I
Logic Devices and Concepts

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser
Copyright © 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-31737-0

1
Non-Conventional Complementary Metal-Oxide-Semiconductor (CMOS) Devices

Lothar Risch

1.1
Nano-Size CMOS and Challenges

The scaling of complementary metal-oxide-semiconductor (CMOS) is key to following Moore's law for higher integration densities, faster switching times, and reduced power consumption at reduced costs. In today's research laboratories MOSFETs with minimum gate lengths below 15 nm have already been demonstrated. An example of such a small transistor is shown in Figure 1.1a, where the transmission electron microscopy (TEM) cross-section shows a functional, fully depleted silicon-on-insulator (SOI) transistor with 14 nm gate length, 20 nm spacers using a 17 nm thin silicon layer and a 1.5-nm gate dielectric. The gate has been defined with electron-beam (e-beam) lithography. For the contacts, elevated source drain regions were grown with selective Si epitaxy to lower the parasitic resistance, and a high dose of dopants was implanted into the epi layer for source and drain. In Figure 1.1b, a TEM cross-section through the fin of a SONOS memory FinFET is shown with a diameter of 8 nm, surrounded by the ONO charge-trapping dielectric. As can be seen, many critical features in Si-MOSFETs are already in the range in the range of 1 to 20 nm.

However, achieving the desired performance gain in electrical parameters from scaling will in time become very challenging, as indicated in the International Technology Roadmap for Semiconductors (ITRS) by many red brick walls [1] (see Figure 1.2).

The three main limiting factors for a performance increase are related to physical laws. Gate leakage stops SiO_2 scaling (see Figure 1.3), while source drain leakage reduction needs higher channel doping and shallower junctions. However, this increases junction capacitance, junction leakage, gate-induced drain currents, reduces carrier mobility and increases parasitic resistance. Because of this, transistors with astoundingly small gate lengths down to 5 nm have been realized [2]; although these are the smallest MOSFETs produced until now, their performance is worse than that of a 20-nm device.

When considering memories, the situation is not much different, and for mainstream DRAM and Floating Gate Flash several constraints can be foreseen. For DRAM, the storage capacitance at small cell size and a low leakage cell transistor

Nanotechnology. Volume 4: Information Technology II. Edited by Rainer Waser
Copyright © 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 978-3-527-31737-0

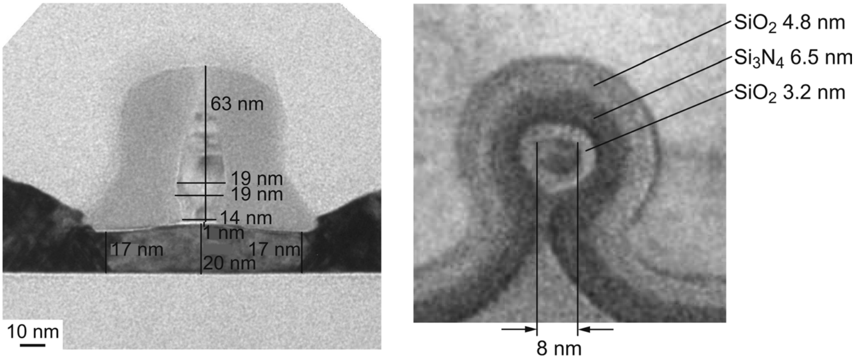
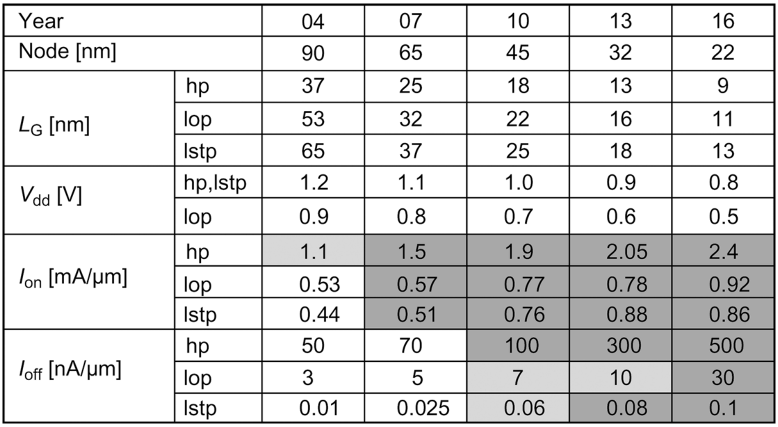


Figure 1.1 (a) A TEM cross-section of a 14-nm gate SOI transistor with raised source/drain (S/D) on 17 nm Si, $t_{ox} = 1.5$ nm. (b) TEM cross-section of a SONOS SOI FinFET across a 8-nm wire-type fin.

become a critical issue. For Floating Gate, the high drain voltages and scaling of the gate dielectric, as well as coupling to neighboring cells, are critical.

Therefore, on the way to better devices, two strategies are proposed by ITRS. The first strategy is to implement new materials as performance boosters. Among these are high-*k* dielectrics and metal gates, high-mobility channels and low-resistivity or



Year		04	07	10	13	16
Node [nm]		90	65	45	32	22
L_G [nm]	hp	37	25	18	13	9
	lop	53	32	22	16	11
	lstp	65	37	25	18	13
V_{dd} [V]	hp,lstp	1.2	1.1	1.0	0.9	0.8
	lop	0.9	0.8	0.7	0.6	0.5
I_{on} [mA/µm]	hp	1.1	1.5	1.9	2.05	2.4
	lop	0.53	0.57	0.77	0.78	0.92
	lstp	0.44	0.51	0.76	0.88	0.86
I_{off} [nA/µm]	hp	50	70	100	300	500
	lop	3	5	7	10	30
	lstp	0.01	0.025	0.06	0.08	0.1

Figure 1.2 ITRS 04 roadmap: gate lengths and currents for high performance, low operation power, low standby power.

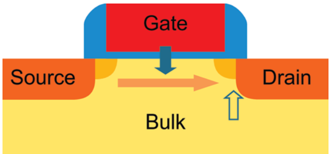


Figure 1.3 Scaling limits of scaled MOSFETs: source to drain, gate dielectric tunneling and junction leakage.

metal source drain junctions. This will lead to a remarkable improvement in the performance of transistors. The second strategy is to develop new device structures with better electrostatic control, such as fully depleted SOI and multi-gate devices. These can also be utilized in DRAMs as low leakage cell transistors, as well as in nanoscale non-volatile Flash memories.

1.2 Mobility Enhancement: SiGe, Strained Layers, Crystal Orientation

Carrier mobility enhancement of electron and holes provide the key to increase the on-currents without higher gate capacitance and without degrading the off-currents. Several methods have been developed, including SiGe heterostructures [3] with a higher hole mobility for the p-channel transistor. This is achieved by growing a thin epitaxial $Si_{1-x}Ge_x$ layer, where x is the Ge concentration, with a thickness of 5–10 nm for the channel region directly on Si (see Figure 1.4). On top of the SiGe layer a thin Si cap layer is deposited with a thickness of 3–5 nm, which is also used for the growth of the gate oxide. This forms a quantum well for the holes due to a step in the valence band of the Si/SiGe/Si heterostructure, with a depth of about 150 mV for a Ge content of 20%. The SiGe layer is under bi-axial compressive strain due to the smaller lattice constant of Si compared to SiGe (see Figure 1.4a). The mobility is enhanced because of the lower effective mass of the holes in SiGe and a split of the degenerated three-valence bands, thus reducing intervalley scattering. Compared to pure Si with a peak hole mobility of about 110 $cm^2 Vs^{-1}$, with 0.25 Ge 210 $m^2 Vs^{-1}$ have been achieved [4],

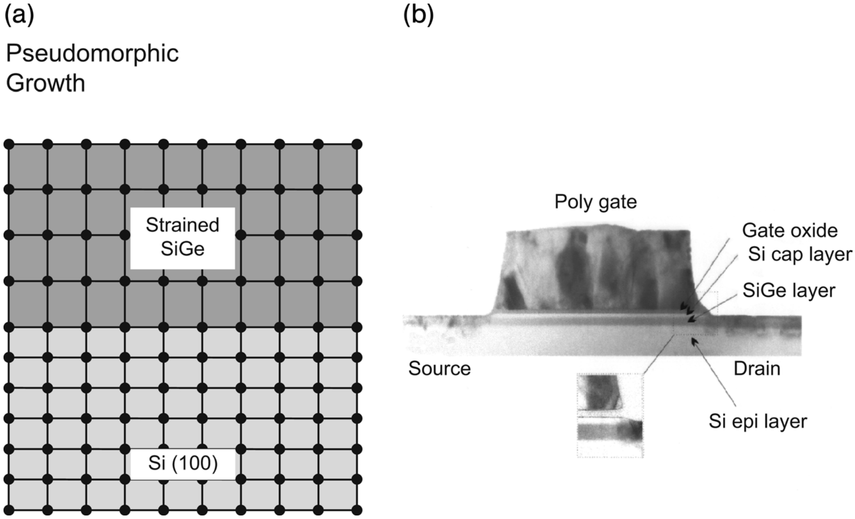


Figure 1.4 (a) Crystal lattice of a Si/SiGe heterostructure. (b) TEM cross-section of a p-channel MOSFET with a Si/SiGe/S quantum well.

extracted from MOSFET measurements. Whereas, the SiGe channel on Si is beneficial for the hole mobility, strained silicon offers both an improved electron and hole mobility, together with a surface channel [5]. The strain is created by a relaxed graded SiGe buffer layer, typically with a thickness of about 3 μm and a Ge concentration of 20–30%. A thin Si layer is grown on top of the relaxed SiGe layer in the range of 10 to 20 nm, which is now under biaxial tensile strain due to the larger lattice constant of the SiGe buffer layer.

Both techniques provide global bi-axial strain on the wafer and are based on Si/SiGe epitaxy. A critical issue here is the increased process complexity, the density of defects and wafer cost. Moreover, the implementation of tensile strain for the n-channel and compressive strain for the p-channel would be desirable, and is difficult to achieve with global strain. Therefore, local uni-axial strain techniques have now become mainstream for mobility enhancement, and these can provide tensile and compressive strain by depositing dedicated layers around the transistor. This method was introduced [6] for the 90-nm CMOS generation. In the n-channel transistor a nitride capping layer with tensile strain is used to improve the drive current by 10–15%. For the p-channel transistor, an embedded SiGe source drain region provides compressive strain and increases the drive current by 25%. TEM cross-sections of the n- and p-channel devices are shown in Figure 1.5 [6].

Another mobility-enhancement technique is based on the crystal orientation dependence of the mobility. Until now, the (1 0 0) surface of silicon wafers has been used with a channel orientation of the transistors in the <0 1 1> direction (see Figure 1.6). This is optimal for the electron mobility but will decrease the hole mobility, which is twice that at the (1 1 0) plane in the <1 0 0> direction. If (1 1 0) wafers are used or rotated (1 0 0) wafers by 45° with the channel in the <1 0 0> direction, the hole mobility is improved remarkably while electron mobility is reduced only moderately (see Figure 1.7) [7].

Therefore, another channel orientation is an effective means to increase p-channel performance, and an improvement of up to 15% has been reported [8]. Unfortunately, the embedded SiGe source drain regions with compressive strain have no remarkable influence in this crystal direction.

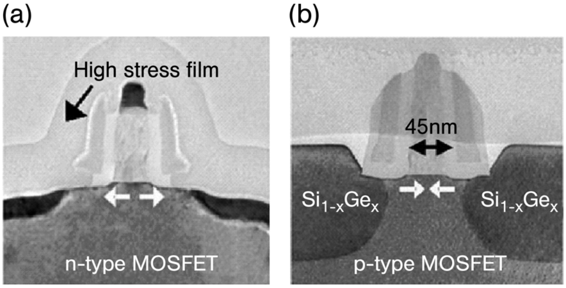


Figure 1.5 (a) 90-nm technology NMOS transistor with tensile stress nitride layer; (b) PMOS, showing heteroepitaxial SiGe source/drain inducing compressive strain [6] .

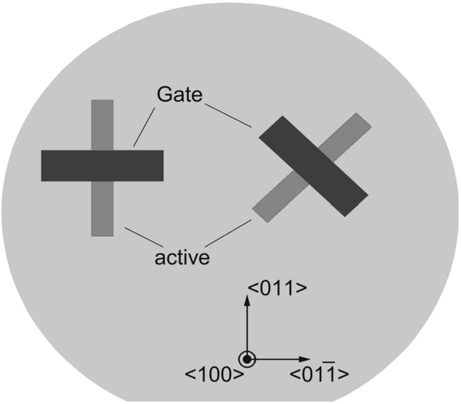


Figure 1.6 Crystal orientation and channel direction on (1 0 0) Si wafers.

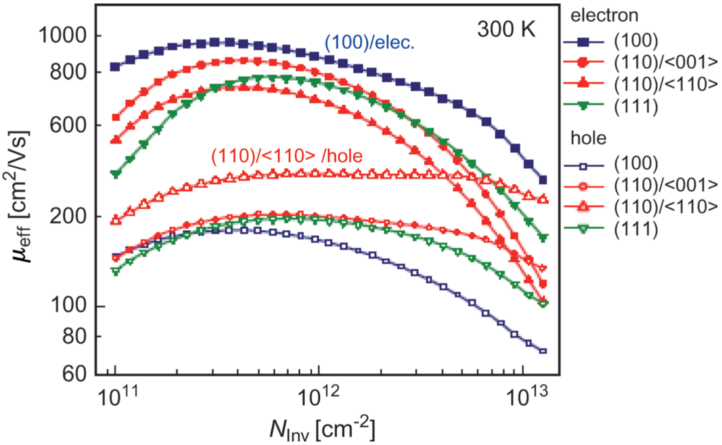


Figure 1.7 Mobility dependence for electrons and holes on crystal orientation and channel direction [7].

1.3 High-*k* Gate Dielectrics and Metal Gate

As indicated in the ITRS roadmap, scaling of the classical SiO_2 gate dielectric and increasing the gate capacitance in order to achieve higher drive currents reached completion at about 2 nm for low standby power, due to Fowler–Nordheim tunneling currents through the gate dielectric. By using nitrided oxides, the minimum thickness could be extended to about 1 nm for high-performance applications with a gate leakage current of about 10^3 A cm^{-2} [9]. The introduction of high-*k* dielectrics allows the use of thicker dielectric layers in order to reduce the tunneling currents at the same equivalent oxide thickness, or to provide thinner dielectrics for continuous scaling. Unfortunately,

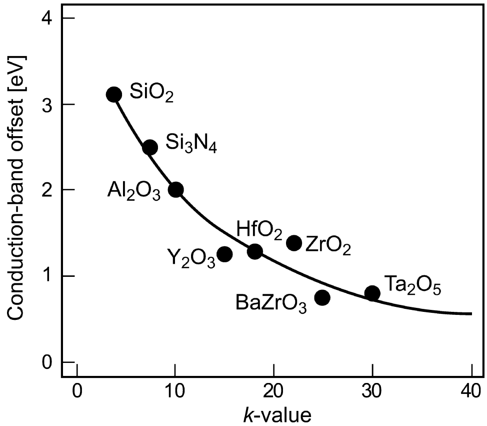


Figure 1.8 Conduction band offset versus k-value for different high-*k* materials [10].

all known high-*k* materials have a smaller bandgap than SiO_2. In Figure 1.8 the conduction band offset as a function of the dielectric constant is shown for different materials [10]. For the highest *k* materials such as Ta_2O_5 ($k = 30$) or TiO_2 ($k = 90$), the bandgap becomes too small and leads to increased gate leakage. Other critical issues are the growth of an interfacial layer during processing. Today, the most mature high-*k* dielectrics are based on Hf. Among these, HfO_2 ($k = 17$–25), HfSiO ($k = 11$) and HfSiON ($k = 9$–11), the latter are the more temperature-stable. An equivalent oxide thickness of below 1 nm has been demonstrated for these high-*k* materials [10]. Other candidates are ZrO_2 and La_2O_3 with dielectric constants between 20 and 30; however, the former is incompatible with a poly silicon gate and requires a metal gate.

For most high-*k* dielectrics a degradation of mobility is observed due to an increased scattering by phonons or a high fixed charge density at the interface. Especially for Al_2O_3, the hole mobility reduction is not acceptable. For the best Hf-based high-*k* dielectrics a 20% lower mobility has been achieved until now, compared to SiO_2.

Closely related to the high-*k* dielectric is a new gate material which avoids the depletion layer of poly silicon gates and the reaction of the high-*k* material with silicon at higher process temperatures. Moreover, metal gates offer the possibility of adjusting the threshold voltage with the workfunction of the gate material instead of doping in the channel, and this decreases the mobility at higher doping concentrations. The desired workfunctions for bulk with n+ poly and p+ poly silicon gates for low-power/high-performance applications with low doped transistor channels are shown in Figure 1.9.

Midgap-like materials such as TiN, TiSiN and W are suitable for n- and p-channel transistors with threshold voltages in the range of 300 to 400 mV, especially for fully depleted SOI or multi-gate transistors with lower channel doping concentrations. For optimized logic processes with low V_t transistors for high performance, in the range of 100 to 200 mV, dual metals with n+ and p+ poly-silicon-like workfunctions must be integrated. For n-channel transistors Ru is a candidate, and for p-channel Ta or RuTa alloys.

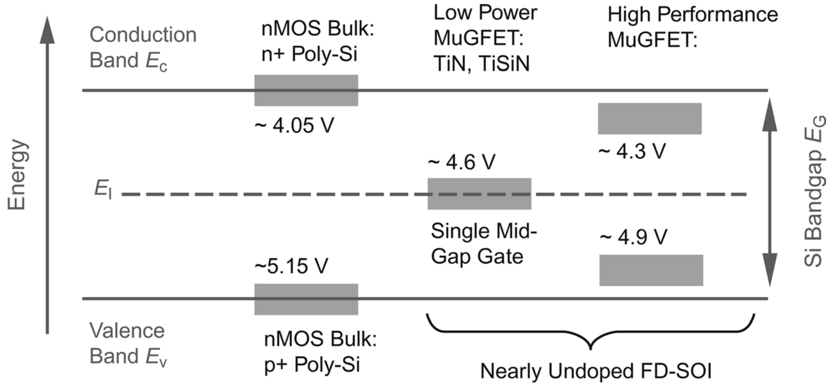


Figure 1.9 Desired workfunction for bulk and FD MOSFETS [24], Pacha ISSCC 2006.

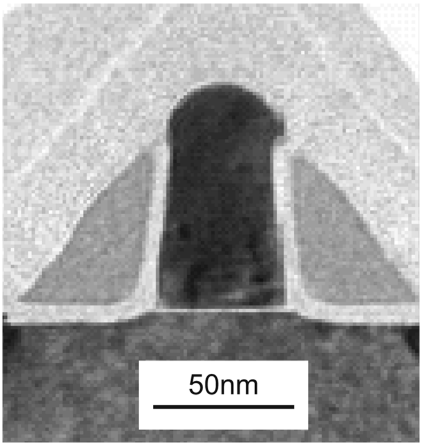


Figure 1.10 A fully silicided NiSi gate transistor [10].

Another gate material option is a tunable workfunction, such as fully silicided NiSi implanted with As and B, or Mo implanted with N. Until now, a shift of the workfunction in a conduction band direction of 200 to 300 mV has been reported [11]. A cross-section of a 50-nm transistor with a fully silicided NiSi gate is shown in Figure 1.10. Here, two approaches have been pursued: the first approach, with Thin Poly, allows the simultaneous silicidation of the source/drain (S/D) and gate, while the second approach uses CMP, offers the independent silicidation of the S/D and gate, and also avoids the formation of thick silicides in the S/D [10].

1.4 Ultra-Thin SOI

Many of the device problems due to short channel effects are related to the silicon bulk. The SOI [12] uses only a thin silicon layer for the channel, which is isolated from

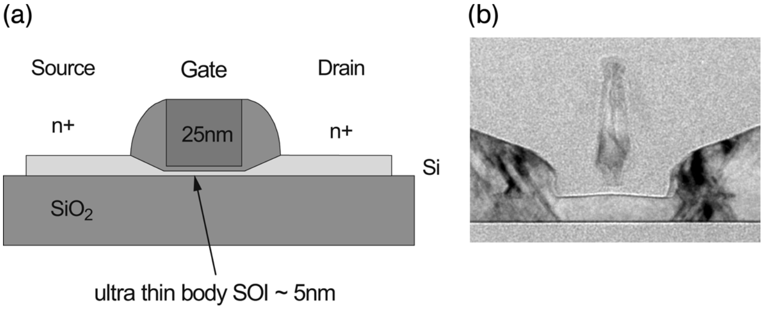


Figure 1.11 (a) A schematic cross-section of a fully depleted SOI transistor with a raised source drain. (b) TEM cross-section of a 12-nm gate fully depleted SOI transistor on 16-nm silicon.

the bulk by a buried oxide. Several companies producing semiconductors have already switched to SOI for high-performance microprocessors or low-power applications. Typically, the thickness of the Si layer is in the range of 50 to 100 nm, and the doping concentrations are comparable to those of bulk devices. This situation, which is referred to as *partially depleted SOI*, has several advantages, most notably a 10–20% higher switching speed. However, further down-scaling faces similar issues as the bulk, and here thinner Si layers [13], which lead to fully depleted channels, are of interest.

A schematic representation and a TEM cross-section of a thin-body SOI transistor with 12-nm gate length and 16-nm Si thickness on 100 nm buried oxide is shown in Figure 1.11. The gate has been defined with e-beam lithography while, for the contacts, raised source drain regions were grown with selective Si epitaxy and a high dose of dopands was implanted into the epi layer.

The experimental current–voltage (I–V) characteristics of n-channel SOI transistors with gate lengths down to 12 nm are shown in Figure 1.12. For gate lengths

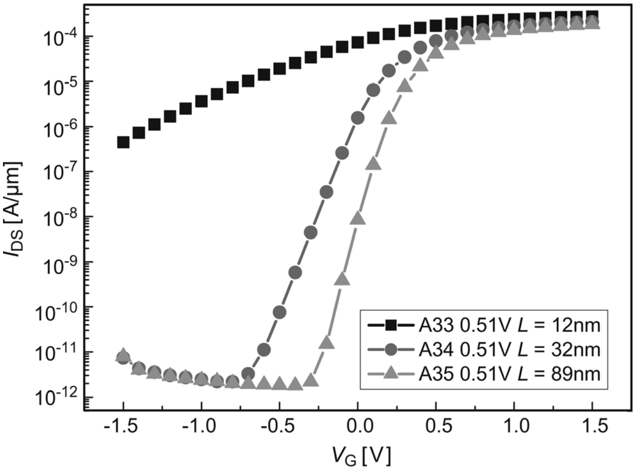


Figure 1.12 Experimental I–V characteristics of 89 to 12-nm SOI transistors on 16-nm silicon with undoped channel, n+ poly gate, $t_{ox} = 1.5$ nm.

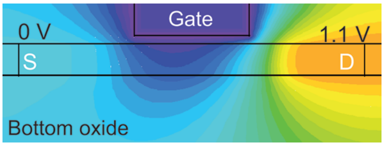


Figure 1.13 Potential distribution in a 30-nm single gate SOI transistor ($t_{Si} = 10$ nm, $t_{ox} = 2$ nm, $V_g = 0$ V, $V_d = 1.1$ V, midgap gate material).

>32 nm, subthreshold slopes of 65 mV dec^{-1} have been reached but, due to the still relatively thick Si body of 16 nm, short channel effects begin to increase below 30 nm gate length, and the transistors with 12 nm gate length cannot easily be turned off.

A two-dimensional (2-D) device simulation of the electrostatic potential of an SOI transistor with undoped channel and a thinner silicon body of 10 nm is shown in Figure 1.13 at a drain voltage of 1.1 V and a gate voltage of 0 V. For a gate length of 30 nm the gate potential controls the channel quite well. However, even with 10 nm Si thickness the potential barrier is slightly lowered at the bottom of the channel.

>This gives rise to an increase in the subthreshold slope as function of gate length, even for 5 nm Si thickness and 1 nm gate oxide (see the device simulation in Figure 1.16). A single-gate SOI exhibits the ideal subthreshold slope of 60 mV dec^{-1} down to about 50-nm gate lengths. In the gate length range of 50 to 20 nm, the turn-off characteristics are still good, and therefore ultra-thin SOI can provide a device architecture which is superior to that of bulk and suitable for the 32-nm node. A simple scaling rule for fully depleted SOI devices proposes a Si thickness of about one-fourth of the gate length in order to achieve good turn-off characteristics.

Whilst in these devices the channel was either low or undoped, this is not feasible in bulk devices because of the punch through from source to drain. The mobility of the charge carriers and the on-current is higher due to lower electric fields; this is shown graphically in Figure 1.14 for different channel doping concentrations. At a

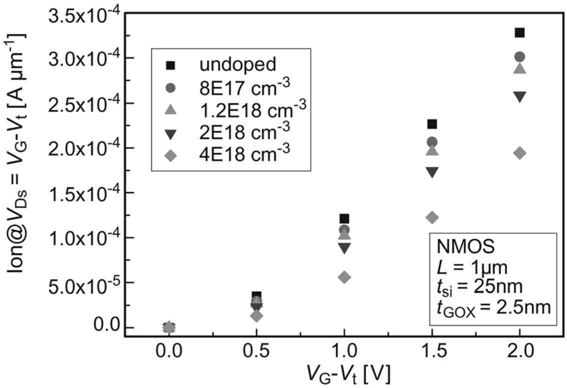


Figure 1.14 Measured on-currents at doped and undoped fully depleted SOI transistors at $V_g - V_t = 1$V.

gate voltage overdrive of 1 V the saturation current of the undoped transistor is twice that of the doped channel, at 4E18 cm^{-3} [14].

Moreover, without channel doping the Zener tunneling currents are reduced as well as electrical parameter variations, due to statistical fluctuations of the doping atoms.

1.5 Multi-Gate Devices

Further reduction of the gate length will require two or more gates for control of the channel, together with thin Si layers. The advantage of a multi-gate is to suppress the drain field much more effectively.

This is illustrated in Figure 1.15, by using the same simulation conditions as in Figure 1.13 and adding a bottom gate to the 30-nm SOI transistor. As shown in Figure 1.15, the electrostatic potential barrier is much higher than in the single-gate device. The better electrostatic control results in a steeper subthreshold slope; this can be seen in Figure 1.16, with a drift diffusion simulation of single- and double-gate transistors. A very thin Si thickness of 5 nm and a equivalent gate oxide thickness of 1 nm has been assumed, with a drain voltage of 1 V. Compared to the single gate, a

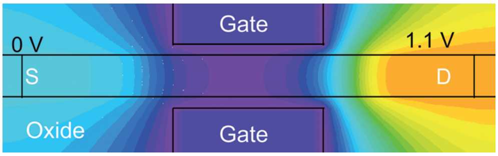


Figure 1.15 Electrostatic potential in a double-gate transistor with 30-nm gate length and 10-nm Si thickness; $V_g = 0\,V$; $V_d = 1.1\,V$; midgap gate material.

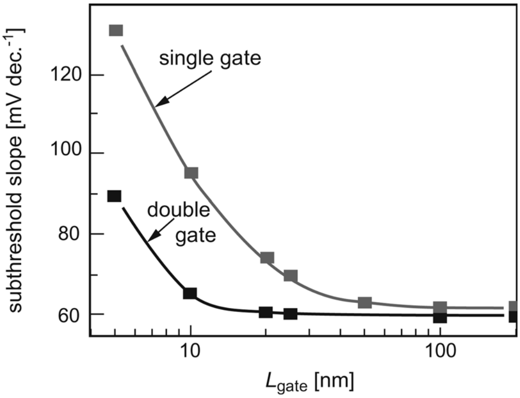


Figure 1.16 Simulated subthreshold slopes of single- and double-gate SOI transistors.

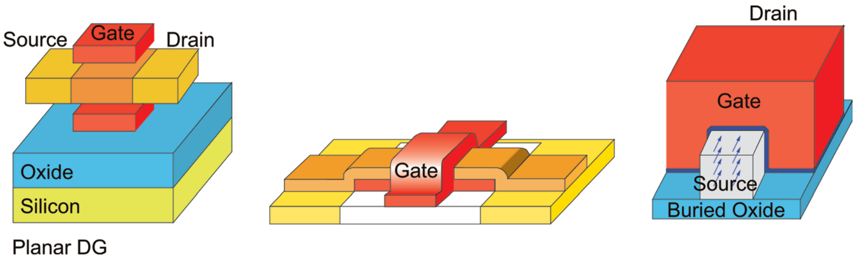


Figure 1.17 Three architectures for multi-gate devices. Left: Planar double-gate wafer-bonded [16]; Center: Gate all-around device [17]; Right: FinFET [18].

10-nm gate length and a subthreshold slope of 65 mV dec^{-1} are predicted for a double gate, and even 5 nm seems feasible with a reasonable subthreshold slope.

The challenge for multi-gate transistors will be to develop a manufacturable process with self-aligned gates to S/D regions. Three promising concepts have been investigated within the EC project NESTOR [15]: the first was a planar double-gate SOI transistor, which uses wafer bonding [16]; the second was a gate all-around device, based on silicon-on-nothing (SON) [17]; and the third was a FinFET type [18] (see Figure 1.17).

1.5.1 Wafer-Bonded Planar Double Gate

The planar double-gate transistor is an evolution of the ultra-thin SOI transistor, with a top and a bottom gate being used for better control of the channel. Processing starts with the bottom gate, spacers and elevated S/D regions using a SOI wafer with a thin silicon layer (see Figure 1.18). The gate is then encapsulated with dielectric layers and planarized with chemical mechanical polishing (CMP). Next, a handle wafer with an oxide layer is bonded onto the wafer with the bottom gate. The bulk Si of the top wafer is then completely removed down to the buried oxide, which acts as an etch stop. After removal of the buried oxide, a gate dielectric is deposited on the thin Si layer. Finally, the top gate and metallization are processed as in a conventional transistor.

An atomic force microscopy (AFM) image of a double-gate transistor test-structure with two separated contacts for the bottom and top gate is depicted in Figure 1.19a, using e-beam litho and an alignment mark for the top gate. A TEM cross-section of the first devices with a p+ poly-Si top and a n+ poly-Si bottom gate for V_t adjustment is shown in Figure 1.19b [19].

Recently, functional double-gate transistors with a TiN metal gate and lengths down to 12 nm and 8 nm for the top and bottom gates have been demonstrated [16] (see Figure 1.20). The 20-nm devices show good short-channel characteristics with $S = 102\ \text{mV}\,\text{dec}^{-1}$, an off-current in the range of $1\ \mu\text{A}\,\mu\text{m}^{-1}$, and an on-current of $1250\ \mu\text{A}\,\mu\text{m}^{-1}$.

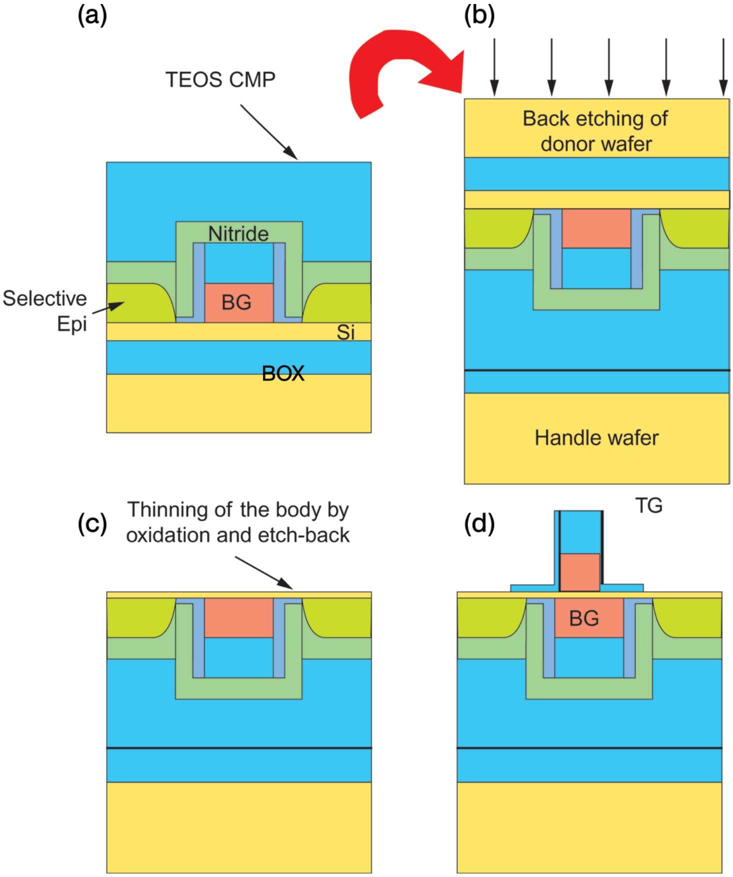

(a)
TEOS CMP
Nitride
Selective Epi
BG
Si
BOX
(b)
Back etching of donor wafer
Handle wafer
(c)
Thinning of the body by oxidation and etch-back
(d)
TG
BG

Figure 1.18 Process flow for a wafer-bonded double-gate transistor: Bottom gate, raised source drain and planarization, wafer bonding and back etch of Si bulk wafer, back etch Si channel, gate dielectric and top gate. BOX = Buried Oxide; BG = Buried Gate; TEOS = tetraethyl orthosilicate; CMP = chemical mechanical polishing.

1.5.2 Silicon-On-Nothing Gate All Around

The second approach for multi-gate architectures is based on silicon-on-nothing, as proposed by [20], which uses bulk Si wafers instead of SOI. A SiGe layer is grown with selective chemical vapor deposition (CVD) epitaxy and on top, non-selectively, a thin Si layer for the channel (see Figure 1.21). Next, the SiGe layer is removed by an isotropic etching process. The gate dielectric is then deposited around the silicon bridge, followed by the gate material, which is either poly-Si or a TiN metal gate. A 40-nm gate length and very thin Si channels down to 15 nm have been successfully

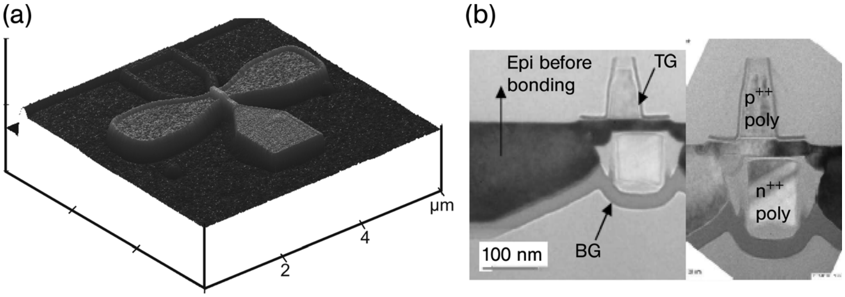


Figure 1.19 (a) AFM image of planar double-gate transistor with top and bottom gate. (b) TEM cross-section of planar double-gate transistor with n+/p+ poly gates.

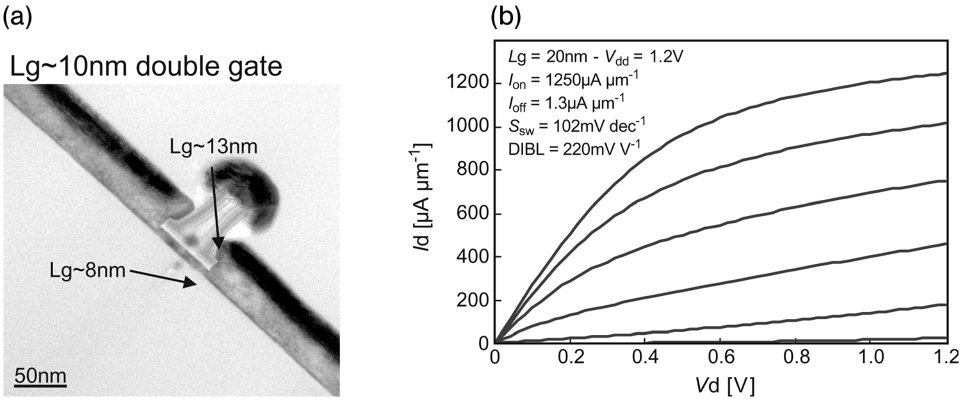


Figure 1.20 (a) TEM cross-section of a 10-nm double-gate transistor realized with wafer bonding [16]. (b) *I–V* characteristics of a 29-nm wafer-bonded double-gate device with a TiN metal gate [16].

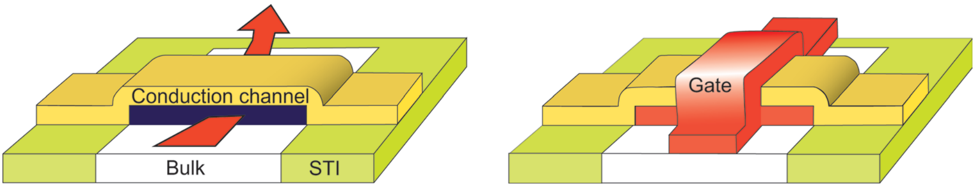


Figure 1.21 Gate all-around transistor processing based on silicon-on-nothing (SON) with a SiGe layer, which is removed for the gate [17].

fabricated [17]. Within the EC project NESTOR, devices with gate lengths of 25 nm have been achieved (see Figure 1.22a). These exhibit excellent short-channel characteristics, with $S = 70\,\text{mV}\,\text{dec}^{-1}$, $\text{DIBL} = 11.8\,\text{mV}$, and high on currents of $1540\,\mu\text{A}\,\mu\text{m}^{-1}$ ($I_{\text{off}} = 2\,\mu\text{A}\,\mu\text{m}^{-1}$, $t_{\text{ox}} = 2\,\text{nm}$) at 1.2 V (see Figure 1.22b). As shown in

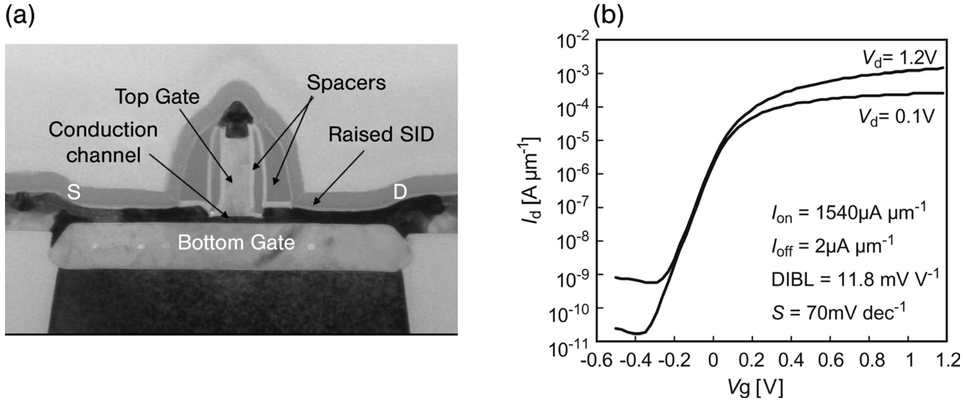


Figure 1.22 (a) TEM cross-section of 25-nm gate all-around SON transistor ($t_{ox} = 2$ nm; $t_{Si} = 10$ nm [15]. (b) Electrical characteristics of 25-nm gate all-around transistor [15].

Figure 1.22a, the bottom gate is still larger than the top gate. Ongoing studies have focused on a reduced bottom gate capacitance and a self-aligned approach.

Recently, multi-bridge transistors [21] have been reported using a similar type of SiGe layer etch technique for the fabrication of two or more channels stacked above each other and with an on-current of up to 4.2 mA μm^{-1} at 1.2 V.

1.5.3 FinFET

The FinFET [18, 22] can provide a double- or triple-gate structure with relatively simple processing (see Figure 1.23). First, the fin on SOI is structured with a tetraethylorthosilicate (TEOS) hardmask (Figure 1.23, left). A Si_3N_4 capping layer shields the top of the fin for a double-gate FinFET, and the same process flow can be used for triple-gate devices, without the capping layer. Next, a gate dielectric and the poly-Si gate are deposited and structured with litho and etching (Figure 1.23, center). The buried oxide provides an etch stop for the definition of the fin height. After this, a gate spacer is formed, raised source/drain regions are grown with epitaxy, and highly doped n+ or p+ regions implanted (Figure 1.23, right). The source/drain regions are enhanced using selective Si epitaxy to lower the sheet resistance. The facet of the Silicon epitaxy has been optimized to reduce the capacitance of drain to gate.

A TEM cross-section of a 20-nm tri-gate FinFET [23] is shown in Figure 1.24. Here, the top of the Si fin is also used for the channel, and no corner effects are observed at low fin doping concentrations. The fin and the gate layer have been processed with e-beam lithography. The smallest fin widths are in the range of 10 nm (see also Figure 1.30).

TEM cross-sections of a tri-gate device with larger fins of about 36 nm are also shown in Figure 1.24. The fin height is in the range of 35 nm, the corners are rounded by sacrificial oxidation, the gate dielectric is 2–3 nm SiO_2, and the poly gate surrounds the fin with a slight under-etch of the buried oxide.

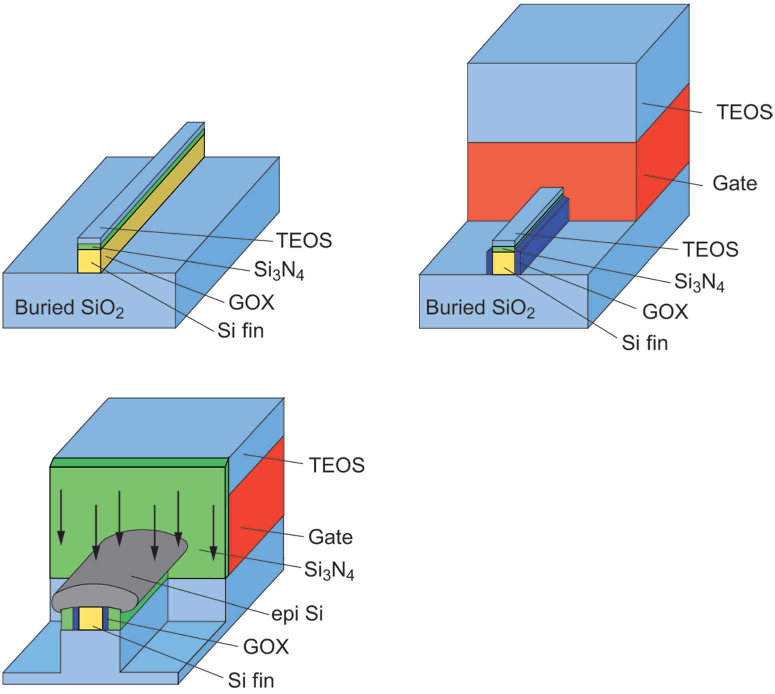


Figure 1.23 Process flow for a FinFET on buried oxide with a capping layer on top of the fin, a poly-Si gate, and raised source/drain regions with implantation. For details, see the text.

The measured I–V_g characteristics of n- and p-channel FinFETs with 20-nm and 30-nm gate length, respectively, are depicted in Figure 1.25. For the n-channel transistor a saturation current of 1.3 mA μm^{-1} (normalized by fin height) at an off-current of 100 nA μm^{-1} has been achieved at a gate voltage of 1.2 V, despite a relaxed

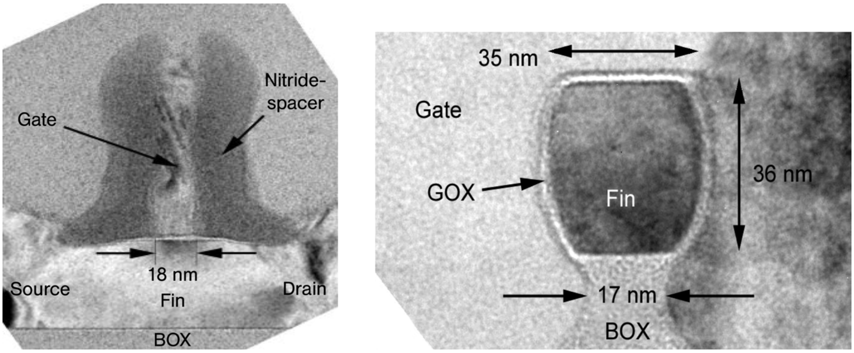


Figure 1.24 TEM cross-sections of a tri-gate FinFET on 100-nm buried oxide along and across the fin. Left: cross section along the fin; only the gate length is visible (18 nm). Right: cross section of the fin; on top it is 35 nm, height 36 nm, bottom ~17 nm.

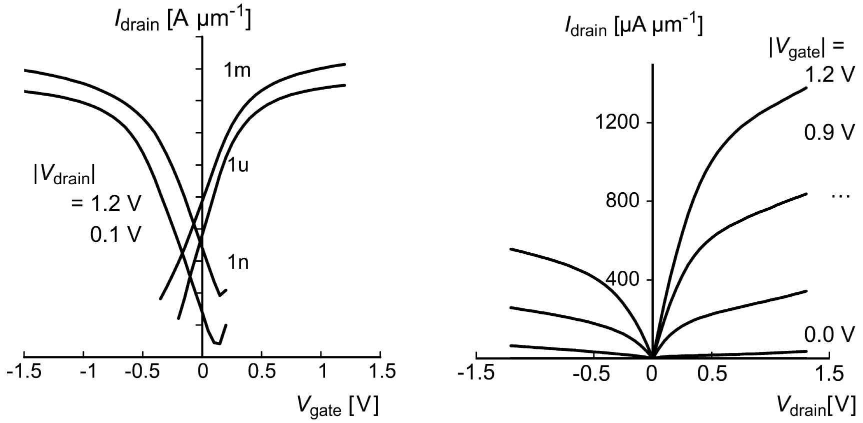


Figure 1.25 Measured *I–V* characteristics of 20-nm n-FinFETs (left) and 30-nm p-FinFETs (right) with $t_{ox} = 3$ nm (n) and 2 nm (p).

gate oxide thickness of 3 nm. For the p-channel, a high on current of 500 µA/µm and an off current in the range of 5 nA µm^{-1} is measured at 30-nm gate length. The FinFET has the advantage of self-aligned source and drain regions.

In Figure 1.25 the current was normalized on the height of a single fin. The electrical width of the device would be 2.2 times larger. For circuit applications, multi-fins are often needed in order to achieve higher drive currents (in Figure 1.26 the device has four fins) [24]. For a comparison with planar transistors it is important how many fins with height, width and pitch can be integrated on the same area as for the conventional device.

With respect to the switching time of multi-gate devices, the drive current together with the gate capacitance must be considered. Here, it was shown by simulation, that multi-gate devices can achieve 10–20% faster delay times compared to single-gate devices, mainly due to the better I_{off}/I_{on} ratio [25]. This was confirmed experimentally in Ref. [24] for inverter FO2 ring oscillators, where tri-gate FinFETs with TiSiN gate,

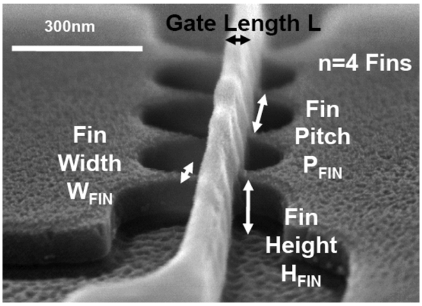


Figure 1.26 Scanning electron microscopy image of a multi-channel FinFET [24] with four fins on SOI. The gate length is 60 nm, fin width 30 nm, and pitch 200 nm.

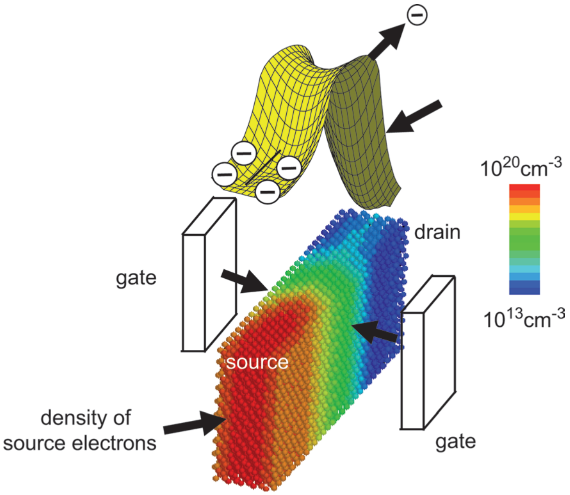


Figure 1.27 Atomistic simulation of a double-gate FinFET using the tight binding method.

55 nm length, and a low-doped channel achieved, with 21 ps, a much better speed performance than comparable planar MOSFETs in a 65 nm low-power CMOS technology, especially for sub-1 V power supply voltages.

1.5.4 Limits of Multi-Gate MOSFETs

The physical limit for the minimum channel length of multi-gate transistors has been investigated with 3-D quantum mechanical simulations using the tight binding method [26]. The device is composed of atoms in the silicon crystal lattice; the current can flow either by thermionic emission across the potential barrier of the channel, or directly via tunneling through the barrier from source to drain (see Figure 1.27).

In Figure 1.28, the simulated source drain current as a function of gate voltage is given with and without band to band tunneling for different gate lengths. An aggressive Si thickness of 2 nm and equivalent oxide thickness of 1 nm has been assumed. For gate lengths of 8 nm the tunneling contribution is on the order of the current over the potential barrier. At 4 nm the current is increased by two orders of magnitude by tunneling, but even 2-nm gates seem possible with off currents in the range of μA μm^{-1}, corresponding to ITRS hp specifications. Gate control is still effective and would achieve a subthreshold slope of about 140 mV dec^{-1}.

1.6 Multi-Gate Flash Cell

Multi-gate transistors are also very suitable for highly integrated memories with small gate lengths. Flash memory cells require thicker gate dielectrics than in logic

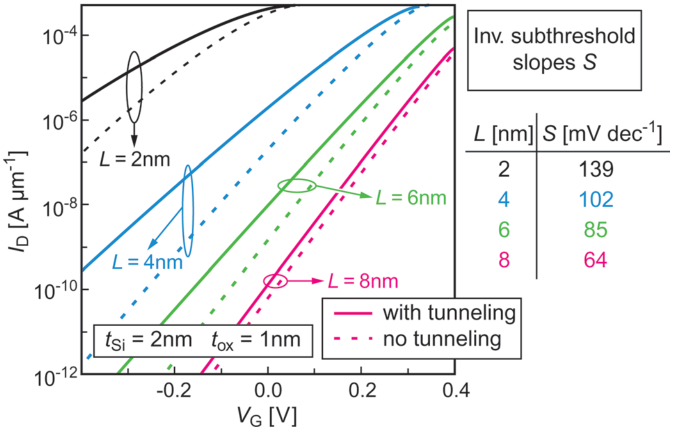


Figure 1.28 Thermionic and total current (+tunneling) of double-gate FinFETs simulated with the tight binding method [26].

applications, and therefore exhibit enhanced short channel effects. Currently, the most widely used Flash cell consists of a transistor with a floating gate [27] or a charge-trapping dielectric [28] sandwiched between the gate electrode and the channel region. A small amount of charge is transferred into the storage region either by tunneling or hot electron injection. This can be stored persistently and read out by a shift in the I–V_g characteristics. A schematic cross-section of a tri-gate FinFET memory transistor with improved electrostatic channel control compared to a planar device is shown in Figure 1.29, where a multilayer ONO gate dielectric around the fin serves as the storage element.

An experimentally realized memory structure [29] with a very small Si fin of 12 nm width and height of 38 nm is shown in Figure 1.30. The multilayer dielectric consisted of 3 nm SiO_2, 4 nm Si_3N_4 and 6.5 nm SiO_2.

The charge is uniformly injected into the nitride trapping layer by Fowler–Nordheim tunneling. The electrical function has been verified experimentally down to

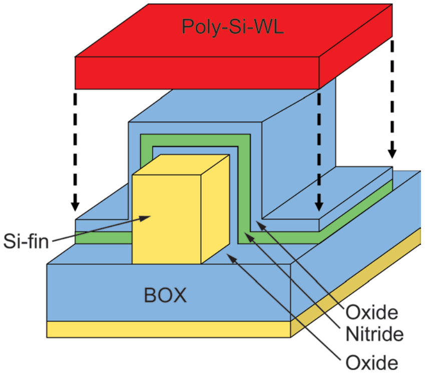


Figure 1.29 Schematic cross-section of a tri-gate charge-trapping memory field-effect transistor (FET).