Materials Science and Engineering B 154–155 (2008) 3–13

Contents lists available at ScienceDirect

Materials Science and Engineering B

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

Advanced activation trends for boron and arsenic by combinations of single, multiple flash anneals and spike rapid thermal annealing

W. Lerch[a,*], S. Paul[a], J. Niess[a], S. McCoy[b], J. Gelpey[b], F. Cristiano[c], F. Severac[c], P. Fazzini[d], A. Martinez-Limia[e], P. Pichler[e], H. Kheyrandish[f], D. Bolze[g]

[a] Mattson Thermal Products GmbH, Daimlerstrasse 10, 89160 Dornstadt, Germany
[b] Mattson Technology Canada Inc., 605 W. Kent Avenue, Vancouver, BC V6P6T7 Canada
[c] LAAS/CNRS, 7 av. Du Col. Roche, 31077 Toulouse, France
[d] CEMES/CNRS, 29 rue G. Marvig, 31055 Toulouse, France
[e] Fraunhofer-IISB, Schottkystrasse 10, 91058 Erlangen, Germany
[f] CSMA–MATS, Queens Road, Stoke on Trent ST4 7LQ, United Kingdom
[g] IHP, Im Technologiepark 25, 15326 Frankfurt (Oder), Germany

ARTICLE INFO

Article history:
Received 16 July 2008
Received in revised form 4 August 2008
Accepted 25 August 2008

Keywords:
Flash annealing
Spike RTA
Multiple flash annealing
Combination flash annealing and spike RTA
Combination spike RTA and flash annealing
Hall-effect measurements
Hall mobility measurement
Defect evolution
Boron
Arsenic
PMOS
NMOS
Crystalline Si
Amorphous Si
OED
Reactive gaseous ambient
Advanced activation thermal cycles
Dopant simulation

ABSTRACT

Millisecond annealing as an equipment technology provides temperature profiles which favour dopant activation but nearly eliminate dopant diffusion to form extremely shallow, highly electrically activated junctions. For the 45-nm technology node and beyond, precisely controlled gate under-diffusion is required for optimum device performance. Therefore, on boron and arsenic beamline-implanted wafers, various annealing schemes were investigated for the formation of ultra-shallow and custom-shaped junctions. The main scheme consisted of flash annealing with peak temperatures ranging from 1250 to 1300 °C, combined with spike rapid thermal annealing with peak temperatures in the range from 900 to 1000 °C to achieve a desired junction depth. As alternative, to reduce the sheet resistance of pMOS and nMOS source–drain extensions, combinations of two or three flash anneals in succession were tested. Finally, the standard flash anneal condition of a 750 °C intermediate temperature followed by the flash anneal was changed to a high intermediate temperature of 950 °C followed by the flash anneal up to 1300 °C. The results of all these annealing schemes were analysed by four-point probe measurement. Selected samples were analysed by Hall-effect measurements for peak activation, and by secondary ion mass spectrometry for profile shape as well as diffusion effects. Transmission electron microscopy was used to study residual defects. Selected boron and arsenic dopant profiles were also compared to predictive simulation results which address the diffusion and activation at extrinsic concentrations.

1. Introduction

In the wake of the continuing miniaturization of devices, there is a more and more pressing need towards shallower and steeper junctions with higher levels of activation to reduce short-channel effects. This is reflected on the one hand by a reduction of implant energies which are now in the sub-1 keV regime. On the other hand, to reduce diffusive broadening, the process times of source/drain annealings and source/drain extension annealings have reduced drastically from soak annealing for several seconds via high ramp-up spike rapid thermal annealing (RTA) to millisecond annealing (MSA). At the same time, increasing peak temperatures were used to raise the dopant activation. Today, MSA either in the form of flash lamp [1] or laser annealing [2] is applied to complementary MOSFETs (CMOS) as the activation method of choice for implanted impurities in source and drain regions to fulfil the above requirements. A simple replacement of conventional RTA by MSA can cause problems because insufficient gate under-diffusion during the activation process results in a reduction of the drive current. On the other hand, MSA reduces the gate depletion in poly silicon [3,4] and

* Corresponding author. Tel.: +49 7348 981 234; fax: +49 7348 981 239.
E-mail address: wilfried.lerch@mattson.com (W. Lerch).

0921-5107/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.mseb.2008.08.017

provides technological advantages like higher solid solubility and, therefore, better activation. One technique to overcome the restrictions is to perform source/drain extension (SDE) implantation with high tilt to keep the overlap. However, this approach has potential disadvantages like increase in gate leakage current and degradation of gate oxide reliability due to implantation damage. The other solution is to control the overlap/gate under-diffusion by ordinary dopant diffusion prior to the highly activating diffusion-less MSA process. Here and in the following, ordinary diffusion refers to Fermi-level dependent diffusion with the intrinsic point defects in equilibrium or near to equilibrium. Other diffusion effects like transient enhanced diffusion (TED) or enhanced diffusion due to reactive gaseous ambient should be well controlled or suppressed.

One of the first papers reporting on spike RTA and flash lamp annealing was from Toshiba using Xe flash lamps [5]. Ito et al. showed for a 50-nm gate length MOSFET superior device characteristics for both the nMOS and the pMOS with this annealing sequence. Chen et al. reported that the technique of using a spike RTA combined with a so-called "super anneal" (either flash lamp or laser anneal) is an alternative solution to overcome the complicated disposable spacer approach when only the "super anneal" is used and shows also advanced device parameters for sub-65 nm designs [6]. But this sequence is not the only possibility. In the literature a variety of papers can be found which show that MSA annealing prior to spike RTA has advantages for the final device performance [7], especially if the MSA approach is non-melt laser based [3,8–12]. Furthermore, a repetition of the MSA either at various positions in the process flows with other steps in between or directly subsequent to each other improves the electrical performance [13,14]. All these annealing schemes are based on the evaluation of electrical device data. In this contribution, the different annealing schemes based on flash lamp MSA and spike RTA are evaluated. Especially single and multiple flash annealing processes, and the sequences of spike RTA and flash MSA and flash MSA and spike RTA were studied on bare silicon wafers implanted with either boron or arsenic to understand the activation mechanisms as well as the evolution of extended defects. Finally, suggestions based on dopant activation and defect structure/dissolution for the integration of implant and anneal receipts into complementary MOSFETs are given.

2. Experimental details

For the experiments, 200 mm prime (1 0 0) Si wafers were used. The resistivity of the n-type wafers for the boron experiments was 10–20 Ω cm, that of the p-type wafers for the arsenic experiments 15–20 Ω cm. Boron was implanted as $^{11}B^+$ (0.5 keV, 1.0×10^{15} cm^{-2}) either into a $^{74}Ge^+$ pre-amorphized layer (30 keV, 1.0×10^{15} cm^{-2}) or into crystalline silicon (c-Si). In the pre-amorphized wafers, the upper edge of the end-of-range damage (EOR) after annealing was at 50 nm. Arsenic was implanted as $^{75}As^+$ (1 keV, 1.0×10^{15} cm^{-2}) into crystalline silicon. All the wafers were implanted on an Applied Materials Quantum batch implanter at tilt and twist angles of 0° with electrostatic deceleration in front of the target. A key issue for these ultra-shallow implants of 10–20 nm depth is the variability of the native oxide. To avoid any influence on the implantation process, prior to implantation, the wafers were cleaned with Piranha, HF, SC1, and SC2 solutions, and an additional HF dip was done less than 15 min before the implantation step.

In Fig. 1, the various measured temperature time profiles of the processes used in this investigation are shown. The technique used to measure the temperature profiles during MSA is described in detail by Stuart et al. [15]. The thermal budgets, defined as time integrals of the temperatures, are considerably different and provide a preliminary estimation of the impact on diffusion. As an example, the integral of the temperature time profiles above 700 °C are: spike RTA (950 °C): 6310 °Cs, flash anneal (750 °C/1300 °C): 1809 °Cs, and flash anneal (950 °C/1250 °C): 4332 °Cs. For a more precise determination of the impact on diffusion, the temperature-time profiles would have to be weighted with the temperature-dependent dopant diffusivity. Locally, dopant diffusion may be reduced by clustering and precipitation while implant damage and ambient effects may lead to increased diffusivities.

The spike RTA processes were performed in a Mattson 3000 Plus RTP system equipped with Mattson's absolute temperature measurement, a temperature controller optimized for spike RTA, and wafer rotation [16]. The ramp-up rate to the pre-stabilization step was 50 K/s. The recipe included pre-stabilization at 650 °C for 10 s followed by a spike with a ramp-up rate set to 250 K/s. The peak temperature was varied between 900 and 1000 °C. The anneals were either performed in 100 ppm oxygen in nitrogen (standard for boron) [17] or in 10% oxygen in nitrogen (standard for arsenic) [18], but also higher and lower partial pressures of oxygen were tested to evaluate the influence of the gaseous ambient on dopant diffusion, retained dose, and surface protection [19,20]. The flash annealing was carried out in a Mattson Millios™ fRTP™ in nitrogen (around 100 ppm oxygen as a standard) and in an oxygen-containing ambient (10% O_2) to evaluate the influence of the gaseous ambient on dopant diffusion and activation. Depending on the particular recipe, after reaching an intermediate temperature of 750 °C (medium) or 950 °C (high intermediate temperature) with a ramp-up rate of 150 K/s, the wafers were flash annealed to peak temperatures of 1250 or 1300 °C [21–23]. The flash peak width is approximately 1.6 ms half maximum full width. Such high peak temperatures for MSA can be safely chosen as they have no impact on long-term reliability in patterned wafers especially considering the gate contact [4].

Sheet resistances were measured by a KLA-Tencor RS-100 four-point-probe (4PP) using probe type D. A measurement grid with a circular 121 site pattern with 3-mm edge exclusion was used. In Table 1, a summary of the sheet resistance values after flash anneals (T_i = 750 °C) with peak temperatures of 1250 and 1300 °C is given for the substrate conditions and dopants investigated.

The chemical profiles were analysed by high resolution secondary ion mass spectrometry (SIMS) using a CAMECA 4600 quadrupole depth profiler with a 500 eV primary O_2 beam under normal incidence for boron, and a 500 eV Cs beam under 45° for

Table 1
Sheet resistance and standard deviation summary for the various combinations of substrate condition, doping element, and flash anneal peak temperature investigated

	α-Si		c-Si			
	Boron[a]		Boron[a]		Arsenic[a]	
	1300 °C[b]	1250 °C[b]	1300 °C[b]	1250 °C[b]	1300 °C[b]	1250 °C[b]
Mean R_s (Ω/sq)	377	508	427	610	640	772
S.D. (%)	2.65	2.93	3.18	3.38	1.12	1.48

The gaseous ambient was 100 ppm oxygen in nitrogen.
[a] Dopant.
[b] Peak temperature of single flash anneal, the intermediate temperature is T_i = 750 °C.