ARTICLE IN PRESS

Materials Science in Semiconductor Processing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Materials Science in Semiconductor Processing

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

# Doping by flash lamp annealing

S. Prucnal*, L. Rebohle, W. Skorupa

*Helmholtz-Zentrum Dresden – Rossendorf, Institute of Ion Beam Physics and Materials Research, Bautzner Landstraße 400, 01328 Dresden, Germany*

ARTICLE INFO

*Keywords:*
Flash lamp annealing
Pulsed light sintering
Semiconductors

ABSTRACT

After a short introduction we will highlight processing issues (setup, comparison of annealing methods, relevant requirements for annealing due to doping, diffusion, activation, recrystallization, defect engineering), as well as doping issues for group IV-semiconductors (shallow junctions, hyperdoping, solar cells, superconductivity) and other semiconductors (manganese doping of GaAs for diluted magnetic semiconductors, doping for transparent conductive oxides). Mostly ion implantation serves as a source of dopants, but also diffusion from deposited layers is of growing importance.

## 1. Introduction

After 2000, flash lamp annealing (FLA) in the millisecond range using xenon-filled lamps has been becoming a key technology for thermal processing in advanced chip technology. Despite this, the starting point of research and development work devoted to this type of short time annealing was in the mid-seventies, at least for semiconductor materials processing. The birth of FLA happened with a first publication in 1977. Kachurin and Nidaev [1] from the Academy of Sciences at Novosibirsk reported results on electrical activation of 80 keV, $10^{16}$ $cm^{-2}$ phosphorus implants where they compared annealing with pulses of 8 ms of a ruby laser to annealing with pulse-discharge pump lamps. The results regarding carrier concentration and mobility were well comparable. Nevertheless, the modification of solid material by pulsed light treatments had already started earlier in the fifties, e.g. by the use of pump lamps later used for igniting lasers. More information about the early days of processing of materials with light sources can be found in Ref. [2].

The industrial need for millisecond thermal processing was not urgent at all during that early time. Halogen lamps and graphite heaters came on stream in the eighties for recrystallizing amorphous silicon, also in the framework of silicon-on-insulator (SOI) approaches, as well as for radiation damage removal and electrical activation purposes after ion implantation, see the related chapters of this issue. This type of annealing, working in the range of several tens of seconds, called Rapid Thermal Processing (RTP) technology, was finally reduced to about one second due to the imperative of following Moore's law in chip technology. With this so-called spike annealing RTP reached its limits at about 2000, at least for the shallow boron implants. Halogen lamps whose function is concerned with a glowing wire inside a glass-tube cannot deliver shorter pulse times. On the other hand, the advanced chip technology called for shorter annealing times below one second. The most important problem to be solved was the suppression of the transient enhanced diffusion of boron mostly related to the source-drain electrodes of the p-type transistors. An extensive review on transient enhanced diffusion of boron was presented by Jain et al. [3].

In the following we will present a short review on FLA for doping applications. Processing and equipment issues will be considered at first, followed by doping aspects in silicon as shallow junction formation, hyperdoping for band gap engineering and doping aspects in solar cells. For other group IV semiconductor materials topics as doping and superconductivity of germanium, and doping of silicon carbide will be reported. Finally the focus is turned to other semiconductors: doping of thin GaAs films, formation of diluted magnetic materials and doping aspects in transparent conductive oxides (TCO).

## 2. Processing issues

According to [4] short-time annealing encompasses annealing times over more than 10 orders of magnitude, namely between 10 ns and 100 s. Although there are plenty of potential annealing methods, there are three main techniques covering this range today, namely RTP in the range of 1–100 s, FLA in the range of 100 µs to 100 ms, and laser annealing (LA) in the ns range and below. The specific requirements of the material and the intended application (e.g. ultra-shallow junctions or doping above the solubility limit) determine the choice of the annealing method. A useful quantity for this determination is the thermal response time $\tau_{th}$, defined as [4]:

* Corresponding author.
*E-mail address:* s.prucnal@hzdr.de (S. Prucnal).

http://dx.doi.org/10.1016/j.mssp.2016.10.040
Received 23 June 2016; Received in revised form 7 October 2016; Accepted 21 October 2016
Available online xxxx
1369-8001/ 

Please cite this article as: Prucnal, S., Materials Science in Semiconductor Processing (2016), http://dx.doi.org/10.1016/j.mssp.2016.10.040

$$\tau_{th} = \frac{s^2}{D_{th}} \quad D_{th} = \frac{\lambda}{\rho C_p} \tag{1}$$

where $s$ is the thermal diffusion length, $D_{th}$ is the thermal diffusivity, $\lambda$ is the thermal conductivity, $\rho$ is the mass density, and $C_P$ is the specific heat capacity. Practically, $s$ is the characteristic length of interest, e.g. the width of a doping profile or the wafer thickness. For silicon, the thermal diffusivity ranges between 0.9 $cm^2 s^{-1}$ at RT and 0.09 $cm^2 s^{-1}$ at 1600 K which equals a thermal response time $\tau_{th}$ of 3 and 30 ms, respectively, for a wafer thickness of 525 µm.

RTA is an isothermal process which means that $\tau_{th}$ is much shorter than the annealing time and that both the front and backside of the processed wafer are at the same temperature. The energy is typically supplied by halogen lamps which emit a broad spectrum in the visible and near infrared (NIR) spectral range with a maximum in the NIR. Interference and pattern effects can be neglected as the annealing time is long enough to level out temperature differences which may arise due to inhomogeneous absorption.

In case of FLA $\tau_{th}$ is of the same order of magnitude as the annealing time. The backside of the wafer is still heated, but its temperature is significantly lower than that of the front side. The energy is supplied by Xe lamps which also emit a broad spectrum, but now with a maximum in the blue-violet spectral region. Interference and pattern effects matter, but they are reduced by averaging over the broad flash lamp spectrum. The resulting temperature profiles depend on the thermal and optical properties of the material system. Due to the optical interaction with the substrate, surface pattern effects can occur. Both RTA and FLA are "one-shot-one-wafer" processes and thus well suited for industrial applications.

In contrast to this, LA is an adiabatic process where $\tau_{th}$ is now much longer than the annealing time. In this case only a thin layer at the surface is heated, whereas the backside and the major part of the bulk remain at room temperature. Exceptions are LA with a continuous beam laser causing dwell times in the ms range and the use of a $CO_2$ laser with penetration depths comparable to the wafer thickness. As a laser is monochromatic, interference and pattern effects are serious and depend even more on the thermal and optical properties of the material system. In addition, the annealing of a whole wafer requires the scanning of the laser beam over the wafer area. The main characteristics of these three annealing techniques are summarized in Fig. 1.

Basically, a FLA tool consists of a process chamber with a substrate holder most probably made of quartz, a bank of Xe flash lamps, a reflector above which directs the light toward the wafer, and an optional preheating system (Fig. 2). The flash lamps are usually separated from the actual process chamber to protect the lamps against evaporated or sublimated sample material and to protect the sample against the lamps in case they explode or fracture. A good homogeneity of the flash light is usually achieved by using properly shaped reflectors, a suitable lamp arrangement (e.g. two overlapping banks of lamps) and the right measure of oversizing with respect to the arc length compared to sample dimensions. The preheating system, typically halogen lamps or a hot plate, enables higher peak temperatures while the temperature gradient between front and backside is reduced. This reduction can prevent the build-up of harmful stress levels inside the substrate caused by the different thermal expansion of the front and backside. Further literature on stress evolution and pattern effects during FLA can be found in [5–7] and [8–10], respectively.

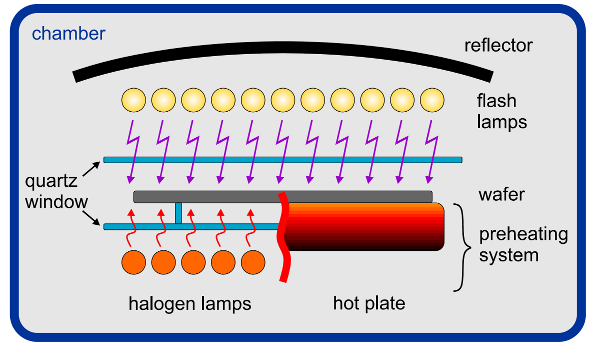


**Fig. 2.** Basic scheme of a FLA tool as used for semiconductor wafer processing.

The flash lamps are usually powered by a pulse forming network made of capacitors and coils which determine shape and duration of the flash pulse. Pulse duration and pulse intensity are the two most important FLA parameters and are approximated by the full profile width at half maximum and the energy density of the flash light which hits the sample surface, respectively. Typical values for semiconductor doping are in the range of 0.1–20 ms and 10–100 J $cm^{-2}$. Another parameter of interest is the emitted spectrum of the flash lamps which is composed of the emission lines of Xe (or another noble gas) and a broad spectral band which ranges from the UV to the IR. Whereas the broad band is caused by collision and recombination events within the plasma, the line features are due to inner transitions within the corresponding noble gas ions. Fig. 3 displays the time-integrated

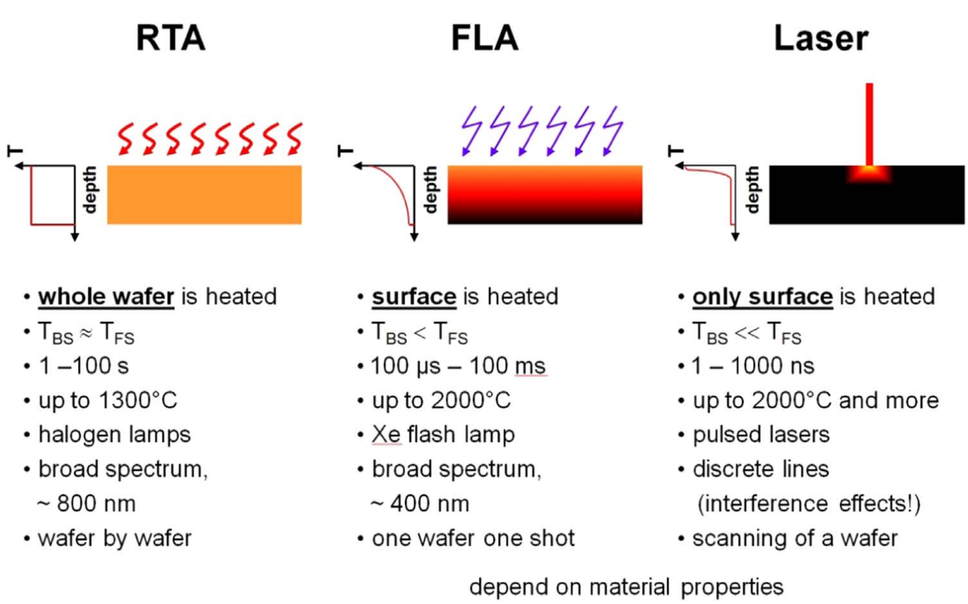


**Fig. 1.** Comparison of different short time annealing techniques.