Can 2D-Nanocrystals Extend the Lifetime of Floating-Gate Transistor Based Nonvolatile Memory?

Wei Cao, *Student Member, IEEE*, Jiahao Kang, *Student Member, IEEE*, Simone Bertolazzi, Andras Kis, *Member, IEEE*, and Kaustav Banerjee, *Fellow, IEEE*

***Abstract*— Conventional floating-gate (FG) transistors (made with Si/poly-Si) that form the building blocks of the widely employed nonvolatile flash memory technology face severe scaling challenges beyond the 12-nm node. In this paper, for the first time, a comprehensive evaluation of the FG transistor made from emerging nanocrystals in the form of 2-dimensional (2D) transition metal dichalcogenides (TMDs) and multilayer graphene (MLG) is presented. It is shown that TMD based 2D channel materials have excellent gate length scaling potential due to their atomic scale thicknesses. On the other hand, employing MLG as FG greatly reduces cell-to-cell interference and alleviates reliability concerns. Moreover, it is also revealed that TMD/MLG heterostructures enable new mechanism for improving charge retention, thereby allowing the effective oxide thickness of gate dielectrics to be scaled to a few nanometers. Thus, this work indicates that judiciously selected 2D-nanocrystals can significantly extend the lifetime of the FG-based memory cell.**

***Index Terms*— 2D materials, CMOS scaling, dichalcogenide, floating-gate transistor, graphene, graphene/TMD heterostructures, memory, MoS_2, $MoSe_2$, NAND flash, transition metal, WS_2, WSe_2.**

I. Introduction

FLOATING-gate (FG) transistors have been playing a vital role in the pervasive flash memory technology, because their nonvolatile nature is demanded by applications in the rapidly growing arena of portable electronics including digital cameras, mobile phones, and universal serial bus drives. Because of their field-effect transistor (FET)-based structures, FG transistors share nearly all the scaling related issues with nanoscale FETs. However, a key difference is that the gate dielectrics including the tunnel oxide (TOX) and control oxide (COX) (Fig. 1) in FG transistors should be much thicker than the gate dielectrics in conventional FETs to prevent any leakage of the stored charge. Therefore, reducing the effective oxide thickness (EOT), which is the most powerful driver in FET scaling, imposes a limit in the case of FG transistor scaling [1]. Another issue hindering FG transistor scaling is cell-to-cell interference (CTCI) [2] caused by the capacitive coupling between adjacent FGs, as shown by the schematically drawn capacitor between two FG transistors in Fig. 1. These two issues represent the major electrical constraints for the scaling of FG transistors apart from lithography-induced limitations. Table I summarizes the degree of CTCI and the latest International Technology Roadmap for Semiconductors (ITRS) projections as well as recently achieved progress in industry on gate-dielectric scaling at different technology nodes [3]–[5]. It can be observed that beyond the 12-nm node, both control of CTCI and gate-dielectric scaling face severe challenges. For better immunity to CTCI, the thickness of FGs that determines the coupling capacitance (schematically shown in Fig. 1) should be scaled to a few nanometers for future nodes. Unfortunately, the commonly used poly-Si FGs in such conditions suffer from ballistic current (during programming) that degrades the reliability of memory cells [6], as well as reduced density of states (DOSs) and thus lower

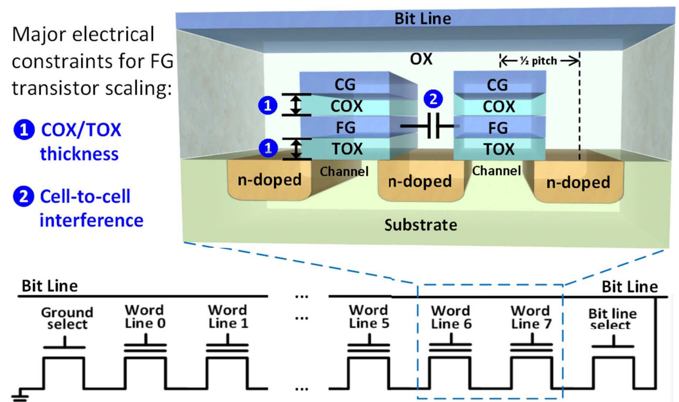


Fig. 1. Schematic of an 8-bit NAND flash string (bottom) and cross-sectional view (top) of two adjacent FG transistors along the bit line direction. Two major electrical constraints for FG transistor scaling: 1) the thicknesses of COX and TOX, which should be sufficient for 10-year retention time and 2) CTCI, which is caused by the coupling between adjacent FGs. OX and CG stand for oxide and control gate, respectively.

Manuscript received November 4, 2013; revised January 18, 2014; accepted August 14, 2014. Date of current version September 18, 2014. This work was supported in part by the U.S. National Science Foundation under Grant CCF-1162633. The review of this paper was arranged by Editor Y.-H. Shih.

W. Cao, J. Kang, and K. Banerjee are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: weicao@ece.ucsb.edu; jiahao_kang@ece.ucsb.edu; kaustav@ece.ucsb.edu).

S. Bertolazzi and A. Kis are with the Department of Electrical Engineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: simone.bertolazzi@epfl.ch; andras.kis@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2014.2350483

0018-9383 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

TABLE I
ITRS PROJECTIONS AND RECENT INDUSTRIAL PROGRESS ON THE GATE-DIELECTRIC SCALING OF PLANAR FG TRANSISTORS AND THE DEGREE OF CTCI [3]–[5]. ONO: OXIDE–NITRIDE–OXIDE

½ pitch (nm)	32	22	21 - 16	12	8
Thickness of TOX (nm)	6-7 SiO_2	6-7 SiO_2	5-7 SiO_2	5-6 SiO_2	5-6 SiO_2
EOT of COX (nm)	10-13 ONO	10-13 ONO	8-13 ONO/High-K	9 High-K	8 High-K
Cell-to-cell interference	Under control	Under control	Under control	severe	severe

Legend: white — Solution exists; orange — Solution known; red — Solution not known

charge-storage capability due to energy quantization along the thickness direction. Metal FGs have been proposed to avoid such problems [6]. However, the diffusive metal atoms in metal FGs contaminate the TOX and cause reduction of retention time [4], [5]. Recently, multilayer graphene (MLG) was shown to be an effective FG material for storing large amounts of charge and for suppressing CTCI [7]. In addition, this sp^2 bonded strong 2D carbon material can sustain very high current densities [8] and its impermeable nature prevents any metal atoms from contaminating the TOX [9]. Moreover, the relatively large interlayer resistance [10] can help suppress ballistic current during programming. For FET and FG transistor scaling, reducing channel thickness is another key component besides reducing EOT. In this regard, Bertolazzi *et al.* [11] recently demonstrated a FG transistor with monolayer molybdenum disulfide [MoS_2, a member of the transition metal dichalcogenide (TMD) family] as channel material and MLG-based FG. From a manufacturability point of view, the planar structure of both MLG and TMD materials provides compatibility with the main-stream planar CMOS platform [12]. It is also worthwhile to note that 2D material-based FG transistor is not mutually exclusive with respect to a 3-D IC architecture [13], [14] (as recently proposed in [15]), i.e., based on it, 3-D NAND flash [16] can be made even more compact and energy efficient (due to lower leakage). Although this new device seems promising, its advantages in terms of overcoming the obstacles in FG transistor scaling beyond 12-nm node remain unclear. Only a comprehensive scaling analysis can reveal its true potential and may help determine the optimal device geometry and material choices. In this paper, gate length, cell-to-cell distance, and gate-dielectric thickness scaling are studied for the first time for a MoS_2/MLG FG transistor based on various simulation techniques. Multilayer graphene nanoribbon (GNR) is proposed as FG to improve the device immunity to CTCI. Subsequently, investigation is extended to other three monolayer TMD materials, WS_2, $MoSe_2$, and WSe_2. It is found that these three materials (in comparison with MoS_2) can provide new mechanism for improving charge retention characteristics and new impetus to gate-dielectric scaling if proper gate-dielectric stack is chosen. A deliberately designed FG transistor based on WSe_2/MLG and high-K gate-dielectric is proposed for sub-10-nm nodes that can significantly extend the lifetime of the FG transistor-based flash memory cell.

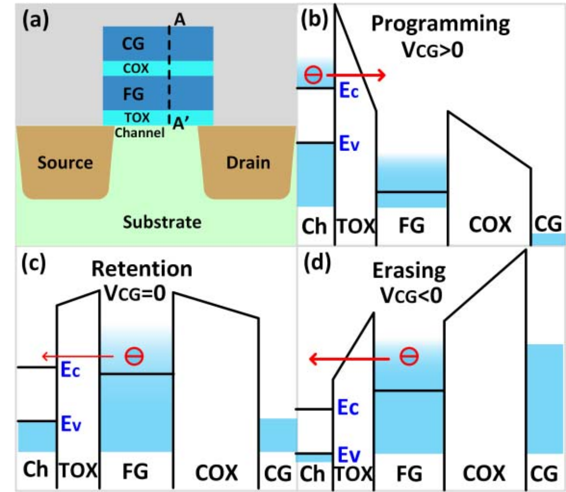


Fig. 2. (a) Cross section of a typical FG transistor. Band diagrams and electron transport directions along the AA′ dashed line in (a) for (b) programming, (c) retention, and (d) erasing states. CG, Ch, E_c, and E_v represent control gate, channel, conduction band minima, and valence band maxima, respectively.

II. FUNDAMENTALS OF FG TRANSISTOR

Fig. 2(a) shows the cross section of a typical FG transistor along the channel direction. Figs. 2(b)–(d) show the band diagrams along the dashed line AA′ in Fig. 2(a) for programming, retention, and erasing states, respectively. During programming, electrons tunnel from the channel into the FG. During erasing, electrons tunnel back. Both programming (P) and erasing (E) operate mainly through Fowler–Nordheim tunneling (FNT). In the retention state, electrons leak into the channel, mainly through localized trap-assisted tunneling, which is also called stress-induced-leakage-current (SILC) [1].

The P/E and leakage current can be expressed as [17]

$$I = \frac{2q}{h} \sum_{k_\perp} \int T(E)[f_{\mathrm{Ch}}(E) - f_{\mathrm{FG}}(E)]dE \qquad (1)$$

where q is the elementary charge, h is Planck's constant, $k_\perp$ is wave vector perpendicular to the tunnel direction, $T(E)$ is the tunneling probability, and $f_{\mathrm{Ch/FG}}(E)$ is the Fermi–Dirac function in the channel/FG region. As a general mechanism in FG transistors, the large tunneling barrier asymmetry between FNT during P/E and direct tunneling (DT) during retention introduces large tunneling probability difference exceeding 10 orders of magnitude, and thus guarantees a fast P/E and 10-year retention time simultaneously.

TOX is designed as the conduction path for P/E in FG transistor. Therefore, applied CG voltage is expected to primarily drop across TOX to ensure high gate efficiency. Gate coupling ratio (GCR) is defined to characterize the gate efficiency of a FG transistor [1]

$$\mathrm{GCR} = \frac{V_{\mathrm{TOX}}}{V_{\mathrm{CG}}} \approx \frac{C_{\mathrm{COX}}}{C_{\mathrm{COX}} + C_{\mathrm{TOX}}} = 1 \bigg/ \left(1 + \frac{T_{\mathrm{COX}}}{T_{\mathrm{TOX}}} \frac{\varepsilon_{\mathrm{TOX}}}{\varepsilon_{\mathrm{COX}}}\right) \qquad (2)$$

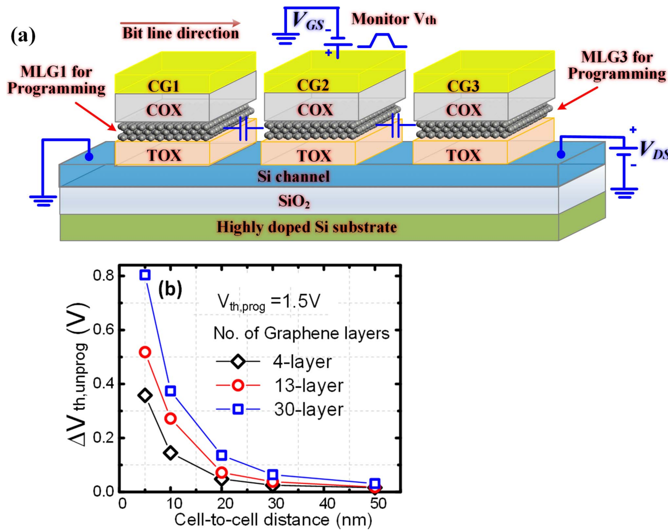


Fig. 3. (a) 3-D view of three MLG-based FG transistors in series along the bit line, which is used to study CTCI in a NAND flash. The worst condition is reached when both the left and right devices are programmed to a threshold voltage $V_{\mathrm{th,prog}}$, and the central device remains unprogrammed with $V_{\mathrm{th,unprog}} = 0$ V. (b) Threshold voltage variation ($\Delta V_{\mathrm{th,unprog}}$) of the unprogrammed central device versus cell-to-cell distance for 4-, 13-, and 30-layer MLG, respectively.

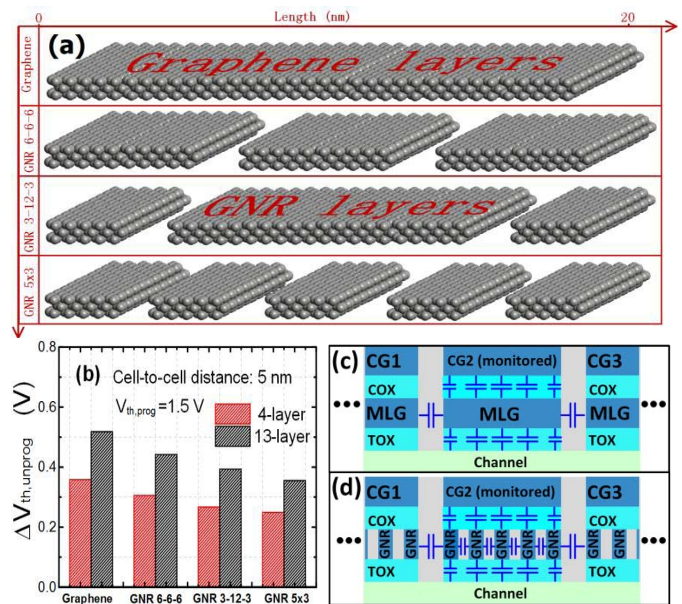


Fig. 4. (a) 3-D schematic view of four types of MLG/GNR FGs. The first (top most) is a whole piece of MLG with a width of 20 nm. This MLG is then divided following three strategies: 6 nm each for three (GNR 6–6–6), 12 nm for one and 3 nm for two (GNR 3–12–3), and 3 nm each for five (GNR 5 × 3). (b) Threshold voltage variations of the chosen four strategies with 5-nm cell-to-cell distance are studied for both 4- and 13-layer cases. Capacitance distributions in (c) MLG-based FGs and (d) GNR-based FGs for understanding the physics behind the CTCI reduction using GNRs as FG.

where V, C, T, and ε denote voltage, capacitance, thickness, and dielectric constant, respectively. This equation does not consider the nonideal depletion effect [1] of poly-Si-based FG. According to the roadmap [3], SiO_2-based TOX has been scaled approaching its lower limit of 5 nm for 10-year retention time. Reducing T_{COX}, as indicated by (2), is a method to increase GCR. However, thin COX may lead to P/E saturation due to charge leakage from FG to CG, which should be suppressed in FG transistor design. Current technology adopts wrap-around structure [1], [3] to increase C_{COX}, thus boosting GCR, while keeping COX thick, eventually leading to a compromised GCR of 0.5–0.6. In this paper, TMD, MLG, and high-K dielectric are investigated as channel, FG, and TOX (COX) of FG transistor, respectively, in comparison with the conventional Si/poly-Si/SiO_2.

III. Graphene as FG Material

For mass storage applications, FG transistors must be packed densely in memory arrays. The interference between adjacent cells, i.e., CTCI, becomes a nonnegligible factor in memory design beyond the 22-nm node. MLG, as a metallic and atomically thin film, has the potential to suppress CTCI if it is used as FGs. In this regard, Hong *et al.* [7] has preliminarily shown that MLG is much better than poly-Si as FGs. However, further exploration of the full advantage of using graphene as FGs is still in demand. Fig. 3(a) shows a schematic of three MLG-based FG transistors in series along the bit line in a NAND flash. Considering the worst condition when two neighboring devices are programmed, threshold voltage shift of the unprogrammed device in the center versus cell-to-cell distance is simulated for 4-, 13-, and 30-layer MLG, as shown in Fig. 3(b). 11.7-nm Al_2O_3 and 5-nm SiO_2 are chosen as COX and TOX, respectively. Gate length for each device is set to be 20 nm. The simulation is carried out with Atlas device simulator [18]. The MLG FG is set as equipotential region in the simulation, which has reasonable accuracy, since the much higher conductivity (along both in-plane and out-of-plane directions) of MLG compared with TOX and COX ensures that any applied bias essentially drops across the TOX and COX. Moreover, the screening length of MLG, i.e., the electric field penetration depth, has been demonstrated to be around 0.6 nm [10], which is close to that of a metal. Threshold voltage is extracted using constant current method [19]. 30-layer MLG has a thickness of $\sim$10 nm, which is at the level of, but still smaller than the thickness of normally used poly-Si FG. As shown in Fig. 3(b): 1) threshold voltage variation increases drastically with scaled cell-to-cell distance and 2) thinner MLG exhibits smaller threshold voltage variation, i.e., smaller CTCI. For $V_{\mathrm{th,prog}}$ of 1.5 V, considering 0.2 V to be the maximum acceptable $\Delta V_{\mathrm{th,unprog}}$, MLG with less than 13 layers is preferred for beyond 12-nm node applications, for which cell-to-cell distance is expected to be also sub-12 nm.

To further suppress CTCI, multilayer discrete GNRs are proposed as FGs. This is to employ the advantage of charge trap memory (CTM) since CTM has better immunity against CTCI and higher reliability than FG transistor due to its discrete charge storage nodes that results in screening effect [20]. Compared with Si-nitride-based CTMs, multilayer GNRs have large DOS, which provides much larger charge storage capacity compared with Si-nitride CTMs [4], [5]. Fig. 4(a) shows four strategies of MLG/GNR distribution for repeating the simulation introduced in Fig. 3. The 20-nm wide graphene

layers (the other dimension is assumed to be infinite) in the first strategy is set as the reference case (graphene). It is then divided into three 6-nm wide GNR layers as the second strategy (GNR 6–6–6), two 3-nm and one 12-nm wide GNR layers as the third strategy (GNR 3–12–3), and five 3-nm wide GNR layers as the forth strategy (GNR 5 × 3). Each strategy has a total width (including space between GNRs) of 20 nm, which corresponds to the chosen gate length of each FG transistor shown in Fig. 3(a). Cell-to-cell distance is set to be 5 nm for both 4- and 13-layer MLG/GNR cases. It is worthwhile to note that for sub-10-nm wide GNRs in FG application, zigzag edge is preferred, since tight-binding calculation indicates that when the width of GNRs becomes less than 10 nm, zigzag-edged GNRs remain metallic or semimetallic, while armchair edged GNRs have large bandgaps, which make them semiconducting [21], [22]. Simulation results are shown in Fig. 4(b). It can be observed that: 1) the distribution of GNRs should be as discrete as possible, which is indicated by the $\Delta V_{\text{th,unprog}}$ reduction from graphene to GNR 6–6–6, then to GNR 5 × 3 and 2) narrower GNRs should be placed at the edges of the FG region, which is concluded from the comparison between GNR 6–6–6 and GNR 3–12–3. The physics of CTCI reduction by discrete GNRs can be explained as follows. Compared with the conductive MLG, the laterally placed discrete GNRs are isolated and only capacitively coupled, which lead to a potential gradient from the center to the edges of the FG of the monitored (center) cell when adjacent cells are programmed, as shown by the capacitor network in Figs. 4(c) and (d). On the other hand, in FETs or FG transistors, the highest channel potential energy that determines the device threshold voltage is generally near the central part of the channel, i.e., right below the central part of the FG. The stability of the potential of GNRs in the central part of FG (due to lower interference from the neighboring FGs) helps stabilize the highest potential energy and thus the threshold voltage in FG transistors, thereby leading to reduced CTCI. Moreover, lateral charge sharing, which is a problem in metal nanocrystal FG [5], is unlikely to occur in GNR FG, since no contaminative atoms can be introduced to form percolation paths between GNRs (due to strong in-plane sp^2 bonds in graphene). In fact, two graphene layers separated by a 3-nm thick dielectric layer have been experimentally demonstrated to be electrically isolated [23].

IV. 2D Semiconductors as Channel Material

Gate length scaling of FETs/FG transistors which are elementary components in IC product can render higher performance and greater compactness (leading to more functionalities and lower cost), and thus usually acts as a technology driver. According to device physics [19], this lateral scaling is generally limited by vertical scaling, i.e., the scaling of gate-dielectric thickness and channel thickness, to suppress short-channel effects (SCEs). Aggressive gate-dielectric scaling risks introducing gate leakage for FETs and degrading charge retention time for FG transistors. Channel thickness scaling is a much safer option. However, scaling of conventional 3D materials, such as Si, Ge, and III–Vs, generally leads to severe quantum confinement (leading to uncontrollable bandgap variation) and carrier mobility reduction [19]. The family of 2D semiconducting crystals, such as monolayer TMD materials featuring atomic-scale thicknesses and pristine surfaces (without dangling bonds), is emerging as a promising solution for beyond-Si electronics [25]–[29]. In this section, monolayer MoS_2 as the most well-known member in the TMD family is studied as channel material in FG transistors. The contact between metal and TMDs in this paper is assumed to be ohmic, which can be realized by proper contact engineering [30], [31]. Threshold voltage roll off and subthreshold swing (SS) degradation are two indicators of SCE during gate length scaling. These two effects are studied in FG transistors for three types of channel material/device structures: 5-nm Si/SOI structure, monolayer MoS_2/SOI structure, and monolayer MoS_2/double-gate (DG) structure (realized by connecting gate and substrate electrodes together), as schematically shown in Figs. 5(a) and (b). The simulation is carried out by an in-house nonequilibrium Green's function (NEGF) device simulator, in which 2-D Poisson's equation and 1-D NEGF transport equation are solved self-consistently. For Si SOI device, three lowest sub-bands induced by quantum confinement in the channel are considered in the mode-space-approach-based NEGF transport equation [32]. To simplify the simulation, a lumped gate dielectric with 6-nm EOT is used to replace the COX/FG/TOX stack. Source/drain contact type is assumed to be ohmic, and meanwhile contact resistance is not considered in the simulation. It is observed in Figs. 5(c) and (d) that the characteristics of Si/SOI device deviate severely with respect to that of long channel case when the gate length shrinks below 40 nm. In contrast, MoS_2-based devices, especially the DG structure, show much better immunity against SCE for sub-22-nm nodes. This advantage

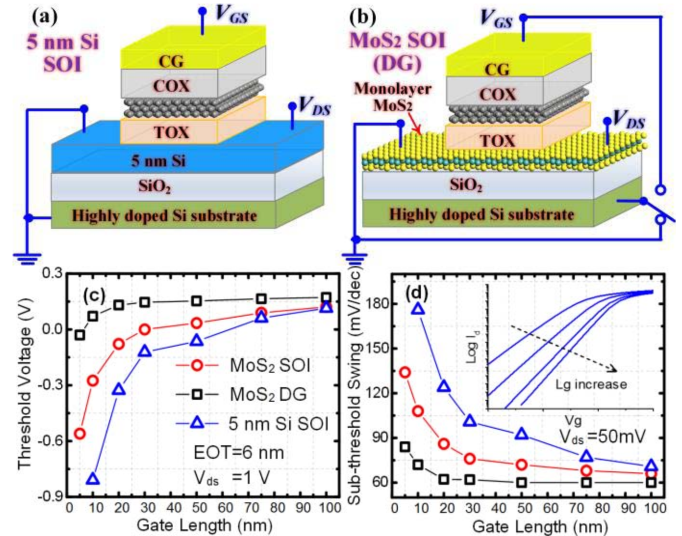


Fig. 5. Three types of channel material/device structures. (a) 5-nm Si/SOI structure, (b) MoS_2/SOI (DG) structure. The two structures in (b) can be interchanged by switching the substrate terminal to ground or to gate electrode. (c) Threshold voltage variation with gate length scaling (known as V_{th} roll-off) for those three devices. (d) SS degradation with gate length scaling. Inset: semilog I–V curves of MoS_2 SOI device. EOT of TOX/COX stack is set to be 6 nm. 5-nm underlap between source/drain and gate is used to improve device electrostatics [24] when device gate length is smaller than 10 nm.

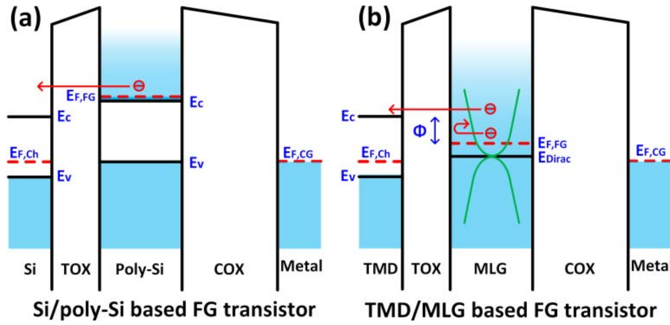


Fig. 6. Band diagrams of (a) Conventional Si/poly-Si and (b) TMD-/MLG-based FG transistor in the retention state. Green lines in (b) represent the band structure of MLG near the Dirac point E_{Dirac}. E_c and E_v represent conduction band minima and valence band maxima, respectively. Φ is the additional barrier height for charge leakage. $E_{F,\text{Ch/FG/CG}}$ is the Fermi level in channel/FG/CG region.

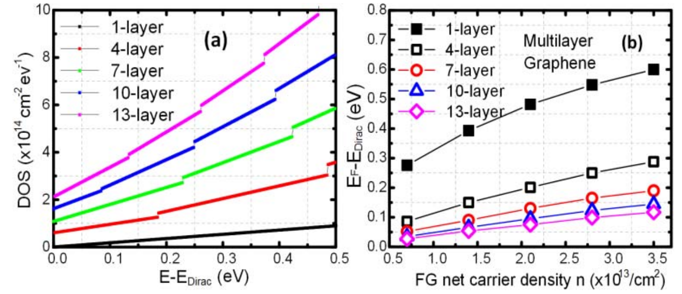


Fig. 7. (a) DOS of MLG for 1–13 layers versus the energy relative to the Dirac point. (b) Elevation of E_F relative to the Dirac point versus the stored net carrier density.

stems from the atomic-scale thickness of monolayer MoS_2 channel that offers excellent gate controllability over the electrostatic potential in the channel region.

V. Beyond Direct-Tunneling and FNT

Although using 2D semiconductors enables the possibility of scaling FG transistors to sub-10-nm nodes, due to increasing fringing effect from source/drain to channel, devices will not be able to perform as well as their long channel counterparts without accordingly scaled gate dielectric, as indicated by the simulation results in Fig. 5. However, the tunneling-based retention mechanism of FG transistors puts a lower limit on the thickness of TOX, and thus on COX, if sufficient gate efficiency or GCR is required. Therefore, a new retention mechanism is desired to extend this limit. Band offset, as a feature of heterostructures [33], can be exploited in MLG/TMD based FG transistor application. Figs. 6(a) and (b) shows schematic band diagrams of conventional Si/poly-Si and TMD/MLG based FG transistors in the retention state. According to the electrostatics of FG transistor, the Coulomb potential energy of negatively charged FG is higher than that of channel in the retention state. In Si/poly-Si based FG transistor, the affinities of Si and poly-Si are the same, and thus, E_c of poly-Si FG must be higher than that of Si channel in the retention state, as shown in Fig. 6(a). Hence, all charges stored in the FG of Si/poly-Si device could leak out via DT. In contrast, the affinity of MLG ($\sim$4.6 eV) in TMD/MLG device is larger than that of TMD materials, thus offering a band offset between E_c of the TMD channel and the Dirac point (E_{Dirac}) of MLG, such that no states are available at the TMD side for tunneling. If the Fermi level ($E_{F,\text{FG}}$) elevation with respect to E_{Dirac} and the Coulomb potential energy increase in MLG are properly controlled, $E_{F,\text{FG}}$ in MLG could be lower than E_c of the TMD channel, thereby only a small amount of stored electrons with energy higher than E_c of the channel material can leak out. Equivalently, this band offset potentially offers an additional barrier for the leaking charge. The energy difference (Φ) between E_c of TMD and $E_{F,\text{FG}}$ in MLG can be seen as the effective barrier height. It is worthwhile to note that stored electrons in the FG with energy lower than the E_c of the TMD channel in the retention state have no way to leak out through the TOX, since no extra energy can be gained by these electrons to transit to E_c of the TMD channel, even with the aid of traps inside the TOX [34]. Therefore, this new retention mechanism greatly reduces the heavy dependence of retention on the quality of the TOX. To achieve a large Φ, many layers of MLG are needed to provide sufficiently high DOS to move $E_{F,\text{FG}}$ as close to E_{Dirac} as possible for fixed charge storage. Meanwhile, FG capacitance should be large to suppress the Coulomb potential energy increase.

Fig. 7(a) shows the DOS of 1–13 layers MLG as a function of the energy level with respect to E_{Dirac}, which is extracted from the band structure of MLG [35]. The discontinuities correspond to the appearance of new sub-bands. It can be observed that DOS becomes larger with more layers of graphene and increases with energy. Considering the Fermi–Dirac distribution of both electrons and holes, $E_{F,\text{FG}}$ elevation with stored net carrier density (n) is calculated and shown in Fig. 7(b). For single layer graphene, $E_F - E_{\text{Dirac}}$ of 0.32 eV is needed to store carrier density of 1×10^{13}/cm^2. In comparison, 4-layer MLG needs 0.12 eV, and 13-layer MLG only needs 0.03 eV, i.e., charge storage capacity becomes stronger with more layers. To make $E_F - E_{\text{Dirac}}$ lower than 0.1 eV for storing carrier density of 1×10^{13}/cm^2, MLG with more than seven layers is preferred. The comprehensive consideration toward suppressing CTCI and enhancing charge storage capacity generate an optimal range of 7–13 for the number of graphene layers.

A deliberate design of TOX is necessary to ensure high FG capacitance and thus small Coulomb potential energy increase of FG. The choice of COX should follow the requirements of decent gate efficiency and negligible leakage through COX, hence does not have much flexibility. A 9-nm HfO_2/12-nm HfO_2 stack is proposed as TOX/COX in this paper. Although HfO_2 is known to suffer from high trap density and thus SILC, when integrated between Si and poly-Si, it may not have such problem in the enclosure of 2D TMD and MLG that blocks any contaminations [9]. In fact, a MoS_2/MLG FG transistor with HfO_2 as TOX has been experimentally demonstrated to have good endurance and 10-year retention time [11]. Conventional SiO_2 TOX with small FG capacitance (5-nm SiO_2 as TOX, 9.4-nm Al_2O_3 as COX) is chosen as a

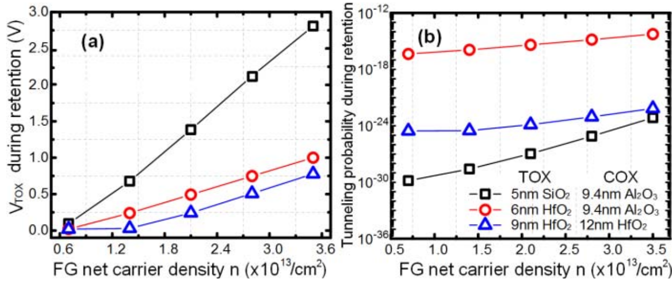


Fig. 8. (a) Coulomb potential energy increase ($=V_{TOX}$) and (b) Tunneling probability through TOX during retention, versus FG net carrier density for three types of TOX/COX stacks.

comparative reference. Aggressively scaled high-K TOX with large FG capacitance (6-nm HfO_2 as TOX, 9.4-nm Al_2O_3 as COX) and the proposed stack are studied. As shown in Fig. 8(a), the Coulomb potential energy increase of the SiO_2 device increases much faster, as expected, than that of the two high-K counterparts versus FG net carrier density. This effect leads to faster increase of tunneling probability during retention, as shown in Fig. 8(b), since the potential energy increase lowers the tunnel barrier. For the condition of very high FG carrier density, 5-nm SiO_2 loses its advantage of lower tunneling leakage over 9-nm HfO_2. The tunneling leakage of 6-nm HfO_2 device is much higher than the other two in the entire range of FG carrier density, indicating that 6 nm is overscaled for HfO_2 TOX. Here, tunneling probability is calculated based on direct-tunneling (DT) mechanism. It is well known that the distribution of oxide traps in energy and space has a statistical behavior that accordingly renders a statistical distribution to SILC. Accurate description of SILC needs a statistical study based on Monte Carlo simulation, which is not the focus of this paper. On the other hand, as discussed above, the proposed new retention mechanism reduces the heavy dependence of retention on the quality of TOX and, thus, is immune to SILC. Therefore, in a simple but physical manner, DT that shares the same trend with SILC, such as the dependence on tunnel barrier height and thickness, is used in this paper to evaluate and compare the upper limits of the retention performances of the devices that have been studied.

As indicated by (2), high-K-based TOX reduces GCR and thus gate efficiency, compared with the conventional SiO_2. It is necessary to examine their controllability over the tunneling barrier of TOX as well as P/E efficiency. As in case of traditional Si/SiO_2/poly-Si devices, the P/E efficiency in TMD/high-K/MLG heterostructure devices is primarily determined by the GCR and the tunneling properties of TOX. The TMD and the MLG are not expected to limit the P/E efficiency since they have large DOS available for tunneling from and tunneling to. As shown in Fig. 9, the 5-nm SiO_2 device has the highest GCR, around 0.57, and provides the largest varying range of tunneling probability, from 10^{-30} to 10^{-7}. The 6-nm HfO_2 device offers the smallest range due to its thin TOX and small GCR (0.17), but it offers comparable programming efficiency (tunneling probability at high overdrive voltage) with respect to the SiO_2 device due to

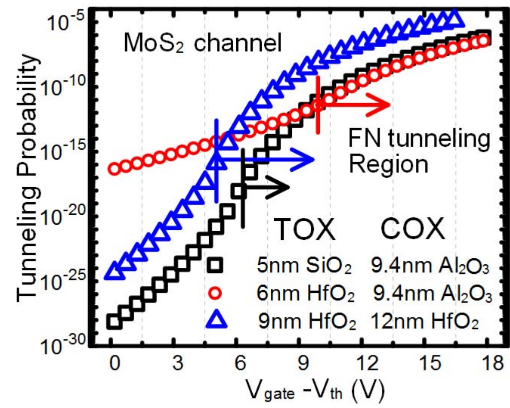


Fig. 9. Tunneling probability versus overdrive gate voltage during programming operation for three types of TOX/COX stacks. Monolayer MoS_2 is used as channel material. The vertical bars divide lines into DT and FNT regions.

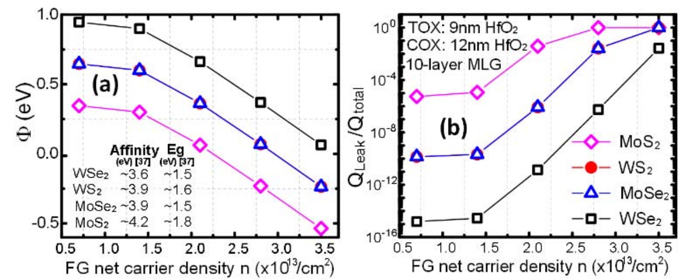


Fig. 10. (a) Additional leakage barrier Φ and (b) ratio of leakable charge density (with energy higher than E_c of the channel material) to the total stored charge density, versus the FG net carrier density for the 9-nm HfO_2 device with 10-layer MLG. Listed electron affinity and bandgap data for the monolayer TMDs in (a) are from [37]. The symbols for WS_2 and $MoSe_2$ coincide in (a) and (b) due to the nearly identical parameters for these two materials.

its small tunneling mass (0.18 m_0) [36] and low tunneling barrier height (1.7 eV for MoS_2 channel [11]). Although the 9-nm HfO_2 device has a relatively low GCR (0.43), it shows the highest turn-ON tunneling probability (10^{-5}) and relatively large varying range of tunneling probability, from 10^{-25} to 10^{-5}, owning to its small EOT (thus stronger electric field in TOX), small tunneling mass, and low tunneling barrier height.

Figs. 10(a) and (b) show the extracted Φ and the ratio of leakable charge (with energy higher than E_c of TMD channel) to the total stored charge, respectively, for the proposed 9-nm HfO_2 stack with four types of TMD channel materials. With increasing FG carrier density, the leakage barrier keeps decreasing until it becomes negative. MoS_2 can only provide a small Φ and reduce the leakable charge density ratio only when the stored charge density is low. Other three materials have smaller affinities and thus smaller leakable charge density ratio. WSe_2 provides the highest Φ among the four studied TMD materials. It allows Φ to remain positive in the entire range of FG carrier density. Under the condition of low FG carrier density, the leakable charge density ratio can be reduced by nearly 15 orders. The introduced novel retention mechanism essentially changes the lower limit of the integral and the Fermi–Dirac function, besides the tunneling probability in the current expression in (1). The leakage currents for the

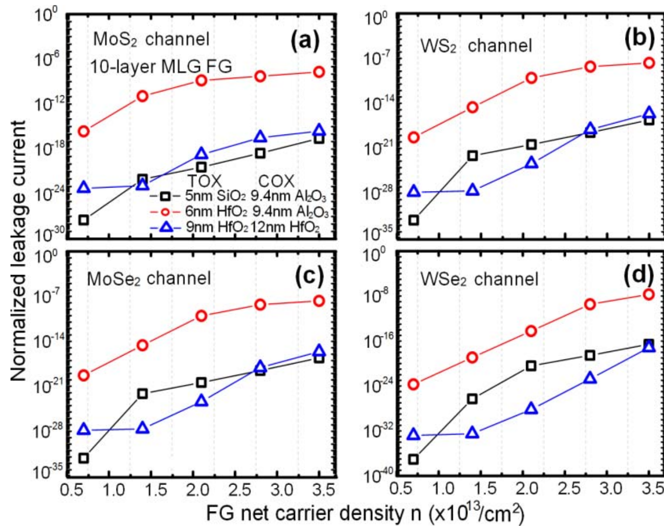


Fig. 11. Normalized (divided by the programming current of the 5-nm SiO_2 device) leakage current versus FG net carrier density (n) for devices with monolayer (a) MoS_2, (b) WS_2, (c) $MoSe_2$, and (d) WSe_2 as channel materials. 10-layer MLG is used as the FG.

three types of TOX/COX stacks with MoS_2, $MoSe_2/WS_2$, and WSe_2 can be calculated using (1).

To make the ratio between programming current and leakage current look straightforward, the calculated leakage currents are normalized by a typical programming current level, specifically, that of the 5-nm SiO_2 device, as shown in Figs. 11(a)–(c). It can be observed that the 6-nm HfO_2 device offers the largest leakage to programming current ratio for all four TMD channels. For MoS_2 channel, the 9-nm HfO_2 device offers limited retention benefit due to small Φ, thus still leakier than the SiO_2 device. $MoSe_2$ or WS_2 channels offer larger Φ, thus the 9-nm HfO_2 device has comparable retention for high FG carrier density and better retention for intermediate FG carrier density, compared with the SiO_2 device. WSe_2 channel provides the largest Φ so that the retention for the 9-nm HfO_2 device is better than the SiO_2 device in most of the range of n. It is worthwhile to mention that the leakage current in the SiO_2 device is always the lowest in the condition of low FG carrier density, because the SiO_2 device is able to reap the retention benefits in this condition. At very high FG charge densities, all three TMD material-based devices lose their retention benefits, and become quite leaky. As a result, device with 9-/12-nm HfO_2 TOX/COX and WSe_2 channel exhibits the best retention characteristics among all the combinations. More importantly, EOT of this device is only 3.5 nm (considering the two series capacitors), which is extremely beneficial for scaling beyond the 10-nm node. Therefore, GCR for TMD/MLG based FG transistors may not need to be very high due to the introduced new mechanism. Table II provides a best-case comparison between conventional Si/poly-Si device and proposed TMD/MLG heterostructure device in terms of scaling and performance, which highlights the advantage of the latter device. For the multibit-storage applications, the total P/E memory window (considering both P/E operation in which the FG is negatively/positively charged) is expected

TABLE II

COMPARISON TABLE FOR BEST-CASE PROJECTIONS ON THE SCALING OF Si/POLY-Si [38], [39] BASED AND TMD/MLG HETEROSTRUCTURE-BASED FG TRANSISTORS (WITH 3.5-nm EOT, WHICH IS APPROXIMATELY EQUIVALENT TO THE DG CASE IN FIG. 5)

	TOX EOT (nm)	COX EOT (nm)	Gate Length (nm)	Cell-to-Cell Distance (nm)	Ratio between P/E current and leakage current
Si/poly-Si	5 (SiO_2)	5 (high-K)	15-16	15-16	10^{17} - 10^{27}
TMD/MLG	1.5 (HfO_2)	2 (HfO_2)	5 - 10	5 - 10	10^{19} - 10^{34}

to be around 12 V. To achieve this, the EOT of COX in our device should increase from 2 to 3.5–4.0 nm for stored carrier density of $3.5 \times 10^{13}/cm^2$. In this situation, the total EOT of 5.0–5.5 nm for TMD/MLG is still much less than the >10-nm EOT for Si/poly-Si device.

Finally, although process and cost related issues, relevant to high-volume manufacturing, are not considered in this proof-of-concept work, it is perhaps worthwhile to mention that all the 2D materials of interest in this paper have been successfully synthesized over large area [40]–[45], which provide sufficient credibility for realizing future high-volume production of the proposed heterostructure FG device in a cost-effective manner.

VI. CONCLUSION

As summarized in Table II, TMD/MLG heterostructure-based FG transistors exhibit strong immunity to CTCI and better electrostatics for gate length scaling, and may take advantage of band offset (if TMD material is judiciously chosen) to improve charge retention and thus facilitate further gate-dielectric scaling. Combined with the proposed high-K gate-dielectric stack for which EOT is only 3.5 nm, WSe_2/MLG heterostructure-based FG transistor shows excellent potential for sub-10-nm nodes, which can significantly extend the lifetime of the FG transistor based flash memory cell in either 2-D or 3-D IC design architecture.

REFERENCES

[1] P. Pavan, L. Larcher, and A. Marmiroli, *Floating Gate Devices: Operation and Compact Modeling*. New York, NY, USA: Springer-Verlag, 2004.

[2] J.-D. Lee, S.-H. Hur, and J.-D. Choi, "Effects of floating-gate interference on NAND flash memory cell operation," *IEEE Electron Device Lett.*, vol. 23, no. 5, pp. 264–266, May 2002.

[3] (2012). *Table PIDS8a, Flash Memory Technology Requirements*. *International Technology Roadmap for Semiconductors*. [Online]. Available: http://www.itrs.net

[4] K. Parat, "Recent developments in NAND flash scaling," in *Proc. Int. Symp. VLSI Technol. Syst., Appl.*, Apr. 2009, pp. 101–102.

[5] N. Ramaswamy *et al.*, "Engineering a planar NAND cell scalable to 20 nm and beyond," in *Proc. 5th IEEE Int. Memory Workshop (IMW)*, May 2013, pp. 5–8.

[6] S. Raghunathan, T. Krishnamohan, K. Parat, and K. Saraswat, "Investigation of ballistic current in scaled floating-gate NAND FLASH and a solution," in *Proc. IEEE IEDM*, Dec. 2009, pp. 819–822.

[7] A. J. Hong *et al.*, "Graphene flash memory," *ACS Nano*, vol. 5, no. 10, pp. 7812–7817, 2011.

[8] H. Li, C. C. Russ, W. Liu, D. Johnsson, H. Gossner, and K. Banerjee, "On the electrostatic discharge robustness of graphene," *IEEE Trans. Electron Devices*, vol. 61, no. 6, pp. 1920–1928, Jun. 2014.

[9] J. S. Bunch *et al.*, "Impermeable atomic membranes from graphene sheets," *Nano Lett.*, vol. 8, no. 8, pp. 2458–2462, 2008.

[10] Y. Sui and J. Appenzeller, "Screening and interlayer coupling in multilayer graphene field-effect transistors," *Nano Lett.*, vol. 9, no. 8, pp. 2973–2977, 2009.

[11] S. Bertolazzi, D. Krasnozhon, and A. Kis, "Nonvolatile memory cells based on MoS_2/graphene heterostructures," *ACS Nano*, vol. 7, no. 4, pp. 3246–3252, 2013.

[12] W. Cao, J. Kang, W. Liu, Y. Khatami, D. Sarkar, and K. Banerjee, "2D electronics: Graphene and beyond," in *Proc. Eur. Solid-State Device Res. Conf. (ESSDERC)*, Sep. 2013, pp. 37–44.

[13] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[14] G.-L. Loi, B. Agrawal, N. Srivastava, S.-C. Lin, T. Sherwood, and K. Banerjee, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *Proc. 43rd IEEE/ACM Design Autom. Conf.*, Jul. 2006, pp. 991–996.

[15] J. Kang, W. Cao, X. Xie, D. Sarkar, W. Liu, and K. Banerjee, "Graphene and beyond-graphene 2D crystals for next-generation green electronics," *Proc. SPIE*, vol. 9083, p. 908305, Jun. 2014.

[16] H. Tanaka *et al.*, "Bit cost scalable technology with punch and plug process for ultra high density flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2007, pp. 14–15.

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

[18] *Atlas User's Manual*, Silvaco, Inc., Santa Clara, CA, USA, Jun. 2009.

[19] S. M. Sze, *Physics of Semiconductor Devices*, 3rd ed. New York, NY, USA: Wiley, 2007, pp. 328–337.

[20] C. Yeh, *Advanced Charge Trap Memory: Stack Design and Cell Characterization*. Ann Arbor, MI, USA: ProQuest, Sep. 2011.

[21] M. Ezawa, "Peculiar width dependence of the electronic properties of carbon nanoribbons," *Phys. Rev. B*, vol. 73, no. 4, pp. 045432-1–045432-8, 2006.

[22] C. Xu, H. Li, and K. Banerjee, "Modeling, analysis, and design of graphene nano-ribbon interconnects," *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1567–1578, Aug. 2009.

[23] T. Georgiou *et al.*, "Vertical field-effect transistor based on graphene–WS_2 heterostructures for flexible and transparent electronics," *Nature Nanotechnol.*, vol. 8, pp. 100–103, Dec. 2012.

[24] K. D. Cantley, Y. Liu, H. S. Pal, T. Low, S. S. Ahmed, and M. S. Lundstrom, "Performance analysis of III-V materials in a double-gate nano-MOSFET," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2007, pp. 113–116.

[25] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer MoS_2 transistors," *Nature Nanotechnol.*, vol. 6, pp. 147–150, Nov. 2011.

[26] H. Fang, S. Chuang, T. C. Chang, K. Takei, T. Takahashi, and A. Javey, "High-performance single layered WSe_2 p-FETs with chemically doped contacts," *Nano Lett.*, vol. 12, no. 7, pp. 3788–3792, 2012.

[27] W. Liu, J. Kang, D. Sarkar, Y. Khatami, D. Jena, and K. Banerjee, "Role of metal contacts in designing high-performance monolayer n-type WSe_2 field effect transistors," *Nano Lett.*, vol. 13, no. 5, pp. 1983–1990, 2013.

[28] D. Sarkar, W. Liu, X. Xie, A. C. Anselmo, S. Mitragotri, and K. Banerjee, "MoS_2 field-effect transistor for next-generation label-free biosensors," *ACS Nano*, vol. 8, no. 4, pp. 3992–4003, 2014.

[29] W. Cao, J. Kang, D. Sarkar, W. Liu, and K. Banerjee, "Performance evaluation and design considerations of 2D semiconductor based FETs for sub-10 nm VLSI," in *Proc. IEEE Int. Electron Device Meeting*, Dec. 2014, pp. 30.5.1–30.5.4.

[30] J. Kang, D. Sarkar, W. Liu, D. Jena, and K. Banerjee, "A computational study of metal-contacts to beyond-graphene 2D semiconductor materials," in *Proc. IEEE Int. Electron Device Meeting*, Dec. 2012, pp. 17.4.1–17.4.4.

[31] J. Kang, W. Liu, D. Sarkar, D. Jena, and K. Banerjee, "Computational study of metal contacts to monolayer transition-metal dichalcogenide semiconductors," *Phys. Rev. X*, vol. 4, no. 3, p. 031005, 2014.

[32] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches," *J. Appl. Phys.*, vol. 92, no. 7, pp. 3730–3739, Oct. 2002.

[33] C. Kittel, *Introduction to Solid State Physics*. New York, NY, USA: Wiley, 2004.

[34] X. Xie *et al.*, "Low-frequency noise in bilayer MoS_2 transistor," *ACS Nano*, vol. 8, no. 6, pp. 5633–5640, 2014.

[35] F. Guinea, A. H. C. Neto, and N. M. R. Peres, "Electronic states and Landau levels in graphene stacks," *Phys. Rev. B*, vol. 73, no. 24, pp. 245426-1–245426-8, 2006.

[36] Y. T. Hou, M.-F. Li, H. Y. Yu, and D. L. Kwong, "Modeling of tunneling currents through HfO_2 and $(HfO_2)_x(Al_2O_3)_{1-x}$ gate stacks," *IEEE Electron Device Lett.*, vol. 24, no. 2, pp. 96–98, Feb. 2003.

[37] J. Kang, S. Tongay, J. Zhou, J. Li, and J. Wu, "Band offsets and heterostructures of two-dimensional semiconductors," *Appl. Phys. Lett.*, vol. 102, no. 1, p. 012111, 2013.

[38] J. Seo *et al.*, "Highly reliable M1X MLC NAND flash memory cell with novel active air-gap and p+ poly process integration technologies," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2013, pp. 3.6.1–3.6.4.

[39] M. Helm *et al.*, "A 128 Gb MLC NAND-flash device using 16 nm planar cell," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2014, pp. 326–328.

[40] W. Liu, H. Li, C. Xu, Y. Khatami, and K. Banerjee, "Synthesis of high-quality monolayer and bilayer graphene on copper using chemical vapor deposition," *Carbon*, vol. 49, no. 13, pp. 4122–4130, Nov. 2011.

[41] W. Liu, S. Krämer, D. Sarkar, H. Li, P. M. Ajayan, and K. Banerjee, "Controllable and rapid synthesis of high-quality and large-area Bernal stacked bilayer graphene using chemical vapor deposition," *ACS Chem. Mater.*, vol. 26, no. 2, pp. 907–915, 2014.

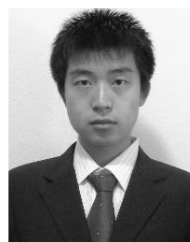
[42] Y. Zhan, Z. Liu, S. Najmaei, P. M. Ajayan, and J. Lou, "Large-area vapor-phase growth and characterization of MoS_2 atomic layers on a SiO_2 substrate," *Small*, vol. 8, no. 7, pp. 966–971, Apr. 2012.

[43] S. Najmaei *et al.*, "Vapour phase growth and grain boundary structure of molybdenum disulphide atomic layers," *Nature Mater.*, vol. 12, pp. 754–759, Jun. 2013.

[44] Y. Yu, C. Li, Y. Liu, L. Su, Y. Zhang, and L. Cao, "Controlled scalable synthesis of uniform, high-quality monolayer and few-layer MoS_2 films," *Sci. Rep.*, vol. 3, p. 1866, May 2013.

[45] J. Huang *et al.*, "Large-area synthesis of highly crystalline WSe_2 monolayers and device applications," *ACS Nano*, vol. 8, no. 1, pp. 923–930, 2014.

Wei Cao (S'12) received the B.S. degree in physics from Nanjing University, Nanjing, China, in 2008, and the M.S. degree in microelectronics and solid-state electronics from Fudan University, Shanghai, China, in 2011. He is currently pursuing the Ph.D. degree with the Nanoelectronics Research Laboratory, Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA, USA.



Jiahao Kang (S'10) received the B.E. degree in microelectronics from the Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing, China, in 2010. He is currently pursuing the Ph.D. degree with the Nanoelectronics Research Laboratory, Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA, USA.

Simone Bertolazzi received the Laurea degree in engineering physics from the Politecnico di Milano, Milan, Italy, in 2007, and the M.Sc. degree in engineering physics from the Politecnico di Milano and the École Polytechnique de Montréal, Montréal, QC, Canada, in 2010. He is currently pursuing the Ph.D. degree with the Laboratory of Nanoscale Electronics and Structures, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Andras Kis (M'13) received the Ph.D. degree in physics from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2003.

He is currently an Assistant Professor of Electrical Engineering with EPFL, where he heads the Laboratory of Nanoscale Electronics and Structures. His current research interests include 2D semiconductors, electronic devices, and circuits based on MoS_2 and related 2D materials.

Kaustav Banerjee (S'92–M'99–SM'03–F'12) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1999.

He is currently a Professor of Electrical and Computer Engineering and the Director of the Nanoelectronics Research Laboratory with the University of California at Santa Barbara, Santa Barbara, CA, USA.