JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

1/2004

Advanced silicon devices and technologies for ULSI era

Special issue edited by Andrzej Jakubowski and Lidia Łukasiak

Part 1

Silicon everywhere
K. J. Jakubowski — Invited paper — 3

Silicon microelectronics: where we have come from and where we are heading
L. Łukasiak, A. Jakubowski, and Z. Pióro — Invited paper — 7

Critical modeling issues of SiGe semiconductor devices
V. Palankovski and S. Selberherr — Invited paper — 15

Recent developments in vertical MOSFETs and SiGe HBTs
S. Hall, O. Buiu, P. Ashburn, and K. de Groot — Invited paper — 26

Modeling SiGe-base HBT using APSYS 2000 - a 2D simulator
A. Linkowski, L. Łukasiak, and A. Jakubowski — Paper — 36

Scattering mechanisms in MOS/SOI devices with ultrathin semiconductor layers
J. Walczak and B. Majkusiak — Invited paper — 39

High-field current transport and charge trapping in buried oxide of SOI materials under high-field electron injection
A. N. Nazarov, Y. Houk, and V. I. Kilchytska — Invited paper — 50

Ultrathin oxynitride films for CMOS technology
R. B. Beck and A. Jakubowski — Invited paper — 62

Closed-form 2D modeling of sub-100 nm MOSFETs in the subthreshold regime
J. Osthaug, T. A. Fjeldly, and B. Iñiguez — Invited paper — 70

Editorial Board

Editor-in Chief: .... *Paweł Szczepański*

Associate Editors: .... *Krzysztof Borzycki*
Marek Jaworski

Managing Editor: .... *Maria Łopuszniak*

Technical Editor: .... *Anna Tyszka-Zawadzka*

Editorial Advisory Board

Chairman: .... *Andrzej Jajszczyk*
Marek Amanowicz
Daniel Bem
Andrzej Hildebrandt
Witold Hołubowicz
Andrzej Jakubowski
Alina Karwowska-Lamparska
Marian Kowalewski
Andrzej Kowalski
Józef Lubacz
Krzysztof Malinowski
Marian Marciniak
Józef Modelski
Ewa Orłowska
Andrzej Pach
Zdzisław Papir
Janusz Stokłosa
Wiesław Traczyk
Andrzej P. Wierzbicki
Tadeusz Więckowski
Tadeusz A. Wysocki
Jan Zabrodzki
Andrzej Zieliński

ISSN 1509-4553
© Copyright by National Institute of Telecommunications
Warsaw 2004

Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

Despite the fact that a range of limitations are beginning to appear as CMOS technology is being raised to ever higher levels of perfection, it is anticipated that silicon will be the dominant material of the semiconductor industry for at least the first half of the 21st century. The forecast for microelectronics development published in 2001 by SIA (Semiconductor Industry Association) reaches ahead to the years 2010–2016. Moreover, a comparison with former SIA forecasts indicates that they become more aggressive (that is more optimistic) with time.

While the development of silicon microelectronics in the past could be attributed mostly to the reduction of the feature size (progress in lithography), today it relies more on new material solutions, such as SOI, SON, SiGe or SiC. The combination of this trend with continuous miniaturization provides the opportunity of improving IC functionality and speed of operation.

Telecommunications and information technology are arguably the most powerful drivers behind microelectronics product development nowadays. Thus plenty of new applications are being created for fast analog and RF circuits, as well as for information processing ones. It is clear that with the anticipated $f_{\max} = 130$ GHz and $f_T = 65$ GHz to be reached by RF bipolar transistors in 2005, according to the 2001 issue of ITRS (it is interesting to note that in 1999 ITRS predicted only $f_{\max} = 50$ GHz and $f_T = 40$ GHz for 2005), a lot of effort must be put into the development of appropriate material, processing, characterization and modeling. Such an outstanding progress, however, will not happen without increased speed offered by new materials solutions. SiGe-base HBTs may serve as an example here – a device with $f_T = 350$ GHz has already been reported in the literature.

High-speed isn't, however, everything. Portable wireless products push, for obvious reasons, for low-power solutions. This trend requires new architectural solutions (e.g. channel thinning), and in consequence, new material, such as SOI (or its possible successor SON – silicon-on-nothing), where current driveability is considerably higher than in conventional MOSFETs.

In this issue the Reader will find a selection of papers and lectures presented during the 6th Symposium Diagnostics & Yield: Advanced Silicon Devices and Technologies for ULSI Era, which took place in Museum of Earth, Warsaw, Poland on June 22–25, 2003. A number of the papers are devoted to the most important issues concerning semiconductor technology, performance of state-of-the-art devices and advanced materials, such as SOI, SiGe and SiGeC. Several papers address the problems of modeling and characterization and the broad subject of gate dielectrics is covered, too.

We hope the Readers will find these Proceedings useful and interesting.

Organization of this Symposium would not be possible without the support of Polish Committee for Scientific Research (research projects no. 4T11B03523, 8T11B07519, 8T11B07419). The organizers acknowledge also the contribution of the Committee of Electronics and Telecommunications of Polish Academy of Sciences, and the NEXUS Microsystems Association. Finally, we are also grateful to the National Institute of Telecommunications for making this journal available for the publication of the Symposium papers.

Andrzej Jakubowski – General Chairman
Lidia Łukasiak – Programme Chairman
Guest Editors

Invited paper

Silicon everywhere

Krzysztof J. Jakubowski

Abstract — The paper discusses the ubiquity of silicon and its compounds in the geo-chemical structure of the Earth. The dominant role of mineral oxidized compounds of silicon as the building material of the primordial magma, metamorphic and sedimentary rocks is also shown. The reader's attention is also attracted to the numerous varieties of silicon compounds valued for their intrinsic beauty. Finally, methods of silicon production for electronics purposes are briefly addressed.

Keywords — *silicon, silica, quartz, geo-chemical structure.*

The ubiquity of silicon is a characteristic feature of the geo-chemical structure of the Earth. The popular slogan: "Earth – silicon planet" is thus fully justified. Silicon is, beside oxygen, the most common element of the external, solid part of the Earth also known as the earth crust or lithosphere. Its average content is estimated to be in the range of 26–29 per cent by weight depending on the adopted calculation method. The mineral silicate fraction is also the dominant component of the chemical composition of meteorites. This indicates a privileged position of silicon in the primordial matter of the Solar System.

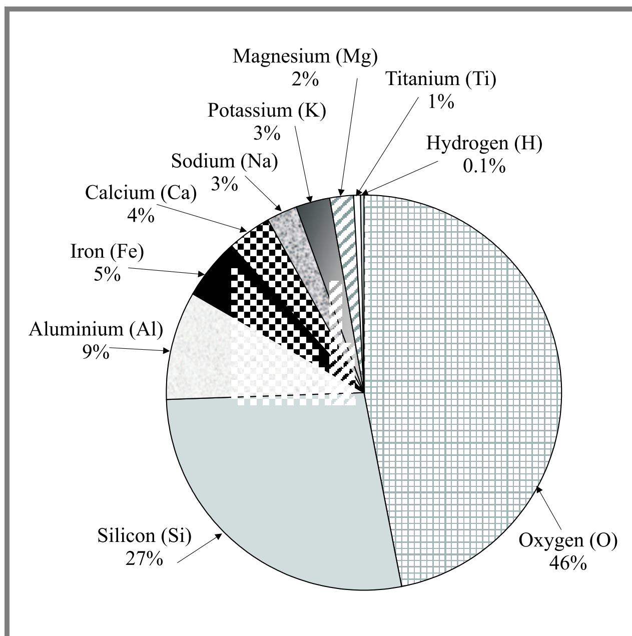


Fig. 1. Composition of the lithosphere.

It should be noted that few elements only contribute significantly to the lithosphere. Based on thousands of analyses it is estimated that the average content of those elements (expressed in per cent by weight – see Fig. 1) is following.

Oxygen and silicon have, therefore, an enormous advantage over the next 6 elements that constitute merely a few per cent of the lithosphere. It is, thus, no accident that these elements are considered to be the main rock-forming components.

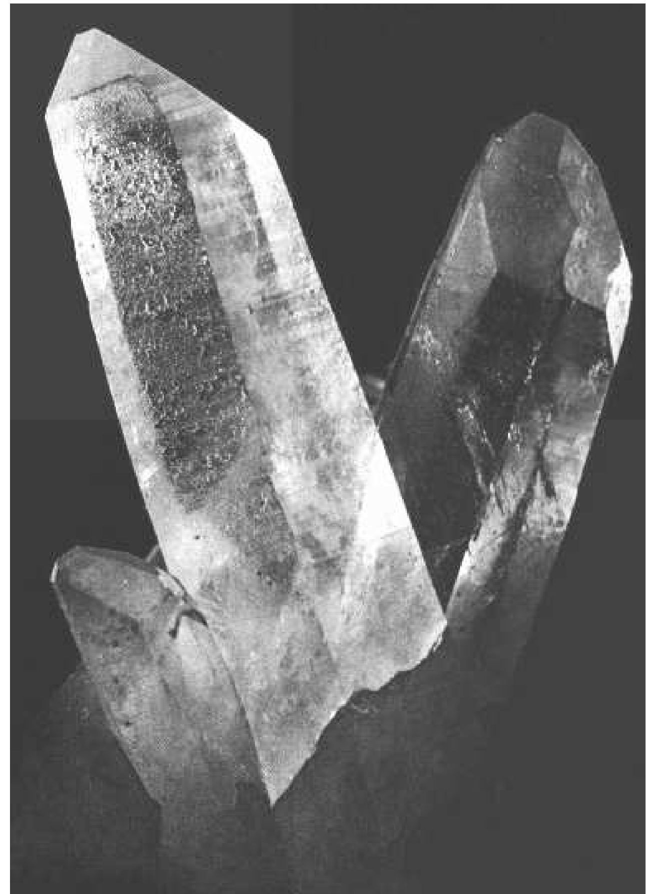


Fig. 2. Natural quartz crystals.

Silicon does not exist naturally in its pure form. It is almost always oxidized, the most often as silicon dioxide SiO_2 also known as silica (Fig. 2). Silicon is extremely oxyphilic and its bonds to oxygen in silicate crystal lattices are very durable in all geo-chemical conditions. $[SiO_2]^4$-anions are the basic component of the silicate crystal structure (Fig. 3).

Together with aluminosilicates, silicates form the biggest family of minerals, constituting 87 per cent (!) of the components building the lithosphere. Rough estimates of the average content of the rock-forming minerals in the lithosphere are thus unambiguous, as indicated in Table 1.

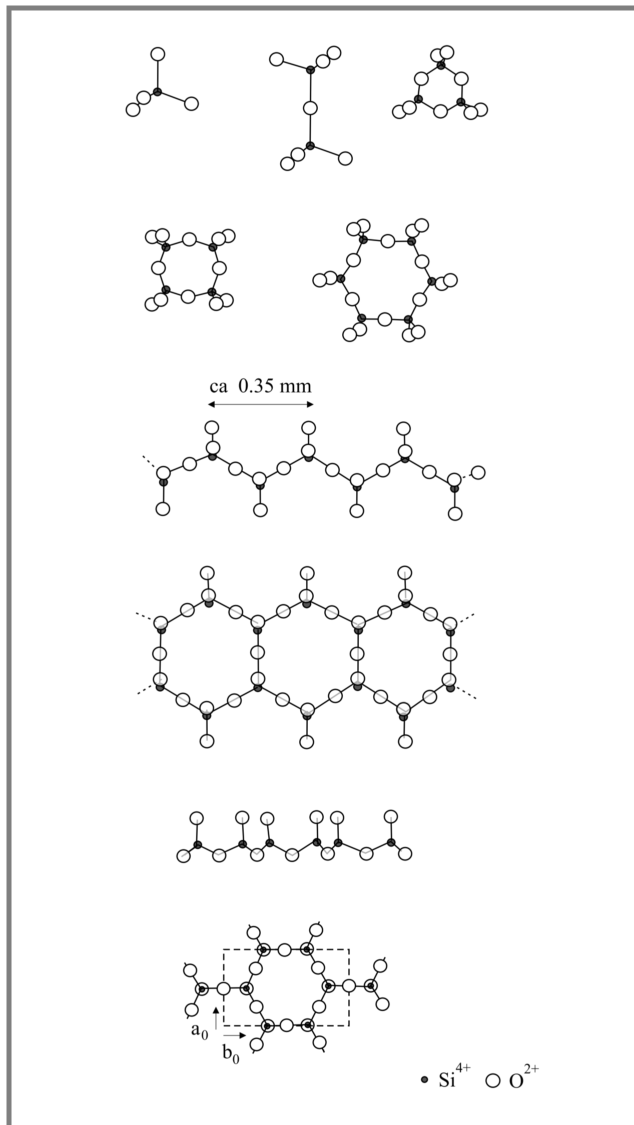


Fig. 3. Basic structures of the silicon-oxygen anions in silicate crystal lattices.

The real "power" of silicon may be fully appreciated only when the information presented in Table 1 is taken into account. It is thus clear that silicon is rightly considered as one of the most important lithophilic elements. In other words, mineral oxidized compounds of silicon are the dominant building material of the primordial magmatic rocks, such as the most widespread granite and basalt that form the core of the lithosphere. Silica is also abundant in metamorphic rocks and is easily found in sedimentary ones. In short, it is present in every geo-chemical environment. In oceanic and fluvial waters significant amounts of extremely diluted silica (fluvial waters – 25–30 g/t, oceanic waters – 6 g/t) are present, too. Silica may reach higher concentrations with the help of certain live organisms building silica skeletons. Such valuable minerals as diatomaceous earth or diatomites are created from clusters of the skeletons of certain single-cell algae known as diatoms (*diatomae*). The organogenic siliceous rocks include also radiolarites (formed of clusters of protozoa called radiolaria) and spongiolithes (consisting of *silicispongiae*). The presence of silicon in the tissues of live organisms is a separate issue. Silicon is especially abundant in certain plants where unusual combinations of silicon and organic compounds are observed.

Table 1
Average content of rock-forming minerals in the lithosphere

Mineral	Simplified chemical composition	Percentage by weight
Feldspars	$Ca[Al_2Si_2O_8]$; $Na[AlSi_3O_8]$; $K[AlSiO_3O_8]$	58
Quartz	SiO_2	12.5
Pyroxenes	Ca, $(Mg,Fe)_2[Si_2O_6]$	12.0
Mica	$K(Mg,Fe,Mn)_3$ $[(OH)_2AlSi_3O_{10}]$	3.5
Olivines (chrysolites)	$(Mg,Fe)_2[SiO_4]$	2.6
Amphiboles	$Ca_2(Mg,Fe)_5$ $[(OH)Si_4O_{11}]_2$	1.7
Silty minerals	$Al_4[OH_8/Si_4O_{10}]$ (example)	1.0

Compounds of oxidized silicon are the basis of extremely numerous mineral forms with very different structural, morphologic, physical and chemical properties. Minerals belonging to the group of silicates and aluminosilicates are the most common rock-forming minerals that are also valued in practical applications that include all areas of material culture and technology. The very beginnings of the use of silica go back to the Paleolithic Age, where the ability to fabricate flint tools was the measure of the technological progress. Nowadays synthetic single crystals of quartz and silicon wafers are simply indispensable for the sophisticated technology of the materials required by modern microelectronics and optoelectronics.

One of the most efficient ways to obtain pure silicon is to use minerals belonging to the class of spatial silicates, especially SiO_2, which forms in itself a separate group of minerals. Numerous polymorphic varieties include **tridymite**, **cristobalite** and the most common **quartz** taking a high-temperature form (hexagonal α-quartz stable in the range of 573–850°C) and a low-temperature one (trigonal β-quartz stable at temperatures below 573°C) (Fig. 4).

Let us also mention numerous varieties valued for their color, that is transparent and colorless *rock crystal*, dark-brown *smoky quartz*, black *morion*, violet *amethyst*, yellow *citrin* and *rose quartz*. Many colorful varieties of quartz take their colors from the inclusions of other minerals, e.g. asbestos (*cat's eye*), krokidolit (*tiger's eye*), actinolite (*prase*), green micas (*aventurine quartz*) or rutile spine (*rutilated quartz – hair of Venus*). Cryptocrystalline varieties of quartz belonging to the chalcedon group take many

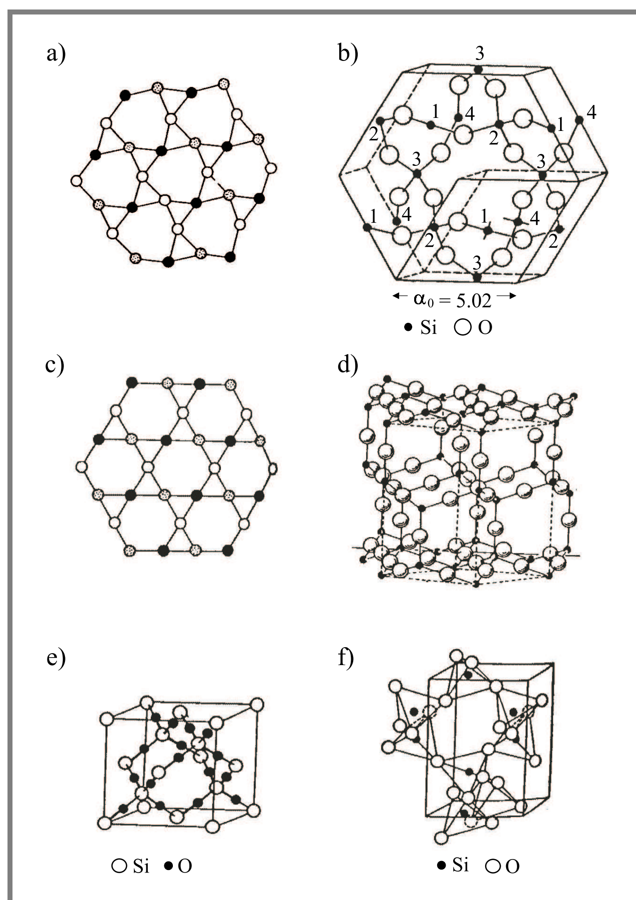


Fig. 4. Fundamental structures of polymorphic SiO_2 varieties: (a),(b) quartz β; (c) quartz α; (d) trydymite α; (e) cristobalite α; (f) cristobalite β.

different forms, such as multicolored and layered agate (Fig. 5) and onyx, jasper with diversified colors, green *chryzopras*, red *carnelian* and *sardonyx*, and finally common *flint*. In the world of minerals amorphous varieties of pure silica are found, too, mostly in the form of opals, and significantly less often as a silica glass formed as a result of melting of quartz sand during atmospheric discharge (the so-called *fulgurite*). Finally, let us mention very specific mineral forms of silica clusters – *geyserites*, deposited around certain hot, volcanic sources and geysers.

Fabrication of high-purity silicon (above 99.997 per cent of Si) for electronics requires, of course, the use of complex and time-consuming technologies of extensive processing of raw materials, that is production of second generation minerals. Regardless of the intricacy of the processing or synthesis, natural minerals are always the basis. Vein deposits of quartz, quartzite or quartz sands are the fundamental and most often used source of metallic silicon. It should be noted that the content of silica in the sands and quartz sandstone may sometimes reach a peak value of 99 per cent due to the processes of erosion, water transport and sedimentation.

Technologically pure, metallic silicon is usually obtained by means of electro-thermal reduction of molten quartz or quartzite with carbon. It is used as a component of alloys and is also the basis for the production of silicon carbide (SiC), silicides and silicones. Silicon with semiconductor properties, as well as single crystals of silicon (doped or not), are obtained from polycrystalline silicon.

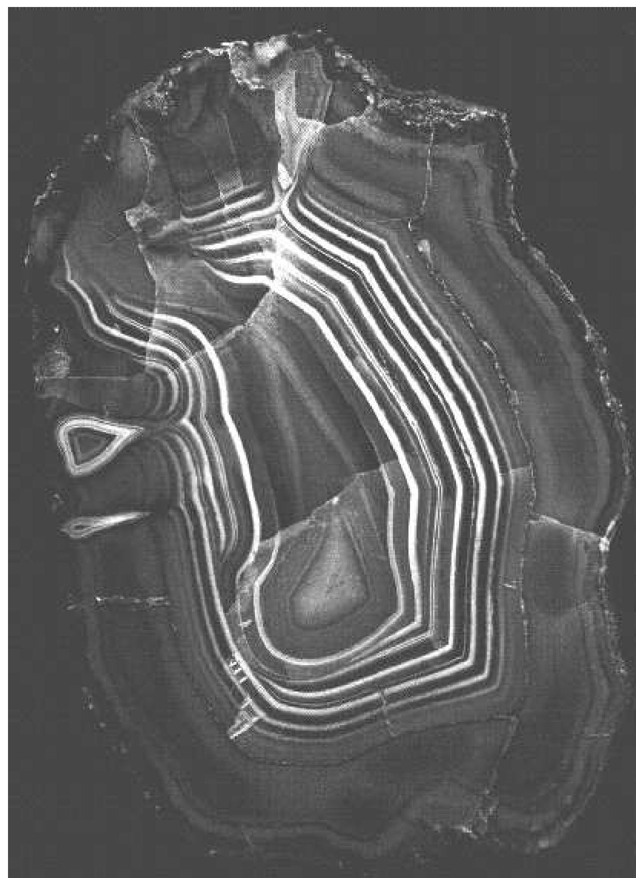
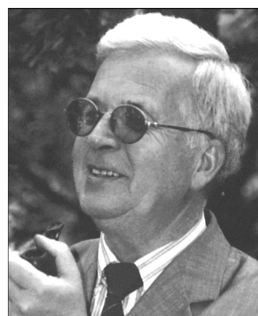


Fig. 5. Agate.

World resources of raw materials needed for the production of metallic silicon (quartz, quartzite, quartzite sands and sandstone) are enormous and practically unlimited. The only problem is high consumption of energy during the fabrication of pure silicon, which is important from the economic point of view.



Krzysztof J. Jakubowski was born in 1937. He graduated from Warsaw University in 1960. In 1967 he received the Ph.D. degree from the Faculty of Geology, Warsaw University. In 1974 he joined Polish Academy of Sciences and became the Director of the Museum of Earth. His research

interests include dynamic geology (gravitational shifts of rock masses), conservation of the nature and its resources, history of Earth-related sciences and museology, as well as popularization of science. He is the author and co-author of more than 200 publications. In 2001 he received a Hugo Steinhaus Award for science popularization. Since 2002 he is the vice-Chairman of Polish National Committee of ICOM (International Council of Museums).

e-mail: mzpaleo@warman.com.pl
Museum of Earth
Polish Academy of Sciences
Al. na Skarpie 20/26
00-488 Warsaw, Poland

Invited paper

Silicon microelectronics: where we have come from and where we are heading

Lidia Łukasiak, Andrzej Jakubowski, and Zbigniew Pióro

Abstract — The paper briefly presents the history of microelectronics and the limitations of its further progress, as well as possible solutions. The discussion includes the consequences of the reduction of gate-stack capacitance and difficulties associated with supply-voltage scaling, minimization of parasitic resistance, increased channel doping and small size. Novel device architectures (e.g. SON, double-gate transistor) and the advantages of silicon-germanium are considered, too.

Keywords — MOSFET, scaling, SiGe, SOI.

1. Introduction

The silicon semiconductor industry has been a strategic part of the worldwide economy for quite some time. In the past 25 years integration, expressed as the number of transistors in an LSI (VLSI, ULSI) chip, increased approximately by a factor of 10^4 and the computational capability of an LSI chip, understood as the MIPS value of a microprocessor, increased by a factor of about 10^3. This process has had a significant impact on system performance (speed of operation and power consumption) and manufacturing costs. On the other hand, it opened new markets, where performance is of crucial importance (portable equipment, automotive industry, communications, etc.). It is being anticipated that the continuous progress in microelectronics will bring about wide-ranging social and cultural changes in the future.

The paper presents briefly the history of microelectronics development and then proceeds to discuss its limitations and possible solutions.

2. History

It can be reasonably argued that solid-state electronics began in 1945, when Bell labs established a group, lead by William Shockley, whose task was to develop a semiconductor equivalent of a vacuum tube. In 1947 John Bardeen and Walter Brattain created an amplifying circuit based on a "transfer resistance" device with point contacts. The name of the device evolved finally into a transistor. A year later William Shockley presented a revolutionary concept of a junction transistor, which had several advantages over its point-contact predecessor. In 1952 a junction field-effect transistor appeared. In 1958 Jack Kilby from Texas Instruments built a simple oscillator circuit with five integrated components. In 1959 Robert Noyce of Fairchild created a crude predecessor of today's integrated circuits using planar technology for the first time. Then, in 1960 MOSFET [1] reached a practical stage, 30 years after its concept was first presented [2]. This set the stage for the rapid development of microelectronics in the years to come.

Another important achievement was Claude Shannon's theory of communication [3]. The theory included the most fundamental issues concerning sending and processing information, that is the essence of telecommunications [4]. Among many other important aspects of the theory, Shannon defined a practical measure of information, that is a bit (binary digit). This concept was closely related to using the binary system for coding the information. Shannon's work included also the improvement of the reliability of communication by means of a certain redundancy of the information being sent over a channel.

Solid-state electronics and Shannon's theory enabled the progress in the area of data processing and communication. Before the birth of a microprocessor the functions that an integrated circuit could perform had to be implemented in the hardware, that is silicon. A microprocessor, on the other hand, could take and execute instructions that defined the function it was performing in a given application. This approach efficiently cut costs and increased enormously the versatility of the chip. The first microprocessor, 4004, was released in 1971, by Intel. The chip counted 2300 transistors and the die size was approximately 13.5 mm^2. Microprocessor 4004 was capable of performing 60000 operations in one second. Its main disadvantage was the width of its data bus – only 4 bits, which was not enough to even code the alphabet. Thus in 1972 the 8-bit version, 8008, was introduced. The chip contained about 3500 transistors. Apart from a wider data bus it could also address much more memory and had a 3–4 times higher processing power. This trend continued ever since. The subsequent microprocessors contained ever more transistors, could address more and more memory locations, their data buses were wider and they were getting faster and faster. The tendency of increasing the functionality of microprocessors is illustrated in Fig. 1, where the number of transistors in subsequent Intel processors is given.

The curve presented in Fig. 1 is a reflection of Moore's law [5]. Originally, Moore noticed that the number of components in an IC corresponding to the minimum cost per component was doubling every year.

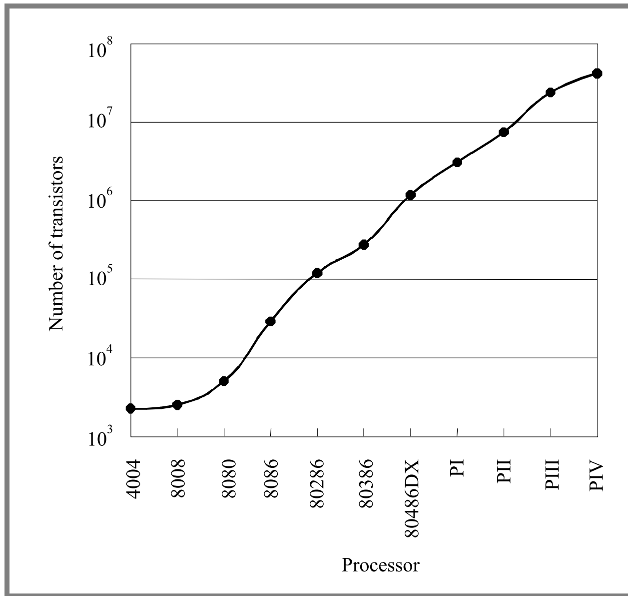


Fig. 1. Number of transistors in subsequent Intel processors (data after www.intel.com).

Later, the law had been reformulated to state that the transistor count on an IC chip doubles every 18 months. This tendency results not only from miniaturization, but also from the increase of the chip area and from improvements in the architecture, as illustrated in Fig. 2.

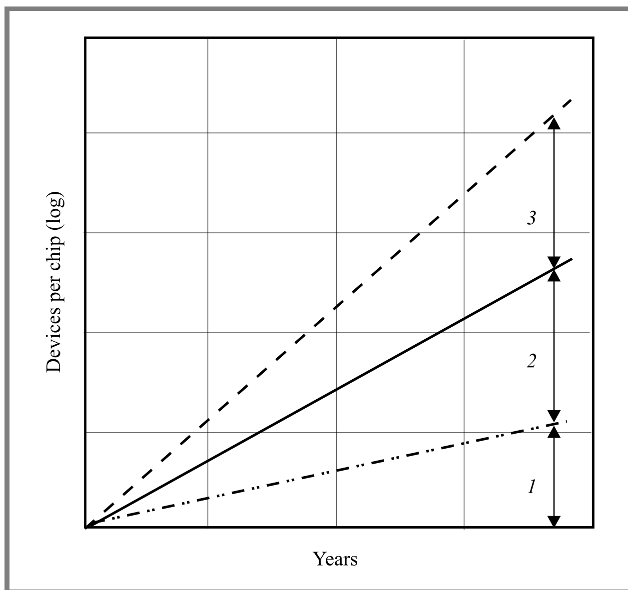


Fig. 2. Moore's law. Explanations: *1* – chip-area increase; *2* – line-width reduction; *3* – improvement of circuits and individual devices.

Moore's law has nothing to do with physics. Its persistence is rather caused by the fact that increased functionality of chips enabled by elevated transistor count causes new applications to appear. These applications, in turn, push for further progress in microelectronics. Thus a positive feedback has been created that made Moore's law a self-fulfilling prediction.

It has been argued in [6] that a shift is taking place from the world where assets were of value to the world where the value is in owning the information about assets. The boom in applications mentioned above is both the reason and the result of this shift. Moore's law is not the only one to govern the development of information technology. Robert Metcalfe, the inventor of Ethernet, predicted that the value of a network is proportional to the square of the number of nodes. This is known as Metcalfe's law [7] (Fig. 3).

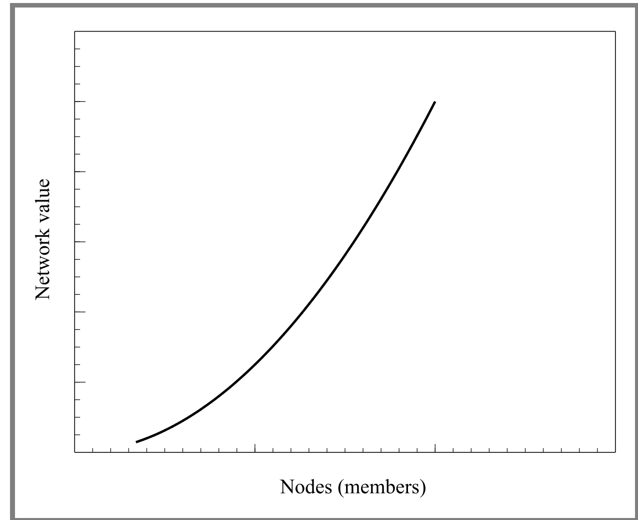


Fig. 3. Illustration of Metcalfe's law.

Another prediction, known as Gilder's law or bandwidth law, says that bandwidth grows at least three times faster than computer power [7] (Fig. 4).

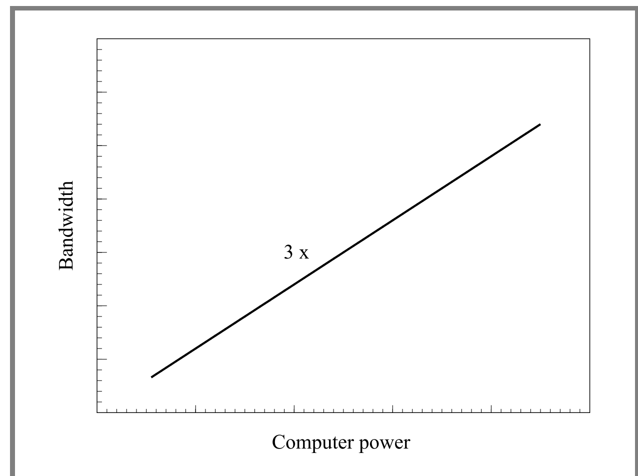


Fig. 4. Illustration of Gilder's law.

These three laws work in synergy to accelerate the development of information technology. As it was mentioned

earlier, a certain positive feedback exists between the needs created by information technology and the means to satisfy those needs, provided by microelectronics. The big question is whether the progress can be maintained forever.

3. Limits to miniaturization and possible solutions

The progress of silicon microelectronics, so far, has been driven mostly by miniaturization. The most obvious result of this process is the reduction of the dimensions of individual semiconductor devices, which, together with the improvement in the fabrication process, allows ever more devices to be crammed into one chip. The feature size F can be expressed as a function of time by the following formula:

$$F \approx \sim (7-8) \exp\left[-0.13 \cdot (\text{Year} - 1971)\right] \; [\mu\text{m}]. \quad (1)$$

This trend is illustrated in Fig. 5 (after [8, 9]).

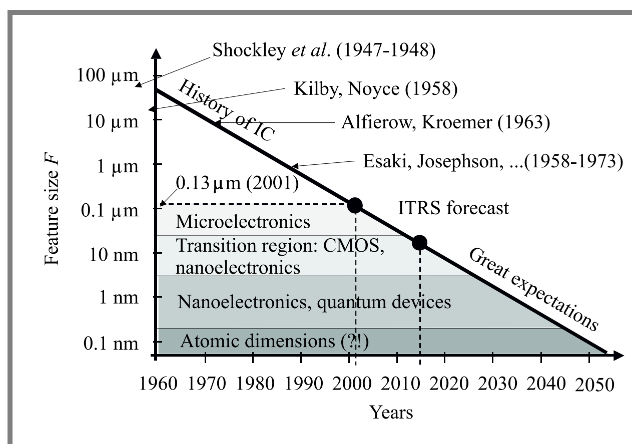


Fig. 5. Changes of feature size with time (the names and dates correspond to Nobel prizes crucial for the development of microelectronics).

The reduction of feature size, of course, increases chip functionality leading finally to a system on a chip. Reduced dimensions have yet another advantage of increased speed of operation. This is due to the shortening of the physical path the carriers have to pass and also to the reduction of the capacitances. The question is, however, how long this trend can be continued.
The main requirements for a healthy MOSFET are: high I_{ON}/I_{OFF} ratio, short-channel effects kept at a reasonable level and finite subthreshold slope. The ways to achieve this goal and the associated dangers are discussed below.

3.1. Reduction of gate-stack capacitance

The value of I_{ON} can be boosted by increasing the gate-stack capacitance. This is achieved mainly by the reduction of the gate-oxide thickness, but also by means of decreasing the gate-electrode capacitance.

Reduction of gate-oxide thickness. This step has another advantage of minimizing short-channel effects as well. The gate-oxide thickness (expressed in nanometers) has changed over the years according to the following formula (illustrated in Fig. 6):

$$t_{SiO_2} = 120 \cdot \exp\left[-0.12 \cdot (\text{Year} - 1970)\right]. \quad (2)$$

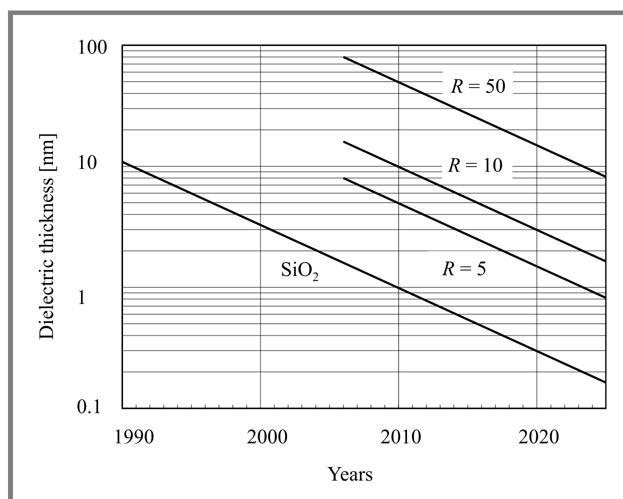


Fig. 6. Gate-oxide (SiO_2) thickness over the years. Thickness of hypothetical dielectric materials with the dielectric constant R times higher than that of SiO_2 is shown for comparison.

From the device point of view, the biggest concern associated with thin gate-oxides is the growing leakage current. Every 0.2-nm reduction between 2 and 1 nm implies a 10-fold increase of the tunneling current [10].
It has been demonstrated that even MOSFETs with the gate-oxide thickness as thin as 1.5 nm [10, 11], and even below 1 nm, can operate appropriately, on condition the channel is short. It seems, however, that increasing power consumption remains a problem, especially in terms of the entire chip rather than in terms of a single transistor. On the other hand, it is being argued [12] that with static power management the strongest limitation to the reduction of gate-oxide thickness will come from reliability, not from power-consumption concerns. While the mechanisms of thin-oxide breakdown are not understood well [13–16], it is clear that the electrical strength of a dielectric is aggravated by the roughness of the interface, especially in the case of thin layers, where a local thinning may be really dangerous. Finally, thin gate-oxides encourage the diffusion of the gate-electrode dopant into the substrate. All the above constraints indicate that there is a certain minimum gate-oxide thickness past which we cannot go.
To be sure, the operation of the final device is not the only source of the limitations of the gate-oxide thickness reduction. The formation of very thin gate-oxide layers is rather difficult [17] mainly due to extremely high reactivity of silicon surface, which cannot be maintained in the state of chemical purity for a sufficiently long time. Sufficiently

long time is understood here as that needed for loading the silicon wafers into the boat and then loading the boat into the furnace processing zone.

The other problem to be solved in the area of ultra-thin oxide formation is oxidation process controllability. In the standard, batch-type furnaces oxide growth at normally used temperatures is too fast for ultra-thin layers. As a result a compromise has to be reached between process control (and reproducibility) and oxide quality.

The answer to the problems with ultrathin SiO_2 might be the use of high-k dielectrics. Their obvious advantage is that, when compared to SiO_2, they allow for the same capacitance at much higher thickness. If the rate of gate--dielectric thinning were to be maintained, then the thickness of a hypothetical material with the dielectric constant R times higher than that of SiO_2 could be expressed with the following formula (in nanometers):

$$t_{diel} = 1.6 \cdot R \cdot \exp\left[-0.12 \cdot (\text{Year} - 2006)\right]. \qquad (3)$$

Curves corresponding to the thickness of dielectric materials with $R = 5, 10, 50$ are shown in Fig. 6. Unfortunately, high-k dielectrics have to fulfill a number of requirements in order to be useful. These are [e.g. 18]: high thermal stability, perfect stoichiometry (to minimize the number of defects) low concentration of interface states and stability of the interface during thermal treatments and external radiation, resistance to dopant diffusion, sufficiently wide bandgap, sufficient barrier height for both electrons and holes. A number of high-k dielectric have been tested [e.g. 19], but it is not clear which is going to replace SiO_2 as the dominant gate-oxide material.

Reduction of gate capacitance. The other component of gate capacitance is the gate electrode made of doped polysilicon, which has the advantage of allowing for the work function adjustment by means of doping. Increased doping concentrations are also used to reduce its resistance and minimize the depletion effects. The gate capacitance degradation due to the depletion of polysilicon accounts for 0.4–0.5 nm of the equivalent oxide thickness of the total gate capacitance at inversion according to [20]. This disadvantage could be removed if polysilicon were replaced with metal. The difficulty here is, however, that the workfunction of a given metal is fixed. If a metal with midgap workfunction is used for both NMOS and PMOS, the threshold voltage will be increased by about half the bandgap [8]. To prevent this, two different metals would have to be used, with appropriate workfunctions. This would, however, complicate the fabrication process and increase costs. To make things even more ambiguous, it is being argued in [12] that poly-depletion alone, while undesirable, is not harmful enough to be the reason for replacing poly-gates with metallic ones. On the other hand, [12] proposes that poly-SiGe gate with low Ge moll fraction be used instead of poly-Si, because it allows for up to 2× reduction in poly-depletion.

3.2. Supply voltage

Another possibility of increasing the I_{ON}/I_{OFF} ratio is to increase the ratio of the supply voltage to the threshold voltage (V_{DD}/V_{TH}). Unfortunately, the V_{DD}/V_{TH} ratio is decreasing with time. It used to be around 10 for older technologies but drops to merely 4 for 0.18 μm and to less than 2 for 0.07 μm low-performance CMOS [12]. This is mostly due to the fact that oxide reliability requires appropriate reduction of the supply voltage. On the other hand, the reduction of threshold voltage is not easy because the built-in voltages inherent in the semiconductor structures are not well scalable. It is being predicted that the decrease of V_{TH} below approximately 0.25 V would result in a considerable increase of the subthreshold current [19]. Moreover, it has been demonstrated that scaling according to ITRS will lead to insufficient I_{ON} for CMOS generations of 70 nm and below [12]. Thus it is being anticipated that V_{DD} scaling will be slower than previously expected.

3.3. Minimizing the resistance

Still another way to increase I_{ON} is to minimize the parasitic series resistance, coming mainly from source and drain. The structure of a source/drain is schematically shown in Fig. 7, together with the main components of the resistance.

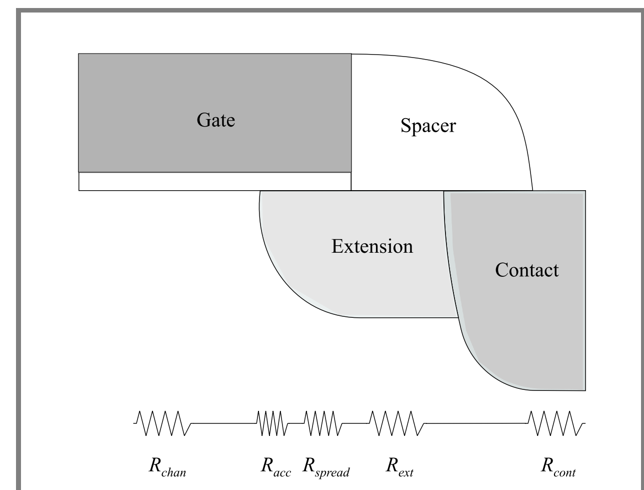


Fig. 7. Typical source/drain structure with main resistance components.

The reduction of junction depth is one of the more effective ways of controlling short-channel effects [19, 21]. This, however, increases the series resistance of a MOSFET, which can only be minimized by means of increased doping, obviously limited by maximum solid solubility. Moreover, high dopant concentrations required to minimize the resistance increase the dopant diffusivity, thus increasing the difficulties in the formation of shallow junctions resulting from thermally enhanced diffusion associated with post-implantation annealing [8]. Metastable dopant concentrations, exceeding the maximum solid solubility, could

solve the problem on condition that device processing is modified so as to minimize thermal cycles capable of dopant deactivation or that methods are found to suppress the formation of point defects facilitating deactivation [8]. A more radical solution would be to propose a different device structure.

The source and drain contacts are another source of unwanted resistance. The silicidation process, one that is used the most often to form these contacts, consumes the most highly doped portion of silicon thus further aggravating the resistance problem. This is because the resistivity of an ohmic connection between a silicide and silicon depends strongly on the doping level at the semiconductor surface. These difficulties could be potentially solved if the silicide was formed before dopant implantation or if low-temperature epitaxy could be used for silicide formation (e.g. in the case of $CoSi_2$ or $NiSi_2$) [8].

3.4. Increased channel doping

A sufficient I_{ON}/I_{OFF} ratio requires the OFF current to be kept at reasonable level. The control of its value can be performed by means of increased channel doping [e.g. 12]. When pushed too far, however, this may cause the source-substrate and drain-substrate junctions to act as tunneling diodes [19], with obvious consequences for device operation, not to mention mobility degradation.

3.5. Consequences of small size

The most commonly used transport models correspond to the situation where the dimensions of the devices are far greater than the mean free path. If these conditions are not fulfilled, other types of transport, usually referred to as ballistic, have to be considered [22].

On the other hand, continuous miniaturization decreases the number of atoms that an individual device consists of. The number of dopant atoms is of particular importance here, because these atoms are distributed rather randomly. Thus unintended doping non-uniformity may become a problem both in terms of a single device and in terms of unwanted differences between devices in the same integrated circuit [22].

3.6. Changing the device architecture

In view of the difficulties that seem to impede further scaling of conventional CMOS, novel device architectures have been proposed. Some of them are discussed below.

Silicon-on-insulator (SOI). A schematic cross-section of an SOI CMOS inverter is shown in Fig. 8.

The difference between conventional bulk MOSFETs and their SOI counterparts lies in the fact that the active region of a SOI MOSFET is a thin, monocrystalline silicon film separated from the rest of the substrate by a layer of buried SiO_2 or BOX. Such a design has a number of advantages. First of all due to a limited thickness of the active region the area of the source and drain junctions is much smaller, which means that the capacitance is considerably lower, therefore speed is higher. Smaller junction areas mean also that the leakage currents are lower.

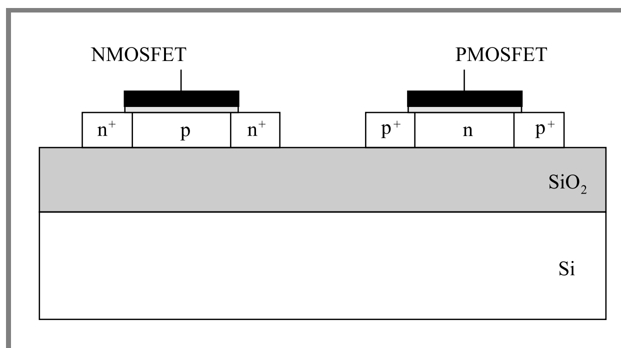


Fig. 8. A cross-section of a SOI CMOS inverter.

The SOI technology facilitates the isolation of individual devices from the rest of the circuit and thus has a considerable potential for high packaging density [23]. Another advantage is that the current driveability increases with decreasing active film, which is illustrated in Fig. 9.

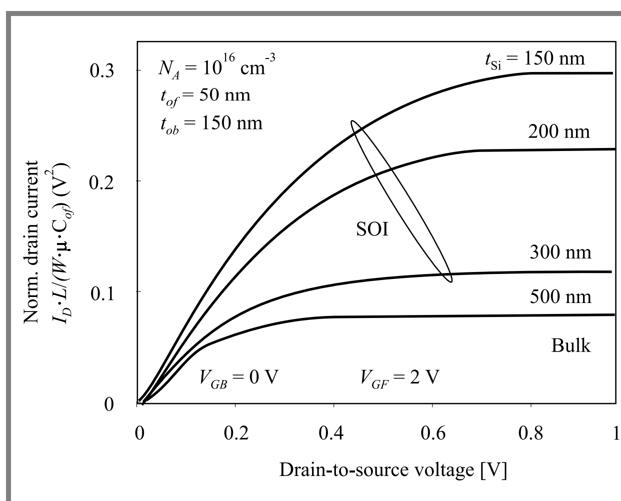


Fig. 9. Output characteristics of SOI and bulk MOSFETs [24].

While the parameters of SOI MOSFETs simulated in Fig. 8 correspond to the very beginning of SOI history, the advantages resulting from thin active region are clearly visible.

Due to the reduction of the vertical dimensions fully-depleted (FD) SOI was shown to suppress short channel effects [23]. It has been demonstrated, however, in [24] that in short-channel ($L < 0.1$ μm) FD SOI MOSFETs an electrostatic coupling between source and drain taking place via the channel and the BOX relaxes considerably the control over short-channel effects (especially the subthreshold slope).

Reducing the thickness of BOX down to 10 nm together with the reduction of the channel thickness to a similar value results in the suppression of this coupling, even for

devices with very short channels [25]. It is, however, difficult to fabricate such thin silicon and BOX layers by means of SIMOX or wafer bonding. Thus a new technology has been proposed, called silicon-on-nothing (SON). SON is aimed at fabrication of localized SOI areas under the gates of transistors within the bulk CMOS flow [26]. A schematical cross section of a SON transistor is shown in Fig. 10.

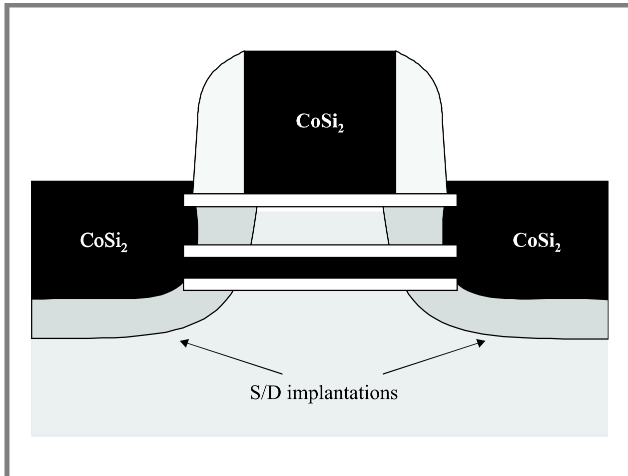


Fig. 10. A schematic cross-section of a SON MOSFET (after [25]).

This process makes it possible to form very thin layers of silicon channel and BOX (with the thickness controlled by epitaxy). Another advantage of this solution is that the buried dielectric does not reach the highly doped regions of source and drain. Therefore, the depth of the extensions is controlled by the channel thickness (which helps suppress short-channel effects), while the highly doped regions can be sufficiently deep to reduce the series resistance [25].

Double-gate MOSFET. The advantages of the double-gate transistor lie in the fact that short-channel effects are controlled by device geometry instead of doping as is the case with bulk devices. A schematic view of the double-gate transistor is shown in Fig. 11.

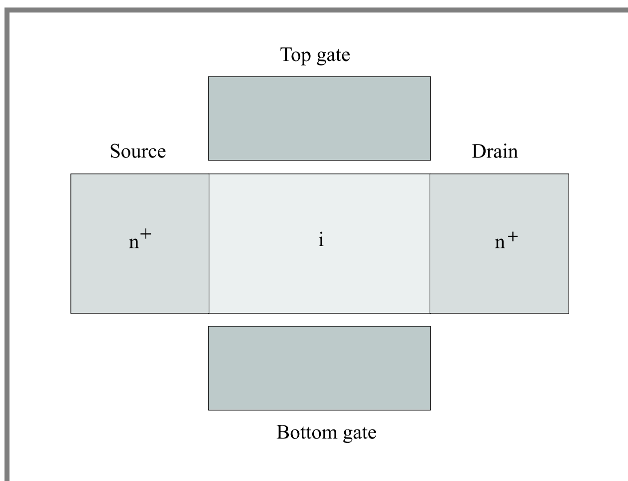


Fig. 11. A schematic view of a double-gate transistor.

The junction depth is controlled by the channel thickness. As a result, the control of short-channel effects is tighter in a DG MOSFET allowing aggressive reduction of the channel length. Thanks to the reduction of the channel doping the carrier transport is improved and the tunneling current flowing between drain and body is suppressed [20]. It is being estimated that the DG MOSFET can be scaled up to 50% further than the bulk MOSFET for some applications [27].

3.7. Silicon-germanium – increasing speed

Silicon and germanium both have the structure of a diamond. Their lattice constants are similar, there is, however, a certain lattice misfit between the two, which amounts to 4.17% at room temperature. This imposes certain constraints on the growth of SiGe layers. An important feature of SiGe is the fact that the width of its bandgap depends on the amount of germanium added (the bandgap is reduced about 7.5 mV per percent Ge). Grading Ge contents from 0 to 15% over a distance of 50–60 nm one obtains built-in electric fields of 15–20 kV/cm. Such electric fields may easily accelerate charge carriers to the saturation velocity.

This makes silicon-germanium an ideal material for the base of a bipolar transistor. The fact that the SiGe bandgap is lower than that of Si is very favourable in the case of Si emitter and SiGe base, because it enhances the injection process. Therefore the current gain of a silicon-SiGe heterojunction bipolar transistor (HBT) may be 100–1000 times higher than that of silicon BJTs (Hitachi reached $h_{FE\,\max} = 29000$). This is the case with the so-called "real" HBTs, where the germanium contents are high and constant throughout the base. Since such a high gain is not needed in practice, part of it may be sacrificed in order to allow the doping concentration in the base to be higher. This has a beneficial effect of lowering base resistance meaning higher maximum frequency of oscillations. This trade-off was impossible in conventional BJTs. Moreover, since the doping concentration of the Si base was low, the base itself could not be too thin, because of punch-through. In contrast, the thickness of a SiGe base could be scaled down, which was beneficial for the cut-off frequency. Moreover, if the germanium content is graded throughout the base from a low value at the emitter side to a high one at the collector side the transport of carriers through the base is additionally assisted by a built-in electric field, which considerably reduces transit time, leading to higher cut-off frequencies. One of the highest f_T reported so far is 350 GHz (with $f_{\max}$ of 170 GHz) and in an optimized design $f_{\max}$ of 285 GHz and f_T of 270 GHz were achieved [28]. It should be noted that in conventional BJT technology the cut-off frequency was growing at a rate of barely 4% per year. The introduction of SiGe boosted that growth to no less than 30% per year. The achieved

maximum frequencies are comparable to those obtained in HBTs based on $A_{III}B_V$ compounds and fabrication costs are significantly lower. It is being expected that SiGe will also increase the speed of operation of MOSFETs.

4. Summary

Despite increasing difficulties, the silicon technology is steadily progressing towards better functionality and higher speed of operation. Even if conventional technology reaches its limits, there will still be some room to maneuver in the form of e.g. new device architecture, especially in view of the fact that a normally operating ultra-thin channel PMOSFET with gate length of 6 nm has already been reported [29]. It should be remembered, however, that miniaturization, apart from simple reduction of size, brings about also significant qualitative changes. Some of those are illustrated in Table 1. The reduction of the number of dopants in the active region poses obvious problems for those involved in device fabrication. Those involved in device modeling will have to find a way to describe the operation of such devices. Finally, the incredible speed they offer will surely have huge impact on the design of integrated circuits, since the delay problem will increasingly shift from devices to interconnects.

Table 1
MOSFET evolution

Parameter	Past	Present	Future
Channel length	1 μm	0.1 μm	0.01 μm
Number of dopant atoms in the active region	$\sim 10^6$	$\sim 10^6$	e.g. 3
Number of electrons participating in the switching process	$\sim 10^7$	$\sim 10^4$	e.g. 30
Cut-off frequency	1 GHz	100 GHz	$>$ 1 THz

On the other hand, it may be argued that microelectronics is slowly reaching the stage where the value moves from the technology itself to its application. In such circumstances, it is rather difficult to predict the future, and maybe the best approach towards it is that of Albert Einstein, who once said: "I never worry about the future. It comes soon enough".

References

[1] D. Kahng and M. M. Atalla, "Silicon-silicon dioxide field induced surface devices", in *IREE Solid-State Dev. Res. Conf.*, Pittsburgh, USA, 1960.

[2] J. E. Lilienfeld, "Method and apparatus for controlling electric currents", US patent, no. 1 745 175, 1930.

[3] C. Shannon, "A mathematical theory of communication", *Bell Syst. Techn. J.*, vol. 27, July 1948.

[4] J. Szabatin, "Era informacyjna a teoria Shannona", *Przegląd Telekomunikacyjny*, no. 4, pp. 278–287, 2000 (in Polish).

[5] G. E. Moore, "Cramming more components onto integrated circuits", *Electronics*, vol. 38, 1965.

[6] N. Negroponte, *Being Digital*. Alfred Knopf, 1995.

[7] G. Gilder, *Telecosm: How Infinite Bandwidth will Revolutionize the World*. Free Press, 2000.

[8] J. D. Plummer and P. B. Griffin, "Material and process limits in silicon VLSI technology", *Proc. IEEE*, vol. 89, pp. 240–257, 2001.

[9] A. Jakubowski and L. Łukasiak, "O telekomunikacyjnych pożytkach z elektroniki wynikających", *Przegląd Telekomunikacyjny*, no. 1, pp. 5–11, 2003 (in Polish).

[10] C. T. Liu, "Circuit requirement and integration challenges of thin gate dielectrics for ultra small MOSFETs", *IEDM Tech. Dig.*, pp. 747–750, 1998.

[11] H. S. Momose *et al.*, "Tunneling gate oxide approach to ultra-high current drive in small-geometry MOSFETs", *IEDM Tech. Dig.*, pp. 593–596, 1994.

[12] T. Skotnicki, "Heading for decananometer CMOS – is navigation among icebergs still a viable strategy?", in *30th Eur. Solid-State Dev. Res. Conf.*, Cork, Ireland, 2000.

[13] H. Iwai and H. S. Momose, "Ultra-thin gate oxides – performance and reliability", *IEDM Tech. Dig.*, pp. 163–166, 1998.

[14] B. E. Weir *et al.*, "Ultra-thin gate dielectric: they break down, but do they fail?", *IEDM Tech. Dig.*, pp. 73–76, 1997.

[15] J. H. Stathis and D. J. DiMaria, "Reliability projection for ultra-thin oxides at low voltage", *IEDM Tech. Dig.*, pp. 167–170, 1998.

[16] G. Groeseneken *et al.*, "Reliability of ultra-thin oxides for the giga-bit generations", in *29th Eur. Solid-State Dev. Res. Conf.*, Leuven, Belgium, 1999, p. 72.

[17] A. M. Stoneham and C. J. Sofield, "Modeling the oxide and the oxidation process: can silicon oxidation be solved?", in *Fundamental Aspects of Ultrathin Dielectrics on Si-based Devices*, E. Garfunkel *et al.*, Eds. Kluwer, 1998.

[18] L. C. Feldman *et al.*, "Ultrathin dielectrics in silicon microelectronics – an overview", in *Fundamental Aspects of Ultrathin Dielectrics on Si-based Devices*, E. Garfunkel *et al.*, Eds. Kluwer, 1998.

[19] H. Iwai and S. Ohmi, "CMOS downsizing and high-*k* gate insulator technology", in *4th IEEE Int. Caracas Conf. Dev., Circ. Syst. ICCDCS'2002*, pp. D049-1–D049-8.

[20] H.-S. P. Wong, "Beyond the conventional transistor", *IBM J. Res. Dev.*, vol. 46, no. 2/3, 2002.

[21] J. R. Brews, W. Fichtner, E. Nicollian, and S. M. Sze, "Generalized guide for MOSFET miniaturization", *IEEE Electron. Dev. Lett.*, vol. EDL-1, pp. 2–4, 1980.

[22] R. W. Keyes, "Fundamental limits of silicon technology", *IEEE Proc.*, vol. 89, pp. 227–239, 2001.

[23] J.-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI*, Kluwer, 1995.

[24] A. Jakubowski, M. Jurczak, and L. Łukasiak, "Krzem na izolatorze – przyrządy", *Elektronika*, vol. 37, no. 4, pp. 15–20, 1996 (in Polish).

[25] T. Skotnicki, S. Monfray, and C. Fenouillet-Beranger, "Emerging silicon-on-nothing (SON) devices technology", in *Proc. Int. Symp. SOI Technol. Dev. XI, Electrochem. Soc. Proc.*, vol. 2003–05, pp. 133–148, 2003.

[26] S. Monfray and T. Skotnicki *et al.*, "First 80 nm SON (silicon-on-nothing) MOSFETs with perfect morphology and high electrical performance", *IEDM Tech. Dig.*, pp. 645–648, 2001.

[27] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device scaling limits of Si MOSFETs and their application dependencies", *IEEE Proc.*, vol. 89, pp. 259–288, 2001.

[28] J.-S. Rieh *et al.*, "SiGe HBTs with cut-off frequency of 350 GHz", *IEDM Tech. Dig.*, pp. 771–774, 2002.

[29] B. Doris *et al.*, "Extreme scaling with ultra-thin SOI channel MOSFETs", *IEDM Tech. Dig.*, pp. 267–270, 2002.

Lidia Łukasiak graduated from the Faculty of Electronics (Warsaw University of Technology) in 1988 and joined the Institute of Microelectronics and Optoelectronics the same year. She received the Ph.D. and D.Sc. degrees from Warsaw University of Technology in 1988, 1994 and 2002. In 1995 she received Prime Minister's Award for a distinguished Ph.D. thesis. Her research interests include modeling and characterization of semiconductor devices, as well as microprocessor techniques. She was a member of the Technical Programme Committee of 4th International Caracas Conference on Devices, Circuits and Systems, ICCDCS'2002.
e-mail: lukasiak@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Andrzej Jakubowski received the M.Sc., Ph.D. and D.Sc. degrees from the Faculty of Electronics, Warsaw University of Technology. At present he is the head of the Division of Microelectronics and Optoelectronics Devices in the Institute of Microelectronics and Optoelectronics (WUT). In the years 1989–1992 he held the position of the Director of the Institute of Electron Technology in Warsaw. From 1989 to 1991 he was the head of the Government Program for Polish Microelectronics. In the years 1990–1991 he headed Polish Government Committee on Applied Sciences. The main area of his research interests is modeling, characterization and technology of semiconductor structures and integrated circuits. He is also the author and co-authorof approximately 400 technical papers.
e-mail: jakubowski@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Zbigniew Pióro was born in Nadkole, Poland, in 1948. He received the M.Sc. and Ph.D. degrees from Warsaw University of Technology in 1971 and 1975, respectively, both in microelectronics. He joined the Institute of Microelectronics and Optoelectronics of Warsaw University of Technology in 1971. In the years 1994–2001 he held the position of R&D Director of AB-micro s.c., a company involved in automation of industrial processes. His research interests include metrology, digital signal processing and microprocessor techniques.
e-mail: pioro@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Invited paper

Critical modeling issues of SiGe semiconductor devices

Vassil Palankovski and Siegfried Selberherr

Abstract — We present the state-of-the-art in simulation of silicon-germanium (SiGe) semiconductor devices. The work includes a detailed comparison of device simulators and current transport models. Among the critical modeling issues addressed in the paper, special attention is focused on the description of the anisotropic majority/minority electron mobility in strained SiGe grown on Si. We use a direct approach to obtain scattering parameters (*S*-parameters) and other derived figures of merit of SiGe heterojunction bipolar transistors (HBTs) by means of small-signal AC-analysis. Results from two-dimensional hydrodynamic simulations of SiGe HBTs are presented in good agreement with measured data. The examples are chosen to demonstrate technologically important issues which can be addressed and solved by device simulation.

Keywords — SiGe HBT, numerical simulation, modeling, bandgap, mobility, small-signal simulation, S-parameters.

1. Introduction

SiGe HBTs progressively replace III-V devices for their typical applications, such as low noise amplifiers and frequency dividers up to 99 GHz [1], and are considered essential for 40 Gbit/s optical communication systems. Transit frequencies, f_T, of 350 GHz [2], maximum oscillation frequencies, $f_{\max}$, of 285 GHz, and ring oscillator delays of 4.2 ps [3] have been reported. Figure 1 shows the rapid progress of peak-f_T of SiGe HBTs over the last couple of years. The devices are fully compatible with the existing state-of-the-art 0.13 μm CMOS technology [3, 4]. Digital application-specific integrated circuits (ASICs) are combined with SiGe HBT circuits in the so-called SiGe BiCMOS technology and are in volume production.

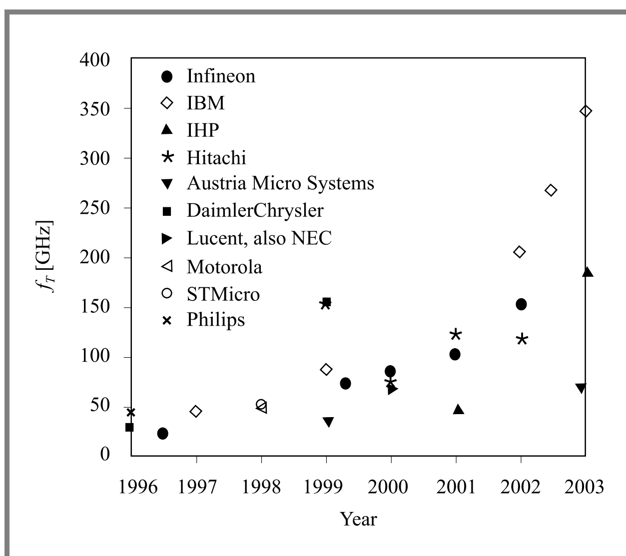


Fig. 1. Current gain cutoff frequency f_T of SiGe HBTs over time.

With the shrinking of device dimensions and replacement of hybrid mounted transistors by MMICs, rigorous physical device simulation and circuit simulation with distributed devices has to be carried out by simulation tools which account for physical effects on a microscopic level. Optimization of geometry, doping, materials, and material composition is targeting high power, high breakdown, high speed (high f_T, $f_{\max}$), low leakage (low power consumption), low noise, etc. This is a challenging task that requires significant efforts in device modeling.

Section 2 gives a review of state-of-the-art device simulators and discusses the choice of current transport models to be used. In Section 3 critical modeling issues are addressed, such as bandgap narrowing, anisotropic electron minority mobility in strained SiGe, carrier transport through heterointerfaces, carrier generation/recombination, and lattice self-heating.

Section 4 presents numerical simulation results compared to the experimental data for SiGe HBTs. The examples are chosen to demonstrate technologically important issues which can be addressed and solved by device simulation. In particular, examples were chosen, where physical effects are of importance for both the DC-, and the AC- device behavior, e.g. forward characteristics of SiGe HBTs with different Ge contents considering band gap narrowing and anisotropic mobility effects, output characteristics including self-heating and impact-ionization generation effects, and f_T vs. I_C plots accounting for hot-carrier effects and anisotropic transport. All obtained results are in good agreement with the measured data.

2. Device simulators

The continuously increasing computational power of computer systems allows the use of technology computer aided design (TCAD) tools on a very large scale. Several commercial device simulators, e.g. [5–10], company-developed simulators, e.g. [11, 12], and university developed simulators, e.g. [13–19], claim the capability to handle SiGe devices. These simulators differ considerably in dimensionality (one-, /quasi-/two-, or /quasi-/three-dimensional), in the choice of carrier transport model (drift-diffusion, energy-

Table 1
Comparison of device simulators

Simulator	Dimension	Model	Features
NEMO	1D		Schrödinger-Poisson solver
BIPOLE3	Quasi-2D	DD	Polysilicon
ATLAS	2D	DD, ET	TE heterojunction model
APSYS	2D	HD	Optical, interfaces
Jungemann	2D	DD, HD, MC	Rigorous transport modeling
PISCES	2D	DD, ET	Polysilicon, harmonic balance
MEDICI	2D	DD, HD	Anisotropic properties
FIELDAY	2D, 3D	DD	Electrothermal
Minimos-NT	2D, 3D	DD, HD	(See Section 3)
DESSIS	2D, 3D	DD, HD	Trap modeling, TFE model
DD – drift-diffusion, ET – energy-transport, HD – hydrodynamic			

transport, or Monte Carlo statistical solution of the Boltzmann equation), and in the capability of including electrothermal effects. The drift-diffusion transport model [20] is by now the most popular model used for device simulation. With down-scaling feature sizes, non-local effects become more pronounced and must be accounted for by applying an energy-transport or hydrodynamic transport model [21]. During the last two decades Monte Carlo methods for solving the Boltzmann transport equation have been developed [22, 23] and applied for device simulation [24–26]. However, reduction of the demand on computational resources is still an issue and, therefore, Monte Carlo device simulation is still not feasible for industrial application on a daily basis. A way to preserve the accuracy at lower computational cost is to calibrate lower order transport models to Monte Carlo simulation data.

In addition, quantum mechanical effects are often neglected or accounted for only by simple models for quantum corrections [27, 28], as solving the Schrödinger or the Wigner equation is extremely expensive in terms of computational resources.

The limited feedback from technological state-of-the-art process development to simulator development is a common drawback. The quality of the physical models can be questioned as the model parameters for SiGe are often simply inherited from parameters for silicon. Critical issues concerning simulation of heterostructures are frequently not considered, such as interface modeling at heterojunctions and at silicon/polysilicon interfaces. Hydrodynamic and high field effects, such as carrier energy relaxation, impact ionization, and self-heating effects, are often ignored.

The two-dimensional device simulator PISCES [13], developed at Stanford University, incorporates modeling capabilities for SiGe based devices, e.g. for silicon/polysilicon interfaces. One of its versions, PISCES-HB, includes harmonic balance for large signal simulation.

The device simulator MEDICI from Synopsis [10], which is also based on PISCES, offers simulation features for SiGe/Si HBTs. Advantages of this simulator are hydrodynamic simulation capabilities and a rigorous approach to generation/recombination processes. In addition, it includes a module treating anisotropic material properties. This simulator has some weakness in the capability of mixed-mode device/circuit simulation.

At the quantum level, among others, a one-dimensional Schrödinger-Poisson solver NEMO [12], based on non-equilibrium Green's functions, is offered for sub-0.1 μm SiGe structures.

The two- and three-dimensional device simulator DESSIS from ISE [8] has demonstrated a rigorous approach to semiconductor physics modeling. Various critical issues, such as extensive trap modeling, are solved.

Quasi-two-dimensional approaches using a simplified one-dimensional current equation are demonstrated, among others, by BIPOLE3 from BIPSIM [7] which additionally features good models for polysilicon.

The two-dimensional Fast Blaze from Silvaco [6] has capabilities of simulating heterostructure devices. Simulations of SiGe HBTs were announced, based on a simulator originally developed at the University of Ilmenau, PROSA [18]. However, in the latter no material interfaces are considered. Several good optimization results for SiGe HBTs were achieved with another university developed simulator, SCORPIO [29].

Table 1 summarizes features of SiGe device simulators discussed in this paper.

3. Critical issues of modeling SiGe devices

This section discusses critical modeling issues for SiGe semiconductor devices. We have addressed these issues in our three-dimensional device simulator Minimos-NT [19], which can deal with different complex structures and materials, such as SiGe and various III-V binary and ternary compounds, with arbitrary material composition profiles in a wide temperature range.

The models are based on experimental or Monte Carlo simulation data and employ analytical functional forms which cover the whole material composition range. The model parameters are checked against several independent technolo-

gies to obtain a concise set used for all simulations. Reviewing simulation of HBTs and submicron heterojunction field-effect transistors with gate-lengths down to 100 nm, solutions of energy transport equations are necessary to account for non-local effects, such as velocity overshoot. A model for carrier temperature dependent energy relaxation times [30] has been developed as well as a model for lattice temperature dependent saturation velocities [31].

Heterointerface modeling is a key issue for devices which include abrupt junctions. Thermionic emission and field emission effects critically determine the current transport parallel and perpendicular to the heterointerfaces.

All important physical effects, such as bandgap narrowing, anisotropic electron minority mobility in strained SiGe, Shockley-Read-Hall recombination, surface and Auger recombination, and impact ionization are taken into account. III-V materials and SiGe are known to have a reduced heat conductivity in comparison to silicon [32]. Self-heating effects are accounted for by solving the lattice heat flow equation self-consistently with the energy transport equations. Examples are given in Section 4 for SiGe HBTs.

Advanced device simulation allows a precise physics-based extraction of small-signal parameters [33, 34]. Measured bias dependent *S*-parameters serve as a valuable source of information when compared at different bias points to simulated *S*-parameters from a device simulator, such as Minimos-NT. This procedure reflects the full RF-information contained in the *S*-parameters and allows process control beyond the comparison of DC-quantities.

3.1. Bandgap and bandgap narrowing

Modeling of strained SiGe is not a trivial task, since attention has to be focused on the stress-dependent change of the bandgap due to Ge content [35].

The temperature-dependent bandgaps of the constituents, E_g^{Si} and E_g^{Ge}, are calculated by the commonly used model of Varshni [36]

$$E_g = E_{g,0} - \frac{\alpha \cdot T_L^2}{\beta + T_L}, \tag{1}$$

where $E_{g,0}$ is the bandgap at T_L = 0 K. The parameter values are summarized in Table 2. The dependence on the material composition x is then introduced by

$$E_g^{\mathrm{SiGe}} = E_g^{\mathrm{Si}} \cdot (1-x) + E_g^{\mathrm{Ge}} \cdot x + C_g \cdot (1-x) \cdot x \tag{2}$$

with a bowing parameter $C_g = -0.4$ eV. This one-valley bandgap fit can be applied to the case of the technologically important strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ grown on Si (see Fig. 2).

Table 2
Parameter values for modeling the bandgap energy

Material	$E_{g,0}$ [eV]	α [eV/K]	β [K]
Si	1.1695	$4.73 \cdot 10^{-4}$	636
Ge	0.7437	$4.774 \cdot 10^{-4}$	235

Depending on the strain the bandgap can become smaller than the one of pure Ge [37] in certain cases. In the unstrained case, however, an X-to-L gap transition is observed at about $x = 0.85$, which has to be accounted by the model as well.

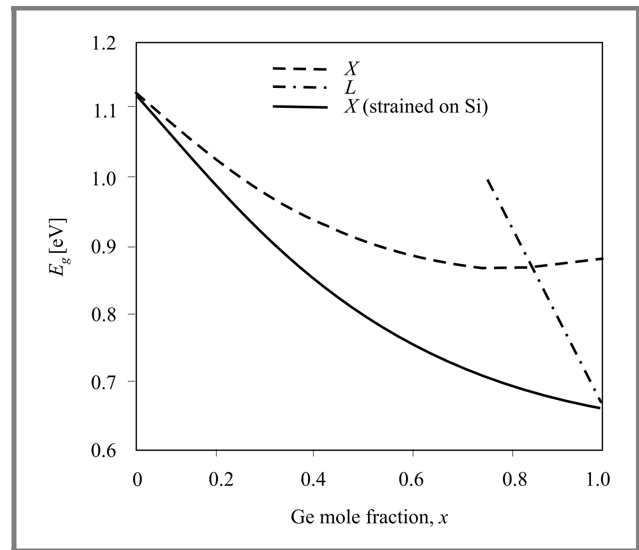


Fig. 2. Material composition dependence of the L and X-bandgaps in $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ at 300 K.

The stress-dependent change of the bandgap is an effect which must be separated from dopant-dependent bandgap narrowing (BGN) which for itself depends on the semiconductor material composition, the doping concentration, and the lattice temperature [38].

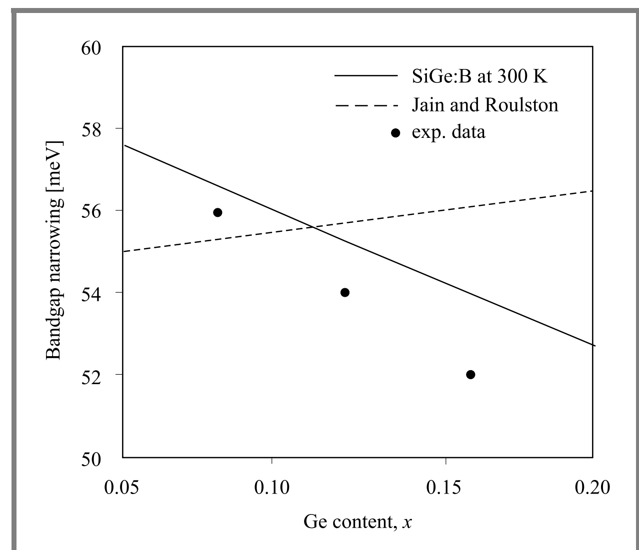


Fig. 3. Doping-dependent bandgap narrowing versus Ge content in p-SiGe compared to experimental data.

In Fig. 3 BGN versus material composition in boron-doped $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ is compared to another model [39]. The decrease of the BGN with increase of the Ge fraction was already experimentally observed [40, 41]. Our theoretical approach

explains this effect by the decreased density of states in the valence band and an increase of the relative permittivity in the strained SiGe alloy.

3.2. Carrier mobility

As the minority carrier mobility is of considerable importance for bipolar transistors, an analytical low field mobility model which distinguishes between majority and minority electron mobilities has been developed [38] using Monte Carlo simulation data for electrons in Si. A similar expression is currently implemented in Minimos-NT:

$$\mu_n^{\mathrm{maj}} = \frac{\mu_n^{\mathrm{L}} - \mu_{mid}^{\mathrm{maj}}}{1 + \left(\frac{N_{\mathrm{D}}}{C_{mid}}\right)^{\alpha}} + \frac{\mu_{mid}^{\mathrm{maj}} - \mu_{hi}^{\mathrm{maj}}}{1 + \left(\frac{N_{\mathrm{D}}}{C_{hi}^{\mathrm{maj}}}\right)^{\beta}} + \mu_{hi}^{\mathrm{maj}}, \quad (3)$$

$$\mu_n^{\mathrm{min}} = \frac{\mu_n^{\mathrm{L}} - \mu_{mid}^{\mathrm{min}}}{1 + \left(\frac{N_{\mathrm{A}}}{C_{mid}}\right)^{\alpha}} + \frac{\mu_{mid}^{\mathrm{min}} - \mu_{hi}^{\mathrm{min}}}{1 + \left(\frac{N_{\mathrm{A}}}{C_{hi}^{\mathrm{min}}}\right)^{\beta}} + \mu_{hi}^{\mathrm{min}}, \quad (4)$$

where μ^{L} is the mobility for undoped material, μ_{hi} is the mobility at the highest doping concentration. μ_{mid}^{maj}, μ_{hi}^{maj}, μ_{mid}^{min}, μ_{hi}^{min}, C_{mid}, C_{hi}^{maj}, C_{hi}^{min}, α, and β are used as fitting parameters. The final low-field electron mobility μ^{LI}, which accounts for a combination of both acceptor and donor doping is given by

$$\mu_n^{\mathrm{LI}} = \left(\frac{1}{\mu_n^{\mathrm{maj}}} + \frac{1}{\mu_n^{\mathrm{min}}} - \frac{1}{\mu_n^{\mathrm{L}}}\right)^{-1}. \quad (5)$$

Figure 4 demonstrates a good match between the analytical model, our Monte Carlo simulation data, and measurements from [42–45] at 300 K for Si.

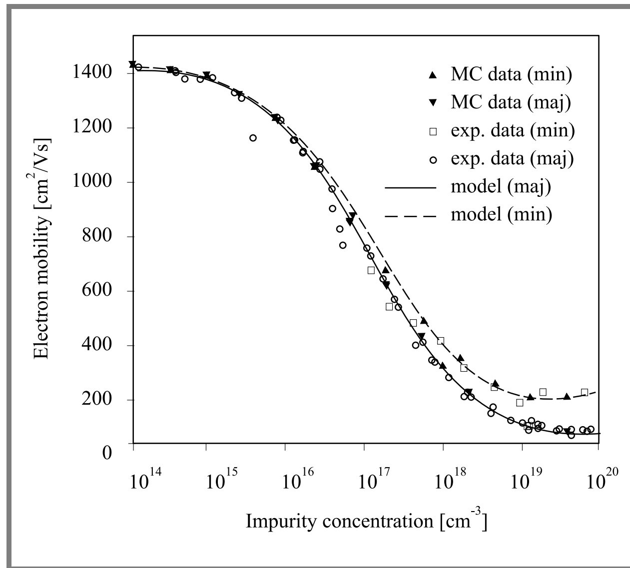


Fig. 4. Majority and minority electron mobility in Si at 300 K: comparison between Monte Carlo simulation data and experimental data.

Monte Carlo simulation which accounts for alloy scattering and the splitting of the anisotropic conduction band valleys due to strain [46] in combination with an accurate ionized impurity scattering model [47] allowed us to obtain results for SiGe for the complete range of donor and acceptor concentrations and Ge contents x. We use the same functional form to fit the doping dependence of the in-plane mobility component for $x = 0$ and $x = 1$ (Si and strained Ge on Si). The material composition dependence is modeled by

$$\frac{1}{\mu(x)} = \frac{1-x}{\mu^{\mathrm{Si}}} + \frac{x}{\mu^{\mathrm{Ge}}} + \frac{(1-x)\cdot x}{C_{\mu}} \quad (6)$$

C_{μ} is a bowing parameter which equals 140 cm^2/Vs and 110 cm^2/Vs for doping levels below and above C_{mid}, respectively. Figure 5 shows the in-plane minority electron mobility in $Si_{1-x}Ge_x$ as a function of x at 300 K for different acceptor doping concentrations. The model parameters used for SiGe at 300 K are summarized in Table 3.

Table 3
Parameter values for the majority/minority electron mobility at 300 K

Parameter	Si	Ge(on Si)	Unit
μ_n^{L}	1430	560	cm^2/Vs
μ_{mid}^{maj}	44	80	cm^2/Vs
μ_{hi}^{maj}	58	59	cm^2/Vs
μ_{mid}^{min}	141	124	cm^2/Vs
μ_{hi}^{min}	218	158	cm^2/Vs
α	0.65	0.65	
β	2.0	2.0	
C_{mid}	$1.12 \cdot 10^{17}$	$4.0 \cdot 10^{17}$	cm^{-3}
C_{hi}^{maj}	$1.18 \cdot 10^{20}$	$4.9 \cdot 10^{18}$	cm^{-3}
C_{hi}^{min}	$4.35 \cdot 10^{19}$	$5.4 \cdot 10^{19}$	cm^{-3}

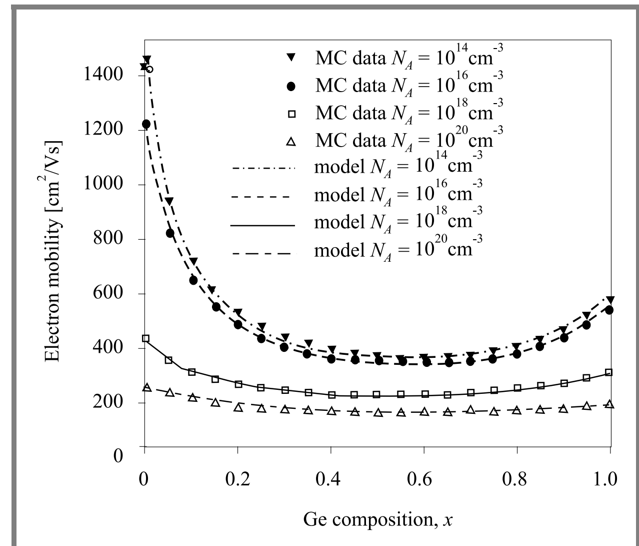


Fig. 5. Minority electron mobility in $Si_{1-x}Ge_x$ as a function of x for in-plane direction: the model is in good agreement with Monte Carlo simulation data.

The component of the mobility perpendicular to the surface is then obtained by a multiplication factor given by the ratio of the two mobility components. The good agreement of the model with the measured and the Monte Carlo simulation data, both for in-plane and perpendicular to the surface directions, is illustrated in Fig. 6.

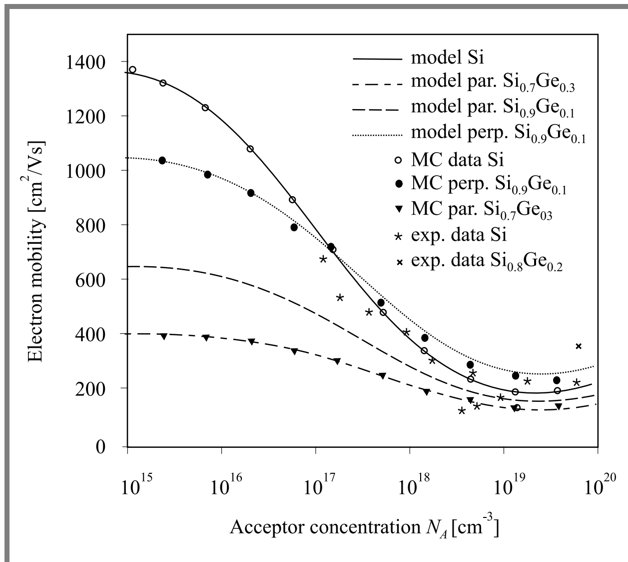


Fig. 6. Minority electron mobility in $Si_{1-x}Ge_x$ as a function of N_A and x: the model is in good agreement with measurements and Monte Carlo simulation data both for in-plane and perpendicular to the surface directions.

4. Analyzed SiGe HBT structures

In this section we analyze SiGe HBTs from an industrial vendor. The devices are part of proven 0.8 μm and 0.35 μm BiCMOS technologies which include CMOS process and high-performance analog-oriented HBT module. The applications reach from circuits for mobile communication to high-speed networks.

Our methodology for characterization and optimization of SiGe HBTs involves process calibration, device calibration employing two-dimensional device simulation, and automated technology computer aided design optimization.

4.1. Device fabrication and process simulation

The devices under investigation are polysilicon-emitter double-base SiGe HBTs epitaxially grown by a chemical vapor deposition process. An implanted n-well, similar to the one used in the standard CMOS technology, is used. The buried layer is connected to a sinker to conduct the electron current from the buried layer to the collector contact. The base consists of an intrinsic base (below the emitter window) and an extrinsic base. The germanium content has a triangular shape. The base-emitter junction is formed by rapid thermal processing which causes outdiffusion of arsenic from the polysilicon emitter layer into the crystalline silicon.

The process simulation with DIOS [8] starts from the blank wafer to the final device and reflects real device fabrication as accurately as possible. The implant profiles as well as annealing steps are calibrated to one-dimensional SIMS profiles. To save computational resources the simulation domain covers only one half of the real device which is symmetric and the collector-sinker is not included in the structure.

4.2. SiGe HBT from the 0.8 μm technology node

The influence of the selectively-implanted-collector (SIC) doping on device performance was studied in order to obtain an optimal profile for specific requirements (high speed or high breakdown voltage). For that purpose, four SiGe HBT structures with emitter areas of 6×0.8 μm^2 have been investigated both experimentally and by means of process simulation, followed by two-dimensional device simulation. The simulated device structure with the phosphorus SIC implant is shown in Fig. 7.

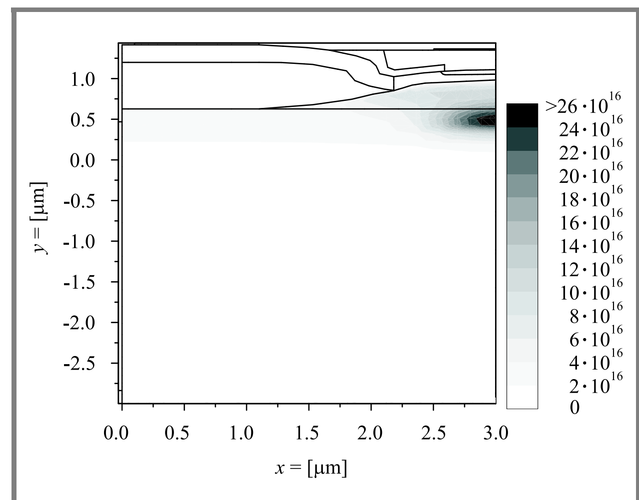


Fig. 7. Simulated device structure (0.8 μm technology) and phosphorus collector implant [cm^{-3}].

The only process step in which the four HBTs (hereafter referred to as dev. 1, dev. 2, dev. 3, dev. 4) differ is the combination of energy and dose used for the SIC implants, as summarized in Table 4. The resulting phosphorus doping profiles in vertical cuts under the emitter windows of the four devices are shown in Fig. 8.

Table 4
Summary of key process and device parameters

Device	Energy [keV]	Dose [cm^{-2}]	f_T [GHz]	BV_{CE0} [V]	$f_T \cdot BV_{CE0}$ [GHz·V]
Dev. 1	480	$7\cdot10^{12}$	32	4.0	128
Dev. 2	480	$3\cdot10^{13}$	40	3.7	148
Dev. 3	300	$7\cdot10^{12}$	33	3.1	102
Dev. 4	300	$3\cdot10^{13}$	42	2.3	97

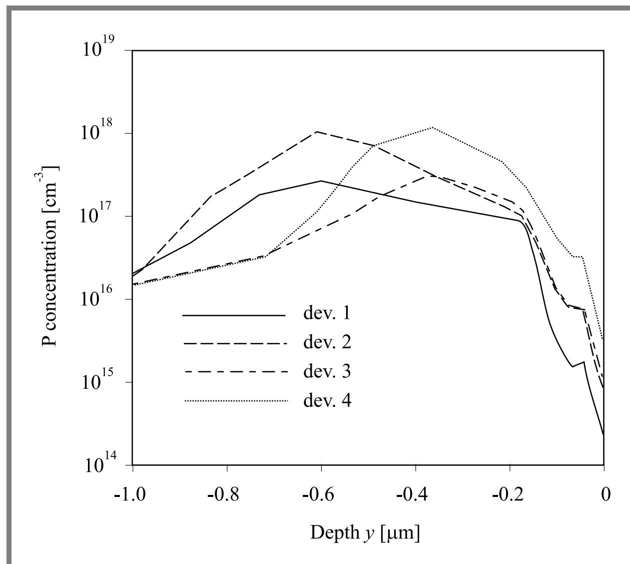


Fig. 8. Phosphorus doping profile under the emitter contact for all four devices.

A comparative Monte Carlo simulation of ion implantation [48] of phosphorus in silicon and SiGe was performed to check the accuracy of the process simulation with respect to SiGe (Fig. 9).

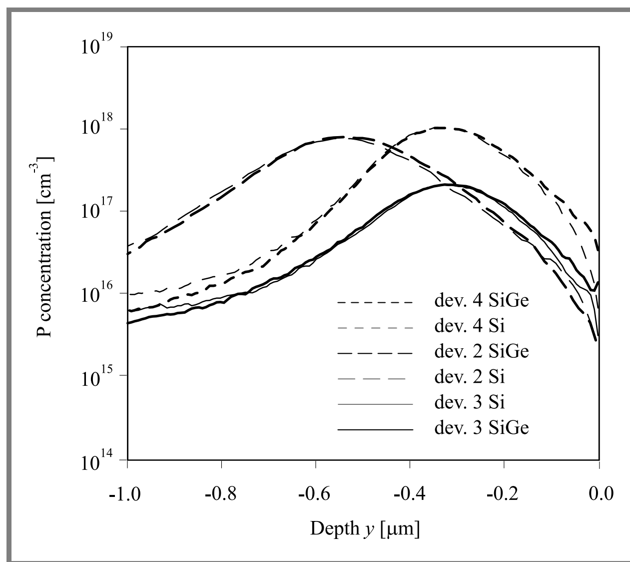


Fig. 9. Comparative simulation of Monte Carlo ion implantation of phosphorus in Si and SiGe.

The physical models in Minimos-NT are well calibrated [49]. The same is true for DESSIS, used for comparison. Both device simulators correctly reproduce the measured forward Gummel plot at 300 K (Fig. 10) with default models. The slight increase of collector current I_C with dose and energy at high bias is due to the differences in the base push-out effect. f_T is extracted by small-signal AC-analysis.

The only fitting parameters used in the simulation are the contribution of bandgap narrowing to the conduction band (here about 80% and 20% for donor and acceptor doping, respectively), and the concentration of traps in the Shockley-Read-Hall recombination model (here 10^{13} cm^{-3}).

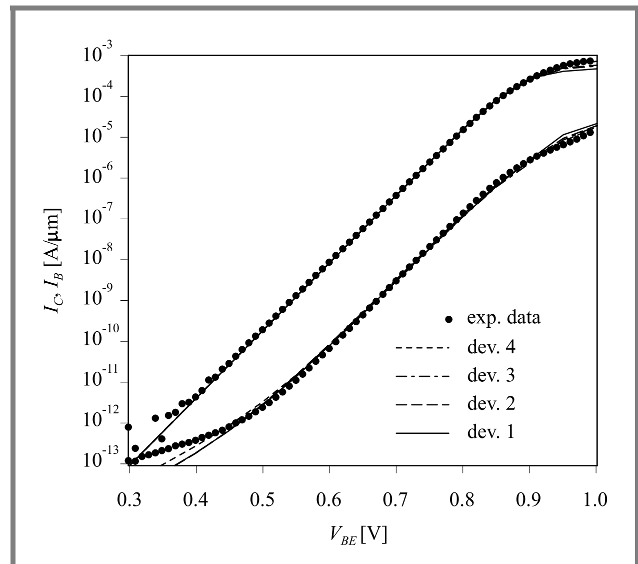


Fig. 10. Forward Gummel plots at V_{CB} = 0 V. Comparison between measurement and simulation.

However, as can be seen in Figs. 11 and 12, both DESSIS and Minimos-NT failed to explain the experimentally observed similarity in peak f_T for dev. 1 and dev. 3 and, respectively, for dev. 2 and dev. 4. This again turned our attention to the SIC implant. An automated device calibration with our TCAD framework [50] was performed. It turned out that 50% more phosphorus in the collector of the two low-dose devices (dev. 1 and dev. 3) already gives an acceptable qualitative agreement.

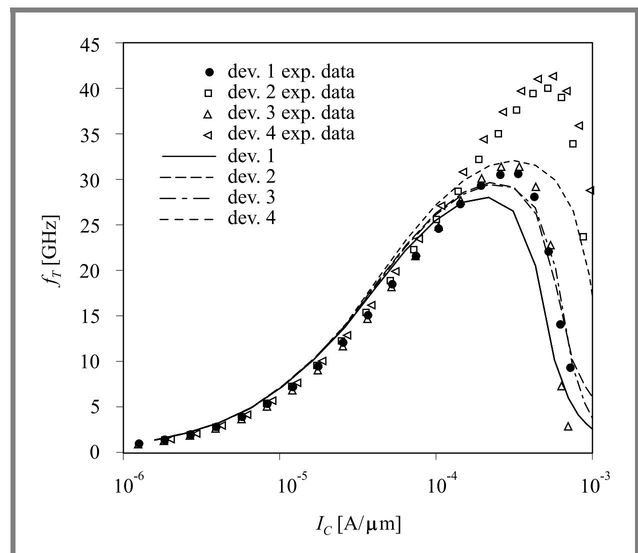


Fig. 11. Frequency f_T versus I_C at V_{CE} = 1.5 V. Comparison between measurement and drift-diffusion simulation with DESSIS.

It is known that with shrinking device dimensions nonlocal effects, such as velocity overshoot, become more pronounced. Neglecting these effects can be a reason for un-

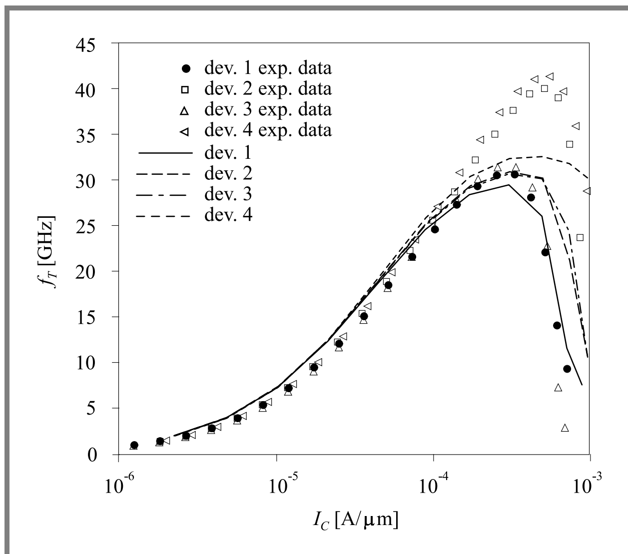


Fig. 12. Frequency f_T versus I_C at V_{CE} = 1.5 V. Comparison between measurement and drift-diffusion simulation with Minimos-NT.

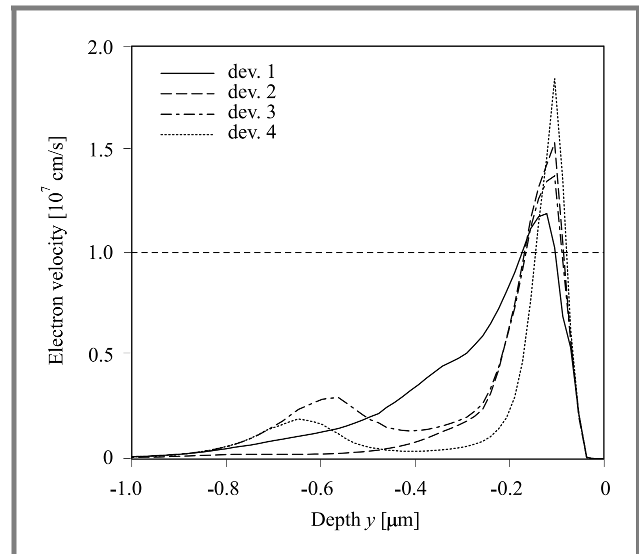


Fig. 14. Electron velocity overshoot in the base-collector space charge region at $V_{CE} = V_{BE}$ = 0.88 V.

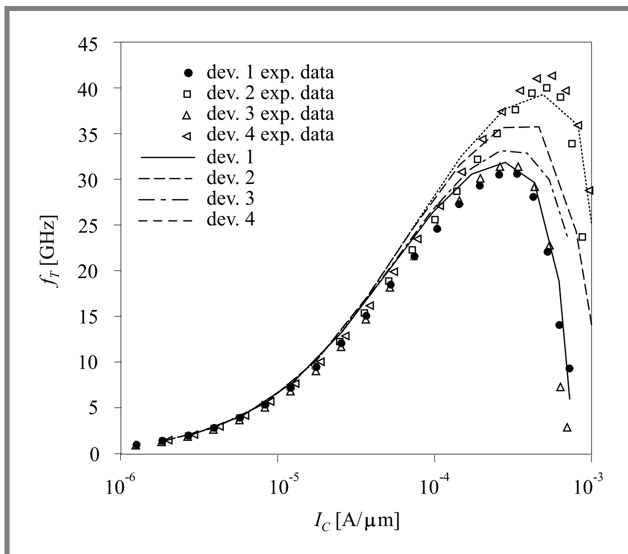


Fig. 13. Frequency f_T versus I_C at V_{CE} = 1.5 V. Comparison between measurement and hydrodynamic simulation with Minimos-NT.

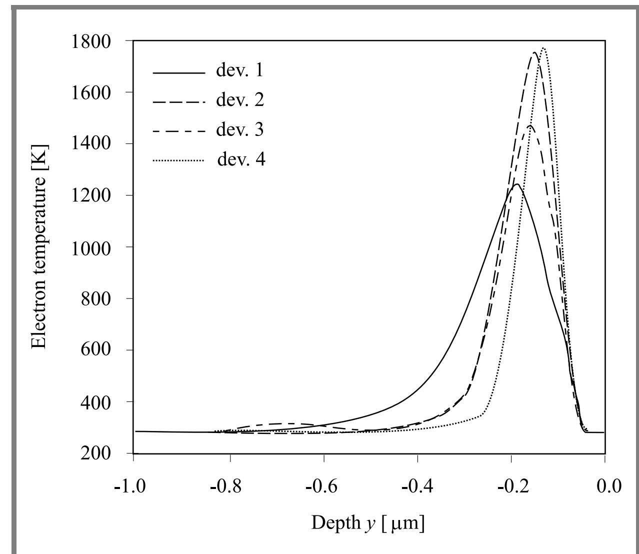


Fig. 15. Electron temperature distribution in the four simulated devices at $V_{CE} = V_{BE}$ = 0.88 V.

derestimating f_T [51]. For that purpose, we performed simulations with the hydrodynamic transport model which improved the results quantitatively (Fig. 13). Figure 14 shows the velocity overshoot over the greater part of the base region which is about twice the saturation velocity limit in the drift-diffusion case (10^7 cm/s). This correlates to the higher electron energy (Fig. 15) in the collector and explains the increase of f_T in comparison to drift-diffusion simulations (see Figs. 11 and 12). The good agreement at low currents is very important since HBTs typically operate at much lower frequencies than at the maximum f_T. Simulations prove that in this range optimizations of the SIC implant do not have the influence on f_T, i.e. the base-emitter capacitance and not the base-collector capacitance is dominating. The maximum f_T was found to have a stronger dependence on the dose than on the energy of the implants.

Furthermore, the important figure of merit $BV_{CE0} \cdot f_T$ (see Table 4) reaches a maximum for high SIC implant energies (deep implant) and high SIC doses. We found that the higher f_T for high-dose/low-energy SIC implants is due to a smaller base width and a delayed onset of the base push-out effect due to the higher collector doping.

4.3. SiGe HBT from the 0.35 μm technology node

The investigated SiGe HBTs from the next generation have emitter areas of $12 \cdot 0.4\ \mu m^2$. The device structure with the phosphorus SIC implant is shown in Fig. 16.

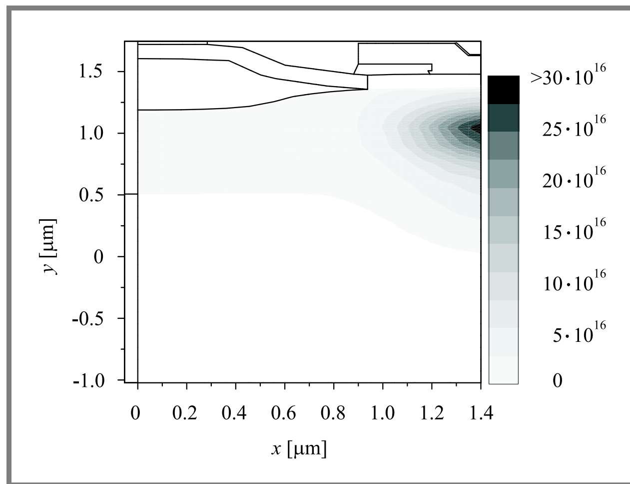


Fig. 16. Simulated device structure (0.35 μm technology) and phosphorus collector implant [cm^{-3}].

All important physical effects, such as surface recombination, impact ionization (II) generation, and self-heating (SH), are properly modeled and accounted for in the simulation in order to get good agreement with the measured forward (Fig. 17) and output characteristics (Fig. 18) using a concise set of models and parameters. In contrast, simulation without including SH effects cannot reproduce the experimental data, especially at high power levels. The only fitting parameters used in the simulation are the contribution of BGN to the conduction band, the trap charge density in the Shockley-Read-Hall recombination model (here 10^{14} cm^{-3}), the velocity recombination for holes in the polysilicon contact model [52] used at the emitter contact, and the substrate thermal resistance.

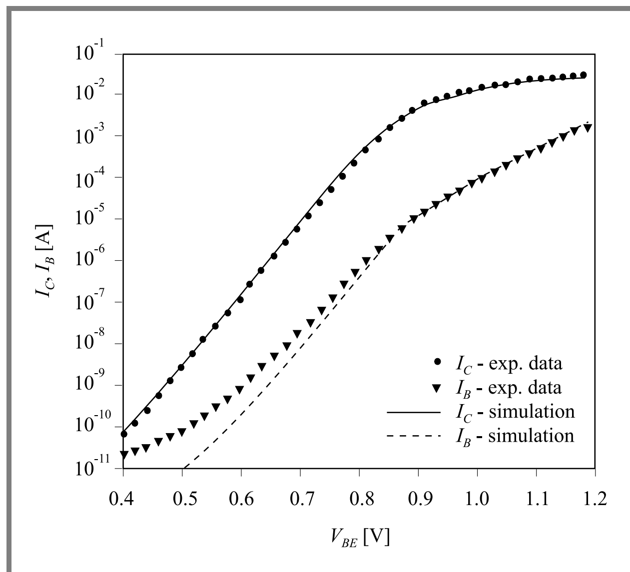


Fig. 17. Forward Gummel plots at V_{CB} = 0 V: comparison between measurement data and simulation at room temperature.

A closer look at the increasing collector current I_C at high collector-to-emitter voltages V_{CE} and constant base current I_B, stepped by 0.4 μA from 0.1 μA to 1.7 μA, reveals

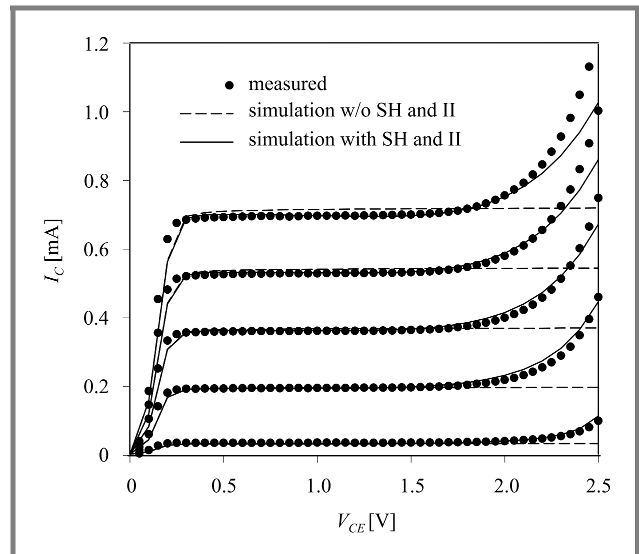


Fig. 18. Output characteristics: simulation with and without self-heating (SH) and impact ionization (II) compared to measurement data. I_B is stepped by 0.4 μA from 0.1 μA to 1.7 μA.

the interplay between self-heating and impact ionization (Fig. 19). While impact ionization leads to a strong increase of I_C, self-heating decreases it. In fact, both I_C and I_B increase due to self-heating at a given bias condition. As the change is relatively higher for I_B, in order to maintain it at the same level, V_{BE} and, therefore, I_C decrease.

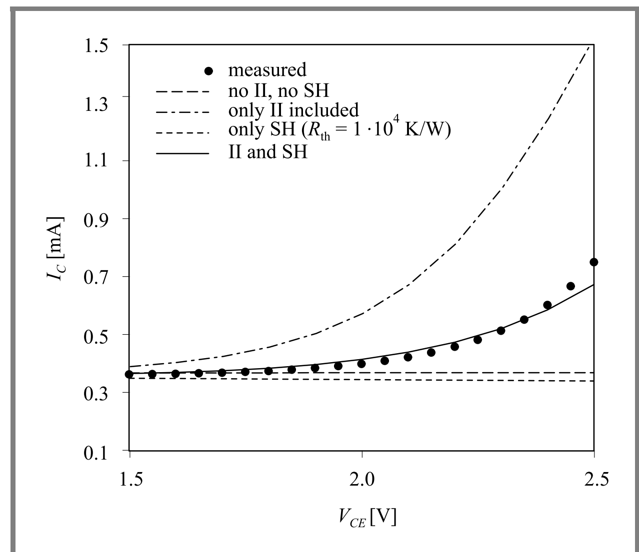


Fig. 19. Output characteristics for $I_B = 0.9$ μA: a closer look at the increasing I_C at high V_{CE} reveals the interplay between self-heating (SH) effect and impact ionization (II) generation.

A proper DC calibration is an important prerequisite for AC simulation (Fig. 17) Note that it is absolutely necessary for AC simulations to take the complete device structure into account in order to consider the capacitances between collector and substrate C_{CS} as well as between base and collector C_{BC}.

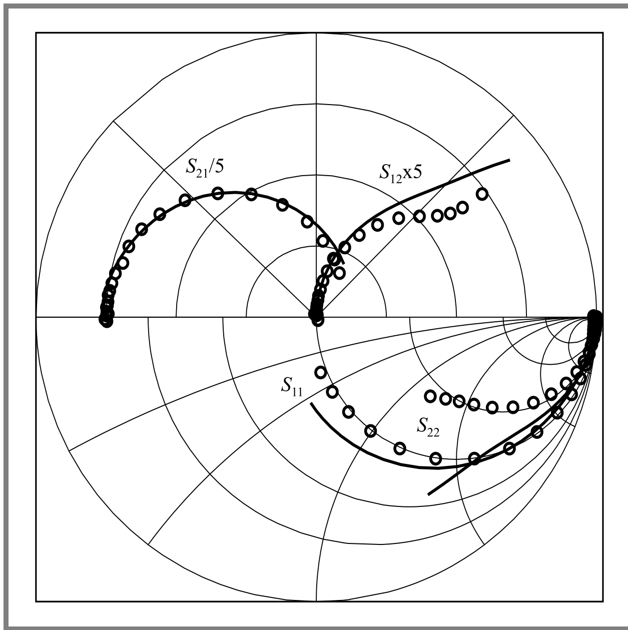


Fig. 20. S-parameters in a combined Smith chart (radius = 1) from 50 MHz to 31 GHz at V_{CE} = 1 V and current density J_C = 28 kA/cm^2 (measurements with circles).

The quality of the simulated (intrinsic) Y-parameters is proven by calculating the row and column sums of the admittance matrix, which have to be zero according to Kirchhoff's laws. The simulation yields errors of about 10^{-16} A/V for typical matrix entries of 10^{-3} A/V. The transformation to intrinsic S-parameters is completely analytical and, thus, the results can be directly compared to the measurement data. Since the measurement environment accounts for the parasitics, no transformation to extrinsic parameters is necessary.

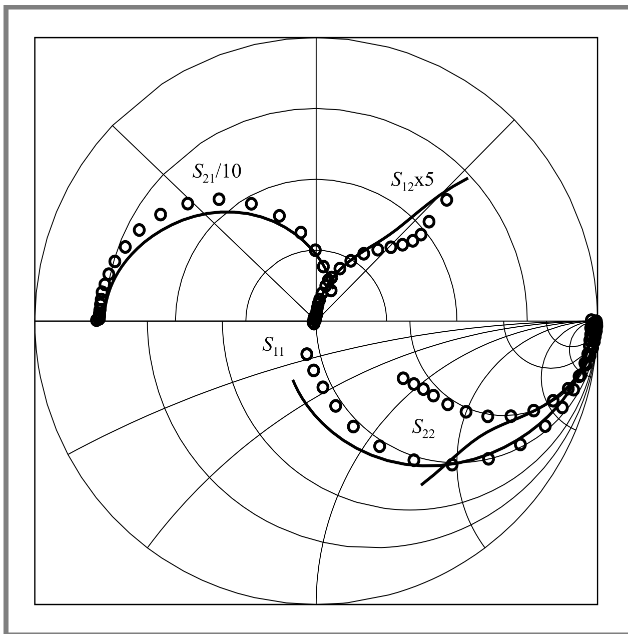


Fig. 21. S-parameters in a combined Smith chart (radius = 1) from 50 MHz to 31 GHz at V_{CE} = 1 V and current density J_C = 76 kA/cm^2 (measurements with circles).

Figures 20 and 21 show a comparison of simulated and measured S-parameters at V_{CE} = 1 V and current densities J_C = 28 kA/cm^2 and J_C = 76 kA/cm^2 in the frequency range between 50 MHz and 31 GHz. For the same device we calculated the matched gain g_m and the short-circuit current gain h_{21} in order to extract the figures of merit f_T and $f_{\max}$ found at the respective unity-gain points.

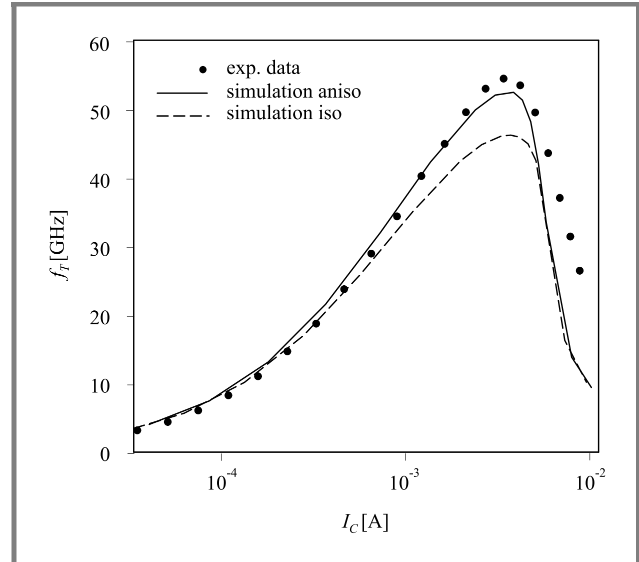


Fig. 22. Cut-off frequency f_T versus collector current I_C at $V_{CE} = 1$ V (anisotropic with solid line, isotropic with dashed line, measurements with circles).

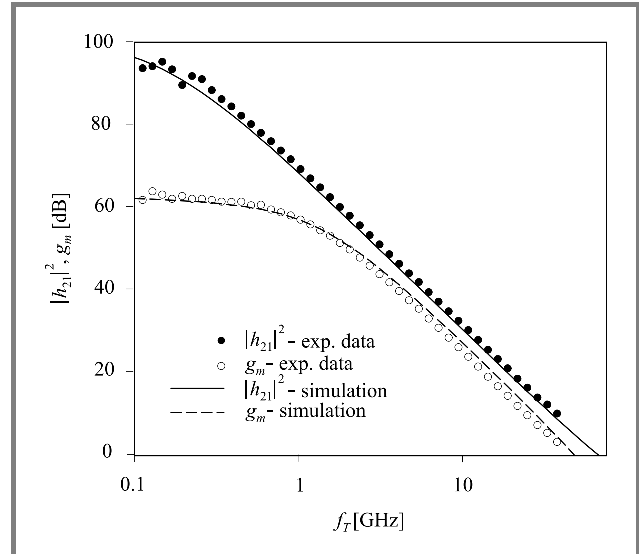


Fig. 23. Short-circuit current gain h_{21} and matched gain g_m versus frequency at V_{CE} = 1 V and current density J_C = 76 kA/cm^2 (measurements with circles).

Figures 22 and 23 show the comparison of our results and the corresponding measurement data. While the measurement covers a range up to 31 GHz the simulation is extended to frequencies beyond the unity-gain point. The peak of the f_T-curve in Fig. 22 corresponds exactly to the frequency at the respective intersection in Fig. 23. In ad-

dition, the effect of the introduction of anisotropic electron mobility is demonstrated in Fig. 22.

5. Conclusion

A brief overview of the state-of-the-art of simulation tools for SiGe HBTs has been given. Critical issues for numerical modeling of SiGe devices have been discussed including accurate models for bandgap narrowing and minority/majority electron mobility in strained SiGe. We have presented experiments and simulations of SiGe HBTs. Good agreement was achieved both with experimental DC-results (forward and output characteristics) and with high-frequency data. We were able to extract various sets of small-signal parameters as well as related figures of merit by means of simulation with Minimos-NT. The newly established models are beneficial for future process development.

Acknowledgment

The authors acknowledge the ion implantation simulation from A. Hössinger and valuable inputs from G. Röhrer, S. Wagner, T. Grasser, and H. Kosina. The work is supported by the Austrian Science Fund (FWF), Project P14483-MAT, by CLPP BAS Center of Excellence in IT, by austriamicrosystems AG, Unterpremstätten, Austria, and by Infineon Technologies AG, Munich, Germany.

References

[1] J. Böck, H. Schäfer, H. Knapp, D. Zöschg, K. Aufinger, M. Wurzer, S. Boguth, M. Rest, R. Schreiter, R. Stengl, and T. Meister, "Sub 5 ps SiGe bipolar technology", *IEDM Tech. Dig.*, pp. 763–766, 2002.

[2] J.-S. Rieh, B. Jagannathan, H. Chen, K. Schonenberg, D. Angell, A. Chinthakindi, J. Florkey, F. Golan, D. Greenberg, S.-J. Jeng, M. Khater, F. Pagette, C. Schnabel, P. Smith, A. Stricker, K. Volant, D. Ahlgren, G. Freeman, K. Stein, and S. Subbanna, "SiGe HBTs with cut-off frequency near 300 GHz", *IEDM Tech. Dig.*, pp. 771–774, 2002.

[3] B. Jagannathan, M. Meghelli, A. V. Rylyakov, R. A. Groves, A. K. Chinthakindi, C. M. Schnabel, D. A. Ahlgren, G. G. Freemann, K. J. Stein, and S. Subbanna, "A 4.2-ps ECL ring-oscillator in a 285 GHz f_{max} SiGe technology", *IEEE Electron Dev. Lett.*, vol. 23, no. 9, pp. 541–543, 2002.

[4] T. Hashimoto, Y. Nonaka, T. Saito, K. Sasahara, T. Tominari, K. Sakai, K. Tokunaga, T. Fujiwara, S. Wada, T. Udo, T. Jinbo, K. Washio, and H. Hosoe, "Integration of a 0.13-μm CMOS and a high performance self-aligned SiGe HBT featuring low base resistance", *IEDM Tech. Dig.*, pp. 779–782, 2002.

[5] APSYS, http://www.crosslight.com/downloads/downloads.html

[6] ATLAS/Blaze, http://www.silvaco.com/products/vwf/atlas/

[7] BIPOLE3, http://www.bipsim.com/mainframe.html

[8] DESSIS and DIOS, http://www.ise.com/products/index.html

[9] G-PISCES-2B, http://www.gateway-modeling.com/products.htm

[10] MEDICI, www.synopsys.com/products/avmrg/device_sim_ds.html

[11] E. Buturla, P. Cottrell, B. Grossman, and K. Salsburg, "Finite-element analysis of semiconductor devices: the FIELDAY program," http://www.research.ibm.com/journal/rd/441/buturla.pdf

[12] NEMO, http://www.cfdrc.com/nemo/

[13] PISCES-ET, http://www-tcad.stanford.edu/tcad.html

[14] FLOODS and FLOOPS, http://www.tec.ufl.edu/ flooxs/

[15] C. Jungemann, B. Neinhüs, and B. Meinerzhagen, "Full-band Monte Carlo device simulation of a SiGe/Si HBT with a realistic Ge profile", *IEICE Trans. Electron.*, vol. E83-C, no. 8, pp. 1228–1234, 2000.

[16] DEVICE, http://www.uv.ruhr-uni-bochum.de/

[17] nextnano3, http://www.webplexity.de/nextnano3.php

[18] J. Geßner, F. Schwierz, H. Mau, D. Nuernbergk, M. Roßberg, and D. Schipanski, "Simulation of the frequency limits of SiGe HBTs", in *Proc. Model. Simul. Microsyst.*, Puerto Rico, 1999, pp. 407–410.

[19] Minimos-NT 2.0 User's Guide, Institut für Mikroelektronik, Technische Universität Wien, Austria, http://www.iue.tuwien.ac.at/software/minimos-nt

[20] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*. Wien, New York: Springer, 1984.

[21] W. Hänsch, *The Drift Diffusion Equation and its Application in MOSFET Modeling*. Wien, New York: Springer, 1991.

[22] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*. Wien, New York: Springer, 1989.

[23] K. Hess, Ed., *Monte Carlo Device Simulation: Full Band and Beyond*. Boston, Dordrecht, London: Kluwer, 1991.

[24] H. Kosina and S. Selberherr, "A hybrid device simulator that combines Monte Carlo and drift-diffusion analysis", *IEEE Trans. Comput. Aid. Des.*, vol. 13, no. 2, pp. 201–210, 1994.

[25] W. Engl, A. Emunds, B. Meinerzhagen, H. Peifer, and T. Thoma, "Bridging the gap between the hydrodynamic and the Monte Carlo model – an attempt", in *Proc. VLSI Process./Dev. Model. Worksh.*, Osaka, 1989, pp. 32–33.

[26] S. Laux and M. Fischetti, "The DAMOCLES Monte Carlo device simulation program", in *Computational Electronics*, K. Hess, J. Leburton, and U. Ravaioli, Eds. Kluwer, 1991, pp. 87–92.

[27] W. Hänsch, T. Vogelsang, R. Kircher, and M. Orlowski, "Carrier transport near the Si/SiO_2 interface of a MOSFET", *Solid State Electron.*, vol. 32, no. 10, pp. 839–849, 1989.

[28] K. Dragosits, V. Palankovski, and S. Selberherr, "Two-dimensional modeling of quantum mechanical effects in ultra-short CMOS devices", in *Advances in Simulation, Systems Theory and Systems Engineering*, N. Mastorakis, V. Kluev, and D. Koruga, Eds. WSEAS Press, 2002, pp. 113–116.

[29] D. Richey, J. Cressler, and A. Joseph, "Scaling issues and Ge profile optimization in advanced UHV/CVD SiGe HBT's", *IEEE Trans. Electron Dev.*, vol. 44, no. 3, pp. 431–440, 1997.

[30] B. Gonzales, V. Palankovski, H. Kosina, A. Hernandez, and S. Selberherr, "An energy relaxation time model for device simulation", *Solid State Electron.*, vol. 43, pp. 1791–1795, 1999.

[31] R. Quay, C. Moglestue, V. Palankovski, and S. Selberherr, "A temperature dependent model for the saturation velocity in semiconductor materials", *Mat. Sci. Semicond. Process.*, vol. 3, no. 1–2, pp. 149–155, 2000.

[32] V. Palankovski and S. Selberherr, "Thermal models for semiconductor device simulation", in *Proc. Eur. Conf. High Temper. Electron.*, Berlin, 1999, pp. 25–28.

[33] R. Quay, R. Reuter, V. Palankovski, and S. Selberherr, "*S*-parameter simulation of RF-HEMTs", in *Proc. High Perform. Electron Dev. Microw. Opt. Appl. EDMO*, Manchester, 1998, pp. 13–18.

[34] S. Wagner, V. Palankovski, T. Grasser, R. Schultheis, and S. Selberherr, "Small-signal analysis and direct *S*-parameter extraction", in *Proc. Int. Symp. Electron Dev. Microw. Opt. Appl. EDMO*, Manchester, 2002, pp. 50–55.

[35] J. Eberhardt and E. Kasper, "Bandgap narrowing in strained SiGe on the basis of electrical measurements on Si/SiGe/Si hetero bipolar transistors", *Mat. Sci. Eng.*, vol. B89, pp. 93–96, 2002.

[36] Y. Varshni, "Temperature dependence of the energy gap in semiconductors", *Physica*, vol. 34, pp. 149–154, 1967.

[37] S. Jain, *Germanium-Silicon Strained Layers and Heterostructures*, vol. 24 of *Advances in Electronics and Electron Physics*. Academic Press, 1994.

[38] V. Palankovski, G. Kaiblinger-Grujin, and S. Selberherr, "Implications of dopant-dependent low-field mobility and band gap narrowing on the bipolar device performance", *J. Phys. IV*, vol. 8, pp. 91–94, 1998.

[39] S. Jain and D. Roulston, "A simple expression for band gap narrowing (BGN) in heavily doped Si, Ge, GaAs and Ge_xSi_{1-x} strained layers", *Solid State Electron.*, vol. 34, no. 5, pp. 453–465, 1991.

[40] Ž. Matutinović-Krstelj, V. Venkataraman, E. Prinz, J. Sturm, and C. W. Magee, "A comprehensive study of lateral and vertical current transport in $Si/Si_{1-x}Ge_x/Si$ HBT's", *IEDM Tech. Dig.*, pp. 87–90, 1993.

[41] M. Libezny, S. Jain, J. Poortmans, M. Caymax, J. Nijs, R. Mertens, K. Werner, and P. Balk, "Photoluminiscence determination of the Fermi energy in heavily doped strained $Si_{1-x}Ge_x$ layers", *Appl. Phys. Lett.*, vol. 64, no. 15, pp. 1953–1955, 1994.

[42] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus- and boron-doped silicon", *IEEE Trans. Electron Dev.*, vol. ED-30, no. 7, pp. 764–769, 1983.

[43] K. Wolfstirn, "Hole and electron mobilities in doped silicon from radiochemical and conductivity measurements", *J. Phys. Chem. Solids*, vol. 16, pp. 279–284, 1960.

[44] S. E. Swirhun, D. E. Kane, and R. M. Swanson, "Measurements of electron lifetime, electron mobility and band-gap narrowing in heavily doped p-type silicon", *IEDM Tech. Dig.*, pp. 24–27, 1986.

[45] I. Y. Leu and A. Neugroschel, "Minority-carrier transport parameters in heavily doped p-type silicon at 296 and 77 K", *IEEE Trans. Electron Dev.*, vol. 40, no. 10, pp. 1872–1875, 1993.

[46] S. Smirnov, H. Kosina, and S. Selberherr, "Investigation of the electron mobility in strained $Si_{1-x}Ge_x$ at high Ge composition", in *Proc. Simul. Semicond. Process. Dev.*, pp. 29–32, 2002.

[47] H. Kosina and G. Kaiblinger-Grujin, "Ionized-impurity scattering of majority electrons in silicon", *Solid State Electron.*, vol. 42, no. 3, pp. 331–338, 1998.

[48] A. Hössinger and S. Selberherr, "Accurate three-dimensional simulation of damage caused by ion implantation", in *Proc. Model. Simul. Microsyst.*, Puerto Rico, 1999, pp. 363–366.

[49] S. Selberherr, W. Hänsch, M. Seavey, and J. Slotboom, "The evolution of the MINIMOS mobility model", *Solid State Electron.*, vol. 33, no. 11, pp. 1425–1436, 1990.

[50] T. Grasser, R. Strasser, M. Knaipp, K. Tsuneno, H. Masuda, and S. Selberherr, "Device simulator calibration for quartermicron CMOS devices", in *Proc. Simul. Semicond. Process. Dev.*, K. De Meyer and S. Biesemans, Eds., Leuven, 1998, pp. 93–96.

[51] C. Jungemann, B. Neinhüs, and B. Meinerzhagen, "Comparative study of electron transit times evaluated by DD, HD, and MC device simulation for a SiGe HBT", *IEEE Trans. Electron Dev.*, vol. 48, no. 10, pp. 2216–2220, 2001.

[52] Z. Yu, B. Ricco, and R. Dutton, "A comprehensive analytical and numerical model of polysilicon emitter contacts in bipolar transistors", *IEEE Trans. Electron Dev.*, vol. 31, no. 6, pp. 773–784, 1984.

Vassil Palankovski was born in Sofia, Bulgaria, in 1969. He received the diploma degree in electronics from the Technical University of Sofia in 1993. Afterwards he worked in the telecommunications field for three years. He joined the Institut für Mikroelektronik at the Technische Universität Wien in 1997, where he received the doctoral degree in 2000, and is currently working as a postdoctoral researcher. In summer 2000 Dr. Palankovski held a visiting research position at LSI Logic Corp., Milpitas, California. His scientific interests include device and circuit simulation, heterostructure device modeling, and physical aspects in general.

e-mail: Palankovski@iue.tuwien.ac.at
Institute for Microelectronics, TU Vienna
Gusshausstr. 27–29, A-1040 Vienna, Austria

Siegfried Selberherr (IEEE Fellow) was born in Klosterneuburg, Austria, in 1955. He received the degree of diplom-ingenieur in electrical engineering and the doctoral degree in technical sciences from the Technische Universität Wien in 1978 and 1981, respectively. Dr. Selberherr has been holding the venia docendi on computer-aided design since 1984. Since 1988 he has been the head of the Institut für Mikroelektronik and since 1999 he is dean of the Fakultät für Elektrotechnik und Informationstechnik at the Technische Universität Wien, Austria. His current topics are modeling and simulation of problems for microelectronics engineering.

e-mail: Selberherr@iue.tuwien.ac.at
Institute for Microelectronics, TU Vienna
Gusshausstr. 27–29, A-1040 Vienna, Austria

Invited paper

Recent developments in vertical MOSFETs and SiGe HBTs

Stephen Hall, Octavian Buiu, Peter Ashburn, and Kees de Groot

Abstract — There is a well recognised need to introduce new materials and device architectures to Si technology to achieve the objectives set by the international roadmap. This paper summarises our work in two areas: vertical MOSFETs, which can allow increased current drive per unit area of Si chip and SiGe HBT's in silicon-on-insulator technology, which bring together and promise to extend the very high frequency performance of SiGe HBT's with SOI-CMOS.

Keywords* — *vertical MOSFET, HBT, SOI.

1. Introduction

The ITRS roadmap recognizes the need for innovative device concepts and architectures in addition to the more predictable scaling of MOSFETs. Furthermore, the incorporation of the heterojunction bipolar transistor (HBT) into CMOS processes and further scaling of MOSFETs opens up the r.f. market to mainstream silicon technology. In this paper, we report the status of our work in two related areas. Firstly, we consider innovative vertical MOSFET architectures and secondly, we address work on SiGe HBT's on silicon-on-insulator (SOI).

One vertical MOST device concept features a so-called dielectric pocket (DP) which sits on top of a Si turret. The pocket serves a number of functions; it reduces greatly the influence of the large area parasitic bipolar transistor in the vertical structure and also prevents encroachment of the doping from the extrinsic drain, so reducing electrical bulk punch-through effects. Perhaps most importantly, it reduces charge-sharing effects (SCE) associated with the reverse biased drain and so improves threshold voltage control. The device offers dual channel or gate all around for high current drive with reduced footprint and can be readily scaled to the decananometer regime. We consider also strategies for reducing parasitic capacitance using thickened oxide regions and inhibiting parasitic bipolar transistor action using a poly-SiGe extrinsic source contact.

The SiGe HBT (on SOI) process has the advantage that all the device layers (collector base and emitter) are realized in a single epitaxial growth stage. This involves selective and non-selective growth. We will briefly review our work and report the results of a recent study concerning high injection effects and associated current crowding in the low-doped emitter region of the device. A major showstopper for SiGe HBT's is the transient enhanced diffusion of boron from the very heavily doped base. The incorporation of carbon in substitutional sites during growth of the base has been shown to suppress TED very effectively. Various growth techniques are usually used, with conditions far from equilibrium so leading to metastable alloys with C concentrations exceeding the equilibrium solubility (C supersaturation). Further post growth device processing requires that the structural properties and alloy compositions remain stable during any annealing steps. We comment here on out proposed methodologies to address this new area of activity.

2. Vertical MOSFETs

It is recognised that vertical transistors can overcome scaling problems due to lithography resolution, whereby decananometer channels can be realised with relaxed lithography as the channel length is determined by the accuracy of ion-implantation or epitaxial growth. Vertical transistors also allow double gate or gate all-around structures thus increasing current drive albeit at the expense of increased device capacitance. The architectures are however compact and hence give a high drive with reduced footprint compared to an equivalent lateral architecture, as demonstrated in [1]. Much of the work reported previously is concerned with discrete device fabrication often using fabrication techniques that could not be easily integrated into an advanced CMOS process [2–5]. Rather than present a detailed account of the large body of work gathered on vertical MOSFETs and double gate MOSFETs, we choose to summarise the key developments by identifying three basic approaches to vertical MOSFET realization and these are depicted in Fig. 1.

In the first approach, the device is realised by epitaxy [2]. This is the easiest way to realise devices but has a number of important disadvantages. Perhaps most importantly, the approach does not easily allow the production of p- and n-channel devices required for CMOS. Furthermore, parasitic overlap capacitances are very high and there is a strong parasitic bipolar transistor. The gain of this transistor can be high because the effective base will be very thin (set by the channel length) and also the currents themselves will be high as the effective emitter/collector regions extend across the turret. The device features very deep source/drain junctions and the body region must be highly doped. It will suffer from severe short channel effects. The second approach uses selective epitaxy [3, 4] and by this means parasitic capacitances can be reduced, particularly in the replacement gate approach of Hergenrother *et al.* [4]. Parasitic bipolar gain and currents remain high, however, short channel effects remain severe and there may be reliability problems with the gate oxide which is re-

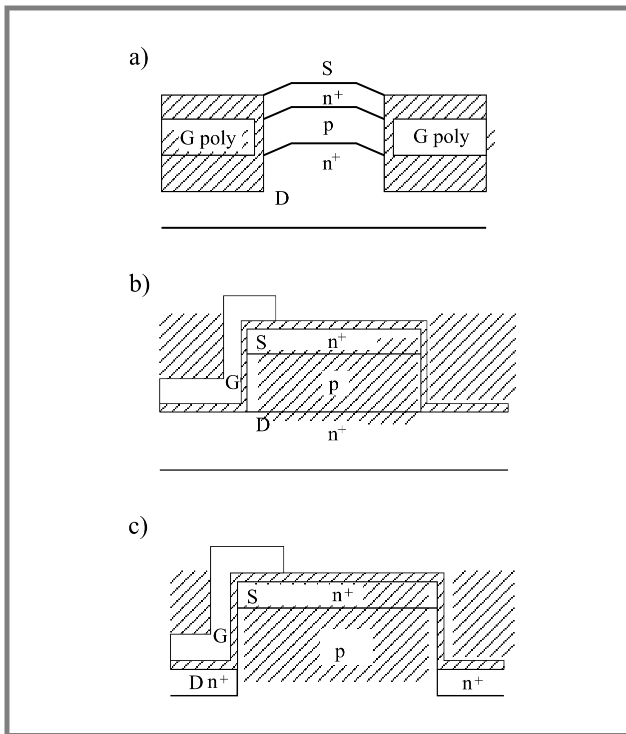


Fig. 1. Vertical MOSFET types: (a) epitaxy; (b) selective epitaxy; (c) ion-implanted.

alized in a non-standard way. The third approach employs ion implantation [5] and allows the realisation of architectures with reduced parasitic bipolar gain and improved short channel effects but the parasitic capacitances remain high. A key advantage of this approach is the ability to devise CMOS comparable processes.

We identify therefore three important research issues for VMOS transistors namely reduction of parasitic capacitance, control of short channel effects and reduction of parasitic bipolar gain. We now investigate the means of addressing these three key issues.

We have reported a novel vertical transistor architecture incorporating a so-called dielectric pocket [6, 7] for control of short channel effects. This concept was first demonstrated successfully within a lateral architecture [8], but implementation in a vertical architecture is considerably simpler. The structure is compatible with strategies reported previously to reduce parasitic capacitances in the device [7, 9].

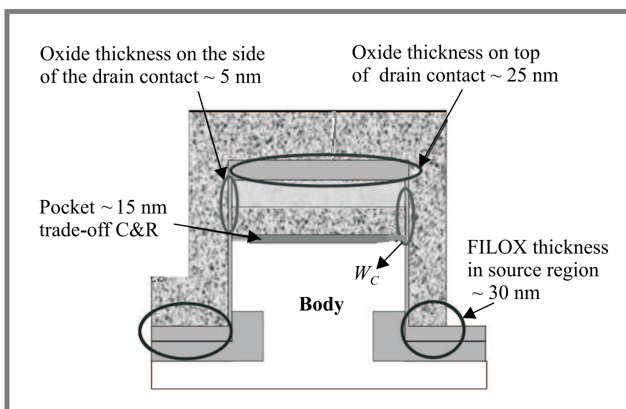


Fig. 2. Vertical MOSFET architecture, featuring the dielectric pocket and thickened oxide regions to minimise parasitic capacitance.

The basic structure of the VMOST device concept featuring the dielectric pocket is shown in Fig. 2. The pocket serves a number of functions; it reduces greatly the influence of the large area parasitic bipolar transistor in the vertical structure and also prevents encroachment of the doping from the extrinsic drain, so reducing electrical bulk punch-through effects. Perhaps most importantly, it reduces charge sharing effects associated with the reverse biased drain and so improves threshold voltage control. Note the silicidation of the top contact, which we designate the drain, to reduce significantly the associated parasitic resistance. The intrinsic drain contact is formed by out-diffusion from the poly-Si extrinsic drain contact in a manner not unlike that of a poly-emitter on a bipolar transistor process. Care is required with the thermal budget to ensure that only slight out-diffusion occurs from this extrinsic drain contact such that there is no encroachment of drain dopant below the DP into the channel region. Such encroachment would remove the electrostatic influence of the DP on the channel and so mitigate its influence on the SCE. Notice that dual channel operation is achieved by the gate poly-contact running over the pillar width which is set to the minimum feature size to reduce drain/body overlap capacitance.

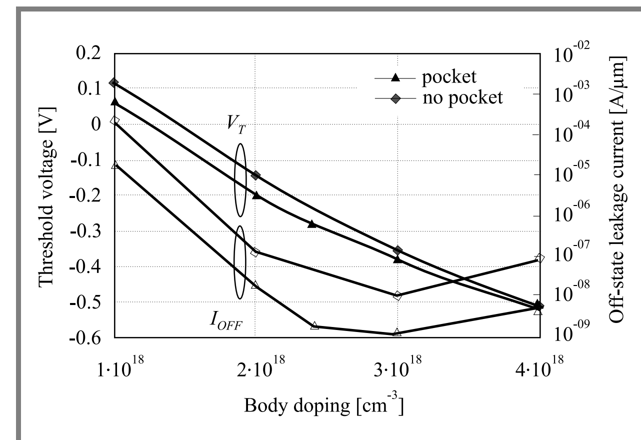


Fig. 3. Threshold voltage and off-state leakage current versus body doping for 50 nm vertical MOSTs with and without dielectric pocket. The gate oxide thickness was 2 nm.

The ISE device simulator was used to obtain electrical characteristics for the DP-VMOST. The Van Dort quantum correction model, hydrodynamic model, avalanche and band-to-band tunnelling were all switched on to provide self-consistent data for short-channel, highly doped devices. The dielectric pocket thickness was set at 15 nm to minimise its parasitic capacitance [7]. The gate oxide thickness was set at a conservative value of 2 nm as the formation of gate oxides on sidewalls with associated corners is likely to prove problematic. The body doping was then varied to

produce electrical characteristics for the PMOS with and without the DP, as summarised in Fig. 3.

The leakage current is seen to be at minimum for a body doping of $3 \cdot 10^{18}\,\mathrm{cm}^{-3}$; band-to-band tunnelling becomes increasingly dominant beyond the minimum whereas punch-through causes the increase in leakage for lower doping levels. A body doping of $2.4 \cdot 10^{18}\,\mathrm{cm}^{-3}$ results in a threshold voltage of -0.28 V and a leakage current of 1.5 nA/μm at a drain to source voltage of -1 V.

Further simulations serve to investigate the influence of the spacing between the pocket and the gate oxide (referred to as the contact region width, W_C) on the threshold voltage. The DP serves to increase the absolute magnitude of threshold voltage because it inhibits charge sharing by the drain. For $W_C = 50$ nm, the V_T equals that of the non-DP device indicating that the DP no longer exerts electrostatic influence. The results are summarised in Fig. 4, where it is shown that decreasing W_C results in significant reduction in I_{OFF} with little reduction in I_{ON}. The latter arises because of the influence on V_T. For $W_C = 20$ nm and a body doping of $2 \cdot 10^{18}\,\mathrm{cm}^{-3}$, an on-current of 1 mA/$\mu$m and off-current 10 nA/$\mu$m is obtained. Simulated transistor characteristics (not shown) reveal a DIBL of 80 mV for a body doping of $2.4 \cdot 10^{18}\,\mathrm{cm}^{-3}$. Repeating the simulations with a gate oxide thickness of 1.2 nm results in a DIBL of 30 mV, $I_{OFF} = 7$ nA/μm, $I_{ON} = 1.2$ mA/μm.

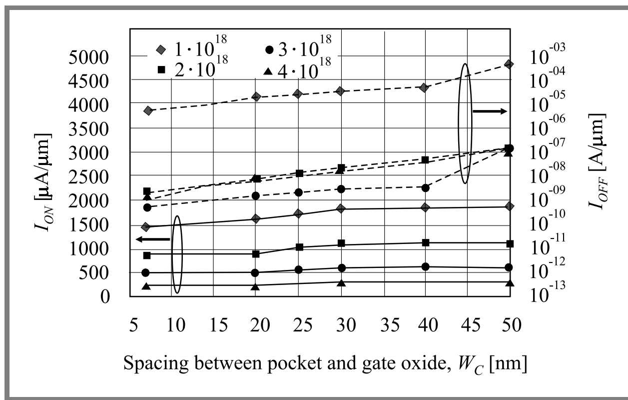


Fig. 4. I_{ON} and I_{OFF} for different contact widths, W_C.

3. Parasitic capacitance reduction

The approaches to reduce the parasitic capacitances are indicated in Fig. 1 and involve thickening oxide regions using the following techniques. Reduction of gate-source overlap capacitance (bottom of pillar) is achieved by a LOCOS type process FILOX [9, 10]. Incorporation of a deposited oxide region on top of the pillar and the thicker oxide grown on the highly doped poly-Si extrinsic drain contact can reduce significantly the gate-drain capacitance. A previous study using MOS-capacitors has demonstrated a 5-fold reduction in parasitic overlap capacitance using FILOX and a thick top oxide [10].

A semi-analytical transient model has been developed to assist in the optimization of the device with regard to parasitic capacitances [7]. The methodology is to consider an inverter loaded by n other inverters. A unity fan-out CMOS inverter circuit is shown in Fig. 5, where the parasitic capacitances are identified. These capacitive components of the device structure are readily related to the device geometries and physical parameters – doping levels, oxide thicknesses, etc. using standard equations.

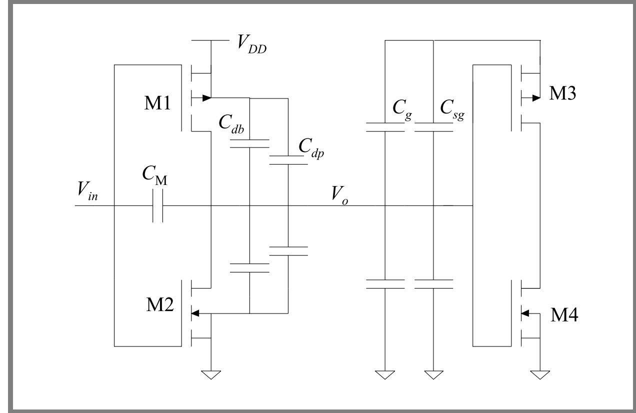


Fig. 5. Circuit used as a test bench.

The capacitances can be lumped together into one single capacitor, C_T assigned to the output node, V_o by recognising series and parallel configurations and treating V_{in} and V_{DD} as virtual earths. As the gate to source/drain overlap capacitances (C_{gso}/C_{gdo}) are responsible for a significant portion of C_T in a vertical structure, these capacitances were isolated from the total gate source/drain capacitance, (C_{gs}/C_{gd}). Other capacitances considered were gate to substrate (included with C_{sg}), drain to bulk depletion region C_{db}, dielectric pocket capacitance, C_{dp} taking into account depletion under the oxide, and finally the capacitance due to the drain contact. The voltage dependence of depletion capacitance is linearised in the usual manner. Note that vertical devices benefit from being operated "source down" because this reduces the switched depletion capacitance compared to "drain down", i.e. source at the top of the turret.

We define a circuit performance metric $\tau = 3\tau_f$ where τ_f is the fall time of V_o following an abrupt input voltage step. The fall time can be calculated by considering the discharge of the effective load capacitance, C_T, via the NMOS pull-down transistor, between 10 and 90% of the full voltage swing as indicated in the model equations below, where $i(t)$ is the transient current through the NMOS and other symbols have their usual meaning. Gate field dependence of mobility is included also:

$$\tau = 3\left(\tau_{fs} + \tau_{fus}\right) = -3C_T(V_o) \int_{0.1\,V_{DD}}^{0.9\,V_{DD}} \frac{dt}{i(t)},$$

$$\tau_{fs} = \frac{C_T}{2C_{ox}Wv_{sat}}\left(1+\frac{L\cdot\xi_{sat}}{V_{DD}-V_{th}}\right)\int_{V_{Dsat}}^{V_H}\frac{1}{(V_{DD}-V_{th})(1+\lambda V_o)}dV_o,$$

$$\tau_{fus} = \frac{C_T L}{2C_{ox}W\mu_{eff}(V_{DD})}\int_{V_{Lt}}^{V_{Dsat}}\frac{1}{(V_{DD}-V_{th})V_o-\frac{V_o^2}{2}}\,dV_o\,.$$

The latter two equations represent the time the device is saturated and unsaturated, respectively. The equations are easily solved numerically. To validate the model, a unity fan-out inverter circuit was simulated using the SpectreS simulator in Cadence v4.4.3 and a production 350 nm lateral device SPICE deck. The simulated fall time was equal to 58 ps while the calculated fall time was equal to 49 ps, an error of 15% that represents reasonable agreement for a comparative study.

Using this methodology, we have identified the approximate thicknesses of dielectric regions required to optimise the VMOST. The results of the study are indicated in Fig. 1 (thickness values quoted).

4. Reduction of parasitic BJT gain

The DP concept brings the added benefit of reducing considerably the area of the effective emitter and collector formed by the source and drain of the VMOSFET and so reduces the magnitude of base and collector currents of the PBT. The DP device is however inherently "drain up".

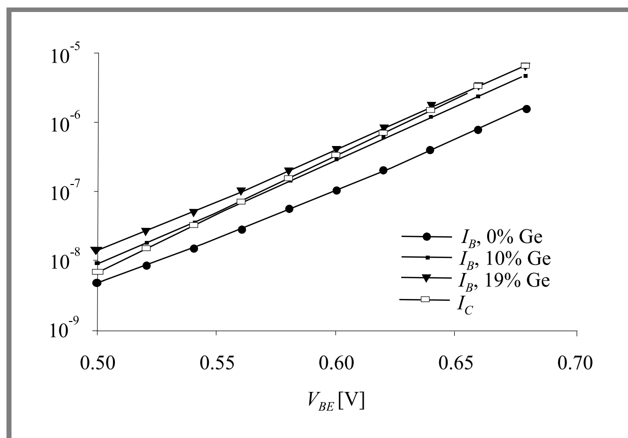


Fig. 6. Increase in base current as a result of Ge incorporation into the emitter.

An alternative approach to minimise parasitic bipolar transistor gain for drain-down configurations (source at the top), is to include a poly-SiGe extrinsic source contact [11]. This serves to steepen the profile of minority carriers injected into the parasitic emitter so increasing the base current and reducing the gain. Figure 6 shows Gummel plots from test bipolar transistors where a considerable reduction in BJT gain is evident.

A theoretical model for the base current of such a poly-SiGe emitter has been developed, which combines the effects of the poly-SiGe grains, the grain boundaries and the interfacial layer at the poly-SiGe/Si interface into an expression for the effective surface recombination velocity of a poly-SiGe emitter [11]. The model is equally valid for the parasitic BJT in vertical MOSFETs. Silicon bipolar transistors were fabricated with 0, 10 and 19% Ge in the poly-SiGe emitter and the variation of base current with Ge content characterised. The measured base current for a poly-SiGe emitter was seen to increase by a factor of 3.2 for 10% Ge and 4.0 for 19% Ge compared with a control transistor containing no germanium, in good agreement with the theoretical predictions. The competing mechanisms of base current increase by Ge incorporation into the polysilicon and base current decrease due to an interfacial oxide layer were investigated. The size of the base current increase with Ge content depends on the thickness of the interfacial layer, with larger increases being obtained for thinner interfacial layers. The introduction of germanium into a polysilicon emitter therefore allows the base current, and hence the gain, to be controlled by means of the Ge content in the poly-SiGe parasitic emitter of a vertical MOSFET.

5. Test transistors

Both p-channel and n-channel vertical MOS transistors have been successfully fabricated to demonstrate some of the concepts outlined above. Figure 7 shows a field effect SEM of a vertical transistor, demonstrating the FILOX concept whereby thickened oxide regions are achieved in the source and drain gate overlap regions, to minimise overlap capacitance. Both pillar top and bottom have a 60 nm thick FILOX oxide as visible under the gate electrode and this oxide encroachment reduces the gate to top electrode capacitance. Calculations based on the electron micrographs, and process and device simulations give an estimate for capacitance reduction of a factor of 3.

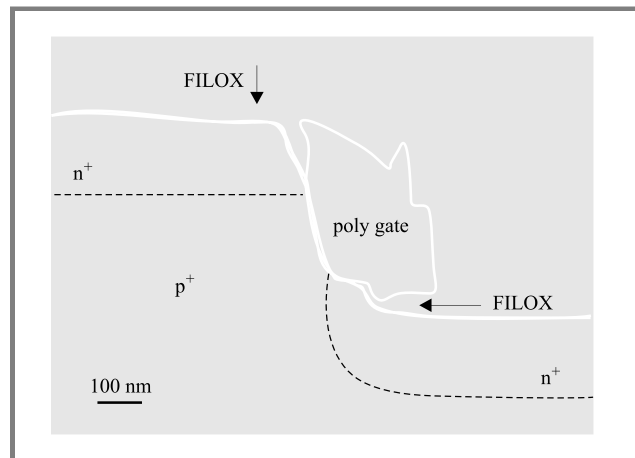


Fig. 7. Field emission SEM cross-section of a surround gate ion-implanted vertical NMOS.

Figure 8 shows transfer characteristics of vertical NMOS transistors similar to those shown in Fig. 7. Gate ox-

ide thickness is 3 nm and channel length is estimated to be 100 nm. Characteristics on the right employ a single gate whereas those on the left have a surround gate and hence considerably higher drive.

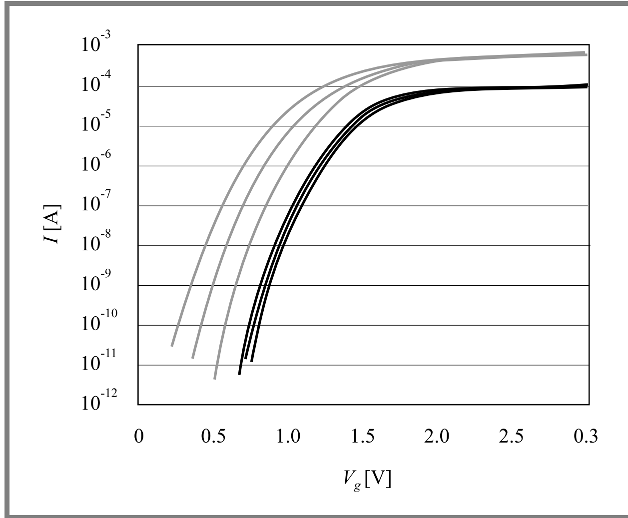


Fig. 8. Transfer characteristics of vertical n-channel transistors.

Transfer characteristics of surround gate vertical PMOS transistors are shown in Fig. 9. The threshold voltage is higher than normal since the gate oxide thickness of 10 nm was chosen very conservatively to prevent shorts in the corners of the transistors. Otherwise the transistors are of very high quality with negligible variation over the wafer.

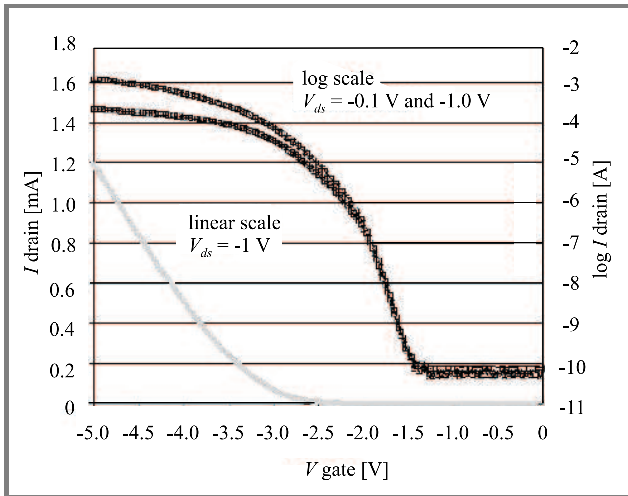


Fig. 9. Transfer characteristics of a 300 nm surround gate ion-implanted vertical PMOSFET.

6. Performance appraisal

The VMOS inverter with 2 nm gate oxide was compared to equivalent minimum sized lateral CMOS inverters taken from ITRS 2001. The width of the PMOST, W_p was set to 2 W_n. The source and drain resistances of the VMOST

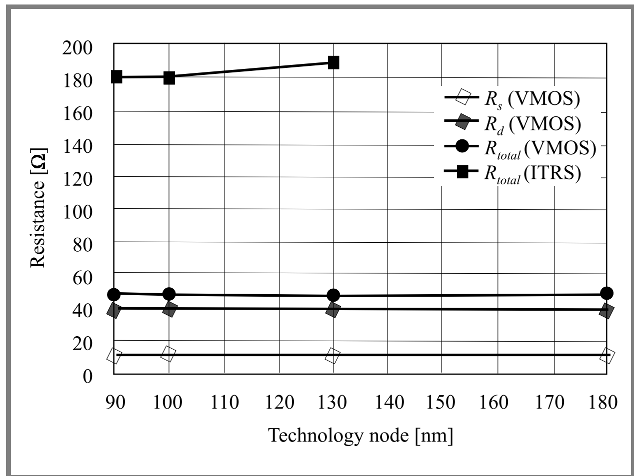


Fig. 10. Scaling of series resistance.

(extracted from ISE simulation) are 3 to 4 times smaller than the ITRS values and these resistances do not significantly increase when the device is scaled as shown in Fig. 10. This arises essentially because the dielectric pocket precludes the need for pockets or extensions.

Figure 11 serves to quantify the performance of the optimised VMOS up to the 45 nm technology node, using the delay metric $C_T V_{DD}/I_{ON}$. The capacitance C_T obtained

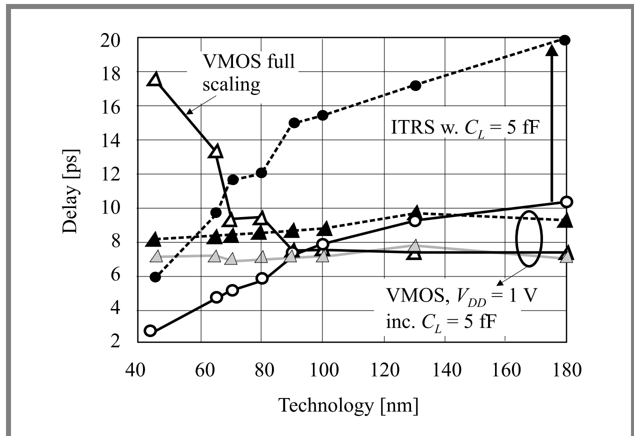


Fig. 11. Projected performance of VMOS in context of the roadmap. Delay approaching that of the 40 nm ITRS node can be realised with earlier generation rules for loaded inverters (5 fF).

from ISE simulations includes gate overlap and depletion capacitances [7]. The VMOS shows a clear advantage down to the 90 nm node (note that the channel length of the VMOS is not scaled). Thereafter the performance deteriorates largely because of the reduced V_{DD}. The effect of maintaining V_{DD} for the VMOST is shown also. Beyond the 100 nm node, we can say that load capacitance due to interconnections will have a very dominant role and so the additional drive capability of the VMOS should continue to offer considerable advantage. We demonstrate this by including a fixed load capacitance of 5 fF in the calculation

of delay. In this context, the smaller relative degradation in performance of the VMOST compared with the ITRS roadmap demonstrates the potential of the latter device.

7. Heterojunction bipolar devices

Silicon-germanium heterojunction transistors allow Si technology to penetrate lucrative radio-frequency application markets of mobile communications and local area networks. For the bipolar transistor, the silicon-on-insulator technology brings advantages of reduced collector capacitance and reduced cross-talk for mixed signal circuits. In the context of BiCMOS, the CMOS circuitry gains much advantage from SOI; namely simplified process flow and latch-up immunity together with enhanced MOSFET device performance. In particular, the lower threshold voltage and low junction capacitance associated with SOI-CMOS is particularly suited to low-voltage, low power applications. The SiGe HBT also offers enhanced performance at lower current levels due to the collector current enhancement thus there are overall benefits for lower power, high performance SOI-BiCMOS circuits. Bond and etch back SOI (BESOI) is preferred for HBT application due to the better Si and SiO_2 quality.

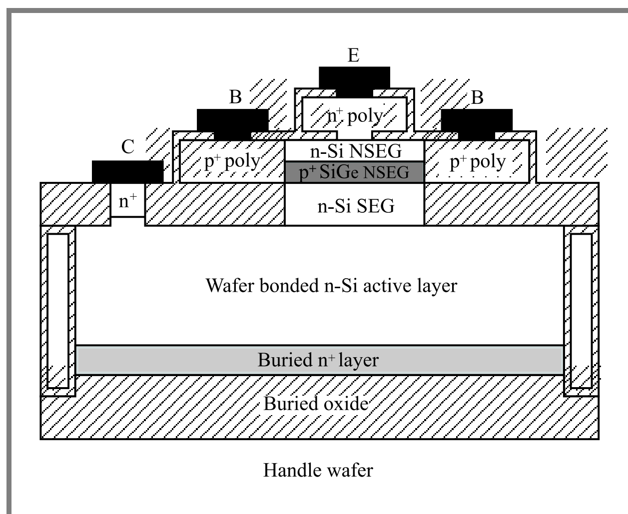


Fig. 12. Schematic diagram of the bonded wafer SOI device architecture.

A schematic diagram of the SiGe HBT on SOI is shown in Fig. 12. The bonded SOI wafers used feature a 1 μm buried oxide layer with the surface Si layer thinned to a nominal thickness of 1.5 μm. Deep, poly-Si filled trenches provide isolation through to the buried oxide layer. The patterned SOI layer is used to provide the heavily doped buried collector as well as the crystalline seed layer for subsequent epitaxial layer growth of the silicon collector and SiGe base layers. Before the silicon epitaxial growth the SOI layer is implanted with a $5 \cdot 10^{15}\text{cm}^{-2}$ dose of arsenic ions at an energy of 160 keV to form the buried collector contact layer. The transistor layers are grown using selective epitaxial growth (SEG) for the Si collector, followed in the same growth step by non-selective epitaxial growth (NSEG) for the p+ SiGe base (nominally 12% Ge) and the n-Si emitter cap [12]. The selective Si collector was grown at 900°C and the non-selective SiGe base and Si emitter cap at 750°C.

The advantages of this approach are that the basic transistor structure is grown in a single epitaxy step and the growth interface is kept away from the transistor active regions. Transistor characteristics show good uniformity across the wafer implying good control of the Ge and B concentrations across the wafer during the non-selective SiGe growth.

Achieving full optimisation of HBT's is facilitated by this growth of collector, base and emitter layers in the same growth step. The incorporation of a low-doped emitter (LDE), however, implies a propensity for high emitter resistance particularly in devices where the emitter overlay is large. Furthermore, issues arise as to the realisation of the buried layer and associated collector resistance for HBTs on SOI. We have identified a new parasitic mechanism whereby we explain anomalies in SOI-HBT device characteristics to be the result of current confinement caused by high lateral resistance in the low doped emitter region of the device.

Figure 13 shows sets of Gummel plots for SiGe HBTs (a) and Si controls (b), where the parameter b is varied. Considering first the SiGe HBTs (a), for $b = 0$, we observe high leakage current which we have explained earlier [12]. For small b, we observe correlated turn-over of collector and base current at relatively low values of $V_{BE} (\approx 0.64\text{ V})$ presumably due to series resistance induced, lateral voltage drops. As the base current level is small, we suggest at this stage that the turn-over is a result of resistive drops in the emitter due to the high collector current. At higher bias levels $(0.88 - 1.0\text{ V})$, pronounced I_B-kinks and hard limiting of I_C are evident. The onset of the base current kinks correlate with the hard limiting of I_C, suggesting quasi-saturation due to internal forward biasing of the collector-base junction due to collector series resistance.

Alternatively, the kinks may be due to a combination of emitter resistance and modified kink effect (MKE) [14]. The MKE is an effect associated with the collector-base junction. At high collector current densities, the image charge of holes accumulating at the collector end of the base become sufficiently large to cause the band-bending in the base depletion region to change sign. This has the effect of "mirroring" the valence band edge discontinuity into the conduction band thus causing the appearance of a potential barrier, which results in quasi-saturation of I_C. We will discuss further, which of these two effects (collector resistance or MKE) is the likely explanation. Turning to the Si devices of Fig. 13b, I_C is of course lower than that of the HBT's for a given V_{BE} so collector resistance induced kinks might not be evident. MKE is of course not to be expected in homojunction devices. For higher b-values (5.5, 6 μm), I_C is seen to turn-over at $V_{BE} \approx 0.7$ V and exhibit an ideality factor (m) of two with coincident severe limiting of the base current. This latter trend is evident in both SiGe and Si base devices.

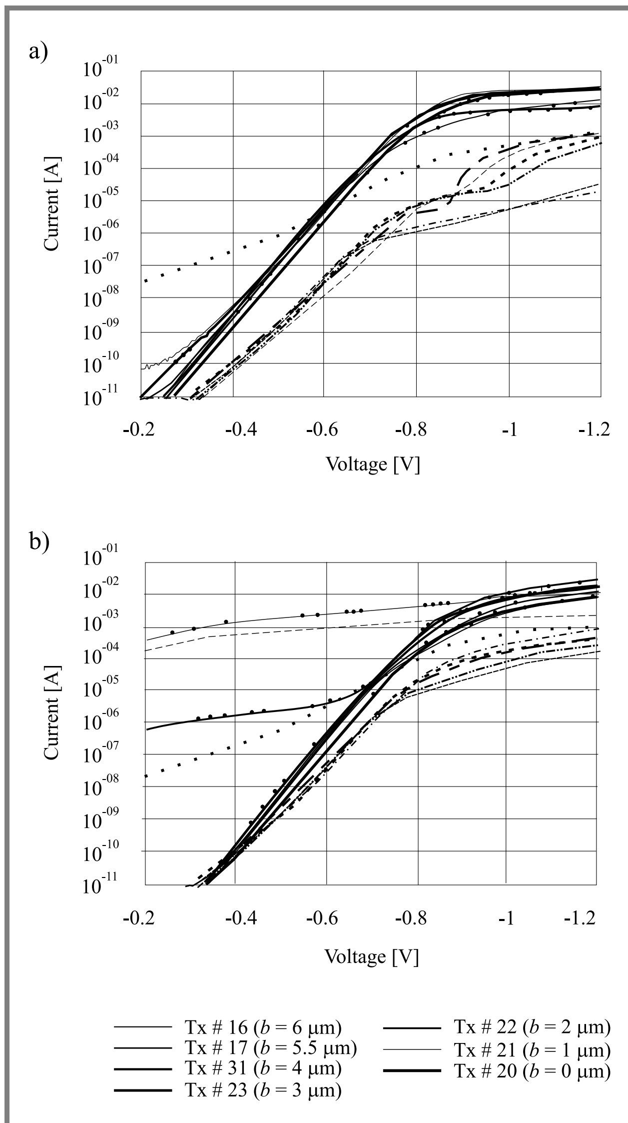


Fig. 13. Measured Gummel characteristics from (a) SiGe HBT (wafer # 6); (b) Si base (wafer # 9).

Figure 14 shows a plot of the collector current dependence of devices with large values of b. It is seen that at low bias the current exhibits an areal dependence with relation to the emitter polysilicon (EP): for these devices the active area is set by the emitter polysilicon. However, at approximately 0.6 V for the HBT and 0.7 V for the BJT the current dependence changes from an EP areal dependence to an emitter window (EW) areal dependence. This bias level corresponds to the onset of the $m = 2$ regime. For devices with smaller b, the emitter polysilicon areal dependence is maintained until higher biases.

We have shown in previous work [13] that the collector current in the $(m = 2)$ – regime shows an activation energy of half the band-gap and the current gain β shows a negative coefficient of temperature. These latter three properties are strong indicators of high injection. The bias level at which the proposed high injection effect occurs is not consistent with the doping levels in either base or emitter so the $(m = 2)$ – region on the characteristics and the (presumably) associated turn over in base current can be considered anomalous.

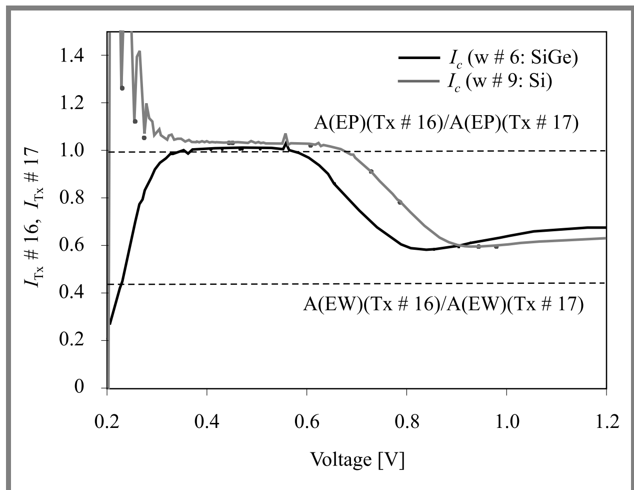


Fig. 14. Device dimensional collector current dependence for devices with large values of b.

Silvaco ATLAS was used to simulate devices with $b = 2\ \mu$m and $b = 6\ \mu$m. The emitter and collector regions were set at uniform doping level and the base was set as a Gaussian function, with a peak doping of $4 \cdot 10^{18}\,\mathrm{cm}^{-3}$. The collector doping was $1 \cdot 10^{17}$ and the emitter doping $N_E = 1 \cdot 10^{18}\,\mathrm{cm}^{-3}$. The collector contact was placed at the back of the device. We consider simulations for $V_{BE} = 0.8$ V, $V_{BC} = 0$ V. This is the bias level at which simulated Gummel pots begin to show the turn-over in I_B. (Note that our simulator is not fully calibrated against experimental results.) For the device with $b = 2\ \mu$m the current flow was seen to be uniform across the whole of the active area at this bias level. Figure 15 shows the simulated electron current density for a device with $b = 6\ \mu$m. Confinement of the current to the centre of the device is clearly evident.

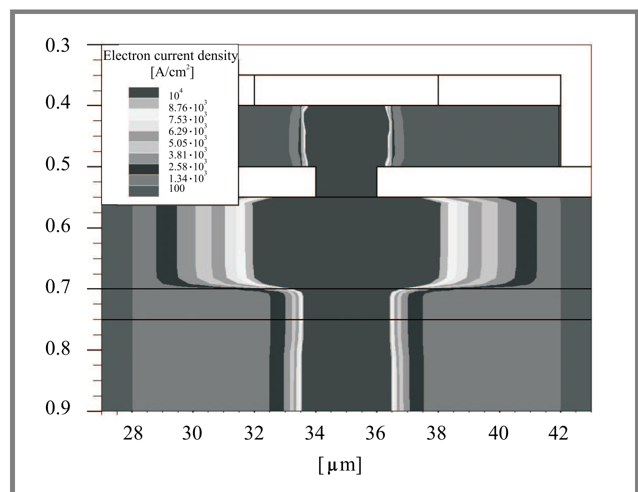


Fig. 15. Simulated electron current density with device biased at $V_{be} = 0.8$ V for $b = 6\,\mu$m.

Figure 16 shows emitter resistance extracted using the Ning-Tang technique [15]. The results for larger b-values will be in error (by about a factor of 2 from Fig. 13) because of the increasing lack of areal dependence. A linear extrapolation from the small values gives an indication of the likely emitter spreading resistance for large b-values.

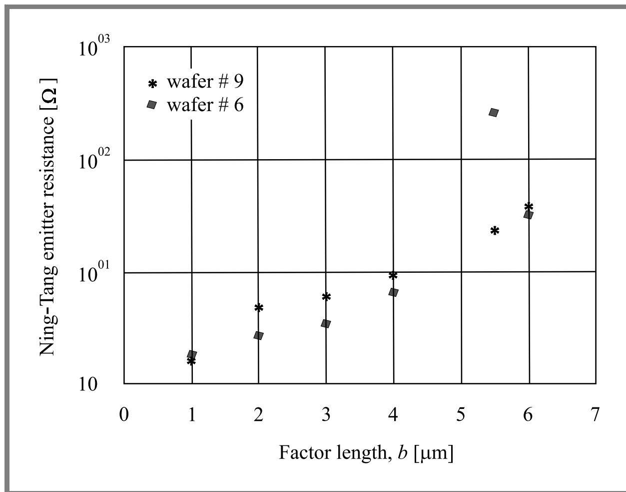


Fig. 16. Emitter resistance extracted using the Ning-Tang technique.

Next, we resolve the issue as to the nature of the kinks for devices with smaller b-values. Applying the methodology of [14] to the $b = 1\ \mu$m curve in Fig. 13, does not produce an ideal I_C and I_B curve with quasi-saturation of I_C. We conclude that the kinks arise from quasi-saturation due to R_C, therefore. This is consistent with the high value of I_C and the estimated value of R_C ($\sim$ 100 Ω). This highlights the problems of realising low resistance access to the collector for SOI-HBTs. Wafer bonding offers the possibility of buried silicide layer to address this issue [16].

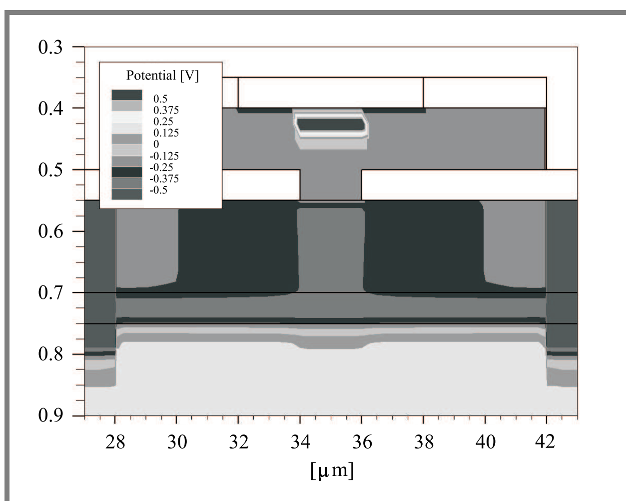


Fig. 17. Simulated potential with device biased at $V_{be} = 0.8$ V for emitter overlay $= 6\ \mu$m.

Finally, Fig. 17 shows a 2D potential distribution, and the de-biaisng effect of the emitter resistance is clearly evident.

8. SiGeC HBT technology

In addition to the parasitic mechanism of the previous section, a major issue with HBTs is parasitic barrier formation by the out-diffusion of boron from the heavily doped base. This is exacerbated by the flow of interstitials arising from implant damage or oxidation of the surface: "transient enhanced diffusion". A strategy to limit this effect is to introduce a small percentage of substitutional carbon (C_s) into the lattice, which allows band engineering and strain control. However, carbon has low solubility in Si and 2D growth conditions must be maintained to achieve substitutional rather than interstitial carbon (C_i) incorporation whilst increasing the critical thickness, decreasing the strain and improving the bandgap shift. The SiGeC layer can be obtained at low temperatures using MBE and CVD epitaxial growth techniques but we are concerned with further processing steps, such as thermal oxidation, and post implant activation anneals. The challenge then is to obtain an alloy layer (SiGeC) with sufficiently high C_s content and to ensure that this is maintained during the subsequent HBT process. Thus thermal budget and Si cap thickness play an important role as well as any source or sink sites formed by interstitials as these have a strong influence on the diffusion and clustering behaviour of carbon. For instance, an oxidation process will generate higher carbon loss rates than a vacuum or inert gas annealing. Redistribution or complete loss of carbon from the SiGeC layer can occur due to C-diffusion and clustering can result in interstitial carbon (C_i) leading to the formation of β – SiC precipitates. The situation is further complicated because these processes are inter-dependent.

The total concentration of carbon inside the alloy can be measured with secondary ion mass spectrometry (SIMS) whereas C_s is typically obtained from high resolution X-Ray diffraction (HRXRD), phonon based techniques such as Raman and Fourier transform infra-red (FTIR) spectroscopy. Cluster formation can be viewed using high resolution TEM (HRTEM), HRXRD, FTIR and Raman spectroscopy. No single technique provides all the necessary information therefore and so we adopt the following methodology. We characterise the layers post-growth, using SIMS for carbon, boron, and germanium total concentration profiles, scanning ellipsometry (SE) which yields layer thickness and carbon concentration (as well as optical properties), FTIR with the C_s line at 610/cm, O_i line at 1100/cm, Si-C lines 700–900/cm, and XRD. Figure 18 shows initial results from SE and XRD showing a disparity between C concentrations, which can be interpreted in terms of C_s and C_i.

Figure 19 shows corresponding layer thickness derived from SE and XRD. Further calibration will be sought from HRTEM.

The results from the various techniques are correlated to provide a self-consistent assessment of layer content and morphology in particular the I/S ratio. If the XRD and if the results from the various techniques are correlated to

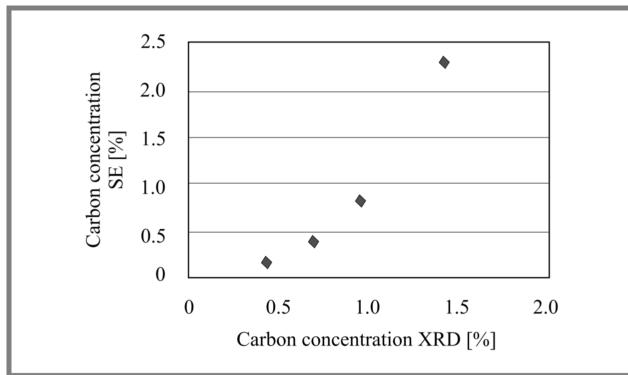


Fig. 18. C concentration derived from SE and XRD.

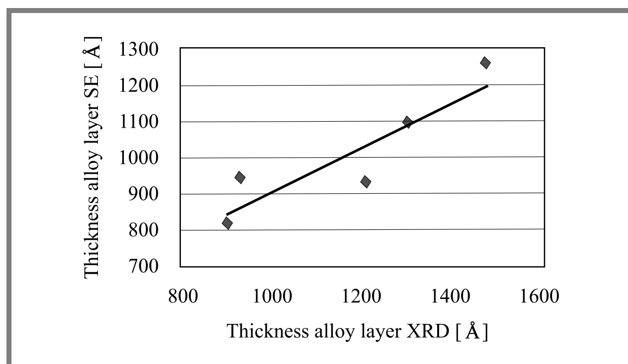


Fig. 19. Layer thickness derived from SE and XRD.

provide a self-consistent assessment of layer content and morphology. If the XRD and SIMS results do not remain correlated during subsequent heat treatments, we look for evidence of carbon clustering using HRTEM. Further correlation is sought with electrical results from test transistors. The ultimate aim of the experiments is to optimise the thermal budget.

9. Conclusions

In this paper we, have summarised our recent work in two areas: the first relates to novel vertical transistor concepts whereby we have demonstrated by simulation and experimentally some potential advantages of our ideas and concepts. The second area concerns explanation and modelling of some anomalous effects in experimental characteristics of SiGe transistors on SOI substrates and a discussion of our strategy for realising SiGeC layers into HBTs.

Acknowledgements

We acknowledge our collaborators at the Universities of Queen's Belfast, Surrey and Imperial College UK. Furthermore, we acknowledge graduate students and research assistants who have contributed to the work: Dr AC Lamb, D. C. Donaghy, H. A. W. El Mubarek, V. D. Kunz, T. Uschino, and all at the Southampton Central Facility for production of experimental devices. Silvaco are thanked for provision of modelling software and EPSRC, and the EU for funding the work.

References

[1] T. Schulz, W. Rosner, L. Risch, A. Korbel, and U. Langmann, "Short channel vertical sidewall MOSFETs", *IEEE TED*, vol. 48, no. 8, pp. 1783–1788, 2001.

[2] V. R. Rao, F. Wittmann, H. Gossner, and I. Eisele, "Hysteresis behaviour in 85-mm channel length n-MOSFET's grown by MBE", *IEEE Trans. Electron. Dev.*, vol. 43, pp. 973–976, 1996.

[3] J. Moers, D. Klaes, A. Tonnesmann, L. Vescan, S. Wickenhauseer, M. Marso, P. Kordos, and H. Luth, "19 GHz vertical Si p-channel MOSFET", *IEEE Electron. Lett.*, vol 35, pp. 239–240, 1999.

[4] J. M. Hergenrother, "The vertical replacement-gate (VRG) MOSFET: a 50 nm MOSFET with lithography-independent gate length", *IEDM Tech. Dig.*, pp. 75–78, 1999.

[5] K. C. Liu, T. Chin, Q. Z. Lui, T. Nakamura, P. Yu, and P. Asbeck, "A deep submicron Si/Sub 1-x/Ge/Sub/Sub x/Si vertical PMOSFET fabricated by Ge ion implantation", *IEEE Electron. Dev. Lett.*, vol. 19, pp. 13–15, 1998.

[6] A. C. Lamb, L. S. Riley, S. Hall, V. D. Kunz, C. H. de Groot, and P. Ashburn, "A 50 nm vertical MOSFET concept incorporating a retrograde channel and a dielectric pocket", in *Proc. ESSDERC*, 2000, pp. 347–350.

[7] D. Donaghy, S. Hall, V. D. Kunz, C. H. de Groot, and P. Ashburn, "Investigating 50 nm channel length vertical MOSFETs containing a dielectric pocket in a circuit environment", in *Proc. ESSDERC*, 2002, pp. 499–503.

[8] M. Jurczak, T. Skotnicki, R. Gwoziecki, M. Paoli, B. Tormen, P. Ribot, D. Dutartre, S. Monfay, and J. Galvier, "Dielectric pockets – a new concept of the junctions for deca-nanometric CMOS devices", *IEEE TED*, vol. 48, no. 8, pp. 1770–1774, 2001.

[9] V. D. Kunz, C. H. de Groot, and P. Ashburn, "Method for local oxidation in vertical semiconductor devices using fillets". UK patent application, no. 0128414.0, 2001.

[10] V. D. Kunz, T. Uchino, C. H. de Groot, P. Ashburn, D. C. Donaghy, S. Hall, Y. Wang, and P. L. F. Hemment, "Reduction of parasitic capacitance in vertical MOSFET's by fillet local oxidation (FILOX)", accepted in *IEEE Trans. Electron. Dev.*

[11] V. D. Kunz, C. H. de Groot, P. Ashburn, and S. Hall, "Gain control in SiGe HBTs by the introduction of germanium into polysilicon emitters", accepted in *IEEE Trans. Electron. Dev.*

[12] J. F. W. Schiz, J. M. Bonar, A. C. Lamb, F. Christiano, P. Ashburn, P. L. F. Hemment, and S. Hall, "Leakage current mechanisms associated with selective epitaxy in SiGe heterojunction bipolar transistors", *IEEE Trans. Electron. Dev.*, vol. 48, no. 11, pp. 2492–2499, 2001.

[13] S. Hall, A. C. Lamb, M. Bain, B. M. Armstrong, H. Gamble, H. A. W. El Mubarek, and P. Ashburn, "SiGe HBTs on bonded wafer substrates", *Microelectron. Eng.*, vol. 59, pp. 449–454, 2001.

[14] M. S. Peter, J. W. Slotboom, D. Terpstra, J. H. Klootwijk, F. van Rijs, and W. B. de Boer, "Base current kink effect in SiGe HBT's', in *Proc. ESSDERC*, 1999, pp. 716–719.

[15] T. H. Ning and D. D. Tang, "Method for determining the emitter and base series resistance of bipolar transistors", *IEEE Trans. Electron. Dev.*, vol. 31, no. 4, pp. 409–413, 1984.

[16] H. S. Gamble, B. M. Armstrong, P. Baine, M. Bain, and D. W. McNeill, "Silicon-on-insulator substrates with buried tungsten silicide layer", *Solid State Electron.*, vol. 45, no. 4, pp. 551–557, 2001.

Stephen Hall received the Ph.D. degree from the University of Liverpool, U.K., in 1987, for work on a new form of integrated injection logic in the GaAs/AlGaAs materials system. He then joined the lecturing staff, University of Liverpool, where he began work on SOI materials characterization and device physics, subsequently obtaining funding to work in these areas on the SIMOX and oxidised porous Si systems. The work was in collaboration with U.K. university and industrial collaborators. Other areas of activity concerned fabrication and electrical assessment of cobalt disilicide Schottky diodes together with silicon-germanium materials characterization and device physics for both bipolar transistor and MOSFET. The SiGe work currently concerns high performance analogue bipolar and new forms of low voltage logic with particular interest in modelling and associated materials characterization. Current SOI work is in the area of low voltage/low power integrated circuits both SiGe HBT and MOS. His other major activity concerns novel design concepts for very deep sub-micron vertical MOSFETs. He is currently head of the Department of Electrical Engineering & Electronics and serves on a number of U.K. national advisory groups.

e-mail: s.hall@liverpool.ac.uk
Department of Electrical Engineering & Electronics
University of Liverpol
Liverpool L69 3GJ, U.K.

Octavian Buiu graduated from University of Bucharest (Romania) with a degree in applied physics; he got his Ph.D. in atomic and molecular physics from "Babes-Boylai" University of Cluj (Romania). After graduation in 1987, he worked in several research institutes in the fields of: nuclear power reactors, electronic devices and microtechnology. Currently, he is a lecturer in the Department of Electrical Engineering & Electronics at the University of Liverpool. He is a member of the Solid State Electronics research group. His research interests include: SiGe and SiGeC alloys for use in HBT architectures, electrical properties of ultrathin silicon dioxide films and high *k* dielectrics, low power microelectronic devices for bio medical applications. He is a member of IEEE and of Institute of Physics, U.K.

e-mail: o.buiu@liverpool.ac.uk
Department of Electrical Engineering & Electronics
University of Liverpol
Liverpool L69 3GJ, U.K.

Peter Ashburn was born in Rotherham, England, in 1950. He received the B.Sc. degree in electrical and electronic engineering, in 1971, and the Ph.D. degree in 1974, both from the University of Leeds. His dissertation topic was an experimental and theoretical study of radiation damage in silicon p-n junctions. In 1974, he joined the technical staff of Philips Research Laboratories and worked initially on ion implanted integrated circuit bipolar transistors, and then on electron lithography for sub-micron integrated circuits. In 1978 he joined the academic staff of the Department of Electronics & Computer Science of the University of Southampton as a lecturer, and currently is the holder of a Personal Chair in Microelectronics. Since taking up a post at University of Southampton, Professor Ashburn has worked on polysilicon emitter bipolar transistors, high-speed bipolar and BiCMOS technologies, gate delay expressions for bipolar circuits and the effects of fluorine in bipolar transistors. His current research interests include SiGe HBTs on SOI, SiGeC devices and ultimate CMOS devices. He has authored and co-authored 170 papers in the technical literature, given invited papers on polysilicon emitters and SiGe heterojunction bipolar transistors and has authored two books on bipolar transistors and SiGe heterojunction bipolar transistors.

e-mail: pa@ecs.soton.ac.uk
Department of Electronics & Computer Science
University of Southampton, Highfield SO17 1BJ, U.K.

Kees de Groot received a M.S. degree in physics in 1994 from the University of Groningen, the Netherlands, and a Ph.D. degree in 1998 from the University of Amsterdam for research carried out at the Philips Research Laboratories in Eindhoven. His Ph.D. thesis was on NdFeB permanent magnets and magnetism in rare-earth intermetallic compounds. From 1998 to 2000, he was a Research Fellow at the Francis Bitter Magnet Lab at the Masschussets Institute of Technology, Cambridge, where he conducted research on spin tunnel junction and photovoltaic materials. Since 2000 he is with the department of Electronics and Computer Science of the University of Southampton, U.K., where he is a tenured Research Lecturer. His current research interest are architecture and materials improvement of field effect transistors, including vertical MOSFET's, high-*k* dielectrics, and tunnel transistors.

e-mail: chdg@ecs.soton.ac.uk
Department of Electronics & Computer Science
University of Southampton, Highfield SO17 1BJ, U.K.

Paper

Modeling SiGe-base HBT using APSYS 2000 – a 2D simulator

Adam Linkowski, Lidia Łukasiak, and Andrzej Jakubowski

Abstract — The paper is devoted to optimization of SiGe-base HBT with respect to operation speed by means of numerical simulation. The influence of design parameters on f_T is studied.

Keywords — base transit time, cut-off frequency, HBT, SiGe-base.

1. Introduction

Rapid progress in microelectronics created conditions for the development and widespread use of simulation techniques in the design of semiconductor devices. These techniques are based on numerical methods of solving a set of transport equations. The advantage of simulations lies in the fact that the number of costly experiments may be significantly reduced, moreover it is possible to analyze completely novel designs. Unfortunately, commercially available simulators are far from ideal. Often they are based on simplified calculations through the omission of certain second- and third-order phenomena.

This work is devoted to the illustration of the possibilities offered by APSYS 2000 (a 2D simulator [1]) to optimize the design of SiGe HBT's with respect to their speed of operation.

2. Results

A cross-section of the simulated transistor is shown in Fig. 1. This device is the both the result of initial optimization and the basis for further work on increasing f_T.

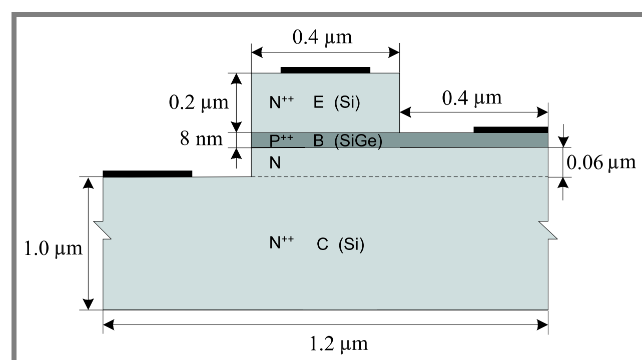


Fig. 1. A cross-section of a SiGe HBT.

The doping profile of the considered transistor is shown in Fig. 2.

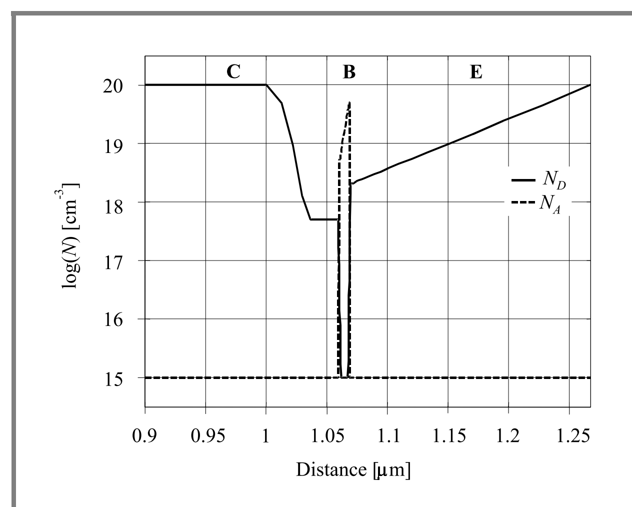


Fig. 2. HBT doping profile.

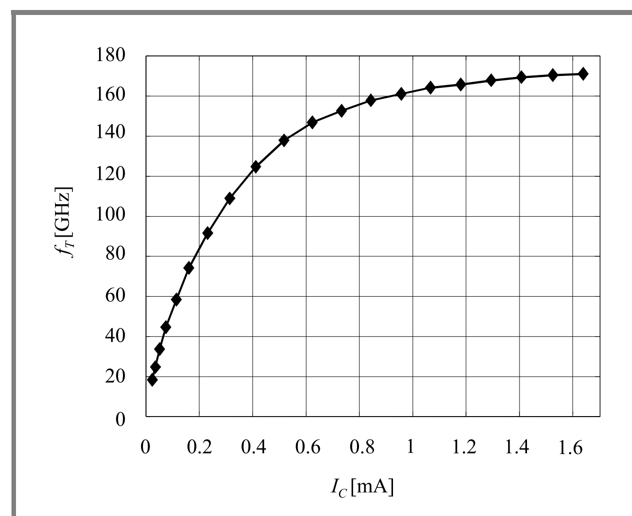


Fig. 3. Cut-off frequency f_T as a function of the collector current I_C.

The maximum cut-off frequency f_T determined from the $f_T = f(I_C)$ dependence shown in Fig. 3 is about 170 GHz. This is by no means the highest possible value, the designer is, however, limited by the fact that the calculations performed by the simulator do not always converge.

Despite this difficulty a designer often has at his disposal more than one way to improve a given parameter.

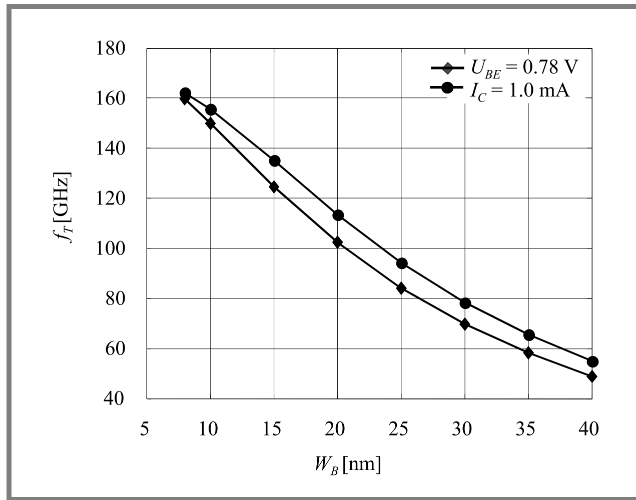


Fig. 4. Cut-off frequency f_T as a function of base width W_B (values of U_{BE} and I_C are kept constant).

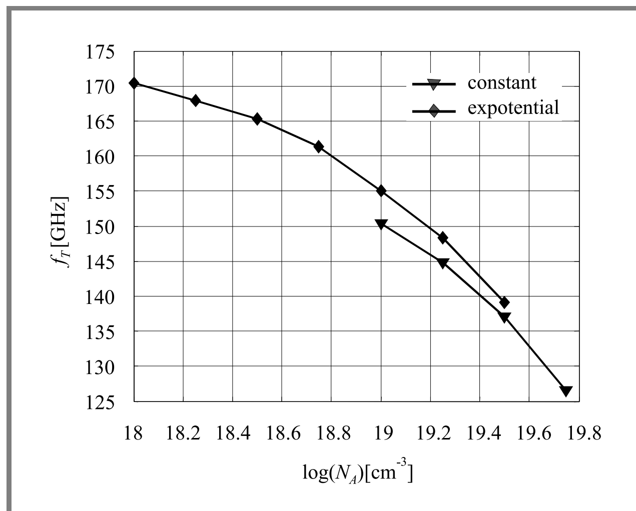


Fig. 5. Cut-off frequency f_T as a function of base doping level (constant and exponential doping profile).

The cut-off frequency is defined mostly by the carrier transit time through the active region, that is the base and the depleted regions of the E-B and B-C junctions. Thus it may be expected that the most important parameter affecting f_T is the effective base width. This is also confirmed in Fig. 4. The effective base width may be changed through fabrication of ever thinner SiGe layers or through the manipulation of the base doping profile, which affects the thickness of the region consumed by the B-C junction. Minimization of the width modulation effect requires relatively high doping concentration in the base, which also decreases the width of the B-C junction. The speed of operation is further improved by the built-in field associated with the exponential doping profile in the base (Fig. 5).

The built-in field in the base may be also obtained using a linear gradation of the Ge content but, as can be seen in Fig. 6, f_T decreases with increasing Ge content (lower carrier mobility, e.g. [2]), while the additional built-in field is of little importance.

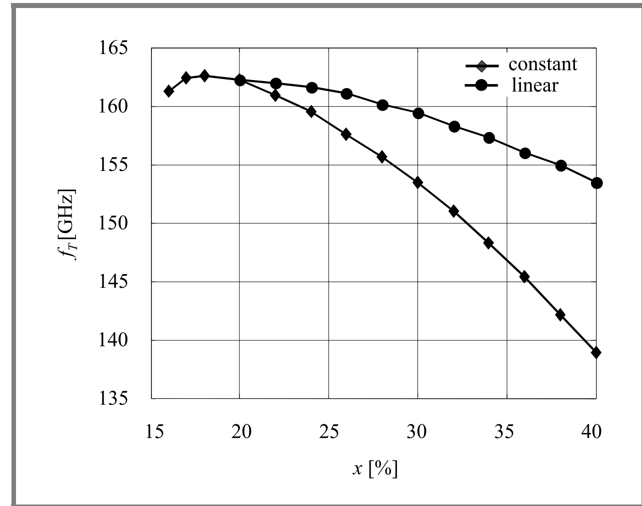


Fig. 6. Cut-off frequency f_T as a function of Ge content x for constant and linear Ge profile.

Junction capacitances are the other two important parameters affecting f_T. The simplest way to reduce them is to decrease the junction area, which may be achieved through the reduction of the dimensions L_1 and L_2, marked in Fig. 1. This, indeed, leads to higher f_T (Fig. 7).

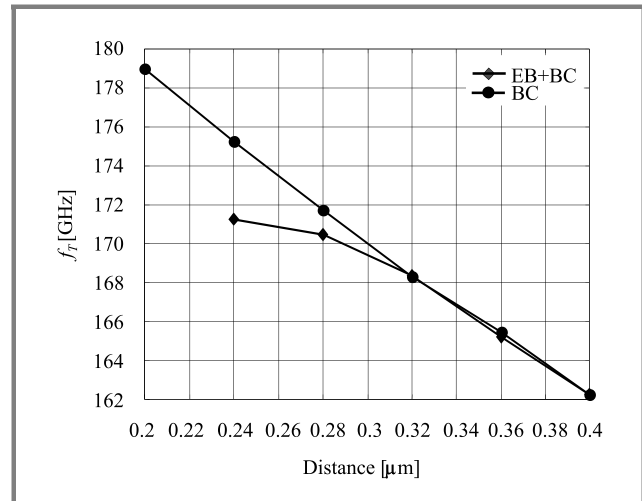


Fig. 7. Cut-off frequency f_T as a function of dimensions L_1 and L_2.

The cut-off frequency may be also improved by means of reducing the emitter width (Fig. 8). Of course, the effects obtained through the reduction of individual dimensions

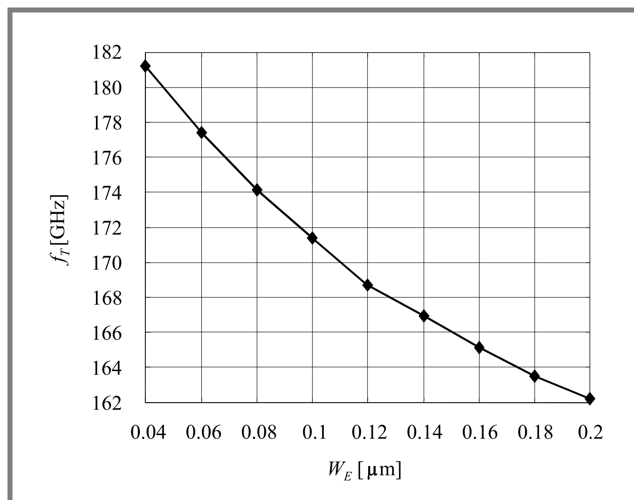


Fig. 8. Cut-off frequency f_T as a function of emitter width W_E.

cumulate. Thus, for example, a transistor with $W_E = 60$ nm, L_1 and $L_2 = 0.24$ μm (the calculations do not converge for L_1 lower than 0.24 μm) achieves f_T of 204 GHz.

3. Conclusions

The maximum f_T obtained by means of simulation using APSYS 2000 was approximately 200 GHz. This is considerably lower than f_T obtained experimentally [3]. The reason for this lies probably in the internal limitations of the simulator used in this work.

Acknowledgment

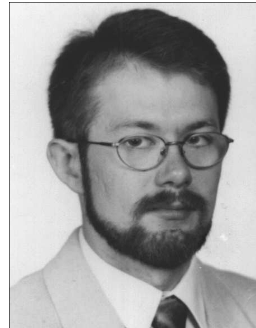
This work was supported by research project no. 8T11B07519 of the State Committee for Scientific Research.

References

[1] "APSYS user's manual, version 3.6", Crosslight Software Inc., Gloucester, Canada, 2000.

[2] V. Palankovski and S. Selberherr, "Critical modeling issues of SiGe semicondutor devices", *J. Telecommun. Inform. Technol.*, no. 1, pp. 15–25, 2004.

[3] J. S. Rieh, B. Jagannathan, H. Chen, K. T. Schonenberg, D. Angell, A. Chinthakindi, J. Florkey, F. Golan, D. Greenberg, S. J. Jeng, M. Khater, F. Pagette, C. Schnabel, P. Smith, A. Stricker, K. Vaed, R. Volant, D. Ahlgren, G. Freeman, K. Stein, and S. Subbanna, "SiGe HBTs with cut-off frequency of 350 GHz", *IEDM Tech. Dig.*, 2002.



Adam Linkowski received M.Sc. degree in 2000 from Warsaw University of Technology. He is currently working towards a Ph.D. His thesis is devoted to the design and modeling of SiGe-base HBT's. His research interests include also processing of radiolocation signals and optimization of aircraft detection and tracking algorithms.

e-mail: alinkows@o2.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Lidia Łukasiak, **Andrzej Jakubowski** – for biography, see this issue, p. 14.

Invited paper

Scattering mechanisms in MOS/SOI devices with ultrathin semiconductor layers

Jakub Walczak and Bogdan Majkusiak

Abstract — Main scattering mechanisms affecting electron transport in MOS/SOI devices are considered within the quantum-mechanical approach. Electron mobility components (i.e., phonon, Coulomb and interface roughness limited mobilities) are calculated for ultrathin symmetrical DG SOI transistor, employing the relaxation time approximation, and the effective electron mobility is obtained showing possible mobility increase relative to the conventional MOSFET in the range of the active semiconductor layer thickness of about 3 nm.

Keywords — ultrathin SOI, scattering mechanisms, electron mobility.

1. Introduction

Silicon devices based on the MOS structure are still the dominating group of semiconductor devices that are subjected to continuous technological progress – in terms of both dimension downscaling and modifying the basic structure of the device and/or developing new ones. The latter tendency is exemplified by silicon-on-insulator (SOI) technology, which is a growing up advantageous alternative to the conventional bulk CMOS technology. In addition to overcoming some of the drawbacks and limitations of the CMOS technology, the SOI concept gives direct entry to nanoelectronics, which is achieved in various forms of ultrathin SOI devices (especially in fully depleted single gate (SG) or double-gate (DG) SOI transistors) by decreasing the thickness of the semiconductor active layer far below 10 nm [1, 2]. Figure 1 shows schematically structural differences between conventional n-channel bulk MOS, SG SOI and DG SOI transistors.

Since transport properties of charge carriers determine their mobility and the overall current – voltage behavior of the device, scattering processes are of special interest when modeling the device operation. Therefore channel carriers mobility models have been continuously developed for decades along with MOS technology growth, with main goals to reflect the actual/measured device performance as well as to investigate the physical phenomena experienced by carriers in the channel area. The physical approach is particularly important since it allows for better understanding of the device operation and is necessary if one wants to make reliable predictions when changing the device operation conditions and/or its configuration, as is the case with the transition from the conventional CMOS to SOI technology.

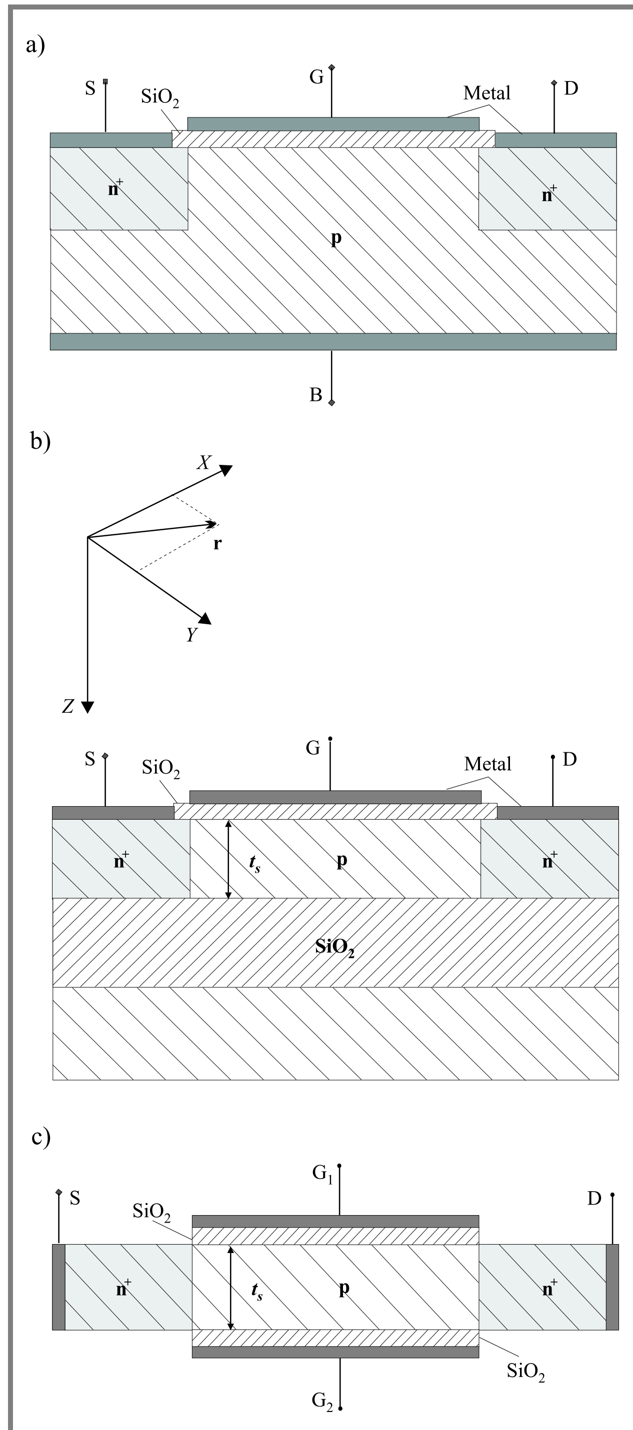


Fig. 1. Schematic cross-sections of (a) bulk MOS, (b) SG SOI and (c) DG SOI transistors.

Because in MOS transistors carriers are spatially confined within a near-surface layer of very small thickness (tens of nanometers), quantum-mechanical effects occur affecting the spatial distribution of the carriers and imposing quantization of their energy [3, 4]. Therefore one of the milestones in MOS transport modeling was the inclusion of the quantum-mechanical phenomena into the device theory after which one could distinguish between "classical" and "quantum-mechanical" descriptions.

Whereas the "classical" approach based on the solution to the Poisson's equation alone was acceptable to a certain extent when modeling the conventional bulk MOS transistors, the quantum-mechanical approach is rigorously imposed in the case of SOI devices with ultrathin active layers and solution to the Schrödinger equation must be included self-consistently with that to Poisson's equation. In the nanometric range of the active layer thickness, quantum effects determine the behavior of the SOI device and substantial differences arise between the bulk MOSFETs and ultrathin SOI transistors with respect to potential and carriers distributions, as well as the dynamics of scattering mechanisms, which additionally depend now on the semiconductor thickness [5–8].

In this paper we present some aspects of scattering processes in MOS/SOI transistors following the quantum-mechanical modeling approach and concentrating on electron mobility in ultrathin n-channel symmetrical double-Al-gate SOI transistors with uniform substrate doping. The main mechanisms, i.e. the phonon, Coulomb and interface roughness scattering, limiting the channel electron mobility, will be discussed briefly, pointing out the most important differences between the cases of bulk NMOSFETs and ultrathin DG SOI devices, as well as showing the influence of the semiconductor thickness variation. The scattering rates will be obtained and analyzed within the perturbation theory after obtaining the self-consistent solution within the effective mass approximation. Next, electron mobilities corresponding to particular scatterers will be obtained utilizing the relaxation time approximation, and eventually the effective electron mobility will be assembled and plotted vs. the transverse effective field in the channel.

2. Two-dimensional electron gas (2DEG) in MOS/DG SOI transistor channel

As already mentioned, the spatial confinement of electrons within a very narrow layer of MOSFET as well as SOI transistor channel area, defined in terms of energy by the potential energy barrier of Si/SiO_2 interface and the conduction band bottom, results in quantization of kinetic energy corresponding to motion in the direction z perpendicular to the surface. The total electron kinetic energy wave vector relationship near the energy band minimum may be approximated as:

$$E(\mathbf{k}) = E_i + \frac{\hbar^2 k_x^2}{2m_x} + \frac{\hbar^2 k_y^2}{2m_y} \tag{1}$$

being a set of energy subbands with discrete minima E_i and quasi-continuous energies of motion in the two remaining directions (x,y) parallel to the surface (hence the 2DEG term). Moreover, the electron spatial distribution within the potential well follows now quantum-mechanical principles and is defined by envelope wave functions $\xi_i(z)$. Assuming that the potential energy $V(z)$ varies in the z direction only, the eigenstates $(E_n, \xi_n(z))$ may be obtained by the solution to the system of Poisson's and the time-independent Schrödinger one-dimensional equations:

$$\frac{d^2\phi}{dz^2} = \frac{e}{\varepsilon_{Si}}\Big(N_A(z) + n(z) - p(z)\Big), \tag{2}$$

$$\left(-\frac{\hbar^2}{2m}\frac{d^2}{dz^2} + V(z)\right)\xi_i(z) = E_i\xi_i(z) \tag{3}$$

along with the equations describing the electron distribution $n(z)$ and the number of electrons N_i in each subband:

$$n(z) = \sum_i N_i\left|\xi_i(z)\right|^2 + \sum_i N_i'\left|\xi_i'(z)\right|^2, \tag{4}$$

$$N_i = \frac{n_v m_d K_B T}{\pi\hbar^2}\ln\left(1+\exp\left(\frac{E_{Fn}-E_i}{K_B T}\right)\right). \tag{5}$$

Because of the conduction band energy minima configuration in **k**-space (Fig. 2), one obtains two series of eigenstates when the $Si\text{-}SiO_2$ interface is the (100) plane – the first one (nonprimed) corresponding to twofold valleys of the effective mass in the quantization direction $m = m_l = 0.916\, m_0$ and the second one (primed) corresponding to fourfold valleys of the effective mass in the quantization direction $m = m_t = 0.19\, m_0$.

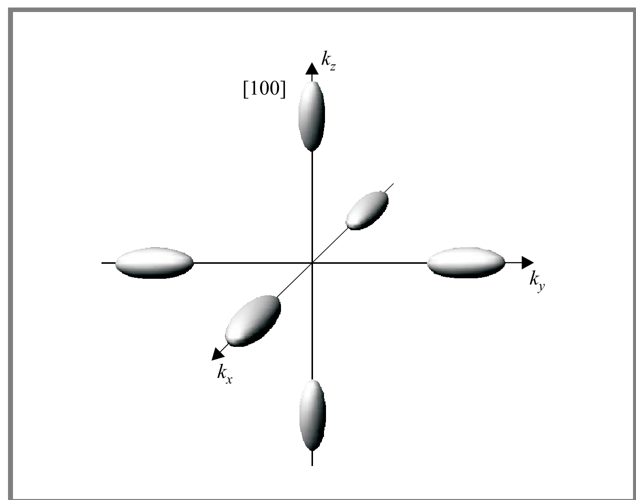


Fig. 2. Constant energy surfaces ($E(\mathbf{k})$ = const) in the vicinity of the silicon conduction band minima.

In this paper we assume infinite potential barrier height at the SiO_2-Si interface, which is a common practice. In other words we do not consider here a possible penetration of electron wave functions into the dielectric layer. Figures 3 and 4 show examples of the distributions of energy levels and eigenfunctions obtained within this approximation for both the conventional bulk NMOS and DG SOI transistors, respectively. The envelope functions have been scaled for illustration purposes, since:

$$\int_0^\infty |\xi_i(z)|^2 dz = 1 . \qquad (6)$$

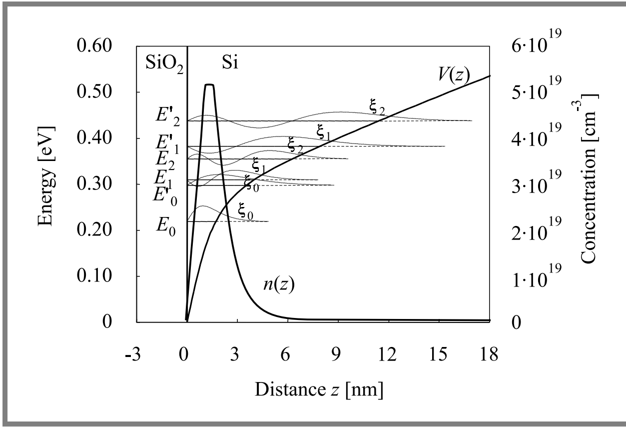


Fig. 3. Energy diagram of a bulk NMOS transistor ($T = 300$ K, $N_A = 10^{17}$ cm^{-3}, $t_{ox} = 3$ nm, $V_{GS} = 1.8$ V, $V_{FB} = -0.922$ V, Al- gate).

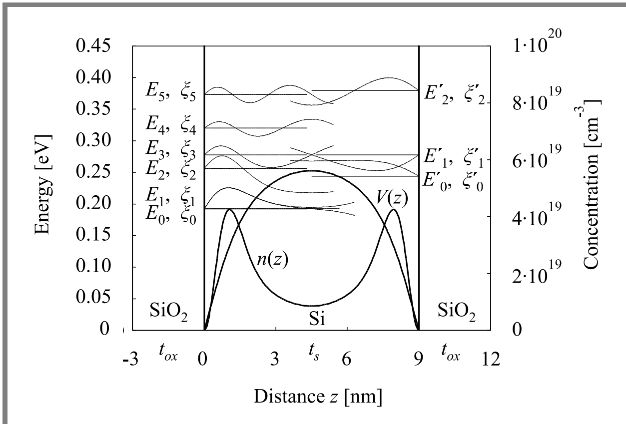


Fig. 4. Energy diagram of a symmetrical DG SOI transistor ($T = 300$ K, $N_A = 10^{17}$ cm^{-3}, $t_{ox} = 3$ nm, $t_{ox} = 9$ nm, $V_{GS} = 1.5$ V, $V_{FB} = -0.922$ V, Al- gates).

The difference between the m_l and m_t effective masses results in a split between the nonprimed and primed series in the energy scale and hence leads to different electron occupation of the corresponding energy subbands. Moreover, the density of states masses differ for both series since $m_{d2} = m_t = 0.19 \cdot m_0$ for the twofold valleys, while $m_{d4} = (m_l \cdot m_t)^{1/2} = 0.417 \cdot m_0$ for the four-fold valleys, and the corresponding average conduction effective masses are: $m_{c2} = 0.19 \cdot m_0$, $m_{c4} = 2/(1/m_l + 1/m_t) \approx 0.315 \cdot m_0$. These differences will strongly affect the resultant transport properties.

The substantial difference between the bulk MOS and the symmetrical DG SOI structure in terms of quantum effects is the channel potential well shape. In the former case the potential well is triangle-like, while in the latter it may be additionally affected by the bias applied to both gates, as well as the thickness of the semiconductor layer. Volume inversion occurs in ultrathin structures of DG SOI, meaning that the inversion charge distribution centroid moves towards the center of the structure.

3. Theoretical approach

In this paper we focus on the main, usually considered scattering mechanisms, i.e. the phonon, Coulomb and interface roughness scattering, since these are believed to influence the device performance to the highest extent at room temperatures [4, 9, 10]. We do not consider here neither high-energy nor ballistic transport problems.

Scattering processes are typically analyzed and modeled within the perturbation theory treating a scattering mechanism as a source of perturbation energy $\mathbf{H}'$ introduced into the regular periodic potential of the lattice, which generates transitions of carriers between different quantum states. According to this approach, the transition probability per unit time between the initial and final states φ_i and φ_f, is given in a general form by the Fermi Golden Rule:

$$P_{fi} = \frac{2\pi}{\hbar} \left| M_{fi} \right|^2 \delta\Big(E_f - \big(E_i \pm \hbar\omega\big)\Big), \qquad (7)$$

where the matrix element M_{fi} depends on a particular scatterer and is defined as:

$$M_{fi} = \int d^3 r \varphi_f^* \mathbf{H}' \varphi_i . \qquad (8)$$

The δ-function in (7) corresponds to the energy conservation law including the energy absorption/emission in the case of inelastic scattering (e.g. with optical phonons), as well as no energy exchange in the case of elastic scattering when $\hbar\omega \approx 0$ (e.g. scattering by low-energy acoustic modes, Coulomb centers and interface roughness). Electron transitions in silicon conduction band may be additionally divided into intravalley scattering processes and intervalley scattering processes, depending on the target valley. Intervalley scattering involves relatively high transition wave vectors and is mainly assigned to optical phonons and high-energy acoustic phonons.

Furthermore, intrasubband and intersubband transitions may be defined in the case of 2DEG in the MOS/SOI type transistor channel. An example of a transition scheme between the initial electron wave vector $\mathbf{k}$, the final wave vector $\mathbf{k}'$ and the scattering wave vector $\mathbf{q}$ is shown in Fig. 5 for an intersubband elastic transition.

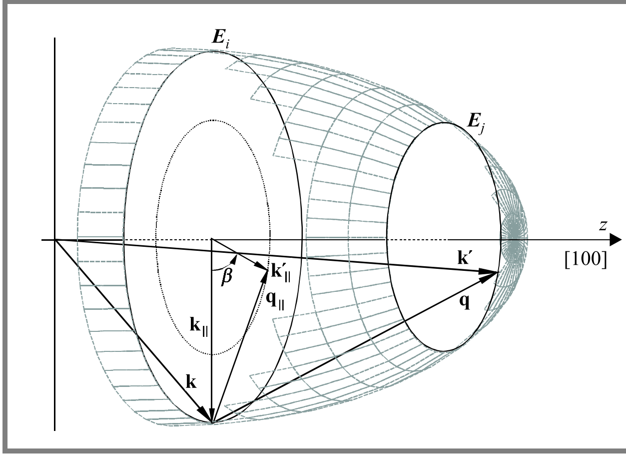


Fig. 5. Illustration of wave vector exchange in 2DEG.

As opposed to a three-dimensional case, the momentum conservation law in the above 2DEG system is fulfilled only for momentum components parallel to the surface being the plane of the considered transport, or in other words, for the components perpendicular to the direction of quantization:

$$\mathbf{k}'_{||} = \mathbf{k}_{||} \pm \mathbf{q}_{||}\,. \tag{9}$$

One can define the energy dependent rate $1/\tau_{ij}(E)$ of scattering by a given scattering mechanism for transition between subbands i and j by adding all possible final states $\mathbf{k}'$. Similarly, the total scattering rate $1/\tau_i(E)$ associated with the ith subband is determined by adding transitions into other subbands including intrasubband transitions ($i = j$) and taking into account intervalley transitions by the application of the corresponding valley degeneration factors.

Having the scattering rates, the Boltzmann equation may be solved and electron mobility may be obtained either by means of a precise full Monte Carlo analysis or within much simpler relaxation time approximation. We follow the second approach believing that in terms of the basic scattering mechanisms it reflects the mobility behavior well enough to recognize the most significant trends. Thus, we calculate ith subband mobility for a given scatterer:

$$\mu_i = \frac{e}{m_{ci}}\langle\tau_i\rangle \tag{10}$$

by averaging the subband relaxation times over the energy distribution:

$$\langle\tau_i\rangle = \frac{\sum\limits_{\mathbf{k}} E(k)\tau_i\big(E(k)\big)\Big(-\frac{\partial f}{\partial E(k)}\Big)}{\sum\limits_{\mathbf{k}} E(k)\Big(-\frac{\partial f}{\partial E(k)}\Big)}. \tag{11}$$

Then, the average rth-scattering mechanism limited mobility is calculated depending on the relative subband occupation:

$$\mu_r = \frac{\sum\limits_i \mu_{r,i} N_i}{N_{inv}}\,, \tag{12}$$

where N_{inv} denotes the inversion charge density. Finally, the total effective electron mobility due to considered scattering mechanisms may be derived by applying Mathiessen's rule:

$$\frac{1}{\mu_{eff}} \approx \frac{1}{\mu_{ph}} + \frac{1}{\mu_{imp}} + \frac{1}{\mu_{sr}} \tag{13}$$

and expressed as a function of effective transverse electric field defined as:

$$E_{eff} = \frac{\int\limits_0^{Z_f} n(z)\big(\partial\phi(z)/\partial z\big)dz}{\int\limits_0^{Z_f} n(z)dz}\,, \tag{14}$$

where Z_f is the integration range equal to $Z_f = \infty$ for MOSFET and $Z_f = t_s/2$ for the symmetrical DG SOI structure.

4. Phonon scattering

Due to lattice vibrations, which occur even at 0 K, phonon scattering is an inherent mechanism limiting electron mobility and is of particular importance at room temperature. This type of scattering is described as collisions between electrons and the vibrating lattice during which the lattice energy quantum $\hbar\omega_{\mathbf{q}}$ (a phonon) and quasi-momentum $\hbar\mathbf{q}$ transfer occurs. Phonons are believed to cause local perturbations of the band structure affecting the electron transport. This effect is conventionally described within the deformation potential approximation. Because of relatively small energies of long wavelength acoustic phonons, electron transitions induced by those phonons are typically treated as elastic and are believed to be mainly intravalley processes.

For the z-direction momentum components, the conservation law is not fulfilled (see Eq. (9)). Instead, the matrix element depends on the form-factor [11]:

$$\mathrm{I}_{ij}(q_z) = \int\limits_0^{\infty} \xi_i(z)e^{iq_z z}\xi_j(z)dz\,. \tag{15}$$

We stay here within frequently used energy equipartition and isotropic approximations applying effective deformation potential Ξ_{eff} value, irrespective of the phonon wave vector direction [12, 13]. In such a case, the following identity is utilized:

$$\frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\Big|\mathrm{I}_{ij}(q_z)\Big|^2 dq_z \equiv \int\limits_0^{\infty}\xi_i(z)^2\xi_j(z)^2 dz = F_{ij} \tag{16}$$

and the ith subband intravalley scattering rate is obtained:

$$\frac{1}{\tau_i(E)} = \frac{K_B T}{\hbar^3\rho u^2}\sum_j \theta\big(E - E_j\big) m_d \Xi_{eff}^2 F_{ij}\,, \tag{17}$$

where ρ is crystal density, u – sound velocity, and $\theta(E)$ is the step function.

Similarly, deformation potential constants D_r and phonon energies E_r are defined for intervalley transitions of different types – g transitions between parallel valleys and f transitions between perpendicular valleys. The corresponding intervalley scattering rate is given by:

$$\frac{1}{\tau_i(E)} = \sum_r \frac{n_v (D_r)^2 m_d}{\hbar \rho E_r} \binom{n_r}{n_r+1} \sum_j \theta\left(E \pm E_r - E_j\right) F_{ij}, \tag{18}$$

where n_r denotes a number of rth type phonons and n_v is the degeneracy factor of the final valleys.

The upper/lower symbols correspond to phonon absorption/emission. The deformation potentials are still a matter of investigation and are taken as empirical or fitting parameters [13, 14]. Here we utilize typical phonon parameters [15] (Table 1).

Table 1
Phonon parameters

Phonon type	Energy E_r [meV]	Deformation constant D_r [10^8 eV/cm]
f_1	19	0.3
f_2	47.5	2.0
f_3	59.1	2.0
g_1	12.1	0.5
g_2	18.6	0.8
g_3	62.2	11.0
Acoustic deformation potential Ξ_{eff} [eV]		12.0

Following the theory outlined above, we have determined phonon limited mobility of electrons for conventional bulk NMOS transistor and DG SOI transistor, varying the active semiconductor layer thickness. The results were then plotted in Fig. 6 as a function of the effective field.

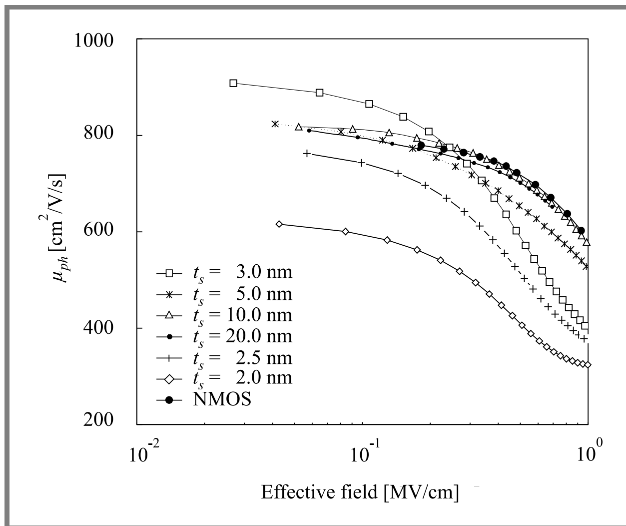


Fig. 6. Phonon mobility for NMOS (black dots) and DG SOI transistor for different values of the semiconductor thickness.

One can observe that the obtained phonon-limited mobility values are generally smaller for DG SOI than for NMOS except for a particular semiconductor thickness, about 3 nm, and in the range of lower fields. This tendency in mobility behavior may be better seen when plotted as a function of the semiconductor thickness (Fig. 7).

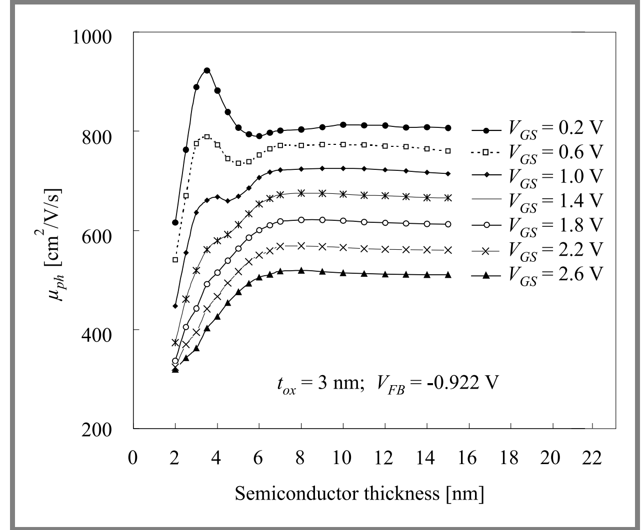


Fig. 7. Phonon-limited mobility of electrons in DG SOI transistor vs. semiconductor thickness t_s for different gate voltage biases ($t_{ox} = 3$ nm).

Here, it can be seen that the calculated phonon-limited mobility reaches a local maximum at the semiconductor thickness of about 3 nm and for low gate voltages. The maximum decreases then gradually with increasing gate voltage and finally vanishes, following the general trend. The phonon-limited mobility decrease for ultrathin SOI relative to the bulk NMOSFET values results from stronger spatial confinement of carriers, which is a direct consequence of the semiconductor thickness decrease for the SOI case. The confinement broadens the electron wave function spectrum in the direction z allowing for the interaction with a wider range of phonons, which increases the coefficient (16). The other observed tendency, i.e. the mobility increase for $t_s \approx 3$ nm, is the effect of relative subband configuration/occupation which, for the case of DG SOI, favors the nonprimed series of lower density-of-states and conduction masses when decreasing the semiconductor thickness and/or the gate voltage bias. Thus, the two counteracting effects determine the resultant phonon-limited electron mobility in ultrathin DG SOI transistor, giving a range of the semiconductor thickness and the gate voltage values where some increase of the mobility may be predicted.

5. Coulomb scattering

This type of scattering is associated with the presence of electrostatic centers affecting the motion of channel carriers. The scatterers are believed to be mainly impurity ions, interface states, charges in the oxide and the gate material.

Scattering potential of Coulomb centers is neutralized to some extent by free carriers, which modify their spatial distribution and screen the bare external scattering potential. Therefore, this screening effect substantially reduces the scattering strength, particularly in the range of higher inversion charge concentrations. Another phenomenon inherently connected with the electrostatic interaction between the scatterers and carriers in a MOSFET channel is the image potential induced due to a sandwich-like structure of the device consisting of several layers with different dielectric constants. Additional effects of Coulomb nature are predicted and investigated when scaling down device dimensions to the nanometric range, such as remote charge scattering, and a change of the image model is suggested [16].

Here we concentrate on the image charge effect and briefly report differences between the image potential influence on the impurity-ion-limited mobility in conventional MOS and DG SOI transistors [17].

Conventionally, semi-infinite SiO_2-Si layers are assumed and only one image charge $Q' = (\varepsilon_{Si} - \varepsilon_{ox})/(\varepsilon_{Si} + \varepsilon_{ox}) \cdot Q$ of a scattering center Q is taken into account in the MOSFET theory, optionally with an image charge due to the finite width of the space charge layer in the semiconductor [9].

In an ultrathin SOI structure, however, additional interfaces are present and the image charge model becomes more complicated. Let us assume semi-infinite oxide layers as a first approximation, which corresponds to the above-mentioned conventional MOS case. Due to the other Si-SiO_2 interface, the total image potential ϕ induced by an impurity ion Q present in the channel at coordinates ($\mathbf{r}_0$, z_0) may be developed into an infinite series:

$$\phi(\mathbf{r}, z) = \frac{Q}{4\pi\varepsilon_{Si}} \left\{ \frac{\widetilde{\varepsilon}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + (z+z_0)^2\right)^{\frac{1}{2}}} + \right.$$
$$+ \sum_{n=1} \left(\frac{\widetilde{\varepsilon}^{2n-1}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + |z-(2nt_s - z_0)|^2\right)^{\frac{1}{2}}} + \right.$$
$$+ \frac{\widetilde{\varepsilon}^{2n}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + |z-(2nt_s + z_0)|^2\right)^{\frac{1}{2}}} +$$
$$+ \frac{\widetilde{\varepsilon}^{2n}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + |z-(z_0 - 2nt_s)|^2\right)^{\frac{1}{2}}} +$$
$$\left.\left. + \frac{\widetilde{\varepsilon}^{2n+1}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + (z+z_0+2nt_s)^2\right)^{\frac{1}{2}}} \right) \right\}, \qquad (19)$$

where $\widetilde{\varepsilon} = (\varepsilon_{Si} - \varepsilon_{ox})/(\varepsilon_{Si} + \varepsilon_{ox})$, which can be illustrated by the series of induced image charges shown in Fig. 8.

As can be seen, one could expect the scattering image potential to be higher for the case of a SOI device than for a conventional MOSFET. The situation becomes however more complex if the device is a truly ultrathin DG SOI device, i.e. if the oxide layers are of ultrathin range. Then oxide layers with the thickness comparable to that of the semiconductor layer, or even thinner, must be taken into account, as well as the gate material influence. We utilize the following solution for the image potential in the Metal-SiO_2-Si system (Fig. 9) [18]:

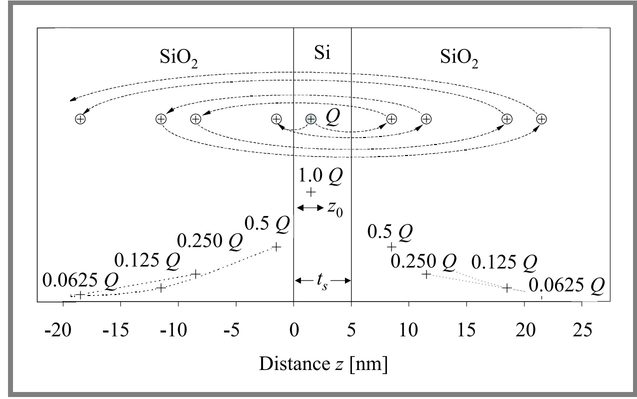


Fig. 8. Image charges seen in SOI device with $t_{ox} = \infty$.

$$\phi(\mathbf{r}, z) = \frac{Q}{4\pi\varepsilon_{Si}} \sum_{n=0} \left(\frac{\widetilde{\varepsilon}^{n+1}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + (z+2nt_{ox}+z_0)^2\right)^{\frac{1}{2}}} + \right.$$
$$\left. - \frac{\widetilde{\varepsilon}^{n}}{\left((\mathbf{r}-\mathbf{r}_0)^2 + (z+2(n+1)t_{ox}+z_0)^2\right)^{\frac{1}{2}}} \right). \qquad (20)$$

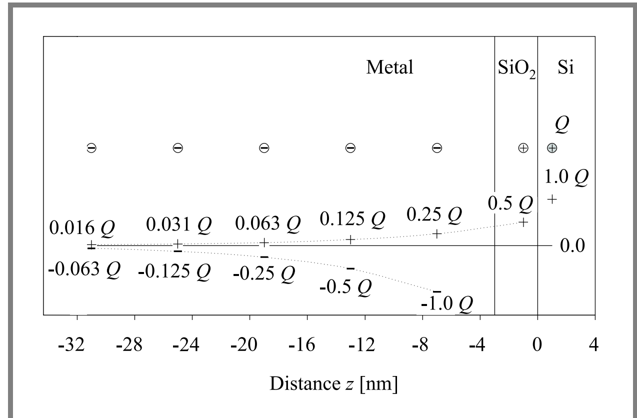


Fig. 9. Image charge affecting the potential felt in Si (Me-SiO_2-Si system).

Then we perform the superposition of the solutions (19) and (20) in order to obtain the image-charge distribution in the Me-SiO_2-Si-SiO_2-Me system, as shown in Fig. 10. Now, both metallic gates affect the resultant image potential, inducing image charge of opposite sign, and thus reducing the total scattering potential.

Next, we derive Coulomb scattering rates and the electron mobility limited by impurity ions (of assumed concentration of $N_A = 10^{17}$ cm^{-3}) located in the channel area, employing the model [4] and including the screening effect, which is also affected by the modified image-charge

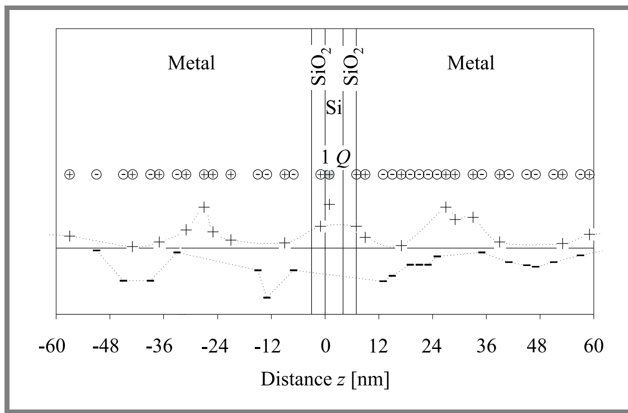


Fig. 10. Image charge affecting the potential felt in Si (Me-SiO_2-Si-SiO_2-Me system).

distribution. The calculated mobilities are plotted versus effective transverse electric field for DG SOI with various values of semiconductor thickness (from 15 nm to 2 nm) and for NMOSFET (Fig. 11), as well as for various gate oxides (Fig. 12). As can easily be seen, the obtained mobilities are much higher for ultrathin DG SOI devices than for the NMOSFET. As already mentioned, this is the consequence of energy level configuration in DG SOI, which favors electron occupation of the non-primed E_0 subband when thinning the semiconductor layer and/or decreasing the gate bias. Moreover, **the surface density of impurity ions is reduced in thinner semiconductor layers**. However, no interface states have been included here, which may change the overall picture due to doubled interface influence.

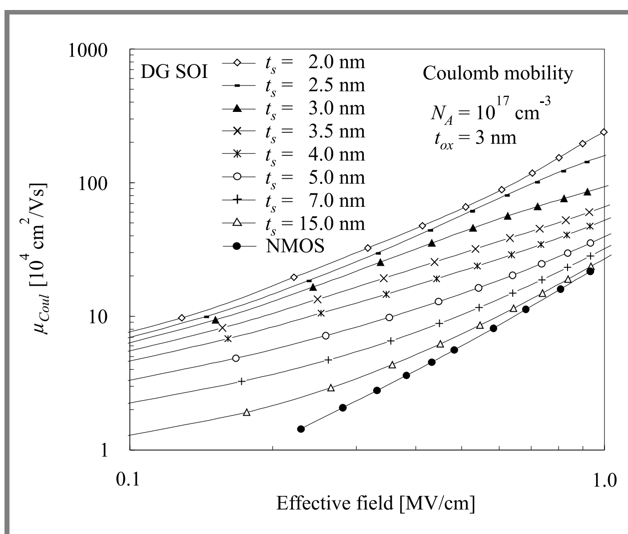


Fig. 11. Coulomb electron mobility for the DG SOI and NMOSFET transistors.

Figure 12 shows the mobility with the gate oxide as a parameter and it presents a comparison of different image charge models in the case of a NMOS transistor and a DG SOI transistor with the silicon layer thickness equal to 2 nm. The "conv." points (black dots) correspond to the conventional model of "one real – one image" charge, typically employed in MOSFETs, while the curve "∞" (for DG SOI only) corresponds to the infinite thickness of both oxide layers, as shown in Fig. 8. The curves for different oxide layers were obtained including the influence of a metallic gate on the image potential. The image

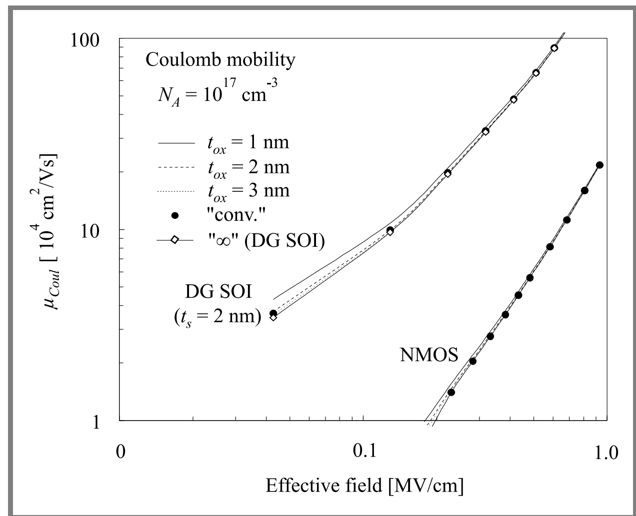


Fig. 12. Coulomb electron mobility for different t_{ox}.

charge reduces the scattering potential introduced by an impurity ion, but it also affects the inversion charge, and thus reduces the screening effect controlled by the inversion charge. Considering the screening, the image charge induced in the gate material replaces the inversion charge. Therefore, only a slight increase of the Coulomb mobility relative to the conventional model is noticeable for oxides below about 3 nm and mainly in the range of lower fields, since for higher fields the inversion charge of high concentration dominates as a screening factor. Here the model taking into account the specificity of the DG SOI structure gives results similar to those obtained with the conventional MOSFET-dedicated model.

6. Interface roughness scattering

Since carrier transport in MOS/SOI devices takes place in the near-surface layer of the semiconductor at the border with the dielectric layer, surely all irregularities of this interface, called surface roughness, must contribute to scattering, thus limiting the effective carrier mobility. This scattering mechanism is believed to dominate in the range of higher fields, where the electrons are forced towards the surface.

Statistical parameters of the Si-SiO_2 interface, which determine its scattering properties, are typically described by the autocovariance function $C(\mathbf{r}) = \langle \Delta(\mathbf{r}')\Delta(\mathbf{r}' - \mathbf{r}) \rangle$ and the corresponding spectrum $S(\mathbf{q}_{||}) = |\Delta(\mathbf{q}_{||})|^2$ of the function $\Delta(x, y)$ (usually unknown) of local deviations from the ideally smooth interface. The form of the auto-covariance function is still a matter of investigation, and usually

Gaussian or exponential forms are employed with some modifications [19, 20]. Here we assume the exponential one:

$$C(\mathbf{r}) = \Delta^2 \exp\left(-\frac{\sqrt{2}r}{\lambda}\right), \tag{21}$$

$$S(\mathbf{q}_{||}) = \pi\Delta^2\lambda^2\left(1+\frac{q_{||}^2\lambda^2}{2}\right)^{-3/2}. \tag{22}$$

In each case the key parameters are Δ – the rms value of the deviations and λ – the correlation length.
The spatial deviations of the interface result in deviations of the potential experienced by electrons in the channel. The corresponding spatial shift of the potential energy:

$$\Delta V(z) = V'(z) - V(z) \tag{23}$$

(see Fig. 13) is treated as the perturbation and usually approximated by:

$$\Delta V(z) \approx \Delta(\mathbf{r})\frac{\partial V(z)}{\partial z}. \tag{24}$$

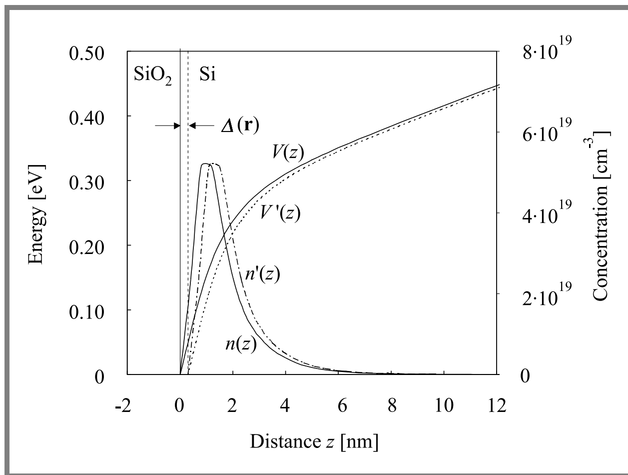


Fig. 13. Illustration of local deviation of the Si-SiO$_2$ interface from the ideal plane.

Therefore, taking into account this "geometrical" effect only, the matrix element is expressed in terms of the transverse electric field $E(z)$:

$$M_{ij} = e\int \xi_i(z)\Delta(\mathbf{r})E(z)\xi_j(z)dz. \tag{25}$$

Assuming that only the lowest subband is occupied (as in the low-temperature regime), one obtains a direct dependence of the matrix element on the effective field:

$$M_{00} = e\Delta(\mathbf{r})E_{eff} \tag{26}$$

which is a commonly employed approximation when modeling the surface roughness mobility.
It has been reported, however, that in the case of ultrathin SOI devices the approximation (24)-(25) cannot be further employed and the perturbation potential should be rather derived based on the definition (23) [21]. On the other hand, a still more complex description of surface roughness scattering includes additional effects due to electron distribution shift and extra electric field due to the deformed interface, as well as the image-charge modification [4]. This "full" matrix element is given by:

$$M_{ij}\left[q_{||}(\beta)\right] = \int_0^\infty dz\Big(\xi_i(z)\frac{\partial V(z)}{\partial z}\xi_j(z)\Big) + \\ +\frac{e^2}{\varepsilon_{\mathrm{Si}}}\frac{\varepsilon_{\mathrm{Si}}-\varepsilon_{ox}}{\varepsilon_{\mathrm{Si}}+\varepsilon_{ox}}\int_0^\infty dz\xi_i(z)\gamma(q_{||})\xi_j(z), \tag{27}$$

where:

$$\gamma(q_{||}) = \left(N_{inv}+N_{depl}\right)e^{-q_{||}z} + \\ +\frac{q_{||}^2}{16\pi}\left(\frac{K_1(q_{||}z)}{q_{||}z} - \frac{(\varepsilon_{\mathrm{Si}}-\varepsilon_{ox})}{2(\varepsilon_{\mathrm{Si}}+\varepsilon_{ox})}K_0(q_{||}z)\right) \tag{28}$$

and K_1, K_0 are the modified Bessel functions.
The surface roughness scattering rate is given by:

$$\frac{1}{\tau_{ij}\big(E(k)\big)} = \frac{2\pi}{\hbar}\frac{1}{2\pi\hbar^2}\theta\big(E(k)-E_j\big)\times \\ \times\int_0^{2\pi}\frac{d\beta}{2\pi}\,m_d\,S\left[q_{||}(\beta)\right]\left|M_{ij}\right|^2(1-\cos\beta). \tag{29}$$

In this work we compare the three models mentioned above employing them in the case of an ultrathin DG SOI transistor. We denote them as: Model 1, corresponding to the simplified form (25) of the matrix element, Model 2, corresponding to the matrix element derived using definition (23), and Model 3, which not only employs the matrix element according to Model 2 but also includes additional

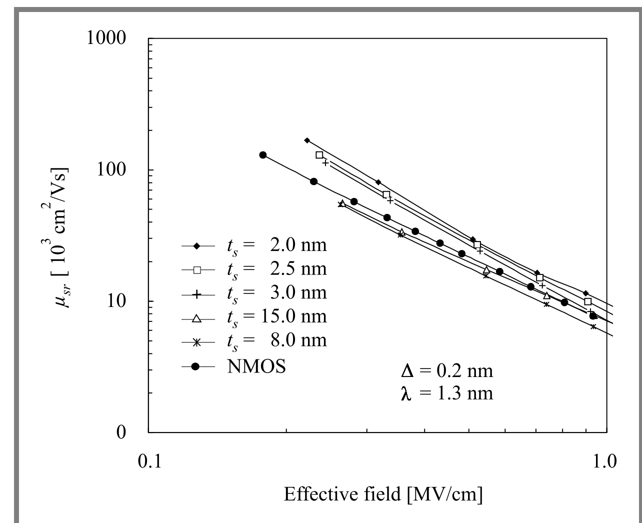


Fig. 14. Surface roughness mobility – Model 1.

effects (expressions (27)-(28)). In all these cases screening by inversion electrons is included. The interface parameters employed are $\Delta = 0.2$ nm and $\lambda = 1.3$ nm. The results

are shown as surface roughness mobility plotted vs. effective field for various semiconductor thicknesses of DG SOI. The results corresponding to a NMOS transistor serve as a reference. Figure 14 shows the mobility obtained using Model 1. As can be seen, the calculated dependencies for the DG SOI transistor follow approximately the $(E_{eff})^{-2}$ slope, as is typically recognized for MOSFETs. However, when analyzing these results in terms of the active semiconductor layer thickness, the observed tendency is somewhat surprising and doubtful. Here, the mobilities increase with the thinning of the semiconductor layer in DG SOI.

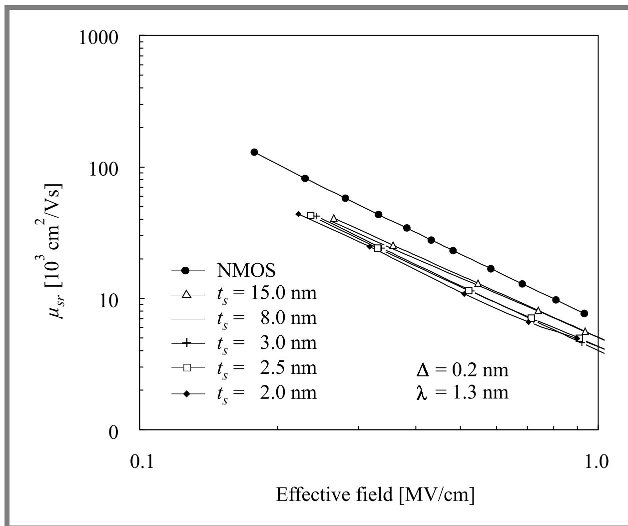


Fig. 15. Surface roughness mobility – Model 2.

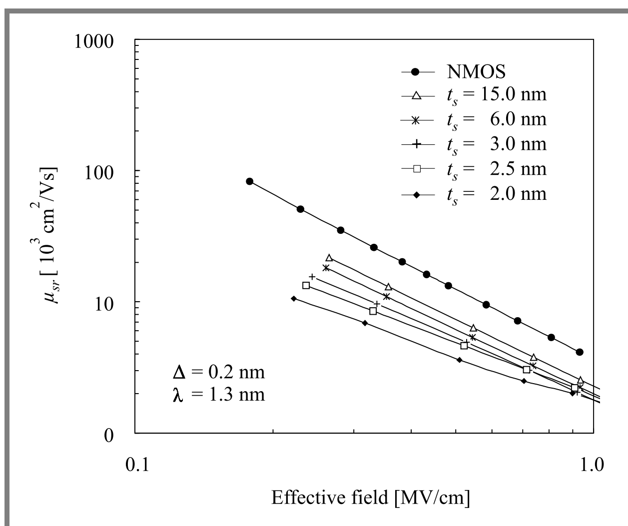


Fig. 16. Surface roughness mobility – Model 3.

Moreover, they exceed the mobility of the NMOS, even in the range of the thinnest layers (2–3 nm). One may intuitionally expect the SOI devices to exhibit a quite opposite trend, since the second SiO_2-Si interface is also rough and surely contributes to the scattering. Additionally, a given roughness height Δ becomes of higher importance when decreasing the semiconductor thickness, particularly in the ultrathin range. Better results, in this regard, are obtained with a more precise Model 2 (Fig. 15), since all the DG SOI points lie below the NMOS curve. However, still no unequivocal and strong enough trend is observed in terms of the semiconductor thickness. Model 3, being of the highest complexity, "improves" the picture accordingly to our expectations – the calculated surface roughness mobility strongly decreases with the semiconductor layer thickness decrease (Fig. 16).

7. Conclusions

Having derived the mobility components, we compose, according to (13), the effective electron mobility for the symmetrical DG SOI and NMOS transistors and plot it vs. the effective field (Fig. 17) and as a function of the semiconductor thickness (Fig. 18).

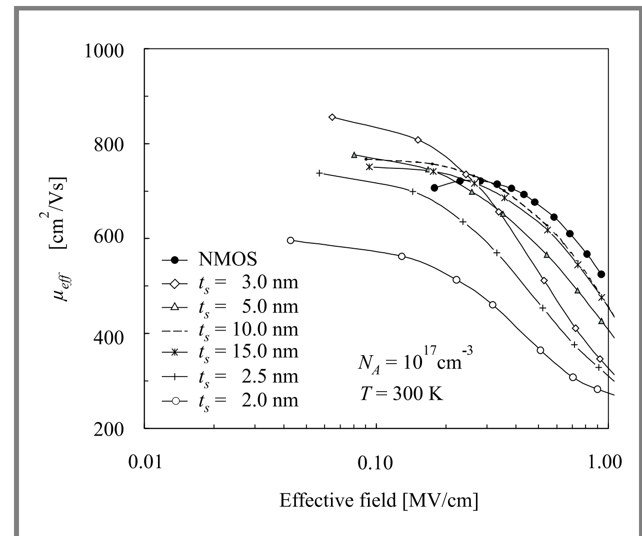


Fig. 17. Effective electron mobility versus transverse effective field.

The effective mobility is mainly determined by the phonon scattering at room temperatures, and it is additionally suppressed by the Coulomb scattering in the range of lower fields and by the interface roughness scattering in the range of higher transverse fields.

The phonon scattering is generally more intense in ultrathin DG SOI structures than in the conventional bulk MOS transistor and its impact increases with the thinning of the active semiconductor layer due to stronger spatial confinement of electrons in narrowed potential bounds.

However, another effect affecting transport properties of electrons may be observed in ultrathin symmetrical DG SOI devices oriented in the (100) plane. The energy separation between the primed and nonprimed eigenstates series increases when the semiconductor layer becomes thinner and/or the gate bias is relaxed, thus favoring the occupation of the lower nonprimed series (the lowest energy level E_0 subband, to be precise) of lower electron mass

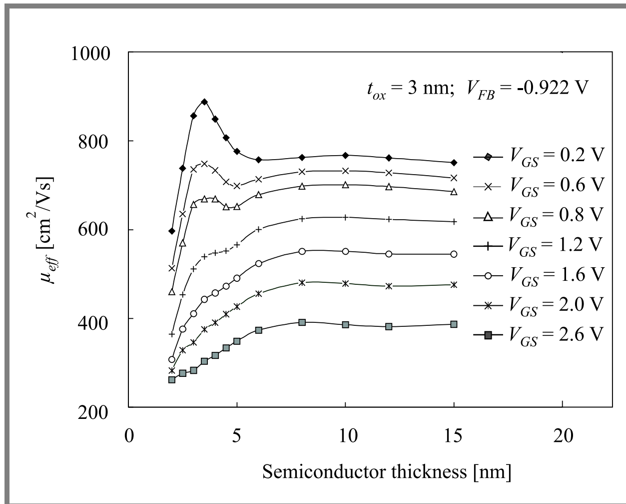


Fig. 18. Effective electron mobility in DG SOI transistor versus semiconductor layer thickness.

in the transport direction. This is illustrated in Fig. 19, where it can be seen that in the ultrathin range of the semiconductor thickness the combined population of the subbands E_0 and E_1 relative to the total inversion charge, reaches nearly 100%. Because of the two counteracting effects mentioned above, a certain increase of the phonon mobility may be expected to occur in the ultrathin range (here, according to our calculations, for the semiconductor thickness of about 3 nm), which has also been reported by other researchers [6, 8].

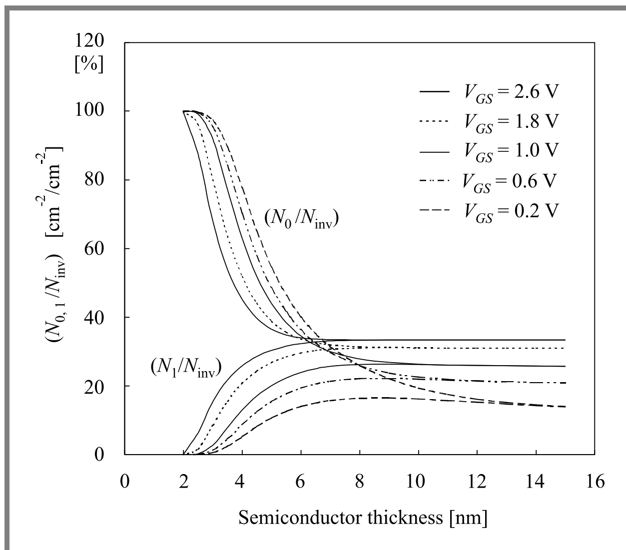


Fig. 19. Relative fraction of electrons in subbands E_0, E_1 (the nonprimed series) in the total inversion charge for DG SOI versus semiconductor thickness ($t_{ox} = 3$ nm).

The discussed "transport-facilitating" effect of the subband configuration plays an important role also in the case of Coulomb scattering, which suppresses the mobility more strongly in the conventional bulk MOSFET than in the DG SOI transistor, as analyzed here for the impurity-ion scattering. On the other hand, ultrathin DG SOI devices suffer much more from the interface roughness scattering than bulk MOSFETs due to the presence of another interface. This effect increases strongly when the semiconductor thickness is scaled down.

In this work we obtained results suggesting a possible increase of electron mobility in the symmetrical DG SOI transistor for the range of semiconductor layer thickness of about 3 nm. This effect is justified by theoretical postulates.

Furthermore, we have shown that conventional MOSFET dedicated models, particularly simplified ones, need to be reconsidered when applied to ultrathin SOI devices, although the modeling results may not be very sensitive to the model employed (as here in the case of image charge models), or, on the contrary, very sensitive (as for the interface roughness modeling).

Therefore, due to structural specificity, which allows for the semiconductor layer thickness variation and the corresponding change of the quantum structure across the channel area, ultrathin SOI and in particular DG SOI devices may be very efficient tools enabling verification of the carrier scattering theory.

References

[1] F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, and T. Elewa, "Double-gate silicon-on-insulator transistor with volume inversion: a new device with greatly enhanced performance", *IEEE Electron Dev. Lett.*, vol. 8, p. 410, 1987.

[2] J. P. Colinge, X. Baie, V. Bayot, and E. Grive, "A silicon-on-insulator quantum wire", *Solid State Electron.*, vol. 39, p. 49, 1996.

[3] F. Stern and W. E. Howard, "Properties of semiconductor surface inversion layers in the electric quantum limit", *Phys. Rev.*, vol. 163, p. 816, 1967.

[4] T. Ando, A. Fowler, and F. Stern, "Electronic properties of two-dimensional systems", *Rev. Mod. Phys.*, vol. 54, p. 437, 1982.

[5] B. Majkusiak, T. Janik, and J. Walczak, "Semiconductor thickness effects in the double-gate SOI MOSFET", *IEEE Trans. Electron Dev.*, vol. 45, p. 1127, 1998.

[6] M. Shoji and S. Horiguchi, "Electronic structures and phonon-limited electron mobility of double-gate silicon-on-insulator Si inversion layers", *J. Appl. Phys.*, vol. 85, p. 2722, 1999.

[7] B. Majkusiak, "Quantum mechanical effects in SOI devices", *Solid State Electron.*, vol. 45, p. 607, 2001.

[8] F. Gámiz, J. B. Roldán, J. A. López-Villanueva, P. Cartujo-Casinello, J. E. Carceller, and P. Cartujo, "Monte Carlo simulation of electron transport in silicon-on-insulator devices", *Electrochem. Soc. Proc.*, vol. 2001-3, p. 157, 2001.

[9] M. V. Fischetti and S. E. Laux, "Monte Carlo study of electron transport in silicon inversion layers", *Phys. Rev. B*, vol. 48, p. 2244, 1993.

[10] F. Gámiz, J. B. Roldán, J. A. López-Villanueva, J. Banqueri, J. E. Carceller, and P. Cartujo, "Universality of electron mobility curves in MOSFETs: a Monte Carlo study", *IEEE Trans. Electron Dev.*, vol. 42, p. 258, 1995.

[11] P. J. Price, "Two-dimensional electron transport in semiconductor layers. I. Phonon scattering", *Ann. Phys.*, vol. 133, p. 217, 1981.

[12] K. Masaki, C. Hamaguchi, K. Taniguchi, and M. Iwase, "Electron mobility in Si inversion layers", *Jpn. J. Appl. Phys.*, vol. 28, p. 1856, 1989.

[13] S. Takagi, J. L. Hoyt, J. J. Welser, and J. F. Gibbons, "Comparative study of phonon limited mobility of two-dimensional electron in strained and unstrained Si metal-oxide-semiconductor field effect transistors", *J. Appl. Phys.*, vol. 80, p. 1567, 1996.

[14] S. E. Laux and M. V. Fischetti, "Issues in modeling small devices", *IEDM Tech. Dig.*, p. 523, 1997.

[15] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials", *Rev. Mod. Phys.*, vol. 55, p. 645, 1983.

[16] I. Kawashima, Y. Kamakura, and K. Taniguchi, "Ensemble Monte Carlo/molecular dynamics simulation of gate remote charge effects in small geometry MOSFETs", *IEDM Tech. Dig.*, p. 113, 2000.

[17] J. Walczak and B. Majkusiak, "Modeling of Coulomb scattering of electrons in ultrathin symmetrical DG SOI transistor", *Electrochem. Soc. Proc.*, vol. 2003-05, p. 355, 2003.

[18] M. Kleefstra and G. C. Herman, "Influence of the image force on the band gap in semiconductors and insulators", *J. Appl. Phys.*, vol. 51, p. 4923, 1980.

[19] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, "Surface roughness at the Si(100)-SiO_2 interface", *Phys. Rev.*, vol. 32, p. 8171, 1985.

[20] A. Pirovano, A. L. Lacaita, G. Zandler, and R. Oberhuber, "Explaining the dependences of the hole and electron mobilities in Si inversion layers", *IEEE Trans. Electron Dev.*, vol. 47, p. 718, 2000.

[21] F. Gámiz, J. B. Roldán, J. A. López-Villanueva, P. Cartujo-Casinello, and J. E. Carceller, "Surface roughness at the Si-SiO_2 interfaces in fully depleted silicon-on-insulator inversion layers", *J. Appl. Phys.*, vol. 86, p. 6854, 1999.

Jakub Walczak born in Mikołów, Poland, in 1971. Graduated from Warsaw University of Technology (WUT) in 1996, received the Ph.D. degree in 2002. Employed at the Institute of Microelectronics and Optoelectronics, Faculty of Electronics and Information Technology, WUT. His research area concentrates on transport and scattering processes in ultrathin semiconductor devices.

e-mail: walczak@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Bogdan Majkusiak was born in Warsaw, Poland, in 1955. He received the M.Sc., Ph.D. and D.Sc. degrees from Warsaw University of Technology (WUT), in 1979, 1985, and 1991, respectively. He received the title of Professor in 2003. He joined the Institute of Microelectronics and Optoelectronics, Faculty of Electronics and Information Technology, Warsaw University of Technology, in 1978, where he has worked as a Professor since 1995. In 1992 he spent 6 months at Carnegie Mellon University. From 1993 to 1996, he was a head of the microelectronics specialty at the Faculty of Electronics and Information Technology, WUT. From 1996 to 1999, he was an Associate Dean responsible for academic affairs, and from 1996 to 2003 a Senior Associate Dean at the Faculty of Electronics and Information Technology, WUT. His current research interest includes physics, modeling, and characterization of the metal-insulator-semiconductor devices with emphasis on problems accompanying ultrathin insulators and quantum-mechanical phenomena, as well as physics of nanoelectronics devices.

e-mail: majkusiak@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Invited paper

High-field current transport and charge trapping in buried oxide of SOI materials under high-field electron injection

Alexei N. Nazarov, Yuri Houk, and Valeriya I. Kilchytska

Abstract — Mechanisms of the charge transfer, the charge trapping, and the generation of positive charge during the high-field electron injection into buried oxide of silicon-on-insulator structures fabricated by different technologies are analyzed based on the data obtained from current-voltage, injection current-time, and capacitance-voltage characteristics together with SIMS data. Electron injection both from the Si film and the Si substrate is considered. The possibility of using the trap-assisted electron tunneling mechanisms to explain the high-field charge transfer through the buried oxides of UNIBOND and SIMOX SOI materials is considered. It is shown that considerable positive charge is accumulated near the buried oxide/substrate interface independently from the direction of the injection (from the film or from the silicon substrate) for UNIBOND and SIMOX SOI structures. Thermal stability of the charge trapped in the buried oxides is studied at temperatures ranging from 20 to 400°C. The theory is compared with the experimental data to find out the mechanisms of the generation of positive charge in UNIBOND and SIMOX buried oxides.

Keywords* — *Fowler-Nordheim current, trap-assisted tunneling, silicon-on-insulator, buried oxide, SIMOX, UNIBOND, anode hole injection, band-to-band impact ionization.

1. Introduction

Silicon-on-insulator (SOI) structures are intensively introduced in large-scale fabrication of reliable integrated circuits and devices. The reliability and stability of such devices strongly depend upon the quality of the buried oxide (BOX) layer and its interfaces. This is especially important for such an attractive application of SOI as high-power devices, where the processes of charge injection into the BOX play a key role in the stability of operation. It is known [1] that the electrical and structural properties of the buried oxide silicon interfaces in SIMOX SOI structures are worse when compared to those of the gate oxide/silicon ones. This, in turn, results in increased charge trapping in the BOX and its degradation during SOI device operation, especially at high voltages. Such behavior often limits promising prospective applications of SOI structures, especially for high-voltage and high-temperature devices.

Thus in the present work we study the nature of electrical conductance, charge trapping and positive charge generation processes in the BOX of SOI substrates made using the two most advanced technologies, namely UNIBOND® and SIMOX. Possible mechanisms of charge injection and positive charge build-up into the BOX of the SIMOX and experimental UNIBOND SOI wafers during high-field electron injection are discussed.

2. Experimental

The experimental UNIBOND SOI structures were fabricated by CEA-LETI using the Smart Cut® process [2]. UNIBOND structures were made on p-type silicon substrates with the doping level of $6.5 \cdot 10^{13}$ cm^{-3}. The thickness of buried oxide and silicon film was 400 and 200 nm, respectively.

The SIMOX SOI structures were formed on p-type Si with the doping of $8 \cdot 10^{14}$ cm^{-3} by means of a single implantation process with the oxygen dose of $1.8 \cdot 10^{18}$ O^+/cm^2, beam energy of 200 keV at 600°C. In the following these SOI structures will be denoted simply as SIMOX ones. Post implantation annealing was performed at 1320°C in an Ar+2%O_2 ambient for 6 hours. The thickness of the BOX and the silicon film was 360 and 200 nm, respectively.

For the purposes of electrical investigation SOI capacitors were formed by mesa isolation. The investigated Al-Si-SiO_2-Si mesa capacitors had variable areas, ranging from 0.95 to $2.75 \cdot 10^{-3}$ cm^2.

Electrical injection into the BOX was performed using constant voltage stress at electric fields in the range of 3–7 MV/cm. Both positive and negative stress voltages were applied to the silicon film, so that electrons were injected from the silicon substrate and the silicon film, respectively. During such injection current vs. time (I-t) measurements were performed. Other electrical characterization has been carried out by means of high-field current-voltage (I-V) and high-frequency (1 MHz) capacitance-voltage (C-V) techniques.

3. Results and discussion

3.1. High-field electron injection

As seen in Fig. 1a, both experimental UNIBOND and SIMOX SOI structures exhibit asymmetry in high-field conduction, which indicates different quality of the BOX/film and the BOX/substrate interfaces. Indeed, in a lot of studies devoted to the investigation of the material quality and defects in SIMOX structures it was shown that the BOX

contains a great number of silicon inclusions [3], crystalline coesite phase of SiO_2 [4, 5], as well as dangling [6] and strained [7] bonds, all located predominantly near the BOX/substrate interface. The presence of these defects may result in worse electrical properties of this interface compared to those of the BOX/film interface. In the case of the experimental UNIBOND structure the bonding of the initial wafers occurs at the bottom BOX interface [2] (i.e. at the BOX/substrate interface). It is thus natural to suggest that its electrical properties are somewhat worse than those of the top BOX interface. In fact, as seen in Fig. 1a, the threshold voltage of high-field electron injection from the substrate is much lower than that from the film for both experimental UNIBOND and SIMOX structures, attesting to the inferior electrical quality of the BOX/substrate interface in both structures.

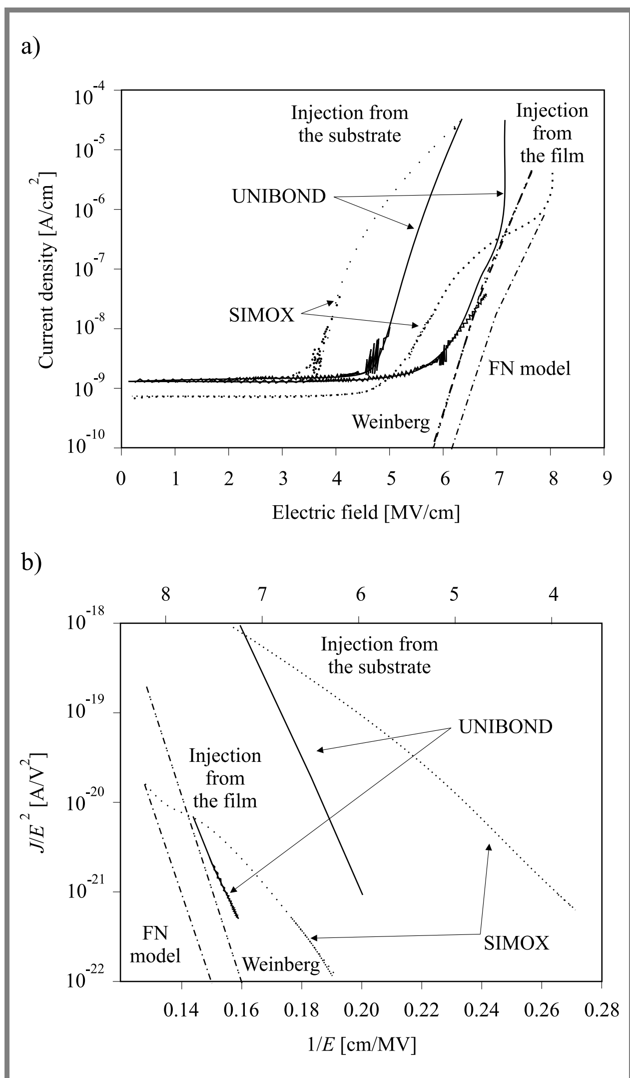


Fig. 1. Current-voltage characteristics of the experimental UNIBOND and SIMOX SOI structures: (a) standard coordinates; (b) Fowler-Nordheim plot. Theoretical curves describing the tunneling current in the BOX of the experimental UNIBOND structure calculated according to the Fowler-Nordheim [9] and Weinberg [10] models are also presented.

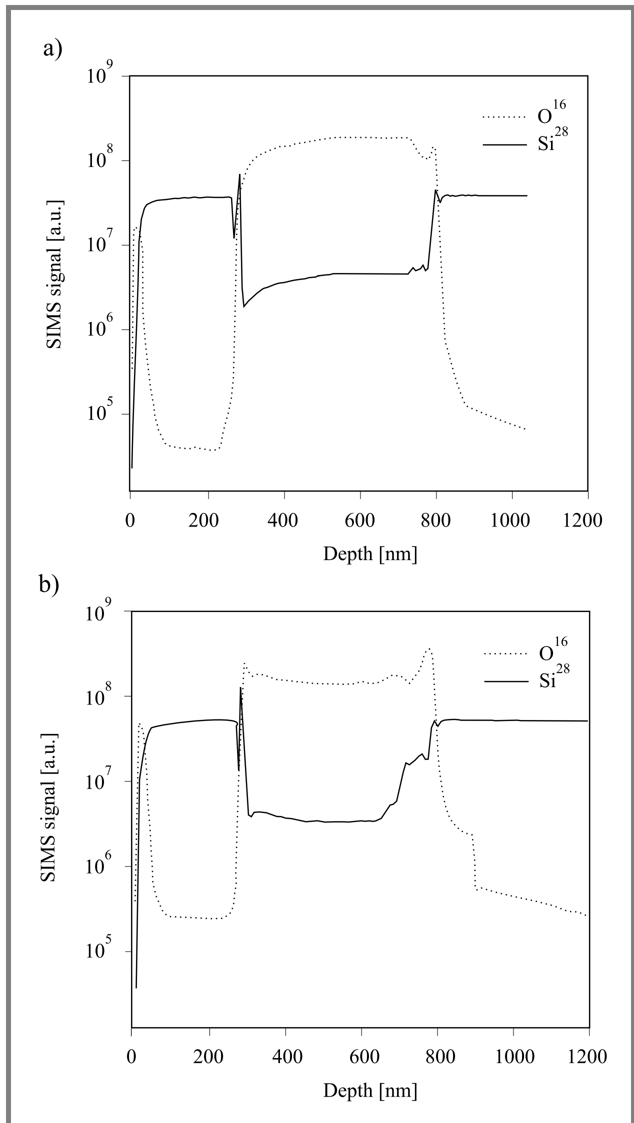


Fig. 2. SIMS investigations of UNIBOND (a) and SIMOX (b) SOI structures.

Indeed, SIMS measurements of atom distribution in the BOX, presented in Fig. 2, indicate the presence of non-stoichiometric composition near the BOX/silicon substrate interface for both experimental UNIBOND and SIMOX structures. It should be noted that in the case of SIMOX structures, considerable increase of Si atom concentration is observed at a distance of more than 100 nm from the BOX/silicon substrate interface (Fig. 2b). Such increase of Si concentration can be associated with silicon inclusions, observed in BOX of similar SIMOX material [3] near the BOX/substrate interface. The nature of such inclusions in the BOX has been proposed in the paper by Afanas'ev *et al.* [8]. In the case of experimental UNIBOND SOI structures, detectable decrease of oxygen atom concentration in the BOX at the distance of 50 nm from the BOX/substrate interface has been found. Since this interface is the bonded one we can suggest that the decrease of

oxygen concentration in this region could be associated with oxygen outdiffusion into the bonded interface. Additionally, it should be noted that the observed non-stoichiometry of the BOX near the BOX/Si substrate interface in the case of UNIBOND material is considerably lower than that of the SIMOX material.

The high-field electric conduction curves of both UNIBOND and SIMOX structures are almost linear in the Fowler-Nordheim (FN) coordinates (Fig. 1b). This attests that the dominant conduction mechanism in both cases is electron tunneling from cathode to the BOX through a triangular potential barrier. This assumption is also supported by a weak temperature dependence of high-field conductivity.

By the classical FN theory the current density, J_{FN}, can be written as [9]:

$$J_{FN} = C \cdot E_c^2 \cdot \exp\left(-\frac{\beta \cdot \Phi^{3/2}}{E_c}\right), \qquad (1)$$

where $C = \frac{e^3 m_0}{8\pi h m^* \Phi}$ and $\beta = \frac{4\sqrt{2m^*}}{3e\hbar}$. E_c is the electric field in the BOX near the cathode, Φ is the potential barrier height, m_0 is the mass of free electron in vacuum, m^* is the effective electron mass in SiO_2 band gap ($m^* = 0.5m_0$), other values have their usual meaning.

Thus the application of the simple FN analysis to the I-V characteristics allows the determination of the effective potential barrier heights for electron emission from the silicon film and the silicon substrate into the BOX. The obtained values of the effective barrier heights (Φ_{FN}) and the threshold electric field of high-field electron injection (E_{th}) for both interfaces of the experimental UNIBOND and SIMOX materials are summarized in Table 1.

Table 1
Effective barrier heights (Φ_{FN}), calculated by using FN analysis, and threshold electric fields (E_{th}) of high-field electron injection into the BOX from the silicon film and the substrate in different SOI structures

Material	Direction of injection	Φ_{FN} [eV]	E_{th} [MV/cm]
UNIBOND	Film	2.4	5.8
	Substrate	2.3	4.8
SIMOX	Film	1.3	4.7
	Substrate	1.2	3.3
Gate oxides	Si substrate [9]	2.9	6.0
	Si substrate [11]	3.1	6.0

In both cases (SIMOX and experimental UNIBOND structures) the effective barrier heights for the BOX/film interface (about 1.3 and 2.4 eV, respectively) are higher than those for the BOX/Si substrate interface (about 1.2 and 2.3 eV, respectively) indicating worse electrical properties of the BOX/substrate interface for both materials. However, for the experimental UNIBOND SOI structures the effective potential barrier heights are considerably higher than those observed in SIMOX, and slightly below the typical values obtained for gate oxides (2.9–3.1 eV) [11, 12]. Moreover, the threshold electric field of high-field electron injection from the Si film into the BOX of the experimental UNIBOND material is just the same as for good thermal oxide (see Table 1). Thus, the electrical properties of the Si film/BOX interface in the experimental UNIBOND structure are just similar to those of the gate oxide/silicon interface.

Reduced values of the potential barrier heights determined by the FN analysis of the I-V characteristics of both the experimental UNIBOND and SIMOX structures indicate that the mechanism of charge transfer through the BOX differs in some way from the ordinary FN mechanism. The observed decrease of the effective potential barrier heights for high-field electron injection from Si into the BOX for the studied SOI structures should be interpreted as an increase of the probability of electrons tunneling from Si into the BOX as compared with that of electrons tunneling from Si into gate oxide for metal-thermal oxide-silicon (MOS) structures. In order to explain such effects in [13] a model, considering the appearance of interfacial asperities, has been developed. However, for the experimental UNIBOND structures this approach is physically unjustified, especially for the Si film/BOX interface, and for the SIMOX structures the use of this approach gives unreasonable large values of asperity separation [14].

We suggest that an increase of the probability of electron tunneling through the potential barrier in the case of the SOI BOX can be connected to the trap-assisted tunneling (TAT) mechanism [15].

TAT model employment. In the case of the TAT mechanism the slope of the high-field I-V characteristic (and therefore the threshold electric field), presented in FN coordinates, is linked to the energetic position of a trap in the oxide band gap ϕ_t, while its intercept is connected both with the energetic position and the trap concentration N_t. Therefore, we can numerically fit TAT current to our experimental data, which allows us to extract trap parameters.

Direct employment of the TAT model yields unreliable values of trap concentrations in the experimental UNIBOND, as well as SIMOX, structures for injection at both interfaces.

Due to a high quality of the silicon film/BOX interface it is necessary in the case of the experimental UNIBOND material to take into account the possibility of direct FN tunneling of electrons from the silicon film into the BOX simultaneously with the trap-assisted tunneling. Indeed, Fig. 1a demonstrates that the experimental I-V characteristics of the electron injection from the silicon film into the BOX of the experimental UNIBOND structure are close enough to the classical FN curve calculated according to (1). Equation (1) was, however, initially derived [9] for the case of electrons tunneling from a metal into vacuum, and is based on the Sommerfeld model of metal. In this model the free electrons are assumed to form a three-

dimensional Fermi gas. In our case of tunneling from semiconductor into the BOX at high fields, however, very strong band banding at the $Si\text{-}SiO_2$ interface is observed, therefore we have a large density of electrons which are confined to a narrow layer at this interface. Such layers have been studied extensively and have been characterized by a two-dimensional model with quantized levels or subbands [16]. Weinberg showed [10] that the tunneling current from the lowest discrete subband may be expressed as follows:

$$J_B(E_0) = \theta \left[1.13 \left(1 + \frac{m_{ox}}{m_s} \frac{E_0}{e\phi_B} \right)^{-1} \right] \sqrt{\frac{e\, m_{ox}}{m_s \phi_B}} \frac{\varepsilon_{ox}^2}{\varepsilon_s} F^2 \times \times \exp\left[-\frac{\beta(\phi_B - E_0)}{F} \right] = C_W F^2 \exp\left[-\frac{\beta(\phi_B - E_0)}{F} \right], \quad (2)$$

where $\theta = 0.626$ is the fraction of carriers in this subband, $E_0 = 0.209$ eV is the energy of this subband relatively to the edge of semiconductor conduction band at the interfaces.
One can see from Fig. 1a that this current is much closer to the experimental current of electrons injected from Si film into the BOX of UNIBOND structure than the classical FN one and its subtraction from the experimental data yields a significantly different curve. If the TAT model is fit to the result of this subtraction, one obtains a reasonable estimation of trap parameters in the BOX of the experimental UNIBOND structures in the vicinity of the Si film/BOX interface: $\phi_t = 1.48$ eV, $N_t = 4.38 \cdot 10^{15}\,\text{cm}^{-3}$.
For the injection from the Si substrate into the BOX of the experimental UNIBOND material we obtain good approximation $\phi_t = 0.82$ eV, $N_t = 2.78 \cdot 10^{18}\,\text{cm}^{-3}$ by assuming a 0.5 eV lowering of the potential barrier at the Si substrate/BOX interface due to worse quality of this interface (see the discussion of the SIMS results). Thus we obtained a lower trap concentration in the vicinity of the BOX/silicon film interface than in the vicinity of the BOX/substrate interface, which is physically justified.
As seen in Fig. 1b, the high-field I-V characteristics of SIMOX structures differ significantly from straight lines when plotted in FN coordinates. It should be noted, however, that it is possible to split them into two straight-line sections (Fig. 3). Thus, we suppose that there are two stages in the process of electron injection into the BOX of SIMOX structure, each with different potential barriers. Indeed, classical FN estimations of the effective barrier heights for these sections (Table 2) show that they decrease with the injection time. This indicates that the barriers are dynamically changing during the injection process. Therefore, we suggest the following model. As it is showed below, a significant positive charge is rapidly created in the BOX of a SIMOX structure exposed to high-voltage stress. Moreover, its centroid moves during the injection towards the Si substrate/BOX interface. This charge affects the electric field in the BOX, and the total field differs considerably from the external one. To employ the TAT model properly we have to use just this total field in the BOX in tunneling region.
As the centroid moves toward the substrate/BOX interface during the injection, the shape of the barrier changes [17].

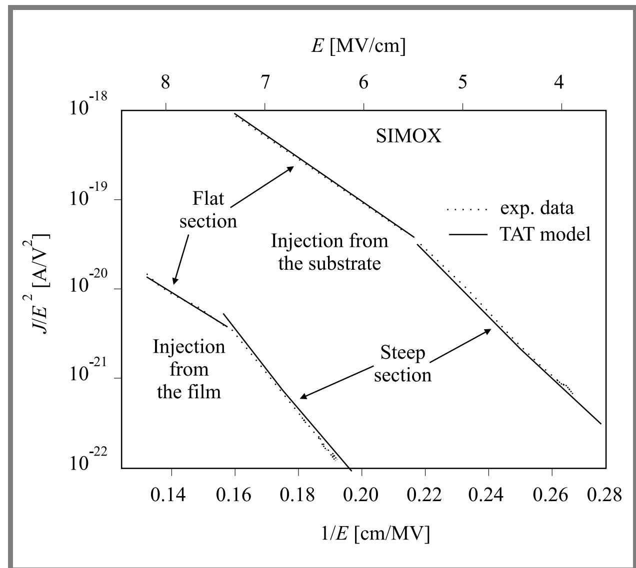


Fig. 3. TAT modeling of high-field injection current into the BOX of SIMOX SOI structure in FN axes.

As a consequence, the tunneling probability changes too, which may cause a distortion of I-V characteristics of SIMOX structures. To estimate trap parameters we fit the first straight (steep) section of the I-V characteristics plotted in FN coordinates, while taking into account the presence of positive charge in the BOX with a fixed centroid. The obtained results are presented in Table 2. The values of

Table 2
Effective barrier heights (Φ_{FN}) for two different sections of high-field I-V characteristics of SIMOX structures and the parameters of traps in the BOX near both interfaces

Interface	Section (see Fig. 3)	Φ_{FN} [eV]	ϕ_t [eV]	N_t [cm^{-3}]
Substrate	Steep	1.4	1.9	$1.6 \cdot 10^{16}$
	Flat	1.1		
Film	Steep	1.6	1.9	$1.4 \cdot 10^{14}$
	Flat	1.0		

trap concentrations are close enough to those of the experimental UNIBOND structure, which does not correlate with other information about the quality of SIMOX BOX. This shows that it is necessary to carry out a more comprehensive analysis of charge injection into such a complicated structure.

3.2. Charge build-up

It is well established that high-field electron injection into thermal oxide [18], BOX of SIMOX [13] and wafer bonding (WB) SOI structures [19] results in electron and hole trapping in these oxides.
High-frequency (HF) C-V measurements along with I-t measurements were used to study the charge trapping

in the buried oxides under high-field electron injection. The C-V technique is more useful in the case of SOI capacitors in comparison with MOS ones, because it allows to obtain more comprehensive information about the charge in the BOX and its changing during the injection due to the possibility of electrical potential control at both interfaces of BOX. Therefore, it is easy to determine the net accumulated charge in the BOX and its centroid [20].

Figures 4a and 5a present the C-V characteristics obtained from the experimental UNIBOND and SIMOX SOI structures before and after electron injection into the BOX from the Si film with injection times up to $2 \cdot 10^4$ s. Similar C-V characteristics are also observed in the case of injection from the Si substrate.

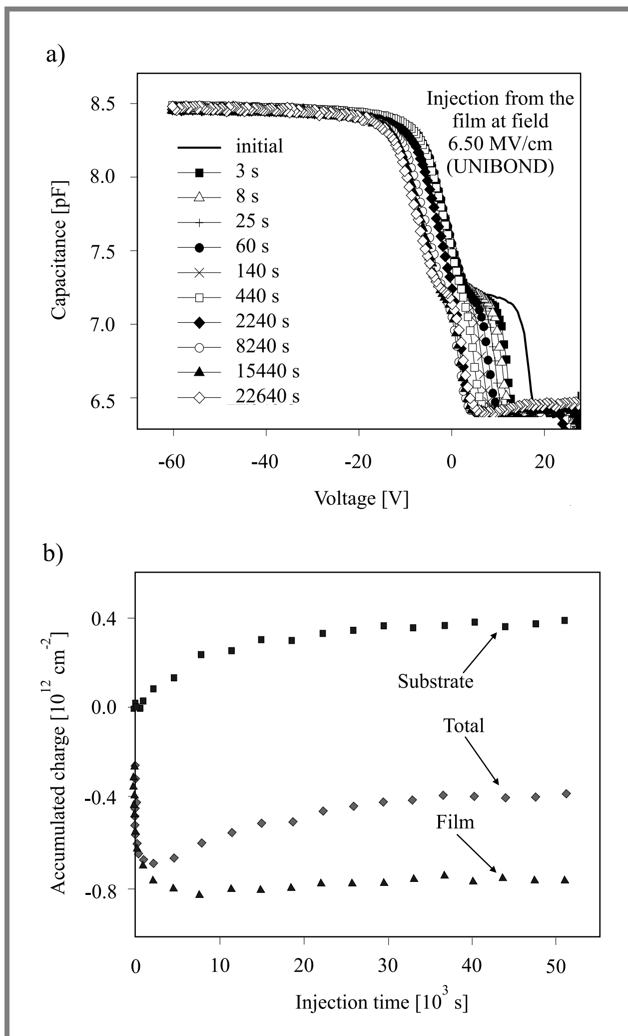


Fig. 4. (a) Capacitance-voltage characteristics of the UNIBOND SOI structure with the electron injection time as a parameter; (b) charge accumulated in the BOX as a function of electron injection time.

The difference in the shape of the C-V plots obtained from the investigated experimental UNIBOND and SIMOX SOI capacitors arises from the difference dopant types of the Si film: n-type in the case of UNIBOND structures and p-type in the case of SIMOX ones. In spite of the different shape, in both cases the C-V plots exhibit two well-defined regions with varying capacitance related to the changing of the electrical potential at the BOX/substrate interface (negative voltages – left part of the plots) and at the BOX/film interface (positive voltages).

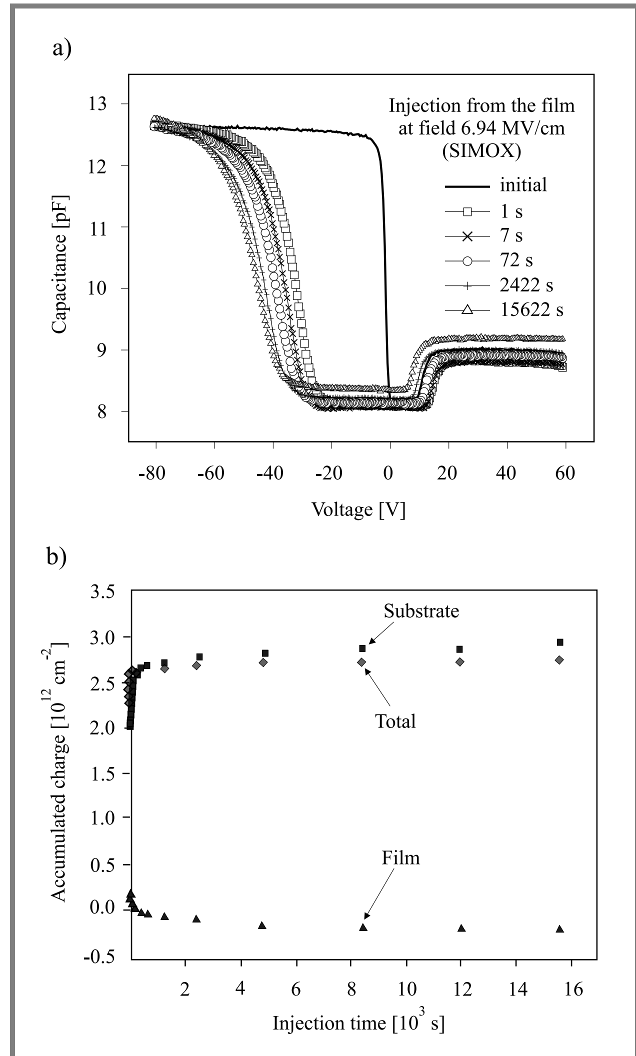


Fig. 5. (a) Capacitance-voltage characteristics of the SIMOX SOI structure with the electron injection time as a parameter; (b) charge accumulated in the BOX as a function of electron injection time.

The C-V characteristics allow the determination of the average doping concentration in Si film both for the experimental UNIBOND and SIMOX SOI structures ($N_d = 2.7 \cdot 10^{16}$ cm^{-3} and $N_a = 2.4 \cdot 10^{16}$ cm^{-3}, respectively). Additionally, it is possible to find the flat-band and mid-gap voltages for the BOX/substrate and the BOX/film interface. The peculiarities of the calculations needed to determine the above mentioned parameters for the n-Si film/BOX/p-Si substrate and the p-Si film/BOX/p-Si substrate SOI structures have been presented elsewhere [20, 21].

Table 3
Parameters of generated positive charge for different oxides (at field $E = 6.25 \cdot 10^6$ V/cm)

Material	Direction of injection	Q_{acc}^{∞} $[10^{11}\mathrm{cm}^{-2}]$	σ^{+} $[10^{-16}\mathrm{cm}^{2}]$	$\eta(0)$ $[10^{-4}]$	T_{ann} [°C]
UNIBOND	Film	2.6	12.0	3.0	400
	Substrate	4.0	4.3	1.7	
SIMOX	Film	24.0	28.0	67.0	> 400
	Substrate	9.6	7.4	7.1	
Gate oxides	Si substrate [22]	78.0	0.25	2.0	150–400
	Si substrate [17]	0.17	5.8	0.09	

It is worth mentioning that the value of the initial positive charge in the BOX (determined from the difference in the flat-band voltages of the BOX/Si film and the BOX/Si substrate interfaces) was about the same for both the experimental UNIBOND and the SIMOX SOI capacitors and was not higher than $1 \cdot 10^{12}\,\mathrm{cm}^{-2}$. This positive charge was initially located mainly near the BOX/silicon film interface. Figures 4a and 5a demonstrate a rather different behavior of the experimental UNIBOND and the SIMOX structures under high-field electron injection, but, as it is described later, both materials are similar in the main trends of the charge build-up. From the C-V characteristics of SOI capacitors the accumulated charge in the BOX during high-field electron injection was calculated using the technique described in [24]. The charge accumulated in the BOX of experimental UNIBOND and SIMOX SOI structures during the electron injection from the Si film, is shown as a function of the injection time in Figs. 4b and 5b, respectively.

In the case of the experimental UNIBOND structure only a decrease of the positive charge at the BOX/film interface due to the trapping of negative charge is observed during first minutes of injection (Fig. 4). Then, some positive charge begins to build up near the BOX/substrate interface. Thus the dynamic characteristics of the charge accumulation in such structures can be divided into two regions, one of which is related to negative charge trapping near the film and the other to positive charge accumulation near the interface with the substrate (see Fig. 4b). It should be noted that for the experimental UNIBOND structures positive charge accumulation occurs predominantly near the BOX/substrate interface during the electron injection from both the film and the substrate. The concentration of this positive charge in the UNIBOND structures at the electric field of 6.25 MV/cm is near $3 \cdot 10$ cm^{-2} for the injection from the film and about $4 \cdot 10^{11}$ cm^{-2} for the injection from the substrate (Table 3).

In the case of SIMOX structures the C-V plots after stress indicate shifts mainly due to the build-up of positive charge which is localized predominantly near the BOX/substrate interface (see Fig. 5a). After 2 seconds of electron injection from the Si film the value of positive charge at the BOX/substrate interface is about $2.2 \cdot 10^{12}\mathrm{cm}^{-2}$ in contrast to about $2 \cdot 10^{11}\mathrm{cm}^{-2}$ at the BOX/silicon film one. Moreover, from Fig. 5b it is clearly seen, that most of the positive charge is created very quickly, during the first seconds, in strong contrast to the experimental UNIBOND material. It should also be pointed out that positive charge accumulates in the BOX predominantly near the interface with the substrate during electron injection from both the Si film and the Si substrate, just as in experimental UNIBOND structures. This indicates that this interface is more imperfect than the BOX/silicon film one for both SOI materials. Then, at the later stages of the injection, the positive charge accumulation in the BOX is also accompanied by the trapping of the negative charge near the film interface, or a shift of the positive charge from the film to the substrate interface (see Fig. 5b). This may be seen in the C-V plots as a small negative shift of the part related to the BOX/film interface. The first phenomenon is probably dominant because the I-*t* characteristics, presented in Fig. 6, show consid-

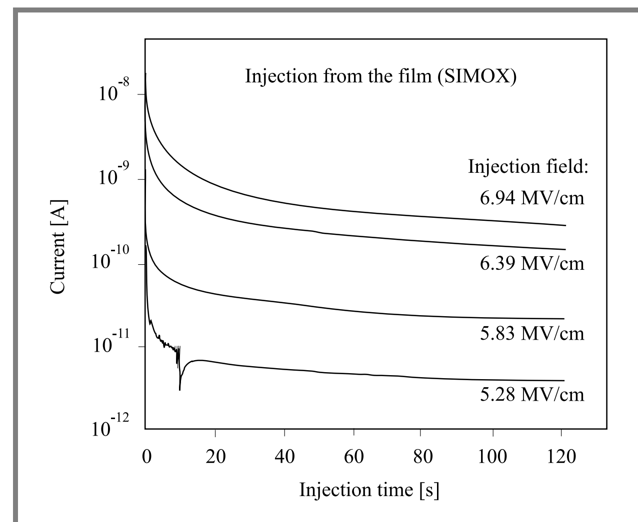


Fig. 6. Current relaxation during the injection of electrons from the film into the SIMOX BOX with the injection field as a parameter.

erable current decrease during the electron injection from the Si film into the BOX. Thus, the positive charge generation and negative charge trapping occur at the same time during the electron injection in the BOX of SIMOX SOI structure, and in this case the concept of centroid charge does not work. However, since these processes occur at

different BOX interfaces they can be studied independently using the BOX/substrate and the BOX/film parts of the C-V characteristics.

In the case of low trapping, electron capture and positive charge accumulation can be expressed in the following way [23]:

$$Q_{acc} = Q_{acc}^{\infty}\left\{1 - \exp[-\sigma_i Q_{inj}(t)]\right\}, \tag{3}$$

where Q_{acc}^{∞} is the maximum concentration of accumulated charge determined from the C-V curves shifts, σ_i is the effective cross section of the process, $Q_{inj}(t)$ is the number of injected electrons and can be calculated by the current integration of I-t characteristics registered during electron injection.

From the Q_{acc} vs. Q_{inj} characteristics it is possible to determine independently both the maximum concentration of trapped or generated charge and the process cross section. The value of the maximum concentration is determined from the saturation level of these characteristics. Once the maximum concentration is known, one can calculate the cross section from the slope of the initial part of such a dependence in a following manner:

$$\eta = \left.\frac{d\,Q_{acc}(t)}{d\,Q_{inj}(t)}\right|_{Q_{inj}\to 0} = Q_{acc}^{\infty}\,\sigma_i\,, \tag{4}$$

where η is the efficiency of charge accumulation.

Since electron trapping occurs mainly near the BOX/film interface, electron trap parameters estimated from the $Q_{acc}^{film} = f(Q_{inj})$ dependence were approximately the same for both the experimental UNIBOND and the SIMOX structures ($Q_{acc,el}^{\infty} \sim (4-6)\cdot 10^{11}\text{cm}^{-2}$ and $\sigma_e \sim (2-4)\cdot 10^{15}\text{cm}^{2}$). Electron traps with similar cross sections ($\sim (2-4)\cdot 10^{15}\text{cm}^{2}$) were found in the gate oxide subjected to different types of radiation [24] and in the BOX of vacuum ultraviolet irradiated UNIBOND and triple implanted SIMOX SOI structures [25]. Traps with similar cross-sections have not been observed in virgin BOX, when silicon film was removed before metal electrode deposition. In our case, we may be possibly dealing with negative charge traps generated during metal gate deposition performed in the laboratory. This process may have resulted in a rather high concentration of electron traps at the top BOX interface. We think, therefore, that these traps are not related to the nature of the buried oxide of the experimental UNIBOND and SIMOX SOI structures.

The parameters of positive charge creation in the BOX of experimental UNIBOND and SIMOX structures for electron injection both from the film and from the substrate obtained using the $Q_{acc} = f(Q_{inj})$ dependence are summarized in Table 3. It can be seen that the efficiency of positive charge creation in the BOX of the SIMOX material is higher than in the UNIBOND structures, which further confirms that the quality of SIMOX buried oxides is worse. Moreover, it should be pointed out that the maximum concentration of the injection-induced positive charge in the SIMOX BOX is higher than in UNIBOND ones by an order of magnitude, which is consistent with the above conclusion, too.

Thermal annealing of generated positive charge. Removal of the generated positive charge in the BOX of experimental UNIBOND and SIMOX SOI structures by thermal heating was studied by annealing the samples both in isochronal and isothermal modes [26]. After each annealing step, the sample was cooled rapidly in order to record the room-temperature C-V characteristics from which the midgap voltage shifts were determined. The anneal-induced removal of the generated positive charge for different samples is depicted in Fig. 7 as a function of the annealing temperature.

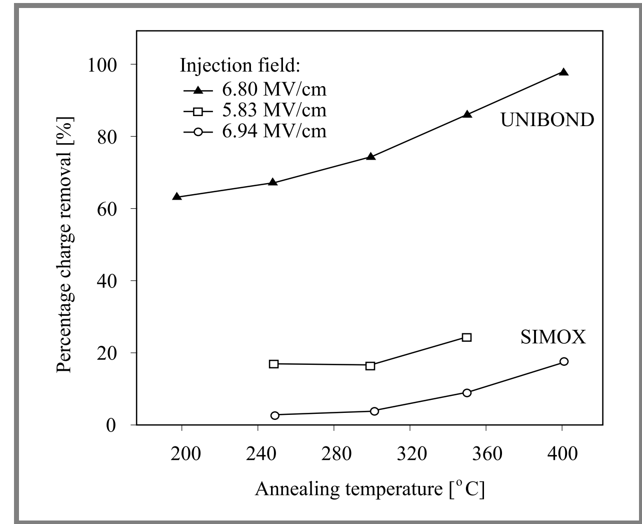


Fig. 7. Annealing of the injection-induced positive charge in the BOX of UNIBOND and SIMOX SOI structures.

From Fig. 7 it is obvious that the positive charge created during the electron injection into the experimental UNIBOND BOX has been almost completely removed after a 15 min. anneal at temperatures up to 400°C. The positive charge generated in a thermal gate oxide is also completely removed when the temperatures of annealing range from 150 to 400°C. In the case of the SIMOX BOX no more than 30% of the total accumulated charge is removed even if the annealing temperature is as high as 400°C. Furthermore, the positive charge generated in the SIMOX BOX at higher electric field is more thermally stable than that generated at lower field.

High thermal stability of positively charged defects created during the electron injection in the SIMOX BOX indicates a more complicated structure of these defects in comparison with that in the thermal gate oxide and in the experimental UNIBOND BOX. Additionally, increased thermal stability of the positive charge generated in the SIMOX BOX at high electric fields suggests the creation of more complicated defects in the BOX than those at lower fields. Thus, it is reasonable to assume that the generation of positive charge during high-field electron injection into the SIMOX BOX occurs simultaneously with the creation of new defects from the precursor sites.

Mechanisms of positive charge generation. First of all, it should be noted that for experimental UNIBOND and SIMOX materials electron injection both from the film and the substrate leads to positive charge generation in the BOX always near the BOX/substrate interface. The obtained results can testify to the absence of trap creation mechanism [27], because in that case the positive charge has to be created always near the anode, independently of the direction of electron injection.

As it can be seen from Table 3, for the **experimental UNIBOND SOI material** the maximum concentration of the generated positive charge is approximately equal and even a little higher for electron injection from the substrate than from the film. In this case it is possible to suggest that **anode hole injection (AHI)** [28] or **band-to-band impact ionization (BTBI)** [29, 30] mechanisms exist in the experimental UNIBOND SOI structure at high electric fields. Detailed overview of positive charge generation mechanisms in the BOX of SOI structures is presented in [14].

Since there was no analytic expression for the hole-generation probability in the case of anode hole injection mechanism in the paper by DiMaria *et al.* [28], we numerically approximated the data presented in Fig. 6 of this paper, and obtained the following result:

$$\alpha_{AHI}(E) = \alpha_{AHI}^{0}\left(E/E_{th}^{AHI} - 1\right)^{6.2}, \qquad (5)$$

where $\alpha_{AHI}^{0} = 1.58 \cdot 10^{-8}$ and $E_{th}^{AHI} = 1.1$ MV/cm, which coincides with the threshold of the exponential field dependence in Fishetti's paper [17].

Holes injected from the anode are swept across the whole oxide film for both directions of FN electron injection. Therefore, regardless of the polarity of electron injection into the oxide and hole trap localization in the oxide, no asymmetry is expected for the AHI mechanism. Thus, in this case $P_{neg}/P_{pos} \approx 1$. Some asymmetry may arise from the variation in the efficiency of hole injection for various contacts.

The existence of AHI in thick oxides has been reported in some papers [17, 28]. AHI modeling, performed by DiMaria *et al.* [28], predicts a sufficiently strong dependence of the probability of hole generation (see Eq. (5)). Therefore, the observed increase of the maximum concentration of the generated positive charge, $Q_{acc,holes}^{\infty}$, with the growth of the electric field may be associated with an increase of the hole-generation probability and a decrease of the recombination cross-section for electrons, $\sigma_r \sim E^{-3}$.

In the case of the BTBI mechanism, the hole-generation probability has been shown to be strongly dependent on the electric field [29]. Moreover, this dependence is considerably different from that for anode hole injection:

$$\alpha_{BTBI}(E) = \alpha_{BTBI}^{0}\left(E/E_{th}^{BTBI} - 1\right)^{4}, \qquad (6)$$

where $\alpha_{BTBI}^{0}(E) = 3.6$ (for thick oxide, $d = 400$ nm), and E_{th}^{BTBI} is the threshold electric field for the BTBI ($E_{th}^{BTBI} = 6.4$ MV/cm).

Since BTBI holes are created in the bulk of thick oxides, in the case of asymmetric hole-trap distribution the asymmetry in hole trapping will also be observed during electron injection with different FN electric field polarity. In the case of the UNIBOND structures, for example, where a non-uniform distribution of hole traps with a maximum near the BOX/silicon substrate interface is observed, $P_{neg}/P_{pos} << 1$ for the BTBI mechanism. Usually, in thick oxides the processes of the BTBI occur simultaneously with the AHI [22].

Experimental results and theoretical curves describing AHI and BTBI as functions of the average electric field in the BOX are shown in Fig. 8. The theoretical curves for AHI and BTBI are extracted from the papers of DiMaria *et al.* [28] and Arnold *et al.* [29], respectively. The efficiency of positive charge generation, η, was obtained directly from the experimental $Q_{acc} = f(Q)_{inj}$ characteristics. The generation cross section, σ^{+}, was calculated from Eq. (4) and the known values of the maximum generated charge presented in Fig. 8a.

The theoretical generation-efficiency curves were calculated from the following expression [14]:

$$\eta(E) = \alpha(E)\sigma_p Q_p = \alpha_0\left(E/E_{th} - 1\right)^{n}\sigma_p Q_p \,. \qquad (7)$$

The curves describing the maximum generated charge, as well as the characteristics of the generation cross-section, were obtained using equations (28a) and (28b) of [14]. The hole-capture cross-section was $\sigma_p = 1 \cdot 10^{-14}$ cm^{-2}. Other parameters, such as α_0, E_{th}, and n for both AHI and BTBI mechanisms were presented above, $Q_p = 5 \cdot 10^{12}$ cm^{-2}.

A comparison of the experimental data with the theoretical curves for AHI and BTBI for all three parameters of positive-charge generation in the case of electron injection from the film shows that the experimental values are significantly higher than the theoretical ones (see Fig. 8). This indicates that the quality of the experimental UNIBOND material investigated in this paper is not at the level of thermal oxides. Simultaneously, one may notice that the qualitative behavior of the experimental data is similar to the theoretical calculations.

Fitting the theoretical AHI characteristics to the experimental data presented in Fig. 8b allows the extraction of the generation efficiency in the form $\eta(E) = \eta_0(E/E_{th}^{\exp} - 1)^{14.3}$, where $\eta_0 = 1.9 \cdot 10^{-20}$ and $E_{th}^{\exp} = 0.4$ MV/cm. From the experimental data for generation cross-section at low fields, presented in Fig. 8c, the recombination cross-section has been obtained as $\sigma_r(E) = \sigma_r^0 E^{-1.5}$, where $\sigma_r^0 = 1.4 \cdot 10^{-14}$ MV$^{1.5}$ cm$^{0.5}$. As it can be seen from Eq. (7), the dependence of the generation efficiency on the electric field is determined by the dependence of hole-generation probability, α, on the electric field. Thus, we can write, that $\alpha_{AHI}(E) = \alpha_{AHI}^{0}(E/E_{th}^{\exp} - 1)^{14.3}$, where the exponential is considerably higher than that extracted from DiMaria's data [28] and present in Eq. (5).

However by employment of both AHI and BTBI mechanisms it is possible to obtain a good agreement between the experimental results and theory (see Fig. 8). In this case the best agreement was obtained using Eqs. (5) and (6)

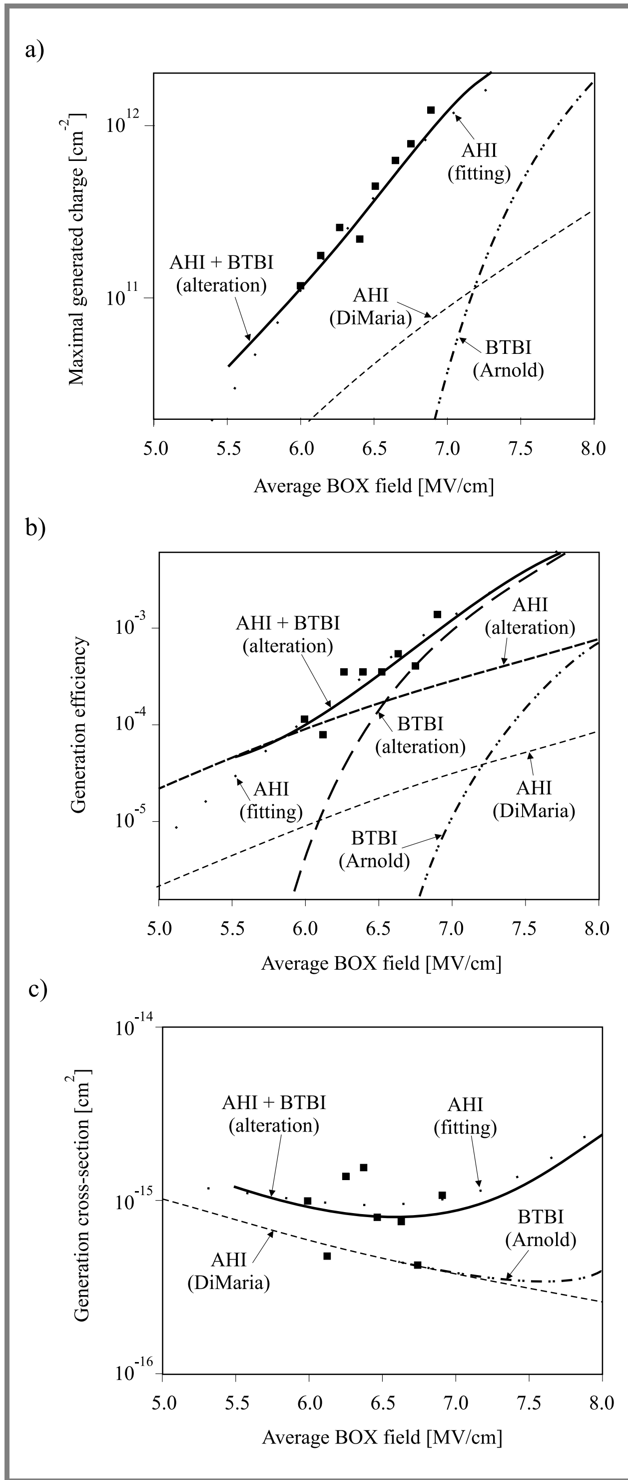


Fig. 8. Parameters of positive charge generation in the BOX of the experimental UNIBOND structures as a function of average BOX electric field: (a) maximum generated charge; (b) generation efficiency; (c) generation cross-section. Theoretical curves for AHI and BTBI mechanisms are plotted after DiMaria *et al.* [28] and Arnold *et al.* [29], respectively.

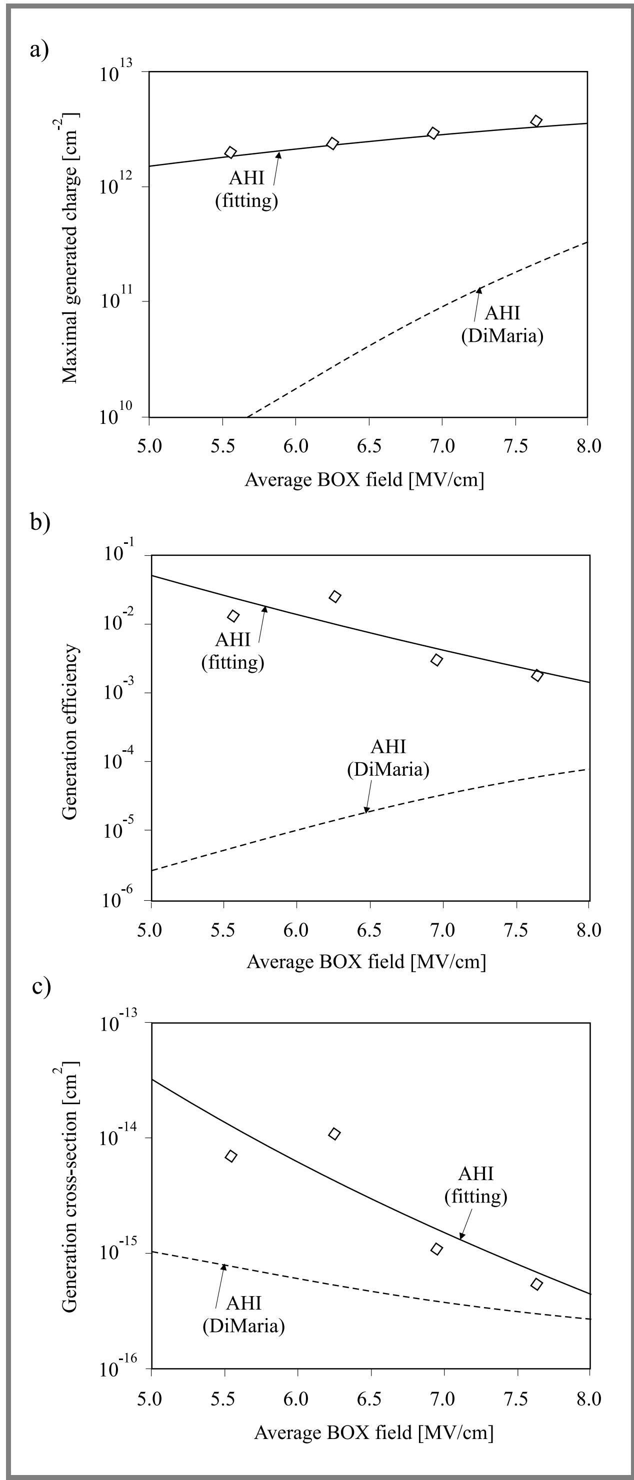


Fig. 9. Parameters of positive charge generation in the BOX of SIMOX structures as a function of the average BOX electric field: (a) maximum generated charge; (b) generation efficiency; (c) generation cross section. Theoretical curves for AHI mechanism are plotted after DiMaria *et al.* [28].

to calculate the hole-generation probability, with the following parameters: $\alpha^0_{AHI} = 1.4 \cdot 10^{-8}$, $E^{AHI}_{th} = 0.8$ MV/cm, $\alpha^0_{BTBI} = 3.6$, $E^{BTBI}_{th} = 5.6$ MV/cm. One can see that in the case of the experimental UNIBOND SOI structure the threshold electric fields for the beginning of AHI and BTBI processes are slightly lower than those reported in [17, 30]. This can be explained by a considerable thickness of the BOX (nearly 400 nm) and

disturbed BOX/substrate interface. Other parameters in the model are similar to those proposed in the paper of DiMaria *et al.* [28]: $\sigma_p = 1.5 \cdot 10^{-14}\,\text{cm}^2$, $\sigma_r^0 = 2 \cdot 10^{-13}\,\text{MV}^3\text{cm}^{-1}$, and $Q_p = 4 \cdot 10^{12}\,\text{cm}^{-2}$.

In the case of the **SIMOX material** the maximum charge generated during the electron injection from the film is much higher than that generated during the electron injection from the substrate (see Table 2). Since for this material the structural quality of the BOX/substrate interface is considerably worse than that of the BOX/film interface, the AHI mechanism can explain the observed behavior. As it is shown in Fig. 9, the fitting of the theoretical AHI curves to the experimental data allows good agreement to be obtained. In this case, however, the hole-capture cross-section σ_p and the electron recombination cross-section σ_r have to be strongly dependent on the electric field ($\sigma_p \sim E^{-14.9}$ and $\sigma_r \sim E^{-9.3}$), which is physically impossible. It should be noted, though, that in the SIMOX material an increase of the electric field results in an increase of the thermal stability of the trapped positive charge (see above). This indicates that a new defect complex is created. Thus, we may conclude that the apparent strong dependence of σ_p and σ_r on the electric field is possibly associated with the transformation of the initial traps or generation of new hole and electron traps for with smaller capture cross-section than that of the initial traps. Trap transformation in the SIMOX BOX during the high field electron injection is a process that is physically probable. Indeed, the amorphous oxide network in the SIMOX BOX is very strained [5, 7] and contains a lot of precursor sites for charge trapping. In the trapping process, carriers can break the weak and strained bonds, creating trapped charge [31, 32].

4. Conclusions

Processes of charge transfer and positive charge generation during high-field electron injection in the buried oxides of experimental UNIBOND and SIMOX SOI structures have been investigated.

It was shown that high-field electron injection into the BOX of both kinds of SOI structures may be described using trap-assisted tunneling mechanism on condition that the peculiarities of each kind of SOI structures are taken into account. In the case of the injection from the Si film into the BOX of the experimental UNIBOND structure it is necessary to consider simultaneously direct tunneling because the quality of this interface is very high. For the SIMOX structure it is necessary to take into account the influence of the positive charge generated in the BOX on the processes of injection. Estimation of the trap concentrations in the vicinity of each interface of both experimental UNIBOND and SIMOX SOI structures obtained from the TAT fitting confirms that the quality of the substrate/BOX interface in comparison with that of the film/BOX interface is worse for both types of structures.

It was demonstrated that in both types of SOI structures positive charge in the BOX is generated mainly near the BOX/substrate interface, which confirms that the quality of this interface is worse. In the BOX of the SOI structures fabricated using the experimental UNIBOND technique the positive charge is generated by anode hole injection and band-to-band impact ionization mechanisms with lower threshold voltages than in the case of thin gate oxides. In the case of the SIMOX material the predominant generation mechanism is determined to be the anode hole injection with simultaneous dynamic transformation of the precursor sites into traps in the strained structure of the BOX under the influence of hot electrons. Positive charge is generated more effectively in the SIMOX SOI structures than in the experimental UNIBOND ones, and this charge is shown to be much more thermally stable in the BOX of SIMOX structures. These facts confirm that the quality of the SIMOX BOX is worse when compared to that of the experimental UNIBOND BOX.

Acknowledgments

This work was supported by STCU project no. 2332. The authors thank Dr. M. Bruel and Dr. B. Aspar (CEA/LETI, Grenoble, France) for providing the experimental UNIBOND SOI wafers. The authors also acknowledge Dr. H. Moriceau and Dr. B. Aspar for helpful discussions.

References

[1] T. Ouisse, S. Cristoloveanu, and G. Borel, "Hot-carrier-induced degradation of the back interface in short-channel silicon-on-insulator MOSFETs", *IEEE Electron. Dev. Lett.*, vol. 12, pp. 290–292, 1991.

[2] M. Bruel, "Smart-Cut® technology: basic mechanisms and applications", in *Perspectives, Science and Technology for Novel Silicon-on-Insulator Devices*, P. L. F. Hemment, V. S. Lysenko, and A. N. Nazarov, Eds. Dordrecht: Kluwer, 2000, pp. 1–15.

[3] S. Kranse, M. Anc, and P. Roitman, "Evaluation and future trends of SIMOX material", *MRS Bull.*, vol. 23, p. 25–29, 1998.

[4] V. V. Afanas'ev, A. Stesmans, and H. E. Twigg, "Epitaxial growth of SiO_2 produced in silicon by oxygen ion implantation", *Phys. Rev. Lett.*, vol. 77, pp. 4206–4209, 1996.

[5] V. V. Afanas'ev, A. Stesmans, A. G. Revesz, and H. L. Hughes, "Structural inhomogenity and silicon enrichment of buried SiO_2 layers formed by oxygen implantation in silicon", *J. Appl. Phys.*, vol. 82, pp. 2184–2199, 1997.

[6] K. Vanhensden and A. Stesmans, "Similiarities between separation by implanted oxygen and bonded and etch-back silicon-on-insulator material as revealed by electron spin resonanse", in *Silicon-on-Insulator Technology and Devices V,* S. Cristoloveanu *et al.*, Eds. Electrochemical Society, 1994, pp. 197–202.

[7] A. G. Revesz and H. L. Hughes, "Properties of the buried oxide layer in SIMOX structures", *Microelectron. Eng.*, vol. 36, pp. 343–350, 1997.

[8] V. V. Afanas'ev, A. Stesmans, A. G. Revesz, and H. L. Hughes, "Mechanism for Si island retention in buried SiO_2 layers formed by oxygen ion implantation", *Appl. Phys. Lett.*, vol. 71, no. 15, pp. 2106–2108, 1997.

[9] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2", *J. Appl. Phys.*, vol. 40, pp. 278–283, 1969.

[10] Z. A. Weinberg, "Tunneling of electrons from Si into thermally grown SiO_2", *Solid State Electron.*, vol. 20, pp. 11–18, 1976.

[11] Z. A. Weinberg and A. Harstein, “Effect of silicon orientation and hydrogen annealing on tunneling from Si into SiO_2”, *J. Appl. Phys.*, vol. 54, pp. 2517–2521, 1983.

[12] C. M. Osburn and E. J. Weitzman, “Electrical conduction and dielectric breakdown in silicon dioxide films on silicon”, *J. Electrochem. Soc.*, vol. 119, pp. 603–609, 1972.

[13] S. Hall and S. P. Wainwright, “On electron conduction and trapping in SIMOX dielectric”, *J. Electrochem. Soc.*, vol. 143, pp. 3354–3358, 1996.

[14] A. N. Nazarov, V. I. Kilchytska, and I. P. Barchuk, “Charge carrier injection and trapping in the buried oxides of SOI structures”, in *Progress in SOI Structures and Devices Operating at Extreme Conditions*, F. Balestra, A. Nazarov, and V. S. Lysenko, Eds. Dordrecht: Kluwer, 2002, pp. 139–158.

[15] C. Svensson and I. Lundström, “Trap-assisted charge injection in MNOS structure”, *J. Appl. Phys.*, vol. 44, pp. 4657–4663, 1973.

[16] F. Stern, “Self-consistent results for *n*-type Si inversion layers”, *Phys. Rev. B*, vol. 5, pp. 4891–4899, 1972.

[17] M. V. Fischetti, “Generation of positive charge in silicon dioxide during avalanche and tunnel electron injection”, *J. Appl. Phys.*, vol. 57, pp. 2860–2878, 1985.

[18] C. Chen, W. L. Wilson, and M. Smayling, “Tunneling induced charge generation in SiO_2 thin film”, *J. Appl. Phys.*, vol. 83, pp. 3898–3905, 1998.

[19] S. Bengtsson, A. Jauhiainen, and O. Engstrom, “Oxide degradation of wafer bonded metal-oxide-semiconductor capacitors following Fowler-Nordheim electron injection”, *J. Electrochem. Soc.*, vol. 139, pp. 2302–2306, 1992.

[20] A. N. Nazarov, V. S. Lysenko, V. A. Gusev, and V. I. Kilchitskaya, “C-V and thermally activated investigations of ZMR SOI meza structures”, in *Silicon-on-Insulator Techology and Devices V*, S. Cristoloveanu *et al.*, Eds. Electrochemical Society, 1994, pp. 236–244.

[21] T. E. Rudenko, A. N. Rudenko, A. N. Nazarov, V. S. Lysenko, and V. I. Kilchitskaya, “Ehlektrofizicheskie svojjstva KNI-struktur: metody issledovanija i ehksperimentalnye rezultaty”, *Mikroelektronika*, vol. 23, no. 6, pp. 18–31, 1994.

[22] M. V. Fischetti, “Model for the generation of positive charge at the Si-SiO_2 interface based on hot-hole injection from anode”, *Phys. Rev. B*, vol. 31, pp. 2099–2113, 1985.

[23] V. V. Afanas’ev and V. K. Adamchuk, “Injection spectroscopy of localized states in thin insulating layers on semiconductor surfaces”, *Prog. Surf. Sci.*, vol. 47, pp. 301–394, 1994.

[24] J. M. Aitken and D. R. Young, “Electron trapping by radiation-induced charge in MOS devices”, *J. Appl. Phys.*, vol. 47, p. 1196, 1976.

[25] V. V. Afanas’ev, A. Stesmans, A. G. Revesz, and H. L. Hughes, “Trap generation in buried oxides of silicon-on-insulator structures by vacuum ultraviolet radiation”, *J. Electrochem. Soc.*, vol. 144, pp. 749–753, 1997.

[26] A. N. Nazarov, V. I. Kilchytska, I. P. Barchuk, A. S. Tkachenko, and S. Ashok, “Radio frequency plasma annealing of positive charge generated by Fowler-Nordheim electron injection in buried oxides in silicon”, *J. Vac. Sci. Technol.*, vol. 18, pp. 1254–1261, 2000.

[27] D. A. Buchanan, A. D. Marwick, D. J. DiMaria, and L. Dori, “Hot-electron-induced hydrogen redistribution and defect generation in metal-oxide-semiconductor capacitors”, *J. Appl. Phys.*, vol. 76, pp. 3595–3608, 1994.

[28] D. J. DiMaria, E. Cartier, and D. A. Buchanan, “Anode hole injection and trapping in silicon dioxide”, *J. Appl. Phys.*, vol. 80, pp. 304–317, 1996.

[29] D. Arnold, E. Cartier, and D. J. DiMaria, “Theory of high-field electron transport and impact ionization in silicon dioxide”, *Phys. Rev. B*, vol. 49, pp. 10278–10297, 1994.

[30] D. J. DiMaria, “Defect production, degradation, and breakdown of silicon dioxide films”, *Solid State Electron.*, vol. 41, pp. 957–965, 1999.

[31] S. K. Lai, “Interface trap generation in silicon dioxide when electrons are captured by trapped holes”, *J. Appl. Phys.*, vol. 54, p. 2540, 1983.

[32] G. W. McPherson, R. B. Khamankar, and A. Shanware, “Complementary model for intrinsic time-dependent dielectric breakdown in SiO_2 dielectrics”, *J. Appl. Phys.*, vol. 88, pp. 5351–5359, 2000.

Alexei N. Nazarov graduated summa *cum laude* from Kiev Polytechnical Institute (Kiev, Ukraine) in 1968, in the field of electronics engineering. He received the Ph.D. and D.Sc. degrees in physics and mathematics from the Institute of Semiconductor Physics (ISP), NASU, Kiev, Ukraine, in 1982 and 1993, respectively. His D.Sc. research was on physics of hydrogen plasma interaction with semiconductor materials and devices. He is now with the Department of Optoelectronics, ISP NASU as leading reseacher and a Professor at National Technical University (KPI) giving courses on “Microelectronics and Nanotechnology”. Dr. Nazarov is a Senior Member of IEEE Society and a Member of Electrochemical Society. He is currently involved in the development of silicon-on-insulator CMOS devices operating in harsh conditions and R&D of Si-based optoelectronic devices.

e-mail: nazarov@lab15.kiev.ua
Institute of Semiconductor Physics
National Academy of Sciences of Ukraine
03028, Prospect Nauky, 45, Kiev, Ukraine

Yuri Houk received the M.Sc. degree in applied physics from National Technical University of Ukraine in 2002. His Master’s project was carried out in the Institute of Semiconductor Physics, National Academy of Sciences of Ukraine (ISP, NASU), and was devoted to charge transport and trapping in buried oxide of SOI structures. Since 2002 he has been participating in the Ph.D. programme of ISP, NASU. The subject of his Ph.D. investigations is charge transport and trapping in the buried oxides of SOI structures and MOSFETs. In 2001 he joined ISP as a researcher assistant. He is involved in the investigation of charge injection and trapping processes in buried oxide of SOI structures and in the investigation of radiation hardness of MOSFETs fabricated on the SOI basis.

e-mail: houk@lab15.kiev.ua
Institute of Semiconductor Physics
National Academy of Sciences of Ukraine
03028, Prospect Nauky, 45, Kiev, Ukraine

Valeriya I. Kilchytska received the M.Sc. degree in solid-state electronics and Ph.D. degree in semiconductors and dielectrics physics, from Kiev University in 1992 and in 1997, respectively. She performed her Ph.D. research in Institute of Semiconductor Physics, Kiev, Ukraine. Her Ph.D. work was devoted to the investigation of electrical and radiation properties of SOI structures, particularly to the improving of the radiation hardness of buried dielectric. From 1997 to 2001 she continued to work in the Institute of Semiconductor Physics, first as researcher assistant and then as researcher. She has been involved in the high-temperature characterization of SOI devices and in the investigation of bias-temperature and injection processes in the buried oxides in SOI structures. She is currently working in the Microelectronics Laboratory of the Université Catholique de Louvain, Louvain-la-Neuve, Belgium. Her current research interest is focused on the room- and high-temperature characterization and simulation of advanced deep sub-micrometer SOI devices.

e-mail: lerka@dice.ucl.ac.be
Institute of Semiconductor Physics
National Academy of Sciences of Ukraine
03028, Prospect Nauky, 45, Kiev, Ukraine

Invited paper

Ultrathin oxynitride films for CMOS technology

Romuald B. Beck and Andrzej Jakubowski

Abstract — In this work, a review of possible methods of oxynitride film formation will be given. These are different combinations of methods applying high-temperature oxidation and nitridation, as well as ion implantation and deposition techniques. The layers obtained using these methods differ, among other aspects in: nitrogen content, its profile across the ultrathin layer, ... etc., which have considerable impact on device properties, such as leakage current, channel mobility, device stability and its reliability. Unlike high-temperature processes, which (understood as a single process step) usually do not allow the control of the nitrogen content at the silicon-oxynitride layer interface, different types of deposition techniques allow certain freedom in this respect. However, deposition techniques have been believed for many years not to be suitable for such a responsible task as the formation of gate dielectrics in MOS devices. Nowadays, this belief seems unjustified. On the contrary, these methods often allow the formation of the layers not only with a uniquely high content of nitrogen but also a very unusual nitrogen profile, both at exceptionally low temperatures. This advantage is invaluable in the times of tight restrictions imposed on the thermal budget (especially for high performance devices). Certain specific features of these methods also allow unique solutions in certain technologies (leading to simplifications of the manufacturing process and/or higher performance and reliability), such as dual gate technology for system-on-chip (SOC) manufacturing.

Keywords — MOS technology, gate stack, ultrathin oxynitride layers, high temperature processing, plasma processing.

1. Introduction

The ITRS Roadmap predicts that the successive technology nodes to be reached in the future still require aggressive scaling down of ultrathin dielectric layers serving as gate dielectrics in MOSFET's. It is commonly believed that as a result silicon dioxide (used since the beginning of the silicon technology era) will have to be abandoned, as the dielectric layers with higher dielectric constants would allow the use of thicker layers while preserving the gate capacitance.

It should be remembered, however, that there is yet another parameter that has to be taken into account while choosing the new gate-dielectric layer, namely the conduction band offset. Its value determines the current flow through the ultrathin dielectric layer (unless direct tunnelling currents dominate the leakage current): the lower the conduction band offset, the higher the current flowing through the MOS system.

A comparison of the materials considered as possible replacements for SiO_2 shows that the materials with dielectric constant tend to have lower values of conduction band offset (Fig. 1), which means that the choice is not obvious and some sort of compromise will have to be reached. Such difficulties will not occur if we stay with silicon dioxide and related compounds. In fact, careful studies prove that scaled-down pure silicon dioxide layers would perform well (in terms of leakage current) in high-performance devices, even down to the 22 nm node [1], and only low power and low standby power devices require serious technological changes.

Combinations of nitride and oxide layers have been successfully used for years in a number of applications; including even combinations of relatively thin but still separate layers, e.g. oxide-nitride (ON) or oxide-nitride-oxide (ONO) DRAM capacitor dielectric, or metal-nitride-oxide-silicon (MNOS) stacks for memory devices. In fact, an oxynitride layer is some sort of a curiosity, as it is not yet known at all why nitrogen atoms get incorporated into the oxide while nitride and oxide phases never coexist in the bulk under equilibrium conditions.

It was, therefore, obvious and tempting to investigate the possibility of the application of nitride and oxynitride layers as ultrathin gate dielectric in MOS devices.

In this thickness regime, no technology allows obtaining pure silicon nitride. In other words, some form of silicon oxide is always present. Thus, hereafter, we will use the terms "oxynitride layers" or SiO_xN_y for all ultrathin dielectrics discussed in this paper, keeping in mind that they may contain different amounts of silicon nitride.

The most important advantages expected of the application of oxynitride layers are as follows:

- much experience gained in the nitride/oxynitride processing during the years;
- compatibility of the material and its formation technologies with the standard silicon technology;
- high temperature stability;
- many possible techniques to choose from.

The aim of this work is to review the methods of the ultrathin oxynitride formation investigated so far and discuss their characteristic features, advantages and, therefore, the chances for application to ICs technology.

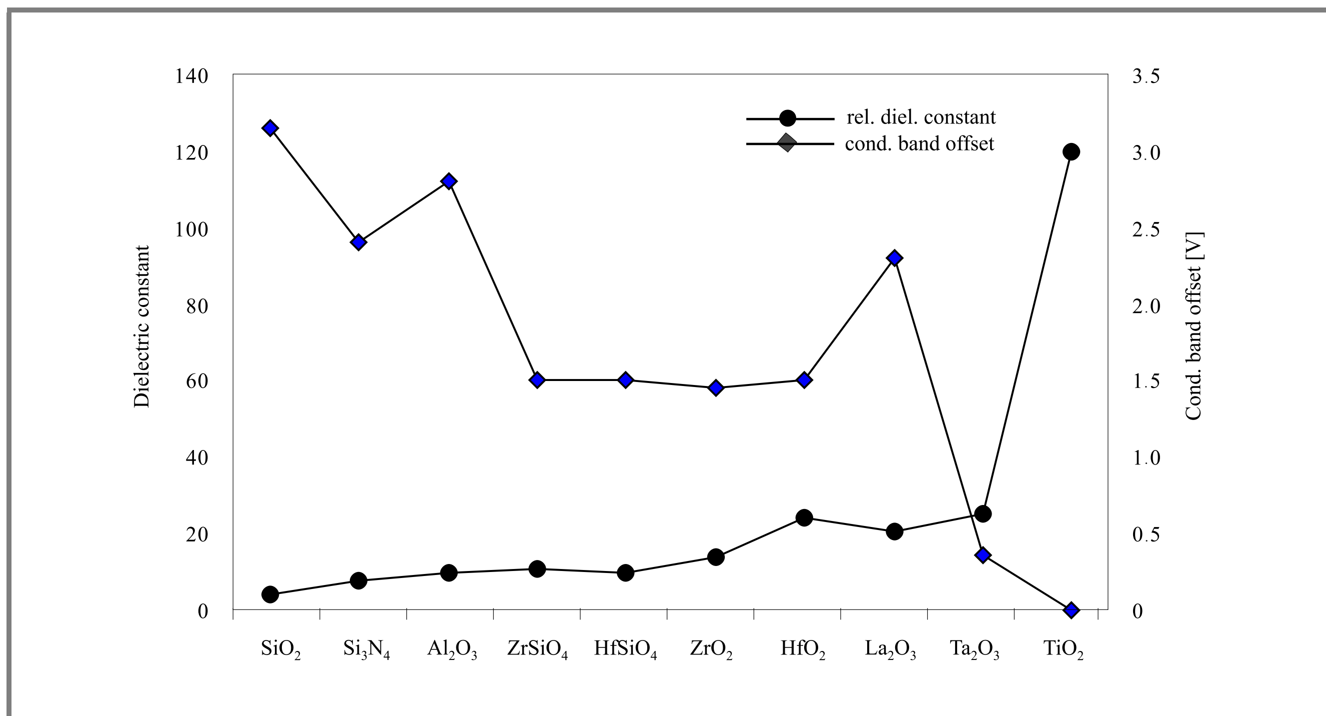


Fig. 1. Dielectric constant and conduction band offset of different dielectric materials.

2. "Ideal" ultrathin oxynitride layer

In most cases, the application of materials different than SiO_2 results in the deterioration of the silicon-insulator interface properties. The situation is, however, different in the case of the introduction of nitrogen into SiO_2 layers. It has been proved that the incorporation of some nitrogen into the silicon oxide gate dielectric can be beneficial for the technology and MOSFET device performance. Thus, ideal ultrathin oxynitride gate dielectric should exhibit:

1) nitrogen pile-up near or at the oxide-silicon interface – in order to improve the hot-electron immunity,

2) even higher nitrogen pile-up at the oxide-poly silicon interface – to suppress boron penetration from the poly-Si gate and to enhance device reliability.

It was also experimentally confirmed that the performance of MOS devices depends strongly on the concentration and the distribution of the nitrogen throughout the oxynitride layer (e.g. [2–5]).
Unfortunately, too much nitrogen at the interface is detrimental for MOSFETs' performance, as excessive nitrogen at the oxide-silicon interface may lead to a reduction of the peak carrier mobility in the channel and allow boron accumulation in the oxide, leading even to device instabilities. Hence, many teams seem to prefer believing in the rule: "better no nitride pile-up at the interface, than too much".

3. Oxynitride ultrathin layers formation

The list of the methods that could possibly be used to form ultrathin oxynitride layers is long. There are high-temperature methods, deposition methods (CVD or PVD) or even combinations of the two. Oxynitride layers may be produced in single process step but more often they are built up in two or even more steps in order to obtain the "ideal oxynitride layer". These are:

- Thermal processes
  - silicon oxynitridation in NO,
  - silicon oxynitridation in N_2O,
  - nitridation of previously grown oxide in NH_3,
  - nitridation of previously grown oxide in N_2.
- Deposition (CVD or PVD) and plasma assisted techniques
  - nitrogen implantation into silicon and then thermal or plasma oxidation (or the same in reverse order),
  - plasma nitridation in e.g. remote plasma,
  - PECVD deposition of an oxynitride layer,
  - atomic layer deposition (ALD),
  - jet vapour deposition (JVD).

3.1. Thermal processes

Thermal processes are performed at high temperatures not only due to the fact that the oxynitride-growth rate is relatively low when compared to that of the oxide, but also because of the observed dependence of the nitrogen content in the layer on the process temperature (the higher the temperature, the more nitrogen incorporated in the layer – see Fig. 2). Typically, however, the concentration of nitrogen (at the SiO_xN_y-Si interface) is of the order of 10^{15} cm^{-2}. When averaged across the whole layer it gives considerably less than 20% of the nitride in the oxynitride.

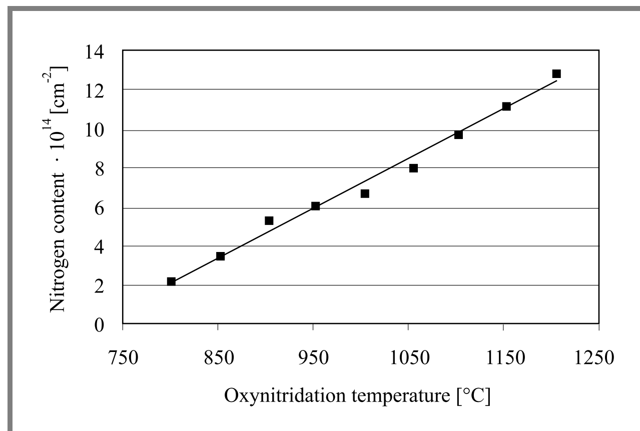


Fig. 2. Nitrogen content in 10 nm oxynitride films grown in N_2O by means of RTP (after [26]).

In order to obtain more nitrogen within the layer deposition methods have to be used. There are, however, several consequences of this change. First, low-temperature processing (deposition temperatures are usually $< 400°C$) results in a different structure of the layer (often non-equilibrium films). Moreover, (in contrast to high-temperature processes) a lot of hydrogen remains in the deposited layer. Both these features have much influence on the electrophysical properties of the whole gate system (silicon/oxynitride/poly-Si). This, in turn, may have considerable impact on the reliability.

Second, the layer is deposited onto the silicon surface, which is inevitably covered by a natural oxide. The thickness and nature of this oxide strongly depend on the method of chemical cleaning used. The oxide layer, rarely thinner than 0.4 nm, does not undergo much change during a low-temperature deposition process. Thus, unless special treatment (e.g. high temperature post deposition annealing) is used, it is precisely this layer that forms the silicon-oxynitride interface, which may obviously lead to poor interface properties of as-deposited layers. On the other hand, it may be expected that any high temperature process performed afterwards may considerably change the interface properties and composition profile of these layers.

As far as the formation of ultrathin oxynitride layers is concerned, high-temperature methods used to be the most obvious solution, especially since they have been used successfully for many years to form ultrathin oxide layers. Thus, they were also the first methods to be studied very carefully in the late nineties. The results were summarized in excellent reviews (e.g. [6, 7]).

There are important differences between depositing oxynitride layers in NO and N_2O ambients, although they might not be quite obvious at a glance. They result from the fact that the behaviour of NO and N_2O at high temperatures is different. The former is stable and diffuses to the interface, where the oxynitride is formed, while the latter rapidly decomposes and the reaction may follow many different paths (several tens of possible reactions), making the whole process difficult to control. Oxynitridation in NO is advantageous to process carried out in N_2O also due to the fact that it results in more nitrogen incorporated in the final layer (approximately < 15 at.% vs. < 5 at.% only for NO and N_2O, respectively). From the device point of view the advantages of oxynitridation in NO are lower leakage current and lower interface defect densities, as well as improved electrical stress properties. With a temperature rise, the concentration of incorporated nitrogen increases, and so does the width of the nitrogen containing region.

The kinetics of the growth of oxynitride films also make the process carried out in NO advantageous when compared to that performed in N_2O. In the former, the growth rate of the oxynitride film is much lower than in the latter. Thus, the process carried out in NO allows easier control (better process repeatability and layer uniformity) of the formation of extremely ultrathin oxynitride layers at high temperatures.

The differences in the nitrogen profile control are also observed between films grown in RTP (RTN) reactors and those grown in classical furnaces. The nitrogen profiles in the layers grown in the former are rather narrow and located at the interface, while in the latter much broader profiles are obtained. These effects are attributed to the location where the decomposition of the reactive gas takes place. In the case of furnaces it is the gas inlet of the hot wall tube, whereas in RTP reactive gas decomposes only in the vicinity of the hot wafer.

When nitridation of the previously grown ultrathin layer is considered, two types of ambient gas may be used, namely N_2 or NH_3. Pure nitrogen is believed to be a relatively inert gas, but annealing of SiO_2 layers in N_2 has been proved (many years ago, already) to reduce the fixed charge in the oxide-silicon system, which has been attributed to the formation of a SiO_xN_y layer at the interface. Forming an oxynitride layer by means of a nitridation in N_2 requires, however, a very high temperature – usually too high for the thermal budget typical of modern technologies. Replacing the furnace with an RTN may improve the situation a little, but not much (Fig. 3).

Nitridation of oxide layers in NH_3 is definitely a more effective method of oxynitride formation. It allows incorporation of some 10–15 at.% of nitrogen into the layer. In

this case, however, the nitrogen piles up at both interfaces of the layer, namely: at silicon-oxynitride and oxynitride-poly-Si. Increasing the nitridation time results in more uniform nitrogen distribution, while almost no film growth is observed. The layers formed by this method exhibit, however, high concentrations of hydrogen that can play a role of traps. The concentration of hydrogen seems to follow that of nitrogen: the higher the temperature and the longer the process, the higher the concentration of both nitrogen and hydrogen. In order to reduce the hydrogen concentration within the layer, a high-temperature annealing process has to be performed.

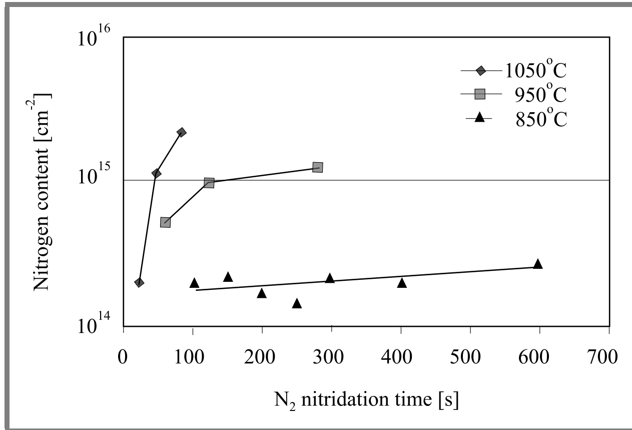


Fig. 3. Kinetics of nitrogen incorporation in an oxynitride layer during RTN (in N_2) at different temperatures (after [27]).

3.2. Deposition and plasma assisted techniques

A number of methods involving plasma or ion beam techniques (i.e. ion implantation) are studied for the application in oxynitride formation. Except for PECVD and JVD methods, where the oxynitride layers are formed in a single process, all the rest of these methods require a few processing steps (applying, except of ALD – different techniques).

Implantation of nitrogen atoms may be used to form a nitrogen rich layer in a silicon substrate, which then can be oxidised, either thermally or by means of a plasma-assisted method, to form the final oxynitride layer. This approach is particularly interesting for mixed IC or SOC technologies, where different gate oxide thicknessses are required depending on the role played by the individual devices (digital circuit, low power or memory). It has been shown (e.g. [8–10]) that the dose of the implanted nitrogen may be used to control the oxidation rate, and consequently, the final oxynitride thickness. This way, instead of individual processes of gate oxide formation for each type of devices, local implantations may be used to predetermine the final oxynitride thickness. Thus, oxynitride layers with different thickness are obtained in single a oxidation process (approach often referred to as “dual gate process”). This sequence has some limitations that should be always kept in mind. Forming ultrathin oxynitride layers requires ultra-shallow implantation of nitrogen, which would suggest the use of very-low-energy implantation. Most implanters cannot, however, produce reasonable ion currents at low energies, which considerably reduces the process throughput. This is why attempts are being made to use alternative techniques, such as ion immersion implantation (III) (e.g. [11]) or even implantation from plasma sources, e.g. ECR (e.g. [12]), or r.f. plasma (e.g. [13]). These techniques lead to the formation of ultrathin oxynitride films with the peak of the nitrogen concentration located at the silicon-oxynitride interface (exactly as for all single-step high-temperature processes).

The reverse order of processing steps, i.e. oxidation first, nitrogen implantation later, may, on the other hand, offer some other technological possibilities. First, it is believed to result in lower concentration of defects in the oxynitride layer. This is most probably the consequence of the amorphization of silicon by implanted nitrogen, which does not get annealed out neither during the following low temperature plasma oxidation (temperature too low), nor even during a very short high-temperature oxidation (time too short). The effects of silicon surface damage due to ion bombardment during r.f. plasma exposure were studied e.g. in [13]. Even short exposure to low-power argon plasma in parallel symmetric electrode system, which is commonly believed to result in little radiation damage, appeared to increase considerably the oxidation rate of the subsequent plasma oxidation. This was attributed to the silicon amorphization during the exposure to argon plasma (see Fig. 4 where the probable distributions of very low energy argon and oxygen in silicon are shown).

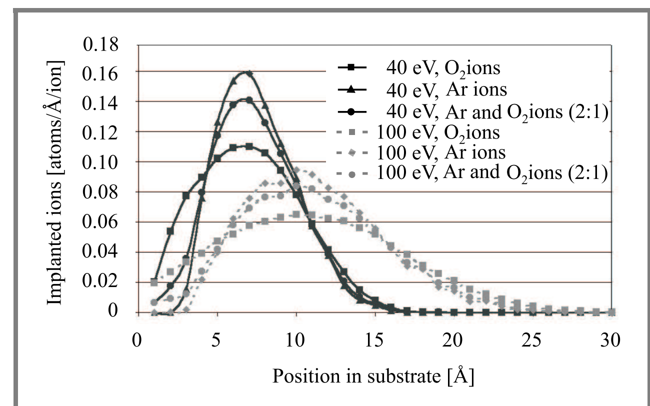


Fig. 4. The distribution of Ar and O atoms implanted into Si substrate during plasma oxidation in a parallel plate symmetric reactor in pure O_2, Ar and Ar:O_2 (2:1), as calculated with SRIM (after [28]).

On the other hand, even this low thermal budget annealing is enough to reduce considerably the initial nitrogen content in the film (especially during high temperature oxidation). The second approach enables also the silicon region to re-

main almost untouched, on condition that the implantation process is carried out very carefully (the right choice of energy and dose). This creates even more interest in very low-energy implantations performed from plasma sources (e.g. ECR). These processes yield totally different profiles of the nitrogen incorporation in the film – the maximum of the nitrogen profile stays away from the interface. As a consequence, the danger of carrier mobility degradation due to excess nitrogen at the interface is eliminated. In some extreme cases a pure- SiO_2 interface may remain, exactly as before the implantation process (e.g. [14]).

Plasma processes can be performed in different modes, such as planar reactors or remote plasma assisted process (RPAP), and using various plasma ignition methods, such as: microwaves, electron-cyclotron-resonance (ECR) or simply – radio-frequency (r.f.).

Among them, remote-plasma-assisted processing seems to have been studied particularly often in the past, e.g. [15–17]. Its potential advantages are:

– selective excitation of source and carrier gases (determined by their point of injection into the reactor chamber),

– substrate is located outside the plasma glow region,

– gas flow and pressure prevent backstreaming of source gases into the plasma generation region.

In this type of experimental set up two types of processes can be performed, namely: remote-plasma-oxidation (RPO) (or nitridation – RPN), or remote plasma enhanced chemical vapour deposition (RPECVD). In the first group, the active species react with the silicon substrate forming either oxide (RPAO) or nitride (RPAN) layers. In the latter process type (RPECVD), the species activated in plasma undergo a chemical reaction and form oxide or nitride layers without consuming silicon atoms from the substrate, however, because these are supplied with the reactive gases (SiH_4 is typically used for this purpose).

The source gases used in the nitridation step are usually N_2:He [16] or N_2:Ar [16, 17], where both He and Ar particles serve as nitrogen excitation promoters. The resulting nitrogen concentration in the ultrathin oxynitride layer can be of the order of 8–10 at.% [17], which is similar to thermal nitridation in NO. It is, however, suggested that in order to obtain the best possible properties of the ultrathin oxynitride layer and, consequently, MOSFET devices, multi-step processing should be applied. An exemple of a set of 6 processing steps presented in [15] includes: surface etching, sacrificial oxide thermal oxidation, annealing, etching off sacrificial oxide, remote plasma oxidation, and remote plasma nitridation.

PECVD process allows basically a single-step deposition of an oxynitride layer. The main difficulty in this case is the reduction of the growth rate to the level allowing for good control of the ultrathin deposition process. The obvious move is to decrease the applied r.f. power and reactive-gas pressure, but in the case of an r.f. process a limit exists that results form the need to keep the plasma stable and uniform across the regions larger than the size of the silicon wafer.

Another possible solution to this problem is diluting the reactive gases in an inert gas, e.g. argon or helium. This can lead to precise control of the process within the whole ultrathin thickness range (Fig. 5) [18]. The bombardment of inert-gas ions can, however, also contribute to the final properties of the oxynitride layer. This effect is not always taken into account, as primary consequences of substrate amorphisation, i.e. higher oxidation (nitridation) rate, are difficult to distinguish from the dependence of the nitrogen excitation rate on the type of an inert gas used. In fact, it has been shown in [17] that even in remote-plasma reactors, believed to produce growth conditions almost free of ion bombardment, diluting N_2 with Ar may lead to significant changes not only in the deposition rate of the layer, but also in the nitrogen incorporation during the nitridation process. If other process conditions are the same, the nitrogen incorporation resulting from the use of N_2 or N_2/Ar plasma is 8 at.% and 22 at.%, respectively. Other observations presented in this work prove that ion bombardment is essential to understand the process of plasma nitridation of oxide layers formed previously, because oxygen removal (e.g. by bombarding Ar ions) from the silicon oxide lattice seems to be a limiting step in the nitridation of such a film. Bombardment of the substrate with Ar ions during the deposition process (already mentioned above) leads to radiation damage in the deposited layers, that can be detected e.g. by means of XPS, as shown in [19].

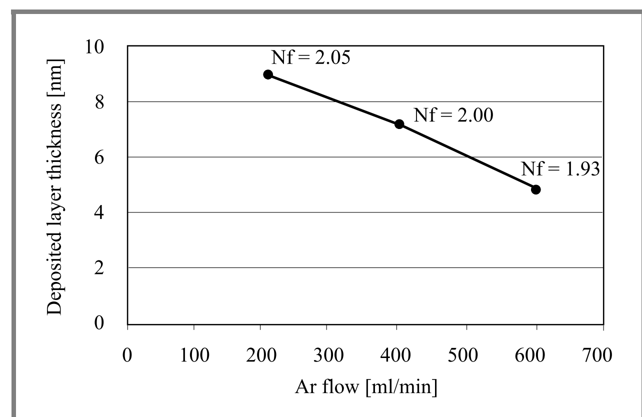


Fig. 5. PECVD deposited nitride layer thickness as a function of argon flow. The process duration is 2 min. The respective refractive index values (@ 632 nm) are also shown (after [18]).

The primary difference between plasma deposition methods and the techniques of oxynitride formation discussed above is that they allow very high concentration of nitride (far beyond the reach of any other processes) to be obtained in oxynitride layers. Depending on the reactive gases used, the nitride concentration in an as-deposited ultrathin layer can be even as high as 95% and as low as 60%, or be-

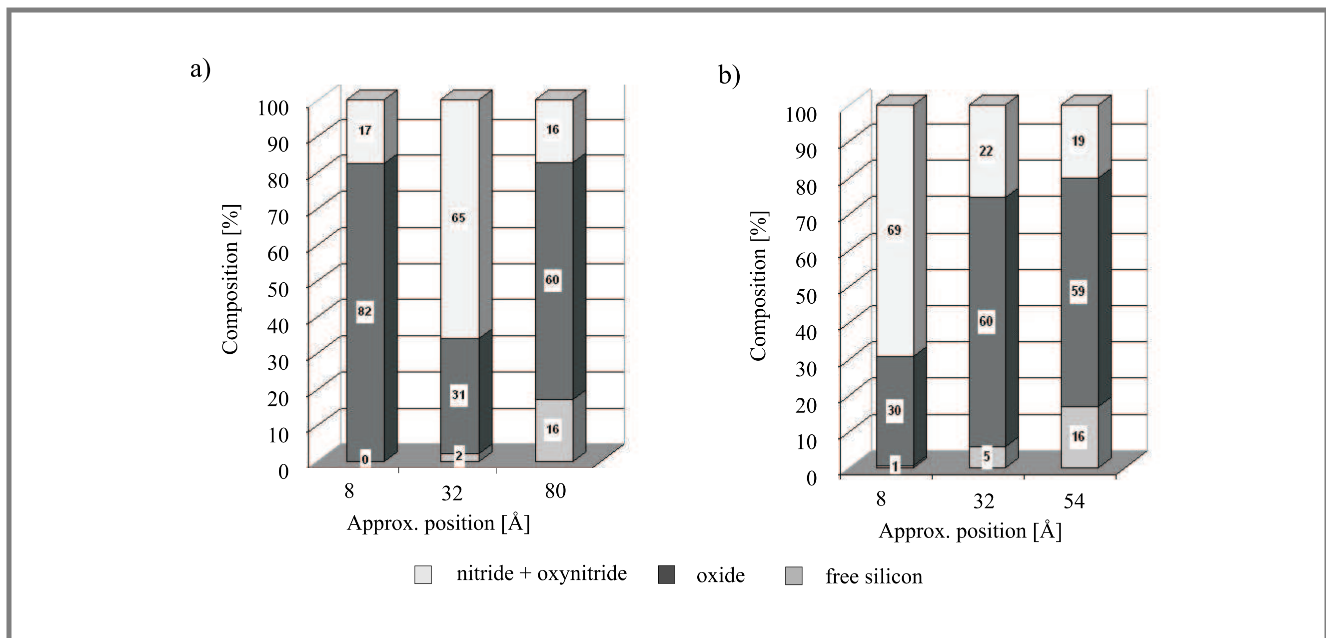


Fig. 6. Comparison of XPS determined compositions at several selected positions within the layer for two different sets of PECVD parameters: (a) 83 Å layer (parameter set A) and (b) 60 Å layer (parameter set B) (after [20]).

tween 40% and 60%, as demonstrated by spectroscopic ellipsometry studies. Such high concentrations of nitride are often attributed to non-equilibrium, or nitrogen excess in the oxynitride structure.

On top of that, it has been shown that, in contrast to thermal processes, one can obtain different type of the nitrogen profile within the film by means of choosing appropriate PECVD process parameters [20]. In fact, as it can be seen in Fig. 6, PECVD appears to be very sensitive to process parameters in this respect.

Silane – SiH_4 – is usually used as silicon source, although some papers report good results also with Si_2F_6 or SiH_2Cl_2. In order to get high concentrations of nitride, ammonia (NH_3) has to be used as the nitrogen source in plasma discharge, while for lower concentrations – NH_3 and N_2O. Nitrogen, which is used for diluting SiH_4 most often (in order to have a non-flammable mixture N_2 98%: SiH_4 2% is often used), is not effective in this respect (exactly as in the case of thermal processing). The use of ammonia in a low-temperature process has, however, another consequence, too. Oxynitride layers formed in this way exhibit unusually high concentration of hydrogen. Some reports claim even up to 30% of hydrogen in oxynitride layers. Due to the possible detrimental effects of hydrogen present in the dielectric-semiconductor system, a concentration of hydrogen this high cannot be tolerated. Thus, in order to improve the quality of the dielectric films in question, high-temperature annealing (usually RTA) is recommended.

Annealing of the ultrathin PECVD oxynitride layers at 600°C and 800°C in nitrogen was studied in [21, 22]. XPS studies, as well as optical and electrical characterization, have proved that the behaviour of the films is quite complex, as many effects take place simultaneously (massive hydrogen outdiffusion, film densification, nitrogen substitution by oxygen, etc.). Understanding these mechanisms still requires a lot of research to be done.

ALD is a type of chemical vapour deposition that typically allows deposition of less than a single monolayer per process cycle and has been successfully studied for a number of applications. Although the use of ALD to form ultrathin silicon-nitride gate-dielectric layers has been shown in a number of papers (e.g. [23, 24]), to our knowledge not much has been done so far in terms of identifying the obtained nitrogen profile or nitrogen content in the films. Thus, we will not comment on this method in this review.

The silicon nitride deposition by means of JVD is usually carried out in a co-axial dual-nozzle jet vapour source, in which highly diluted silane and $Ne:H_2$ mixture flow into microwave discharge (e.g. [25]). Similarly to the case of ALD, most of the effort has been spent, so far, to prove the quality of the layer and its performance in the test devices and to our knowledge, no data on nitrogen content and profile are available.

4. Conclusions

Device requirements allow the concept of "ideal ultrathin oxynitride" gate dielectric film to be formulated. This concept should take advantage of both, the increase of hot-electron resistance caused by a limited (!!) nitrogen incorporation at the silicon-oxynitride interface, and the resistance to boron diffusion – due to a high content of nitrogen at the top of the dielectric film.

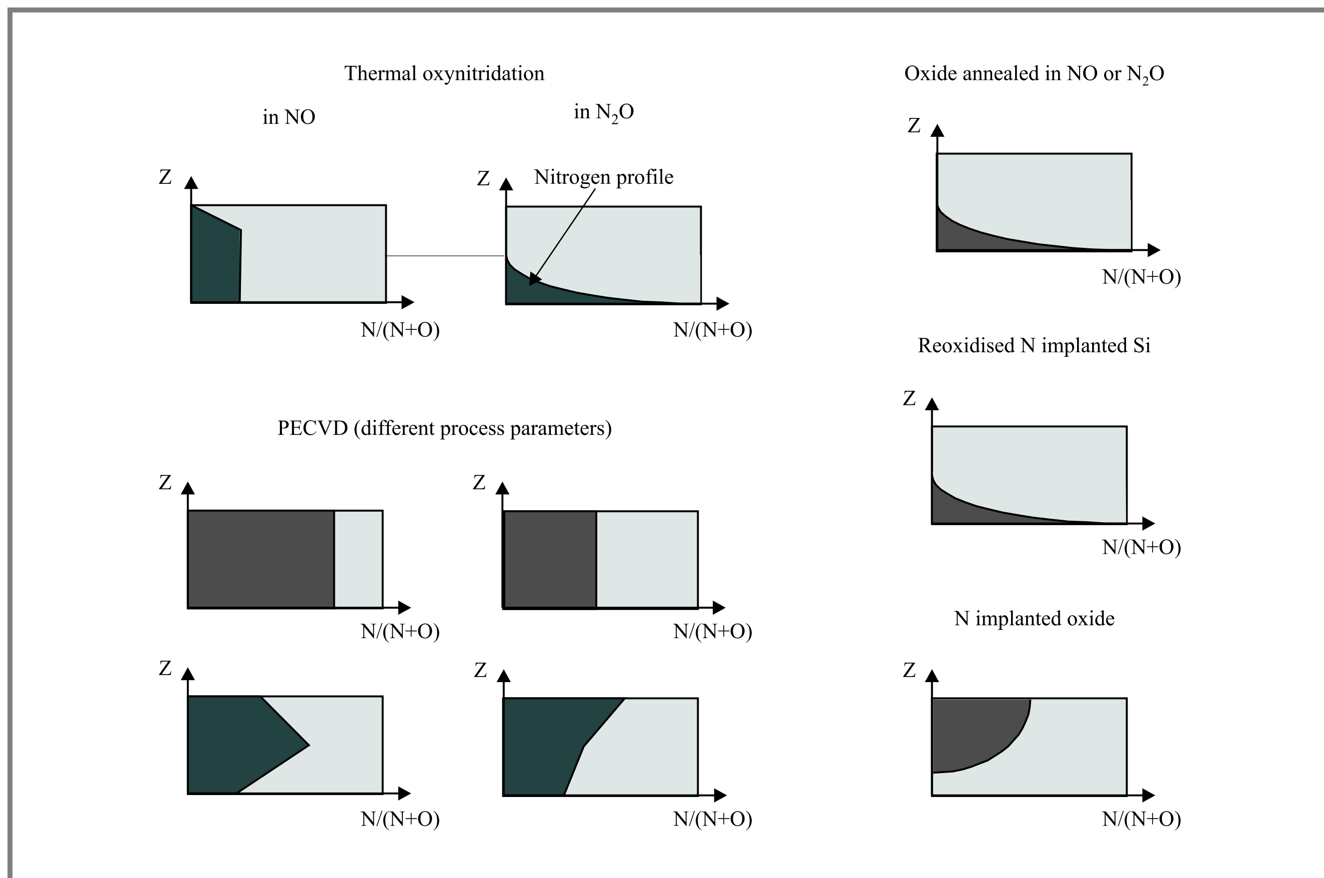


Fig. 7. Typical nitrogen incorporation profiles obtained by different technological methods of ultrathin oxynitride film formation.

In this review we have presented a variety of the methods that may be used to form ultrathin oxynitride and discussed nitrogen profiles obtained typically in the oxynitride layers formed (see Fig. 7). Thermal processes, so far studied the most heavily, do not provide nitrogen profiles of interest. The overall content of nitrogen in the layers formed in this way is lower, while the interface (in most cases) tends to be heavily nitrided. Multi-step processing has to be performed in order to overcome this limitation.

The methods involving ion implantation are more flexible in terms of the obtained nitrogen profiles. Due to the possibility of radiation damage, the technological difficulty is reducing the energy of ions reaching the sample, while simultaneously maintaning the control and the repeatability of the process. This group of methods is of particular interest due to the fact that oxynitride layers with different thickness may be manufactured in a single oxynitridation process – a feature exceptionally interesting for SOC technologies.

Growing oxynitride layers be means of PECVD appears the most flexible in this comparison. These methods may produce many different nitrogen profiles in the film in a single technological process, depending on the process parameters. On top of that, the incorporation of nitrogen in these layers can be the highest. Thus, even if it is remembered that the concentration of the nitrogen in the film is reduced after high-temperature annealing, PECVD oxynitride layers still "hold the record".

5. Acknowledgements

The authors would like to acknowledge also the contribution to this work of: T. Bieniek, M. Cuch and M. Giedz.

This work was supported by research project no. 8T11B 074 19 financed by the State Committee for Scientific Research.

References

[1] C. M. Osburn *et al.*, "Vertically scaled MOSFET gate stacks and junctions: how far are we likely to go?", *IBM J. Res. Develop.*, vol. 46, no. 2-3, pp. 299–315, 2002.

[2] M. Bhat *et al.*, "Electrical properties and reliability of MOSFET's with rapid thermal NO-nitrided SiO_2 gate dielectrics", *IEEE Trans. Electron Dev.*, vol. 42, no. 5, pp. 907–914, 1995.

[3] W. L. Hill *et al.*, "Low-pressure rapid thermal chemical vapor deposition of oxynitride gate dielectrics for n-channel and p-channel MOSFETs", *IEEE Trans. Electron Dev.*, vol. 43, no. 1, pp. 15–22, 1996.

[4] D. Wristers *et al.*, "Degradation of oxynitride gate dielectric reliability due to boron diffusion", *Appl. Phys. Lett.*, vol. 68, no. 15, pp. 2094–2096, 1996.

[5] D. Bouvet *et al.*, "Influence of nitrogen profile on electrical characteristics of furnace – or rapid thermally nitrided silicon dioxide films", *J. Appl. Phys.*, vol. 79, no. 9, pp. 7114–7122, 1996.

[6] I. J. R. Baumvol, "Atomic transport during growth of ultrathin dielectrics on silicon", *Surf. Sci. Rep.*, vol. 36, no. 1–8, pp. 5–166, 1999.

[7] M. L. Green *et al.*, "Ultrathin (< 4 nm) SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics: understanding the processing, structure, and physical and electrical limits", *J. Appl. Phys.*, vol. 90, no. 5, pp. 2057–2121, 2001.

[8] P. Berruyer and M. Bruel, "Nitrogen implantation for local inhibition of oxidation", *Appl. Phys. Lett.*, vol. 50, no. 2, pp. 89–91, 1987.

[9] N. In-Ho *et al.*, "Nitrogen profile effects on the growth rate of gate oxides grown on nitrogen-implanted silicon", *J. Vac. Sci. Technol. B*, vol. 19, no. 1, pp. 299–304, 2001.

[10] C. H. Lee and D. L. Kwong, "Multiple gate oxide technology using nitrogen implantation and high-pressure O_2 oxidation", *Semicond. Sci. Technol.*, vol. 18, no. 2, pp. 88–91, 2003.

[11] I. J. R. Baumvol *et al.*, "Ultrathin silicon oxynitride film formation by plasma immersion nitrogen implantation", *Appl. Phys. Lett.*, vol. 74, no. 6, pp. 806–808, 1999.

[12] T. Seino, T. Matsuura, and J. Murota, "Contribution of radicals and ions in atomic-order plasma nitridation of Si", *Appl. Phys. Lett.*, vol. 76, no. 3, pp. 342–344, 2000.

[13] T. Bieniek *et al.*, "Ultra shallow nitrogen plasma implantation for ultrathin silicon oxynitride (SiO_xN_y) layers formation", in *6th Symp. Diagn. Yield – Adv. Silic. Dev. Technol. ULSI Era*, Warsaw, Poland, 2003.

[14] A. Raveh, J. Brewer, and E. A. Irene, "Nitridation of thermal SiO_2 films by radio-frequency plasma assisted electron cyclotron resonance: layer structure and composition", *J. Vac. Sci. Technol. A – Vac. Surf. Films*, vol. 19, no. 1, pp. 17–24, 2001.

[15] G. Lucovsky, "Ultrathin nitrided gate dielectrics: plasma processing chemical characterisation, performance, and reliability", *IBM J. Res. Develop.*, vol. 43, no. 3 pp. 301–326, 1999.

[16] A. Khandelwal, B. C. Smith, and H. H. Lamb, "Nitrogen incorporation in ultrathin gate dielectrics: a comparison of He/N_2O and He/N_2 remote plasma processes", *J. Appl. Phys.*, vol. 90, no. 6, pp. 3100–3108, 2001.

[17] A. Khandelwal *et al.*, "Low-temperature Ar/N/sub 2/ remote plasma nitridation of SiO_2 thin films", *J. Vac. Sci. Technol. A – Vac. Surf. Films*, vol. 20, no. 6, pp. 1989–1996, 2002.

[18] R. B. Beck *et al.*, "PECVD formation of ultrathin silicon nitride layers for CMOS technology", Vacuum *(Surf. Eng., Surf. Instrum. Vac. Technol.)*, vol. 70, pp. 323–329, 2003.

[19] R. B. Beck *et al.*, "Electrophysical properties of formed by PECVD ultrathin (< 6.0 nm) silicon nitride layers", in *6th Symp. Diagn. Yield – Adv. Silic. Dev. Technol. ULSI Era*, Warsaw, Poland, 2003.

[20] R. B. Beck *et al.*, "Very thin (< 10 nm) silicon oxynitride (SiO_xN_y) layers fromed by PECVD", in *Electrochem. Soc. Meet. – Chem. Vap. Deposit. XVI & Symp. EUROCVD-14*, Paris, France, 2003.

[21] R. B. Beck *et al.*, "Thermal stability of formed by PECVD ultrathin (< 6.0 nm) silicon oxynitride layers determined by optical and electrical measurements", in *6th Symp. Diagn. Yield – Adv. Silic. Dev. Technol. ULSI Era*, Warsaw, Poland, 2003.

[22] P. Hoffmann *et al.*, "Photoemission studies of very thin (< 10 nm) silicon oxynitride (SiO_xN_y) layers formed by PECVD", in *Proc. E-MRS*, Warsaw, Poland, 2003.

[23] Q. D. M. Khosru *et al.*, "High-quality NH/sub 3/-annealed atomic layer deposited Si-nitride/SiO_2 stack gate dielectrics for sub-100 nm technology generations", *Elsev. Solid State Electron.*, vol. 46, no. 10, pp. 1659–1664, 2002.

[24] A. Nakajima *et al.*, "NH_3-annealed atomic-layer-deposited silicon nitride as a high-*k* gate dielectric with high reliability", *Appl. Phys. Lett.*, vol. 80, no. 7, pp. 1252–1254, 2002.

[25] T. P. Ma, "Gate dielectric properties of silicon nitride films formed by jet vapor deposition", *Elsev. Appl. Surf. Sci.*, vol. 117–118, pp. 259–267, 1997.

[26] M. L. Green *et al.*, "Rapid thermal oxidation of silicon in N_2O between 800 and 1200 degrees C: incorporated nitrogen and interfacial roughness", *Appl. Phys. Lett.*, vol. 65, no. 7, pp. 848–850, 1994.

[27] M. L. Green et al., "Ultrathin SiO_xN_y by rapid thermal heating of silicon in N_2 at $T = 760 - 1050$ degrees C", *Appl. Phys. Lett.*, vol. 71, no. 20, pp. 2978–2980, 1997.

[28] R. B. Beck *et al.*, "Low temperature plasma oxidation of silicon – possibility of application in CMOS-ULSI technology", in *Conf. Plasma Chem.*, Kazimierz Dolny, Poland, 2002.

Romuald B. Beck graduated from the Faculty of Electronics, Warsaw University of Technology, in 1977 and joined the Institute of Microelectronics and Optoelectronics the same year. He received the Ph.D. and D.Sc. degrees in 1985 and 1996, respectively. His research activities are concentrated in the area of modeling, diagnostics and technology of metal-insulator-semiconductor structures, especially with very thin and ultra thin oxide. His work is focused on theoretical and experimental studies of different methods of dielectric-layer formation (including those involving DC and AC plasma) and their kinetics, as well as the yield, electrophysical properties and reliability of the resulting devices. His research interests include also dry etching methods and their implementation in modern IC and MEMS technology. He authored and co-authored more than 90 technical publications.

e-mail: beck@imio.pw.edu.pl
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Andrzej Jakubowski – for biography, see this issue, p. 14.

Invited paper

Closed-form 2D modeling of sub-100 nm MOSFETs in the subthreshold regime

Jarle Østhaug, Tor A. Fjeldly, and Benjamin Iñiguez

Abstract — Closed-form 2D modeling of deep-submicron and sub-100 nm MOSFETs is explored using a conformal mapping technique where the 2D Poisson equation in the depletion regions is separated into a 1D long-channel case and a 2D Laplace equation. The 1D solution defines the boundary potential values for the Laplacian, which in turn provides a 2D correction of the channel potential. The model has been tested for classical MOSFETs with gate lengths in the range 200–250 nm, and for a super-steep retrograde MOSFET with a gate length of 70 nm. With a minimal parameter set, the present modeling reproduces both qualitatively and quantitatively the experimental data obtained for such devices.

Keywords — sub-100 nm MOSFET, two-dimensional device modeling, conformal mapping, threshold voltage, subthreshold regime, leakage current.

1. Introduction

Very important issues in the development of modern semiconductor device technology are the increasing levels of complexity in the fabrication process and the many subtle mechanisms that govern the properties of advanced devices. These mechanisms, which are explained by a careful consideration of the device physics, have to be formulated and implemented into process modeling and circuit design tools, to empower the circuit designers with means to fully utilize the potentials of modern technology.

However, it is recognized that the development of commercial circuit design tools lags significantly behind the pace of progress in device processing. This time lag creates a costly delay in the adoption of new technology in circuit designs, especially for the many companies that depend on chip foundries for the fabrication of their proprietary circuits. They consistently have to contend with designs based on relatively conservative design rules. Hence, a strong impetus exists for focusing on advanced device modeling for circuit simulators and other computer aided design (CAD) tools.

For silicon CMOS technology, the present industry modeling standards for circuit simulation, such as BSIM3 [1, 2], BSIM4 [3, 4], EKV [5], MOS Model 9 [6], and MOSA1 [7, 8], are based on one-dimensional (1D) theory, initially developed for long-channel FETs. However, the steady reduction in feature size, with gate lengths presently well into the sub-100 nm regime [9], has strongly enhanced a number of phenomena, collectively known as short-channel effects, related to the two- and even the three-dimensionality of the device structures. To keep pace with technology, this has necessitated extensive, phenomenological modifications of the 1D models, resulting in a steady erosion of their physical basis and a plethora of model parameters, many of which are of an obscure origin.

Therefore, as a prerequisite to obtaining very precise descriptions of the next generations of MOSFETs, the consideration of 2D models based on a self-consistent solution of the 2D field-pattern in the device [10–15] has become necessary. In such an approach, short-channel effects and scaling properties will be intrinsic to the model, which, accordingly, may require a minimum set of parameters of clear physical origin. Hence, a close accord is established with the fabrication process. For the same reason, such models will represent a significant and needed improvement for use in CAD tools aimed at the next generation of circuit design. In particular, the 2D strategy is expected to yields precise scaling information on important properties, such as threshold voltage and the subthreshold leakage current. Above threshold properties will have to be solved self-consistently using appropriate transport formalisms for the channel current.

Here, we consider closed-form 2D solutions of the subthreshold properties of deep submicron and of sub-100 nm MOSFETs. Following the procedure by Klös and Kostka [15], the 2D Poisson equation for the depletion regions is separated into a 1D long-channel problem and a Laplacian with well-defined boundary conditions for the 2D region under the gate. The latter is solved using conformal mapping techniques [16]. In this work, the definition of the 2D boundary conditions and the mapping functions used in the Laplace problem have been improved compared to those of [15], and the technique has been extended to sub-100 nm range MOSFETs.

This method applies to classical MOSFETs as well as to non-classical structures [17]. We note that the classical MOSFET approach is directly applicable to super-steep-retrograde (SSR) channel doped devices, to Si-Si/Ge strained devices, and to sidewall vertical MOSFETs with a partially depleted body. Likewise, the same approach, with somewhat different boundary conditions for the Laplacian, may be applicable to double-gate fully depleted vertical MOSFETs and double-gate SOI MOSFETs.

The device geometry for the classical MOSFET and the corresponding boundary conditions of the Laplacian are discussed first. Next, the conformal mapping technique is

presented. This approach is then applied to classical MOSFETs with gate lengths between 200 and 250 nm, where results for the channel potential profile, the scaling of the threshold voltage with the gate length, the drain induced barrier lowering (DIBL) effect, and the subthreshold leakage current are presented and compared to experiments. Finally, corresponding results for sub-100 nm SSR MOSFET are presented.

2. Modeling approach

Consider the schematic classical MOSFET geometry of Fig. 1, which shows the actual source and drain contact regions and the depletion region outlined by dashed lines. Starting from the work by Klös and Kostka [15], the contact regions in the present analysis are approximated by rectangular boxes, and the potential distributions of the source and drain depletion regions (Regions 2 and 3) are calculated by means of a 1D Poisson equation.

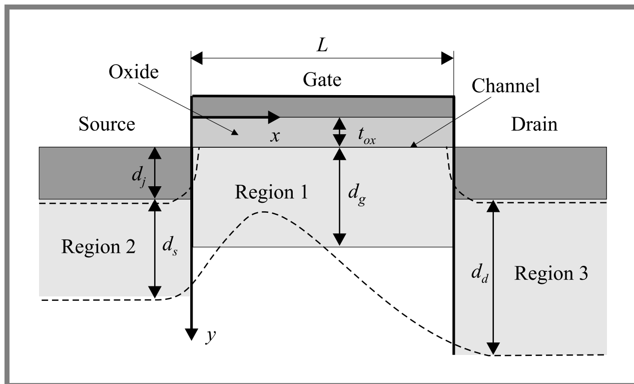


Fig. 1. Schematic MOSFET geometry. The lower dashed line indicates the depletion boundary resulting from a 2D analysis.

A 2D analysis may be performed in the central region under the gate, where the superposition principle allows us to partition the actual 2D potential distribution into that corresponding to a 1D Poisson equation and that of a Laplacian. We note that the only adjustable parameter in this procedure is the effective thickness used for the source and drain contacts (to compensate for the rounded shape of the actual contacts near the gate). The threshold and subthreshold properties of this device are largely determined by the channel potential profile $\Phi_o(x) = \phi_o = \varphi_o(x)$, where ϕ_o is the channel potential according to the 1D long-channel case and $\varphi_o(x)$ is the contribution from the 2D Laplacian in Region 1. All the potentials are measured relative to that of the substrate interior.

The 2D contribution to the channel potential profile can be determined by either considering Region 1 as consisting of only the semiconductor slab under the oxide or by including the gate oxide as well [15]. In the first case, the channel is a boundary for the Laplacian. In the second case, the oxide-metal interface is the boundary. It is therefore necessary to scale the actual oxide thickness t_{ox} to $t'_{ox} = t_{ox}\varepsilon_s/\varepsilon_i$, where ε_s and ε_i are the semiconductor and the oxide permittivities, respectively. Also the "oxide" part must remain neutral and non-conducting. Since the second approach provides a better definition of the boundary condition for the Laplacian, it will be used in the present analysis.

From the channel potential profile $\Phi_o(x)$ resulting from our analysis, we can determine the threshold voltage V_T and the subthreshold current I_{sub}. V_T corresponds to the gate-source voltage V_{GS} for which the minimum value in the channel potential satisfies the threshold condition, $\Phi_{o,\min} = 2\varphi_b$. Here $\varphi_b = V_{th}\ln(N_s/n_i)$, where V_{th} is the thermal voltage, N_s is the substrate doping, and n_i is the intrinsic carrier density in silicon. I_{sub} for a given V_{GS} is determined by the channel potential profile and, in particular, by the value of $\Phi_{o,\min}$.

In practice, the channel potential $\Phi_o(x)$ has to be determined from the normal electrical field $E_n(x)$ pointing into Region 1 from the channel. As will be shown below, $E_n(x)$ consists of the field contribution E_o from the 1D analysis and the contribution $E_{2D}(x)$ from the 2D analysis. The latter is found by performing an integration over all the boundaries of Region 1.

Here we assume for simplicity that the substrate is uniformly doped. Non-uniform doping profiles in the vertical direction may be approximated by two or more horizontal layers of different but uniform doping (see below). The same would apply for strained Si-Si/Ge MOSFETs.

2.1. 1D boundary conditions

The boundary conditions for the 2D Laplace problem in Region 1 (including the oxide) are defined in terms of the potential distributions along at the oxide-metal interface ($y = 0$) and at the vertical boundaries at $x = 0$ and $x = L$. At the vertical boundaries, the potential distributions are derived from the 1D Poisson equations in Regions 1–3, the potentials of the source and drain contacts, and the potential at the vertical sidewalls of the oxide.

1D potential distribution in Region 1. In the 1D approximation, the potential distribution $\phi_g(y)$ relative to the substrate interior in Region 1 becomes:

$$\phi_g(y) = \begin{cases} \varphi_g + (\phi_o - \varphi_g)\dfrac{y}{t'_{ox}}, & 0 \le y < t'_{ox} \\ \phi_o + \dfrac{qN_s}{2\varepsilon_s}(y - t'_{ox})^2 - E_o(y - t'_{ox}) & \\ & t'_{ox} \le y < d_g + t'_{ox} \\ 0, & y \ge d_g + t'_{ox} \end{cases}, \quad (1)$$

where φ_g is the potential at the gate-"oxide" interface,

$$d_g = \sqrt{\frac{2\varepsilon_s\phi_o}{qN_s}} \quad (2)$$

is the depletion depth, and

$$E_o = \frac{qN_s}{\varepsilon_s} d_g = \sqrt{\frac{2qN_s\phi_o}{\varepsilon_s}} \tag{3}$$

is the magnitude of the vertical 1D electric field contribution at the semiconductor-oxide interface.

At threshold, the 1D channel potential is $\phi_o = V_{SB} + 2\varphi_b$, and below threshold it becomes:

$$\phi_o = V_{GB} - V_{FB} + \frac{qN_s}{\varepsilon_s} t'^{\,2}_{ox} + -2t'_{ox}\sqrt{\frac{qN_s}{2\varepsilon_s}\left(V_{GB} - V_{FB} + \frac{qN_s}{2\varepsilon_s} t'^{\,2}_{ox}\right)}, \tag{4}$$

where V_{GB} is the gate-substrate voltage and V_{FB} is the flat band voltage. It is assumed in Eq. (1) that the oxide thickness is much smaller than the gate length.

We note that the total 2D potential distribution in Region 1 is

$$\Phi_g(x, y) = \phi_g(y) + \varphi_g(x, y), \tag{5}$$

where $\varphi_g(x, y)$ is the solution of the Laplacian.

1D potential distribution in Regions 2 and 3. The potential distributions in Regions 2 and 3 are initially approximated by 1D distributions of the following form:

$$\phi_{s,d}(y) = \begin{cases} \phi_g + (V_{s,d} - \phi_g)\dfrac{y}{t'_{ox}}, & 0 \le y < t'_{ox} \\ V_{S,D}, & t'_{ox} \le y \le d_j + t'_{ox} \\ V_{S,D} + \dfrac{qN_s}{2\varepsilon_s}\left(y - d_j - t'_{ox}\right)^2 - E_{o(s,d)}(y - d_j - t'_{ox}) & d_j \le y \le d_j + d_{s,d} \\ 0, & y > d_j + d_{s,d} + t'_{ox} \end{cases} \tag{6}$$

Here $V_S \equiv V_{SB} + V_{bi}$ and $V_D \equiv V_S + V_{DS}$, are the potentials at the source and drain contact regions relative to the substrate interior, V_{bi} is the built-in voltage, and V_{DS} is the drain-source bias. The 1D depletion widths of the two regions are given by

$$d_{s,d} = \sqrt{\frac{2\varepsilon_s V_{S,D}}{qN_s}}. \tag{7}$$

E_{os} and E_{od} are the vertical electric fields at the interface of the source and drain contact regions, respectively, at the effective contact depth $y = d_j$ (see Fig. 1):

$$E_{o(s,d)} = \frac{qN_s}{\varepsilon_s} d_{s,d} = \sqrt{\frac{2qN_s V_{S,D}}{\varepsilon_s}}. \tag{8}$$

Boundary conditions of the Laplacian in Region 1. To solve the Laplacian in Region 1, we must determine the boundary conditions for the three interfaces indicated in Fig. 2.

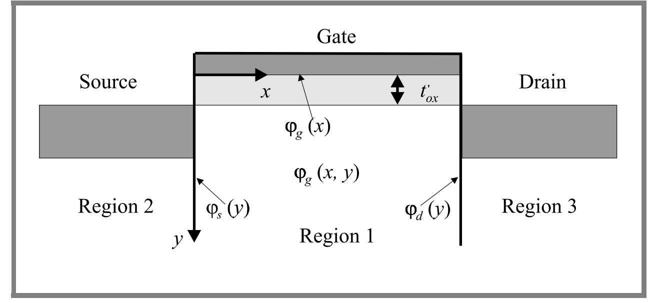


Fig. 2. Boundary conditions for the Laplacian of Region 1.

At the metal-"oxide" interface, we have

$$\varphi_g(x) \equiv \varphi_g(x, 0) = V_{GB} - V_{FB}. \tag{9}$$

At the vertical boundaries, we require the potential to be continuous. Hence, from Eqs. (1) and (5), we obtain for the source and drain side interfaces:

$$\varphi_{s,d}(y) = \phi_{s,d}(y) - \phi_g(y) = $$

$$= \begin{cases} (V_{S,D} - \phi_o)\dfrac{y}{t'_{ox}}, & 0 \le y \le t'_{ox} \\ V_{S,D} - \phi_o + E_o(y - t'_{ox}) - \dfrac{qN_s}{2\varepsilon_s}\left(y - t'_{ox}\right)^2, & t'_{ox} \le y \le d_j + t'_{ox} \\ V_{S,D} - \phi_o + E_{o(s,d)}d_j + \dfrac{qN_s}{2\varepsilon_s} d_j^2 + +\left(E_o - E_{o(s,d)} - \dfrac{qN_s}{\varepsilon_s} d_j\right)(y - t'_{ox}), & d_j + t'_{ox} \le y \le d_g + t'_{ox} \\ V_{S,D} - E_{o(s,d)}(y - d_j - t'_{ox}) + \dfrac{qN_s}{2\varepsilon_s}(y - d_j - t'_{ox})^2, & d_g + t'_{ox} \le y \le d_j + d_{s,d} + t'_{ox} \\ 0, & y > d_j + d_{s,d} + t'_{ox} \end{cases} \tag{10}$$

We can see from Eq. (1) that the intitial 1D estimate for the channel potential is $\Phi_o(x) = \phi_o$, or $\varphi_o(x) = 0$. As will be discussed below, solving for the above boundary conditions, we eventually will arrive at an improved $\Phi_o(x)$ that is properly adjusted for the 2D effects.

3. Conformal mapping

As discussed by Klös and Kostka [15], the Laplacian in the semi-infinite slab denoted as Region 1, can be solved by conformal mapping, given the boundary conditions discussed above. This is done by considering the (x, y) plane as a complex plane, and mapping Region 1 of this plane (see Fig. 2) into the upper half of a transformed, complex (u, v) plane, as shown in Fig. 3. The relative simplicity of the boundary conditions in the transformed plane allows us to derive potential distributions in this plane, from which they can be transformed back to the (x, y) plane.

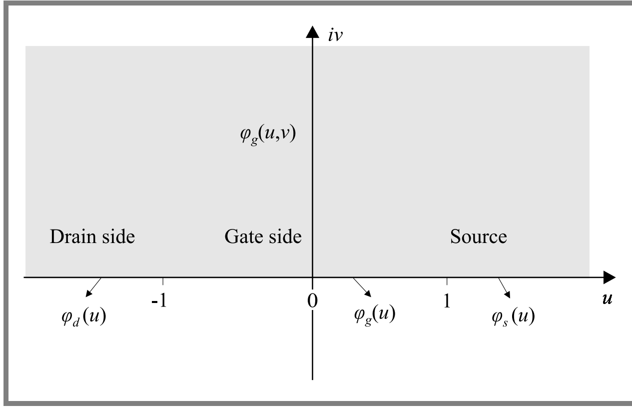


Fig. 3. Conformal mapping of Region 1 of the (x, y) plane into the upper half of the complex (u, v) plane.

The mapping between the two planes is achieved by means of the following Schwartz-Christoffel transformation [16]:

$$z = \int \frac{dz}{dw}\,dw = \frac{L}{\pi}\int \frac{dw}{\sqrt{w-1}\sqrt{w+1}} = \\ = 2\,\frac{L}{\pi}\,\ln\left(\frac{\sqrt{w-1}+\sqrt{w+1}}{\sqrt{2}}\right). \quad (11)$$

The solution of the Laplacian in the (u, v) plane is given by the following integral along the u-axis:

$$\varphi(u, v) = \frac{v}{\pi}\int_{-\infty}^{+\infty} \frac{\varphi(u')}{(u-u')^2+v^2}\,du', \quad (12)$$

where $\varphi(u)$ is the boundary condition transformed to the u-axis. This result can then be transformed back to the (x, y) plane by means of Eq. (11). Note that along the u-axis, the transformation in Eq. (11) can be written as:

$$x = \frac{L}{\pi}\mathrm{Arcos}(u), \quad y = 0, \quad \text{for} \quad |u| \le 1, \quad (13)$$

$$x = 0, \quad y = \frac{L}{\pi}\,\ln\left(|u| + \sqrt{u^2-1}\right) \equiv \\ \equiv \frac{L}{\pi}\,\mathrm{Arcosh}(|u|), \quad \text{for} \quad |u| > 1. \quad (14)$$

For MOSFET modeling, we will only be needing the electric field component E_n normal to the channel. This field includes the 1D electrical field E_o and the 2D vertical field component $E_{2D}(x)$ associated with the Laplacian. Assuming that the oxide thickness is much smaller than the gate length, the latter can be taken to be the same at the channel as at the metal interface. Hence, in the (x, y)-plane, we have:

$$E_n(x) = E_o + E_{2D}(x). \quad (15)$$

The contribution $E_{2D}(u)$ from the Laplacian can easily be obtained from Eq. (12) as the mapped derivative of $\varphi(u, v)$ with respect to v in the limit of $v \to 0$ (see also [15, 16]):

$$E_{2D}(u) = \lim_{v\to 0}\left|\frac{\partial w}{\partial z}\right| \frac{d\varphi(u, v)}{dv} = \\ = \frac{1}{\pi}\left|\frac{\partial w}{\partial z}\right| \int_{-\infty}^{+\infty} \frac{1}{u-u'}\,\frac{\partial \varphi}{\partial u}\bigg|_{u'} du'. \quad (16)$$

The integral in this expression can be solved piecewise for the various sections of the boundaries indicated in Fig. 1. Along the u-axis, the integration limits are as follows:

$$-\infty, -u_4 = -\cosh\left(\frac{\pi(d_j+d_d+t'_{ox})}{L}\right), -u_2, -u_1, -1, \\ 1, u_1 = \cosh\left(\frac{\pi(d_j+t'_{ox})}{L}\right), u_2 = \cosh\left(\frac{\pi(d_g+t'_{ox})}{L}\right), \\ u_3 = \cosh\left(\frac{\pi(d_j+d_s+t'_{ox})}{L}\right), \infty \quad (17)$$

We note that since $\varphi(u) = 0$ for $u < -u_4$ and for $u > u_3$, we have no contributions to the integral from these intervals.

3.1. Some approximations

In order to find analytical solutions to the integral of (16), we have to introduce some additional approximations. Specifically, we have to replace the mapping functions for x and y in the integrand by more manageable functions.
For the channel region $(-1 \le u \le 1)$, we propose to use the following approximation:

$$x = \frac{L}{\pi}\mathrm{Arcos}(u) \approx \\ \approx \frac{L}{\pi}\left[\sqrt{2}\left(\sqrt{1-u} - \sqrt{1+u}\right) + \frac{\pi}{2} + \left(2 - \frac{\pi}{2}\right)u\right]. \quad (18)$$

The transformation and the error of the approximate function are shown in Fig. 4. We note that the maximum error is about 0.2%.
We note that the derivative of x with respect to u has the following exact form:

$$\frac{dx}{du} = -\frac{L/\pi}{\sqrt{1-u^2}}. \quad (19)$$

Outside the channel region $(|u| > 1)$, we have the exact mapping function shown in Eq. (14). Here we propose to use the following approximate function:

$$y = \frac{L}{\pi}\mathrm{Arcosh}(|u|) \approx \\ \approx \frac{L}{\pi} \times \begin{cases} \sqrt{2(|u|-1)}, & \text{for } |u| < u_o \\ \sqrt{2(|u|-1)} - k(|u|-u_o), & \text{for } |u| \ge u_o \end{cases}, \quad (20)$$

where $u_o = 1.2$ and $k = 0.14$.

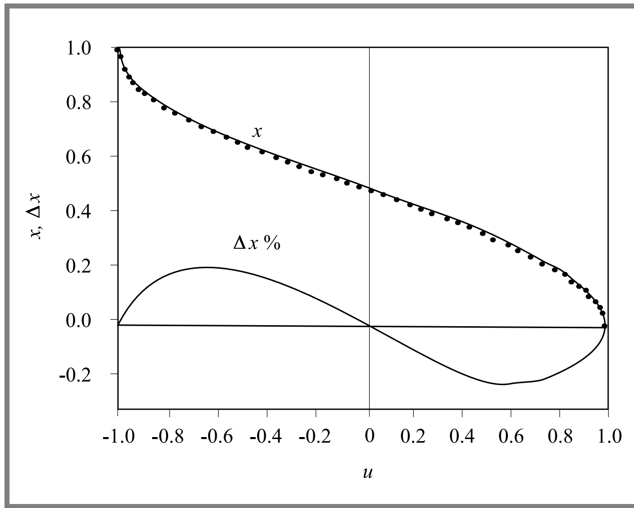


Fig. 4. Comparison between the exact (dots) and the approximate (top solid curve) mapping functions for the channel region. The error of the approximate function is shown in the lower curve.

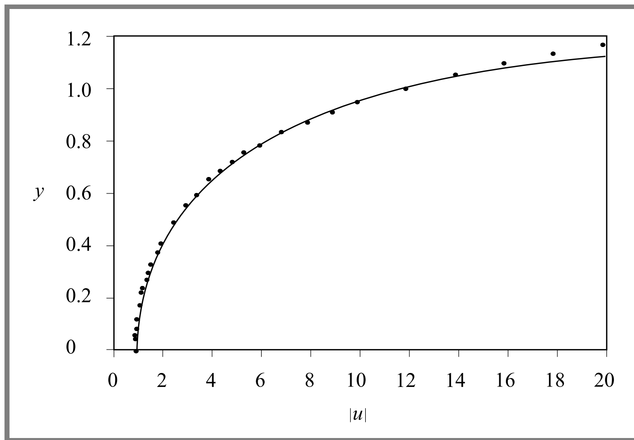


Fig. 5. Comparison between the exact (dots) and the approximate (solid curve) mapping functions for the channel region.

A comparison of these expressions is shown in Fig. 5 for $1 \leq |u| < 20$. We note that the derivative of y with respect to u has the exact form:

$$\frac{dy}{du} = \frac{L/\pi}{\sqrt{u^2-1}}. \tag{21}$$

The error associated with the approximate mapping is very small, less than 0.01 for $|u|$ less than about 15, then increases to about 0.05 at $|u| = 20$, and continues to increase for higher values of $|u|$. However, we note that the contributions from the boundaries to the potential and the electrical field at or near the channel vanishes for values of $|u| > u_3$ (source side) and $|u| > u_4$ (drain side) (see Eqs. (16) and (17) and Fig. 1). For well-designed transistors, typical values of u_3 and u_4 are within the range of u shown in Fig. 5.

3.2. Channel potential profile

Using the above formalism, the partial integrals from all the boundary sections of Eq. (16) have analytical solutions (although with somewhat lengthy expressions in most cases). The transformation from the (u, v)-plane to the (x, y)-plane is straightforward using the inverse of Eqs. (13) and (14) for the coordinate.

In order to determine the channel potential profile, $\Phi_o(x) = \phi_o + \varphi_o(x)$, in the subthreshold regime, we have to consider the layer of mobile channel charge with sheet density $qn(x)$. Using Gauss' law on a small section of this sheet charge, we find the following relationship (see Eqs. (15) and (16)):

$$qn(x) = \left[E_i(x) - E_n(x)\right]\varepsilon_s, \tag{22}$$

where E_i is the total normal field in the "oxide". Again, noting that the oxide thickness is very small compared to the channel length, we can safely assume that the vertical field is constant inside the "oxide", which means that it can be expressed in terms of the difference between the potentials at the metal boundary and at the channel, i.e.:

$$E_i(x) = \left[V_{GS} - V_{FB} - V_{BS} - \Phi_o(x)\right]/t'_{ox}, \tag{23}$$

where V_{GS} is the gate-source voltage. Combining (22) and (23) and using elementary electron statistics, we obtain the following expression for the total potential profile $\Phi_o(x)$ in the channel:

$$q\frac{n_i^2}{N_s}\exp\left(-\frac{\Phi_o(x,0)}{V_{th}}\right) = = \left\{\frac{\left[V_{GS}-V_{FB}-\Phi_o(x)\right]}{t'_{ox}} - E_n(x)\right\}\varepsilon_s, \tag{24}$$

where n_i is the intrinsic carrier concentration and V_{th} is the thermal voltage. We observe that, except for the portions of the channel close to the source and drain, the term on the left side of this expression can be ignored. Hence, for the central part of the channel, we obtain the following potential profile:

$$\Phi_o(x) \approx V_{GS} - V_{FB} - E_n(x)t'_{ox} = = V_{GS} - V_{FB} - \frac{\varepsilon_s}{\varepsilon_i}\left[E_o + E_{2D}(x)\right]t_{ox}. \tag{25}$$

A satisfactory, approximate, analytical solution of Eq. (24), that covers the full length of the channel, is also available using the "generalized diode equation" approach discussed in [18].

3.3. Threshold voltage

We recall that the threshold condition is satisfied when the minimum of the channel potential just reaches the value

$\Phi_{o,\min} = V_{SB} + 2\varphi_b$. From Eq. (25), we find the following expression for the threshold voltage:

$$V_T = V_{FB} + V_{SB} + 2\varphi_b + E_{n,\max} t'_{ox}, \qquad (26)$$

where $E_{n,\max}$ is the maximum of the normal electric field in the channel at threshold.

3.4. Subthreshold current

The subthreshold drain current I_{sub} can be expressed either in terms of a thermionic emission current I_{tem} in short-channel devices, or as a drift-diffusion current I_{dd} for longer channels. To cover a wide range of channel lengths, we may apply the following unified approximation [19], which is always dominated by the "rate limiting" transport mechanism:

$$I_{sub} \approx \left(\frac{1}{I_{tem}} + \frac{1}{I_{dd}} \right)^{-1}. \qquad (27)$$

The thermionic emission current is given by:

$$I_{tem} \approx W \delta A^* T^2 \exp\left(\frac{\Phi_{g,\min} - V_{bi}}{V_{th}} \right), \qquad (28)$$

where $\Phi_{g,\min}$ is the minimum value of the channel potential, δ is the effective thickness of the channel at the potential minimum, W is the width of the channel, A^* is Richardson's constant. An expression for the drift-diffusion current was discussed in [19], for which an approximation applicable to subthreshold MOSFETs was presented in [15].

4. Quarter-micron MOSFET

We first consider conventional n-channel MOSFETs with device lengths of 250 nm and 210 nm by comparing our results with the experimental data by Chung *et al.* [20]. The experimental data used here came from devices with $N_s = 2 \cdot 10^{17}$ cm^{-3} and $t_{ox} = 8.6$ nm, except for the subthreshold characteristics that came from a device with $N_s = 4 \cdot 10^{17}$ cm^{-3} and $t_{ox} = 5.6$ nm. The junction depth in all cases was about 0.16 μm, and the gate-junction overlap was 0.1 μm. The devices from [20] were chosen since they did not include any halo doping or LDD. Hence, the devices displayed the type of short-channel behavior expected also for classical sub-100 nm MOSFETs (see below).

Figure 6 shows comparisons of the modeled and experimental results for the threshold voltage versus applied drain-source bias and versus the gate length, respectively, for MOSFETs with $t_{ox} = 8.6$ nm. To arrive at the experimental data in the latter case, a calculated long-channel threshold voltage of 0.42 V was used.

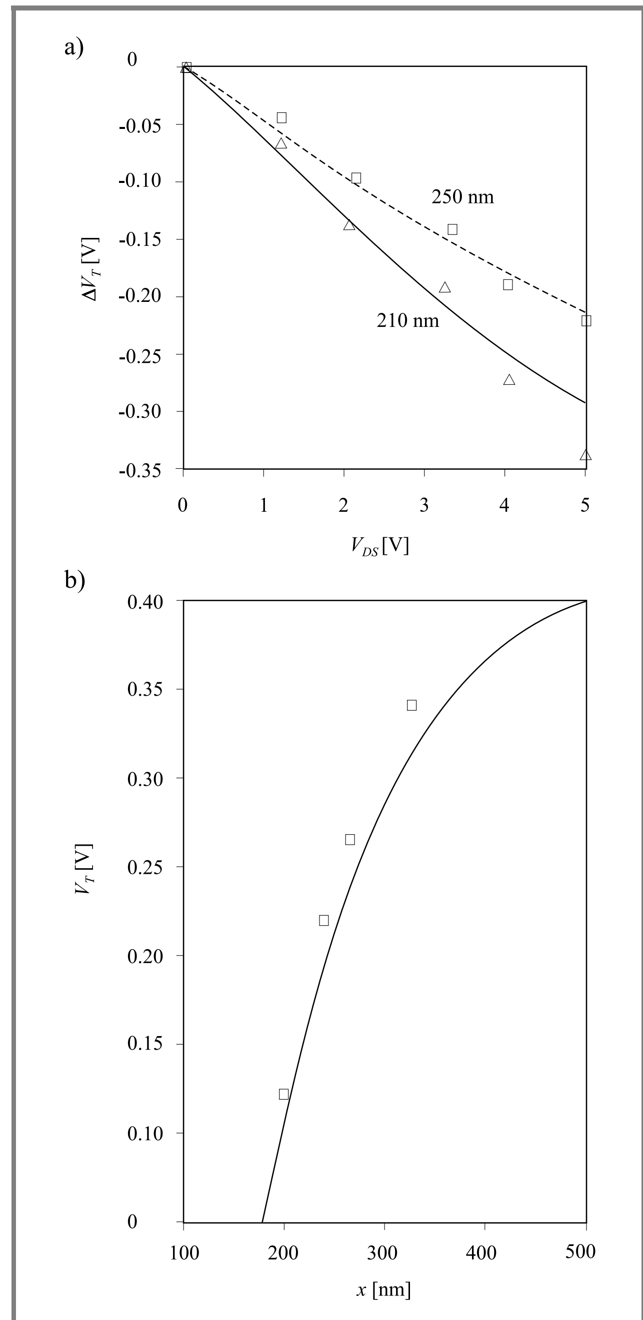


Fig. 6. Comparison of experimental (symbols) and modeled (lines) variation in threshold voltage: (a) versus applied drain-source bias for $t_{ox} = 8.6$ nm MOSFET with 210 nm and 250 nm gate lengths, and (b) versus gate length for $V_{DS} = 0.05$ V.

The modeled threshold voltages were determined from the minimum channel potential for a set of gate-source voltages, selecting the ones that comply with the threshold condition. The calculations were based on processing and geometric parameters for the two devices.

Figure 7 shows the corresponding central parts of the potential distributions in the channel at threshold, calculated using Eq. (27) for $V_{DS} = 0.05$ V and at 3 V.

These curves clearly illustrate the lowering of the threshold voltage (DIBL-effect) related to the reduction of

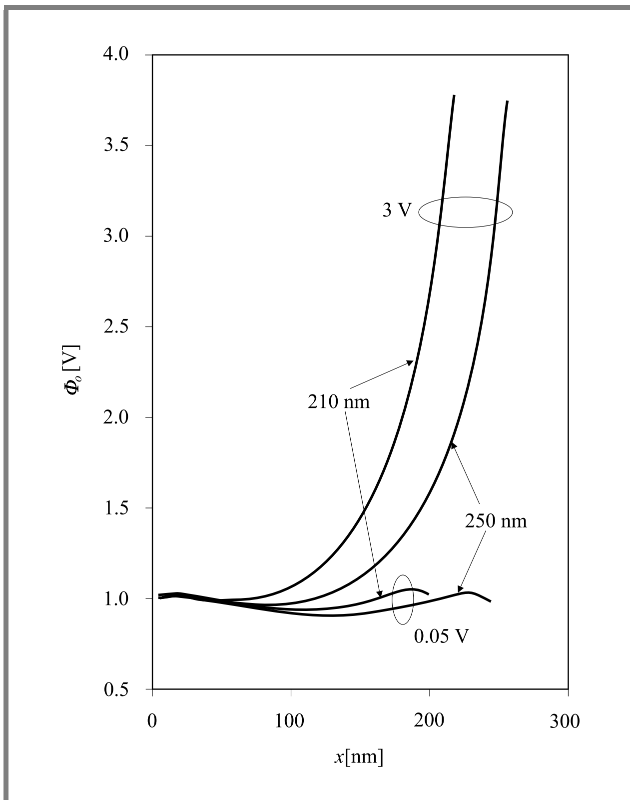


Fig. 7. Model calculations of the channel potential relative to the substrate for $V_{DS} = 0.05$ V and 3 V for gate lengths of 210 nm and 250 nm and $t_{ox} = 8.6$ nm.

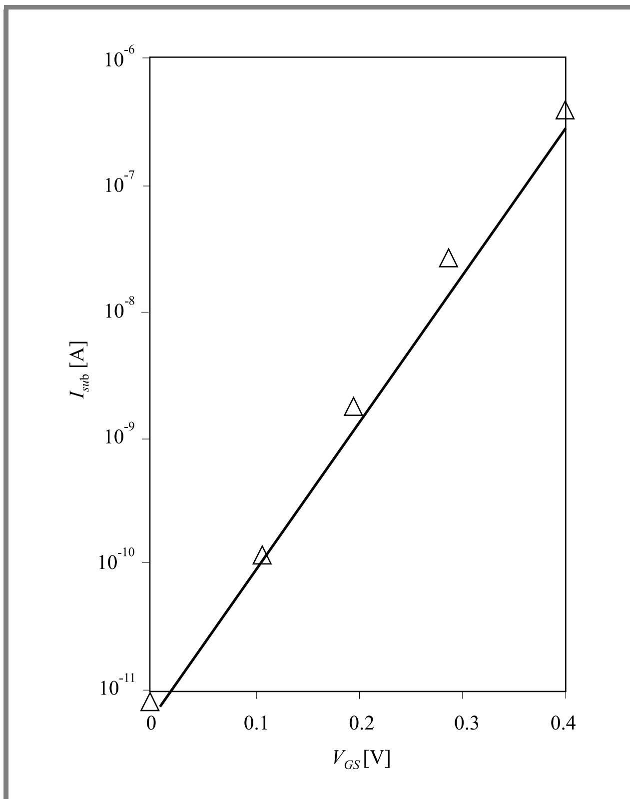


Fig. 8. Measured (symbols) and modeled (line) subthreshold transfer characteristics for a 250 nm MOSFET with and $t_{ox} = 5.6$ nm at $V_{DS} = 0.05$ V.

the energy barrier with increasing drain-source bias. Also the shift of the potential minimum in the direction of the source with increasing drain-source bias is indicated.

A comparison of an experimental and modeled subthreshold transfer characteristics for a 250 nm device with $t_{ox} = 5.6$ nm is shown in Fig. 8 for $V_{DS} = 0.05$ V. Note that this device has a different oxide thickness and doping than the 250 nm device discussed in Figs. 6 and 7. However, except for N_s and t_{ox}, the same parameter values were used in the simulations.

The data shown here indicate that the present 2D modeling strategy is quite suitable for deep submicron MOSFETs operating in the subthreshold regime. The deviations observed can mostly be attributed to the deviations of the processing variables from those reported. As indicated above, the only adjustable parameter in the present modeling is the effective depth source and drain contact, which accounts for the rounding of the contacts towards Region 1 (see Fig. 1). Next, we will also test the approach for an experimental MOSFET in the sub-100 nm range.

5. Sub-100 nm MOSFET

One of the problems of scaling classical MOSFETs into the sub-100 nm range is that the substrate doping density has to be increased into the 10^{18} cm^{-3} range in order to contain the source and drain depletion layers. However, this high doping has several detrimental effects on the MOSFET properties, such as degradation of the channel mobility and too large threshold voltages.

One solution to this problem is to use a much lower doping in the substrate and instead ion implant a higher doping of desired concentration under the surface. After annealing, a thin layer of lower doping concentration remains at the surface, typically to a depth of less than 100 nm. The much higher doping needed to prevent severe short-channel effects and punch-through stretches for another 100 nm or so into the substrate. This is the so-called SSR (super-steep-retrograde) channel MOSFET structure.

Here, we consider such an n-channel MOSFET with a 70 nm gate length, reported by Xu *et al.* [21]. The doping at the surface of this device is about $2.8 \cdot 10^{17}$ cm^{-3} and increases almost exponentially to about $1.8 \cdot 10^{18}$ cm^{-3}

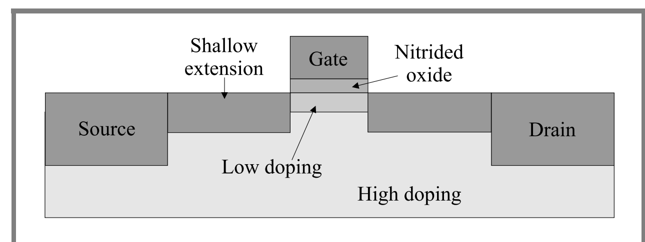


Fig. 9. Schematic view of a super-steep-retrograde channel MOSFET. For the modeling, we assume that the low-doped and the high-doped regions each have a constant doping density.

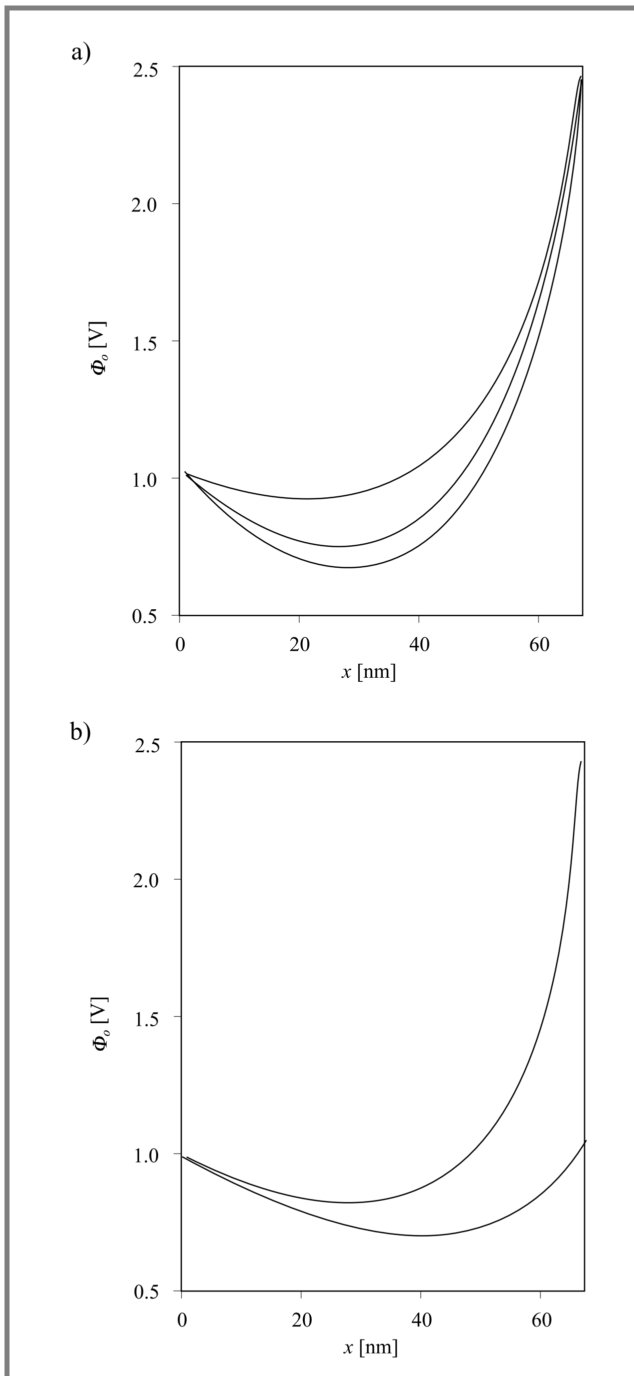


Fig. 10. Model calculations of the channel potential profiles in 70 nm SSR MOSFET relative to the substrate for (a) $V_{DS} = 1.6$ V at $V_{GS} = 0$ V (lower curve), 0.1 V (middle curve), and 0.31 V (upper curve), and (b) $V_{GS} = 0.1$ V at $V_{DS} = 0$ V (lower curve) and 1.5 V (upper curve).

at a depth of 60 nm. It also has the added benefit of very shallow source and drain extensions, with a thickness of about 50 nm, realized by means of a SALICIDE process.

The variable substrate doping of the SSR channel MOSFET is modeled as a two-layer structure, as indicated in Fig. 9, where each layer has a constant doping. The thickness and the doping concentration in the shallow top region are chosen to be 35 nm and $9.2 \cdot 10^{17}$ cm^{-3}, respectively, which accounts for all the doping atoms in the graded region near the surface. At threshold and zero drain-source bias, the depletion layer stretches into the highly doped region. The thickness of the highly doped layer is such that it is never penetrated by the gate depletion layer.

The modeling of the SSR MOSFET proceeds as described above for the 250 nm MOSFET, except that the vertical boundaries of the central region under the gate (Region 1) now include an extra section to account for the shallow layer of reduced substrate doping. Region 1, where the Laplacian is defined, is still assumed to include the gate oxide and a semi-infinite slab of semiconductor below (see Fig. 2).

5.1. Modeling results for 70 nm SSR MOSFET

Again, the threshold voltages were determined from the minimum channel potentials for a set of gate-source voltages, selecting the ones that comply with the threshold condition. Figure 10 shows the calculated channel potential profiles for some combinations of drain-source and gate-source biases in the subthreshold regime.

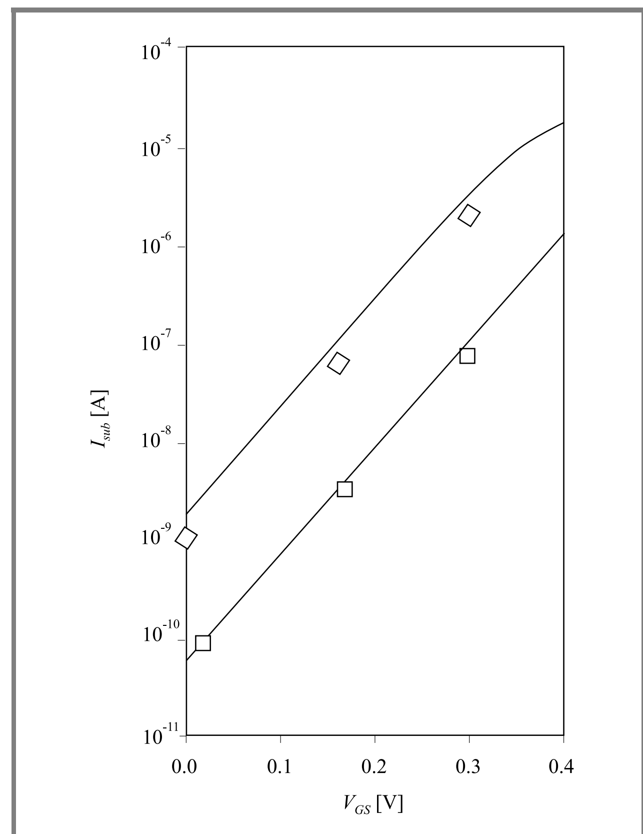


Fig. 11. Experimental (symbols) and modeled (lines) subthreshold transfer characteristics for a 70 nm SSR MOSFET with $t_{ox} = 3$ nm. $V_{DS} = 0.1$ V (lower curve) and 1.6 V (upper curve).

At zero drain-source bias we obtain the threshold voltage $V_T = 0.3$ V, in close agreement with the value reported in [21]. From the calculated variation of the threshold voltage with the applied drain-source bias, we find a DIBL parameter of $\sigma = \Delta T/V_{DS} = 80$ mV/V. This value is reasonably close to the value of about 70 mV/V obtained from the experimental subthreshold transfer NMOS characteristics of [21].

Figure 11 shows a comparison of the experimental and modeled subthreshold characteristics of the 70 nm SSR MOSFET.

6. Summary

A closed-form 2D modeling technique for deep-submicron and sub-100 nm MOSFETs has been investigated. The technique is based on conformal mapping, where the 2D Poisson equation in the depletion regions is separated into a 1D long-channel case and a 2D Laplace equation. From the straightforward 1D solution, the boundary potential values of the Laplacian is obtained, from which a 2D correction to the channel potential is derived. The model has been tested for classical MOSFETs with gate lengths in the range 200–250 nm, and for a super-steep retrograde MOSFET with a gate length of 70 nm. With a minimum set of parameters, the present modeling reproduces both qualitatively and quantitatively the experimental data obtained for such devices.

This method applies to classical MOSFETs as well as to non-classical structures. As shown here, the classical MOSFET approach is directly applicable to SSR MOSFETs, and may also be applied to Si-Si/Ge strained devices and to sidewall vertical MOSFETs with a partially depleted body. We foresee that the same approach, with somewhat adjusted boundary conditions for the Laplacian, may also be applicable to double-gate fully depleted vertical MOSFETs and to double-gate SOI MOSFETs.

References

[1] Y. Cheng and C. Hu, *MOSFET Modeling & BSIM3 User's Guide*. Boston: Kluwer, 1999.

[2] W. Liu and C. Hu, "BSIM3V3 MOSFET model", in *Silicon and Beyond: Advanced Device Models for Circuit Simulators*, M. S. Shur and T. A. Fjeldly, Eds. Singapore: World Scientific, 2000, pp. 1–31.

[3] W. Liu *et al.*, *BSIM4 User's Manual*. Berkeley: University of California, 2000.

[4] T. Ytterdal, Y. Cheng, and T. A. Fjeldly, *Device Modeling for Analog and RF CMOS Circuit Design.* London: Wiley, 2003.

[5] C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications", in *Analog Integrated Circuits and Signal Processing.* Boston: Kluwer, 1995, vol. 8, pp. 83–114.

[6] R. Velghe, D. Klaassen, and F. Klaassen, "MOS Model 9". Unclassified Report NL-UR 003/94, Philips Electronics N.V, 1994.

[7] K. Lee, M. Shur, T. A. Fjeldly, and T. Ytterdal, *Semiconductor Device Modeling for VLSI.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[8] T. A. Fjeldly, T. Ytterdal, and M. Shur, *Introduction to Device Modeling and Circuit Simulation*. New York: Wiley, 1998.

[9] ITRS – International Technology Roadmap for Semiconductor. Austin: Semiconductor Industry Assoc., 2001.

[10] K. N. Ratnakumar and J. D. Meindl, "Short-channel MOSFET threshold voltage model", *IEEE J. Solid-State Circ.*, vol. 17, pp. 937–948, 1982.

[11] Y. A. El-Mansy and A. R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region", *IEEE Trans. Electron Dev.*, vol. 40, pp. 254–262, 1977.

[12] V. K. De and J. D. Meindl, "An analytical threshold voltage and subthershold current model for short-channel MESFETs", *IEEE J. Solid-State Circ.*, vol. 28, pp. 169–172, 1993.

[13] S.-H. Oh, D. Monroe, and J. M. Hergenrother, "Analytic description of short-channel effects in fully-depleted double gate and cylindrical, surrounding-gate MOSFETs", *IEEE Electron Dev. Lett.*, vol. 21, no. 9, 2000.

[14] Y.-S. Pang and J. R. Brews, "Analytical subthreshold surface potential model for pocket n-MOSFETs", *IEEE Trans. Electron Dev.*, vol. 49, pp. 2209–2216, 2002.

[15] A. Klös and A. Kostka, "A new analytical method of solving 2D Poisson's equation in MOS devices applied to threshold voltage and subthreshold modeling", *Solid-State Electron.*, vol. 39, pp. 1761–1775, 1996.

[16] E. Weber, *Electromagnetic Fields*, vol. 1: *Mapping of Fields*. New York: Wiley, 1950.

[17] P. M. Zeitzoff, J. A. Huchby, and H. R. Huff, "MOSFET and front-end process integration: scaling trends, challenges, and potential solutions through the end of the roadmap", in *Int. J. High Speed Electron. Syst.*, 2002, vol. 12, pp. 267–293.

[18] T. A. Fjeldly, B. J. Moon, and M. S. Shur, "Approximate analytical solution of the generalized diode equation", *IEEE Trans. Electron Dev.*, vol. 38, pp. 1976–1977, 1991.

[19] T. A. Fjeldly and M. S. Shur, "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs", *IEEE Trans. Electron Dev.*, vol. 40, pp. 137–145, 1993.

[20] J. E. Chung, M.-C. Jeng, J. E. Moon, P.-K. Ko, and C. Hu, "Performance and reliability design issues for deep-submicrometer MOSFET's", *IEEE Trans. Electron Dev.*, vol. 38, no. 3, pp. 545–554, 1991.

[21] Q. Xu, H. Qian, H. Yin, H. Ji, B. Chen, Y. Zhu, M. Liu, Z. Han, H. Hu, Y. Qiu, and D. Wu, "The investigation of key technologies for sub-0.1-μm CMOS device fabrication", *IEEE Trans. Electron Dev.*, vol. 48, no. 7, 2001.

Jarle Østhaug is a master degree student at the Physics Department, University of Oslo, Norway, and UniK, the University Graduate Center at Kjeller, Norway. He is expected to graduate in 2004.

e-mail: jarleo@unik.no

UniK – University Graduate Center

N-2027 Kjeller, Norway

Tor A. Fjeldly received his M.Sc. degree from Norwegian Institute of Technology (1967) and his Ph.D. degree from Brown University (1972). He was a Scientist at Max-Plack Institute for Solid State Physics, Germany (1972–1974) and at SINTEF research organization, Norway (1974–1983). He has been Professor at the Norwegian University of Science and Technology since 1973, and is presently at UniK – the University Graduate Center, Norway. He is author of about 200 scientific and technical works, including three books, and is editor of seven books. His current research interests are electronic device modeling and circuit simulation. Dr. Fjeldly is Fellow of IEEE.
e-mail: torfj@unik.no
UniK – University Graduate Center
N-2027 Kjeller, Norway

Benjamin Iñiguez received his B.S., M.S. and Ph.D. degrees in physics from the University of the Balearic Islands (UIB), Spain, in 1989, 1992 and 1996, respectively. From 1997 to 1998 he was a postdoctoral research scientist at the ECSE Department, Rensselaer Polytechnic Institute, Troy, NY. From 1998 to 2001 he was a Research Scientist (Postdoctoral Marie-Curie Grant Holder) in the Microelectronics Laboratory, Université catholique de Louvain (UCL), Belgium. In February 2001 he joined the Universitat Rovira i Virgili (URV), Tarragona, Spain, as Associate Professor. His current research interests are characterization and modeling of advanced electron devices.
e-mail: binyigue@etse.urv.es
Universitat Rovira i Virgili (URV)
Tarragona, E-43001, Spain

Invited paper

New approach to power semiconductor devices modeling

Andrzej Napieralski and Małgorzata Napieralska

Abstract — The main problems occurring during high power device modeling are discussed in this paper. Unipolar and bipolar device properties are compared and the problems concerning high time-constant values related to the diffusion phenomena in the large base are explained. Traditional and novel concepts of power device simulation are presented. In order to make accurate and modern semiconductor device models widely accessible, a website has been designed and made available to Internet users[1], allowing them to perform simulations of electronic circuits containing high power semiconductor devices. In this software, a new distributed model of power diode has been included. Together with the existing VDMOS macromodel library, the presented approach can facilitate the design process of power circuits. In the future, distributed models of IGBT, BJT and thyristor will be added.

Keywords — power device modeling, SPICE, circuit simulation, VDMOS, PIN diode, IGBT, web-based simulation.

1. Introduction

Many CAD programs have been developed in microelectronics and successfully applied to circuit analysis. The question arises, why there are no CAD programs for power-circuit analysis? In this paper, we will discuss the need for a development of a new kind of power semiconductor device models. Additionally, we will present a new concept of a free, Internet-based software development, dedicated to power semiconductor circuit analysis. Until now, high prices and hardware requirements limited the access to professional CAD tools for the majority of educational institutions, students and small enterprises. Free versions of commercial software have limited functionality and do not contain the modern numerical algorithms and the libraries of device models.

This is especially true in the case of power electronics, as it requires advanced device models to obtain simulation results of good accuracy. Such models are not available now, even in the commercial versions of the CAD software.

The increasing popularity of the Internet can help to solve this problem. Wilamowski, Malinowski and Regnier [33] were probably the first to see the possibility of performing circuit simulations over the Internet by means of dedicated Internet applications. Simulation software may run on remote servers and results may be sent to the user in the form of numerical data or graphics. Wilamowski *et al.* [34] have put special emphasis on the pay-per-use access to simulation software and platform-independent user interface provided by a web page. In this paper, a development of a free simulation environment is discussed that includes new numerical algorithms and modern device models. The Internet is not only considered as a data transmission system but first of all as a medium for international cooperation of scientific centres, enterprises and individual users that would enable the development of better and freely accessible simulation tools and models [27–30, 34].

[1] http://www.dmcs.p.lodz.pl/dmcs-spice

2. Power device modeling

The main parameters of power semiconductor devices are:

– voltage-blocking capability,

– current capability,

– switching performance,

– safe-operation limits.

The basic parameters of modern power devices are presented in Table 1. As can be seen from this table, the switching performance of a VDMOS (a unipolar device) is very good, but its current capability and maximum blocking voltage are much smaller than in the case of bipolar devices (GTO [14, 31] and GCT). The IGBT transistor is a kind of fusion of the best properties of unipolar and bipolar devices. Voltage and current rating is much better than in the case of VDMOS and switching properties are much better than in the case of GTO and GCT.

Table 1
The basic parameters of the most commonly applied modern power devices

Parameters	VDMOS	IGBT	GTO	GCT
V_M [V]	1000	4500	6000	6000
I_T [A]	350	900	6000	6000
t_{ON} [μs]	0.1	0.2	10	–
Storage time [μs]	0	6	30	5
f_T [kHz]	2000	100	1	1
V_{ON} [V]	Very high	Low	Very low	Very low

Figure 1 presents the product of device V-I ratings for power semiconductor devices. The solid line represents the present state and the dotted line the future trends in power semiconductor device development.

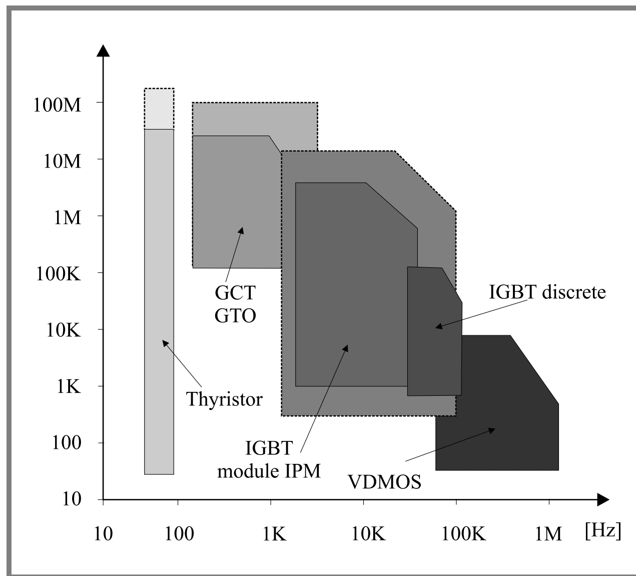


Fig. 1. Product of device V-I ratings [VA] as a function of operating frequency.

In all types of power devices, the blocking capability is a function of the large base width and its doping concentration. Expression (1) gives the value of the maximum blocking voltage (breakdown voltage) as a function of the large base doping concentration N_D:

$$V_B = 60\left(\frac{10^{16}}{N_D}\right)^{\frac{3}{4}}, \tag{1}$$

where: N_D – doping concentration [cm^{-3}],
V_B – blocking voltage [V].

Expression (2) gives the value of the large-base width as a function of blocking voltage V_B and doping concentration N_D without taking the punch trough effect into account:

$$W = \sqrt{\frac{2\varepsilon(\Phi + |V|)}{qN_D}} \cong \sqrt{\frac{2\varepsilon|V|}{qN_D}}, \tag{2}$$

where: W – large base width.

The conclusion from these two expressions is that for a high value of blocking voltage, the semiconductor devices must have a low doping and a large width of the base.

The necessary base width values have been calculated from expressions (1) and (2) for three different values of large base doping with the corresponding maximum blocking voltage V_B. The results are presented in Table 2.

The most important conclusion from this table is that for the high power semiconductor devices the voltage must be supported in a large base. Therefore, the charge transport through the base cannot be instantaneous and the voltage drop over such a structure cannot be neglected. Let us consider now two types of power devices – a unipolar one and a bipolar one. In the case of the unipolar device, the current capability, or in other words, the value of R_{ON} resistance, can be found from a very simple consideration depicted in Fig. 2.

Table 2
The maximum blocking voltage
and necessary base width for three different values
of large-base doping concentration

Device	N_D	V_B [V]	W [μm]
1	$1.25\cdot10^{14}$ cm^{-3}	1600	125
2	$1\cdot10^{14}$ cm^{-3}	1895	157
3	$1\cdot10^{13}$ cm^{-3}	10700	1175

According to Table 2, for a given value of the large base doping concentration, the corresponding maximum blocking voltage V_B and the necessary base width can be calculated. In the case presented in Fig. 2, the value of N_D is equal to $1.25\cdot10^{14}$ cm^{-3}. The corresponding maximum blocking voltage is 1600 V and the base width is equal to 125 μm. Current density in the case of the unipolar device can be expressed by:

$$J_n = q\mu_n nE = q\mu_n n\frac{V}{W}, \tag{3}$$

where: E is the electric field, n is the carrier concentration (equal to N_D in this case), μ_n is the electron mobility, q is the elementary charge.

Taking into account Fig. 2, the value of R_{ON} for the unipolar device having the surface equal to 1 cm^2 can be found from the following expression:

$$R_{\text{cm}^2} = \frac{V}{J_n\text{cm}^2} = \frac{W}{q\mu_n N_D\text{cm}^2}, \tag{4}$$

In the case of the device from Fig. 2, this value will be equal to 0,462 Ω. In the case of high current this value of on-state resistance is definitely too high. Even for a relatively low current of 10 A, the voltage drop will be 46 V. We can thus conclude that unipolar devices cannot be applied in the case of high voltages. This conclusion is in accordance with the real device parameters presented in Table 1.

In the case of a bipolar device, the situation is completely different. Two mechanisms are responsible for the current density and it can be expressed by a set of the following equations:

$$J_n = q\mu_n nE + qD_n\frac{\partial n}{\partial x}, \tag{5a}$$

$$J_p = q\mu_p pE - qD_p\frac{\partial p}{\partial x}, \tag{5b}$$

$$J = J_n + J_p + \varepsilon\frac{\partial E}{\partial t}, \tag{5c}$$

where: D_n and D_p are electron and hole diffusion constants, respectively, and ε is the permittivity.

As can be seen from Eqs. (5a–5c), the current capability in the case of bipolar devices is higher than in the previous case. Due to the diffusion, the maximum current density is much higher and consequently the corresponding voltage drop much lower than in the case of unipolar devices. Additionally, in the bipolar structure some effects resulting from the changes in the effective base doping can appear, for example: the base widening effect, and thyristor effect rendering in the additional current capability.

Now, we will analyze how the mechanism of charge transport through the large base influences the switching performance of the device.

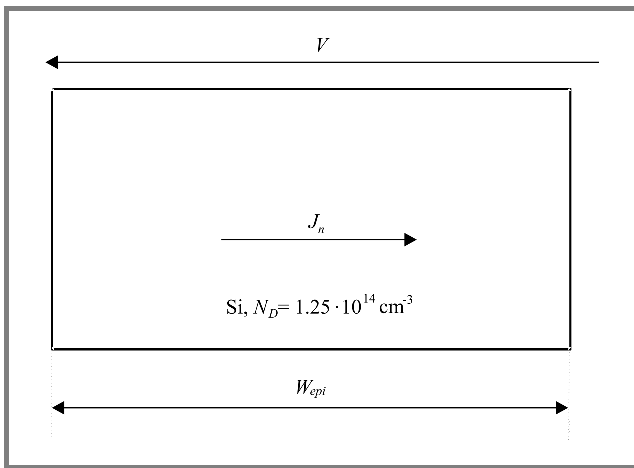


Fig. 2. Illustration of R_{ON} calculation in the case of unipolar devices.

According to Fig. 2 the charge transport time through the large base in the case of the unipolar device can be expressed by:

$$t = \frac{W_{epi}}{v_{nsat}}, \tag{6}$$

where: v_{nsat} is the electron saturation velocity:

$$v_{nsat} = 10^7 \ \frac{\text{cm}}{\text{s}} \left[\frac{T}{300\ \text{K}}\right]^{-0.87}. \tag{7}$$

In the case of power devices operating in conditions of high electric field, one can assume that all the charge carriers are able to attain their maximum speed. For the temperature equal to 300 K, the transit time through the large base can be easily calculated. In Table 3 transit times for three values of base width are presented.

As can be seen from Table 3, in the case of unipolar devices even for a very large base the transit time is very short and, in the majority of cases, the internal time constant of the device can be neglected, compared with the time constant of the external circuit.

The situation is different in the case of bipolar devices. As a first order simplification, we can assume that the transit time of minority carriers through a large base can be expressed by [25]:

$$\tau_D = \frac{W_{epi}^2}{2D} = \frac{W_{epi}^2}{4D_n}, \tag{8}$$

where: $D = 2D_n$ at a high injection level.

Table 3
Electron transit time through the large base of a unipolar device for three different base widths

Device	W [μm]	Transit time [ns]
1	125	1.25
2	157	1.57
3	1175	11.75

In Table 4 the transit time for three values of base width is presented.

Table 4
Electron transit time through the large base of a bipolar device for three different base widths

Device	W [μm]	Transit time [μs]
1	125	1.12
2	157	1.76
3	1175	98.62

In this case the transit time through the large base is very long and the internal time constant of bipolar devices can be much higher than those of the external circuit. In such a case, a simple lumped model cannot be used any longer. It is necessary to take into account the carrier transport inside the bipolar devices and the distributed model should be applied. Until now such a model did not exist in the circuit CAD simulation programs.

3. Contemporary methods of power device simulation

Because of the problems mentioned above, correct power device simulation is not possible in standard simulation programs. The first question is why power-device models are necessary? Their main areas of application are as follows:

- power integration – simulation of power functionality in integrated circuits [24],
- correct simulation of power dissipation inside power semiconductor devices (with special emphasis on the losses during the switching period),
- computation of the temperature distribution inside the semiconductor structure,

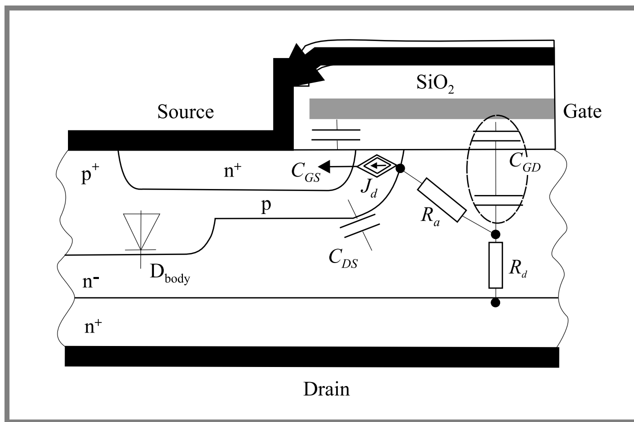


Fig. 3. The internal structure of VDMOS power transistor.

– correct design of the cooling environment,

– prediction of the second thermal breakdown.

In the case of unipolar devices, such as VDMOS (Fig. 3), minority carriers are responsible for current conduction in the normal mode of operation. Therefore, it is possible to build a relatively simple macromodel [17–23, 25] taking into account all the internal elements of the presented structure.

A macromodel of this type, developed for SPICE-like simulators, is presented in Fig. 4 [17–23]. In the case when the body diode D_{GD} is not conducting (VDMOS is not reverse biased), this model is quite accurate and can be applied for a simulation of all types of circuits. The situation is different when the body diode D_{GD} starts to conduct current and diffusion phenomena have to be considered.

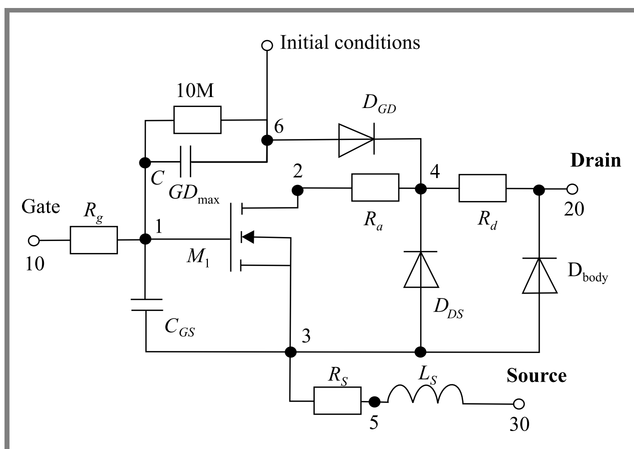


Fig. 4. VDMOS SPICE-like macromodel [2, 23].

In Fig. 5 a half-bridge circuit configuration is presented. The corresponding results of the simulation and measurement are shown in Fig. 6. The voltage and current waveforms are in good agreement with the experiment, and even the power dissipation inside the structure can be correctly predicted.

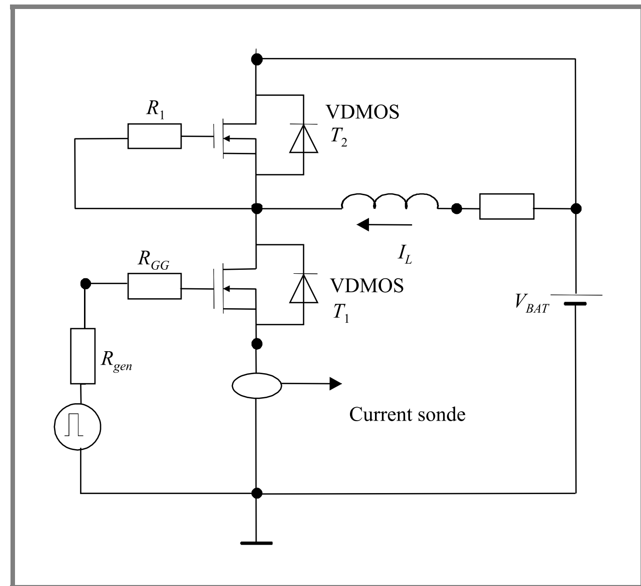


Fig. 5. Half-bridge circuit configuration [23].

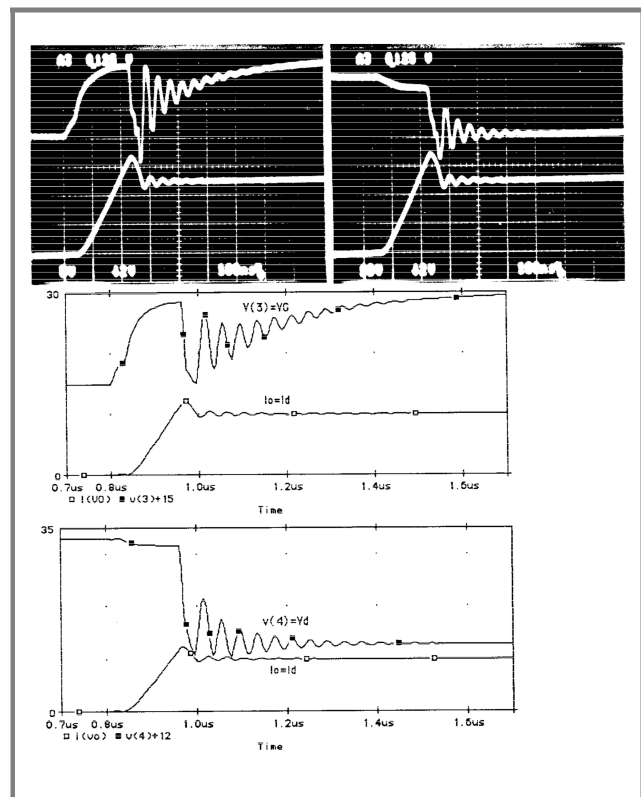


Fig. 6. The simulation and measurement results of a half-bridge with VDMOS transistors [23].

The above considerations lead us to the conclusion that in the case of unipolar devices we have:

– no problem with the diffusion phenomena (minority carriers only),

– very good accuracy obtained by means of lumped models and fast simulation,

– clear physical/geometrical interpretation of model parameters,

– distributed models are necessary only to model the operation of the body diode D_{GD}.

Such a simple approach cannot be applied in the case of a bipolar structure, such as PIN diode, thyristor or IGBT. In Fig. 7, the cross-section of an IGBT elementary cell is presented. All the important phenomena occur in the large base (in this case in the n^- region). In this punch-through IGBT transistor, the large base is relatively short, but even in such a case, one cannot neglect the minority carrier transport phenomena.

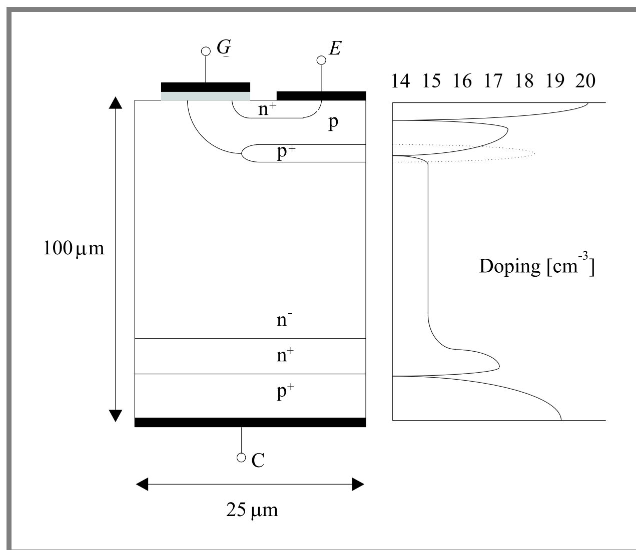


Fig. 7. The internal structure of IGBT power transistor (cross-section of an elementary cell) [5–8].

The most commonly adopted approach to solve this problem is the application of a finite element, finite difference or finite box method.

In Fig. 8 the discretisation of one IGBT cell for 2D simulation is presented. The application of such models enables not only correct current and voltage response simulation but even evaluation of the internal power losses, energy dissipation, temperature distribution (with an additional thermal submodel [6, 10–13]) and current density inside the structure.

As an example of this type of simulation, collector-emitter voltage and collector current as well as dissipated power and energy losses during IGBT switching are presented in Figs. 9 and 10.

The application of an additional thermal model enables the computation of the temperature rise inside the semiconductor structure. This additional thermal model has to be a three-dimensional one. In Fig. 11, the time waveforms of

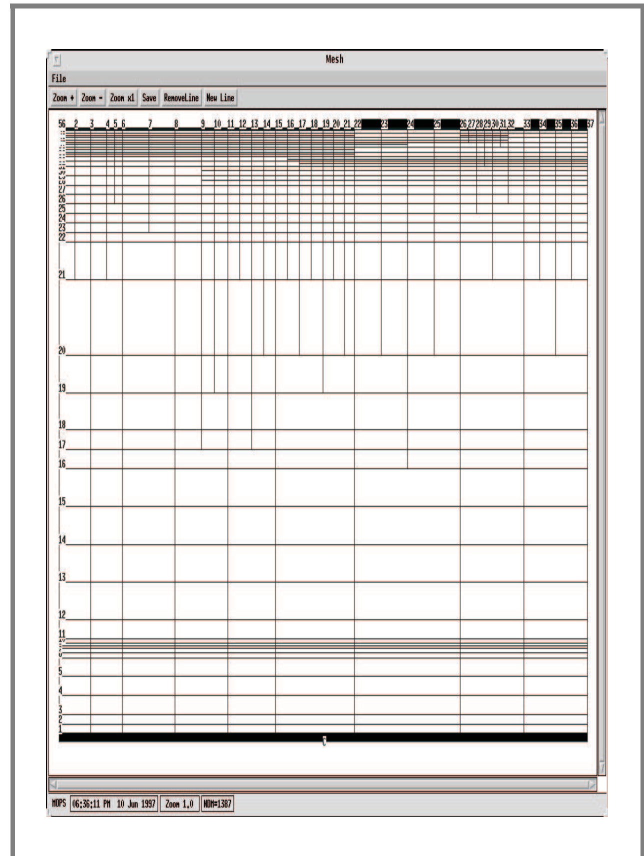


Fig. 8. Discretisation mesh used during 2D analysis of an IGBT structure [6].

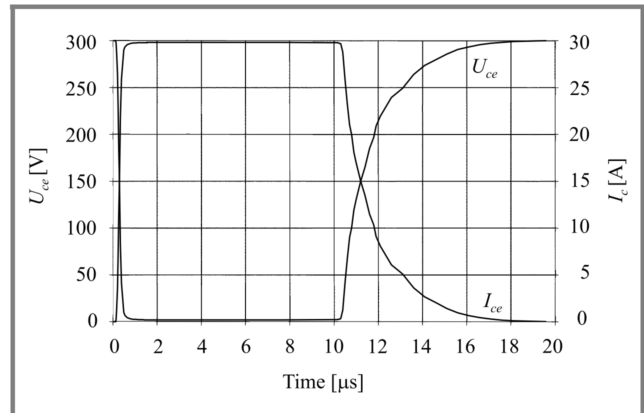


Fig. 9. V_{CE} and I_{CE} during IGBT switching [6].

total power dissipation and the maximum temperature rise inside the studied IGBT structure, are presented.

The main drawback of such an approach is the very long time of simulation and difficulty in simulating more than one or two power devices in the same circuit. The power circuits designers need relatively simple models which can be applied in standard SPICE-like circuit simulation programs with a very short time of simulation and the possibility of high power circuit simulation. In the next section a new approach to this problem will be presented.

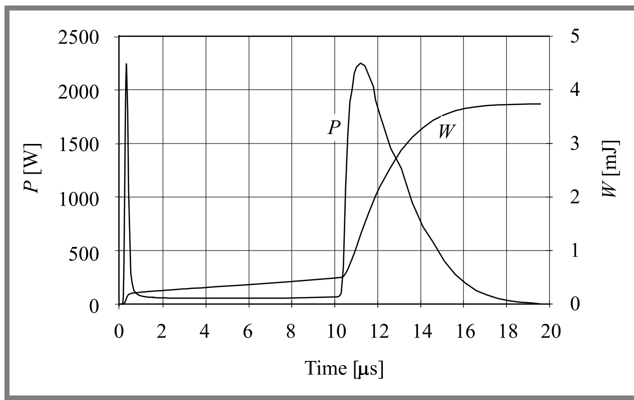


Fig. 10. Power and energy losses during IGBT switching [6].

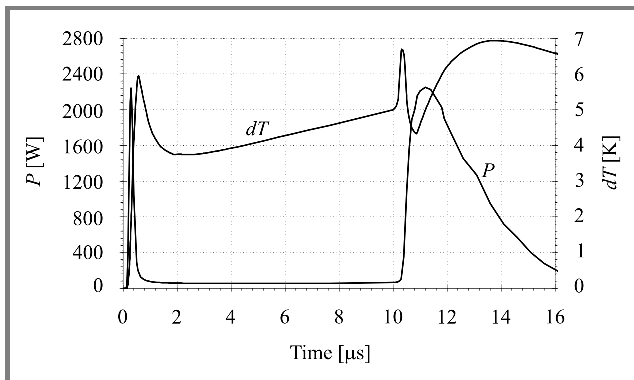


Fig. 11. Time waveforms of total power dissipation and maximum temperature rise inside the IGBT structure [6].

4. Developing the new type of bipolar power device model

Lumped semiconductor devices models are applied in all versions of SPICE programs. Such models can be used in the case of low-power devices, where the base is relatively short and device internal time constants are much lower than those of the external circuit. For the reasons that will be explained later, our model will be implemented in SPICE3F5 simulator.

In order to enable simulation of circuits containing power semiconductor devices, we will try to develop separate device models, taking into account the distributed nature of the phenomena occurring in bipolar power devices and determining their dynamic response. In this section, a model of PIN diode will be presented [27]. Its development is the key point for future design of this type of models for BJT, IGBT or thyristor.

In order to simplify the modeling procedure, we will apply the so-called "modular" approach. In the semiconductor structure considered here, we will distinguish several regions of different physical and/or electrical nature. Then, a simplified sub-model will be assigned to each of them. This approach allows for decreasing the simulation time considerably without any important loss of accuracy [4, 15].

As explained above, in the case of power semiconductor devices the most important is the large base region, which assures high voltage blocking capability during the off-state and where the excess carriers are stored during the on-state (Fig. 12). The simplest device in which one can consider all important physical phenomena is the PIN diode. The well known, one-dimensional Benda-Spenke model [1] has been adopted for this purpose. The behaviour of stored charge carriers can be described by means of the ambipolar diffusion equation:

$$\frac{\partial^2 p(x,t)}{\partial x^2} - \frac{1}{D}\left(\frac{p(x,t)}{\tau} + \frac{\partial p(x,t)}{\partial t}\right) = 0 \tag{9}$$

where: p is the carrier concentration, D is the ambipolar diffusion constant and τ is the common electron and hole lifetime.

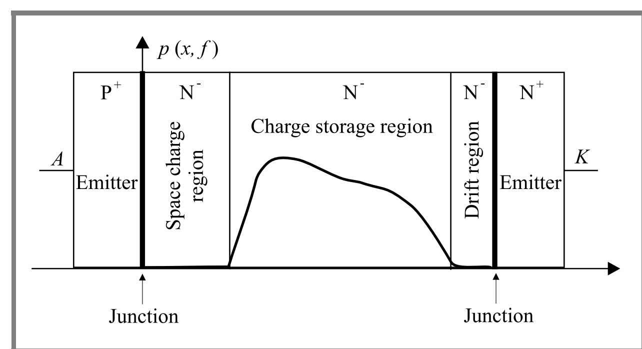


Fig. 12. Modular modeling concept and its application to PIN diode.

Equation (9) cannot be solved analytically, therefore many different approaches using numerical solutions have been presented in the literature. The proposed model is based on the algorithmic approach, i.e. the solution is obtained with a numerical algorithm. A new efficient centered weighted essentially non-oscillatory (CWENO) scheme [16] is used to ensure high computational stability and good accuracy. After the solution of Eq. (9) is obtained, the negative voltage drop in the space charge region can be calculated applying Poisson's equation [29]:

$$\frac{dE(x)}{dx} = -\frac{\rho(x)}{\varepsilon}, \tag{10}$$

where: E is the electric field, ρ is the charge density and ε is the permittivity.

In order to solve this nonlinear equation, the Newton-Raphson method has been applied [29].

5. Simulation environment

In the proposed approach, the simulation software runs on a network server and the user interface is provided by a web page ensuring data entry point and result presentation. One should be carefully consider operations to be performed on

the client side and on the server side [30, 33]. Current server performance and network bandwidths enable all the operations to be performed on the server side. The user receives simulation results in the form he requested and all he needs to use to develop a simulation environment is a web browser. This makes the proposed solution as portable and platform-independent as possible, which can ease the cooperation between different users. However, in some situations it may be more suitable to do some data processing on the client side. Thus, we are considering a development of some client-side software that could be used when desired.

The designed simulation environment comprises four main modules as illustrated in Fig. 13. Computational resources are provided by means of Apache server running under Linux operating system. Nevertheless, the code is portable to Windows and Unix operating systems. Circuit analysis is performed with a batch-executed simulator that may include additionally implemented device models. Simulation results are processed with GNUPlot, providing graphical data representation. Finally, graphical user interface functions (circuit description and simulation parameters input point as well as results visualization) have been implemented in PHP code that dynamically generates HTML pages rendered by the user's web browser (Fig. 14).

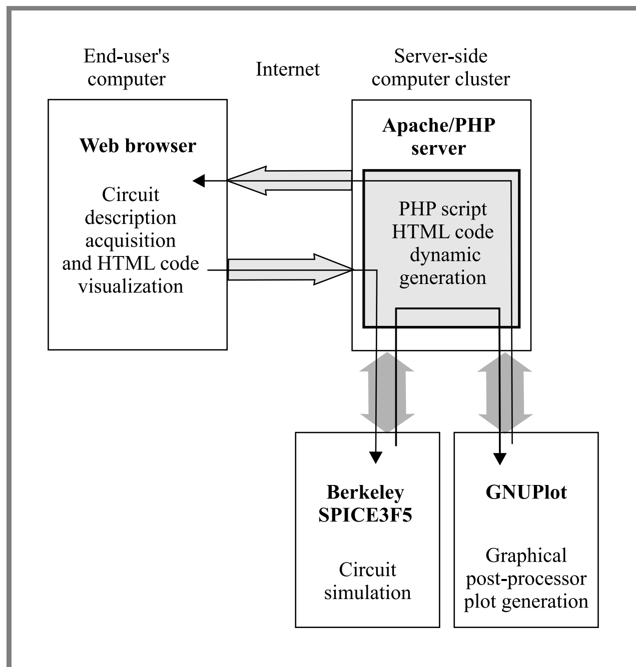


Fig. 13. Structure and data flow of the developed simulation environment.

Thanks to the proposed solution, simulation and data processing can be performed on dedicated servers, therefore not engaging the computers of end-users. Another advantage is that no additional software has to be installed on the end-user side. In order to ensure free access to the environment, it has been based on open source and GNU-licenced software.

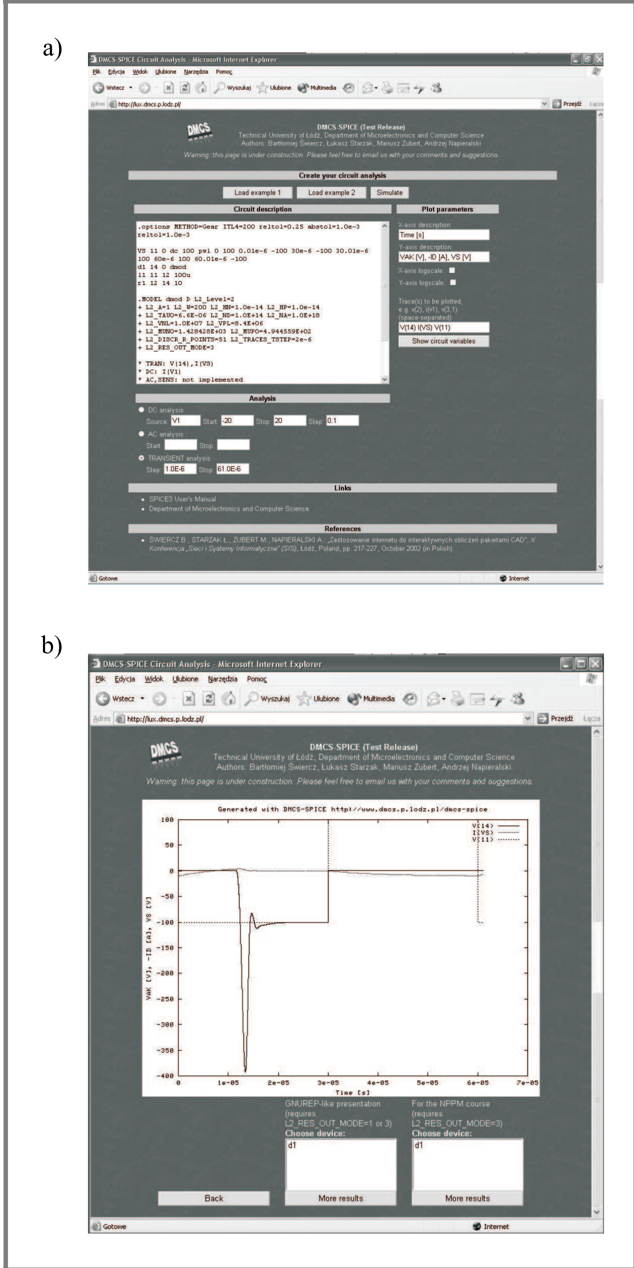


Fig. 14. Designed website – circuit data input (a) and plotted simulation results (b).

The circuit simulation core has been based on Simulation Program with Integrated Circuit Emphasis (SPICE) because of high popularity and strong position of SPICE-like simulators. This choice ensures wide accessibility of the environment.

Because of its open source character, Berkeley SPICE3F5 [26] has been used. This enables constant improvement of the designed environment by implementation of new device models and more efficient numerical algorithms. Also, the simulation core could be customized to meet the requirements of the project.

During the development of the presented environment, it turned out that linearization algorithms implemented in

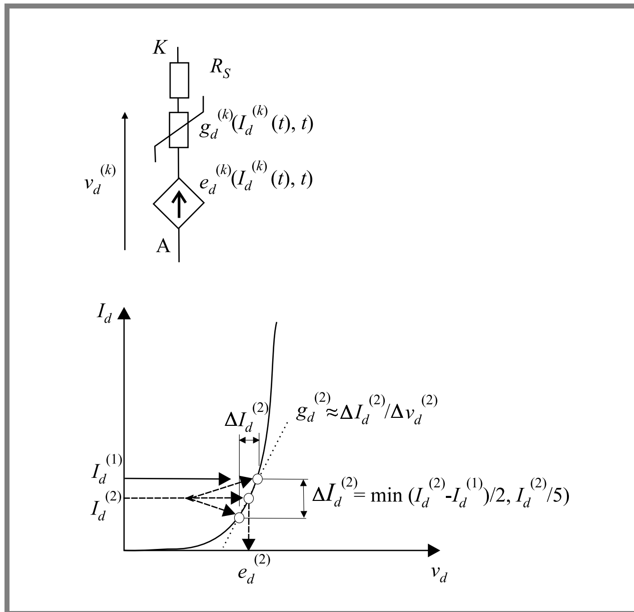


Fig. 15. The implemented linearization algorithm – power diode example.

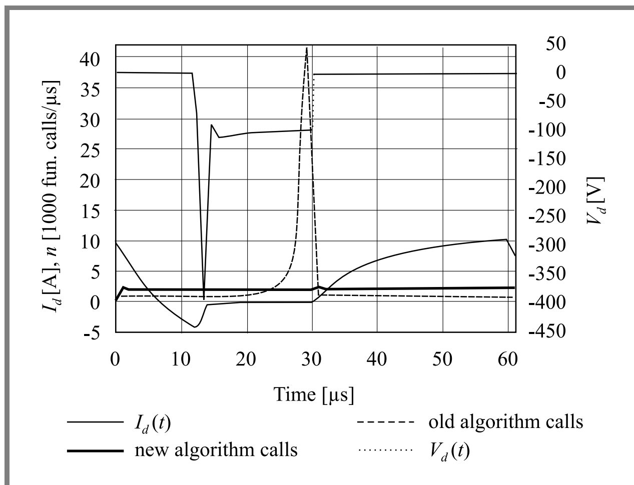


Fig. 16. Number of diode voltage evaluations using the original SPICE3F5 and the new linearization algorithm (calls per microsecond of simulation time). The calculations were made for diode reverse and forward recovery (see device current I_d and voltage V_d waveforms).

the original SPICE3F5 were unable to assure numerical convergence during the simulation of circuits containing highly nonlinear elements, such as the power diode model presented. Thus, a better-suited algorithm has been proposed and included in the developed software. During each iteration, the conductance value is calculated based on two points from its closest neighbourhood, as presented in Fig. 15. If convergence problems occur, a simplified algorithm is used. The new algorithm has permitted the number of calls of the function implementing the diode characteristics to be decreased, especially when fast dynamic processes take place in the device structure (Fig. 16).

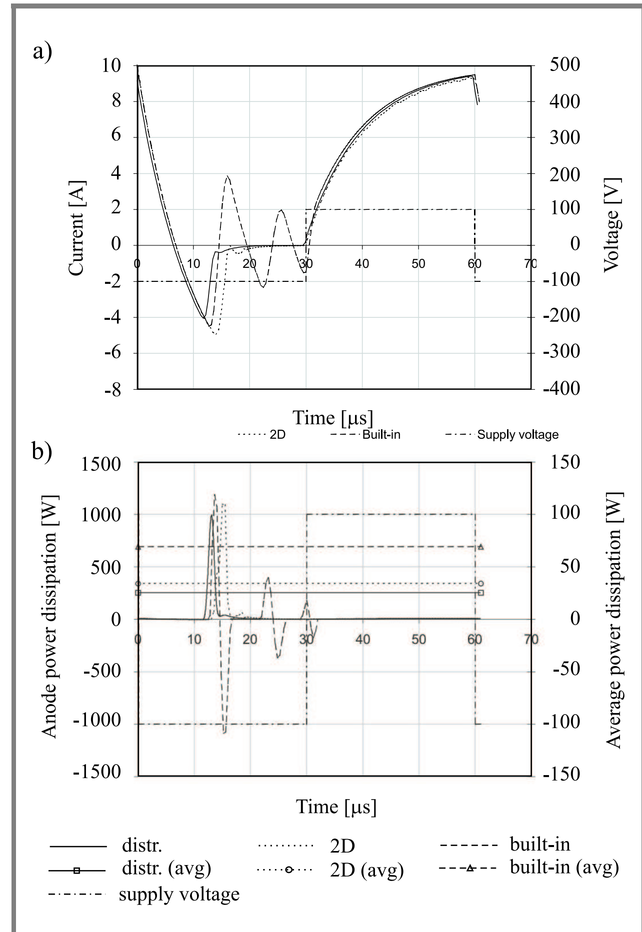


Fig. 17. Simulation results – PIN diode reverse and forward recovery with inductive load: (a) current waveforms; (b) power dissipation waveforms and average values: 2D – bidimensional model; built-in – built-in PSPICE power diode model with fitted parameters; distr. – distributed model included in the presented package.

Simulation results obtained with the model and the developed simulation environment described above are presented in Fig. 17. They show a good agreement with 2D simulations performed with the simulator MOPS [5, 8, 9, 32] based on the finite-box method, and demonstrate the erroneous results obtained with the built-in SPICE diode model with fitted parameters.

6. Conclusion

In this paper, the most important phenomena in power semiconductor devices have been discussed. As a conclusion concerning power bipolar devices we found that:

- there is a problem with the diffusion (minority carriers),
- the accuracy of computation is very poor when lumped models are used, therefore new compact models are very welcome [35],

– for correct simulation, distributed models are still necessary.

In order to make the simulation of power devices possible, a new type of a PIN diode model has been proposed in this paper. The results obtained seem to be very promising as far as the accuracy and simulation time are concerned.

In the last part of this paper, a free environment allowing the electronic circuit design by means of Internet has been presented. Thanks to the implemented PIN diode model, users have been given the possibility to simulate a realistic behaviour of circuits containing power diodes and VDMOS transistors. One should note that a professional design process might require the designer to consider electro-thermal couplings in power devices [3, 35]. However, it seems that the developed tool offers new possibilities as compared to the commercial CAD packages currently available.

The client-server system architecture and the Berkeley SPICE basis enable the application of the environment in the education of electronics and facilitate the future development of such tools in cooperation with other scientific, educational or industrial centres, as well as with individual users worldwide. Future extensions of the environment are being considered by including models of other power semiconductor devices.

Acknowledgements

I would like to express my gratitude to Dr. Marek Turowski, Dr. Mariusz Grecki, Dr. Mariusz Zubert, Dr. Grzegorz Jabłoński, Dr. Michał Furmańczyk, Dr. Marcin Janicki, Mr. Łukasz Starzak and Mr. Bartłomiej Świercz, for all scientific discussions during our common work on the fascinating topic concerning power semiconductor device modeling.

References

[1] H. Benda, A. Hoffmann, and E. Spenke, "Switching processes in alloyed PIN rectifiers", *Solid State Electron.*, vol. 8, pp. 887–906, 1965.

[2] K. Djellabi, M. Napieralska, H. Tranduc, P. Rossel, and K. Kassmi, "Modéles du transistor MOS de puissance", *Rev. Gén. Eléectr.*, no. 6, pp. 8–16, 1992.

[3] M. Furmańczyk, "Elektrotermiczna symulacja układów VLSI ze szczególnym uwzględnieniem integracji w środowisku projektowania". Politechnika Łódzka, Łódź, 1999 (Ph.D. thesis in Polish).

[4] H. Goebel, "A unified method for modeling semiconductor power devices", *IEEE Trans. Pow. Electron.*, vol. 9, no. 5, pp. 497–505, 1994.

[5] M. Grecki, "Dwuwymiarowe modelowanie bipolarnych i unipolarnych struktur półprzewodnikowych". Politechnika Łódzka, Łódź, 1995 (Ph.D. thesis in Polish).

[6] M. Grecki, G. Jabłoński, and A. Napieralski, "Transient temperature evaluation during switching process in IGBT transistor", in *Proc. 4th Int. Sem. Pow. Semicond. ISPS*, Prague, Czech Republic, 1998, pp. 119–124.

[7] M. Grecki, M. Turowski, and A. Napieralski, "2D analysis of more then one semiconductor devices together with external circuit", in *Proc. Fifth IEE Int. Conf. Pow. Electron. Var.-Speed Driv. PEVD*, London, England, 1994, pp. 441–446.

[8] M. Grecki, G. Jabłoński, and A. Napieralski, "MOPS – parallel environment for simulation of electronic circuits using physical models of semiconductor devices", in *Proc. 4th Eur. PVM/MPI Users' Group Meet.*, Kraków, Poland, 1997, pp. 478–485.

[9] G. Jabłoński: "Komputerowa analiza układów elektronicznych z zastosowaniem wielowymiarowych modeli fizycznych przyrządów półprzewodnikowych mocy". Politechnika Łódzka, Łódź, 1999 (Ph.D. thesis in Polish).

[10] M. Janicki, "Thermal modelling of semiconductor structures with special consideration of inverse problem methods for parameter estimation". Ph.D. thesis, Technical University of Łódź, Łódź, Poland, 1999.

[11] M. Janicki, M. Zubert, and A. Napieralski, "Application of inverse problem algorithms for integrated circuit temperature estimation", *Microelectron. J.*, vol. 30, pp. 1099–1107, 1999.

[12] M. Janicki, G. De Mey, and A. Napieralski, "Transient thermal analysis of multilayered structures using Green's functions", *Microelectron. Reliab.*, vol. 42, no. 7, pp. 1059–1064, 2002.

[13] M. Janicki, G. De Mey, and A. Napieralski, "Application of Green's functions for analysis of transient thermal states in electronic circuits", *Microelectron. J.*, vol. 33, no. 9, pp. 733–738, 2002.

[14] Ph. Leturcq, J. Gaubert, A. Napieralski, and Z. Khatir, "Simulation sur microcalculateur de l'ouverture des thyristors GTO", *Rev. Gén. Eléctr.*, no. 1, pp. 40–48, 1990.

[15] Ph. Leturcq, M. O. Berraies, J.-P. Laur, and P. Austin, "Full dynamic power bipolar device models for circuit simulation", in *Proc. Record. Pow. Electron. Special. Conf. PESC*, Fukuoka, Japan, 1998, pp. 1695–1703.

[16] D. Levy, G. Puppo, and G. Russo, "Central WENO schemes for hyperbolic systems of conservation laws", *Math. Model. Numer. Anal.*, vol. 33, no. 3, pp. 547–571, 1999.

[17] R. Maimouni, M. Belabadia, P. Rossel, H. Tranduc, D. Allain, C. E. Cordonnier, M. Napieralska, and J. Costa Freire, "Modèle SPICE du transistor V.DMOS pour circuits de commutations", in *Proc. L'Electronique de Puissance du Futur*, Bordeaux, France, 1988.

[18] R. Maimouni, D. Allain, M. Napieralska, H. Tranduc, P. Rossel, and C. E. Cordonnier, "Model library for power MOSFETs in switching circuits and converters", in *Proc. Mediterr. Electrotech. Conf. ELECON*, Lisbon, Portugal, 1989.

[19] R. Maimouni, P. Rossel, D. Allain, M. Napieralska, H. Tranduc, and C. E. Cordonnier, "Modele "universel" du transistor MOS de puissance pour le logiciel SPICE", in *J. d'Etud. SEE "Simulation en Electronique de Puissance du Composant au Systeme"*, Marseille, France, 1989.

[20] R. Maimouni, P. Rossel, D. Allain, M. Napieralska, H. Tranduc, and C. E. Cordonnier, "Modèle universel du transistor MOS de puissance pour le logiciel SPICE", *Rev. Gén. Eléctr.*, no. 1, pp. 49–59, 1990.

[21] M. Napieralska, H. Tranduc, P. Rossel, and C. E. Cordonnier, "Bibliotheque de transistors MOS de puissance dans le couplage des logiciels "HyperCard" et "Spice", in *Proc. 3ème Coll. Natl. SEE "L'Electronique de Puissance du Futur" EPF*, Toulouse, France, 1990.

[22] M. Napieralska, H. Tranduc, C. E. Cordonnier, J. P. Berry, and P. Rossel, "Power MOSFET's library builder for switching circuits simulation and design", in *Proc. Symp. Mater. Dev. Pow. Electron. MADEP*, Florence, Italy, 1991.

[23] M. Napieralska, "Modélisation du transistor V.DMOS pour simulation de circuits en électronique de puissance". Ph.D. thesis, INSA, Toulouse, France, 1991.

[24] A. Napieralski, M. Napieralska, and M. Grecki, "Computer laboratory for CAD of power integrated circuits", in *Computer-Aided Design and Computer-Aided Engineering in Electronic Engineering Education*, A. Filipkowski, Ed. Warszawa: WUT, 1996, pp. 135–184.

[25] A. Napieralski and M. Napieralska, *Polowe półprzewodnikowe przyrządy dużej mocy*. Warszawa: WNT, 1995 (in Polish).

[26] T. L. Quarles, A. R. Newton, D. O. Pederson, and A. Sangiovanni-Vincentelli, "SPICE3 version 3F5 user's manual", Department of EECS, University of California, Berkeley, USA, 1994.

[27] L. Starzak, M. Zubert, A. Napieralski, P. Austin, G. Bonnet, M. Marmouget, J.-L. Sanchez, and T. Bordignon, "Physical power diode model and its implementation in SABER environment", in *Proc. 8th Int. Conf. Mix. Des. Integr. Circ. Syst. MIXDES*, Zakopane, Poland, 2001, pp. 213–220.

[28] L. Starzak, A. Napieralski, and J. J. Charlot, "VHDL-AMS: a competitor for SPICE in modeling of semiconductor devices", in *Proc. Int. Conf. "Modern Probl. Radio Eng., Telecommun. Comput. Sci." TCSET*, Lviv-Slavske, Ukraine, 2002, pp. 353–356.

[29] L. Starzak, M. Zubert, and A. Napieralski, "The new approach to the power semiconductor devices modelling", in *Proc. Int. Conf. Model. Simul. Microsyst. MSM*, San Juan, Puerto Rico, 2002, pp. 640–644.

[30] B. Swiercz, L. Starzak, M. Zubert, and A. Napieralski, "An interactive website as a tool for CAD of power circuits", in *Tech. Proc. 2003 Nanotech. Conf. Trade Show (NanoTech)*, San Francisco, USA, 2003, pp. 346–349.

[31] M. Turowski and A. Napieralski, "Two-dimensional analysis of GTO switching under the influence of external circuit", *IEE Proc.-G, Circ., Dev. Syst.*, vol. 141, no. 6, pp. 483–488, 1994.

[32] M. Turowski, "Dwuwymiarowe modelowanie półprzewodnikowych przyrządów dużej mocy". Politechnika Łódzka, Łódź, 1992 (Ph.D. thesis in Polish).

[33] B. Wilamowski, A. Malinowski, and J. Regnier, "Internet as a new graphical user interface for the SPICE circuit simulator", *IEEE Trans. Ind. Electron.*, vol. 48, no. 6, pp. 1266–1268, 2001.

[34] B. Wilamowski, A. Malinowski, and J. Regnier, "SPICE based circuit analysis using web pages", in *Proc. 2000 ASEE Ann. Conf.*, Saint Louis, USA, 2000.

[35] M. Zubert, "Wielowymiarowe modelowanie sprzężonych zjawisk fizycznych w nowoczesnych strukturach półprzewodnikowych, z zastosowaniem języka VHDL-AMS i modeli o stałych rozłożonych, ze szczególnym uwzględnieniem modelowania elektrotermicznego układów VLSI oraz Smart Power". Politechnika Łódzka, Łódź, 1999 (Ph.D. thesis in Polish).

Małgorzata Napieralska received the M.Sc. degree in electronics in 1982 from Technical University of Łódź, from 1987 to 1991 she was a Ph.D. student in Laboratoire d'Automatique et d'Analyse des Systèmes du CNRS, Toulouse, France, researching in the field of modeling of power devices. In 1991 she received the Ph.D. in microelectronics from Institut Nationale des Sciences Appliquées (France). Currently she is with the Department of Microelectronics and Computer Science, Technical University of Łódź working on modeling and macromodeling of semiconductor devices, integrated circuits and micro-electro-mechanical structures and the CAD of electronic circuits. She is author or co-author of over 60 publications in scientific journals and conference proceedings, 2 books and over 20 80 scientific reports.

e-mail: MNapier@dmcs.p.lodz.pl
Department of Microelectronics and Computer Science
Technical University of Łódź
av. Politechniki 11
93-590 Łódź, Poland

Andrzej Napieralski received the M.Sc. and Ph.D. degrees from TUL in 1973 and 1977 respectively, and D.Sc. (habilitation) degree in electronics from Warsaw University of Technology (Poland) in January 1989, and in microelectronics from Université de Paul Sabatié (France) in May 1989. He received the title of Professor 1995, and become the Tenured Professor in 1999. In 2002 he has been elected as the Vice-Rector for Promotion and International Co-operation. He is author or co-author of 376 publications in scientific journals and conference proceedings, 24 books and textbooks, 15 chapters in books, 2 patents and over 90 scientific reports and expertises. Since July 1985 until September 1991, he has been a member of the research team on power devices at the CNRS LAAS, Toulouse, and Visiting Professor at the INSA of Toulouse. Since October 1991 he has been Professor at TUL. From 1992 until 1995 he was the Vice-Director of Institute of Electronics, and since 1996 he has been Director of DMCS. In 2000 he was appointment as the Chair of IEEE Poland ED Chapter and in 2001 the Editor of IEEE EDS Newsletter for Scandinavia and Central Europe.

e-mail: Napier@dmcs.p.lodz.pl
Department of Microelectronics and Computer Science
Technical University of Łódź
av. Politechniki 11
93-590 Łódź, Poland

Invited paper

Process and device requirements for mixed-signal integrated circuits in broadband networking

Krzysztof Iniewski, Marek Syrzycki, and Sorin P. Voinigescu

Abstract — The paper describes the present status of the broadband wireline infrastructure consisting of the backbone core, metro rings, access network, local and storage area networks. Examples of various mixed-signal integrated circuits are described. Based on these considerations required process and device performance is extrapolated.

Keywords — CMOS, SiGe, InP, networking, WAN, MAN, SAN, LAN, OEO conversion, cut-off frequency, manufacturability.

1. Introduction – broadband network today

The convergence of voice, data and video onto a single network combined with large global consumer and corporate appetite for internetworking is leading to installation and upgrading of communication infrastructure worldwide. The paper reviews recent trends in broadband infrastructure deployment and shows future possible directions. Networking integrated circuits (ICs), being the nuts and bolts of this infrastructure build-up, are continuously evolving leading to very complex electronic devices. The main theme of the paper is to link dramatic changes in Internet network with demands placed on IC design, IC processes and device performance that are needed in order to continue fuelling this infrastructure growth.

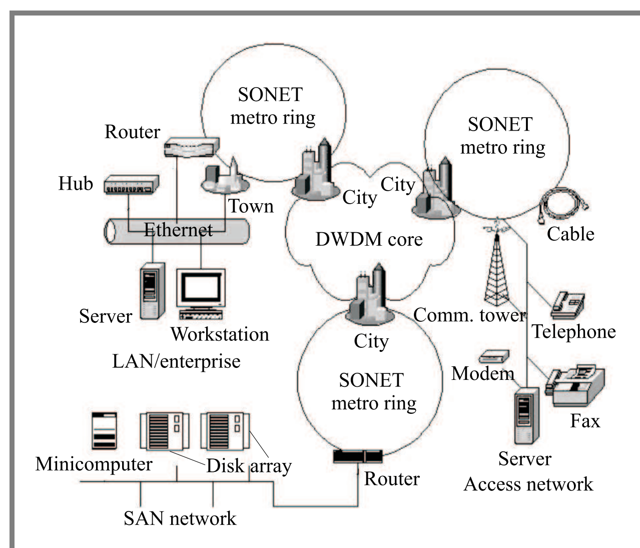


Fig. 1. Conceptual drawing of the broadband network.

The complexity of today's global network infrastructure has been primarily caused by the use of multiple networks, as shown in Fig. 1, and multiple protocols:

- Access networks – networks for both consumers and corporate customers to provide data, video and voice to all required locations using time-domain multiplexing (TDM), synchronous optical network (SONET), and asynchronous transfer mode (ATM) [1].
- LANs – local area networks that connect PCs, workstations, printers and other devices inside a building or campus, traditionally using Ethernet (10 Mbit/s, 100 Mbit/s and 1 Gbit/s) connections.
- SANs – storage area networks that connect backend storage disks via high-speed interfaces using primarily fiber channel (1 Gbit/s and 2 Gbit/s) protocols.
- MANs – metropolitan area networks that connect data and voice traffic at the city level typically using SONET rings (OC-48/OC-192[1]).
- WANs (core, backbone) – wide area networks which connect multiple corporate locations or cities across long distances. ATM, SONET and IP are used in the core of the network [2].

Access. Access networks need to carry both voice and data. Voice is carried using well-known circuit-switching techniques. Data is carried using multiple different technologies like TDM, integrated services digital network (ISDN), ATM, plesio-synchronous digital hierarchy (PDH), frame relay or digital subscriber loop (DSL). As data needs to be merged with voice at some point in the network, ATM is well suited for that purpose, being able to transport reliably voice over SONET (fiber) or over DSL (copper wire).

The access networks use different media for data transmission: copper twisted pair wires, coaxial cables, optical fibers or simply air in the case of wireless access. Digital subscriber loop technology is becoming the dominant technology for the transmission over the copper wire, while cable modems are used for transmission over

[1] OC in SONET indicates optical carrier rate. The bandwidth of OC-48 is 2.488 Gbit/s while that of OC-192 is 9.953 Gbit/s.

the coaxial cable. As very small percentage of businesses and households has a direct fiber connection, fiber to the home (FTTH) technology cannot be effectively deployed. Instead, fiber to the curb (FTTC) is being considered due to the progress in passive optical networks (PON) technology which uses passive optical splitters to split the optical signal from fiber to copper wires.

Enterprise. Ethernet rules local area networks. This well known technology has managed to evolve from 10 Mbit/s to 100 Mbit/s (fast Ethernet) to 1 Gbit/s (gigabit Ethernet), and now to 10 Gbit/s, in a backwards compatible fashion. Ethernet architects are already discussing 40 and 100 Gbit/s versions of this technology although practical implementations are probably years away. Due its popularity and massive deployment in LANs, Ethernet is cheap. As a result it threatens other protocols in areas outside LAN. In particular, optical Ethernet (running Ethernet over optical fiber) might become popular in the future in the access and metro networks shunning away ATM and SONET.

Storage. Storage area networking is currently the fastest growing segment of broadband deployment. While discussions continue on merits of network attached storage (NAS) vs. storage area networks, both remain universally deployed in large and very large corporations to store massive amounts of corporate data. Fiber channel is the dominant protocol in SANs although iSCSI[2] implementations might threaten that dominant position in the future.

Metro. Current metropolitan infrastructure is primarily based on SONET rings that carry ATM and IP traffic [3, 4]. Future services might use various other technologies, such as next-generation SONET, optical Ethernet, multi-service dense wavelength division multiplexing (DWDM), or multi protocol Lambda switching (MPLS) [5, 6]. It needs to be pointed out that the important function of metropolitan area networks is also to collect wireless data traffic from third generation (3G) wireless networks.

Core. Core networks require transmission over long distances, typically beyond 500 km, preferably beyond 5000 km. In traditional networks repeaters are used to regenerate signals every 60–80 km [7, 8]. Long span without repeaters can be also accomplished using optical amplification enhanced by pulse shaping and error coding techniques. Optical signals can be restored using Raman or erbium doped fiber amplifiers (EDFAs). Future optical techniques of ultra-long haul systems might use solitons, which, due to their unique properties, can travel in optical fiber over very long distances. It is also anticipated that optical switching technology, probably based on MEMs structures, will replace electronic switching in the core of the network [9].

[2]iSCSI is a new technology of running IP Internet Protocol over small computer system interface (SCSI) protocol.

2. Mixed-signal integrated circuits for broadband communication

Broadband communication ICs are semiconductor chips that enable processing and transmission of voice, video and data, in either electrical or optical form, from one location to another across broadband Internet network.

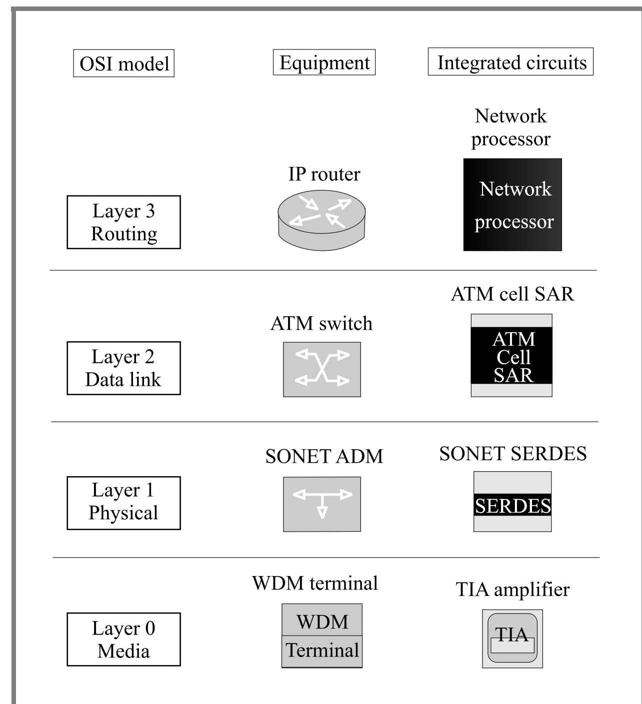


Fig. 2. Classification of integrated circuits (right), networking equipment (middle) using open interconnect system (OSI) model (left).

Figure 2 summarizes the classification of integrated circuits in broadband networking. The left column shows open system interconnect (OSI) model hierarchy consisting of the following layers:

- Layer 0 or media layer. An example of networking technology is wavelength division multiplexing (WDM), an example of networking equipment is a WDM terminal, and an example of an integrated circuit is trans-impedance amplifier (TIA).
- Layer 1 or physical layer. An example of networking technology is SONET transport, an example of networking equipment is SONET add/drop multiplexer (ADM), and an example of an integrated circuit is SONET SERDES[3].
- Layer 2 or data link layer. An example of networking technology is ATM, an example of networking equipment is ATM switch, and an example of an

[3]SERDES – to SERialize and DE-Serialize. MUX/DE-MUX type IC device that converts parralel data stream into a serial one.

integrated circuit is ATM cell segmentation and reassembly (SAR) device.

- Layer 3 or networking layer. An example of networking technology is IP routing, an example of networking equipment is a core router, while an example of an integrated circuit is network processor.

After OSI model has been introduced and networking equipment has been broadly classified, we can discuss the role of integrated circuits in the networking gear. Integrated circuits typically perform the following functions:

- Coding, modulation, and amplification of electrical signals for transmission through physical medium (optical fiber, twisted pair copper wire, coaxial cable). Also in reverse direction: de-coding, demodulation and equalization for reception of electrical signals. These devices are typically referred to as physical medium devices (PMDs) and physical layer devices (PHYs), and are considered to be Layer 1 devices in the OSI protocol stack. These devices are either completely analog or mixed-signal with significant analog content.

- Data formatting into frames or cells using predefined protocols (ATM, Ethernet, fiber channel, etc.). These devices are typically referred to as framers or mappers, and are considered to be Layer 2 devices. These devices are providing digital processing and the only analog/mixed-signal circuitry they require is high-speed serial links used for chip-to-chip communication.

- Data-packet processing. Processing functions include protocol conversion, packet forwarding, policing, look-up, classification, encryption, and traffic management [10–12]. These devices are typically referred to as network processors, classification engines or traffic managers, and are considered to be Layer 3 devices. Typically they are purely digital devices that use highly parallel I/O interfaces.

2.1. PMD and PHY mixed-signal devices

In general, physical medium dependent devices are being used for I/O interfacing to physical media, like fiber, copper wire or coaxial cable. Typically they are data protocol independent as they deal only with physical effects and electrical signals. PMDs include devices such as amplifiers for photo-detectors or drivers for lasers. PMD devices require analog design expertise.

Physical layer devices are responsible for defining how the traffic will be transported from one location to another. PHY devices include SONET SERDES blocks, clock and data recovery (CDR) circuits [14, 15], Ethernet transceivers [16], cable and xDSL modems chips [17, 18]. The most critical parameters for these devices are related to timing jitter. PHY devices have to comply with elaborate specification for the amount of jitter being generated (intrinsic jitter), the amount of jitter the device can tolerate at its input (jitter tolerance), and the transfer characteristics of the output jitter vs. applied input jitter (jitter transfer). Different jitter specifications exist for different transport technologies (SONET, Ethernet or fiber channel) making it difficult to implement the same phase-lock loop design in different devices.

2.2. Optical to electrical to optical conversion

Most of network communication in wide area network is carried on optical fibers while most networking in local area networks relies on electronic devices and transport along electrical wires [13]. As a result, the optical to electrical to optical (O-E-O) conversion process is required in numerous places in the modern broadband network. We will use the O-E-O link example to illustrate further challenges in mixed-signal IC design. Figure 3 shows an optical module containing a number of PMD and PHY devices.

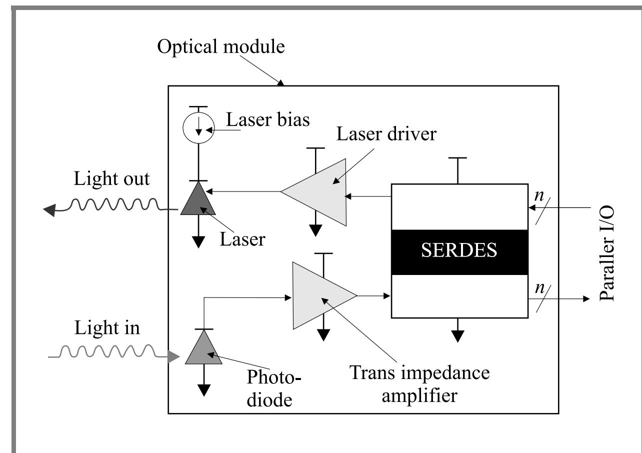


Fig. 3. Optical module for O-E-O conversion containing optical devices (laser and photodiode), analog circuits (laser driver and trans-impedance amplifier), and the mixed-signal IC.

The optical signal is received by a photo-diode. Both PIN and avalanche photo diodes (APD) are used for that purpose. Since the photodiode produces photo-current I while most electronic devices require voltage signal V, the conversion from I to V is done by a PMD device called trans-impedance amplifier. While converting from current to voltage the amplifier also amplifies the signal. It is important to note that TIA is extremely noise sensitive, as it deals with very small currents in the range of a few pico amperes (pA). Due to this noise sensitivity TIA is the hardest element to integrate with other ICs in the optical module.

After an I-to-V conversion a voltage signal is ready for further processing. In some cases the amplitude of the signal coming out of the TIA is too small and requires additional amplification by the limiting amplifier (not shown

in Fig. 3), the output of which will provide constant signal amplitude regardless of the strength of the optical signal being converted. The limiting amplifier is another example of a PMD device.
The signal from the limiting amplifier is received by the clock and data recovery block inside SERDES IC. The CDR block recovers the clock from the NRZ data stream using analog phase-lock loop techniques. The jitter tolerance of a CDR is typically the most challenging parameter to be met in commercial products, because it has to comply with stringent networking standards[4]. After clock recovery, the signal needs to be de-serialized to lower rates. The circuitry that performs this function is called a de-multiplexer (de-mux) or de-serializer, and is also typically a part of the SERDES IC. Using the recovered clock and lower speed, further digital processing of the parallel data can be performed in a follow-up Layer 2 digital device.
Similar processing occurs in the reverse, egress direction. The data generated by a Layer 2 device is being serialized using a multiplexer (mux) or serializer device. Clock multiplication unit (CMU) is used to generate high purity reference clock for data transmission. The entire SERDES operation is summarized in Fig. 4.

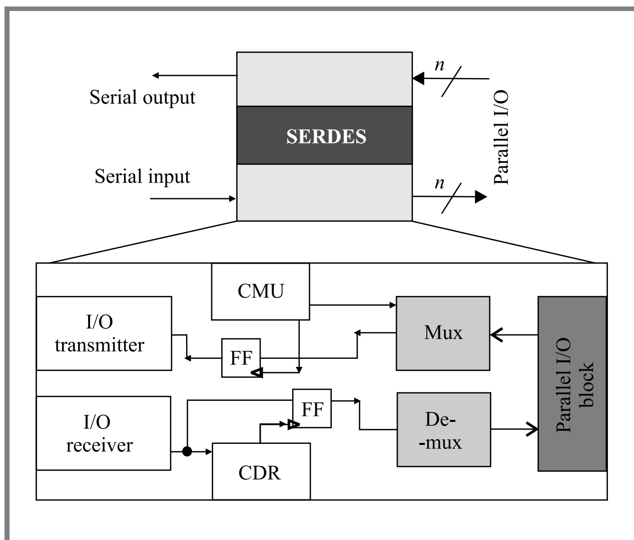


Fig. 4. Schematic block diagram of SERDES IC device that incorporates analog I/O receive and transmit blocks, clock and data recovery (CDR), clock multiplication phase-locked loop (CMU), mux and de-mux, and parallel digital I/O block. FF indicates a flip flop, as two critical flip-flops are shown that are used to re-time the data in the transmit direction and clean up the receive data using the CDR supplied recovered-clock.

The data transmitted from SERDES reaches a PMD device called a laser driver. The laser driver in turn modulates the laser itself, effectively converting the electrical signal into an optical one.
A laser driver has to meet very specific requirements that are somewhat different from those imposed on other PMD components. It has to operate at the same high-date rate as other devices but it must deliver higher voltage swings as well. The requirement for higher voltage swing is due to currently available laser structures. For example, conventional Mach-Zehnder (MZ) or eletro-absorption (EA) structures require voltage swings of 5 V or more to achieve very good extinction ratio.

[4]Each networking protocol, like SONET or Ethernet, defines different criteria for jitter requirements.

3. Device requirements for broadband circuits

While the current broadband infrastructure runs at the rates of up to 10 Gbit/s, the future network will require processing at 40 Gbit/s and beyond [19, 20]. It is generally accepted that 40 Gbit/s data rates are currently in the realm of SiGe, GaAs or InP technologies due to their fundamental device properties (high mobility, transconductance and breakdown) [21]. However, for a given technology node, the mainstream CMOS technology shows potential competitive advantages in terms of cost, level of integration and power dissipation as technology scaling enables both increased functionality and speed (frequency) in integrated circuits. It is vital to predict at which point CMOS performance will be sufficient to produce 40 Gbit/s circuits.
Cut-off frequency f_T is a critical parameter for transistor performance. It is defined as the frequency at which the transistor voltage gain becomes unity; f_T is frequently used for speed comparison between various semiconductor processes. Figure 5 presents f_T data for NMOS, PMOS and SiGe HBT transistors as a function of lithographic feature size [20].

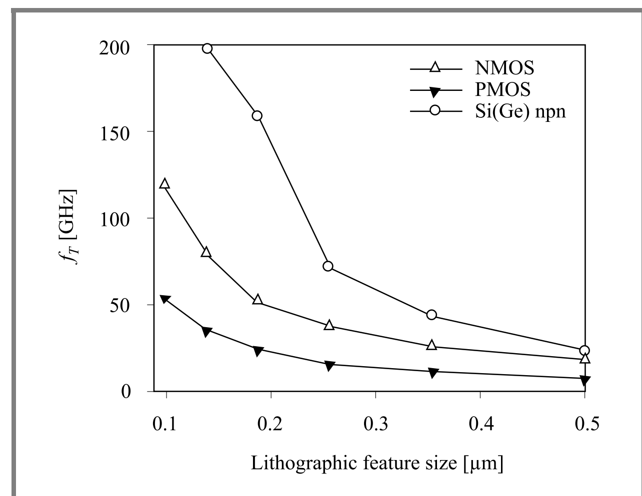


Fig. 5. Cut-off frequency data as a function of lithographic feature size for NMOS, PMOS, and SiGe npn transistors.

The plots in Fig. 5 represent measured data up to 0.13 μm and simulation-based extrapolated data up to 90 nm. All MOSFET and SiGe HBT measured data represent maximum f_T at the best possible bias point. For MOSFETs

the optimum bias point is: $V_{GS} = V_{DS} = V_{DD}$. For example, in a 0.13 μm NMOSFET at $V_{DS} = 0.4$ V and $V_{GS} = 0.65$ V (low bias) the measured f_T is 70 GHz, while the maximum cut-off frequency is around 80 GHz. It is interesting to note that even at low bias MOSFET cut-off frequency is still higher than that of SiGe HBT biased at $V_{CE} = 0.4$ V.

Some interesting observations can be drawn from Fig. 5. First, NMOS device has a 2 $\times$ speed advantage over PMOS. This well-known result reminds us that all fast CMOS circuits will continue to rely on NMOS design in high-speed paths of operation using PMOS devices only for biasing and loading. Second, SiGe devices retain a 2 $\times$ speed advantage over NMOS transistors. Or, to put it in a different perspective, CMOS process is behind SiGe, typically by two generations.

It is instructive to compare the predictions of the International Technology Roadmap for Semiconductors (ITRS) with the measured data [20]. This comparison is presented in Table 1.

Table 1
Cut-off frequency versus process feature size*

Process feature size [nm]	Cut-off frequency [GHz] ITRS 1999	Cut-off frequency [GHz] ITRS 2001	Cut-off frequency [GHz] data from Fig. 2 [20]
180	20	–	50
150	25	132	–
130	30	149	80
90	40	225	–
65	–	371	120

* Data compiled from International Technology Roadmap for Semiconductors presented in 1999 and 2001. The data from reference [20] is included for comparison.

Clearly, the predictions made in 1999 were very pessimistic, while the subsequent correction in 2001 overly optimistic. In fact, a comparison between ITRS 1999 and ITRS 2001 indicates astonishing differences – the values predicted in 2001 are approximately 5 times higher than those predicted in 1999 for the same feature size.

4. Process requirements

In the recent years advanced CMOS processes could compete with early SiGe implementations as the cost of mask making was not a major economic factor. For example, 0.18 μm CMOS was competitive to 0.35 μm SiGe process by virtue of offering similar performance in a well-known CMOS design and process environment. This trend might change in the future as the cost of mask making rapidly increases with feature size reductions[5]. With non-recurring engineering charges (NRE) for mask making reaching over one million dollars for 90 nm CMOS process it will likely be more cost effective to manufacture ICs in 0.18 μm SiGe process rather than in 90 nm CMOS. Based on the cut-off frequency data presented in both processes one can expect similar high-frequency performance. The only application driving the 90 nm CMOS process option would be digital devices with extremely high (over 10 million) gate count.

The operation frequency of real circuits is lower than the pure device speed metrics because real circuits have to drive a heavy fan-out load in addition to device and interconnect parasitics. One popular circuit speed metric for high performance microprocessor design is the FO4 delay, where a CMOS inverter drives a load of fan-out of four. Although FO4 delay is a different parameter than f_T, both are very strongly correlated. For clarity we will continue to use the cut-off frequency f_T as a measure of the speed of the process.

Without design innovations, the achievable operation frequency of circuits is around $f_T/10$. With some design innovations it is possible to design circuits up to $f_T/4$. In fact a review of the papers published in the *Journal of Solid-State Circuits* during the last 5 years indicates that state-of-the-art circuits have their frequency of operation somewhere between $f_T/10$ and $f_T/2$. However, it has to be pointed out that some of the circuits published in the literature are of academic nature, and are not proven to work across PVT (process, voltage, temperature) or lack experimental evidence as they are simulation based only. It is therefore reasonable to assume that for commercial products with reasonable yield the maximum frequency of operation is limited by $f_T/4$.

Using this $f_T/4$ metric and the results of Fig. 5 we can draw the following conclusions:

- 10 Gbit/s circuits can be fabricated in the 0.18 μm CMOS process with f_T of 50 GHz (there are in fact commercial products which accomplish this performance).
- 20 Gbit/s circuits can be fabricated in the 0.13 μm CMOS process with f_T of 80 GHz (although commercial demand for 20 Gbit/s circuits is not clear).
- Even the 65 nm CMOS process with f_T of 120 GHz might not be fast enough to manufacture 40 Gbit/s circuits. On the other hand advanced SiGe process at 0.18 μm has f_T of 160 GHz, which should be sufficient for 40 Gbit/s applications.

The simple $f_T/4$ metric can obviously be used only as a basic figure of merit. More detailed considerations are required, in particular for modern CMOS processes which use strained silicon and silicon on insulator (SOI) options.

[5]Typically NRE charges for mask making double from generation to generation.

In fact, one can argue that for analog circuits the maximum frequency of oscillations $f_{\max}$[6] is equally important in the determination of the maximum frequency of circuit operation.

Further improvements in MOSFET cutoff and, especially, oscillation frequencies can be achieved using SOI substrates, which increase channel mobility, gate control of the channel and significantly reduce source-bulk and drain-bulk capacitance [22]. The latter two are the dominant capacitive elements in Si MOSFETs at 130 nm and below, and cannot be properly scaled in bulk CMOS processes. Further 30% to 50% improvement in both electron and hole mobilities can be accomplished by using strained Si and SiGe channels on virtual substrate wafers. It is likely that a 90 nm SOI CMOS process incorporating strained silicon channels and reduced parasitic back-end will be adequate for 40 Gbit/s CDR blocks and SERDES ICs operating with full-rate 40 GHz clocks and sub 1.2 V supply.

CMOS scaling down to 90 nm exacerbates several well-known challenges. For digital applications, these include short channel effects, controlling threshold voltage and exponentially increasing leakage currents due to both source/drain and gate currents. For analog circuits additional challenges include IV linearity, low noise characteristics, and transistor matching. To solve these challenges new transistor structures will likely be needed for the 65 nm process generation and below. These include, but are not limited to, ultra-thin body SOI, band-gap engineered transistors (strained Si and SiGe), FinFETs or vertical structures. Only these new structures offer hope of reaching 100 Gbit/s rates in silicon, otherwise InP based devices will have to be used to overcome that speed barrier.

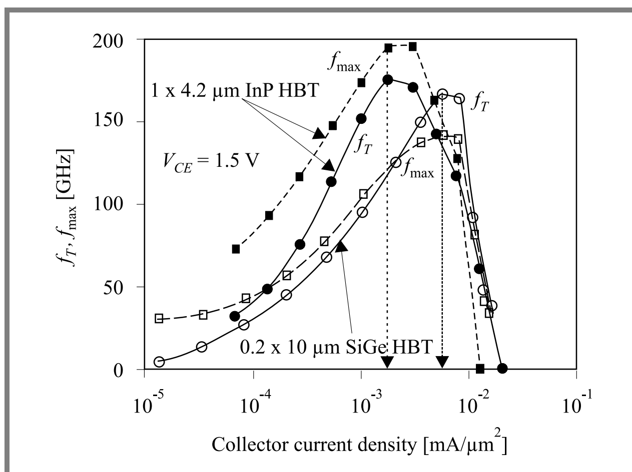


Fig. 6. The cut-off frequency f_T and the maximum oscillation frequency $f_{\max}$ for 1 μm $\times$ 4.2 μm InP and 0.2 μm $\times$ 10 μm SiGe HBTs.

To show the potential behind InP technology Figs. 6 and 7 show the measured f_T and $f_{\max}$ data for various high-speed devices. Figure 6 contains the data for 1 μm $\times$ 4.2 μm InP and 0.2 μm $\times$ 10 μm SiGe HBTs. The InP device achieves higher maximum cut-off frequency of 170 GHz compared to 160 GHz for the SiGe device, and significantly higher maximum oscillation frequency of 200 GHz compared to 130 GHz for the SiGe device.

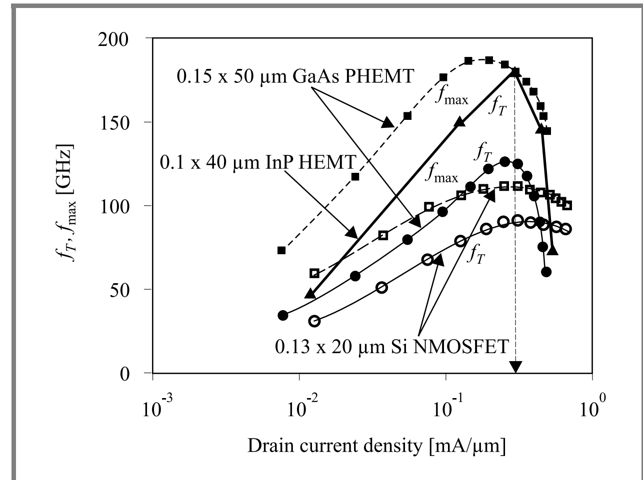


Fig. 7. The cut-off frequency f_T and the maximum oscillation frequency $f_{\max}$ 0.13 μm $\times$ 20 μm silicon NMOSFET, 0.15μm $\times$ 50 μm GaAs PHEMT and 1 μm $\times$ 4.2 μm InP HEMT.

Figure 7 contains similar data for 0.13 μm $\times$ 20 μm silicon NMOSFET, 0.15 μm $\times$ 50 μm GaAs PHEMT[7] and 1 μm $\times$ 4.2 μm InP HEMT. Although the slowest among all the devices shown, NMOSFET exhibits a pretty good high-frequency performance with the peak cut-off frequency of over 80 GHz and the maximum frequency of oscillations of 110 GHz. Finally, although not shown in Figs. 6 and 7, it needs to be mentioned that 0.1 μm InP HEMT device achieves $f_{\max}$ of over 300 GHz.

5. Design for manufacturability (DFM) of mixed-signal IC's

Intensive scaling down of transistor/wire feature size and supply and signal voltage levels in ICs generates enhanced interest in circuit manufacturability and yield. This has become of particular importance for analog and mixed-signal blocks of CMOS IC systems that are much more sensitive to noise and process variations than digital parts.

The classical yield-optimization approaches developed for digital ICs focus on taking into account the effects of manufacturing defects (lithography defects or spot defects) and manufacturing process variations that result in a spread of electrical performance. Spot defects are quite reliably modeled as spots of extra or missing material, introduced into the process by the inaccuracies of technological operations.

[6] The maximum frequency of oscillations $f_{\max}$ is defined as the frequency at which power gain is equal to 1.

[7] HEMT – high electron mobility transistor, a popular high-speed structure for III-V based transistors.

These spot defects can cause shorts or/and breaks in ICs, causing catastrophic defects that degrade IC yield. Possible yield degradation due to spot defects can be evaluated using the critical area approach [23, 24]. This methodology requires adetailed knowledge of defect size distribution in a given technology and circuit layout. Currently, this method finds application in evaluating the potential yield of digital cell libraries offered by different vendors [25].

Estimation of IC yield that takes manufacturing process variations into account should be based on measured process statistics and physical layer (device and circuit level) models that allow statistical modeling of circuit performance versus statistical process variations. This methodology is focused on predicting parametric yield, i.e. the yield resulting from the variation of electrical performance. The Monte-Carlo simulation [26] approach is usually computationally-intensive, which limits its application to circuits containing a relatively small number of MOS transistors. The corner analysis is a simplified way that takes manufacturing process variations into account and includes them sufficiently early in the circuit simulation; it is the one that is practically used throughout the industry.

In digital circuitry spot defects are the major source of catastrophic yield loss. However, the mixed-signal circuits featuring layout with usually larger feature sizes and not as densely packed as digital circuits, are much less affected by spot defects. Instead, they are more prone to parametric faults (and, as a result, parametric yield loss) that originate from such factors as manufacturing process variations, circuit layout (producing various parasitics), or digital noise propagation through the substrate. Consequently, the predictability of the parametric yield of mixed-signal blocks should include a concerted effort totake suspected yield-degrading effects into account as early as possible in the design flow.

It is expected that corner analysis will remain a valuable design tool to estimate the parametric yield of mixed-signal circuits. The validity of this analysis depends to a large extend on the quality of corner models. Most of the corner models used in the industry reflect 3-s (s = standard deviation) changes in the digital circuit performance (usually speed vs. power consumption) due to process variations. These models are also routinely used for evaluating the manufacturability of mixed-signal circuits. How adequate they are for the mixed-signal circuits, remains an open question. New corner models reflecting changes in mixed-signal performance would be more adequate for more accurate parametric yield prediction.

Our experience in mixed-signal CMOS design shows that circuit layout has been one of the major factors contributing to circuit performance and circuit sensitivity [27]. Different layout styles may result in different parasitics and different defect tolerances. Considering different layout solutions requires multiple layout designs, followed by layout extraction and post-layout simulation – a scenario that can not be easily achieved under the time to market pressure. It is believed, however, that this limitation can be mitigated through the automation of layout design of CMOS mixed-signal blocks.

Designing a high quality analog and mixed-signal circuit layout requires usually a human expert and is not a quick procedure. On the other hand, any attempt to automate the process must rely on some degree of layout regularity. But the regularity may be hard to achieve in mixed-signal circuits, where most of the MOS transistors have unique aspect ratios (W/L), and many of the transistors must have their I-V characteristics matched. The solution lies in adapting the fully-stacked analog CMOS layout approach [28]. In this technique, MOS transistors are placed in stacks, frequently sharing their sources and drains to minimize parasitic capacitances associated with implanted source/drain regions. Placing MOS transistors in stacks becomes possible due to initial transistor splitting along the channel width, so that a single MOS transistor featuring the channel width, W, is implemented as a parallel connection of NMOS transistor segments, each featuring channel width of W/N. The fully-stacked CMOS layout technique exploits grouping of MOS transistor segments with similar channel widths into stacks of equal height, that can be later positioned using one of the known placement and routing techniques. The fully-stacked CMOS layout allows also precise MOS transistor matching using common-centroid geometry.

Early attempts [29] to automate mixed-signal CMOS IC layout design aimed at making use of the fully-stacked layout that – by its own nature – minimized drain-substrate and source-substrate p-n junction capacitances. But it is not only the capacitances that need to be minimized. A very compact fully-stacked layout usually requires a more complex metal interconnection pattern. Since metal interconnects constitute another source of parasitic capacitances, the challenge is in making an intelligent tradeoff between the compactness of the fully-stacked layout and the number and lengths of metal interconnects, which is a tradeoff between p-n junction capacitances, and metal wire-to-substrate capacitances. The new methodology [30] proposes to use an automated layout design tool to generate multiple layouts of the same circuit, followed by a post-layout extraction and circuit simulation to evaluate circuit performance, and to choose the most effective layout solution. The same approach can be coupled in future with the manufacturability analysis of mixed-signal CMOS ICs.

As mixed-signal circuits grow in complexity, the need for development of behaviour-level models of mixed-signal blocks rises dramatically. These models can aid in efficient mixed-signal system design through proper system partitioning and mixing behaviour-level and circuit-level simulation within the same CAD tool. Behavioural models can be customized to allow statistical simulation, as well as noise modeling. However, technology-specific system implementations do require substantial model building and model characterization activity to fully utilize potential benefits that come from the use of behaviour-level design for manufacturability.

References

[1] L. Cloetens, "Broadband access: the last mile", in *Int. Solid-State Circ. Conf.*, 2001, Paper 1.2.

[2] R. Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*. Academic Press, 1998.

[3] N. Alvarez and K. Iniewski, "CMOS FEC: building cost-effective 10 Gbit/s metro networks", *Fiberopt. Prod. News*, Nov. 2002.

[4] K. Iniewski, "Dispersion compensation, forward error correction and 10 Gbit/s transponders for metro networks", *Lightw. Mag.*, Jan. 2003.

[5] A. Banerjee *et al.*, "Generalized multiprotocol label switching: an overview of routing and management enhancements", *IEEE Commun. Mag.*, pp. 144–150, Jan. 2001.

[6] D. Awduche and Y. Rekhter, "Multiprotocol Lambda switching: combining MPLS traffic engineering control with optical crossconnects", *IEEE Commun. Mag.*, pp. 111–116, Mar. 2001.

[7] K. Iniewski, "OEO conversion still a fact of life in telecom nets", *Lightwave*, May 2002.

[8] T. Ellermeyer *et al.*, "A 10 Gbit/s eye-opening monitor IC for decision-guided adaptation of the frequency response of an optical receiver", *J. Solid-State Circ.*, vol. 35, pp. 1958–1963, 2000.

[9] A. Neukermans *et al.*, "MEMS technology for optical networking applications", *IEEE Commun. Mag.*, pp. 63–69, Jan. 2001.

[10] N. McKeown, M. Izzard, A. Mekkittikul, B. Ellersick, and M. Horowitz, "The tiny tera: a packet switch core", in Hot Interconnects V, Stanford University, Aug. 1996 (also: *IEEE Micro*, pp. 26–33, Jan./Feb. 1997).

[11] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches", *IEEE/ACM Trans. Net.*, vol. 7, pp. 188–201, 1999.

[12] W. Bux *et al.*, "Technologies and building blocks for fast packet forwarding", *IEEE Commun. Mag.*, pp. 70–77, Jan. 2001.

[13] U. Black, *Emerging Communication Technologies*. Prentice Hall, 1997.

[14] J. Savoj and B. Rezavi, "A 10 Gbit/s CMOS clock and data recovery circuit with frequency detection", in *Int. Solid-State Circ. Conf.*, 2001, Paper 5.3.

[15] M. Reinhold *et al.*, "40 Gbit/s clock and data recovery/1:4 DEMUX IC in SiGe technology", in *Int. Solid-State Circ. Conf.*, 2001, Paper 5.6.

[16] R. He *et al.*, "A DSP based receiver for 1000 BASE-T PHY", in *Int. Solid-State Circ. Conf.*, 2001, Paper 19.6.

[17] A. Gattani *et al.*, "A CMOS HDSL2 analog front-end", *J. Solid-State Circ.*, vol. 35, pp. 1964–1975, 2000.

[18] P. Laaser *et al.*, "A 285 mW CMOS single chip analog front end for G.SHDSL", in *Int. Solid-State Circ. Conf.*, 2001, Paper 19.1.

[19] S. Voinigescu, P. Popescu, P. Banens, M. Copeland, G. Fortier, K. Howlett, M. Herod, D. Marchesan, J. Showell, S. Szilagyi, H. Tran, and J. Weng, "Circuits and technologies for highly integrated optical networking IC's at 10 Gbit/s and 40 Gbit/s", in *Proc. IEEE CICC*, 2001, pp. 331–338.

[20] S. P. Voinigescu, D. S. Mc.Pherson, F. Pera, S. Szilagyi, M. Tazlauanu, and H. Tran, "A comparison of silicon and III-V technology performance and building block implementations for 10 and 40 Gbit/s optical networking ICs", *Int. J. High Speed Electron. Syst.*, vol. 13, no. 1, 2003.

[21] E. Sano, "High-speed lightwave communication ICs based on III-V compound semiconductors", *IEEE Commun. Mag.*, pp. 154–158, Jan. 2001.

[22] M. Jurczak, T. Skotnicki, M. Paoli, B. Tormen, J. Martins, J. Regolini, D. Dutarte, P. Ribot, D. Lenoble, R. Pantel, and S. Monfray, "Silicon-on-nothing (SON) – an innovative process for advanced CMOS", *IEEE Trans. Electron Dev.*, vol. 47, p. 2179, 2000.

[23] W. Maly, "Modeling of lithography related yield losses for CAD of VLSI", *IEEE Trans. CAD*, vol. CAD-4, no. 3, pp. 166–177, 1985.

[24] S. W. Director, W. Maly, and A. Strojwas, *VLSI Design for Manufacturing: Yield Enhancement*. Boston: Kluwer, 1989.

[25] H. T. Heineken, J. Khare, and M. d'Abreau, "Manufacturability analysis of standard cell libraries", in *Proc. IEEE Custom ICs Conf.*, 1998, pp. 321–324.

[26] D. H. M. Walker, *Yield Simulation for Integrated Circuits*. Kluwer, 1987.

[27] R. Ockey and M. Syrzycki, "Analysis of manufacturability factors for analog CMOS A/D converter building blocks", *Analog Integr. Circ. Sig. Proc. Int. J.*, Kluwer, vol. 26, no. 3, pp. 239–255, 2001.

[28] U. Gatti, F. Maloberti, and V. Liberali, "Full stacked layout of analogue cells", in *Proc. IEEE Int. Symp. Circ. Syst.*, 1989, pp. 1123–1126.

[29] R. Naiknaware and T. S. Fiez, "Automated hierarchical CMOS analog circuit stack generation with intramodule connectivity and matching consideration", *IEEE J. Solid-State Circ.*, vol. 34, no. 3, pp. 304–316, 1999.

[30] P. Khademsameni and M. Syrzycki, "Manufacturability analysis of analog CMOS ICs through examination of multiple layout solutions", in *Proc.17th IEEE Int. Symp. Def. Fault Toler. VLSI Syst.*, Vancouver, Canada, 2002, pp. 3–11.

Krzysztof Iniewski is a founder of silicoMOS a consulting firm in Vancouver (Canada) and has over 15 years of technical experience in semiconductor and communication IC industry. From 1995 to 2003, he was with PMC-Sierra and held various technical and management roles in analog development, design services and strategic marketing. Prior to joining PMC-Sierra, he was an Assistant Professor at the University of Toronto's Electrical Engineering and Computer Engineering Department. Dr. Iniewski has published over 64 research papers in international journals and numerous conference presentations. He holds 10 international patents granted in the USA, Canada, France, Germany, and Japan. In 1988 he received his Ph.D. degree in electronics (honours) from the Warsaw University of Technology (Poland).

e-mail: iniewski@ieee.org
School of Engineering Science
Simon Frazer University
8888 University Drive
Burnaby, British Columbia, V5A 1S6, Canada

Marek Syrzycki received a M.Sc. degree in electronic engineering from the Warsaw University of Technology (Poland) in 1970, and a Ph.D. degree in electrical engineering from the Institute of Electron Technology, Research and Production Centre of Semiconductor Devices (Poland) in 1978. In 1979, he joined the Warsaw University of Technology as an Assistant Professor. From 1986 to 1988 he was a visiting research associate in the Department of Electrical and Computer

Engineering, Carnegie-Mellon University, Pittsburgh (USA), where he worked on IC layout extraction, process simulation, and defect modeling. Dr. Syrzycki joined the Simon Fraser University in 1988, where he is currently a Professor in the School of Engineering Science. He is an author or co-author of 2 patents and over 60 scientific publications in journals, books and at conferences. His research interests are in analog and mixed-signal VLSI IC's design, integrated microsensors, intelligent vision sensors, IC fabrication defects, design for manufacturability of analog CMOS IC's, and design of CAD tools for VLSI IC's. He is a member of the IEEE and the IEEE Electron Devices Society.

e-mail: marek@cs.sfu.ca
Centre for Systems Science
Simon Frazer University
8888 University Drive
Burnaby, British Columbia, V5A 1S6, Canada

Sorin P. Voinigescu graduated in 1984 with a M.Sc. degree in electronics from the Polytechnic Institute of Bucharest (Romania). He received the Ph.D. degree in electrical and computer engineering from the University of Toronto (Canada) in 1994. His Ph.D. dissertation was on the design and fabrication of VLSI compatible Si/SiGe PMOSFET's. Between 1984 and 1991 he worked in R&D and in Academia in Bucharest, designing and lecturing on microwave semiconductor devices and microwave integrated circuits. Between 1994 and 2000 he was with Nortel Networks in Ottawa, where he was responsible for projects in high frequency characterization and statistical scalable compact model development for Si, SiGe and III-V heterostructure devices. He spearheaded the modeling infrastructure development for, and was involved in the prototyping of wireless and broadband fiber optics transceivers in emerging semiconductor technologies. In April 2000 he co-founded Quake Technologies Inc. an Ottawa-area fabless semiconductor company focussing on the design and fabrication of 10 Gbit/s and 40 Gbit/s physical layer ICs. As chief technology officer at Quake he coordinated the access and characterization of Si, SiGe, GaAs and InP technologies, high frequency package design and electro-optical interface product development. In September 2002 he joined the Electrical and Computer Engineering Department at the University of Toronto as an Associate Professor. His research and teaching interests focus on the modeling and characterization of very deep submicron semiconductor devices and on novel design techniques and low-voltage, low-power topologies for wireless, optical fiber and wireline data communication physical layer integrated circuits in the 10 GHz to 100 GHz range. Dr. Voinigescu has authored or co-authored more than 30 refereed and invited technical papers spanning the simulation, modeling, design and fabrication of GaAs- InP- and Si-based heterostructure devices and circuits and holds two US patents in these areas. He is the co-recipient of the Best Paper Award at the 2001 IEEE Custom Integrated Circuits Conference, a Senior Member of the IEEE and a member of the TPC of the GaAs IC Symposium.

e-mail: sorinv@eecg.toronto.edu
Department of Electrical & Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario, M5S 3G4, Canada

Invited paper

Evolution and recent advances in RF/microwave transistors

Juin J. Liou and Frank Schwierz

Abstract — Most applications for radio frequency/microwave (thereafter called RF) transistors had been military oriented in the early 1980s. Recently, this has been changed drastically due to the explosive growth of the markets for civil wireless communication systems. This paper gives an overview on the evolution, current status, and future trend of transistors used in RF electronic systems. Important background, development and major milestones leading to modern RF transistors are presented. The concept of heterostructure, a feature frequently used in RF transistors, is discussed. The different transistor types and their figures of merit are then addressed. Finally an outlook of expected future developments and applications of RF transistors is given.

Keywords — microwave devices, RF devices, heterostructures, HEMT, HBT, frequency limits, RF CMOS.

1. Introduction

Currently RF electronics is most likely the fastest growing segment of semiconductor industry. This is due to the explosive growth in the wireless communications market during the past 10 years. Currently, there is a string of further applications which are either already commercially available or are expected to come to market in the near future. Examples are the 3rd generation cellular phones with extended functionality (e.g. mobile internet access), satellite communication services such as direct broadcast satellite (DBS) and local multipoint distribution system (LMDS), and local area networks such as wireless local area network (WLAN) and wireless personal area network (WPAN), also known as Bluetooth.

About 20 years ago, the situation was much different. During that time, RF electronics was somewhat mysterious, and their applications had been mainly military (e.g. secure communications, electronic warfare systems, missile guidance, control electronics for smart ammunition, radars) and the funding for their development came mainly from government agencies. In the first half of the 1980s, satellite television using low-noise transistors operating at 12 GHz in the receiver front-ends emerged as the first civil application of RF transistors with a market volume worth mentioning.

The backbone of RF systems is high-speed transistors with the capability of operating at GHz frequencies. In the following sections we will introduce important figures of merit (FOMs) of RF transistors and cover the evolution of these devices. This will be followed by the discussions of different types of RF transistors and major milestones of their development. Finally, the year 2003 state of the art of RF transistors will be highlighted, and an outlook of expected future development will be given.

2. RF transistor FOMs

The term RF stands for radio frequency and is commonly designated as electromagnetic waves with frequencies around and above 1 GHz. Thus RF transistors are devices with the capability to operate and amplify signals at GHz frequencies.

RF transistors are used in a large number of different circuits, such as low-noise and power amplifiers, mixers, oscillators, frequency converters, etc. Although the requirements on transistor performance differ from application to application, RF transistors in general can be divided into two groups: small-signal low-noise transistors and power transistors. For low-noise transistors, very low noise in the transistor and high operating frequency are desired, while power transistors are designed for high output power at high operating frequency.

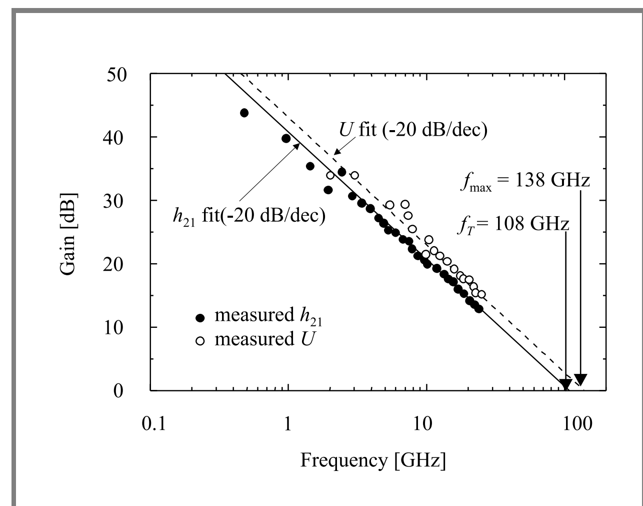


Fig. 1. Current gain, unilateral power gain, and extrapolated f_T and f_{max} of a GaAs MESFET. Data taken from [4].

RF engineers use several different FOMs to characterize transistors. These FOMs are the characteristic frequencies including the cutoff frequency f_T and maximum frequency for oscillation f_{max}, the minimum noise figure NF_{min}, and the RF output power P_{out} [1, 2]. The cutoff frequency is the frequency at which the small signal current gain of the transistor, h_{21}, becomes unity (i.e., 0 dB). The max-

imum frequency of oscillation, on the other hand, is the frequency at which the unilateral power gain of the transistor, U, becomes unity. Both h_{21} and U are frequency dependent and roll off with a slope of -20 dB/dec at high frequencies. Figure 1 shows the measured h_{21} and U of a typical RF transistor. As can be seen, f_T and $f_{\max}$ can be estimated by extrapolation of the lower frequency data to the frequency axis using the known -20 dB/dec slope. This practice is not only convenient but in some cases inevitable because of the frequency limit of the measurement equipment. P. Greiling [3] stated in his review of the history of GaAs field-effect transistor (FET): "For those of us associated with this technology, this measurement problem always seems to exist. We are in a catch 22 situation in which we are developing circuits for instruments that are needed to measure the circuits we are developing".

3. Period between 1960 and 1980

Since the invention of the bipolar junction transistor (BJT) in 1947, device engineers have devoted a lot of efforts to increase the speed and operating frequency of transistors. The first transistors capable of amplifying signals in the frequency range around 1 GHz were Ge BJTs developed in the late 1950s. Soon after that, Si and GaAs BJTs had been exploited for high-frequency applications, and the Si BJT became the dominating transistor type in RF electronics [5]. In 1970, the state of the art Si BJTs showed minimum noise figures of 1.3, 2.6 and 4 dB at frequencies of 1, 2, and 4 GHz, respectively, and output powers of 100, 20, and 5 W at frequencies of 1.2, 2, and 4 GHz, respectively [5].

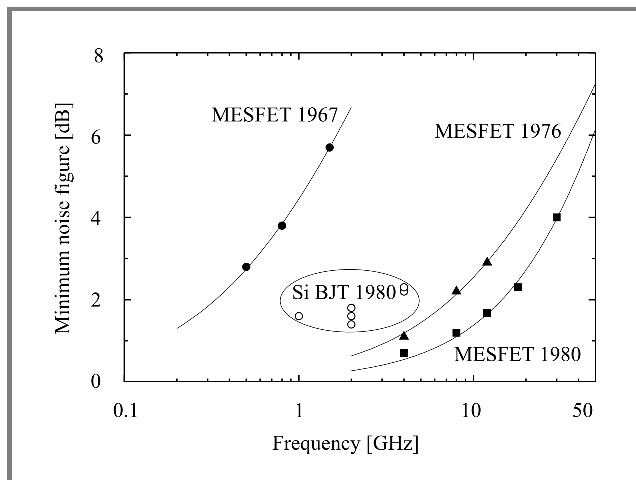


Fig. 2. State of the art Si BJTs and GaAs MESFETs in terms of the minimum noise figure reported in 1980.

The development of GaAs BJTs, however, had only limited success. By 1968, the interest in these transistors declined and the research activities stopped [6]. More success had achieved for GaAs FETs, however. In 1966, C. Mead presented the first GaAs metal-semiconductor FET (MESFET) and laid the foundation for a revolution in RF electronics [7]. One year later, a GaAs MESFET with $f_{\max}$ of 3 GHz was reported [8]. In 1970, a GaAs MESFET with a record $f_{\max}$ around 30 GHz was obtained [9], which clearly exceeded the performance of any other transistor type at that time, and in 1973 the 100 GHz $f_{\max}$ mark was reached [10]. Both low-noise and power GaAs MESFETs became commercially available in the mid 1970s.

Si BJTs and GaAs MESFETs were the only RF transistor types available in the late 1970s. Si BJTs were commonly used at frequencies below 4 GHz, whereas in the frequency range between 4 and 18 GHz the GaAs MESFET was the device of choice. Figure 2 shows the minimum noise figures of Si BJTs and GaAs MESFETs developed in this period.

4. Period between 1980 and 2000

4.1. Development of III-V HEMTs

In the late 1970s experiments at Bell Labs revealed the existence of a two-dimensional electron gas (2DEG) in epitaxially grown heterostructures consisting of undoped GaAs and n-doped AlGaAs. Both materials have the same lattice constant, thus resulting in a lattice matched heterostructure. The measured electron mobility in the 2DEGs was much higher than that in bulk GaAs [11]. The underlying physics is shown in Fig. 3. Electrons transfer from the conduction band of the doped AlGaAs to the energetically lower conduction band of the undoped GaAs. This transfer creates an electric field and band bendings at the heterointerface. The transferred electrons are confined in a narrow potential well on the GaAs side and are spatially separated from the donor ions. Thus ionized impurity scattering is largely suppressed, a mechanism leading to a high electron mobility.

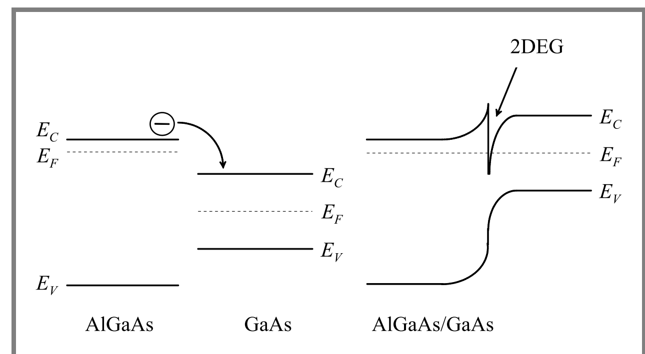


Fig. 3. Band diagram of the AlGaAs/GaAs heterostructure in HEMTs.

Engineers were interested in developing a transistor structure taking the advantage of high 2DEG mobility. The basic

idea came again from Bell Labs [12], but the first successful realization of such a device was reported by researchers at Fujitsu [13]. The Bell group called their device selectively doped heterostructure transistor (SDHT), while the name of the Fujitsu device was high electron mobility transistor (HEMT). The naming of this transistor became even more confusing because other groups reporting experimental transistors of the same kind called their devices modulation doped FET (MODFET, University of Illinois, Rockwell) and two-dimensional electron gas FET (TEGFET, Thomson). The name HEMT prevails and is widely used by the RF community.

Early HEMTs consisted of the AlGaAs/GaAs material system. They showed better RF performance compared to GaAs MESFETs, especially in terms of minimum noise figure and output power, but the performance improvement was less than anticipated. One of the targets in HEMT design is the combination of a high electron mobility μ_0 with a high 2DEG electron sheet density n_s. It was found later that, by replacing the GaAs layer with an InGaAs layer, the product $\mu_0 \times n_s$ can be considerably increased. Thus, in the mid 1980s, the AlGaAs/InGaAs heterostructure was introduced in HEMTs, and the most prominent types are the AlGaAs/InGaAs/GaAs and InAlAs/InGaAs/InP HEMTs.

The lattice constant of InGaAs is larger than that of AlGaAs and GaAs. When grown on GaAs substrate, the atoms of the InGaAs layer can be adjusted to accommodate the GaAs lattice, thus resulting in a strained layer (frequently called pseudomorphic layer), provided the InGaAs layer is thinner than the so-called critical thickness t_c. For $In_{0.2}Ga_{0.8}As$ layers, which are typical for the GaAs pseudomorphic HEMT (PHEMT), t_c is about 20 nm. The heterostructure system $In_{0.52}Al_{0.48}As/In_xGa_{1-x}As$ grown on InP substrate is lattice matched for $x = 0.53$. A further increase in the In content, i.e. $x > 0.53$, results in a strained layer as well. Figure 4 shows the reported $\mu_0 \times n_s$ products for various heterostructures, which clearly demonstrates the advantage of high In-content layers [14].

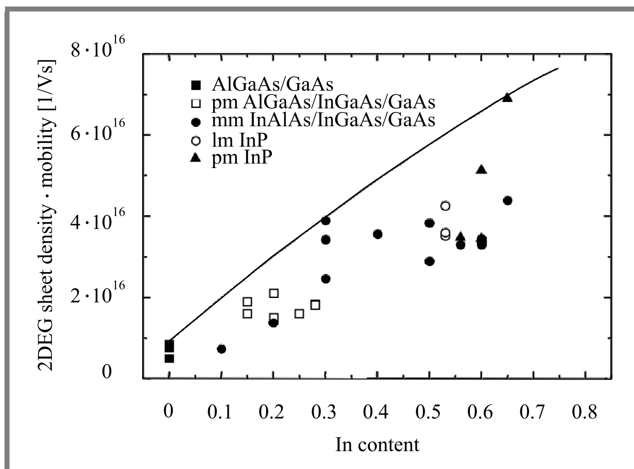


Fig. 4. Product $\mu_0 \times n_s$ versus In content in different HEMTs: lm – lattice matched, pm – pseudomorphic, mm – metamorphic.

GaAs PHEMT became commercially available in the early 1990s and are now in widespread use for both low-noise and power amplifications. Figure 5 shows the reported f_T and f_{max} of GaAs PHEMTs [14].

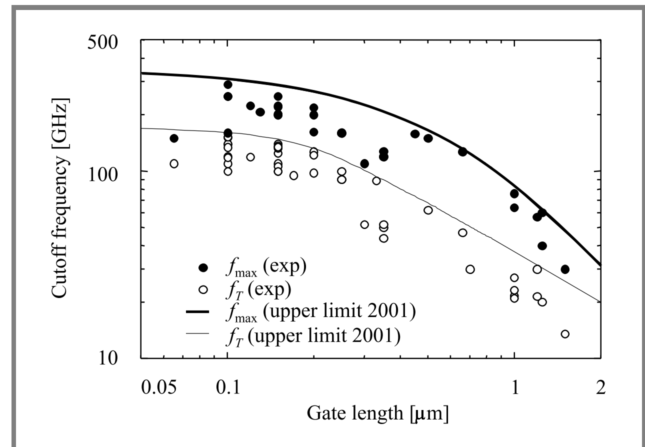


Fig. 5. Reported f_T and f_{max} versus gate length for different GaAs PHEMTs.

Although InP HEMTs show even better RF performance compared to GaAs PHEMTs, these transistors still await for commercialization. The main reasons are the low degree of maturity of InP technology and the InP substrates, which are expensive and available only in small diameters. Nevertheless, InP HEMTs possess the lowest noise figures and the highest operating frequencies among all field-effect transistors.

4.2. Development of III-V HBTs

The basic idea to use a heterostructure in bipolar transistors is almost as old as the bipolar transistor itself. In 1948, W. Shockley described the advantage of a bipolar transistor consisting of a wide bandgap emitter and a narrow bandgap base [15]. The physical effect exploited in a heterostructure bipolar transistor (HBT) is shown in Fig. 6. Because of the bandgap difference at the emitter-base heterjunction, electrons moving from the emitter to the base encounter a smaller energy barrier to be surmounted compared to that to be surmounted by holes moving from the base to the emitter. Thus, hole injection into the emitter is effectively reduced. This effect allows for the realization of a very thin and highly doped base layer, thereby leading to a short base transit time and low base resistance. As a consequence, HBTs with extremely high f_T and f_{max} values are possible.

It took more than 30 years to materialize Shockley's idea in practical devices, as the advance in epitaxial growth technology, especially the development of molecular beam epitaxy (MBE), allowed for the growth of high-quality heterostructures and the realization of GaAs HBTs in the early 1980s. To date, GaAs HBTs with AlGaAs and InGaP emitters are commercially available and used mostly

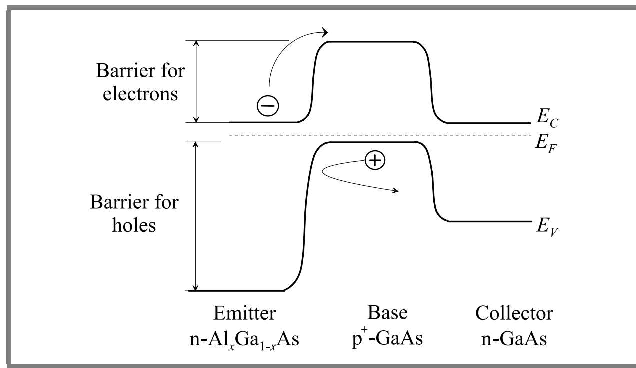


Fig. 6. Band diagram of the AlGaAs/GaAs heterostructure in HBTs.

for power amplification in wireless communication systems.

Much effort has also been spent on the development of the InP HBT, which consists of an InAlAs emitter, InGaAs base, and either InGaAs or InP collector. InP HBTs show higher $f_T's$ and $f_{max}'s$ than GaAs HBTs but are not yet commercially available. Recently an interesting and novel InP HBT utilizing the substrate transfer has been reported [16]. This concept dramatically reduces the size of the extrinsic transistor, which minimizes the collector-base capacitance, and results in an extremely high f_{max}. A transferred substrate HBT with an extrapolated f_{max} of more than 1 THz has been reported [17]. This is the highest f_{max} ever obtained from a three-terminal semiconductor device.

Because of economical reasons, it is always desirable to use Si-based devices instead of III-V devices, provided that the performance of the Si-based devices is adequate. A major step to use Si-based transistors at frequencies above 4 GHz was the development of SiGe HBTs. These transistors consist of a strained SiGe base layer embedded between the Si emitter and collector. The first SiGe HBT was reported in 1987 [18], and the RF performance of SiGe HBTs has improved continuously since. Currently SiGe HBTs are commercially available, and both f_T and f_{max} of advanced laboratory SiGe HBTs are exceeding the 300 GHz mark.

4.3. Further developments

Three directions in RF transistor research during the 1990s are worth mentioning. The first is the availability of the Si MOSFET as an RF device. Despite the fact that the Si MOSFET had not been considered seriously in the past for RF applications due to its relatively low speed, the continuous scaling and increasing maturity of short-gate Si MOS technology in recent years has led the MOSFET to become a strong candidate for applications in the lower GHz range. In fact, the topic RF CMOS was frequently discussed at all major device conferences around the world since the mid-1990s. Meanwhile Si laterally diffused MOSFET (LDMOSFET) for high-power applications up to 2.5 GHz and small-signal RF CMOS circuits are commercially available.

The second direction is the investigation of wide bandgap semiconductors, such as SiC and III-nitrides, for use in RF power transistors with large output powers in the GHz range. The wide bandgaps of these materials (3.2 eV for SiC and 3.4 eV for GaN compared to 1.1 eV for Si and 1.4 eV for GaAs) result in high breakdown fields and high operating temperatures for wide bandgap transistors. Most prominent devices are SiC MESFETs and AlGaN/GaN HEMTs. SiC MESFETs became commercially available in 1999, and AlGaN/GaN HEMTs with f_T and f_{max} exceeding 100 GHz and extremely high output power densities (output power per mm gate width) have been reported.

Finally, the so-called GaAs metamorphic HEMT (MHEMT) should be mentioned [19]. The key feature of this transistor is an InGaAs channel layer grown on GaAs substrates with an In content higher than that used in GaAs PHEMTs. This is done using a thick relaxed InGaAs buffer layer serving as a relaxed pseudosubstrate for the actual device layer grown on top of the buffer. The main advantage of the metamorphic approach is that inexpensive GaAs substrates can be used to obtain InP HEMT like performance.

Furthermore, the conventional Si BJTs and GaAs MESFETs have been improved in terms of RF performance and maturity.

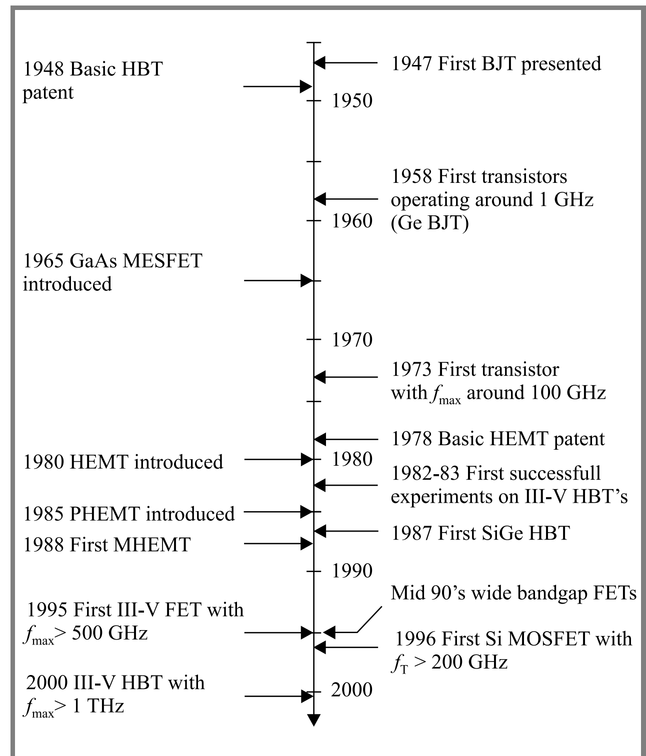


Fig. 7. Major milestones for the evolution of RF transistors.

Figure 7 summaries the major milestones of the evolution of RF transistors during the past four decades.

5. State of the art of RF transistors – 2003

During the more than 40 years of RF transistor development, the operating frequency has been increased continuously. This became possible by shrinking the critical device dimensions, introducing heterostructures, and exploiting the properties of new semiconductor materials.
Table 1 lists the state of the art performance in terms of f_T and $f_{\max}$ for different types of RF transistors. Two remarks should be made in regard to the values in Table 1. First, the $f_T - f_{\max}$ pairs do not necessarily belong to the same transistor. Second, the values represent the performance of laboratory test devices. Commercial devices possess lower f_T and $f_{\max}$ values, because for these devices not only high performance but also cost, yield, and reproducibility are of concern.

Table 1
State of the art of RF transistors in terms of f_T and $f_{\max}$

Transistor type	f_T [GHz]	Ref.	$f_{\max}$ [GHz]	Ref.
GaAs MESFET	168	20	177	21
AlGaAs/GaAs HEMT	113	22	151	23
GaAs PHEMT	152	24	290	25
InP HEMT	562	26	600	27
SiC MESFET	22	28	50	28
AlGaN/GaN HEMT	121	29	195	30
Si MOSFET	245	31	193	32
SiGe HBT	350	33	338	34
GaAs HBT	156	35	350	36
InP HBT	377	37	478	38
InP HBT (TS)	300	39	1080	17

TS – transferred substrate.

To date, InP transistors possess the highest $f_T^{\prime}s$, the highest $f_{\max}^{\prime}s$, and thus the highest operating frequencies of all transistor types. InP HEMTs show the lowest noise figure among all RF transistors. Minimum noise figures below 1 dB at 60 GHz and of 1.2 dB at 94 GHz have been reported for InP HEMTs having a 0.1 μm gate length [40]. GaAs MHEMTs show only slightly higher noise figures than InP HEMTs. State of the art GaAs PHEMTs with noise figures less than 1 dB up to 30 GHz have been reported [41]. In general, FETs are less noisy compared to bipolar transistors. On the other hand, bipolar transistors possess higher output power densities (both per unit chip area and per mm device width) than field effect transistors. For example, output power densities of 10 mW per μm^2 emitter area and 30 W per mm emitter length have been achieved in a GaAs HBT [42]. In the case of power FETs, wide bandgap FETs demonstrate the highest output power densities up to 20 GHz. An AlGaN/GaN HEMT with 11.2 W/mm at 10 GHz has been realized [43]. Currently the total output power of AlGaN/GaN HEMTs is, however, still lower compared to SiC MESFETs and III-V power HEMTs. Work is under way to realize large area AlGaN/GaN HEMTs with high output powers. The only transistor types delivering useful output power above 60 GHz are GaAs PHEMTs and InP HEMTs. A main drawback of InP HEMTs is the low breakdown voltage stemmed from the narrow bandgap of the In-rich InGaAs layers.

6. Future outlook and developments

During the 1970s and 1980s, military applications dominated RF electronics, and the device performance was the major concern while the economical factor played only a secondary role. The situation has changed dramatically recently, as the new global political situation in the 1990s has led to considerable cuttings in the military budgets. Furthermore, a strong shift to consumer applications is taking place. Thus, the design philosophy for most RF systems has changed from "performance at any price" to "sufficient performance at lowest cost".
Prior to 1980, only two transistor types (Si BJT and GaAs MESFET) existed. In 2003, a large variety of different competing devices and technologies is available, including Si CMOS, Si BJT, SiGe HBT, GaAs MESFET, GaAs HEMT, GaAs HBT, InP HEMT, InP HBT, and wide bandgap FETs. Each of the different transistor types has certain advantages and disadvantages in terms of maturity, cost, and performance. The situation of the circuit designer can be described by "the wider the choice, the greater the difficulty". In the mass consumer markets (operating frequencies up to 2.5 GHz), all technologies can compete, but Si-based technologies have a clear cost advantage. Most applications above 2.5 GHz belong to GaAs-based transistors (MESFET, HEMT, HBT). High-performance applications above 40 GHz are dominated by InP-based transistors. The role of RF CMOS is expected to grow in the future. The upper frequency limits of Si-based technologies will increase, due to the MOSFET scaling and the use of SiGe HBTs. The importance of GaAs PHEMTs and HBTs will continue to grow as well. For some specific applications, the commercial use of InP devices can be expected. On the other hand, the market share of the GaAs MESFET is shrinking, and this trend will continue for many years to come.
We conclude with two simple statements. First, in the foreseeable future, the dynamic growth of RF electronics will continue and new applications will emerge. Second, RF transistors are no longer exotic but becoming more and more mainstream devices.

References

[1] F. Schwierz and J. J. Liou, "Semiconductor devices for RF applications: evolution and current status", *Microelectron. Reliab.*, vol. 41, pp. 145–168, 2001.

[2] F. Schwierz and J. J. Liou, *Modern Microwave Transistors: Theory, Design, and Performance*. Hoboken, New Jersey: Wiley, 2003.

[3] P. Greiling, "The historical development of GaAs FET digital IC technology", *IEEE Trans. Microw. Theory Techn.*, vol. 32, pp. 1144–1156, 1984.

[4] K. Onodera, M. Tokumitsu, M. Tomizawa, and K. Asai, "Effects of neutral buried p-layer on high-frequency performance of GaAs MESFET's", *IEEE Trans. Electron Dev.*, vol. 38, pp. 429–436, 1991.

[5] H. F. Cooke, "Microwave transistors: theory and design", *Proc. IEEE*, vol. 59, pp. 1163–1181, 1971.

[6] H. T. Yuan, W. V. McLevige, and H. D. Shih, "GaAs bipolar digital integrated circuits", in *VLSI Electronics: Microstructure Science*, N. G. Einspruch and W. R. Wisseman, Eds. Orlando: Academic-Press, 1985, vol. 11, pp. 173–213.

[7] C. A. Mead, "Schottky barrier gate field effect transistor", *Proc. IEEE*, vol. 59, pp. 307–308, 1966.

[8] W. W. Hooper and W. I. Lehrer, "An epitaxial GaAs field-effect transistor", *Proc. IEEE*, vol. 55, pp. 1237–1238, 1967.

[9] K. Drangeid, R. Sommerhalder, and W. Lehrer, "High-speed gallium-arsenide Schottky-barrier field-effect transistors", *Electron. Lett.*, vol. 6, pp. 228–229, 1970.

[10] W. Baechtold, K. Daetwyler, T. Forster, T. O. Mohr, W. Walter, and P. Wolf, "Si and GaAs 0.5 μm gate Schottky-barrier field-effect transistors", *Electron. Lett.*, vol. 9, pp. 232–234, 1973.

[11] R. Dingle, H. L. Störmer, A. C. Gossard, and W. Wiegman, "Electron mobilities in modulation-doped semiconductor heterojunction superlattices", *Appl. Phys. Lett.*, vol. 33, pp. 665–668, 1978.

[12] R. Dingle, A. C. Gossard, and H. L. Stormer, U.S. patent 4,163,237, filed Apr. 24, 1978, issued July 31, 1979.

[13] T. Mimura, S. Hiyamizu, T. Fujii, and K. Nanbu, "A new field-effect transistor with selectively doped GaAs/n-$Al_xGa_{1-x}As$ heterojunctions", *Jpn. J. Appl. Phys.*, vol. 19, pp. L225–L227, 1980.

[14] F. Schwierz, "Microwave transistors: state of the art in the 1980s, 1990s, and 2000s" (unpublished).

[15] W. Shockley, U.S. patent 2,569,347, filed June 26, 1948, issued Sept. 25, 1951.

[16] Q. Lee, B. Agarwal, D. Mensa, R. Pullela, J. Guthrie, L. Samoska, and M. J. W. Rodwell, "A > 400 GHz $f_{\max}$ transferred-substrate heterojunction bipolar transistor IC technology", *IEEE Electron Dev. Lett.*, vol. 19, pp. 77–79, 1998.

[17] M. J. W. Rodwell, M. Urteaga, T. Mathew, D. Scott, D. Mensa, Q. Lee, J. Guthrie, Y. Betser, S. C. Martin, R. P. Smith, S. Jaganathan, S. Krishnan, S. I. Long, R. Pullela, B. Agarwal, U. Bhattacharya, L. Samoska, and M. Dahlstrom, "Submicron scaling of HBTs", in *IEEE Trans. Electron Dev.*, vol. 48, pp. 2606–2624, 2001.

[18] S. S. Iyer, G. L. Patton, S. L. Delage, S. Tiwari, and J. M. C. Storck, "Silicon-germanium base heterojunction bipolar transistors by molecular beam epitaxy", in *Tech. Dig. IEDM*, Washington, 1987, pp. 874–877.

[19] W. E. Hoke, P. J. Lemonias, J. J. Mosca, P. S. Lyman, A. Torabi, P. F. Marsh, R. A. McTaggart, S. M. Lardizabad, and K. Helzar, "Molecular beam epitaxial growth and device performance of metamorphic high electron mobility transistor structures fabricated on GaAs substrates", in *J. Vac. Sci. Technol.*, vol. B17, pp. 1131–1135, 1999.

[20] K. Tokumitsu, M. Hirano, T. Otsuji, S. Yamaguchi, and K. Yamasaki, "A 0.1-μm self-aligned-gate GaAs MESFET with multilayer interconnection structure for ultra-high-speed IC's", in *Tech. Dig. IEDM*, San Francisco, 1996, pp. 211–214.

[21] D. C. Caruth, R. L. Shimon, M. S. Heins, H. Hsia, Z. Tang, S. C. Chen, D. Becher, J. J. Huang, and M. Feng, "Low-cost 38 and 77 GHz CPW MMICS using ion-implanted GaAs MESFETs", in *IEEE MTT-S Dig.*, Boston, 2000, pp. 995–998.

[22] A. N. Lepore, H. M. Levy, R. C. Tiberio, P. J. Tasker, H. Lee, E. D. Wolf, L. F. Eastman, and E. Kohn, "0.1 μm gate length MODFETs with unity current gain cutoff frequency above 110 GHz", *Electron. Lett.*, vol. 24, pp. 364–366, 1988.

[23] I. Hanyu, S. Asai, M. Nukokawa, K. Joshin, Y. Hirachi, S. Ohmura, Y. Aoki, and T. Aigo, "Super low-noise HEMTs with a T-shaped WSi_x gate", *Electron. Lett.*, vol. 24, pp. 1327–1328, 1988.

[24] L. D. Nguyen, P. J. Tasker, D. C. Radulescu, and L. F. Eastman, "Characterization of ultra-high-speed pseudomorphic AlGaAs/InGaAs (on GaAs) MODFETs", *IEEE Trans. Electron Dev.*, vol. 36, pp. 2243–2248, 1989.

[25] K. L. Tan, R. M. Dia, D. C. Streit, T. Liu, T. Q. Trinh, A. C. Han, P. H. Liu, P.-M. D. Chow, and H.-C. Yun, "94-GHz 0.1-μm T-gate low-noise pseudomorphic InGaAs HEMT's", *IEEE Electron Dev. Lett.*, vol. 11, pp. 585–587, 1990.

[26] Y. Yamashita, A. Endoh, K. Shinohara, K. Hikosaka, T. Matsui, S. Hiyamizu, and T. Mimura, "Pseudomorphic $In_{0.52}Al_{0.48}As$/$In_{0.7}Ga_{0.3}As$ HEMTs with an ultrahigh f_T of 562 GHz", *IEEE Electron Dev. Lett.*, vol. 23, pp. 573–575, 2002.

[27] P. M. Smith, S.-M. J. Liu, M.-Y. Kao, P. Ho, S. C. Wang, K. H. G. Duh, S. T. Fu, and P. C. Chao, "W-band high efficiency InP-based power HEMT with 600 GHz $f_{\max}$", *IEEE Microw. Guid. Wave Lett.*, vol. 5, pp. 230–232, 1995.

[28] S. T. Allen, R. A. Sadler, T. S. Alcorn, J. Sumakeris, R. C. Glass, C. H. Carter Jr., and J. W. Palmour, "Silicon carbide MESFET's for high-power S-band applications", *Mat. Sci. Forum*, vol. 264-268, pp. 953–956, 1998.

[29] V. Kumar, W. Lu, R. Schwindt, A. Kuliev, G. Simin, J. Yang, M. A. Khan, and I. Adesida, "AlGaN/GaN HEMTs on SiC with f_T of over 120 GHz", *IEEE Electron Dev. Lett.*, vol. 23, pp. 455–457, 2002.

[30] R. Quay, R. Kiefer, F. van Raay, H. Massler, S. Ramberger, S. Müller, M. Dammann, M. Mikulla, M. Schlechtweg, and G. Weimann, "AlGaN/GaN HEMTs on SiC operating at 40 GHz", in *Tech. Dig. IEDM*, San Francisco, 2002, pp. 673–676.

[31] H. S. Momose, E. Morifuji, T. Yoshimoto, T. Ohguro, M. Saito, and H. Iwai, "Cutoff frequency and propagation delay time of 1.5-nm gate oxide CMOS", *IEEE Trans. Electron Dev.*, vol. 48, pp. 1165–1174, 2001.

[32] S. Narashima, A. Ajmera, H. Park, D. Schepis, N. Zamdmer, K. A. Jenkins, J.-O. Plouchart, W.-H. Lee, J. Mezzapelle, J. Bruley, B. Doris, J. W. Sleight, S. K. Fung, S. H. Ku, A. C. Mocuta, I. Yang, P. V. Gilbert, K. P. Muller, P. Agnello, and J. Welser, "High performance sub-40 nm CMOS devices for the 70 nm technology node", in *Tech. Dig. IEDM*, Washington, 2001, pp. 625–628.

[33] J.-S. Rieh, B. Jagannathan, H. Chen, K. T. Schonenberg, D. Angell, A. Chinthakindi, J. Florkey, F. Golan, D. Greenberg, S.-J. Jeng, M. Khater, F. Pagette, C. Schnabel, P. Smith, A. Stricker, K. Vaed, R. Volant, D. Ahlgreen, G. Freeman, K. Stein, and S. Subbanna, "SiGe HBTs with cut-off frequency of 350 GHz", in *Tech. Dig. IEDM*, San Francisco, 2002, pp. 771–774.

[34] B. Jagannathan, M. Meghelli, K. Chan, J.-S. Rieh, K. Schonenberg, D. Ahlgreen, S. Subbanna, and G. Freeman, "3.9 ps SiGe HBT ECL ring oscillator and transistor design for minimum gate delay", *IEEE Electron Dev. Lett.*, vol. 24, pp. 324–326, 2003.

[35] T. Oka, K. Hirata, K. Ouchi, H. Uchiyama, T. Taniguchi, K. Mochizuki, and T. Nakamura, "Advanced performance of small-scaled InGaP/GaAs HBT's with f_T over 150 GHz and $f_{\max}$ over 250 GHz", in *Tech. Dig. IEDM*, San Francisco, 1998, pp. 653–656.

[36] W. J. Ho, N. L. Wang, M. F. Chang, A. Sailer, and J. A. Higgins, "Self-aligned, emitter-edge-passivated AlGaAs/GaAs heterojunction bipolar transistors with extrapolated maximum oscillation frequency of 350 GHz", *IEEE Trans. Electron Dev.*, vol. 39, p. 2655, 1992.

[37] W. Hafez, J.-W. Lai, and M. Feng, "Submicron InP-InGaAs single heterojunction bipolar transistors with f_T of 377 GHz", *IEEE Electron Dev. Lett.*, vol. 24, pp. 292–294, 2003.

[38] D. Yu, K. Lee, B. Kim, D. Ontiveros, K. Vargason, J. M. Kuo, and Y. C. Kao, "Ultra high-speed InP-InGaAs SHBTs with $f_{\max}$ of 478 GHz", *IEEE Electron Dev. Lett.*, vol. 24, pp. 384–386, 2003.

[39] Y. Betser, D. Scott, D. Mensa, S. Jaganathan, T. Mathew, and M. J. Rodwell, "InAlAs/InGaAs HBTs with simultaneously high values of f_T and $f_{\max}$ for mixed analog/digital applications", *IEEE Electron Dev. Lett.*, vol. 22, pp. 56–58, 2001.

[40] K. H. G. Duh, "A super low-noise 0.1 μm T-gate InAlAs-InGaAs-InP HEMT", *IEEE Microw. Guid. Wave Lett.*, vol. 1, pp. 114–116, 1991.

[41] J. H. Lee, H.-S. Yoon, B.-S. Park, S.-J. Maeng, C.-W. Lee, H.-T. Choi, C.-E Yun, and C.-S. Park, "Noise performance of pseudomorphic AlGaAs/InGaAs/GaAs high electron mobility transistors with wide head T-shaped gate recessed by electron cyclotron resonance plasma etching", *Jpn. J. Appl. Phys.*, vol. 38, pp. 654–657, 1999.

[42] B. Bayraktaroglu, J. Barette, L. Kehias, C. I. Huang, R. Fitch, R. Neidhard, and R. Scherer, "Very high-power-density CW operation of GaAs/AlGaAs microwave heterojunction bipolar transistors", *IEEE Electron Dev. Lett.*, vol. 14, pp. 493–495, 1993.

[43] J. R. Shealy, V. Kaper, V. Tilak, T. Prunty, J. A. Smart, B. Green, and L. F. Eastman, "An AlGaN/GaN high-electron-mobility transistor with an AlN sub-buffer layer", *J. Phys.: Condens. Matter*, vol. 14, pp. 3499–3509, 2002.

Juin J. Liou received the B.Sc. (honors), M.Sc., and Ph.D. degrees in electrical engineering from the University of Florida, Gainesville, in 1982, 1983, and 1987, respectively. In 1987, he joined the Department of Electrical and Computer Engineering at the University of Central Florida, Orlando, Florida, where he is now a Professor. Dr. Liou has published six textbooks (another in progress), more than 180 journal papers (including 11 invited articles), and more than 130 papers (including 35 keynote or invited papers) in international and national conference proceedings. He has been awarded more than $4.0 million of research grants, and he is an associate editor for the *Simulation Journal* and a regional editor for the *Microelectronics Reliability.* Dr. Liou received ten different awards on excellence in teaching and research from the University of Central Florida and two different awards from the IEEE Electron Device Society. Dr. Liou is an IEEE EDS Distinguished Lecturer, an IEEE EDS Administrative Committee Member, and a Senior Member of the IEEE.
e-mail: liou@pegasus.cc.ucf.edu
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32186, USA
Department of Electronics Science and Technology
Huazhong University of Science and Technology
Wuhan, P. R. China

Frank Schwierz received his Ph.D. degree in electrical engineering from the Technical University Ilmenau, Ilmenau, Germany, in 1983. He is currently Head of the RF and Nanoelectronic Device Group at Technical University Ilmenau, Germany. Dr. Schwierz's research interests include semiconductor device physics, simulation, and modeling. Dr. Schwierz has published 1 book and more than 60 journal and conference papers. He is a member of the Council of the Center of Micro- and Nanotechnology, Germany, and a Senior Member of the IEEE.
e-mail: frank.schwierz@tu-ilmenau.de
Fachgebiet Festkörperelektronik
Technische Universität Ilmenau
P.O. Box 100565, 98684 Ilmenau, Germany

Regular paper

Photonic crystal fibre characterisation with the method of lines

Igor A. Goncharenko and Marian Marciniak

Abstract — Photonic crystal fibres are longitudinally uniform fibres in which in lateral directions periodic refractive index changes occur. Two basically different light guiding mechanisms occur in crystal fibres: index guiding and bandgap guiding. In the paper different modelling methods have been evaluated when applied to photonic crystal fibres. In particular, the method of lines has been shown to be effective and reliable for both classes of photonic crystal fibres. High accuracy results for optical field distribution and dispersion characteristics in a photonic crystal fibre have been achieved with the method of lines.

Keywords* — *optical crystal fibres, photonic bandgap, method of lines.

1. Introduction

Optical fibres are actually regarded as the best available transmission medium in telecommunications. In addition, a growing number of very successful attempts to exploit fibres for optical signal processing purposes has to be noted recently. In silica glass fibres used actually in telecommunications the guiding of the light is assured by refractive index increase within the core region in comparison with surrounding cladding of the fibre. This is referred to in the literature as the "index guiding" principle and it results in specific transmission characteristics of the fibres as dispersion, single-mode vs. multimode transmission depending on the wavelength of the light, and several other.

Recently a new class of fibres based in "bandgap guiding" principle receive a lot of attention. Those are photonic crystal fibres, in which a new operational principle of optical fibres is possible, namely guidance due to photonic bandgap (PBG) effect [1–4]. Photonic bandgaps are forbidden photon energy intervals, which may be displayed by periodic dielectric structures (photonic crystals), and correspond to the electronic bandgaps of semiconductor crystals. Such PBGs may exist in periodic silica/air structures. One of the most promising application areas where photonic crystals are finding use is in optical fibre technology. These fibres of a new type are often called photonic crystal fibres (PCFs). PCFs are single material optical fibres with periodic array of air holes running along their entire length. Typical PCF cross section is shown in Fig. 1. PCFs have properties that can differ substantially from conventional step-index fibers, such as unusual dispersion characteristics (the "endlessly single mode" guidance could be achieved in a fibre with a periodic air-silica cladding), low or high effective nonlinearities, and many others.

In fibre like the one shown in Fig. 1, light can be guided using either one of two quite different mechanisms, so they can be divided into two very different groups [1–4]. The first is fibres having a high-index core (typically solid silica) surrounded by a two-dimensional photonic-crystal-

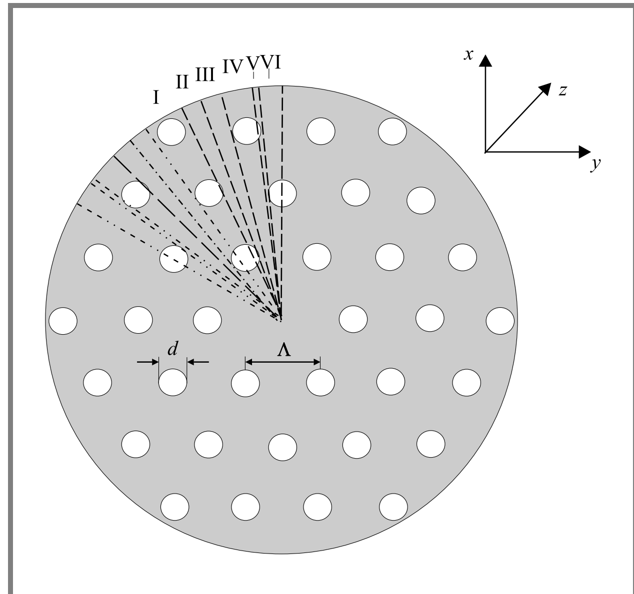


Fig. 1. Photonic crystal fibre cross-section; air holes are arranged in a hexagonal lattice in the cladding region.

cladding structure. These fibres have properties, which partially resembles those of conventional fibres due to the fact that the guidance is caused by total internal reflection (TIR). The higher refractive index of the core compared to the effective index of the photonic crystal cladding allows for traditional index guiding (effective index guidance). These fibres (TIR-PCFs) do not in fact rely on PBG effect at all.

Radically different to TIR-PCFs are fibres, where photonic-crystal-cladding structure is exhibiting PBG effect, and where this effect is utilised to confine light in core region. A full 2D photonic band gap can be created when the holes are arranged in a hexagonal lattice as in Fig. 1 or in honeycomb configuration, i.e., there exists a frequency range in which light cannot propagate in the transverse plane. Numerical simulations show that a band gap only forms when the air holes are quite large, that is when the air hole diameter d is at least 40% of the hole separation Λ. When a defect is introduced into such a structure (for instance, the absence of the hole in the centre), a localised state is created within the bandgap, and so it becomes possible for

PCF to guide light along the length of the fibre making use this state. These fibres, PBG-PCFs, show remarkable properties among which being the ability to confine and guide light along a core region having refractive index below that of the cladding structure. It was recently proven that it is possible to guide light almost entirely within an air-core using PBG fibres. The transmission losses in such fibres (in case they have the ideal structure) are in 10–100 times lower than the one in conventional fibres.

2. Review of modelling methods

The appearance of the new class of optical waveguides represented by photonic crystal fibres not only have opened up for new waveguiding properties, but it has also placed new and stronger demands on fibre modelling.

The effective-index approach should primarily be seen as a rapid method for gaining a qualitative impression of the waveguiding properties of TIR-PCFs [5–8]. In this model, the waveguide consists of a core and cladding region that have refractive indices n_{co} and n_{cl}. The core is pure silica, but the definition of the refractive index in microstructured cladding region is defined in terms of the propagation constant of the lowest order mode that could propagate in the infinite cladding material. This cladding mode field, ψ, is determined by solving the scalar-wave equation within a unit cell centred on one of the holes. By reflection symmetry the boundary condition at the cell edge (at radius $\Lambda/2$) is $\mathrm{d}\psi/\mathrm{d}s = 0$, where s is the coordinate normal to the edge. The propagation constant of the resulting fundamental space-filling mode, β_{FSM}, is used to define the effective index of the cladding as $n_{ef} = \beta_{FSM}/k$ (where k is the free-space propagation constant of light with wavelength λ). Now having determined the cladding- and core-index values, the approximate propagation properties of the PCF may be calculated as for step-index fibre with core index n_{co}, core radius $\Lambda/2$, and cladding index $n_{cl} = n_{ef}$. The cladding-index model can be modified by the addition of a wavelength dependent refractive index for silica through the Sellmeier formula. The dispersion properties of the PCFs are calculated by numerical differentiation.

However the effective-index approach ignores the complex refractive index profiles within PCFs. Hence, this reduced model is unable to accurately predict modal properties, which depend critically on the arrangement and size of the air holes.

To accurately model PCFs with large air holes, it is crucial to use a full vector model. In full vector technique based on plane-wave expansion method, the modal fields and refractive index profile are decomposed into plane wave components, and by doing this the wave equation is reduced to an eigenvalue problem [4, 8, 9]. This equation is then solved to find the modes and their corresponding propagation constants. As this approach can account for any kind of complicated cladding structure, it can accurately model PCF. However it is not efficient, as it does not take advantage of the localisation of the guided modes, and so many terms are needed to obtain an accurate description. Also, this technique involves defining the refractive index profile over a restricted region and using periodic boundary conditions to extend the structure over all space in the transverse plane. Hence, an additional periodicity is imposed on the system (i.e., a periodic distribution of both defects and air holes), which therefore somewhat restricts its applicability to TIR-PCFs, which, unlike PBG-PCFs, do not need to be periodic.

To overcome these disadvantages some alternative approaches have been proposed. One of them is based on modal decomposition using Hermite-Gaussian functions (localised function method) [10–13]. This technique takes advantage of mode localisation, and so can be more efficient than the plane-wave method. However, it cannot be accurate unless the refractive index is also represented well. In multipole method modal expansion on cylindrical functions is used [14–17]. However the localisation function method and multipole method cannot efficiently describe an extended hexagonal lattice structure. Biorthogonal modal method is based on the non-self-adjoin character of the electromagnetic propagation in a fibre [18, 19]. This technique is somewhat complicated as well as numerical method based upon the calculation of the multiple scattering between all the air holes that form the cladding of the PCF [20].

A 3D full-vectorial beam propagation method is successfully applied to investigate longitudinally varying structures or propagation and polarisation effects, which are of practical interest for advanced optical applications [21, 22].

Another method based on space-domain-division-type technique, such as finite element method (FEM) with locally variable mesh is also useful for design and modelling PCFs [23–26]. Contrarily to others computational methods, the FEM does not need any approximation. It models the propagation characteristics of the field taking into account the actual structure of the guide. In order to obtain a precise description of the field distribution over PCF cross-section, and especially near the holes, the classical Maxwell differential equations system must be solved for a large set of properly chosen elementary subspaces, taking into account the conditions of continuity of the fields. The first step consists in splitting the cross section of the modelled guide into distinct homogeneous subspaces. This parcelling results in a mesh of simple finite elements (triangles and quadrilaterals in the 2 dimensions case). For a better description of the fields, the shorter distance of the subspaces to the centre is, the smaller their dimensions are chosen. Then, the Maxwell equations system is discretised for each element, leading to a set of elementary matrices. The combination of the latter creates a global matrix system for the whole studied structure. Finally, the effective index, the distributions of the amplitudes and of the polarisations of the modes are numerically computed, taking into the conditions of continuity at the boundary of each subspace. However finite element method is strictly numerical and very time consumable. In calculations by the pure numerical methods casual solutions can appear besides

the right ones. Thus, the additional efforts have to be spent for definition of the correct solution between the number of solutions obtained.

3. Method of lines

To reduce the numerical efforts another technique, co-called method of lines, can be applied for calculation of parameters and field distribution of the guided modes in PCFs. Method of lines (MoL) introduced first by Reinhold Pregla [27] is a special finite difference method, which make it possible to analyse the wave propagation in multilayer waveguide structures. This method uses the semi-analytical approach, which yields accurate results with less computational effect compared to other techniques. In the MoL the mode coupling is automatically taken into account. Non-physical or spurious modes do not appear, and the method has no problem with relative convergence phenomena [28, 29]. Disadvantage of the MoL is in reduced flexibility: different geometries require new algorithms.

Following the MoL we divide the cross-section of PCF into a sequence of regions (layers) homogeneous in angular direction for any fixed radial coordinate (see Fig. 1).Then we look out for a solution of system of coupled partial differential equations in each layer:

$$-j\varepsilon[\mathbf{E}] = \bar{r}^{-1}\frac{\partial}{\partial\phi}[\mathbf{H}] + \begin{bmatrix} D_z \\ -\bar{r}^{-1}D_r\bar{r} \end{bmatrix}\widetilde{H}_\phi , \tag{1}$$

$$-j\mu[\mathbf{H}] = \bar{r}^{-1}\frac{\partial}{\partial\phi}[\mathbf{E}] - \begin{bmatrix} \bar{r}^{-1}D_r\bar{r} \\ D_z \end{bmatrix}E_\phi , \tag{2}$$

where $\mathbf{H} = \left[-\widetilde{H}_z^t, \widetilde{H}_r^t\right]^t$, $\mathbf{E} = \left[E_r^t, E_z^t\right]^t$, superscripts t stand for matrix transposition, $\widetilde{H}_{r,z,\phi} = \eta_0 H_{r,z,\phi}$,$\bar{r} = k_0\, r$, k_0 and η_0 are the wave number and wave impedance of free space, $E_{r,z}$ and $H_{r,z}$ are the components of electric and magnetic fields respectively, ε and μ are dielectric and magnetic permittivities, $D_z = jn_{ef}$, $D_r = \partial/\partial r$, n_{ef} is the effective refractive index of the fibre mode. These equations can be easily derived from Maxwell's equations. The azimuthal fields components are obtained from

$$\begin{aligned} \widetilde{H}_\phi &= j\mu^{-1}\left[D_z, -D_r\right][\mathbf{E}] , \\ E_\phi &= -j\varepsilon^{-1}\left[D_r, D_z\right][\mathbf{H}] . \end{aligned} \tag{3}$$

After some mathematical manipulation system (1)–(2) can be rewritten in matrix notation as

$$\frac{\partial^2}{\partial\phi^2}[\mathbf{H}] + \left[Q_H^\phi\right][\mathbf{H}] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \tag{4}$$

$$\frac{\partial^2}{\partial\phi^2}[\mathbf{E}] + \left[Q_E^\phi\right][\mathbf{E}] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \tag{5}$$

with

$$\begin{aligned} Q_{E11}^\phi &= D_r\varepsilon^{-1}\bar{r}D_r\varepsilon\bar{r} + \mu\bar{r}D_z\mu^{-1}\bar{r}D_z + \varepsilon\mu\bar{r}^2 , \\ Q_{E12}^\phi &= D_r\varepsilon^{-1}\bar{r}D_z\varepsilon\bar{r} - \mu\bar{r}D_z\mu^{-1}\bar{r}D_r , \\ Q_{E21}^\phi &= D_z\varepsilon^{-1}\bar{r}D_r\varepsilon\bar{r} - \mu\bar{r}D_r\mu^{-1}\bar{r}D_z , \\ Q_{E22}^\phi &= D_z\varepsilon^{-1}\bar{r}D_z\varepsilon\bar{r} + \mu\bar{r}D_r\mu^{-1}\bar{r}D_r + \varepsilon\mu\bar{r}^2 \end{aligned} \tag{6}$$

and

$$\begin{aligned} Q_{H11}^\phi &= \varepsilon\bar{r}D_r\varepsilon^{-1}\bar{r}D_r + D_z\mu^{-1}\bar{r}D_z\mu\bar{r} + \varepsilon\mu\bar{r}^2 , \\ Q_{H12}^\phi &= \varepsilon\bar{r}D_r\varepsilon^{-1}\bar{r}D_z - D_z\mu^{-1}\bar{r}D_r\mu\bar{r} , \\ Q_{H21}^\phi &= \varepsilon\bar{r}D_z\varepsilon^{-1}\bar{r}D_r - D_r\mu^{-1}\bar{r}D_z\mu\bar{r} , \\ Q_{H22}^\phi &= \varepsilon\bar{r}D_z\varepsilon^{-1}\bar{r}D_z + D_r\mu^{-1}\bar{r}D_r\mu\bar{r} + \varepsilon\mu\bar{r}^2 . \end{aligned} \tag{7}$$

These partial differential equations have to be discretised with respect to the radial coordinate using finite differences. This results in a system of ordinary differential equations, which can be solved analytically. In this case, all potentials and dielectric permittivities, as well as, the radial coordinate r have to be discretised. To fulfil the interface conditions the discretisation is done on two different line systems.

In transform domain the discretised wave equations are

$$\frac{d}{d\phi^2}\overline{\mathbf{F}} - \Gamma^2\overline{\mathbf{F}} = \mathbf{0} , \tag{8}$$

where $\mathbf{F}$ is $\mathbf{H}$ or $\mathbf{E}$, $\mathbf{H} = \mathbf{T}_H\overline{\mathbf{H}}$, $\mathbf{E} = \mathbf{T}_E\overline{\mathbf{E}}$, Γ is the vector of eigenvalues and derived from eigenvalue problem

$$\mathbf{T}_{H,E}^{-1}Q_{H,E}^\phi\mathbf{T}_{H,E} = -\Gamma^2 . \tag{9}$$

General solution of the Eqs. (8) is

$$\overline{\mathbf{F}} = A\cosh\left(\Gamma\phi\right) + B\sinh\left(\Gamma\phi\right) . \tag{10}$$

Coefficients A and B can be obtained from the relation between magnetic and electric fields.

The analytical solution should be performed in ϕ-direction. The electric and magnetic fields components should be matched at the boundaries of the layers. Thus we can transform fields components from inner side of the layer VI (see Fig. 1) to its outer side. By such a way determinant equation for defining effective refractive index n_{ef} is derived. The computation efforts can be reduced if we take into account the azimuthal periodicity of the structure and will use the Floquet's theorem.

4. Modelling and results

In the following we present preliminary results as the proof of applicability of MoL for calculation of PCFs structures. Figure 2 shows the dispersion characteristics of fundamental mode of PCF (curve *1*) with $\Lambda = 2.3$ μm, $d = 0.6$ μm, refractive index $n = 1.46$. Curve *2* presents dispersion characteristics of conventional fibre with cladding refractive index n_{cl}, where $n_{cl} = (nS_s + S_h)/(S_s + S_h)$, S_s and S_h are the cross-section squares of silica and air holes respectively.

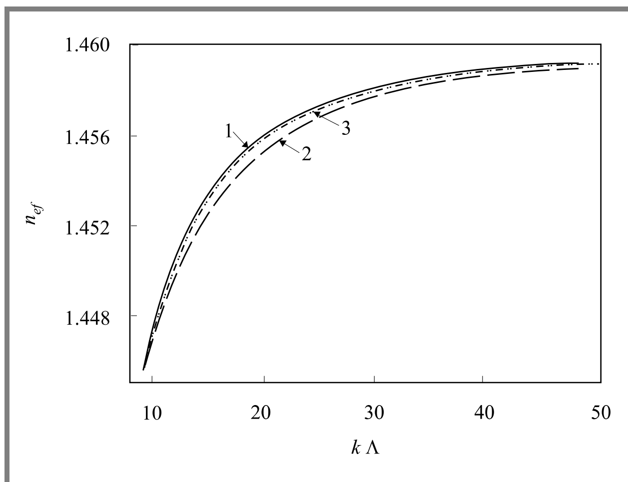


Fig. 2. Dispersion characteristics of fundamental modes of PCF (curve *1*) and conventional fibre (curve *2*), calculated by MoL. Curve *3* presents the dispersion characteristic of the same PCF obtained by full-vector method in [19].

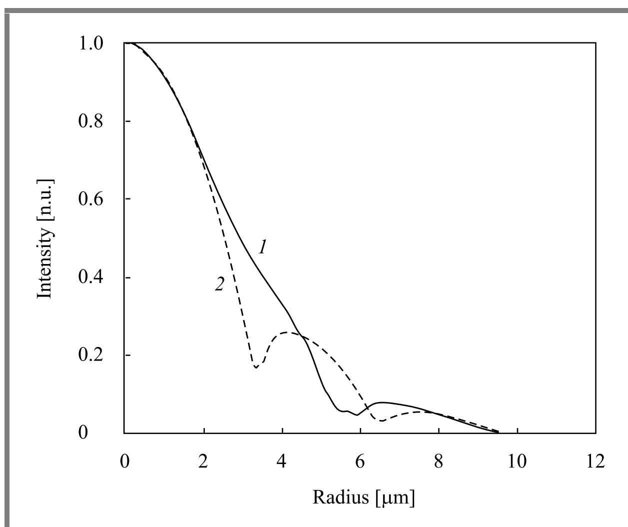


Fig. 3. Radial distribution of the transverse electric fields of fundamental mode of PCF in x (curve *1*) and y (curve *2*) directions.

Curve *3* presents the dispersion characteristic of the same PCF calculated by full-vector method in [19]. As can be seen from the figure, the results obtained by MoL are in very good agreement with those obtained by full-vector analysis. Moreover, the comparison carried out in [21] shows the good agreement between the dispersion values from [19] and the ones obtained by full-vector BMP method in [21]. Figure 3 shows the radial distribution of transverse electric field of fundamental mode in x- and y-directions.

5. Conclusions

Photonic crystal fibres have been classified with respect to basics of guiding mechanism, index guiding or bandgap guiding. Different modelling methods have been evaluated when applied to photonic crystal fibres with special emphasis to the method of lines. Mathematical background of the method of lines has been discussed, and the method has been successfully implemented for calculation of the parameters of modal fiend distribution and dispersion characteristics in photonic crystal fibres. It has been shown that results obtained by the method of lines are in a good agreement with the ones obtained by other full-vector numerical methods.

Acknowledgement

The authors acknowledge the highest influence of the personal contacts with and of the pioneering work of Professor Dr.-Ing. Reinhold Pregla from Allgemeine und Theoretische Elektrotechnik, FernUniversität, Hagen, Germany, the inventor of the method of lines for applications in the microwaves and photonics.

This work has been developed at the National Institute of Telecommunications as the scientific output of a joint study on “Telecom applications of photonic periodic structures” carried out within the framework of the European COST 268 Project *Wavelength Scale Photonic Components for Telecommunications Project*.

References

[1] J. C. Knight, J. Broeng, T. A. Birks, and P. St. J. Russell, “Photonic bandgap guidance in optical fibers”, *Science*, vol. 282, pp. 1476–1478, 1998.

[2] J. Broeng, D. Mogilevstev, S. E. Barkou, and A. Bjarklev, “Photonic crystal fibers: a new class of optical waveguides”, *Opt. Fib. Technol.*, vol. 5, no. 3, pp. 305–330, 1999.

[3] J. Broeng, T. Sondergaard, S. E. Barkou, P. M. Barbeito, and A. Bjarklev, “Waveguidance by the photonic bandgap effect in optical fibres”, *J. Opt. A: Pure Appl. Opt.*, vol. 1, no. 4, pp. 477–482, 1999.

[4] T. M. Monro, D. J. Richardson, N. G. R. Broderick, and P. J. Bennett, “Holey optical fibers: an efficient modal model”, *J. Lightw. Technol.*, vol. 17, no. 6, pp. 1093–1102, 1999.

[5] J. C. Knight, T. A. Birks, P. St. J. Russell, and J. P. de Sandro, “Properties of photonic crystal fiber and the effective index model”, *J. Opt. Soc. Am. A-Opt. Image Sci.*, vol. 15, no. 3, pp. 748–752, 1998.

[6] A. Bjarklev, J. Broeng, K. Dridi, and S. E. Barkou, “Dispersion properties of photonic crystal fibres”, in *24th Eur. Conf. Opt. Commun. ECOC’98*, Madrid, Spain, 1998, vol. 1, pp. 135–136.

[7] T. A. Birks, J. C. Knight, and P. St. J. Russell, “Endlessly single-mode photonic crystal fiber”, *Opt. Lett.*, vol. 22, pp. 961–963, 1997.

[8] A. Bjarklev, J. Broeng, and S. E. Barkou, “Modelling of photonic crystal fibres”, in *25th Eur. Conf. Opt. Commun. ECOC’99*, Nice, France, 1999, vol. 1, pp. 16–19.

[9] M. J. Steel and R. M. Jr. Osgood, “Polarization and dispersive properties of elliptical-hole photonic crystal fibers”, *J. Lightw. Technol.*, vol. 19, no. 4, pp. 495–503, 2001.

[10] D. Mogilevtsev, T. A. Birks, and P. St. J. Russell, “Group-velocity dispersion in photonic crystal fibers”, *Opt. Lett.*, vol. 23, no. 21, pp. 1662–1664, 1998.

[11] D. Mogilevtsev, T. A. Birks, and P. St. J. Russell, “Localized function method for modeling defect modes in 2D photonic crystals”, *J. Lightw. Technol.*, vol. 17, pp. 2078–2081, 1999.

[12] T. M. Monro, D. J. Richardson, and P. J. Bennett, “Developing holey fibres for evanescent field devices”, *Electron. Lett.*, vol. 35, no. 14, pp. 1188–1189, 1999.

[13] T. M. Monro, D. J. Richardson, N. G. R. Broderick, and P. J. Bennett, “Modeling large air fraction holey optical fibers”, *J. Lightw. Technol.*, vol. 18, no. 1, pp. 50–56, 2000.

[14] T. P. White, R. C. McPhedran, C. M. de Sterke, and L. C. Botten, "Multipole method for efficient microstructured optical fiber calculations. Technical Digest", in *Conf. Lasers Electro-Opt. CLEO'2001*, Baltimore, USA, 2001, pp. 597–598.

[15] M. A. van Eijkelenborg, M. C. J. Large, A. Argyros, J. Zagari, S. Manos, N. A. Issa, I. Bassett, S. Fleming, R. C. McPhedran, C. M. De Sterke, and N. A. P. Nicorovici, "Microstructured polymer optical fibre", *Opt. Expr.*, vol. 9, no. 7, pp. 319–327, 2001.

[16] M. J. Steel, T. P. White, C. M. de Sterke, R. C. McPhedran, and L. C. Botten, "Symmetry and degeneracy in microstructured optical fibers", *Opt. Lett.*, vol. 26, pp. 488–490, 2001.

[17] T. P. White, R. C. McPhedran, L. C. Botten, G. H. Smith, and C. M. de Sterke, "Calculations of air-guided modes in photonic crystal fibers using the multipole method", *Opt. Expr.*, vol. 9, no. 13, pp. 721–732, 2001.

[18] E. Silvestre, M. V. Andres, and P. Andres, "Biorthonormal-basis method for vector description of optical fiber modes", *J. Lightw. Technol.*, vol. 16, pp. 923–928, 1998.

[19] A. Ferrando, E. Silvestre, J. J. Miret, P. Andres, and M. V. Andres, "Full-vector analysis of a realistic photonic crystal fiber", *Opt. Lett.*, vol. 24, no. 5, pp. 276–278, 1999.

[20] P. J. Roberts and T. J. Shepherd, "The guidance properties of multicore photonic crystal fibres", *J. Opt. A: Pure Appl. Opt.*, vol. 3, no. 6, pp. S133–140, 2001.

[21] F. Fogli, L. Saccomandi, P. Bassi, G. Bellanca, and S. Trillo, "Full vectorial BPM modeling of index-guiding photonic crystal fibers and couplers", *Opt. Expr.*, vol. 10, no. 1, pp. 54–59, 2002.

[22] B. J. Eggleton, C. Kerbage, P. S. Westbrook, R. S. Windeler, and A. Hale, "Microstructured optical fiber devices", *Opt. Expr.*, vol. 9, no. 13, pp. 698–713, 2001.

[23] F. Brechet, J. Marcou, and D. Pagnoux, "Accurate computation of the chromatic dispersion in unimodal photonic crystal fibres", in *25th Eur. Conf. Opt. Commun. ECOC'99*, Nice, France, 1999, vol. 1, pp. 26–27.

[24] F. Brechet, J. Marcou, and D. Pagnoux, "Analysis of photonic crystal fibres modes by the finite element method", in *25th Eur. Conf. Opt. Commun. ECOC'99*, Nice, France, 1999, vol. 1, pp. 282–283.

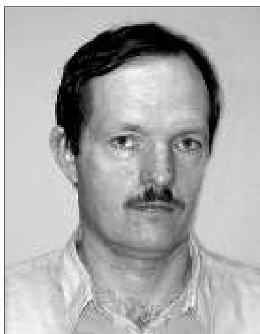
[25] R. Ghosh, A. Kumar, J.-P. Meunier, and E. Marin, "Modal characteristics of few-mode silica-based photonic crystal fibres", *Opt. Quant. Electron.*, vol. 32, no. 6–8, pp. 963–970, 2000.

[26] M. Koshiba and K. Saitoh, "Numerical verification of degeneracy in hexagonal photonic crystal fibers", *IEEE Photon. Technol. Lett.*, vol. 13, no. 12, pp. 1313–1315, 2001.

[27] R. Pregla, "About the nature of the method of lines", *Int. J. Electron. Commun. (AEÜ)*, vol. 41, pp. 369–370, 1987.

[28] R. Pregla, "The method of lines as generalized transmission line technique for the analysis of multilayered structures", *Int. J. Electron. Commun. (AEÜ)*, vol. 50, no. 3, pp. 293–300, 1996.

[29] R. Pregla and W. Pascher, "The method of lines", in *Numerical Techniques for Microwave and Millimeter Wave Passive Structures*. T. Itoh, Ed. New York: Wiley, 1989, pp. 381–446.



Igor A. Goncharenko was born in Minsk, Belarus. He graduated from the Physics Department of the Byelorussian State University in 1981. He received the Ph.D. degree in physics and mathematics from the USSR Academy of Sciences (Moscow) in 1985 and Dr.Sci. degree from the National Academy of Sciences of Belarus (Minsk) in 2001. Since 1985, he works in Institute of Electronics of the National Academy of Sciences of Belarus in Minsk. Since 2002 he has the position of Principal Research Fellow. In 1994 he was awarded a Royal Society Fellowship and worked with the Optoelectronics Research Centre, University of Southampton, UK. In 1996 he was awarded the Alexander von Humboldt Research Fellowship and carried out research at the FernUniversitaet in Hagen, Germany. His fields of investigation include theory of complicated optical fiber and waveguide structures, linear and nonlinear (soliton) pulse propagation in optical structures, optical information processing. Dr. Goncharenko is a member of the IEEE.

e-mail: lomoi@inel.bas-net.by
Institute of Electronics
National Academy of Sciences of Belarus
22 Logoisky Trakt
Minsk 220090, Belarus

Marian Marciniak Associate Professor has been graduated in solid state physics from Marie-Curie Sklodovska University in Lublin, Poland, in 1977. From 1985 to 1989 he performed Ph.D. studies in electromagnetic wave theory at the Institute of Fundamental Technological Research, Polish Academy of Sciences, followed by Ph.D. degree (with distinction) in optoelectronics received from Military University of Technology in Warsaw. In 1997 he received his Doctor of Sciences (habilitation) degree in physics/optics from Warsaw University of Technology. From 1978 to 1997 he held an academic position in the Military Academy of Telecommunications in Zegrze, Poland. In 1996 he joined the National Institute of Telecommunications in Warsaw where he actually leads the Department of Transmission and Fibre Technology. Previous activities have included extended studies of optical waveguiding linear and nonlinear phenomena with analytic and numerical methods including beam-propagation methods. Actual research interests include photonic crystal technology and phenomena, optical packet-switched networks, and the future global optical and wireless network. Recently he has introduced and developed a concept of a hybrid real-time service end photonic packet network. He is an author or co-author of over 190 technical publications, including a number of conference invited presentations and 13 books authored, co-authored and/or edited by himself. He is a Senior Member of the IEEE – Lasers & Electro-Optics, Communications, and Computer Societies, a member of The New York Academy of Sciences, The Optical Society of America, SPIE – The International Society for Optical Engineering and its Technical Group on Optical Networks, and of the American Association for

the Advancement of Science. In early 2001 he originated the IEEE/LEOS Poland Chapter and he has served as the Chairman of that Chapter until July 2003. He is widely involved in the European research for optical telecommunication networks, systems and devices. He was the originator of accession of Poland to European Research Programs in the optical telecommunications domain, in chronological order: COST 240 *Modelling and Measuring of Advanced Photonic Telecommunication Components*, COST P2 *Applications of Nonlinear Optical Phenomena*, COST 266 *Advanced Infrastructure for Photonic Networks*, COST 268 *Wavelength-Scale Photonic Components for Telecommunications*, COST 270 *Reliability of Optical Components and Devices in Communications Systems and Networks*, COST 273 *Towards Mobile Broadband Multimedia Networks*, and very recently two new starting actions COST 288 *Nanoscale and Ultrafast Photonics* and COST P11 *Physics of Linear, Nonlinear and Active Photonic Crystals*. In all but two those projects he acted as one of the originators at the European level. He has been appointed to Management Committees of all those Projects as the Delegate of Poland. In addition, he has been appointed as the Evaluator of the European Union's 5th Framework Program proposals in the Action Line *All-Optical and Terabit Networks*. He is a Delegate to the International Telecommunication Union, Study Group 15: *Optical and Other Transport Networks*, and to the International Electrotechnical Commission, Technical Committee 86 *Fibre Optics* and its two sub-Committees. He served as a member of Polish Delegation to the World Telecommunication Standards Assembly WTSA 2000. From 2002 he participates in the work of the URSI – *International Union of Radio Science, Commission D – Electronics and Photonics*. In 2000 he originated and actually serves as the Chairman of the Technical Committee 282 on *Fibre Optics* of the National Committee for Standardisation. Since May 2003 he serves as the Vice-President of the Delegation of Poland to the Intergovernmental Ukrainian-Polish Working Group for Co-operation in Telecommunications. He is the originator and the main organiser of the *International Conference on Transparent Optical Networks* ICTON starting in 1999, and a co-located events the *European Symposium on Photonic Crystals* ESPC and *Workshop on All-Optical Routing* WAOR since 2002. He is the Technical Program Committee Co-Chair of the *International Conference on Advanced Optoelectronics and Lasers* CAOL, and he participates in Program Committees of the *Conference on the Optical Internet & Australian Conference on Optical Fibre Technology* COIN/ACOFT, the *International Conference on Mathematical Methods in Electromagnetic Theory* MMET, the *International Workshop on Laser and Fiber-Optical Network Modeling* LFNM, and the *International School for Young Scientists and Students on Optics, Laser Physics and Biophysics/Workshop on Laser Physics and Photonics*. He serves as a reviewer for several international scientific journals, and he is a Member of the Editorial Board of *Microwave & Optoelectronics Technology Letters* journal, Wiley, USA, and the *Journal of Telecommunications and Information Technology*, National Institute of Telecommunications, Poland. Languages spoken: Polish (native), English, French, and Russian. His biography has been cited in Marquis *Who's Who in the World, Who's Who in Science and Engineering*, and in the *International Directory of Distinguished Leadership of the American Biographical Institute*.

e-mail: M.Marciniak@itl.waw.pl
e-mail: marian.marciniak@ieee.org
Department of Transmission and Fibre Technology
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Regular paper

Sensitivity of microwave radiometers with square – law and linear detectors

Bronisław Stec, Andrzej Dobrowolski, and Waldemar Susek

Abstract — Stochastic analysis of modulation microwave radiometers with square – law and linear detectors is presented in the paper. Assuming ideal detector characteristics it is shown that in typical applications, i.e., in very low power measurements, a type of detector used is of no influence on total radiometer sensitivity. Other aspects of use of a particular detector are also presented.

Keywords — microwave thermograph, radiometer, thermal radiation, sensitivity.

1. Introduction

A modulation radiometer [3], presented in Fig. 1, was used for analysis. A high frequency (HF) amplifier can be described by the statistically equivalent pass band B_{HF} defined as [1, 4, 5]:

$$B_{HF} \overset{df.}{=} \frac{1}{2\pi} \frac{\left[\int\limits_{-\infty}^{\infty} G(\omega)\mathrm{d}\omega\right]^2}{2\int\limits_{-\infty}^{\infty} G^2(\omega)\mathrm{d}\omega}, \tag{1}$$

where $G(\omega)$ is the double-sided characteristics of RMS power gain of a HF amplifier.
B_{HF} is a pass band of a hypothetical rectangular filter that transfers signal with the same statistical error of mean-square value as a real filter when white noise is present at its input.
Low frequency (LF) amplifier is characterised by the noise pass band B_{LF} defined as [1, 4]:

$$B_{LF} \overset{df.}{=} \frac{1}{2\pi} \frac{\int\limits_{0}^{\infty} |H(\omega)|^2 \mathrm{d}\omega}{|H(0)|^2}, \tag{2}$$

where $H(\omega)$ is the transmittance function of an LF amplifier and output filter system, defined for real frequencies.
B_{LF} is a pass band of a hypothetical rectangular filter that transfers signal with the same mean-square value as a real filter, when white noise is present at its input. Therefore, a band defined like this can be a useful measure of a bandwidth applied in narrowband measurements of mean-square values.
A nonlinear inertialess system was used as a detector. It was characterised by a double-half square function in the form of

$$U_{det} = \beta U_{in}^2 \tag{3}$$

or by single-half linear function in the form of

$$U_{det} = \begin{cases} \gamma U_{in} & \text{for} \quad U_{in} \geqslant 0 \\ 0 & \text{for} \quad U_{in} < 0 \end{cases}. \tag{4}$$

A thermal noise of the equivalent noise temperature T is an input signal of the detector. It is assumed that this noise formed by HF amplifier is narrowband, has normal distribution and, moreover, the noise entering detector has no DC component.
Double-sided characteristics of RMS power gain of a HF amplifier is approximated by the exponential function:

$$G(\omega) = G_{ps\,\max}\left[\mathrm{e}^{-\alpha(\omega+\omega_0)^2} + \mathrm{e}^{-\alpha(\omega-\omega_0)^2}\right]. \tag{5}$$

Consequently, a double-sided spectral power density of a detector input-signal, presented in the normalised form in Fig. 2, can be described as

$$S_{in}(\omega) = \frac{N}{2} G(\omega) = \\ = \frac{N_{\max}}{2}\left[\mathrm{e}^{-\alpha(\omega+\omega_0)^2} + \mathrm{e}^{-\alpha(\omega-\omega_0)^2}\right] \quad \left[\frac{\mathrm{W}}{\mathrm{Hz}}\right], \tag{6}$$

where

$$N_{\max} = G_{ps\,\max} k T. \tag{7}$$

Using definition (1) and relation (5) we get

$$B_{HF} = \frac{1}{\sqrt{2\pi\alpha}}. \tag{8}$$

According to Wiener-Kintchine theorem, the autocorrelation function of input noise is equal to

$$R_{in}(\tau) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} S_{in}(\omega)\mathrm{e}^{\mathrm{j}\omega\tau}\mathrm{d}\omega = \frac{N_{\max}}{2\sqrt{\pi\alpha}}\mathrm{e}^{\frac{-\tau^2}{4\alpha}}\cos\omega_0\tau. \tag{9}$$

Assuming that input noise is an ergodic process and because of absence of DC component, we can estimate its variance as

$$\sigma_{in}^2 = R_{in}(0) = \frac{N_{\max}}{2\sqrt{\pi\alpha}}. \tag{10}$$

Taking dependence (8) into account we find relation between the band width B_{HF} and input noise variance

$$\sigma_{in}^2 = \frac{N_{\max} B_{HF}}{\sqrt{2}}. \tag{11}$$

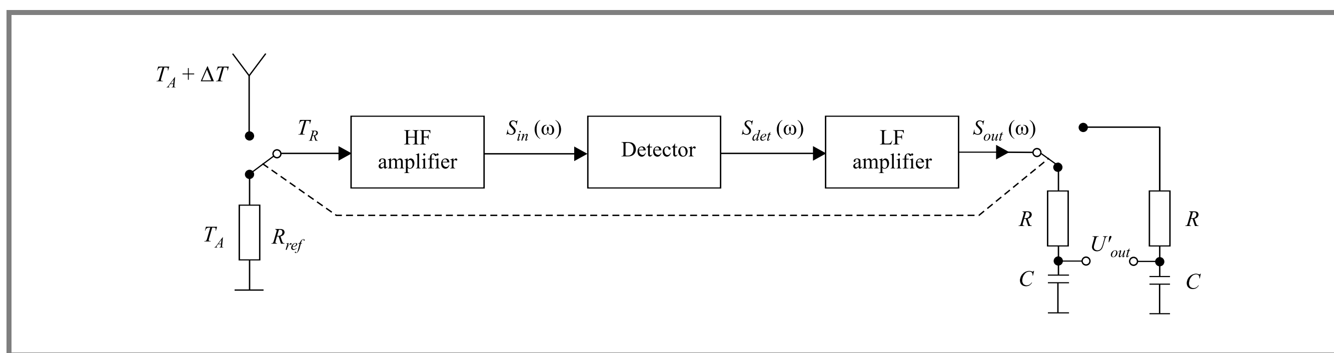


Fig. 1. Simplified block diagram of the modulation radiometer.

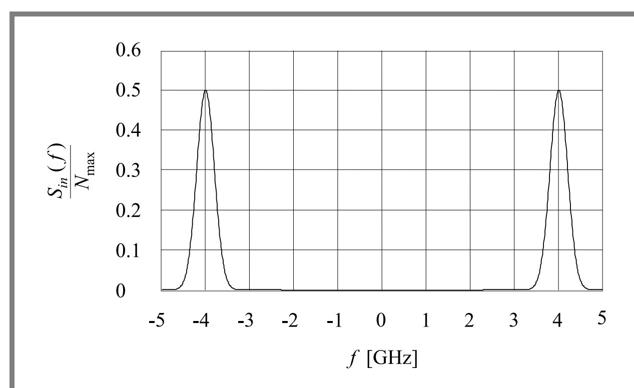


Fig. 2. Double-sided spectral power density of input noise.

Introducing the autocorrelation function envelope concept

$$\rho_{in}(\tau) = \mathrm{e}^{\frac{-\tau^2}{4\alpha}} \tag{12}$$

we finally get

$$R_{in}(\tau) = \sigma_{in}^2 \rho_{in}(\tau) \cos \omega_0 \tau\,. \tag{13}$$

2. Square – law detector

Considering transmission of stationary normal noise without component through a double-half inertialess system with square characteristics specified by relation (3) we find an expression which describes autocorrelation function of a process at the detector output [2]:

$$R_{det}(\tau) = \beta^2 \sigma_{in}^4 + 2\beta^2 R_{in}^2(\tau)\,. \tag{14}$$

Using relation (13) we have

$$R_{det}(\tau) = \beta^2 \sigma_{in}^4 \left[1 + \rho_{in}^2(\tau) + \rho_{in}^2(\tau) \cos 2\,\omega_0 \tau\right]. \tag{15}$$

If a fast-varying component that does not appear at an LF amplifier output (playing a role of a post-detector filter) is neglected, an autocorrelation function of an output process can be written as

$$R_{det}(\tau) = \beta^2 \sigma_{in}^4 \left[1 + \rho_{in}^2(\tau)\right]. \tag{16}$$

According to Wiener-Kintchine theorem a noise spectrum at the detector output is

$$S_{det}(\omega) = \int_{-\infty}^{\infty} R_{det}(\tau)\,\mathrm{e}^{-\mathrm{j}\omega\tau}\mathrm{d}\tau = $$

$$= \beta^2 \sigma_{in}^4 \left[2\pi\delta(\omega) + \sqrt{2\pi\alpha}\,\mathrm{e}^{\frac{-\alpha\omega^2}{2}}\right]. \tag{17}$$

The first component describes power of DC component. Its value at the LF amplifier output equals:

$$P_{DET} = \left|H(0)\right|^2 \frac{1}{2\pi} \int_{-\infty}^{\infty} \beta^2 \sigma_{in}^4 2\pi\delta(\omega)\,\mathrm{d}\omega = $$

$$= \frac{1}{2}\left|H(0)\right|^2 \beta^2 N_{\max}^2 B_{HF}^2\,. \tag{18}$$

The second component represents spectrum of a low-varying component of output noise shown for real frequencies in Fig. 3. In a very narrow pass band of an LF amplifier as compared to the band B_{HF} (typically $B_{LF}/B_{HF} \cong 10^{-8}$) a varying exponential factor practically equals unity and a spectral density of output noise can be considered as constant

$$S_{det\,LF}(\omega) = \beta^2 \sigma_{in}^4 \sqrt{2\pi\alpha}\,\mathrm{e}^{\frac{-\alpha\omega^2}{2}} \approx$$

$$\approx \beta^2 \sigma_{in}^4 \sqrt{2\pi\alpha} = S_{det\,0}\,. \tag{19}$$

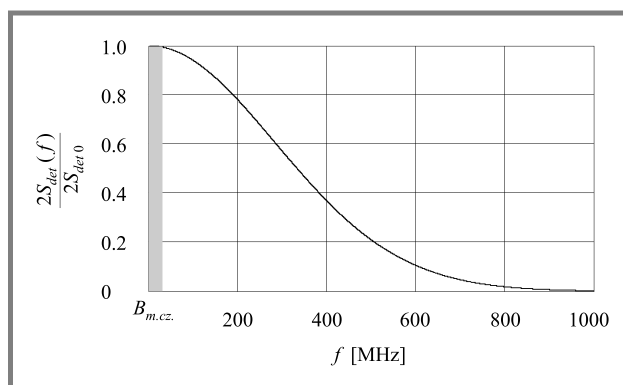


Fig. 3. Normalised one-sided noise spectral density at detector output.

Using relations (8) and (11), Eq. (19) can be expressed as

$$S_{det\,0} = \frac{1}{2}\beta^2 N_{\max}^2 B_{HF}\,. \tag{20}$$

Power of variable component of a noise at LF amplifier output is described by integral:

$$P_{det} = \frac{1}{2\pi}\int_0^\infty 2S_{det\,0}\left|H(\omega)\right|^2 \mathrm{d}\omega = \\ = 2S_{det\,0}\frac{1}{2\pi}\int_0^\infty \left|H(\omega)\right|^2 \mathrm{d}\omega\,. \tag{21}$$

Using definition (2) of a noise pass band for a low frequency channel, we get

$$P_{det} = 2S_{det\,0}B_{LF}\left|H(0)\right|^2 = \\ = \left|H(0)\right|^2\beta^2 N_{\max}^2 B_{HF}B_{LF}\,. \tag{22}$$

Including Eqs. (18) and (21), the ratio of constant and variable components can be expressed as

$$\rho_{square} = \frac{P_{DET}}{P_{det}} = \frac{B_{HF}}{2B_{LF}}. \tag{23}$$

DC voltage component at the LF amplifier output can be written down as

$$U_{OUT} = a_s\sqrt{P_{DET}} = \frac{b_s}{\sqrt{2}}N_{\max}B_{HF}\,, \\ b_s = a_s\beta\left|H(0)\right|, \tag{24}$$

where a_s is the proportionality factor depending on particular system design and b_s additionally includes also detector characteristics and LF channel amplification for DC and very low frequencies components.

An RMS value of a noise variable component at LF amplifier output – resulting from low-pass filtration – is equal to

$$U_{out} = a_s\sqrt{P_{det}} = b_s N_{\max}\sqrt{B_{HF}B_{LF}}\,. \tag{25}$$

Introducing a constant

$$c_s = b_s G_{ps\,\max}k \tag{26}$$

we get the following final expressions describing DC component and RMS value of AC component of the LF amplifier output voltage:

$$\begin{cases} U_{OUT} = \dfrac{c_s}{\sqrt{2}}T B_{HF} \\ U_{out} = c_s T\sqrt{B_{HF}B_{LF}} \end{cases}. \tag{27}$$

Equivalent noise temperature of input signal is a sum of RMS input temperature of radiometer noise, antenna noise, and measured temperature changes if antenna is connected to the system input

$$T = T_R + T_A + \Delta T = T_{sys} + \Delta T \tag{28}$$

or it equals

$$T = T_{sys} \tag{29}$$

if properly selected the reference resistance R_{ref} is connected to the input.

In a modulation receiver, constant voltage related to system noise is eliminated in an output differential circuit and as a consequence, DC output voltage depends only on antenna temperature rise and is described by relation:

$$U'_{OUT} = \frac{c_s}{\sqrt{2}}\left(T_{sys}+\Delta T\right)B_{HF} - \frac{c_s}{\sqrt{2}}T_{sys}B_{HF} = \\ = \frac{c_s}{\sqrt{2}}\Delta T\,B_{HF}\,. \tag{30}$$

On the other hand, we can neglect – as a very small – the component ΔT in expression describing fluctuation component and then the RMS value of AC component at the LF amplifier output is

$$U_{out} = c_s\left(T_{sys}+\Delta T\right)\sqrt{B_{HF}B_{LF}} \overset{\Delta T << T_{sys}}{\approx} c_s T_{sys}\sqrt{B_{HF}B_{LF}}. \tag{31}$$

At the differential output, because of summation of noncorrelated fluctuation components, the RMS value of AC component is twice as large; as result we have:

$$U'_{out} = 2\,c_s\,T_{sys}\sqrt{B_{HF}\,B_{LF}}\,. \tag{32}$$

Radiometer sensitivity is defined as the minimum input signal rise (in temperature units) $\Delta T_{\min}$ that ensures that DC component and AC component RMS value of output voltage are equal [4]:

$$U'_{OUT}\left(\Delta T_{\min}\right) = U'_{out}. \tag{33}$$

Thus, using (30) and (32) we get

$$\Delta T_{\min} = 2\,T_{sys}\sqrt{\frac{2\,B_{LF}}{B_{HF}}}\,. \tag{34}$$

3. Linear detector

Qualitatively different problem appears for single-half linear detector. Since such detector reproduces input noise envelope (envelope is described by Rayleigh distribution for narrowband Gaussian noise), and assuming that detector characteristics slope equals 45°($\gamma = 1$), as well as neglecting components close to central frequency of HF amplifier and its harmonics, we obtain the following expression for spectral density of the output noise power [2]:

$$S_{det}(\omega) = \frac{\pi\sigma_{in}^2}{2}\left[2\pi\delta(\omega) + \frac{1}{4}\sqrt{2\pi\alpha}\,\mathrm{e}^{\frac{-\alpha\omega^2}{2}}\right]. \tag{35}$$

As previously, we find expressions for powers of DC and AC components:

$$P_{DET} = \left|H(0)\right|^2\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\pi\sigma_{in}^2}{2}2\pi\delta(\omega)\mathrm{d}\omega = \\ = \frac{\pi}{2\sqrt{2}}\left|H(0)\right|^2 N_{\max}B_{HF}\,, \tag{36}$$

$$P_{det} = \frac{1}{2\pi}\int_0^\infty 2\frac{\pi\sigma_{in}^2}{8}\sqrt{2\pi\alpha}\left|H(\omega)\right|^2 \mathrm{d}\omega = \\ = \frac{1}{2}\frac{\pi}{2\sqrt{2}}\left|H(0)\right|^2 N_{\max} B_{LF}\,. \quad (37)$$

Using Eqs. (36) and (37) we express the ratio of DC and AC component powers as

$$\rho_{linear} = \frac{P_{DET}}{P_{det}} = \frac{2B_{HF}}{B_{LF}}\,. \quad (38)$$

As in the previous section, we get:

$$\begin{cases} U_{OUT} = a_l\sqrt{P_{DET}} = c_l\sqrt{TB_{HF}} \\ U_{out} = a_l\sqrt{P_{det}} = \frac{c_l}{\sqrt{2}}\sqrt{TB_{LF}} \end{cases}, \quad (39)$$

where:

$$c_l = a_l\left|H(0)\right|\sqrt{\frac{\pi}{2\sqrt{2}}G_{ps\,\max}k}\,. \quad (40)$$

At the differential output:

$$\begin{cases} U'_{OUT} = c_l\sqrt{\left(T_{sys}+\Delta T\right)B_{HF}} - c_l\sqrt{T_{sys}B_{HF}} \\ U'_{out} = 2\frac{c_l}{\sqrt{2}}\sqrt{\left(T_{sys}+\Delta T\right)B_{LF}} \approx \sqrt{2}c_l\sqrt{T_{sys}B_{LF}} \end{cases}. \quad (41)$$

Using definition (33), we obtain:

$$\sqrt{1+\frac{\Delta T_{\min}}{T_{sys}}} = 1 + \sqrt{\frac{2B_{LF}}{B_{HF}}}\,. \quad (42)$$

It is common in radiometry that the ratio $\Delta T_{\min}/T_{sys}$ is very small. Thus, we can perform power series expansion of the left side of above expression and, without any significant error, limit to the first two terms:

$$\sqrt{1+\frac{\Delta T_{\min}}{T_{sys}}} = 1 + \frac{1}{2}\frac{\Delta T_{\min}}{T_{sys}} - \frac{1}{8}\left(\frac{\Delta T_{\min}}{T_{sys}}\right)^2 + \ldots \approx \\ \approx 1 + \frac{1}{2}\frac{\Delta T_{\min}}{T_{sys}}\,. \quad (43)$$

Consequently, we have

$$\frac{1}{2}\frac{\Delta T_{\min}}{T_{sys}} \approx \sqrt{\frac{2B_{LF}}{B_{HF}}} \quad (44)$$

and, finally, the expression for sensitivity of a radiometer with a linear detector takes the following form:

$$\Delta T_{\min} \approx 2T_{sys}\sqrt{\frac{2B_{LF}}{B_{HF}}}\,. \quad (45)$$

4. Summary

Comparison of relations (23) and (38):

$$\frac{\rho_{linear}}{\rho_{square}} = \frac{\dfrac{2B_{HF}}{B_{LF}}}{\dfrac{B_{HF}}{2B_{LF}}} = 4 \quad (46)$$

shows that independently of bands ratio the output ratio of DC and AC components in linear detector is 6 dB better than for square detector. However, the sensitivities described by dependencies (34) and (45), most important in radiometry are almost the identical.

In real conditions, radiometer sensitivity does not depend on type of detector used, while it is a function of system temperature and a bandwidth ratio of pre- and post-detection channels, respectively.

Linear detector is characterised by linear dependence of output power on measured temperature while for square detector output voltage depends linearly on temperature. Since voltage measurement is related to lower measurement error than power measurement and because of good availability of square detectors such as semiconductor diodes for very low signals, single-half or double-half square detectors are commonly used in microwave amplifiers with direct amplification.

References

[1] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*. New York: Wiley Interscience, 1971.

[2] I. S. Gonorowski, *Radiomechaničeskie cepi i signaly*. Moskwa: Radio i svjaz', 1986 (in Russian).

[3] B. Stec, "Microwave thermography: methods and equipment", in *Proc. 11th Int. Conf. Microw., Radar Wirel. Commun. MIKON'96, Worksh.*, Warsaw, Poland, 1996, pp. 97–114.

[4] M. E. Tiuri, "Radio astronomy receivers", *IEEE Trans. Antenn. Propag.*, vol. 12, pp. 930–938, 1964.

[5] D. F. Wait, "The sensitivity of the Dicke radiometer", *J. Res. Natl. Bur. Stand. – C. Eng. Instrum.*, vol. 71C, no. 2, 1967.

Bronisław Stec has worked at the Military University of Technology since 1962. He participates in works of Scientific Council of the Faculty of Electronics and the Senate of the Military University of Technology, Warsaw, Poland. He implemented a microwave frequency detector, a monopulse direction finding system and a microwave thermograph for non-invasive measurement of temperature distribution inside biological objects. A military warning device and a receiver for recognition of communication satellites were created under his leadership. From 1980 to 1994 he was a Deputy Director and later Director of the Institute of Electronic Circuits. In 1991 he became an Associate Professor at the Military University of Technology.

e-mail: BStec@wel.wat.edu.pl
Faculty of Electronics
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Andrzej Dobrowolski was born in 1966. He graduated from the secondary school in 1985, and then joined the army. He had studied at the Military University of Technology, Warsaw, Poland, for five years. His individual study was focused on infrared sensors. In 1990 – after graduating from the University – he received the M.Sc. degree in the field of radar technology and started working as an engineer in the Institute of Electronic Circuits. He has carried out his research works on gathering an image of full spatial temperature distribution inside biological objects. In 2001 he defended his thesis and received Ph.D. degree in the field of telecommunications. Currently, he works at the Institute of Fundamentals of Electronics, Military University of Technology. His current scientific interests focus on applications of digital signal processors in the multifrequency microwave thermograph and on biomedical signal processing.

e-mail: ADobrowolski@wel.wat.edu.pl
Faculty of Electronics
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Waldemar Susek was born in 1963. He graduated from the secondary school in 1983. He had studied at the Military University of Technology for five years. In 1990 – after completing the University – he received on M.Sc. degree of electronics in the field of radar technology. He worked as an radar engineer in military units. In 1995 he returned to the Military University of Technology and started to work as an assistant in the Institute of Radar Technology. He has carried out research work on radiometric receivers. Nowadays he works in the same Institute and deals with applications of the multifrequency microwave thermography.

e-mail: WSusek@wel.wat.edu.pl
Faculty of Electronics
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Regular paper

Multifrequency microwave thermograph for biomedical applications

Bronisław Stec, Andrzej Dobrowolski, and Waldemar Susek

Abstract — This paper presents problems related to thermal radiation of human bodies in microwave range in aspect of diagnosis of breast carcinoma. A mathematical model of thermal radiation transfer through tissues is introduced and methods of measurement of temperature, depth and size of a heat source, by means of multifrequency microwave thermography are described. Theoretical considerations are supplemented by presentation of experimental results.

Keywords — microwave thermograph, radiometer, thermal radiation, breast carcinoma.

1. Introduction

The passive microwave thermography is based on measurement of thermal radiation emitted by each body, which has the temperature higher than the absolute zero [1–8]. The greatest intensity of radiation is in the infrared, but high attenuation of tissue in this range limits application of the infrared thermography to measurements of skin temperature only. In the microwave frequency range the intensity of radiation is about ten million times lower but attenuation of tissue is low, too. Moreover in this range the intensity of radiation is directly proportional to absolute temperature.

Thermal radiation from a biological body is attenuated by each layer of tissue. Moreover it is reflected and at the same time refracted on the interfaces between different layers [9]. The initial analysis indicates that attenuation, characterised by depth of penetration, is the most important factor.

Analysis of characteristics of radiation and attenuation makes it possible to conclude, that measurements on several frequencies will enable us to estimate the depth and size of the heat source [10–14]. Computer simulation shows that for anatomical depths of the heat source, the maximum intensity of the received radiation is in the frequency range between 1 GHz and 5 GHz. Therefore, we use radiometers working in this range.

In microwave radiometers temperature measurement is made essentially by measuring the thermal noise power [15]. Considering the method of thermal noise power measurement, microwave radiometers can be divided into two groups: total power (compensatory) radiometers and modulation radiometers (Dicke radiometers). Modulation radiometers can be further divided into self-balancing ones and those with compensation of reflection coefficient.

In Dicke radiometers, the output signal carrying information about temperature difference is directly proportional to the gain of high frequency section. This requires high stability and periodical calibration with reference noise source. It is also important that temperature reading depends not only on temperature of examined tissues but also on imperfect matching between antenna and measured object. As a result, Dicke radiometers have to be recalibrated before every single measurement, which is very strenuous in clinical applications.

Using self-balancing radiometer with internal, controlled noise source can eliminate errors resulting from gain changes. Block diagram of such a device with additional compensation of reflection coefficient of antenna-tissue interface is presented in [16]. In such radiometer, the input power is compared with power from internal noise source. As a result of self-balancing procedure voltage at the output of low-pass filter goes to zero. Thus the result of measurement is independent from radiometer's gain.

In previous designs of radiometers constructed in Military University of Technology in Warsaw internal noise source was controlled by analogue regulation loop. Error signal from the output of synchronous detector was transmitted to regulation elements of internal noise generator, which were nonlinear PIN diodes. To ensure quasi-linear conversion characteristics of radiometer, the temperature range had to be limited.

Limitation mentioned above can be successfully overcome using computer-controlled digital regulation loop [17] presented in Fig. 1. Regulation voltage of the internal noise generator changes until zero output voltage is reached. Temperature is calculated by software, which takes into account effects nonlinear characteristics of regulation elements.

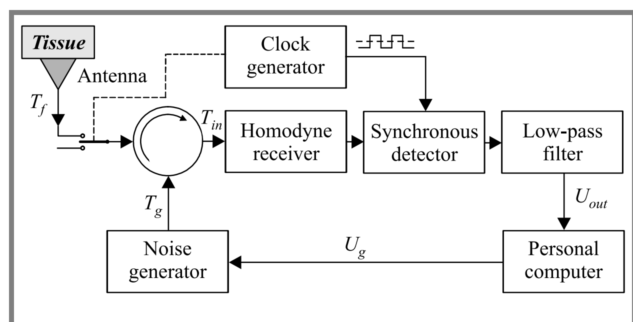


Fig. 1. Block diagram of a radiometer with compensation of reflection coefficient and digitally controlled internal noise source.

Application of optimal control and conversion algorithm reduced the sensitivity of temperature measurement to changes of environmental parameters such as electrical properties and ambient temperature. A homodyne receiver

was used in measurement unit described above, which reduces noise level and improves temperature resolution. Further performance improvement is obtained by fully digital synchronous detection and low-pass filtering by digital signal processor.

Monofrequency radiometry enables measurement of average temperature of a certain area. Therefore, we do not know whether the heat source is cool and just under the skin or perhaps it is hot, but deeply located. In both cases the thermal brightness of the external surface may be equal. This is illustrated in Fig. 2. We only know that there is an area of increased temperature under the antenna, which may indicate the presence of tumour.

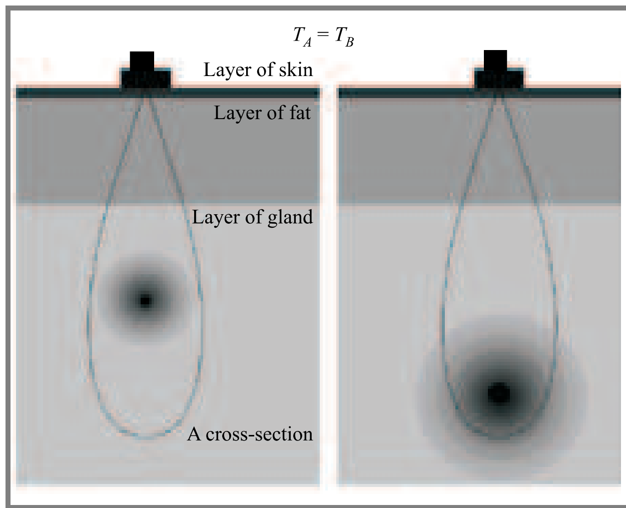


Fig. 2. Monofrequency microwave thermography.

Estimation of spatial temperature distribution inside the investigated object is particularly interesting from the practical point of view. The solution presented below uses the power thermography at different frequencies [18, 19]. This method is based on increase of intensity of thermal radiation and decrease of penetration depth into biological tissues with frequency.

Spatial temperature distribution can be determined by measurements taken in adjacent points. Therefore in next chapters the uni-dimensional analysis of deep-seated profile of temperature distribution will be presented as well as algorithm of inverse transformation converting the results of multiband measurements of temperature brightness on the external surface into deep-seated temperature distribution.

2. Temperature brightness of the external surface

As presented in [20, 21] three-layer model of tissue, increase of physical temperature of the internal heat source T results in increase of the temperature brightness T' of the external surface of tissue in accordance with the following formula

$$T' = T\,\mathrm{e}^{-\frac{d_g}{\delta_g}}\,t_{gf}\,\mathrm{e}^{-\frac{d_f}{\delta_f}}\,t_{fs}\,\mathrm{e}^{-\frac{d_s}{\delta_s}}, \tag{1}$$

where: d_g, d_f, d_s – distances in gland, fat and skin layers; δ_g, δ_f, δ_s – power penetration depths in each layer; t_{gf}, t_{fs} – coefficients of power transmission at layer interfaces.

Moreover, increase of the temperature brightness T_f measured by a radiometer working at frequency f depends on reflection coefficient of the antenna – skin interface Γ_f and can be described as follows

$$T_f = T'\left(1 - \left|\Gamma_f\right|^2\right)K_T = T' K_f\,, \tag{2}$$

where: K_T – a factor defined as displayed temperature increase/real temperature increase ratio; K_f – a resultant coefficient included – moreover – actual antenna – tissue matching.

In medicine, particularly in oncology, there is no point source of heat but unknown temperature distribution $T(z)$. The authors aim at finding $T(z)$ distribution on the basis of discrete measurements of T_{fi} values. This typical problem has no unique solution. Analysis of this problem shows that in order to obtain estimate solution it is necessary to take some simplifying assumptions regarding the source of thermal radiation and type of estimated temperature distribution.

From the structure of tissue and properties of an initial stage of a tumour [22, 23], one can conclude that cancer has a spherical shape and the distribution of temperature originating from it exponentially decreases to zero in the layer of gland. Therefore, the Gauss curve has been assumed as the fitting function for deep-seated temperature distribution:

$$T(z) = T_S\,\mathrm{e}^{-\left(\frac{z-d_g}{\sigma}\right)^2}. \tag{3}$$

This is illustrated in Fig. 3. In this situation, increase of physical temperature of the internal heat source causes increase of temperature brightness of the external surface T_f, and is expressed by the following equation:

$$T_f = T_g\,t_{gf}\,\mathrm{e}^{-\frac{d_f}{\delta_f}}\,t_{fs}\,\mathrm{e}^{-\frac{d_s}{\delta_s}}\,K_f\,. \tag{4}$$

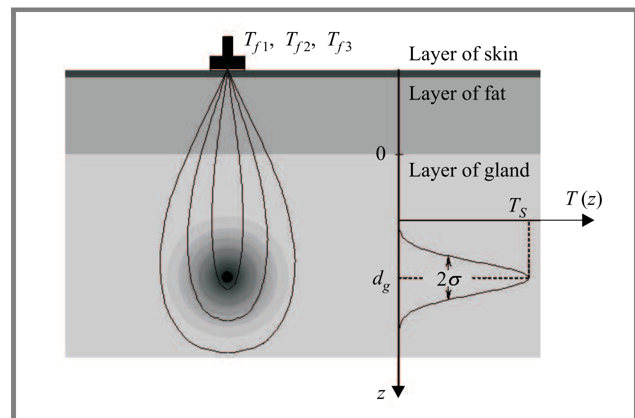


Fig. 3. Distribution of temperature inside biological tissue.

In this equation T_g is the effective temperature on the gland – fat interface and it is defined by integral of temperature distribution $T(z)$ in a range from zero to infinity, with regard to transfer coefficient:

$$\xi(z) = \mathrm{e}^{-\frac{z}{\delta_g}}. \quad (5)$$

From Eqs. (3) and (5) we obtain:

$$T(\xi) = T_S \mathrm{e}^{-\left(\frac{d_g}{\sigma}\right)^2} \xi^{-\frac{\delta_g}{\sigma^2}\left(\delta_g \ln \xi + 2d_g\right)}. \quad (6)$$

Integrating in relation to z from zero to infinity is equivalent to integrating in relation to ξ from zero to one, and consequently the temperature brightness on the gland – fat interface can be written as

$$T_g = \int_0^1 T(\xi) \mathrm{d}\xi. \quad (7)$$

Unfortunately there is no expression, which describes antiderivative of a function shown above, but it is possible to derive a formula for a definite integral as follows:

$$T_g = \frac{T_S \sigma \sqrt{\pi}}{2\delta_g} \mathrm{e}^{\frac{\sigma^2 - 4\delta_g d_g}{4\delta_g^2}} \left[\mathrm{erf}\left(\frac{2\delta_g d_g - \sigma^2}{2\delta_g \sigma}\right) + 1\right]. \quad (8)$$

As according to the initial assumptions, the following relationship is valid across analysed spectral range (1.5–4.4 GHz):

$$\mathrm{erf}\left(\frac{2\delta_g d_g - \sigma^2}{2\delta_g \sigma}\right) \approx 1 \quad (9)$$

and Eq. (8) simplifies to

$$T_g = \frac{T_S \sigma \sqrt{\pi}}{\delta_g} \mathrm{e}^{\frac{\sigma^2 - 4\delta_g d_g}{4\delta_g^2}}. \quad (10)$$

Substituting Eq. (10) into Eq. (4), we obtain the following formula:

$$T_f = \frac{T_S \sigma \sqrt{\pi}}{\delta_g} \mathrm{e}^{\frac{\sigma^2 - 4\delta_g d_g}{4\delta_g^2}} t_{gf} \mathrm{e}^{-\frac{d_f}{\delta_f}} t_{fs} \mathrm{e}^{-\frac{d_s}{\delta_s}} K_f. \quad (11)$$

This formula defines relationship between increase of the physical temperature of the internal heat source and increase of temperature brightness on the external surface measured by a radiometer operating at frequency f.

3. The inverse transformation

3.1. The homogeneous layer of tissue

In case of a single muscle layer, formula (11), which describes temperature brightness of the external surface, measured at frequency f_i as a function of deep-seated temperature distribution can be expressed as

$$T_{fi} = \frac{T_s \sigma \sqrt{\pi}}{\delta_{mi}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{mi} d_m}{4\delta_{mi}^2}} K_{fi}. \quad (12)$$

As a result of a calibration process we obtained coefficients K_{fi} $(i = 1,2,3)$ for three radiometers working at different frequencies. Using these coefficients and proper power penetration depths δ_{mi} we can estimate the real temperature distribution by means of multifrequency measurement.
By solving the set consisting of Eq. (12) for three frequencies (f_1, f_2, f_3):

$$\begin{cases} T_{f1} = \frac{T_s \sigma \sqrt{\pi}}{\delta_{m1}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{m1} d_m}{4\delta_{m1}^2}} K_{f1}, \\ T_{f2} = \frac{T_s \sigma \sqrt{\pi}}{\delta_{m2}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{m2} d_m}{4\delta_{m2}^2}} K_{f2}, \\ T_{f3} = \frac{T_s \sigma \sqrt{\pi}}{\delta_{m3}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{m3} d_m}{4\delta_{m3}^2}} K_{f3}, \end{cases} \quad (13)$$

we obtain expressions describing the real temperature distribution inside investigated tissue:

$$d_m = \frac{\alpha \cdot \delta_{m1}^2\left(\delta_{m2}^2 - \delta_{m3}^2\right) - \beta \cdot \delta_{m3}^2\left(\delta_{m1}^2 - \delta_{m2}^2\right)}{\delta_{m1}^2(\delta_{m2} - \delta_{m3}) - \delta_{m2}^2(\delta_{m1} - \delta_{m3}) + \delta_{m3}^2(\delta_{m1} - \delta_{m2})},$$

$$\sigma = 2\delta_{m1}\delta_{m2}\sqrt{\frac{d_m\left(\delta_{m2}^{-1} - \delta_{m1}^{-1}\right) - \alpha}{\delta_{m1}^2 - \delta_{m2}^2}},$$

$$T_S = \frac{T_{f1}\,\delta_{m1}}{\sigma K_{f1}\sqrt{\pi}} \mathrm{e}^{\frac{4\delta_{m1} d_m - \sigma^2}{4\delta_{m1}^2}}, \quad (14)$$

where:

$$\alpha = \ln\left(\frac{T_{f1} K_{f2}\,\delta_{m1}}{T_{f2} K_{f1}\,\delta_{m2}}\right), \quad \beta = \ln\left(\frac{T_{f2} K_{f3}\,\delta_{m2}}{T_{f3} K_{f2}\,\delta_{m3}}\right).$$

The equations presented above were used for experimental verification of described method used to estimate the temperature, depth and size of an internal heat source, by means of multifrequency microwave thermograph.

3.2. The tissue consist of gland, fat and skin layers

When the measured tissue consists of separate layers of gland, fat and skin it is necessary to perform four-band measurement. In this case, formula (11), which describes temperature brightness on the external surface, measured at frequency f_i as a function of deep-seated temperature distribution can be written as

$$T_{fi} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{gi}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{gi} d_g}{4\delta_{gi}^2}} t_{gfi} \mathrm{e}^{-\frac{d_f}{\delta_{fi}}} t_{fsi} \mathrm{e}^{-\frac{d_s}{\delta_{si}}} K_{fi}. \quad (15)$$

The layer of skin in tested place, taking into account its thickness, can be treated as a thin layer. Consequently, Eq. (15) assumes the following form:

$$T_{fi} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{gi}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{gi} d_g}{4\delta_{gi}^2}} t_{gfi} \mathrm{e}^{-\frac{d_f}{\delta_{fi}}} K_{fi}. \quad (16)$$

Based on the Fresnel's formulas and using dielectric data of tissues from [24], coefficient of power transmission on the gland and fat interface t_{gfi} is about 0.96 across analysed spectral range and for an angle of incidence within $\pm 25°$. As a consequence, formula (16) can be expressed as:

$$T_{fi} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{gi}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{gi} d_g}{4\delta_{gi}^2}} \mathrm{e}^{-\frac{d_f}{\delta_{fi}}} 0.96 K_{fi} . \qquad (17)$$

Conducting four-band measurement using four radiometers and applying formula (17), the following set of nonlinear equations is obtained:

$$\begin{cases} T_{f1} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g1}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g1} d_g}{4\delta_{g1}^2}} \mathrm{e}^{-\frac{d_f}{\delta_{f1}}} 0.96 K_{f1} \\ T_{f2} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g2}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g2} d_g}{4\delta_{g2}^2}} \mathrm{e}^{-\frac{d_f}{\delta_{f2}}} 0.96 K_{f2} \\ T_{f3} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g3}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g3} d_g}{4\delta_{g3}^2}} \mathrm{e}^{-\frac{d_f}{\delta_{f3}}} 0.96 K_{f3} \\ T_{f4} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g4}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g4} d_g}{4\delta_{g4}^2}} \mathrm{e}^{-\frac{d_f}{\delta_{f4}}} 0.96 K_{f4} \end{cases} . \qquad (18)$$

Numerical solution of that set of equations makes it possible to determine all parameters of temperature distribution in the examined tissue (T_S, σ, d_g, d_f). Complete depth of the heat source is expressed by the following equation:

$$d = d_g + d_f + d_s \approx d_g + d_f . \qquad (19)$$

Because it is relatively easy to estimate the thickness of a fat layer, it is possible to determine temperature distribution using three-band measurement introducing into relation (16) a resultant factor defined by equation:

$$K_{ri} = t_{gfi} \mathrm{e}^{-\frac{d_f}{\delta_{fi}}} K_{fi} . \qquad (20)$$

Formula (16) simplifies to

$$T_{fi} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{gi}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{gi} d_g}{4\delta_{gi}^2}} K_{ri} . \qquad (21)$$

As a result of three-band measurement and using formula (21), the following set of equations is obtained:

$$\begin{cases} T_{f1} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g1}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g1} d_g}{4\delta_{g1}^2}} K_{r1} \\ T_{f2} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g2}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g2} d_g}{4\delta_{g2}^2}} K_{r2} \\ T_{f3} = \frac{T_S \sigma \sqrt{\pi}}{\delta_{g3}} \mathrm{e}^{\frac{\sigma^2 - 4\delta_{g3} d_g}{4\delta_{g3}^2}} K_{r3} \end{cases} . \qquad (22)$$

By solving above equations, formulas describing real temperature distribution in examined tissue can be written as:

$$d_g = \frac{\alpha \cdot \delta_{g1}^2 \left(\delta_{g2}^2 - \delta_{g3}^2\right) - \beta \cdot \delta_{g3}^2 \left(\delta_{g1}^2 - \delta_{g2}^2\right)}{\delta_{g1}^2 \left(\delta_{g2} - \delta_{g3}\right) - \delta_{g2}^2 \left(\delta_{g1} - \delta_{g3}\right) + \delta_{g3}^2 \left(\delta_{g1} - \delta_{g2}\right)} ,$$

$$\sigma = 2\delta_{g1}\delta_{g2} \sqrt{\frac{d_g \left(\delta_{g2}^{-1} - \delta_{g1}^{-1}\right) - \alpha}{\delta_{g1}^2 - \delta_{g2}^2}} ,$$

$$T_S = \frac{T_{f1} \delta_{g1}}{\sigma K_{r1} \sqrt{\pi}} \mathrm{e}^{\frac{4\delta_{g1} d_g - \sigma^2}{4\delta_{g1}^2}} , \qquad (23)$$

where:

$$\alpha = \ln\left(\frac{T_{f1} K_{r2} \delta_{g1}}{T_{f2} K_{r1} \delta_{g2}}\right) , \quad \beta = \ln\left(\frac{T_{f2} K_{r3} \delta_{g2}}{T_{f3} K_{r2} \delta_{g3}}\right) .$$

Complete depth of the heat source is increased by width of fat and skin layers and can be calculated using formula (19).

Numerical analysis of formulae shown in this chapter leads to a conclusion that results of estimation of unknown temperature distribution strongly depend on accuracy of T_{fi} temperature measurements. This is a result of rapidly changing exponential functions appearing in final formulae. Therefore it seams reasonable to conduct extra measurements on other frequencies to precise the results. It requires the choice of an optimal algorithm of solving redundant set of nonlinear equations.

4. Experimental results

The setup we used for verification of this method, is presented in Fig. 4. A chunk of beef was used as tissue. We used a poly-propylene tube of diameter 5 mm, containing 1.5% saline solution as the heat source.

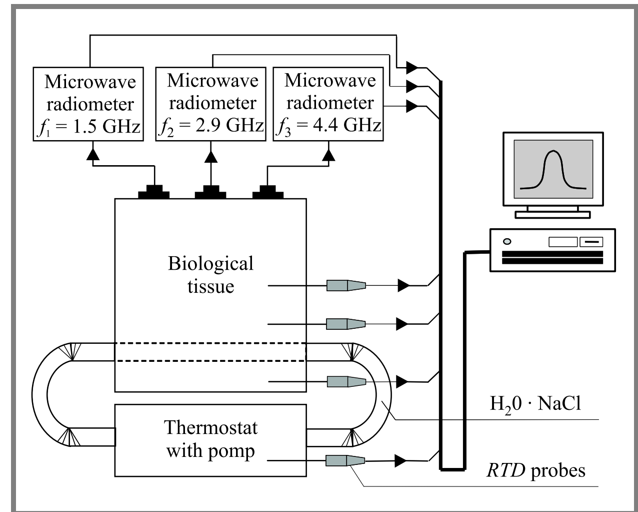


Fig. 4. Scheme of the measurement.

Temperature of solution was regulated by the thermostat. To inspect the physical temperature distribution inside the tissue we used mini hypodermic probes with platinum RTD element. In the range of temperatures from 30°C to 45°C, conductivity of the solution is about 2 S/m, and relative permittivity falls into a range from 70 to 75. Such parameters assure very good coefficient of power transmission across the solution – tissue interface, equal to 0.99 for all measurement frequencies. Side of tube can be omitted in analysis because its thickness is only 0.1 mm.

To test the non-invasive thermometry based on the principles described in this paper, we used the experimental three-band radiometer system, which measured temperature brightness at $f_1 = 1.5$ GHz, $f_2 = 2.9$ GHz and $f_3 = 4.4$ GHz. Main parameters of the radiometric measurement system are given in Table 1. Measurements were made automatically and the results were displayed and stored using a PC.

Table 1
Main parameters of the radiometric system

Parameters	f_i [GHz]		
	1.5	2.9	4.4
RF bandwidth, B_{HF} [MHz]	180	390	485
Noise temperature, T_R [K]	215	425	558
Equivalent integrating time [15], $\tau = \frac{1}{2B_{LF}}$ [s]	6	6	6
Integration time, t [s]	10	10	10
Resolution*, $\Delta T = 2\frac{T_A + T_R}{\sqrt{B_{HF}\tau}}$ [K]	0.032	0.030	0.032

* Brightness temperature resolution (sensitivity) calculated according to [15, 25] for $T_A = 310$ K.

Radiometer working at the highest frequency determines the overall capability of the measurement system to estimate depth and size of a heat source. It is because high frequency microwaves cannot deeply penetrate tissues. Then again, radiometer working at the lowest frequency determines probability of detection of a heat source in antenna's field of view.

As a result of a calibration process we obtain coefficients: $K_{f1} = 1.05$, $K_{f2} = 1.2$ and $K_{f3} = 0.8$ and power penetration depths for the investigated tissue: $\delta_{m1} = 6.6$ mm, $\delta_{m2} = 3.8$ mm, $\delta_{m3} = 3.0$ mm.

Experiment involved wide distribution of temperature. Distribution should be sufficiently wide in relation to dimensions of tube, so that its influence on results was prevailing.

Figure 5 shows typical results obtained for temperature distribution with the following parameters: $T_S = 20$°C, $d_m = 20$ mm and $\sigma = 6.2$ mm. The result of radiometric measurements, calculated on the basis of Eqs. (14) and (3), is shown as a solid curve and the results of direct measurements by the RTD probes are shown as points.

As the brightness temperatures T_{fi} fluctuate randomly due to the nature of thermal radiation, the deep-seated profile of temperature distribution estimated from them also fluctuates randomly. We computed numerically the error in estimating this profile by means of a Monte Carlo technique. We used a 95.5% confidence level to define the accuracy of tissue temperature measurement. It was illustrated by dashed lines in Fig. 5.

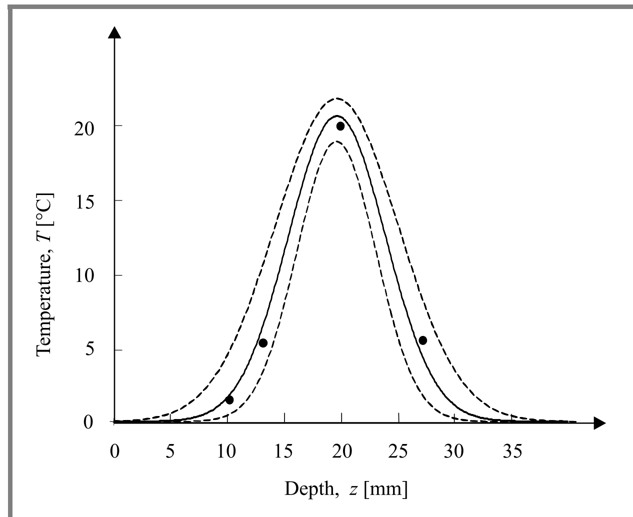


Fig. 5. Results of measurements.

From this chart, we can conclude that the correctness of the presented analysis for single homogenous muscle layer has been confirmed by experiment. Results obtained indicate a possibility of non-invasive detection and measurement of spatial temperature distribution inside a human body by means of multifrequency microwave thermograph.

5. Conclusion

The objective of this article was to present a measurement method suitable to estimate deep-seated temperature distribution by means of a multifrequency microwave thermograph. Theoretical analysis justified initial assumptions, on which the method of estimating temperature distribution inside biological tissues was based. Experimental results confirmed that for single homogenous muscle layer it is possible to estimate the deep-seated profile of temperature distribution after a three-band measurement. For three layers (gland, fat and skin) a multi-band measurement is necessary with four radiometers working at different frequencies. Additional conditions make this measurement possible with only three radiometers.

Considering measurement error, a kind of redundancy in data acquisition is advisable. This makes the whole system less sensitive to random fluctuations of brightness temperatures T_{fi}. Research suggests adding a 5th frequency around 1 GHz (to obtain the deepest penetration) in the radiometer system.

References

[1] F. Bardati, M. Bertero, M. Mongiardo, and D. Solimini, "Singular system analysis of the inversion of microwave radiometric data: applications to biological temperature retrieval", *Inv. Probl.*, vol. 3, pp. 347–370, 1987.

[2] B. Bocquet, J. C. Van De Velde, A. Mamouni, Y. Leroy, G. Giaux, J. Delannoy, and D. Delvalee, "Microwave radiometric imaging at 3 GHz for the exploration of breast tumours", *IEEE Trans. MTT*, vol. 38, no. 6, pp. 791–793, 1990.

[3] D. V. Land, "A clinical microwave thermography system", *IEE Proc.*, vol. 134, no. 2, pp. 193–200, 1987.

[4] Y. Leroy, A. Mamouni, J. C. Van De Velde, B. Bocquet, and B. Dujardin, "Microwave radiometry for non-invasive thermometry", *Automedica*, vol. 8, no. 4, pp. 181–202, 1987.

[5] Y. Leroy, A. Mamouni, J. C. Van De Velde, B. Bocquet, G. Giaux, and J. Dellanoy, "Non-invasive measurement of subcutaneous tissue temperature by microwave radiometry", *Innov. Tech. Biol. Med.*, vol. 12 (special 1), pp. 155–162, 1991.

[6] A. Mamouni, Y. Leroy, B. Bocquet, J. C. Van De Velde, and F. Gelin, "Computation of near – field microwave radiometric signals: definition and experimental verification", *IEEE Trans. MTT*, vol. 39, no. 1, pp. 124–132, 1991.

[7] B. Stec, "Microwave thermography: methods and equipment", in *11th Int. Conf. Microw., Radar Wirel. Commun., Worksh.*, Warsaw, Poland, 1996, pp. 97–114.

[8] F. Sterzer, "Microwave radiometers for non-invasive measurements of subsurface tissue temperatures", *Automedica*, vol. 8, no. 4, pp. 203–211, 1987.

[9] B. Bocquet, P. DeHour, A. Mamouni, J. C. Van De Velde, and Y. Leroy, "Near field microwave radiometric weighting functions for multilayered materials", *J. Electromagnet. Wav. Appl.*, vol. 7, no. 11, pp. 1497–1514, 1993.

[10] L. Dubois, J. Pribetich, J. Falbre, M. Chive, and Y. Moschetto, "Non-invasive microwave multifrequency radiometry used in microwave hyperthermia for bidimensional reconstruction of temperature patterns", *Int. J. Hyperther.*, vol. 9, pp. 415–431, 1993.

[11] K. Maruyma, S. Mizushina, T. Sugiura, G. M. J. Van Leeuwen, J. W. Hand, G. Marrocco, F. Bardati, A. D. Edwards, D. Azzopardi, and D. Land, "Feasibility of non-invasive measurement of deep brain temperature in newborn infants by multifrequency microwave radiometry", *IEEE Trans. MTT*, vol. 48, no. 11, pp. 2141–2147, 2000.

[12] S. Mizushina, T. Shimizu, K. Suzuki, M. Kinomura, H. Ohba, and T. Sugiura, "Retrieval of temperature-depth profiles in biological objects from multifrequency microwave radiometric data", *J. Electromagnet. Wav. Appl.*, vol. 7, no. 11, pp. 1515–1548, 1993.

[13] H. Ohba, M. Kinomura, M. Ito, T. Sugiura, and S. Mizushina, "Multifrequency microwave radiometry for non-invasive thermometry using a new temperature profile model function", in *Asia Pacif. Microw. Conf.*, Tokyo, Japan, 1994, pp. 401–404.

[14] F. Bardati, V. J. Brown, and P. Tognolatti, "Temperature reconstructions in a dielectric cylinder by multifrequency microwave radiometry", *J. Electromagnet. Wav. Appl.*, vol. 7, no. 11, pp. 1549–1571, 1993.

[15] M. E. Tiuri, "Radio astronomy receivers", *IEEE Trans. Antenn. Propag.*, vol. 12, pp. 930–938, 1964.

[16] B. Stec and M. Żurawski, "Compensated microwave thermometer for medical applications", in *Asia Pacif. Microw. Conf.*, Tokyo, Japan, 1994, pp. 405–408.

[17] B. Stec and W. Susek, "A 4.4 GHz microwave thermometer with compensation of reflection coefficient", in *13th Int. Conf. Microw., Radar Wirel. Commun.*, Wrocław, Poland, 2000, pp. 453–456.

[18] B. Stec and A. Dobrowolski, "Estimation of internal distribution of temperature inside biological tissues by means of multifrequency microwave thermograph", in *13th Int. Conf. Microw., Radar Wirel. Commun.*, Wrocław, Poland, 2000, pp. 577–580.

[19] B. Stec and A. Dobrowolski, "Estimation of internal distribution of temperature inside biological tissues by means of multifrequency microwave thermograph", *J. Telecommun. Inform. Technol.*, no. 1, pp. 39–42, 2002.

[20] B. Stec, A. Dobrowolski, and W. Susek, "Estimation of temperature distribution inside biological tissues by means of multifrequency microwave thermograph", in *23rd Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Istanbul, Turkey, 2001.

[21] B. Stec, A. Dobrowolski, and W. Susek, "Estimation of deep-seated profile of temperature distribution inside biological tissues by means of multifrequency microwave thermograph", in *IEEE MTT-S Int. Microw. Symp.*, Seattle, USA, 2002, pp. 2261–2264.

[22] W. Kazanowska, P. Knapp, A. Mazurek, T. Filipkowski, and B. Stec, "An attempt to employ microwave thermography for the evaluation of potential danger of breast cancer", in *8th Int. Meet. Gynaecol. Oncol.*, Barcelona, Spain, 1993.

[23] T. Maeda, "Diagnosis of mammary gland diseases using microwave thermography: a study on adjunctive diagnostic methods in relation to different factors in breast diseases", *Nippon-Geka-Gakkai-Zasshi*, vol. 91, no. 5, pp. 622–630, 1990.

[24] M. A. Stuchly and S. S. Stuchly, "Dielectric properties of biological substances – tabulated", *J. Microw. Pow.*, vol. 15, no. 1, pp. 19–26, 1980.

[25] D. F. Wait, "The sensitivity of the Dicke radiometer", *J. Res. Natl. Bur. Stand. – C. Eng. Instrum.*, vol. 71C, no. 2, 1967.

Bronisław Stec – for biography, see this issue, p. 115.

Andrzej Dobrowolski, **Waldemar Susek** – for biography, see this issue, p. 116.

INFORMATION FOR AUTHORS

The *Journal of Telecommunications and Information Technology* is published quarterly. It comprises original contributions, both regular papers and letters, dealing with a broad range of topics related to telecommunications and information technology. Items included in the journal report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

The *Journal* is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, speech devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology. Suitable research-related manuscripts should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published by invitation only.

Papers published by invitation and regular papers should contain up to 15 and 8 printed pages respectively (one printed page corresponds approximately to 3 double-space pages of manuscript, where one page contains approximately 2000 characters).

Manuscript: An original and two copies of the manuscript must be submitted, each completed with all illustrations and tables attached at the end of the papers. Tables and figures have to be numbered consecutively with Arabic numerals. The manuscript must include an abstract limited to approximately 100 words. The abstract should contain four points: statement of the problem, assumptions and methodology, results and conclusion, or discussion, of the importance of the results. The manuscript should be double-spaced on only one side of each A4 sheet (210 × 297 mm). Computer notation such as Fortran, Matlab, Mathematica etc., for formulae, indices, etc., is not acceptable and will result in automatic rejection of the manuscript. The style of references, abbreviations, etc., should follow the standard IEEE format.

References should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. The **reference entry** (correctly punctuated according to the following rules and examples) **has to contain.**

From journals and other serial publications: initial(s) and second name(s) of the author(s), full title of publication (transliterated into Latin characters in case it is in Russian, possibly preceded by the title in Russian characters), appropriately abbreviated title of periodical, volume number, first and last page number, year. E.g.:

[1] Y. Namihira, "Relationship between nonlinear effective area and modefield diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

From non-periodical, collective publications: as above, but after title – the name(s) of editor(s), title of volume and/or edition number, publisher(s) name(s) and place of edition, inclusive pages of article, year. E.g.:

[2] S. Demri, E. Orłowska, "Informational representability: Abstract models versus concrete models" in *Fuzzy Sets, Logics and Reasoning about Knowledge*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

From books: initial(s) and name(s) of the author(s), place of edition, title, publisher(s), year. E.g.:

[3] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

Figure captions should be started on separate sheet of papers and must be double-spaced.

Illustration: Original illustrations should be submitted. All line drawings should be prepared on white drawing paper in black India ink. Drawings in Corel Draw and Postscript formats are preferred. Colour illustrations are accepted only in exceptional circumstances. Lettering should be large enough to be readily legible when drawing is reduced to two- or one-column width – as much as 4:1 reduction from the original. Photographs should be used sparingly. All photographs must be gloss prints. All materials, including drawings and photographs, should be no larger than 175 × 260 mm.

Page number: Number all pages, including tables and illustrations (which should be grouped at the end), in a single series, with no omitted numbers.

Electronic form: A floppy disk together with the hard copy of the manuscript should be submitted. It is important to ensure that the diskette version and the printed version are identical. The diskette should be labelled with the following information: a) the operating system and word-processing software used, b) in case of UNIX media, the method of extraction (i.e. tar) applied, c) file name(s) related to manuscript. The diskette should be properly packed in order to avoid possible damage during transit.

Among various acceptable word processor formats, $\TeX$ and $\LaTeX$ are preferable. The *Journal's* style file is available to authors.

Galley proofs: Proofs should be returned by authors as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within two weeks after receiving the offprints.

The copy of the "Journal" shall be provided to each author of papers.

Copyright: Manuscript submitted to this journal may not have been published and will not be simultaneously submitted or published elsewhere. Submitting a manuscript, the authors agree to automatically transfer the copyright for their article to the publisher if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and also all translation rights. No part of the present journal may be reproduced in any form nor transmitted or translated into a machine language without permission in written form from the publisher.

Biographies and photographs of authors are printed with each paper. Send a brief professional biography not exceeding 100 words and a gloss photo of each author with the manuscript.

New approach to power semiconductor devices modeling

A. Napieralski and M. Napieralska — Invited paper — 80

Process and device requirements for mixed-signal integrated circuits in broadband networking

K. Iniewski, M. Syrzycki, and S. P. Voinigescu — Invited paper — 90

Evolution and recent advances in RF/microwave transistors

J. J. Liou and F. Schwierz — Invited paper — 99

Regular papers

Photonic crystal fibre characterisation with the method of lines

I. A. Goncharenko and M. Marciniak — Regular paper — 106

Sensitivity of microwave radiometers with square - law and linear detectors

B. Stec, A. Dobrowolski, and W. Susek — Regular paper — 112

Multifrequency microwave thermograph for biomedical applications

B. Stec, A. Dobrowolski, and W. Susek — Regular paper — 117

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 872 43 88
tel./fax: +48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit