2014-11

Negative Differential Transconductance in Silicon Quantum Well Metal-Oxide-Semiconductor Field Effect/Bipolar Hybrid Transistors

UTD AUTHOR(S): Clint Naquin and Mark Lee

Naquin, C., M. Lee, H. Edwards, G. Mathur, et al. 2014. "Negative differential transconductance in silicon quantum well metal-oxide-semiconductor field effect/bipolar hybrid transistors." Applied Physics Letters 105: 213507-1 to -4.

Negative differential transconductance in silicon quantum well metal-oxide-semiconductor field effect/bipolar hybrid transistors

Clint Naquin,[1] Mark Lee,[1] Hal Edwards,[2] Guru Mathur,[2] Tathagata Chatterjee,[2] and Ken Maggio[2]
[1]*Department of Physics, University of Texas at Dallas, Richardson, Texas 75080, USA*
[2]*Texas Instruments, Inc., Richardson, Texas 75243, USA*

(Received 28 August 2014; accepted 16 November 2014; published online 25 November 2014)

Introducing explicit quantum transport into Si transistors in a manner amenable to industrial fabrication has proven challenging. Hybrid field-effect/bipolar Si transistors fabricated on an industrial 45 nm process line are shown to demonstrate explicit quantum transport signatures. These transistors incorporate a lateral ion implantation-defined quantum well (QW) whose potential depth is controlled by a gate voltage (V_G). Quantum transport in the form of negative differential transconductance (NDTC) is observed to temperatures >200 K. The NDTC is tied to a non-monotonic dependence of bipolar current gain on V_G that reduces drain-source current through the QW. These devices establish the feasibility of exploiting quantum transport to transform the performance horizons of Si devices fabricated in an industrially scalable manner. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4902919]

The 2011 International Technology Roadmap for Semiconductors (ITRS) challenges the silicon semiconductor industry with aggressive targets for high-frequency and low-noise device performance by the 2026 end-of-roadmap.[1] In addition, the ITRS requires that devices meeting these goals must be fabricated within the progression of industrial standard process nodes (45 nm, 28 nm, 22 nm, etc.) to guarantee economically scalable manufacturing. Devices fabricated using one-off, low yield, or high marginal cost methods are not adequate solutions. Even considering cutting-edge industrial Si device architectures, such as SiGe Bi-CMOS ("Bi" = bipolar, "C" = complementary, "MOS" = metal-oxide-semiconductor),[2] ultra-thin body (UTB),[3] and multi-gate (MG)[4] field effect transistors (FETs), the ITRS admits that it is unknown whether the evolution of any current technology will achieve its end-of-roadmap targets.[1]

One assumption of the ITRS is that Si CMOS transistors will continue to operate using the semi-classical device physics principles established in the 1940s.[5] "Semi-classical" means that charge transport in response to built-in and applied electric fields is governed by the Boltzmann equation description of drift and diffusion. Quantum mechanics is only implicit–hidden in the band structure and Fermi energy. However, it has been recognized for over forty years[6] that transport through explicitly quantum states (i.e., states of discrete energy, conductance, charge, etc.) can yield transformative capabilities beyond semi-classical device physics. Examples include maximum oscillation frequencies >1 THz,[7,8] ultra-low noise performance,[9] and fundamentally quantum limited sensitivity.[10]

Quantum transport is manifested when a potential profile confines electrons to a length scale comparable to their de Broglie wavelength, λ. Most well developed techniques for building such small potential profiles use some combination of the following: (1) vertical transport across thin epitaxial layers of heterogeneous band gap materials, (2) lateral transport through structures defined by serial electron-beam lithography, or (3) various one-off fabrication methods. Much success has been achieved in explicitly quantum semiconductor devices, particularly using III-V materials at cryogenic temperatures. However, the processing and materials used are considered too slow, too high in marginal cost, or materially incompatible for integration with industrial Si CMOS processing. For this reason Si CMOS has not moved beyond semi-classical physics.

Beginning with the industrial "45 nm" process node, the lateral dimensions of mass-produced Si CMOS transistors are approaching the size required for quantum transport at or near room temperature. This potentially opens a quantum approach towards achieving ITRS performance goals in industrial Si CMOS. In a Si MOSFET channel the longest electron λ is ultimately limited by the mean distance between momentum-randomizing scattering, i.e., the elastic mean free path ℓ.[11] Thus the longest $\lambda \sim \ell = v_F\tau = v_F m^*\mu/e$, where m^* is the effective mass, v_F is the Fermi velocity, τ is the scattering lifetime, μ is the mobility, and e is the charge. For a Si MOSFET at 300 K with $\mu \approx 300\,\mathrm{cm^2/V{\cdot}s}$ (Ref. 12) and $10^{18}\,\mathrm{cm^{-3}}$ carrier density, $\lambda \sim \ell \approx 10$ to 50 nm. This sets the length scale for the potential confinement profile needed to produce explicitly quantum transport.

Existing literature contains several demonstrations of Si quantum devices. Examples include quantum point contacts,[13,14] a quantum dot single electron detector,[15] low $1/f$ noise single-electron transistors[16] and memory,[17] and electron turnstiles.[18,19] These reports show that Si devices confining electrons within a 50–100 nm dimension demonstrate quantum transport at $T \leq 100$ K. Devices with 10–30 nm dimension have shown quantum operation near 300 K, consistent with the simple physical estimate above. These works provide proof-of-principle that quantum operation in Si can be achieved to 300 K at <50 nm length scale. However, the fabrication methods used are unsuitable to industrial scale Si CMOS processing. Thus these demonstrations do not provide adequate solutions to the ITRS challenges.

Here, we describe the fabrication and basic operating characteristics of a class of explicitly quantum Si transistors fabricated by industrially standard processing. Unlike almost all existing quantum semiconductor devices, here a potential profile small enough to define an electron quantum well (QW) in a two-dimensional MOSFET interface is built by lateral ion implantation doping. The potential depth of the QW in the channel between source (S) and drain (D) can then be modulated by a gate voltage V_G. Definitive evidence that a lateral QW can be constructed in this manner is shown by observation of quantum transport signatures in the form of negative differential transconductances (NDTCs) (*i.e.*, $g_m = \partial I_{DS}/\partial V_G < 0$) at temperatures >200 K. These NDTCs are similar to what occurs in III-V resonant tunneling transistors,[20] where vertical QWs are defined by heteroepitaxial growth. Our Si QW transistors were fabricated by Texas Instruments (TI) on a commercial process line. Consequently, they are fully compatible with standard industrial manufacturing and are scalable to smaller generations of processing nodes.

Both conventional (no QW) and QW *n*-channel MOS (NMOS) transistors were fabricated simultaneously as neighboring devices on the same wafer die. All transistors underwent the same photomasking steps in the same standard low-voltage gate oxide "45 nm" generation CMOS process flow. This process can produce gate lengths of 45, 40, and 35 nm. To maximize chance of success in constructing a QW, all devices studied here had the smallest possible mask-defined gate length of 35 nm (and 1 μm width). All devices used polysilicon gate electrodes on a SiO_2 gate dielectric (2 nm equivalent oxide thickness). Si_3N_4 sidewall spacers 40 nm thick were deposited on the gate stack walls to ensure gate electrical isolation. The *p*-type body was doped with acceptor implants to control threshold voltage, and a separate ohmic body contact was made. Arsenic and phosphorus implants ($\sim 6 \times 10^{20}\,cm^{-3}$) formed n^+ source/drain (SD) contacts. Following standard short channel NMOS design principles,[21] lightly doped drain (LDD) implants were made to form the shallowest portion of the SD contact regions (see Fig. 1).

The *sole* difference between conventional and QW NMOS is the LDD implant dopant polarity. For conventional NMOS, *n*-type LDD implants were used. For QW NMOS, *p*-type LDD implants ($\sim 2 \times 10^{20}\,cm^{-3}$ boron) were used. Because both NMOS and *p*-channel MOS devices are normally fabricated on the same wafer die, both *n*- and *p*-type implants were available in the process line. Fabrication of QW and conventional NMOS transistors was differentiated only through the layout of the LDD implant mask, where some conventional *n*-LDD mask openings were replaced with *p*-LDD openings. There were no interruptions or additions to the normal process flow. Both *n*- and *p*-LDD implants were activated using the same standard rapid thermal anneal step. Current-voltage measurements of the body-to-drain/source diodes for both conventional and QW devices showed diode behavior with identical ideality factor of 1.2. This indicates that the recombination rate due to traps and defects in the body-to-drain/source depletion regions is unaffected by the polarity of LDD dopant implants used.

Measurements of basic transistor characteristics were performed in a probe station where device temperature could be varied from 5 K to 300 K with the test die in vacuum. In all measurements S was held at ground. Measurements of I_{DS} vs. V_G at various drain voltages V_D and either body current (I_B) or body voltage (V_B) bias were done using an Agilent 4156 C semiconductor parameter analyzer. Measurement protocol dictated monitoring currents independently in all four probe leads (S, D, G, B) to ensure that changes in I_{DS} were due to current between drain and source, rather than body and gate. With V_G above threshold, $I_S = -I_D$ to a high degree of precision with I_B and I_G constituting < 1% of I_D. Thus there were no significant spurious source-drain current paths. In all data presented total current in all leads summed accurately to zero, indicating no current leakage paths.

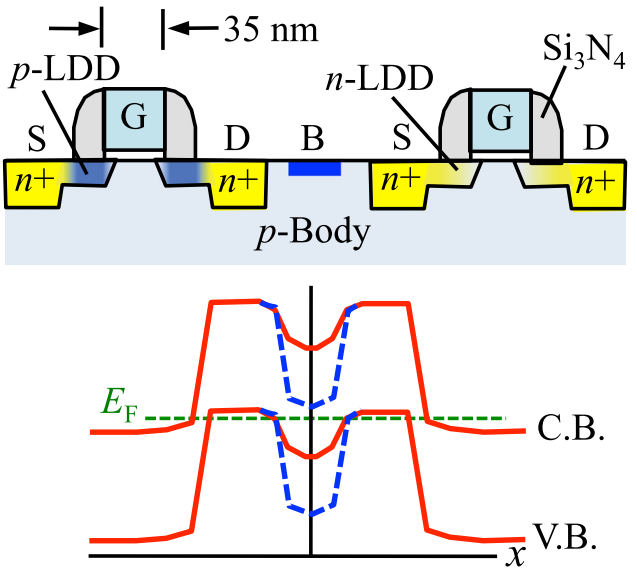

FIG. 1. (Top) Schematic cross-section (not to scale) of two NMOS FETs showing source (S), gate (G), drain (D), and body (B) contacts. Conventional (right) and quantum well (QW) (left) differ only in LDD dopant polarity: *n*-type for conventional and *p*-type for QW. Both types may share a body contact. (Bottom) Equilibrium conduction band (C.B.) and valence band (V.B.) potential profiles and Fermi level (E_F) of the interface channel beneath a QW NMOS gate with $V_G = 0$ (red solid) and $V_G > 0$ (blue dashed), depicting formation of an *n*-channel QW (based on TCAD process simulations).

Evidence of quantum transport through a lateral QW is given by observation of NDTC regimes, i.e., discrete intervals of V_G for which $g_m < 0$. Fig. 2(a) shows how a series of three NDTCs at 10 K develops with body current bias I_B in a typical 35 nm gate length QW NMOS. The body bias drives the threshold gate voltage negative. For comparison, the inset of Fig. 2(b) shows I_{DS}–V_G measured at 10 K on a 35 nm conventional (*n*-LDD) NMOS. Here, I_{DS} is a monotonic function of V_G so $g_m > 0$ at all V_G. No NDTC was observed in any conventional NMOS at any temperature and bias used. By contrast, twenty-five QW NMOS were measured and all showed a series of NDTCs for $0\,V < V_G < 1.5\,V$ similar to Fig. 2(a). The NDTCs were observed when: (1) temperature was $\leq$220 K and (2) body current $I_B > 0$ and voltage $V_B \geq 0.5\,V$ using either body current bias or voltage bias. Specifically, no NDTC was observed when $I_B = 0$ (open body). Whether the body was voltage biased or current biased at $I_B > 0$ (usually between 10 to 100 pA), NDTCs occurred only when the body potential reached $V_B \geq 0.5\,V$.

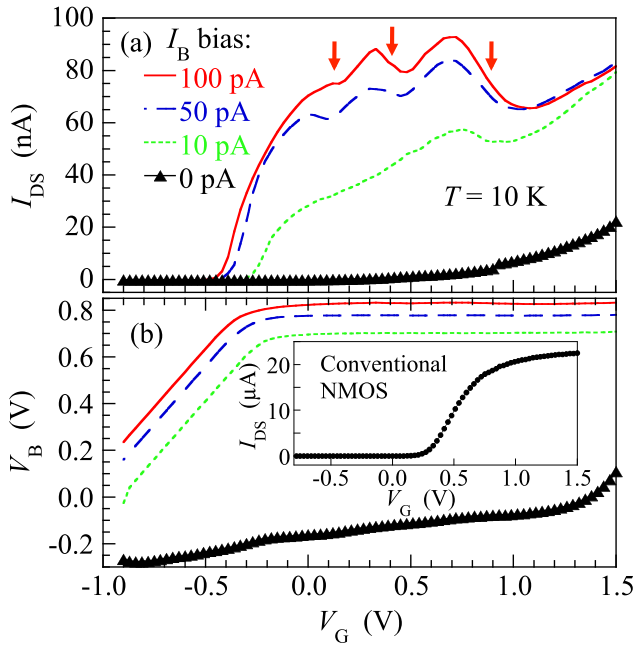

FIG. 2. (a) Evolution of NDTC peaks in I_{DS} vs. V_G on a typical 35 nm QW NMOS for body bias currents of 0 pA (solid black line/triangles), 10 pA (short dashed green line), 50 pA (long dashed blue line) and 100 pA (solid red line) at 10 K and $V_{DS} = 30$ mV. Red arrows indicate regions of NDTC where $g_m = \partial I_{DS}/\partial V_G < 0$. (b) Body voltage developed in response to each body current bias shown in (a) and V_G. Inset: I_{DS} vs. V_G on a 35 nm conventional NMOS at 10 K, where $g_m > 0$ for all V_G.

The NDTC dependence on body bias is summarized in Fig. 2 which presents (a) I_{DS}–V_G data on a QW NMOS at 10 K plotted for body current biases of $I_B = 0$, 10, 50, and 100 pA, and (b) the body voltages V_B developed in response to these I_B biases. For $I_B = 0$ pA no NDTC appears in I_{DS}–V_G, with V_B rising with V_G but not saturating. For $I_B = 10$, 50, and 100 pA, NDTCs in I_{DS}–V_G appear only when V_B saturates at ≥ 0.5 V. The strength of the largest NDTCs, as measured by the current peak-to-valley ratio (PVR, the ratio of I_{DS} values at local maximum to local minimum of the NDTC) also tend to increase with I_B and V_B.

It is clear that body bias has a large influence on I_{DS} and the NDTC. In fact, requiring $I_B > 0$ and $V_B \geq 0.5$ V to observe NDTC is essentially the condition for turning on an *npn* bipolar transistor formed by the body as base and source/drain as emitter/collector. The bipolar nature of this operation is demonstrated in the Gummel plot of Fig. 3. Here, I_{DS} (analog of bipolar I_{CE}) and I_B are shown as a function of applied V_B, in this case taken at several constant V_G values between 0 and 1.26 V. It is clear that the body bias significantly modulates I_{DS}. The modulating current I_B is orders-of-magnitude smaller than I_{DS}, so the effect of V_B and I_B on I_{DS} can be described as bipolar transistor control. This also means that additive drain-body diode or leakage current cannot account for the large current swings generating the NDTCs.

Consequently, we define a bipolar transistor current gain $\beta = I_{DS}/I_B$ for these devices. Fig. 4 shows both β and I_{DS} vs. V_G on a QW NMOS at 50 K. In this measurement the body (base) was voltage biased at $V_B = 0.56$ V, sufficient to turn the effective bipolar *npn* "on." Here, I_B is not constant but varies with V_G to maintain the constant V_B. The resulting β can be $>10^3$ when V_G is positive. Moreover, there are clear

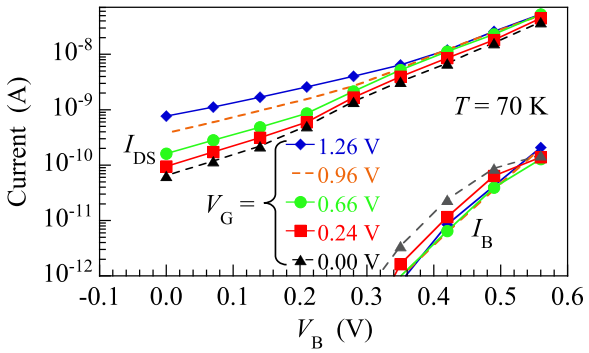

FIG. 3. Gummel plot of I_{DS} and I_B vs. body voltage bias V_B at fixed $V_G = 0.00$ V (dashed black line/triangles), 0.24 V (solid red line/squares), 0.66 V (solid green line/circles), 0.96 V (dashed orange line), and 1.26 V (solid blue line/diamonds) on a QW NMOS at 70 K with fixed $V_{DS} = 30$ mV.

intervals of V_G where $\partial\beta/\partial V_G < 0$. (Conventional NMOS showed only $\partial\beta/\partial V_G > 0$.) These negative $\partial\beta/\partial V_G$ regions correspond closely to the NDTCs that appear in the simultaneous measurement of I_{DS}–V_G. Therefore, the NDTC behavior appears to arise from a non-monotonic dependence of effective bipolar gain on the 2D QW interface channel conduction controlled by V_G.

These empirical observations strongly suggest that the NDTC results from a combination of both MOSFET and bipolar effects. The NDTC is explicitly controlled by V_G and occurs only when V_G is sufficiently high (at given V_B). This means that the current responsible for NDTC must flow through an inverted 2D interface channel under the gate oxide between source and drain. Simultaneously, since NDTC only occurs when $I_B > 0$ *and* $V_B \geq 0.5$ V, the effective *npn* bipolar formed by source (emitter), body (base), and drain (collector) must be "on" as well.

A physical explanation of the NDTC mechanism consistent with the experimental phenomenology treats the device as an *npn* bipolar where the *p*-body (base) bias controls I_{DS} through both bulk and interface channels. When an interface channel QW is present, 2D transport through the QW between source and drain is allowed only when V_G aligns QW states connecting source and drain. At those values of V_G, I_{DS} can transit via the interface channel. Oxide and interface traps at the Si/SiO_2 interface are known to greatly reduce the carrier recombination lifetime τ^* in the interface

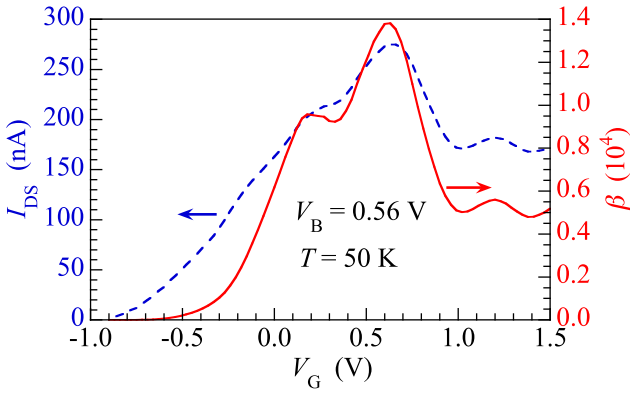

FIG. 4. Plot of both I_{DS} (blue dashed line, left axis) and effective bipolar current gain $\beta = I_{DS}/I_B$ (red solid line, right axis) vs. V_G on one QW NMOS with fixed body voltage $V_B = 0.56$ V and $V_{DS} = 30$ mV.

channel.[22] In standard bipolar device theory,[23] a reduction in τ^* directly decreases β. Consequently, as V_G is increased, whenever a quantized interface channel connects source and drain, τ^* decreases and β falls, resulting in negative $\partial\beta/\partial V_G$ and NDTC from the decreasing bipolar gain.

Multiple NDTCs, as in Fig. 2, then represent formation of consecutive quantized interface channels as V_G modulates the QW depth. The typical spacing, ΔV_G, between NDTCs in a 35 nm gate length device is $\sim$300 mV. With this ΔV_G, a simple model of the QW as a finite square potential well estimates, using a standard calculation,[24] the effective QW length to be $L_{eff} \approx 10$ to 15 nm. This is significantly shorter than the nominal 35 nm gate length. TCAD simulations suggest that a smaller L_{eff} results from diffusion of p-LDD acceptor implants into the channel and formation of a lateral depletion region between the p-LDD barriers and an inverted n-channel. Quantum transport simulations using more realistic potential profiles will be carried out to refine these details.

In summary, we have observed signatures of explicit quantum transport, in the form of NDTCs, in a Si QW transistor where a lateral QW is generated by industrially standard ion implantation doping. The phenomenology of QW NMOS operation suggests that the NDTCs result from a decrease in an effective bipolar current gain when a quantized state connects source to drain in the QW interface channel. These devices were fabricated using an industrially standard 45 nm technology process and so are fully compatible with the requirements of industrial scale production.

The authors thank S. Loveless, C. Taylor, and E. Pryor (TI) for measurement support; K. Benaissa and G. Baldwin (TI) for fabrication support; K. K. O, M. Fischetti, and S. Shichijo (UTD) for useful conversations about device physics; and A. Bowling, B. Doering, R. Wise, M. Denissen, G. Frantz, and B. Haroun (TI) for making critical resources available. Work at UTD was supported by the Semiconductor Research Corporation through the Texas Analog Center of Excellence Task 1836.145 and by the National Science Foundation under grant ECCS-1403421.

[1]See http://www.itrs.net/Links/2011ITRS/2011Chapters/2011RFAMS.pdf and http://www.itrs.net/Links/2011ITRS/2011Tables/RFAMS_2011Tables.xlxs for explicit ITRS device performance targets and commentary.
[2]U. R. Pfeiffer, E. Ojefors, A. Lisauskas, and H. G. Roskos, Bipolar/BiCMOS Circuits and Technology Meeting, BCTM 2008 (2008), p. 149.
[3]Y.-K. Choi, K. Asano, N. LIndert, V. Subramanian, T. J. King, J. Bokor, and C. Hu, IEEE Electron Device Lett. **21**, 254 (2000).
[4]I. Ferain, C. A. Conline, and J.-P. Colinge, Nature **479**, 310 (2011).
[5]S. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. (Wiley-Interscience, New York, 2006).
[6]*Nanostructure Physics and Fabrication*, edited by M. A. Reed and W. P. Kirk(Academic Press Inc., New York, 1989).
[7]T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, Appl. Phys. Lett. **43**, 588 (1983).
[8]R. Köhler, A. Tredicucci, F. Beltram, H. E. Beere, E. H. Linfield, A. G. Davies, D. A. Ritchie, R. C. Iotti, and F. Rossi, Nature **417**, 156 (2002).
[9]R. J. Schoelkopf, P. Wahlgren, A. A. Kozhevnikov, P. Delsing, and D. E. Prober, Science **280**, 1238 (1998).
[10]O. Astafiev, S. Komiyama, T. Kutsuwa, V. Antonov, Y. Kawaguchi, and K. Hirakawa, Appl. Phys. Lett. **80**, 4250 (2002).
[11]M. Lee, L. N. Pfeiffer, and K. W. West, Appl. Phys. Lett. **81**, 1243 (2002).
[12]F. Lime, K. Oshima, M. Cassé, G. Ghibaudo, S. Cristoloveanu, B. Guillaumot, and H. Iwai, Solid State Electron. **47**, 1617 (2003).
[13]J. H. F. Scott-Thomas, S. B. Field, M. A. Kastner, H. I. Smith, and D. A. Antoniadis, Phys. Rev. Lett. **62**, 583 (1989).
[14]M. Prunnila, S. Eranen, J. Ahopelto, A. Manninen, M. Kamp, M. Emmerling, A. Forchel, A. Kristensen, B. S. Sorensen, P. E. Lindelof, and A. Gustafsson, Mater. Sci. Eng. B **74**, 193 (2000).
[15]L. P. Rokhinson, L. J. Guo, S. Y. Chou, and D. C. Tsui, Phys. Rev. B **63**, 035321 (2001).
[16]N. M. Zimmerman, W. H. Huber, A. Fujiwara, and Y. Takahashi, Appl. Phys. Lett. **79**, 3188 (2001).
[17]K. Nishiguchi, H. Inokawa, Y. Ono, A. Fujiwara, and Y. Takahashi, Appl. Phys. Lett. **85**, 1277 (2004).
[18]K. Nishiguchi, A. Fujiwara, Y. Ono, H. Inokawa, and Y. Takahashi, Appl. Phys. Lett. **88**, 183101 (2006).
[19]M. Pierre, B. Roche, X. Jehl, R. Wacquez, M. Sanquer, M. Vinet, N. Feltin, and L. Devoille, in *Proceedings of 2010 Conf. Precision Electromagnetic Measurements*, Daejeon, Korea (2010), p. 755.
[20]F. Beltram, F. Capasso, S. Luryi, S. N. G. Chu, A. Y. Cho, and D. L. Sivco, Appl. Phys. Lett. **53**, 219 (1988).
[21]R. Jacob Baker, *CMOS: Circuit Design, Layout, and Simulation*, 3rd ed. (Wiley–IEEE Press, New York, 2011), chap. 6.
[22]A. Neugroschel, C.-T. Sah, K. M. Han, M. S. Carroll, T. Nishida, J. T. Kavalieros, and Y. Lu, IEEE Trans. Electron Dev. **42**, 1657 (1995).
[23]P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 5th ed. (John Wiley & Sons, Hoboken NJ, 2009), pp. 8–14.
[24]E. Merzbacher, *Quantum Mechanics*, 2nd ed. (John Wiley & Sons, New York, 1970), pp. 105–114.