Co-Implantation and Other Spike Anneal Solutions

Mater. Res. Soc. Symp. Proc. Vol. 912 © 2006 Materials Research Society 0912-C01-01

Millisecond Annealing: Past, Present and Future

Paul Timans[1], Jeff Gelpey[2], Steve McCoy[2], Wilfried Lerch[3], and Silke Paul[3]
[1]Mattson Technology, Inc., 47131 Bayside Parkway, Fremont, California, 94538
[2]Mattson Technology Canada, Inc., 605 West Kent Avenue, Vancouver, V6P 6T7, Canada
[3]Mattson Thermal Products GmbH, 10 Daimlerstrasse, Dornstadt, 89160, Germany

ABSTRACT

The challenge of achieving maximal dopant activation with minimal diffusion has re-awakened interest in millisecond-duration annealing processes, almost two decades after the initial research in this field. Millisecond annealing with pulsed flash-lamps or scanned energy beams can create very shallow and abrupt junctions with high concentrations of electrically active carriers, but solutions for volume manufacturing must also meet formidable process control requirements and economic metrics. The repeatability and uniformity of the temperature cycle is the key for viable manufacturing technology, and the lessons from the development of commercial rapid thermal processing (RTP) tools are especially relevant. Advances in the process capability require a fuller understanding of the trade-off between dopant activation, defect annealing. diffusion and deactivation phenomena. There is a strong need for a significant expansion of materials science research into the fundamental physical processes that occur at the short time scales and high temperatures provided by millisecond annealing.

INTRODUCTION

The continuing scaling of CMOS devices poses special challenges for the incorporation of electrically active dopants. In particular, the control of short channel effects demands ultra-shallow source and drain junctions with extremely shallow and abrupt doping profiles. These regions must contain very high concentrations of electrically active dopants to minimize the parasitic resistance of the transistor. The latter point is crucial because improvements in mobility of the channel combined with reductions in its length are making the on-state resistance increasingly limited by the parasitic resistance [1]. Polysilicon gate electrode doping must also be maximized, to limit depletion effects that increase the series capacitance and raise the effective dielectric thickness.

Ion implantation can provide very high concentrations of dopants in shallow layers but it has become increasingly difficult to make the dopants electrically active while restricting dopant diffusion during annealing. As a result, in recent years there has been a rise in interest in advanced methods for forming ultra-shallow junctions (USJ), with an especially strong focus on the challenge of boron doping, because of its fast diffusion characteristics [2]. Relatively conventional paths include RTP spike annealing, especially in combination with co-implantation of species that restrict the diffusion or improve the activation of the dopants. An alternative is solid-phase epitaxy (SPE), where the epitaxial recrystallization of amorphous layers created by the implant is accompanied by the incorporation of dopants on lattice sites. SPE occurs rapidly even at relatively low temperatures (<650°C), gives a very high degree of electrical activation and introduces little dopant diffusion [3]. However, SPE does not remove defects that form in the crystalline silicon just below the original position of the amorphous/crystalline interface.

These defects can greatly increase junction leakage current [4]. They can also release silicon interstitials during subsequent thermal processing, causing transient-enhanced diffusion of dopants and deactivating previously active dopants [3].

The limitations of conventional approaches and SPE have stimulated research into much shorter anneals performed at very high temperatures. Conventional RTP systems heat the wafer isothermally and the rate of heat loss from the wafer's surfaces limits spike anneals to > 0.5 s duration [5]. A shorter heating cycle can be achieved by delivering a pulse of heat to the wafer's surface, where the devices are formed, followed by very rapid conductive cooling to the substrate, which acts as a heat sink [6,7]. This method gives heating cycles with durations from tens of milliseconds all the way down to nanoseconds. Nanosecond-duration heating typically requires the use of pulsed lasers, which can deliver the extremely high power needed. However, the timescale is too short for many solid-state phenomena to progress significantly, and most nanosecond processing depends on forming a molten layer. When ion-implanted silicon melts and then regrows epitaxially on freezing, a very large concentration of dopants can be incorporated on lattice sites, forming extremely abrupt and highly activated junctions [2]. This has stimulated a tremendous amount of research and revealed many interesting phenomena, but turns out to be extraordinarily difficult to use in device manufacturing. The multitude of problems includes non-uniform heating of the device structures and the fundamental difficulty of working with a liquid phase that can often lead to surface deformation and damage [2].

Fortunately, it is possible to improve dopant activation by millisecond-duration heating at temperatures just below the melting point of silicon. This "millisecond annealing" approach was foreseen in the early 80s as being a natural extension of RTP for an era when diffusion should be constrained to ~10 nm [8]. Approaches explored included heating with flash lamps, scanned lasers and electron beams [7-15]. Much of the initial research was reported at early conferences of the Materials Research Society, but the device dimensions of the time did not demand such advanced capabilities and conventional RTP based on isothermal lamp heating emerged as the dominant manufacturing technology. For almost two decades millisecond annealing remained a scientific curiosity, but the approaching crisis in dopant activation has led to a renaissance in this technology [16-30]. However, this renaissance is accompanied by a much greater focus on the needs of volume manufacturing, bringing a host of new challenges, including requirements for process repeatability, uniformity, temperature measurement, defect control and cost-effectiveness.

KEY ISSUES IN THE MATERIALS SCIENCE OF MILLISECOND ANNEALING

How to maximize dopant activation?

Fig. 1 illustrates a fundamental concept underlying millisecond annealing [30]. The solid curves showing diffusion lengths of B atoms were calculated from $[4D(T)t]^{1/2}$, where D(T) is the intrinsic diffusion coefficient for B at the process temperature T and t is the process time. The activation energy for intrinsic B diffusion is 3.46 eV. The dashed curve represents the time taken to activate 50% of the carriers that can be introduced by an implant of 10^{15} B/cm^2 at 250 eV [31]. The latter process has an activation energy of ~ 4.7 eV, which is higher than that for diffusion, and hence it is kinetically favoured at higher temperatures. Hence, we can achieve better activation with less diffusion by annealing for a shorter time at a higher temperature, a trend that

has been confirmed in several studies [8,18,31,32]. Fig. 1 shows that restricting diffusion to < 2 nm requires annealing for < ~ 4 ms at > ~ 1250°C. This argument is based on kinetics, but higher temperature processing may also incorporate more dopant because of the greater solubility of dopants at higher temperatures [33]. However, the solid-solubility limit is essentially an equilibrium concept, making its significance less clear for ion-implanted layers and very short annealing times. Incorporation of dopants at moving crystal interfaces can provide concentrations of electrically active dopants that vastly exceed solubility limits, as seen in pulsed laser anneals that melt the surface, SPE of amorphous layers and in vapour deposition [2,3,34]. However, metastable doping may deactivate during subsequent thermal processing. Deactivation is especially significant for implanted layers because residual defects can release point defects that accelerate the process [3].

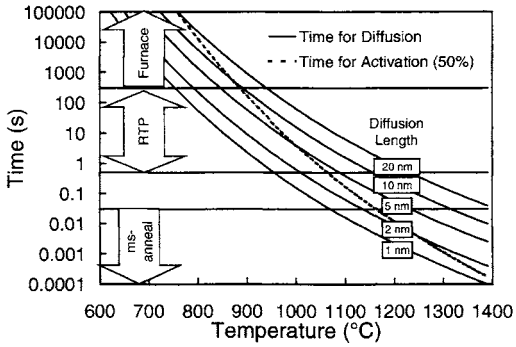

Fig. 1. Illustration of the thermal budget criteria for various degrees of B diffusion and for electrical activation of 50% of 10^{15} B/cm^2 implanted at 250 eV [30,31]. Millisecond duration anneals at temperatures just below the melting point of silicon allow electrical activation without significant diffusion. Reprinted with permission from Ref. 30 (©2004 IEEE).

Fig. 2 shows the heating cycle for a millisecond anneal performed using the flash-assisted RTP™ (fRTP™) method, where the wafer is rapidly heated to an intermediate temperature and its front surface is then heated by a pulse of energy from a bank of flash-lamps [16-23]. An ultra-fast pyrometer measured the temperature of the top surface of the wafer and a second pyrometer monitored the bottom surface. The temperatures of the two surfaces converge ~15 ms after the pulse, as the heat diffuses through the wafer. Fig. 3 compares conventional spike annealing and millisecond annealing through the trade-off between junction depth (X_J) and sheet resistance (R_S), which is a convenient metric for tracking how good an annealing technology is at activating dopants without causing excessive diffusion. Results for various B-doping implants consistently demonstrate the advantage of fRTP over conventional RTP [16,23,35]. Fig. 3 includes predictions for box-shaped doping profiles with concentrations of electrically active boron chosen to match the experimental results for the two annealing methods. There is ~ 100% improvement in electrical activation with the fRTP approach. Fig. 3 also shows the PMOS X_J-R_S requirement for 45 nm-node logic technology derived from the 2005 International Technology Roadmap for Semiconductors (ITRS) [36]. Improved ion-implantation and co-doping methods, combined with advances in millisecond annealing such as the flash-SPE approach, can be expected to meet these aggressive X_J / R_S targets [19].

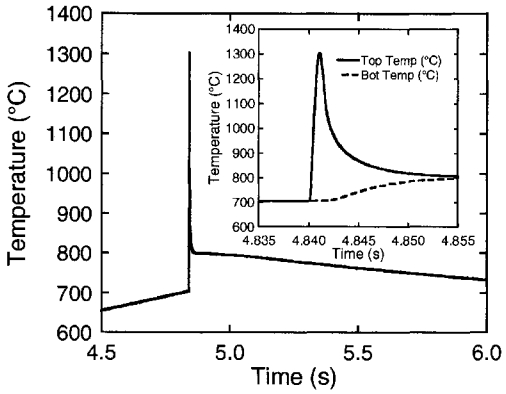

Fig. 2. Flash-Assisted RTP™ heating cycles. The temperatures of the two surfaces of the wafer are tracked by high speed pyrometers as the wafer is isothermally preheated to 700℃ and then a pulse of energy heats the front surface to 1300℃.

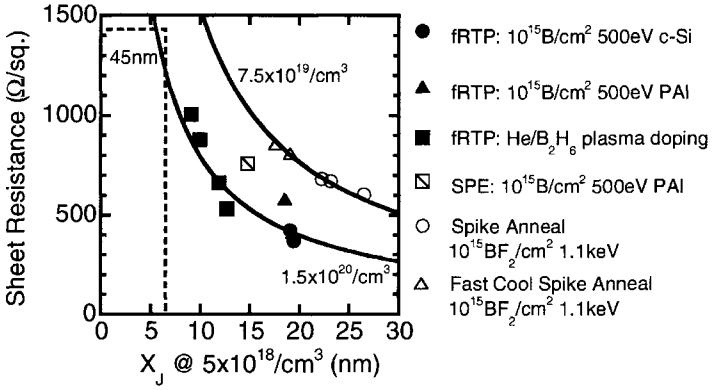

Fig. 3. A comparison of X_J/R_S results for flash-assisted RTP, conventional spike anneals and SPE, together with the ITRS target for 45 nm PMOS [16,23,35]. The solid curves are predictions for ideal, box-shaped boron doping profiles (B concentrations are marked on the curves).

How effectively does millisecond annealing remove defects?

Optimization of implantation and annealing schemes must also consider the acceptable limits for residual defects. The impact of residual defects on device performance strongly depends on their nature, concentration and location as well as on other issues such as impurity gettering phenomena. For example, residual defects in the depletion regions of p-n junctions can greatly increase junction leakage current [4]. It has been suggested that as devices scale down in size, junction leakage contributes less to the overall leakage power problem [37]. However, recent reports of low-power CMOS requirements have placed renewed emphasis on reducing junction leakage, and indeed specifically on reducing implantation damage [38]. Furthermore,

dopant deactivation concerns also drive the need to limit residual defects. A series of experiments were performed to probe the relative efficacy of spike annealing, millisecond annealing and SPE regrowth in removing damage. Measurements were performed using a novel non-contact metrology scheme, the RsL™ method from Frontier Semiconductor. The RsL approach uses dynamic measurements of photo-induced voltages in junctions to simultaneously measure sheet resistance and junction leakage [39]. Non-contact metrology is becoming essential because of the severe difficulties in using traditional contact-based probes to establish properties of USJ [40]. The study included comparisons of implants of 10^{15} B/cm^2 at 500 eV into crystalline and amorphous silicon, the effects of halo doping and the impact of annealing the halo implant before implanting the USJ. The preamorphization implants (PAI) were 10^{15} Ge/cm^2 at 30 keV, and the halo implants were $4x10^{13}$ As/cm^2 at 40 keV. In some samples the halo implants were annealed for 10 s at 1050°C before the PAI or B implants. The final anneals were either a spike anneal at 1050°C, an SPE anneal of 5 s at 650°C, fRTP with preheating to 700°C followed by a jump to 1250°C, or fRTP with preheating to 750°C followed by a jump to 1300°C.

Fig. 4 shows the trends in junction leakage for several of the conditions explored. The results show that SPE produces very leaky junctions, while spike anneals consistently give low leakage, regardless of the implant scheme adopted. Leakage for spike-annealed implants without halos was below the measurement limit of 0.1 μA/cm^2. B implants into crystalline silicon also resulted in low leakage, regardless of annealing approach, whereas PAI implants increased leakage. Halo doping consistently increased leakage, as may be expected, because it greatly reduces depletion region width. Annealing the halo implant before the B doping steps significantly improved leakage. PAI cases still showed significant leakage after millisecond anneals at 1250°C, but 1300°C anneals were more effective. The results show that optimized millisecond annealing can provide significantly improved damage annealing relative to low temperature anneals based SPE regrowth alone.

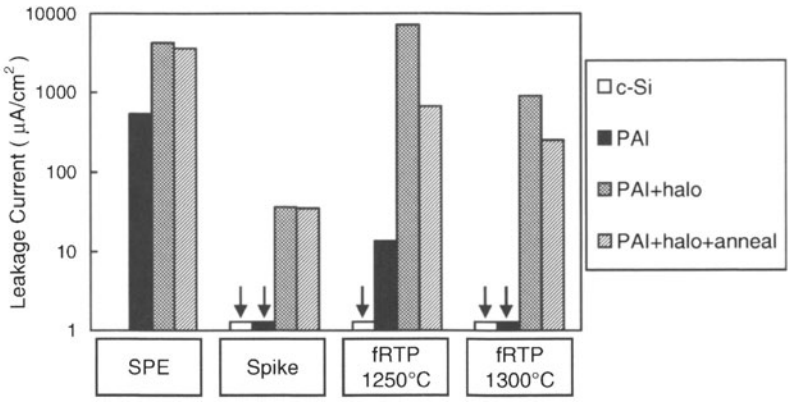

Fig. 4. Leakage current measurements from junctions formed by 500 eV B implants into various samples, annealed by various methods. Measurements were performed by the RsL™ method from Frontier Semiconductor. Arrows indicate cases where the leakage current was below the measurement limit (< ~ 0.1 μA/cm^2).

The impact of residual damage is also evident in the effects of thermal cycles on annealed junctions. Fig. 5 shows how the sheet resistances of implanted layers formed by three implant and annealing schemes evolve with 30 s anneals at various temperatures in a Mattson 2900 RTP tool [17]. The junctions were all formed by B implants into c-Si or amorphized silicon as described above, and the activating cycles were fRTP anneals at 1300℃ or SPE at 650℃. No deactivation is evident for implants into c-Si. Junctions formed by SPE of PAI layers show a strong deactivation above 750℃. Millisecond annealing of the PAI samples improved resistance to deactivation, as is consistent with more effective damage removal.

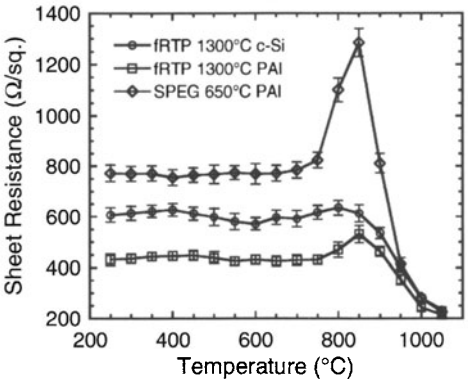

Fig. 5. Deactivation of B-implanted and annealed samples by subsequent thermal processing for 30 s at a range of temperatures [17].

MILLISECOND ANNEALING FOR VOLUME MANUFACTURING

The early literature on millisecond annealing is dominated by studies employing scanned CW laser beams because this approach was convenient for R&D, even though the laser beam powers available were too small for an economically viable wafer throughput. Indeed for the ~20 W ion laser that was often used, the throughput was expected to be four 3" wafers per hour [7]. As interest in millisecond annealing waned such problems ceased to be of interest, but manufacturing requirements are now very important, and questions of process control and economics must be resolved.

Uniformity, Repeatability and Process Control

Process uniformity and repeatability are key concerns in volume manufacturing. Here there are major differences between approaches that expose the entire wafer to broad-area pulses of radiation, such as flash-lamp annealing, and methods that sequentially process regions of the wafer, such as CW laser annealing. Flash-lamp annealing demands very uniform delivery of optical energy across the whole surface of the wafer. Advanced designs that employ optical integrators can provide exceeding uniform illumination, as illustrated by the results in Fig. 6(a), which is a histogram of 625 temperature values deduced from sheet resistance values measured on a 300 mm diameter wafer. The 1-sigma variation in temperature is only 3.3 K.

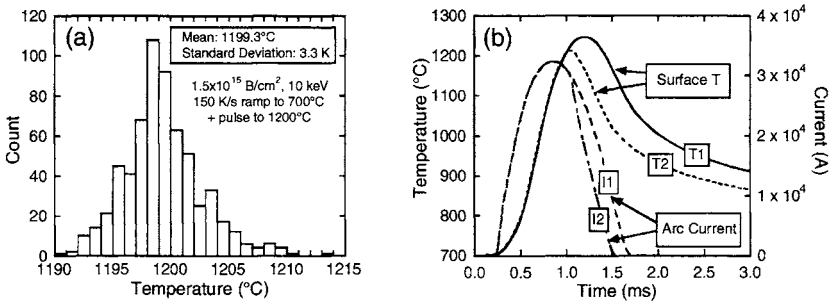

Fig. 6. Process control in fRTP millisecond annealing (a) Uniformity of large area pulsed annealing. The results are from a 625-point map of sheet resistance of a 300 mm implant monitor wafer annealed by the fRTP approach. (b) Temperature control in fRTP. Two temperature profiles (T1 & T2) are shown with corresponding lamp current profiles (I1 & I2).

The key to repeatability in thermal processing technologies lies in accurate temperature measurement and closed-loop control. The challenges are very similar to those for temperature control in conventional RTP, but the measurements and control actions must be delivered ~1000 times faster. Temperature monitoring in millisecond annealing is possible with advanced high-speed pyrometers, as shown in Fig. 2. Indeed, sufficiently fast temperature sensing enables closed-loop control and repeatability of the peak temperature. Fig. 6(b) illustrates how control of the current flowing in a bank of high power arc lamps can be used to set the maximum temperature reached at the wafer surface.

For scanned beam processing, wafer uniformity is coupled to repeatability, since the beam power or shape can vary both spatially, for example along the length of a line-shaped beam, as well as temporally. Typically scan fields must be "stitched" together to cover the whole wafer. Early studies of CW laser annealing showed that the amount of scan overlap must be optimized for reasonable uniformity [12-14]. Fig. 7 illustrates the challenge. Fig. 7(a) shows the predicted surface temperature for an elliptically-shaped Gaussian beam with a 1/e width of 100 μm and a 1/e length of 5 mm that is scanned in the direction of its width at a speed of 10 cm/s, giving a dwell time of 1 ms at the centre of the beam [41]. The beam power was set to give a peak temperature of 1300℃, for a preheat temperature of 500℃, and the thermal properties of silicon were fixed at those for 500℃. The integrated effect of a heating cycle on a thermally activated process can be treated as being equivalent to an effective time, t_{eff}, at an arbitrary reference temperature [31,32,42]. Here, t_{eff} was calculated for electrical activation of a B implant, using an activation energy is 4.7 eV and a reference temperature of 1050℃ [31]. Fig. 7(a) illustrates how t_{eff} evolves with time and shows that annealing occurs very near the maximum in temperature. Fig. 7(b) shows the distributions in temperature and t_{eff} along the long axis of the beam. The very strong temperature sensitivity makes the t_{eff} distribution much narrower than the temperature distribution. The practical consequence is a need for a large overlap of scans to assure reasonable uniformity. By summing t_{eff} contributions from successive scans we can predict the uniformity for any given strategy in beam shape and overlap. Fig. 7(c) shows the effect of overlapping multiple scans on uniformity. The peak-to-peak variation in t_{eff} is reduced to 5% of the average t_{eff} by using an 82% overlap. This amount of overlap is consistent with the

approaches adopted by the early laser annealers [7,12-14]. This analysis does not include other subtle effects, such as the effect of partial regrowth of amorphous layers during one scan on subsequent activation/deactivation by subsequent scans [7]. The area that can be processed per unit time is limited by overlap requirements and by the laser power available. Fig. 7(d) shows the wafer throughput predicted for a range of beam power, for a peak-to-peak uniformity criterion of 5%. Throughput issues will be discussed in more detail below.

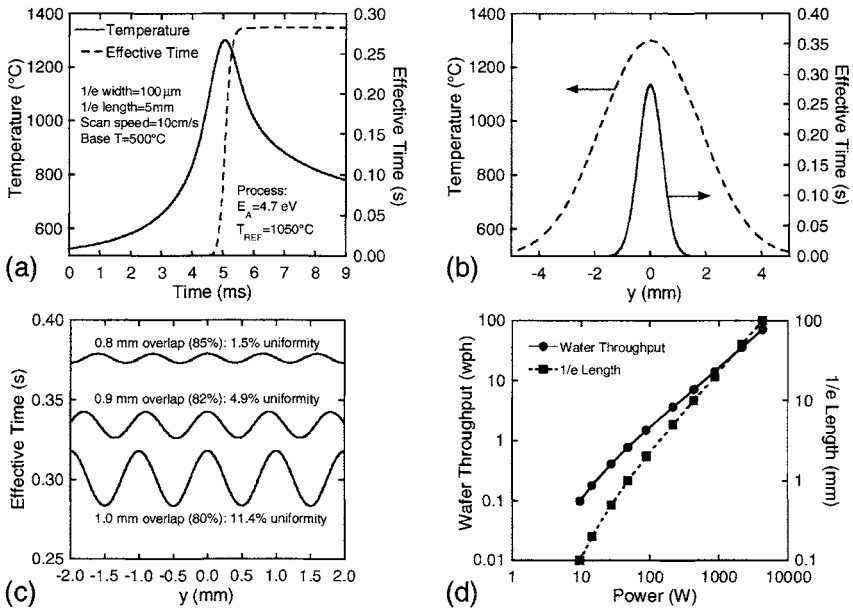

Fig. 7 (a) Temperature-time cycle for scanned laser heating, together with the evolution of the effective anneal time, t_{eff}. (b) Comparison of the temperature profile along the length of the beam with the t_{eff} distribution. (c) Residual non-uniformity in t_{eff} distributions when multiple scans are "stitched" together, for different overlap conditions. (d) Wafer throughput as a function of beam power, together with the corresponding beam lengths that can be employed. The throughput was calculated for overlap criteria that gave 5% range in t_{eff}.

Pattern effects in millisecond annealing

Whenever a wafer is not in thermal equilibrium with its surroundings, patterns on the wafer's surfaces can lead to temperature non-uniformity. This "pattern effect" arises because the pattern affects the absorption of electromagnetic energy and the emission of thermal radiation [43]. Thermal diffusion parallel to the wafer surface smoothes out temperature non-uniformity and reduces pattern effects. Hence, their magnitude depends strongly on the length scale of the pattern relative to that of thermal diffusion. In conventional RTP the length scale that is significant is typically > 5 mm, but in millisecond annealing it is much shorter, because heat has less time to diffuse. For example, the diffusivity of silicon, D_{Si}, is 1.05×10^{-5} $m^2 s^{-1}$ at 1100°C, and