Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Process Control 14 (2004) 423–430
JOURNAL OF PROCESS CONTROL
www.elsevier.com/locate/jprocont

Optimal control of rapid thermal annealing in a semiconductor process ☆

R. Gunawan, M.Y.L. Jung, E.G. Seebauer, R.D. Braatz *

Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, 93 Roger Adams Laboratory, Box C-3, Urbana, IL 61801-3792, USA

Received 9 May 2003; received in revised form 21 July 2003; accepted 21 July 2003

Abstract

This study focuses on the optimal control of rapid thermal annealing (RTA) used in the formation of ultrashallow junctions needed in next-generation microelectronic devices. Comparison of different parameterizations of the optimal trajectories shows that linear profiles give the best combination of minimizing junction depth and sheet resistance. Worst-case robustness analysis of the optimal control trajectory motivates improvements in feedback control implementations for these processes. This is the first time that the effects of model uncertainties and control implementation inaccuracies are rigorously quantified for RTA.

Keywords: Optimal control; Batch control; Uncertainty analysis; Robustness analysis; Microelectronics processes; Semiconductor processing; Rapid thermal annealing

1. Introduction

Moore's law requires a continued shrinkage of feature sizes in microelectronic devices. For example, advanced CMOS devices will require junction depths between 13 and 22 nm in the source and drain extension region by the year 2005 according to the 2001 International Technology Roadmap for Semiconductors. The current technology for formation of such ultrashallow junctions depends on ion implantation of dopant, such as boron, into silicon. Although the junction depth can be made shallower by reducing the implant energy, the effectiveness of this approach is limited by the need to anneal out the point and/or extended defects generated by ion implantation. Silicon self-interstitial defects can mediate the diffusion of dopants during the annealing process, which leads to a significant increase of the junction depth. This phenomenon is known as "transient enhanced diffusion" (TED).

The state-of-the-art in post-implant annealing employs a lamp-based rapid thermal annealing (RTA). Fig. 1 shows a typical RTA "spike" anneal program, which consists of a stabilization step at constant temperature (~650 °C), followed by a spike anneal comprising a linear heating step at a constant rate (~100 °C/s) reaching a maximum temperature (~1000 °C), and finally a radiative cooling step at a initial rate of several tens of degrees per second. In the literature, there exists conflicting experimental evidence on the efficacy of using high heating rates (up to 400 °C/s) in the spike anneal profile to reduce TED [1,2]. However, recent results [3,4] tend to confirm the benefit of using high heating rates. The results also suggest that the reduction in the junction depth comes at the expense of an undesired increase in the sheet resistance. The trade-off in reducing the junction depth without sacrificing the sheet resistance presents a major roadblock in transistor miniaturization and motivates a careful optimization of the post-implant annealing temperature program. For this reason, considerable efforts have been put forth in the design of the optimal post-implant annealing programs to produce the desired junction depth while achieving low sheet resistance (see [5] and references therein).

In the literature, there have been extensive developments in the control of rapid thermal processing (RTP) to provide temperature tracking and wafer thermal

☆ A preliminary abbreviated version of this manuscript was accepted for presentation at the IFAC Symposium on Advanced Control of Chemical Processes Hong Kong, 2003.

* Corresponding author. Tel.: +1-217-333-5073; fax: +1-217-333-5052.
E-mail address: braatz@uiuc.edu (R.D. Braatz).

0959-1524/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2003.07.005

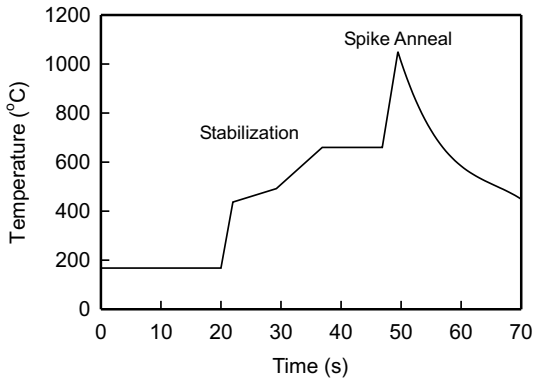

Fig. 1. Typical rapid thermal anneal temperature program, which consists of a stabilization step and a spike anneal (i.e., a fast linear heating step starting ~650 °C, followed by a natural cool down step).

uniformity (see [6] and references therein). However, the design of the optimal RTA program that minimizes TED has adopted an ad hoc approach through trial and error and heuristics [5], due to incomplete understanding of TED mechanisms and correspondingly undependable models for describing dopant diffusion and activation. A recent comprehensive TED modeling in conjunction with the maximum a posteriori estimation provided not only a correlative model but also a model for a priori predictive purposes [7,8]. This paper focuses on the design of the spike anneal program that optimizes the junction depth subject to a constraint on the sheet resistance. A model-based optimal control using the aforementioned TED model was formulated utilizing different parameterizations of the optimal trajectory to elucidate the true optimal annealing program. The optimal annealing program provides the temperature setpoints for lower level RTP/RTA controllers.

In addition, finite time control policies such as the spike annealing programs can result in control performance that is highly sensitive to model uncertainties and control implementation inaccuracies [9]. Worst-case robustness analysis of the resulting optimal trajectory quantifies the performance degradation with respect to model parameter uncertainties and control implementation inaccuracies [10]. Such analysis can provide guiding directions in improving the TED model (through additional experimentations) and in the selection of the RTA control instrumentation (sensors and actuators) and strategies. This is the first time that the effects of model uncertainties and control implementation inaccuracies are rigorously quantified for RTA.

2. Transient enhanced diffusion model

2.1. Model formulation

TED arises from reaction–diffusion processes consisting of Fickian diffusion, electrical drift motion, and reaction networks including boron activation and interstitial clustering. The model comprises coupled continuity equations (i.e., mass balances) for each species and Poisson's equation to incorporate the electrical field effect on the charged species. The general continuity equation is

$$\frac{\partial C_i}{\partial t} = -\frac{\partial J_i}{\partial x} + G_i \tag{1}$$

where C_i denote the concentration, J_i is the flux, and G_i is the net generation rate of species i. The flux J_i includes terms describing the Fickian diffusion and the electric field drift motion. The electric field arises from the spatial charge imbalance according to Poisson's equation:

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{Q(x)}{\varepsilon} \tag{2}$$

where ε denote the dielectric constant and the charge density $Q(x)$ is given by

$$Q(x) = p - n + \sum_i \gamma_i C_i \tag{3}$$

with p and n denoting the hole and electron concentrations, respectively, and γ_i denoting the average charge of species i (e.g., a species A which possesses a charge distribution of 30% A^{+2} and 70% A^{-} has an average charge of $30\% \times 2 + 70\% \times -1 = -0.1$). The charge distributions adhere to Fermi–Dirac statistics. The concentrations p and n are assumed to be in thermal equilibrium at all times.

The generation term G_i includes the formation and annihilation rates due to the boron activation and clusters formation and dissociation reactions. The boron activation reactions provide a pathway between mobile interstitial boron B_i to and from immobile activated boron (i.e., substitutional boron, B_s):

$$\begin{array}{c} \xrightarrow{\text{Kick-out}} \\ B_s + Si_i \underset{dis}{\overset{assoc}{\rightleftharpoons}} (B_s - Si_i) \underset{ki}{\overset{ko}{\rightleftharpoons}} B_i + Si_s \\ \xleftarrow{\text{Kick-in}} \end{array}$$

In addition, the intermediate ($B_s - Si_i$) also acts as a nucleation center for mixed boron–silicon clusters (see below). The rates of reactions follow reactant-limited rate expressions, where the reaction rate constants adhere to the Arrhenius law.

The above view of free B_i as the main pathway of boron diffusion agrees with computational density functional theorem (DFT) calculations [11–13] and experimental results [14,15]. In 1999, two simultaneous DFT-based studies appeared [16,17] providing evidence for the intermediate ($B_s - Si_i$) as the main pathway of boron diffusion, which was supported by subsequent publications [18–20]. This led to an explicit debate regarding the dominant mechanism of boron diffusion in

silicon. However, our Monte Carlo simulations using maximum likelihood parameters [21,22] showed that the diffusion of boron observed in TED is most likely mediated by motion of free B_i.

Clusters of interstitial atoms have been shown to form during TED [23,24]. There exists evidence supporting the formation of clusters consisting of pure boron [24], pure silicon [23], and mixed boron–silicon [25]. During thermal annealing, the clusters can act as reservoir (during stabilization step) and source (during heating step) for mobile interstitial boron and silicon. The formation and dissociation of pure interstitial clusters follow the reactions

$$I + I \underset{d}{\overset{f}{\rightleftharpoons}} I_2$$

$$I_{m-1} + I \underset{d}{\overset{g}{\rightleftharpoons}} I_m$$

where I denotes the interstitial (boron and silicon) and the indices m denote the sizes of the clusters. The cluster formation and growth rates assume diffusion-limited kinetics in agreement with much of the literature (see for example [26]). On the other hand, the dissociation rate follows a first-order kinetic expression with a rate constant according to the Arrhenius law.

The following reaction network describes the formation and dissociation of mixed boron–silicon clusters:

$$B_{p-1}Si_q + B_i \underset{\hat{d}}{\overset{\hat{g}}{\rightleftharpoons}} B_pSi_q$$

$$B_pSi_{q-1} + Si_i \underset{\hat{d}}{\overset{\hat{g}}{\rightleftharpoons}} B_pSi_q$$

where p, q are integers larger than or equal to 1. The formation/growth and dissociation rates of mixed clusters again follow diffusion-limited and first-order kinetics, respectively, as in the pure cluster dynamics. Detailed description of the model is available elsewhere [8].

2.2. Parameter estimates

The TED model requires a set of pre-exponential factors and activation energies associated with the diffusivities and kinetic rate constants for the boron activation reaction and cluster dissociation dynamics. Theoretical and experimental works of TED have paid little attention to the pre-exponential factors, focusing instead on the activation energies. In fact, pre-factors are notoriously difficult to estimate a priori. In simple kinetic models of elementary reactions, pre-factors represent modified attempt or collision frequencies. However, in solid there exists a distribution of vibrational frequencies, and thus defining a single attempt frequency is difficult. One commonly used estimate is the Debye frequency, which for Si is about 6×10^{12} s^{-1}. However, it is well known from other branches of kinetics that such a simple picture is often inadequate. For example, a survey [27] of pre-factors for gas desorption from semiconductor surfaces shows that, while the average value indeed lies near 10^{13} s^{-1}, only ten percent of individual cases fall within an order of magnitude of this range. The story is roughly similar for hopping diffusion on semiconductors [28]. Although the Si Debye frequency will be employed in the present work, clearly the use of a priori estimates of pre-factors requires caution.

Determinations of the activation energies have also proven difficult. Nevertheless, experimental and ab initio DFT estimates of the activation energies are scattered throughout the literature. For many of the activation energies, the published values show significant variation. For example, experimental results for the Si self-interstitial suggest diffusivities that vary by more than 10 orders of magnitude [29]. The reliability of the experimental results is questionable because the mobile species, including many point and extended defects, are difficult to observe because they exist at low concentration. Ab initio calculations based on density functional theory (DFT) have also proven problematic. Most of the DFT calculations are valid only at 0 K. However, the assumed mechanism may change at high temperatures. The expressions employed here tacitly assume a rate law based on a transition state formulation, in which an atom attempts to cross a well-defined potential barrier at a frequency comparable to a lattice vibrational frequency. Such formulations are usually applied under conditions at modest temperatures where kT is much less than the barrier height. At 1050 °C, however, $kT = 0.11$ eV, which approaches the same magnitude as several barriers in Table 1. Under such conditions, it becomes questionable whether the potential barrier remains well defined. Other possible high-temperature mechanisms have been discussed for other kinds of point defects by Van Vechten [30]. Another problem with DFT calculations is that they typically do not account for the effects of entropy, which can change the pre-exponential factor by orders of magnitude [31].

Table 1 lists the maximum a posteriori estimates for the most important TED energetics [8]. Maximum a posteriori estimation takes a Bayesian approach which combines experimental data with a priori information, in this case, from the maximum likelihood estimation applied to published experimental and/or DFT values. Fig. 2 presents the after-anneal experimental data from various RTA programs along with the corresponding simulation profiles using the TED model in conjunction with the MAP estimates. In addition, Fig. 3 shows the agreement of the TED simulations using the MAP estimates with the experimental observations compiled from the literature [32,33], and with the Sematech curve, which consolidates the best junction depth and sheet resistance pairs that can be achieved through conventional spike annealing.

Table 1
Maximum a posteriori estimates of TED energetics

TED parameters	MAP estimate	Std. dev.
E_{diff,B_i^+} (boron diffusivity)	0.336	0.013
$E_{\mathrm{diff},Si_i^{+2}}$ (Si diffusivity)	0.72	0.03
E_{ko} (kick-in/kick-out)	0.439	0.041
E_{ki} (kick-in/kick-out)	0.50	0.03
E_{dis} (kick-in/kick-out)	0.547	0.005
$E_{2,\mathrm{B}}$ (cluster dissociation—pure boron)	1.806	0.023
E_2 (cluster dissociation—pure Si)	1.4	0.03
E_3 (cluster dissociation—size 3)	2.184	0.018
E_4 (cluster dissociation—size 4)	3.055	0.002
E_{large} (cluster dissociation—large Si)	3.7	0.1
$E_{\mathrm{large,mix}}$ (cluster dissociation—large mixed)	3.5	n/a[a]

[a] Energy is from ML estimation of a single published value.

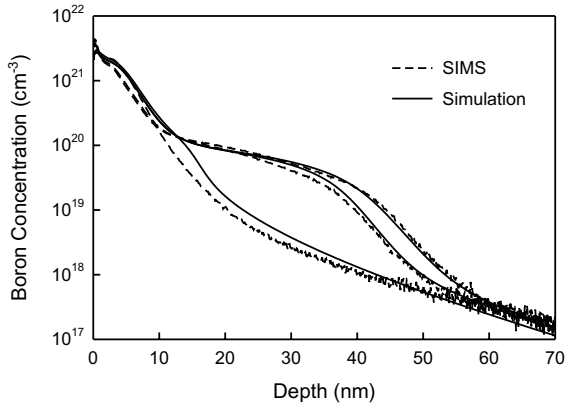

Fig. 2. Comparison of experimental and simulated boron profiles using the TED model in conjunction with the MAP parameters. The junction depth is defined as the depth from the surface at which the total boron concentration reaches 10^{18} atoms/cm^3: on the order of 50 nm in the present cases.

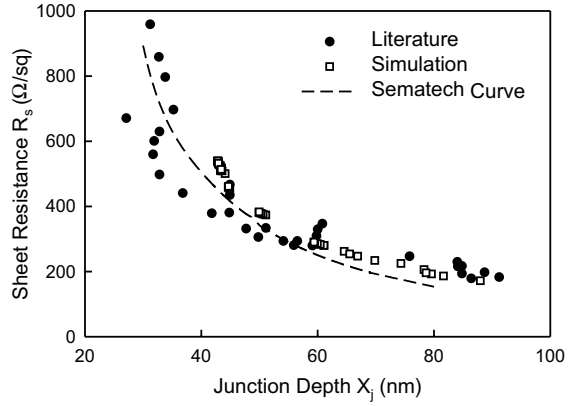

Fig. 3. Comparison of the junction depth and sheet resistance data from experimental studies employing various annealing conditions, the Sematech curve, and the TED simulations.

3. Optimal control formulation

The performance of a CMOS transistor depends particularly heavily upon the junction depth and sheet resistance. (Performance variables can also include junction abruptness, profile stability to subsequent process steps, and others.) Shallower junction depths give improved short channel performance to minimize leakage current, while low sheet resistance reduces power consumption. There are many avenues to reducing the junction depth, e.g., low implant energy [34] combined with RTA with high ramp-up rates [2,4] and alternative doping technology such as plasma inversion ion implantation or plasma gas-immersion laser doping [35]. Nevertheless, RTA currently provides the most dependable and economical path to satisfy the current needs of ultrashallow junctions.

This paper presents a model-based control formulation for designing the optimal temperature program that minimizes the junction depth while maintaining a suitable sheet resistance. The optimization variable is the RTA temperature profile, in particular, the heating and cooling profiles and the annealing (maximum) temperature. The optimal control objective chosen here is to minimize the junction depth while maintaining a satisfactory sheet resistance, which can be formulated into a constrained optimization problem:

$$\min_{\substack{T(t)\\ R_s \leqslant R_{s,\max}\\ \beta_{\min} \leqslant \frac{dT}{dt} \leqslant \beta_{\max}\\ T \leqslant T_{\max}}} \quad X_j \tag{4}$$

where X_j denotes the junction depth, R_s denotes the sheet resistance, and $T(t)$ is the RTA temperature trajectory. The sheet resistance is given by:

$$R_s = \frac{1}{q \int \mu(C) C(x)\, dx} \ [\Omega/\mathrm{sq}] \tag{5}$$

where q denotes the carrier charge, $\mu(C)$ denotes the mobility (concentration dependent), and $C(x)$ is the spatial concentration of charge carrier (i.e., activated dopant). The following formula gives the mobility $\mu(C)$ for boron [36]:

$$\mu(C) = 44.9 + \frac{425.6}{1 + \left(\frac{C}{2.23\times 10^{17}}\right)^{0.719}} \ [\mathrm{cm}^2/\mathrm{V\,s}] \tag{6}$$

In this work, it is desired to produce junctions with the sheet resistance below $R_{s,max}$ of 350 Ω/sq. The constraints on the temperature gradient, β_{min} and β_{max}, describe the limits for the cooling and heating rates, respectively. Lamp-based RTA can produce heating rates up to 400 °C/s [2], while recent advances in RTA technology using arc lamps [37] can achieve cooling rates up to 200 °C/s [38]. The maximum temperature of the thermal anneal T_{max} is set to the melting point temperature of silicon at 1410 °C.

4. Worst-case analysis

Worst-case analysis [9,10] provides tools for quantifying the robustness of the optimal control performance to uncertainties in model parameters and control implementation. The parametric and control uncertainties are described as norm bounded perturbations $\boldsymbol{\delta u}$ and $\boldsymbol{\delta\theta}$ such that

$$E_\theta = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \boldsymbol{\delta\theta}, \|\boldsymbol{W}_\theta \boldsymbol{\delta\theta}\| \leqslant 1 \right\} \tag{7}$$

$$E_u = \left\{ \boldsymbol{u} : \boldsymbol{u} = \hat{\boldsymbol{u}} + \boldsymbol{\delta u}, \|\boldsymbol{W}_u \boldsymbol{\delta u}\| \leqslant 1 \right\} \tag{8}$$

where $\boldsymbol{W}_\theta$ and $\boldsymbol{W}_u$ are positive-definite weighting matrices. This formulation includes uncertain parameters lying within a hyperellipsoid

$$E_{\theta,\mathrm{MAP}} = \left\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{V}_\theta^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leqslant r^2(\alpha) \right\} \tag{9}$$

where $r^2(\alpha)$ denotes chi squared distribution with α confidence level, which is equivalent to Eq. (7) with $\boldsymbol{W}_\theta = (1/r(\alpha))\boldsymbol{V}_\theta^{-1/2}$ for the Euclidean norm [10]. The uncertainty formulation also includes independent bounds on each element for the optimal control trajectory

$$\boldsymbol{u}_{min} \leqslant \boldsymbol{u} \leqslant \boldsymbol{u}_{max} \tag{10}$$

A first-order series expansion of the junction depth with respect to the MAP parameter estimates gives

$$\delta X_j = \boldsymbol{L}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \boldsymbol{L}\,\boldsymbol{\delta\theta} \tag{11}$$

where $\boldsymbol{\delta\theta}$ belongs to $E_{\theta,\mathrm{MAP}}$ and L denote the sensitivity coefficients given by

$$L_i = \left. \frac{\partial X_j}{\partial \theta_i} \right|_{\theta=\hat{\theta}} \tag{12}$$

Based on this expansion, the worst-case deviation of the junction depth is given by

$$\delta X_{j,wc} = \max_{\|\boldsymbol{W}_\theta \boldsymbol{\delta\theta}\| \leqslant 1} |\boldsymbol{L}\,\boldsymbol{\delta\theta}| \tag{13}$$

where the solution to the minimization problem gives the worst-case parameter uncertainty vector $\boldsymbol{\delta\theta}_{wc}$

$$\boldsymbol{\delta\theta}_{wc} = \pm \frac{r(\alpha)}{\|\boldsymbol{L}\boldsymbol{V}_\theta^{1/2}\|} \boldsymbol{V}_\theta \boldsymbol{L}^{\mathrm{T}} \tag{14}$$

and the maximum junction depth increase of

$$\delta X_{j,wc} = r(\alpha)\|\boldsymbol{L}\boldsymbol{V}_\theta^{1/2}\| \tag{15}$$

On the other hand, the worst-case analysis with respect to the control implementation inaccuracies requires a second-order series expansion:

$$\delta X_j = \boldsymbol{M}\,\boldsymbol{\delta u} + \boldsymbol{\delta u}^{\mathrm{T}} \boldsymbol{H}\,\boldsymbol{\delta u} \tag{16}$$

where

$$M_i = \left. \frac{\partial X_j}{\partial u_i} \right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} \tag{17}$$

$$H_{ik} = \left. \frac{\partial^2 X_j}{\partial u_i \partial u_k} \right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} \tag{18}$$

because the elements of M that are not on active constraints are equal to zero and thus the effect of such control uncertainties only shows up in the second-order terms. Here the worst-case junction depth increase is given by

$$\delta X_{j,wc} = \max_{\boldsymbol{\delta u}_{min} \leqslant \boldsymbol{\delta u} \leqslant \boldsymbol{\delta u}_{max}} |\boldsymbol{M}\,\boldsymbol{\delta u} + \boldsymbol{\delta u}^{\mathrm{T}} \boldsymbol{H}\,\boldsymbol{\delta u}| \tag{19}$$

This optimization problem is equivalent to [10]

$$\max_{\mu_\Delta(\boldsymbol{N}) \geqslant k} k \tag{20}$$

where k is any real number, μ is the structured singular value, the perturbation $\Delta = \mathrm{diag}\{\Delta_r, \Delta_r, \delta_c\}$ consists of independent real scalar blocks Δ_r and a complex scalar δ_c, and

$$\boldsymbol{N} = \begin{bmatrix} 0 & 0 & k\boldsymbol{w} \\ k\boldsymbol{H} & 0 & \boldsymbol{kHz} \\ \boldsymbol{z}^{\mathrm{T}}\boldsymbol{H} + \boldsymbol{M} & \boldsymbol{w}^{\mathrm{T}} & \boldsymbol{z}^{\mathrm{T}}\boldsymbol{Hz} + \boldsymbol{Mz} \end{bmatrix} \tag{21}$$

with

$$\boldsymbol{w} = \tfrac{1}{2}(\boldsymbol{\delta u}_{max} - \boldsymbol{\delta u}_{min}) \tag{22}$$

and

$$\boldsymbol{z} = \tfrac{1}{2}(\boldsymbol{\delta u}_{max} + \boldsymbol{\delta u}_{min}) \tag{23}$$

Upper and lower bounds for k can be efficiently computed using iterative μ computations or skewed-μ analysis (for details, see [10]).

5. Numerical implementation

The TED model was implemented using Alagator scripts running on a public domain FLOOPS 2000 [39]. FLOOPS spatially discretizes the continuity equations to give an implicit system of ordinary differential equations. To solve the resulting ODEs, the FLOOPS differential equation solver uses the TR/BDF2 [39] composite method for time integration, which combines the trapezoidal rule (TR) and the backward differentiation formula (BDF). The advantages of TR/BDF2 method are its second-order accuracy and ease of im-

plementation (one-step algorithm). In addition, it also has a desirable stability condition (L-stability) for solving stiff differential equations [40].

An experimental as-implanted SIMS profile provided by International Sematech was used as the initial boron profile. The ion implantation utilized 2×10^{15} ions/cm^2 of boron at 0.60 keV with 0° tilt into intrinsic silicon. The total boron was assumed to contain 20% substitutional boron and 80% interstitial boron as suggested by experimental observations [41,42]. The initial conditions for silicon interstitials were set to track the total boron concentration. The clusters and the B_s–Si_i complex were assumed not present initially. Boundary conditions at the surface for all species assumed no flux (i.e., $J_i|_{x=0} = 0$) with no surface Fermi level pinning [43].

The parameterizations of the optimal trajectory include linear and quadratic profiles. In the case of linear parameterization, the design variables consisted of the heating and cooling rates and the annealing temperature, which were solved by extending the golden search method [44] to multidimensional problems. The motivation of using this algorithm came from the small number of optimization variables and the ease of incorporating the output constraint arising from the sheet resistance.

The quadratic parameterization defines optimal heating and cooling profiles with minimum heating and cooling rates equal to the rates obtained in the linear parameterization. In this case, the maximum of the heating and cooling rates were relaxed to 1000 and −400 °C/s, respectively. Again, the optimization used the aforementioned extension of the golden search method. In addition, only concave quadratic profiles (i.e., positive quadratic coefficients) were considered in accord to the literature evidence showing that concave RTA program was more effective than convex profiles in reducing TED [45].

The worst-case analysis required computing sensitivity matrices with respect to the parameters and optimal control trajectories. The sensitivity coefficients were estimated using the finite difference method [46]. The structured singular value problem in Eq. (20) was computed using the μ-Analysis and Synthesis Matlab Toolbox [47].

6. Results and discussion

Fig. 4 presents the optimal RTA programs using linear and quadratic parameterizations of the temperature trajectory, which give junction depths of 51.3 and 48.2 nm, respectively (shown in Fig. 5), and the same sheet resistance of 350 Ω/sq. The optimal linear heating and cooling rates were 400 and 200 °C/s, respectively, indicating that the optimal RTA program was to heat and cool as quickly as possible to the annealing temperature of 1111 °C, in agreement with experimental studies [3,34]. The use of a high annealing temperature with fast heating and cooling is needed since the effective activation energy of TED is lower than that for boron activation. Although experimental studies [3,34] have suggested using fast heating and cooling rates, the determination of the maximum annealing temperature is currently done through trial-and-error experimentation and heuristics. In contrast, the optimal control formulation using the TED model provides an estimate of the annealing temperature, and therefore will reduce the number of costly experiments.

The quadratic parameterization led to the optimal quadratic trajectory with an annealing temperature of 1144 °C. The use of quadratic profiles only gave minimal improvement of the junction depth over the optimal linear program. In fact, if the optimization problem for linear parameterization was solved using relaxed constraints on the heating rate and cooling rates to equal those of the quadratic case, the optimal annealing

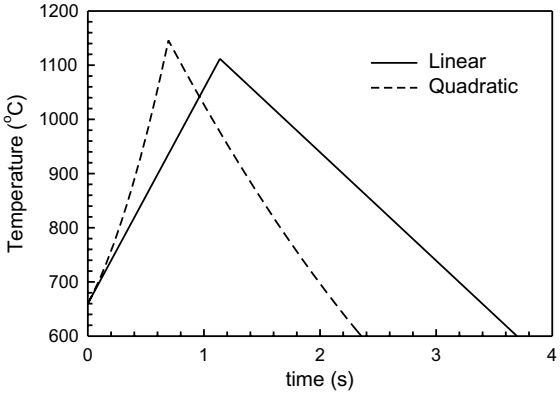

Fig. 4. The optimal spike anneal programs employing linear and quadratic parameterizations.

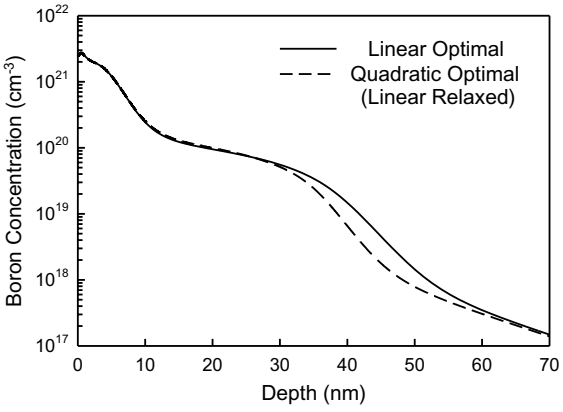

Fig. 5. Simulations of after-anneal boron profiles employing the optimal RTA programs.

temperature increased to 1146 °C to give a junction depth of 48.2 nm (see Fig. 5). That is, if the same maximum heating and cooling rates are used, the quadratic and linear parameterizations lead to the same minimum junction depth. Since heating and cooling rates are the true constraints in practice, there is no benefit of using quadratic over linear heating and cooling profiles. In addition, the results indicate that the use of even higher heating rates than the existing RTA technology (>400 °C) will not bring significant reduction in the TED.

Worst-case analysis was applied to the linear optimal trajectory. The model parameter uncertainties were quantified by the MAP covariance estimate [8]. The analysis of control implementation inaccuracies used control trajectory perturbations of 5, 10, and 15 °C at five temperatures along the heating and cooling ramps (660, 800, 950, 1050, 1100, 1112 °C). Table 2 presents the worst-case junction depth increases due to uncertainties in model parameters and control implementation.

The analysis indicates that the deviations from the optimal junction depth from model parameter uncertainties are minimal, whereas the junction depth increase due to control errors are moderate to significant. The result suggests that the MAP estimation provided parameter estimates with sufficient accuracy for use in the optimal control study. In practice, control errors can arise from many sources. Inaccuracies from temperature measurements using a pyrometer are in the order of 3–15 °C. The typical RTA controllers make approximately 100 temperature measurements every second [48], which translates to a corrective control action every 4 °C at a heating rate of 400 °C/s. Temperature non-uniformity across the wafer on average can cause control error of 10 °C. Another source of control implementation errors which is not easily quantified as a temperature deviation, is the rounding of the temperature trajectory at the top of spike annealing due to physical limitations. Accounting for all sources of control inaccuracies, the control implementation errors could easily exceed 15 °C at any given time. From the worst-case analysis, control errors of this magnitude can lead to junction depth increase as high as 10 nm, which is in similar magnitude to the junction depth increase from the TED itself. This finding indicates that existing metrology and controllers for implementing RTA programs need improvement as future junction depth requirements necessitate further reduction of junction depth, and thus smaller control errors.

Table 2
Worst-case analysis of junction depth deviations (in nm) with respect to uncertainties in model parameters and control implementation

$\delta\theta$	$\lvert\delta u\rvert \leqslant 5$ °C	$\lvert\delta u\rvert \leqslant 10$ °C	$\lvert\delta u\rvert \leqslant 15$ °C
0.13	1.89	5.15	9.78

7. Conclusions

This paper has shown that the optimal RTA program for minimizing TED while achieving the desired sheet resistance consists of fast linear heating and cooling profiles, as suggested by many experimental studies. The model based optimal control directly computes the maximum annealing temperature, which avoids the heuristics and trial-and-error approach commonly done in literature and in industrial practice, reducing the number of experimentations and the cost to identify the optimal annealing program. Worst-case analysis on the optimal junction depth suggests the need to improve existing metrology and RTA controllers to minimize control implementation inaccuracies.

Acknowledgements

Support from International Sematech and the National Center for Supercomputing Applications (NCSA) is gratefully acknowledged.

References

[1] D.F. Downey, S.W. Falk, A. Bertuch, F.S.D. Marcus, Effects of fast rapid thermal anneals on sub-keV boron and BF_2 ion implants, Journal of Electronic Materials 28 (1999) 1340–1344.
[2] S. Shishiguchi, A. Mineji, T. Hayashi, S. Saito, Boron implanted shallow junction formation by high-temperature/short-time/high-ramping-rate (400 C/sec) RTA, in: Symposium on VLSI Technology, Kyoto, Japan, 1997. Paper 7B-3.
[3] G. Mannino, P.A. Stolk, N.E.B. Cowern, W.B. de Boer, A.G. Dirks, F. Roozeboom, J.G.M. van Berkum, P.H. Woerlee, N.N. Toan, Effect of heating ramp rates on transient enhanced diffusion of ion-implanted silicon, Applied Physics Letters 78 (2001) 889–891.
[4] G. Gelpey, K. Elliot, D. Camm, S. McCoy, J. Ross, D.F. Downey, E.A. Arevalo, Advanced annealing for sub-130 nm junction formation, in: 201st Meeting of the Electrochemical Society, Philadelphia, PA, 2002. Abstract 735.
[5] S.C. Jain, W. Schoenmaker, R. Lindsay, P.A. Stolk, S. Decoutere, M. Willander, H.E. Maes, Transient enhanced diffusion of boron in Si, Journal of Applied Physics 91 (2002) 8919–8941.
[6] T.F. Edgar, S.W. Butler, W.J. Campbell, C. Pfeiffer, C. Bode, S.B. Hwang, K.S. Balakrishnan, J. Hahn, Automatic control in microelectronics manufacturing: practices, challenges, and possibilities, Automatica 36 (2000) 1567–1603.
[7] M.Y.L. Jung, E.G. Seebauer, New physics for modeling transient enhanced diffusion in RTP, in: Proceedings of the Rapid Thermal and other Short-Time Processing Technologies, vol. 2000-9, ECS, Toronto, Canada, 2000, pp. 15–20.
[8] R. Gunawan, M.Y.L. Jung, E.G. Seebauer, R.D. Braatz, Maximum a posteriori estimation of transient enhanced diffusion energetics, AIChE Journal 49 (2003) 2114–2123.
[9] D.L. Ma, S.H. Chung, R.D. Braatz, Worst-case performance analysis of optimal batch control trajectories, AIChE Journal 45 (1999) 1469–1476.
[10] D.L. Ma, R.D. Braatz, Worst-case analysis of finite-time control policies, IEEE Transactions on Control Systems Technology 9 (2001) 766–774.

[11] C.S. Nichols, G.F.A. van de Walle, S.T. Pantelides, Mechanisms of dopant impurity diffusion in silicon, Physical Review B 40 (1989) 5484–5496.
[12] J. Zhu, T.D. de la Rubia, L.H. Yang, C. Mailhiot, G.H. Gilmer, Ab initio pseudopotential calculations of B diffusion and pairing in Si, Physical Review B 54 (1996) 4741–4747.
[13] J. Zhu, Ab initio pseudopotential calculations of dopant diffusion in Si, Computational Materials Science 12 (1998) 309–318.
[14] N.E.B. Cowern, G.F.A. van de Walle, D.J. Gravesteijn, C.J. Vriezema, Experiments on atomic-scale mechanisms of diffusion, Physical Review Letters 67 (1991) 212–215.
[15] N.E.B. Cowern, G.F.A. van de Walle, P.C. Zalm, D.J. Oostra, Reactions of point defects and dopant atoms in silicon, Physical Review Letters 69 (1992) 116–119.
[16] W. Windl, M.M. Bunea, R. Stumpf, S.T. Dunham, M.P. Masquelier, First-principles study of boron diffusion in silicon, Physical Review Letters 83 (1999) 4345–4348.
[17] B. Sadigh, T.J. Lenosky, S.K. Theiss, M.-J. Caturla, T.D. de la Rubia, M.A. Foad, Mechanism of boron diffusion in silicon: an ab initio and kinetic monte carlo study, Physical Review Letters 83 (1999) 4341–4344.
[18] P. Alippi, L. Colombo, P. Ruggerone, Energetics and diffusivity of atomic boron in silicon by density-functional-based tight-binding simulations, Computational Materials Science 22 (2001) 44–48.
[19] P. Alippi, L. Colombo, P. Ruggerone, A. Sieck, G. Seifert, T. Frauenheim, Atomic-scale characterization of boron diffusion in silicon, Physical Review B 64 (2001) 075207.
[20] J.-W. Jeong, A. Oshiyama, Atomic and electronic structures of a boron impurity and its diffusion pathways in crystalline Si, Physical Review B 64 (2001) 235204.
[21] M.Y.L. Jung, R. Gunawan, R.D. Braatz, E.G. Seebauer, Pair diffusion and kick-out: relative contributions to diffusion of boron in silicon, Journal of Electrochemical Society, submitted for publication.
[22] R. Gunawan, M.Y.L. Jung, R.D. Braatz, E.G. Seebauer, Parameter sensitivity analysis applied to modeling transient enhanced diffusion and activation of boron in silicon, Journal of Electrochemical Society, in press.
[23] P.A. Stolk, H.J. Gossmann, D.J. Eaglesham, D.C. Jacobson, C.S. Rafferty, G.H. Gilmer, M. Jaraiz, J.M. Poate, H.S. Luftman, T.E. Haynes, Physical mechanisms of transient-enhanced dopant diffusion in ion-implanted silicon, Journal of Applied Physics 81 (1997) 6031–6050.
[24] E.J.H. Collart, A.J. Murrell, M.A. Foad, J.A. van den Berg, S. Zhang, D. Armour, R.D. Goldberg, T.S. Wang, A.G. Cullis, T. Clarysse, W. Vandervorst, Cluster formation during annealing of ultra-low-energy boron-implanted silicon, Journal of Vacuum Science and Technology B 18 (2000) 435–439.
[25] T.E. Haynes, D.J. Eaglesham, P.A. Stolk, H.J. Gossmann, D.C. Jacobson, J.M. Poate, Interactions of ion-implantation-induced interstitials with boron at high concentrations in silicon, Applied Physics Letters 69 (1996) 1376–1378.
[26] K.J. Laidler, Chemical Kinetics, Harper & Row, New York, NY, 1987.
[27] Z. Wang, E.G. Seebauer, Estimating pre-exponential factors for desorption from semiconductors: consequences for a priori process modeling, Applied Surface Science 181 (2001) 111–120.
[28] E.G. Seebauer, M.Y.L. Jung, Surface diffusion on metals, semiconductors and insulators, Chapter II, in: H.P. Benzel (Ed.), Landolt-Börnstein Numerical Data and Functional Relationships, Vol. III/42A, Springer Verlag, New York, 2001.
[29] D. Eaglesham, Dopants defects and diffusion, Physics World 8 (1995) 41–45.
[30] J.A. Van Vechten, Activation enthalpy of recombination-enhanced vacancy migration in Si, Physical Review B—Condensed Matter 38 (1988) 9913–9919.
[31] J.A. Van Vechten, C.D. Thurmond, Entropy of ionisation and temperature variation of ionisation levels of defects in semiconductors, Physical Review B (Solid State) 14 (1976) 3539–3550.
[32] A. Agarwal, H.-J. Gossmann, A.T. Fiory, Effect of ramp rates during rapid thermal annealing of ion implanted boron for formation of ultra-shallow junctions, Journal of Electronic Materials 28 (1999) 1333–1339.
[33] A. Agarwal, H.-J. Gossmann, A.T. Fiory, Ultra-shallow junctions by ion implantation and rapid thermal annealing: spike-anneals, ramp rate effects, Material Research Society Symposium Proceedings 568 (1999) 19–30.
[34] A. Agarwal, Ultra-shallow junction formation using conventional ion implantation and rapid thermal annealing, in: International Conference on Ion Implantation Technology, Tyrol, Austria, 2000, pp. 293–299.
[35] E.C. Jones, E. Ishida, Shallow junction doping technologies for ULSI, Materials Science and Engineering R 24 (1998) 1–80.
[36] B.V. Zeghbroeck, Principles of Semiconductor Devices, 2002. Available from <http://ece-www.colorado.edu/~bart/book/>.
[37] B. Peuse, A. Rosekrans, J. Gelpey, P. Dick, Dynamic temperature uniformity of RTP using single zone illumination, in: 1st International Rapid Thermal Processing Conference, 1993, p. 443.
[38] Vortek Industries Ltd., private communications, 2002.
[39] M.E. Law, S.M. Cea, Continuum based modeling of silicon integrated circuit processing: An object oriented approach, Computational Materials Science 12 (1998) 289–308.
[40] R.E. Bank, W.M. Coughran Jr., W. Fichtner, E.H. Grosse, D.J. Rose, R.K. Smith, Transient simulation of silicon devices and circuits, IEEE Transactions on Electron Devices ED-32 (1985) 1992–2007.
[41] M.J. Caturla, M.D. Johnson, T.D. de la Rubia, The fraction of substitutional boron in silicon during ion implantation and thermal annealing, Applied Physics Letters 72 (1998) 2736–2738.
[42] H. Kobayashi, I. Nomachi, S. Kusanagi, F. Nishiyama, Lattice site location of ultra-shallow implanted B in Si using ion beam analysis, in: K.S. Jones, M.D. Giles, P. Stolk, J. Matsuo, E.C. Jones (Eds.), Si Front-End Processing—Physics and Technology of Dopant-Defect Interactions III, Materials Research Society, Inc., Warrendale, PA, 2001, p. J5.3.
[43] M.Y.L. Jung, R. Gunawan, R.D. Braatz, E.G. Seebauer, Surface fermi level pinning: An electrical "valve" in transient enhanced diffusion, in: Materials Research Society Spring Meeting, San Francisco, CA, 2001. Presentation J4.21.
[44] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, New York, NY, 1992.
[45] Z. Nényei, C. Grunwald, W. Lerch, J. Niess, D.F. Downey, R. Ostermeir, Competitive process analysis and some new optimization methods in RTP, in: Proceedings of the 6th International Conference on Advanced Thermal Processing of Semiconductors, Kyoto, Japan, 1998.
[46] A. Varma, M. Morbidelli, H. Wu, Parametric Sensitivity in Chemical Systems, Cambridge University Press, Melbourne, Australia, 1999.
[47] G.J. Balas, J.C. Doyle, K. Glover, A. Packard, R. Smith, μ-Analysis and Synthesis Toolbox, The MathWorks, Inc., Natick, MA, 2001.
[48] P.J. Timans, Mattson Technology, private communications, 2002.