SCIENTIFIC ISSUES

JAN DŁUGOSZ UNIVERSITY
in CZĘSTOCHOWA

MATHEMATICS XXII

Scientific Editor
Andrzej Zbrzezny

2017

Secretary: **Katarzyna Domańska**
Computer Typesetting and Technical Editing: **Andrzej Zbrzezny**
Graphical Project of a Cover: **Sławomir Sadowski**
ISSN 2450–9302

Cooperating Reviewers

Piotr Artiemjew — University of Warmia and Mazury in Olsztyn
Grzegorz Bryll — University of Opole
Yevhen Chaplia — National Academy of Sciences of Ukraine, Lviv, Ukraine
Imed El Fray — Warsaw University of Life Sciences
Roman Frič — Slovak Academy of Science, Košice;
& Catholic University in Ružomberok, Slovak Republic
Roman Ger — Silesian University in Katowice
Katarzyna Grygiel — Jagiellonian University
Ján Gunčaga — Catholic University in Ružomberok, Slowak Republic
Jacek Jędrzejewski — Jan Długosz University in Częstochowa (retired)
Magda Kacprzak — Bialystok University of Technology
Zygfryd Kominek — Silesian University in Katowice
Mirosław Kurkowski — Cardinal Stefan Wyszyński University
Janusz Matkowski — University of Zielona Góra
Tomasz Połacik — University of Silesia, Katowice
Tadeusz Poreda — Łódź University of Technology
Yuriy Povstenko — Jan Długosz University in Częstochowa
Yaroslav P'yanylo — Ukrainian National Academy of Sciences
Paweł Róg — Częstochowa University of Technology
Grażyna Rygał — Jan Długosz University in Częstochowa
Robert Sochacki — University of Opole
Marcin Szpyrka — AGH University of Science and Technology
Oleg Tikhonenko — Cardinal Stefan Wyszyński University in Warsaw
Szymon Wąsowicz — The University of Bielsko-Biała
Władysław Wilczyński — University of Łódź
Wacław Zawadowski — Warsaw University

Contents

Part I
Education in Mathematics

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
11–22
DOI http://dx.doi.org/10.16926/m.2017.22.01

HOW HUNGARIAN TEACHERS THINK ABOUT COMBINATORICS AND ITS TEACHING

GABRIELLA AMBRUS, NÓRA KATALIN RÁCZ, AND ÖDÖN VANCSÓ

Abstract

Teachers' attitudes and beliefs have a fundamental influence on the way of teaching and the efficiency of learning. This topic was also brought up in the didactic research in recent years, in connection with mathematics teachers as well.– To the best of our knowledge, however, a study that investigates this specific area (combinatorics) was not included. As combinatorics has a special role in connecting the development of mathematical thinking and teaching different areas, our short-time Project (Didactics Grant of the Hungarian Academy: Complex Mathematics Education in the 21st Century) focused on the teaching of combinatorics in Hungary. Our aim was to examine some aspects of this field in connection with Tamás Varga's method and also teachers' beliefs towards combinatorics with the help of a questionnaire. The main results of this survey are discussed below.

1. Introduction

In the last fifty years of Hungarian mathematics education the ideas of Tamás Varga's "Complex Mathematics Education Experiment" have played an important role, and the results of the experiment were realized primarily in the first eight grades, and partly in grades 9 to12. Our research group's[1] goal is to rethink traditions based Tamás Varga's teaching methods, and to

• *Gabriella Ambrus* — e-mail: ambrusg@cs.elte.hu
Loránd Eötvös University in Budapest, Faculty of Science, Mathematics Institute.
• *Nóra Katalin Rácz* — e-mail: noriracz16@gmail.com
Loránd Eötvös University in Budapest, Faculty of Science, Mathematics Institute.
• *Ödön Vancsó* — e-mail: vancso.odon@gmail.com
Loránd Eötvös University in Budapest, Faculty of Science, Mathematics Institute.

[1]Ödön Vancsó, Gabriella Ambrus, Csaba Csapodi (ELTE Univ.), József Kosztolányi, Klára Pintér (Univ. Szeged), Judit Szitányi (ELTE Univ.), Eleonóra Stettner (Univ. Kaposvár); György Emese (Újpesti Bilingual Technical HS), primary and secondary mathematics teachers.

implement his ideas in secondary schools to a greater extent, first considering the teaching of combinatorics in grades 5 to 12[2].

Our choice for combinatorics as a topic had several reason. Firstly, combinatorics has a specific role in the teaching of mathematics, because it is important in the teaching of thinking methods, and it features in several areas of mathematics. Secondly, a number of combinatorial problems (e.g. ordering), even if at different levels, can be taught to different grades so combinatorics is particularly suitable for comparing the problem solving ability of different age groups see [12]. What is more, this topic also had priority in Tamás Varga's experiment and finally in his teaching method.

Several factors affect the effectiveness of teaching in school. for instance, the mathematics teacher's personal attitude concerning the subject (area) and its teaching has high priority. The inquiry into teachers' beliefs towards combinatorics is part of the research of our research group.

2. Theoretical background

2.1. **Tamás Varga and his teaching method.** After spending a short period working as a schoolteacher, Tamás Varga, a Hungarian mathematics and physics teacher, worked in the National Institute of Education in Budapest, then was a lecturer at Eötvös Loránd University (ELTE) between 1951 and 1967. His thoughts presented at UNESCO's Research Symposium on Mathematics Education (1962, Budapest) resonated well with the reform efforts initiated by the so-called Sputnik Shock in the Western World. The trend represented by him and his colleagues, set off in 1964 and summarized in their Complex Mathematics Teaching Experiment, can be taken as a kind of response to the weaknesses of the New Math.

Tamás Varga had excellent international contacts, among others with Hans Freudenthal. He published and gave lectures in German, English and French as well.

In Varga's method the *genetic approach* based on discovery and prior knowledge and experiments of the students was decisive, see [3].

In 1978 the new mathematics curriculum for primary schools (grades 1 to 8) was introduced on the grounds of Varga's above mentioned experiment, which resulted in a complete change in mathematics teaching see, [13].

Tamás Varga emphasized the importance of the teaching of combinatorics and probability for instance see, [24], [25], [26], [27]. While in the

[2]Now, according to our results in the short-time Project we obtained the opportunity to go on with the research of Tamás Varga's method and extend our research group within the Content Pedagogy Research Program of the Hungarian Academy – 2016–2020, Complex Mathematics Education in the 21st Century – Improving Mathematical Thinking Based on the Most Recent Research Results

curriculums before 1978 the teaching of combinatorics started at the age of 17 years, in the new concept it started as early as in the first class, at the age of 6 years.

In the new mathematics curriculum for primary schools (1978), combinatorics was not a separate unit because combinatorial thinking is basically important in the different subject areas of mathematics with a special role in probability, see [4]. Problems with more or less combinatorial content appeared at every grade in the primary school.

Although the activity of Tamás Varga and his colleagues has had a great impact on Hungarian mathematics teaching since the seventies of the last century, in the last decades this effect seems to be diminishing. The long-term aim of the work of our research group is the extensive investigation and right-on adaptation of Varga's method. First we had the opportunity to begin this work on the subject of the teaching of combinatorics within a didactics grant for a year of the Hungarian Academy of Sciences in 2015. In the relatively short time of our project we tried to prepare and realize our investigation, but the circumstances did not allow us to obtain precise results in some cases.

2.2. **Teacher's attitudes and beliefs in connection with their teaching.** In the definition of Richardson, see [18], "*attitudes and beliefs are a subset of a group of constructs that name, define, and describe the structure and content of mental states that are thought to drive a person's actions. Other constructs in this set include conceptions, perspectives, perceptions, orientations, theories, and stances. (p.3)".*

In a further distinction, concentrating now on a definition in the teaching, attitude means the "affective" side and belief can be considered as a *major construct of interest in studying teachers' ways of thinking and classroom practices (p.3).*

Richardson emphasizes *that the difference between the two terms remained somewhat unclear in the empirical literature (p.3).* For example, Pajares suggests that among others attitudes (and values, preconceptions, theories, and images) are really "beliefs in disguise" see [17].

Considering secondary mathematics teachers' domain specific belief systems, the term *belief is used as an individual's personal conviction*, see [6], we will use belief in the same meaning in the following – and the organisation of beliefs can be considered as a system of individual ideas (belief system). The teacher's ideas (beliefs) appear in the context of "attitude theory" see, [10] (Theorie von Einstellungen, see, [8]).

Hannula states that beliefs about mathematics should be examined in the context of cognition, emotion and motivation. He emphasizes the cognitive aspect of belief as he mention it in the cognitive domain in his three dimensional model see [11].

Although, according, for example, to Pajares see [17] a total separation between the professional knowledge and teaching beliefs is problematic, some authors mention that it is worth to analyze these areas separately see [19], [14].

Teachers' conceptions (beliefs) influence teaching practice. However, the specific relationship between teacher's beliefs and their teaching practice is not known see [15], [22].

It should be emphasized that the teachers' beliefs have a decisive influence on their students' beliefs and what is more, the image about mathematics is largely decided in the school see [10]. So, teachers' beliefs in connection with a mathematical subject (e.g. algebra or geometry) may have a decisive role in the relevant student's beliefs which also emphasize the importance of our investigation, although this question is not directly considered in the present research.

Teacher's ideas have been studied in many practical contexts, for example, on the use of modeling tasks see [16], [1] or on problem solving. It seems that teachers can think differently about mathematics when teaching mathematics in different areas (e.g calculus, geometry, probability or statistics) see [6]. There are teachers' belief research results related e.g. to the teaching of statistics and calculus [5], [6], [7], but the authors are unaware of the existence of such studies specifically on combinatorics.

3. Methodology

In this study, we do not examine teachers' beliefs about mathematics in general, only teachers' ideas about one area, i.e. combinatorics; although it is obvious that the teacher's basic orientation (e.g. static or process-oriented way of thinking, see [10] influences the teaching of this topic as well.

In order to obtain an image of Hungarian mathematics teachers' ideas concerning combinatorics, we prepared a questionnaire consisting of statements on four main issues with four statements each. The four main issues are the following: *1. How positive is the attitude towards combinatorics. 2. Combinatorics problems are difficult. 3. The formulas and standard problems are important in teaching combinatorics. 4. Knowing combinatorics is useful in other areas of life as well.*

The questionnaire – asking for grading statements – is in our opinion capable of measuring emotional attitudes (not only cognitive attitudes), even if only to a limited extent.

While filling out the questionnaire, the teachers had to classify the 16 statements belonging to four topics on a 5-level Likert scale.

In order to better appraise individual teachers' attitudes and orientation about teaching combinatorics, two extra tables were added to the previous 16 questions. In these tables teachers declared in the system of specific topics (Algebraic expressions, Equations, Functions, Sets, Logic, Graphs, Combinatorics, Coordinate geometry, Plane geometry, Sequences, Statistics, Number Theory, Word problems, Spatial Geometry, Trigonometry, Probability) how much they like teaching combinatorics and how successful they are in teaching combinatorics. When specifying the topics, we gave as narrow areas as possible, aiming to gain a more accurate picture. For example, graphs were not included in combinatorics (graphs are mentioned separately) and plane and spatial geometry were also handled separately.

When we assembled the questionnaire, it was important that

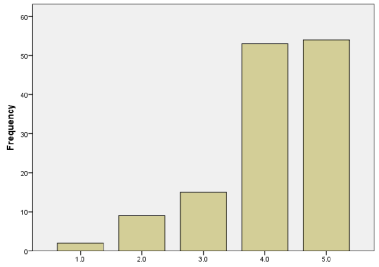
- The questionnaire should not be too long (since we wanted it to be filled out by many teachers)
- Statements should be kept in the shortest possible form and should be as clear as possible in order to get responses that can be evaluated easily
- Statements should be objective, wording should not be only positive (if possible, the wording should not influence the choice of the participant)
- All topics should have the same number of statements (so it can be evaluated easily)
- Preferably there should be statements of varying strength within one topic (so it can be evaluated easily)

The full questionnaire is available for viewing at the authors.

We analyzed the responses and examined some questions in connection with the results. In this study the following questions are in the focus:

- Do teachers like teaching combinatorics?
- Do teachers teach combinatorics willingly?
- How do they relate to some of the fundamental problems in teaching combinatorics (usefulness and problems of the subject, the question of teaching standard problems)?

3.1. **The conditions of the study.** The people who filled out the questionnaire (either on paper or electronically) were teachers who taught in Hungarian middle schools or secondary schools. They could fill out the test anonymously and voluntarily. Until the evaluation date, 156 teachers filled out the test, from which 136 were fully filled out, (31 male and 102 female, and 3 not declared). Analysis was done on these 136 questionnaires. This



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.0	2	1.5	1.5	1.5
	2.0	9	6.7	6.8	8.3
	3.0	15	11.1	11.3	19.5
	4.0	53	39.3	39.8	59.4
	5.0	54	40.0	40.6	100.0
	Total	133	98.5	100.0	
Missing	System	2	1.5		
Total		135	100.0		

FIGURE 1. Answers based on the statement "I like teaching combinatorics".

relatively small number set some limitations in our results. Most of the participants had been in teaching for at least 15 to 25 years.

3.2. **The mathematical background.** In the analysis, one important aspect was to assign an objective capability parameter to each teacher. To do so, an extended version of the Rasch model was used as a tool, which assigns a parameter to the individuals and also to the questions. This is important because this way we also find out which issues are more important and less important and which responses indicate similar capabilities.

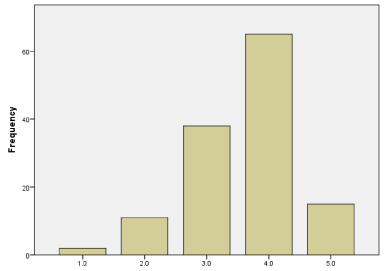
The parameters for the teachers were based on the four statements mentioned before, by which with the help of descriptive statistics and ANOVA we examined how the teachers' opinions related to one another's on the following: usefulness and difficulty of combinatorics, how positive their attitudes are towards the topic and how important they think standard problems are. Of the results of this study we pick only the ones relevant to this paper. (The detailed statistical evaluation of the teacher's answers is available for viewing at the authors in Hungarian.)

4. RESULTS

The results come from the short-time investigation and can be the basis of further and more precise investigations. The following part concentrates on the questions (see Methodology, section 3) and the limits of the evaluation of the teachers' answers are also considered. Looking at the statements' basic statistics we recorded the distribution of the answers.

For example, from the answers to the first question we see that most participants enjoy teaching combinatorics (Figure 1) and most of the time they think they are successful (Figure 2).

Based on the basics statistics we also found that participants think combinatorics is useful and not only the standard problems. The substantial majority of teachers think that combinatorics is not only for individuals



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.0	2	1.5	1.5	1.5
	2.0	11	8.1	8.4	9.9
	3.0	38	28.1	29.0	38.9
	4.0	65	48.1	49.6	88.5
	5.0	15	11.1	11.5	100.0
	Total	131	97.0	100.0	
Missing	System	4	3.0		
Total		135	100.0		

FIGURE 2. Answers based on the statement "I teach combinatorics successfully".

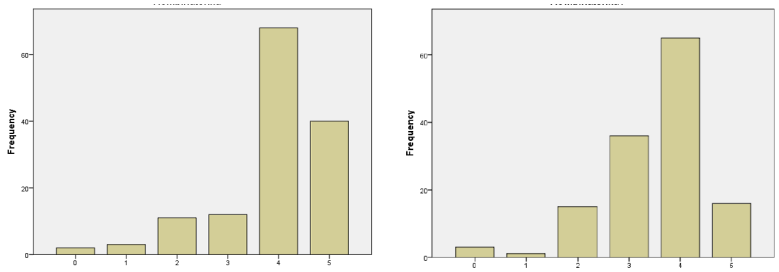


FIGURE 3. How willingly (left figure) and how successfully (right figure) mathematics teachers teach combinatorics "in the context of other areas".

who excel in mathematics, but it is important for every pupil. This is interesting because the majority of the teachers (even though less than before) think that if someone is not good at problem solving, then he/she will not be able to handle combinatorics problems.

Furthermore, we examined how much teachers like to teach combinatorics, compared to 16 other areas of mathematics (see earlier).

It turned out that teachers ***enjoy teaching the most*** *the following four areas: equations, functions, sequences and algebraic expressions, while the four areas which* ***they least enjoy teaching*** *are graphs, logic, trigonometry and probability.* Although we should note that the average was above 3 on the scale of 0-5 (5 is the best rate) for every topic.

The four areas which are ***taught most successfully*** *are equations, functions, sets and statistics, while the ones which are* ***taught least successfully*** *are trigonometry (the only one with an average of below 3), graphs, logic and coordinate geometry.*

Looking at combinatorics compared to other areas of mathematics (Figure 3) the picture is a little different to the one on Figures 1 and 2 where

the teachers had to make statements about combinatorics alone. They do not like it so much compared to other topics but the successfulness of their teaching of combinatorics is considered similarly as before.

We found the following results using the modified *Rasch model in our* four areas.

a) How important formulas and standard problems are in combinatorics: In the course of the person parameters fit test several teachers stood out, and the person- item map show that statements 4 *(It is good in teaching combinatorics that the standard problems get the emphasis.)* and 8 *(It is good for the student to realize clearly which type of combinatorial task is to be solved (permutation, combination or variation), this way he/she can easily find the appropriate term)* stand out of the model. However, statement number 4 is the one which *divides the teachers:* The item's characteristic curves show the same. We can also see that *there is strong agreement among the teachers that the primary goal of teaching combinatorics is not the teaching of the standard problems.*

b) How useful the teachers think combinatorics is: The items fit well into the model, statement 15 stood out the most: *how important combinatorics is in terms of model development.* Item parameters are homogeneous according to the Andersen test, but the person parameters are not, according to the Wald test. The person parameter fit test show that teachers fit better than at the previous area (attitudes towards standard problems), but there are some outstanders here too.

We can see on the person- item map that the items are shifted to the left as *the vast majority of the teachers find the teaching of combinatorics useful.* We can see at the item parameter fit and also on the map, that question 15 does not fit perfectly into the model, and also it can be observed that the individuals do not fit in either.

c) How difficult teachers think combinatorics problems are: While examining the item parameters the fit was not perfect, but all four items were usable. The Andersen test did not show that the items are homogeneous, it left out two statements in the Rasch test because of inappropriate responses, the Wald test is not interpretable either. It can be concluded from the person parameter test and also from person and item maps that several teachers do not fit the model and neither do two statements.

We can see on the person and item maps that *most teachers do not agree with statement 7 (combinatorics is mostly for people who excel in mathematics). Statement 12 (combinatorics problems are often difficult to teachers) is the most divisive statement in this area.*

d) Teachers' attitudes towards combinatorics: Interestingly, the first statement (*how much the teacher likes combinatorics*) stood out the most from the rest. But even this statement fell under the usable category.

The Wald test and the Andersen likelihood test could not be used again as it was in the case of the items 1 and 2. At the same time it cannot be said based on items 3 and 4 either that the item or person parameters would be homogeneous. It can be concluded from the person and item map that *the vast majority of teachers have a positive attitude towards combinatorics because for statement 1 most participants of the test gave high values. Statement 11 (whether the teacher teaches combinatorics successfully), split the teachers and many gave low scores. They similarly gave only few points for statement 2 (how much they bring combinatorics problems into their other lessons).*

5. Discussion

It can be said that the majority of responding teachers have a positive attitude towards combinatorics. They teach combinatorics willingly, but the extent of their willingness varies compared to the other topics. At first sight this may be because of the appearance of the more popular topics as well as the separation of the topics. However, it can be seen that although the teaching of graphs is less popular in Hungary, and this area has been listed as a separate topic, it did not improve further the popularity of combinatorics. It may be that several teachers did not consider graphs as part of combinatorics and this way mentioned separately the topic could not influence the popularity of combinatorics.

In case of combinatorics and also of other topics, the success of teaching does not reach the extent to which they like to teach these topics – according to the responses to the second table of the survey. Combinatorics was not among the best here either.

The opinion about "Teaching formulas and standard problems" divides teachers but most of the teachers consider "learning combinatorics" useful; this can be partly related to the opinion which was chosen by most of them, "the basic aim of teaching this topic is not the teaching of standard problems".

Although it is generally believed by the teachers that combinatorics is not only for those with better performance, still this topic is often labeled as complicated.

Due to the circumstances of completing the survey (i.e. the respondents voluntarily presented their opinions on combinatorics) the responses cannot be considered representative among Hungarian mathematics teachers. In spite of the positive attitude, our findings pointed at several contradictory

(or seemingly contradictory) phenomena in connection with teachers' beliefs and attitudes towards combinatorics and its teaching, which may have their causes mainly in the present state of Hungarian combinatorics teaching. Comparative (similar) and more precise studies among teachers in other countries would be needed to examine this statement.

Acknowledgement. This article was supported by the 2015 Methodological Grant of the Hungarian Academy of Sciences.

References

[1] G. Ambrus, Ö. Vancsó. *Ist eine schnelle tiefgehende (und nachhaltige) Änderung in der Vorstellung von Mathematiklehrern möglich? - Reflexion der Erfahrungen eines Fortbildungskurses im Bereich der mathematischen Modellierung.* TMCS, 13 (1), (2015) 1–20.

[2] H. Burkhardt, – with contributions by H. Pollak. *Modelling in Mathematics Classrooms: reflections on past developments and the future.* ZDM, (38) (2006), 178–195.

[3] E. Deák. *Die besondere Verflechtung der mathematischen Forschung, des Mathematik-Unterrichts und der Mathematikdidaktik Ungarns im 19. und 20. Jahrhundert.* Mitteilungen der GDM, Nr. 74 (2002)

[4] I. Eglesz I, Cs. Kovács, J. Radnainé Szendrei, V. Sztrókayné Földvári. *Matematika 5. Tanári Kézikönyv*, Tankönyvkiadó, Budapest, (1981).

[5] A. Eichler. *Statistics teachers and Classroom practices.* In. C. Batanero, G. Burril & C. Reading (Eds.). Teaching statistics in school mathematics-challenges for teaching and teacher education, New ICMI study series, Vol. 15, Heidelberg/New York: Springer, (2011), 175–186.

[6] A. Eichler, R. Ehrens. *Domain-specific belief systems of secondary mathematics teachers.* In. Pepin, Birgit et al. (Eds.). From beliefs to dynamic affect systems in mathematics education. Exploring a mosaic of relationships and interactions. Cham: Springer, (2015), 179–200.

[7] R. Ehrens, A. Eichler. *Teachers' curricular beliefs referring to calculus.* In. Proceedings of the 12$^{\text{th}}$ International Congress on Mathematics Education (ICME), Seoul, Korea (2012).

[8] F. Furinghetti, E. Pehkonen. *Rethinking Characterizations of Beliefs.* In. G. Leder, E. Pehkonen, G. Törner (Eds.). Beliefs: A hidden variable in mathematics education? Dordrecht: Kluwer Academic Publishers, (2002), 39–57.

[9] M. Glaymann, T. Varga. *Zwischen unmöglich und sicher: Exemplarische Kapitel der Kombinatorik, Wahrscheinlichkeitsrechnung und Simulation im Unterricht der Sekundarstufe I.* Freiburg: Herder (1975).

[10] S. Grigutsch, U. Raatz, G. Törner. *Einstellungen gegenüber Mathematik bei Mathematiklehrer,* Journal für Mathematik-Didaktik, (19) (1998) 3–45.

[11] M. S. Hannula. *Exploring new dimensions of mathematics-related affect: embodied and social theories.* Research in Mathematics Education 14 (2), (2012) 137–161.

[12] A. Hoffmann. *Vergleich der Problemlösefähigkeit von Primar-und Sekundarstufenschülern im Hinblick auf Lerneffekte während eines Interviews*, Journal für Mathematik-Didaktik, 26 (3/4), (2005), 176–199.

[13] S. Klein. *The Effects of Modern Mathematics*, Akadémiai Kiadó, Budapest (1987).

[14] S. Kuntze, L. Zöttl. *Auf Aufgaben bezogene Überzeugungen und übergreifende Beliefs von Lehramtsstudierenden*, Beiträge zum Mathematikunterricht, (2007).

[15] K. Maaß. *How can teachers' belief affect their professional development?* ZDM (43) (2011) 573–586.
[16] K. Maaß, J. Gurlitt. *Designing a Teacher-Questionnaire to Evaluate Professional Development about Modelling*, Conference-Volume der CERME, (2009), Lyon.
[17] F. M. Pajares. *Teachers' Belief and Educational Research: Cleaning up a Messy Construct.* Review of Educational Research, 62(3), (1992) 307–332.
[18] V. Richardson. *The Role of Attitudes and Beliefs in Learning to Teach,* In. J. Sikula (Ed.).Handbook of research on teacher education,second edition, New York: Macmillan. (1996) 102–119
[19] L. Shulman. *Those who understand: Knowledge growth in teaching.* Educational Researcher, 15(2), (1986) 4 -14.
[20] P. Szebenyi (Ed.). *Az általános iskolai nevelés és oktatás terve 1978, [Mathematics syllabus for primary school, 1978],* Országos Pedagógiai Intézet, Budapest, (1978).
[21] J. Szendrei. *In memory of Tamás Varga (2007)*
`http://www.cieaem.org/?q=system/files/varga.pdf`
[22] A. G. Thompson. *Teachers' beliefs and conceptions: a synthesis of the research Handbook of research on mathematics teaching and learning,* A. Douglas (Ed.) Grouws, Macmillan Publishing Company, (1992) 127–146.
[23] G. Törner. *Mathematical beliefs – A search for a Common Ground.* In. Leder G., E. Pehkonen, G. Törner (Eds.) Beliefs: A Hidden variable in Mathematics Education? Dordrecht: Kluwer. (2002) 73 – 94.
[24] T. Varga. *Mathematics Education in Hungary Today*, Educational Studies in Mathematics 19 (1988), 291–298.
[25] T. Varga, M. Dumon. *Combinatoire Statistiques et Probabilités de 6 á 14 ans. Fiches de travail.* Guide at Commentaires. O.C.D.L. Paris (1973).
[26] T. Varga. *Combinatorics and probability for young children*, I. Sherbrooke Mathematics Project. Canada, University of Sherbrooke, (1967).
[27] T. Varga. *Notes on mathematics reform in Hungarian schools. Pilot work in Budapest. Higher mathematics in lower grades.* Prepubescent Mathematical Learning. New York, London, Paris, Gordon and Breach Science Publishers, (1969).

Received: April 2017

Gabriella Ambrus
Loránd Eötvös University in Budapest
Faculty of Science, Mathematics Institute
H-1117 Budapest, Pázmány P. s. 1/c
E-mail address: `ambrusg@cs.elte.hu`

Nóra Katalin Rácz
Loránd Eötvös University in Budapest
Faculty of Science, Mathematics Institute
H-1117 Budapest, Pázmány P. s. 1/c
E-mail address: `noriracz16@gmail.com`

Ödön Vancsó
Loránd Eötvös University in Budapest
Faculty of Science, Mathematics Institute
H-1117 Budapest, Pázmány P. s. 1/c
E-mail address: vancso.odon@gmail.com

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
23–29
DOI http://dx.doi.org/10.16926/m.2017.22.02

GEOGEBRA AND PROBLEMS OF TANGENT CIRCLES

MARTIN BILLICH

Abstract

There are many mathematical programmes for teaching and learning geometry from middle school to the university level. Dynamic software package Geogebra can help students to explore and understand more concepts in geometry on their own. In this paper we are concentrating in one of the most interesting and less known two-dimensional transformations like inversion with respect to a circle supported by GeoGebra. The use of this transformation through GeoGebra makes possible a number of elegant solutions to classical construction problems in geometry. Contribution of this paper is presentation of some problems which require the construction of the circle tangent to given circles.

1. Introduction

By computers with dynamic geometry software we are able to solve problems which are difficult to solve by classical approach. GeoGebra is a multiplatform mathematical software that is simple to use and which is successfully applied at all levels of education in geometry, algebra and calculus. The basic idea of GeoGebra's interface is to provide two presentations of each mathematical object in its algebra and graphics windows. If we change an object in one of these windows, its presentation in the other one will be immediately updated. The main advantages of using GeoGebra are:

- GeoGebra offers easy-to-use interface, multilingual menus, commands and help.
- The use of GeoGebra in the classroom does not require any prior training for students. Students can learn how to use the main features directly during problem-solving activities and exercises.
- In GeoGebra environment students can manipulate variables easily by simply dragging free objects around the plane of drawing, or by using sliders and, in fact, students can see how the dependent objects will be affected.

• *Martin Billich* — e-mail: martin.billich@ku.sk
Catholic University in Ružomberok.

- The use of GeoGebra allows students to explore a wider range of function types, and provides students to make the connections between symbolic and visual representations.
- The algebra input allows the user to generate new objects or to modify those already existing, by the command line.
- The worksheet files can easily be exported or published as web pages.

One of the transformations in GeoGebra environment is the so-called *inversion* (or "*reflection about circle*"). This map was first introduced by J. Steiner about 1830. The inversion have many properties in common with line reflection. An important property of line reflection is that it preserves basic geometric classifications, straight lines go into straight lines and circles go into circles. Although the inversion preserves the class of lines and circles but can transform a line into a circle and a circle into a line. This and other exceptional properties of inversion serve as a foundation for its effectiveness in solving various problems in geometry.

In next section we discuss the main characteristics of inversion as well as some properties to be used throughout the paper. The use of inversion allows us to develop a unified method of solution for many problems in elementary geometry, especially those concerning construction of circles tangent to one or several circles and, supported by GeoGebra.

2. The Inversion

In plane, suppose ω is a circle with center O and radius R. We call the transformation that sends an arbitrary point P distinct from O into point P' lying on the ray $\overrightarrow{OP}$ such that $|OP| \cdot |OP'| = R^2$ the *inversion relative to circle* ω. The inversion relative ω will be also called the inversion with center O and degree R^2 and ω will be called the circle of inversion.

In this paper we shall fix $\mathcal{I}$ as the inversion relative to circle of radius R with center O. First, we will establish the following basic properties of inversion $\mathcal{I}$.

- Inversion is a transformation from $\mathbb{R}^2 \backslash \{O\}$ onto itself.
- Inversion is a bijection of order 2, that is if P' is the image of P, then P is the image of P'.
- Points on the circle of inversion stay fixed.
- Points inside [outside] of the circle of inversion are moved outside [inside].
- A line l going through the center of inversion is invariant under the inversion.
- Let $P \neq O$ be a point inside circle ω of inversion. To determine the position of P' on the ray $\overrightarrow{OP}$, draw a chord through P perpendicular

to OP meeting ω at T and S. Then the tangents to ω at T and S meet at the inverse point P'.

- If P is a point outside ω, then the two tangents from P to ω determine the chord TS and, the intersection between TS and the ray $\overrightarrow{OP}$ gives the inverse point P'.
- Suppose the points P and Q are different from each other and from the point O and, that the points O, P and Q are noncollinear. If $P' = \mathcal{I}(P)$ and $Q' = \mathcal{I}(Q)$, then the triangles OPQ and $OQ'P'$ are similar. That is, $\angle POQ \cong \angle Q'OP'$.

The following properties regarding the inversion of lines and circles hold (for details, see [2, 3, 6]):

- The inverse of any line l not through the center O of inversion, is a circle through O (minus the point O itself), and the diameter through O of this circle is perpendicular to l.
- The inverse of any circle through the center O of inversion (with O omitted) is a line perpendicular to the diameter through O, that is a line parallel to the tangent at O to the circle.
- The inverse of a circle not passing through the center O of inversion is a circle not passing through O. In particular, every circle orthogonal to the circle of inversion is its own inverse.
- Under the inversion the angle between lines [circles] is equal to the angle between their images.
- The angle between a circle and a line is equal to the angle between the images of these figures under the inversion.

3. Problems of tangent circles

In this section we will illustrate the type of problems that can be solved by method of inversion relative to circle in the GeoGebra environment. It is obvious that all of the construction problems of this section could be solved by students in the classroom with the help of a ruler and a compass too.

Now consider two problems (see also [1, 2]) which require the construction of circles tangent to several circles.

Problem 1. Construct all circles which are tangent to two given circles c_1, c_2 and pass through a given point O, lying outside c_1 and c_2.

Suppose c is one of the desired circles. Let $\mathcal{I}$ be an inversion with center O. Then $\mathcal{I}$ maps the circles c_1, c_2 into the circles c_1', c_2' and the circle c into a common tangent c'. It is obvious that the solution to the problem are circles which are the images of the common tangents of the circles c_1', c_2' under the inversion $\mathcal{I}$. Since in generally there are four of these tangents (in figure 1 they are shown by dotted lines), the problem has four solutions.

All the steps above may be demonstrated in the environment of software GeoGebra.

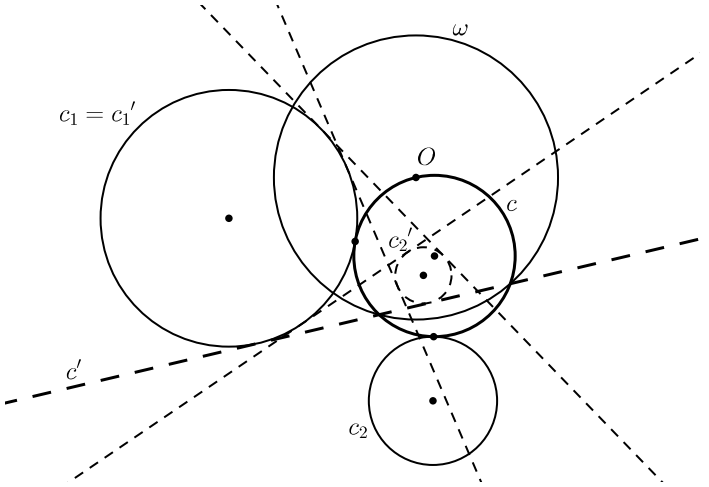


Figure 1

In our case a circle c_1 is orthogonal to the circle of inversion ω and therefore c_1 is invariant under this inversion ($c_1 = c_1'$).

Problem 2. (*Problem of Apollonius*). Construct all circles which are tangent to three given circles c_1, c_2 and c_3 of radii r_1, r_2 and r_3 around the points O_1, O_2 and O_3, respectively.

The three-circle case generally has eight solutions. Number of solution circles depend on the arrangement of the given circles. We shall present solutions to this problem for two different position of the circles c_1 and c_2:

A. The circles c_1 and c_2 are tangent (externally) at a single point P.
Solution. Suppose the circle c of radius r and centered at O is one of the desired circles. Let $\mathcal{I}$ be an inversion with center P and radius R such that the circle of inversion intersect the circles c_1 and c_2 in pairs of distinct points or/and the circle c_3 is orthogonal to c. The inversion $\mathcal{I}$ maps the circles c_1 and c_2 into a pair of parallel lines c_1' and c_2', and the circle c_3 into a circle c_3'. In the case of orthogonality of circles c_3 and c, the circle c_3 is invariant under this inversion ($c_3 = c_3'$). The circle c is taken by the inversion $\mathcal{I}$ into a circle c', which is tangent to circle c_3' and to both the parallel lines c_1' and c_2'. In this way, the solution of our problem has been reduced to a simpler construction problem: to construct all circles tangent to a given pair of parallel lines and to a given circle. Since in generally there are four of these tangent circles (in figure 2 they are shown by dotted lines), the problem has four solutions. As shown in figure 2, all the previsious steps

may be demonstrated in the GeoGebra environment. With the help of this software, by dragging the center O_3 of given circle c_3 around the plane of drawing, we can see how the solution circles will be changed.

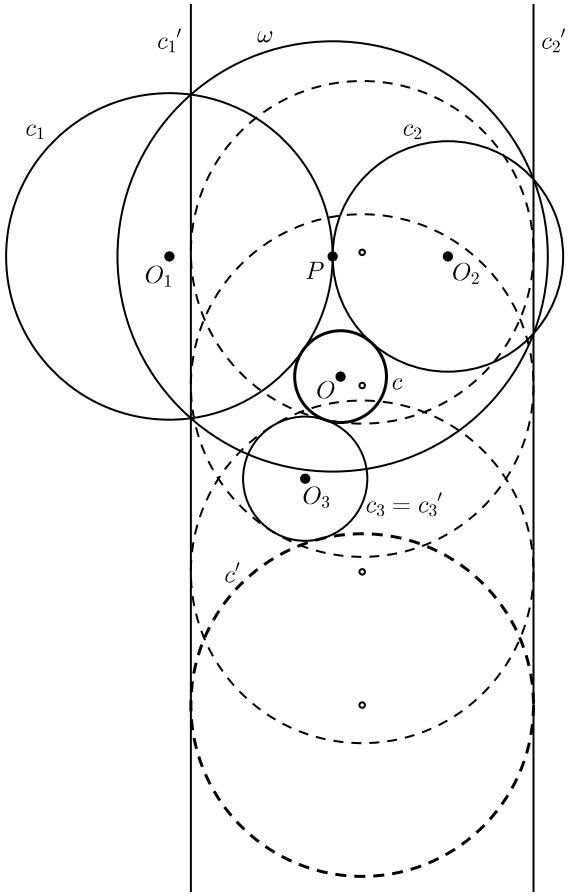


Figure 2

B. The circles c_1 and c_2 have no points in common (and c_3 is a circle outside c_1 and c_2).

Solution. Suppose the circle c of radius r and centered at O is one of the desired circles (figure 3). We connect the segment O_1O_2 from the centers of the circles c_1 and c_2 and draw circles of radii $r_1 + d$, $r_2 + d$, $r_3 + d$ with centers O_1, O_2, O_3, respectively, where

$$d = \frac{|O_1O_2| - (r_1 + r_2)}{2}.$$

We denote the constructed circles by $\overline{c}_1$, $\overline{c}_2$, $\overline{c}_3$ respectively (in figure 3 they are shown by dotted lines). Let $\overline{c}$ be the circle concentric with a circle c of radius $\overline{r} = r - d$. It is obvious that if we can construct the circle $\overline{c}$, we can easily construct the circle c. It is easy to see that $\overline{c}$ is tangent to the circles $\overline{c}_1$, $\overline{c}_2$, $\overline{c}_3$. The circles $\overline{c}_1$, $\overline{c}_2$ are constructed so that they are tangent to one another at some point P. This reduces it to a case **A**.

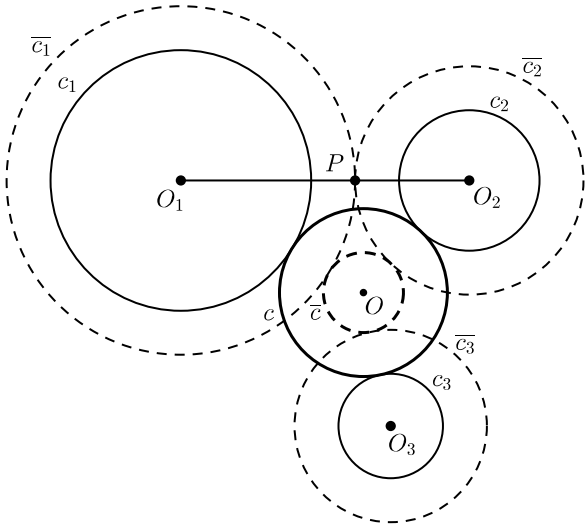


Figure 3

We leave the finalization of solution of this problem to the reader and suggest that the reader verify that the pair of the circles c_1 and c_3 or c_2 and c_3 could be used in place of the pair c_1 and c_2 in the above construction. Using the software GeoGebra it could be obtained all eight solution circles (see figure 4, [4]).

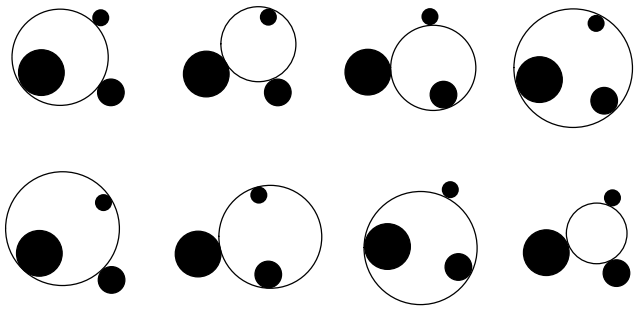


Figure 4

4. Conclusion

In this paper, some construction problems in plane geometry which require the construction of the tangent circles, by the inversion relative to circle and with the help of software GeoGebra, are introduced. The main strategy of inversion is to transform a given tangency problem into another problem of tangent circles that is simpler to solve. Subsequently, the solutions to the original problem are found from the solutions of the transformed problem by undoing the transformation. The usefulness of inversion can be increased significantly by resizing of given circles (see also [1, 5]). One usually inverts in a circle centered at a point where given circles intersect, since these will all be sent to lines. Quite often, only the center of inversion matters, and not the radius of the circle. Inversion, supported by dynamic geometry software like GeoGebra, is useful for solving various problems in elementary geometry.

References

[1] I. Ya. Bakel'man, *Inversions*, Chicago, IL: University of Chicago Press, 1974.
[2] M. Billich, *Circle inversion and problems on tangent circles*, In: Proceedings of International Congress IMEM 2009, Catholic University in Ružomberok, 2009, 492-496.
[3] H. S. M. Coxeter, *Introduction to geometry*, John Wiley & Sons., 1961.
[4] D. Gisch; J. M. Ribando, *Apollonius' Problem: A Study of Solutions and Their Connections*, American Journal of Undergraduate Research 3, 2004, 15–25.
[5] Z. Sklenáriková, *K metódam riešenia Apolloniovej úlohy*, In: Matematika v proměnách veku III, Edícia Dějiny matematiky, Praha, 2004, 45-55.
[6] Z. Sklenáriková, J. Čižmár, *Elementárna geometria euklidovskej roviny*, FMFI UK Bratislava, 2002.

Received: July 2017

Martin Billich
Catholic University in Ružomberok,
Department of Mathematics, Faculty of Education,
Place A. Hlinka 56/1, 034 01 Ružomberok, Slovakia
E-mail address: martin.billich@ku.sk

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
31–45
DOI http://dx.doi.org/10.16926/m.2017.22.03

APPLICATIONS OF PYTHON PROGRAMS IN SOLVING OF EQUATIONS BASED ON SELECTED NUMERICAL METHODS

MARCIN ZIÓŁKOWSKI, LIDIA STĘPIEŃ, MARCIN RYSZARD STĘPIEŃ, AND ARTUR GOLA

Abstract

In this paper we present the mathematical background of the four most used numerical methods of solving equations and few examples of Python applications that find the approximations of the roots of the given equations. We also compare the exact and approximate solutions of polynomial equations of third degree. Exact solutions are obtained with usage of Cardano formulae by the help of Mathematica environment, the approximate ones – based on the selected numerical methods by the help of applications written in Python language.

1. Introduction

From the mathematical point of view, it is always very important to find the exact solution of a given equation. Unfortunately, the effective algorithms exist only for a small group of equations. For example, we can solve only polynomial equations whose degree is less than five. But even in this case, the obtained formulae may be very complicated and they cannot be used in practice, especially in the case of third or fourth degree polynomial equations. For the other groups of equations, that contain in addition logarithmic, exponential or trigonometric functions, it is usually impossible to find exact solutions. In such situations, we use numerical methods to find approximations of solutions. Of course, almost all numerical methods have

• *Marcin Ziółkowski* — e-mail: marcin_ ziolkowski@sggw.pl
Warsaw University of Life Sciences.
• *Lidia Stępień* — e-mail: l.stepien@ajd.czest.pl
Jan Długosz University in Częstochowa.
• *Marcin Ryszard Stępień* — e-mail: mstepien@tu.kielce.pl
Kielce University of Technology.
• *Artur Gola* — e-mail: a.gola@ajd.czest.pl
Jan Długosz University in Częstochowa.

their own limitations that relate to the properties of the function appearing in the equation. Moreover, we usually must know in which intervals the roots of a given equation are. On the other hand, we luckily may use many different methods in the process of finding the approximations. The most important methods are: bisection, secant method, tangent method and iteration method [1]. Computations that are made in this case are also very laborious because of many loops appearing. Therefore, we should use computer programs. One of the most effective programming language that let write short and simple applications that solve mathematical problems is Python language. This language additionally has no limitation of the introduced numbers [4].

All analysed numerical methods are based on Bolzano – Cauchy theorem:

Theorem 1. *Let $f(x)$ be a function which is continuous on bounded interval $[a, b]$. If $f(a) \cdot f(b) < 0$, then there exists in interval (a, b) at least one root of the equation $f(x) = 0$.*

2. Bisection method

Now we discuss the first and simplest method which is called bisection. Let f be a continuous function defined on the interval $[a, b]$ and let $f(a) \cdot f(b) < 0$ what means that the sign of the values of this function changes in interval $[a, b]$. We will look for zero of the function f (or the root of an equation $f(x) = 0$). The idea of bisection is connected with the division the analysed interval onto two identical parts using the midpoint of an interval. Each time we choose this part of an interval where the value of the function changes the sign. The cut is continued until we get the exact solution or the required precision of the approximation is reached. If the function f fulfils the assumptions mentioned above, then the root of an equation exists and the bisection method can be used for finding it. In the next steps we determine the point x_0, which is a center of an interval $[a, b]$, taking: $x_0 = \frac{a+b}{2}$. Next, we count the value of the function at this point. If the $f(x_0)$ is sufficiently close to 0, i.e. $|f(x_0)| < \epsilon$, then x_0 can be accepted as a proper approximation of the root of an equation and the algorithm ends. In the other case, we take a new interval for searching the root – it will be the interval $[a, x_0]$ or $[x_0, b]$ depending on in which interval the function f changes the sign at its endpoints. The algorithm is repeated from the beginning until required precision is reached.

Example 1. *Find the approximation of solution of the equation $x^3 - 17x - 13 = 0$ in the interval $[4, 5]$*

We will use the Python application that is written on the base of the bisection method. The program is listed below.

```
BISECTION METHOD - PROGRAM IN PYTHON LANGUAGE
def fun(x):
    return x*x*x-17*x-13
def main():
    a=float(input("Enter the left endpoint of interval: "))
    b=float(input("Enter the right endpoint of interval: "))
    precision=float(input("Enter precision of solution: "))
    x_0=(a+b)/2
    while abs(fun(x_0)) >= precision:
        x_0=(a+b)/2
        if fun(x_0)==0:
            break
        elif fun(a)*fun(x_0)<0:
            b=x_0
        else:
            a=x_0
        print(x_0)
    print("Approximation of solution is:",x_0)
main()
```

Analysed equation has three real solutions. For example, the only positive solution can be found in the interval $[4, 5]$. The exact solution is equal to (*Mathematica environment*):

$$x_0 = \frac{\sqrt[3]{\frac{1}{2}\left(117 + \sqrt{45267}i\right)}}{\sqrt[3]{9}} + \frac{17}{\sqrt[3]{\frac{3}{2}\left(117 + \sqrt{45267}i\right)}}.$$

As it is seen the solution has a very complicated form consisting complex numbers and we are not able to find even the approximation of it. Using the Python program we find an approximation of the chosen precision (the precision is understood as the difference between the value $f(x_0)$ and zero). The sequence of the approximations of the above solution is presented in Table 1. Precision of the approximation is 10^{-8}.

The advantage of using the bisection method is its simplicity. This method is always convergent, but slowly converges. The interval in which the root is located is always divided into half. If the determined approximation is close to the real root, then the speed of the method decreases.

3. Secant method

Now we remind the second numerical method of finding the roots of non-linear equations which is called secant method. This method uses (during

STEP	APPROXIMATION
1	4.5
2	4.25
3	4.375
4	4.4375
5	4.46875
6	4.453125
7	4.4609375
8	4.46484375
9	4.462890625
10	4.4619140625
11	4.46240234375
12	4.46264648438
13	4.46252441406
14	4.46246337891
15	4.46243286133
16	4.46241760254
17	4.46242523193
18	4.46242141724
19	4.46242332458
20	4.46242237091
21	4.46242284775
22	4.46242260933
23	4.46242272854
24	4.46242266893
25	4.46242263913
26	4.46242262423
27	4.46242263168
28	4.46242263541
29	4.46242263354
30	4.46242263447
31	4.46242263494

TABLE 1. Approximations of the solution of the equation $x^3 - 17x - 13 = 0$ in the $[4, 5]$ interval – bisection method

generating successive approximations of the value of the sought root of an equation) the linear interpolation. The linear interpolation strategy is built on the basis of the known values of the two recently calculated ordinates of function f. Assume that we have a given function f, two starting points x_1

and x_2, and an interval $[a, b]$ in which we search the root, where the points x_1 and x_2 belong. In this interval, the function must fulfil the following conditions:

(1) f is differentiable on the interval $[a, b]$ and it is continuous there,

(2) f has different signs at the endpoints of the interval $[a, b]$ (this does not apply to the points x_1 and x_2),

Since the function is continuous, then by Darboux theorem it accepts in the interval $[a, b]$ all values between $f(a)$ and $f(b)$. These values have different signs (or they lie on different sides of OX axis), so there must be a point x_0 in the interval $[a, b]$, for which one of the two possibilities holds: $f(a) < f(x_0) < f(b)$ or $f(b) < f(x_0) < f(a)$.

(3) $f'(x) \neq 0$ for $x \in [a, b]$,

Therefore, there is no local minimum or maximum. This condition assures that the secant will not be parallel to the OX axis, which would make it impossible to determine its intersection with this OX axis.

If the function f satisfies the given conditions, there exists a root of a given equation in the interval $[a, b]$. In the secant method, two previously designated points are used to determine the next approximation of the function root, as follows: we create in a given interval $[a, b]$ the sequence of secant lines whose zeroes converge to the solution of the equation $f(x) = 0$. Depending on the signs of the function $f(x)$ and $f''(x)$ at the points a and b we use two different formulae. If $f(b) \cdot f''(b) > 0$, then we are looking for solution with the formula:

$$x_0 = a,\ x_{n+1} = x_n - \frac{f(x_n)}{f(b) - f(x_n)}\,(b - x_n)\,,\ n \geq 1. \tag{1}$$

However, if $f(a) \cdot f''(a) > 0$, then we find solution basing on the following formula:

$$x_0 = b,\ x_{n+1} = x_n - \frac{f(x_n)}{f(x_n) - f(a)}\,(x_n - a)\,,\ n \geq 1. \tag{2}$$

Now we will illustrate the method using some example.

Example 2. *Find the approximation of the solution of an equation* $x^3 + 7x^2 - 11x + 4 = 0$ *in the interval* $[-9, -8]$.

We will use the Python application that is written on the base of the secant method and formulae (1) and (2). Here is the listing of the program.

```
SECANT METHOD - PROGRAM IN PYTHON LANGUAGE
def fun(x):
    return x*x*x+7*x*x-11*x+4
def diff2(x):
    return 6*x+14
def main():
    a=float(input("Enter the left endpoint of interval: "))
    b=float(input("Enter the right endpoint of interval: "))
    precision=float(input("Enter precision of solution: "))
    if fun(b)*diff2(b)>0:
        x_0=a
        while abs(fun(x_0))>=precision:
            x_0=x_0-(fun(x_0)/(fun(b)-fun(x_0)))*(b-x_0)
            print(x_0)
    elif fun(a)*diff2(a)>0:
        x_0=b
        while abs(fun(x_0))>=precision:
            x_0=x_0-(fun(x_0)/(fun(x_0)-fun(a)))*(x_0-a)
            print(x_0)
    print("Approximation of the solution is:",x_0)
main()
```

Analysed equation has a real solution located in the interval $[-9,-8]$. The exact solution is equal to (*Mathematica environment*):

$$x_0 = \frac{1}{3}\left(-82\sqrt[3]{\frac{2}{1487-3\sqrt{633}}} - \sqrt[3]{\frac{1}{2}\left(1487-3\sqrt{633}\right)}\right) - \frac{7}{3}.$$

As it is seen the solution has also a very complicated form. Using the Python program we find an approximation of the chosen precision. The sequence of the approximations of the above solution is presented in Table 2. The precision is 10^{-12}.

The advantage of the secant method is that we do not need to know the analytical form of the derivative of the function. It is not as fast as the Newton method (tangent method that will be discussed in the next section), but usually faster than the bisection method as its convergence takes into

STEP	APPROXIMATION
1	-8.32183908045977
2	-8.364885832734824
3	-8.370338145447704
4	-8.371023837640712
5	-8.371109993976804
6	-8.371120818185480
7	-8.371122178060274
8	-8.371122348904741
9	-8.371122370368353
10	-8.371122373064878
11	-8.371122373403650
12	-8.371122373446212
13	-8.371122373451557
14	-8.371122373452230
15	-8.371122373452314
16	-8.371122373452325

TABLE 2. Approximations of the solution of the equation $x^3+7x^2-11x+4=0$ in the [-9,-8] interval – secant method

account the shape of the graph of the function (the interval is not divided on two identical parts).

4. Newton tangent method

Newton method which is called also the tangent method allow to find the approximations of the roots of the continuos function f. We can also use it in the case when we are not able to find exact solutions of the equation $f(x)=0$ or when exact solutions are complicated. This method is one of the best methods as its convergence is very quick but unfortunately the function f in the equation must satisfy the following strong conditions [2]:

Theorem 2. *If the function f is continuous in the interval $[a,b]$ and:*

- *signs of the function f are different on the endopints of the interval $[a,b]$ i. e. $f(a)\cdot f(b)<0$,*
- *there exists only one root of the function f in the analysed interval,*
- *first and second derivative of the function f do not change their signs in the interval $[a,b]$,*

then equation $f(x)=0$ has only one solution $x_z\in[a,b]$.

If the function f has more zeroes in the interval $[a, b]$ we may use this method dividing the interval $[a, b]$ onto smaller ones in which the conditions given in the above theorem are satisfied.

The Newton method means constructing the sequence of the tangents to the graph of the function f. Points of intersection the secants with OX axis are the following approximations of the root of the function f. The method is convergent to this solution x_z independent on the choice of a startpoint $x_0 \in [a, b]$. The idea of the Newton method is presented in Fig. 1. and Fig. 2.

In the first step we choose the starting point x_0 and we construct the first tangent to the graph of the function f in the point $P_0 = (x_0, f(x_0))$. We usually take as x_0 one of the points a or b in which signs of the function f and its second derivative f'' are the same ($f(a) \cdot f''(a) > 0$ or $f(b) \cdot f''(b) > 0$). The tangent to the graph of the function f in the point P_0 has the intersection with the OX axis in the point $(x_1, 0)$. Then we construct the next tangent to the graph of the function f in the point $P_1 = (x_1, f(x_1))$ and we find its intersection with the OX axis finding the point $(x_2, 0)$ and so on. This recurentive procedure let find the approximations of the solution of the equation $f(x) = 0$. The sequence $x_0, x_1, \ldots$ converges to the sought solution.

The procedure can be presented as the recurentive formula:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{3}$$

The choice of the starting point is very important and has an influence on the convergence of the procedure. The Newton method usually converges faster than secant method. [3].

Every numerical method generates errors. The error of n-th approximation in this case can be found using the following formula :

$$|x_z - x_n| \leq \frac{f(x_n)}{\min_{x\in[a,b]} |f'(x)|}.$$

or

$$|x_z - x_n| \leq \frac{\max_{x\in[a,b]} |f''(x)|}{2 \min_{x\in[a,b]} |f'(x)|}(x_z - x_{n-1})^2.$$

where x_z is the exact value of the zero of the function f.

Practically, to end the algorithm we choose one of the following conditions:

- $|f(x_n)| < \epsilon$ – the value of the function f in the point x_n is less than ϵ

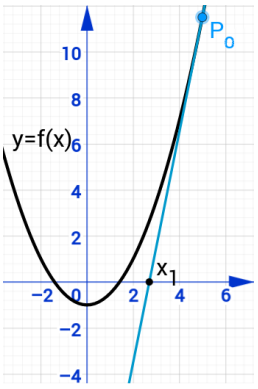


FIGURE 1. First approximation – point x_1 [own work]

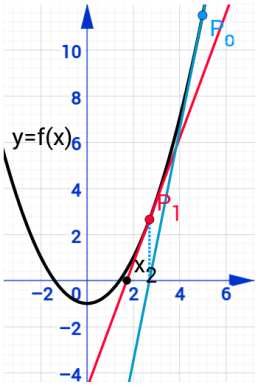


FIGURE 2. Second approximation – point x_2 [own work]

- $|x_{n+1} - x_n| < \epsilon$ – the difference between approximations is less than ϵ
- $\frac{M}{2m}(x_{n+1} - x_n)^2 \leq \epsilon$ – the error is less than ϵ.

Example 3. *Find the approximation of the root of the equation* $\frac{1}{10}(x^3 + 4x^2 - 9x + 1) = 0$ *in the interval* $[1, 2]$.

The function f satisfies the conditions of the theorem 2: it is continuous, the values on the endpoints have different signs: $f(1) = -\frac{3}{10}, f(2) = \frac{7}{10}$. First derivative is equal to $f'(x) = \frac{1}{10}(3x^2 + 8x - 9)$ and its zeroes are equal to $x_1 = -\frac{1}{3}(4 + \sqrt{43})$ and $x_2 = \frac{1}{3}(\sqrt{43} - 4) < 1$, so $f'(x) > 0$ for $x \in (-\infty, x_1) \cup (x_2, \infty)$ – first derivative of the function has the same sign in the interval $[1, 2]$. Second derivative is equal to $f''(x) = \frac{1}{5}(3x + 4)$, so $f''(x) > 0$ for $x \in (-\frac{4}{3}, \infty)$ – second derivative of the function has the same sign in the interval $[1, 2]$ (see also Fig. 3.).

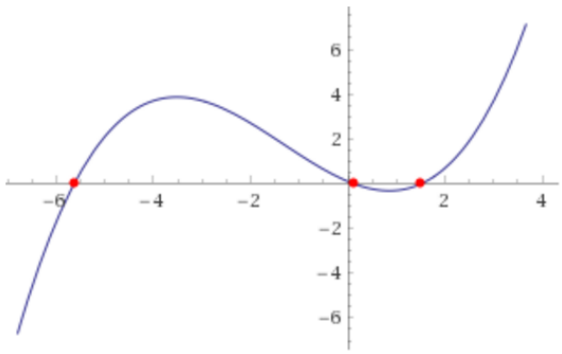


FIGURE 3. Graph of the function $f(x)=\frac{1}{10}(x^3+4x^2-9x+1)$ [www.wolframalpha.com]

We can find the exact solutions of the analysed equation. The solution in the interval $[1,2]$ (*Mathematica environment*) has the form:

$$x=-\frac{4}{3}+\frac{1}{3}\left(\frac{43}{\sqrt[3]{\frac{1}{2}\left(-479+3i\sqrt{9843}\right)}}+\sqrt[3]{\frac{1}{2}\left(-479+3i\sqrt{9843}\right)}\right)$$

The exact solution is not practical, so we use the Newton method to find the approximation of it.

The Python program can be written using formula (3). Its code is listed below.

```
TANGENT (NEWTON) METHOD - PROGRAM IN PYTHON LANGUAGE
import math
def f(x):
    return (x*x*x+4*x*x-9*x+1) / 10
def dev_f(x):
    return (3*x*x+8*x-9) / 10
def main():
    a = 1
    b = 2
    x_0 = b
    precision = float(input("Determine precision:"))
    while abs(f(x_0))>=precision:
        x_0 = x_0 - f(x_0) /dev_f(x_0)
```

```
        print(x_0)
    print("Approximation of the solution is:",x_0)
main()
```

The approximation of the solution with precision 10^{-15} we find in fifth iteration. The approximations are presented in Fig. 4.

```
[artur@ag-msi Metoda stycznych]$ python3 newtonAG.py
1.631578947368421
1.5229854433598955
1.5127697483555214
1.5126801384140627
1.5126801315345624
Solution: 1.5126801315345624
f(x):     0.00000000000000000000
epsilon:  0.00000000100000000000
```

FIGURE 4. The following approximations of the equation [own work]

5. Iteration method

Iteration method is a very practical application of Banach fixed-point theorem to finding the approximations of the solutions of nonlinear equations. The basic advantage of this method is its fast convergence. So we do not need many iterations to find the satisfactionable approximation. But unfortunatelly this method can be used only if the following conditions are satisfied:

(1) The nonlinear equation $f(x) = 0$ must be presented in the form $g(x) = x$, where g is differentiable in the analysed interval $[a, b]$;
(2) The function g is a contraction mapping in the interval $[a, b]$.

Let us remind the definition of the contraction mapping.

Definition 1. *The function f is said to be a contraction mapping in interval $[a, b]$ if there exists a constant $M < 1$, such that for all $x, y \in [a, b]$ the following condition is satisfied:*

$$|f(x) - f(y)| \leq M |x - y| .$$

Remark. It is clear (and can be eaisily proved) that if the function f is differentiable in the interval $[a, b]$ and $|f'(x)| < 1$ for all $x \in [a, b]$ then the function f is a contraction mapping. We also use this practical condition to check if the function is a contraction map.

Now we remind the theorem which is the base of the iteration method.

Theorem 3. *If the function f is a contraction mapping in the interval $[a, b]$ then there exists only one point x_0 such that $f(x_0) = x_0$. In addition the sequence given by the following recurentive formula: $a, f(a), f(f(a)), \ldots$ is convergent to x_0 independent on the choice of the starting point a.*

As it may be noticed, in this case there exists only one solution of the equation $f(x) = x$.

Basing on the above theorem we can present the following algorithm of the iteration method:

(1) We rewrite the equation $f(x) = 0$ in the form $g(x) = x$;
(2) We check if in the interval $[a, b]$ the function f is such that $f(a) \cdot f(b) < 0$ and function g is a contraction mapping;
(3) We create the sequence: $t, g(t), g(g(t)), g(g(g(t))), \ldots$ that is convergent to the sought solution, the starting point $t \in [a, b]$ can be chosen arbitraly.

Now we will illustarate the method with two examples.

Example 4. *Find the solution of the equation $\cos x = 2x$ in the interval $[0, \frac{\pi}{2}]$ using iteration method .*

Let us notice that this equation is equivalent to the equation $\cos x - 2x = 0$. Let f be the function given by the formula $f(x) = \cos x - 2x$. It is clear that this function has at least one zero in the interval $[0, \frac{\pi}{2}]$ as $f(0) = 1 > 0$ and $f(\frac{\pi}{2}) = -\pi < 0$. In addition it is not a polynomial equation, we cannot find the exact solution. Let us rewrite the equation in the form $\frac{\cos x}{2} = x$. The function g given by the formula $g(x) = \frac{\cos x}{2}$ is a contraction mapping in every interval as we have an inequality $|g'(x)| = \frac{1}{2}|\sin x| < 1$. Now we can find the approximation of the only one solution using the Python program that is listed below.

```
ITERATION METHOD - PROGRAM IN PYTHON LANGUAGE
import math
x=0
print("Determine precision:")
precision=float(input())
while abs(math.cos(x)-2*x)>=precision:
    x=0.5*math.cos(x)
    print(x)
print("Approximation of the solution is:",x)
```

In fifth step we obtain the approximation $x \approx 0,45$ with precision 10^{-3} and in ninth step approximation $x \approx 0,450184$ with precision 10^{-6}.

Example 5. *Find the solution of the equation* $x^3+4x-1=0$ *in the interval* $[0,1]$.

Let W be a polynomnial given by the formula $W(x)=x^3+4x-1$. We easily notice that $W(0)=-1$ and $W(1)=4$. So at least one solution of the equation exists in the interval $[0,1]$. We can use Cardano formulae and by the help of *Mathematica environment* find the exact solution:

$$x_0=\frac{\sqrt[3]{\frac{1}{2}\left(9+\sqrt{849}\right)}}{\sqrt[3]{9}}-4\sqrt[3]{\frac{2}{3\left(9+\sqrt{849}\right)}}.$$

The exact solution has very complicated form so we use iteration method in this case. We can rewrite the equation in the form $\frac{1-x^3}{4}=x$. The function$g$ given by the formula $g(x)=\frac{1-x^3}{4}$ is a contraction mapping in the interval $[0,1]$ as we have the inequality: $|g'(x)|=\frac{3}{4}x^2<1$.

Now we can find the approximations of the only one solution using Python program.

```
ITERATION METHOD - PROGRAM IN PYTHON LANGUAGE
import math
def fun(x):
    return 0.25*(1-x*x*x)
def main():
    x=0
    print("Determine precision:")
    precision=float(input())
    while abs(x*x*x+4*x-1)>=precision:
        x=fun(x)
        print(x)
    print("Approximation of the solution is:",x)
main()
```

It is interesting that we obtain the approximation $x\approx 0,246$ with precision 10^{-3} in the second step, and the approximation $x\approx 0,246266$ with precision 10^{-6} in the fifth step, so the iteration method in this case is fast convergent.

6. Final remarks

In this article we have presented the possiblilities of using the computer programs written in Python language in solving of nonlinear equations. In many real situations it is the best way to find at least approximations of the solutions of such equations as in many cases there does not exist the

way to obtain the exact form of the solutions or exact solutions have very compilcated impractical forms consisting for example complex numbers. Then we may use numerical methods to find approximations. We remind four important methods: bisection method, secant method, tangent method and iteration method and we show how to use them to write computer programs that find the approximations of the solutions machanically. On one hand Python language programs have a very simple structure so writing programs is very quick and simple, on the other hand programs let to avoid long and complicated computations. Of course, every method have its own advantages and disadvantages. Bisection method is very simple and we can use it to every equation, but it is slowly convergent, secant and tangent methods are usually faster but the functions present in equations must satisfy adiitional, often strict, conditions. The iteration method seems to be the fastest in many cases but we can use it very seldom as it is not always possible to present the equation in a form $g(x) = x$ and even if we do that, the function g must be a contraction mapping in the analysed interval.

References

[1] Povstenko J. *Wprowadzenie do metod numerycznych*, Wydawnictwo EXIT, Warszawa 2005.
[2] Fortuna Z., Macukow B., Wąsowski J., *Metody numeryczne*, Wydawnictwo Naukowo-Techniczne, Warszawa 2002.
[3] Ryaben'kii V.S.; Tsynkov S.V., *A Theoretical Introduction to Numerical Analysis*, CRC Press, Boca Raton 2007.
[4] Dawson M. *Python dla każdego. Podstawy programowania*, Helion, Warszawa 2014.

Received: October 2017

Marcin Ziółkowski
Warsaw University of Life Sciences,
Department of Informatics,
Ul. Nowoursynowska 159, 02-787 Warszawa, Poland
E-mail address: marcin_ ziolkowski@sggw.pl

Lidia Stępień
Jan Długosz University in Częstochowa,
Institute of Mathematics and Computer Science,
Al. Armii Krajowej 13/15, 42-200 Częstochowa, Poland
E-mail address: l.stepien@ajd.czest.pl

Marcin Ryszard Stępień
Kielce University of Technology,
Department of Mathematics and Physics,
Al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland
E-mail address: mstepien@tu.kielce.pl

Artur Gola
Jan Długosz University in Częstochowa,
Institute of Mathematics and Computer Science,
Al. Armii Krajowej 13/15, 42-200 Częstochowa, Poland
E-mail address: a.gola@ajd.czest.pl

Part II
Mathematics and its Application

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
49–60
DOI http://dx.doi.org/10.16926/m.2017.22.04

ON MONOTONICITY OF REAL FUNCTIONS

JACEK MAREK JĘDRZEJEWSKI

Abstract

Monotonicity of functions were of great interest of many mathematicians. Starting from the well known theorem of monotonicity of a differentiable function one can get quite sophisticated results. We give a survey of results when thesis of them is *continuous and monotone function*. Someone can ask why it should be continuous. Even a differentiable functions but not at the only point of its domain with positive derivative need not be non-decreasing. That is why we want to look for theorems for continuous functions.

Well known theorems on monotonicity

Starting his mathematical way any student knows that:

Theorem 1[1]**.** *If $f : (a, b) \longrightarrow \mathbb{R}$ fulfils the following conditions:*

- *f' exists at every point of (a, b),*
- *$f'(x) \geq 0$ if $x \in (a, b)$,*

then f is non-decreasing (and continuous).

It is not difficult to find out that:

If a continuous function $f : (a, b) \longrightarrow \mathbb{R}$ is non-decreasing, then it is not necessary for f to be differentiable everywhere.

We shall try to give some week conditions for a function to be continuous and non-decreasing.

Let us remind that a monotone function (say non-decreasing) must have a non-negative derivative in a big set of points.

Theorem 1. *If $f : (a, b) \longrightarrow \mathbb{R}$ is non-decreasing function, then f' exists almost everywhere in (a, b) and $f'(x) \geq 0$.*

• *Jacek M. Jędrzejewski* — jacek.m.jedrzejewski@gmail.com
Retired from Pomeranian Academy in Słupsk.

[1] I know nobody, who knows the author of this theorem.

It means that condition $\mathscr{M}$ is fulfilled almost everywhere in the set E if the set of points at which this condition is not fulfilled has Lebesgue measure 0.

The converse statement is not true! There are monotone functions which are differentiable only in the set which complement has measure 0. For example:

Example 1.

Let K be the classical Cantor ternary set. Denote by (a_n, b_n) all components of the complement of the set K and define the function f as follows:

$$f(x) = \begin{cases} 0 & \text{if} \quad x \in \{a_n : n \in \mathbb{N}\}, \\ 1 & \text{if} \quad x \in \{b_n : n \in \mathbb{N}\}, \\ \text{linear and continuous} & \text{in} \quad \text{each interval } [a_n, b_n]. \end{cases}$$

Let us discuss some generalizations of those theorem. There are different ways of generalizations:

(1) $f'(x) \geq 0$ not everywhere,
(2) instead of usual derivative one can consider generalized derivative,
(3) both of them.

1. Generalizations of the first kind

First series of theorems deals with ordinary derivatives. Their generalizations concern to the smaller set of points where the derivative exists.

Theorem 2. (G. Goldowsky – [2], 1928, L. Tonelli – [12], 1930)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ is continuous, f' exists nearly everywhere in (a, b) and $f'(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing.

The term *nearly everywhere* means that such property holds for the set which complement is at most countable.

Next theorem needs the idea of absolute continuity of a function. Let us remind this notion. We say that a function $f : (a, b) \longrightarrow \mathbb{R}$ is absolutely continuous if for each positive ε there is a positive δ such that

$$\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \varepsilon$$

for each sequence of intervals (a_k, b_k) for which

$$\sum_{k=1}^{\infty} |b_k - a_k| < \delta.$$

Theorem 3. (S. Saks – [7], 1937)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ is absolutely continuous and $f'(x) \geq 0$ almost everywhere, then f is non-decreasing.

When looking at the previous theorems we can observe that continuity of a function is contained in assumptions. Z. Zahorski proved interesting theorem for functions which were from the I class of Baire with Darboux property. But still the thesis says that such a function must be continuous.

Theorem 4. (Z. Zahorski – [13], 1950) *If a function $f : (a, b) \longrightarrow \mathbb{R}$ belongs to the I class of Baire, is a Darboux function, f' exists nearly everywhere in (a, b) and $f'(x) \geq 0$ almost everywhere, then f is non-decreasing (and continuous).*

It is worth to add that none of the assumptions can be omitted. Here we give some counterexamples for it.

Example 2. Darboux property.

The function $f : [-1, 1] \longrightarrow \mathbb{R}$ defined in a way

$$f(x) = \begin{cases} 1 + x & \text{if} \quad x \in [-1, 0), \\ x & \text{if} \quad x \in [0, 1]. \end{cases}$$

has all Zahorski's assumptions but Darboux property.

Example 3. Existence of a derivative.

The function $f : [0, 1] \longrightarrow \mathbb{R}$ defined in a way

$$f(x) = \begin{cases} 0 & \text{if} \quad x \in \{a_n : n \in \mathbb{N}\}, \\ 1 & \text{if} \quad x \in \{b_n : n \in \mathbb{N}\}, \\ \frac{1}{b_n - a_n} \cdot x - \frac{a_n}{b_n - a_n} & \text{if} \quad x \in (a_n, b_n), \\ 0 & \text{if} \quad x \in K \setminus (\{a_n : n \in \mathbb{N}\} \cup \{b_n : n \in \mathbb{N}\}), \end{cases}$$

where (a_n, b_n) are all complement of the Cantor set K, has Darboux property, belongs to the first class of Baire and $f'(x) > 0$ for each x from the set $\bigcup_{n=1}^{\infty}(a_n, b_n)$. This function is differentiable at no point of the (uncountable) Cantor set.

2. Generalization of the second kind

2.1. **Approximate derivative.** First attemptions to generalize theorems on monotonicity were given by Tolstoff in 1939. He applied approximate derivative instead of usual derivative, but he assumed a bit more than Zahorski did, it means that Tolstoff assumed that a function should be approximately continuous.

Theorem 5. (G. Tolstoff – [10], 1939) *If a function $f : (a, b) \longrightarrow \mathbb{R}$ is approximately continuous, f'_{ap} exists nearly everywhere in (a, b) and $f'_{ap}(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing (and continuous).*

Z. Zahorski asked then: Is the next statement true?

Zahorski's hypothesis
If a function $f : (a, b) \longrightarrow \mathbb{R}$ belongs to the I class of Baire, is a Darboux function,f'_{ap} exists nearly everywhere in (a, b) and $f'_{ap}(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing.

Positive answer has been given independently by A. M. Bruckner and T. Świątkowski in 1966. Then the proper theorem should be named Bruckner-Świątkowski Theorem.

Theorem 6. (T. Świątkowski – [8], 1966, A. M. Bruckner – [1], 1966)
If a function $f : (a, b) \longrightarrow \mathbb{R}$

(1) *belongs to the I class of Baire,*
(2) *fulfils Darboux condition,*
(3) *f'_{ap} exists nearly everywhere in (a, b),*
(4) *$f'_{ap}(x) \geq 0$ almost everywhere,*

then f is non-decreasing (and continuous).

The proof of this theorem given by Świątkowski is as simple as can be. He used only fundamental properties of Darboux functions and approximate derivative and used no special tools of real functions theory. In spite of this proof, Bruckner involved several properties of real functions for example: Banach condition, VB and VBG functions, but the statement is the same. I am not able to say which of those proofs is better or more general. Both of them are brilliant and both of them are pretty long and complicated.

Approximate continuity can be regarded as continuity with respect to density topology. Thus approximate derivative is also the limit of appropriate quotient with respect to topology stronger than the natural one.

Thus we can come to next series of theorems.

2.2. **Qualitative derivative.**

Definition 1. *If $f : (a, b) \longrightarrow \mathbb{R}$ is any function and $x_0 \in (a, b)$, then g is called the qualitative limit of f at the point x_0 if there exists a residual subset E of (a, b) such that*

$$\lim_{x \to x_0} f_{\restriction E}(x) = g.$$

If we apply limit of this kind to differential quotient of a function, we get the idea of qualitative derivative of this function. Using this kind of

generalization of the usual derivative one can get the next theorem in our series.

Theorem 7. (J. L. Leonard – [4], 1972)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ fulfils Darboux condition, belongs to the I class of Baire, f'_q exists nearly everywhere in (a, b) and $f'_q(x) \geq 0$ almost everywhere, then f is non-decreasing (and continuous).

One can observe that this kind of a limit can be obtained as limit with respect to some topology. This topology is not a generalization of density topology, but it also is stronger than the natural one. This topology can be defined in the following way:

Definition 2. *A subset U of $\mathbb{R}$ is called qualitatively open if it can be represented in the form*

$$U = G \triangle E,$$

where G is open in natural topology and E is of the first category.

There are several different topologies defined in a similar way. For example: let $\mathscr{J}$ be a σ-ideal of subsets of $\mathbb{R}$. If a set U is regarded as open set in the topology generated by $\mathscr{J}$ if it is a symmetrical difference of open set in natural topology and a set from the ideal $\mathscr{J}$, then we obtain a topology in $\mathbb{R}$.

Applying this topology to the idea of a limit of a function and to the differential quotient one can get quite a big class of theorems like the last one.

Generalizations of this idea will be find in further part of article.

3. Generalizations of the third kind

3.1. **Preponderant derivative.**

Definition 3. *A number g is called a preponderant limit of a function $f : (a, b) \longrightarrow \mathbb{R}$ at a point x_0 if there exists a measurable set E such that*

$$\liminf_{h \to 0^-} \frac{\mu\left(E \cap (x_0 - h, x_0)\right)}{h} > \frac{1}{2}$$

$$\liminf_{h \to 0^+} \frac{\mu\left(E \cap (x_0, x_0 + h)\right)}{h} > \frac{1}{2}$$

and

$$\lim_{x \to x_0} f_{\restriction E}(x) = g.$$

Theorem 8. (J. Jędrzejewski – master's thesis, 1969)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ is from the I class of Baire, fulfils Darboux condition, f'_{pr} exists nearly everywhere in (a, b) and $f'_{pr}(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing (and continuous).

Theorem 9. (J. L. Leonard – [4], 1972)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ is preponderantly continuous, f'_{pr} exists nearly everywhere in (a, b) and $f'_{pr}(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing (and continuous).

The sets applied for preponderant limit do not generate any topology, then this kind of generalization is quite different from the previously considered manners.

3.2. **Selective derivative.**

Definition 4. (R. J. O'Malley – [5], 1977)
By a selection we mean a real function p of two variables which associates to each pair of points x and y a point $p(x, y)$ fulfilling the following conditions:

(1) *$p(x, y) = p(y, x)$ for each x and y from $\mathbb{R}$,*
(2) *if $x < y$, then $x < p(x, y) < y$.*

Definition 5. *We say that a number g is a selective limit with respect to the selection p of a function f at a point x_0 if*

$$g = \lim_{y \to x_0} p(x_0, y).$$

Selective limit operation applied to differential quotient gives s-derivative.

Definition 6. *A number $f'_s(x_0)$ is called selective derivative of a function f at the point x_0 if*

$$f'_s(x_0) = \lim_{x \to x_0} \frac{f(p(x, x_0)) - f(x_0)}{p(x, x_0) - x_0}$$

for a given selection p.

Theorem 10. (R. J. O'Malley – [5], 1977)
If a function $f : (a, b) \longrightarrow \mathbb{R}$ belongs to the I class of Baire, fulfils Darboux condition, f'_s exists nearly everywhere in (a, b) and $f'_s(x) \geq 0$ almost everywhere in (a, b), then f is non-decreasing (and continuous).

3.3. **Świątkowski's τ-derivative.**

Definition 7. (T. Świątkowski – [9], 1972)
For an $x \in \mathbb{R}$ let τ_x be a class of sets fulfilling the following conditions:

(1) *if $A \in \tau_x$ and $B \in \tau_x$ then $A \cap B \in \tau_x$,*
(2) *if δ is a positive number and $E \in \tau_x$, then $E \cap (x - \delta, x + \delta) \in \tau_x$,*
(3) *$\bigcap \tau_x = \{x\}$,*
(4) *if $\delta > 0$ and $E \in \tau_x$, then $\big(E \cap (x - \delta, x + \delta)\big) \setminus \{x\} \neq \varnothing$.*

Definition 8. *We say that a class* $\{\tau_x : x \in \mathbb{R}\}$ *of sets fulfilling the above conditions fulfils Khintchine's condition if* x_0 *is* τ*-accumulation point of the set*

$$\bigcup_{n=1}^{\infty} (x_n - \delta_n, x_n + \delta_n)$$

for every sequences $(x_n)_{n=1}^{\infty}$ *and* $(\delta_n)_{n=1}^{\infty}$ *such that*

- $\lim\limits_{n\to\infty} x_n = x_0$,
- $\lim\limits_{n\to\infty} \delta_n = 0$, *and* $\delta_n > 0$, *for each* $n \in \mathbb{N}$,
- $\lim\limits_{n\to\infty} \dfrac{\delta_n}{|x_n - x_0|} > 0$

Any class τ_x of subsets of the set of real numbers has some properties of a system of neighbourhoods of the point x. Applying the topological terminology, we say that a point x is called to be a τ-accumulation point of a set E if

$$E \cup \{x\} \in \tau_x.$$

Applying this τ-limit operation to the differential quotient of a function we obtain τ-derivative.

Theorem 11. (T. Świątkowski – [9], 1972)
A class of sets $\{\tau_x : x \in \mathbb{R}\}$ *fulfils the Khintchine condition if and only if for each monotone function* f *from the existence of the* τ*-derivative of* f *implies the existence of* f'.

Before we come to theorem on monotonicity, w have to define also condition (W).

Definition 9. *We say that a function* f *and a class* $\tau = \{\tau_x : x \in (a, b)\}$ *satisfy condition* (W) *if*

(1) f *fulfils Darboux condition,*
(2) f *is nearly everywhere continuous in* (a, b),
(3) τ *fulfils Khintchine's condition nearly everywhere in* (a, b),
(4) f' *exists nearly everywhere in* (a, b).

Theorem 12. (M. Mastalerz-Wawrzyńczak – [6], 1977)
Let a class of sets $\tau = \{\tau_x : x \in \mathbb{R}\}$ *fulfils conditions of Definition* 7 *and the Khintchine condition. Assume moreover that a function* $f : (a, b) \longrightarrow \mathbb{R}$ *fulfils condition* (W) *with the class* τ.

Under those assumptions, if $f'_\tau(x) \geq 0$ *almost everywhere in* (a, b), *then* f *is non-decreasing (and continuous).*

3.4. **Local systems.**

Definition 10. (B. S. Thomson – [11], 1985)
By a local system we mean a class $\mathcal{S}$ consisting of non-empty collections $\mathcal{S}(x)$ for each real number x, fulfilling the following conditions:

(1) $\{x\} \notin \mathcal{S}(x)$,
(2) $E \in \mathcal{S}(x) \to x \in E$,
(3) $E \in \mathcal{S}(x) \wedge F \supset E \to F \in \mathcal{S}(x)$,
(4) $E \in \mathcal{S}(x) \wedge \delta > 0 \Longrightarrow E \cap (x-\delta, x+\delta) \in \mathcal{S}(x)$.

A local system is called filtering at a point x if

$$E \cap F \in \mathcal{S}(x) \ \text{ whenever} E \in \mathcal{S}(x) \ \text{ and } \ F \in \mathcal{S}(x).$$

A local system is called filtering if it is filtering at each x in $\mathbb{R}$.

A local system is called bilateral if

$$E \cap (x-\delta, x) \neq \varnothing \quad \textit{and} \quad E \cap (x, x+\delta) \neq \varnothing$$

for each $x \in \mathbb{R}$, $E \in \mathcal{S}(x)$ and $\delta > 0$.

Definition 11. (B. S. Thomson)
A number g is called $\mathcal{S}$-limit of a function f at a point x if

$$f^{-1}(g-\varepsilon, g+\varepsilon) \cup \{x\} \in \mathcal{S}(x)$$

for each positive ε. We shall write then

$$g = (\mathcal{S}) \lim_{t \to x} f(t).$$

If we apply the $\mathcal{S}$-limit operation to the differential quotient of a function f at a point x, then we get $\mathcal{S}$-derivative of the function f at x.

One of monotonicity criterion (involving generalized derivatives) is given by B. S. Thomson:

Theorem 13. (B. S. Thomson)
Let $\mathcal{S}$ be a bilateral and filtering system fulfilling

- *intersection condition i.e. for each class of sets*

  $$\{E_x \in \mathcal{S}(x) \,:\, x \in \mathbb{R}\}$$

  there is a positive function $\delta : \mathbb{R} \longrightarrow \mathbb{R}$ such that

  $$E_x \cap E_y \cap [x, y] \neq \varnothing$$

  whenever $0 < y - x < \min\{\delta(x), \delta(y)\}$,
- *variation condition i.e. $\psi(I) \leq V_I(\psi, \mathcal{S})$ whenever ψ is a non-negative, sub-additive interval function.*

Then if a function $f : (a, b) \longrightarrow \mathbb{R}$ belongs to the I class of Baire, fulfils Darboux condition, $f'_{\mathcal{S}}$ exists nearly everywhere in (a, b) and $f'_{\mathcal{S}}(x) \geq 0$ almost everywhere, then f is non-decreasing (and continuous).

3.5. **$\mathfrak{B}$-systems.** Let us start from defining $\mathfrak{B}$ classes.

Definition 12. *For each $x \in \mathbb{R}$ let $\mathfrak{B}_x^+$ be a class of non-empty sets fulfilling the following properties:*

(1) $B_1 \cup B_2 \in \mathfrak{B}_x^+ \iff (B_1 \in \mathfrak{B}_x^+ \vee B_2 \in \mathfrak{B}_x^+)$,
(2) $B \cap (x, x + t) \in \mathfrak{B}_x^+$ *for each* $B \in \mathfrak{B}_x^+$ *and* $t > 0$.

For each $x \in \mathbb{R}$ let $\mathfrak{B}_x^-$ be a class of non-empty sets fulfilling the following properties:

(1) $B_1 \cup B_2 \in \mathfrak{B}_x^- \iff (B_1 \in \mathfrak{B}_x^- \vee B_2 \in \mathfrak{B}_x^-)$,
(2) $B \cap (x, x + t) \in \mathfrak{B}_x^-$ *for each* $B \in \mathfrak{B}_x^-$ *and* $t > 0$.

Let $\mathfrak{B}_x = \mathfrak{B}_x^- \cup \mathfrak{B}_x^+$.

Definition 13. *If f defined in some (a, b) is a real function, then a number (or $+\infty$ or $-\infty$) is called $\mathfrak{B}$-limit number of f at x_0 from (a, b) if*

$$\left\{x \in (a, b) \,:\, f^{-1}(U_g)\right\} \in \mathfrak{B}_{x_0}$$

for any neighbourhood U_g of the point g.

Definition 14. *If*

$$\left\{x \in (a, b) \,:\, f^{-1}(U_g) \in \mathfrak{B}_{x_0}^-\right\}$$

for any neighbourhood U_g of the point g, then g is called the left $\mathfrak{B}$-limit number of a function f at a point x_0.

Similarly we define right $\mathfrak{B}$-limit numbers of a function f at a point x_0.

- By $L_{\mathfrak{B}}^+(f, x_0)$ we denote the set of right $\mathfrak{B}$-limit numbers of f at x_0.
- By $L_{\mathfrak{B}}^-(f, x_0)$ we denote the set of left $\mathfrak{B}$-limit numbers of f at x_0.
- By $L_{\mathfrak{B}}(f, x_0)$ we denote the set of all $\mathfrak{B}$-limit numbers of f at x_0.

Then, as for usual limit numbers, one can state:

Theorem 14. *For arbitrary real function f on the interval (a, b) and any x_0 from (a, b) the sets $L_{\mathfrak{B}}(f, x_0)$, $L_{\mathfrak{B}}^-(f, x_0)$ and $L_{\mathfrak{B}}^+(f, x_0)$ are non-empty, closed and*

$$L_{\mathfrak{B}}(f, x_0) = L_{\mathfrak{B}}^-(f, x_0) \cup L_{\mathfrak{B}}^+(f, x_0).$$

Up to now we have defined $\mathfrak{B}$-limit numbers of a function, we shall apply rather $\mathfrak{B}$-limits instead limit numbers. Let us define them.

Definition 15. *A number g is called $\mathfrak{B}$-limit of a function at a point x_0 from (a, b) if*

$$\{g\} = L_{\mathfrak{B}}(f, x_0).$$

There is another possibility to characterized $\mathfrak{B}$-limits of a function. But before we do this we shall have to define the second class of sets denoted by $\mathfrak{B}_x^*$ for all $x \in \mathbb{R}$.

Definition 16. *A subset E of $\mathbb{R}$ belongs to the class $\mathfrak{B}_x^*$ if $\mathbb{R} \setminus E \notin \mathfrak{B}_x$.*

Now we can give the following characterization of $\mathfrak{B}$-limit of a function.

Theorem 15. *A number g is $\mathfrak{B}$-limit of a function at a point $x_0 \in (a,b)$ if and only if*

$$\left\{x \in (a,b) \,:\, f^{-1}(U_g) \in \mathfrak{B}_{x_0}^*\right\}$$

for any neighbourhood U_g of the point g.

We know from previous theorems that any function has $\mathfrak{B}$-limit number at any point of domain of f. This time we are not able to state that any function has $\mathfrak{B}$-limit, as it is evident for usual limits, but if a $\mathfrak{B}$-limit of a function exists it must be only one.

Let us remark yet that the class $\mathfrak{B}^*$ that is applied to our considerations is very similar to the class τ considered by T. Świątkowski.

Next properties will be of some use in the further theory.

Definition 17. *We say that the class $\mathfrak{B}$ fulfils condition $\mathscr{M}$ if*

$$\bigcup_{n=1}^{\infty} E_n \in \mathfrak{B}_{x_0}$$

for any: $x_0 \in (a,b)$, sequence $(x_n)_{n=1}^{\infty}$ converging to x_0 and every sequence of sets $(E_n)_{n=1}^{\infty}$ such that $E_n \in \mathfrak{B}_{x_n}$.

Definition 18. *We say that the class $\mathfrak{B}$ fulfils condition $\mathscr{M}'$ if*

$$\bigcup_{n=1}^{\infty} E_n \in \mathfrak{B}_{x_0}$$

for any: $x_0 \in (a,b)$, sequence $(x_n)_{n=1}^{\infty}$ converging to x_0 and every sequence of intervals $(E_n)_{n=1}^{\infty}$ such that $E_n \in \mathfrak{B}_{x_n}$.

Now we want to compare $\mathfrak{B}$-derivatives with usual derivatives for monotone and continuous functions. Of course, if a function is differentiable it must be $\mathfrak{B}$-differentiable.

Assume now that a system $\mathfrak{B}$ fulfils condition $\mathscr{M}'$ and an increasing and continuous function f is $\mathfrak{B}$-differentiable and at some point $x \in (a,b)$ is not differentiable, for example it is not right differentiable. Then there are two numbers α, β and sequences $(u_n)_{n=1}^{\infty}$, $(w_n)_{n=1}^{\infty}$ converging to x and such that

$$x < u_{n+1} < w_{n+1} < u_n < w_n,$$

$$\frac{f(u_n)-f(x)}{u_n-x}<\alpha<\beta<\frac{f(w_n)-f(x)}{w_n-x}$$

for each positive integer n.

Then there are non-empty intervals $(u_n, u_n+\gamma_n)$, $(w_n-\delta_n, w_n)$ such that

$$\frac{f(t)-f(x)}{t-x}<\alpha<\beta<\frac{f(s)-f(x)}{s-x}$$

for all $t\in(w_n-\delta_n, w_n)$ and $s\in(u_n, u_n+\gamma_n)$ and $n\in\mathbb{N}$. In view of condition $\mathscr{M}'$ of the system $\mathfrak{B}$ the conditions

$$\bigcup_{n=1}^{\infty}(w_n-\delta_n, w_n)\in\mathfrak{B}_x \quad\text{and}\quad \bigcup_{n=1}^{\infty}(u_n, u_n-\gamma_n)\in\mathfrak{B}_x.$$

hold. Then there are $\mathfrak{B}$-limit numbers of the function $\dfrac{f(y)-f(x)}{y-x}$ at x, one of them less than α and the second greater than β, what means that the function f is not $\mathfrak{B}$-differentiable at x.

In that way we have proved:

Theorem 16. *If a system $\mathfrak{B}$ fulfils condition $\mathscr{M}'$ then $\mathfrak{B}$-differentiability of a monotone function is equivalent to differentiability of that function.*

4. Topological approach

Let $\mathcal{T}_0$ be the natural topology in the set $\mathbb{R}$. Assume that $\mathcal{T}$ is a topology in $\mathbb{R}$ stronger than $\mathcal{T}_0$ and such that each x in $\mathbb{R}$ is a bilateral $\mathcal{T}$-accumulation point of $\mathbb{R}$, then the class $\mathfrak{B}=\{\mathfrak{B}_x : x\in\mathbb{R}\}$ is a class fulfilling conditions of J. Jędrzejewski, (see Definition 12). and conditions of T. Świątkowski (Definition 7). Moreover it forms also local system of B. Thomson, not every local system but only filtering one. Since those ideas are derived from a topology, then it is logical to use topological terminology.

Let us observe that there are several topologies in $\mathbb{R}$ generating system $\mathfrak{B}$ which fulfils condition $\mathscr{M}'$ and consequently $\mathfrak{B}$-derivative of an increasing function coincides with usual derivative.

Theorem 17. *If $\mathcal{T}$ is a topology in $\mathbb{R}$ fulfilling the above conditions and for any monotone function from the existence of $\mathcal{T}$-derivative implies the existence of the natural derivative, then each interval is a connected set (with respect to the topology $\mathcal{T}$).*

For the proof of this theorem see Theorem 16 or Theorem 12 in [6].

In the end, let us see that if a non-empty interval I or $\mathbb{R}$ is a disconnected set in topology $\mathcal{T}$, then even continuity of a function does not force a function f fulfilling conditions (2) – (4) from Theorem 6.

Here a problem has arisen:

PROBLEM.
Suppose that $\mathcal{T}$ is a topology in $\mathbb{R}$ for which each interval is a connected set (with respect to $\mathcal{T}$). Let $f'_{\mathcal{T}}$ denotes the derivative of f with respect to $\mathcal{T}$. Is it true that each function f defined in an interval (a, b) (or in $\mathbb{R}$) fulfilling the following conditions:

- f is Baire class 1,
- f fulfils Darboux condition,
- $f'_{\mathcal{T}}$ exists nearly everywhere,
- $f'_{\mathcal{T}}(x) \geq 0$ almost everywhere,

is monotone?

References

[1] Bruckner A. M., *An affirmative answer to a problem of Zahorski*, Mich. Math. Journ. 13 (1966), p. 15–26.
[2] Goldovski G., *Note sur les dérivées éxactes*, Rec. Math. Soc. Math. Moscou 35 (1928), p. 35–36.
[3] Jędrzejewski J. M., *O monotoniczności funkcji*, Master Thesis, Łódź University (1969).
[4] Leonard J. L., *Some conditions implying the monotonicity of a real function*, Rev. Roum. Math. Pures et Appl., 12 (1972), p. 757–780.
[5] O'Malley R. J., *Selective derivatives*, Acta Math. Acad. Sci. Hung., 29 (1977), p. 77–97.
[6] Mastalerz-Wawrzyńczak M., *On a certain condition of the monotonicity of functions*, Fund. Math. XCVII (1977), p. 188–198.
[7] Saks S., *Theory of the integral*, Monografie Matematyczne 7, Warszawa–Lwów, 1937.
[8] Świątkowski T., *On the conditions of monotonicity of functions*, Fund. Math. 59, (1966), p. 189–201.
[9] Świątkowski T., *On a certain generalization of the notion of the derivative*, Zesz. Nauk. P. Ł. 149, Matematyka 2.1 (1972), p. 89–103. (in Polish).
[10] Tolstov G., *Sur quelques propriétés des fonctions approximativement continues*, Rec. Math. N. S. 5 (1939), p. 637–645. (Mat. Sb.)
[11] Thomson B. S., *Real Functions*, Lectures Notes in Mathematics, No 1170, Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1985.
[12] Tonelli L., *Sulle derivate esatte*, Mem. Inst. Bologna (8) 8 (1930-31), p. 13–15.
[13] Zahorski Z., *Sur la première dérivée* Trans. Amer. Math. Soc. 69 (1950), p. 1–54.

Received: August 2017

E-mail address: jacek.m.jedrzejewski@gmail.com

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
61–71
DOI http://dx.doi.org/10.16926/m.2017.22.05

FRACTIONAL HEAT CONDUCTION IN AN INFINITE ROD WITH HEAT ABSORPTION PROPORTIONAL TO TEMPERATURE

YURIY POVSTENKO AND JOANNA KLEKOT

Abstract

The one-dimensional time-fractional heat conduction equation with heat absorption (heat release) proportional to temperature is considered. The Caputo time-fractional derivative is utilyzed. The fundamental solutions to the Cauchy and source problems are obtained using the Laplace transform with respect to time and the exponential Fourier transform with respect to the spatial coordinate. The numerical results are illustrated graphically.

1. Introduction

The classical heat conduction is based on the Fourier law which constitutes the linear dependence between the heat flux and the temperature gradient. As a result, the temperature T satisfies the standard parabolic equation

$$\frac{\partial T}{\partial t} = a\,\Delta T, \tag{1}$$

where t is time, Δ denotes the Laplace operator, a is the heat diffusivity coefficient.

If volume heat absorption proportional to temperature occurs in the medium, then instead of (1) we obtain [10]

$$\frac{\partial T}{\partial t} = a\,\Delta T - bT. \tag{2}$$

• *Yuriy Povstenko* — j.povstenko@ajd.czest.pl
Jan Długosz University in Częstochowa.
• *Joanna Klekot* — joanna.klekot@im.pcz.pl
Częstochowa University of Technology.

The cases $b > 0$ and $b < 0$ correspond to absorption and release of heat, respectively.

Time-nonlocal generalizations of the Fourier law were studied by many authors (see, for example, [5], [8], [14], [15], [18] and references therein). The time-nonlocal dependence between the heat flux and the temperature gradient with the "long-tail" power kernel [11], [12], [13] can be interpreted in terms of fractional integrals and derivatives and leads to the time-fractional heat conduction equation

$$\frac{\partial^\alpha T}{\partial t^\alpha} = a\,\Delta T, \quad 0 < \alpha \le 2. \tag{3}$$

The time-fractional counterpart of equation (2) has the following form

$$\frac{\partial^\alpha T}{\partial t^\alpha} = a\,\Delta T - bT, \quad 0 < \alpha \le 2. \tag{4}$$

Here $\frac{\partial^\alpha f}{\partial t^\alpha}$ is the Caputo fractional derivative [3], [6], [9]:

$$\frac{\mathrm{d}^\alpha f(t)}{\mathrm{d}t^\alpha} = \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-\tau)^{n-\alpha-1} \frac{\mathrm{d}^n f(\tau)}{\mathrm{d}\tau^n}\,\mathrm{d}\tau, \quad n-1 < \alpha < n,$$

where $\Gamma(x)$ is the gamma function, n is an integer.

2. Fundamental solution to the Cauchy problem

Consider the time-fractional heat conduction equation (4) in an infinite rod:

$$\frac{\partial^\alpha T}{\partial t^\alpha} = a\,\frac{\partial^2 T}{\partial x^2} - bT, \quad -\infty < x < \infty, \quad 0 < t < \infty, \quad 0 < \alpha \le 2, \tag{5}$$

under the initial conditions

$$t = 0: \quad T = p_0\,\delta(x), \quad 0 < \alpha \le 2, \tag{6}$$

$$t = 0: \quad \frac{\partial T}{\partial t} = 0, \qquad 1 < \alpha \le 2, \tag{7}$$

where $\delta(x)$ is the Dirac delta function. The constant coefficient p_0 is introduced to obtain the nondimensional quantities used in numerical calculations (see Eq. (14)).

The zero condition at infinity

$$\lim_{x \to \pm\infty} T(x,t) = 0 \tag{8}$$

is also assumed.

To solve the problem (5)–(8), the integral transform technique will be employed. Recall that the Laplace transform with respect to time t is defined as [2]

$$\mathcal{L}\{f(t)\} = f^*(s) = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,\mathrm{d}t,$$

$$\mathcal{L}^{-1}\{f^*(s)\} = f(t) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f^*(s)\,\mathrm{e}^{st}\,\mathrm{d}s, \quad t>0,$$

where the asterisk denotes the transform, s is the Laplace transform variable, c is the positive fixed number such that all the singularities of $f^*(s)$ lie to the left of the vertical line known as the Bromwich path of integration. The Caputo fractional derivative has the following Laplace transform rule [3], [6], [9]:

$$\mathcal{L}\left\{\frac{\mathrm{d}^\alpha f(t)}{\mathrm{d}t^\alpha}\right\} = s^\alpha f^*(s) - \sum_{k=0}^{n-1} f^{(k)}(0^+)s^{\alpha-1-k}, \quad n-1<\alpha<n.$$

The exponential Fourier transform with respect to the spatial coordinate x is used in the domain $-\infty<x<\infty$ and has the form [17]:

$$\mathcal{F}\{f(x)\} = \widetilde{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{ix\xi}\,\mathrm{d}x,$$

$$\mathcal{F}^{-1}\left\{\widetilde{f}(\xi)\right\} = f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \widetilde{f}(\xi)\,\mathrm{e}^{-ix\xi}\,\mathrm{d}\xi,$$

$$\mathcal{F}\left\{\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right\} = -\xi^2\widetilde{f}(\xi).$$

Applying the integral transforms to (5)–(8), we get

$$\widetilde{T}^* = \frac{p_0}{\sqrt{2\pi}}\,\frac{s^{\alpha-1}}{s^\alpha + a\xi^2 + b}. \tag{9}$$

To invert the Laplace transform, the following equation [3], [6], [9]

$$\mathcal{L}^{-1}\left\{\frac{s^{\alpha-1}}{s^\alpha + c}\right\} = E_\alpha(-c\,t^\alpha)$$

is used, where

$$E_\alpha(z) = \sum_{n=0}^{\infty}\frac{z^n}{\Gamma(\alpha\,n+1)}, \qquad \alpha>0, \quad z\in C,$$

is the Mittag-Leffler function in one parameter α.

Hence, the solution is expressed as

$$T(x,t) = \frac{p_0}{\pi}\int_0^\infty E_\alpha\left[-\left(a\xi^2+b\right)t^\alpha\right]\cos(x\xi)\,\mathrm{d}\xi. \tag{10}$$

Consider several particular cases of the solution (10). The classical heat conduction corresponds to the value $\alpha = 1$. In this instance the Mittag-Leffler function reduces to the exponential function

$$E_1\left[-\left(a\xi^2+b\right)t\right] = \exp\left[-\left(a\xi^2+b\right)t\right].$$

The integral in Eq. (10) is evaluated as [16]

$$\int_0^\infty \mathrm{e}^{-ax^2}\cos(bx)\,\mathrm{d}x = \frac{\sqrt{\pi}}{2\sqrt{a}}\exp\left(-\frac{b^2}{4a}\right). \tag{11}$$

Therefore,

$$T(x,t) = \frac{p_0}{2\sqrt{\pi a t}}\,\exp\left(-\frac{x^2}{4at} - bt\right). \tag{12}$$

The solution (12) is presented in [1], [10].

Another particular case is obtained for $\alpha = 1/2$. Using the following expression for the Mittag-Leffler function [14]

$$E_{1/2}(-x) = \frac{2}{\sqrt{\pi}}\int_0^\infty \mathrm{e}^{-u^2-2ux}\,\mathrm{d}u,$$

changing the order of integration and taking into account (11), we get

$$T(x,t) = \frac{p_0}{\pi\sqrt{2a}t^{1/4}}\int_0^\infty \frac{1}{\sqrt{u}}\,\exp\left(-u^2 - \frac{x^2}{8at^{1/2}u} - 2bt^{1/2}u\right)\mathrm{d}u. \tag{13}$$

For $b = 0$ the solution (13) coincides with the corresponding solution to the time-fractional diffusion-wave equation [14].

The results of numerical calculations are shown in Figs. 1–3. In calculations we have used the nondimensional quantitites

$$\overline{T} = \frac{\sqrt{a}t^{\alpha/2}}{p_0}\,T, \qquad \overline{x} = \frac{x}{\sqrt{a}t^{\alpha/2}}, \qquad \overline{b} = bt^{\alpha}. \tag{14}$$

3. Fundamental solution to the source problem

Next, we consider the time-fractional heat conduction equation with the source term

$$\frac{\partial^\alpha T}{\partial t^\alpha} = a\,\frac{\partial^2 T}{\partial x^2} - bT + w_0\,\delta(x)\,\delta(t), \quad -\infty < x < \infty, \quad 0 < \alpha \le 2, \tag{15}$$

under zero initial conditions

$$t = 0: \quad T = 0, \qquad 0 < \alpha \le 2, \tag{16}$$

$$t = 0: \quad \frac{\partial T}{\partial t} = 0, \qquad 1 < \alpha \le 2. \tag{17}$$

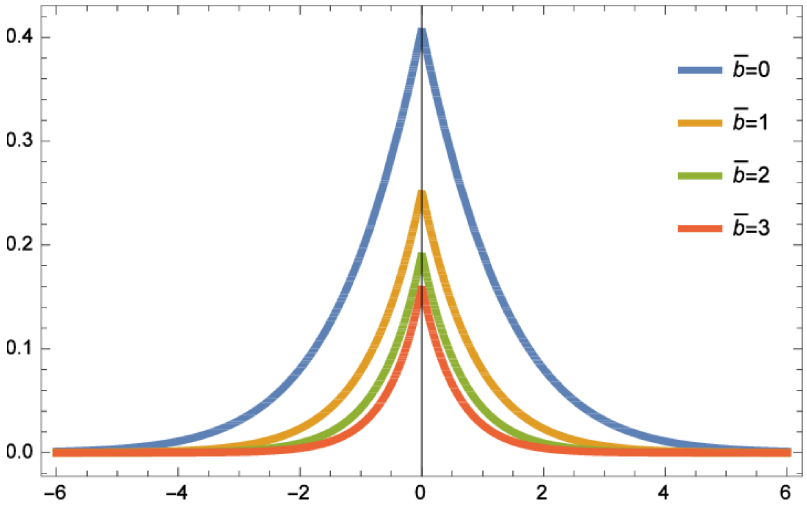


Figure 1 Solution to the Cauchy problem for $\alpha = 0.5$ and various values of $\bar{b}$.

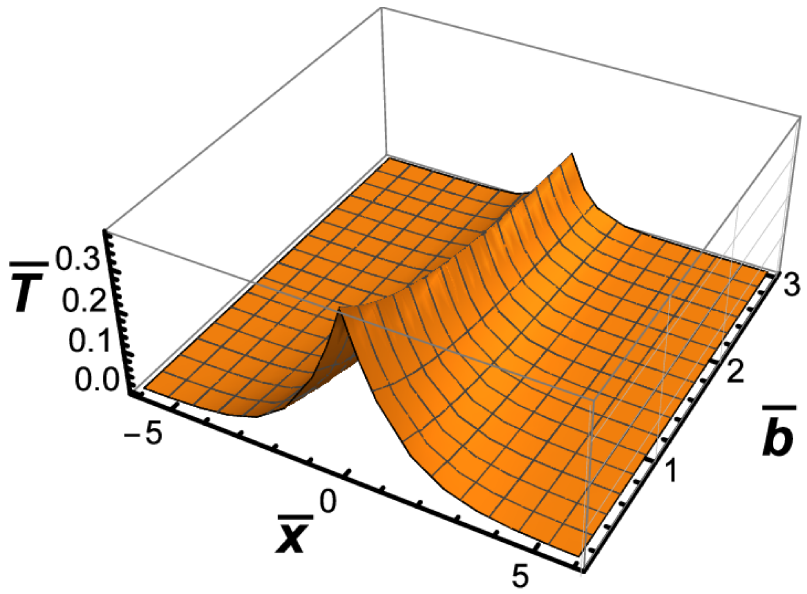


Figure 2 Solution to the Cauchy problem for $\alpha = 0.5$.

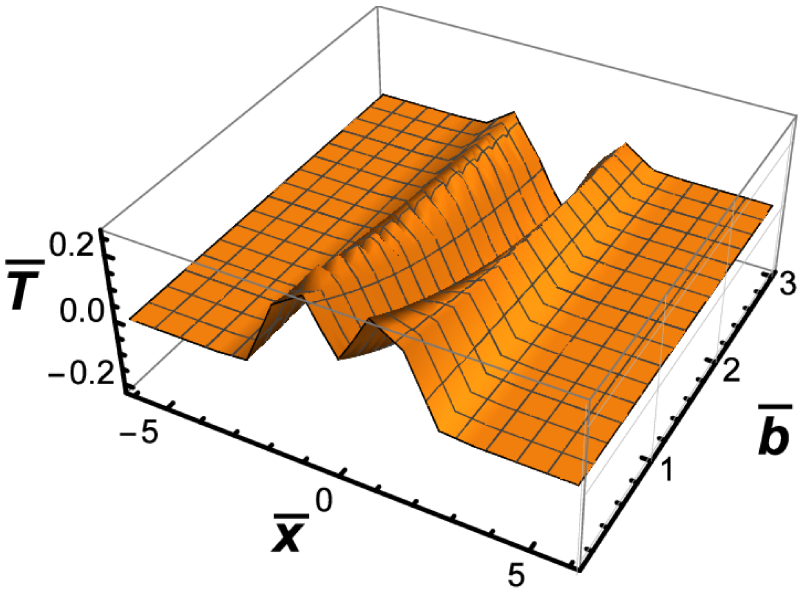


Figure 3 Solution to the Cauchy problem for $\alpha = 1.5$.

The constanmt coefficient w_0 is introduced to get the nondimensional quantity used in numerical calculations (see Eq. (24)).

The integral transform technique gives

$$\widetilde{T}^* = \frac{w_0}{\sqrt{2\pi}} \frac{1}{s^\alpha + a\xi^2 + b}. \tag{18}$$

By virtue of the fact that [3], [6], [9]

$$\mathcal{L}^{-1}\left\{\frac{s^{\alpha-\beta}}{s^\alpha + c}\right\} = t^{\beta-1} E_{\alpha,\beta}\left(-c\, t^\alpha\right),$$

where $E_{\alpha,\beta}(z)$ is the Mittag-Leffler function in two parameters α and β

$$E_{\alpha,\beta}(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(\alpha\, n + \beta)}, \qquad \alpha > 0, \quad \beta > 0, \quad z \in C,$$

the fundamental solution to the source problems is expressed as

$$T(x,t) = \frac{w_0\, t^{\alpha-1}}{\pi} \int_0^\infty E_{\alpha,\alpha}\left[-\left(a\xi^2 + b\right) t^\alpha\right] \cos(x\xi)\, \mathrm{d}\xi. \tag{19}$$

In the case of the classical heat conduction equation ($\alpha = 1$), the solutions (10) and (19) coincide and yield (12).

For heat conduction with $\alpha = 1/2$, we have [14]

$$E_{1/2,\,1/2}(-x) = \frac{2}{\sqrt{\pi}} \int_0^\infty \mathrm{e}^{-u^2-2ux}\, u \,\mathrm{d}u$$

and

$$T(x,t) = \frac{w_0}{\pi\sqrt{2a}t^{3/4}} \int_0^\infty \sqrt{u}\, \exp\left(-u^2 - \frac{x^2}{8aut^{1/2}} - 2bt^{1/2}u\right) \mathrm{d}u. \tag{20}$$

When the coefficient $b = 0$, the solution (20) coincides with the corresponding solution to the time-fractional diffusion-wave equation presented in the book [14].

For the wave equation corresponding to $\alpha = 2$, the Mittag-Leffler function is expressed as

$$E_{2,2}\left[-\left(a\,\xi^2 + b\right) t^2\right] = \frac{\sin\left(t\sqrt{a\,\xi^2 + b}\right)}{t\sqrt{a\,\xi^2 + b}}\,. \tag{21}$$

Taking into account the following integral [16]

$$\int_0^\infty \frac{\sin\left(c\sqrt{x^2+\gamma^2}\right)}{\sqrt{x^2+\gamma^2}}\, \cos\left(\beta x\right) \mathrm{d}x$$

$$= \begin{cases} \dfrac{\pi}{2}\, J_0\left(\gamma\sqrt{c^2-\beta^2}\right), & 0 < \beta < c\,, \\ 0, & 0 < c < \beta\,, \end{cases} \tag{22}$$

where J_0 is the Bessel function, from the general expression (19) we get the corresponding solution

$$T(x,t) = \begin{cases} \dfrac{w_0}{2\sqrt{a}}\, J_0\left[\sqrt{b\left(at^2 - x^2\right)/a}\right], & 0 < |x| < \sqrt{a}t, \\ 0, & \sqrt{a}t < |x| < \infty. \end{cases} \tag{23}$$

Dependence of the nondimensional fundamental solution to the source problem for the time-fractional heat conduction equation

$$\overline{T} = \frac{\sqrt{a}t^{\alpha/2-1}}{w_0}\, T \tag{24}$$

on the similarity variable $\overline{x}$ is presented in Figs. 4–7 for different values of the order of fractional derivative α and the parameter $\overline{b}$. Recall that other nondimensional quantities are the same as in Eq. (14).

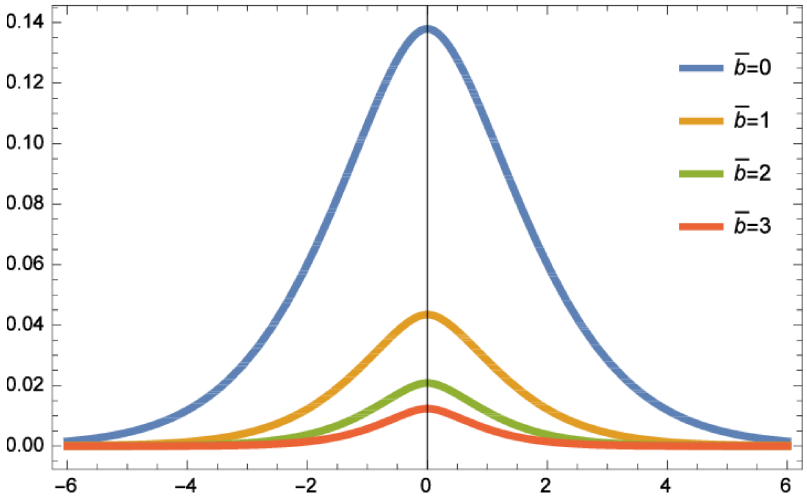


Figure 4 Solution to the source problem for $\alpha = 0.5$ and various values of $\bar{b}$.

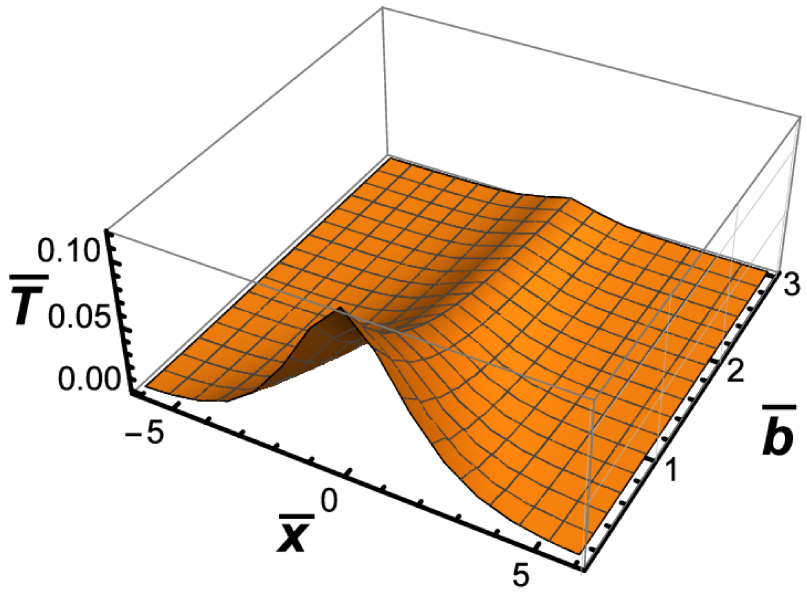


Figure 5 Solution to the source problem for $\alpha = 0.5$.

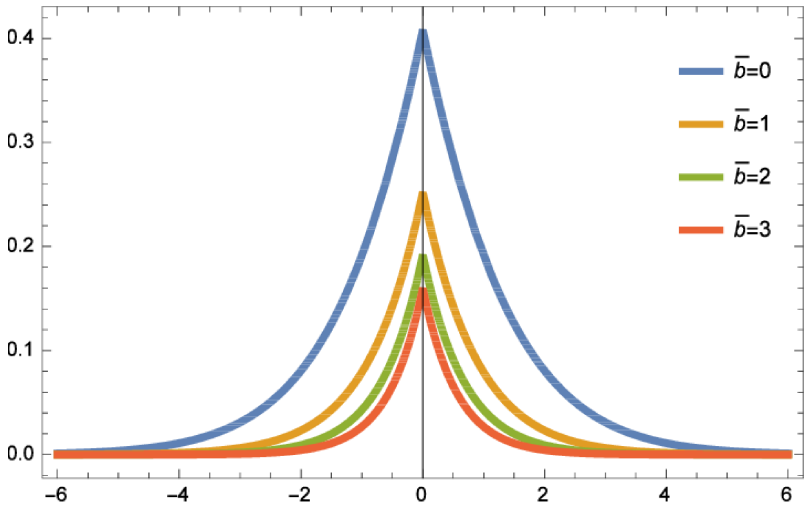


Figure 6 Solution to the source problem for $\alpha = 1.5$.

4. Conclusions

We have studied the time-fractional heat conduction equation with heat absorption proportional to temperature in the case of one Cartesian spatial coordinate. The fundamental solutions to the Cauchy and source problems have been obtained. Several particular cases have been considered, specifically the solutions to the classical heat conduction equation ($\alpha = 1$) and to the wave equation ($\alpha = 2$).

It should be emphasized that the time-fractional heat conduction equation with the Caputo fractional derivative of order $0 < \alpha \leq 2$ for $0 < \alpha \leq 1$ interpolates between the so-called localized heat conduction ($\alpha = 0$) and the standard heat conduction ($\alpha = 1$), whereas for $1 < \alpha \leq 2$ interpolates between the classical heat conduction ($\alpha = 1$) and the theory of heat conduction described by the wave equation ($\alpha = 2$). The wave equation for temperature is obtained as a consequence of time-nonlocal generalization of the Fourier law with constant memory kernel and no fading of memory; such theory was developed by Nigmatullin [7] and Green and Naghdi [4]. It is evident from the solution (23) that for the wave equation there appear two wave fronts at $x = \pm\sqrt{a}t$. The results of numerical calculations have been presented graphically.

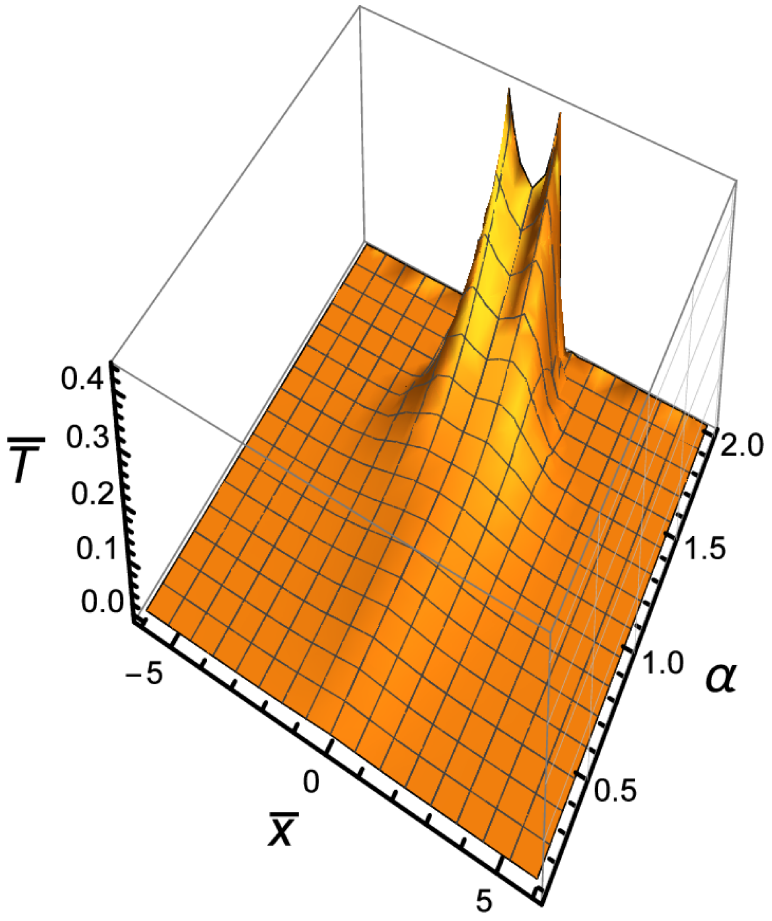


Figure 7 Solution to the source problem for $\bar{b} = 2$.

References

[1] A. G. Butkovskiy, *Characteristics of Systems with Distributed Parameters*. Nauka, Moscow, 1979 (In Russian).

[2] G. Doetsch, *Anleitung zum praktischer Gebrauch der Laplace-Transformation und der Z-Transformation*. Springer, München, 1967.

[3] R. Gorenflo, F. Mainardi, *Fractional calculus: Integral and differential equations of fractional order*, In: A. Carpinteri, F. Mainardi (Eds.), *Fractals and Fractional Calculus in Continuum Mechanics*, pp. 223–276. Springer, Wien, 1997.

[4] A. E. Green, P. M. Naghdi, *Thermoelasticity without energy dissipation*. J. Elast. **31** (1993), 189–208.

[5] M. E. Gurtin, A. C. Pipkin, *A general theory of heat conduction with finite wave speeds*. Arch. Rational Mech. Anal. **31** (1968), 113–126.

[6] A. A. Kilbas, H. M. Srivastava, J. J. Trujillo, *Theory and Applications of Fractional Differential Equations*. Elsevier, Amsterdam, 2006.
[7] R. R. Nigmatullin, *To the theoretical explanation of the "universal response"*. Phys. Stat. Sol. (b) **123** (1984), 739–745.
[8] R. R. Nigmatullin, *On the theory of relaxation with "remnant" temperature*. Phys. Stat. Sol. (b) **124** (1984), 389–393.
[9] I. Podlubny, *Fractional Differential Equations*. Academic Press, San Diego, 1999.
[10] A. D. Polyanin, *Handbook of Linear Partial Differential Equations for Engineers and Scientists*, Chapman & Hall/CRC, Boca Raton, 2002.
[11] Y. Povstenko, *Fractional heat conduction equation and associated thermal stresses*, J. Thermal Stresses **28** (2005), 83–102.
[12] Y. Povstenko, *Thermoelasticity which uses fractional heat conduction equation*, J. Math. Sci. **162** (2009), 296–305.
[13] Y. Povstenko, *Non-axisymmetric solutions to time-fractional diffusion-wave equation in an infinite cylinder*, Fract. Calc. Appl. Anal. **14** (2011), 418–435.
[14] Y. Povstenko, *Linear Fractional Diffusion-Wave Equation for Scientists and Engineers*. Birkhäser, New York, 2015.
[15] Y. Povstenko, *Fractional Thermoelasticity*. Springer, New York, 2015.
[16] A. P. Prudnikov, Yu. A. Brychkov, O. I. Marichev, *Integrals and Series, Vol. 1: Elementary Functions*. Gordon and Breach Science Publishers, Amsterdam, 1986.
[17] I. N. Sneddon, *The Use of Integral Transforms*. McGraw-Hill, New York, 1972.
[18] K. K. Tamma, X. Zhou, *Macroscale and microscale thermal transport and thermomechanical interaction: some noteworthy perspectives*. J. Thermal Stresses **21** (1998), 405–449.

Received: June 2017

Yuriy Povstenko
Jan Długosz University in Częstochowa,
Faculty of Mathematics and Narural Sciences,
Institute of Mathematics and Computer Science,
al. Armii Krajowej 13/15,
42-200, Częstochowa, Poland
E-mail address: j.povstenko@ajd.czest.pl

Joanna Klekot
Częstochowa University of Technology,
Faculty of Mechanical Engineering and Computer Science,
Institute of Mathematics,
al. Armii Krajowej 21,
42-200, Częstochowa, Poland
E-mail address: joanna.klekot@im.pcz.pl

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
73–81
DOI http://dx.doi.org/10.16926/m.2017.22.06

SOME PROPERTIES OF GENERALIZED TRIBONACCI QUATERNIONS

ANETTA SZYNAL-LIANA AND IWONA WŁOCH

Abstract

In this paper we introduce distinct types of Tribonacci quaternions. We describe dependences between them and we give some their properties also related to a matrix representation.

1. Introduction

Let $\mathbb{H}$ be the set of quaternions z of the form

$$z = a + bi + cj + dk, \tag{1}$$

where $a, b, c, d \in \mathbb{R}$ and i, j, k are complex operators such that

$$i^2 = j^2 = k^2 = ijk = -1 \tag{2}$$

and

$$ij = -ji = k,\ jk = -kj = i,\ ki = -ik = j. \tag{3}$$

If $z_1 = a_1 + b_1 i + c_1 j + d_1 k$ and $z_2 = a_2 + b_2 i + c_2 j + d_2 k$ are any two quaternions then the equality, the addition, the substraction and the multiplication by scalar are defined as follows.

Equality: $z_1 = z_2$ only if $a_1 = a_2$, $b_1 = b_2$, $c_1 = c_2$, $d_1 = d_2$,
addition: $z_1 + z_2 = (a_1 + a_2) + (b_1 + b_2)\, i + (c_1 + c_2)\, j + (d_1 + d_2)\, k$,
substraction: $z_1 - z_2 = (a_1 - a_2) + (b_1 - b_2)\, i + (c_1 - c_2)\, j + (d_1 - d_2)\, k$,
multiplication by scalar $s \in \mathbb{R}$: $sz_1 = sa_1 + sb_1 i + sc_1 j + sd_1 k$.
The quaternion multiplication is defined using (2).
The conjugate of a quaternion z is defined by

$$z^* = (a + bi + cj + dk)^* = a - bi - cj - dk. \tag{4}$$

• *Anetta Szynal-Liana* — e-mail: aszynal@prz.edu.pl
Rzeszow University of Technology.
• *Iwona Włoch* — e-mail: iwloch@prz.edu.pl
Rzeszow University of Technology.

Moreover we use the following notation for
real part: $\Re z = (z + z^*)/2 = a \in \mathbb{R}$,
imaginary part: $\Im z = (z - z^*)/2 = bi + cj + dk \in \mathbb{H}$.
The norm of a quaternion z is defined by

$$N(z) = a^2 + b^2 + c^2 + d^2. \tag{5}$$

For the basics on the quaternions theory, see [12].

Let F_n be the nth Fibonacci number defined recursively by $F_n = F_{n-1} + F_{n-2}$ for $n \geq 2$ with the initial terms $F_0 = F_1 = 1$. There are many numbers defined by linear recurrence relations and they are also named as numbers of the Fibonacci type. We list some of them.

- $L_n = L_{n-1} + L_{n-2}$, for $n \geq 2$ with $L_0 = 2$, $L_1 = 1$ – Lucas numbers,
- $P_n = 2P_{n-1} + P_{n-2}$, for $n \geq 2$ with $P_0 = 0$, $P_1 = 1$ – Pell numbers,
- $Q_n = 2Q_{n-1} + Q_{n-2}$, for $n \geq 2$ with $Q_0 = 2$, $Q_1 = 2$ – Pell-Lucas numbers,
- $J_n = J_{n-1} + 2J_{n-2}$, for $n \geq 2$ with $J_0 = 0$, $J_1 = 1$ – Jacobsthal numbers,
- $j_n = j_{n-1} + 2j_{n-2}$, for $n \geq 2$ with $j_0 = 2$, $j_1 = 1$ – Jacobsthal-Lucas numbers.

These numbers have many applications in distinct areas of mathematics also in the quaternions theory.

In 1963 Horadam [5] introduced nth Fibonacci and Lucas quaternions. Three decades later in [6] Horadam mentioned about the possibility of introducing Pell quaternions and generalized Pell quaternions. Interesting results concerning Pell quaternions, Pell-Lucas quaternions have been obtained quite recently and can be found in [2], [11]. Jacobsthal quaternions and Jacobsthal-Lucas quaternions were introduced in [10].

In the most recent paper of G. Cerda-Morales (see [1]) we can find the definition of generalized Tribonacci quaternions due to their coefficients. The definition of these quaternions is based on the definition of generalized Tribonacci numbers V_n

$$V_n = rV_{n-1} + sV_{n-2} + tV_{n-3}, \ for \ n \geq 3,$$

where $V_0 = a$, $V_1 = b$, $V_2 = c$ are arbitrary integers and r, s, t are real numbers. For $r = s = t = 1$ we have the set of quaternions defined in this paper. We present some properties of generalized Tribonacci quaternions, in particular relations between them.

2. The Tribonacci numbers

Let $n \geq 0$ be integer. The nth Tribonacci number T_n is defined by $T_0 = 1$, $T_1 = 1$, $T_2 = 2$, and

$$T_n = T_{n-1} + T_{n-2} + T_{n-3}, \text{ for } n \geq 3. \tag{6}$$

Tribonacci numbers have been firstly defined by Feinberg in 1963, see [3].

The characteristic equation of (6) has the form $x^3 - x^2 - x - 1 = 0$ and it has roots

$$\alpha = \frac{1 + \sqrt[3]{19 + 3\sqrt{33}} + \sqrt[3]{19 - 3\sqrt{33}}}{3},$$
$$\beta = \frac{1 + \omega\sqrt[3]{19 + 3\sqrt{33}} + \omega^2\sqrt[3]{19 - 3\sqrt{33}}}{3},$$
$$\gamma = \frac{1 + \omega^2\sqrt[3]{19 + 3\sqrt{33}} + \omega\sqrt[3]{19 - 3\sqrt{33}}}{3},$$

where

$$\omega = \frac{-1 + \epsilon\sqrt{3}}{2},\ \epsilon^2 = -1.$$

Hence the Binet formula for the Tribonacci number T_n has the form

$$T_n = \frac{\alpha^{n+2}}{(\alpha - \beta)(\alpha - \gamma)} + \frac{\beta^{n+2}}{(\beta - \alpha)(\beta - \gamma)} + \frac{\gamma^{n+2}}{(\gamma - \alpha)(\gamma - \beta)}. \tag{7}$$

There are some versions of Tribonacci numbers defined by the same linear recurrence relation as T_n but with different initial conditions.

The nth generalized Tribonacci number t_n is a number defined recursively by the recurrence relation of the form $t_n = t_{n-1} + t_{n-2} + t_{n-3}$ for $n \geq 3$ with fixed t_0, t_1, t_2. For special value of t_0, t_1, t_2 we obtain different kinds of Tribonacci numbers. If $t_0 = 1$, $t_1 = 1$, $t_2 = 2$ then we obtain the definition of T_n. Apart Tribonacci numbers T_n we define other kinds of Tribonacci numbers namely numbers R_n, S_n and U_n. For $n \geq 0$ we define three types of Tribonacci numbers as follows

$R_0 = 3,\ R_1 = 1,\ R_2 = 3$ and $R_n = R_{n-1} + R_{n-2} + R_{n-3}$ for $n \geq 3$,

$S_0 = 3,\ S_1 = 2,\ S_2 = 5$ and $S_n = S_{n-1} + S_{n-2} + S_{n-3}$ for $n \geq 3$,

$U_0 = 0,\ U_1 = 1,\ U_2 = 2$ and $U_n = U_{n-1} + U_{n-2} + U_{n-3}$ for $n \geq 3$.

The Table 1 presents values of these Tribonacci numbers for $n = 0, 1, ..., 10$.

n	0	1	2	3	4	5	6	7	8	9	10
T_n	1	1	2	4	7	13	24	44	81	149	274
R_n	3	1	3	7	11	21	39	71	131	241	443
S_n	3	2	5	10	17	32	59	108	199	366	673
U_n	0	1	2	3	6	11	20	37	68	125	230

Table 1.

Above Tribonacci numbers were considered in [3], [7], [8], [9] where among other Binet formulas for them were found. Moreover in [8] some relations between Tribonacci numbers were given. We recall these dependences

$$R_n = T_{n-1} + 2T_{n-2} + 3T_{n-3}, \ for\ n \geq 3 \tag{8}$$

$$S_n = 3T_n - T_{n-1}, \ for\ n \geq 1 \tag{9}$$

$$U_n = T_{n-1} + T_{n-2}, \ for\ n \geq 2 \tag{10}$$

$$\sum_{l=0}^{n} U_l = T_{n+1} - 1 \tag{11}$$

$$\sum_{l=1}^{n} R_l = 2U_{n+1} + U_{n-1} - 3 \tag{12}$$

$$\sum_{l=0}^{n} S_l = \frac{3U_{n+2} + 2U_{n+1} - U_n - 2}{2} \tag{13}$$

$$\sum_{l=0}^{n} T_l = \frac{U_{n+2} + U_{n+1} - 1}{2}. \tag{14}$$

From the above identities we can obtain other relations

$$2T_n = U_{n+1} + U_{n-1}, \ for\ n \geq 1 \tag{15}$$

$$\sum_{l=1}^{n} R_l = 3T_n + T_{n-1} - 3 \tag{16}$$

$$\sum_{l=0}^{n} S_l = T_{n+2} + 2T_n - 1 \tag{17}$$

$$\sum_{l=0}^{n} T_l = \frac{T_{n+2} + T_n - 1}{2}. \tag{18}$$

3. The Tribonacci quaternions

For $n \geq 0$ the nth generalized Tribonacci quaternion TQ_n is defined as

$$TQ_n = t_n + it_{n+1} + jt_{n+2} + kt_{n+3}. \tag{19}$$

In particular for Tribonacci numbers we obtain distinct Tribonacci quaternions. Using presented earlier Tribonacci numbers T_n, S_n, R_n and U_n we have four types of Tribonacci quaternions. Then

$$TTQ_n = T_n + iT_{n+1} + jT_{n+2} + kT_{n+3} \tag{20}$$

$$TRQ_n = R_n + iR_{n+1} + jR_{n+2} + kR_{n+3} \tag{21}$$

$$TSQ_n = S_n + iS_{n+1} + jS_{n+2} + kS_{n+3} \tag{22}$$

$$TUQ_n = U_n + iU_{n+1} + jU_{n+2} + kU_{n+3} \tag{23}$$

Firstly we give relations between Tribonacci quaternions.

Theorem 1. *Let n be an integer. Then*

(i) $TRQ_n = TTQ_{n-1} + 2TTQ_{n-2} + 3TTQ_{n-3},\ for\ n \geq 3,$
(ii) $TSQ_n = 3TTQ_n - TTQ_{n-1},\ for\ n \geq 1,$
(iii) $TUQ_n = TTQ_{n-1} + TTQ_{n-2},\ for\ n \geq 2,$
(iv) $2TTQ_n = TUQ_{n+1} + TUQ_{n-1},\ for\ n \geq 1.$

Proof. (i) Using (21) and (8) we have

$$\begin{aligned} TRQ_n &= R_n + iR_{n+1} + jR_{n+2} + kR_{n+3} = \\ &= (T_{n-1} + 2T_{n-2} + 3T_{n-3}) + i\,(T_n + 2T_{n-1} + 3T_{n-2}) + \\ &\quad + j\,(T_{n+1} + 2T_n + 3T_{n-1}) + k\,(T_{n+2} + 2T_{n+1} + 3T_n) = \\ &= (T_{n-1} + iT_n + jT_{n+1} + kT_{n+2}) + \\ &\quad + 2\,(T_{n-2} + iT_{n-1} + jT_n + kT_{n+1}) + \\ &\quad + 3\,(T_{n-3} + iT_{n-2} + jT_{n-1} + kT_n) = \\ &= TTQ_{n-1} + 2TTQ_{n-2} + 3TTQ_{n-3}. \end{aligned}$$

In the same way, using (9), (10) and (15) one can easily prove identities (ii)-(iv). □

The next theorem gives formulas for sums of Tribonacci quaternions.

Theorem 2. *Let n be an integer. Then*

(i) $\sum_{l=0}^{n} TUQ_l = TTQ_{n+1} - TTQ_0,$
(ii) $\sum_{l=1}^{n} TRQ_l = 2TUQ_{n+1} + TUQ_{n-1} - (3 + 4i + 7j + 14k),$
(iii) $\sum_{l=0}^{n} TSQ_l = \frac{3TUQ_{n+2} + 2TUQ_{n+1} - TUQ_n}{2} - (1 + 4i + 6j + 11k),$

(iv) $\sum_{l=0}^{n} TTQ_l = \frac{TUQ_{n+2} + TUQ_{n+1} - (1+3i+5j+9k)}{2}$,

(v) $\sum_{l=1}^{n} TRQ_l = 3TTQ_n + TTQ_{n-1} - (3+4i+7j+14k)$,

(vi) $\sum_{l=0}^{n} TSQ_l = TTQ_{n+2} + 2TTQ_n - (1+4i+6j+11k)$,

(vii) $\sum_{l=0}^{n} TTQ_l = \frac{TTQ_{n+2} + TTQ_n - (1+2i+4j+8k)}{2}$.

Proof. (i) Using (23) and (11) we have

$$\begin{aligned}
\sum_{l=0}^{n} TUQ_l &= TUQ_0 + TUQ_1 + \ldots + TUQ_n = \\
&= (U_0 + iU_1 + jU_2 + kU_3) + \\
&+ (U_1 + iU_2 + jU_3 + kU_4) + \ldots + \\
&+ (U_n + iU_{n+1} + jU_{n+2} + kU_{n+3}) = \\
&= (U_0 + U_1 + \ldots + U_n) + \\
&+ i\,(U_1 + U_2 + \ldots + U_{n+1}) + \\
&+ j\,(U_2 + U_3 + \ldots + U_{n+2}) + \\
&+ k\,(U_3 + U_4 + \ldots + U_{n+3}) = \\
&= T_{n+1} - 1 + i(T_{n+2} - 1 - U_0) + j(T_{n+3} - 1 - U_0 - U_1) + \\
&+ k\,(T_{n+4} - 1 - U_0 - U_1 - U_2) = \\
&= T_{n+1} + iT_{n+2} + jT_{n+3} + kT_{n+4} - (1 + i + 2j + 4k),
\end{aligned}$$

which ends the proof. In the same way one can easily prove (ii)-(vii). □

4. The Binet formula and a matrix representation

Using the Binet formula for Tribonacci numbers T_n we can give the direct formula for nth Tribonacci quaternion

$$\begin{aligned}
TTQ_n &= \tfrac{\alpha^{n+2}}{(\alpha-\beta)(\alpha-\gamma)} + \tfrac{\beta^{n+2}}{(\beta-\alpha)(\beta-\gamma)} + \tfrac{\gamma^{n+2}}{(\gamma-\alpha)(\gamma-\beta)} + \\
&+ i\left(\tfrac{\alpha^{n+3}}{(\alpha-\beta)(\alpha-\gamma)} + \tfrac{\beta^{n+3}}{(\beta-\alpha)(\beta-\gamma)} + \tfrac{\gamma^{n+3}}{(\gamma-\alpha)(\gamma-\beta)}\right) + \\
&+ j\left(\tfrac{\alpha^{n+4}}{(\alpha-\beta)(\alpha-\gamma)} + \tfrac{\beta^{n+4}}{(\beta-\alpha)(\beta-\gamma)} + \tfrac{\gamma^{n+4}}{(\gamma-\alpha)(\gamma-\beta)}\right) + \\
&+ k\left(\tfrac{\alpha^{n+5}}{(\alpha-\beta)(\alpha-\gamma)} + \tfrac{\beta^{n+5}}{(\beta-\alpha)(\beta-\gamma)} + \tfrac{\gamma^{n+5}}{(\gamma-\alpha)(\gamma-\beta)}\right).
\end{aligned}$$

For other types of Tribonacci quaternions we can obtain analogous formulas, we omit their presentations.

Matrix representations play an important role in the theory of numbers defined by the recurrence relations, see for example [4]. We give a matrix generator also for Tribonacci quaternions TQ_n.

Theorem 3. *Let*

$$T = \begin{bmatrix} -TQ_0 - TQ_1 + TQ_2 & TQ_1 - TQ_0 & TQ_0 \\ TQ_0 & TQ_2 - TQ_1 & TQ_1 \\ TQ_1 & TQ_0 + TQ_1 & TQ_2 \end{bmatrix} \tag{24}$$

and

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{25}$$

Then

$$TA^n = \begin{bmatrix} TQ_{n-1} & TQ_{n-2} + TQ_{n-1} & TQ_n \\ TQ_n & TQ_{n-1} + TQ_n & TQ_{n+1} \\ TQ_{n+1} & TQ_n + TQ_{n+1} & TQ_{n+2} \end{bmatrix} \text{ for } n \geq 2. \tag{26}$$

Proof. (by induction on n) If $n = 2$ we have

$$A^2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

and

$$TA^2 = \begin{bmatrix} -TQ_0 - TQ_1 + TQ_2 & TQ_1 - TQ_0 & TQ_0 \\ TQ_0 & TQ_2 - TQ_1 & TQ_1 \\ TQ_1 & TQ_0 + TQ_1 & TQ_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \end{bmatrix} =$$

$$= \begin{bmatrix} TQ_1 & TQ_0 + TQ_1 & TQ_2 \\ TQ_2 & TQ_1 + TQ_2 & TQ_0 + TQ_1 + TQ_2 \\ TQ_0 + TQ_1 + TQ_2 & TQ_0 + TQ_1 + 2TQ_2 & TQ_0 + 2TQ_1 + 2TQ_2 \end{bmatrix} =$$

$$= \begin{bmatrix} TQ_1 & TQ_0 + TQ_1 & TQ_2 \\ TQ_2 & TQ_1 + TQ_2 & TQ_3 \\ TQ_3 & TQ_2 + TQ_3 & TQ_4 \end{bmatrix}.$$

Assume that

$$TA^n = \begin{bmatrix} TQ_{n-1} & TQ_{n-2} + TQ_{n-1} & TQ_n \\ TQ_n & TQ_{n-1} + TQ_n & TQ_{n+1} \\ TQ_{n+1} & TQ_n + TQ_{n+1} & TQ_{n+2} \end{bmatrix}.$$

We shall show that

$$TA^{n+1} = \begin{bmatrix} TQ_n & TQ_{n-1} + TQ_n & TQ_{n+1} \\ TQ_{n+1} & TQ_n + TQ_{n+1} & TQ_{n+2} \\ TQ_{n+2} & TQ_{n+1} + TQ_{n+2} & TQ_{n+3} \end{bmatrix}.$$

Using induction's hypothesis we have

$$TA^{n+1} = TA^nA = \begin{bmatrix} TQ_{n-1} & TQ_{n-2}+TQ_{n-1} & TQ_n \\ TQ_n & TQ_{n-1}+TQ_n & TQ_{n+1} \\ TQ_{n+1} & TQ_n+TQ_{n+1} & TQ_{n+2} \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} TQ_n & TQ_{n-1}+TQ_n & TQ_{n-2}+TQ_{n-1}+TQ_n \\ TQ_{n+1} & TQ_n+TQ_{n+1} & TQ_{n-1}+TQ_n+TQ_{n+1} \\ TQ_{n+2} & TQ_{n+1}+TQ_{n+2} & TQ_n+TQ_{n+1}+TQ_{n+2} \end{bmatrix} =$$

$$= \begin{bmatrix} TQ_n & TQ_{n-1}+TQ_n & TQ_{n+1} \\ TQ_{n+1} & TQ_n+TQ_{n+1} & TQ_{n+2} \\ TQ_{n+2} & TQ_{n+1}+TQ_{n+2} & TQ_{n+3} \end{bmatrix},$$

which ends the proof. □

References

[1] G. Cerda-Morales, *On a Generalization for Tribonacci Quaternions*, arXiv:1707.04081.

[2] C. B. Çimen, A. İpek, *On Pell Quaternions and Pell-Lucas Quaternions*, Advances in Applied Clifford Algebras 26 (2016), 39-51. doi: https://doi.org/10.1007/s00006-015-0571-8

[3] M. Feinberg, *Fibonacci - Tribonacci*, The Fibonacci Quarterly 1.3 (1963), 71-74.

[4] V. E. Hoggatt Jr, I. D. Ruggles, *A primer on the Fibonacci sequence, Part III*, The Fibonacci Quarterly 1.3 (1963), 61-65.

[5] A. F. Horadam, *Complex Fibonacci Numbers and Fibonacci Quaternions*, American Mathematical Monthly 70 (1963), 289-291.

[6] A. F. Horadam, *Quaternion recurrence relations*, Ulam Quarterly 2 (1993), 23-33.

[7] G. Köhler, *Generating functions of Fibonacci-like sequences and decimal expansions of some fractions*, The Fibonacci Quarterly 23.1 (1985), 29-35.

[8] P.-Y. Lin, *De Moivre-type identities for the Tribonacci numbers*, The Fibonacci Quarterly 26.2 (1988), 131-134.

[9] W. R. Spickerman, *Binet's formula for the Tribonacci sequence*, The Fibonacci Quarterly 20.2 (1982), 118-120.

[10] A. Szynal-Liana, I. Włoch, *A note on Jacobsthal quaternions*, Advances in Applied Clifford Algebras 26 (2016), 441-447. doi: https://doi.org/10.1007/s00006-015-0622-1

[11] A. Szynal-Liana, I. Włoch, *The Pell quaternions and the Pell octonions*, Advances in Applied Clifford Algebras 26 (2016), 435-440. doi: https://doi.org/10.1007/s00006-015-0570-9

[12] J. P. Ward, *Quaternions and Cayley Numbers: Algebra and Applications*, Kluwer Academic Publishers, London, 1997.

Received: November 2017

Anetta Szynal-Liana
Rzeszow University of Technology,
Faculty of Mathematics and Applied Physics,
al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland
E-mail address: aszynal@prz.edu.pl

Iwona Włoch
Rzeszow University of Technology,
Faculty of Mathematics and Applied Physics,
al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland
E-mail address: iwloch@prz.edu.pl

Part III
Computer Science

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
85–97
DOI http://dx.doi.org/10.16926/m.2017.22.07

A GPGPU–BASED SIMULATOR FOR PRISM: STATISTICAL VERIFICATION OF RESULTS OF PMC

MARCIN COPIK, ARTUR RATAJ, AND BOŻENA WOŹNA-SZCZEŚNIAK

Abstract

We describe a GPGPU–based Monte Carlo simulator integrated with `Prism`. It supports Markov chains with discrete or continuous time and a subset of properties expressible in PCTL, CSL and their variants extended with rewards. The simulator allows an automated statistical verification of results obtained using `Prism`'s formal methods.

Keywords: GPGPU, Monte Carlo simulation, `Prism`, probabilistic model checking, statistical model checking, probabilistic logics.

1. Introduction

We present a GPGPU–based simulator which extends the model checker `Prism` [9]. The simulator uses the Monte Carlo method for a statistical probabilistic model checking [14, 10] (SPMC). SPMC involves a generation of a large number of random paths (i.e. samples) in a probabilistic Markov chain, evaluating a given property on each path, and finally finding an average of these evaluations, which approximate a correct value of the property. Monte Carlo methods typically are able to precisely compute confidence intervals (CI) around the approximated value.

The GPGPU simulator (further called $\mathcal{S}^G$) is integrated with `Prism`, which allows to check a single model implementation using either one of `Prism`'s probabilistic model checking (PMC) methods, or $\mathcal{S}^G$. $\mathcal{S}^G$ supports the same models and properties as the `Prism`'s CPU–based simulator (further denoted $\mathcal{S}^C$), yet the latter, lacking GPGPU acceleration and on–the–fly compilation of the model, is considerably slower.

• *Marcin Copik* — mcopik@gmail.com
RWTH Aachen University.
• *Artur Rataj* — arturrataj@gmail.com
Polish Academy of Sciences, Institute of Theoretical and Applied Informatics.
• *Bożena Woźna-Szcześniak* — b.wozna@ajd.czest.pl
Jan Długosz University in Częstochowa.

Beside the simulator itself, we present its simple application: an automated method of verifying property values computed using `Prism`'s formal methods. Namely, the user may request, that certain property classes be computed in two steps:

(1) A PMC step. A property is computed using one of `Prism` 's formal PMC methods;
(2) An automated statistical verification (ASV) step. The obtained property value v is statistically evaluated using $\mathcal{S}^G$; it is then checked if v fits into a number of confidence intervals (CI) of various confidence levels.

Because a significant imprecision in a PMC method is typically caused by some mechanism for reducing computational complexity, the user might tend to disable such a mechanism, rather than wait for a statistical verification. This is why it is crucial that the verifying simulator be fast: it should effectively save user's time, by providing data in tight CI within a short time.

The paper is constructed as follows. Section 2 describes what is currently supported by $\mathcal{S}^G$. In Section 3 we provide a description of some implementation details of $\mathcal{S}^G$. In Section 4 ASV is discussed. Section 5 presents a case study. Finally, the last section concludes the paper.

2. Supported models and properties

$\mathcal{S}^G$ accepts a specific set of properties, for two finite probabilistic classes of Markov chains: a *discrete-time Markov chain* (DTMC) and a *continuous-time Markov chain* (CTMC). Unlike DTMC, where each transition corresponds to a discrete time-step, in a CTMC transitions occur in continuous time given by a negative exponential distribution. Both of the classes can be enriched with rewards structures, resulting respectively in rDTMC and rCTMC. A reward structure allows to specify two distinct types of rewards: state (instantaneous) and transition (cumulative) ones, assigned respectively to states and transitions by means of a reward function. Formal definitions of all of the above systems can be found e.g. in [8].

The temporal logics *Probabilistic Computation Tree Logic* (PCTL) [6] and *Continuous Stochastic Logic* (CSL) [1] can be used to specify properties for respectively DTMCs and CTMCs. $\mathcal{S}^G$ recognises only flat subsets of each logic. We will refer to these subsets as respectively FlatPCTL and FlatCSL.

Definition 1 (Syntax of FlatPCTL)**.** *Let $a \in \mathcal{AP}$ be an atomic proposition, $\sim \in \{<, \leq, \geq, >\}$, $p \in [0,1]$ a probability bound, and k is a non–negative integer or ∞. The syntax of FlatPCTL is defined inductively as follows:*

$$\phi ::= \mathtt{P}_{\sim p}[\psi], \quad \psi ::= \mathtt{X}\phi_1 \mid \mathtt{G}^{\leq k}\phi_1 \mid \mathtt{F}^{\leq k}\phi_1 \mid \phi_1 \mathtt{U}^{\leq k}\phi_1 \mid \phi_1 \mathtt{R}^{\leq k}\phi_1,$$
$$\phi_1 ::= a \mid \phi_1 \wedge \phi_1 \mid \neg\phi_1.$$

In the syntax above, we distinguish between state formulae ϕ, ϕ_1 and path formulae ψ, which are evaluated over states and paths, respectively. A property of a model is always expressed as a state formula. The path modalities (i.e., *next state* – $\mathtt{X}$, *bounded globally* – $\mathtt{G}$, *bounded eventually* – $\mathtt{F}^{\leq k}$, *bounded until* – $\mathtt{U}^{\leq k}$, and *bounded release* – $\mathtt{R}^{\leq k}$), which are standard in temporal logics, can occur only within the scope of the *probabilistic operator* $\mathtt{P}_{\sim p}[\cdot]$.

Intuitively, a state s satisfies $\mathtt{P}_{\sim p}[\psi]$ if the probability of taking a path from s satisfying path formula ψ meets the bound $\sim p$. Next, $\mathtt{X}\phi$ is true if ϕ is satisfied in the next state; $\mathtt{G}^{\leq k}\phi$ is true if ϕ holds for all time-steps that are less or equal to k; $\mathtt{F}^{\leq k}\phi$ is true if ϕ is satisfied within k time-steps; $\phi_1 \mathtt{U}^{\leq k}\phi_2$ is true if ϕ_2 is satisfied within k time-steps and ϕ_1 is true from now on until ϕ_2 becomes true. $\phi_1 \mathtt{R}^{\leq k}\phi_2$ is true if either ϕ_1 is satisfied within k time-steps and ϕ_2 is true from now on up to the point where ϕ_1 becomes true, or ϕ_2 holds for all time-steps that are less or equal to k. The formal semantics over DTMC can be found e.g. in [6, 8].

$\mathcal{S}^G$ supports also an extension of FlatPCTL allowing specifications over reward structures by means of the following state formulae: $\mathtt{R}_{\sim r}[\mathtt{C}^{\leq k}]$ | $\mathtt{R}_{\sim r}[\mathtt{I}^{=k}]$ | $\mathtt{R}_{\sim r}[\mathtt{F}\phi]$ where $\sim \in \{<, \leq, \geq, >\}$, $r \in \mathbb{R}_{\geq 0}$, $k \in \mathbb{N}$, and ϕ is a FlatPCTL formula.

The formal semantics over rDTMC can be found in [8]. Here we only provide an intuition. Namely, a state s of an rDTMC satisfies $\mathtt{R}_{\sim r}[\mathtt{C}^{\leq k}]$, if from state s the expected reward *cumulated* after k time-steps satisfies $\sim r$. Next, a state s of an rDTMC satisfies $\mathtt{R}_{\sim r}[\mathtt{I}^{=k}]$, if from state s the expected state reward at time-step k satisfies $\sim r$. Finally, a state s of an rDTMC satisfies $\mathtt{R}_{\sim r}[\mathtt{F}\phi]$, if from state s the expected reward cumulated before a state satisfying ϕ is reached meets the bound $\sim r$.

Definition 2 (Syntax of FlatCSL)**.** *Let a and p be as in Definition 1, and I be an interval of $\mathbb{R}_{\geq 0}$. The syntax of FlatCSL is defined inductively as follows:*

$$\phi ::= \mathtt{P}_{\sim\ p}[\psi], \quad \psi ::= \mathtt{X}\phi_1 \mid \mathtt{G}^I\phi_1 \mid \mathtt{F}^I\phi_1 \mid \phi_1\mathtt{U}^I\phi_1 \mid \phi_1\mathtt{R}^I\phi_1,$$
$$\phi_1 ::= a \mid \phi_1 \wedge \phi_1 \mid \neg\phi_1.$$

Satisfying $\mathtt{P}_{\sim p}[\psi]$ and path modalities are the same for FlatCSL as for FlatPCTL, except that the parameter of the modalities is an interval I of the non-negative reals, rather than an integer upper bound. For example, the path formula $\phi_1\mathtt{U}^I\phi_2$ holds if ϕ_2 is satisfied at some time instant in the interval I and always earlier ϕ_1 holds.

$\mathcal{S}^G$ supports an extension of FlatCSL allowing specifications over reward structures in a manner similar to FlatPCTL, the only difference are the time

bounds: $\mathtt{R}_{\sim r}[\mathtt{C}^{\leq t}] \mid \mathtt{R}_{\sim r}[\mathtt{I}^{=t}] \mid \mathtt{R}_{\sim r}[\mathtt{F}\phi]$ where $\sim \in \{<, \leq, \geq, >\}$, $r, t \in \mathbb{R}_{\geq 0}$, and ϕ is a FlatCSL formula.

The extension of CTMC with rewards is analogous to that of DTMC, given above, barring the mentioned differences in time bounds. The formal semantics over rCTMC can be found in [8], see though that $\mathcal{S}^G$does not support therein mentioned steady state.

3. Implementation of $\mathcal{S}^G$

In a case of Markov models implemented in the `Prism` language, a generation of random simulation paths is not computationally expensive. A single transition consists of an evaluation of its guards, an enumeration of updates if viable, a random selection of subsequent transitions and finally an estimation of properties. The syntax of guards and updates allows simple arithmetical and logical operations and a few basic mathematical functions, such as a power or a logarithm. `Prism` comes already with the mentioned CPU–based simulator $\mathcal{S}^C$, well–suited to debugging tasks but slow, as it is sequential and generates each path by reinterpreting a model specification.

Instead of such an on–the–fly reinterpretation, the tool Ymer [15] compiles expressions to a form which is faster to evaluate repeatedly. Another tool APMC [7], in, turn provides a translation of `Prism` models to C programs which are later compiled and executed.

Another obstacle preventing the $\mathcal{S}^C$ from achieving a reasonable performance is its inherent sequentiality. A Monte Carlo simulation is considered to be embarrassingly parallel – it samples the model by generating a large number of independent random paths. `Prism` has approached this problem by providing the ability to perform distributed sampling, Ymer and APMC support distributed sampling as well. The latest version of Ymer implement multi–threaded sampling as well [16].

The improvement in parallel and distributed sampling is limited by the number of threads supported on multi–core CPU systems. A processor with a rich set of instructions and multiple cache levels is a perfect tool for complex and general problems but using it for a simple simulation of a moderate Markov model would be a very expensive over–engineering, both financial cost– and energy–wise. On the contrary, recent advances in GPGPUs made them a very efficient replacement for such computations, which benefit from massive parallelism. This simultaneous execution of hundreds and thousands lightweight threads on a GPGPU comes at a price: well-known limitations include a restriction of a group of threads to execute the same instruction at the same time or a burdensome memory model with a high cost of non–regular memory access patterns. We believe though, that

those restrictions do not play a significant role in a Monte Carlo simulation of `Prism` models. For that purpose we have chosen OpenCL [13] as a framework and programming language for GPGPU simulation.

3.1. **Architecture.** Our decision to implement the simulator engine for `Prism` using OpenCL has been driven by its capability of supporting many types of devices offered by different vendors, the most common being graphic cards and multi–core CPU servers. We will further use an OpenCL–specific terminology.

To simplify the implementation we have used OpenCL bindings for Java, the main source language of `Prism`, provided through the JavaCL library [2].

Fig. 1 presents a general scheme of the simulator. Within `Prism` there is no significant difference between starting a simulation in using $\mathcal{S}^C$ or $\mathcal{S}^G$. In both cases a model and a list of properties is required. Then, a just–in–time source-to-source translation of the model and properties produces a dedicated *compute kernel* (block `Kernel generator` in Fig. 1) which is basically a program for a number of vector processors, which use a single–instruction–multiple–data paradigm. The approach is very different from the reinterpretation scheme implemented in $\mathcal{S}^C$, which stores the model and properties in memory structures, and not as an automatically generated program. $\mathcal{S}^G$ executes kernels implemented in OpenCL C language, a modified version of C99, which has been adapted to OpenCL's device abstraction and stripped from features usually not allowed on a device, such as recursion or function pointers. Those simple programs can be effectively compiled into native bytecode of the device (typically, a graphic card). The syntax of `Prism` language makes the translation process rather straightforward due to the similarity between its expressions and the OpenCL C language. Only minor and automatically applied changes allow to obtain an expression valid in the latter language.

3.2. **Method.** A scheme of generating a single random path (sample) is presented in Algorithm 1. Capitalised identifiers indicate constants which are injected into kernel source. Many details have been omitted in the case of continuous time models. For example, in CTMC both times of entering and leaving state may be required in certain situations, such as a bounded until or a property with cumulative reward. The first argument `idx` is an unique identifier of kernel instance, `offset` specifies how many paths have been processed in kernels previously computed on the device. The argument `seed` seeds the generator of pseudorandom numbers. As the scheme is an equivalent of an OpenCL kernel, the two last arguments are OpenCL storage buffers in device memory, where the kernel is allowed to save results of property verification, and also the length of created path used

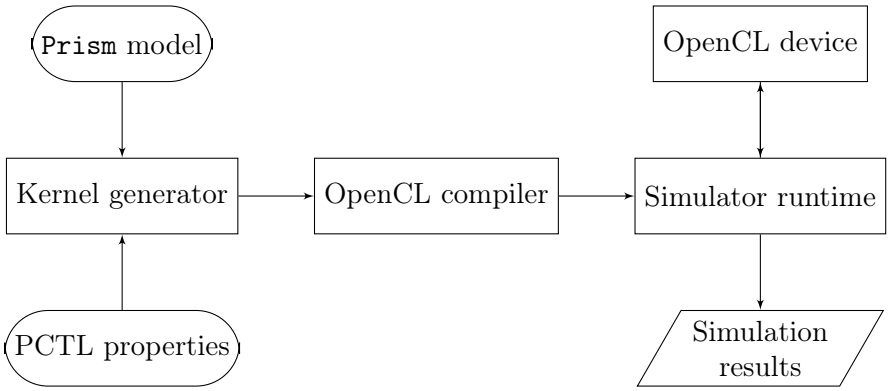


FIGURE 1. The OpenCL-based simulator engine in `Prism`.

to display sampling statistics. The first loop (lines 5–7) resets the collected statistical data about properties, the second loop (lines 8–27) generates the path until any of the following: its maximum length is reached (line 8), a deadlock (line 14) or a self–loop (line 18) occurs, or all properties are verified precisely enough (line 24). The last loop (lines 29–31) copies the collected data into the global memory.

The implementation had to be specially adapted for GPGPU devices, given that there can be a significant slow–down if the kernel diverges from the SIMD paradigm. Another example of such change is choosing always the smallest, most space efficient integer type for holding a state variable, which is possible as `Prism` models specify ranges for each such variable. For efficiency, if a simulated model updates a variable with a value exceeding these ranges, then the behaviour is undefined. A large speed–up can be achieved by handling an update synchronised between `Prism` modules in one step. If such update is performed on the same copy of state vector, it may induce a race condition of the type Read-After-Write. $\mathcal{S}^G$ detects such situations and creates additional copies of the affected variables, if necessary, instead of relying on the OpenCL compiler, which might handle the issue less effectively. Creating only one instance of a state vector is crucial for performance because it decreases significantly memory space used by the kernel, as discussed later with memory complexity of the algorithm. The simulation kernel is also capable of detecting when there is only one transition available, and it does not change the state. Such a behaviour indicates that there is only a single self–loop in the current state, which is interpreted by $\mathcal{S}^G$ as a stop condition. If there is an unbounded property which has not been satisfied yet, its value is not going to change. If there is

a property with a lower bound, which has not been reached yet, it can be evaluated immediately and the process of simulating the current path ends.

Our pseudorandom number generator of choice is Random123 [12], which provides a performance satisfying our needs. For more details on model conversion into a form for GPGPU computation see [3].

Algorithm 1 A generic path generation algorithm for OpenCL kernel.

```
1: procedure KERNEL(idx, offset, seed, path_lengths, property_results)
2:     prng ← initialize_prng(seed, idx, offset) ▷ A distinct seed for each path
3:     state_vector ← INITIAL_STATE_VECTOR
4:     time ← 0
5:     for each property p ∈ PROPERTIES do
6:         reset(results_p)
7:     end for
8:     for i < MAX_PATH_LENGTH do
9:         active_updates ← evaluate_guards(state_vector) ▷ Single and synchronised
10:        time_update(time)                          ▷ More complex for CTMC
11:        for each property p ∈ PROPERTIES do
12:            active_properties ← property_p_update(state_vector, results_p, time)
13:        end for
14:        if active_updates = 0 then
15:            break                                  ▷ Deadlock detected
16:        end if
17:        no_change ← update(prng, state_vector, active_updates)
18:        if no_change ∧ active_updates = 1 then
19:            for each property p ∈ PROPERTIES do
20:                active_properties ← property_p_update(state_vector, results_p, time)
21:            end for
22:            break                                  ▷ Loop detected
23:        end if
24:        if ¬active_properties then
25:            break                                  ▷ Stop sampling
26:        end if
27:    end for
28:    path_lengths[idx + offset] = i   ▷ Save results in global memory
29:    for each property ∈ property_results do
30:        property[idx + offset] = results_p   ▷ Save results in global memory
31:    end for
32: end procedure
```

A generated kernel is passed as a string of source code to a specific device compiler (block `OpenCL compiler` in Fig. 1). This compilation does not add

a significant overhead on modern OpenCL platforms – we have found that it typically takes less than one second for tested models. A kernel represented in device bytecode is sent to simulator's runtime (see again Fig. 1), where a range of *work items* is enqueued on the device, as described in more details in the next section. Each one of them is responsible for producing exactly one random path through the model and its identifier *idx*, is paired with an *offset*, in order to produce a unique key across many separate kernel enqueued on the device. The key is necessary for correct accessing memory storage and unique random seeds for each generated path.

4. Automatic statistical verification

The ASV step is optionally triggered at the user request, in one of the following ways:

- unconditionally;
- if a quantitative property is being computed, like the probability value;
- if a steady state is detected prematurely in an iterative PMC method.

The last criterion is discussed in more detail in the example in Section 5.

The ability to process an extensive number of samples in a very short time often allows for a reliable and fast ASV of a property value v obtained using PMC. In the ASV step, in order to save the user's time, $\mathcal{S}^G$ must finish within a predefined time $T_{\max}$.

After the ASV step, the user is presented with a set of diagnostics, so that he can estimate the correctness of v. Let the simulator estimate a value w of the same property. Let R_{CI} be the ratio of the width of CI at confidence level 90% to w. The following independent diagnostics can be presented:

- $T_{\max}$ too low to reach $R_{CI} < R^{CI}_{\max}$; $R^{CI}_{\max}$ has a default value of $1 \cdot 10^{-2}$ and can be customised;
- v within a CI of a confidence level x, outside a CI of a confidence level y, where $x \in L = \{90\%, 95\%, 99\%\}$, $y \in L \vee \{< 90\%\}$.

5. Case study

An instantaneous reward in a CTMC can be estimated by weighting the reward function over a probabilistic distribution of states at a time instant t. `Prism` computes the distribution by uniformisation (Jensen's method) [5] which discretises the CTMC with respect to a constant rate. Then, probabilities are approximated by a finite summation Z of Poisson–distributed steps of the derived DTMC. The number of these steps depends on the precision required, and is computed using the Fox–Glynn method [4]. Yet,

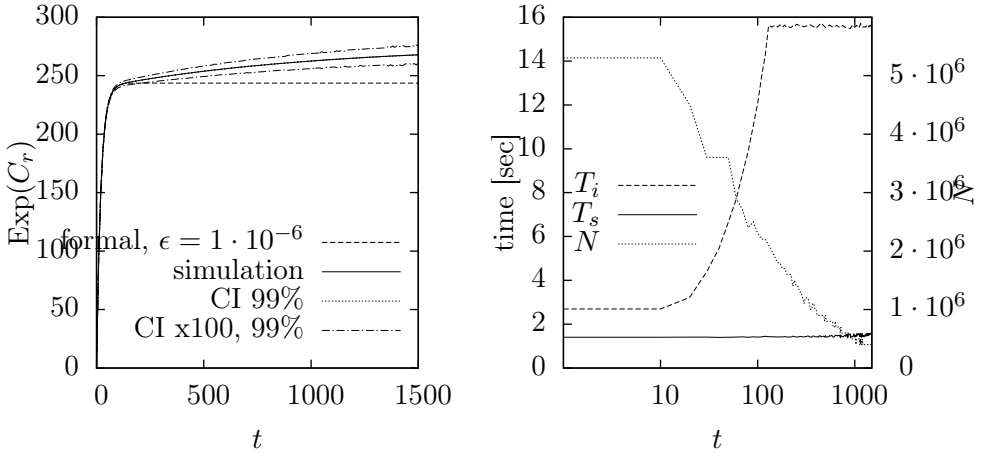


FIGURE 2. (a) Estimates of $\mathrm{Exp}(C_r)$, CIs are narrow enough to be hardly seen, thus their magnification by 100 is also show. (b) elapsed CPU time and the number of samples per single property, T_i and T_s are total times of, respectively, the formal method performing Z, and the statistical estimation on $\mathcal{S}^G$, N is the number of samples chosen by $\mathcal{S}^G$ in order to fit within $T_{\max}$.

in order to shorten computation time, when performing that summation, `Prism` also tests, if a steady state has been reached, by finding a maximum difference, either relative or absolute, between elements in solution vectors from two successive steps. If the difference is smaller than a constant threshold ϵ, the summation is terminated early.

`Prism`'s default criterion of the termination is to use a relative difference and $\epsilon = 10^{-6}$. The criterion can be customised in order to set a compromise between precision and computational complexity.

To illustrate the ASV, we will discuss a model where the detection of a steady state is premature if the default termination criterion is used. In effect, were the automatic SV not enabled, the user would obtain incorrect results without a warning.

We will use a modified Model 2 from [11]. It is a simple CTMC with multiple clients and servers, in which some of the servers can occasionally be broken. To trigger the said premature termination of Z, we modify the model by using constants $r_s = 1, r_l = 500$ (see [11] for details), i.e. the server is slower at initiating a connection with a client, but faster at processing a request. We ask for an expected instantaneous reward value C_r, equal to the number of clients requesting at a given time instant t.

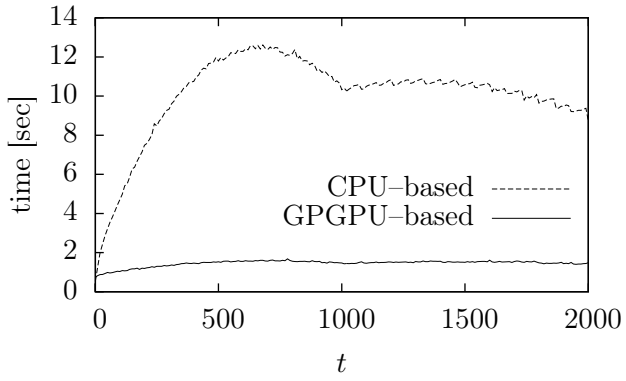


FIGURE 3. Computation time on two OpenCL devices.

Let the user choose the default termination criterion, and let he also request, that SV be used for up to $T_{\max} = 1.5$ sec. after a PMC step such that a steady state detection has terminated early Z (or any other formal iterative method which uses ϵ). $\mathrm{Exp}(C_r)$ against t, found using both a formal PMC and $\mathcal{S}^G$, is depicted in Fig. 2(a). We see that the model undergoes a fast change at $t \approx 100$, during which the increase of requests becomes rapidly slower. The constant $T_{\max}$ allows for a fairly narrow CI at 99% confidence level – its width never reaches 0.1. In the case of the discussed diagram $R_{CI} < 4 \cdot 10^{-4}$ for any t – such a narrow CI makes it probable that following the said change at $t \approx 100$, the PMC results become less and less precise. This would trig respective warnings, that values from the results of the PMC step fall outside a CI of a high confidence level.

The relative temporal overhead of SV for the chosen $T_{\max}$ is illustrated in Fig. 2(b). It is largest for small t and makes `Prism` run for about 50% longer. For $t \gtrsim 200$ `Prism` needs less than 10% of an additional CPU time to perform the SV step. The figure also shows the number of samples which can be computed within $T_{\max}$; wee see that the number decreases for larger t due to longer paths which need to be generated by the simulator.

In order to obtain the CPU times in Fig. 2(b), we have used an AMD R9 Nano, a modern mid-range GPU with 4096 stream processors. Let us compare that GPU to another OpenCL device, containing two Intel Xeon E5-2630 v3 CPUs with clock frequency 2.40GHz, each of them providing 8 cores with HyperThreading, resulting in 32 threads for OpenCL. Fig. 3 compares the simulation time from Fig. 2(b) to that of the CPU OpenCL device, both evaluating the same number of samples. In the case of the latter device, we see times in the range of 8 and 12 seconds, i.e. it is several

times slower. This shows the advantage of streams processors in the case of a large number of independent, non–memory intensive tasks.

$\mathcal{S}^C$ has been excluded from the chart as it is not optimised for speed. At $t = 2000$, where long paths make it especially advantageous to use the on–the–fly compilation, it took $\mathcal{S}^C$ over 130 minutes to evaluate the same reward property on the same CPU, thousands of times slower comparing a respective simulation on a GPGPU.

6. Discussion

The SV limits itself now to diagnostics, but it could be straightforwardly extended to influence on the PMC step. For example, if a property p_i turns out to be computed imprecisely in the PMC step, and the following property to compute p_{i+1} differs only in the time instant t, the SV could automatically decrease ϵ for the computation of p_{i+1}.

We expect to release an open-source version of the tool in the following months. The further development would be focused on a parallelisation across multiple OpenCL devices, with a dynamic and automatic load balancing.

References

[1] Aziz, A., Sanwal, K., Singhal, V., Brayton, R.: Model-checking continuous-time markov chains. Transactions on Computational Logic (TOCL) **1**(1), 162–170 (2000). DOI http://dx.doi.org/10.1145/343369.343402.

[2] Chafik, O.: Javacl. https://github.com/nativelibs4java/JavaCL (2016)

[3] Copik, M.: GPU-accelerated stochastic simulator engine for PRISM model checker. Bachelor's Thesis, Silesian University of Technology, Gliwice, Poland (2014)

[4] Fox, B.L., Glynn, P.W.: Computing poisson probabilities. Commun. ACM **31**(4), 440–445 (1988). DOI http://dx.doi.org/10.1145/42404.42409

[5] Gross, D., Miller, D.: The randomization technique as a modeling tool and solution procedure for transient Markov processes. Operations Research **32**(2), 343–361 (1984). DOI http://dx.doi.org/10.1287/opre.32.2.343

[6] Hansson, H., Jonsson, B.: A logic for reasoning about time and reliability. Formal aspects of computing **6**(5), 512–535 (1994) DOI http://dx.doi.org/10.1007/BF01211866

[7] Herault, T., Lassaigne, R., Peyronnet, S.: Apmc 3.0: Approximate verification of discrete and continuous time markov chains. In: Proceedings of the 3rd International Conference on the Quantitative Evaluation of Systems, QEST '06, pp. 129–130. IEEE Computer Society, (2006). DOI http://dx.doi.org/10.1109/QEST.2006.5

[8] Kwiatkowska, M., Norman, G., Parker, D.: Stochastic model checking. In: Formal Methods for the Design of Computer, Communication and Software Systems: Performance Evaluation (SFM'07), *LNCS*, vol. 4486, pp. 220–270. Springer (2007)

[9] Kwiatkowska, M., Norman, G., Parker, D.: PRISM 4.0: Verification of probabilistic real-time systems. In: G. Gopalakrishnan, S. Qadeer (eds.) Proc. 23rd International Conference on Computer Aided Verification (CAV'11), *LNCS*, vol. 6806, pp. 585–591. Springer (2011). DOI http://dx.doi.org/10.1007/978-3-642-22110-1_47

[10] Legay, A., Delahaye, B., Bensalem, S.: Statistical model checking: An overview. In: International Conference on Runtime Verification, pp. 122–135. Springer (2010). DOI http://dx.doi.org/10.1007/978-3-642-16612-9_11

[11] Pourranjbar, A., Hillston, J., Bortolussi, L.: Don't just go with the flow: Cautionary tales of fluid flow approximation. In: Computer Performance Engineering - 9th European Workshop, pp. 156–171 (2012). DOI http://dx.doi.org/10.1007/978-3-642-36781-6_11

[12] Salmon, J.K., Moraes, M.A., Dror, R.O., Shaw, D.E.: Parallel random numbers: As easy as 1, 2, 3. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, SC '11, pp. 16:1–16:12. ACM, (2011). DOI http://dx.doi.org/10.1145/2063384.2063405.

[13] Stone, J.E., Gohara, D., Shi, G.: OpenCL: A parallel programming standard for heterogeneous computing systems. Computing in science & engineering **12**(1-3), 66–73 (2010). DOI http://dx.doi.org/10.1109/MCSE.2010.69

[14] Younes, H., Kwiatkowska, M., Norman, G., Parker, D.: Numerical vs. statistical probabilistic model checking. International Journal on Software Tools for Technology Transfer (STTT) **8**(3), 216–228 (2006). DOI http://dx.doi.org/10.1007/s10009-005-0187-8

[15] Younes, H.L.S.: Ymer: A statistical model checker. In: Proceedings of the 17th International Conference on Computer Aided Verification, CAV'05, pp. 429–433. Springer-Verlag, Berlin, Heidelberg (2005). DOI http://dx.doi.org/10.1007/11513988_43

[16] Younes, H.L.S.: Ymer. https://github.com/hlsyounes/ymer (2016)

Received: October 2017

Marcin Copik
RWTH Aachen University,
Schinkelstr. 2, 52062 Aachen, Germany
E-mail address: mcopik@gmail.com

Artur Rataj
Polish Academy of Sciences
Institute of Theoretical and Applied Informatics
ul. Bałtycka 5, 44-100 Gliwice, Poland
E-mail address: arturrataj@gmail.com

Bożena Woźna-Szcześniak
Jan Długosz University in Częstochowa
Institute of Mathematics and Computer Science
Al. Armii Krajowej 13/15, 42-200 Częstochowa, Poland
E-mail address: b.wozna@ajd.czest.pl

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
99–116
DOI http://dx.doi.org/10.16926/m.2017.22.08

THE CONCEPT OF IMPLEMENTATION OF THE DECENTRALIZED APPLICATION LEVEL STRUCTURE FOR PROVIDERS OF A TYPICAL NEXT GENERATION NETWORK

N. KNIAZIEVA, S. SHESTOPALOV, W. SUSŁOW, A. YATSKO

Abstract

Based on the IPCC concept, the architecture of the Next Generation Network with the Decentralized Structure Applications Level to improve the quality control of the intellectual services (ISes) provisioning is proposed. It is recommended to evaluate the effectiveness of the control of provisioning ISes by means of the complex criterion, which takes into account the interests of three participants of ISes: process service providers, equipment suppliers and network users. The subcriteria are defined for each participant. The method used for the calculation of each subcriterion has been proposed as well. The problem of the association of subcriteria into the resulting complex criterion has been solved.

1. Introduction

Analysis of the Problem Domain

Nowadays, some Ukrainian operators offer (apart from basic services) a set of intellectual services (ISes) - services that can provide the level of applications of the Next Generation Network (NGN).

A Next Generation Network (NGN) is a packet-based network which is capable of providing telecommunication services, making use of multiple broadband, QoS-enabled transport technologies. It is worth noting that in

• *Nina Kniazieva* — e-mail: knyazeva@ukr.net
Odessa National Academy of Food Technologies, Ukraine.
• *Serhii Shestopalov* — e-mail: shestopalov.s@mail.ru
Odessa National Academy of Food Technologies, Ukraine.
• *Walery Susłow* — e-mail: walery.suslow@tu.koszalin.pl
Koszalin University of Technology, Poland.
• *Andrzej Yatsko* — e-mail: andrzej.jacko@gmail.com
Koszalin University of Technology, Poland.

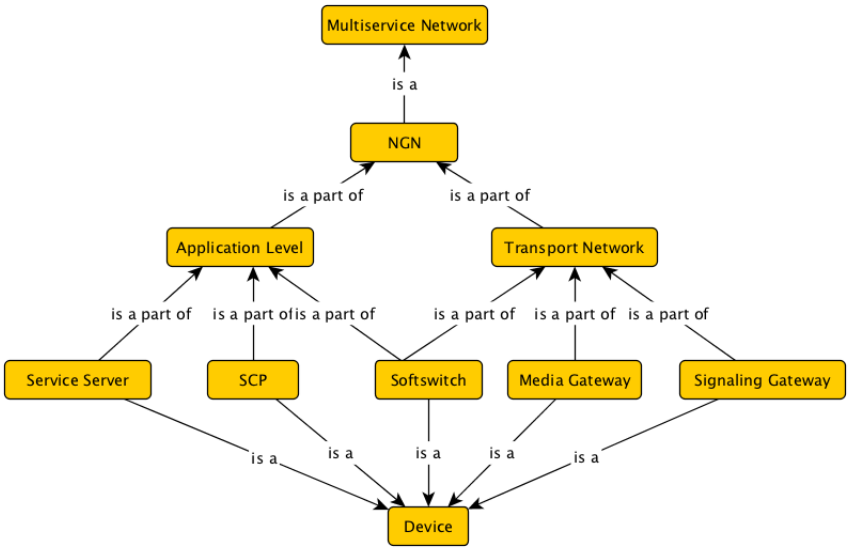


FIGURE 1. The concept map of the NGN domain

NGN the service-related functions are independent from underlying transport related technologies. It enables unfettered user access to subnetworks and to competing service providers or services. It supports mobility of users and operators and allows competitive provision of new services to users [1, 2]. The concept map [3] of the NGN domain is shown in Figure 1.

In Ukraine, as well as in many Eastern European countries, a complete NGN has not been created yet. As it stands, Ukraine does not even have full 3G coverage (see map in Figure 2), while European Union countries typically use LTE communication. However, some operators create networks that can provide Triple Play Service (speech, video, data). The transition to the next generation is slowly taking place.

Nowadays in the countries of Eastern Europe, telecom operators can choose between IMS/TISPAN or IPCC architecture when designing NGN. Most of such networks operated around the world are built in accordance to the IPCC concept. Selection of second architecture is supported by the fact that a large number of new technologies are being supported by IPCC. This allows the operator to make a dedicated network configuration and to expand network which will optimally interact with the already installed network equipment.

Additionally, under the conditions which now exist in Ukraine and many countries of Eastern Europe, the IPCC architecture allows to make a smooth

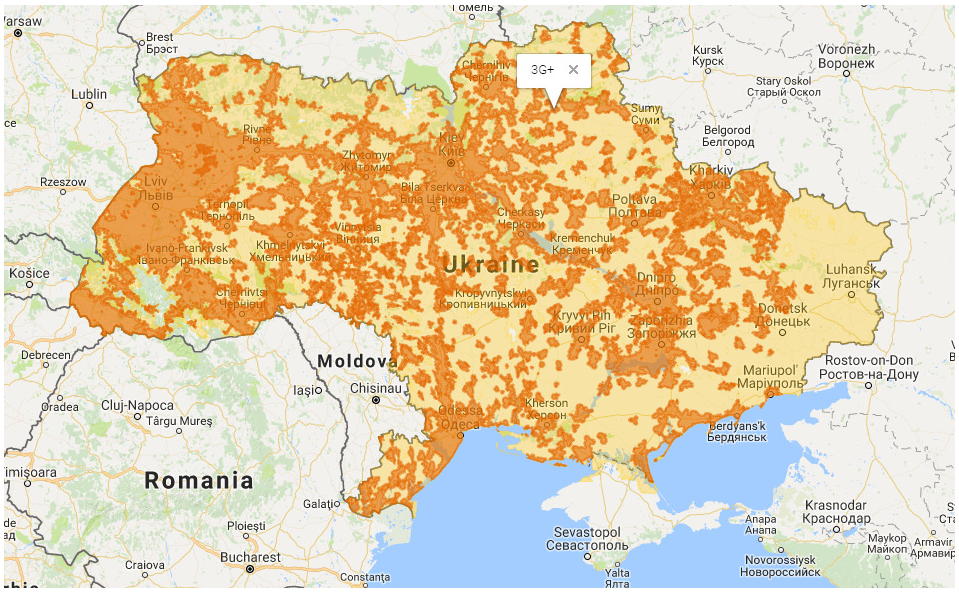


FIGURE 2. The 3G coverage of one of the leading mobile network operators in Ukraine

transition from the public telephone network (PSTN) to NGN. Such a business plan requires a smaller initial investment, while facilitating the gradual switch to modern equipment, while usually being more attractive and acceptable for the owners of the telecommunication network.

The biggest known drawback of the IPCC concept is the problem of compatibility between different equipment. IMS/TISPAN architectures simplify the problems that arise when using the concept of IPCC. In addition, it facilitates the convergence of mobile and fixed communications. Of course, from the point of view of the global network, the TISPAN architecture is an option as well, but not in the form of a simple controller connected to another controller, as the IPCC is regarded to be.

However, the possibilities of NGN with TISPAN architecture should be clarified in the future. But NGN with the IPCC architecture is now quite usual and familiar thing, which is used in many countries. So, a gradual transition to NGN using the already installed equipment is considered to be the optimal solution for Eastern European countries. Therefore, in the near future the NGN will be developed there most likely as a network using the architecture proposed by IPCC.

Intellectual services (ISes), which are granted by the application level of NGN, include Televoting (VOT), Virtual private networks (VPN), Abbreviated dialing (ABD) and many others. The variety of IS is growing day by day and the demand for them is growing as well. The Applications Level with Centralized Structure (ALCS) is usually used to manage the IS in NGN. In this context, one of the major issues regarding Eastern European telecom is the development of the implementation principles of the application level for the NGN to improvements of the provision control quality of ISes.

The review of NGN architecture is discussed in the works by B.S. Goldstein, O.B. Goldstein [4], M.O. Sokolov, A.A. Attsyk, A.V. Pinchuk, Y.S. Kryukov, A. Titov, M. Hlinnykova, O.B. Antonyan, E.M. Skuratovsky, I.G. Baklanov, V.V. Makarova. The NGN-architecture developed by them has been implemented by telecom leaders such as Alcatel, Ericsson, Lucent Technologies, Siemens, and Russian manufacturer - STC "Proteus".

The network service control issues, and of the evaluation of efficiency of the control systems, and also of the design of criteria of control quality are considered in the works of V.K. Steklov [5], V.G. Kryvutsa, L.N. Berkman, N.A. Knyazeva, E.V. Kilchitskaya, V.V. Makarova, B.J. Kostika, E. Steinberg, A.N. Starodub, N.Y. Parshenkov. The quality of service supported by NGN, was discussed by such authors as P. Ferguson, R. Haston.

2. Main results

The purpose and statement of the problem

The quality of NGN services significantly depends on the structure of the application level of the network. Based on the existing structures of the application level and also with the intension to improve the provision control quality of ISes, the authors propose the new way of organizing the application level and suggest comprehensive quality control criteria for NGN providers.

Variants of the structure of the NGN application level

Currently, there are essentially two distinguished concepts of building the application level: with a centralized structure or with a distributed structure.

The application level with a centralized structure

Modern NGNs normally use the Applications Level with the Centralized Structure (ALCS). Details of the network with such architecture are shown in Figure 3.

The ALCS modules responsible for control of ISes consist of Softswitches, which provide an SSF switching service, and also a Server, which provides

the function maintenance service of SCF. It is assumed that there are several geographically separated areas. Network Control of the district of the area is performed by the installed Softswitch. Each Softswitch manages the transport network while being the switching point of IS. Each district has its own data network and its own signaling network.

The schematic in Figure 3 highlights the signaling network. Each district network must be connected to the single Softswitch. This schematic shows details for just one of districts. If the regional network uses not only packet-switched technology, but also switching channels, the Signaling Gateway (SG) in the signaling network and Media Gateway (MG) in the data network should be set.

SG - the device which mediates the interaction between packet transport networks and access networks at the signal level. Its function is to convert formats of the signaling messages. MG provides communication between packet-switched network and circuit-switched network, performing protocol and data format conversion. It can perform information flow processing and, in some cases, it provides support of the subscriber alarm system. MG is divided into transport gateways, which serve as the transition nodes between the PSTN and packet networks, and access gateways serving to connect users.

Intellectual services are given as follows. The requests for the IS comes from the user. If the user is in the IP network, packets are sent directly to the relevant Softswitch. If the user sent requests from outside the IP network, then the packets will be sent only after transformations by MG and SG. The Softswitch determines whether the ordered service is an intellectual service. If it is true, the switching to IS occurs and a request to the dedicated server is sent (this request is transmitted via the signaling network). If the server is not busy, it starts to serve the request. In the other case, it is checked whether there is a place in the buffer queue. If there is no free place, the application request is lost. Otherwise, the application is put into the queue and waits for service.

In [6] the drawbacks of NGN with ALCS are analyzed in detail:

1. There may be a problem regarding limited throughput of the signaling network and productivity of IS control centers. If the number of users using the IS grows, so does the number of requests for the IS and the resultant load per signaling network and control center for sequent providing the IS. This can degrade network productivity, which will either become unacceptable to the users or make it impossible to support the principle of centralized control.

2. Certain types of IS, within their properties, are not designed for centralized performance (for example, IS which forbid certain received calls

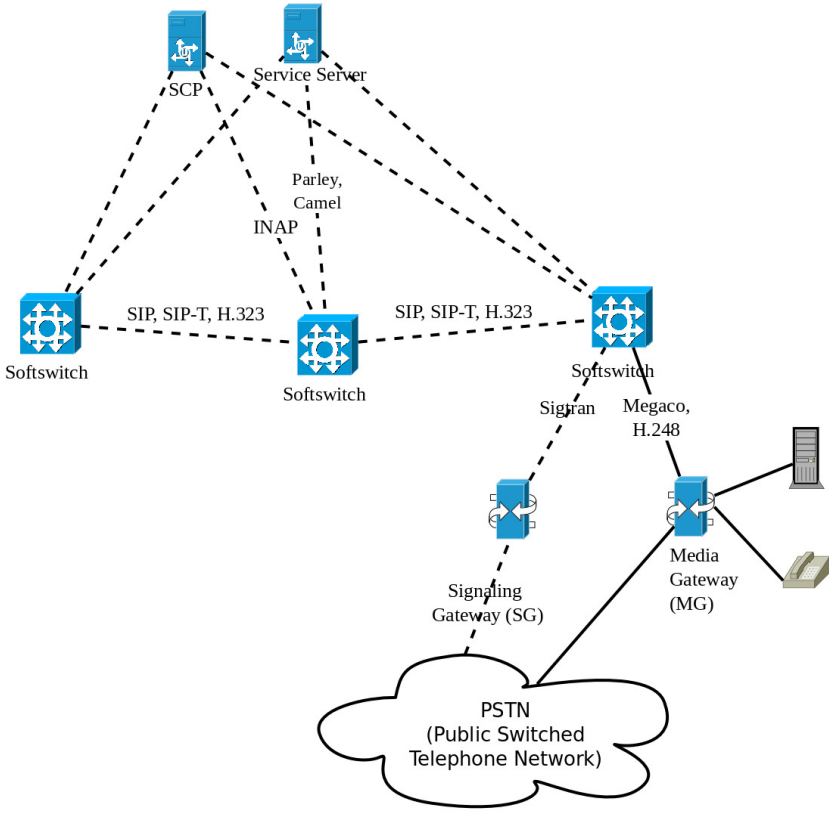


FIGURE 3. The NGN application level built according to the principle of centralized control

according to certain criteria). The criteria may vary depending on network conditions. The throughput limit of the signaling systems and an individual character of the criteria that have to sort calls lead to the need to determine the possibility of providing decentralized management of IS.

3. There exist intellectual services which do not admit performance degradation. Obviously, for the control of such IS it is necessary to establish the possibility to use the decentralized control. Decentralization, most of all, will reduce the delay of IS.

The application level with distributed structure

To manage the ISes in the Intelligent Network (IN), which was the forerunner of NGN and had great influence on it, the standard Applications Level with Distributed Structure (ALDtS) was offered. The architecture of the ALDtS is shown in Figure 4.

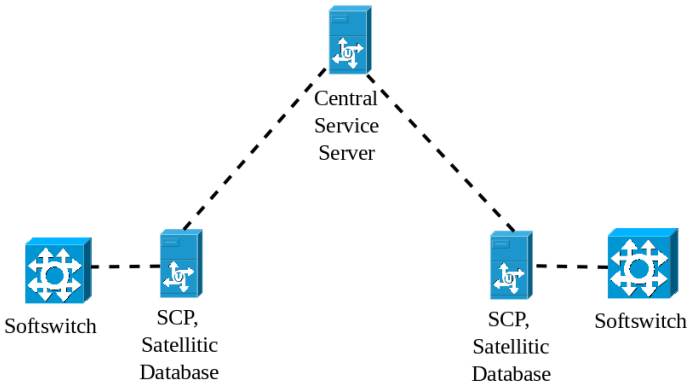


FIGURE 4. The NGN application level built according to the principle of distributed control

This architecture provides for the existence of the central server which contains all the necessary data and the logic of provisioning of ISes. At the same time, it provides for the existence of the Service Control Points (SCPs) with its own databases which are put into a switching points of service. They contain data and service logic for the most commonly used services. There are several districts with their own service control and own databases. Service Switching Point (SSP) and the Call Control Point (CCP) are physically located in a Softswitch. In the case of increasing the number of requests for some ISes, which are not in their own database, the data and the logic for the services can be downloaded from a central server. IS is given as follows. The request, similar to the ALCS case, goes to the Softswitch. The Softswitch asks the SCP and their own database. If the necessary data and service logic for that IS are found, then the request is served. In the other case, the request is sent to the central server for transmitting necessary information to the appropriate SCP. After receiving the data, the service is performed.

The architecture of a network with ALDtS can solve a part of the problems arising in the ALCS case. First of all, it allows reduction of the total time of service of IS requests. However, as it is illustrated in Figure 4, SCPs are not connected directly. Therefore, in the case of failure of the central server this control system is not able to function normally. In this case, at the request of the IS, which is not in its own database, the SCP cannot handle the request, regardless of normal functioning.

To solve the aforementioned problems, the authors proposed an alternative structure, called the Applications Level with Decentralized Structure

(ALDS), which is based on the concept of IPCC. The architecture of the network with the ALDS provides several service switching units (Softswitches) and several service control units (servers). Calculation of the number of servers and Softswitches is a separate problem that is not discussed in this paper.

The authors have proposed two approaches to the realization of ALDS. Using both approaches, first of all, the IS have to be classified. The classification may be performed, for example, according to the IS technologies. Some ISes, however, do not allow the delay of performance, those should be united under a separate class [7-9].

After that, by the first approach, each server has to contain the service logics for all classes of ISes (universal server). That is, each server repeats all the possibilities of the united center that operates in ALCS.

The second approach requires the usage of specialized servers. In such a case, the classes of ISes may be completed in accordance to the requests for their certain types on the corresponding territory. A certain set of classes of ISes and their service logic should be placed in a separate ISes control node (server). For each set of classes there should be its own node. It is necessary to predict the possibility to provide one set of the classes of ISes by several servers in the case of failure of any server. This variant seems to be the optimal. Believing that servers are specialized, the authors propose to consider the NGN with the ALDS.

While NGN architecture with the ALCS is investigated and described in detail, without paying attention to the lack of clear standards, problems of NGN architecture with the ALDS seem not to have been sufficiently investigated. In scientific works one can find only some variants of decentralized control of simplified systems. However, for NGN it has been believed until now that using the ALCS is the best solution.

The network architecture proposed by the authors of NGN with ALDS based on the concept IPCC is shown in Figure 5. As shown in Figure, the existence of multiple geographically separated areas is allowed. Each of them has its own Softswitch which controls district transport network and performs Switching Service Functions (SSF). Each district has its own data network and signaling network as well. In Figure 5, greater emphasis is put on the signaling system because of the subject of the article. A district network must be connected to a separate Softswitch. For simplicity, such a connection is shown in the Figure only for one of Softswitches.

If the district network uses both packet and channel switching technologies, it is necessary to establish the SG in the signaling network and the MG in the data transmission network.

While using the ALDS, the server performing the logic of the IS complex has to be placed next to the Softswitch. Then it can be assumed that the intellectual service is carried out practically at the point of switching service.

Figure 5 shows the distribution of NGN among the data transmission network (the transport platform) and the application level. At the lower layer of pseudo 3D scheme, the authors tried to show the data transmission network, at the higher layer - the application level.

The application level constitutes practically the separate logical network which is needed to exchange information and control messages between servers. This network may be carried out by the signaling network or by the data transmission network. However, as it is seen in Figure 5, the advantage is given by the authors to the signaling system. During an exchange of information between servers a protocol: MGCP, MEGACO, SIP can be used. The connection between the Softswitch and the server may be realized through an open API (Parlay, Camel).

In such architecture, the provision of IS is performed as follows. The request for IS comes from the user. If the user is in the IP-network, the packets are coming directly to a certain Softswitch. If the user is out of IP-network, the packets are coming only after transformations in MG and SG. If the service is not an IS, then it is provisioned by the Softswitch software. If it is the IS, then the Softswitch sends a request to a directly connected server to perform the request. If the server is not busy, it serves the request. Otherwise, a queue buffer check takes place. If it fails, the request will be lost. If there is a space, the request is placed into the queue.

Due to the fact that servers in different areas may contain different sets of classes of services and service logic, a situation where the server of the certain district will be not able to serve the request may occur. In that case, the server should redirect the request to the server of another district which will be able to perform the request. When selecting such a server, it's necessary to use the probability matrix of transitions of requests from the current server to another server for the corresponding class of IS. After examining the situation, the request will be sent from the current server to the selected server. If the selected server is not busy, the service will start. If it is busy, the request is placed into the queue. If there is no place in the server buffer, the request will be lost.

In the case that the server which is connected to the Softswitch breaks down, the Softswitch stops performing the IS switching function and sends the request to another Softswitch.

Comparing the logic of service provision in the ALCS and in the ALDS, one can assume that the architecture of NGN with ALDS will function

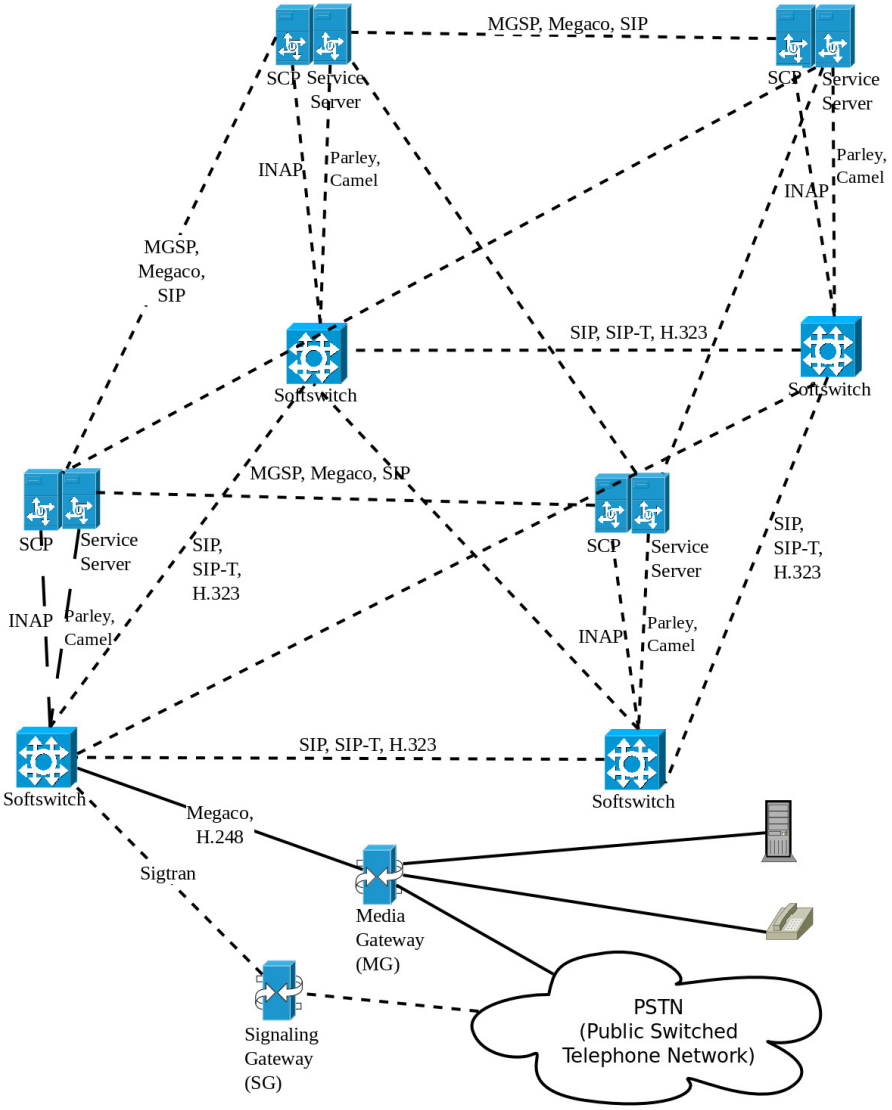


FIGURE 5. The NGN application level built according to the principle of decentralized control

better under significant network loads. It may happen for example, due the upsurge of the number of users and the number of IS applications at the success of the business campaign of telecom operator. However, the final conclusions can be made only after comparing the quality of IS provision control in the case of IALCS and ALDS. For a final resolution, it is necessary

to create a comprehensive criterion of quality that will take into account the interests of all participants of IS provision process - service providers, equipment suppliers, and users.

Evaluation of intelligent services provisioning quality

The authors propose a methodical approach to evaluating of the IS provision control quality.

This approach should be implemented in the following six step sequence:

1. Decision about the degree of influence of each participant of the IS process on the value of the complex criterion of quality of control.
2. Selection of subcriteria for each participant of the process.
3. Creation the weighted coefficients for subcriteria.
4. Assessing the value achieved by each subcriterion.
5. Calculating the resulting scores for each participant of the process.
6. Calculating the complex criteria of the quality.

Let us consider the implementation of steps of the proposed methodical approach to assess the quality of control the IS provision from the perspective of all participants of the process.

Step 1. Decision about the degree of influence of each participant of the IS process on the value of the complex criterion of quality of control.

First of all, it is necessary to take into account the degree of importance of each participant of the process, more precisely the degree of its impact on the complex control quality criterion of the intellectual services provision. This is possible using the method of expert evaluations. Each ith participant of the process has to be assigned "weight" W_i in the accepted assessment system. If the experts conclude that the impact of some of the participants on the result is equal, they can set to them equal "weights" W_i.

Step 2. Selection of subcriteria for each participant of the process

For each participant of the process of forming the complex control quality criterion of the IS provision (for service provider, equipment supplier, and network user) the indexes have to be determined, which should be taken into account forming the complex criterion (hereinafter let us call them subcriteria).

From the user's viewpoint, the quality of control of provision of intellectual services can be most accurately evaluated using the value of users' satisfaction with the quality of received IS, which is formed by combination of different elements: user's equipment $\overline{E_u}$, service transport $\overline{T_s}$, service provision $\overline{P_s}$, and content creation $\overline{C_c}$ [10].

From the service provider's viewpoint, the quality of control may be evaluated by the complexity of the service logic implementation and administration $\overline{S_t}$ and by the quantity of successfully provided $\overline{S}$ that will raise the demand for them. It should be noted that the complexity of the

service logic implementation and administration $\overline{S_t}$ is the subcriterion, the achieved level of which for ALCS and ALDS have to be established by the method of expert evaluations.

From the equipment supplier's viewpoint, the quality of control may be evaluated by indexes determined by the ITU-T Recommendations Y.1541 [10], namely: by time of service provision $\overline{T_{cs}}$ and by the probability of refuse of service provision $\overline{P_B}$. In addition to the specified subcriteria, in accordance with the ITU Recommendations for future networks [1, 12], the following quality indicators are proposed as subcriteria: structural survivability $\overline{P_{st}}$ and reliability of control systems $\overline{R}$. Additionally, it is proposed to take into account the cost of the intelligent superstructure $\overline{C}$.

Step 3. Creating the weighting coefficients for subcriteria

To take into account the degree of importance of each participants' subcriterion, the weighting coefficients K_{ij} have to be determined using expert evaluations. Here, j is a number of subcriterion of the ith participant of the process (see Table 1, column 2). The values of obtained weighting coefficients are normalized for each ith participant of the process, i.e. $\sum_{j=1}^{n_i} K_{ij} = 1$ (n_i is the number of subcriteria of the ith participant of the process, the values n_i can be different for each ith participant).

Step 4. Assessing the value achieved by each subcriterion

For each subcriterion of each participant of the process the acceptable maximum or minimum (depending on the subcriterion type) and current values (in appropriate units or in scores) are determined.

For each subcriterion the relative scores O_{ij} are found. Relative score may be formed as a product of the ratio of the current indicator value and its maximum acceptable value (or based on the ratio of minimum acceptable value and the current indicator value) and a weighting factor K_{ij} of the subcriterion determined in step 3. The method of calculation of each relative score is shown in Table 1 (column 5 is for ALCS and column 8 is for ALDS). Column 3 includes maximal or minimal acceptable values of the quality subcriteria for further definition of relative scores O_{ij}. Column 4 contains the current values of the quality subcriteria. Columns 6 and 7 are filled similarly for ALDS.

Step 5. Calculating the resulting scores for each participant of the process

At this stage, the resulting value is determined: it is an overall weighted score A_i of all the subcriteria for each ith participant of the process:

$$A_i = \sum_{j=1}^{n_i} O_{ij} \tag{1}$$

where O_{ij} is the relative score obtained for the jth subcriterion of the ith participant of the process and determined in step 4; $i = \overline{1, m}$ and m is

the number of participants of the process; $j = \overline{1, n_i}$, n_i is the number of subcriteria for the ith participant of the process.

Step 6. Calculating the complex criterion of the quality

At this stage, the value of the resulting complex quality criterion for the application level with centralized and decentralized control architecture is calculated for each ith participant of the process. Having the values of the resulting score A_i (see formula (1)), the complex quality criterion K for ALCS and ALDS is calculated considering the interest of all the participants of the process:

$$K = \sum_{i=1}^{m} A_i W_i \tag{2}$$

The obtained value of the complex quality criterion K (formula (2)) for the ALCS and ALDS leads to the conclusion about feasibility of their application for control of provision of ISes.

Table 1 summarizes the method of calculating the relative valuations O_{ij} as well as the total weighted assessment of all subcriteria for each ith participant of the A_i process with the following initial data:
- the number of participants of the process $m = 3$;
- the number of subcriteria for the first participant of the process $n_1 = 4$;
- the number of subcriteria for the second participant of the process $n_2 = 2$;
- the number of subcriteria for the third participant of the process $n_3 = 5$.

The complex quality criterion K is calculated based on formula (2).

3. Final remarks

Conclusions and prospects for further research

1. The Next Generation Network with ALDS architecture is based on the IPCC concept in order to improve the quality control of IS provisioning. Two approaches to realizing the ALDS have been proposed. Using the first approach, each server contains the logic of service of all classes of services (universal server). That is, each of the servers repeats all the capabilities of a separate server that functions at ALDS. The second approach predicts the use of specialized servers.

2. On the basis of recommendations of the ITU, according to analysis of modern scientific publications concerning the issue of quality control in the provided services, the researchers recommend to evaluate the effectiveness of the control of provisioning ISes by means of complex criterion which takes into account the interests of three participants of IS process – service providers, equipment suppliers and network users.

Quality subcriterion	Weight of the subcriterion	Application level structure which is analyzed					
		ALCS			ALDS		
		Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})	Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})
1	2	3	4	5	6	7	8
I.From the user's viewpoint							
1. Customer Equipment,$\overline{E}_u$	K_{11}	$\overline{E}_{u_{\mathrm{ALCS}}}(\max)$	$\overline{E}_{u_{\mathrm{ALCS}}}$	$\frac{K_{11}\overline{E}_{u_{\mathrm{ALCS}}}}{\overline{E}_{u_{\mathrm{ALCS}}}(\max)}$	$\overline{E}_{u_{\mathrm{ALDS}}}(\max)$	$\overline{E}_{u_{\mathrm{ALDS}}}$	$\frac{K_{11}\overline{E}_{u_{\mathrm{ALDS}}}}{\overline{E}_{u_{\mathrm{ALDS}}}(\max)}$
2. Service Transport, $\overline{T}_s$	K_{12}	$\overline{T}_{s_{\mathrm{ALCS}}}(\max)$	$\overline{T}_{s_{\mathrm{ALCS}}}$	$\frac{K_{12}\overline{T}_{s_{\mathrm{ALCS}}}}{\overline{T}_{s_{\mathrm{ALCS}}}(\max)}$	$\overline{T}_{s_{\mathrm{ALDS}}}(\max)$	$\overline{T}_{s_{\mathrm{ALDS}}}$	$\frac{K_{12}\overline{T}_{s_{\mathrm{ALDS}}}}{\overline{T}_{s_{\mathrm{ALDS}}}(\max)}$
3. Service Provision, $\overline{P}_s$	K_{13}	$\overline{P}_{s_{\mathrm{ALCS}}}(\max)$	$\overline{P}_{s_{\mathrm{ALCS}}}$	$\frac{K_{13}\overline{P}_{s_{\mathrm{IALCS}}}}{\overline{P}_{s_{\mathrm{ALCS}}}(\max)}$	$\overline{P}_{s_{\mathrm{ALDS}}}(\max)$	$\overline{P}_{s_{\mathrm{ALDS}}}$	$\frac{K_{13}\overline{P}_{s_{\mathrm{ALDS}}}}{\overline{P}_{s_{\mathrm{ALDS}}}(\max)}$
4. Content Creation, $\overline{C}_c$	K_{14}	$\overline{C}_{c_{\mathrm{ALCS}}}(\max)$	$\overline{C}_{c_{\mathrm{ALCS}}}$	$\frac{K_{14}\overline{C}_{c_{\mathrm{ALCS}}}}{\overline{C}_{c_{\mathrm{ALCS}}}(\max)}$	$\overline{C}_{c_{\mathrm{ALDS}}}(\max)$	$\overline{C}_{c_{\mathrm{ALDS}}}$	$\frac{K_{14}\overline{C}_{c_{\mathrm{ALDS}}}}{\overline{C}_{c_{\mathrm{ALDS}}}(\max)}$
Total	$\sum_{j=1}^{4} K_{ij} = 1$			$\sum_{j=1}^{4} O_{ij}$			$\sum_{j=1}^{4} O_{ij}$

Quality subcriterion	Weight of the subcriterion	Application level structure which is analyzed					
		ALCS			ALDS		
		Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})	Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})
1	2	3	4	5	6	7	8
II.From the service provider's viewpoint							
1. The complexity of the logic of implementation and of administration of the service, $\overline{St}$	K_{21}	$\overline{St}_{\mathrm{ALCS}}(\min)$	$\overline{St}_{\mathrm{ALCS}}$	$\frac{K_{21}\overline{St}_{\mathrm{ALCS}}(\min)}{\overline{St}_{\mathrm{ALCS}}}$	$\overline{St}_{\mathrm{ALDS}}(\min)$	$\overline{St}_{\mathrm{ALDS}}$	$\frac{K_{21}\overline{St}_{\mathrm{ALDS}}(\min)}{\overline{St}_{\mathrm{ALDS}}}$
2. The number of successfully provided IS, $\overline{S}$	K_{22}	$\overline{S}_{\mathrm{ALCS}}(\max)$	$\overline{S}_{\mathrm{ALCS}}$	$\frac{K_{22}\overline{S}_{\mathrm{ALCS}}}{\overline{S}_{\mathrm{ALCS}}(\max)}$	$\overline{S}_{\mathrm{ALDS}}(\max)$	$\overline{S}_{\mathrm{ALDS}}$	$\frac{K_{22}\overline{S}_{\mathrm{ALDS}}}{\overline{S}_{\mathrm{ALDS}}(\max)}$
Total	$\sum_{j=1}^{2} K_{ij} = 1$			$\sum_{j=1}^{2} O_{ij}$			$\sum_{j=1}^{2} O_{ij}$

Quality subcriterion	Weight of the subcriterion	Application level structure which is analyzed					
		ALCS			ALDS		
		Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})	Maximal or minimal admissible value	The resulting value	Evaluation (O_{ij})
1	2	3	4	5	6	7	8
III.From the equipment supplier's viewpoint							
1. Time for Service Provision, $\overline{T}_{cs}$	K_{31}	$\overline{T}_{cs_{ALCS}}(\min)$	$\overline{T}_{cs_{ALCS}}$	$\frac{K_{31}\overline{T}_{cs_{ALCS}}(\min)}{\overline{T}_{cs_{ALCS}}}$	$\overline{T}_{cs_{ALDS}}(\min)$	$\overline{T}_{cs_{ALDS}}$	$\frac{K_{31}\overline{T}_{cs_{ALDS}}(\min)}{\overline{T}_{cs_{ALDS}}}$
2. The probability of Refusing at Service Provision, $\overline{P}_B$	K_{32}	$\overline{P}_{B_{\text{ALCS}}}(\min)$	$\overline{P}_{B_{\text{ALCS}}}$	$\frac{K_{32}\overline{P}_{B_{\text{ALCS}}}(\min)}{\overline{P}_{B_{\text{ALCS}}}}$	$\overline{P}_{B_{\text{ALDS}}}(\min)$	$\overline{P}_{B_{\text{ALDS}}}$	$\frac{K_{32}\overline{P}_{B_{\text{ALDS}}}(\min)}{\overline{P}_{B_{\text{ALDS}}}}$
3. Structural survivability, $\overline{P}_{st}$	K_{33}	$\overline{P}_{st_{ALCS}}(\max)$	$\overline{P}_{st_{ALCS}}$	$\frac{K_{33}\overline{P}_{st_{ALCS}}}{\overline{P}_{st_{ALCS}}(\max)}$	$\overline{P}_{st_{ALDS}}(\max)$	$\overline{P}_{st_{ALDS}}$	$\frac{K_{33}\overline{P}_{st_{ALDS}}}{\overline{P}_{st_{ALDS}}(\max)}$
4. Reliability, $\overline{R}$	K_{34}	$\overline{R}_{\text{ALCS}}(\max)$	$\overline{R}_{\text{ALCS}}$	$\frac{K_{34}\overline{R}_{\text{ALCS}}}{\overline{R}_{\text{ALCS}}(\max)}$	$\overline{R}_{\text{ALDS}}(\max)$	$\overline{R}_{\text{ALDS}}$	$\frac{K_{34}\overline{R}_{\text{ALDS}}}{\overline{R}_{\text{ALDS}}(\max)}$
5. Cost of intelligent superstructure, ¦	K_{35}	¦$_{\text{ALCS}}(\min)$	¦$_{\text{ALCS}}$	$\frac{K_{35}¦_{\text{ALCS}}(\min)}{¦_{\text{ALCS}}}$	¦$_{\text{ALDS}}(\min)$	¦$_{\text{ALDS}}$	$\frac{K_{35}¦_{\text{ALDS}}(\min)}{¦_{\text{ALDS}}}$
Total	$\sum_{j=1}^{5} K_{ij} = 1$			$\sum_{j=1}^{5} O_{ij}$			$\sum_{j=1}^{5} O_{ij}$

Table 1. The calculations of the relative and total weighted ratings

3. Subcriteria are defined for each participant which should be taken into account when creating the complex criterion. The method for calculating each subcriterion has been proposed.

4. The problem of association of subcriteria into the resulting complex criterion has been solved. For the formation of complex criterion, a methodical approach has been proposed, which takes into account the performance of the next steps:

- Decision about the degree of influence of each participant of the IS process on the value of the complex criterion of quality of control.
- Selecting subcriteria for each participant of the process.
- Creating the weighting coefficients for subcriteria.
- Assessing the value achieved by each subcriterion.
- Calculating the resulting scores for each participant of the process.
- Calculating the complex criterion of quality.

The suggested methodical approach allows us to determine a complex quality criterion for providing ISes, and it can be used by designers of NGN, as well as by providers of typical multi-service network, for making the decisions concerning the choice of the application level structure.

References

[1] ITU-T: Telecommunication Standardization Sector, *Next Generation Networks: Frameworks and functional architecture models*, In: General overview of NGN, ITU-T Recommendation Y.2001 (12/2004), Printed in Switzerland Geneva, 2005, p. 5.

[2] J.L. Salina, P. Salina, *Next Generation Networks: Perspectives and Potentials*, Wiley, 2008.

[3] A. Yatsko, W. Susłow, *Insight into Theoretical and Applied Informatics. Introduction to information technologies and computer science*, De Gruyter Open, 2015.

[4] B.S. Goldstein, *Post-NGN communication networks*, In: B.S. Goldstein, A.E. Kucheryavy. St. Petersburg: BHV-Petersburg, 2013.

[5] V.G. Krivuca, V.K. Steklov, L.N. Berkman, B.J. Kostik, V.F. Olijnik, *Telecommunications control with the latest technology*, Kiev: Technika, 2007.

[6] S.V. Shestopalov, *Architecture of NGN with centralized and decentralizing control system*, In: 10th Anniversary International Scientific Conference "Modern Problems of Radio Engineering, Telecommunications and Computer Science" TCSET'2010, Lviv, 2010.

[7] N.A. Kniazieva, S.V. Shestopalov, *Improving the quality of service management by using the decentralized management system*, Visnyk DUIKT[1], vol. 8, issue 1, p. 21-28.

[8] S.V. Shestopalov, N.A. Kniazieva, *The method of improvement of the services management process*, In: II International scientific and technical symposium "New technologies in telecommunications" - (DUIKT- Carpathians' 2009, p. Vyshkiv), Kyiv, DUIKT, 2009, pp. 39-40.

[1]State University of Telecommunications, Ukraine

[9] S.V. Shestopalov, *Indicators of services management quality NGN*, In: Scientific Conference "Modern information and communication technologies" COMINFO'2008-Livadia, Theses of the report, Kyiv, DUIKT, 2008, pp. 187-189.

[10] N.A. Kniazieva, *Evaluation of the quality of communication services from the point of view of customer satisfaction*, In: N.A. Kniazieva, A.S. Kalchenko, Science and Education a New Dimension: Natural and Technical Science, vol. 8, Budapest, 2013, p. 156-161.

[11] *Network performance objectives for IP-based services*, ITU-T: Telecommunication Standardization Sector, Y.1541, Geneva, 2007.

[12] *About International Telecommunication Union (ITU)*, http://www.itu.int/en/about/Pages/default.aspx

Received: May 2017

Nina Kniazieva
Odessa National Academy of Food Technologies,
Department of information technology and cyber security,
1/3 Dvoryanskaya Street, 65-082 Odessa, Ukraine
E-mail address: knyazeva@ukr.net

Serhii Shestopalov
Odessa National Academy of Food Technologies,
Department of information technology and cyber security,
1/3 Dvoryanskaya Street, 65-082 Odessa, Ukraine
E-mail address: shestopalov.s@mail.ru

Walery Susłow
Koszalin University of Technology, Poland,
Faculty of Electronics and Computer Science,
2 Śniadeckich Street, 75-453 Koszalin, Poland
E-mail address: walery.suslow@tu.koszalin.pl

Andrzej Yatsko
Koszalin University of Technology, Poland,
Faculty of Civil Engineering Environmental and Geodetic Sciences,
2 Śniadeckich Street, 75-453 Koszalin, Poland
E-mail address: andrzej.jacko@gmail.com

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
117–127
DOI http://dx.doi.org/10.16926/m.2017.22.09

THEORETICAL AND METHODOLOGICAL BASES OF MODULAR TECHNOLOGY OF PARALLEL TABULAR COMPUTATIONS USING UNIVERSAL PROCESSORS

MIKHAIL SELIANINAU

Abstract

In this paper, we deal with the methodology for application of the theory and methods of the tabular technology of modular information processing on the basis of modern computing machinery. The use of the minimal redundant modular number system and the interval-modular form of representation of an integer number determined by its modular code creates the computer-arithmetical basis of a methodology under consideration. The main advantage of the offered methodology consists inincrease of the computation speed and accuracy at the organization of high-precision arithmetic processing of multi-digit data by means of universal processors on the basis of minimal redundant modular encoding method.

1. Introduction

At present, the branch of high-performance computing is developing extremely rapidly. The development of computer science determines a rapid increase of computing scales that leads to qualitatively new requirements imposed on numerical methods and computing algorithms. The main requirement is the receiving in an acceptable time of the correct results of solving the task which are not distorted by rounding errors.

The need for science in high-speed and high-precision computing has always existed and now it is especially urgent. The most important tasks sensitive to the computing speed and accuracy arise in various areas: in physics and cosmology, in experimental mathematics and computational geometry, in the growing fields of high-tech industry and so on [1–5]. A characteristic feature of such problems is their high dimensionality. In this connection, the time to solve the task becomes a very critical parameter.

• *Mikhail Selianinau* — m.selianinov@ajd.czest.pl
Jan Długosz University in Częstochowa.

The main reason affecting the accuracy of calculations is the use of rounding in arithmetic operations that is caused by the fixed and relatively small length of operands in modern processors. At the same time, the choice of the optimal algorithm which is stable to rounding errors is often time-consuming. Therefore, the main way to get the correct results of numerical calculations is the use of the program processing of extremely large numbers for which the digit capacity exceeds the finite computer word length. However, the construction of large number arithmetic on the basis of positional number systems leads to a significant and often unacceptable reduction of the processing speed.

The top manufacturers of modern processor platforms first of all place special emphasis on parallelism. The technologies of vector and multi-core computing are actively developing. Now there are more perfect SIMD-commands that allow us to process simultaneously increasingly more pairs of operands, the compilers can perform automatic vectorization under the condition of the suitable data structure, the number of processor cores grows steadily and moreover there are specialized multicore coprocessors [6]. As a result, there is a task of developing a new approach to representation of multi-bit numbers which allows an easy parallelization of calculations at the level of elementary arithmetic operations.

In view of the sequential internal structure of classical positional arithmetic algorithms the traditional implementation of both high-speed and high-precision computations on their basis is most often inefficient and, in many cases, is just unacceptable. Therefore, research and development of new flexible techniques of reliable parallel data processing in the set of real numbers as well as on more complicated mathematical models are very relevant today.

The use of modular arithmetic (MA) which has a natural internal data parallelism is a promising scientific direction in the field of high-speed high-precision computation [7–10]. The lack of carries between the adjacent digits of modular number representation allows performing the main arithmetic operations easily and quickly, vectorizing them effectively and allocating on computation kernels.

The method of implementing of computer arithmetic on the base of the number representation in the modular computational basis is the most consistent among others with the concept of developing modern high-performance systems and technologies of parallel programming.

2. The basic notation and terminology

Let us introduce the following notation: $\mathbf{Z}$ is the set of all integers;

$\lfloor x \rfloor$ denotes the floor of x, i.e. the greatest integer less than or equal to x, $\lfloor x \rfloor = \max \{ y \in \mathbf{Z} \mid y \leq x \}$;

$\lceil x \rceil$ denotes the ceiling of x, i.e. the smallest integer greater than or equal to x, $\lceil x \rceil = \min \{ y \in \mathbf{Z} \mid y \geq x \}$;

$\mathbf{X} \times \mathbf{Y} = \{ \forall (x, y) \mid x \in \mathbf{X},\ y \in \mathbf{Y} \}$ is the Cartesian product of two sets $\mathbf{X}$ and $\mathbf{Y}$;

$\gcd(A, B)$ stands for the greatest common divisor of two integers A and B;

$\mathbf{Z}_m = \{0, 1, \ldots, m-1\}$ and $\mathbf{Z}_m^- = \left\{ -\left\lfloor \frac{m}{2} \right\rfloor, -\left\lfloor \frac{m}{2} \right\rfloor + 1, \ldots, -\left\lceil \frac{m}{2} \right\rceil - 1 \right\}$ are the sets (rings) of least nonnegative residues and absolutely least residues modulo $m > 1$, respectively;

$|x|_m$ denotes the element of $\mathbf{Z}_m$ congruent to x modulo m;

$\left|X^{-1}\right|_m$ designates the multiplicative inversion of an integer X modulo m ($\gcd(X, m) = 1$);

$m_1, m_2 \ldots, m_k$ are the natural modules ($k \geq 2$);

$M_l = \prod_{i=1}^{l} m_i$, $M_{i,l} = M_l / m_i$ $(l = k-1, k;\ i = 1, 2, \ldots, k)$

$\left(|X|_{m_1}, |X|_{m_2}, \ldots, |X|_{m_k} \right)$ represents a modular code (MC) of the integer $X \in \mathbf{Z}$ with respect to the basis $\{m_1, m_2, \ldots, m_k\}$.

3. The basic principles of modular number processing

As is known, the parallel computing structures play a fundamental role in the modern computer science and its numerous applications. The analysis of the modern directions of development of parallel computing structures shows that in recent years special attention has been paid to the so-called parallel ring structures based on modular number systems (MNS). The theoretical basis of MA is created by abstract algebra and number theory [7, 11–13].

A classical non-redundant MNS on the set $\mathbf{Z}$ is determined by pairwise relatively prime modules $m_1, m_2 \ldots, m_k$ ($\gcd(m_i, m_j) = 1$; $i, j = 1, 2, \ldots, k$; $i \neq j$; $k > 1$) using the definition of a homomorphic mapping $\mathbf{\Phi}$: $\mathbf{Z} \to \mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times \ldots \times \mathbf{Z}_{m_k}$ which assigns the set $(\chi_1, \chi_2, \ldots, \chi_k)$ of residues $\chi_i = |X|_{m_i}$ of the division of an integer X by a modulo m_i $(i = 1, 2, \ldots, k)$ to each $X \in \mathbf{Z}$. In this case the standard notation $X = (\chi_1,\ \chi_2, \ldots, \chi_k)$ is used.

It is known that the set of integers satisfying to the system of simultaneous linear congruences

$$\begin{cases} X \equiv \chi_1 \pmod{m_1} \\ X \equiv \chi_2 \pmod{m_2} \\ \quad \dots \\ X \equiv \chi_k \pmod{m_k} \end{cases} \tag{1}$$

corresponds to an MC of integer X. If $m_1, m_2, \dots, m_k$ are pairwise relatively prime modules, then the simultaneous congruences (1) have a unique solution being the residue class modulo M_k determined by the congruence

$$X \equiv \sum_{i=1}^{k} M_{i,k}\mu_{i,k}\chi_i \pmod{M_k}, \tag{2}$$

where $\mu_{i,k} = \left|M_{i,k-1}^{-1}\right|_{m_i}$. In essence, the formula (2) represents the so-called Chinese remainder theorem (CRT) [7, 11].

The practical application of the MNS assumes that instead of a residue class there is only a single integer corresponding to the MC $(\chi_1,\ \chi_2, \dots, \chi_k)$. Therefore, in the MNS with the basis $\{m_1, m_2, \dots, m_k\}$ one or the other set of representatives of the residue classes is used as the numerical range $\mathbf{D}$ in order to ensure the required one-to-one mapping $X \to (\chi_1,\ \chi_2, \dots, \chi_k)$. In this case, the maximum cardinality of a set of integers $\mathbf{D}$ is equal to M_k.

In computer applications the rings $\mathbf{Z}_{M_k}$ and $\mathbf{Z}_{M_k}^{-}$ are usually used as a numerical range of the MNS. Taking the above into account, in the first case the modular coding is defined as a mapping $\mathbf{\Phi}_{MNS} : \mathbf{Z}_{M_k} \to \mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times \dots \times \mathbf{Z}_{m_k}$ assigns the MC $(\chi_1, \chi_2, \dots, \chi_k)$ to each $X \in \mathbf{Z}_{M_k}$. A decoding mapping $\mathbf{\Phi}_{MNS}^{-1} : \mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times \dots \times \mathbf{Z}_{m_k} \to \mathbf{Z}_{M_k}$ based on the relationship (2) operates according to the following rule

$$X = \left|\sum_{i=1}^{k} M_{i,k}\chi_{i,k}\right|_{M_k} \quad (\chi_{i,k} = |\mu_{i,k}\chi_i|_{m_i}). \tag{3}$$

Let two arbitrary integers A and B be represented in the MNS with the base numbers $m_1, m_2 \dots, m_k$, i.e. $A = (\alpha_1, \alpha_2, \dots, \alpha_k)$, $B = (\beta_1, \beta_2, \dots, \beta_k)$ $\left(\alpha_i = |A|_{m_i},\ \beta_i = |B|_{m_i},\ i = 1, 2, \dots, k\right)$. Then each ring (modular) operation $\circ \in \{+, -, \times\}$ on them has the following general form:

$$A \circ B = \left(|\alpha_1 \circ \beta_1|_{m_1}, |\alpha_2 \circ \beta_2|_{m_2}, \dots, |\alpha_k \circ \beta_k|_{m_k}\right). \tag{4}$$

In other words, the result of the operation " $\circ$ " on two integers A and B in the modular representation is formed by an independent application of " $\circ$ " to all pairs of their residues $(\alpha_i,\ \beta_i)$ with respect to the corresponding modules m_i : $|\alpha_i \circ \beta_i|_{m_i}$, $i = 1, 2, \dots, k$. The main fundamental advantage

of the MA over a position number systems (PNS) arithmetic consists in parallel digit-wise realization of ring operations (see (4)).

In an MNS, the MC does not contain explicit information about the value of the element of the numerical range that corresponds to it. In contrast to the modular operations (4), for the realization in MA of the so-called non-modular operations it is not enough only to have the separate residue values but the evaluation of positional values of the whole number is also required. Therefore, for the performing non-modular operations in MNS it is necessary to use the such forms for number representation (via the residues of MC) which allow us to obtain some characteristics for determining the position of a given number in a numerical range. Such operations include scaling, division (in the general case), multiplication of fractions, sign detection, overflow control, direct and inverse conversions of the positional and modular codes, errors detection and correction.

It is obvious that in the final analysis the complexity of computing of the so-called integral characteristics of MC determines the effectiveness of a MA created on their basis. The optimization of integral characteristic base of a MNS (first of all, MNS with integer ranges in the first place) plays a key role in the development of efficient MA [7, 8, 14-16].

4. Minimal redundant modular coding

Within the framework of development of the theory and applications of modular computing structures (MCS) a special place is devoted to optimization of methods and algorithms of performance of nonmodular operations in MSS with respect to the input redundancy of coding of the elements of the numerical ranges, the implementation time of these operations, the throughput of modular processors, etc.

It is known that the use of code redundancy often allows us to improve significantly the arithmetic or other properties of the number system, including a MNS. The immediate application of expression (3) as the basic form of integer numbers representation for the synthesis of non-modular procedures is practically unacceptable due to the complexity of direct computer implementation, especially in the case of large values of M_k. This is caused by the fact that the CRT demands a large modulo operations which are very labor-consuming, especially when using a wide ranges. In addition, the CRT is not well adapted for performing operations on the elements from the symmetric ranges and is not suitable for design of scaling procedures.

At the same time, the parallel representation forms of the integer numbers on the basis of CRT (3), which provide a significant simplification of non-modular operations and have much better realization properties in

comparison with non-redundant analogs, can be obtained by introducing some code redundancy.

The priority positions in this context belong to a minimal redundant modular coding, an interval index characteristic and an interval modular form (IMF) of integer numbers associated with it [7, 8, 15, 16]. A minimal redundant modular coding $\mathbf{\Phi}_{MRMNS} : \mathbf{D} \to \mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times \ldots \times \mathbf{Z}_{m_k}$ provides the use of a working numerical range $\mathbf{D}$ whose cardinality is less than the cardinality of the range of classical (non-redundant) MNS with modules $m_1, m_2, \ldots, m_k$, i.e. $|\mathbf{D}| < M_k$.

A redundant MNS on the set of integers $\mathbf{Z}$ is determined by means of $k > 1$ pairwise relatively prime modules $m_1, m_2, \ldots, m_k$ and an additional module m_0 satisfying the conditions $\gcd(m_i, m_0) = 1$ $(i = 1, 2, \ldots, k)$. In this case, a set $\mathbf{Z}_{2M}^{-} = \{ -M, -M+1, \ldots, M-1 \}$, where $M = m_0 M_{k-1}$, is usually used as a numerical range $\mathbf{D}$. In view of the condition $2M < M_k$, the additional module m_0 should be chosen on the basis of the following estimation $2m_0 < m_k$. The resulting redundant MNS is naturally a narrowing of the initial non-redundant MNS and possesses all its advantages.

The decoding mapping $\mathbf{\Phi}_{MRMNS}^{-1} : \mathbf{Z}_{m_1} \times \mathbf{Z}_{m_2} \times \ldots \times \mathbf{Z}_{m_k} \to \mathbf{D}$, which is a restoration of a number $X \in \mathbf{D}$ by residues of its MC $(\chi_1, \chi_2, \ldots, \chi_k)$, is carried out using the so-called interval index (II) $I(X)$ and a corresponding IMF of a number X which is represented as

$$X = \sum_{i=1}^{k-1} M_{i,k-1} \left| M_{i,k-1}^{-1} \chi_i \right|_{m_i} + I(X) M_{k-1}. \tag{5}$$

According to a CRT, the interval-index characteristic $I(X)$ of the MC is uniquely determined by the relation (5) [7, 8]. At the same time, an IMF (5) is free from the above-mentioned disadvantages of using a CRT (see (2) and (3)).

In order to achieve the required level of code redundancy of a MNS with the range $\mathbf{D} = \mathbf{Z}_{2M}^{-}$ the conditions under which the remainder $\hat{I}_k(X) = |I(X)|_{m_k}$ uniquely determines the II $I(X)$ of each number $X \in \mathbf{D}$ were obtained in [8, 15]. For this purpose it is necessary and sufficient that the kth module m_k of the MNS satisfies the condition $m_k \geq 2m_0 + \rho$, where $\rho = \max\{\rho_{k-1}(X)\}$ is the maximum value of the rank characteristic of the $(k-1)$th order [7, 8]. As this takes place, the minimum redundancy is achieved in the case when the equality $m_k - 2m_0 - \rho = |m_k - \rho|_2$ is satisfied.

The main advantage of MRMNS over non-redundant analogs consists in a significant simplification of the calculation of the interval index characteristic $I(X)$. In spite of the fact that the introduced redundancy is very small,just owing to it, the calculation of the II $I(X)$ is actually reduced to summation of k residues modulo m_k, i.e. it is a trivial operation. In

contrast, in order to calculate the traditionally used integral characteristics of MC (the mixed-radix digits, the rank characteristics and some others) it is necessary to perform the calculation of the sums of k sets of residues with respect to modules of MNS.

The use of a MRMNS instead of a MNS reduces the complexity of algorithm for computation of a base interval-index characteristic expressed by the number of the necessary of tables and the number of modulo additions from $O(k^2)$ to $O(k)$. In the final analysis, in particular this provides a simplicity of non-modular procedures synthesized on the basis of IMF (5), first of all, of the procedures for conversion of a MRMC to a position code and scaling which play an extremely important role in the technology of modular information processing (TMIP). The realization of the advantages of MRMNS over non-redundant MNS with respect to minimizing computational costs of non-modular operations opens up wide possibilities for creation of systems that are characterized by minimal computational burden in a class of equivalent modular analogs in terms of productivity, calculation accuracy and complexity of the basic reconfiguration mechanism [7, 8].

5. The fundamental advantages of a MCS over the position computing structures

An MNS has the maximum level of internal parallelism among a set of all number systems. Due to this property, the MA allows us:

- to expand the set of used modules without complicating the algorithms of ring operations, i.e. to increase the digit capacity of the elements of the numeric ranges;
- to change the number of the bases of MNS, i.e. the accuracy of data processing;
- to perform efficiently the tabular calculations both at the hardware and software level;
- to apply the principle of formal information processing at the modular segments of computational procedures; in the framework of this principle the cardinality of the numerical range is focused only on the final results of the calculations;
- to organize extremely simply the multiprocessor mode of calculations;
- to rebuild flexibly a system configuration changing the sets of modules and operating tables.

The features listed above as well as some other unique properties of MA open exclusively wide opportunities for execution of parallel high-precision calculations at a qualitatively new level (with respect to performance, simplicity of accuracy change and flexibility of operating modes).

The MCS constitute a unique structural method for the decomposition of computation procedures into a set of subprocesses independent from each other. The natural code parallelism along with the properties listed above provides a number of fundamental advantages of an MNS over a PNS, in particular:

- a complete independence of the time period of modular (ring) operations from the number of bases and hence from the code length of MNS;
- the high computational speed for the ranges of large numbers;
- simplicity of computation pipelining at the level of low-bit operations;
- an exclusively high performance of data processing for mathematical models more complicated than the real ranges, for example, the sets of complex numbers, quaternions, polynomials, etc.;
- efficiency of modular code constructions for error control and correction.

Because of the noted advantages an MCS ideally correlates with the concepts of advanced computer technology including the underlying ideologies of supercomputer, multicore processor, neural network and some similar systems of parallel information processing [7-10].

Taking into account the reasons mentioned above, during the past 10-15 years the modular direction goes through a stage of rapid development in computer science, digital signal processing and other fields of science and technology.

6. The methodology for constructing the multiprocessor information processing in MRMNS

The information processing technology is evolving rapidly and steady expands its applications and scope. Along with a hardware approach for creating high-performance parallel information processing systems (PIPS) based on MNS, the alternative multiprocessor methodology which allows us to realize the advantages of MCS at the software level, can also be applied.

If some computation process is modular in nature, i.e. is composed of modular operations (see (4)), then it can be parallelized by executing subprocesses associated with the different modules of MNS on separate universal processors (cores) of modern computers. At the same time, one of the most important (from a practical point of view) properties of MCS consisting in their unique reconfiguration capability can be used to a great extent.

The ability of MCS to change flexibly its own configuration is conditioned by the table nature of MA and in practice is achieved by trivial changing of sets (arrays) of useful constants, including the bases of MNS and corresponding operation tables.

A multiprocessor PIPS functioning in MRMNS with pairwise relatively prime modules $m_1, m_2, \ldots, m_k$ and the operating range $\mathbf{D} = \mathbf{Z}_{2M}^{-}$ is realized by using the k universal processors (cores) $PC_1, PC_2, \ldots, PC_k$, the ith of which performs the role of the modular datapath with respect to the base $m_i\,(i = 1, 2, \ldots, k)$. Thus, when executing any modular operation of the type (4) on integers A and B represented by their MRMC, the processor PC_i evaluates the ith digit $\gamma_i = |\alpha_i \circ \beta_i|_{m_i}$ of the result $C = A \circ B$ of the operation $\circ \in \{+, -, \times\}$. Therefore, in a multiprocessor PIPS the modular computing processes are reduced to subprocesses executed independently of each other on different processors (cores) according to the corresponding modules of MRMNS.

It is easy to see that an arbitrary computational process that does not contain logical operations in principle can always be transformed into a purely single modular computational process which in the most general case is not a one-step process but has a recursive organization and represents a set of typical (standard) computational procedures (segments). It is necessary to bear in mind that the application of the modular mode (also called the formal calculation mode (FCM)) is justified only when the calculation results for the modular segment of the implementable procedure do not exceed the bounds of the standard operating range of the MRMNS. But at the same time, the results of intermediate calculations may not meet this requirement. The noted circumstance exerts a decisive influence on the choice not only of the range $\mathbf{D}$, but also of the range of input data $\hat{\mathbf{D}} = \mathbf{Z}_{2P} = \{-P, -P+1, \ldots, P\} \subset \mathbf{D}$ (P is a some natural number). In this connection, the problem of choosing the cardinalities $2P$ and $2M$ of the ranges $\hat{\mathbf{D}}$ and $\mathbf{D}$, respectively, is governed by the strategy ensuring the conditions that allow us to realize all the most important advantages of MRMA to the maximum extent [7, 8]. This goal can be achieved if in the FCM the final results for any modular segment of the computational procedures are the elements of the range $\mathbf{D}$ for all permissible values of the input variables.

The multiprocessor PIPS on the basis of the MRMA allow us to use systems of modules $m_1, m_2, \ldots, m_k$ which can theoretically take on any integer values within the range of modern computers including the biggest ones, for example $2^{16} \pm p, 2^{32} \pm p, 2^{64} - p$ $(p = 0, 1, 2, \ldots)$. Thus, the essentially new possibilities (in comparison with traditional approaches) are opened up for extending the used number range in order to ensure the necessary conditions for the wider and more effective application of the formal calculation mode for the implementation of both modular and quasi-modular processes.

Another distinctive and very important feature of the modular multiprocessor PIPS is the extreme simplicity of modification of the used set of modules $m_1, m_2, \ldots, m_k$ and the corresponding package of operating tables

by the purely programmable way, i.e. the simplicity of system reconfiguration. This reconfiguration operation can also be performed taking into account the possible change (increase or decrease) of the number of processors (cores).

7. Conclusions

In conditions of the rapid development and wide distribution of modern computer technology the multiprocessor systems of modular information processing represent a very promising and competitive alternative towards the systems which designed within the framework of the traditional methods that focus mainly on hardware-intensive approach.

The originality of the research and development of the new multiprocessing PIPS is determined by using the MRMNS as a computer-arithmetic basis. The MRMNS allow us to develop a much simpler and more effective arithmetic both on the real ranges and as well as on more complicated mathematical models (the sets of complex numbers, quaternions, polynomials, etc.) in comparison with the classical MNS [7, 8, 15-19].

The implementation of the formulated approach allows us:

- to extend or reduce the internal composition of the current set of modules without changing the execution procedure for the modular segments of computation processes, i.e. to regulate the calculation accuracy;
- to use super-large modules (in the range of values from 2^{16} to 2^{64}) and therefore the MNS operating ranges of an extra-high cardinality;
- to perform efficiently a lookup table data processing by means of arrays having a volume up to 2^{32} or more elements;
- to expand considerably the action limits of the FCM for computer models of computational procedures;
- to reconstruct flexibly the configuration of the modular information processing system by the programmable change of the sets of basic modules and operating tables.

It should be emphasized that with the evolution of multiprocessor computers the realization of the advantages of MNS at the program level is extremely simplified. At the same time, the attainable practical effects increase significantly, this primary concerns the best performance and accuracy.

References

[1] U. Kulisch, *Computer Arithmetic and Validity: Theory, Implementation, and Applications*. De Gruyter Studies in Mathematics 33, Berlin, 2013.

[2] G.B. Arfken, H.J. Weber, F.E. Harris, *Mathematical Methods for Physicists. A Comprehensive Guide. 7th edition*. Academic Press, Boston, 2012.

[3] C.W. Ueberhuber, *Numerical Computation 1: Methods, Software, and Analysis*. Springer-Verlag, Berlin, 2013, DOI 10.1007/978-3-642-59118-1.
[4] C.W. Ueberhuber, *Numerical Computation 2: Methods, Software, and Analysis*. Springer-Verlag, Berlin, 2013.
[5] A.R. Krommer, C.W. Ueberhuber, *Numerical Integration: on Advanced Computer Systems (Lecture Notes in Computer Science)*. Springer-Verlag, Berlin, 2008.
[6] S. Mittal, *A Survey of Techniques for Architecting and Managing Asymmetric Multicore Processors*. ACM Computing Surveys, **48:3** (2016), 45:1–45:38, DOI 10.1145/2856125.
[7] A.A. Kolyada, I.T. Pak, *Modular Structures of Pipeline Digital Information Processing*. University Press, Minsk, 1992 (in Russian).
[8] A.F. Chernyavsky, V.V. Danilevich, A.A. Kolyada, M.Y. Selyaninov, *High-speed Methods and Systems of Digital Information Processing*. Belarusian State University Press, Minsk, 1996 (in Russian).
[9] P.V. Ananda Mohan, *Residue Number Systems: Algorithms and Architectures*. Kluwer Academic Publishers, 2002.
[10] A. Omondi, B. Premkumar, *Residue Number Systems: Theory and Implementation*. Imperial College Press, London, 2007.
[11] I.M. Vinogradov, *Fundamentals of Number Theory*. Nauka, Moscow, 1981 (in Russian).
[12] M.F. Atiyah, I.G. Macdonald, *Introduction to Commutative Algebra*. Addison-Wesley Publishing Co., Reading, Massachusetts, 1969.
[13] D.S. Dummit, R.M. Foote, *Abstract algebra. 3rd edition*. John Wiley & Sons, Inc., Hoboken, NJ, 2004.
[14] A.A. Kolyada, M.Y. Selyaninov, *On the formation of the integral characteristics of the codes of residue number systems with the symmetrical range*. Cybernetics **4** (1986), 20–24 (in Russian).
[15] M. Selyaninov, *Modular technique of parallel information processing*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XIII** (2008), 43–52.
[16] M. Selyaninov, *Construction of modular number system with arbitrary finite ranges*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XIV** (2009), 105–115.
[17] M. Selianinau, *High- speed modular structures for parallel computing in the space of orthogonal projections*. Scientific Issues, Jan Długosz University of Czestochowa, Ser. Technical and IT Education, **V** (2010), 87–96.
[18] M. Selyaninov, *Modular number systems in a complex plane*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XV** (2010), 131–138.
[19] M. Selyaninov, *Arithmetic of quadratic minimal redundant modular number systems*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XVI** (2011), 129–134.

Received: May 2017

Mikhail Selianinau
Jan Długosz University,
Institute of Technics and Safety Systems,
42-200 Częstochowa, Al. Armii Krajowej 13/15, Poland
E-mail address: m.selianinov@ajd.czest.pl

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
129–139
DOI http://dx.doi.org/10.16926/m.2017.22.10

TABULAR MINIMAL REDUNDANT MODULAR STRUCTURES FOR FAST AND HIGH-PRECISION COMPUTATIONS USING GENERAL-PURPOSE COMPUTERS

MIKHAIL SELIANINAU

Abstract

The present paper is a continuation of research in parallel information processing based on the tabular modular computing structures. We deal with the methodology of using a minimal redundant modular number system for high-speed and high-precision computation by means of modern universal multicore processors. Advantages of formal computing mode on the base of modular arithmetic are demonstrated by the example of implementation of digital signal processing procedures. The additive and additive-multiplicative formal computing schemes with the obtained estimations of the cardinality of working ranges for the realization of calculations are presented in the article.

1. Introduction

In recent years, the tabular methods of digital information processing, both at the hardware and software levels, have been widely used to solve the problems of performance improvement, flexible organization of adaptive operating modes and some others in modern computer algebra and arithmetic, as well as in their numerous applications in such fields of science and technology as digital signal processing (DSP), image recognition and image processing, various purpose identification systems, artificial intelligence systems, information protection and similar systems [1-6].

However, within a framework of classical computer-aided algorithmic foundations based on the arithmetic of positional number systems (PNS) there are a number of objective factors that impede intensification of the development and practical implementation of tabular computer structures.

• *Mikhail Selianinau* — m.selianinov@ajd.czest.pl
Jan Długosz University in Częstochowa.

First of all, such factors include the sequential internal nature of PNS which appears in the presence of interdigit carry propagation during the execution of arithmetic operations that most often does not allow us to carry out an efficient decomposition of the positional computer structures (PCS) into the components acceptable for tabular implementation. Thus, the principles of tabular information processing and the PNS arithmetic totally contradict each other. Taking the above into account, the research aimed to eliminate the indicated contradictory situation seems to be very topical.

The analysis of modern developing information technologies from the point of view of their suitability for the organization of tabular adaptive computing indicates that the modular technology provides the widest range of possibilities in this respect [7-10]. This is caused by the unique property of modular number system (MNS) to perform in a natural way the decomposition of computational processes into the independent or quasi-independent subprocesses determined on mathematical models whose facilities have dimensions many times smaller (by the number of bases) in comparison with a very large dynamic range of the applied MNS.

In modern applications of modular arithmetic (MA) the development of high-performance parallel systems, which operate completely in the so-called formal computing mode (FCM), occupies a highly important place [7, 8]. This mode is characterized by the absence of rounding on the modular segments of the computational processes, i.e. on segments consisting only of modular operations: addition, subtraction and multiplication in the MNS without overflow check. At the same time it is assumed that the computation results on each modular segment do not exceed the limits of the used dynamic range. Since the arithmetic operations in the modular code (MC) are carried out independently for each of the modules, then in the framework of the FCM a high performance is achieved due to the parallel nature of MA along with the absence of rounding.

In recent years, the modular information-processing systems (MIPS) are mostly applied to the implementation of high-precision and absolutely exact computations. First of all, this concerns the applications in the field of DSP, pattern recognition and image processing as well as in information protection. Currently, among the systems of the specified class the priority positions are occupied by the MIPS which can be implemented programmatically, i.e. without the use of special hardware. Within the framework of existing and intensively developing computer technologies of parallel processing the facilities to implement the FCM at the program level and to increase its operating limits are steadily expanding.

2. The computer-arithmetic basis of tabular technologies of modular information processing

At present, the modular direction in computer science, digital information processing and in many other fields of science and technology experiences a stage of rapid development directly linked with the outstanding achievements of integral technologies. During the past 15-20 years, the fundamental and applied researches both on the theory and the use of modular computing structures (MCS) have been focused on the implementation of the main advantages of MNS, first of all those which are caused by tabular nature of MA. In so doing, the following problems have received primary attention:

- the design of special purpose very-large-scale-integration (VLSI) chips, which take into account the nature of MNS as much as possible;
- the optimization of non-modular MCS;
- the numerous particular applications of MA.

Varied and comprehensive electronic devices and components including extensive classes of VLSI memories have been developed for the tabular MCS and special processors on their basis. There are numerous publications of well-known specialists in the field of MSC with the specific examples of modular processing units, processors and systems of tabular type. These examples shows clearly that a family of modular VLSI architectures and chips allows us to implement the majority of single-step computing procedures such as convolution and correlation of discrete sequences, adaptive finite impulse response (FIR) filters, mapping of discrete signals into the spaces of orthogonal projections and similar tasks. Within the framework of existing electronic components and circuits the ability of MNS-based devices to flexible modification of their own architecture is implemented simply by reprogramming of the used VLSI chips. At the same time, the design and implementation of reconfigurable high-performance modular processors and digital information-processing systems functioning entirely in the FCM usually requires too much hardware resources, and in the case of multi-step basic procedures it becomes impossible.

Because of the widespread distribution and rapid development of modern computers the software-based tabular MIPS presents an attractive competitive alternative to the traditional modular computing technologies oriented to the primary use of parallel pipeline VLSI architectures.

In terms of the basic optimality criteria of high-speed MCS a minimal redundant modular arithmetic (MRMA) is used as a computer-arithmetic basis of parallel tabular MIPS oriented to universal multi-processor computers. It is known that the most distinctive feature of an MRMA is the more complete and perfect computer arithmetic in comparison with the analogues

of classical MNS [7, 8 11-17]. The software-based versions of MRMA allow us to apply a huge dimension tables (up to 2^{32} words or more), very large modules (for example, from 2^{16} to 2^{64}) and therefore the super-large operating ranges. In the computer models of information processing procedures this circumstance makes it possible to expand significantly the FCM computational limits inherent to conventional modular systems based on VLSI architectures. In turn, this leads to the performance improvement as well as to the considerable increase in the accuracy of the final results.

As is easily seen, the perfect compliance of the MRMA and the tabular principles of the digital information processing appears most clearly in the software-based MCS. First of all, it is expressed in the triviality and flexibility of the programmable substitution mechanism of used operating tables including modification (increase or decrease) and even a complete replacement of the current set of the bases of minimal redundant MNS (MRMNS). Thus, the tabular minimal redundant MCS (MRMCS) and, therefore, the computer models of MIPS based on them have an extremely high level of adaptability and flexibility.

The design of the basic set of computer algorithms and their software models on the grounds of MRMA and modern computers is an inherent development stage of the family of tabular MRMCS, which create a computing environment with the maximum allowable limits of the FMC.

3. The basic model of the computing environment for the technology of modular information processing

The analysis of modern principles and methods used to design efficient algorithms for digital information processing allows us to conclude that usually all of them are aimed at implementing the decomposition concept of computation process organization.

Within the framework of this strategy it is expected the decomposition of the performed computational process (for example, the different discrete transforms, calculation of the convolution, correlation of discrete signals, etc.) into a set of similar procedures of the dimension smaller than the original one, and the construction of a corresponding rule for the formation of a net result by combining all the partial results of the executable procedures. Since most of the frequently used algorithms of digital information processing actually have the same operational structure, then their computer implementation fits into a framework of the common mathematical model.

The digital information processing systems, which target functions are described by the calculating relationships possessing the modular operational spectrum, represent the main application domain for tabular MIPS.

First, let us consider the principles of implementation of the non-recursive (one-step) processes in the FCM (for example, FIR filtering procedures). In this case, the basic elementary computational procedure fits into the framework of generalized computer model which can be described by the following expression

$$Y(l) = \left\rfloor \sum_{n=0}^{N-1} h_{l,n}^{(u)} X_l(n) \right\lfloor \ (l = 0, 1, \ldots, L-1;\ u = 0, 1, \ldots, U-1), \quad (1)$$

where $h_{l,n}^{(u)}$ is a certain real constant; $X_l(n)$ and $Y(l)$ are the integer samples of the input $\{X_l(0), X_l(1), \ldots, X_l(N-1)\}$ and output $\{Y(0), Y(1), \ldots, Y(L-1)\}$ digital signals, respectively; N, L and U are the natural numbers; $\rfloor x \lfloor$ designates the nearest integer to the real number x determined by the rule

$$\rfloor x \lfloor = \begin{cases} \lfloor x \rfloor & \text{if } x < \lfloor x \rfloor + 0,5; \\ \lceil x \rceil & \text{if } x \geq \lfloor x \rfloor + 0,5; \end{cases}$$

where $\lfloor x \rfloor = \max\{y \in \mathbf{Z} \mid y < x\}$, $\lceil x \rceil = \min\{y \in \mathbf{Z} \mid y \geq x\}$, $\mathbf{Z}$ is the set of all integers.

Since the expression (1) should be implemented using the MRMA, then it is necessary to reduce it to the form that is consistent with the principle of formal calculations. This can be made in two ways:

a) by the replacement of the coefficients $h_{l,n}^{(u)}$ by the rational fractional approximations $H_{l,n}^{(u)}/Q$, where $H_{l,n}^{(u)} = \left\rfloor Q\, h_{l,n}^{(u)} \right\lfloor \in \{-Q, -Q+1, \ldots, Q-1\}$, Q is the natural number which determines the accuracy of the approximation of real constants $h_{l,n}^{(u)} \in [-1, 1]$ by simple fractions $H_{l,n}^{(u)}/Q$;

b) by the approximation of the products $h_{l,n}^{(u)} X_l(n)$ by some fractional-rational estimations $Y_l^{(u)}(n)/S$, where

$$Y_l^{(u)}(n) = \left\rfloor S\, h_{l,n}^{(u)} X_l(n) \right\lfloor, \quad (2)$$

S is the selected natural scale. In the first case, the expression (1) is transformed into an approximate model

$$Y(l) \approx \left\rfloor Q^{-1} \sum_{n=0}^{N-1} H_{l,n}^{(u)} X_l(n) \right\lfloor \quad (3)$$

and into a model

$$Y(l) \approx \left\rfloor S^{-1} \sum_{n=0}^{N-1} Y_l^{(u)}(n) \right\lfloor \quad (4)$$

in the second case ($l = 0, 1, \ldots, L-1$; $u = 0, 1, \ldots, U-1$).

The implementation models (3) and (4) of the expression (1) are called the additive-multiplicative (AM) and additive (A) models, respectively. Thus, the generalized computer model is essentially reduced to computing the sum value

$$Y^{(u)}(l) = \begin{cases} \sum\limits_{n=0}^{N-1} H_{l,n}^{(u)} X_l(n) \ \text{ in the case of AM-model,} \\ \\ \sum\limits_{n=0}^{N-1} Y_l^{(u)}(n) \ \text{ in the case of A-model.} \end{cases} \tag{5}$$

The use of FCM is justified only when the computational results obtained on the selected modular segments of the realizable procedure do not exceed the working range of MRMNS. Thus, in order to implement the expression (5) entirely in the FCM it is required that the dynamic range $\mathbf{D} = \mathbf{Z}_{2M}^{-} = \{-M, -M+1, \ldots, M-1\}$ of the default MRMNS should include all the possible values of $Y^{(u)}(l)$ (here, $M = m_0 M_{k-1}$; m_0 is an auxiliary module satisfying the following conditions: $m_0 \geq \rho$ and $m_k \geq 2m_0 + \rho$ [7,8,11]; $M_{k-1} = \prod_{i=1}^{k-1} m_i$; $m_1, m_2, \ldots, m_k$ are the basic modules, $k \geq 2$). This circumstance exerts a decisive influence on the choice not only of the dynamic range $\mathbf{D}$, but also of the source data range $\hat{\mathbf{D}} = \mathbf{Z}_{2P}^{-} = \{-P, -P+1, \ldots, P-1\} \subset \mathbf{D}$ (P is some natural number).

It is assumed that the constant factors from the set $\{H_{l,0}^{(u)}, H_{l,1}^{(u)}, \ldots, H_{l,N-1}^{(u)}\}$ ($l = 0, 1, \ldots, L-1$; $u = 0, 1, \ldots, U-1$), that appeared in (3), are represented in the minimal redundant modular code (MRMC), i.e., by the vectors $H_{l,n}^{(u)} = \left(\left|H_{l,n}^{(u)}\right|_{m_1}, \left|H_{l,n}^{(u)}\right|_{m_2}, \ldots, \left|H_{l,n}^{(u)}\right|_{m_k}\right)$, and the values $X_l(n)$, which in the general case are represented by λ-bits complement binary numbers $X_l(n) = \left(x_l^{(\lambda-1)}(n)\ x_l^{(\lambda-2)}(n)\ \ldots\ x_l^{(0)}(n)\right)_2$ ($x_l^{(j)}(n) \in \{0,1\}$, $j = 0, 1, \ldots, \lambda-1$; $\lambda = 1 + \lceil \log_2 P \rceil$), are transformed into the MRMC ($|X_l(n)|_{m_1}$, $|X_l(n)|_{m_2}$, …, $|X_l(n)|_{m_k}$) during the implementation of the model (3).

With regard to the model (4), for its implementation it is necessary to obtain the MRMC ($\left|Y_l^{(u)}(n)\right|_{m_1}$, $\left|Y_l^{(u)}(n)\right|_{m_2}$, …, $\left|Y_l^{(u)}(n)\right|_{m_k}$) of numbers $Y_l^{(u)}(n)$ representing the scaled values of the numbers $X_l(n)$ (see (2)). This can be carried out using the methodology presented in [8].

Therefore, the residues $\left|Y^{(u)}(l)\right|_{m_i}$ ($i = 0, 1, \ldots, k$) of the MRMC of the number $Y^{(u)}(l)$ is calculated according to the rule

$$\left|Y^{(u)}(l)\right|_{m_i} = \begin{cases} \left|\sum\limits_{n=0}^{N-1} \left|\left|H_{l,n}^{(u)}\right|_{m_i} |X_l(n)|_{m_i}\right|_{m_i}\right|_{m_i} & \text{in the case of AM-model,} \\ \left|\sum\limits_{n=0}^{N-1} \left|Y_l^{(u)}(n)\right|_{m_i}\right|_{m_i} & \text{in the case of A-model} \end{cases} \tag{6}$$

for the implementation of the expression (5) in a FCM.

The tabular data

$$\mathbf{H}_u = \left\{ \left(\left|H_{l,n}^{(u)}\right|_{m_1}, \left|H_{l,n}^{(u)}\right|_{m_2}, \ldots, \left|H_{l,n}^{(u)}\right|_{m_k} \right) \right\}_{n=\overline{0,N-1};\, l=\overline{0,L-1}} \tag{7}$$

are the constituent elements of the computer AM-models (6) of the expression (1). The possibility of programmable partial modification or complete change of the set of MRMC (7) provides a simple adaptive reconfiguration of the computation process. In principle, the number U of the sets $\mathbf{H}_u$ can be an arbitrary number, and the total amount of default constants for tabular information processing is $V_T = N \cdot L \cdot U$ words of length $\lambda_{MNS} = \sum\limits_{i=1}^{k} \lceil \log_2 m_i \rceil$ bits.

Now let us estimate the cardinality $2M$ of the MRMS operating range $\mathbf{D}$ which ensures the correctness of the application of the expression (6).

According to (5), when using the AM-model we have

$$\left|Y^{(u)}(l)\right| \le \sum_{n=0}^{N-1} \left|H_{l,n}^{(u)} X_l(n)\right| \le P \sum_{n=0}^{N-1} \left|H_{l,n}^{(u)}\right|$$

Therefore, in this case the parameter M of the range $\mathbf{D}$ should satisfy the condition

$$M \ge P \max_{l,u} \left\{ \sum_{n=0}^{N-1} \left|H_{l,n}^{(u)}\right| \right\} \tag{8}$$

Without loss of generality, we can assume that $\left|h_{l,n}^{(u)}\right| \le 1$ for all admissible values of n, l and u. Then, in view of the fact that $H_{l,n}^{(u)} \in \{-Q, -Q+1, \ldots, Q-1\}$ and

$$\max_{l,u} \left\{ \sum_{n=0}^{N-1} \left|H_{l,n}^{(u)}\right| \right\} = N \cdot Q$$

we have from (8)

$$M \ge N \cdot P \cdot Q \tag{9}$$

In the case of applying the A-model to the implementation of the expression (1), the necessary estimates for the basic parameter M of the MRMNS range were obtained in [14]:

$$M > \left\lceil \frac{N \cdot S \cdot (2P+1)}{2} \right\rceil .$$

For most modern DSP applications the quite acceptable values of P (a parameter of the data range $\hat{\mathbf{D}}$) and N (a number of digital signal samples) are $2^{16} \leq P \leq 2^{24}$ and $2^8 \leq N \leq 2^{10}$, respectively. Since the calculation of $Y^{(u)}(l)$ (see (5)) in the MRMNS is carried out exactly, i.e. without rounding or truncation errors, then for approximation of the coefficients $H_{l,n}^{(u)}$ we can choose a natural number Q of the size, for example, between 12 and 16 bits, which corresponds to $2^{11} < Q < 2^{16}$. Taking the preceding into account, for the parameter M of the MRMNS operating range we obtain from (9) the following estimate: $2^{35} < M < 2^{50}$.

It is easy to see that implementing FCM within the framework of AM-model, an arbitrary recursive computational process (for example, the Winograd Fourier transform algorithm) represents a set of typical procedures (segments) of the form

$$Y_r^{(l)} = \sum_{n=0}^{N_r-1} C_{r,n,l} X_{r,l}(n) \;\; (r = 0, 1, \ldots, R-1;\; l = 0, 1, \ldots, L_r - 1), \quad (10)$$

where r is the specific number of process stage; R, L_r, N_r are some natural numbers; $C_{r,n,l}$ are the constants from the set $\{-Q, -Q+1, \ldots, Q-1\}$; $\{Y_r(l)\}_{l=\overline{0,L_r-1}}$ and $\{X_{r,l}(n)\}_{n=\overline{0,N_r-1}}$ represent the input and output signals of the rth standard segment of computational process under consideration, respectively. The input signal $\{X_{r,l}(n)\}_{n=\overline{0,N_r-1}}$ is formed from the samples of the input signal $\{X(n)\}_{n=\overline{0,N_r-1}}$ ($N = N_0$ is the signal length) in the case $r = 0$, and from the samples of the output signal $\{Y_{r-1}(l)\}_{l=\overline{0,L_{r-1}-1}}$ if $0 < r < R-1$. The signal $\{Y_{R-1}(l)\}_{l=\overline{0,L_{R-1}-1}}$ on the last stage represents the output signal $\{Y(l)\}_{l=\overline{0,L-1}}$ (L is a natural number) of the basic computational process. In this case, of course, the equality $L_{R-1} = L$ holds.

In the case when the modular computational process has a recursive organization the absolute values of the samples of the output signals of implementable basic procedures steadily increase together with an increase of the number of iterations (see (10)). This leads to the fact that even for a relatively small number of iterations R the cardinality $|\mathbf{D}| = 2M$ of the dynamic range $\mathbf{D}$ of an MRMNS, which ensures the application correctness of the principle of formal calculations, becomes a very large number as it

follows from the estimate of the parameter M:

$$M > P \cdot Q^R \cdot \prod_{r=0}^{R-1} N_r. \tag{11}$$

It should be noted that if the samples of the input signal $\{X(n)\}_{n=\overline{0,N-1}}$ and the constants $C_{r,n,l}$ appearing in (10) are integer complex numbers, then the right-hand side of the inequality (11) increases with an extra 2^R.

Therefore, the use of appropriate MRMNS for hardware implementation of recursive computational processes of the type (11) is associated with significant hardware costs and at present it is hardly advisable. At the same time, the tabular MIPS based on the universal processors allows us to realize in full measure the fundamental advantages of MRMA due to the possibility to use the extremely large dynamic ranges **D** and to ensure the maximum operation limits of the FCM.

4. Conclusions

With the wide distribution and rapid development of modern computing machinery, the considered approach to creating parallel digital information processing systems on the basis of MRMA is a very convenient and advanced alternative in comparison with other approaches based on the primary use of special purpose digital hardware.

Two types of computer MRMA-models of adaptive DSP procedures that fit into the so-called additive and additive-multiplicative formal computing schemes are presented in the article. The proposed models provide an exceptionally wide range of possibilities for flexible regulation of the working limits of the FCM depending on the current values of the accuracy and other parameters of the computational processes. This makes it possible to select optimally the bases of the required MRMNS for the preset parameters: the cardinality of the operating range, the number of bases, etc.

The possibility of using large dynamic ranges, when implementing the fast and high precision DSP-algorithms by means of the FCM in MRMA, allows us to remove from the general computational process the scaling operations following the standard (elementary) procedure of A- or AM-model. It provides:

- a significant performance increase due to a sharp reduction in the number of non-modular operations;
- the minimization of the upper threshold (and in many cases the total absence) of the computational error;
- the possibility of using a minimal basic set of non-modular operations, i.e. a set that includes only code transformations.

The inclusion of the complex, quadratic and polynomial-scalar variants of MRMA, which are oriented to extra-large dynamic ranges, into the computer-arithmetic base of the technology of tabular MIPS will increase significantly the computational speed (according to preliminary estimates, approximately 4-10 times) on the sets of complex numbers and polynomials. At the same time, the total amount of internal memory for storing the calculated tables is significantly reduced (at least by 40%) in comparison with non-redundant analogs of quadratic and polynomial-scalar MA.

References

[1] S.W. Smith, *Digital Signal Processing: A Practical Guide for Engineers and Scientists. 1st Edition*. Newnes, Burlington, MA, 2002.

[2] R.G. Lyons, D.L. Fugal, *The Essential Guide to Digital Signal Processing*. Prentice Hall Press, NJ, 2014.

[3] L. Deligiannidis, H. Arabnia, *Emerging Trends in Image Processing*. Prentice Hall Press, NJ, 2014.

[4] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer-Verlag, NJ, 2010, DOI 10.1007/978-1-84882-935-0.

[5] C.W. Ueberhuber, *Numerical Computation 1: Methods, Software, and Analysis*. Springer-Verlag, Berlin, 2013, DOI 10.1007/978-3-642-59118-1.

[6] C.W. Ueberhuber, *Numerical Computation 2: Methods, Software, and Analysis*. Springer-Verlag, Berlin, 2013.

[7] A.A. Kolyada, I.T. Pak, *Modular Structures of Pipeline Digital Information Processing*. University Press, Minsk, 1992 (in Russian).

[8] A.F. Chernyavsky, V.V. Danilevich, A.A. Kolyada, M.Y. Selyaninov, *High-speed Methods and Systems of Digital Information Processing*. Belarusian State University Press, Minsk, 1996 (in Russian).

[9] P.V. Ananda Mohan, *Residue Number Systems: Algorithms and Architectures*. Kluwer Academic Publishers, 2002.

[10] A. Omondi, B. Premkumar, *Residue Number Systems: Theory and Implementation*. Imperial College Press, London, 2007.

[11] M. Selyaninov, *Modular technique of parallel information processing*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XIII** (2008), 43–52.

[12] M. Selyaninov, *Construction of modular number system with arbitrary finite ranges*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XIV** (2009), 105–115.

[13] M. Selianinau, *High-speed modular structures for parallel computing in the space of orthogonal projections*. Scientific Issues, Jan Długosz University of Czestochowa, Ser. Technical and IT Education, **V** (2010), 87–96.

[14] M. Selianinau, *Modular principles of high-speed adaptive filtration of discrete signals*. Scientific Issues, Jan Długosz University of Czestochowa, Ser. Technical and IT Education, **VI** (2011), 75–84.

[15] M. Selyaninov, *Application of modular computing technique for high-speed implementation of cyclic convolution*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XIX** (2014), 217–226.

[16] M. Selyaninov, *Modular number systems in a complex plane*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XV** (2010), 131–138.

[17] M. Selyaninov, *Arithmetic of quadratic minimal redundant modular number systems*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XVI** (2011), 129–134.

[18] M. Selyaninov, *Modular technique of high-speed parallel computing on the sets of polynomials*. Scientific Issues of Jan Długosz University of Czestochowa, Mathematics **XVII** (2012), 69–76.

Received: May 2017

Mikhail Selianinau
Jan Długosz University,
Institute of Technics and Safety Systems,
42-200 Częstochowa, Al. Armii Krajowej 13/15, Poland
E-mail address: m.selianinov@ajd.czest.pl

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
141–160
DOI http://dx.doi.org/10.16926/m.2017.22.11

EXPECTED VOLUMES OF REQUESTS IN SYSTEMS OF THE QUEUEING NETWORK WITH A LIMITED NUMBER OF WAITING PLACES

PAWEŁ ZAJĄC AND MIKHAIL MATALYTSKI

Abstract

We present a method of finding the expected volume of requests in HM-network with homogeneous requests and bypass of the queueing systems of requests. The case was considered when the volume changes associated with the transitions between the states of the network are deterministic functions, depending on the state of the network and time, and the systems are single line. It is assumed that the probability of the states of the network systems, the parameters of the entrance flow of the requests and the service depend on the time.

Keywords: *HM-networks, queueing network, single-line queueing system, queueing system, queuing time, limited queue, demands total volume, volume of requests, capacity of claims, wireless access point*

1. Introduction

In the information system (IS) the total number of memory volume is bounded by some value, which is usually called memory volume [1]. In the IS designing the main task is the determination of the memory volume so as to take into account the conditions that limit the proportion of the lost information. One of the methods for solving problems of IS design is the use of HM-queueing networks [2]. Further under the IS we will understand systems, converting objects which is the information coming in portions in the form of messages [1]. HM-networks can be used to determine the volume of a buffer storage of systems, which are representing processing IS nodes and transferring messages. Note that considered problem is one of

• *Paweł Zając* — e-mail: pawel_zajac@vp.pl
Czestochowa University of Technology.
• *Mikhail Matalytski* — e-mail: m.matalytski@gmail.com
Czestochowa University of Technology.

the main, for example, in the design of communication centers or hubs in data communication networks.

Neglect time dependence of messages processing of their volumes can lead to serious errors in finding the buffer memory in IS and in calculating of the probability of messages loss. The solution in the general case the above problems can be based on the use of HM-networks with revenues. In such networks, the request during the transition from one queueing system (QS) to another brings some revenue last (which is equal to the volume of this request), and revenue (volume) of the first QS is reduced by this number.

QS with revenues in the stationary regime has been introduced in the consideration in [3], and networks - in [4]. A survey results obtained by queueing systems and queueing networks (QN) in the stationary regime contained in [5]. It is dedicated to finding the mean revenues in the network systems which depend only on their states and do not depend on time, and solving the problem of finding the optimal request service intensities in QS by the method of dynamic programming. QN with revenues in the non-stationary regime has been studied in [6, 7]. Revenues from transitions between network states were depended on the states and their time or were random variables with the given moments of the first and second orders. In a survey article [7] the results of researching, optimization and selection of the optimal strategies in Markov networks with revenues, various applications are described them as a probabilistic models for forecasting expected revenues in the information and telecommunication systems and networks, when, for example, requests service on the server generates revenue for a servicer, as well as insurance companies, logistics transportation systems, industrial systems and other facilities. For the first time application of the HM-networks for estimating the memory volume in the IS has been described in [8].

2. Computer network model with limited number of simultaneous connections

Consider the action model a wireless network. Let us assume that the network is a set of wireless access points S_i each of which gives the user the ability to connect to the network of information by the next available port and use its resources, $i = \overline{1,n}$. Each access point can simultaneously connect multiple users to a network or execute queries users (i.e. for users to make bandwidth sharing for m_i smaller teams). All other requests users do not receive service, create maximum queue requests L_i and are waiting for a reply access point $L_i < \infty$, $i = \overline{1,n}$. From a technical point of view, the value of L_i is determined taking into account the technical characteristics of a wireless access point (WAP). For example, the greater the number of

divided bands of the original panel is more customers may be combined, and the slower the speed of the bands assigned to them. The main limitation is the number of possible IP addresses that are distributed to connect.

We also assume that in the absence free places in the queue, to the access point comes the loss of user requests and redirecting this question to the next WAP. From a practical point of view, redirecting associated with direct customer flow (ie, the device working with computer network), or search for «distant» WAPs do not go beyond the radius of visibility of the device.

At request, we assume a data packet, the sending by the source (ie. client device) and tending to the recipient (by WAP). In computer networks, the package is designed in a way, a block of data sent over the network in batch mode. Computer lines, which does not support packet-mode such as traditional telecommunication point-to-point transfer data simply as a sequence of bytes, characters, or bits individually. The package consists of two types of data: the control information and user data (also called a payload). The control information contains information necessary for the provision of user data: source and destination addresses, error detection codes (such as checksums) and information about the order. Typically, the control information contained in the packet header and tail, and the user data disposed therebetween.

Various communication protocols use different conventions for separating elements and formatting data. Protocol packets "Synchronous binary" formats in 8-bit bytes, and for the separation of the used special characters. In other protocols, such as Ethernet header and the beginning of data elements, their location in relation to the package they are registered. Some protocols format the information at the level of bits and not bytes. In this case, it is assumed that each data packet sent by a user will be deterministic or random length (volume).

Many networks cannot guarantee delivery, no duplicates packages, and order delivery, such as the UDP protocol on the Internet. However, this can be done on top of a transport packet (for one level of the OSI model), which can provide such protection. Packet header identifies the type of data packets, the package, the total number of packets and IP address of the source and destination. In our case, it should be assumed that the request sent by the user will always be delivered to the recipient (AP), which will not be taken into account in case of packet loss between the source and the destination.

Estimation of the total volume of data (data packets) for each wireless access point (WAP) at the given point in time is an important task when designing a wireless network, because it allows to locate the highly loaded AP and distribute the load evenly over them.

Therefore, you must determine the average total volume of data (packets of users) received by the access points to the network information (for example, Internet), taking into account the limited number of «simultaneous» connections on this point. This problem can be solved by using HM-network storage service. According to the queueing network we mean a collection of interconnected queueing systems S_i with a limited buffer size L_i to hold the queue (queues) package requests. By application we mean the user's request to the WAP, which is a data packet.

3. Analysis of queueing network with bypass of queueing systems and time depending parameters of entrance flow and service time

Consider the open exponential queueing network comprising n queueing systems. With some probability requests from have a chance to join the queue, or instantly pass in the matrix of transition probabilities to the other queueing system or leave the network. The probability of attachment to the queueing systems depend on the state and number of queueing system from which the notification is sent.

The message is sent to the network at a given time interval $[t, t + \Delta t]$ with probability $\lambda(t)\Delta t + o(\Delta t)$ and is supported in queueing system S_i in this interval with a probability of $\mu_i(t)\Delta t + o(\Delta t)$.

Let p_{ij} - the probability of transition of requests after servicing from the system S_i to S_j, $i, j = \overline{0, n}$, system S_0 we also understand the external environment. Now consider the case where the stream parameters and service depend on the time. Notification is sent from the external environment in the i-th SOM with the probability of p_{0i}, $\sum_{i=0}^{n} p_{0i} = 1$. A request to the queuing system from outside or from the second system at time t with probability $f^{(i)}(k, t)$, when the network is in state of (k, t), connects to the queue, and the probability of $1 - f^{(i)}(k, t)$ not join the requests queue, counts on service (i.e. the time service with a probability of 1 is equal to zero). If the request ended service in the i-th queueing system, then with the probability of p_{ij} is immediately sent to the j-th system and with the probability p_{i0}leaves the queueing network, $\sum_{j=0}^{n} p_{0i} = 1$, $i = \overline{1, n}$.

State of the network is described vector $k(t) = (k, t) = (k_1(t), \ldots, k_n(t)) = (k_1, \ldots, k_n, t)$, where $k_i(t)$ the number of requests in the system S_i (in a queue and handling), $i = \overline{1, n}$. Let $\varphi_i(k, t)$ - conditional probability that the request comes from outside and to the i-th queueing system at time t, when the network is in state (k, t) cannot be handled by queueing system; $\psi_{ij}(k, t))$ - conditional probability that request by introducing point i-th system the outside, at time t, when the network is in state (k, t), for the

first time will receive service in the j-th system; $\alpha_i(k,t)$ - conditional probability that requests serviced at time t in the i-th queueing system, when the network is to state (k,t), will not be supported further in either a single system; $\beta_{ij}(k,t)$ - conditional probability that the request serviced in the i-th system at time t, when the network is to state (k,t), for the first time will be supported in the j-th queueing system, $i,j=\overline{1,n}$.

The formula for total probability we get:

$$\varphi_i(k,t)=\left(1-f^{(i)}(k,t)\right)\left(p_{i0}+\sum_{j=1}^{n}p_{ij}\varphi_j(k,t)\right),i=\overline{1,n},$$

$$\psi_{ij}=f^{(i)}(k,t)\delta_{ij}+\left(1-f^{(i)}(k,t)\right)\sum_{l=1}^{n}p_{il}\psi_{lj}(k,t),i=\overline{1,n}.$$

If the function $f^{(i)}(k,t)=f^{(i)}(k_i,t)$ depends only on the number of requests, then

$$\varphi_i(k,t)=\left(1-f^{(i)}(k,t)\right)\left(p_{i0}+\sum_{j=1}^{n}p_{ij}\varphi_j(k,t)\right),i=\overline{1,n}, \tag{1}$$

$$\psi_{ij}=f^{(i)}(k_i,t)\delta_{ij}+\left(1-f^{(i)}(k,t)\right)\sum_{l=1}^{n}p_{il}\psi_{lj}(k,t),i,j=\overline{1,n}. \tag{2}$$

Furthermore,

$$\alpha_i(k,t)=p_{i0}+\sum_{j=1}^{n}p_{ij}\varphi_j(k-I_i,t),$$

$$\beta_{ij}(k,t)=\sum_{l=1}^{n}p_{il}\psi_{lj}(k-I_i,t),i,j=\overline{1,n}. \tag{3}$$

We also have equality

$$\varphi_i(k,t)+\sum_{j=1}^{n}\psi_{ij}(k,t)=1,\ i,j=\overline{1,n} \tag{4}$$

$$\alpha_i(k,t)=1-\sum_{j=1}^{n}\beta_{ij}(k,t)=1,\ i,j=\overline{1,n}. \tag{5}$$

From (1) and (3) we find

$$\psi_{ij}(k,t)=f^{(i)}(k,t)\delta_{ij}+\left(1-f^{(i)}(k,t)\right)\beta_{ij}(k-I_i,t),\ i,j=\overline{1,n}. \tag{6}$$

4. The system of difference-differential equations for the expected volume of requests in the HM-systems network

Let $v_i(k,t)$ - expected volume of requests, which accumulate in the system S_i at time t, when, during the initial network is to state k, and assume that the function is differentiable with respect to t; $r_i(k)$ - increasing the volume of requests in the system S_i per unit time when the network is to state k; $r_{0i}(k+I_i,t)$ - volume of requests, which increases the total volume of requests in the system S_i, when the network passes from the state of (k,t) in to state $(k+I_i,t+\Delta t)$ in time Δt; $-R_{i0}(k-I_i,t)$ - reducing the volume size of the system, if the network causes a transition from state (k,t) in to state $(k+I_i,t+\Delta t)$; $r_{ij}(k+I_i-I_j,t)$ - the volume of requests of the system S_i (reducing the volume of requests of the system S_j), to which increases the total volume of requests, the network changes its state from (k,t) to $(k+I_i-I_j,t+\Delta t)$ in time Δt, $i,j=\overline{1,n}$. Note that we consider now the case when the value of r_{0i}, R_{i0}, r_{ij} are deterministic functions dependent states of network and time.

Assume that the network is in a state (k,t). At time Δt it may be in a state k or go to the states $(k-I_i)$, $(k+I_i)$, $(k+I_i-I_j)$, $i,j=\overline{1,n}$. If the network is still in the state $(k,t+\Delta t)$, volume of requests in the system S_i is $r_i(k)\Delta t$ plus volume $v_i(k,t)$, to which increases its volume in the remaining t time units. The probability of this happening is $1-\sum_{i=1}^{n}\left(\lambda(t)p_{0i}\left(1-\varphi_i(k,t)\right)+\mu_i(t)\left(1-\beta_{ii}(k,t)\right)\right)\Delta t+o(\Delta t)$. If the network passes to the state of $k+I_i,t+\Delta t)$ with probability $\lambda(t)p_{0i}\psi_{ij}(k+I_i,t)\Delta t+o(\Delta t)$, then the total volume of requests in the system S_i is $[r_{0i}(k+I_i,t)+v_i(k+I_i,t)]$ and if the state $(k-I_i,t+\Delta t)$ with probability $\mu_i(k,t)\alpha_i(k-I_i,t)u(k_i,t)\Delta t+o(\Delta t)$, then the total volume of requests of this system will be $[-R_{i0}(k-I_i,t)+v_i(k-I_i,t)]$, $i=\overline{1,n}$.

Similarly, if the network goes from (k,t) in to a state $(k+I_i-I_j,t+\Delta t)$ with probability $\mu_j(k,t)\beta_{ji}(k+I_i-I_j,t)u(k_j,t)\Delta t+o(\Delta t)$, then it will increase the total volume of requests in the system S_i by the amount $r_{ij}(k+I_i-I_j,t)$ plus the volume of the remaining time, if the initial state of the network was $(k+I_i-I_j$. Then, using the formula of the total probability for the expected volume of requests in the system, you can get a system of difference-differential equations (DDE) [9].

$$\frac{\mathrm{d}\nu_i(k,t)}{\mathrm{d}t} = r_i(k) - \sum_{i=1}^{n} \left[\lambda(y)p_{0i}\left(1-\varphi_i(k,t)\right) + \mu_i(t)\left(1-\beta_{ii}(k,t)\right)\right]\nu_i(k,t) + \tag{7}$$

$$+\sum_{j=1}^{n}[\lambda(t)p_{0j}\psi_{ij}(k+I_j,t)\nu_i(k+I_j,t) + \tag{8}$$

$$+\mu_j(t)\alpha_j(k-I_j,t)u(k_j(t))\nu_i(k-I_j,t)] + \tag{9}$$

$$+\sum_{\substack{j=1\\ j\neq i}}^{n}[\mu_j(t)\beta_{ji}(k+I_i-I_j,t)u(k_j(t))\nu_i(k+I_i-I_j,t) + \tag{10}$$

$$+\mu_i(t)\beta_{ij}(k-I_i+I_j,t)u(k_i(t))\nu_i(k-I_i+I_j,t)] + \tag{11}$$

$$+\sum_{\substack{j=1\\ j\neq i}}^{n}[u_j(t)\beta_{ji}(k+I_i-I_j,t)u(k_j(t))r_{ij}(k+I_i-I_j,t) - \tag{12}$$

$$-u_i(t)\beta_{ij}(k-I_i+I_j,t)u(k_i(t))r_{ji}(k-I_i+I_j,t)] + \tag{13}$$

$$+\sum_{\substack{c,s=1\\ c,s\neq i}}^{n}\mu_s(t)\beta_{sc}(k+I_c-I_s,t)u(k_s(t))\nu_i(k+I_c-I_s,t) + \tag{14}$$

$$+\lambda(t)p_{0i}\psi_{ij}(k+I_i,t)r_{0i}(k+I_i,t) - \tag{15}$$

$$-\mu_i(t)\alpha_i(k-I_i,t)u(k_i(t))R_{i0}(k-I_i,t). \tag{16}$$

Expressions for conditional probabilities $\varphi_i(k,t)$, $\psi_{ij}(k,t)$, $\alpha_i(k,t)$, $\beta_{ij}(k,t)$, $i,j=\overline{1,n}$, originate from (1)-(5).

For a closed network, ie. if $\lambda(t) = 0$, $p_{0i} = p_{i0} = 0$, $r_{0i}(k_i, t) = R_{i0}(k_i, t) = 0$, $\sum_{i=1}^{n} k_i(t) = K$, $i = \overline{1,n}$, then we have from (7)

$$
\begin{aligned}
(17)\quad \frac{\mathrm{d}\nu_i(k,t)}{\mathrm{d}t} &= r_i(k) - \sum_{i=1}^{n} \mu_i(t)\left(1 - \beta_{ii}(k_i,t)\right)\nu_i(k,t) + \\
&+ \sum_{i=1}^{n} \mu_j(t)\alpha_j(k - I_j, t)u(k_j(t))\nu_i(k - I_j, t) + \\
&+ \sum_{\substack{j=1 \\ j\neq i}}^{n} \mu_j(t)\beta_{ji}(k + I_i - I_j, t)u(k_j(t))\nu_i(k + I_i - I_j, t) + \\
&+ \mu_i(t)\beta_{ij}(k - I_i + I_j, t)u(k_i(t))\nu_i(k - I_i + I_j, t) + \\
&+ \sum_{\substack{j=1 \\ j\neq i}}^{n} [\mu_j(t)\beta_{ji}(k + I_i - I_j, t)u(k_j(t))r_{ij}(k + I_i - I_j, t) - \\
&- \mu_i(t)\beta_{ij}(k - I_i + I_j, t)u(k_i(t))r_{ji}(k - I_i + I_j, t)] + \\
&+ \sum_{\substack{c,s=1 \\ c,s\neq i}}^{n} \mu_s(t)\beta_{sc}(k + I_c - I_s, t)u(k_s(t))\nu_i(k + I_c - I_s, t), i = \overline{1,n}.
\end{aligned}
$$

5. Solution of DDE for the closed network with a limited number of spaces of expectations of requests in queueing systems

Consider the closed queueing network with a limited length of the request for the system S_i, equal L_i, $i = \overline{1,n}$. It is assumed that the notification coming in the system S_i to service, takes place in the queue, $i = \overline{1,n}$. If on arrival request into the system S_i the number of requests is less than L_i , then the incoming request takes place in the queue, otherwise it is immediately operated and goes according to the matrix pass $P = \|p_{ij}\|_{n\times n}$ by other system:

$$
(18)\qquad f_{k_i}^{(i)}(t) = \begin{cases} 1 & \text{if} \quad 0 \le k_i(t) < L_i, \\ 0 & \text{if} \quad k_i(t) \ge L_i. \end{cases}
$$

For the closed network $\psi_{ij}(k,t)$ - conditional probability that the request occupy the i-th system in to a time t , where network is in the state (k,t) immediately after the time of its transfer to the system (Ie, without the transfer of the last report), then will operate in j-th system $\left(\sum_{i=1}^{n} k_i(t) = K - 1, i, j = \overline{1,n}\right)$; $\beta_{ij}(k,t)$ - conditional probability that the

request handled in i -th system at the time of t , where network is in the state of (k,t) immediately before the end of its operation, will first serve in j-th queueing system, $\sum_{i=1}^{n} k_i(t) = K, i, j = \overline{1,n}$.

According to the formula for the probability of complete:

$$\psi_{ij}(k,t) = f^{(i)}(k,t)\delta_{ij} + \left(1 - f^{(i)}(k,t)\sum_{i=1}^{n} p_{il}\psi_{lj}(k,t)\right), \quad \sum_{i=1}^{n} k_i(t) = K-1, j = \overline{1,n}, \tag{19}$$

$$\beta_{ij}(k,t) = \sum_{l=1}^{n} p_{il}\psi_{lj}(k-I_i,t), \sum_{i=1}^{n} k_i(t) = K, j = \overline{1,n}. \tag{20}$$

So

$$\sum_{j=1}^{n} \psi_{ij}(k,t) = 1, \quad \sum_{i=1}^{n} k_i(t) = K-1, \tag{21}$$

$$\beta_{ij}(k,t) = 1, \quad \sum_{i=1}^{n} k_i(t) = K. \tag{22}$$

In a closed network must be: if $L_i < K$, $i = \overline{1,n}$, then $\sum_{i=1}^{n} L_i \geq K$ - the number of requests in the network should not be larger than the maximum number of requests that can be accepted by all queueing systems [Gordon-Newell].

Example 1. *Consider a closed network* $n = 3$, $m_i = 1$, $i = \overline{1,3}$, $L_1 = 1$, $L_2 = 2$, $L_3 = 2$, $K = 6$. *Let's assume that* $f^{(i)}(k,t) = f^{(i)}(k_i,t)$. *The number of states in this network is equal 6 which have the form* $\{(0,3,3), (1,2,3), (1,3,2), (2,1,3), (2,2,2), (2,3,1)\}$.

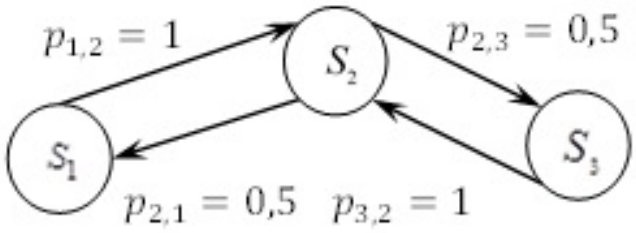


FIGURE 1. Closed queueing network

From the formulas (19), in the case of $k_1(t) + k_2(t) + k_3(t) = K - 1$, we have a system of equations:

(23)
$$\begin{cases} \psi_{11}(k,t) = f^{(1)}(k_1,t) + \left(1 - f^{(1)}(k_1,t)\right)\psi_{21}(k,t), \\ \psi_{12}(k,t) = \left(1 - f^{(1)}(k_1,t)\right)\psi_{22}(k,t), \\ \psi_{13}(k,t) = \left(1 - f^{(1)}(k_1,t)\right)\psi_{23}(k,t), \\ \psi_{21}(k,t) = \left(1 - f^{(2)}(k_2,t)\right)(0,5\psi_{11}(k,t) + 0,5\psi_{31}(k,t)), \\ \psi_{22}(k,t) = f^{(2)}(k_2,t) + \left(1 - f^{(2)}(k_2,t)\right)(0,5\psi_{12}(k,t) + 0,5\psi_{32}(k,t)), \\ \psi_{23}(k,t) = \left(1 - f^{(2)}(k_2,t)\right) + (0,5\psi_{13}(k,t) + 0,5\psi_{33}(k,t)), \\ \psi_{31}(k,t) = \left(1 - f^{(3)}(k_3,t)\right)\psi_{21}(k,t), \\ \psi_{32}(k,t) = \left(1 - f^{(3)}(k_3,t)\right)\psi_{22}(k,t), \\ \psi_{33}(k,t) = f^{(3)}(k_3,t) + \left(1 - f^{(3)}(k_3,t)\right)\psi_{23}(k,t). \end{cases}$$

The system of equations (23) with probabilities $\psi_{11}(k,t)$, $\psi_{12}(k,t)$, $\psi_{13}(k,t)$, $\psi_{21}(k,t)$, $\psi_{22}(k,t)$, $\psi_{23}(k,t)$, $\psi_{31}(k,t)$, $\psi_{32}(k,t)$, $\psi_{33}(k,t)$ can be solved using the Mathematica 8 package. Then we get:

(24)
$$\psi_{11}(k,t) = \\ = \frac{f^{(1)}(k_1,t)\left(0,5 + f^{(2)}(k_2,t)\left(0,5 - 0,5f^{(3)}(k_3,t)\right)\right) + 0,5f^{(3)}(k_3,t)}{0,5f^{(1)}(k_1,t) + f^{(2)}(k_2,t) - 0,5 \cdot f^{(1)}(k_1,t)f^{(2)}(k_2,t) + 0,5f^{(3)}(k_3,t) - 0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)},$$

(25)
$$\psi_{12}(k,t) = \\ = \frac{\left(1 + f^{(1)}(k_1,t)\right)f^{(2)}(k_2,t)}{0,5f^{(1)}(k_1,t) - f^{(2)}(k_2,t) + 0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t) - 0,5f^{(3)}(k_3,t) + 0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)},$$

$$\psi_{13}(k,t) =$$

$$= \frac{\left(0,5f^{(1)}(k_1,t)\left(0,5-0,5f^{(2)}(k_2,t)\right)+0,5f^{(2)}(k_2,t)\right)f^{(3)}(k_3,t)}{1-0,5f^{(1)}(k_1,t)-f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)}+$$

$$+\frac{\left(1+f^{(1)}(k_1,t)\right)f^{(2)}(k_2,t)}{\left(0,5f^{(1)}(k_1,t)+f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)}\times \tag{26}$$

$$\times\frac{1}{\left(\left(1-0,5f^{(1)}(k_1,t)\right)+f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)},$$

$$\psi_{21}(k,t) =$$

$$+\frac{\left(1+f^{(1)}(k_1,t)\right)f^{(2)}(k_2,t)}{\left(0,5f^{(1)}(k_1,t)+f^{(2)}(k_2,t)-0,5f^{(1)}(k_1,t)\right)f^{(2)}(k_2,t)+0,5f^{(3)}(k_3,t)-0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)}\times \tag{27}$$

$$\times\frac{1}{\left(1-f^{(1)}(k_1,t)\right)},$$

$$\psi_{22}(k,t) =$$

$$\frac{f^{(2)}(k_2,t)}{0,5f^{(1)}(k_1,t)+f^{(2)}(k_2,t)-0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)+0,5f^{(3)}(k_3,t)-0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)}, \tag{28}$$

$$\psi_{23}(k,t) =$$

$$\frac{f^{(2)}(k_2,t)+0,25f^{(1)}(k_1,t)f^{(3)}(k_3,t)+\left(0,5+f^{(1)}(k_1,t)\left(0,5-0,25f^{(2)}(k_2,t)-0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)\right)}{\left(\left(0,5f^{(1)}(k_1,t)-f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)\right.}\times \tag{29}$$

$$\times\frac{1}{\left(1-0,5f^{(1)}(k_1,t)-f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)},$$

$$\psi_{31}(k,t) =$$

$$\frac{0,5f^{(1)}(k_1,t)\left(1+f^{(2)}(k_2,t)\left(1-f^{(3)}(k_3,t)\right)+f^{(1)}(k_1,t)\left(1-f^{(2)}(k_2,t)\left(1-f^{(3)}(k_3,t)\right)+f^{(3)}(k_3,t)\right)+f^{(3)}(k_3,t)\right)}{\left(1-0,5f^{(1)}(k_1,t)\right)\left(0,5f^{(1)}(k_1,t)+f^{(2)}(k_2,t)-0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)+0,5f^{(3)}(k_3,t)-0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)\right)}, \tag{30}$$

(31)
$$\psi_{32}(k,t) = \frac{f^{(2)}(k_2,t)\left(1-f^{(3)}(k_3,t)\right)}{0,5f^{(1)}(k_1,t)-f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)},$$

(32)
$$\psi_{33}(k,t) = \frac{0,5\left(1+f^{(1)}(k_1,t)\left(1-f^{(2)}(k_2,t)\right)-0,5f^{(2)}(k_2,t)\right)f^{(3)}(k_3,t)}{1-0,5f^{(1)}(k_1,t)-f^{(2)}(k_2,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2,t)f^{(3)}(k_3,t)}.$$

For queueing network, according to formulas (24)-(32), conditional probabilities $\psi_{ij}(k,t)$ take the form:

$$\begin{aligned}
&\psi_{11}(1,1,3,t)=0, && \psi_{31}(1,1,3,t)=0,\\
&\psi_{11}(2,0,3,t)=0, && \psi_{31}(2,0,3,t)=0,\\
&\psi_{11}(0,2,3,t)=0,25, && \psi_{31}(0,2,3,t)=0,125,\\
&\psi_{11}(2,1,2,t)=0, && \psi_{31}(2,1,2,t)=0,\\
&\psi_{11}(1,2,2,t)=0,25, && \psi_{31}(1,2,2,t)=0,\\
&\psi_{13}(1,1,3,t)=0,75, && \psi_{33}(1,1,3,t)=0,\\
&\psi_{13}(2,0,3,t)=0,5, && \psi_{33}(2,0,3,t)=0,\\
&\psi_{13}(2,1,2,t)=0,5, && \psi_{33}(2,1,2,t)=0,125,\\
&\psi_{13}(2,2,1,t)=0, && \psi_{33}(2,2,1,t)=0,125,\\
&\psi_{22}(0,2,3,t)=0, && \psi_{22}(0,3,2,t)=0,\\
&\psi_{23}(1,2,2,t)=0, && \psi_{23}(1,3,1,t)=0,\\
&\psi_{21}(2,2,1,t)=0, && \psi_{22}(2,3,0,t)=0.
\end{aligned}$$

If $k_1(t)+k_2(t)+k_3(t)=K$, then from pattern (20) we will have:

$$\beta_{11}(k,t) = \psi_{21}(k - I_1, t) =$$
$$= \frac{f^{(1)}(k_1 - 1, t)\left(0,5 + f^{(1)}(k_1 - 1, t)\left(0,5 - 0,5f^{(2)}(k_2, t)\right)\right) - 0,5f^{(2)}(k_2, t)}{\left(1 - f^{(1)}(k_1 - 1, t)\right)\left(0,5f^{(1)}(k_1 - 1, t) + f^{(2)}(k_2, t) - 0,5f^{(1)}(k_1 - 1, t)\right)f^{(2)}(k_2, t) + f^{(3)}(k_3, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3, t)},$$

$$\beta_{12}(k,t) = \psi_{22}(k - I_1, t) =$$
$$= \frac{f^{(2)}(k_2, t)}{0,5f^{(1)}(k_1 - 1, t) + f^{(2)}(k_2, t) - 0,5f^{(1)}(k_1 - 1, t)f^{(2)}(k_2, t) + 0,5f^{(3)}(k_3, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3, t)},$$

$$\beta_{13}(k,t) = \psi_{23}(k - I_1, t) =$$
$$= \frac{f^{(2)}(k_2, t) + 0,25f^{(1)}(k_1 - 1, t)f^{(3)}(k_3, t) + 0,5 + f^{(1)}(k_1 - 1, t)\left(0,5 - 0,25f^{(2)}(k_2, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3, t)\right)}{\left(0,5f^{(1)}(k_1 - 1, t) - f^{(2)}(k_2, t) + 0,5f^{(1)}(k_1 - 1, t)f^{(2)}(k_2, t) - f^{(3)}(k_3, t) + 0,5f^{(2)}(k_2, t)f^{(3)}(k_3, t)\right)} \times$$
$$\times \frac{1}{\left(1 - 0,5f^{(1)}(k_1 - 1, t) - f^{(2)}(k_2, t) + 0,5f^{(1)}(k_1 - 1, t)f^{(2)}(k_2, t) - 0,5f^{(3)}(k_3, t) + 0,5f^{(2)}(k_2, t)f^{(3)}(k_3, t)\right)},$$

$$\beta_{31}(k,t) = \psi_{21}(k - I_3, t) =$$
$$= \frac{0,5\left(f^{(1)}(k_1 - 1, t)\left(1 + f^{(1)}(k_1, t)\left(1 - f^{(2)}(k_2, t)\right)\right) - f^{(2)}(k_2, t)\right)}{\left(1 - f^{(1)}(k_1, t)\right)\left(0,5f^{(1)}(k_1 - 1, t) + f^{(2)}(k_2, t) - 0,5f^{(1)}(k_1, t)\right)f^{(2)}(k_2, t) + 0,5f^{(3)}(k_3 - 1, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3 - 1, t)},$$

$$\beta_{32}(k,t) = \psi_{22}(k - I_3, t) =$$
$$= \frac{f^{(2)}(k_2, t)}{0,5f^{(1)}(k_1, t) + f^{(2)}(k_2, t) - 0,5f^{(1)}(k_1, t)f^{(2)}(k_2, t) + 0,5f^{(3)}(k_3 - 1, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3 - 1, t)},$$

$$\beta_{33}(k,t) = \psi_{23}(k - I_3, t) =$$
$$= \frac{f^{(2)}(k_2, t) + 0,25f^{(1)}(k_1, t)f^{(3)}(k_3 - 1, t) + 0,5 + f^{(1)}(k_1, t)\left(0,5 - 0,25f^{(2)}(k_2, t) - 0,5f^{(2)}(k_2, t)f^{(3)}(k_3 - 1, t)\right)}{\left(0,5f^{(1)}(k_1, t) - f^{(2)}(k_2, t) + 0,5f^{(1)}(k_1, t)f^{(2)}(k_2, t) - 0,5f^{(3)}(k_3 - 1, t) + 0,5f^{(2)}(k_2, t)f^{(3)}(k_3 - 1, t)\right)} \times$$
$$\times \frac{1}{\left(1 - 0,5f^{(1)}(k_1, t) - f^{(2)}(k_2, t) + 0,5f^{(1)}(k_1, t)f^{(2)}(k_2, t) - 0,5f^{(3)}(k_3 - 1, t) + 0,5f^{(2)}(k_2, t)f^{(3)}(k_3 - 1, t)\right)},$$

$$\beta_{21}(k,t) = 0,5\psi_{11}(k-I_2,t) + 0,5\psi_{31}(k-I_2,t) =$$

$$= 0,5\left(\frac{f^{(1)}(k_1,t)\left(0,5+f^{(2)}(k_2-1,t)\left(0,5-0,5f^{(3)}(k_3,t)\right)\right)+0,5f^{(3)}(k_3,t)}{0,5f^{(1)}(k_1,t)+f^{(2)}(k_2,t)-0,5f^{(1)}(k_1,t)f^{(2)}(k_2,t)+0,5f^{(3)}(k_3-1,t)-0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)}\right)+$$

$$+0,5\left(\frac{0,5f^{(1)}(k_1,t)\left(1+f^{(2)}(k_2-1,t)\left(1-f^{(3)}(k_3,t)\right)+f^{(1)}(k_1,t)\left(1-f^{(2)}(k_2-1,t)\left(1-f^{(3)}(k_3,t)\right)+f^{(3)}(k_3,t)\right)+f^{(3)}(k_3,t)\right)}{\left(1-f^{(1)}(k_1,t)\right)\left(0,5f^{(1)}(k_1,t)+f^{(2)}(k_2-1,t)-0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)+0,5f^{(3)}(k_3,t)-0,5f^{(2)}(k_2-1,t)\right)f^{(3)}(k_3,t)}\right),$$

$$\beta_{22}(k,t) = 0,5\psi_{12}(k-I_2,t) + 0,5\psi_{32}(k-I_2,t) =$$

$$= 0,5\left(\frac{\left(1+f^{(1)}(k_1,t)\right)f^{(2)}(k_2-1,t)}{0,5f^{(1)}(k_1,t)-f^{(2)}(k_2-1,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)}\right)+$$

$$+\left(\frac{f^{(2)}(k_2-1,t)\left(1-f^{(3)}(k_3,t)\right)}{0,5f^{(1)}(k_1,t)-f^{(2)}(k_2-1,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)}\right),$$

$$\beta_{23}(k,t) = 0,5\psi_{13}(k-I_2,t) + 0,5\psi_{33}(k-I_2,t) =$$

$$= 0,5\left(\frac{\left(0,5+f^{(1)}(k_1,t)\left(0,5-0,5f^{(2)}(k_2-1,t)\right)+0,5f^{(2)}(k_2-1,t)\right)f^{(3)}(k_3,t)}{1-0,5f^{(1)}(k_1,t)-f^{(2)}(k_2-1,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)}\right)+$$

$$+\frac{\left(1+f^{(1)}(k_1,t)\right)f^{(2)}(k_2-1,t)}{\left(0,5f^{(1)}(k_1,t)+f^{(2)}(k_2-1,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)\right)}\times$$

$$\times\frac{1}{\left(\left(1-f^{(1)}(k_1,t)\right)+f^{(2)}(k_2-1,t)+0,5\left(f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-\left(f^{(3)}(k_3,t)-f^{(2)}(k_2-1,t)f^{(2)}(k_3,t)\right)\right)\right)}+$$

$$+0,5\left(\frac{0,5\left(1+f^{(1)}(k_1,t)\left(1-f^{(2)}(k_2-1,t)\right)-0,5f^{(2)}(k_2-1,t)\right)f^{(3)}(k_3,t)}{1-0,5f^{(1)}(k_1,t)-f^{(2)}(k_2-1,t)+0,5f^{(1)}(k_1,t)f^{(2)}(k_2-1,t)-0,5f^{(3)}(k_3,t)+0,5f^{(2)}(k_2-1,t)f^{(3)}(k_3,t)}\right).$$

Considering the above, you can find conditional probability values $\beta_{ij}(k,t)$ *for each service system for each network state* $k, i, j = \overline{1,3}$. *By definition* $\beta_{ij}(k,t)$ *shows that* $\beta_{ij}(k,t) = 0$, $i = \overline{1,3}$. *In addition, some transitions between network states are impossible, which is due to the probability matrix between the states of the network. Therefore, the corresponding conditional probabilities* $\beta_{ij}(k,t)$ *are also equal to zero. We will write conditional probabilities* $\beta_{ij}(k,t)$, $i = \overline{1,3}$*:*

$$\beta_{21}(1,2,3,t) = 0,5\psi_{11}(1,1,3,t) + 0,5\psi_{31}(1,1,3,t) = 0 + 0 = 0,$$
$$\beta_{21}(2,1,3,t) = 0,5\psi_{11}(2,0,3,t) + 0,5\psi_{31}(2,0,3,t) = 0 + 0 = 0,$$
$$\beta_{21}(0,3,3,t) = 0,5\psi_{11}(0,2,3,t) + 0,5\psi_{31}(0,2,3,t) = 0,25 + 0,125 = 0,375,$$
$$\beta_{21}(2,2,2,t) = 0,5\psi_{11}(2,1,2,t) + 0,5\psi_{31}(2,1,2,t) = 0 + 0 = 0,$$
$$\beta_{21}(1,3,2,t) = 0,5\psi_{11}(1,2,2,t) + 0,5\psi_{31}(1,2,2,t) = 0,25 + 0 = 0,25,$$
$$\beta_{23}(1,2,3,t) = 0,5\psi_{13}(1,1,3,t) + 0,5\psi_{33}(1,1,3,t) = 0,75 + 0 = 0,75,$$
$$\beta_{23}(2,1,3,t) = 0,5\psi_{13}(2,0,3,t) + 0,5\psi_{33}(2,0,3,t) = 0,5 + 0 = 0,5,$$
$$\beta_{23}(2,2,2,t) = 0,5\psi_{13}(2,1,2,t) + 0,5\psi_{33}(2,1,2,t) = 0,5 + 0,125 = 0,625,$$
$$\beta_{23}(2,3,1,t) = 0,5\psi_{13}(2,2,1,t) + 0,5\psi_{31}(2,1,2,t) = 0 + 0,125 = 0,125,$$
$$\beta_{12}(0,3,3,t) = 0,$$
$$\beta_{12}(1,2,3,t) = \psi_{22}(0,2,3,t) = 0,$$
$$\beta_{12}(1,3,2,t) = \psi_{22}(0,3,2,t) = 0,$$
$$\beta_{13}(2,2,2,t) = \psi_{23}(1,2,2,t) = 0,$$
$$\beta_{13}(2,3,1,t) = \psi_{23}(1,3,1,t) = 0,$$
$$\beta_{31}(2,2,2,t) = \psi_{21}(2,2,1,t) = 0,$$
$$\beta_{32}(2,3,1,t) = \psi_{22}(2,3,0,t) = 0.$$

Let $r_i(k) = 4$ *bytes be the volume of one request. Let* $r_{ij}(k,t)$ *- linear because of* t *and independent of state of* k *equals* $20t$ *bytes. The intensity of handling requests in the queueing network system are equal* $\mu_1(t) = \cos(4t) + 1$, $\mu_2(t) = 5\cos(4t) + 5$, $\mu_3(t) = 5\cos(4t) + 1$. *Let's take the initial conditions:* $\nu_1(0,3,3,t) = 0$, $\nu_1(1,2,3,t) = 20$, $\nu_1(1,3,2,t) = 20$, $\nu_1(2,1,3,t) = 40$, $\nu_1(2,2,2,t) = 40$, $\nu_1(2,3,1,t) = 40$, $\nu_2(0,3,3,t) = 60$, $\nu_2(1,2,3,t) = 40$, $\nu_2(1,3,2,t) = 60$, $\nu_2(2,1,3,t) = 10$, $\nu_2(2,2,2,t) = 40$, $\nu_2(2,3,1,t) = 60$, $\nu_3(0,3,3,t) = 60$, $\nu_3(1,2,3,t) = 60$, $\nu_3(1,3,2,t) = 40$, $\nu_3(2,1,3,t) = 60$, $\nu_3(2,2,2,t) = 40$, $\nu_3(2,3,1,t) = 20$.
Then the system of DDE (17) takes the form:

$$\frac{\mathrm{d}\nu_1(0,3,3,t)}{\mathrm{d}t} = -7\left(1+\cos(4t)\right)\nu_1(0,3,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_1(1,2,3,t)}{\mathrm{d}t} = -11\left(1+\cos(4t)\right)\nu_1(1,2,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_1(1,3,2,t)}{\mathrm{d}t} = -11\left(1+\cos(4t)\right)\nu_1(1,3,2,t) + 4,$$
$$\frac{\mathrm{d}\nu_1(2,1,3,t)}{\mathrm{d}t} = -11\left(1+\cos(4t)\right)\nu_1(2,1,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_1(2,2,2,t)}{\mathrm{d}t} = -11\left(1+\cos(4t)\right)\nu_1(2,2,2,t) + 4 - 5t\left(1+\cos(4t)\right),$$

$$\frac{\mathrm{d}\nu_1(2,3,1,t)}{\mathrm{d}t} = -11\,(1+\cos(4t))\,\nu_1(2,3,1,t) + 4,$$
$$\frac{\mathrm{d}\nu_2(0,3,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(0,3,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_2(1,2,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(1,2,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_2(1,3,2,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(1,3,3,t) + 3,75\,(1+\cos(4t))\,\nu_2(1,3,3,t) + 4 - 75t\,(1+\cos(4t))\,,$$
$$\frac{\mathrm{d}\nu_2(2,1,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(2,1,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_2(2,2,2,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(2,2,2,t) + 5\,(1+\cos(4t))\,\nu_2(2,1,3,t) + 4 - 50t\,(1+\cos(4t))\,,$$
$$\frac{\mathrm{d}\nu_2(2,3,1,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(2,3,1,t) + 3,125\,(1+\cos(4t))\,\nu_2(2,2,2,t) + 4 - 12,5t,$$
$$\frac{\mathrm{d}\nu_3(0,3,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_3(0,3,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_2(1,2,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(1,2,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_3(0,3,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(0,3,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_3(1,2,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_2(1,2,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_3(1,3,2,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_3(1,3,2,t) + 7,5\,(1+\cos(4t))\,\nu_3(1,2,3,t) + 75t\,(1+\cos(4t)) + 4,$$
$$\frac{\mathrm{d}\nu_3(2,1,3,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_3(2,1,3,t) + 4,$$
$$\frac{\mathrm{d}\nu_3(2,2,2,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_3(2,2,2,t) + 7,5\,(1+\cos(4t))\,\nu_3(2,3,1,t) + 75t\,(1+\cos(4t)) + 4,$$
$$\frac{\mathrm{d}\nu_3(2,3,1,t)}{\mathrm{d}t} = -7\,(1+\cos(4t))\,\nu_3(2,3,1,t) + 6,25\,(1+\cos(4t))\,\nu_3(2,2,2,t) + 60t\,(1+\cos(4t)) + 4.$$

This system can be solved using the Mathematica 8 package and get a numerical solution. This solution is shown in the form of expected volumes of requests charts for each queueing system of network in the figures 1-4.

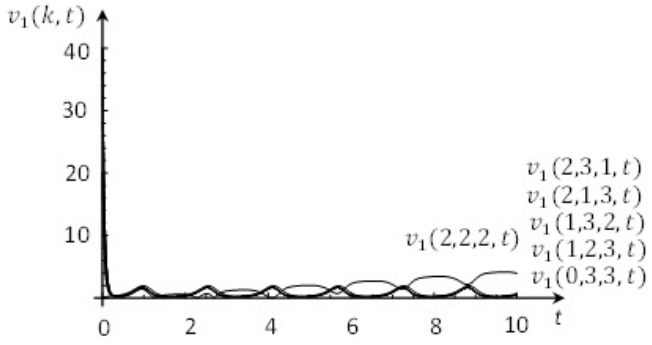


FIGURE 2. The volumes of requests in the system S_1 for $T = [0, 10]$

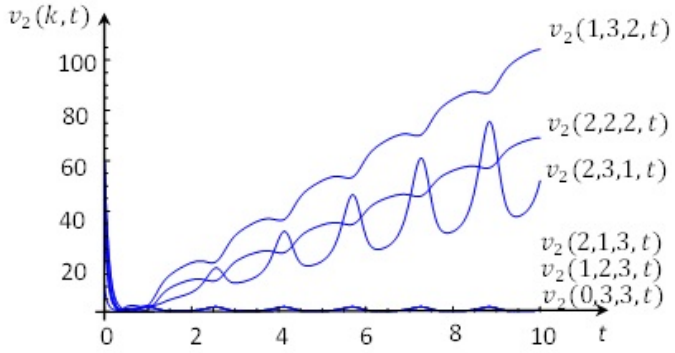


FIGURE 3. The volumes of requests in the system S_2 for $T = [0, 10]$

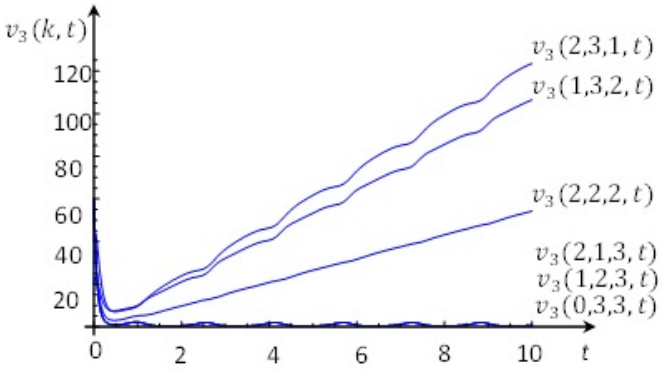


FIGURE 4. The volumes of requests in the system S_3 for $T = [0, 10]$

We can also find analytical solutions for the expected volumes of requests in queueing network systems. Then we get:

$$\nu_1(0,3,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x} \int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x,$$

$$\nu_1(1,2,3,t) = 4\mathrm{e}^{\int_0^t 11(\cos(4x)+1)\mathrm{d}x} \left[5 + \int_0^t \mathrm{e}^{-\int_0^x 11(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(1,3,2,t) = 4\mathrm{e}^{\int_0^t 11(\cos(4x)+1)\mathrm{d}x} \left[5 + \int_0^t \mathrm{e}^{-\int_0^x 11(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(2,1,3,t) = 4\mathrm{e}^{\int_0^t 11(\cos(4x)+1)\mathrm{d}x} \left[10 + \int_0^t \mathrm{e}^{-\int_0^x 11(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(2,2,2,t) = \mathrm{e}^{\int_0^t 11(\cos(4x)+1)\mathrm{d}x} \left[40 + \int_0^t (4 - 5t\,(1+\cos(4t)))\,\mathrm{e}^{-\int_0^x 11(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(2,3,1,t) = 4\mathrm{e}^{\int_0^t 11(\cos(4x)+1)\mathrm{d}x} \left[10 + \int_0^t \mathrm{e}^{-\int_0^x 11(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(0,3,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x} \left[15 + \int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(1,2,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[10+\int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_1(1,3,2,t) = \mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[60+\int_0^t\left(4-75x\left(1+\cos(4x)\right)+\right.\right.$$

$$\left.\left.+3,75\left(1+\cos(4x)\right)\left(4\mathrm{e}^{\int_0^x 7(\cos(4z)+1)\mathrm{d}z}\left[15+\int_0^x \mathrm{e}^{-\int_0^s 7(\cos(4l)+1)\mathrm{d}l}\mathrm{d}s\right]\right)\right)\mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_2(2,1,3,t) = 2\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[5+2\int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right]$$

$$\nu_2(2,2,2,t) = \mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[40+\int_0^t\left(4-50x\left(1+\cos(4x)\right)+\right.\right.$$

$$\left.\left.+5\left(1+\cos(4x)\right)\left(2\mathrm{e}^{\int_0^x 7(\cos(4z)+1)\mathrm{d}z}\left[5+2\int_0^x \mathrm{e}^{-\int_0^s 7(cos(4l)+1)\mathrm{d}l}\mathrm{d}s\right]\right)\right)\mathrm{e}^{-\int_0^x 7(cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_2(2,3,1,t) = \mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[60+\int_0^t\left(4-12,5x+3,125\left(1+\cos(4x)\right)\times\right.\right.$$

$$\times\mathrm{e}^{\int_0^x 7(\cos(4l)+1)\mathrm{d}l}\left[60+\int_0^x\left(4-50h\left(1+\cos(4h)\right)+5\left(1+\cos(4h)\right)\times\right.\right.$$

$$\left.\left.\left.\left.\times\left(4\mathrm{e}^{\int_0^h 7(\cos(4k)+1)\mathrm{d}k}\int_0^h \mathrm{e}^{-\int_0^s 7(\cos(4w)+1)\mathrm{d}w}\mathrm{d}s\right)\right)\mathrm{e}^{-\int_0^h 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}h\right)\right]\right)\mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x,$$

$$\nu_3(0,3,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[15+\int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_3(1,2,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[15+\int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_3(1,3,2,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[10+\int_0^t\left(4+75x\left(1+\cos(4x)\right)+7,5x\left(1+\cos(4x)\right)\times\right.\right.$$

$$\left.\left.\times\left(4\mathrm{e}^{\int_0^x 7(\cos(4z)+1)\mathrm{d}z}\left[15+\int_0^t \mathrm{e}^{-\int_0^q 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}q\right]\right)\right)\mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_3(2,1,3,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[15+\int_0^t \mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right],$$

$$\nu_3(2,2,2,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[10+\int_0^t\left(4+12,5x\left(1+\cos(4x)\right)+1,25\left(1+\cos(4x)\right)\times\right.\right.$$

$$\left.\left.\times\left(\nu_3(2,3,1,t)\right)\mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right]\right.$$

$$\nu_3(2,2,2,t) = 4\mathrm{e}^{\int_0^t 7(\cos(4x)+1)\mathrm{d}x}\left[5+\int_0^t\left(4+62,5x\left(1+\cos(4x)\right)+7,25\left(1+\cos(4x)\right)\times\right.\right.$$

$$\left.\left.\times\left(\nu_3(3,2,2,t)\right)\mathrm{e}^{-\int_0^x 7(\cos(4y)+1)\mathrm{d}y}\mathrm{d}x\right]\right..$$

6. Conclusions

We present a method of finding the expected volume of requests in open HM-network with homogeneous requests, bypass of queueing network systems of requests. Were considered a case where the changes in volumes associated with transitions between states of the network are deterministic functions dependent states of network and time, and service systems are single line, assuming that the probability of network systems states, the parameters of entrance flow of messages and service depend on time.

These results can be used to find the amount of memory in information systems. A model of wireless computer network operation with limited number of concurrent connections is presented. This network was analyzed with homogeneous requests, bypass of service nodes, and time-dependent parameters of stream of requests. The conditional probabilities for the service of requests are given by (1)-(6).

A set of differential equations was obtained for the expected volume of homogeneous requests in network systems with limited number of waiting places and an example of a solution of the equation system for this network is shown in Figure 1, where the probability of joining the request to queues in queueing systems depends only on the number of requests in them. Graphs the expected volume of requests for queueing network systems can be found in Figures 2-4.

References

[1] Tikhonenko O.: *Metody probalistyczne analizy systemów informacyjnych*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2006.

[2] Matalytski M., Tikhonenko O., Koluzaeva E.: *Systems and queueing networks: analysis and application*, GrSU, Grodno 2011 (in Russian).

[3] Crabill, T.: *Optimal Control of a Service Facility with Variable Exponential Service Times and Constant Arrival Rate*, Management Science, V. 18 (9), 1972, p. 560-566.

[4] Foschini, G.: *On heavy traffic diffusion analysis and dynamic routing in packet switched networks*, Computer Performance, 10 (1977), p. 499-514.

[5] Stidham S., Weber R.: *A survey of Markov decision models for control of networks of queue*, Queueing Systems, V.13 (3), 1993, p. 291-314.

[6] Matalytski M., Pankov A.: *Analysis of the stochastic model of the changing of incomes in the open banking network*, Computer Science, v. 3 (5), 2003, p. 19-29.

[7] Matalytski M.: *On some results in analysis and optimization of Markov networks with incomes and their applications*, Automation and Remote Control, v. 70 (10), 2009, p. 1683-1697.

[8] Matalytski M., Naumenko V.: *Zastosowanie HM-sieci kolejkowych dla wyznaczenia objętości pamięci systemów informacyjnych*, Studia Informatica, V. 35 (3), 2014, p. 63-69.

[9] Zając P., Matalytski M.: *Finding of expected volumes of requests in systems of the queueing network with a limited amount of spaces of expectations*, Studia Informatica, V. 35 (3), 2014, p. 63-69.

Received: June 2017

Paweł Zając
Czestochowa University of Technology,
Institute of Mathematics,
al. Armii Krajowej 21, 42-201 Częstochowa, Poland
E-mail address: pawel_zajac@vp.pl

Mikhail Matalytski
Czestochowa University of Technology,
Institute of Mathematics,
al. Armii Krajowej 21, 42-201 Częstochowa, Poland
E-mail address: m.matalytski@gmail.com

Scientific Issues
Jan Długosz University
in Częstochowa
Mathematics XXII (2017)
161–175
DOI http://dx.doi.org/10.16926/m.2017.22.12

COMPARING SAT- AND SMT- BASED BOUNDED MODEL CHECKING FOR ECTL PROPERTIES

AGNIESZKA M. ZBRZEZNY

Abstract

We compare two bounded model checking methods for properties expressed in the existential fragment of Computation Tree Logic (ECTL). To this end we use the generic pipeline paradigm (GPP) and the train controller system (TC), the classic concurrency problems, which we formalise by means of a finite transition system. We consider several properties of the problems that can be expressed in ECTL logic, and we present the performance evaluation of the mentioned bounded model checking methods by means of the running time and the memory used.

1. Introduction

Bounded model checking [2, 3, 9] (BMC) is one of the symbolic model checking techniques designed for finding witnesses for existential properties or counterexamples for universal properties. Its main idea is to consider a model reduced to a specific depth. The method works by mapping a bounded model checking problem to the satisfiability problem (SAT) or to satisfiability modulo theories problem (SMT). For the existential part of Computation Tree Logic (ECTL) [6] and a network of automata the BMC method can by described as follows: given a model $\mathcal{M}$ for a network of automata, an ECTL formula φ, and a bound k, a model checker creates a propositional formula or a quantifier-free first-order formula $[\mathcal{M}, \varphi]_k$ that is satisfiable if and only if the formula φ is true in the model $\mathcal{M}$.

In this paper, we make the following contributions. Firstly, we define and implement an SMT-based BMC method for ECTL and for a network of automata modelled using transition systems. Next, we report on the

• *Agnieszka M. Zbrzezny* — e-mail: agnieszka.zbrzezny@ajd.czest.pl
Jan Długosz University in Częstochowa.

Partly supported by National Science Centre under the grant No. 2014/15/N/ST6/05079.

initial experimental evaluation of our SMT-based BMC method. To this aim, we use two scalable benchmarks: the *generic pipeline paradigm* [8] and the *train controller system* [7].

The structure of the paper is as follows. In Section 2 we shortly recall definition of transition systems. Then, we present syntax and semantics of ECTL. In Section 3 we shortly present BMC technique for ECTL. and we define a new SMT-based BMC method for ECTL. In Section 4 we present experimental evaluation of the SAT-based BMC [12] and SMT-based BMC for ECTL and for a train controller system (TC) and generic pipeline paradigm (GPP). In Section 5 we conclude the paper.

2. Preliminaries

In this section we introduce the basic definitions used in the paper. In particular, we define the semantics of a transition system (TS) and syntax and semantics of ECTL.

2.1. **Transition System.** The transition system [1] (also called a *model*) is a tuple

$$\mathcal{M} = (S, Act, \longrightarrow, s^0, \mathcal{AP}, L) \text{ where:}$$

S is a set of states, Act is a set of actions, $\longrightarrow \subseteq S \times Act \times S$ is a transition relation, $s^0 \in S$ is the initial state, $\mathcal{AP}$ is a set of atomic propositions, and $L : S \to 2^{\mathcal{AP}}$ is a labelling function assigning to each state a set of atomic propositions that are assumed to be true at that state. The transition system is called finite if S, Act, and $\mathcal{AP}$ are finite. We write $s \xrightarrow{a} s'$ instead of $(s, a, s) \in \longrightarrow$. Moreover, we write $s \longrightarrow s'$ if $s \xrightarrow{a} s'$, for some $a \in Act$.

We assume that a considered model has no terminal states, i.e. for every $s \in S$ there exist $s' \in S$ such that $s \longrightarrow s'$. The set of all natural numbers is denoted by $\mathbb{N}$. A *path* in $\mathcal{M}$ is an infinite sequence $\pi = (s_0, s_1, \ldots)$ of states such that $s_i \longrightarrow s_{i+1}$ for each $i \in \mathbb{N}$. For a path $\pi = (s_0, s_1, \ldots)$ and $i \in \mathbb{N}$, the i-th state of π is defined as $\pi(i) = s_i$. By $\Pi(s)$ we denote the set of all the paths starting at $s \in S$.

2.2. **ECTL.** The Existential Computation Tree Logic is a restriction of a propositional branching-time temporal logic CTL that was introduced by Emerson and Clarke in [6]. The syntax of ECTL formulae over the set $\mathcal{AP}$ of atomic propositions is defined by the following grammar: $\varphi := \mathbf{true} \mid \mathbf{false} \mid p \mid \neg p \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \mathbf{EX}\varphi \mid \mathbf{E}(\varphi \mathbf{U} \varphi) \mid \mathbf{EG}\varphi$, where $p \in \mathcal{AP}$ and φ is a formula. The symbols $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{G}$ are the modal operators for "neXt time", "Until" and "Globally", respectively. The symbol $\mathbf{E}$ is the existential path quantifier. The derived basic modalities are defined as follows:

$$\mathbf{EF}\alpha \stackrel{df}{=} \mathbf{E}(\mathbf{true}\,\mathbf{U}\,\alpha), \qquad \mathbf{E}(\alpha\,\mathbf{R}\,\beta) \stackrel{df}{=} \mathbf{E}(\beta\,\mathbf{U}\,(\alpha \wedge \beta)) \vee \mathbf{EG}\beta.$$

Let $\mathcal{M}$ be a model, and φ an ECTL formula. An ECTL formula φ is *true* in the model $\mathcal{M}$ (in symbols $\mathcal{M} \models \varphi$) iff $\mathcal{M}, s^0 \models \varphi$ (i.e., φ is true at the initial state of the model $\mathcal{M}$), where

$\mathcal{M}, s \models \mathbf{true}$,

$\mathcal{M}, s \not\models \mathbf{false}$,

$\mathcal{M}, s \models p$ iff $p \in L(s)$,

$\mathcal{M}, s \models \neg p$ iff $p \notin L(s)$,

$\mathcal{M}, s \models \alpha \wedge \beta$ iff $\mathcal{M}, s \models \alpha$ and $\mathcal{M}, s \models \beta$,

$\mathcal{M}, s \models \alpha \vee \beta$ iff $\mathcal{M}, s \models \alpha$ or $\mathcal{M}, s \models \beta$,

$\mathcal{M}, s \models \mathbf{EX}\alpha$ iff $(\exists \pi \in \Pi(s))(\mathcal{M}, \pi(1) \models \alpha)$,

$\mathcal{M}, s \models \mathbf{E}(\alpha \mathbf{U} \beta)$ iff $(\exists \pi \in \Pi(s))(\exists m \geq 0)(\mathcal{M}, \pi(m) \models \beta$ and $(\forall j < m)\mathcal{M}, \pi(j) \models \alpha)$,

$\mathcal{M}, s \models \mathbf{EG}\alpha$ iff $(\exists \pi \in \Pi(s))(\forall m \geq 0)(\mathcal{M}, \pi(m) \models \alpha)$.

3. Bounded Model Checking

The satisfiability modulo theories (SMT) problem is a decision problem for logical formulas with respect to combinations of background theories expressed in classical first-order logic with equality. The SMT-based Bounded Model Checking (BMC) is a popular model checking technique for the verification of concurrent systems. We have given a model $\mathcal{M}$, an existential modal formula φ, and a non-negative bound k, the SMT-based BMC consists in searching for a non-empty set of paths of length k that constitute a witness for the checked property φ. In particular, the SMT-based bounded model checking algorithms generate a quantifier-free first-order formula which is satisfiable if and only if the mentioned set of paths exist. The quantifier-free first-order formula is usually obtained as a combination of an SMT encoding of the unfolding of the transition relation of the given model, and an SMT encoding of the property in question. If the generated quantifier-free first-order formula is not satisfiable, then k is incremented until either the problem becomes intractable due to the complexity of solving the corresponding SAT instance, or k reaches the upper bound of the bounded model checking problem for the language under consideration.

We have implemented a translation to SMT strictly following the translation to SAT given in [12].

Since $\mathcal{M}$ is a parallel composition of a finite number n of finite transition systems, every state of $\mathcal{M}$ can be encoded as a natural number vector of the length n. Thus, each state of $\mathcal{M}$ can be represented by a valuation

of a vector (called a *symbolic state*) of different individual variables called *individual state variables*. Moreover, every action of $\mathcal{M}$ can be represented by a valuation of an individual variable, and the designated positions l of the k-paths used in the translation can be also be represented by valuations of individual variables. Furthermore, k-paths can be represented as vectors of symbolic states.

The SAT-based BMC method for ECTL was introduced in [9], and then it was improved in [12]. Unfortunately the encoding presented in [12] does not encode actions In our new SMT-based approach we encode actions as well.

3.1. **SMT-based bounded model checking.** Let $\mathcal{M}$ be a model, $k \geq 0$ a bound, φ an ECTL formula, and let $\mathcal{M}, s \models_k \varphi$ denotes that φ is k-true at the state s of $\mathcal{M}$. The formula φ is k-true in $\mathcal{M}$ (in symbols $\mathcal{M} \models_k \varphi$) iff $\mathcal{M}, s^0 \models_k \varphi$ (i.e., φ is k-true at the initial state of the model $\mathcal{M}$).

The *bounded model checking problem* asks whether there exists $k \in \mathbb{N}$ such that $\mathcal{M} \models_k \varphi$. The following theorem states that for a given model and an ECTL formula there exists a bound k such that the model checking problem ($\mathcal{M} \models \varphi$) can be reduced to the BMC problem ($\mathcal{M} \models_k \varphi$). The theorem can be proven by induction on the length of the formula φ.

Theorem 1 ([12])**.** *Let $\mathcal{M}$ be a model and φ an* ECTL *formula. Then, the following equivalence holds: $\mathcal{M} \models \varphi$ iff there exists $k \geq 0$ such that $\mathcal{M} \models_k \varphi$.*

Translation to SMT. The translation to SMT is based on the bounded semantics. Let $\mathcal{M}$ be a model, φ an ECTL formula, and $k \geq 0$ a bound. The presented SMT encoding of the BMC problem for ECTL is based on the SAT encoding of the same problem [12], and it relies on defining the quantifier-free first-order formula $[\mathcal{M}, \varphi]_k := [\mathcal{M}^{\varphi, s^0}]_k \wedge [\varphi]_{\mathcal{M},k}$ that is satisfiable if and only if $\mathcal{M} \models_k \varphi$ holds.

The definition of the formula $[\mathcal{M}^{\varphi,s^0}]_k$ assumes that states of the model $\mathcal{M}$ are encoded in a symbolic way. Such a symbolic encoding is possible, since the set of states of $\mathcal{M}$ is finite. In particular, each state s can be represented by a vector $\overline{\mathbf{w}}$ (called a *symbolic state*) of different individual variables ranging over the natural numbers (called *individual state variables*) and each action can be represented by a valuation of a symbolic action $\overline{\mathbf{a}}$ that is a vector of the individual variables ranging over the natural numbers. The formula $[\mathcal{M}^{\varphi,s^0}]_k$ encodes a rooted tree of $k-$paths of the model $\mathcal{M}$. The number of branches of the tree depends on the value of the auxiliary function $f_k : \text{ECTL} \to \mathbb{N}$ defined in [10].

Given the above, the j-th symbolic k-path $\boldsymbol{\pi}_j$ is defined as the following sequence $(\overline{\mathbf{w}}_{0,j}, \ldots, \overline{\mathbf{w}}_{k,j})$, where $\overline{\mathbf{w}}_{i,j}$ are symbolic states for $0 \leq i \leq k$ and $0 \leq j < f_k(\varphi)$.

Let $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}'$ be two different symbolic states. We assume definitions of the following auxiliary quantifier-free first-order formulae: $I_{s^0}(\overline{\mathbf{w}})$ - encodes the initial state of the model $\mathcal{M}$, $\mathcal{T}(\overline{\mathbf{w}}, \overline{\mathbf{a}}, \overline{\mathbf{w}}')$ - encodes the transition relation of $\mathcal{M}$, and $p(\overline{\mathbf{w}})$ - encodes the set of states of $\mathcal{M}$ in which $p \in \mathcal{AP}$ holds.

The formula $[\mathcal{M}^{\varphi,s^0}]_k$ encoding the unfolding of the transition relation of the model $\mathcal{M}$ $f_k(\varphi)$-times to the depth k is defined as follows:

$$[\mathcal{M}^{\varphi,s^0}]_k := I_{s^0}(\overline{\mathbf{w}}_{0,0}) \wedge \bigwedge_{j=0}^{f_k(\varphi)-1} \bigwedge_{i=0}^{k-1} \mathcal{T}(\overline{\mathbf{w}}_{i,j}, \overline{\mathbf{a}}_{i,j} \overline{\mathbf{w}}_{i+1,j}) \tag{1}$$

For every ECTL formula φ the function f_k determines how many symbolic k-paths are needed for translating the formula φ. Given a formula φ and a set A of k-paths such that $|A| = f_k(\varphi)$, we divide the set A into subsets needed for translating the subformulae of φ. To accomplish this goal we need the auxiliary functions $h_n^{\mathbf{U}}(A, e)$ and $h_n^{\mathbf{R}}(A, e)$ that were defined in [12].

Below we show the translation for the temporal operators **EX**, **EU** and **EG** only.

Let φ be an ECTL formula, $\mathcal{M}$ a model, and $k \in \mathbb{N}$ a bound. The quantifier-free first-order formula $[\varphi]_{\mathcal{M},k} := [\varphi]_k^{[0,0,F_k(\varphi)]}$, where $F_k(\varphi) = \{j \in \mathbb{N} \mid 0 \leqslant j < f_k(\varphi)\}$, encodes the bounded semantics for ECTL, and it is defined inductively as shown below. Namely, let $0 \leqslant n < f_k(\varphi)$, $m \leqslant k$, $n' = min(A)$, $h_{\mathbf{X}} = h_{\mathbf{X}}(A)$, $h_k^{\mathbf{U}} = h_k^{\mathbf{U}}(A, f_k(\beta))$, and $h_k^{\mathbf{G}} = h_k^{\mathbf{G}}(A, f_k(\alpha))$, then:

$$[\mathbf{EX}\alpha]_k^{[m,n,A]} := \overline{\mathbf{w}}_{m,n} = \overline{\mathbf{w}}_{0,n'} \wedge [\alpha]_k^{[1,n',h_{\mathbf{X}}]},$$

$$[\mathbf{EG}\alpha]_k^{[m,n,A]} := \overline{\mathbf{w}}_{m,n} = \overline{\mathbf{w}}_{0,n'} \wedge L_k(n') \wedge \bigwedge_{j=0}^{k} [\alpha]_k^{[j,n',h_{\mathbf{G}}(j)]},$$

$$[\mathbf{E}(\alpha\,\mathbf{U}\,\beta)]_k^{[m,n,A]} := \overline{\mathbf{w}}_{m,n} = \overline{\mathbf{w}}_{0,n'} \wedge \bigvee_{i=0}^{k} \left([\beta]_k^{[i,n',h_{\mathbf{U}}(k)]} \wedge \bigwedge_{j=0}^{i-1} [\alpha]_k^{[j,n',h_{\mathbf{U}}(j)]} \right).$$

Theorem 2. *Let $\mathcal{M}$ be a model, and φ an* ECTL *formula. Then for every $k \in \mathbb{N}$, $M \models_k \varphi$ if, and only if, the formula $[\mathcal{M}, \varphi]_k$ is satisfiable.*

3.2. **Example.** Now we show how to apply our SMT-based BMC method to verify the *generic pipeline paradigm* (GPP) TS model. The model of GPP involves $n + 2$ automata: Producer producing data or being inactive, Consumer receiving data or being inactive, and a chain of n intermediate Nodes which can be ready for receiving data, processing data or sending data.

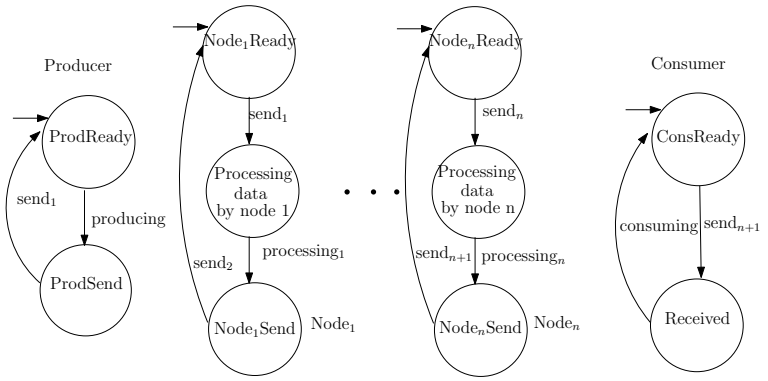


FIGURE 1. The GPP system [11]

As an example we consider the system with one node. We will show the important parts of the unfolding (to the depth 2) of the transition relation and also the translation of the temporal formula $\varphi_5 = \mathbf{EF}(ProdSend \wedge Received)$ on the k-path of the length 2. To this aim we need some auxiliary notations.

Let w denotes the symbolic global state and a denotes the symbolic action. We encode the location $ProdReady$ by 0 and the location $ProdSend$ by 1; the locations $Node_1Ready$ is encoded by 0, the location $Node_1Proc$ by 1, and the location $Node_1Send$ by 2; eventually the location $ConsReady$ is encoded by 0, and the location $ConsReceived$ by 1.

Moreover, we encode the action $Produce$ by 1, the action $Send_1$ by 2, the action $Proc_1$ by 4, the action $Send_2$ by 3, and the action $Consume$ by 5.

Now we are ready to present some of the formulae involved in the unfolding of the transition relation and in the translation of the formula φ.

The only initial state is represented by the following formula: $a_0 = 0 \wedge w_0 = 0 \wedge w_1 = 0 \wedge w_2 = 0$.

The SMT file for the GPP system and the formula φ_5 generated by our implementation is following:

```
(set-logic QF_LIA)
```

```
(declare-fun w0 () Int)
(declare-fun w1 () Int)
(declare-fun w2 () Int)
(declare-fun a0 () Int)
(declare-fun w3 () Int)
(declare-fun w4 () Int)
(declare-fun w5 () Int)
(declare-fun a3 () Int)
(declare-fun w6 () Int)
(declare-fun w7 () Int)
(declare-fun w8 () Int)
(declare-fun a6 () Int)

; Path nr 0: first state
(assert true)

(assert (and (or (and (= w0 0) (= w3 1) (= a3 1)) (and (= w0 1)
(= w3 0) (= a3 2)) (and (not (= a3 1)) (not (= a3 2))
(= w0 w3))) (or (and (= w1 0) (= w4 1) (= a3 2)) (and (= w1 2)
(= w4 0) (= a3 3)) (and (= w1 1) (= w4 2) (= a3 4))
(and (not (= a3 2)) (not (= a3 3)) (not (= a3 4)) (= w1 w4)))
(or (and (= w2 0) (= w5 1) (= a3 3)) (and (= w2 1) (= w5 0)
(= a3 5)) (and (not (= a3 3)) (not (= a3 5)) (= w2 w5)))
(or (= a3 1) (= a3 2) (= a3 3) (= a3 4) (= a3 5))))

(assert (and (or (and (= w3 0) (= w6 1) (= a6 1)) (and (= w3 1)
(= w6 0) (= a6 2)) (and (not (= a6 1)) (not (= a6 2)) (= w3 w6)))
(or (and (= w4 0) (= w7 1) (= a6 2)) (and (= w4 2) (= w7 0)
(= a6 3)) (and (= w4 1) (= w7 2) (= a6 4)) (and (not (= a6 2))
(not (= a6 3)) (not (= a6 4)) (= w4 w7))) (or (and (= w5 0)
(= w8 1) (= a6 3)) (and (= w5 1) (= w8 0) (= a6 5))
(and (not (= a6 3)) (not (= a6 5)) (= w5 w8)))
(or (= a6 1) (= a6 2) (= a6 3) (= a6 4) (= a6 5))))

; Translated formula
(assert (and (or (and (= w0 1) (= w2 1)) (and (= w3 1) (= w5 1))
(and (= w6 1)(= w8 1))) (and (= w0 w0) (= w1 w1) (= w2 w2))))

; Initial path: 0
(assert (and (= a0 0) (= w0 0) (= w1 0) (= w2 0)))

(check-sat)
(get-model)
```

The corresponding DIMACS file is following:

```
p cnf 83 222
-80 81 82 0
-77 82 83 0
-74 -76 77 0
-70 -71 72 0
-69 72 73 0
-66 -67 68 0
-64 68 69 0
-61 -62 63 0
-59 -60 61 0
-57 63 64 0
-52 -54 55 0
-49 -62 67 0
-49 -55 56 0
-48 -62 71 0
-48 -49 50 0
-46 -47 48 0
-45 -50 51 0
-41 -42 43 0
-40 43 44 0
-37 -38 39 0
-35 -58 59 0
-35 -36 37 0
-34 39 40 0
-31 -32 33 0
-29 -30 31 0
-27 -53 54 0
-27 -28 29 0
-26 33 34 0
...
```

4. Experimental Results

In this section we experimentally evaluate the performance of our SMT-based BMC encoding for Ectl over the TS semantics. We compare our experimental results with the experimental results generated using SAT-based [12] We have conducted the experiments using two benchmarks: the generic pipeline paradigm (GPP) TS model [8] and the train controller system (TCS) TS model [7]. We would like to point out that both benchmarks are very useful and scalable examples.

An evaluation of both BMC algorithms, which have been implemented in C++ is given by means of the running time, the memory used, and the number of generated variables and clauses.

We would like to point out that both benchmarks are very useful and scalable examples.

4.1. **A Train Controller System.** To evaluate the BMC techniques for ECTL, we analyse a scalable concurrent system, which is a train controller system (TCS). The system consists of a controller, and n trains (for $n \geq 2$), and it is assumed that each train uses its own circular track for travelling in one direction. All trains have to pass through a tunnel, but because there is only one track in the tunnel, arriving trains cannot use it simultaneously. There are signal lights on both sides of the tunnel, which can be either red or green. All trains notify the controller when they request entry to the tunnel or when they leave the tunnel. The controller controls the colour of the signal lights

An automata model of the TCS system is shown on Figure 2.

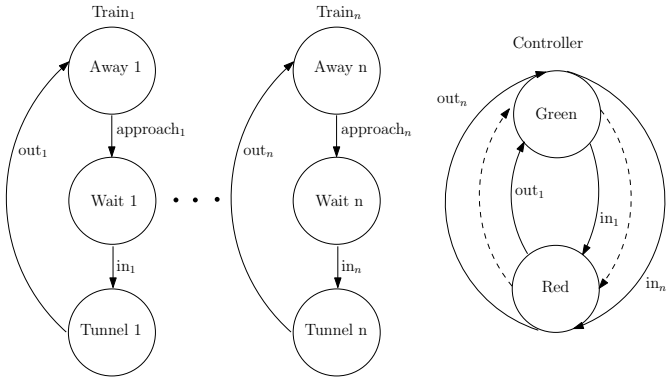


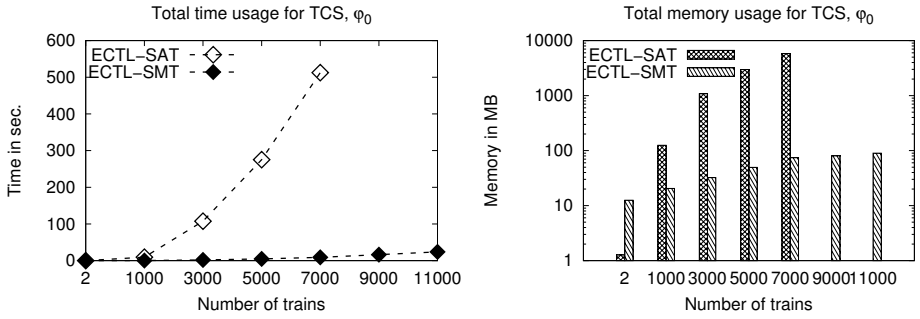
FIGURE 2. A network of automata for a train controller system ([11])

The specifications for it are given in the existential form, i.e., they are expressed in the ECTL language.

Formula φ_0 states that there exists the case that all trains are outside the tunnel and Train n is in the tunnel, where n is the number of trains.

$$\varphi_0 = \mathbf{EF}\big((\bigwedge_{i=1}^{n-1} \neg InTunnel_i) \wedge InTunnel_n\big).$$

In Figures 3(a) and 3(b) we present a comparison of total time usage and total memory usage for the formulae φ_0. As we can see, in this case SMT-BMC is much better than SAT approach. The reason is that we need only one symbolic to verify the formula.



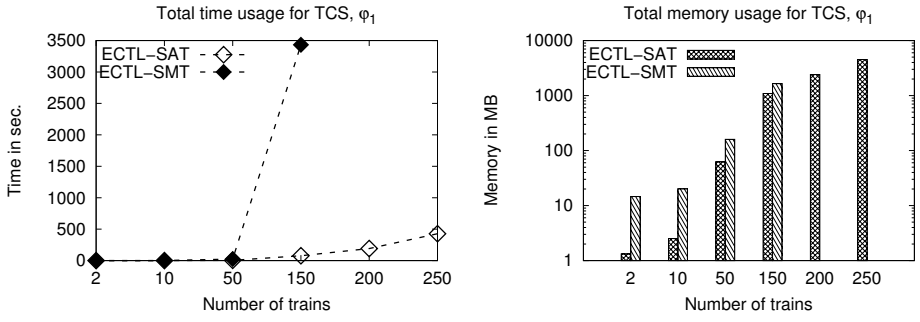
(a) (b)

FIGURE 3. A comparison of total time usage and total memory usage for the formulae φ_0

Formula φ_1 states that there exists the case that Train 1 is in the tunnel and either it and other train will not be in the tunnel during the next $n+1$ steps, where n is the number of trains.

$$\varphi_1 = \mathbf{EF}(InTunnel_1 \wedge \mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \wedge \underbrace{\mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \wedge \mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \ldots))}_{n})).$$

In Figures 4(a) and 4(b) we present a comparison of total time usage and total memory usage for the formulae φ_1.



(a) (b)

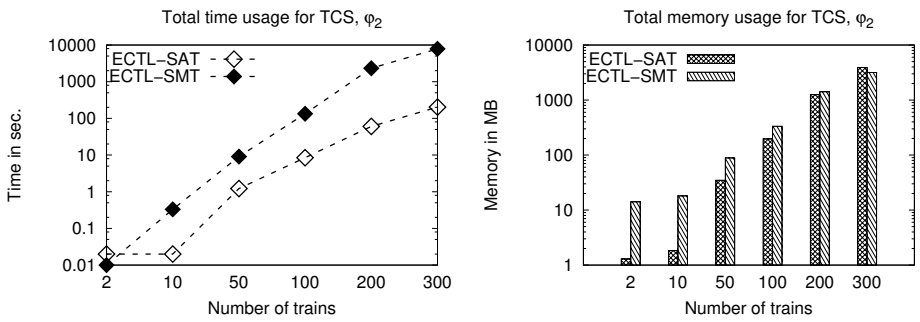
FIGURE 4. A comparison of total time usage and total memory usage for the formulae φ_1

An analysis of experimental results for formula φ_1 leads to the conclusion that SAT-based BMC for ECTL uses less time and memory comparing to SMT-based BMC for ECTL. The reason is that although BMC needs a lot of paths for verification, but these paths are short. The SAT-based algorithm was not able to verify the system with 300 on lack of memory. In the case of SAT-based BMC, generation of DIMACS file consume most of the memory and time.

Formula φ_2 expresses that there exists the case that Train 1 is in the tunnel or either it or other train will not be in the tunnel during the next $n+1$ steps, where n is the number of trains.

$$\varphi_2 = \mathbf{EF}(InTunnel_1 \vee \mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \vee \underbrace{\mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \vee \mathbf{EX}(\bigwedge_{i=1}^{n} \neg InTunnel_i \ldots))}_{n}).$$

In Figures 5(a) and 5(b) we present a comparison of total time usage and total memory usage for the formulae φ_2.



(a) (b)

FIGURE 5. A comparison of total time usage and total memory usage for the formulae φ_2.

An observation of experimental results for formula φ_2 leads to the conclusion that SAT-BMC for ECTL uses less time and memory comparing to SMT-based BMC for ECTLand SAT-based BMC is significantly faster than SMT-based approach. In this case almost all memory was used by SMT- and SAT-solvers. The reason is that BMC needs a lot of short paths for verification. In this case SAT-based BMC is usually better than SMT-based BMC.

4.2. **Generic Pipeline Paradigm.** The benchmark we consider is the generic pipeline paradigm (GPP) [8], which consists of three parts: Producer producing data, Consumer receiving data, and a chain of n intermediate Nodes that transmit data produced by Producer to Consumer. The local states for each component (Producer, Consumer, and intermediate Nodes), and their protocols are shown on Fig. 1.

Formula φ_3 states that there exists a path that always Producer is ready to produce data or either Consumer will receive data in maximum $2n+1$ steps.

$$\varphi_3 = \mathbf{EG}(\neg ProdSend \vee \underbrace{\mathbf{EX}(}_{2n+1} Received))$$

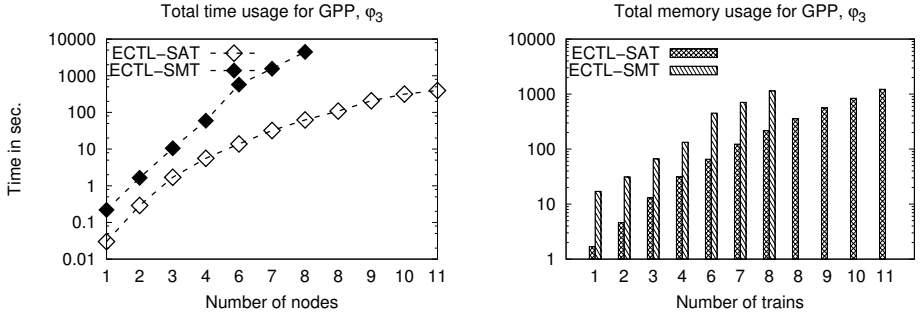
Formula φ_4 states that there exists a path that always in maximum $n^2 + 2n$ steps Consumer will receive the data.

$$\varphi_4 = \underbrace{\mathbf{EX}(Received \wedge \mathbf{EX}(Received \ldots)))}_{n^2+2n}$$

Formula φ_5 states that there exists a path that Producer will produce data and Consumer will receive the data.

$$\varphi_5 = \mathbf{EF}(ProdSend \wedge Received)$$

In Figures 6(a) and 6(b) we present a comparison of total time usage and total memory usage for the formulae φ_3.

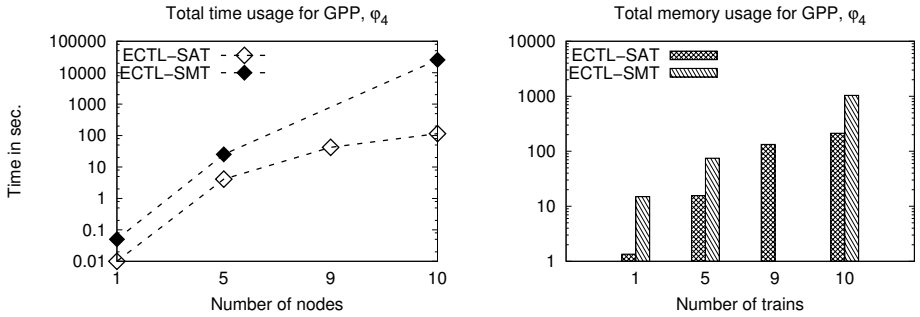


(a) (b)

FIGURE 6. A comparison of total time usage and total memory usage for the formulae φ_3.

An observation of experimental results for formula φ_3 leads to the conclusion that SAT-BMC uses less time and memory comparing to SMT-BMC for ECTL.

In Figures 7(a) and 7(b) we present a comparison of total time usage and total memory usage for the formulae φ_4.

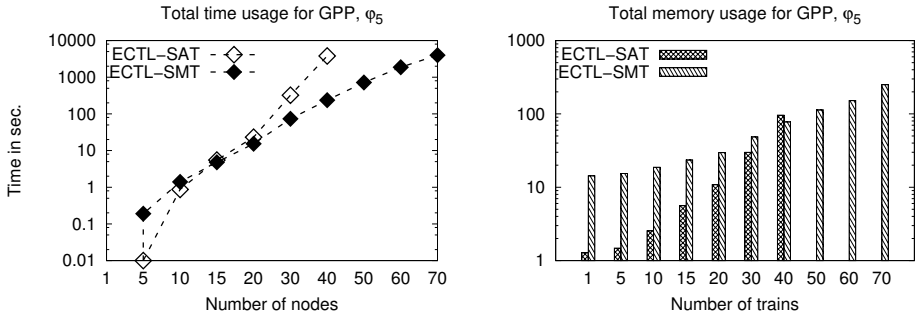


(a) (b)

FIGURE 7. A comparison of total time usage and total memory usage for the formulae φ_4.

An analysis of experimental results for formula φ_4 leads to the conclusion that also in this case SAT-BMC for ECTL uses less time and memory comparing to SMT-BMC. SMT-BMC is worse in this case because we need many short paths to verify the formula.

In Figures 8(a) and 8(b) we present a comparison of total time usage and total memory usage for the formulae φ_5.



(a) (b)

FIGURE 8. A comparison of total time usage and total memory usage for the formulae φ_5.

The SAT-BMC is able to verify the formula φ_5 for GPP with 40 nodes and the SMT-BMC is able to verify the formula for GPP with 70 nodes memory usage for the SAT-BMC is lower than for SMT-BMC.

For the tests we have used a computer equipped with Intel Core i7-5500U 2.4 GHz x 4 processor and 12 GB of RAM, running Ubuntu Linux. We have used the state of the art SAT-solver MiniSat5 [4] and Z3 SMT-solver [5].

5. Conclusions

In this paper we presented an SMT encoding for ECTL. We implemented the new method and we showed a comparison between the SAT- and SMT-based BMC methods for ECTL. The experimental results showed that, in general, SAT-based approach is better for tested systems and properties than SMT-based approach. In general, results show that the approaches are complementary, and that the SMT-based BMC approach appears to be superior for the short paths. This is a novel and interesting result, which shows that the choice of the BMC method should depend on the considered system and formula.

References

[1] C. Baier and J.-P. Katoen. *Principles of model checking*. MIT Press, 2008.

[2] A. Biere, A. Cimatti, E. Clarke, and Y. Zhu. Symbolic model checking without BDDs. In *Proceedings of the 5th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'99)*, volume 1579 of *LNCS*, pages 193–207. Springer-Verlag, 1999.

[3] Armin Biere, Alessandro Cimatti, Edmund M. Clarke, Ofer Strichman, and Yunshan Zhu. Bounded model checking. *Advances in Computers*, 58:117–148, 2003.

[4] N. Eén and N. Sörensson. MiniSat. http://minisat.se/MiniSat.html.

[5] N. Eén and N. Sörensson. MiniSat - A SAT Solver with Conflict-Clause Minimization. In *Proceedings of 8th International Conference on Theory and Applications of Satisfiability Testing(SAT'05)*, LNCS. Springer-Verlag, 2005.

[6] E. A. Emerson and E. Clarke. Using branching-time temporal logic to synthesize synchronization skeletons. *Science of Computer Programming*, 2(3):241–266, 1982.

[7] W. van der Hoek and M. Wooldridge. Model checking knowledge and time. In *Proceedings of the 9th Int. SPIN Workshop (SPIN'02)*, volume 2318 of *LNCS*, pages 95–111. Springer-Verlag, 2002.

[8] D. Peled. All from one, one for all: On model checking using representatives. In *Proceedings of the 5th Int. Conf. on Computer Aided Verification (CAV'93)*, volume 697 of *LNCS*, pages 409–423. Springer-Verlag, 1993.

[9] W. Penczek, B. Woźna, and A. Zbrzezny. Bounded model checking for the universal fragment of CTL. *Fundamenta Informaticae*, 51(1-2):135–156, 2002.

[10] B. Woźna. ACTL* properties and bounded model checking. *Fundamenta Informaticae*, 63(1):65–87, 2004.

[11] B. Woźna-Szcześniak, A. M. Zbrzezny, and A. Zbrzezny. The BMC method for the existential part of RTCTLK and interleaved interpreted systems. In *Proceedings of*

the 15th Portuguese Conference on Artificial Intelligence (EPIA'2011), volume 7026 of *LNAI*, pages 551–565. Springer-Verlag, 2011.

[12] A. Zbrzezny. Improving the translation from ECTL to SAT. *Fundamenta Informaticae*, 85(1-4):513–531, 2008.

Received: June 2017

Agnieszka M. Zbrzezny
Jan Długosz University in Częstochowa
Institute of Mathematics and Computer Science
al. Armii Krajowej 13/15, 42-200 Częstochowa, Poland
E-mail address: agnieszka.zbrzezny@ajd.czest.pl

All papers submitted to *Scientific Issues of Jan Długosz University in Częstochowa, Mathematics* should be written in English, following the LaTeX template to be found at

`http://www.scientificissues.ajd.czest.pl`.

Submissions should be made electronically to the following address:

`sci.iss.ajd@ajd.czest.pl`.

The `.tex` source file of the manuscript, along with its `.pdf` version, must be included with any submission. All figures should be converted to either `.jpg` or `.pdf` format.

A manuscript submitted to Scientific Issues of Jan Długosz University in Częstochowa is first considered by the Scientific Board. If the subject of the manuscript is not consistent with the journal profile or its text contains distinct defects, the article can be rejected.

Upon preliminary acceptance by the Scientific Board, the manuscript is sent to reviewers. The choice of reviewers is confidential. A review evaluates the manuscript and indicates unambiguously whether or not the manuscript can be published. When the required number of reviews have been received, usually two, the Editor will consider the experts' opinions and make an initial decision to accept, reject, or request a revision. In the case of the latter, the author should correct the manuscript according to reviewers' suggestions and resubmit the revised version. The Scientific Board reserves the right for making small corrections in the text of the final version of the manuscript. All manuscripts accepted for publication are prepared in both electronic and print versions.

Each submission of a manuscript to *Scientific Issues of Jan Długosz University in Częstochowa, Mathematics* implies the acceptance of the following requirements.

- Submission of a manuscript is taken to imply that it has not been previously published and that it is not being considered for publication elsewhere.
- Upon acceptance of a manuscript, the author is asked to transfer copyright to the publisher.

Please pay attention that in March 2016 the ISSN number was changed from ISSN 1896-0286 to ISSN 2450–9302.