Recent Technologies for Industry and Computer Science

Lublin University of Technology
Faculty of Electrical Engineering
and Computer Science
ul. Nadbystrzycka 38A
20-618 Lublin

Recent Technologies for Industry and Computer Science

edited by
Jan Sikora
Waldemar Wójcik

Politechnika Lubelska
Lublin 2011

Reviewers:
dr hab. inż. Oleksandra Hotra, prof. Politechniki Lubelskiej
prof. dr hab. inż. Wiktor Łozbin

Publication approved by Rector of Lublin University of Technology

© Copyright by Lublin University of Technology 2011

ISBN: 978-83-62596-35-5

Publisher: Lublin University of Technology
ul. Nadbystrzycka 38D, 20-618 Lublin, Poland
Realization: Lublin University of Technology Library
ul. Nadbystrzycka 36A, 20-618 Lublin, Poland
tel. (81) 538-46-59, email: wydawca@pollub.pl
www.biblioteka.pollub.pl
Printed by : ESUS Tomasz Przybylak, Poznań, Poland
www.esus.pl

The digital version is available at the Digital Library of Lublin University of Technology: www.bc.pollub.pl
Impression: 100 copies

Contents

1. Multilevel and near-field optical data storage – prospect from theoretical point of view

A. Lapchuk, V. Petrov, A. Kryuchyn
Institute for Information Recording
2, Shpak str., Kyiev, Ukraine
Tel. 380-44 -454-8389, e-mail: petrov@ipri.kiev.ua

The principles of development of multilevel and near-field optical data storage are discussed. The synthesis of the optimal structure of multilevel optical data storage is reduced to a solution of a simple differential equation. The analysis of solutions of a differential equation enabled to obtain simple formulae for optical data storage capacity and to obtain simple criteria for applicability of a recording medium for multilevel data storage. The microstip near-field probe and microstrip near-field probe with a metal tip are proposed for near field optical data storage. It is found that microstrip near field probe has strong near-field interaction with recording layer in the case of medium with large dissipative losses. It is shown that recording layers with great dissipative losses and using illumination-collection mode of scanning near-field optical microscope (the most convenient mode for optical data storage) provide the highest optical efficiency and data storage capacity for near-field optical data storage.

1.1 Introduction

Optical data storage has led to revolution in information technology and storage. Consequently compact disk (CD)[1-3], digital video-disk (DVD)[4-5] and Blu –Ray disk (BD) [6-7] have appeared as compact, portable devices that have high storage density and high resistance to intense electromagnetic radiation. Given its high tolerance to vibration and high reliability the optical disk with bit-by-bit principle of information recording and retrieval

has an advantage over holographic memory. However in recent year the flash memory [8] and magnetic storage devices [9] have a significant increase of data density in comparison with optical data storage. Therefore the main challenge for optical data storage is to meet the rapid growth in demand for storage capacity.

There are two main ways for large increase in optical data storage density: volume storage and near field optical data storage. The near field optical storage uses small hole in opaque screen or sharp metal edge near recording surface to create a tiny bright spot with a size significantly smaller than a half of wavelength. The size of a light spot determines the resolution of optical system and therefore the near field method gives a large increase in optical data storage capacity. However near field method has very poor optical efficiency. Due to small optical efficiency data exchange rate is very slow and method needs a large improvement to be useful for technological application. Current optical disks have only one or two recording layers and therefore it uses inefficiently the volume of the disk. The large increase in storage capacity can be obtained by using all volume for data storage – 3D optical data storage. There are two types of 3D optical data storages: holographic data storage [10-11] and multilevel data storage [12-16].

Holographic data storage is based on page principle when information is read and recorded simultaneously from a two dimensional array of pits (page) [11]. Information on pages is not localized; it is distributed in entire 3D recording medium. The holographic data storage requires special medium and special optical system and electronics and is not under consideration in this paper. The multilevel optical data storage has many semitransparent recording layers and information from an inner layer is recorded and read through upper recording layers. There are two types of multilevel disks. The first type uses reflection from recording layer to get recording information. The recording process for this disk is the same as in conventional optical disk. During information recording the reflectivity or relief of recording layers is changed. The second type uses a fluorescent medium of recording layers for information recording [17-18].

Since multilevel optical storage uses the same principles of data recording and reading as in a conventional optical disk it is compatible with modern optical storage devices. This method also provides good data protection by a protective layer and it allows easy disk ejection from computer since reading devices has no mechanical contact with a disk (it preserves far-field method for information reading). In spite of great importance of multilevel optical data storage till now there is no mathematical model which can accurately evaluates required recording layer properties and a data storage capacity.

1.2. Multilevel optical data storage

Below we will present a simple method for calculation of focused laser beam propagation in thin multi-layered medium. By calculating of beam intensity inside a multilevel disk we will obtain the dependence of data capacity and signal level on recording layers properties. To derive a simple formula for a data capacity of a multilevel disk we will assume that every recording layer is monolayer (has no sublayers structure).

In the multilevel optical data storage a laser beam is focused inside a recording medium on a layer with desired information, as show in Fig. 1. Recording layers are inside transparent medium which has no dissipative losses. During the propagation through recording layers the intensity of a beam decreases due to dissipative losses in recording layers and reflection from recording layers. It is clear that every recording layer should have the same signal amplitude. To receive the same signal from all recording layers the reflection coefficient from recording layers should satisfy a condition:

$$p = T_n^2 r_n \tag{1}$$

where p is relative power of light reflected from the n-th recording layer after it returned back from an optical disk to a photodetector, T_n is the intensity transmission coefficient of a beam from the disk surface to the n-th recording layer, r_n –is the intensity reflection coefficient of this layer.

The value of p determines the maximum possible signal amplitude, therefore we denote it as signal level. Here we assumed that recording layers are thin and therefore the reflection coefficient is low $r_n << 1$. Due to small reflection coefficient it is possible to neglect the re-reflection part of light with small loss in accuracy of calculation. However the bottom recording layers should have a relatively large reflection coefficient since a signal from these layers is strongly attenuated by upper recording layers. Therefore the last several layers need a special consideration. Since the last recording layers have a large reflection coefficient a medium of recording layer should have a relatively large refractive index. A large refractive index of recording layers results in thinner recording layer and in decrease of difference in transmission coefficients for rays propagating at different angles.

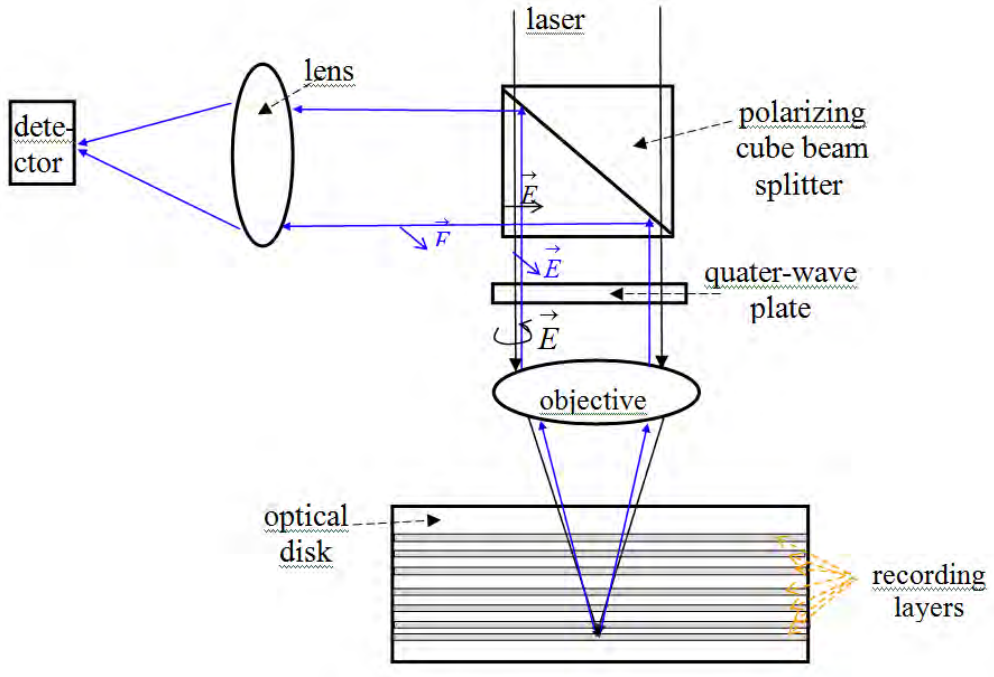


Fig. 1. Principal scheme of a multilevel data storage.

The reflection and transmission coefficients have strong dependence on recording layers parameters [19] and we want to estimate multilevel optical storage capacity depending on medium properties of recording layer. Our analysis will be based on the analysis of plane wave propagated normally to the disk surface. The modern optical systems have an objective with large numerical aperture and therefore the large part of laser beam will propagate at large angle to the disk surface. Therefore this part of the beam will have different reflection and transmission coefficients relatively to an axial ray. Since recording layers are placed inside a plastic medium with a refractive index approximately equal to 1.5, the convergence angle of a beam will be significantly decreased in disk medium. However it will be not small enough do not have any influence on beam parameters. Since we will not take into account difference

in reflection and transmission for rays propagating at different angles our estimation would give only approximate value of multilevel optical data storage capacity.

1.3. Recording layers with large dissipative losses and fluorescent disk

Medium of recording layers used in the modern optical data storage have large dissipative losses. The large imaginary part of refractive index of recording layers medium provides a large reflection from a recording layer and hence relatively large signal level. In read only memory (ROM) optical disk, used for mass production of video and audio information, an aluminum layer is used as recording layers to obtain high reflection from a recording layer. In rewritable (RW) optical disk GeSbTe layer is used as a medium with very fast transformation from glass to crystal and vice versa to write and erase information. Is it possible to use these mediums for multilevel optical data storage?

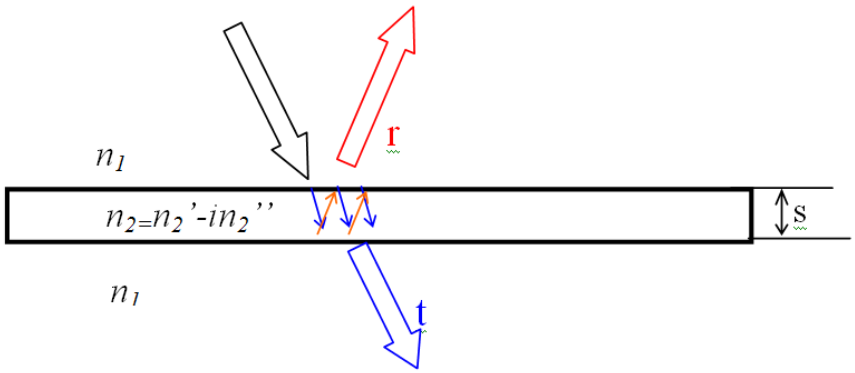


Fig. 2. Optical beam reflection from a thin layer with high dissipative losses.

To get an answer we should calculate the reflection and transmission coefficients of thin layers with high dissipative losses (see Fig. 2). In our calculation it is assumed that $|n_2| >> n_1$, which is true for recoding layers used in optical data storage.

By using Fresnel coefficient and summation of re-reflected light it is easy to obtain a simple formula for the reflection and transmission coefficients of thin recording layer:

$$|r|^2 = |r_0|^2 \left(1 + \frac{2n_1 \bar{n}_2 * (n_1 - n_2)^2 + 2n_1 n_2 * \left(n_1 - \bar{n}_2 \right)^2 + \left| \left(n_1 - \bar{n}_2 \right) \right|^4}{4n_1^2 |n_2|^2} \right) 4k^2 s^2 \approx$$

$$\approx |r_0|^2 \frac{|n_2|^2}{n_1^2} k^2 s^2 \left(1 + O\left(\frac{n_1}{|n_2|} \right)\right) \qquad (2)$$

$$|t|^2 = 1 - 2k_0 s n_2'' - i2k_0 s \left(i2n_1 n_2'' - i2n_2' n_2'' \right) / 2n_1 \approx 1 - 2k_0 s n_2' n_2'' / n_1 + O\left((ks)^2\right) \approx 1 - \beta s \,, \qquad (3)$$

where: r_0 is reflection coefficient from interface of optical disk /recording layer mediums, k_0 is wavenumber, $\beta = 2k_0 n_2' n_2'' / n_1$ n_2' and n_2'' are real and imaginary part of refractive index of thin layer, consequently.

From Eq. (1) and Eq. (2) follows that reflection increases square and transmission decreases linearly with layer thickness. Hence for thin layers almost all energy losses of a beam propagating inside a disk will be due to dissipation in recording layers and with good accuracy we can neglect the reflection losses. In this approximation the beam intensity decrease can be written as follows:

$$\left|T_N\right|^2{}_N = t_1 * t_2 * ... * t_n = \left(1-\beta s_1\right)*\left(1-\beta s_2\right)*...(1-\beta s_n) \approx \exp(-\beta*\left(s_1+s_2+..s_n\right)) = \exp(-\beta * S_N), \quad (4)$$

where $\left|T_N\right|^2$ - is the transmission coefficient of a beam through N recording layers and s_i is a thickness of i-th recording layer.

In the case of many recording layers and of slow changing thickness of recording layers with high accuracy we could make the next approximation:

$$s_i = S(i) - S(i-1) \approx \frac{dS(i)}{di}\left(i-(i-1)\right) = \frac{dS(i)}{di}, \quad (5)$$

where $S(i)$ is total thickness of i-th recording layers.

With the help of Eq. (2), Eq. (4) and Eq. (5) we could rewrite Eq. (1) as:

$$\left|r_0\right|^2 \frac{\left|n_2\right|^2}{\left|n_1\right|^2} k_0{}^2 \left(\frac{dS}{dn}\right)^2 \exp\left(-\beta S\right)^2 = p. \quad (6)$$

Nonlinear differential Eq. (6) can be easily transformed to linear integral equation:

$$\frac{dS}{di}\exp(-\beta S(i)) = \sqrt{p} \Big/ \left(\left|r_0\right|\frac{\left|n_2\right|}{n_1}k_0\right). \quad (7)$$

The solution of Eq. (7) can be written as follows:

$$\exp(-\beta S) = -\beta * \sqrt{p} * \left(n - n_0\right) \Big/ \left(\left|r_0\right|\frac{\left|n_2\right|}{n_1}k_0\right), \quad (8)$$

where n_0 is a constant of integration. From Eq. (8) follows that:

$$S(n) = -\ln\left(\beta * \sqrt{\gamma} * \left(n_0 - n\right) \Big/ \left(\left|r_0\right|\frac{\left|n_2\right|}{n_1}k_0\right)\right) \Big/ \beta. \quad (9)$$

From Eq. (4) and Eq. (9) the thickness of the i-th recording layer can be found:

$$s_n = \frac{dS(n)}{dn} = \frac{1}{\beta\left(n_0 - n\right)}. \quad (10)$$

The total thickness of all recording layers can be calculated by a simple formula:

$$S_{total} = -\ln\left(\beta * \sqrt{p} * \left(n_0 - N_0\right) \Big/ \left(\left|r_0\right|\frac{\left|n_2\right|}{n_1}k_0\right)\right) \Big/ \beta, \quad (11)$$

where N_0 is a total number of recording layers on one optical disk. Since the layer thickness should be positive we could write inequality:

$$\beta^* \sqrt{p} * (n_0 - 1) / \left(|r_0| \frac{|n_2|}{n_1} k_0 \right) < 1, \tag{12}$$

and hence

$$n_0 < |r_0| \frac{|n_2|}{n_1} k_0 / \left(\beta^* \sqrt{p} \right) + 1. \tag{13}$$

Since $n_0 > N_0$ we can write:

$$N_0 < |r_0| \frac{|n_2|}{|n_1|} k_0 / \left(\beta^* \sqrt{p} \right). \tag{14}$$

On the assumption of high reflection coefficient $|r_0| \approx 1$ (large $|n_2| / n_1 >> 1$ ratio) Eq. (14) could be written as follows:

$$N_0 < \frac{|n_2| k_0}{n_1 \beta^* \sqrt{p}} = 0,5 \frac{|n_2| k_0}{n_1 k \frac{n_2' n_2''}{n_1} \sqrt{p}} \approx 0,5 \frac{|n_2|}{n_2' n_2'' \sqrt{p}}. \tag{15}$$

From Eq. (15) follows that maximum number of recording layers which could be written on one disk for a given signal level is inverse to dissipative losses of recording layer and to square root of a signal level. Hence when someone decreases signal level 100 times he can obtain only tenfold increase in number of recording layers. However decrease in losses 10 times gives increase in number of recording layers in 10 times. From mentioned above follows that the best way to get large information capacity in one disk is to use recording layers with small dissipative losses. The value of $0,5|n_2| / (n_2' n_2'')$ of recording medium we have a notation as a quality factor for multilayer recording medium.

From data in Table 1 one can see that GeSbTe medium used for RW disk has low quality factor. Therefore for multilevel disk with GeSbTe recording layers it is possible to obtain only 3 recording layers even for the case of rather low signal level p=0.01. From data presented in Table 1 follows that silver is the best medium for multilelel data storage. Direct calculation of multilevel disk properties has shown that Eq. (15) gives the only qualitative estimation for the number of recording layers. The recording layer thickness fast increase with increase layer number (depth of layer) and the reflection from recording layer rapidly increase and soon reflection becomes too large to be neglected. Therefore Eq. (15) overestimates the number of recording layers and can be applied only for qualitative analysis of multilevel optical disk.

Table 1. Quality factor for conventional recording layer medium (λ=400nm).

Medium	Al	Ag	Au	GeSbTe
n'	0.49	0.179	1.24	3.5
n''	4.86	1.95	1.79	1.7
quality factor	2	5.6	1	0.65
N_0(p=0.01)	10	28	5	3

The method worked out above can be applied to the fluorescence multilevel optical data storage. In this case the radiation from a recording layer is proportional to layer thickness and the field intensity decrease can be approximated accurately by the exponential law. Therefore the condition for constant signal can be written as follows:

$$\alpha\left(\frac{dS}{dn}\right)\exp\left(-\beta S\right)^2 = \alpha\left(\frac{dS}{dn}\right)\exp\left(-2\beta S\right) = p\,, \tag{16}$$

where α is the optical efficiency, β is attenuation coefficient (due to absorption by fluorescent medium), n – recording layer number, S - is total length of n upper recording layers.

The optical efficiency is the product of fluorescence efficiency and optical system efficiency (efficiency to catch fluorescent light, radiated homogeneously in all directions (α<0,1). Applying the same method as above we obtain the next formula for a number of recording layers:

$$N_0 = \alpha / 2p\,. \tag{17}$$

The Eq. (17) provides accurate estimation for fluorescent disk layer number since recording layers in this disk have very low reflectance. From Eq. (17) that the number of total recording layers for fluorescent disk depends only on optical efficiency and signal level.

For example for signal level p=0.01 and optical efficiency α=0,10 the disk can have only five recording layers. However since in a fluorescent disk it is possible to separate fluorescent light from background radiation by using filter, this method cans use very low signal levels to detect a useful signal and therefore it can have a large number of recording layers.

1.4. Recording layer with low dissipative losses

From the previous part follows that the irecording layers of a disk should have low dissipative losses to obtain the large number of recording layers in one optical disk. However for R and RW optical disk information recording is carried out by heating of recoding layers and therefore the recording layer should have the dissipative losses sufficiently large to reach a temperature sufficient for information recording. Since the dissipative losses should be small the recording layers should have low working temperature.

In optical band almost all media with a high refractive index have high losses. Therefore it is very difficult to create a multi-level optical disk with a high signal level. Table 2 gives complex refractive indices for mediums which could be used as recording layers for a multilevel optical disk (have large refractive index and low dissipative losses).

Table 2. Complex refractive indice of medium for multilevel data storage (λ =400 nm).

Mediuma	Complex refractive index *n*	
	Re *n*	Im *n*
GaP	4.19600	0.275000
Si	5.57000	0.387000
TiO_2	3.40000	0.00000
AlSb	4.57000	2.12000
ZnTe	3.40000	0.950000
Si_8Ge2	5.79000	1.34000
PbSe	4.98000	0.173000
InP	4.41500	1.73500
InAs	3.10800	1.95700
ZnS	2.56000	0.00000
GaAs	4.37300	2.14600
AlSb	4.57000	2.12000

For ROM disk the best case for data storage capacity is to have a recording layer without any dissipative losses. However in this case a real part of complex refractive index of recording medium should be sufficiently high to enable an increase of reflection from bottom recording layers to obtain a required signal level. From data on Table 2 it is clear that TiO_2 has the highest refractive index for blue light for optical media and has not dissipative losses. Therefore, TiO_2 is very attractive for a multi-level ROM disk as medium for recording layers.

Since TiO_2 is mechanically and chemically stable, a multi-level TiO_2 based disk can be used for long-term information storage. TiO_2 has two possible crystal structures of anatase and rutile with different refractive indices. The structure of anatase after heating to 400^0C irreversibly transformed into rutile and therefore this medium could be used for R-disk. However since TiO_2 has not dissipative losses it needs an additional sublayer that has a low dissipative losses to heat TiO_2 layer to temperature of 400^0C. There are two ways of using an additional medium for recording layers (see Fig. 3): a) a layer of TiO_2 matrix with embedded inside nanoparticles (b); recording layer with two sublayers. A homogeneous layer is more easy for manufacturing and homogeneous structure will have significantly smaller scattering therefore it looks more promising.

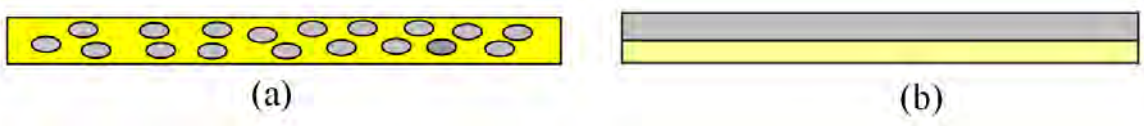


Fig. 3. Structures of recording level with two mediums.

It is impossible to apply the mathematical model worked out in previous paragraph for the case of ROM disk with recording layers without dissipative losses (the best case for ROM disk). A new approach is needed to determine relation between recording layer parameters and information capacity of a muli-level disk for this case. Here it will be assumed, exactly as above, that the layer parameters (reflection and transmission coefficients) are changing smoothly from a layer to a layer and therefore it is possible to change summation to integration for calculating beam energy during light propagation inside a disk. It will be also assumed that all recording layers should have the same signal level and therefore the reflection and transmission coefficient should satisfy Eq. (1). Differences of transmission coefficients through n and *n*-1 recording layers can be written as follows:

$$T(n)-T(n-1)=\frac{P(n)}{P_0}-\frac{P(n-1)}{P_0}=\frac{P(n-1)}{P_0}\left(\frac{P(n)}{P(n-1)}-1\right)=T_0(n-1)\left(t(n)-1\right), \quad (18)$$

where $t(n)$ and $T(n)$ are transmission coefficients (intensity) of the *n*-th and *n* recording layers, consequently, *P(n)* - beam power after transmission through *n* recording layer, P_0 - initial power of the beam. Since a layer has no lossless we can write:

$$r(n)+t(n)=1, \quad (19)$$

or:

$$r(n)-1=-r(n), \quad (20)$$

where $r(n)$ is the reflection coefficient (intensity) of the *n*-th recording layer. Now we can rewrite Eq. (18) as follows:

$$T(n)-T(n-1)\approx\frac{dT(n)}{dn}=-T(n-1)R(n)=-T(n)\frac{r(n)}{(1-t(n))}. \quad (21)$$

After taking the logarithm of the left and right parts of Eq. (1) it transforms to:

$$2\ln(T(n))+\ln(r(n))=\ln p, \quad (22)$$

and after taking the derivative:

$$2\frac{dT(n)}{dn}/T(n)+\frac{dr(n)}{dn}/r(n)=0. \quad (23)$$

After substitution Eq. (21) into Eq. (23) it transforms to:

$$-2\frac{r(n)}{(1-r(n))}+\frac{dr(n)}{dn}/r(n)=0. \quad (24)$$

It is easy to find solution of Eq. (24):

$$1/r(n)+\ln r(n)=2(n_0-n), \quad (25)$$

where n_0 is the constant of integration. The transcendental Eq. (24) can only be solved by a numerical method. However for thin recording layers reflection coefficient is small ($r<<1$) and hence its value satisfies the next inequality $|\ln r|<<1/r$ and therefore in a first approximation we can neglect a term with a logarithm in Eq. (25):

$$-1/R\approx 2(n-n_0), \quad (26)$$

or :

$$r(n)=-0,5/(n-n_0). \quad (27)$$

A constant of integration can be found from condition that every recording layer should provide a proper signal level. The condition is most simply to write for the first recording layer:

$$r(1) = \gamma = -0,5 / (1 - n_0), \tag{28}$$

or:

$$n_0 = 1 + 0,5 / p. \tag{29}$$

After substitution Eq. (29) into Eq. (27) it could be rewritten as follows:

$$r(n) = 0,5 / (n_0 - n) = 0,5(1 + 0,5 / p - n) = 0,5p / ((1 - n)p + 0,5). \tag{30}$$

Now we can obtain a maximum possible number of recording layers from a condition that bottom layer should reflects all incident light:

$$R(N) = 0,5 / (n_0 - N_0) = 1, \tag{31}$$

and finally:

$$N_0 = n_0 - 0,5 = 1 + 0,5 / p - 0.5 = 0,5(1 + 1 / p). \tag{32}$$

We do not use the parameters of medium of recording layer to derive the equation of Eq. (32). The only condition imposed on the lossless recording layer medium is sufficiently high refractive index to satisfy the Eq. (3). From Eq. (32) follows that a maximum number of recording layers is smaller by a factor of two compared with the ideal case when energy of the propagating beam distributed homogeneously between all layers without reflection losses from other layers:

$$N_0' = 1 / p \ ; \tag{33}$$

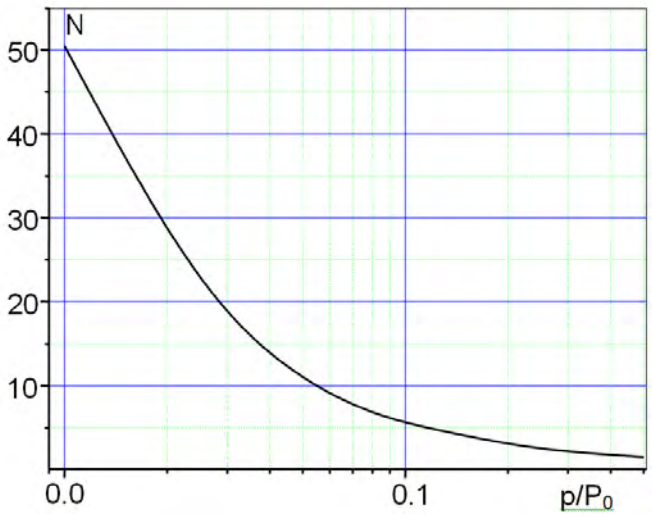


Fig. 4. Dependence of maximum number of recording layers on margin signal level.

Fig. 4 shows the dependence of the recording layers number (maximum possible) on a signal level calculated by using Eq. (31). Fig. 5 and Fig. 6 show dependence of the reflection coefficient on layer number for the case of 15 and 30 layers and with a signal level of 3.3% and 1.5%, respectively. From data presented in Fig. 5 and 6 one can see that the reflection coefficient has smooth dependence on a layer number and is small ($r<<1$) almost for all layers, as it was assumed in our mathematical model. Therefore the proposed mathematical model should provide sufficiently accurate calculation of multilevel optical disk parameters. However the assumption of small value and smooth dependence of a reflection coefficient is not true for the last recording layers and therefore a more rigorous approach is needed for simulation of the parameters of these recording layers.

From the curves in Fig. 5 and Fig. 6 it is clear that due to a fast change in reflectivity of the last recording layers, only the last layer has a reflection coefficient over 0.4. But such a low reflection coefficient (<0.4) can be easily obtained by a layer of recording medium that has a relatively low refractive index such as titanium oxide TiO_2.

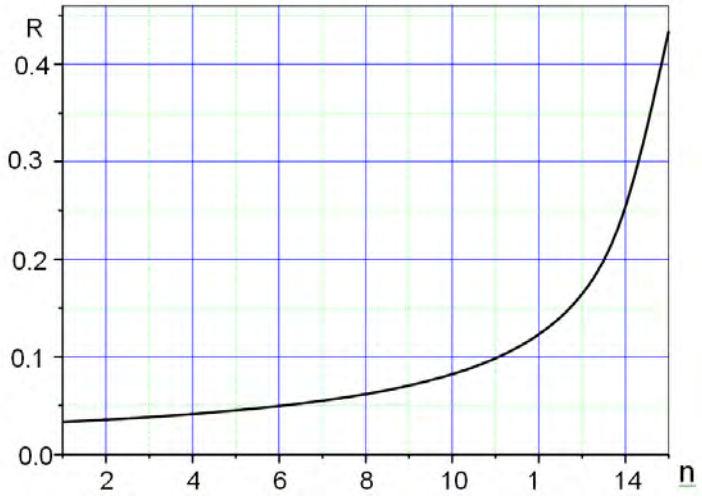


Fig. 5. Reflection coefficient of different recording layer (n is layer number) for the case of margin signal level of 3,3%.

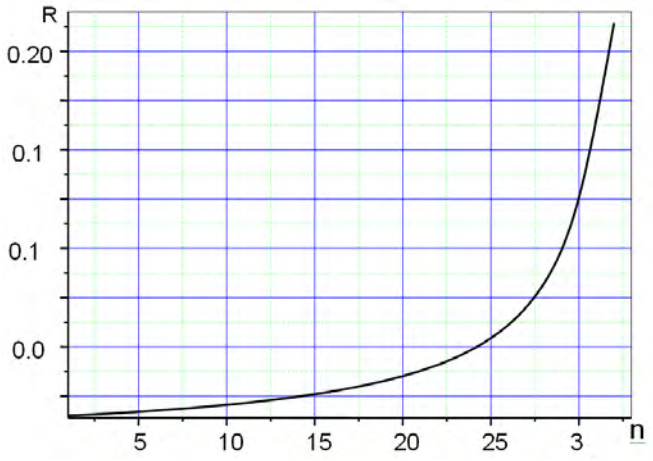


Fig. 6. Reflection coefficient of different recording layer (n is layer number) for the case of margin signal level of 1,5%.

1.5. Simulation of focused-beam propagation in a multilevel medium near the focal plane

Fig. 7 shows the optical scheme of a beam focused by objective lens. It is assumed that the impinging light beam at the entrance pupil of the objective lens is Gaussian beam with a plane waterfront. In our algorithm, the objective lens is expected to be an object which transfers a plane waterfront to a spherical waterfront with appropriate polarization rotation and with specified aberration. In such a situation, the focused-beam field can be calculated by the Straton-Chu integral [4]:

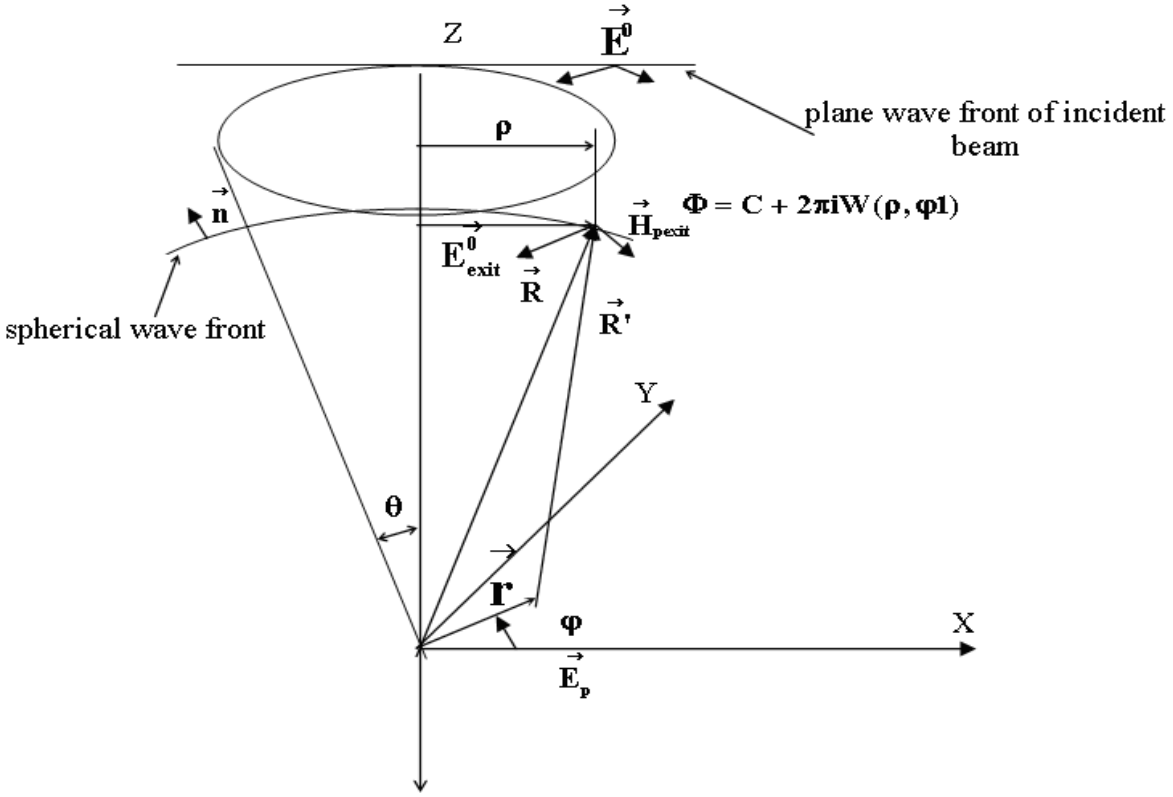


Fig. 7. Schematic diagram of a beam focused by an objective lens.

$$\vec{E}_r = \frac{1}{4\pi}\oint_{Sexit}\left[\vec{n}\times\vec{E}_{pexit}\right]\times\left(-\frac{ik\left(\vec{r}-\vec{R}\right)}{\left|\vec{r}-\vec{R}\right|}f\left(\vec{R}-\vec{r}\right)dS\right)+\left(\vec{n}\cdot\vec{E}_{pexit}\right)\frac{ik\left(\vec{r}-\vec{R}\right)}{\left|\vec{r}-\vec{R}\right|}f\left(\vec{R}-\vec{r}\right)+ \tag{34}$$
$$ik\rho_1\left[\vec{n}\times\vec{H}_{pexit}\right]f\left(\vec{R}-\vec{r}\right)dS,$$

where the integration is taken over areal of exit pupil plane, k is wavenumber, E_{pexit} and H_{pexit} are electric and magnetic fields at the exit-pupil plane, and:

$$f\left(\vec{R}-\vec{r}\right)=\frac{\exp\left(-i\,k\left|\vec{R}-\vec{r}\right|\right)}{\left|\vec{R}-\vec{r}\right|},$$

where $\vec{n}$ is unit vector orthogonal to surface of integration, square brackets means vector product.

Since $\vec{H}_{pexit}=-\frac{k}{\omega\mu_0}\left[\vec{n}\times\vec{E}_{pexit}\right]$ the Eq. (34) can be written as:

$$\vec{E}_r=\frac{1}{4\pi}\oint_{Sexit}\left[\vec{n}\times\vec{E}_{pexit}\right]x\left(-\frac{i\,k\left(\vec{r}-\vec{R}\right)}{\left|\vec{r}-\vec{R}\right|}f\left(\vec{R}-\vec{r}\right)dS\right)+\left(\vec{n}\cdot\vec{E}_{pexit}\right)\frac{i\,k\left(\vec{r}-\vec{R}\right)}{\left|\vec{r}-\vec{R}\right|}f\left(\vec{R}-\vec{r}\right)+ \tag{35}$$
$$i\,k\vec{E}_{pexit}\,f\left(\vec{R}-\vec{r}\right)dS.$$

The Eq. (35) can be simplified, by using the approximate equality $\left(\vec{r}-\vec{R}\right)/\left|\vec{r}-\vec{R}\right|=-\vec{R}/\left|\vec{R}\right|=-\vec{n}$ the error of which is smaller than $r/R\sim\lambda/R<0{,}001$ to:

$$\vec{E}_r=\frac{i\,k}{2\pi}\oint_S\vec{E}_{pexit}\,f\left(\vec{R}-\vec{r}\right)ds. \tag{36}$$

For integral (36) calculation we need to have the field at the exit pupil plane. This field can be calculated from the incident field at entrance pupil by the formula:

$$\vec{E}^{\,0}_{pexit}=\begin{pmatrix}\vec{s} & \vec{p}\end{pmatrix}\begin{bmatrix}P_{sx} & P_{xp}\\ P_{sy} & P_{py}\end{bmatrix}\begin{pmatrix}E_x^0\\ E_y^0\end{pmatrix}\exp\left(2\pi i W(\rho,\varphi)\right), \tag{37}$$

where E_x^0 and E_y^0 are the electric field components located in entrance pupil,

$$\begin{bmatrix}P_{sx} & P_{xp}\\ P_{sy} & P_{py}\end{bmatrix}=\begin{bmatrix}\dfrac{\beta}{\sqrt{1-\gamma^2}} & -\dfrac{\alpha}{\sqrt{1-\gamma^2}}\\ \dfrac{\alpha}{\sqrt{1-\gamma^2}} & \dfrac{\beta}{\sqrt{1-\gamma^2}}\end{bmatrix},$$

is the polarization transformation matrix of the lens, $W(\rho,\varphi)=\sum C_{nk}\left(\frac{\rho}{\rho_0}\right)^n\cos\left(k\left(\varphi-\varphi_{nk}\right)\right)$ is the wave-front error due to lens aberrations, expanded in the set of Zernike's circle polynomials, $\alpha=\rho\cos(\phi)/R$, $\beta=\rho\sin(\phi)/R$, $\gamma=\sqrt{1^2-\alpha^2-\beta^2}$ (Fig. 7).

The simplest method for field calculation in a multilevel medium near focal plane is to represent it as a set of plane waves (application of Fourier transform) since it is easy to calculate plane wave propagation in multilayer medium. It is possible to obtain the electromagnetic field in Fourier space without calculation the field distribution at the focal plane by the use of stationary-phase integration method:

$$\vec{E}_{pexit} = \vec{E}_{spot}\left(\frac{kx}{r},\frac{ky}{r}\right)\frac{2\pi kz\exp\left(i\,kr\right)}{r^2} = \vec{E}_{spot}\left(\alpha,\beta\right)\frac{2\pi kz\exp\left(i\,kr\right)}{r^2} \tag{38}$$

where $\vec{E}_{spot}\left(\alpha,\beta\right)$ is the Fourier image of the field at the focal plane, $\alpha = kx/r$, $\beta = ky/r$ are *x*- and *y*- component of vector wavenumber. Eq. (38) after simple transformation can be used for the calculation of the field in Fourier space at focal plane:

$$\vec{E}_{spot}\left(\alpha,\beta\right) = \frac{f\ \lambda}{\left(2\pi\right)^2}\vec{E}\left(\frac{kx}{r},\frac{ky}{r}\right)\exp\left(i\,kf\right). \tag{39}$$

The factor $\exp\left(2\pi iW\left(\rho,\phi\right)\right)$ can be added to Eq. (39) to take into account lens aberrations. Thus, knowledge of the electric field distribution in the exit-pupil allows easy to calculate the electric field in Fourier space. We verified by numerical calculation that for a lens with high numerical aperture *NA* < 0.9 the direct field calculations in the focal plane by using the Sratton-Chu method and calculation by using the approximate Fourier spectrum obtained from (39) with consequent inverse Fourier transform give for the main light spot the same electric field intensity distribution (with an error less than 0.1 %)

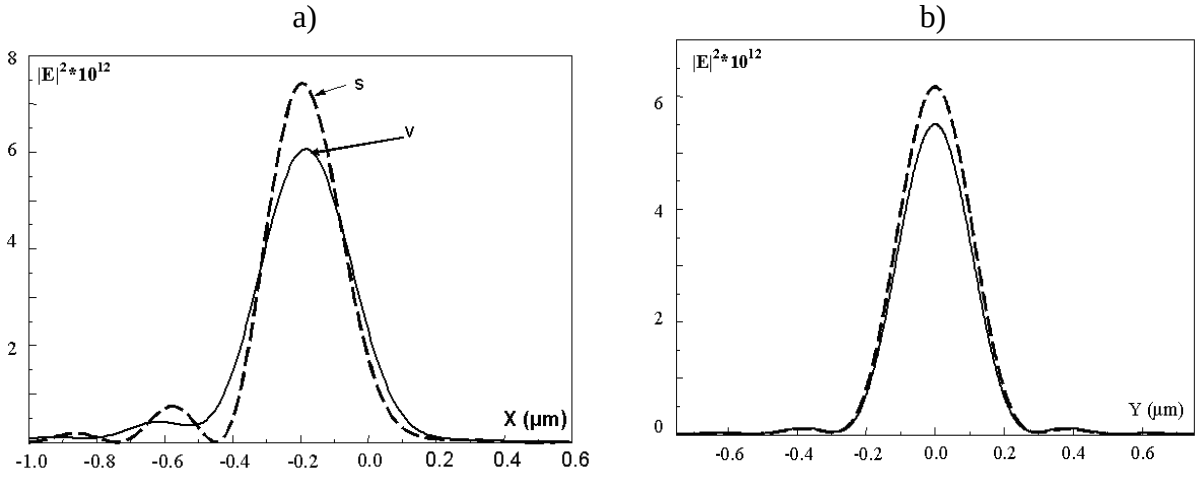


Fig. 8. Electric field intensity in two orthogonal cross sections (along X and Y axes) for the case of x - polarized homogeneous field of incident beam at entrance-pupil plane for a lens with coma aberration. __ = 0.2 _ rms; - - - scalar theory; —— vector theory; *NA* = 0.85. (a) Field intensity along X axis. (b) Field intensity along Y axis.

Comparison of field intensity distribution obtained from numerical simulation by using the algorithm elaborated above with the data from [1] has shown that they agree with an error slightly above 1%. From a comparison of field intensities at the focal plane obtained from scalar and vector theories for a lens with large numerical aperture, which are shown in Fig. 8 and 9, it is clear that vector theory gives a wider main spot with decreased field intensity at the spot center and smaller side-lobe intensity around it. For example, from the curves in

Fig. 8 one can see that scalar theory gives a more intense and narrower main spot and side lobes with larger intensity than vector theory for the case of a lens with coma aberration. Fig. 9 demonstrates field intensity calculated by scalar and vector theory for a lens with spherical aberration. As one can see, the tendency of smaller intensity for the main spot is also valid for the lens with the above-mentioned aberration. The spot shape for the lens with spherical aberration has a different distribution along a different cross section, depending on polarization of the incident beam, as is clear from a comparison of the intensity plots in Fig.s 9(a, b) obtained by vector theory.

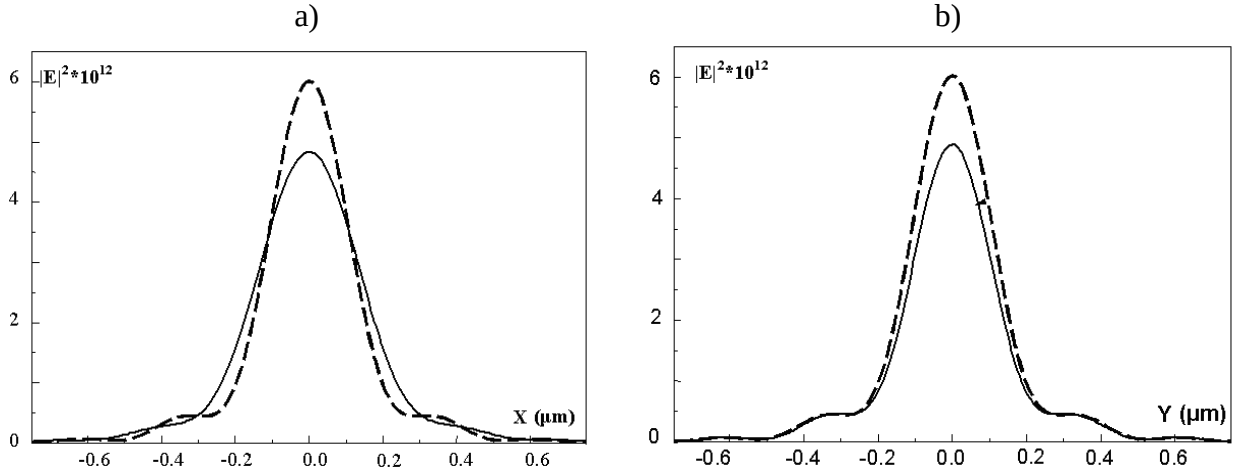


Fig. 9. Electric field intensity in two orthogonal cross sections (along X and Y axes) for the case of x - polarized homogeneous field of incident beam at entrance-pupil plane for a lens with spherical aberration __ = 0.1 _ rms; - - - scalar theory; —— vector theory; *NA* = 0.85. (a) Field intensity along X axis. (b) Field intensity along Y axis a lens with numerical aperture *NA* < 0.9 and possessing Zernike's coefficients $|C_{ik}| < 1$ (for a lens with small aberration).

The algorithm for simulation of focused beam propagation in multilayer medium is based on expansion of electromagnetic fields of this beam in a set of plane waves. The focused beam is represented as a sum (integral) over plane waves by the use of Fourier transform. Every plane wave is represented as sum of *s*− and *p*−polarized beams (see Fig. 10). For every *s*−and *p*−polarized beam, the problem of its propagation through a parallel multilevel structure is solved by calculation of Fresnel coefficients at every interface followed by application of scattering-matrix theory for calculation to calculate the general scattering matrix for all multilayer structure. From general scattering matrix we can calculate plane wave amplitude at top and bottom surfaces of the structure. Then, the field in any layer can be calculated by the using the theory of transmission lines:

$$\begin{aligned} E_t(z_j - L_j) &= E_t(z_j)\cos(\gamma_j L_j) + i\rho_j H_t(z_j)\sin(\gamma_j L_j), \\ H_t(z_j - L_j) &= i\frac{E_t(z_j)}{\rho_j}\sin(\gamma_j L_j) + H_t(z_j)\cos(\gamma_j L_j), \end{aligned} \tag{40}$$

where E_t and H_t are tangential components of electric and magnetic field, respectively; ρ_j is the wave resistance of the *j*-th layer, L_j is length of *j*-th layer.

The fields of all plane waves should be summed to calculate the field distribution of propagating beam inside volume of multilayer structure. The field calculations utilizing the Fourier transform for a focused beam has good convergence only near the focal plane (the distance from the focal plane does not exceed several wavelengths). Other methods should be applied to field calculations far from the focal plane.

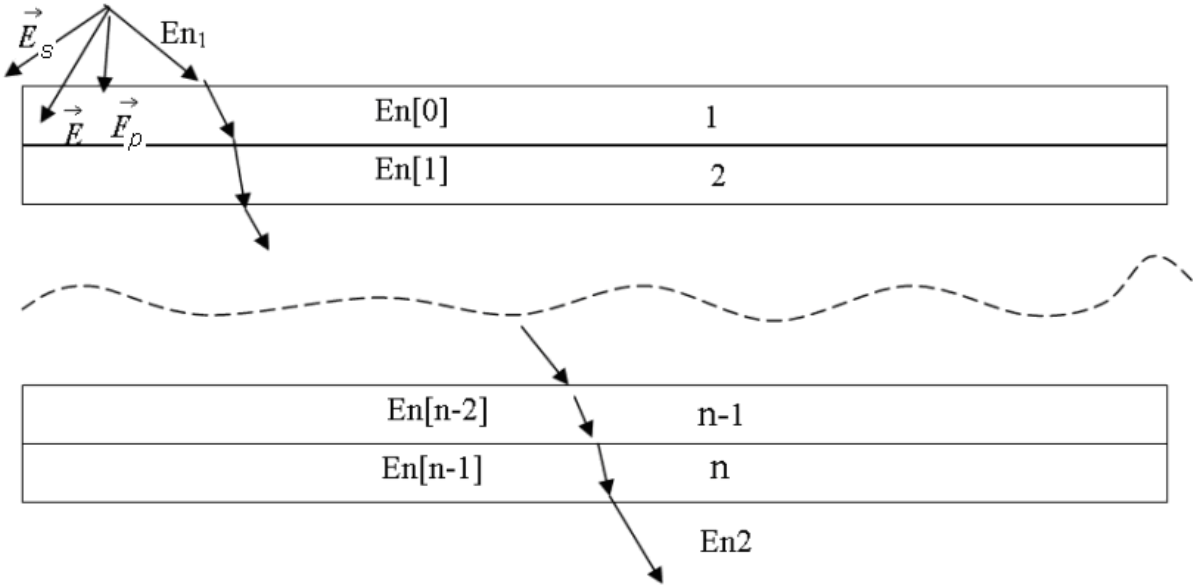


Fig. 10. Optical scheme of plane-wave propagation in a multi-layer medium.

1.6. Features of the optical system of multi-level data storage

Multilevel method has several features, which specifies special requirements to the optical system of optical head and on an optical disk layer structure. The each layer is located at various depths from the surface of a disk. It is well known that a beam waterfront passed through a layer of the thickness h with a refraction index n obtained a spherical aberration of nNA^4h/λ. The optical system for multilevel optical data storage should have an opportunity to move a focal plane up to 1 mm in the depth of an optical disk and therefore the value of spherical aberration of the beam will be changed with a change of the depth. However the problem of compensation of spherical aberration is already solved by using a liquid crystal plate with a circular phase adjuster [20, 21], and therefore it will be not analyzed here.

Fig. 11 shows the principal optical scheme of information reading from a multilevel optical disk. It is clear that the beam reflects from all recording layers and therefore the reflected beam in addition to a useful signal has strong background radiation. Since the optical disk structure should provide approximately the same signal level from all recording layers, the power of background radiation will N0-1 times exceed the power of the useful beam at the surface of the disk. However the only useful beam will be exactly focused at a photodetector and all its power will be detected. Only a small part of power of the beam, reflected from the i-th recording layer will be received by a detector:

$$P_i = P_0 S_d / S_i, \tag{41}$$

where P_0 is power of the beam reflected from i-th layer on the surface of the disk, S_i – the area of light spot of the beam reflected from the i-th recording layer (image plane) in detector plane, S_d is the detector area. Now we could write a signal to background ratio as:

$$\eta_{fon} = 2\sum_{i=1}^{(N0-1)/2} \frac{I_0 S_d}{S_i} / \left(I_0 \frac{S_d}{S_0} \right). \tag{42}$$

where S_0 is the light spot area in a detector plane of the beam reflected from a current recording layer. Provided that $S_0 = S_d$ and since:

$$S_0 / S_i = \left(\frac{\lambda}{NA} \right)^2 / H^2 (iNA)^2 = \left(\frac{\lambda}{H} \right)^2 \frac{1}{(iNA)^4}. \tag{43}$$

The Eq. (42) could be rewritten as follows:

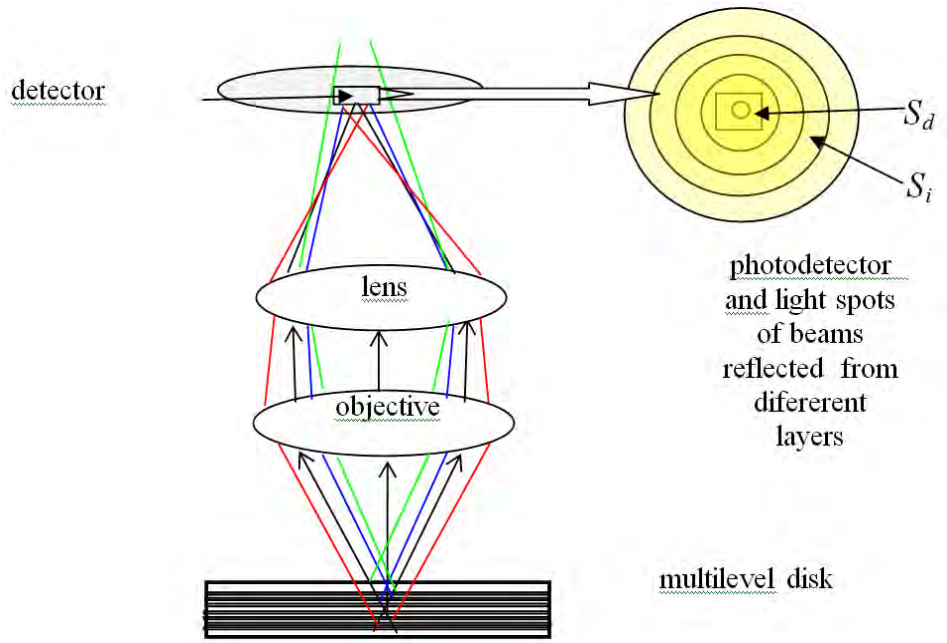


Fig. 11. Principal optical scheme of light intensity distribution in a detector plane for the beams reflected from different recording layers.

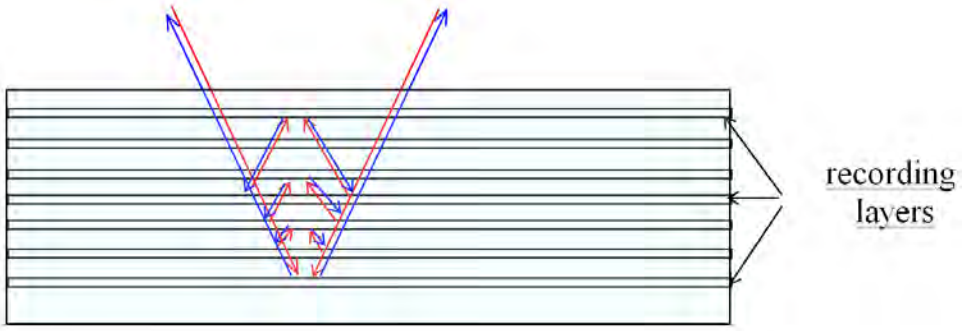


Fig. 12. Optical scheme of beam reflection. It is shown only re-reflected beams which have image plane close to image plane of beam reflected from recording layer.

$$\eta_{fon} = 2\left(\frac{\lambda}{H} \right)^2 \sum_{i=1}^{(N0-1)/2} \frac{1}{(iNA)^4} \tag{44}$$

From equation above it is clear that in the case of sufficiently small detector area and relatively large distance between recording layers $\left(\frac{\lambda}{H\ NA^2}\right)^2 << 1$, the background signal can be reduced to low level do obtain sufficiently low signal to noise ratio. However the partial coherency of useful and background beam could significantly decrease signal quality due to interference effect:

$$\begin{aligned} I \sim \mathrm{Re}\left(\left(E_{sig} + E_{noise}\right)\left(\bar{E}_{sig} + \bar{E}_{noise}\right)\right) = \\ = I_{sig} + I_{noise} + 2\sqrt{I_{sig} I_{noise}}\ \bar{\cos}(\varphi) = I_{sig}\left(1 + 2\ \bar{\cos}(\varphi)\sqrt{I_{noise}/I_{sig}} + I_{noise}/I_{sig}\right) \end{aligned} \quad (45)$$

where φ is phase angle shift between the fields of beam with signal and background field, $\bar{\cos}(\varphi)$ average value of cos(φ).

For incoherent light the averaging results in close to zero value and therefore in this case the noise increase due to interference effect will be small. Therefore in multilevel data storage it is better to use a wideband laser diode or several laser diodes with slightly different wavelengths to decrease beam coherence and hence to suppress signal noise due to interference effect.

The numerical aperture of objective increase allows to decrease the distance between layers as:

$$H_{\min} = 1/NA^2, \quad (46)$$

due to decrease of depth of focus. Therefore, the using of objective with large numerical aperture allows significantly decrease the distance between recording layers and increase the data storage capacity.

There is also a background light from re-reflected beams. The signal from re-reflected beams can be estimated as follows. The reflection coefficient of recording layer increases with its depth and should be close to 1 for the last layer. Therefore, the amount of energy that penetrates through all layers will be close to zero. On the other hand, it was found above that about half of beam energy will be used to create a signal (energy reflected from all recording layers). Thus, for the lossless layers, the beam energy will be equally divides between the energy used to create a signal and energy of re-reflected light. Since the energy that goes to a signal is split equally between N_0 recording layers, the energy of signal will be less than the energy of re-reflected light by a factor $1/N_0$. The above implies that re-reflected beams will create approximately the same background signal as the beams reflected from inactive recording layers. Since the light spots from re-reflected beams also will be blurred only a small part of their energy hits the photodetector. However, for some re-reflected beams will have image plane close to photodetector. These are recording layers symmetrically situated relatively working recording layer (every second of them). For these recording layers the shift of their image plane is not determined by the distance between the layers but by the difference in distance between layers (Fig. 12). The most numbers of such re-reflected beam exist for the case of reading information from the lowest layer (N-1) / 2. Fig. 12 shows this case for the multilevel disk with seven recording layers. The total energy of this type of re-reflected beams will not be larger than:

$$P / P_0 = R^3 (N_0 - 1) / 2 \sim \frac{1}{2}(N_0 - 1) / N_0^{\,3} \approx \frac{1}{2} / N_0^{\,2}, \qquad (47)$$

where R is reflection coefficient of one recording layer, P_0 – initial energy of the beam, N_0 – total number of recording layers .

Thus, this part of re-reflection energy is smaller by a factor $2N$ than the energy of the signal ($1/2N_0$). However, since these re-reflected beams will create signal with the same frequency as the information signal, they can significantly affect the signal quality. Therefore, the special attention should be paid to achieve unequal distances between the recording layers to minimize cross signal from the neighboring recording layers. The crosstalk signals from the lower layers are much greater than from the upper layers since they have sufficinlly larger reflection coefficients. Therefore the special attention should be paid to situated the lower recording layers with large distances difference between them.

Re-reflection, due to interference effect, can lead to the degradation of the main maximum of a focused beam and to appearance of wide side maximum. The wide side maximum results in increase of a cross-talk signal from adjacent pits. Different rays of a focused beam have different directions of propagation and hence have different phase shift when propagating from one recording layer to another (Fig. 13). In the case of equal distances between recordings layers a ray with a phase shift:

$$\gamma d = kd \ \cos(\varphi) = (n + 0.5)\frac{\lambda}{2}, \qquad (48)$$

reflected back rays from two adjacent layers interfere negatively and propagated forward ray has no power losses. For rays which have a direction of propagation satisfying Bragg's law:

$$\gamma d = kd \ \cos(\varphi) = n\frac{\lambda}{2}, \qquad (49)$$

the reflected rays increase each other, and on the contrary, the amplitude of propagated forward ray decreases strongly. Hence the spatial frequencies corresponding to a Bragg's law will disappear in propagating forward beam. The rays with close directions to these angles, due to many re-reflections, will move in the lateral direction and create wide side maximum in a propagating forward beam.

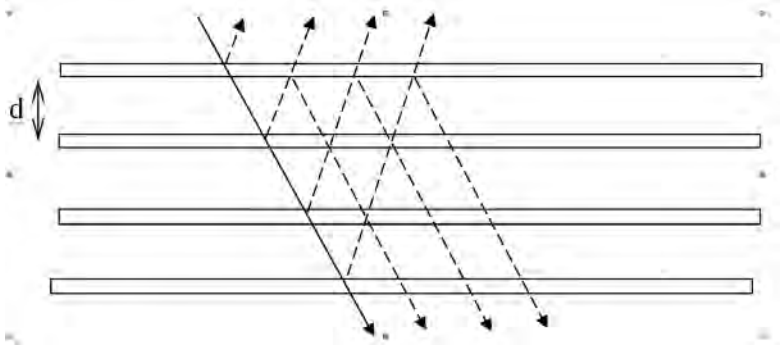
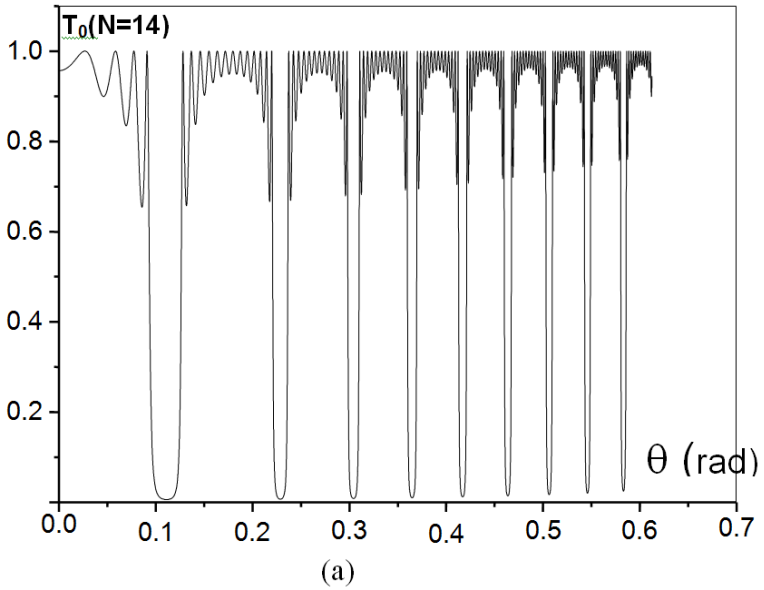
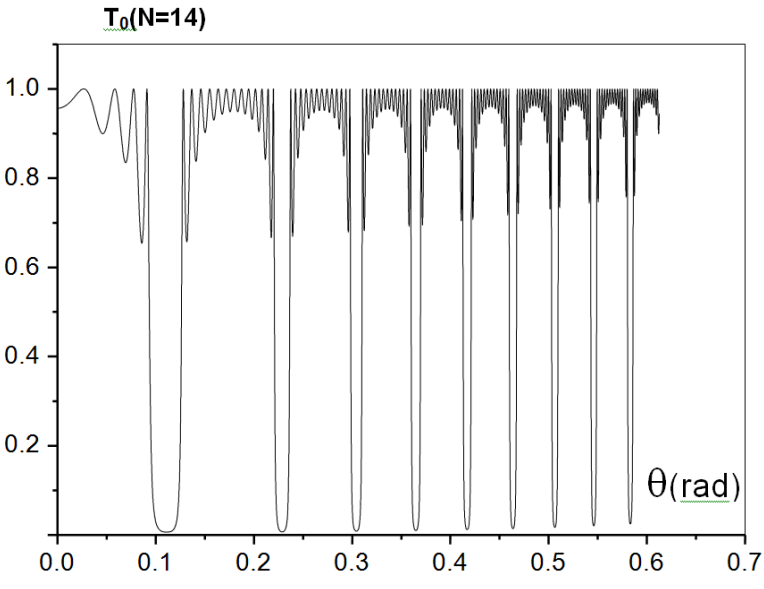


Fig. 13. Optical scheme of ray re-reflection in multilayer medium.



(a)



(b)

Fig. 14. Reflection coefficient (intensity) through 14 4,5 nm thick TiO_2 layers in a plastic of medium (n=1.4) with interlayer distance d=7100 nm; λ=400 nm: a) p- polarization; b) s-polarization.

Fig. 14 shows the dependence of the transmission coefficient of the plane wave through 14 identical equidistance thin TiO_2 layers on the angle of plane wave propagation. As it was expected, the multilayer structure is a comb filter of spatial frequencies, which erases the narrow spatial frequencies band in propagating forward beam. The numerical simulation has shown that in the case of different distances between recording layers the result is similar to the constant distance case.

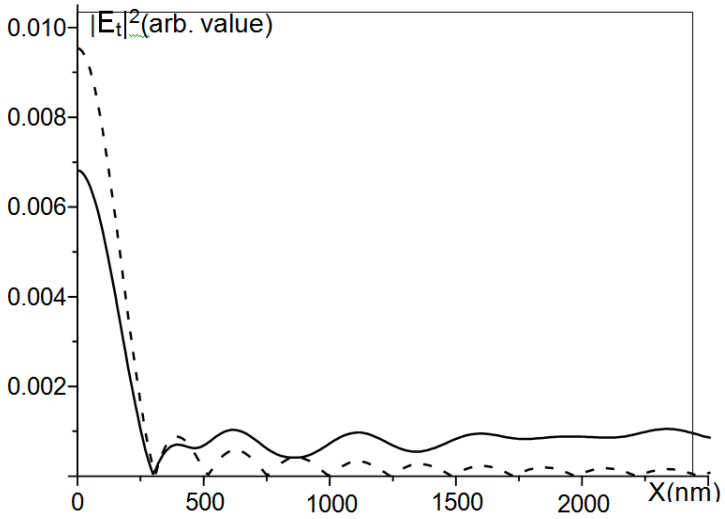


Fig. 15. Light spot intensity of p-polarized beam in the focal plane (2D case, *NA*0,8 with compensation of spherical aberration), which is located inside optical disk: dash line – homogenous medium; solid line– a medium with 14 thin (4.5 nm) TiO_2 layers with separation distances between them of *d*=7100 nm.

Fig. 15 shows the field intensity distribution of Gaussian focused beam (at the edges of the lens field intensity decreases to I_0/e^2) with numerical aperture 0.8 in the focal plane for 2D case (the homogeneous field along the second transverse coordinates). All layers have the same thickness and the same distances between them. Fig. 15 also shows the focal plane beam shape for the case of an optical disk without recording layers to show clear the influence of light reflection and interference.

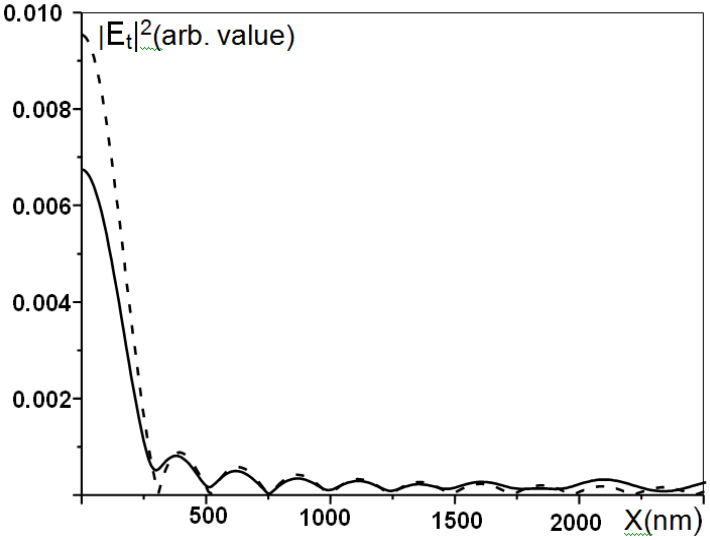


Fig. 16. Light spot intensity of p-polarized beam in the focal plane (2D case, *NA*=0.8 with compensation of spherical aberration), which is located inside optical disk: dash line –a homogenous medium; solid line– medium with 14 thin (4.5 nm) TiO_2 layers with initial separation distance between them of *d*=7100 nm which increases by 5 nm with every shift by one layer.

Fig. 16 shows the distribution of the field intensity in the focal plane for the case when the distance between layers constantly increases by 5 nm (the distance between the first and second layers is d = 7100 nm). One can see from curves on Fig. 16 that a small change of distance between adjacent layers results in a substantial decrease in the intensity of side background radiation. Further increase in the distance difference between adjacent layers does not substantially change the intensity of field distribution in the focal plane.

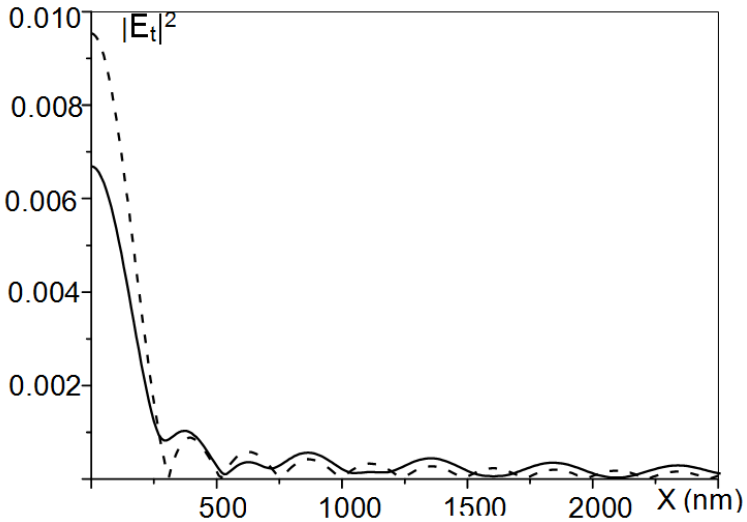


Fig. 17. Light spot intensity of p-polarized beam in the focal plane (2D case, *NA*=0.8 with compensation of spherical aberration), which is located inside optical disk: dash line –a homogenous medium; solid line – a medium with 14 thin (4.5 nm) TiO_2 layers with initial separation distance between them of *d*=7100 nm which increases by 20 nm with every shift by one layer.

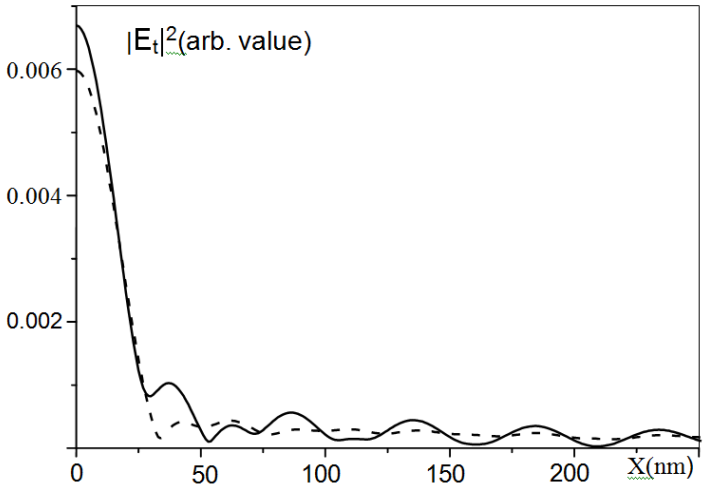


Fig. 18. Light spot intensity of p-polarized beam in the focal plane (2D case, *NA*=0.8 with compensation of spherical aberration), which is located inside optical disk with 14 thin (4.5 nm) TiO_2 layers with initial separation distance between them of *d* =7100 nm which increases by 20 nm with every shift by one layer: dash line – *s*-polarization; solid line – *p*-polarization.

This can be seen in Fig. 17, which shows the intensity distribution in the focus for *p*-polarized wave when the distance between adjacent layers increases immediately to 20nm (while maintaining all other parameters of the structure of the disc and the beam). The intensity distribution of *s*-polarized beam has the same width of the main maximum and approximately the same intensity of side lobes (Fig. 18), except amplitude of the first side lobe which is substantially less than intensity of *p*-polarized beam and the field intensity in the side spot for *s*-polarization, contrary to p-polarization, has not significant oscillations. An increase of intensity of a side light spot is inevitable price for a multi-level optical storage, because in multilayer medium the ray re-reflection is unavoidable, which leads to a lateral displacement of the part of beam energy. Expanding the size and increasing intensity of light of side lobes leads inevitably to an increase of crosstalk between pits. The pits in the area of side light spot reflect light back. According to the reciprocity principle the part of reflected light from nearby pits would fall on the detector independently of its size in the same way as the part of light of a focused beam creates a lateral maximum in the focal plane. Lower levels of a signal and stronger background light, strong crosstalk signal call for a powerful method of error protection.

1.7. Conclusion

We analyzed a multi-level optical disk with two different disk structures:
1) a multi-level optical disk having thin recording layers made from conventional media used in an optical disk for recording layers: metal layers (ROM media type) and ultrathin layers GeSbTe (R and RW media types);
2) a multi-level optical disk with recording layers made from a transparent thin dielectric with a large refraction index.

A differential equation for the optimal reflection coefficient of recording layers was derived for the case when recording layers had large dissipative losses (recording layers on the basis of thin metal and GeSbTe films). The solution of this equation gave a simple formula for the maximum number of recording layers depending on the permissible level of a useful signal and recording layers parameters. It is shown that this method can not get more than 5-15-layers on one side of an optical disk.

A differential equation for the optimal reflection coefficient of recording layers for the case of recording layers without dissipative losses is derived. By solving this equation we obtained a simple formula for dependence of a maximum number of the recording layers on the permissible level of a useful signal. It is shown that in the case of recording layers without dissipation it is possible to obtain the ROM optical disk having up to 100 recording layers with total information capacity of 200-2000 Gbytse.

It is shown that a low crosstalk signal from adjacent layers can be received by changing separation distance between recording layers, especially between the lower recording layers and choosing proper transverse size of a detector.

1.8. Near-field optical data storage

1.8.1. Scanning near-field microscope

Near-field optical data storage uses scanning near field microscope [22, 23] for read/write data on an optical disk. The main part of the scanning near-field optical microscope (SNOM) is a probe that consists of a cone-type optical transparent dielectric with the side surface covered by an opaque metal film. On the apex of the cone, there is a small hole in the metal coating. The light beam, excited either by an incident focused beam, or by a dielectric

waveguide situated at the wide side of the cone, is propagating along the cone and strikes a hole at the apex of the probe. The beam size outside the probe, close to the hole, is approximately equal to the diameter of the hole, and therefore, can be smaller than a half of the wavelength. The beam size determines the lateral resolution of microscope and therefore it can be as small as hole diameter. Unfortunately, high spatial resolution of SNOM implies too low optical efficiency coefficient, since it drastically reduces with decrease of the hole size. There are two parameters used for calculation of the optical efficiency in SNOM: 1) far-field transmission coefficient, or simply transmission coefficient; 2) field enhancement.

For usual probe with a large convergence angle and a small aperture, the transmission coefficient, k_f, decreases rapidly with increasing the wavelength, λ, as [24- 26]:

$$k_f \sim \left(d / \lambda\right)^6, \tag{50}$$

where d is the hole diameter.

Usually, the transmission coefficient is much lower than 10^{-4} even for a relatively large hole size of 100 nm [27]. Low optical efficiency is due to a low transmission coefficient of the light propagating through a small hole in an opaque screen, as well as due to an evanescent region near the apex of the probe.

The field intensity enhancement, k_{fe}, is used to characterize the near-field interaction of SNOM with the scanned sample, which is the ratio of the field intensity at the SNOM aperture, $\left|\vec{E}_a\right|^2$, to the field intensity of an incident beam, $\left|\vec{E}_{inc}\right|^2$ [22, 27]:

$$k_{fe} = \left|\vec{E}_a\right|^2 / \left|\vec{E}_{inc}\right|^2. \tag{51}$$

Here, we also use a near-field transmission coefficient, k_n, defined as follows (for the case of SNOM scanning an object with losses) [28]:

$$k_n = P_a / P_{inc}, \tag{52}$$

where P_a is the total energy flow through the aperture (due to the near-field interaction, almost all energy is absorbed by a scanned object), and P_{inc} is the energy of an incident beam (this definition is similar to that for the far-field transmission case).

The near-field transmission coefficient is dependent upon the value of losses in the medium of the scanned sample; it is very low (<<0.001) for a conventional SNOM scanning of any medium. Fig. 19 shows an optical scheme which can be used for an optical data storage [29]. An optical scheme in Fig. 19a), c) illuminates an optical disk by near-field and detects a signal in far-field and on the optical scheme in Fig. 19b), to the contrary, an optical disk is illuminated by far-field and a signal is detected in near-field. Only an optical scheme shown in Fig. 19 uses near-field for illumination and signal detection. Using far-field for illumination or for signal detection causes a strong noise due to an interaction of far-field with a large area of an optical disk. A scheme using near-field for illumination and signal detection (Fig. 19c) has a significantly weaker noise but a very high background signal from a beam reflected back from probe apex. Weak probe beam interaction with a disk and results in a very weak useful signal and a strong background signal. That fact makes practically impossible to

separate a useful from background signals in this mode of operation. Therefore all experimental data in near-field data recording were obtained in modes which use both near-field and far-field for data recording: reflection mode, reflection – collection mode, illumination mode. For this case signal amplitude is determined by the far-field transmission coefficient. The first experimental results in near-field optical data storage were obtained by Betzig [30]. Betzig recorded magnetic marks in Co/Pt multi-layers magnetic medium (normal polarization of magnetic medium) with ten repetitions of $4A^0$ Co i $10A^0$ Pt sputter coated onto a glass substrate. A recording mark was obtained by heating a magnetic medium over Curie temperature (300^0 C) of a magnetic medium. There was obtained a magnetic mark structure with periodicity of 120 nm in both directions corresponding to a storage density of 45 Gbits/in^2, what is 100 times more dense than on CD disc and is several times more dense than on BD. Several other laboratories demonstrated near-field optical recording on a polymer film [31-33]. However due to low optical throughput the data transfer rate is extremely low, it is about 10 kHz/s. This rate is lower by a factor of 10^3 than in modern optical data storage.

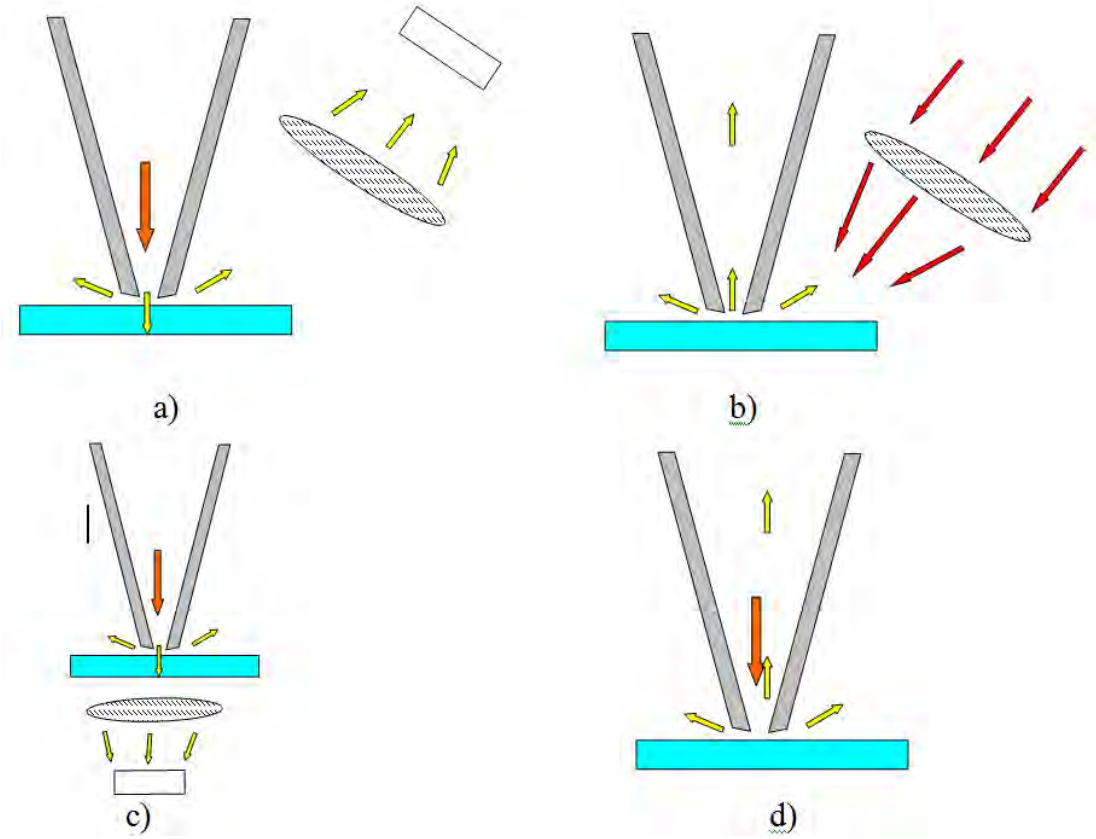


Fig. 19. Principal optical schemes of a near-field optical data storage using scanning near-field optical microscope operating in different modes: a) reflection; b) reflection - collection; c) illumination; d) illumination - collection.

Practical application of near-field microscope for optical data storage requires a great increase in the data transfer rate and hence a new type of near-field probe with a significantly higher optical efficiency. Several new probe structures which show a huge increase in optical efficiency were proposed in last decade [34-38].

However till now there are no experimental results with appropriate for technological application data exchange rate and theoretical foundation for viable near-field optical data storage.

1.9. Pyramidal shape near-field microstrip probe

We proposed a SNOM with a microstrip probe for significant improvement of optical efficiency [39-50]. One of possible types of a microstrip probe is the pyramid-type microstrip probe (PTMP) (see schematic drawing in Fig. 20). The PTMP has a transparent pyramid-like core with a truncated corner. Metal strips coat two opposite sidewalls of the pyramid. The transparent body and two metal strips form a tapering microstrip line, similar to an ordinary microstrip line where two opposite sides of a dielectric rectangular slab are coated with metal films, as shown in Fig. 21. The incident beam (either a focused beam or a dielectric waveguide mode) couples to the probe through its wide end, and propagates along the probe, reaching the narrow end that forms the aperture. The light passing through the narrow end interacts with the scanned sample. In far-infrared band metal strips can be represented with high accuracy as perfect conductors which can support quasi-TEM wave which has no cut-off size. The incident light should have electric field polarization orthogonal to the metal strips in order to excite the quasi-TEM mode that has no cut-off size.

Numerical simulations based on a simplified near-field probe model, which represents a probe as tapered microstrip line with low reflection and dissipative losses has shown that the microstrip probe has a high far-field transmission coefficient [28, 39-40] which can be calculated by using simple formula:

$$k_f = P_r / P_0 = (ka)^2 / k_l \tag{53}$$

where k is wavenumber, a is linear aperture size, and k_l is the power attenuation coefficient of the incident beam in metal strips, P_0, P_r are energies of initial and radiated beams, respectively.

For the case of scanning surface with dissipative losses the near field interaction of a microstrip probe with the scanning surface results in energy transfer from probe to surface through induction current. The near-field transmission coefficient (calculated by using simplified model) [28, 39-40] can be written as follows:

$$k_f = P_l / P_0 \sim A\varepsilon_r ka / k_l \tag{54}$$

where P_l is energy losses in scanning sample due to current induced by near-field interaction (due to near-field interaction), A is the dimensionless coefficient which depends on permittivity of probe dielectric core.

From Eq. (53) follows that a microstrip probe has a significant advantage over a conventional near field probe in far field transmission coefficient, especially for the small aperture size (a<100 nm) since it decreases with decrease of the aperture size a as a square of the aperture diameter. Comparison of far-ield (Eq. 53) and near-field (Eq. 54) transmission coefficients shows that for the small aperture size the latter will decrease slower with decreasing the aperture size a and for small aperture size.

Therefore, the near-field interaction will dominate in probe interaction with scanning surface in the case of scanning area with large dissipative losses. Since the near-field interaction results in dissipative losses (no radiation) it can not be detected in far-field. Therefore for detection of strong near-field interaction of a microstrip probe with the sample with large dissipative losses the illumination–collection mode of operation of SNOM should

be used. The illumination–collection mode of operation of SNOM does not require a lens and uses only a near-field probe to illuminate and detect a signal and therefore is simpler and more easy for fabrication.

It follows from analysis above that for optical data recording with a microstrip probe the strongest signal can be obtained for using a recording medium with large dissipative losses and in operation in illumination – collection mode. Since conventional recording layers used in optical data storage such as GeSbTe and metal films have great losses they can be well applied for near-field data storage based on microstrip probe.

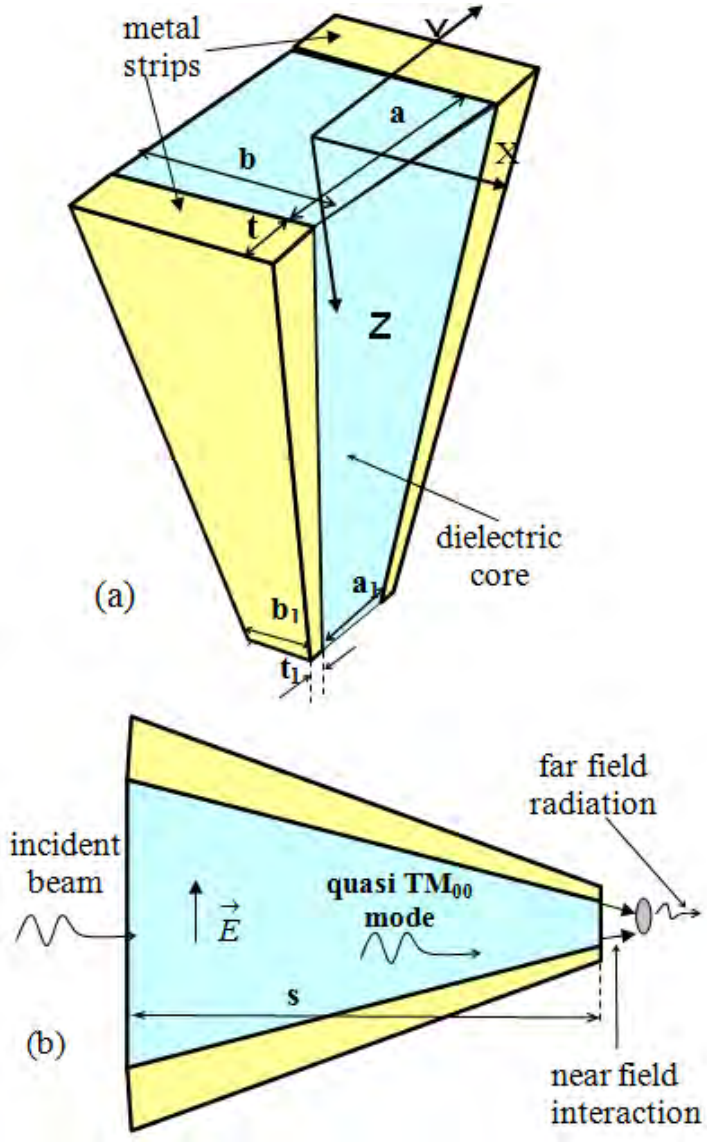


Fig. 20. (a) Schematic representation of a cone-type SNOM probe; (b) Diagram showing light propagation in the probe.

However, it is unlikely that a simplified model, based on the properties of quasi-TEM wave of a microstrip line with ideal metal strips (Fig. 21), used in Refs. [28, 39-40], is sufficient for accurate calculation of characteristics of PTMP in optical band due to surface plasmon resonance phenomena. Therefore, a numerical simulation on the bases of a rigorous theory is required to obtain reliable theoretical results.

1.10. Quasi TM_{00} modes of microstrip line in optical band

The near-field probe can be represented as a tapered waveguide, and in our case as a tapered microstrip line. Therefore for clear understanding of beam propagation in a near-field probe and its interaction with a scanning object through aperture it is very important to know dispersion characteristics and the field structure of waveguide modes. The microstrip line structure is shown in Fig. 21 [51].

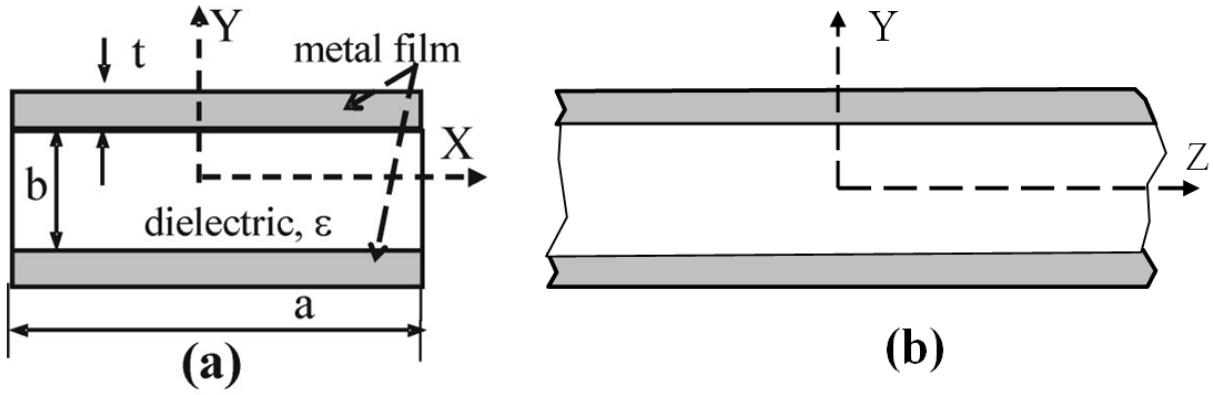


Fig. 21. Cross-sectional view of an optical microstrip line.

We are interested mainly in a several first quasi-TM_{nm} modes (surface plasmon waves), which can propagate in the region near the probe apex of the NSOM, where the first subscript *n* denotes the field dependence on the *x* axis (along the layer in the microstrip's cross section) and the second subscript *m* on the *y* axis (coordinate across the layer). It should be noted that for the complex structure such as a microstrip line this notation is somewhat ambiguous because of the complexity of the field structure of microstrip modes. For small microstrip sizes, however, only a several waves can propagates in the line.

The properties of these modes are assumed to be derived from the properties of the TM_n waves propagating in the infinite two-dimensional dielectric–metal-layer structures analyzed above, and thus our notation for the microstrip mode is based on this fact. Since the structure has two symmetry planes at x=0 and y=0 and it is assumed that the excitation field has the same symmetry it is possible to place magnetic wall at x=0 and electric wall at y=0. We used finite difference time domain method for accurate simulation of waveguide modes in microstrip line.

The excitation of a microstrip nanowaveguide by the field of a mode of a microstrip line with perfectly conducting strips was used to excite waveguide modes. It was found that for this symmetry and for small microstip sizes ($a \le \lambda/2$; $b \le \lambda/2$) only one quasi-TM_{00} microstrip mode can propagate in a microstrip line. The rigorous simulation has shown that properties of fundamental quasi quasi-TM_{00} can be derived from TM_0 modes of the plane metal dielectric structure by replacing the side walls of microstrip line by impedance plane (mode matching method).

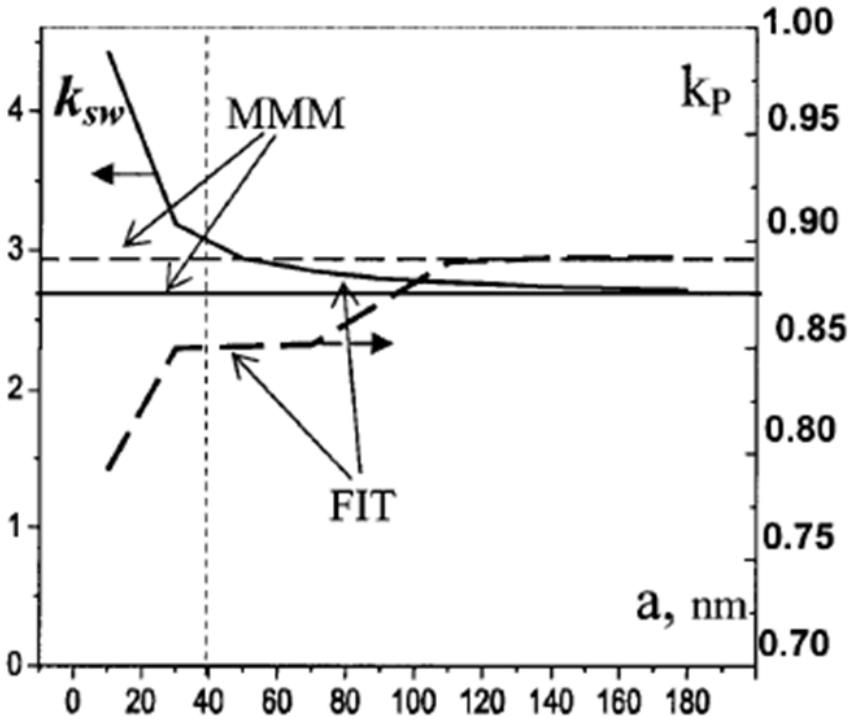


Fig. 22. Slow-wave factor (solid curve) and attenuation coefficient (dashed curve) of the quasi-TM_{00} mode as a function of microstrip width. The microstrip has a core layer ε_1= 2.25 and 20 nm thick silver strips with a height of b= 40 nm operating at λ=790 nm.

Fig. 22 shows the characteristics of a fundamental quasi-TM_{00} microstrip mode calculated by the rigorous FIT algorithm and by the simplified MMM in which the waveguide side walls are replaced by magnetic walls (φ= 0, n= 0).

The simplified MMM with magnetic sidewalls measures any dependence of the slow-wave factor and the attenuation coefficient on microstrip width (TEM mode approximation), whereas a rigorous FIT model gives a sharp increase in the slow-wave factor for a narrow ($a >> b$) microstrip line.

The difference between mode characteristics calculated by two methods decreases quickly with increasing microstrip width and approaches zero. For a wide microstrip line ($a/b>>1$) the characteristics of quasi-TM_{00} modes (the slow-wave factor and the attenuation coefficient) obtained by both methods agree well. Note that the slow-wave factor and the attenuation coefficient calculated by the FIT are always larger than those obtained by the MMM.

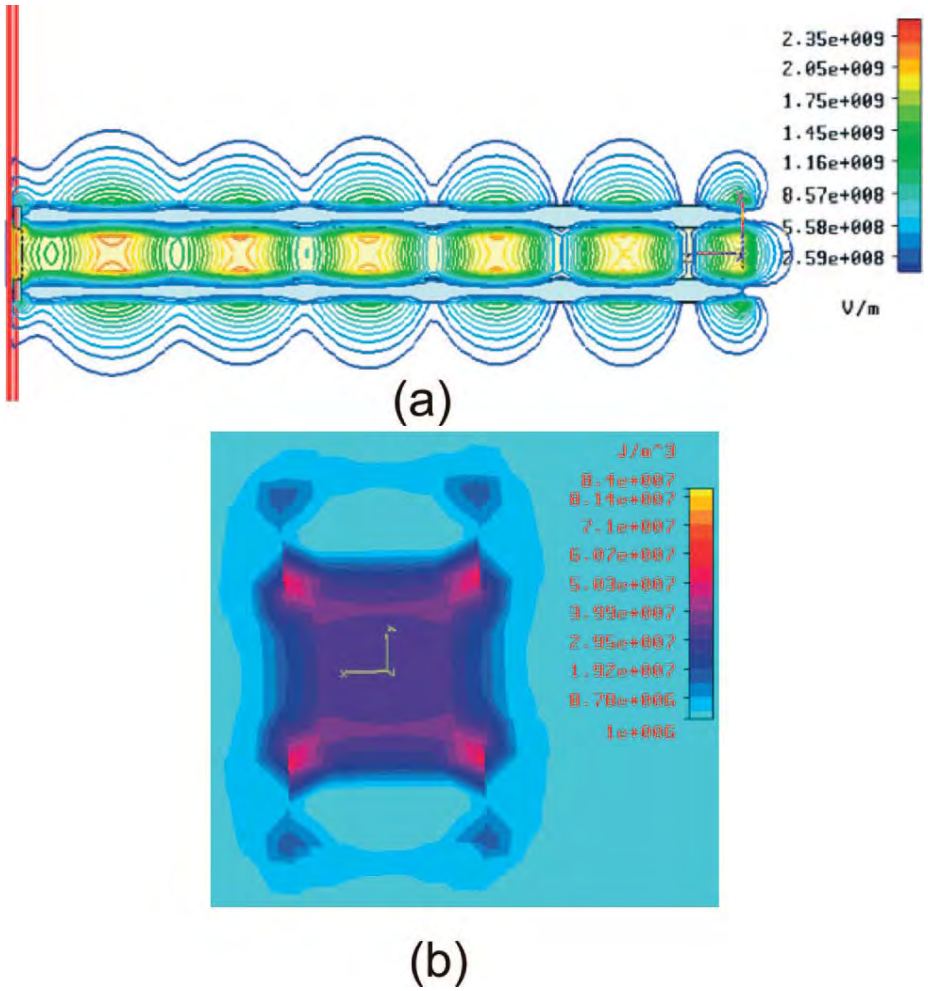


Fig. 23. (a) Electric field amplitude of the quasi-TM_{00} mode along a microstrip line (at x = 0) and (b) electric field energy density at the cross section (at z = 30 nm) in a microstrip line with a length of s =160 nm covered with silver metal strips: a = b= 10 nm, t = 4.5 nm, ε_1 = 4, λ= 650 nm , γ/k = 14.0 calculated by the FIT and k/γ = 12.4 by a MMM).

Fig. 23 shows the electric field amplitude and the energy density distribution of the quasi-TM_{00} mode of the microstrip line. One can see that the cross-sectional size of the quasi-TM_{00} field is approximately equal to the size of a microstrip's cross section. From the electric field distribution in Fig. 23 one can see that it has large peaks at the metal strip ends, with large field penetration into the metal. It is these field intensity peaks in the areas of the metal edges that make the slow-wave factor and the attenuation coefficient of the TM_{00} mode of the microstrip line larger than those of the TM_0 mode of the planar structure.

Therefore the simplified MMM to microstrip lines can be applied only for qualitative analysis of the characteristics of a TM_{00} mode of an optical microstrip, and a more rigorous model is needed for accurate calculation. It should also be noted that the electric field of the fundamental mode of an optical microstrip has a loop of the electric field near the microstrip end, similar to that for the conventional microstrip line (Fig. 23(a)).

As a width of the fundamental mode of an optical microstrip is approximately equal to the microstrip line transverse structure size, one can expect that it will be possible to obtain a beam diameter of several nanometers for a tiny microstrip. A field pattern with a beam size smaller than 10 nm propagating in a dielectric rod covered with two golden strips is shown in Fig. 24. This tiny-sized wave has a slow-wave factor as large as 31.1 and great power loss that is due to strong field penetration in the metal.

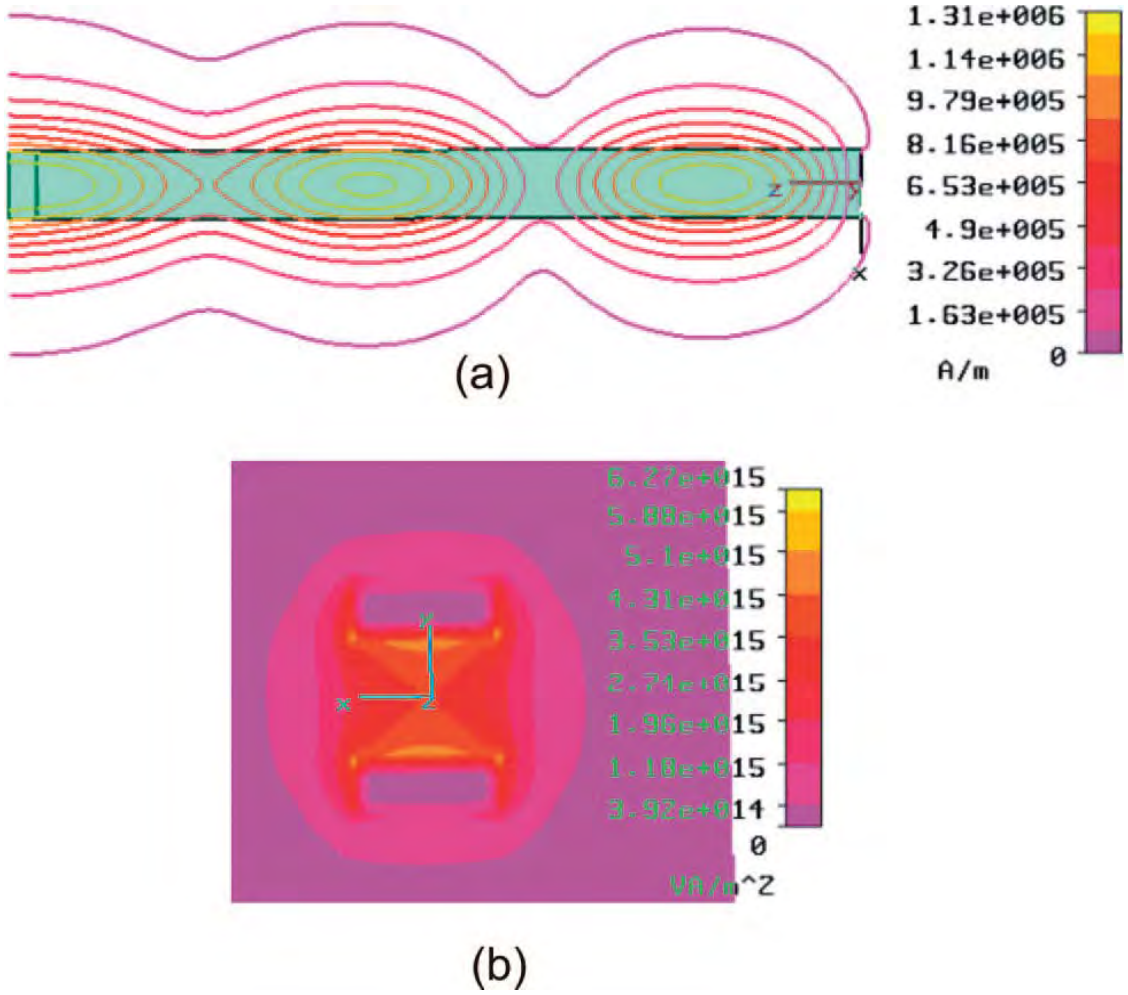


Fig. 24. (a) Magnetic field amplitude of the quasi-TM_{00} mode along a microstrip (at $y = 0$) (b) Power flow (z components) at the microstrip cross section (at $z = 20$ nm) in an optical microstrip line with a length of $s = 60$ nm covered with gold metal strips: $a = b = 5$ nm, $t = 1.5$ nm, $\varepsilon_1 = 4$, $\lambda = 1520$ nm.

1.11. Pyramid-type microstrip probe properties

As mentioned above the near-field probe may be represented as a tapering waveguide, or, in the case of the pyramid-type microstrip probe (PTMP), as a tapering microstrip line. Therefore, knowledge of the waveguide modes is of great importance for understanding features of beam propagation in the SNOM probe and for design a probe with optimized characteristics. It was found [51] that the quasi-TM_{00} has no cut-off frequency, and therefore, can propagate along the whole probe (from its wide end to the apex), which should result in a high field enhancement and high optical throughput.

Numerical simulation has shown [51] that the quasi-TM_{00} mode is confined mostly to the region of the dielectric slab (see Fig. 23, 24); as a result, the characteristic size of the space occupied by the mode in the direction across the layered structure is approximately equal to the height b (see Fig. 21) of a microstrip line. The quasi-TM_{00} mode is the best one for using as a working mode in the PTMP in order to obtain high spatial resolution (owing to a small effective diameter of the beam).

This mode can be easily excited by a focused light spot in the center of the wide end of the PTMP, or by a fundamental mode of an optical waveguide coupled to the wide end of the PTMP. The Microwave Studio Package cannot simulate a focused beam spot, and therefore, to simulate excitation of the beam in a probe, we studied the quasi-TM_{00} fundamental mode propagating in an ideal microstrip line with perfect conducting strips. In all calculations described below, we shall use a symmetrical PTMP (with x=0 and y=0 being the two planes of symmetry of a probe), and the probe excitation by the quasi-TM_{00} wave of a microstrip line with ideally conductive metal strips.

In order to have high optical efficiency, a near-field probe should satisfy the following conditions: (1) High coupling efficiency of an incident laser beam to the quasi-TM_{00} working mode; (2) The ability for the working mode to propagate along the whole length of a probe; (3) Low energy losses and good interaction of the working mode with the aperture.

In view of the fact that, practically, only common-optics devices can be used for the probe excitation, the transverse dimension of the wide side of a probe has to be at least $\lambda/2$. On the other hand, the quasi-TM_{00} mode has a beam width approximately equal to or smaller than one wavelength; furthermore, the beam width decreases rapidly with increasing the dielectric constant and with decreasing the thickness of metal strips [51].

However, to obtain good coupling efficiency between an incident beam and the quasi-TM_{00} mode at the wide end of a probe, the size of an incident beam should be approximately equal to the characteristic size of the beam associated with the quasi-TM_{00} mode. Hence, the dielectric core of a probe should have a small permittivity, and the thickness of the dielectric slab in the plane of interaction, as well as the characteristic size of the excited beam, should not exceed one wavelength.

While propagating through a probe, the quasi-TM_{00} mode experiences a backward reflection, because of the probe tapering. The backward reflection can significantly deteriorate the optical efficiency. In order to reduce this parasitic reflection, the relative change of the microstrip transverse size along the probe must be small on a distance equal to a half of the wavelength of the quasi-TM_{00} mode (i.e., the distance of positive interference with the waves reflected back). This condition may be written as follows:

$$\frac{\lambda}{2a(z)}\frac{da(z)}{dz} = \frac{\lambda_0}{2a(z)}\frac{\mathrm{tg}(\theta_1)}{k_{sw}} << 1, \tag{55a}$$

$$\frac{\lambda}{2b(z)}\frac{db(z)}{dz} = \frac{\lambda_0}{2b(z)}\frac{\mathrm{tg}(\theta_2)}{k_{sw}} << 1, \tag{55b}$$

where λ is the wavelength of the quasi-TM_{00} mode and λ_0 is the wavelength in the free space; θ_1 and θ_2 are the tapering angles of the probe within different planes; k_{sw} is the slow-wave factor [51]; and a and b are the cross-sectional dimensions of the dielectric slab (cf. Figs. 21).

It follows from Eq. (55) that for the same tapering angle, the reflection is larger for small a and b, and hence, the losses of the quasi-TM_{00} mode due to the back-reflection will increase in the vicinity of the probe apex. However, the fast increase of k_{sw} with decreasing the microstrip line height a [51] should significantly decrease the reflection from the region near the probe apex, and therefore, reduce the energy losses.

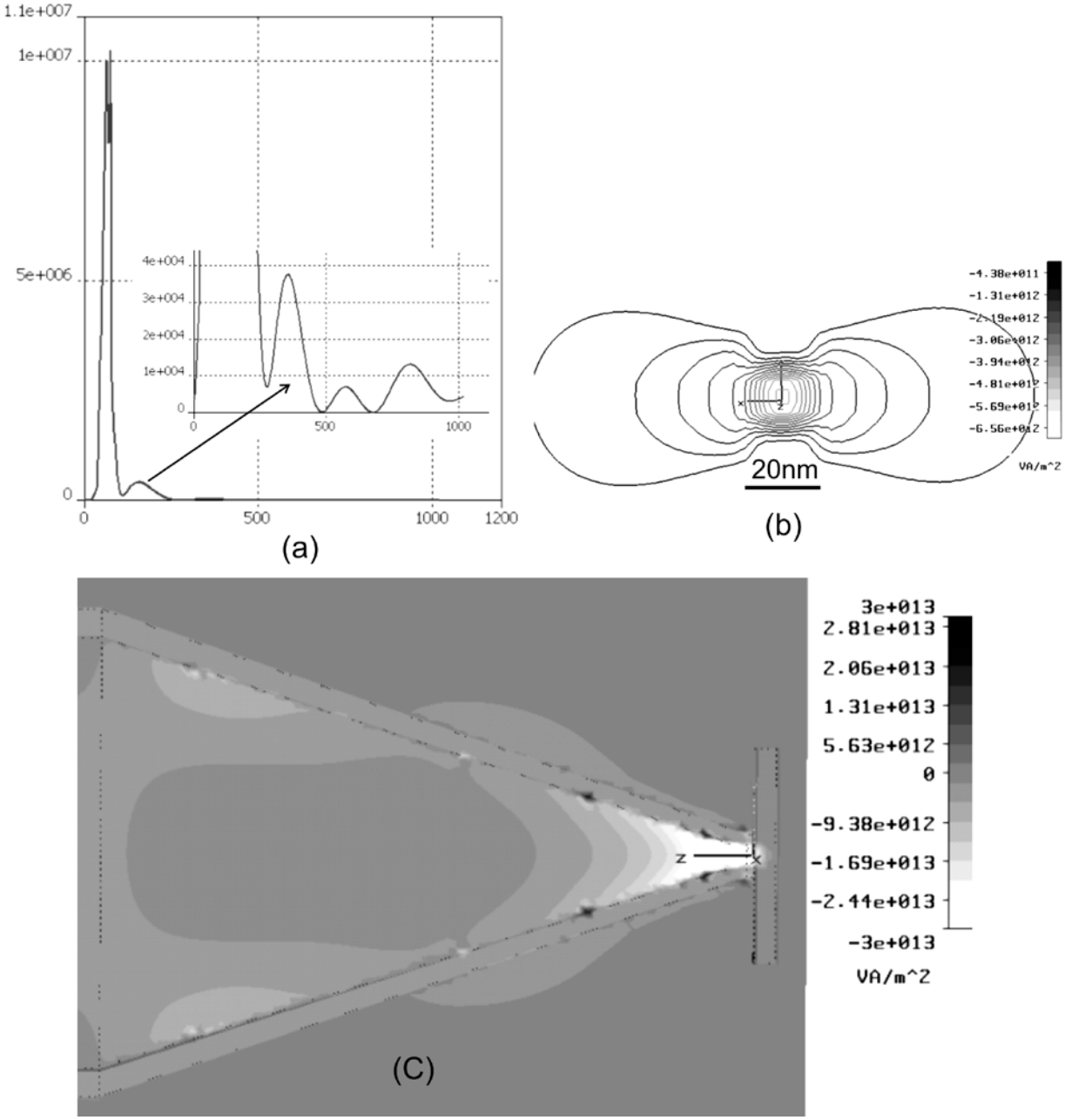


Fig. 25. Parameters of the PTMP simulated by MWS: (a) Electric field intensity along the probe axes (x=y=0); (b) Power flow P_z at the probe aperture; (c) Power flow P_z along the probe in the YZ plane). The calculation is for the following parameters: a_1=b_1=600, s=100, a=b=20nm, t=20, t_1=-40, λ=780 nm, and ε=2.25.

Fig. 25 shows the field distribution in the PTMP with an aperture of, a=b=20nm, t=20nm. As one can see from Fig. 25 (a), the field distribution along the probe has a structure characteristic of a static wave with a high, sharp peak near the probe apex, where intensity of the field is enhanced by a factor of approximately 2500.

This result is in good agreement with a simple formula $4k_t(a/a_1)^2$ derived from a simplified model of the PTMT [28, 39-40]; here k_t is the energy loss coefficient (i.e., the ratio of the beam power that reached the apex to the power of an incident wave), and definition of a and a_1 is clear from Fig. 20(a). It should be noted that due to energy losses, the field distribution

does not exactly correspond to that for the static wave (i.e., the field amplitude is not equal to zero at the wave nodes).

The spatial resolution of the SNOM can be calculated as a size of the light spot at the probe aperture. The probe of a conventional SNOM, as is pointed out above, has a spot size approximately equal to the diameter of the hole. Since the PTMT is an open structure, there is no easy way to calculate a spatial resolution for this case.
From the fact that the tapered microstrip line has only one propagating mode (quasi-TM_{00} mode) in a small region near the probe apex (in the case of symmetrical excitation), one may suggest that the size of the light spot formed close to the probe apex can be estimated from the transverse dimensions characterizing spatial localization of the quasi-TM_{00} mode; these dimensions, as is estimated above, are equal to the transverse sizes (thickness and width) of the dielectric slab of the microstrip line [51].
However, an incident wave may excite strong local surface plasmon waves at the edges of the metal strips at the probe apex, and therefore, the spot size can increase at least by a factor of $2t_1$ (see Fig. 20) in the transverse plane (i.e., along the *y*-axis) due to a large field intensity of the local plasmon waves.
Fig. 26 shows spatial distribution of the main component of the electric field (E_y) and the electric field energy distribution at the probe aperture simulated by the MWS. One can see in Fig. 26(*a*) that the electric field amplitude is strongly inhomogeneous in the horizontal plane (i.e., in a transverse plane perpendicular to the structure symmetry axis), displaying sharp peaks at the dielectric-metal and the metal-air interfaces.
The amplitude of the E_y component of the field is small above the ends of the metal strips and is rapidly decreasing in the air with distance from the metal strips. In addition, E_y decreases monotonically along the *x*-axis with distance from the center of the dielectric slab. However, the spatial distribution of E_y does not give a true picture of the electric field intensity at the aperture, because it does not display a peak of the electric field intensity at the end of the metal strips, which is mainly due to the E_z component of the electric field.
Fig. 26(*c*) shows the electric field energy distribution (in a horizontal plane located 6 nm off the probe end), which displays two high peaks just below the metal strip edges. Therefore, the area of the light spot at the aperture may be calculated by an approximate formula:

$$S_{spot} = (a_1 + 2t_1) \cdot b_1 \tag{56}$$

where a_1 and b_1 are cross-sectional dimensions of the dielectric slab, and t_1 is the hickness of the metal strips at the apex (see Fig. 26).

It follows from Eq. (58) that in order to obtain high horizontal spatial resolution, both the dielectric and metal layers must be thin at the probe apex. Additional numerical simulation has shown that the Eq. (58) is not accurate for thick metal strips and the light spot size along the *y*-axis being less than $a_1 + 2t_1$.

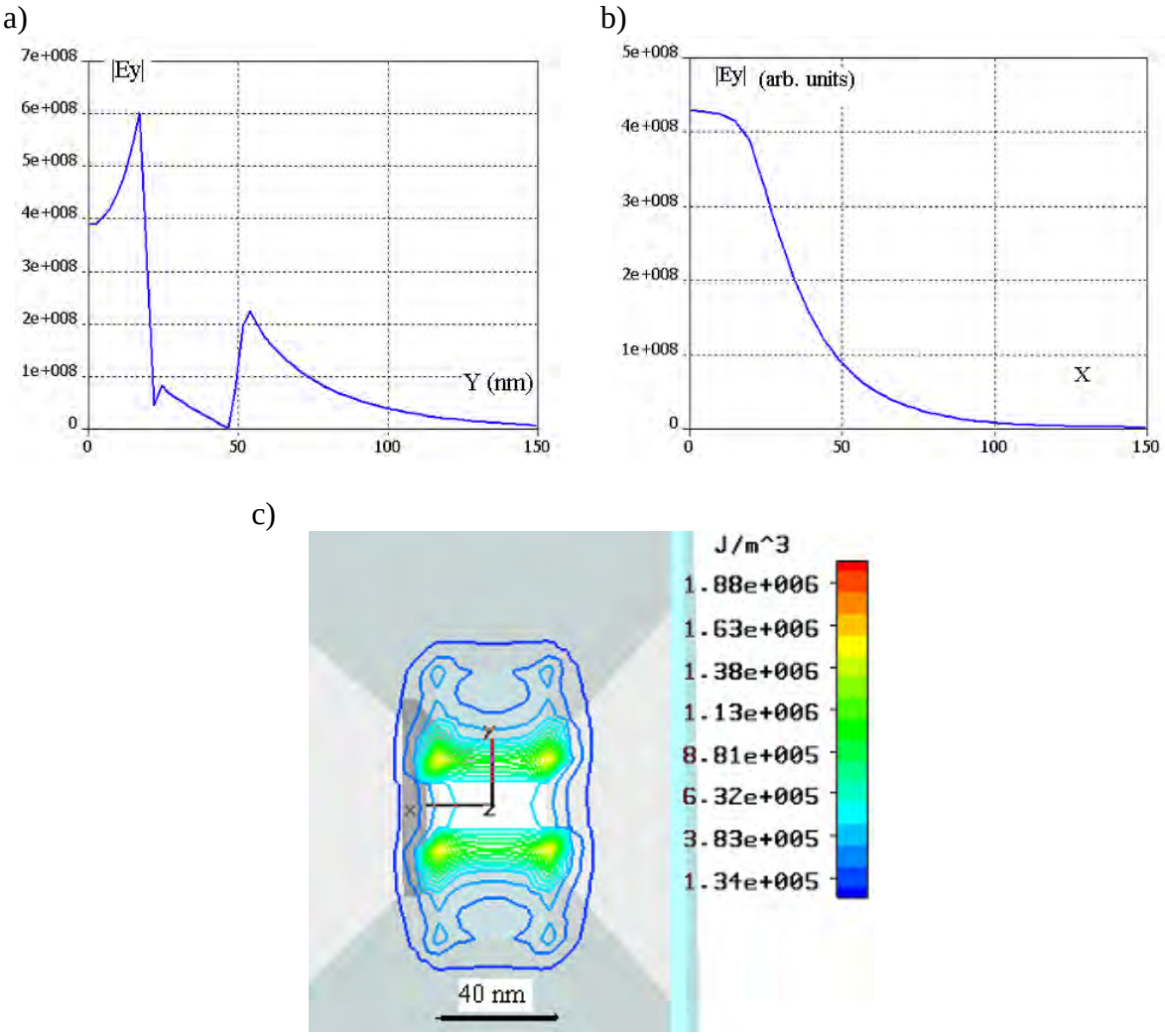


Fig. 26. Amplitude of the E_y at the microstrip probe apex: (a) along the y-axis (x=0); (b) along the x-axis (y=0). (c) Spatial distribution of the electric field energy in a horizontal plane located 6 nm off the microstrip probe apex. The following parameters were used in the calculation: $a = b$ = 400 nm; a_1=b_1=40 nm; t=70 nm; t_1=25 nm; dielectric constant of the glass core ε=2.25; gold strips; s=1000 nm; and λ=780 nm.

It follows from Eq. (56) that in order to obtain high horizontal spatial resolution, both the dielectric and metal layers must be thin at the probe apex. Additional numerical simulation has shown that the Eq. (56) is not accurate for thick metal strips and the light spot size along the y-axis being less than $a_1 + 2t_1$.

The working quasi-TM_{00} mode of an optical microstrip line has high coupling efficiency to the aperture of the PTMP and has no cut-off frequency. Hence, a high optical efficiency of the PTMP may be achieved assuming a high coupling efficiency of an incident (exciting) wave to the quasi-TM_{00} mode of the probe. A simple formula for the far-field transmission coefficient for the PTMP was obtained in Refs. [28, 39- 40] by using a simplified model:

$$k_f \approx 0.4 \frac{\rho_0}{k_l \sqrt{\varepsilon}} \left(\frac{a_1}{3\lambda} \right)^2, \qquad (57)$$

where a_1 is the height of the tapering microstrip line at the apex of the PTMP (see Fig. 20), ρ_0 and ε are the free-space impedance and effective dielectric constant of the micro strip line, respectively, and k_l is the attenuation coefficient of an incident wave.

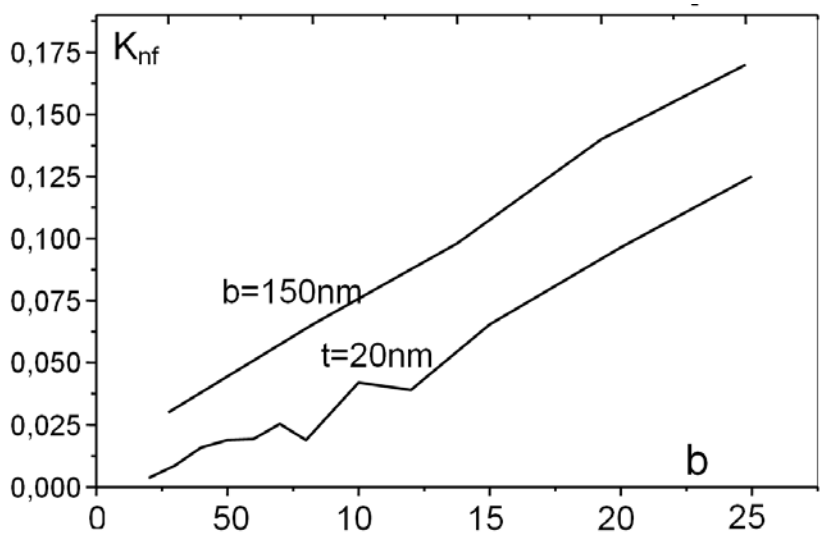


Fig. 27. Near-field transmission coefficient for the case of the PTMP with constant cross-sectional dimensions (PTMP as a portion of a regular microstrip line). The probe parameters used in the calculation are as follows: $a=b=a_1=b_1$; $t=t_1$; ε=2.25; silver strips; and λ=780 nm.

Fig. 27 demonstrates the far-field transmission coefficient calculated by the MWS for the case of the PTMP being a piece of a regular microstrip line (i. e., the PTMP with the constant cross-sectional dimensions). The far-field transmission coefficient was calculated as a power radiated to the free space from the end of an optical microstrip line excited by the quasi-TEM mode (with unit power) of an ideal microstrip line. The numerical results for the case of a small aperture size are close to those yielded by the simplified Eq. (57). However, with an increase of the transverse microstrip line sizes *a* and *b* (see Fig. 20), the transmission coefficient is not increasing as fast as predicted by Eq. (57). We suggest that the deviation from the results obtained by the simplified theory is due to transformation of the quasi-TM_{00} mode into a surface plasmon wave as the microstrip-line height *a* increases, as is mentioned above. With increasing the aperture size, the far-transmission coefficient should become close to the transmission coefficient of a single metal-strip plasmon wave at the edge of the strip end (which should be smaller than 1).

For the FIT (FDTD) method, the tapered metal film to some extent is similar to a corrugated metal layer with the surface roughness equal to the linear size of Yee cell. Therefore, the field near the tapered metal layer calculated by the FIT method is modulated due to this virtual corrugated metal surface. Power flow calculation requires knowledge of the electrical and magnetic fields, which in the FIT method are located in different vertices of Yee cell (some of them may be in the dielectric, while others in the metal environment). Therefore, one could expect that the FIT calculation algorithm may give not accurate result for the power flow in some points near to a metal surface. Fig. 28 shows the distribution of the power flow P_z along the PTMP to demonstrate inaccuracy in calculation of the power flow at a tapered metal surface by the MWS package. It can be seen from Fig. 28 that there are some points near the metal surface, where the power flow displays abnormal behavior (i. e.,

propagation of the power flow in reverse direction). Because of limited computational capabilities, in the numerical simulation presented in this paper, only several Yee cells are located between the two metal surfaces in the probe apex region. Therefore, due to somewhat inaccurate (in consequence of a small number of Yee cells used) power flow calculation for points close to metal surfaces, our direct calculation of the far-field transmission coefficient of the PTMP by the MWS package yields unstable results. Nevertheless, in the numerical simulations that do not display "reverse" points for the power flow in the aperture region, the derived far-transmission coefficients were close to those obtained for the case of radiation from the end of a regular microstrip line. We suggest therefore that the far-field transmission coefficient of the PTMP should be close to that numerically calculated for the case of a regular microstrip line.

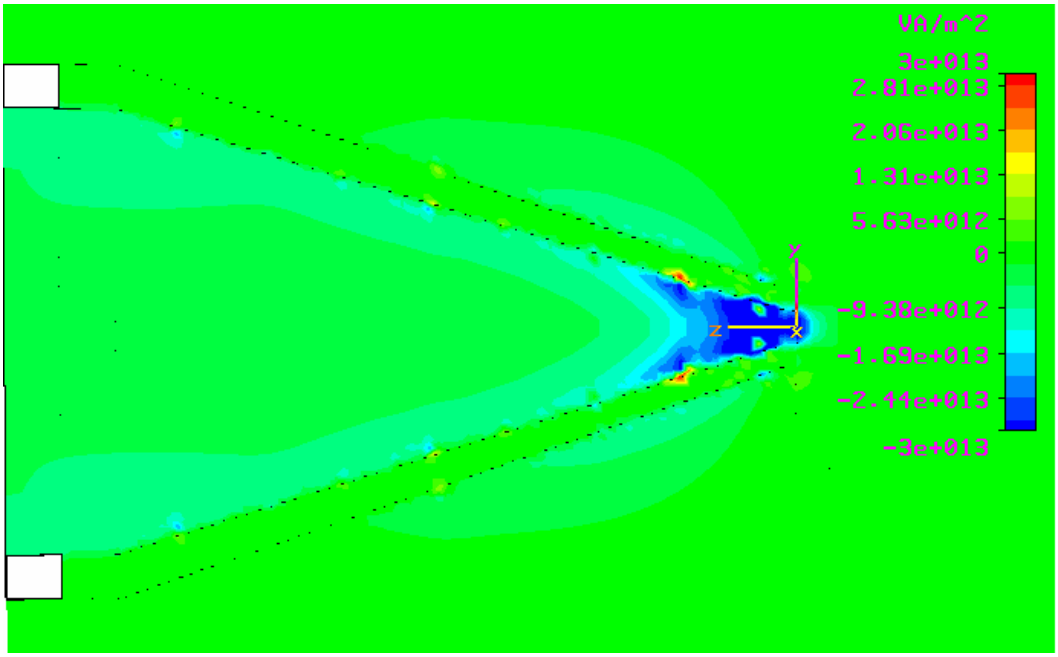


Fig. 28. Spatial distribution of P_z component of the power flow along z-axis in the PTMP. Sample parameters are as follows: a_1=b_1=50 nm; s=600 nm; a=b=400 nm; t=70 nm; t_1=25 nm; dielectric core is glass with ε=2.25; silver electrodes; λ=780 nm.

As is shown above, the PTMP has a high field enhancement, which should lead to a strong near-field interaction between the probe and a scanned sample. Hence, the probe should have a large near-field energy transmission coefficient for the case of scanning the sample with high losses. The lossy medium near the apex of the probe "forces" the MWS package to form fine Yee cells in the PTMP apex region, and therefore, the transmission coefficient is calculated for this case with acceptable accuracy.

Fig. 29 shows the total energy flow and power distribution for the case when the probe interacts with a thin GeSbTe plate, which is used for rewritable optical data storage. One can infer from Fig. 29 that a high near-field optical transmission can be obtained (24% for the crystalline and 20% for the amorphous states) for the probe with a very small aperture size of 40 nm. Also, it should be noted that the difference in energy transmission for different phase states of GeSbTe plate, obtained in the numerical simulations, is sufficiently large for reading applications (due to a high signal contrast), and, at the same time, is small enough to be used for re-recording the data (because of a small difference in amount of energy transmitted to the recording film in the crystalline and amorphous states).

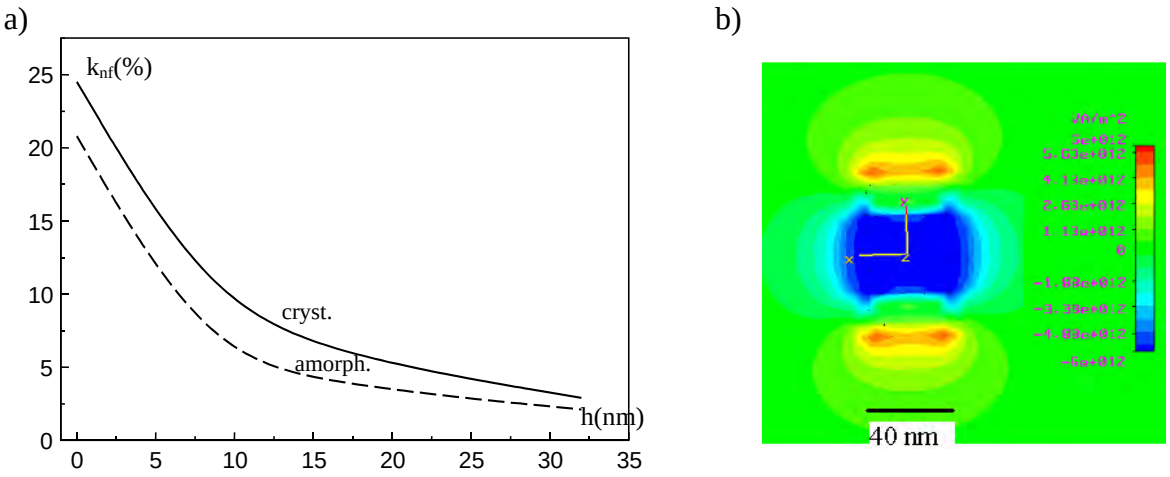


Fig. 29. (a) Dependence of the near-field transmission coefficient on the distance between the probe and a recording medium; (b) power distribution in the aperture plane for the case where the probe is interacting with 100 nm × 100 nm × 15 nm GeSbTe plate (n=4.68+4.16i for the crystalline state and n=4.34+1.75i for the amorphous state). Probe characteristics are as follows: a=b=400 nm; a_1=b_1=40 nm; t=70 nm; t_1=15 nm; h=1000 nm; ε=2.25; silver strips; and λ=780 nm.

The numerical simulation of the near-field interaction of the PTMT with a thin iron plate has shown that it has the near-field transmission coefficient as large as in the case of a thin GeSbTe film. Therefore, such a probe can be applied for heat-assisted magnetic data recording in a new generation of super-dense magnetic data storage devices.

1.12. Microstrip probe with metal tip

The PTMP produces a light spot size of which in the horizontal plane approximately equals to the total thickness of the metal-dielectric layered structure, and therefore, for obtaining a high spatial resolution, the probe should have a very thin and narrow all the three medium layers (one dielectric and two metals). However, using this approach only, it would be extremely difficult to obtain a spatial resolution higher than 40 nm for the PTMP. Moreover, the light spot intensity in the area under the probe apex, as is shown above, is strongly inhomogeneous with the two peaks at the dielectric-metal and metal-air interfaces. One may ask a question as whether or not is it possible to improve the spatial resolution by modifying the probe design so that only one peak appears in the light spot. One way to realize such a possibility is described below.

It is well known that a characteristic diameter of the spatial localization of the TM_0 mode for a circular metallic cylinder waveguide tends to zero when the cylinder diameter tends to zero [52-54]. At the same time, the cylinder HE_1 mode (that also has no cut-off frequency) is twice degenerated, and the characteristic diameter of the light spot associated with this mode increases with decreasing the cylinder diameter [52-54]. On the other hand, it was shown by a numerical simulation [55] that four modes can propagate in a finite-width metal film waveguide, denoted as ss_b^0, sa_b^0, as_b^0, and aa_b^0 in accordance with the symmetry of their electric field structure. Unfortunately, the waveguide spectrum was not investigated in [55] for the case when both the thickness and width of the metal film tend to zero simultaneously. However, in the case of small cross-sectional sizes of the rectangular waveguide (t<<λ; a<<λ, see Fig. 30 c), it follows from the quasistatic approach that the asymptotic behavior of the first mode of the waveguide should be similar to that for a circular

metal waveguide. Hence, the rectangular metal waveguide should also have two propagating modes with characteristics close to those of TM_0 (quasi-TM_0) and HE_1 (quasi-HE_1) modes of the circular metallic cylinder. Fig. 30 shows the electric field structure of the quasi-TM_0 (aa_b^0) and HE_1 (sa_b^0) modes, obtained in a numerical simulation using the MWS for the case of a metal strip situated between two dielectric slabs. Our numerical simulation has confirmed that: (i) the quasi-TM_0 mode of a very small rectangular waveguide is similar to the TM_0 mode of a circular metallic rod; and (ii) the quasi-HE_1 mode spreads over a large area in the cross-sectional plane and has a low (close to 1) slow-wave factor (see Fig. 30(b).

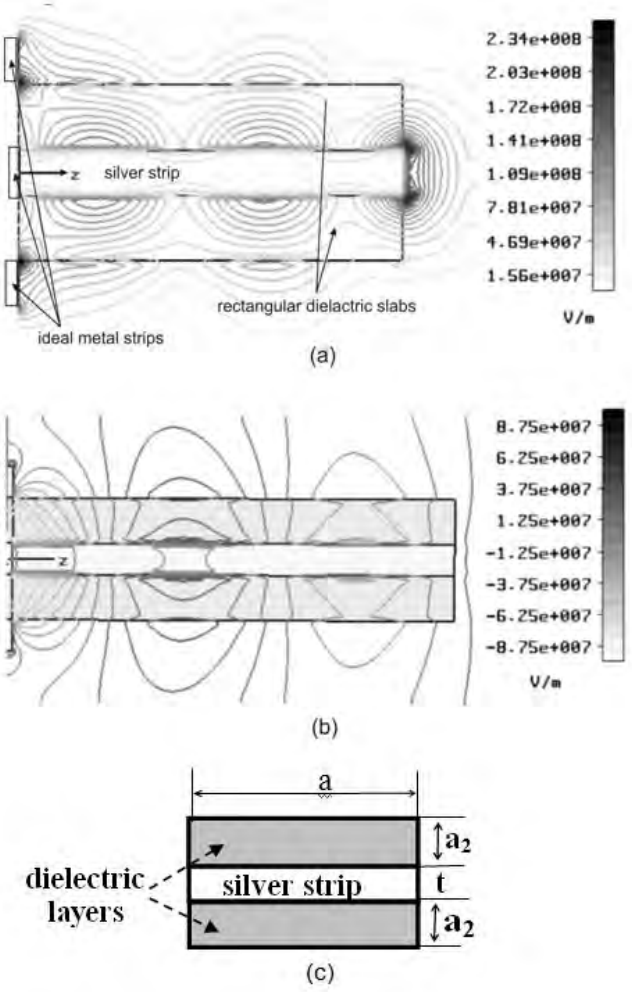


Fig. 30. The electric field amplitude distribution along a rectangular silver strip line: (a) asymmetric quasi-TM_0 mode, and (b) symmetric quasi-HE_1 mode. (c) Cross-sectional view of the structure. The following parameters were used in calculations: b= 100nm; t=70nm, s (structure length) = 1000 nm, ε = 2.25, and λ = 780 nm. Ideal metal strips are used only for mode excitataion.

In a first approximation, as is shown above, the quasi-TM_{00} microstripe-line mode can be represented as a symmetric sum of two plasmon waves propagating along the two metal-dielectric interfaces. Therefore, one may expect that trimming a portion of one of the metal strips of the PTMP at some distance from the aperture should transform the microstrip TM_{00} mode into the quasi-TM_{00} (aa_b^0) surface plasmon mode propagating along the other metal strip with optical efficiency approximately equal to 50%. As a result, in the case of very small cross-sectional dimensions of the metal strip edge, one may obtain very small spot size, strong field enhancement due to edge singularity [56] and a high spatial resolution. The structure of this new probe with only one metal tip is schematically shown in Fig. 31.

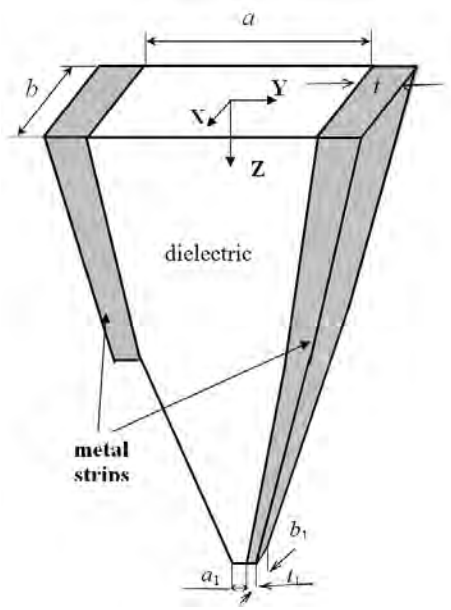


Fig. 31. Pyramid-type microstrip probe with only one metal tip.

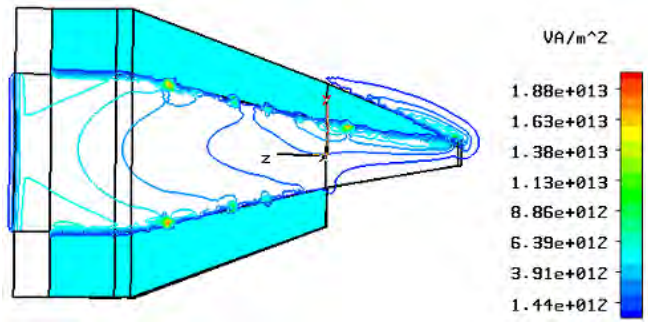


Fig. 32. Power flow (P_z component) in the pyramid-type microstrip probe with only one metal tip. The probe consists of a dielectric core with n = 1.5, covered with gold strips with the cross-sectional dimension of the tip edge equal to 5 nm; λ=780 nm.

Fig. 32 shows the power flow distribution along the probe (P_z component) and demonstrates the transformation of the quasi-TM_0 mode of the microstrip line to the $aa_b{}^0$-mode of a metal rectangular strip. In addition, Fig. 33 illustrates the beam size and electric-field energy distribution for the case of tip edge size of 5 nm, obtained using the MWS simulation. The resultant spot size, as one can infer from Fig. 33, is approximately equal to 15 nm (at the field intensity of the order of $1/e^2$), and the probe has field intensity enhancement well above 2000. The field intensity presented in Fig. 33 (b) is calculated for a direction along the symmetry axis of the dielectric pyramid.

a) b)

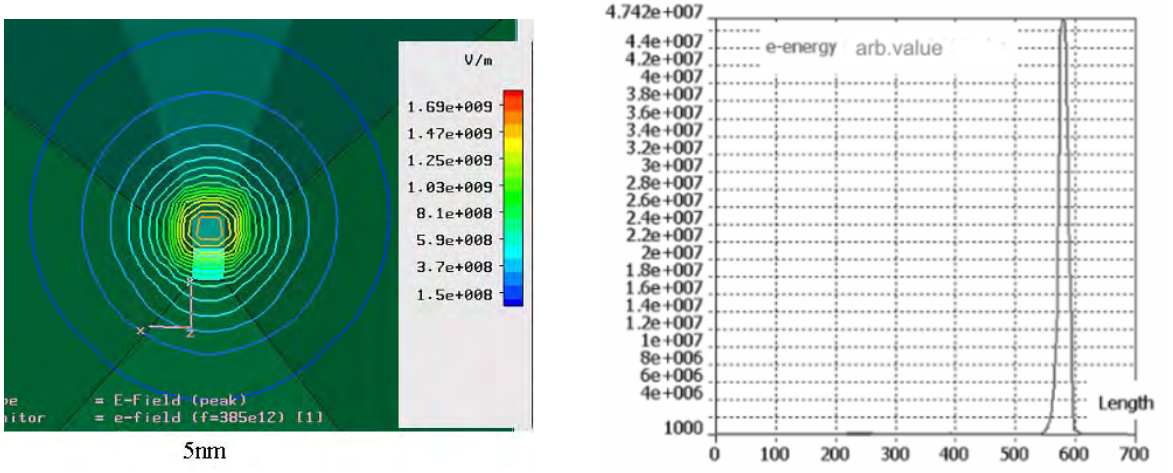


Fig. 33. (a) Distribution of the amplitude of the electric field in the plane that coincides with the surface of the edge of the metal tip; (b) Density of the electric field energy along the centerline of the dielectric core. The data are for a probe with gold strips, n=1.5 for the dielectric core, and the cross-sectional dimension of the tip edge equal to 5 nm; λ=780 nm.

The PTMP with a single metal tip has a strong field enhancement, and therefore, it may promote a strong near-field interaction with the scanned sample, yielding a high near-field transmission coefficient if the sample is characterized by great losses. Fig. 34 shows the dependence of the near-field energy transmission on a distance between the tip edge and a thin $Ge_2Sb_2Te_5$ monocrystalline film. It follows from Fig. 34 that the modified probe has approximately a near-field transmission coefficient smaller by a factor of 2 than that for an ordinary PTMP (which is expected due to 50% losses in a junction connecting the microstrip line with the metal-strip waveguide), but higher spatial resolution (approximately 40 nm).

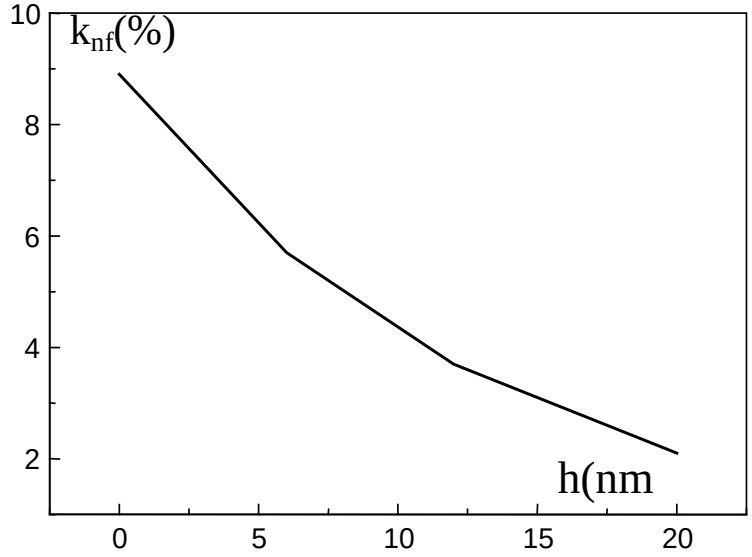


Fig. 34. Dependence of the near-field transmission coefficient on the distance to a recording media for the case of crystalline GeSbTe layer (n=4.68+4.16i). The probe has a gold tip with the edge size 15 × 15 nm; the beam size equals to 40nm.

1.13. Conclusion

The numerical simulation has demonstrated that the modified PTMP exhibits a high spatial resolution down to 40 nm. Our simulation based on rigorous 3D FIT method, yields a high far-field transmission coefficient and field enhancement for the PTMP. A high near-field energy transmission is obtained for the case of scanning the samples with great losses. The field in the probe is excited by a fundamental mode of a microstrip line with perfect metal strips, which cannot be realized in the experiment. In order to obtain better comparison between the numerical results and the experimental data, the excitation by focused beam or fiber beam should be simulated. More accurate numerical method should to be used for obtaining reliable data on the far-field transmission coefficient. The spatial resolution can be increased down to several nanometers by modifying the PTMP to have a single metal tip. Our numerical simulation using the MWS studio has shown that the PTMP with a single metal tip has high optical efficiency and high field enhancement.

Owing to improved characteristics mentioned above, both the ordinary PTMP and the PTMP with a single metal tip may be practically used in applications such as optical and magnetic data storage, nanolithography, and other types of nanotechnology where light is utilized for modification of a thin surface layer.

References

1. Bouwhuis G., Braat J., Haijser A., J. Pasman, van Rosmalen G., Immink K. S.: Principles of Optical Disc Systems, Bristol: Adam Hilger, 1985.
2. Marchant A.B.: Optical recording. A Technical Overview. Massachusetts: Addison-Wesley Company, 1990.
3. Petrov V.V., Kryuchyn A.A., Tokar A.P.: Optiko-mechanical storage devices., Kiev: Naukova Dumka, 1992.
4. Gore Ch., Salamoff P. J.: The Complete DVD Book: Designing, Producing, and Marketing Your Independent Film on DVD, California, USA: Michael Weise Productions, 2005.
5. Woerlee P., Koppers W., Martens H., Nijboer J., van den Oetelaar R., Spruit H., Weijenbergh P.: Format of an 8.5 GB double-layer DVD recordable disc, Proc. SPIE , Vol. 5380, 2004, pp. 15–20.
6. Katayama R., Komatsu Y.: Blue/DVD/CD compatible optical head, Appl. Opt., Vol. 47, No 22, 2008, pp. 4045-4054.
7. Neijzen J. H. M., Meinders E. R., van Santen H.: Liquid Immersion Deep-UV Optical Disc Mastering for Blu-ray Disc Read-Only Memory, Jpn. J. Appl. Phys., Vol. 43, No. 7B, 2004, pp. 5047-5052.
8. Brewer J., Gill M.: Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using Flash Memory Devices., Wiley-IEEE, 2007.
9. Moser A., Takano K., Margulies D.T., Albrecht M., Sonobe Y., Ikeda Y., Sun Sh., Fullerton E.E.: Magnetic recording: advancing into the future., J. Phys. D: Appl. Phys., Vol 35, 2002, R157–R167.
10. Hong J. H., McMichael I., Chang T. Y., Christian W., Paek E. G.: Volume holographic memory systems: techniques and architectures., Opt. Eng., Vol. 34. 1995, pp. 2193–2203.
11. Ayres M., Hoskins A., Curtis K.: Image oversampling for page-oriented optical data storage., Appl. Opt., Vol. 45, No. 11, 2006, pp. 2459-2464.
12. Kikukawa T., Inoue M., Mishima K., Ushida T.: Recording characteristics of 10-layers recodable optical disc and a prospect for over 500G-byte recording., Jpn. J. Appl. Phys., Vol. 49, 2010.
13. Shirashi J., Kobayashi S., Miyashita H., Hino H.: New signal Quality Evaluation Method for 33.4 GB/Layer BDs., ISOM2009 Tech. Dig., 2009, pp. 74-75.
14. Inoue M., Kosuda A., Mishima K., Ushida T., Kikukawa T.: 512Gb recording on 16-layer optical disc with Blu-Ray Disk based optics. Proc. SPIE, Vol. 7730, 2010, D-1-D6.
15. Lapchuk A.S., Kryuchin A. A., Klimenko V. A., Kolesnikov M.U., Petrov V.V.: Diffraction of Gaussian laser beam by three-dimensional grating of dielectric spheres. Proc. SPIE, Vol. 3055, 1996, pp. 160-169.
16. Shylo S. A., Lapchuk A. S., Song J. S., Kim K. S.: Optical Parameters of Light Beam in Multilayer Nano-Structures., J. of the Korean Physical Society, Vol. 47, Aug. 2005, pp. 18-23.
17. Glushko B.A., Levich E.B.: USA patent No 6071671, Fluorescent optical memory, 2006.
18. Wang M., Esener S.: USA patent No 7439009, Three-dimensional optical data storage in fluorescent dye-doped photopolymer, 2008.

19. Born M., Wolf E.: Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, 7-th edition, Cambridge University Press, 1999.
20. Somalingam S., Dressbach K., Hain M., Stankovic S., Tschudi Th., Knittel J., Richter H.: Effective Spherical Aberration Compensation by Use of a Nematic Liquid-Crystal Device., Appl. Opt. , Vol 43, No. 13, 2004, pp. 2722-2729.
21. Knittel J., Richter H., Hain M., Somalingam S., Tschudi T.: Liquid crystal lens for spherical aberration compensation in a Blu-ray disc system., IEE Proceedings, Vol. 152, No. 1, 2005, pp. 15 – 18.
22. Courjon D.: Near-field microscope and near-field optics., Imperial College Press, 2003.
23. Mironov V.L.: Fundamentals of scanning probe microscopy., Moscow: Russian Natioal Academy of Scinces, 2004.
24. Bethe H.A.: Theory of diffraction by small holes., Phys. Rev., Vol. 66, 1944, pp. 163-170.
25. Bouwkamp C.J.: On the diffraction of electromagnetic waves by small circular disks and holes. Philips Res. Rep., Vol. 5, 1950, pp. 401-408.
26. Roberts A.: Small-hole coupling of radiation into a near-field probe., J. Appl. Phys., Vol. 70, 1991, pp. 4045-4050.
27. Pohl D.W., Novotny L., Hecht B., Heinzelmann H.: Radiation coupling and image formation in scanning near-field optical microscopy., Thin solid films, Vol. 273, 1996, pp. 161-167.
28. Lapchuk A.S., Kryuchin A.A.: Near-field optical microscope working on TEM wave, Ultramicroscopy, Vol. 99, No. 2-3, 2004, pp. 143-157.
29. Kaupp G.: Atomic Force Microscopy, Scanning Nearfield Optical Microscopy and Nanoscratching: Application to Rough and Natural Surfaces., Springer: Heidelberg-Verlag- London, 2006.
30. Betzig E., Trautman J. K., Wolfe R., Gyorgy E. M., Finn P. L., Kryder M. H., Chang C.-H.: Near-field magneto-optics and high-density data storage., Appl. Phys. Lett., Vol. 61, 1992, pp. 142–144.
31. Patan S., Arena A., Allegrini M., Andreozzi L., Faetti M., Giordano M.: Near-field optical writing on azo-polymethacrylate spin-coated films., Opt. Commun., Vol. 210, 2002, pp. 37–41.
32. Lee H. W., Kim Y. M., Jeon D. J., Kim E., Kim J., Park K.: Rewritable organic films for near-field recording., Optical Materials, Vol. 21, 2002, pp. 289–293.
33. Likodimos V., Labardi M., Pardi L., Allegrini M., Giordano M.: Optical nanowriting on azobenzene side-chain polymethacrylate thin films by near-field scanning optical microscopy, Appl. Phys. Lett., Vol. 82, 2003, pp. 3313-3315.
34. Fangl J.-Y., Tien Ch.-H., Shieh H.-P.: Dual-probe near-field fiber head with gap servo control for data storage applications., Opt. Express, Vol. 15, No 22, 2007, pp. 14619-14628.
35. Sendur K., Challener W., Peng Ch.: Ridge waveguide as a near field aperture for high density data storage, J. Appl. Phys., Vol. 96, No. 5, 2004, pp. 2743-2752.
36. Denk W., Pohl D. W.: Near-field optics: microscopy with nanometer-size fields., J. Vac. Sci. Technol., Vol. B 9, 1991, pp. 510–513.
37. Chen X.-W., Sandoghdar V., Agio M.: Highly efficient interfacing of guided plasmons and photons in nanowires., Nano Lett., Vol. 9, 2009, pp. 3756–3761.
38. Chen X.-W., Sandoghdar V., Agio M.: Nanofocusing radially-polarized beams for high-throughput funneling of optical energy to the near field., Opt. Express, Vol. 18, No. 10, 2010, pp. 10878-10887.

39. Lapchuk A.S.: Estimation of optical efficiency of a near-field optical microscope on the basis of a simplified mathematical model., J. Opt. A: Pure Appl. Opt., Vol. 3, No. 6, 2001, pp. 455-459.
40. Lapchuk A.S., Kryuchyn A.A., Shikhovets O.V.: Optical efficiency of near-field microscopes., Data recording, storage and processing, Vol. 3, No. 4, 2001, pp. 19-31.
41. Lapchuk A., Kryuchin A.: The theoretical investigation for improvement of scanning near-field optical microscope. Proc. SPIE, Vol. 4779, 2002, pp. 80-189.
42. Lapchuk A.S., Kryuchin A.A.: Patent of Ukraine No 58634, A probe of the near-field microscope for making polarization measurements., Bulletin No. 8, 2003.
43. Lapchuk A.S.: The theoretical investigation of characters of SNOM probe operating on TEM wave, Technical Digest of Optical Data Storage Conference, Vancouver, Canada, 2003, pp. 137-139.
44. Lapchuk A.S.: The theoretical investigation of characters of SNOM probe operating on TEM wave., Proc. SPIE, Vol. 5069, 2003, pp. 319-329.
45. Lapchuk A.S., Baek S.W., Jeong H.S., Kyong C.S.: The microstrip probe with golden strips in infrared optical waveband for optical and magnetic data storage., Technical Digest of ISOM, Nara, Japan, 2003, pp. 200-201.
46. Lapchuk A.S., Yun S.-K., Yurlov V., Song J.-H., An S., Nevirkovets I.: Numerical simulation of characteristics of near-field microstrip probe having pyramidal shape, JOSA A, Vol. 24, No. 8, 2007, pp. 2407-2417.
47. Lapchuk A.S., Choi M.-G.: USA paent No 7327665, Optical fiber probe using an electrical potential difference and an optical recorder using the same A., 2008.
48. Lapchuk A.S., Jeong H. S., Shin D. I.: USA patent No 7312445, Pyramid-shaped near field probe using surface plasmon wave, 2007.
49. Lapchuk A.S., Kryuchin A.A.: Numerical simulation of a piramidal-shape microstrip probe. Data recording, storage and processing, Vol. 10, No. 1, 2008, pp. 16-33.
50. Oh J., Kim Y.-J., Lapchuk A.S., Kyong Ch. S., Goto K.: Near-field optical microprobe array of waveguide mode for the optical data storage application, Intern. Symp. on Optical Memory and Optical Data Storage (ISOM/ODS), 2005, paper: WP21.
51. Lapchuk A.S., Shin D., Jeong H.-S., Kyong Ch. S., Shin D.-I.: Mode propagation in optical nanowaveguides with dielectric cores and surrounding metal layers, Appl. Opt., Vol. 44, No. 35, 2005, pp. 7522-7531.
52. Novotny L., Hafner C.: Light propagation in a cylindrical waveguide with a complex, metallic, dielectric function, Phys. Rev E, Vol. 50, No. 3, 1994, pp. 4094-3106.
53. Khosravi H., Tilley D. R., Loudon R.: Surface polaritons in cylindrical optical fibers, J. Opt. Soc. Am. A., Vol. 8, No. 1, 1991, pp. 112-122.
54. Aerst G.C., Boardman A.D., Paranjapet B.V.: Non radiative surface plasmon-polariton modes of inhomogeneous metal circular cylinders, J. Phys. F: Metal Phys.,Vol. 10, 1980, pp. 53-65.
55. Berini P.: Plasmon-polariton waves guided by thin lossy metal films of finite width: Bound modes of symmetric structures, Phys. Rev. B, Vol. 61, No. 15, 2000, pp. 10484-10503.
56. Lapchuk A.S., Shylo S.A., Nevirkovets I.P.: Local plasmon resonance at metal wedge, JOSA A, Vol. 25, No. 7, 2008, pp. 1535-1540.

2. Stabilność i stabilizacja dodatnich układów liniowych niecałkowitego rzędu za pomocą sprzężenia zwrotnego od wektora stanu

Tadeusz Kaczorek

Politechnika Białostocka
Wydział Elektryczny
Wiejska 45D, 15-351 Białystok
email: kaczorek@isep.pw.edu.pl

Streszczenie: Wprowadzono nowe pojęcie stabilności praktycznej i stabilności asymptotycznej układów dodatnich, stożkowych jednowymiarowych (1D) i dwuwymiarowych (2D) niecałkowitego rzędu. Podano warunki konieczne i wystarczające stabilności praktycznej i stabilności asymptotycznej tych układów dodatnich i stożkowych niecałkowitego rzędu. Wykazano, że badanie stabilności praktycznej i stabilności asymptotycznej układów dodatnich 2D można sprowadzić do badania stabilności odpowiednich układów dodatnich 1D. Podano trzy metody liniowych nierówności macierzowych (LMI) badania stabilności dodatnich układów liniowych. Zastosowano również metodę LMI do wyznaczania macierzy sprzężeń zwrotnych dla wektora stanu tak, aby układ zamknięty był dodatni i asymptotycznie stabilny. Efektywność tych metod została pokazana na przykładach numerycznych.

Słowa kluczowe: układ dodatni, 1D, 2D, stożkowy, stabilność praktyczna, stabilność asymptotyczna, stabilizacja, sprzężenie zwrotne, LMI.

2.1. Wprowadzenie

W układach dodatnich zmienne stanu, wymuszenia i odpowiedzi przyjmują tylko wartości nieujemne. Przykładami układów dodatnich są procesy w reaktorach chemicznych, wymiennikach ciepła, kolumnach destylacyjnych oraz modele zanieczyszczenia wody

i atmosfery, układy kompartmentalne itp. Układy dodatnie występują w technice, ekonomi, biologii medycynie i naukach społecznych.

Układy dodatnie są określone na stożkach. Dlatego teoria tych układów jest bardziej złożona i mniej rozwinięta. Aktualny stan rozwoju teorii układów dodatnich jest przedstawiony w monografii [3, 6]. Pojęcie układów stożkowych zostało wprowadzone w pracach [8, 19]

Podstawowymi modelami liniowych układów dwuwymiarowych (2D) są modele wprowadzone przez Roessera [34], Fornasiniego i Marchesiniego [4] oraz Kurka [26]. Modele te zostały uogólnione na układy dodatnie w pracach [6, 14, 20, 37]. Osiągalność i sterowanie z minimalną energią układów liniowych standardowych i dodatnich 2D były rozpatrywane w pracach [14, 18, 24, 25]. Pojęcie układów wewnętrznie dodatnich 2D z opóźnieniami w wektorze stanu i wymuszeniach zostało wprowadzone w pracach [14, 20, 25]. Podano w nich warunki konieczne i wystarczające wewnętrznej dodatniości, osiągalności, sterowalności, obserwowalności i sterowania z minimalną energią. Stabilność dodatnich układów liniowych 1D i 2D była badana w pracach [2, 5, 10, 15, 17, 37], a stabilność odporna w pracy [1].

Podstawy matematyczne układów niecałkowitego rzędu są podane w monografiach [27-30, 32, 33]. Dodatnie układy liniowe niecałkowitego rzędu były rozpatrywane w pracach [7, 11], a ich stabilność była badana w pracach [9, 10, 21, 22]. Metody liniowych nierówności macierzowych (LMI) do badania stabilności dodatnich liniowych układów 2D zostały zaproponowane w pracach [15, 36]. Dodatnie układy liniowe 2D niecałkowitego rzędu zostały wprowadzone w pracach [12, 13, 16]. Pojęcie praktycznej stabilności dodatnich dyskretnych układów liniowych 1D zostało wprowadzone w pracy [21]. Zastosowania układów niecałkowitego rzędu są podane w pracach [31, 32, 35, 38].

Praca ta jest poświęcona stabilności i stabilizacji dodatnich układów liniowych niecałkowitego rzędu ze sprzężeniem zwrotnym od wektora stanu. Układ tej pracy jest następujący.

W punkcie 2 podano podstawowe definicje i twierdzenia dotyczące stabilności dodatnich układów liniowych 1D niecałkowitego rzędu oraz wprowadzono pojęcie praktycznej i asymptotycznej stabilności układów niecałkowitego rzędu i układów stożkowych. Praktyczna i asymptotyczna stabilność dodatnich układów 2D niecałkowitego rzędu jest rozpatrywana w punkcie 3. Podano tu warunki konieczne i wystarczające stabilności oraz pokazano, że badanie stabilności dodatnich układów 2D można sprowadzić do badania stabilności odpowiednich układów dodatnich 1D.

W punkcie 4 zaproponowano zastosowanie metod LMI do badania stabilności dodatnich układów niecałkowitego rzędu oraz do wyznaczania macierzy wzmocnień sprzężenia zwrotnego od wektora stanu tak, aby układ zamknięty był dodatni i stabilny asymptotycznie. W punkcie 5 podano uwagi końcowe i możliwości uogólnienia tej pracy.

W pracy będą stosowane następujące oznaczenia. Zbiór macierzy rzeczywistych o wymiarach $n \times m$ i elementach nieujemnych będziemy oznaczać przez $\Re_+^{n \times m}$ oraz $\Re_+^n = \Re_+^{n \times 1}$. Macierz $A = [a_{ij}] \in \Re_+^{n \times m}$ (wektor x) o wszystkich elementach dodatnich $a_{ij} > 0$ dla $i = 1, \ldots, n, j = 1, \ldots, m$ (o wszystkich składowych dodatnich) będziemy oznaczać przez $A > 0$ ($x > 0$). Zbiór liczb całkowitych nieujemnych będziemy oznaczać przez Z_+ a macierz jednostkowa wymiaru $n \times n$ przez I_n.

2.2. Stabilność dodatnich układów liniowych 1D niecałkowitego rzedu

2.2.1. Dodatnie układy 1D

Weźmy pod uwagę dyskretny układ liniowy:

$$x_{i+1} = Ax_i + Bu_i \tag{1a}$$

$$y_i = Cx_i + Du_i \tag{1b}$$

gdzie, $x_i \in \Re^n$, $u_i \in \Re^m$, $y_i \in \Re^p$, $i \in Z_+$ są odpowiednio wektorem stanu, wymuszenia i odpowiedzi natomiast $A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$, $C \in \Re^{p\times n}$, $D \in \Re^{p\times m}$ są macierzami stanu.

Definicja 1. Układ (1) jest nazywany (wewnętrznie) dodatnim wtedy, gdy $x_i \in \Re_+^n$, $y_i \in \Re_+^p$, $i \in Z_+$ dla dowolnych warunków początkowych $x_0 \in \Re_+^n$ oraz wszystkich wymuszeń $u_i \in \Re_+^m$, $i \in Z_+$.

Twierdzenie 1. [3, 6] Układ (1) jest dodatni wtedy i tylko wtedy gdy

$$A \in \Re_+^{n\times n}, \quad B \in \Re_+^{n\times m}, \quad C \in \Re_+^{p\times n}, \quad D \in \Re_+^{p\times m} \tag{2}$$

Układ dodatni (1) jest nazywany stabilnym asymptotycznie jeżeli rozwiązanie

$$x_i = A^i x_0 \tag{3}$$

równania

$$x_{i+1} = Ax_i, \quad A \in \Re_+^{n\times n}, \quad i \in Z_+ \tag{4}$$

spełnia następujący warunek

$$\lim_{i\to\infty} x_i = 0 \quad \text{dla wszystkich } x_0 \in \Re_+^n \tag{5}$$

Twierdzenie 2. [3, 10] Dla układu dodatniego (4) następujące stwierdzenia są równoważne:

1) Układ jest stabilny asymptotycznie,
2) Wartości własne $z_1, z_2, \ldots, z_n$ macierzy A mają moduły mniejsze od jedności, tj. $|z_k| < 1$ dla $k = 1, \ldots, n$,
3) $\det[I_n z - A] \neq 0$ dla $|z| \geq 1$,
4) $\rho(A) < 1$, gdzie $\rho(A)$ jest promieniem spektralnym macierzy A definiowanym przez $\rho(A) = \max_{1\leq k\leq n} \{|z_k|\}$,
5) Wszystkie współczynniki $\hat{a}_i$, $i = 0,1,\ldots,n-1$ wielomianu charakterystycznego

$$p_{\hat{A}}(z) = \det[I_n z - \hat{A}] = z^n + \hat{a}_{n-1} z^{n-1} + \ldots + \hat{a}_1 z + \hat{a}_0 \tag{6}$$

macierzy $\hat{A} = A - I_n$ są dodatnie,

6) Wszystkie minory główne macierzy

$$\overline{A} = I_n - A = \begin{bmatrix} \overline{a}_{11} & \overline{a}_{12} & \cdots & \overline{a}_{1n} \\ \overline{a}_{21} & \overline{a}_{22} & \cdots & \overline{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{a}_{n1} & \overline{a}_{n2} & \cdots & \overline{a}_{nn} \end{bmatrix} \tag{7a}$$

są dodatnie, tj.,

$$\left|\overline{a}_{11}\right| > 0, \begin{vmatrix} \overline{a}_{11} & \overline{a}_{12} \\ \overline{a}_{21} & \overline{a}_{22} \end{vmatrix} > 0, ..., \det \overline{A} > 0 \tag{7b}$$

7) Istnieje ściśle dodatni wektor $\overline{x} > 0$ taki, że

$$\left[A - I_n\right]\overline{x} < 0 \tag{8}$$

Twierdzenie 3. [6] Układ dodatni (4) jest niestabilny jeżeli przynajmniej jeden element na głównej przekątnej macierzy *A* jest większy niż 1.

2.2.2. Dodatnie układy niecałkowitego rzędu

W pracy tej korzystać będziemy z następującej definicji różnicy niecałkowitego rzędu

$$\Delta^{\alpha} x_k = \sum_{j=0}^{k} (-1)^j \binom{\alpha}{j} x_{k-j}\ , \quad 0 < \alpha < 1 \tag{9}$$

gdzie $\alpha \in R$ jest rzędem różnicy niecałkowitej, oraz

$$\binom{\alpha}{j} = \begin{cases} 1 \quad \text{for} \quad j = 0 \\ \dfrac{\alpha(\alpha-1)\cdots(\alpha-j+1)}{j!} \quad \text{for} \quad j = 1, 2, ... \end{cases} \tag{10}$$

Weźmy pod uwagę układ dyskretny niecałkowitego rzędu opisany w przestrzeni stanu równaniami

$$\Delta^{\alpha} x_{k+1} = Ax_k + Bu_k \tag{11a}$$

$$y_k = Cx_k + Du_k \tag{11b}$$

gdzie, $x_k \in \Re^n$, $u_k \in \Re^m$, $y_k \in \Re^p$, $k \in Z_+$ są odpowiednio wektorem stanu, wymuszenia i odpowiedzi, $A \in \Re^{n \times n}$, $B \in \Re^{n \times m}$, $C \in \Re^{p \times n}$, $D \in \Re^{p \times m}$.

Korzystając z (9) możemy równania (11) napisać w postaci

$$x_{k+1} + \sum_{j=1}^{k+1} (-1)^j \binom{\alpha}{j} x_{k-j+1} = Ax_k + Bu_k, \ k \in Z_+ \tag{12a}$$

$$y_k = Cx_k + Du_k \tag{12b}$$

Definicja 2. Układ (12) jest nazywany (wewnętrznie) dodatnim układem niecałkowitego rzędu wtedy i tylko wtedy, gdy $x_k \in \Re_+^n$ oraz $y_k \in \Re_+^p$ $k \in Z_+$ dla dowolnych warunków brzegowych $x_0 \in \Re_+^n$ oraz dla wszystkich ciągów wymuszeń $u_k \in \Re_+^m$, $k \in Z_+$.

Twierdzenie 4. [7] Rozwiązanie równania (12a) dane jest zależnością

$$x_k = \Phi_k x_0 + \sum_{i=0}^{k-1} \Phi_{k-i-1} B u_i \tag{13}$$

gdzie Φ_k jest określone równaniem

$$\Phi_{k+1} = (A + I_n \alpha)\Phi_k + \sum_{i=2}^{k+1} (-1)^{i+1} \binom{\alpha}{i} \Phi_{k-i+1} \tag{14}$$

przy czym $\Phi_0 = I_n$.

Lemat 1. [7] Jeżeli

$$0 < \alpha \leq 1 \tag{15}$$

to

$$(-1)^{i+1} \binom{\alpha}{i} > 0 \quad \text{dla} \quad i = 1, 2, \ldots \tag{16}$$

Twierdzenie 5. [7] Niech $0 < \alpha < 1$. Układ niecałkowitego rzędu (12) jest dodatni wtedy i tylko wtedy, gdy

$$A + I_n \alpha \in \Re_+^{n \times n}, \quad B \in \Re_+^{n \times m}, \quad C \in \Re_+^{p \times n}, \quad D \in \Re_+^{p \times m} \tag{17}$$

2.2.3. Praktyczna stabilność układów niecałkowitego rzędu

Z zależności (10) i (16) wynika, że współczynniki

$$c_j = c_j(\alpha) = (-1)^j \binom{\alpha}{j+1}, \quad j = 1, 2, \ldots \tag{18}$$

szybko maleją ze wzrostem *j* oraz są one dodatnie dla $0 < \alpha < 1$. W praktyce zakłada się, że *j* jest ograniczone przez liczbę naturalną *h*.

W takim przypadku równanie (12a) przyjmuje postać

$$x_{k+1} = A_\alpha x_k + \sum_{j=1}^{h} c_j x_{k-j} + B u_k, \quad k \in Z_+ \tag{19}$$

gdzie

$$A_\alpha = A + I_n \alpha \tag{20}$$

Zauważyć należy, że równanie (19) opisuje dyskretny układ liniowy z *h* opóźnieniami w wektorze stanu.

Definicja 3. Dodatni układ niecałkowitego rzędu (12) jest nazywany stabilnym praktycznie wtedy i tylko wtedy, gdy układ (19) jest stabilny asymptotycznie.

Definiując nowy wektor stanu

$$\tilde{x}_k = \begin{bmatrix} x_k \\ x_{k-1} \\ \vdots \\ x_{k-h} \end{bmatrix} \tag{21}$$

możemy równania (19) i (12b) napisać w postaci

$$\tilde{x}_{k+1} = \tilde{A}\tilde{x}_k + \tilde{B}u_k, \quad k \in Z_+ \tag{22a}$$

$$y_k = \tilde{C}x_k + \tilde{D}u_k \tag{22b}$$

gdzie

$$\tilde{A} = \begin{bmatrix} A_\alpha & c_1 I_n & c_2 I_n & \dots & c_{h-1} I_n & c_h I_n \\ I_n & 0 & 0 & \dots & 0 & 0 \\ 0 & I_n & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & I_n & 0 \end{bmatrix} \in \Re_+^{\tilde{n}\times\tilde{n}}, \quad \tilde{B} = \begin{bmatrix} B \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \Re_+^{\tilde{n}\times m}$$

$$\tilde{C} = \begin{bmatrix} C & 0 & \dots & 0 \end{bmatrix} \in \Re_+^{p\times\tilde{n}}, \quad \tilde{D} = D \in \Re_+^{p\times m}, \quad \tilde{n} = (1+h)n \tag{22c}$$

Do sprawdzenia stabilności praktycznej dodatniego układu niecałkowitego rzędu (12) można więc wykorzystać twierdzenie 2.

Twierdzenie 6. Dodatni układ niecałkowitego rzędu (12) jest stabilny praktycznie wtedy i tylko wtedy, gdy spełniony jest jeden z równoważnych warunków:

1) Wartości własne $\tilde{z}_k, \quad k = 1,\dots,\tilde{n}$ macierzy $\tilde{A}$ mają moduły mniejsze od jedności, tj.

$$|\tilde{z}_k| < 1 \quad \text{dla} \quad k = 1,\dots,\tilde{n} \tag{23}$$

2) $\det[I_{\tilde{n}} z - \tilde{A}] \neq 0$ dla $|z| \geq 1$,

3) $\rho(\tilde{A}) < 1$, gdzie $\rho(\tilde{A})$ jest promieniem spektralnym macierzy $\tilde{A}$ definiowanym przez $\rho(\tilde{A}) = \max_{1\leq k\leq\tilde{n}}\{|\tilde{z}_k|\}$,

4) Wszystkie współczynniki $\tilde{a}_i, \quad i = 0,1,\dots,\tilde{n}-1$ wielomianu charakterystycznego

$$p_{\tilde{A}}(z) = \det[I_n(z+1) - \tilde{A}] = z^n + \tilde{a}_{n-1}z^{n-1} + \dots + \tilde{a}_1 z + \tilde{a}_0 \tag{24}$$

macierzy $[\tilde{A} - I_{\tilde{n}}]$ są dodatnie,

5) Wszystkie minory główne macierzy

$$[I_{\tilde{n}} - A] = \begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \dots & \tilde{a}_{1\tilde{n}} \\ \tilde{a}_{21} & \tilde{a}_{22} & \dots & \tilde{a}_{2\tilde{n}} \\ \dots & \dots & \dots & \dots \\ \tilde{a}_{\tilde{n}1} & \tilde{a}_{\tilde{n}2} & \dots & \tilde{a}_{\tilde{n}\tilde{n}} \end{bmatrix} \tag{25a}$$

są dodatnie, tj.

$$|\tilde{a}_{11}|>0,\quad \begin{vmatrix}\tilde{a}_{11} & \tilde{a}_{12}\\ \tilde{a}_{21} & \tilde{a}_{22}\end{vmatrix}>0\ ,...,\ \det[I_{\tilde{n}}-\tilde{A}]>0 \tag{25b}$$

6) Istnieją ściśle dodatnie wektory $\overline{x}_i \in \Re_+^n,\ i=0,1,...,h$ spełniające zależności

$$\overline{x}_0 < \overline{x}_1,\ \overline{x}_1 < \overline{x}_2,...,\overline{x}_{h-1} < \overline{x}_h \tag{26a}$$

takie, że

$$A_\alpha \overline{x}_0 + c_1\overline{x}_1 + ... + c_h\overline{x}_h < \overline{x}_0 \tag{26b}$$

Dowód. Pierwsze pięć warunków 1) - 5) wynika natychmiast z twierdzenia 2. Wykorzystując zależność (8) do macierzy $\tilde{A}$ otrzymamy

$$\begin{bmatrix} A_\alpha & c_1I_n & c_2I_n & ... & c_{h-1}I_n & c_hI_n\\ I_n & 0 & 0 & ... & 0 & 0\\ 0 & I_n & 0 & ... & 0 & 0\\ ... & ... & ... & ... & ... & ...\\ 0 & 0 & 0 & ... & I_n & 0 \end{bmatrix}\begin{bmatrix}\overline{x}_0\\ \overline{x}_1\\ \overline{x}_2\\ \vdots\\ \overline{x}_{h-1}\\ \overline{x}_h\end{bmatrix} < \begin{bmatrix}\overline{x}_0\\ \overline{x}_1\\ \overline{x}_2\\ \vdots\\ \overline{x}_h\end{bmatrix} \tag{27}$$

Z zależności (27) wynika warunek (26).

Twierdzenie 7. Jeżeli dodatni układ niecałkowitego rzędu (12) jest stabilny asymptotycznie, to wtedy suma elementów w każdym wierszu macierzy dołączonej $\mathrm{Adj}[I_{\tilde{n}}-\tilde{A}]$ jest ściśle dodatnia, tj.

$$\mathrm{Adj}\,[I_{\tilde{n}}-\tilde{A}]^{-1}\mathbf{1}_{\tilde{n}}>0 \tag{28}$$

gdzie $\mathbf{1}_{\tilde{n}}=[1\ \ 1\ \ ...\ \ 1]^T \in \Re_+^{\tilde{n}}$, T oznacza transpozycje.

Dowód. Jak wiadomo [9, 14], jeżeli układ (22) jest stabilny asymptotycznie to wtedy wektor

$$\overline{x}=[I_{\tilde{n}}-\tilde{A}]^{-1}\mathbf{1}_{\tilde{n}} \tag{29}$$

jest ściśle dodatnim ($\overline{x}>0$) punktem równowagi dla $\tilde{B}u=\mathbf{1}_{\tilde{n}}$. Zauważmy, że

$$\det[I_{\tilde{n}}-\tilde{A}]>0 \tag{30}$$

Ponieważ wszystkie wartości własne macierzy $[I_{\tilde{n}}-\tilde{A}]$ są dodatnie. Warunki (29) i (30) implikują więc (28). □

Przykład 1. Sprawdzić stabilność praktyczną dodatniego układu niecałkowitego rzędu

$$\Delta^\alpha x_{k+1}=0.1x_k,\quad k\in Z_+ \tag{31}$$

dla $\alpha=0.5$ oraz $h=2$.

Korzystając z (18), (20) i (22c) otrzymamy

$$c_1 = -\frac{\alpha(\alpha-1)}{2} = \frac{1}{8}, \quad c_2 = \frac{1}{16}, \quad A_\alpha = 0.6$$

oraz

$$\tilde{A} = \begin{bmatrix} A_\alpha & c_1 & c_2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & \frac{1}{8} & \frac{1}{16} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

W tym przypadku wielomian charakterystyczny (24) przyjmuje postać

$$p_{\tilde{A}}(z) = \det[I_{\tilde{n}}(z+1) - \tilde{A}] = \begin{vmatrix} z+0.4 & -\frac{1}{8} & -\frac{1}{16} \\ -1 & z+1 & 0 \\ 0 & -1 & z+1 \end{vmatrix} = z^3 + 2.4z^2 + 1.675z + 0.2125 \tag{32}$$

Wszystkie współczynniki wielomianu (32) są dodatnie, z twierdzenia 6 wynika więc, że układ (31) jest stabilny praktycznie.

Korzystając z (28) otrzymamy

$$\mathrm{Adj}[I_{\tilde{n}} - \tilde{A}]\mathbf{1}_{\tilde{n}} = \left(\mathrm{Adj} \begin{bmatrix} 0.4 & -\frac{1}{8} & -\frac{1}{16} \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.2500 \\ 1.4625 \\ 1.6750 \end{bmatrix}$$

Warunek (28) jest więc spełniony.

Twierdzenie 8. Dodatni układ niecałkowitego rzędu (12) jest stabilny praktycznie tylko wtedy, gdy układ dodatni

$$x_{k+1} = A_\alpha x_k, \quad k \in Z_+ \tag{33}$$

jest stabilny asymptotycznie.

Dowód. Z zależności (26b) mamy

$$(A_\alpha - I_n)\bar{x}_0 + c_1\bar{x}_1 + ... + c_h\bar{x}_h < 0 \tag{34}$$

Zauważmy, że nierówność (34) może być spełniona tylko, wtedy gdy istnieje ściśle dodatni wektor $\bar{x}_0 \in \Re_+^n$ taki, że

$$(A_\alpha - I_n)\bar{x}_0 < 0 \tag{35}$$

ponieważ $c_1\bar{x}_1 + ... + c_h\bar{x}_h > 0$.

Z twierdzenia 2 wynika, że warunek (35) implikuje stabilność asymptotyczną układu dodatniego (33). □

Z twierdzenia 8 wynika następujący ważny wniosek.

Wniosek 1. Dodatni układ niecałkowitego rzędu (12) jest niestabilny praktycznie dla dowolnej skończonej liczby *h*, jeżeli układ dodatni (33) jest niestabilny.

Twierdzenie 9. Dodatni układ niecałkowitego rzędu (12) jest niestabilny praktycznie jeżeli przynajmniej jeden element na głównej przekątnej macierzy A_α jest większy od jedności.

Dowód. Dowód wynika natychmiast z twierdzeń 8 i 3. □

Przykład 2. Weźmy pod uwagę autonomiczny dodatni układ niecałkowitego rzędu opisany równaniem

$$\Delta^\alpha x_{k+1} = \begin{bmatrix} -0.5 & 1 \\ 2 & 0.5 \end{bmatrix} x_k, \quad k \in Z_+ \tag{36}$$

dla $\alpha = 0.8$ i dowolnej liczby *h*.
W tym przypadku $n = 2$ oraz

$$A_\alpha = A + I_n \alpha = \begin{bmatrix} 0.3 & 1 \\ 2 & 1.3 \end{bmatrix} \tag{37}$$

Z twierdzenia 9 wynika, że dodatni układ niecałkowitego rzędu jest niestabilny praktycznie dla dowolnej liczby *h* ponieważ element (2,2) macierzy (37) jest większy od 1.
Ten sam wynik otrzymamy z warunku 5) twierdzenia 2, ponieważ wielomian charakterystyczny macierzy $A_\alpha - I_n$

$$p_{\tilde{A}}(z) = \det[I_{\tilde{n}}(z+1) - A_\alpha] = \begin{bmatrix} z+0.7 & -1 \\ -2 & z-0.3 \end{bmatrix} = z^2 + 0.4z - 2.21$$

ma jeden ujemny współczynnik ($\hat{a}_0 = -2.21$).

2.2.4. Asymptotyczna stabilność układów niecałkowitego rzędu

W tym podrozdziale będziemy rozpatrywać stabilność praktyczną układów dodatnich dla $h \to \infty$.

Definicja 4. Dodatni układ niecałkowitego rzędu (12) jest nazywany stabilnym asymptotycznie jeżeli układ jest stabilny praktycznie dla $h \to \infty$.

Lemat 2. Jeżeli $0 < \alpha < 1$ to $\quad \sum_{j=1}^{\infty} c_j = 1 - \alpha \tag{38}$

gdzie współczynniki c_j zdefiniowane są zależnością (18).

Dowód. Korzystając z rozwinięcia w szereg Maclaurina, łatwo wykazać, że $(1-z)^{\alpha} = \sum_{j=0}^{\infty}(-1)^j \binom{\alpha}{j} z^j$. Następnie podstawiając $z = 1$ otrzymujemy $\sum_{j=0}^{\infty}(-1)^j \binom{\alpha}{j} = 0$. Z tego równania oraz z (38) otrzymujemy

$$1-\alpha+\sum_{j=2}^{\infty}(-1)^j \binom{\alpha}{j} = 1-\alpha-\sum_{j=1}^{\infty}(-1)^j \binom{\alpha}{j+1} = 1-\alpha-\sum_{j=1}^{\infty} c_j = 0 \,.\ \square$$

Twierdzenie 10. Dodatni układ niecałkowitego rzędu (12) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy układ dodatni

$$x_{i+1} = (A + I_n)x_i \tag{39}$$

jest stabilny asymptotycznie.

Dowód. Jak wiadomo [2], układ dodatni (19) dla $h \to \infty$ jest stabilny asymptotycznie wtedy i tylko wtedy, gdy układ dodatni

$$x_{i+1} = (A_{\alpha} + \sum_{j=1}^{\infty} c_j I_n)x_i \tag{40}$$

jest stabilny asymptotycznie. Układy dodatnie (39) i (40) są równoważne ponieważ z (38) i (20) mamy

$$A_{\alpha} + \sum_{j=1}^{\infty} c_j I_n = A + I_n\alpha + I_n(1-\alpha) = A + I_n \,.\ \square$$

Stosując twierdzenie 6 do układu dodatniego (39) otrzymamy następujące twierdzenie.

Twierdzenie 11. Dodatni układ niecałkowitego rzędu (12) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy spełniony jest jeden z równoważnych warunków:

1) Wartości własne $z_1, z_2, ..., z_k$ macierzy $A + I_n$ mają moduł mniejszy od jedności, tj. $|z_k| < 1$ dla $k = 1,...,n$,
2) Wszystkie współczynniki wielomianu charakterystycznego macierzy A są dodatnie,
3) Wszystkie minory główne macierzy $-A$ są dodatnie.

Twierdzenie 12. Dodatni układ niecałkowitego rzędu (12) jest niestabilny jeżeli, przynajmniej jeden element na głównej przekątnej macierzy A jest dodatni.

Dowód. Jeżeli przynajmniej jeden element na głównej przekątnej macierzy A jest dodatni, wtedy przynajmniej jeden element na głównej przekątnej macierzy $A + I_n$ jest większy od jedności, a jak wiadomo [6, 10], taki układ jest niestabilny. □

Przykład 3. Korzystając z twierdzenia 11 znaleźć wartości współczynnika c, dla których dodatni układ niecałkowitego rzędu (12) dla

$$A = \begin{bmatrix} -0.5 & 1 \\ 0.2 & c \end{bmatrix} \quad \text{i} \quad \alpha = 0.8 \tag{41}$$

jest stabilny asymptotycznie.
Układ niecałkowitego rzędu jest dodatni jeżeli wszystkie elementy macierzy

$$A_\alpha = A + I_n \alpha = \begin{bmatrix} 0.3 & 1 \\ 0.2 & c + \alpha \end{bmatrix} \tag{42}$$

są nieujemne, tj. $c + \alpha \geq 0$ oraz $c \geq -\alpha = -0.8$.
Korzystając z warunku 2) twierdzenia 11 dla macierzy (41) otrzymamy

$$\det[I_n z - A] = \begin{vmatrix} z + 0.5 & -1 \\ -0.2 & z - c \end{vmatrix} = z^2 + (0.5 - c)z - (0.5c + 0.2)$$

oraz $c < -0.4$. Układ niecałkowitego rzędu (12) z (41) jest więc dodatni i stabilny asymptotycznie dla $-0.8 \leq c < -0.4$. Ten sam wynik otrzymamy stosując warunek 3) twierdzenia 11.

2.2.5. Układy stożkowe niecałkowitego rzędu

Definicja 5. [8, 19] Niech $P = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \in \Re^{n \times n}$ będzie nieosobliwa oraz p_k będzie k-tym ($k = 1, ..., n$) wierszem macierzy P. Zbiór

$$\mathcal{P} := \left\{ x \in \Re^n : \bigcap_{k=1}^{n} p_k x \geq 0 \right\} \tag{43}$$

jest nazywany liniowym stożkiem generowanym przez macierz P.
W podobny sposób można zdefiniować liniowy stożek dla wymuszeń u

$$\mathcal{Q} := \left\{ u \in \Re^m : \bigcap_{k=1}^{m} q_k u \geq 0 \right\} \tag{44}$$

generowany przez nieosobliwą macierz $Q = \begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix} \in \Re^{m \times m}$ oraz dla wyjść y, liniowy stożek

$$\mathcal{V} := \left\{ y \in \Re^p : \bigcap_{k=1}^{p} v_k y \geq 0 \right\} \tag{45}$$

generowany przez nieosobliwą macierz $V = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} \in \Re^{p \times p}$.

Definicja 6. Układ niecałkowitego rzędu (12) nazywamy $(\mathcal{P}, \mathcal{Q}, \mathcal{V})$ stożkowym układem niecałkowitego rzędu, jeżeli $x_i \in \mathcal{P}$ oraz $y_i \in \mathcal{V}$, $i \in Z_+$ dla wszystkich $x_0 \in \mathcal{P}$, $u_i \in \mathcal{Q}$, $i \in Z_+$.

Stożkowy $(\mathcal{P}, \mathcal{Q}, \mathcal{V})$ układ niecałkowitego rzędu (12) będziemy w skrócie nazywać stożkowym układem niecałkowitego rzędu.

Zauważyć należy, że jeżeli $\mathcal{P} = \Re_+^n$, $\mathcal{Q} = \Re_+^m$, $\mathcal{V} = \Re_+^n$ wtedy $(\Re_+^n, \Re_+^m, \Re_+^p)$ układ stożkowy jest równoważy klasycznemu układowi dodatniemu [8, 19].

Twierdzenie 13. Układ niecałkowitego rzędu (12) jest $(\mathcal{P}, \mathcal{Q}, \mathcal{V})$ stożkowym układem niecałkowitego rzędu wtedy i tylko wtedy, gdy

$$\overline{A} = PAP^{-1} \in \Re_+^{n \times n}, \ \overline{B} = PBQ^{-1} \in \Re_+^{n \times m}, \ \overline{C} = VCP^{-1} \in \Re_+^{p \times n}, \ \overline{D} = VDQ^{-1} \in \Re_+^{p \times m} \tag{46}$$

Dowód. Niech

$$\overline{x}_i = Px_i, \ \ \overline{u}_i = Qu_i \ \text{ i } \ \overline{y}_i = Vy_i, \ i \in Z_+ \tag{47}$$

Z definicji 5 wynika, że jeżeli $x_i \in \mathcal{P}$ wtedy $\overline{x}_i \in \Re_+^n$, jeżeli $u_i \in \mathcal{Q}$ wtedy $\overline{u}_i \in \Re_+^m$ jeżeli $y_i \in \mathcal{V}$ wtedy $\overline{y}_i \in \Re_+^p$. Z (12) i (47) mamy

$$\begin{aligned} \overline{x}_{k+1} + \sum_{j=1}^{k+1} (-1)^j \binom{\alpha}{j} \overline{x}_{k-j+1} &= Px_{k+1} + \sum_{j=1}^{k+1} (-1)^j \binom{\alpha}{j} Px_{k-j+1} = PAx_k + PBu_k \\ &= PAP^{-1}\overline{x}_k + PBQ^{-1}\overline{u}_k = \overline{A}\overline{x}_k + \overline{B}\overline{u}_k, \ k \in Z_+ \end{aligned} \tag{48a}$$

oraz

$$\overline{y}_k = Vy_k = VCx_k + VDu_k = VCP^{-1}\overline{x}_k + VDQ^{-1}\overline{u}_k = \overline{C}\overline{x}_k + \overline{D}\overline{u}_k, \ k \in Z_+ \tag{48b}$$

Jak wiadomo [6], układ (48) jest dodatni wtedy i tylko wtedy, gdy warunki (46) są spełnione. □

Twierdzenie 14. Stożkowy układ niecałkowitego rzędu (12) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy dodatni układ niecałkowitego rzędu (48) jest stabilny asymptotycznie.

Dowód. Z zależności (46) mamy

$$\begin{aligned}\det[Iz-\overline{A}] &= \det[Iz-PAP^{-1}] = \det[P(Iz-A)P^{-1}] = \\ &= \det[Iz-A]\det P\det P^{-1} = \det[Iz-A]\end{aligned} \tag{49}$$

ponieważ $\det P \det P^{-1} = 1$.

Z twierdzenia 14 wynika następujący ważny wniosek.

Wniosek 2. Stożkowy układ niecałkowitego rzędu (12) jest stabilny praktycznie (stabilny asymptotycznie) wtedy i tylko wtedy, gdy dodatni układ niecałkowitego rzędu jest stabilny praktycznie (stabilny asymptotycznie).

W celu sprawdzenia stabilności praktycznej i stabilności asymptotycznej stożkowych układów niecałkowitego rzędu możemy korzystać z twierdzenia 2 oraz 6.

2.3. Stabilność dodatnich układów liniowych 2D niecałkowitego rzędu

2.3.1. Dodatnie układy liniowe 2D niecałkowitego rzędu

Definicja 7. [13] Różnica niecałkowitego rzędu (α,β) dwuwymiarowej funkcji dyskretnej x_{ij} nazywamy funkcję określoną zależnością

$$\Delta^{\alpha,\beta}x_{ij} = \sum_{k=0}^{i}\sum_{l=0}^{j} c_{\alpha\beta}(k,l)x_{i-k,j-l},\ \ n_1-1<\alpha<n_1,\quad n_2-1<\beta<n_2;\quad n_1,n_2\in N=\{1,2,...\} \tag{50a}$$

gdzie $\Delta^{\alpha,\beta}x_{ij} = \Delta_i^{\alpha}\Delta_j^{\beta}x_{ij}$ oraz

$$c_{\alpha,\beta}(k,l) = \begin{cases} 1 \quad \text{dla}\ \ k=0\ \ \text{i}\ \ l=0 \\ (-1)^{k+l}\dfrac{\alpha(\alpha-1)...(\alpha-k+1)\beta(\beta-1)...(\beta-l+1)}{k!l!} \quad \text{dla}\ \ k,l\geq 0\ \ \text{oraz}\ \ k+l>0 \end{cases} \tag{50b}$$

Weźmy pod uwagę dwuwymiarowy układ niecałkowitego rzędu (α,β) opisany równaniami

$$\Delta^{\alpha,\beta}x_{i+1,j+1} = A_0x_{ij} + A_1x_{i+1,j} + A_2x_{i,j+1} + B_0u_{ij} + B_1u_{i+1,j} + B_2u_{i,j+1} \tag{51a}$$

$$y_{ij} = Cx_{ij} + Du_{ij} \tag{51b}$$

gdzie $x_{ij} \in \Re^n$, $u_{ij} \in \Re^m$, $y_{ij} \in \Re^p$ są odpowiednio wektorem stanu, wymuszenia i odpowiedzi oraz $A_k \in \Re^{n\times n}$, $B_k \in \Re^{n\times m}$, $k = 0,1,2$, $C \in \Re^{p\times n}$, $D \in \Re^{p\times m}$.

Korzystając z definicji 7 równanie (51a) można napisać w postaci

$$x_{i+1,j+1} = \overline{A}_0 x_{ij} + \overline{A}_1 x_{i+1,j} + \overline{A}_2 x_{i,j+1} - \sum_{k,l\in D_{i+1,j+1}\setminus D_{11}} \sum c_{\alpha,\beta}(k,l) x_{i-k+1,j-l+1} + B_0 u_{ij} + B_1 u_{i+1,j} + B_2 u_{i,j+1} \tag{52}$$

gdzie $D_{pq} := \{(i,j): 0 \le i \le p, 0 \le j \le q, i, j \in Z_+\}$ i $\overline{A}_0 = A_0 - I_n\alpha\beta$, $\overline{A}_1 = A_1 + I_n\beta$, $\overline{A}_2 = A_2 + I_n\alpha$.

Warunki brzegowe dla (52) mają postać

$$x_{i0},\ i \in Z_+ \quad \text{i} \quad x_{0j},\ j \in Z_+ \tag{53}$$

Definicja 8. Układ (51) (oraz (52)) jest nazywany (wewnętrznie) dodatnim dwuwymiarowym układem niecałkowitego rzędu jeżeli $x_{ij} \in \Re_+^n$ oraz $y_{ij} \in \Re_+^p$, $i, j \in Z_+$ dla dowolnych warunków brzegowych $x_{i0} \in \Re_+^n, i \in Z_+$ $x_{0j} \in \Re_+^n$, $j \in Z_+$ i wszystkich ciągów wymuszeń $u_{ij} \in \Re_+^m$, $i, j \in Z_+$.

W [13] zostało wykazane, że:

a) Jeżeli $0 < \alpha < 1$ oraz $1 < \beta < 2$ wtedy

$$c_{\alpha,\beta}(k,l) < 0 \text{ dla } k = 1,2,...;\ l = 2,3,... \text{ oraz } c_{\alpha,\beta}(k,1) > 0,\ k = 1,2,...;\ c_{\alpha,\beta}(0,l) > 0,\ l = 2,3,... \tag{54a}$$

b) Jeżeli $1 < \alpha < 2$ oraz $0 < \beta < 1$ wtedy

$$c_{\alpha,\beta}(k,l) < 0 \text{ dla } k = 2,3,...;\ l = 1,2,... \text{ oraz } c_{\alpha,\beta}(k,0) > 0,\ k = 2,3,...;,\ c_{\alpha,\beta}(1,l) > 0,\ l = 1,2,... \tag{54b}$$

Twierdzenie 15. [13] Dwuwymiarowy układ niecałkowitego rzędu (51) dla $0 < \alpha < 1$ oraz $1 < \beta < 2$ (lub $1 < \alpha < 2$ oraz $0 < \beta < 1$) jest dodatni wtedy i tylko wtedy, gdy[1]

$$\overline{A}_k \in \Re_+^{n\times n},\ B_k \in \Re_+^{n\times m},\ k = 0,1,2;\ C \in \Re_+^{p\times n},\ D \in \Re_+^{p\times m} \tag{55}$$

---

[1] Zakładamy, że $\sum_{k=2}^{i+1} c_{k,1} x_{i-k+1,j} = 0$ oraz $\sum_{l=2}^{j+1} c_{0,l} x_{i+1,j-l+1} = 0$ gdyż $c_{k,1} > 0$, $k = 1,2,...$ oraz $c_{0,l} > 0$, $l = 2,3,...$

2.3.2. Stabilność praktyczna

Zauważmy, że układ (52) jest dwuwymiarowym układem liniowym z liczbą opóźnień w wektorze stanu zwiększającą się do nieskończoności dla $i, j \to \infty$.

Z (50b) wynika, że współczynniki

$$c_{k,l} = -c_{\alpha,\beta}(k,l) = (-1)^{k+l-1} \frac{\alpha(\alpha-1)...(\alpha-k+1)\beta(\beta-1)...(\beta-l+1)}{k!l!} \text{ dla } k,l \geq 0 \text{ oraz } k+l>0 \quad (56)$$

szybko maleją dla rosnących k oraz l. W przypadkach praktycznych zakłada się, że liczby k oraz l są ograniczone przez liczby naturalne L_1 oraz L_2. W takim przypadku równanie (52) dla $B_0 = B_1 = B_2 = 0$ przyjmuje postać

$$x_{i+1,j+1} = \overline{A}_0 x_{ij} + \overline{A}_1 x_{i+1,j} + \overline{A}_2 x_{i,j+1} + \sum_{k.l \in D_{i+1,j+1} \backslash D_{11}} \sum c_{k,l} x_{i-k+1,j-l+1} \quad (57)$$

gdzie $D_{pq} := \{(i,j): 0 \leq i \leq p, 0 \leq j \leq q, i,j \in Z_+\}$.

Równanie (57) opisuje dwuwymiarowy układ liniowy ze skończoną liczbą opóźnień w wektorze stanu. Układ (57) otrzymano poprzez opuszczenie wszystkich opóźnień dla $i > L_1$ oraz $j > L_2$ układu (52).

Definiując nowy wektor stanu

$$\begin{aligned} \tilde{x}_{ij} &= [x_{ij}^T \; x_{i-1j}^T \; ... \; x_{i-L_1 j}^T \; x_{ij-1}^T \; ... \; x_{i-L_1 j-1}^T \; x_{ij-2}^T \; ... \; x_{i-L_1 j-2}^T \; ... \; x_{i-L_1 j-L_2}^T] \in R^{\tilde{N}} \\ \tilde{N} &= (L_1+1)(L_2+1)n; \quad i,j \in Z_+ \end{aligned} \quad (58)$$

Równanie (57) można napisać w postaci

$$\tilde{x}_{i+1,j+1} = \tilde{A}_0 \tilde{x}_{ij} + \tilde{A}_1 \tilde{x}_{i+1,j} + \tilde{A}_2 \tilde{x}_{i,j+1} \quad i,j \in Z_+ \quad (59)$$

gdzie

$$\tilde{A}_0 = \begin{bmatrix} \overline{A}_0 & I_n c_{21} & ... & I_n c_{L_1,1} & I_n c_{L_1+1,1} & I_n c_{12} & ... & I_n c_{L_1,2} & I_n c_{L_1+1,2} & I_n c_{13} & ... & I_n c_{L_1,L_2+1} & I_n c_{L_1+1,L_2+1} \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \end{bmatrix}$$

$$\tilde{A}_1 = \begin{bmatrix} \overline{A}_1 & 0 & ... & 0 & 0 & I_n c_{02} & ... & 0 & 0 & I_n c_{03} & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ I_n & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & I_n & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & I_n & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \end{bmatrix} \tag{60}$$

$$\tilde{A}_2 = \begin{bmatrix} \overline{A}_2 & I_n c_{20} & ... & I_n c_{L_1,0} & I_n c_{L_1+1,0} & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ I_n & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & I_n & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & ... & 0 & 0 & 0 & ... & I_n & 0 \end{bmatrix}$$

Tym samym dwuwymiarowy układ niecałkowitego rzędu (52) został sprowadzony do standardowego układu dwuwymiarowego bez opóźnień ale o większych wymiarach.

Twierdzenie 16. Dwuwymiarowy układ (59) jest dodatni wtedy i tylko wtedy, gdy

$$\overline{A}_k \in \Re_+^{n \times n}, \ k = 0,1,2 \tag{61}$$

Dowód. Dowód wynika z (59), (60) oraz faktu, że układ jest układem dodatnim wtedy i tylko wtedy, gdy wszystkie jego macierze mają nieujemne elementy. □

Definicja 9. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) nazywamy praktycznie stabilnym jeżeli, układ opisany równaniem (57) jest stabilny asymptotycznie.

Twierdzenie 17. [10, 17] Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest praktycznie stabilnym wtedy i tylko wtedy, gdy jest spełniony jeden z poniższych warunków:

1) $\det(I_{\tilde{N}} - \tilde{A}_0 z_1 z_2 - \tilde{A}_1 z_2 - \tilde{A}_2 z_1) \neq 0 \ \ \forall (z_1, z_2) \in B := \{(z_1, z_2) : | z_1 | \leq 1, | z_2 | \leq 1\}$ (62)

2) Istnieje ściśle dodatni wektor $\lambda \in \Re_+^{\tilde{N}}$ taki, że

$$[\tilde{A}_0 + \tilde{A}_1 + \tilde{A}_2 - I_{\tilde{N}}]\lambda < 0 \tag{63}$$

3) Jednowymiarowy układ dodatni

$$x_{i+1} = (\tilde{A}_0 + \tilde{A}_1 + \tilde{A}_2)x_i, \quad i \in Z_+ \tag{64}$$

jest stabilny asymptotycznie,

4) Jednowymiarowy układ dodatni

$$x_{i+1} = \begin{bmatrix} \tilde{A}_1 + \tilde{A}_2 & \tilde{A}_0 \\ I_{\tilde{N}} & 0 \end{bmatrix} x_i \quad i \in Z_+ \tag{65}$$

jest stabilny asymptotycznie.

Twierdzenie 18. [10, 17] Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest praktycznie stabilny jeżeli, dwuwymiarowy układ dodatni

$$\tilde{x}_{i+1,j+1} = \tilde{A}_0 \tilde{x}_{ij} + \tilde{A}_1 \tilde{x}_{i+1,j} + \tilde{A}_2 \tilde{x}_{i,j+1} \tag{66}$$

jest stabilny asymptotycznie.

Z twierdzenia 18 wynika następujący ważny wniosek.

Wniosek 3. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest praktycznie niestabilny dla dowolnych skończonych L_1 oraz L_2, jeżeli dwuwymiarowy układ dodatni (66) jest niestabilny.

Twierdzenie 19. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest praktycznie niestabilny jeżeli, przynajmniej jeden element na głównej przekątnej macierzy $\overline{A}_1 + \overline{A}_2$ jest większy od jedności.

Dowód. Jak wiadomo [6], jednowymiarowy układ dodatni (65) jest niestabilny asymptotycznie jeżeli przynajmniej jeden element na głównej przekątnej macierzy $\tilde{A}_1 + \tilde{A}_2$ jest większy od jedności. Ze struktury macierzy $\tilde{A}_1$ i $\tilde{A}_2$ zdefiniowanych przez (60) wynika, że przynajmniej jeden element na głównej przekątnej macierzy $\tilde{A}_1 + \tilde{A}_2$ jest większy od jedności wtedy i tylko wtedy, gdy przynajmniej jeden element na głównej przekątnej macierzy $\overline{A}_1 + \overline{A}_2$ jest większy od jedności. Z twierdzenia 17 wynika, że dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest stabilny praktycznie wtedy i tylko wtedy, gdy jednowymiarowy układ dodatni (65) jest stabilny asymptotycznie. □

Twierdzenie 20. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest niestabilny praktycznie, jeżeli

$$A_k \in \Re_+^{n \times n} \text{ dla } k = 1,2 \tag{67}$$

Dowód. Z twierdzenia 15 dwuwymiarowy układ niecałkowitego rzędu (51) dla $0 < \alpha < 1$ oraz $1 < \beta < 2$ (lub $1 < \alpha < 2$ oraz $0 < \beta < 1$) jest dodatni wtedy i tylko wtedy, gdy warunek (55) jest spełniony. Z (52) wynika, że macierz

$$\overline{A}_1 + \overline{A}_2 = A_1 + A_2 + (\alpha + \beta) I_n \tag{68}$$

ma wszystkie elementy na głównej przekątnej większe od jedności, jeżeli jest spełniony warunek (67). W takim przypadku z twierdzenia 19 wynika, że dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest praktycznie niestabilny. □

2.3.3. Stabilność asymptotyczna

W tym podrozdziale zostanie rozpatrzona praktyczna stabilność dodatnich dwuwymiarowych układów niecałkowitego rzędu dla $L_1 \to \infty$ oraz $L_2 \to \infty$.

Definicja 10. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest nazywany stabilnym asymptotycznie, jeżeli układ ten jest praktycznie stabilny dla $L_1 \to \infty$ oraz $L_2 \to \infty$.

W dowodzie głównego wyniku tego podrozdziału, zostanie wykorzystany następujący lemat.

Lemat 3. Jeżeli $0 < \alpha < 1$ oraz $1 < \beta < 2$ (lub $1 < \alpha < 2$ oraz $0 < \beta < 1$) wtedy

$$\sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{\alpha,\beta}(k,l) = 0 \tag{69}$$

Dowód. W sposób analogiczny do dowodu z lematu 2 może być wykazane, że

$$\sum_{i=0}^{\infty} (-1)^i \binom{\alpha}{i} = \sum_{i=0}^{\infty} (-1)^i \frac{\alpha(\alpha-1)...(\alpha-i+1)}{i!} = 0 \text{ dla } \alpha > 0 \tag{70}$$

Korzystając z zależności (50b) oraz (70) otrzymamy

$$\begin{aligned} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{\alpha,\beta}(k,l) &= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} (-1)^{k+l} \frac{\alpha(\alpha-1)...(\alpha-k+1)\beta(\beta-1)...(\beta-l+1)}{k!\,l!} = \\ &\left(\sum_{k=0}^{\infty} (-1)^k \frac{\alpha(\alpha-1)...(\alpha-k+1)}{k!} \right) \left(\sum_{l=0}^{\infty} (-1)^l \frac{\beta(\beta-1)...(\beta-l+1)}{l!} \right) = 0 \end{aligned} \tag{71}$$

Twierdzenie 21. [10, 20] Dodatni dwuwymiarowy model ogólny z opóźnieniami

$$x_{i+1,j+1} = \sum_{k=0}^{p} \sum_{l=0}^{q} \left(A_{kl}^0 x_{i-k,j-l} + A_{kl}^1 x_{i-k+1,j-l} + A_{kl}^2 x_{i-k,j-l+1} \right) \text{ dla } i, j \in Z_+ \tag{72}$$

gdzie $x_{ij} \in \Re_+^n$ jest wektorem stanu oraz $A_{kl}^t \in \Re_+^{n \times n}$, $k = 0,1,\ldots,p$; $l = 0,1,\ldots,q$; $t = 0,1,2$ jest stabilny asymptotycznie wtedy i tylko wtedy, gdy jednowymiarowy układ dodatni

$$x_{i+1} = \left(\sum_{k=0}^{p} \sum_{l=0}^{q} \left(A_{kl}^0 + A_{kl}^1 + A_{kl}^2 \right) \right) x_i \text{ dla } x_i \in \Re_+^n,\ i \in Z_+ \tag{73}$$

jest stabilny asymptotycznie.

Twierdzenie 22. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy jednowymiarowy układ dodatni

$$x_{i+1} = \left(\hat{A} + I_n \right) x_i,\ \hat{A} = A_0 + A_1 + A_2,\ x_i \in \Re_+^n,\ i \in Z_+ \tag{74}$$

jest stabilny asymptotycznie.

Dowód. Z zależności (51) dla $B_0 = B_1 = B_2 = 0$ oraz (50) otrzymamy

$$x_{i+1,j+1} = A_0 x_{ij} + A_1 x_{i+1,j} + A_2 x_{i,j+1} + \sum_{\substack{k=0 \\ k+l>0}}^{i+1} \sum_{l=0}^{j+1} c_{k,l} x_{i-k+1,j-l+1} \tag{75}$$

gdzie $c_{k,l} = -c_{\alpha,\beta}(k,l)$ dla $k,l \geq 0$ oraz $k + l > 0$.

Z twierdzenia 21 dwuwymiarowy układ dodatni z opóźnieniami jest stabilny asymptotycznie wtedy i tylko wtedy, gdy jednowymiarowy układ dodatni

$$x_{i+1} = \left(\hat{A} + \sum_{\substack{k=0 \\ k+l>0}}^{\infty} \sum_{l=0}^{\infty} c_{k,l} I_n \right) x_i,\ x_i \in \Re_+^n,\ i \in Z_+ \tag{76}$$

jest stabilny asymptotycznie. Z zależności (50b) mamy $c_{00} = -1$ oraz z (69) otrzymamy

$$\sum_{\substack{k=0 \\ k+l>0}}^{\infty} \sum_{l=0}^{\infty} c_{k,l} I_n = I_n \tag{77}$$

Podstawiając (77) do (76) otrzymamy (74). □

Twierdzenie 23. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy jest spełniony jeden z poniższych równoważnych warunków:

1) Wartości własne $z_1, ..., z_n$ macierzy $\hat{A} + I_n$ mają moduły mniejsze od jedności,
2) Wszystkie współczynniki wielomianu charakterystycznego macierzy $\hat{A}$ są dodatnie,
3) Wszystkie minory główne macierzy $-\hat{A}$ są dodatnie.

Twierdzenie 24. Dodatni dwuwymiarowy układ niecałkowitego rzędu (51) jest niestabilny, jeżeli przynajmniej jeden element na głównej przekątnej macierzy $\hat{A}$ jest dodatni.

Dowód. Jeżeli przynajmniej jeden element na głównej przekątnej macierzy $\hat{A}$ jest dodatni wtedy przynajmniej jeden element na głównej przekątnej macierzy $\hat{A}+I_n$ jest większy od jedności, jak wiadomo [4, 8, 20] taki układ (74) jest niestabilny. □

Przykład 4. Korzystając z twierdzenia 23 sprawdzić stabilność asymptotyczną dodatniego dwuwymiarowego układu niecałkowitego rzędu (51) dla $\alpha = 0.3$ i $\beta = 1.2$ oraz

$$A_0 = \begin{bmatrix} 0.4 & 0 \\ 0.1 & 0.5 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -1 & 0 \\ 0.2 & -1.1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.2 & 0 \\ 0.2 & 0.1 \end{bmatrix} \tag{78}$$

Zauważmy, że układ niecałkowitego rzędu jest dodatni, ponieważ macierze

$$\overline{A}_0 = A_0 - I_n\alpha\beta = \begin{bmatrix} 0.04 & 0 \\ 0.1 & 0.14 \end{bmatrix}, \quad \overline{A}_1 = A_1 + I_n\beta = \begin{bmatrix} 0.2 & 0 \\ 0.2 & 0.1 \end{bmatrix}, \quad \overline{A}_2 = A_2 + I_n\alpha = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0.4 \end{bmatrix} \tag{79}$$

mają nieujemne elementy.

W tym przypadku

$$\hat{A} = A_0 + A_1 + A_2 = \begin{bmatrix} -0.8 & 0 \\ 0.5 & -0.5 \end{bmatrix} \tag{80}$$

Pierwszy warunek twierdzenia 23 jest spełniony, gdyż macierz

$$\hat{A} + I_n = \begin{bmatrix} 0.2 & 0 \\ 0.5 & 0.5 \end{bmatrix} \tag{81}$$

ma wartości własne $z_1 = 0.2, \quad z_2 = 0.5$, których moduły są mniejsze od jedności.

Drugi warunek twierdzenia 23 jest również spełniony, ponieważ wielomian charakterystyczny macierzy (80)

$$\det[I_n z - \hat{A}] = \begin{vmatrix} z+0.8 & 0 \\ -0.5 & z+0.5 \end{vmatrix} = z^2 + 1.3z + 0.4 \tag{82}$$

ma dodatnie współczynniki.

Wszystkie minory główne macierzy

$$-\hat{A} = \begin{bmatrix} 0.8 & 0 \\ -0.5 & 0.5 \end{bmatrix} \tag{83}$$

są dodatnie, tj. $\Delta_1 = 0.8, \quad \Delta_2 = 0.4$.

Wszystkie trzy warunki twierdzenia 23 są spełnione, więc dodatni dwuwymiarowy układ niecałkowitego rzędu o macierzach (78) jest stabilny asymptotycznie.

Przykład 5. Korzystając z twierdzenia 24 wykażemy, że dodatni dwuwymiarowy układ niecałkowitego rzędu (51) dla $\alpha = 0.5$ i $\beta = 1.2$ oraz o macierzach

$$A_0 = \begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.7 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -0.1 & 0.3 \\ 0 & -0.2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.4 & 0.2 \\ 0 & -0.5 \end{bmatrix} \tag{84}$$

jest niestabilny.
W tym przypadku macierz

$$\hat{A} = A_0 + A_1 + A_2 = \begin{bmatrix} 0.1 & 0.6 \\ 0.1 & 0 \end{bmatrix} \tag{85}$$

ma jeden dodatni element na głównej przekątnej. Zgodnie z twierdzeniem 24 dodatni dwuwymiarowy układ niecałkowitego rzędu jest więc niestabilny. Ten sam wynik otrzymamy stosując jeden z warunków twierdzenia 23.

2.4. Wykorzystanie liniowych nierówności macierzowych (LMI)

2.4.1. Układy 1D niecałkowitego rzędu

Definicja 11. [36] Nierówność w postaci

$$\mathrm{F}(x) + F > 0 \tag{86}$$

gdzie x przyjmuje wartości w rzeczywistej przestrzeni wektorowej V, odwzorowanie $\mathrm{F}: \mathrm{V} \to S^n$ jest liniowe, oraz $F \in S^n$ (zbiór macierzy symetrycznych), nazywamy liniową nierównością macierzową (LMI).

LMI jest wykonalne (spełnione), jeżeli istnieje taki $x \in \mathrm{V}$, że nierówność (86) jest spełniona; w przeciwnym wypadku LMI jest nazywane niewykonalnym.

Macierz $A = [a_{ij}] \in \Re^{n \times n}$ nazywamy macierzą Metzlera, jeżeli jej elementy poza główną diagonalą są nieujemne, tj. $a_{ij} \geq 0$ dla $i \neq j$, $i, j = 1,...,n$. Macierz $A = [a_{ij}] \in \Re^{n \times n}$ nazywamy macierzą Hurwitza, jeżeli jej wartości własne mają ujemne części rzeczywiste (układ $\dot{x} = Ax$ jest stabilny asymptotycznie). Macierz $A = [a_{ij}] \in \Re^{n \times n}$ nazywamy macierzą Schura, jeżeli jej wartości własne mają moduły mniejsze od jedności (układ $x_{i+1} = Ax_i$ jest stabilny asymptotycznie).

Lemat 4. [15, 36] Macierz $A = \Re_+^{n \times n}$ jest macierzą Schura wtedy i tylko wtedy, gdy LMI

$$blockdiag\,[P - A^T PA, \quad P] \succ 0 \tag{87}$$

jest spełnione dla diagonalnej macierzy P.

Lemat 5. [15, 36] Macierz Metzlera $A = \Re^{n \times n}$ jest macierzą Hurwitza wtedy i tylko wtedy, gdy LMI

$$blockdiag\,[-(A^T P + PA), \quad P] \succ 0 \tag{88}$$

jest spełnione dla diagonalnej macierzy P.

Jak wiadomo, macierz $A = \Re_+^{n\times n}$ jest macierzą Schura wtedy i tylko wtedy, gdy $(A - I_n)$ jest macierzą Hurwitza.

Lemat 6. [15, 36] Macierz nieujemna $A = \Re_+^{n\times n}$ jest macierzą Schura wtedy i tylko wtedy, gdy LMI

$$blockdiag\,[-\left((A-I_n)^T P + P(A-I_n)\right), \quad P] \succ 0 \tag{89}$$

jest spełnione dla diagonalnej macierzy P.

Lemat 7. Macierz nieujemna $A = \Re_+^{n\times n}$ jest macierzą Schura wtedy i tylko wtedy, gdy LMI

$$blockdiag\left\{\begin{bmatrix} P & -A^T P \\ -PA & P \end{bmatrix}, \quad P\right\} \succ 0 \tag{90}$$

jest spełnione dla diagonalnej macierzy P.

Dowód. Weźmy pod uwagę przekształcenie przez kongruencję

$$\begin{bmatrix} I & A^T \\ 0 & I \end{bmatrix}\begin{bmatrix} P & -A^T P \\ -PA & P \end{bmatrix}\begin{bmatrix} I & 0 \\ A & I \end{bmatrix} = \begin{bmatrix} P - A^T PA & 0 \\ 0 & P \end{bmatrix}$$

Jak wiadomo, dodatnia określonosć macierzy jest niezmiennicza względem przekształcenia przez kongruencję. Warunek (90) jest więc równoważny warunkowi (87). □

Twierdzenie 25. Dodatni układ niecałkowitego rzędu (12) jest stabilny praktycznie wtedy i tylko wtedy, gdy jest spełniony jeden z poniższych równoważnych warunków:

1) LMI

$$blockdiag\left\{\begin{bmatrix} P_1 - P_2 - A_\alpha^T P_1 A_\alpha & -c_1 A_\alpha^T P_1 & \dots & -c_{h-1} A_\alpha^T P_1 & -c_h A_\alpha^T P_1 \\ -c_1 P_1 A_\alpha & P_2 - P_3 - c_1^2 P_1 & \dots & -c_1 c_{h-1} P_1 & -c_1 c_h P_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -c_{h-1} P_1 A_\alpha & -c_1 c_{h-1} P_1 & \dots & P_h - P_{h+1} - c_{h-1}^2 P_1 & -c_{h-1} c_h P_1 \\ -c_h P_1 A_\alpha & -c_1 c_h P_1 & \dots & -c_{h-1} c_h P_1 & P_{h+1} - c_h^2 P_1 \end{bmatrix}, \begin{bmatrix} P_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & P_h & 0 \\ 0 & 0 & \dots & 0 & P_{h+1} \end{bmatrix}\right\} \succ 0$$

(91)

jest spełnione dla diagonalnych macierzy $P_1, ..., P_{h+1}$.

2) LMI

$$blockdiag\left\{-\begin{bmatrix} A_\alpha^T P_1 + P_1 A_\alpha - 2P_1 & P_2 + c_1 P_1 & \dots & c_{h-1} P_1 & c_h P_1 \\ P_2 + c_1 P_1 & -2P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{h-1} P_1 & 0 & \dots & -2P_{h-1} & P_{h+1} \\ c_h P_1 & 0 & \dots & P_{h+1} & -2P_h \end{bmatrix}, \begin{bmatrix} P_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & P_h & 0 \\ 0 & 0 & \dots & 0 & P_{h+1} \end{bmatrix}\right\} \succ 0 \tag{92}$$

jest spełnione dla diagonalnych macierzy $P_1,...,P_{h+1}$.

3) LMI

$$blockdiag\left\{\begin{bmatrix} P_1 & 0 & \dots & 0 & -A_\alpha^T P_1 & -P_2 & \dots & 0 \\ 0 & P_2 & \dots & 0 & -c_1 P_1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & P_{h+1} & -c_h P_1 & 0 & \dots & -P_{h+1} \\ -P_1 A_\alpha & -c_1 P_1 & \dots & -c_h P_1 & P_1 & 0 & \dots & 0 \\ -P_2 & 0 & \dots & 0 & 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & -P_{h+1} & 0 & 0 & \dots & P_{h+1} \end{bmatrix}, \begin{bmatrix} P_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & P_h & 0 \\ 0 & 0 & \dots & 0 & P_{h+1} \end{bmatrix}\right\} \succ 0 \tag{93}$$

jest spełnione dla diagonalnych macierzy $P_1,...,P_{h+1}$.

Dowód. Dodatni układ niecałkowitego rzędu (12) jest praktycznie stabilny wtedy i tylko wtedy, gdy macierz $\overline{A}$ jest macierzą Schura. Stosując lemat 4 do układu (22a) otrzymamy LMI (91), ponieważ

$$blockdiag\left[P-\tilde{A}^T P\tilde{A},\quad P\right]=blokdiag\left\{\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\right.$$

$$\left.-\begin{bmatrix} A_\alpha^T & I_n & \dots & 0 & 0\\ c_1 I_n & 0 & \dots & 0 & 0\\ \vdots & \vdots & \dots & \vdots & \vdots\\ c_{h-1} I_n & 0 & \dots & 0 & I_n\\ c_h I_n & 0 & \dots & 0 & 0\end{bmatrix}\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\begin{bmatrix} A_\alpha & c_1 I_n & \dots & c_{h-1} I_n & c_h I_n\\ I_n & 0 & \dots & 0 & 0\\ \vdots & \vdots & \dots & \vdots & \vdots\\ 0 & 0 & \dots & 0 & 0\\ 0 & 0 & \dots & I_n & 0\end{bmatrix},\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\right\}=$$

$$=blockdiag\left\{\begin{bmatrix} P_1-P_2-A_\alpha^T P_1 A_\alpha & -c_1 A_\alpha^T P_1 & \dots & -c_{h-1} A_\alpha^T P_1 & -c_h A_\alpha^T P_1\\ -c_1 P_1 A_\alpha & P_2-P_3-c_1^2 P_1 & \dots & -c_1 c_{h-1} P_1 & -c_1 c_h P_1\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ -c_{h-1} P_1 A_\alpha & -c_1 c_{h-1} P_1 & \dots & P_h-P_{h+1}-c_{h-1}^2 P_1 & -c_{h-1} c_h P_1\\ -c_h P_1 A_\alpha & -c_1 c_h P_1 & \dots & -c_{h-1} c_h P_1 & P_{h+1}-c_h^2 P_1\end{bmatrix},\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\right\}\succ 0$$

Podobnie, stosując lemat 6 do układu (22a) otrzymamy LMI (92), ponieważ

$$blockdiag\left[-\left((\overline{A}-I_n)^T P+P(\overline{A}-I_n)\right),\quad P\right]=$$

$$=blockdiag\left\{-\begin{bmatrix} A_\alpha^T-I_n & I_n & \dots & 0 & 0\\ c_1 I_n & -I_n & \dots & 0 & 0\\ \vdots & \vdots & \dots & \vdots & \vdots\\ c_{h-1} I_n & 0 & \dots & -I_n & I_n\\ c_h I_n & 0 & \dots & 0 & -I_n\end{bmatrix}\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}+\right.$$

$$\left.-\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\begin{bmatrix} A_\alpha-I_n & c_1 I_n & \dots & c_{h-1} I_n & c_h I_n\\ I_n & -I_n & \dots & 0 & 0\\ \vdots & \vdots & \dots & \vdots & \vdots\\ 0 & 0 & \dots & -I_n & 0\\ 0 & 0 & \dots & I_n & -I_n\end{bmatrix},\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\right\}=$$

$$=blockdiag\left\{-\begin{bmatrix} A_\alpha^T P_1+P_1 A_\alpha-2P_1 & P_2+c_1 P_1 & \dots & c_{h-1} P_1 & c_h P_1\\ P_2+c_1 P_1 & -2P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ c_{h-1} P_1 & 0 & \dots & -2P_{h-1} & P_{h+1}\\ c_h P_1 & 0 & \dots & P_{h+1} & -2P_h\end{bmatrix},\begin{bmatrix} P_1 & 0 & \dots & 0 & 0\\ 0 & P_2 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & P_h & 0\\ 0 & 0 & \dots & 0 & P_{h+1}\end{bmatrix}\right\}\succ 0 \tag{94}$$

Stosując lemat 7 do układu (22a) otrzymamy LMI (93), ponieważ

$$blockdiag\left\{W,\ \begin{bmatrix} P_1 & 0 & ... & 0 \\ 0 & P_2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & P_{h+1} \end{bmatrix}\right\} \succ 0$$

$$W = \begin{bmatrix} \begin{bmatrix} P_1 & 0 & ... & 0 \\ 0 & P_2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & P_{h+1} \end{bmatrix} & -\begin{bmatrix} A_\alpha^T & I_n & ... & 0 \\ c_1 I_n & 0 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ c_{h-1} I_n & 0 & ... & I_n \\ c_h I_n & 0 & ... & 0 \end{bmatrix}\begin{bmatrix} P_1 & 0 & ... & 0 \\ 0 & P_2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & P_{h+1} \end{bmatrix} \\ -\begin{bmatrix} P_1 & 0 & ... & 0 \\ 0 & P_2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & P_{h+1} \end{bmatrix}\begin{bmatrix} A_\alpha & c_1 I_n & ... & c_{h-1} I_n & c_h I_n \\ I_n & 0 & ... & 0 & 0 \\ \vdots & \vdots & ... & \vdots & \vdots \\ 0 & 0 & ... & I_n & 0 \end{bmatrix} & \begin{bmatrix} P_1 & 0 & ... & 0 \\ 0 & P_2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & P_{h+1} \end{bmatrix} \end{bmatrix}$$

Przykład 6. Korzystając z LMI sprawdzić stabilność praktyczną dodatniego układu niecałkowitego rzędu

$$\Delta^\alpha x_{k+1} = 0.1 x_k, \quad k \in Z_+ \tag{95}$$

dla $\alpha = 0.5$ oraz $h = 2$.

Korzystając z (18) oraz (22c) otrzymamy

$$c_1 = \frac{\alpha(1-\alpha)}{2} = \frac{1}{8},\ c_2 = \frac{\alpha(\alpha-1)(\alpha-2)}{3!} = \frac{1}{16},\ A_\alpha = 0.6$$

i

$$\tilde{A} = \begin{bmatrix} A_\alpha & c_1 & c_2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & \frac{1}{8} & \frac{1}{16} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Korzystając z twierdzenia 25 oraz z obliczeń w środowisku MATLAB® otrzymamy dla LMI (91)

$$blockdiag\left\{\begin{bmatrix} P_1 - P_2 - A_\alpha^T P_1 A_\alpha & -c_1 A_\alpha^T P_1 & -c_2 A_\alpha^T P_1 \\ -c_1 P_1 A_\alpha & P_2 - P_3 - c_1^2 P_1 & -c_1 c_2 P_1 \\ -c_2 P_1 A_\alpha & -c_1 c_2 P_1 & P_3 - c_2^2 P_1 \end{bmatrix},\ \begin{bmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix}\right\} \succ 0$$

gdzie

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\,[7.8921 \quad 3.5026 \quad 2.1132]$$

dla LMI (92)

$$blockdiag\left\{-\begin{bmatrix} A_\alpha^T P_1 + P_1 A_\alpha - 2P_1 & P_2 + c_1 P_1 & c_2 P_1 \\ P_2 + c_1 P_1 & -2P_1 & P_3 \\ c_2 P_1 & P_3 & -2P_2 \end{bmatrix}, \begin{bmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix}\right\} \succ 0$$

gdzie

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\,[6.9266 \quad 3.1155 \quad 2.6096]$$

oraz dla LMI (93)

$$blockdiag\left\{\begin{bmatrix} P_1 & 0 & 0 & -A_\alpha^T P_1 & -P_2 & 0 \\ 0 & P_2 & 0 & -c_1 P_1 & 0 & -P_3 \\ 0 & 0 & P_3 & -c_2 P_1 & 0 & 0 \\ -P_1 A_\alpha & -c_1 P_1 & -c_2 P_1 & P_1 & 0 & 0 \\ -P_2 & 0 & 0 & 0 & P_2 & 0 \\ 0 & -P_3 & 0 & 0 & 0 & P_3 \end{bmatrix}, \begin{bmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix}\right\} \succ 0$$

gdzie

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\,[7.7203 \quad 3.6738 \quad 2.2765]$$

Wszystkie LMI są spełnione dla macierzy P_1, P_2, P_3 co oznacza, że dodatni układ niecałkowitego rzędu (95) jest stabilny asymptotycznie.

Przykład 7. Korzystając z LMI sprawdzić stabilność praktyczną dodatniego układu niecałkowitego rzędu

$$\Delta^\alpha x_{k+1} = \begin{bmatrix} -0.2 & 1 \\ 0.1 & b \end{bmatrix} x_k, \quad k \in Z_+ \tag{96}$$

dla $\alpha = 0.8$ i $h = 2$, oraz dwóch następujących wartości współczynnika *b*:
Przypadek 1: $b = -0.5$, przypadek 2: $b = 0.5$.

Korzystając z zależności (18) oraz (22c) otrzymamy

$$c_1 = \frac{\alpha(1-\alpha)}{2!} = 0.08, \; c_2 = \frac{\alpha(\alpha-1)(\alpha-2)}{3!} = 0.032$$

oraz

Przypadek 1. $A_{\alpha_1} = A + I_n\alpha = \begin{bmatrix} 0.6 & 1 \\ 0.1 & 0.3 \end{bmatrix}$

i

$$\tilde{A}_1 = \begin{bmatrix} A_{\alpha_1} & c_1 I_2 & c_2 I_2 \\ I_2 & 0 & 0 \\ 0 & I_2 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & 1 & 0.08 & 0 & 0.032 & 0 \\ 0.1 & 0.3 & 0 & 0.08 & 0 & 0.032 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Przypadek 2. $A_{\alpha_2} = A + I_n\alpha = \begin{bmatrix} 0.6 & 1 \\ 0.1 & 1.3 \end{bmatrix}$

i

$$\tilde{A}_2 = \begin{bmatrix} A_{\alpha_2} & c_1 I_2 & c_2 I_2 \\ I_2 & 0 & 0 \\ 0 & I_2 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & 1 & 0.08 & 0 & 0.032 & 0 \\ 0.1 & 1.3 & 0 & 0.08 & 0 & 0.032 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Dla przypadku 1 korzystając z twierdzenia 25 oraz z obliczeń w środowisku MATLAB® otrzymamy dla LMI (91)

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\left[\begin{bmatrix} 16.0915 & 0 \\ 0 & 84.3680 \end{bmatrix}, \begin{bmatrix} 4.2540 & 0 \\ 0 & 16.3556 \end{bmatrix}, \begin{bmatrix} 2.5726 & 0 \\ 0 & 8.6007 \end{bmatrix}\right]$$

dla LMI (92)

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\left[\begin{bmatrix} 8.8848 & 0 \\ 0 & 35.5971 \end{bmatrix}, \begin{bmatrix} 2.5601 & 0 \\ 0 & 7.2962 \end{bmatrix}, \begin{bmatrix} 2.2771 & 0 \\ 0 & 5.2364 \end{bmatrix}\right]$$

oraz dla LMI (93)

$$blockdiag\,[P_1, P_2, P_3] = blockdiag\left[\begin{bmatrix} 13.3199 & 0 \\ 0 & 70.8279 \end{bmatrix}, \begin{bmatrix} 3.537 & 0 \\ 0 & 13.1042 \end{bmatrix}, \begin{bmatrix} 2.2117 & 0 \\ 0 & 7.2682 \end{bmatrix}\right]$$

W przypadku 2 dodatni układ niecałkowitego rzędu (96) jest niestabilny dla każdego *h* (nie tylko dla *h* = 2) ponieważ macierz A_{α_2} posiada jeden element na głównej przekątnej większy od jedności.

Wielomian charakterystyczny macierzy $A_{\alpha_2} - I_n$

$$p(z) = \det[I_n(z+1) - A_{\alpha_2}] = \begin{vmatrix} z-0.4 & -1 \\ -0.1 & z-0.3 \end{vmatrix} = z^2 - 0.7z - 0.22$$

ma dwa ujemne współczynniki. Układ (96) jest więc również niestabilny dla dowolnego *h*.

2.4.2. Układy 2D niecałkowitego rzędu

2.4.2.1. Dwuwymiarowy model Roessera układu niecałkowitego rzędu

W dalszych rozważaniach będziemy korzystać z następującej definicji horyzontalnej i wertykalnej różnicy niecałkowitego rzędu funkcji dwuwymiarowej [23].

Definicja 12. Horyzontalną różnicą niecałkowitego rzędu α funkcji dwuwymiarowej x_{ij}, $i, j \in Z_+$ nazywamy funkcję określoną zależnością

$$\Delta_\alpha^h x_{ij} = \sum_{k=0}^{i} c_\alpha(k) x_{i-k,j} \tag{97a}$$

gdzie $\alpha \in \Re$, $n_1 - 1 < \alpha < n_1 \in N = \{1,2,...\}$ oraz

$$c_\alpha(k) = \begin{cases} 1 & \text{dla} \quad k = 0 \\ (-1)^k \binom{\alpha}{k} = (-1)^k \dfrac{\alpha(\alpha-1)...(\alpha-k+1)}{k!} & \text{dla} \quad k > 0 \end{cases} \tag{97b}$$

Definicja 13. Wertykalną różnicą niecałkowitego rzędu β funkcji dwuwymiarowej x_{ij}, $i, j \in Z_+$ nazywamy funkcję określoną zależnością

$$\Delta_\beta^v x_{ij} = \sum_{l=0}^{j} c_\beta(l) x_{i,j-l} \tag{98a}$$

gdzie $\beta \in \Re$, $n_2 - 1 < \beta < n_2 \in N = \{1,2,...\}$ oraz

$$c_\beta(l) = \begin{cases} 1 & \text{dla} \quad l = 0 \\ (-1)^l \binom{\beta}{l} = (-1)^l \dfrac{\beta(\beta-1)...(\beta-l+1)}{l!} & \text{dla} \quad l > 0 \end{cases} \tag{98b}$$

Lemat 8. [7] Jeżeli $0 < \alpha < 1$ ($0 < \beta < 1$), to

$$c_\alpha(k) < 0 \quad \left(c_\beta(k) < 0\right) \text{ dla } k = 1,2,\ldots \tag{99}$$

Weźmy pod uwagę dwuwymiarowy liniowy układ niecałkowitego rzędu opisany równaniami

$$\begin{bmatrix} \Delta_\alpha^h x_{i+1,j}^h \\ \Delta_\beta^h x_{i,j+1}^v \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_{ij}^h \\ x_{ij}^v \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u_{ij} \tag{100a}$$

$$y_{ij} = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_{i,j}^h \\ x_{i,j}^v \end{bmatrix} + Du_{ij} \quad i, j \in Z_+ \tag{100b}$$

gdzie $x_{ij}^h \in \Re^{n_1}$, $x_{ij}^v \in \Re^{n_2}$ są odpowiednio horyzontalnym i wertykalnym wektorem stanu w punkcie (i, j), $u_{ij} \in \Re^m$ jest wektorem wymuszenia, $y_{ij} \in \Re^p$ jest wektorem odpowiedzi w punkcie (i, j) oraz $A_{11} \in \Re^{n_1 \times n_1}$, $A_{12} \in \Re^{n_1 \times n_2}$, $A_{21} \in \Re^{n_2 \times n_1}$, $A_{22} \in \Re^{n_2 \times n_2}$, $B_1 \in \Re^{n_1 \times m}$, $B_2 \in \Re^{n_2 \times m}$, $C_1 \in \Re^{p \times n_1}$, $C_2 \in \Re^{p \times n_2}$, $D \in \Re^{p \times m}$.

Korzystając z definicji 12 i 13 możemy równanie (100a) napisać w postaci

$$\begin{bmatrix} x_{i+1,j}^h \\ x_{i,j+1}^v \end{bmatrix} = \begin{bmatrix} \overline{A}_{11} & A_{12} \\ A_{21} & \overline{A}_{22} \end{bmatrix} \begin{bmatrix} x_{ij}^h \\ x_{ij}^v \end{bmatrix} - \begin{bmatrix} \sum_{k=2}^{i+1} c_\alpha(k) x_{i-k+1,j}^h \\ \sum_{l=2}^{j+1} c_\beta(l) x_{i,j-l+1}^h \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u_{ij} \tag{101}$$

gdzie $\overline{A}_{11} = A_{11} + \alpha I_{n_1}$ oraz $\overline{A}_{22} = A_{22} + \beta I_{n_2}$.

Z zależności (101) wynika, że dwuwymiarowy układ niecałkowitego rzędu jest układem dwuwymiarowym z opóźnieniami rosnącymi wraz z *i* oraz *j*. Z (97b) oraz (98b) wynika, że współczynniki $c_\alpha(k)$ oraz $c_\beta(l)$ w (101) silnie maleją wraz ze wzrostem *k* oraz *l*. W przypadku praktycznym przyjmuje się, że *k* oraz *l* są ograniczone przez liczby naturalne L_1 oraz L_2. W tym przypadku równanie (101) przyjmuje postać

$$\begin{bmatrix} x_{i+1,j}^h \\ x_{i,j+1}^v \end{bmatrix} = \begin{bmatrix} \overline{A}_{11} & A_{12} \\ A_{21} & \overline{A}_{22} \end{bmatrix} \begin{bmatrix} x_{ij}^h \\ x_{ij}^v \end{bmatrix} - \begin{bmatrix} \sum_{k=2}^{L_1+1} c_\alpha(k) x_{i-k+1,j}^h \\ \sum_{l=2}^{L_2+1} c_\beta(l) x_{i,j-l+1}^h \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u_{ij} \tag{102}$$

Warunki brzegowe dla równań (100a), (101) i (102) mają postać

$$x_{0j}^h \quad \text{dla} \quad j \in Z_+, \quad x_{i0}^v \quad \text{dla} \quad i \in Z_+ \tag{103}$$

Twierdzenie 26. [23] Rozwiązanie równania (101) z warunkami brzegowymi (103) ma postać

$$\begin{bmatrix} x_{i,j}^h \\ x_{i,j}^v \end{bmatrix} = \sum_{p=0}^{i} T_{i-p,j} \begin{bmatrix} 0 \\ x_{p0}^v \end{bmatrix} + \sum_{q=0}^{j} T_{i,j-q} \begin{bmatrix} x_{0q}^h \\ 0 \end{bmatrix} + \sum_{p=0}^{i} \sum_{q=0}^{j} (T_{i-p-1,j-q} B^{10} + T_{i-p,j-q-1} B^{01}) u_{pq} \tag{104a}$$

gdzie

$$B^{10} = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, B^{01} = \begin{bmatrix} 0 \\ B_2 \end{bmatrix} \tag{104b}$$

oraz macierz tranzycji $T_{pq} \in \Re^{n \times n}$ jest zdefiniowana przez

$$T_{pq} = \begin{cases} I_n & \text{dla } p = 0, q = 0 \\ T_{10}T_{p-1,q} + T_{01}T_{p,q-1} - \sum_{k=2}^{p}[c_\alpha(k)I_{n_1} \quad 0]T_{p-k,q} - \sum_{l=2}^{q}[0 \quad c_\beta(l)I_{n_2}]T_{p,q-l} & \text{dla } p+q>0\ (p,q \in Z_+) \\ 0 \quad \text{(zmacierz zerowa)} & \text{dla } p<0 \text{ i/lub } q<0 \end{cases} \tag{105a}$$

gdzie

$$T_{10} = \begin{bmatrix} \overline{A}_{11} & A_{12} \\ 0 & 0 \end{bmatrix}, T_{01} = \begin{bmatrix} 0 & 0 \\ A_{21} & \overline{A}_{22} \end{bmatrix} \tag{105b}$$

Weźmy pod uwagę układ (102) ograniczony przez dwie liczby naturalne L_1 oraz L_2

$$\overline{G}(z_1, z_2) = \begin{bmatrix} I_{n_1} - z_1^{-1}\overline{A}_{11} + \sum_{k=2}^{L_1} c_\alpha(k) z_1^{-k} I_{n_1} & -z_1^{-1}A_{12} \\ -z_2^{-1}A_{21} & I_{n_2} - z_2^{-1}\overline{A}_{22} + \sum_{l=2}^{L_2} c_\beta(l) z_2^{-l} I_{n_2} \end{bmatrix} \tag{106}$$

Niech

$$\det \overline{G}(z_1, z_2) = \sum_{p=0}^{N_1}\sum_{q=0}^{N_2} a_{N_1-p,N_2-q} z_1^{-p} z_2^{-q} \tag{107}$$

gdzie $N_1, N_2 \in Z_+$ są określone przez liczby L_1 ora L_2 w (102).

Twierdzenie 27. [23] Niech (107) będzie wielomianem charakterystycznym układu (102). Wtedy macierze T_{pq} spełniają zależność

$$\sum_{p=0}^{N_1}\sum_{q=0}^{N_2} a_{pq} T_{pq} = 0 \tag{108}$$

Twierdzenie 27 jest uogólnieniem klasycznego twierdzenia Cayleya-Hamiltona na dwuwymiarowe układy niecałkowitego rzędu opisane modelem Roessera (101).

2.4.2.2. Dwuwymiarowy model Roessera dodatniego układu niecałkowitego rzędu

Definicja 13. Układ (100) nazywamy (wewnętrznie) dodatnim dwuwymiarowym układem niecałkowitego rzędu wtedy i tylko wtedy, gdy $x_{ij}^h \in \Re_+^{n_1}$, $x_{ij}^v \in \Re_+^{n_2}$ oraz $y_{ij} \in \Re_+^p$ $i, j \in Z_+$ dla dowolnych warunków brzegowych $x_{0j}^h \in \Re_+^{n_1}$, $j \in Z_+$ i $x_{i0}^v \in \Re_+^{n_2}$, $i \in Z_+$ oraz wszystkich wymuszeń $i, j \in Z_+$.

Twierdzenie 28. [23] Dwuwymiarowy układ niecałkowitego rzędu (101) dla $\alpha, \beta \in \Re$, $0 < \alpha \le 1$, $0 < \beta \le 1$ jest dodatni wtedy i tylko wtedy gdy

$$\begin{bmatrix} \overline{A}_{11} & A_{12} \\ A_{21} & \overline{A}_{22} \end{bmatrix} \in \Re_+^{n\times n}, \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \in \Re_+^{n\times m}, \quad [C_1 \quad C_2] \in \Re_+^{p\times n}, \quad D \in \Re_+^{p\times m} \tag{109}$$

Weźmy pod uwagę dodatni układ niecałkowitego rzędu złożony z modelu Roessera (101) ze sprzężeniem zwrotnym od wektora stanu

$$u_{ij} = [K_1 \quad K_2] \begin{bmatrix} x_{i,j}^h \\ x_{i,j}^v \end{bmatrix} \tag{110}$$

gdzie $K = [K_1 \quad K_2] \in \Re^{m\times n}$, $K_j \in \Re^{m\times n_j}$, $j = 1,2$ jest macierzą wzmocnień,

Poszukiwać będziemy takiej macierzy wzmocnień K, dla której układ zamknięty

$$\begin{bmatrix} x_{i+1,j}^h \\ x_{i,j+1}^v \end{bmatrix} = \begin{bmatrix} \overline{A}_{11} + B_1K_1 & A_{12} + B_1K_2 \\ A_{21} + B_2K_1 & \overline{A}_{22} + B_2K_2 \end{bmatrix} \begin{bmatrix} x_{ij}^h \\ x_{ij}^v \end{bmatrix} - \begin{bmatrix} \sum_{k=2}^{i+1} c_\alpha(k) x_{i-k+1,j}^h \\ \sum_{l=2}^{j+1} c_\beta(l) x_{i,j-l+1}^h \end{bmatrix} \tag{111}$$

jest dodatni i stabilny asymptotycznie.

Twierdzenie 29. Dodatni zamknięty układ niecałkowitego rzędu (111) jest dodatni i stabilny asymptotycznie wtedy i tylko wtedy, gdy istnieje macierz blokowo diagonalna

$$\Lambda = blockdiag\,[\Lambda_1, \Lambda_2], \quad \Lambda_k = diag\,[\lambda_{k1},...,\lambda_{kn_k}], \quad \lambda_{kj} > 0, \quad k = 1,2; \quad j = 1,...,n_k \tag{112}$$

oraz macierz rzeczywista

$$D = [D_1 \quad D_2], \quad D_k \in \Re^{m\times n_k}, \quad k = 1,2 \tag{113}$$

spełniające warunki

$$\begin{bmatrix} \overline{A}_{11}\Lambda_1 + B_1D_1 & A_{12}\Lambda_2 + B_1D_2 \\ A_{21}\Lambda_1 + B_2D_1 & \overline{A}_{22}\Lambda_2 + B_2D_2 \end{bmatrix} \in \Re_+^{n\times n} \tag{114}$$

$$\begin{bmatrix} A_{11}\Lambda_1 + B_1D_1 & A_{12}\Lambda_2 + B_1D_2 \\ A_{21}\Lambda_1 + B_2D_1 & A_{22}\Lambda_2 + B_2D_2 \end{bmatrix} \begin{bmatrix} 1_{n_1} \\ 1_{n_2} \end{bmatrix} < \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{115}$$

gdzie $1_{n_k} = [1 \quad ... \quad 1]^T \in \Re_+^{n_k}$, $k = 1,2$.

Macierz wzmocnień dana jest zależnością

$$K = [K_1 \quad K_2] = [D_1 \quad D_2]\Lambda^{-1} = [D_1\Lambda_1^{-1} \quad D_2\Lambda_2^{-1}] \tag{116}$$

Dowód oraz procedura wyznaczania macierzy wzmocnień K dane są w [23].

Jak wiadomo [10, 23], dodatni układ zamknięty (111) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy jednowymiarowy układ dodatni o macierzy

$$\begin{bmatrix} \overline{A}_{11}+B_1K_1 & A_{12}+B_1K_2 \\ A_{21}+B_2K_1 & \overline{A}_{22}+B_2K_2 \end{bmatrix} - \sum_{k=2}^{\infty} \begin{bmatrix} I_{n_1} c_\alpha(k) & 0 \\ 0 & I_{n_2} c_\beta(k) \end{bmatrix} \tag{117}$$

jest stabilny asymptotycznie.
Biorąc pod uwagę, że [23]

$$\sum_{k=2}^{\infty} c_\alpha(k) = \alpha - 1, \quad \sum_{k=2}^{\infty} c_\beta(k) = \beta - 1$$

oraz $\overline{A}_{11} = A_{11} + \alpha I_{n_1}$ i $\overline{A}_{22} = A_{22} + \beta I_{n_2}$ możemy macierz (117) napisać w postaci

$$\begin{bmatrix} \hat{A}_{11}+B_1K_1 & A_{12}+B_1K_2 \\ A_{21}+B_2K_1 & \hat{A}_{22}+B_2K_2 \end{bmatrix} = A + BK \tag{118}$$

gdzie $\hat{A}_{11} = A_{11} + I_{n_1}$ i $\hat{A}_{22} = A_{22} + I_{n_2}$ oraz

$$A = \begin{bmatrix} \hat{A}_{11} & A_{12} \\ A_{21} & \hat{A}_{22} \end{bmatrix}, B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \tag{119}$$

Twierdzenie 30. Zamknięty układ niecałkowitego rzędu (111) jest dodatni i stabilny asymptotycznie wtedy i tylko wtedy, gdy istnieje dodatnio określona macierz blokowa (112) oraz macierz rzeczywista (113) takie, że warunek (114) jest spełniony i LMI

$$\begin{bmatrix} -\Lambda & A\Lambda + BD \\ (A\Lambda + BD)^T & -\Lambda \end{bmatrix} \prec 0 \tag{120}$$

jest wykonalne względem dodatnio określonej diagonalnej macierzy Λ.

Dowód. Układ zamknięty (111) jest dodatni wtedy i tylko wtedy, gdy warunek (114) jest spełniony, ponieważ warunek

$$\begin{bmatrix} \overline{A}_{11}+B_1K_1 & A_{12}+B_1K_2 \\ A_{21}+B_2K_1 & \overline{A}_{22}+B_2K_2 \end{bmatrix} = \begin{bmatrix} \overline{A}_{11}+B_1D_1\Lambda_1^{-1} & A_{12}+B_1D_2\Lambda_2^{-1} \\ A_{21}+B_2D_1\Lambda_1^{-1} & \overline{A}_{22}+B_2D_2\Lambda_2^{-1} \end{bmatrix} =$$

$$= \begin{bmatrix} \overline{A}_{11}\Lambda_1+B_1D_1 & A_{12}\Lambda_2+B_1D_2 \\ A_{21}\Lambda_1+B_2D_1 & \overline{A}_{22}\Lambda_2+B_2D_2 \end{bmatrix} \begin{bmatrix} \Lambda_1^{-1} & 0 \\ 0 & \Lambda_2^{-1} \end{bmatrix} \in \Re_+^{n\times n}$$

jest równoważny do warunku (114).
Dodatni układ zamknięty (111) jest stabilny asymptotycznie wtedy i tylko wtedy, gdy LMI [15]

$$P - (A+BK)^T P (A+BK) \succ 0 \tag{121}$$

jest spełnione dla dodatnio określonej diagonalnej macierzy *P*.
Korzystając z uzupełnienia Schur możemy warunek (121) zapisać w postaci

$$\begin{bmatrix} -P & P(A+BK) \\ (A+BK)^T P & -P \end{bmatrix} \prec 0 \tag{122}$$

Podstawiając (116) i $P = \Lambda^{-1}$ do (122) otrzymamy

$$\begin{bmatrix} -\Lambda^{-1} & \Lambda^{-1}(A+BD\Lambda^{-1}) \\ (A+BD\Lambda^{-1})^T \Lambda^{-1} & -\Lambda^{-1} \end{bmatrix} =$$

$$= blockdiag\,[\Lambda^{-1}, \Lambda^{-1}] \begin{bmatrix} -\Lambda & A\Lambda + BD \\ (A\Lambda + BD)^T & -\Lambda \end{bmatrix} blockdiag\,[\Lambda^{-1}, \Lambda^{-1}] \prec 0 \tag{123}$$

Stosując przekształcenie przez kongruencję z macierzą przekształcenia $blockdiag\,[\Lambda, \Lambda]$ otrzymamy warunek (120).

Przykład 8. Dany jest dwuwymiarowy model Roessera niecałkowitego rzędu dla $\alpha = 0.4$, $\beta = 0.5$ oraz

$$A_{11} = \begin{bmatrix} -0.5 & -0.1 \\ 0.1 & 0.01 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} -0.1 & -0.1 \\ 0.2 & 0.1 \end{bmatrix},$$

$$A_{21} = \begin{bmatrix} -0.3 & -0.1 \\ 0.2 & 0.1 \end{bmatrix}, \quad A_{22} = \begin{bmatrix} -1 & -0.1 \\ 0.4 & 0.1 \end{bmatrix}, \tag{124}$$

$$B_1 = \begin{bmatrix} -0.2 \\ 0.1 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} -0.3 \\ 0.2 \end{bmatrix}.$$

Należy wyznaczyć macierz wzmocnień $K = [K_1 \quad K_2]$, $K_i \in \Re^{1\times 2}$, i = 1,2 tak, aby układ zamknięty był dodatni i stabilny asymptotycznie.

Dwuwymiarowy modelu Roessera niecałkowitego rzędu z macierzami (124) nie jest dodatni, ponieważ macierze stanu posiadają elementy ujemne. Układ jest również niestabilny ponieważ macierz

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} -0.5 & -0.1 & -0.1 & -0.1 \\ 0.1 & 0.01 & 0.2 & 0.1 \\ -0.3 & -0.1 & -1 & -0.1 \\ 0.2 & 0.1 & 0.4 & 0.1 \end{bmatrix} \tag{125}$$

ma dodatnie elementy na głównej przekątnej.

Wybieramy

$$D = [D_1 \quad D_2], \qquad D_1 = [-0.4 \quad -0.2], \qquad D_2 = [-0.4 \quad -0.2] \tag{126}$$

Korzystając z twierdzenia 30 oraz z obliczeń w środowisku MATLAB® otrzymamy dla LMI (120)

$$\Lambda = blockdiag\,[\Lambda_1, \Lambda_2], \quad \Lambda_1 = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0.2258 & 0 \\ 0 & 0.2413 \end{bmatrix} \tag{127}$$

co oznacza, że LMI jest spełnione dla diagonalnej macierzy Λ.
Korzystając z zależności (116) otrzymamy macierz wzmocnień w postaci

$$K = [K_1 \quad K_2] = [D_1\Lambda_1^{-1} \quad D_2\Lambda_2^{-1}] = [-1 \quad -0.5 \quad -1.7715 \quad -0.8288] \tag{128}$$

Układ zamknięty jest dodatni, ponieważ macierze

$$\overline{A}_{11} + B_1K_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.36 \end{bmatrix}, \quad A_{12} + B_1K_2 = \begin{bmatrix} 0.2543 & 0.0658 \\ 0.0229 & 0.0171 \end{bmatrix},$$

$$A_{21} + B_2K_1 = \begin{bmatrix} 0 & 0.05 \\ 0 & 0 \end{bmatrix}, \quad \overline{A}_{22} + B_2K_2 = \begin{bmatrix} 0.0314 & 0.1487 \\ 0.0457 & 0.4342 \end{bmatrix}$$

mają tylko nieujemne elementy.
Układ zamknięty jest też stabilny asymptotycznie, ponieważ wielomian charakterystyczny

$$\det \begin{bmatrix} I_{n_1} z - (A_{11} + B_1K_1) & -(A_{12} + B_1K_2) \\ -(A_{21} + B_2K_1) & I_{n_2} z - (A_{22} + B_2K_2) \end{bmatrix} = \begin{vmatrix} z + 0.3 & 0 & -0.2543 & -0.0658 \\ 0 & z + 0.04 & -0.0229 & -0.0171 \\ 0 & -0.05 & z + 0.4686 & -0.1487 \\ 0 & 0 & -0.0457 & z + 0.0658 \end{vmatrix} =$$

$$= z^4 + 0.8743z^3 + 0.2166z^2 + 0.0141z + 0.0003$$

ma dodatnie współczynniki.

Przykład 9. Dany jest dodatni dwuwymiarowy model Roessera niecałkowitego rzędu, w którym $\alpha = 0.4$, $\beta = 0.9$ oraz

$$A_{11}=\begin{bmatrix}-0.4 & 0.01\\ 0.03 & 0.001\end{bmatrix},\qquad A_{12}=\begin{bmatrix}0.01 & 0.01\\ 0.01 & 0.2\end{bmatrix},$$

$$A_{21}=\begin{bmatrix}0.01 & 0.2\\ 0 & 0.01\end{bmatrix},\qquad A_{22}=\begin{bmatrix}-0.9 & 0.01\\ 0.01 & -0.8\end{bmatrix},\tag{129}$$

$$B_1=\begin{bmatrix}0\\ 0.001\end{bmatrix},\qquad B_2=\begin{bmatrix}0\\ 0.002\end{bmatrix}.$$

Należy wyznaczyć macierz wzmocnień $K=[K_1 \quad K_2]$, $K_i \in \Re^{1\times 2}$, i = 1,2 tak, aby układ zamknięty był dodatni i stabilny asymptotycznie.

Dwuwymiarowy układ niecałkowitego rzędu o modelu Roessera z macierzami (129) jest niestabilny, ponieważ macierz

$$\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}=\begin{bmatrix}-0.4 & 0.01 & 0.01 & 0.01\\ 0.03 & 0.001 & 0.01 & 0.2\\ 0.01 & 0.2 & -0.9 & 0.01\\ 0 & 0.01 & 0.01 & -0.8\end{bmatrix}\tag{130}$$

ma dodatnie elementy na głównej przekątnej.

Wybieramy

$$D=[D_1 \quad D_2],\quad D_1=[0.13 \quad -0.37],\quad D_2=[-3.19 \quad -0.11]\tag{131}$$

Korzystać z twierdzenia 30 oraz z obliczeń w środowisku MATLAB® otrzymamy dla LMI (120) otrzymamy

$$\Lambda=blockdiag\,[\Lambda_1,\Lambda_2],\quad \Lambda_1=\begin{bmatrix}0.0554 & 0\\ 0 & 0.0755\end{bmatrix},\quad \Lambda_2=\begin{bmatrix}0.8659 & 0\\ 0 & 0.0032\end{bmatrix}\tag{132}$$

co oznacza, że LMI jest spełniony dla diagonalnej macierzy Λ.

Korzystając z zależności (116) otrzymamy macierz wzmocnień w postaci

$$K=[K_1 \quad K_2]=[D_1\Lambda_1^{-1} \quad D_2\Lambda_2^{-1}]=[2.3466 \quad -4.9007 \quad -3.6840 \quad -34.3750]\tag{133}$$

Układ zamknięty jest dodatni, ponieważ macierze

$$\overline{A}_{11}+B_1K_1=\begin{bmatrix}0 & 0.01\\ 0.0323 & 0.3961\end{bmatrix},\quad A_{12}+B_1K_2=\begin{bmatrix}0.01 & 0.01\\ 0.0063 & 0.1659\end{bmatrix},$$

$$A_{21}+B_2K_1=\begin{bmatrix}0.01 & 0.2\\ 0.0047 & 0.0002\end{bmatrix},\quad \overline{A}_{22}+B_2K_2=\begin{bmatrix}0 & 0.01\\ 0.0026 & 0.7513\end{bmatrix},$$

mają tylko nieujemne elementy.

Układ zamknięty jest też stabilny asymptotycznie ponieważ, wielomian charakterystyczny

$$\det\begin{bmatrix} I_{n_1}z-(A_{11}+B_1K_1) & -(A_{12}+B_1K_2) \\ -(A_{21}+B_2K_1) & I_{n_2}z-(A_{22}+B_2K_2) \end{bmatrix} = \begin{vmatrix} z+0.4 & -0.01 & -0.01 & -0.01 \\ -0.0323 & z+0.0039 & -0.0063 & -0.1659 \\ -0.01 & -0.2 & z+0.9 & -0.01 \\ -0.0047 & -0.0002 & -0.0026 & z+0.1487 \end{vmatrix} =$$

$$= z^4 + 1.4527z^3 + 0.5572z^2 + 0.0544z + 0.0001$$

ma dodatnie współczynniki.

2.5. Uwagi końcowe

Wprowadzono pojęcie stabilności praktycznej i stabilności asymptotycznej dyskretnych dodatnich układów i układów stożkowych niecałkowitego rzędu. Podano warunki konieczne i wystarczające stabilności dla tych układów niecałkowitego rzędu. Wykazano, że badanie stabilności dodatnich układów 2D można sprowadzić do badania stabilności odpowiadających dodatnich układów 1D. Podano trzy metody LMI badania stabilności układów dodatnich niecałkowitego rzędu.

Pokazano możliwość zastosowania metody LMI do wyznaczania macierzy wzmocnień sprzężeń zwrotnych od wektora stanu dla modelu Roessera, tak aby układ zamknięty był dodatni i stabilny asymptotycznie. Podano warunki konieczne i wystarczające istnienie rozwiązania tego problemu. Efektywność proponowanych metod LMI pokazano na przykładach numerycznych modelu Roessera niecałkowitego rzędu. Podane rozważania można łatwo uogólnić na układy dodatnie 1D i 2D z opóźnieniami niecałkowitego rzędu. Uogólnienie tych rozważań na układy ciągłe 1D i 2D niecałkowitego rzędu jest problemem otwartym.

Praca wykonana w ramach pracy statutowej S/WE/1/106. Wydziału Elektrycznego Politechniki Białostockiej

Bibliografia

1. Busłowicz M., Robust stability of positive discrete-time linear systems with multiple delays with unity rank uncertainty structure or non-negative perturbation matrices, *Bull. Pol. Acad. Sci. Techn.* Vol. 55, No. 1, 2007, 347-350.
2. Busłowicz M., Simple stability conditions for linear positive discrete-time systems with delays, *Bull. Pol. Acad. Sci. Techn.*, Vol. 56, No. 4, 2008
3. Farina L., Rinaldi S., Positive Linear Systems; Theory and Applications, J. Wiley, New York 2000.
4. Fornasini E., Marchesini G., Double indexed dynamical systems, in *Math. Sys. Theory*, 12, 1978, 59-72.
5. Gałkowski K., Kummert A., Fractional polynomials and *n*D systems. *Proc IEEE Int. Symp. Circuits and Systems*, ISCAS'2005, Kobe, Japan, 2005, CD-ROM.
6. Kaczorek T., Positive 1D and 2D Systems, Springer-Verlag, London 2002.
7. Kaczorek T., Reachability and controllability to zero of positive fractional discrete-time systems. *Machine Intelligence and Robotics Control*, Vol. 6, No. 4, 2007.

8. Kaczorek T., Reachability and controllability to zero of cone fractional linear systems, *Archives of Control Sciences*, Vol. 17, No. 3, 2007, 357-367.
9. Kaczorek T., Asymptotic stability of positive 2D linear systems, *Proc. of 13th Scientific Conf. Computer Applications in Electrical Engineering*, April 14-16, 2008, Poznan, Poland.
10. Kaczorek T., Asymptotic stability of positive 1D and 2D linear systems, *Recent Advances in Control and Automation*, Acad. Publ. House EXIT, 2008, 41-52.
11. Kaczorek T., Fractional positive continuous-time linear systems and their reachability, *Int. J. Appl. Math. Comput. Sci.*, Vol. 18, No. 2, 2008, 223-228.
12. Kaczorek T., Fractional 2D linear systems, *Journal of Automation, Mobile Robotics & Intelligent Systems*, Vol. 2, No. 2, 2008, 5-9.
13. Kaczorek T., Positive different orders fractional 2D linear systems, *Acta Mechanica et Automatica*, Vol. 2, No. 2, 2008, 51-58.
14. Kaczorek T., Reachability and minimum energy control of positive 2D systems with delays, *Control and Cybernetics*, Vol. 34, No 2, 2005, 411-423.
15. Kaczorek T., LMI approach to stability of 2D positive systems, *Multidimensional Systems and Signal Processing*, Vol. 20, No. 1, 2009, 39-54.
16. Kaczorek T., Positive 2D fractional linear systems, *COMPEL*, Vol. 28, No. 2, 2009, 341-352.
17. Kaczorek T., Asymptotic stability of positive 2D linear systems with delays, *Bull. Pol. Acad. Sci. Techn*. Vol. 57, No. 2, 2009, 133-137.
18. Kaczorek T., Reachability and controllability to zero tests for standard and positive fractional discrete-time systems, *Journal of Automation and System Engineering*, Vol.42, No. 6-7-8, 2008, 769-787.
19. Kaczorek T., Computation of realizations of discrete-time cone systems, *Bull. Pol. Acad. Sci. Techn*. Vol. 54, No. 3, 2006, 347-350.
20. Kaczorek T., Positive 2D systems with delays, *MMAR 2006, 12th IEEE IFAC International Conference on Methods in Automation and Robotics*, 2006, Poland.
21. Kaczorek T., Practical stability of positive fractional discrete-time linear systems, *Bull. Pol. Acad. Sci. Techn.*, Vol. 56, No. 4, 2008, 313-317.
22. Kaczorek T., Independence of the asymptotic stability of the 2D linear systems with delays of their delays, *Intern. J. Applied Math. and Comp. Sci.*, Vol. 19, No.2 (in Press).
23. Kaczorek T., Rogowski K., Positivity and stabilization of fractional 2D linear systems described by Roesser model, *Proc. of Conf. Methods and Models in Automation and Robotics*, 2009, Miedzyzdroje, Poland.
24. Klamka J., Controllability of dynamical systems, Kluwer Academic Publ., Dordrecht, 1991.
25. Klamka J., Positive controllability of positive systems, *Proc. of American Control Conference, ACC-2002*, Anchorage, 2002, (CD-ROM).
26. Kurek J., The general state-space model for a two-dimensional linear digital systems, *IEEE Trans. Autom. Contr*. AC-30, 1985, 600-602.
27. Miller K.S., Ross B., An Introduction to the Fractional Calculus and Fractional Differential Equations, Willey, New York 1993.
28. Nashimoto K., Fractional Calculus, Descartes Press, Koriyama, 1984.
29. Oldham K. B., Spanier J., The Fractional Calculus, New York: Academic Press, 1974.
30. Ostalczyk P., Epitome of the Fractional Calculus, Theory and its Applications in Automatics, Wydawnictwo Politechniki Lodzkiej, Lodz 2008 (in Press).
31. Ostalczyk P., The non-integer difference of the discrete-time function and its application to the control system synthesis, *Int. J. Syst, Sci.* Vol. 31, No. 12, 2000, 1551-1561.

32. Oustaloup A., Commande CRONE. Paris, Hermés, 1993.
33. Podlubny I., Fractional Differential Equations, San Diego: Academic Press, 1999.
34. Roesser R. P., A discrete state-space model for linear image processing, *IEEE Trans. Autom. Contr.*, AC-20(1), 1975,1-10.
35. Sierociuk D., Dzieliński D., Fractional Kalman filter algorithm for the states, parameters and order of fractional system estimation. *Int. J. Appl. Math. Comp. Sci*., Vol. 16, No. 1, 2006, 129-140.
36. Twardy M., An LMI approach to checking stability of 2D positive systems, *Bull. Pol. Acad. Techn. Sci.* Vol. 55, No. 4, 2007, 379-383.
37. Valcher M. E., On the internal stability and asymptotic behavior of 2D positive systems. *IEEE Trans. on Circuits and Systems – I*, Vol. 44, No. 7, 1997, 602-613.
38. Vinagre M., Feliu V., Modeling and control of dynamic system using fractional calculus: Application to electrochemical processes and flexible structures, *Proc. 41st IEEE Conf. Decision and Control*, Las Vegas, NV, 2002, 214-239.

3. Ионно-лучевые технологии в микро-, нано- и оптоэлектронике, в ядерно-физических методах анализа материалов и приборных структур

Ф.Ф. Комаров, О.В. Мильчанин, А.М. Миронов, А.С. Камышан

НИИ прикладных физических проблем имени А.Н. Севченко Белорусского государственного университета, ул. Курчатова 7, Минск, 220108, Беларусь
тел. +375(17)2124833, KomarovF@bsu.by

Представлен ряд разработок, доведенных до технологического исполнения, позволяющих получать уникальные структуры микро-, нано- и оптоэлектроники с использованием ионных пучков. Рассмотрена уникальная система неразрушающего элементного и структурного анализа материалов с нанометровым разрешением по глубине.

A few important implementations of ion beams to produce unique structures of microelectronics and optoelectronics are presented. A novel system for non-destructive elemental and structural analysis with the nanometer depth resolution is also treated.

3.1. Введение

В последнее время, в современных технологиях создания СБИС и оптоэлектронных полупроводниковых приборов, можно уверенно выделить тенденцию использования ионной имплантации примесей, не относящихся к легирующим. В данном случае используются эффекты накопления и трансформации дефектов для создания локальных областей полупроводника, обладающих специфическими (требуемыми) свойствами. Можно говорить и о появлении целого класса технологий, где используются протонные пучки. Использование протонной имплантации обусловлено уникальными свойствами атомов водорода. Благодаря большой химической активности, водород может образовывать специфичный тип дефектов как с атомами матрицы, так и с атомами легирующей примеси, а также собственные водородо-индуцированные дефекты, которые при определенных условиях остаются стабильными даже при высоких температурах отжига. С другой стороны, благодаря

малой массе ионов водорода, имплантационные слои за исключением области остановки ионов (вблизи проективного пробега ионов R_p) остаются практически бездефектными. Все эти уникальные свойства протонных пучков позволили разработать ряд новых технологий, использующих эффекты примесно-дефектной инженерии [1,2]. В данной работе авторами представлен ряд разработок, доведенных до технологического исполнения, позволяющих получать уникальные структуры с использованием протонных пучков. Рассмотрены также реализованные в нашей лаборатории методы элементного и структурного анализа материалов приборных структур с нанометровым разрешением по глубине.

3.2. Создание структур кремний-на-изоляторе

Структуры кремний на изоляторе обладают следующими преимуществами перед другими известными типами КНИ-структур и обычными полупроводниковыми подложками:

– повышение радиационной и термической стойкости;

– увеличение выхода годных (меньшая площадь элементов и чипов снижает вероятность попадания ростового дефекта в активную область);

– увеличение плотности компоновки элементов из-за формирования на поверхности полупроводниковых элементов субмикропрофильных “канавок” шириной 0,1–0,2 мкм;

– упрощение: полное отсутствие эффекта защелкивания за счет отсутствия изоляцией *p-n* переходами, создаваемыми имплантацией; упрощение изоляции и уменьшение ее размеров в 3–5 раз (и всего чипа); сочетание многих функций на одном чипе (например, логических, телекоммуникационных и силовых для мобильной связи или управления);

– лучшими свойствами: двух-трёхкратное увеличение быстродействие и/или снижение энергопотребление при низком напряжении (менее 1,5 В); работа при ультранизком напряжении питания (порядка 5 В); работа при высоких температурах, вплоть до 500 °С за счет уменьшения токов утечек переходов меньших размеров.

Предлагаемая технология КНИ-структур может быть реализована на типовом оборудовании, имеющемся практически на любом предприятии – изготовителе электронных изделий. Использование изготавливаемых КНИ структур в указанных приборах обеспечит существенную экономию энергоресурсов, даст дополнительную экономию за счет большей надежности, позволит решить некоторые принципиально новые задачи при создании объектов, работающих в экстремальных условиях.

Среди десятка различных методов производства КНИ-пластин можно выделить два доминирующих и достигших промышленного освоения: SIMOX и Smart-Cut. При оценке стоимости конечных КНИ-пластин, возможности варьирования их свойств, использования стандартного технологического оборудования и техпроцессов, наиболее простым и перспективным выглядит метод Smart-Cut, сочетающий процессы прямого соединения окисленных пластин и прецизионного ионного отслоения. Авторами выполнены работы по отработке, развитию и адаптации к существующему технологическому оборудованию на НПО «Интеграл» технологии производства КНИ-пластин («по мотивам» метода Smart-Cut) [3-10].

В качестве исходных использовали стандартные пластины кремния диаметром 100 и 150 мм (100)-ориентации, легированные бором (КДБ-12). Исходные пластины отбирались с учетом минимальных прогиба (< 5 мкм) и неплоскостности (< 10 мкм) пластин. Имплантация ионов водорода проводилась на ускорителе Skanibal 128S, имеющем газовый источник. Режимы имплантации: энергия 80–100 кэВ, доза $4–5\times10^{16}$

$H_2^+/см^2$, температура мишени < 50 °С. Имплантация проводилась в структуры SiO_2(20–300 нм)/Si. Набор дозы осуществлялся поэтапно, в несколько шагов, с выдержкой пластин между режимами имплантации в вакууме при комнатной температуре в течение 6–8 часов.

Для очистки пластин и формирования гидрофильных поверхностей использовались в различных вариациях процедуры плазменной обработки, химической и гидромеханической очистки. Были отработаны оригинальные методики подготовки химически чистых оксидированных поверхностей пластин с высокой степенью гидрофильности [8]. Процедура связывания имплантированных пластин со структурами SiO_2(20–200 нм)/Si проводилась вручную с использованием специально разработанной оснастки. Контроль качества связывания осуществлялся на-просвет в ближнем ИК-диапазоне электромагнитного излучения [10].

Для усиления связи между пластинами, а также для исследований влияния дополнительных низкотемпературных обработок (НТО) на качество получаемых КНИ-пластин, отжиги проводили при температурах 80–200 °С с различными длительностями (вплоть до 24 часов). Для термически-вызываемого полного скола по водородо-индуцированному дефектному слою проводили отжиг при температурах 450–550 °С в течение 20–60 минут. Для части образцов использовался неполный отжиг (длительность 1–15 минут) с последующим механическим сколом по дефектному слою. Финишная термообработка получаемых КНИ-структур (для полного отжига дефектов структуры в верхнем кристаллическом слое кремния) проводилась при температурах 1050–1100 °С в среде кислорода или азота.

На рис.1 представлены ПЭМ микрофотографии поперечных сечений образцов после имплантации ионов водорода и последующей термообработки [8]. Дефектный слой, формируемый имплантацией ионов водорода с дозами $4–5\times10^{16}$ $H_2^+/см^2$, достаточно протяженный и состоит из кластеров точечных дефектов. Отжиг при 450 °С уже в течение 5 минут приводит к существенному уменьшению толщины дефектного слоя. При этом по всей толщине этого слоя формируются микротрещины (рис. 1Б). Увеличение длительности термообработки приводит к формированию макротрещины параллельно поверхности пластин (рис. 1В). Отжиг имплантированных пластин кремния ионами водорода (с дозами $4–5\times10^{16}$ $H_2^+/см^2$) при температурах 400 °С не всегда приводил к формированию макротрещин даже при больших длительностях (60 минут). При использовании термообработок при более высоких температурах (500–600 °С) макротрещины формируются быстрее, однако в этом случае наблюдается больший разброс их местоположения в дефектном слое. Отжиг пакета связанных пластин (имплантированной водородом и окисленной) при температуре 450 °С и длительности до 15 минут не приводил к разъединению по водородо-индуцированному слою. Но пластины удавалось разделить механически. Исследования методом РЭМ сразу после отжига не выявили отличий в шероховатости поверхностей образцов КНИ-структур, полученных как полным термическим, так и механическим сколом. Но при этом в образцах сколотых механически были обнаружены протяженные дефекты (трещины) в переносимом слое, по-видимому, за счет напряжений при механическом воздействии.

Финишная термообработка сколотых от пластин доноров КНИ-структур при температурах 1050–1100 °С приводит к полному отжигу структурных дефектов в поверхностном слое (рис. 2А). На рис. 2Б представлен типичный профиль элементного состава в КНИ структуре, полученный методов ОЖЕ-спектроскопии в сочетании с послойным травлением образца [7]. Можно выделить резкие ступеньки атомной концентрации на границах раздела в КНИ структуре. Симметричность профиля примеси относительно захороненного оксидного слоя свидетельствует

о сравнимых границах раздела Si/SiO_2 и SiO_2/Si. Результаты ПЭМ исследований хорошо согласуются с ОЖЕ-профилями. Можно отметить высокое качество и гомогенность слоев и границ раздела в КНИ-структуре. По структурным свойствам верхний кристаллический слой КНИ пластин сравним с исходными пластинами кремния — не выявлено образования дополнительных дефектов структуры в слое при формировании КНИ структур. Исследования КНИ-образцов методом АСМ показывают, что шероховатость верхнего кристаллического слоя кремния в образцах, изготавливаемых без дополнительного низкотемпературного отжига, не превышает 7–7,5 нм. Это значение ниже, чем известные из литературы (10–20 нм).

Данный факт, вероятно, связан с большим отжигом радиационных дефектов во время имплантации за счет ступенчатого набора дозы и длительной выдержки между этапами имплантации. И, как результат, финальный дефектный слой становится более узким (по сравнению с одностадийным набором дозы). Минимальная шероховатость поверхности (на уровне 2 нм) зарегистрирована для образцов, где использовались также и режимы НТО.

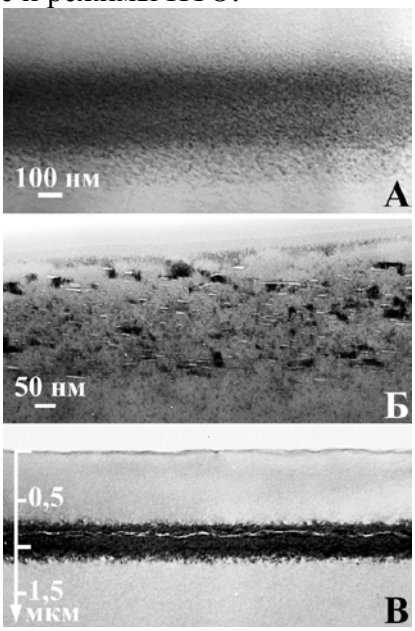


Рис. 1. Светлопольные ПЭМ микрофотографии дефектного слоя в кремниевых пластинах после имплантации ионов водорода (А) и отжига при 450 °С: 5 минут (Б) и 10 минут (В)

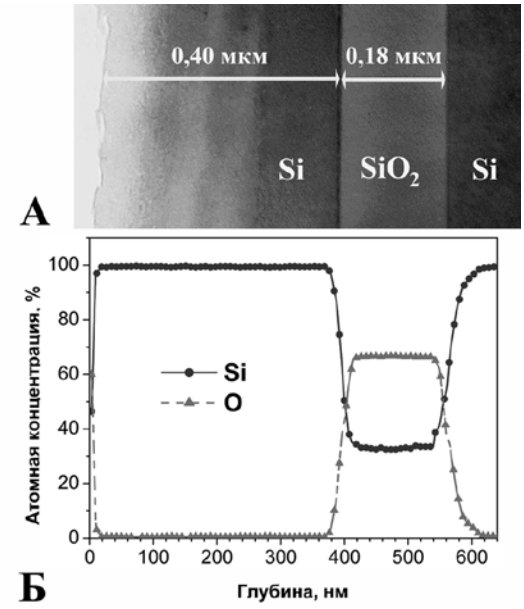


Рис. 2. ПЭМ фотография сечения КНИ-пластины (А) и ОЖЕ-профиль элементного состава (Б)

Таким образом, продемонстрирована возможность создания качественных КНИ-пластин с использованием стандартного технологического оборудования. С использованием многоступенчатого набора дозы имплантации водорода, а также дополнительных низкотемпературных отжигов, в работе показана возможность существенного снижения шероховатости поверхности КНИ-пластин (изготавливаемых в сочетании методов прямого связывания окисленных пластин и прецизионного ионного скола) вплоть до 2 нм. К настоящему времени получено два патента Республики Беларусь на указанную технологию.

3.3. Создание внутренних геттерирующих слоев в кремнии

Проектируемая спецификация будущих кремниевых приборов предполагает максимальное содержание примесей металлов не более $2{,}5\times10^9$ ат./см^2. Поэтому, при

создании Si-приборов, производители все чаще дополняют строгий регламент чистых комнат процессами геттерирования. Традиционно, для очистки активных областей от примесей металлов и дефектов применяются геттерирующие слои, создаваемые на непланарной стороне пластины. Но, для уменьшения глубины залегания *p-n*-переходов требуются более низкие температуры термообработки и короткие времена отжига, что приводит к снижению эффективности геттерирующих слоев, сформированных на непланарной стороне пластины.

Поэтому, создание геттерирующих слоев, локализованных в непосредственной близости к активным элементам, стало в последнее время объектом интенсивных исследований. Интересными для промышленного применения являются методы создания внутреннего геттерирующего слоя основанные на внедрении средних доз ионов H^+ и He^+. Недавние исследования по формированию микропустот (пор) при имплантации ионов водорода или гелия в кремний продемонстрировали высокую эффективность геттерирования таких примесных металлов как Cu, Ni, Co, Fe, Ag, Au, Pd. Но, следует отметить, что эффективность геттерирования, как правило, определяют на модельных экспериментах. Авторами работ [4,6,7] было исследовано влияние создаваемого геттерирующего слоя, с использованием протонной имплантации, на работу тестовых структур, максимально приближенных к реальным полупроводниковым приборам.

В качестве исходных использовались пластины Cz-Si (001)-ориентации n-типа проводимости (КЭФ-4,5) и структуры, содержащие эпитаксиальные слои Si (2 мкм, 1 Ом·см) на (111)-Si (КЭФ-0,01). Режимы формирования внутренних геттерирующих слоев представлены в табл. 1. Для конкретной энергии имплантации водорода выбирали дозы, при которых концентрация водорода в области максимума распределения достаточно высока для создания структурных дефектов, но ниже критической (~$1{,}5–2\times10^{21}$см$^{-3}$), когда происходит выделение большого количества газовых пузырьков уже после имплантации, что приводит к появлению микротрещин и других нежелательных дефектов структуры.

Для определения оптимальных условий формирования геттерирующих слоев, последующий термический отжиг образцов проводился при различных температурах. Для исследований образцов с геттерирующими слоями были изготовлены специальные тестовые структуры, содержащие диоды Шоттки. Режимы получения тестовых структур представлены в табл. 2.

Таблица 1 — Режимы формирования геттерирующих слоев

№	Тип подложки	Энергия и доза имплантации H^+	Режимы термического отжига
1	Si (2 мкм, 1 Ом·см)/(111)Cz-Si (КЭФ-0,01)	215 (H_2^+) кэВ, $2{,}5\times10^{16}$ см$^{-2}$	1) 800 °С, 5 мин. 2) 800 °С, 30 мин. 3) 900 °С, 5 мин. 4) 1000 °С, 5 мин. 5) 900 °С, 15 мин.
2	Si (2 мкм, 1 Ом·см)/(111)Cz-Si (КЭФ-0,01)	215 (H_2^+) кэВ, 3×10^{16} см$^{-2}$	1) 800 °С, 5 мин. 2) 800 °С, 30 мин. 3) 900 °С, 5 мин. 4) 1000 °С, 5 мин. 5) 900 °С, 15 мин.
3	(001)Cz-Si (КЭФ-4,5)	75 (H^+) кэВ, 2,5×1016 см$^{-2}$	900 °С, 30 мин.
4	(001)Cz-Si (КЭФ-4,5)	75 (H^+) кэВ, $3{,}5\times10^{16}$ см$^{-2}$	900 °С, 30 мин.

Таблица 2 — Режимы формирования тестовых структур

№	Проводимая операция	Описание режимов операции	Температура обработки
1	Формирование диодов Шоттки	1. Напыление пленки металла (Ni) испарения 2. Формирование силицида никеля (NiSi)	< 50 °C 450 °C
2	Формирование контактных областей	1. Напыление пленки металла (Al) 2. Вжигание алюминия для формирования контакта	< 50 °C 450 °C

Были проведены структурные исследования слоев кремния, содержащих водородо-индуцированные слои и дефекты как после имплантации, так и после термообработок. На рис. 3а представлено светлопольное ПЭМ изображение структуры внутреннего дефектного слоя, сформированного имплантацией ионов водорода с выбранными режимами (табл. 1).

Детальный анализ структуры показывает, что данный барьерный слой состоит из мелких водородо-вакансионных комплексов и кластеров точечных дефектов. Последующая термическая обработка при температурах ниже 600 °C (в течение 15–120 минут) не приводит к существенным изменениям структуры и перераспределению примеси в имплантированных водородом слоях кремния. Лишь при температурах отжига выше 650 °C наблюдается отжиг радиационно-индуцированных дефектов вблизи R_p.

Плотность дефектов в области R_p значительно уменьшается и наблюдается формирование только двух типов водородо-индуцированных дефектов: «пластинчатых» и «петлеподобных» дефектов структуры, размер которых составляет 10–60 нм и 150–250 нм соответственно. Дальнейшее увеличение температуры отжига до 800 °C (в течение 5 минут) приводит к сильному снижению концентрации «пластинчатых» дефектов (рис.3б). С другой стороны, зарегистрировано увеличение слоевой плотности и размеров петлеподобных дефектов. При этом, в центральной части отдельных больших петлеподобных дефектов наблюдается формирование микропустот.

При более высоких температурах обработки или увеличении длительности отжига происходит практически полный отжиг водородо-индуцированных дефектов вблизи R_p. На рис.3в представлено изображение структуры кремния имплантированного ионами водорода после отжига при температуре 900 °C в течение 15 минут. Наблюдается узкий дефектный слой вблизи R_p, содержащий большое количество микропустот, которые соединены дефектами дислокационного типа. При этом, структурное качество кремниевой матрицы от поверхности до дефектного слоя сравнимо с чистыми исходными подложками, что подтверждается РОР исследованиями в сочетании с каналированием.

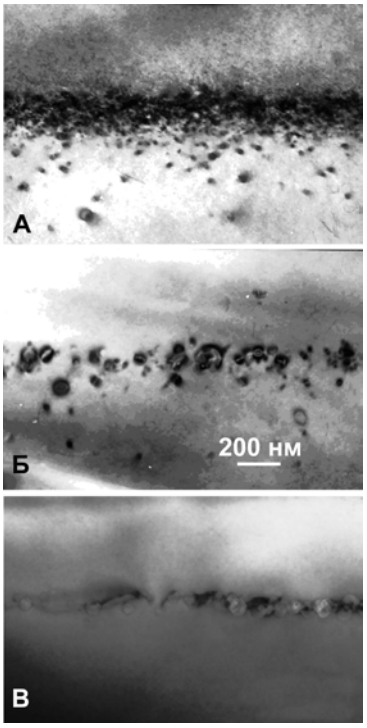


Рис. 3. Светлопольные ПЭМ изображения в поперечном сечении структуры внутреннего дефектного слоя в кремнии после имплантации (140 кэВ, 10^{16} $см^{-2}$) ионов водорода (А) и последующего термического отжига: 800 °С, 5 минут (Б); 900 °С, 15 минут (В)

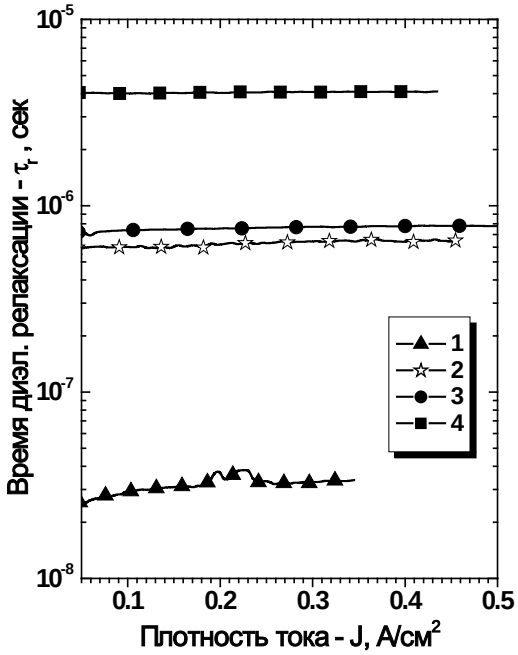


Рис. 4. Зависимость времени жизни неосновных носителей заряда от плотности тока обратносмещенного диода Шоттки в образцах без (1) и с внутренним геттером (2–4), полученным имплантацией ионов водорода (215 кэВ, $2{,}5\times10^{16}$ $см^{-2}$) и последующим термическим отжигом: 2 – 1000 °С, 5 минут; 3 – 800 °С, 30 минут; 4 – 900 °С, 15 минут.

Исследования тестовых диодов Шоттки, сформированных в эпитаксиальном кремнии, содержащем внутренние геттерирующие слои, проводили с использованием CV-измерений. Результаты исследований можно обобщить следующим образом:

– Наблюдается сильное уменьшение высокочастотной проводимости при измерениях в обратносмещенных диодах Шоттки в образцах с геттерирующими слоями по сравнению с исходными, что связано с уменьшением центров захвата для неосновных носителей заряда.

– Геттерирующая эффективность увеличивается как с повышением температуры, так и длительности отжига (рис. 4). Время жизни неосновных носителей заряда существенно увеличивается (в 160 раз, что более чем на 2 порядка величины больше чем для исходных структур) в случае формирования геттера в режимах: имплантация ионов водорода – 215 кэВ, $2{,}5\times10^{16}$ $см^{-2}$, последующий отжиг в среде N_2 при температуре 900 °С в течение 15 минут.

Данные DLTS исследований находятся в хорошем согласии с результатами CV-измерений. В тестовых структурах диодов Шоттки, созданных на эпитаксиальных слоях кремния, обнаруживаются глубокие уровни, связанные с ловушечными состояниями для носителей заряда, что приводит к появлению пиков DLTS спектров. При наличии геттера наблюдается практически полная аннигиляция глубоких уровней в тестовых диодах Шоттки.

Таким образом, были разработаны основные режимы формирования в кремниевых пластинах внутреннего геттера, состоящего из узкого барьерного слоя, содержащего большое количество микропустот. Результаты исследований методами

DLTS и C-V измерений свидетельствует о повышении структурного совершенства эпитаксиальных слоев кремния (за счет геттерирования) в тестовых диодах Шоттки. Установлено, что использование геттерирующих слоев позволяет на 2 порядка снизить концентрацию глубоких уровней в эпитаксиальных слоях кремния, связанных с наличием дефектов и нежелательных металлических примесей.

3.4. Применение имплантации протонов для изоляции приборов на полупроводниках A^3B^5

Ионная модификация полупроводниковых кристаллов бинарных и тройных полупроводниковых соединений, таких как GaN, GaAs, InP, AlGaAs делает возможным формирование в них изолирующих областей. Внедряемые ионы, передавая энергию атомам материала, создают дефекты структуры кристалла, которым соответствуют глубокие уровни-ловушки в запрещенной зоне полупроводника, захватывающие свободные носители заряда, в результате чего материал становится изолирующим. Преимущество ионной имплантации перед традиционным методом изоляции — мезатравлением, состоит в сохранении планарности, даже при интеграции приборов вертикальной структуры (PIN-диоды) с планарными приборами (FET). В настоящее время технология изоляции с помощью имплантации протонов используется для создания гетероэпитаксиальных транзисторов, фотодетекторов, лазеров, волноводов, а также для электрической изоляции приборов в монолитных интегральных схемах.

Для создания изоляции необходимого качества требуются равномерные по толщине эпитаксиального слоя распределения дефектов с концентрацией, специфической для типа полупроводника и уровня его легирования. Дозы внедряемых ионов при этом должны соответствовать оптимальной концентрации создаваемых дефектов. При более низких дозах ионного облучения скорость удаления носителей оказывается недостаточной для подавления проводимости; при повышенных дозах плотность дефектов становится настолько высокой, что сопротивление уменьшается из-за включения механизма прыжковой проводимости. Чем выше уровень легирования, тем большая концентрация дефектов требуется для создания изолирующей области.

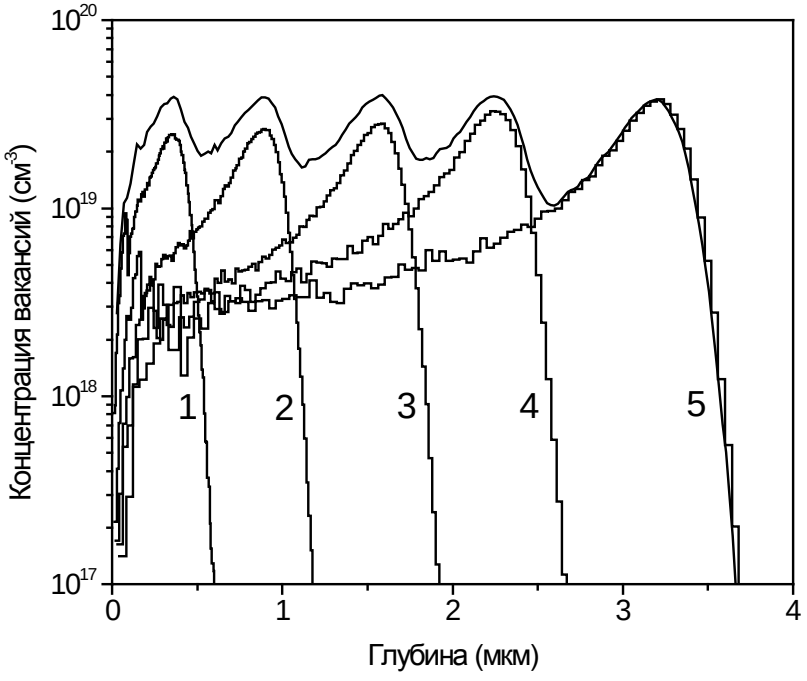


Рис. 5. Распределения вакансий, сформированных в GaAs полиэнергетической имплантацией ионов H^+ в арсенид галлия с энергиями 50 (1), 130 (2), 220 (3), 300 (4) и 400 (5) кэВ и дозами соответственно $4{,}2\times10^{13}$, $5{,}3\times10^{13}$, $6{,}2\times10^{13}$, $8{,}0\times10^{13}$ и $1{,}0\times10^{14}$ см$^{-2}$

Таким образом, широкое практическое использование ионной имплантации для формирования изоляции в бинарных и тройных полупроводниковых соединениях возможно только при условии, что для заданных типов полупроводников определены оптимальные параметры ионного легирования и термообработок. В связи с этим, нами разработана физико-математическая модель, позволяющая рассчитать оптимальные энергии и дозы для имплантации заданного типа ионов в заданный материал исходя из формы профиля радиационных дефектов (обратная задача полиэнергетической ионной имплантации) [11,12].

Применительно к задаче формирования изолирующих областей в полупроводниках бинарных и тройных полупроводниковых соединений, искомый профиль является равномерным распределением дефектов кристаллической структуры до заданной глубины эпитаксиального слоя.

Пример расчета распределения вакансий при имплантации протонов с оптимальными энергиями и дозами приведен на рис. 5. Видно, что для получения равномерного распределения вакансий по глубине от 0 до 3,5 мкм достаточно провести полиэнергетическую имплантацию ионов H^+ с энергиями от 50 до 400 кэВ.

Изготовлены экспериментальные образцы на эпитакисальных структурах GaAs n-типа (рис. 6) и измерены электрофизические параметры сформированных изолирующих областей [13-17]. Проведены испытания стабильности созданной изоляции. В допустимых пределах не обнаружено изменения электрофизических свойств созданной изоляции. Результаты испытаний: термостабильность изоляции — не менее 300 °С; пробивное напряжение при ширине изолирующего слоя не менее 4 мкм — не менее 200 В; ток утечки при напряжении 5В — не более 10 нА.

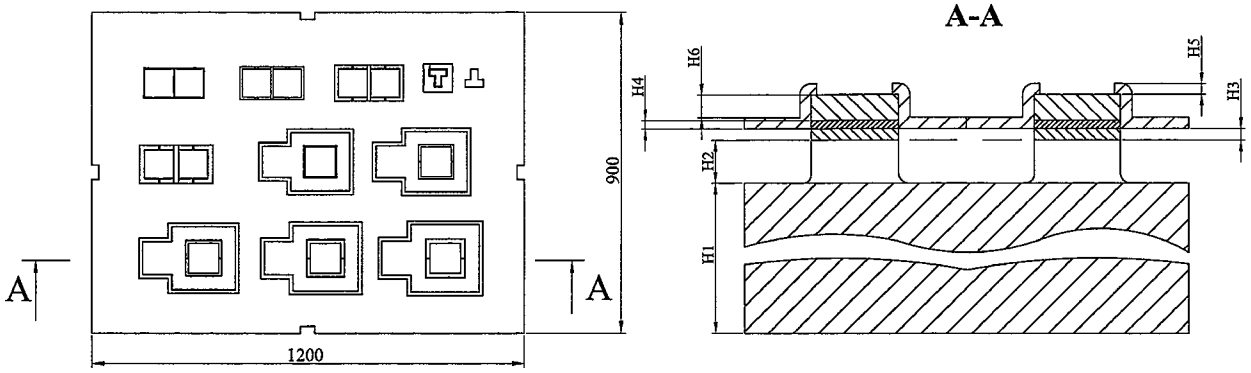


Рис. 6. Тестовая структура для оценки качества изоляции

Нами также исследовалась зависимость слоевого сопротивления изолирующих слоев, полученных имплантацией протонов в GaAs *n*-типа, от температуры постимплантационного отжига, а также частотная зависимость проводимости до и после отжига.

Рис. 7. Ускоритель ионов AN-2500 системы Van de Graaf

Полиэнергетическая имплантация протонов в образцы кристаллического GaAs n-типа (толщина 400±20 мкм, сопротивление 0,55±0,05 Ом·см, концентрация носителей $2{,}7\pm0{,}4\times10^{15}$ см$^{-3}$, подвижность носителей 4180 см^{2}В$^{-1}$с$^{-1}$) проводилась на ускорителе Van de Graaf (рис. 7), с поддержанием плотности ионного тока 0,15 μA/см^2. Предварительно на поверхность пластин GaAs были нанесены слои омических контактов (слой эвтектического сплава 88 ат.% Au + 12 ат.% Ge толщиной 0,15 мкм, слой Ni толщиной 0,5 мкм и слой Au толщиной 0,1 мкм) и дополнительно слой Au толщиной 1 мкм на обратную сторону пластин. Энергии (400, 300, 220, 130 и 60 кэВ) и дозы (2×10^{14}, $1{,}5\times10^{14}$, $1{,}2\times10^{14}$, $1{,}1\times10^{14}$ и $1{,}0\times10^{14}$ см$^{-2}$ соответственно) имплантируемых протонов были рассчитаны с помощью разработанной нами программы PROFCON исходя из условия получения равномерного распределения радиационных повреждений на глубине до 3,7 мкм от поверхности GaAs.

Отжиг производился в течение 15 минут при температурах в диапазоне 50–500 °C. Проводимость при постоянном и переменном токе измерялась с помощью системы HIOKI 3532, с ошибкой не более 0,1%. Частота переменного тока изменялась в пределах от 50 Гц до 1 МГц.

На рис. 8 приведены зависимости слоевого сопротивления от температуры отжига, измеренные при частотах 1, 10, 100 кГц и 1 МГц переменного тока (AC) и при постоянном токе (DC). Видно, что сразу после имплантации протонами слоевое сопротивление образцов составляет примерно 10^8 Ом·см. При отжиге сопротивление увеличивается и достигает максимума 5×10^8 Ом·см при 320 °C для постоянного тока. Из частотной зависимости сопротивления можно сделать вывод о том, что проводимость обусловлена прыжковым механизмом. При увеличении температуры отжига прыжковая проводимость подавляется, при температуре более 380 °C зонный механизм проводимости становится основным, а роль прыжкового механизма снижается, что соответствует отжигу радиационных дефектов.

Для групп образцов, отожженных при различных температурах, на частоте 1 МГц была измерена зависимость проводимости от температуры, что позволило определить энергию активации ΔE прыжковой проводимости. Для образцов, отожженных при

низких температурах (160°С), получена величина ΔE = 0,4 эВ. Температуре 240 °С соответствует две величины энергии активации, 0,2 эВ и 0,5 эВ. Для 340 °С энергия активации составляет 0,65 эВ. Такие значения энергии активации прыжковой проводимости характерны для переходов электронов через потенциальный барьер между уровнями дефектов. Энергия активации 0,65 эВ соответствует уровню Е4 (0,67 эВ), связанному с дефектным комплексом $As_{Ga} + V_{As}$. Уровень с энергией активации 0,39–0,40 эВ, наблюдавшийся в облученном протонами GaAs *n*-типа, также связан с комплексным дефектом.

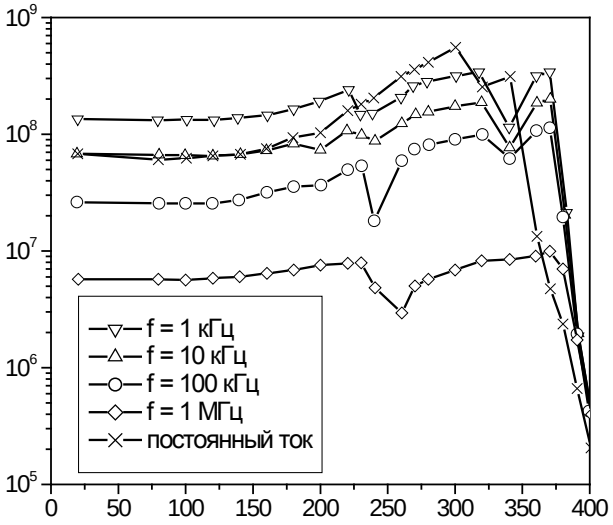


Рис. 8. Зависимость слоевого сопротивления от температуры отжига, измеренная при различных частотах переменного тока и при постоянном токе

Таким образом, проведенные исследования имплантированного протонами GaAs *n*-типа, а также на линейке фотодетекторов ИК-диапазона на этом материале, позволили получить изоляцию достаточного качества и сделать выводы о механизмах переноса заряда в облученных слоях.

3.5. Комплекс для элементного анализа твердотельных материалов

В основу функционирования разработанного комплекса для элементного анализа примесного состава твердотельных материалов положен принцип регистрации и математической обработки энергетических спектров обратного рассеяния ионов[18]. При измерении энергетических спектров заряженных частиц с использованием газовых или твердотельных детекторов одним из основных измерительных узлов является анализатор импульсов. Это связано с тем, что амплитуда импульса на выходе вышеуказанных типов детекторов прямо пропорциональна выделенной в них энергии и задачей измерения является построение зависимости скорости счета (количества зарегистрированных импульсов в единицу времени) от их амплитуды, т.е. энергии частиц.

В случае использования электростатического анализатора энергии ионов (ЭСА) измеряется скорость счета в зависимости от напряжения питания ЭСА. При фиксированном питании ЭСА вырезает из спектра полосу, ширина которой (ширина канала) определяется конструкцией анализатора. Амплитуда импульса на выходе

полупроводникового детектора, работающего в режиме счета импульсов, пропорциональна энергии регистрируемых частиц, поэтому для расширения энергетического диапазона в сторону меньших энергий необходимо максимально уменьшить уровень шумов спектрометрического тракта.

Известно, что движение заряженных частиц в однородных электростатических полях неизбежно сопровождается изменением энергии иона. Поэтому, для получения преломляющей системы, не меняющей энергии ионного пучка, используется неоднородное поле цилиндрического конденсатора. На рис. 9 приведен схематический чертеж ЭСА.

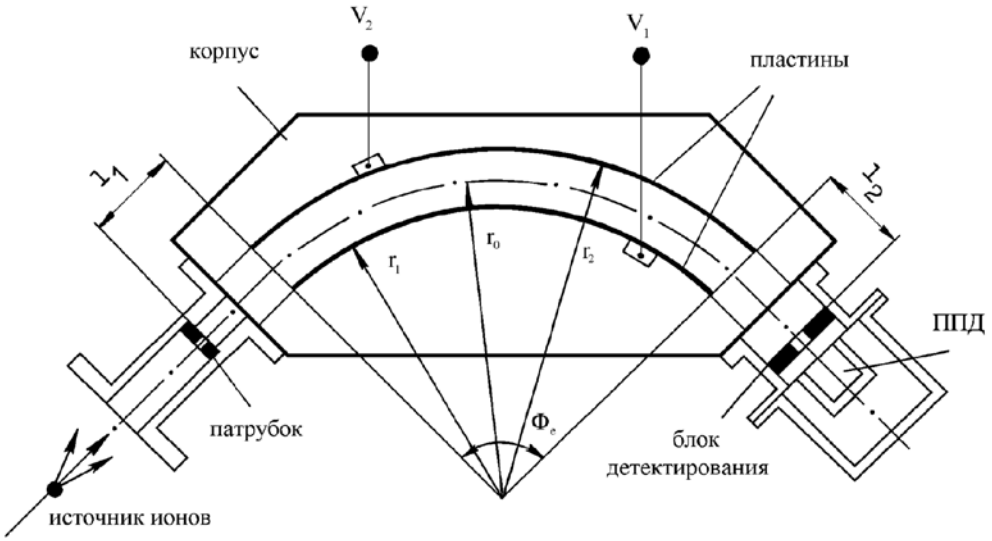


Рис. 9. Схема ЭСА

Известно, что потенциал, подаваемый на обкладки цилиндрического конденсатора, линейно связан с энергией частицы, двигающейся по окружности с радиусом r_0 Коэффициент пропорциональности определяется только геометрическими размерами конденсатора, то есть:

$$U_{\text{ЭСА}} = 2U_{\text{уск.}} \cdot \ln (r_2 / r_1),$$

где $U_{\text{уск.}}$ – ускоряющий потенциал ускорителя ионов.

Секторное электрическое поле может фокусировать по направлению ионный пучок определенной энергии с малым углом раствора при любых значениях угла Φ_e. Для построения хода лучей необходимо принять, что угол сектора равен $\Phi_e / 2^{0,5}$, а радиус кривизны траектории сектора равен $r_0/2^{0,5}$. Данная система эквивалентна оптической линзе в сочетании с призмой, преломляющий угол которой равен $\Phi_e / 2^{0,5}$. Таким образом, задавая значения r_0 , r_2 , r_1 и Φ_e , получим расстояние от границ сектора ($l_1 = l_2$), на котором необходимо устанавливать входную и выходную диафрагмы, а также коэффициент пропорциональности между кинетической энергией частицы и потенциалом, подаваемым на пластины анализатора.

Энергетическое разрешение ЭСА определяется выражением:

$$\frac{\Delta E}{E} = \frac{s_1 + s_2}{r_0}, \qquad (1)$$

где s_1, s_2 – ширины входной и выходной диафрагм, соответственно; r_0 – радиус центральной траектории движения ионов в ЭСА.

Конструктивно ЭСА состоит из трех узлов: входного переходника, корпуса с электродами и блока детектирования.

Таким образом, основные отличия при измерении энергетического спектра заряженных частиц при использовании ЭСА по сравнению с твердотельными и газовыми спектрометрическими детекторами заключается в следующем:

– электростатический анализатор энергии ионов селектирует частицы определенной энергии интенсиметром, а не анализатором импульсов;

– для реализации измерения спектра требуется прецизионный, регулируемый с малым шагом, высоковольтный источник питания;

– нормировка натекающего на мишень заряда в случае ЭСА должна проводиться для каждого канала, что обусловлено слабой светосилой и поканальным измерением спектра, на что требуются большие затраты времени измерений, в течение которых колебания тока ионного пучка могут быть достаточно большими (более 1 %).

В состав комплекса входят: ионопровод с системой коллимации пучка; электростатический анализатор энергии ионов; вакуумная камера с гониометром, на котором установлен держатель образца; система регистрации и управления[18]. На рис. 10 приведена структурная блок-схема аналитического модуля.

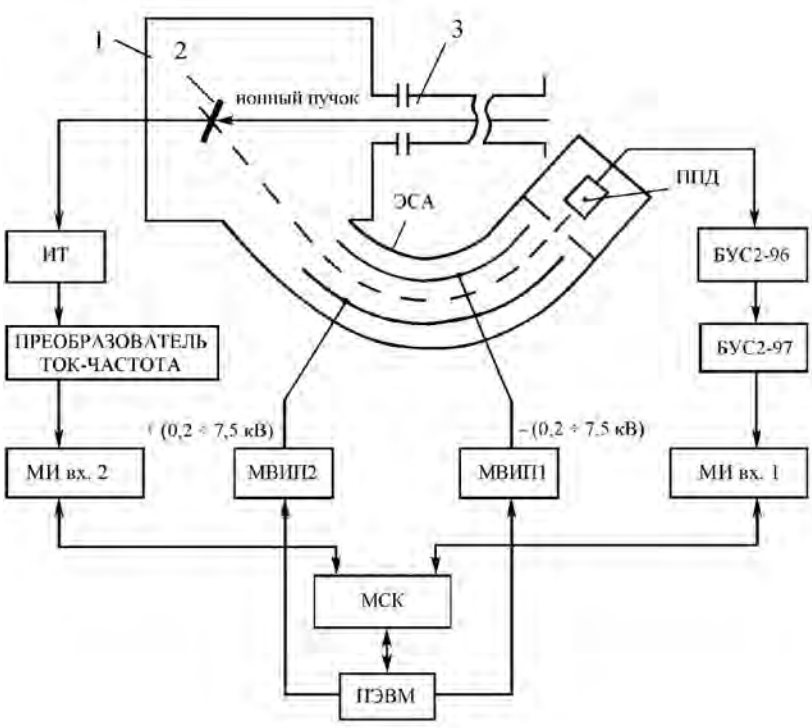


Рис. 10. Структурная блок-схема аналитического модуля:

ЭСА – электростатический анализатор; ППД – полупроводниковый детектор; БУС2-96 – предварительный блок усиления спектрометрический; БУС2-97 – основной блок усиления спектрометрический; МИ – модуль программно-управляемого интенсиметра; МВИП – модуль программно-управляемого высоковольтного источника питания; ИТ – измеритель тока; МСК – модуль системного контроллера; 1 – вакуумная камера; 2 – образец; 3 – ионопровод

Функционально электрическая часть комплекса делится на две подсистемы – регистрации и управления. Первая из них составляет основу информационно-измерительного комплекса. Она включает в себя объединенные локальной информационной шиной (системной магистралью): БУС2-97 – спектрометрический

усилитель; МОД – модуль программируемого одноканального дискриминатора; МИ – модуль программируемого интенсиметра; ИП – источник питания подсистемы.

Подсистема управления включает в себя два программируемых модуля высоковольтных источников питания МВИП (до 7,5 кВ) и источник питания ИП.

Связь подсистем с управляющей персональной ЭВМ реализуется через интерфейс USB при помощи концентратора. Предусмотрена возможность подключения модулей высоковольтных источников и системного контроллера подсистемы регистрации непосредственно к USB портам ПЭВМ.

Для определения соответствия значения подаваемого на ЭСА высокого напряжения энергии регистрируемых ионов проводилась его калибровка. Имеются два варианта проведения этой процедуры.

Первый заключается в следующем. На место мишени ставится образец, представляющий собой пленку достаточно тяжелого металла, напыленную на более легкую подложку. Затем измеряются энергетические спектры рассеянных пленкой ионов при четырех-пяти значениях энергии падающего на образец ионного пучка и проводится оценка соответствия между положением точки на половине высоты высокоэнергетической границы измеренного спектра и значением потенциала, подаваемого на анализатор. Разность между положениями этих точек при разных значениях энергии анализирующего пучка на шкале энергий деленные на разности потенциалов, соответствующих этим точкам, $\Delta U/7{,}5$ дадут искомую ширину канала анализатора.

Во втором способе при фиксированной энергии измеряются спектры с нескольких образцов, на которые напылена пленка металла с разным Z, например, Au и Cr, и определяются расстоянием между границами спектров в единицах энергии ионов и напряжения, подаваемого на анализатор. Разделив эту разность в шкале энергий на $\Delta U/7{,}5$, определяем энергетическую ширину канала анализатора, т.к. 7,5 B – минимальный шаг, с которым может изменяться высокое напряжение на ЭСА. Эта операция проделывается для трех-четырех значений энергии. После этого ширина канала определяется как среднее этих измерений.

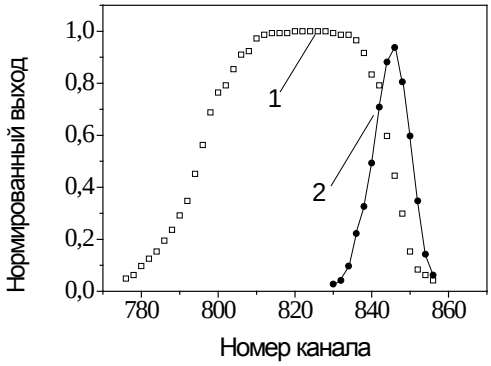


Рис. 11. Энергетический спектр протонов с энергией 240 кэВ, рассеянных пленкой Au

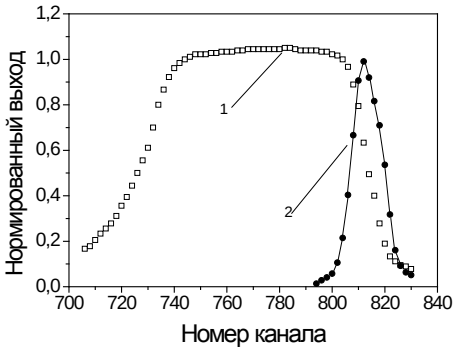


Рис. 12. Энергетический спектр протонов с энергией 240 кэВ, рассеянных пленкой Cr

Второй способ предпочтительнее, т.к. два измерения ведутся при фиксированной (с точностью установки) энергии, и в слагаемые ошибок войдет только одно значение погрешности установки энергии, а не два, как в первом случае. При калибровке анализатора мы использовали второй способ. В качестве мишеней использовали кремниевые пластины с напыленными на них пленками золота и хрома толщиной 200 и 300 ангстрем соответственно.

Изменение энергии пучка дает возможность определить зависимость ширины канала от энергии. На рис. 11 и 12 приведены экспериментальные спектры от пленок золота и хрома на кремнии для энергий пучка ионов водорода 240 кэВ. Такие же спектры были получены для энергий протонов 190, 220, 260 и 280 кэВ. Во всем измеренном диапазоне энергий регистрируемых рассеянных ионов ширина канала постоянна и составляет 281±5 эВ.

Для определения энергетического разрешения всего спектрометрического тракта, включая анализатор ЭСА, воспользуемся рис. 11. Разрешение определяется крутизной спада высокоэнергетической границы спектра, которая соответствует однократному рассеянию падающих на мишень ионов атомами поверхности и приповерхностного слоя толщиной в несколько десятков ангстрем. За его величину принимается ширина пика производной dN/dE на полувысоте. На рис. 11, 12 приведены эти производные. Таким образом, измеренные ширины спектров для пленок Au и Cr соответственно равны: ΔE_{Au}= 3,1 кэВ (1,3 %) и ΔE_{Cr} = 3,9 кэВ (1,6 %). Отличие в ширине спектров обусловлено различным качеством поверхности пленок золота и хрома.

Рис. 13. Измерительный комплекс для элементного анализа твердотельных материалов методом резерфордовского обратного рассеяния ионов с нанометровым разрешением по глубине

Таким образом, разработан и изготовлен измерительный комплекс для элементного анализа твердотельных материалов (рис. 13), позволяющий регистрировать спектры резерфордовского обратного рассеяния с энергетическим разрешением 1,3%, что соответствует минимальной толщине регистрируемого слоя 1,5 нм. В целом. он позволяет вести элементный анализ структурного состояния материалов и приборных структур (включая наноматериалы) с разрешением 1–2 нм[18].

Литература

1. A.R.Chelyadinskii, F.F.Komarov. Defect-impurity engineering in implanted silicon. Physics-Uspekhi, v.46, No.8, pp.789 – 820.

2. К.К.Кадыржанов, Ф.Ф.Комаров, А.Д.Погребняк, В.Ф.Русаков, Т. Э.Туркебаев. Ионно-лучевая и ионно-плазменная модификация материалов. Москва, изд-во МГУ, 2005, 642с.
3. F.Komarov, O.Milchanin Formation of silicon-on-insulator structures with low surface roughness. Proceed. Intern. Conf. “Micro- and nanoelectronics 2005”, Zvenigorod, Russia, October 3 – 7th 2005, pp.02 – 35.
4. Ф.Ф.Комаров, О.В.Мильчанин, А.М.Миронов, А.И.Купчишин. Применение протонных пучков в современной микро- и оптоэлектронике. Электроника, 2006, No.10, с. 30-35
5. E.B.Boyko, A.S.Kamyshan, F.F.Komarov, A.E.Lagutin. Interaction of fast hydrogen ions with silicon surface at glancing angles of incidence. Nicl. Instr. Meth., 2007, B256, No. 2, pp. 359 – 362.
6. Ф.Ф.Комаров, О.В.Мильчанин, А.М.Миронов. Формирование водородно-индуцированных дефектов и их применение в технологии микро- и оптоэлектроники. Вестник БГУ сер. 1, 2006, No. 3, с. 56 – 62.
7. Ф.Ф.Комаров, О.В.Мильчанин, А.М.Миронов, А.И.Купчишин. Формирование структур микро- и оптоэлектроники с использованием протонных пучков. Материалы и структуры современной электроники. 2006, БГУ, с.123 – 133.
8. Ф.Ф.Комаров, О.В.Мильчанин. Развитие технологии формирования структур кремний-на-изоляторе. Доклады НАН Беларуси, 2006, т.50, No. 2, с.41 – 45
9. Ф.Ф.Комаров, О.В.Мильчанин, В.В.Пилько, Ю.Г.Фоков. Формирование протяженных дефектов в кремнии при высокодозной имплантации ионов водорода. Поверхность, 2008, No. 4, с. 27 – 30.
10. Ф.Ф.Комаров, О.В.Мильчанин, В.В.Пилько. Устройство для соединения пластин. Патент РБ на полезную модель, №5197, 2009.01.05.
11. Ф.Ф.Комаров, А.Ф.Комаров, А.М.Миронов. Формирование однородных легированных слоев в металлах и полупроводниках методом полиэнергетической высокодозной ионной имплантации. Доклады НАН Беларуси, 2007, т. 51, No. 3, с. 52 – 56.
12. F.F.Komarov, A.M.Mironov, P.Zukovski. Characteristics of the implant-isolated GaAs layers after annealing. Proceed. Intern. Conf. “Micro- and nanoelectronics 2005”, Zvenigorod, Russia, October 3 – 7th 2005, pp. 03 – 28.
13. Т.Колтунович, П.Жуковски, П.Венгерек, Я.Партыка, Ф.Ф.Комаров, Л.А.Власукова. Температурная стабильность вертикальной изоляции GaAs, созданной имплантацией ионов водорода. Материалы и структуры современной электроники. 2006, БГУ, с.193 – 196.
14. F.F.Komarov, A.M.Mironov, P.Zukowski, T.Koltunowicz. Characteristics of the implant-isolated GaAs layers. Mater. of the 5th Intern. Conf. “New Electrical and Electronic Technologics and Their Industrial Implementation”, Lublin University of Technology, Zakopane, Poland, June 12 – 15, 2007, pp. 126 – 127.
15. A.Didyk, F.Komarov, L.Vlasukova, V.Yuvchenko. Structure changes in InP and GaAs crystals double irradieted with electrons and swift heavy ions. Vacuum, 2007, v.81, No. 10, pp. 1175 – 1179.
16. P.Zukowski, T.Koltunowicz, J.Partyka, P.Wegierek, F.F.Komarov, A.M.Mironov, N.Butkievith, D.Freik. Dielectric properties and model of hopping conductivity of GaAs irradiated by H^+ ions. Vacuum, 2007, v. 81, No. 10, pp. 1137 – 1141.
17. Ф.Ф.Комаров, А.М.Миронов, П.Жуковски. Формирование дефектов структуры в GaAs, облученном быстрыми протонами и их применение для создания межприборной изоляции. Сб. докл. Межд. конф. «Актуальные проблемы физики твердого тела». 2009, Минск, изд. А.Н.Вараскин, т.3, с.79-81.

4. Плазменная модификация структуры и свойств материалов

В.М. Анищик

Белорусский государственный университет, 220050, Минск, Беларусь
e-mail: Anishchik@bsu.by

В последнее время для модификации свойств материалов и изделий широко используются пучки различных излучений – ионные, плазменные, электронные и т.д. Это во многом обусловлено тем, что часто технологические свойства изделия определяются свойствами (и, естественно, составом) поверхности. А поскольку облучение – это поверхностная обработка, то изучение влияния облучения на поверхностные слои материалов представляет не только научный, но и практический интерес.

Ниже будет представлен обзор работ по плазменной модификации поверхности, проводимых на кафедре физики твердого тела физического факультета Белорусского государственного университета.

При обработке поверхности материалов потоками заряженных частиц можно выделить три области воздействия, приводящих к различным эффектам:

1. режим слабого воздействия, при котором происходит нагревание верхнего поверхностного слоя до температуры ниже точки плавления материала. В случае обработки высокоэнергетическими ионными пучками модификация микростуктуры происходит за счет имплантации ионов;
2. режим среднего воздействия, при котором происходит плавление слоя материала и последующее быстрое повторное затвердевание (в основном за счет теплопроводности подложки). Полученная микростуктура обусловлена процессами быстрой закалки из расплава;

3. режим сильного воздействия, при котором температура нагрева может достигать точки кипения расплавленного материала, вследствие чего происходит абляция части вещества с поверхности мишени. Механические ударные волны проникают глубоко в объем образца и создают дефектную структуру.

Все эти эффекты могут быть получены при обработке материалов компрессионными плазменными потоками. Особенностью данных потоков является относительно большое время существования импульса плазмы (≈ 100 мкс), достаточное для протекания физико-химических процессов. Такие потоки получают с помощью квазистационарных плазменных ускорителей с собственным азимутальным магнитным полем, в которых ускорение плазмы сопровождается формированием на выходе разрядного устройства компрессионного потока, параметры плазмы в котором существенно выше, чем в межэлектродном промежутке.

Компрессионные потоки средних энергий получают с помощью квазистационарных плазменных ускорителей типа “магнитоплазменный компрессор” (МПК) [1]. Преимуществом МПК по сравнению с другими типами ускорителей является высокая устойчивость генерируемого компрессионного потока, возможность управления его составом, размерами и параметрами плазмы при достаточной для практического применения длительности разряда.

Высокоэнергетические компрессионные потоки получают в принципиально новых двухступенчатых квазистационарных сильноточных плазменных ускорителях (КСПУ), в которых реализуется ионно-дрейфовое ускорение замагниченной плазмы. Основной принцип работы КСПУ заключается в пространственном разделении зон ионизации и ускорения, переходе на ионный токоперенос, создании анодного и катодного трансформеров (электродов), обеспечивающих магнитную экранировку твердотельных элементов их конструкции и инверсию носителей тока (электронов в подводящей цепи и ионов – в ускорительном канале).

Внешний вид и схема разрядного устройства МПК представлен на рис. 1. Эксперименты проводились в режиме “остаточного газа”, при котором предварительно откачанную камеру МПК заполняли рабочим газом (азотом) до давления 400 Па. Амплитудное значение разрядного тока МПК достигало 70 кА при длительности разряда ~ 80 мкс.

На выходе ускорителя формировался компрессионный плазменный поток диаметром 0,7 см и длиной 10 см, угол полураскрытия плазменной струи составлял 10-15°. Скорость плазменных образований компрессионного потока составляла $(5\text{-}6)\cdot 10^6$ см/с, концентрация электронов плазмы $(4\text{-}7)\cdot 10^{17}$ см$^{-3}$, а ее температура – 2-3 эВ.

По данным калориметрических измерений плотность энергии, поглощаемой поверхностью мишени в зависимости от ее удаления от среза разрядного устройства МПК составляет 5-25 Дж/см^2 за импульс, что в условиях экспериментов соответствует плотности мощности потока в диапазоне $(0{,}6\text{-}3)\cdot 10^5$ Вт/см^2.

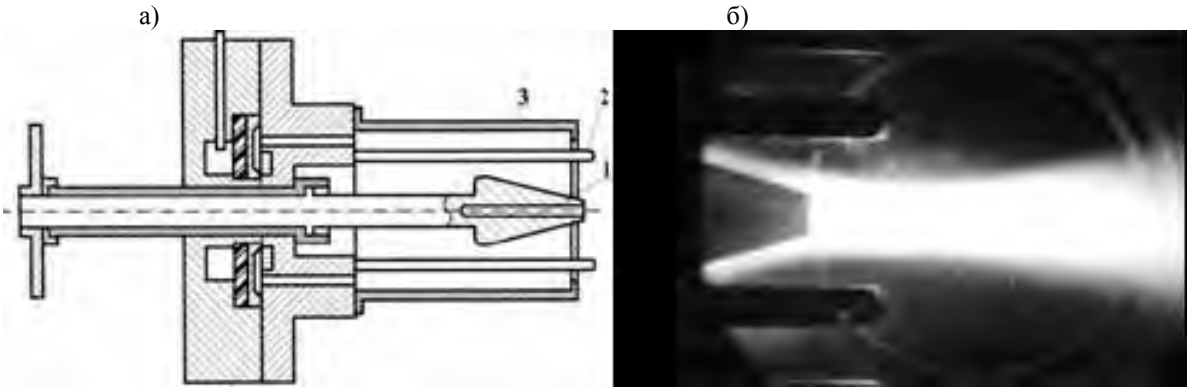


Рис. 1. Схема разрядного устройства МПК:
а) 1-катод, 2-анод, 3-кожух,
б) фотография плазменного пучка (время экспозиции 2 мкс).

Исследование в режиме сильного воздействия проводились на установке КСПУ типа П-50М с суммарной энергией емкостных накопителей 215 кДж. Схема разрядного устройства КСПУ представлена на рис. 2. КСПУ П-50М генерирует плазменные потоки длиной 35 см и диаметром ~ 3 см, угол полураскрытия плазменной струи после выхода из разрядного устройства составлял ~20°. Параметры плазмы в зависимости от сорта рабочего газа и начальных параметров ускорителя изменялись в следующих пределах: концентрация электронов плазмы 10^{16}-10^{18} $см^{-3}$, температура плазмы –5-15 эВ, скорость плазмы (5-20)·10^{7} см/с.

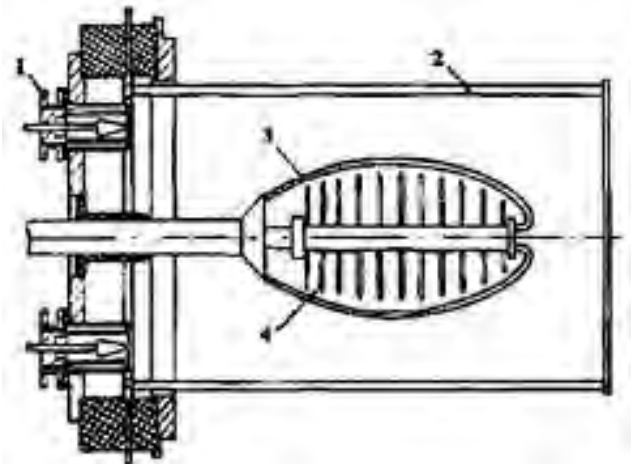


Рис. 2. Схема разрядного устройства КСПУ П-50М:
1-ускоритель первой ступени, 2-анодный трансформер, 3-катодный трансформер,
4-“струйные катоды”.

Микроструктура и топография обработанных поверхностей исследовались с помощью электронного сканирующего микроскопа LEO 1455 VP, элементный состав с помощью рентгеновского энергодисперсионного микроанализатора и Оже-спектрометра PHI-660. Структурное состояние изучалось с помощью рентгеновского дифрактометра. Микротвердость измерялась по методу Виккерса с использованием нагрузки от 0,2 до 2,0 Н. Трибологические исследования проводились по методу

“острие-поверхность” при скорости индентора и нагрузки на него 4 мм/с и 1,0 Н, соответственно.

Исследовались [2, 3] монокристаллические пластинки кремния (10х10х0,28 мм) с ориентацией (111) и (100), которые устанавливались нормально к плазменному пучку на расстоянии 6-16 см от разрядного устройства МПК. Калориметрические исследования показали, что поглощенная поверхностью образца энергия находится в области 5-25 Дж/см^2 (в зависимости от положения образца), что соответствует изменению плотности мощности плазменного потока в диапазоне $(0,5\text{-}3)\cdot10^5$ Вт·см$^{-2}$.

Взаимодействие компрессионного потока на образец приводит к образованию вблизи поверхности ударно-сжатого плазменного слоя. Необходимо отметить, что торможение такого компрессионного потока, в плазму которого вморожено магнитное поле, сопровождается образованием замкнутых токовых петель (вихрей) [4].

Экспериментальные данные показывают, что в результате воздействия плазменного потока глубина модифицированного слоя может достигать 6 мкм. На микрофотографиях поверхности образцов хорошо видны периодические структуры, фрагменты которых имеют цилиндрическую форму (рис. 3, 4). Длина этих фрагментов составляет 50-100 мкм, диаметр - 0,7-1,5 мкм, а поверхностная плотность - $(2\text{-}6)\cdot10^6$ см^2.

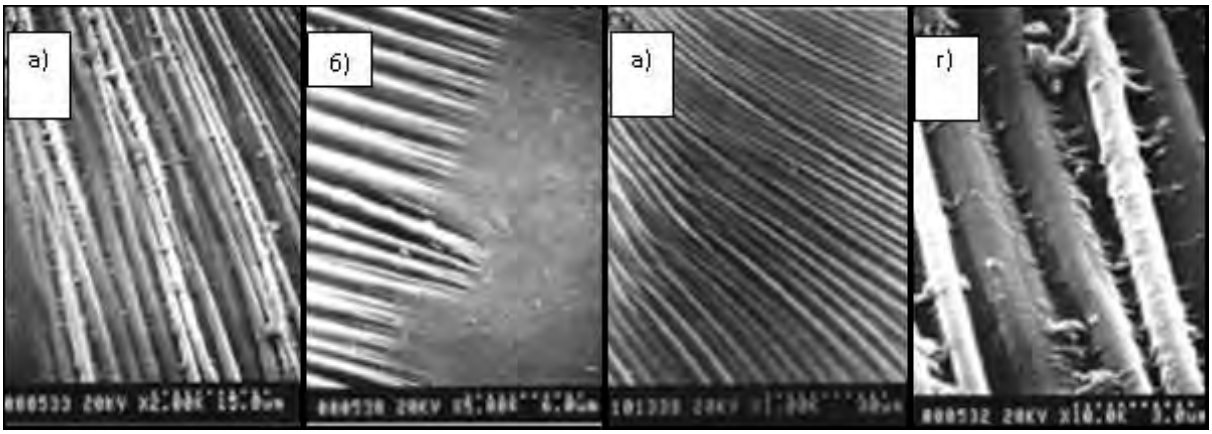


Рис. 3. Морфология поверхности пластины кремния (111)
после обработки плазменным пучком.

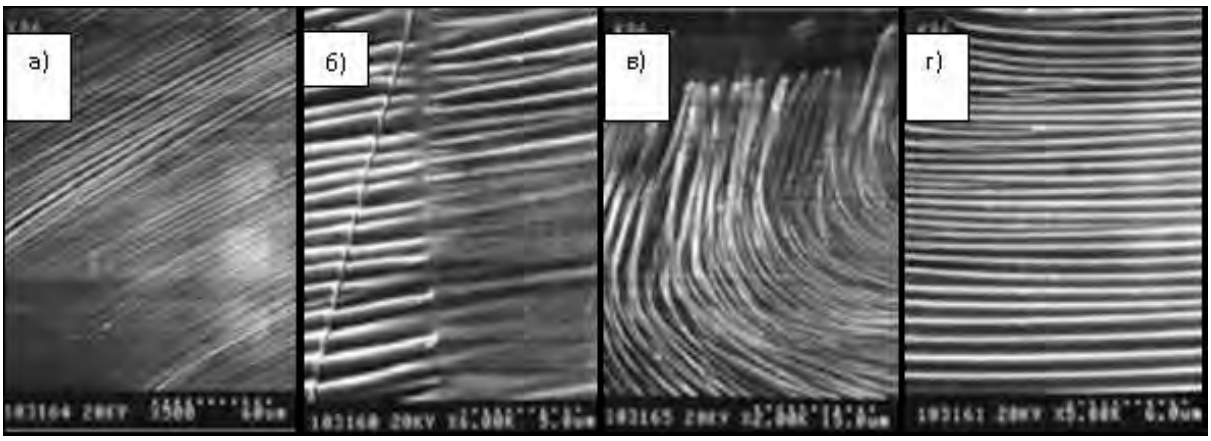


Рис. 4. Морфология поверхности пластины кремния (100)
после обработки плазменным пучком.

Несмотря на схожесть структур, приведенных на рис. 3 и рис.4 можно заметить и некоторые различия, связанные с кристаллографической ориентацией. Так, на пластинах с ориентацией (111) между цилиндрическими образованиями наблюдались «шпоры» (рис. 3г), что может быть связано с влиянием внешних силовых факторов на процессы кристаллизации расплавленного поверхностного слоя. Приведенные на рис. 3а прямолинейные цилиндрические фрагменты имеют диаметр 1 мкм и достигают длины 100 мкм. Наблюдаются и образования другой формы (рис. 3в и 4в), форма которых может быть обусловлена различием в деформации поверхности образца и турбулизацией плазменного пучка на мишени. Рис.3б и 4б показывают границы между структурно упорядоченными и разупорядоченными областями, влияющими на формирование цилиндрических структур. Другими словами, наличие границ предполагает существование значительных градиентов термодинамических параметров по поверхности образца при воздействии плазменного пучка. Следует также отметить наличие переплетающихся цилиндрических фрагментов, рис. 3в и 4в.

Если в плазменный пучок ввести высокодиспергированные металлические частицы, то на поверхности мишени формируется наноструктурированное покрытие [5, 6]. На рис. 5 приведены микрофотографии железоникелевого покрытия на поверхности (100) пластины кремния.

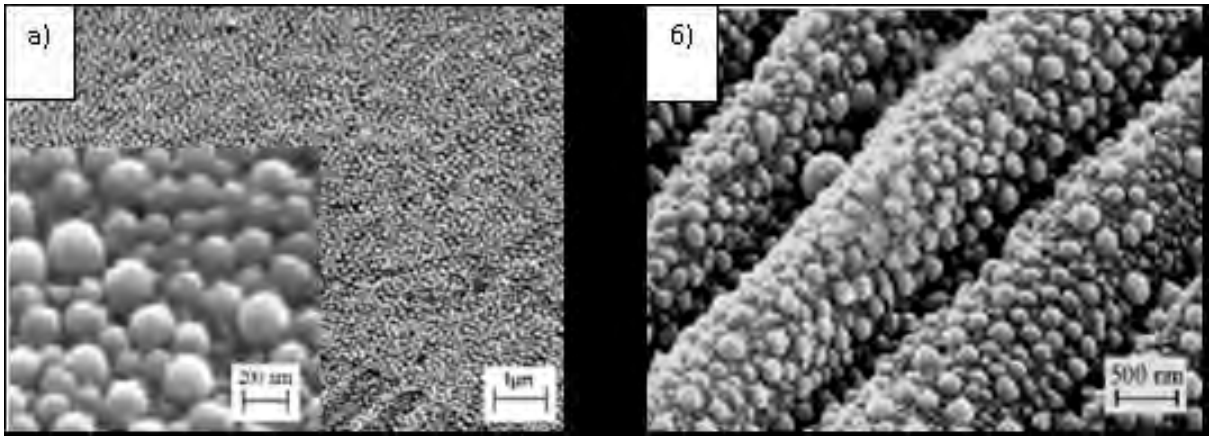


Рис. 5. Микрофотографии покрытия железоникелевого покрытия на поверхности (100) пластины кремния

Покрытие представляет собой однородный наноструктурированный слой из сцепленных между собой сферических частиц размерами от 20 до 300 нм, в свою очередь состоящих из еще более мелких частиц размерами порядка 10-30 нм. Толщина покрытия сравнима с размером частиц и составляет 100-200 нм. Покрытие формируется на всей поверхности мишени, в том числе и на периодических цилиндрических структурах.

В [7] была рассмотрена простая модель, связывающая образование периодических структур с развитием гидродинамической неустойчивости Кельвина-Гельмгольца (НКГ) в системе плазма-расплав. В рамках этой модели сделаны оценки характерных пространственных масштабов формирующейся структуры. Движение расплава и плазмы, а также их взаимодействие описывается с помощью уравнений идеальной магнитной гидродинамики.

Полученные в [7] оценки расстояния между соседними цилиндрами в одном пакете (то есть пространственный период их расположения в пакете) удовлетворительно согласуется с экспериментальными результатами [2, 3, 8]. Однако

из идеальной (не учитывающей диссипацию) модели [7] следует, что на границе плазма-расплав всегда присутствует неустойчивость Кельвина-Гельмгольца.

Рассмотрим эту проблему (НКГ) с учетом вязкости, электропроводности и конечной глубины слоя расплава [9].

Свойства жидкого кремния близки к свойствам типичных металлов [10]. Это касается, в частности, кинетических коэффициентов. Следовательно, течение кремниевого расплава можно, как и течение металла, описать с помощью уравнений магнитной гидродинамики [11, 12]:

$$\rho\left[\partial_t \vec{v} + \left(\vec{v}\cdot\vec{\nabla}\right)\vec{v}\right] = \partial_j\left[\eta\left(\partial_j \vec{v} + \vec{\nabla} v_j\right)\right] - \vec{\nabla}\left(p + \frac{\vec{B}\cdot\vec{H}}{2}\right) + \left(\vec{H}\cdot\vec{\nabla}\right)\vec{B} \qquad (1)$$

$$\operatorname{div}\vec{v} = 0 \qquad (2)$$

$$\partial_t \vec{B} = \operatorname{rot}\left(\vec{v}\times\vec{B} - \nu \operatorname{rot}\vec{B}\right) \qquad (3)$$

$$\operatorname{div}\vec{B} = 0 \qquad (4)$$

где $\vec{v}$ - скорость течения среды; p - давление; ρ - плотность; $\vec{B} = \mu_0 \vec{H}$ - индукция магнитного поля; $\vec{H}$ - напряженность магнитного поля; μ_0 - магнитная проницаемость вакуума; $\nu = \dfrac{1}{\mu_0 \sigma}$ - магнитная вязкость; σ - удельная проводимость; η - динамическая вязкость, $\partial_t = \dfrac{\partial}{\partial t}$ - производная по времени; $\vec{\nabla}$ - оператор градиента; v_j и ∂_j – компоненты векторов $\vec{v}$ и $\vec{\nabla}$ соответственно. По повторяющемуся индексу подразумевается суммирование.

Уравнения (1)-(4) используются и для описания динамики плазмы (под плазмой в данном случае понимается не вся компрессионная плазма, а плазма, находящаяся вблизи поверхности расплава). Следуя [7], предполагается, что в (1) не учитывается сила тяжести, любая среда считается несжимаемой, а ее характеристики (плотность, вязкость и т.п.) и поверхностное натяжение на границе раздела не зависят от температуры. В этом случае система (1)-(4) становится замкнутой при условии, что характеристики любой среды известны. Наложение начальных и граничных условий завершает общую математическую формулировку задачи. Считается, что твердая часть мишени является непроницаемой для расплава.

В декартовой системе координат, плоскость $X0Y$ которой совпадает с плоской границей раздела плазма-расплав, исследуется устойчивость стационарного состояния, при котором вектор скорости течения обеих сред $\vec{V}$ параллелен плоскости $X0Y$, а значения их плотности ρ, давления p_0, вязкости η, магнитной вязкости ν и инду кции магнитного поля $\vec{B}$ зависят только от координаты z.

Магнитное поле также параллельно плоскости $X0Y$. Условие, что все эти параметры зависят только от одной координаты, не приводит к существенному ограничению общности задачи, если их изменения на расстояниях порядка нескольких периодов поверхностных структур вдоль плоскости $X0Y$ (то есть нескольких микрометров) достаточно малы.

Данное стационарное состояние должно удовлетворять уравнениям (1)-(4), поэтому все рас- смотренные параметры в общем случае связаны между собой. Далее

на стационарное состояние накладываются малые возмущения скорости $\vec{u}$, давления p и магнитного поля $\vec{b}$. Граница раздела в стационарном состоянии описывается уравнением z=0. Возмущенная (модулированная) граница имеет вид $z=\xi(x,y,t)$, плазме соответствует область $z>\xi$, расплаву — $h(t)<z<\xi$. Функция $h(t)\geq 0$ задает глубину расплава. Сохраняя в (1)-(4) только члены не выше первого порядка по малым возмущениям и их производным, можно записать:

$$\rho\left(\partial_t + \vec{V}\cdot\vec{\nabla}\right)\vec{u} = \partial_j\left[\eta\left(\partial_j\vec{u} + \vec{\nabla}u_j\right)\right] - \rho\left(\vec{u}\cdot\vec{\nabla}\right)\vec{V} - \vec{\nabla}\left(p + \vec{b}\cdot\vec{H}\right) + \left(\vec{H}\cdot\vec{\nabla}\right)\vec{b} + \left(\vec{b}\cdot\vec{\nabla}\right)\vec{H} + \vec{S} \quad (5)$$

$$\operatorname{div}\vec{u} = 0 \quad (6)$$

$$\left(\partial_t + \vec{V}\cdot\vec{\nabla}\right)\vec{b} = \left(\vec{b}\cdot\vec{\nabla}\right)\vec{B} - \left(\vec{u}\cdot\vec{\nabla}\right)\vec{B} + \left(\vec{B}\cdot\vec{\nabla}\right)\vec{u} - \operatorname{rot}\left(\nu\operatorname{rot}\vec{b}\right) \quad (7)$$

$$\operatorname{div}\vec{b} = 0 \quad (8)$$

Добавочный член $\vec{S} = \vec{n}\,\alpha\,\delta(z)\Delta\xi$ в (5) учитывает поверхностное натяжение α на границе раздела сред; $\delta(z)$- дельта-функция; Δ - оператор Лапласа, $\vec{n}$ - единичный вектор вдоль оси $0Z$. В таком случае на границе плазма-расплав (z=0) должно выполняться условие $\left(\partial_t + \vec{V}\cdot\vec{\nabla}\right)\xi = \vec{n}\cdot\vec{u}$, а на границе расплава с твердой частью мишени $z=-h(t)$ и $\vec{u}$=0. Недостающее граничное условие легко получить, интегрируя (5) вдоль нормали к границе раздела плазма-расплав [10].

Дальнейшие математические преобразования аналогичны [7]. Сначала в (5)-(8) и граничных условиях проводится фурье-преобразование всех возмущений по x и по y. При этом вводится новый параметр $\vec{k}$ - волновой вектор, лежащий в плоскости $X0Y$. Для простоты все преобразованные величины будут обозначаться теми же самыми символам. Проекция преобразованного уравнения (5) на ось $0Z$ приводит к дифференциальному уравнению относительно u_z.

Возможны два случая. Если $\vec{k}\cdot\vec{B}(z)\neq 0$, из полученного уравнения можно исключить u_x, u_y, b_x, b_y, а u_z выразить через b_z. Обозначая $\psi = \left[\vec{k}\cdot\vec{B}(z)\right]^{-1} b_z$, это уравнение можно записать

$$\left(\mathrm{L}_V\Delta_\nu + \mathrm{L}_B\right)\boldsymbol{\psi} = i\,k^4\,\alpha\,\delta(z)\xi(t) \quad (9)$$

В (9) введены линейные дифференциальные операторы

$$L_z = \partial_z^2 - k^2,\ \mathrm{D}_z = \partial_t - i\,\vec{k}\cdot\vec{V}(z),$$

$$\Delta_\eta = \rho(z)D_z - \eta'(z)\partial_z - \eta(z)L_z,$$

$$L_\eta = \partial_z\Delta_\eta\partial_z - k^2\Delta_\eta,\ \Delta_\nu = \left[\vec{k}\cdot\vec{B}(z)\right]^{-1}\left[D_z - \nu(z)L_z\right]\left[\vec{k}\cdot\vec{B}(z)\right],$$

$$\mathrm{L}_V = L_{\boldsymbol{\eta}} - k^2\,\eta''(z) - i\,\partial_z\rho(z)\left[\vec{k}\cdot\vec{V}'(z)\right],$$

$$L_B = \partial_z \left[\vec{k}\cdot\vec{H}(z)\right]\left[\vec{k}\cdot\vec{B}(z)\right]\partial_z - k^2\left[\vec{k}\cdot\vec{H}(z)\right]\left[\vec{k}\cdot\vec{B}(z)\right].$$

Штрихи над функциями обозначают производные по *z*. Граничным условием при *z*=0 служит $D_z\xi(t) + i\Delta_\nu\psi = 0$, при *z*=–*h*(*t*) в нуль обращается $\Delta_\nu\psi$. Еще одно условие при *z*=0 получится, если проинтегрировать (9) по *z* в окрестности *z*=0.

Если $\vec{k}\cdot\vec{B}(z) = 0$, аналогичным образом получается уравнение для u_z:

$$\mathrm{L}_V u_z + i\,\partial_z\left[\vec{k}\cdot\vec{H}'(z)\right] b_z = \alpha\, k^4\,\delta(z)\xi(t) \qquad (10)$$

где магнитное поле b_z является решением

$$\left[\hat{D}_z - \nu(z)\hat{L}_z\right] b_z = 0.$$

Уравнения (9), (10) громоздки и содержат частные производные. Следует, также, не забывать, что решаются эти уравнения для двух областей пространства с изменяющейся геометрией, при этом необходимо различать случаи $\vec{k}\cdot\vec{B}(z) = 0$ и $\vec{k}\cdot\vec{B}(z) \neq 0$. Все это существенно затрудняет расчеты, поэтому, следуя [7], будем полагать что $\vec{H}$, $\vec{B}$, ρ, $\vec{V}$, η и ν постоянны в каждой из областей, где *z*>0 или –*h*(*t*)<*z*<0. Заметим, что случай $\vec{k}\cdot\vec{B}(z) = 0$ рассматривать в таком приближении не имеет смысла, так как при этом из уравнения (10) исчезает связь скорости и магнитного поля.

При $\vec{k}\cdot\vec{B}(z) \neq 0$ можно получить уравнение, описывающее изменение во времени фурье-амплитуды деформации поверхности расплава:

$$\left[\alpha\,\mathrm{k}^3 + I_+ + I_- D_-^{-1}\coth(kh)\,D_-\right]\xi(t) = J(t), \qquad (11)$$

где $\mathrm{I}_\pm = \left(\vec{k}\cdot\vec{H}_\pm\right)\left(\vec{k}\cdot\vec{B}_\pm\right) + \left(\rho_\pm D_\pm + k^2\eta_\pm\right)\left(D_\pm + k^2\nu_\pm\right)$, $\mathrm{D}_\pm = \dfrac{d}{dt} + i\,\vec{k}\cdot\vec{V}_\pm$, а D_-^{-1} - означает оператор, обратный оператору D_-. Знаками “+” и “–” отмечены величины, относящиеся к плазме и расплаву соответственно. Неоднородная часть *J*(*t*) сложным образом зависит от времени и от вида функции *h*(*t*) и функций, определяющих начальные условия для η, но не зависит от ξ(*t*). Поэтому (11) является линейным неоднородным дифференциальным уравнением относительно ξ(*t*).

Для исследования устойчивости необходимо проанализировать спектр соответствующего однородного уравнения. В случае, когда coth(*kh*) меняется медленно, его можно считать постоянной величиной, и однородная часть (11) будет дифференциальным уравнением с постоянными коэффициентами,

решение ξ(*t*) которого имеет вид $\xi_0\exp(\mathrm{i}\,\omega t)$. Таким образом, получается дисперсионное соотношение

$$\rho_0\omega^2 + c_1\omega - c_0(t). \qquad (12)$$

Коэффициенты в уравнении (12) сложным образом зависят от волнового вектора *k* и от *h*:

$$\rho_0 = \rho_+ + \rho_- \coth(kh),$$

$$c_1 = 2\rho_0\left(\vec{k}\cdot\vec{V}_0\right) - i\,k^2\eta_0,$$

$$c_0 = \left[\eta_+\nu_+ + \eta_-\nu_- \coth(kh)\right]k^4 + \alpha\,k^3 + \left(\vec{k}\cdot\vec{H}_+\right)\left(\vec{k}\cdot\vec{B}_+\right) +$$
$$+\left(\vec{k}\cdot\vec{H}_-\right)\left(\vec{k}\cdot\vec{B}_-\right)\coth(kh) - \rho\left[\vec{k}\cdot\left(\vec{V}_+ - \vec{V}_-\right)\right]^2 - \rho_0\left(\vec{k}\cdot\vec{V}_0\right)^2 + i\,k^2\eta_0\left(\vec{k}\cdot\vec{V}_1\right)$$

$$\vec{V}_0 = \rho_0^{-1}\left[\rho_+\vec{V}_+ + \rho_-\vec{V}_- \coth(kh)\right],$$

$$\eta_0 = \eta_+ + \rho_+\nu_+ + \left(\eta_- + \rho_-\nu_-\right)\coth(kh),$$

$$\vec{V}_1 = \eta_0^{-1}\left[\left(\eta_+ + \rho_+\nu_+\right)\vec{V}_+ + \left(\eta_- + \rho_-\nu_-\right)\vec{V}_- \coth(kh)\right],$$

$$\rho = \rho_0^{-1}\rho_+\rho_- \coth(kh).$$

Квазистационарный спектр (12) дает всю информацию, необходимую для анализа устойчивости системы плазма-расплав. Если *h* меняется, то в (12) можно подставлять ее мгновенные значения. Такой подход оправдан до тех пор, пока малы производные coth(*kh*), что хорошо выполняется для больших *h*, на практике — для *kh*>0,5.

Решая квадратное уравнение (12) относительно ω, можно получить две ветви комплексных частот. Решение (11) искали в виде $\xi_0\exp(i\omega t)$, поэтому наличие положительной мнимой части у частоты приводит к экспоненциальному затуханию возмущения. Возмущения, частоты которых имеют отрицательные мнимые части, экспоненциально растут, и в этом случае возникает НКГ.

Те возмущения, у которых частоты вещественны, носят чисто колебательный характер и называются нейтральными. Как и в классическом случае [10], оказывается, что всегда присутствуют растущие возмущения, поэтому НКГ всегда проявляется и не может быть подавлена.

Анализ уравнения (12) показывает, что на плоскости волновых векторов нейтральным возмущениям отвечает замкнутая кривая, примыкающая к началу координат. Уравнение этой кривой можно получить из решения (12) с условием Imω=0. Все точки (*kx*,*ky*), изображающие нарастающие возмущения, охватываются этой кривой, а все затухающие возмущения располагаются снаружи. Соответственно, нейтральная мода с минимальной длиной волны является характеристической: более короткие волны обязательно затухают, а среди более длинных волн обязательно найдутся неустойчивые.

Расчетная зависимость характеристической длины волны от глубины расплава кремния приведена на рис. 6. При расчетах использовались значения параметров для кремния из [9], вязкость и удельное сопротивление плазмы считались пренебрежимо малыми, также малой по сравнению со скоростью течения плазмы принималась начальная скорость течения расплава, магнитное поле было направлено перпендикулярно скорости течения плазмы. Параметры плазмы взяты из [2, 3, 8].

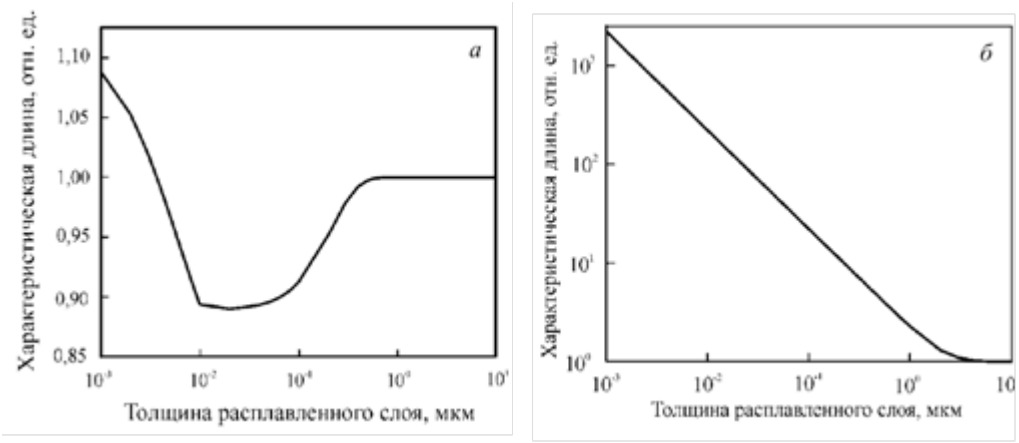


Рис. 6. Зависимость характеристической длины волны от толщины расплавленного слоя кремния. За единицу измерения принята длина волны при бесконечной толщине расплавленного слоя λ:
а - идеальный случай, λ_0=0,2 мкм; б - диссипативный случай, λ_0=6,4 мкм.

Учет диссипативных эффектов меняет спектр малых возмущений поверхности. Из рис. 6 легко видеть, что в идеальном случае (вязкость и омическое сопротивление жидкого кремния пренебрежимо малы) характеристическая длина меняется незначительно (на 10-20%), в то время как в диссипативном случае при уменьшении глубины характеристическая длина резко возрастает на несколько порядков. Это можно интерпретировать как “вытеснение” НКГ в длинноволновую область, так как нейтральные и растущие возмущения уже не проявляются на размерных масштабах 1-100 мкм.

Поэтому в рамках принятой модели необходимым для наблюдения исследуемого явления условием является проплавление мишени на достаточную глубину (0,1 мкм и более). При выполнении этого условия наименьшая длина волны неустойчивой моды уменьшается до 6,4 мкм (рис. 6б). Согласно [2, 3, 8], период расположения цилиндров в пакете составляет 1-2 мкм. Заметная разница между этими величинами может объясняться существенной ролью нелинейных эффектов при формировании цилиндров из гармонического профиля волны на поздних стадиях процесса или влиянием градиентов физических параметров плазмы и расплава.

Механизм формирования регулярных структур, связанный с возникновением НКГ является достаточно универсальным. Действительно, возможность возникновения НКГ при заданном режиме работы МПК определяется глубиной проплавления мишени и свойствами расплава. Для многих металлов эти показатели достаточно близки к показателям жидкого кремния. На рис. 7 представлены результаты наших экспериментов по воздействию КП на моно- и поликристаллический алюминий. Сравнивая рис. 7 и рис. 3, легко обнаружить сходную “пакетную” структуру модифицированной поверхности с близкими характерными размерами образований, составляющих пакет.

Существенным является отличие формы образований на поверхности алюминия от цилиндрической, характерной для кремния. Это различие может возникать на более поздней, нелинейной стадии их формирования и не может быть описано в рамках рассмотренной линейной модели.

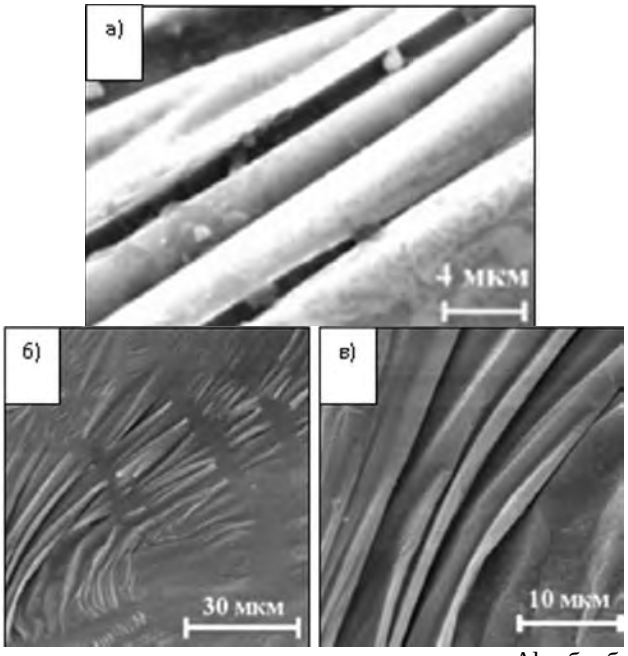


Рис. 7. РЭМ-изображение поверхности моно- и поликристалла Al, обработанной компрессионной плазмой:
а - цилиндрические структуры на поверхности монокристалла;
б - характерные филаментарные структуры на поверхности поликристалла, объединенные в “пакеты”;
в - структура одного из пакетов, увеличенный фрагмент поверхности (б).

Как было показано выше, действие компрессионного плазменного пучка вызывает появление на поверхности кремниевой пластины характерных цилиндрических образований, а в случае введения в плазму дисперсных металлических порошков – формирование на поверхности (и на цилиндрических структурах) наноструктурированных покрытий. Представляло интерес выяснить, что произойдет при воздействии плазменного пучка на систему “металлическая пленка – полупроводниковая пластина” [13].

Исследуемые образцы представляли собой пластину кремния (100) с нанесенной на нее пленкой железа (толщина 1,2 мкм). Скорость плазменного пучка, его температура и электронов были $5 \cdot 10^6$ см$^{-1}$, 2 эВ и $3 \cdot 10^{17}$ см$^{-3}$ соответственно. Плотность энергии, поглощенной мишенью варьировалась в интервале 0,6 – 1,4 ГВт·м^2.

Исследования, проведенные с помощью растрового электронного микроскопа показали, что в результате взаимодействия плазменного пучка с мишенью образуется модифицированный слой толщиной от 4 до 25 мкм, причем железо и кремний распределены по этому слою равномерно. Образование такого слоя обусловлено плавлением железа и углерода, их перемешиванием и кристаллизацией, при этом он имеет дендритную структуру (рис. 8). Обнаружено, что в форме дендритов кристаллизуется кремний, а железо преимущественно находится в междендритном пространстве.

Формирование дендритной структуры обусловлено концентрационным переохлаждением. Поскольку растворимость железа в кремнии в твердом состоянии очень

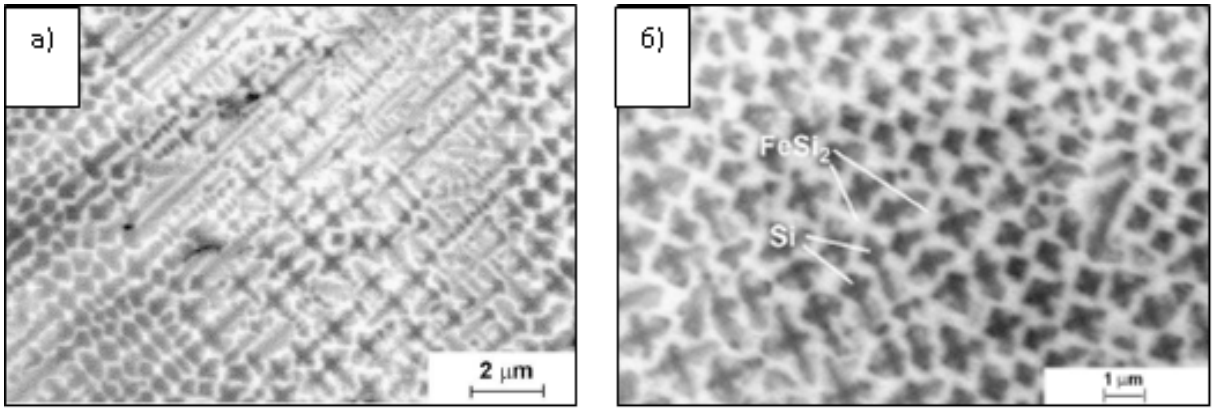


Рис. 8. Поверхность системы железо-кремний после обработки плазменным пучком с различной плотностью энергии:
а - 0,7 ГВт/м^2, б - 1,3 ГВт/м^2.

мала, то в неравновесных условиях атомы железа скапливаются в жидкости перед фронтом кристаллизации, в результате чего возникает градиент концентрации, приводящий в данном случае к возникновению концентрационного переохлаждения. Поскольку охлаждение и кристаллизация протекают при высоких скоростях ~10^7 К/с формирование дендритов можно описать моделью Курца-Фишера [14]. Модель предполагает следующие соотношения между геометрическими размерами дендрита и параметрами процесса кристаллизации.

$$R = 2\pi\sqrt{\frac{\Gamma D}{m(k-1)C_0 V}} \tag{13}$$

$$\lambda_1 = 4{,}3\sqrt[4]{\frac{m(k-1)\Gamma D C_0}{VG^2}} \tag{14}$$

где R – радиус закругления, $\Gamma = \gamma_{sl} / \Delta S_f$ - коэффициент Гиббса-Томсона ($6{,}0\cdot10^{-7}$ К/м), γ_{sl} - коэффициент поверхностного напряжения расплава (1,62 Н/м), ΔS_f - энтропия кристаллизации на единицу объема ($2{,}7\cdot10^6$ Дж/(К·м^3)), D – коэффициент диффузии в жидком состоянии ($10{,}8\cdot10^{-9}$ м^2/с), m – наклон линии ликвидуса, k – коэффициент разделения, C_0 – средняя концентрация атомов железа, V – скорость роста дендрита, λ_1 - расстояние между дендритами, G – температурный градиент вблизи фронта кристаллизации.

Проведенные расчеты показали, что скорость роста дендрита примерно 0,04 м/с и температурный градиент вблизи фронта кристаллизации уменьшается с ростом плотности энергии от $1{,}73\cdot10^8$ до $0{,}08\cdot10^8$ К/м (табл.1).

Таблица 1.
Расстояния между дендритами и температурный градиент в расплаве для различных плотностей энергии

W, ГВт/м^2	**λ_1, мкм**	**G, 10^8К/м**
0,6	0,6	1,73
0,7	0,9	1,15
1,1	1,3	0,37
1,3	1,7	0,11
1,4	2,0	0,08

Локализация атомов железа в междендритном пространстве обусловлена объемной диффузией, что увеличивает вероятность образования силицидов. Данные рентгеноструктурных исследований (рис. 9) свидетельствует о формировании высокотемпературного дисилицида железа α-$FeSi_2$. При плотности энергии 0,6 ГВт/м^2 наблюдаются рефлексы как железа, так и дисилицида, а при больших плотностях энергии на дифрактограмме присутствуют только рефлексы дисилицида.

Как показали проведенные исследования [15-21], воздействие компрессионных плазменных потоков приводит к существенной модификации структуры и свойств поверхности металлов.

В процессе обработки происходит внедрение азота в решетку α-Fe [15]. Так как температура плазмы у обрабатываемой поверхности составляет ~(3-5)-10^4 K, то насыщение легирующим элементом происходит в e-области. По данным оже-электронной спектроскопии, в приповерхностных слоях обработанных образцов содержится до 20 ат.% азота, что соответствует эвтектоидному составу. Наблюдающееся уменьшение поверхностной концентрации азота с увеличением энергии обусловлено дополнительным распылением формирующихся слоев набегающим плазменным потоком. Неравномерное распределение легирующего элемента по глубине приводит к появлению как α'-Fe фазы (концентрация азота <8 ат.%), так и g-Fe фазы (концентрация азота >8 ат.%).

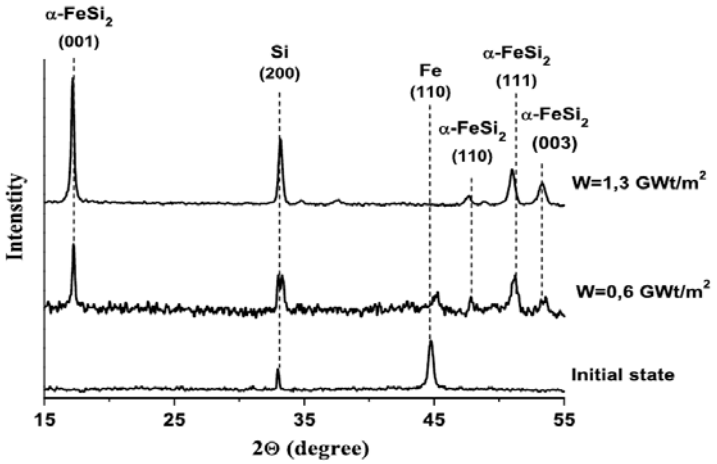


Рис. 9. Участки рентгенограмм поперечного сечения образца.

С помощью растровой электронной микроскопии на поверхности всех обработанных образцов обнаружена ячеистая структура, представляющая собой полигональные ячейки со средним размером 400 нм, которая формируется при затвердевании из концентрационно-переохлажденного расплава. Согласно общепринятым моделям [21, 22], поверхностная неоднородность расплава вызывает боковую диффузию примесей, вследствие чего их концентрация вблизи неоднородности уменьшается, а равновесная температура плавления, соответственно, возрастает. При определенной степени концентрационного переохлаждения образовавшаяся ячеистая структура становится устойчивой.

Конечная микроструктура модифицированного слоя, формирующегося в результате быстрого охлаждения, представлена на рис. 10, 11. Модифицированный слой состоит из двух зон, толщины которых увеличиваются с увеличением энергии набегающего потока (рис. 10). При воздействии серией импульсов толщина этих зон остается постоянной, однако дисперсность составляющих их компонентов меняется (рис. 11). С уменьшением размера зерен возрастает сопротивление пластической

деформации, что проявляется в увеличении микротвердости и повышении износостойкости.

Каждая из модифицированных зон имеет примерно одинаковый набор фазовых составляющих, но их соотношение в разных зонах различно. Было установлено, что эти фазы представляют собой азотистые аустенит и мартенсит, а также нитрид переменного состава e-Fe_2+$_x$N (0<x<1).

Общая схема модифицированного слоя представлена на рис. 12. Основной фазовой составляющей первой зоны толщиной ~6-7 мкм является высокодисперсный мартенсит твердостью ~7 ГПа с включениями остаточного азотистого аустенита. С ростом энергии потока толщина данной зоны увеличивается более чем в 2 раза, а однородность мартенсита повышается. Вторая зона размером ~12-15 мкм состоит из фаз, имеющих столбчатое строение

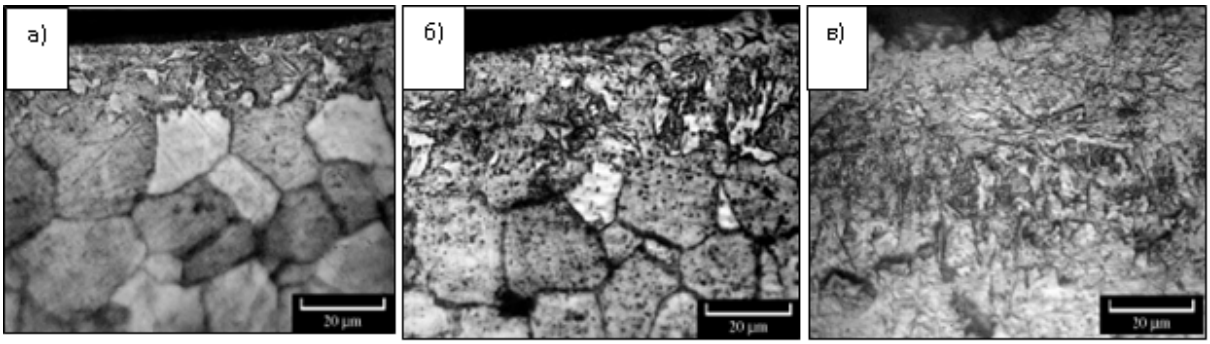


Рис. 10. Микроструктура модифицированного слоя образцов железа, обработанных плазменными потоками при плотности энергии: а — 7 Дж/см^2; б — 12 Дж/см^2; в — 17 Дж/см^2.

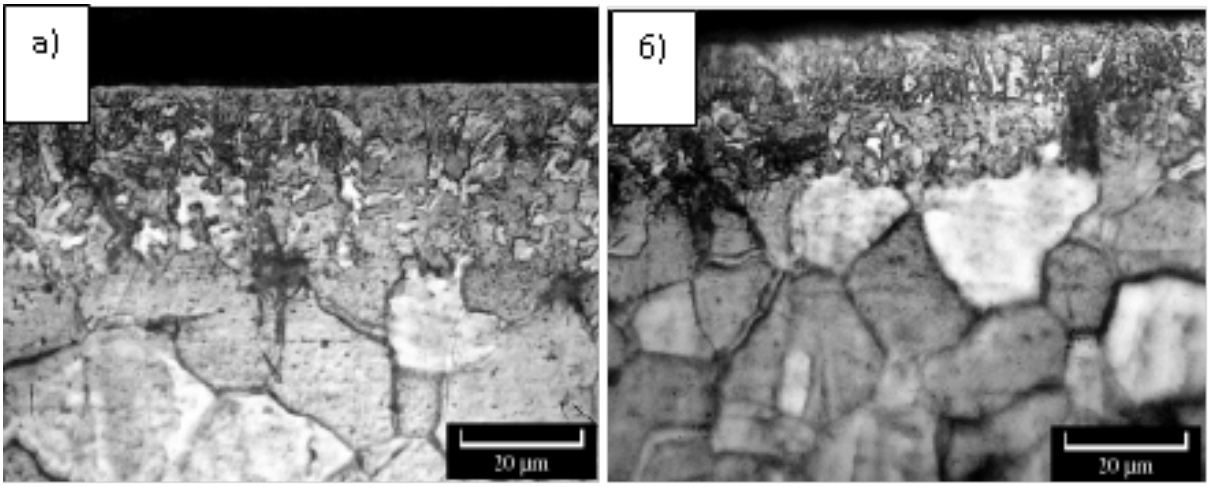


Рис. 11. Микроструктура модифицированного слоя образцов железа, обработанных серией импульсов плазменного потока: а - 5 импульсов; б - 10 импульсов.

и твердость ~4 ГПа, что хорошо согласуется с литературными данными о формировании под поверхностным слоем столбчатых нитридных фаз [23]. С увеличением количества импульсов воздействия повышается ориентированность зерен в направлении теплоотвода. Состав этих фаз идентифицировать достаточно сложно (может варьироваться от Z-Fe_2N до g-Fe_4N). Толщина этой зоны явным образом не

зависит от энергии плазменного потока. В более глубоких слоях образца наблюдается переходная зона толщиной несколько десятков микрон, отличающаяся от структуры исходного материала незначительным присутствием всех вышеперечисленных фаз.

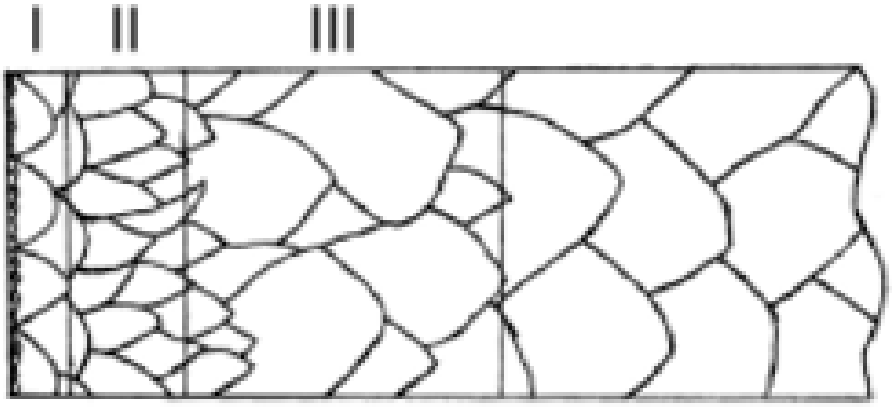


Рис. 12. Схема микроструктуры модифицированного слоя железа, обработанного плазменным потоком:
I — зона равноосных кристаллов; II — зона столбчатых кристаллов; III — переходная зона.

Механизм образования микроструктуры при воздействии компрессионного плазменного потока, по-видимому, подобен механизму затвердевания из жидкой фазы. В течение первых 100 мкс обработки происходит плавление поверхностного слоя за счет термализации кинетической энергии плазменного потока при его торможении на мишени, после чего начинается быстрое охлаждение расплава за счет теплоотвода в объем образца. По мере возрастания степени переохлаждения на границе жидкость-твердое тело начинают образовываться зародыши кристаллизации, причем на процесс зародыше-образования оказывает катализирующее действие внедряемая примесь.

Образовавшиеся кристаллы приобретают столбчатую форму, так как частота зарождения новых зерен в расплаве перед движущимся фронтом кристаллизации, как правило, недостаточна для препятствования росту первоначальных кристаллов. Чем больше число частиц, вызывающих зародышеобразование, тем меньше будет размер столбчатых кристаллов. По мере повышения скорости образования и роста новых центров кристаллизации происходит увеличение количества теплоты кристаллизации, выделяемой этими кристаллитами, что приводит к замедлению движения фронта роста столбчатых кристаллов и образованию у поверхности зоны равноосных кристаллов.

Данные рентгеноструктурного анализа (рис. 13, а-в) свидетельствуют, что импульсная плазменная обработка обеих сталей приводит к формированию аустенитной фазы и частичному или полному растворению упрочняющих карбидов. На рентгенограмме обработанной стали Р6М5 (рис. 13а, б) наблюдается появление дифракционных линий g-Fe (111) и g-Fe(200). В стали Х12М происходит практически полный переход мартенсита в аустенит (рис. 13в).

Повышенное содержание аустенита (до 85%) при плазменной обработке ранее наблюдалось в хромистых сталях [24] и связано, возможно, с распадом карбидов хрома. Захват азота атомами хрома затрудняют его диффузию вглубь образца и обеспечивает повышенную концентрацию азота в поверхностном слое, что положительно сказывается на стабилизации аустенита [25]. Кроме того, атомы хрома могут легировать аустенит, что, возможно, затрудняет или останавливает аустенит-мартенситное превращение при охлаждении.

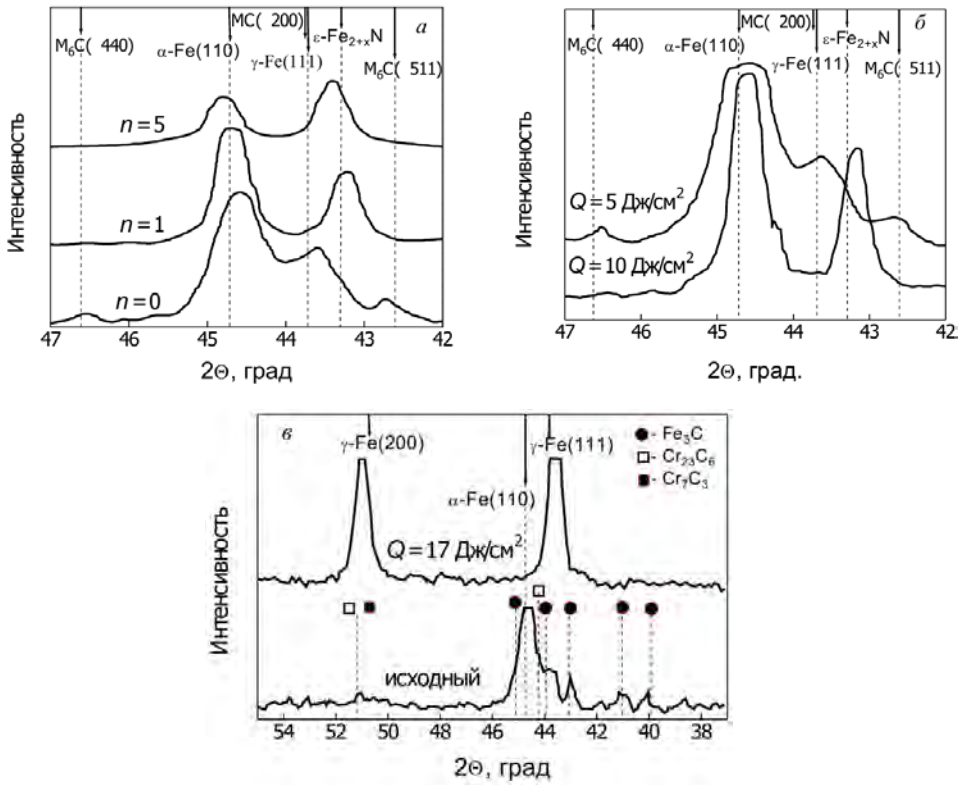


Рис. 13. Участки рентгенограмм стали Р6М5 (а, б) и Х12М (в)
в исходном состоянии и после импульсного воздействия азотной плазмы.
Q — плотность энергии плазмы в импульсе; n — число импульсов.
а - Q=13 Дж/см^2; б - n=1; в - Q=17 Дж/см2, n=1.

Для стали Р6М5 при плотности энергии плазменного импульса выше 5 Дж/см^2 наблюдается растворение карбидных фаз M_6C=$(Fe,Cr)_3(W, Mo)_3C$ (рис. 13а, б), однако совпадение рефлексов VC(200) и g-Fe(111) не позволяет однозначно говорить о растворении карбидов VC, которые, как это следует из литературных данных [25], практически нерастворимы в сталях, содержащих сложные карбиды M_6C. В то же время, в стали Х12М происходит распад карбидов M_3C, M_7C_3 и $M_{23}C_6$ (рис. 13в). Так как температура распада этих карбидов известна и составляет 1000-1300°С [25], то можно предположить, что температуры поверхностного слоя образцов сталей при плазменной обработке достигает таких значений.

В стали Р6М5, в отличие от Х12М, наблюдается значительное смещение рефлексов, соответствующих аустениту, от табличного значения в сторону низших углов, что указывает на расширение решетки вследствие введения в решетку атомов азота и подтверждается данными ОЭС. С увеличением количества импульсов параметр решетки аустенита уменьшается до табличного значения (рис. 13а), при этом параметр решетки мартенсита также уменьшается. Увеличение количества импульсов приводит также к изменению соотношения интенсивностей линий g-Fe(111) и a-Fe(110) за счет увеличения концентрации аустенита в зондируемом рентгеновскими лучами слое, что ранее наблюдалось в углеродистых сталях [24].

Полученные данные показывают, что при плазменной обработке происходит легирование аустенита и мартенсита атомами азота и, кроме того, возможно формирование карбонитридов $M_6(C,N)$ и M(C,N). При растворении карбидов в слое материала, где происходит наиболее сильный распад вторых фаз, будет происходить замена атомов железа в решетке аустенита атомами других металлов. Высокая концентрация азота в поверхностном слое способствует формированию нитридов

(карбонитридов) e-Fe_2+_xN ($0<x<1$) и y'-Fe_4N. Хотя данные РСА и не подтверждают образование нитридов, однако это может быть связано не с их отсутствием, а с неоднородностью их распределения по глубине, высокой дисперсностью и перекрытием рефлексов е-нитрида переменного состава и аустенита. Для обнаружения данных фаз необходимо использование просвечивающей электронной микроскопии.

Плазменная обработка приводит к существенной модификации поверхностных областей, при этом из-за разной температуры нагрева материала на разной глубине зона воздействия имеет слоистое строение. В стали Р6М5 при плотности энергии выше 5 Дж/см^2 первый слой толщиной ~6 мкм содержит аустенит и карбид VC, а второй переходной слой толщиной также около 6 мкм содержит все составляющие, соответствующие исходному состоянию стали, и, кроме того, образовавшийся при обработке аустенит.

При плотности энергии 5 Дж/см^2 общая толщина модифицированного слоя снижается до ~8 мкм, при этом в верхней оплавленной зоне этого слоя содержится небольшое количество аустенита и отсутствуют карбиды M_6C, крупные включения которых наблюдаются в более глубоких слоях модифицированной области. Данные РЭМ свидетельствуют, что карбид VC не распадается даже вблизи облученной поверхности. Толщина модифицированной области увеличивается с повышением энергии падающего потока и увеличением количества импульсов азотной плазмы до 27 мкм при Q=13 Дж/см^2 и n=5.

Распределение микротвердости по глубине модифицированного слоя коррелирует с наблюдающимися изменениями его микроструктуры. В стали Р6М5 поверхностный слой азотистого аустенита, обедненный карбидами, имеет более низкую твердость, чем слой, содержащий упрочняющие карбиды. Соответственно, твердость поверхностного слоя быстрорежущей стали после плазменной обработке с Q=5 Дж/см^2 больше, чем после обработки с Q=10 Дж/см^2, что обусловлено большее высокой концентрацией карбидов в обработанном слое. Изменение микротвердости с увеличением количества импульсов хорошо согласуется с изменением фазового состава модифицированного слоя. Для стали Х12М микротвердость также уменьшается вследствие растворения вторых фаз.

Трибологические испытания показали заметное улучшение износостойкости обработанных образцов, при этом следует отметить, что визуально треки износа на облученных образцах не наблюдаются, что свидетельствует о незначительном уносе массы при трении. Обнаружено, что коэффициент трения гораздо слабее зависит от режима обработки, чем микротвердость.

Таким образом, полученные результаты свидетельствуют, что "плазменную закалку" инструментальных сталей следует проводить в узком интервале значений плотности энергии плазменного потока, при котором реализуется начальная стадия растворения карбидов. Оптимальные характеристики сталей достигаются при плазменной обработке, вызывающей минимальное оплавление поверхности.

Как и в случае полупроводников, исследовалось изменение структуры и свойств систем “пленка металла-сталь” [26-29]. Рассмотрим это на примере системы “титан-сталь” [27]. Толщина пленки титана на стали Ст3 (0,2 C; 0,2 Si; 0,5 Mn) составляла 0,4-1,0 мкм. Плотность энергии - 13 Дж/см^2.

Воздействие компрессионного плазменного потока азота приводит к расплавлению верхнего поверхностного слоя образцов. В результате обработки существенно изменяется рельеф поверхности. Появляются хорошо заметные, хаотично расположенные волны, характерные при данном виде обработки для быстро затвердевшей жидкой фазы. РЭМ исследования при большем увеличении показали, что поверхность образцов, подвергнутых обработке, содержит области с ячеистой

структурой (рис. 14), причем на поверхности углеродистой стали без покрытия (рис. 14а) наблюдаются ячейки размером ~0,5 мкм, а в случае перемешанной системы титан-сталь (рис. 14б) образуются ячейки размером 0,5-1,0 мкм, отличающиеся не только размером, но и формой. Формирование ячеистой структуры, вероятнее всего, вызвано проходящей на стадии затвердевания ячеистой кристаллизацией.

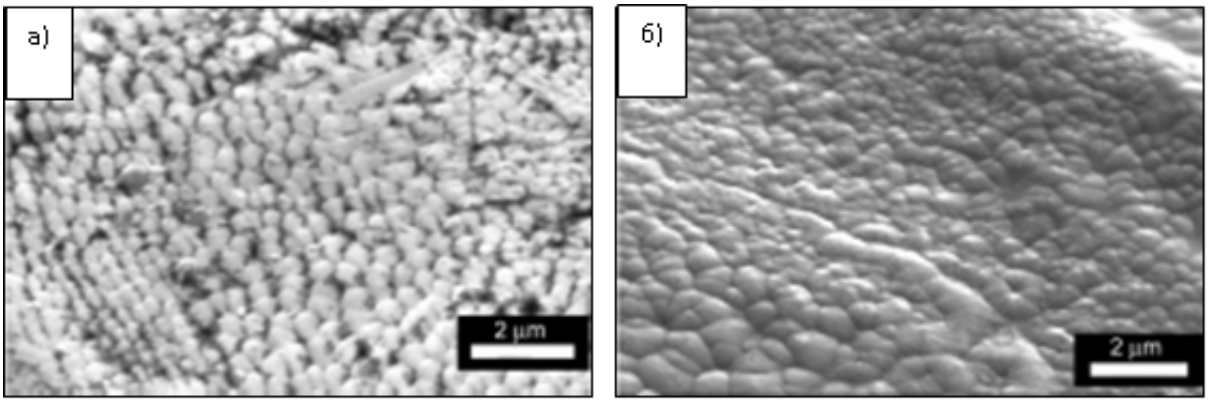


Рис. 14. РЭМ изображения ячеистой структуры поверхности обработанных образцов:
а - углеродистая сталь; б - система титан-сталь (толщина покрытия Ti ~1 мкм).

Развитие структурной неустойчивости на фронте кристаллизации (как и вслучае полупроводников) можно описать моделью концентрационного переохлаждения, приводящей к неустойчивости плоского фронта кристаллизации, так как любой выступ, образующийся на поверхности, оказывается в переохлажденной жидкости и поэтому не исчезает. Эти неустойчивости при определенных условиях могут развиваться в стационарные периодические ячеистые структуры. Формирование ячеистой структуры обычно приводит к образованию примесных сегрегационных зон, являющихся границами столбчатых ячеек-кристаллов. При этом в примесных сегрегационных зонах вследствие различия периода решетки материала в областях сегрегаций и в объеме кристалла-ячейки появляются внутренние напряжения, достаточные для зарождения дислокаций. Плотность дислокаций в промежутках между ячейками может достигать ~10^{10} $см^{-2}$.

В исходном состоянии углеродистая сталь имеет ферритно-перлитную структуру. На рентгенограмме (рис. 15а) наблюдаются дифракционные линии феррита и не выявляется цементитная оставляющая перлита вследствие высокой дисперсности карбида 8-Fe_3C. Рентгенограмма образца стали с титановым покрытием (рис. 15б) представляет собой суперпозицию дифракционных пиков феррита и титана. Было обнаружено, что последующая обработка компрессионным плазменным потоком приводит к изменению фазового состава.

Импульсное воздействие высокотемпературной плазмы на тонкопленочную систему приводит к расплавлению покрытия и слоя подложки и последующему жидкофазному перемешиванию обоих компонентов под действием давления плазмы. Как только действие импульса прекращается, происходит остывание и повторное затвердевание перемешанной системы.

В результате обработки системы титан-сталь на рентгенограмме (рис. 15в) исчезают дифракционные отражения титана, а дифракционные линии феррита смещаются в сторону меньших углов. Исследования элементного состава поверхности необработанной системы титан-сталь методом рентгеновского спектрального микроанализа показывают лишь наличие 3-5 ат.% железа вследствие превышения

размера области генерации над толщиной пленки титана. В результате обработки содержание Fe на поверхности достигает 70-80 ат.%. Такое изменение концентрации элементов может быть вызвано либо испарением части титанового покрытия, либо жидкофазным перемешиванием обоих компонентов под действием импульса плазмы и взаимной диффузией атомов титана и железа. Естественно предположить, что отсутствие на рентгенограммах дифракционных линий Ti связано с образованием твердого раствора замещения, в который титан входит в качестве надразмерной примеси.

На это указывает и значительное смещение дифракционных линий феррита в сторону меньших углов, что обусловлено замещением атомов железа атомами титана, имеющими больший атомный радиус. Исследования относительного изменения параметра решетки с ростом толщины титанового покрытия показали, что увеличение содержания Ti в плазменно-переме-шанном слое приводит к деформации решетки твердого раствора железо-титан.

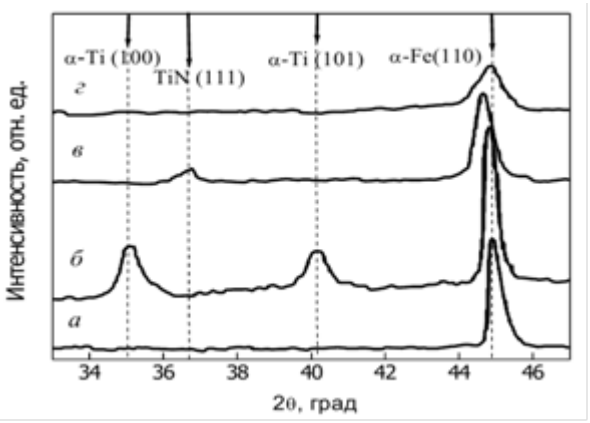


Рис. 15. Участки рентгенограмм исследуемых образцов:
а - углеродистая сталь; б - система титан-сталь (толщина покрытия Ti ~1 мкм);
в - обработанная система титан-сталь (толщина покрытия Ti ~1 мкм); г - обработанная сталь.

Установлено также взаимодействие азотной плазмы с элементами системы, в частности, формирование нитрида титана (рис. 15в). Следует отметить, что формирование TiN после обработки было отмечено на всех образцах с покрытием любой толщины. Доказательством формирования TiN является визуально наблюдаемое появление характерного золотистого цвета поверхности. Однако объемная доля этой фазы невелика, и методом РСА зафиксировать ее появление можно лишь при толщине предварительно нанесенного покрытия титана около 1 мкм. Таким образом, согласно данным рентгеновского зондирования, в модифицированном слое перемешанной системы "титановое покрытие-стальная подложка" содержится высокая концентрация азота, достаточная для формирования нитрида титана. Во время остывания и повторного затвердевания смеси компонентов железо-титан Ti входит в решетку Fe в качестве надразмерной примеси.

Литература

1. Морозов А.И. Физика плазмы. 1975, т. 1, № 2, с. 179-191.
2. Углов В.В., Анищик В.М., Асташинский В.В., Ананин С.И., Аскерко В.В., Асташинский В.М. и др. Письма в ЖЭТФ, 2001, т.74, №4, с.234-236.
3. Uglov V.V., Anishchik V.M., Astashynski V.V., Astashynski V.M. et al. Surf.Coat.Technol., 2002, v.158-159, p.273-276.

4. Ананин С.И., Асташинский В.М., Костюкевич Е.А. и др. Физика плазмы, 1998, т. 24, с. 1003.
5. Astashynski V.M., Ananin S.I., Kostyukevich E.A., Kuzmitski A.M. et al. High Temperature Material Processes, 2007, v. 24, p. 537-548.
6. Углов В.В., Черенда Н.Н., Анищик В.М., Свешников Ю.В. Физика и химия обработки материалов, 2005, № 4, с. 31-35.
7. Квасов Н.Т., Шедко Ю.Г., Углов В.В., Асташинский В.М. Доклады БГУИР, 2007, №4, с. 101-107.
8. Astashynski V.M., Ananin S.I., Askerko V.V., Kostyukevich E.A. et al. Surf.Coat.Technol., 2004, v.180-181, p.392-395.
9. Анищик В.М., Асташинский В.М., Квасов Н.Т., Углов В.В. и др. Физика и химия обработки материалов, 2008, № 5, с. 27-33.
10. Глазов В.М., Чижевская С.Н., Глаголева Н.Н. Жидкие полупроводники. М.: Наука, 1967, 244 с.
11. 11.Chandrasekhar S. Hydrodynamic and hydromagnetic stability. Oxford: University Press, 1961, 655 p.
12. Ландау Л.Д., Лифшиц Е.М. Теоретическая физика. Т.8. Электродинамика сплошных сред. М.: Физматлит, 2003, 656 с.
13. Uglov V.V., Anishchik V.M., Kvasov N.T., Petukhou Yu. A. et al. Vacuum, 2009, v.83, p. 1152-1154.
14. Kaya H., Cadirli E., Gunduz M. Journal of Material ngineering and Perfomance, 2006, v. 16(1), p. 12-21.
15. Углов В.В., Анищик В.М., Асташинский В.В., Свешников Ю.В. и др. Физика и химия обработки материалов, 2004, № 4, с. 37-42.
16. Углов В.В., Анищик В.М., Асташинский В.В., Свешников Ю.В. и др. Физика и химия обработки материалов, 2002, № 3, с. 23-28.
17. Anishchik V.M., Uglov V.V., Astashynski V.V., Astashynski V.M. et al. Vacuum, 2003, v.70, p.269-274.
18. Cherenda N.N., Uglov V.V., Anishchik V.M., Stalmoshenok E.K. et al. Vacuum, 2005, v.78, p.483-487.
19. Anishchik V.M., Astashynski V.V., Uglov V.V., Fedotova J.A. et al. Vacuum, 2005, v.78, p.589-592.
20. Углов В.В., Анищик В.М., Стальмошенок Е.К., Черенда Н.Н. Физика и химия обработки материалов, 2004, № 5, с. 44-49.
21. Cherenda N.N., Uglov V.V., Anishchik V.M., Stalmoshenok E.K. et al. Surf. Coat.Technol., 2006, v.200, p.5534-5542.
22. Лифшиц Е.М., Питаевский Л.П. Физическая кинетика. М.: Наука, 1979, 528 с.
23. Конторович И.Е., Совалова А.А. ЖТФ, 1950, т.20, №1, с.53-65.
24. Якушин В.Л., Калин Б.Л., Скрытный В.И., Буланов И.А. Тр. X Межнац. совещ. "Радиационная физика твердого тела", 2000, с.273-279.
24. Геллер Ю.А. Инструментальные стали. М.: Металлургия, 1983, 527 с.
25. Uglov V.V., Anishchik V.M., Cherenda N.N., Stalmoshenok E.K. et al. Vacuum, 2005, v.78, p.489-493.
26. Углов В.В., Анищик В.М., Черенда Н.Н. Стальмошенок Е.К. и др. Физика и химия обработки материалов, 2005, № 2, с. 36-41.
27. Uglov V.V., Cherenda N.N., Anishchik V.M., Stalmoshenok E.K. et al. Vacuum, 2007, v.81, p.1341-1344.
28. Углов В.В., Черенда Н.Н., Стальмошенок Е.К., Полуянова М.Г. и др. Перспективные материалы, 2009, № 3, с. 69-76.

5. BEMLAB - open source, objective Boundary Element Method library

J. Sikora[1,2], P. Wieleba[2], W. Wójcik[1]

[1)]Lublin University of Technology,
Faculty of Electrical Engineering and Computer Science,
38a Nadbystrzycka Str., 20-618 Lublin, Poland,

[2)]Department of Metrology and Nondestructive Testing,
Electrotechnical Institute,
28 Pożaryskiego Str., 04-703 Warsaw, Poland.

Nomenclature

Abbreviations
BEM – Boundary Element Method
FEM – Finite Element Method
DOT – Diffuse Optical Tomography
EIT – Electrical Impedance Tomography
CT – Computer Tomography
NMRI – Nuclear Magnetic Resonance Imaging
NIR – Near Infra Red
PDE – Partial Differential Equation
BIE – Boundary Integral Equation
RTE – Radiative Transfer Equation

Symbols

Ω – Domain
Γ – Boundary
Γ_j – Boundary element j
Φ – Potential [V]; Photon density [$1/m^3$]
$\vec{n}$ – Normal vector outward the domain Ω
k – Wave number
G – Fundamental solution, Green function
∂/∂n – Normal derivative
ω – Frequency of source light intensity modulation

5.1. Introduction

Tomography imaging techniques require proper numerical models. Data gathered from the hardware are used to create an image of the examined object internal structure. The inverse problem using the adequate numerical model is solved and its results compose the picture of the object internal.

The inverse problem allows to find object internal model parameters m using data *d* gathered form the boundary of the model using the scan hardware. The relationship between *d* and *m* can be written as:

$$d = \aleph(m) \tag{1}$$

where $\aleph$ is a non-linear operator which represents the numerical model of the physical problem.

The base unit of the inverse problem is the forward problem which allows to calculate d based on known m. It is very important to calculate forward problems as fast as possible to process further image reconstruction effectively. It is particularly important in Diffuse Optical Tomography (DOT) and Electrical Impedance Tomography (EIT) where calculations are very time consuming [11].

The most common tomography techniques in medicine are the X-ray based Computer Tomography and the Nuclear Magnetic Resonance Imaging (NMRI). Computer Tomography uses X-rays which are the ionizating radiation. Long tissues exposure on X-ray based radiation is very dangerous therefore CT cannot be used frequently. Whereas Nuclear Magnetic Resonance Imaging uses magnetic field from 0.5 to 2 Tesla. Larger values of magnetic field are prohibited in medicine because it is not neutral for living organisms. Both mentioned techniques are volumetric, therefore precise image is obtained and exact object interior structure is presented.

In contrast to DOT commonly used X-ray based Computer Tomography and NMRI use fast algorithms. CT uses back-projection while NMRI uses Fourier Transform based algorithms. Availability of fast algorithms to reconstruct the image made it possible to popularize these methods in medicine testing.

However CT and NMRI are not ideal testing methods. X-ray based CT can only be used rarely because of its dangerous influence on tissues. NMRI is also not neutral for living organisms. Moreover both methods require devices of big dimensions. NMRI requires extensive cooling, therefore special installations have to be applied. Diffuse Optical Tomography does not have drawbacks of CT and NMRI. DOT uses near infrared (NIR) light which is safe for tissues, which can be exposed to it permanently. The size of optical scanners is relatively small – they are portable. However DOT cannot be used for precise volumetric

imaging. It also requires much more efficient processing units. Nowadays used algorithms are very slow therefore they make it impossible to introduce DOT to every day medical testing.

The main areas of DOT application in medicine are:

- neonatal head testing of brain haemorrhage,
- breast testing for detecting tumours.

Usage of DOT for testing infants brain haemorrhage is especially important. It is required to test the brain with haemorrhage permanently, so doctors could know if it increases or decreases and if applied treatment is appropriate. It is also beneficial to use DOT for detecting tumours in breasts. Nowadays popularly used mammography uses X-ray based ionizating radiation and as testing should be done regularly it is not neutral. Moreover while the mammography test is taken, breasts are deformed whereas while using DOT scanners they are not. Despite of fact that the image is not as precise as that reconstructed using CT or NMRI it is desired to introduce DOT imaging in the mentioned areas. However the main drawback of DOT image reconstruction, which is the long time of image reconstruction has to be solved.

Further sections describe universal, open source and objective software implementing Boundary Element Method (BEM) for solving partial differential equations, which can be used in tomography applications.

5.2. Radiative Transport Equation in Diffuse Optical Tomography

Firstly the numerical model for Diffuse Optical Tomography have to be introduced. Near infrared light used in DOT is an electromagnetic wave. Therefore the light transport phenomenon can be described using Radiative Transport Equation (RTE). Depending on the type of scanner or its work mode the source of near infrared light can be described as defined in:

- time domain – the signal is in the form of ultra fast impulses,
- frequency domain – the light intensity modulation.

RTE defined in the time domain has the following form:

$$\left(\breve{s}\cdot\nabla+\mu_a(r)+\mu_s(r)+\frac{1}{c}\frac{\partial}{\partial t}\right)\varphi(r,\breve{s},t)=\mu_s(r)\int_s^{n-1}\Theta(\breve{s},\breve{s}')\varphi(r,\breve{s},t)d\breve{s}'+q(r,\breve{s},t) \qquad (2)$$

and in the frequency domain [13]:

$$\left(\breve{s}\cdot\nabla+\mu_a(r)+\mu_s(r)+i\frac{\omega}{c}\right)\varphi(r,\breve{s},\omega)=\mu_s(r)\int_{s^{n-1}}\Theta(\breve{s},\breve{s}')\varphi(r,\breve{s},\omega)d\breve{s}'+q(r,\breve{s},\omega) \qquad (3)$$

where: $\breve{s}$ – directional vector; $\Theta(\breve{s},\breve{s}')$ – dispersing phase function describing probability that photon with the beginning direction $\breve{s}'$ will have direction $\breve{s}$ after the dispersing event occurs; q – source inside the examined domain Ω, φ – photon density; μ_a – absorption coefficient; μ_s – disperse coefficient; r – geometrical coordinates of the examined point; c – the speed of light in the medium; t – time; ω – frequency.

The above RTE equation is a precise numerical model for light transport phenomenon including NIR. However the numerical model based on RTE is difficult to solve, because of a

long time required to obtain results using nowadays hardware. Monte Carlo is one of the methods which can be used to solve RTE.

However, the Diffusion Optical Tomography operates on testing objects consisting of tissues, which characterize the following relation:

$$\mu_s^{'} >> \mu_a \tag{4}$$

Tissues absorption coefficient is much smaller then the dispersing coefficient. Thanks to this fact RTE can be reduced to the diffusion equation without loss of results quality. Then the diffusion equation in the time domain can be formed as [1, p. 1535] [9, p. 1780] [10, p. 896]:

$$\left(\nabla \cdot D\nabla - \mu_a - \frac{\partial}{\partial t}\right)\varphi(r,t) = q_o(r,t) \tag{5}$$

and in the frequency domain [11, p. 139] [10, p. 896]:

$$\left(\nabla \cdot D\nabla - \mu_a - \frac{i\omega}{c}\right)\varphi(r,\omega) = q_o(r,\omega) \tag{6}$$

where D – the diffusion coefficient:

$$D = \frac{1}{3\left(\mu_a + \mu_s^{'}\right)} \tag{7}$$

Further discussion will concentrate on a frequency domain because then the medical testing is shorter. The diffusion equation in the frequency domain (6) can be presented as the Helmholtz equation including the source q_o:

$$\nabla^2\varphi(r,\omega) - k^2\varphi(r,\omega) = -\frac{q_o(r,\omega)}{D}, \quad \text{where} \quad k^2 = \frac{\mu_a}{D} - i\frac{\omega}{cD} \tag{8}$$

where: $k \in C$ – is the complex wavenumber.

The collimated source NIR light is supplied to the baby head or the surface of the breast through the fibre-optic. In the numerical model it is modeled as the point source located under the surface at the distance of $\frac{1}{\mu_s^{'}}$.

Forward problem definition requires also setting boundary conditions (BC). The third kind BC, also known as Robin BC, are used in the DOT model and they represent the following relationship:

$$\frac{\partial\varphi}{\partial n} = m_{RBC}\varphi + n_{RBC} \tag{9}$$

where m_{RBC}, n_{RBC} – parameters.

The above Helmholtz Partial Differential Equation (PDE) can be solved using any applicable method and also Boundary Element Method with title BEMLAB software.

5.3. Governing equation in Electrical Impedance Tomography

Another considered type of tomography imaging technique is Electrical Impedance Tomography (EIT). EIT uses electrical properties of examined materials like electrical conductivity σ. The examined object is stimulated using voltage or current source and the layout of potential on the surface is collected using sensors. These data are used by the reconstruction algorithm which gives a layout of objects internal. EIT numerical model involves the Laplace equation [12, p. 112]:

$$\operatorname{div}(\sigma \operatorname{grad}\varphi) = 0. \tag{10}$$

The Laplace PDE can also be solved using Boundary Element Method and the title BEMLAB software.

5.4. Boundary Element Method

Diffusive Optical Tomography and Electrical Impedance Tomography problems are popularly solved using Finite Element Method (FEM) [14]. FEM is the domain method which means that the whole object domain Ω has to be discretized. One of the FEM alternatives is the Boundary Element Method. BEM requires only the surface Γ of the examined object to be discretized, therefore its dimension is smaller by one than in domain methods. The good quality boundary mesh creation task is much simpler than creating a domain mesh. BEM is the method characterized by the square computational complexity $0(N^2)$ [8].

Moreover calculation of φ in any point inside the examined domain Ω is done without remeshing the domain. The number of equations in BEM is usually much less than in FEM. BEM has advantages comparing to FEM, but also there are some drawbacks. When the problem is characterized with the unsuitable geometry, which means that the number of boundary elements Γ_j is close to the number of domain elements Ω_j in FEM, then BEM calculations are slower than in FEM. This is strengthened by the fact that the left hand side matrix a in the set of linear equations (20) is dense in BEM in contrast to the sparse one in FEM. Laplace equation (10) or Helmholtz equation (8) can be solved using BEM when they are defined as Boundary Integral Equation (BIE). BIE has the following analytical form:

$$c_i\varphi + \int_\Gamma \varphi \frac{\partial G}{\partial n} d\Gamma = \int_\Gamma \frac{\partial \varphi}{\partial n} G d\Gamma + \int_\Omega f\ G\ d\Omega \tag{11}$$

where: Ω – the problem domain, Γ – boundary of domain Ω, φ – field function potential or photon density in DOT, G – Green function, so-called fundamental solution, $n = |\vec{n}|$ – length of the normal vector directed outwards to the boundary element, ∂ / ∂n – normal derivative, f – domain (source) function, c_i – coefficient removing or restricting the singularity from the primitive function of PDE solution.

Table 1. Green Functions

Space dimension	Green functions *G(R)* for:	
	Laplace / Poisson PDE	Helmholtz PDE
1D	$G(R)=-\frac{1}{2}R$ (12)	$G(R)=-\frac{1}{2k}\sinh(kR)$ (13)
2D	$G(R)=\frac{1}{2\pi}\ln\frac{1}{R}$ (14)	$G(R)=\frac{1}{2\pi}K_0(kR)$ (15)
3D	$G(R)=\frac{1}{2\pi R}$ (16)	$G(R)=\frac{1}{2\pi R}e^{-kR}$ (17)

The Green function G mentioned above varies for each type of PDE and dimension of the space in which the problem is defined. Selected Green functions for Laplace/Poisson and Helmholtz equations were gathered in table 1. It is worth to see that when the value of R decreases to zero (R $\rightarrow$ 0), the singularity occurs and special integration procedure has to be taken.

If the problem is to be calculated using BEM, only the boundary Γ of the examined domain Ω has to be discretized.

In figure 1 there is presented an example 2D object, which boundary Γ was discretized with linear boundary elements. The boundary Γ around the domain Ω was discretized with 12 boundary elements Γ_j, where j ∈ ⟨1, 12⟩. There are also 12 boundary nodes, where i ∈ ⟨1, 12⟩. Boundary conditions were also marked in the model. Dirichlet Boundary Conditions (DBC) were applied on the top and bottom border (known potential φ), whereas Neumann Boundary Conditions were applied on the left and right border of the model (known potential normal derivative $\partial\varphi/\partial n$. As can be noticed the domain was not discretized. While creating the boundary mesh it is required that the normal vector to the boundary element $\vec{n}$ is directed outward the examined domain Ω.

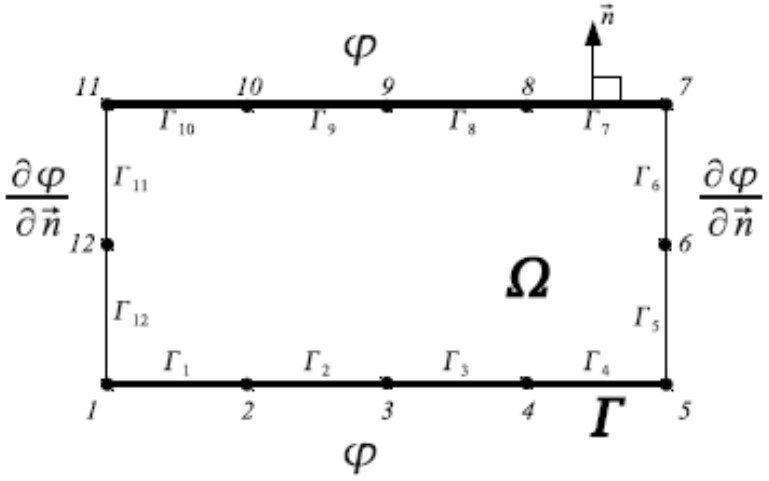


Figure 1: Example 2D boundary mesh with the marked boundary conditions

When the domain boundary Γ is discretized the Boundary Integral Equation (11) has to be written in the discretized form. The boundary was discretized into *J* boundary elements and it has *I* nodes. Each element has *K* nodes which interpolate potential φ and *L* nodes which interpolate its normal derivative $\partial\varphi/\partial n$. When the source is defined in the domain Ω, it can be

discretized into D domain elements. The discretized BIE for one domain problem has the following form:

$$c_i(\mathbf{r}_i)u_i(\mathbf{r}_i) + \sum_{j=1}^{J}\sum_{k=1}^{K} u_k^{(j)} \int_{\Gamma_j(\xi)} N_k^{(u)}(\xi) \frac{\partial G(\mathbf{r}(\xi),\mathbf{r}_i)}{\partial n} \left| J_{\Gamma_j}^{(g)}(\xi) \right| d\Gamma_j(\xi) =$$

$$= \sum_{j=1}^{J}\sum_{l=1}^{L} q_l^{(j)} \int_{\Gamma_j(\xi)} N_l^{(q)}(\xi) G(\mathbf{r}(\xi),\mathbf{r}_i) \left| J_{\Gamma_j}^{(g)}(\xi) \right| d\Gamma_j(\xi) + \tag{18}$$

$$+ \sum_{d}^{D} \int_{\Omega_d(\varsigma)} f(\mathbf{r}(\varsigma)) G(\mathbf{r}(\varsigma),\mathbf{r}_i) \left| J_{\Omega_d}^{(g\Omega)}(\varsigma) \right| d\Omega_d(\varsigma),$$

where: $N_k^{(u)}$ – base interpolation functions for potential φ; $N_l^{(q)}$ – base interpolation functions for potential derivative $\partial\varphi/\partial$n ; $\left|J_{\Gamma_j}^{(g)}(\xi)\right|$ – Jacobian of transformation of geometrical boundary element from the global coordinate system to the local coordinate system; $\left|J_{\Omega_d}^{(g\Omega)}(\xi)\right|$ – Jacobian of transformation of geometrical domain element from the global coordinate system to the local coordinate system; ξ – boundary point local coordinates of transformed boundary element; ς – domain point local coordinates of transformed domain element; $u_k^{(j)}$ – potential φ value in the node k of the element j; $q_l^{(j)}$ – potential normal derivative $\partial\varphi/\partial$n in the node l of the element j.

Normally potential φ and its normal derivative $\partial\varphi/\partial$n are interpolated by boundary elements of the same type, which results the following relation:

$K = L$ – number of boundary element nodes is the same for both interpolation functions φ and $\partial\varphi/\partial$n .

Matrices are good containers and matrix calculations are clear, therefore BIE (18) can be presented in the matrix form:

$$\mathbf{A}\varphi = \mathbf{B}\frac{\partial}{\partial n}\varphi + \mathbf{F}. \tag{19}$$

When the Laplace PDE is calculated the vertical vector **F** responsible for domain source potential is equal zero: **F** = **0**. Introducing boundary conditions is the next step. BCs are applied to the equation (19) which results in the following set of linear equations:

$$\mathbf{ax} = \mathbf{b}, \tag{20}$$

where: **x** – is the unknown vector composed from unknown boundary potentials $u = \varphi$ and/or potential normal derivatives $q = \partial\varphi/\partial$n : $\mathbf{x} = \begin{bmatrix} \varphi \\ \dfrac{\partial\varphi}{\partial n} \end{bmatrix}$. It is worth to notice that the left hand side matrix a is dense.

The set of equations (20) can be calculated using solvers based on LU decomposition [6], faster GMRES [7, 18] or any other available. As the result of BEM calculations all

potential $u = \varphi$ and its normal derivative $q = \partial\varphi/\partial$n are known in the boundary nodes. Internal node values do not have to be calculated but if needed their values are calculated using potential and its normal derivative values from all boundary nodes.

5.5. Multi domain problems

The above discretized form of BIE (18) is applicable for one domain problem. But BEM can also solve multi domain problems. Then BIE should be modified so regions are included.

Two domain problem is presented in figure 2. There is a boundary $\Gamma^{(1:2)}$ named an interface between domains (regions) $\Omega^{(1)}$ and $\Omega^{(2)}$. There are also marked external boundaries for particular domains: $\Gamma^{(1)}$ for $\Omega^{(1)}$ and $\Gamma^{(2)}$ for $\Omega^{(1)}$. This particular notation of boundaries and consequently BIE was created specially for the BEMLAB library. Normal vectors $\vec{n}$ to the boundaries were also marked. When the domain $\Omega^{(1)}$ is considered the vector $\vec{n}_1$ is taken, when the domain $\Omega^{(2)}$ is considered the vector $\vec{n}_2$ is taken.

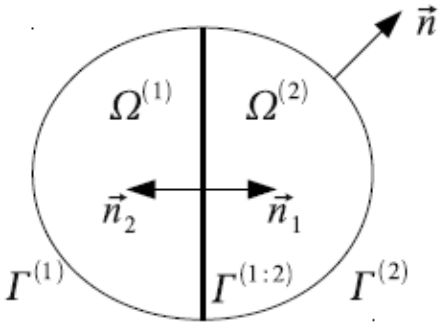


Figure 2: Two domain ($\Omega^{(1)}$ and $\Omega^{(2)}$) problem with a marked interface $\Gamma^{(1:2)}$ between them

The Boundary Integral Equation for the multi domain problem made of $\Re$ domains created for BEMLAB numerical package has the following form:

$$\sum_{r=1}^{\Re}\sum_{b=1}^{B_r} c_i(\mathbf{r}_i)u_i(\mathbf{r}_i) +$$

$$+\sum_{r=1}^{\Re}\sum_{b=1}^{B_r}\left(\sum_{j=1}^{J_{b_r}}\sum_{k=1}^{K} u_k^{(j)} \int_{\Gamma_j(\xi)} N_k^{(u)}(\xi)\frac{\partial G(\mathbf{r}(\xi),\mathbf{r}_i)}{\partial n}\left|J_{\Gamma_j}^{(g)}(\xi)\right| d\Gamma_j(\xi)\right) =$$

$$= \sum_{r=1}^{\Re}\sum_{b=1}^{B_r}\left(\sum_{j=1}^{J_{b_r}}\sum_{l=1}^{L} q_l^{(j)} \int_{\Gamma_j(\xi)} N_l^{(q)}(\xi)G(\mathbf{r}(\xi),\mathbf{r}_i)\left|J_{\Gamma_j}^{(g)}(\xi)\right| d\Gamma_j(\xi) + \right. \tag{21}$$

$$\left. + \sum_{d}^{D_r} \int_{\Omega_d^{(r)}(\varsigma)} f_r(\mathbf{r}(\varsigma))G_r(\mathbf{r}(\varsigma),\mathbf{r}_{i_r})\left|J_{\Omega_d}^{(g\Omega)}(\varsigma)\right| d\Omega_d^{(r)}(\varsigma)\right)$$

where: B_r – is the number of boundaries neighbouring the current region (domain) r; b_r – is the current boundary between the current region r and the neighbour (or external region marked as Ω', where $\Omega' \notin \Omega$).

The matrix BIE equation (19) and the set of equations (20) has the same form for multi domain problems. However before applying boundary conditions, interface conditions have to be additionally applied to the equation (21). There are two interface continuity conditions on the interface:

Potential continuity – In the interface node the following relationship occurs:

$$\varphi_1\big|_{\Gamma^{(1:2)}} = \varphi_2\big|_{\Gamma^{(1:2)}} \tag{22}$$

Potential normal derivative continuity – In the interface node the following relationship occurs:

$$m_1 \frac{\partial \varphi_1}{\partial n}\bigg|_{\Gamma^{(1:2)}} = -m_2 \frac{\partial \varphi_2}{\partial n}\bigg|_{\Gamma^{(1:2)}} \tag{23}$$

where: m_1, m_2 – are material parameters of particular domains. The − (minus) sign exists because normal vectors has the same direction but opposite turns as is marked in Fig. 2.

After solving the set of equation (20) for multi domain problems, all potential $u = \varphi$ and its normal derivative $q = \partial\varphi/\partial$n values in nodes of external boundaries $\Gamma^{(r)}$ and interfaces $\Gamma^{(r_1:r_2)}$ are known.

This section includes some basic information about Boundary Element Method, which were required to model and implement BEMLAB library. There were presented notation and modelling procedure which make possible to create the universal BEM software applicable to diverse problems.

5.6. BEMLAB software

The name Boundary Element Method was proposed by C. A. Brebbia in 1970s [2]. The number of applications increased since then and further areas are being investigated. At the same time development of FEM was much more rapid and nowadays its role is indisputable. When the computers become more common and high level programming languages arise, lots of applications implementing FEM were created. There is a broad choice of open source as well as commercial FEM software. Everyone can choose the one which is the best for particular applications. This is also one of the reasons why the FEM is much more popular method for solving PDEs than BEM.

The BEM software availability is much smaller than the broad choice of FEM one [5, 15]. And unfortunately, only little BEM codes were created for all the time since the BEM arose. One of the reasons may be the fact that BEM mathematical description is more complicated. Another one probable cause is the existence of problematic singularities which have to be taken into consideration and solved. Because of lack of BEM software applicable to tomography applications the BEMLAB [17] project was initiated.

5.6.1. Technology

Licensing

It was decided that the BEMLAB software will be developed using the open source licence. Some of the reasons why the code is publicly available are: the intention of creating the community around the project, Boundary Element Method popularization among engineers and scientists, its further development and acceleration.

BEMLAB binary packages and the source code are distributed under GNU LGPL (Lesser General Public License) license terms. The project provides universal library and the reference application, which is the easiest way for solving problems using BEM. There are also auxiliary programs provided to facilitate engineer's tasks.

Technology main goals

The licensing and technology were chosen in the way so the BEMLAB project could be named as a "good open source project", which comply criteria described in e.g. [3]. The following aims were set against the project:

- calculation correctness,
- usage easiness,
- further development easiness, by choosing well known technologies.

The numerical software is a special type of software which target group is relatively small comparing to the system or application software. Therefore the chosen technology, tools, modelling and development procedures should be already known to the potential users, so they can be easily engaged to use the created software. It is particularly important in case of BEM software, because the method by itself is not widely known.

Objectivity

BEMLAB software has an objective architecture because it simplifies the process of modelling, development and further hosting. This also decreases the entrance level for new developers so they can faster and easier get to know the project architecture. Moreover there is a Unified Modelling Language (UML) [4] available, which allows creation of standardized diagrams made of unified symbols. UML allows creation software requirements and architecture during the whole process of modelling and development. UML diagrams are unambiguous therefore all projects participants has a clear view of how the code will be implemented or is already implemented. Furthermore "pictures" like diagrams are generally easier assimilated than the code by itself. The objectivity allows modelling and creating the code which is better adapted to the reality, than the functional one.

Programming language

Nowadays, many programming languages support object-oriented programming. C++ was chosen as the main programming language used in BEMLAB software. Other considered were Java and C#, but both require virtual machines and their programmes are slower than those written in C++. It is important that those three programming languages are popular and are taught in all computer engineering studies of the undergraduate and graduate courses. Nowadays computer companies use them to create software.

Some available BEM codes [15] are written in Fortran which was a popular language of numerical software. Nowadays it is not popular, commercial software does not use it and finally it is taught in a scarce scope if not at all. Moreover functional C and Objective C++ are popular in the open source community.

Compiler

C++ compilers are available on almost any platform and operating system. Almost all C/C++ software in Linux/Unix like operating systems use GNU compilers. BEMLAB reference implementation also uses GNU C++ Compiler which is the base compiler among open source software. There are multithreaded algorithms specially designed and implemented for BEMLAB, which fasten BEM calculations on multicore processors and on multiprocessor computers. BEMLAB library uses threads introduced in C++0x specification.

Therefore it is recommended to use GNU Compiler version 4.4.0 or never, because C++0x threads are available from that version only. However up to now it is possible to use previous versions of GNU compilers (tested all major versions since GNU Compiler 3.3), but with the restriction that multithreaded algorithms are turned off. This is deprecated but makes possible to use BEMLAB on older operating systems where newer versions of GNU Compiler are not yet available. However the code without multithreading will be completely removed in the future.

Code manager

The most important case for the end user is possibility of using the software. The user wants to run the software in the known and the easiest way possible. Similarly the programmer wants to compile and build the software efficiently. Code managers come with help to fulfil these requirements. Code managers allow to automate such tasks like environment configuration, code compilation, binaries building, installation or source package creation. All tasks are very important but the environment configuration is worth noting, because thanks to it the developer does not have to bother how the end user environment and the operating system are configured. Differences between platforms and systems distributions are transparent and the code manager manages with them automatically. Moreover it provides the possibility to provide user's special configuration options in the unified way e.g. install the application in a non standard location. Two code managers were taken into account:

- the suite of tools: automake, autoconf and libtool, also known as autotools,
- cmake.

Finally the first one was used despite of its disadvantages like e.g. difficulty to programme. However it is much more common, users are used to it and simply is fair enough. The second one (cmake) is not so popular and maybe it will not become. The code manager is required to develop the code effectively.

Version control system

Almost none programming project can obey without Version Control System (VCS). And not only programming one e.g. this chapter creation (writing) was carried out with the help of version control system. VCS is designed to store all versions of particular files. It allows to compare changes, create branches, merge codes versions. It has the full history of changes made in the code or document. It allows a group of people to develop the code effectively.

Without the VCS it would be impossible. When the BEMLAB project was started out two version control systems engines were considered:

• CVS – popular and widely used,

• SVN – newer, somewhat less popular than CVS, but with substantial advantages and much modern.

Nowadays GIT is obtaining consecutive applications because of its modern architecture. However it was not mature enough that time and is not yet as popular as both mentioned above together – support on various operating systems is still restricted. Finally the SVN was chosen, which has been used for the whole processed development. Publicly the repository is available under the address:

svn://svn.bemlab.org/bemlab

Until now SVN is sufficient but the repository migration to GIT cannot be excluded in the future.

Website

Many open source projects have websites and every “good open source project” must have one. The website is easily available and is the most popular place for distributing software and documentation. The community is gathering around the projects website and available services. This is a very important part of the project. BEMLAB project has its own domain:

http://bemlab.org/

There are many web applications available but only actively developed and with a good support (bug fixes deployment) were considered to be used with a BEMLAB project. Mainly Content Management Systems (CMS) such as Joomla or Drupal were considered, and Wiki applications such as MoinMoin and MediaWiki. Finally MediaWiki [20] was chosen – it is broadly used (among others by Wikipedia) and bug fixes are systematically made available. Moreover a lot of people know MediaWiki interpreter and it supports LaTeX equations.

The website and the software also require the logo so they can be easily recognized. Figure 3 presents the logo specially designed for the BEMLAB project – it presents a discretized surface of sphere, the boundary.

5.6.2. Data Input/Output format

While creating specification for the project it was decided that the calculated problems will be defined fully using one file format. It was also decided that the problem can be split among many files. Exchanging data with external files was a priority.

File format

Taking above into account the text file format compatible with Matlab M-files was chosen. The example matrix definition named matrix consisting of 2 rows and 3 columns with a marked comment is as follows:

```
matrix=[1 3 7; 4 6 8]; % comment (2,3)
```

The choosen file format can easily be converted into Scilab sci-format, by changing the comment string.

```
shell% sed -e 's/%/\/\/g' file_matlab.m >
file_scilab.sci
```

The converted file can then be read into scilab:

```
scilab-> exec('file_scilab.sci')
```

Furthermore a text file format is more user friendly. It simplifies tasks related with mesh creation. Text files can easily be edited and manipulated using standard tools such as text editors (e.g. vi) or line editors (e.g. sed, awk). It is also easer to write programmes which process data files in script languages like perl, python, ruby or using the shell interpreter.

5.7. Data format

The BEMLAB data format is based on the file format presented above. The base unit is the matrix. Mesh files are composed of many matrices with the strictly defined names. The huge advantage of the file format is its universality. Any number of domains, elements, sources, etc. can be defined for the calculated problem. The matrix definitions can be specified in files in any order. A basic example is presented in figure 4.

Figure 3: BEMLAB logo

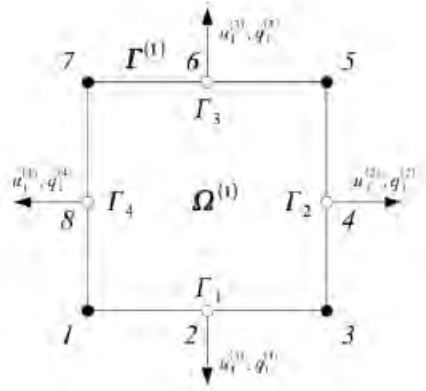


Figure 4: 2D boundary mesh

This is a two dimensional problem discretized with $J = 4$ boundary elements. Potential u_k^j and its normal derivative q_l^j were marked in the $I = 4$ interpolation nodes. One node, constant elements are used to interpolate potential and its normal derivative $K = L = 1$ – see equation (18). Geometry is discretized with 2 node linear boundary elements.

The mesh from figure 4 can be defined in BEMLAB data format in the following form:

```
nodes=[0,0; 5,0; 10,0; 10,5; 10,10; 5,10; 0,10; 0,5]; %

elementsGeom=[1,3; 3,5; 5,7; 7,1]; % 4
elementsU=[2; 4; 6; 8]; % 4
elementsQ=elementsU;

elementTypeGeom=['LineLinear'];
elementTypeU=['LineConst'];
```

```
elementTypeQ=elementTypeU;

dirichletElements=[ 1, 3 ]; % 2

dirichletBC=[0; 10]; % 2

neumannElements=[ 2, 4 ]; % 2

neumannBC=[0; 0]; % 2

angleCoefficients=[ 2, 4, 6, 8 ]; % 4
angleCoefficientValues=[ 0.5, 0.5, 0.5, 0.5 ]; % 4

methodType_1=['poisson'];
```

where:

- nodes – is the matrix which includes all nodes coordinates defined in the problem,
- elementsGeom – is the matrix which includes geometrical boundary elements definitions – contains indexes to nodes defined in nodes matrix,
- elementsU – is the matrix which includes definitions of elements interpolating potential $u = \varphi$ – contains indexes to nodes defined in nodes matrix,
- elementsQ – is the matrix which includes definitions of elements interpolating potential normal derivative $q = \partial\varphi/\partial n$ – contains indexes to nodes defined in nodes matrix, and usually is equal to elementsU matrix,
- elementTypeGeom – is the matrix which defines geometrical element type – here first order linear element is defined: ['LineLinear'],
- elementTypeU – is the matrix which defines type of element interpolating potential $u = \varphi$ – here first order linear element is defined: ['LineConst'],
- elementTypeQ – is the matrix which defines type of element interpolating potential normal derivative $q = \partial\varphi/\partial n$ – usually is equal to elementTypeU matrix,
- dirichletElements – is the matrix which defines elements which has Dirichlet Boundary Condition defined – here two elements 1 and 3,
- dirichletBC – is the matrix which defines Dirichlet Boundary Condition values (potential u) in consecutive elements defined in matrix named dirichletElements – here element 1 has the potential $u_1^{(1)} = 0$ defined in node 2 and element 3 has the potential $u_1^{(3)} = 10$ in node 6,
- neumannElements – is the matrix which defines elements which has Neumann Boundary Condition defined – here two elements 2 and 4,

- neumannBC – is the matrix which defines Neumann Boundary Condition values (potential normal derivative q) in consecutive elements defined in matrix named neumannElements – here elements 2 has the value $q_1^{(2)} = 0$ defined in node 4 and element 4 has the value $q_1^{(4)} = 0$ in node 8,

- angleCoefficients – is the matrix which contains indices of nodes which interpolate potential u and its normal derivative q and will have c_i coefficient value manually defined – see equation (18) – this matrix is auxiliary for this model,

- angleCoefficientValues – is the matrix which contains values of c_i coefficient for nodes defined in matrix angleCoefficients – here all c_i values are equal 0.5, where i = 1, 2, 3, 4. c_i = 0.5 is the default value therefore angleCoefficients and angleCoefficientValues matrices don't have to be defined in this case,

- methodType_1 – is the matrix which contains the name of an integral kernel which is used to set up BEM matrices: **A,B,F**.

The above problem defined in the presented file example_2d.m can be solved using BEMLAB software by issuing the following command:

```
% obem_solve -i example_2d.m -m 1234 -o
solution_output.m
```

The results are written to the output file solution_output.m.
The BEMLAB package also provides programs which can be used for post-processing tasks like drawing a plot.

5.8. BEMLAB architecture

This section presents only selected information about the BEMLAB software architecture.

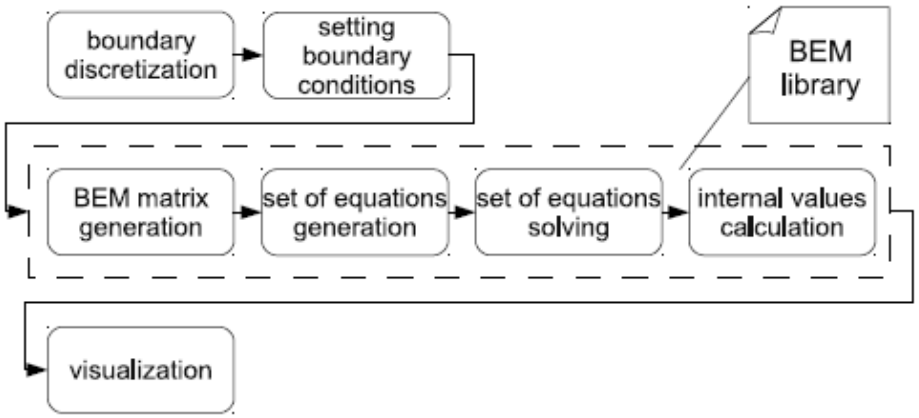


Figure 5: Activity diagram containing main activities taken while solving problems defined by Partial Differential Equations, which include preprocessing (first row), chosen numerical method calculations (second row), and finally postprocessing (third row)

Before creating the BEMLAB library architecture the mathematical description of the numerical method has been done (section 4). Among others use case models has been created. One of the diagrams which show the whole process of any problem modelling is the activity diagram presented in figure 5.

The diagram contains main activities taken during solving problems defined by Partial Differential Equations. The whole process is divided into three stages from the numerical software point of view. The following stages include basic BEM activities presented in the one domain model case for the simplicity, but extendible in the multi domain one:

1. Preprocessing – Includes mainly:
   - *boundary discretization* Γ of the problem domain Ω,
   - *setting boundary conditions* on the external boundary of the problem.

2. Numerical method calculations – Includes activities involving Boundary Element Method tasks, which mainly include:
   - *BEM matrix generation* **A,B,F** – equation (19),
   - *set of equations generation* – the left hand side **a** and the right hand side **b** matrices are generated using among others generated previously **A,B,F** matrices and boundary conditions – equation (20),
   - *set of equations solving* – this activity is the most time consuming and includes running the solver – as a result all unknown boundary values of potential *u* and its normal derivative *q* are known,
   - *internal values calculations* – this activity is optional and run only when needed, it uses the BEM engine and needs the data required for previous tasks and calculated boundary values of potential *u* and its normal derivative *q*.

3. Postprocessing – Includes tasks involving usage of results obtained with numerical method in the previous stage like:
   - *visualization.*

The basic activities of the BEM core included in the BEMLAB library are marked with a dashed line.

Another diagram which was created during modelling the BEMLAB software is the component diagram presented in figure 6.

The software was divided into: *lib* – the library, application – reference application and tests – testing module. The main logic is included in the library *lib*. As the BEMLAB software implements BEM comprehensively providing the multi physics package, the *lib* was divided into several components:

- *base* – Includes containers used accross the library, methods implementing required algorithms like matrix calculations or iterators, implements methods responsible for the data file format compatible with Matlab M-files,
- *bem* – Includes algorithms wich implement boundary element method. There are main algorithms implementing activities from the diagram 5, the equation (21), Greens functions, integration kernels calculations, boundary and interface conditions application, internal point calculations, etc.,
- *integration* – Includes required integration algorithms e.g. Gauss Quadrature,
- *solver* – Includes algorithms used for solving sets of linear equations (20): $\mathbf{a}\,\mathbf{x} = \mathbf{b}$,
- *auxiliary* – Includes auxiliary algorithms not available in Standard Template Library (STL) distributed with C++ but required by the BEMLAB software.

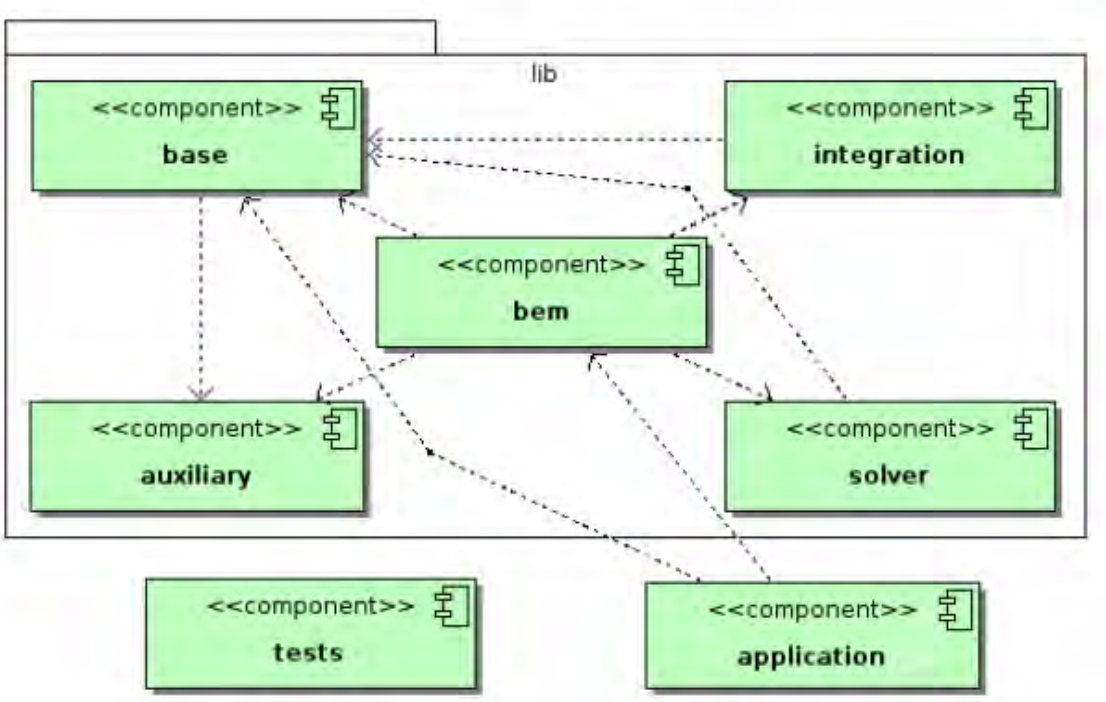


Figure 6: BEMLAB component diagram

The application component consists of several classes which use the library to provide its functionality for the end-user. Using application is the easiest way to proceede calculations by the end-user.

A very important is the test component. The development cannot be efficient without broad range of tests. It is especially important in the numerical software where a small change may have a big impact on the correctness of calculations. Tests allow to detect mistakes and bugs on the very early stage.

There are several type of tests. Some of them test particular methods and classes, where the others the library and the application as a whole – acceptance testing. The whole bunch of tests is run before every commit to the version control system.

CSparskit2

The longest task which is proceeded during BEM calculations is solving the set of linear equations as presented in diagram 5 and equation (20). Moreover the left hand side matrix a is dense in opposite to FEM where it is sparse and symmetric. Therefore fast algorithms known from FEM cannot be used.

However calculations can be fastened using GMRES algorithms. The Generalized Minimal Residual Method (GMRES) was proposed by Yousef Saad [7] in 1980s.

One of the GMRES implementations is the SPARSKIT2 package [19] by Yousef Saad. The source code is written in Fortran. It was essential to write a wrapper in C++ to make it use the BEMLAB containers implementing matrix format compatible with Matlab M-files. The C++ wraper was named CSparskit2 and is available at [18]. CSparskit2 uses BEMLAB base component (Fig. 6) and depends on SPARSKIT2 package.

5.9. The Diffuse Optical Tomography problem described by means of the baby head model

Diagnosing and controling head haemorhages in newborn infants especially premature babies, and woman breast tumour detection are the main areas of scientists intersest for Diffuse Optical Tomography application (DOT). DOT uses near infrared light which wavelength λ is usually from 760 to 830mm.

A three dimensional baby head model is presented in this section. Figure 7 presents the model of baby head.

The head is divided into three domains: $\Omega^{(1)}$ – scalp, $\Omega^{(2)}$ – skull and $\Omega^{(3)}$ – brain. Each domain $\Omega^{(r)}$, where r ∈ ⟨1, 2, 3⟩ differs in tissue optical parameters. Boundary Element Method requires only boundaries to be discretized, therefore only meshes for $\Gamma^{(1)}$, $\Gamma^{(2)}$ and $\Gamma^{(3)}$ boundaries have to be generated. The boundaries were discretized using six node quadratic triangle.

The NIR source characterized by the frequency modulation of source light intensity ω = 100MHz is used in this example. The NIR light source is modelled as the collimated point source placed $\frac{1}{\mu_s'}$ under the model surface.

This is the most accurate mapping of light source [9] as the light dispersion starts only when the light ray passes the $\frac{1}{\mu_s'}$ length. The point source is located inside $\Omega^{(1)}$ domain in the presented model.

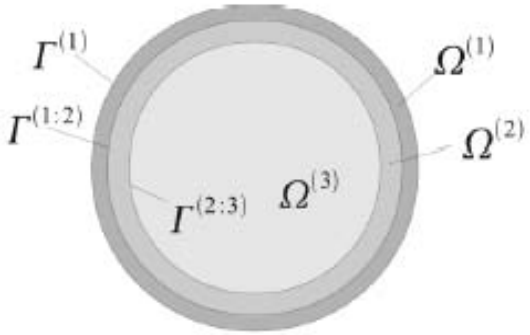


Figure 7: Schematic three layers model of baby head

All calculations were done using BEMLAB software. The visualization was done using BEMLAB programmes. The following figures present direct results of the modelled forward problem. Figure 8a) and b) presents the layout of photon density on the head surface $\Gamma^{(1)}$, c) and d) on the skull surface $\Gamma^{(1:2)}$ and e) and f) on the brain surface for the amplitude and the phase shift respectively.

All figures consist of two subfigures were the first one presents amplitude of photon density and the second one its phase – logarithmic scale was used. The yellow marker shows the placement of the source point. It is impossible to calculate such model analytically, because of a non regular geometry.

However it can be stated that the obtained results are correct based on the correct results obtained for the example geometries where analytical solution is known and calculated using the same process. Moreover the obtained photon density changes on the boundaries in the expected way – amplitude decreases when the distance from the source increases and the phase value increases when the distance from the source increases.

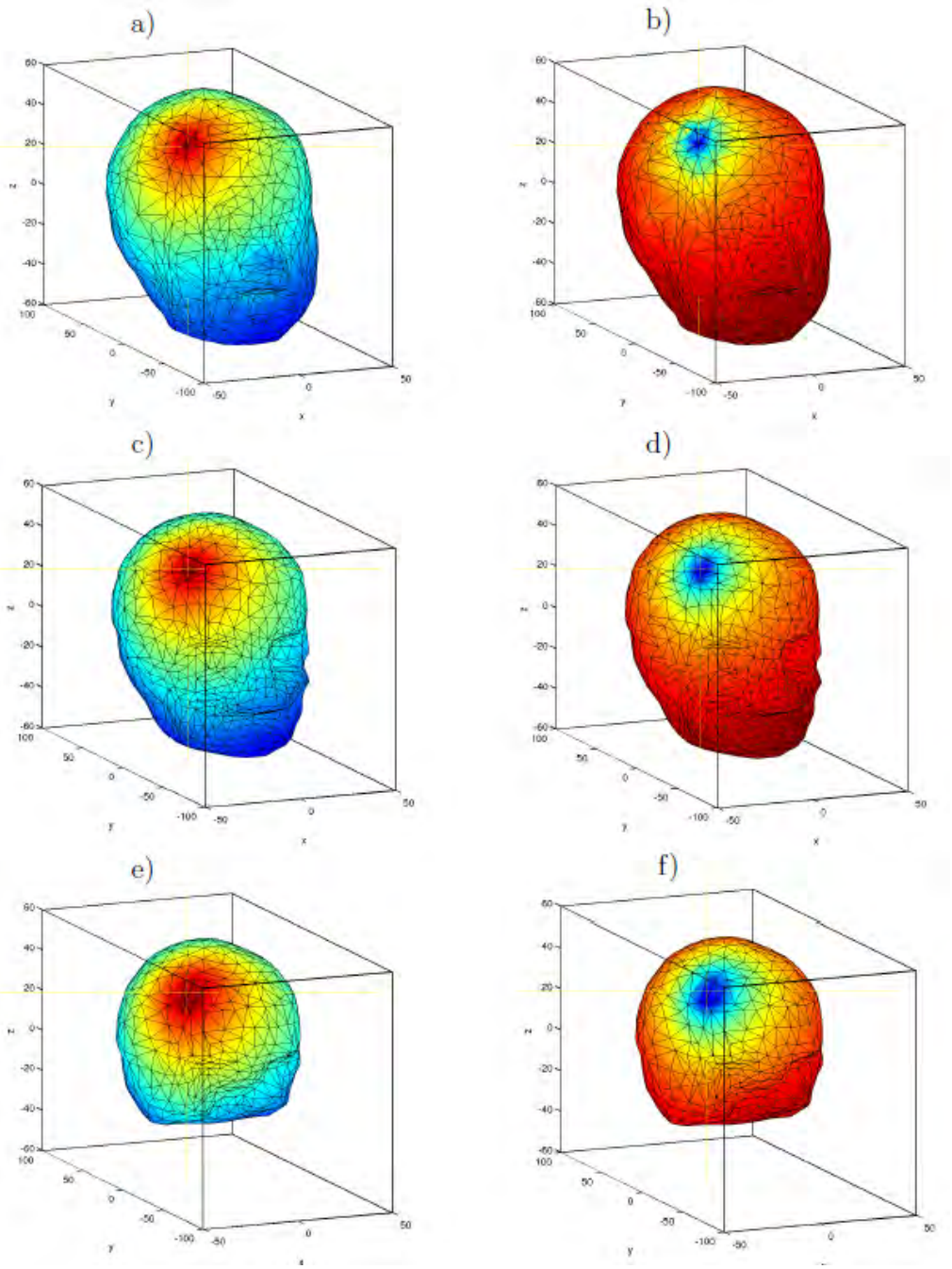


Figure 8: Photon density layout presented in logarithmic scale on: a) the head surface $\Gamma^{(1)}$, c) the skull surface $\Gamma^{(1:2)}$, e) the brain surface $\Gamma^{(2:3)}$; the left column presents the amplitude and the right column the phase shift

5.10. Summary

The BEMLAB software is designed to solve Diffuse Optical Tomography and Electrical Impedance Tomography problems. Among others it can also be used for solving electromagnetic problems. BEMLAB is protected by the open source license, which means that can be freely distributed (binaries as well as the source code). It has an objective architecture which eases modelling and development. It uses multi-threaded BEM algorithms which accelerate calculations on multicore or multiprocessor computers.

The BEMLAB software was designed to be an universal BEM package which implements various types of boundary elements, Partial Differential Equations. It can be used to calculate multi domain problems of any geometry and with any number of domains. It provides an easy to use data format compatible with Matlab m-files. There are also some auxiliary programmes for preprocessing and postprocessing provided.

The BEMLAB project aspires to be the platform for Boundary Element Method improvement. The projects created infrastructure allows to run the dispersed development. Now, BEMALB is a ready to use software for solving problems described by PDEs including tomography problems as it was presented.

Bibliography

1. S.R. Arridge, M. Cope, D.T. Delpy: The theoretical basis for the determination of optical pathlengths in tissue: temporal and frequency analysis, Physics of Medical Biology, Vol. 37, No. 7, 1992, pp. 1531-1560.
2. C.A. Brebbia: The boundary element method for engineers, Wiley, 1978.
3. K. Fogel: Producing Open Source Software: How to Run a Successful Free Software Project, Publisher: O'Reilly, 2005.
4. M. Fowler: UML Distilled: A brief guide to the standard object modeling language, Addison-Wesley, ed. 3, 2003.
5. J. Mackerle: Object-oriented programming in FEM and BEM: a bibliography (1990-2003), Advances in Engineering Software, Vol. 35, 2004, pp. 325-336.
6. W.H. Press, P.B. Flannery, S.A. Teukolsky, W.T. Vetterling: Numerical Recipes in FORTRAN: The Art of Scientific Computing: LU Decomposition and Its Applications, Cambridge University Press, 1992, pp. 34-42.
7. Y. Saad, M.H. Schultz: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM Journal on Scientific and Statistical Computing, Vol. 7, No. 3, 1986, pp. 856-869.
8. M. Schanz, O. Steinbach: Boundary element analysis: Mathematical aspects and applications, Springer-Verlag, Berlin, 2007.
9. M. Schweiger, S.R. Arridge, M. Hiraoka, D.T. Delpy: Finite element method for the propagation of light in scattering media: boundary and source conditions, Medical Physics, Vol. 22, No. 11, 1995, pp. 1779-1792.
10. M. Schweiger, S.R. Arridge: Finite element method for the propagation of light in scattering media: frequency domain case, Medical Physics, Vol. 24, No. 6, 1997, pp. 895-902.
11. J. Sikora: Boundary Element Method for Impedance and Optical Tomography, Oficyna Wydawnicza Politechniki Warszawskiej, 2007.
12. J. Starzynski: Laboratorium podstaw elektromagnetyzmu, Oficyna Wydwanicza Politechniki Warszawskiej, Warszawa, 2005, (in Polish).
13. T. Tarvainen and M. Vauhkonen and V. Kolehmainen and J.P. Kaipio and S.R. Arridge: Utilizing the radiative transfer equation in optical tomography, PIERS Online, http://piers.mit.edu, Vol. 4, No. 6, 2008, pp. 655-660.
14. O.C. Zienkiewicz, R.L. Taylor: The finite element method, Butterworth- Heinemann, ed. 6, 2000.
15. P. Wieleba, J. Sikora: Open Source BEM Library, Advances in Engineering Software, 2009, issn=0965-9978, Vol. 40, No. 8, 2008, pp. 564-569.
16. P. Wieleba, J. Sikora: Open Source Boundary Element Method Library for Diffusion Optical Tomography, Electrotechnical Review, Vol. II/2007, 2007.

17. BEMLAB homepage, http://bemlab.org/
18. CSparskit2 homepage, http://bemlab.org/csparskit2/
19. Sparskit2 homepage, http://www-users.cs.umn.edu/saad/software/SPARSKIT/
20. sparskit.html
21. MediaWiki homepage, http://mediawiki.org/