SIMULATION OF SEMICONDUCTOR DEVICES AND PROCESSES Vol. 4
Edited by W. Fichtner, D. Aemmer - Zurich (Switzerland) September 12-14, 1991 - Hartung-Gorre

Challenges for Process Modeling and Simulation in the 90's - an Industrial Perspective

Marius Orlowski
MOTOROLA INC.,Advanced Products Research and Development Laboratory
3501 Ed Bluestein Blvd, Austin, Texas 78721, USA

1 Global Perspective

1.1 Introduction

This paper looks as much into the past as it tries to peer into the next few years, in order to identify the needs and requirements for process simulation in an industrial environment. The look into the immediate and not too distant future aims at the assessment of rapidly changing and emerging technologies to gauge the challenges they pose for process modeling and simulation. The look into the past taken here is necessarily a selective one since it will be dictated by the previously identified challenges that face us today and which we will encounter in the future. Thus, the accomplishments of the past are not enumerated but screened with respect to their utility in addressing the challenges lying ahead. Besides, there are excellent reviews in this field to which the reader is referred, and thus it seems quite pointless to compile the same published material in a slightly different manner, offering a slightly different emphasis. After all, one has to bear in mind that this field is not an academic discipline in its own right but is mainly driven and justified by the needs of the semiconductor industry. The difficult issues and problems, exciting as they may be in terms of scientific research, often become obsolete and irrelevant from the industrial point of view before they are fully understood, let alone solved, because of the transitory nature of many of the pertinent technologies.

Another aspect, along the same theme, is the often encountered circumstance that a theoretical answer (solution) to the whys and to the control requirements of a specific technology is not achieved before the technology itself is already mature in engineering terms and well under control. It is not disputed that even then simulation has its rights and place in the industrial environment. It has been proven in practice that even if the basic mechanisms and phenomena are satisfactorily understood, the sheer complexity and intricacy of modern semiconductor processing makes the use of simulation tools not only beneficial but - given the fierce global competition - imperative. Rather what is emphasized here is that process simulation - or, in this context, more precisely modeling - can be more efficiently applied than presently done. What is meant by this is something one could call in military language the need for a *rapid deployment modeling force* providing an engineer with general concepts, guidance, and insights at a **very early stage** of process or technology development. First order approaches coupled with computer modeling capability in the critical stage between invention and application are crucial to the industry.

1.2 Are Modeling and Simulation becoming separate disciplines?

At this point, the distinction between modeling and simulation which till now appeared to be more or less a linguistic one, should be discussed. This distinction will very likely become a differentiation between separate disciplines. Let me first define what process modeling and process simulation are. Process modeling can be viewed as an activity consisting of two stages. First, there must be a concept of mechanisms and relations which capture the essence of the actual phenomenon. Second, the concept of mechanisms and relations has to be translated into a set of equations or computational operations. Simulation is not simply computation, i.e. it is not only evaluation of these equations and operations. Today, simulation is a complex discipline involving computer science, numerical mathematics, informatics, computer graphics, and presumably in the not too distant future, artificial intelligence. We might add here also statistics concepts integrated into the overall simulation fabric. Statistical methodology can provide sets of recipes to be evaluated by traditional simulation tools, allowing one to estimate model and/or process parameters and to make error statements at specified levels of confidence. Thus simulation today and in the future sets out to be a sophisticated, multidisciplinary, and powerful engineering tool. Two examples may serve as an illustration: i) visualization, and ii) software interfaces.. As to the first example, obviously, a tremendous advantage of simulation as opposed to the experiment is that simulation allows one to *see* the inner-workings of a device or of a reaction chamber, for example. It allows to *see* the internal electric field or temperature distribution at any point in space and time. But how does one *see* this?

1.3 Visualization

The net result of a simulation is numbers, column after column and page after page of numbers. In a case of a three-dimensional calculation it is not unusual to deal with 100K data points for one quantity. This, together with 300K spatial coordinates multiplied by the number of possible time steps exceeds the grasp of any human's imaginative faculty. To *see* these data one has to bring them into context. Visualization does precisely this; it creates context, and generates the relations necessary for understanding. So, graphics as condensed and structured information is nowadays a necessity rather than a convenience or luxury. In the case of transient simulation a video capability will be increasingly important. This stems from the fact that we are not only interested in the final outcome; we are equally interested in the intermediate stages through which the system evolves, because this knowledge can be crucial for the assessment of the process feasibility when process conditions are subject to significant modifications. A good practical example of this is the coupled dynamics during the rapid thermal anneal of dopants and crystal defects. How seriously the visualization is already taken in some places is exemplified by the establishment of an entire visualization group and visualization laboratory at the University of Illinois and at the National Center for Supercomputing Applications (NCSA) in Urbana-Champaign.

Graphical representation is also required in a further respect as a structuring and administrative element to indicate in an easy manner the current state of simulation within a given process setup. A related issue is the educated, intelligent retrieval of relevant or consolidated numbers out of the plethora of simulation results. This capability might be realized within an expert system integrated with existing tools taking advantage of the techniques provided by artificial intelligence.

1.4 Tool Integration and Virtual Fab

The second example mentioned in section 1.2 is the problem of communication between various simulators. This includes the issue of representation of data, geometries and attributes. In the US, industrial, university, and national laboratories launched a task force, the *TCAD Initiative*, to coordinate and to streamline the particular activities of the participating labs in a joint effort to establish standards for data representation and software interfaces and to avoid duplication.

Another pressing subject, particularly for three-dimensional applications, is the development of improved, faster algorithms (grid generators and solvers) taking advantage of specific computer architectures, the integration of different time and spatial scales, the treatment of moving nonplanar boundaries in three-dimensions - just to mention some of the important issues. In logical consequence, the overall objective of this integration effort is to develop tools and technologies to enhance reproducible, high yield, low cost, and flexible chip manufacturing by constructing something daring now being called a *virtual factory*, a factory which can be run in simulation. This might become reality sooner than one is presently inclined to think. The progress already achieved in this ambitious project, particularly at Stanford University, is noteworthy but cannot be discussed here. However, there are two utilities of practical importance which must be mentioned in this context. The first is intelligent software which helps to analyze the simulation output under various aspects. Second, it can be expected that progress in the computer sciences will provide designers and engineers with the inverse modeling capability subject to defineable constraints, at least in selected areas where the technological requirements can be translated "backwards" directly into device design, for example.

1.5 Environment Required for Modeling

As to the conditions for successful process modeling, clearly, process modeling must not only have **direct** access to experimental data but also requires, very often, an experimental facility with fast turnaround to understand or to verify conceived mechanisms or to simplify the experimental setup in such a manner as to be able to extract relevant model parameters. This suggests that the modeling effort should be closely tied to experimental material research and process development. Equally important for process modeling is the availibility of appropriate software to accommodate the input of a new model for its evaluation. Here, the software will be used for model development or generally for insight, whereas within the simulation framework it serves as a tool for analysis and/or for optimization. The speedy evaluation of a new model requires simulation tools with flexible and modular architecture.

This statement is also valid to a lesser degree for the simulation itself. The need for swift adaptation of simulation capabilities to changing technological requirements can be addressed by software packages which allow an exchange of modules and easy implementation of new or updated models and which can be handed out again to a device or process engineer. There exists already a first generation of prototypes, with such general-solver software as ZOMBIE and PROMIS from the University of Vienna and PEPPER from MCC in Austin or on a more descriptive level the tree architecture of PREDICT 2 from MCNC. The next generation of process simulators must have capabilities reaching far beyond those of the traditional workhorses in process simulation as SUPREM III and SUPREM IV from Stanford University, combining at least the model flexibility of ZOMBIE, PROMIS and PEPPER, with the architectural features of PREDICT 2. First, they have to have a utility shell which allows an easy input of process flow and flexible choice of modules and models. Second, they should possess a flexible interface to other process simulators, to topography simulators and possibly to equipment simulation tools as well. Third, their flexibility and modularity should include a library of primitives for easy definition of new model equations, geometries, and attributes. Stanford University has recently embarked on such a project in two-dimensional process simulation.

1.6 Why is There a Need for a Hierarchy of Models?

Another issue in the context of process modeling is the need for a hierarchy of process models in order to bridge the gap between microscopic, macroscopic, and eventually, technological parameters to provide guidance to an engineer. This is of practical importance in the industrial environment. Consider an elaborate Monte Carlo code supplied with extensive knowledge of atomic and molecular forces, of crystal structure and electron shell properties of the relevant impurities, a code which can handle implantation and diffusion from first principles. For the sake of argument one can even be bolder and assume that the code's results luckily match the experimental profiles, i.e. we are considering a software tool with truely predictive capability and not merely a fitting machine. What is the use of such software to an engineer designing a device? A very small one, I presume, if any. For, in the extreme case, he might possibly be in utter confusion as to how his final simulation data (for example, final impurity model after an anneal) are to be **understood**. True, he can look at the damage creation during the implantation, he might analyse the initial dynamics of excess point defects, he might study the proximity effects of the interaction of end-of-range dislocations with the Si surface, and so on. But is he sufficiently equipped to make such inquiries let alone to perform such kind of analysis? Probably not. Probably, he is not even supposed to be and to do so, given his distraction with the chores of supervising his lots in the process line. Or, how can he possibly assess the significance of some esoteric model parameters whose meaning and range of applicability is known only to a handful of experts? In brief, all this suggests that such a superb simulation tool in academic terms does not generate any **usable, i.e operational**, knowledge or **understanding** for the application engineer. Understanding is generated by relations to items of which we have a sufficient grasp.

These remarks should help to justify the need for *short cut* models, which might be either descriptive, phenomenlogical, or physical in nature [1]. At any rate, process simulation, apart from the attempt to understand fundamental mechanisms, must offer a hierarchy of models relating microscopic quantities gradually to the more macroscopic ones. This school of thought is particularly advocated at MCNC in North Carolina and is an extremely valuable approach for characterizing the phenomenology of novel process conditions or of emerging technologies.

However, a new element will be required in the near future. Apart from the need of translating direct experimental observations into descriptive, phenomenological models, there is also an urgent need to devise similarly simple models coming, however, from a different side, namely by **condensing** sophisticated microscopic concepts into models which are easy to grasp while retaining the essential physical relations for specific technology applications. Right now there is hardly anyone doing this. The researchers involved in modeling efforts probing into deeper levels of fundamental understanding of the mechanisms are often too consumed with their work and too busy to embark on another level of **modeling, viz. of the relations between the data obtained from microscopic simulation**. This postulate is not as far fetched as it may appear at first. It is known in device simulation that simulation tools are extensively used to help generate simplified design models, or models for circuit designers. Device simulation adresses a **particular** device, it does not give a **perspective** on a **class** of devices.

[1] "Physical model" by no means implies the employment of elementary mechanisms. A physical system can be characterized by macroscopic observables, they themselves being an outcome of some microscopic variables, and described by suitable relations between these observables that we call *physical laws*.

This effort, strongly advocated here, almost automatically entails an often forgotten responsibility of the modeling to assess the validity range of a given model with respect to a specific application and to reexamine the basic assumptions under which the model equation has been derived. In doing so, despite - or as it should be clear by now just because of - the hierarchy of models of various degree of complexity one could arrive at a cohesive body of physical description resulting in clear dispositions for an engineer seeking support.

Finally, I wish to resume the point mentioned last in the introduction of this section. A wealth of concepts and techniques have been developed in other engineering sciences like engineering mechanics, material sciences, metallurgy or even geology, as well in physics, chemistry and mathematics. Many of these techniques are virtually unknown to the semiconductor community but they lend themselves to be used or slightly adapted to semiconductor applications. Some of these examples will be mentioned in section 3 of this paper. They can be an invaluable help, particularly at the initial stage of technology development at which materials and processes have to be understood and characterized.

1.7 Merging of Process and Equipment Modeling/Simulation

Traditionally process modeling was concerned with processes within the silicon crystal such as diffusion or with reactions in intimate contact with the crystal surface, like oxidation. All processes of deposition and etching of materials not reacting with the crystal silicon surface have been termed topography simulation and more recently equipment simulation - in recognition of the impact of the physics and chemistry present in a reaction chamber, i.e. of the entire equipment setup. However, it becomes more and more evident that the "below" and "above" silicon division is quite arbitrary and artificial. For example, the temperature distribution in a furnace used for oxidation determines the oxide uniformity not only from wafer to wafer but also within the wafer. The control of the oxide thickness is as important from an engineering point of view as the thickness variation due to variations in dopant concentrations in the silicon at the crystal surface. The thickness variation due to temperature distribution effects has hitherto been considered to be the subject of equipment simulation whereas the thickness dependence on dopant levels is the traditional domain of process simulation. An even more drastic case is rapid thermal oxidation. Here, the oxide growth depends on the heat source spectrum, on gas dynamics, convective cooling, surface chemical cleaning conditions, in addition to the more "conventional" effects like doping concentration levels and thermally induced stress. In both cases the line which divides process and equipment simulation is the issue of boundary conditions and how they are brought about. Until today, the boundary conditions for process simulation were estimated based on the good judgment of the engineer.

With the inclusion of equipment simulation it will be possible to account for variations in processing conditions that are intrinsic to equipment geometry. The two disciplines are even more entangled if we consider the effect of the heating up and cooling down phases of the furnace on the oxidation or on the dopant activation process, because of the growing importance of transient phenomena. Another example is the deposition of metals and subsequent silicidation of the silicon crystal with concomittant diffusion of impurities. The development of advanced or new materials will not only require the full thrust of the joint capabilities of process and equipment simulation but, beyond that, also new techniques and new concepts. The scope of process and equipment modeling on the one hand and process and equipment simulation on the other hand will keep expanding. So we are facing a division between modeling and simulation and at the same the time integration of process with equipment modeling and of process with equipment simulation.

1.8 Elements of Interdisciplinary and Cross-institutional Collaboration

Historically simulation evolved side by side with modeling, but today and in the near future as indicated above, a restructuring and integration of these and other disciplines on a large scale will become necessary. The demands facing process modeling and simulation, the enormous complexity of processes, the vast variety of materials, the multitude of techniques and concepts - all this makes an inter-disciplinary and inter-institutional collaboration mandatory. This is a fact which everybody is aware of. What is less clear is that the existing network of collaboration and interdependence can no longer be left to itself, but must be backed by some concept and administered and monitored accordingly - not only on a national level, but in view of the scarce resources and immense costs, on a multinational and perhaps global level. In the US, organizations in existence such as Sematech in Austin and the Semiconductor Research Corporation (SRC) exemplify such coordinating bodies with and without core research facilities. Major US semiconductor companies are trying to overcome technological challenges by harnessing the combined research strengths of universities, both in the US and in Europe, national labs, and industrial research partnerships. For example, Motorola is an active member of SRC, and of Sematech; participates actively in joint projects with several universities such as Stanford University, University of California at Berkeley, MIT, University of Texas at Austin, University of Illinois, MCNC, and the University of Florida for the US, and in Europe has close collaborations with ETH Zurich, with Fraunhofer Gesellschaft in Erlangen, with the Technical University of Vienna, with Technical University of Aachen, and also maintains contacts to university researchers in Japan.

This shift in paradigms in the conduct of research will generate a new vision and a need for new organizational arrangements. One possibility is to concentrate all the disciplines involved in fewer super research centers. This might not always be a desirable or feasible solution for political, economic, educational and regional reasons, as well for reasons of wearout and of complacency within one super-institution. The alternative for successful research in this field is a closer, structured, coordinated, and monitored collaboration between expert groups, centers of excellence, and industrial laboratories. This approach is not attractive, especially to those who like to have one dominant factor or explanation of any goal or effect (or rather often of an absence of such factor). There is, however, some evidence that the success of such an undertaking relies on **many factors** carefully blended and balanced with one another and on the achievable degree of consensus between the parties involved in defining the overall sustainability of a system.

The key elements in the research planning and management are the identification of strategically important thrust areas **along with** the identification of orthogonal technological opportunities in conjunction with research capabilities, and pertinent resource allocation. After a thorough process of consultations the phase of coordination and of streamlining of the research efforts can take place. Then, in order to make this network of inter-institutional arrangements work, a continual exchange of experience on a personal basis between researchers through periodic workshops, an institutionalized easily accessible circulation of progress and status reports, active feedback from user's groups, and regular assessment of the status of a particular project are indispensible. This system of meetings, workshops and so on keeps the technology and expertise transfer, the blood of the system, alive.

Major universities and national labs in the US offer courses, training classes, and workshops on a regular schedule to those who intend to be introduced in an efficient manner to fields as well as to experts working directly in the related field, wishing to update and/or enhance their knowledge. In Europe, only CEI-Europe/Elsevier Company offers such courses, but they are often poorly attended. It goes without saying, that only engineers with appropriate background and training will be able to transfer expertise and technology from participating research centers to their labs and to tap resources outside of their companies.

Finally, returning to the specific topic of this talk of process modeling and simulation, it has to be mentioned that process modeling and simulation are particularly suited to assume a critical role as an agent or medium of continuous education and training and that they lend themselves to serve as an interface to research and technology transfer from other research centers.

2 Technological Challenges for Process Simulation

2.1 Overview

The continuous downscaling of feature sizes in integrated circuits brings about reduction of vertical and lateral dimensions and, due to the constraint of diffusion kinetics in silicon, a simultaneous reduction of heat cycles. Reduced dimensions require much better control over the evolution of dopant profiles and reaction kinetics leading to layer formation. As stated recently by Poncet [2] the challenge consists in much tighter control requirements than previously, because many physical and geometrical effects considered second or third order effects a few years ago, are nowadays becoming first order effects. Downscaling acts as magnifier for subtle physical phenomena, and consequently effects which could be neglected on a larger scale must now be controlled on a smaller one. A good illustration of this is the control of dopant diffusion kinetics during rapid thermal annealing which requires control of transient phenomena related to damage generation, excess nonuniform point defect concentrations and dopant and damage anneal under various circumstances, including implantation damage, preamorphization, oxidation modified diffusion behavior, damage annealing under point defect injection condition, self-diffusion, gettering, and oxygen precipitation.

Another effect of reduced dimensions is the decreasing thickness of deposited or grown layers. As the granularity of these thin films approaches the dimension of the film thickness new physical behaviors come into effect. Modeling of multilayer structures of thin films requires full analysis of interfacial phenomena like segregation, interface passivation properties, contamination, and of stress at the surface as well as in the bulk.

Low thermal budgets can be achieved either by processing at lower temperatures (that means for the most part temperatures of 850^oC or lower) or by decreasing the cycle time, i.e. by rapid thermal processing. Low temperature processing involves a new regime of crystal damage anneal, leading to more diffusional broadening of dopant profile at lower than at higher temperatures [3] due to retarded point defect recombination at low temperatures. Low thermal silicidation has produced the most startling effects, eg. the recent report [4] by IBM researchers of anomalous asymetrical dopant diffusion at temperatures as low as 200^oC, which presents a new challenge to present diffusion theories. In section 3.2 of this paper a novel explanation will be put forward explaining this phenomenon by unidirectional dopant migration in high transient stress gradients.

A related issue already mentioned in section 1.7 is the determination of temperature boundary conditions in an RTA furnace which can be only addressed by the joint effort of process

[2] A. Poncet, "Recent trends in multilayer process simulation for submicron technologies", *Proceedings of ESSDERC '90*, p. 277, 1990

[3] P. Packan, Ph.D. dissertation, Stanford University, 1991

[4] M. Wittmer et al, Phys. Rev. Lett. **66**, p. 632, Feb. 1991

and equipment modeling. Considerable effort has been made at Stanford and at the Fraunhofer Gesellschaft in Erlangen [5] in this area.

Looking further into the future, process modeling is faced with the advent of new materials related to existing technologies such as the potential use of copper for the interconnects or the use of new ferroelectric materials such as PLZT, for DRAM applications, or related to adjunct technologies such as optoelectronics and the recent renaissance of superconductivity as a substitute for existing semiconductor interconnect technology. It is also likely that the emphasis of materials research will shift to lower (liquid nitrogen) temperatures, because of the diminishing engineering returns of further downscaling with simultaneous increase of power consumption and circuit degradation. Furthermore, it can be observed that already device design has begun to challenge the statistical or continuum treatment of electronic behavior in solids, so process simulation presumably will have to part with simulation based on partial differential equations and rely increasingly on Monte Carlo [6] based algorithms. This is already a technique of choice in low pressure reactive ion etching and chemical vapor deposition. In process simulation an area likely to convert entirely to Monte Carlo based calculations is ion implantation into a sandwich system of thin layers, particularly in compound semiconductors. In the case of silicon, knowledge of the depth distribution of implant-induced damage and of point defects is required to understand activation and diffusion mechanisms occurring during the subsequent anneal, particularly for very short time scales. This complex information which has to be provided for a sound process modeling of advanced annealing techniques, in addition to the as-implanted profile of the impurity is not likely to be provided reliably by means other than Monte Carlo and Boltzmann transport based techniques.

Finally, it has to be mentioned that a major obstacle for process model development or for model calibration is the unsatisfactory extraction and poor characterization of relevant physical parameters. Particularly, the problem of measurement of two-dimensional implanted and diffusion profiles is a critical one. Unfortunately, the shrinkage of semiconductor devices to subquater micron levels has not been accompanied by equal progress in the measurement techniques used to characterize these devices. The discussion of this topic is beyond the scope of this paper and the reader is referred to an excellent review [7] by Subrahmanyan in which he also lists the relevant literature on this subject.

2.2 Transient Diffusion-Reaction Phenomena for Dopants and Point Defects

Four areas can be distinguished in diffusion modeling: 1) the mechanisms by which impurities and point defects diffuse, 2) the reactions of dopants and particularly of point defects with all sorts of lattice imperfections such as extended defects, oxygen precipitates, and presence of other species such as hydrogen, traps, dopant and point defect clusters, 3) boundary conditions for dopants, including segregation phenomena, and for point defects, including recombination velocities and conditions for point defect injection during interfacial reactions such as oxidation, nitridation, and silicidation, and 4) electrical activation and deactivation mechanisms of dopants at high concentration. Even though this classification may be considered quite arbitrary, since

[5] see for example B. Hu et al "Process Simulation for Laser Recrystallization", talk given at *1991 International Workshop on VLSI Process and Device Modeling* , Oiso, Japan

[6] A Monte Carlo program for the simulation of point defect and impurity diffusion has been developed by Van Vechten and collaborators: U. Schmid et al, Comp. Phys. Comm., **58**, 329, 1990

[7] R. Subrahmanyan, "Methods for the Measurement of Two-dimensional Doping Profiles", *Proceedings of the International Workshop on the Measurement and Characterization* , edited by C.M. Osburn and G.E. McGuire, Research Triangle Park, NC, 1991, MCNC, to appear in JVST B, 1991

all the four areas are so tighlty coupled to one another that almost any other classification scheme would be as good as this one, it has the advantage of reflecting the present interests of researchers in process modeling.

2.2.1 Diffusion Mechanisms

As to the first item of diffusion mechanisms, it is now known that dopants in silicon diffuse with both vacancies and interstitials and considerable amount of recent work has clarified the relative importance of each point defect species for boron, phosphorus, arsenic, and antimony **intrinsic** diffusion. Studies of the influence of point defects, i.e. of self-interstitials and vacancies, on the diffusion of substitutional dopants in silicon have led to the conclusion that substitutional dopants conform to the dual-diffusion model, diffusing with preference f_I via interstitialcy and with complementary preference $f_V = 1 - f_I$ via vacancy mechanism. Given a diffusion equation $\partial_t C = \nabla(D\nabla C)$ the diffusivity is given by $D = D_o(f_I C_I / C_I^* + f_V C_V / C_V^*)$ where C_I and C_V are self-interstitial and vacancy concentrations, respectively, and asterisks refer to quantities under equilibrium point-defect conditions. It has been established that phosphorus and boron diffuse predominantly via interstitials, whereas arsenic and antimony via vacancies. It has been also found [8] that certain diffusion phenomena can be only explained by postulating the existence of dopant fluxes driven by the gradients in point defect concentrations which, under some conditions, can lead to up-hill diffusion. A model which accounts for both observations is the so called dopant-point defect pair diffusion model [9] and related models which can be reduced to the concept of to the pair diffusion if evaluated in the limit of the characteristic time for pairing reactions or for the kick-out mechanism being much faster than the elementary diffusional jump. Whether the pair diffusion model is the unique descriptive model which accounts for the two basic observations ($D \sim C_I$ *or* $\sim C_V$ and $J \sim \nabla C_I$ *or* ∇C_V) mentioned above is still an open question. One of the unsettled issues is, for example, the relative concentration of point defects and of dopant-defect pairs in various charge states. In a recent paper Giles [10] has investigated the dependence of transient diffusion effect on background doping. It has been found that arsenic background greatly enhances P diffusion. A high concentration boron background completely suppresses the transient damage effect on P diffusion. This effect arises because negatively charged interstitials are more effective at promoting the diffusion of positively charged substitutional phosphorus than neutral or positively charged interstitials. The extracted energy level assignments of the respective charge states were succesfully used in explaining transient phosphorus diffusion due to phosphorus implantation damage. A similar analysis of the implantation damage on arsenic-phosphorus codiffusion has been performed by Law and Pfiester [11]. However, a large degree of ambiguity as to the precise value of energy levels, dopant-pair binding energies, reaction constants and diffusivities still persists [12]

Another kind of complication arises from some indications of dopant diffusion dependence on the presence of other dopants which cannot be explained by simple Fermi level effects. These

[8] M. Orlowski, Appl.Phys. Lett. **58**, 1991, p.1479

[9] Despite its recent popularity, there is a long history of pair diffusion first invoked by Yoshida et al., J. Appl. Phys., **45**, 1498, (1981), subsequently, based on the percolation assumption, modified by Mathiot and Pfister, J. Appl. Phys.,**55**, 3518, 1984, and most recently revived by Morehead and Lever, Appl. Phys. Lett.,**48**, 151,1986, and others.

[10] M. Giles, *"Transient Phosohorus Diffusion Below the Amorphization Threshold"*, preprint 1990, submitted to J. Appl. Phys., see also M. Giles, IEEE-CAD **8**, 5, 1989.

[11] M. Law and J. Pfiester, *"Low Temperature Annealing of Arsenic/Phosphorus Junctions"*, preprint 1990, submitted to IEEE-ED.

[12] M. Orlowski, review paper on *" Mechanisms of Dopant Diffusion in Si"*, in preparation

effects have been rationalized by invoking dopant-dopant pairing effects [13]. Also, in a series of papers and most recently [14] Aronowitz has shown that current diffusion theories that couple active dopant with point defects are inadequate to deal with dopant-dopant interactions. The theory he advances is based on quantum mechanical calculations of an extended Hückel theory of model silicon lattice and is capable of modeling the experimental profiles as well as successfully predicting diffusion patterns that are then observed.

2.2.2 Dopant - Crystal Defect Interactions

To complicate the picture even further and to shift attention to the second point of the classification given above, it has recently been found [15] that not only do the crystal damage and its dynamics greatly influence dopant diffusion and activation, but also that impurities at sufficient concentrations affect the evolution of extended defects in silicon. One is therefore dealing with an extremely sensitive system of interactions and dependencies. This extreme sensitivity is responsible for the agonizing discrepancies in extracted values for point defects equilibrium concentrations and diffusivities. Early work in this area by researchers such as Gösele, Tan, Taniguchi, Hu, Fahey and others yielded values which differed by several orders of magnitude [16]. Although models will presumably always remain incomplete to some extent, it has to be conceded that recently, as far the characterization of point defect dynamics is concerned, significant progress has been achieved, notobly due to the work by Griffin at Stanford University, Law at the University of Florida and by Giles at the Univiersity of Michigan. The results of this work are consistent values for variables such as bulk and interface recombination velocities, diffusivities and equilibrium concentrations covering a technologicall relevant range of experimental situations including oxidation and nitridation for the point defects [17]. The key to the reconciliation of the point defect parameters has been careful modeling of point defect interface and bulk dynamics. The bulk silicon even under intrinsic conditions is characterized by various levels of vacancy traps depending on the silicon material (CZ versus FZ versus epitaxial) and its processing history. Another factor influencing the bulk dynamics is the oxygen interaction with point defects. The majority of the oxygen($\sim$ 95%) is atomically dissolved and accupies interstitial sites. The high diffusivity and low solubility make oxygen the most important precipitate-forming element in CZ silicon. Current modeling work on oxygen precipitation is being done at the Technical University of Vienna [18] and at Stanford [19]. On the theoretical side of oxygen precipitation Schrems calculates the growth and decay of oxygen precipitates combining a staistical approach with chemical rate equations and with Fokker-Planck equation which describes larger precipitates more efficently. The Fokker- Planck equation is solved in continuous variable of oxygen atom numbers and time. From the calculated size distribution of precipitates, quantities of technological relevance such as the total precipitated amount of oxygen, the precipitate density and the average precipitate radius versus time can be determined. On the experimental side, in order to determine the precipitation rate R as a function of nucleation rate N and of growth rate G, $R = f(N, G)$, has to be determined as a function of time and

[13] M.Orlowski, Phys. Lett. A, **137**, 115, 1989 and N. Cowern, Appl. Phys. Lett. **54**, 703, 1989

[14] S. Aronowitz, J. Appl. Phys., **69**, 3901, 1991

[15] S. Coffa et al, Appl. Phys. Lett., **56**, 2405, 1990

[16] see for example P. Fahey et al, Rev. Mod. Phys., **61**, 289, 1989 and references therein

[17] P. Griffin et al, *"Consistent Models for Point Defects in Silicon"*, 1991 International Workshop on VLSI Process and Device Modeling, Oiso, Japan

[18] M. Schrems et al, Mat. Sci. Eng., B4, 393, 1989, and Proceedings of 20th ESSDERC 1990, p.201

[19] H. Kennel et al, current work at Stanford University. Also J. Plummer, *"Silicon Process Modeling"*, talk given Process Simulation and Modeling Workshop, MCNC, Research Triangle, 1990

temperature. It has been found that the nucleation rate is maximum around 650°C irrespective of oxygen concentration levels and that the growth rate increases with temperature displaying a characteristic maximum at specific time in the range between 1 and a few thousand minutes and decreasing with increasing temperature. However, although some oxygen may be trapped in some type of precipitates or clusters, the rest of this trapped impurity is dispersed on interstitial sites. Only upon going to lower temperatures, where the solubility of oxygen is much lower, precipitation and probably formation of SiO_2 micro-islands occurs. Kennel at Stanford has found that in FZ material the phosphorus diffusivity increases with increasing average rate of interstial oxygen concentration decrease. Under the same conditions the diffusivity of antimony decreases. These preliminary results do not at present reveal the interactions between oxygen precipitates, interstitial oxygen, and SiO_2 inclusions with point defects and dopants. Even more elusive are transient effects of precipitation and nucleation. Part of the explanation might well be due to microstresses generated by clusters of interstitials and of SiO_2 micro-domains. A hint along this lines is provided by an experiment [20] of the effect of stress of $SiNx$ on dopant diffusion in FZ and CZ silicon. The discrepancy between the dopant diffusivity for high stress (less diffusion) and low stress (more diffusion) is larger for FZ than for CZ material. This can be explained by assuming that in case of CZ the surface stress effect is weaker than in the case of FZ crystal, because of internal stresses due to oxygen precipitation in case of CZ.

2.2.3 Boundary Conditions for Dopant and Point Defects

It is well known that the nature of boundary conditions determines to a large extent the behavior of species not only near interfaces but also in the bulk. Consistent parameters for bulk behavior cannot be therefore obtained without proper description of boundary conditions. Boundary dopant conditions are commonly related to segregation phenomena. It is known, for example, that at the Si/SiO_2 interface arsenic piles up whereas boron piles down. Until recently the dopant transport across the interface has been modeled by a first-order kinetic model, yielding the total interface transport flux $F_s = h(C_1 - C_2/m)$, where h denotes the transport coefficient, $m = C_2^{eq}/C_1^{eq}$, the equilibrium segregation coefficient, and C_1 and C_2 denote the dopant concentrations at the interface that separates the bul phases 1 and 2, respectively. This model became increasingly unreliable, because it was not able to account for the properties of the interface itself and of the transient behavior of the segregation phenomenon. Recently, a new dynamic model [21] for dopant redistribution at interfaces has been proposed, in which a third phase, the interface layer itself, is considered in addition to the adjacent bulk phases. This model not only successfully describes segregation phenomena at various interfaces, including the polysilicon/silicon interface as in the case of emitter poly-outdiffusion, the transient behavior of the build-up of phosphorus accumulation (pileup) at the Si/SiO_2 interface, but it also gives a unambiguous prescription of the coupling between the dopant redistribution at the interface and the diffusion in bulk phases. Moreover it reproduces the first-order kinetic segregation model mentioned above in the limit of global and detailed balance conditions. In this case, it can be shown that the transport coefficient h is no longer a constant, but depends not only on absorption and emission coefficients but also on (equilibrium) concentrations on both sides of the interface. A simplified model [22] has been recently successfully used in modeling of polysilicon diffusion sources [23]. The problem with these models is, essentially, adequate parameter charac-

[20] P. Packan, Ph. D. dissertation, Stanford University, 1991

[21] M. Orlowski, Appl. Phys. Lett. **55**, 1762, 1989

[22] F. Lau et al, Appl. Phys.(Springer) A **49**, 671, 1989

[23] F. Lau, IEDM'90 Proceedings, p. 737, 1990

terization. Presently available experimental data is insuficient for the required calibration and novel experimental setups are required to obtain orthogonal sets of model parameters. Another formidable challenge facing the numerical evalution of this kind of segregation model is the regridding problem [24] of the interface region for moving nonplanar boundaries. This problem might perhaps be successfully addressed by a "continuum" approach to layer growth and to the resulting moving boundary.

This technique has been used to describe titanium silicide film growth [25] and is based on a smooth definition of the transition zone between two phases. Since this technique is perceived to be a promising one some more detailed explanation might be helpful. The model is capable of modeling arbitrarily abrupt transitions between $TiSi_2$ and Ti. The evolving layer can be divided into three regions: a surface region that consists mostly of $TiSi_2$ and a small amount of diffusing silicon, and an intermediate region of nearly constant width that contains Ti, Si, $TiSi$, and $TiSi_2$. The intermediate region is approximately delineated by the $TiSi$ distribution, which has a shape given by the approximate analytic formula:$C_{TiSi} \sim (e^{-k_1 Rx^2} - e^{-k_2 Rx^2})/(1 - k_2/k_1)$, where R is slowly changing parameter depending on silicon concentration and diffusivity, and k_1 and k_2 are reaction constants describing the silicidation process. It can be seen that the width of the transition region can be made very small by sufficiently large k_1 and k_2.

A similar approach has been applied by Rank [26] to the oxidation problem, considering the interface between silicon and silicon dioxide not as a sharp line but as a smooth transition layer. The interface is described by a normalized silicon concentration distribution constrained by the lower and upper bounds, 0 and 1. This distribution function is subject to an equation describing its evolution. The main advantage of this approach is that the finite element mesh remains *topologically invariant* during the simulation time, because nodes are only displaced with the growth of the oxide.

It has long been realized that oxidation enhanced and nitridation retarded diffusion of dopants such as phosphorus or boron is due to significant injection of silicon self-interstitials and vacanies, respectively. Great effort has been spent to determine the pertinent boundary conditions for point defects, the magnitude of the diffusivities, strength of possible bulk sources and sinks for point defects such as the presence and dynamics of stacking faults, the influnce of the wafer thickness, the width of the oxidizing window and so forth. The reader can find the relevant literature on this topic in the references of recent studies which will be quoted below. Here, for the sake of completeness it has to be mentioned, that in the last few years it has been found that silicon formation perturbs point defect concentrations dramatically (see also section 3.2) and as a result, changes dopant diffusion coefficients. One of the first studies on this topic by Wen [27] indicates that silicide formation is associated with enormous vacancy injection which is sufficient to annihilate completely end-of-range dislocation loops. The precise mechanism of vacancy injection during silicidation is presently unknown.

In contrast to silicidation, significant progress has recently been achieved in the modeling of self-interstitial kinetics near oxidizing silicon surfaces [28]. Due to the volume expansion concomitant with the oxidation reaction, excess silicon self-interstitials are generated at the Si/SiO_2 interface. The new model assumes that the generation rate of self-interstitials is proportional to the chemical reaction. To avoid a continuous buildup of interstitials at the interface, the

[24] M. Orlowski, Proceedings of NASECODE VI, p. 526, 1989

[25] L. Borucki et al, IEDM'88 Proceedings, p. 348, 1988

[26] E. Rank, Proceedings of NASECODE VI, p.40, 1989

[27] D.S. Wen et al, Appl. Phys. Lett., **51**, 1182, 1987

[28] K. Taniguchi et al, J. Appl. Phys., **65**, 2723, 1989, and S. Dunham, J. Electrochem. Soc., **136**, 250, 1989

following three anihilation mechanisms have been invoked: 1) interstitial flux into the oxide, 2) interstitial flux into the bulk silicon, and 3) surface recombination at surface kinks. What is new in this model is that the bulk of generated interstitials flows **into the oxide** where they quickly react with the incoming oxygen molecules or atoms. At the same time surface regrowth also annihilates self-interstitials at the interface. The flux into the bulk silicon, causing the anomalous impurity diffusion is negligibly small compared with the first two fluxes. Under these conditions, the concentration of self-interstitials is expressed as a function of the oxygen concentration at the interface

$$\frac{C_I}{C_I^{eq}} = \frac{\alpha R + 1}{\beta R^{1/2m} + 1} \tag{1}$$

where R is the oxidation rate and α and β are physical parameters composed of self-interstitial diffusivity, chemical reaction constants, and of the number of silicon atoms involved in a unit volume of SiO_2. This model explains also the stripe width dependence of nitridation enhanced diffusion of antimony, oxidation enhanced(OED) and oxidation retarded diffusion (ORD) of boron in HCl oxidation, ORD of boron and phosphorus at high temperature, sublinear oxygen pressure dependence on oxidation rate constant, and the crossover of oxidation rate between (100) and (111) orientation at low oxygen pressure.

This brief survey of boundary condition modeling shows an interesting characteristic of present process modeling activity that boundary conditions which have to be supplied to diffusion models as in the discussed cases, are subject themselves to considerable modeling effort. Additional challenge for the immediate future consists in an integration of the "stand-alone" models. For example, a unification of the segregation models for dopant redistribution at a moving interface with the oxidation model discussed above is badly needed.

2.2.4 High Concentration and Activation/Deactivation Effects

The processes of activation and deactivation of dopants, particularly at high concentrations are one of the least understood areas of process modeling despite the existence of much literature on the subject [29]. Until now it has been assumed that the discrepancy between chemical and electric dopant distributions is due to cluster formation, i.e. the aggregation of some number of dopant atoms. Besides the fact that the cluster size, or in the case of various sizes, the size disribution, the charge state of the clusters, their formation and decay kinetics are not well known, there is even no consensus on whether the cluster model is the dominant mechanism, if at all, for dopant deactivation. It could well be that inactivation is due to the formation of a new phase with possibly long range interaction within the crystal lattice. In technical terms the difference might consist of unknown interactions between clusters, which bring the dynamics of such a phase beyond that of pure cluster modeling. Other issues like the interaction of deactivated dopant with point defects, with the annealing of primary crystalline structure, with the annealing of amorphous or partially amorphous layers, with interfaces under various conditions, with lattice stresses, with band gap narrowing effects, or with other impurities, like other dopants and oxygen, for example, and finally the issue of dynamic effects at short annealing times (30 min and less) are, presently, even more difficult to address. In short, despite

[29] A lot of work on clustering has been done at the University of Bologna, see, for example, R. Angelucci et at,J. Electrochem. Soc. **132**, 2726, 1985; in Vienna, see, for example, E. Guerrero et al, J. Electrochem. Soc., **129**, 1826, 1982; by Fair and collaborators, see, for example, R.B. Fair in "Semiconductor Silicon", edited by H.R. Huff et al, Proceedings of Electrochem. Soc., 963, 1981; at Stanford University, see, for example, A. Lietoila et al, Appl. Phys. Lett., **36**, 765 (1980), and in other places including significant work in Japan.

a large body of experimental data, at the present stage the modeling of dopant deactivation has to be classified as purely phenomenological and is mostly based hitherto on the cluster model [30]. In the past, characterized by large dimensions and high thermal budgets the cluster model was a nice looking "physical" model to parametrize the temperature solubility limit of a specific dopant in silicon as a function of temperature. Today, it is known that in most cases either times or temperatures are not large enough to drive activation and deactivation processes into equilibrium and that the actual level of activation might also depend on thermal history. The transient effects of deactivation are however important for the diffusion phenomena because the concentration gradients of substitutional dopants determine for most part the diffusion fluxes at every point in time during the anneal. If one assumes that clustering takes place on a much smaller time scale than the diffusion, then in high concentration regions the gradient of substitutional dopant vanishes and consequently the dopant flux is zero. If this is however incorrect, because deactivation is not instantaneous but operates on a time scale comparable to that of diffusion, then this kind of modelling entails an artificial enhancement of dopant diffusivity necessary to reproduce the experimental results. Overall, one gets an ill-calibrated model with no predictive capability.

In view of these uncertainties and lack of fundamental understanding Subrahmanyan et. al. [31] offered a new methodological approach to study arsenic deactivation based on a dynamic equation characterizing the two competing mechanisms of activation and deactivation in a most general way, leaving open what these mechanisms might be:

$$\frac{\partial C_{active}}{\partial t} = K_A \mathcal{F}(C_{inactive}) - K_D \mathcal{R}(C_{active}) \tag{2}$$

where K_A and K_D are activation and deactivation coefficients, respectively, and $\mathcal{F}$ and $\mathcal{R}$ are model dependent functions of the active and inactive concentrations [32]. Under these premises and given the experimental data any appropriate scaling of K_A for comparison with K_D should lead to the same conclusion once the above equation has been calibrated and suitable values for K_A and K_D have been obtained for the specific model-dependent functions $\mathcal{F}$ and $\mathcal{R}$. The general form of the equation suggested three questions: 1) What is the initial condition for an isothermal clustering model? Is $C_{act}(t = 0) = C_{chem}$, or is $C_{act}(t = 0) = C_{solid.sol.}$, or is it something else? 2) What are the characteristic times for clustering and declustering rates as a function of temperature and local concentration? And related to this, is the equilibriun state reached during typical diffusion times? 3) If not, what is the influence of the annealing time and of ramp-down conditions on the clustering and hence on the electrical activation?

From suitably designed experiments it has been found that a) the initial activation level is set during the solid-phase epitaxial regrowth almost instantaneously at 700°C and is higher than the solid solubility limit but lower than the total chemical concentration. Because of the rapidity of the regrowth at a low temperature this initial condition is the **same** for all isothermal anneals if the ramp-up rate is sufficiently fast. Furthermore, it has been found that during isothermal anneal the activation rate is higher than the deactivation rate at high temperatures, but lower at low temperatures, with a crossover at about 840°C. Using the model of Tsai et al [33] as a convenient parametrization it has been found that less than 5 *min* are needed at 1000°C to reach equilibrium electrical concentration, 60 *min* at 900°C, and even longer at lower temperatures. This implies that for typical anneals of less than 30 *min* at 900°C, it is not

[30] See SUPREM III manual, Stanford University, for the conventional cluster models

[31] R. Subrahmanyan et al, Proceedings IEDM'90, p.749, 1990

[32] In the most general case $\mathcal{F}$ could also depend on active and $\mathcal{R}$ on inactive concentration.

[33] M. Tsai et al, J. Appl. Phys. **51**, 3230, 1980

correct to assume that the electrical concentration is at equilibrium, with all the implications for diffusion dynamics discussed above. Based on this analysis and recalibration of the model a prediction has been made that the sheet resistance might vary up to 30% depending only on the ramp-down condition (slow versus rapid ramp down) which has been confirmed by experiments.

The relation between inactive/immobile [34] and substitutional boron atoms at high concentration in conjunction with transient effects has been recently investigated by Cowern [35]. It has been found that the transient diffusion is a multi-step migration process. It can be pictured accurately as "normal" diffusion, accelerated by the interstitial supersaturation $C_I/C_I^{eq} >> 1$. Anomalous boron transient diffusion characterized by static peak and enhanced tail diffusion arise from the presence of two components: one diffusing and electrically active (B_s), the other static and inactive (B_i). Both components become mixed during the anneal according to the kick-out and its inverse mechanism: $B_i \rightleftarrows B_s + I$, supplemented with a clustering reaction of the following type: $mB_s + nI \rightleftarrows IDC$. IDC is treated here as a cluster containing m B atoms, formed with net absorption of n Si interstitials. The evolution during the anneal towards normal diffusion occurs on **two** time scales: a) time for transient diffusion due to primary damage and b) time scale for trapped boron release, which is much larger than the time scale for transient diffusion. The transient diffusion is confined to concentrations below a critical concentration value. This leads to a static peak and a broader diffused region at lower concentrations. Secondary transient effects, such as point defect generation during the release process could be also important.

The preceding discussion of technology issues in the area of diffusion-reaction phenomena shows an extraordinary complexity of mechanisms which are rarely in equilibrium for current technologies thus calling for a transient simulation capability. Moreover, this complex "internal" behavior is coupled with no less involved behavior of external boundary conditions which as mentioned in section 1.7 make coupling of process with equipment simulation imperative. Since other areas of simulation which are briefly mentioned in the following section are no less demanding the fomidable challenges confronting process modeling in the 90's can well be appreciated.

2.3 Summary of other Technological Issues

Reaction-diffusion phenomena related to silicon have been the traditional realm of process modeling and simulation. However, issues related to diffusion stand in close and often causal relationship with other engineering issues. For example, silicide formation raises the question of elastic and thermal stability properties, of epitaxial growth versus nucleation, intra- and inter-granular precipitation, slip mechanisms, grain misfit dislocation, creep properties and so on. All boundary conditions are affected to some extent by surface and interfacial properties. These properties in turn are influenced by passivation or surface migration properties. Another important area of process modeling will be the engineering of thermal expansion coefficients on a microscopic scale. Here, process modeling will be challenged to understand the role of dopant admixtures in order to match thermal expansion coefficients as far as possible. In enumerating these areas the scope of this review is restricted to silicon based technology. But it is clear that arduous tasks lie ahead for process simulation in GaAs based and in adjunct technologies such as optoelectronics and high temperature superconductivity which might be merged with silicon technology before too long.

[34] inactive and immobile dopants are not always synonymous

[35] N. Cowern et al, *"Transient diffusion of ion-implanted B in Si: dose, time and matrix dependence of atomic and electrical profiles"*, to be published in J. Appl. Phys.

3 Novel Modeling Principles

In the preceeding sections some novel techniques like the "continuum" method of treating moving boundaries due to interface reactions have been already mentioned. Thermodynamics-based concepts are advantageous in characterizing complex phenomena which are statistical in nature, like phase transition, nucleation, or densification effects. Crystallographic theories can help to understand strained layer (super)lattices, slip, dislocation or fracture effects. These concepts are not well known to the process modeling community. In the following, two examples, familiar to the author, the concept of fractal mathematics and of a model stress-induced migration based on Fokker-Planck equation will be presented to illustrate their potential benefits.

3.1 Concept of Fractal Dimension

Science often advances through the introduction of new ideas which simplify the understanding of complex problems. Concepts such as nucleation, aggregation, and spinodal decomposition, have played an essential role in the modern understanding of the structure of materials. More recently, fractal geometry has emerged as an essential idea for understanding the kinetic growth and properties of disordered materials. Since this concept was absent in the traditional scope of semiconductor process modeling and simulation a more detailed exposition of this concept seems to be in order.

Although the field is in its infancy, it seems clear already that fractal geometry is one of the key simplifing concepts that will allow material scientists to understand and to **control** a variety of disorderly growth processes that can dramatically modify the structure and properties of materials [36]. It turns out that complicated, seemingly disorderly structures can be characterized through a single number d, the fractal dimension. The fractal dimension can be defined as the exponent which relates the mass M, of an object to its size S.

$$M \sim S^d \tag{3}$$

For fractal objects, the exponent d need not to be integral. Fractal objects may differ drastically from one another, but they share a common feature - self-similarity. Self-similarity or dilatation symmetry means that the object looks the same under transformation of scales, such as increasing resolution of the measurement. Within certain limits, the essential geometric features of a fractal object repeat themselves on some scale, or are not changed upon magnification. To be more specific: A line is one dimensional, a square two-dimensional, a cube three-dimensional. But a crinky line, like the coastline of Britain, is more than one-dimensional and less than two-dimensional, it has actually around 1.25 dimensions. Roughly speaking, if something has more than one but less than two fractal dimensions it is better at filling the space than a one-dimensional object, but not quite so good as a two-dimensional one. It is clear, how this concept can be applied to issues of material densification. An initial material with a fractal dimension 2.2 can be more densified than a material with a fractal dimension of 2.8. Armed with a technique for measuring the irregularity of shapes, the theory of fractals is now being applied to protein structure, acid rain, earthquakes, the fluctuation of exchange rates, oil extraction, conduction in porous materials, epidemics, corrosion, brittle materials, music, distribution of galaxies, the level of water in the rivers, the shapes of clouds, mountains, lakes, trees, and snow flakes. Nearly every aspect of science can be addressed by the concept of fractals, because all aspects of nature involve some roughness and irregularity.

[36] For an introducton into fractal concepts see for example: *Fractals in Physics*, edited by L. Pietronero and E. Tosatti (North-Holland, Amsterdam, 1986)

Returning to the discussion of material sciences, the fractal dimension is not a purely theoretic concept, but fortunately can be directly measured through the use of scattering techniques. In scattering techniques, an incident beam (light, neutrons, or x-rays) impinges on a sample and the angular dependence of the scattered intensity is measured. For fractal geometry, the intensity profiles always exhibit power-law dependence, when plotted versus the wave vector k,

$$I \sim k^p \tag{4}$$

The exponent p, the so called *Porod slope* is directly related to the fractal dimension. Through Bragg's law, the wave factor k can be related to a characteristic length l and the scattering angle. By scanning the scattering angle, one effectively studies an object at different length scales. Using a combination of different beams, it is possible to probe length scales from 1 $\AA$ to 1 μm. This measurement technique is based on a simple idea of scaling proposed by Hausdorff in 1919 to measure the same object with different units of measurement. If one reduces the measuring unit by an integer N then the number of times the reduced measuring unit fits into the object under measurement will increase by N for a one-dimensional, by N^2 for a two-dimensional, and by N^3 for a three-dimensional object; however, for a object of fractal dimension it will - as might be expected by now - increase by N^d and d being no longer an integer but a number $1 \leq d \leq 3$.

The concept of fractals can be introduced into process modeling to describe the transport properties of any species being transported in irregular media such as a network of grain boundaries. An example for diffusion on networks is the recent data on fluorine redistribution in polysilicon layers obtained at Motorola's laboratories [37]. After fluorine implantation and during the subsequent anneal at elevated temperatures the fluorine profile becomes narrower instead of broader as the time or/and the anneal temperature (i.e. so called $D \cdot t$ product) increase. Outside of the narrow peak the fluorine distribution is essentially flat. This unusual behavior which cannot be explained by any bulk diffusion phenomenon can be rationalized by considering two fluorine species, one inside the grains, and another within the grain boundaries. It is assumed that that the mobility of fluorine in the grain boundaries is much higher than inside the grains. So in order for fluorine to be transported away from its original site the fluorine first has to be ejected from the grains into the grain boundary, where it can move on the grain boundaries like in a network of channels. In order to model the fluorine diffusivity on the grain boundary network consistently, a law by Archie, which was used in the late 40's to model the electrical conductivity of salt-water saturated porous rocks, must be invoked: $D = D_o \phi^m$, where D_o describes the intrinsic fluorine diffusion of the grain boundary and the exponent m and ϕ, are parameters reflecting the network connectivity or its *fractal dimension* [38]. The strength of the concept of the diffusion on fractal networks is that it characterizes the highly complex medium consisting of grains and grain boundaries with just two well understood parameters and that it casts the entire problem into a wellknown framework of diffusion theory.

3.2 Modeling of Stress Effects

Stress analysis in process simulation has been applied almost exclusively in studies of silicon oxidation and to some extent in reliability studies of interconnect lines. For some reason, however, the latter topic has never been considered to be in the mainstream of process simulation

[37] H. Tseng et al, Proceedings of the Spring Meeting of the Electrochem. Society in Washington D.C., abstract no.402, p. 613

[38] M. Orlowski et al, submitted to IEDM'91.

in its traditional understanding. As already alluded to in the section 1.8 on interdisciplinary collaboration, an impressive body of theoretical methods to determine the internal forces within a solid as the effect of externally applied forces has been already developed in mechanical engineering and such different sciences such as geology and metallurgy. These concepts have to be evaluated and when occasion arises, applied to problems encountered in semiconductor engineering. However, it has to be expected that some new model development is in order, because of the advent of engineering on microscopic scale. Fields of potential application of stress analysis are briefly reviewed in the following.

The important field of stresses in thin films has been persistently ignored in the past by the process simulation community, despite numerous experimental papers and despite formidable challenges for the process engineers. In this context the stress analysis of multilayer structures calls for particular attention. Significant work has been done on stresses in superlattices in compound semiconductors, but this knowledge didn't spread much to the research community involved in silicon based devices. It is quite certain that stresses on microscopic scale are an important element in the overall description of mass impurity transport in materials used in silicon-based technologies. There is, for example, growing evidence that microstresses caused by impurity clustering, particularly by oxygen precipitation, accounts for differences in species transport in CZ and FZ silicon as mentioned in section 2.2. Finally, despite substantial progress in diffusion theory, the transport of species driven by gradients of stress distribution has not been properly formulated. At the end of this section a novel approach to describe migration in stress gradients derived self-consistently from first principle in the framework of the diffusion models based on point defect kinetics will be attempted for the first time. Moreover, recent experimental evidence (for reference, see footnote 2) of anomalous asymetrical diffusion at temperatures low as 200^oC will be presented and it will be argued that this phenomenon can be readily explained by the present approach of stress gradient induced migration. A completely untouched area of great relevance is a modeling and simulation capability to predict stress induced material (bulk and interface) defects.

Stress effects in silicon oxidation have been long suspected since the observation of oxidation rate retardation around the convex and cocave corners of oxidized trench walls in silicon. The cylindrical oxide structures of Kao [39] helped quantify this phenomenon, and showed that the retardation is greater at convex than concave corners. The experiments showed also that there is feedback from stresses to the oxidation rate coefficients, thus introducing a nonlinearity which must be considered in a self-consistent description of the oxidation process. This effect is even significant for planar oxides. It has been later found that the diffusivity of oxidant in SiO_2 fims depends not only on the temperature and, on the stress, but also on the thermal history of the films. Low temperature oxides grow in a state of compressive stress and are denser than oxides grown at high temperatures. A second kind of nonlinearity is associated with the oxide flow. It has been found that if the measured viscosity of SiO_2 is substituted in a model using constant viscosity, the predicted stresses around silicon corners are large enough to cause mechanical rupture. Since such failures are not observed experimentally, viscosity must be reduced significantly by the flow. The challenges facing two-dimensional modeling of oxide flow are to experimentally determine the stress dependence of the Deal-Grove growth coefficients , the low-stress viscosity and its modification under high stress. This program has been pursued most successfully at Stanford University, where most recently Griffin and Rafferty [40] have used the known properties of the oxide layer to gauge the properties of the Si_3N_4 layer at

[39] D-B. Kao, Ph.D. Dissertation, Stanford Univesity, 1986

[40] P. Griffin and C. Rafferty, IEDM'90 Proceedings, p. 741

processing temperatures. Until now nitride films have been treated as elastic solids at processing temperatures. An analytical estimate of the relaxation time τ based on elastic nitride model leads to $\tau \approx E/\mu \times L^4/T^3 d$ where μ is the oxide viscosity, E is the nitride Young's modulus, T, L and d are the nitride thickness, deformed length and displacement from equilibrium, which amounts to $\approx$ one hour at $1100°C$. Experimentally no nitride relaxation has been observed. Griffin and Rafferty's results therefore favor a viscous model of nitride deformation over an elastic model, and provide the first experimental estimates of thin-film Si_3N_4 viscosity as a function of temperature. The viscosity of thin nitride films depends on the stoichiometry of the deposited films and is well modeled by an Arrhenius dependence over typical processing temperatures. This extremenly valuable technique can be now extended to other materials and will help determine their elastic and viscous properties.

Another challenging issue is to incorporate the impact of stress into current point defect based transport models. The present derivation is based on the technique [41] of evaluation of elementary jump frequencies within the framework of the master equation and employing the theory of absolute reaction rates in conjunction with crystal lattice strain energy model invoked by Clarence Zener [42]. First the diffusion in spatially nonuniform stress fields will be considered. For a spatially nonuniform stress field $\sigma = \sigma(\vec{x})$, it is shown, that the commonly used Fick-type equation to describe stress dependent diffusion of atomic species C, $\partial C/\partial t = \nabla(D_{eff}\nabla C)$ with $D_{eff} = D_o F$, and F being a function of the stress field, $F = F(\sigma)$, is incorrect - except for special cases such as of hydrostatic pressure - and has to be replaced by an equation of a Fokker-Planck type proposed here.

$$\frac{\partial C}{\partial t} = \nabla(D_{eff}\nabla C + \mu_{stress}(\sigma(\vec{x}), C)\nabla\sigma) \tag{5}$$

where μ_{stress} can be defined as a dopant mobility with respect to stress gradient in units of $[cm \cdot sec \cdot MPa]^{-1}$. Following the strain energy model, the effective diffusivity is given by $D_{eff} = D_o F_o exp(-\sigma\Delta V/kT)$, where D_o represents dopant diffusivity unpertubed by stress fields. ΔV is the activation volume defined as the partial derivative of the free energy, ΔG, with respect to the stress at constant temperature: $\Delta V = (\partial G/\partial\sigma)_T$. If one assumes that the diffusion process under consideration involves both creation and migration of point defects, the free energy ΔG will become a sum of terms, $\Delta G_f + \Delta G_m$, and correspondingly, the activation volume will become a sum of terms ΔV_f, the change in volume of the point defect formation, and ΔV_m, the lattice dilation attending the elementary diffusion jump. Under these circumstances the stress mobility can be written in the following way:

$$\mu_{stress}(\sigma(\vec{x}, C)) = D_{eff} \cdot C\frac{\Delta V_m - \Delta V_f}{kT} \tag{6}$$

The last result is interesting because the mobility can change its sign, but this cannot be discussed here [43]. The present derivation provides also, for the first time, a direct link between the absolute reaction rate theory of the elementary diffusional jump and the phenomenological thermodynamic theory describing species redistribution in terms of gradients of a complete thermodynamic potential, i.e. sum of chemical, thermal, electrical, and mechanical potentials with purely phenomenological coefficients being only constrained by the Onsager relations. In particular, the coefficient of the stress gradient field is shown to be proportional to the difference

[41] M.Orlowski, IEDM'90 Proceedings, p. 729

[42] C. Zener, Acta Cryst., vol.2, p.163, 1949

[43] M. Orlowski, in preparation

of the activation volumes associated with the lattice dilatation attending the defect migration in excited state to the destination site, ΔV_m, and with the transition from ground state to the excited state, ΔV_f, whereas the coefficient of the concentration gradient, D_{eff}, is a function of the sum of the activation volumes: $\Delta V = \Delta V_m + \Delta V_f$. It can easily be shown that in the case of hydrostatic pressure the Fokker-Planck equation derived above, reduces to a diffusion equation with effective diffusion equation coefficent D_{eff}, as assumed by almost all researchers in this field.

It will be now shown that the additional term in the Fokker-Planck equation containing the gradient of the stress field, the *friction term* [44] , is responsible for asymetric broadening of dopant diffusion profiles observed by IBM reserachers [4]. They have observed during Pd_2Si formation at $200°C$ substantial asymmetric diffusional broadening of buried marker layers, with diffusion occuring preferentially towards the siliciding interface. In discussing possible explanations mainly based on the assumption of point defect (here vacancy) injection they concluded that the extant diffusion models cannot account for the new diffusion phenomenon. Here, an alternative explanation is put forward. The possibility of creating excess vacanies in quantities to cause such a sizeable diffusion effects during silicidation at temperatures around $200°C$ can easily be dismissed. A clear indication of this is that the same effect is seen for antimony boron, and gallium. Since antimony diffuses mainly via vacancies, and boron and gallium via interstitials, only antimony should diffuse, and the diffusion for boron and gallium due to the undersaturation of interstitials should be strongly surpressed. The result of the experiment was that in all cases the same asymetrical broadening occurs indicates that the cause of it is not due to point defects, but to some other common driving force. One can easily imagine that this driving force is due to the gradient in the stress generated during the initial stage of silicidation. The reaction of Pd_2Si formation is rapid even at low temperatures and it creates high tensile stress at the interface. Since at low temperature the silicon crystal is very stiff in terms of elasticity this results in a very steep gradient in the surface region. Since the flux is proportional to the gradient of the stress the flux is approximately constant over significant portions of the surface silicon layer. This explains the results of the experiment with double marker layer $0.15\mu m$ apart, where both markers show the same amount of asymetrical broadening. The gradient created by tensile stress during the initial stage of silicidation travels into the bulk crystal with considerable speed. This explains the observation that enhanced dopant diffusion occurs at the beginning of the silicide formation and does not seem to continue for the entire silicidation process. At elevated temperatures this effect is weak, if present at all, because the heated crystal does not allow for a buildup of a sufficiently strong stress gradient. This phenomenon is bound to play an important role in thin film processing at low temperatures.

Acknowledgements

The author would like to acknowledge many helpful discussions on process simulation with his colleagues at Motorola, in particular with Dr. Ravi Subrahmanyan, Dr. Bob Lojek, and Dr. Len Borucki, and with many researchers outside of Motorola, in particular with Dr. Sheldon Aronowitz, Dr. Nick Cowern, Prof. Robert Dutton, Prof. Richard Fair, Prof. Martin Giles, Prof. Ulrich Gösele, Dr. Peter Griffin, Dr. Frank Lau, Prof. Mark Law, Dr. Jürgen Lorenz, Dr. F. Morehead, Dr. Brian Mulvaney, Dr. K. Nishi, Dr. Peter Pichler, Prof. James Plummer, Prof. Hans Pötzl, Prof. Siegfried Selberherr, Prof. Y. Tan, Prof. Kenji Taniguchi, Prof. Al Tasch, Prof. James van Vechten, Prof. Masayuki Yoshida, and many others, as well as the kind support of Dr. Charles Meyer and Dr. Lou Parrillo at Motorola's Advanced Products Research and Development Laboratory.

[44] H. Risken, *The Fokker-Planck Equation*, Springer, New York, 1984