Solid-State Electronics 65–66 (2011) 184–190

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Quasi-planar bulk CMOS technology for improved SRAM scalability

Changhwan Shin [a,*], Chen Hua Tsai [b], Mei Hsuan Wu [b], Chung Fu Chang [b], You Ren Liu [b], Chih Yang Kao [b], Guan Shyan Lin [b], Kai Ling Chiu [b], Chuan-Shian Fu [b], Cheng-tzung Tsai [b], Chia Wen Liang [b], Borivoje Nikolić [a], Tsu-Jae King Liu [a]

[a] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA
[b] United Microelectronics Corporation, Hsinchu, Taiwan, ROC

ARTICLE INFO

Article history:
Available online 18 July 2011

Keywords:
Variability
MOSFET
SRAM
CMOS

ABSTRACT

A simple approach for manufacturing quasi-planar bulk MOSFET structures is demonstrated and shown to be effective not only for improving device performance but also for reducing variation in 6T-SRAM read and write margins, in an early 28 nm CMOS technology. With optimization of the pocket implant doses, voltage scaling is facilitated. Since its benefits increase with decreasing channel width, quasi-planar bulk MOSFET technology should be advantageous for future CMOS technology generations (22 nm and beyond).

1. Introduction

Following Moore's Law, transistor density has roughly doubled with each new CMOS technology generation largely due to the steady miniaturization of the transistor. Variation in transistor threshold voltage (V_T) due to random dopant fluctuations and line-edge-roughness [1] and gate work-function variation [2] become more significant as the transistor gate length (L_G) is reduced below 30 nm, so that continued transistor scaling poses a growing challenge, particularly for static random-access memory (SRAM) arrays which typically employ the smallest transistors and have the most stringent yield requirement [3]. V_T mismatch makes it difficult to lower the SRAM operating voltage [4], so that increasing power density has become a critical issue. Therefore, an improved transistor design that provides for reduced short-channel effects (*i.e.* improved gate control over the channel potential) and hence reduced V_T sensitivity to process-induced variations is needed to facilitate voltage scaling. Examples include the fully depleted silicon-on-insulator (FD-SOI) MOSFET with thin buried-oxide (thin-BOX) [5] and multiple-gate transistor structures (*e.g.* FinFET, MuGFET, Tri-Gate FET) [6]; but these require either expensive SOI substrates and/or more complex fabrication processes that pose significant barriers to their widespread adoption. Recently, a low-cost quasi-planar bulk CMOS technology was proposed and demonstrated to provide for improved performance and reduced variability [7,8]. In contrast with FinFET/MuGFET/Tri-Gate FET structures which employ a narrow body region to suppress short-channel effects, the quasi-planar bulk MOSFET structure uses the conventional retrograde channel doping of the planar bulk MOSFET structure to suppress leakage current, in addition to a quasi-planar gate electrode and gate fringing electric fields, to achieve improved gate control.

This paper presents more details of the study of quasi-planar bulk CMOS technology for improved SRAM scalability [8]. In Section 2, the device fabrication process is described. In Section 3, the benefits of the quasi-planar MOSFET design for improving transistor performance and reducing variability to improve SRAM yield are presented. Section 4 presents the conclusions from this study.

2. Device fabrication

(1 0 0) epi-Si wafers were used as the starting substrates for fabricating MOSFETs with ⟨1 1 0⟩ channel orientation in an early 28 nm-generation bulk CMOS logic technology. The sequence of front-end-of-line fabrication process steps is outlined in Fig. 1. After conventional shallow-trench-isolation (STI) processing, N/P well and V_T-adjust ion implantation steps were performed, followed by high-temperature rapid thermal annealing (RTA). Subsequently, dilute hydrofluoric acid (DHF) was used to remove residual sacrificial oxide, as well as to recess the STI oxide by a small amount (15 nm) prior to gate stack formation to achieve quasi-planar MOSFETs. A shorter DHF dip was used for the control (planar MOSFET) devices. The gate stack was formed by plasma nitridation of a thermal oxide layer of 1.45 nm physical thickness followed by deposition of an undoped polycrystalline silicon layer of 70 nm thickness. To define the gate electrodes with tight control of physical gate length (as small as 30 nm) for logic transistors and

[*] Corresponding author. Tel.: +1 510 664 4202; fax: +1 510 643 2636.
E-mail address: shinch@eecs.berkeley.edu (C. Shin).

0038-1101/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sse.2011.06.022

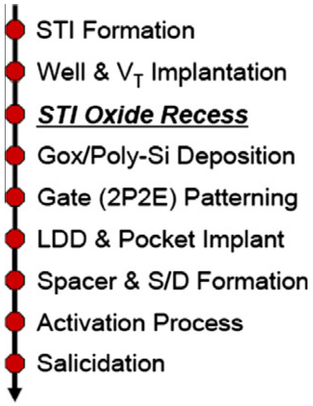

Fig. 1. Sequence of front-end-of-line CMOS fabrication process steps used to fabricate logic devices and SRAM arrays in this work.

0.149 μm^2 6-T SRAM bit cells, a double-patterning/double-etch (2P2E) process employing 193 nm immersion lithography and advanced hard-mask etching techniques was used. After gate stack patterning, pocket ion implantation was performed. An experimental split was included to explore lighter pocket doping, in which the implant dose was lowered by 10^{13} cm^{-2}. Gate-sidewall spacers were formed prior to source/drain ion implantation. To activate the implanted dopants, a rapid thermal process (RTP) followed by laser spike annealing (LSA) was used to enhance the electrical conductivity in the source/drain regions. Afterwards, a nickel silicidation (NiSi) process was applied. Subsequently, dual contact etch stop layers (CESL) of SiN_x – highly compressive stress liner for PMOS devices, and highly tensile stress liner for NMOS devices – for performance enhancement were formed by plasma-enhanced chemical vapor deposition (PECVD). After interlayer dielectric (ILD) oxide deposition, contact hole definition, tungsten plug formation and chemical mechanical planarization (CMP), a standard copper metal interconnection process was followed.

A standard test-chip mask set was used to fabricate individual logic transistors and 6T-SRAM arrays, ~2500 cells per device-under-test (DUT). Fig. 2 shows plan-view scanning electron microscopy and cross-sectional transmission electron microscopy images of a fabricated SRAM cell.

3. Results and discussion

3.1. Quasi-planar vs. planar MOSFETs

3.1.1. Improved performance

Due to improved gate control and increased effective channel width, quasi-planar MOSFETs (in which the STI oxide is recessed by 15 nm) have higher on-state drive current (I_{ON}) for comparable off-state leakage current (I_{OFF}), as shown in Fig. 3. Lower pocket doping results in lower V_T as well as higher average effective mobility, and hence even higher I_{ON}. Because the benefit of sidewall gating increases as the layout width decreases, the pass-gate (PG) devices show greater improvement (2.4×) in I_{ON} than the pull-down (PD) devices (2.1× improvement). The performance enhancement (4.5×) is greatest for the PMOS devices not only because they have the narrowest layout width and but also because hole mobility is higher for the (1 1 0) sidewall channel surfaces, whereas electron mobility is lower [9].

3.1.2. Suppressed V_T variation

V_T statistics are shown in Fig. 4 for the PG/PD/PU devices. Improved gate control results in steeper subthreshold swing and hence lower V_T for the quasi-planar MOSFETs. In this early 28 nm CMOS process, the standard pocket implant dose is relatively high for the n-channel devices. As a result, variation in V_T is slightly larger for the quasi-planar PG and PD devices, due to more significant impact of random dopant fluctuations (RDF) for the gated sidewalls. This undesirable effect is eliminated by using a lighter pocket implant dose, as shown in Fig. 4a and b, which further lowers V_T without significantly increasing I_{OFF} (See Fig. 3d and e). The standard pocket implant dose is lower for the p-channel devices, so that the impact of RDF for the gated sidewalls is not an issue. Thus, PMOS V_T variation is reduced when the STI oxide is recessed, due to the superior electrostatic integrity of the quasi-planar structure (Fig. 4c). If an even lighter pocket implant dose is used, then V_T variation is slightly larger due to degraded short-channel effect. In short, V_T variation in quasi-planar devices can be lower than in planar devices if the channel/pocket doping level is optimized.

Pelgrom plots [10] showing how V_T variation increases with decreasing channel area, for logic devices, are shown in Fig. 5. Pelgrom's coefficient (A_{VT}) is reduced by 8% and 7% for the NMOS and PMOS quasi-planar devices with lower pocket doping, respectively. This improvement is consistent with the SRAM device results shown in Fig. 4.

3.1.3. Improved short-channel effect

Fig. 6 shows the short-channel effect for logic devices with 250 nm drawn width. It can be seen that V_T roll-off is reduced for the quasi-planar structures, even though the channel is much

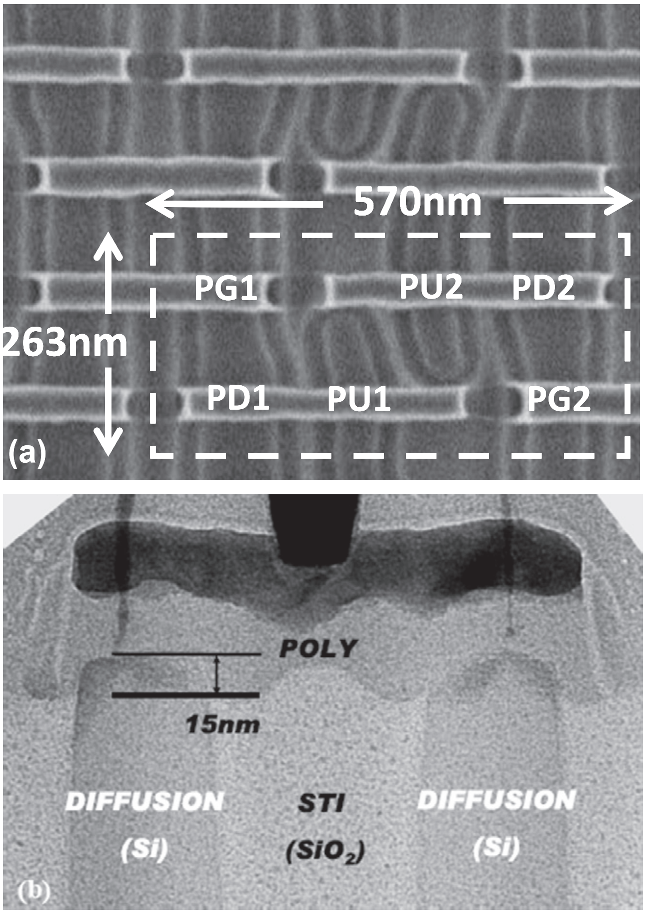

Fig. 2. (a) 0.149 μm^2 SRAM cell plan-view CDSEM image after gate patterning. (b) XTEM taken along a poly-Si gate electrode in an SRAM array, for 15 nm nominal STI-oxide recess depth.

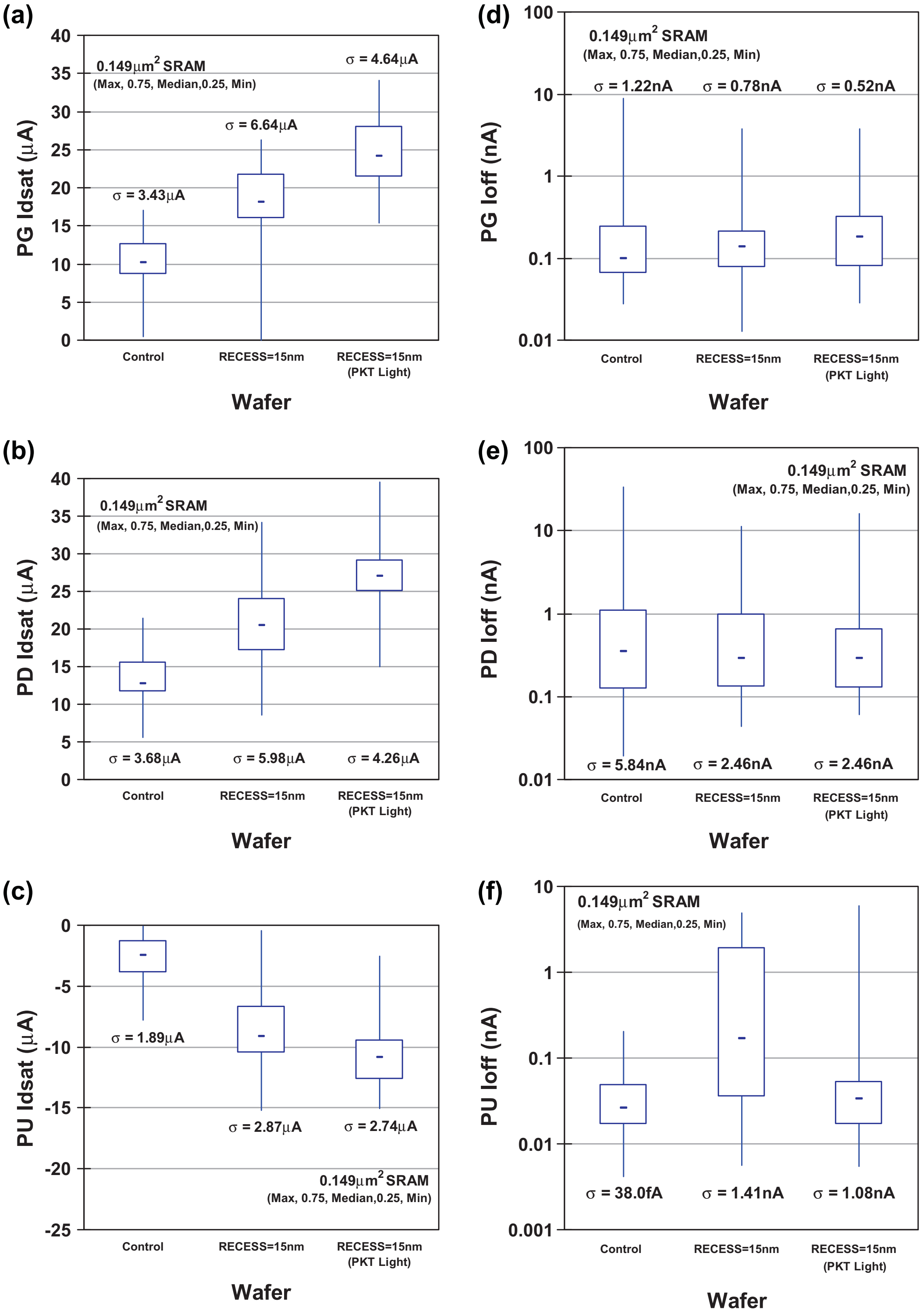

Fig. 3. Comparison of ON/OFF current statistics for planar (control) *vs.* quasi-planar (recess = 15 nm) bulk MOSFETs in SRAM cells. (a) pass-gate NMOS I_{ON} (b) pull-down NMOS I_{ON} (c) pull-up PMOS I_{ON} (d) pass-gate NMOS I_{OFF} (e) pull-down NMOS I_{OFF} (f) pull-up PMOS I_{OFF}.

wider (by >16×) than the STI oxide recess depth. Reasonable short-channel control is maintained by the quasi-planar structure even with lighter pocket doping.

3.1.4. Increased narrow width effect

The reverse narrow width effect, *i.e.* V_T reduction with decreasing channel width (W), stems from increased gate control for

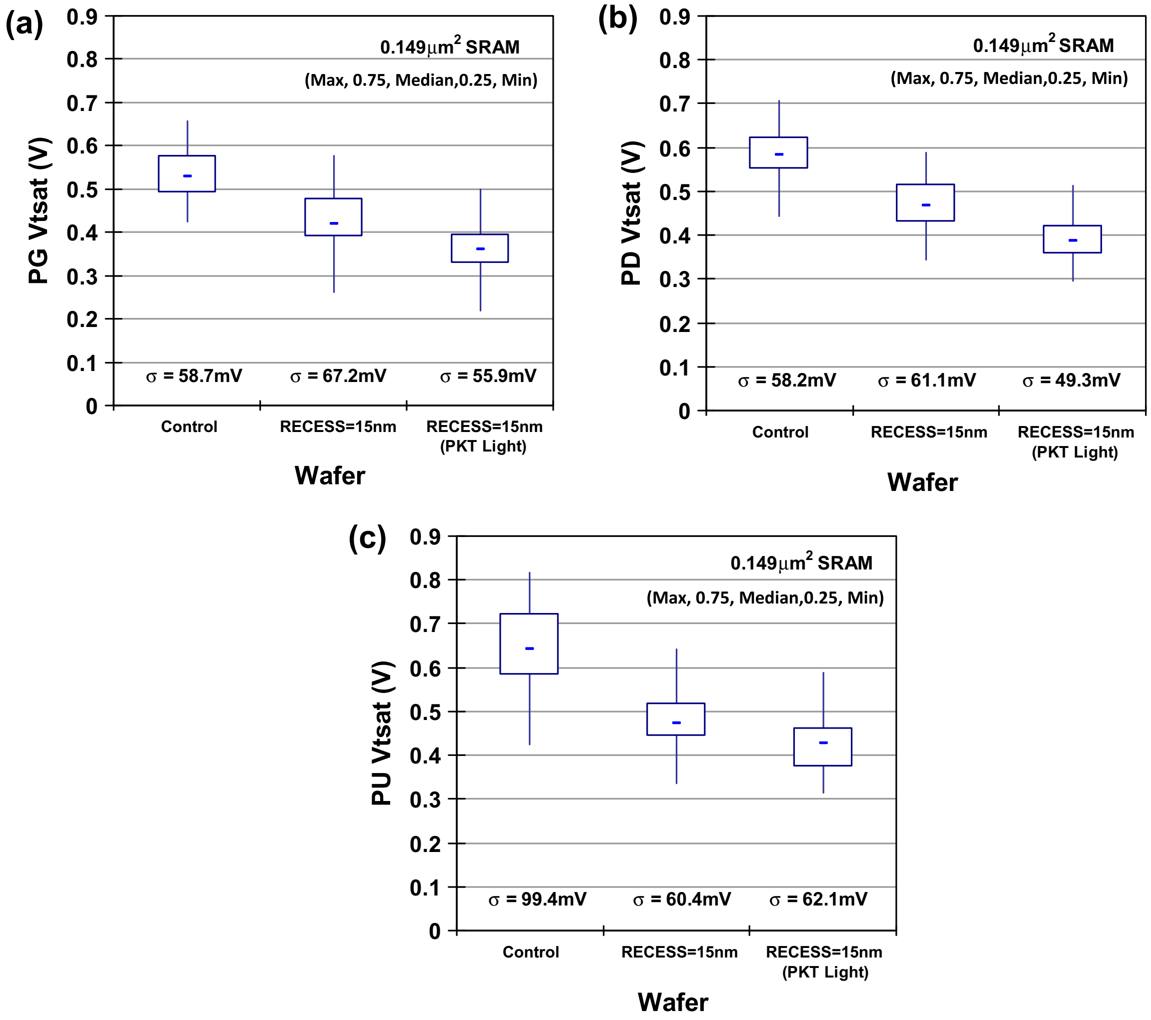

Fig. 4. Comparison of saturation V_T statistics for planar (control) *vs.* quasi-planar (recess = 15 nm) bulk MOSFETs in SRAM cells: (a) pass-gate NMOS, (b) pull-down NMOS, (c) pull-up PMOS.

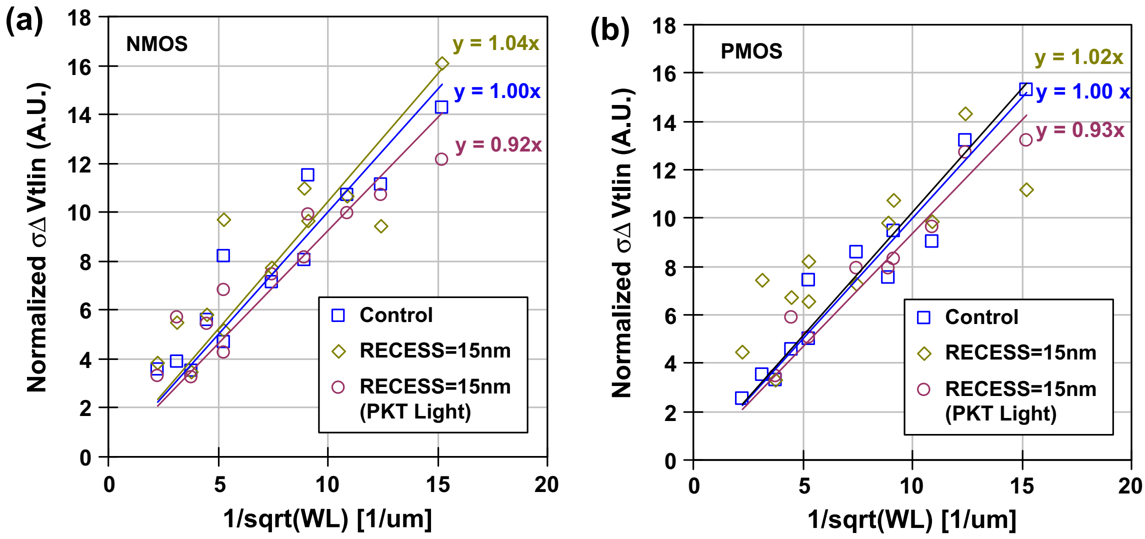

Fig. 5. Pelgrom plots for (a) NMOS and (b) PMOS logic devices with drawn width ranging from 120 nm to 1 μm and drawn gate length ranging from 36 nm to 0.2 μm.

narrower channel width due to fringing electric fields between the gate electrode and channel sidewalls. This effect is intensified in quasi-planar devices, as shown in Fig. 7. It should be noted that, overall, V_T variation is lower for quasi-planar devices – even with

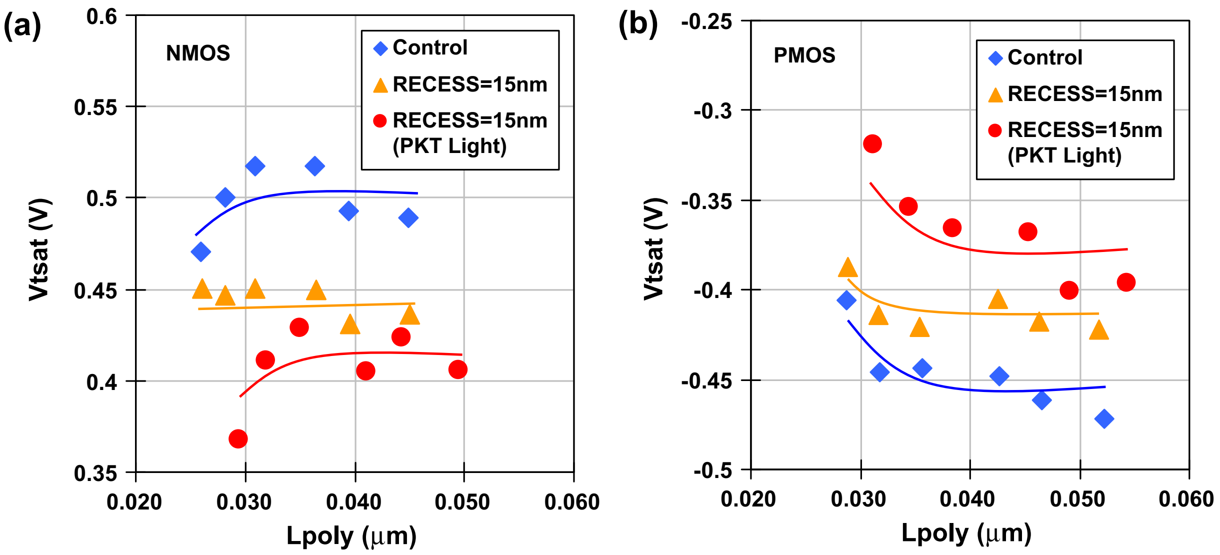

Fig. 6. Saturation threshold voltage with decreasing gate length, for logic devices with 0.25 µm drawn width. (a) NMOS and (b) PMOS.

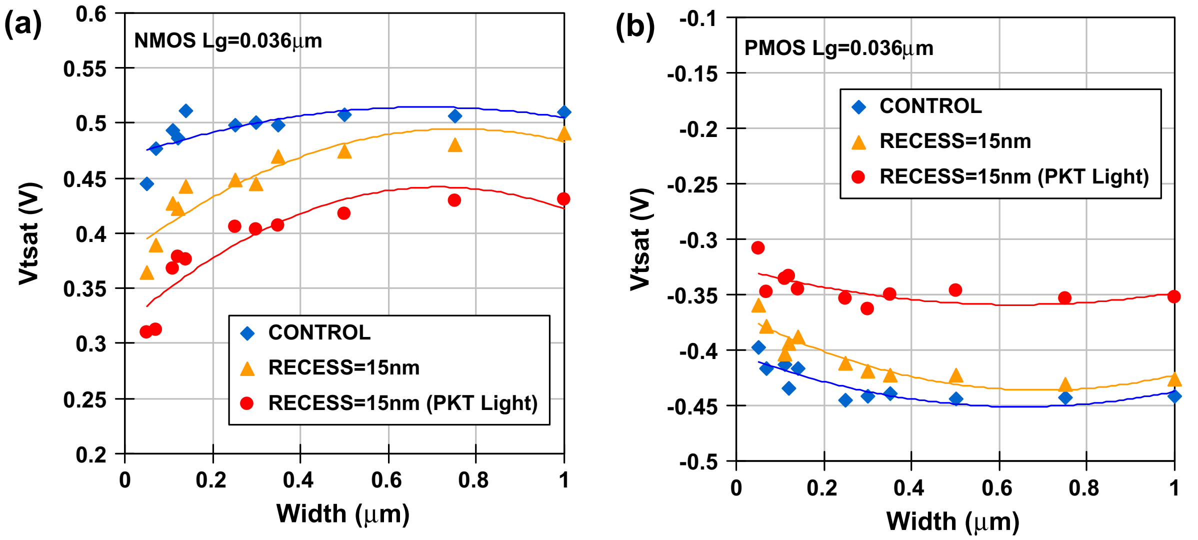

Fig. 7. Measured reverse narrow width effect for devices with 36 nm gate length: (a) NMOS and (b) PMOS. Median V_T is lower when the STI oxide is recessed, due to improved gate control over the channel potential.

reduced pocket doping – than for the planar devices, due to improved short channel control.

To maximize the benefits of quasi-planar CMOS technology, wider transistors should be segmented into stripes of uniform width less than or equal to ~$2L_G$ [11]. A double-patterning approach [12] similar to that used for gate patterning in state-of-the-art CMOS processes can be used to form channel segments of highly uniform width without the need for forming high-aspect-ratio isolation trenches.

3.1.5. Compact transistor model

One of the advantages of quasi-planar CMOS technology over FD-SOI and FinFET/MuGFET/Tri-Gate FET technologies is that it is compatible with standard bulk MOSFET compact models used for circuit design. In this work, the BSIM4.6 compact model was calibrated to the electrical characteristics of quasi-planar bulk MOSFETs, with fitting parameters including electrical and physical gate-oxide thickness, gate length offset, and the number of fingers in the device. Fig. 8 shows that the compact model can be well-fitted to quasi-planar bulk MOSFET characteristics, including the body effect. This illustrates another advantage of quasi-planar CMOS technology, which is that it allows for adaptive body biasing, *i.e.* dynamic optimization of the trade-off between performance (delay) and power consumption (energy).

3.2. Benefits of quasi-planar bulk CMOS technology for 6T-SRAM

3.2.1. Cell yield enhancement

In this early 28 nm CMOS technology, SRAM yield (gauged by 3-sigma/mean values for DC read and write noise margins, SNM and WRM, respectively) was slightly diminished by recessing the STI oxide, because of the aforementioned increase in NMOS V_T variation (Ref. Fig. 4a and b). If lighter pocket doping is used, however, variability is reduced so that yield is superior for the quasi-planar CMOS technology, as shown in Fig. 9. The nominal SNM is degraded by recessing the STI oxide because the cell beta ratio is degraded

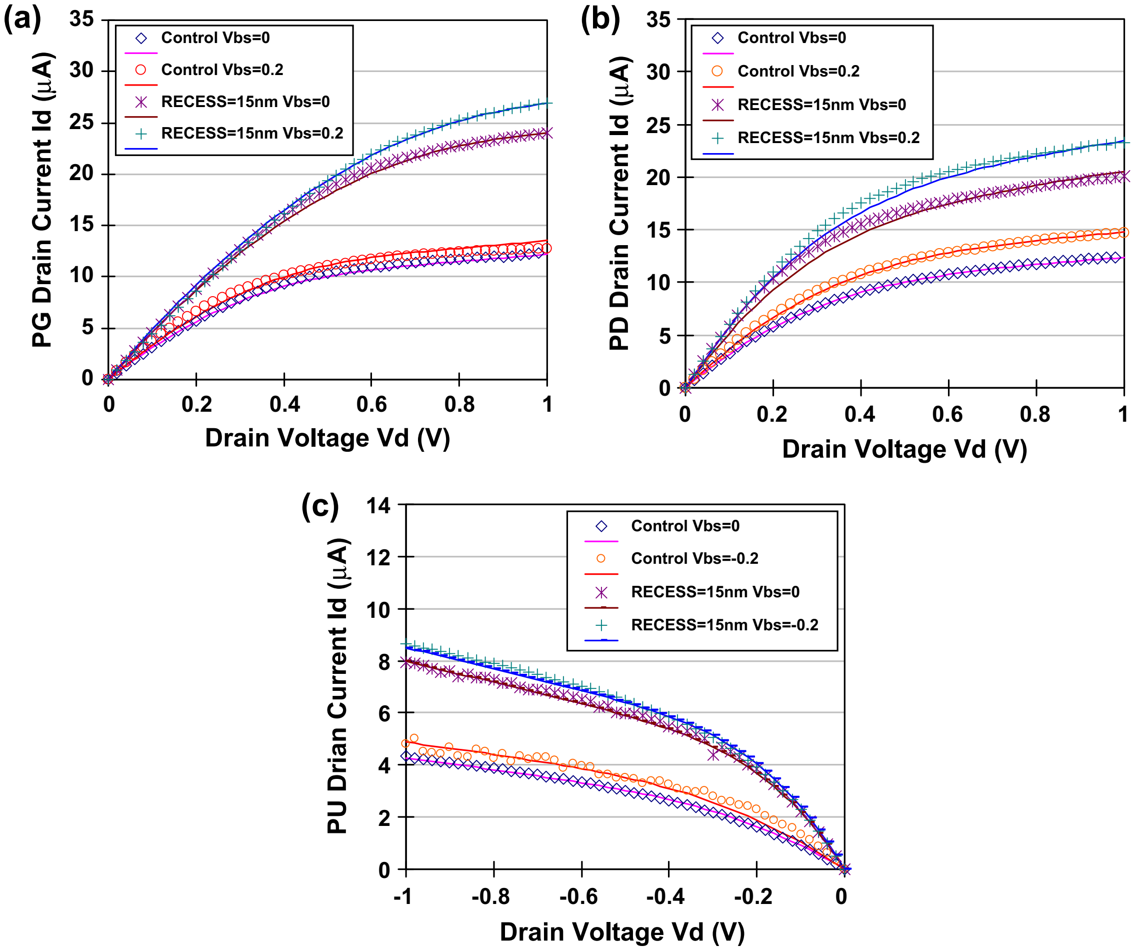

Fig. 8. Comparison of measured output characteristics for planar (control) *vs.* quasi-planar (recess = 15 nm) bulk MOSFETs in SRAM cells, for $|V_{GS}| = 1.0$ V. The effect of forward body biasing is also shown. (a) Pass-gate NMOS, (b) pull-down NMOS, and (c) pull-up PMOS. The symbols are measured data; the lines show the fitted compact model.

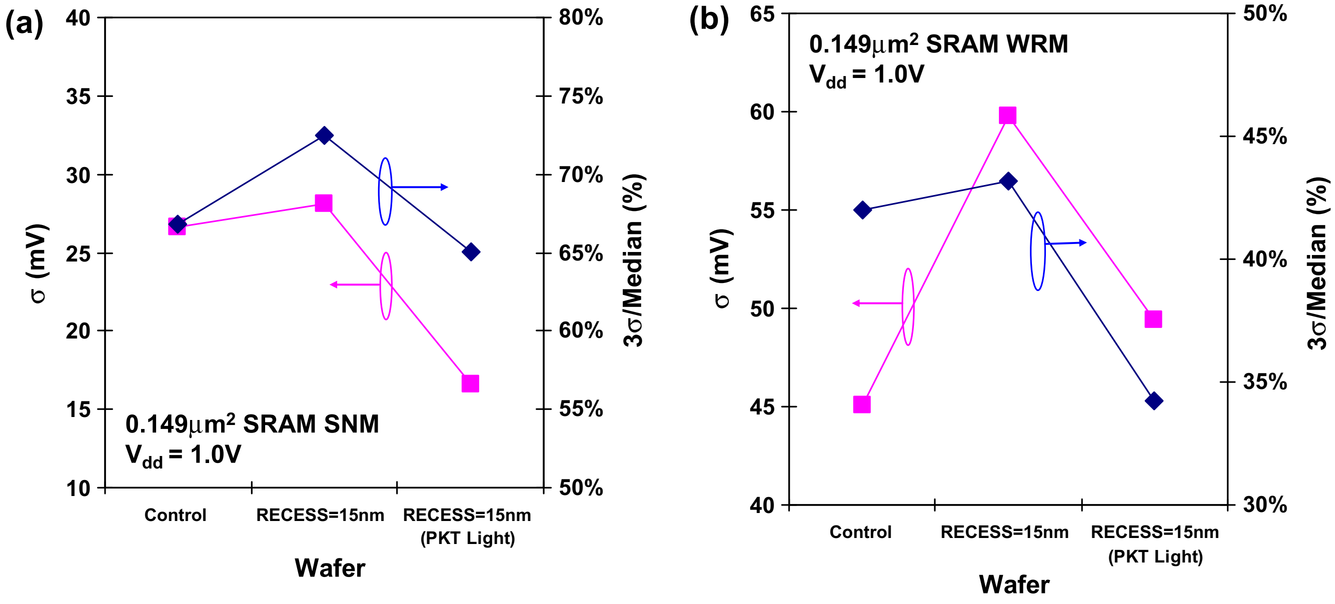

Fig. 9. Sigma and 3-sigma/median values for (a) read margin (SNM) and (b) write margin (WRM). $V_{dd} = 1.0$ V.

(even though the drive strength of the PU device is improved), which is why the increase in 3sigma/median is larger than for sigma (Fig. 9a). The fact that nominal WRM is improved by recessing the STI oxide accounts for the observation that 3sigma/median

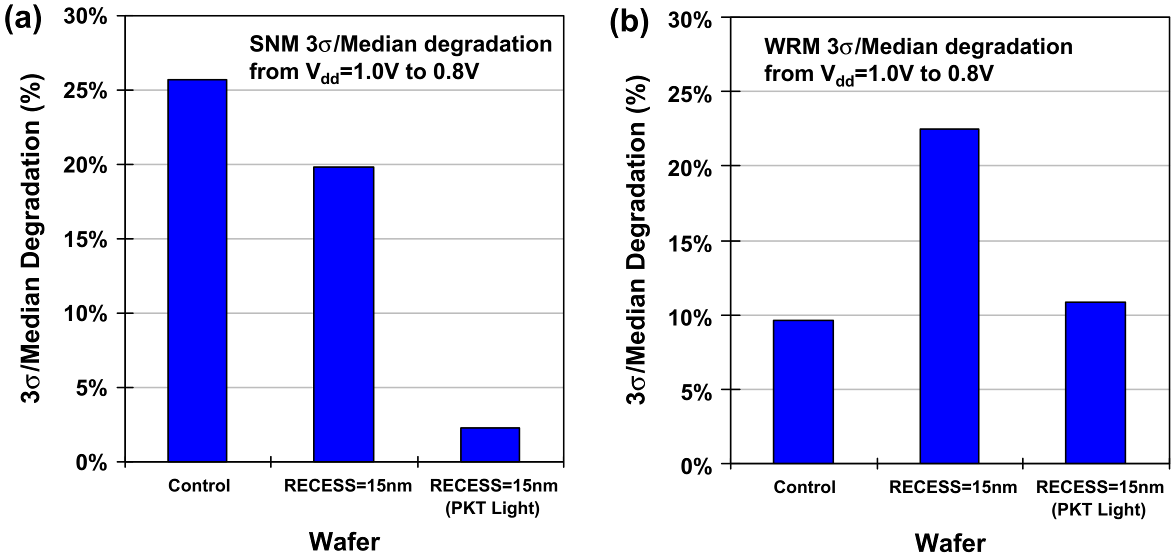

Fig. 10. Degradation in 3-sigma/median for (a) SNM and (b) WRM as V_{dd} is reduced from 1.0 V to 0.8 V.

does not increase by very much, even though sigma increases significantly due to increased PG V_T variation (Fig. 9b).

3.2.2. Voltage reduction

Supply-voltage (V_{dd}) reduction is desirable to reduce power density and to facilitate increased transistor density. Generally, however, relative variability increases as the gate overdrive ($V_{dd} - V_T$) decreases, so that yield is degraded. Fig. 10 shows that the degradation in SNM yield with V_{dd} reduction (from 1.0 V to 0.8 V) is dramatically reduced for quasi-planar bulk CMOS technology with reduced pocket doping, while the degradation in WRM yield with V_{dd} reduction is not significantly worse for quasi-planar bulk CMOS technology. With separately optimized pocket implant doses for the NMOS and PMOS devices, reduced degradation in both SNM yield and WRM yield with V_{dd} scaling can be achieved.

4. Conclusion

With optimized pocket doping, quasi-planar MOSFETs achieved by slightly recessing the STI oxide prior to gate-stack formation in an otherwise conventional CMOS fabrication process can provide for improved performance and reduced variability, and hence can facilitate the scaling of SRAM operating voltage. The benefits of the quasi-planar bulk MOSFET design increase with decreasing channel width, so that quasi-planar CMOS technology is a compelling solution for future generations (22 nm and beyond).

Acknowledgments

This work was supported in part by the Center for Circuit & System Solutions (C2S2) Focus Center, one of six research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation program. C. Shin appreciates the support of the Korea Foundation for Advanced Studies (KFAS) and helpful comments from Dr. Reinaldo Vega.

References

[1] Asenov A, Brown AR, Davies JH, Kaya S, Slavcheva G. Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs. IEEE Trans Elect Dev 2003;50(9):1837–52.
[2] Dadgour H, Endo K, De V, Banerjee K, Modeling and analysis of grain-orientation effects in emerging metal-gate devices and implications for SRAM reliability. In: IEDM tech dig; December. 2008, p. 705–8.
[3] Bowman KA, Tang X, Eble JC, Meindl JD. Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance. IEEE J SolidState Circ 2000;35(8):1186–93.
[4] Nii K, Yabuuchi M, Tsukamoto Y, Ohbayashi S, Oda Y, Usui K, et al. A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment. In: VLSI symp circuit dig, June. 2008, p. 212–3.
[5] Fenouillet-Beranger C, Denorme S, Perreau P, Buj C, Faynot O, Andrieu F, et al. FDSOI devices with thin BOX and ground plane integration for 32 nm node and below. Solid State Electron 2009;53(7):730–4.
[6] Kawasaki H, Khater M, Guillorn M, Fuller N, Chang J, Kanakasabapathy S, et al. Demonstration of highly scaled FinFET SRAM cells with high-K/metal gate and investigation of characteristic variability for the 32 nm node and beyond. In: IEDM tech dig; 2008, p 237–40.
[7] Tsai CH, King Liu T-J, Tsai SH, Chang CF, Tseng YM, Liao R, et al. Segmented tri-gate CMOS technology for device variability improvement, In: Proc IEEE VLSI-TSA; April 2010, p. 114–5.
[8] Shin C, Tsai CH, Wu MH, Chang CF, Liu YR, Kao CY, et al. Tri-gate bulk CMOS technology for improved SRAM scalability. In: Proc. IEEE european solid-state device research conf; September 2010. p. 142–5.
[9] Chang L, Ieong M, Yang M. CMOS circuit performance enhancement by surface orientation optimization. IEEE Trans Electron Dev 2004;51(10):1621–7.
[10] Pelgrom MJM, Duinmaijer A, Welbers A. Matching properties of MOS transistors. IEEE J SolidState Circ 1989;24(5):1433–40.
[11] Sun X, Liu Q, Moroz V, Takeuchi H, Gebara G, Wetzel J, et al. Tri-gate bulk MOSFET design for CMOS scaling to the end of the roadmap. IEEE Electron Dev Lett 2008;29(5):491–3.
[12] US patent 7190,050.