Purdue University
Purdue e-Pubs

Other Nanotechnology Publications | Birck Nanotechnology Center

4-1-2008

Outperforming the conventional scaling rules in the quantum-capacitance limit

Joachim Knoch
IBM Res GmbH, Zurich Res Lab

W Riess
IBM Res GmbH, Zurich Res Lab

Joerg Appenzeller
Birck Nanotechnology Center, Purdue University, appenzeller@purdue.edu

Follow this and additional works at: http://docs.lib.purdue.edu/nanodocs

Part of the Nanoscience and Nanotechnology Commons

Knoch, Joachim; Riess, W; and Appenzeller, Joerg, "Outperforming the conventional scaling rules in the quantum-capacitance limit" (2008). *Other Nanotechnology Publications.* Paper 172.
http://docs.lib.purdue.edu/nanodocs/172

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Outperforming the Conventional Scaling Rules in the Quantum-Capacitance Limit

J. Knoch, W. Riess, and J. Appenzeller, *Senior Member, IEEE*

***Abstract*—We present a study on the scaling behavior of field-effect transistors in the quantum-capacitance limit (QCL). It will be shown that a significant performance improvement in terms of the power delay product can be obtained in devices scaled toward the QCL. As a result, nanowires or nanotubes exhibiting a 1-D transport are a premier choice as active channel materials for transistor devices since the QCL can be attained in such systems.**

***Index Terms*—Gate delay, MOSFET, nanowire/tube, one-dimensional (1-D), quantum capacitance, scaling.**

I. Introduction

The driving force for generations of chip designs has been Denard's scaling rules of device miniaturization [1]. During device scaling, the gate capacitance (not normalized) is kept almost constant by decreasing the gate length L and the gate-oxide thickness d_{ox} simultaneously. When applying the same approach to novel nanotransistors based on, e.g., nanotubes or nanowires exhibiting a 1-D transport, the so-called quantum-capacitance limit (QCL) [2] can be reached—a regime that is not accessible in conventional 2- or 3-D FETs. The reason for this is the density of states (DOS) within the channel that increases in bulk FETs but decreases in the case of 1-D structures. In the QCL, the potential within the channel is determined by the gate potential, and as such, short channel effects are suppressed. On the other hand, in the QCL, the charge in the channel no longer increases with decreasing d_{ox} contrary to the usually encountered classical limit (CL). At a first glance, one may therefore assume that due to the absence of any d_{ox} dependence, the QCL is detrimental to the transistor performance. However, our analysis demonstrates that the opposite is indeed the case. Here, we address the question of how scaling manifests itself in the QCL using the gate delay and the power delay product as relevant figures of merit to quantify the ON-state performance of the scaled transistor devices. We will show that improving the device performance in terms of gate delay and power delay product occurs faster in 1-D transistors in the QCL than predicted according to the conventional scaling rules.

Manuscript received November 19, 2007; revised January 16, 2008. The review of this letter was arranged by Editor M. Ostling.

J. Knoch and W. Riess are with the IBM Research GmbH, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland (e-mail: jkn@zurich.ibm.com).

J. Appenzeller is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: appenzeller@purdue.edu).

Digital Object Identifier 10.1109/LED.2008.917816

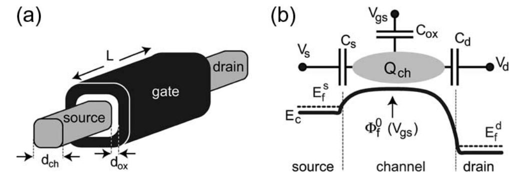

Fig. 1. (a) Schematics of the wrap-gate transistor design under consideration. (b) Surface potential along the direction of current transport. The total charge in the channel is determined by all terminal voltages.

II. Analytical Considerations

Consider the wrap-gate device geometry shown in Fig. 1(a) with a nanowire/tube of diameter d_{ch} and an oxide thickness of d_{ox}. The source/drain contacts are degenerately doped, whereas the channel of length L is considered as intrinsic. Fig. 1(b) shows the surface potential $\Phi_f(x)$ along the device together with the source, drain, and geometrical oxide capacitances. The potential maximum in the channel Φ_f^0 determines the carrier injection from the source Fermi distribution into the channel [3] and, hence, determines the current flow as well as the amount of mobile charge within the channel. Φ_f^0 can be obtained by noting that the channel charge is $Q_{\mathrm{tot}} = -\Phi_f^0/e \cdot (C_s + C_{\mathrm{ox}} + C_d)$, where the C's refer to the source, gate oxide, and drain capacitances. At the same time, $Q_{\mathrm{tot}} = C_s V_s + C_d V_d + C_{\mathrm{ox}} V_g + Q_{\mathrm{ch}}$, where Q_{ch} is the mobile charge injected by the contacts. Solving for the gate potential $\Phi_g = -eV_g$ and with the so-called quantum capacitance $C_q = e\partial Q_{\mathrm{ch}}/\partial \Phi_f^0$ [2], [4], one obtains $\delta\Phi_f^0 = C_{\mathrm{ox}}/C_\Sigma \cdot \delta\Phi_g + C_d/C_\Sigma \cdot \Phi_d$, where $C_\Sigma = C_s + C_{\mathrm{ox}} + C_d + C_q$, and $V_d = -e\Phi_d$. In the following, we only consider electrostatically well-behaved devices where $C_{\mathrm{s,d}} \ll C_{\mathrm{ox}}$, leading to $\Phi_f^0 = C_{\mathrm{ox}}/(C_{\mathrm{ox}} + C_q)\Phi_g + \Phi_{\mathrm{bi}}$ with Φ_{bi} being the built-in potential.

In order to obtain first-order expressions for the gate delay τ and the power delay product $P \cdot \tau$, we employ the Landauer approach of current transport [3]. The following approximations are made. First, since Φ_f^0 determines the injection of carriers, the transmission probability for carriers to flow from source to drain $T(E) = 0$ for energies $E < \Phi_f^0$. For $E \geq \Phi_f^0$, we approximate $T(E) = l_{\mathrm{scat}}/(l_{\mathrm{scat}} + L)$, where l_{scat} is the mean free path for scattering [5]. Although this expression is strictly valid only in the field-free case [3], it is applicable here since the fields are rather small in the case of long channel lengths. In devices with a short channel length as considered here, the fields are also small since in the QCL, the potential distribution in the channel is determined by the gate potential (rather than the channel charge) which enforces a constant potential within the channel. Second, V_{ds} is considered to be large enough so that $f(E_f^d) \approx 0$ for $E > \Phi_f^0$. With these approximations and by

0741-3106/$25.00 © 2008 IEEE

Authorized licensed use limited to: Purdue University. Downloaded on August 13, 2009 at 13:36 from IEEE Xplore. Restrictions apply.

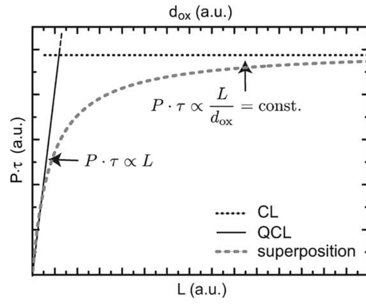

Fig. 2. Power delay product as a function of L and $d_{\rm ox}$. The horizontal black dotted line shows the CL, where $P \cdot \tau = \text{const.}$, and the straight black line belongs to the QCL with $P \cdot \tau \propto L$. The gray dashed line is the superposition of the two limiting curves showing an increasing scaling benefit in the QCL.

using the expression for Φ_f^0, the drain current can be calculated analytically to be

$$I_d \propto \frac{l_{\rm scat}}{l_{\rm scat}+L} \cdot \frac{1}{L} \cdot \frac{C_{\rm ox}C_q}{C_{\rm ox}+C_q}(V_{\rm gs}-V_{\rm th})v_{\rm inj}. \quad (1)$$

Here, $V_{\rm th}$ is the threshold voltage, and $v_{\rm inj}$ is the maximum velocity at which carriers are injected into the channel from the source contact; in addition, $C_{\rm ox} \propto L/d_{\rm ox}$, and $C_q \propto L \times 1/\sqrt{E_f^s - \Phi_f^0}$, i.e., C_q is approximately proportional to the DOS in the channel (see below). One can now distinguish between four different cases: the classical as well as the QCL, where either $C_{\rm ox}$ or C_q becomes dominant. Furthermore, in the diffusive transport regime, $l_{\rm scat}/(l_{\rm scat}+L) \approx l_{\rm scat}/L$, or in the case of ballistic transport, $l_{\rm scat}/(l_{\rm scat}+L) \approx 1$. By calculating $\tau = C_g V_{\rm dd}/I_d$ and the power delay product $P \cdot \tau$ for the different cases and transport regimes, we obtain the following results: 1) In case of diffusive transport, $\tau_{\rm diff} \propto L^2$, and for ballistic transport, $\tau_{\rm ball} \propto L$. This is true in the CL and the QCL suggesting that scaling toward the QCL does not negatively impact the device performance as measured by τ. 2) A difference between the QCL and the CL is expected for $P \cdot \tau$ since in the QCL, the total gate capacitance $C_g \approx C_q$ becomes independent of $d_{\rm ox}$ such that $P \cdot \tau$ linearly decreases when the device dimensions are shrunk. In the CL, however, $P \cdot \tau = \text{const.}$ when L and $d_{\rm ox}$ are scaled simultaneously. Fig. 2 shows $P \cdot \tau$ in the CL and the QCL. Scaling a device from the CL toward the QCL is expected to result in a smooth transition with an improving power delay product in the QCL, implying a substantial scaling benefit.

III. Device Simulations and Discussion

In order to verify the results of the analytical considerations, we performed simulations based on a self-consistent solution of the Poisson and Schrödinger equations. To keep the computational burden as small as possible, the surface potential approach of Yan *et al.* [6] and Auth and Plummer [7] is used. The following modified Poisson equation is obtained which captures all aspects related to the scaling of $d_{\rm ox}$ and the appearance of short channel effects in laterally scaled devices

$$\frac{d^2\Phi_f}{dx^2} - \frac{\Phi_f - \Phi_g + \Phi_{\rm bi}}{\lambda^2} = -\frac{e(\rho \pm N)}{\varepsilon_0 \varepsilon_{\rm ch}}. \quad (2)$$

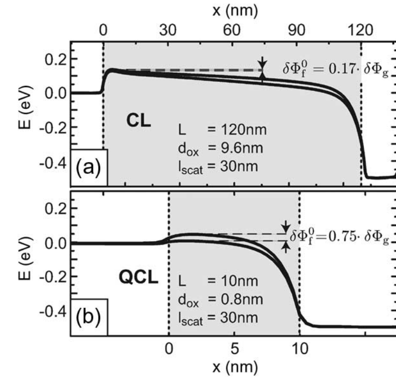

Fig. 3. Conduction band profile for (a) a device with $L = 120$ nm and $d_{\rm ox} = 9.6$ nm and (b) for a FET with $L = 10$ nm and $d_{\rm ox} = 0.8$ nm showing that (a) is rather in the CL, whereas (b) is scaled toward the QCL. Here, $V_{\rm ds} = 0.5$ V, and $V_{\rm gs} = 0.5$ and 0.55 V.

Here, $\lambda = \sqrt{\varepsilon_{\rm ch}/\varepsilon_{\rm ox} \cdot d_{\rm ox}/4 \cdot d_{\rm ch}}$ is the relevant length scale of potential variations, and Φ_g and $\Phi_{\rm bi}$ are the gate and built-in potentials.[1] To calculate the charge in and current through the device, we use the nonequilibrium Green's function formalism [5]. An effective mass approximation is employed, and scattering is taken into account via Buettiker contacts [8].

Simulations of devices were carried out with the following parameters: $E_f^{\rm s,d} = 0.1$ eV, $m^\star = 0.1 m_0$, and $d_{\rm ch} = 3$ nm; a midgap work-function metal was assumed as gate electrode, and $\Phi_{\rm bi} = 0.4$ eV. Furthermore, $\varepsilon_{\rm ox} = 3.9$, and $\varepsilon_{\rm ch} = 12$. Transfer characteristics of several devices were simulated starting with a rather long channel length of $L = 140$ nm and an oxide thickness of $d_{\rm ox} = 11.2$ nm. The channel length L and the gate-oxide thickness $d_{\rm ox}$ are then scaled down simultaneously to $L = 10$ nm and $d_{\rm ox} = 0.8$ nm, whereas the supply voltage was kept constant at a value of $V_{\rm dd} = 0.5$ V. The inset of Fig. 4(a) shows the I_d–$V_{\rm gs}$ curves for three different devices with $L = 10$, 40, and 140 nm.

From $\delta\Phi_f^0 = C_{\rm ox}/C_\Sigma \cdot \delta\Phi_g + C_d/C_\Sigma \cdot \Phi_d$ (see Section II), it is apparent that in the device's ON-state, $\partial\Phi_f^0/\partial\Phi_g$ approaches zero in the CL and unity in the QCL. Fig. 3(a) shows the conduction band for two different $V_{\rm gs}$'s in a device with $L = 120$ nm and $d_{\rm ox} = 9.6$ nm. It is apparent that Φ_f^0 hardly changes with $V_{\rm gs}$ (i.e., $\delta\Phi_f^0/\delta\Phi_g$ becomes small) showing that the device in Fig. 3(a) is rather in the CL. On the other hand, in the case of $L = 10$ nm and $d_{\rm ox} = 0.8$ nm [Fig. 3(b)], the bands are moved much more efficiently for the same $\delta\Phi_g$. The large band movement is a result of the scaling of $d_{\rm ox}$ in relation to the DOS in the channel, leading to $C_{\rm ox} > C_q$ which shows that the device is scaled toward the QCL. Note that in a bulk FET, this would not happen since the large DOS always leads to $C_q \gg C_{\rm ox}$ even for the thinnest $d_{\rm ox}$ considered here. Consequently, the QCL is (for reasonable $d_{\rm ox}$) a unique feature of 1-D transistor structures.

When scaling the devices from the CL toward the QCL, the gate capacitance $C_g = C_{\rm ox}C_q/(C_{\rm ox}+C_q)$ becomes increasingly dependent on $V_{\rm gs}$. This is due to the fact that $C_q \propto L \times (\partial/\partial\Phi_f^0) \int dE (1/\sqrt{E - \Phi_f^0}) f_s(E - E_f^s) = L \times \int dE$

[1]The expression for λ used in the present analysis is an approximation for a surround-gate FET (see also [7]).

Authorized licensed use limited to: Purdue University. Downloaded on August 13, 2009 at 13:36 from IEEE Xplore. Restrictions apply.

$(1/\sqrt{E})\ (\partial/\partial\Phi_f^0)\ f_s\ (E+\Phi_f^0-E_f^s) \propto L \times 1/\sqrt{E_f^s-\Phi_f^0}$, where Φ_f^0, in turn, depends on $V_{\rm gs}$. Note that the thermal broadening of the Fermi function prevents C_q from diverging once the conduction band in the channel aligns with the Fermi level of the source contact. Here, the maximum value $C_q^{\rm max} \propto 1/\sqrt{4k_BT}$ which is roughly a factor of five smaller than $C_{\rm ox}$ for $d_{\rm ox} = 0.8$ nm. With the chosen parameters, $C_q^{\rm max}$ lies within the chosen range of $V_{\rm dd}$ such that the QCL can actually be reached for the smallest oxide thicknesses. Since $P \cdot \tau$ represents the energy needed for switching a device, it was calculated according to $\int QdV_{\rm gs}$, where Q is the total channel charge extracted from the simulations. In the same fashion, the gate delay is obtained as $\tau = \int QdV_{\rm gs}/(V_{\rm dd}I_d)$. The main panel of Fig. 4(a) shows the extracted τ values as a function of L and $d_{\rm ox}$ in the case of (hollow circles) scattering and (solid squares) ballistic transport. The dashed line in Fig. 4(a) is a second-order polynomial fit, and the black line is a linear fit showing that the gate delay exactly shows the dependence on L expected from the analytical considerations aforementioned. This is true over the entire range, i.e., from the CL toward the QCL, as has been pointed out previously. As a result, continued scaling yields improved gate delays irrespective of the scaling regime, i.e., CL or QCL. On the contrary, in terms of power delay product, the analytical considerations predict a performance benefit in the QCL. Fig. 4(b) shows the simulated data exhibiting the same behavior in both cases: the scattering and the ballistic transport. To increase the channel length and the gate-oxide thickness, the device tends toward the CL, and the power delay product approaches a constant value. However, toward the QCL, the devices exhibit a significantly decreasing power delay product, as was predicted previously. In the present case, an $\sim$40% lower $P \cdot \tau$ compared with the CL is observed for a technologically feasible $d_{\rm ox} = 1.6$ nm.[2]

The important implication of the present analysis is the following. Nanowires/tubes with a very small diameter enable ultimately scaled transistor devices in a wrap-gate architecture since electrostatic integrity is preserved down to the smallest dimensions. However, besides this pure geometrical argument, this letter shows that the nanowires/tubes offer an additional scaling benefit. In the case of a 1-D transport, devices can be scaled toward the QCL which shows a clear scaling advantage in terms of the power delay product, i.e., the energy needed for switching the transistors. In practical cases, where parasitic source/drain capacitances $C_{\rm par}$ exist, this scaling advantage is somewhat diminished. Since $P \cdot \tau$ is proportional to the capacitance of the FET (if $V_{\rm dd}$ is constant) and by taking $C_{\rm par}$ into account, one obtains $(P\cdot\tau)_{\rm par}^{\rm QCL}/(P\cdot\tau)_{\rm par}^{\rm CL} = ((1 + C_{\rm QCL}/C_{\rm par})/(1+C_{\rm ox}/C_{\rm par}))$, where $Q_{\rm QCL} = \int QdV_{\rm gs}/V_{\rm dd}^2$ is an average quantum capacitance. This ratio is always smaller than one, which means that if $C_{\rm par}$ is not too large compared with $C_{\rm ox}$, a significant scaling benefit is obtained in practical cases as well. As a result, nanowires/tubes exhibiting 1-D transport are a premier choice as the channel materials for high-performance ultimately scaled FET devices.

[2]The reason for the lower $P \cdot \tau$ in case of $d_{\rm ox} = 0.8$ nm and scattering is that the conduction band in the channel is moved substantially below the (quasi)-Fermi level in source due to the tight gate control. Therefore, scattering in the source extension leads to a somewhat lower carrier density and, thus, lower capacitance within the channel compared with the ballistic case.

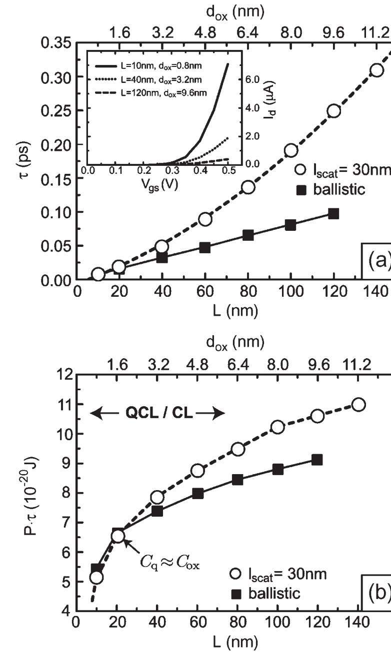

Fig. 4. (a) Gate delay extracted from simulations versus L, $d_{\rm ox}$ for (hollow circles) scattering and (solid black squares) ballistic transport. The dashed black line is a quadratic fit to the data points. In the case of ballistic transport, the data points lie on a straight line. (b) Simulated $P \cdot \tau$ as a function of L and $d_{\rm ox}$. The curves exhibit the same behavior as shown in Fig. 2, implying a significant scaling benefit in the QCL.

IV. Conclusion

We studied the performance of the nanotube/nanowire FETs scaled into the QCL. Continued scaling leads to improved τ independent of whether the device is in the CL or the QCL. However, in approaching the QCL, a FET yields a significantly improved performance in terms of $P \cdot \tau$, which means that 1-D devices offer a significant performance advantage when compared with their 2- or 3-D counterparts.

References

[1] R. H. Denard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid State Circuits*, vol. SSC-9, no. 4, pp. 256–268, Apr. 1974.

[2] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1853–1864, Sep. 2003.

[3] M. S. Lundstrom and J.-H. Rhew, "A Landauer approach to nanoscale MOSFETs," *J. Comput. Electron.*, vol. 1, no. 4, pp. 481–489, Dec. 2002.

[4] S. Luryi, "Quantum capacitance devices," *Appl. Phys. Lett.*, vol. 52, no. 6, pp. 501–503, Feb. 1988.

[5] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[6] R.-H. Yan, A. Ourmazd, and F. Lee, "Scaling the Si MOSFET: From bulk to SOI to bulk," *IEEE Trans. Electron Devices*, vol. 39, no. 7, pp. 1704–1710, Jul. 1992.

[7] C. P. Auth and J. D. Plummer, "Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFETs," *IEEE Electron Device Lett.*, vol. 18, no. 2, pp. 74–76, Feb. 1997.

[8] R. Venugopal, M. Paullson, S. Goasguen, S. Datta, and M. S. Lundstrom, "A simple quantum mechanical treatment of scattering in nanoscale transistors," *J. Appl. Phys.*, vol. 93, no. 9, pp. 5613–5625, May 2003.

Authorized licensed use limited to: Purdue University. Downloaded on August 13, 2009 at 13:36 from IEEE Xplore. Restrictions apply.