \quad \frac{L_p}{L_n} = \frac{\mu_{thp}}{\mu_{thn}}. \tag{13}
\]

The length of the channel is scaled proportionally to the channel mobility and the width of the channel is scaled inversely proportional to the saturation velocity of the channel.

6. Conclusions

In a nanoscale channel, it is highly likely that the mobility and saturation velocity are scattering-independent and hence ballistic. The pseudo channel length modulation in the absence of quasi-pinchoff and finite output conductance is attributed to the drain velocity increasing towards the ultimate saturation velocity as electric field increase due to increase in the drain voltage. New CMOS scaling law that scales width with the saturation velocity and the length with the low-field Ohmic mobility is proposed.

Acknowledgments

The authors would like to acknowledge the financial support from E-Science grant of the Ministry of Higher Education (MOHE), I.S. and V.K.A. are thankful to the Universiti Teknologi Malaysia (UTM) for wonderful hospitality. V.K.A. acknowledges the UTM grant 77506 as a distinguished visiting professor. M.L.P.T. thanks the University of Cambridge for access to its facilities in completion of his PhD work.

References