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

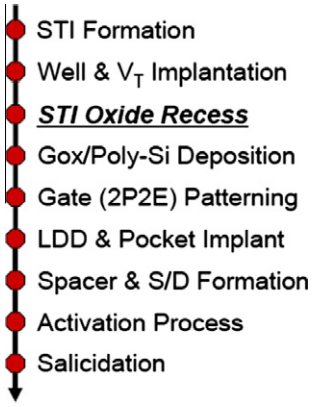

Fig. 1. Sequence of front-end-of-line CMOS fabrication process steps used to fabricate logic devices and SRAM arrays in this work.

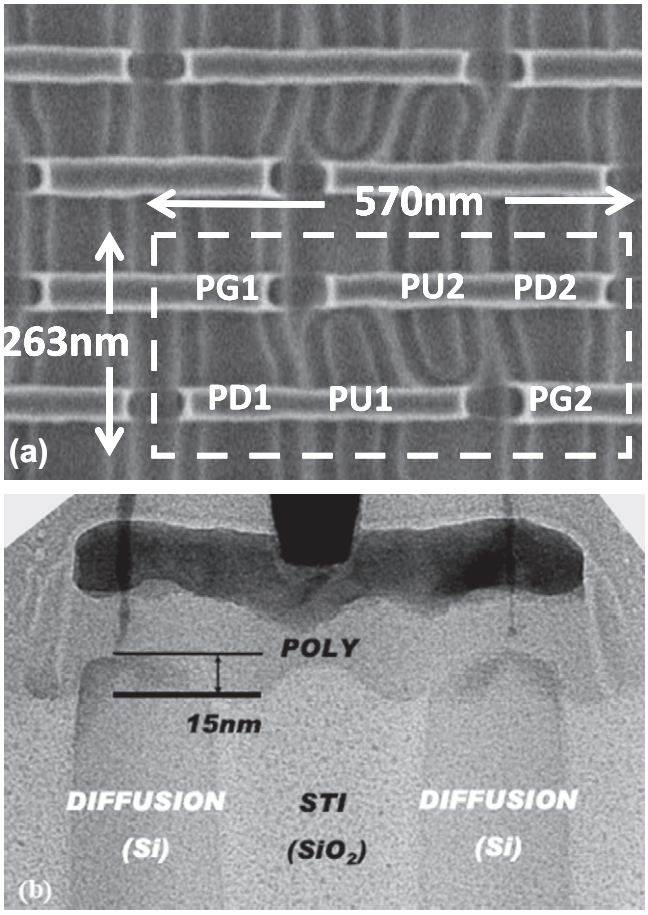

Fig. 2. (a) 0.149 μm^2 SRAM cell plan-view CDSEM image after gate patterning. (b) XTEM taken along a poly-Si gate electrode in an SRAM array, for 15 nm nominal STI-oxide recess depth.

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