Graphene Flash Memory

Augustin J. Hong, †,‡,* Emil B. Song, §,∥,‡ Hyung Suk Yu, § Matthew J. Allen, § Jiyoung Kim, † Jesse D. Fowler, † Jonathan K. Wassei, §,∥ Youngju Park, § Yong Wang, § Jin Zou, § Richard B. Kaner, § Bruce H. Weiller, † and Kang L. Wang §

†IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, United States, ‡Department of Electrical Engineering, The Department of Materials Science & Engineering, and The Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States, §Micro/Nano Technology Department, Space Materials Laboratory, The Aerospace Corporation, Los Angeles, California 90009, United States, †Device Solutions Business, Samsung Electronics Co., Korea, and ∥Materials Engineering and Centre for Microscopy and Microanalysis, University of Queensland, Brisbane, St Lucia, Queensland 4072, Australia. These authors contributed equally to this work.

Graphene has emerged as a fascinating platform because its planar atomic structure and high symmetry lead to a variety of superlative electronic and physical properties. Most recent achievements fall broadly into one of two categories: first, those that elucidate or exploit the intrinsic in-plane characteristics of graphene (e.g., field-effect transistors [FETs]) and high tensile strength films, and second, those that utilize the material’s single atomic profile as the low-thickness limit of some scalable thin-film systems such as ultracapacitors or transparent conductors. Here we show that graphene is an excellent platform for flash memory, one that may indeed help overcome several challenges faced by current industry standards.

To date, graphene has been incorporated into several types of nonvolatile memory structures, where each memory type operates on a unique physical mechanism. For example, Echtermeyer et al. show bistable state operation through chemical modification of graphene to form insulating graphene derivatives in graphene FETs, while Standley et al. and Son et al. utilize the filament effect to produce memory functionality in a two-terminal resistor structure. Nano/micro-electromechanical switches (NEMS/MEMS) exploiting graphene’s low mass density and high Young’s modulus have also been realized by Milaninia et al. and Kim et al., which show promise for low power memory applications. Furthermore, the hysteresis effect in graphene ferroelectric-field-effect transistors (FFETs) comprised of a ferroelectric gate oxide and a graphene channel has proven to be an alternative to conventional semiconducting FFETs.

All of the aforementioned memory types are currently being considered as emerging technologies for nonvolatile memory applications because of the limitations encountered in flash technology when attempting to increase storage capacity by miniaturization. In order to recognize the bottleneck of conventional flash technology, we must first understand the structures and operating mechanisms of flash memory. There are two types of flash memory structures: floating-gate flash memory (FG), which is the current industry standard, and charge-trap flash memory (CT), which is an emerging technology. Flash memory, in general, is comprised of a p-type silicon channel substrate, a tunnel oxide, a semiconductor highly doped n-type polysilicon (FG) or an insulating silicon nitride (CT) data storage layer, a control oxide, and a gate electrode. Writing is achieved by applying a voltage pulse on the gate electrode, which allows electrons to tunnel through the tunnel oxide from the silicon channel to the storage layer. This causes a positive shift in the threshold voltage ($V_{th}$) of the memory device and is simply the additional voltage required to compensate the stored charges in the storage layer. The binary values are defined by the current upon a read.

**ABSTRACT** Graphene’s single atomic layer of sp$^2$ carbon has recently garnered much attention for its potential use in electronic applications. Here, we report a memory application for graphene, which we call graphene flash memory (GFM). GFM has the potential to exceed the performance of current flash memory technology by utilizing the intrinsic properties of graphene, such as high density of states, high work function, and low dimensionality. To this end, we have grown large-area graphene sheets by chemical vapor deposition and integrated them into a floating gate structure. GFM displays a wide memory window of $\sim 6$ V at significantly low program/erase voltages of $\pm 7$ V.

GFM also shows a long retention time of more than 10 years at room temperature. Additionally, simulations suggest that GFM suffers very little from cell-to-cell interference, potentially enabling scaling down far beyond current state-of-the-art flash memory devices.

**KEYWORDS:** graphene · nonvolatile flash memory · memory window · retention time · cell-to-cell interference

*Address correspondence to ajhong@us.ibm.com, emil@ee.ucla.edu.

Received for review May 17, 2011 and accepted August 21, 2011.

Published online August 22, 2011 10.1021/nn201809k

© 2011 American Chemical Society
cycle at a voltage within the width of the threshold voltage shift.

The primary goal of most research in flash memory is to increase the density of storage elements such that their parent devices can be miniaturized, which leads to confronting several challenges that would normally jeopardize device performance at the reduced scale. As the device dimensions become smaller (<30 nm), FG suffers from a lower gate-coupling ratio (GCR), the ratio of the voltage drop across the tunnel oxide and the total voltage across the channel and gate, i.e., the amount of capacitive coupling from the gate to channel) and increasing crosstalk between neighboring floating gates.\(^1\)\(^0\) CT, on the other hand, has a single gate to control the channel directly, which allows a high GCR, and an insulating nitride storage layer, which has insignificant interference. Upon scaling, however, CT can become hindered through variability in \(V_{\text{th}}\), caused by the implementation limit of trap density (~\(10^{19}\) cm\(^{-3}\)) and uniformity,\(^1\)\(^0\) and short retention times due to shallow trap energy levels that can promote trap-assisted Poole–Frenkel conduction during the retention state.\(^1\)\(^9\)

Graphene has the potential to exceed the performance of current flash memory technology by utilizing the exceptional intrinsic properties of graphene, such as high density of states, high work function, and low dimensionality when compared to the conventional FG and CT materials. Recently, it was shown that incorporation of an insulating form of graphene, known as graphene oxide, acts as an effective charge storing layer in CT devices.\(^2\)\(^0\) In its natural semimetallic form, graphene can act as the floating gate in FG devices. Although memory devices can be characterized along a wide variety of metrics, in this study, we specifically address the benefits of metallic graphene in a FG structure with an eye toward the width of the memory window, retention time, and cell-to-cell crosstalk at low operating voltages.

RESULTS AND DISCUSSION

In order to investigate the electrical characteristics of graphene flash memory (GFM), a number of devices were fabricated with the following process flow as shown in Figure 1. First, a 4 in. boron-doped (~\(10^{15}\) cm\(^{-3}\)) Si wafer was cleaned sequentially by a Piranha solution (\(\text{H}_2\text{SO}_4 + \text{H}_2\text{O}_2\)) and then dipped into a buffered oxide etchant (BOE, HF + NH_4F + H_2O) to remove any residual native oxide (Figure 1a). Second, a SiO\(_2\) tunnel oxide of 5 ± 0.2 nm was grown for 7 s through rapid thermal oxidation (RTO) at 1000 °C, under an oxygen flow of 40 sccm (Figure 1b). After the graphene sheets were transferred (see Materials and Methods) onto the tunnel oxide surface (Figure 1c), Raman spectroscopy was used to ensure the quality of the transferred single layered graphene (SLG) and multilayered graphene (MLG) (Figure 2e and f).\(^2\)\(^1\) Third, a control oxide of 35 ± 1 nm was formed by evaporating a 1.1 nm thick Al layer, oxidizing in air for 2 days, and depositing 300 cycles of additional Al\(_2\)O\(_3\) through atomic layer deposition (ALD) (Figure 1d).\(^2\)\(^2\) Gate electrodes of Ti/Al/Au (10 nm/500 nm/50 nm) with various areas (2.5 \(\times\) 10\(^{-5}\) to 7 \(\times\) 10\(^{-4}\) cm\(^2\)) were defined and deposited using photolithography and an e-beam evaporator (Figure 1e). In order to isolate each memory device, the Ti/Al/Au gate electrodes were used as a mask to etch the gate stacks. The Al\(_2\)O\(_3\) layer and graphene present outside the device area were removed with a 30 s Cl\(_2\) reactive ion etching (UNAXIS SLR770), followed by 3 min of O\(_2\) plasma. Finally, a 50 nm thick Pt back contact was made by e-beam evaporation (Figure 1f).

Figure 1a–d show transmission electron micrograph (TEM) cross sections of both as-fabricated single-layer and multilayer GFM devices.

An important figure of merit for flash memory is the memory window, which refers to the shift in threshold voltage of the memory device when switching from the 0 to 1 binary states. Industry standards suggest that a minimum width of 1.5 V is necessary to produce a reasonable on/off ratio for reliable memory functionality. For standard FG devices using polysilicon, this requires a program/erase voltage of around ±20 V.\(^1\)\(^0\) This large voltage requirement is due to the low density of states (DOS) in the degenerately doped polysilicon, which leads to the necessity of high GCRs.
In a conductive medium, the stored charges redistribute themselves on the surface to minimize the coulomb potential energy. The characteristic thickness in which the charges reside is called the Thomas–Fermi (T–F) screening length. In general, metals have an extremely short T–F screening length of only several atomic layers, whereas the one atom thick graphene stores charge in a single layer. By adopting a theoretical model and using the ratio of the stored charges (memory window) between our MLG-FM and SLG-FM, \( N_{\text{MLG-FM}}/N_{\text{SLG-FM}} \approx 1.0 \times 10^{13} \text{ cm}^{-2}/3.33 \times 10^{12} \text{ cm}^{-2} \approx 3 \) (\( \Delta V_{\text{MLG-FM}}/\Delta V_{\text{SLG-FM}} \approx 6 \text{ V}/2 \text{ V} \approx 3 \)), we find the screening length to be \( \lambda = 0.8 \text{ nm} \), which is consistent with reported values.

Given the insignificant memory (≈20 mV) observed in the control device (inset of Figure 3b) and a small unintentional hysteresis induced by interface traps (≈10^{-10}–10^{-11} cm^{-2}) in graphene/SiO\(_2\) FETs, we conclude that the charges responsible for the wide memory window are stored in the graphene layer(s).

The counterclockwise memory effects, as indicated by arrows in each plot in Figure 3a and b, show that electron transfer through the tunnel oxide dominates the electron charge–discharge rate. Conversely, a clockwise memory effect can arise due to a leaky control oxide, resulting in electron transfer through the control oxide, and thus cannot be used in integrated circuits. A minimal leakage current of \( 7 \times 10^{-7} \text{ A/cm}^2 \) at a gate voltage \( V_G \) of –10 V corroborates charge transfer through the tunnel oxide (see Supporting Information).

We noticed that the initial threshold voltage (forward bias regime) of the SLG-FM shows a large negative value (–3.89 V) compared to the threshold voltage of the MLG-FM (–1.33 V). This large negative
threshold voltage in the SLG-FM is observed as we need to apply an additional electric field to compensate for the holes and invert the p-type Si substrate. This indicates that SLG is inherently p-type, which is consistent with reports that graphene is doped by atmospheric molecules, photoresist residue, metal etchants, and Al2O3. Additionally, interface states at the oxide/graphene interface created by defects in graphene (the D band in Raman spectra) or dangling bonds of the oxide can induce additional positive charges inside the gate stack. All of the above can be applied to both SLG and MLG. However, MLG will be less sensitive to charge doping effects since the additional layers will screen and lessen the effects of these charges. Accurate contributions from defects and doping to initial threshold voltages should be further investigated.

The next figure of merit for flash memory devices is retention time, which refers to the potential lifetime of nonvolatile storage. Generally, retention time requirements are more than 10 years before the device loses 50% of stored charge. Loss of storage is typically the result of charge tunneling from the floating gate through the tunnel oxide. The rate of tunneling is dependent on the height and thickness of the electronic barrier presented by the tunnel oxide. In traditional polysilicon/SiO2 FG devices, the barrier height is fixed and simply the difference between the SiO2 conduction band edge (0.95 eV) and that of polysilicon (4.02 eV), which implies that the minimum tunnel oxide thickness is limited to 7–8 nm. In contrast, the graphene/SiO2 system offers a larger electronic barrier height due to the higher work function of graphene (~4.6 ± 0.05 V near the Dirac point), which allows the tunnel oxide to be further thinned.

Figure 4 shows the energy band diagram of our MLG/SiO2/Si junction and experimental data for retention time of MLG-FM at room temperature with a SiO2 tunnel oxide of only 5 ± 0.2 nm (i.e., a 30–40% thinner tunnel oxide than that used for polysilicon devices). Measurements indicate that a charge loss of only 8% should occur after 10 years, which is more than adequate for practical devices (Figure 4a). In order to understand the long retention time, multiple charge loss mechanisms can be considered; Schottky emission (SE), Fowler–Nordheim tunneling (FNT), Poole–Frenkel tunneling (PFT), and direct tunneling (DT). Under a normal retention state (Vt = 0) the electronic structure at the device interface is different before and after programming, which is illustrated in the band diagrams of Figure 4b and c. This is a result of the creation of an internal field across the tunnel oxide by stored charges in MLG (NMLGFM ≈ 1013 cm−2). Both the existence of the large barrier height (~3.65 eV) and low electric-field across the tunnel oxide make DT the most likely candidate for charge loss since the other mechanisms require high electric fields to contribute significantly. However, DT is a low-probability event. Thus, we believe that the DT mechanism is attributable to the long retention time of the constructed GFM devices.

The final figure of merit that we examine for flash memory is cell-to-cell interference. For a 2D planar configuration, a given storage element is mainly influenced by its two nearest neighbors, which share the same word line to their gates. This leads to variability in Vt, and accordingly, a bit error can easily occur between nearest neighbors when interference is significant. This is a complex interplay because the increasing thickness of the charge storage layer increases the GCR and hence lowers the operating voltage, but also

---

Figure 4. Retention characteristics of GFM. (a) Retention measurement of MLG-FM showing only 8% of charge loss in 10 years at room temperature. (b) Energy band diagram of MLG/SiO2/Si junctions before programming. (a) Energy band diagram after programming, which shows three possible mechanisms for charge loss during retention state: SE, FNT, and DT. All three mechanisms depend exponentially on the barrier height (φB) between the work function of graphene and the SiO2 tunnel oxide.
increases the interference mainly due to an increase in side-wall capacitance. A careful optimization must take place in order to minimize both the operating voltage and cell-to-cell interference. Graphene offers a unique solution to this problem due to its thinness, which minimizes the resulting field upon charging by nearest neighbors and, hence, lowers the crosstalk between nearest neighbors. In some sense, lower interference is a result of graphene’s lower GCR.

Simulations of both traditional polysilicon/SiO2 and thin film metal-based devices are presented in Figure 5. The simulation is based on a thin film metal of 1 nm (~ trilayer graphene), which is the T-F screening length of MLG with a work function of graphene corresponding to 4.6 eV. The simulation results provide an upper limit for graphene devices, since graphene’s DOS is less than conventional metallic systems. To understand the limit for graphene devices, since graphene’s DOS is less than conventional metallic systems. To understand the maximum interference in a given cell, we simulate the situation where two nearest neighbors are programmed at high gate voltages, while a given cell remains unprogrammed. We then monitor the Vfl shift of the unprogrammed cell due to its nearest neighbors. Simulation results show that the conventional FG interference further suggest that graphene may be instrumental in the next round of miniaturization of flash memory.

In summary, our experiments demonstrate the benefits of graphene as a platform for flash memory. The high density of states, high work function, and low dimensionality positively influence device performance, leading to a wide window of operation at low voltages, long retention time, and low cell-to-cell interference. The simulations pertaining to cell-to-cell interference further suggest that graphene may be instrumental in the next round of miniaturization of flash memory.