Efficiency and reliability of Fowler-Nordheim tunnelling in CMOS floating-gate transistors

B. Rumberg and D.W. Graham

Floating-gate transistors are increasingly used for digital and/or analogue non-volatile memory in standard CMOS integrated circuits. The mask design of the floating-gate's tunnelling junction, where erasure and/or writing occur, is examined. Aided by static and transient tunnelling current measurements for a variety of tunnelling junctions, recommendations for constructing these junctions to minimise the duration, power consumption and oxide degradation of programming are presented.

Introduction: CMOS floating-gate (FG) transistors are important for including non-volatile memory in systems-on-chips and for creating dense, electronically tunable analogue systems. A FG transistor is a MOSFET that has no resistive connection to its gate; instead, a 'control gate' couples capacitively onto the transistor's 'floating gate' (Fig. 1*a*). As a result, the FG's charge, which can be modified by using Fowler-Nordheim (FN) tunnelling and hot-electron injection, creates a programmable and non-volatile threshold-voltage shift from the perspective of the control gate.

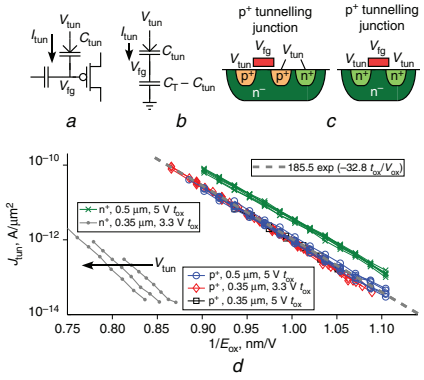

Fig. 1 *Fowler-Nordheim tunnelling characteristics*
a Schematic of floating-gate transistor
b Equivalent tunnelling circuit
c Structure of tunnelling junctions
d Fowler-Nordheim voltage–current measurements

Erasure, and often writing, is achieved by tunnelling electrons through the tunnelling junction, C_{tun}. The design of this junction has significant implications on the speed, efficiency and long-term reliability of writing and erasure. In this work, the characteristics of the two basic junction structures are compared, and the junction size that achieves minimal tunnelling duration and oxide degradation is derived.

FN tunnelling current: FN tunnelling occurs when the electric field across the C_{tun} dielectric is sufficiently high to distort the energy band such that the effective barrier thickness is reduced to 5 nm [1]. An electric field of 0.64 V/nm is required to initiate tunnelling for the 3.2 eV Si–SiO_2 interface (from the floating gate to the oxide dielectric) [1]. This tunnelling current is approximated by [2]

$$J_{tun} = \alpha \exp\left(-\beta t_{ox}/V_{ox}\right) \quad (1)$$

where t_{ox} is the thickness of the oxide barrier, V_{ox} is the voltage across the barrier, and α and β are constants related to the fabrication process and junction type. A thin oxide is desired to minimise the tunnelling voltage. In standard CMOS, the gate oxide is typically used because it is thin and also of high quality, which benefits reliability and predictability. Oxides thinner than 5 nm should be avoided to deter direct tunnelling; consequently, higher voltage I/O devices for 2.5 V (5 nm), 3.3 V (7–8 nm) or 5 V (14–15 nm) operation with thicker oxides are typically used in fine-geometry processes [3]. Thus, our results using 3.3 and 5 V devices from 0.35 and 0.5 µm processes provide a relevant insight into tunnelling in new processes, as well.

To remove electrons from the FG, V_{tun} is raised to a high voltage, typically higher than the reverse breakdown voltage of the source/drain diffusions, but less than the breakdown of the well-to-substrate junction. To avoid reverse breakdown, tunnelling junctions are generally placed within a well. Fig. 1*c* shows the two basic types of tunnelling junctions: a p^+ MOS capacitor formed as a standard pFET and an n^+ MOS capacitor formed with n^+ diffusions along the gate. The n^+ junctions have traditionally been favoured for analogue memory applications [2], but p^+ junctions are becoming common for standard CMOS Flash applications [3]. In this work, the static and transient characteristics of p^+ and n^+ junctions are compared with determine recommendations for junction design.

FG programming characteristics can be engineered via the design of the tunnelling junction – the width, length, t_{ox} and diffusion type. Fig. 1*d* shows the measured FN characteristics for a variety of junction designs. Each trace was obtained by reading V_{fg} through a buffer during a pulse to V_{tun} (i.e. typical tunnelling conditions). All the terminals except V_{tun} and V_{fg} were held fixed, so the circuit can be modelled by Fig. 1*b*. By reading V_{fg}, we could obtain $E_{ox} = (V_{tun} - V_{fg})/t_{ox}$ and $I_{tun} = C_T(d/dt)(V_{fg})$, where C_T is the total capacitance connected to the node. Four different C_{tun} dimensions were used on a 0.5-µm process (µm × µm): 1.5 × 0.6, 3 × 0.6, 1.5 × 1.2 and 3 × 1.2. Five dimensions were used for the 3.3 V and 5 V 0.35 µm p^+ junctions (µm × µm): 0.5 × 0.5, 0.5 × 1, 0.5 × 2, 1 × 0.5 and 2 × 0.5. The 0.35 µm n^+ junction was 0.4 × 0.35 µm.

The p^+ junction curves in Fig. 1*d* all align and are excellently described by $\alpha = 185.5$ A/µm^2 and $\beta = 32.8$ V/nm. The traces align when normalised by the area (i.e. plotted as current density), which illustrates that, at least for large enough V_{ox} to achieve fast tunnelling, the current comes from the full junction area rather than from the edges [2].

The curves for the n^+ junctions correspond to the time after which the junctions have recovered from depletion and have begun to tunnel (more details are provided in the following Section). For 0.5 µm, the difference between the p^+ and n^+ junctions is probably caused by their different flat-band voltages. The low current and V_{tun}-dependence of the 0.35 µm n^+ junction may be explained by variations in the effective oxide thickness due to finite charge depth [4].

In summary, p^+ junctions are more consistent from process to process and the p^+ tunnelling current is significantly higher in the 0.35 µm process.

Temporal dynamics of tunnelling junctions: In addition to the FN traces, the temporal dynamics of the junctions must also be considered. The variable capacitance of the MOS capacitor structures can cause complex transient characteristics. Fig. 2*a* shows the measured transient responses of 0.5 µm FGs for 20 V V_{tun} pulses. This experiment is analogous to block erasure in which all the FGs, regardless of their initial value, should tunnel to approximately the same value. The pulse duration for the p^+ junctions is 340 µs and the durations for the n^+ junctions have been adjusted to achieve an approximately equal amount of tunnelling. Based on the equivalent circuit in Fig. 1*b*, V_{fg} will rise as the electrons tunnel through C_{tun}. As V_{fg} rises, the tunnelling rate decreases due to a decreasing V_{ox}. As a result, FGs with different initial voltages approach the same final voltage (see the p^+ junctions in Fig. 2*a*). The p^+ junctions perform as expected given (1). The n^+ junctions, however, experience a voltage-dependent delay before they begin to tunnel. This delay is a result of the depletion region that is formed underneath the gate in response to the V_{tun} pulse. Most of the tunnelling voltage is dropped across the depletion capacitance, resulting in a small oxide voltage and thus no tunnelling current. The depletion region collapses slowly as carriers are generated from thermal generation and band-to-band tunnelling, after which tunnelling begins [5]. In both processes, the p^+ junctions had no measurable delay.

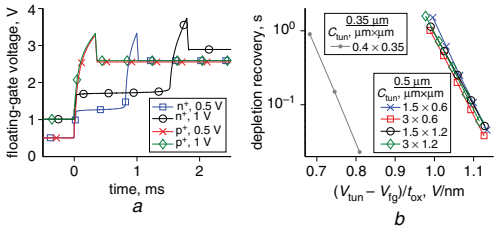

Fig. 2 *Tunnelling junction transient characteristics*
a Transient characteristics of n^+ and p^+ junctions with different initial voltages
b Depletion recovery time of n^+ junctions

Fig. 2*b* shows that the depletion recovery time of the n^+ junctions (duration between the beginning of the tunnelling pulse and the start of tunnelling) is independent of the junction area when plotted in terms of $V_{tun} - V_{fg}$. However, larger junctions have a smaller $V_{tun} - V_{fg}$ due to the capacitive division (Fig. 1*b*); as a result, the delay time increases with C_{tun}. Another result of the voltage-dependence is that the delay time is exponentially related to the initial FG voltage. This is problematic for the memory arrays because, for short erase times, the post-erase distribution of FG values can have a complex and non-monotonic relation to the initial distribution of FG voltages. These transient depletion characteristics are more pronounced for typical tunnelling voltages in the 0.5 µm process than in the 0.35 µm process. However, in both processes, the p^+ junctions achieve faster tunnelling times because of their higher I_{tun} in 0.35 µm and because of their lack of a depletion recovery delay.

Overall, we suggest that p^+ junctions are the best choice because they are faster, require less power to operate (since they are faster, the high-voltage generation circuitry operates for less time), are more consistent from process to process and are always available in CMOS process design kits.

To verify the reliability of using high-voltage tunnelling pulses, we have performed 100 k write/erase cycles on a 0.35 µm FG with a p^+ tunnelling junction and a 200 fF gate capacitor. The FG's threshold voltage was shifted 1 V up and down in each cycle, transferring an accumulated 20 nC of charge through C_{tun}. We observed only a 30% reduction in tunnelling current.

Sizing of junctions for speed and reliability: Tunnelling junctions are often made to be of minimum size to minimise the coupling from V_{tun} to V_{fg}. However, we will derive the optimal C_{tun}/C_T ratio that minimises the time to tunnel to the post-erasure FG voltage, $V_{fg,e}$. Increasing the junction size has two opposing trends: larger area increases the tunnelling current, but it also increases the coupling onto the FG which reduces the final voltage when V_{tun} steps down. To find the junction size that tunnels to the final voltage in the shortest duration, we first write the tunnelling current in terms of C_{tun} and $V_{fg,e}$ as

$$I_{tun} = \alpha(C_{tun}/\gamma)\exp\left[-\frac{\beta t_{ox}}{\left(1-\left(C_{tun}/C_T\right)\right)V_{tun} - V_{fg,e}}\right] \quad (2)$$

where γ is the unit capacitance (aF/µm^2) of C_{tun}. By taking the derivative with respect to C_{tun} and setting the LHS to '0', we find that tunnelling is maximised for the following coupling ratio

$$\frac{C_{tun}}{C_T} = \frac{\beta t_{ox}}{2V_{tun}}\left(1 + 2\frac{V_{tun} - V_{fg,e}}{\beta t_{ox}} - \sqrt{1 + 4\frac{V_{tun} - V_{fg,e}}{\beta t_{ox}}}\right) \quad (3)$$

This equation is verified in Fig. 3 for 0.5 µm FGs. For $V_{tun} = 20$ V, $V_{fg,e} = 2.5$ V, $t_{ox} \simeq 14$ nm and $\beta = 32.8$ V/nm, the optimal coupling ratio is calculated to be approximately 3.1%. It can be seen that the junction with 3.2% coupling reaches the final voltage 23% faster than the larger junction (which suffers from excessive coupling) and 44% faster than the smaller junction (which suffers from insufficient tunnelling area, thus limiting I_{tun}). In addition to reducing the duration compared with a minimum-sized tunnelling junction, the larger sizing also increases the long-term reliability. This is because oxide degradation is related to the charge-density that has passed through the junction [6]. By using a larger junction, the charge-density is reduced, which contributes to an increase in long-term reliability. For digital Flash applications, C_T may not be large enough to achieve such a ratio, but analogue FG applications use large control-gate capacitors, often 100 fF or more, and so will benefit from this sizing.

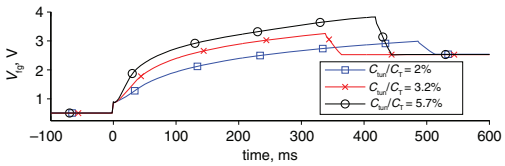

Fig. 3 *Optimal tunnelling junction sizing*

Conclusion: Non-volatile memory is increasingly being included in standard CMOS products. Tunnelling is used in most of these non-volatile memories, and so design methods for tunnelling junctions are of interest if they can improve speed, reliability and/or energy efficiency. We have presented answers to tunnelling junction design decisions that offer improvement in all the three categories.

Acknowledgments: This material is based upon work supported by the National Science Foundation under award no. 1148815. The authors thank J. Hasler for helpful conversations and a historical perspective on tunnelling junctions.

18 July 2013
doi: 10.1049/el.2013.2401
One or more of the Figures in this Letter are available in colour online.

B. Rumberg and D.W. Graham (*Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA*)

E-mail: david.graham@mail.wvu.edu

References

1 Carley, L.: 'Trimming analog circuits using floating-gate analog MOS memory', *IEEE J. Solid-State Circuits*, 1989, **24**, (6), pp. 1569–1575
2 Hasler, P., Minch, B., and Diorio, C.: 'Adaptive circuits using pFET floating-gate devices'. Proc. IEEE ARVLSI, Atlanta, GA, USA, March 1999, pp. 215–229
3 Song, S., Chun, K., and Kim, C.: 'A logic-compatible embedded flash memory for zero-standby power system-on-chips featuring a multi-story high voltage switch and a selective refresh scheme', *IEEE J. Solid-State Circuits*, 2013, **48**, (5), pp. 1302–1314
4 Yang, K., King, Y., and Hu, C.: 'Quantum effect in oxide thickness determination from capacitance measurement'. Proc. IEEE VLSIT, Kyoto, Japan, 1999, pp. 77–78
5 Kolodny, A., Nieh, S., Eitan, B., and Shappir, J.: 'Analysis and modeling of floating-gate EEPROM cells', *IEEE Trans. Electron Devices*, 1986, **33**, (6), pp. 835–844
6 Park, Y., and Schroder, D.: 'Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a Flash EEPROM', *IEEE Trans. Electron Devices*, 1998, **45**, (6), pp. 1361–1368