The Evolution of Theory on Drain Current Saturation Mechanism of MOSFETs from the Early Days to the Present Day

Peizhen Yang[1], W.S. Lau[1], Seow Wei Lai[2], V.L. Lo[2], S.Y. Siah[2] and L. Chan[2]
[1]*Nanyang Technological University,*
[2]*Chartered Semiconductor Manufacturing*
Singapore

1. Introduction

Metal-oxide-semiconductor (MOS) digital logic is based on the enhancement-mode MOS transistors. During the past 40 years, the gate length of Si-based MOS transistors has been scaled down from about 10 μm to below 0.1 μm (100 nm). Currently, MOS transistors fabricated by 45 nm CMOS technology are readily available from various silicon foundries. Moreover, Taiwan Semiconductor Manufacturing Company (TSMC) has successfully developed 28 nm CMOS technology using the conventional silicon oxynitride as the gate insulator with polysilicon gate (Wu et al., 2009). IBM has demonstrated the use of high-K dielectric as the gate insulator with metal gate for their sub-22 nm CMOS technology (Choi et al., 2009). SEMATECH has developed their 16 nm CMOS technology using high-K/metal gate (Huang et al., 2009). Furthermore, several research groups have already reported on the development of 10 nm planar bulk MOS transistors (Wakabayashi et al., 2004; Wakabayashi et al., 2006; Kawaura et al., 2000). It has been reported using a hypothetical double-gate MOS transistor that a direct source-drain (S/D) tunneling sets an ultimate scaling limit for transistor with gate length below 10 nm (Jing & Lundstrom, 2002). Aggressive scaling brings about significant improvement in the integration level of Si-based MOS logic circuits. In addition, it also improves the switching speed because the drain current is increased when a smaller gate length and a smaller effective gate dielectric thickness are used. According to the conventional MOS transistor theory based on the constant electron mobility, the linear drain current (i.e. drain current at low drain voltage) will increase with the reduction of the gate length. Based on the classical concept of velocity saturation, the saturation drain current (i.e. drain current at high drain voltage) will not increase when the gate length is decreased. This theory is obviously contradictory to the experimental observation. Experimentally, we observe that the linear drain current and the saturation drain current are increased when the gate length is reduced. Hence, there is a need to investigate the drain current saturation mechanism in the nanoscale MOS transistors. First and foremost, we need to know the type of electrical conduction between the source and drain (S/D) regions for the state-of-the-art MOS transistors ($L \geq 32$ nm). Fig. 1 shows the various types of electrical conduction between the source and the drain of a n-channel MOS (NMOS) transistor (i) thermionic emission, (ii) thermally assisted S/D tunneling and (iii) direct S/D tunneling

Source: Solid State Circuits Technologies, Book edited by: Jacobus W. Swart,
ISBN 978-953-307-045-2, pp. 462, January 2010, INTECH, Croatia, downloaded from SCIYO.COM

(Kawaura & Baba, 2003). In the thermionic emission, carriers are thermally excited in the source, and then they go over the potential barrier beneath the gate. In the thermally-assisted S/D tunneling, carriers are thermally excited in the source, and then they tunnel slightly beneath the top of the potential barrier. Both thermionic emission and thermally-assisted S/D tunneling have strong temperature dependence. In contrast, the direct S/D tunneling does not need any thermal excitation and thus it has a weak dependence on temperature. Since the tunneling probability increases exponentially with decreasing potential barrier width, a decrease in the gate length will significantly increase the direct S/D tunneling and thus increase the subthreshold current (Kawaura & Baba, 2003). Fortunately, the tunneling current will only exceed the thermal current and degrade the subthreshold slope when the gate length is less than 5 nm (experimentally 4 nm and theoretically 6 nm) (Kawaura et al., 2000). Therefore, we only need to be concerned with thermionic emission between the source and the drain for the state-of-the-art MOS transistors ($L \geq 32$ nm). This chapter will discuss the evolution of theory on drain current saturation mechanism of MOS transistors from the early days to the present day. Section 2 will give an overview of the classical drain current equations that involve the concepts of velocity saturation and pinchoff. Section 3 will address the ambiguity involving the occurrence of velocity saturation and the presence of velocity overshoot in the nanoscale transistors. Section 4 will discuss the newer drain current transport concepts such as ballistic transport and quasi-ballistic transport. Section 5 will discuss the physics behind the apparent velocity saturation observed during transistor scaling and how it differs from the classical concept of velocity saturation within a transistor. Finally, Section 6 will discuss the actual mechanism behind the drain current saturation in nanoscale transistors.

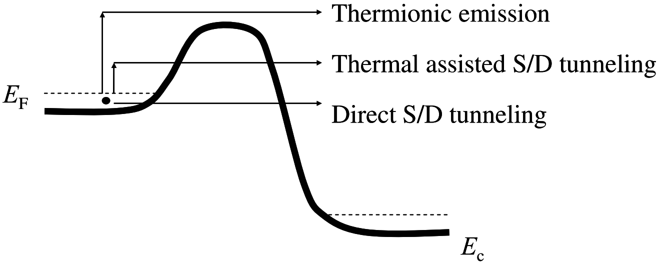


Fig. 1. Various types of electrical conduction between the source and the drain of a NMOS transistor (i) thermionic emission, (ii) thermally assisted S/D tunneling and (iii) direct S/D tunneling. Note that E_F refers to the Fermi level. E_c refers to the conduction band edge.

2. Classical drain current equations for MOS transistors

For long-channel MOS transistors (L = 10 μm), the drain current saturation is related to pinchoff (Hofstein & Heiman, 1963). A qualitative discussion of MOS transistor operation is useful, with the help of Fig.2. For NMOS transistor, a positive gate voltage (V_{GS}) will cause inversion at the Si/SiO$_2$ interface. When the drain voltage (V_{DS}) is small, the channel acts as a resistor and the drain current (I_{ds}) is proportional to V_{DS} (see Fig.3). This is known as the linear operation of the MOS transistor. The equation of the linear drain current is given by (Sah, 1991, b),

$$I_{ds} = \frac{\mu_{eff} W C_{ox}}{L_{eff}} \left[(V_{GS} - V_{th}) V_{DS} - 0.5 V_{DS}{}^{2} \right] \quad (1)$$

where μ_{eff} is the low-field mobility. W is gate width. L_{eff} is the effective channel length. C_{ox} is the gate oxide capacitance per unit area. V_{th} is the threshold voltage.

By taking the partial derivative of equation (1) with respect to V_{DS}, the expression for the drain conductance (g_d) is as follows,

$$g_d = \left. \frac{\partial I_{DS}}{\partial V_{DS}} \right|_{V_{GS}} = \frac{\mu_{eff} W C_{ox}}{L_{eff}} (V_{GS} - V_{th} - V_{DS}) \quad (2)$$

Note that g_d decreases linearly with increasing V_{DS}. At $V_{DS} = V_{GS} - V_{th}$, g_d becomes zero and thus V_{DS} loses its influence on the number of electrons that can be injected by the source. This is because the depletion layer at the drain prevents the drain electric field from pulling out more electrons from source into the channel. Since V_{GS} can decrease the potential barrier of the source-to-channel pn junction, I_{ds} can be increased by using a bigger V_{GS}. Pinchoff point occurs when the electron density in the channel dropped to around zero. The current-saturation drain voltage (V_{Dsat}) is given by,

$$V_{Dsat} = V_{GS} - V_{th,sat} \quad (3)$$

where $V_{th,sat}$ is the saturation threshold voltage.

The saturation drain current (I_{ds}) is then given by (Sah,1991,b),

$$I_{ds} = \frac{\mu_{eff} W C_{ox}}{2 L_{eff}} (V_{GS} - V_{th})^{2} \quad (4)$$

However, the constancy of I_{ds} at high V_{DS} is not maintained in the short-channel MOS transistors because the additional V_{DS} beyond (V_{GS} -V_{th}) will cause the pinchoff point to move slightly towards the source in order to deplete more electrons. This slight reduction in L_{eff} can be considered negligible for the long channel transistors but it becomes significant for the short channel transistors and thus results in a small g_d when $V_{DS} \geq V_{GS} - V_{th}$.

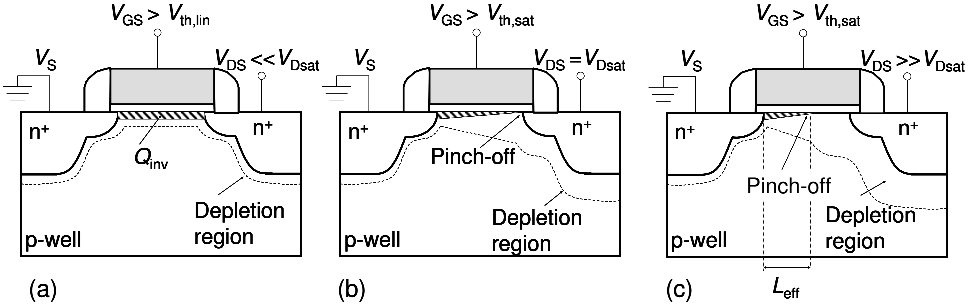


Fig. 2. NMOS transistor operating in (a) the linear mode, (b) the onset of saturation, and (c) beyond saturation where the effective channel length (L_{eff}) is reduced. $V_{th,lin}$ and $V_{th,sat}$ are the linear threshold voltage and the saturation threshold voltage , respectively. Q_{inv} is the inversion charge.

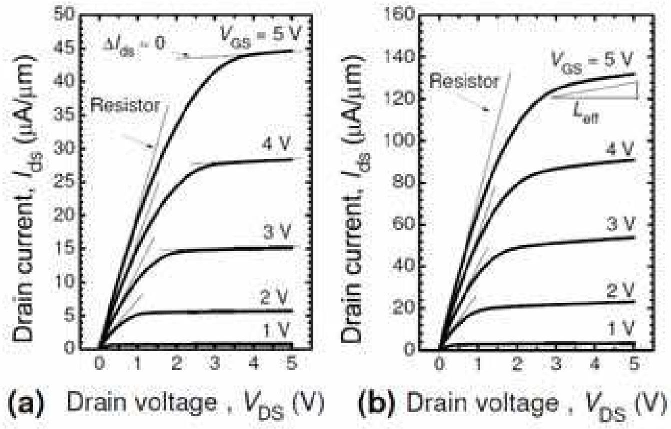
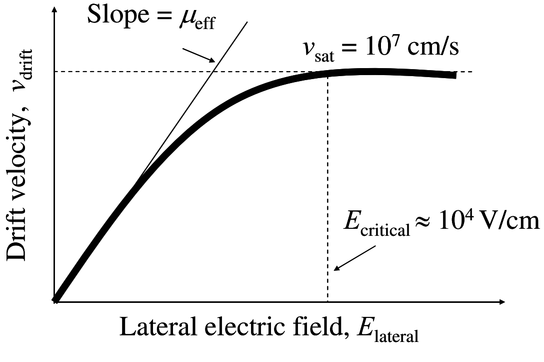


Fig. 3. Experimental I_{ds} versus V_{DS} characteristics of the NMOS transistor with physical gate oxide thickness of 300 Å (a) L =10 μm, W =10 μm, (b) L = 3 μm, W =10 μm.

For short-channel MOS transistors (L < 1 μm), (Taur et al., 1993) proposed that the drain current saturation, which occurs at V_{DS} smaller than the long-channel current-saturation drain voltage (V_{Dsat} = V_{GS} - $V_{th,sat}$), is caused by velocity saturation. From Fig.4, when the lateral electric field ($E_{lateral}$) is small (i.e. V_{DS} is low), the drift velocity (v_{drift}) is proportional to $E_{lateral}$ with μ_{eff} as the proportionality constant. When $E_{lateral}$ is further increased to the critical electric field ($E_{critical}$) that is around 10^4 V/cm, v_{drift} approaches a constant known as the saturation velocity (v_{sat}) (Thornber, 1980). Based on the time-of-flight measurement, at temperature of 300 K, v_{sat} for electrons in silicon is 10^7 cm/s while v_{sat} for holes in silicon is 6×10^6 cm/s (Norris & Gibbons, 1967).



Slope = μ_{eff}
v_{sat} = 10^7 cm/s
Drift velocity, v_{drift}
$E_{critical} \approx 10^4$ V/cm
Lateral electric field, $E_{lateral}$

Fig. 4. Schematic diagram of the drift velocity (v_{eff}) as a function of the lateral electric field ($E_{lateral}$). Note that $E_{lateral} \approx V_{DS}/\ L_{eff}$.

According to the velocity saturation model, the equation of the saturation I_{ds} for the nanoscale MOS transistor is given by (Taur & Ning, 1998, c),

$$I_{ds} = v_{sat} W C_{ox} \left(V_{GS} - V_{th,sat} \right) \tag{5}$$

In contrast with the theoretical predictions that v_{sat} is independent of μ_{eff} (Thornber, 1980), the experimental data show that the carrier velocity in the nanoscale transistor and the low-field mobility are actually related (Khakifirooz & Antoniadis, 2006). This can be better understood as follows. The effects of strain on μ_{eff} can be investigated qualitatively in a simple way through Drude model, $\mu_{eff} = q\tau / m^*$ where τ is the momentum relaxation time, m^* is the effective conductivity mass, and q is the electron charge (Sun et al., 2007). For <110> NMOS transistors that are fabricated on (100) Si substrate, there are four in-plane conduction band valleys (1, 2, 3, 4) and two out-of-plane conduction band valleys (5, 6), as shown in Fig. 5(a). The application of <110> uniaxial tensile stress will remove the degeneracy of the conduction band valleys such that the out-of-plane valleys (5, 6) will have a lower electron energy state that the in-plane valleys (1, 2, 3, 4). Since electrons will preferentially occupy the lower electron energy state, there will be more electrons in valleys (5, 6) compared to valleys (1, 2, 3, 4) and thus the effective in-plane mass becomes smaller. Besides the strain-induced splitting of the conduction band valleys, the strain-induced warping of the out-of-plane valleys (5, 6) in (100) silicon plane also plays a part in the electron mobility enhancement. In the absence of mechanical stress, the energy surface of the out-of-plane valleys (5, 6) is “ circle“ shaped and the effective mass of valleys (5,6) is m_T. When <110> tensile stress is applied, the effective mass of valleys (5, 6) along the stress direction ($m_{T,//}$) is decreased but the effective mass of valleys (5, 6) that is perpendicular to the stress direction ($m_{T,\perp}$) is increased (Uchida et al., 2005). By taking into account the change in the effective mass of the out-of-plane valleys (5, 6) and the strain-induced conduction subband splitting , the low-field mobility enhancement of the bulk <110> NMOS transistors under uniaxial <110> tensile stress can be modeled (Uchida et al., 2005).

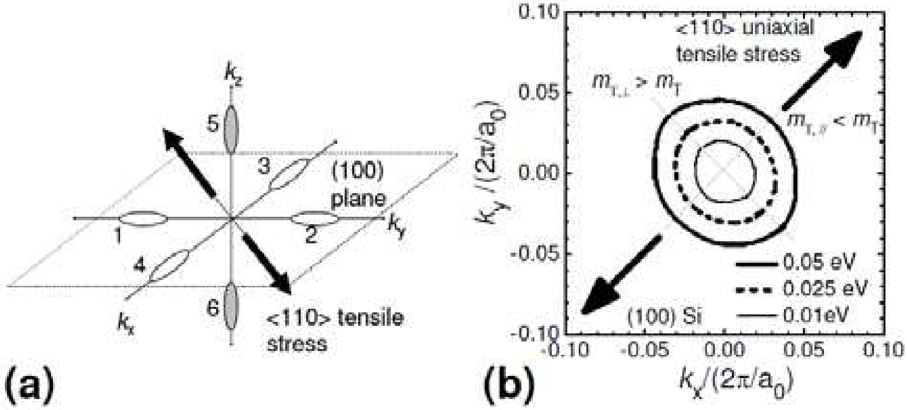


Fig. 5. Effects of <110> uniaxial tensile stress on the conduction band valleys of (100) silicon plane (a) Four in-plane valleys (1, 2, 3, 4) and two out-of-plane valleys (5,6), (b) Energy contours of the out-of-plane valleys (5, 6) , which is modified from (Uchida et al., 2005). Note that a_0 is the unstrained silicon lattice constant. k_x, k_y and k_z are the wave vectors along x direction, y direction and z direction , respectively. $m_{T,//}$ is the effective mass of valleys (5,6) along the stress direction ,and $m_{T,\perp}$ is the effective mass of valleys (5,6) in the direction that is perpendicular to the stress direction. m_T is the effective mass of valleys (5,6) in the absence of mechanical stress.

For <110> p-channel MOS (PMOS) transistors that are fabricated on (100) Si substrate, the lowest energy valence band edge has four in-plane wings (I1, I2, I3, I4) and eight out-of-plane wings (O1, O2, O3, O4). Fig.6, which is modified from (Wang et al., 2006), shows the effects of mechanical stress on the iso-energy contours of the valence band edge. In the absence of mechanical stress, the innermost contours are "star" shaped. When uniaxial compressive stress is applied along <110> channel direction, the innermost contours become oval shaped. In addition, the spacing between the contours increases for I1 and I3 wings while decreases for I2 and I4 wings. This indicates the hole energy lowering of I1 and I3 wings, and the hole energy rise of I2 and I4 wings. Since holes will preferentially occupy the lower hole energy state, there will be a carrier repopulation from I2 and I4 wings to I1 and I3 wings. As the channel length is along the direction of I2 and I4 wings, the hole mobility of <110> PMOS transistor will be improved. On the other hand, the application of uniaxial tensile stress along <110> channel direction leads to the opposite conclusion. The carriers are redistributed from I1 and I3 wings to I2 and I4 wings, leading to a hole mobility degradation in <110> PMOS transistor.

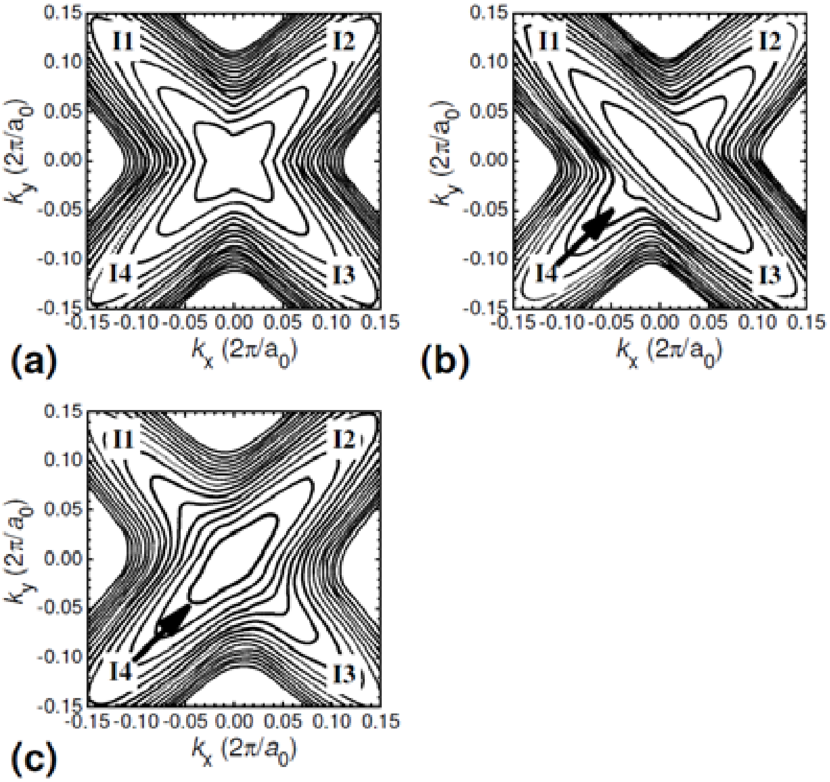


Fig. 6. Iso-energy contours separated by 25 meV in (100) silicon substrate for valence band edge, modified from (Wang et al., 2006). (a) No mechanical stress, (b) Uniaxial compressive stress along <110> direction, (c) Uniaxial tensile stress along <110> direction. Note that a_0 is the unstrained silicon lattice constant. k_x and k_y are the wavevectors along x direction and y direction, respectively. The arrow indicates the direction of the mechanical stress.

In addition to the simulation results of the strain-induced variation to the conduction band edge and the valence band edge, the change in the effective carrier mass by mechanical stress can also be studied by piezoresistance measurements. Device-level piezoresistance measurements in the channel plane can be readily done. From Table I, which is modified from (Chiang et al., 2007), the piezoresistance coefficient along the channel direction (π_L) is negative for NMOS transistor and is positive for PMOS transistor. This indicates that uniaxial tensile stress will decrease the effective carrier mass along the channel direction (m_x) for NMOS transistor but will increase m_x for PMOS transistor. In the other words, <110> tensile stress will increase the electron mobility of <110> NMOS transistor while <110> compressive stress will increase the hole mobility of <110> PMOS transistor. Since the on-state current (I_{on}) enhancement is observed in the nanoscale transistors with the implementation of various strain engineering techniques (Yang et al., 2004; C-H. Chen et al., 2004; Yang et al., 2008; Wang et al. , 2007), the carrier velocity in the nanoscale transistor must be related to the low-field mobility, and thus equation (5) needs to be modified so as to account for the strain-induced I_{on} enhancement.

Table I Device-level piezoresistance coefficients in the longitudinal direction (π_L), the tranverse direction (π_T), and the out-of-plane (π_{out}) direction for <110> channel MOS transistors that are fabricated on (100) Si substrate (Chiang et al., 2007). The units are in 10^{-11} m^2/N. Note that "longitudinal" means parallel to the direction of channel length in the channel plane, "transverse" means perpendicular to the direction of channel length in the channel plane, and "out-of-plane" means in the direction of the normal to the channel plane.

	NMOS transistor	PMOS transistor
π_L	-49	+90
π_T	-16	-46
π_{out}	+87	-44

However, for short channel transistors, the experimental V_{Dsat} is smaller than that predicted by equation (3) (Taur et al., 1993). Using the concept of velocity saturation, (Suzuki & Usuki, 2004) proposed an equation for V_{Dsat} that can account for the disparity between the experimental V_{DS} and the V_{Dsat} that is predicted by equation (3).

$$V_{Dsat} = \frac{V_{GS} - V_{th,sat}}{0.5 + \sqrt{0.25 + \frac{\mu_{eff}\left(V_{GS} - V_{th,sat}\right)}{v_{sat}L_{eff}}}} \tag{6}$$

Since velocity overshoot occurs in the nanoscale transistor (Kim et al., 2008; Ruch, 1972), equation (6) needs to be modified. In the physics-based model for MOS transistors developed by (Hauser, 2005), v_{sat} is treated as a fitting parameter that can be increased to 2.06×10^7 cm/s so as to fit the experimental I_{ds} versus V_{DS} characteristics of the nanoscale NMOS transistor (L = 90 nm). Although this approach is conceptually wrong, it serves as an easy way to avoid detailed discussion in velocity overshoot and quasi-ballistic transport. Hence, the resulting equation is as follows,

$$V_{\mathrm{Dsat}} = \frac{V_{\mathrm{GS}} - V_{\mathrm{th,sat}}}{0.5 + \sqrt{0.25 + \frac{\mu_{\mathrm{eff}}(L_{\mathrm{eff}})}{v_{\mathrm{sat}}(L_{\mathrm{eff}})} \frac{V_{\mathrm{GS}} - V_{\mathrm{th,sat}}}{L_{\mathrm{eff}}}}} \tag{7}$$

where μ_{eff} and v_{sat} are functions of L_{eff}. To avoid confusion, we introduce another parameter called the effective saturation velocity ($v_{\mathrm{sat\_eff}}$). According to (Lau et al., 2008, b), $v_{\mathrm{sat\_eff}}$ is taken to be the average value of the carrier velocity (v_{eff}) when V_{GS} is close to the power supply voltage (V_{DD}). When uniaxial tensile stress is applied, both μ_{eff} and $v_{\mathrm{sat\_eff}}$ of NMOS transistor will be increased. By replacing $v_{\mathrm{sat}}(L_{\mathrm{eff}})$ in equation (7) by $v_{\mathrm{sat\_eff}}(\mu_{\mathrm{eff}}, L_{\mathrm{eff}})$,

$$V_{\mathrm{Dsat}} = \frac{V_{\mathrm{GS}} - V_{\mathrm{th,sat}}}{0.5 + \sqrt{0.25 + \frac{\mu_{\mathrm{eff}}(L_{\mathrm{eff}})}{v_{\mathrm{sat\_eff}}(\mu_{\mathrm{eff}}, L_{\mathrm{eff}})} \frac{V_{\mathrm{GS}} - V_{\mathrm{th,sat}}}{L_{\mathrm{eff}}}}} \tag{8}$$

For long channel MOS transistors, the large L_{eff} will make the third term in the denominator of equation (8) negligible and thus $V_{\mathrm{Dsat}} \approx (V_{\mathrm{GS}} - V_{\mathrm{th,sat}})$. For the short channel MOS transistors, the third term in the denominator of equation (8) must be considered and thus V_{Dsat} is expected to be smaller than $(V_{\mathrm{GS}} - V_{\mathrm{th,sat}})$. According to conventional MOS transistor theory (Taur & Ning, 1998, a), V_{Dsat} is given by $(V_{\mathrm{GS}} - V_{\mathrm{th,sat}})/m$ where the body effect coefficient (m) is typically between 1.1 and 1.4.

3. Does velocity saturation occur in the nanoscale MOS transistor?

For NMOS transistor, the electrons are accelerated by the lateral electric field (E_{lateral}) and thus the drift velocity (v_{drift}) increases. For (100) Si substrate, the optical phonon energy is bigger than 60 meV (Sah, 1991, a). When the kinetic energy of the electron exceeds 60 meV, the optical phonons are generated. However, the generation rate of optical phonon is very large and thus only a few electrons can have energy higher than 60 meV. An equilibrium is reached when the rate of energy gain from E_{lateral} is equal to the rate of energy loss to phonon scattering. This corresponds to the maximum v_{drift} that occurs at E_{lateral} around 10^4 V/cm. The maximum v_{drift} is known as the velocity saturation (v_{sat}). Based on the Monte Carlo simulation by (Ruch, 1972), the distance over which v_{drift} will overshoot the electron v_{sat} is less than 100 nm but this transient in velocity will only last for 0.8 ps before reaching its equilibrium value of 10^7 cm/s. According to (Mizuno, 2000), the amount of channel doping concentration (N_{ch}) will determine if velocity overshoot can be observed in bulk MOS transistors. For NMOS transistor with L = 80 nm, velocity overshoot can occur if $N_{\mathrm{ch}} < 10^{17}$ cm^{-3}. For NMOS transistor with L = 30 nm, velocity overshoot can occur even if $N_{\mathrm{ch}} \approx 10^{18}$ cm^{-3}. This can be attributed to the effective channel length (L_{eff}), which is a function of both the mask gate length (L) and N_{ch}. In fact, (Kim et al., 2008) has reported that the experimental findings of electron velocity overshoot in 36 nm bulk Si-based NMOS transistor at room temperature. Furthermore, the Monte Carlo simulation performed by (Miyata et al., 1993) show that electron velocity overshoot actually increases when the tensile stress is increased. This can account for the strain-induced I_{on} enhancement in the nanoscale NMOS transistors (Yang et al., 2004; C-H. Chen et al., 2004; Yang et al., 2008). Hence, it is more likely that velocity overshoot occur in the nanoscale transistor rather than velocity saturation.

Here, we will like to point out another misconception about the occurrence of velocity saturation in the nanoscale MOS transistors. Based on the classical concept of velocity saturation, the saturation I_{ds} of the short channel MOS transistor has a linear relationship with V_{GS} (see equation 5), and thus the saturation I_{ds} versus V_{DS} characteristics is expected to have constant spacing for equal V_{GS} step (Sze & Ng, 2007). On the other hand, the saturation I_{ds} of the long channel MOS transistor is controlled by pinchoff (Hofstein & Heiman, 1963). Based on the constant mobility assumption, equation 4 predicts that the saturation I_{ds} of long channel MOS transistor has a quadratic relationship with V_{GS} and thus the saturation I_{ds} versus V_{DS} characteristics is expected to have increasing spacing for equal V_{GS} step (Sze & Ng, 2007). However, constant spacing for equal V_{GS} step is often observed in the experimental I_{ds} versus V_{DS} characteristics of the long channel MOS transistor, as shown in Fig.3. This can be understood from the validity of the constant mobility assumption. Experimental data have shown that mobility is actually a function of V_{GS} (Takagi et al., 1994). From Fig.7, μ_{eff} first increases with increasing V_{GS} owing to Coulombic scattering and then decreases owing to phonon scattering and surface roughness scattering. To further investigate, we measured the I_{ds} versus V_{DS} characteristics and the I_{ds} versus V_{GS} characteristics of a long-channel NMOS transistor. Considering equal V_{GS} step, we observed an increasing spacing for 1 V≤ V_{GS} ≤ 3 V but constant spacing for 3 V ≤ V_{GS} ≤ 5V in the saturation I_{ds} versus V_{DS} characteristics of the NMOS transistor (see Fig.8). Since the transconductance (g_m) is a measure of the low-field mobility (μ_{eff}) (Schroder, 1998), the g_m versus V_{GS} characteristics is expected to have the same features as the mobility versus V_{GS} characteristics. From Fig. 8(a), the drain current saturation of the NMOS transistor occurs at V_{DS} around 3 V. With reference to Fig. 8(b), when V_{DS} = 3 V and 0 V ≤ V_{GS} ≤ 3 V, g_m increases monotonically with increasing V_{GS} owing to Coulombic scattering. When V_{GS} is further increased to beyond 3 V, surface roughness scattering will start to dominate and then g_m will decrease with increasing V_{GS}. Hence, for 1 V ≤ V_{GS} ≤ 3 V, the saturation I_{ds} versus V_{DS} characteristics has increasing spacing for equal V_{GS} step. For 3 V ≤ V_{GS} ≤ 5 V, the saturation I_{ds} versus V_{DS} characteristics has constant spacing for equal V_{GS} step. Since velocity saturation does not occur in long channel transistor, the constant spacing observed in the saturation I_{ds} versus V_{DS} characteristics at high V_{GS} cannot be used as an indicator of the onset of velocity saturation.

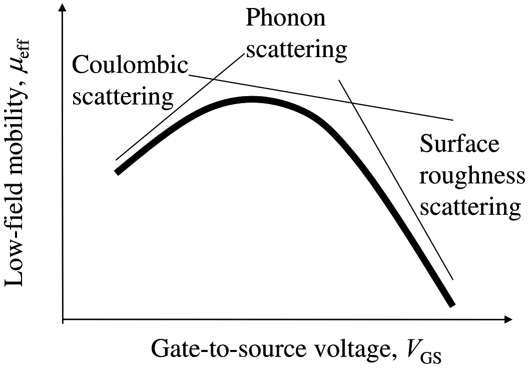


Fig. 7. Effects of the scattering mechanisms on the μ_{eff} versus V_{GS} characteristics of MOS transistor.

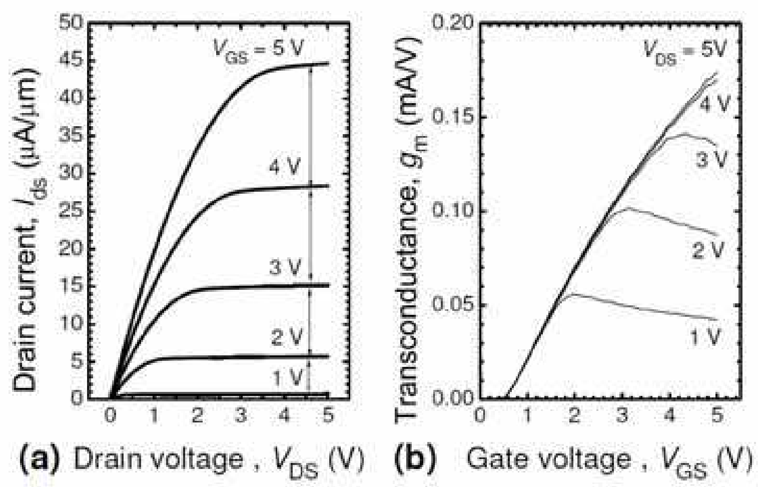


Fig. 8. Constant spacing is observed in the saturation I_{ds} versus V_{DS} characteristics of a NMOS transistor (L = 10 μm, W = 10 μm, physical gate oxide thickness of 300 Å) for equal V_{GS} step.

Here, it is interesting to note that it is common for the saturation I_{ds} versus V_{DS} characteristics of the zinc oxide thin-film transistors to have increasing spacing for equal V_{GS} step (Cheong et al., 2009; Yaglioglu et al., 2005). The mobility of these materials (~ 10 to 20 cm^2/V.s) is only one tenth of the mobility of silicon (~ 100 to 300 cm^2/Vs). In Fig.9, which is modified from (Cheong et al., 2009), the drain current saturation occurs at V_{DS} around 15 V. The increasing spacing observed in the saturation I_{ds} versus V_{DS} characteristics of the thin-

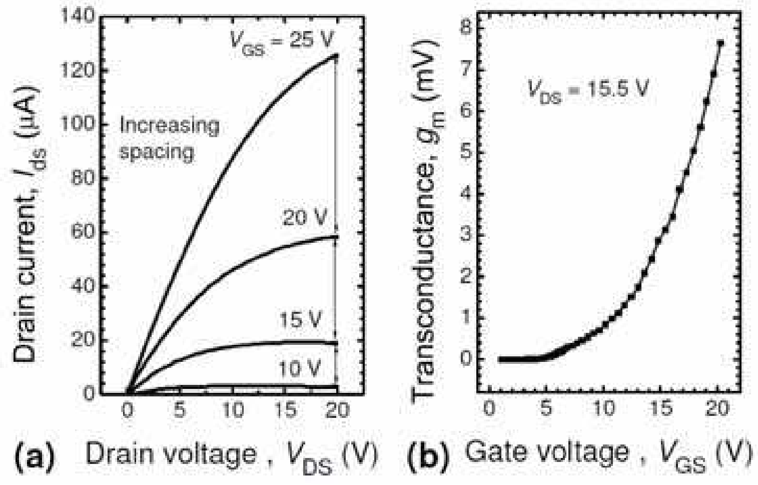


Fig. 9. Zinc oxide thin-film transistors with L = 20 μm and W = 40 μm (a) Increasing spacing observed in the experimental I_{ds} versus V_{DS} characteristics of, (b) Monotonically increasing g_m. Modified from (Cheong et al., 2009).

film transistor is related to the monotonically increasing g_m with increasing V_{GS}. Next, we will study the dependency of the saturation I_{ds} of the thin film transistor on V_{GS}. From Fig. 10, if I_{ds} and V_{GS} have linear dependency, $V_{th,sat}$ extracted by linear interpolation is around 17.5 V. If I_{ds} and V_{GS} have quadratic dependency, $V_{th,sat}$ extracted by extrapolating the linear portion of the $I_{ds}^{0.5}$ versus V_{GS} plot is around 10 V. As seen in the I_{ds} versus V_{DS} characteristics of the thin-film transistor (see Fig.9), the transistor is in cutoff mode when V_{GS} ≤ 10 V. Hence, it is more appropriate to say that I_{ds} of thin-film transistor and V_{GS} have quadratic dependency rather than linear dependency.

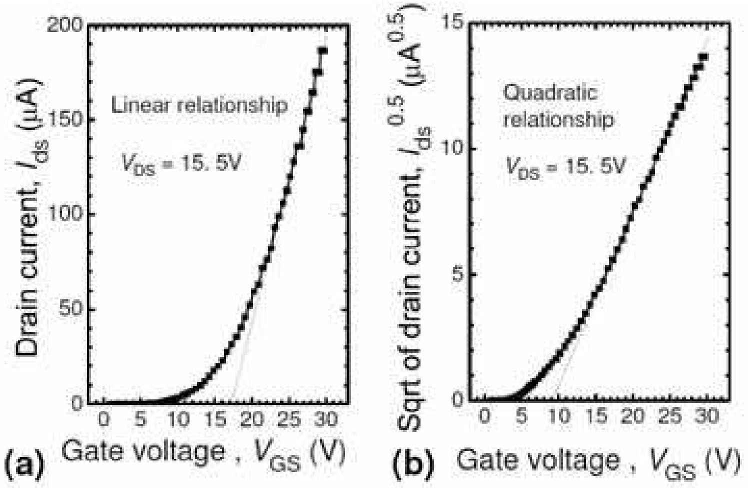


Fig. 10. Relationship between I_{ds} and V_{GS} of the zinc oxide thin-film transistors (L = 20 μm and W = 40 μm) (a) Linear dependency (b) Quadratic dependency. Modified from (Cheong et al., 2009).

4. Newer theories on the saturation drain current equations of the nanoscale MOS transistor

According to (Natori, 2008), the type of carrier transport in the MOS transistor depends on the relative dimension between the gate length (L) and the mean free path (λ), as illustrated in Fig. 11. Qualitatively, λ is the average distance covered by the channel carrier between the successive collisions. When L is much bigger than λ, the channel carriers will experience diffusive transport. When L is comparable to λ, the carriers undergo only a small number of scattering events from the source to the drain and thus the carriers will experience quasi-ballistic transport. Ballistic transport will only occur when $L < \lambda$. The experimentally extracted λ is in the range of 10 nm for the nanoscale transistor (M-J. Chen et al., 2004; Barral et al., 2009). Hence, the state-of-the-art MOS transistor ($L \geq 32$ nm) is more likely to experience quasi-ballistic transport rather than ballistic transport. This section will discuss the main concepts of ballistic transport and then proceed to discuss about the existing quasi-ballistic theories. The emphasis of this section is to introduce a simplified equation for the saturation drain current of the nanoscale MOS transistor that is able to address quasi-ballistic transport while having electrical parameters that are obtainable from the standard device measurements. Here, we will introduce two equations that can satisfy the above criteria (i) Based on the concept of the effective saturation velocity ($v_{\mathrm{sat\_eff}}$) , which is a function of μ_{eff} and temperature (Lau et al. , 2008, b) and (ii) Based on the virtual source model (Khakifirooz et al., 2009).

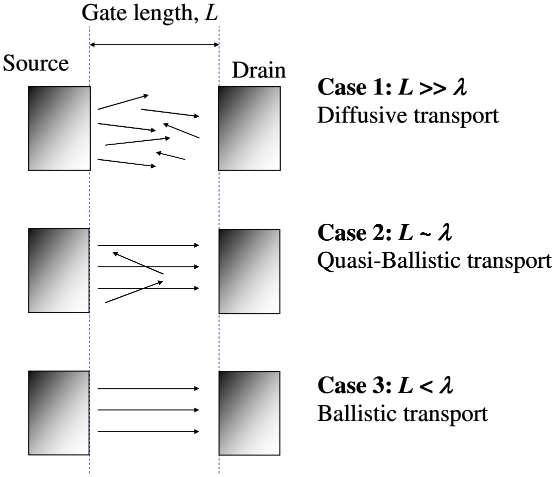

Gate length, L
Source
Drain
Case 1: $L >> \lambda$
Diffusive transport
Case 2: $L \sim \lambda$
Quasi-Ballistic transport
Case 3: $L < \lambda$
Ballistic transport

Fig. 11. Types of carrier transport in MOS transistors, which is modified from Fig. 1 in (Natori, 2008). Note that λ is the mean free path of the carrier.

4.1 Ballistic transport

In vacuum, electrons will move under the influence of electric field according to Newton's second law of motion,

$$F = m_e a = -qE \tag{9}$$

where F, m_e, a, q and E are the resultant force acting on the electron, the electron mass, the acceleration of the electron, the electronic charge , and the electric field ,respectively. Under such a situation, if the applied electric field is constant in both magnitude and direction, the electrons will accelerate in the direction opposite to that of the electric field. This type of transport is known as the ballistic transport. In the other words, if there is no obstacle to scatter the electrons, the electrons will experience ballistic transport (Heiblum & Eastman, 1987). Furthermore, (Bloch, 1928) postulated that the wave-particle duality of electron allows it to move without scattering in the densely packed atoms of a crystalline solid if (i) the crystal lattice is perfect and (ii) there is no lattice vibration. However, doping impurities such as boron, arsenic and phosphorus are added to the silicon crystal so as to tune the electrical parameters such as the threshold voltage and the off-state current (I_{off}). These dopants will disrupt the periodic arrangement of the crystal lattice and thus results in collisions with the impurity ions and the crystalline defects. Moreover, the atoms in crystals are always in constant motion according to the Particle Theory of Matter. These thermal vibrations cause waves of compression and expansion to move through the crystal and thus scatter the electrons (Heiblum & Eastman, 1987). Therefore, achieving ballistic transport in Si-based MOS transistors is only an ideal situation (Natori, 2008).

4.2 Quasi-ballistic transport

Having established that thermionic emission from the source to the channel is still relevant in the state-of-the-art MOS transistor ($L \geq 32$ nm) in Section 1, we will proceed to discuss the main concepts behind quasi-ballistic transport. (Lundstrom, 1997) derived an equation that relates the saturation I_{ds} of the nanoscale transistor to μ_{eff} as follows,

$$I_{\mathrm{ds}} = \left[\frac{C_{\mathrm{ox}}\, W}{\frac{1}{v_{\mathrm{T}}} + \frac{1}{\mu_{\mathrm{eff}}\ \varepsilon(0^{+})}} \right] \left(V_{\mathrm{GS}} - V_{\mathrm{th,sat}} \right) \tag{10}$$

where the random thermal velocity of the carriers (v_{T}) does not depend V_{GS}. The only variable in the v_{T} equation is the temperature (T).

$$v_{\mathrm{T}} = v_{\mathrm{T}}(T) = \sqrt{(2k_B T)/(\pi\ m_t)} \tag{11}$$

where the transverse electron mass of silicon (m_{t}) is equal to 0.19 m_0 where the free electron mass (m_0) is equal to 9.11×10^{-31} kg (Singh, 1993). Using equation (11), v_{T} is approximately equal to 1.2×10^{7} cm/s at temperature of 25 °C. k_{B} is the Boltzmann constant. T is the absolute temperature. $\varepsilon(0^{+})$ is defined as the average electric field within the length ℓ where a $k_{\mathrm{B}}T/q$ potential drop occurs, as shown in Fig.12 in (Lundstrom & Ren, 2002). Despite the lack of equation for $\varepsilon(0^{+})$ (Lundstrom, 1997; Lundstrom & Ren, 2002), Lundstrom has made an important contribution to relate the low-field mobility (μ_{eff}) to I_{on} of the deep submicron MOS transistors, and thus his theory is able to account for the strain-induced enhancement in I_{on} (Yang et al., 2004; C-H. Chen et al., 2004; Yang et al. 2008; Wang et al., 2007).

According to (Lundstrom, 1997), if a carrier backscatters beyond ℓ, it is likely to exit from the drain and is unlikely to return back to the source (see Fig. 12). For NMOS transistor, ℓ is the distance between the top of the conduction band edge and the point along the channel where channel potential drops by $k_{\mathrm{B}}T/q$.

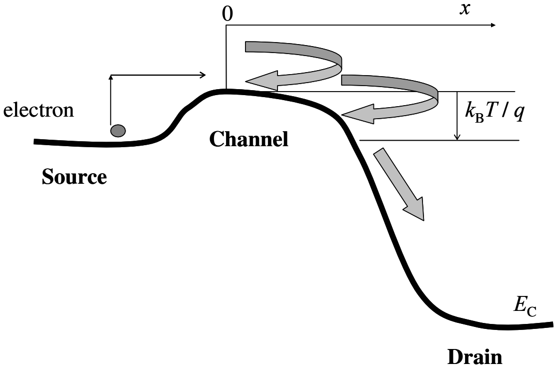


Fig. 12. Definition of the critical length (ℓ) for NMOS transistor. ℓ is defined to be the distance between the top of the conduction band edge and the point along the channel where channel potential drops by $k_{\mathrm{B}}T/q$. Beyond ℓ, the carriers are unlikely to return to the source.

By inspection of equations (10) and (11), a loop-hole can be found in Lundstrom's 1997 theory. If equations (10) and (11) are correct, MOS transistors will function very poorly when the temperature is lowered from room temperature to very low temperature such as liquid helium temperature. However, there are numerous reports that MOS transistors and CMOS integrated circuits can function quite well at the liquid helium temperature (Chou et al., 1985; Ghibaudo & Balestra, 1997; Yoshikawa et al., 2005). Hence, there is a need to modify Lundstrom's 1997 theory. Indeed, (Lundstrom & Ren, 2002) made an attempt to incorporate Natori's 1994 theory into their theory. However, the resulting theory is very much not similar to equation (10) and has not been compared with real device performance. Based on equation (24) in (Natori, 1994), the saturation drain current of the nanoscale MOS transistor is as follows,

$$I_{\mathrm{ds}} = \frac{8\hbar W\left[C_{\mathrm{ox}}\left(V_{\mathrm{GS}}\text{-}V_{\mathrm{th,sat}}\right)\right]^{3/2}}{3m_{\mathrm{t}}\sqrt{q\pi M_{\mathrm{v}}}} \tag{12a}$$

where $\hbar$ is the reduced Planck's constant. M_{v} is the product of the lowest valley degeneracy and the reciprocal of the fraction of the carrier population in the lowest energy level. For a NMOS transistor that is fabricated on (100) Si substrate, the fraction of the carrier population at the strong inversion is around 0.8 at 77 K but it decreases to around 0.4 at 300 K (Stern, 1972). In the other words, M_{v} is a function of temperature (T).
Rearranging equation (12a) results in,

$$I_{\mathrm{ds}} = \left[\frac{C_{\mathrm{ox}}\ W}{\dfrac{1}{v_{\mathrm{inj}}(V_{\mathrm{GS}},T)}}\right]\left(V_{\mathrm{GS}} - V_{\mathrm{th,sat}}\right) \tag{12b}$$

where the injection velocity (v_{inj}) is given by (Natori, 1994),

$$v_{\mathrm{inj}}(V_{\mathrm{GS}},T) = \frac{8\hbar\sqrt{C_{\mathrm{ox}}\left(V_{\mathrm{GS}}\text{-}V_{\mathrm{th,sat}}\right)}}{3m_{\mathrm{t}}\sqrt{q\pi M_{\mathrm{v}}(T)}} \tag{12c}$$

With reference to Fig.8 in (Natori, 1994), v_{inj} increases with increasing temperature (T) and increasing V_{GS}. If Natori's theory is true, v_{inj} can be very high even though the temperature is very low. We propose that this feature of Natori's 1994 theory can be used to cover the shortcomings of Lundstrom's 1997 theory. However, there are some aspects of Natori's 1994 theory that contradict the experimental data. From Fig. 8 in (Natori, 1994), his theory, which disregards the channel scattering, predicted that the saturation I_{ds} of the nanoscale NMOS transistor will increase when temperature increases. However, this is contradictory to the experimental data. Fig. 13 shows that the experimental I_{ds} of a NMOS transistor (L= 60 nm) actually decreases when temperature increases. This can be explained by the increase in channel scattering when temperature increases (Takagi et al., 1994; Kondo & Tanimoto, 2001; Mazzoni et al., 1999). Moreover, equation (12b) cannot account for the strain-induced enhancement in I_{on} (Yang et al, 2004; C-H. Chen et al, 2004; Yang et al., 2008; Wang et al., 2007). Hence, without the help of Lundstrom's 1997 theory, Natori's 1994 theory is contradictory to the experimental data.

In addition, Natori's 1994 theory predicts that the saturation I_{ds} of the nanoscale MOS transistors will follow a $(V_{GS} - V_{th,sat})^{3/2}$ relationship. Fig. 14a shows the saturation $I_{ds}^{2/3}$ versus V_{GS} characteristics of a NMOS transistor (L = 60 nm). The threshold voltage extracted by the linear extrapolation is smaller than the threshold voltage of conduction. This shows that the saturation I_{ds} of the nanoscale MOS transistors does not follow a $(V_{GS} - V_{th,sat})^{3/2}$ relationship. Fig. 14b shows the saturation I_{ds} versus V_{GS} characteristics of the same NMOS transistor. In this case, the extracted threshold voltage is close to the threshold voltage of conduction. Hence, the saturation I_{ds} of nanoscale transistors is more likely to follow a $(V_{GS} - V_{th,sat})$ relationship.

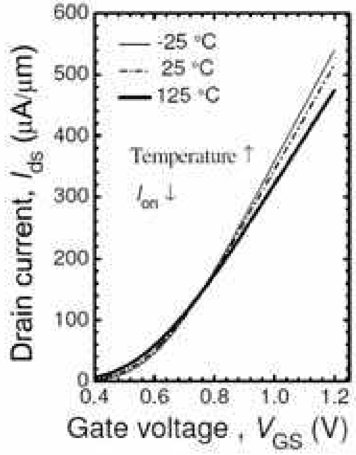


Fig. 13. Effects of temperature on the saturation I_{ds} versus V_{GS} characteristics of a NMOS transistor (L = 60 nm, W = 5 μm).

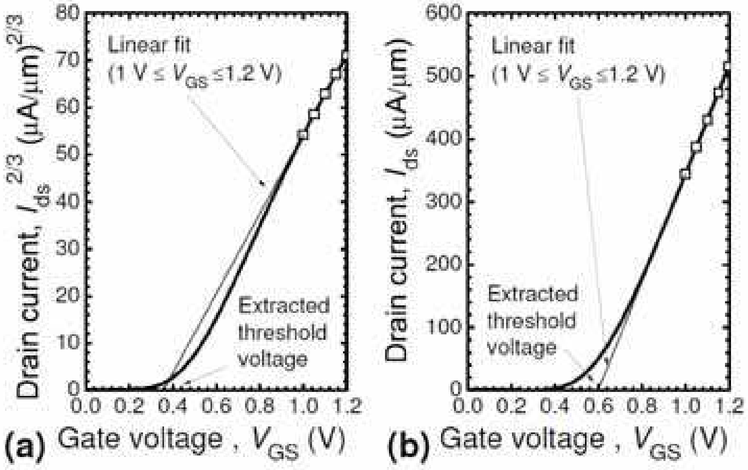


Fig. 14. As opposed to Natori's 1994 theory, the saturation I_{ds} of the short channel NMOS transistor does not follow a $(V_{GS} - V_{th,sat})^{3/2}$ relationship.

4.3 New equation that unifies Natori's 1994 theory and Lundstrom's 1997 theory

We propose a simplified equation that can unify both Natori's 1994 theory and Lundstrom's 1997 theory, as follows (Lau et al., 2008, b),

$$I_{\mathrm{ds}} = \left[\frac{C_{\mathrm{ox}} W}{\dfrac{1}{v_1(V_{\mathrm{GS}},T)} + \dfrac{1}{v_1(V_{\mathrm{GS}},T)}} \right] \left(V_{\mathrm{GS}} - V_{\mathrm{th,sat}} \right) \tag{13}$$

where

$$v_1(\ V_{\mathrm{GS}},\ T) = v_{\mathrm{inj}}(\ V_{\mathrm{GS}},\ T) \tag{14}$$

$$v_2(\ V_{\mathrm{GS}},\ T) = \mu_{\mathrm{eff}}(\ V_{\mathrm{GS}},\ T)\ \varepsilon\left(0^{+}\right) \tag{15}$$

(Lundstrom, 1997) proposed that v_1 is equal to v_T that is only dependent on T, as shown in equation (11). On the other hand, our theory proposed that v_1 is a function of both V_{GS} and T, and v_1 can be higher than v_T given by equation (11). (Natori, 1994) proposed that v_1 is equal to v_{inj}, which is a function of both V_{GS} and T. Recently, (Natori et al., 2003; Natori et al., 2005) simulated the v_{inj} characteristics using the multi-subband model (MSM). In weak inversion, v_{inj} is almost independent of V_{GS} and is approximately equal to 1.2 x 10^7 cm/s, which is equal to v_T. In strong inversion, v_{inj} will increase due to carrier degeneration but is confined within a narrow range from 1.2 x 10^7 cm/s to 1.6 x 10^7 cm/s.

Here, we would like to highlight that both Lundstrom's 1997 theory and Natori's 1994 theory did not consider the series resistance (R_{sd}). Although the conduction band edge (E_c) profile in the n-channel will be the same with or without R_{sd} (Martinie et al., 2008), the E_c within S/D regions will be different when the effects of R_{sd} is considered. If the effects of R_{sd} are disregarded, E_c within S/D regions will appear as a horizontal line, as illustrated in Fig. 12. However, the presence of R_{sd} will cause a potential drop in the S/D regions, resulting in a built-in electric field within the S/D regions (see Fig. 15). This electric field in the source region will accelerate the electrons. Since scattering decreases when temperature decreases (Takagi et al., 1994; Kondo & Tanimoto, 2001; Mazzoni et al., 1999), one would expect that there will be minimal scattering in the source when the temperature is very low. Hence, the presence of R_{sd} will allow the electrons to attain higher energy prior to thermionic emission into the channel. According to (M-J. Chen et al., 2004), the source series resistance (R_s) is about 75 Ω-µm. If the drain current (I_{ds}) is about 800 µA/µm, the voltage drop due to R_s is about 800 µA/µm x 75 Ω-µm = 60 mV. (Note that the thermal voltage, $k_B T/q$ is approximately 26 meV at room temperature.) We proposed that the electrons are "heated" up by the 60 meV energy due to R_{sd} and thus their velocities can be significantly larger than 1.2 x 10^7 cm/s (as predicted by equation 12c). Moreover, this extra energy is expected to increase with increasing V_{GS} because higher V_{GS} implies a bigger I_{ds}. With this extra energy from electron heating in the R_{sd} region, the carriers can overcome the potential barrier at the liquid nitrogen temperature despite not being able to gain energy from the surrounding. The significance of v_2 term is that it establishes a link between I_{on} and μ_{eff}. This provides a better compatibility between theory and I_{on} enhancement in the nanoscale transistors by various stress engineering techniques (Yang et al., 2004; C-H. Chen et al., 2004; Yang et al., 2008; Wang et al., 2007). However, there is no v_2 term in Natori's 1994 theory, as shown in equation (12b). Nevertheless, v_2 is covered by Lundstrom's 1997 theory, as shown in equation (10). Hence, we incorporate v_2 in Lundstrom's 1997 theory into equation (13).

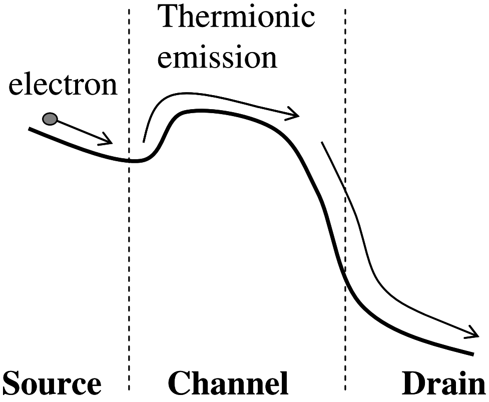


Fig. 15. The effects of S/D series resistance on the conduction band edge of a NMOS transistor in the saturation operation.

Another loop-hole in Lundstrom's 1997 theory is that there is no equation for $\varepsilon(0^+)$. From Fig.9 in (M-J. Chen et al., 2004), the slope of the near-source channel conduction band increases when V_{GS} increases. In the other words, the electric field near the top of potential barrier, $\varepsilon(0^+)$ increases with increasing V_{GS}. Hence, we deduce that $\varepsilon(0^+)$ is a function of both V_{GS} and V_{DS} such that $\varepsilon(0^+, V_{GS}, V_{DS} = V_{DD})$ is approximately equal to $\varepsilon(0^+, V_{GS}, V_{DS} = V_{Dsat})$. Note that V_{DD} is the power supply voltage. This is consistent with Fig. 5 in (Fuchs et al., 2005). Therefore, we propose that $\varepsilon(0^+)$ can be expressed as follows,

$$\varepsilon(0^+) = \frac{\alpha_1 V_{\text{Dsat}}}{L_{\text{eff}}} \tag{16a}$$

where the correction factor (α_1) is smaller than 1. Based on the conventional MOS transistor theory (Taur & Ning, 1998, a), V_{Dsat} is given by (V_{GS} – $V_{th,sat}$)/m where 1.1 ≤ m ≤ 1.4. Furthermore, (Suzuki & Usuki, 2004) proposed a drain current model that shows that V_{Dsat} is smaller than (V_{GS} – $V_{th,sat}$) for the short-channel MOS transistors. This shows that the relationship of V_{Dsat} = (V_{GS} – $V_{th,sat}$)/m is still reasonably correct for very short MOS transistors. Therefore, $\varepsilon(0^+)$ can also be expressed by,.

$$\varepsilon(0^+) = \frac{\alpha_2 \left(V_{\text{GS}} - V_{\text{th,sat}}\right)}{L_{\text{eff}}} \tag{16b}$$

where the correction factor (α_2) is smaller than 1. The value of α_2 can be estimated from the effective carrier velocity (v_{eff}) versus V_{GS} characteristics and the μ_{eff} versus V_{GS} characteristics. Using the saturated transconductance method suggested by (Lochtefeld et al., 2002), v_{eff} was extracted as a function of V_{GS} as shown in Fig.16 (a). For the contact etch stop layer (CESL) with a tensile stress of 1.2 GPa, v_{sat_eff} of the NMOS transistor (L = 60 nm) was 7.3 × 10^6 cm/s. Using the constant current method with reference current, I_{ref}