Performance Limitations of Si Bulk CMOS and Alternatives for future ULSI

Krishna C. Saraswat, Donghyun Kim, Tejas Krishnamohan AND Abhijit Pethe

Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA

Abstract | Diminishing improvement in the on current (I_{ON}) and increase in off current (I_{OFF}) may limit the scaling of bulk Si CMOS. There are several technical issues that make proper device scaling increasingly difficult. Various techniques like ultra-thin gate dielectrics, shallow source/drain junctions and high channel doping are used to mitigate the short channel effects and improve the device performance. Most of these approaches however directly conflict with the goal of obtaining high carrier mobility, low subthreshold swing, low series resistance, and therefore large I_{ON} and low I_{OFF} at low supply voltage. To continue scaling beyond the 22 nm node, various architectural and material changes in the traditional MOSFET would be required for efficient operation of the transistor as a switch. A channel material with high mobility and therefore high injection velocity can increase on current and reduce delay. Currently, strained-Si bulk CMOS is the dominant technology and increasing the strain provides a viable solution to scaling. However, looking into future scaling of nanoscale MOSFETs it becomes important to look at higher mobility materials, like Ge and III-V materials together with innovative device structures, like the multi-gate FETs (MuGFETs) with high-κ dielectrics, metal gate and Schottky source/drain. For both Ge and III-V devices problems of leakage need to be solved. Novel heterostructures will be needed to exploit the promised advantages of Ge and III-V based devices.

MOSFET: Metal Oxide Semiconductor Field Effect Transistor – the basic building block of the electronics chips.

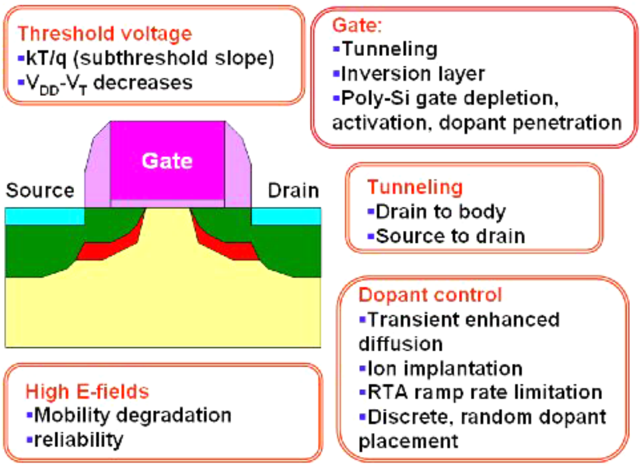
1. Introduction

There are several technical issues that make proper Si bulk MOS transistor scaling increasingly difficult below 45-nm node, as shown schematically in Fig. 1. In a long channel bulk metal-oxide-semiconductor field-effect transistor (MOSFET) the physics of transistor operation can be partitioned into two independent portions, i.e. gate-controlled charge formation in the channel, and drain-controlled charge transport. The threshold voltage V_T, at which the device turns on, is dependent only on the gate voltage and is independent of the drain voltage. Application of the gate voltage lowers a potential barrier near the source and allows electrons to flow from source to drain. Under normal transistor operation fundamental thermodynamics constrains the subthreshold swing to be greater than 60 mV/decade at room temperature. Degraded 2-D electrostatics at short gate lengths worsen (increase) this value – leading to higher off-state leakage current for the same V_T.

DIBL: Drain Induced Barrier Lowering which reduces threshold voltage for increase in drain voltage. DIBL increases with shrinking gate length as drain-electric field can couple to the source efficiently.

In reality, the potential barrier at the source is controlled by the gate as well as the drain through their respective capacitive coupling to that point. As the gate length is reduced, the drain influence becomes stronger. As a result, it becomes harder for the gate to control the source barrier and turn off the channel. The 2-D effects are manifested in various ways:

Figure 1: Problems associated with the scaling of CMOS transistors limited by physical and technological reasons.


Threshold voltage
▪kT/q (subthreshold slope)
▪VDD-VT decreases
Gate:
▪Tunneling
▪Inversion layer
▪Poly-Si gate depletion, activation, dopant penetration
Gate
Source
Drain
Tunneling
▪Drain to body
▪Source to drain
Dopant control
▪Transient enhanced diffusion
▪Ion implantation
▪RTA ramp rate limitation
▪Discrete, random dopant placement
High E-fields
▪Mobility degradation
▪reliability

Quantum Mechanical Tunneling: If an occupied state and an available state of carriers are separated by a thin potential barrier, carriers can flow through the barrier due to wave function overlap, which can not be explained classically.

Band-to-band tunneling: Tunneling of carriers from one band (conduction/valence) to another (valence/conduction) - either directly (no change in crystal momentum) or indirectly (changing crystal momentum - phonon or trap assisted).

Sub-threshold swing: The inverse of rate of change drain current with respect to gate voltage such that the gate voltage is below threshold voltage. This has a minimum value of 60mV/decade at room temperature.

1. reduction in threshold voltage with shrinking gate length (V_T roll-off),
2. V_T reduction with increasing drain voltage (drain induced barrier lowering – DIBL),
3. degraded subthreshold swing.

Collectively, these phenomena known as ‘short channel effects (SCE)’ tend to increase the off-state static leakage power. Thus far, device designers have tried to suppress SCE in short gate length devices by a number of methods:

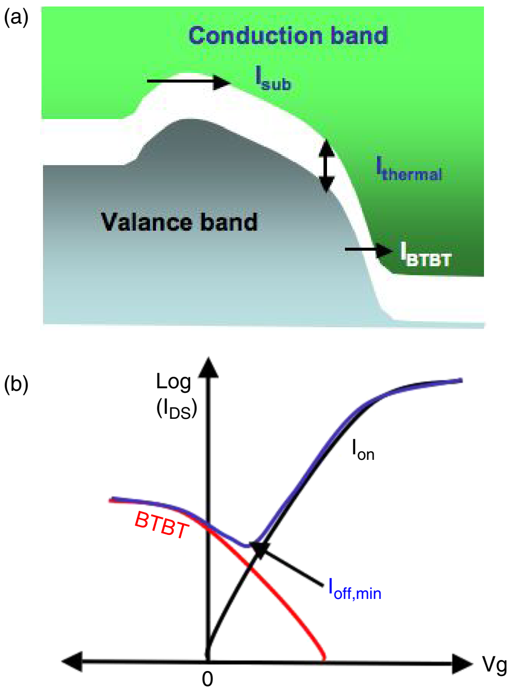
(a) reducing the gate oxide thickness to improve the gate control over the channel,
(b) lowering the source/drain junction depth (especially near the gate edge, where the source/drain regions are called ‘extensions’) to reduce the drain coupling to the source barrier,
(c) increasing the channel doping to terminate the electric field lines which originate from the drain and propagate towards the source. In modern bulk MOSFETs, the channel doping is tailored to have complicated vertical and lateral profiles so as to minimize the impact of gate length variations on the short channel effects.

Each of these approaches comes at a cost which either degrades transistor performance (speed) or introduces a new static leakage mechanism:

(a) As the gate oxide gets very thin, quantum mechanical tunneling allows a leakage gate current to flow. In the direct-tunneling regime, encountered for oxides thinner than about 3 nm, the gate leakage current increases dramatically (∼3X for every 1 A of thickness reduction). The gate leakage can increase standby power as well as compromise proper logic gate operation[1]. Many people have proposed replacing the silicon dioxide (SiO_2) with higher permittivity (high-k) gate dielectrics[2] such as zirconia (ZrO_2) or hafnia (HfO_2). These enable high gate capacitance with physically thick insulators through which tunneling is low. However, the introduction of such new materials without the accompanying degradation of mobility and reliability is very challenging and remains an area of intensive ongoing research.
(b) As the source/drain junction depths get shallow, their doping must be increased so as to keep the sheet resistance constant. Solid solubility of dopants puts an upper limit ($\sim 10^{20}$ cm^{-3}) on the doping density. Therefore, further reduction in junction depth causes an increase in the series resistance encountered in accessing. This degrades the overall transistor performance. Also, from a technological point of view, it becomes difficult to form ultrashallow junctions that remain abrupt after the annealing steps needed to activate the dopants and achieve low resistivity[3].
(c) As the doping density in the channel is increased for SCE suppression, the carrier mobility is degraded due to increased scattering from the ionized dopant atoms. Besides, the subthreshold swing gets worse due to higher depletion capacitance that ‘steals’ away part of the gate voltage from the surface potential. For very high channel doping near the source/drain extensions, another component of static leakage, band-to-band tunneling (BTBT), becomes important. Finally, as the channel volume reduces in extremely scaled transistors, the random placement of discrete dopant atoms cause stochastic inter-device variations [4].

The need to enhance drive currents while scaling the transistor size and decreasing supply voltage, has been accompanied with an exponential increase in the static, off-state leakage of the device. While the active power density on the chip has steadily increased with gate length scaling, the static power density has grown at a much faster rate. The active power arises due to the dissipative switching of charge between the transistor gates and supply/ground terminals during logic switching. The sub-threshold power, also known as static or standby power, is dissipated even in the absence of any switching operation. It arises due to the fact that the MOSFET is not an ideal switch – there

Figure 2: (a) Various leakage mechanisms in a MOSFET: subthreshold conduction, thermal generation and band-to-band tunneling. (b) Minimum off state leakage is decided by the higher of these mechanisms.



Floating body and history effect: Floating body effect is dependence of body potential of SOI transistor on the history of biasing and carrier recombination. It causes history effect which means dependence of transistor threshold voltage (in turn, gate delay) on previous states.

is still some leakage current that flows through it in the off-state. Various leakage mechanisms are shown in Fig. 2. For long channel devices the subthreshold leakage was the primary mechanism responsible for leakage. However for nanoscale devices and for higher mobility materials BTBT leakage becomes important. Static power dissipation was a relatively insignificant component of the total chip power just a few generations back, but it is now comparable in magnitude to the active power. Management and suppression of static power is one of the major challenges to continued gate length reduction for higher performance. Once the scaling of the conventional bulk Si MOSFET starts slowing down, the insertion of performance boosters, like novel materials and non-classical device structures, will become necessary to continue to improve performance.

2. Novel Device Structures

2.1. *Partially Depleted SOI-MOSFETs*

In partially depleted SOI (PDSOI) MOSFETs (Fig. 3), a layer of insulating SiO_2 separates the upper device-containing layer from the bulk Si below. The PDSOI MOSFET is designed similar to a bulk MOSFET with the same dimensions. In the OFF state, the depletion width under the gate is smaller than the thickness of device layer. The PDSOI offers lower parasitic capacitance at the Source/Drain nodes in an inverter circuit, and lower self-capacitances and higher switching speeds. The potential of the floating body is determined dynamically by capacitive coupling of the various electrodes connected to this layer. Hence charge can accumulate in this region, which modifies the device characteristics (floating-body and history effects). Using an implanted body contact may mitigate the floating body effects. Using clever circuit design, the history effects may be used to speed up circuits. PDSOI technology has been successfully ported into high volume manufacturing; however PDSOI runs into similar problems as the bulk MOSFET with respect to scaling and hence may not be a scalable technology for future generations.

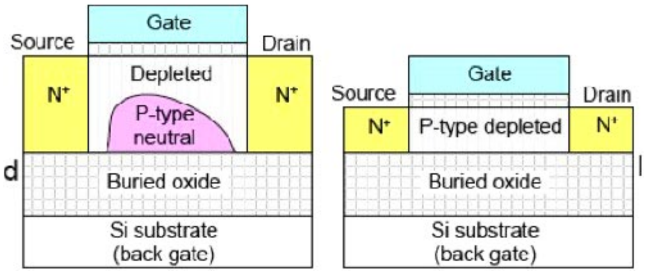
2.2. *Fully Depleted SOI MOSFETs*

If the top film in the PDSOI MOSFET is thinned down so that the semiconductor film under the gate is completely depleted in the OFF state of the device, then it is referred to as a fully depleted (FDSOI) SOI MOSFET (Fig. 3). By eliminating the thin-doped region, the history effects and the floating body effects are suppressed in the FD-SOI MOSFET. Since reducing the film thickness can now control the short channel effects, lower channel doping densities may be employed. Also, the channel vertical electric fields are reduced for the same channel carrier concentration. Hence the inversion layer mobility may be increased without compromising the OFF state current. By increasing the buried oxide thickness, essentially ideal sub-threshold slopes of 60 mV/decade may be achieved. However, this increases the drain control on the source-channel barrier through the buried oxide. Using a thin buried oxide can mitigate this, by terminating the drain fields on the back substrate at the expense of degraded sub-threshold slope.

2.3. *Multi-Gate MOSFETs*

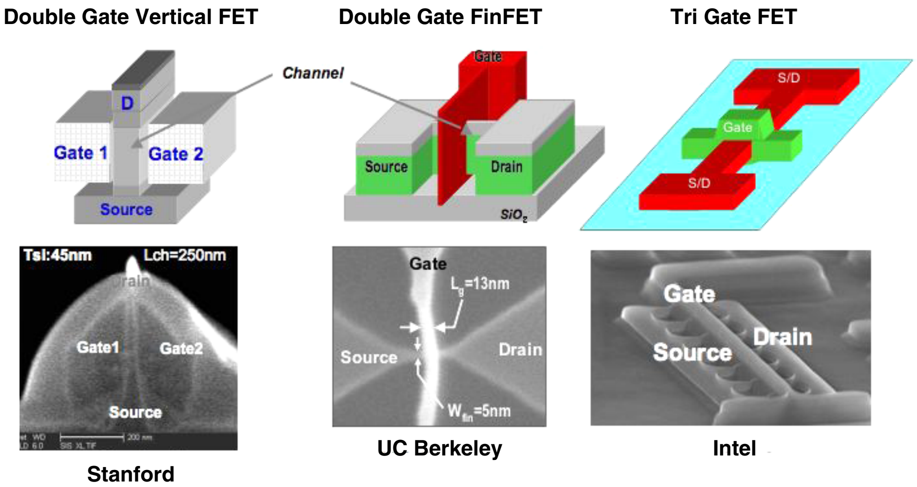
The gate control on the channel is increased by geometrically placing the gate close to the channel. Tighter gate coupling may be achieved by increasing the number of gates from single-gate FDSOI to a double-gate SOI, tri-gate and a gate all around MOSFET. Examples of multigate device structures are shown in Fig. 4. The double-gate MOSFET[7] provides for a symmetric device architecture where the channel is controlled by gate on either side of the Si film. Since the gate control is increased, the requirements on the Si film thickness are relaxed as compared to a FD-SOI with the same gate length to achieve similar OFF state performance. By suitably designing the device, volume inversion is achieved

Figure 3: Evolution of the architecture of the single-gate MOSFET from the PDSOI to the FDSOI structure.



V_T roll-off: Reduction of threshold voltage of transistors as a function of gate length, below 1 micron.

Figure 4: Various multi-gate MOSFET structures.



Thermal injection velocity: Lower limit of source injection velocity arising from thermal energy of carriers at nonzero temperature.

Source injection velocity: Mean carrier velocity at which carriers are injected from source to channel overcoming the source barrier - this limits the current in ballistic transistors.

in the film, where most of the inversion charge resides in the center of the Si film, which has very low vertical field and hence provides for higher mobility. On the other hand, because of the two gates, the channel length can be further reduced. Advanced architectures like the tri-gate MOSFET[8] may be used to increase the effective number of gates allowing for further reduction in the gate length.

3. The Need for High Mobility Channel

Looking into the past 15–20 years, the scaling of planar bulk silicon MOSFETs has been very successful that guaranteed a roughly 17% device performance enhancement per year[5]. In the coming 15–20 years, the International Technology Roadmap for Semiconductors (ITRS) projects the future device performance to closely follow this historical trend. However, theoretical calculations suggest that by simply maintaining the same planar bulk geometry and/or Si channel the MOSFETs may not be able to keep up with the required performance beyond the year of 2010.

The saturation of bulk Si MOSFET drive current (I_{Dsat}) upon dimension shrinkage is limiting the prospect of future scaling as illustrated in Fig. 5[6]. By reading the x-axis backward, the I_{Dsat} from both n-MOSFETs and p-MOSFETs stop to improve beyond a drawn gate length of 0.25 μm. In this analysis the off current was kept constant. To understand this saturation phenomenon, numerous theoretical and experimental analyses were carried out[9–13]. First of all, the I_{Dsat} (and transconductance) in very short-channel MOSFETs is believed to be limited by carrier injection from the source into the channel[9]. In order words, the source injection velocity (v_{src}) saturates during scaling and that its limit is set by thermal injection velocity (v_{inj})[10] as depicted in Fig. 6. The drain current in saturation normalized to channel width can be given as $I_{\text{Dsat}}/W = C'_{\text{ox,inv}}(V_{\text{GS}} - V_t)v$, where v is the effective carrier velocity at the virtual source. This virtual source point is located at the top of the potential barrier between the source and the

Figure 5: The saturation of bulk Si MOSFET drive current (I_{Dsat}) upon dimension shrinkage is limiting the prospect of future scaling[10].

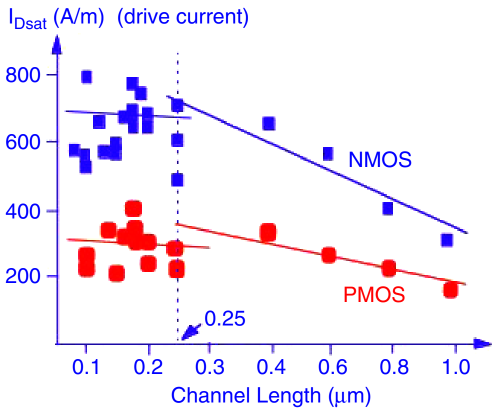
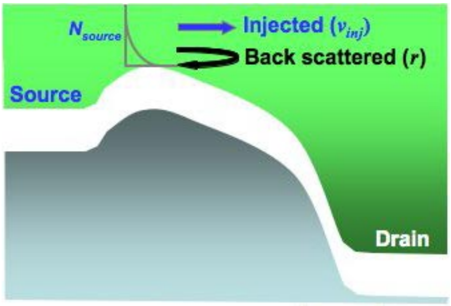


Figure 6: Models used to describe the drive current in nanoscale MOSFETs. The drive current (I_{Dsat}) is proportional to the velocity v_{inj} of the carriers at the source and the backscattering rate r, both of which are determined by the low-field mobility.



$$I_{sat} = qN_{source}v_{inj}\left(\frac{1-r}{1+r}\right)$$

$$\left.\begin{array}{l}\text{High } v_{inj}\\ \text{Low } r\end{array}\right\} \Rightarrow \text{Low } m^*_{transport}$$

$$m^*_{transport} \;\alpha\; \text{mobility}$$

Low-field effective inversion-layer mobility: Mobility of carriers in the inversion layer at low field which is classically well defined with Boltzmann transport model.

Off-equilibrium transport: Carrier transport under high electric field where carriers are not in thermal equilibrium with lattice causing so-called hot-carrier transport.

Quasi-ballistic transport model: Transport model which assumes carriers to move from source to drain almost without any scattering in the channel.

channel[14]. Hence for extremely scaled MOSFETs the velocity of the carriers at the source determines the net drive current. By using new materials in the channel, which have lower carrier effective masses in the length direction, the injection velocity in these materials may be increased leading to higher drive currents. The values of v_{inj} extracted from literature[15] are shown in Fig. 7. Carrier velocities have been increasing primarily because of the reduction of the characteristic length of the potential barrier near the source, as L_G is scaled, and therefore a reduction in the backscattering. Increasing the mobility reduces this backscattering rate, hence providing for higher carrier velocities and hence higher drive currents.

By corroborating measured velocity and mobility dependencies on deeply scaled MOSFETs, the carrier velocity was shown to have a direct proportionality with the low-field effective inversion-layer mobility[12]. Unfortunately, mobility is not a well-defined quantity in a nanoscale MOSFET under high drain bias where off-equilibrium transport dominates; however, in Ref.[9], it was demonstrated that the drain current of a nanoscale MOSFET is directly related to the near-equilibrium mean-free-path for backscattering, which can be deduced from measurements on a corresponding long-channel MOSFET for which the mobility is well defined. In brief, these results suggested that mobility continues to be of crucial importance to saturated transconductance and I_{Dsat} as channel lengths decrease below ∼ 100 nm[13].

Current flow in ultra-short channel transistors may be described by a small number of scattering events referred to as a quasi-ballistic transport model. Fig. 6 depicts the factors that dominate current drive in the classical drift model and the quasi-ballistic model respectively. In the drift model, the velocity near the source region is strongly affected under non-stationary transport by the low-field mobility near the source region, while the injection velocity and the back-scattering rate at the source determine the velocity near the source region in the quasi-ballistic model. Both these factors are strongly related to the low-field mobility at the source. As a result, both the models predict an increase in the current in nanoscale MOSFETs by increasing the low-field mobility near the source region.

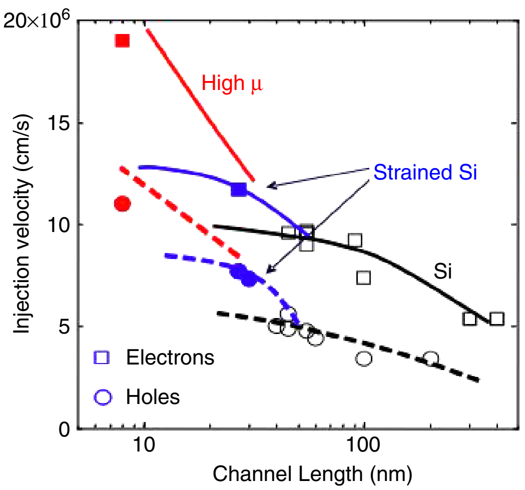
Fig. 7 from the work of Antoniadis at MIT[15] shows that the carrier velocity increase has saturated with scaling of bulk CMOS. A channel material with high μ and therefore high injection velocity (v_{inj}) can increase $\mathrm{I_{ON}}$ and reduce delay and thus allow continued scaling. There are several different approaches to enhance the carrier transport in future MOSFETs to obtain high drive currents. Currently, strained-Si is the dominant technology for high performance MOSFETs and increasing the strain provides a viable solution to scaling[16,17]. MOSFET delay has continued to decrease by use of Si strain to boost velocity, however, velocity boosting will also saturate with strain-based Si band engineering. High mobility is not the only parameter which guarantees high $\mathrm{I_{ON}}$ and low $\mathrm{I_{OFF}}$. In order to obtain these characteristics many other boundary conditions must be met, e.g., high density of states, small bandgap, etc. For example, even though, most

Effective conductivity mass: Effective mass value of carriers used for conductivity calculations.

Atomic layer deposition (ALD): Self-limiting, sequential process that deposits conformal, pin-hole free thin films on substrate.

Rapid Thermal Nitridation (RTN): Exposing SiO_2 to NH_3, N_2O or NO ambient at high temperature for a very short time in a rapid thermal processing (RTP) system. This results in introduction of Nitrogen in SiO_2 matrix, preferentially at the interfaces, resulting in improved reliability.

Figure 7: Historical evolution of the virtual source velocity as a function of reducing channel length. The advantages that were gained with strained-Si are also shown. The points in red depict the velocity target required for continued performance improvement.[17]



III-V materials have high bulk electron mobilities, the drive currents in NMOS in these materials may not be necessarily high. This is because in strong quantization due to either space or electric field, the current is limited by the transport properties of the L-valley, which has a much higher mass in the transport direction than the Γ-valley. Further more, the higher mobility materials have small bandgaps and hence lead to higher BTBT limited OFF current in these transistors. Therefore, looking into future scaling of nanoscale MOSFETs it becomes important to look at higher mobility materials (Table 1), like Ge and III-V *together with innovative device structures* which may perform better than even very highly strained Si. This approach will be described in rest of this paper.

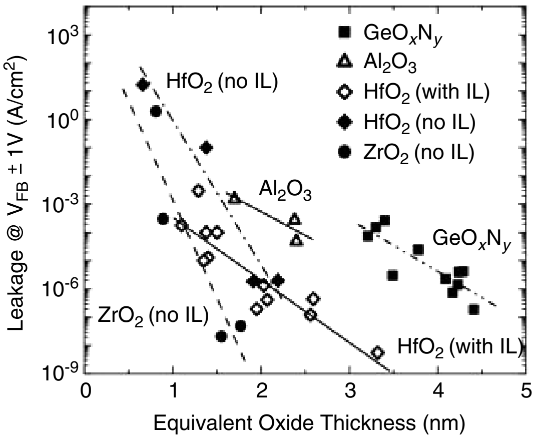
4. Ge MOSFETs

In Ge, the lower effective conductivity mass (m*) of electrons and holes is responsible, respectively, for higher μ_n and μ_p. Historically, Ge had been one of the most important semiconductors as the first transistor and the IC were fabricated in it. However, Ge gave way to Si due to many problems. The native Ge oxides, GeO and GeO_2, are either water soluble or volatile. Ge has much smaller direct band gap (E_G) giving rise to high I_{OFF}. For Ge to become mainstream, surface passivation, innovative device structures to overcome leakage and heterogeneous integration of crystalline Ge layers on Si must be achieved.

Table 1: Properties of various prospective MOSFET channel materials

Material Property	Si	Ge	GaAs	InAs	InSb
Electron mobility	1600	3900	9200	40000	77000
Hole mobility	430	1900	400	500	850
Bandgap (eV)	1.12	0.66	1.424	0.36	0.17
Dielectric constant	11.8	16	12.4	14.8	17.7

Figure 8: Gate leakage measured at V_{FB} + 1V vs. EOT for various dielectrics on Ge.

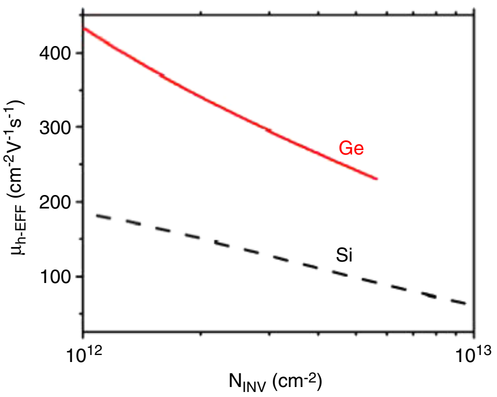


5. Bulk Ge MOSFETs

We demonstrated for the first time in 2002 passivation of Ge with ZrO_2[18]. Since then surface passivation of Ge has been extensively investigated by many researchers with high-k metal oxides of Zr, Hf, Al, La, and Er. In our recent work Ge passivation with its native oxynitride (GeO_xN_y)[19,20] and HfO_2 or ZrO_2 deposited by atomic-layer deposition (ALD) system has been studied[21]. The optimum dielectric stack could be attained by rapid thermal nitridation (RTN) of Ge in ammonia to form GeO_xN_y followed by ALD of the hi-κ film. Excellent electrical characteristics were obtained from MOSCAPs for both techniques with low leakage (Fig. 8), good C–V characteristics and reasonably low interface state density. The RTN technique was also employed to passivate the Ge surface prior to the deposition of SiO_2 for field isolation.

p-MOSFETs have been demonstrated in bulk Ge with high-k gate dielectrics and metal gates showing low gate leakage and high μ_p (Fig. 10)[19–21]. However, the n-MOS device performance is far below the expected theoretical predictions. Higher density of interface states near the conduction band

Figure 9: Hole mobility measured in bulk (100) Ge p-MOSFETs



Full-band Monte-Carlo: Monte-Carlo simulation of carrier transport using full E-k band structure which is required to account for high energy carriers.

1-D Poisson-Schrodinger: Solving one dimensional coupled Poisson and Schrodinger equation to take into account quantization effect.

Quantum effect spintronic devices: Devices that exploit controlling the quantum spin states of electrons.

Double-Gate SOI, tri-gate, gate-all-around MOSFET: In double-gate, Silicon On Insulator technology is used to build transistors with two gates instead of one to have better gate control over channel, in turn better on to off current ratio. Tri gate structure has gate on three sides of channel and gate-all-around transistor has gate surrounding the channel to achieve the same goal.

may explain this[19,22]. Better surface passivation technology needs to be developed to improve n-MOS performance.

5.1. *Strained Ge and Si$_x$Ge$_{1-x}$*

Straining Ge and Si$_x$Ge$_{1-x}$ can significantly increase μ_p because of a reduction in m* and the band splitting due to strain[23]. In extremely scaled MOSFETs, the relation between the short-channel I_{on} and μ is not direct or obvious. Through detailed BTBT including band structure and quantum effects, full-band Monte-Carlo and 1-D Poisson-Schrodinger simulations on ultra-thin, nanoscale DG FET structures shown in Fig. 10, we have systematically compared different high hole mobility p-MOSFET channel materials in terms of the drive current, intrinsic delay and off-state leakage[23]. Fig. 11 shows I_{on} enhancement of the different high-μ_p materials for structure 1 of the Fig. 10. The highest drive currents are obtained from the compressive biaxial strained-Ge (s-Ge) substrates, and for tensile strained-Si (s-Si) channels.

5.2. *Si/Ge/Si Heterostructure FET (H-FET)*

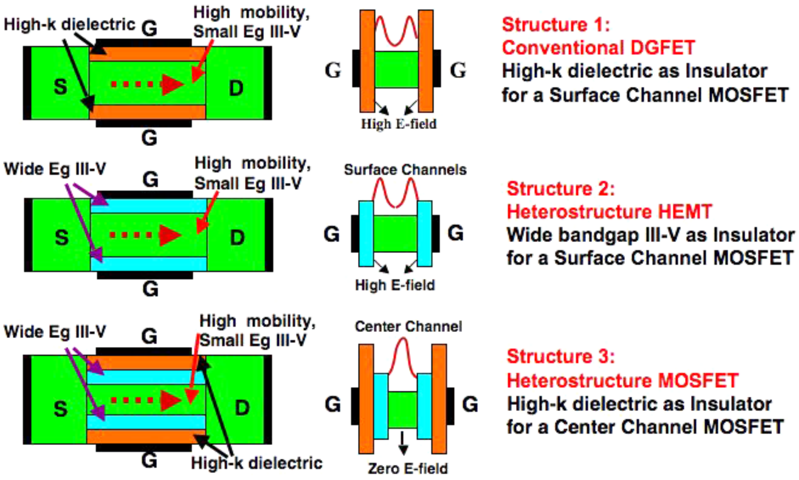
In long channel devices minimum standby off-state current ($I_{OFF,min}$) is determined by subthreshold conduction. However, in a nanoscale transistor generally it is determined by the BTBT leakage, I_{BTBT}. We have developed a new Band to Band Tunneling model[28], which captures band structure information, all possible transitions between different valleys, energy quantization and quantized density of states as schematically shown in Fig. 12.

Strain in general results in reduction in the E_G and hence enhanced I_{BTBT}. Confinement on the other hand results in increased E_G as shown in Fig. 13 and hence reduced I_{BTBT}. Structure 2 and 3 of Fig. 10 propose novel devices to combine strain and quantum mechanical confinement to obtain desired transport properties with reduced off-state leakage. In these structures the transport can be confined to the center of the channel in a high mobility material flanked by a high E_G material. The mobility is further enhanced due to strain, reduced electric field in the center of the double gate structure due to symmetry and the channel being away from the dielectric interface. The bandgap of the center channel can be increased due to confinement by keeping it very thin. We have demonstrated[24,25] a novel Si/s-Ge/Si heterostructure FET (H-FET), in which the transport occurs in high μ_p s-Ge and leakage in wider E_G Si (structure 3 of Fig. 10). This reduces the I_{BTBT}, while retaining high μ_p of Ge. The confinement of thin Ge between Si results in an increase in the E_G and hence reduction in I_{BTBT}, while strain keeps μ_p high. Experimentally, the resulting optimal structure obtained was an ultra-thin, low defect, fully strained Ge epi channel on relaxed Si (r-Si) (Fig. 14). H-FETs on bulk Si show a ~2X μ_p enhancement over Si, while H-FETs on SOI show even higher μ enhancements of >4X over Si (Fig. 15). Both types of H-FETs show reduction in I_{OFF} compared to bulk Ge devices. In particular H-FETs on SOI show significant reduction in I_{OFF} as shown in Fig. 16, due to reduced E-field in Ge and E_g increase due to confinement.

5.3. *Metal Source/Drain*

Parasitic resistance in the S/D regions of the conventional MOSFET has been identified as one of the primary problems of the non-scaling of drive currents in transistor scaling. Replacing the S/D regions of transistors with metals has been suggested as one of the techniques, which might help reduce this effect. Use of the metal S/D offers additional advantages of low-temperature processing for S/D formation, elimination of the parasitic bipolar action and inherent physical scalability of the gate lengths due to the abrupt silicide-silicon interface. However, for the I_{ON} of the metal S/D to be better than diffused S/D the Schottky barrier to channel needs to be very small. In the case of Ge we have found that the Fermi level at metal-Ge Schottky barriers is pinned near the valence band of Ge for a variety of metals, including, Ni, Co and Ti[26]. This provides a very small barrier to the holes in the channel in a PMOS and a large barrier to the n-type substrate. Due to the small barrier height to holes coupled with the high inversion hole mobility, Schottky S/D Ge transistors provide for very high drive currents. We have built high performance PMOSFETs on Si substrates

Figure 10: Different FET structures studied. Structure 3 shows the best promise to get high I_{ON} and low I_{OFF}. Structure 2 may be a good compromise if surface passivation can't be made to work.


High-k dielectric
G
High mobility, Small Eg III-V
S
D
G
G
G
High E-field
Structure 1:
Conventional DGFET
High-k dielectric as Insulator
for a Surface Channel MOSFET
Wide Eg III-V
G
High mobility, Small Eg III-V
S
D
G
Surface Channels
G
G
High E-field
Structure 2:
Heterostructure HEMT
Wide bandgap III-V as Insulator
for a Surface Channel MOSFET
Wide Eg III-V
G
High mobility, Small Eg III-V
S
D
G
High-k dielectric
Center Channel
G
G
Zero E-field
Structure 3:
Heterostructure MOSFET
High-k dielectric as Insulator
for a Center Channel MOSFET

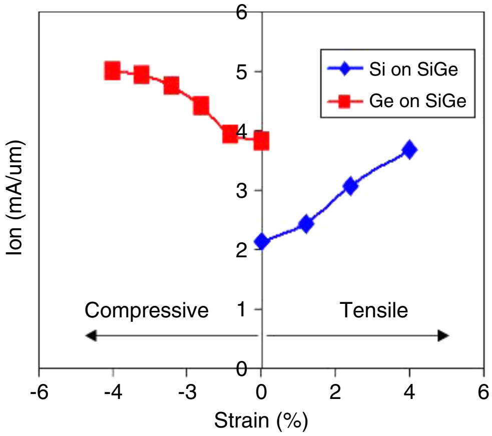
Intrinsic gate delay: The gate delay not considering parasitic component of the circuit.

using a Si/Ge/Si heterostructure channel and NiSi Source/Drain regions[26]. Schematic of the transistor is shown in Fig. 17. Excellent PMOS characteristics were achieved when the doped S/D region in Ge transistors was replaced with metallic NiGe. Using a thin Ge layer within the inversion region of a Schottky Si – PMOSFET provides for higher hole mobility (∼2X), as shown in Fig. 18, and much higher drive currents due to almost zero barrier height to holes in the channel. Also the OFF state leakage is maintained at a low value because it is limited by the large barrier height in the wider bandgap Si and Ge quantization. The transistor hence, combines the advantages of high mobility, and low parasitic resistance and is an attractive candidate for scaling PMOSFETs into the sub-20nm regime.

$I_{OFF,MIN}$ is investigated in Double Gate MOSFETs (structure 1 of Fig. 10) with various high mobility p-MOSFET materials using the new band to band tunneling model. As shown in Fig. 19, the leakage current for s-Si increases monotonically with increasing strain due to the rapid reduction in the E_G and m*, whereas the dependence of I_{OFF} for s-Ge is not monotonic and reveals an optimum point of minimum leakage[23].

Ultimately, device performance is determined by intrinsic gate delay (CV/I) and $I_{OFF,\min}$ achievable. A common terminology used in Fig. 20 is a channel material (x, y) where, x denotes the Ge content in the channel material and y denotes the Ge content in an imaginary relaxed (r) substrate to which the channel is strained (s). For lower values of strain the performance of s-Si and s-SiGe p-MOSFETs are very comparable. However, as we scale to higher strain materials, s-SiGe rapidly outperforms s-Si. Further, by using a s-Ge H-FET, the switching frequency can be increased $(>4x)$ and I_{OFF} can be further effectively reduced. The reduced I_{BTBT} while retaining high I_{ON} makes the Ge p-channel H-FET suitable for scaling into the sub-20nm regime. Similar performance tradeoffs for uniaxial strained Si and Ge are presented in[26].

Figure 11: Full-band Monte Carlo simulation of drive current for strained Ge and Si for structure 1 of Fig. 9 for a p-MOSFET.



6. *n*-Channel MOSFET

The n-MOS device performance in Ge so far has been far below the expected theoretical predictions. Higher density of interface states near the conduction band[19,22] may explain this.

Figure 12: I_{ON} modeling in III–V materials considers low density of states in the Γ-valley, reducing $Q_{\mathrm{inversion}}$ and quantization due to thin body and high E-field causing charge to spill into L and X valleys where μ is low. BTBT modeling considers band information, all possible transitions between bands (direct & indirect), energy quantization and quantized density of states.

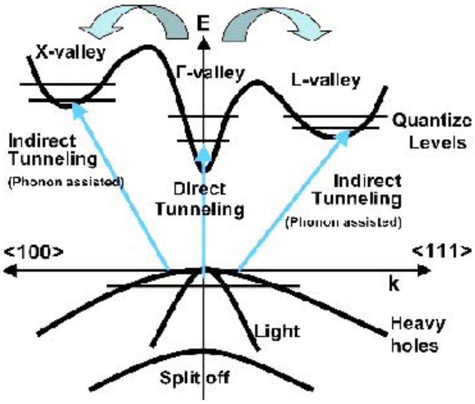
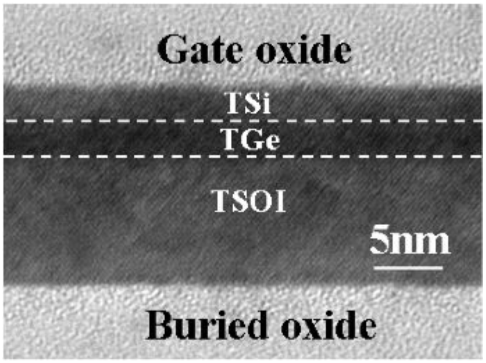


Figure 14: Cross section TEM of defect free ultra-thin strained Ge on relaxed Si on a SOI substrate.



Another problem is lower n-type dopant electrical activity[27] causing high source/drain resistance. Better surface passivation and dopant incorporation technologies need to be developed to improve n-MOS performance. Failing that III-V materials should be investigated.

Due to their small Γ-valley electron mass (m*), III-V materials like GaAs, InAs, InSb and other ternary compounds like InGaAs are being investigated as high μ channel materials for high performance NMOS. The main advantage of a semiconductor with a small m* is its high v_{inj}. However, very high μ materials like InAs and InSb have a very low density of states in the Γ-valley, which tends to greatly reduce the inversion charge for a fixed gate overdrive. At high gate fields due to quantization the energy levels in Γ-valley rise faster than L and X-valley, and the current is largely carried in these heavier mass valleys (Fig. 12), thus reducing the advantage of high μ_n. Materials like InAs and InSb also have a high dielectric constant and hence are more prone to short-channel effects (SCE).

The materials such as strained Si and strained Ge and many III-V materials, have larger carrier

Figure 13: Ultra-thin body and larger quantization increases the effective bandgap and lowers the tunneling rate.

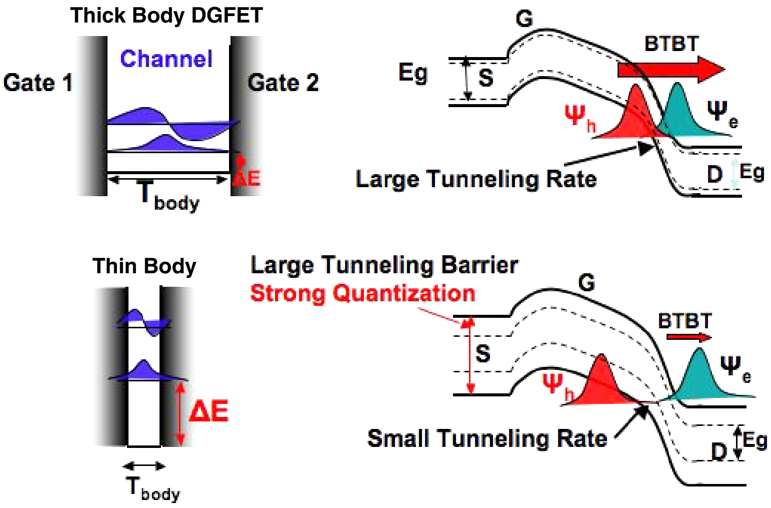


Figure 15: Experimental μ_p vs. N_{inv} for different materials and structures. H-FETs on bulk Si show a ~2X μ_p enhancement over Si, while H-FETs on SOI show even higher μ enhancements of >4X over Si.

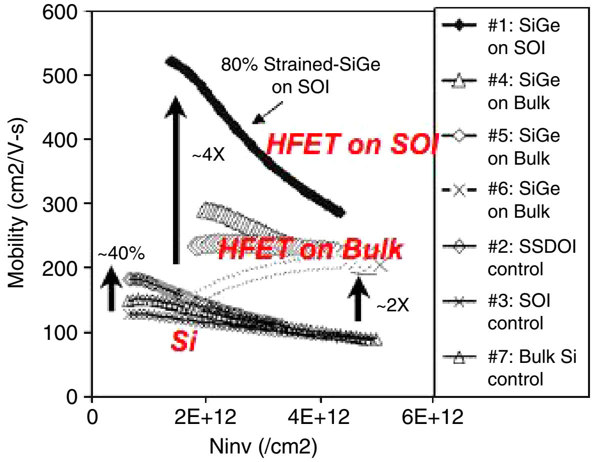


Figure 16: Experimental I_d–V_g characteristics of the s-SiGe H-FETs showing reduction in leakage and good electrostatic control. It has a degraded subthreshold slope due to the thicker gate oxide used in this study and the Ge-related defects at the interface, which can be improved.

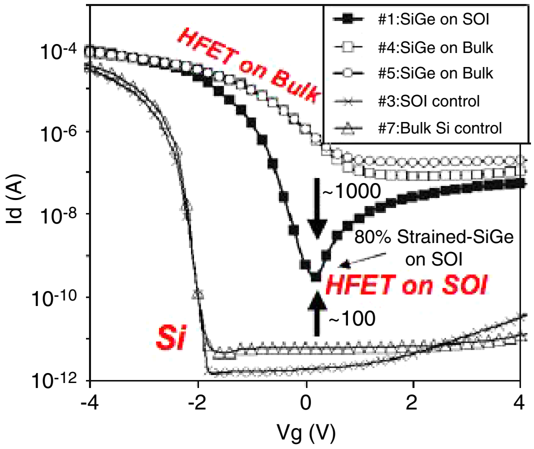
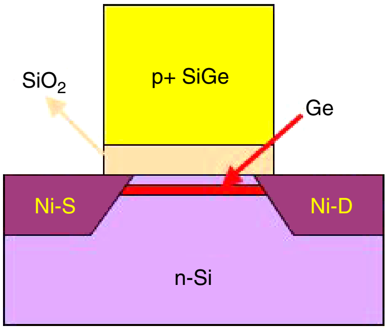


Figure 17: Structure of the Si/Ge/Si Heterostructure Channel Schottky Source/Drain PMOSFET.



mobility than silicon, but the enhanced leakage because of their smaller bandgap or direct band gap may limit their scalability. In a nanoscale transistor generally the minimum standby off-state current, $I_{\text{OFF,min}}$, is determined by the BTBT leakage, I_{BTBT}[24,25,28]. This is especially true for smaller bandgap materials like Ge and InAs. Strain in general results in reduction in the E_G and hence enhanced I_{BTBT}. Confinement on the other hand results in increased E_G and hence reduced I_{BTBT}.

In a simulation study we have investigated[29] GaAs, InAs, InSb, r-Ge and r-Si channel double gate n-MOS structure 1 of Fig. 10 with channel length of 15nm. Fig. 21 shows a comparison of I_{ON} of various channel materials. It is evident that various high μ materials perform better than Si but not too significantly. Furthermore, III-V materials appear to be similar to Ge.

In another simulation study[28] GaAs, InAs, 100% tensile strained Si, 100% compressive-s-Ge, r-Ge and r-Si I_{OFF} primarily due to BTBT leakage was investigated using the double gate n-MOS structure 1 of Fig. 10 with channel length of 15nm. Fig. 22 shows the results that the body thickness strongly affects the I_{BTBT} in these new high μ_n/small E_G materials. GaAs and Si have low I_{BTBT} due to their large E_G. InAs shows large I_{BTBT} due to its small E_G. s-Si suffers large I_{BTBT} and quantization effect doesn't help reducing it much. Although s-Ge has a smaller E_G than r-Ge, it shows a much lower I_{BTBT} than r-Ge due to its large quantization and indirect bandgap. By reducing body thickness, quantization effect can suppress I_{BTBT} making materials like InAs and InSb usable.

To take advantage of high μ_n III-V materials hetero-structure FETs where the transport occurs in a high μ material and leakage in wider E_G material should be investigated, similar to the case of Ge.

7. Beyond CMOS

While there does not appear to be a hard limit to the scaling of FETs down to sub-10-nm gate lengths (22 nm node), it is generally agreed that commensurate performance-dimension scaling will require significant innovations to enhance the transport characteristics of channel materials and to decrease the influence of parasitic components and currents[14,15]. Suppressing the short-channel

Figure 18: Hole mobility measured for a fixed $T_{capi} = 3$ nm for varying T_{Ge}. For small T_{Ge} most of the inversion charge is located in Si and hence has less μ_p. The thick Ge-layers are defective and hence have less μ_p

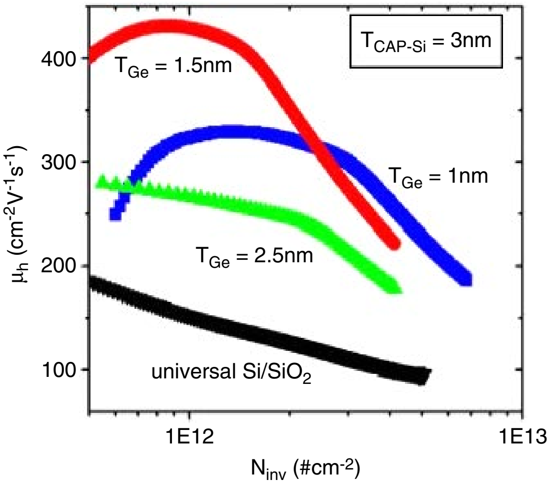
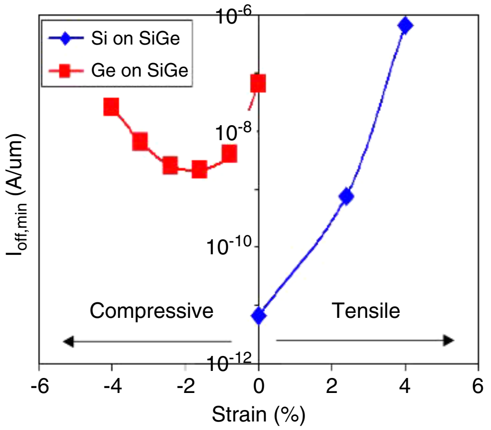


Figure 19: $I_{OFF,min}$ for various channel materials. s-Ge shows a dramatic reduction in off-state leakage compared to r-Ge because of the lower leakage from the Γ-valley.



effects and subthreshold leakage in conventional MOS devices will increasingly become problematic. To break out of these constraints, new CMOS device structures and materials will be introduced in the future as they are proven economically necessary or advantageous[31]. Heterogeneous integration of the Ge and III–V materials on Si with novel device structures and materials can take us to sub-10 nm gate lengths with commensurate improvement in performance. Beyond this we will require new devices and technology. FETs in semiconductor nanowires (SNWs), carbon nanotubes (CNTs)[31], graphene, etc. are some of the evolutionary devices, which may allow us to improve the device performance beyond the 22 nm node. CNTs and SNWs are one-dimensional structures that can be grown with uniform diameter in the nanoscale, without the use of nanometer patterning or lithography technology. In the formation of CNTs and SNWs, the diameter and in some cases the electrical length of these structures are determined during the chemical synthesis process, and not by lithography. However it is well understood by now that undesired variations can also occur during this synthesis process.

For CNTs there is the additional feature that both metallic and semiconducting tubes can be formed, depending on the tube chirality. In order to establish a usable CNT FET technology, it is necessary to achieve precise control of the tube diameter, and metal/semiconductor property, if not indeed complete control of the detailed chirality of the tubes. Also, while single CNT FETs have been demonstrated to be capable of very high current drive, more than 20 μA/tube, it is clearly necessary to have the ability to fabricate CNT transistors with a dense array of parallel tubes with a common gate, in order to emulate the capability of conventional CMOS technology to provide transistors with variable width/length (W/L) ratio.

Silicon nanowires can be synthesized with a somewhat greater degree of control compared to CNTs, and there is no complication comparable to the metallic CNTs. However, the carrier mobility in silicon nanowires is often not even up to that of bulk silicon, let alone as high as what has been observed for CNTs. Therefore the experimental effort on

Figure 20: Simulation of switching frequency vs. I_{OFF} to benchmark different materials and p-MOS device structures.

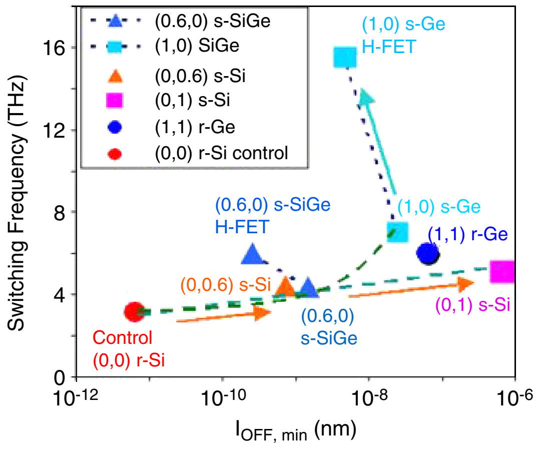


Figure 21: Simulations of drive current vs. body thickness for a double gate NMOS to benchmark the different channel materials for $L_G = 15$ nm, channel body thickness = 5 nm, $T_{ox} = 1$ nm, and $I_{OFF} = 0.1\ \mu$ A/μm.

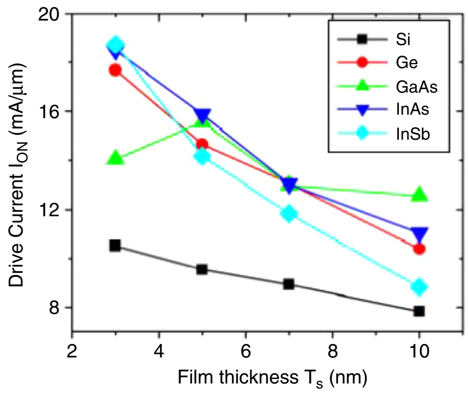
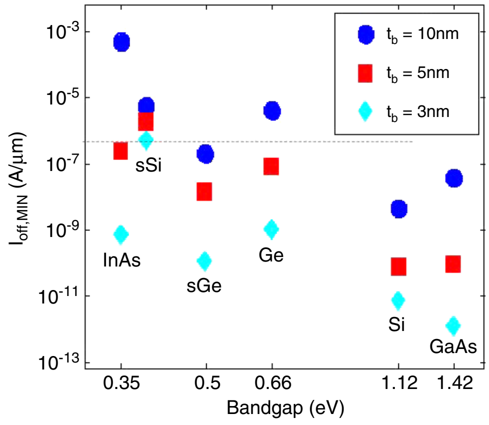


Figure 22: Simulations of I_{off} vs. bandgap for a double gate P/NMOS to benchmark the different channel materials. ($L_G = 15$ nm, $V_{DD} = 0.9$ V, $T_{ox} = 0.9$ nm)



Single electron transistor (SET): A transistor that exploits the movement of single electron through tunneling and can be used in ultra low power circuits.

Quantum dot: A nanostructure that confines the motion of carrier in all three spatial dimensions.

Resonant tunneling devices: Semiconductor devices made of quantum wells surrounded by barriers and can be used in very high speed circuits.

SNWs in future should be to improve transport properties, along with the issues of variations, placement and device design.

During the past decade, there has been extensive research on many other *seemingly useful* devices, e.g., quantum-effect spintronic devices[32], single electron transistor (SET)[33], quantum dots, resonant tunneling devices [34], nanoparticle and molecule based devices[35], and many more. Most of this research is in fabrication and physical, electronic and magnetic properties of nano- and molecular structures that has produced remarkable science, but it is not yet clear how this new science will lead to new technologies. Of these quantum-effect spintronic devices for classic logic gate implementation at vastly reduced power dissipation, in the long-term, and potentially quantum computing applications in the much longer term appear to be most promising. Development of new solid-state spintronic devices will significantly enhance system speed and reduce power consumption as compared to traditional CMOS technologies. While current technology performs memory, logic, and communications using distinct components on separate chips that communicate with each other via external inputs and outputs, this nascent degree of freedom in nanoelectronics offers the potential of integrating nominally discrete operational functions within a single device. These developments will rely on extending the untapped properties of conventional semiconductor materials as well constructing new "hybrid" ferromagnetic/semiconductor structures and novel ferromagnetic semiconductor systems.

8. Conclusion

To continue the scaling of Si CMOS in the sub-65 nm node, innovative device structures and new materials have to be created in order to continue the historic progress in information processing and transmission. Examples of novel device structures being investigated are double gate or surround gate MOS and examples of novel materials are high mobility channel materials like strained Si and Ge, III-V semiconductors, high-k gate dielectrics and metal gate electrodes. Heterogeneous integration of these materials on Si with novel device structures may take us to sub-10 nm channel length, but will require new fabrication technology solutions that are generally compatible with current and forecasted installed Si manufacturing. Beyond that we will need a set of potentially entirely different information processing and transmission devices from the *transistor as we know it*, e.g. silicon-based quantum-effect devices, nanotube electronics and molecular electronics.

Acknowledgements

This work was supported by MARCO Materials, Structure and Devices (MSD) Focus Center, Stanford University Initiative for Nanoscale Materials and Processes (INMP) and an Intel Fellowship.

Received 13 July 2007; revised 23 August 2007.

References

1. P. J. Wright and K. C. Saraswat, Thickness limitations of SiO_2 gate dielectrics for MOS ULSI, IEEE Trans. Electron Devices., vol. 37, no. 8, pp. 1884–1892, 1990.
2. G. D. Wilk, R. M. Wallace, and J. M. Anthony, High-k gate dielectrics: Current status and materials properties considerations, *J. Appl. Phys.*, vol. 89, no. 10, pp. 5243–5275, 2001.

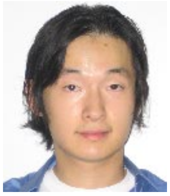
3. J. D. Plummer and P. B. Griffin, Material and process limits in silicon VLSI technology, *Proc. IEEE*, Vol. 89 , No. 3 , pp. 240–258, March 2001
4. A. Asenov, Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 μm MOSFETs: A 3D atomistic simulation study,' *IEEE Trans. Electron Devices*, vol. 45, pp. 2505–2513. 1998.
5. The International Technology Roadmap for Semiconductors, Semiconductor Industry Association. (http://public.itrs.net/)
6. C. Choi, Modeling of Nanoscale MOSFETs (Ph.D. dissertation), Stanford University, 2002.
7. D. Hisamoto, W.-C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, K. Asano, T.-J. King, J. Bokor and C. Hu, A folded channel MOSFET for deep-sub-tenth micron era, in IEDM Technical Digest, 1998, pp. 1032–1034.
8. R. Chau, B. Doyle, J. Kavelieros, D. Barlage, A. Murthy, M. Doczy, R. Arghavani and S. Datta, Advanced Depleted-Substrate Transistors: Single-Gate, Double-Gate and Tri-Gate, in Extended Abstracts of the International Conference on Solid-State Devices and Materials (SSDM), Nagoya, Japan, 2002, pp. 68–69.
9. M. Lundstrom, Elementary scattering theory of the Si MOSFET, *IEEE Electron Device Lett.*, vol. 18, pp. 361–363, 1997.
10. M. Lundstrom and Z. Ren, Essential physics of carrier transport in nanoscale MOSFETs, *IEEE Trans. Electron Devices*, vol. 49, pp. 133–141, 2002.
11. A. Lochtefeld and D. A. Antoniadis, On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit? *IEEE Electron Device Lett.*, vol. 22, pp. 95–97, 2001.
12. A. Lochtefeld and D. A. Antoniadis, Investigating the relationship between electron mobility and velocity in deeply scaled NMOS via mechanical stress, *IEEE Electron Device Lett.*, vol. 22, pp. 591–593, 2001.
13. M. Lundstrom, On the mobility versus drain current relation for a nanoscale MOSFET, *IEEE Electron Device Lett.*, vol. 22, pp. 293–295, 2001.
14. D. A. Antoniadis, I. Aberg, C. Ni, Chleirigh, O. M. Nayfeh, A. Khakifirooz and J. L. Hoyt, Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations, in *IBM Journal of Research and Development*, Vol. 50, No. 4/5, July/September 2006, pp. 363–376.
15. A. Khakifirooz and D. A. Antoniadis, Transistor Performance Scaling: The Role of Virtual Source Velocity and Its Mobility Dependence, *IEEE Int. Electron Dev. Meet. (IEDM) Tech. Digest*, p. 667, 2006
16. J. Welser, J.L. Hoyt and J. F. Gibbons, Electron mobility enhancement in strained-Si n-type metal-oxide-semiconductor field – effect transistors, in *IEEE Electron Device Letters*, Vol. 15. No. 3, March 1994, pp. 100–102.
17. S. Thompson, et al., A 90nm Logic Technology featuring 50nm Strained Silicon Channel Transistors, 7 layers of Cu Interconnects, low-k ILD, and 1 μm^2 SRAM Cell, in *IEDM Technical Digest*, 2002.
18. C. O. Chui, H. Kim, D. Chi, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat, A Sub-400^{o}C Germanium MOSFET Technology with High-k Dielectric and Metal Gate, *IEEE Int. Electron Dev. Meet. 2002 Technical Digest*, pp. 437–440, San Francisco, CA, Dec. 8–11, 2002.
19. Chi On Chui, F. Ito and K. C. Saraswat, Nanoscale Germanium MOS Dielectrics - Part I: Germanium Oxynitrides, *IEEE Trans. Electron Dev.* Vol. 53, No. 7, pp. 1501–1508, 2006.
20. A. Pethe and K. C. Saraswat, Interface state Density measurement at GeO$_x$N$_y$-Ge interface for Ge MIS Application, *IEEE SISC, Dec.* 2006.
21. Chi On Chui, H. Kim, D. Chi, Paul C. McIntyre and K. C. Saraswat, Nanoscale Germanium MOS Dielectrics - Part II: High-k Gate Dielectrics, *IEEE Trans. Electron Dev.* Vol. 53, No. 7, pp. 1509–1516, 2006.
22. K. Martens, et al., Germanium MOS Specific C-V Interpretation and Interface State Density Extraction Pitfalls and a Full Conductance Solution, *IEEE Electron Dev. Lett.*, (Submitted)
23. Tejas Krishnamohan, C. Jungemann, D. Kim, E. Ungersboeck, S. Selberherr, P. Wong, Y Nishi and K. C. Saraswat, Theoretical Investigation Of Performance In Uniaxially- and Biaxially-Strained Si, SiGe and Ge Double-Gate p-MOSFETs' IEEE Int. Electron Dev. Meet. San Francisco, Dec. 2006, pp. 937–940.
24. T. Krishnamohan, D. Kim, C. Nguyen, C. Jungemann, Y. Nish and K. C. Saraswat, High Mobility, Low Band To Band Tunneling (BTBT), Strained Germanium, Double Gate (DG), Heterostructure FETs : Simulations, *IEEE Trans. Electron Dev.*, Vol. 53, No. 5, May 2006, pp. 1000–1009. Invited
25. T. Krishnamohan, Z. Krivokapic, K. Uchida, Y. Nish and K. C. Saraswat, High Mobility, Ultra Thin (UT), Strained Ge MOSFETs On Bulk and SOI With Low Band To Band Tunneling (BTBT) Leakage : Experiments, *IEEE Trans. Electron Dev.* Vol. 53, No. 5, May 2006, pp. 990–999. Invited
26. A. Pethe and K. Saraswat, High – Mobility, Low Parasitic Resistance Si/Ge/Si Heterostructure Channel Schottky Source/Drain PMOSFETs, IEEE Device Research Conf., South Bend, Indiana, June 2007.
27. Chi On Chui, K. Gopalakrishnan, P. B. Griffin, J. D.Plummer and K. C. Saraswat, Activation and Diffusion Studies of Ion-implanted p and n Dopants in Germanium *Appl Phys. Lett.*, Vol. 83, No. 16, 20 October 2003, pp. 3275–3277.
28. D. Kim, T. Krishnamohan, Y. Nishi, K. C. Saraswat, Band to Band Tunneling limited Off state Current in Ultra-thin Body Double Gate FETs with High Mobility Materials : III-V, Ge and strained Si/Ge, *IEEE SISPAD*, Monterey, CA, Sept. 2006, pp. 389–382.
29. A. Pethe, T. Krishnamohan, D. Kim, S. Oh, H.–S.P. Wong, Y. Nishi and K. Saraswat, Investigation of the Performance Limits of III-V Double-Gate NMOSFETs, *IEEE Int. Electron Dev. Meet. (IEDM) Tech. Digest*, pp. 619–622, Washington, D.C., Dec. 2005.
30. J. A. Hutchby, G. I. Bourianoff, V. V. Zhirnov and Joe E. Brewer, Extending the Road beyond CMOS, *IEEE Circuits & Devices Magazine*, pp. 28–41, March 2002.
31. P.G. Collins and P. Avouris, Nanotubes for electronics, *Sci. Amer.*, pp. 62–69, Dec. 2000.
32. B.E. Kane, A silicon-based nuclear spin quantum computer, *Nature*, Vol. 393, pp. 133–137, 1998.
33. K. Yano, T. Ishii, T. Hashimotoi, F. Murai, and K. Seki, Room-temperature single-electron memory, *IEEE Trans. Electron Dev.*, Vol. 41, pp. 1628–1638, 1994.
34. T. Ohshima and R.A. Kiehl, Operation of bistable phase-locked single-electron tunneling logic elements, *J. Appl. Phys.*, Vol. 80, pp. 912–923, July 1996.
35. E. Emberly and G. Kirczenov, Principles for the design and operation of a molecular wire transistor, *J. Appl. Phys.*, Vol. 88 pp. 5280–5282, 2000.

Abhijit Pethe received the B. E. (hons.) degree in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 2001 and the M.S and Ph.D. degrees in electrical engineering from Stanford University, Stanford CA, in 2003 and 2007, respectively.

He is currently a Process Technology Development Engineer with the Intel Corporation, Hillsboro OR. His current research interests include high mobility materials and device technologies for low power high performance CMOS logic applications.

Prof. Krishna Saraswat is Rickey/Nielsen Professor in the School of Engineering, Professor of Electrical Engineering and Professor of Materials Science & Engineering (by courtesy) at Stanford University. He also has an honorary appointment of an Adjunct Professor at the Birla Institute of Technology and Science, Pilani, India since January 2004, and a Visiting Professorship during the summer of 2007 at IIT Bombay, India. He received his B.E. degree in Electronics in 1968 from the Birla Institute of Technology and Science, Pilani, India, and his M.S. and Ph.D. degrees in Electrical Engineering in 1969 and 1974 respectively from Stanford University, Stanford, CA. Professor Saraswat's research interests are in new and innovative materials, structures, and process technology of silicon and germanium devices and interconnects for nanoelectronics, 3-D ICs with multiple layers of heterogeneous devices, and Environmentally Benign Semiconductor Manufacturing. His past work includes modeling of CVD of silicon, conduction in polysilicon, diffusion in silicides, contact resistance, interconnect delay and 2-D oxidation effects in silicon. He pioneered the technologies for aluminum/titanium layered interconnects, CVD of tungsten silicide MOS gates, CVD tungsten MOS gates and tunable workfunction SiGe MOS gates. He developed equipment and simulators for single wafer thermal processing, deposition and etching and technology for the in-situ measurements and real-time control. Jointly with Texas Instruments a microfactory for single wafer manufacturing was demonstrated in 1993.

Prof. Saraswat has graduated more than 50 doctoral students and has authored or co-authored over 500 technical papers, of which six have received Best Paper Award. He is a Fellow of the IEEE, and a member of both The Electrochemical Society and The Materials Research Society. He received the Thomas Callinan Award from The Electrochemical Society in 2000, the Andrew Grove Award from IEEE in 2004 and the Technovisionary Award from the India Semiconductor Association in 2007. He received two gold medals for academic excellence during his education in India.



Donghyun Kim was born in South Korea. He received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 2002 and M.S. degree in electrical engineering from the Stanford University in 2004. He is currently pursuing the Ph.D. degree in electrical engineering at the Stanford University. He has worked on physics of Band-to-Band Tunneling (BTBT), ballistic transport in nano-scale devices. His interests include quantum devices, optoelectronics devices and nanophotonics.

Tejas Krishnamohan was born in Bombay, India. He received the B.S degree from the Indian Institute of Technology, Bombay, India and the M.S and Ph.D degrees in Electrical Engineering from Stanford University, Stanford, CA.

He has worked on physics and technology of high mobility channel MOSFETs, strain engineering, quantum well memory, Band-To-Band Tunneling (BTBT) and ballistic transport in nano-scale devices. His research interests are exploring novel materials, structures and physics for memory and logic devices.