JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2009

SOI Technology: An Opportunity for RF Designers?
J.-P. Raskin — *Invited Paper* — 3

The Impact of Externally Applied Mechanical Stress on Analog and RF Performances of SOI MOSFETs
M. Emam, S. Houri, D. Vanhoenacker-Janvier, and J.-P. Raskin — *Paper* — 18

Prospects and Development of Vertical Normally-off JFETs in SiC
M. Bakowski — *Invited Paper* — 25

Variation Analysis of CMOS Technologies Using Surface-Potential MOSFET Model
H. J. Mattausch et al. — *Invited Paper* — 37

Analysis of the Dispersion of Electrical Parameters and Characteristics of FinFET Devices
A. Malinowski et al. — *Paper* — 45

Rare Earth Silicate Formation: A Route Towards High-*k* for the 22 nm Node and Beyond
I. Z. Mitrovic and S. Hall — *Invited Paper* — 51

Technology of MISFET with $SiO_2/BaTiO_3$ System as a Gate Insulator
P. Firek and J. Szmidt — *Paper* — 61

Modeling, Simulation and Calibration of Silicon Wet Etching
A. Kociubiński, M. Duk, T. Bieniek, and P. Janus — *Paper* — 65

Si-Based Electrodes for Potentiometric Measurements of Aqueous Solutions
M. Zaborowski, D. Tomaszewski, B. Jaroszewicz, and P. Grabiec — *Paper* — 71

(*Contents Continued on Back Cover*)

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 updated in 2008 by SIA (Semiconductor Industry Association) reaches ahead to the years 2016–2022. Unfortunately, a comparison with former SIA forecasts indicates that in certain aspects they become less aggressive (that is less 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 (SOI, SON, SiGe or SiC) and architecture (ultra-thin body, double-gate, multiple-gate) solutions. 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. Plenty of new applications are being for fast analog and RF circuits, as well as for information processing ones. It is clear that with the anticipated peak $f_{\max} = 425$ GHz and $f_T = 395$ GHz to be reached by RF SiGe-base bipolar transistors in 2014, according to the 2008 update of ITRS, a lot of effort must be put into the development of appropriate material, processing, characterization and modeling. While progress in the bipolar technology is impressive, the increase of MOSFET speed is even more so. The same issue of ITRS predicts on-chip clock of ~ 14 GHz for 2022.

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 papers devoted to RF operation of SOI technology, device concepts, numerical analysis of device and circuit parameter variation, fabrication, process simulation, sensors and characterization.

The ongoing increase of operating frequencies applies to satellite communication systems, too, where millimeter band use is expected soon. This raises the issue of adverse propagation phenomena, in particular weather-dependent atmospheric attenuation. As this attenuation rises with frequency, it must be mitigated with proper satellite and antenna design.

Effective use of limited pool of radio frequencies requires proper spectrum management; relevant activities are largely carried out within ITU. The existing spectrum management arrangements are criticized as inefficient in many respects, and several proposals for improvements are presented, taking into account recent advances in technology.

Broadband wireless systems like WiMAX face the problem of limited radio spectrum in a particularly acute way. Spectrum efficiency and system capacity can be improved by dynamic resource management in so called flexible radios, allowing optimized use of limited channel capacities and computing resources. Flexible radio architecture must also deal with interactive use of spectrum by multiple applications, while meeting stringent quality of service (QoS) requirements. The key solution is a software defined radio with embedded intelligence, able to sense the current environment defined as spectrum occupancy, interference, etc., and adapt to it, to ensure the best performance possible with variable spectrum constraints. A cognitive radio with human-like intelligence is required, working in environment whose parameters are not known a priori and can rapidly change. The mathematical foundations for such devices include Bayesian probability theory, maximum entropy principle, etc., to optimize signal sensing and use of multiple antennae.

Advanced as it is, current RF technology fails to deliver coherent, high power CW terahertz radiation, however. As the need to utilize THz frequencies in communications, medicine, security, inspection, etc., grows, the intersubband-based quantum cascade laser may become the source of choice. Before such devices become practical, they must be engineered to work without cryogenic cooling required today.

The quality of multimedia service as perceived by the user of a next generation network (NGN) can be greatly affected by the performance of signaling system during connection setup phase, in particular by delays incurred when heterogenous multi-domain network provides the service. The impact of signaling system performance on user quality of experience (QoE) has been analyzed, with focus on the signaling system and procedures defined within the EuQoS project.

The next paper in this issue deals with the recommendations and regulations regarding the pan-European eCall programme. This is a road safety improvement effort aimed to reduce the current number of road fatalities of over 40,000 a year, providing a standardized in-vehicle emergency call service, with automated feed of accident location and other information to relevant public safety answering point (PSAP). Besides right technology, effective solution needs a harmonized pan-European regulatory framework.

Secure identification of user is of critical importance to many services, as ID theft and other security treats increase. Digital signatures are often user ID-based: user's e-mail, phone number, etc., serves as a public key. Unfortunately, the analysis presented shows such architectures can provide only medium level of security, albeit some improvements are possible.

We hope the Readers will find this issue of the *Journal of Telecommunications and Information Technology* useful and interesting.

Andrzej Jakubowski
Lidia Łukasiak
Guest Editors

Invited paper

SOI Technology: An Opportunity for RF Designers?

Jean-Pierre Raskin

Abstract— This last decade silicon-on-insulator (SOI) MOSFET technology has demonstrated its potentialities for high frequency (reaching cutoff frequencies close to 500 GHz for n-MOSFETs) and for harsh environments (high temperature, radiation) commercial applications. For RF and system-on-chip applications, SOI also presents the major advantage of providing high resistivity substrate capabilities, leading to substantially reduced substrate losses. Substrate resistivity values higher than 1 kΩ cm can easily be achieved and high resistivity silicon (HRS) is commonly foreseen as a promising substrate for radio frequency integrated circuits (RFIC) and mixed signal applications. In this paper, based on several experimental and simulation results the interest, limitations but also possible future improvements of the SOI MOS technology are presented.

Keywords— crosstalk, high resistivity silicon substrate, MOSFET, nonlinearities, silicon-on-insulator, wideband characterization.

1. Introduction

The semiconductor technology has been progressing enormously these last decades, such evolution has been driven by the continuous look for the increase of the operation speed and the integration density of complex digital circuits [1]. In the early 70's a scaling-down procedure of the transistor dimensions established by Dennard and coworkers [2] was proposed to pave the way to reaching both objectives. From those days to now, the keystone of the semiconductor industry has been the optimization of this scaling-down procedure.

The communication industry has always been a very challenging and profitable market for the semiconductor companies. The new communication systems are today very demanding; high frequency, high degree of integration, multistandards, low power consumption, and they have to present good performance even under harsh environment such as high temperature, radiation, etc. The integration and power consumption reduction of the digital part will further improve with the continued downscaling of technologies. The bottleneck for further advancement is the analog front-end. Present-day transceivers often consist of three or four chipset solutions combined with several external components. A reduction of the external components is essential to obtain lower cost, power consumption and weight, but it will lead to a fundamental change in the design of analog front-end architectures. The analog front-end requires a high performance technology, like GaAs or silicon bipolar, with devices that can easily achieve operating frequencies in the GHz range. For the digital signal processor a small device feature size is essential for the implementation of complex algorithms. Therefore, it appears that only the best submicron CMOS technologies could provide a feasible and cost-effective integration of the communication systems.

This last decade metal oxide semiconductor (MOS) transistors have reached amazingly high operation speed and the semiconductor community has started to notice the radio frequency (RF) possibilities of such mainstream devices. Silicon-on-insulator (SOI) MOSFET technology has demonstrated its potentialities for high frequency (reaching cutoff frequencies close to 500 GHz for n-MOSFETs [3]) and for harsh environments (high temperature, radiations) commercial applications.

From its early development phase till recent years, SOI has grown from a mere scientific curiosity into a mature technology. Partially depleted (PD) SOI is now massively serving the 45-nm digital market where it is seen as a low cost – low power alternative to bulk silicon. Fully depleted (FD) devices are also widely spread as they outperform existing semiconductor technologies for extremely low power analog applications [4].

For RF and system-on-chip applications, SOI also presents the major advantage of providing high resistivity substrate capabilities, leading to substantially reduced substrate losses. Substrate resistivity values higher than 1 kΩcm can easily be achieved and high resistivity silicon (HRS) is commonly foreseen as a promising substrate for radio frequency integrated circuits (RFIC) and mixed signal applications [5].

In this paper, based on several experimental and simulation results the interest, limitations but also possible future improvements of the SOI MOSFET technology for microwaves and millimeter-waves applications are presented.

2. State of the Art RF Performance

Since the invention of the bipolar transistor in 1947, the operating frequencies of integrated transistors have been improved every year. In 1958, a cut-off frequency above 1 GHz is reached with a germanium bipolar transistor [6]. Since that date, several integrated technologies have been investigated and improved to further increase the operating frequency of transistors. In 1965, a GaAs metal semiconductor field effect transistor (MESFET) appears in the literature [7]. In 1973, a maximum oscillation frequency ($f_{\max}$) of 100 GHz is measured for a FET [8]. In 1980, a new architecture of field effect transistor with high electron mo-

bility (HEMT) is proposed and fabricated [9]. In 1995, a cutoff frequency $f_{\max}$ higher than 500 GHz is extrapolated for a HEMT [10]. In 2000, the limit of 1 THz is reached with III-V heterostructure bipolar transistor (HBT) [11] and even overpassed by HEMT in 2007 [12].

It is only in 1996, thanks to the successful downscaling of the silicon MOSFET gate, that cutoff frequencies higher than 200 GHz are presented [13]. Since that date, the interest in MOSFETs for low voltage, low power, high integration mixed-mode ICs (digital and analog parts on the same chip) in the field of microwaves and millimeter-waves applications has been constantly growing. MOSFET is a well-known, well-controlled and mature technology, as well as cost effective, which makes it the key technology for mass production.

Nowadays, thanks to the introduction of mobility booster such as strained silicon channel, cutoff frequencies close to 500 GHz and 350 GHz are achieved, respectively, for n- and p-MOSFETs [3] with the channel length of 30 nm.

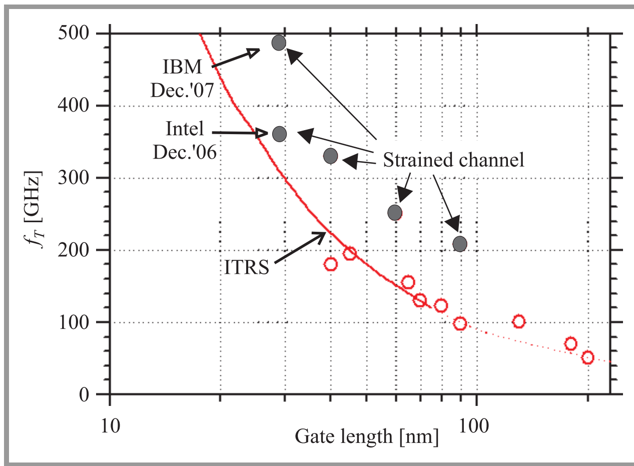


Fig. 1. State of the art current gain cutoff frequency as a function of gate length for unstrained and strained Si and SOI NMOSFETs.

Figure 1 presents the state of the art current gain cutoff frequency (f_T) for n-type MOSFETs as a function of gate length. In that graph, the continuous line represents the prediction from the International Technology Roadmap for Semiconductors (ITRS) published in 2006 [14]. Despite the poor carrier mobility of electrons in silicon compared to III-V materials, silicon MOSFET can be considered as a competitive technology for high frequency applications. It is worth to notice that strained channel silicon MOSFETs even overcome the ITRS roadmap values which gives quite good prospects for silicon technology still for certainly more than 15 years from now on.

3. Main Limiting Factors

Historically, device scaling remains the primary method by which the semiconductor industry has improved productivity and performance. From the 100-nm technology node, CMOS technologies have been facing many grand technological challenges. In this context, the most critical issue consists in the so-called short-channel effects (SCE).

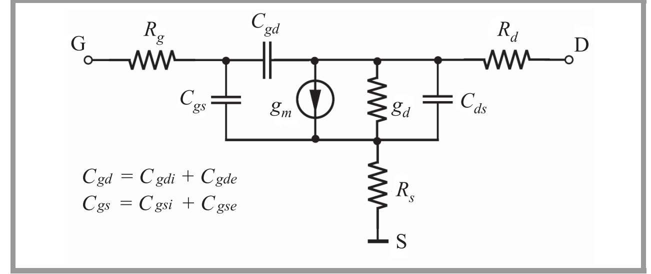


Fig. 2. Small-signal lumped equivalent circuit of MOSFET.

These parasitic effects tend to degrade the subthreshold characteristic, increase the leakage current and lead to a dependence of threshold voltage with respect to the channel length. Those static SCE have been reported theoretically and experimentally in the literature and solutions have been proposed. However, only a few publications have analyzed the limitation or degradation of high frequency characteristics versus the downscaling of the channel length. Considering a classical small-signal equivalent circuit for MOSFET as presented in Fig. 2, we can define the cutoff frequencies f_c, f_T and $f_{\max}$ representing the intrinsic (related to the useful MOSFET effect), the current gain and the available power gain cutoff frequencies, by expressions (1) to (3), respectively:

$$f_c = \frac{g_m}{2\pi C_{gs}}, \tag{1}$$

$$f_T \approx \frac{f_c}{\left(1+\frac{C_{gd}}{C_{gs}}\right)+\left(R_s+R_d\right)\left(\frac{C_{gd}}{C_{gs}}\left(g_m+g_d\right)+g_d\right)}, \tag{2}$$

$$f_{\max} \approx \frac{f_c}{2\left(1+\frac{C_{gd}}{C_{gs}}\right)\sqrt{g_d\left(R_g+R_s\right)+\frac{1}{2}\frac{C_{gd}}{C_{gs}}\left(R_s g_m+\frac{C_{gd}}{C_{gs}}\right)}}, \tag{3}$$

where: g_m – the gate transconductance, g_d – the ouput conductance, C_{gs}, C_{gd} and C_{ds} – the gate-to-source, gate-to-drain, and drain-to-source capacitances, respectively, R_g, R_d and R_s – the gate, drain and source access resistances, respectively.

Figure 3 represents a schematic cross-section of a classical silicon MOSFET where the different components of parasitic source and drain resistances and capacitances are illustrated.

The intrinsic cutoff frequency, f_c, measures the intrinsic ability of a field effect transistor (FET) to amplify high frequency signals. As reported in [15], the f_c values are a factor of 1.5 to 2 higher for HEMTs than for silicon MOSFETs with comparable gate length, and this is mainly explained by the respective dynamic properties of the two types of semiconductors (difference of g_m which is directly proportional to the carrier mobility). In order to enhance

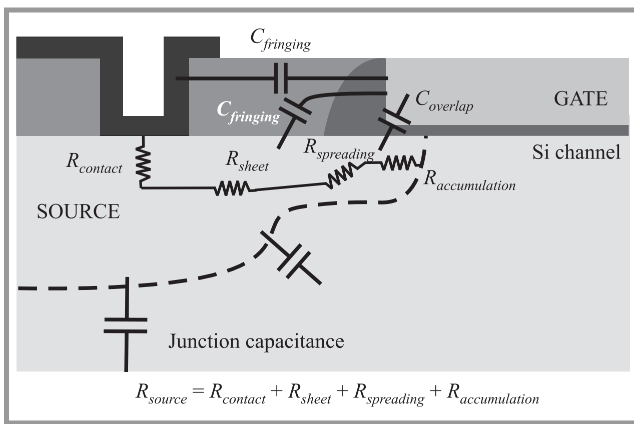


Fig. 3. Schematic cross-section of a Si MOSFET illustrating the different access resistances to the channel and the surrounding overlap and fringing capacitances.

the carrier mobility in silicon channel and then to improve the current drive and high frequency characteristics [3] of MOSFETs, strained n- and p-MOSFETs have been investigated these last years. Beside the carrier mobility difference between Si and III-V materials, it has been demonstrated that the $f_{\max}/f_T$ ratio is lower in the case of Si devices. As explained in [15], besides the well-know degradation of high frequency characteristics due to access resistances (R_g, R_d and R_s), the decrease of the ratios g_m/g_d and C_{gs}/C_{gd} in CMOS technology strongly contributes to the limiting improvement of f_T and $f_{\max}$ with the transistor channel length skrinkage. The increase of the output conductance, g_d, with the reduction of gate channel length is one of the well-known short channel effects of FET devices. The degradation of the ratio C_{gs}/C_{gd} means a loss of channel charge control by the gate and an increase of the direct coupling capacitance between gate (input) and drain (output) terminals. The self-aligned source and drain regions, one of the main advantages of MOSFET structure, are also a reason for the increase of parasitic capacitances between source and gate and more importantly drain and gate. As demonstrated in [15], from extraction results the C_{gs}/C_{gd} ratio is equal to 7.8 for the HEMT and only to 1.5–1.6 in the case of a MOSFET with 90 nm gate length.

It is therefore obvious that the optimization of these internal parameters will be crucial in order to further improve cutoff frequencies of ultra deep submicron MOSFETs. The impact of lightly doped drain (LDD) dose and energy implant as well as annealing temperature and time on C_{gs}/C_{gd} ratio, g_m and g_d and then on $f_{\max}$ has been investigated in [16]. The results demonstrate that LDD implant can indeed be considered as an optimization parameter for improving $f_{\max}$ and especially the ratio $G_{ass}/NF_{\min}$ (G_{ass} and $NF_{\min}$ being the associated power gain and the minimum RF noise figure, respectively), which is the most important figure of merit for low noise microwave applications. However, the optimization window is quite narrow and it seems difficult for a given technological node to get higher C_{gs}/C_{gd} and g_m/g_d ratios than 2 and 6, respectively, for a classical sub-100-nm gate length MOSFET structure. It is the main reason why $f_{\max}$ is almost equal to f_T in the case of MOSFETs and not 1.5 to 2 times higher as in the case of HEMTs with similar gate length and characterized by C_{gs}/C_{gd} and g_m/g_d ratios of 8 and 20, respectively.

In order to further improve the microwave performance of deep submicrometer MOSFETs, it seems crucial to keep the parasitic resistances and capacitances as low as possible, as predicted by ITRS and shown in Fig. 4 and to consider alternative MOS structures for which the C_{gs}/C_{gd} and g_m/g_d ratios (analog SCE) are improved.

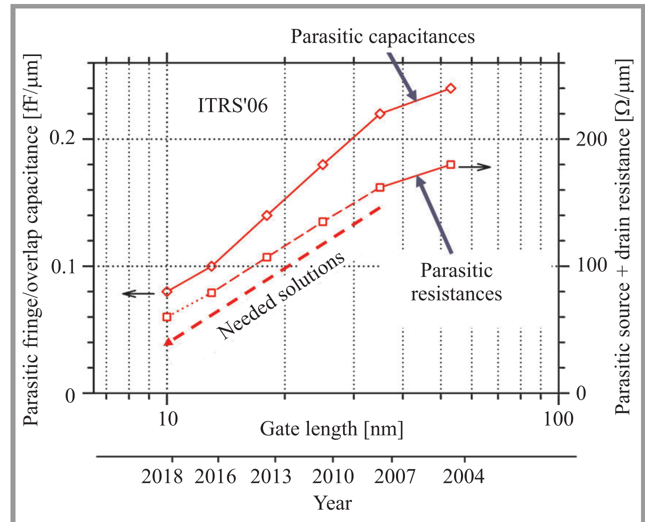


Fig. 4. Parasitic capacitances and source, and drain resistances as a function of the gate length published in ITRS'06 [12].

Several technological options have been presented in the litterature those last years to push further the digital and analog performance limits of single gate Si MOSFETs such as:

- Move from bulk Si MOSFETs to partially depleted [17] or fully depleted [18] SOI MOSFETs to enhance the gate electrostatic control on the channel carriers and thus minimize the SCE. Nowadays, ultra thin body (UTB) MOSFET in SOI technology with a silicon body thickness less that 10 nm has been proposed [19], [20]. Thanks to the buried oxide layer (BOX) underneath the SOI transistors, their junction capacitances (noted *Junction capacitance* in Fig. 3) to the Si substrate are drastically reduced.
- Strained MOSFETs have been largely investigated lately to improve the carrier mobility. The mechanical stress in the channel orginates from specific process steps [21] added into the classical CMOS process flow. Nowadays, strained SOI wafers are produced as well for which the top silicon layer is under a certain level of stress [22], [23].
- Low Schottky barrier contacts [24]–[28] are foreseen as a very interesting candidate to lower the source/drain contact resistances, to form abrupt junctions (no overlap), and drastically reduce the thermal budget for CMOS process.

- Metal gate allows to get rid of loss of electrostatic gate control related to the polysilicon gate depletion [29], [30], as well as to reduce the gate sheet resistance.
- Low-k and air gap [31], [32] should be introduced to reduce fringing capacitances between gate-to-source and gate-to-drain electrodes.
- SOI wafers with thin BOX have been proposed these last years to reduce SCE (for instance, DIBL) but also to lower self-heating issues [19], [20], [33], [34].
- High resistivity silicon substrate has demonstrated superior characteristics for the integration of high quality passive elements such as transmission lines [35], inductors [36], etc., as well as for reduction of the crosstalk between circuit blocks integrated on the same silicon chip [5].

This last point will be developed in detail in Section 5. Figure 5 schematically presents the cross-section of what we can call an ultimate single gate MOSFET basically including the technological options listed above. Unstrained p-type MOSFET including a metal gate and low Schottky barrier source and drain contacts has been built and characterized over a wide frequency band in [27]. The device architecture features a 20 nm thick SOI channel, a 2.4 nm SiO_2 gate oxide, a metallic tungsten gate and 15 nm-wide SiN spacers. The integration of a low Schottky barrier silicide (PtSi) coupled to boron segregation demonstrates a 50% improvement on the current drive accompanied by reinforced immunity against SCE when compared to the dopant-free approach. This constitutes the first implementation of a dopant segregated band-edge silicide obtained by implant-to-silicide (ITS) and activated at low temperature (500°C). The RF characterization unveils a unity current gain cut-off frequency f_T of 180 GHz for a 30 nm gate long device as shown in Fig. 6. This constitutes the best result reported in literature [37] for unstrained channel fully depleted SOI p-MOSFETs.

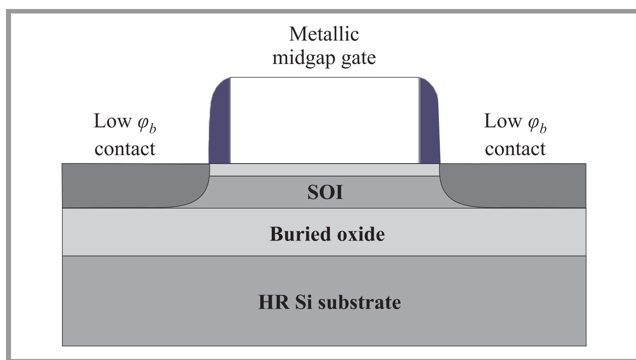


Fig. 5. Schematic cross-section view of an optimized single fully depleted SOI MOSFET.

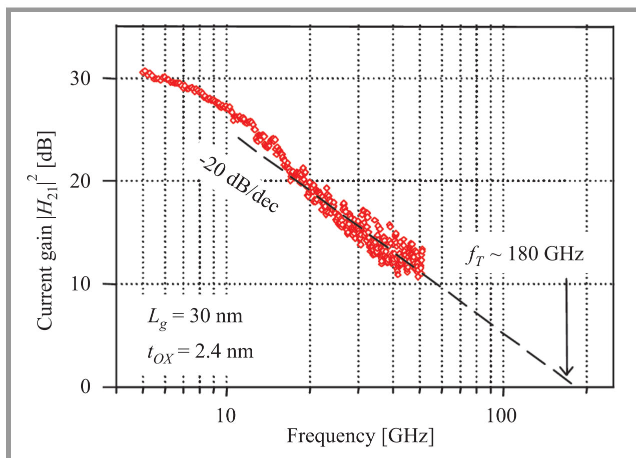


Fig. 6. Current gain as a function of frequency for a 30 nm p-type segregated PtSi Schottky barrier MOSFET.

Multiple gate MOSFETs are often cited as the ultimate MOS devices to reduce the SCE. The analog and RF performances of FinFETs are presented in the following section.

4. RF Performance of a Multigate MOSFET: FinFET

To reduce the SCE in nanometer scale MOSFETs, multiple-gate architectures emerge as one of the most promising novel device structures, thanks to the simultaneous control of the channel by more than one gate. The idea of the double-gate (DG) MOSFET was first introduced by J.-P. Colinge [38]. Starting by the FinFET [39], other multiple-gate SOI MOSFETs have been introduced since [40] such as triple-gate (TG), FinFET, pi-gate (PG), quadruple-gate (QG), omega-gate (Ω-G), etc. Many works have investigated and demonstrated the great potential of multiple-gate devices to comply with the I_{on}/I_{off} requirements of the ITRS for logic operation [40], [41].

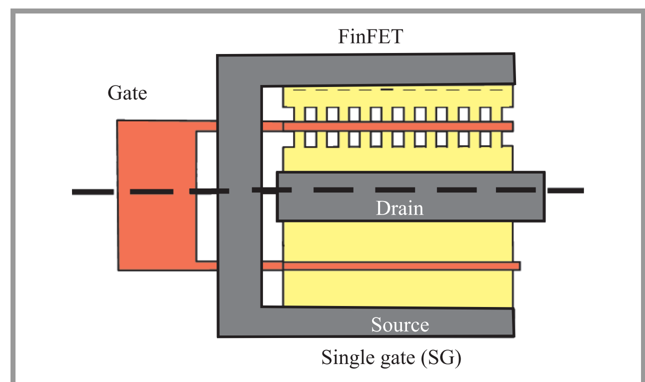


Fig. 7. Schematic top view of a FinFET composed of 10 fins (upper) and SG MOSFET (lower) occupying the same active silicon foot print.

Indeed, FinFETs are known to be promising devices for high density digital applications in the sub-65 nm nodes due to their high immunity to short channel effects and their excellent compatibility with planar CMOS process. Most

Fig. 8. DC and RF characteristics of 60 nm gate length FinFET for various fin widths (W_{fin}): (a) transfer characteristic in log scale; (b) extracted access resistances; (c) transfer characteristic in linear scale and gate transconductance; (d) current gain and maximum available power gain versus frequency.

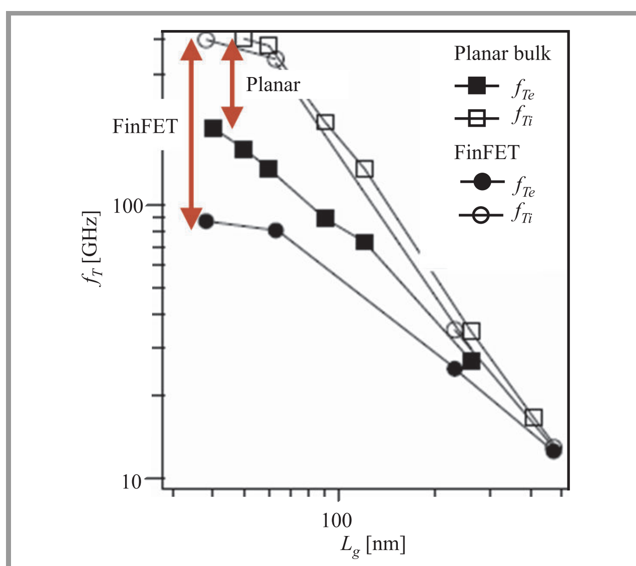


Fig. 9. Extracted intrinsic (f_{Ti}) and extrinsic (f_{Te}) current gain cutoff frequencies for a SG MOSFET and FinFET as a function of the channel length.

of the investigations performed on FinFETs have focused on their technological aspects and perspectives for digital applications [42], [43], while only a few have assessed their analog figures of merit [44], [45]. In this section, the RF performance of FinFETs with various geometries is presented.

FinFETs are fabricated on a SOI wafer with 60 nm Si film on 145 nm of buried oxide, with $< 100 >$ and $< 110 >$ Si planes for top and lateral channels, respectively. The silicon active area is patterned using 193 nm lithography with aggressive resist and oxide hard mask trimming to define narrow silicon fins. A hydrogen anneal and a sidewall oxidation are used for surface smoothening and corner rounding. The fin patterning resulted in a fin height (H_{fin}) of 60 nm, fin width (W_{fin}) of 22, 32 and 42 nm, and fin spacing (S_{fin}) of 328 nm. The gate stack consisting of a plasma nitrided oxide with equivalent oxide thickness equal to 1.8 nm, as measured on planar devices, and 100 nm polysilicon is deposited. Gate lengths (L_g) of 40, 60 and 120 nm are fabricated. High angle As/BF_2 extensions are then implanted and a 40 nm-thick selective epitaxial growth (SEG) is per-

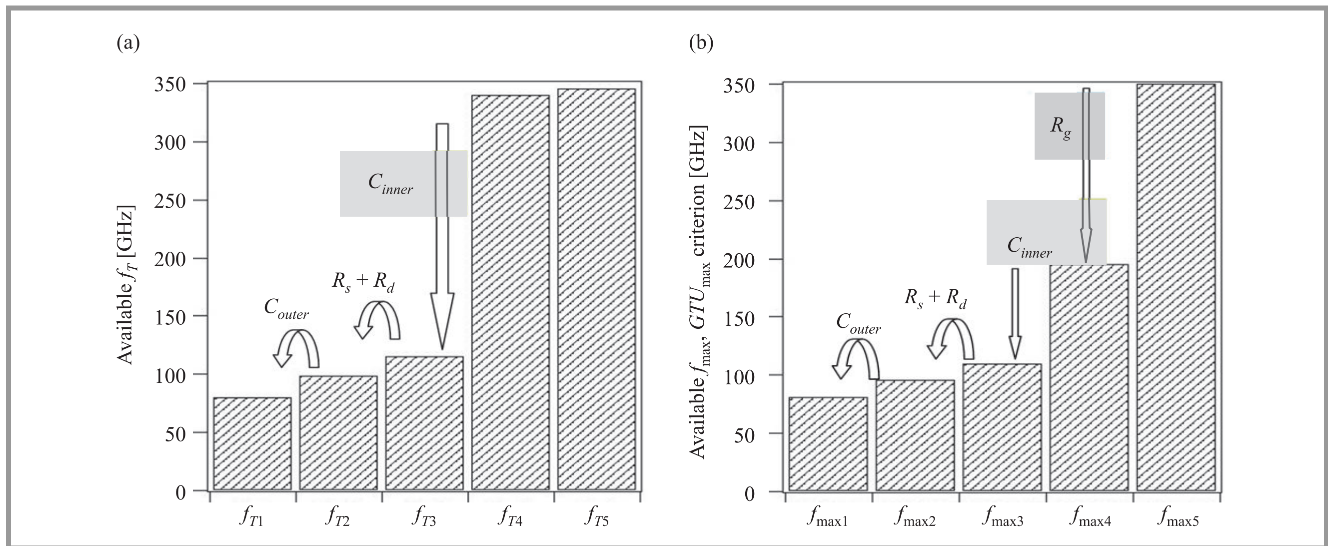


Fig. 10. Analysis of the relative impact of each lumped extrinsic parameters on (a) the current gain cutoff frequency (f_T) and on (b) the maximum available gain cutoff frequency (f_{max}) for a 60 nm long FinFET.

formed on the source and drain regions. After the heavily doped drain (HDD) implantations and rapid thermal annealing (RTA), NiSi is used as silicide and only one metal level is deposited.

The DC and RF analyses are performed on RF FinFETs (Fig. 7) composed of 50 gate fingers (N_{finger}) controling 6 fins (N_{fin}) each. As shown in Fig. 8(a) the 60 nm technology investigated here outlines a good control over SCE, with a subthreshold slope (S) close to 73.5 mV/dec. This value is even closer to ideal for $L_g = 120$ nm ($S =$ 62.9 mV/dec). Data in Fig. 8 are normalized by considering the total gate width $W_{tot} = N_{finger} N_{fin} (W_{fin} + 2H_{fin})$. No threshold voltage ($V_T$) roll off was observed with respect to $L_g (V_T \sim 260$ mV) and only small V_T variations (within 30 mV) are recorded as a function of W_{fin}. As expected, the devices also exhibit reduced SCE as the fin width is reduced. This is shown in Fig. 8(a), which indicates lower S values for narrower fins. However, reducing W_{fin} is also expected to increase the source (R_s) and drain (R_d) resistance [46], as shown in Fig. 8(b), which leads to a reduction of the normalized drain current as well as the effective gate transconductance (Fig. 8(c)).

The S-parameters of the devices are measured with a 110 GHz VNA from Agilent. An open-short de-embedding step is performed to remove the parasitics associated with the access pads. The current gain ($|H_{21}|$) as a function of frequency which yields the device transition frequency (f_T) is presented in Fig. 8(d) for FinFETs with different fin widths. Unfortunately, we can observe a reduction of the cutoff frequency with the shrinkage of W_{fin}. This degradation is mainly related to the increase of the source and drain resistances with the thinning down of the fin width (Fig. 8(b)).

The DC and RF performances of planar MOSFETs with similar dimensions (Fig. 7) have been measured for comparison purposes. Figure 9 presents the extracted RF cutoff frequencies of planar and FinFET devices as a function of channel length. The so-called intrinsic (f_{Ti}) and extrinsic (f_{Te}) cutoff frequencies stand, respectively, for the current gain cutoff frequency related to only the intrinsic lumped parameter elements (g_m, g_d, C_{gsi} and C_{gdi}) and the complete small-signal equivalent circuit presented in Fig. 2 (including the parasitic capacitances, C_{gse} and C_{gde}, as well as the accesss resistances R_s, R_d, and R_g). It is quite interesting to see that both devices present similar intrinsic cutoff frequencies (around 400 GHz for a channel length of 60 nm) but the extrinsic cutoff frequency, f_{Te}, of FinFET (90 GHz) is nearly twice lower than that of the planar MOSFET (180 GHz). A possible explanation for the latter effect might be the more relevant impact of extrinsic capacitances and resistances in the case of short gate length FinFETs.

Based on a wideband analysis, the lumped small-signal equivalent circuit parameters (Fig. 2) are extracted from the measured S-parameters according to the methods described in [47] and [48]. Figure 10 shows the relative impact of each parasitic parameter on the current gain (f_T in Fig. 10(a)) and maximum available power gain (f_{max}, Fig. 10(b)) cutoff frequencies of a 60 nm long FinFET. As expected from the expressions (1)–(3) and the published results for SG MOSFETs [49] the gate resistance has an important impact on f_{max} whereas f_T is unchanged. The sum of fringing capacitances C_{inner} directly linked to the FinFET three-dimensional (3D) architecture has a huge impact on both cutoff frequencies. In fact, f_T and f_{max} drop down, respectively, by a factor of 3 and 2. Finally, the source and drain resistances as well as the parasitic capacitances related to the feed connexions outside the active area of the transistor slightly decrease both cutoff frequencies. Based on that analysis, it is quite clear that the fringing capacitances inside the active area of the FinFET are the most important limiting factor for this type of non-planar multiple gate transistors.

In [50], Wu and Chan analyze the geometry-dependent parasitic components in multifin FinFETs. Parasitic fringing capacitance and overlap capacitance are physically modeled as functions of gate geometry parameters using the conformal mapping method. The relative contribution from each part of the 3D geometry of the FinFET is calculated. They subdivide the fringing capacitances in 3 distinct components noted C_1, C_2 and C_3 in Fig. 11(a). They demonstrate the importance of the fringing capacitance C_3 (Fig. 11(b)) which orginates from the capacitive coupling between the source and drain regions of the fins (side walls) and the gate electrode located between fins assuring the electrical connection between the gates wrapping the different fins connected in parallel through the source and drain contacts.

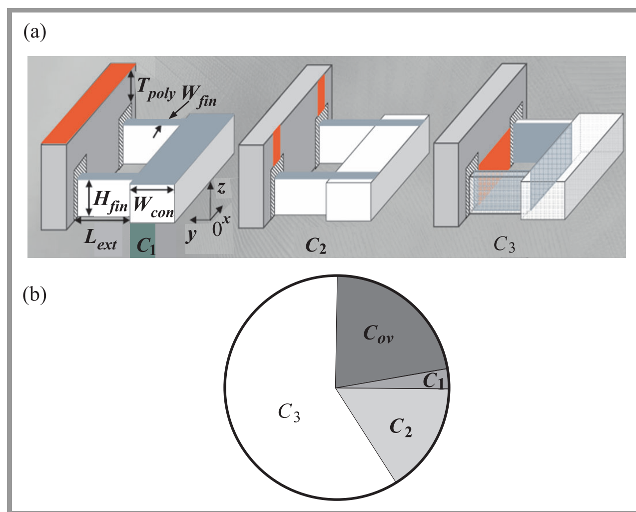


Fig. 11. (a) Three-dimensional schematic presentation of the various contributions of the fringing capacitances for a FinFET (C_1, C_2, C_3); (b) relative importance of each fringing capacitance (C_1, C_2, C_3) and overlap capacitance (C_{ov}).

In [51] and [52], the authors have demonstrated based on finite element numerical simulations the possibility to reduce C_{inner} and thus its impact on the FinFET cutoff frequencies by reducing the fin spacing (Fig. 12(a)) or by increasing the aspect ratio of the fin (higher H_{fin}/W_{fin} – see, Fig. 12(b)), respectively.

Figure 13 shows the extracted input capacitance ($C_{gg} = C_{gs} + C_{gd}$) in strong inversion ($V_g = 1.7$ V and $V_d = 0$ V) as a function of the active gate width (W_{tot}) for a FinFET and a SG MOSFET with 60 nm gate length. Both devices are built simultanuously on the same SOI wafer. A first order extrapolation of the measured data yields C_{gg} values of 1.33 fF/μm for the FinFET devices and only 1.09 fF/μm of active gate width for the SG, indicating a 20% increase of input capacitance in the case of FinFETs. Assuming that the normalized oxide capacitance is equal in both SG and FinFET devices, this increase is solely due to additional fringing in FinFETs. Using additional capacitance data measured in deep depletion, the extrinsic gate capacitance is actually found to be 40% higher for FinFETs. As explained above, this higher normalized input capacitance for FinFET can be explained by the fact that the gate fingers must run over non active area between each pair

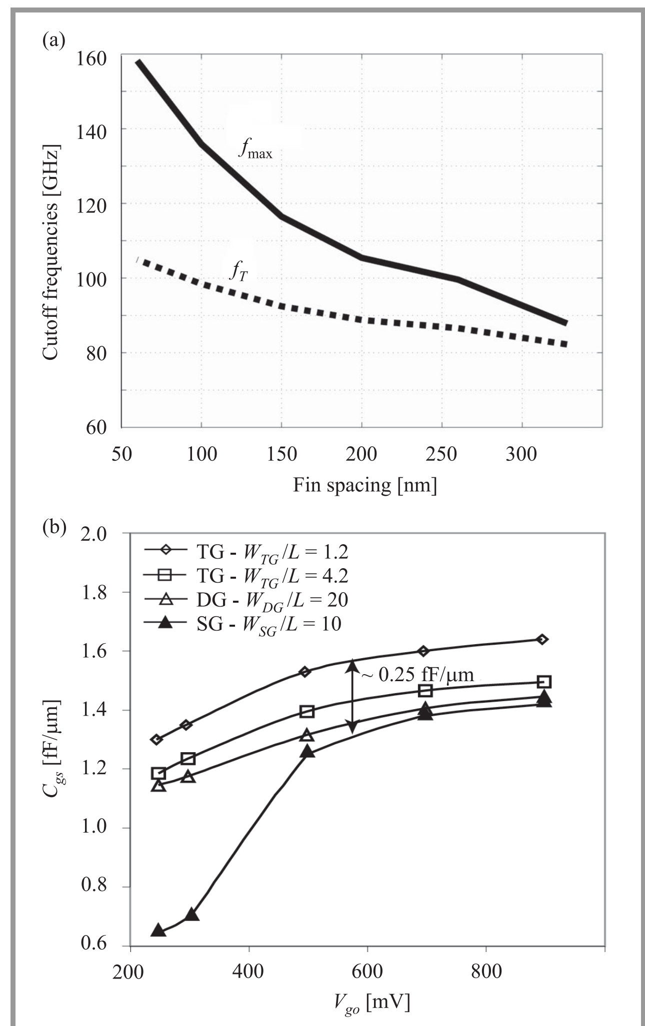


Fig. 12. (a) Cutoff frequencies of FinFETs versus fin spacing; (b) effect of W/L ratio on the normalized C_{gs} extracted at $V_d = 1$ V at various V_{go} and $L = 100$ nm.

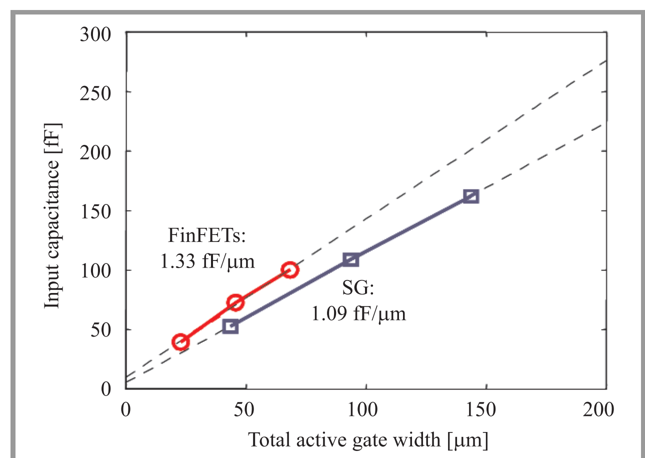


Fig. 13. Extracted input capacitance in strong inversion ($V_g = 1.7$ V and $V_d = 0$ V) as a function of W_{tot} for 60 nm SG MOSFET and 60 nm FinFET.

of parallel fins, a situation that is not encountered in SG MOSFETs.

To summarize, the simulation and experimental results indicate that FinFET is a multiple gate structure of interest to reduce digital short channel effects and then assure a lower threshold voltage roll-off, a better subthreshold slope and then higher I_{on}/I_{off} ratio, but the high frequency performance such as the cutoff frequencies as well as RF noise figure as presented in [53] are degraded compared to its SG MOSFET counterpart because of the increased fringing capacitance linked to its complex 3D non-planar architecture. Consequently, a trade-off exists regarding W_{fin} between high f_T and $f_{\max}$ (large W_{fin}) and good control of SCE (small W_{fin}).

5. High Resistivity SOI Substrate

5.1. Coplanar Waveguides Transmission Lines

The use of high resistivity silicon substrate is mandatory to reduce as much as possible the high frequency losses associated with the substrate conductivity. High resistivity silicon substrate cannot be introduced in the case of bulk Si MOSFETs due to the problem related to latch-up between devices. In SOI technology, thanks to the buried oxide the thin top silicon layer in which the transistors are implemented is electrically isolated from the Si substrate which can have high resistivity without impacting the good behavior of the MOS integrated circuits (ICs). Recently, high quality coplanar waveguides (CPW) presenting insertion loss of less than 2 dB/mm at 200 GHz as well as low- and high-pass filters at millimeter waves have been succefully built in an industrial SOI CMOS process environment [54].

The insertion loss of a CPW line lying on a lossy silicon substrate depends on the conductor loss (α_{cond}) and the substrate loss (α_{sub}) which is inversely proportional to the effective resistivity of the substrate. The effective resistivity represents the value of the substrate resistivity that is actually seen by the coplanar devices. This parameter accounts for the wafer inhomogeneities (i.e., oxide covering and space charge effects) and corresponds to the resistivity that a uniform (without oxide nor space charge effects) silicon wafer should have in order to sustain identical RF substrate losses. The effective resistivity is extracted from the measured S-parameters of the CPW line with a method depicted in [55].

Simulation results displayed in Fig. 14 outline how this parameter affects substrate and total losses for a 50 Ω CPW with 1 μm-thick Al line, the central conductor width of 40 μm and spacing between conductors of 24 μm. These data are obtained with analytical formulas presented in [56] and assuming metal conductivity of $3 \cdot 10^7$ S/m. It is seen that substrate losses (α_{sub}) are small (~ 0.1 dB/cm) when ρ_{eff} is close to 3 kΩcm and become clearly meaningless compared to conductor losses (α_{cond}) when ρ_{eff} reaches 10 kΩcm.

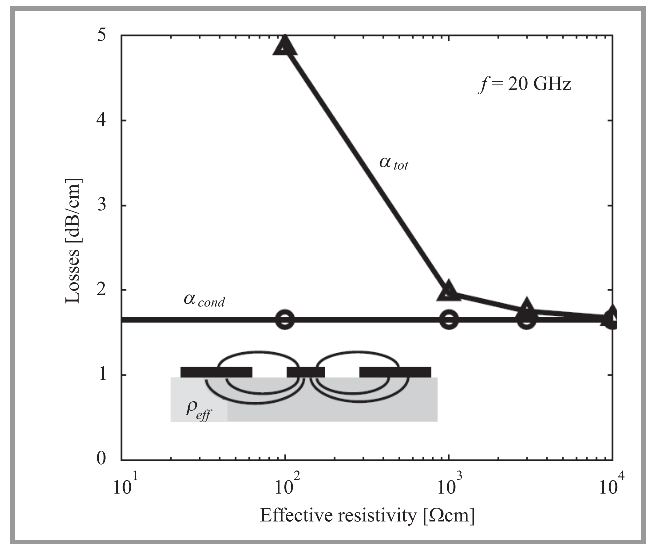


Fig. 14. Total (α_{tot}) and conductor (α_{cond}) losses as a function of ρ_{eff} at 20 GHz for a CPW line geometry according to [56].

Keeping substrate losses at low levels is a priority target when designing high performance integrated silicon systems. In this field, high resistivity (> 3 kΩcm) silicon wafers are foreseen as promising candidates for radio frequency integrated circuits [57] and mixed signal applications [58]. However, oxide passivated high resistivity (HR) wafers are known to suffer from parasitic surface conduction due to fixed charges (Q_{ox}) in the oxide [59]. Indeed, charges within the oxide attract free carriers near the substrate surface, reducing the effective resistivity (ρ_{eff}) seen by coplanar devices and increasing substrate losses. It has been recently shown in [60] that values as low as $Q_{ox} = 10^{10}/\mathrm{cm}^2$ could lower the value of resistivity by more than one order of magnitude in the case of 50 Ω CPW transmission line. The parasitic surface conduction can also be formed underneath metallic lines with the application of a DC bias (V_a) [61].

The extracted line loss and effective substrate resistivity as a function of the DC bias applied to the central conductor of a CPW line are, respectively, presented in Figs. 15(a) and 15(b) for different substrates, oxide layers and metallic lines as summarized in Table 1. Techno A and B are wafers coming from the industry while the three other wafers named C, D and E are home processes with one metal layer. In all cases, the metallic structures are patterned on either oxidized p-type HR unibond SOI (techno A, B, C) or oxidized p-type HR bulk Si (techno D and E) substrates.

The total RF losses (α_{tot}) of the CPW lines are extracted from the measured S-parameters with a thru-line-reflect method [62]. They are reported at 10 GHz in Fig. 15(a) as a function of V_a, where it is seen that α_{tot} may be significantly affected by V_a when the oxide thickness (t_{ox}) is in the several hundreds of nanometers (techno C). Indeed, in that case highly positive or negative biases have a large impact on the free carrier concentration below the oxide, thereby strongly affecting substrate losses. This effect is attenu-

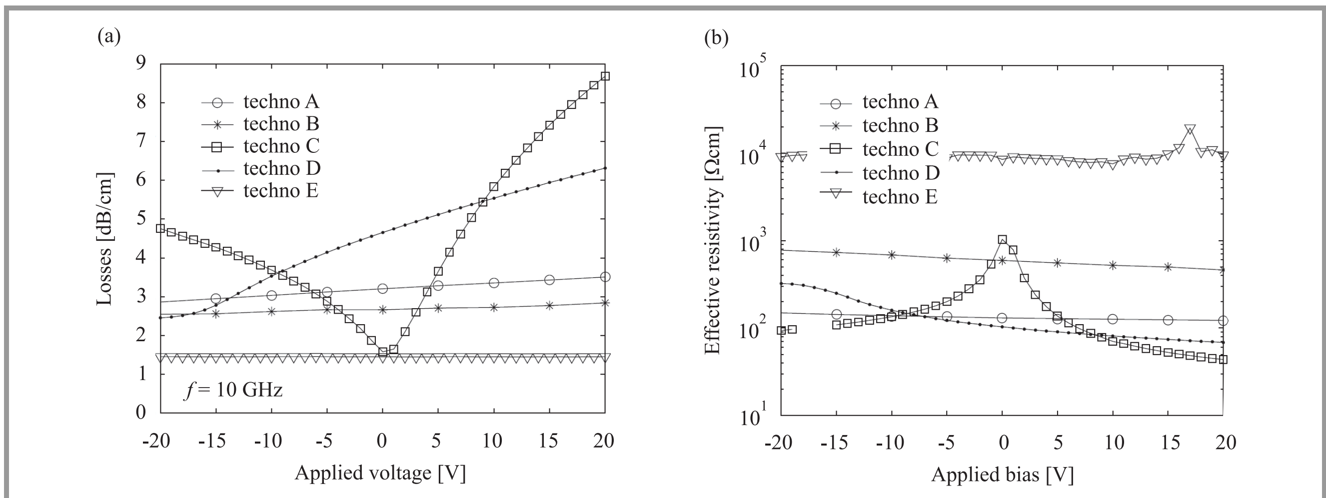


Fig. 15. (a) CPW losses and (b) effective substrate resistivity measured for different technologies described in Table 1 as a function of DC bias applied to the CPW central conductor.

Table 1
Additional information on the different technologies investigated in Fig. 15.

Techno	Starting wafer	Metal layers	Oxide thickness [μm]	Si passivation	Oxide type
A	HR SOI	M3	3	No	BOX + oxidized SOI + interlayer dielectrics
B	HR SOI	M5–M6	4.1	No	BOX + oxidized SOI + interlayer dielectrics
C	HR SOI	M1	0.3	No	BOX + oxidized SOI
D	HR Si bulk	M1	1	No	PECVD
E	HR Si bulk	M1	1	Polysilicon	PECVD
The data in columns 3 and 4, respectively, indicate the metal levels that were used and the total equivalent oxide thickness for CPW lines.					

ated for thicker oxides (techno A, B and D). The V_a value for which losses are minimum ($V_{a,\min}$) corresponds to the state of deep depletion underneath the oxide. As shown in Fig. 15, the V_a depends on the flatband voltage (V_{FB}) of the structure and is therefore dependent on t_{ox} as well as the oxide charge density (Q_{ox}).

The parasitic surface conduction can be reduced or even suppressed if the silicon substrate is passivated before oxidation with a trap-rich, highly resistive layer.

Figure 16 illustrates the impact of trap density (D_{it}) at the HR Si substrate/SiO_2 interface on the value of ρ_{eff} at 0 V for several Q_{ox} densities. It is seen with no surprise that the minimum D_{it} level that is required to obtain lossless substrates (i.e., $\rho_{eff} = 10$ kΩcm) is an increasing function of the fixed charge density in the oxide. This is because for higher positive densities, a higher concentration of electrons is attracted near the substrate surface and a higher density of traps is required to absorb those charges.

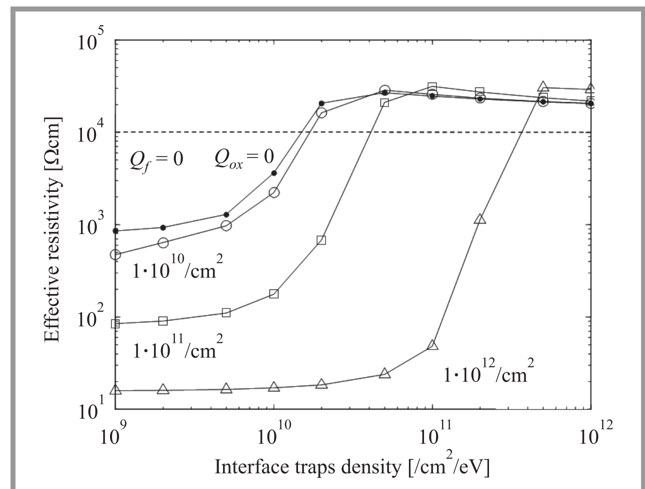


Fig. 16. Simulated effective resistivity values ρ_{eff} as a function of the trap density D_{it} for several fixed charges densities Q_{ox} and an applied bias value of 0 V.

The introduction of a high density of traps at the Si/SiO_2 interface has been successfully achieved using low-pressure chemical vapor-deposited (LPCVD) polysilicon (polySi) and amorphous silicon (α-Si) in [63] and [64], respectively. In the context of SOI technology, substrate passivation could also be an efficient technique to reduce substrate losses. To be compatible with a HR SOI wafer fabrication

process, the passivation layer should be included within the SOI structure by bonding an oxidized silicon wafer with a passivated HR substrate.

In [65], the proposed method consists in the LPCVD-deposition of amorphous silicon followed by Si-crystallization at 900°C with RTA. This method was compared with previously published techniques (passivation with amorphous silicon in [64] or LPCVD-polysilicon in [63] and was demonstrated to perform better in terms of substrate loss reduction: effective resistivity values higher than 10 kΩcm were reported, compared to 3 and 6 kΩcm in the case of amorphous Si and LPCVD polySi passivation, respectively. The new passivation method was also shown to present better rms surface roughness ($\sigma = 0.37$ nm) and to remain effective after long thermal anneals (4 hours at 900°C). A successful bonding of this layer with an oxidized substrate was achieved, showing that this new passivation technique could be introduced at reduced cost inside a smartcut or BESOI process in order to fabricate SOI wafers with enhanced resistivity, i.e., higher than 10 kΩcm.

Figure 15(a) indicates that substrate passivation with polysilicon (techno E) significantly reduces RF losses while getting rid of the V_a influence. This is because traps present inside the polySi layer can absorb free carriers and pin the surface potential to a value independent on V_a [63]. Figure 15(b) presents the effective resistivity (ρ_{eff}) extracted according to a method depicted in [55]. Not surprisingly, the highest ρ_{eff} value is observed for the passivated substrate, while at 0 V, the lowest value is obtained for the low quality (Q_{ox}-rich) PECVD oxide. It should also be noticed that due to the inverted layer underneath the BOX in techno A and B, the extracted values of ρ_{eff} do not exceed 130 and 580 Ωcm, respectively. These values are both more than one order of magnitude lower than the nominal substrate resistivity.

5.2. Crosstalk

In recent years, rapid progress of integrated circuit technology has enabled the co-integration of analog front-end and digital baseband processing circuits of communication systems onto the same chip. Such mixed-signal systems-on-chip (SoCs) allow more functionality, higher performance, lower power and higher reliability than non-integrated solutions, where at least two chips are needed, one for digital and one for the analog applications. Moreover, thanks to CMOS technology scaling and its associated increasing integration level, SoCs have become the way to achieve cost effectiveness for demanding applications such as home entertainment and graphics, mobile consumer devices, networking and storage equipment.

Such a rising integration level of mixed-signal ICs raises new issues for circuit designers. One of these issues is the substrate noise (Fig. 19(a)) generated by switching digital circuits, called digital substrate noise (DSN), which may degrade the behavior of adjacent analog circuits [66].

DSN issues become more and more important with IC evolution as

- digital parts get more noisy due to increasing complexity and clock frequencies;
- digital and analog parts get closer;
- analog parts get more sensitive because of V_{dd} scaling for power concern issues.

In general, substrate noise can be decomposed in three different mechanisms: noise generation, injection/propagation into the substrate and reception by the analog part [67]. Improvement in the reduction of any of these three mechanisms, or in all of them, will lead to a reduction of the DSN and in a relaxation of the design requirements. Typically, guard rings and overdesigned structures are adopted to limit the effect of substrate noise, thereby reducing the advantages of the introduction of new technologies. It is thus a major issue for the semiconductor industry to find area-efficient design/technology solutions to reduce the impact of substrate noise in mixed-signal ICs.

This last decade several publications have demonstrated theoretically and experimentally the interest of high resistivity SOI substrate to greatly reduce the crosstalk level between integrated circuits [5]. Figure 17 shows how the crosstalk between two 50 μm spaced metallic pads is affected by ρ_{eff} and indicates that ρ_{eff} must be at least in the several kΩcm range to get rid of conductive coupling inside the substrate for frequencies around 100 MHz and lower.

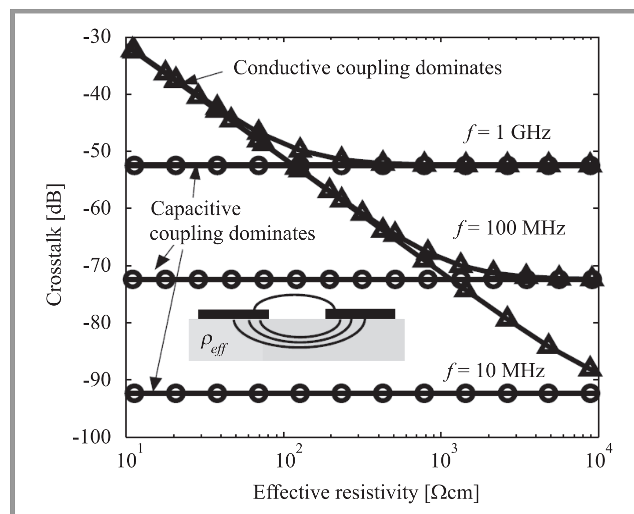


Fig. 17. Simulated crosstalk level at 10 MHz, 100 MHz and 1 GHz as a function of ρ_{eff} according to model presented in [5].

The result of the substrate crosstalk measurements using a classical double-pad structure in which both pads are connected to separate RF probing pads [5] is shown in Fig. 18 in the form of $|S_{21}|$ versus frequency curves. The measurements are performed by using the low-frequency VNA up to 4 GHz and by applying various bias conditions on the coupling pads. The figure shows significantly

higher (∼ 13 dB at 0 V) crosstalk level below 1 GHz for the standard HR SOI wafer, due to conductive effects in the substrate associated with parasitic surface conduction [68]. It also highlights a significant dependence with respect to the applied bias. The crosstalk level is strongly reduced for negative bias and when deep depletion is formed below the BOX, whereas it is enhanced and exhibits higher cutoff frequencies for positive bias and increased inversion below the oxide. On the other hand, the passivated wafer exhibits:

– no effect of the applied bias due to the presence of the trap-rich polysilicon layer below the BOX [55];

– a perfect 20 dB/dec slope which shows that purely capacitive coupling occurs in the measurable frequency range (i.e., above the noise floor of the VNA).

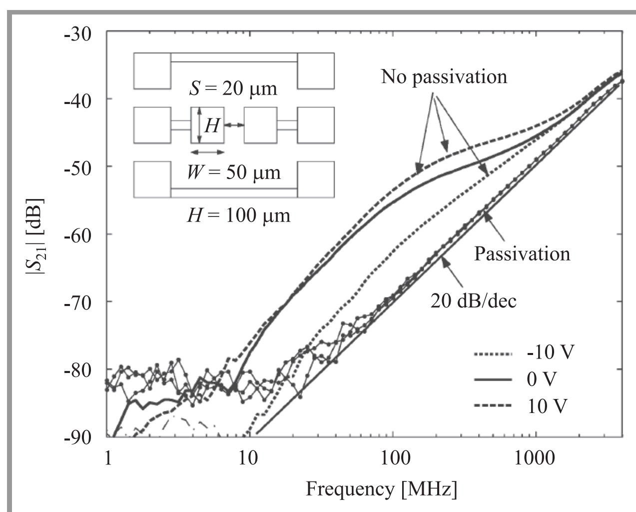


Fig. 18. Crosstalk measured as a function of frequency and under distinct bias conditions on the unpassivated and passivated HR Si wafers.

A reduction of crosstalk below 1 GHz is of particular interest for mixed signal applications, since it is known from previous studies that the frequency spectrum of the noise generated by digital logic typically expands to several hundreds of megahertz, corresponding to multiples of the clock signal [69], [70] or circuit internal resonance frequencies [71]. The generation of noise in that frequency range has also been shown to strongly increase the jitter in phase-locked loops (PLLs), which seem to be particularly sensitive to substrate noise injected at the PLL reference frequency, i.e., in the few hundreds of megahertz range [72]. It is further believed that in terms of crosstalk, the benefits gained by substrate passivation will even increase in the future. Indeed, a reduction of the BOX thickness for the next generations of active SOI devices will be required to reduce short channel effects and self-heating [73].

In [74], we compare experimental DSN characterizations of CMOS circuits lying on SOI and bulk Si substrates. Current injected into the substrate creates substrate voltage fluctuations (substrate noise). It is mainly created by two mechanisms [66]: coupling from the noisy digital power supply circuit and from switching drains.

The DSN for 8 switching inverter trees biased at either 0.8 or 1.2 V and for an input clock frequency of 225 MHz has been measured in the case SOI and Si bulk substrates. DSN for SOI circuit presents a quite different frequency response (Fig. 19(b)). At low frequency, SOI and Si bulk present the same kind of response, with the SOI DSN level decreasing faster with increasing frequency. At higher frequency, the SOI DSN presents a kind of "pass-band filter" shape, which is not visible in the case of the bulk circuit.

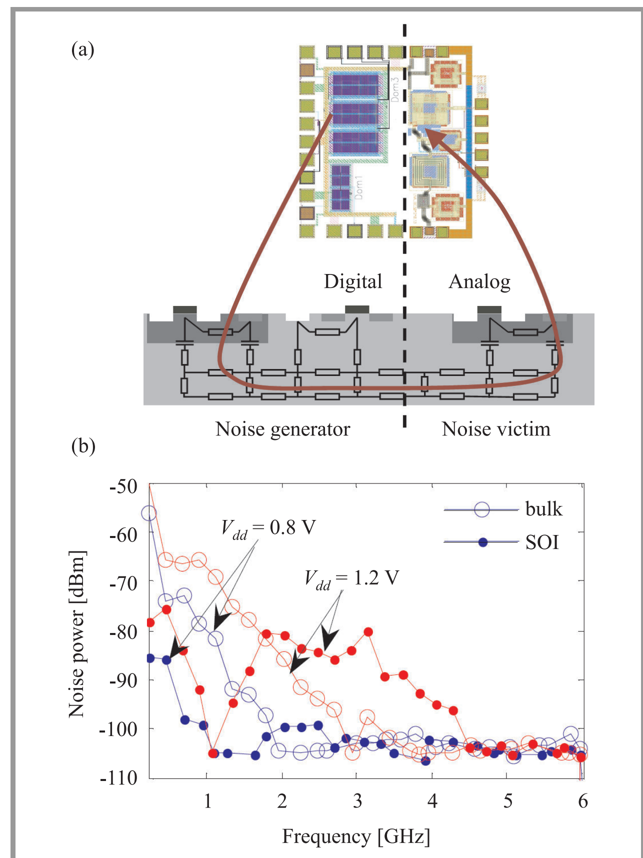


Fig. 19. (a) Schematic representation of the substrate crosstalk between digital and analog parts of a SoC; (b) comparison of the frequency envelope of the measured DSN (clock frequency = 225 MHz, 8 inverter trees).

We have shown in our previous work that this second part of the frequency response is due to ringing on supply rail, due to parasitic capacitances and inductances [75]. For the 1.2 V supply voltage, the SOI technology allows an important reduction of DSN up to 1 GHz. At higher frequency, the noise due to ringing on supply rail becomes dominant, and the bulk circuit shows a lower DSN level. This conclusion is in agreement with the results of studies on the supply noise showing that special attention should be paid to supply rail for SOI technology, due to lower intrinsic decoupling capacitances [76]. At lower power supply (0.8 V), as for the bulk, high frequency noise generation decreases. The ringing supply noise tends thus to be negli-

gible. The SOI technology presents then better DSN results than bulk for frequency up to 2 GHz, and similar DSN level for upper frequency.

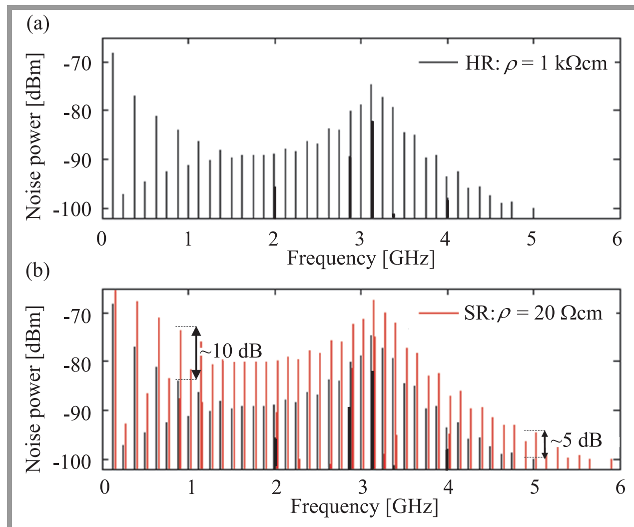


Fig. 20. Frequency spectrum of the measured DSN in (a) standard and (b) high resistivity SOI substrate (clock frequency = 225 MHz, V_{dd} = 1.2 V, 8 inverter trees).

Figure 20 shows the reduction of the DSN thanks to the use of HR SOI substrate compared to standard resistivity SOI. The decrease of the DSN should be even more pronounced if a passivation layer (trap-rich layer) is introduced underneatch the BOX.

5.3. Nonlinearities Along CPW Lines

High-resistivity silicon substrates are promising for RF applications due to their reduced substrate loss and coupling, as presented in the two previous subsections, which helps to enable RF cellular transmit switches on SOI using HRS handle wafers [77], [78]. RF switches have high linearity requirements: for instance, a recent III-V RF switch product specifies less than −45 and −40 dBm for 2nd and 3rd harmonic power (H2 and H3), respectively, at +35 dBm input power [79]. As requirements become even more stringent for advanced multimode phones and 3G standards, it is important to investigate even small contributions to harmonic distortion (HD).

As explained above, when the CPW line is biased the distribution of potential and free carriers inside the Si substrate changes like in the case of a classical MOS capacitor. The variation of carriers distribution in the Si substrate with the applied bias or large RF signal will thus lead to the existence of nonlinear capacitance (C) and conductance (G) associated with the Si substrate. Those variable C and G are at the origin of the harmonics formation inside the Si substrate.

Figure 21 shows the harmonic distortion of Al metal lines on thermally oxidized HRS p-type substrates of different resistivities. The 1 kΩcm substrate presents lower HD than the 20 Ωcm substrate over most of the power sweep.

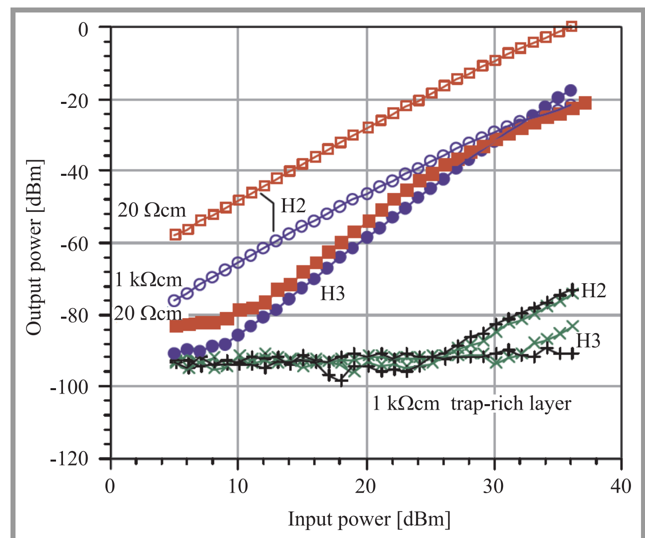


Fig. 21. Measured harmonic distortion for low- and high-resistivity silicon substrates, and high-resistivity silicon substrates with trap-rich layers. CPW metal is aluminum on 60 nm of oxide with length of 2.1 mm. The trap-rich layer significantly reduces HD.

A drastic drop of the HD is observed when the HR Si substrate is passivated with a trap-rich layer (as-deposited amorphous silicon), that is, by at least 50 and 65 dB in H2 and H3, respectively, or to the noise floor. As explained above, thanks to the high density of traps in the polycrystalline silicon or as-deposited amorphous silicon layer located at the Si-SiO_2 interface, the surface potential at this interface is nearly fixed, and the external DC bias or large amplitude RF signal applied to the line does not impact the distribution of carriers inside the Si substrate.

6. Conclusions

The performance of SOI MOSFET technology in microwaves and millimeter waves has been presented. Nowadays, strained SOI N-MOSFET which exhibits a cutoff frequency close to 500 GHz is really competing with the III-V technologies. Thanks to the introduction of high resistivity SOI substrate, the integration of high quality passives is a reality and the reduction of the substrate crosstalk is a real advantage compared to Si bulk for the development of high integration low voltage mixed-mode applications. Major semiconductor companies such as IBM, RFMD, Honeywell, OKI, etc., have already produced several products for the telecommunication market based on SOI RF technologies.

As demonstrated in the present paper, by the introduction of a trap-rich layer underneath the BOX, HR SOI substrate can still be improved. Having a polysilicon-based layer with the thickness of approximately 300 nm sandwiched between the BOX and the HT Si substrate, CPW insertion loss, crosstalk, DSN, as well as harmonic distortion are greatly reduced.

To summarize, present and future HR SOI MOSFET technologies are very good candidatea for mixed-mode low voltage low power RF and even millimeter waves applications.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits", *Electronics*, vol. 38, pp. 114–117, 1965.

[2] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions", *IEEE J. Solid-State Circ.*, vol. SC-9, pp. 256–268, 1974.

[3] S. Lee, B. Jagannathan, S. Narasimha, A. Chou, N. Zamdmer, J. Johnson, R. Williams, L. Wagner, J. Kim, J.-O. Plouchart, J. Pekarik, S. Springer, and G. Freeman, "Record RF performance of 45-nm SOI CMOS technology", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 2007*, Washington, USA, 2007, pp. 255–258.

[4] T. Sakurai, A. Matsuzawa, and T. Douseki, *Fully-Depleted SOI CMOS Circuits and Technology for Ultralow-Power Applications*. New Jersey: Springer, 2006.

[5] J.-P. Raskin, A. Viviani, D. Flandre, and J.-P. Colinge, "Substrate crosstalk reduction using SOI technology", *IEEE Trans. Electron Dev.*, vol. 44, no. 12, pp. 2252–2261, 1997.

[6] H. F. Cooke, "Microwave transistors: theory and design", *Proc. IEEE*, vol. 59, pp. 1163–1181, 1971.

[7] C. A. Mead, "Schottky barrier gate field effect transistor", *Proc. IEEE*, vol. 59, pp. 307–308, 1966.

[8] 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.

[9] 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.

[10] 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, no. 7, pp. 230–232, 1995.

[11] 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", *IEEE Trans. Electron Dev.*, vol. 48, pp. 2606–2624, 2001.

[12] R. Lai, X. B. Mei, W. R. Deal, W. Yoshida, Y. M. Kim, P. H. Liu, J. Lee, J. Uyeda, V. Radisic, M. Lange, T. Gaier, L. Samoska, and A. Fung, "Sub 50 nm InP HEMT device with f_{max} greater than 1 THz", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 2007*, Washington, USA, 2007, pp. 609–611.

[13] H. S. Momose, E. Morifuji, T. Yoshitomi, T. Ohguro, I. Saito, T. Morimoto, Y. Katsumata, and H. Iwai, "High-frequency AC characteristics of 1.5 nm gate oxide MOSFETs", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 1996*, San Francisco, USA, 1996, pp. 105–108.

[14] "International Technology Roadmap for Semiconductors", 2006, [Online]. Available: http://www.itrs.net/Common/2006ITRS/Home2006.html

[15] G. Dambrine, C. Raynaud, D. Lederer, M. Dehan, O. Rozeaux, M. Vanmackelberg, F. Danneville, S. Lepilliet, and J.-P. Raskin, "What are the limiting parameters of deep-submicron MOSFETs for high frequency applications?", *IEEE Electron Dev. Lett.*, vol. 24, no. 3, pp. 189–191, 2003.

[16] G. Pailloncy, C. Raynaud, M. Vanmackelberg, F. Danneville, S. Lepilliet, J.-P. Raskin, and G. Dambrine, "Impact of down scaling on high frequency noise performance of bulk and SOI MOSFETs", *IEEE Trans. Electron Dev.*, vol. 51, no. 10, pp. 1605–1612, 2004.

[17] V. Kilchytska, A. Nève, L. Vancaillie, D. Levacq, S. Adriaensen, H. van Meer, K. De Mayer, C. Raynaud, M. Dehan, J.-P. Raskin, and D. Flandre, "Influence of device engineering on the analog and RF performances of SOI MOSFETs", *IEEE Trans. Electron Dev.*, vol. 50, no. 3, pp. 577–588, 2003.

[18] M. Vanmackelberg, C. Raynaud, O. Faynot, J.-L. Pelloie, C. Tabone, A. Grouillet, F. Martin, G. Dambrine, L. Picheta, E. Mackowiak, P. Llinares, J. Sevenhans, E. Compagne, G. Fletcher, D. Flandre, V. Dessard, D. Vanhoenacker, and J.-P. Raskin, "0.25 μm fully-depleted SOI MOSFET's for RF mixed analog-digital circuits, including a comparison with partially-depleted devices for high frequency noise parameters", *Solid-State Electron.*, vol. 46, iss. 3, pp. 379–386, 2002.

[19] S. Burignat, D. Flandre, V. Kilchytska, F. Andrieux, O. Faynot, and J.-P. Raskin, "Substrate impact on sub-32 nm ultra thin SOI MOSFETs with thin buried oxide", in *Proc EUROSOI 2009, Fifth Worksh. Them. Netw. Sil. Insul. Technol. Dev. Circ.*, Göteborg, Sweden, 2009, pp. 27–28.

[20] T. Rudenko, V. Kilchytska, S. Burignat, J.-P. Raskin, F. Andrieu, O. Faynot, A. Nazarov, and D. Flandre, "Transconductance and mobility behaviors in UTB SOI MOSFETs with standard and thin BOX", in *Proc. EUROSOI 2009, Fifth Worksh. Them. Netw. Sil. Insul. Technol. Dev. Circ.*, Göteborg, Sweden, 2009, pp. 111–112.

[21] K.-W. Ang, J. Lin, C.-H. Tung, N. Balasubramanian, G. S. Samudra, and Y.-C. Yeo, "Strained n-MOSFET with embedded source/drain stressors and strain-transfer structure (STS) for enhanced transistor performance", *IEEE Trans. Electron Dev.*, vol. 55, no. 3, pp. 850–857, 2008.

[22] G. Néau, F. Martinez, M. Valenza, J. C. Vildeuil, E. Vincent, F. Boeuf, F. Payet, and K. Rochereau, "Impact of strained-channe n-MOSFETs with a SiGe virtual substrate on dielectric interface quality evaluated by low frequency noise measurements", *Microelectron. Reliab.*, vol. 47, pp. 567–572, 2007.

[23] S. H. Olsen, E. Escobedo-Cousin, J. B. Varzgar, R. Agaiby, J. Seger, P. Dobrosz, S. Chattopadhyay, S. J. Bull, A. G. O'Neill, P.-E. Hellstrom, J. Edholm, M. Ostling, K. L. Lyutovich, M. Oehme, and E. Kasper, "Control of self-heating in thin virtual substrate strained Si MOSFETs", *IEEE Trans. Electron Dev.*, vol. 53, no. 9, pp. 2296–2305, 2006.

[24] J. M. Larson and J. Snyder, "Overview and status of metal S/D Schottky barrier MOSFET technology", *IEEE Trans. Electron Dev.*, vol. 53, no. 5, pp. 1048–1058, 2006.

[25] D. J. Pearman, G. Pailloncy, J.-P. Raskin, J. M. Larson, and T. E. Whall, "Static and high-frequency behavior and performance of Schottky barrier p-MOSFET devices", *IEEE Trans. Electron Dev.*, vol. 54, no. 10, pp. 2796–2802, 2007.

[26] J.-P. Raskin, D. J. Pearman, G. Pailloncy, J. M. Larson, J. Snyder, D. L. Leadley, and T. E. Whall, "High-frequency performance of Schottky barrier p-MOSFET devices", *IEEE Electron Dev. Lett.*, vol. 29, no. 4, pp. 396–398, 2008.

[27] G. Larrieu, E. Dubois, R. Valentin, N. Breil, F. Danneville, G. Dambrine, J.-P. Raskin, and J.-C. Pesant, "Low temperature implementation of dopant-segregated band-edge metallic S/D junctions in thin-body SOI p-MOSFETs", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 2007*, Washington, USA, 2007, pp. 147–150.

[28] R. Valentin, E. Dubois, J.-P. Raskin, G. Larrieu, G. Dambrine, T. C. Lim, N. Breil, and F. Danneville, "RF small signal analysis of Schottky-barrier p-MOSFET", *IEEE Trans. Electron Dev.*, vol. 55, no. 5, pp. 1192–1202, 2008.

[29] B. Ricco, R. Versari, and D. Esseni, "Characterization of polysilicon-gate depletion in MOS structures", *IEEE Electron Dev. Lett.*, vol. 17, no. 3, pp. 103–105, 1996.

[30] A. Vandooren, A. V. Y. Thean, Y. Du, I. To, J. Hughes, T. Stephens, M. Huang, S. Egley, M. Zavala, K. Sphabmixay, A. Barr, T. White, S. Samavedam, L. Mathew, J. Schaeffer, D. Triyoso, M. Rossow, D. Roan, D. Pham, R. Rai, B.-Y. Nguyen, B. White, M. Orlowski, A. Duvallet, T. Dao, and J. Mogab, "Mixed-signal performance of sub-100 nm fully-depleted SOI devices with metal gate, high K (HfO_2) dielectric and elevated source/drain extensions", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 2003*, Washington, USA, 2003, pp. 11.5.1–11.5.3.

[31] C. H. Ko, T. M. Kuan, K. Zhang, G. Tsai, S. M. Seutter, C. H. Wu, T. J. Wang, C. N. Ye, H. W. Chen, C. H. Ge, K. H. Wu, and W. C. Lee, "A novel CVD-SiBCN low-K spacer technology for high-speed applications", in *Proc. Int. Symp. VLSI Technol. 2008*, Honolulu, Hawaii, USA, 2008, pp. 108–109.

[32] T. I. Bao, H. C. Chen, C. J. Lee, H. H. Lu, S. L. Shue, and C. H. Yu, "Low capacitance approaches for 22 nm generation Cu interconnect", in *Proc. Int. Symp. VLSI Technol. Syst. Appl. VLSI-TSA 2009*, Hsinchu, Taiwan, 2009, pp. 51–56.

[33] T. Ernst, C. Tinella, C. Raynaud, and S. Cristoloveanu, "Fringing fields in sub-0.1 μm fully depleted SOI MOSFET's: optimization of the device architecture", *Solid-State Electron.*, vol. 46, pp. 373–378, 2002.

[34] M. Fujiwara *et al.*, "Impact of BOX scaling on 30 nm gate length FD SOI MOSFET", *IEEE Int. SOI Conf.*, Honolulu, Hawaii, USA, 2005, pp. 180–182.

[35] F. Gianesello, D. Gloria, C. Raynaud, S. Montusclat, S. Boret, C. Clement, P. Benech, J. M. Fournier, and G. Dambrine, "State of the art 200 GHz passive components and circuits integrated in advanced thin SOI CMOS technology on high resistivity substrate", in *Proc. IEEE Int. SOI Conf.*, Niagara Falls, USA, 2006, pp. 121–122.

[36] F. Gianesello, D. Gloria, C. Raynaud, S. Montusclat, S. Boret, and P. Touret, "On the design of high performance RF integrated inductors on high resistively thin film 65 nm SOI CMOS technology", in *Proc. IEEE 8th Top. Meet. Sil. Monolit. Integr. Circ. RF Syst. SiRF 2008*, Orlando, USA, 2008, pp. 98–101.

[37] I. Post, M. Akbar, G. Curello, S. Gannavaram, W. Hafez, U. Jalan, K. Komeyii, J. Lin, N. Lindert, J. Park, J. Rizk, G. Sacks, C. Tsai, D. Yeh, P. Bai, and C.-H. Jan, "A 65 m CMOS SOC technology featuring strained silicon transistors for RF applications", in *Proc. Int. Electron Dev. Meet. IEDM 2006*, San Francisco, USA, 2006, pp. 1–3.

[38] J.-P. Colinge, M.-H. Gao, A. Romano, H. Maes, and C. Claeys, "Silicon-on-insulator "gate-all-around" MOS device", in *Proc. IEEE SOS/SOI Tech. Conf.*, Key West, USA, 1990, pp. 137–138.

[39] D. Hisamoto *et al.*, "FinFET – a self-aligned double-gate MOSFET scalable to 20 nm", *IEEE Trans. Electron Dev.*, vol. 47, no. 12, pp. 2320–2325, 2000.

[40] S. Cristoloveanu, "Silicon on insulator technologies and devices: from present to future", *Solid-State Electron.*, vol. 45, no. 8, pp. 1403–1411, 2001.

[41] J.-T. Park and J.-P. Colinge, "Multiple-gate SOI MOSFETs: device design guidelines", *IEEE Trans. Electron Dev.*, vol. 49, no. 12, pp. 2222–2229, 2002.

[42] J. Kedzierski *et al.*, "High performance symmetric-gate and CMOS-compatible V_t asymmetric-gate FinFET devices", in *Proc. IEEE Int. Electron Dev. Meet. IEDM 2001*, Washington, USA, 2001, pp. 437–440.

[43] D. Woo *et al.*, "Electrical characteristics of FinFET with vertically nonuniform source/drain profile", *IEEE Trans. Nanotech.*, vol. 1, no. 4, pp. 233–237, 2002.

[44] V. Kilchytska, N. Collaert, R. Rooyackers, D. Lederer, J.-P. Raskin, and D. Flandre, "Perspective of FinFETs for analog applications", in *Proc. 34th Eur. Solid-State Dev. Res. Conf. ESSDERC 2004*, Leuven, Belgium, 2004, pp. 65–68.

[45] D. Lederer *et al.*, "FinFet analogue characterization from DC to 110 GHz", *Solild-State Electron.*, vol. 49, pp. 1488–1496, 2005.

[46] A. Dixit *et al.*, "Analysis of the parasitic source/drain resistance in multiple gate field effect transistors", *IEEE Trans. Electron Dev.*, vol. 52, no. 6, pp. 1131–1140, 2005.

[47] J. P. Raskin *et al.*, "Accurate MOSFET characterization at microwaver frequencies for device optimization and analog modeling", *IEEE Trans. Electron Dev.*, vol. 45, pp. 1017–1025, 1998.

[48] A. Bracale *et al.*, "A new approach for SOI device small-signal parameter extraction", *Analog Integr. Circ. Sig. Process.*, vol. 25, pp. 159–167, 2000.

[49] B. Razavi, R.-H. Yan, and K. F. Lee, "Impact of distributed gate resistance on the performance of MOS devices", *IEEE Trans. Circ. Syst. I: Fund. Theory Appl.*, vol. 41, no. 11, pp. 750–754, 1994.

[50] W. Wu and M. Chan, "Analysis of geometry-dependent parasitics in multifin double-gate FinFETs", *IEEE Trans. Electron Dev.*, vol. 54, no. 4, pp. 692–698, 2007.

[51] O. Moldovan, D. Lederer, B. Iniguez, and J.-P. Raskin, "Finite element simulations of parasitic capacitances related to multiple-gate field-effect transistors architectures", in *Proc. 8th Top. Meet. Sil. Monolit. Integr. Circ. RF Syst. SiRF 2008*, Orlando, USA, 2008, pp. 183–186.

[52] J.-P. Raskin, T. M. Chung, V. Kilchytska, D. Lederer, and D. Flandre, "Analog/RF performance of multiple-gate SOI devices: wideband simulations and characterization", *IEEE Trans. Electron Dev.*, vol. 53, no. 5, pp. 1088–1094, 2006.

[53] J.-P. Raskin, G. Pailloncy, D. Lederer, F. Danneville, G. Dambrine, S. Decoutere, A. Mercha, and B. Parvais, "High frequency noise performance of 60 nm gate length FinFETs", *IEEE Trans. Electron Dev.*, vol. 55, no. 10, pp. 2718–2727, 2008.

[54] F. Gianesello *et al.*, "1.8 dB insertion loss 200 GHz CPW band pass filter integrated in HR SOI CMOS technology", in *Proc. Conf. IEEE MTT-S*, Honolulu, Hawaii, USA, 2007.

[55] D. Lederer and J.-P. Raskin, "Effective resistivity of fully-processed high resistivity wafers", *Solid-State Electron.*, vol. 49, pp. 491–496, 2005.

[56] W. Heinrich, "Quasi-TEM description of MMIC coplanar lines including conductor-loss effects", *IEEE Trans. Microw. Theory Tech.*, vol. 41, no. 1, pp. 45–52, 1993.

[57] A. C. Reyes, S. M. El-Ghazaly, S. J. Dom, M. Dydyk, D. K. Schroeder, and H. Patterson, "Coplanar waveguides and microwave inductors on silicon substrates", *IEEE Trans. Microw. Theory Tech.*, vol. 43, no. 9, pp. 2016–2021, 1995.

[58] K. Benaissa, J.-T. Yuan, D. Crenshaw, B. Williams, S. Sridhar, J. Ai, G. Boselli, S. Zhao, S. Tang, S. Ashbun, P. Madhani, T. Blythe, N. Mahalingam, and H. Schichijo, "RF CMOS high-resistivity substrates for systems-on-chip applications", *IEEE Trans. Electron Dev.*, vol. 50, no. 3, pp. 567–576, 2003.

[59] Y. Wu, H. S. Gamble, B. M. Armstrong, V. F. Fusco, and J. A. C. Stewart, "SiO_2 interface layer effects on microwave loss of high-resistivity CPW line", *IEEE Microw. Guid. Wave Lett.*, vol. 9, no. 1, pp. 10–12, 1999.

[60] D. Lederer, C. Desrumeaux, F. Brunier, and J.-P. Raskin, "High resistivity SOI substrates: how high should we go?", in *Proc. IEEE Int. SOI Conf.*, Newport Beach, USA, 2003, pp. 50–51.

[61] C. Schollhorn, W. Zhao, M. Morschbach, and E. Kasper, "Attenuation mechanisms of aluminum millimeter-wave coplanar waveguides on silicon", *IEEE Trans. Electron Dev.*, vol. 50, no. 3, pp. 740–746, 2003.

[62] H.-C. Lu and T.-H. Chu, "The thru-line-symmetry (TLS) calibration method for on-wafer scattering matrix measurement of four-port networks", in *Proc. IEEE MTT-S Int. Microw. Symp. Dig.*, Ford Worth, USA, 2004, vol. 3, pp. 1801–1804.

[63] H. Gamble, B. M. Armstrong, S. J. N. Mitchell, Y. Wu, V. F. Fusco, and J. A. C. Stewart, "Low-loss CPW lines on surface stabilized high resistivity silicon", *IEEE Microw. Guid. Wave Lett.*, vol. 9, no. 10, pp. 395–397, 1999.

[64] B. Wong, J. N. Burghartz, L. K. Natives, B. Rejaei, and M. van der Zwan, "Surface-passivated high resistivity silicon substrates tor RFICs", *IEEE Electron Dev. Lett.*, vol. 25, no. 4, pp. 176–178, 2004.

[65] D. Lederer and J.-P. Raskin, "New substrate passivation method dedicated to high resistivity SOI wafer fabrication with increase substrate resistivity", *IEEE Electron Dev. Lett.*, vol. 26, no. 11, pp. 805–807, 2005.

[66] F. Calmon, C. Andrei, O. Valorge, J.-C. Nunez Perez, J. Verdier, and C. Gontrand, "Impact of low-frequency substrate disturbances on a 4.5 GHz VCO", *Microelectron. J.*, vol. 37, no. 1, pp. 1119–1127, 2006.

[67] M. van Heijningen, M. Badaroglu, S. Donnay, M. Engels, and I. Bolsen, "High-level simulation of substrate noise generation including power supply noise coupling", in *Proc. 37th Conf. Des. Automat. DAC 2000*, Los Angeles, USA, 2000, pp. 446–451.

[68] D. Lederer and J.-P. Raskin, "Bias effects on RF passive structures in HR Si substrates", in *Proc. 6th Top. Meet. Sil. Microw. Integr. Circ. RF Syst.*, San Diego, USA, 2006, pp. 8–11.

[69] M. van Heijningen, J. Compiet, P. Wambacq, S. Donnay, M. G. E. Engels, and I. Bolsens, "Analysis and experimental verification of digital substrate noise generation for epi-type substrates", *IEEE J. Solid-State Circ.*, vol. 35, no. 7, pp. 1002–1008, 2000.

[70] M. van Heijningen, M. Badaroglu, S. Donnay, G. G. E. Gielen, and H. J. De Man, "Substrate noise generation in complex digital systems: efficient modeling and simulation methodology and experimental verification", *IEEE J. Solid-State Circ.*, vol. 37, no. 8, pp. 1065–1072, 2002.

[71] M. Badaroglu, S. Donnay, H. J. De Man, Y. A. Zinzius, G. G. E. Gielen, W. Sansen, T. Fonden, and S. Signell, "Modeling and experimental verification of substrate noise generation in a 220-k gates WLAN system-on-chip with multiple supplies", *IEEE J. Solid-State Circ.*, vol. 38, no. 7, pp. 1250–1260, 2003.

[72] K. A. Jenkins, W. Rhee, J. Liobe, and H. Ainspan, "Experimental analysis of the effect of substrate noise on PLL", in *Proc. 6th Top. Meet. Sil. Monolit. Integr. Circ. RF Syst.*, San Diego, USA, 2006, pp. 54–57.

[73] "International Technology Roadmap for Semiconductors: Front end processes", 2005 [Online]. Available: http://www.itrs.net/Common/2005ITRS/FEP2005.pdf

[74] C. Roda Neve, D. Bol, R. Ambroise, D. Flandre, and J.-P. Raskin, "Comparison of digital substrate noise in SOI and bulk Si CMOS technologies", in *Proc. 7th Worksh. Low-Volt. Low Power Des.*, Louvain-la-Neuve, Belgium, 2008, pp. 23–28.

[75] D. Bol, R. Ambroise, C. Roda Neve, J.-P. Raskin, and D. Flandre, "Wide-band simulation and characterization of digital substrate noise in SOI technology", in *Proc. IEEE Int. SOI Conf.*, Indian Wells, USA, 2007, pp. 133–134.

[76] H. H. Chen and D. D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design", in *Proc. 34th Des. Automat.*, Anaheim, USA, 1997, pp. 638–643.

[77] C. Tinella, O. Richard, A. Cathelin, F. Reaute, S. Majcherczak, F. Blanchet, and D. Belot, "0.13 μm CMOS SOI SP6T antenna switch for multi-standard handsets", in *Proc. 6th Top. Meet. Sil. Monolit. Integr. Circ. RF Syst.*, San Diego, USA, 2006, p. 58.

[78] T. G. McKay, M. S. Carroll, J. Costa, C. Iversen, D. C. Kerr, and Y. Remoundos, "Linear cellular antenna switch for highly integrated SOI front-end", in *Proc. IEEE Int. SOI Conf.*, Indian Wells, USA, 2007.

[79] "Single-pole four-throw high-power switch", RF1450 Data sheet [Online]. Available: http://www.rfmd.com/pdfs/1450DS.pdf

Jean-Pierre Raskin was born in Aye, Belgium, in 1971. He received the industrial engineer degree from the Institut Supérieur Industriel d'Arlon, Belgium, in 1993, and the M.Sc. and Ph.D. degrees in applied sciences from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1994 and 1997, respectively. From 1994 to 1997, he was a research engineer at the Microwave Laboratory of UCL. He worked on the modeling, characterization and realization of MMIC's in silicon-on-insulator technology for low-power, low-voltage applications. In 1998, he joined the EECS Department of the University of Michigan, Ann Arbor, USA. He has been involved in the development and characterization of micromachining fabrication techniques for microwave and millimeter-wave circuits and microelectromechanical transducers/amplifiers working in harsh environments. In 2000, he joined the Microwave Laboratory of UCL as Associate Professor. Since 2007, he has been a Full Professor and Head of the Microwave Laboratory of UCL. His research interests are the modeling, wideband characterization and fabrication of advanced SOI MOSFETs as well as micro and nanofabrication of MEMS/NEMS sensors and actuators. He is an IEEE senior member, EuMA associate member and member of the Research Center in Micro and Nanoscopic Materials and Electronic Devices of UCL. He is author or co-author of more than 350 scientific articles.

e-mail: jean-pierre.raskin@uclouvain.be
Université catholique de Louvain
Microwave Laboratory
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium

Paper

The Impact of Externally Applied Mechanical Stress on Analog and RF Performances of SOI MOSFETs

Mostafa Emam, Samer Houri, Danielle Vanhoenacker-Janvier, and Jean-Pierre Raskin

Abstract— This paper presents a complete study of the impact of mechanical stress on the performance of SOI MOSFETs. This investigation includes dc, analog and RF characteristics. Parameters of a small-signal equivalent circuit are also extracted as a function of applied mechanical stress. Piezoresistance coefficient is shown to be a key element in describing the enhancement in the characteristics of the device due to mechanical stress.

Keywords— cutoff frequency f_T, intrinsic gain, mechanical stress, piezoresistance coefficient, SOI MOSFET.

1. Introduction

Scaling, channel engineering, high-*k* metal gate, etc., are different technological means to improve digital as well as analog performances of modern metal-oxide-semiconductor field effect transistor (MOSFET). Process induced strain whether tensile or compressive applied to the device during the fabrication process is also receiving increasing attention as another alternative to enhance the MOS transistor performance [1]. However, the impact of mechanical stress on the analog and RF characteristics of the transistor is rarely addressed in the literature [2], [3]. This work provides a complete study of the impact of mechanical stress on the dc, analog and RF characteristics of MOSFET transistors. This investigation is supported by the extraction of the different parameters of a small-signal equivalent circuit as a function of the applied mechanical stress. Piezoresistance coefficient is also calculated based on both dc and RF measurements. The mechanical stress is applied externally by means of a 4-point bending measurement setup coupled with dc and RF probe station.

Externally applied mechanical stress cannot in general reach the high values of process induced stress. However, the external application of mechanical stress provides high precision controllable values of stress, thus providing a valuable tool for the study of the device properties as a function of both tensile and compressive stress. Consequently, the results obtained from this study can easily be extrapolated to higher values of stress, applied either externally or internally.

This approach has been adopted repeatedly in the literature, with different variations in setup [4]–[13] knowing that Colman *et al.* [14] were the first to introduce this measurement setup with a single-points bending.

In this work, detailed investigation and results of strained devices are presented in such a way as to be easy to compare with the literature. Such study, important as it is for the design of analog and RF circuits has not been presented before.

2. Measurement Setup and Devices

A four-point bending setup is used to apply external mechanical stress from compressive (−250 MPa) to tensile (250 MPa), i.e., over a range of 500 MPa. A schematic representation of the 4-point bending setup is shown in Fig. 1.

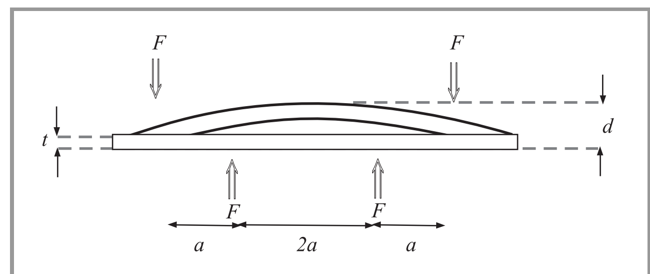


Fig. 1. Four-point bending setup.

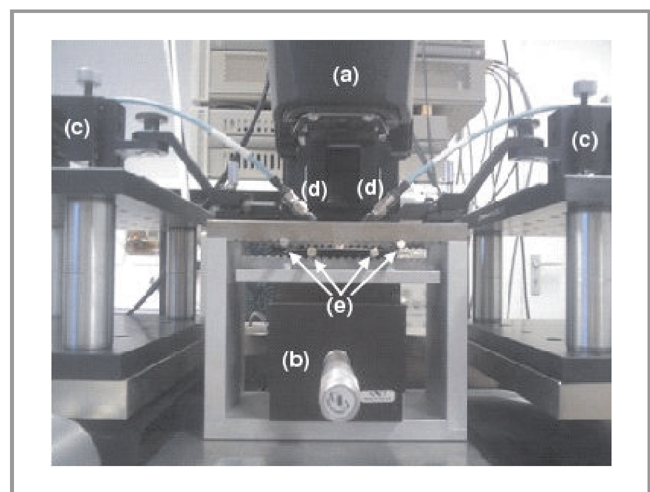


Fig. 2. A photograph of the used four-point bending setup with (a) the microscope, (b) the micrometer screw, (c) the RF probe holders, (d) the RF probes, (e) the metallic rods used to apply the mechanical stress on the 4 inch silicon wafer lying in between.

The value of the applied stress/strain can be calculated using the following formula [10]:

$$\sigma = \frac{3Edt}{8a^2}, \quad (1)$$

where σ is the applied mechanical stress in Pa, E is the wafer's Young's modulus, d is the maximum displacement due to the applied force, t is the wafer thickness, and $2a$ is the distance between the inner contact points.

Displacement is applied through a micrometer screw and measured by an optical microscope thus providing the precision of a few microns (Fig. 2).

Both n- and p-type fully-depleted (FD) silicon-on-insulator (SOI) MOSFETs are studied, bosth featuring 12 gate fingers (each 24 μm wide) connected in parallel. The channel length is 3 μm. The use of long channel devices helps to avoid short channel effects and hence results in a more accurate and less error prone extraction of equivalent circuit parameters.

3. Piezoresistance Coefficient

Piezoresistance coefficient in a transistor has been always defined with regards to the variation of the channel resistivity (or conductivity) as a function of the applied stress [14]–[16]. It was also introduced as the slope of the variation of the transconductance with respect to the transconductance at zero stress ($\Delta G_m/G_{m0}$) [4], [6]. In both cases, the resulting variation with stress is attributed to the dependence of carrier mobility on the applied mechanical stress. As will be shown in the next section, the same values of the piezoresistance coefficient could also be obtained from the variation of the output conductance with respect to the output conductance at zero stress ($\Delta G_d/G_{d0}$). This confirms the fact that the piezoresistance coefficient in a MOSFET device is mainly dominated by the variation of carrier mobility with the applied mechanical stress.

Table 1
Piezoresisitive coefficients π
for $a < 100 >$ wafer [%/kBar] for N- and PMOSFETs
in parallel and perpendicular orientations [7]

NMOS		PMOS	
$\pi \perp$	$\pi \parallel$	$\pi \perp$	$\pi \parallel$
−2.30	−4.97	−4.66	+6.48

The value of the piezoresitance coefficient depends on both the crystalline orientation, and the current orientation with respect to the applied strain [7], this is shown in Table 1 for $a < 100 >$ Si wafer. In this study, the mechanical stress is applied transversally with respect to the direction of the current, while the device channel orientation is <110>.

4. DC Characterization

Based on dc measurements (performed using a HP4145 device parameter analyzer), the piezoresistance coefficient is calculated using the variation in transconductance G_m (at V_{DS} = ±1.2 V and various V_{GS}, in the saturation regime) and also the variation in output conductance G_d (at V_{DS} = 50 mV and V_{GS} = ±2 V; i.e., in the linear regime) with the applied mechanical stress, as shown in Figs. 3 and 4. In both cases, a piezoresistance coefficient of 2 and 1 (10^{-4} MPa^{-1}), is found for P- and NMOSFETs, respectively.

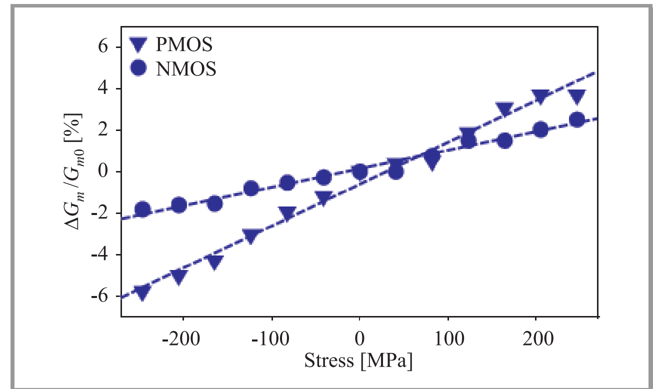


Fig. 3. Relative variation of the maximum dc transconductance G_m in saturation (V_{DS} = ±1.2 V) for P- and NMOSFETs.

The absolute variation of G_m with the applied mechanical stress shows is a 2.5 and 0.84% per 100 MPa for P- and NMOSFETs, respectively, in the saturation region (V_{DS} = ±1.2 V and V_{GS} = ±1.5 V).

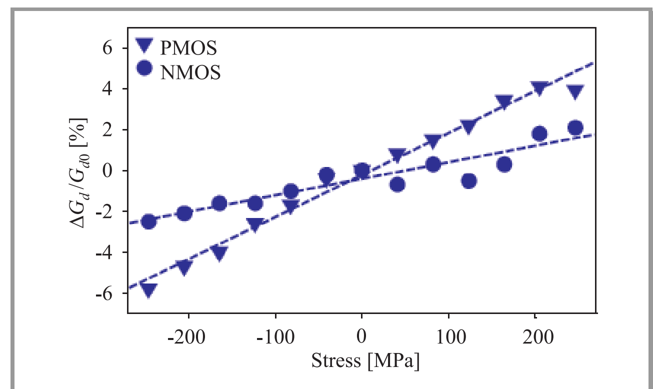


Fig. 4. Relative variation of output conductance G_d in the linear regime (V_{DS} = 50 mV and V_{GS} = ±2 V) for P- and NMOSFETs.

The dc open-loop gain (A_{V0}) is also improved by applying mechanical stress. This can be seen through the improvement of the early voltage V_{EA} with the applied mechanical stress since [17]:

$$A_{V0} = \frac{g_m}{I_{DS}} \cdot V_{EA}. \quad (2)$$

The first term (g_m/I_{DS}) is constant with stress, since the mobility is canceled out. An increase of ∼0.8 and

$\sim$0.7% per 100 MPa is noticed in V_{EA} for P- and NMOSFETs, respectively, at V_{GS} = ±1.8 V, as shown in Fig. 5. However, this increase drops to $\sim$0.3 and $\sim$0.2% per 100 MPa for P- and NMOSFETs, respectively, at V_{GS} = ±0.6 V. This variation is not dependent on the sensitivity of mobility to the applied mechanical stress, as is the case with the piezoresistance coefficient, since V_{EA} can be approximated by [18]

$$V_{EA} = \frac{I_{DS}}{g_D}, \tag{3}$$

where the effect of mobility is simplified between the numerator and denominator.

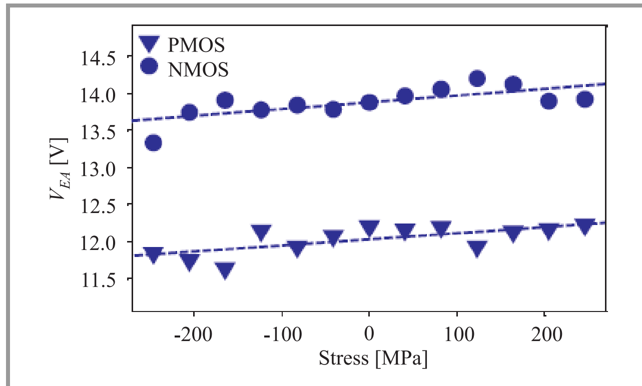


Fig. 5. Variation of early voltage V_{EA} with the applied stress at V_{GS} = ±1.8 V for P- and NMOSFETs.

It is also worth to notice that the threshold voltage V_{th} is quite constant with the applied mechanical stress, as can be seen from Fig. 6. The same applies to the subthreshold slope S as shown in Fig. 7.

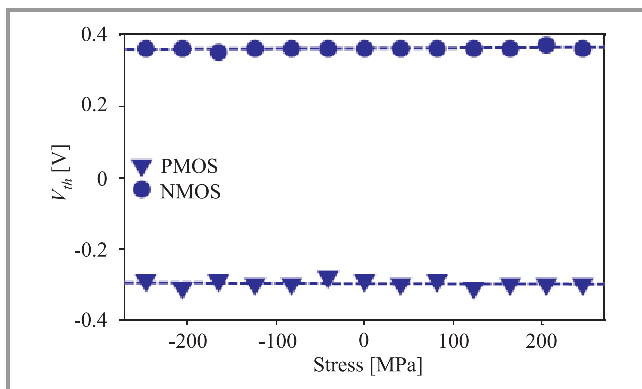


Fig. 6. Variation of threshold voltage V_{th} with the applied stress for P- and NMOSFETs.

The variation of G_m with the applied mechanical stress is also studied as a function of gate voltage V_{GS}. In the saturation region (V_{DS} = ±1.2 V), the piezoresistance coefficient shows an interesting reduction at V_{GS} values close to V_{th} (−0.3 and 0.36 V for P- and NMOSFETs, respectively). Figures 8 and 9 show values of less than 1 and 0.5% per 100 MPa of the applied mechanical stress for

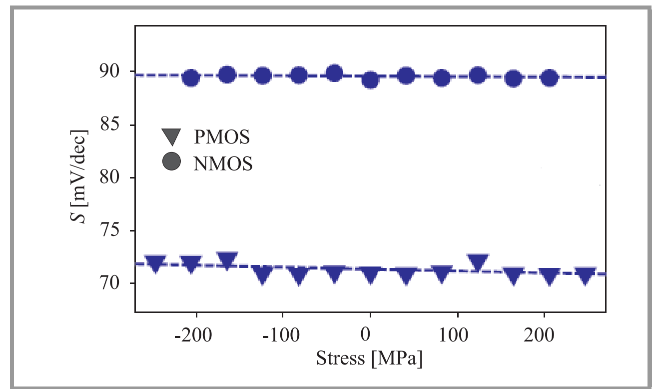


Fig. 7. Variation of subthreshold slope with the applied stress for P- and NMOSFETs.

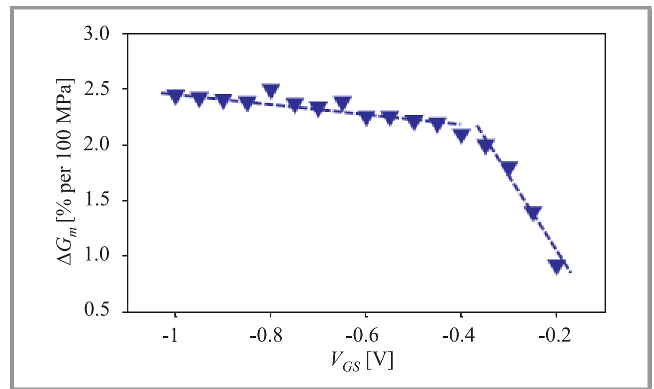


Fig. 8. Variation of dc transconductance G_m with the applied stress for PMOSFETs as a function of V_{GS} at $V_{DS} = -1.2$ V.

P- and NMOSFETs, respectively. These results show a direct relation between the piezoresistance coefficient and the density of carriers in the channel. This relation is further confirmed when studying the variation of G_m in the linear region (V_{DS} = ±50 mV). When the gate bias passes from $|V_{GS}| > |V_{th}|$ to $|V_{GS}| < |V_{th}|$, the channel passes from inversion to depletion, hence the dominant carriers in the channel change from holes to electrons and from electrons to holes for P- and NMOSFET, respectively. As a direct

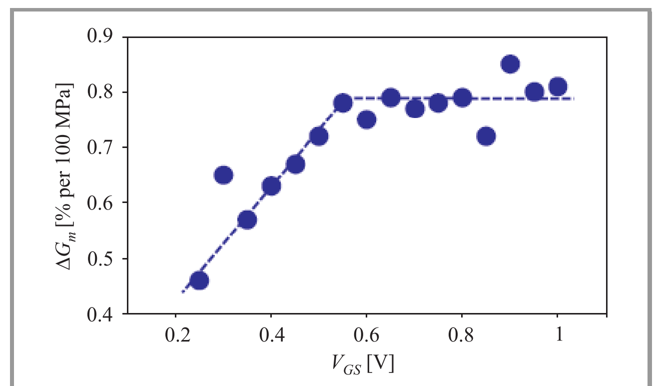


Fig. 9. Variation of dc transconductance G_m with the applied stress for NMOSFETs as a function of V_{GS} at $V_{DS} = 1.2$ V.

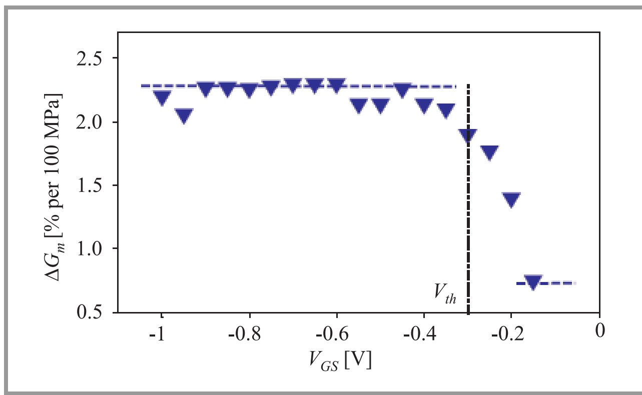


Fig. 10. Variation of dc transconductance G_m with the applied stress for PMOSFETs as a function of V_{GS} at $V_{DS} = -50$ mV.

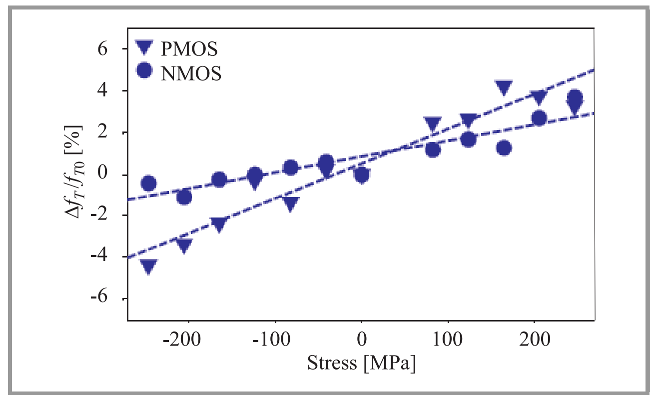


Fig. 12. Relative variation of cutoff frequency f_T with the applied mechanical stress for P- and NMOSFETs.

consequence, the piezoresistance coefficient value follows this transformation of the dominant carrier type in the channel. In PMOSFET (Fig. 10), ΔG_m goes from 2.3 to 0.75% per 100 MPa, with the latter value being close to that calculated earlier for NMOSFET. The inverse can be noticed for NMOSFET as shown in Fig. 11.

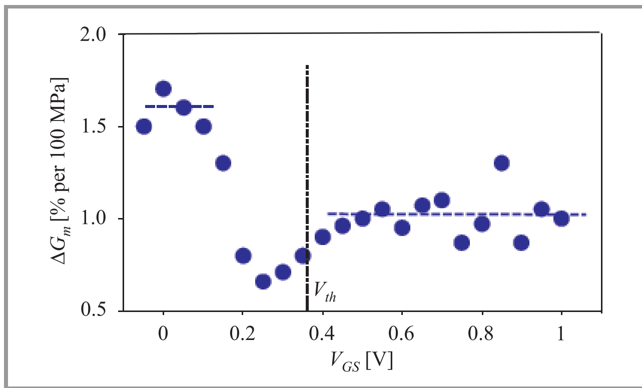


Fig. 11. Variation of dc transconductance G_m with the applied stress for NMOSFETs as a function of V_{GS} at V_{DS} = 50 mV.

This interesting shift in piezoresistance coefficient around V_{th} could be very useful for applications such as piezoresistance gages or switches.

5. RF Characterization

A 2-port Anritsu 37369A^{TM} vector network analyzer (VNA) is used to measure the S-parameters as a function of applied mechanical stress for both the P- and NMOSFETs. An open structure is used for a 1-step de-embedding procedure. Cutoff frequency f_T is extracted from the de-embedded $|H_{21}|$. The maximum of f_T is found at $V_{GS} = \pm 1.6$ V for PMOS and NMOS in the saturation region ($V_{DS} = 1.2$ V). The analysis of the dependence of f_T on the applied mechanical stress is conducted at this maximum f_T point.

Figure 12 shows the relative variation of f_T with respect to f_{T0} as a function of mechanical stress. The slope is found to be 1.7 and 0.8 (10^{-4} MPa^{-1}) for P- and NMOSFETs, respectively. This is slightly lower than the slopes found for $\Delta G_m/G_{m0}$ calculated from dc measurements as shown in Section 4. The absolute variation in f_T is 1.6 and 0.8% per 100 MPa for P- and NMOSFETs, respectively. It is interesting to notice that the ratio is 2:1, which is consistent with the piezoresistance coefficients ratio.

On the other hand, the ratio $f_{\max}/f_T$ shows a negligible variation with the applied mechanical stress, as can be seen from Fig. 13. This important figure of merit [19] depends

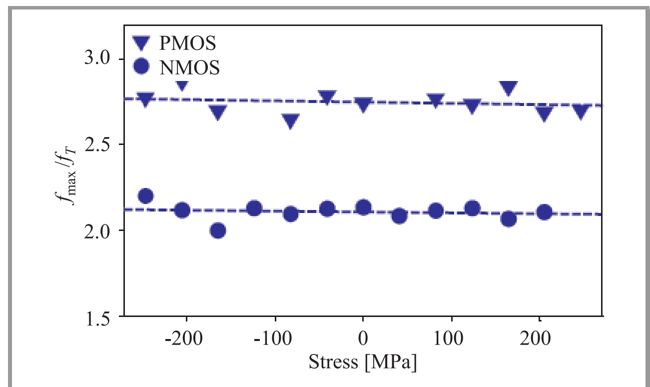


Fig. 13. Variation of (f_{max}/f_T) ratio with the applied mechanical stress for P- and NMOSFETs.

basically on the gate and source resistances (R_g and R_s). As will be shown later, these resistances show very slight variation with the applied mechanical stress.

5.1. Small-Signal Equivalent Circuit

It is of interest at this point to investigate the effect of mechanical stress on the various extrinsic and intrinsic parameters of the small-signal equivalent circuit. A typical small-signal equivalent circuit is shown in Fig. 14, where the elements outside the dashed box are the extrinsic elements whereas the elements inside the dashed box are the intrinsic elements. The term extrinsic refers to those elements which are independent of the bias condition but are scalable with the active zone. The term intrinsic denotes

the elements which are dependent on the bias condition and the size of the active region, thus representing the transistor behavior [20]. Extrinsic capacitances and inductances are neglected. Access elements are removed during the 1-step de-embedding procedure.

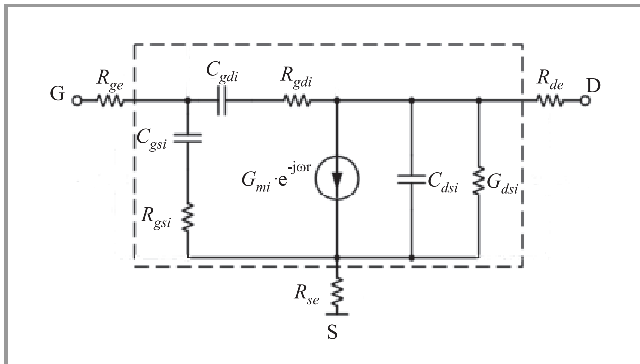


Fig. 14. Small-signal equivalent circuit for MOSFETs. The dashed box contains the intrinsic parameters.

Extrinsic resistances (R_{ge}, R_{se} and R_{de}) are first extracted using the cold-FET method [21]. The variation of these extrinsic resistances with the applied mechanical stress is shown in Fig. 15. Nearly constant behavior with stress can be clearly seen due to the highly doped drain and source areas resulting in low piezoresistance coefficient [22].

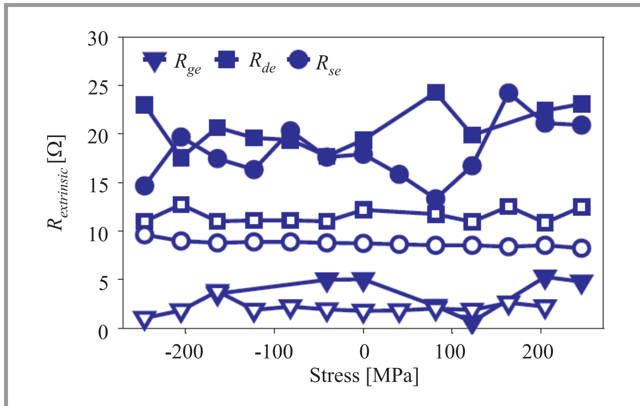


Fig. 15. Variation of extrinsic resistances with the applied mechanical stress for P- (solid symbols) and NMOSFETs (empty symbols).

After removing the effect of extrinsic resistances, the next step is to extract the intrinsic elements of the small-signal equivalent circuit using the direct extraction method proposed in [20]. The extraction is performed in the saturation region at $V_{GS} = \pm 1.5$ V and $V_{DS} = 1.2$ V.

The total gate capacitance ($C_{gg} = C_{gs} + C_{gd}$) shows a slight variation with the applied mechanical stress, namely 0.3% per 100 MPa for both P- and NMOSFETs, as shown in Fig. 16.

The variation of the intrinsic transconductance G_{mi} (extracted from RF measurements) with the applied mechanical stress shows a good agreement with the dc extracted values, taking into account the errors related to the extraction procedures (extrinsic and intrinsic). A variation of 2.3 and 0.8% per 100 MPa is calculated for P- and NMOSFETs, respectively, as shown in Fig. 17. On the other hand, the intrinsic output conductance shows a slight shift from the values extracted for the transconductance, showing a variation with the applied mechanical stress of 2.5 and 0.65% per 100 MPa for P- and NMOSFETs, respectively, as shown in Fig. 18.

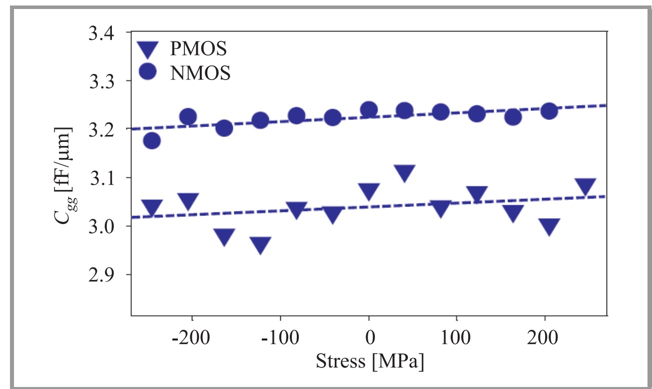


Fig. 16. Variation of intrinsic total gate capacitance C_{gg} with the applied mechanical stress for P- and NMOSFETs.

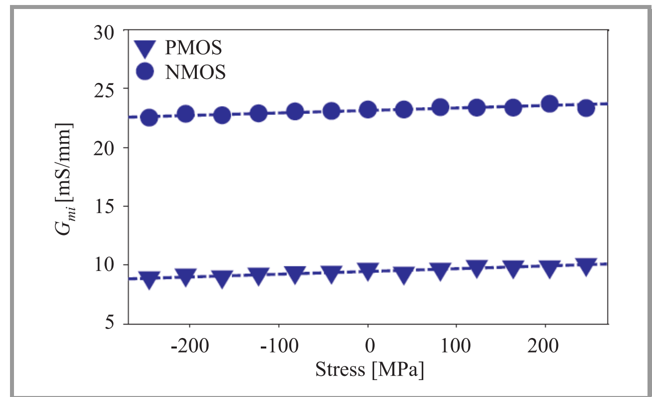


Fig. 17. Variation of intrinsic transconductance G_{mi} with the applied mechanical stress for P- and NMOSFETs.

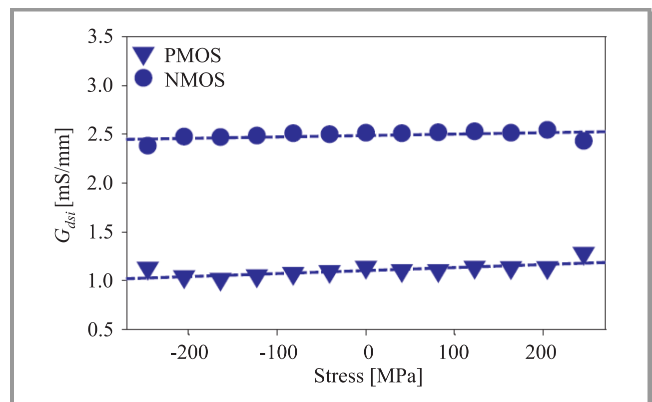


Fig. 18. Variation of intrinsic output conductance G_{dsi} with the applied mechanical stress for P- and NMOSFETs.

The extraction of all intrinsic parameters is performed for devices in saturation, i.e., $V_{DS} = \pm 1.2$ V and $V_{GS} = \pm 1.5$ V. Based on the previous results for C_{gg} and G_{mi}, and knowing that the cutoff frequency f_T is usually approximated by

$$f_T = \frac{G_m}{2\pi C_{gg}} \quad (4)$$

it can be seen that the variation in C_{gg} with the applied mechanical stress is not negligible, however, it is of secondary importance, whereas the major effect comes from the variation in G_{mi}.

5.2. Piezoresistance Coefficient from RF Extraction

The calculation of the piezoresistance coefficient based on RF measurements could be another important tool to characterize the MOSFET behavior under mechanical stress. DC measurements of SOI devices, especially G_m and G_{ds} could suffer from some shift due to the self-heating effect. It is possible to avoid these problems by using the intrinsic parameters, like G_{mi} and G_{dsi}, extracted from RF measured data, as presented in the previous section.

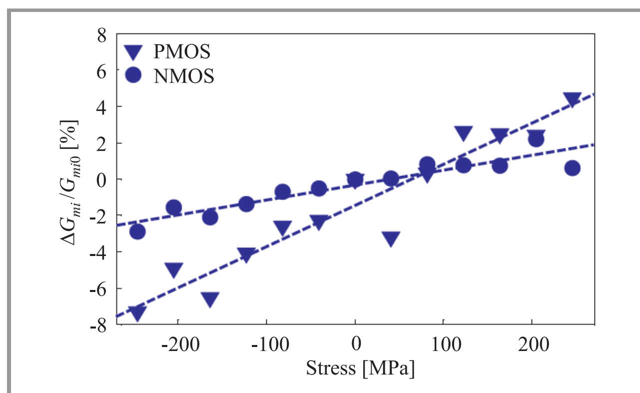


Fig. 19. Relative variation of intrinsic transconductance G_{mi} in saturation ($V_{DS} = \pm 1.2$ V and $V_{GS} = \pm 1.5$ V) for P- and NMOSFETs.

Figure 19 shows the relative variation of the intrinsic transconductance G_{mi} with the applied mechanical stress for both P- and NMOSFETs. A piezoresistance coefficient of 2.25 and 0.82 (10^{-4} MPa^{-1}) for P- and NMOSFETs, respectively, can be calculated from the slope of the corresponding variation. These values are close to the values extracted from dc measurements. The small difference is related to the corrected transconductance by removing the self-heating effect.

6. Conclusion

Based on dc and RF measurements, the mechanical stress is shown to directly affect the dc, analog and RF performances of P- and NMOS transistors. Most of these effects are related to the variation of carrier mobility with the applied mechanical stress, but it was shown that some other effects are also related to the variation of the carrier density inside the channel with the applied mechanical stress.

Cutoff frequencies were shown to vary with the applied mechanical stress as a direct result of the variation of transconductance, while the gate capacitance would still have a slight secondary effect on the variation of cutoff frequency.

On the other hand, the ratio $f_{\max}/f_T$ was shown to slightly vary with stress since it is dominated by the relatively stable extrinsic resistances of the transistors.

The ratio between the performance variation in PMOSFET to the performance variation in NMOSFET with the applied mechanical stress, was shown to be equal to the ratio of piezoresistance coefficients of P- to NMOSFETs.

This characterization methodology being limited in this study to 500 MPa of externally applied mechanical stress, could be extrapolated to higher values of stress/strain, applied internally due to fabrication process steps.

Acknowledgements

This work was supported in part by the Walloon Region (Convention no. 516125 CORMORAN).

References

[1] V. Chan, K. Rim, M. Ieong, S. Yang, R. Malik, Y. W. Teh, M. Yang, and Q. C. Ouyang, "Strain for CMOS performance improvement", in *Proc. IEEE Custom Integr. Circ. Conf.*, San Jose, USA, 2005, pp. 667–673.

[2] D. V. Singh, K. A. Jenkins, J. Sleight, Z. Ren, M. Ieong, and W. Haensch, "Strained ultrahigh performance fully depleted nMOSFETs with f_t of 330 GHz and sub-30 nm gate lengths", *IEEE Electron Dev. Lett.*, vol. 27, no. 3, pp. 191–193, 2006.

[3] S. Houri, M. Emam, and J.-P. Raskin, "RF behavior of strained fully depleted SOI MOSFETs", in *Proc. EUROSOI Conf.*, Cork, Ireland, 2008, pp. 55–56.

[4] A. Hamada, T. Furusawa, N. Saito, and E. Takeda, "A new aspect of mechanical stress effects in scaled MOS devices", *IEEE Electron Dev. Lett.*, vol. 38, no. 4, pp. 895–900, 1991.

[5] C.-L. Huang, H. R. Soleimani, G. J. Grula, J. W. Sleight, A. Villani, H. Ali, and D. A. Antoniadis, "LOCOS-induced stress effects on thin-film SOI devices", *IEEE Trans. Electron Dev.*, vol. 44, no. 4, pp. 646–650, 2004.

[6] R. Degraeve, G. Groeseneken, I. De Wolf, and H. E. Maes, "The effect of externally imposed mechanical stress on the hot-carrier-induced degradation of deep-sub micron nMOSFET's", *IEEE Trans. Electron Dev.*, vol. 44, no. 6, pp. 943–950, 1997.

[7] C. Gallon, G. Reimbold, G. Ghibaudo, R. A. Bianchi, R. Gwoziecki, S. Orain, E. Robilliart, C. Raynaud, and H. Dansas, "Electrical analysis of mechanical stress induced by STI in short MOSFETs using externally applied stress", *IEEE Trans. Electron Dev.*, vol. 51, no. 8, pp. 1254–1261, 2004.

[8] S. E. Thompson, S. Suthram, Y. Sun, G. Sun, S. Parthasarathy, M. Chu, and T. Nishida, "Future of strained Si/semiconductors in nanoscale MOSFETs", in *Proc. Int. Electron Dev. Meet. IEDM*, San Francisco, USA, 2006, pp. 1–4.

[9] F. Rochette, M. Cassé, M. Mouis, G. Reimbold, D. Blachier, C. Leroux, B. Guillaumot, and F. Boulanger, "Experimental evidence and extraction of the electron mass variation in [110] uniaxially strained MOSFETs", *Solid-State Electron.*, vol. 51, no. 11-12, pp. 1458–1465, 2007.

[10] S. Suthram, J. C. Ziegert, T. Nishida, and S. E. Thompson, "Piezoresistance coefficients of (100) silicon nMOSFETs measured at low and high (∼1.5 GPa) channel stress", *IEEE Electron Dev. Lett.*, vol. 28, no. 1, pp. 58–61, 2007.

[11] Y. S. Choi, T. Numata, T. Nishida, R. Harris, and S. E. Thompson, "Impact of mechanical stress on gate tunneling currents of germanium and silicon p-type metal-oxide-semiconductor field-effect transistors and metal gate work function", *J. Appl. Phys.*, vol. 103, no. 64510, pp. 1–5, 2008.

[12] J.-S. Lim, A. Acosta, S. E. Thompson, G. Bosman, E. Simoen, and T. Nishida, "Effect of mechanical strain on $1/f$ noise in metal-oxide semiconductor field-effect transistors", *J. Appl. Phys.*, vol. 105, no. 54504, pp. 1–11, 2009.

[13] Y. J. Kuo, T. C. Chang, P. H. Yeh, S. C. Chen, C. H. Dai, C. H. Chao, T. F. Young, O. Cheng, and C. T. Huang, "Substrate current enhancement in 65 nm metal-oxide-silicon field-effect transistor under external mechanical stress", *Thin-Solid Films*, vol. 517, no. 5, pp. 1715–1718, 2009.

[14] D. Colman, R. T. Bate, and J. P. Mize, "Mobility anisotropy and piezoresistance in silicon p-type inversion layers", *J. Appl. Phys.*, vol. 39, no. 4, pp. 1923–1931, 1968.

[15] G. Dorda, "Piezoresistance in quantized conduction bands in silicon inversion layers", *J. Appl. Phys.*, vol. 42, no. 5, pp. 2053–2060, 1971.

[16] B. Borchert and G. E. Dorda, "Hot-electron effects on short-channel MOSFET's determined by the piezoresistance effect", *IEEE Trans. Electron Dev.*, vol. 35, no. 4, pp. 483–488, 1988.

[17] D. Flandre, J.-P. Eggermont, D. De Ceuster, and P. Jespers, "Comparison of SO1 versus bulk performances of CMOS micropower single-stage OTAs", *IEEE Electron. Lett.*, vol. 30, no. 23, pp. 1933–1934, 1994.

[18] S. M. Sze, *Semiconductor Devices Physics and Technology*. New York: Wiley, 1985.

[19] G. Dambrine, C. Raynaud, D. Lederer, M. Dehan, O. Rozeaux, M. Vanmackelberg, F. Danneville, S. Lepilliet, and J.-P. Raskin, "What are the limiting parameters of deep-submicron MOSFETs for high frequency applications?", *IEEE Electron Dev. Lett.*, vol. 24, no. 3, pp. 189–191, 2003.

[20] J.-P. Raskin, R. Gillon, J. Chen, D. Vanhoenacker-Janvier, and J.-P. Colinge, "Accurate SOI MOSFET characterization at microwave frequencies for device performance optimization and analog modeling", *IEEE Trans. Electron Dev.*, vol. 45, no. 5, pp. 1017–1025, 1998.

[21] A. Bracale *et al.*, "A new approach for SOI devices small-signal parameters extraction", *Anal. Integr. Circ. Sig. Process.*, vol. 25, no. 2, pp. 157–169, 2000.

[22] Y. Kanda, "A graphical representation of the piezoresistance coefficients in silicon", *IEEE Trans. Electron Dev.*, vol. ED-29, no. 1, pp. 64–70, 1982.

Mostafa Emam received the B.Sc. degree in electronics and communication engineering from the Ain Shams University, Cairo, Egypt, in 2001, the Diplome d'Ingénieur degree in electronics and signal processing, and the M.Sc. degree in design of microelectronics circuits and systems both from the Institute National Polytechnique, Toulouse, France, in 2005. He has been working toward the Ph.D. degree in engineering sciences in the Microwave Laboratory, Ecole Polytechnique de Louvain, Université catholique de Louvain, Belgium, since 2006. He worked for the Analog/Mixed Signal Group, Mentor Graphics (2005), George Mason University, Fairfax and AMD, Sunnyvale, USA (2006). His research interests include the characterization and modeling of SOI devices in dc, RF, large-signal, and high frequency noise, for harsh-environment applications and under mechanical stress conditions as well as the design and simulation of RF SOI circuits.

e-mail: mostafa.emam@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium

Samer Houri received his B.E. in electrical engineering from the Beirut Arab University, Lebanon, in 2005, and his M.Sc. degree in nano-electronics from the Université de Provence, Marseille, France, in 2006. Currently working towards a Ph.D. degree in engineering sciences in the Université catholique de Louvain, Belgium. Research interests include RF MEMS and devices.

e-mail: samer.houri@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium

Danielle Vanhoenacker-Janvier received the electrical engineer degree and the Ph.D. degree in applied sciences from the Université catholique de Louvain (UCL), Belgium, in 1978 and 1987, respectively. She is currently with UCL, where she was an Assistant (1979–1987), a Senior Scientist (1987–1994), an Associate Professor (1994–2000), a Professor (2000–2007), and has been a Full Professor since 2007 with the Microwave Laboratory, where she was the Head (2001–2007). She has been involved in the study of atmospheric effects on propagation above 10 GHz for over 30 years. She was engaged in the analysis, design, and measurement of microwave planar passive and active circuits with a special interest, since 1994, in microwave ICs on SOI. She has authored over 140 technical papers and coauthored a book. She is a Reviewer for various international conferences and IEEE and IET journals. She is a member of the evaluation committee of various Laboratories and Research Centers.

e-mail: danielle.vanhoenacker@uclouvain.be
Université catholique de Louvain
Place du Levant, 3, Maxwell Building
B-1348 Louvain-la-Neuve, Belgium

Jean-Pierre Raskin – for biography, see this issue, p. 17.

Invited paper

Prospects and Development of Vertical Normally-off JFETs in SiC

Mietek Bakowski

Abstract— This paper reviews the prospects of normally-off (N-*off*) JFET switch in SiC. The potential of selected vertical JFET concepts and all-JFET cascode solutions for N-*off* operation is analyzed using simulations. The performance of analyzed concepts is compared in terms of blocking voltage, specific on-state resistance, maximum output current density and switching performance in the temperature range from 25°C to 250°C. The main objective of the analysis is to ascertain consequences of different design and technology options for the total losses and high temperature performance of the devices.

Keywords— JFET cascode, normally-off, SiC, vertical JFET.

1. Introduction

A voltage-controlled, normally-off (N-*off*) SiC switch with specific on-state resistance of around $1.0\,\mathrm{m\Omega\,cm^2}$ is desired by many applications including the high volume automotive market. During the recent years several concepts including metal oxide semiconductor field effect transistor (MOSFET) and junction field effect transistor (JFET) designs have been developed to realize such an ideal SiC switch [1]–[4]. The SiC MOSFET would be the device of choice as soon as the SiO_2/SiC interface and reliability issues are solved. Meanwhile, SiC JFET designs are becoming more and more interesting, because of their ruggedness and the achievable low on-state resistance.

2. SiC Switch Concepts

The JFET is presently the most mature switch concept in SiC transistor technology. There are historically several factors accounting for that. First of all, the high electric field strength and the reasonably high electron mobility of SiC make unipolar SiC devices interesting for high voltage applications [5]. As a unipolar device the JFET is forgiving with respect to material quality and does not suffer from bipolar instability [6].

In addition, it does not require as high SiO_2/SiC interface quality as that needed in the MOSFET, since it's function depends on the conducting channel situated in the bulk of the device and controlled by the reverse biased p-n junction. Finally, the wide band gap of the SiC material gives SiC JFETs advantage of the high temperature operation and facilitates N-*off* design due to the high value of the build-in voltage.

Figure 1 illustrates schematically the status of SiC switch concepts during the last decade.

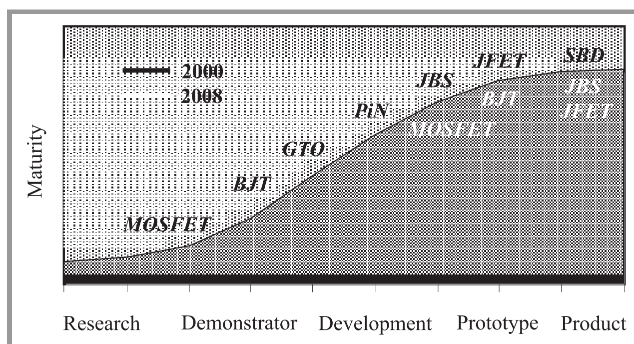


Fig. 1. Maturity of main SiC rectifiers and switches. Explanations: BJT – bipolar junction transistor, GTO – gate turn-off thyristor, PiN – p-i-n rectifier, JBS – junction barrier-controlled Schottky rectifier, SBD – Schottky barrier diode.

3. Advantages of JFET Technology

The JFET can be realized using epitaxial growth for the main voltage controling junctions and for the conduction channel. In this way ion-implantation can be avoided. Ion implantation is in these two cases a performance limiting technology due to excessive damage especially for high doping concentrations. The high quality of the conduction channel and good control of the channel dimensions and doping are crucial for the JFET performance. The normally-on (N-*on*) JFET design is capable of extremely low on-state resistance. Only the SiC BJT has the potential of comparably low on-state voltage due to the even number of p-n junctions configured in such a way that they compensate each other's built-in voltage in saturation. Finally, the JFET does not require the use of an anti-parallel diode reducing the number of semiconductor components in a system [7], [8].

Normally-on JFETs are not easily accepted by the market due to system safety requirements, regardless of their excellent on-resistance. N-*off* JFETs on the other hand require a narrow and relatively low doped channel to ensure the N-*off* performance, and thus pay a penalty in terms of the on-state performance. N-*off* JFETs are also vulnerable to the electromagnetic interference (EMI) noise due to the small range of the gate control voltage. Hence, the gate control circuitry for JFETs requires special attention to ensure reliable operation. In the case of N-*on* JFETs the development of inherently safe gate drivers is particularly desired in order to guarantee the safety of the whole system [9].

4. N-off Epitaxial JFET Concepts

This contribution presents the review of the main epitaxial SiC JFET concepts and analyzes their prospects for N-off performance based on simulations. The vertical JFET types discussed here are the lateral channel JFET (LCJFET) [1], the buried grid JFET (BGJFET) [2], and the dual gate vertical channel trench JFET (DGTJFET) [3]. Furthermore, the comparison of N-off and N-*on* SiC JFET designs of these selected epitaxial concepts in terms of blocking voltage, specific on-state resistance, current density, and switching performance trade-offs and limitations is done.

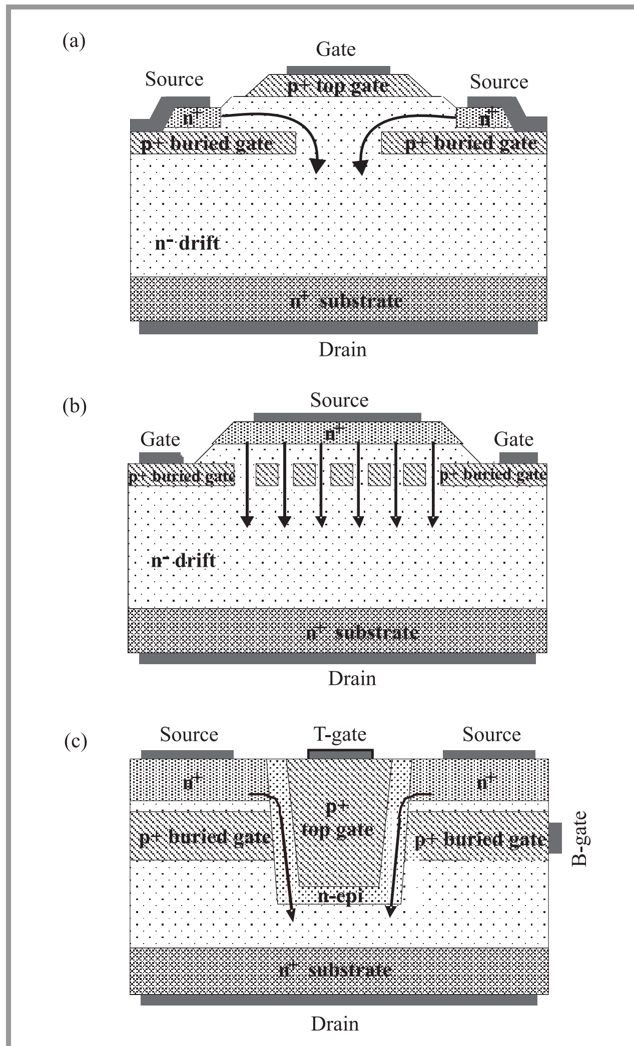


Fig. 2. Schematic drawings of the vertical epitaxial JFET concepts (a) lateral channel JFET, (b) buried grid JFET and (c) double gate vertical channel trench JFET.

The most successful JFET type in terms of voltage and current ratings has been the lateral channel JFET developed by SiCED [10]. A schematic drawing of the LCJFET design is shown in Fig. 2(a). The LCJFET allows optimal control of the channel parameters and offers the largest ease of fabrication compared to other concepts. It also offers the use of the inherent body diode as an anti-parallel diode in switching applications since the buried gate is preferably connected to source. This is necessary in order to reduce the Miller capacitance and thus maintain the high speed of operation as will be shown later in the paper.

The original LCJFET structure uses ion-implantation for the gate and the base region, and planar epitaxial growth for the defect-free channel layer. This leads to advantages in terms of ease of fabrication, freedom of parameter choice due to a wide design window, and small fabrication tolerances. The disadvantage is a relative large specific on-resistance, which is related to the large cell pitch due to the lateral configuration of the channel. In addition, the large cell pitch of 10 to 16 μm makes the use of both gates for the conduction control not feasible due to the prohibitively large gate charge required during switching. The concept is also characterized by relatively low saturation current levels and in order to achieve low on-state resistance, the demonstrated LCJFET designs are typically of N-*on* type.

As will be shown in this paper, the N-off design is not feasible with the LCJFET concept. The single gate drive with the buried gate connected to the source is necessary to mitigate the large Miller effect related to the large cell pitch that otherwise dominates the turn-off behavior. The large cell pitch and the single gate drive make the saturation current levels prohibitively low for any power switching application.

A schematic drawing of the BGJFET design is shown in Fig. 2(b). The main advantage of the vertical BGJFET concept is the small cell pitch that makes low specific on-state resistance and high saturation current densities possible. Furthermore, the inherent symmetric gate drive and the wide design window for the channel length, width, and doping make the N-off design feasible. The disadvantage is that optimization of the channel doping is not as easy as in the case of the lateral channel growth. Trenches have to be etched in the p-doped grid layer. These etched trenches must be epitaxially refilled to full extent. In the case of an implanted grid the possible use of higher doping in the channel is limited by the necessity to compensate the higher doped n-layer on top of the drift layer by the p-type grid implant. The channel doping in this concept is ultimately limited by the tolerances of the photolithography and trench etching process. Another disadvantage is that the use of the integral gate to drain body diode is not readily available with this concept.

A schematic drawing of the DGTJFET structure suggested by DENSO [3] is shown in Fig. 2(c). The DGTJFET offers high current rating capabilities for N-off mode operation. This design combines the advantages of the LCJFET and the BGJFET concepts by using epitaxial regrowth in trenches to define the pitch and the direction of the channel, transforming it from lateral to vertical. The epitaxial channel is grown on the vertical trench walls with the tolerances and the wide design window comparable to the LCJFET.

The DGTJFET is basically the same concept as the LCJFET, but allows a dramatic reduction of the cell pitch

and of the Miller capacitances due to the vertical channel. The low gate to drain capacitance makes fast switching possible even under dual gate driving conditions, while the small cell pitch and the dual gate control results in very low specific on-resistance. The dual gate drive and the wide design window for the channel optimization gives also exceptionally high saturation current levels for N-*off* designs. In addition, the negative temperature dependence of the saturation current is greatly reduced due to the possibility of using highly doped channels. It has thus the advantages of the BGJFET, but it also can surpass its performance due to the larger design window for the channel doping and width. The use of the integral gate to drain body p-n diode is possible in this concept. It is, however, a matter of trade-off with the possible saturation current density as will be shown later in this paper. The disadvantage of this concept is the complex process involving epitaxial regrowth in trenches and planarization techniques.

5. All-JFET Integral Cascodes

Also, the prospects of integral JFET/JFET cascodes are discussed. The integral cascode consisting of a high voltage (HV) N-*on* SiC power JFET and a control low voltage (LV) N-*off* JFET is a powerful concept for a N-*off* SiC switch [8]. The main advantage of the cascode solution is the greatly increased speed of switching due to the fact that the buried gate of the high voltage device, which is connected to the source (ground) of the cascode (source of the LV device), shields the low voltage device, which is driven by the control gate [11]. The gate to drain capacitance is thus reduced. In this way the Miller capacitance is being charged by the main circuit and not by the gate circuit.
Another advantage is the possibility of utilizing the built-in body p-n diode formed between the buried gate of the high voltage N-*on* device and the drain of the cascode as an anti-parallel diode in the switching applications. Two integral cascode concepts considered for analysis are shown in Fig 3. Both are based on a HV BGJFET controlled in first case by a LV N-*off* recessed gate JFET (RGJFET) [8] and in the second case by a LV N-*off* BGJFET. It is of interest to analyze the prospects of these integral cascode solutions for power applications. The hybrid cascode is difficult to optimize, which results in degraded on-state and switching performance [12]. In addition, the cascode configuration with a Si MOSFET compromises the high temperature capability of the SiC JFET [13].
The integral cascode concept allows the optimization of the cascode performance and achieves an on-state voltage comparable to the stand alone N-*on* JFET with equal voltage rating. This is due to several factors. First of all the LV N-*off* JFET can be made less N-*off* by shortening the channel. This is facilitated by the electric field shielding effect due to the buried gate of the HV JFET. The limit is set by the highest tolerated value for the leakage current. It is very important to have as high doping in the channel of the LV JFET as possible since the LV N-*off* JFET

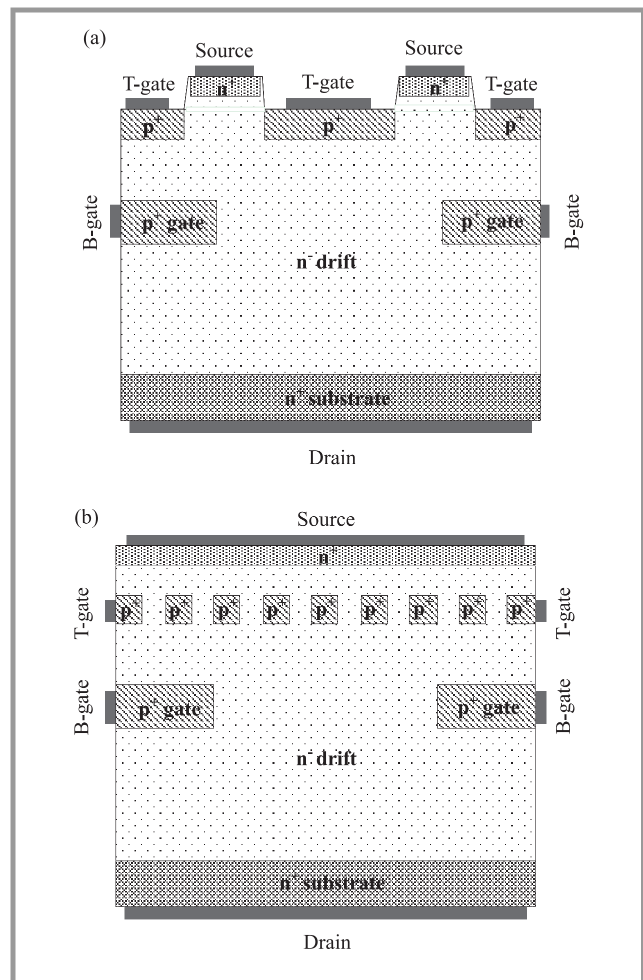


Fig. 3. Schematic drawings of the all-JFET integral cascode concepts (a) recessed gate JFET controlled cascode, (b) buried grid JFET controlled cascode.

determines the current throughput of the whole cascode. Secondly, the HV N-*on* JFET can be made more conductive by increasing the spacing of the buried gate grid. The limit is set by the electric field crowding at the edges of the buried grid when spacing becomes too large.
The grid of the N-*on* HV JFET section must support the full high voltage. This means the grid spacing has to be chosen so that the premature breakdown due to the enhanced electric field at the grid corners is avoided. As a result the cascode with output pentode characteristics may be controlled by a short channel JFET with triode characteristics. In this case the negative bias appearing on the buried gate of the HV JFET is beneficial in obtaining output characteristics with saturation at high current densities and with high value of the saturation current due to the very low on-state voltage of the optimized LV JFET. This will be further exemplified below using the concept from Fig. 3(a). It will be shown furthermore that the cascode optimization potential increases especially with increasing design voltage. This is due to the larger field shielding effect and wider grid spacing range available at lower doping concentrations of the drift region. The on-resistance values lower than those

of the N-*on* JFETs are feasible in the voltage range above 1000 V, as will be shown below.

6. N-off Design and High-Temperature Operation Considerations

The JFET is a unipolar device and as such should show a significant increase of the on-resistance with temperature. In the pure case of the resistivity being controlled by the drift region it should follow the relation of mobility degradation with temperature due to the phonon scattering [14].

When operating the JFET at higher junction temperatures one has also to consider the reduction of the built-in potential with temperature. The reduction of the built-in voltage V_{bi} with temperature is shown in Fig. 4, V_{bi} defines the limit of the unipolar operation for the JFET. A reduction of V_{bi} by 0.4 V is observed when increasing the temperature from room temperature to 250°C. For that reason all the values in this article have been calculated with the applied gate voltage of 2.4 V which includes also a 10% margin for the process tolerances.

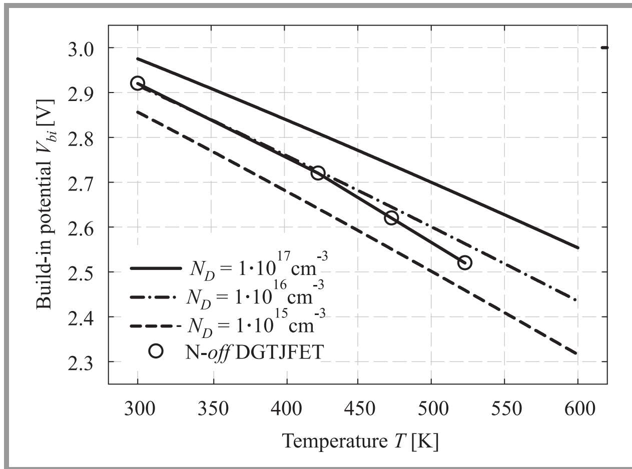


Fig. 4. Built-in potential as a function of the temperature calculated for different channel doping concentrations (see legend) and obtained from simulations (symbols) of the N-off DGTJFET.

In order to realize the N-off SiC JFET the channel of the device has to be fully depleted by the gate to source potential with no applied voltage. This means that the threshold voltage V_{th} has to be equal or higher than zero ($V_{th} \geq 0$). The potential at a p-n junction at zero applied voltage is equal to the so called built-in voltage V_{bi} being the function of the material band gap E_g and the acceptor and donor doping densities N_A and N_D at both sides of the junction (Eq. (1)):

$$V_{bi} = \frac{\mathrm{k}\cdot T}{q}\cdot\ln\left(\frac{N_A\cdot N_D}{N_V\cdot N_C}\right)+\frac{E_g}{q}. \tag{1}$$

A wide band gap material is characterized by a higher value of V_{bi}. For 4H-SiC V_{bi} is at least 2.5 V as compared to 0.6 V for Si. For JFETs, the highest gate voltage that can be applied in forward direction without entering the bipolar mode of operation is given by the built-in potential of the gate-source junction. The gate-source built-in voltage calculated using Eq. (1) is shown in Fig. 1 for the gate region doping of 10^{19}cm^{-3} and the channel doping of 10^{15}cm^{-3}, 10^{16}cm^{-3}, and 10^{17}cm^{-3}. For comparison the applied gate voltages, obtained from simulations of the N-off DGTJFET with a channel doping of $4\cdot10^{16}\text{cm}^{-3}$, at which a significant injection current in the channel is observed, are included in Fig. 4. A correction of $2\mathrm{k}T/q$ due to the majority carrier distribution tails has been subtracted from the calculated V_{bi} values [15]. The values of other parameters are after [14].

The channel doping and width have to be selected satisfying the following condition for the symmetrical gate configuration

$$V_{th} = V_{bi} - \frac{q\cdot N_D\cdot T_{ch}^2}{2\cdot\varepsilon\cdot\varepsilon_0}, \tag{2}$$

where: V_{th} is the threshold voltage, N_D is the channel doping and T_{ch} is the half-width of the channel. The second term in Eq. (2) is the so called pinch-off voltage as obtained from the Poisson equation in the case of an abrupt asymmetrical junction.

In this case the gate doping concentration is much higher than that of the channel region. Partial derivation and normalization of Eq. (2) yields:

$$\frac{\Delta V_{th}}{V_{th}} = \frac{\Delta N_D}{N_D} + \frac{2\cdot\Delta T_{ch}}{T_{ch}}, \tag{3}$$

where: ΔN_D and ΔT_{ch} are technological tolerances of channel doping and channel width.

Since the N-off operation requires $V_{th} \geq 0$, it is clearly seen from Eq. (3) that the current handling capability of the N-off device is by necessity derated due to process tolerances. A design with smaller doping and channel width compared to the one with maximized channel conductivity has to be used to realize a stable N-off device (see Eq. (4)).

In a similar way it can be shown that the choice of the highest possible channel doping is beneficial for compensating the resistivity degradation with temperature caused by phonon scattering. The channel conductivity is proportional to three temperature dependant parameters, the electron mobility μ_n, the electron concentration in the channel assumed to be equal to the ionized doping concentration N_D^+, and the active channel width $2\left(T_{ch}-w_n\right)$:

$$\sigma \sim \mu_n\cdot N_D^+\cdot 2\left(T_{ch}-w_n\right). \tag{4}$$

Derivation and normalization of Eq. (4) gives:

$$\frac{\Delta\sigma}{\sigma_{RT}} = \frac{\Delta\mu_n}{\mu_{n,RT}} + \frac{\Delta N_D^+}{N_{D,RT}^+} - \frac{\Delta w_n}{T_{ch}-w_{n,RT}}, \tag{5}$$

where: σ_{RT}, $\mu_{n,RT}$, $N_{D,RT}$, $w_{n,RT}$ are the room temperature (27°C) values of the conductivity, the electron mobility, the donor density, and the space charge width, respec-

tively, $\Delta\sigma$, $\Delta\mu_n$, ΔN_D^+, Δw_n are the differences of the temperature dependent parameters with respect to their room temperature values.
The sensitivity of the channel resistivity to the listed parameters can now be evaluated using the following temperature dependencies:

$$\mu_n \sim T^{-2.15}, \tag{6a}$$

$$N_D^+ = N_D \cdot \left[1 - \frac{N_D}{N_c(T)} \cdot \exp\left(\frac{0.065}{\mathrm{k}T}\right)\right], \tag{6b}$$

and

$$w_n \sim \sqrt{\frac{V_{bi}(T) - V_G}{N_D}}. \tag{6c}$$

The temperature dependent parameters electron mobility $\mu_n(T)$, density of states of the conduction band $N_C(T)$, density of states of the valence band $N_V(T)$, band gap $E_g(T)$ and donor activation energy $E_d = 0.065$ eV are after [14] and $V_{bi}(T)$ is given by Eq. (1). Incomplete ionization applies in the channel region only and in the space charge region 100% ionization is assumed.
Equation (5) has been evaluated for the gate region doping of 10^{19}cm^{-3} and the channel doping of 10^{15}cm^{-3}, 10^{16}cm^{-3}, and 10^{17}cm^{-3} with $V_G = 2.4$ V. The relative change of the channel resistivity with temperature normalized to the room temperature value is shown in Fig. 5. The compensating effect of the high donor density becomes significant in lowering the temperature degradation of the channel resistance for donor densities above 10^{16}cm^{-3}. This is because the effect of incomplete ionization becomes more significant at high doping densities. The contribution of the last term in Eq. (5) can be neglected.

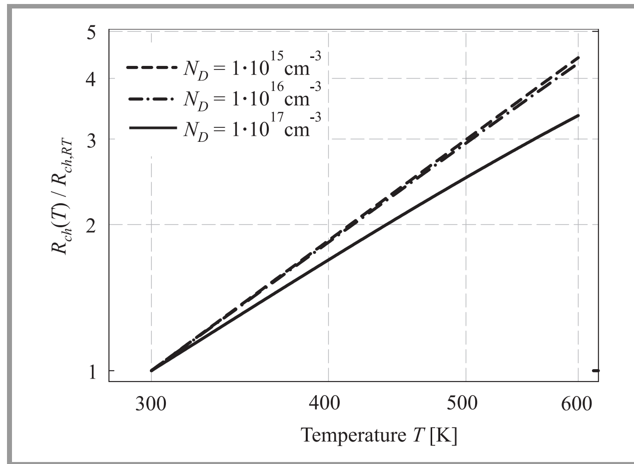


Fig. 5. Channel resistance $R_{ch}(T)$ normalized to the room temperature value $R_{ch,RT}$ as a function of the temperature calculated for different channel doping concentrations (see legend) using Eq. (5).

Low on-resistance and high output current density are two critical demands for a N-*off* design. Both require channel doping and width, and gate control voltage as high as possible. The maximum allowed control gate voltage value is however compromised by the maximum operating temperature of the device according to Eq. (1). Maximization of both the channel width and the channel doping is not possible with respect to the V_{th} condition (Eq. (2)). The optimization of the channel width and doping must be performed with respect to R_{on} and V_{th} as discussed later in this paper. At the same time the optimal choice of the doping and of the width of the channel is compromised by the actual process tolerances given by the selected technology according to Eqs. (2) and (3).

7. Simulation Study of N-*off* JFETs and All-JFET Cascodes

We have analyzed the potential for N-*off* operation of selected epitaxial vertical JFET concepts using simulations. The breakdown voltages, the specific on-state resistance, the maximum controllable current value at the applied drain voltages of 1.0 V and 10 V have been used as evaluation parameters in the temperature range from room temperature (27°C) to 250°C. The simulated structures are of both N-*on* and N-*off* type and cover the design voltage range from 600 V to 4500 V. When it comes to N-*off* behavior,

Table 1
Summary of design parameters of simulated N-*on* and N-*off* JFET and cascode structures with state of the art dimensions

Device type	B_V [V]	L_{ch} [μm]	$2T_{ch}$ [μm]	N_{ch} [cm^{-3}]	Cell pitch [μm]	N_D [cm^{-3}]	W_D [μm]
N-*off* JFETs							
DGTJFET	850	1.8	0.4	$4\cdot10^{16}$	3.6	$7.5\cdot10^{15}$	7.5
	1900	1.8	0.4	$4\cdot10^{16}$	3.6	$7.5\cdot10^{15}$	12
BGJFET	600	1.0	0.27	$4\cdot10^{16}$	2.69	$4\cdot10^{16}$	6
	1200	2.0	0.6	$1.5\cdot10^{16}$	3.0	$1.5\cdot10^{16}$	7.5
	2500	2.0	0.75	$3\cdot10^{15}$	3.15	$3\cdot10^{15}$	12.5
LCJFET1	750	1.0	0.6	$1.2\cdot10^{16}$	10	$4\cdot10^{16}$	7
10 μm	1350	1.0	0.6	$1.2\cdot10^{16}$	10	$1.5\cdot10^{16}$	7
pitch	3000	1.0	0.6	$1.2\cdot10^{16}$	10	$3\cdot10^{15}$	15
LCJFET2	750	2.0	0.6	$1.2\cdot10^{16}$	16	$4\cdot10^{16}$	7
16 μm	1350	2.0	0.6	$1.2\cdot10^{16}$	16	$1.5\cdot10^{16}$	7
pitch	3000	2.0	0.6	$1.2\cdot10^{16}$	16	$3\cdot10^{15}$	15
Cascodes							
RGJFET	all	0.35	0.35	$7\cdot10^{15}$	4.8	$7\cdot10^{15}$	2.4
BGJFET	550	0.6	2.4	$4\cdot10^{16}$	4.8	$4\cdot10^{16}$	6.4
	1250	0.6	2.4	$1.5\cdot10^{16}$	4.8	$1.5\cdot10^{16}$	11.4
	4200	0.6	2.4	$4\cdot10^{15}$	4.8	$4\cdot10^{15}$	31.4
N-*on* JFETs							
BGJFET	600	1.0	1.0	$4\cdot10^{16}$	4.0	$4\cdot10^{16}$	6
	1200	2.0	2.0	$1.5\cdot10^{16}$	5.0	$1.5\cdot10^{16}$	7.5
	2500	2.0	3.0	$3\cdot10^{15}$	5.5	$3\cdot10^{15}$	12.5
	4500	2.0	3.0	$2\cdot10^{15}$	6.0	$2\cdot10^{15}$	30

each JFET concept has its own limitations depending on the applied design and process technologies. First a number of idealized N-off BGJFET, DGTJFET, LCJFET and cascode structures with possibly wide and practically achievable channel dimensions have been designed and evaluated. The cascode structure is based on a HV BGJFET controlled by a short channel LV RGJFET (see Fig. 3(a)).

The design parameters of these structures are summarized in Table 1. Secondly the stand alone BGJFET and RGJFET structures are optimized by adjusting the channel doping separately from the drift region doping and by reducing the channel dimensions in the submicron region. The optimized stand alone JFET structures are compared to the corresponding cascode concept presented in Fig 3.

The optimization procedure is explained with reference to Fig. 6 showing the case of the N-*on* BGJFET with $V_{th} = -10$ V [16]. In general a lower channel doping density means that the potential barrier can be formed with a lower negative gate bias. This will however result in higher specific on-resistance for the structure. On the other hand too high channel doping results in reduced blocking voltage capability. This means that the optimum N_{ch} value for a given W_{ch} is the highest doping density giving full blocking voltage at specified threshold voltage, which can be seen in Fig. 6. The optimization of the integral cascodes, on the other hand, involves the selection of the largest possible spacing for the high voltage grid and of the shortest possible channel length for the low voltage JFET.

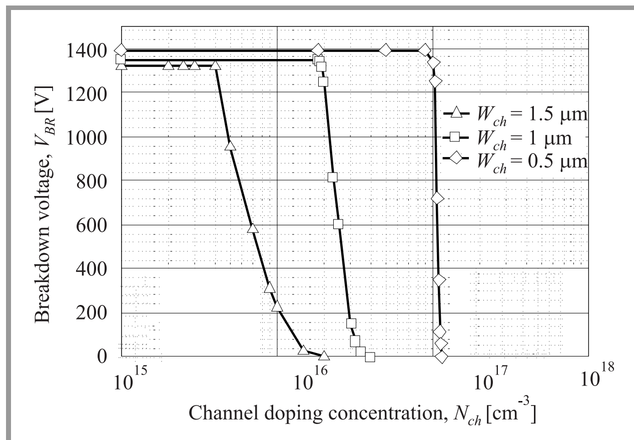


Fig. 6. Reverse blocking voltage, V_{BR} as a function of channel doping concentration, N_{ch} for N-*on* BGJFET with $V_{th} = -10$ V at 250°C.

In Fig. 7 the specific on-resistance is shown as a function of the channel length (a) for the 600 V stand alone RGJFET and RGJFET controlled cascode and (b) for the 600 V stand alone BGJFET and BGJFET controlled cascode for different channel widths. The calculations are done at 250°C and all the structures have the optimal channel doping as explained above.

The design parameters of this second set of structures are summarized in Table 2. All simulated structures satisfy the condition of the leakage current being well below 1 μA/cm^2 at zero applied gate voltage and 250°C.

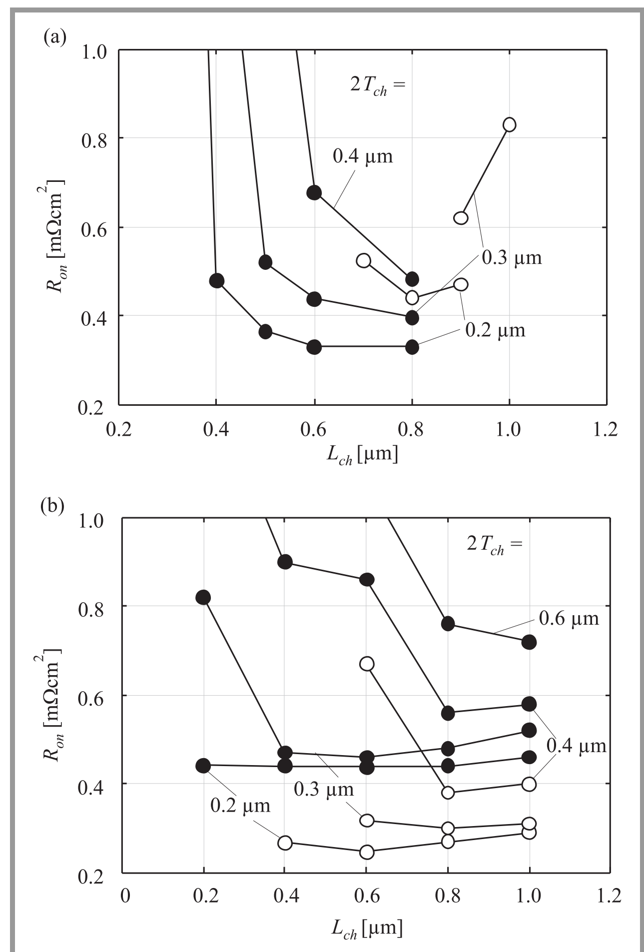


Fig. 7. Specific on-resistance for (a) the stand alone 600 V RGJFET (empty symbols) and RGJFET controlled cascode (filled symbols) and (b) the stand alone 600 V BGJFET (empty symbols) and BGJFET controlled cascode (filled symbols) as a function of channel length, L_{ch} at 25°C with channel width, $2T_{ch}$, as a parameter.

The specific on-resistance values for the state of the art structures are presented in Fig. 8 and for the submicron channel-length structures in Fig. 9.

The maximum controllable current density values are summarized in Tables 3 and 4, respectively. Only dual gate data are shown in Table 3 for LCJFETs, since the saturation current density values for all investigated LCJFET structures with 16 μm cell size and single gate drive are below 10 A/cm^2. The corresponding single gate drive current density values for the structures with 10 μm cell pitch are lower than 270 A/cm^2 and 150 A/cm^2 at 27°C and 250°C, respectively. These prohibitively low current density values make the single gate N-off LCJFETs not useful for power switching applications.

As can be seen in Fig. 8 the N-off designs have a weaker temperature dependence compared to the N-*on* designs. The on-state resistance of the N-off structures is to a large extent dominated by the channel resistance. The drift region resistance dominates in the high voltage N-*on* structures where the ideal phonon scattering related type of tem-

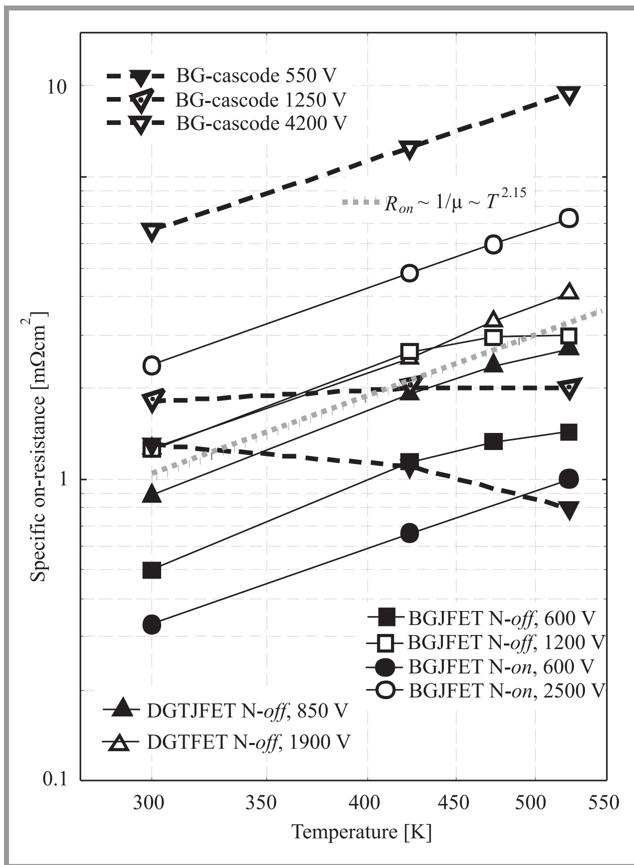


Fig. 8. Temperature dependence of the specific on-resistance for different JFET structures as described in Table 1. The on-resistance based on the phonon scattering limited carrier mobility is shown as dotted line.

Table 2
Summary of design parameters of optimized N-off JFET and cascode structures with submicron dimensions

Device type	B_V [V]	L_{ch} [μm]	$2T_{ch}$ [μm]	N_{ch} [cm^{-3}]	Cell pitch [μm]	N_D [cm^{-3}]	W_D [μm]
N-off JFETs							
RGJFET	600	0.8	0.2	$9\cdot10^{16}$	2.4	$4\cdot10^{16}$	7.0
	1200	0.8	0.2	$9\cdot10^{16}$	2.4	$1.5\cdot10^{16}$	10.0
	3000	0.8	0.2	$9\cdot10^{16}$	2.4	$3\cdot10^{15}$	15.0
BGJFET	600	0.6	0.2	$1.5\cdot10^{17}$	3.6	$4\cdot10^{16}$	7.0
	1200	0.6	0.2	$1.5\cdot10^{17}$	3.6	$1.5\cdot10^{16}$	10.0
	3000	0.6	0.2	$1.5\cdot10^{17}$	3.6	$3\cdot10^{15}$	15.0
Cascodes							
RGJFET	all	0.6	0.2	$9\cdot10^{16}$	2.4	$9\cdot10^{16}$	1.4
BGJFET	600	1.0	1.4	$4\cdot10^{16}$	2.4	$4\cdot10^{16}$	7.0
	1200	1.0	2.0	$1.5\cdot10^{16}$	3.0	$1.5\cdot10^{16}$	10.0
	3000	1.0	2.8	$3\cdot10^{15}$	3.8	$3\cdot10^{15}$	15.0
RGJFET	all	0.4	0.2	$1.6\cdot10^{17}$	3.6	$1.6\cdot10^{17}$	1.0
BGJFET	600	1.0	1.4	$4\cdot10^{16}$	3.6	$4\cdot10^{16}$	7.0
	1200	1.0	2.0	$1.5\cdot10^{16}$	3.6	$1.5\cdot10^{16}$	10.0
	3000	1.0	2.8	$3\cdot10^{15}$	3.6	$3\cdot10^{15}$	15.0

perature dependence is observed. The contribution of the highly doped substrate is not included in the data of Figs. 8 and 9. The substrate will dominate at lower voltage levels and reduce the temperature dependence due to the ionisation of dopants in the same way as demonstrated for high channel doping in Fig. 5. It is interesting to note that the simulated cascode structures show a negative temperature dependence of the on-resistance up to the design voltage of about 1200 V. This behavior is related to the potential barrier present in the case of short and relatively low doped channel in recessed gate JFET.

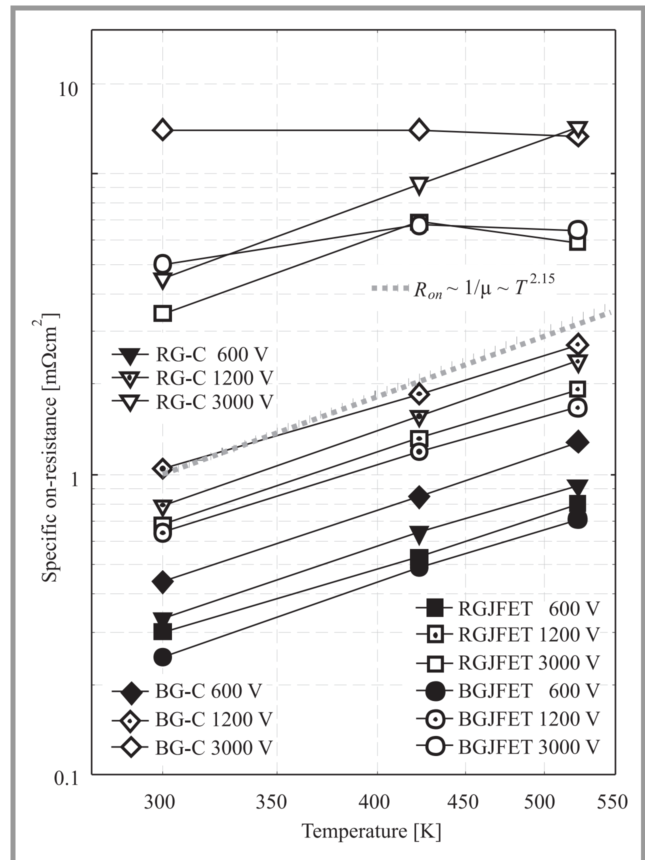


Fig. 9. Temperature dependence of the specific on-resistance for different JFET structures as described in Table 2. The on-resistance based on the phonon scattering limited carrier mobility is shown as dotted line.

The current flow in the presence of such a potential barrier is strongly temperature dependent and has a positive temperature coefficient thus compensating the resistivity increase due to the phonon scattering mechanism. Temperature dependence of the current flow over the 2D potential barrier is similar to that of the current flow through a p-n junction. The reason for the change in behavior with the increasing design voltage is the competition between the negative temperature dependence of the on-resistance due to the lowering of the potential barrier present in the channel of the recessed gate LV JFET and the positive temperature dependence due to the mobility degradation in the drift region.

Table 3
The output current densities of simulated N-*on* and N-*off* JFET and cascode structures with state of the art dimensions at drain voltages of 1 V and 10 V for different temperatures

Device type	Gate drive V_{GS} (2.4 V)	B_V [V]	27°C		150°C		250°C	
			J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]	J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]	J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]
N-*off* JFETs								
DGTJFET	DG	850	1000	2600	1000	2600	350	1000
	DG	1900	760	2600	500	2600	250	980
BGJFET	SG	600	1330	2000	830	1400	500	800
		1200	625	1100	430	750	230	380
		2500	200	285	90	150	60	190
LCJFET 10/16 μm pitch	DG	750	210/70	460/80	130/44	220/60	120/44	190/60
		1350	190/60	270/70	120/40	190/50	110/40	170/50
		3000	110/50	230/70	60/30	160/50	50/30	150/50
Cascode JFET	SG	550	450	2000	650	2200	800	2400
		1250	390	1800	440	1950	500	2000
		4200	150	1000	80	750	60	650
N-*on* JFETs								
BGJFET	SG	600	1000	1300	600	950	490	800
		1200	1200	9000	600	4500	400	3000
		2500	420	1000	200	1600	140	3200
		4500	90	700	40	350	30	250

Table 4
The output current densities of optimized N-*off* JFET and cascode structures with submicron dimensions at drain voltages of 1 V and 10 V for different temperatures

Device type	Gate drive V_{GS} (2.4 V)	B_V [V]	27°C		150°C		250°C	
			J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]	J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]	J_{DS} $V_{DS}=1$ V [A/cm^2]	J_{DS} $V_{DS}=10$ V [A/cm^2]
N-*off* JFETs								
RGJFET	SG	600	2600	4900	1600	3900	1200	3400
		1200	1380	3700	730	2900	600	2500
		3000	290	2400	145	1300	110	900
BGJFET	SG	600	3200	7700	1900	5700	1300	4700
		1200	1500	7000	800	4900	550	3900
		3000	210	2300	145	1300	125	910
Cascodes								
RGJFET/ BGJFET	SG	600	2650	7400	1500	5600	1000	4250
		1200	1200	5500	6250	4000	410	3000
		3000	220	1750	105	920	75	600
BGJFET/ BGJFET	SG	600	2100	8200	1120	5600	760	4400
		1200	880	5300	510	3500	360	2500
		3000	90	630	90	620	70	520

The drift region mobility degradation dominates for voltage designs above 1200 V while the channel barrier lowering dominates at lower design voltages. This is further illustrated by the output characteristics shown in Fig. 10.

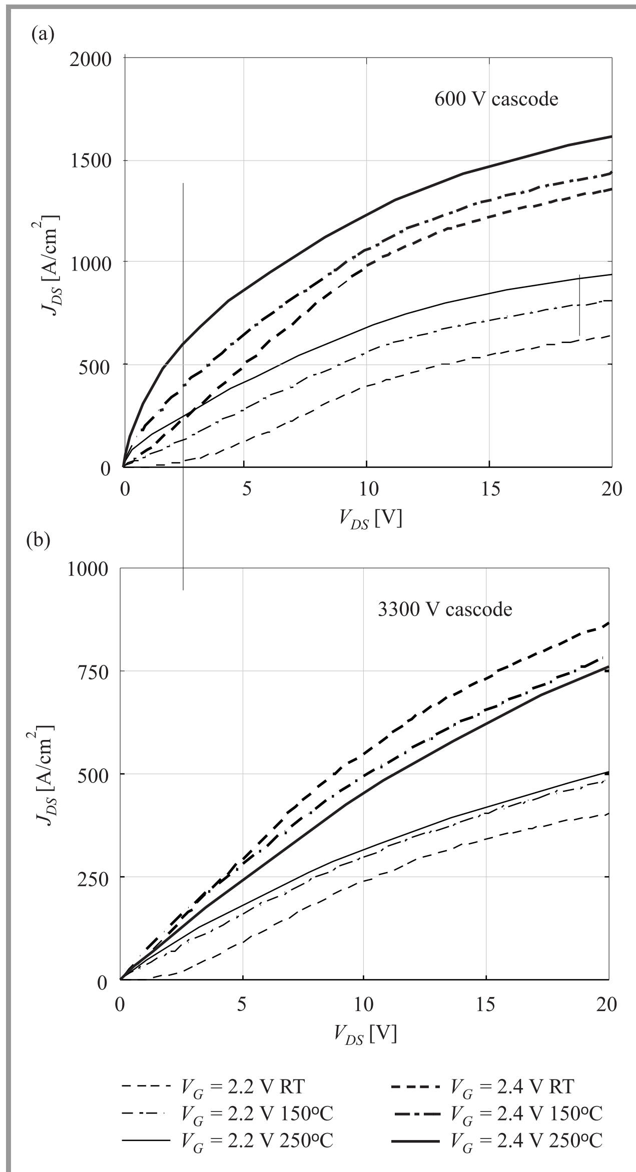


Fig. 10. Output characteristics of (a) 600 V and (b) 3.3 kV integral cascodes based on the high voltage BGJFET and controlled by low voltage RGJFET.

It can be seen that the cascode devices display a triode like type of the current voltage behavior at lower current densities, which begins to resemble pentode like output at higher current densities. The tendency towards the triode like characteristic is also more pronounced at lower temperatures and the transition to the pentode like characteristics is promoted by increased temperature. This behavior is consistent with the existence of the potential barrier in the channel of the LV RGJFET substructure. The transition to the pentode behavior is due to the biasing of the gate of the HV JFET by on-state voltage generated in the upper part of the structure.

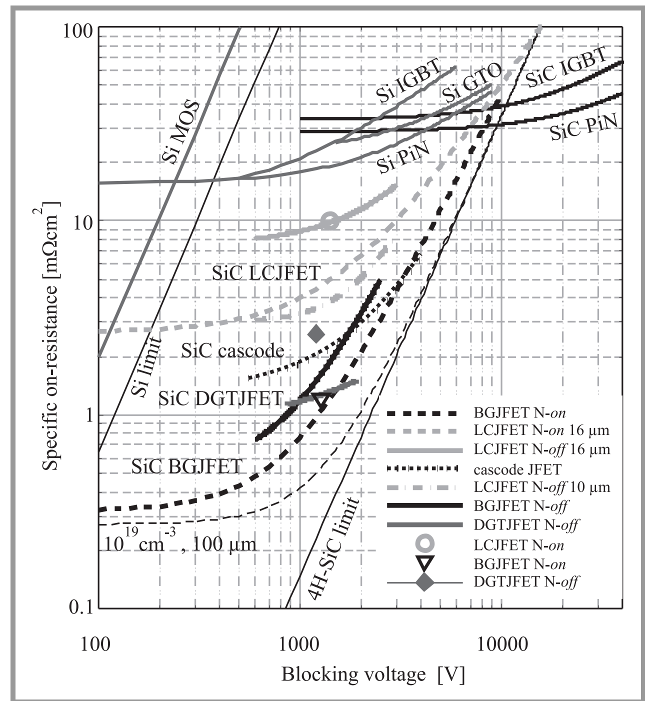


Fig. 11. Simulated and experimental (symbols) specific on-resistance versus blocking voltage for various N-*on* and N-*off* JFET and cascode structures with state of the art dimensions.

In Fig. 11 the specific on-resistance values for state of the art structures including contribution of the thinned 100 μm thick substrate are shown versus blocking voltage together with best published experimental data for LCJFET [13],

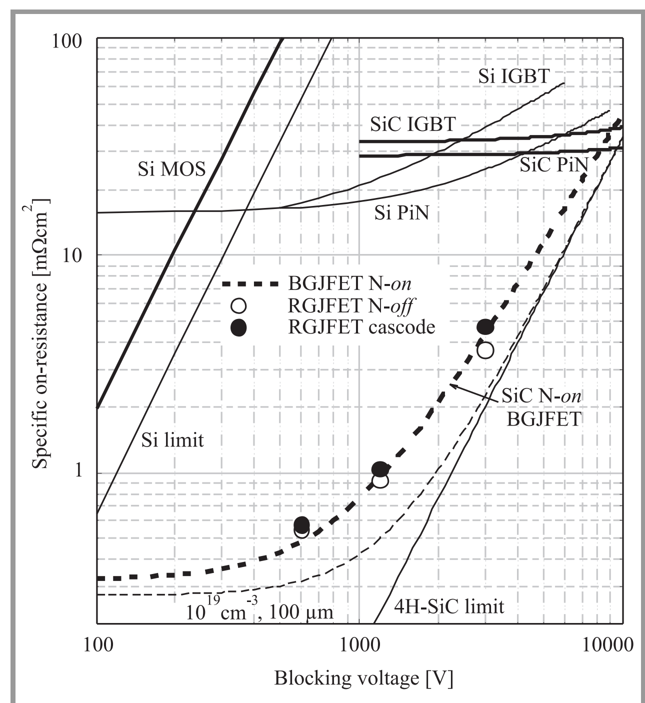


Fig. 12. Simulated specific on-resistance versus blocking voltage for stand-alone N-*off* RGJFET and RGJFET controlled cascode structures with submicron dimensions.

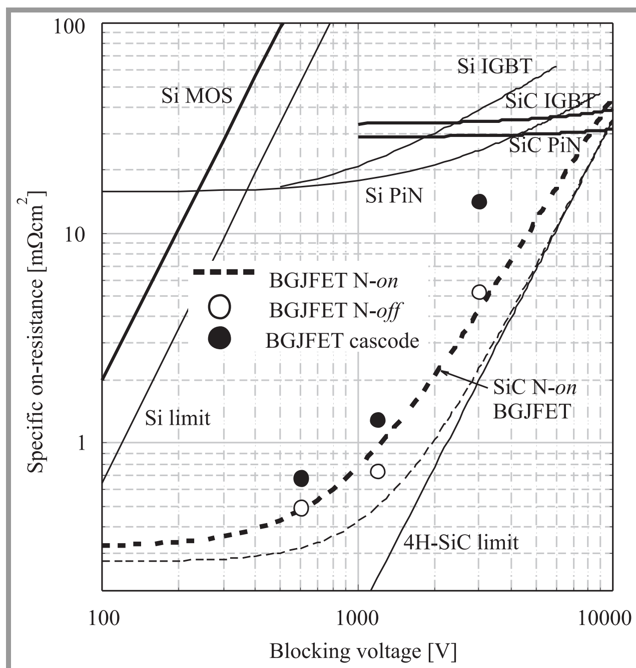


Fig. 13. Simulated specific on-resistance versus blocking voltage for stand-alone N-off BGJFET and BGJFET controlled cascode structures with submicron dimensions.

BGJFET [17], and DGTJFET [18] and theoretical curves for N-*on* LCJFET and BGJFET structures with threshold voltage of -50 V [5]. Finally, in Figs. 12 and 13 the specific on-resistance values for optimized N-off RGJFET and RGJFET controlled cascode and N-off BGJFET and BGJFET controlled cascode are shown versus blocking voltage. It is demonstrated that on-resistance values lower than those for the N-*on* BGJFET with $V_{th} = -50$ V are feasible for optimized N-off stand alone RGJFET and BGJFET structures. It can also be seen that on-resistance values comparable to those for the N-*on* BGJFET with $V_{th} = -50$ V are feasible for optimized integral cascodes of both investigated types especially in the voltage range above 1.0 kV.

8. N-off Design and Switching Considerations

The JFET concepts containing both the buried gate and the top gate have the possibility of single gate or double gate operation. In the first case only one gate is utilized for device control while the other is connected to source (ground). In the second case both gates are connected to the gate unit and used in parallel. In the case when the buried gate is connected to the source the devices have the ability to utilize the internal body diode removing the necessity of an external anti-parallel diode in many inverter and converter applications. This capacity is also inherent to the cascode concept [8].

Connecting the buried gate to the source makes switching much faster and reduces both the charge supplied by the gate unit and turn-off losses since the charge necessary to charge the Miller capacitance is supplied by the main circuit and not by the gate drive unit. The effect can be understood as a buried grid shielding of the top gate contact so that the Miller capacitance seen by the gate unit is greatly reduced.

The comparison of the turn-off switching characteristics of the single gate and the dual gate switching in the case of the N-*on* LCJFET structure is shown in Fig. 14. The gate charge Q_g and the turn-off losses E_{off} are $3.3 \cdot 10^{-6}$ C and $8.3 \cdot 10^{-2}$ J for the top gate control and $2.1 \cdot 10^{-7}$ C and $3.5 \cdot 10^{-3}$ J for the double gate control case, respectively. The driving conditions are the same. The conclusion from Fig. 14 is that the N-off LCJFET will necessarily suffer from slow switching speed since it requires dual gate control in order to pass reasonable forward currents as can be seen in Table 3.

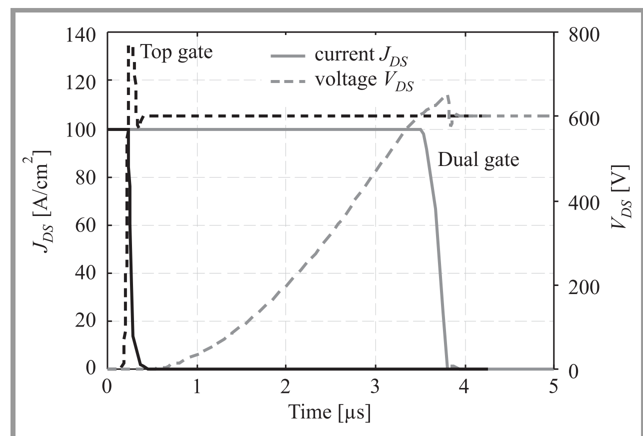


Fig. 14. A comparison of the top gate and double gate switching characteristics of the N-*on* LCJFET structure with $V_{th} = -50$ V and 16 μm cell pitch. Turn-off in inductive circuit with $R_G = 50$ Ω.

In Fig. 15 the same comparison is done for the N-off DGTJFET structure and N-off LCJFET structure with cell pitch of 10 μm. The channel length is 1.0 μm in both cases. The gate charge Q_g and the turn-off losses E_{off} are $1.6 \cdot 10^{-6}$ C and $7.5 \cdot 10^{-3}$ J for the DGTJFET with buried gate control, $1.2 \cdot 10^{-6}$ C and $1.3 \cdot 10^{-2}$ J for the DGTJFET with top gate control, $2.5 \cdot 10^{6}$ C and $2.2 \cdot 10^{-2}$ J for the DGTJFET with buried and top gate control, and $3.3 \cdot 10^{-6}$ C and $8.3 \cdot 10^{-2}$ J for the N-off LCJFET with dual gate control under the same driving conditions. It shows that the sacrifice of the switching speed is much less severe in the case of the DGTJFET structure due to the much smaller cell pitch and thus the smaller Miller capacitance attributed to both the buried and top gates.

The turn-on of N-off JFET devices tends to be slow depending on the very low voltage swing between the threshold voltage and the maximum allowed gate voltage given by the V_{bi} available for charging the gate. Due to the low charging voltage the channel resistance plays a crucial role in controling the speed of the turn-on process as well as the RC constant of the device. Based on numerical simulations, the increase of the channel doping in the range

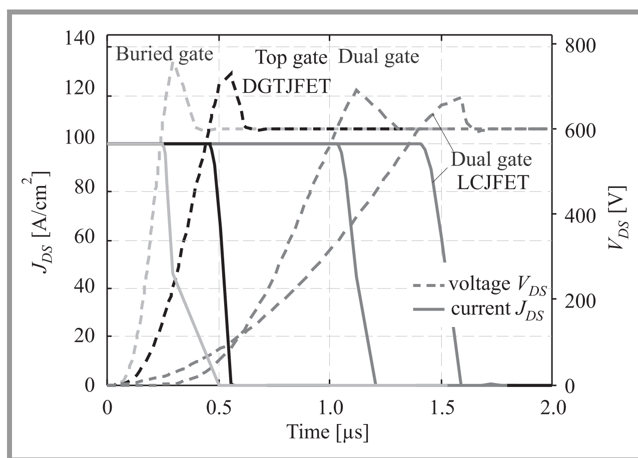


Fig. 15. A comparison of the switching characteristics of the N-off DGTJFET using the buried gate, the top gate and both gates with a N-off LCJFET with 10 μm cell pitch and double gate control. Turn-off in inductive circuit with $R_G = 2\ \Omega$.

$1 \cdot 10^{16}$ cm^{-3} to $1 \cdot 10^{17}$ cm^{-3} reduces the turn-on time (and turn-on losses) during inductive switching of the DGTJFET by one order of magnitude [18]. The effect of the channel doping on the top gate capacitance and turn-off switching is opposite, however the influence of the increased gate capacitance due to the increased channel doping on the turn-off time and turn-off losses is much smaller [19]. The requirement of maximizing the channel doping for the increased turn-on speed of the N-off device coincides with the requirements for the improved high temperature operation as discussed earlier.

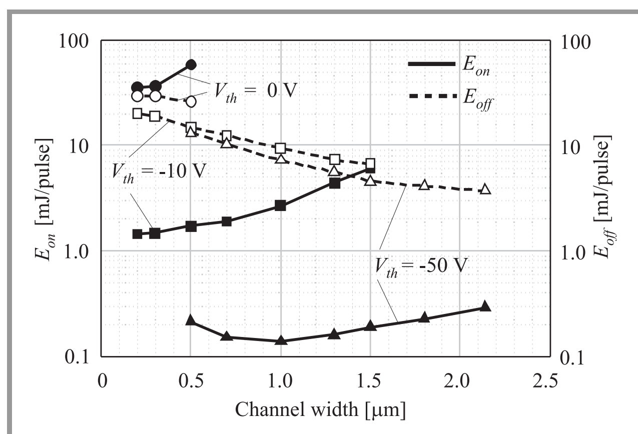


Fig. 16. Turn-on and turn-off energy for N-on and N-off 1.2 kV BGJFET structures with $I_{DS} = 100$ A/cm^2, $V_{DS} = 600$ V and $R_G = 1\ \Omega$ as a function of the channel width at 250°C.

The example of switching energy dependence on the device design is given in Fig. 16. The turn-on and turn-off energy is shown for 1.2 kV N-on BGJFET designs with threshold voltage values of $V_{th} = -50$ V and -10 V and for N-off BGJFET design ($V_{th} = 0$ V) as a function of channel width at switching conditions 100 A/cm^2, 600 V with $R_G = 1\ \Omega$ at 250°C [16]. The devices have a channel length of 1.6 μm. Switching properties and switching losses E_{on} and E_{off} are related to channel doping concentration, N_{ch} and channel width, W_{ch} as discussed above. In addition, with conventional gate drive, the charging time of JFET capacitances depends strongly on the available gate driving voltage which is related to the V_{th} value. Because of that the N-on BGJFET devices will have significantly lower turn-on and turn-off losses compared to the N-off design. The difference in the turn-on losses is especially large and more than two orders of magnitude between the N-on device with $V_{th} = -50$ V and N-off device. The difference in turn-off losses is at the same time about a factor of two.

9. Summary and Conclusions

Prospects of N-off SiC JFET switch are studied by simulation. Several possible epitaxial switch designs are reviewed and compared. The influence of the device design on the conduction and switching performance is investigated in the wide range of N-on to N-off designs with the threshold voltage values from -50 V to 0 V. The advantages of the epitaxial JFET designs for high voltage high temperature operation in the voltage and temperature range of 0.6 kV to 3.0 kV and 25°C to 250°C, respectively, are demonstrated.

The relative merits and limitations of LCJFET, BGJFET and DGTJFET concepts are clearly demonstrated. The stand alone RGJFET and BGJFET concepts are compared to the integral cascode solutions. It is shown that integral cascodes can be optimized to be competitive with and superior to stand-alone switches. It is furthermore demonstrated that the utilization of the N-off JFET and JFET cascode switch requires development of efficient gate driving methods to overcome the limitation of slow turn-on.

References

[1] P. Friedrichs, H. Mitlehner, K. O. Dohnke, D. Peters, R. Schörner, U. Weinert, E. Baudelot, and D. Stephani, “SiC power devices with low on-resistance for fast switching applications”, in *Proc. 12th Int. Symp. Pow. Semicond. Dev. ICs*, Toulouse, France, 2000, p. 213.

[2] Y. Tanaka, M. Okamotu, A. Takatsuka, K. Arai, T. Yatsuo, K. Yano, and M. Kasuga,“700 V 1.0 m$\Omega \cdot$cm^2 buried gate SiC-SIT (SiC-BGSIT)”, *IEEE Electron Dev. Lett.*, vol. 27, no. 11, pp. 908–910, 2006.

[3] R. K. Malhan, Y. Takeuchi, M. Kataoka, A.-P. Mihaila, S. J. Rashid, F. Udrea, and G. A. J. Amaratunga, “Normally-off trench JFET technology in 4H silicon carbide”, *Microelectron. Eng.*, vol. 83, iss. 1, pp. 107–110, 2006.

[4] S. Krishnaswami, A. Agarwal, S. H. Ryu, C. Capell, J. Richmond, J. Palmour, S. Balachandran, T. P. Chow, S. Bayne, B. Gail, K. Jones, and C. Scozzie, “1000 V, 30 A 4H-SiC BJTs with high current gain”, *IEEE Electron Dev. Lett.*, vol 26, no. 3, pp. 175–177, 2005.

[5] M. Bakowski, “Status and prospects of SiC power devices”, *IEEE Trans. Ind. Appl.*, vol. 126, no. 4, pp. 391–399, 2006.

[6] R. K. Malhan, H. Nakamura, S. Onda, D. Nakamura, and K. Hara, “Impact of SiC structural defects on the degradation phenomenon of bipolar SiC devices”, *Mater. Sci. Forum*, vol. 433–436, pp. 917–920, 2003.

[7] B. Ållebrand and H.-P. Nee, "On the possibility to use SiC JFETs in power electronic circuits", in *Proc. 9th Conf. Pow. Electron. Appl., EPE'2001*, Graz, Austria, 2001.

[8] M. Bakowski and U. Gustafsson, "Unipolar and bipolar SiC integral cascoded switches with MOS and junction gate – simulation study", *Mater. Sci. Forum*, vol. 389–393, pp. 1321–1324, 2002.

[9] M. L. Heldwein and J. W. Kolar, "A silicon carbide JFET gate driver circuit allowing short commutation times for sparse matrix converter applications", in *Proc. Nineteenth Ann. IEEE Appl. Pow. Electron. Conf. Expos.*, Anaheim, USA, 2004, vol. 1, pp. 116–121.

[10] P. Friedrichs, H. Mitlehner, K. W. Bartsch, O. Dohnke, R. Kattschmidt, U. Weinert, B. Weis, and D. Stephani, "Static and dynamic characteristic of 4H-SiC JFETs designed for different blocking categories", *Mater. Sci. Forum*, vol. 338–342, pp. 1243–1246, 2000.

[11] M. Bakowski, "Analysis of unipolar and bipolar SiC cascoded switches with MOS gate", *Mater. Sci. Forum*, vol. 433–436, pp. 801–804, 2003.

[12] V. Veliadis, T. McNutt, M. Snook, H. Hearne, P. Potyraj, J. Junghans, and C. Scozzie, "Large area silicon carbide vertical JFETs for 1200 V cascode switch operation", *Int. J. Pow. Manag. Electron.*, vol. 2008, art. id. 523721, 8 p., 2008.

[13] D. Stephani and P. Friedrichs, "Silicon carbide junction field effect transistors", *Int. J. High Speed Electron. Syst.*, vol. 16, no. 3, pp. 825–854, 2006.

[14] M. Bakowski, U. Gustafsson, and U. Lindefelt, "Simulation of SiC high power devices", *Phys. Stat. Sol. A*, vol. 162, pp. 421–440, 1997.

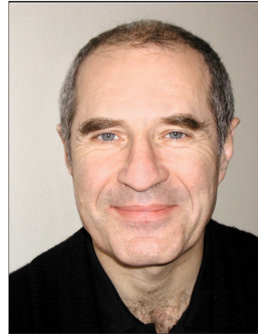
[15] S. M. Sze, *Physics of Semiconductor Devices*. New York: Wiley, 1981.

[16] J. K. Lim and M. Bakowski, "Analysis of 1.2 kV SiC buried grid VJFETs" (to be published in *Phys. Scripta T*, 2009).

[17] Y. Tanaka, K. Yano, M. Okamoto, A. Takatsuka, K. Arai, and T. Yatsuo, "1270 V, 1.21 mΩcm^2 SiC buried gate static induction transistors (SiC-BGSITs)", *Mater. Sci. Forum*, vol. 600–603, pp. 1071–1074, 2009.

[18] R. K. Malhan, M. Bakowski, Y. Takeuchi, N. Sugiyama, and A. Schöner, "Design, process, and performance of all-epitaxial normally-off SiC JFETs", *Phys. Stat. Sol. A*, vol. 206, iss. 10, pp. 2308–2328, 2009.

[19] R. K. Malhan, S. J. Rashid, M. Kataoka, Y. Takeuchi, N. Sugiyama, F. Udrea, G. A. J. Amaratunga, and T. Reimann, "Switching performance of epitraxially grown normally-off 4H-SiC JFET", *Mater. Sci. Forum*, vol. 600–603, pp. 1067–1070, 2009.



Mietek Bakowski was born in Bydgoszcz, Poland, in 1946. He completed M.Sc. studies at the Faculty of Electronics, Warsaw University of Technology, in 1969. He received his Ph.D. and the Assistant Professor competence from the Chalmers University of Technology, Gothenburg, Sweden, in 1974 and in 1981, respectively, and works presently as a senior expert at the Acreo AB, Kista, Sweden. He has worked with the development of silicon bipolar and BiMOS power devices and since 1994 with the design, simulation and electrical evaluation of SiC power devices. In 2000–2003 he has been appointed Adjunct Professor at the Royal Institute of Technology, KTH, Kista.

e-mail: mietek.bakowski@acreo.se
Acreo AB
Electrum 236
SE-164 40 Kista, Sweden

Invited paper

Variation Analysis of CMOS Technologies Using Surface-Potential MOSFET Model

Hans Jürgen Mattausch, Akihiro Yumisaki, Norio Sadachika, Akihiro Kaya, Koh Johguchi, Tetsushi Koide, and Mitiko Miura-Mattausch

Abstract— An analysis of the measured macroscopic within-wafer variations for threshold voltage (V_{th}) and on-current (I_{on}) over several technology generations (180 nm, 100 nm and 65 nm) is reported. It is verified that the dominant microscopic variations of the MOSFET device can be extracted quantitatively from these macroscopic variation data by applying the surface-potential compact model Hiroshima University STARC IGFET model 2 (HiSIM2), which is presently brought into industrial application. Only a small number of microscopic parameters, representing substrate doping (NSUBC), pocket-implantation doping (NSUBP), carrier-mobility degradation due to gate-interface roughness (MUESR1) and channel-length variation during the gate formation (XLD) are found sufficient to quantitatively reproduce the measured macroscopic within-wafer variations of V_{th} and I_{on} for all channel length L_g and all technology generations. Quantitative improvements from 180 nm to 65 nm are confirmed to be quite large for MUESR1 (about 70%) and $L_{\min}$(XLD) (55%) variations, related to the gate-oxide interface and the gate-stack structuring, respectively. On the other hand, doping-related technology advances, which are reflected by the variation magnitudes of NSUBC (30%) and NSUBP (25%), are found to be considerably smaller. Furthermore, specific combinations of extreme microscopic parameter-variation values are able to represent the boundaries of macroscopic fabrication inaccuracies for V_{th} and I_{on}. These combinations are found to remain identical, not only for all L_g of a given technology node, but also for all investigated technologies with minimum L_g of 180 nm, 100 nm and 65 nm.

Keywords— compact model, fabrication inaccuracy, field-effect transistor, macroscopic, microscopic, potential at channel surface, silicon, within wafer.

1. Introduction

As the dimensions of metal oxide semiconductor field effect transistors (MOSFETs) are scaled down, the effects of microscopic fabrication inaccuracies due to process variations are increasingly affecting the macroscopic variation of MOSFET performances and turn out to be increasingly difficult to manage and mitigate [1]. Experimental and theoretical variation analysis is normally based on macroscopic MOSFET properties like threshold voltage V_{th} or on-current I_{on}. Determining the correlation of the macroscopic variation data with microscopic MOSFET properties, like doping concentrations, structure parameters or carrier mobility, has been always a difficult task. An important reason is that the standard V_{th}-based compact models for circuit simulation, like BSIM4 [2], don't provide a sufficiently physical correlation between macroscopic MOSFET performance and microscopic MOSFET parameters, so that numerical technology computer aided design (TCAD) software has to be applied for this purpose. One recent method for improving this situation is the application of predictive technology models (PTM) in combination with a V_{th}-based compact model [3], but having the predictability already included in the compact model is of course preferable. Many studies also focus on just one type of the macroscopic variation aspects like V_{th} [4], because a comprehensive analysis of several macroscopic-variation aspects is more difficult and hard to accomplish.

We report an analysis of V_{th} and I_{on} variations over 3 process generations and verify that microscopic variations can be extracted by applying surface-potential compact models like Hiroshima University STARC IGFET model 2 (HiSIM2) [5], [6], which are presently in the process of being brought into industrial application. Only 4 microscopic parameters, which are related to substrate doping (NSUBC), pocket-implantation doping (NSUBP), gate-interface-roughness carrier-mobility degradation (MUESR1) and channel-length variation during gate structuring (XLD) are found sufficient to quantitatively reproduce within-wafer variation of V_{th} and I_{on} over all channel length (L_g) and all 3 technology generations with a minimum L_g of 180 nm, 100 nm and 65 nm. All of the 3 analyzed technologies apply a pocket-implantation technology for suppressing the performance degradations due to the short-channel effects and are chosen from different manufactures to obtain a more general picture of the technology trends.

2. Surface-Potential Compact Models

A new generation of compact models for circuit simulation, like HiSIM2 [5], [6] or PSP [7], uses a model concept

based on the surface-potential ϕ_S in the MOSFET channel below the gate oxide and incorporates the complete drift-diffusion theory for the MOSFET device [6], [8]. These surface-potential compact models replace the regional approach of V_{th}-based models like BSIM4 with a single equation, which is valid for all operating conditions and whose parameters have a close correlation to the structural and physical MOSFET parameters.

For example, the V_{th} variation due to the substrate doping NSUBC (called NDEP in BSIM4) of a long and wide MOSFET, depicted for BSIM4 and HiSIM2 in Fig. 1,

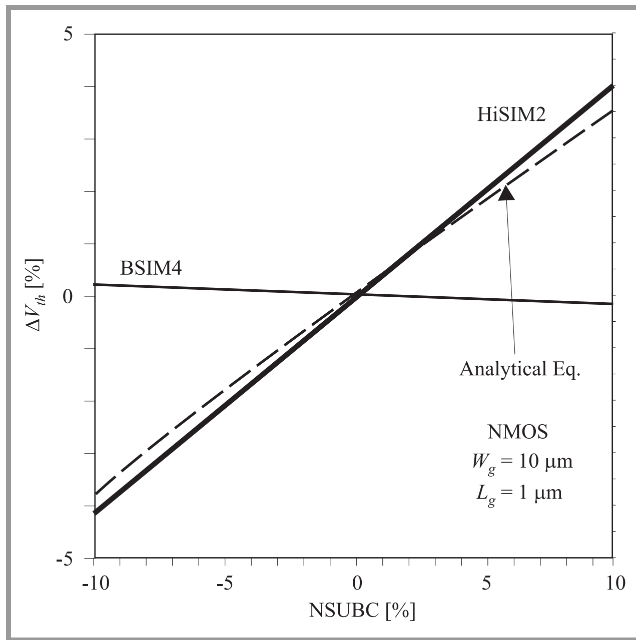


Fig. 1. Comparison of V_{th} as a function of substrate impurity variations for a long and wide MOSFET as predicted by the surface-potential model HiSIM2 and by the threshold-voltage-based model BSIM4 for a 100 nm technology. The result with the conventional analytical V_{th} equation, which neglects the pocket-implantation contributions, is also shown by a dashed line.

with model parameters extracted from a 100 nm CMOS technology, is obviously not correctly reproduced by the V_{th}-based model BSIM4. On the other hand, HiSIM2 is close to the standard analytical solution with a slight deviation explained by the fact that the analytical solution neglects the pocket implantation. Correct scaling properties of the HiSIM2 model parameters are confirmed with 3 wafers fabricated in the same 100 nm process, but with deliberately different substrate doping.

Figure 2 verifies, that the extraction of HiSIM2 parameters from the typically-implanted wafer's (wafer CCC) nominal chip is sufficient to predict the $I_{on} - V_{th}$ characteristics of the 2 differently implanted wafers, namely wafer FFC with reduced substrate doping and wafer SSC with increased substrate doping, by only adjusting the respective parameter NSUBC in the HiSIM2 model appropriately. This predictability of MOSFET characteristics through adjustment of the correlated physical parameter is even true for current derivatives with respect

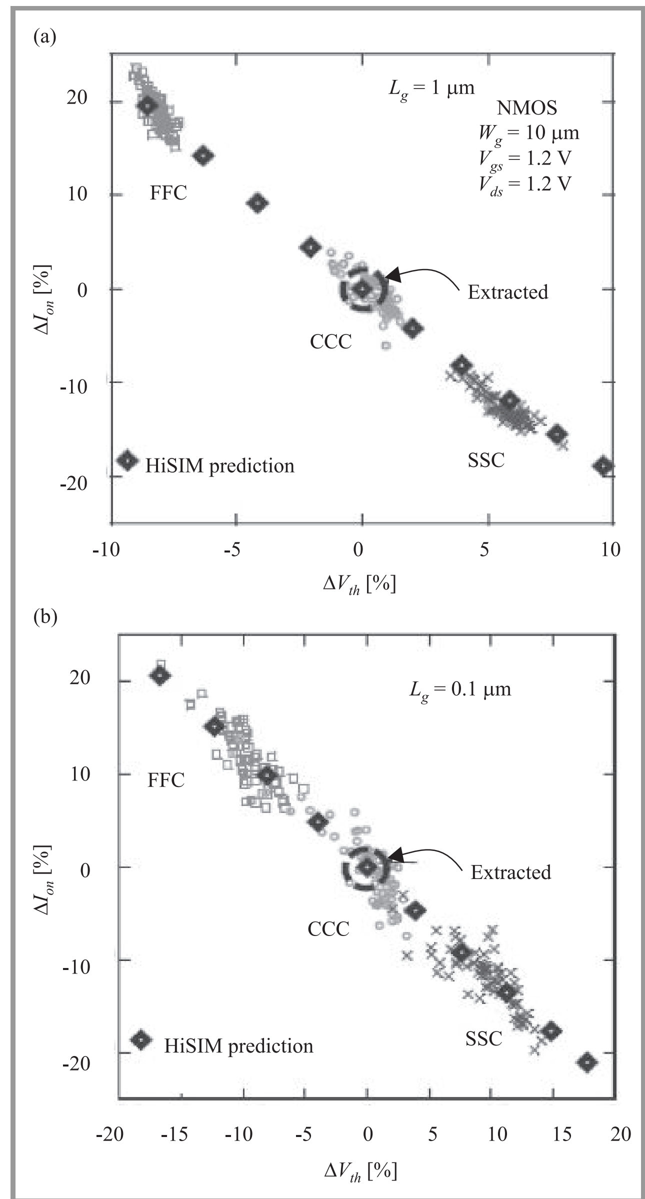


Fig. 2. Prediction of the inter-wafer variations of V_{th} and I_{on} due to the microscopic variation of the substrate doping (NSUBC) with the surface-potential-based compact model HiSIM2 for $L_g = 1$ μm (a) and for $L_g = 0.1$ μm (b). Measurements are shown by square symbols (wafer FFC with reduced substrate doping), circle symbols (wafer CCC with typical substrate doping) and cross symbols (wafer SSC with increased substrate doping), respectively. The model parameter extraction has been performed for the chip denoted by "extracted" from the wafer denoted by CCC. The large black symbols are calculated results with the surface-potential model HiSIM2 by only varying the NSUBC parameter in the extracted model card of parameters.

to V_{gs} (called transconductance g_m) or V_{ds} (called output conductance g_{ds}) as verified in Fig. 3 for the transconductance g_m.

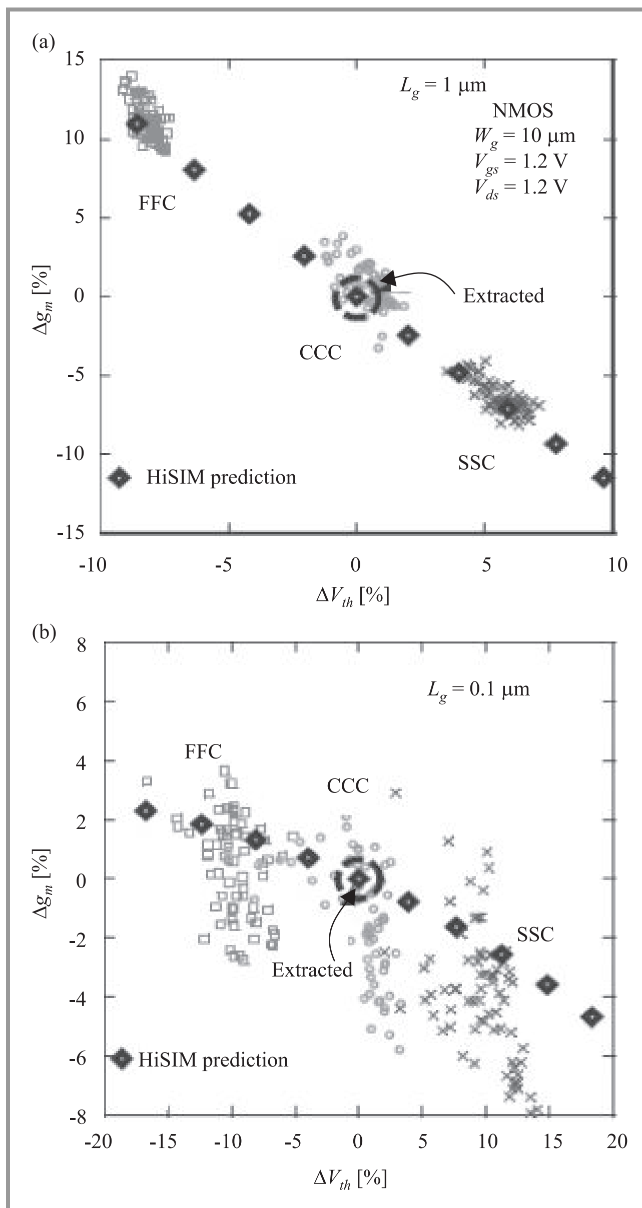


Fig. 3. Comparison of predicted and measured transconductance g_m variations on 3 different wafers (FFC, CCC, and SSC) with different substrate impurity concentrations (modeled with the parameter NSUBC) for $L_g = 1\ \mu$m (a) and $L_g = 0.1\ \mu$m (b). The meanings of the symbols are the same as in Fig. 2. These data verify the predictive capabilities of the surface-potential-based compact model HiSIM2 even for the inter-wafer variations of the V_{gs} derivatives of macroscopic $I-V$ characteristics of the MOSFET.

3. Sensitivity of Model Parameters and Variation-Extraction Method

For correlating the macroscopic variation of the MOSFET characteristics to the microscopic physical MOSFET parameters of the surface-potential model, it is necessary to determine those parameters which most sensitively influence the MOSFET characteristics.

Figure 4 shows the parameter-sensitivity analysis of simulated V_{th} and I_{on} for an 180 nm technology. The results

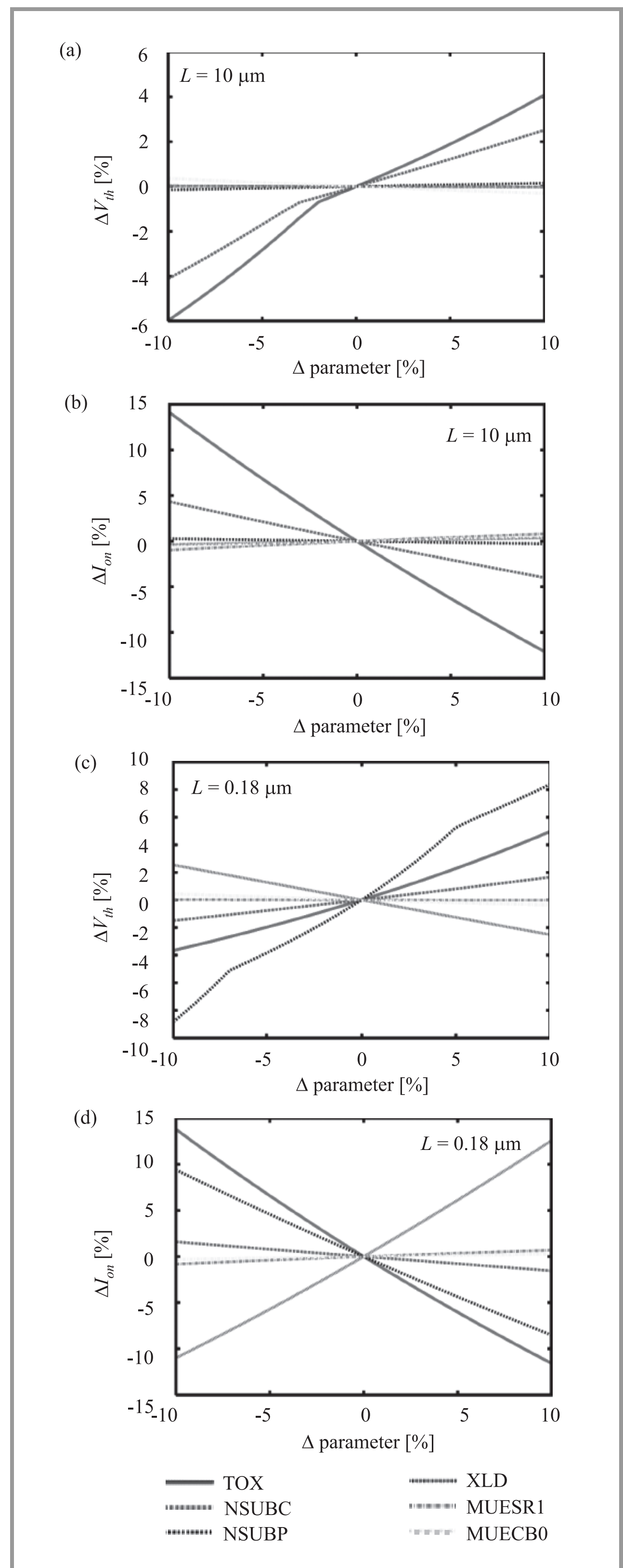


Fig. 4. Sensitivity of surface-potential model parameters for the threshold voltage V_{th} (a), (c) and the on-current I_{on} (b), (d) of NMOSFETs fabricated in a 180 nm CMOS technology. The most sensitive parameters are different for MOSFETs with long channel length $L_g = 10\ \mu$m (a), (b) and short channel length $L_g = 0.18\ \mu$m (c), (d).

for NMOSFETs with long channel lengths ($L_g = 10\ \mu$m) and short channel length ($L_g = 0.18\ \mu$m) are listed in Fig. 5 and show a large TOX sensitivity in all cases. However, the other sensitive parameters are different for $L_g = 10\ \mu$m (NSUBC, MUESR1, MUECB0) and $L_g = 0.18\ \mu$m (XLD, NSUBP). For the average oxide thickness (determining variation of TOX) and the substrate-lattice quality (determining the variation of MUECB0), process control during fabrication is known to be very good. In particular, the average TOX variation within a wafer is usually kept much smaller than an atomic layer, which means that variations of the properties of the gate-oxide interface can be expected to be of main importance in practice. Therefore, we first restrict the variation analysis of the microscopic parameters to the other 4 parameters NSUBC, MUESR1, NSUBP and XLD, to see whether this is already sufficient for reproducing the measured macroscopic V_{th} and I_{on} variations on the basis of the microscopic parameters in the HiSIM2 compact model.

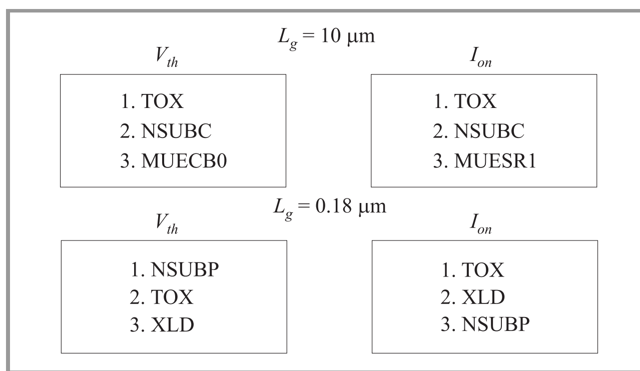
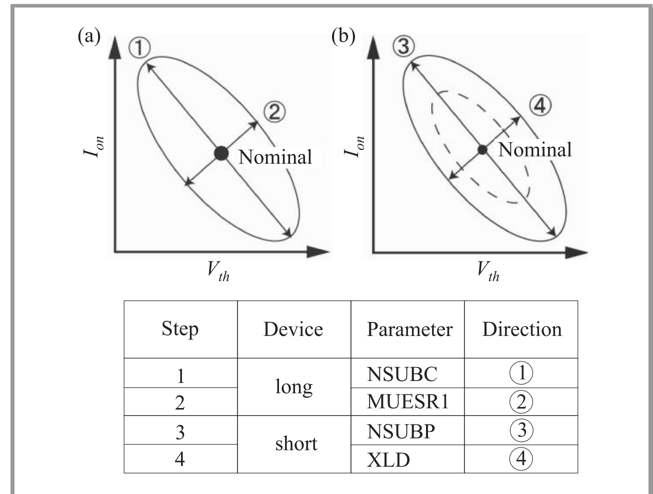


Fig. 5. Most sensitive model parameters from the sensitivity analysis of Fig. 4. The sensitivity with respect to gate-oxide thickness (TOX) is large in all cases. Other sensitive parameters are different for long channel ($L_g = 10\ \mu$m) and short channel ($L_g = 0.18\ \mu$m) MOSFETs. For $L_g = 10\ \mu$m, substrate doping (NSUBC) as well as the mobility parameters for phonon scattering (MUECB0) and surface roughness (MUESR1) are sensitive. For $L_g = 0.18\ \mu$m, the pocket doping (NSUBP) and the fabrication-process-related channel-length change (XLD) in comparison to the designed channel length are sensitive.

For carrying out the extraction procedure of microscopic variations, a typical die is first selected based on the condition that MOSFETs on this die occupy a central position in the measured $I_{on} - V_{th}$ variation for all channel length L_g. Then a nominal HiSIM2 reference-parameter set is extracted, which reproduces the $I - V$ characteristics of the MOSFETs on this selected die for all L_g. The variation of microscopic HiSIM2 parameters with respect to this nominal parameter set is then determined from the measured within-wafer $I_{on} - V_{th}$ variation data according to the 4-step procedure of Fig. 6. Taking into account the L_g-dependence of the respective sensitivities and their influence on the distribution shape, the microscopic variations are first determined for NSUBC and MUESR1 form the long L_g data and then for NSUBP and XLD from the short L_g data. By applying this method, changes in the effective channel doping concentration and the effective



Step	Device	Parameter	Direction
1	long	NSUBC	①
2		MUESR1	②
3	short	NSUBP	③
4		XLD	④

Fig. 6. Extraction strategy of the microscopic parameter variation from the measured macroscopic V_{th} and I_{on} variation data. As the process control is very good for oxide thickness and phonon scattering, these 2 parameters are assumed to have negligible influence on macroscopic variation and a 2 step procedure for extracting the variation of the other sensitive parameters is applied. In the first step the variation of substrate doping (NSUBC) and surface roughness (MUESR1) are determined from the long-channel data (a). In the second step the short-channel data (b) are used to extract the variation of pocket doping (NSUBP) and fabrication related channel-length change (XLD).

mobility degradation as a function of L_g in comparison to their long L_g values, represented by the NSUBC and MUESR1 variations, are correctly captured with the surface-potential model HiSIM2.

4. Microscopic-Variation Extraction for 3 Technology Generations

The microscopic within-wafer parameter variations of NMOSFETs for the 180 nm, 100 nm and 65 nm technology generations have been analyzed according to the method described in Section 3. A nominal die of each wafer has been selected for each technology generation, based on the condition that the MOSFETs on this die occupy a central position (determined by the median values) in the measured $I_{on} - V_{th}$ variation for all L_g. The HiSIM2 reference-parameter set has then been obtained from a fit to the electrical MOSFET characteristics on this selected die. The fitting results of the $I - V$ characteristics of the 100 nm technology are shown for long and short channel length L_g in Fig. 7 as an example for the good reproduction of the nominal electrical MOSFET characteristics. With this reference-parameter set, the variation of microscopic parameters has then been deter-

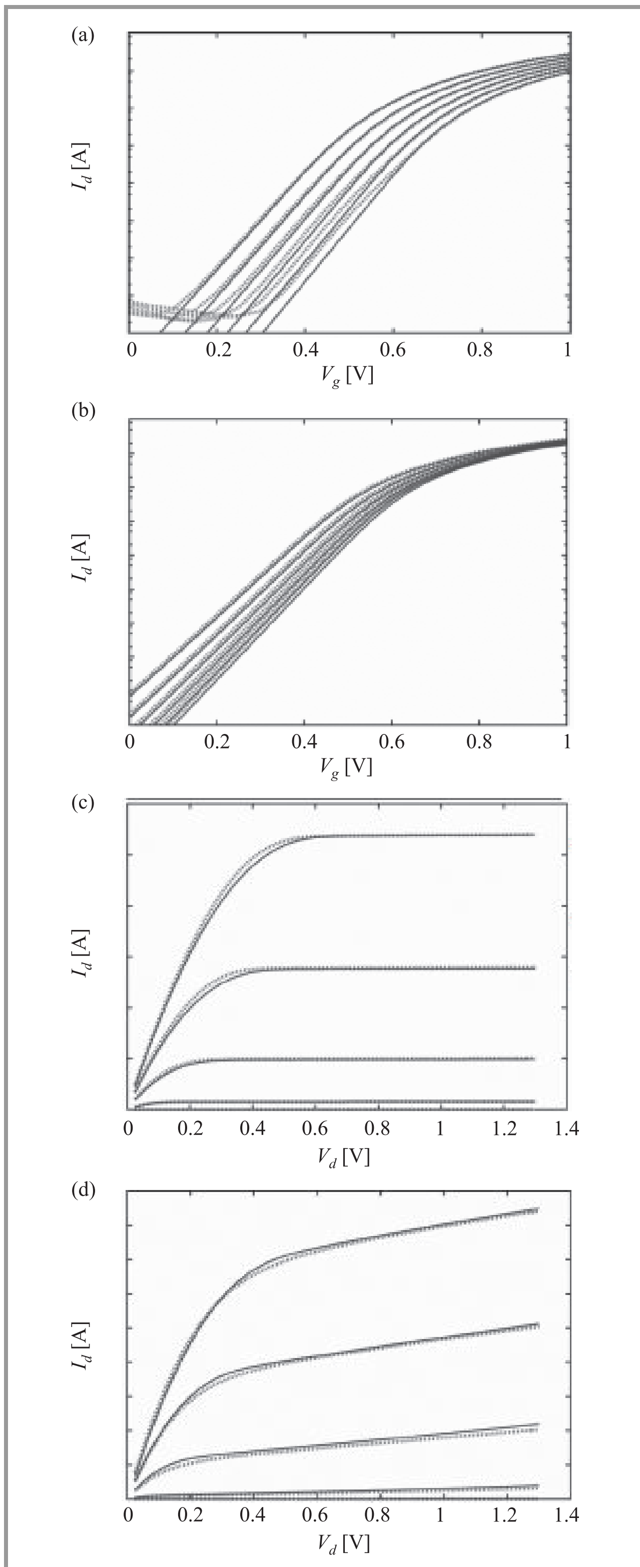
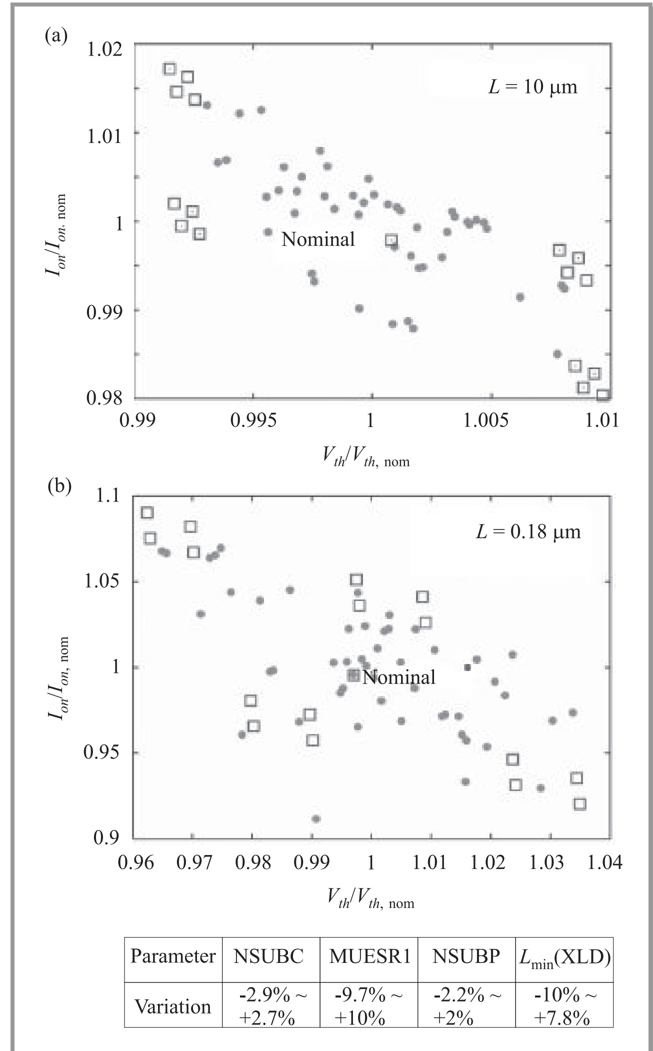


Fig. 7. Example of the parameter extraction results for the selected nominal MOSFET of the analyzed 100 nm CMOS technology by fitting the parameters of the surface-potential model HiSIM2 so that the measured $I-V$ data (solid lines) are accurately reproduced by the simulated data with the compact model (dotted lines). $I_d - V_g$ plots are on logarithmic scale for long L_g in (a), short L_g in (b) and on linear scale for long L_g in (c), short L_g in (d).

mined from the measured within-wafer $I_{on} - V_{th}$ variation according to the 4-step procedure of Fig. 6 as explained in the previous section.

The extracted microscopic variations are shown in Figs. 8, 9, and 10 for the 180 nm, 100 nm, and 65 nm technologies, respectively.

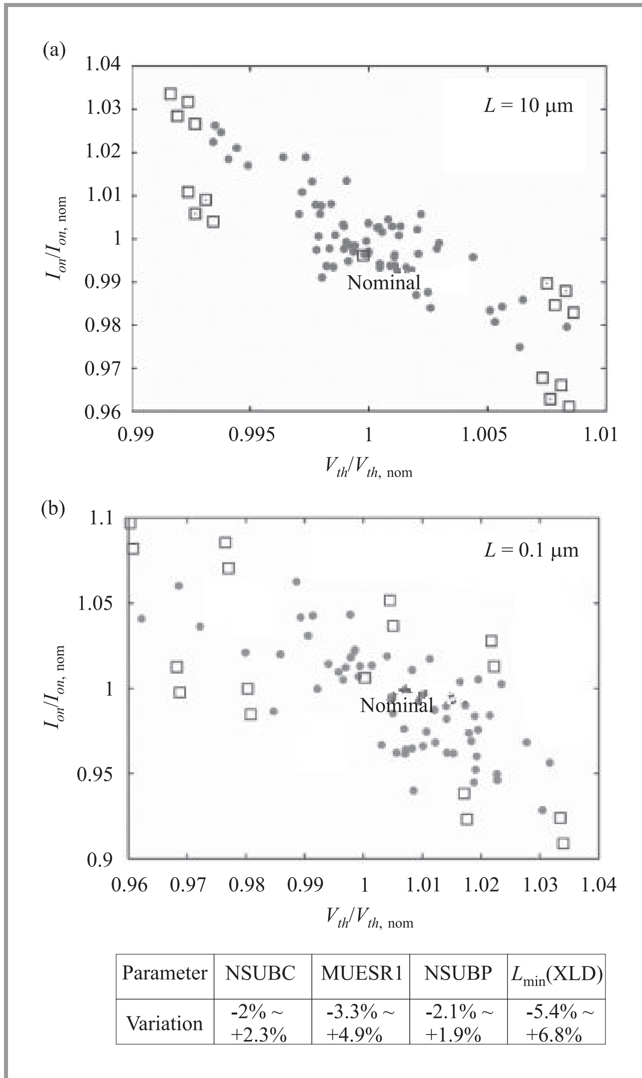


Parameter	NSUBC	MUESR1	NSUBP	L_{min}(XLD)
Variation	-2.9% ~ +2.7%	-9.7% ~ +10%	-2.2% ~ +2%	-10% ~ +7.8%

Fig. 8. Variation extraction of the most sensitive microscopic parameters for the NMOSFETs of a 180 nm technology applying the pocket implantation. Data points in the $I_{on} - V_{th}$ graphs for long (a) ($L_g = 10\ \mu$m) and short (b) ($L_g = 0.18\ \mu$m) MOSFETs are plotted on a scale relative to the selected nominal measured devices. Points show measured data and open squares show the nominal compact model data as well as all 16 combinations of the variation boundaries for the microscopic parameters.

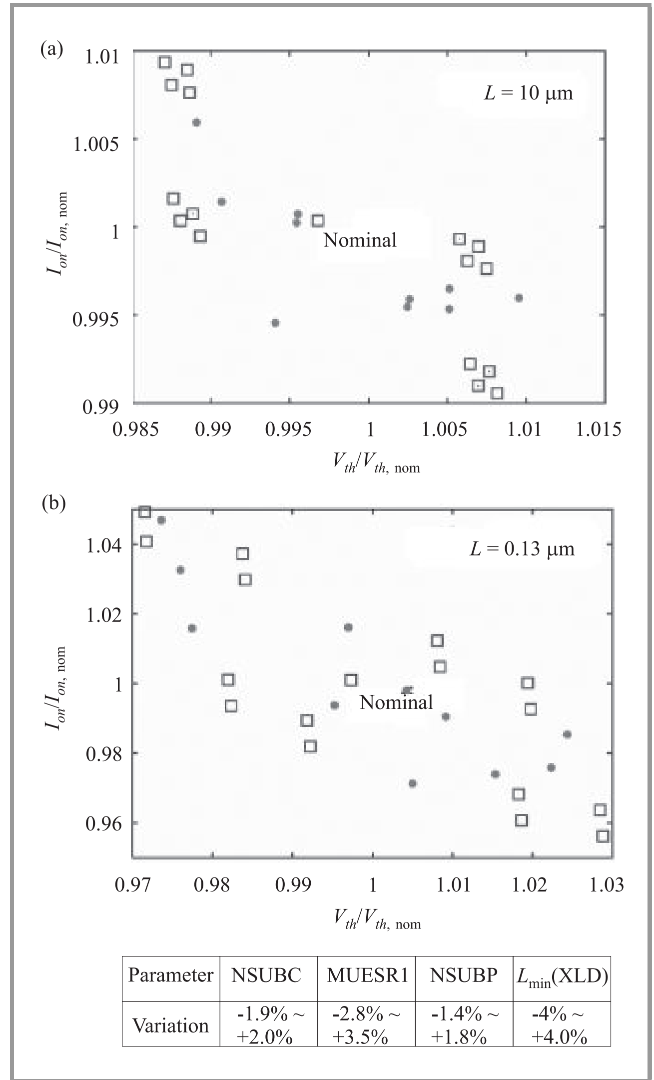
Asymmetric properties of the variation boundaries reflect the systematic component of the within-wafer variation across the wafer. Although only measured transistors down to 130 nm were available for the 65 nm technology, this has no practical negative influence on the accuracy of the extracted microscopic parameter variations of this technology. The 2 graphs in each figure are plotting measured long and short-channel data for

$I_{on} - V_{th}$ variations (points) together with simulated data for the nominal parameter set and the 16 extreme combinations of the microscopic-parameter variations (open squares).



Parameter	NSUBC	MUESR1	NSUBP	L_{min}(XLD)
Variation	-2% ~ +2.3%	-3.3% ~ +4.9%	-2.1% ~ +1.9%	-5.4% ~ +6.8%

Fig. 9. Variation extraction of the most sensitive microscopic parameters for the NMOSFETs of a 100 nm technology applying the pocket implantation. Data points in the $I_{on} - V_{th}$ plots for long (a) ($L_g = 10\ \mu$m) and short (b) ($L_g = 0.1\ \mu$m) MOSFETs are plotted in the same manner as in Fig. 8.



Parameter	NSUBC	MUESR1	NSUBP	L_{min}(XLD)
Variation	-1.9% ~ +2.0%	-2.8% ~ +3.5%	-1.4% ~ +1.8%	-4% ~ +4.0%

Fig. 10. Variation extraction of the most sensitive microscopic parameters for the NMOSFETs of a 65 nm technology applying the pocket implantation. Data points in the $I_{on} - V_{th}$ plots for long (a) ($L_g = 10\ \mu$m) and short (b) ($L_g = 0.13\ \mu$m, which was the smallest measured device) MOSFETs are plotted in the same style as in Fig. 8.

Figures 8, 9, and 10 confirm good agreement between measured and simulated within-wafer variations and verify that the main sources of measured $I_{on} - V_{th}$ variations can be correlated to the microscopic variations of just 4 physical MOSFET properties over 3 technology generations from 180 nm to 65 nm, namely, to substrate and pocket doping concentrations (NSUBC, NSUBP), mobility degradation due to gate-oxide surface-roughness scattering (MUESR1), and channel-length variation due to gate-stack structuring (XLD).

Quantitative improvements of these microscopic parameter variations with the technological advances in each generation of advancing fabrication technologies are plotted in Fig. 11.

Improvements for MUESR1 (70%) and L_{min}(XLD) (55%), related to gate-oxide interface and gate-stack structuring have been quite large in the course of fabrication-technology advances. On the other hand, doping-related technology improvements, reflected by NSUBC (30%) and NSUBP (25%), have been considerably smaller. The average gate-oxide thickness TOX seems to be well controlled, to a level much below the atomic-layer thickness. On the other hand, gate-oxide-interface properties seem to be much more difficult to control. Consequently, the influence of gate-oxide roughness and charged interface traps on the inversion-layer carrier mobility, captured by MUESR1,

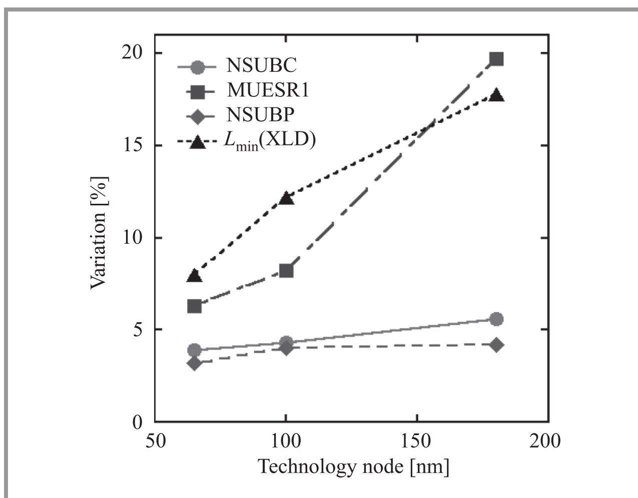


Fig. 11. Extracted variation trends of microscopic device parameters, which are most relevant for within-wafer variations of the macroscopic NMOSFET characteristics of threshold voltage V_{th} and on-current I_{on}, from the 180 nm to the 65 nm technology node.

remain substantial and reflect the variability effects due to the imperfect gate-oxide interface.

5. Conclusion

Surface-potential-based compact models can quantitatively correlate variations of macroscopic MOSFET-performance characteristics with microscopic variations of the physical MOSFET parameters. The most variation-sensitive microscopic parameters are substrate and pocket doping concentrations (NSUBC, NSUBP), the carrier-mobility degradation due to gate-oxide-surface roughness (MUESR1) and effective channel length due to the gate structuring (XLD). Quantitative analysis of $I_{on} - V_{th}$ variation data for NMOSFETs over technology nodes from 180 nm to 65 nm, applying the pocket-implantation technology, confirms that microscopic parameter variations have been continuously improved. The magnitude of these improvements could be extracted from the measured MOSFET-performance data, quantitatively amounting to 70%, 55%, 30% and 25% improvement for MUESR1, $L_{\min}$ (XLD), NSUBC and NSUBP, respectively. The gate-oxide thickness variation is found negligible for the within-wafer variation of I_{on} and V_{th}, but the effect of gate-oxide surface-roughness variations, including the effect of trapped charges, is found to be substantial.

References

[1] K. J. Kuhn, "Reducing variation in advanced logic technologies: approaches to process and design for manufacturability of nanoscale CMOS", in *Proc. IEEE IEDM Tech. Dig.*, Washington, USA, 2007, pp. 471–474.

[2] BSIM3, BSIM4, BSIMSOI [Online]. Available: http://www-device.eecs.berkeley.edu/∼bsim3/bsim4.html

[3] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration", *IEEE Trans. Electron Dev.*, vol. 53, no. 11, pp. 2816–2823, 2006.

[4] K. Takeuchi *et al.*, "Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies", in *Proc. IEEE IEDM Tech. Dig.*, Washington, USA, 2007, pp. 467–470.

[5] M. Miura-Mattausch *et al.*, "HiSIM2: advanced MOSFET model valid for RF circuit simulation", *IEEE Trans. Electron Dev.*, vol. 53, no. 9, pp. 1994–2007, 2006.

[6] M. Miura-Mattausch, H. J. Mattausch, and T. Ezaki, *The Physics and Modeling of MOSFETs: Surface-Potential Model HiSIM*. Singapur: World Scientific, 2008.

[7] G. Gildenblat *et al.*, "PSP: an advanced surface-potential-based MOSFET model for circuit simulation", *IEEE Trans. Electron Dev.*, vol. 53, no. 9, pp. 1979–1993, 2006.

[8] H. C. Pao and C. T. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator) semiconductor transistor (MOST)", *Solid-State Electron.*, vol. 9, no. 10, pp. 927–937, 1966.

Hans Jürgen Mattausch received the Ph.D. degree in physics from the University of Stuttgart, Germany, in 1981. From 1982 to 1995 he was with the Research Laboratories of Siemens AG in Munich, Germany, where he was involved in the development of CMOS technology, memory and telecommunication circuits, power semiconductor devices and compact modeling. From 1995 to 1996 he was with the Siemens Semiconductor Group as Department Head for Product Analysis and Improvement in the Chip-Card IC Division. Since 1996 he is is a Professor at the Research Institute for Nanodevice and Bio Systems and the Graduate School for Advanced Sciences of Matter, Hiroshima University, Higashi-Hiroshima, Japan. He is a senior member of IEEE and a member of IEICE (Institute of Electronics, Information and Communication Engineers of Japan).

Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Akihiro Yumisaki received the B.E. and M.E. degrees in electrical engineering from the Hiroshima University, Japan, in 2006 and 2008, respectively. He is currently with the Sanyo Electric Co., Japan. His main research interest is the quantitative understanding of performance variations in MOS transistors.

Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Norio Sadachika received the B.E., M.E. and Ph.D. degrees in electrical engineering from the Hiroshima University, Japan, in 2004, 2006 and 2008, respectively. Since 2009 he is an Assistant Professor at the HiSIM Research Center, Hiroshima University. His research interests include process variations in CMOS fabrication as well as circuit simulation models for the silicon-on-insulator MOSFETs and multi-gate MOSFETs.
Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Akihiro Kaya was born in Hiroshima, Japan, in 1986. He received the B.E. degree in electrical engineering from the Hiroshima University, Japan, in 2008. Since 2008, he has been studying for his M.E. degree in the Graduate School of Advanced Science of Matter, Hiroshima University. His present research interest is to analyze the influence of process variations on the performance of fabricated integrated circuits.
Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Koh Johguchi received the B.E., M.E. and Ph.D. degrees in electrical engineering from the Hiroshima University, Japan, in 2002, 2004 and 2007, respectively. He is currently an Assistant Professor at the HiSIM Research Center, Hiroshima University. He was a postdoctoral researcher of the "Interdisciplinary Research on Integration of Semiconductor and Biotechnology" project from 2007 to 2009. His research interests are low power memory design, process variation analysis, and device modeling.
Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Tetsushi Koide received the B.E. degree in physical electronics, M.E. and Ph.D. degrees in systems engineering from Hiroshima University, Japan, in 1990, 1992, and 1998, respectively. He was a Research Associate and an Associate Professor in the Faculty of Engineering at the Hiroshima University in 1992–1999 and 1999, respectively. From 1999 to 2001 he was with the VLSI Design and Education Center (VDEC), the University of Tokyo as an Associate Professor. From 2001 to 2008 he was an Associate Professor in the Research Center for Nanodevices and Systems, Hiroshima University. Since 2008 he has been an Associate Professor in the Research Institute for Nanodevice and Bio Systems (RNBS) and the Graduate School of Advanced Sciences of Matter, Hiroshima University. His research interests include system design and architecture issues for memory-based systems, real-time image processing, VLSI CAD/DA, genetic algorithms, and combinatorial optimization. He is a member of the Institute of Electrical and Electronics Engineers, the Association for Computing Machinery, the Institute of Electronics, Information and Communication Engineers of Japan, and the Information Processing Society of Japan.
Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Mitiko Miura-Mattausch received the Ph.D. degree from the Hiroshima University, Japan. She joined the Max-Planck-Institute for solid-state physics in Stuttgart, Germany, as a Researcher from 1981 to 1984. From 1984 to 1996, she was with Corporate Research and Development, Siemens AG, Munich, Germany, working on hot-electron problems in MOSFETs, the development of bipolar transistors, and analytical modeling of deep submicron MOSFETs for circuit simulation. Since 1996, she is a Professor in the Department of Electrical Engineering, Graduate School of Advanced Sciences of Matter at the Hiroshima University, leading the ultra-scaled devices laboratory.
Graduate School of Advanced Sciences of Matter
Hiroshima University
Kagamiyama 1-4-2
Higashi-Hiroshima 739-8527, Japan

Paper

Analysis of the Dispersion of Electrical Parameters and Characteristics of FinFET Devices

Arkadiusz Malinowski, Daniel Tomaszewski, Lidia Łukasiak, Andrzej Jakubowski, Makoto Sekine, Masaru Hori, and Michael L. Korwin-Pawlowski

Abstract— Extensive numerical simulations of FinFET structures have been carried out using commercial TCAD tools. A series of plasma etching steps has been simulated for different process conditions in order to evaluate the influence of plasma pressure, composition and powering on the FinFET topography. Next, the most important geometric parameters of the FinFETs have been varied and the electrical characteristics have been calculated in order to evaluate the sensitivity of the FinFET electrical parameters on possible FinFET structure variability.

Keywords— FinFET, line edge roughness, parameter variability, plasma etching, technology computer aided design (TCAD).

1. Introduction

As the dimensions of MOS transistors are shrunk, the close proximity between the source and drain reduces the ability of the gate electrode to control the potential distribution and the flow of current in the channel. Undesirable short channel effects (SCE) induce higher subthreshold slope, threshold voltage roll-off, and punch-through between the drain and source. Multi-gate MOS SOI transistors, e.g., fin field effect transistors (FinFETs), are expected to be promising candidates for the next generation CMOS devices [1]. Because of their structure, FinFETs suppress short channel effects thus leading to further improvement of CMOS circuit performance [2].

However, small-size FinFETs are sensitive to technological process variations, which disturb the electrical characteristics and lower manufacturing yield. Numerical simulation and modeling of the effect of process parameter modifications and random variations becomes a very relevant task. In the presented work, key processes of fin formation, and their influence on manufacturing yield have been discussed. A methodology of plasma modeling and simulation for advanced silicon devices has been presented. Different plasma-related effects (such as loading effects, sidewall bowing and aspect ratio dependent etching) occurring during fin formation and affecting the fin size and shape have also been explained. Finally, the influence of polysilicon overetch on FinFET performance is discussed.

2. Process Simulations

Most plasma etch processes are based on either dc discharge or radio frequency (RF)-excited plasma, typically driven at a frequency of 13.56 MHz. During such discharge, electrons, ions, and reactive species are generated mainly in the bulk of the plasma. Three fundamental reactions may occur when an ion strikes a molecule: electron attachment, ionization and dissociation. The ions are transported towards the surface via a sheath area and impinge on the surface. The fidelity of pattern transfer during etching depends on important process characteristics, that is ion-energy distribution functions (IEDFs) and ion-angular distribution functions (IADFs). Theoretical study of sheath phenomena is therefore critical to developing appropriate models that will increase understanding of the influence of reactor conditions on plasma etching behavior. The IEDFs and IADFs are calculated using the Monte Carlo (MC) method.

A key aspect for fin formation is the anisotropy of the reactive ion etching (RIE) process. It is directly responsible for the shape and size of the fin area. The RIE process conditions have been chosen to etch a fin with the height of 60 nm and width of 20 nm as the reference model shown in Fig. 1(a). We have used CF4 plasma with the composition of gas 69 a.u./ion 19 a.u. under 100 mTr pressure and excited by a 50 V(dc)/55 V(ac) power source. However, the plasma etching process gives rise to a number of undesirable effects, which may be noticed, if the fin sidewall is enlarged (Fig. 1(b)). First, a random distortion of the sidewall edge may be observed. This phenomenon is called surface roughness and may play a significant role in FinFET device performance. The fin surface roughness is caused by charging effects in the plasma. Other plasma-related effects that have to be taken into account during fin formation are the so-called loading effects. They occur when the total area of the material exposed to the etchant decreases. This occurs obviously in the vicinity of the corners between the buried SiO_2 layer and the sidewalls of the projected fin. Then the reactive species become consumed. and the etch rate decreases, making the process lose its anisotropic character. The resulting fin sidewall becomes rounded instead of being perpendicular to the substrate.

The following parameters of the RIE process and their influence on the FinFET profile have been considered: gas pressure, gas composition and plasma powering.

2.1. Gas Pressure

The pressure of the gas in the chamber strongly influences the spatial distribution of the active plasma. As shown in Fig. 2 (obtained using the MC method), at lower pressures

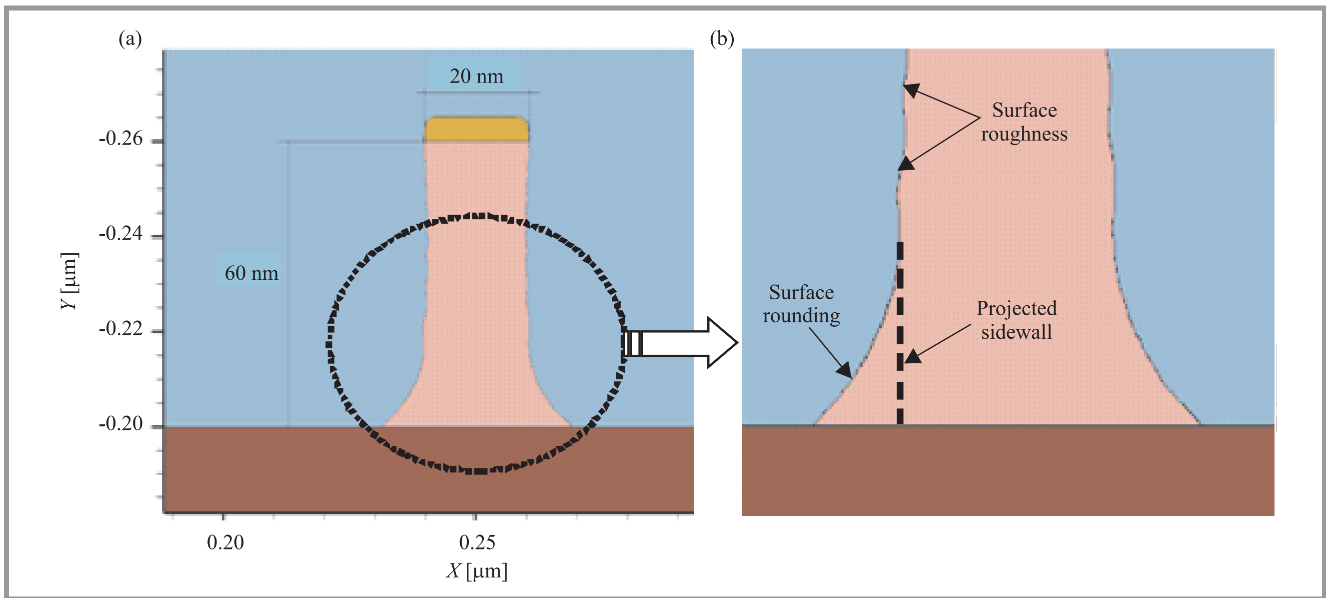


Fig. 1. Fin cross-section after RIE process: (a) overall view, (b) details of the fin sidewall topography: surface roughness at the sidewalls and surface roundings at the corners.

the gas flow acts as an anisotropic beam while at higher pressures spontaneous ion-molecule collisions make the gas flow more like an isotropic cloud. Apart from weaker etching anisotropy the increased gas pressure lowers the etching rate.

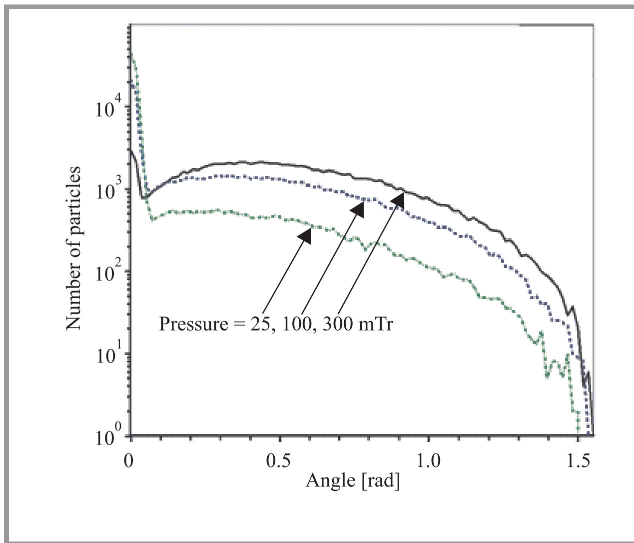


Fig. 2. Ion angular distribution for different chamber gas pressures calculated using the MC method.

The weak anisotropy at higher pressures is followed by lower quality of the patterning. For example, it influences the ratio between the projected fin height and the calculated width of the silicon area etched via the window opened in the mask. However, the effect of the lower etching rate seems to be somewhat ambiguous, because it may be helpful for better control of the total layer thickness etched.

In our case the change of plasma pressure from 25 to 100 mTr has had approximately the same effect on the fin profile as the pressure variation from 100 mTr to 300 mTr. The effect of the gas pressure on the etching rate is illustrated in Table 1.

Table 1
$\mathrm{Fin_{width}}$ versus chamber gas pressure

Pressure [mTr]	Ratio	Rate [nm/min]	$\mathrm{Fin_{width}}$ [nm]
25	3.8	647	25
100	3.46	661	20
300	2.92	587	16

2.2. Gas Composition

The RIE processes with three types of gas mixtures, i.e., gas 69 a.u./ion 19 a.u. (CF4); gas 96 a.u./ion 19 a.u. (SF6) and gas 80 a.u./ion 80 a.u. (HBr) have been evaluated. The results are shown in Table 2. We have considered the neutral to ion flux ratio: the lower the neutral to ion flux ratio the better the anisotropy thus the fin area obtained using HBr is thicker than the one obtained using SF6 or CF4.

Table 2
$\mathrm{Fin_{width}}$ versus gas and ion composition

Gas/ion [a.u.]	Ratio	Rate [nm/min]	$\mathrm{Fin_{width}}$ [nm]
69/19	3.46	661	20
96/19	3.03	598	17
80/80	4.86	730	34

In Fig. 3 the fin etch by gas/ion 80/80 a.u. has been shown. The effect of bowing of the sidewalls in the etched profile is visible. It may be induced by ion deflection (ion

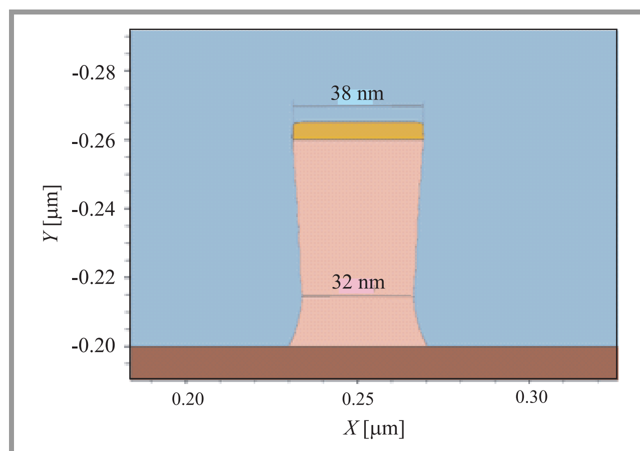


Fig. 3. Fin sidewall bowing.

trajectory distortion) in very narrow spaces between the fins. The angular distribution of ions impacting and subsequently scattered by the etching feature is expected to be the primary cause of non-vertical sidewalls.

2.3. Plasma Powering

Three different powering configurations have been tested. The results are shown in Table 3. In general, increasing power leads to higher etching rate and anisotropy. However, we have found maxima of the etching rate as well as of the anisotropy for V(dc) potential 50 V and V(ac) potential 55 V, respectively. Higher V(dc) and V(ac) potentials result in lower etching rates and ratios. The maximum etching rate is limited either by the chemical reaction rate at the surface or the flow of arriving ions. Higher potentials result in spontaneous collisions thus decreasing anisotropy.

Table 3
Fin_{width} versus power settings

V(dc)/V(ac) [V]	Ratio	Rate [nm/min]	Fin_{width} [nm]
20/25	3.27	615	20
50/55	3.46	661	20
100/105	3.18	611	19

Therefore, anisotropic etch is controlled by the shadowing effect and the directionality of the incoming ions. The lower V(dc) potential is responsible for domination of the isotropic chemical etch.

3. Device Simulations

In order to estimate the variability of fin dimensions in the FinFET caused by the RIE process dispersion Synopsys Sentaurus Structure Editor and Sentaurus Device applications have been used.

A three-dimensional FinFET model has been built (Fig. 4). The structure has been created on a SOI substrate with 60 nm p-type device layer (boron conc.$= 1 \cdot 10^{16}$ cm^{-3}). Source and drain have been doped with arsenic (conc.$=$ $5 \cdot 10^{19}$ cm^{-3}). Fin dimensions are as follows: $Fin_{width} =$ 20 nm, $Fin_{height} = 60$ nm. Over the fin a thin (2 nm) HfO_2 gate dielectric layer has been deposited. Over the gate dielectric silicon nitride spacers have been formed thus defining $Gate_{length} = 25$ nm. The polysilicon layer has been heavily doped with arsenic (conc.$= 1 \cdot 10^{20}$ cm^{-3}).

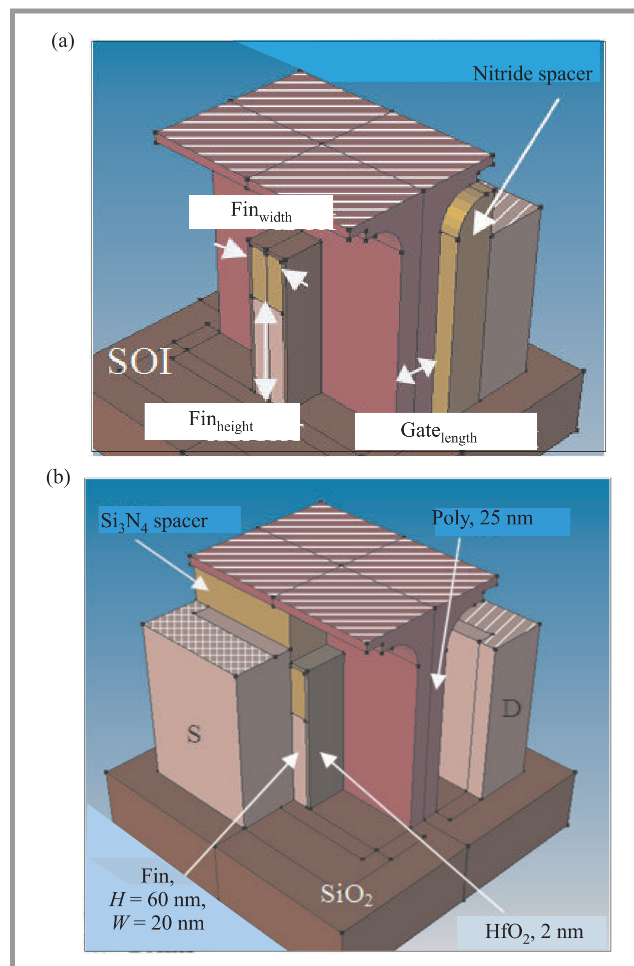


Fig. 4. FinFET structure: (a) details of the fin, (b) details of the spacer.

The electrical characteristics and parameters of the FinFET have been calculated in order to evaluate the influence of RIE dispersion on the device operation as a switch for integrated circuit applications. The FinFET has been biased as follows: gate-source voltage $V_{GS} = -0.5$ V $-$ 1 V, drain-source voltage $V_{DS} = 0.1$ V. An example of the obtained $I_D(V_{GS})$ transfer characteristics is shown in Fig. 5.

The following electrical parameters have been taken into account: threshold voltage (V_T), transconductance (g_M), and subthreshold swing (SS). Table 4 presents the variation of the FinFET electrical parameters due to Fin_{width} change. The following remarks may be formulated. Firstly, the threshold voltage values are very low. They result directly from the shape of the $I_D - V_{GS}$ curves and are related to the non-optimized gate stack structure. Due to the very low (in terms of the gate-stack structure) channel doping

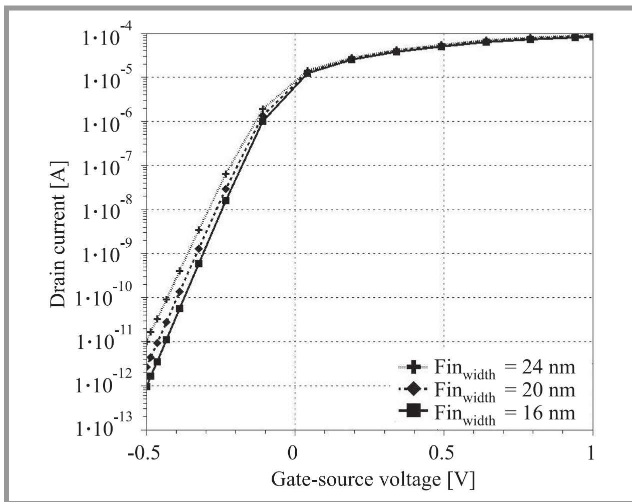


Fig. 5. FinFET $I_D - V_{GS}$ characteristics for a series of fin widths.

concentration in the fin area the FinFETs can be switched off only by negative gate bias. Nevertheless, one may easily notice an improvement of the subthreshold slope with a decrease of the $\text{Fin}_{\text{width}}$. This is most likely due to the fact that in narrow FinFETs the gate control over the channel conduction is better [3].

Table 4
FinFET electrical parameters for different values of $\text{Fin}_{\text{width}}$

$\text{Fin}_{\text{width}}$ [nm]	V_T [V]	g_M [μS]	SS [mV/dec]
16	−0.102	133.1	67.8
20	−0.104	134.3	70.8
24	−0.106	126.1	73.9

As shown in Fig. 4 the polysilicon gate has been created using the silicon nitride spacers. Spacer lithography technology is attractive for overcoming the limits of conventional lithography techniques in terms of pattern fidelity and critical dimension (CD) variation. Simulations of the spacer lithography variations causing $\text{Gate}_{\text{length}}$ variations have been also carried out. The results are shown in Table 5.

Table 5
FinFET electrical parameters for different values of $\text{Gate}_{\text{length}}$

$\text{Gate}_{\text{length}}$ [nm]	V_T [V]	g_M [μS]	SS [mV/dec]
20	−0.109	112.4	73.3
25	−0.104	134.3	70.8
30	−0.088	189.5	69.2

It may be noticed, that as expected a decrease of $\text{Gate}_{\text{length}}$ leads to the threshold voltage lowering. It may be also stated, that the shortening of the gate induces lowering of the transconductance. This somewhat unexpected FinFET behavior has been caused by the method to generate device structures considered in this paper. We have assumed that the distance between the heavily doped source and drain areas is constant. Different values of the $\text{Gate}_{\text{length}}$ parameter have been obtained by the variation of the spacer thickness. However, an increase of the latter (and decrease of the $\text{Gate}_{\text{length}}$) induces large increase of the series resistance. This, in turn, strongly degrades current conduction and transconductance.

4. Results and Conclusions

The ultra-thin fin formation with good uniformity is still a challenging task for FinFET manufacturing. The uniformity of silicon fin width ($\text{Fin}_{\text{width}}$) is especially critical for the FinFET because its variation may cause a change in channel potential and subband structure, which governs short-channel behavior and quantum confinement effects of inversion charges. Also if $\text{Gate}_{\text{length}}/\text{Fin}_{\text{width}}$ ratio is smaller than 1.5, drain induced barrier lowering (DIBL), subthreshold swing, and off-state leakage current increase significantly. Thus, a small change of fin width may result in large variation of device characteristics for short gate lengths [4].

It has been shown, that non-optimized FinFET structure leads to a number of undesired effects, e.g., incorrect threshold voltage and degraded $I - V$ characteristics.

It has been also shown in our paper that the uniformity of silicon fin width strictly depends on the RIE parameters, such as pressure, gas composition, and RF power. Therefore, a precise control of these parameters during process is critical.

Acknowledgements

This work was partially supported by Polish Ministry of Science and Higher Education under project NN 515 4449 33, “Modeling and Characterization of Multigate MOS SOI Structures”, and by the Center for Advanced Studies Warsaw University Technology (WUT) under scholarship for Ph.D. CAS/6/POKL.

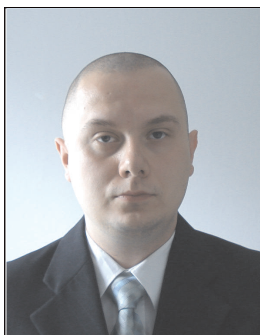
References

[1] J. P. Colinge, *FinFETs and Other Multi-Gate Transistors*. New York: Springer, 2008.

[2] H. Kawasaki *et al.*, “FinFET process and integration technology for high performance LSI in 22 nm node and beyond”, in *Proc. Junct. Technol. 2007 Int. Worksh.*, Kyoto, Japan, 2007, pp. 3–8.

[3] A. Yagishita, “FinFET SRAM process technology for hp 32 nm node and beyond”, in *Proc. Integr. Circ. Des. Technol. ICICDT'07. IEEE Int. Conf.*, Austin, USA, 2007, pp. 1–4.

[4] Y. K. Choi *et al.*, “Spacer FinFET: nanoscale double-gate CMOS technology for the terabit era”, *Solid-State Electron.*, vol. 46, pp. 1595–1601, 2002.



Arkadiusz Malinowski received the B.Sc. and M.Sc. degrees from the Warsaw University of Technology, Poland, in 2005 and 2007, respectively. He currently pursues the Ph.D. degree in electrical and computer engineering at the Warsaw University of Technology and the Nagoya University, Japan. In March 2004 he joined the Institute of Electron Technology, Warsaw, working in the area of TCAD semiconductor process and device simulation. His research interests include FinFET CMOS technology scaling, nanocarbon based FET and plasma nanoprocessing.

e-mail: m_arkadi@nuee.nagoya-u.ac.jp
Department of Electrical Engineering
and Computer Science
School of Engineering
Nagoya University
IB building 3F 331, Furo-cho, Chikusa-ku
Nagoya City, Aichi, 464-8603, Japan

Daniel Tomaszewski received the M.Sc. degree from the Warsaw University of Technology, Poland, in 1980. Since then he is with the Institute of Electron Technology, Warsaw. In 1998 he received the Ph.D. degree in electrical engineering. His research interests include modeling and characterization of silicon and silicon-on-insulator devices for the purpose of IC diagnostics and design. He participated in several conferences and workshops related to these fields.

e-mail: dtomasz@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Lidia Łukasiak graduated from the Faculty of Electronics, Warsaw University of Technology, Poland, in 1988 and joined the Institute of Microelectronics and Optoelectronics the same year. She received the Ph.D. and D.Sc. degrees from the same university in 1994 and 2002, respectively. Since 2004 she has been the Vice-Director for Teaching of the Institute of Microelectronics and Optoelectronics. Her research interests include modeling and characterization of semiconductor devices and microprocessor systems.

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

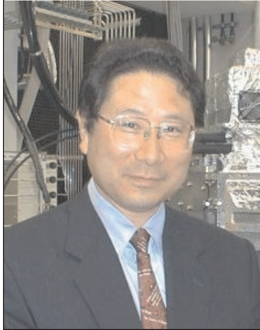
Andrzej Jakubowski received the M.Sc., Ph.D. and D.Sc. degrees in electrical engineering from the Warsaw University of Technology (WUT), Poland. At present Professor Jakubowski works at the Institute of Microelectronics and Optoelectronics (WUT). His main research interests include modeling and characterization of semiconductor devices and integrated circuits. He is author and co-author of more than 500 papers, several books and textbooks.

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

Makoto Sekine is a Professor at the Plasma Nanotechnology Research Center, Nagoya University, Japan. After graduating from the Waseda University, Tokyo, in 1982 he joined the Toshiba, where he developed new plasma etch tools that were commercialized by the Tokyo Electron. Ltd. and continue to be prominent in the semiconductor industry. As a Visiting Researcher at the University of California, Berkeley, from 1989–1991, he focused on diagnostics for plasma-surface interactions. In 1998–2001, he was the Research Manager of the Plasma Laboratory of the Association of Super-Advanced Electronics Technologies (ASET, a Japanese national consortium for electronics technology development). His work there pioneered the application of fundamental diagnostics and modeling to the understanding of dielectric etch mechanisms in wide classes of commercial plasma etchers. While at the ASET, he initiated a new laboratory dedicated to the reduction of global warming gases (e.g., PFC's) through plasma processes. Upon his return to Toshiba post-ASET, he worked on 3D high density chip packaging technology. After retiring Toshiba in 2004, he was a Vice-President of a venture firm in Connecticut, US, and a visiting scientist at the De-

partment of Chemistry, University of Connecticut. Then he moved to Nagoya University in 2006.

e-mail: sekine@plasma.engg.nagoya-u.ac.jp
Department of Electrical Engineering
and Computer Science
School of Engineering
Nagoya University
IB building 3F 331, Furo-cho, Chikusa-ku
Nagoya City, Aichi, 464-8603, Japan



Masaru Hori received the B.Sc., M.Sc. degrees from the Waseda University, Tokyo, Japan, in 1981 and 1983, respectively, and Ph.D. degree from the Nagoya University in 1986. He is a Professor of Graduate School of Engineering at the Nagoya University and the Director of the Center for Plasma Nano Engineering. He has held staff positions at the Toshiba, the Cavendish Laboratory of Cambridge, the Meijo University, the NU Eco Engineering and the University of Tsukuba. His research interests focus on fabrication nanometer thin next generation ULSI devices, research on atomic and molecular radicals measured by quantum optics, manufacturing and device applications of carbon nanowalls, low-temperature formation of microcrystalline silicon thin films and applications to flexible devices manufacturing, development of autonomous nanomanufacturing equipment, research on dry process using synchrotron radiation. He has published more than 200 journal papers. He is a member of Japan Society of Applied Physics, The Institute of Electrical Engineers of Japan, and American Vacuum Society. In 2003 he received an award Plasma Electronics Award and JJAP Editorial Contribution Award in 2004. He is also a recipient 17th World Interfinish Congress & Exposition, Invited Presentation Award in 2008. He served as the Chair of Executive Committee of 27th International Symposium on Dry Process and a member of Organizing Committee of 6th International Conference on Reactive Plasmas. He is also the President of Intellectual Cluster Creation Project.

e-mail: hori@nuee.nagoya-u.ac.jp
Department of Electrical Engineering
and Computer Science
School of Engineering
Nagoya University
IB building 3F 331, Furo-cho, Chikusa-ku
Nagoya City, Aichi, 464-8603, Japan

Michael L. Korwin-Pawlowski received the M.Sc. degree in electronics from the Warsaw University of Technology, Poland, in 1963, and Ph.D. in electrical engineering from the University of Waterloo, Canada, in 1974, and the diplome MBA in finance from the Long Island University, USA, in 1993. He joined the Université du Québec en Outaouais, Canada, in 2004 as a Professor at the Département d'informatique et d'ingénierie after over 30 years of career at high management levels in R&D, engineering, marketing and operations management in the international semiconductor and microelectronic industry, including 16 years (1982–1999) at the General Instrument Corporation and General Semiconductor Corporation in Taiwan, USA and Ireland, and at the X-ion S.A. in France (1999–2002). Current research interests are in fiber optic capillary sensors, pulsed laser deposited thin films, multicharged ion technology and microelectronic device modeling. He is the author or co-author of over 80 publications, 2 book translations and 18 USA, French and international patents.

e-mail: michael.korwin-pawlowski@uqo.ca
Département d'informatique et d'ingénierie
Université du Québec en Outaouais
101 rue Saint-Jean-Bosco
Gatineau, QC, J8X 3X7, Canada

Invited paper

Rare Earth Silicate Formation: A Route Towards High-*k* for the 22 nm Node and Beyond

Ivona Z. Mitrovic and Stephen Hall

Abstract— Over the last decade there has been a significant amount of research dedicated to finding a suitable high-*k*/metal gate stack to replace conventional SiON/poly-Si electrodes. Materials innovations and dedicated engineering work has enabled the transition from research lab to 300 mm production a reality, thereby making high-*k*/metal gate technology a pathway for continued transistor scaling. In this paper, we will present current status and trends in rare earth-based materials innovations; in particular Gd-based, for the high-*k*/metal gate technology in the 22 nm node. Key issues and challenges for the 22 nm node and beyond are also highlighted.

Keywords*— *gadolinium silicate, interfacial layer, lanthanides, rare earth oxides.

1. Introduction

A 32 nm process technology, including a high-*k* dielectric and metal-gate has been announced [1]–[5]. The equivalent oxide thickness (EOT) of the high-*k* dielectric has been reduced from 1.0 nm on 45 nm node to 0.9 nm [1] on the 32 nm process, while gate length has been reduced to 30 nm. Transistor gate pitch continues to scale 0.7× every two years – with 32 nm providing the tightest gate pitch in the industry. The key industrial players are Intel Corporation, IBM alliance (with AMD, Chartered, Freescale, Infineon, Samsung, ST and Toshiba), TSMC, NEC, Panasonic in collaboration with Renesas. Different integration strategies have been employed by various parties involved, namely a replacement gate [6], [7] or a conventional gate-first approach [8], [9]. In the latter, the gate stack is formed before the source and drain, as in a conventional complementary metal oxide semiconductor (CMOS) process, while the former is a gate-last approach. Moving beyond the planar transistor, IBM, AMD, Freescale and Toshiba have recently presented a fin field effect transistor (FinFET) with high-*k* and metal gates for the 32 nm node and beyond [3].

A static random access memory (SRAM) cell was devised at areas down to 0.128 μm^2. Using 22 nm design rules, the cell was fabricated using a CMOS process flow and e-beam lithography. In the cell, fin pitch was 80 nm, gate pitch 110 nm and physical gate length 30 nm. To enable high-*k* and metal gates, chemical vapour deposition (CVD)-based HfO_2, physical vapour deposition (PVD) TiN and polysilicon were deposited on the fin portion of the device. TSMC has also announced a 32 nm process [4], which includes a high-*k*/metal-gate scheme based on a gate-first technology. A 0.15 μm^2 SRAM cell was developed by using a Hf-based material and 193 nm immersion lithography with a numerical aperture of 1.35. The high-*k* material has been scaled to 1 nm, at 30 nm physical gate length and 130 nm gate pitch. The team of NEC and Toshiba Corporation has developed a 32 nm process, with a single-exposure lithography technology and a gate-first high-*k*/metal-gate process [5]. They demonstrated a SRAM cell of 0.124 μm^2 and a gate density of 3650 KGate/mm^2. From the relevant publications and press releases, it is evident that Hf-based dielectrics are retained for the 32 nm node leaving other dielectrics for consideration at further technology nodes. Physical gate length scaling, from 35 nm, in the 45 nm generation, to 30 nm in the 32 nm generation, is enabled by high-*k* dielectric scaling and shallow junction engineering [10].

Low standby power (LSTP) technology requirements in the near term years according to the International Technology Roadmap for Semiconductors (ITRS) [11] are listed in Table 1. As can be seen, more aggressively scaled EOT is required in order to reach the specification for the 22 nm node. This leads to stringent control of interfacial oxide layer, either by reducing its thickness or by increasing its *k*-value. A good interface requires either that the oxide is amorphous, or that it is epitaxial and lattice-matched to the underlying silicon [12]. Amorphous oxides represent a low-cost solution; nonetheless, the challenge is to keep these materials amorphous even after post-deposition high temperature processing in order to avoid increased surface roughness and additional leakage due to the formation of grain boundaries, as shown in many investigations [13]–[16]. Another approach is based on the development of epitaxial metal oxides grown directly on silicon surfaces [12]. It is known that for given values of MOS leakage current and insulator thickness, the dielectric constant, *k*, and the offset value between oxide and silicon energy bands, ΔE, are bound roughly by a hyperbolic relation $k \cdot \Delta E = C_E$, where C_E is a constant. Engstrom *et al.* [17] have suggested that a value of $C_E \approx 70$ eV is necessary for the 22 nm bulk LSTP node, while the corresponding figure for silicon-on-insulator (SOI) technology is in the range of 30–40 eV. Rare earth (RE) oxides offer interesting properties to fulfil such requirements: a high dielectric constant [17], a high band gap and suitable band offsets with respect to Si [18], sufficiently high breakdown strength,

Table 1
Low standby power technology requirements – near term years [11]

Year in production	2008	2009	2010	2011	2012	2013	2014	2015
MPU/ASIC metal $1\frac{1}{2}$ pitch [nm] (contacted)	59	52	45	40	36	32	28	25
Physical length gate for high performance logic [nm]	29	27	24	22	20	18	17	15
Physical gate length for LSTP (Lg)								
Extended planar bulk and DG [nm]	38	32	29	27	22	18	17	15
UTB FD [nm]	*	*	*	*	*	20	18	17
EOT								
Extended planar bulk [nm]	1.6	1.5	1.4	1.3	1.2	1.1	*	*
UTB FD [nm]	*	*	*	*	*	1.2	1.1	1.0
DG [nm]	*	*	*	*	*	*	*	1.1
Maximum gate leakage current density ($J_{g,\text{limit}}$)								
Extended planar bulk [mA/cm^2]	81	94	110	120	140	150	*	*
UTB FD [A/cm^2]	*	*	*	*	*	150	170	180
DG [A/cm^2]	*	*	*	*	*	*	*	190

* delineate one of two time periods: either before initial production ramp has started for ultra-thin body fully depleted (UTB FD) SOI or double-gate (DG) MOSFETs, or beyond when planar bulk or UTB FD MOSFETs have reached the limits of practical scaling.

extremely low leakage current, and well-behaved interface properties. Their key advantages for advanced CMOS are:

- thermal stability;
- a feature to shift the work function of a metal gate towards n-type and thus engineer a transistor threshold voltage [19], [20];
- a reduction of the low-*k* interfacial layer (IL) thickness.

An important consideration in choosing an alternative high-*k* dielectric is its compatibility with Si, and metal silicates have attracted much recent attention [21]–[23]. The presence of silicon leads to improved metal oxide/silicon interface stability and reduced leakage currents. This in turn has generated interest in the deposition of lanthanide silicates, such as Gd-silicate [24], La-silicate [25], Pr-silicate [26], [27]. A silicate formation is known to occur when a rare earth oxide is in contact with a Si-containing dielectric or a silicon substrate in the presence of oxygen [23]. As such, it can be used to consume the typical SiO_2-like interfacial layer between high-*k* and silicon substrate or to enhance its *k*-value. In this paper, we will show that RE silicate formation is one of the possible pathways towards scaling beyond the 22 nm node. In particular, our recent work on optimization of GdSiO-based gate stack is reviewed and results presented on thermal stability and mechanisms of silicate formation.

2. Rare Earth Oxides and Silicates

Rare earth oxides are attractive materials for gate dielectric application, and in particular the lanthanide oxides group (LnO's). Lanthanides refer to a series of 15 elements from La to Lu in the periodic table, which have similar but gradually changing characteristics [28]. The LnO's can have different stoichiometries due to the multiple oxidation states (+2, +3, and +4) of the metals. This leads to oxides with different structural phases including two cubic phases, namely, the calcium fluoride structure for the Ln(IV) only, and the bixbyite structure for Ln(III) [29]. Lanthanide oxides with more than one valence state are not the best choice for CMOS processing because of the coexistence of phases with different oxygen contents and possible transformations between them, or even the occurrence of mixed valence-state structures [30], [31]. A summary of the key properties of RE LnO's is given in Table 2 ([17], [29] and references therein). Those deemed suitable for application are now described.

The RE oxides, such as La_2O_3 [32], [33], Pr_2O_3 [34]–[37], Gd_2O_3 [38], and Nd_2O_3 [39], [40] have been investigated. They are good insulators due to their large band gaps, high dielectric constants (13–16 for Gd_2O_3, 25–30 for La_2O_3 [17], 15–25 for Pr_2O_3 [41]) and good thermodynamic stability on silicon even at high temperatures, i.e., when heated in contact with silicon will not directly react to form silicide, metal, or silicon oxide [42]. Another attractive feature of Pr_2O_3, Gd_2O_3, and Nd_2O_3 is their relatively close lattice match to silicon ($2a_{Si} = 1.090$ nm), which offers the possibility of epitaxial growth, eliminating problems related to grain boundaries in polycrystalline films.

Silicate formation of binary lanthanide oxides in contact with SiO_2 has been the subject of many studies [24], [40], [43]–[46] whereby the SiO_2 IL is consumed during a high temperature step. The process has been demonstrated for La_2O_3 [47], [48] or Gd_2O_3 [49], [50] deposited on SiO_2.

Table 2
Properties of RE Ln_2O_3 oxides [17], [29]

Compound	Structure	a[Å]	c[Å]	Band gap [eV]	Dielectric constant
Er_2O_3	Bixbyite	10.548		5.3, 5	13
La_2O_3	Bixbyite	11.327		5.5, 6	25, 27–30
La_2O_3 A type polymorph	Hexagonal	3.937	6.129	5.5	25
Nd_2O_3	Bixbyite	11.08		4.4	16
Nd_2O_3 A type polymorph	Hexagonal	3.829	5.997	4.4	16
Sm_2O_3	Bixbyite	10.92		5	13, 11.4–15
Ho_2O_3	Bixbyite	10.606		5, 5.2	13.1
Gd_2O_3	Bixbyite	10.813		5.3, 6.4	13.7, 13.6
Dy_2O_3	Bixbyite	10.665		4.9	13.1
Lu_2O_3	Bixbyite	10.391		5.4, 6	12.5
Pr_2O_3	Hexagonal	3.857	6.916	4.6	14.9, 25.4 [41]

The extent of silicate formation depends on temperature, time (spike anneal) and the ion radius of the RE element [31], [51], [52]. A unique property of RE elements is "lanthanide contraction", a term which refers to the observation that the ion radii of rare earth elements decrease gradually as the atomic number increases. The quantity of Si-O-Ln bonds increases as the post-annealing temperature rises, and this increase depends strongly on the ion radii of the RE elements. As a result, metal-oxides with larger ion radii easily form LnSiO (silicate) layers. An alternate explanation can be deduced from thermodynamic arguments [52]. Furthermore, the quantity of Si-O-Si bonds increases after annealing independent of the element.

Rare earth oxide and silicate films have been deposited by various methods including atomic and electron beam evaporation [24], [38], [49], [50], [53], molecular beam epitaxy (MBE) [40], metal organic chemical vapour deposition (MOCVD) and atomic layer deposition (ALD) [33], [35], [54], [55]. CVD and ALD techniques allow the controlled growth of highly conformal films on planar and high-aspect ratio substrates. Recent developments in precursors for the MOCVD and ALD of lanthanide oxides LnO_x and silicates $LnSi_xO_y$ (Ln = Pr, Gd, La and Nd) have been described in [55], with emphasis on the effect of the precursors molecular structure on process chemistry and layer purity.

3. GdSiO: High-*k* Material for the 22 nm Node

3.1. Gd_2O_3 Studies

The Gd_2O_3 layers have been studied extensively [38], [52], [56]–[58]. Gd_2O_3 has been deposited by MOCVD [57], anodic oxidation [59], thermal oxidation [60], [61], MBE [62], [63], e-beam evaporation [49], [50] and ALD [55], [64], [65]. The use of Gd_2O_3 dielectric layers has been reported for Si [52], [63] and III-V compounds, such as GaN [66], [67] and GaAs [68], [69]. Based on thermodynamic consideration, Gd_2O_3 formed according to equation $2Gd_2O_3 + 3\,Si = 4Gd + 3SiO_2$ with $\Delta G > 100$ kJ, is expected to be stable on Si up to 1000°C [42]. Furthermore, Gd is a single valence metal ion (+3); therefore it forms only a single oxide (Gd_2O_3) from the reaction with oxygen. This oxide does not exhibit any intermediate metastable states while reacting with oxygen. The effective dielectric constant of the Gd_2O_3 films is in the range of 7–23 [57]. The reported values for a band gap vary from 5.2 eV [70], [71], 5.3 [29], to 6.4 [17]. The conduction and valence band offsets to Si are larger than 2 eV [62], [71]. The lattice parameter of Gd_2O_3 in its bixbyite phase is 1.081 nm (Table 2), while Si has a lattice constant a_{Si} of 0.545 nm, where $2 \times a_{Si}$ is 0.4% larger than $a_{Gd_2O_3}$, which allows for epitaxial growth.

There are reports on electrical properties of epitaxial Gd_2O_3 thin films grown by MBE [62], [72], with EOT$<$ 1 nm and leakage current density below 1 mA/cm^2. As can be seen from Table 1, these numbers exceed the requirements for LSTP application predicted for 2012. A careful control of the thermodynamic parameters, such as oxygen chemical potential allows the interface layer change from oxide-like to a silicate-like, and thus leads to larger *k* and lower leakage for the latter one [62]. The impact of rapid thermal anneal (RTA) on structural and electrical properties of crystalline Gd_2O_3 layers grown on Si has been discussed in [73]. Any degradation of the layers can be significantly reduced by capping with amorphous-Si prior to RTA.

3.2. Academic Cluster Work

The so-called Academic Cluster comprises of four member institutions, namely – Chalmers University of Technology (Sweden), AMO GmbH (Germany), Tyndall National Institute (Ireland) and Liverpool University (UK). We have recently reported an optimized process based on GdSiO for the gate dielectric, consistent with the 22 nm LSTP target [49]. The Gd_2O_3 layers are deposited by

e-beam evaporation [49], [50] and atomic layer deposition [64], [65] on different interfacial silicon dioxide layers (thermal ∼1–4 nm, and native ∼1 nm). Detailed material properties of the layers are assessed by variable angle (65–75°) spectroscopic ellipsometry (VASE), medium energy ion scattering (MEIS), X-ray diffraction (XRD), high resolution transmission electron microscopy (HRTEM) and X-ray photoelectron spectroscopy (XPS). Electrical characterization including capacitance-voltage and current-voltage techniques is conducted on metal insulator semiconductor (MIS) capacitors with TiN and Au electrodes. The formation of gadolinium silicates has been achieved by RTA annealing of various Gd_2O_3/SiO_2 gate stacks in the temperature range of 800 to 1050°C and anneal time 1–100 s.

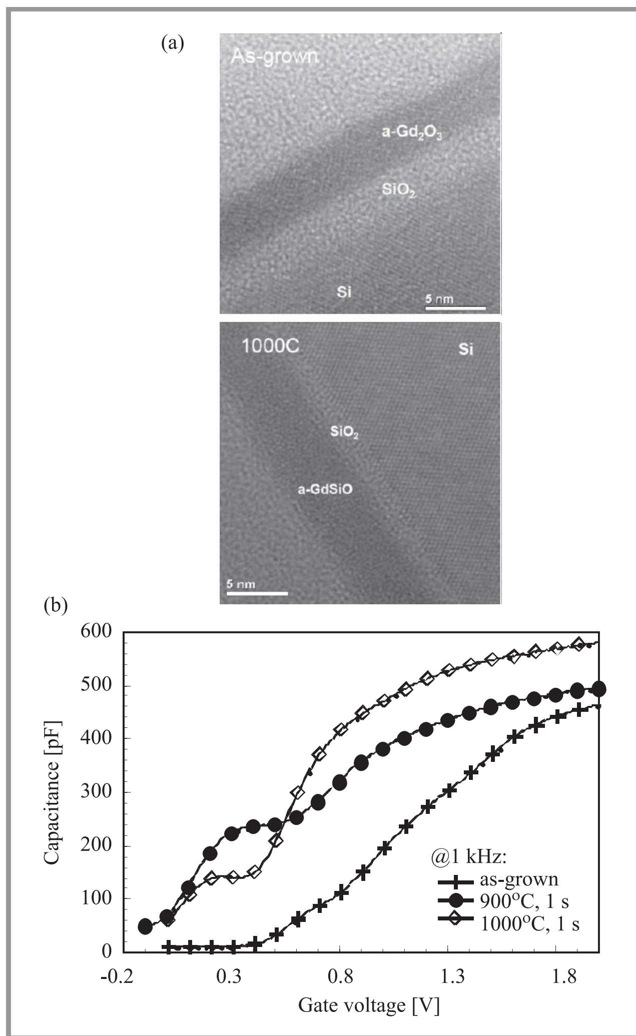


Fig. 1. (a) HRTEM images of Gd_2O_3/SiO_2 gate stacks as-deposited by ALD (top) and after RTA at 1000°C for 1 s in N_2 (bottom); (b) *C-V* plots of corresponding MIS devices (Au/*hi-k*/SiO_2/Si(100)).

The HRTEM images of ALD Gd_2O_3/SiO_2 gate stacks as-deposited and after the RTA @1000°C are shown in Fig. 1(a). The high-*k* dielectric is deposited in amorphous form. Following a 1000°C anneal for 1 s, the intermixing is complete resulting in a 4.6 nm amorphous GdSiO layer. The consumption of interfacial SiO_2 layer is evident after RTA anneal. More importantly, the silicate reaction causes an increase of accumulation capacitance in the associated $C-V$ plots (Fig. 1(b)), and hence the capacitance equivalent thickness (CET) of the gate stack is reduced after RTA. Furthermore, k is ~ 16 and amorphization of the film annealed at 1000°C reduces the leakage current density ($9.9 \cdot 10^{-7}$ Acm^{-2}) by an order of magnitude compared to the 900°C sample [65].

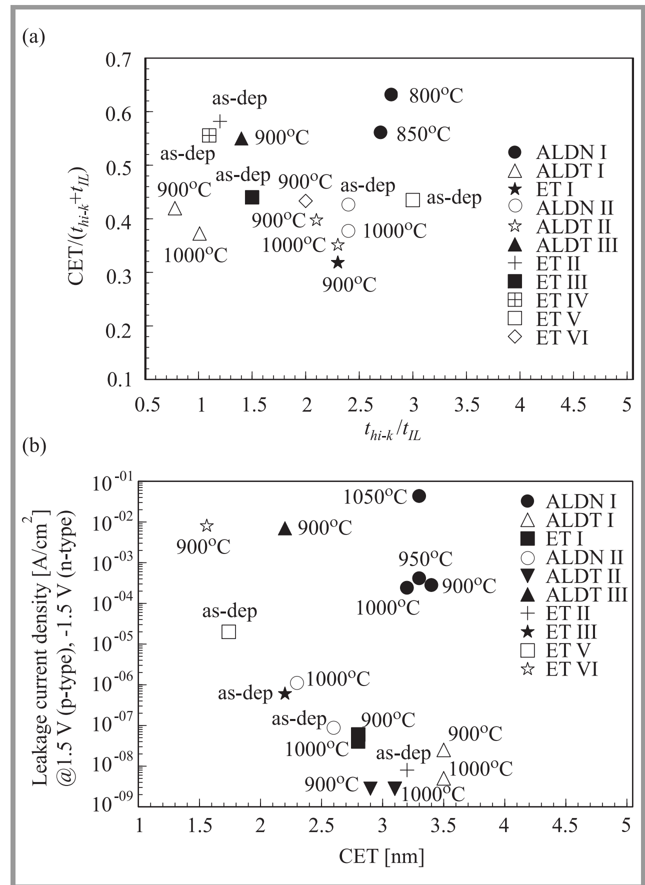


Fig. 2. (a) CET/total deposited thickness versus t_{hi-k}/t_{IL} ratio and (b) leakage versus CET for various Gd_2O_3/SiO_2 gate stacks deposited by ALD and e-beam evaporation, on thermal (T) or native oxide (N). I-VI refer to different runs, i.e., various t_{hi-k}/t_{IL} ratios.

Various ratios of the physical thicknesses of the Gd_2O_3/SiO_2 (t_{hi-k}/t_{IL}) gate stack have been explored [64] as shown in Fig. 2. By plotting CET/($t_{hi-k}+t_{IL}$) versus t_{hi-k}/t_{IL} for 1 s anneal time, it can be seen that a value $t_{hi-k}/t_{IL} \sim 2-2.5$ gives the optimal scaling, that is, the smallest CET of the GdSiO gate dielectric stack. However, this observation needs to be taken equally with the effects of mechanisms involved during RTA anneals on scaling, that is, the way the RTA is performed has shown to be of crucial importance [49]. Specifically, for the same thickness ratios achieved by two deposition processes, e-beam evaporation and ALD, different scaling/CET can be seen after RTA from Fig. 2(b). For ALD processes, RTA was performed in an inert ambient (N_2 or Ar) – RTA1, while

for e-beam evaporation in addition to RTA1, an inert gas is used but with open vacuum valve referred to as RTA2 [49]. GdSiO formation at 900°C by RTA1 results only in slight reduction of CET and this is found to be due to residual oxygen in the process chamber that can diffuse through the films and react at the Si/SiO_2 interface. This is a parasitic effect rather than an intrinsic feature of RTA1; as a result, CET cannot be further reduced as IL re-growth counteracts the consumption due to silicate formation. This explains an increased CET for ALD deposited stacks. The use of RTA2 with open vacuum valve was seen to cause significant reduction of IL re-growth and thus consumption of the IL dominates during silicate formation which further enhances scaling. Using the RTA2 in combination with a preceding post deposition annealing (PDA) treatment enabled to achieve GdSiO gate stack with an EOT of 1.3 nm and $j = 0.02$ Acm^{-2}, in line with ITRS LSTP targets for the 22 nm node [49]. This stack can be optimized further for scaling requirements beyond the 22 nm node as GdSiO has been successfully introduced into fully functional SOI n-MOSFETs with TiN metal gate electrodes [74].

3.2.1. Thermal Stability and Mechanism of GdSiO Formation

The thermal behavior of these Gd_2O_3/SiO_2 gate stacks can be assessed by plotting MIS device leakage current and CET versus RTA process temperature as shown in Fig. 3(a) and Fig. 3(b), respectively. It can be seen that very low leakage currents ($< 10^{-7}$ A/cm^2) are obtained for stacks deposited by both techniques (evaporation and ALD), in particular when there is a thermal oxide as an IL. Furthermore, an increase from 900°C to 1000°C does not compromise the leakage current; it is further reduced. The observed trend is a significant reduction of the capacitance equivalent thickness after annealing as outlined in Fig. 3(b). When observing the RTA time series, very low leakage currents of $\sim 10^{-9}$ A/cm^2 were obtained for 10 s and 100 s anneal time (Fig. 3(a)), however the CET is largely increased after 100 s anneal (Fig. 3(b)). The CET is reduced further when anneal time varies from 1 to 10 s.

Oxygen has been found to diffuse into the film eliminating oxygen vacancies, but Si diffusion was absent after oxygen and vacuum annealing at temperatures up to 800°C for GdSiO films on Si(100) [75]. It has been suggested [23] that silicate formation for rare earth based materials occurs through inter-diffusion with underlying SiO_2 to form silicates, rather than by diffusion of Si and subsequent oxidation, as this can explain absence of silicates when capped structures are used [76]–[78]. It seems that excluding oxygen and preventing the oxidation of the silicon substrate can prevent silicate formation. There are also reports on GdSiO gate dielectric films deposited on Si(001) substrates using UHV (ultra high-vacuum) e-beam evaporation from pressed-powder targets [24]. These films were amorphous as deposited and remained amorphous when annealed to temperatures up to 900°C, and showed $k \sim 16$ and low leakage currents of $5.7 \cdot 10^{-3}$ A/cm^2 at 1 V.

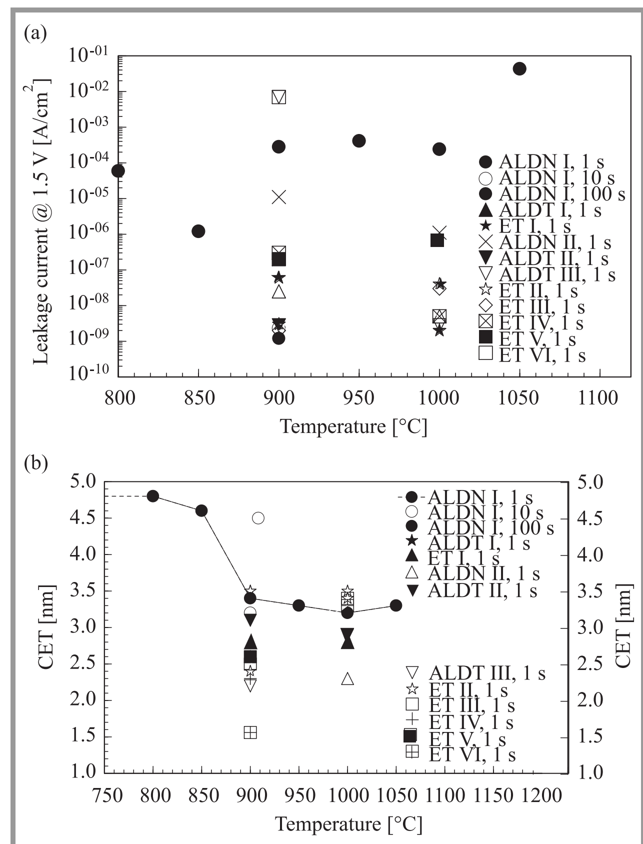


Fig. 3. (a) Leakage current and (b) CET versus anneal temperature for various Gd_2O_3/SiO_2 gate stacks deposited by ALD and e-beam evaporation.

For the ALD Gd_2O_3/SiO_2 gate stacks discussed here, annealing causes substantial reordering of the layers and the effects are apparent by plotting the MEIS energy spectra as depth profiles for Gd, Si and O shown in Fig. 4(a). It can be seen that higher temperatures drive more silicon into the high-k layer. The source of the incorporated Si is likely to be from inter-diffusion with the native SiO_2 layer, with some contribution from the substrate [65]. Diffraction patterns (XRD) from samples annealed in the temperature range from 800–1050°C contain a peak at 28.6° referring to cubic Gd_2O_3. The most pronounced peaks, corresponding to the highest level of crystallinity in the films, occur around 900°C to 950°C. The MEIS and XRD results suggest that two competing mechanisms occur during annealing [65]. The ALD Gd_2O_3 layer crystallizes into the cubic phase at all temperatures studied and the extent of crystallinity increases with increasing temperature. The second mechanism is the diffusion of Si into the layer. As Si is swept into the polycrystalline Gd_2O_3 the crystalline grains become amorphised. This reduces the thickness of the crystalline layer and thus the intensity of the XRD diffraction features. Crystallisation dominates at lower temperatures and silicate formation dominates at higher temperatures. It follows that the final state of a film after annealing depends on the film thickness as well as the anneal temperature and duration, as expected for a diffusion driven process [23].

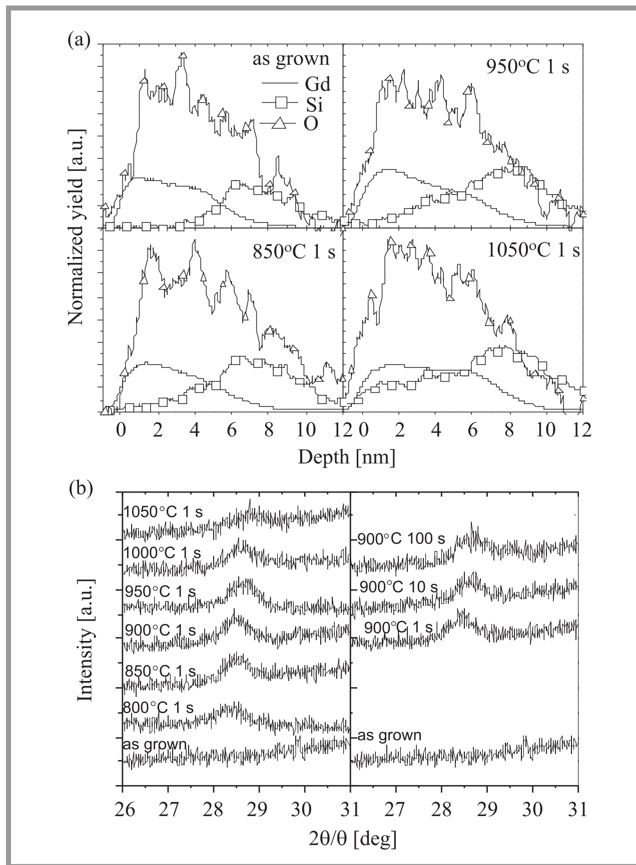


Fig. 4. (a) MEIS and (b) XRD profiles for ALD grown Gd_2O_3/ SiO_2 gate stacks under different RTA conditions [65].

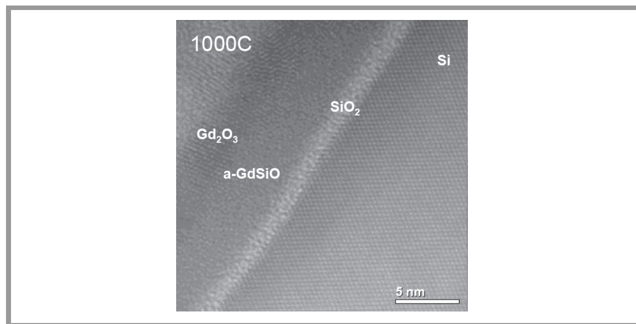


Fig. 5. HRTEM image of ALD bi-layer structure Gd_2O_3/ $GdSiO/SiO_2$ after RTA at 1000°C for 1 s in N_2.

Figure 5 depicts the case when, even after 1000°C anneal, the layer is not fully transformed into the silicate but contains bi-layer structure due to the initially deposited thicker (∼ 8 nm) Gd_2O_3 film. Note also that the interfacial SiO_2 layer is present after the anneal. Similar has been observed for other RE based silicates [23], [50], [52].

3.2.2. Band Offsets for GdSiO

An insight into the energy band line-ups of ALD GdSiO gate stacks is provided by XPS measurements. The onset of the excitation from the valence to conduction bands can be observed at an energy corresponding to the band gap energy below the XPS O 1 s core signal [79], [80]. From the threshold energy of an energy loss spectrum for O 1 s photoelectrons, the bandgap of GdSiO is determined to be 6.3 eV [81]. The measured XPS valence band spectrum for $GdSiO/SiO_2/Si$ is shown in Fig. 6(a). The valence band offset between GdSiO and Si is determined by evaluating the energies of the valence band maxima of GdSiO and Si. These energies were determined by analytically finding the intersection of the regression-determined line segment defining the leading edge of the valence band and the flat energy distribution curve [82], and taking into account charging effects on the XPS spectra [83], [84].

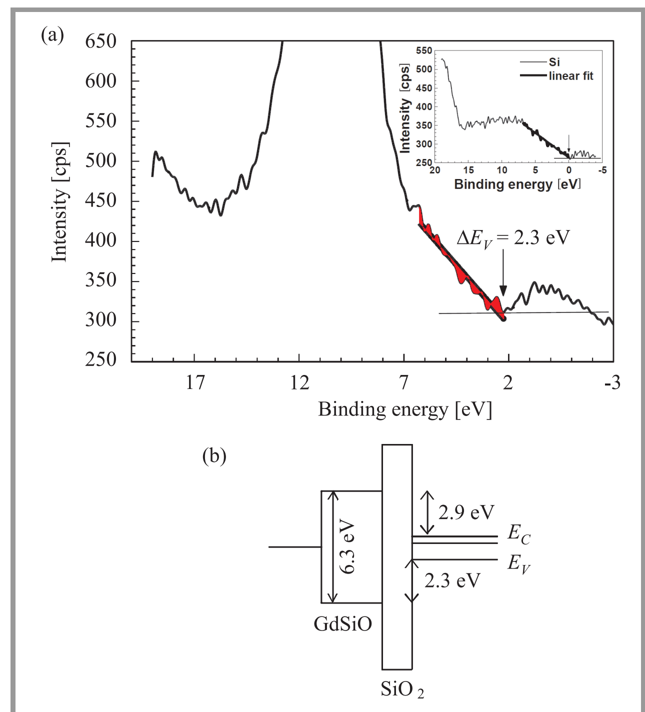


Fig. 6. (a) The XPS valence band spectrum and (b) the band diagram of ALD $GdSiO/SiO_2/Si(100)$ structure.

The values of 2.31 eV for GdSiO and 0.01 eV for Si(100) are estimated. The difference of these values corresponds to a valence band offset of 2.3 eV. Thus the conduction band offset is 2.9 eV. The resulting energy band alignment for a $GdSiO/SiO_2/Si(100)$ is shown in Fig. 6(b). The results are comparable to the ones published by Hattori *et al.* [85] for Gd_2O_3/silicate/SiO_2/Si, where $E_g = 6.4$ eV, $\Delta E_V = 2.2$ eV and $\Delta E_C = 3.1$ eV. For advanced CMOS application, the band offsets to Si above 2 eV and roughly equal ($\Delta E_C \approx \Delta E_V$) represent one of the most desirable features for future gate stack.

4. Beyond the 22 nm Node

Recent reports indicate that lanthanum-based ternary oxides are likely to have major role in meeting the ITRS requirements for scaling beyond the 22 nm node [86]. The growth methods proposed include molecular beam deposition (MBD) [87], [88], pulsed laser deposition (PLD) [89], or ALD [90]. The interfacial layers can be avoided when

ternary rare earth oxide films ($La_xM_{2-x}O_3$, M = Sc, Lu, or Y) are deposited on Si by ALD from metal amidinate precursors and H_2O [86]. Both $LaScO_3$ and $LaLuO_3$ films are found to be amorphous and free of interfacial layers, with high dielectric constants (~ 23 for $LaScO_3$ and 28 ± 1 for $LaLuO_3$), low leakage current density, and are scalable to EOT< 1 nm. $La_{1.23}Y_{0.77}O_3$ films have polycrystalline structures with moderately high $k = 17 \pm 1.3$ and low leakage current [86]. The growth of stoichiometric and smooth $LaLuO_3$ films ($C_E \sim 65$ eV) that remain amorphous up to 1000°C has been reported [89]. The band gap has been found to be 5.2 ± 0.1 eV, with symmetrical conduction and valence band offsets of 2.1 eV, a dielectric constant of ~ 32, and low leakage current density levels. Amorphous $GdScO_3$ films have also demonstrated a high permittivity of 22, the EOT of ~ 1 nm, and the leakage current density less than 2 mA/cm^2 [91].

$LaAlO_3$-based heterostructures are also expected to fulfil the requirements of ITRS beyond 22 nm [92]. Very aggressive scaling < 1 nm EOT with reasonable leakage currents can be achieved for amorphous $LaAlO_3$ (LAO) films on Si(001), but interfacial SiO_2 grows at anneals above 400°C. The addition of an Al_2O_3 interlayer increases the thermal stability up to 600°C. Both a-$LaAlO_3$/Si(001) and a-$LaAlO_3/\gamma-Al_2O_3$/Si(001) systems should be appropriate for gate-first processes, when optimized. Suzuki *et al.* at Toshiba Labs have reported [93] 0.3 nm EOT for amorphous $LaAlO_3$ grown by PLD. It should be noted that LAO often exhibits structural instability and interface reaction during high temperature treatment [94], [95].

There are few publications about ultra-thin high-*k* films with EOTs lower than 1 nm on crystalline Gd_2O_3 grown by MBE [62], [72]. There has been a recent attempt to grow epitaxially ternary $(Nd_{1-x}Gd_x)_2O_3$ (NGO) thin films, with the idea that a combination of Gd_2O_3 and Nd_2O_3 would create a system exhibiting exact lattice matching with Si [96]. The NGO films show promising electrical features, the CET of 0.9 nm and leakage currents below 1 mA/cm^2. The key parameter for IL control is found to be oxygen partial pressure during the interface formation and/or MBE growth; it prevents silicide inclusions, while avoiding the formation of interfacial SiO_x [63], [71]. GdSiO material, grown by e-beam evaporation or ALD, as discussed in this paper, has potential for further scaling and is also contender as a high-*k* material beyond the 22 nm node.

Although there seem to be several possibilities to engineer the gate stack for ultimate scaling, most of the high-*k* materials reported above show high mid-gap density of interface states ($D_{it} = 3 \cdot 10^{11}$ eV^{-1}cm^{-2} [91], $1.4 \cdot 10^{12}$ eV^{-1}cm^{-2} [96], $5 \cdot 10^{12}$ eV^{-1}cm^{-2} [95]), and bulk fixed charge density $> 10^{11}$ cm^{-2} [91], [96]. On the pathway of scaling, it is critical to obtain an optimal high-*k*-Si interface with acceptably low $D_{it} < 10^{11}$ eV^{-1}cm^{-2} and mobility in MOSFET channels approaching the universal curve. At the same time, the problem of reliability and stability (charge trapping) [97], [98] of these new dielectrics still remains acute and requires further studies.

5. Summary

In this paper, recent work on rare earth based oxides and silicates has been reviewed with an emphasis on materials suitable for integration according to the ITRS LSTP targets for the 22 nm node and beyond. Understanding the mechanisms that create interfacial layers is a key requirement for further scaling of these high-*k* dielectrics. There are several approaches to achieve elimination of the interfacial SiO_2 layer and thus ultimate scaling, employed by various research labs worldwide. Reaction of the SiO_2 with a RE oxide causes an effective increase for the *k*-value of the interfacial layer. Furthermore, full reaction of the interfacial layer with cap can be used to form a higher-*k* dielectric, or the formation of epitaxial high-*k* dielectrics can be utilised. The scaling potential of GdSiO-based gate stacks which exhibit excellent thermal stability, low leakage currents and sufficiently high band offsets to be employed in the devices for technological nodes beyond 22 nm has been demonstrated.

Acknowledgements

The authors collaborate under the banner of the high-*k* gang (http://www.high-k-gang.eu/). The work has benefited from funding provided by the EC PULLNANO (Academic Cluster), the Network of Excellence NANOSIL and EPSRC, UK. The authors are part of the SINANO Institute.

References

[1] S. Natarajan, M. Armstrong, M. Bost, R. Brain, and M. Brazier, "A 32 nm logic technology featuring 2nd generation high-*k* metal-gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291 Mb array", in *Proc. IEEE IEDM 2008 Conf.*, San Francisco, USA, 2008, pp. 941–943.

[2] H. S. Yang *et al.*, "Scaling of 32 nm low power SRAM with high-*k* metal gate", in *Proc. IEEE IEDM 2008 Conf.*, San Francisco, USA, 2008, pp. 233–236.

[3] H. Kawasaki *et al.*, "Demonstration of highly scaled FinFET SRAM cells with high-*k*/metal gate and investigation of characteristic variability for the 32 nm node and beyond", in *Proc. IEEE IEDM 2008 Conf.*, San Francisco, USA, 2008, pp. 237–240.

[4] C. H. Diaz *et al.*, "32 nm gate-first high-*k*/metal-gate technology for high performance low power applications", in *Proc. IEEE IEDM Tech. Dig.*, San Francisco, USA, 2008, pp. 629–632.

[5] S. Hasegawa *et al.*, "A cost-conscious 32 nm CMOS platform technology with advanced single exposure lithography and gate-first metal gate/high-*k* process", in *Proc. IEEE IEDM 2008 Conf.*, San Francisco, USA, 2008, pp. 938–940.

[6] K. Mistry *et al.*, "A 45 nm logic technology with high-*k*+ metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging", in *Proc. IEEE IEDM Tech. Dig.*, Washington, USA, 2007, pp. 247–250.

[7] S. Mayuzumi, S. Yamakawa, Y. Tateshita, T. Hirano, M. Nakata, S. Yamaguchi, K. Tai, H. Wakabayashi, M. Tsukamoto, and N. Nagashima, "High-performance metal/high-*k* n- and p-MOSFETs with top-cut dual stress liners using gate-last damascene process on (100) substrates", *IEEE Trans. Electron Dev.*, vol. 56, no. 4, pp. 620–626, 2009.

[8] X. Chen *et al.*, "A cost effective 32 nm high-*k*/metal gate CMOS technology for low power applications with single-metal/gate-first process", in *Proc. IEEE VLSI Tech. Symp.*, Honolulu, Hawaii, USA, 2008, pp. 88–89.

[9] M. Chudzik *et al.*, "High-performance high-*k* metal gates for 45 nm CMOS and beyond with gate-first processing", in *Proc. IEEE VLSI Tech. Symp.*, Kyoto, Japan, 2007, pp. 194–195.

[10] S. Tyagi *et al.*, "Future device scaling – beyond traditional CMOS", in *Proc. IEDST'09 Conf.*, Mumbai, India, 2009, pp. 1–4.

[11] "International Technology Roadmap for Semiconductors (ITRS)" [Online]. Available: http://public.itrs.net

[12] H. J. Osten, "Epitaxial high-*k* dielectrics on silicon", in *Proc. IEEE ASDAM 2004 Conf.*, Smolenice, Slovakia, 2004, pp. 155–162.

[13] Y. Ma, Y. Ono, L. Stecker, D. R. Evans, and S. T. Hsu, "Zirconium oxide based gate dielectrics with equivalent oxide thickness of less than 1.0 nm and performance of submicron MOSFET using a nitride gate replacement process", in *Proc. IEEE IEDM Conf.*, Washington, USA, 1999, pp. 149–152.

[14] B. H. Lee, L. Kang, W.-J. Qi, R. Nieh, Y. Jeon, K. Onishi, and J. C. Lee, "Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application", in *Proc. IEEE IEDM Conf.*, Washington, USA, 1999, pp. 133–136.

[15] S. Jeon, C.-J. Choi, T.-Y. Seong, and H. Hwang, "Electrical characteristics of ZrO_xN_y prepared by NH_3 annealing of ZrO_2", *Appl. Phys. Lett.*, vol. 79, iss. 2, pp. 245–247, 2001.

[16] J. Kwo *et al.*, "Properties of high *k* gate dielectrics Gd_2O_3 and Y_2O_3 for Si", *J. Appl. Phys.*, vol. 89, iss. 7, pp. 3920–3927, 2001.

[17] O. Engstrom, B. Raeissi, S. Hall, O. Buiu, M. C. Lemme, H. D. B. Gottlob, P. K. Hurley, and K. Cherkaoui, "Navigation aids in the search for future future high-*k* dielectrics: physical and electrical trends", *Solid-State Electron.*, vol. 51, iss. 4, pp. 622–626, 2007.

[18] J. Robertson, "High dielectric constant oxides", *Eur. Phys. J. Appl. Phys.*, vol. 28, no. 3, pp. 265–291, 2004.

[19] L. Pantisano, T. Schram, B. O'Sullivan, T. Conard, S. De Gendt, G. Groeseneken, P. Zimmerman, A. Akheyar, M. M. Heyns, S. Shamuilla, V. V. Afanas'ev, and A. Stesmans, "Effective work function modulation by controlled dielectric monolayer deposition", *Appl. Phys. Lett.*, vol. 89, iss. 11, pp. 113505-1–113505-3, 2006.

[20] L.-A. Ragnarsson, V. S. Chang, H. Y. Yu, H.-J. Cho, T. Conard, K. M. Yin, A. Delabie, J. Swerts, T. Schram, S. De Gendt, and S. Biesemans, "Achieving conduction band-edge effective work functions by La_2O_3 capping of hafnium silicates", *IEEE Electron Dev. Lett.*, vol. 28, no. 6, pp. 486–488, 2007.

[21] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-*k* gate dielectrics: current status and materials properties considerations", *J. Appl. Phys.*, vol. 89, iss. 10, pp. 5243–5275, 2001.

[22] I. Z. Mitrovic, O. Buiu, S. Hall, C. Bungey, T. Wagner, W. Davey, and Y. Lu, "Electrical and structural properties of hafnium silicate thin films", *Microelectron. Reliab.*, vol. 47, iss. 4–5, pp. 645–648, 2007.

[23] S. Van Elshocht, C. Adelmann, T. Conard, A. Delabie, A. Franquet, L. Nyns, O. Richard, P. Lehnen, J. Swerts, and S. De Gendt, "Silicate formation and thermal stability of ternary rare earth oxides as high-*k* dielectrics", *J. Vac. Sci. Technol. A*, vol. 26, no. 4, pp. 724–730, 2008.

[24] J. A. Gupta, D. Landheer, J. P. McCaffrey, and G. I. Sproule, "Gadolinium silicate gate dielectric films with sub-1.5 nm equivalent oxide thickness", *Appl. Phys. Lett.*, vol. 78, iss. 12, pp. 1718–1720, 2001.

[25] M. Copel, E. Cartier, and F. M. Ross, "Formation of a stratified lanthanum silicate dielectric by reaction with Si(001)", *Appl. Phys. Lett.*, vol. 78, iss. 11, pp. 1607–1609, 2001.

[26] G. Lupina, T. Schroeder, C. Wenger, J. Dabrowski, and H.-J. Müssig, "Thermal stability of Pr silicate high-*k* layers on Si(001)", *Appl. Phys. Lett.*, vol. 89, iss. 22, pp. 222909-1–222909-3, 2006.

[27] A. Sakai, S. Sakashita, M. Sakashita, Y. Yasuda, S. Zaima, and S. Miyazaki, "Praseodymium silicate formed by postdeposition high-temperature annealing", *Appl. Phys. Lett.*, vol. 85, no. 22, pp. 5322–5324, 2004.

[28] H. Iwai, S. Ohmi, S. Akama, C. Ohshima, A. Kikuchi, I. Kashiwagi, J. Taguchi, H. Yamamoto, J. Tonotani, Y. Kim, I. Ueda, A. Kuriyama, and Y. Yoshihara, "Advanced gate dielectric materials for sub-100 nm CMOS", in *Proc. IEEE IEDM Conf.*, San Francisco, USA, 2002, pp. 625–628.

[29] D. P. Norton, "Synthesis and properties of epitaxial electronic oxide thin-film materials", *Mater. Sci. Eng. R*, vol. 43, iss. 5–6, pp. 139–247, 2004.

[30] P. Delugas and V. Fiorentini, "Dielectric properties of two phases of crystalline lutetium oxide", *Microelectron. Reliab.*, vol. 45, iss. 5–6, pp. 831–833, 2005.

[31] H. Ono and T. Katsumata, "Interfacial reactions between thin rare-earth-metal oxide films and Si substrates", *Appl. Phys. Lett.*, vol. 78, no. 13, pp. 1832–1834, 2001.

[32] J. Wu, M. Y. Yang, A. Chin, W. J. Chen, and C. M. Kwei, "Electrical characteristics of high quality La_2O_3 gate dielectric with equivalent oxide thickness of 5 Å", *IEEE Electron. Dev. Lett.*, vol. 21, no. 7, pp. 341–343, 2000.

[33] J.-B. Cheng, A.-D. Li, Q.-Y. Shao, H.-Q. Ling, D. Wu, Y. Wang, Y.-J. Bao, M. Wang, Z.-G. Liu, and N.-B. Ming, "Growth and characteristics of La_2O_3 gate dielectric prepared by low pressure metalorganic chemical vapor deposition", *Appl. Surf. Sci.*, vol. 233, iss. 1–4, pp. 91–98, 2004.

[34] H. J. Osten, J. P. Liu, and H. J. Mussig, "Band gap and band discontinuities at crystalline Pr_2O_3/Si(001) heterojunctions", *Appl. Phys. Lett.*, vol. 80, iss. 2, pp. 297–299, 2002.

[35] R. Lo Nigro, V. Raineri, C. Bongiorno, R. Toro, G. Malandrino, and I. L. Fragala, "Dielectric properties of Pr_2O_3 high-*k* films grown by metalorganic chemical vapor deposition on silicon", *Appl. Phys. Lett.*, vol. 83, iss. 1, pp. 129–131, 2003.

[36] R. Lo Nigro, R. G. Toro, G. Malandrino, V. Raineri, and I. L. Fragala, "A simple route to the synthesis of Pr_2O_3 high-*k* thin films", *Adv. Mater.*, vol. 15, iss. 13, pp. 1071–1075, 2003.

[37] T.-M. Pan, F.-J. Tsai, C.-I. Hsieh, and T.-W. Wu, "Structural properties and electrical characteristics of praseodymium oxide gate dielectrics", *Electrochem. Solid-State Lett.*, vol. 10, no. 4, pp. G21–G24, 2007.

[38] J. Kwo *et al.*, "High ε gate dielectrics Gd_2O_3 and Y_2O_3 for silicon", *Appl. Phys. Lett.*, vol. 77, iss. 1, pp. 130–132, 2000.

[39] M. D. Kannan, S. K. Narayandass, C. Balasubramanian, and D. Mangalaraj, "Structure and electrical properties of thermally evaporated Nd_2O_3 thin films", *Phys. Stat. Sol. A*, vol. 128, iss. 2, pp. 427–433, 1991.

[40] A. Fissel, Z. Elassar, O. Kirfel, E. Bugiel, M. Czernohorsky, and H. J. Osten, "Interface formation during molecular beam epitaxial growth of neodymium oxide on silicon", *J. Appl. Phys.*, vol. 99, iss. 7, pp. 074105-1–074105-6, 2006.

[41] T. Busani and R. A. B. Devine, "The importance of network structure in high-*k* dielectrics: $LaAlO_3$, Pr_2O_3 and Ta_2O_5", *J. Appl. Phys.*, vol. 98, iss. 4, pp. 044102-1–044102-5, 2005.

[42] K. J. Hubbart and D. G. Schlom, "Thermodynamic stability of binary oxides in contact with silicon", *J. Mater. Res.*, vol. 11, no. 11, pp. 2757–2776, 1996.

[43] M. Copel, E. Cartier, V. Narayanan, M. C. Reuter, S. Guha, and N. Bojarczuk, "Characterization of silicate/Si(001) interfaces", *Appl. Phys. Lett.*, vol. 81, iss. 22, pp. 4227–4229, 2002.

[44] C.-J. Choi, M.-G. Jang, Y.-Y. Kim, M.-S. Jun, T.-Y. Kim, and M.-H. Song, "Electrical and structural properties of high-*k* Er-silicate gate dielectric formed by interfacial reaction between Er and SiO_2 films", *Appl. Phys. Lett.*, vol. 91, iss. 1, pp. 012903-1–012903-3, 2007.

[45] G. Lupina, T. Schroeder, J. Dabrowski, C. Wenger, A. U. Mane, H.-J. Mussig, P. Hoffmann, and D. Schmeisser, "Praseodymium silicate films on S(100) for gate dielectric applications: physical and electrical characterization", *J. Appl. Phys.*, vol. 99, iss. 11, pp. 114109-1–114109-5, 2006.

[46] M. Copel, "Selective desorption of interfacial SiO_2", *Appl. Phys. Lett.*, vol. 82, iss. 10, pp. 1580–1582, 2003.

[47] D. J. Lichtenwalner *et al.*, “Lanthanum silicate gate dielectric stacks with subnanometer equivalent oxide thickness utilizing an interfacial silica consumption reaction”, *J. Appl. Phys.*, vol. 98, iss. 2, pp. 024314-1–024314-6, 2005.

[48] A. Laha, A. Fissel, and H. J. Osten, “Engineering the interface between epitaxial lanthanide oxide thin films and Si substrates: a route towards tuning the electrical properties”, *Microelectron. Eng.*, vol. 84, iss. 9–10, pp. 2282–2285, 2007.

[49] H. D. B. Gottlob, M. Schmidt, A. Stefani, M. C. Lemme, H. Kurz, I. Z. Mitrovic, W. M. Davey, S. Hall, M. Werner, P. R. Chalker, K. Cherkaoui, P. K. Hurley, J. Piscator, O. Engstrom, and S. B. Newcomb, “Scaling potential and MOSFET integration of thermally stable Gd silicate dielectrics”, *Microelectron. Eng.*, vol. 86, iss. 7–9, pp. 1642–1645, 2009.

[50] H. D. B. Gottlob, M. Schmidt, M. C. Lemme, H. Kurz, I. Z. Mitrovic, M. Werner, W. M. Davey, S. Hall, P. R. Chalker, K. Cherkaoui, P. K. Hurley, B. Raeissi, O. Engstrom, and S. B. Newcomb, “Gd silicate: a high-*k* dielectric compatible with high temperature annealing”, *J. Vac. Sci. Technol. B*, vol. 27, no. 1, pp. 249–252, 2009.

[51] D. Eom, S. Y. No, C. S. Hwang, and H. J. Kim, “Deposition characteristics and annealing effect of La_2O_3 films prepared using $La(iPrCp)_3$ precursor”, *J. Electrochem. Soc.*, vol. 154, iss. 3, pp. G49–G53, 2007.

[52] X. Wu, D. Landheer, G. I. Sproule, T. Quance, M. J. Graham, and G. A. Botton, “Characterization of gadolinium and lanthanum oxide films on Si (100)”, *J. Vac. Sci. Technol. A*, vol. 20, no. 3, pp. 1141–1144, 2002.

[53] S. Guha, E. Cartier, M. A. Gribelyuk, N. A. Bojarczuk, and M. C. Copel, “Atomic beam deposition of lanthanum- and yttrium-based oxide thin films for gate dielectrics”, *Appl. Phys. Lett.*, vol. 77, iss. 17, pp. 2710–2712, 2000.

[54] A. C. Jones, “Molecular design of improved precursors for the MOCVD of electroceramic oxides”, *J. Mater. Chem.*, vol. 12, no. 9, pp. 2576–2590, 2002.

[55] A. C. Jones, H. C. Aspinall, P. R. Chalker, R. J. Potter, K. Kukli, A. Rahtu, M. Ritala, and M. Leskala, “Recent developments in the MOCVD and ALD of rare earth oxides and silicates”, *Mater. Sci. Eng. B*, vol. 118, iss. 1–3, pp. 97–104, 2005.

[56] R. Lupták, K. Fröhlich, A. Rosová, K. Hušeková, M. Tapajna, D. Machajdík, M. Jergel, J. P. Espinós, and C. Mansilla, “Growth of gadolinium oxide films for advanced MOS structure”, *Microelectron. Eng.*, vol. 80, pp. 154–157, 2005.

[57] M. P. Singh, C. S. Thakur, K. Shalini, S. Banerjee, N. Bhat, and S. A. Shivashankar, “Structural, optical, and electrical characterization of gadolinium oxide films deposited by low-pressure metalorganic chemical vapour deposition”, *J. Appl. Phys.*, vol. 96, no. 10, pp. 5631–5637, 2004.

[58] B. A. Orlowski, E. Guziewicz, N. E. Orlowska, A. Bukowski, and R. L. Johnson, “Photoemission study of Gd on clean Si(111) surface”, *Surf. Sci.*, vol. 507–510, pp. 218–222, 2002.

[59] P. Y. Kuei and C. C. Hu, “Gadolinium oxide high-*k* gate dielectrics prepared by anodic oxidation”, *Appl. Surf. Sci.*, vol. 254, iss. 17, pp. 5487–5491, 2008.

[60] H.-H. Ko, L.-B. Chang, M.-J. Jeng, P.-Y. Kuei, and K.-Y. Horng, “Properties of thermal gadolinium oxide films on silicon”, *Jap. J. Appl. Phys.*, vol. 44, no. 5A, pp. 3205–3208, 2005.

[61] L.-Z. Hsieh, H.-H. Ko, P.-Y. Kuei, L.-B. Chang, and M.-J. Jeng, “Hysteresis in gadolinium oxide metal-oxide-semiconductor capacitors”, *J. Appl. Phys.*, vol. 98, iss. 7, pp. 076110-1–076110-3, 2005.

[62] A. Laha, H. J. Osten, and A. Fissel, “Influence of interface layer composition on the electrical properties of epitaxial Gd_2O_3 thin films for high-*k* application”, *Appl. Phys. Lett.*, vol. 90, iss. 11, pp. 113508-1–113508-3, 2007.

[63] M. Czernohorsky, E. Bugiel, H. J. Osten, A. Fissel, and O. Kirfel, “Impact of oxygen supply during growth on the electrical properties of crystalline Gd_2O_3 thin films on Si(001)”, *Appl. Phys. Lett.*, vol. 88, iss. 15, pp. 152905-1–152905-3, 2006.

[64] I. Z. Mitrovic, M. Werner, W. M. Davey, S. Hall, P. R. Chalker, H. D. B. Gottlob, M. C. Lemme, O. Engstrom, K. Cherkaoui, and P. K. Hurley, “Quest for an optimal gadolinium silicate gate dielectric stack”, in *39th Conf. IEEE SISC 2008*, San Diego, USA, 2008.

[65] M. Werner, P. R. Chalker, W. M. Davey, I. Z. Mitrovic, S. Hall, and I. Alexandrou, “Formation of high-*k* gadolinium silicate via silicon oxide inter-diffusion into gadolinium oxide”, *Appl. Phys. Lett.*, 2009 (submitted).

[66] J. W. Johnson *et al.*, “Gd_2O_3/GaN metal-oxide-semiconductor field-effect transistor”, *Appl. Phys. Lett.*, vol. 77, iss. 20, pp. 3230–3232, 2000.

[67] M. Hong *et al.*, “Single-crystal GaN/Gd_2O_3/GaN heterostructure”, *J. Vac. Sci. Technol. B*, vol. 20, iss. 3, pp. 1274–1277, 2002.

[68] M. Hong, M. Passlack, J. P. Mannaerts, J. Kwo, S. N. G. Chu, N. Moriya, S. Y. Hou, and V. J. Fratello, “Low interface state density oxide-GaAs structures fabricated by in situ molecular beam epitaxy”, *J. Vac. Sci. Technol. B*, vol. 14, iss. 3, pp. 2297–2300, 1996.

[69] M. Hong, J. Kwo, A. R. Kortan, J. P. Mannaerts, and A. M. Sergent, “Epitaxial cubic gadolinium oxide as a dielectric for gallium arsenide passivation”, *Science*, vol. 283, no. 5409, pp. 1897–1900, 1999.

[70] D. Jia, L. Lu, and W. M. Yu, “Erbium energy levels relative to the band gap of gadolinium oxide”, *Opt. Commun.*, vol. 212, iss. 1–3, pp. 97–100, 2002.

[71] A. Fissel, M. Czernohorsky, and H. J. Osten, “Characterization of crystalline rare-earth oxide high-*k* dielectrics grown by molecular beam epitaxy on silicon carbide”, *J. Vac. Sci. Technol. B*, vol. 24, no. 4, pp. 2115–2118, 2006.

[72] H. D. B. Gottlob *et al.*, “0.86-nm CET gate stacks with epitaxial Gd_2O_3 high-*k* dielectrics and FUSI NiSi metal electrodes”, *IEEE Electron Dev. Lett.*, vol. 27, no. 10, pp. 814–816, 2006.

[73] M. Czernohorsky, D. Tetzlaff, E. Bugiel, R. Dargis, H. J. Osten, H. D. B. Gottlob, M. Schmidt, M. C. Lemme, and H. Kurz, “Stability of crystalline Gd_2O_3 thin films on silicon during rapid thermal annealing”, *Semicond. Sci. Technol.*, vol. 23, no. 3, pp. 035010-1–035010-4, 2008.

[74] M. Schmidt, A. Stefani, H. D. B. Gottlob, and H. Kurz, “Integration of Gd silicate/TiN gate stacks into SOI n-MOSFETs”, *Microelectron. Eng.*, vol. 86, iss. 7–9, pp. 1683–1685, 2009.

[75] D. Landheer, X. Wu, J. Morais, I. J. R. Baumvol, R. P. Pezzi, L. Miotti, W. N. Lennard, and J. K. Kim, “Thermal stability and diffusion in gadolinium silicate gate dielectric films”, *Appl. Phys. Lett.*, vol. 79, no. 16, pp. 2618–2620, 2001.

[76] B. W. Busch, J. Kwo, M. Hong, J. P. Mannaerts, B. J. Sapjeta, W. H. Schulte, E. Garfunkel, and T. Gustafsson, “Interface reactions of high-*k* Y_2O_3 gate oxides with Si”, *Appl. Phys. Lett.*, vol. 79, iss. 15, pp. 2447–2449, 2001.

[77] D. Niu, R. W. Ashcraft, and G. N. Parsons, “Water absorption and interface reactivity of yttrium oxide gate dielectrics on silicon”, *Appl. Phys. Lett.*, vol. 80, iss. 19, pp. 3575–3577, 2002.

[78] A. Goryachko, J. P. Liu, D. Krüger, H. J. Osten, E. Bugiel, R. Kurps, and V. Melnik, “Thermal stability of Pr_2O_3 films grown on Si(100) substrate”, *J. Vac. Sci. Technol. A*, vol. 20, iss. 6, pp. 1860–1866, 2002.

[79] S. Miyazaki, H. Nishimura, M. Fukuda, L. Ley, and J. Ristein, “Structure and electronic states of ultrathin SiO_2 thermally grown on Si(100) and Si(111) surfaces”, *Appl. Surf. Sci.*, vol. 113–114, pp. 585–589, 1997.

[80] S. Miyazaki, “Characterization of high-*k* gate dielectric/silicon interfaces”, *Appl. Surf. Sci.*, vol. 190, iss. 1–4, pp. 66–74, 2002.

[81] I. Z. Mitrovic *et al.*, “Shift in the band-offsets and dominant trap levels in Gd-based high-*k* gate stacks”, *J. Appl. Phys.*, 2009 (in preparation).

[82] A. Fissel, J. Dabrowski, and H. J. Osten, “Photoemission and ab initio theoretical study of interface and film formation during epitaxial growth and annealing of praseodymium oxide on Si(001)”, *J. Appl. Phys.*, vol. 91, iss. 11, pp. 8986–8991, 2002.

[83] S. Toyoda *et al.*, “Precise determination of band offsets and chemical states in SiN/Si studied by photoemission spectroscopy and x-ray absorption spectroscopy”, *Appl. Phys. Lett.*, vol. 87, iss. 10, pp. 102901-1–102901-3, 2005.

[84] Y. Liu, T. P. Chen, L. Ding, S. Zhang, Y. Q. Fu, and S. Fung, "Charging mechanism in a SiO_2 matrix embedded with Si nanocrystals", *J. Appl. Phys.*, vol. 100, iss. 9, pp. 096111-1–096111-3, 2006.

[85] T. Hattori, T. Yoshida, T. Shiraishi, K. Takahashi, H. Nohira, S. Joumori, K. Nakajima, M. Suzuki, K. Kimura, I. Kashiwagi, C. Ohshima, S. Ohmi, and H. Iwai, "Composition, chemical structure, and electronic band structure of rare earth oxide/Si(100) interfacial transition layer", *Microelectron. Eng.*, vol. 72, iss. 1–4, pp. 283–287, 2004.

[86] H. Wang, J.-J. Wang, R. Gordon, J.-S. M. Lehn, H. Li, D. Hong, and D. V. Shenai, "Atomic layer deposition of lanthanum-based ternary oxides", *Electrochem. Solid-State Lett.*, vol. 12, no. 4, pp. G13–G15, 2009.

[87] J. M. J. Lopes, U. Littmark, M. Roeckerath, S. Lenk, J. Schubert, and S. Mantl, "Effects of annealing on the electrical and interfacial properties of amorphous lanthanum scandate high-*k* films prepared by molecular beam deposition", *J. Appl. Phys.*, vol. 101, iss. 10, pp. 104109-1–104109-5, 2007.

[88] J. M. J. Lopes, M. Roeckerath, T. Heeg, U. Littmark, J. Schubert, S. Mantl, Y. Jia, and D. G. Schlom, "La-based ternary rare-earth oxides as alternative high-*k* dielectrics", *Microelectron. Eng.*, vol. 84, iss. 9–10, pp. 1890–1893, 2007.

[89] J. M. J. Lopes *et al.*, "Amorphous lanthanum lutetium oxide thin films as an alternative high-*k* gate dielectric", *Appl. Phys. Lett.*, vol. 89, iss. 22, pp. 222902-1–222902-3, 2006.

[90] J. M. J. Lopes, M. Roeckerath, T. Heeg, J. Schubert, S. Mantl, and V. V. Afanas'ev, "Amorphous lantahnum lutetium oxide thin films as an alternative high-*k* material", *ECS Trans.*, vol. 11, no. 4, pp. 311–318, 2007.

[91] K. H. Kim, D. B. Farmer, J.-S. M. Lehn, P. V. Rao, and R. G. Gordon, "Atomic layer deposition of gadolinium scandate films with high dielectric constant and low leakage current", *Appl. Phys. Lett.*, vol. 89, iss. 13, pp. 133512-1–133512-3, 2006.

[92] L. F. Edge, D. G. Schlom, P. Sivasubramani, R. M. Wallace, B. Holländer, and J. Schubert, "Electrical characterization of amorphous lanthanum aluminate thin films grown by molecular-beam deposition on silicon", *Appl. Phys. Lett.*, vol. 88, iss. 11, pp. 112907-1–112907-3, 2006.

[93] M. Suzuki, T. Yamaguchi, N. Fukushima, and M. Koyama, "$LaAlO_3$ gate dielectric with ultrathin equivalent oxide thickness and ultralow leakage current directly deposited on Si substrate", *J. Appl. Phys.*, vol. 103, iss. 3, pp. 034118-1–034118-5, 2008.

[94] X. B. Lu, Z. G. Liu, Y. P. Wang, Y. Yang, X. P. Wang, H. W. Zhou, and B. Y. Nguyen, "Structure and dielectric properties of amorphous $LaAlO_3$ and $LaAlO_xN_y$ films as alternative gate dielectric materials", *J. Appl. Phys.*, vol. 94, iss. 2, pp. 1229–1234, 2003.

[95] M. Suzuki, M. Tomita, T. Yamaguchi, and N. Fukushima, "Ultra-thin (EOT = 3Å) and low leakage dielectrics of La-aluminate directly on Si substrate fabricated by high temperature deposition", in *Proc. IEEE IEDM Tech. Dig.*, Washington, USA, 2005, pp. 433–436.

[96] A. Laha, A. Fissel, E. Bugiel, H. J. Osten, "Crystalline ternary rare earth oxide with capacitance equivalent thickness below 1 nm for high-*k* application", *Appl. Phys. Lett.*, vol. 88, iss. 17, pp. 172107-1–172107-3, 2006.

[97] T.-M. Pan, C.-S. Liao, H.-H. Hsu, C.-L. Chen, J.-D. Lee, and K.-T. Wang, "Excellent frequency dispersion of thin gadolinium oxide high-*k* gate dielectrics", *Appl. Phys. Lett.*, vol. 87, iss. 26, pp. 262908-1–262908-3, 2005.

[98] A. N. Nazarov *et al.*, "Charge trapping in ultrathin Gd_2O_3 high-*k* dielectric", *Microelectron. Eng.*, vol. 84, iss. 9–10, pp. 1968–1971, 2007.

Ivona Z. Mitrovic received the Ph.D. degree in electronic engineering from the University of Liverpool, UK, in 2007, the M.Sc. degree in materials science from the University of Belgrade in 2002, and Dipl.-Ing. degree in microelectronics from the Faculty of Electronic Engineering, University of Nis, Serbia, Yugoslavia, in 1997. She took part in a research project concerning $BaTiO_3$ ceramics (1997–2001), worked as a Research Assistant (2001–2007) and a Research Associate at the University of Liverpool (2000–2009). Since June 2009, she is a Lecturer in the Solid State Electronics Research Group, Department of Electrical Engineering and Electronics, University of Liverpool. Her research interests span materials for beyond 22 nm technological node targeting energy harvesting products for medical, automotive and aerospace applications, as well as emerging technologies for energy conversion and storage.

e-mail: ivona@liverpool.ac.uk
Department of Electrical Engineering and Electronics
University of Liverpool
Brownlow Hill, Liverpool L69 3GJ, United Kingdom

Stephen Hall has been Head of the Department of Electrical Engineering and Electronics at the University of Liverpool, UK, from 2001 to the present date. He has interests spanning materials characterization, device physics and innovative device design and gate level circuits. He has over 200 conference and journal papers in the area of silicon technology, devices and circuits. These include novel measurements and contributions to the understanding of MOS related interfaces and materials quality. He has successfully designed and built novel MOS and bipolar devices in silicon for about 20 years. More recently, his work encompasses gate level circuits relating to low voltage/low power SOI, micro-power and biologically inspired concepts. He was Technical Programme Chair of ESSDERC 2008, and currently sits on the Steering Committee of ESSDERC/ESSCIRC and INFOS, for which he was vice-Chair in 2009 and is a member of the Steering Committee from 2009.

e-mail: s.hall@liverpool.ac.uk
Department of Electrical Engineering and Electronics
University of Liverpool
Brownlow Hill, Liverpool L69 3GJ, United Kingdom

Paper

Technology of MISFET with $SiO_2/BaTiO_3$ System as a Gate Insulator

Piotr Firek and Jan Szmidt

Abstract— The properties of barium titanate ($BaTiO_3$, BT), such as high dielectric constant and resistivity, allow it to find numerous applications in the field of microelectronics. In this work silicon metal-insulator-semiconductor field effect transistor (MISFET) structures with $BaTiO_3$ thin films (containing La_2O_3 admixture) acting as gate insulator were investigated. The films were produced by means of radio frequency plasma sputtering (RF PS) of sintered $BaTiO_3$ + La_2O_3 (2% wt.) target. In the paper transfer and output $I-V$, transconductance and output conductance characteristics of the obtained transistors are presented and discussed. Basic parameters of these devices, such as threshold voltage (V_{TH}) are determined and discussed.

Keywords— barium titanate, I–V characteristics, MISFET structures, radio frequency plasma sputtering.

1. Introduction

Barium titanate ($BaTiO_3$, BT) ceramics have been extensively used in the field of electronic applications. Multilayer ceramic capacitors (MLCCs) [1], [2], embedded capacitances in printed circuit boards [2], optical waveguides [3], electrooptic modulators [4], micromechanical [5] and humidity sensor [6] devices, positive temperature coefficient of resistivity thermistors [7], gas sensors [8] were fabricated using BT. In all those applications $BaTiO_3$ was used in the form of either bulk material or thick layer. BT shows ferroelectric and piezoelectric properties as well as a high dielectric constant that make it a promising material for potential applications in dynamic access random memories (DRAM) [9], [10] or non-volatile memories (NVM) [9], [11].

Thin barium titanate films for microelectronic applications are usually either amorphous or polycrystalline and have significantly worse electrical properties than bulk or thick-film material. It is difficult, for example, to obtain uniform composition, the piezoelectric effect is weaker, and the values of the dielectric constant are lower (typically less then 50) [12]. On the other hand, its dielectric constant is usually still much higher than that of silicon dioxide although thin BT layers are typically plagued with higher leakage current and lower dielectric strength.

2. Experimental Details

The fabrication process of metal-insulator-semiconductor field effect transistor (MISFET) structures is presented

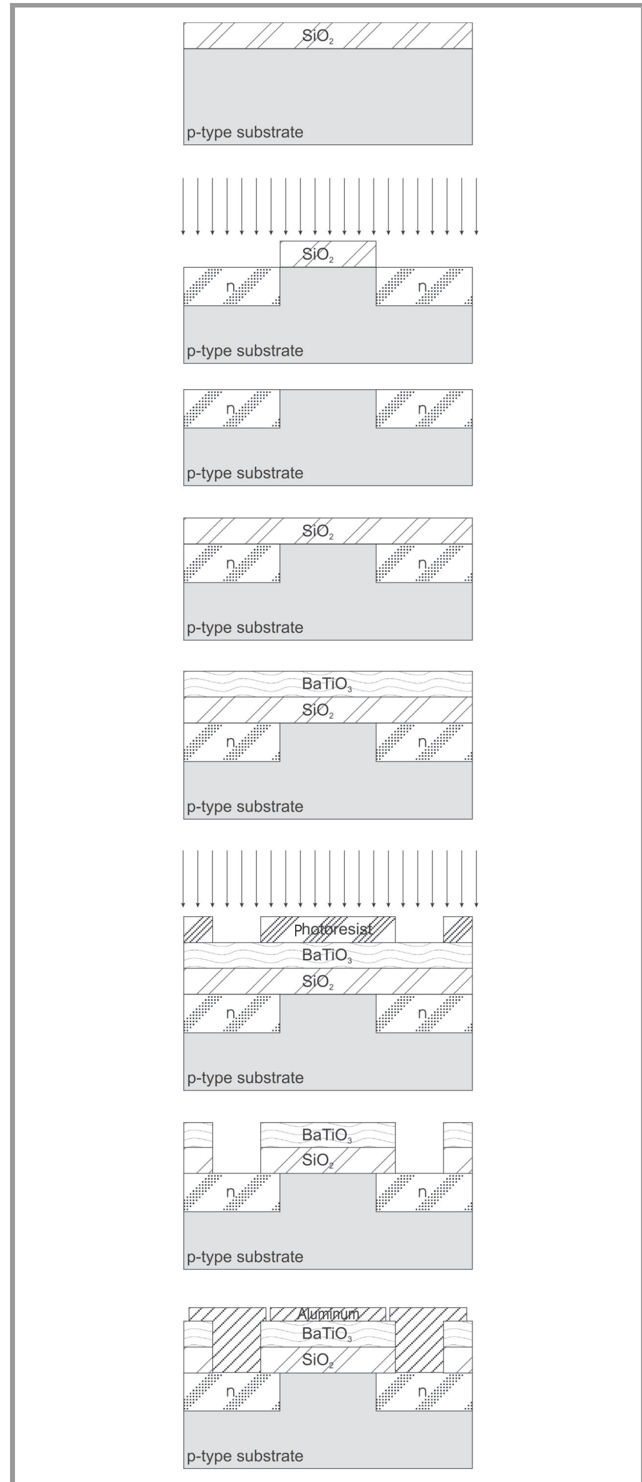


Fig. 1. MISFET fabrication process with cross-sectional view of the structures.

in Fig. 1. Its first step is thermal oxidation in order to obtain field oxide of about 440 nm. The p-type silicon $< 100 >$ oriented substrate with the resistivity of $6 - 8\ \Omega$cm was used. After cleaning processes 40 nm thick SiO_2 film was grown thermally and then a thin (approximately 80 nm) barium titanite film was deposited by means of radio frequency plasma sputtering (RF PS) of sintered $BaTiO_3 + La_2O_3$ (2% wt.) target.

A schematic diagram of the RF PS setup is shown in Fig. 2. The $BaTiO_3$ layer was deposited as a result of 30 min long process (280 V self-bias voltage, argon flow rate of 10 ml/min and 15 mm distance between the Si substrate and the sputtered target). Next, a photoresist mask for etching in a buffer solution of hydrofluoric acid was prepared by means of photolithography. As a last step contacts for metallization were opened and aluminum was evaporated. The described fabrication process is presented in Fig. 2. Silicon wafer and transistor topography are shown in Fig. 3.

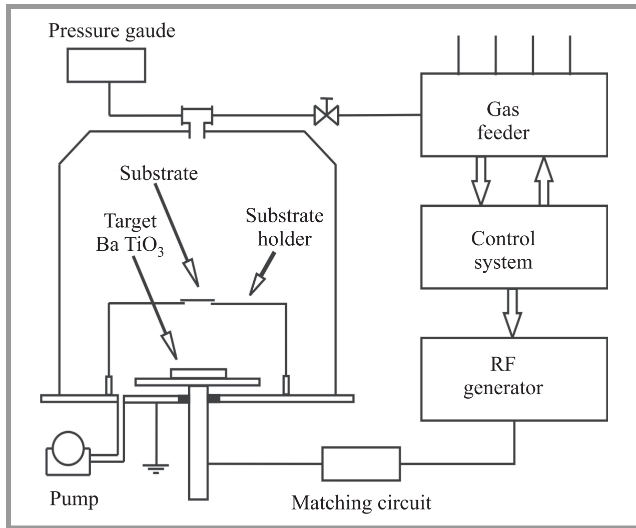


Fig. 2. Schematic diagram of the setup for radio frequency plasma sputtering deposition processes.

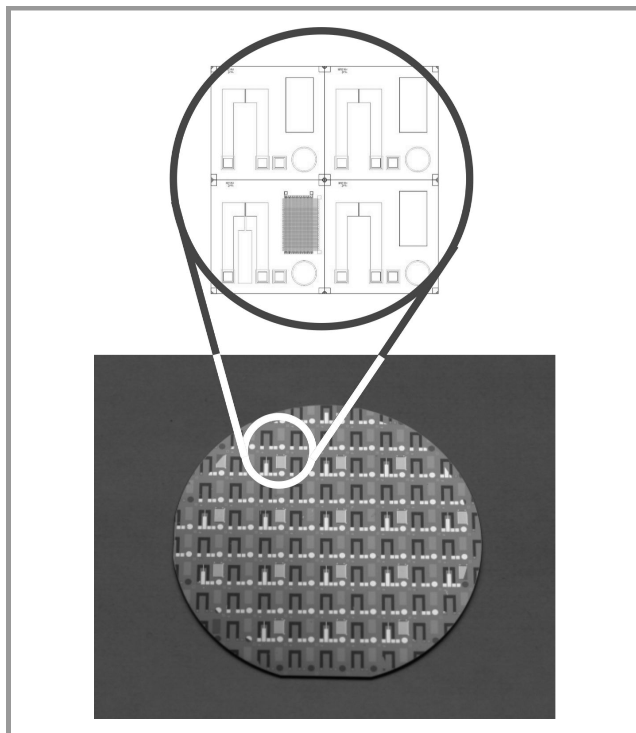


Fig. 3. Silicon wafer and MISFET topography.

3. Results and Discusions

The dielectric constant (k) of about 20 was extracted from capacitance-voltage measurements of a MIS structure

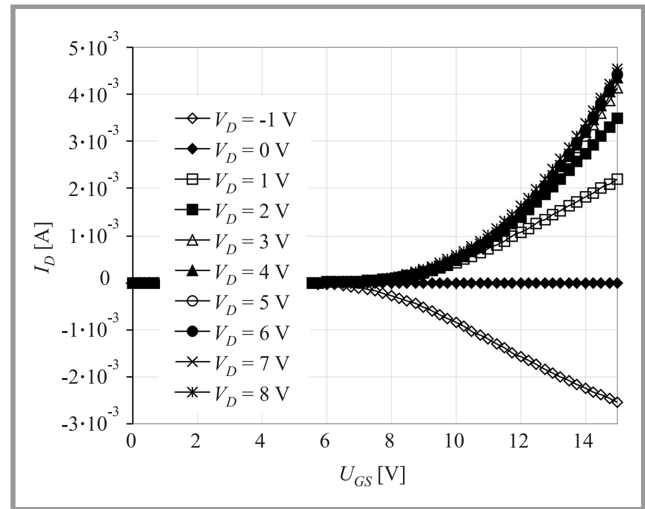


Fig. 4. Transfer current-voltage characteristics of the fabricated structures.

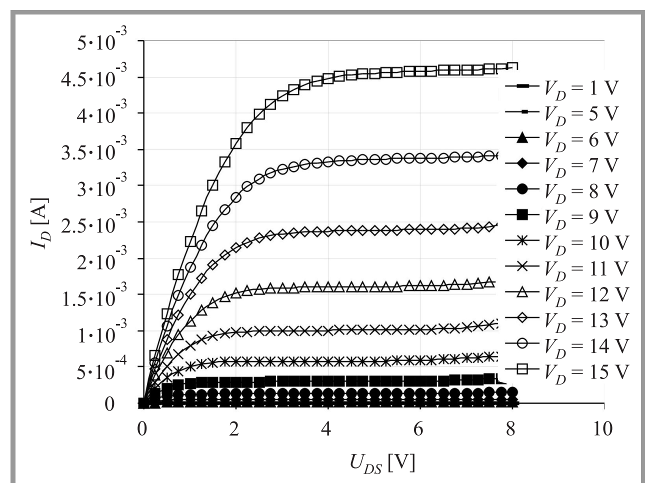


Fig. 5. Output current-voltage characteristics of the fabricated structures.

containing $BaTiO_3$ dielectric. The current-voltage ($I - V$) characteristics of MISFETs were measured with Keithley SMU 236/237/238. The obtained transfer and output characteristics are presented in Figs. 4 and 5.

Threshold voltage (U_T) is one of the most important parameters of a transistor since it represents the gate voltage at which the MISFET channel is turned on. The threshold voltage ranged from -6 V to -8 V.

The transconductance g_m and output conductance of the structures are presented in Figs. 6 and 7, respectively.

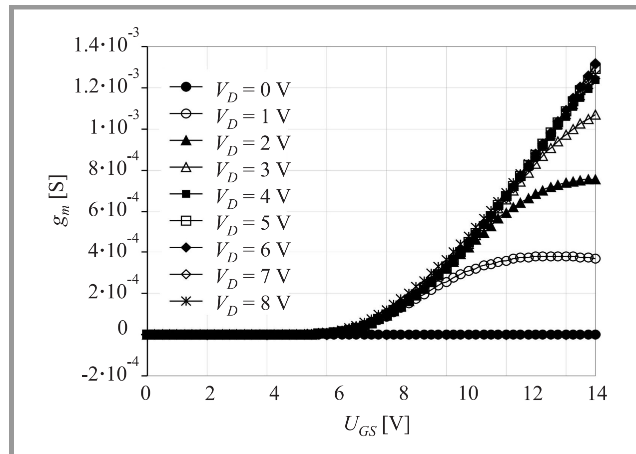


Fig. 6. Transconductance characteristics of the fabricated structures.

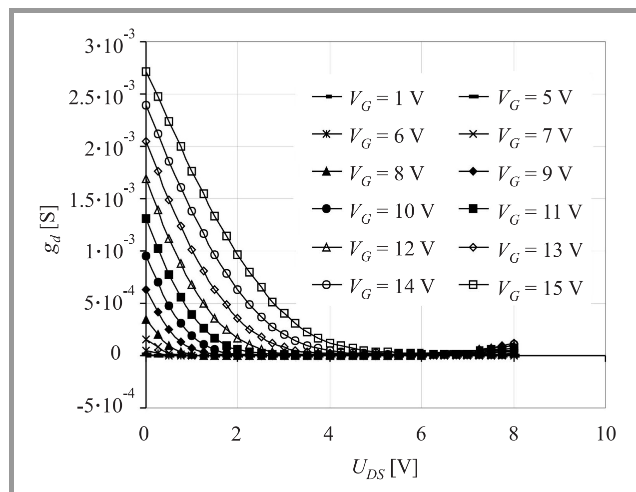


Fig. 7. Output conductance characteristics of the fabricated structures.

It can be seen that the values of transconductance are relatively low when compared to a typical silicon MOSFET. Taking into consideration that these structures contain material ($BaTiO_3$) that is relatively novel for this type of applications and demanding from the technological point of view, the obtained results in our opinion are very satisfying. Postprocessing (e.g., annealing) or better purity of the films should improve the results significantly.

The dispersion of the obtained parameters (e.g., threshold voltage, trans- and output conductance) may be caused mainly by the variations of BT thickness and its composition in the area under the gate. The structure of the layer, i.e., the grain size or the presence of amorphous phase, the quality of the interface between the BT layer and SiO_2 and the influence of plasma during fabrication may lead to different levels of the effective charge in the dielectric and at the aforementioned interface. As a consequence, the transistors show different values of flat-band voltage and threshold voltage.

4. Conclusions

The obtained BT films show good adhesion to SiO_2 layers on silicon substrate. Their relatively low dielectric constant (k) (for BT) is due to their amorphous nature. High values of the threshold voltage are a consequence of charge presence at the $SiO_2/BaTiO_3$ interface. A better control of the deposition process (e.g., purity) may significantly improve the film properties. Our investigations confirm that the RF PS method is suitable for obtaining BT layers that may exhibit several very interesting electronic properties, especially for future IS (ion sensitive) FET structures.

References

[1] J. Zhen, Z. Yue, G. Yousong, W. Sen, L. Lingfeng, X. Zhigang, and W. Yanbin, "Non-reducible $BaTiO_3$-based dielectric ceramics for Ni-MLCC synthesized by soft chemical method", *Ceram. Int.*, vol. 32, no. 4, pp. 447–450, 2006.

[2] A. Rae, M. Chu, and V. Ganine, "Barium titanate-past, present and future", *Ceram. Trans.*, vol. 100, pp. 1–12, 1999.

[3] D.-G. Sun, Z. Liu, Y. Huang, S.-T. Ho, D. J. Towner, and B. W. Wessels, "Performance simulation for ferroelectric thin-film based waveguide electro-optic modulators", *Opt. Commun.*, vol. 255, no. 4–6, pp. 319–330, 2005.

[4] P. Tang, D. J. Towner, T. Hamano, A. L. Meier, and B. W. Wessels, "Electrooptic modulation up to 40 GHz in a barium titanate thin film waveguide modulator", *Opt. Expr.*, vol. 12, no. 24, pp. 5962–5967, 2004.

[5] D. L. Polla and L. F. Francis, "Ferroelectric thin films in microelectromechanical systems applications", *MRS Bull.*, vol. 21, no. 7, pp. 59–65, 1996.

[6] V. M. Fuenzalida, M. E. Pilleux, and I. Eisele, "Adsorbed water on hydrothermal $BaTiO_3$ films: work function measurements", *Vacuum*, vol. 55, no. 1, pp. 81–83, 1999.

[7] L. Affleck and C. Leach, "Microstructures of $BaTiO_3$ based PTC thermistors with Ca, Sr and Pb additions", *J. Eur. Ceram. Soc.*, vol. 25, no. 12, pp. 3017–3020, 2005.

[8] M. Kumar, S. Rani, M. C. Bhatnagar, and S. C. Roy, "Structure, ferroelectric and gas sensing properties of sol-gel derived (Ba, Sr)(Ti, Zr)O_3 thin films", *Mater. Chem. Phys.*, vol. 107, no. 2–3, pp. 399–403, 2008.

[9] J. F. Scott, "Device physics of ferroelectric thin-film memories", *Jap. J. Appl. Phys.*, vol. 38, no. 4B, pp. 2272–2274, 1999.

[10] R. Ramesh, S. Aggarwal, and O. Auciello, "Science and technology of ferroelectric films and heterostructures for non-volatile ferroelectric memories", *Mater. Sci. Eng.*, vol. 32, no. 6, pp. 191–236, 2001.

[11] T. Kuroiwa *et al.*, "Dielectric properties of $(Ba_xSr_{1-x})TiO_3$ thin films prepared by RF sputtering for dynamic random access memory application", *Jap. J. Appl. Phys.*, vol. 33, no. 9B, pp. 5187–5191, 1994.

[12] R. Thomas, D. C. Dube, M. N. Kamalasanan, and N. Deepak Kumar, "Electrical properties of sol-gel processed amorphous $BaTiO_3$ thin films", *J. Sol-Gel Sci. Technol.*, vol. 16, no. 1–2, pp. 101–107, 1999.

Piotr Firek was born in Rawa Mazowiecka, Poland, in 1977. He received the M.Sc. degree in microelectronics from the Faculty of Electronics and Information Technology, Warsaw University of Technology (WUT), Poland, in 2004, where he is currently finishing a Ph.D. thesis. His research concentrates on fabrication, characterization, processing and application of thin and thick film materials (e.g., barium titanate, boron nitride, DLC, diamond) in microelectronic devices.

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

Jan Szmidt received the M.Sc. degree in electronics from the Faculty of Electronics, Warsaw University of Technology (WUT), Poland, in 1976. From the same university, he received the Ph.D. and D.Sc. degrees in 1984 and 1995, respectively. In 1999, he became an Associate Professor. In 2002 he became an Associate Dean for Development of the Faculty of Electronics and Information Technology, Warsaw University of Technology. Since 2006 he has been the Head of Electronic Materials and Microsystem Technology Division, Institute of Microelectronics and Optoelectronics, WUT and since 2008 – the Dean of the Faculty of Electronics and Information Technology WUT. His research interests concentrate on technology and characterization of thin films for electronics, especially for microelectronics and nanoelectronics, as well as on their application in microelectronic and nanoelectronic structures.

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

Paper

Modeling, Simulation and Calibration of Silicon Wet Etching

Andrzej Kociubiński, Mariusz Duk, Tomasz Bieniek, and Paweł Janus

Abstract— The methods of parameter optimization in Etch3D^{TM} simulator and the results of the comparison of simulations of silicon etching in KOH with experiments are presented. The aim of this study was to calibrate the tool to a set of process conditions that is offered by Institute of Electron Technology (ITE). The Taguchi approach was used to analyze the influence of every remove probability function (RPF) parameter on one or more output parameters. This allowed tuning the results of simulation to the results of real etching performed in ITE.

Keywords— anisotropic wet etching, KOH, silicon technology.

1. Introduction

Anisotropic wet chemical etching of single crystalline silicon (Si) in KOH/TMAH is the standard process technology to fabricate three-dimensional structures for microelectromechanical devices. However, the etch process is depends on the crystal orientation, etching temperature, etchant concentration, and the length of time the wafer remains in the etchant [1]. The final structure determined by Si anisotropic etching is difficult to predict precisely. Etch3D^{TM} simulator addresses this challenge by wet etching simulation [2]. Performing simulations with Etch3D^{TM} prior to going to the fab helps users reduce the time-consuming and expensive process of iteratively refining the masks and processing parameters.

The Etch3D^{TM} is a silicon wet etching simulator based on a first-principles, atomistic simulation method [3]. The wafer is represented by a matrix of "atoms" arranged with the same geometry and connectivity as the 18-atom cells in crystalline silicon. The simulator uses a "voxel" (volumetric pixel) based process emulation tool that takes 2D masks and a description of the fabrication process, to build highly detailed, realistic-looking virtual prototypes. The cell size that is used in an Etch3D^{TM} simulation is determined from the resolution, in voxels per micron. Scale invariance means that the simulations can be run with the cell arbitrarily larger size than the size of the actual silicon crystal cells (which is 0.543 nm) and produce the same macroscopic behavior, at least up to the size required to resolve macroscopic features of the wafer surface.

After initialization, which usually involves applying a mask to one or both sides of the wafer, the simulation goes through a series of time steps (also called frames). During each frame, the simulator computes the removal probability for each atom on the wafer surface that is exposed to the etchant (i.e., not masked). The simulator uses a Monte Carlo approach – it compares the removal probability to a random number between 0 and 1. Using this approach, it is possible to duplicate certain microscopic topographies that occur in actual anisotropic etching [2].

However, the primary purpose of Etch3D^{TM} is to model the macroscopic evolution of the wafer surface geometry. The atomistic method, as currently implemented, does not explicitly account for etchant concentration. Rather, it is assumed that the etch rates of all crystal planes vary linearly within a range of concentration. Etch3D^{TM} includes several sets of recommended parameter values for specific concentration ranges of KOH and TMAH. Users can further tune these parameter values for a specific concentration level.

The remove probability function (RPF) defines the probability p of removing an atom with n_1 first neighbors (in the crystal lattice) and n_2 second neighbors by the function

$$p(n_1,n_2) = p_0(n_1)\frac{1+\mathrm{e}^{\varepsilon(-n_2^0)/k_BT}}{1+\mathrm{e}^{\varepsilon(n_2-n_2^0)/k_BT}}, \tag{1}$$

where: ε is the average energy of interaction between second neighbors, and where for $n_1 = 2,3$ the constants $p_0(n_1)$, and $n_2^0 = n_2^0(n_1)$ are parameterized as follows:

$$p_0(2) = \frac{a_{21}(1+\mathrm{e}^{-a_{22}a_{23}})}{1+\mathrm{e}^{a_{22}(c-a_{23})}}, \tag{2}$$

$$n_2^0(2) = b_{21} - b_{22}c\,, \tag{3}$$

$$p_0(3) = \frac{1.0 - a_{31}(1+\mathrm{e}^{-a_{32}a_{33}})}{1+\mathrm{e}^{a_{32}(c-a_{33})}}, \tag{4}$$

$$n_2^0(3) = b_{31} - b_{32}c\,. \tag{5}$$

These values were obtained by a curve-fitting approach by Coventor, but these parameters require further adjustment in order to match the user-measured data.

2. Fabrication and Measurement of Test Structures

A dedicated mask for test etching was designed and manufactured (Fig. 1). The mask contains several structures, designed so as to represent a variety of geometries that allow exposition of multiple crystallographic planes, enabling the analysis of crystallographic planes of etched silicon.

N-type (resistivity $2-4$ Ω-cm) single crystal $< 100 >$ silicon substrates with 4-inch diameter were used for

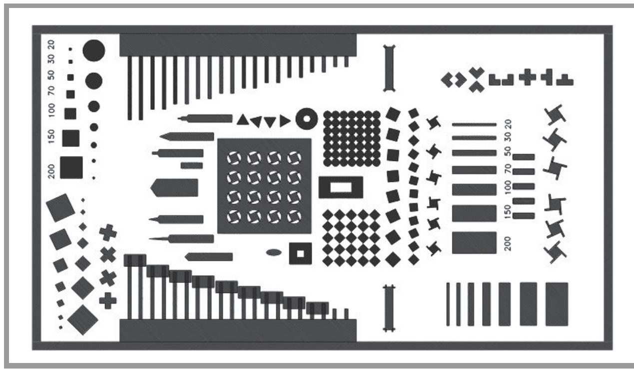


Fig. 1. Test structure for KOH etching.

anisotropic etching. Initially, a 50-nm silicon dioxide layer was grown thermally. Thereafter a 100 nm thick silicon nitride layer was deposited and photolithographically patterned. Before insertion of the wafers in etching solution, native oxide was removed. The wafers were etched in KOH (concentration 30%) at 60°C for 10, 20, 30 and 40 min. Etch depth was measured using a profilometer(typical results are given in Table 1).

Table 1
Etch depth versus time for KOH etching carried out in ITE

Time [min]	Etch depth [μm]
10	3.22
20	7.00
30	10.05
40	13.43

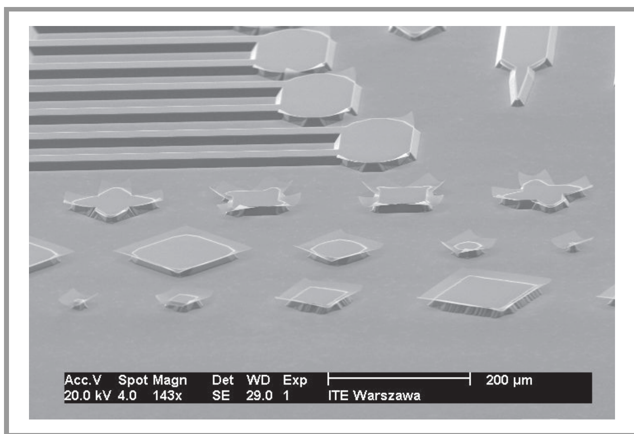


Fig. 2. SEM microphotograph of test structures after 40 min etching in KOH.

The structures were then subjected to scanning electron microscope (SEM) observations. A SEM micrograph of test structures after 40 min etching in KOH is shown in Fig. 2. The etch depth as well as characterized shapes of etched structures were used as an input to Etch3D^{TM}.

3. Simulation and Calibration of Etch3D^{TM}

Due to the high number of test structures, simulating the whole mask would take a very long time. In order to reduce the overall simulation time, only a subset of the structures was selected for initial simulations. The simulation results were then compared in detail with the experimentally fabricated structures, making use of SEM techniques.

The Etch3D^{TM} simulation software gives the user access to 10 so-called RPF parameters. It turned out that tuning all parameters in order to find the optimal values, for different etching times, concentrations and temperatures and using different test structures would result in a huge amount of data to be processed.

As a first step, in order to reduce the amount of data, a set of four crosses, rotated at different angles, was selected for tuning the RPF parameters (Fig. 3). This structure offers several interesting etching angles and is also quite simple, allowing time-effective simulations. This decision allowed the number of voxels, and thus the resolution of simulation, to be increased, which results in an improved quality of the simulation.

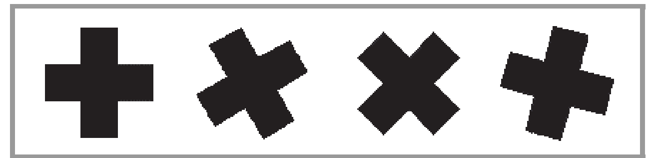


Fig. 3. Layout of the selected 4-cross test structure.

We have selected separate sets of input and output parameters for the experiment in order to allow comparison not only in quality but also quantity:

– input parameters: 10 RPF parameters, concentration and temperature of bath;

– output parameters: 5 selected geometry details revealing selected crystallographic planes (Fig. 4).

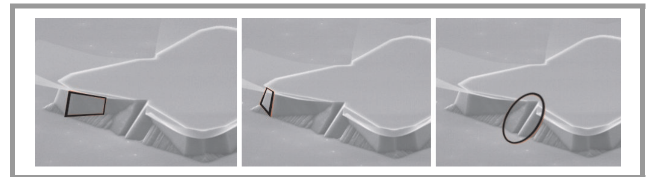


Fig. 4. Example of geometry details defined as output parameters.

The first set of simulations resulted in an enormous amount of data to analyze mainly due to the fact that:

– the 10 RPF input parameters had to be modified for each process setup;

– each input parameter affected the results in a different way;

– one input parameter influenced more than one output parameters.

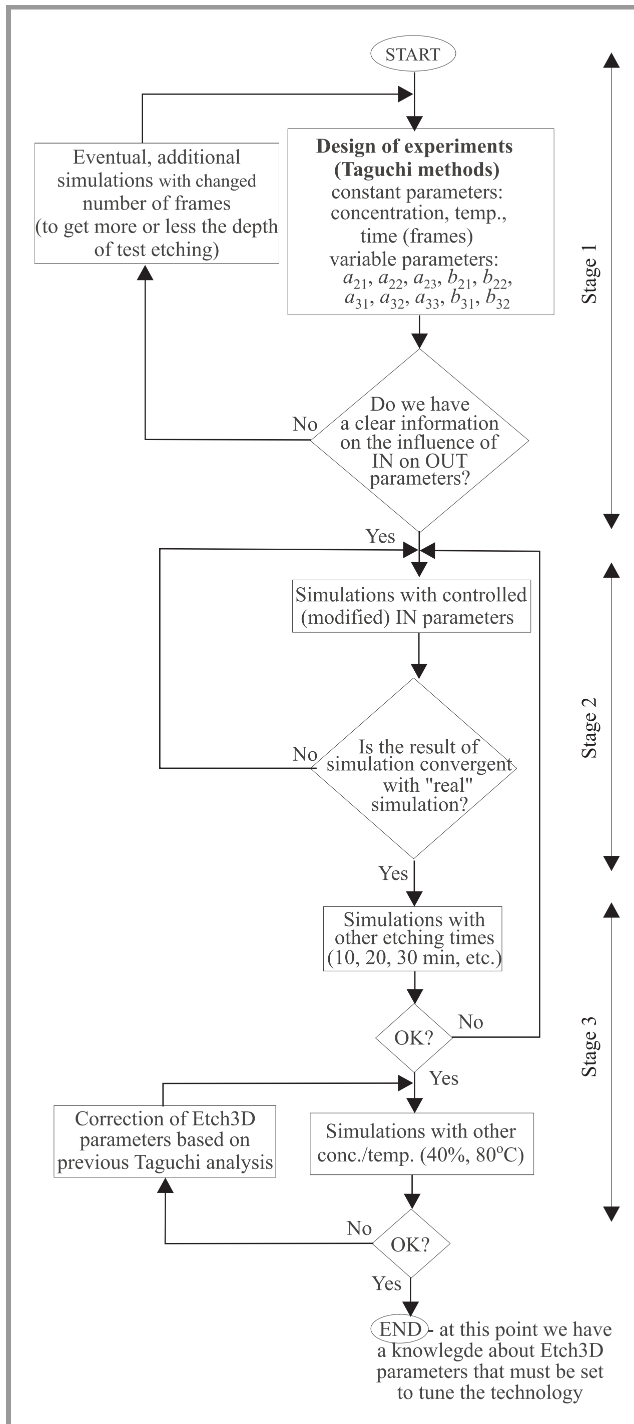


Fig. 5. Algorithm of tuning Etch3D^{TM} RPF parameters to existing KOH/TMAH technology.

Due to a high number of possible parameter combinations, a reliable statistical method of planning further simulations was necessary. The Taguchi approach was proposed to design a sequence of simulations and analyze results. The Taguchi method [4], [5] is an approach to designing experiments using statistical analysis. It allows for the process and product design to be improved through the identification of controllable factors and their settings, minimizing the variation of a product around a defined target response. The Taguchi approach in Etch3D^{TM} was used to analyze the influence of every RPF parameter on one or more output parameters. This allowed tuning the results of the simulation to those of real etching performed in ITE.

To solve this calibration problem, a simulation optimization procedure was developed (Fig. 5). The whole tuning sequence was divided into three stages.

In stage 1 the Taguchi table was created and a set of simulations was performed. The purpose of this stage is to get the information on the influence of input parameters on the results of simulation (Fig. 6).

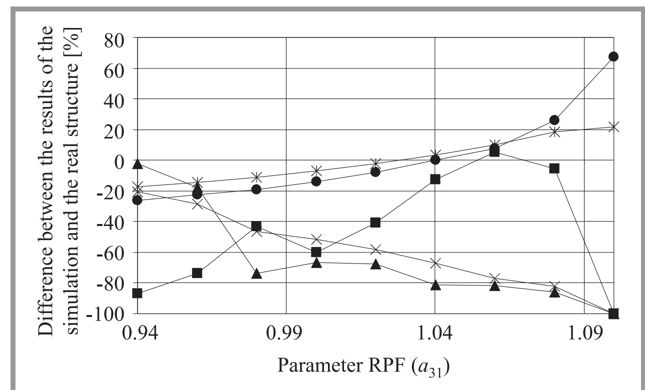


Fig. 6. Graphs presenting dependence of various geometrical details on the RPF parameter (a_{31}).

In stage 2, based on the information from stage 1, a new set of simulations was designed. The purpose of this stage was to perform more detailed simulations with modified input parameters (however, concentration, temperature and number of frames remained constant).

At this stage the results of the simulations are tuned to the results of etched test structures. The results of the simulations of test structures with tuned parameters of Etch3D^{TM} are presented in Figs. 7–11.

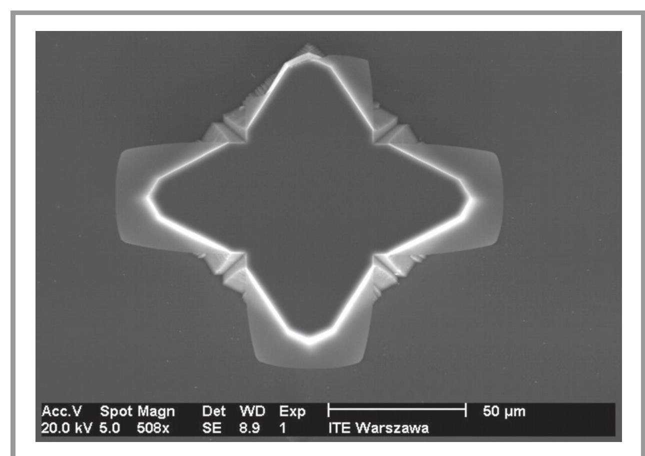


Fig. 7. SEM microphotograph of the real structure (top view).

Comparing the results obtained with default and tuned parameters good agreement with the crystallographic planes

of the etched test structures may be noticed (higher-order crystallographic planes marked in Fig. 4). Moreover, the difference in the depths of the simulated and etched groove is below 5%.

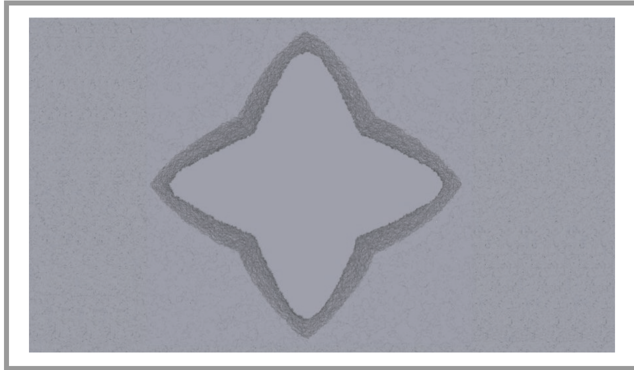


Fig. 8. Results of the simulations using the default a_{ij}, b_{ij} coefficients.

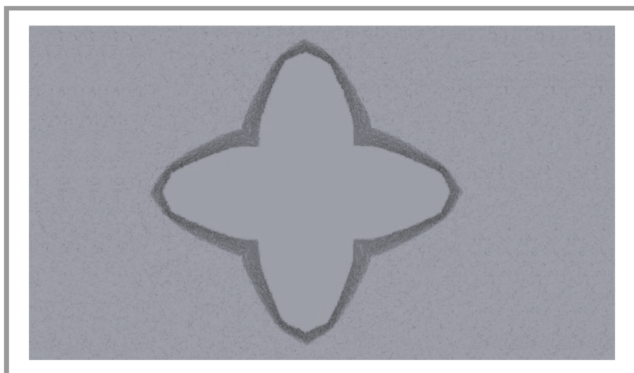


Fig. 9. Results of the simulations – using tuned Etch3D™ RPF parameters.

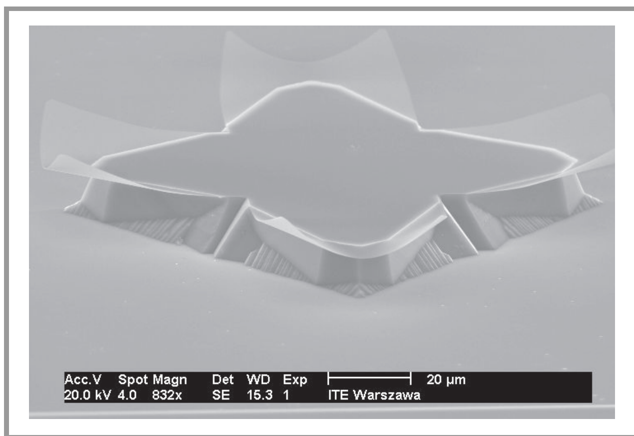


Fig. 10. Main view of the real structure.

It was found that some of the parameters have significant influence on the output parameters (angles, etch depth, etc.). For example, changing the b_{31} parameter results in the greatest increase of the etch depth than any other RPF parameter.

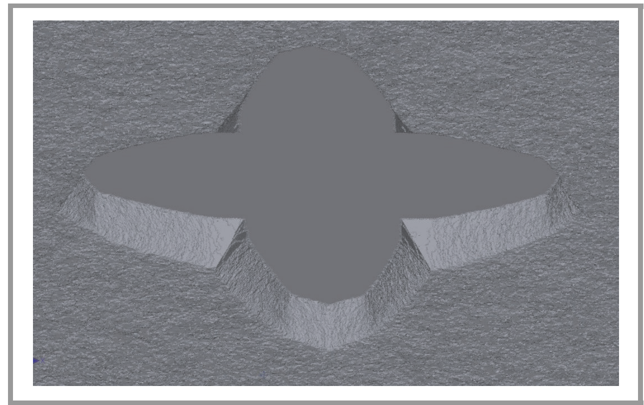


Fig. 11. Main view of the simulated structure with tuned Etch3D™ RPF parameters.

Finally, tuning of only 5 out of 10 parameters was necessary for tuning. In the case of the remaining 5 parameters their default values could be treated as sufficiently accurate.

Table 2
Remove probability function parameters

RPF parameter	Default	Tuned
a_{21}	0.5	0.35
a_{22}	25.0	25.0
a_{23}	0.7	0.7
b_{21}	7.0	7.0
b_{22}	0	1.2
a_{31}	1.0	0.94
a_{32}	7.5	7.5
a_{33}	0.5	0.5
b_{31}	7.3	7.7
b_{32}	0.6	0.1

The values of RPF parameters used for tuned simulations are presented in Table 2.

4. Process Emulation of MEMS and IC structures

The Etch3D™ is intended to be used in conjunction with SEMulator3D™-MEMulator™ [2]. The results of virtual manufacturing of silicon accelerometer using both programs [6] are shown in Fig. 12.

MEMulator™ is useful for preprocessing or postprocessing a wafer, and for visualizing the results of an Etch3D™ simulation. Preprocessing with MEMulator™ may entail creating a wafer, then using one or more of the available etch steps in MEMulator™ to create some initial features on the wafer, or using one or more of the available deposition steps to create etch stops. The material that is deposited on the wafer with MEMulator™ may be etched

in Etch3D^{TM}, but this would be unusual because the actual process steps (other than wafer bonding) do not produce crystalline silicon. Postprocessing with MEMulator^{TM}, on

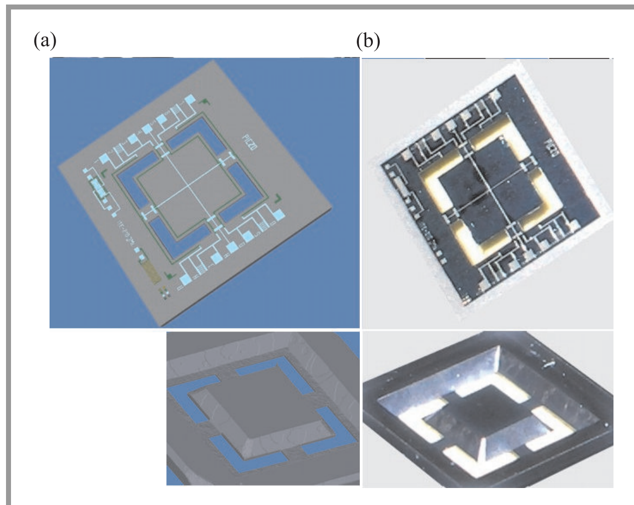


Fig. 12. Emulation of fabrication steps of piezoresistive accelerometer. Results of virtual manufacturing (a) and real structure (b).

the other hand, may entail an arbitrary series of deposition and etch steps on top of the wet-etched surface that was produced by Etch3D^{TM}.

5. Conclusions

The presented modeling, simulation and real etching results demonstrate successful calibration of wet silicon etching processes. Using the Taguchi method it is possible to tune the etch process to the fab reality and a given technology specification.

Tuned software like Etch3D^{TM} is a perfect tool to optimize the wet etching of silicon processes and use it for virtual, rapid prototyping. The precise 3D model generated using described approach may be further analyzed using finite element method.

Virtual prototyping is a powerful tool that may be used to reduce the time consuming and expensive process of refining masks and many processing parameters. This is of great importance since TTM (time to market) is a key parameter of the MEMS product development process.

References

[1] H. Seidel, L. Csepregi, A. Heuberger, and H. Baumgartel, "Anisotropic etching of crystalline silicon in alkaline solutions", *J. Electrochem. Soc.*, vol. 137, no. 11, pp. 3612–3625, 1990.

[2] "Etch3D User Guide Version 2006.5", Sept. 2006 [Online]. Available: http://www.coventor.com

[3] M. A. Gosalvez, R. M. Nieminen, P. Kilpinen, E. Haimi, and V. Lindroos, "Atomistic wet chemical etching of crystalline silicon: atomistic Monte-Carlo simulations and experiments", *Appl. Surf. Sci.*, vol. 178, pp. 7–26, 2001.

[4] P. J. Ross, *Taguchi Techniques for Quality Engineering*. New York: McGraw-Hill, 1988.

[5] G. Z. Yin and D. W. Jillie, "Orthogonal design for process optimization and its application in plasma-etching", *Solid-State Technol.*, vol. 30, no. 5, pp. 127–132, 1987.

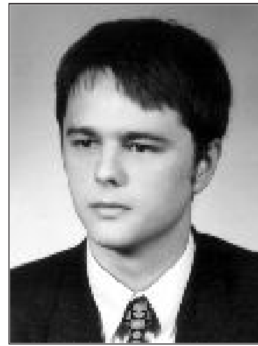
[6] P. Janus, T. Bieniek, A Kociubiński, P. Grabiec, and G. Schrópfer, "Modeling and co-simulation of integrated micro- and nanosystems", in *14th Int. Conf. MIXDES 2007*, Ciechocinek, Poland, 2007.

Andrzej Kociubiński received the M.Sc. and Ph.D. degrees in electronic engineering from the Warsaw University of Technology, Poland, in 2002 and 2007, respectively. In 2001 he joined the Department of Silicon Microsystems and Nanostructure Technology of the Institute of Electron Technology, Warsaw. Since 2007, he has been with the Lublin University of Technology. His research interests include diagnostics, characterization, simulation and modeling of silicon devices. His recent works are related to SOI devices, pixel detectors on SOI wafers and integrated microsystems.
e-mail: akociub@semiconductor.pl
Lublin University of Technology
Nadbystrzycka st 38a
20-618 Lublin, Poland

Mariusz Duk received the M.Sc. and Ph.D. degrees in electrical engineering from the Lublin University of Technology, Poland, in 1999 and 2008, respectively. The doctor thesis was devoted to the application of wavelet transform to the compression of signals obtained from the optical flame monitoring system. Since the graduation he has been working in the Department of Electronics. In the course of his work he created dozens of electronic and optoelectronics designs. His most important designs are certain versions of optical flame monitoring systems, and laboratory fuel injection system for engines.
e-mail: m.duk@pollub.pl
Lublin University of Technology
Nadbystrzycka st 38a
20-618 Lublin, Poland

Tomasz Bieniek received the M.Sc. and Ph.D. degrees in electronic engineering from the Warsaw University of Technology, Poland, in 2002 and 2007, respectively. The doctor thesis was devoted to plasma technologies for ultrathin dielectric layer formation. In 2006 he joined the Department of Silicon Microsystems and Nanostructure Technology of the Institute of Electron Technology, Warsaw. His research is focused on MEMS and complementary metal oxide semiconductor silicon technologies and multidomain modeling, as well as the simulation of micro- and nanostructures. He is an author and co-author of more than 40 technical papers and presentations presented in journals and at conferences.

e-mail: tbieniek@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland



Paweł Janus received the B.Sc. and M.Sc. degrees from the Technical University in Wrocław, Poland, in 1996 and 1998, respectively. In 2003 he received the Ph.D. degree in electrical engineering from the Wrocław University of Technology. Since 2003, he has been with the Institute of Electron Technology, Warsaw. His research interests include studies of materials for MEMS applications, micromachining of silicon microstructures, microsensors, and microactuators. His current research is silicon force sensors development for nanomechanical applications. He is an author and co-author of more than 30 technical papers presented in journals and at conferences.

e-mail: janus@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Paper

Si-Based Electrodes for Potentiometric Measurements of Aqueous Solutions

Michał Zaborowski, Daniel Tomaszewski, Bohdan Jaroszewicz, and Piotr Grabiec

Abstract— Three sensors for chemical and physical examination of aqueous solutions were presented in the paper. An Au potentiometric electrode, an AgCl chlorine ion sensor and a p-n junction thermometer were developed. Their layout and internal structure were explained in the light of the manufacturing process. The device characteristics were measured in conditions corresponding to normal operation of the devices. Basic electrical parameters of the developed structures, as well as their sensitivity to environmental parameter variation were estimated.

Keywords— chlorine ion sensor, potentiometric sensor, p-n junction thermometer.

1. Introduction

Potentiometric sensors allow for easy measurements of ion concentration in aqueous solutions. They work in a range of sensitivity that is important in nature sciences [1]. The sensors have been developed on the basis of gold electrodes and covered with membranes sensitive to dedicated ions. Measurements are carried out in a circuit containing a reference electrode placed in the solution together with the sensor. The resulting electromotive force (EMF) is directly dependent on the ion concentration. Changes of ambient temperature usually influence the sensor output signal level and its sensitivity. In the presented work, an Au potentiometric electrode, a Cl^- ion sensor and a thermometer chip have been developed.

2. Device Fabrication

Gold electrodes were designed as basic devices suitable for potentiometric measurements. They were manufactured using n-type $< 100 >$ oriented silicon wafers of low resistivity ranging from 0.01 to 0.001 Ωcm. A simple planar technology based on the complementary metal oxide semiconductor (CMOS) process was developed. First, a 100 nm thick thermal silicon dioxide layer was grown and a 65 nm thick silicon nitride layer was deposited using low-pressure chemical vapor deposition (LPCVD). Next, square 1×1 mm contact holes were photolithographically defined and opened. The Si_3N_4 layer was removed by plasma etching, whereas the SiO_2 layer was removed by wet etching. Afterwards, a Cr/Au double layer was sputtered to cover the top and bottom wafer surfaces. The thickness of the chromium and gold layers was 10 nm and 500 nm, respectively. The electrodes were patterned at the top surface by means of photolithography followed by wet etching of Au and Cr layers. Post-metallization annealing at 340°C in nitrogen was performed in order to minimize the series resistance of the electrodes. A cross-section and simple topology of the gold electrode is shown in Fig. 1. Finally, the wafers were diced into separate square 4.9×4.9 mm chips.

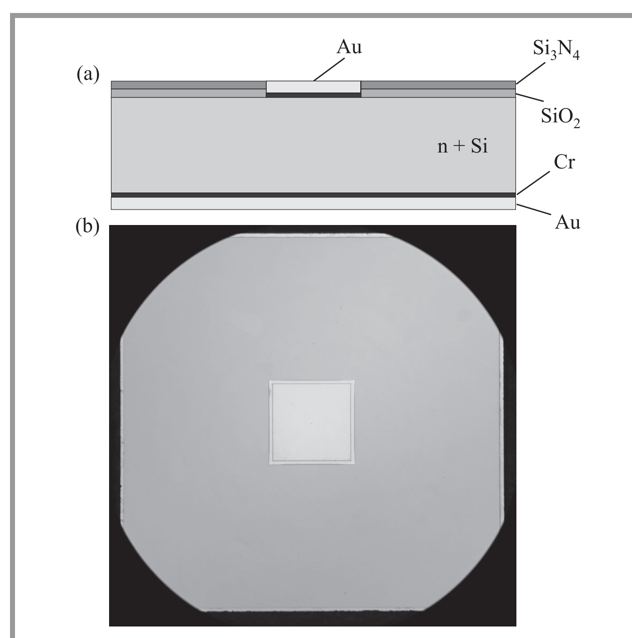


Fig. 1. A schematic cross-section (a) and top view of gold potentiometric electrode (b).

The gold electrodes were designed for operation in aqueous solutions. They could be covered with a variety of ion-sensitive membranes by sensor developers and users. Our efforts were concentrated on a small and stable value of series resistance of the device. An example of the on-wafer distribution of the measured Au electrode resistance has been shown in Fig. 2. The mean value of the resistance was about 30 mΩ. Also the deviation of the distribution from its mean value was reasonably small (about 5 mΩ).

Another type of the device for potentiometric measurements was the chlorine potentiometric sensor, designed based on the above-described gold electrodes. After the process of annealing in nitrogen ambient the silicon wafers

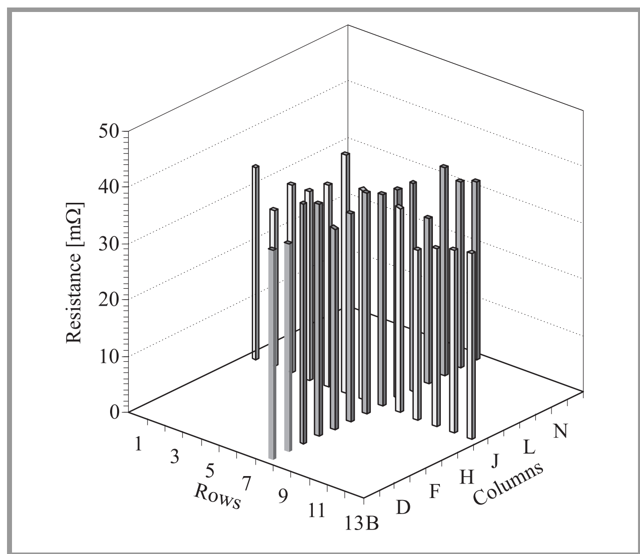


Fig. 2. On-wafer distribution of Au potentiometric electrode resistance.

with Au electrodes were subjected to a short etching step in a buffer HF solution. Next, two electrochemical processes were performed. First, backsides of the wafers were masked with a resist layer while the patterns at the front sides were Ag-electroplated. The process was optimized towards fine morphology of Ag grains (Fig. 3). Next, the 5 μm thick silver layer was partially electrochlorinated.

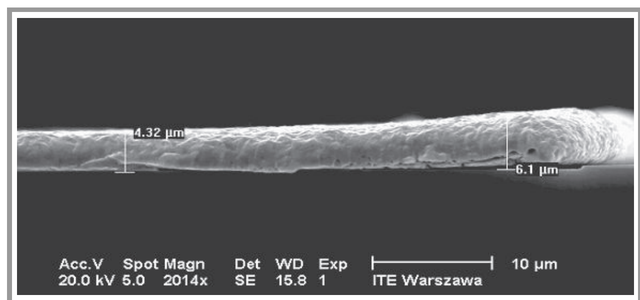


Fig. 3. Side view of the chlorine potentiometric sensor. Fine Ag layer morphology is visible.

Finally, the resist layer covering the backsides was removed in acetone bath, and the wafers were diced into chips. The topology of the top side of the device together with its cross-section has been presented in Fig. 4.

As mentioned above the potentiometric measurements required parallel observations of the temperature, which might influence measurements of ions. For this reason a p-n junction-based thermometer was designed and fabricated. It is known that thermal properties of the forward-biased p-n junction depend on the supplied current and on the doping distributions. Thus, in the first step a series of numerical calculations using the ATHENA/SSUPREM4 were done. Their purpose was to optimize the sensitivity (which in theory is of the order of –(2 – 3) mV/deg) and linearity. The results of these calculations are shown in Fig. 5.

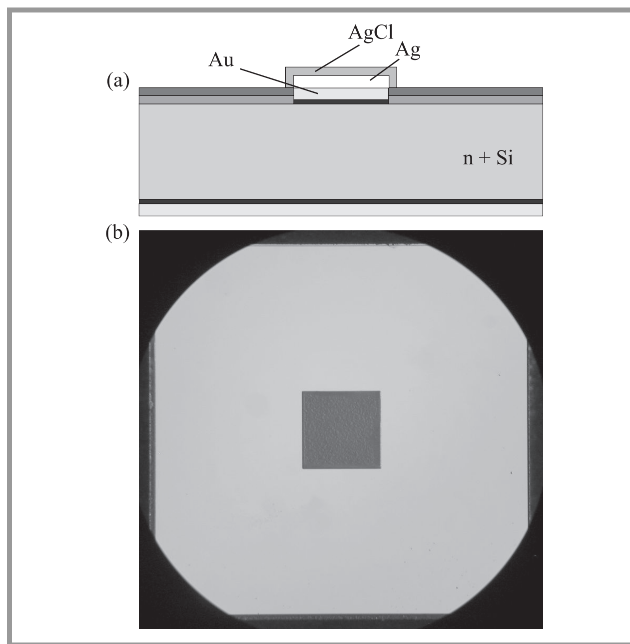


Fig. 4. A cross-section (a) and top view of the chlorine potentiometric sensor (b).

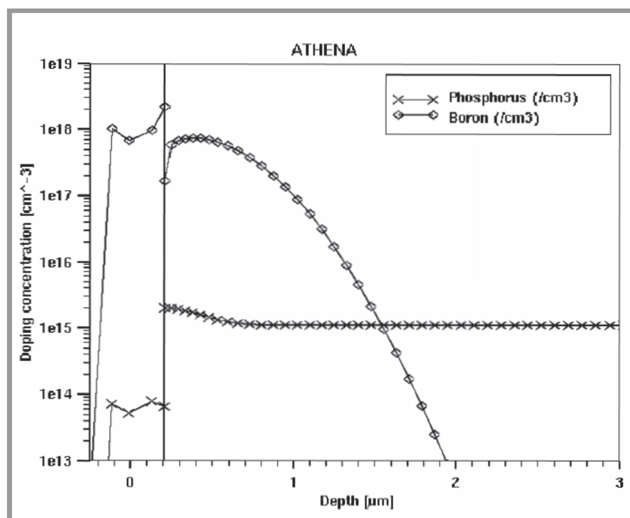


Fig. 5. Doping distribution in the p-n junction based thermometer calculated numerically using the ATHENA/SSUPREM4 solver.

The thermometer design was based on the following requirements: low electrical resistance between the sensor and the contact areas, and low thermal resistance between the sensor and a tested fluid. Therefore, the thermometer structures were fabricated in silicon membranes (Fig. 6). The chips were manufactured using $3-5\ \Omega$cm $< 100 >$ oriented n-type silicon wafers. In the first step the silicon was thinned to a 100 μm membrane at the centre of the device by means of an anisotropic etch. A Si_3N_4 layer served as the mask against the etching. Next, the anode region was defined at the bottom side of the device. A $8 \cdot 10^{13}$ cm^{-2} dose of boron ions was implanted into the anode area with 130 keV energy. Diffusion at 1000°C was used subsequently to form the p-n junction.

Next, phosphorus implantation (80 keV, $5 \cdot 10^{15}$ cm^{-2}) was employed to create the subcontact layer to the cathode of the sensor.

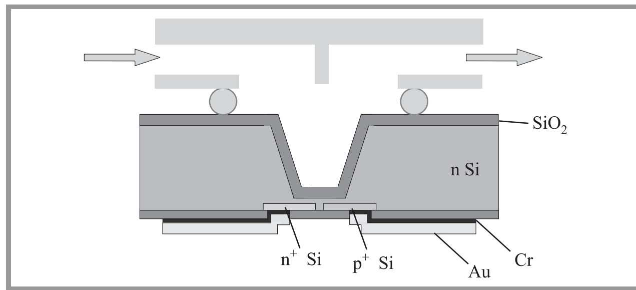


Fig. 6. Cross-sections of the p-n junction based thermometer working in a dedicated flow package.

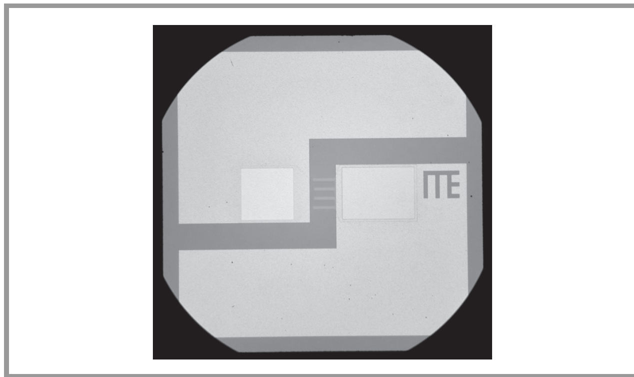


Fig. 7. View of the thermometer based on p-n junction.

Finally, Al contacts of the sensors were formed and the wafers were diced into chips. The thermometer chip layout is shown in Fig. 7.

3. Electric Measurements

The devices described in the previous section had a common topology (size, placement of contacts). Therefore, they were very well suited for operation in identical packages. Such a package was designed for measurements of ions under fluid flow conditions. Five sensor cells together with the reference electrode cell are presented in Fig. 8.

The AgCl potentiometric electrodes were destined for operation as Cl^- ion sensors in aqueous solutions. Four water solutions of KCl with different Cl^- ion concentrations were used in the experiments. The sensors worked in the flow packages, which were joined in series together with the reference electrode (as shown in Fig. 8). Obviously the sensors were uniformly distributed along the fluid flow. The output signals, i.e., the electromotive force (*EMF*) between the electrodes and the reference electrode were registered by a control application prepared in the LabView environment. The results are shown in Fig. 9.

The AgCl sensors exhibited log-lin reversible response to chlorine ion concentration. The sensitivity of the sensors in the higher range of Cl^- concentration was found to be

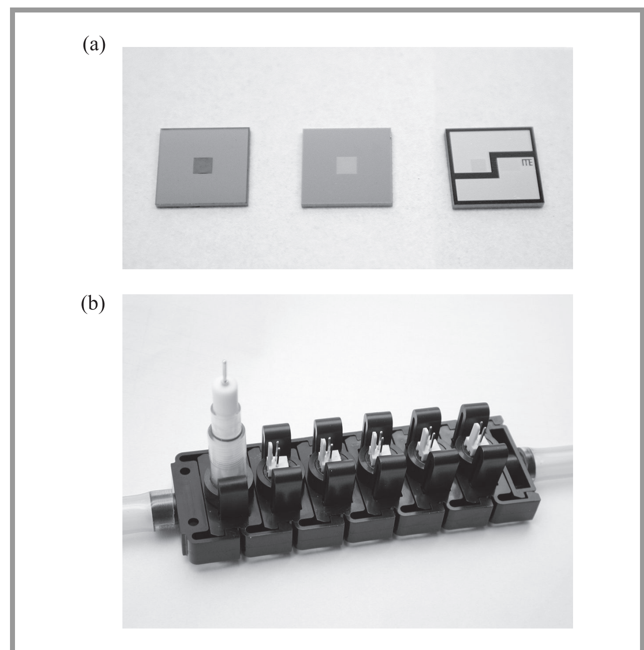


Fig. 8. Sensor devices (a) and a package for potentiometric measurements under fluid flow conditions (b).

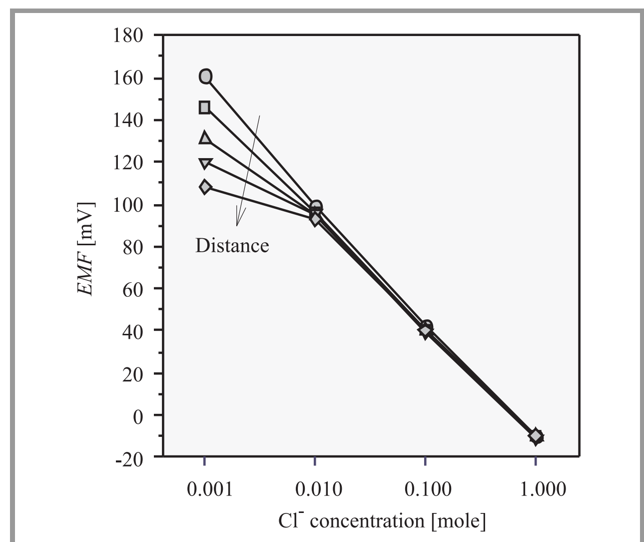


Fig. 9. Output signal versus Cl^- ion concentration for five AgCl electrodes assembled in the package for potentiometric measurements under fluid flow conditions.

close to the theoretical value of 57 mV/dec. However, in the case of sensors more distant from the reference electrode and for very diluted solutions a discrepancy from the log-lin behavior might be noticed. It may be explained by a simple model shown in Fig. 10. Here the variable meaning is as follows: V is an ideal voltmeter, R_V is its internal resistance, n denotes ion concentration, E is the measured electromotive force, R_{sol} denotes the resistance of the solution between the reference and the AgCl electrodes, X is the distance between them, and α, β, n_0 denote constants. According to this simple model the measured voltage starts to differ from the electromotive force if measured in a very

diluted solution or if the sensor is placed far away from the reference electrode. In both cases the solution resistance becomes non-negligible as compared to the internal resistance of the voltmeter. A giga-ohm class voltmeter is recommended for potentiometric measurements.

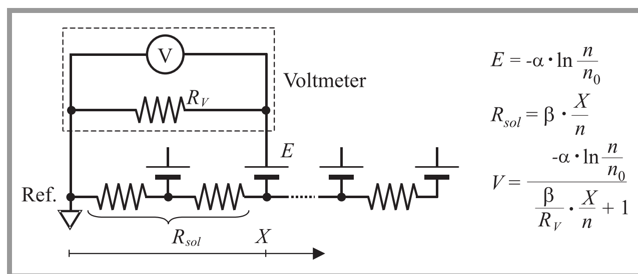


Fig. 10. A simple model of the Cl^- ion measurement setup using a number of cells.

The temperature sensors were the third set of devices to be subjected to electrical measurements. They were tested using different setups, i.e., before dicing on a hot plate, and after dicing and assembling in the package shown in Fig. 8. The temperature of the hot plate was controlled by a thermocouple within the chuck. Therefore, there was a small difference between the thermocouple and the wafer surface temperatures.

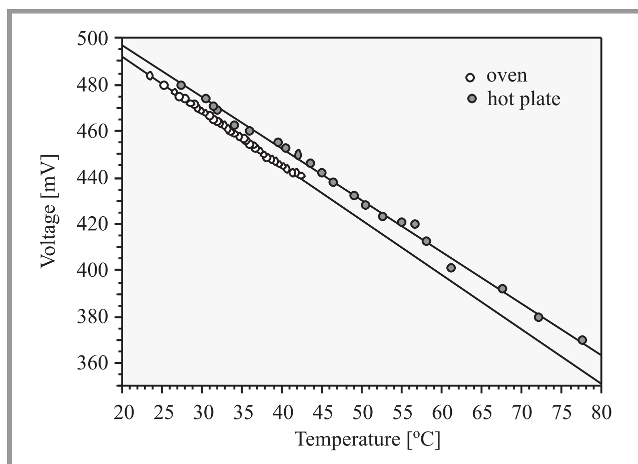


Fig. 11. Output signals versus temperature for the p-n junction-based thermometer measured on a hot plate (closed dots); in an oven (open dots).

Direct testing of the temperature sensors in the flow cell under water flow conditions was found to be difficult because of ambiguous recording of device temperature. Inlet and outlet temperatures differed by more than 3°C in experiments with 40°C water temperature. There was no possibility to install any temperature control equipment inside the package. As thermosensors might be used not only in the aqueous environment but in the air as well, they were calibrated in the packages placed in an air oven (Fig. 11). The obtained calibration characteristics $V(T)|_{I=100\,\mu\text{A}}$ revealed the satisfactory linearity with the -2.38 mV/deg slope, which was close to the theoretical value. Small heat capacity and low heat resistance between the p-n junction and device surface should be emphasized.

4. Summary

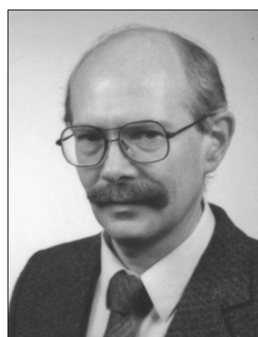
In the presented work, the development of an Au potentiometric electrode, a Cl^- ion sensor and a thermometer chip is described. The technological process sequence, measurement procedures and results are reported. The electrical parameters and their spread have been examined in conditions corresponding to normal operation of the devices. The satisfactory on-wafer spread of series resistance of the gold electrodes, the measured sensitivity of AgCl electrodes to chlorine ion concentration, and the observed response of the thermometer structures have confirmed advantages of the developed technologies. Moreover, the standardization of the sensor chip layout and packaging make them a useful tool set suitable for environmental parameter observation.

Acknowledgments

The authors are grateful to partners of Water Risk Management in Europe (WARMER) project for helpful discussions.
The work was partially supported by the Commission of the European Communities under contract no. 034472 FP6-2005-IST-5 – WARMER.

References

[1] M. Zaborowski, B. Jaroszewicz, D. Tomaszewski, P. Prokaryn, E. Malinowska, E. Grygołowicz-Pawlak, and P. Grabiec, "Fabrication of MOS – compatible ion – sensitive devices for water pollution monitoring (WARMER)", in *Proc. 14th Int. Conf. Mix. Des. Integr. Circ. Syst. 2007*, Ciechocinek, Poland, 2007, pp. 477–481.



Michał Zaborowski received the M.Sc. degree in solid-state electronics from the Warsaw University of Technology, Poland, in 1974. In 1976 he joined the Institute of Electron Technology (IET), Warsaw. He worked on IC metallization technology and received the Ph.D. degree in the IET, in 1998. His research interests include nanotechnology, micromachining and microsensor technology, particularly for biological applications. He is the author and co-author of more than 60 papers and conference presentations.

e-mail: mzab@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Piotr B. Grabiec graduated from the Warsaw University of Technology, Poland, in 1973, and received the Ph.D. degree in chemistry from the same university in 1985. In 1974 he joined the Institute of Electron Technology, Warsaw, where he was involved in CVD and diffusion technology research. Since 1999 he has been the Head of Silicon Microsystem and Nanostructure Department. His present activity involves fabrication of silicon ASICs, optoelectronic devices and MEMS and their integration. He has been involved in 10 EU projects and has been awarded numerous awards for development and commercialization of advanced micro-devices. He is the member of IEEE and Electrochemical Society. He is the author and co-author of more than 300 scientific papers and conference presentations. He holds 19 patents.

e-mail: grabiec@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Bohdan Jaroszewicz received the M.Sc. degree in electronics from the Warsaw University of Technology, Poland, in 1968. He joined UNITRA-CEMI (Semiconductor Device Factory) in 1968 and was involved in a variety of research, engineering and fabrication activities related to MOS integrated circuits. As the engineering section manager he was responsible for optimization of MOS IC yield. He joined the Institute of Electron Technology in Warsaw, in 1997. His main research interests are ion implantation engineering, as well as design, fabrication and characterization of ISFETs and ionising radiation detectors.

e-mail: bjarosz@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Daniel Tomaszewski – for biography, see this issue, p. 49.

Paper

LPT and SLPT Measurement Methods of Flat-Band Voltage (V_{FB}) in MOS Devices

Krzysztof Piskorski and Henryk M. Przewłocki

Abstract— The photoelectric techniques are often used for the measurements of metal oxide semiconductor (MOS) structure parameters. These methods, which consist in illuminating the MOS structure with a semitransparent metal gate by a UV light beam, are often competitive for typical electric measurements. The results obtained by different photoelectric methods are, in many cases, more accurate and reproducible than the results of other measurements. The flat-band voltage V_{FB} is an important parameter of any MOS structure since its value influences the threshold voltage V_T, which decides for example about power consumption of MOS transistors. One of the methods to measure the V_{FB} value is the electric method of $C(V)$ characteristic. This method involves certain calculations and requires the knowledge about parameters of the investigated sample. The accuracy of this method is rarely better than ±100 mV (for higher doping of the substrates the accuracy is worse). The other method of V_{FB} value determination, outlined in this article, is the photoelectric light pulse technique (LPT) method. This method based on the idea proposed by Yun is currently being optimized and verified experimentally.

Keywords*— *flat-band voltage, light pulse technique, MOS system, photoelectric methods, scanned light pulse technique.

1. Introduction

The light pulse technique (LPT) and the scanning light pulse technique (SLPT) are photoelectric methods used to determine electrical parameters of metal oxide semiconductor (MOS) structures. In this article, determination of the flat-band voltage V_{FB} of MOS structures using these two methods will be discussed. The LPT method may be used to determine the V_{FB} value of the entire MOS device, while the SLPT method allows determination of the distribution of local V_{FB} values over the gate area. In the latter case this is done by scanning the gate area with a light beam of small diameter (small in comparison with gate dimensions).

At present, the best results of LPT and SLPT measurements are experimentally obtained making use of a digital lock-in amplifier. The DC signal u at the output of the lock-in amplifier is a function of the potential V_G applied to the gate of the MOS device under investigation. This is illustrated in Fig. 1(b), where the dependence of the signal u on the gate voltage V_G is shown. In Fig. 1(a) the capacitance-voltage $C(V_G)$ curve of the same structure is shown for comparison.

As will be shown later, the gate voltage V_G at which u changes sign is the flat-band voltage of the MOS structure under investigation. Hence, by adjusting gate voltage V_G to obtain $u = 0$ the V_{FB} value can be determined. The main difficulty in practical application of the LPT and SLPT measurement methods results from the fact that the signal $|u|$ in the accumulation range is in most of the cases several orders of magnitude smaller than in inversion and depletion (as schematically shown in Fig. 1). Hence, very high sensitivity of the measurement setup is required to correctly determine the V_G value corresponding to $u = 0$.

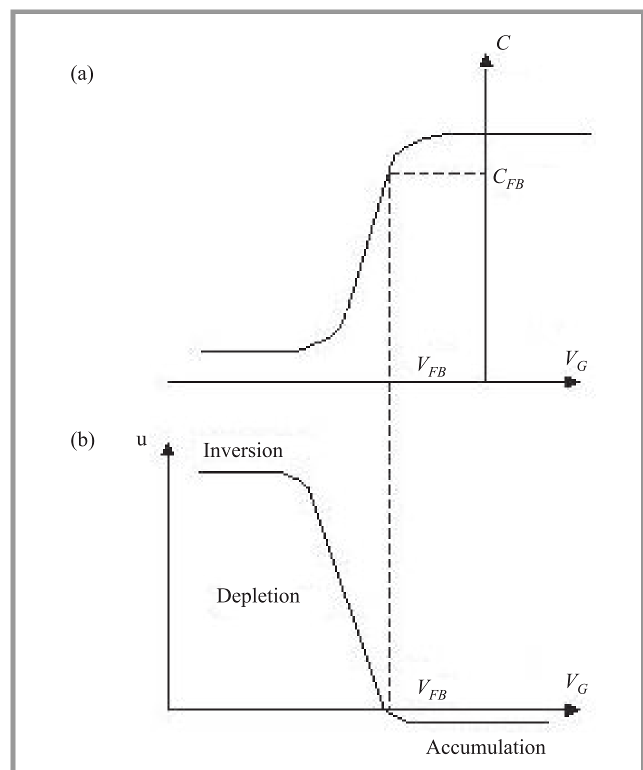


Fig. 1. (a) The $C(V_G)$ curve of a MOS device and (b) the $u = f(V_G)$ characteristic of the same device. The V_G value at which u changes sign is the flat-band voltage V_{FB}.

The idea of using the LPT method to determine the flat-band voltage V_{FB} was first proposed by Yun [1]. Using the oscilloscope, he observed the dependence of the current pulses resulting from light pulses illuminating the MOS device under investigation on the gate voltage V_G. He correctly concluded, that by changing the gate bias V_G from

inversion and depletion towards accumulation of the MOS structure, one observes diminishing current pulses that disappear at $V_G = V_{FB}$. Incorrectly, however, he stated that at V_G values corresponding to accumulation, the MOS structure does not respond with any electrical signal to the excitation by the light pulses.

A more rigorous and comprehensive treatment of the problems related with the LPT method of V_{FB} determination was given by Jakubowski and Krawczyk [2], [3]. These authors proved, both theoretically and experimentally, that the electrical signal with which the MOS structure responds to the light pulses, changes sign at $V_G = V_{FB}$ and that there exists a finite and measurable electrical response of the MOS structure in the state of accumulation. The polarity of this response is opposite to the polarity observed in the inversion and depletion ranges, as shown in Fig. 1.

The authors of [3] used the lock-in amplifier in their experimental work, which allowed them to obtain $u(V_G)$ characteristics similar to the one shown in Fig. 1. However, they did not try to take full advantage of this technique as a V_{FB} determination method.

It is the purpose of this work to improve the understanding and the experimental implementation of this measurement technique, to make it more sensitive, more precise and more accurate than the commonly used method of $C(V_G)$ characteristics.

2. Physical Background of the Method

Consider an MOS structure illuminated by a series of light pulses as shown in Fig. 2(a). The photon energy of this light hv should be larger that the band gap E_G of the semiconductor substrate (to generate electron-hole pairs), but smaller than the barrier heights at gate-dielectric and dielectric-semiconductor interfaces (not to generate current flow across the dielectric).

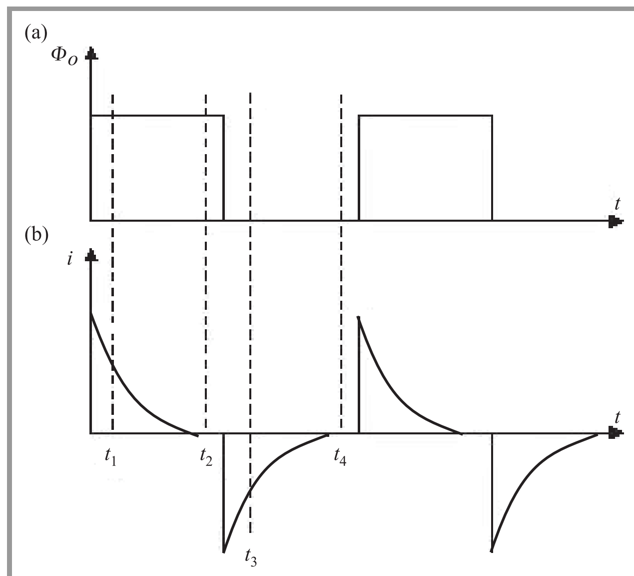


Fig. 2. (a) The intensity of photon flux Φ_0 versus time t and (b) the current i versus time t.

As a result of such pulsed illumination, a series of current pulses can be detected in the external circuit of the investigated structure (Fig. 2(b)). The magnitude of these pulses depends on the semiconductor surface potential ϕ_S and the pulses disappear when $\phi_S = 0$, as shown in [3].

Hence, by finding the dependence of the magnitude of these current peaks on the gate bias V_G one may find the flat-band voltage value V_{FB}, at which the current peaks disappear.

Following [4] in Fig. 3 the movement of charges is illustrated in different time intervals $t = t_1$, t_2, t_3 and t_4, marked both in Figs. 2 and 3.

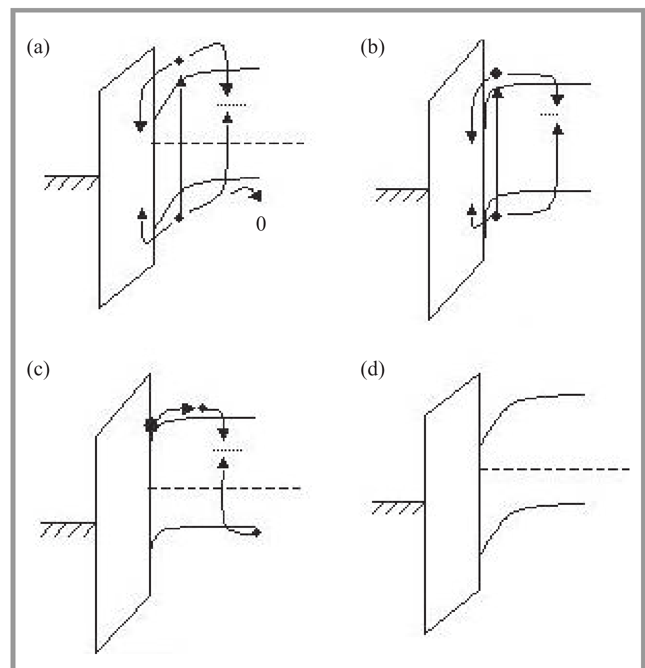


Fig. 3. Band diagrams of the MOS system for different times illustrating the current flux at the dielectric-semiconductor interface: (a) $t = t_1$; (b) $t = t_2$; (c) $t = t_3$; (d) $t = t_4$.

At $t = t_1$, the light beam causes increased generation of carriers in the semiconductor. The photo-generated electrons may either recombine through the interface states, or in the semiconductor bulk, or they may accumulate in the potential well at the semiconductor-dielectric interface. At the same time, the photo-generated holes can either recombine with the electrons, at the interface or in the bulk, or they can leave the semiconductor through the back contact, causing current flow in the external circuit and accumulate in the MOS gate. After some time, electrons and holes are accumulated on both sides of the dielectric, thus, increasing the voltage drop in the dielectric. Such an increase of this voltage drop must be balanced by a decrease of the semiconductor surface potential ϕ_S, which causes a decrease of the width w of the space charge region at the semiconductor surface. As a result, the carrier generation rate decreases, leading to the situation in which the carrier generation is

balanced by recombination and the current in the external circuit disappears, as illustrated in Figs. 2 and 3(b) for $t = t_2$.

When illumination disappears, electrons accumulated at the semiconductor surface are emitted into the semiconductor bulk where they recombine with the holes "returning" from the gate through the external circuit, as illustrated in Fig. 3(c). These "returning" holes create a negative current pulse in the external circuit, as shown in Fig. 2, at $t = t_3$. This process continues until thermal equilibrium is reached at the semiconductor surface, as illustrated in Fig. 3(d) and the current in the external circuit disappears, as shown in Fig. 2, at $t = t_4$.

3. Calculation of the $u = f(V_G)$ Characteristics

In an attempt to fully exploit the advantages of LPT and SLPT methods of flat-band voltage determination in MOS structures we have developed a method which allows the dependence of the DC signal u at the output of the lock-in amplifier to be calculated as a function of the potential V_G applied to the gate of the MOS structure under investigation. This method is based on the theory developed in [2], [3].

The effective light generation level is defined as

$$\xi = \frac{\Delta n}{n_i} = \frac{\Delta p}{n_i}, \tag{1}$$

where: Δn, Δp are the excess electron and hole concentrations generated by light and n_i is the intrinsic concentration. As shown in [2], [3] the main parameters of MOS structures under illumination (in quasi-equilibrium) may be expressed by the same expressions that apply in equilibrium if the Fermi potential $u_F = \frac{q\phi_F}{kT}$ is replaced by $u_F^* = \frac{q\phi_F^*}{kT}$, and the intrinsic concentration n_i is replaced by n_i^*, where:

$$u_F^* = \frac{1}{2}\ln\frac{\xi + e^{u_F}}{\xi + e^{-u_F}} \tag{2}$$

and

$$n_i^* = n_i\sqrt{\left(\xi + e^{u_F}\right)\left(\xi + e^{-u_F}\right)}. \tag{3}$$

The DC signal at the output of lock-in amplifier is proportional to the difference ΔQ_S of the semiconductor surface charge in equilibrium Q_S and under illumination Q_S^*:

$$u \sim \Delta Q_S = Q_S - Q_S^*, \tag{4}$$

where Q_S is given by [5]

$$Q_S = \sqrt{2\,kT\,\varepsilon_{Si}\,\varepsilon_0\,n_i}\cdot F_S, \tag{5}$$

which under illumination becomes:

$$Q_S^* = \sqrt{2\,kT\,\varepsilon_{Si}\,\varepsilon_0\,n_i^*}\cdot F_S^*, \tag{5a}$$

where: k – Boltzmann's constant, T – absolute temperature, ε_{Si} – relative electrical permittivity of the semiconductor, ε_0 – permittivity of free space and F_S is the Kingston function, given by

$$F_S = -\frac{u_S}{|u_S|}\sqrt{e^{u_F}\left(e^{-u_S} + u_S - 1\right) + e^{-u_F}\left(e^{u_S} - u_S - 1\right)}, \tag{6}$$

which under illumination becomes:

$$F_S^* = -\frac{u_S^*}{|u_S^*|}\sqrt{e^{u_F^*}\left(e^{-u_S^*} + u_S^* - 1\right) + e^{-u_F^*}\left(e^{u_S^*} - u_S^* - 1\right)}, \tag{6a}$$

where: u_s is the normalized surface potential in equilibrium:

$$u_S = \frac{q\phi_S}{kT}, \tag{7}$$

becoming:

$$u_S^* = \frac{q\phi_S^*}{kT} \tag{7a}$$

under illumination.

The gate voltage V_G is the same when the MOS structure is in equilibrium and when it is illuminated. In the case of equilibrium, V_G may be expressed as [5]

$$V_G - V_{FB} = \phi_S - \frac{Q_S}{C_I}, \tag{8}$$

which under illumination becomes:

$$V_G - V_{FB} = \phi_S^* - \frac{Q_S^*}{C_I}, \tag{8a}$$

where: C_I is the capacitance of the dielectric.

Since $Q_S(Q_S^*)$ are the functions of $u_S(u_S^*)$ and $\phi_S(\phi_S^*)$ given above one may calculate ϕ_S, ϕ_S^*, Q_S and Q_S^* values for any V_G value, if the effective light generation level ξ is known. Hence, for any ξ value the $\Delta Q_S = f(V_G)$ characteristics may be calculated. These characteristics should have the same shape as the experimental $u = f(V_G)$ characteristics.

For given experimental conditions the value of ξ may be determined by taking the $C(V_G)$ characteristics of the MOS structure in the darkness, as well as, under illumination and by calculating the u_F^* value using the expression [2], [3]:

$$u_F^* = u_F\left[\frac{C_{INV}\left(C_I - C_{INV}^*\right)}{C_{INV}^*\left(C_I - C_{INV}\right)}\right]^2, \tag{9}$$

where: C_{INV}, C_{INV}^* are the inversion capacitance values in the darkness and under illumination, respectively.

Once the value of u_F^* is known, ξ may be calculated using:

$$\xi = \mathrm{e}^{-u_F} \frac{\mathrm{e}^{2u_F} - \mathrm{e}^{2u_F^*}}{\mathrm{e}^{2u_F^*} - 1}. \quad (10)$$

The calculated $u(V_G)$ characteristics in inversion/depletion and accumulation are presented in Figs. 4(a) and 4(b), respectively. An ideal n-type MOS structure with the following parameters: $t_{OX} = 63.2$ nm, $N_D = 1.487 \cdot 10^{15}$ cm^{-3}, $C_I = 5.4637 \cdot 10^{-8}$ F/cm^2 was assumed.

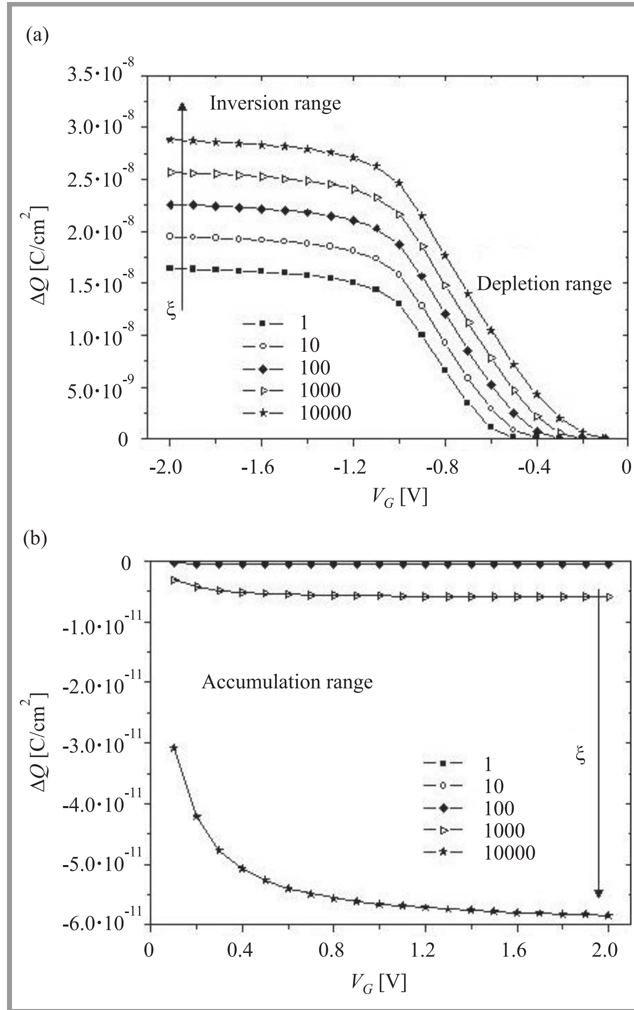


Fig. 4. Calculated $\Delta Q_S = f(V_G)$ characteristics for different ξ values in: (a) inversion/depletion and (b) accumulation.

The influence of ξ on the calculated ΔQ (hence also on the u signal) is much stronger in accumulation than in inversion/depletion, as clearly seen in Fig. 4. The ratio of the signal in accumulation to the signal in inversion ($|u_{accu}|/u_{inv}$) is shown in Fig. 5 as a function of ξ.
Hence, the ξ value may be directly determined from the $|u_{accu}|/u_{inv} = f(\xi)$ plot:

$$\xi = 10^{\frac{m}{n}} \left(\frac{|u_{accu}|}{u_{inv}} \right)^{\frac{1}{n}}, \quad (11)$$

where: m, n are the parameters of the interpolation line $y = mx + n$ approximating the $|u_{accu}|/u_{inv}$ dependence.

Hence, having the measured $u(V_G)$ characteristics and comparing the $|u_{accu}|/u_{inv}$ ratio with the calculated dependence of $|u_{accu}|/u_{inv}$ on ξ (for the same structure parameters) one may determine the effective light generation level ξ.

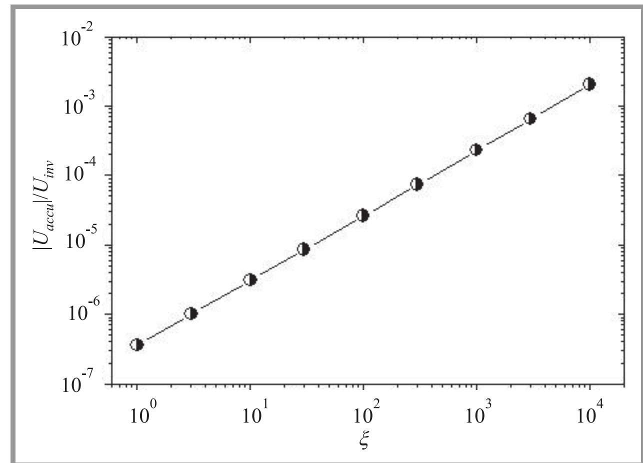


Fig. 5. The ratio of $|u_{accu}|/u_{inv}$ for different ξ. In the range of 1 to 10^4 the ratio shows the linear dependence of log $|u_{accu}|/u_{inv}$ on log ξ.

4. Implementation of the Measurement Method

The measurement setup for the flat-band voltage V_{FB} determination is shown in Fig. 6. The light beam is chopped and reflected onto the gate of MOS structure by the mirror.

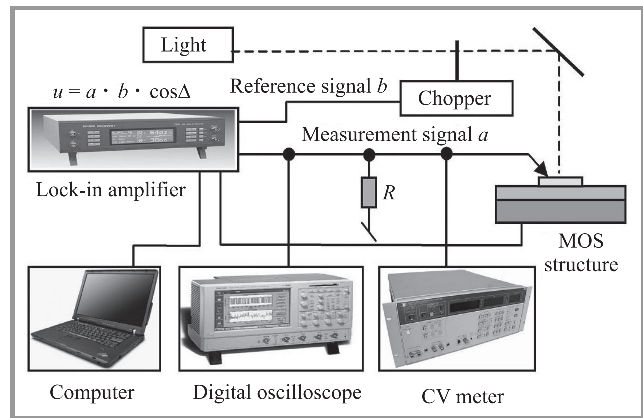


Fig. 6. The measurement setup for the flat-band V_{FB} voltage determination.

The measured signal a from the structure is detected by a digital oscilloscope and a lock-in amplifier, which allows very small signals to be measured, even if they are below the noise level. The reference signal b from the chopper is also fed into the lock-in amplifier. The output signal u of the lock-in amplifier is a product of a and b signals

and cosine of the phase difference between these signals ($\cos\Delta$). Changing the phase of signal b allows the maximum value of the output signal ($\cos(0, 180°) = 1$) to be obtained.

The comparison of the measured and calculated characteristics is shown in Fig. 7. The parameters used in the calculation model ($N_D = 1.354 \cdot 10^{15}$ cm^{-3}, $t_{OX} = 61.83$ nm) were determined by $C(V_G)$ measurements. The ξ value is assumed to be 10^4. The measured $u(V_G)$ characteristics are shifted along the horizontal axis so that the signal u reaches zero at zero gate voltage.

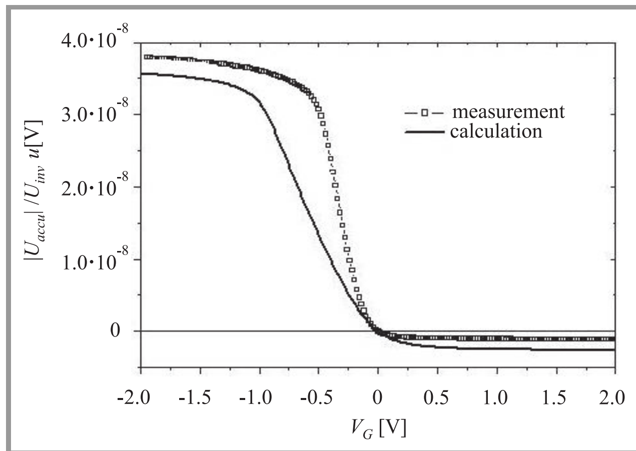


Fig. 7. The comparison between measured and calculated $u(V_G)$ characteristics.

The differences between both (measured and calculated) characteristics are clearly seen, although the shape of these curves is quite similar. The inaccuracies (e.g., different slope in depletion range, different level in accumulation and inversion ranges) are probably caused by the fact that calculations were done assuming an ideal MOS structure, by the errors in measuring the structure parameters and, possibly, by the inaccuracy of the determination of the zero level of the signal u.

5. Results

The measurements of $u = f(V_G)$ characteristics made on dozens of MOS structures with different parameters (e.g., different t_{OX}, different annealing time $t(N_2)$) confirmed, that the photoelectric LPT method is a very promising technique of the flat-band voltage V_{FB} determination. Numerous measurements made on the same structure show excellent reproducibility. The spread of points, at which $u(V_G)$ characteristics change sign (V_{FB} values), is not greater than ± 5 mV in this case. The sensitivity of the LPT method is very good. Due to the use of a lock-in amplifier high precision measurements of very small signal values ($\sim$ nV) in accumulation may be performed. Hence, the LPT method seems to allow much more accurate V_{FB} determination than the method of $C(V_G)$ characteristics.

Despite many advantages of the LPT method the determination of its absolute accuracy is still problematic. In accordance with our present knowledge and assuming for the moment, that all the measurements of $u(V_G)$ characteristics are taken at the same temperature T, the $u(V_G)$ characteristics taken at different values of the light beam power P should have one common point at which they intersect one another. This point corresponds to $\phi_S = 0$ and $V_G = V_{FB}$. Hence, this point of intersection is expected to lie at the $u = 0$ axis.

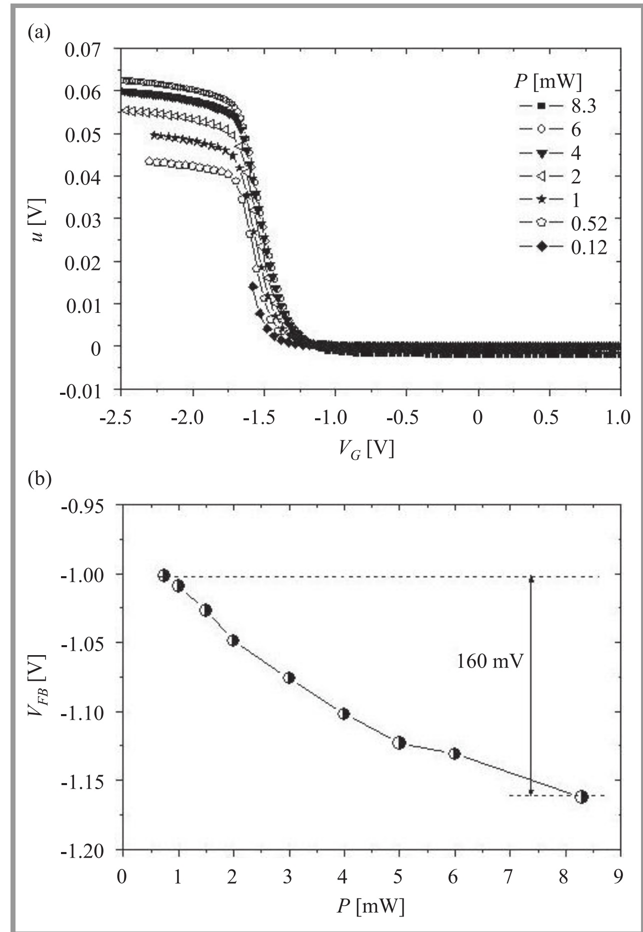


Fig. 8. (a) The $u(V_G)$ characteristics measured at different light power P and (b) V_{FB} values shifted by 0.11 mV along vertical axis as a function of light power.

Our measurement results show, however, that this "point", or rather a spot of final dimensions, where the family of $u(V_G)$ characteristics taken at different P values is very narrow (< 200 mV), lies not at the $u = 0$ axis, but a little (~ 0.1 mV) above it. This may result in considerable differences in V_G values at which these characteristics intersect the $u = 0$ axis. This effect results probably from the zeroing accuracy of our lock-in amplifier, which is insufficient for this application.

Therefore it is assumed further that the position of the minimum width of the band of $u(V_G)$ characteristics and not the point of intersection with the $u = 0$ axis determines

the V_{FB} value. The fact that the $u(V_G)$ characteristics taken over a wide range of light power P do not cross one another at one point, but rather at a spot of finite dimensions, may be explained by the differences of the temperature T of the structure, resulting from illumination with different light beam power P.

The $u(V_G)$ characteristics measured at different P are presented in Fig. 8(a). The influence of T on the V_{FB} value is illustrated in Fig. 8b, by the V_{FB} versus P characteristics. In Fig. 8(b) the amplitude (difference between highest and lowest V_{FB} values) is approximately equal to 160 mV. The V_{FB} value measured on the same MOS structure by $C(V_G)$ characteristics method is equal to -0.955 V, so it is clearly seen that in in the case of LPT photoelectric measurements the V_{FB} values obtained at different power of light are more negative than those obtained from the the $C(V_G)$ measurements. Moreover a decreasing tendency of $V_{FB} = f(P)$ for $P > 0.75$ mW is observed. This phenomenon may be explained by the increasing temperature caused by the high power of the light beam illuminating the sample [6], [7]. Further investigations will be focused on detailed understanding of this problem.

Although, the absolute accuracy of the LPT method is not satisfactory yet, the SLPT method which is a modification of the LPT method was used to measure local V_{FB} values. By scanning the gate area with a light beam of small diameter the spatial distribution of the local V_{FB} values may be determined. Using this method a series of local V_{FB} values was determined along the diagonal of a square gate, as shown in Fig. 9.

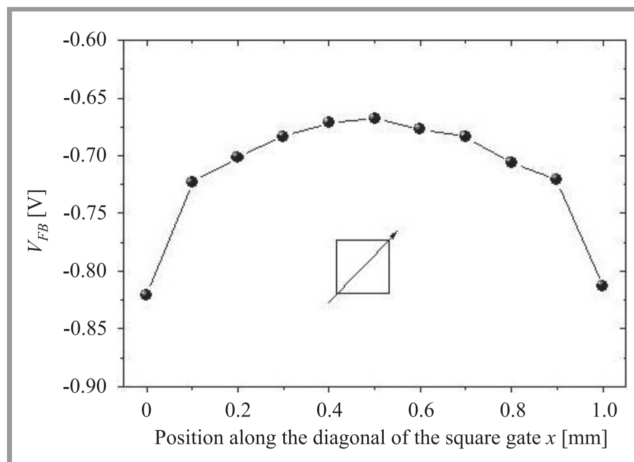


Fig. 9. One-dimensional distributions of local V_{FB} values in a MOS structure.

Although high absolute accuracy of each local V_{FB} value, represented by a point in Fig. 9, may not be guaranteed, there is no doubt that the shape of the V_{FB} distribution over the gate area is correctly represented. This shape of V_{FB} distribution is not surprising taking into account the previously determined distributions over the gate area of the effective contact potential difference ϕ_{MS} [8]–[11] and of the barrier height E_{BG} at the gate-dielectric interface [12]–[15]. Local values of all these parameters have "dome-like" distributions over gate areas of metal gate MOS structures. It is our hypothesis that this shape of distributions of electrical parameters is caused by the non uniform distribution of the mechanical stress at the metal-dielectric interface [16]–[20].

6. Conclusions

A new, high precision photoelectric measurement method of the flat-band V_{FB} voltage in MOS structures is studied. This method, called light pulse technique (LPT) consists in illuminating a semitransparent gate of a MOS structure by a series of light pulses and measuring the output current signal which is a function of the potential V_G applied to the investigated structure. The magnitude of these current pulses depends on semiconductor surface potential ϕ_S and when $\phi_S = 0$ the pulses disappear. This situation defines the flat-band state in semiconductor.

The measurement results of V_{FB} values confirmed that the LPT method is characterized by a good precision and good reproducibility. The problem of the absolute accuracy of this method has not been solved yet and it is going to be the main purpose of our further investigations.

The SLPT method which is a modification of the LPT method allows to the local values of V_{FB} at different points over the gate of the MOS structure to be measured. It was proved that, as expected, the V_{FB} values have a characteristic dome-like distribution over the gate area. This shape is similar to those of ϕ_{MS} and E_{BG} distributions measured previously and is characterized by the highest values in the middle of the gate and lower values at the gate corners. It is our hypothesis that the mechanical stress existing in the oxide under the metal gate has a dominant influence on the shape of the distribution of the above mentioned electrical parameters.

References

[1] B. H. Yun, "Direct measurement of flat-band voltage in MOS by infrared excitation", *Appl. Phys. Lett.*, vol. 21, no. 5, pp. 194–195, 1972.

[2] A. Jakubowski and S. Krawczyk, "Electrical properties of the MIS capacitor under illumination", *Electron Technol.*, vol. 11, no. 1/2, pp. 3–22, 1978.

[3] A. Jakubowski and S. Krawczyk, "Photoelectric method of the MIS flat-band voltage determination", *Electron Technol.*, vol. 11, no. 1/2, pp. 23–35, 1978.

[4] O. Engstrom and A. Carlsson, "Scanned light pulse technique for the investigation of insulator – semiconductor interfaces", *J. Appl. Phys.*, vol. 54, no. 9, pp. 5245–5251, 1983.

[5] E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*. New York: Wiley, 1982.

[6] M. Leśko and H. M. Przewłocki, "Badanie wpływu temperatury na parametry elektryczne struktur MOS", *Elektronika*, no. 2–3, pp. 56–57, 2005 (in Polish).

[7] M. Leśko, "Badanie wpływu temperatury na parametry i charakterystyki C(V) struktur MOS". M.Sc. thesis, Warsaw University of Technology, Institute of Electron Technology, Warsaw, 2004 (in Polish).

[8] H. M. Przewłocki *et al.*, "The lateral distribution of the effective contact potential difference over the gate area of MOS structures", *Internet J. Electron Technol.*, vol. 35, no. 6, pp. 1–6, 2003.

[9] H. M. Przewłocki *et al.*, "Distribution of the contact potential local values over the gate area of MOS structures", *Microelectron. Eng.*, vol. 72, pp. 165–173, 2004.

[10] A. Kudła *et al.*, "Photoelectric measurements of the local value of the contact potential difference in the metal-insulator-semiconductor MIS structures", *Thin Solid Films*, vol. 450, pp. 203–206, 2004.

[11] H. M. Przewłocki *et al.*, "Variability of the local ϕ_{MS} values over the gate area of MOS devices", *J. Telecommun. Inform. Technol.*, no. 1, pp. 34–39, 2005.

[12] K. Piskorski and H. M. Przewłocki, "Distribution of potential barrier height local values at Al-SiO_2 and Si-SiO_2 interfaces of the metal-oxide-semiconductor (MOS) structures", *Internet J. Electron Technol.*, vol. 36, no. 5, pp. 1–5, 2004.

[13] K. Piskorski and H. M. Przewłocki, "Distribution of potential barrier height local values at Al-SiO_2 and Si-SiO_2 interfaces of the metal-oxide-semiconductor (MOS) structures", *Bull. Polish Acad. Sci.*, vol. 54, no. 4, pp. 461–468, 2006.

[14] K. Piskorski and H. M. Przewłocki, "Investigation of barrier height distributions over the gate area of Al-SiO_2-Si structures", *J. Telecommun. Inform. Technol.*, no. 3, pp. 49–54, 2007.

[15] H. M. Przewłocki, K. Piskorski, A. Kudła, and D. Brzezińska, "Distributions of barrier heights, difference of effective contact potential, and local values of flat-band voltage in Al-SiO_2-Si and poly-Si-SiO_2-Si structures", *Thin Solid Films*, vol. 516, pp. 4184–4189, 2008.

[16] C. H. Bjorkman, J. T. Fitch, and G. Lucovsky, "Correlation between midgap interface state density and thickness-averaged oxide stress and strain at Si/SiO_2 interfaces formed by thermal oxidation of Si", *Appl. Phys. Lett.*, vol. 56, no. 20, pp. 1983–1986, 1990.

[17] S. M. Hu, "Stress-related problems in silicon technology", *J. Appl. Phys.*, vol. 70, no. 6, pp. 53–80, 1991.

[18] I. De Wolf, H. E. Maes, and S. K. Jones, "Micro-Raman study of stress distribution in local isolation structures and correlation with transmission electron microscopy", *J. Appl. Phys.*, vol. 71, no. 2, pp. 898–906, 1991.

[19] I. De Wolf, H. E. Maes, and S. K. Jones, "Stress measurements in silicon devices through Raman spectroscopy: bridging the gap between theory and experiment", *J. Appl. Phys.*, vol. 79, no. 9, pp. 7148–7156, 1996.

[20] K. F. Dombrowski, I. De Wolf, and B. Dietrich, "Stress measurements using ultraviolet micro-Raman spectroscopy", *J. Appl. Phys.*, vol. 75, no. 16, pp. 2450–2451, 1999.

Krzysztof Piskorski was born in Zgierz, Poland, in 1976. He received the M.Sc. degree from the Technical University of Łódź, Poland, in 2002. His Masters project in "Dry etching of A_{III}–B_V nitrides" was carried out at the Ecole Centrale de Lyon, France, in 2001. He is currently working as a Research Assistant at the Institute of Electron Technology (IET) in Warsaw, Poland, with the Department of MOS System Studies. His research interests include photoelectric measurements for MOS structures.

e-mail: kpisk@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Henryk M. Przewłocki obtained the M.Sc., Ph.D., and D.Sc. degrees in 1959, 1969, and 2001, respectively, specializing in physics, technology, and measurements of MOS structures. In the period of 1965–1983, he also taught various courses to graduate and undergraduate students of the Technical University of Warsaw. At present he is Head of a research group at the Institute of Electron Technology (IET) in Warsaw, specializing in photoelectric properties and measurement methods of MOS structures. He has served as consultant and committeeman to numerous scientific, industrial and educational organizations in Poland and worldwide and has received a number of prestigious awards, including the highest Polish award, the National Award for Outstanding Achievements in Technical Sciences.

e-mail: hmp@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland

Paper

Lateral Force Calibration Method Used for Calibration of Atomic Force Microscope

Magdalena Ekwińska and Zygmunt Rymuza

Abstract—Modern heterogeneous micro- and nanostructures usually integrate modules fabricated using various materials and technologies. Moreover, it has to be emphasized that the macro and micro nanoscale material parameters are not the same. For this reason it has become crucial to identify the nanomechanical properties of the materials commonly used in micro- and nanostructure technology. One of such tests is a nanowear test performed using the atomic force microscope (AFM). However, to obtain quantitative measurement results a precision calibration step is necessary. In this paper a novel approach to calibration of lateral force acting on the tip of an AFM cantilever is discussed. Presented method is based on application of known lateral force directly on the tip using a special test structure. Such an approach allows for measurements of nanowear parameters (force, displacement) with the uncertainty better than ±3%. The calibration structure designed specifically for this calibration method is also presented.

***Keywords*— AFM, calibration structure, cantilever, MEMS.**

1. Introduction

The current trend for miniaturization of mechanical components brought not only the shift of manufacturing technology from conventional to silicon, but most of all made it necessary to describe the behavior of microelectromechanical system (MEMS) in the scale in which they operate. In micro and nanoscale the forces applied as well as the areas of contact are much smaller than in macroscale. In addition the microscale influence of such forces as adhesion or capillary forces is much more significant than the macroscale one. Under these circumstances the well known macroscale material parameters are not applicable in microscale and the easiest way to describe the microscale material properties is to perform an experiment in the same scale. Such investigations may be done with the use of atomic force microscope (AFM), which is a powerful device for estimation of, e.g., micro- and nanoscale wear resistance of materials. AFM is composed of a probe scanner, probe displacement detector, electronics connected with a computer and a system of isolation from vibrations. The scanner, which is the heart of the system, enables movement between the sample and the probe to be achieved. The scanner is usually a ceramic piezoelectric device that may move a sample or a probe. The probe is a cantilever, that is a lever with a sharp (cone or pyramidal) tip at one end. During wear test the cantilever tip is in contact with the sample surface. Depending on the applied scanning direction the lever is bending or twisting. The detecting system of cantilever

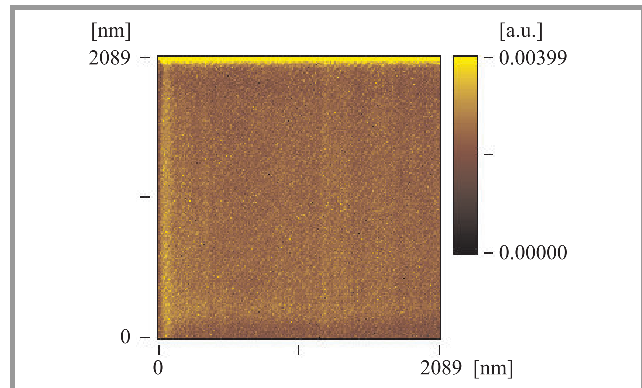


Fig. 1. Lateral force signal obtained during nanowear test.

displacement is usually a laser beam and a four-segment photodiode. The laser beam is focused on one end of the cantilever, where it is deflected and falls on the detector. The change of the laser beam position on the detector is later converted to the force (lateral or normal) signal in arbitrary units [a.u.] (Fig. 1). Therefore, a calibration is needed.

2. Calibration Method

In the literature there are several calibration techniques for lateral forces. The most known are analytical and geometrical [1] methods, two-step calibrating method for a scan parallel to the long axis of the cantilever [2], improved wedge calibration [3], etc.

The first method is a simple analytical one in which the cantilever stiffness is established from the macroscale equation. This method involves precise measurements of cantilever dimensions. These dimensions may be either taken from a catalog or measured experimentally. The maximum uncertainty may reach 100% of the measured value.

The second method is a geometric one [1] in which the cantilever torsion angle is established. Conversion of the torsion angle into lateral force is obtained by the analysis of the optical geometry of the laser-beam path. In this calculation precise estimation of the cantilever tip height is also required. The uncertainty of the discussed method is around 30%.

The two-step method of calibration described in [2], [4] for a scan parallel to the long axis of the cantilever unfortunately cannot be considered as a quantitative one.
The wedge calibration method [3] is a direct calibration of the lateral force by applying a known turning moment or by the definition of the torsion moment of the cantilever on a substrate with well-defined slopes. An improved wedge calibration method utilizes a commercial calibration grating TGF11 (manufactured by NT – MDT Inc. in Moscow, Russia) takes into account the effect of the tip radius of curvature and eliminates the need for multiple measurements with different loads. The inconveniences of this method are quite complicated calculations and the need for stiffness calibration in normal direction. Moreover, during calibration additional parameters such as friction coefficient and adhesion force between cantilever and sample surface should be used. The uncertainty of this method is between 3 and 11% (95% level of confidence) depending on the applied normal force.
There is also a method in which the resonance frequency of cantilever with additional mass is used. The main disadvantage in this case is the application of the additional mass, which may lead to cantilever tip damage or change its parameters. The error in this method is 15%.
The present work deals with a new method of lateral force calibration. In this method the whole AFM is treated as a black box. Extortion is a well known value of lateral force applied at the end of the cantilever tip. This force causes cantilever torsion, then a change of the position of the laser beam spot on the detector, and a change of the signal from the detector, and finally a change of the output parameter: namely, the lateral force signal in arbitrary units. Comparison of the applied force in nN and the force in a.u. received from the device enables estimation of calibration parameter K in nN/a.u.

3. Calibration Sample

In order to calibrate the lateral force, a system with elastic element was elaborated [5], [6] (Fig. 2). There are three main parts of the elastic sample: surface on which the cantilever tip is located (*1*) with specially prepared roughness, flat spring which is bended in S shape (*2*), surface to which AFM table is mounted (*3*). During calibration of the lateral force this calibration device is placed on the AFM table and the cantilever tip is approached to the surface (*1*). After obtaining contact a known value of displacement is applied by the AFM scanner. The surface to which the AFM table is mounted (*3*) moves according to the displacement applied by the piezoscanner. If surface (*1*) of the calibration sample were free then it would move. During calibration the surface cannot move because it is held in one position by the cantilever tip. That causes bending of the element (*2*) of the calibration sample. It is worth underlining that the displacement of the bending element is smaller than the one applied by the AFM table. The small displacement of the element (*2*) multiplied by the stiffness of the calibrating device (which has to be obtained earlier) is the real value of the applied force acting on the cantilever tip trapped between microasperities of the surface (*1*). This force causes twisting of the cantilever and the change of the lateral force signal in arbitrary units. This procedure is repeated from 100 to 500 times. As a result of the calibration of the AFM a cloud of points representing the dependence of the lateral force (in a.u.) between cantilever and elastic element from bending (caused by the scanner movements), is obtained (Fig. 3).

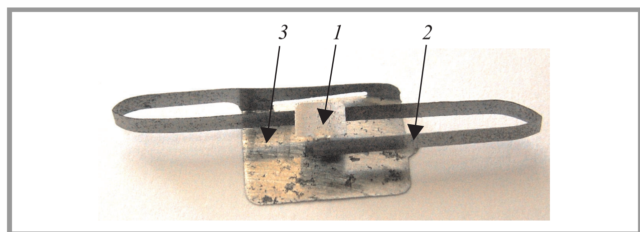


Fig. 2. Lateral force calibration sample. Explanations: *1* – element, which is elastically deformed during calibration, *2* – area where AFM's tip stands during calibration, *3* – device holder.

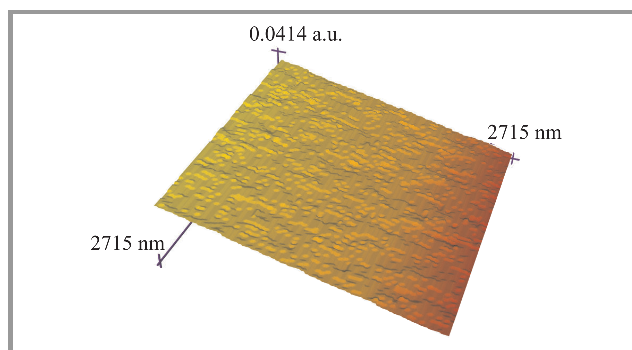


Fig. 3. AFM image presenting the dependence of lateral force between cantilever and elastic element on its bending (caused by scanner movements).

Due to the monolithic structure of the calibration sample and the use of the AFM scanner as an element responsible for the movements of the calibration sample the calibration procedure is simple and may be done directly on a cantilever mounted on AFM just before or just after wear test measurements. The uncertainty of the presented calibration method does not exceed ±3%. Figure 4 presents schematic estimation of the uncertainty of the calibration method. Apart from uncertainty of the AFM measurements and that of the calibration of lateral force, the range of force has big influence on the measurement results.
The results of uncertainty estimation for different force ranges is presented in Fig. 5. It can be seen that the uncertainty do not exceed 40% in the worst case. The worst case is when the calibrating procedure is made for very small force ranges, e.g., several nanometers and several dozen, the stiffness of the cantilever differs from the stiffness of the calibration sample a hundred times and the forces

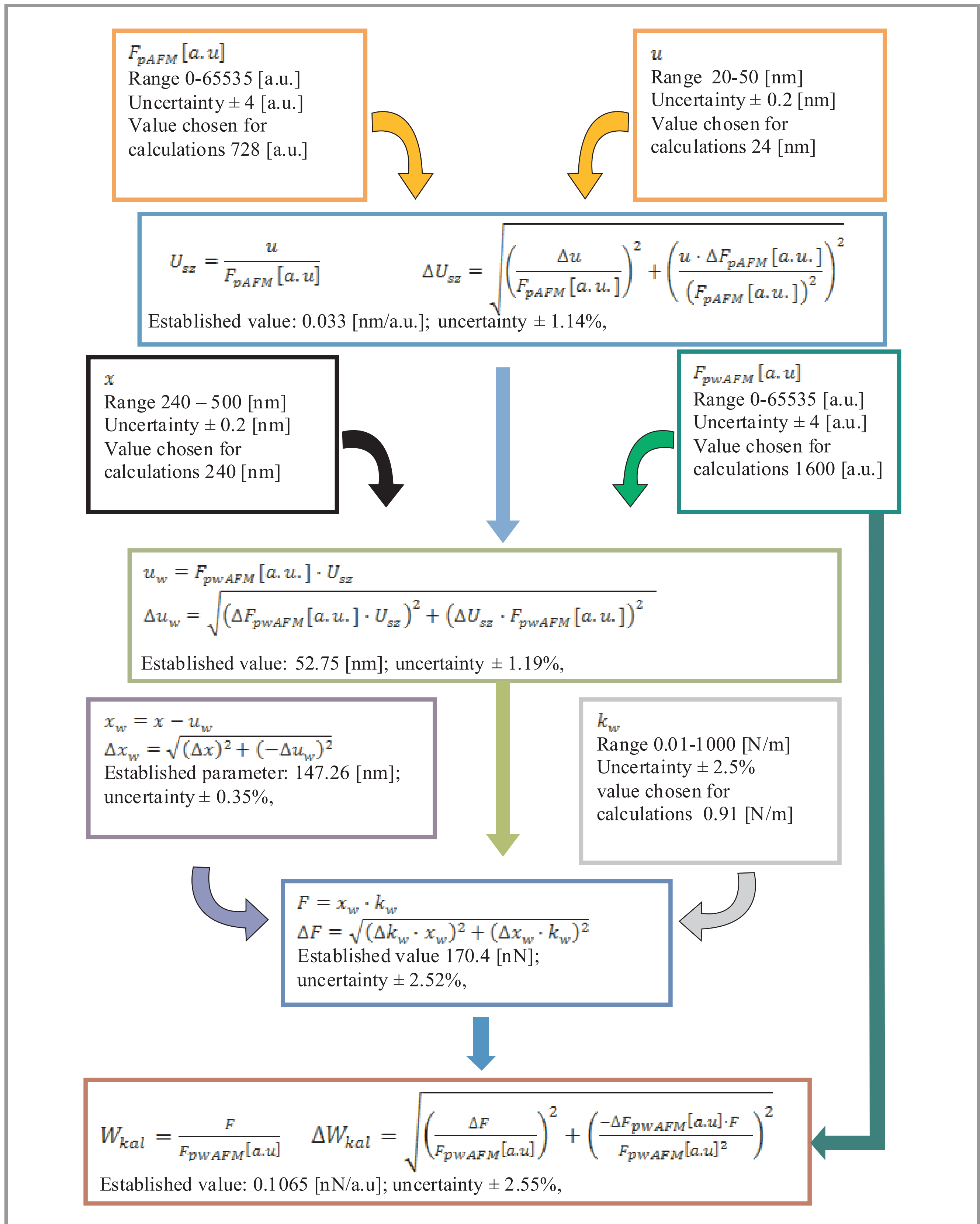


Fig. 4. Estimation of the uncertainty of the lateral force calibration method. Explanations: u – displacement of the cantilever tip during calibration procedure, F_{pAFM} – lateral force, U_{sz} – multiplier describing how big displacement of the cantilever tip causes a change of 1 a.u. of the lateral force during rigid/stiff sample test, F_{pwAFM} – change of the lateral force value, u_w – displacement of the cantilever tip during calibration on the calibration sample, x_w – real displacement, x – applied displacement, k_w – stiffness of the calibration sample, W_{kal} – calibration coefficient, F – lateral force in real units [nN].

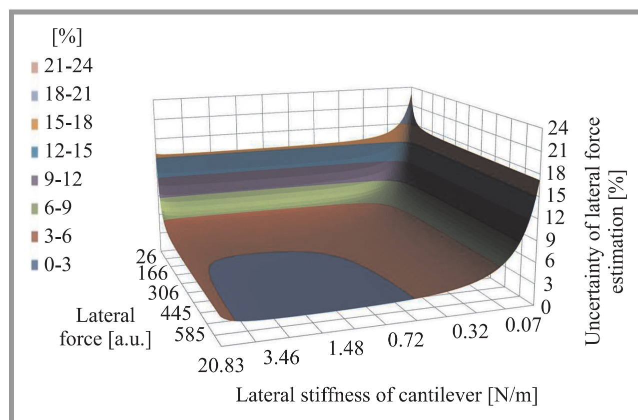


Fig. 5. Uncertainty of the real value of the lateral force estimation as a function of the measured lateral force and cantilever stiffness, with the calibration sample stiffness $k_w = 1$ N/m, uncertainty of the stiffness of the calibration sample ± 0.025 N/m (2.5% of stiffness value).

during real measurements (measurement after calibration) are around several dozen [a.u.].

4. Summary

A new method of calibration of lateral force in AFM was presented. This method may be used for most of the cantilevers available on the market. In this new calibration approach a new additional calibration device is used. During the calibration procedure the cantilever is fixed in the microscope in the same way as during measurements. A dependence of the lateral force in arbitrary units on the applied force is obtained as a result of the two-step calibration. This relationship depends on lateral stiffness of the whole measuring system in the microscope, lateral stiffness of the cantilever and the stiffness of the fixation of the cantilever to the microscope. The calibration should carried out for every newly mounted cantilever.

The presented calibration method is novel, easy to use and applicable to a cantilever that is already mounted in the AFM system. The primary source of the calibration uncertainty as well as the uncertainty during later lateral force measurements is the uncertainty of the calibration of the calibration sample itself. The uncertainty of the stiffness measurement of the calibration sample is ± 2 to 3%. The uncertainty of the lateral force measurement on the AFM is smaller than 0.01% of the measuring range.

Using the presented method the uncertainty of the calibration of lateral force in AFM does not exceed ± 3%. To achieve this the following conditions should be fulfilled:

- uncertainty of the stiffness of lateral force calibration sample should not exceed ± 2.5%;
- stiffness of the measuring cantilever should not depart far from the stiffness of the calibration sample (it should not be smaller than 0.7 k_w and simultaneously it should not be higher than 10 k_w, where k_w = stiffness of the calibration sample);
- force range during and after calibration should not be lower than 0.5% of the AFM measuring range;
- displacement range applied during and after calibration should not be lower than 0.5% of measuring range of the AFM.

This method enables better precision of measurements to be obtained than with the use of the methods known from the literature. Besides it may be used for all types of cantilevers and does not depend on clamping of the cantilever or stiffness of the whole measuring unit.

References

[1] E. Liu, B. Blanpain, and J. P. Celis, "Calibration procedures for frictional measurements with lateral force microscopy", *Wear*, vol. 192, iss. 1–2, pp. 141–150, 1996.

[2] J. Ruan and B. Bhushan, "Atomic-scale friction measurements using friction force microscopy: part I – general principles and new measurements techniques", *ASME J. Tribol.*, vol. 116, no. 2, pp. 378–388, 1994.

[3] M. Varenberg, I. Etison, and G. Halperin, "An improved wedge calibration method for lateral force in atomic force microscopy", *Rev. Sci. Instr.*, vol. 74, iss. 7, pp. 3362–3367, 2003.

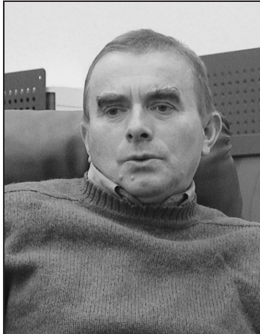
[4] R. G. Cain, S. Biggs, and N. W. Page, "Force calibration in lateral force microscopy", *J. Coll. Interf. Sci.*, vol. 227, iss. 1, pp. 55–65, 2000.

[5] M. Ekwińska, "A new method of calibration of lateral force in atomic force microscope (AFM)", *Mach. Dyn. Probl.*, vol. 28, no. 3, pp. 89–94, 2004.

[6] M. Ekwińska, "The work of friction during nanowear process", *Elektronika*, no. 8–9, pp. 206–209, 2004.

Magdalena Aleksandra Ekwińska received the M.Sc. and Ph.D. degrees in micromechanics from the Warsaw University of Technology, Poland, in 2002 and 2009, respectively. The doctoral thesis was devoted to investigation of nanoscale material properties and comparison of two commonly used nanoscale tests: nanowear and nanoindentation. One of the results was the method of prediction of the results of the time consuming nanowear test based on the information from nanoindentation test only. In 2001 she was an Intern with Hysitron Company (manufacturer of nanoindentation equipment) and in 2003 with the Mechanical-Mathematical Department of the University of Belarus. From 2007 she is a member of Polish Tribology Society (PTT), which is a founder member of International Tribology Council (ITC). In 2008 she joined the Institute of Electron Technology to the Department of Silicon Microsystems and Nanostructure Technology.

Her work is focused to MEMS multidomain modeling and simulation of micro- and nanostructures.
e-mail: ekwinska@ite.waw.pl
Institute of Electron Technology
Lotników av. 32/46
02-668 Warsaw, Poland



Zygmunt Rymuza received Ph.D. and habilitation (D.Sc.) degrees in the design of fine mechanisms and tribology of miniature systems at the Warsaw University of Technology, Poland. He is Professor of the Warsaw University of Technology (Department of Mechatronics) in the Institute of Micromechanics and Photonics and the Leader of microtribology research group and Head of two laboratories: Laboratory of Micro/Nanotechnology and Laboratory of Microtribology. His research field now concerns mainly nanomechanics and micro/nanotribology of ultrathin films and MEMS/NEMS/nanotechnology materials and miniature tribosystems embodied in mechatronic devices. He is author of monographs on tribology of miniature systems as well as tribology of polymeric and many technical papers published in many international and Polish journals and presented during most important conferences. He organized and was the chairman of a series of unique, well-known in the world international conferences on microtribology. He is the member of board and vice-president of Polish Tribology Society. His educational activities relate to the lectures and classes for graduate and undergraduate students in "Micromechatronics", "Microtribology", "Nanotechnology", "Surface Engineering in Microtechnology", "Design of Precise Instruments".
e-mail: z.rymuza@mchtr.pw.edu.pl
Institute of Micromechanics and Photonics
Warsaw University of Technology
Św. A. Boboli st 8
02-525 Warsaw, Poland

Paper

The Influence of Meteorological Phenomena on Modern Satellite Systems

Jan Bogucki, Jacek Jarkowski, and Ewa Wielowieyska

Abstract— The areas of attention, described in this paper, extend throughout the modern satellite systems. Future satellite systems are to be planned for the millimeter band, which has greater weather attenuation effects than until now used bands. This paper provides a brief overview of propagation factors on millimeter-band earth-satellite paths and requirements in relation to the need for specific types of propagation data.

Keywords— millimeter waves, propagation, satellite link.

1. Introduction

A variety of commercial organizations have recently expressed an intent to provide commercial earth-space service via millimeter-band satellite systems. Both mobile and fixed services have been proposed using geostationary orbit or non-geostationary orbit satellite systems, and non-commercial systems are also planned.

Satellite telecommunication has grown to be the most important commercial space application. In terms of business volume, industrial activity and employment generated, satellite communication is so far the most important segment of the industry.

The modern satellite technology has shown the world new ways to use orbital space and radio spectrum resources. It offers alternatives in voice, video and data communications networking to distant places where there is a little or no ground infrastructure. Satellite systems will play an important complementary role in providing the global coverage for both fixed and mobile communication. There are four important satellite industry segments:

- satellite service (mobile, fixed and broadcasting);
- satellite manufacturing and their components (for commercial and government customers);
- launch industry (launch services, production of vehicle and their components);
- ground equipment (networks and consumer equipment).

Satellites are a key component of the world's communication infrastructure, including technically advanced and developing countries alike. For example, the UK space industry sector growing by nearly 8% in 2006/2007 and overall turnover of 5.8 GBP billion with almost 19,000 people employed in high-tech, high value jobs [1].

Overall worldwide industry revenue growth was 16% from 2006 to 2007. Satellite television and direct broadcast satellite (DBS), representing three-quarters of total satellite services revenues in 2007, increased 18% overall to 55.4 billion dollars. The present crisis can slow down this growth [2].

These economic factors make a last decade saw a big increase in the number of satellites and satellite operators and also make innovation in development of satellite systems. Nowadays, multibeam satellites are introduced as an attractive means to reduce the size and cost of the earth stations. Regenerative satellites and onboard processing are also covered as they fulfill the same objective. Actually, these techniques bring more complexity to the routing of information compared to the "bent pipe" routing associated with single beam satellites.

2. Performance Parameters of Modern Satellite Systems

There is a trend in the satellite telecommunication towards larger effective apertures, a significantly higher number of smaller beams, a higher effective isotropic radiated power, a higher gain-to-temperature ratio, more complex switching functions, and onboard processing functions. Satisfying those demands will require an antenna technology significantly more advanced than that employed by current wide area coverage transponder systems [3].

There are very important parameters of antennas:

- *antenna gain*: high gain antennas minimize the terminal size and maximize the capacity;
- *number of beams*: determines the percentage of the desired coverage area available;
- *coverage flexibility*: defines the ability of the antenna to provide high performance coverage across a wide field of view.

Most of the new systems propose to employ new technologies such as multiple narrow spot-beam antennas, on-board demodulation and routing of traffic between beams, inter-satellite links, and in some cases scanning beams to continuously illuminate the service area as the satellite flies by. It gives the satellite communication systems a huge potential to offer, promising high-capacity transmission capabilities over wide areas.

Modern satellite systems have to operate at the higher frequencies, i.e., in the millimeter range. Communication between the RF terminal and the satellite is governed by the basic principles of electromagnetic wave propagation. This

spectrum of radiation covers everything from AM radio to light, but satellite systems operate in microwave frequency between about 1 GHz and 80 GHz (the segment above 30 GHz is more aptly called millimeter waves). These frequencies are not so crowded, and channels are wider there. All frequency bands are allocated by the International Telecommunication Union (ITU) and its committees and conferences.

Contemporary communication satellite systems have entered a period of transition from point-to-point high-capacity trunk communications between large, costly ground terminals to multipoint-to-multipoint communications between small, low-cost stations. A technique called frequency reuse allows satellites to communicate with a number of ground stations using the same frequency by transmitting in narrow beams pointed toward each of the stations. Beam widths can be adjusted to cover areas as large as the entire Europe or as small as Mazovia Province. Two stations far enough apart can receive different messages transmitted on the same frequency. Satellite antennas have been designed to transmit several beams in different directions, using the same reflector [4].

3. Propagation on Satellite Paths

Propagation impairments produced by the troposphere are limiting factors for the effective use of the millimeter range. Use of smaller earth terminals, while very attractive for consumer and transportable applications, make it difficult to provide sufficient link margin for propagation related to outages.

Future satellite systems are planned for the millimeter range, which has greater troposphere and weather attenuation effects than C- (6/4 GHz) and Ku- (14/12 GHz) bands [5]. In the emergence of new satellite communication systems operating in the millimeter range, the role of atmospheric effects on propagation paths has gained increased significance. The impairing factors of rain have always been considered when designing links at up to centimeter range.

Meteorological statistics are abundantly available for locations around the world. But often, the parameters maintained in weather records do not reflect the information require by propagation analysis (Table 1).

Table 1
Standard meteorological parameters of interest to meteorologists and propagation analysis

Typical weather records		Propagation interest	
Max/min temperature	[°C]	Mean temperature	[°C]
Relative humidity	[%]	Water vapour density	[g/m^3]
Cloud cover	[%]	Cloud liquid water	[kg/m^2]
Rain accumulation	[mm]	Rain rate	[mm/hr]

The ITU-R has worked out the tables of needed parameters and they are accessible on the Internet pages of ITU-R.

The optimal performance of satellite path is when it works with 99.99% availability. Accurate estimates of the propagation impairments that the effect link quality and availability and determine signal interference fields are essential for the reliable design of telecommunication systems and the efficient use of the electromagnetic spectrum.

At the National Institute of Telecommunications (NIT), Warsaw, a computer program for the system power budget analysis of satellite radio links was developed. As known, such analysis is very important in the radio link network planning and the optimization of the existing transmission networks.

This paper addresses the issues related to predicting different types of propagation impairments as well as combining them together to determine the overall impact on satellite links over a wide range of outage probabilities.

4. Equations for Satellite Paths

Every communication link through satellite includes a transmission from an earth station to the satellite and a transmission from the satellite to the earth station. Therefore to calculate the system performance two sets equations are used; one for the uplink and one for the downlink.

The carrier power to noise power spectral density ratio at the satellite receiver input (uplink) can be calculated as follows:

$$\frac{C_u}{N_{ou}} = \frac{w\, P_T\, G_T\, G_{ru}\, \lambda^2}{(4\pi R_u)^2\, \mathrm{k} T_d\, L_d} \qquad (1)$$

or expressing in dB:

$$\left(\frac{C_u}{N_{ou}}\right)_{[\mathrm{dB}]} = P_{T[\mathrm{dBW}]} + G_{T[\mathrm{dBi}]} + G_{ru[\mathrm{dBi}]} + 20\lg\frac{\lambda}{4\pi R_u} + 10\lg\frac{w}{\mathrm{k}T_d} + 10\lg\frac{1}{L_d}\,, \qquad (2)$$

where: C_u – power of the received carrier at the input to the satellite transponder, N_{ou} – noise power spectral density, P_T – power fed to the transmitting antenna, G_T – earth station transmit antenna gain in pertinent direction, R_u – distance, earth station to satellite, λ – wave length emitted at earth station to satellite path, G_{ru} – the satellite receive antenna gain in pertinent direction, T_d – the uplink system noise temperature, k – the Boltzman constant, $\mathrm{k} = 1.38 \cdot 10^{-23}$ [JK^{-1}], w – antenna efficiency, L_d – additional losses – attenuation due to propagation conditions in troposphere.

A set of six elements is used to fully describe the carrier power to noise power spectral density ratio. The first element is often called *effective isotropical radiated power* (EIRP) and the second *free space loss*. The last element shows additional losses which are attenuations due to propagation conditions in troposphere – due to rain, atmospheric gases, clouds and fog. They can be described only statistically.

The downlink is similar to the uplink. The equations are the same, but there are differences in the numbers and the emphasis.

5. Propagation Effects for Satellite Paths

The atmospheric loss is the result of the combined effects of attenuation due to atmospheric gases and attenuation due to water (rain, clouds, snow and ice) [6], [7]. The results of calculation of free space loss and additional losses for several of elevation angles (path distance) and frequencies from 10 GHz to 50 GHz are presented below. It is assumed that the ground station is situated at NIT and receives signal from several GEO satellites which position changes from 7° up to 30°.

5.1. Free Space Loss Versus Elevation Angle

The free space loss L_{FS} depends on the frequency and on the distance d between the earth station (ES) and the satellite station (SS). The distance of satellite path depends on elevation angle:

$$d = \sqrt{(\mathrm{R}+h)^2 - (\mathrm{R}\cos\theta)^2} - \mathrm{R}\sin\theta\,, \tag{3}$$

where: R – the radius of Earth, 6371 km, h – the satellite orbit height, θ – the elevation angle.

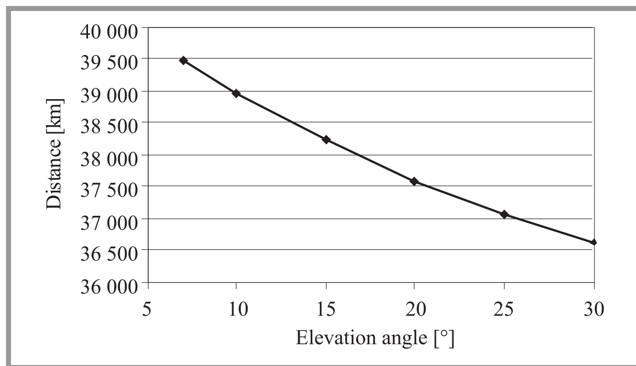


Fig. 1. Distance of GEO – Earth satellite path.

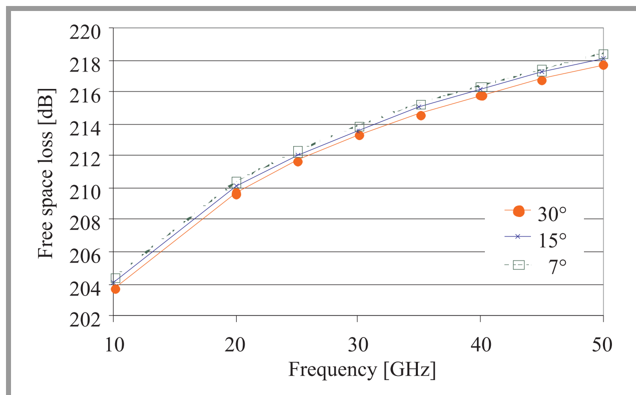


Fig. 2. Free space loss for satellite paths.

The maximum distance from a satellite to an earth station as seen on the edge-of-earth is 41 679 km. The minimum distance, when ES is located on equator "under" the SS is 35 786 km equal $d = h$, when $\theta = 90°$. The length of GEO-Earth satellite path depends on elevation angle is shown in Fig. 1.

The free space loss depends on elevation angle and frequency. When the elevation angle decreases by 5°, free space loss increases only by 0.1 dB for each frequency points at the x axis. It is connected with path lengthen (Fig. 2).

5.2. Attenuation Due to Gas Versus Elevation Angle

Oxygen and water vapour are the main atmospheric gases affecting the signal at millimeter waves. Oxygen concentration is almost constant during the day and during the year and slightly varies over the globe. The oxygen specific attenuation depends on frequency, ground temperature and atmospheric pressure. The amount of water vapour is highly variable being the function of temperature and of atmospheric conditions. ITU-R Rec. P.676-6 [8] indicates how to calculate yearly average gaseous attenuation.

Pressure, temperature and water vapour are functions of the height and consequently these parameters depend on elevation angle.

When the elevation angle θ includes from 5° up to 90°, attenuation due to gas can then be written as follows:

$$A = \frac{h_o\gamma_o + h_w\gamma_w}{\sin\theta}, \tag{4}$$

where: γ_o – specific attenuations due to dry air [dB/km], γ_w – specific attenuations due to water vapour [dB/km], h_o – equivalent height for dry air, h_w – equivalent height for water vapour.

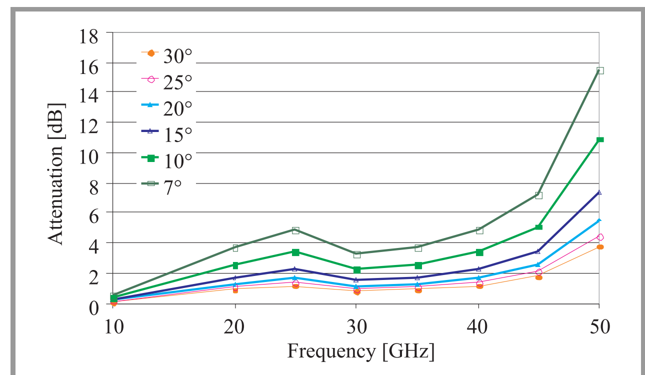


Fig. 3. Attenuation due to dry air and water vapour for satellite paths.

Attenuation due to dry air and water vapour depends on elevation angle and frequency (see Fig. 3).

5.3. Rain Attenuation Versus Elevation Angle

The specific attenuation of rain depends on temperature, terminal velocity and shape (mainly radius) of the raindrops. One of the popular performing models of rain attenuation at frequencies up to 50 GHz is ITU-R Rec. P.618-8 [9]. When measured distributions of rain intensity are not available, the global map of the parameters of the rain intensity recommended by ITU-R P.837-4 [10] can be used, without substantial degradation of the performance of rain attenuation models.

At first the mean rain height above the mean sea level h_R may be obtained from the 0°C isotherm as [11]

$$h_R = h_o + 0.36\,, \tag{5}$$

where: h_o – the height the 0°C isotherm above mean sea level [km].

For $\theta \geq 5°$ compute the slant-path length L_S below the rain height from:

$$L_S = \frac{h_R - h_s}{\sin\theta}\,, \tag{6}$$

where: h_s – the height above mean sea level of the ground station [km].

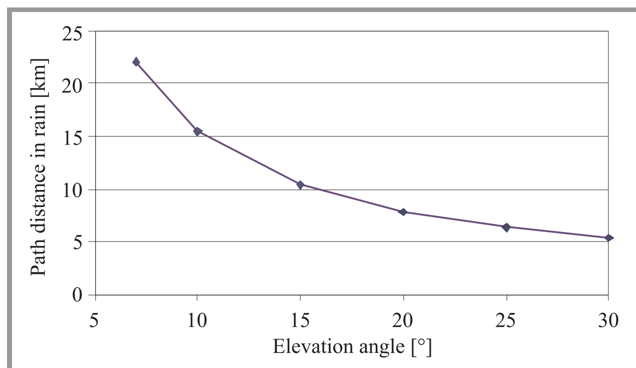


Fig. 4. Effective path length in the rain.

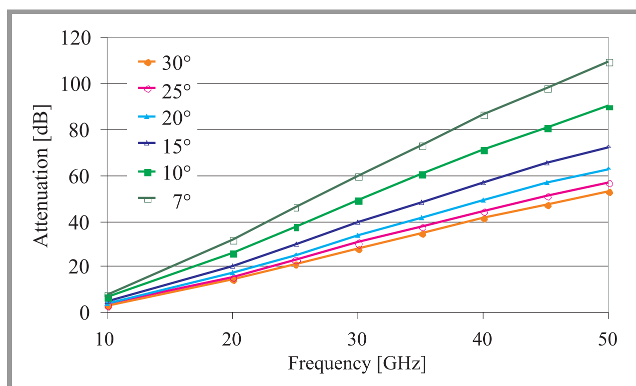


Fig. 5. Attenuation due to rain for satellite paths.

The predicted attenuation $A_{0.01}$ [dB] exceeded for 0.01% of an average year is obtained from:

$$A_{0.01} = \gamma_R L_E\,, \tag{7}$$

where: L_E – the effective path length, γ_R – the specific attenuation,

$$\gamma_R = \mathrm{k}(R_{0.01})^{\alpha}\,, \tag{8}$$

where: $R_{0.01}$ – the rainfall rate, exceeded for 0.01% of an average year (with an integration time of 1 min), γ, α – the frequency-dependent coefficients.

Figure 4 shows the effective path length in rain and Fig. 5 presents how the attenuation due to rain depends on the elevation angle and frequency.

5.4. Attenuation Due to Clouds and Fog Versus Elevation Angle

Clouds and fog consist of suspended water droplets of size smaller than the wavelength for frequencies up to V band. Clouds attenuation is highly variable, depending on the presence or not of clouds along the link and on their liquid water content. One of the commonly used models to compute attenuation is ITU-R Rec. P.676-6 [8] and Rec. P.840-3 [12], which indicates how to calculate clouds attenuation as function of the integrated reduced liquid water content, the frequency, the elevation angle and the dielectric constant of the water.

For clouds or fog consisting entirely of small droplets, generally less than 0.01 cm, the Rayleigh approximation is valid for frequencies below 200 GHz and it is possible to express the attenuation in terms of the total water content per unit volume. Thus the specific attenuation within a cloud or fog can be written as

$$\gamma_c = K_l M\,, \tag{9}$$

where: γ_c – specific attenuation within the cloud [dB/km], K_l – specific attenuation coefficient [(dB/km)/(g/m^3)], M – liquid water density in the cloud or fog [g/m^3].

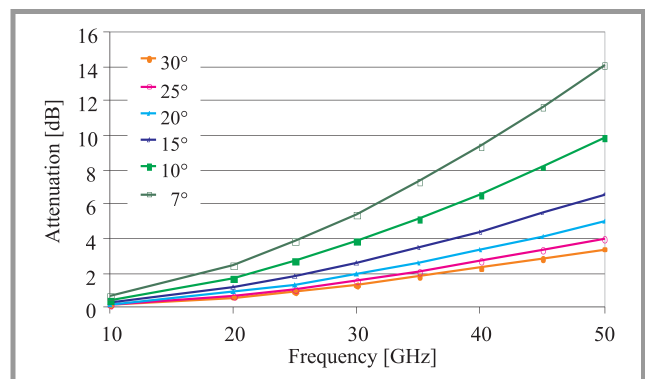


Fig. 6. Attenuation due to clouds or fog for satellite paths.

Attenuation due to clouds and fog depends on elevation angle and frequency (see Fig. 6).

5.5. Total Attenuation

Figures 7 and 8 restore our sense of proportion each part of attenuation takes in total loss and show how total attenuation depends on frequency and elevation angle of GEO path.

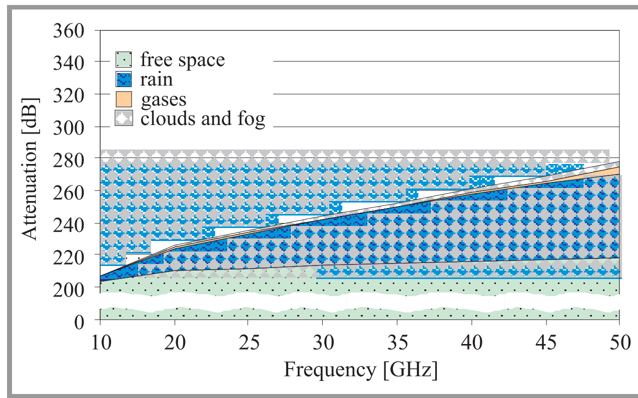


Fig. 7. Total attenuation of GEO path (30° elevation angle).

The ground station is located at NIT with antenna directed on the 30° elevation angle. It is "easy" path because elevation angle is fairly high although propagation conditions are difficult. Total attenuation is 279 dB at 50 GHz within 0.01% time. To achieve availability of 99.99% of time for this path on the frequency of 50 GHz the system requires an excess margin of approximately 60 dB.

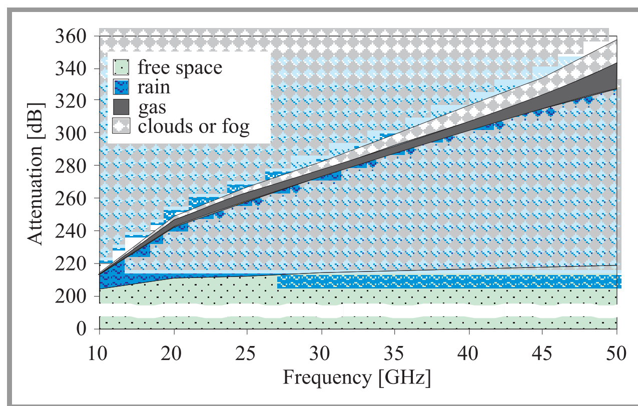


Fig. 8. Total attenuation of GEO path (7° elevation angle).

The reduction of the elevation angle form 30° to 7° causes the increase of the total attenuation about 80 dB at 50 GHz.

Tropospheric scintillations due to small-scale refractive index inhomogenities induced by atmospheric turbulence along the propagation path which causes rapid fluctuations of the received signal amplitude are not taken into consideration in this paper.

6. Example of Prediction and Empirical Data at Satellite Link

The measurements of the 12.5 GHz beacon signal from Lucz 1 were conducted and simultaneously of 1-minute average rain rate under the Earth-Lucz path. Antenna elevation angle was 22° and azimuth 224.3°, location NIT-Warsaw.

Figure 9 shows the rain attenuation statistics, the percentages of the year in 4 consecutive years that attenuation level has been exceeded in case of rain on this path and the average of 4 years. During a storm with very intense rainfall the signal exceeded 20 dB level during 10 min.

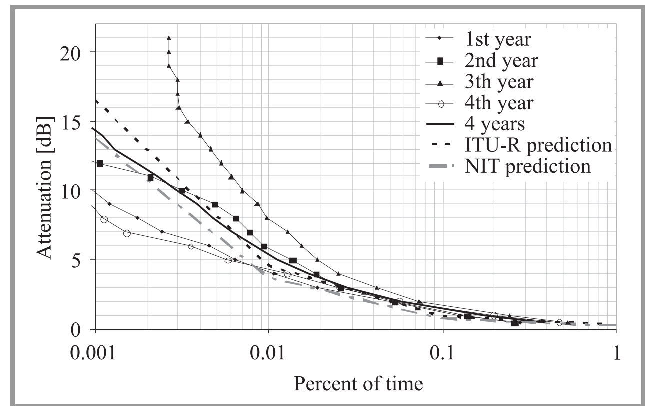


Fig. 9. Measured and calculated distributions of rain attenuation at satellite path.

Empirical annual attenuation distributions were compared with the predicted distribution, based on ITU-R model [9].

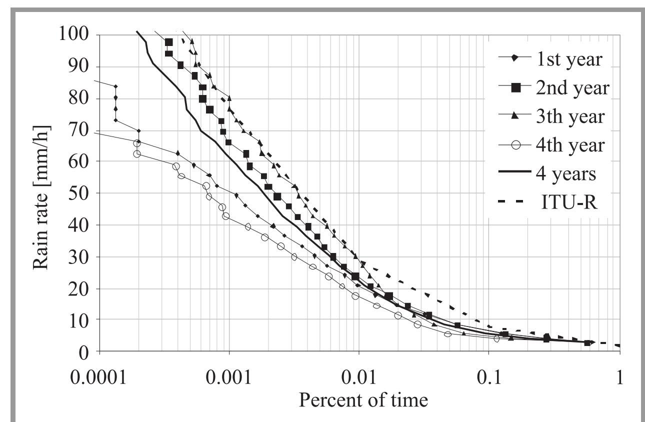


Fig. 10. Measured and calculated distributions of rain rate on the path under satellite link.

The dashed line ITU-R prediction shows calculated rain attenuation exceeded with chosen probability using formula recommended by ITU for average year and the parameter $R_{0.01}$ obtained from the ITU tables. The dashed line NIT prediction shows calculated rain attenuation for chosen time percent using formula recommended by ITU for average year but the parameter $R_{0.01}$ was obtained from NIT measurements [13].

Empirical annual distributions of rain and prediction annual distributions of rain for Warsaw – H region, are presented in Fig. 10.

In our measurement system typing bucket gauges were applied. Their parameters were:

- 1 tip/min corresponded to rain rate of 2.8 mm/h;
- rain rates from this value down to 0.28 mm/h were processed by dedicated software, which averaged single tips in the gaps shorter than 10 min, longer gaps were considered as the breaks between the rain events;
- rain intensity was measured in millimeters per hour with an integration time of 1 min.

Figure 11 shows three distributions of rain rates:

- ITU-R (H): defined by ITU for H region which was recommended by ITU-R Rec. P.837-1 [14] which was in force up to 1994;
- ITU-R: calculated from the tables recommended by ITU-R Rec. P.837-5 [15] which is now in force;
- NIT 4 years: calculated from the empirical data obtained in NIT, averaged for the period of four years.

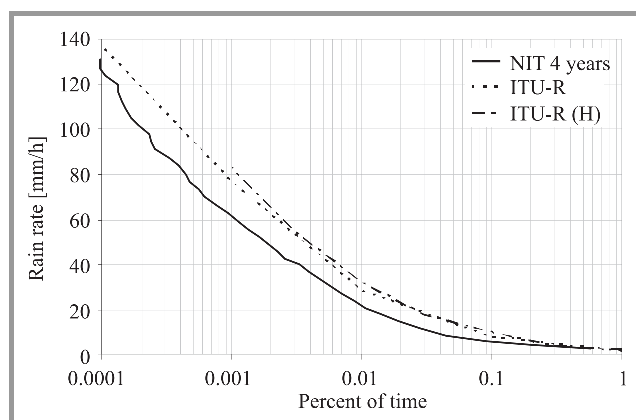


Fig. 11. The comparison of empirical and predicted annual rain distribution.

It is indicated that in recent years theoretical predicted satellite attenuations changed in accordance with real conditions. Difference observed between measured and predicted attenuation is smaller nowadays (ITU-R Rec. P.837.5 [15]) than a dozen years ago (ITU-R Rec. P.837.1 [14]).

7. Conclusions

Considerable variation of attenuation across the satellite coverage area is fairly common and the system design can be optimized by analyzing such variations. In this context the reliable prediction of propagation impairments for millimeter-wave systems becomes important. Modern satellite links can be properly and precisely engineered to overcome potentially detrimental propagation effects. Knowledge of fading estimation is extremely important for the design of millimeter-wave satellite systems. If reception frequently cuts in and out during light rainstorm or other atmospheric events, this is a good indication that the system has not been peaked to maximum performance.

The role of atmospheric effects on propagation paths gained increased significance with increase of frequency new satellite systems operate on.

Until now only the impairing factor of rain has been considered when designing satellite links. Gases, clouds and fog were previously considered “secondary affects” and now they are appearing to play a significant part in total losses; particularly at higher frequencies and lower elevation angles (Fig. 8).

One of consequences of operating the satellite path millimeter region close to the molecular absorption line is that there can be a significant difference in the atmospheric attenuation between the two edges of the band.

The change of elevation angle by 23° has caused the increase of attenuation due to rain by 56 dB at frequency of 50 GHz and for 0.01% of time (Fig. 5), while the effective path length in rain increase only 16.65 km.

Modern satellite systems ought to operate with high exceedance probability levels up to 1% and then link margins are economically practical.

Signal-to-noise of modern satellite systems required to achieve BER of 10^{-10} (bit error ratio of ten to the power minus ten) can be only 5 dB to comply with suitable coding.

Measured distributions of attenuation at satellite path even at 12 GHz indicated that atmospheric events influence is significant at quality of slant links. For accurate prediction of attenuation on slant paths it is necessary to perform long-term data measurements of the above factors different regions.

Our studies indicate that ITU-R model with ITU rain parameters corresponds to experimental few years average data of attenuation caused by rain; maximum difference is 2 dB and the ITU-R model with local rain parameters is even better. The maximum difference is less then 1 dB.

Empirical annual attenuation distributions differ from each other a lot.

In this work, computer program *TraSat* has been used to make calculations for all propagation parameters of satellite links. The program uses algorithms which are recom-

mended by ITU-R and can apply parameters recommended by ITU-R as well as parameters from others resources, e.g., kown local values for chosen parameters.

Benefit of modern satellite systems is their ability to operate with broader bandwidths in the millimeter and the spacecraft is thus able to accommodate smaller antennas. The role of the space industry is very important in meeting our connectivity access telecommunication targets and to contribute to the well-being of the world's population.

It may be that new mobile voice, data and video applications capture consumer interest to growth rates similar to that of satellite television and radio.

Overall satellite industry growth of 16% indicates a fundamental robustness and flexibility to weather business cycles. The prospects of a credit crunch and low stock values will raise the interest of new investors or those returning to the satellite industry sector, with attention focusing on business fundamentals and quality of operations.

References

[1] "Science Minister reveals new figures on growth in UK space industry" [Online]. Available: http://www.bnsc.gov.uk/7286.aspx

[2] *Data from the Satellite Industry Association*. Washington: Satellite Industry Association, 2008.

[3] J. Bogucki and J. Jarkowski, "Kierunki rozwoju satelitarnej techniki telekomunikacyjnej", *Zeszyty Naukowe Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej – Radiokomunikacja, Radiofonia, Telewizja*, no. 1, pp. 183–186, 2007 (in Polish).

[4] J. Bogucki, J. Jarkowski, and E. Wielowieyska, "Propagation on modern satellite paths", in *Proc. Nineteenth Int. Wrocław Symp. Exhib., Electromagn. Compat.*, Wrocław, Poland, 2008, pp. 374–377.

[5] B. R. Elbert, *The Satellite Communication Ground Segment and Earth Station Handbook*. London: Artech House, 2001.

[6] J. Bogucki, "Trasy nachylone w zakresie fal milimetrowych". *Telekomunikacja i Techniki Informacyjne*, no. 3–4, pp. 66–92, 2003 (in Polish).

[7] J. Bogucki and E. Wielowieyska, "Tłumienie dodatkowe w łączu satelitarnym", *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, no. 1, pp. 48–51, 1998 (in Polish).

[8] "Attenuation by atmospheric gases", ITU-R Rec. P.676-6 (03/2005).

[9] "Propagation data and prediction methods required for the design of Earth-space telecommunication systems", ITU-R Rec. P.618-8 (04/2003).

[10] "Characteristics of precipitation for propagation modelling", ITU-R Rec. P.837-4 (04/2003).

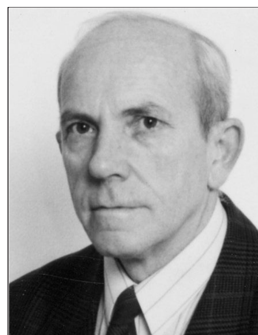
[11] "Rain height model for prediction methods", ITU-R Rec. P.839-3 (02/2001).

[12] "Attenuation due to clouds and fog", ITU-R Rec. P.840-3 (10/1999).

[13] A. Kawecki, "Wieloletnie charakterystyki intensywności deszczu w Miedzeszynie na potrzeby radiokomunikacji", *Prace IŁ*, no. 106, pp. 69–84, 1996 (in Polish).

[14] "Characteristics of precipitation for propagation medelling", ITU-R Rec. P.837-1 (08/1994).

[15] "Characteristics of precipitation for propagation medelling", ITU-R Rec. P.837-5 (08/2007).



Jan Bogucki was born in Warsaw, Poland. He received Eng. degree from the Warsaw Technical University of Technology in 1972. Since 1973 he has been employed at the National Institute of Telecommunications, Warsaw, where he has been engaged in research of digital radiolinks, digital television, EMC capability and microwave propagation in the troposphere.

e-mail: J.Bogucki@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Jacek Jarkowski was born in Warsaw, Poland. He received M.Sc. from the Warsaw University of Technology (WUT) in 1963 and received his Ph.D. degree on radiocommunication science in 1975. Since 1962 was employed in the Faculty of Electronics WUT, and since 2003 he is with the National Institute of Telecommunications, Warsaw. His primary research interests are antennas, propagation and radiocommunication systems and currently wireless sensor networks.

e-mail: J.Jarkowski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Ewa Wielowieyska was born in Warsaw, Poland. She finished the Mathematics Faculty of Warsaw University. Since 1981 she has been employed at the National Institute of Telecommunications, Warsaw, where she has been engaged in research of microwave propagation in the troposphere, propagation digital radio signals on short, medium and long waves.

e-mail: E.Wielowieyska@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Trends in Use of RF Spectrum

Ryszard Strużak

Abstract—This paper reviews possible improvements in the use of radio waves for carrying information from an engineering viewpoint. A few new concepts are proposed, which reduce the problem to an arrangement of solids in a multidimensional space.

Keywords—radio communication hyperspace, radio frequency spectrum resources, radio regulations, spectrum engineering, spectrum management.

1. Introduction

This paper deals with the use of the radio waves for information transmission, a problem under discussion since many years. It is also known also as spectrum management, or management of the radio frequency (RF), or RF spectrum resources. (We use interchangeably the terms *radio waves* and *RF spectrum resources*.) There is opinion that these resources are not used as effectively as they should and could be, and various improvements have been proposed. The issue attracts more and more attention as new technologies appear that can change dramatically the way the RF spectrum has been used and managed.

We seek here general trends rather than detailed analysis of specific improvement proposals. We begin with a short summary of the present spectrum management and its improvement proposals. Next, we move to engineering aspects. We generalize classic notions of the signal mask, receiver selectivity, and transmitter coverage and reduce the problem to an arrangement of multidimensional solids that should be "packed" tightly together, but not too close. Then, we discuss some possible improvements due to scientific and technological progress. The paper is based on the presentation prepared for the URSI General Assembly [1]. Opinions expressed here reflect personal views of the author.

2. Radio Regulations

The first radio regulations were created at the 1903 Berlin conference. All those interested gathered there to assure that the uses made of the radio waves will not infringe their vital interests. It was only two years after the first wireless communication across Atlantic astonished the world. Since then, the significance of radio increased enormously. The International Telecommunication Union (ITU) has been created. The Berlin agreement has been replaced by the ITU convention and radio regulations. These are now legally binding in some 190 countries around the world and this section offers their short summary. More details can be found elsewhere, e.g., in [2], [3].

2.1. International Regulations

The ITU regulations represent the collective wisdom of all the ITU members. The radio waves and the geostationary satellite orbit (GSO) (since 1963) are treated as *common heritage of mankind* that must be fairly shared among all countries. On the international arena they can be used freely, without any restrictions, as long as it *"shall not cause harmful interference to, and shall not claim protection from harmful interference"* [3]. If however there is an interference threat, the regulations define principles on how, when, where and under what conditions their use should be allowed or denied. In essence, the radio regulations are a collection of the principles, requirements, procedures, and plans, commonly agreed among all those interested as necessary for the fair shared use of the RF spectrum and geostationary satellite orbit. As that use depends on the progress in science and technology, the ITU regulations are reviewed every few years. Governments and private sector entities work together toward a compromise, and the participation in the work is open to all those interested. For instance, the preparations for the 2007 World Radiocommunication Conference started 5 years before the conference and resulted in a few thousands modification proposals. The proposals were considered by some 3000 experts from 161 countries, gathered in Geneva, Switzerland. The conference documents count some 12000 pages and the final acts alone – 500 pages. These numbers illustrate how the task is important and difficult.

A conflict-free use of any shared resource implies the exchange of information about the actual and planned uses, collaboration and compromise. The ITU procedures and common planning exercises create a practical framework for that. The system makes it possible to use the RF spectrum (and the GSO) in an orderly and transparent way. Additionally, the ITU standardization activities lead to standard transmission methods, which in turn assures a mass market for equipment and facilitates radio services across the globe, making them cheap.

Uses made of the RF spectrum (and the GSO) that require international recognition are to be registered in the ITU databases. Every new proposal is examined in view of potential conflicts with those notified earlier in the database. The system bases on formal declarations, and does not foresee any monitoring by an independent body. The declared use can differ significantly from the actual use, and most often it does differ, which leads to apparent scarcity of free frequency bands and orbital positions. It is a weak point of the ITU system. Finally, we note that the protection of passive services depends totally on administrative regulations.

2.2. National Regulations

The ITU regulations are a part of an international treaty that governments commit to observe. In most countries they are complemented by national regulations. The purpose is to achieve specific political, economic, or social objectives. Assuring universal access and protecting the investments made, are examples of such objectives. Most national regulations introduce licenses and fees for the access to the RF resources. Only minuscule parts of the RF spectrum (so-called ISM frequency bands) are exempted from the licensing obligations. These are allocated for industrial, scientific, medical and domestic applications of RF waves.

Various licensing criteria are in use around the world. One is the seniority or first-come, first-served principle. Another one is the criterion of *merits* determined through comparative hearings, known also as *the beauty contest*. In some countries licensing is treated a source of revenue and licenses are auctioned. The access to the resources is given to those who pay the most, which is a form of the wealth criterion. In reality it is an extra tax, which increases the governmental budget without any effort (an idea popular among politicians). It also augments the price of radio services and detours money that otherwise could directly be invested in the telecommunications.

2.3. Private Spectrum

A number of economists consider privatization as the best way to improve the use of the RF spectrum resources. They request the spectrum should be ruled by market forces only, as it is with the real estate. It should be traded, aggregated, divided and freely used for a wide range of owner-selected services [4]. However, Elinor Ostrom, the 2009 Nobel Prize laureate in Economic Sciences, challenged that view, showing that common resources can be successfully managed without government regulation or privatization. Privatization is not a new approach, as it prevailed in the pre-regulation era. At that time, it resulted in the power race and chaos and was abandoned. In view of their crucial role played in the safety, security, and wealth of nations, doubts are expressed if the privatization of RF resources serves as well the society as its advocates claim. A patchwork of incompatible proprietary solutions can replace the worldwide standards and wipe out all benefits they offer. Analogy to real estate is misleading, as it ignores the electromagnetic interactions [5]. The borders of tradable piece of the spectrum cannot be unambiguously determined as in real estate.

2.4. Spectrum Commons

We mentioned that every nation enjoys free access to the RF spectrum resources as they are considered a common heritage of the humanity. Advocates of the open spectrum doctrine propose that principle to be extended over every citizen. The concept is not new, like the privatization idea. Indeed, before the regulations, radio waves were a commons and were appropriated as needed, without any formalities. The concept of free (unregulated) use was abandoned because of a flood of litigations. At that time, the radio science and technology were too primitive to cope with the spectrum sharing problem. However, knowledge and technology of today make it possible to come back to the idea, at least in some applications. Wireless computer networks based on the IEEE 802.11 standards is the best example. They are extremely successful because their use does not require any license (they operate in ISM – industrial, scientific, and medical frequency bands).

The doctrine does not sweep away the market. It only restricts market forces to the equipment market, leaving aside the radio waves; much like in the sea transport, where ships and harbors are privately owned, but the use of ocean waters is free and open to everybody. With current technology, the open access can offer only the *best effort* level of the quality of service, which decreases as the number of users increases. Instead of denying the access to the latecomers or those less wealthy, it forces all the users to share degradation of the quality of service.

3. Radio Communication Hyperspace

Three signal-related concepts are discussed in this section: signal solid, propagation mapping, and reception window.

3.1. Signal Solids

Message to be transmitted from source to destination is a continuous function of continuous time, or a sequence of symbols, i.e., a function of discrete time. The transmitter maps it into a radio wave signal, using modulation and other operations. Examples are the analog-to-digital conversion, signal compression, and scrambling. That process involves a number of independent variables. For instance, the signal engages four variables: frequency, amplitude, phase, and polarization of the carrier radio wave. The widths of the frequency band and time slots the signal occupies are other variables. These variables create a multidimensional space, in which the transmitter concentrates the signal energy in some specific regions. We call them *signal solid*. It is easy to notice that it is a generalization of the signal mask which, in our convention, is a projection of the signal solid on the plane frequency-energy. A multidimensional solid cannot be shown on a plane sheet without deformations, but it can be depicted in a series of plane projections, or cross section cuts, assuming fixed values of other variables. The same reasoning is applicable to all signal-related concepts such as the station coverage for instance.

3.2. Propagation

The radio wave propagates along its path loosing energy due to the absorption, spatial spreading, diffusion, and shadowing. The reflection and scattering alter it too, as does so the Doppler effect, if the transmitter, receiver, or reflecting objects are moving. The finite velocity of the wave causes the signal latency. In summary, signal solids change dynamically the form, size, and position. Most of

these effects have random components that can be described only by their probability distributions. Propagation-related effects distort the signal so that the received message differs from that originally sent. The difference, often expressed as bit error rate, is closely related to system performance measures such as transmission range or speed. The transmitting and receiving stations correct the predictable transmission errors by applying appropriate algorithms and data built-in in their hardware and software. In the anticipation of signal distortions, a multitude of operations are used for that purpose, such as modulation and demodulation, coding and decoding, or spreading and despreading. Clearly, the more the system "knows" about the expected distortions, the more efficient the error correction.

3.3. Reception Window

The receiving station creates a multidimensional "window" that – ideally – is transparent to the intended signal and opaque to all other signals. It is easy to note that this is a generalized selectivity mask of the receiver, whose classic definition is restricted to frequency and energy. In analog systems, the reception window is usually a single opening. In digital systems, it consists of a series of non-contiguous openings, stationary or changing in time. All physically realizable systems have also spurious windows through which unintended signals can penetrate. Normally, the intended signal must fit exactly the receiver window in all dimensions as foreseen by the system design. Any overflow or underflow leads to undesired effects. All signals that do not fit that window at least in one dimension are to be rejected. Examples of such signals are those that arrive at wrong times, at wrong frequencies, from wrong direction, or with wrong polarization.

4. Technological Improvements

This section presents some comments on possible improvements and on the role of science and technology.

4.1. Electromagnetic Compatibility

Improved use of the RF resources requires improved electromagnetic compatibility (EMC) engineering methods and tools. They are necessary to identify, analyze, predict, prevent, and reduce signal collisions. It involves radio wave propagation because, for instance, shadowing by natural obstacles can efficiently attenuate unwanted signals in terrestrial microwave applications at no extra cost. It is the least expensive reduction method of unwanted signals, but it requires detailed digital models of the terrain irregularities and man-made objects, and careful selection of the station locations.

4.2. Redundancy

There are techniques that improve the spectrum use by optimizing the quantity of information sent per hertz, or by reducing the quantity of spectrum needed to send bits at a specified rate. Signal compression, for instance, removes redundant bits. The radio wave propagation process can do the opposite – can add extra redundancy due to reflections. The received wave is composed of a number of replicas of the original sent. Each replica arrives at a different time, from different direction, and with different polarization, and interferes with the others. The interference can be destructive, when the waves cancel each other, or constructive, when the waves add together. Various techniques exploit it. The diversity reception, intelligent antennas, single-input-multiple-output (SIMO) systems, and multiple-input-multiple-output (MIMO) systems are examples. The idea of signal summation lies also behind the concept of multiple-input-single-output (MISO) systems, as well as of single-frequency networks (SFN), where two (or more) stations transmitting the same content share a common frequency band coverage region.

4.3. Unwanted Signals

The radio wave carries the intended message to its receiver but does not stop there. It continues farther, and can reach other receivers, where it is neither expected nor wanted. Unwanted signals interact with the intended signal and add to the transmission error. As the propagation medium is shared by all the constituencies of the environment, all the radio waves they radiate coexist and the resultant wave follows activities of the individual radiation sources. That introduces additional random factors to the process. If the characteristics of these signals are known or foreseeable, their negative effects can be eliminated or reduced by appropriate system design.

4.4. Spurious Radiations

Radio equipment produces spurious radiations due to technical imperfections and laws of physics. These do not carry any useful information, but occupy the spectrum. (Man-made ISM emissions fall in that category.) Their reduction increases the equipment cost, which does not translate directly into any tangible benefit for the equipment owner. The beneficiaries are the owner's neighbors. The spurious radiations are thus never completely removed. On the other end, receiving stations have spurious windows that make it possible for unwanted radiations to penetrate and disturb the system operation. Like the unwanted signals from transmitter, they result from technical imperfections and are tolerated because of the cost factor. It is the role of the spectrum managers to keep strict restrictions on such radiations and responses for common benefit. It is not an easy task as it usually requires equipment replacing, which always create an economic loss to someone.

4.5. False Signals

Radio waves interact when applied to a non-linear element. They mix together, and produce a number of false signals

at frequencies $F = (nF_1 \pm mF_2)$ in the case of two signals at frequencies F_1 and F_2, where n and m are integers. Such mixing can involve more than two true signals and produce more spurious products. Usually, their amplitudes are much less than the original true signals and decrease as the order of the mixing product $(n+m)$ increases. Although the spurious do not carry any useful information, they can interfere with useful signals as they were true signals from transmitters.

4.6. Signal Separation

It follows from previous sections that any unintended signal at the receiver must lay outside of the receiver window, at a safe distance from it, at least in one dimension. The safe distance is application and technology dependent and is defined by the system design. It may be the frequency separation, code separation, direction separation, etc. In ultra wide band (UWB) systems, for instance, it is the difference in the signal (spectral) power density: the UWB signal has power density much smaller than that of the (intended) narrow-band signals. The signal separation is often realized in terms of station coverage. Traditionally, that coverage is defined as the geographical area within which service from a radio communications facility can be received. In this context, the service and signal are interchangeable. As a consequence, the concept of signal multidimensionality leads us to the idea of multidimensional coverage solids (Fig. 1). Only three variables are shown in the figure: two geographical coordinates and frequency. The cylinders represent the coverage of three omni directional radio communication systems A, B, and C.

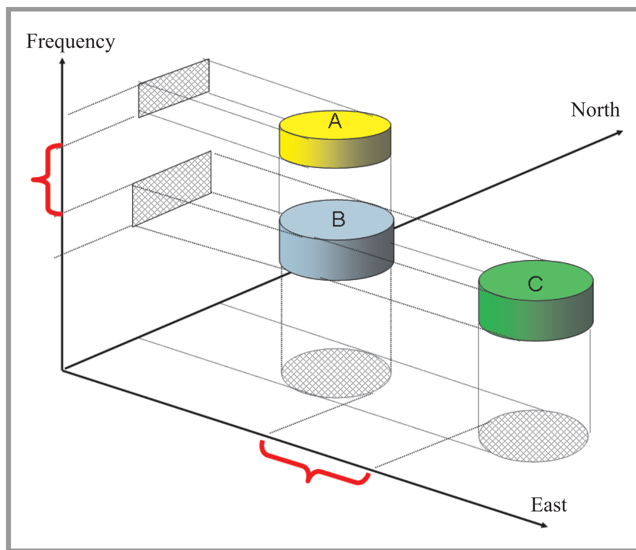


Fig. 1. Separated coverage solids of tree stations A, B, C.

Stations A and B share a common geographical area but are separated in the frequency domain, B and C share a common frequency band but serve separate geographical areas. The coordinate axes may represent also other variables such as time or polarization. This representation applies to all variables, by which one radio signal can be distinguished from others.

The coverage is determined by the energy relations among the desired and unwanted signals and system noise. As a result, the coverage of an isolated station and the coverage of the same station in the presence of neighboring stations can differ significantly (see Fig. 2). The figure presents results of computer simulation of that effect assuming all stations are identical. In this example, the increase of the number of stations can result in the reduction of the original coverage up to 25% of its original value. To lower the coverage

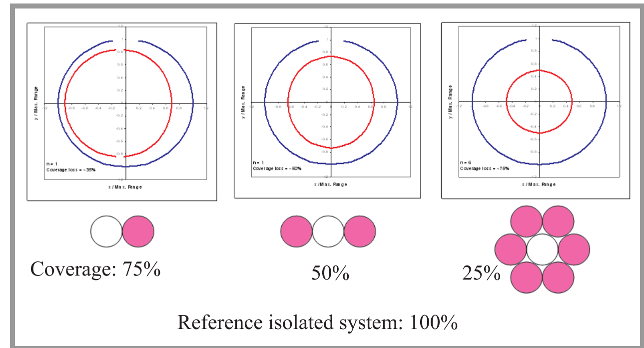


Fig. 2. Station coverage depends on signal environment.

losses, separation distances must be increased. The example shows that the station coverage cannot be determined confidently once for ever. It depends on, and varies with, the signal environment that changes unpredictably. If the station is a part of network, this remark applies to the whole network.

4.7. Filling Holes

Digital signals appear in the time domain as disjoint forms: pulses, or groups of pulses (packets), separated by empty spaces. Time-division multiplex access (TDMA) uses these spaces to transfer two or more bit streams in the same (shared) frequency band. Band-limited signals are similarly represented by disjoint figures in the frequency domain. Empty spaces (*spectrum holes*) can be used to allow multiple users to transmit at the same (shared) time period. Frequency-division multiple access (FDMA) systems base on that principle. The same principle applies to systems that use directional antennas to fill in geographical coverage holes.

4.8. Packing Solids

To avoid harmful interactions, the signal solids must be separated by safe distances. On the other hand, they should be packed tightly together, to leave more space to other signals. Keeping the solids closely, but not too close, requires the signal sizes, forms, and positions to be precisely matched against each other (except for randomized systems). A common reference frame, such as the frequency and time standards, synchronization and harmonized access coordination schemes are crucial here.

4.9. Collaboration

In our hyperspace convention, the engineering task consists in the best arrangement of solids in a hyperspace. That is not possible without collaboration. Activities of ITU and other organizations have served that purpose since many years. Special techniques have been developed to optimize the frequency planning, coverage coordination, and communication protocols.

In modern systems, many of the common ITU collaborative coexistence rules are embedded in the equipment hardware and software. Neighbor-friendly medium access protocols, such as the scheduled access, access on-demand, or self-adaptive access are examples.

The adaptive approach implies a real-time monitoring of the signal environment. For instance, the carrier-sense multiple access (CSMA) systems verify the absence of other signals before transmitting. If a signal is sensed, the transmitter waits for the transmission in progress to finish before initiating its own transmission. Such a friendly approach is not new. It has been invented many years ago with primitive technology, in not-for-profit applications and passed a practical test of life in the radio amateur services.

4.10. More Variables

The signal frequency is one of physical variables, or dimensions, that decide on how the radio medium is used. It decides on propagation effects, on antenna performances and on the system cost, to a large degree. For that reason it first focused engineering efforts since the very beginning. The vital role of time became clear later, after the digital technology was developed. The number of signal dimensions is not fixed in advance for any specific radio system, as it depends on the technology applied by the system. Any physical variable by which radio signals can be distinguished one from another, can be exploited. The more dimensions used, the more degrees of freedom in the radio spectrum utilization, with potentially significant impact on the system costs. Future systems are expected to make wider use of greater number of these variables, which creates new research challenges.

4.11. Optimization

The geometrical abstractions proposed here may help in designing improved use of RF resources by future systems. The problem is related to mathematical optimization problems such as the tessellation, packing, or knapsack problem. A two-dimensional tessellation is a collection of plane figures that fills the plane with no overlaps and no gaps. In a packing problem, given objects are to be packed tightly into a container with minimal gaps.

In a knapsack problem, a collection of objects is to be selected from a given set in such a way that their total value is as large as possible. Also the development of the communication robots and intelligent networks is a challenge, where the mathematical game theory can be helpful. Appropriate mathematical models and optimization tools can facilitate the engineering task enormously.

4.12. Communication Robots

Signal sensing mentioned earlier is only a step towards flexible systems that adapt themselves to the changing environment and a number of concepts have been proposed. The software-defined radio, agile-radio, policy-defined-radio, dynamic spectrum access, and cognitive radios are examples [6], [7]. They monitor the environment (individually or in a group), sense the spectrum holes and estimate the channel-state. They use predictive modeling and interference threat analysis and select the best signal parameters, providing dynamic and fair spectrum sharing with other radios. They can also follow the local regulatory restrictions, if these and location information is stored in their memories.

This leads us to the concept of *intelligent communication robots* that work together and negotiate to assure the best possible use of the radio frequency spectrum according to given criteria. A further step is a network of such robots. Future self-organizing and self-learning networks, to which today's wireless ad hoc mesh networks evolve, have capacity to overpass all what we imagine today. This is possible thanks to *swarm intelligence*, based on the collective actions of the component systems. The individual robots will collaborate and exchange real-time information about the traffic and radio environment, like ants or bees, much better and much faster than spectrum mangers can. Signals will flow in three planes: data plane, control plane and knowledge plane.

4.13. Science and Technology

Today radio communication technologies base on the James Clerk Maxwell's (1831–1879) theory. However, physics has progressed enormously since the 19th century. Max Planck (1858–1947) introduced quantum theory. Albert Einstein (1879–1955) replaced the traditional notion of space, time, energy and matter by the interchangeability principle of matter and energy in four-dimensional space-time.

New ideas continue to appear changing our understanding of the world, as, for instance, the concept of the dark matter, or string theory. According to the theory, the electron is no more a material particle or wave function as we were taught, but consists of oscillating "strings". And string theory requires the universe of ten dimensions or eleven, and not four, as Einstein told us [8]. To derive tangible results from that progress, a number of practical application-oriented research programs have started, among which the quest for quantum computer is probably most popular.

5. Concluding Remarks

"There is no more spectrum available" – declared Herbert Hoover, then the US Secretary of Commerce, in 1925. As a matter of fact, it never was sufficient. At the Hoover's times, the new radio regulations were expected to solve the spectrum scarcity problem. Later, spectrum engineering was expected to do the same [9]. Today, some economists propose market forces as universal medicine to solve the spectrum scarcity problem. However, the scarcity relates to physical processes that do not change when the spectrum is public, or is privately owned. Economic mechanisms (and administrative rules) deal only with the way the RF spectrum is accessed, i.e., who has the right to use it and for what purpose. How efficient is that use decides the technology applied.

Radio technologies in use are determined by the status of science, by investment opportunities and business inertia, and by the balance between competing interests. The society is composed of various groups, each with its own world-views, deeply rooted in the past experience. Their interests and goals are often conflicting. What is the best for one group is not necessarily good for the others. Those, whose needs have been satisfied, are against any change that would threaten their acquired benefits. Newcomers, with no access to the radio spectrum, press for changes. Engaged are systems of values and preferences and there are no universally accepted criteria in dealing with such issues.

We have discussed here some technological improvements possible in the use of RF spectrum. In spite of all their advantages, the new technologies have a little chance for quick and universal introduction. Their high costs and enormous investments in old technologies, which still work and bring profits, are the main obstacles. For some time to come, we will thus continue to live with the present regulatory arrangements, most likely mixed with other approaches.

In spite of the spectrum shortage claimed since the beginning of radio, radio applications have developed enormously. It has been possible thanks to the progress in science and technology that did not say the last word yet. Not so long ago, in 2007, the first nano-radio has been built [10]. It is a single carbon nanotube, one ten-thousandth the diameter of a human hair that requires only a battery and earphones to tune in to a radio station. To what degree will quantum physics and nanotechnology impact the use of the RF spectrum resources? Nobody can answer such questions today. With the future, the only sure thing is that it will surprise us…

References

[1] R. Strużak, "Improved utilization of the radio spectrum respecting physical laws", in *Proc. XXIXth URSI Gener. Assem. Conf.*, Chicago, USA, 2008, Sess. E05.

[2] J. A. Stine and D. L. Portigal, "Spectrum 101 – An Introduction to Spectrum Management", MITRE Tech. Rep. MTR 04W0000048, March 2004.

[3] R. Strużak, *Introduction to International Radio Regulations*. Trieste: ICTP, 2003; also R. Strużak, "Introduction to spectrum management", in *Methods and Algorithms for Radio Channel Assignment*, R. Leese and S. Hurley, Eds. Oxford: Oxford University Press, 2002, pp. 7–21.

[4] R. J. Matheson, "Flexible spectrum use rights", in *Proc. Symp. Adv. Radio Technol. ISART 2005*, Boulder, USA, 2005, pp. 21–36.

[5] R. Strużak, "Flexible spectrum use and laws of physics", in *Proc. ITU Worksh. Market Mech. Spect. Manag.*, Geneva, Switzerland, 2007.

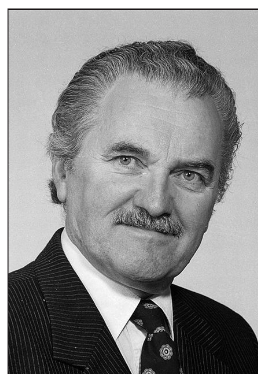
[6] O. Ileri and N. B. Mandayam, "Dynamic spectrum access models: toward an engineering perspective in the spectrum debate", *IEEE Commun. Mag.*, vol. 46, no. 1, pp. 153–160, 2008.

[7] S. Haykin, "Cognitive radio: brain-empowered wireless communications", *IEEE J. Select. Areas Commun.*, vol. 23, no. 2, pp. 201–220, 2005.

[8] B. Greene, *The Elegant Universe. Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*. New York: Norton, 2003.

[9] *JITAC: Spectrum Engineering – the Key to Progress*, New York: IEEE, 1968.

[10] R. Sanders, "Single nanotube makes world's smallest radio", UC Media Rel., Oct. 2007 [Online]. Available (accessed: 29.04.2008): http://www.berkeley.edu/news/media/releases/-2007/10/31_Nano-Radio.shtml



Ryszard Strużak is a Full Professor at the National Institute of Telecommunications (NIT), Poland, and a Co-Director of ICTP Schools on Wireless Networking, Italy. He is the author or co-author of some 200 papers as well as 10 patents in the areas of radio communications, spectrum management, and electromagnetic compatibility. He was a consultant to ITU, IUCAF, PWC, UNOCHA, WB and to industry and governmental agencies in Poland, Switzerland, Turkey, United Kingdom, the United States, and other countries. He served as the Head of Technical Department and Acting Assistant Director at CCIR-ITU, a Visiting Professor at Institut National Polytechnique de Toulouse, a Full Professor at the University of Information Technology and Management at Rzeszów, a Professor a Wrocław University of Technology, a Professor and the Head of NIT Wrocław Branch, the Editor-in-Chief and the Chair of Editorial Board, Global Communications. He co-founded and chaired International Wroclaw Symposium on EMC. He was active in international organizations: ITU, URSI, IEC and CISPR, where he was elected to leading positions, among others the Vice Chairman of ITU Radio Regulations Board. He was honored by the ITU Silver Medal, two international awards, and by highest national awards and decorations. He was elected a member of New York Academy of Science and an Academician of the International Telecommunication Academy. He is a Life Fellow of IEEE and was awarded by an IEEE EMCS Acknowledgment of Gratitude.

e-mail: r.struzak@ieee.org

National Institute of Telecommunications Wrocław Branch
Swojczycka st 38
51-501 Wrocław, Poland

Paper

Adaptive Resource Management and Flexible Radios for WiMAX

José Salazar, Ismael Gómez, and Antoni Gelonch

Abstract—The availability of dynamic resource management will be crucial for the deployment of future wireless systems characterized by high data rate services with rigid quality of service demands. Flexible radios appear as the technological answer required to achieve constraint goals under different channel conditions and transmission scenarios. This paper is focused on enhancing another step of flexibility within the resource management by including an efficient handling of computing resources. This concept towards flexible architectures represents a key word for a real successful implementation due to the relationship between the radio applications, which face the scarcity of resources within a heterogeneous environment, and the processing power needed to execute them.

Keywords—computing resource management, radio resource management, reconfigurability, software radio platforms.

1. Introduction

Some important threads for new wireless technologies such as long term evolution (LTE) and worldwide interoperability for microwave access (WiMAX) mobile (802.16e standard) are high demands of flexibility, and reliability. These future mobile networks are defined by several radio access technologies (RAT's) with cells in different hierarchical levels and frequencies offering the same or nearly the same services [1]. For multistandard/multimode terminals the preference of one standard over another will be defined by different criteria. For example required quality of service (QoS), link conditions, or network traffic; but could also be dictated taking into account the terminal capabilities such as battery life and energy consumption.

Concerning the deployment of radio interfaces of both terminals and base stations becomes a hazardous task to obtain solutions where are considered all the possible scenarios, and parameters. Different scenarios for instance, imply that the assignment of services and its quality restrictions will depend on the user preferences and capabilities, or could take part as a seamless roaming depending on the area of location. However, a reconfiguration process is inherited and this will be dictated by a set of parameter selection towards an optimization target.

Software radio (SR) exploits reconfigurability and adaptive parameterization to achieve flexibility [2], [3]. Adaptive parameterization means that the signal processing structure may be switched by parameters to realize different standards. Nevertheless, the reconfigurable capacity of a SR platform rely on the processing computational resources consumed by the standards is able to execute [4]. Flexible radio systems which can be able to dynamically establish an optimal management of resources of SR platforms will play a main role. The idea behind this work is a framework which brings together a dynamic awareness and control of the resources involved in the reconfiguration process. The goal in this paper is to test a fair usage of resources across radio and hardware environments, focused on trading efficient computational and radio resource assignation but providing constant QoS.

The rest of this paper is organized as follows: Section 2 deals with the concepts of advanced resource management and orthogonal frequency division multiplexing (OFDM) systems. Section 3 describes a framework for flexible radio platform. Section 4 describes a case of study defining system specifications adopted for simulations. Follows simulation results and discussion in Section 5. Finally, Section 6 concludes the paper.

2. Advanced Resource Management and OFDM

2.1. Radio Resource Management

The scenario in modern mobile networks introduce new tasks within the radio resource management (RRM) problematic, an efficient use of radio resources is determined by the general characteristics of the networks their selves and by a common management within the whole system. Reconfigurable baseband RF/IF platforms include the use of advance media access control (MAC) and RRM functionalities. Such solutions incorporate suitable control of the reconfiguration process necessary to assure the selection of the RAT which can offer the desired service [1], they manage dynamically the allocation and de-allocation of radio resources (e.g., time slots, codes, frequency carriers, etc.) [5]. On the other hand in order to achieve vertical and horizontal handovers, radios with cognitive centric properties, such as spectrum monitoring, localization, tracking, and others features [1], appear to handle necessary cross-layer interactions leading to increase multiuser capacities.

2.2. Computing Resource Management

Hardware platforms on the mobile devices will impose limits on processing operation even under optimization processes of cognitive algorithms. Therefore for a more sustainable approach another feature must be added: the awareness of the processing capabilities, this is a real way

towards flexibility and granularity. The management of computing resources refers to the hardware that implements the (software-defined) signal processing chains for radio communications.

2.3. Adaptive Modulation in OFDM-Based Systems

Orthogonal frequency division multiplexing systems are conceived for data transmission in frequency selective channels, and have as mathematical background the fast Fourier transform (FFT). OFDM advantages are spectral efficiency, ability to fight multipath distortion and delay spread, and resilience to intersymbol interference (ISI). Adaptive modulation in OFDM entails some decision taking about the service data rate, bandwidth of transmission, by selecting some transmission mode. It presents several advantages compared to techniques that use a unique carrier signal. Based on the channel prediction the transmitter selects the appropriate modulation schemes for next time slot. The constant throughput of an adaptive OFDM scheme exploits the frequency selectivity of the channel while offering a constant throughput [6]. Higher order modulation schemes, which are spatially more efficient such as quadrature amplitude modulation (QAM), need better channel conditions that the lower order modulation schemes accordingly to a bit error rate (BER) criterion.

2.4. The 802.16e Standard (WiMAX Mobile)

The IEEE 802.16e standard release defines the WiMAX mobile system. WiMAX is short of worldwide interoperability for microwave access officially known as wireless metropolitan area network (MAN) [7], [8]. Quality of service in WiMAX is delivered using adaptive modulation and variable forward error correction (FEC) encoding per RF burst.

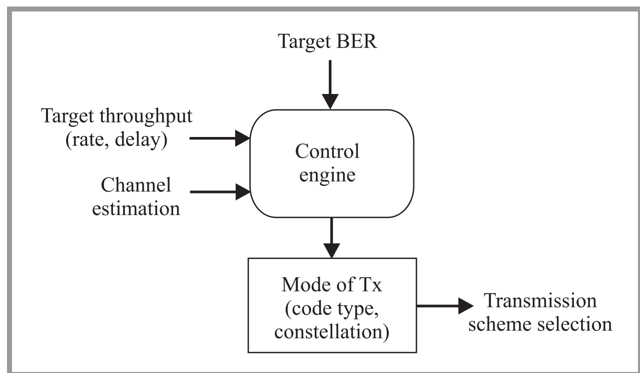


Fig. 1. Algorithm for adaptive modulation in WiMAX.

The transmission modes are defined by the MAC layer in order to adjust the user quality of service parameters depending on the channel characteristics, such as the delay spread, the fading, the user speed, etc. This ensures a robust link while maximizing the number of bits/second for each user unit, i.e., enables an optimal average data rate, improving system capacity and also allow being able to deal with service level agreements [9]. Switching levels for modulations schemes are set to achieve some BER performance, so that if the estimated channel condition is within the stored switching level for the modulation scheme under operation it maintain the same modulation format [10], [11].
The algorithm is shown in Fig. 1. Evaluates the channel, the required throughput to achieve the target BER, and switch the transmission mode, if the channel conditions (signal-to-noise ratio – SNR) are below some threshold, then renegotiation of target QoS takes place.

3. Integrated Resource Management Framework

Adaptive modulation for WiMAX accomplish to have self awareness deciding about the data rate, the transmission mode, and therefore about the bandwidth of the transmission. Nevertheless, some selections might be achieved to support the user's required QoS, or the network load, at the expenses of rising computational load and inefficient energy consumption. For example, let suppose a scenario where a user of a certain service will require certain BER $> 10^{-4}$ then considering the link conditions, a more efficient spectral scheme shall be selected such 64 QAM; nevertheless such terminal has limited computing resources, or less battery life in order to execute the signal processing tasks, so that the system capacity will be affected. Conservative approaches, tending to reduce sustainable rate through negotiation, or even not taking into account QoS constraints, will maximize radio capacity at the expenses of computing costs.
Trade-offs between capacity, robustness and computing capabilities might be accomplished in real time. The concept of flexible radios introduces automated radio, com-

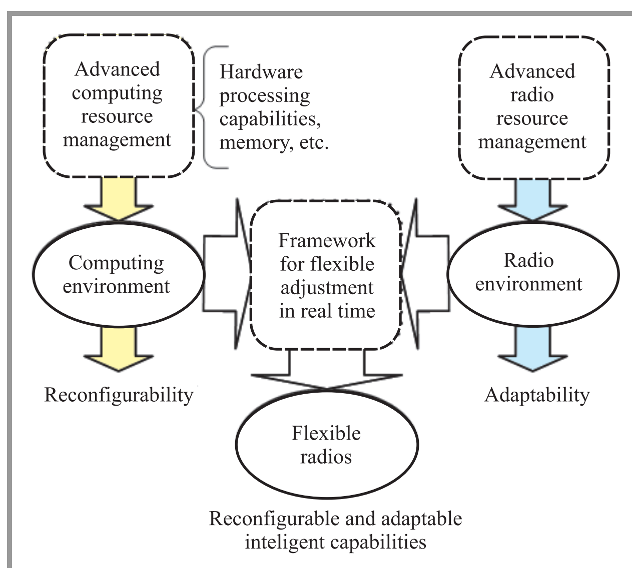


Fig. 2. Flexible radio framework.

puting and energy resource management of the SR equipment. Flexible radios ease an exchange of radio for computing resources and vice versa [4]. Parameterization might be achieved by learning standards and its features, i.e., defining the similarities and dissimilarities among their signal processing chains [3], thus aassembling platform-specific, defining the parameters dependencies of the different modular blocks of a communications standard chain, including the computational costs (million instructions per second/million operations per second (MIPS/MOPS) counts).

In the framework depicted in Fig. 2, an engine based on dynamic wave-form design, continuously seek the best set of resources available, taking decisions for parameter optimization maintaining QoS, based on radio, and computing environments monitoring and even future predictions of behaviors. This dynamic wave-form-design varies from the adaptive modulation process described before because the parameter changing depends on feedback information searching a fair distribution on resources utilization across radio, hardware and application environments.

4. Case Study

4.1. The IEEE 802.16e System Model

As an OFDM system, the high data rate sequence of symbols is divided into multiple parallel low data-rate sequences, each of is used to modulate the orthogonal subcarrier. The transmitted baseband signal can be represented as

$$x(t) = \sum_{i=1}^{L-1} s[i] \cdot \mathrm{e}^{-\mathrm{j}2\pi(\Delta f + iB_c)t}, \ 0 \le t \le T,$$

where: $s[i]$ is transmitted symbol of the ith subcarrier, L – total number of subcarriers, B_c – subcarrier bandwidth, Δf – frequency offset, T – useful symbol duration (without cyclix prefix).

Table 1

Parameters for OFDM symbols

Parameter	Value	Definition
B	1.25, 1.75, 5, 10, 15	Channel bandwidth (adjustable) [MHz]
L	128, 512, 1024, 2048	Total number of subcarriers, including the user subcarriers N_S, guard subcarriers N_G and pilot subcarriers N_P, normally is equal to the FFT length (N_{FFT})
n	28/25	Oversampling factor
G	1/4, 1/8, 1/16, 1/32	Ratio of cyclic prefix time to useful time

At the receiver, the symbol $s[i]$ is calculated integrating the received signal with a complex conjugate of the corresponding subcarrier:

$$\widehat{s}[i] = \int_0^T x(t) \cdot \mathrm{e}^{\mathrm{j}2\pi(\Delta f + iB_c)t} \mathrm{d}t.$$

As it was described in Section 2, WiMAX physical layer allows making an optimum choice of various parameters, such as prefix cycle length, number of subcarriers, subcarrier spacing, preamble interval, in order to achieve some performance goal. The respective values defined for the primitives of an OFDM symbol are condensed in Table 1 [6].

Based on these parameters the time duration of an OFDM symbol T_s and thus the corresponding data rate R can be acquired as it follows:

$$T_s = (1/\Delta f)(1+G),$$

$$T_s = [1/(f_s/N_{FFT})]\,(1+G),$$

$$T_s = [1/(nB/N_{FFT})]\,(1+G),$$

$$R = \frac{N_s N_b C_{rate}}{\frac{N_{FFT}}{nB}(1+G)},$$

where N_b, C_{rate} are the transmission parameters; C_{rate} is coding rate used, N_b is number of bits, depending on the constellation used.

4.2. Performance of the IEEE 802.16e Physical Layer

WiMAX physical layer advantages depend significantly from the timing, and frequency synchronization at the receiver side which is achieved by the using of pilots and the cyclic prefix insertion [6], [12]. BER coding and modulation schemes listed in Table 2 have been evaluated using Monte-Carlo simulations developed in MATLAB.

Table 2

Simulated coded and modulation schemes

Scheme	Modulation	RS code	Conv. code
1	QPSK	(32, 24, 4)	2/3
2	QPSK	(40, 36, 2)	5/6
3	16 QAM	(64, 48, 8)	2/3
4	16 QAM	(80, 72, 4)	5/6
5	64 QAM	(108, 96, 6)	3/4
6	64 QAM	(120, 108, 46)	5/6

4.3. WiMAX Computational Complexity

Another advantage of OFDM to single-carrier modulation schemes is that it requires lower computing costs for higher data rate links.

Typical equalizers complexity is proportional to the product of the bandwidth – delay spread product: $B \cdot T_m$, thus

the number of taps required to achieve ISI free communication is given by $\xi = B \cdot T_m$.
This means that such equalizer of ξ taps have to complete ξ number of multiply-accumulate operations per symbol (MACs). The processing costs grow as the square of the data, and the total complexity is of the order of:

$$O(B^2 T_m).$$

Now focusing on an OFDM-based system susch as WiMAX, one of the most demanding tasks to execute in terms of processing is the fast Fourier transform algorithm, which encloses computing costs of the order:

$$O(N \log_2 N)$$

thus considering that B/N_{FFT} is number of symbols per second, $1/T_m$ is complexity order per symbol.
WiMAX will present nearly linear behavior of computing complexity even under the demanding tasks of forward error correction blocks chosen and the highest data rates:

$$O(B \log_2 B \cdot T_m).$$

Taken into account a channel model, as in [13], [14] where the delay spread is $T_m = 0.202$, the order of complexity of transmission of the whole orthogonal frequency division multiple access (OFDMA) system will have a linear tendency for different bandwidths, as it shown in Fig. 3.

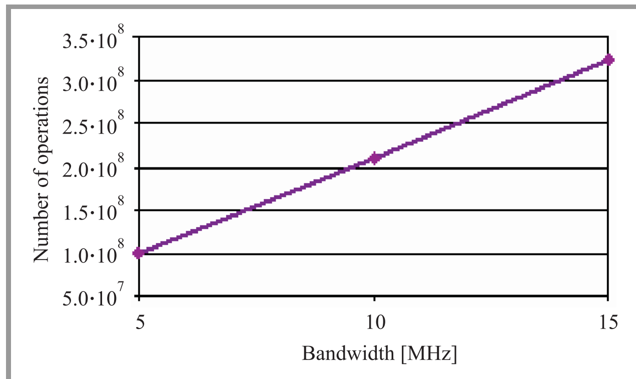


Fig. 3. Order of computational complexity for different transmission bandwidths.

This fact is also true when some FEC as turbo decoder, block turbo codes are used at the receiver side, there are of course variations concerning the system performance, and the computing costs are raised, but the behavior still linear.

4.4. Simulated Processing Complexity

Execution times of the processing blocks on the whole chain were obtained for the different modulation schemes and the code rates selected at every time slot considering downlink transmission and reception. The computational costs in terms of central processing unit (CPU) time rise depending on the number of bits corresponding to modulation scheme as consequence of the number of bits processed to maintain the throughput. The results are shown in Table 3.

Table 3
CPU requirements for different modulation-codes (MCs)

Scheme	CPU time [μs]	
	transmission	reception
1	527.97	1226.29
2	849.41	2349.33
3	1130.79	3792.67
4	1214.68	3999.96
5	1225.93	4038.19
6	1765.09	6594.06

The characteristics of the processor used for the simulations are revealed in Table 4.

Table 4
CPU specifications

Characteristics of the processor	
Processor class	AMDS Athlon 64 X2
Processor speed	3.2 GHz
Processor cores	2
Bus speed	1000 MHz
Performance/speed	19376 MOPS @ 3.2 GHz
Speed efficiency	6.05 MOPS/MHz
Perfomrance/power	19376 MOPS @ 125 W
Power efficiency	155 MOPS/W
Bytes/cycle	15.7
Average latency	17 ns

4.5. Scenarios for Case Study Simulation

The following scenario is defined:

- Downlink transmission.
- Analysis for bit error rate services demand (BER target) of 10^{-4}.
- Path delays, and power of channel taps have been selected according to SUI channel 2 [14].
- The bit error rate performance of the OFDMA system is depicted in Fig. 4. Is important to notice that on the transmitter side, turbo codes are selected together with Reed-Solomon decoder to achieve the quality of service demands. This turbo decoder has a block length of 1024, and it use a stopping criteria. It halts to iterate when the BER target is reached.

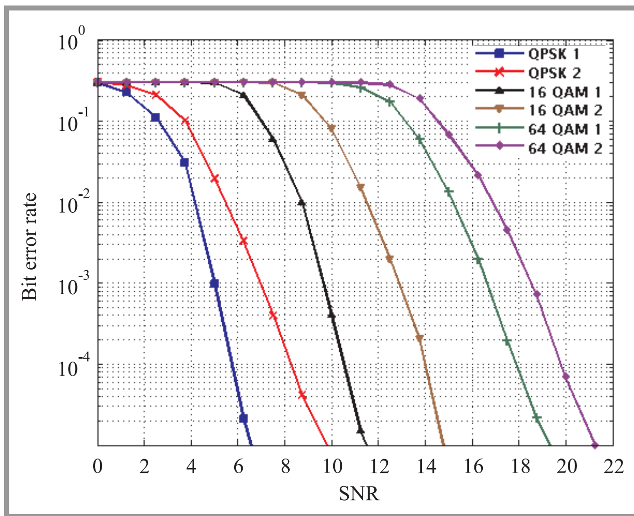


Fig. 4. BER performance of OFDMA system.

Therefore appear two cases that traditional approaches can deploy in time:

- The base station (BS) will choose to optimize its spectral efficiency. The normalized spectral efficiency for the different transmission modes is enclosed in Table 5. The BS will choose for transmission the most efficient modulation-coding scheme for all the users that shall be within its range, in this case 64 QAM 3/4. From the terminal perspective this will be the worst case because of the demanding processing requirements.

- The second strategy is to go towards the other extreme condition: the BS decides that the most suitable transmission mode is the one with the optimal consumption of computing resources. So it selects quadrature phase shift keying (QPSK) as transmission mode for all users to whose is giving service. The minimum spectral efficiency at the base station results in an optimal use of processing resources. There is a difference of 4.5 times between the maximum of computing costs in the strategy 1 (172576.30 MOPS) to the strategy 2 (38351.30 MOPS).

Table 5
Spectral efficiency for different modulation-code schemes

Modulation	QPSK		16 QAM		64 QAM	
Code rate	1/2	3/4	1/1	3/4	2/3	3/4
Spectral efficiency [%]	6.25	9.38	12.5	18.75	25	28.13

In Fig. 5 are illustrated the computational complexity for bandwidths 5 MHz and 10 MHz.
If the hardware platform constraints for the user terminals are known in terms of the maximum processing operations that the CPU can accomplish at the time when the application has to be executed, then some cost function can be defined to achieve the desired goal for a more balanced strategy of decision. Let suppose that the turbo decoder now won't have any halt condition based on the instantaneous BER, but it stops when some maximum number

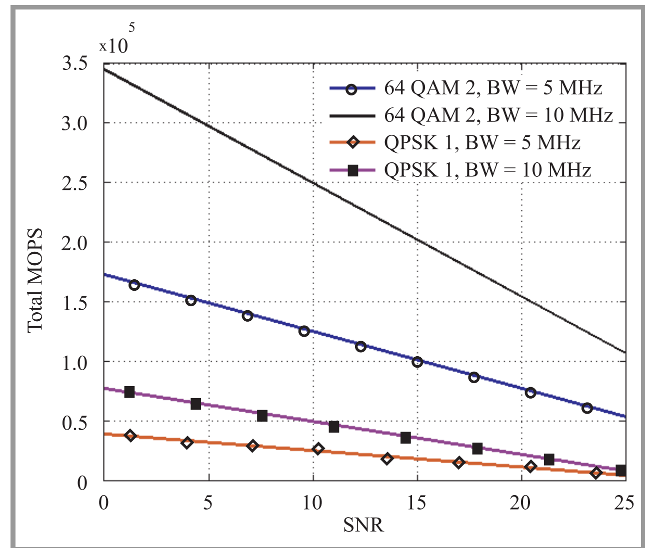


Fig. 5. Computational complexity for the two strategies.

of iterations (NOI) is reached. We take into account two cases: NOI = 20 and NOI = 5 (Table 6).

Table 6
Consumed MOPS for a transmission bandwidth $B = 10$ MHz

Max NOI	QPSK 1/2	64 QAM 3/4
5	19837.9	89223.07
20	345165.472	76713.988

Is important to notice that a terminal will force to achieve the required quality of service constraint (BER target) at the expenses of these costs, but is clear that some terminals can reach this goal taking into account a better spectral efficiency or an improvement in terms of computational complexity and energy consumption.

4.6. WiMAX Waveform Adaptive Function

The next step is to introduce a version of a real adaptive function which guarantees a desired QoS for some channel realization. The idea is to maintain a sustainable BER target according to the channel conditions defined by an SNR level. It will tend to choose suitable code/modulation mapping scheme taking into account the availability of computing resources for the user and spectral efficiency for the base station. It will obtain the best conditions for the requested QoS, and thus can either transmit with a minimum of computing costs, or negotiate for a lower QoS when channel conditions or the computing costs do not allow that transmission mode. This process is shown in Fig. 6.

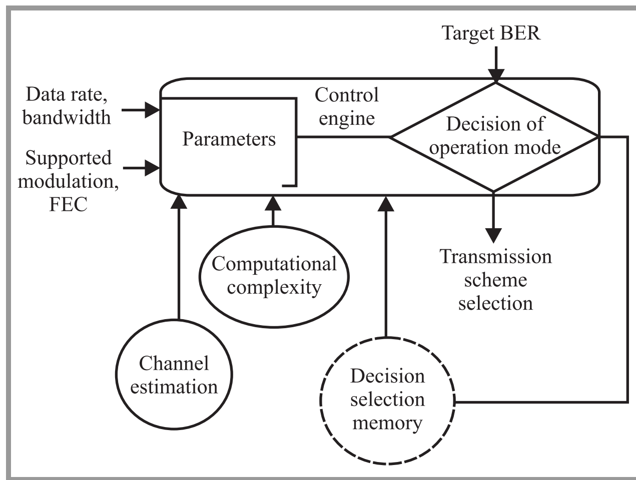


Fig. 6. Algorithm for WiMAX wave-form design.

In order to accomplish this task, we define the following cost function:

$$Cost = 0.7S_{eff} + 0.3C_{eff},$$

where S_{eff} is spectral efficiency, C_{eff} is computational efficiency.

5. Results

On our final analysis we suppose that all user terminals in the whole system are equidistant to the base station. Their computing capabilities are defined so its receiver can deploy either 5 or 20 iterations. The bandwidth of transmission can be chosen between 5 and 10 MHz. Several combinations of parameters can be achieved depending on the resources availability for each user device.

The algorithm use the cost function defined in last section, to set the number of user terminals to whose will assign one (constellation code throughput) triplet and thus certain bandwidth, and certain data rate. The different modulation code schemes are assigned to the two types of terminals that execute 5 or 20 iterations in the FEC in order to achieve

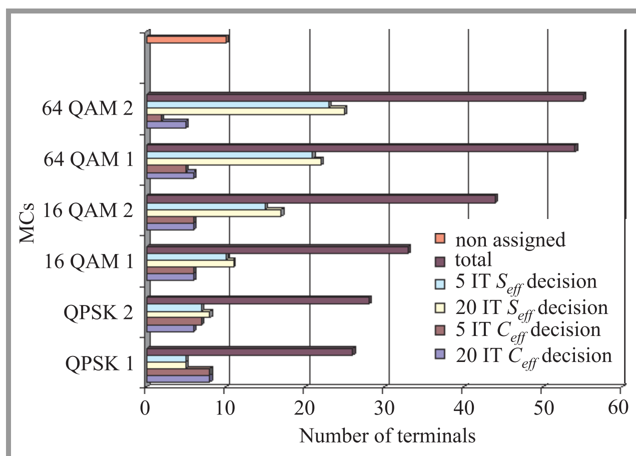


Fig. 7. Modulation-coding schemes assigned to the terminals (BER 10^{-4}) – scenario 1.

the quality of service constraint that was defined. Therefore a suitable allocation is deployed within the system. Results are depicted in Fig. 7.

In order to validate this results, we define a second cost function which to the contrary set more specific weight on the computing resources than to the spectral efficiency:

$$Cost = 0.3S_{eff} + 0.7C_{eff}.$$

The results appear in Fig. 8.

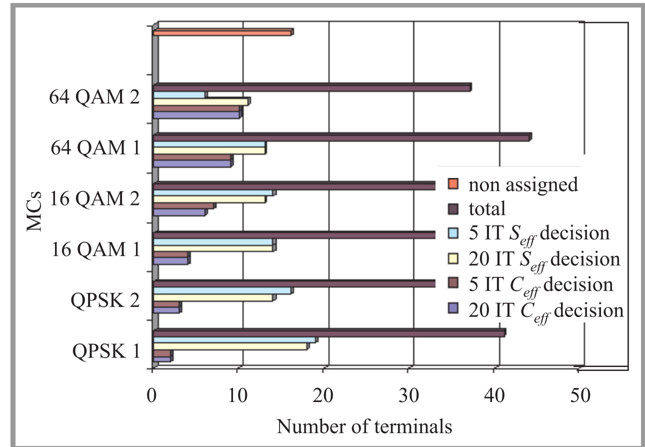


Fig. 8. Modulation-coding schemes assigned to the terminals (BER 10^{-4}) – scenario 2.

Comparing both figures is clear that for some users there will be no transmission settings, because either from the spectral or the computing resources point of view, there are not enough resources available for its service. Then the weight of the cost function defines the tendency of resources and thus the number of terminals to which is assigned each of modulation-coding schemes (Table 7). Better spectral-efficient MC's are assigned mostly for the terminals in Fig. 7, while in Fig. 8, is the contrary case.

Table 7
Terminals scenario

Number of terminals	B [MHz]	Max NOI permited
250	5	20
	10	5

6. Conclusions

In this paper has been illustrated that in order to achieve the necessary level of reliability, granularity and flexibility defined for future radio terminals, it is of major importance to include terminal capacity in terms of its processing power and power consumption. It also has been exposed how is possible to asses' tradeoffs between QoS parameters and computing costs dynamically. Thus, it can be consider that the engine can be updated with tracked information every time slot (in ms).

Several research challenges remain to be overcome before a truly practical system might be deployed. In spite of the technical challenge, it is clear that flexible radios offer hope to meet the resources demands in terms of spectrum, robustness and computing and of new wireless communications.
It is clear that optimization can not be done in devices (only event detection). Engines seeking the optimum set of resources available, must take into account the computing capabilities having its major effect in case when the services are latency limited and a dynamical seamless switch between RAT's (for example UMTS/WiMAX) is conceived.

Acknowledgements

This work was supported by the Spanish National Science Council CYCIT under Grant TEC2006-09109, which is partially financed from the European Community through the FEDER program.

References

[1] R. Ferrús, A. Gelonch, O. Sallent, J. Pérez-Romero, N. Nafisi, and M. Dohler, "A feasible approach for E2E QoS management in coordinated heterogeneous radio access networks", in *Proc. IEEE Int. Perf. Comp. Commun. Conf.*, Phoenix, USA, 2005.

[2] J. Mitola, "The software radio architecture", *IEEE Commun. Mag.*, vol. 33, no. 5, pp. 26–28, 1995.

[3] F. K. Jondral, "Software-defined radio – basics and evolution to cognitive radio", *EURASIP J. Wirel. Commun. Netw.*, no. 3, pp. 275–283, 2005.

[4] V. Marojevic, J. Salazar, I. Gomez, and A. Gelonch, "Computing system modeling for flexible radios", in *Proc. Karlsruhe Worksh. Softw. Radio WSR08*, Karlsruhe, Germany, 2008.

[5] J. Mitola and G. Maguire, "Cognitive radio, making software radios more personal", *IEEE Pers. Commun. Mag.*, vol. 6, no. 4, pp. 13–18, 1999.

[6] J. Andrews, A. Gosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking.* Englewood Clifs: Prentice Hall, 2007.

[7] "IEEE Standard for Local and Metropolitan Area Networks" – Part 8: "PHY", IEEE Standard 802.16-2004, http://ieee802.org/16

[8] "IEEE Standard for Local and Metropolitan Area Networks. Admendment 2 and Corregendum 1 to IEEE Std 802.16-2004", IEEE Standard 802.16e-2005, http://ieee802.org/16

[9] J. Escoda *et al.*, "QoS and computational cost in wireless WiMAX". Escola Superior de Castelldefels, Universitat Politècnica de Catalunya, Spain, June 2008.

[10] A. J. Goldsmith and S.-G. Chua, "Variable-rate variablepower MQAM for fading channels", *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, 1997.

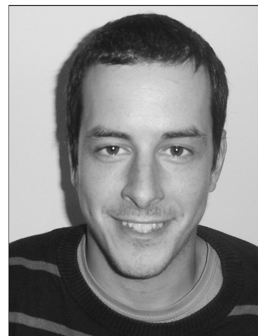
[11] T. Keller and L. Hanzo, "Adaptive modulation techniques for duplex OFDM transmission", *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1893–1906, 2000.

[12] WiMAX Forum, "Mobile WiMAX" – Part I: "A Technical Overview and Performance Evaluation", 2005, http://www.wimaxforum.org/home/

[13] M. Komara, "SDR architecture ideally suited for evolving 802.16 WiMAX standards", Airnet Communications Corporation Press, Melbourne, USA, 2006.

[14] D. S. Baum, "Simulation the SUI channel models", IEEE 802.16 Broadband Wireless Access Working Group, 2004.

José Salazar received the degree in electrical engineering from the Universitat Autonoma Metropolitana, Mexico, in 2000, and the M.Sc. in electrical engineering from the Dresden University of Technology, Germany, in 2003. His preferred area lays in mobile communications. On the basis of his activities and experiences at the university and the industry is fully interested to play a part in the development of the 4th generation of mobile communication systems. Since 2004 join the Group Radio Communications at the Universitat Politècnica de Catalunya, Spain, where is doing a Ph.D. in the field of resource management combined with software defined radio.
e-mail: jose.salazar@tsc.upc.edu
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
C/Jordi Girona 1–3, 08034 Barcelona, Spain



Ismael Gómez received the engineer of telecommunication degree from the Universitat Politècnica de Catalunya, Spain, in 2008, where he is working towards his Ph.D. in the area of digital communications, supported by a grant from the Spanish government. His main research topic is on the design of execution environments for high performance reconfigurable computers for signal processing applications.
e-mail: ismael.gomez@tsc.upc.edu
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
C/Jordi Girona 1–3, 08034 Barcelona, Spain

Antoni Gelonch is Associate Professor at the Universitat Politècnica de Catalunya, Spain. His research activities are in the mobile communications framework. His efforts have been focused in implementation methodologies to incorporate software radio concept as a mechanism to support cognitive architectures, strategies and algorithms for an efficient and flexible resource management. He has contributed to several papers in international journals and conferences and he has participated in research projects funded by public and private organizations.
e-mail: antoni@tsc.upc.edu
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
C/Jordi Girona 1–3, 08034 Barcelona, Spain

Paper

Mathematical Foundations of Cognitive Radios

Romain Couillet and Mérouane Debbah

Abstract—Recently, much interest has been directed towards software defined radios and embedded intelligence in telecommunication devices. However, no fundamental basis for cognitive radios has ever been proposed. In this paper, we introduce a fundamental vision of cognitive radios from a physical layer viewpoint. Specifically, our motivation in this work is to embed human-like intelligence in mobile wireless devices, following the three century-old work on Bayesian probability theory, the maximum entropy principle and minimal probability update. This allows us to partially answer such questions as, what are the signal detection capabilities of a wireless device, when facing a situation in which most parameters are missing, how to react and so on. As an introductory example, we will present previous works from the same authors following the cognitive framework, and especially the multi-antenna channel modeling and signal sensing.

Keywords— Bayesian inference, cognitive radio, maximum entropy.

1. Introduction

In 1948, Claude Shannon introduced a mathematical theory of communications [1], allowing two to three generations of research to design increasingly sophisticated telecommunication tools, whose purpose is to constantly increase the achievable transmission rate over various communication channels. One of the key conclusions of Shannon was to show that a linear increase in the transmission bandwidths is expected to provide linear growth in the channel transmission capacity, while linear transmit power increase only provide sublinear capacity growth.

As a consequence, the last decades of research in telecommunications led to a situation in which the available transmission bandwidth became dramatically scarce and can only be acquired by service providers at extraordinarily high prices. Then, in the end of the nineties, the conclusions of Foschini [2] and Telatar [3] on their work on multiple antenna (MIMO) systems came as a salvation: when increasing the number of embedded antennas in both transmit and receive devices, a potential linear growth (with the number of antennas) in capacity was expected. Since the exploitation of the space dimension can come virtually at a zero cost compared to the exploitation of the frequency dimension, these stunning results rapidly generated lots of research work in the early years of the twenty-first century.

However, practical applications of multiple antenna systems took a long time to be put in place, when it was clearly realized that the exceptional predicted capacity gain could only come at a very strong signal to noise ratio (SNR) and for low correlated channels; for instance, line of sight components in a transmission almost completely annihilates the gain of multiple antenna systems. However, up to this point in the evolution of wireless devices, the initial result from Shannon was still applicable to the most advanced technologies.

After the MIMO delusion, Joseph Mitola [4] realized that a new virtual dimension could be exploited to increase the achievable transmission rate: making the radio smarter. The basic insight of Mitola was to observe that most allocated bandwidth is not efficiently used in the sense that, most of the time, large pieces of bandwidth are left unoccupied. Emabling the wireless devices to sense the frequency spectrum in a decentralized manner[1] allows for a potentially high increase of *spectral efficiency*, which we define here as the actual averaged transmission rate over the theoretical capacity. These large-scope ideas from Mitola recently motivated a wide range of research with common denominator the introduction of intelligence in wireless devices. For instance, Haykin [5] introduces the concept of interference temperature, which allows to control the level of interference allowed in a network, i.e., if a given user has a rate constraint largely inferior to the effective channel capacity, the excess unused rate could be used by another device, as long as this device does not request more than the available excess rate. This interference temperature brought the new idea of primary and secondary users in a wireless network: primary users are those subscribers who are charged a high price to communicate with high quality of service, while secondary users pay a lower price to communicate over opportunistic excess rates left unused by the primary users, e.g., [6], [7].

However, all these ideas, revolutionary as they may seem, only scratch the surface of a larger entity that is the cognitive radio. Indeed, if the cognitive radio is defined, as was supposedly the prior idea of Mitola or even more certainly the basic view of Haykin[2], as a radio in which all entities are capable of cognition, then the limitations in the capabilities of these radios is still unknown and not really explored. Concrete works on smart devices date back to Shannon's time as well. Claude Shannon was already interested in ideas such as a robot capable of playing chess [8]; he provided an original viewpoint of the cognitive abilities of future computers back in 1953 [9] and even constructed a mind-reading machine, the circuitry of which is depicted in [10].

[1] So to limit the needs for control signaling.

[2] Remember that the title of his main contribution on cognitive radios [5] refers to "brain-empowered" radios.

In this work, we propose to define a fundamental basis for cognitive radios on a physical layer viewpoint, which enables human-like intelligence in wireless devices. This work comes as a rupture compared to previous telecommunication work, as we will no longer rely on Shannon's work, but will rather extend it. The reasons why we escape from Shannon's framework will be explained and justified in the following sections. The additional mathematical tools needed to extend Shannon's theory of information are the theory of Bayesian probabilities, the maximum entropy principle [11] and the minimal cross-entropy principle [12], [13], among others.

The remainder of this paper unfolds as follows: in Section 2, we present the key philosophical ideas which lead from Shannon's classical information theory to Jaynes' more general probability theory. In Section 3, we provide two examples of direct application of Jaynes' maximum entropy principle to the problems of channel modeling and signal sensing. Then in Section 4, we discuss the present advantages and limitations of cognitive radios, and provide our conclusions in Section 5.

Notation: In the following, boldface lower-case symbols represent vectors, capital boldface characters denote matrices ($\mathbf{I}_N$ is the $N \times N$ identity matrix). X_{ij} denotes the (i, j) entry of $\mathbf{X}$. The Hermitian transpose is denoted $(\cdot)^{\mathsf{H}}$. The operators tr$\mathbf{X}$ and $|\mathbf{X}|$ represent the trace and determinant, respectively. The symbol $\mathrm{E}[\cdot]$ denotes expectation. The operator $\mathrm{vec}(\cdot)$ turns a matrix $\mathbf{X}$ into a vector of the concatenated columns of $\mathbf{X}$. Finally, the notation $P_x(y)$ denotes the probability density function of the variable x in position $x = y$.

2. From Shannon to Jaynes

We will first present a simple example to show the inherent limitations of Shannon's theory of information.

2.1. Channel Capacity Revisited

Let us consider the simplest communication scheme, modeled as

$$y = x + n, \tag{1}$$

for some transmit signal x, additive background noise n and receive signal y. The Shannon capacity C of such a system reads

$$C = \sup_{p_x} I(x;y), \tag{2}$$

with p_x the probability distribution of the variable x taken in the set of single-variable probability distributions, and I denotes the mutual information [1]. The Eq. (2) can only be computed if the distribution of n is known. In practice, n is often taken as Gaussian, both for simplicity reasons and because this is somehow *often* close to the reality. However, there is no actual way to predict the distribution of the noise before transmitting data, and in reality the expression (2) is impossible to compute. This leads to the conclusion that all capacity computations are in fact only *approximations* of Eq. (2).

Moreover, it is important to observe that what we call noise is in effect the sum contribution of interfering waves with different properties. If part of this noise can be analyzed by the cognitive device[3], then the capacity will increase. All these primary observations lead to realize that the channel capacity is largely dependent on the prior information available at the receiver. In particular, two identical receivers, facing the same channel, may have different actual capacities depending on the individual channel state information.

Assuming the noise is known to be Gaussian with zero mean, the receiver is left to estimate the noise variance. In general, only approximative values of the SNR are available. Therefore, the channel capacity might be better seen as a *rate vector*, with entries indexed by every possible values of the SNR and taking different degrees of probability. These degrees of probability differ for each receiver, making the capacity again information-dependent and user-dependent. As a matter of fact, what one would call "real capacity", that would correspond to the capacity if the receiver knows exactly the noise variance, does not carry in itself any actual significance: as recalled by Jaynes [11; p. 634], the channel capacity is not an intrinsic value of the channel but an intrinsic value of the level of knowledge of the system designer[4].

2.2. Limitations of Information Theory

Already in 1963, Leon Brillouin [14] realized the fundamental limitation of Shannon's information theory. In his own words [14], "*The methods of [information] theory can be successfully applied to all technical problems concerning information: coding, telecommunication, mechanical computers, etc. In all of these problems we are actually processing information or transmitting it from one place to another, and the present theory is extremely useful in setting up rules and stating exact limits for what can and cannot be done. But we are in no position to investigate the process of thought, and we cannot, for the moment, introduce into our theory any element involving the human value of the information. This elimination of the human element is a very serious limitation, but this is the price we have so far had to pay for being able to set up this body of scientific knowledge. The restrictions that we have introduced enable us to give a quantitative definition of information and to treat information as a physically measurable quantity. This definition cannot distinguish between information of great importance and a piece of news of no great value for the person who receives it.*"

Within the realm of cognitive devices, this situation in which information carries relevance, which depends on whom receives it, typically arises. Let us go back to

[3]There is no reason why a cognitive device would not be able to infer on what the noise is made of.

[4]The system designer can be seen as a virtual entity sharing the knowledge of both transmitter and receiver.

the channel capacity example above. If the receiver is provided with some additional information concerning the transmission medium, like the typical channel delay spread, the channel Doppler spread, the number of reflections, the presence of buildings in the neighborhood, how much does this affect the channel capacity is an open issue, which cannot be solved within Shannon's framework. And if the receiver experiences a poor decoding rate, what kind of information should it request to the transmitter in order to increase its performance is also an open question, e.g., should the receiver request more pilot symbols at the risk of a huge waste in spectral efficiency, should the receiver request some deterministic information regarding a given parameter of the channel? All these problems do not have deterministic channel-dependent answers but depend on the specific knowledge of the transmitter/receiver pair to which some piece of additional information might or might not be valuable.

To partially answer those questions, we propose in the following to introduce first the notion of *degrees of belief*, which turns every deterministic measurable entity, e.g., the value of the channel capacity, the value of the SNR or the value of the channel fading, into a random variable with an assigned probability distribution: this probability distribution will translate the confidence of the cognitive devices regarding the estimation of the measurable entity in question. Then, we will introduce the notion of *relevance* which enables to estimate the relative importance of information. Finally, we will discuss our general view of the capabilities of a cognitive radio.

2.3. The Bayesian Approach

As briefly stated in the previous section, we aim at extending the classical Shannon's information theory to enable cognitive devices with the ability of *plausible reasoning*. That is, a cognitive radio should not rely on empirical (often erroneous) decisions, but rather should be able to express doubt and to reason honestly when provided with limited knowledge. A first step in this approach is to turn empirical decisions into degrees of belief.

2.3.1. Degrees of Belief and the Maximum Entropy Principle

In the Bayesian philosophy, contrary to the orthodox probability philosophy, deterministic parameters of a system, e.g., a weight, a height, the channel delay spread, which a cognitive entity needs to evaluate, must be characterized by the degrees of belief attached to all possible values for this parameter. Therefore this gives a clear meaning, for instance, to the probability that the height of the Eiffel Tower is 50 m. As a consequence, assuming a cognitive telecommunication device is not aware of the intensity σ^2 of the background noise, instead of expressing the achievable transmission rate as the scalar $C = \log(1+\sigma^2)$, which is therefore irrelevant to the communication device, it would be more adequate to consider the "vector" $C(x) = \log(1+x)$, $x \geq 0$, attached to a degrees of belief function, i.e., a probability density function, $p(x)$ for each potential noise variance x. Two fundamental questions arise at this point: (i) how to use the vector $C(x)$?, and (ii) how to compute $p(x)$?

Answering (i) is a matter of *decision theory*, in the sense that different requirements might come into play to decide on the actual transmission rate to use: if reliability is needed, one will decide to transmit at a rate $\log(1+x)$ such that $\int_0^x p(t)dt$ is less than a given (small) value, while if performance with low reliability is sought for, then x will take a larger value. This part of the cognitive radio spectrum will not be covered in this contribution.

Question (ii), on the contrary, is the point of interest in the present paper. Given the total amount of prior information at the cognitive device, how to assign degrees of belief in a systematic way? The answer to this question partially appears in the work of Shannon [1] but is better explained and developed by Jaynes [15] thanks to the introduction of the maximum entropy principle (MaxEnt) [16]. The key idea behind MaxEnt is to find a density function p, which fulfills the constraints imposed by the prior information I while introducing no additional (unwanted) information. In other words, this density function should maximize the ignorance about unknown parameters of the cognitive device, while satisfying the constraints given in I. In Jaynes' terms, this density function is maximally non-committal regarding missing information. This function translating ignorance is proven by Jaynes and more accurately later by Shore and Johnson [17] to be the entropy function H:

$$H(p) = -\int \log(p(t))p(t)dt\,. \tag{3}$$

When the information contained in I is of statistical nature, such as first or second order statistics, the function p which maximizes the entropy while satisfying the constraints in I is unique and can be computed with Lagrangian multipliers. An example will be given in Subsection 3.1.

2.3.2. Relevance

The problem of relevance of information is a second topic in the establishment of foundations for cognitive radios. If cognitive devices were to act like human beings, they should be able to request additional information when they do not have enough evidence to take decisions. For instance, to obtain a more accurate estimate of the noise variance σ^2 in order to have more confidence on the achievable transmission rates, an intelligent device could require the transmitter to stop transmitting so that it can estimate σ^2. But this would be an expensive waste in spectral efficiency, so it could alternatively request deterministic information on a dedicated channel from the transmitter. How accurate this information must be is then another problem. To be able to decide on what *question* to ask to the transmitter, the cognitive device needs to be able to judge the *relevance* of every possible question.

This notion of *questions*, or *inquiries*, is a philosophical topic upon which little literature and very few concrete

results exist. In 1978, Cox [18], who is also at the origin of the immense work from Jaynes on Bayesian probability theory [19], mathematically defined a question as the set of possible answers to this question. Therefore, a question will be relevant if its answers carry valuable information. Assuming the set of questions is seen as an ordered set, with the largest questions being the most relevant (since their answers carry potentially more cogent information), a cognitive device can decide which appropriate request to formulate to the transmitter. The work on relevance and questions is however still in its infancy, but we insist that those are fundamental needs to the cognitive radio field; for instance, interesting contributions are found in the works of Knuth *et al*. [20], who uses lattice theory to create partial orders of finite sets of questions, which is seen as the dual (in the lattice theory terminology) of the set of answers to those questions.

2.3.3. What is a Cognitive Radio?

In our viewpoint, a cognitive radio must ideally be able to adapt to its environment, by gathering all cogent information about the propagation channel, the transmitted signal, etc., while never producing undesirable empirical information. This would therefore relieve the telecommunication field from all ad hoc methods, based on empirical decisions concerning unknown parameters. This does not mean that a cognitive device is not prone to making errors; however, these potential errors will never originate from erroneous system assumptions, but rather from lack of information, which would generate *broad* maximum entropy distributions[5]. If more cogent information is provided to a cognitive device, it will integrate it and increase its decision capabilities. In a way, the more signals a cognitive communication device is fed with, the more efficient it is; this would mean, for instance, that cognitive devices age wisely: the older the cognitive device, the more efficient.

Regarding for instance signal sensing, the first steps of which will be detailed in Section 3, we expect a cognitive device to process the received signals as follows:

1. **Initialization**: integrate all cogent information about the communication channel, the properties of the supposedly received informative signal, etc., and compute the degrees of belief associated to all relevant variables.

2. **Update loop**: when the cognitive device is fed with incoming signals, it shall update its degrees of belief regarding all the previous variables and provide the overall probability that the received signal originates from a coherent data source.

3. **Decision**: using some criterion from decision theory, e.g., the evidence for the presence/absence of a coherent data source is more than a given threshold, the cognitive device declares whether data originating from a coherent source have been received.

[5]When little is known on a given parameter, the maximum entropy distribution attached to this parameter will be broad in the sense that no specific value is preferred to any other, while when more information is available on this parameter, the maximum entropy distribution will be very peaky around the exact value of the parameter.

This protocol does not necessarily provide the most efficient sensing strategy in specific situations (sometimes it might provide a quick response, sometimes traditional algorithms might provide faster responses), but it provides the most *honest* way to treat the signal sensing problem. It is important to note that no signal detection strategy can be proven superior to any other as long as too much information on the communication environment is missing. If a given algorithm could be proven better than the Bayesian strategy, this would mean this algorithm has an information advantage; honesty would then require that the Bayesian strategy be aware of this additional piece of information. The significant advantage of the Bayesian philosophy and the maximum entropy principle over classical methods is that they do not to take any empirical guess to solve a problem. Therefore, instead of being either luckily very good, or unluckily very bad depending on the accuracy of this "guess", they perform as best as their prior information allows them to.

Also, a cognitive device ought to be capable of requesting information when it faces a situation where it crucially lacks cogent information; for instance, a cognitive mobile phone in a low network coverage situation, should be able to request information (or even help) to the neighboring cell phones which enjoy better coverage. The interest of this request would be measured by its relevance. Adding the possibilities of formulating inquiries might eventually lead to enabling cognitive devices with the ability of *discussing*, instead of just *transmitting* and *receiving*. Bidirectional communications used to be a point of deep interest when it was realized that Shannon's theory of communication is in fact precisely a theory of *transmission*, in which past transmitted information is assumed uncorrelated with subsequent transmitted information. In 1973, Marko proposed a generalization of Shannon's information theory framework to encompass bidirectional communications [21], in the objective to accurately model the social interactions among animals and especially human beings. The lead was then followed by Massey [22] who extends information theory to include feedback in the expression of Shannon's mutual information.

3. Examples of Application

The most elementary requirement of a cognitive radio lies in its sensing capabilities. When a waveform is received at the cognitive device, it must be capable of deciding whether this waveform originates from a coherent source of information or if this waveform is pure background noise. When little is known by the receiver concerning the surrounding environment, this problem is very intricate and has led to lots of different ad hoc techniques. Our purpose in the following is to provide a unique way of deciding on

the presence of a coherent data source given a specific amount of prior information at the receiver. First, we will discuss channel modeling, which is a necessary step to properly handle the Bayesian signal detection method.

3.1. MaxEnt Channel Modeling

3.1.1. Introduction

Channel modeling is an entire field of research in telecommunications, which produces every year lots of new contributions. However, this huge amount of previous work on channel models leads to the following paradoxical conclusion: for a given total information gathered by a cognitive device, there exist many different channel models proposed in the literature. In such a situation, which of those channel models is the cognitive device supposed to trust? In reality, the fundamental difference between all those models lies in the additional hypothesis each of them, explicitly or implicitly, carries; some models might implicitly suggest that channels usually have a short delay spread for a given communication technology, or might suggest that it is very likely to have a strong line of sight component, etc.
However, if the receiver is not aware of that implicit information, this very information should honestly *not* be taken into account. What we will provide in the following is a systematic way to model channels, given some cogent information I, which fulfill the constraints imposed by I while being non-committal regarding unknown parameters. In brief, we will provide the most elementary models compliant with I, without introducing unwanted hypothesis.

3.1.2. Gaussian i.i.d. Channels

Surprisingly enough, we will realize that most of the classical channels in the basic literature fall into the maximum entropy channel modeling methodology. This is the case of Gaussian i.i.d. Indeed, let us assume that the information I known to the cognitive device gathers the following:

– the transmitter is equipped with n_T transmit antennas,

– the receiver is equipped with n_R receive antennas,

– the channel carries an energy E.

The transmission model is

$$\mathbf{y} = \sqrt{\frac{\rho}{n_\mathrm{T}}}\mathbf{H}\mathbf{x} + \mathbf{n}, \qquad (4)$$

where $\mathbf{x} \in \mathbb{C}^{n_\mathrm{T}}$ is the transmitted symbol vector, $\mathbf{n} \in \mathbb{C}^{n_\mathrm{R}}$ the thermal or interfering noise, ρ the signal to noise ratio and $\mathbf{H} \in \mathbb{C}^{n_\mathrm{T} \times n_\mathrm{R}}$ the channel we want to model.
In mathematical terms, based on the fact that

$$\int d\mathbf{H} \sum_{i=1}^{n_\mathrm{R}} \sum_{j=1}^{n_\mathrm{T}} |h_{ij}|^2 P_\mathbf{H}(\mathbf{H}) = n_\mathrm{T} n_\mathrm{R} E \qquad \text{(finite energy)}, \quad (5)$$

$$\int dP_\mathbf{H}(\mathbf{H}) = 1 \qquad (P_\mathbf{H}(\mathbf{H}) \text{ is a probability distribution}), \quad (6)$$

what distribution $P_\mathbf{H}$[6] should the modeler assign to the channel? The modeler would like to derive the most general model complying with those constraints, in other words the one which maximizes our uncertainty while being consistent with the energy constraint. This statement is mathematically expressed by the maximization of the following expression involving Lagrange multipliers with respect to $P_\mathbf{H}$:

$$\begin{aligned} L(P_\mathbf{H}) = &- \int d\mathbf{H} P_\mathbf{H}(\mathbf{H}) \log P_\mathbf{H}(\mathbf{H}) \\ &+ \gamma \sum_{i=1}^{n_\mathrm{R}} \sum_{j=1}^{n_\mathrm{T}} \left[E - \int d\mathbf{H} |h_{ij}|^2 P_\mathbf{H}(\mathbf{H})\right] \\ &+ \beta \left[1 - \int d\mathbf{H} P_\mathbf{H}(\mathbf{H})\right]. \end{aligned} \qquad (7)$$

If we derive $L(P_\mathbf{H})$ with respect to $P_\mathbf{H}$, we get

$$\frac{dL(P_\mathbf{H})}{dP_\mathbf{H}} = -1 - \log P_\mathbf{H}(\mathbf{H}) - \gamma \sum_{i=1}^{n_r} \sum_{j=1}^{n_t} |h_{ij}|^2 - \beta = 0, \quad (8)$$

which yields

$$\begin{aligned} P_\mathbf{H}(\mathbf{H}) &= \mathrm{e}^{-(\beta + \gamma \sum_{i=1}^{n_\mathrm{R}} \sum_{j=1}^{n_\mathrm{T}} |h_{ij}|^2} \\ &= \mathrm{e}^{-(\beta)} \prod_{i=1}^{n_\mathrm{R}} \prod_{j=1}^{n_\mathrm{T}} \exp -(\gamma \mid h_{ij} \mid^2) \\ &= \prod_{i=1}^{n_\mathrm{R}} \prod_{j=1}^{n_\mathrm{T}} P_{h_{ij}}(h_{ij}) \end{aligned}$$

with

$$P_{h_{ij}}(x) = \mathrm{e}^{-(\gamma |x|^2 + \frac{\beta+1}{n_\mathrm{R} n_\mathrm{T}})}. \qquad (9)$$

One of the most important conclusions of the maximum entropy principle is that, while we have only assumed the knowledge about the variance, this assumption naturally implies independent entries since the joint probability distribution $P_\mathbf{H}$ simplifies into products of $P_{h_{ij}}$. Therefore, based on the previous state of knowledge, the only solution to the maximization of the entropy is the Gaussian i.i.d. channel. This does not mean that we have supposed independence of the channel fades in the model, nor does it mean that real channels ought to be i.i.d. if those are known to be of energy E. However, in the generalized $L(P_\mathbf{P})$ expression, there exists no constraint on the dependence of the channel entries and this leads to natural independence as an honest guess on the behavior of the channel entries. Another surprising result is that the distribution achieved is Gaussian. Once again, Gaussianity is not an assumption but a consequence of the fact that the channel has finite energy.

[6] It is important to note that we are concerned with $P_{\mathbf{H}|I}$, where I represents the general background knowledge (here the variance) used to formulate the problem. However, for the sake of readability, $P_{\mathbf{H}|I}$ will be denoted $P_\mathbf{H}$.

3.1.3. Other Channel Models

In [23], a more complete survey on MaxEnt channel models is proposed. We will gather in the following the main results.
If the information I at the receiver is the same as previously but the receiver is not aware of the exact value of the channel energy E but knows that it is contained in the interval $[0, E_{\max}]$, then

$$P_{\mathbf{H}}(\mathbf{H}) = \int P_{\mathbf{H},E}(\mathbf{H},E)dE \quad (10)$$

$$= \int P_{\mathbf{H}|E}(\mathbf{H})P_E(E)dE\,. \quad (11)$$

If P_E is assigned a uniform prior on the set $[0, E_{\max}]$, then we obtain[7]:

$$P_{\mathbf{H}}(\mathbf{H}) = \frac{1}{E_{\max}\pi^{n_{\mathsf{R}}n_{\mathsf{T}}}} \int_{\frac{1}{E_{\max}}}^{\infty} u^{n_{\mathsf{R}}n_{\mathsf{T}}-2} \mathrm{e}^{-\sum_{i=1}^{n_{\mathsf{R}}}\sum_{j=1}^{n_{\mathsf{T}}}|h_{ij}|^2 u} du\,. \quad (12)$$

Note that the distribution is invariant to unitary transformations, is not Gaussian and moreover the entries are not independent when the modeler has no knowledge on the amount of energy carried by the channel. This point is critical and shows the effect of the lack of information on the exact energy.
If the channel covariance matrix $\mathbf{Q} = \mathrm{E}(\mathrm{vec}(\mathbf{H})\mathrm{vec}(\mathbf{H})^{\mathsf{H}})$ is known to the receiver, and therefore is part of the side information I, then, denoting $\mathbf{Q} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{H}}$ the spectral decomposition of $\mathbf{Q}$, with $\mathbf{V} = \left[\mathbf{v}_1, ..., \mathbf{v}_{n_{\mathsf{R}}n_{\mathsf{T}}}\right]$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_{n_{\mathsf{R}}n_{\mathsf{T}}})$:

$$P_{\mathbf{H}}(\mathbf{H}) = \frac{1}{\prod_{i=1}^{n_{\mathsf{R}}n_{\mathsf{T}}} \pi\lambda_i} \exp\left\{ \sum_{i=1}^{n_{\mathsf{R}}n_{\mathsf{T}}} \frac{|\,\mathbf{v}_i^{\mathsf{H}}\mathrm{vec}(\mathbf{H})\,|^2}{\lambda_i} \right\}. \quad (13)$$

3.2. Signal Detection

Now that channel modeling has been investigated, the multiple antenna signal sensing problem can be completely handled.

3.2.1. Channel State Information

In this problem, the cogent information at the receiver is divided into known parameters:

- S-i) the receiver has n_{R} antennas;
- S-ii) the receiver samples as many as L times the input from the RF interface;
- S-iii) the signal sent by the transmitter has a constant unit mean power; it is quite important to note that this hypothesis is very weak and should be made more accurate for communications schemes that are known only to use either QPSK (quadrature phase shift keying), 16-QAM (quadrature amplitude modulation), 64-QAM modulations for instance;
- S-iv) the MIMO channel has a constant mean power.

We similarly define additional information the receiver may be aware of

- V-i) the transmitter possesses (and uses) n_{T} antennas;
- V-ii) the noise variance σ^2 is known.

3.2.2. Signal Model

Given a certain amount of sampled signals, the objective of the signal detection methods is to be able to optimally infer on the following hypothesis:

- $\mathcal{H}_0$: only background noise is received;
- $\mathcal{H}_1$: informative data added to background noise is received.

Given hypothesis S-iii), the only information on the transmitted signal (under $\mathcal{H}_1$) is its unit variance. The maximum entropy principle claims that, under this limited state of knowledge, the transmitted data must be modeled as i.i.d. Gaussian [11]. The data vector, at time $l \in \{1, \ldots, L\}$, is denoted $\mathbf{s}^{(l)} = (s_1^{(l)}, \ldots, s_{n_{\mathsf{T}}}^{(l)})^{\mathsf{T}} \in \mathbb{C}^{n_{\mathsf{T}}}$. The data vectors are stacked into the receive matrix $\mathbf{S} = [\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(L)}]$.
If the noise level σ^2 is known, then either under $\mathcal{H}_0$ or $\mathcal{H}_1$, the background noise must be represented, due to the same maximum entropy argument as before, by a *complex standard Gaussian matrix* $\boldsymbol{\Theta} \in \mathbb{C}^{n_{\mathsf{R}} \times L}$ (i.e., a matrix with i.i.d. standard complex Gaussian entries θ_{ij}) [25]. Under $\mathcal{H}_1$, the channel matrix is denoted $\mathbf{H} \in \mathbb{C}^{n_{\mathsf{R}} \times n_{\mathsf{T}}}$ with entry h_{ij} being the link between the jth transmitting antenna and the ith receiving antenna. The model for $\mathbf{H}$ follows the MaxEnt channel modeling rules. In the present situation, only the constant mean power (or equivalently, the energy) of the channel is known. Therefore $\mathbf{H}$ will be modeled as i.i.d. Gaussian, following the reasoning in the previous section. The received data at sampling time l are given by the $n_{\mathsf{T}} \times 1$ vector $\mathbf{y}^{(l)}$ that we stack, over the L sampling periods, into the matrix $\mathbf{Y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}] \in \mathbb{C}^{n_{\mathsf{R}} \times L}$.
This leads for $\mathcal{H}_0$ to the model:

$$\mathbf{Y} = \sigma\boldsymbol{\Theta}\,. \quad (14)$$

And for $\mathcal{H}_1$ to

$$\mathbf{Y} = [\mathbf{H}\ ,\ \sigma\mathbf{I}_N] \begin{bmatrix} \mathbf{S} \\ \boldsymbol{\Theta} \end{bmatrix}. \quad (15)$$

We also denote by $\boldsymbol{\Sigma}$ the covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{E}[\mathbf{Y}\mathbf{Y}^{\mathsf{H}}] \quad (16)$$

$$= L\left(\mathbf{H}\mathbf{H}^{\mathsf{H}} + \sigma^2\mathbf{I}_{n_{\mathsf{R}}}\right) \quad (17)$$

$$= \mathbf{U}\left(L\boldsymbol{\Lambda}\right)\mathbf{U}^{\mathsf{H}}, \quad (18)$$

[7] The assignment of uniform priors on variables defined on a continuous space is a very controversial point of the maximum entropy theory, which is longly discussed in [11]. Another classically used prior, which solves the problem of invariance to variable change is the so-called Jeffreys' uninformative prior [24].

where $\boldsymbol{\Lambda} = \mathrm{diag}\left(v_1+\sigma^2,\ldots,v_{n_\mathsf{R}}+\sigma^2\right)$, with $\left\{v_i, i \in \{1,\ldots,n_\mathsf{R}\}\right\}$ the eigenvalues of $\mathbf{H}\mathbf{H}^\mathsf{H}$ and $\mathbf{U}$ a certain unitary matrix.

Our intention is to make a decision on whether, given the received data matrix $\mathbf{Y}$, the probability for $\mathcal{H}_1$ is greater than the probability for $\mathcal{H}_0$. This problem is usually referred to as *hypothesis testing* [11]. The decision criterion is based on the ratio

$$C(\mathbf{Y}) = \frac{P_{\mathcal{H}_1|\mathbf{Y}}(\mathbf{Y})}{P_{\mathcal{H}_0|\mathbf{Y}}(\mathbf{Y})}, \tag{19}$$

which we need to decide is whether lesser or greater than 1.

3.2.3. Results and Experiments

At this point in the derivation, computing C resorts to mere mathematical integration. The details of the calculus are given in [26]. We only provide here the results. First, assume σ^2 and n_T are known, then, denoting $x_1,\ldots,x_{n_\mathsf{R}}$ the eigenvalues of $\mathbf{Y}\mathbf{Y}^\mathsf{H}$:

$$P_{\mathbf{Y}|\mathcal{H}_0}(\mathbf{Y}) = \frac{1}{(\pi\sigma^2)^{n_\mathsf{R}L}}\mathrm{e}^{-\frac{1}{\sigma^2}\mathrm{tr}\mathbf{Y}\mathbf{Y}^\mathsf{H}} \tag{20}$$

$$= \frac{1}{(\pi\sigma^2)^{n_\mathsf{R}L}}\mathrm{e}^{-\frac{1}{\sigma^2}\sum_{i=1}^{n_\mathsf{R}} x_i} \tag{21}$$

and

$$P_{\mathbf{Y}|\mathcal{H}_1}(\mathbf{Y}) = \int_{\boldsymbol{\Sigma}} P_{\mathbf{Y}|\boldsymbol{\Sigma}\mathcal{H}_1}(\mathbf{Y},\boldsymbol{\Sigma})P_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})d\boldsymbol{\Sigma} \tag{22}$$

$$= \int_{\mathcal{U}(n_\mathsf{R})\times(\mathbb{R}^+)^{n_\mathsf{R}}} P_{\mathbf{Y}|\boldsymbol{\Sigma},\mathcal{H}_1}(\mathbf{Y},\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{H})P_{\boldsymbol{\Lambda}}(\boldsymbol{\Lambda})d\mathbf{U}d\boldsymbol{\Lambda}, \tag{23}$$

which, after complete derivation, using in particular the Harish-Chandra identity [27], gives

$$P_{\mathbf{Y}|\mathcal{H}_1}(\mathbf{Y}) = \alpha \sum_{\mathbf{a}\subset[1,n_\mathsf{R}]} \frac{\mathrm{e}^{\frac{\sum_{i=1}^{n_\mathsf{T}} x_{a_i}}{\sigma^2}}}{\prod_{a_i}\prod_{\substack{j\neq a_1\\ \cdots\\ j\neq a_i}}(x_{a_i}-x_j)} \times \sum_{\mathbf{b}\in\mathcal{P}(n_\mathsf{T})}(-1)^{\mathrm{sgn}(\mathbf{b})+1}\prod_{l=1}^{n_\mathsf{T}} J_{n_\mathsf{R}-L-2+b_l}(n_\mathsf{T}\sigma^2, n_\mathsf{T}x_{a_l}) \tag{24}$$

with $\mathcal{P}(k)$ the ensemble of permutations of k, $\mathrm{sgn}(\mathbf{b})$ the sign of the permutation $\mathbf{b}$:

$$J_k(x,y) = \int_x^{+\infty} t^k \mathrm{e}^{-t-\frac{y}{t}}dt \tag{25}$$

and

$$\alpha = \frac{(n_\mathsf{R}-n_\mathsf{T})!\,n_\mathsf{T}^{(2L-n_\mathsf{T}+1)n_\mathsf{T}/2}\mathrm{e}^{n_\mathsf{T}^2\sigma^2-\frac{\sum_{i=1}^{n_\mathsf{R}} x_i}{\sigma^2}}}{n_\mathsf{R}!\,\pi^{n_\mathsf{R}L}\sigma^{2(n_\mathsf{R}-n_\mathsf{T})(L-n_\mathsf{T})}\prod_{j=1}^{n_\mathsf{T}-1} j!}. \tag{26}$$

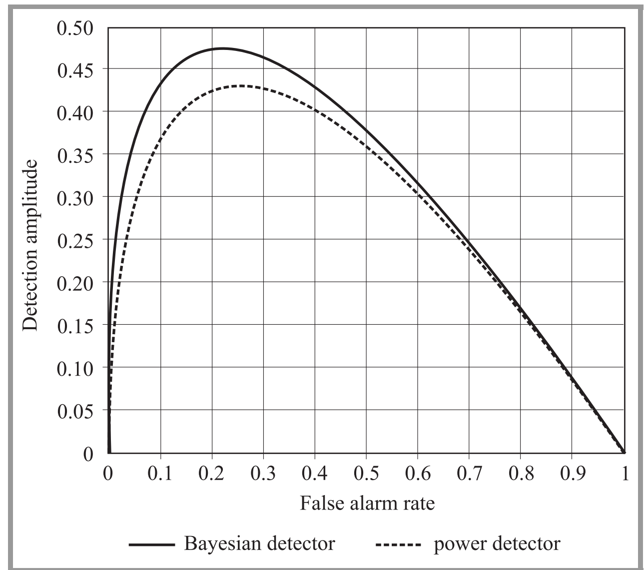


Fig. 1. Detection amplitude comparison in MIMO; $M=1$, $N=8$, $L=20$, $SNR=-10$ dB.

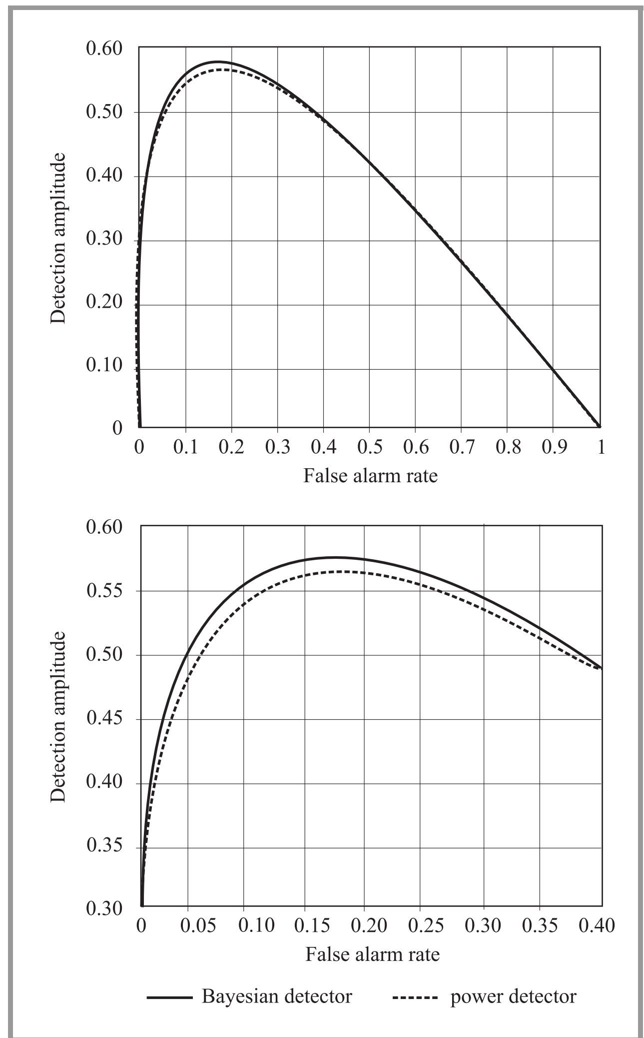


Fig. 2. Detection amplitude comparison in MIMO; $M=2$, $N=8$, $L=10$, $SNR=-10$ dB.

These expressions are rather complex but show that the Bayesian signal detection, within the state of knowledge I, only depends on the eigenvalues $x_1, \ldots, x_{n_R}$ of the Gram matrix $\mathbf{YY}^H$ of the received data $\mathbf{Y}$.

A comparison with the classical power detector, e.g., [28]–[30], which consists in summing all individual powers received on the antenna array is provided in Fig. 1. In the latter, $n_T = 1$ and the comparison is made between the difference "correct detection rate minus false alarm rate" computed from Monte Carlo simulations for both Bayesian and classical signal detectors.

We observe a slight gain in performance due to the novel Bayesian detector. Especially, for a low false alarm rate (which is often demanded in practice), we observe a large gain in correct detection rate provided by the Bayesian detector. This statement is however only valid for $n_T = 1$. When n_T is larger, then the channel hardening effect reduces the gain of the Bayesian detector. This is shown in Fig. 2 in which $n_T = 2$.

Now, if the noise power σ^2 is not perfectly known (this is classically the situation since knowledge of the noise power implies prior identification of the background noise), the probability distribution must be updated by marginalizing over σ^2, from the lower bound σ_-^2 to the upper bound σ_+^2 on σ^2. Therefore:

$$P_{\mathbf{Y}|I} = \frac{1}{\sigma_+^2 - \sigma_-^2} \int_{\sigma_-^2}^{\sigma_+^2} P_{\mathbf{Y}|\sigma^2, I}(\mathbf{Y}, \sigma^2) d\sigma^2, \qquad (27)$$

which is too involved to compute, but can be numerically estimated. An example is provided in Fig. 3 in which

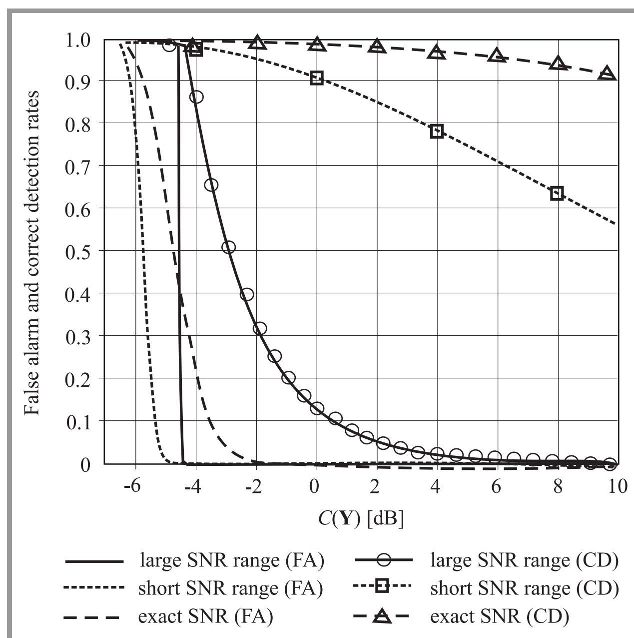


Fig. 3. False alarm (FA) and correct detection (CD) rates for exact, short discrete range ($\{0, 2.5, 5\}$) and large discrete range ($\{-5, -2.5, 0, 2.5, 5\}$) SNR; $M = 1$, $N = 8$, $L = 10$, $SNR = 2.5$ dB.

the intervals $[\sigma_-^2, \sigma_+^2]$ are taken increasingly large. In the latter, correct detection rate against false alarm rate is depicted for different values of σ_-^2 and σ_+^2. It is observed that the range of ensured correct detection gets increasingly narrower when $[\sigma_-^2, \sigma_+^2]$ is large. Note that this situation cannot be compared against classical power detection methods which do not provide solutions when σ^2 is not perfectly known.

4. Discussion

In addition to these first two studies on maximum entropy considerations for cognitive radios, the authors proposed more practical studies on maximum entropy orthogonal frequency division multiplexing (OFDM) channel estimation [31], maximum entropy carrier frequency offset estimation [32; Chapter 13], minimal update channel estimation [33], etc. From all those studies, we draw the following conclusions.

- Quite often, classical techniques, in particular in the channel estimation field, are rediscovered using MaxEnt. However, it is important to note that, even if the final formulas are the same in the classical and Bayesian MaxEnt approach, the philosophical conclusions are very different. Usually classical methods derive from empirical parameter settings, which could have been chosen differently, while Bayesian approaches give unique deterministic solutions, which stem from honesty in the treatment of prior information.

- The MaxEnt principle allows one to marginalize over all parameters when those are not perfectly known. As a consequence, while classical solutions are found anew, those methods can usually be extended to cope with the lack of information on some key variables. For instance, in the signal sensing proposed in Section 3 and completed in [26], the situations where noise variance and number of transmit antennas are not perfectly known can be easily handled, whereas classical methods stumble on these problems and solve them by empirical (possibly largely erroneous) parametrizations.

On the other hand, MaxEnt calculus and final solutions can turn very rapidly extremely mathematically involved, as exemplified by the final signal sensing formula in Section 3. This is a major problem, and the subject of most criticism towards Bayesian approaches. A missing part in these MaxEnt approach would be a systematic method which, from the general (very involved) solution, would provide approximate solutions. Quite remarkably, Caticha provides a vision of the maximum entropy principle, or more precisely the minimum cross entropy principle, which might help decide on the optimal approximation taken from a set of possible approximations [34]. These considerations might lead to such systematic approximation methods.

Another point of concern in the MaxEnt framework lies in the many integrals that may need to be computed when lit-

tle is known on the surrounding environment. With the increasing capabilities of modern computers, numerical approximations might help to compute those integrals, but these approximations would only be valid if not so many integrals are considered; two reasons explain this fact: first, the complexity increase due to additional integrals is exponential in the number of integrals and second, small errors in inner integrals tend to lead to large errors when integrated many times (this is often referred to as the curse of dimensionality).

As a consequence, while the first MaxEnt results provided by the authors show significant performance increase, many problems remain to be solved for cognitive radios to be fully intelligent, both on fundamental philosophical considerations (many questions raised in the introduction of the present paper are left unanswered) and on practical applications.

5. Conclusion

In this paper, we introduced the fundamentals of cognitive radios under a physical layer viewpoint. These fundamentals are based on the extension of Shannon's information theory to the Bayesian probability theory and the maximum entropy principle, which enables the cognitive devices with plausible (human-like) reasoning. Through the first-step studies of maximum entropy channel modeling and signal sensing, we paved the path to the establishment of strong theoretical grounds to the realm of cognitive radios.

Acknowledgment

This work is partially supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless Communications NEWCOM++.

References

[1] C. Shannon, "A mathematical theory of communications", *Bell Syst. Tech. J.*, vol. 27, no. 7, pp. 379–423, 1948.

[2] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", *Wirel. Pers. Commun.*, vol. 6, no. 3, pp. 311–335, 1998.

[3] E. Telatar, "Capacity of multi-antenna Gaussian channels", *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999.

[4] J. Mitola III and G. Q. Maguire Jr, "Cognitive radio: making software radios more personal", *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, 1999.

[5] S. Haykin, "Cognitive radio: brain-empowered wireless communications", *IEEE J. Selec. Areas in Commun.*, vol. 23, no. 2, pp. 201–220, 2005.

[6] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments", in *First IEEE Int. Sympo. New Front. Dynam. Spectr. Acc. Netw.*, Baltimore, USA, 2005, pp. 131–136.

[7] L. S. Cardoso, M. Kobayashi, Ø. Ryan, and M. Debbah, "Vandermonde frequency division multiplexing for cognitive radio", in *IEEE 9th Worksh. Sig. Proces. Adv. Wirel. Commun.*, Recife, Brazil, 2008, pp. 421–425.

[8] C. Shannon, "Programming a computer for playing chess", *Phil. Mag.*, Ser. 7, vol. 41, no. 314, pp. 256–275, 1950.

[9] C. Shannon, "Computers and automata", Proc. IRE, vol. 41, iss. 10, pp. 1234–1241, 1953.

[10] C. Shannon, "A mind-reading machine", Bell Laboratories memorandum, 1953 (reprinted in *Claude Elwood Shannon Collected Papers*. New York: IEEE Press, 1993).

[11] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press, June 2003.

[12] A. Caticha, "Maximum entropy, fluctuations and priors", in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Ed. A. Mohammad-Djafari, vol. 568, no. 94, 2001, http://arxiv.org/abs/math-ph/0008017

[13] A. Caticha, "Lectures on probability, entropy and statistical physics", 2008, http://arxiv.org//abs/0808.0012 [physics.data-an]

[14] L. Brillouin, *Science and Information Theory*. New York: Dover Publ., 1963.

[15] E. T. Jaynes, "On the rationale of maximum-entropy methods", *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[16] E. T. Jaynes, "Information theory and statistical mechanics", *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[17] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Trans. Inform. Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[18] R. T. Cox, "Of inference and inquiry, an essay in inductive logic", in *Proceedings 1978 Maximum Entropy Formalism Conference*. Cambridge: MIT Press, 1979, pp. 119–167.

[19] R. T. Cox, "Probability, frequency and reasonable expectation", *Amer. J. Phys.*, vol. 14, pp. 1–13, 1946.

[20] K. H. Knuth, A. Caticha, J. L. Center, A. Giffin, and C. C. Rodriguez, "Lattice theory, measures and probability", in *27th Int. Worksh. Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, *AIP Conf. Proc.*, Melville, USA, 2007, vol. 954, pp. 23–36.

[21] H. Marko, "The bidirectional communication theory – a generalization of information theory", *IEEE Trans. Commun.*, vol. COM-21, no. 12, pp. 1345–1351, 1973.

[22] J. Massey, "Causality, feedback and directed information", in Proc. Int. Symp. Inform. Theory Appl., Waikiki, Hawaii, 1990.

[23] M. Guillaud, M. Debbah, and A. L. Moustakas, "Maximum entropy MIMO wireless channel models", *IEEE Trans. Inform. Theory*, Dec. 2006, http://arxiv.org/abs/cs.IT/0612101

[24] H. Jeffreys, "An invariant form for the prior probability in estimation problems", *Proc. Royal Soc. London, Ser. A, Math. Phys. Sci.*, vol. 186, no. 1007, pp. 453–461, 1946.

[25] A. M. Tulino and S. Verdú, *Random Matrix Theory and Wireless Communications*. Hanover (USA): Now Publishers, 2004, vol. 1, iss. 1.

[26] R. Couillet and M. Debbah, "Bayesian inference for multiple antenna cognitive receivers", *IEEE Wirel. Commun. Netwo. Conf.*, Budapest, Hungary, 2009.

[27] A. B. Balantekin, "Character expansions, Itzykson-Zuber integrals, and the QCD partition function", *Phys. Rev. D,* vol. 62, no. 8, p. 085017, 2000.

[28] H. Urkowitz, "Energy detection of unknown deterministic signals", *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[29] V. I. Kostylev, "Energy detection of a signal with random amplitude", in *Proc. IEEE Int. Conf. Commun. ICC'02*, New York, USA, 2002, pp. 1606–1610.

[30] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Spatial-spectral joint detection for wideband spectrum sensing in cognitive radio networks", in *Proc. ICASSP Conf.*, Las Vegas, USA, 2008.

[31] R. Couillet and M. Debbah, "A maximum entropy approach to OFDM channel estimation", in *IEEE Sig. Proces. Adv. Wirel. Commun. Conf.*, Perugia, Italy, 2009.

[32] L. Song, T. Jiang, and Y. Zhang, *Orthogonal Frequency Division Multiple Access (OFDMA)* to be published by Auerbach Publ., CRC Press, Taylor & Francis Group.

[33] R. Couillet, A. Ancora and M. Debbah, "Minimal update for OFDM channel estimation" to be submitted.

[34] C. Y. Tseng and A. Caticha, "Maximum entropy approach to the theory of simple fluids", 2003, Arxiv preprint cond-mat/0310746.

Romain Couillet was born in Abbeville, France. He received his Master of Science degree (specialization in mobile telecommunication) from the Eurecom Institute, France in 2007. He received his M.Sc. degree (specialization in communication systems) from the Telecom ParisTech, France, in 2007. In Sept. 2007, he joined NXP, founded by Philips, which eventually merged into the ST-Ericsson company. In ST-Ericsson, he works as an algorithm development engineer on the Long Term Evolution Advanced (LTE-A) project. In parallel to his position in ST-Ericsson, he is currently pursuing his Ph.D. degree at the Supélec, France. His research interests include wireless communications, multi-user MIMO detection techniques, cognitive radio and random matrix theory. He is the recipient of the Valuetools 2008 best student paper award.

ST-Ericsson
505 route des Lucioles, 06650 Sophia Antipolis, France
e-mail: romain.couillet@supelec.fr
Alcatel-Lucent Chair, Supélec
3 rue Joliot Curie, 91192 Gif sur Yvette, France

Mérouane Debbah Mérouane Debbah was born in Madrid, Spain. He entered the Ecole Normale Suprieure de Cachan, France, in 1996, where he received his M.Sc. and Ph.D. degrees, respectively, in 1999 and 2002. From 1999 to 2002, he worked for the Motorola Labs on Wireless Local Area Networks and prospective fourth generation systems. From 2002 until 2003, he was appointed Senior Researcher at the Vienna Research Center for Telecommunications (FTW), Austria, working on MIMO wireless channel modeling issues. From 2003 until 2007, he joined the Mobile Communications Department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor. He is presently a Professor at the Supelec (Gif-sur-Yvette, France), holder of the Alcatel-Lucent Chair on Flexible Radio. His research interests are in information theory, signal processing and wireless communications. He is the recipient of the 2005 Mario Boella Prize Award, the 2007 General Symposium IEEE GLOBECOM best paper award, the Wi-Opt 2009 best paper award as well as the Valuetools 2007,Valuetools 2008 and CrownCom 2009 best student paper awards. He is a WWRF fellow.

e-mail: merouane.debbah@supelec.fr
Alcatel-Lucent Chair, Supélec
3 rue Joliot Curie, 91192 Gif sur Yvette, France

Paper

A Microscopic Approach for THz Intersubband Challenges

Mauro F. Pereira

Abstract—The main candidate to be a practical and low cost high power THz source is the intersubband-based quantum cascade laser, which can have a tremendous impact in many practical applications, including last mile and indoor telecommunication systems. In this review we discuss current challenges for THz intersubband device development from a microscopic point of view. Next summarize the search for new mechanisms and structure designs that can lead to intersubband gain without population inversion. This is a very important topic of current research, since is both an extremely elegant phenomenon from the basic physics of view and crucial for effective lasing in the THz range. The reason is that scattering phenomena can lead to level broadenings of the same order of magnitude of the lasing transitions, making population inversion by carrier injection in upper lasing subbands extremely difficult. Previous work in the literature is compared and contrasted with a new scheme that may lead to high temperature lasing by engineering the nonequilibrium population inversion with a combination of band structure and many body effects mediated by a k-space filter.

Keywords—band structure engineering, coupled valence bands, intersubband laser, intersubband transitions, lasing without inversion, terahertz radiation.

1. Introduction

Intersubband optics and the terahertz (THz) range of the electromagnetic spectrum are the current frontiers in semiconductor science from both fundamental and applications points of view. To date the exploitation of THz waves has been largely underdeveloped due to the lack of a compact coherent source; providing high output power preferably with continuous wave (cw) operation. This is despite the huge potential THz technology has in a varied list of applications; detecting tumours and skin cancers, pharmaceutical applications, detecting and discriminating different explosive threat materials, environmental sensing and gas monitoring, industrial process control, as well as applications in astronomy, semiconductor imaging, security and medical imaging and telecommunication applications.

Due to the increasing demand for bandwidth it is expected that THz communication systems will be developed in a few years time. American and Japanese companies have already taken the first steps in this direction. Short-range indoor communication systems which work at carrier frequencies of a few hundred gigahertz will represent the wireless local area network (LAN) systems of the future. Such systems could also solve the current "last mile telecommunication" problem.

As a matter of fact, the demand for bandwidth in wireless short-range communication systems has doubled every 18 months over the last 25 years and there is no reason to expect that this trend would come to an end [1].

There are several different meanings for bandwidth and here it means the rate at which information can be transmitted over a given medium (telephone cables, cable TV, microwave relays, fiber optics, satellite links, etc.) In general the permissible bandwidth is about 0.1% to 1% of the carrier frequency. This means that the available bandwidth grows with the carrier frequency. In other words, the higher the frequency, the greater the volume of information that can be transmitted.

The current short-range communications systems Bluetooth and wireless LANs will not be able to deliver the bandwidth needed in 10 years because they operate with carrier frequencies of only a few gigahertz and the same happens for the currently emerging ultra wide band technology. Those systems are limited to rates below 1 Gbit/s. Future systems delivering data rates above 10 Gbit/s will have to work at several ten or preferably a few hundred GHz and this will require higher carrier frequencies. The use of bandwidth is regulated in the USA up to frequencies of 300 GHz with a window from 275 to 300 GHz which is reserved for communications [2]. In Europe the allocation is the same but ends at 275 GHz. In other words, there is free communication bandwidth available above 275 GHz. Systems operating at those frequencies are already in the THz range. Thus, compact and inexpensive devices emitting in this range are required. The quantum cascade laser (QCL) is a potential candidate.

The first mid infra red (MIR) QCL was demonstrated in 1994 by Faist *et al.* at Bell Labs [3] and the extension of the QCL design to emission frequencies below the Restrahlen band was achieved in 2002 when Köhler *et al.* [4] demonstrated the first THz QCL, at 4.4 THz. Since then they have achieved a maximum operating temperature of 186 K [5].

Even though intersubband emitters are the best candidate as compact high power sources for THz radiation, a current review of the field seems to indicate that conventional designs for QCLs emitting at a given energy $\hbar\omega$ appear to have an maximum operating temperature $T_{\max} \approx \hbar\omega/k_B$ [6]. Thus attainment of room temperature operation seems unlikely without external cooling. Dramatic progress can only be achieved with deep understanding and unconventional manipulation of microscopic mechanisms.

Figure 1 illustrates some of the microscopic difficulties that prevent current THz QCLs to operate at room temperature at high output power. The most critical issue currently

limiting THz QCLs as widely exploitable imaging sources is achieving population inversion at high temperatures depicted in Fig. 1(a)–(d).

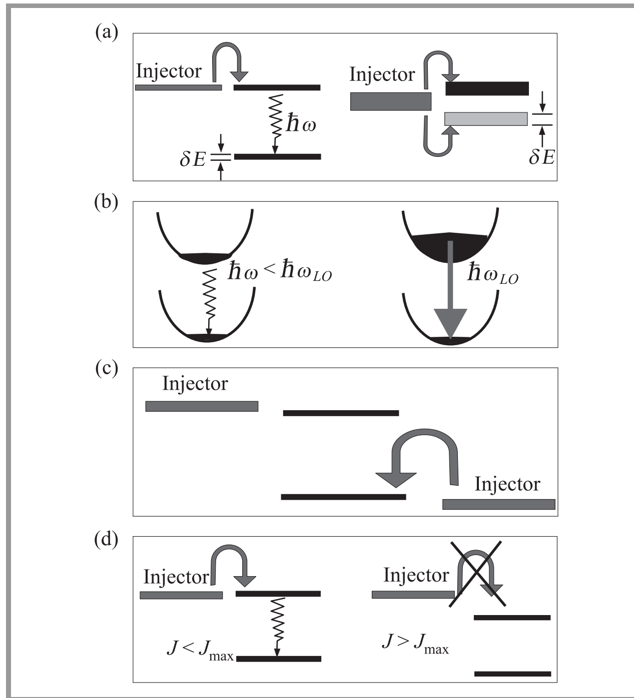


Fig. 1. Cartoon of the most critical issue currently limiting THz QCLs as widely exploitable imaging sources is achieving population inversion at high temperatures.

In Fig. 1(a), in contrast to the mid-IR case (left diagram), for THz transitions (right diagram) the energy difference between the lasing levels is not much larger than the level broadening. The electrons can tunnel from the injector to the upper or lower levels reducing population inversion. In Fig. 1(b), in the right-hand diagram the temperature is high. The upper subband is occupied up to electronic states with sufficient energy to emit longitudinal-optic (LO) phonons, and thus decay nonradiatively to the lower lasing level, reducing the population inversion. In Fig. 1(c), thermal backfilling of the lower radiative state by carriers from the injector states reduces population inversion at high temperatures. In Fig. 1(d), at high temperatures beyond the peak bias level when the maximum achievable current J_{max} is reached, the injector and upper levels get misaligned and the device ceases to operate.

2. Lasing Without Population Inversion

Lasing without inversion (LWI) is a new solution for producing laser light and it has been achieved in gas and semiconductor lasers, opening new possibilities for cleverly sidestepping traditional difficulties of producing radiation in both extremes of the spectrum: ultraviolet and x-ray lasing on the high energy side and far infrared (TERA-MIR) in the low energy range. In a gas of atoms, laser light buildup begins when a single photon, emitted by an atom in a high-energy (excited) state, stimulates other excited atoms to emit photons with identical attributes. Ordinary lasers normally require the energy-intensive process of "population inversion", in which a majority of the atoms must be excited into a high-energy state. Promoting atoms into excited states prepares them for participating in the laser process, but it also serves to prevent them from soaking up the light and thereby sabotaging the laser process.

However, maintaining a population inversion in ultraviolet and x-ray lasers is extremely difficult because the high-lying excited states necessary to produce such light are so short-lived. To the best of the author's knowledge, the first demonstration of lasing without inversion appeared in atomic optics and was based on quantum interference in Rb atoms [7]. Previous experiments had produced nanosecond bursts of light without population inversion, but the paper [7] was the first to report a sustained laser beam through LWI. In these experiments, an external laser beam essentially creates two pathways for the atoms to get from the ground state ($|a\rangle$) to the excited state ($|b\rangle$). In the rubidium experiment, for example, the probability of getting atoms from ($|a\rangle$) to ($|b\rangle$) becomes the overlap of the likelihood of getting from state ($|a\rangle$) to state ($|b\rangle$) directly and going from state a to an even higher excited state ($|c\rangle$) then decaying to state ($|b\rangle$). Under the proper conditions, the overlapping likelihoods can interfere so as to cancel each other out, preventing absorption. In other words, effective LWI in atomic systems has been based on quantum interference. Future goals are to achieve LWI in inexpensive diode lasers (like those in CD players) and to produce x-ray and UV light through LWI. Here we analyse a very different scenario in intersubband emitters [8]–[10]. This may turn out crucial in the search for room temperature THz QCLs. The main reason for this is that dephasing and scattering phenomena can lead to level broadenings of the same order of magnitude of the lasing transitions, making population inversion by carrier injection in upper lasing subbands extremely difficult, as depicted in the cartoons of Fig. 1.

The first experimental realization of intersubband lasers without inversion exploited the nonparabolicity of the conduction subbands and local population inversion near $k = 0$ even though the lowest subband may have larger global occupation [10].

Later on, valence-band-based designs have been proposed [11]–[13]. Intervalence band emitters based on Si-Ge structures have been investigated. However, lasing on quantum cascade structures has never been demonstrated. Only electroluminescence has been measured so far [14]–[16]. A complete set parameters required for predictive calculations of optical properties of Si-Ge devices is still not known, although progress in this direction has been recently achieved [17]. On the other hand, the material parameters for the III-V system investigated here are very well known. This review complements the results given recently in [18], [19] comparing and contrasting the results for two different well widths.

The following features are unique of the approach introduced in [18] and were not found in previous studies in the literature:

- The strongly k-dependent transverse-electric (TE) transition is used to create a k-space filtering effect that enhances the gain due to local population inversion that arises due to the strongly nonparabolicity of the valence bands.
- The nonequilibrium Keldysh Green's function method is used allowing the consideration of dephasing effects.
- Detrimental cross-absorption due to multiple transitions is taken into account.

Another useful feature of the approach is that simple surface emitting designs can be constructed due to the TE polarization of the emitted field.

3. Numerical Results and Discussion

Necessary conditions to obtain lasing without inversion exploiting nonparabolicity are that the upper conduction subband should either have smaller effective mass or cooler electrons. For unstrained GaAs-AlGaAs wells, even though the conduction bands can be characterized by effective masses (parabolic) "nonparabolicity" appears with different effective masses per subband. And usually the lowest band has a lighter effective mass. The valence bands in contrast have typically lower averaged effective masses in the upper lasing subband which is an advantage. However, as the nonparabolicity can be in some cases so strong that an effective mass does not make sense and the full dispersions must be used, which can complicate the numerical calculations enormously, specially if many particle and dephasing effects are taken into account [18], [19].
The model system investigated in this paper is globally out of equilibrium but the holes are assumed to be thermalized within each subband with occupation functions characterized by different temperatures. The temperatures of the electrons (or holes) in different subbands can be different. This fact has as determined experimentally for conduction band designs by means of microprobe photoluminescence experiments [20]. It is a design challenge to cool the upper subband electrons for more efficient lasing. Furthermore the number of carriers in each subband in an intersubband system depends on the injection scheme and can be chosen independently of the temperature in simulations, see, e.g., [18].
The results shown next are all for non-global inversion conditions and $GaAs/Al_{0.3}Ga_{0.7}As$ quantum wells. Only the top two valence subbands 1 and 2 are occupied and in all cases the lowest hole band has equal or more total carrier density than the upper (lasing) hole subband, i.e., $n_2 \leq n_1$. Detrimental cross absorption due to higher (empty) hole subbands is taken into account. Before proceeding further the following point should be highlighted: very good agreement with experiments has been obtained by considering electron-phonon and electron-impurity scattering as the main origin of dephasing in mid infrared systems [21]–[23]. Interface roughness scattering also plays a role even in THz systems, however, this effect depends on sample quality [24], [25] and the inclusion of sample quality considerations would be beyond the scope of this paper.
The nature of the main non-radiative channel between subbands, which is crucial to the device performance, remains a matter of controversy for THz systems. The typical energy spacing for one active region period is lower than the optical phonon energy (36 meV in GaAs). Thus, optical phonon emission may not always be the dominant non-radiative mechanism unless the electrons are very hot. Moreover, many-body effects (electron-electron scattering resonances) have been observed in recent experiments for THz QCLs operating in the quantum Hall regime [26]. Thus in the theoretical limit analyzed in this paper, the dephasing is due only to electron-electron mechanisms [27]. Consideration of other scattering channels will be the subject of future research.
The optical properties in which we are interested are calculated in our approach through the optical susceptibility, χ, which can be directly obtained from the Green's function G. The absorption $\alpha(\omega)$ and gain spectra $g(\omega) = -\alpha(\omega)$ are calculated from the imaginary part of the optical susceptibility $\chi(\omega)$ (see, [18]):

$$\begin{aligned} \alpha(\omega) &= \frac{4\pi\omega}{cn_b}\Im\{\chi(\omega)\}, \\ \chi(\omega) &= 2\sum_{\mu\neq\nu,\vec{k}} \wp_{\mu\nu}(k)\chi_{\nu,\mu}(k,\omega). \end{aligned} \tag{1}$$

Here n_b denotes the background refractive index, c is the speed of light, $\wp_{\nu\mu}(k) = ed_{\nu\mu}(k)$ is the transition dipole moment between the subbands ν and μ, which are labelled $\mu = 1, 2, \ldots$ from the top valence band. Thus in the discussion that follows, if holes are injected in the subband 2 and make a transition to subband 1 creating a photon, this actually means that an electron made a transition from valence subband 1 to subband 2 and will be called a (2,1) transition.
The nonequilibrium steady-state susceptibility function $\chi_{\nu\mu}(k,\omega)$ is evaluated through the carriers Keldysh Green's function G, whose time evolution is described by a Dyson equation. The resulting integro-differential equation for $\chi_{\nu\mu}(k,\omega)$ is solved numerically in this paper including many body effects at the Hartree-Fock level, complex nonparabolic band structure, correlation and dephasing mechanisms.
The numerical scheme used here can be summarized as follows: the first step is the solution of the 8×8 $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian [27]. The Green's functions and self-energies are expanded using eigenstates and eigenvalues of this Hamiltonian. Next, by assuming thermalized holes, the full nonequilibrium Green's function (NEGF) scheme is simplified and reduces to the self-consistent evaluation of chemical potentials and self-energy matrix elements which lead

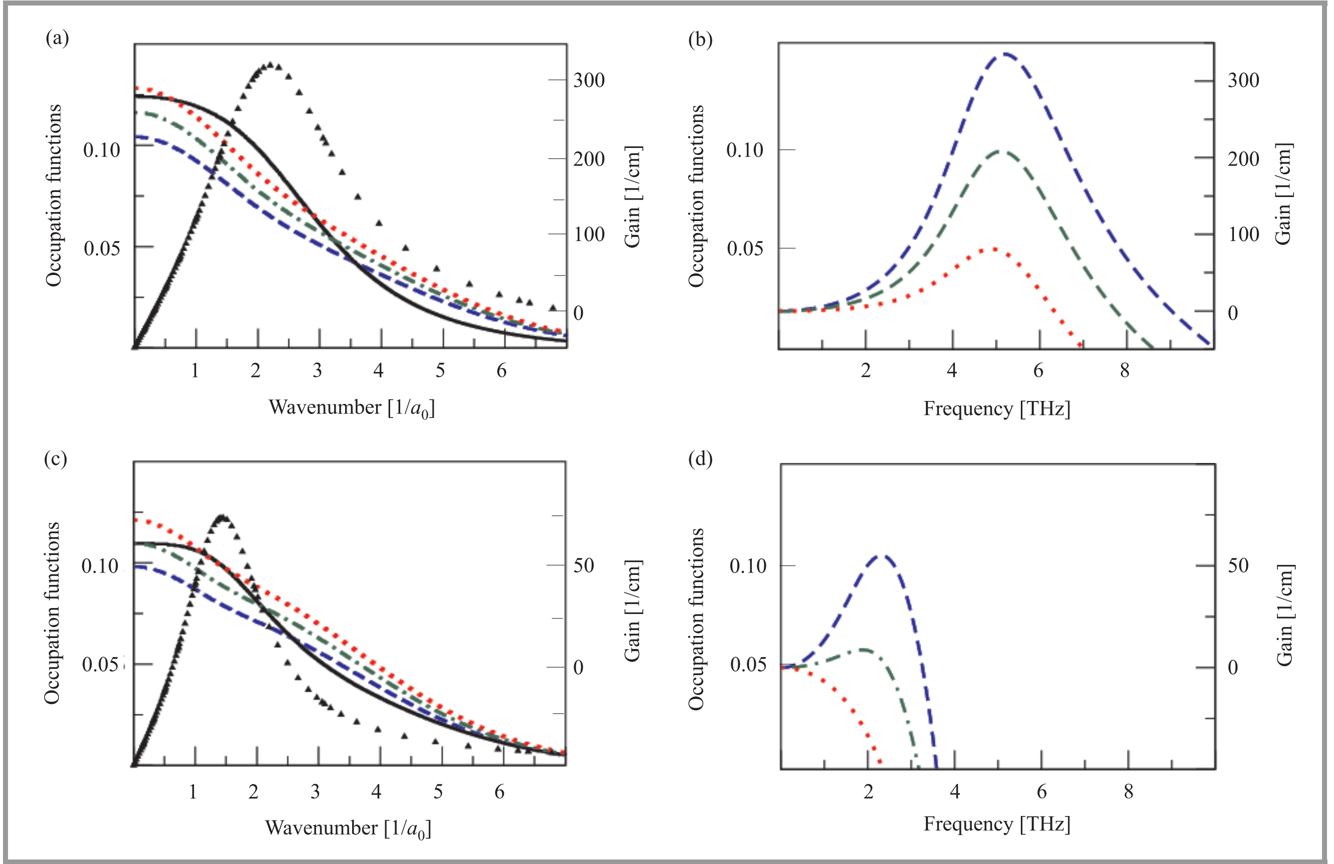


Fig. 2. Comparison of occupation functions and transition dipole moments (symbols) in (a), (c) against gain spectra in (b), (d). The two quantum wells considered have widths of 5 nm (a), (b) and 10 nm (c), (d). The wavenumber is presented in units of $1/a_0$, where a_0 is the 2D Bohr radius.

to subband energy renormalizations, dephasing constants and occupation functions. Only carrier-carrier scattering is considered here and details of the corresponding self-energy are given in [27].

Finally, absorption and gain are given by the solution of the integro-differential equation obtained from the carriers Green's function by numerical matrix inversion. The susceptibility function can be written as

$$\chi_{\nu\mu}(k,\omega) = \sum_{\mathbf{k}'\neq\mathbf{k}} \left(M^{\nu\mu}_{kk'}\right)^{-1} \wp_{\nu\mu}(k')\delta n_{\nu\mu k'}, \quad (2)$$

where $\left(M^{\nu\mu}_{k'k}\right)^{-1}$ is the inverse of

$$M^{\nu\mu}_{kk'} = \left(\hbar\omega - e_{\nu\mu}(k) + i\Gamma_{\nu\mu}\right)\delta_{k,k'} + \left(1-\delta_{k,k'}\right)\delta n_{\nu\mu k}\tilde{V}^{\nu\mu}_{\mathbf{k}-\mathbf{k}'}. \quad (3)$$

Here $\delta n_{\nu\mu k} = n_\nu(k) - n_\mu(k)$ denotes the nonequilibrium population difference between subbands ν and μ. Further details of the renormalized energies $e_{\nu\mu}$, electron-electron scattering broadening $\Gamma_{\nu\mu}$ and the Coulomb matrix elements $\tilde{V}^{\nu\mu}_{\mathbf{k}-\mathbf{k}'}$ are given in [21], [27]. The role of the dipole function $\wp_{\nu\mu}(k')$ as a k-space filter is then clear in Eq. (2). The overlap of a strongly peaked dipole function exactly where local population inversion $\delta n_{\nu\mu k'}$ takes place enhances the gain without inversion effect as illustrated by Fig. 2. More details of the numerical method are given in [18].

Figure 2 compares and contrasts the occupation functions and gain spectra for two different quantum wells. An important remark should be made before analyzing the numerical results. The gain spectra are inversely proportional to the period length. The period used here is the quantum well width, $L_c = 5$ nm or 10 nm consistently with the model. In actual QCLs, the period is extended to include barriers, injector, and collectors, and it can easily be at least 5 as large. Since the number of photons emitted per period remains constant this means that the gain spectra calculated here are very high in comparison to what should be expected in actual QCL structures. Note, however, that the qualitative analysis that follows is fully consistent with the single quantum well model.

The occupation functions are given on the left and corresponding gain spectra on the right panels for 5 nm (top) and 10 nm (bottom) GaAs-$Al_{0.3}Ga_{0.7}As$ quantum wells. Only the two top subbands are occupied and in both subbands the electrons are thermalized at $T = 300$ K. In both panels in the left the triangle symbols are the corresponding transition dipole moments for the (1,2) transition respon-

sible for the gain. The dipoles have been scaled in the y-axis to fit in the plot. The solid curve is the occupation of the upper subband, fixed in all gain calculations at $n_2 = 4 \cdot 10^{11}$ carriers/cm^2. From bottom to top, the dashed, dot-dashed and dotted curves are for increasing lower hole subband total occupation density $n_1 = 4$, 4.5 and $5 \cdot 10^{11}$ carriers/cm^2.

The corresponding gain curves are easily identified on the right and correspond to the densities in the second subband. In other words, from top to bottom, the dashed, dot-dashed and dotted curves are for increasing lower hole subband total occupation density $n_1 = 4$, 4.5 and $5 \cdot 10^{11}$ carriers/cm^2. The gain decreases as the lower occupation increases. The transition dipole moment, which has a strong filtering effect in k-space is scaled to fit in the plots. Following intuition, the gain decreases as the global occupation of the lower band increases. The 5 nm quantum well leads to better local inversion assuming that the same 2D density can be achieved in both samples and leads to more robust gain against increases in the lower subband carrier density.

However, in an actual structure it may be more difficult to inject the same global density of electrons for 5 nm then for 10 nm. More realistic calculations will be performed with a new generation of our NEGFs simulator, which will be capable of describing valence band cascaded structures and will be the subject of future publications.

In summary, this manuscript reviewed a possible solution for intersubband emitters that can lead to revolutionary devices for the completely open field of THz telecommunications and many other applications.

References

[1] S. Cherry, "Edholm's law of bandwidth", *IEEE Spectr.*, vol. 41, pp. 58–59, 2004.

[2] M. Koch, "IN-door THz communications: a vision for 2020", in *Terahertz Frequency Detection and Identification of Materials and Objects*, R. E. Miles *et al.*, Eds. Nato Security Through Science Series. Berlin: Springer, 2007, pp. 325–338.

[3] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser", *Science*, vol. 264, no. 5158, pp. 553–556, 1994.

[4] R. Köhler, A. Tredicucci, F. Beltran, H. E. Beere, E. H. Linfield, A. Davies, D. A. Ritchie, R. Iotti, and F. Rossi, "Terahertz semiconductor-heterostructure laser", *Nature*, vol. 417, no. 6885, pp. 156–159, 2002.

[5] S. Kumar, Q. Hu, and J. Reno, "186 K operation of terahertz quantum-cascade lasers based on a diagonal design", *Appl. Phys. Lett.*, vol. 94, no. 13, pp. 131105-1–1331105-3, 2009.

[6] B. S. Williams, "Terahertz quantum-cascade lasers", *Nat. Phot.*, vol. 1, no. 9, pp. 517–525, 2007.

[7] A. S. Zibrov, M. D. Lukin, D. E. Nikonov, L. Hollberg, M. O. Scully, V. L. Velichansky, and H. G. Robinson, "Experimental demonstration of laser oscillation without population-inversion via quantum interference and in RB", *Phys. Rev. Lett.*, vol. 75, no. 8, pp. 1499–1502, 1995.

[8] A. Wacker, "Coexistence of gain and absorption", *Nat. Phys.*, vol. 3, no. 5, pp. 298–299, 2007.

[9] R. Terazzi, T. Gresch, M. Giovanni, N. Hoyler, F. Faist, and N. Sekine, "Bloch gain in quantum cascade laser", *Nat. Phys.*, vol. 3, no. 5, pp. 329–333, 2007.

[10] J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, A. L. Hutchinson, M. S. Hybertsen, and A. Y. Cho, "Quantum cascade lasers without intersubband population inversion", *Phys. Rev. Lett.*, vol. 76, no. 3, pp. 411–415, 1996.

[11] G. Sun, A. Liu, and J. B. Khurgin, "Valence intersubband lasers with inverted light-hole effective mass", *Appl. Phys. Lett.*, vol. 72, no. 12, pp. 1481–1483, 1998.

[12] L. Friedman, G. Sun, and A. Soref, "SiGe/Si THz laser based on transitions between inverted mass light-hole and heavy-hole subbands", *Appl. Phys. Lett.*, vol. 78, no. 4, pp. 401–403, 2001.

[13] R. A. Soref and G. Sun, "Terahertz gain in a SiGe/Si quantum staircase utilizing the heavy-hole inverted effective mass", *Appl. Phys. Lett.*, vol. 79, no. 22, pp. 3639–3641, 2001.

[14] G. Dehlinger, L. Diehl, U. Gennser, H. Sigg, J. Faist, K. Ensslin, D. Grutzmacher, and E. Muller, "Intersubband electroluminescence from silicon-based quantum cascade structures", *Science*, vol. 290, no. 5500, p. 2277, 2000.

[15] L. Diehl, S. Mentese, E. Müller, D. Grützmacher, H. Sigg, U. Gennser, I. Sagnes, Y. Campidelli, O. Kermarrec, D. Bensahel, and J. Faist, "Electroluminescence from strain-compensated $Si_{0.2}Ge_{0.8}$/Si quantum-cascade structures based on a bound-to-continuum transition", *Appl. Phys. Lett.*, vol. 81, no. 25, pp. 4700–4702, 2002.

[16] R. Bates, S. A. Lynch, D. Paul, Z. Ikonic, R. W. Kelsall, P. Harrison, S. L. Liew, D. J. Norris, A. G. Cullis, W. R. Tribe, and D. D. Arnone, "Interwell intersubband electroluminescence from Si/SiGe quantum cascade emitters", *Appl. Phys. Lett.*, vol. 83, no. 20, pp. 4092–4094, 2003.

[17] D. J. Paul, "8-band $k \cdot p$ modeling of the quantum confined Stark effect in Ge quantum wells on Si substrates", *Phys. Rev. B*, vol. 77, no. 15, pp. 155323-1–155323-7, 2008.

[18] M. F. Pereira Jr., "Intervalence transverse-electric mode terahertz lasing without population inversion", *Phys. Rev. B*, vol. 78, no. 24, pp. 245305-1–245305-5, 2008.

[19] M. F. Pereira, "Valence intersubband gain without population inversion", *Centr. Eur. J. Phys.*, DOI: 10.2478/s11534-009-0061-5 (in press) [Online]. Available: http://www.springerlink.com/content/j9nm0k6179g142n6/

[20] M. S. Vitiello, G. Scamarcio, V. Spagnolo, B. S. Williams, S. Kumar, Q. Hu, and J. L. Reno, "Measurement of subband electronic temperatures and population inversion in THz quantum-cascade lasers", *Appl. Phys. Lett.*, vol. 86, no. 111115–111117, 2005.

[21] M. F. Pereira Jr., S.-C. Lee, and A. Wacker, "Controlling many-body effects in the midinfrared gain and terahertz absorption of quantum cascade laser structures", *Phys. Rev. B*, vol. 69, p. 205310, 2004.

[22] R. Nelander, A. Wacker, M. F. Pereira Jr., D. G. Revin, M. R. Soulby, L. R. Wilson, J. W. Cockburn, A. B. Krysa, J. S. Roberts, and R. J. Airey, "Fingerprints of spatial charge transfer in quantum cascade lasers", *J. Appl. Phys.*, vol. 102, no. 11, pp. 113104-1–113104-5, 2007.

[23] M. F. Pereira Jr, R. Nelander, A. Wacker, D. G. Revin, M. R. Soulby, L. R. Wilson, J. W. Cockburn, A. B. Krysa, J. S. Roberts, and R. J. Airey, "Characterization of intersubband devices combining a nonequilibrium many body theory with transmission spectroscopy experiments", *J. Mater. Sci. Mater. Electron.*, vol. 18, no. 7, pp. 689–694, 2007.

[24] A. Wacker, "Coherence and spatial resolution of transport in quantum cascade lasers", *Phys. Stat. Sol. C*, vol. 5, no. 1, pp. 215–220, 2008.

[25] T. Kubis, "Quantum theory of transport and optical gain in quantum cascade lasers", *Phys. Stat. Sol. C*, vol. 5, no. 1, pp. 232–235, 2008.

[26] G. Scalari, S. Blaser, J. Faist, H. Beere, E. Linfield, D. Ritchie, and G. Davies, "Terahertz emission from quantum cascade lasers in the quantum Hall regime: evidence for many body resonances and localization effects", *Phys. Rev. Lett.*, vol. 93, no. 23, pp. 237403-1–237403-4, 2004.

[27] M. F. Pereira Jr. and H. Wenzel, "Interplay of Coulomb and nonparabolicity effects in the intersubband absorption of electrons and holes in quantum wells", *Phys. Rev. B*, vol. 70, no. 20, pp. 205331-1–205331-8, 2004.

Mauro F. Pereira was born in Rio de Janeiro, Brazil. He received the B.Sc. in physics at PUC/RJ (1983) and the M.Sc. in physics (1985). He completed the Ph.D. in optical sciences at the Optical Sciences Center in Tucson/AZ in 1992 and received an equivalent Dr. Sci. degree in physics from UFRJ in 1993. He was a Research Associate at PUC/RJ, CBPF, Uni-Rostock, and the TU-Berlin, an invited lecturer in Bremen, an Associate Professor at UFBA and a Senior Researcher at Tyndall National Institute before joining the Materials and Electrical Engineering Research Institute of Sheffield Hallam University as a Professor. His expertise includes nonlinear and quantum optics, exciton and polariton effects, band structure engineering, many-body effects, semiconductor lasers including quantum cascade structures, nonequilibrium Green's functions, quantum transport and numerical methods. The results are aimed at fundamental understanding and as input for the design and simulation of novel optical and electrooptical devices.
e-mail: M.Pereira@shu.ac.uk
Materials and Engineering Research Institute
Sheffiled Hallam University
S1 1WB, Sheffield, United Kingdom

Paper

Impact of Signaling System Performance on QoE in Next Generation Networks

Jordi Mongay Batalla, Jarosław Śliwiński, Halina Tarasiuk, and Wojciech Burakowski

Abstract—The first experience of quality by multimedia applications' users takes place during the setup phase of a new connection. If the setup phase is not accepted or "slowly accepted", the confidence of the user decreases. The user becomes more sensitive when he/she pays the connections with assured quality of service (QoS). In this case, the process of call request should be also accomplished with QoS guarantees. This paper presents the signaling sub-system implemented within the EuQoS system. The EuQoS signaling process follows main assumptions of next generation networks (NGN) architecture and performs tasks related with codec agreement between multimedia end users, admission control and resource reservation functions. In this paper, we present analytical, simulation and experimental results showing the impact of signaling system performance on quality of experience (QoE) for the potential users of multi-layer EuQoS system. In particular, the presented approach aims at ensuring user QoE of the connection setup phase by ensuring QoS for transferring signaling messages by the network.

Keywords*—*call setup delay, class of service, heterogeneous networks, next generation networks, quality of experience, quality of service, signaling system.

1. Introduction

This paper deals with an impact of the signaling system on user's quality of experience (QoE) in next generation networks (NGN). In particular we focus on signaling system and its procedures implemented within the EuQoS[1] project [1]. The aim of the EuQoS system [2], [3] is to guarantee end-to-end quality of service (QoS) in heterogeneous multi-domain networks. For this purpose the EuQoS system combines a complete architecture and a full framework [3] to provide absolute end-to-end QoS guarantees for a number of end-to-end classes of service (E2E CoS) [4], [5]. The architecture considers heterogeneous network scenario, addressing the multi-core IP network, as well as, the current access network technologies, such as WiFi (wireless fidelity), xDSL (digital subscriber line), LAN/Ethernet (local area network/Ethernet) and UMTS (universal mobile communication systems). Moreover, the generic architecture is open to add new network technologies. The applications, for which finally QoS is provided are, among others, voice over IP, video on demand, data transfer, and interactive game. All of them require call setup phase before data transfer.

The call setup phase is accomplished by the EuQoS signaling system, which has been designed in compliance with the ITU-T recommendations about signaling requirements for IP QoS networks [6] and requirements for resource and admission control functions, as defined in scope of NGN activities [7]. The signaling system follows a "push mode" approach of NGN architecture [8], i.e., the system requires that in order to start the setup procedure [9], the applications must send resource reservation requests in an explicit way to the control plane architecture.

The setup procedure starts at sending the calling user's request to the system and finishes when receiving the corresponding response indicating that the new call can be admitted and the called user is ready to initiate the connection. The time between these two events is well known as call setup delay [10]. In [11], ITU-T imposes limitations on expected values of setup delay. Among other requirements, it states that for international connections and under normal load conditions [12], the setup delay for the 95% of the setup procedures should be less or equal to 11 s.

In fact, the first experience of quality by the multimedia application users takes place during the request of a new call. If the request is not accepted or "slowly accepted", the confidence of the user decreases. The notion of QoE is already evolving. In the last years, the ITU-T Standardization Group 12 (SG12) specified the QoE requirements as definable end-to-end parameters, which provide information not only about the network layer (as performed by the QoS parameters) but also about the transport and application layers behavior [13].

We follow this notion of QoE and in our studies the definable end-to-end parameter is the call setup delay as suggested in [14]. The aim of this paper is to show the impact of signaling system performance on setup delay in EuQoS system based on NGN architecture.

In order to guarantee target values of call setup delay, the system should dedicate some resources to the signaling process (setup and release procedures). These resources are both server and network resources. The server resources are intended for the processing of the signaling messages within the signaling servers, which manage new calls. Whereas, the network resources are intended

[1]EuQoS – IST 6 FR EU project "End-to-end Quality of Service Support over Heterogeneous Networks".

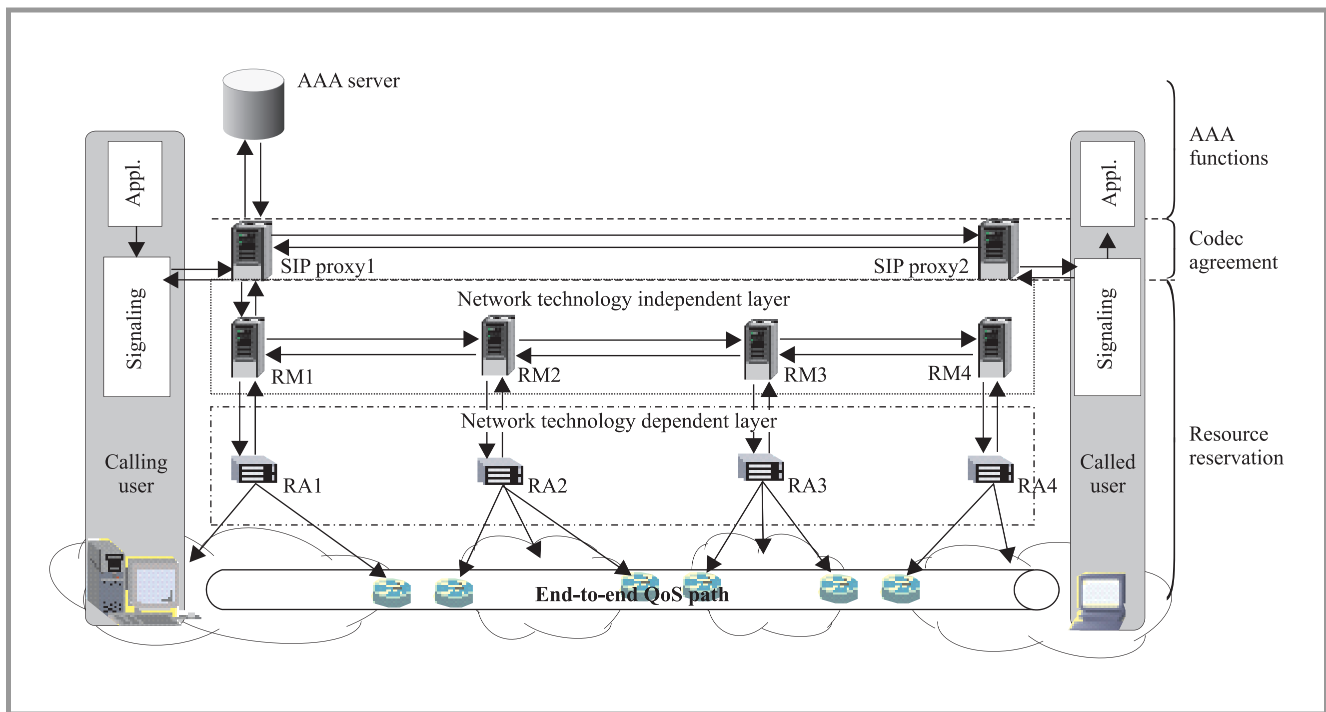


Fig. 1. The EuQoS system.

for transferring the signaling messages by the network between signaling servers. For reserving the necessary server and network resources, we should correctly dimension the signaling system. We propose to use the decomposition approach, which separately considers three phases of the signaling procedure in EuQoS system – all the operations performed:

– at the application layer, it means in session initiation protocol (SIP) proxies and authentication, authorization and accounting (AAA) servers for, e.g., voice over IP application;

– at the technology independent/technology dependent layer (TI/TD layer), for performing the admission control and resource allocation functions;

– at the network layer for transferring the signaling messages by dedicated signaling CoS.

Since the above three phases are distributed in time, i.e., they do not carry out simultaneously, we may enforce delay constrains for each one, so that the sum of the delay constrains of all the phases is less or equal to 11 s for the 95% of setup procedures. For illustration purposes, in our analysis we assume that the target setup delay value introduced by each of considered phases should be not greater than 3.5 s for the 95% of setup procedures.

Let us focus on the first phase of setup procedure, i.e., the operations performed at application layer. In previous studies of the EuQoS project [15], we obtained required performance parameters for the involved SIP proxy devices, which actually are less demanding than commercial ones. In fact, based on the literature we can observe that the SIP proxies are capable of running a high number of calls per second. For example, CISCO SIP proxies is able to handle about 100 call/s [16] and the direct routing approach allows even higher rates [17]. In the same way, the current AAA servers may manage until thousands of calls per second, e.g., the host intrusion detection system (HIDS) v3.1 of HP [18]. Measurements of mean establishment times of voice over IP calls related with AAA and SIP proxies functions were presented in [19]. Under normal load conditions, the mean establishment time is 0.94 s. Taking into account the state of the art, in our studies we focus on performance evaluation of the TI/TD layer and the transfer of signaling messages by the network (signaling class of service), which has been studied with less attention in the literature. These processes are the center of our studies in Sections 3 and 4, respectively. Previously, we present main features of the EuQoS system in Section 2. Finally, Section 5 summarizes the paper.

2. Overview of EuQoS System

The EuQoS system consists of three main layers, which are application layer, technology independent/technology dependent layer, and network layer as indicated in Fig. 1. The signaling procedure to set up a new connection (e.g., voice over IP) in the network is the following.

When any user sends a new call request to the EuQoS system, this initiates security functions with the AAA server.

These functions check if the user sending a call is authorized to use the EuQoS system. Next, the SIP proxy initiates the codec agreement with the called user. After the codec agreement, the first resource manager (RM) in the way checks whether the end-to-end QoS path to the called user exists and could provide the QoS guarantees required by the associated end-to-end CoS. The RM in the access domain periodically receives from the QoS routing protocol, the information about the QoS paths, i.e., the end-to-end paths with predefined QoS guarantees (in the form of target values for QoS parameters as IPTD, IPDV, IPLR). The QoS routing protocol is an enhanced version of border gateway protocol called EQ-BGP [2], [20], which builds end-to-end QoS paths on multi-domain network.

Afterwards, the system checks if, currently, there are available network resources (bandwidth, buffer capacities) to accept the call in consecutive domains. This is performed by the admission control functions implemented in resource managers and resource allocators (RA) as it will be indicated in the next section. If the process positively concludes, the system (SIP proxy) sends an affirmative response to the calling user allowing the communication.

Figure 1 presents the EuQoS signaling system, which is divided into the servers involved in AAA functions, the servers (SIP proxy) involved into the codec agreement and the servers (RM and RA) involved into the admission control function and resource allocation. The last ones, in turn, are divided into the servers, which are independent of the network technology (RM) and the servers, which functionalities depend on the network technology (RA). The purpose of this specialization is to allow each domain for custom implementation of its own QoS mechanisms without requiring other domains to be aware of their specific details.

Using the ITU-T terminology for NGN [21], [22], the SIP proxy would fulfill the service control function (SCF) at the service stratum; whereas, the resource manager and resource allocator would carry transport control function (TCF) at the transport stratum. The resource manager we would call policy decision functional entity (PD-FE), which is independent of the network technology and the resource allocator we would call the transport resource control functional entity (TRC-FE), which depends on the network technology.

A set of signaling elements (functions, databases, interfaces, etc.), as well as the signaling protocols involved in the different signaling layers were defined and implemented in the system. The diameter base protocol [23] communicates the SIP proxy with the AAA server in the access domain; the SIP protocol communicates end users and proxies, next steps in signaling (NSIS) protocol [24] connects the RMs and, at last, common open policy service (COPS [25]) communicates the RMs with the RAs and these ones with the network devices.

3. Performance Evaluation of TI/TD Layer

In this part, we analyze the impact of call handling scenario and the processing of messages at TI/TD layer on the part of the setup delay related to the TI/TD layer, which we denote as T_{ti-td}.
We focus on results of T_{ti-td} only for simulation studies, because the prototype implementation of RM and RA servers in the project was not optimized to take adequate conclusions. We present details about the call handling scenario at the TI/TD layer, a methodology to calculate T_{ti-td} for single and multi-domain scenario as well as the assumed simulation model. The aim of our studies is to show how many calls the system can handle respecting target values of T_{ti-td} (3.5 s for the 95% of setup procedures). In [15] we presented simulation results of performance evaluation on TI/TD layer. In this paper we enhance these studies showing detailed simulation results of analyzed queuing network.

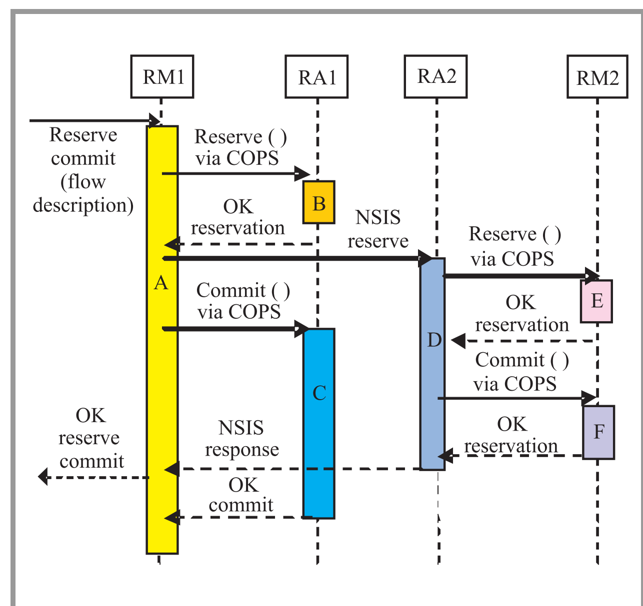


Fig. 2. Call scenario for two domains (domain 1: RA1, RM1; domain 2: RA2, RM2), one direction.

Figure 2 shows the horizontal (between RMs) and vertical (between RM and RA) signaling exchange for the successful setup of a call traversing two domains (domain 1 and domain 2). The call handling scenario is as follows: RM1 receives a QoS request and asks to the connection admission control (CAC) module in RA1 if there are enough resources to handle the new connection. If so, the requested resources are reserved and RA1 sends a confirmation to RM1. Next, RM1 forwards the QoS request to its peer RM2 and in parallel sends a request to RA1 to actually allocate the reserved resources in the associated access network equipment. When RM1 receives the confirmation from both RA1 and RM2, it replies to SIP proxy that the new call

can be admitted. As we can see in Fig. 2 the call handling scenario in RM2 and RA2 (i.e., in the egress domain) is the same as in RM1 and RA1 (ingress domain).

When the path has more than two domains, the resources are reserved also in transit domains. However, only the ingress domain RM checks whether an end-to-end path with feasible QoS exists. For bi-directional calls, the call handling process is performed in parallel during the setup procedure. So, in order to calculate the T_{ti-td} for a successful scenario, it is sufficient to simulate it only in one direction.

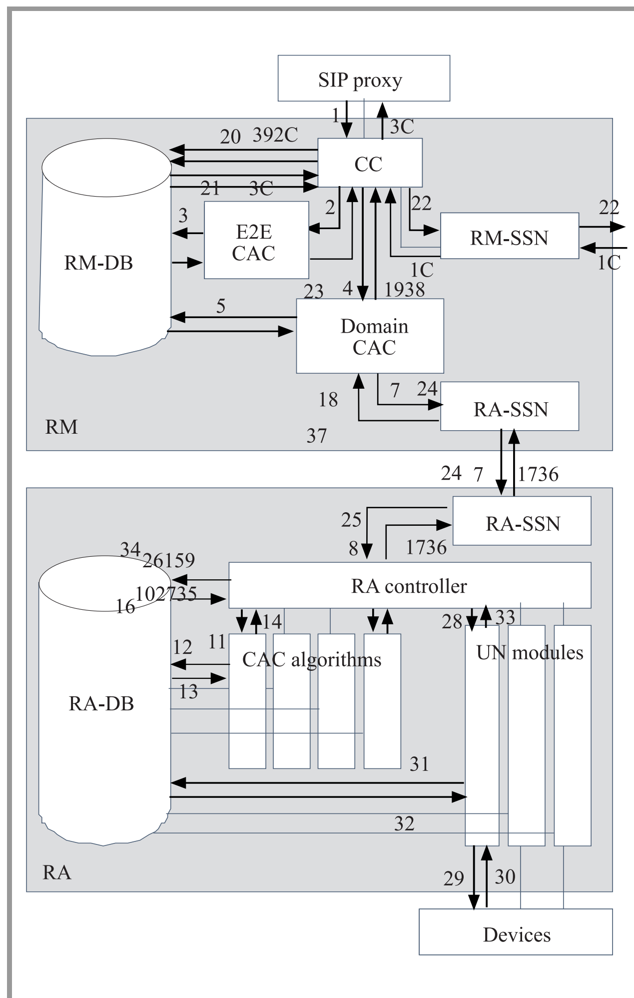


Fig. 3. The RM and RA architecture and call handling scenario for ingress domain calls and confirmations. Explanations: DB – data base, UN – underlining network, $1, 2 \ldots 39$ task numbers, C – confirmation task.

The detailed architecture of the RM and RA, together with the tasks identifications (task ID) are presented in Fig. 3. In the RM, the call controller (CC) receives QoS requests and controls all the CAC submodules. The end-to-end CAC (E2E CAC) is in charge of checking whether an end-to-end path with feasible QoS characteristics actually exists (by looking at its EQ-BGP routing information base). Then, the CC asks the domain CAC to check the operator policies and to look for an intra-domain path. The domain CAC has two submodules, one for the intra-domain part and the other one for the inter-domain link. Finally, the RA controller receives the QoS request via signaling module interconnecting the RM with the RA signaling and service negotiation (RA SSN). The RA controller runs a different CAC algorithm for each CoS. The RA CAC algorithms check the amount of available resources by querying the RA data base (RA-DB) and decide about the acceptation of the new call. If the call is accepted, the RA controller asks the appropriate access network UN module to configure its network devices for handling the new call. Different mechanisms must be configured depending on the type of network device (IP router, LAN/Ethernet switch, WiFi access point, etc.). Note that UN modules are involved in the call handling process only in *access* domains. For *transit* domains, resources are reserved basing only on the decision of the RA CAC algorithm, and there is no dynamic configuration of resources in inter-domain network devices.

We model the RM and RA architecture (Fig. 3) as the queuing network shown in Fig. 4. The simulation model is composed of a chain of servers, each one associated to an infinite-length FIFO (first in first out) queue. We distinguish eight types of servers: RM-CC (1), RM-DB (2), RM-RA link (3), RM-RM link (4), RA-C (5), RA-DB (6), RM-RA link (7), devices (8).

We assume that new calls and call acceptance confirmations independently arrive to the RM according to Poissonian processes with given mean arrival rates. Service processing times are the following deterministic values:

- RM-CC, RA-CC: $t_{RM-CC} = t_{RA-CC} = 100\ \mu$s;
- database access in the RM and RA: $t_{RM-DB} = t_{RA-DB} = 1$ ms;
- RM-RA link, RA-RM link, RM-RM link: $t_{RM-RA} = t_{RA-RM} = t_{RM-RM} = 1$ ms;
- access to the device: $t_{DEV} = 600$ ms (WiFi) and $t_{DEV} = 100$ ms (IP).

The processing times for WiFi and IP devices are based on test bed network equipment (LINKSYS WiFi access point WRT54G, CISCO 1841 router or Linux router kernel 2.6.18 IMQ). For other elements as call controllers or data bases, we tried to infer the expected processing time based on today's high speed servers.

We consider three event types in the simulation model: *new call arrival, new confirmation arrival*, and *departure of a task from a server*. Arrivals and confirmations invoke a sequence of tasks, whereas, the departure of a task from one server determines its arrival to another server. Furthermore, we distinguish call events processed at *ingress, transit* or *egress* domains. The maximum number of tasks is performed when processing calls at the ingress domain, since the E2E CAC is checked. Fewer tasks are required to

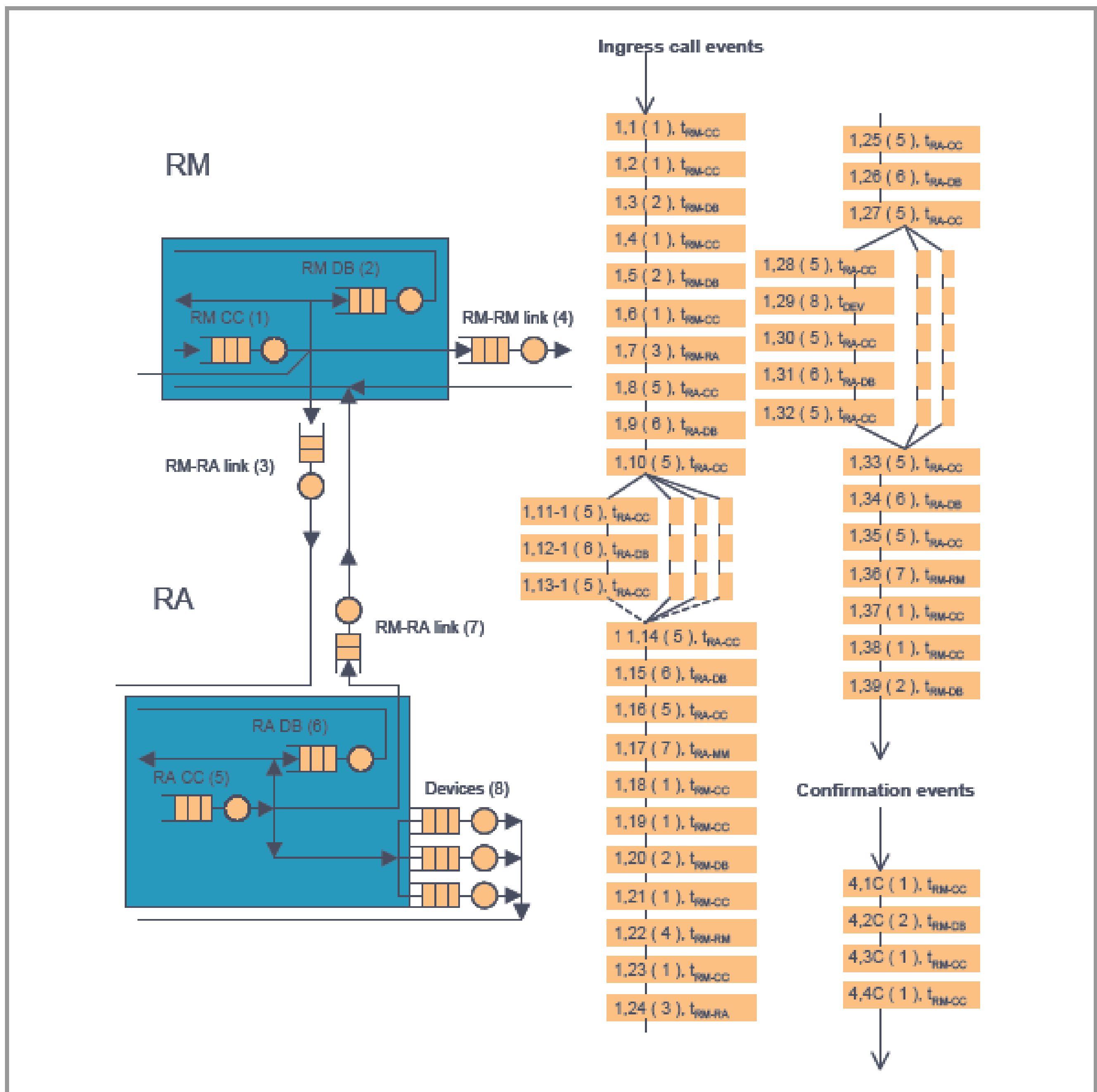


Fig. 4. The RM and RA for the ingress domain – simulation model.

process confirmation arrivals. The mean call arrival rates are $\lambda_{ingress}$, λ_{egress}, $\lambda_{transit}$, and the mean arrival rate of confirmations for ingress and transit domains are $\lambda_{conf-ingress}$ and $\lambda_{conf-transit}$. Due to space limitations, we only present events related to ingress domain.

Note that, in these simulations, we do not model the state machine of NSIS or COPS (see Fig. 2). The detailed message exchange of NSIS is modeled in simulation studies of signaling class of service (see Section 4). For COPS protocol, we model its performances by t_{RM-RA}, t_{RA-RM} times only.

The above simulation model has been coded in a discrete event simulator written in C++. Each event is represented by the quadruple <*Sequence ID, Task ID, Server ID, Processing time*>. The Sequence ID denotes whether the event is a call or a confirmation arrival.

We begin our studies with single domain performance evaluation. In particular, we simulate two scenarios. First, we assume that the single access domain handles both ingress and egress calls and confirmations assuming $\lambda_{ingress} = \lambda_{egress} = \lambda_{conf}$. We simulate WiFi and IP access domains in isolation. For each test we vary the call and confirma-

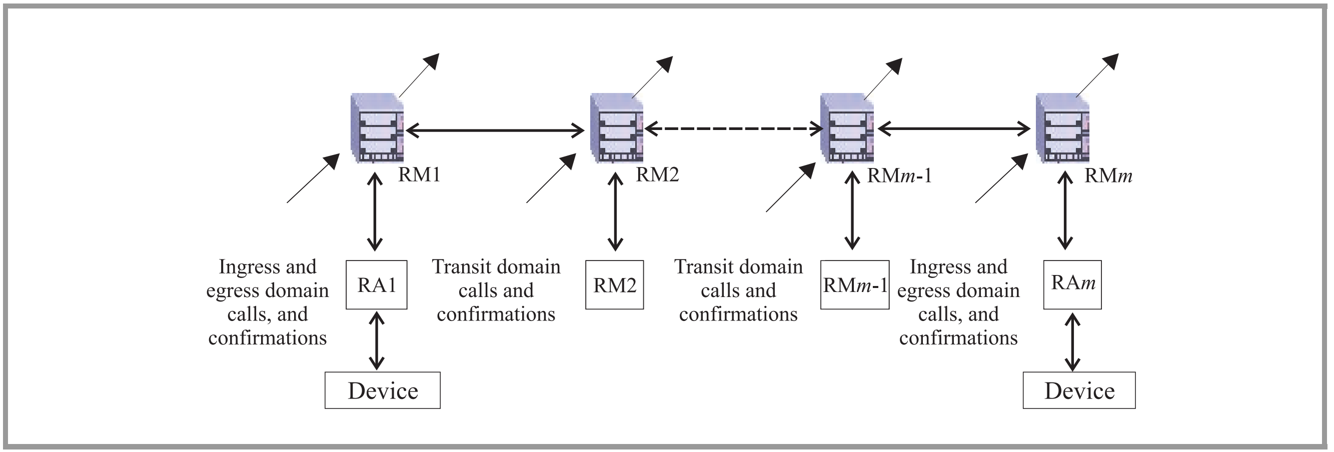


Fig. 5. Multi-domain call handling scenario with m domains, where $m = 2+n$, 2 access domains and n transit domains.

tion arrival rates up to the 95% resource utilization of the RM-RA elements (in this case the bottleneck element is the device itself).

In the next step, we analyze performances of single transit domain assuming that RM and RA handle only transit calls and confirmations assuming $\lambda_{transit} = \lambda_{conf}$; in this case the bottlenecks are RM-DB and RA-DB, it means the access to the databases.

Finally, based on the results obtained for single ingress/egress and transit domains, we are able to calculate the quantiles of T_{ti-td} for a multi-domain scenario (see Fig. 5).

For this purpose we distinguish among the following sequences of tasks performed by RM and RA in the ingress domain:

- $S1_{ingress}$: resource reservation tasks in RM1<–>RA1 performed before forwarding the request to the next RM;
- $S2_{ingress}$: tasks in RM1<–>RA1<–>DEV1, i.e., resource allocation tasks;
- $S3_{ingress}$: RM tasks invoked by confirmation arrival from neighboring domain.

Thus, we define the delay T_{ti-td} as follows:

$$T_{ti-td} = t^{S1}_{ingress} + \max\left\{t^{S2}_{ingress}, \left(n \cdot t_{transit} + t_{egress} + t^{S3}_{ingress}\right)\right\}, \quad (1)$$

where $t^{S\#}_{ingress}$ is the delay introduced by sequence S#, t_{egress} is the delay in the egress domain, n is the number of transit domains, $t_{transit}$ is the delay introduced by a transit domain.

Based on the detailed simulation results obtained for single domain scenarios (for ingress/egress and transit) we can derive:

$$t^{S2}_{ingress} < t^{S3}_{ingress} + t_{egress} + n \cdot t_{transit}. \quad (2)$$

By applying (1) and (2) the T_{ti-td} is calculated:

$$T_{ti-td} = t^{S1}_{ingress} + t^{S3}_{ingress} + t_{egress} + n \cdot t_{transit}. \quad (3)$$

In our studies, we calculate results when the access domain are both WiFi or both IP. The T_{ti-td} quantiles for 2, 10 and 20 domains is presented in Figs. 6–8.

In particular, Fig. 6 shows the results for two WiFi and two IP access domains. For both scenarios, the curves show a significant difference between the delays obtained for the two technologies. With reference to the 3.5 s target for the 0.95-quantile of T_{ti-td}, we observe that, in WiFi, $\lambda_{ingress}$, λ_{egress}, and λ_{conf} should not exceed about 0.6 call/s. In IP domain, the limit is about 4.75 call/s. For scenarios with 10 and 20 domains we assume that two domains are access domains while the others are transit domains.

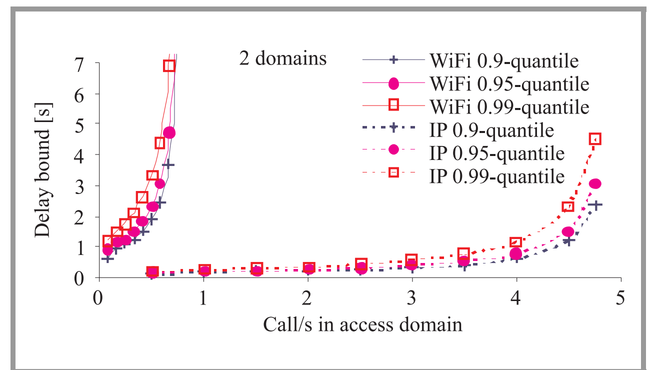


Fig. 6. Quantiles of T_{ti-td} for 2 WiFi or 2 IP domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress} = \lambda_{conf}$).

Figures 7 and 8 show the quantile values of T_{ti-td} for WiFi and IP access domains, respectively. The transit call rates are 233 or 300 call/s in each transit domain. The results show that the number of transit domains has smaller impact on the T_{ti-td} than the transit call arrival rate. Moreover,

acceptable call arrival rates are larger for IP access domains than for WiFi.

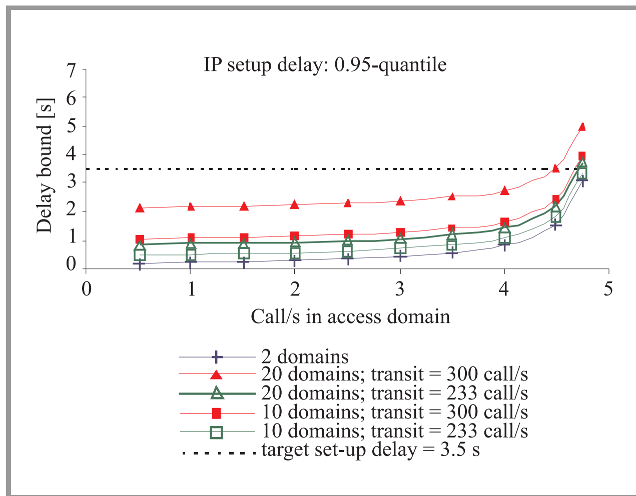


Fig. 7. The 0.95-quantile of T_{ti-td} for 2 IP access domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress}$ and $\lambda_{ingress} = \lambda_{conf}$).

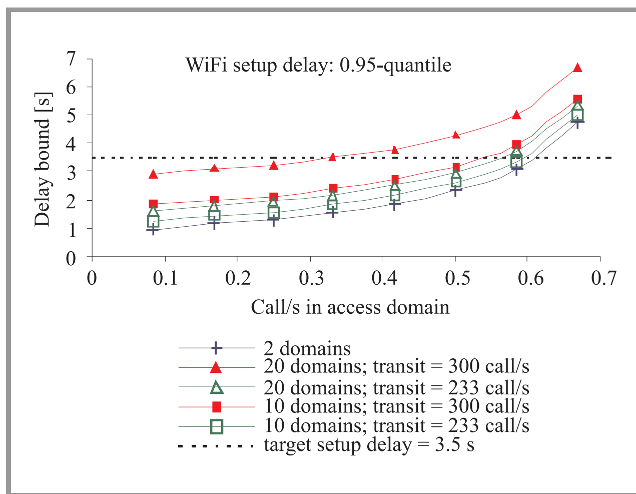


Fig. 8. The 0.95-quantile of T_{ti-td} for 2 WiFi access domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress}$ and $\lambda_{ingress} = \lambda_{conf}$).

Figure 9 shows some characteristics of the transit RM and transit RA. In particular, we collected characteristics of server utilization, mean delay in the queues, and mean queue sizes. As we can observe, the main bottlenecks of RM and RA elements for transit domains are RM-DB and RA-DB servers. The characteristics of these two servers are very close to each other.

Concluding, the signaling system at the TI/TD level is able to handle about 300 call/s in transit domains and it correctly scales with the number of transit domains (for values of 10 and 20 transit domains). An important open point is the call arrival process for NGN multi-service networks.

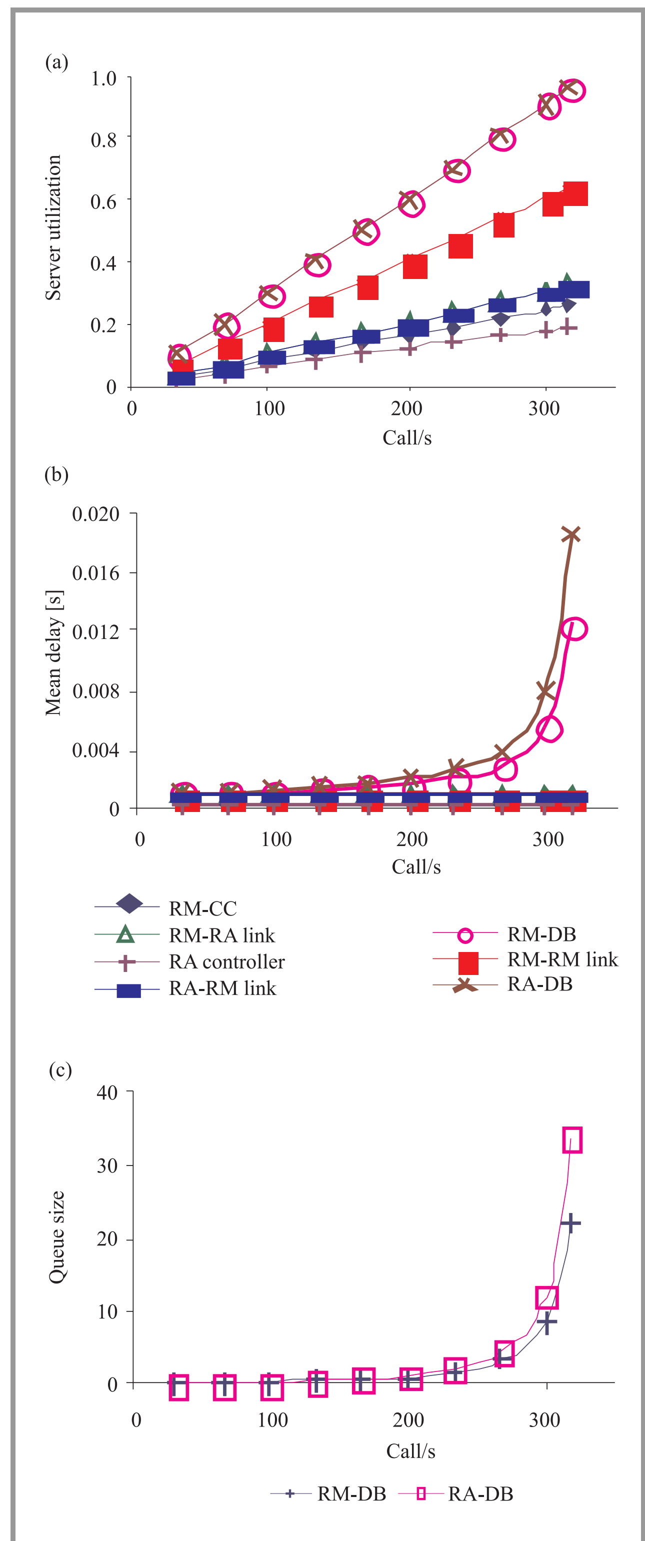


Fig. 9. Transit domain characteristics: (a) utilization of RM and RA servers, (b) mean delay, and (c) mean queue size for RM and RA elements versus transit call arrival rates ($\lambda_{transit}$) $\lambda_{transit} = \lambda_{conf-transit}$.

In these studies, we commonly considered telephony arrival model only but, in further studies, we are also facing up multi-service models.

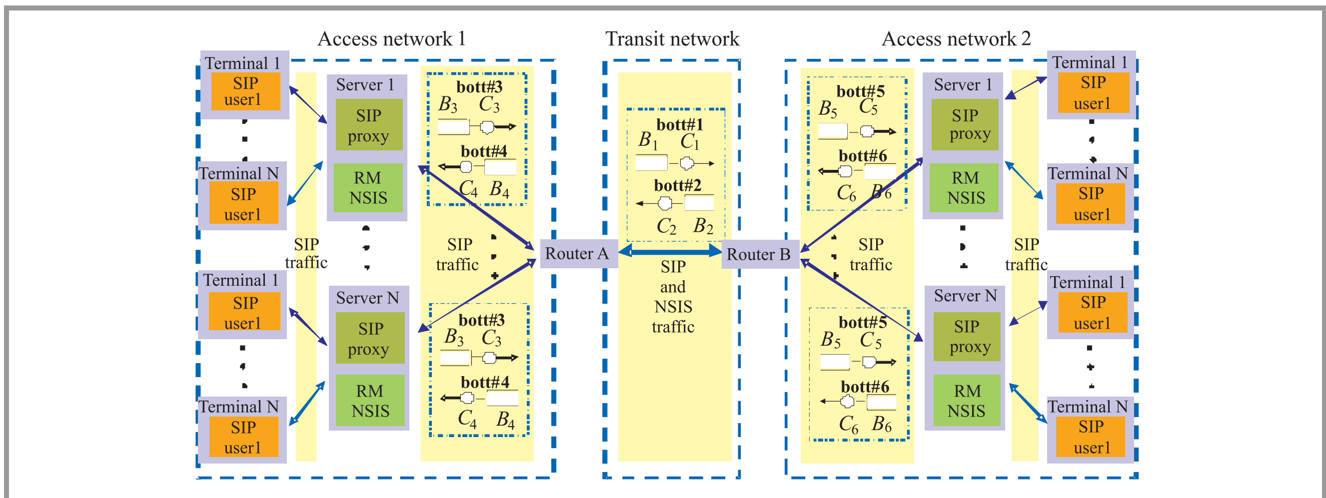


Fig. 10. Model of the EuQoS signaling system.

4. Performance Evaluation of the Signaling Class of Service

At the network layer we propose to implement the signaling class of service. The IETF document [4] defines, among others, the signaling class of service (S-CoS) designated to carry signaling traffic related with setup and release procedures. The objective of the S-CoS is to ensure target values of the part of the call setup delay related to the transfer of the signaling packets. We call this delay as *transfer packet call setup delay* or briefly T_s *delay*. We expect to obtain target values of T_s *delay* by transferring the signaling packets with adequate quality of service expressed in target maximum IPTD. In fact, according to [4] we can assume that the S-CoS tolerates the delay variation of the signaling traffic. One of the solutions to guarantee the QoS objectives for the signaling traffic is a correct resource provisioning [26].

For the aim of provisioning the S-CoS, we investigate the bottlenecks within the signaling path. The bottlenecks are the slowest links [27] in whose entrance packets of different setup procedures gather and where we should provision appropriate resources for each of the setup procedures.

To compute the necessary amount of bandwidth within the S-CoS for one setup procedure, we propose to ensure the same maximum IPTD (maxIPTD) in all the bottlenecks along the path. From the knowledge of the sizes of packets submitted to each bottleneck, and more precisely from the length of the longest burst of packets of the setup procedure submitted to each bottleneck we obtain the maxIPTD for 1 isolated setup procedure as (4)

$$\max IPTD = \frac{length\ of\ longest\ packet\ burst_{bott\ 1} \times 8}{C_{bott\ 1}} = \ldots = \frac{length\ of\ longest\ packet\ burst_{bott\ i} \times 8}{C_{bott\ i}}. \qquad (4)$$

The value of $C_{bott\ i}$ indicates the necessary amount of bandwidth for one setup procedure in the bottleneck i of S-CoS.

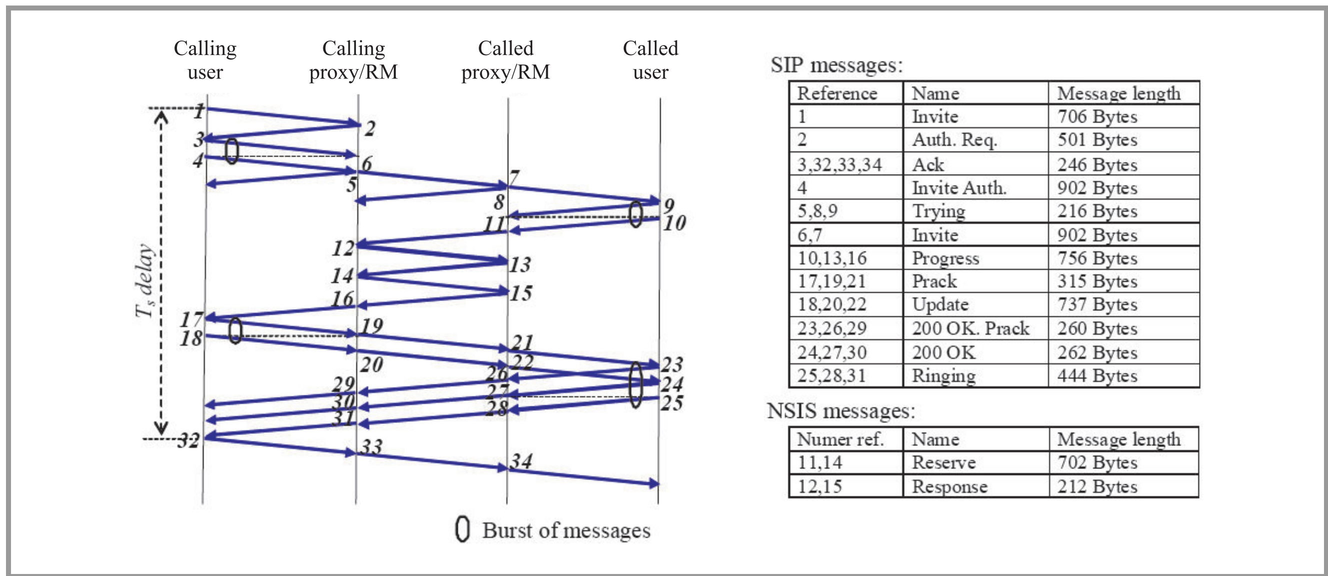
The equation, which completes the consistency of the system of linear equations presented in (4), comes from the value of T_s *delay*. In fact, we defined that the setup procedure should finish in a time equal to 3.5 s as presented in equation:

$$T_s\ delay = \sum_{all\ bottleneck\ s} \frac{\sum_{all\ packets\ in\ bottleneck} packet\ length \times 8}{C_{bott\ i}}. \qquad (5)$$

In the EuQoS system we assume that bottlenecks only may appear at the exit of SIP proxies and next inter-domain border routers as presented in the model of Fig. 10. Therefore, we consider only SIP and NSIS traffic, which is the unique signaling traffic in these bottlenecks.

As we may deduce from Eqs. (4) and (5), the necessary resources for signaling traffic ($C_{bott\ i}$) depend on the message sequence. Figure 11 shows the setup message sequence in the EuQoS system for which we implement the S-CoS. The lengths of the packets referred in Fig. 11 do not consider user data protocol (UDP), IP and link layer headers. Based on the above setup message sequence and assumptions and by the way of example, we will present further down simulation and measurement results.

Since packet losses cause retransmissions and, as an undesired result, increase delay of the whole setup procedure, we provision the buffer resources to ensure no losses in the queues (IPLR $= 0$) over normal operation of the network (no link failures). Since the biggest burst of packets in one setup procedure (EuQoS system) contains 2 packets (see Fig. 11) in the direction from calling to called and 3 packets from called to calling, then, the buffer size for M simultaneous setup procedures should equal $2 \times M$ packets in the calling-to-called direction and $3 \times M$ packets in the opposite direction.



Reference	Name	Message length
1	Invite	706 Bytes
2	Auth. Req.	501 Bytes
3,32,33,34	Ack	246 Bytes
4	Invite Auth.	902 Bytes
5,8,9	Trying	216 Bytes
6,7	Invite	902 Bytes
10,13,16	Progress	756 Bytes
17,19,21	Prack	315 Bytes
18,20,22	Update	737 Bytes
23,26,29	200 OK. Prack	260 Bytes
24,27,30	200 OK	262 Bytes
25,28,31	Ringing	444 Bytes

Numer ref.	Name	Message length
11,14	Reserve	702 Bytes
12,15	Response	212 Bytes

Fig. 11. SIP and NSIS message sequence in EuQoS system.

Let us remark that the transport protocol used to carry the signaling messages is not excessively important since, by provisioning the buffers, we assume no losses. Anyway, if we transport the signaling messages using the UDP protocol, we have to pay attention to the retransmission timers of the signaling protocols because these values are manually set and do not consider the round trip time (RTT) values of the network. The SIP protocol defines the default value of the retransmission timer equal to 500 ms when we use UDP protocol [28; Section 17.1.1.1]. The timers in the calling user will stop when the messages and its responses cross the whole network. This time is usually higher than 500 ms, so, such a value of retransmission timer results in the unnecessary retransmission of signaling packets (packets are not lost). This could cause a dangerous overload of the S-CoS. Therefore, in EuQoS implementation, we set the default values of SIP timers over UDP to 3 s.

4.1. Evaluation by Simulations

The performed tests aim at validating the proposed provisioning method to ensure target T_s *delay*. In our simulations we assume the same model as presented in Fig. 10. Moreover, we assume that the value N is equal to 5, this implies 5 servers and 25 terminals in each access network. The terminal i of the access network 1 (AN1) initiates a setup procedure with the terminal i of the access network 2 (AN2) and also the terminal i of the AN2 initiates another setup procedure with the terminal i of the AN1. All the terminals (50) initiate setup procedures at the same time and when the setup procedure corresponding to the terminal i finishes, then, the terminal i instantly initiates a new setup procedure. This is the worst-case scenario because finally, any new setup procedure finds the system full. Note that, in a real scenario, the setup procedures randomly arrive to the system and may find the system not completely full. Therefore, the values of T_s *delay* presented in these simulations are the upper bound.

Table 1
Bottleneck link capacity for 1 unique setup procedure

Initiated in AN1						
C [kbit/s]	C_1	C_2	C_3	C_4	C_5	C_6
bott#1–bott#6	38.3	32.3	49.6	32.3	38.3	42.3
bott#1–bott#4	28.7	24.2	37.2	24.2	–	–
bott#1–bott#2	15.8	13.3	–	–	–	–
Initiated in AN2						
bott#1–bott#6	32.3	38.3	42.3	38.3	32.3	49.6
bott#1–bott#4	21.4	25.3	28.0	25.3	–	–
bott#1–bott#2	13.3	15.8	–	–	–	–

Table 2
Bottleneck buffer size for 1 unique setup procedure

Initiated in AN1						
B [packets]	B_1	B_2	B_3	B_4	B_5	B_6
bott#1–bott#6	2	3	2	3	2	3
bott#1–bott#4	2	3	2	3	–	–
bott#1–bott#2	2	3	–	–	–	–
Initiated in AN2						
bott#1–bott#6	3	2	3	2	3	2
bott#1–bott#4	3	2	3	2	–	–
bott#1–bott#2	3	2	–	–	–	–

Table 3
Link capacities for 25 setup procedures initiated in AN1 and 25 initiated in AN2 (values set in the simulation scenario)

C [kbit/s] B [packets]	C_1 B_1	C_2 B_2	C_3 B_3	C_4 B_4	C_5 B_5	C_6 B_6
bott#1–bott#6	$25\times(38.3+32.3)=$ **1765.0** $25\times(2+3)=$ **125**	$25\times(32.3+38.3)=$ **1765.0** $25\times(2+3)=$ **125**	$5\times(49.6+42.3)=$ **459.5** $25\times(2+3)=$ **125**	$5\times(32.3+38.3)=$ **353.0** $25\times(2+3)=$ **125**	$5\times(38.3+32.3)=$ **353.0** $25\times(2+3)=$ **125**	$5\times(42.3+49.6)=$ **459.5** $25\times(2+3)=$ **125**
bott#1–bott#4	$25\times(28.7+21.4)=$ **1252.5** $25\times(2+3)=$ **125**	$25\times(24.2+25.3)=$ **1237.5** $25\times(2+3)=$ **125**	$5\times(37.2+28.0)=$ **326.0** $25\times(2+3)=$ **125**	$5\times(24.2+25.3)=$ **247.5** $25\times(2+3)=$ **125**	**100 000** **125**	**100 000** **125**
bott#1–bott#2	$25\times(15.8+13.3)=$ **727.5** $25\times(2+3)=$ **125**	$25\times(13.3+15.8)=$ **727.5** $25\times(2+3)=$ **125**	**100 000** **125**	**100 000** **125**	**100 000** **125**	**100 000** **125**

The simulation environment is ns-2 [29] where we developed modules to simulate the signaling process and integrated them into the EuQoS simulation model [30]. The integration of these modules permits its use in any network scenario (even in environments with different network techniques). In our own-implemented modules we model the signaling protocols: SIP and NSIS.

We introduce a bottleneck in the respective link by setting the appropriate value of the link capacity obtained from the provisioning method presented above. The links that are not considered bottlenecks in a certain test are set to 100 Mbit/s. Also the links between terminals and servers are 100 Mbit/s links. Next, we perform three tests, each one with 10 000 setup procedures. The number of bottlenecks change from one test to another. The first test considers 6 bottlenecks, from bott#1 and bott#6 as indicated in Fig. 10. The next test considers 4 bottlenecks (from bott#1 to bott#4) and the last one considers only the bottlenecks in the both directions of the link between router A and router B, i.e., bott#1 and bott#2.

We expect that each bottleneck will require a different provisioning, since the volume of signaling traffic submitted to them is different (see Fig. 11). Tables 1 and 2 show the necessary values of capacity in the bottleneck links and size of the bottleneck buffers for one unique setup procedure initiated in the AN1 and one unique setup procedure initiated in AN2 for the cases of 6, 4 and 2 bottlenecks in the signaling path. Values of this tables we exploit to complete Table 3, which presents the values set in the links and buffers of the simulation scenario considering that all the 50 terminals initiate setup procedures.

For each test, we calculate the time that packets last on transferring the 100 Mbit/s links. This time is not considered within the provisioning method and we subtract it from the total T_s *delay* value obtained in each test. The T_s *delay* values of the test (after subtraction) are presented in Table 4. Specifically, Table 4 shows the T_s *delay* of the shortest (min) and longest (max) setup procedure. In our simulations, we discard the first setup procedures of the tests because, at the beginning, the system is empty and these first setup procedures finish earlier, i.e., we only consider the T_s *delay* values of the setup procedures after striking the balance of the simulation process. We may observe that the values of T_s *delay* respect the target value equal to 3.5 s in all the cases. In opposition to the simulations performed in [26], in the presented simulation approach there is no multiplexing gain in the network because the system is permanently occupied. In fact, the setup procedures arriving to the system synchronize ourselves and transfer the network in the same order. This simulation approach allowed us to validate the provisioning method presented above.

Table 4
Value of T_S *delay* for the scenarios
with 6, 4 and 2 bottlenecks (simulation results)

Bottleneck	bott#1–bott#6 min/max	bott#1–bott#4 min/max	bott#1–bott#2 min/max
T_s *delay* [s]	3.499/3.500	3.499/3.500	3.500/3.500

4.2. Evaluation by Measurements

In this section we present measurement results of S-CoS performance evaluation. The aim of the measurements is to demonstrate that the S-CoS provisioned by the method presented in this paper may ensure target T_s *delay* in the test bed environment. On the other hand, we will also demonstrate that we cannot ensure delay requirements for the signaling traffic not carried by the S-CoS, if the network is heavy loaded.

Figure 12 presents the measurement test bed, where we emulate a scenario with two domains that are interconnected by an inter-domain link. The capacity of the inter-domain link equals 8 Mbit/s, while links inside each domain offer 100 Mbit/s capacity.

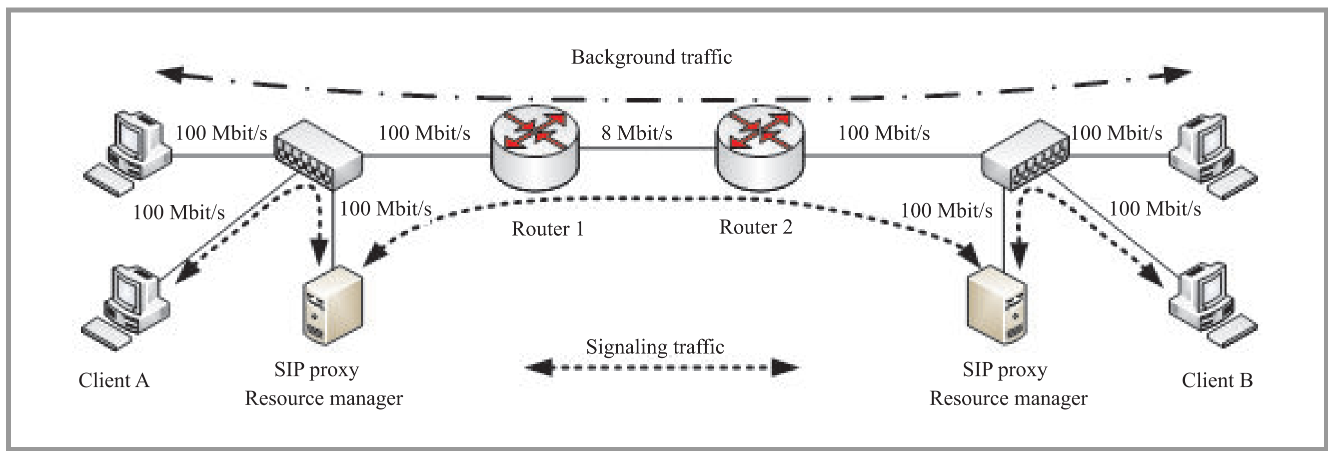


Fig. 12. Measurement test bed.

For implementing the S-CoS, all the signaling servers must "mark" their own packets before sending them into the network. This is done by setting the differentiated services code point (DSCP) field in the IP header of packets. The DSCP value assumed for the signaling packets is equal to binary value 101000 [4].

Figure 13 presents the model of the outgoing port of two routers: router 1 and router 2. In router 1 it is implemented in the direction from client A to client B and, in router 2, in the opposite direction.

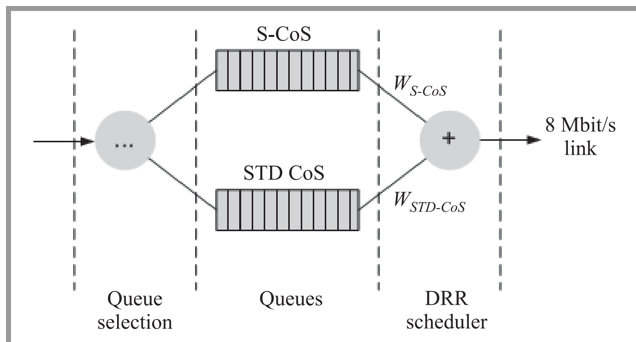


Fig. 13. Model of the outgoing port of the routers.

For our measurements, we consider only two classes of service: the S-CoS and the standard CoS (STD CoS), which is the best effort service. To ensure bandwidth separation between considered classes in IP router we use the deficit round robin (DRR) scheduler. We provision the S-CoS with 1000 and 2000 kbit/s in the two tests, respectively. The scheduler's weight for S-CoS traffic (w_{S-CoS}) is set to guarantee assumed amount of bandwidth, whereas, the scheduler's weight for STD CoS traffic ($w_{STD-CoS}$) complements the link capacity up to 8 Mbit/s.

By the provisioning method, we may calculate how many simultaneous setup procedures can be admitted, as a maximum, in the S-CoS not to exceed 1000 and 2000 kbit/s. When applying the provisioning method, we consider one unique bottleneck between proxy/RMs in the direction from client A to client B since the setup procedures are initiated only in client A. Moreover, we do not consider bottlenecks between clients and proxy/RMs because there is only signaling traffic. Table 5 presents the maximum number of simultaneous setup procedures for the provisioned S-CoS capacities, as well as the necessary buffer size set in the S-CoS to avoid losses for this number of simultaneous setup procedures.

Table 5
Class of service capacities and buffer sizes
(experiment configuration)

S-CoS capacity	1000 kbit/s	2000 kbit/s
Maximum number of simultaneous setup procedures	131	262
S-CoS buffer size	262 packets	524 packets

An artificial call generator in the client A initiates the setup procedures and maintains a constant number of them running in the system. When a setup procedure finishes, the call generator instantly initiates a new one.

On the other hand, the background traffic (STD CoS) is composed by 10 TCP connections in each direction, which fill the 8 Mbit/s link. We set the buffer size in the STD CoS equal to 1000 packets, which is sufficiently large to prevent any packet loss and introduces a high delay in the STD CoS traffic.

All measurements that we take, are performed at the signaling application level (T_s *delay*). We simplified the RM implementation with main focus on processing of signaling messages in order to reduce the delay introduced by state management operations, e.g., database access and storage and policy algorithms.

In the first experiment we compare the performance of signaling system with and without the S-CoS. We measure the T_s *delay* for cases with 131 or 262 running setup procedures, with capacity assigned to S-CoS equal to 1 Mbit/s or 2 Mbit/s, respectively. In Fig. 14 we show the maximum and minimum values of T_s *delay* measured in the test bed. The measurement consisted of at least 1000 setup procedures after warm up period.

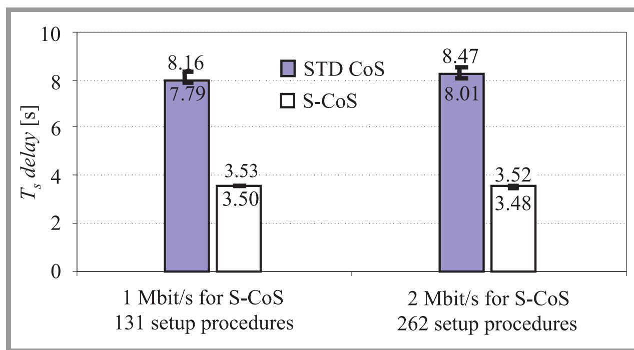


Fig. 14. Measurements of T_s *delay* within the EuQoS test bed.

By introducing the S-CoS in the network we are able to guarantee the value of the T_s *delay*; for both considered cases it is lightly higher than assumed target of 3.5 s. The tiny differences are due to the transfer of signaling packets by the not-bottleneck links. On the other hand, when signaling messages shares the capacity with the background traffic in STD CoS, the observed average T_s *delay* exceeds 8 s. As we expected, without the implementation of the S-CoS, we are not able to ensure the setup delay requirements for Internet telephony presented in [11]. Therefore, we argue that signaling traffic and best effort traffic must not be handled by the same network resources when the network is heavily loaded.

In the next experiment we evaluate the characteristics of T_s *delay* when the number of simultaneous setup procedures is lower than the calculated maximum value (for given S-CoS capacity). Figure 15 shows this characteristic for the case when 2 Mbit/s capacity is dedicated for S-CoS; we may see maximum and minimum measured values of T_s *delay*. The relation between T_s *delay* and the number of simultaneous setup procedures is linear, as we could expect. Moreover, for given number of simultaneous setup procedures, the T_s *delay* of different setup procedures does not vary so much (intervals are very small).

On the basis of the above results, we strongly recommend to introduce the dedicated class of service to ensure QoS at the packet transfer of signaling messages. Unfortunately, the inherent characteristics of the signaling traffic (bursts of packets, sporadic traffic, etc.) do not allow to effectively control it and, because of this, it is not possible to carry the signaling traffic by other CoSs on aggregation with other kinds of traffics.

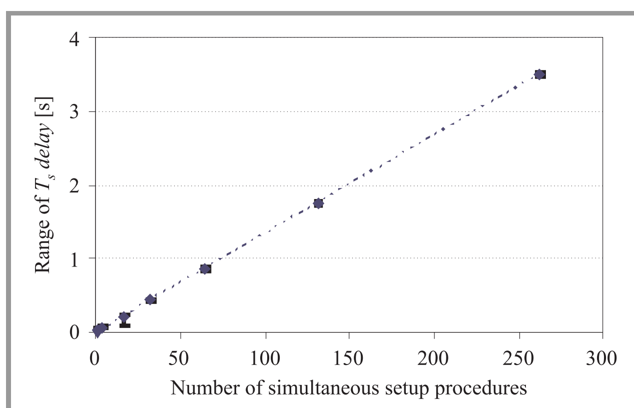


Fig. 15. Measured characteristics of T_s *delay* within the EuQoS test bed.

5. Conclusions

In this paper we studied the signaling performance of one example of next generation network system: the EuQoS system. We considered main processes inside the signaling system, as call handling scenario and transferring of the signaling messages by the network. Both these factors may affect into the call setup delay and, as a consequence, into the user quality of experience. We decided to confront them separately, using decomposition approach. As a result, we achieved a system capable of establishing end-to-end connections with the same requirements for call setup delay as assured in the ISDN network [10].

The key for the correct design of the system lies in reserving enough resources, both for signaling servers (processing power) and for the transfer of signaling messages in the network (bandwidth and buffer). For the signaling servers, we calculated the maximum call arrival rate (number of setup procedures per second) that the system can handle to assure given call processing delay. We considered all the signaling servers involved into the signaling system, paying special attention to the signaling servers designed to the resource reservation and allocation process over exemplary technologies. Following the analysis and relations between signaling servers, we evaluated system performance by means of dedicated simulation and measurement tools.

To complete the provisioning of resources, we investigated how the traffic generated by the signaling servers is handled in the network. We assumed that the number of simultaneous setup procedures handled by the signaling system is limited. Thanks to this assumption, we were able to provide dimensioning rules for signaling class of service. The rules provide a value of bandwidth and buffer space that should be dedicated for signaling traffic at the bottleneck points in order to guarantee given setup delay requirements. The devised provisioning method was verified by the simulation of the signaling system and by measurements in the test bed scenario.

Though we strived to do general analysis of the signaling system features, some results may be particular for the EuQoS system. Anyway, the presented proposals and methods can be used as a good guideline for the design of other signaling systems.

Future works in this field should be directed to find solutions for signaling congestion control, i.e., how to ensure a maximum number of simultaneous setup procedures within the system. It is also important to consider traffic models at call level, however, it requires an insight into measurements in multiservice networks. Other problem is related with the inherent complexity of the signaling systems in next generation networks. In this work, we con-

sidered only the main signaling servers in each domain, but in reality we should expect multiple ways of system decomposition and distribution.

Acknowledgements

We would like to thank all the partners of the EuQoS consortium for their cooperation and their work on developing the EuQoS system.

References

[1] "The EuQoS Consortium" [Online]. Available: http://www.euqos.eu/

[2] X. Masip-Bruin *et al.*, "The EuQoS system: a solution for QoS routing in heterogeneous networks", *IEEE Commun. Mag.*, vol. 45, no. 2, pp. 96–103, 2007.

[3] W. Burakowski *et al.*, "Provision of end-to-end QoS in heterogeneous multi-domain networks", *Ann. Telecommun. Springer*, vol. 63, iss. 11, p. 559, 2008.

[4] J. Babiarz and F. Baker, "Configuration Guidelines for DiffServ Service Classes". Internet RFC 4594, Aug. 2006.

[5] K. Chan, J. Babiarz, and F. Baker, "Aggregation of Diffserv Service Classes". Internet RFC 5127, Febr. 2008.

[6] "Signaling Requirements for IP QoS", ITU-T TR Q-Series Supplement 51 (10/2004).

[7] "Resource and admission control functions in next generation networks", ITU-T Rec. Y.2111 (09/2006).

[8] "General overview of NGN", ITU-T Rec. Y.2001 (10/2004).

[9] "Framework for the network management of IP-based networks", ITU-T Rec. E.417 (02/2001).

[10] "Network grade of service parameters and target values for circuit-switched services in the evolving ISDN", ITU-T Rec. E.721 (05/1999).

[11] "Network post-selection delay in PSTN/ISDN networks using Internet telephony for a portion of the connection", ITU-T Rec. E.671 (03/2000).

[12] "Network grade of service parameters and target values for maritime and aeronautical mobile services", ITU-T Rec. E.774 (10/1996).

[13] "Quality of experience requirements for IPTV services", ITU-T Rec. G.1080 (06/2008).

[14] A. Moorsel, "Metrics for the Internet age: quality of experience and quality of business", in *Fifth Perform. Worksh.*, Nuremberg, Germany, 2001.

[15] H. Tarasiuk *et al.*, "Designing the simulative evaluation of an architecture for supporting QoS on a large scale", in *Proc. QoSim 2008 Conf.*, Marseille, France, 2008 (to appear ACM Digital Library).

[16] CISCO systems document [Online]. Available: http://www.cisco.com/en/US/tech/tk652/tk701/-technologies_white_paper09186a00800a9818.shtml.

[17] V. Matic, I. Franicevic, and D. Sekalec, "Parallel SIP proxy servers using direct routing approach", in *Proc. Int. Conf. Softw. Telecommun. Comput. Netw. SoftCOM 2006*, Split-Dubrovnik, Croatia, 2006.

[18] "Host Intrusion Detection System (HIDS) v3.1. Sizing Guidelines and Tuning Primer", HP Company internal document, Sept. 2005.

[19] Y. Bai *et al.*, "A novel distributed wireless VoIP server based on SIP", in *Proc. Multimed. Ubiquit. Eng. Conf. MUE'07*, Seoul, Korea, 2007.

[20] A. Beben, "EQ-BGP: an efficient inter-domain QoS routing protocol", in *Proc. 20th IEEE Int. Conf. Adv. Inform. Netw. Appl.*, Los Alamitos, USA, 2006, pp. 560–564.

[21] N. Morita, H. Imanaka, O. Kamatani, T. Oba, and K. Tanida, "Overview and status of NGN standardization activities at ITU-T", *NTT Tech. Rev.*, Nov. 2007.

[22] J. Song *et al.*, "Overview of ITU-T NGN QoS control", *IEEE Commun. Mag.*, vol. 45, no. 9, pp. 116–123, 2007.

[23] P. Calhoun *et al.*, "Diameter Base Protocol". Internet RFC 5127, Sept. 2003.

[24] R. Hancock *et al.*, "Next Steps in Signaling (NSIS): Framework". Internet RFC 4080, June 2005.

[25] D. Durham *et al.*, "The COPS (Common Open Policy Service) Protocol". Internet RFC 2748, Jan. 2000.

[26] J. Mongay Batalla and R. Janowski, "Provisioning dedicated class of service for reliable transfer of signaling traffic", in *Proceedings 20th International Teletraffic Congress, June 2007, Ottawa, Canada*, Lecture Notes in Computer Science, vol. 4516. Berlin-Heidelberg: Springer: 2007, pp. 853–864.

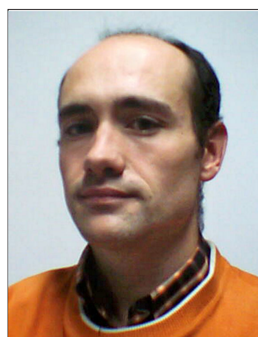
[27] P. Rodriguez and E. W. Biersack, "Dynamic parallel access to replicated content in the Internet", *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 455–465, 2002.

[28] J. Rosenberget *et al.*, "SIP: Session Initiation Protocol". Internet RFC 3261, June 2002.

[29] "The Network Simulator ns-2" [Online]. Available: http://www.isi.edu/nsnam/ns/

[30] "The ns-2 signaling modules" [Online]. Available: http://tnt.tele.pw.edu.pl/include/tools/sim-euqos-ptl.tgz

---



Jordi Mongay Batalla was born in Barcelona, Spain, in 1975. He received the M.Sc. degree in telecommunications from Universitat Politècnica de València in 2000. He worked one year in Centro Nazionale di Astrofisica in Bologna, Italy, as research scientist specializing in qualities and capacities of diffserv networks. Nowadays, he is finishing Ph.D. studies in the Warsaw University of Technology, Poland. He is with Telecommunications Network Technologies (TNT) Group from 2004. His research interest focus mainly on quality of service in diffserv networks and next generation network architecture.
e-mail: jordim@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Jarosław Śliwiński received the M.Sc. and Ph.D. degrees from the Warsaw University of Technology, Poland, in 2003 and 2008, respectively. His research interests cover traffic control, systems' design and implementation methodology.
e-mail: jareks@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Halina Tarasiuk received the M.Sc. degree in computer science Szczecin University of Technology, Poland, in 1996 and Ph.D. degree in telecommunications from the Warsaw University of Technology, in 2004. From 1998 she is with Telecommunication Network Technologies Group at the Institute of Telecommunications, Warsaw University of Technology. In 2003 as a member of the group she received Rector's Award for scientific achievements. From 2004 she is an Assistant Professor at the Warsaw University of Technology. From 1999 to 2003 she was collaborated with Polish Telecom R&D Centre. She participated in several European and national projects (2004–2008). Her research interests focus on NGN and NWGN architectures, node and network virtualization, signaling system performance, admission control and resource allocation methods and queueing mechanisms. She is the author and co-author of more than 40 research papers presented in books, journals, and conference proceedings and about 30 technical reports.

e-mail: halina@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Wojciech Burakowski was born in Warsaw, Poland, in 1951. He received his M.Sc., Ph.D. and D.Sc. degrees in telecommunications from the Warsaw University of Technology in 1975, 1982 and 1992, respectively. Now he works as Full Professor at the Institute of Telecommunications, Warsaw University of Technology and the National Institute of Telecommunications, Warsaw. He leads TNT research group. Since 1990 he has been involved in many COST and EU Framework Projects. He is a member of Telecommunications Section of the Polish Academy of Sciences and an expert in 7 FR Programme. He was a chairman and a member of many technical programme committees of national and international conferences. He is the author or co-author of about 170 papers published in books, international and national journals and conference proceedings and about 70 technical reports. His research areas include new networks techniques, ATM, IP, heterogeneous networks (fixed and wireless), network architecture, traffic engineering, simulation techniques, network mechanisms and algorithms.

e-mail: wojtek@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Recommendations and Regulations of the European Commission Regarding the Pan-European eCall

Wojciech Michalski

Abstract— The paper presents the main actions conducted by the European Commission in the context of the eCall programme (the initiatives of Driving Group), the recommendations and requirements for introduction of the pan-European eCall (architecture, handling procedures, minimum set of data (MSD) content, performance criteria, etc.) as well as regulations concerning the pan-European eCall, particularly the status of the eCall project covering also the legal situation of eCall in the Member States including Poland.

Keywords— eCall, eSafety, in-vehicle system, minimum set of data.

1. Introduction

According to investigation and estimates made by the European Commission (EC) around 40, 000 people were killed and more than 1.7 million people injured in 2008 in road accidents across the European Union (EU). Road fatalities in EU have fallen by more than 27% since 2001, when EC published its *White Paper on European Transport Policy* [1]. Significant impact on this positive trend had *European Road Safety Action Programme* [2] and the *Intelligent Car Initiative* [3]. However, these fatality numbers indicate that the current actions towards reducing the number of accidents are not sufficient. Therefore the EC proposed that EU should set the target of decreasing the number of road fatalities by 50% by 2010 and start actions described in this paper.

2. The eCall Definition

The eCall is an emergency call generated automatically via activation of in-vehicle sensors or manually by vehicle occupants. When activated, the in-vehicle eCall system (IVS) will establish a voice connection directly with the relevant public safety answering points (PSAP). At the same time, a minimum set of accident data (MSD) will be sent to the eCall operator receiving the voice call. They will be sent during the call processing or immediately after the set up of emergency call.

The eCall is treated as a part of the E112 service (supplement of TS12) concerning data transmission from the IVSs to the PSAPs, especially including accurate location information (e.g., precise location, time and type of the accident).

The eCall is also the name of the emergency call project developed by the EC Member States to obtain a pan-European solution.

The eCall itself will not reduce the number of accidents but it is expected to improve response times in case of traffic accident and save lives by faster help.

3. The eCall Initiatives Conducted by the European Commission

From the beginning the European Commission has been coordinating actions concerning eCall and originated several initiatives related to it. The first initiatives addressed establishment of the eSafety Forum as well as working groups. Moreover, the EC organized high level meetings, elaborated regulations and introduced the eCall Memorandum of Understanding (MoU).

3.1. The eSafety Forum

Establishment of the eSafety Forum was a fundamental action of the EU concerning the eCall. It was a joint industry and public initiative for improving road safety by using information and communication technology (ICT). The main idea was to join forces and create a European strategy to accelerate the research and development, deployment and use of intelligent integrated safety systems including advanced driver assistance systems (ADAS) to improve road safety in Europe.

3.2. The Driving Group eCall

One of the working groups acting under the eSafety Forum is the eCall Driving Group (DG), established at the end of 2002, which identified the key stakeholders involved in the eCall process and outlined the functionalities of the interfaces to be established between the stakeholders. The DG eCall classified its members into the following four categories:

- automotive industry;
- mobile telecommunication industry;
- public emergency authorities and associated or cooperating service organizations;
- public social security organizations, private insurance companies and automobile clubs.

The DG eCall consists of several working subgroups established to solve different issues related to both the service chain and value chain. Objectives to solve by these subgroups include:

- performance criteria related to the eCall chain;
- functional requirements and the specifications for the eCall generator;
- PSAP requirements regarding receiving and handling eCall;
- cost and benefits for the insurance industry;
- cost of the in-vehicle system;
- overview of available studies.

Currently, DG eCall has 138 members.

3.3. Regulations on the Emergency Calls and High Level Meetings

Another initiative of the EC was publication of the directives, communications, recommendations as well as requirements on the emergency calls (112, E112 and eCall) implemented in the EU (see [1]–[17]).
Moreover, under the auspices of the eSafety Forum, 28 recommendations were developed on how the road safety could be improved through new technologies. These recommendations have led to establishment of dedicated working groups developing requirements towards the implementation of specific technologies or applications.
The DG eCall together with European Commission organize meetings in order to build consensus among participants, in particularly to specify the system, define the functional architecture as well as solve organizational issues. Some of them were dedicated to defining a road map for implementation of the eCall, some as a reaction to the communications from the EC calling for the Member States to sign the MoU and take necessary steps to deploy the eCall.

3.4. The eCall Memorandum of Understanding

The MoU was introduced in the half of 2004 by the DG eCall to actively investigate eCall solution and business cases. The main purpose of this initiative was acceleration of deployment as well as implementation of eCall built on the single pan-European emergency call number 112 by the end of 2010.
The MoU comprises necessary arrangements for implementing the eCall action plan and sets out the measures to be taken by the EC, Member States, automotive industry, telecoms and insurance industry.

4. The eCall Service Chain

The DG eCall identified six domains with separate responsibilities and tasks for the eCall service chain [4]. These domains are depicted in Fig. 1.

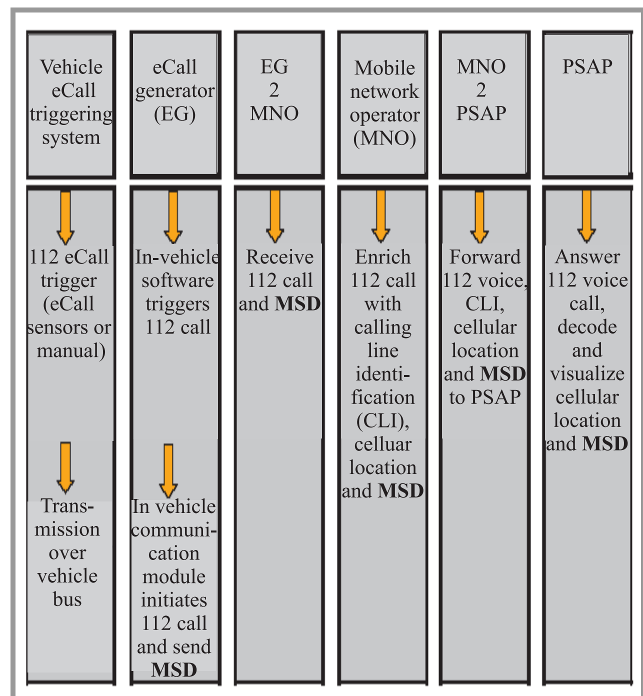


Fig. 1. The domains of the eCall service chain [4].

5. The eCall Architecture and Handling Procedure

The architecture of eCall system recommended by the DG eCall is illustrated in Fig. 2.

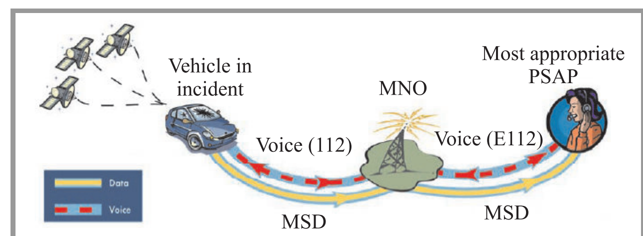


Fig. 2. The eCall system architecture [4].

Procedure for handling eCall:

- In the event of an accident the eCall generator should initiate eCall by sensor triggering and send eCall consisting of a pure voice (audio) telephone call based on 112 and the minimum set of data to the PSAP in automatic or manual manner.
- Handled by the mobile network, the eCall (voice and data) should be recognized as an 112 emergency call and enriched by mobile network operator (MNO) with the calling line identification (CLI) as well as location information (based on the best effort principle) and delivered to the appropriate PSAP.
- If MSD has been properly received, the PSAP should transmit an acknowledgement to the eCall generator.

Table 1
Revised MSD [5]

Offset	Name	Size [byte]	Type	Unit	Description
0	Control	1	Bitfield		Bit 7: 1 = automatic activation Bit 6 : 1 = manual activation Bit 5 : 1 = test call Bit 4 : 1 = no confidence in position Bit 3 – 0 : 1 = reserved
1	VIN	15	Bitfield		Up to 20 VIN characters stored using 6 bits per character with no padding
16	Timestamp	4	Unsigned integer	Seconds	Timestamp of incident event. Seconds elapsed since midnight, January 1st, 1970 UTC
20	Latitude	4	Signed integer	Milliarcsecs	GNSS position latitude (WGS84)
24	Longitude	4	Signed integer	Milliarcsecs	GNSS position longitude (WGS84)
28	Direction	1	Unsigned integer	Degrees	Direction of travel (based on last 3 positions)
29	Service provider	4	Unsigned integer	IPv4	Service provider IP address (optional)
33	Optional data	107	String	To be defined	Additional data, i.e., crash information (optional)
	Sum	140			

Explanations: VIN – vehicle identification number, UTC – universal coordinate time, a.k.a. Greenwich mean time, GNSS – global navigation satellite system, WGS84 – world geodetic system 1984 (a standard for identifying global position).

The eCall should be a "sleeping" application run by the eCall generator which acts only when the generator detects an incident serious enough for triggering an automatic eCall or passengers in the vehicle generate a manual eCall.

5.1. The eCall Activation

An automatic eCall trigger signal should be assigned to different crash types, e.g., front, rear, side and roll crashes. It should be generated by the airbag control module and associated or not with other sensors data (e.g., radar, speed, gyro, axle load). The eCall generator is responsible for sending an automatic eCall trigger signal to PSAP and the vehicle manufacturers are responsible for determination of the parameters of this signal. It is required that a number of false eCalls sent by the eCall generator shall be reduced to minimum, because activation of emergency calls should be safe and robust.

Manual activation depends on the specific human machine interface for the eCall generator. The eCall may be initiated by holding the eCall button down for three seconds or pushing the button twice within five seconds. The eCall system in vehicle should be designed in a way minimizing probability of unintended actions and number of false calls.

5.2. The eCall Minimum Set of Data

The DG eCall recommends that the MSD includes important information to help send the emergency services to the site of the accident and to speed up the response. The first proposal of MSD content for eCall was given in [4]. The revisions to this proposal given in [5] are illustrated in Table 1.

The proposed revisions include definitions of: signed and unsigned status of integers, big-endian byte order, encoding for degrees as well as an initial time for the timestamp field. According to these revisions, the last 4 bits of control field should be zero. The reserved bits should be cleared to allow future MSD receiving implementations to detect an earlier version MSD. The vehicle identification number (VIN) is stored using 6 bits per character and a terminating character defined to distinguish between VINs of differing lengths. They also specify requirement added to pad the service provider field when optional data present. This

field contains the IPv4 address of service provider stored as series of 4 1-byte, unsigned integers. Optional data field should not be padded out to its full length. Although this field may contain up to 107 bytes, unused bytes in the field do not need to be padded to reach the full length of 107 bytes. All of these revisions are described in greater detail in [5].
According to DG eCall recommendation, the Committee European de Normalization (CEN) is responsible for finally standardization of MSD content.

5.3. The PSAP Structures and Scenarios

Emergency call structure may include two levels of PSAP (PSAP 1 and PSAP 2) served by the same public body; where PSAP 1 means the first point of contact for eCall and PSAP 2 is the actual emergency operator handling the emergency situation. It may also be built as publicly operated PSAP 1 as well as service provider or telecom operator operating as PSAP 1 under control of emergency agency/public authority. To minimize the necessary investment, the DG eCall recommends that current PSAP structures were revised, e.g., through public/private partnerships.

5.4. Performance Criteria

The DG eCall recommends end-to-end performance criteria related to timing, quality of service (QoS), PSAP, mobile networks, location and map accuracy. In the DG eCall recommendations the experiences from comparable automatic and manual vehicle emergency or assistance calling systems and current PSAP systems and emergency response systems have been taken into account.
Short response time is critical for the efficient handling of the eCall. Performance criteria related to end-to-end timing in the eCall service chain are illustrated in Fig. 3.

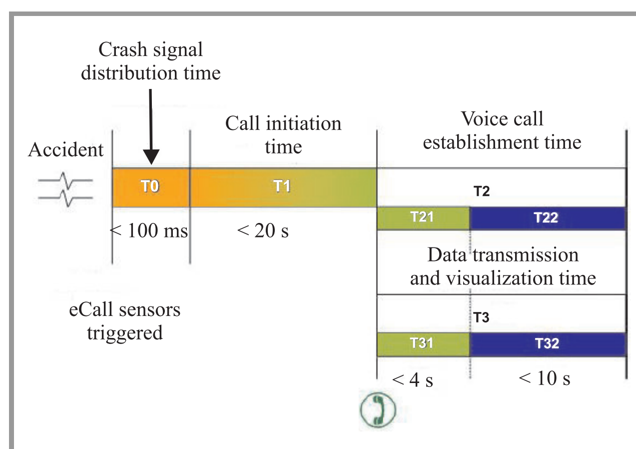


Fig. 3. End-to-end criteria for eCall service chain [4].

The end-to end QoS specifications include a requirements that 85% of all activated and sent eCalls should successfully reach the PSAP by 2010, than 89% all of them by 2015 and 92% by 2020. Moreover, 90% of all accidents of severity crossing the thresholds for triggering an eCall should be successfully reported by the eCall generator to the mobile network by 2010, 95% by 2015 and 98% by 2020. The eCall generator should be reprogrammable in order to change the MSD structure, e.g., service provider and vehicle identification fields.
For the eCall as the pan-European emergency service, the PSAPs in any Member States should handle 99% of all received calls.
Mobile network operators should provide full roaming capabilities in their core networks and treat the eCall as a 112 call with the same priority and reliability.
Precise location of the vehicle involved in an accident should be based on the best performance that satellite based location can provide at any time. For correct location a satellite system should currently guarantee accuracy equal or better than 50 m (for 50% of cases) and equal or better than 150 m (for 95% of cases).
For mapping accuracy, completeness of the road geometry should be kept down to the lowest local level and accuracy of the road geometry should be precise to 15 meters. No less than 99.9% roads of categories 1 to 4 must have a name; for category 5 the requirement is 97%.

5.5. Privacy

All Member States involved in development of the eCall should be obliged to comply with the directives related to protection of data and privacy of citizens as in case of the 112 service.
The DG eCall recommends to treat the eCall as a public service built on top of the pan-European single emergency number 112. Moreover, the citizens should be informed about eCall capabilities when buying a vehicle, especially about the data transferred to the PSAP.
From the point of view of the DG eCall, the data should be controlled by PSAP, or by private organization indicated by the public authorities to perform that role.

5.6. The eCall Certification

According to the DG eCall recommendation, the following entities should be responsible for the certification of components of the eCall system:

– the vehicle manufactures: for certification of the eCall generator (using existing certification procedures);

– the mobile network operators: for certification of their networks;

– the PSAPs: for certification of the PSAP system (along with the PSAP operators' procedures for handling eCall);

– all stakeholders involved in the eCall chain: for the certification of interoperability of the eCall service.

5.7. The eCall Deployment Plan

As agreed by the DG, the eCall road map assumes that all Member States will be ready to upgrade their PSAPs by September 2009 and to introduce of eCall as standard option in all vehicles granted type approval from 1 September 2010.

Moreover, it was assumed that all key stakeholders had signed the MoU by the end of 2006. Actually only half of the Member States have already signed the MoU.

According to the road map a full specification was ready and deployment of the eCall system began in mid-2007. Actually, the final ETSI/3GPP (The European Telecommunications Standards Institute/Third Generation Partnership Project) standards on emergency call handling and eCall data transfer as well as CEN standard regarding MSD content for eCall have not yet been published.

Full-scale field tests had been performed from the beginning of 2008, but in fact such tests were performed only in some Member States: in the Scandinavian countries, Germany and Austria.

6. The Status and the Basic Regulations of the eCall

The eCall project is treated as a part of the European eSafety programme built as a supplement to the single European emergency call number system (112). According to this, the regulations regarding eCall consist of two levels.

The first level comprises of the EU law for emergency services, especially for the enhanced 112 (E112). The fundamental regulations on emergency services are included in the Universal Service Directive [6]. It includes the obligation for the Member States to ensure an appropriate answer and handling of the calls made to 112, as well as the obligation for public telephone network operators to make caller location information available to authorities handling emergencies. The obligations concerning calling location information are included in Commission Recommendation [7]. According to this document, for every European emergency number 112 call, the public telephone network operators should, initiated by the network, forward (push) to public safety answering point the best information available on the location of the caller, to the extent technically feasible. All location information should be accompanied by an identification of the network from which the call originated.

The second level includes documents directly addressing the eCall. Currently, the eCall project is defined only in the optional/additional documents of the European Commission (e.g., communications, recommendations, programs, etc.) but it is not described in the obligatory EU low level acts (e.g., directives).

6.1. The EU Regulations Regarding the eCall Project

In communication from the Commission to the Council and the European Parliament [8], the eCall is presented as a project of emergency call initiated from the vehicle based on E112 with delivery of precise location information to the PSAP.

In the next EC communication [9], the eCall project was presented as a part of i2010 strategy and an element of information society included in the *Intelligent Vehicles* project. Moreover, the scope of the eCall functionality, possibility of integration with 112 and eCall deployment plan by 2009 as well as responsibility of the DG eCall, a list of the stakeholders involved in eCall and precise obligations for Member States was given.

In the third EC communication, a summary of introduction of a single European emergency call number as well as eCall in particular Member States was done [10]. The purpose of this document was mobilization of the Member States and branch environment to accelerate deployment of the eCall project. Moreover, it was focused on the influence of the Member States on adoption of a PSAP to handle the eCall and signing of the MoU by all EU Members.

In communication [11], EC is considering three possible policy options for progress of the eCall deployment:

- not intervening and leaving the introduction to market forces;
- supporting voluntary introduction by industry;
- mandating introduction through regulatory measures.

One of the regulatory measures, which EC will plan to take in 2010, is a recommendation to the Member States targeting MNO on the transmission of eCall, including MSD from the in-vehicle system to the PSAP. The second is a proposal for a regulation under the vehicle type-approval legislation [12] for the mandatory introduction of the in-vehicle part of the eCall service in new type-approved vehicles in Europe starting with certain categories (i.e., first in passenger cars and light commercial vehicles (categories M1 and N1) for which an appropriate triggering mechanism exist, and later in other vehicle categories). The last is the assessment of a potential regulatory measure for the necessary upgrading of the PSAP infrastructure required for proper receipt and handling of eCalls, in the framework of the proposed directive on the deployment of intelligent transport system (ITS) in Europe.

Paralelly to the above mentioned EC deliverables, other documents have been also elaborated and published. One of them was report on road safety [13] accepted by the European Parliament in April 2006 that indicates, e.g., on the costs regarding eCall, especially costs level from point of view of vehicle manufacturers. In June 2008 the European Parliament accepted the intelligent car report [14]. The main issue of this report is a problem concerning the introduction of eCall, especially in the context of the pilot tests performed in 2007–2008 and the signing of MoU by selected Members States.

In December 2008, the EC adopted the ITS action plan [15], in which support for eCall deployment is treated as one of EU actions. In the same time was adopted the ITS directive proposal, which provides for a legal instrument (i.e., a regulatory committee) to impose measures on the Member States, notably for the *"harmonized introduction of pan-European aCall"* [16]. Moreover, an announcement of the final version of standards necessary for deployment of eCall as well as the proposals of new regulations for introduction eCall project was given.

6.2. Voluntary Character of the eCall Project

Analyzing the documents elaborated by EC one can say that currently participation in the eCall project is voluntary. This note regards all stakeholders involved in eCall project – the Member States, the automotive industry, the mobile and fixed telecommunication industry, the public emergency authorities, the public social security, automobile clubs, etc. The Member States were only obliged to meet the requirements related to the emergency call number E112. They should apply harmonized conditions and principles to the provision of caller location information to emergency services for all calls to the single European emergency call number 112. Moreover, they should enable to make emergency calls using the single European emergency call number 112 and other national emergency number as well as to make all calls free of charge and without having to use any means of payment. The Member States should also provide that emergency calls should be routed to, and handled within, the appropriate emergency control centre. For each emergency call for which the subscriber number has be identified, the public telephone network operators should provide the capability to public safety answering points and renewing the location information through a call back functionality for the purpose of handling the emergency. The whole of eCall project has been defined as a pan-European public-private partnership, what means that this project has a voluntary character.

6.3. Legal Situation of the eCall Project in the Member States

The EC periodically controls how the obligations concerning handling emergency to the E112 are met by the Member States, but for the eCall project the EC monitors progress only in the countries which have signed MoU.

Currently, 15 Member States have signed the eCall MoU: Austria, Czech Republic, Cyprus, Estonia, Finland, Germany, Greece, Italy, Lithuania, Netherland, Portugal, Slovenia, Spain, Sweden and Slovakia as well as Norway, Suisse and Iceland. The six countries have announced willingness to sign shortly (e.g., Belgium, Bulgaria, Luxemburg, Poland, Romania and Hungary). With their signature, they commit themselves to actively support the timely implementation of the pan-European in vehicle emergency call system. The Member States that have not signed MoU inform EC about studies and analysis conducted by the appropriate public administration units or the interested institutions.

The EU countries that launched the eCall project have adopted their own implementation approaches. In France the eCall project is developing based on SMS communications instead of eCall data transfer – in band modem solution. Due to different priorities concerning development of emergency call system the UK is not interested in the eCall project in a form defined by the EC.

In each country being MoU signatory the structure of the eCall project management is different. Generally, the unit responsible for eCall project is an appropriate minister (e.g., minister of transport) as well as some central administration organ. Moreover, for purposes of the project the operational coordinator is indicated (e.g., PSAP administrator, trust of interested stakeholders, etc.). The governments of the Member States nominate their delegates to the eSafety Forum and to the eCall Driving Group.

In all the European countries the eCall project has mixed public-private partnership character. On the side of private institutions the project is introduced on the voluntary basis. In no country participation in the eCall project by players that are necessary to implementation of the eCall (e.g., vehicle manufacturers, telecom operators, etc.) is not obligatory. The same principle holds for the eCall users – the plans concerning obligatory participation in the eCall do not exist for them.

6.4. Legal Situation Regarding eCall Project in Poland

Polish law was adjusted in 2008 to enable implementation of regulations of directive on electronic communications concerning the emergency calls [17]. It was possible due to change of the telecommunications law as well as the law on state medical rescue. According to these new regulations the network operators are obliged to ensure that emergency call routed to 112 will be delivered to the appropriate PSAP. From 31 December 2010 calls may be delivered to other subjects that were chosen to perform PSAP role by regional administrative authority (governor of a province). Moreover, network operators were obliged to pass the information related to location of network terminations initiating calls to number 112 to the Polish telecom regulator (UKE). Mobile network operators should pass these informations in real time. Management of data base with location information and other user's data is the responsibility of president of UKE. Moreover, The regulation of the Ministry of Interior Affairs and Administration (MSWiA) of 17 September 2007 on the detailed organization of PSAP was published; it includes, e.g., the requirements for equipment and functionality of PSAPs.

Thanks to these changes the Polish law is currently adjusted to the EC recommendation mentioned above. It doesn't mean however that full practical solutions included in this recommendation were implemented in Poland. It should be noted that no separate regulation for the eCall exists in Poland. Generally, the Polish government supports initia-

tive to build electronic system to inform emergency services about road accidents.

7. Conclusion

The EC recommendations and requirements for the pan-European eCall already exist. Part of the standardization works concerning some technical problems for eCall reaches final versions, e.g., ETSI requirements for communications of citizens with authorities/organizations in case of distress (e.g., for emergency call handling) and 3GPP requirements for eCall data transfer based on an in-band modem solution.

Several problems are still waiting to be addressed and solved. One of them is a question whether a subscriber identity module (SIM) card is required for the eCall set up. About half of the Member States are in favor and a half are against. In this situation the analysis of cost of SIM management versus business case should be performed. Based on the results of that estimation an appropriate solution should be selected, for example – provide "eCall flag" and allow in-vehicle emergency calls in all European States, either with or without a SIM. It may be also allowed to use basic low-cost eCall devices without a SIM or provide an ad hoc low cost eCall SIM. Moreover, it is possible to select solution with one SIM number for all vehicles as no ID is needed.

The second problem is related to the "eCall flag". According to an opinion of the stakeholders involved in project, the "eCall flag" is needed and it should be included in the update of the telecommunications regulations (USD).

The next one is embedded or brought-in/nomadic IVS. The embedded solution provides reliability and robustness. The brought-in solution may not be reliable or robust enough in all electrical and mechanical environments but allows bundling with other services, solves life-time issue as well as SIM management. The stakeholders suggest support of a dual architecture with built-in safety and security functions and all other possibilities via telematics terminal, e.g., high end: built-in, lower end: nomadic.

Other identified problems regard PSAP workload, elaboration of standards for MSD transmission, privacy, etc. Although an emergency service center can act as a filter, it may cause delay, while allowing calls directly could increase PSAP workload.

The eCall as a standard in all cars can spearhead upgrade of the emergency services in Europe. An authoritative Eurobarometer study investigated users' attitudes towards intelligent vehicle safety systems in Europe finding that over 70% of the respondents in the EU want to have eCall in their car.

References

[1] "White Paper on European transport Policy for 2010: time to decide", COM (2001) 370 [Online]. Available: http://ec.europa.eu/transport/strategies/2001_white_paper_en.htm

[2] "Communication from the Commission – European Road Safety Action Programme – Halving the number of road accident victims in the European Union by 2010. A shared responsibility", COM (2003) 311 [Online]. Available: http://europa.eu/legislation_summaries/internal_market/single_market_for_goods/motor_vehicles/technical_implications_road_safety/l24257_en.htm

[3] "Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions – Intelligent Car Initiative – Raising Awareness of ICT for Smater, Safer and Cleaner Vehicles", COM (2006) 59 final [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2006:0059:FIN:EN:PDF

[4] "Recommendations of the DG eCall for the introduction of the pan-European eCall", Safety Forum, eCall Driving Group, Apr. 2006 [Online]. Available: http://www.ecall.fi/Position_papers_DG_eCall_-v2.pdf

[5] "Proposed MSD content for eCall", eSafety Forum, DG eCall, Jan. 2007 [Online]. Available: http://www.kokom.no/kokomsoek/20070312%20eCall_MSD_Content_v2%20(2).pdf

[6] "Directive 2002/22/EC of the European Parliament and of the Council of 7 March 2002 on universal service and user's rights relating to electronic communications networks and services (Universal Service Directive)" [Online]. Available: http://eur-lex.-europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:108:0051:-0077:EN:PDF

[7] "Commission Recommendation on processing of caller location information in electronic communication networks for the purpose of location enhanced emergency call service", C(2003) 2657 [Online]. Available: http://eur-lex.europa.eu/LexUri-Serv/LexUriServ.do?uri=OJ:L:2003:189:0049:0051:EN:PDF

[8] "Communication from the Commission to the Council and the European Parliament Information and Communications Technologies for Safe and Intelligent Vehicles" (SEC(2003) 963), Brussels, 15.09.2003, COM(2003) 542 final [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2003:0542:-FIN:EN:PDF

[9] "Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions – The 2nd eSafety Communication – Bringing eCall to Citizen", COM(2005) 431 [Online]. Available: http://eur-lex.europa.eu/Lex-UriServ/LexUriServ.do?uri=COM:2005:0431:FIN:EN:PDF

[10] "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – Bringing eCall back on track – Action Plan (3rd Safety Communication)", COM(2006) 723 [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52006DC0723:EN:NOT

[11] "Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions – eCall: Time to deployment", Brussels, 21.08.2009, COM(2009) 434 final [Online]. Available: http://eur-lex.europa.eu/Notice.do?mode=dbl&lang=en&ihmlang=en&lng1=en,pl&lng2=bg,cs,da,de,el,en,es,et,fi,fr,hu,it,lt,lv,mt,nl,pl,pt,ro,sk,sl,sv,&val=499932:cs&page

[12] "Directive 2007/46/EC establishing a framework for approval of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles", OJ L263/1, 9.10.2007 [Online]. Available: http://www.dft.gov.uk/pgr/roads/vehicles/sectionecwholevehicletype/eudirective200746ec.pdf

[13] "Report on road safety: Bringing eCall to citizens", (2005/2211(INI)), Committee on Transport and Tourism, Rapporteur: Gary Titley, FINAL A6-0072/2006 [Online]. Available: http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&-reference=A6-2006-0072&language=EN

[14] "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions "Towards Europe-wide Safer, Cleaner and Efficient Mobility", The First Intelligent Car Report", Brussels, 17.09.2007, COM(2007) 541 final [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri= COM:2007:0541:-FIN:EN:PDF

[15] "Action Plan for the Deployment of Intelligent Transport Systems in Europe", COM (2008) 886 [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri= COM:2008:0886:-FIN:EN:PDF

[16] "Proposal for a Directive laying down the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other transport modes", COM (2008) 887 [Online]. Available: http://eur-lex.-europa.eu/LexUriServ/ LexUriServ.do?uri=COM:2008:0887:FIN:-EN:PDF

[17] "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications (directive on privacy and electronic communications)" [Online]. Available: http://eur-lex.europa.-eu/pri/en/oj/dat/ 2002/l_201/l_20120020731en00370047.pdf

Wojciech Michalski was born in Bogate, in Poland, in 1952. He received the M.Sc degree in telecommunications engineering from the Technical University in Warsaw in 1977. He has been with the Switching Systems Department of National Institute of Telecommunications (NIT) since 1977, currently as a senior specialist. His research interests and work are related to PSTN backbone and access networks, GSM networks and IP networks. He is an author and co-author of technical requirements and many documents concerning telecommunication networks, services and protocols.

e-mail: W.Michalski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

ID-Based Digital Signatures with Security Enhanced Approach

Jacek Pomykała

Abstract—In the paper the ID-based digital signatures with signer's protection in case of the private key compromising is investigated. The proposed protocols have two main ingredients. First is the application of the credential system for the suitable verification key approval. Second is the application of the subliminal channel together with the interactive generation of the secret key, to obtain the increased resistance of the system against the powerful adversary. The particular interest was turned towards the significance of the deniable encryption in creation of the corresponding protocols.

Keywords—cryptography, deniable encryption, ID-based schemes.

1. Introduction

The critical point joining the functionality of the digital signatures is the management and authentication of the corresponding public keys. The potential way of cheating of (certificated) public key causes the risk that the identity of the user may be stolen. The concept of the ID-based public key cryptography introduced by Shamir implied the significant simplification of the corresponding management and authentication process.

In this concept the role of public key has been replaced by the user identity on the network (user-ID) like e-mail address, phone number, etc. More precisely there is the secret key of the private key generator (PKG), called the master key that is involved in the creation of the entity's secret key $sk = sk(\mathrm{ID}, s)$, by means of some trapdoor function. In order to keep the consistency, the public key Ω of PKG (known by each entity) should be related to the secret key $sk(\mathrm{ID}, s)$, so that the proof of knowledge of $sk(\mathrm{ID}, s)$ could be checked by any verifier with the aid of Ω.

When comparing with the certificate based cryptosystems the elimination of the public key certificates results here with the evident drawback. It implies that PKG knows the user secret key $sk(\mathrm{ID}, s)$. Moreover, loosing the master key s would compromise the secret keys of all entities. This obstacle make favorable the application of the ID-based schemes only in the systems with intermediate level of security.

In this paper we will encounter this problem suggesting some further ideas towards enhancing the security of some ID-based signature schemes. The first idea is to include in the entity's secret key its private part k generated by the user. The corresponding public part K sent in advance to PKG allows him to approve the corresponding verification key $vk = vk(\mathrm{ID}, K)$ related to Ω. As a result, in the subsequent step the signer is able to compute the new secret key $sk = sk(\mathrm{ID}, s, k)$ related to the suitable verification key. Such general scheme is described in Section 6.

Many ID-based signature schemes use a random parameter r in the signature generation process. This admits to hide the suitable private value k in the corresponding pseudorandom value $r = r(m, k)$. Certainly the corresponding commitment $R = R(r, m, \mathrm{ID}, \Omega, K)$ depends implicitly on k. Recovering this dependence allows the suitable party T (sharing the key k with the signer) to read the hidden (in the subliminal channel) information, with the aid of some trapdoor information t (known only by T). Since R is a one way function of r, its extraction from R is rather unrealistic task. We will show that the one-wayness may be replaced by the suitable trapdoor function depending on the parameters k and t that allows to recover the hidden information by the trustee from the corresponding signature Sig.

The above idea can be further enhanced in order to protect the signer against quite powerful adversary (that forces the signer to unveil his secret key $sk = sk(\mathrm{ID}, s, k)$). We assume that $r = r(m, k, \rho)$ with a random value ρ and the value of k that can be verified only on the basis of some trapdoor information t known only by T. In case of attack the signer show the "fake" values k' and ρ' instead of k and ρ leading to the same value $r = r(m, k', \rho')$.

Therefore the signer, when forced to unveil the signature parameters is able to get persuaded the adversary to believe that all of them have a real-looking data. Concluding, any verifier (except T) is not able to distinguish between the random r and a pseudorandom $r(m, k, \rho)$ even if he knows the signer secret key $sk(\mathrm{ID}) = sk(\mathrm{ID}, s, k)$. The resulting signature is to be called the deniable signature.

To construct the suitable pseudorandom function, the notion of deniable encryption $r = E_{den}(h(m, k))$ is applied similarly as in [1], with h being the secure hash function. The corresponding trapdoor information t applied in the decryption algorithm D, allows the trustee to check if $D_t(r) = h(m, k)$. On the other hand, even the signer (who does not known the trapdoor value t) is not able to find the evidence if a given pseudorandom value has actually the form $r = E_{den}(h(m, k))$. Hence the deniable signature protects the signer against the coercion attack when the adversary demands him to unveil the secret key $sk = sk(\mathrm{ID})$ and the corresponding pseudorandom parameters. The suitable protocol is constructed in Section 8. Summing up we are able not only to extend the basic ID-based signature scheme against the compromising the PKG master key s, but also the signer private key $sk = sk(\mathrm{ID}, s, k)$ when being coerced by the adversary. Throughout the paper we will illustrate the related ideas on digital signatures with secu-

rity based on hardness either of the factorization problem or the computational Diffie-Hellman problem in the gap Diffie-Hellman groups (GDH groups).

2. Related Work

The ID-based cryptosystems were introduced by Shamir [2]. The idea of gap Diffie-Hellman group based on the Weil pairing has its origin in the paper [3]. Boneh and Franklin [4] have proposed the first provably secure ID-based cryptosystem relating to GDH groups.

In [5]–[7] the ID-based digital signatures from the gap Diffie-Hellman groups were given. The proxy ID-based digital signature with derandomized Weil pairing computation was proposed in [8]. The general concept of transforming the standard signature schemes into the corresponding identity-based signatures (IBS) was the subject of paper [9].

In this paper we investigate the extensions of ID-based signature schemes having in mind the security requirements. The suitable improvements are based on the idea of subliminal channels investigated by Simmons [10] and applied in [11] for IBS scheme from the bilinear pairing. This approach was enhanced in [1] for the standard (certificate-based) signature schemes, referring to the concept of deniable encryption [12], [13].

The approach developed here deals with the ID-based digital signatures of the suitable form. We start from the standard IBS schemes and investigate the subsequent improvements by adding some "space" for the subliminal transfer and applying subsequently the concept of deniable signature.

The development is illustrated on the standard IBS schemes referring to [14] and [15]. At first we propose the extension of the standard IBS by including the signer private part to the secret key $sk(\mathrm{ID})$ and the corresponding approval by PKG (c.f. [16], [17]). Next we investigate the deniable encryption idea in standard signatures [1] to make the relevant transformation to IBS with the suitable security requirements.

3. Gap Diffie-Hellman Groups and Weil Pairing

Let $G = (G,+)$ be a group of prime order q and P, Q be any nontrivial elements of G. The discrete logarithm problem (DLP) in G may be stated as follows:

$$\text{Find } a \in Z_p \text{ such that } aP = Q.$$

Let us formulate the following related problems.

- The computational Diffie-Hellman problem (C-DH) – given the triple (P, aP, bP) find the element abP.
- The decisional Diffie-Hellman problem (D-DH) – given a quadruple (P, aP, bP, cP) decide whether $c = ab(\bmod q)$ (in which case we shall write that $(P, aP, bP, cP) = \mathrm{DH}$).

We call the group $G = (G,+)$ a gap Diffie-Hellman (GDH) group if (roughly speaking) the D-DH problem is computationally easy, while the C-DH problem is hard.

Let us recall briefly the construction of the gap Diffie-Hellman group based on the elliptic group structure applying the Weil pairing [4]. Let E be an elliptic curve over a finite field K of characteristic p and let n an integer not divisible by p. Denote by $\mathrm{cl}(K)$ the algebraic closure of K. It can be shown that the group $E[n]$ of n-torsion points of $E/\mathrm{cl}(K)$ is isomorphic to $Z_n \times Z_n$. The Weil pairing is a map

$$e : E[n] \times E[n] \to \mathrm{cl}(K)^*,$$

satisfying the following properties:

- alternation: for all $P, Q \in E[n]$, $e(PQ) = e(QP)^{-1}$;
- bilinearity: for any $P, Q, R \in E[n]$ we have $e(P + Q, R) = e(P, R)e(Q, R)$;
- non-degeneracy: if $P \in E[n]$ is such that for all $Q \in E[n]$, $e(PQ) = 1$, then $P = O$;
- computability: there exist an efficient algorithm to compute $e(P, Q)$ for any $P, Q \in E[n]$.

We now turn our attention to a more concrete situation. Let p be prime, $a \in Z_p^*$. Consider the elliptic curve E over F_p and the map $\Phi : E/\mathrm{cl}(F_p) \to E/\mathrm{cl}(Fp)$ defined by

$$E : Y^2 = X^3 + a, \Phi(O) = O;\ \Phi(x, y) = (\zeta x, y),$$

where

$$\zeta \in F_{p^2}^* \setminus \{1\}, \zeta^3 = 1,\ p = 2 \ (\mathrm{mod}\ 3)$$

or

$$E : Y^2 = X^3 + aX, \Phi(O) = O;\ \Phi(x, y) = (-x, iy),$$

where

$$i \in F_{p^2}^*,\ i^2 = -1,\ p = 3 \ (\mathrm{mod}\ 4).$$

One can easily check that Φ is an automorphism. Pick up a point $P \in E/F_p$ of prime order $q, q|p+1 = \mathrm{card}\, E/F_p$. Then $E[q] = \langle P, \Phi(P)\rangle$. We define the modified Weil pairing $\hat{e}$ by

$$\hat{e} : G \times G \to G',\ \hat{e}(R,\ S) = e(R, \Phi(S)),$$

where

$$G_1 = \langle P\rangle,\ G' = F_{p^2}^*.$$

It easy to show that for every $R \in \langle P\rangle$ such that $\hat{e}(R, P) = 1$, we have $R = 0$. It is known that the C-DH problem in G is hard (cf. [4]), but as it is shown in [18] not harder than the DLP in G'. The existence of Weil pairing implies directly that D-DH problem is easy in G. The randomized algorithm computing the Wail pairing was first proposed in [19]. The corresponding derandomized algorithm was shown in [8]. In what follows we will consider the bilinear structure (G, G', e, P) with G, G' and P as defined above and e being the suitable modified Weil pairing.

4. ID-Based Signature Schemes

The standard ID-based signature scheme considered in this paper is the tuple: IBS = (Setup, Extract, Sign, Verify). The corresponding algorithms are described below:

Setup of the system
The algorithm takes as input the security parameter and returns the description of the system, algebraic structure, the suitable hash functions and the pair (s,Ω), where s is the master key and Ω the corresponding PKG-public key.

Extract
Given the user identity (ID), PKG computes and sends (by a secure channel) the corresponding secret key for the signer $sk(\mathrm{ID}) = sk(\mathrm{ID},s)$.

Signing
Having as input the message m, the secret key $sk(\mathrm{ID})$ and a random element r, the signer computes the signature of m : $\mathrm{Sig} = [m,R,\sigma]$, where $R = R(m,r,\mathrm{ID})$ is the suitable commitment of r and $\sigma = \sigma(m,r,sk(\mathrm{ID}))$.

Verification
Any entity (verifier) having as input the signer identity ID, the signature Sig and PKG-public key Ω outputs "accept" or "reject" according to the verification process.

As an example we give the ID-based signature scheme from G-DH groups proposed in [7] (see also [15]), which will be the initial bilinear pairing based protocol for the further improvements.

Example 1: ID-based signature schemes from bilinear pairing
The protocol (IBSBP) consists of the following algorithms:

Setup of the system
Having as input the security parameter, it returns:
the bilinear structure (G,G',P,e), the pair $(s,\Omega = sP)$ and the suitable hash functions h and Q

$$h : \{0,1\}^* \to Z_q,$$

$$Q : \{0,1\}^* \to G.$$

Extract
Private key generator takes as input the identity ID of the user and returns the secret key $sk(\mathrm{ID}) = sQ(\mathrm{ID})$.

Signing
Given a message $m \in \{0,1\}^*$ the signer generates a random $r \in Z_q$ and computes the signature of $m : \mathrm{Sig} = [m,R,\sigma]$, where $R = rP$ and $\sigma = r\Omega + h(m,R)sk(\mathrm{ID})$.

Verification
To verify the signature $[m,R,\sigma]$ any user checks if $e(P,\sigma) = e(\Omega,R + h(m,R)Q(\mathrm{ID}))$.

5. Strong ID-Based Signature Scheme

In this section the enhance the IBS scheme to protect the signer against the compromising of the PKG master key s. This is due to the additional interactive protocol between the user U and PKG. The user generates a random (secret) k and the corresponding commitment K sends to PKG. This is the input for PKG to compute the suitable (public) verification key vk and the corresponding secret part for the signer. Consequently, the signer computes the final secret key $sk = sk(k,s,\mathrm{ID})$ not available for PKG. The protocol SIBS consists of the following algorithms: SIBS = (Setup, U-PKG, PKG-U, Keygen, Sign, Verify).

Setup
Having as input the security parameter, the algorithm returns the descryption of the system with the suitable hash functions and the pair (s,Ω).

***U*-PKG**
The signer U having as input the identity ID and Ω, generates the random secret key $k = k_{\mathrm{ID}}$ and publish the corresponding commitment $K = K(k,\mathrm{ID},\Omega)$.

PKG-*U*
PKG having as input the master key s and K computes and publish the verification key $vk = vk(s,\mathrm{ID},K)$ and the corresponding secret part $s_{\mathrm{PKG}-U}$ sends to the signer by the secure channel.

Keygen
The signer U having as input the private value $k = k(\mathrm{ID})$ and the secret part $s_{\mathrm{PKG}-U}$ computes the secret key $sk_U = sk_U(k,s_{\mathrm{PKG}-U})$ relating to the verification key $vk(s,\mathrm{ID},K)$.

Sign
The signer having as input the message m and the secret key sk_U computes the suitable signature $\mathrm{Sig} = \mathrm{Sig}(m,sk_U)$.

Verify
Any entity having as input the signature and the verification key $vk = vk(\mathrm{ID},K,\Omega)$ returns as output: accept or reject according to the verification process.

Example 2: Strong bilinear pairing based signature with credential delegation
Referring to Example 1 the above protocol is specified as follows:

Setup
Having as input the security parameter the algorithm returns the bilinear structure (G,G',e,P), the master key $s \in Z_q$ together with PKG-public key $\Omega = sP$ and suitable hash functions h,H and Q as below:

$$h : \{0,1\}^* \to Z_q,$$

$$H : G \to G,$$

$$Q : \{0,1\}^* \to G.$$

U-PKG
The algorithm is performed by the signer. Having as input P and random element $k \in Z_q$ the value $K = kP$ is computed and sent to PKG.

PKG-U
Having as input the master key s and the pair: (K,ID), PKG computes the credential approval $sH(vk) = sH(k\Omega) = sH(ksP) = sH(sK)$, the corresponding secret part sQ(ID) and send them to the signer.

Keygen
The algorithm is performed by the signer. Having as input the tuple $(k, sQ(\text{ID}))$ the signer computes the secret key $sk = sk(\text{ID}, k, s) = ksQ(\text{ID})$

Sign
The algorithm is performed by the signer. It has as input the message and signer's secret key $sk = ksQ(\text{ID})$ and returnes the signature: $[\text{Sig}, vk(\text{ID}),\ sH(vk(\text{ID}))]$, where $vk(\text{ID}) = k\Omega$, $\text{Sig} = [m, R, \sigma]$, with $R = rP$ and $\sigma = r[vk(\text{ID})] + h(m,R)sk(\text{ID})$.

Verify
The algorithm is performed by any verifier. First he checks the credential: $(vk(\text{ID}), sH(vk(\text{ID}))$ using the PKG-public key Ω. The credential is accepted provided $e(H(vk(\text{ID})), \Omega) = e(sH(vk(\text{ID})), P)$. If so he verifies the signature returning "accept" as output provided $e(P, \sigma) = e(vk(\text{ID}),\ R + h(m,R)Q(\text{ID}))$.

Example 3: Strong Fiat-Feige-Shamir ID-based signature scheme

The above protocol is the tuple: SFFSIBS = (Setup, PKGKeygen, U-PKG, PKG-U, Keygen, Sign, Verify) and is specified as follows:

Setup
Having as input the security parameter the algorithm returns the ring Z_n, the secure hash functions: $g, H : \{0,1\}^* \to Z_n$ with n to be specified latter on and the pair $(s, \Omega) = ((p,q), pq)$, where p, q are random prime numbers of suitable size.

U-PKG
Having as input the identity ID of the signer, the algorithm returns the private key $k = (p', q')$ and the corresponding commitment $K = p'q'$.

PKG-U
The algorithm is performed by PKG. Having as input the triple (ID, Ω, K) and the master key $s = (p,q)$, it returns the verification key $vk = vk(\text{ID}, K, \Omega) = (v_1, v_2, \ldots, v_l)$ mod $K\Omega$, with $v_j = H(\text{ID}\|j)$ and the secret value $s_{\text{PKG}-U} = (s'_1, \ldots, s'_l)$ satisfying the equalities $(s'_j)^2 = v_j$ mod Ω, $j = 1, 2, \ldots, l$. The secret value $s_{\text{PKG}-U}$ is sent to the signer by the secure channel. Here H is the given hash function with the corresponding value of n being equal to $K\Omega$.

Keygen
Having as input the value of $s_{\text{PKG}-U}$ and $k = (p', q')$ the signer computes the secret key $sk = sk(\text{ID}, k, s) = (s_1, \ldots, s_l)$, satisfying the inequalities $(s_j)^2 = v_j$ mod $K\Omega$, $j = 1, 2, \ldots, l$.

Sign
The algorithm has as input the message m, secret key $sk = (s_1, \ldots, s_l)$ and a random element $r \in Z_{K\Omega}$. As an output it returns $\text{Sig} = [m, \boldsymbol{R}, \sigma]$. Here $\sigma = r(s_1^{b_1}, \ldots, s_j^{b_l})$ with $\boldsymbol{R} = (b_1, \ldots, b_l)$ and $b_j (j = 1, 2, \ldots, l)$ are the subsequent bits of $g(m, U)$, with $U = r^2$ mod $K\Omega$, where g is the given hash function with the corresponding value $n = K\Omega$.

Verify
Having as input the message the signature $\text{Sig} = [m, \boldsymbol{R}, \sigma]$ and the verification key $vk = vk(\text{ID}, K, \Omega)$ the algorithm outputs "accept" provided $\sigma^2 (v_1^{b_1}, v_2^{b_2}, \ldots, v_l^{b_l})^{-1}$ mod $K\Omega$ is equal to U', such that $g(m, U')$ has the subsequent bits equal to b_j, $j = 1, 2, \ldots, l$.

6. T-Shared Key ID-Based Signature Scheme (T-SKIBS)

Some types of digital signatures require the presence of the selected third party in the verification process (see, eg., [20] and [21] – relating to the IBS schemes from bilinear pairing). We recall that we consider the signatures of the form $\text{Sig} = [m, R, \sigma]$, where R denotes the suitable commitment of the pseudorandom element $r = r(m, k, \text{ID})$ involved in the generation of σ. Now we will create the relevant subliminal channel to hide in R the information readable only for some third party T that shares the secret key k with the signer U. We assume that R is the commitment of r derived by the application of one-way (trapdoor) homomorphism Φ and H is a suitable hash function with the image included in the domain of Φ. The general scheme is the following: T-SKIBS = (Setup, Share$U$-T, Extract, Sign, Verify, Verify*), where the algorithms are described as follows:

Setup
Having as input the security parameter, the corresponding algebraic structure, the suitable hash functions and the pair (s, Ω) are given as output.

Share$U$-T
This is an interactive protocol between the signer U and the trustee T, at the end of which the secret shared key k is computed.

Extract
The algorithm is performed by PKG. It takes as input the pair (ID,s) and outputs the secret key of the signer $sk = sk(\text{ID}, s)$.

Sign

The input is the tuple $(m,k,sk(\mathrm{ID}),r)$, where $r = H(m,k,\Omega)$. The output is the signature that has the form $\mathrm{Sig} = [m,R,\sigma]$, where $\sigma = \sigma(m,k,sk(\mathrm{ID}),r)$.

Verify

The algorithm is performed by any entity. Having as input the signature Sig and the PKG-public key Ω. The algorithm outputs "accept" if σ is consistent with the pseudorandom parameter R and the message m.

Verify*

The algorithm is performed by the trustee T. Having as input the tuple (σ,k,Ω), the output is "accept" provided R is consistent with the value of $\Phi(H(m,k))$.

The algorithm Verify* provides us with the additional protection of the signer against the compromising of the secret key $sk = sk(\mathrm{ID})$, since it is no more equivalent to loosing the identity of the signer.

Example 4: *T*-shared key ID-based signature with bilinear pairing

Let (G,G',e,P) be a bilinear structure and $\Phi : Z_q \to G$ be the corresponding additive one-way group homomorphism. We let $R = \Phi(r)$ and $H : \{0,1\}^* \to Z_q$ be the suitable hash function. Using this specification for the protocol from Section 4, we see that the suitable changes will concern only the algorithms: Sign, Verify and Verify*. Namely the signer selects a random $r' \in Z_q$ and computes the commitment $R' = r'P$. Next he computes $r'' = H(m,k,R')$ and the generated signature has the form: $\mathrm{Sig} = [m,R',r'',\sigma^*]$, where $\sigma^* = \sigma(m,sk(\mathrm{ID}),r'+r'')$. The verification algorithm (Verify) is actually the same as in the basic scheme with the commitment R replaced by $R'+R''$. The additional (strong) verification (algorithm Verify*) checks if actually $\Phi(H(m,k,R')) = R''$.

The additional information r'' (superfluous for any verifier except T) is just the "hint" for the trustee to decide if the signature was generated by the authorized signer or not (algorithm Verify*).

7. Strong T-Shared ID-Based Signature Scheme (ST-SIBS)

Joining the concept of T-shared key (Section 6) with the extended communication (U-PKG, PKG-U) between the signer and PKG (Section 5) we arrive at the following protocol: ST-SIBS = (Setup, Share$U$-T, U-PKG, PKG-U, Keygen, Sign, Verify, Verify*).

Setup

Having as input the public data, the algorithm returns the corresponding algebraic structure suitable hash functions and the pair (s,Ω).

Share*U*-*T*

This is an interactive protocol between the signer U and the trustee T, at the end of which the secret shared key κ is computed.

***U*-PKG**

The signer U having as input the identity ID and Ω, generates the random secret key $k = k_{\mathrm{ID}}$ and publish the corresponding commitment $K = K(k,\mathrm{ID},\Omega)$.

PKG-*U*

PKG having as input the master key s and K computes and publish the verification key $vk = vk(s,\mathrm{ID},K)$ and the corresponding secret part $s_{\mathrm{PKG}-U}$ sends to the signer by the secure channel.

Keygen

Having as input the private value $k = k(\mathrm{ID})$ and the secret part $s_{\mathrm{PKG}-U}$ the signer U computes the secret key $sk_U = sk_U(\mathrm{ID},k,s_{\mathrm{PKG}-U})$ relating to the verification key $vk(s,\mathrm{ID},K)$.

Sign

Having as input the message m, the secret key sk_U and the corresponding pseudorandom value $r = r(m,k)$ the signer computes the suitable signature $\mathrm{Sig} = [m,R,\sigma]$, where $\sigma = \sigma(m,k,r,sk_U)$ and $R = R(m,k,r,\Omega)$ is the corresponding commitment of r.

Verify

Any entity having as input the verification key $vk = vk(\mathrm{ID},K,\Omega)$ and signature Sig returns as output accept or reject according to the verification process.

Verify*

Having as input the tuple (σ,κ,Ω), the trustee (T) outputs "accept" provided R is consistent with the value of $\Phi_\kappa(m)$ and "reject" otherwise.

The validity of the signature is checked by two kind of verification-weak verification which can be made by any entity and strong verification performed only by the trustee. The secret key of the signer $sk(\mathrm{ID}) = sk(\mathrm{ID},s_{\mathrm{PKG}-U},k)$ is computed with the aid of PKG master key s and the secret key κ shared with T. Therefore neither PKG nor T is able to forge the signature. Moreover, even in the case when the secret key $sk(\mathrm{ID})$ is compromising (but not the shared key κ), trustee can still distinguish between the signature generated by the real signer and the forger. This is due to the algorithm Verify* in the above protocol. Below we show the suitable example based on the Fiat-Feige-Shamir signature.

Example 5: Strong FFS *T*-shared key ID-based signature scheme

The protocol follows the above steps with the suitable specifications (see Example 3), i.e., SFFST-SKIBS = (Setup, Share*U-T, U*-PKG, PKG-*U*, Keygen, Sign, Verify, Verify*).

The only changes when compared with Example 3 concern the additional algorithms Share*U-T*, Verify* and the suitable modification in algorithm Sign (related to the dependence of the pseudorandom value r on the shared key κ). Namely:

Share*U-T*
The signer (U) and trustee (T) proceed the interactive protocol which outputs the shared secret key κ.

Sign
The algorithm takes as input the message m, secret key $sk = (s_1, \ldots, s_l)$ and a random element $r' \in Z_\Omega$. Applying the chineese remainder theorem we compute $r \bmod K\Omega : r = h(m,\kappa) \bmod K$ and $r = r' \bmod \Omega$. As an output it returnes $\mathrm{Sig} = [m, \boldsymbol{R}, \sigma]$. Here $\sigma = r(s_1^{b_1}, \ldots, s_l^{b_l}) \bmod K\Omega$, where $\boldsymbol{R} = (b_1, \ldots, b_l)$ and $b_j (j = 1, 2, \ldots, l)$ are the subsequent bits of $g(m,U)$, with $U = r^2 \bmod K\Omega$ and g being the suitable hash function $g : \{0,1\}^* \to Z_n (n = K\Omega)$.

Verify*
The trustee applying the verification key $vk = vk(\mathrm{ID}, K, \Omega) = (v_1, v_2, \ldots, v_l) \bmod K\Omega$ computes first the value $r^2 \bmod K\Omega$ then using the secret value $k = (p', q')$ corresponding to the commitment K computes the square root $r \bmod K$ and check if $r = h(m,\kappa)$, where κ is the secret key shared between the signer and trustee.

8. Deniable T-Shared Key ID-Based Signature (DT-SKIBS)

The signature scheme considered in Section 7 has the following (sometimes undesired) properties:

- the same message signed twice has the same signature;
- given the valid signature (satisfying both algorithms Verify and Verify*) the signer can check (prove another party) that the signature satisfies indeed the strong verification algorithm.

The second property might be used by the adversary in the so called coercion attack considered in the context of an encryption process in [12]. Suppose that the sender encrypts the message and sends it to the receiver. After some time the sender can be coerced by an adversary to give up the plaintext message together with the random choices involved in the encryption process. We can pose the question: can the sender protect himself against such an attack? The original idea of translucent sets applied in [12] was then exploited in [13] to give the more practical solution. Here we adopt the idea of deniable encryption in the context of the signature schemes that allows to avoid the above weakness. The idea is as follows. We let the pseudorandom element r be depending on some "random" factor ρ, so that for any fixed m, κ, κ' and ρ, there exists a corresponding ρ' satisfying $r = r(m,\kappa,\rho) = r(m,\kappa',\rho')$. In the case of the coercion attack the signer could recover the "fake" value κ' still keeping the real value of κ secret. This idea was investigated in [1] in the case of the certificate based signature schemes. Below we adopt it for the ID-based signature schemes. In the following protocol we incorporate the deniable encryption affecting merely the algorithms Share*U-T*, Sign and Verify* in the above ST-SIBS scheme. The corresponding protocol runs as follows:

Setup
Having as input the security parameter the algorithm returns the corresponding algebraic structure of the system together with the suitable hash function $h : \{0,1\}^* \to \{0,1\}^*$, and the pair (s, Ω).

Share*U-T*
In this part the suitable deniable encryption function $E_{\mathrm{den}} : Z_n \to Z_n$ and the corresponding decryption function $D_t : \mathrm{Im}(E_{\mathrm{den}}) \to Z_n$ are defined. Moreover, the signer and trustee (T) compute the shared secret key κ, while the corresponding trapdoor information t is known only for T.

***U*-PKG**
The signer U having as input the identity ID and Ω, generates the random secret key $k = k_{\mathrm{ID}}$ and publish the corresponding commitment $K = K(k, \mathrm{ID}, \Omega)$.

PKG-*U*
PKG having as input the master key s and the commitment K computes and publish the verification key $vk = vk(s, \mathrm{ID}, K)$ and the corresponding secret part $s_{\mathrm{PKG}-U}$ sends to the signer by the secure channel.

Keygen
Having as input the private value $k = k_{\mathrm{ID}}$ and the secret part $s_{\mathrm{PKG}-U}$ the signer U computes the secret key $sk_U = sk_U(\mathrm{ID}, k, s_{\mathrm{PKG}-U})$ relating to the verification key $vk(s, \mathrm{ID}, K)$.

Sign
Having as input the message m the secret key sk_U and the pseudorandom value $r = E_{\mathrm{den}}[h(m,\kappa,\rho)]$ the suitable signature $\mathrm{Sig} = [m, R, \sigma]$, with $\sigma = \sigma(m, k, r, sk_U)$ and the commitment $R = R(m, k, r, \Omega)$ of r is computed as output.

Verify
Any entity having as input the verification key $vk = vk(\mathrm{ID}, K, \Omega)$ and signature Sig returns as output "accept" if σ is consistent with (m, R) or "reject" otherwise.

Verify*
Having as input the tuple $(t, \sigma, \kappa, \Omega)$ the trustee (T) outputs "accept" provided the (t,k)-"trapdoor" inverse of R agrees with $h[(m,\kappa)]$ and "reject" otherwise.
Below we illustrate the above protocol using the standard ID-based signatures given in [7] and [14].

Example 6: Deniable *T*-shared ID-based signature from bilinear pairing
The protocol consists of the following algorithms: Setup, Share*U-T*, Extract, Sign, Verify, Verify*.

Setup
Given the security parameter the bilinear structure (G, G', e, P), the suitable hash functions $h : \{0,1\}^* \to Z_n$, $H : Z_n \to Z_q$, $Q : \{0,1\}^* \to G$ and the pair (s, Ω) are returned.

Share*U-T*
Here the suitable deniable encryption function $E_{\text{den}} : Z_n \to Z_q$ and the corresponding decryption function $D_t : \text{Im}(E_{\text{den}}) \to Z_n$ are defined. The signer (U) and trustee (T) compute the shared secret key κ, while the corresponding trapdoor information t (allowing to decrypt the deniably encrypted message) is known only for T.

Extract
Having as input the signer identity ID the algorithm returns the secret key for the signer $sk(\text{ID}) = sQ(\text{ID})$.

Sign
The signer selects a random $r' \in Z_q$ and computes the commitment $R' = r'P$. Next he applies the deniable encryption algorithm to compute $r'' = E_{\text{den}}[h(m, \kappa, R')]$. The signature is: $\text{Sig} = [m, R', r'', \sigma]$, where $\sigma = (r' + r'')\Omega + H(m, R' + r''P)sQ(\text{ID})$.

Verify
Any user checks if $e(P, \sigma) = e(\Omega, R' + r''P + H(m, R' + r''P)Q(\text{ID}))$.

Verify*
Using the trapdoor information t the trustee decrypts the value r'' and checks if it is equal to $h(m, \kappa, R')$.

Example 7: Strong FFS deniable *T*-shared key ID-based signature

Specifying the above general DT-SKIBS scheme with the SFFST-SKIBS protocol (see Example 5) we obtain the following protocol SFFSDT-SKIBS = (Setup, Share*U-T*, PKGKeygen, PKG-*U*, Keygen, Sign, Verify, Verify*). The corresponding algorithms are as follows:

Setup
Having as input the security parameter the ring Z_n together with the suitable hash functions: $H, g, h : \{0,1\}^* \to Z_n$ (with the values $n = n(H)$, $n = n(g)$, $n = n(h)$ to be specified, respectively) and the pair $(s, \Omega) = ((p,q), pq)$ with the suitable prime numbers p, q are given as output with the image of h contained in the interval $[M, M+K]$ for the suitable values of M and K.

Share*U-T*
Here the suitable deniable encryption function $E_{\text{den}} : Z_{n(h)} \to Z_N$ and the corresponding decryption function $D_t : \text{Im}(E_{\text{den}}) \to Z_n$ are defined. The signer (U) and trustee (T) compute the shared secret key κ, while the corresponding trapdoor information t (allowing to decrypt the deniably encrypted message) is known only for T.

***U*-PKG**
Having as input the identity ID of the signer, the algorithm returns the private key $k = (p', q')$ and the corresponding commitment $K = p'q'$ is sent to PKG.

PKG-*U*
The algorithm is performed by PKG. Having as input the triple (ID, Ω, K) and the master key $s = (p, q)$, it returns the verification key $vk = vk(\text{ID}, K, \Omega) = (v_1, v_2, \ldots, v_l)$ mod $K\Omega$, with $v_j = H(\text{ID}\|_j)$ and the secret value $s_{\text{PKG}-U} = (s'_1, \ldots, s'_j)$ satisfying the equalities $(s'_j)^2 = v_j \mod \Omega$, $j = 1, 2, \ldots, l$. The secret value is sent to the signer. Here H is the given hash function with the corresponding value of $n(H)$ being equal to $K\Omega$.

Keygen
Having as input the value of $s_{\text{PKG}-U}$ and $k = (p', q')$ the signer computes the secret key $sk = sk(\text{ID}, k, s) = (s_1, \ldots, s_l)$, satisfying the inequalities $(s_j)^2 = v_j \mod K\Omega$, $j = 1, 2, \ldots, l$.

Sign
The algorithm has as input the message m, secret key $sk = (s_1, \ldots, s_l)$ and a random element $r' \in Z_\Omega$. Applying the chineese remainder theorem we compute $r \mod K\Omega : r = E_{\text{den}}(h(m, \kappa)) \mod K$ and $r = r' \mod \Omega$ (with r belonging to the interval $[M, M+K]$). As an output it returns $\text{Sig} = [m, \boldsymbol{R}, \sigma]$. Here $\sigma = r(s_1^{b_1}, \ldots s_l^{b_l})$, where $\boldsymbol{R} = (b_1, \ldots, b_l)$ and $b_l (j = 1, 2, \ldots, l)$ are the subsequent bits of $g(m, U)$, with $U = r^2 \mod K\Omega$ and $n(g) = 2^{1+1}$.

Verify
Having as input the signature $\text{Sig} = [m, R, \sigma]$ and the verification key $vk = vk(\text{ID}, K, \Omega) = (v_1, v_2, \ldots, v_l) \mod K\Omega$, the algorithm outputs "accept" provided $\sigma^2 (v_1^{b_1}, v_2^{b_2}, \ldots, v_l^{b_l})^{-1}$ mod $K\Omega$ is equal to U', such that $g(m, U')$ has the subsequent bits equal to b_j, $j = 1, 2, \ldots, l$.

Verify*
Applying the verification key $vk = vk(\text{ID}, K, \Omega) = (v_1, v_2, \ldots, v_l) \mod K\Omega$ the trustee computes first the value r^2 mod $K\Omega$. Then using the secret value $k = (p', q')$ corresponding to the commitment K computes the square root r mod K and check if $D_t(a)$ is equal to $h(m, \kappa)$, where κ is the secret key shared between the signer and trustee, while a is the unique number congruent to r mod K belonging to the interval $[M, M+K]$.

9. Concluding Remarks

In the paper we have investigated the possible improvements of the ID-based signature schemes in a successive way, from the simpler protocols to the more advanced ones. To increase the clarity of presentation we have illustrated the ideas by the examples of two basic schemes due to Fiat-Feige-Shamir (see [14]) and bilinear pairing based protocol due to X.Yi [7]. The security of the underlying schemes relies on different computational problems namely the integer factorization problem and C-DH problem in the group of n-torsion points of elliptic curve over the finite field, respectively. Both the complexity and security of the basic schemes were studied in details in the literature. Here we have focused our attention towards the protection of the investigated schemes against the risk of compromising

the private key of the signer and the so called coercion attack (see, e.g., [1]). The main ingredient applied in this approach was the construction of the suitable subliminal channel in the underlying digital signatures.

This channel can be used to protect the signer in case of force, blackmail, etc. The critical information leaked subliminally can be read only by the pointed third party that knows some trapdoor information. The given evidence prove that some weakness of the ID-based digital signatures could be overcome by application of the interactive secret key generation and the idea of the deniable encryption.

Acknowledgement

The author is grateful to dr Jerzy Pejaś for reading of the first version of the article and his valuable remarks concerning the paper.

References

[1] K. Durnoga, J. Pomykała, and T. Trabszys, "Signature scheme with blackmail warning" (preprint).

[2] A. Shamir, "Identity-based cryptosystems and digital signatures", in *Proc. Crypto'87*, Santa Barbara, USA, 1987, pp. 47–53.

[3] A. Joux, "A one-round protocol for tripartite Diffie-Hellman", *J. Cryptol.*, vol. 17, no. 4, pp. 263–276, 2004.

[4] D. Boneh and M. Franklin, "Identity-based encryption from the Weil pairing", *SIAM J. Comp.*, vol. 32, no. 3, pp. 586–615, 2003.

[5] J. C. Cha and J. H. Cheon, *An Identity-Based Signature from Gap Diffie-Hellman Groups*, LNCS, vol. 2567. Berlin: Springer, 2003, pp.18–30.

[6] R. Sakai and M. Kasahara, "ID based cryptosystems with pairing on elliptic curve", in *Symp. Cryptogr. Inform. Secur. SCIS'2003*, Hamamatsu, Japan, 2003.

[7] X. Yi, "An identity based signature scheme from the Weil pairing", *IEEE Commun. Lett.*, vol. 7, no. 2, pp. 76–78, 2003.

[8] J. Pomykała and B. Źrałek, "A model of Id-based proxy signature scheme", in *Proc. 6th Coll. Iberoam. Collab. Electron. Commun. eCommerce Tech. Res. Conf.*, Madrid, Spain, 2008.

[9] M. Bellare, C. Namprempre, and G. Neven, *Security Proofs for Identity-Based Identification and Signature Schemes*, LNCS, vol. 3027. Berlin: Springer, 2004, pp. 268–286.

[10] G. J. Simmons, "The subliminal channel and digital signatures", in *Proc. EUROCRYPT'84 Worksh. Adv. Cryptol. Theory Appl.*, Paris, France, 1985, pp. 364–378.

[11] J. Pomykała and T. Trabszys, "Blackmail warning verifiably encrypted signatures from bilinear pairing", *Bull. WAT*, vol. LVII, no. 4, pp. 167–182, 2008.

[12] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, *Deniable Encryption*, LNCS, vol. 1294. Berlin: Springer, 1997, pp. 90–104.

[13] M. Klonowski, P. Kubiak, and M. Kutyłowski, *Practical Deniable Encryption*, LNCS, vol. 4910. Berlin: Springer, 2008, pp. 599–609.

[14] U. Feige, A. Fiat, and A. Shamir, "Zero-knowledge proofs of identity", *J. Cryptol.*, vol. 1, iss. 2, pp. 77–94, 1988.

[15] X. Cheng, J. Liu, and X. Wang, *Identity-Based Aggregate and Verifiably Encrypted Signatures from Bilinear Pairing*, LNCS, vol. 3483. Berlin: Springer, 2005, pp. 1046–1054.

[16] R. Tamassia and D. Yao, "Cascaded authorization with anonymous – signer aggregated signatures", in *Proc. IEEE Inform. Assur. Worksh.*, Royal Holloway, UK, 2006, pp. 84–91.

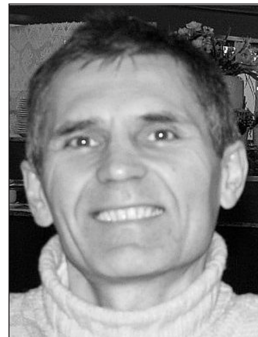
[17] A. Boldyreva, A. Palacio, and B. Warinschi, "Secure proxy signature schemes for delegation of signing rights", *Cryptol. ePrint Arch.*, 2003, Rep. 2003/096.

[18] A. Menezes, T. Okamoto, and S. Vanstone, "Reducing elliptic curve logarithms to logarithms in a finite field", *IEEE Trans. Inform. Theory*, vol. 39, pp. 1639–1646, 1993.

[19] V. S. Miller, "The Weil pairing, and its efficient calculation", *J. Cryptol.*, vol. 17, no. 4, pp. 233–334, 2004.

[20] J. Pejaś, "ID-based directed threshold signcryption scheme using a bilinear pairing", *Polish J. Envir. Stud.*, vol. 17, no. 4C, pp. 335–341, 2008.

[21] J. Pomykała and B. Źrałek, "Electronic signature, developement perspective", in *Electronic Signature and Biometric Authentication*, B. Hołyst and J. Pomykała, Eds. Warsaw: WSM, 2009 – in print (in Polish).



Jacek Pomykała (born in 1958) has obtained the M.Sc.degree in 1981 and Ph.D. degree in 1986 at the Faculty of Mathematics Informatics and Mechanics of University of Warsaw. His habilitation was done in 1997 at the Institute of Mathematics of Polish Academy of Sciences. He works in the Mathematical Institute at the Faculty of Mathematics Informatics and Mechanics of Warsaw University. His scientific interests are number theory, cryptology, computational complexity, security of computer systems. He was the invited speaker of many international conferences in the area of mathematics, computer science and security systems, the author of over 30 publications in the various international journals of these domains. He is an author of one book on the modeling and security of information systems, one of the editors of the book concerning the cryptographic and biometrics authentication and the special volume of the International Journal of Biometrics entitled "Digital signature and biometric in theory and practice". At present he is the leader of the international research seminar on Computational Number Theory and Cryptology at the Faculty of Mathematics Informatics and Mechanics of Warsaw University.

e-mail: pomykala@mimuw.edu.pl
Institute of Mathematics
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Banacha st 2
02-097 Warsaw, Poland