INTRODUCTION

Silicon CMOS technology is pervasive in electronics and it has become crucial to support nearly all applications, from computation to communications and controls. The main reason for the success of the technology is the aggressive device scaling which has produced high performance chips nearing one billion devices with characteristic dimensions in the deca-nanometer range. This spectacular growth has continued on an exponential trend for nearly four decades, following Moore’s law, and silicon devices dominate the majority of the semiconductor market, having resisted the competition of many other technologies based on different semiconductor materials which have at best found a small specialized niche. The reasons for this success may actually appear very simple but are rooted in properties that for a long time were considered drawbacks. Silicon has an indirect bandgap, just above one volt, and a relatively large effective mass. As a result, a saturation velocity of nearly $10^7$ cm/s is established uniformly over a wide range of electric fields, velocity overshoot is small even in widely non-homogeneous structures. In addition, while size quantization effects are measurable in the channel cross-section, since a narrow well confines the electrons near the oxide interface, quantum coherence effects in the direction of propagation are not relevant and may not play a considerable role until perhaps channel lengths below 10 nm are reached. Therefore, although device scaling had to overcome increasing fabrication challenges, throughout its evolution the MOSFET has essentially behaved in a nearly classical way with very predictable performance. In reaching the nanoscale, however, the situation has started to change because intrinsic physical limitations are being approached.

NANOMETER MOSFET

A major problem in scaled MOS devices with nanoscale gate length is the difficulty in controlling leakage current between source and drain contacts when the device is OFF. In standard planar MOS structures doping profiles may be used to mitigate in part this problem, but the drawback of increased channel doping is a reduction of mobility. Another problem is the need to scale also the oxide thickness to maintain device performance, but once thicknesses below 1.5 nm are used, tunneling gate current increases beyond tolerable levels. One trend is to consider new dielectric materials with a larger dielectric constant (high-$\kappa$ dielectric), to yield similar interface field with larger thickness at the same voltage. This solution has also problems, however, since experiments show that in general the use of these dielectrics is associated with a drastic drop in mobility, related to remote optical interaction which has little effect in regular silicon dioxide.

Advanced device design has focused on highly confined channels to improve source-drain isolation. Typically, a double gate MOS structure is considered, where a narrow slab of silicon has a similar gate on the top and bottom interfaces. While this is a useful design template, a planar double gate device is of difficult realization. Practical structures utilize instead a vertical configuration, for instance the FinFET [1]. A variation uses all three top interfaces to form as many conduction channels and is called a trigate MOSFET. Such devices in general provide good source-drain isolation, but offer somewhat modest current drive. There are a number of possible reasons for this. A major structural feature of the FinFET family is that instead of an overlap between the gate and the contact regions they present an extension region (or underlap) separating physically gate and contacts. The current drive properties of the channel degrade with increasing extension length, because too much of the source-drain applied voltage drops there rather than in the channel under the gate. Figure 1 shows the results of particle Monte Carlo simulations for FinFETs with increasing extension length from 5 to 20 nm for a 20 nm gate length. The voltage applied to drain and gate is 1.0 V for this example. One can see how the potential become flat in the channel for large extension lengths, with a marked degradation of the velocity that consequently leads to poor current drive.

This work was supported by the National Science Foundation and the Semiconductor Research Corp.
The use of high-κ dielectric materials for the gate raises also some issues in the realization of a narrow conduction channel. When a particle approaches the dielectric gate structure, an image charge is formed which generates a force. In a self-consistent simulation the Poisson equation is solved and it is difficult to isolate the image force behavior to assess the impact of the dielectric choice. To investigate this issue, a particle steering model was implemented, where a particle is dragged at different channel locations. The 3D Poisson equation is solved for each position, obtaining by integration the potential energy that needs to be overcome to enter the channel. Figure 2 shows results for a square channel structure of decreasing diameter for silicon dioxide (permittivity of 3.9) and for a high-κ dielectric (permittivity of 30). One can see that the potential barrier for silicon dioxide increases considerably for narrower channels, reaching several multiples of kT. Conversely, for a high-κ dielectric with permittivity larger than that of silicon (about 11.9) the image force changes sign, resulting in an attractive rather than repulsive field that would tend to draw the particle into the channel. These results apply to an empty channel for a single particle attempting traversal, therefore they are representative of OFF or subthreshold condition and suggest how naturally the channel opposes leakage currents when few particles would enter the channel. Obviously, a high-κ dielectric material does not favor isolation with respect to regular silicon dioxide.

Size quantization in the channel is also an important issue to consider for nanoscale MOS devices. In the context of particle Monte Carlo simulation it is possible to develop a formalism with the complete detail of quantum transverse subbands in the channel. This leads, however, to a very costly simulation since a large number of intra- and inter-subband scattering mechanisms need to be considered. In the range of practical MOS devices, a quantum correction strategy has proven to be efficient, where the electrostatic potential is modified by adding an appropriate correction that imparts to the particle ensemble a space distribution equivalent to the quantum one, while leaving a semi-classical description and a simpler algorithm.
The application of quantum corrections to Monte Carlo simulation provides an additional interface force that repels the maximum of charge density away. An example is shown in Fig. 3 for a FinFET structure. Interestingly, although the charge is considerably shifted, current simulation results for the device

Fig. 2 – Potential energy for a particle entering a square silicon channel surrounded by silicon dioxide or a high-κ dielectric.

Fig. 3 – Example of quantum corrections in the crosssection of a FinFET.
considered in Fig. 1 do not deviate appreciably when fully classical or quantum corrected simulations are compared, due to an interplay of reduction of interface scattering and density variation. Quantum corrections are evaluated from solutions of Schrödinger equation in the transverse cross-section [2].

For high frequency applications, the improved isolation properties of the FinFET structure are not of primary importance. The overall capacitances of the structure are influential. One should expect the wrapping geometry of the gate to limit high frequency performance, with lower cutoff frequency than the regular planar MOSFET structure with comparable size.

The ultimate limitations of integrated technologies may actually be set by thermal considerations. ON devices with extremely short channels exhibit a quasi-ballistic transport in the channel. Nearly all the energy gained is redistributed in the drain contact region by charge-charge interaction and dissipated by phonon emission [3]. Hot spots are therefore typically found in a relatively small region of the device and the device structure needs to allow for sufficient heat dissipation away from the drain. Collectively, also the leakage in OFF devices is very important, since even the smallest leakage current multiplied by millions or billions of devices on a single chip may produce a global increase of the chip temperature that in future technologies with aggressive device packing may lead to the melting point. One should also keep in mind that a temperature increase increases the phonon statistics and leads to a degradation of mobility. Future device design should couple the electrical analysis to thermal analysis. Therefore, particle simulation remains important because it allows one to track the generation of individual phonons, the transport of which can be followed by adding an appropriate flow equation.

In conclusion, while scaling of silicon device has remains successful for several decades, the challenges related to a transition from semi-classical to quasi-ballistic/quantum regime are great, requiring an ever increasing complexity of models. For high frequency applications there remain many unknowns, because the most promising structures from a digital design point of view have certainly limitations due to large capacitances related to the gate structures.

References

