CAMBRIDGE

Cambridge University Press
978-1-107-40853-1 - Materials Science and Technology for Nonvolatile Memories: Materials Research Society Symposium Proceedings: Volume 1071
Editors: Dirk J. Wouters, Seungbum Hong, Steven Soss and Orlando Auciello
Excerpt
More information

Advanced Flash Memory

Mater. Res. Soc. Symp. Proc. Vol. 1071 © 2008 Materials Research Society 1071-F02-01

Flash Memory Scaling: From Material Selection to Performance Improvement

Tuo-Hung Hou, Jaegoo Lee, Jonathan T. Shaw, and Edwin C. Kan
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853

ABSTRACT

Below the 65-nm technology node, scaling of Flash memory, NAND, NOR or embedded, needs smart and heterogeneous integration of materials in the entire device structure. In addition to maintaining retention, in the order of importance, we need to continuously make functional density (bits/cm^2) higher, cycling endurance longer, program/erase (P/E) voltage lower (negated by the read disturbance, multi-level possibility and noise margin), and P/E time faster (helped by inserting SRAM buffer at system interface). From both theory and experiments, we will compare the advantages and disadvantages in various material choices in view of 3D electrostatics, quantum transport and CMOS process compatibility.

INTRODUCTION

Battery-powered portable electronics, such as mobile phones, MP3 players, digital cameras etc., have fuelled skyrocketing demand for nonvolatile Flash memory since late 1990's. The advance in technology is even more impressive. The Flash technology has demonstrated its outstanding scaling capability in the last decade. A two-fold increase in bit-density of NAND Flash has been realized every year for the past seven years [1]. Today 16-gagabit density with 50-nm design rule is in mass production. This trend far exceeds the projection of the Moore's law in logic integrated circuits. Therefore, Flash is arguably the present technology driver of the semiconductor industry. However, this great momentum, mainly relying on the straightforward geometrical shrinkage, has been expected to slow down for technology nodes of 40 nm and beyond due to several challenging roadblocks in device scaling [1-4]. First, the thickness of tunnel oxide is not easily scaleable in order for satisfactory charge retention, especially after many program/erase (P/E) cycles. The stress induced leakage current (SILC) gives rise to unacceptable statistical distribution in retention for a high-density memory array, which limits the thickness of tunnel oxide to be 7-8 nm [2, 3]. The non-scalable tunnel oxide deteriorates the short channel effects (SCE) and impedes further gate-length scaling. This is particularly severe in NOR-type Flash where the large drain voltage (> 3.2V) is necessary for hot-carrier programming. Second, the distance between adjacent float-gates (FGs) has become extremely narrow due to aggressive scaling. As a result, the cell-to-cell interference is no longer negligible. This in part can be mitigated by reducing the FG height and by utilizing a low-κ spacer between FGs. However, these inevitably hurt the coupling ratio (CR) necessary for decent P/E efficiency. In conventional designs, while the thickness of inter-poly oxide or so called control oxide is also reaching its scaling limit, the CR can still be engineered by the additional capacitance provided by FG sidewalls. The better immunity to the cell-to-cell interference by reducing the FG height is at the expense of the dwindling CR, and as a consequence even higher P/E voltage is required. P/E voltages are projected still at 15 V for NAND Flash until the end of roadmap in 2018 [5]. Higher P/E voltage leads to higher power dissipation and adversely affects the parallel writing process. It also adds tremendous overhead on power consumption and area of the peripheral circuit for both stand-alone and embedded memory [6]. Even more importantly, the endurance under many P/E cycles is deteriorated by the high field in the thin tunnel oxide. The resulting threshold voltage V_{th} shifting and SILC in short-retention bits are the key reliability concerns. Therefore, a fundamentally new approach to scale cell size without compromising memory performance is of great importance in Flash memory technology.

Meanwhile, there has been very active research on alternative nonvolatile memories that do not employ charge storage in FG. Among the most mature are ferroelectric random access memory (FRAM) [7], magnetoresistive random access memory (MRAM) [8], and phase-change random access memory (PRAM) [9]. Although enormous progress has been made, none of them have stood up to completely address the strict requirements for low-cost, high-density, and high-speed nonvolatile storage. FRAM relying on the charge polarization in small capacitors has limited scaling potential. Its destructive read is also undesirable. MRAM and PRAM are still under active investigation to realize an efficient P/E scheme compatible with the current drive capacity of scaled access transistors in the one-transistor-one-magnetic-tunnel-junction (1T1MTJ) and one-transistor-one resistor (1T1R) cells. In addition, any emerging technology has to be a cost-effective replacement, a daunting challenge considering the maturity of today's Flash technology as well as the prevalent implementation of multiple bits per cell. Therefore, it is safe to project that Flash memory will still be the main workhorse of the portable nonvolatile storage for many years to come [10-12]. The question is how we are able to extend its longevity by overcoming aforementioned scaling challenges before any viable alternative becoming a reality.

In an attempt to address this, in this paper we highlight the importance of smart and heterogeneous integration of materials throughout the entire device structure, including charge storage medium, tunnel oxide, control oxide, control gate, and sensing channel. From both theory and experiments, we will compare the advantages and disadvantages in various material choices in view of three-dimensional (3D) electrostatics, quantum transport and CMOS process compatibility. We will limit our discussion mainly on NAND-type Flash memory owing to its better scalability and dominate role in portable massive storage. However, many viewpoints presented here may apply to NOR-type Flash as well.

CHARGE STORAGE MEDIUM

Flash memory relies on the static-charge storage in an isolated FG. The conventional choice of material for FG has been n-doped poly-Si because of its process compatibility in the Si process. However, many aforementioned scaling challenges stem from the continuous poly-Si FG. Non-scaleable thickness of the tunnel oxide due to the poor immunity against SILC and significant cell-to-cell interference are two main inherent disadvantages. Although the industry has every reason to push the continuous FG to its limit, with recent demonstration in the 43nm-node technology [13], it is of little doubt that at some point discrete charge storage, which consists of multiple discrete FGs instead of a continuous one, has to be utilized in order to fundamentally resolve these issues [1, 2, 14]. The discreteness among FGs prevents complete loss of memory states through localized SILC, and greatly suppresses the FG-to-FG coupling. This enables both tunnel oxide and cell size scaling. Proposed device implementation is basically divided into two major categories, silicon-oxide-nitride-oxide-silicon (SONOS) or SONOS-type memories [15-25] and nanocrystal (NC) memories [26-42]. SONOS-type memories utilize natural traps in dielectrics while NC memories utilize semiconductor or metal NCs embedded inside dielectrics for charge storage. Here we are interested in the best option available to address the remaining roadblock, the high P/E voltage.

Cambridge University Press
978-1-107-40853-1 - Materials Science and Technology for Nonvolatile Memories: Materials Research Society Symposium Proceedings: Volume 1071
Editors: Dirk J. Wouters, Seungbum Hong, Steven Soss and Orlando Auciello
Excerpt
More information

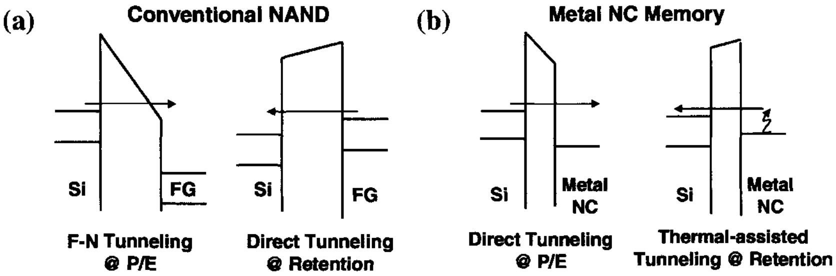

Figure 1 Energy band diagram representation at P/E and retention in the nonvolatile memory cells with (a) thick tunnel oxide and poly-Si FG and (b) thinner tunnel oxide and metal NC

In the present Flash memory, the ratio between retention time t_R and P/E time t_{PE} is about 10^{12}-10^{14}. In order to realize this tremendous ratio, field-asymmetric tunneling processes in the tunnel barrier have to be deliberately engineered between retention and P/E. The asymmetry in conventional Flash is most often provided by external P/E voltage. For example, in NAND Flash, the asymmetry between the Fowler-Nordheim (FN) tunneling under P/E and the direct tunneling (DT) during retention is exploited as illustrated in Fig. 1(a). However, this approach also limits the scalability of P/E voltage. Metal NC memory [32-37] has been proposed to enhance the tunneling asymmetry at low P/E voltage. The material-dependent FG work function of metal NCs provides additional band offset to the Si band edges of the channel. During retention, only a small portion of thermally excited charge in metal NCs is able to directly tunnel back to Si channel due to the Si forbidden bandgap. This greatly improves memory retention even with a thinner tunnel oxide. Meanwhile, the thinner tunnel oxide allows fast P/E operation through DT at low P/E voltage. The asymmetry between the DT under P/E and the thermal-assisted tunneling during retention as illustrated in Fig. 1(b) is fundamentally different from that in the conventional NAND. On the contrary, semiconductor NCs, such as Si, Ge, and SiGe NCs [26-30], provides little or none band offset to the channel. The quantum-size effect of semiconductor NCs further broadens bandgap larger than that in bulk Si. In metal NCs, this bandgap broadening is suppressed by the large density of states in metal for the size of metal NCs we are generally interested in [43]. Furthermore, previous studies suggested that charge retention in semiconductor NC memories is governed by interface traps surrounding NCs with deep energy level inside the Si bandgap [44, 45]. However, this mechanism is less reliable because the interface traps are subjected to many process variations such as the backend forming gas annealing, and there is no known method to reliably engineer deep-level traps. So are true for SONOS-type memories relying on bulk traps in dielectrics. Many studies have shown that retention at high temperature is problematic for SONOS with shallow-level traps [14, 20, 46]. In brief summary, the metal NC memory is a unique approach to further scale down the tunnel oxide without compromising retention. Therefore, the cell size scaling, low P/E voltage, and robust memory reliability may be realized simultaneously.

Electrostatics is another important consideration to achieve low P/E voltage. For better P/E efficiency, any potential drop on FGs has to be minimized especially with aggressive scaling on the thickness of tunnel and control oxide. In the conventional n-dope poly-Si FG, the poly depletion is present. In SONOS memories, the voltage drop on nitride is substantial because the nitride permittivity is only two times larger than SiO_2 and the thickness of nitride is comparable

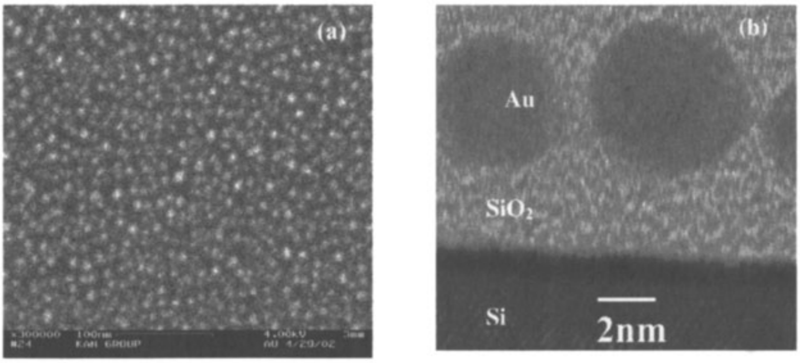

Figure 2 (a) A SEM plane-view image of Au NCs with area density of $4\times10^{11}/cm^2$, and (b) A STEM cross-sectional image of Au nanocrystals embedded in SiO_2 [34].

to oxide barriers. That is one of the reasons why higher-κ trap layers, such as Al_2O_3 [20], HfAlO [21], HfSiO [22], AlN [23], HfO_2 [24], and Ta_2O_5 [25], are more desirable. With the relatively higher Si permittivity and the small NC size, semiconductor NC memories seem to mitigate this concern. Nevertheless, as discussed in the later sections, the integration of high-κ dielectrics such as HfO_2 with κ = 20 as both the tunnel and control oxide makes the voltage drop on semiconductor NCs unavoidable. Therefore, metal NCs are the best option to eliminate the voltage drop with the orders of magnitude higher free electron concentration than the semiconductor counterparts. In addition, the above analysis is solely based on one-dimensional (1D) electrostatic approximation and too simplistic for NC memories because of the nature of the 3D spherical NCs and their two-dimensional (2D) placement. This is highlighted in the cross-sectional TEM and plain-view SEM in Fig. 2 [34]. Detail examination based on 3D electrostatics reveals the field-enhancement effects around NCs [47, 48]. For a typical design of metal NC memory, the potential drop in the tunnel oxide can be more than 40% higher than that in the continuous FG memory, resulting in great improvement on the P/E efficiency. This field enhancement is subject to not only geometrical parameters, many times being able to be solved only by numerical simulation, but also the choice of materials of NC and surrounding dielectric. Considering a simplified case when the top gate, the sensing channel, and other NCs are relatively far away and a NC with charge amount of Q stored is placed in a uniform field E_0, the analytical solution of the electric field intensity exists and can be expressed as [47, 48]:

$$E_r = E_0\left(1+\frac{2a^3}{r^3}\left(\frac{\varepsilon_{NC}-\varepsilon}{\varepsilon_{NC}+2\varepsilon}\right)\right)\cos\theta+\frac{\Sigma_i Q_i}{4\pi\varepsilon r^2} \quad (1)$$

$$E_\theta = -E_0\left(1-\frac{a^3}{r^3}\left(\frac{\varepsilon_{NC}-\varepsilon}{\varepsilon_{NC}+2\varepsilon}\right)\right)\sin\theta \quad (2)$$

where the origin of the spherical coordinate (r,θ) is at the center of the NC, a is the NC radius, ε_{NC} is the NC permittivity, ε is the dielectric permittivity, and θ is the angle between r and E_0. For a metal NC with infinite ε_{NC}, the field-enhancement term is reduced to $(a/r)^3$ even with high-κ dielectrics. On the other hand, for a Si NC with ε_{NC} = 11.7 embedded in SiO_2, the field-

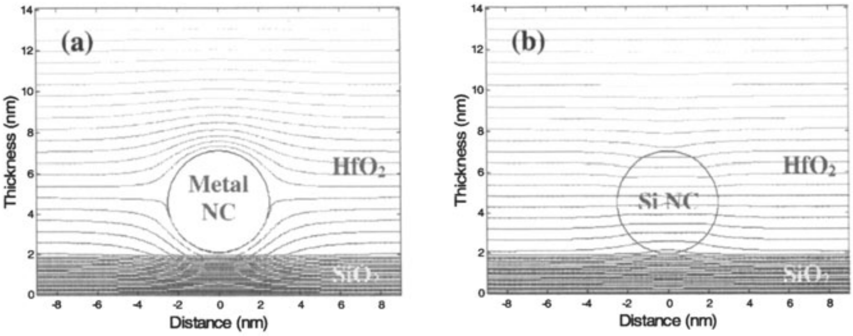

Figure 3 The cross-section view of the 3D electrostatic potential contours in NC memory unit cells with (a) a metal NC and (b) a Si NC. The NC diameter, the thickness of SiO_2 tunnel oxide, and the thickness of HfO_2 control oxide are 5 nm, 2 nm, and 7 nm, respectively. $V_G = 8$ V and no charge is stored in the NCs. The potential is monotonic from top to bottom, and the contour spacing is 0.2 V.

enhancement is merely $0.4\times(a/r)^3$ and it gets less or even becomes negative when embedded in a high-κ matrix. The numerical simulation of 3D potential contours in a unit cell of metal and semiconductor NC memories is shown in Fig. 3 with a high-κ HfO_2 control oxide. The potential drop inside the Si NC and the electric field decrease around it are in strong contrast with the unit cell of metal NC memory. Therefore, metal NCs are preferable choice over semiconductor NCs as the integration with high-κ dielectrics is inevitable for future scaled memory devices [49]. Meanwhile, due to the infinitesimal physical size of traps, the $(a/r)^3$ term vanishes. Therefore, SONOS memories remain similar to the conventional continuous FG memory without additional field-enhancement from 3D electrostatics.

Despite aforementioned advantages, the discreetness of FGs also poses fundamental challenges in maintaining P/E efficiency. First, the charging energy arising from shrinking capacitance in the discrete FGs becomes substantial. Single-electron charging energy E_{CH} is the electrostatic energy required to store an additional electron in a small capacitor due to the Coulomb repulsion energy. It can be expressed as e^2/C where e is the elemental charge and C is the self-capacitance of the charge storage node from the 3D electrostatic calculation. C is a strong function of the NC size, and E_{CH} increases dramatically with the NC size scaling [50]. Therefore, in a typical design with a NC diameter of 5 nm, the maximum number of charges every NC can stably hold is around 10. In the SONOS-type memory, because of the infinitesimal size of traps, every trap can hold at most one charge. As a result, to warrant sufficient memory window and P/E efficiency, the NC or trap density has to be deliberately engineered to provide charge storage capability comparable to the conventional continuous FG. One interesting example is by stacking multiple layers of NC vertically to provide additional storage capacity [33, 39, 40]. Furthermore, small physical size of discrete FGs, also true for the extreme cases of traps, associates with small capture cross-sectional area σ during charge injection. This may adversely affect P/E efficiency. Lastly, the partial coverage of NCs over the surface of the Si channel results in less control on the channel potential, *i.e.* less memory window. A channel-control factor R between 0 and 1 is usually adopted in comparison with a continuous FG cell with R equal to 1 [48]. A smart design to boost R without increasing NC density will be further

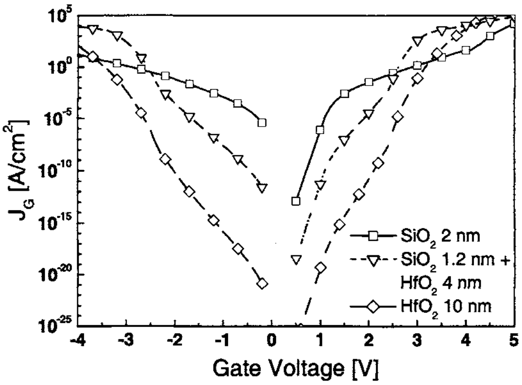

Figure 4 Calculated tunneling current from the WKB approximation for three tunnel oxide with the same 2-nm EOT.

discussed later by utilizing high-κ control oxide or small sensing channel. Overall NC with its moderate size provides low E_{CH}, large σ, and sufficient R, which greatly suppress adverse effects on P/E efficiency.

From the aspect of manufacturability, controlling tight V_{th} distribution at P/E states in a large memory array is very critical. It is a major drawback of the scaled NC memory cell as the fluctuation in the NC size and the NC number in each cell become substantial. However, it was projected that NC memory technology still has strong potential to scale beyond 65-nm node with current NC self assembly methods [14, 51, 52]. Recent efforts on ordered placement of NCs with controllable spacing of 3-15 nm [41, 42] may push the scaling limit even further. On the other hand, SONOS may provide better immunity to device variations owing to the large number density of traps. A heterogeneous NC/nitride stack may improve both V_{th} distribution and P/E efficiency for superior scalability [53, 54]. Both the semiconductor NC memory and the SONOS memory are fully compatible with the conventional Flash technology. They have been demonstrated for embedded applications to be fully compatible with CMOS, using even less masking steps compared with the embedded FG memory [16, 27]. High-κ trap layers and metal NCs are less compatible due to the concern of thermal stability and contamination. However, as high-κ dielectrics and metal gates become inevitable in the future Si technology, this may be less critical with better understanding and control on new material integration.

TUNNEL OXIDE

Tailoring the band structure of the tunnel barrier is another effective way to achieve significant tunneling asymmetry. High-κ dielectrics with lower electron / hole barriers are better field-sensitive tunnel barriers than SiO_2 [35]. In Fig. 4, tunneling current calculation based on Wentzel-Kramer-Brillouin (WKB) approximation [55] is shown for SiO_2 and HfO_2 with the same 2-nm effective oxide thickness (EOT). The WKB approximation of transmission probability T_{WKB} at the DT regime is expressed as:

$$T_{WKB} = \exp(-2\int_0^{t_{ox}} \sqrt{\Delta E - qF_{ox} \cdot x}\; dx) \qquad (3)$$

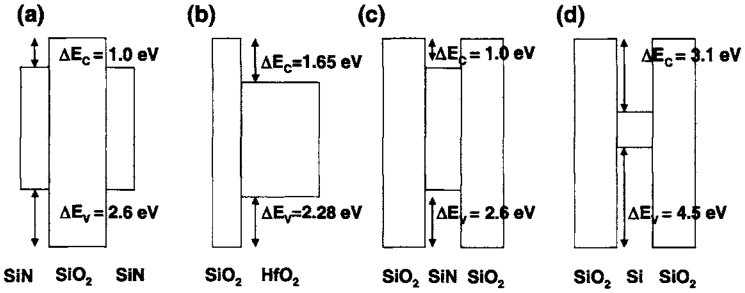

Figure 5 Energy band diagram representation of (a) crested barrier, (b) asymmetric layered barrier, (c) bandgap-engineered ONO, and (d) double tunnel junction.

where ΔE, F_{ox}, and t_{ox} are the dielectric / Si band offset, oxide electric field and oxide thickness, respectively. dT_{WKB}/dF_{ox} suggests T_{WKB} has stronger field dependence with smaller ΔE. Therefore, the current of HfO_2 has much stronger field dependence than that of SiO_2 at the DT regime. The lower transition voltage from DT to FN due to the lower ΔE further enhances the overall asymmetry. Composite tunnel barriers with multiple layers of dielectrics such as crested tunnel barriers [29, 56] and asymmetric layered barriers [49, 57, 58] as illustrated in Fig. 5 are designed by the same principle. In Fig. 4, a 1.2-nm SiO_2 + 4-nm HfO_2 with EOT of 2 nm exhibits similar field-sensitivity as a pure HfO_2 dielectric. The interfacial SiO_2 between high-κ dielectrics and the Si channel exists at many high-κ deposition processes, and also desirable to ease the severe mobility degradation caused by the remote phonon scattering [59] and reduce the interface traps that can affect cycling endurance.

The other class of field-sensitive tunnel barriers such as bandgap-engineered Oxide-Nitride-Oxide (ONO) [17], and double tunnel junction [19, 39] is also illustrated in Fig. 5. The structure consists of a small bandgap dielectric layer (SBL) sandwiched between two large bandgap dielectric layers (LBL). Resonant tunneling through the bound states at SBL is utilized to enhance the transmission probability at high field. However, this process is quenched at low field with bound state energy at SBL higher than the energy of injecting carriers. The only remaining transport is the DT current through the composite LBG/SBG/LBG, which is very low. Therefore, superior t_R / t_{PE} ratio at low P/E voltage has been demonstrated at highly scaled memory cells [39].

The employment of high-κ tunnel dielectrics is hampered by other disadvantages, such as mobility degradation in the channel and more importantly insufficient reliability caused by interface states D_{it} and dielectric traps. Transport mechanism of many high-κ dielectrics at low field is governed by the trap-assisted tunneling or interface-state assisted tunneling. Hence the large field-asymmetry estimated from an ideal high-κ dielectric is over optimistic. Furthermore, both natural and stress-induced traps in high-κ may degrade the cycling endurance and V_{th} distribution. However, through the advance of process technology, high-κ gate dielectrics have met strict reliability requirements for future CMOS [60]. Continuous P/E voltage scaling of Flash memory may eventually make reliable high-κ tunnel oxide feasible.

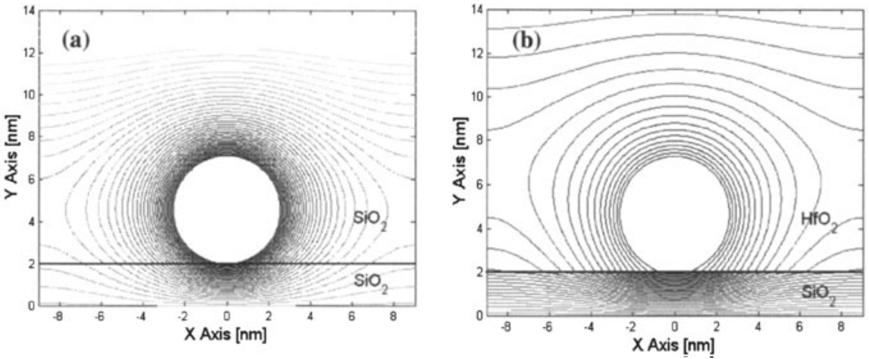

Figure 6 The cross-section view of the 3D electrostatic potential contours in the NC memory unit cell with (a) 7 nm SiO_2, and (b) 35 nm HfO_2 as the control oxide. Only part of the HfO_2 is shown in (b). The NC potential is set as -1 V while V_G = 0 V. The potential increases monotonically from the NC with the contour spacing of 25 mV [49].

CONTROL OXIDE & CONTROL GATE

Under P/E, the electric field at the control oxide increases significantly with the charge build-up at the FG, and so is the inter-poly leakage current. P/E saturation occurs when the inter-poly leakage current is comparable to the injecting current from the channel. As shown in Fig. 4, except under very high bias, high-κ dielectrics have substantially less leakage current than SiO_2 at the same EOT due to the physical thickness. Therefore, high-κ control oxide may be exploited to reduce the inter-poly leakage current and to increase CR simultaneously. This enables large memory window at lower P/E voltage or at higher P/E speed [61]. Combining with a metal electrode of high work function, the inter-poly current can be even further suppressed during erase [62].

In addition, spherical NCs are discretely placed on top of a 2D channel in NC memories. The coupling between NCs and the channel is subjected to 3D electrostatics. The detail of this coupling is important to determine the NC self capacitance, *i.e.* E_{CH}. It is also important to determine the channel-control factor *R*. Smaller E_{CH} allows more charges being stably stored in NCs, and larger *R* provides wider memory window where *R* = 1 representing the upper limit of a continuous FG. Both are critical in optimizing memory P/E and retention characteristics. The cross-sections of the 3D potential contours in the NC unit cell with SiO_2 and HfO_2 control oxide are plotted in Fig. 6. The EOT remains the same for both stacks. It is obvious that the fringing field through HfO_2 to the Si channel is much stronger due to the higher permittivity of HfO_2. As a result, E_{CH} with HfO_2 is only a half of that with SiO_2, and *R* increases from 0.55 to 0.85 dramatically. This leads to the increase of t_R / t_{PE} ratio by more than four orders of magnitude [49].

Fermi-level pinning is known to shift the effective gate work functions of metal/high-κ and polysilicon/high-κ gate stacks substantially [63]. Similar effects have been found critical in NC memories integrated with high-κ control oxide [64]. The effective NC work function is not only a bulk property of the NC, but also governed by the interface with the surrounding dielectric